Узлы моделирования IBM SPSS Modeler 16

Примечание

Прежде чем использовать эту информацию и продукт, описанный в ней, прочтите сведения в разделе “Уведомления” на стр. 291.

Информация о продукте

Это издание применимо к версии 16, выпуску 0, модификации 0 IBM(r) SPSS(r) и ко всем последующим выпускам и модификациям до тех пор, пока в новых изданиях не будет указано иное.

Содержание

Предисловие

IBM® SPSS Modeler - это инструментальная среда исследования данных IBM Corp., рассчитанная на работу с предприятием. SPSS Modeler помогает организациям улучшить взаимосвязи с клиентами и отдельными лицами, обеспечивая глубокое понимание данных. Организации используют приобретенные с помощью SPSS Modeler глубокие знания для сохранения выгодных заказчиков, обнаружения возможностей дополнительных покупок, привлечения новых клиентов, обнаружения ошибок, сокращения рисков и улучшений в обеспечении государственных служб.

Наглядный интерфейс SPSS Modeler дает пользователям возможность применить свой конкретный опыт в бизнесе, что способствует разработке более мощных предсказывающих моделей и сокращает время принятия решения. SPSS Modeler предлагает много способов моделирования, таких как алгоритмы предсказания, классификации, сегментации и ассоциативного обнаружения. Когда моделей IBM SPSS Modeler Solution Publisher поддерживает их распространение на уровне организации для принимающих решение сотрудников или для применения к базе данных.

О бизнес аналитике IBM

Программное обеспечение IBM для бизнес аналитики предоставляет полную, последовательную и точную информацию, которая повышает эффективность ведения бизнеса. Полный набор программного обеспечения для business intelligence, прогностической аналитики, управления финансовой эффективностью и стратегией и аналитических приложений позволяет ясно видеть текущую ситуацию, а также делать прогнозы, позволяющие предпринимать практические действия. В сочетании с решениями для конкретных отраслей, проверенной практикой и услугами бизнес аналитика IBM позволяет организациям любых размеров достигать наивысшей производительности, уверенно автоматизировать процессы принятия решений и добиться лучших результатов.

Как составная часть этого набора, программное обеспечение IBM SPSS Predictive Analytics помогает организациям предсказывать будущие события и предпринимать практические действия непосредственно на основе этих предсказаний. Коммерческие, правительственные и академические организации всего мира, полагаются на технологию IBM SPSS, обеспечивающую конкурентное преимущество в привлечении, удержании и повышении отдачи от клиентов. Включая программное обеспечение IBM SPSS в свои ежедневные операции, организации могут прогнозировать будущие события, направлять и автоматизировать решения для соответствия бизнес-целям и достигать ощутимых конкурентных преимуществ. Чтобы получить дальнейшую информацию или связаться с представителем, зайдите на http://www.ibm.com/spss.

Техническая поддержка

Техническая поддержка предоставляется клиентам, оплачивающим обновительные взносы. Пользователи могут обращаться в службу технической поддержки, если у них возникают какие-либо проблемы с использованием или установкой программного обеспечения IBM Corp.. За технической поддержкой обращайтесь на сайт IBM Corp.: http://www.ibm.com/support. При обращении за поддержкой будьте готовы назвать себя и организацию, в которой вы работаете.

Глава 1. О программе IBM SPSS Modeler

IBM SPSS Modeler - это комплект инструментов исследования данных, при помощи которого можно быстро разрабатывать прогнозные модели, использующие деловые знания и опыт, и внедрять их в деловые операции для усовершенствования процесса принятия решений. Разработанный на основе модели промышленного стандарта CRISP-DM, IBM SPSS Modeler поддерживает весь процесс исследования данных, от обработки исходных данных до получения лучших деловых результатов.

IBM SPSS Modeler предлагает ряд методов моделирования, взятых из таких областей, как обучение машин, искусственный интеллект и статистика. При помощи методов, доступных на палитре Моделирование, можно извлечь новую информацию из данных и разработать прогнозные модели. У каждого из методов есть свои сильные стороны и типы задач, для решения которых он лучше всего подходит.

SPSS Modeler можно приобрести как отдельный продукт или использовать как клиент в сочетании с сервер SPSS Modeler. Кроме того, доступен ряд дополнительных возможностей, сводка которых дается в следующих разделах. Дополнительную информацию смотрите в разделе http://www.ibm.com/software/analytics/spss/products/modeler/.

Продукты IBM SPSS Modeler

В семейство продуктов IBM SPSS Modeler и связанные с этим семейством программы входят следующие продукты:

- IBM SPSS Modeler
- сервер IBM SPSS Modeler
- IBM SPSS Modeler Administration Console
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- Адаптеры сервер IBM SPSS Modeler для IBM SPSS Collaboration and Deployment Services

IBM SPSS Modeler

SPSS Modeler - это полнофункциональная версия продукта, устанавливаемая и запускаемая на персональном компьютере. SPSS Modeler можно запустить в локальном режиме, как автономный продукт, или в распределенном режиме вместе с сервер IBM SPSS Modeler, чтобы повысить производительность на больших наборах данных.

Используя SPSS Modeler, можно быстро и интуитивно строить точные прогнозные модели, не прибегая к программированию. Используя уникальный визуальный интерфейс, можно легко визуализировать процесс анализа данных. В продукт встроены расширенные функции аналитики, при поддержке которых можно обнаруживать в данных скрытые структуры и тенденции. Можно моделировать результаты и выяснять, какие факторы на них влияют, чтобы полностью использовать деловые возможности и ограничивать риски.

SPSS Modeler доступен в двух версиях: SPSS Modeler Professional и SPSS Modeler Premium. Дополнительную информацию смотрите в разделе “Выпуски IBM SPSS Modeler” на стр. 2.

сервер IBM SPSS Modeler

SPSS Modeler пользуется архитектурой клиент - сервер, чтобы распределять требования ресурсоемких операций по мощным серверным программам, что повышает производительность для больших наборов данных.

сервер SPSS Modeler - это отдельно лицензируемый продукт, который непрерывно работает в режиме распределенного анализа на хосте сервера совместно с одной или несколькими установками IBM SPSS Modeler. При этом сервер SPSS Modeler обеспечивает высокую производительность для больших наборов данных, поскольку ресурсоемкие операции можно выполнять на сервере без скачивания данных на компьютер клиента. Кроме того, сервер IBM SPSS Modeler обеспечивает поддержку для возможностей оптимизации SQL и моделирования в базе данных, что дает дополнительный выигрыш в производительности и автоматизации.

IBM SPSS Modeler Administration Console

Modeler Administration Console - это графическая программа для управления многочисленными опциями конфигурации сервер SPSS Modeler, который также можно конфигурировать посредством файла опций. Эта прикладная программа содержит консольный пользовательский интерфейс для отслеживания и конфигурирования установок сервер SPSS Modeler installations, and is available free-of-charge сервер SPSS Modeler. Эту прикладную программу можно установить только на компьютерах Windows; однако она может управлять сервером на любой поддерживаемой платформе.

IBM SPSS Modeler Batch

Хотя обычно исследование данных - интерактивный процесс, можно также запустить SPSS Modeler из командной строки, не открывая графический интерфейс. Например, у вас могут быть продолжительные или повторяющиеся задачи, которые желательно выполнить без участия пользователя. SPSS Modeler Batch - это особая версия продукта, предоставляющая поддержку всех аналитических возможностей SPSS Modeler без вызова обычного пользовательского интерфейса. Для использования SPSS Modeler Batch требуется лицензия сервер SPSS Modeler.

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher - это инструмент, при помощи которого можно создать пакетную версию потока SPSS Modeler; такую версию можно запускать внешним механизмом времени выполнения или встроить во внешнюю прикладную программу. Этим способом можно публиковать и внедрять полные потоки SPSS Modeler для использования в средах, где SPSS Modeler не установлен. SPSS Modeler Solution Publisher распространяется в составе службы IBM SPSS Collaboration and Deployment Services - Scoring, для которой требуется отдельная лицензия. С этой лицензией вы получаете модуль времени выполнения SPSS Modeler Solution Publisher, при помощи которого можете запускать опубликованные потоки.

Адаптеры сервер IBM SPSS Modeler для IBM SPSS Collaboration and Deployment Services

Для IBM SPSS Collaboration and Deployment Services доступен ряд адаптеров, при посредстве которых SPSS Modeler и сервер SPSS Modeler могут взаимодействовать с репозиторием IBM SPSS Collaboration and Deployment Services. При этом поток SPSS Modeler, внедренный в репозиторий, доступен для совместного использования несколькими пользователями или для обращения из прикладной программы IBM SPSS Modeler Advantage тонкого клиента. Адаптер устанавливается в той системе, в которой находится репозиторий.

Выпуски IBM SPSS Modeler

SPSS Modeler доступен в следующих выпусках.

SPSS Modeler Professional

SPSS Modeler Professional содержит все инструменты, необходимые для работы с большинством типов структурированных данных, таких как трассировка поведения и взаимодействия в системах CRM, демографии, поведения покупателей и данных о продажах.

SPSS Modeler Premium

SPSS Modeler Premium - это отдельно лицензируемый продукт, расширяющий SPSS Modeler Professional для работы с такими специальными данными, как данные в аналитике объектов или социальных сетях, и с неструктурированными текстовыми данными. SPSS Modeler Premium состоит из следующих компонентов.

IBM SPSS Modeler Entity Analytics добавляет дополнительное измерение к прогностической аналитике IBM SPSS Modeler. Прогностическая аналитика пытается предсказать будущее поведение данных из прошлого, а объектная аналитика направлена на улучшение связности и согласованности текущих данных посредством устранения конфликтов идентичности в самих записях. Идентичность может относиться к индивидууму, организации, а также к любому другому объекту, для которого возможна неоднозначность. Разрешение идентичности может оказаться крайне необходимым для ряда полей, в том числе для управления отношениями с клиентами, обнаружения мошенничества, противодействия отмыванию денег или для национальной и международной безопасности.

IBM SPSS Modeler Social Network Analysis преобразует информацию о взаимосвязях в поля, характеризующие социальное поведение отдельных лиц и групп. При помощи данных, описывающих взаимосвязи, в основе которых лежат социальные сети, IBM SPSS Modeler Social Network Analysis определяет социальных лидеров, влияющих на поведение других участников сети. Кроме того, вы можете определить, какие люди наиболее подвержены влиянию других участников сети. Сочетая полученные результаты с результатами других измерений, можно создать исчерпывающие профили отдельных лиц, на которых будут основаны ваши прогнозные модели. Модели, содержащие эту социальную информацию, выполняются лучше моделей, которые ее не содержат.

IBM SPSS Modeler Text Analytics использует новейшие лингвистические технологии и обработку естественного языка (NLP) для быстрой обработки самых разнообразных неструктурированных текстовых данных, для извлечения и организации ключевых понятий и группирования этих понятий в категории. Извлеченные понятия и категории можно сочетать с существующими структурированными данными, такими как демографические, и применять к моделированию при помощи полного комплекта инструментов исследования данных IBM SPSS Modeler для получения более качественных и специализированных решений.

Документация IBM SPSS Modeler

Документация в формате электронной справки доступна через меню Справка SPSS Modeler. Сюда входит документация по SPSS Modeler, сервер SPSS Modeler и SPSS Modeler Solution Publisher, а также Руководство по прикладным программам и другие сопроводительные материалы.

Полная документация по каждому продукту (включая указания по установке) доступна в формате PDF в подпапках *\Documentation* каждого продукта DVD. Документы по установке также можно скачать с веб-сайта по адресу http://www-01.ibm.com/support/docview.wss?uid=swg27038316.

Кроме того, документация в обоих этих форматах доступна в Информационном центре SPSS Modeler по адресу http://publib.boulder.ibm.com/infocenter/spssmodl/v16r0m0/.

Документация SPSS Modeler Professional

В комплект документации SPSS Modeler Professional (включая указания по установке) входят:

- **Руководство пользователя IBM SPSS Modeler.** Общее введение в использование SPSS Modeler, в том числе о создании потоков данных, обработке пропущенных значений, построению выражений CLEM, работе с проектами и отчетами и составлению пакетов потоков для внедрения в IBM SPSS Collaboration and Deployment Services, прогнозирующие прикладные программы или IBM SPSS Modeler Advantage.
- **Узлы источников, обработки и вывода IBM SPSS Modeler.** Описания всех узлов, служащих для чтения, обработки и вывода данных в различных форматах. По существу это все узлы, кроме узлов моделирования.
- **Узлы моделирования IBM SPSS Modeler.** Описания всех узлов, служащих для создания моделей исследования данных. IBM SPSS Modeler предлагает ряд методов моделирования, взятых из таких областей, как обучение машин, искусственный интеллект и статистика.

- **Руководство по алгоритмам IBM SPSS Modeler.** Описание математических основ методов моделирования, используемых в IBM SPSS Modeler. Это руководство доступно только в формате PDF.
- **Руководство по прикладным программам IBM SPSS Modeler.** Примеры в этом руководстве служат кратким специализированным введением к тем или иным методам и технологиям моделирования. Это руководство доступно также в электронном виде в меню Справка. Дополнительную информацию смотрите в разделе “Примеры прикладных программ”.
- **Сценарии и автоматизация IBM SPSS Modeler.** Информация об автоматизации системы путем создания сценариев, включая сценарии свойств, которые могут использоваться для работы с узлами и потоками.
- **Руководство по внедрению IBM SPSS Modeler .** Информация о выполнении IBM SPSS Modeler потоков и сценариев как шагов обработки заданий под управлением IBM SPSS Collaboration and Deployment Services Deployment Manager.
- **Руководство разработчика IBM SPSS Modeler CLEF .** CLEF предоставляет возможности интеграции с программами других производителей, таких как подпрограммы обработки данных или алгоритмы моделирования, как с узлами в IBM SPSS Modeler.
- **Руководство по исследованию данных в базе данных IBM SPSS Modeler.** Информация о том, как использовать мощности вашей базы данных для повышения производительности и расширения диапазона возможностей анализа с привлечением алгоритмов от сторонних производителей.
- **Руководство администратора и руководство по производительности сервер IBM SPSS Modeler .** Информация о том, как сконфигурировать и администрировать сервер IBM SPSS Modeler.
- **Руководство пользователя по консоли администратора IBM SPSS Modeler .** Информация об установке и использовании пользовательского интерфейса консоли для мониторинга и конфигурирования сервер IBM SPSS Modeler. Консоль реализована как подключаемый модуль прикладной программы Диспетчер развертывания .
- **Руководство по CRISP-DM IBM SPSS Modeler.** Пошаговое руководство к использованию методологии CRISP-DM для исследования данных SPSS Modeler.
- **Руководство пользователя IBM SPSS Modeler Batch .** Полное руководство по использованию IBM SPSS Modeler в пакетном режиме, включая подробности выполнения в пакетном режиме и аргументы командной строки. Это руководство доступно только в формате PDF.

Документация SPSS Modeler Premium

В комплект документации SPSS Modeler Premium (включая указания по установке) входят:

- **IBM SPSS Modeler Entity Analytics Руководство пользователя.** Информация об использовании аналитики объектов совместно с SPSS Modeler, в том числе по установке и конфигурированию репозитория, узлам аналитики объектов и задачам управления.
- **IBM SPSS Modeler Social Network Analysis Руководство пользователя.** Руководство по выполнению анализа социальной сети совместно с SPSS Modeler, включая анализ групп и анализ распространения.
- **Руководство пользователя SPSS Modeler Text Analytics .** Информация об использовании аналитики текстов совместно с SPSS Modeler, в том числе по узлам исследования текстов, интерактивной инструментальной среде, шаблонам и другим ресурсам.
- **Руководство пользователя по консоли администратора IBM SPSS Modeler Text Analytics.** Информация об установке и использовании пользовательского интерфейса консоли для мониторинга и конфигурирования сервер IBM SPSS Modeler для использования совместно с SPSS Modeler Text Analytics . Консоль реализована как подключаемый модуль прикладной программы Диспетчер развертывания .

Примеры прикладных программ

Инструменты исследования данных в SPSS Modeler помогают разрешить широкий спектр деловых и организационных проблем, а примеры прикладных программ предоставляют краткие, целевые введения в конкретные методы и способы моделирования. Используемые здесь наборы данных намного меньше огромных складов данных, которыми управляют некоторые исследователи данных, но применяемые понятия и методы должны масштабироваться до реальных прикладных программ.

Обратиться к примерам можно, выбрав **Примеры прикладных программ** в меню Справка в SPSS Modeler. Файлы данных и потоки примеров устанавливаются в папке *Demos* в каталоге установки продукта. Дополнительную информацию смотрите в разделе “Папка demos”.

Примеры моделирования баз данных. Смотрите эти примеры в руководстве *IBM SPSS Modeler In-Database Mining Guide*.

Примеры сценариев. Смотрите эти примеры в руководстве *IBM SPSS Modeler Scripting and Automation Guide*.

Папка demos

Файлы данных и примеры потоков, используемые с примерами прикладных программ, устанавливаются в папке *Demos* в каталоге установки продукта. К этой папке можно также обратиться из группы программ IBM SPSS Modeler в меню Пуск Windows или, щелкнув по *Demos* в списке недавно использовавшихся каталогов в диалоговом окне Открыть файл.

Глава 2. Введение в моделирование

Модель - это набор правил, формул или уравнений, которые можно использовать для предсказания выходных данных на основании набора входных полей или переменных. Например, финансовая компания может использовать модель для предсказания рисков выдачи кредита обратившимся за ней клиентам на основании информации, уже известной о бывших клиентах.

Возможность предсказания выходных данных - это основная цель предсказательной аналитики, а понимание процесса моделирования - это ключ к использованию IBM SPSS Modeler.

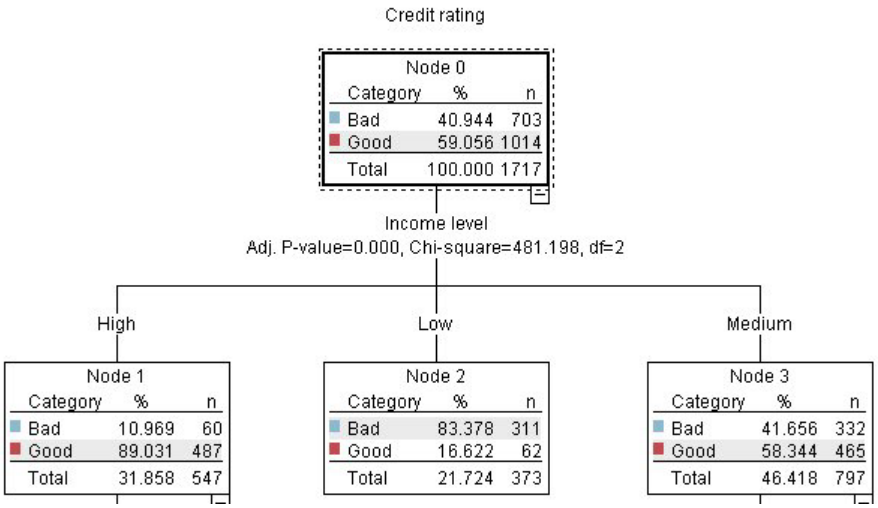


Рисунок 1. Простая модель дерева решений

В этом примере используется модель **дерева решений**, которая классифицирует записи (и предсказывает отклик), используя ряд правил решений, например:

```
IF income = Medium
AND cards <5
THEN -> 'Good'
```

Хотя в этом примере используется модель CHAID (Chi-squared Automatic Interaction Detection - автоматическое обнаружение взаимодействия хи-квадрат), он предназначен для общего введения, и большинство понятий широко применяются в других типах моделирования в IBM SPSS Modeler.

Для понимания любой модели сначала нужно понять, какие данные в нее поступают. В этом примере данные содержат информацию о клиентах банка. Используются следующие поля:

Имя поля	**Описание**
Credit_rating	Кредитный рейтинг: 0=Плохой, 1=Хороший, 9=значения отсутствия
Age	Возраст в годах
Income	Уровень дохода: 1=Низкий, 2=Средний, 3=Высокий
Credit_cards	Количество кредитных карт у клиента: 1=меньше пяти, 2=пять или больше
Education	Уровень образования: 1=Высшая школа, 2=Колледж
Car_loans	Количество выданных ссуд на покупку автомобиля: 1=Не было или одна, 2=Две или больше

Банк поддерживает базу данных с хронологической информацией о клиентах, которые брали ссуды в этом банке, в частности, выполняли ли они свои обязательства (Кредитный рейтинг = Хороший) или отказывались от платежей (Кредитный рейтинг = Плохой). Используя эти существующие данные, банк хочет построить модель, которая позволит предсказать, насколько вероятно, что будущие обращения за ссудами приведут к неплатежам.

Используя модель дерева решений, вы можете проанализировать характеристики двух групп клиентов и предсказать вероятность отказа от платежей по ссуде.

В этом примере используется поток с именем *modelingintro.str*, доступный в папке *Demos* в подпапке *streams*. Файл данных - это *tree_credit.sav*. Дополнительную информацию смотрите в разделе “Папка demos” на стр. 5.

Давайте посмотрим на этот поток.

1. В главном меню выберите:
   Файл > Открыть поток
2. Щелкните по золотому значку слепка на панели инструментов диалогового окна Открыть и выберите папку Demos.
3. Дважды щелкните по папке *streams*.
4. Дважды щелкните по файлу с именем *modelingintro.str*.

Построение потока

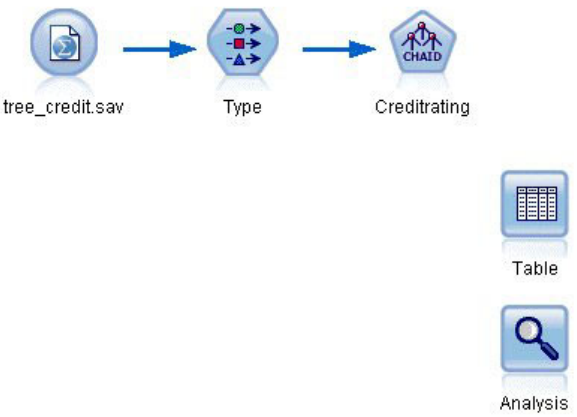


Рисунок 2. Поток моделирования

Для построения потока, который будет создавать модель, нам нужно по крайней мере три элемента:

- Узел источника, читающий данные с некоторого внешнего источника, в данном случае - из файла данных IBM SPSS Statistics.
- Узел источника или дополнительный узел Тип, где задаются свойства полей, такие как уровень измерения (тип данных, содержащихся в поле) и роль каждого поля (входное поле или поле назначения для моделирования).
- Узел моделирования, генерирующий слепок модели, когда запущен поток.

В этом примере мы используем узел моделирования CHAID. CHAID (Chi-squared Automatic Interaction Detection - автоматическое обнаружение взаимодействия хи-квадрат) - это метод классификации для построения деревьев решений с использованием конкретного типа статистики, известного как статистика хи-квадрат, для нахождения оптимальных точек расщепления в дереве решений.

Если уровни измерений заданы на узле источника, отдельный узел Тип можно исключить. Функционально результат будет таким же.

У этого потока есть также узлы Таблица и Анализ, которые будут использоваться для просмотра результатов скоринга после создания слепка модели и добавления его к потоку.

Узел источника Файл статистики читает данные в формате IBM SPSS Statistics из файла данных *tree_credit.sav*, установленного в папке *Demos*. (Специальная переменная с именем *$CLEO_DEMOS* используется для указания на положение этой папки в текущей установке IBM SPSS Modeler. Это обеспечивает правильный путь независимо от папки или версии текущей установки).

Рисунок 3. Чтение данных при помощи узла источника Файл статистики

Узел Тип задает **уровень измерения** для каждого поля. Уровень измерения - это категория, обозначающая тип данных в поле. Наш файл данных использует три разных уровня измерения.

Количественное поле (такое как поле *Возраст*) содержит количественные численные значения, а у **Номинального** поля (такого как поле *Кредитный рейтинг*) есть два или более отдельных значения, например, *Плохой*, *Хороший* или *Нет кредитной истории*. **Порядковое** поле (такое как поле *Уровень дохода*) описывает данные с несколькими отдельными значениями, для которых есть естественный присущий им порядок, в данном случае - *Низкий*, *Средний* и *Высокий*.

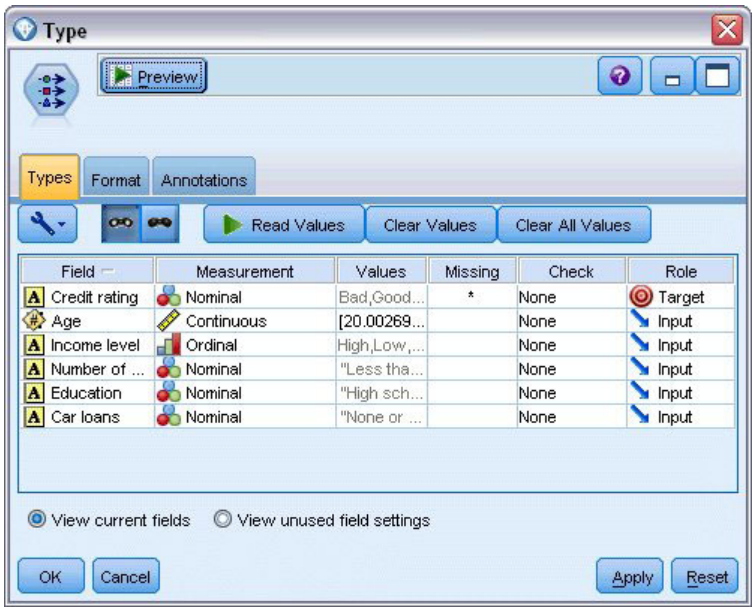


Рисунок 4. Задание полей назначения и входных полей на узле Тип

Для каждого поля узел Тип задает также **роль**, которую каждое поле играет при моделировании. Роль *Назначение* задается для поля *Кредитный рейтинг*, которое обозначает, выполнит ли конкретный клиент свои обязательства по кредиту. Это **назначение** моделирования, поле, значение в котором мы хотим предсказать.

Для других полей задается роль *Входные* поля. Входные поля иногда называют **предикторами**, то есть полями, значения которых используются для алгоритма моделирования, чтобы предсказать значение в поле назначения.

Узел моделирования CHAID генерирует модель.

На вкладке Поля узла моделирования выбирается опция **Использовать предварительно определенные роли**, что означает использование поля назначения и входных полей, заданных на узле Тип. В этом месте мы могли бы изменить роли полей, но для этого примера будем использовать их, как есть.

1. Откройте вкладку Опции построения.

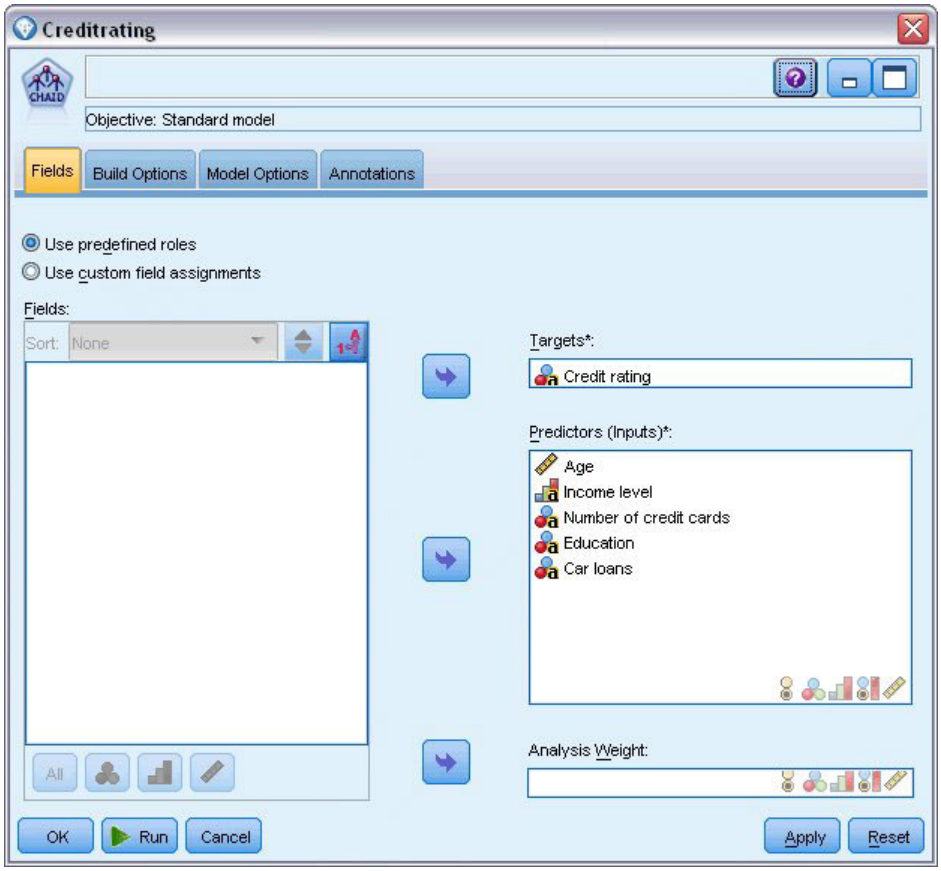


Рисунок 5. Вкладка Поля узла Моделирование CHAID

Здесь есть несколько опций, с помощью которых можно задать тип нужной модели для построения.

Мы хотим получить еще не применявшуюся модель, поэтому будем использовать опцию по умолчанию **Построить новую модель**.

Также мы хотим получить просто одну стандартную модель дерева решений без каких-то усовершенствований, поэтому оставим опцию цели по умолчанию - **Построить одно дерево**.

Хотя дополнительно мы можем запустить сеанс интерактивного моделирования, позволяющий уточнить модель, в этом примере генерируется модель с использованием режима по умолчанию - **Построить модель**.

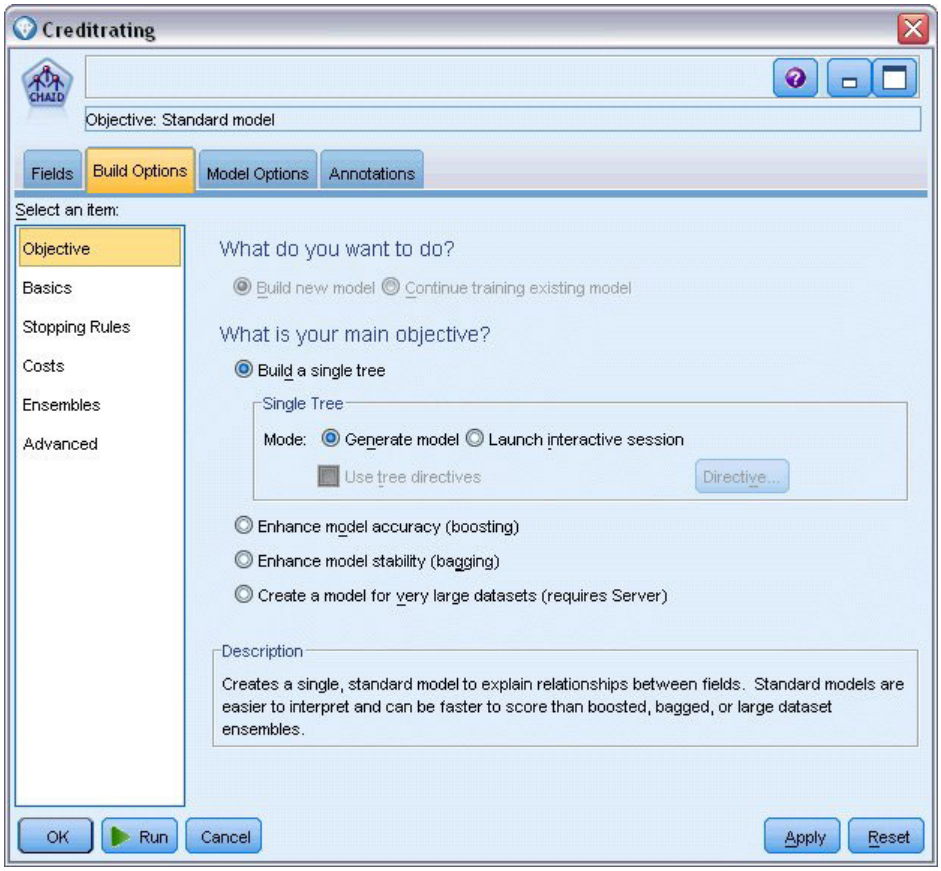


Рисунок 6. Вкладка Опции построения узла Моделирование CHAID

В этом примере мы хотим построить понятное и простое дерево, поэтому ограничим рост дерева минимальным числом наблюдений для родительских и дочерних узлов.

2. На вкладке Опции построения выберите **Правила остановки** слева на панели навигатора.
3. Выберите опцию **Использовать абсолютное значение**.
4. Задайте для параметра **Минимальное число записей в родительской ветви** значение 400.
5. Задайте для параметра **Минимальное число записей в дочерней ветви** значение 200.

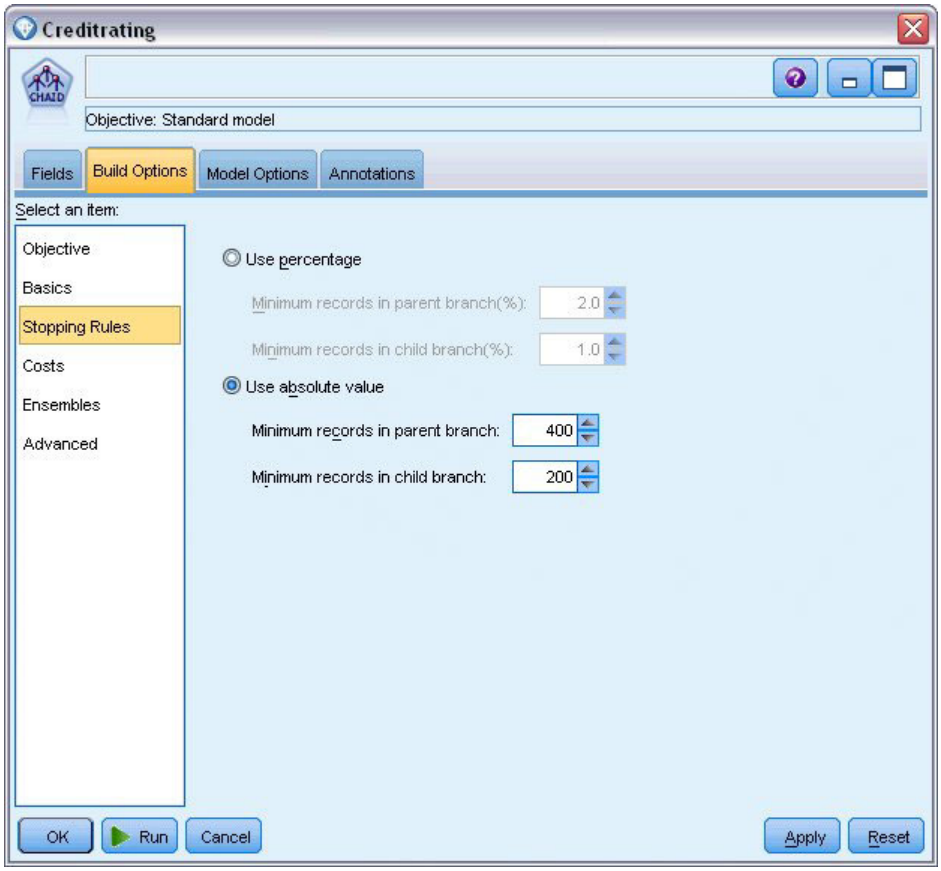


Рисунок 7. Задание критериев остановки для построения дерева решений

Для этого примера значения всех других опций мы можем использовать по умолчанию, поэтому нажмите кнопку **Выполнить**, чтобы создать модель. (Возможные варианты: щелкните правой кнопкой мыши по узлу и в контекстном меню выберите **Выполнить** или выберите узел и перейдите в меню Инструменты к пункту **Выполнить**).

Просмотр модели

После завершения выполнения слепок модели добавляется на палитру Модели в верхнем правом углу окна прикладных программ, а также размещается на холсте потока со ссылкой на узел моделирования, где он был создан. Для просмотра подробностей модели щелкните правой кнопкой мыши по слепку модели и выберите **Просмотр** (на палитре моделей) или **Изменить** (на холсте).

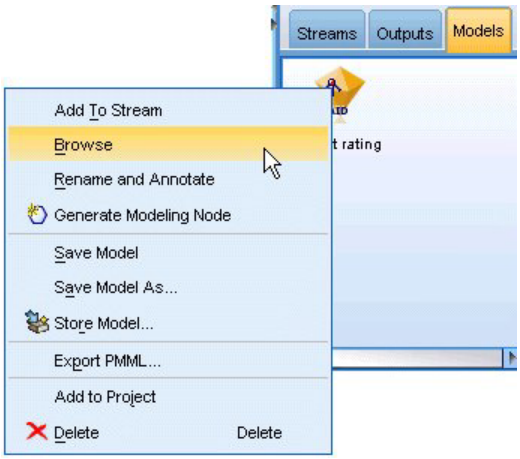


Рисунок 8. Палитра моделей

В случае слепка CHAID на вкладке Модель подробности выводятся в виде набора правил, и важно, что приводится ряд правил, которые можно использовать для назначения индивидуальных записей дочерним узлам на основе значений различных входных полей.

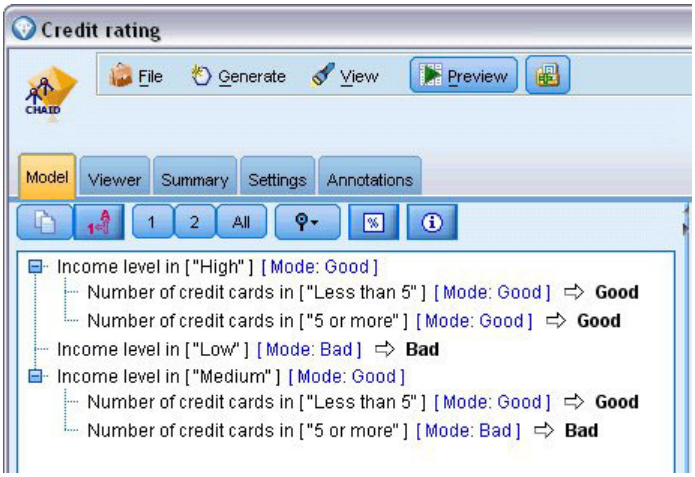


Рисунок 9. Слепок модели CHAID, набор правил

Для каждого конечного узла дерева решений, то есть для такого узла дерева, который нельзя расщепить дальше, возвращается предсказание *Хороший* или *Плохой*. В каждом случае предсказание определяется **модой**, или наиболее частым ответом, для записей, попавших на этот узел.

Справа от набора правил на вкладке Модель выводится диаграмма Важности, которая показывает относительную важность каждого предиктора при оценке модели. Отсюда легко видеть, что *Уровень дохода* - наиболее важный предиктор в данном случае, и есть еще только один важный фактор *Количество кредитных карт*.

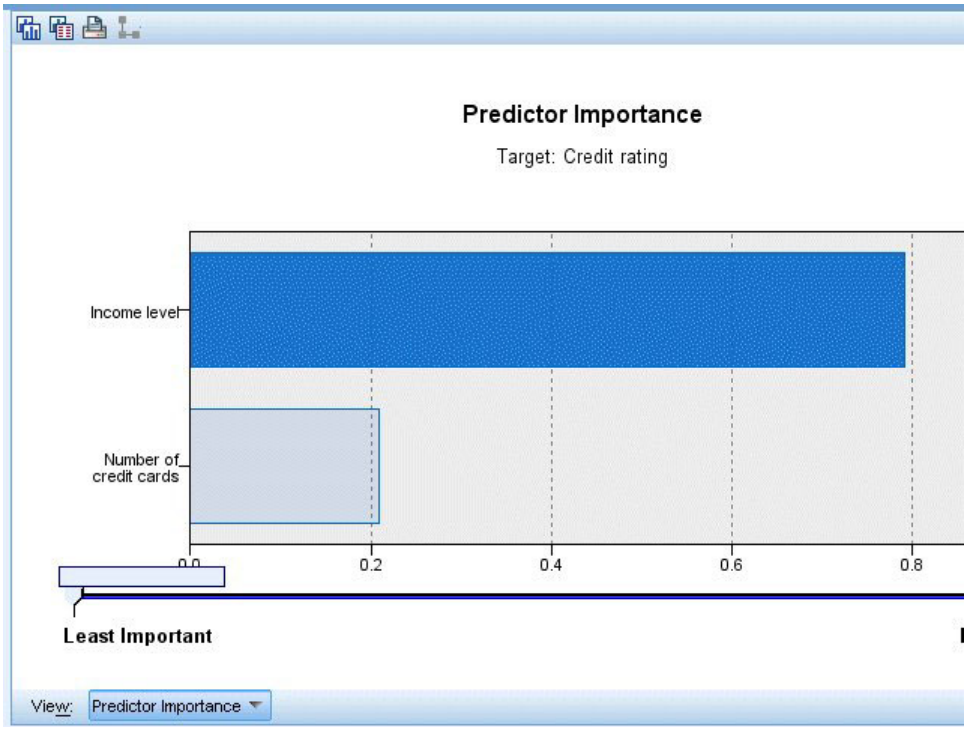


Рисунок 10. Диаграмма важности предикторов

На вкладке Программа просмотра слепка модели та же модель выводится в форме дерева с узлом в каждой точке решения. Используйте управляющие элементы масштабирования на панели инструментов, чтобы приблизить конкретный узел или уменьшить детализацию и увидеть большую часть дерева.

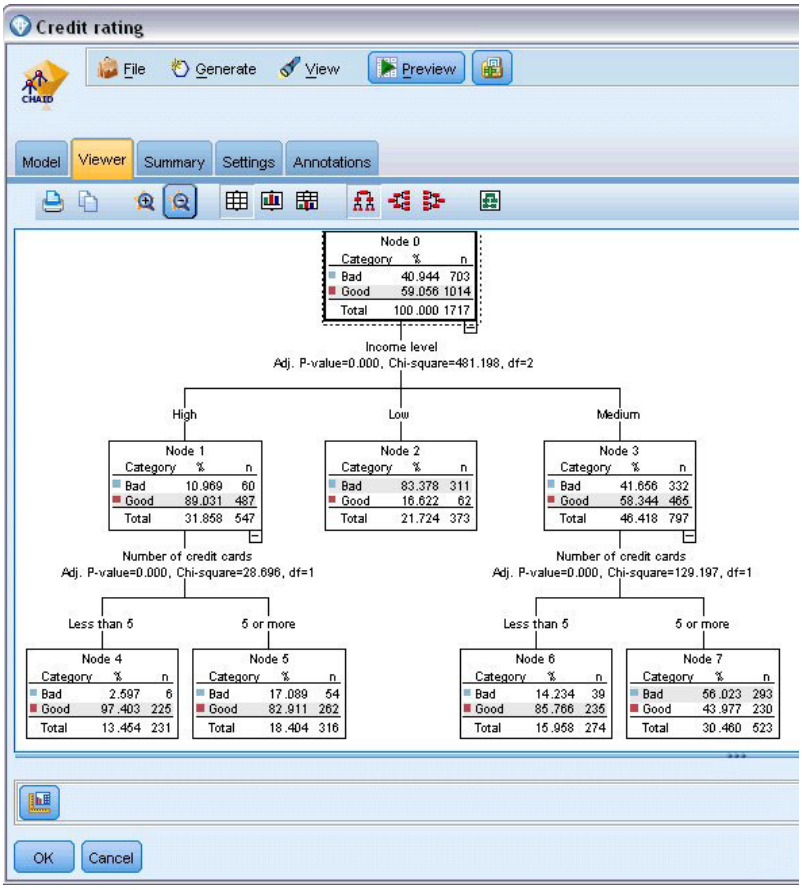


Рисунок 11. Вкладка Программа просмотра в слепке модели, выбор уменьшенной детализации

В верхней части дерева первый узел (Узел номер 0) дает сводку по всем записям в наборе данных. Более 40% наблюдений в наборе данных классифицируются как плохие (рискованные). Это очень большая часть, поэтому посмотрим, может ли дерево подсказать, какие факторы могут быть за это ответственны.

Видно, что первое расщепление производится по показателю *Уровень дохода*. Записи, для которых уровень дохода принадлежит категории *Низкий*, назначаются узлу 2, и неудивительно, что эта категория содержит максимальную процентную долю лиц, не выполняющих своих обязательств по кредитам. Это с очевидностью приводит к выводу, что клиенты в этой категории связываются с наибольшим риском.

Однако на самом деле 16% клиентов в этой категории *не* относятся к неплательщикам, то есть предсказание не всегда будет правильным. Никакая модель не может предсказать каждый отклик, но хорошая модель должна позволить вам предсказать *наиболее вероятный* отклик для каждой записи на основе доступных данных.

Аналогично, если посмотреть на клиентов с самым высоким доходом (Узел номер 1), мы увидим, что абсолютное большинство клиентов связаны с наименьшим риском (89%). Но более десятой части из них тоже не выполняли обязательств. Можно ли уточнить критерии, чтобы минимизировать для них риск?

Посмотрим, как модель разделила этих клиентов на две подкатегории (Узлы 4 и 5) на основании количества имеющихся у них кредитных карт. Если давать взаймы только клиентам с высоким доходом, у которых меньше пяти кредитных карт, успешность операции повысится с 89% до 97% и даже больше.

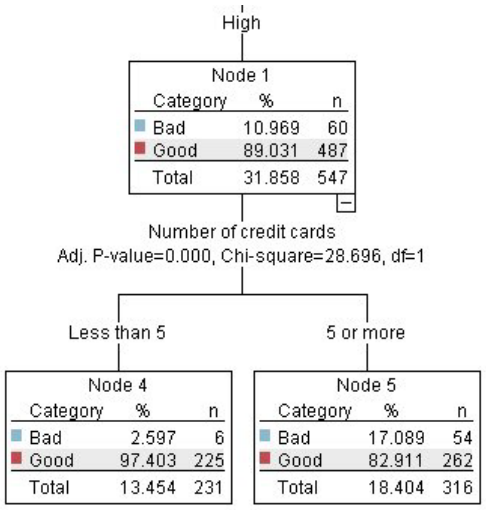


Рисунок 12. Представление дерева для клиентов с высоким доходом

Но что можно сказать о клиентах в категории дохода Средний (Узел 3)? Они более явно разделены между рейтингами Хороший и Плохой.

Нам и здесь могут помочь подкатегории (Узлы 6 и 7). На этот раз, ограничив займы только клиентами, у которых меньше пяти кредитных карт, мы увеличим показатель Хороших операций с 58% до 85%, существенное повышение.

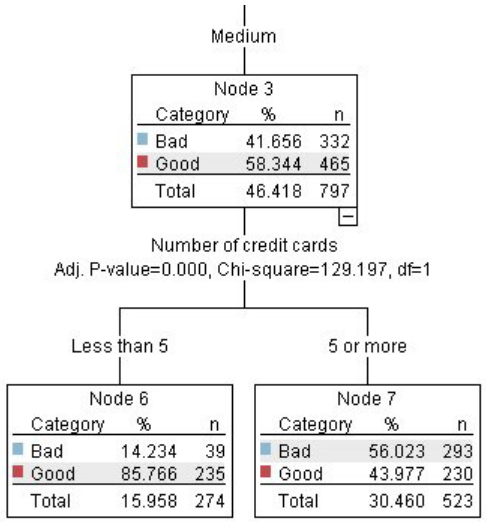


Рисунок 13. Представление дерева для клиентов со средним доходом

Итак, мы выяснили, что каждая запись, представляющая входную информацию модели, будет назначена конкретному узлу, и ей будет назначено предсказание *Хороший* или *Плохой* на основании наиболее частого отклика для этого узла.

Этот процесс назначения предсказаний индивидуальным записям известен как **скоринг**. Проводя скоринг для тех же записей, по которым оценивалась модель, мы можем выяснить, насколько точно он выполняется на данных обучения, то есть данных, для которых известен результат. Посмотрим, как это делается.

Оценка модели

Мы просмотрели модель, чтобы понять, как работает скоринг. Но для оценки, *насколько точно* от работает, нужно оценить некоторые записи и сравнить предсказанные моделью отклики с действительными результатами. Мы собираемся исследовать те же записи, которые использовались для оценки модели, что позволит сравнить наблюдаемые и предсказанные отклики.

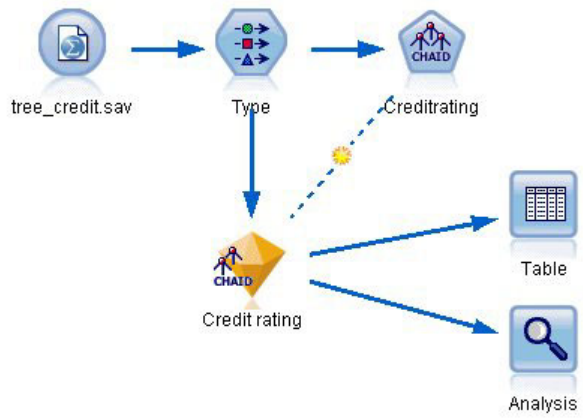
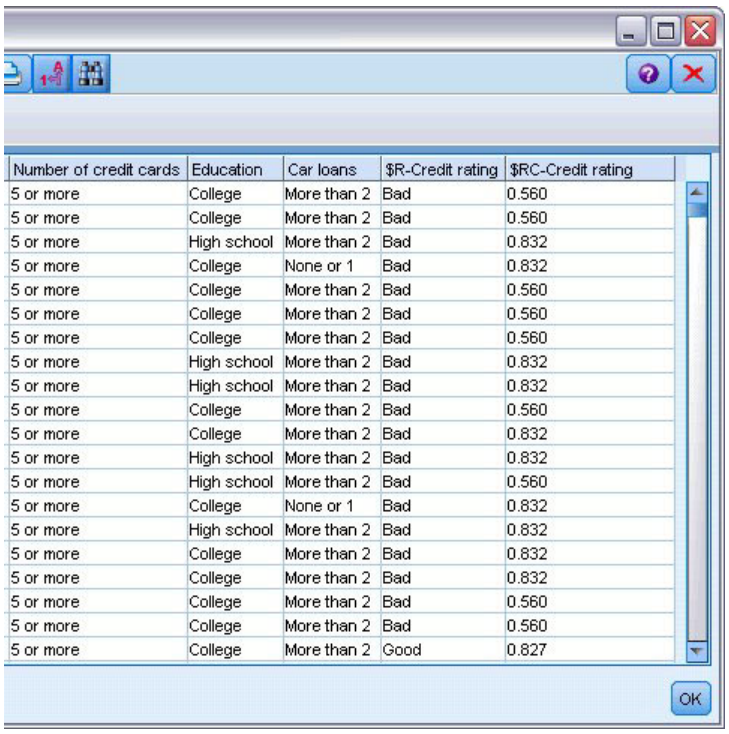


Рисунок 14. Присоединение слепка модели к узлу выходных данных для оценки модели

1. Чтобы увидеть оценки или предсказания, присоедините узел Таблица к слепку модели, дважды щелкните по узлу Таблица и нажмите кнопку **Выполнить**.

   Таблица покажет предсказанные оценки в поле с именем *$R-Кредитный рейтинг*, которое было создано моделью. Эти значения можно сравнить с исходным полем *Кредитный рейтинг*, которое содержит настоящие отклики.

   По договоренности имена сгенерированных при скоринге полей состоят из имени поля назначения со стандартным префиксом, таким как *$R-* для предсказаний или *$RC-* для значений достоверности. Модели разных типов используют разные наборы префиксов. **Значение достоверности** - это собственная оценка модели в диапазоне от 0,0 до 1,0, насколько точно предсказано каждое значение.



Number of credit cards	Education	Car loans	$R-Credit rating	$RC-Credit rating
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	High school	More than 2	Bad	0.832
5 or more	College	None or 1	Bad	0.832
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	High school	More than 2	Bad	0.832
5 or more	High school	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.832
5 or more	High school	More than 2	Bad	0.832
5 or more	High school	More than 2	Bad	0.560
5 or more	College	None or 1	Bad	0.832
5 or more	High school	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Good	0.827

OK

Рисунок 15. Таблица, показывающая сгенерированные оценки и значения достоверности

Как и ожидалось, предсказанное значение совпадает с фактическими откликами для многих записей, но не для всех. Причина этого в том, что у каждого конечного узла CHAID есть смесь откликов. Предсказание совпадает с *самым общим* откликом, но оно неправильно для всех остальных откликов этого узла. (Вспомним о 16%-ном меньшинстве клиентов с низким доходом, которые не отказывались выполнять обязательства по кредиту).

Для исключения этой ситуации мы можем продолжить расщепление дерева на всё меньшие и меньшие ветви, пока на каждом узле не окажется по 100% абсолютно *Хороших* или абсолютно *Плохих* клиентов без примеси других откликов. Но такая модель может быть чрезвычайно усложненной и скорее всего не будет хорошо обобщаться на другие наборы данных.

Чтобы точно подсчитать, сколько есть правильных предсказаний, мы можем пройти по таблице и учесть количество записей, для которых значение в предсказанном поле *$R-Кредитный рейтинг* совпадает со значением в поле *Кредитный рейтинг*. К счастью, есть гораздо более простой способ - использовать узел Анализ, который делает это автоматически.

2. Соедините слепок модели с узлом Анализ.
3. Дважды щелкните по узлу Анализ и нажмите кнопку **Выполнить**.

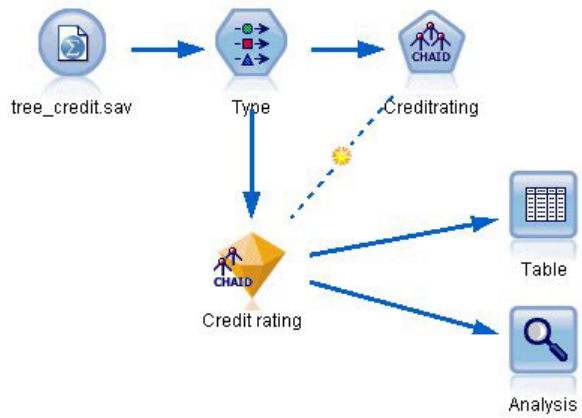


Рисунок 16. Присоединение узла Анализ

Анализ показывает, что для 1899 из 2464 записей (более 77%) предсказанное моделью значение совпадает с действительным откликом.

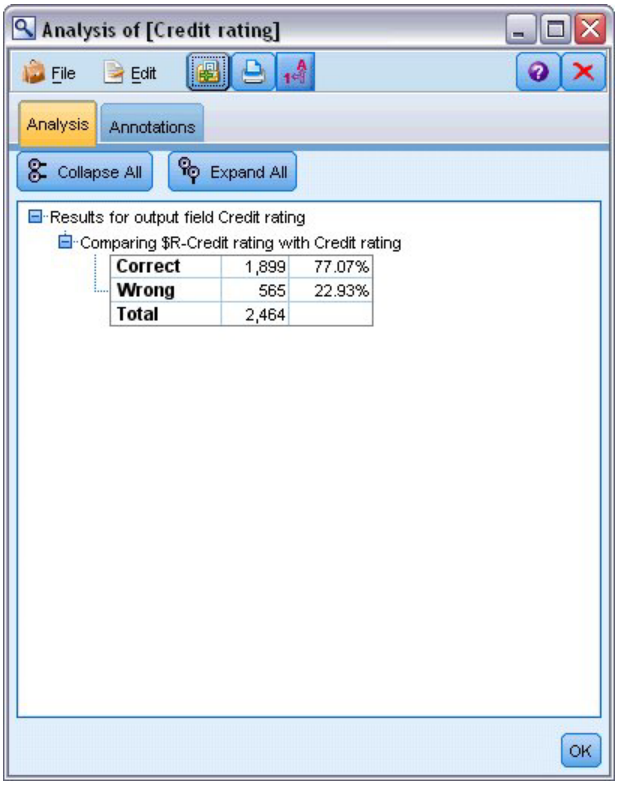


Рисунок 17. Результаты анализа, сравнивающие наблюдаемые и предсказанные отклики

Этот результат ограничен тем фактом, что оцениваемые записи были теми же, что использовались для оценки самой модели. В реальной ситуации можно было использовать узел Разделение, чтобы разбить данные на две отдельные выборки, для обучения и оценки.

Используя один раздел выборки для генерирования модели и другой - для ее испытания, можно получить гораздо более точный показатель, насколько хорошо модель обобщается на другие наборы данных.

Узел Анализ помогает испытать модель на записях, для которых мы уже знаем фактический результат. Следующая стадия иллюстрирует, как можно использовать модель, чтобы оценить записи, для которых мы

не знаем выходных значений. Например, сюда могут быть включены потенциальные клиенты, которые еще не работают с банком, но представляют из себя будущих получателей рекламной рассылки.

Скоринг записей

Ранее мы проводили скоринг тех же записей, которые использовались для оценки модели, чтобы понять, насколько точно построена модель. Теперь мы собираемся оценить другой набор записей, отличный от использованного для создания модели. Это цель моделирования с использованием поля назначения: изучить записи, для которых известны выходные данные, чтобы идентифицировать структуры, которые позволят предсказать еще не известные выходные данные.

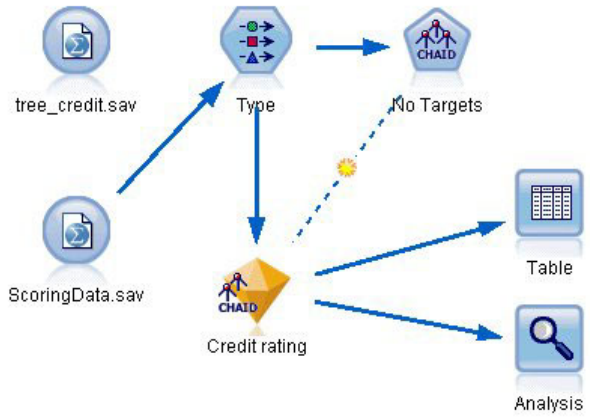


Рисунок 18. Присоединение новых данных для скоринга

Можно изменить узел источника Файл статистики для указания на другой файл данных или добавить новый узел источника, читающий данные, которые вы хотите оценить. В любом случае новый набор данных должен содержать те же входные поля, что использовала модель (*Возраст*, *Уровень дохода*, *Образование* и так далее), но не поле назначения *Кредитный рейтинг*.

Другой вариант - добавить слепок модели к любому потоку, включающему в себя ожидаемые входные поля. Тип источника (чтение из файла или базы данных) не имеет значения, если имена и типы полей совпадают с используемыми в модели.

Можно сохранить также слепок модели как отдельный файл, экспортировать модель в формате PMML для использования с другими прикладными программами, использующими этот формат, или сохранить эту модель в репозитории IBM SPSS Collaboration and Deployment Services, который на уровне предприятия обеспечивает внедрение, скоринг и управление моделями.

Сама модель независимо от используемой инфраструктуры работает одинаково.

Итог

Этот пример демонстрирует основные шаги по созданию, исследованию качества и скорингу модели.

- Узел моделирования оценивает модель, изучая записи, для которых известны выходные данные, и создает слепок модели. Иногда это называется обучением модели.
- Слепок модели можно добавить к любому потоку с ожидаемыми полями для скоринга записей. По скорингу записей, для которых известны выходные данные (например, для существующих клиентов), можно оценить, насколько хорошо все работает.
- После того, как вы будете удовлетворены качеством модели, можно оценивать новые данные (например, возможных клиентов) для предсказания их отклика.

- Данные, которые использовались для обучения или оценки модели, можно назвать аналитическими или хронологическими данными; данные скоринга можно назвать также операционными.

Глава 3. Обзор моделирования

Обзор узлов моделирования

IBM SPSS Modeler предлагает ряд методов моделирования, взятых из таких областей, как обучение машин, искусственный интеллект и статистика. При помощи методов, доступных на палитре Моделирование, можно извлечь новую информацию из данных и разработать прогнозные модели. У каждого из методов есть свои сильные стороны и типы задач, для решения которых он лучше всего подходит.

Руководство по прикладным программам IBM SPSS Modeler предоставляет примеры многих таких методов вместе с общей информацией о процессе моделирования. Это руководство доступно и в виде оперативного учебного пособия, и в формате PDF. Дополнительную информацию смотрите в разделе “Примеры прикладных программ” на стр. 4

Методы моделирования разделены на три категории:

- Классификация
- Взаимосвязь
- Сегментация

Модели классификации

Модели классификации используют значения одного или нескольких **входных** полей, чтобы предсказать значения одного или нескольких выходных полей, или полей **назначения**. Некоторые примеры таких способов следующие: деревья решений (алгоритмы дерева C&R, QUEST, CHAID и C5.0), регрессии (линейная, логистическая, обобщенная линейная и Кокса), нейронные сети, модели опорных векторов и Байесовские сети.

Модели классификации позволяют организациям предсказать известные результаты, например, купит ли покупатель товар или уйдет без покупки, или похожа ли транзакция на известный шаблон мошенничества. Способы моделирования включают в себя компьютерное обучение, вывод правил по индукции, идентификацию подгрупп, статистические способы и генерирование нескольких моделей.

Узлы классификации

Узел автоклассификации создает и сравнивает несколько различных моделей для двоичных выходных данных (да или нет, уйдет клиент или останется и так далее), что позволяет выбрать лучший подход для данного анализа. Поддерживается несколько алгоритмов моделирования, что делает возможным выбор желательных для использования способов, конкретных опций для каждого из них и критериев сравнения результатов. Этот узел генерирует набор моделей на основе заданных опций и ранжирует лучших кандидатов в соответствии с заданными вами критериями.

Узел автонумерации оценивает и сравнивает модели для выходных данных в количественном числовом диапазоне при помощи нескольких разных способов. Этот узел работает аналогично другим узлам автоклассификации, допуская выбор алгоритмов для использования и экспериментирование с несколькими комбинациями опций при одном проходе моделирования. Поддерживаемые алгоритмы включают в себя нейросети, дерево C&R, CHAID, линейную регрессию, обобщенную линейную регрессию и механизмы опорных векторов (support vector machines, SVM). Модели можно сравнивать на основе корреляции, относительной ошибки или числа используемых переменных.

Узел дерева классификации и регрессии (Classification and Regression, C&R) генерирует дерево решений, позволяющее предсказывать или классифицировать будущие наблюдения. Этот метод использует рекурсивное разделение, чтобы расщепить обучающие записи на сегменты, на каждом шаге минимизируя неоднородность, причем узел дерева считается “чистым”, если все 100% наблюдений в узле попадают в конкретную категорию поля назначения. Входные поля и поля назначения могут быть из числового диапазона или категориальными (номинальными, порядковыми или флагами); все расщепления бинарны (только две подгруппы).

Узел QUEST предоставляет метод бинарной классификации для построения деревьев решений, разработанный для уменьшения времени обработки, требуемого для анализа больших деревьев C&R, при одновременном подавлении обнаруженного в способах деревьев классификации предпочтения входных полей, допускающих больше расщеплений. Входные поля могут быть в числовом диапазоне (количественными), но поле назначения должно быть категориальным. Все расщепления бинарные.

Узел CHAID генерирует деревья решений, используя статистику хи-квадрат для определения оптимальных расщеплений. В отличие от узлов дерева C&R и QUEST, CHAID может генерировать не только бинарные деревья, то есть у некоторых расщеплений может быть больше двух ветвей. Входные поля и поле назначения могут быть количественными (числовой диапазон) или категориальными. Исчерпывающий CHAID - это модификация метода CHAID, при котором проделывается более тщательная работа по изучению всех возможных расщеплений для каждого предиктора, но это требует больше времени для вычислений.

Узел C5.0 строит или дерево решений, или набор правил. Эта модель работает, разделяя выборку на основании значения в поле, дающего максимальный информационный выигрыш на каждом уровне. Поле назначения должно быть категориальным. Разрешено несколько разделений на подгруппы, и таких подгрупп может быть больше двух.

Узел списка решений определяет подгруппы или сегменты, которые показывают более высокое или более низкое правдоподобие для данного бинарного результата по сравнению с полной совокупностью. Например, вы могли бы искать клиентов с низкой вероятностью оттока или с высокой вероятностью отклика на кампанию. Вы можете включить свои знания о бизнесе в модель, добавляя свои собственные пользовательские сегменты и параллельно просматривая альтернативные модели, чтобы сравнить результаты. Модели списка решений состоят из списка правил, в котором каждое правило имеет условие и следствие. Правила применяются по очереди, и первое подходящее правило определяет результат.

Модели линейной регрессии предсказывают значения непрерывного целевого поля на основе линейных взаимосвязей между целевым полем и одним или несколькими предикторами.

Узел PCA/фактора предоставляет мощные средства сокращения числа данных для уменьшения сложности ваших данных. Анализ главных компонентов (principal components analysis, PCA) находит линейные комбинации входных полей, которыми главным образом определяются изменения в целом наборе полей, где компоненты ортогональны друг другу. Факторный анализ направлен на выявление скрытых факторов, объясняющих структуру корреляций в наборе наблюдаемых полей. Цель обоих подходов - найти небольшое количество производных полей, которые эффективно суммируют информацию исходного набора входных полей.

Узел выбора возможностей изучает входные поля на возможность удаления, основываясь на наборе критериев (таких как процентная доля пропущенных значений); затем этот узел ранжирует важность оставшихся полей по отношению к заданному полю назначения. Например, если у набора данных сотни потенциальных входных полей, какие из них потенциально наиболее полезны при моделировании исхода лечения пациента?

Дискриминантный анализ делает более строгие предположения, чем логистическая регрессия, но он может быть ценной альтернативой или дополнением к анализу логистической регрессии, когда эти предположения оказываются правильными.

Логистическая регрессия - это статистический метод для классификации записей на основании значений входных полей. Она аналогична линейной регрессии, но логистическая регрессия использует категориальные поля назначения вместо численных.

Обобщенная линейная модель расширяет общую линейную модель, так что зависимая переменная считается линейно связанной с факторами и ковариатами через заданную функцию связи. Более того, модель допускает наличие у зависимой переменной распределения, отличающегося от нормального. Она включает в себя функциональные возможности большого количества статистических моделей, в том числе линейной регрессии, логистической регрессии, логлинейных моделей для количества данных и интервал-цензурированных моделей выживания.

Обобщенная линейная смешанная модель (generalized linear mixed model, GLMM) обобщает линейную модель таким образом, что у значений назначения может быть отличное от нормального распределение и оно будет линейно связано с факторами и ковариатами через задаваемую функцию связи, так что наблюдения могут быть скоррелированными. Обобщенные линейные смешанные модели включают широкий набор моделей, начиная от простой линейной регрессии и кончая сложными многоуровневыми моделями для не нормально распределенных данных с повторными измерениями.

Узел регрессии Кокса позволяет построить модель дожития для данных времени-до-события в присутствии цензурируемых записей. Эта модель создает функцию дожития, которая предсказывает вероятность, что изучаемое событие произойдет в данное время (*t*) для данных значений входных переменных.

Узел механизма опорных векторов (Support Vector Machine, SVM) позволяет классифицировать данные по одной или двум группам без переобучения. SVM хорошо работает с широкими наборами данных, в частности, в случае очень большого числа входных полей.

Узел Байесовская сеть позволяет построить вероятностную модель, комбинируя наблюдаемые и записанные сведения с очевидными с точки зрения здравого смысла данными, чтобы установить правдоподобие возникновения событий. Этот узел в основном работает с усиленными деревом наивными байесовскими сетями (Tree Augmented Naïve Bayes, TAN) и полными марковскими сетями, которые изначально используются для классификации.

Узел Самообучаемая модель откликов (Self-Learning Response Model, SLRM) позволяет построить модель, в которой одно новое наблюдение или всего несколько наблюдений могут быть использованы для повторной оценки модели без необходимости повторного обучения модели с использованием всех данных.

Узел временных рядов оценивает экспоненциальное сглаживание, а также одномерные и многомерные модели авторегрессии и проинтегрированного скользящего среднего (Autoregressive Integrated Moving Average, ARIMA) для временных рядов и создает прогнозы будущего выполнения. Предшественником узла временных рядов всегда должен быть узел Интервалы времени.

Узел *k* ближайших соседей (k-Nearest Neighbor, KNN) связывает новое наблюдение с категорией или значением *k* объектов, ближайших к нему в пространстве предикторов, где *k* - это целое число. Подобные наблюдения близки друг к другу, а непохожие наблюдения, наоборот, удалены друг от друга.

Модели связывания

Модели связывания находят структуры в ваших данных, в которых один или несколько объектов (таких как события, покупки или атрибуты) связаны с одним или несколькими другими объектами. Модели конструируют наборы правил, определяющие эти взаимосвязи. Здесь поля среди данных могут быть и входными полями, и полями назначения. Вы могли бы найти эти связи вручную, но алгоритмы правил связывания делают это гораздо быстрее и могут изучить более сложные структуры. Модели Априори и CARMA - это примеры использования таких алгоритмов. Еще один тип модели связывания - это модель обнаружения последовательностей, которая находит последовательные шаблоны в структурированных по времени данных.

Модели связывания наиболее полезны при предсказании нескольких выходных значений, например, покупатель, купивший товар X, купил также Y и Z. Модели связывания связывают конкретное следствие (например, решение что-либо купить) с набором условий. Преимущество алгоритмов правил связывания перед более стандартными алгоритмами дерева решений (C5.0 и деревья C&R) состоит в том, что связывание может существовать между любыми атрибутами. Алгоритм дерева решений построит правила с возможностью одного следствия, в то время как алгоритмы связывания стараются найти несколько правил, у каждого из которых может быть отдельное следствие.

Узлы связывания

Узел Априори извлекает набор правил из данных, выделяя правила с наибольшим информационным содержимым. Узел Априори предлагает пять различных способов выбора правил и использует сложные схемы индексирования для эффективной обработки больших наборов данных. Для больших задач узел Априори обычно быстрее при обучении; у него нет произвольного ограничения количества правил, которые можно сохранить, и он может обрабатывать правила с количеством предварительных условий до 32. Для узла Априори требуются категориальные входные и выходные поля, он был оптимизирован для полей такого типа и показывает с ними высокую производительность.

Модель CARMA извлекает из данных набор правил, не требуя, чтобы вы задавали входные или выходные поля. В отличие от узла Априори, узел CARMA предлагает параметры построения для поддержки правил (поддержка относится и к антецедентам, и к консеквентам), а не только для поддержки антецедентов. Это означает, что сгенерированные правила можно использовать в более широком наборе прикладных программ, например, чтобы найти список продуктов или услуг (антецедентов), консеквент которых - это товар, который вы хотите продвигать в этом летнем сезоне.

Узел последовательности обнаруживает правила связывания для последовательных или зависящих от времени данных. Последовательность - это список наборов элементов с тенденцией появления в предсказуемом порядке. Например, покупатель, который приобрел лезвия и лосьон после бритья, с большой вероятностью в следующий раз купит крем для бритья. Узел последовательности основан на алгоритме правил связывания CARMA, использующем эффективный двухпроходный способ обнаружения последовательностей.

Модели сегментации

Модели сегментации делят данные на сегменты, или кластеры, записей с одинаковыми структурами входных полей. Так как эти модели работают только с входными полями, у них нет отношения к выходным полям (полям назначения). Примеры моделей сегментации - это сети Коонена, кластеризация K-средних, двухшаговая кластеризация и выявление аномалий.

Модели сегментации (их называют также "моделями кластеризации") полезны в тех случаях, когда конкретный результат неизвестен (например, при идентификации нового шаблона мошенничества или интересующих вас групп в базе данных клиентов). Модели кластеризации уделяют главное внимание идентификации групп сходных записей и присвоению меток записям в соответствии с группами, к которым они принадлежат. Это делается без использования преимуществ предварительного знания о группах и их характеристиках, что отличает модели кластеризации от других способов моделирования, то есть отсутствие предварительно определенного выходного поля (поля назначения), значение в котором предсказывалось бы моделью. У этих моделей нет правильных или неправильных ответов. Их ценность в способности захватывать интересные группировки данных и представлять полезные описания этих группировок. Модели кластеризации часто используются для создания кластеров или сегментов, которые используются в качестве входных данных при последующем анализе (например, при разделении потенциальных покупателей по однородным подгруппам).

Узлы сегментации

Узел автоматической кластеризации оценивает и сравнивает модели кластеризации, идентифицирующие группы записей со сходными характеристиками. Этот узел работает аналогично другим узлам автоматического моделирования, допуская экспериментирование с несколькими комбинациями опций при одном проходе моделирования. Модели можно сравнивать при помощи базовых показателей, пытаясь фильтровать и ранжировать с их использованием полезность моделей кластеризации и предоставить показатель на основе важности конкретных полей.

Узел K-средних кластеризует набор данных в отдельные группы (или кластеры). Этот метод определяет фиксированное количество кластеров, итерационно распределяет записи по кластерам и настраивает центры кластеров, пока дальнейшие уточнения более не улучшают модель. Вместо попытки предсказать выходное значение *k*-средние используют процесс, называемый неконтролируемым обучением, чтобы обнаружить структуры в наборе входных полей.

Узел Коонена генерирует тип нейросети, которую можно использовать для кластеризации набора данных в отдельные группы. Когда сеть полностью обучена, похожие записи должны быть близко друг от друга на выходной карте, а отличающиеся записи должны быть сильно разделены. По количеству наблюдений, захваченных каждым нейроном в слепке модели, можно определить сильные нейроны. Это может дать представление об оправданном количестве кластеров.

Узел Двухшаговый использует метод двухшаговой кластеризации. На первом шаге проводится первый проход по данным, при котором необработанные входные данные сжимаются в управляемый набор подкластеров. На втором шаге используется способ иерархической кластеризации для все большего слияния подкластеров в крупные и еще более крупные кластеры. У двухшагового метода есть преимущество автоматической оценки оптимального числа кластеров для обучающих данных. Он может эффективно обрабатывать поля смешанных типов и большие наборы данных.

Узел выявления аномалий определяет необычные наблюдения, или выбросы, которые не соответствуют структуре “нормальных” данных. При помощи этого узла можно находить выбросы даже в том случае, если они не подходят ни под какие ранее известные шаблоны или вы точно не уверены, что именно ищете.

Модели исследования данных в базе данных

IBM SPSS Modeler поддерживает интеграцию с инструментами исследования данных и моделирования, доступных от поставщиков баз данных, в том числе Oracle Data Miner, IBM DB2 InfoSphere Warehouse и Microsoft Analysis Services. Построение моделей, их скоринг и сохранение в базе данных - все эти операции возможны в прикладной программе IBM SPSS Modeler. Полную информацию смотрите в *Руководстве по исследованию данных в базах данных IBM SPSS Modeler*, доступном на установочном носителе.

IBM SPSS StatisticsМодели

Если есть копия IBM SPSS Statistics, установленная и лицензированная на вашем компьютере, можно получить доступ к определенным подпрограммам IBM SPSS Statistics из IBM SPSS Modeler, чтобы построить и оценить модели.

Дополнительная информация

Доступна также подробная документация об алгоритмах моделирования. Дополнительную информацию смотрите в *Руководстве по алгоритмам IBM SPSS Modeler*, доступном на установочном носителе продукта.

Построение моделей расщепления

Моделирование расщеплений позволяет использовать один поток, чтобы построить отдельные модели для каждого возможного значения флагового, номинального или количественного входного поля, так что итоговые модели доступны из одного слепка модели. Возможные значения для входных полей могут очень по-разному воздействовать на модель. При помощи моделирования расщеплений можно легко построить наиболее подходящую модель для каждого возможного значения поля при одном выполнении потока.

Обратите внимание, что интерактивные сеансы моделирования расщеплять нельзя. При интерактивном моделировании вы задаете каждую модель индивидуально, поэтому не будет никаких преимуществ при использовании расщепления, при котором несколько моделей строится автоматически.

Моделирование расщеплений работает, назначая конкретное входное поле как поле расщепления. Это можно сделать, задав для роли поля значение **Расщепление** в спецификации Тип.

В качестве полей расщепления можно назначать только поля с типом измерения **Флаг**, **Номинальное**, **Порядковый номер** или **Непрерывное**.

В качестве полей расщепления можно назначить несколько входных полей. Однако в этом случае сильно возрастет количество создаваемых моделей. Модель строится для каждой возможной комбинации значений выбранных полей расщепления. Например, если в качестве полей расщепления назначаются три входных поля и у каждого из них по три значения, в результате будет создано 27 моделей.

Даже после назначения одного или нескольких полей в качестве полей расщепления вы еще можете выбрать, создавать ли модели расщепления, или одинарную модуль, используя переключатель в диалоговом окне узла моделирования.

Если поля расщепления определены, но переключатель не включен, генерируется только одна модель. Аналогично, если этот переключатель включен, но поля расщепления не определены, расщепление игнорируется и генерируется одна модель.

При запуске потока отдельные модели генерируются в фоновом режиме для каждого возможного значения полей расщепления, и только один слепок модели помещается на палитру моделей и на холст потока. Слепок модели расщепления помечается знаком расщепления; это два серых треугольника, наложенных на изображение слепка.

При просмотре слепка модели расщепления будут показаны все отдельные модели, которые были построены.

Индивидуальную модель из списка можно изучить, дважды щелкнув по ее значку слепка в средстве просмотра. При этом откроется стандартное окно браузера для индивидуальной модели. Если слепок расположен на холсте, при двойном щелчке по миниизображению графика откроется полноразмерный график. Дополнительную информацию смотрите в разделе “Средство просмотра расщепленных моделей” на стр. 46.

После создания модели как модели расщепления вы не можете удалить из нее обработку расщеплений или отменить расщепление далее по потоку из узла или слепка моделирования расщепления.

Пример. Крупная розничная торговая сеть хочет оценить продажи по категориям товаров в каждом из магазинов по стране. Используя моделирование расщеплений пользователи назначают поле Магазин для своих входных данных как поле расщепления, что позволяет построить отдельные модели для каждой категории в каждом магазине в одной операции. Итоговую информацию можно использовать для гораздо более точного управления уровнями запасов, чем в одной модели.

Расщепление и разделы

У расщепления есть несколько общих с созданием моделей возможностей, но расщепления и разделы используются очень по-разному.

Разделение равномерно делит набор данных на две или три части: обучение, испытание и (необязательно) проверка; оно используется для испытания производительности одной модели.

Расщепление делит набор данных столько частей, сколько доступно по существующим значениям для поля расщепления; оно используется для построения нескольких моделей.

Разделение и расщепление работают полностью независимо друг от друга. На узле моделирование можно выбрать одно и/или другое.

Узлы моделирования, поддерживающие модели расщепления

Многие узлы моделирования могут создавать модели расщепления. Исключения - это узлы автокластеризации, временных рядов, PCA/факторов, выбора показателей, SLRM, узлы моделей связывания (Априори, CARMS и Последовательность), узлы моделей кластеризации (K-средние, Коонена, двухшаговые и аномальные), Statistics Модели и узлы, используемык для моделирования In-database.

Моделирование расщепления поддерживают следующие узлы:

Дерево C&R

Байесовская сеть

QUEST

GenLin

CHAID

KNN

C5.0

Модель Кокса

Нейросеть

Автоклассификатор

Список решений

Автонумерация

Регрессия

Логистическое

Дискриминантный

SVM

Затрагиваемые расщеплением возможности

Использование моделей расщепления по-разному влияет на многие возможности IBM SPSS Modeler. В этом разделе представлено руководство по использованию моделей расщепления в связи с другими узлами в потоке.

Узлы операций с записями

При использовании моделей расщепления в потоке, содержащем узел **Выборка**, стратифицируйте записи по полям расщепления, чтобы получилась ровная выборка данных. Эта опция доступна при выборе способа выборки Сложный.

Обратите внимание, что при наличии в потоке узла **Баланс** балансировка применяется ко всему набору входных записей, а не к подмножеству записей внутри расщепления.

При агрегировании записей с использованием узла **Агрегация** задайте, что поля расщепления - это ключевые поля, если вы хотите вычислить агрегаты для каждого расщепления.

Узлы операций с полями

Узел **Тип** расположен там, где вы задаете, какие поля используются в качестве полей расщепления.

Обратите внимание на то, что, хотя узел **Ансамбль** используется для объединения двух или более слепков моделей, он не может использоваться для обращения действия расщепления, так как модели расщепления встроены в один слепок модели.

Узлы моделирования

Модели расщепления не поддерживают вычисление важности предикторов (относительную важность предикторных входных полей при оценке модели). При построении моделей расщепления параметры важности предикторов игнорируются.

Узел **KNN** (ближайшие соседи) поддерживает модели расщепления только в том случае, если он задан для предсказания поля назначения. При другой настройке (только определить ближайших соседей) модель не создается. Если выбрана опция "Автоматически определить k", у каждой из моделей расщепления может быть различное количество ближайших соседей. Таким образом, у общей модели количество сгенерированных столбцов будет равно максимальному числу ближайших соседей среди всех моделей

расщепления. Для тех моделей расщепления, у которых число ближайших соседей меньше этого максимума, соответствующее число столбцов будет заполнено значениями $null. Дополнительную информацию смотрите в разделе “Узел KNN” на стр. 281.

Узлы моделирования баз данных

Узлы моделирования In-database не поддерживают моделей расщепления.

Слепки моделей

Экспорт в PMML из слепка модели расщепления невозможен, так как слепок содержит несколько моделей, а PMML не поддерживает такое пакетирование. Однако экспорт в текстовый формат или формат HTML возможен.

Моделирование опций полей узла

Все узлы моделирования содержат вкладку Поля, где можно задать поля для использования при построении модели.

Перед построением модели необходимо указать поля, которые должны служить полями назначения и входными полями. За немногими исключениями, все узлы моделирования будут использовать информацию о полях из узла Тип, расположенного выше. Если вы используете узел Тип для выбора входных полей и полей назначения, на этой вкладке можно ничего не менять. (Узел Последовательность и узел Извлечение текста - это исключения, и для них требуется, чтобы параметры полей задавались на узле моделирования).

Использовать параметры узла типа. Эта опция указывает узлу, что следует использовать информацию о полях из узла Тип, расположенного выше. Это опция по умолчанию.

Использовать пользовательские параметры. Эта опция указывает узлу, что следует использовать информацию о полях, заданную здесь, а не ту, что задана на расположенных выше узлах Тип. После выбора этого варианта задайте приведенные ниже поля, как это потребуется.

Примечание: Не все поля выводятся для всех узлов.

- **Использовать транзакционный формат (только узлы Априори, CARMA, Правила связывания MS и Априори Oracle).** Включите этот переключатель, если у исходных данных **транзакционный формат**. У записей в этом формате есть два поля, один для ID и один для содержимого. Каждая запись представляет единственную транзакцию или элемент, и связанные элементы связаны наличием одинакового ID. Выключите этот переключатель, если у данных **табличный формат**, в котором элементы представлены отдельными флагами, где каждое поле флага указывает на наличие или отсутствие конкретного элемента, а каждая запись представляет полный набор связанных элементов. Дополнительную информацию смотрите в разделе “Сравнение табличных и транзакционных данных” на стр. 226.
  - **ID.** Для транзакционных данных выберите из списка поле ID. Значения в поле ID могут быть числовыми или символическими. Каждое уникальное значение в этом поле должно обозначать конкретный объект анализа. Например, в прикладной программе Корзина покупок каждый ID может представлять одного покупателя. В прикладной программе Анализ Web-журнала каждый ID может представлять отдельный компьютер (по IP-адресу) или одного пользователя (по регистрационным данным).
  - **Последовательные ID.** (только для узлов Априори и CARMA) Если ваши данные предварительно отсортированы, так что все записи с одинаковым ID сгруппированы совместно в потоке данных, выберите эту опцию для ускорения обработки. Если ваши данные предварительно не отсортированы (или вы в этом не уверены), оставьте эту опцию выключенной, и узел отсортирует данные автоматически.

    Примечание: Если данные не отсортированы, но выбирается эта опция, можно получить недопустимые результаты для вашей модели.

– **Содержимое.** Задайте поле или поля содержимого для модели. Эти поля содержат нужные элементы при моделировании связывания. Можно задать несколько полей флагов (если данные в табличном формате) или одно номинальное поле (если данные в транзакционном формате).

- **Цель.** Для тех моделей, для которых требуется одна или несколько полей назначения, выберите поле назначения или несколько полей. Это аналогично заданию для поля роли *Поле назначения* на узле Тип.
- **Оценка.** (Только для моделей автокластеризации.) Для кластерных моделей поле назначения не задается; однако можно выбрать поле оценки для идентификации его уровня важности. Кроме этого, можно оценить, насколько хорошо кластеры различают значения этого поля, что в свою очередь означает, можно ли использовать кластеры для предсказаний этого поля.
  – **Поля ввода.** Выберите одно или несколько входных полей. Это аналогично заданию для поля роли *Входное* на узле Тип.
  – **Подмножества.** Это поле позволяет позволяет задать поле, с помощью которого данные будут разделяться на отдельные выборки для этапов обучения, испытания и проверки при построении модели. Сгенерировав модель при помощи одной выборки и испытав ее при помощи другой, можно получить надежный показатель качества обобщения модели на более крупные наборы данных, подобные текущим. Если при помощи узлов Тип или Раздел были определено несколько полей разделов, на вкладке Поля на каждом узле моделирования, где используется разделение, должно быть выбрано одно поле раздела. (Если представлен только один раздел, он будет использоваться автоматически при всяком включении разделения.) Кроме того, учтите, что для применения выбранного раздела в анализе на вкладке Параметры модели для узла должно быть также включено разделение. (Выключение этой опции делает возможным отключение разделения без изменения значений параметров полей.)
- **Разбиения.** Для разбиения моделей выберите поле или поля разбиения. Это аналогично заданию для поля роли *Расщепление* на узле Тип. В качестве полей разбиения можно назначать только поля с типом измерения **Флаг**, **Номинальное**, **Порядковый номер** или **Непрерывное**. Поля, выбранные как поля разбиения, нельзя использовать в качестве полей назначения, разделов, частоты, веса или входных полей. Дополнительную информацию смотрите в разделе “Построение моделей расщепления” на стр. 28.
- **Использовать поле частоты.** Это опция позволяет вам выбрать поле для веса частоты. Его можно использовать, если каждая запись в ваших данных обучения представляет несколько блоков, например, если используются агрегированные данные. Значениями в этом поле должны быть количества блоков, представленных каждой записью. Дополнительную информацию смотрите в разделе “Использование полей частоты и веса”.

Примечание: Если появится сообщение об ошибке **Недопустимые метаданные (для входных/выходных полей)**, убедитесь, что вы задали все необходимые поля, в частности поле частоты.

- **Использовать поле веса.** Это опция позволяет вам выбрать поле как вес наблюдения. Веса наблюдений используются для учета различий в дисперсии разных уровней поля назначения. Дополнительную информацию смотрите в разделе “Использование полей частоты и веса”.
- **Последующие.** Для узлов вывода правил (Априори) выберите поля, которые будут использоваться как консеквенты в итоговом наборе правил. (Это соответствует полям с ролью *Назначение* или *Оба* на узле Тип).
- **Противоположные.** Для узлов вывода правил (Априори) выберите поля, которые будут использоваться как антецеденты в итоговом наборе правил. (Это соответствует полям с ролью *Входное* или *Оба* на узле Тип).

У некоторых моделей вкладка Поля отличается от описанной в этом разделе.

- Дополнительную информацию смотрите в разделе “Опции полей узла Последовательность” на стр. 241.
- Дополнительную информацию смотрите в разделе “Опции полей узла CARMA” на стр. 230.

Использование полей частоты и веса

Поля частоты и веса используются для предоставления дополнительной важности некоторым записям по сравнению с остальными, так как, например, вы знаете, что некоторая часть возможных данных недостаточно представлена в данных обучения (вес), или одна запись представляет несколько идентичных наблюдений (частота).

- Значениями для поля частоты должны быть положительные целые числа. Записи с отрицательным или нулевым весом частоты исключаются из анализа. Нецелочисленные веса частот округляются до ближайшего целого.
- Значения веса наблюдений должны быть положительными, но не обязательно целыми значениями. Записи с отрицательным или нулевым весом наблюдений исключаются из анализа.

Скоринг полей частоты и веса

Поля частоты и веса используются при обучении моделей, но не используются при скоринге, так как оценка для каждой записи основана на ее характеристиках и не зависит от того, сколько наблюдений она представляет. Допустим, например, что ваши данные расположены в следующей таблице.

Таблица 1. Пример данных

Состоит в браке	**Отклик**
Да	Да
Да	Да
Да	Да
Да	Нет
Нет	Да
Нет	Нет
Нет	Нет

На основании этих данных можно заключить, что трое из четверых клиентов, которые состоят в браке, ответили на предложение, а двое из троих не состоящих в браке не ответили. Поэтому вы будете оценивать все новые записи соответственно, как показано в следующей таблице.

Таблица 2. Пример записей для скоринга

Состоит в браке	**$-Отклик**	**$RP-Отклик**
Да	Да	0,75 (три из четырех)
Нет	Нет	0,67 (два из трех)

Другой вариант - сохранять ваши данные обучения более компактно, используя поле частоты, как показано в следующей таблице.

Таблица 3. Вариант примера записей для скоринга

Состоит в браке	**Отклик**	**Частота**
Да	Да	3
Да	Нет	1
Нет	Да	1
Нет	Нет	2

Так как здесь представлен тот же самый набор данных, вы построите ту же модель и предскажете отклики на основании только матримониального положения. Если при скоринге данных есть десять клиентов в браке, для каждого из них вы предскажете *Да* независимо от того, как они представлены, десятью отдельными записями или одной со значением частоты 10. Хотя вес обычно представлен не целым числом, его можно рассматривать аналогично, как обозначение важности записи. Поэтому при скоринге записей поля частоты и веса не используются.

Оценка и сравнение модели

Некоторые типы моделей поддерживают поля частоты, другие - поля веса, а некоторые модели поддерживают обе опции. Но во всех случаях при применении этих полей они используются только для построения моделей и не рассматриваются при оценке модели на узле Оценка или на узле Анализ или при ранжировании моделей с помощью большинства методов, используемых узлами Автоклассификация и Автонумерация.

- При сравнении моделей (например, по диаграммам оценки) значения частоты и веса будут игнорироваться. Это позволяет провести поуровневое сравнение моделей, использующих и не использующих эти поля, но означает также, что для точной оценки нужно использовать набор данных, который точно воспроизводит заполнение, не основываясь на полях частоты или веса. С практической точки зрения это можно сделать, убедившись, что модели оцениваются с использованием выборки испытания, в которой значения полей частоты и веса всегда равны 0 или 1. (Это ограничение применимо только при оценке моделей; если значения частоты и веса были равны 1 для выборок обучения и испытания, нет необходимости первоочередного использования этих полей).
- Если используется автоклассификация, частоту можно учитывать при ранжировании моделей на основании дохода, поэтому в данном случае рекомендуется этот метод.
- При необходимости данные можно разделить на выборки обучения и испытания, используя узел Разделы.

Опции анализа узлов моделирования

Многие узлы моделирования содержат вкладку Анализ, позволяющую получить информацию о важности предикторов вместе с простыми и скорректированными оценками склонности.

Оценка модели

Вычислить важность предикторов. Для моделей, производящих соответствующую меру важности, можно вывести диаграмму, показывающую относительную важность каждого предиктора при оценке модели. Обычно вам потребуется направить усилия по моделированию на рассмотрение наиболее важных предикторов, отбрасывая или игнорируя наименее важные. Обратите внимание на то, что для некоторых моделей вычисление важности предикторов - это длительный процесс, особенно при работе с большими наборами данных, и в результате по умолчанию для некоторых моделей эта опция будет выключена. Важность предикторов недоступна для моделей списка решений. Дополнительную информацию смотрите в разделе “Важность предиктора” на стр. 42.

Оценки склонностей

Оценки склонностей можно включить на узле моделирования и на вкладке Параметры слепка модели. Эта функциональная возможность доступна только при выборе поля назначения флагового типа. Дополнительную информацию смотрите в разделе “Оценки склонностей” на стр. 35.

Вычислить простые оценки склонности. Простые оценки склонности получаются из модели на основе только обучающих данных. Если модель предсказывает значение *true* (будет отклик), склонность совпадает с P, где P - это вероятность предсказания. Если модель предсказывает значение false, склонность вычисляется как (1 – P).

- При выборе этой опции при построении модели оценки склонности будут включены в слепок модели по умолчанию. Однако вы всегда можете включить простые оценки склонности в слепке модели независимо от выбора их на узле моделирования.
- При скоринге модели простые оценки склонности будут добавлены в поле с буквами *RP*, присоединенными к стандартному префиксу. Например, если предсказания содержатся в поле с именем *$R-churn*, именем поля оценки склонности будет *$RRP-churn*.

Вычислить скорректированные оценки склонности. Простые оценки склонности основаны исключительно на оценках модели, которая может быть переобучена, что приведет к чрезмерно оптимистическим оценкам

склонности. Скорректированные оценки вносят компенсацию, изучая выполнение модели на испытательном и проверочном разделах и уточняя склонности для улучшения в соответствии с этим оценки.

- Для этого раздела требуется, чтобы в потоке присутствовало допустимое поле раздела.
- В отличие от простых оценок достоверности, скорректированные оценки склонностей нужно вычислять при построении модели; в противном случае они будут недоступны при скоринге слепка модели.
- При скоринге модели скорректированные оценки склонности будут добавлены в поле с буквами *AP*, присоединенными к стандартному префиксу. Например, если предсказания содержатся в поле с именем *$R-churn*, именем поля оценки склонности будет *$RAP-churn*. Скорректированные оценки склонности недоступны для моделей логистической регрессии.
- При вычислении скорректированных оценок склонности испытательный или проверочный раздел, используемый для вычисления, не должен быть сбалансирован. Чтобы исключить это, убедитесь, что выбрана опция **Только сбалансированные данные обучения** на любом вышележащем узле Баланс. Кроме этого, если на вышележащем уровне взята сложная выборка, это может сделать скорректированные оценки склонностей неприемлемыми.
- Скорректированные оценки склонности недоступны для моделей моделей дерева или набора правил с "бустингом". Дополнительную информацию смотрите в разделе “Усиленные модели C5.0” на стр. 111.

Основываясь на. Для вычисления скорректированных оценок склонности в потоке должно присутствовать поле раздела. Вы должны указать, какой из разделов использовать для этих вычислений - проверочный или испытательный. Для получения наилучших результатов испытательный или проверочный раздел должен включать в себя по крайней мере столько записей, сколько и раздел, использованный для обучения исходной модели.

Оценки склонностей

Для моделей, возвращающих предсказание *да* или *нет*, в дополнение к стандартным значениям предсказаний и доверительным значениям можно затребовать оценки склонностей. Оценки склонностей обозначают правдоподобие конкретного выхода или отклика. Пример представлен в следующей таблице.

Таблица 4. Оценки склонности

Покупатель	**Склонность к отклику**
Джо Смит	35%
Джейн Смит	15%

Оценки склонностей доступны только для моделей с флаговыми полями назначения и обозначают правдоподобие значения *True*, определенного для поля, как задано на узле источника или узле Тип.

Сравнение оценок склонности с оценками достоверности

Оценки склонности отличаются от оценок достоверности, применимых к текущему предсказанию для его значений *да* или *нет*. Например, в случае предсказания *нет* высокая оценка достоверности означает высокое правдоподобие отклика *нет*. Оценки склонности обходят это ограничение для включения более простого сравнения по всем записям. Например, предсказание *нет* с оценкой достоверности *0.85* переводится в простую оценку склонности *0.15* (или *1 минус 0.85*).

Таблица 5. Оценки достоверности

Покупатель	**Прогноз**	**Показатель доверия**
Джо Смит	Откликнется	0,35
Джейн Смит	Не откликнется	0,85

Получение оценок достоверности

- Оценки склонности можно включить на вкладке Анализ узла моделирования или на вкладке Параметры слепка модели. Эта функциональная возможность доступна только при выборе поля назначения флагового типа. Дополнительную информацию смотрите в разделе “Опции анализа узлов моделирования” на стр. 34.
- В зависимости от используемого способа ансамбля оценки склонности можно вычислить также при помощи узла Ансамбль.

Вычисление скорректированных оценок склонности

Скорректированные оценки склонности вычисляются как часть процесса построения модели, а в противном случае они недоступны. После построения модели она оценивается с использованием данных из раздела испытания или оценки, а новая модель для получения скорректированных оценок склонности конструируется при анализе производительности исходной модели для этого раздела. В зависимости от типа модели для вычисления скорректированных оценок склонности используется один из двух способов.

- Для моделей набора правил и типа дерева скорректированные оценки склонности генерируются при повторном вычислении частоты каждой категории на каждом узле дерева (для моделей типа дерева) или поддержки и достоверности каждого правила (для моделей набора правил). Этот результат в модели набора правил или типа дерева сохраняется с исходной моделью, чтобы использоваться при всяком требовании скорректированных оценок свойств. При всяком применении исходной модели к новым данным можно последовательно применить новую модель к простым оценкам склонности, чтобы сгенерировать скорректированные оценки.
- Для других моделей записи, созданные при скоринге исходной модели на разделе испытания или проверки, затем категоризуются по их простой оценке склонности. Далее обучается нейронная сеть, которая определяет нелинейную функцию, отображающую среднюю простую склонность в каждой категории на среднюю наблюдаемую склонность в той же категории. Как отмечалось ранее для моделей типа дерева, итоговая нейронная сеть применяться к простым оценкам склонности при всяком требовании скорректированных оценок свойств.

Меры предосторожности относительно пропущенных значений в испытательном разделе. Обработка входных значений отсутствия в разделе испытания/проверки зависит от модели (подробности смотрите в индивидуальных алгоритмах скоринга моделей). Модель C5 не может вычислить скорректированные склонности, когда есть входные значения отсутствия.

Слепки моделей

Рисунок 19. Слепок модели

Слепок модели - это контейнер для модели, то есть набор правил, формул или уравнений, представляющих результаты операций построения модели в IBM SPSS Modeler. Основная цель слепка - оценивание данных для прогнозирования и дальнейший анализ свойств модели. При открытии слепка модели на экране вы можете просмотреть различные сведения о модели, такие как относительная важность входных полей при создании модели. Для просмотра предсказаний необходимо присоединить и выполнить следующий узел процесса или вывода. Дополнительную информацию смотрите в разделе “Использование слепков моделей в потоках” на стр. 47.

Рисунок 20. Ссылка на модель из узла моделирования в слепок модели

После успешного выполнения узла моделирования соответствующий слепок модели размещается на холсте потока, где он представлен желтым ромбовидным значком в виде кристалла. На холсте потока слепок показан с соединением (сплошная линия) с ближайшим подходящим узлом до узла моделирования, и со ссылкой (пунктирная линия) на сам узел моделирования.

Слепок располагается также на палитре моделей в верхнем правом углу окна IBM SPSS Modeler. В любом положении слепки можно выбрать и просмотреть в них подробности моделей.

После успешного выполнения узла моделирования слепки всегда размещаются на палитре моделей. Можно задать пользовательскую опцию для управления - будет ли слепок дополнительно размещаться на холсте потока.

В следующих разделах представлена информация об использовании слепков моделей в IBM SPSS Modeler. Для глубокого понимания используемых алгоритмов смотрите *Руководство по алгоритмам IBM SPSS Modeler*, доступное в папке *\Documentation* на DVD для IBM SPSS Modeler.

Ссылки на модели

По умолчанию слепок модели показывается на холсте со ссылкой на узел моделирования, который ее создал. Это особенно полезно в сложных потоках с несколькими слепками, и позволяет вам идентифицировать слепок, который будет изменен каждым узлом моделирования. Каждая ссылка содержит знак для обозначения, будет ли заменена модель при выполнении узла моделирования. Дополнительную информацию смотрите в разделе “Замена модели” на стр. 39.

Определение и удаление ссылок на модели

Ссылки можно определить и удалить вручную на холсте. При определении новой ссылки указатель изменяется на указатель ссылки.

Рисунок 21. Указатель ссылки

Определение новой ссылки (контекстное меню)

1. Щелкните правой кнопкой мыши по узлу моделирования, с которого вы хотите активировать ссылку.
2. В контекстном меню выберите пункт **Определить ссылку на модель**.
3. Щелкните по слепку, на котором вы хотите заканчивать ссылку.

Определение новой ссылки (главное меню)

1. Щелкните правой кнопкой мыши по узлу моделирования, с которого вы хотите активировать ссылку.
2. В основном меню выберите:

   Правка > Узел > Определить ссылку для модели
3. Щелкните по слепку, на котором вы хотите заканчивать ссылку.

Удаление существующей ссылки (контекстное меню)

1. Щелкните правой кнопкой мыши по слепку в конце ссылки.
2. В контекстном меню выберите **Удалить ссылку на модель**.

Альтернативный вариант:

1. Щелкните правой кнопкой мыши по знаку в середине ссылки.
2. В контекстном меню выберите **Удалить ссылку**.

Удаление существующей ссылки (главное меню)

1. Щелкните по узлу моделирования или слепку, с которых вы хотите удалить ссылку.
2. В основном меню выберите:

   Правка > Узел > Удалить ссылку для модели

Копирование и вставка ссылок на модель

Если вы копируете связанный слепок без его узла моделирования и вставляете этот слепок в тот же поток, этот слепок вставляется со ссылкой на узел моделирования. У новой ссылки то же состояние замены модели (смотрите раздел “Замена модели” на стр. 39), что и у исходной ссылки.

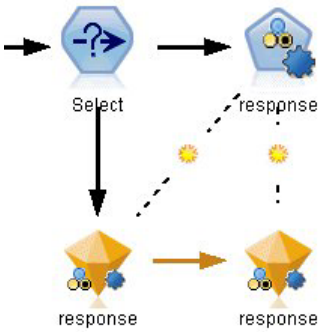


Рисунок 22. Копирование и вставка связанного слепка

Если вы копируете и вставляете слепок вместе с его связанным узлом моделирования, сохраняется ссылка, куда вставлены объекты, в тот же поток или в новый.

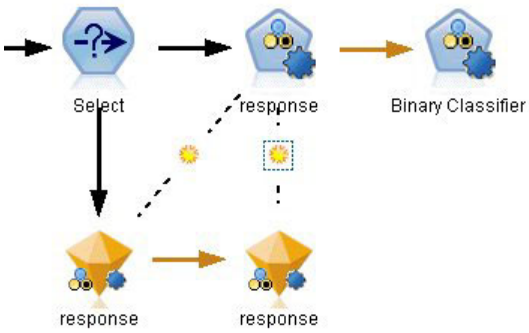


Рисунок 23. Копирование и вставка связанного слепка

Примечание: Если вы копируете связанный слепок без его узла моделирования и вставляете этот слепок в новый поток (или в суперузел, который не содержит узла моделирования), ссылка разрывается и вставляется только слепок.

Ссылки на модели и суперузлы

Если вы определили, что суперузел будет включать или узел моделирования, или слепок связанной модели, но не одновременно, ссылка разрывается. Расширение суперузла не восстанавливает ссылку; этого можно достичь, только отменив создание суперузла.

Замена модели

Вы можете выбрать, заменить ли (то есть, изменить) существующий слепок, при повторном выполнении узла моделирования, создавшего этот слепок. Если выключить опцию замены, при повторном выполнении узла моделирования создается новый слепок.

Примечание: Замена модели отличается от обновления, которое относится к изменению модели в сценарии.

Каждая ссылка из узла моделирования на слепок содержит знак для обозначения, будет ли заменена модель при повторном выполнении узла моделирования.

Рисунок 24. Включена ссылка на модель с заменой модели

Изначально ссылка показывается с включенной заменой модели, что обозначается маленьким значком лучистого солнца в ссылке. В этом состоянии повторное выполнение узла моделирования на одном конце ссылки просто изменяет слепок на другом конце.

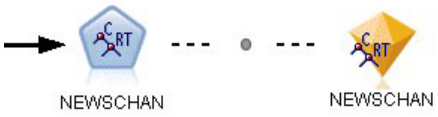


Рисунок 25. Выключена ссылка на модель с заменой модели

Если замена модели выключена, знак ссылки заменяется на серую точку. В этом состоянии повторное выполнение узла моделирования на одном конце ссылки добавляет новую, измененную версию слепка на холст.

В любом случае, на палитре Модели существующий слепок изменяется или новый слепок добавляется, в зависимости от настройки системной опции **Заменить предыдущую модель**.

Порядок выполнения

При выполнении потока с несколькими ветвями, содержащими слепки моделей, поток сначала оценивается для подтверждения, что ветвь с включенной заменой модели выполняется перед всеми ветвями, использующими слепок итоговой модели.

Если у вас более сложные требования, можно задать порядок выполнения вручную через сценарий.

Изменение параметра замены модели

Чтобы изменить параметр для замены модели:

1. Щелкните правой кнопкой мыши по знаку на ссылке.
2. Выберите требуемую опцию **Включить (выключить) замену модели**.

Примечание: Параметр замены модели для ссылки на модель перезаписывает параметр на вкладке Уведомления диалогового окна Пользовательские опции (Инструменты > Опции > Пользовательские опции).

Палитра моделей

Палитра моделей (на вкладке Модели в окне менеджеров) позволяет различными способами использовать, изучать и изменять слепки моделей.

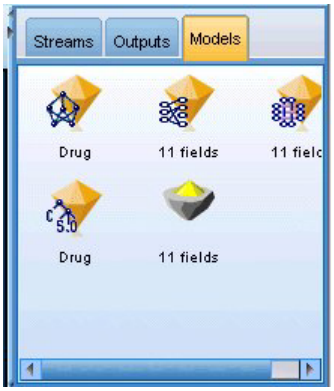


Рисунок 26. Палитра моделей

При щелчке правой кнопкой мыши по слепку модели на палитре моделей открывается контекстное меню со следующими опциями:

- **Добавить в поток.** Добавляет слепок модели в текущий активный поток. Если в потоке есть выбранный узел, слепок модели будет соединен с выбранным узлом, когда такое соединение станет возможно, а в противном случае - с ближайшим возможным узлом. Слепок выводится со ссылкой на узел моделирования, создавший модель, если этот узел все еще находится в потоке.
- **Обзор.** Открывает для слепка браузер моделей.
- **Переименовать и аннотировать.** Позволяет переименовать слепок модели и/или изменить аннотацию для слепка.
- **Создать узел моделирования.** Если у вас есть слепок модели, который нужно изменить или обновить, а используемый для создания модели поток недоступен, можно использовать эту опцию для повторного создания узла моделирования с теми же опциями, что и для создания исходной модели.
- **Сохранить модель, Сохранить модель как.** Сохраняет модель во внешний двоичный файл сгенерированной модели (.gm).
- **Запомнить модель.** Запоминает слепок модели в IBM SPSS Collaboration and Deployment Services Repository.
- **Экспорт PMML.** Экспортирует слепок модели в формате PMML, который можно использовать для скоринга новых данных вне IBM SPSS Modeler. **Экспорт PMML** доступен для всех сгенерированных узлов моделей. *Примечание*: Для использования этой возможности требуется лицензия на сервер IBM SPSS Modeler.
- **Добавить в проект.** Сохраняет слепок модели и добавляет его в текущий проект. На вкладке Классы слепок будет добавлен в папку Сгенерированные модели. На вкладке CRISP-DM он будет добавлен в фазу проекта по умолчанию.
- **Delete.** Удаляет слепок модели с палитры.

При щелчке правой кнопкой мыши по незанятой области на палитре модели открывается контекстное меню со следующими опциями:

- **Открыть модель.** Загружает слепок модели, ранее созданный в IBM SPSS Modeler.
- **Извлечь модель.** Извлекает хранимую модель из репозитория IBM SPSS Collaboration and Deployment Services.
- **Загрузить палитру.** Загружает сохраненную палитру моделей из внешнего файла.
- **Извлечь палитру.** Извлекает хранимую палитру моделей из репозитория IBM SPSS Collaboration and Deployment Services.

- **Сохранить палитру.** Сохраняет полное содержимое палитры моделей во внешний сгенерированный файл палитры моделей (.gen).
- **Запомнить палитру.** Запоминает полное содержимое палитры моделей в репозитории IBM SPSS Collaboration and Deployment Services.
- **Очистить палитру.** Удаляет все слепки из палитры.
- **Добавить палитру к проекту.** Сохраняет палитру моделей и добавляет ее к текущему проекту. На вкладке Классы слепок будет добавлен в папку сгенерированных моделей. На вкладке CRISP-DM он будет добавлен в фазу проекта по умолчанию.
- **Импорт PMML.** Загружает модель из внешнего файла. Вы можете открывать, просматривать и оценивать модели PMML, созданные IBM SPSS Statistics или другими прикладными программами, поддерживающими этот формат. Дополнительную информацию смотрите в разделе “Импорт и экспорт моделей в виде PMML” на стр. 48.

Просмотр слепков моделей

Браузеры слепков моделей позволяют исследовать и использовать результаты ваших моделей. В браузере можно сохранить, распечатать или экспортировать сгенерированную модель, изучить сводку модели и просмотреть или изменить аннотации для модели. Для некоторых типов слепков моделей можно также сгенерировать новые узлы, такие как узлы Фильтр или Набор правил. Для некоторых моделей можно просмотреть также параметры моделей, такие как правила или центры кластеров. Для некоторых типов моделей (модели на основе дерева и кластерные модели) можно просмотреть графическое представление структуры модели. Элементы управления для использования браузеров слепков моделей описаны ниже.

Меню

Меню Файл. У всех слепков моделей есть меню Файл, содержащее некоторое подмножество следующих опций:

- **Сохранить узел.** Сохраняет слепок модели в файле узла (.nod).
- **Запомнить узел.** Запоминает слепок модели в репозитории IBM SPSS Collaboration and Deployment Services.
- **Заголовок и комментарий.** Позволяет изменить верхний и нижний колонтитул страницы для печати из слепка.
- **Параметры страницы.** Позволяет изменить настройки страницы для печати из слепка.
- **Предварительный просмотр печати.** Выводит предварительный макет, как будет выглядеть слепок на печати. Выберите в подменю информацию для вывода при предварительном просмотре.
- **Печать.** Печатает содержимое слепка. Выберите в подменю информацию для печати.
- **Печатать представления.** Печатает текущее представление или все представления.
- **Экспортировать текст.** Экспортирует содержимое слепка в текстовый файл. Выберите в подменю информацию для экспорта.
- **Экспорт HTML.** Экспортирует содержимое слепка в файл HTML. Выберите в подменю информацию для экспорта.
- **Экспорт PMML.** Экспортирует слепок модели в формате PMML, который можно использовать с другими PMML-совместимыми программами. Дополнительную информацию смотрите в разделе “Импорт и экспорт моделей в виде PMML” на стр. 48. *Примечание*: Для использования этой возможности требуется лицензия на сервер IBM SPSS Modeler.
- **Экспорт SQL.** Экспортирует модель с использованием языка структурированных запросов (structured query language, SQL), который можно изменять и использовать с другими базами данных.

  Примечание: Экспорт SQL доступен только из следующих моделей: C5, C&RT, CHAID, QUEST, Линейная регрессия, Логистическая регрессия, Нейронная сеть, PCA/факторная модель и Список решений.

- **Опубликовать для адаптера скоринга сервера.** Публикует модель в базу данных с установленным адаптером скоринга, позволяя выполнять скоринг модели непосредственно в базе данных. Дополнительную информацию смотрите в разделе “Публикация моделей для адаптера скоринга” на стр. 50.

Меню Создать. У большинства слепков моделей есть также меню Создать, позволяющее генерировать новые узлы на основе слепка модели. Опции, доступные в этом меню, будут зависеть от типа модели, которую вы просматриваете. Обратитесь к конкретному слепку модели, чтобы узнать подробности, что именно можно сгенерировать из конкретной модели.

Меню Вид. На вкладке слепка Модель это меню позволяет вывести или скрыть различные панели инструментов визуализации, доступные в текущем режиме. Чтобы сделать доступным полный набор панелей инструментов, выберите Режим изменений (значок кисти) на панели инструментов Общие.

Кнопка Предварительный просмотр. У некоторых слепков моделей есть кнопка Предварительный просмотр, позволяющая вам увидеть образец данных модели, в том числе дополнительные поля, созданные в процессе моделирования. По умолчанию выводится десять строк, однако это значение можно изменить в свойствах потока.

Кнопка Добавить в текущий проект. Сохраняет слепок модели и добавляет его в текущий проект. На вкладке Классы слепок будет добавлен в папку Сгенерированные модели. На вкладке CRISP-DM он будет добавлен в фазу проекта по умолчанию.

Сводка слепков моделей / Информация

На вкладке Сводка или в представлении Информация выводится информация о полях, параметрах компоновки и процессе оценки модели. Результаты представляются в виде дерева, которое можно раскрывать или сворачивать, щелкнув по конкретному элементу.

Анализ. Выводится информация о модели. Конкретные подробности различаются для разных типов моделей; они представлены в разделах для каждого слепка модели. Кроме этого, если вы запускали узел Анализ, присоединенный к этому узлу моделирования, информация этого анализа также будет выводится в данном разделе.

Поля. Список полей, используемых в качестве полей назначения и входных полей при построении модели. Для моделей с расщеплениями выводится также список полей, определивших расщепления.

Параметры компоновки / Опции. Содержит информацию об используемых при построении модели параметрах.

Сводная информация по обучению. Выводится тип модели, поток, используемый для ее создания, создавший ее пользователь, отметка времени построения модели и время, затраченное на ее построение.

Важность предиктора

Обычно при моделировании сосредотачивают внимание на наиболее важных предикторах и исключают или игнорируют наименее важные. Это помогает сделать диаграмма важности предикторов, показывая относительную важность каждого предиктора при оценке модели. Поскольку значения важности являются относительными, сумма этих значений для всех показанных предикторов равна 1,0. Важность переменных не связана с точностью модели. Она лишь связана с важностью каждого предиктора для предсказания, а не с точностью этого предсказания.

Важность предикторов доступна для моделей, которые создают соответствующие статистические показатели важности, в том числе для моделей нейронных сетей, деревьев решений (дерево C&R, C5.0, CHAID и QUEST), Байесовских сетей, дискриминанта, SVM и SLRM, линейной и логистической регрессии,

обобщенной линейной модели и модели ближайших соседей (KNN). Для большинства этих моделей важность предикторов можно включить на вкладке Анализ узла моделирования. Дополнительную информацию смотрите в разделе “Опции анализа узлов моделирования” на стр. 34. Для моделей KNN смотрите раздел “Соседи” на стр. 283.

Примечание: Важность предикторов не поддерживается для моделей расщепления. При построении моделей расщепления параметры важности предикторов игнорируются. Дополнительную информацию смотрите в разделе “Построение моделей расщепления” на стр. 28.

Вычисление важности предикторов может занять существенно больше времени, чем построение модели, особенно при использовании больших наборов данных. Для моделей SVM и логистической регрессии требуется больше времени, чем для других моделей, и по умолчанию эта опция отключена для данных моделей. Если используется набор данных с большим числом предикторов, начальное экранирование с помощью узла Выбор показателей может ускорить получение результатов (смотрите ниже).

- Важность предикторов вычисляется в испытательном разделе, если он доступен. В противном случае используются данные обучения.
- Для моделей SLRM важность предикторов доступна, но вычисляется по алгоритму. Дополнительную информацию смотрите в разделе “Слепки моделей SLRM” на стр. 272.
- Вы можете использовать инструменты диаграмм IBM SPSS Modeler, чтобы работать с графом, редактировать и сохранять его.
- Дополнительно можно сгенерировать узел Фильтр на основании информации, содержащейся в диаграмме важности предикторов. Дополнительную информацию смотрите в разделе “Фильтрование переменных на основании важности” на стр. 44.

Важность предикторов и выбор показателей

Может показаться, что диаграмма важности предикторов, выведенная в слепке модели, в некоторых случаях дает результаты, аналогичные узлу Выбор показателей. Но выбор показателей ранжирует каждое входное поле на основании силы его взаимосвязи с заданным полем назначения и не зависит от других входных полей, а диаграмма важности предикторов обозначает относительную важность каждого входного поля для *этой* конкретной модели. Таким образом, выбор показателей будет более консервативен в экранировании входных полей. Например, если оба поля *название задания* и *категория задания* сильно связаны с вознаграждением, выбор показателей обозначит важность обоих этих полей. Но при моделировании принимаются во внимание также взаимодействия и корреляции. Таким образом, вы можете обнаружить, что используется только одно из этих входных полей, если в обоих из них повторяется одинаковая информация. На практике выбор показателей наиболее полезен для предварительного экранирования, в частности, при работе с большими наборами данных, у которых много переменных, а важность предикторов полезнее при точной настройке модели.

Различия в важности между отдельными моделями и узлами автоматического моделирования

Вы можете заметить небольшие различия в важности предикторов в зависимости от того, создаете ли вы отдельную модель из одного узла или используете для генерирования результатов узел автоматического моделирования. Такие различия в реализации обусловлены некоторыми ограничениями в реализации.

Например, для отдельных классификаторов, таких как CHAID, при вычислении значений важности применяется правило остановки и используются значения вероятностей. В отличие от этого, автоматический классификатор не использует правило остановки и непосредственно использует предсказанные метки в вычислении. Эти различия могут привести к тому, что при генерировании отдельной модели с помощью автоклассификатора значение важности можно рассматривать как грубую оценку по сравнению со значением, полученным для отдельного классификатора. Чтобы получить самые точные значения важности предикторов, мы советуем использовать отдельный узел, а не узлы автоматического моделирования.

Фильтрование переменных на основании важности

Дополнительно можно сгенерировать узел Фильтр на основании информации, содержащейся в диаграмме важности предикторов.

Отметьте предикторы, которые вы хотите включить в диаграмму (если это применимо), и в меню выберите:

Создать > Узел Фильтр (Важность предикторов)

OR

> Отбор полей (Важность предикторов)

Максимальное количество переменных. Наиболее важные предикторы включаются и исключаются на основе указанного количества.

Важность больше чем. Включает и исключает все предикторы с относительной важностью большей указанного значения.

Средство просмотра ансамблей

Модели для ансамблей

Модель для ансамбля предоставляет информацию о моделях компонентов в ансамбле и эффективности ансамбля как единого целого.

Главная (независимая от вида) панель инструментов позволяет выбирать, использовать ли для скоринга ансамбль или опорную модель. Если для скоринга используется ансамбль, то можно также выбрать правило объединения. Эти изменения не требуют повторного выполнения модели, однако эти варианты выбора сохраняются в (слепке) модели для скоринга и/или нисходящего анализа модели. Они также влияют на PMML, экспортируемый из средства просмотра ансамблей.

Правило объединения. Это правило, которое применяется при скоринге ансамбля, чтобы объединить предсказанные значения для базовых моделей с целью вычисления значений скоринга для ансамбля.

- Предсказанные значения для ансамбля в случае **категориальных** целевых переменных можно объединить, используя голосование, наибольшую вероятность или наибольшую среднюю вероятность. **Голосование** позволяет выбрать категорию, которая наиболее часто имеет наибольшую вероятность в базовых моделях. **Наибольшая вероятность** позволяет выбрать категорию, которая имеет наибольшую вероятность среди всех базовых моделей. **Наибольшая средняя вероятность** позволяет выбрать категорию, которая имеет наибольшую вероятность при усреднении вероятностей категорий по базовым моделям.
- Предсказанные значения для ансамбля в случае **непрерывных** целевых переменных можно объединить, используя среднее или медиану предсказанных значений для базовых моделей

Выбор по умолчанию основывается на спецификациях заданных во время построения моделей. Изменение правила объединения приводит к пересчету точности моделей и обновлению всех изображений точности моделей. Обновляется также диаграмма важностей предикторов. Этот элемент управления недоступен, если для скоринга выбрана опорная модель.

Показать все правила объединения. Если выбран этот пункт, то на диаграмме качества модели выводятся результаты для всех имеющихся правил объединения. Также обновляется диаграмма точности моделей компонентов, чтобы показать опорные линии для каждого метода голосования.

Сводка для модели: Вид Сводка для модели - это мгновенная визуальная сводка качества и разнородности ансамбля.

Качество. Эта диаграмма выводит точность окончательной модели в сравнении с опорной моделью и наивной моделью. Точность представляется в формате "больше значит лучше"; "наилучшая" модель будет иметь наибольшую точность. Для категориальной целевой переменной точность - это просто процент записей, для которых предсказанное значение совпадает с наблюденным значением. Для непрерывной целевой переменной точность - это 1 минус отношение средней абсолютной ошибки предсказания (среднего абсолютных значений разностей предсказанных и наблюденных значений) к диапазону предсказанных значений (разности максимального и минимального предсказанных значений).

Для ансамблей, созданных с помощью бэггинга, опорная модель - это стандартная модель, построенная по всему обучающему разбиению. Для ансамблей, созданных с помощью бустинга, опорная модель - это первая компонентная модель.

Наивная модель представляет точность в случае, когда модель не была построена, и относит все записи к модальной категории. Наивная модель не вычисляется для непрерывных целевых переменных.

Разнородность. Эта диаграмма выводит "разброс мнений" среди моделей компонент, используемых для построения ансамбля, представленный в формате "больше значит более разнородный". Это есть мера того, насколько сильно предсказания различаются в базовых моделях. Разнородность недоступна для моделей ансамблей, созданных с помощью бустинга, и также не выводится для непрерывных целевых переменных.

Важность предикторов: Обычно при моделировании сосредотачивают внимание на наиболее важных предикторах и исключают или игнорируют наименее важные. Это помогает сделать диаграмма важности предикторов, показывая относительную важность каждого предиктора при оценке модели. Поскольку значения важности являются относительными, сумма этих значений для всех показанных предикторов равна 1,0. Важность переменных не связана с точностью модели. Она лишь связана с важностью каждого предиктора для предсказания, а не с точностью этого предсказания.

Важность предикторов недоступна для всех моделей ансамблей. Набор предикторов может варьироваться по моделям компонентов, но важность может быть вычислена для предикторов, используемых, по крайней мере, в одной модели компонента.

Частота предикторов: Набор предикторов может варьироваться по моделям компонентов в силу выбора метода моделирования или выбора предикторов. Диаграмма частоты предикторов представляет собой точечную диаграмму, показывающую распределение предикторов по моделям компонент в ансамбле. Каждая точка представляет одну или несколько моделей компонент содержащих конкретный предиктор. Предикторы изображаются графически вдоль оси у и сортируются в порядке убывания частоты; таким образом, самый верхний предиктор - это тот, который используется в наибольшем числе моделей компонент, а самый нижний - это тот, который был использован в наименьшем числе моделей. Показываются 10 верхних предикторов.

Предикторы, которые используются наиболее часто, обычно являются наиболее важными. Эта диаграмма бесполезна для методов, в которых набор предикторов не может меняться по моделям компонентов.

Точность моделей компонентов: Данная диаграмма является точечной диаграммой точности предсказания для моделей компонентов. Каждая точка представляет одну или несколько моделей компонентов с уровнем точности, изображенном графически вдоль оси у. Наведите указатель мыши на любую точку, чтобы получить информацию о соответствующей отдельной модели компонента.

Опорные линии. Для этого графика используются цветные линии для ансамбля, а также опорная модель и простые модели. Рядом с линией, соответствующей модели, которая будет использована для скоринга, стоит переключатель.

Интерактивность. Диаграмма обновится, если изменить правило объединения.

Ансамбли, созданные с помощью бустинга. Для ансамблей, созданные с помощью бустинга, выводится диаграмма с линиями.

Подробности о моделях компонентов: Эта таблица выводит информацию о моделях компонентов, представленных построчно. По умолчанию модели компонентов отсортированы в порядке возрастания номеров модели. Строки можно отсортировать в возрастающем или убывающем порядке по значениям любого столбца.

Модель. Номер, показывающий порядок, в котором модели компонентов были созданы.

Точность. Общая точность, выраженная в виде процента.

Метод. Метод моделирования.

Предикторы. Число предикторов, использованных в модели компонента.

Размер модели. Размер модели зависит от метода моделирования: Для деревьев это количество узлов в дереве; для линейных моделей это количество коэффициентов; для нейронных сетей это количество синапсов.

Записи. Взвешенное число входных записей в обучающей выборке.

Автоматическая подготовка данных:

Этот вид выводит информацию о том, какие поля были исключены и как преобразованные поля были получены на этапе автоматической подготовки данных (ADP). Для каждого поля, которое было преобразовано или исключено, в таблице перечисляется имя поля, его роль в анализе и действие, совершенное на этапе ADP. Поля сортируются в алфавитном порядке имен полей по возрастанию.

Действие **Урезать выбросы**, если показано, означает, что те значения непрерывных предикторов, которые лежат вне границ отсечения (определяемых тремя стандартными отклонениями от среднего значения), заменяются значением границы отсечения.

Слепки моделей расщепления

Слепок модели расщепления предоставляет доступ ко всем отдельным моделям, созданным расщеплениями.

Слепок модели расщепления содержит:

- - список всех созданных моделей расщепления вместе с набором статистических данных для каждой из этих моделей
- - информацию об общей модели

Для дальнейшего изучения каждой индивидуальной модели ее можно открыть в списке моделей расщепления.

Средство просмотра расщепленных моделей

На вкладке Модели перечисляются все содержащиеся в слепке модели и в различных формах предоставляется статистика о моделях расщепления. На этой вкладке есть две общие формы, зависящие от узла моделирования.

Сортировать по. Используйте этот список для выбора, в каком порядке будут перечисляться модели. Этот список можно отсортировать на основе значений любого из столбцов вывода в возрастающем или убывающем порядке. Как вариант, щелкните по заголовку столбца, чтобы отсортировать список по этому столбцу. По умолчанию используется общая точность в убывающем порядке.

Меню Показать/скрыть столбцы. Нажмите эту кнопку, чтобы вывести меню, в котором можно выбрать индивидуальные столбцы, которые будут показаны или скрыты.

Вид. Если вы используете разные разделы, можно выбрать просмотр результатов или для данных обучения, или для данных испытания.

Для каждого расщепления подробности показываются следующим образом:

График. Миниизображение, показывающее распределение данных для этой модели. Если слепок находится на холсте, дважды щелкните по миниизображению, чтобы открыть полноразмерный график.

Модель. Значок типа модели. Дважды щелкните по этому значку, чтобы открыть слепок модели для этого конкретного расщепления.

Поля расщепления. Поля, обозначенные на узле моделирования как поля расщепления, с их различными возможными значениями.

Число записей в расщеплении. Количество записей, включенных в это конкретное расщепление.

Число используемых полей. Ранжирует модели расщепления на основании числа используемых входных полей.

Общая точность (%). Процентная доля записей, правильно предсказанных моделью расщепления, относительно общего количества записей в этом расщеплении.

Расщепление. Заголовок столбца показывает поля, используемые для создания расщепления, и ячейки являются значениями расщепления. Дважды щелкните по любому расщеплению, чтобы открыть средство просмотра моделей для построения данного расщепления.

Точность. Общая точность, выраженная в виде процента.

Размер модели. Размер модели зависит от метода моделирования: для деревьев это количество узлов в дереве; для линейных моделей это количество коэффициентов; для нейронных сетей это количество синапсов.

Записи. Взвешенное число входных записей в обучающей выборке.

Использование слепков моделей в потоках

Слепки моделей помещаются в поток, позволяя оценить новые данные и сгенерировать новые узлы. **Скоринг** данных позволяют вам использовать информацию, добытую при построении моделей, чтобы создать предсказания для новых записей. Для просмотра результатов скоринга нужно присоединить к слепку конечный узел (то есть, узел обработки или вывода) и выполнить этот конечный узел.

Для некоторых моделей слепки могут предоставить также дополнительную информацию о качестве предсказания, например, значения достоверности или расстояния от центров кластеров. В процессе генерирования новых узлов их можно создать на основе структуры сгенерированной модели. Например, большинство моделей, выполняющих отбор входных полей, позволяют генерировать узлы Фильтр, через которые будут передаваться только входные поля, идентифицированные моделью как важные.

Использовать слепок модели для скоринга данных

1. Соедините слепок модели с источником данных или с потоком, который передаст на него данные.
2. Добавьте или присоедините к слепку модели один или несколько узлов обработки или вывода (таких как узел Таблица или узел Анализ).
3. В слепке модели выполните исходящий поток одного из узлов.

Примечание: Для скоринга данных нельзя использовать узел неуточненных данных. Для оценки данных на основании модели правил связывания используйте узел неуточненных правил для генерирования слепка

набора правил и используйте этот слепок для скоринга. Дополнительную информацию смотрите в разделе “Генерирование набора правил из слепка модели связывания” на стр. 237.

Использовать слепок модели для генерирования узлов обработки

1. Просмотрите модель на палитре или измените ее на холсте потока.
2. Выберите нужный тип узла в меню Создать окна браузера слепков моделей. Доступные опции будут различными в зависимости от типа слепка модели. Обратитесь к конкретному слепку модели, чтобы узнать подробности, что именно можно сгенерировать из конкретной модели.

Повторное генерирование узла моделирования

Если у вас есть слепок модели, который нужно изменить или обновить, а используемый для создания модели поток недоступен, можно использовать эту опцию для повторного создания узла моделирования с теми же опциями, что и для создания исходной модели.

Чтобы повторно построить модель, щелкните правой кнопкой мыши по модели на палитре моделей и выберите пункт **Генерировать узел моделирования**.

Как вариант, при просмотре любой модели выберите пункт **Генерировать узел моделирования** в меню Создать.

В большинстве случаев сгенерированный узел моделирования должен быть функционально идентичен узлу, использованному для создания исходной модели.

- Для моделей Дерево решений в узле можно сохранить также дополнительные параметры, заданные в интерактивном сеансе, и в повторно сгенерированном узле моделирования будет включена опция **Использовать директивы дерева**.
- Для моделей Список решений будет включена опция **Использовать сохраненную информацию интерактивного сеанса**. Дополнительную информацию смотрите в разделе “Опции модели списка решений” на стр. 138.
- Для моделей Временные ряды включена опция **Продолжить оценку с использованием существующих моделей**, позволяющая повторно сгенерировать предыдущую модель с текущими данными. Дополнительную информацию смотрите в разделе “Опции модели временных рядов” на стр. 257.

Импорт и экспорт моделей в виде PMML

Язык разметки предсказательных моделей (predictive model markup language, PMML) - это формат XML, описывающий исследование данных и статистические модели, в том числе входные данные моделей, а также преобразования, используемые для подготовки данных к исследованию, и параметры, определяющие сами модели. IBM SPSS Modeler может экспортировать и импортировать PMML, делая возможным совместное использование моделей несколькими поддерживающими этот формат прикладными программами, например, IBM SPSS Statistics.

Более подробную информацию о PMML смотрите на сайте группы исследования данных (*http://www.dmg.org*).

Экспортировать модель

Экспорт PMML поддерживается для большинства типов моделей, сгенерированных в IBM SPSS Modeler. Дополнительную информацию смотрите в разделе “Типы моделей, поддерживающие PMML” на стр. 49.

1. Щелкните правой кнопкой мыши по слепку модели на палитре моделей. (Можно также дважды щелкнуть по слепку модели на холсте и выбрать меню Файл).
2. В меню щелкните по **Экспортировать PMML**.
3. В диалоговом окне Экспорт (или Сохранить) укажите каталог назначения и уникальное имя для модели.

Примечание: Опции для экспорта PMML можно изменить в диалоговом окне Опции пользователя. В главном меню выберите:

Инструменты > Опции > Опции пользователя

и перейдите на вкладку PMML.

Импортировать модель, сохраненную как PMML

Модели, экспортированные в виде PMML из IBM SPSS Modeler или других прикладных программ, можно импортировать на палитру моделей. Дополнительную информацию смотрите в разделе “Типы моделей, поддерживающие PMML”.

1. На палитре моделей щелкните по ней правой кнопкой мыши и выберите из меню **Импортировать PMML**.
2. Выберите файл для импорта и нужным образом задайте опции для меток переменных.
3. Нажмите кнопку **Открыть**.

Использовать метки переменных, если они есть в модели. Для переменных в словаре данных PMML может задавать и имена, и метки переменных (например, Referrer ID для *RefID*). Выберите эту опцию для использования меток переменных, если они присутствуют в исходном экспортированном PMML.

Если выбрана опция меток переменных, но в PMML нет меток переменных, имена переменных используются обычным способом.

Типы моделей, поддерживающие PMML

Экспорт PMML

IBM SPSS Modelerмодели. Следующие созданные в IBM SPSS Modeler модели можно экспортировать как PMML 4.0:

- Дерево C&R
- QUEST
- CHAID
- линейная регрессия
- Нейросеть
- C5.0
- логистическая регрессия
- Обобщенная линейная регрессия
- SVM
- Априорный анализ
- Carma
- K-средние
- Коонена
- TwoStep
- GLMM (поддерживается только для моделей GLMM с только фиксированными эффектами)
- Список решений
- Модель Кокса
- Последовательность (скоринг для моделей PMML последовательностей не поддерживается)
- StatisticsМодель

Внутренние модели баз данных. Если модели сгенерированы с помощью внутренних алгоритмов баз данных, экспорт PMML доступен только для моделей IBM InfoSphere Warehouse. Модели, созданные при помощи

Analysis Services от Microsoft или Oracle Data Miner, экспортировать нельзя. Обратите внимание также на то, что модели IBM, экспортированные как PMML, нельзя импортировать обратно в IBM SPSS Modeler.

Импорт PMML

IBM SPSS Modeler может импортировать и оценивать модели PMML, сгенерированные текущими версиями всех продуктов IBM SPSS Statistics, в том числе модели, экспортированные из IBM SPSS Modeler, а также модель или ее преобразованную в PMML форму, сгенерированную IBM SPSS Statistics 17.0 или более новой версии. Существенно, что это относится к любому PMML, который может оценить механизм скоринга, со следующими исключениями:

- Нельзя импортировать модели Априори, CARMA, обнаружения аномалий и последовательности.
- После импорта в IBM SPSS Modeler модели PMML нельзя просматривать, хотя их можно использовать при скоринге. (Обратите внимание на то, что в первую очередь это относится к моделям, экспортированным из IBM SPSS Modeler. Чтобы избежать этого ограничения, экспортируйте модель как файл сгенерированной модели [*.*gm*], а не PMML).
- Модели IBM InfoSphere Warehouse, экспортированные как PMML, нельзя импортировать.
- При импорте происходит ограниченная проверка, но при попытке проведения скоринга модели - полная. Таким образом, возможны ситуации, когда импорт прошел успешно, но скоринг завершился неудачно или привел к неправильным результатам.

Публикация моделей для адаптера скоринга

Вы можете опубликовать модели на сервере баз данных, для которых установлен адаптер скоринга. Адаптер скоринга позволяет выполнить скоринг моделей в базе данных с использованием возможностей пользовательских функций (user-defined function, UDF) базы данных. Выполнение скоринга в базе данных исключает необходимость извлекать данные перед скорингом. Публикация в адаптер скоринга генерирует также несколько примеров SQL для выполнения UDF.

Чтобы опубликовать в адаптер скоринга:

1. Дважды щелкните по слепку модели, чтобы открыть ее.
2. В меню слепка модели выберите:

   Файл > Опубликовать для адаптера скоринга сервера
3. В диалоговом окне заполните соответствующие поля и нажмите кнопку **OK**.

Соединение с базой данных. Подробности соединения с базой данных, которую вы хотите использовать для модели.

ID опубликования. (Только для баз данных DB2 for z/OS) Идентификатор для модели. Если вы повторно строите ту же модель и используете тот же ID опубликования, сгенерированный SQL остается тем же, поэтому можно перепостроить модель без необходимости изменять прикладную программу, использующую ранее сгенерированный SQL. (Для других баз данных сгенерированный SQL уникален для модели).

Сгенерировать SQL примера. Если выбрана эта опция, генерируется пример SQL в файл, заданный в поле **Файл**.

Неуточненные модели

Неуточненная модель содержит информацию, извлеченную из данных, но она не предназначена для непосредственного генерирования предсказаний. Это означает, что такую модель нельзя добавлять в потоки. Неуточненные модели выводятся на палитре сгенерированных моделей в виде значка “необработанных алмазов”.

Рисунок 27. Значок неуточненной модели

Для просмотра информации о неуточненной модели правил щелкните правой кнопкой мыши по модели и выберите из контекстного меню пункт **Просмотр**. Как и для других моделей, сгенерированных в IBM SPSS Modeler, на различных вкладках выводится сводная информация и правила о созданной модели.

Генерирование узлов. При помощи меню Создать вы можете создавать новые узлы на основе правил.

- **Узел выбора.** Создает новый узел выбора для выбора записей, к которым применимо текущее выбранное правило. Если правило не выбрано, эта опция отключается.
- **Набор правил.** Генерирует узел Набор правил, чтобы предсказывать значения для одного поля назначения. Дополнительную информацию смотрите в разделе “Генерирование набора правил из слепка модели связывания” на стр. 237.

Глава 4. Модели экранирования

Экранирование полей и записей

Во время предварительных стадий анализа можно использовать несколько узлов моделирования, чтобы найти поля и записи, которые с наибольшей вероятностью будут представлять интерес при моделировании. Узел Выбор показателей можно использовать для экранирования полей и их ранжирования по важности, а узел Выявление аномалий - для поиска необычных записей, которые не соответствуют известным структурам "нормальных" данных.

Узел выбора возможностей изучает входные поля на возможность удаления, основываясь на наборе критериев (таких как процентная доля пропущенных значений); затем этот узел ранжирует важность оставшихся полей по отношению к заданному полю назначения. Например, если у набора данных сотни потенциальных входных полей, какие из них потенциально наиболее полезны при моделировании исхода лечения пациента?

Узел выявления аномалий определяет необычные наблюдения, или выбросы, которые не соответствуют структуре “нормальных” данных. При помощи этого узла можно находить выбросы даже в том случае, если они не подходят ни под какие ранее известные шаблоны или вы точно не уверены, что именно ищете.

Обратите внимание на то, что Выявление аномалий идентифицирует необычные записи или наблюдения через кластерный анализ на основе набора полей, выбранных в модели безотносительно какого-либо поля назначения (зависимой величины) и независимо от того, соответствуют ли эти поля структуре, которую вы хотите предсказать. Поэтому вам может потребоваться использовать обнаружение аномалий в комбинации с выбором показателей или другим способом экранирования и ранжирования полей. Например, можно использовать набор показателей для определения наиболее важных полей, связанных с конкретным полем назначения, а затем использовать обнаружение аномалий для нахождения записей, которые наиболее необычны по отношению к этим полям. (Альтернативным подходом может быть построение модели дерева решений с последующим изучением всех неправильно классифицированных данных как потенциальных аномалий. Однако этот способ будет более трудным для репликации или автоматизации на большом масштабе).

Узел выбора возможностей

Задачи исследования данных могут включать в себя сотни и даже тысячи полей, которые потенциально можно использовать как входные. В результате большая доля времени и усилий будет потрачена на проверку, какие именно поля или переменные будут включены в модель. Чтобы сузить выбор, можно использовать алгоритм выбора возможностей для идентификации наиболее важных в данном анализе полей. Например, если вы пытаетесь предсказать исход лечения пациента на основании большого числа факторов, какие факторы с наибольшей вероятностью окажутся важными?

Выбор возможностей состоит из трех шагов:

- **Экранирование.** Удаляет неважные или проблемные входные данные (записи) или наблюдения, для которых во входных полях пропущено слишком много значений, или же изменение которых или слишком велико, или слишком мало, чтобы быть полезным.
- **Ранжирование.** Сортирует оставшиеся входные поля и назначает их ранги на основании важности.
- **Выбор.** Определяет подмножество возможностей для использования в последовательных моделях, например, сохраняя только самые важные входные поля и фильтруя или исключая все остальные.

В наше время, когда многие организации перегружены слишком большим количеством данных, выигрыш от выбора возможностей для упрощения и ускорения процесса моделирования может быть существенным.

Быстро фокусируясь на наиболее важных полях, вы можете сократить объем необходимых вычислений; проще обнаруживаются небольшие, но важные взаимосвязи, которые в противном случае можно пропустить; и, в конечном итоге, получить более простые и точные модели, допускающие более понятные объяснения. Сократив количество используемых в модели полей, вы можете обеспечить возможность сокращения времени скоринга, а также объема данных, собираемых при последующих итерациях.

Пример. У телефонной компании есть хранилище данных, содержащее информацию об откликах на специальные предложения от 5000 клиентов компании. Эти данные включают в себя большое число полей, содержащих сведения о возрасте, занятости и доходе клиентов, а также статистику их телефонных звонков. Три поля назначения показывают, откликнулся ли клиент на каждое из трех предложений. Компания хочет использовать эти данные для помощи в предсказаниях, какие клиенты наиболее вероятно откликнутся на аналогичные предложения в будущем.

Требования. Одно поле назначение (поле, для которого задана роль *Назначение*) вместе с несколькими входными полями, которые вы хотите экранировать или ранжировать относительно назначения. И у поля назначения, и у входных полей может быть уровень измерения *Количественный* (числовой диапазон) или *Категориальный*.

Параметры моделей выбора возможностей

Параметры на вкладке Модель включают в себя стандартные опции модели, а также параметры, позволяющие точнее настраивать критерии для экранирования входных полей.

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Экранирование входных полей

Экранирование включает в себя удаление входных полей или наблюдений, которые не добавляют какой-либо полезной информации по отношению к взаимосвязи полей - входных и назначения. Опции экранирования основываются на атрибутах полей, в связи с которыми возникают вопросы по поводу их прогнозирующей способности относительно выбранного поля назначения. Экранированные поля исключаются из вычислений, используемых для ранжирования входных полей, а также их можно отфильтровать или удалить из данных, используемых для моделирования.

Поля можно экранировать на основе следующих критериев:

- **Максимальный процент пропущенных значений.** Экранирует поля с чрезмерным количеством пропущенных значений, которое выражается процентной долей от полного числа записей. Поля с большой процентной долей пропущенных значений предоставляют мало информации для предсказаний.
- **Максимальный процент записей в одной категории.** Экранирует поля, у которых чрезмерно много записей (относительно их общего количества) попадает в одну категорию. Например, если у 95% покупателей в базе данных машины одного типа, включение этой информации не помогает отличить одного покупателя от другого. Экранируются все поля, превышающие заданный максимум для этого показателя. Эта опция применима только к категориальным полям.
- **Максимальное число категорий как процент записей.** Экранирует поля с чрезмерно большим количеством категорий относительно общего количества записей. Если большая процентная доля категорий содержит всего по одному наблюдению, использование такого поля может быть ограниченным. Например, если все покупатели носят разные шляпы, эта информация не поможет смоделировать шаблоны поведения. Эта опция применима только к категориальным полям.
- **Минимальный коэффициент вариации.** Экранирует поля, для которых коэффициент изменчивости не больше заданного минимума. Этот показатель равен отношению среднеквадратичного отклонения значений входного поля к его среднему значению. Если это значение близко к нулю, у значений переменной нет большой изменчивости. Эта опция применима только к количественным полям (числовой диапазон).

- **Минимальное стандартное отклонение.** Экранирует поля, для которых среднеквадратичные отклонения не больше заданного минимума. Эта опция применима только к количественным полям (числовой диапазон).

Записи с пропущенными данными. Записи или наблюдения у которых отсутствуют значения для поля назначения или для всех входных полей, автоматически исключаются из всех вычислений, используемых для ранжирования.

Опции выбора возможностей

Вкладка Опции позволяет задать значения параметров по умолчанию для выбора или исключения входных полей в слепке модели. Затем модель можно добавить в поток для выбора поднабора полей, которые будут использоваться при дальнейшем построении модели. Кроме того, эти значения можно переопределить, выбрав дополнительные поля или отменив выбор таковых в браузере моделей поле генерирования модели. Однако значения параметров по умолчанию делают возможным применение слепка модели без дополнительных изменений, что может оказаться особенно полезным для работы со сценариями.

Дополнительную информацию смотрите в разделе “Результаты модели выбора возможностей” на стр. 56.

Доступны следующие параметры:

Все ранжированные поля. Выбирает поля на основе их ранжирования как *важных*, *пограничных* или *маловажных*. Для каждого ранга можно отредактировать метку, а также значения отсечения, при помощи которых будут назначаться записи для того или иного ранга.

Максимальное число полей. Выбирает *n* верхних полей на основе важности.

Важность больше чем. Выбирает все поля с важностью больше указанного значения.

Поле назначения всегда сохраняется, независимо от варианта выбора.

Опции ранжирования важности

Все категориальные. Когда все входные поля и поля назначения категориальные, важность можно ранжировать на основе любого из четырех показателей:

- **Хи-квадрат Пирсона.** Проверяет независимость полей (входных и назначения) без обозначения силы или направления любой существующей взаимосвязи.
- **Хи-квадрат отношения правдоподобия.** Аналогично хи-квадрат Пирсона, но проверяется также независимость между полями назначения и входными.
- **V Крамера.** Показатель связи, основанный на статистике хи-квадрат Пирсона. Значения изменяются от 0 (нет связи) до 1 (абсолютная связь).
- **Лямбда.** Показатель связи, отображающий пропорциональное сокращение ошибки, когда переменная используется для предсказания значения назначения. Значение 1 означает, что входное поле идеально предсказывает поле назначения, а значение 0 - что входное поле не дает никакой информации о поле назначения.

Некоторые категориальные. Когда некоторые, но не все, поля категориальные и поле назначения также категориальное, важность можно ранжировать на основе хи-квадрат Пирсона или отношения правдоподобия. (*V* Крамера и лямбда недоступны, если не все входные поля категориальные).

Категориальные и количественные поля. При ранжировании категориальных входных полей для количественного поля назначения или наоборот (одни или другие категориальные, но не одновременно), используется *F*-статистика.

Оба количественные. При ранжировании количественных входных полей для количественных полей назначения используется *t*-статистика на основании корреляционного коэффициента.

Слепки моделей выбора возможностей

Слепки моделей выбора возможностей выводят важность каждого входного поля для выбранного поля назначения по ранжированию, выполненному узлом выбора возможностей. Перечисляются также и поля, экранированные до ранжирования. Дополнительную информацию смотрите в разделе “Узел выбора возможностей” на стр. 53.

При запуске потока, содержащего слепок модели выбора возможностей модель действует как фильтр, сохраняющий только выбранные входные поля, как это обозначено при текущем выборе на вкладке Модель. Например, можно выбрать все поля, ранжированные как важные (одна из опций по умолчанию), или вручную выбрать подмножество полей на вкладке Модель. Поле назначения также сохраняется, независимо от варианта выбора. Все остальные поля исключаются.

Фильтрование основано только на имени поля; например, если выбраны входные поля *возраст* и *доход*, будет сохранено любое поле, у которого одно из этих имен. Эта модель не изменяет ранжирования полей на основании новых данных; она просто фильтрует поля по выбранным именам. Поэтому нужно с осторожностью применять эту модель к новым или измененным данным. При сомнениях рекомендуется повторно сгенерировать модель.

Результаты модели выбора возможностей

На вкладке Модель слепка модели Выбор возможностей ранг и важность для всех входных полей выводятся на верхней панели, и это позволяет вам выбрать поля для фильтрования, используя переключатели в левом столбце. При запуске потока сохраняются только выбранные поля; остальные поля отбрасываются. Варианты выбора по умолчанию основаны на опциях, заданных для узла построения модели, но вы можете выбрать дополнительные поля или отменить их выбор нужным вам образом.

На нижней панели перечисляются входные поля, исключенные из ранжирования из-за процентной доли значений отсутствия или по другому критерию, заданному на узле моделирования. Как и для ранжированных полей, вы можете выбрать включение или отбрасывания этих полей, используя переключатели в левом столбце. Дополнительную информацию смотрите в разделе “Параметры моделей выбора возможностей” на стр. 54.

- Чтобы отсортировать список по столбцу ранга, имени поля, важности или иному другому из выводимых столбцов, щелкните по его заголовку. Или используйте панель инструментов, выберите нужный элемент из списка Сортировать по и с помощью кнопок со стрелками вверх-вниз измените направление сортировки.
- При помощи панели инструментов можно включить или выключить все поля и открыть диалоговое окно Включить поля, где можно отобрать поля по рангу или важности. Расширить выбор можно, удерживая клавиши Shift или Ctrl при щелчке по полям, а с помощью клавиши табуляции можно включать и выключать группу выбранных полей. Дополнительную информацию смотрите в разделе “Выбор полей по важности”.
- В пояснении под таблицей выводятся значения порогов для ранжирования входных полей как важных, пограничных или маловажных. Эти значения задаются в режиме моделирования. Дополнительную информацию смотрите в разделе “Опции выбора возможностей” на стр. 55.

Выбор полей по важности

При скоринге данных с использованием слепка модели Выбор возможностей будут сохранены все поля, выбранные из списка ранжированных или экранированных полей, отмеченных переключателями в левом столбце. Остальные поля будут отброшены. Чтобы изменить выбор, можно использовать панель инструментов для доступа к диалоговому полю Проверить поля, где можно выбрать поля по рангу или важности.

Все маркированные поля. Выбирает все поля, отмеченные как важные, пограничные или не важные.

Максимальное число полей. Позволяет выбрать *n* первых полей на основании важности.

Важность больше чем. Выбирает все поля с важностью больше указанного порога.

Генерирование фильтра из модели выбора возможностей

На основании результатов модели Выбор возможностей вы можете использовать диалоговое окно Генерировать фильтр из возможности, чтобы сгенерировать один или несколько узлов Фильтр, которые включают или исключают подмножества полей на основании важности относительно заданного поля назначения. Хотя сам слепок модели можно использовать как фильтр, данная опция обеспечивает гибкость в экспериментировании с различными подмножествами полей без копирования или изменения модели. Поле назначения всегда сохраняется фильтром независимо от выбора его включения или исключения.

Включить/исключить. Вы можете выбрать включение или исключение полей, например, включить первые в списке 10 полей или исключить все поля, отмеченные как не важные.

Выбранные поля. Включает или исключает все поля, выбранные в настоящее время в таблице.

Все отмеченные поля. Выбирает все поля, помеченные как важные, пограничные или не важные.

Максимальное число полей. Позволяет выбрать первые *n* полей в списке на основании важности.

Важность больше чем. Выбирает все поля с важностью, большей заданного порога.

Узел выявления аномалий

Модели выявления аномалий служат для обнаружения выбросов, необычных наблюдений в данных. В отличии от других методов моделирования, записывающих правила касательно необычных наблюдений, модели выявления аномалий записывают информацию о чертах нормального поведения. Это дает возможность находить даже такие выбросы, которые не отвечают известным схемам, и это особенно полезно в таких областях, как выявление мошенничества, где постоянно появляются новые схемы. Выявление аномалий - неконтролируемый метод, то есть для него не требуется набор обучающих данных с известными случаями мошенничества в качестве отправной точки.

В отличии от традиционных методов поиска выбросов, где обычно отслеживается всего лишь одна или две переменных в одном просмотре, метод выявления аномалий способен исследовать большое число полей, выявляя кластеры или равноправные группы, в которые попадают сходные записи. Затем, чтобы выявить возможные аномалии, каждую запись сравнивают с другими записями той же равноправной группы. Чем дальше отстоит наблюдение от нормального центра, тем вероятнее, что это необычный случай. Например, алгоритм может разложить записи по трем кластерам и затем пометить флагом те записи, которые оказались далеки от центров кластеров.

Каждой записи присваивается индекс аномальности, равный отношению показателя отклонения от группы к среднему этих показателей по данному кластеру. Чем больше этот индекс, тем больше отклонение по сравнению со средним. При обычных обстоятельствах наблюдения с индексом аномальности меньше 1 да и вплоть до 1,5 не следует считать аномалиями, поскольку их отклонение близко к среднему или лишь немногим больше среднего. Однако наблюдения, у которых этот индекс превышает 2, можно считать основательными кандидатами в аномалии, поскольку их отклонение как минимум вдвое превышает среднее.

Выявление аномалий - это исследовательский метод, предназначенный для быстрого выявления необычных наблюдений или записей, которые становятся кандидатами для дальнейшего анализа. Выявляются лишь *подозрительные* на аномалии случаи, и пристальное исследование может такого рода случай подтвердить или не подтвердить. Даже если запись оказалась допустимой, вы все же можете заэкранировать ее от

остальных данных при построении моделей. Кроме того, если алгоритм неоднократно выдает ложные аномалии, это может указывать на ошибку или артефакт в процессе сбора данных.

Обратите внимание на то, что выявление аномалий обнаруживает необычные записи или наблюдений посредством кластерного анализа, основанного на наборе полей, выбранных в модели безотносительно к какому-либо конкретному полю назначения (зависимому полю) или проверки, какое отношение имеют эти поля к подозрительным структурам, которые вы пытаетесь предсказать. По этой причине есть смысл использовать выявление аномалий совместно с отбором показателей или иной технологией экранирования или ранжирования полей. Например, при помощи отбора показателей можно определить наиболее важные поля по отношению к заданному полю назначения, а затем использовать выявление аномалий, чтобы найти поля, наиболее необычные по этим полям. (Другой подход - построить модель дерева решений и затем изучить ошибочно классифицированные записи как потенциальные аномалии. Но такой метод будет трудно воспроизвести или автоматизировать в больших масштабах.)

Пример. При проверке грантов на развитие сельского хозяйства на возможные случаи мошенничества обнаружение аномалий помогает выявить отклонения от нормы, выделяя аномальные записи, требующие дальнейшего исследования. Особенный интерес представляют заявки на гранты, в которых запрашивается несоразмерно большая (или слишком маленькая) сумма для фермы указанного типа и размера.

Требования. Одно или несколько входных полей. Обратите внимание на то, что в качестве входных полей могут использоваться только поля, для которых задана роль **Входное** при помощи узла источника или типа. Поля назначения (задана роль **Назначение** или **Обе**) игнорируются.

Достоинства. Отмечая наблюдения, которые *не* соответствуют известному набору правил, в противоположность тем, которые соответствуют этому набору, модели выявления аномалий могут идентифицировать необычные наблюдения, даже если их необычность не укладывается в ранее выявленные закономерности. При использовании в сочетании с выбором функций выявление аномалий позволяет проверять большие объемы данных для сравнительно быстрого обнаружения записей, представляющих наибольший интерес.

Опции моделей выявления аномалий

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Определить значение отсечения для аномалии на основе. Задает метод, используемый для определения значения отсечения, при котором следует отмечать аномалии. Доступны следующие параметры:

- **Минимальный уровень индекса аномальности.** Задает минимальное пороговое значение для отметки аномалий. Будут отмечены записи, отвечающие этому порогу или превышающие его.
- **Процент наиболее аномальных записей в данных обучения.** Автоматически устанавливает порог на уровне, при котором отмечается заданный процент записей в обучающих данных. Полученный в результате порог отсечения включается в модель в качестве параметра. Обратите внимание на то, что эта опция определяет, как задается пороговое значение, а *не* фактический процент записей, отмечаемых при скоринге. Фактические результаты скоринга могут быть различными в зависимости от данных.
- **Число наиболее аномальных записей в данных обучения.** Автоматически устанавливает порог на уровне, при котором отмечается заданное число записей в обучающих данных. Полученный в результате порог включается в модель в качестве параметра. Обратите внимание на то, что эта опция определяет, как задается пороговое значение, а *не* конкретное число записей, отмечаемых при скоринга. Фактические результаты скоринга могут быть различными в зависимости от данных.

Примечание: Независимо от того, как определяется значение отсечения, оно не влияет на исходное значение индекса аномальности, выводимое для каждой записи. Оно просто задает порог, при котором записи отмечаются как аномальные при оценке или скоринге модели. Если в дальнейшем вы захотите проверить большее или меньшее количество записей, воспользуйтесь узлом Выбрать для определения поднабора записей по значению индекса аномальности (`$O-AnomalyIndex > X`).

Число полей аномалий к отчету. Задает число выводимых в отчете полей как указание, почему конкретная запись отмечена как аномальная. В отчете выводятся самые аномальные поля, то есть демонстрирующие наибольшее отклонение от нормального значения поля для кластера, которому назначена запись.

Дополнительные опции выявления аномалий

Чтобы задать опции для пропущенных значений и другие параметры, включите режим **Дополнительно** на вкладке Дополнительно.

Поправочный коэффициент. Значение, используемое для балансирования относительного веса, присвоенного непрерывным (числовой диапазон) и категориальным полям при вычислении расстояния. Чем выше это значение, тем сильнее влияние непрерывных полей. Это значение должно быть ненулевым.

Автоматически вычислять число равноправных групп. Обнаружение аномалий может использоваться для быстрого анализа большого количества возможных решений по выбору оптимального числа равноправных групп для обучающих данных. Диапазон решений можно расширить или сузить, задав минимальное и максимальное число равноправных групп. При больших значениях система сможет исследовать более широкий диапазон возможных решений, однако за это придется расплачиваться возрастанием времени обработки.

Задать число равноправных групп. Если вы знаете, сколько кластеров нужно включить в модель, выберите эту опцию и введите число равноправных групп. Выбор этой опции обычно приводит к повышению производительности.

Уровень и коэффициент шума. Эти параметры определяют, как обрабатываются выбросы во время двухэтапной кластеризации. На первой стадии дерево свойств кластеров используется для сжатия данных с очень большим количеством отдельных записей до приемлемого числа кластеров. Это дерево строится на основе мер сходства, и когда на узле дерева скапливается слишком много записей, он расщепляется на дочерние узлы. На втором этапе на терминальных узлах дерева свойств кластера начинается иерархическая кластеризация. Обработка шумов включена на первом проходе по данным и выключена на втором проходе. Наблюдения в кластере шумов после первого прохода по данным назначаются обычным кластерам во втором проходе.

- **Уровень шума.** Укажите значение от 0 до 0,5. Этот параметр применим, только если дерево свойств кластеров заполняется на фазе роста, то есть оно не может больше принимать наблюдения в терминальный узел и никакой терминальный узел не может быть расщеплен.

  Если дерево свойств кластеров заполнено и для уровня шума задано значение 0, порог будет увеличен и дерево будет перестроено по всем наблюдениям. После окончательного разбиения на кластеры, значения, которые не могут быть приписаны к кластерам, помечаются как выбросы. Кластеру выбросов дается идентификационный номер –1. Кластер выбросов не включен в число кластеров, то есть если задать *n* кластеров и обработку шума, алгоритм выдаст *n* и один кластер шума. На практике увеличение этого значения дает алгоритму больше свободы, чтобы поместить необычные записи в дерево, а не присваивать их отдельному кластеру выбросов.

  Если дерево свойств кластеров заполняется и уровень шума больше 0, это дерево будет перестроено после помещения данных из разреженных терминальных элементов в их собственный терминальный элемент шума. Терминальный элемент считается разреженным, если отношение числа наблюдений в таком элементе к числу наблюдений в самом большом элементе меньше уровня шума. После того, как дерево построено, выбросы будут помещены в дерево свойств кластеров (СК), если это возможно. В противном случае выбросы отбрасываются для второй фазы кластеризации.
- **Коэффициент шума.** Определяет часть памяти, выделенной для компонента, который должен использоваться для буферизации шума. Возможные значения: от 0 до 0,5. Если включение данного наблюдения в терминальный элемент дерева даст плотность, меньшую порогового значения, терминальный элемент не расщепляется. Если плотность превышает порог, терминальный элемент будет расщеплен, и в дереве свойств кластеров появится новый маленький кластер. На практике увеличение этого параметра может заставить алгоритм быстрее стремиться к более простому дереву.

Заполнить пропущенные значения. Для непрерывных полей заменяет пропущенные значения на средние для поля. Для категориальных полей пропущенные категории объединяются и рассматриваются как допустимая категория. Если отключить эту опцию, любые записи с пропущенными значениями будут исключены из анализа.

Слепки моделей выявления аномалий

Слепки моделей выявления аномалий содержат всю информацию, захваченную моделью выявления аномалий, а также информацию об обучающих данных и процессе оценки.

При выполнении потока, содержащего слепок модели выявления аномалий, к потоку добавляются новые поля в соответствии с опциями, выбранными на вкладке Параметры в слепке модели. Дополнительную информацию смотрите в разделе “Параметры моделей выявления аномалий”. Новые имена полей основываются на названии модели с префиксом *$O*, как показано в следующей таблице.

Таблица 6. Генерирование имени нового поля.

Имя поля	Описание
$O-Anomaly	Поле флага, указывающее, является ли запись аномальной.
$O-AnomalyIndex	Значение индекса аномальности для записи.
$O-PeerGroup	Задает равноправную группу, которой назначена запись.
$O-Field-n	Имя *n*-го из полей с наибольшими аномальностями по отклонению от кластерной нормы.
$O-FieldImpact-n	Индекс отклонения переменной для поля. Это значение выражает отклонение от нормального значения поля для кластера, которому назначена запись.

Можно (необязательно) подавить оценки для неаномальных записей, чтобы результаты легче читались. Дополнительную информацию смотрите в разделе “Параметры моделей выявления аномалий”.

Подробности моделей выявления аномалий

На вкладке Модель для сгенерированной модели Обнаружение аномалий выводится информация о равноправных группах в этой модели.

Обратите внимание на то, что выводимые размеры и статистические показатели равноправных групп - это оценки на основе обучающих данных, которые могут несколько отличаться от фактических результатов скоринга, даже если они выполнялись на тех же данных.

Сводка моделей выявления аномалий

На вкладке Сводка для слепка модели выявления аномалий выводится информация о полях, параметрах построения и процессе оценки. Выводится также число равноправных групп вместе со значением отсечения, используемым для отметки записей как аномальных.

Параметры моделей выявления аномалий

Вкладка Параметры позволяет задать опции для оценки слепка модели.

Указать аномальные записи, содержащие. Задает, как аномальные записи обрабатываются в выводе.

- **Флаг и индекс.** Создает поле флага со значением *True* для всех записей, превышающих пороговое значение, которое заложено в модели. Для каждой записи в отдельном поле выводится также индекс аномальности. Дополнительную информацию смотрите в разделе “Опции моделей выявления аномалий” на стр. 58.
- **Только флаг.** Создает поле флага без вывода индекса аномальности для каждой записи.

- **Только индекс.** Выводит индекс аномальности без создания поля флага.

Число полей аномалий к отчету. Задает число выводимых в отчете полей как указание, почему конкретная запись отмечена как аномальная. В отчете выводятся наиболее аномальные поля, то есть демонстрирующие наибольшее отклонение от нормального значения поля для кластера, которому назначена запись.

Отклонить записи. Выберите эту опцию для отбрасывания всех неаномальных записей из потока, что помогает сосредоточиться на потенциальных аномалиях в нижележащих узлах. Другой вариант - отбросить все аномальные записи, чтобы ограничить дальнейший анализ записями, которые в данной модели не отмечены как потенциальные аномалии.

Примечание: Из-за небольших различий в округлении фактическое число записей, отмеченных при оценке, может не совпадать с числом записей, отмеченных при обучении модели, даже если используются те же данные.

Глава 5. Узлы автоматического моделировани

Узлы автоклассификации оценивают и сравнивают несколько различных способов моделирования, позволяя вам испытать разнообразные подходы в одном запуске моделирования. Вы можете выбрать используемые алгоритмы и конкретные опции для каждого из них, в том числе сочетания, которые в иных случаях были бы взаимно исключающими. Например, вместо выбора одного из вариантов нейросети (быстрого, динамического или усеченного) вы можете испытать все из них. Этот узел исследует все возможные сочетания опций, ранжирует модели-кандидаты на основании заданного вами показателя и сохраняет лучшие модели для использования при скоринге или для дальнейшего анализа.

Вы можете выбирать из трех узлов автоматического моделирования, основываясь на потребностях анализа:

Узел автоклассификации создает и сравнивает несколько различных моделей для двоичных выходных данных (да или нет, уйдет клиент или останется и так далее), что позволяет выбрать лучший подход для данного анализа. Поддерживается несколько алгоритмов моделирования, что делает возможным выбор желательных для использования способов, конкретных опций для каждого из них и критериев сравнения результатов. Этот узел генерирует набор моделей на основе заданных опций и ранжирует лучших кандидатов в соответствии с заданными вами критериями.

Узел автонумерации оценивает и сравнивает модели для выходных данных в количественном числовом диапазоне при помощи нескольких разных способов. Этот узел работает аналогично другим узлам автоклассификации, допуская выбор алгоритмов для использования и экспериментирование с несколькими комбинациями опций при одном проходе моделирования. Поддерживаемые алгоритмы включают в себя нейросети, дерево C&R, CHAID, линейную регрессию, обобщенную линейную регрессию и механизмы опорных векторов (support vector machines, SVM). Модели можно сравнивать на основе корреляции, относительной ошибки или числа используемых переменных.

Узел автоматической кластеризации оценивает и сравнивает модели кластеризации, идентифицирующие группы записей со сходными характеристиками. Этот узел работает аналогично другим узлам автоматического моделирования, допуская экспериментирование с несколькими комбинациями опций при одном проходе моделирования. Модели можно сравнивать при помощи базовых показателей, пытаясь фильтровать и ранжировать с их использованием полезность моделей кластеризации и предоставить показатель на основе важности конкретных полей.

Лучшие модели сохраняются в одном слепке моделей, что позволяет просматривать и сравнивать эти модели и выбирать, какие из них будут использоваться при скоринге.

- Только для двоичных, номинальных и числовых полей назначения вы можете выбрать несколько моделей скоринга и комбинировать оценки в одном ансамбле моделей. При комбинировании предсказаний от нескольких моделей можно избежать ограничений, накладываемых на отдельную модель, и часто получить в результате более высокую общую точность, чем можно бы было достигнуть в любой единичной модели.
- Дополнительно вы можете выбрать подробное изучение результатов и генерирование узлов моделирования или слепков моделей для любой из индивидуальных моделей, которую вы хотите использовать или изучать далее.

Модели и время выполнения

В зависимости от набора данных и числа моделей узлам автоматического моделирования для выполнения может потребоваться несколько часов или даже больше. При выборе опций обратите внимание на

количество моделей, которые создаются. С практической точки зрения у вас может возникнуть желание запланировать запуски моделирования на ночное время или на выходные, когда системные ресурсы могут быть менее затребованы.

- При необходимости можно использовать узел Раздел или узел Выборка для сокращения количества записей, включенных в начальный обучающий проход. После сужения выбора до нескольких моделей-кандидатов можно восстановить полный набор данных.
- Для сокращения количества входных полей используйте выбор возможностей. Дополнительную информацию смотрите в разделе “Узел выбора возможностей” на стр. 53. Другой вариант - использовать начальные запуски моделирования для идентификации полей и опций, которые заслуживают дальнейшего изучения. Например, если лучшие из выполняемых моделей используют три одинаковых поля, это явное указание на то, что эти поля следует сохранить для окончательной модели.
- Если хотите, можно ограничить время, затрачиваемое на оценку каждой модели и задать показатели оценивания, используемые для обследования и ранжирования моделей.

Параметры алгоритма для узла автоматического моделирования

Для каждого типа модели есть свои настройки по умолчанию; вы также можете изменить эти опции, отдельно для каждого типа модели. Конкретные опции подобны доступным в отдельных узлах моделирования с той разницей, что вместо выбора той или иной настройки можно, в большинстве случаев, выбрать применение любого числа настроек. Например, при сравнении моделей нейросети можно выбрать несколько разных методов обучения, и опробовать каждый, используя или не используя начальное значение генератора псевдослучайных чисел. Будут использованы все возможные сочетания выбранных опций, что дает возможность легко сгенерировать множество различных моделей за один запуск. Однако соблюдайте осторожность, выбирая несколько значений, потому что число моделей может быстро вырасти.

Чтобы опции для каждого типа модели

1. На узле автоматического моделирования выберите вкладку **Дополнительно**.
2. Чтобы выбрать тип модели, сначала щелкните по столбцу **Параметры модели**.
3. В выпадающем меню выберите **Задать**.
4. В диалоговом окне **Параметры алгоритма** выберите опции в столбце **Опции**.

Примечание: Дополнительные опции доступны на вкладке Дополнительно диалогового окна **Параметры алгоритма**.

Правила остановки для узла автоматического моделирования

Правила остановки для узлов автоматического моделирования относятся к выполнению всего узла, а не к отдельным моделям, которые строятся этим узлом.

Ограничить общее время выполнения. (Только для моделей нейронной сети, k-средних, Коонена, двухшаговой, опорных векторов, k ближайших соседей, байесовской сети и деревьев C&R) Останавливает выполнение после заданного числа часов. В слепок модели войдут все модели, сгенерированные на данный момент; остальные модели созданы не будут.

Остановить, как только будут получены допустимые модели. Останавливает выполнение, когда модель удовлетворяет всем критериям, заданным на вкладке Отбрасывание (для узлов автоматической классификации и автоматической кластеризации) или на вкладке Модель (для узла автоматической нумерации). Дополнительную информацию смотрите в разделе “Опции отклонения узла автоклассификации” на стр. 69. Дополнительную информацию смотрите в разделе “Опции отбрасывания узла автокластеризации” на стр. 76.

Узел автоклассификации

Узел автоклассификации оценивает и сравнивает модели для номинальных (набор) и бинарных (да/нет) объектов назначения, используя несколько различных способов, что позволяет испытать разнообразные подходы при одном запуске моделирования. Вы можете выбрать используемые алгоритмы и экспериментировать с несколькими сочетаниями опций. Например, вместо выбора одного из вариантов нейросети (быстрого, динамического или усеченного) вы можете испытать все из них. Этот узел исследует все возможные сочетания опций, ранжирует модели-кандидаты на основании заданной вами меры и сохраняет лучшие модели для использования при скоринге или при дальнейшем анализе. Дополнительную информацию смотрите в разделе Глава 5, “Узлы автоматического моделировани”, на стр. 63.

Пример. У компании по розничной торговле есть хронологические данные, отслеживающие предложения, сделанные для конкретных покупателей в прошлых кампаниях. Теперь компания хочет достичь более выгодных результатов, подбирая правильные предложения для каждого из покупателей.

Требования. Поле назначения с уровнем измерения *Номинал* или *Флаг* (с заданной ролью **Назначение**) и по крайней мере одно входное поле (с заданной ролью **Ввод**). Для поля флага предполагается, что заданное для поля назначения значение *True* представляет попадание при вычислении дохода, роста и связанной статистики. У входных полей может быть уровень измерения *Количественный* или *Категориальный* с тем ограничением, что некоторые входные данные могут не подходить для некоторых типов моделей. Например, у порядковых полей, используемых как входные в моделях дерева C&R, CHAID и QUEST, должна быть числовая (а не строковая) система хранения, и они будут игнорироваться этими моделями, если задано иное. Аналогично, в некоторых случаях количественные входные поля могут быть дискретизированы. Это те же требования, что и при использовании отдельных узлов моделирования; например, модель байесовской сети работает одинаково при генерировании и из узла Байесовская сеть, и из узла Автоклассификация.

Поля частоты и веса. Частота и вес используются для предоставления дополнительной важности некоторым записям по сравнению с остальными, так как, например, пользователь знает, что набор данных сборки недостаточно представляет раздел заполнения родительских элементов (вес) или одна запись представляет несколько идентичных наблюдений (частота). Если задано поле частоты, оно может использоваться моделями дерева C&R, CHAID, QUEST, списка решений и байесовской сети. Поле веса может использоваться моделями C&RT, CHAID и C5.0. Другие типы моделей будут игнорировать эти поля, но в любом случае построят свои модели. Поля частоты и веса используются только для построения моделей и не рассматриваются при оценке или скоринге моделей. Дополнительную информацию смотрите в разделе “Использование полей частоты и веса” на стр. 32.

Поддерживаемые типы моделей

К поддерживаемым типам моделей относятся нейросети, дерево C&R, QUEST, CHAID, C5.0, логистическая регрессия, список решений, байесовская сеть, дискриминант, ближайшие соседи и SVM. Дополнительную информацию смотрите в разделе “Дополнительные опции узла автоклассификации” на стр. 67.

Опции моделей узла автоклассификации

На вкладке Модель узла Автоклассификация можно задать несколько создаваемых моделей вместе с критериями, используемыми для сравнения моделей.

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Создание моделей расщепления. Строит отдельную модель для каждого возможного значения входных полей, заданных как поля расщепления. Дополнительную информацию смотрите в разделе “Построение моделей расщепления” на стр. 28.

Ранжировать модели, используя. Задает критерии, используемые для сравнения и ранжирования моделей. В число опций входят общая точность, площадь под кривой ROC, прибыль, рост и количество полей. Обратите внимание на то, что все эти измерения будут доступны в сводном отчете независимо от того, что здесь выбрано.

Примечание: Для номинального (набор) поля назначения в качестве ранжирования можно выбрать только **Общую точность** и **Число полей**.

При вычислении прибыли, роста и связанной статистики предполагается, что определенное для поля назначения значение *True* представляет попадание.

- **Общая точность** Процентная доля записей, правильно предсказанных моделью, по отношению к общему числу записей.
- **Площадь под кривой ROC** Кривая ROC предоставляет индекс для производительности модели. Чем выше эта кривая лежит над базовой линией, тем точнее критерий.
- **Прибыль (кумулятивная)** Сумма прибылей по всем кумулятивным процентилям (отсортированным в терминах достоверности для предсказания), вычисленная на основании заданных критериев стоимости, дохода и веса. Обычно прибыль начинается около нуля для верхней процентили, устойчиво увеличивается, а затем уменьшается. Для хорошей модели прибыли покажут хорошо выраженный пик, который входит в отчет вместе с процентилью, в которой он отмечается. Для модели, не дающей информации, кривая прибыли будет относительно ровной, может расти, уменьшаться или идти на одном уровне в зависимости от примененной структуры стоимость/доход.
- **Подъем (кумулятивный)** Коэффициент попадания в кумулятивные квантили по отношению ко всей выборке (здесь квантили сортируются по достоверности предсказаний). Например, значение 3 для подъема в верхней квантили означает, что коэффициент попадания здесь в три раза больше, чем для выборки в целом. Для хорошей модели подъем должен быть заметно больше 1,0 для верхних квантилей, а затем резко падать до значения около 1,0 для нижних квантилей. Для модели, не дающей информации, значение подъема будет повсюду около 1,0.
- **Количество полей** Ранжирует модели на основании количества используемых входных полей.

Ранжировать модели, применив. Если используются разделы, можно задать, как именно ранжируются модели, на основании обучающего набора данных или набора тестирования. Для больших наборов данных использование раздела для предварительного просмотра моделей может существенно повысить производительность.

Число используемых моделей. Задает максимальное количество моделей, которые будут перечислены в слепке моделей, создаваемым узлом. Высшие по рангу модели перечисляются в соответствии с критерием ранжирования. Обратите внимание на то, что увеличение этого ограничивающего значения может понизить производительность. Максимальное разрешенное значение - 100.

Вычислить важность предикторов. Для моделей, производящих соответствующую меру важности, можно вывести диаграмму, показывающую относительную важность каждого предиктора при оценке модели. Обычно вам потребуется направить усилия по моделированию на рассмотрение наиболее важных предикторов, отбрасывая или игнорируя наименее важные. Обратите внимание на то, что важность предикторов может увеличить время обсчета некоторых моделей, поэтому эта характеристика не рекомендуется, если вы просто хотите провести более широкое сравнение среди разных моделей. Важность становится более полезной, когда вы уже сузите свой анализ до небольшого количества моделей, которые нужно будет изучить более подробно. Дополнительную информацию смотрите в разделе “Важность предиктора” на стр. 42.

Критерии прибыли. *Примечание*. Только для флаговых полей назначения. Прибыль равна доходу каждой записи за вычетом ее стоимости. Прибыли для квантили - просто сумма прибылей для всех записей в квантили. Предполагается, что прибыли применяются только к попаданиям, а стоимости - ко всем записям.

- **Стоимости.** Задайте стоимость, связанную с каждой записью. Для стоимости можно выбрать опции **Фиксированная** или **Переменная**. Для фиксированных стоимостей задайте нужные значения. Для переменных стоимостей нажмите кнопку Выбор полей, чтобы выбрать поле стоимостей. (Для диаграмм ROC поле **Стоимости** недоступно.)
- **Прибыль.** Задайте прибыль, связанную с каждой записью, представляющей попадание. Для стоимости можно выбрать опции **Фиксированная** или **Переменная**. Для фиксированных прибылей задайте нужные значения. Для переменных прибылей нажмите кнопку Выбор полей, чтобы выбрать поле прибыли. (Для диаграмм ROC поле **Прибыль** недоступно.)
- **Вес.** Если записи в ваших данных представляют несколько блоков, можно использовать частотные веса, чтобы скорректировать результаты. Задайте вес, связанный с каждой записью, используя **Фиксированные** или **Переменные** веса. Для фиксированных весов задайте значение веса (количество блоков на запись). Для переменных весов нажмите кнопку Выбор полей, чтобы выбрать поле весов. (Для диаграмм ROC поле **Вес** недоступно.)

Критерии подъема. *Примечание*. Только для флаговых полей назначения. Задает процентиль, которая будет использоваться для вычислений подъема. Обратите внимание на то, что это значение можно изменить и при сравнении результатов. Дополнительную информацию смотрите в разделе “Слепки автоматизированных моделей” на стр. 77.

Дополнительные опции узла автоклассификации

На вкладке Дополнительно узла автоклассификации можно применить раздел (если это доступно), выбрать используемый алгоритм и задать правила остановки.

Выберите модели. По умолчанию для построения выбраны все модели; но если у вас есть Analytic Server (сервер аналитических служб), вы можете ограничиться теми моделями, которые можете запустить в Analytic Server (сервер аналитических служб) и задать для них либо построение моделей расщепления, либо готовность к обработке больших объемов данных.

Используемые модели. Используйте переключатели в столбце слева, чтобы выбрать типы моделей (алгоритмы) для включения в сравнение. Чем больше типов вы выберете, тем больше моделей будет создано и тем больше времени займет их обработка.

Тип модели. Перечисляет доступные алгоритмы (смотрите ниже).

Параметры модели. Для каждого типа модели можно использовать значения по умолчанию или нажать кнопку **Задать**, чтобы выбрать опции для каждого типа моделей. Конкретные опции аналогичны доступным на других узлах моделирования, однако здесь можно выбрать несколько опций или их комбинаций. Например, при использовании моделей нейросетей вместо выбора одного из шести способов обучения можно указать все из них, чтобы при одном проходе обучить шесть моделей.

Количество моделей. Содержит количество моделей, созданных для каждого алгоритма на основе текущих параметров. При комбинировании опций количество моделей может быстро возрастать, поэтому настоятельно рекомендуется особенно внимательно следить за этим значением, в частности, при использовании больших наборов данных.

Ограничить максимальное время, уходящее на построение одной модели. (Только для моделей K-средних, Коонена, двухшаговой оценки, SVM, KNN, байесовской сети и списка решений). Задает предельное максимальное время на каждую модель. Например, если некоторой конкретной модели потребуется неожиданно большое время на обучение из-за какого-то сложного взаимодействия, задержка выполнения полного прохода моделирования может быть нежелательной.

Примечание: Если поле назначения номинальное (набор), опция Список решений недоступна.

Поддерживаемые алгоритмы

Узел нейросетей использует упрощенную модель обработки информации человеческим мозгом. Нейросети работают, обсчитывая большое количество связанных между собой обрабатываемых элементов, которые представляют абстрактную версию нейронов. Нейросети - это мощные средства оценки общих функциональных зависимостей, требующие минимальных знаний статистики и математики для их обучения и применения.

Узел C5.0 строит или дерево решений, или набор правил. Эта модель работает, разделяя выборку на основании значения в поле, дающего максимальный информационный выигрыш на каждом уровне. Поле назначения должно быть категориальным. Разрешено несколько разделений на подгруппы, и таких подгрупп может быть больше двух.

Узел дерева классификации и регрессии (Classification and Regression, C&R) генерирует дерево решений, позволяющее предсказывать или классифицировать будущие наблюдения. Этот метод использует рекурсивное разделение, чтобы расщепить обучающие записи на сегменты, на каждом шаге минимизируя неоднородность, причем узел дерева считается “чистым”, если все 100% наблюдений в узле попадают в конкретную категорию поля назначения. Входные поля и поля назначения могут быть из числового диапазона или категориальными (номинальными, порядковыми или флагами); все расщепления бинарны (только две подгруппы).

Узел QUEST предоставляет метод бинарной классификации для построения деревьев решений, разработанный для уменьшения времени обработки, требуемого для анализа больших деревьев C&R, при одновременном подавлении обнаруженного в способах деревьев классификации предпочтения входных полей, допускающих больше расщеплений. Входные поля могут быть в числовом диапазоне (количественными), но поле назначения должно быть категориальным. Все расщепления бинарные.

Узел CHAID генерирует деревья решений, используя статистику хи-квадрат для определения оптимальных расщеплений. В отличие от узлов дерева C&R и QUEST, CHAID может генерировать не только бинарные деревья, то есть у некоторых расщеплений может быть больше двух ветвей. Входные поля и поле назначения могут быть количественными (числовой диапазон) или категориальными. Исчерпывающий CHAID - это модификация метода CHAID, при котором проделывается более тщательная работа по изучению всех возможных расщеплений для каждого предиктора, но это требует больше времени для вычислений.

Логистическая регрессия - это статистический метод для классификации записей на основании значений входных полей. Она аналогична линейной регрессии, но логистическая регрессия использует категориальные поля назначения вместо численных.

Узел списка решений определяет подгруппы или сегменты, которые показывают более высокое или более низкое правдоподобие для данного бинарного результата по сравнению с полной совокупностью. Например, вы могли бы искать клиентов с низкой вероятностью оттока или с высокой вероятностью отклика на кампанию. Вы можете включить свои знания о бизнесе в модель, добавляя свои собственные пользовательские сегменты и параллельно просматривая альтернативные модели, чтобы сравнить результаты. Модели списка решений состоят из списка правил, в котором каждое правило имеет условие и следствие. Правила применяются по очереди, и первое подходящее правило определяет результат.

Узел Байесовская сеть позволяет построить вероятностную модель, комбинируя наблюдаемые и записанные сведения с очевидными с точки зрения здравого смысла данными, чтобы установить правдоподобие возникновения событий. Этот узел в основном работает с усиленными деревом наивными байесовскими сетями (Tree Augmented Naïve Bayes, TAN) и полными марковскими сетями, которые изначально используются для классификации.

Дискриминантный анализ делает более строгие предположения, чем логистическая регрессия, но он может быть ценной альтернативой или дополнением к анализу логистической регрессии, когда эти предположения оказываются правильными.

Узел *k* ближайших соседей (k-Nearest Neighbor, KNN) связывает новое наблюдение с категорией или значением *k* объектов, ближайших к нему в пространстве предикторов, где *k* - это целое число. Подобные наблюдения близки друг к другу, а непохожие наблюдения, наоборот, удалены друг от друга.

Узел механизма опорных векторов (Support Vector Machine, SVM) позволяет классифицировать данные по одной или двум группам без переобучения. SVM хорошо работает с широкими наборами данных, в частности, в случае очень большого числа входных полей.

Стоимости ошибочной классификации

В некоторых контекстах определенные виды ошибок обходятся пользователю дороже других. Например, может оказаться более дорогостоящим классифицировать претендента на кредит с высоким уровнем риска, как с низким уровнем риска (один вид ошибки), чем классифицировать претендента на кредит с низким уровнем риска риском как с высокими уровнем риска (другой вид ошибки). Стоимости ошибочной классификации позволяют задать относительную важность различных видов ошибок предсказания.

Стоимости ошибочной классификации - это по существу веса, применяемые к конкретным исходам. Эти веса факторизуются в модель и могут фактически изменить предсказание (в качестве способа защиты от дорогостоящих ошибок).

За исключением моделей C5.0, стоимости ошибочной классификации при скоринге моделей не применяются, и при ранжировании или сравнении моделей во внимание не принимаются. Модель, включающая в себя стоимости, не может дать меньше ошибок, чем та, которая не ранжируется и не может ранжироваться хоть сколь-нибудь выше в единицах общей точности, но, скорее всего, она будет выполняться лучше на практике, поскольку в ней заложено предусмотренное смещение в пользу *менее дорогостоящих* ошибок.

Матрица стоимостей показывает стоимость для каждого возможного сочетания предсказанной и действительной категорий. По умолчанию для всех стоимостей ошибочной классификации задается значение 1,0. Чтобы ввести пользовательские значения стоимостей, выберите **Использовать стоимости ошибочной классификации** и введите в матрицу стоимостей нужные вам значения.

Чтобы изменить стоимость ошибочной классификации, выберите ячейку, соответствующую нужному сочетанию предсказанного и действительного значений, удалите существующее содержание ячейки и введите для нее желаемую стоимость. Стоимости не являются автоматически симметричными. Например, если для стоимости ошибочной классификации *A* как *B* задать значение 2,0, у стоимости ошибочной классификации *B* как *A* все равно будет значение по умолчанию 1,0, пока вы не измените также и его явным образом.

Опции отклонения узла автоклассификации

На вкладке Отклонение узла автоклассификации можно автоматически отклонить модели, которые не соответствуют определенным критериям. Эти модели не будут перечисляться в сводном отчете.

Можно задать минимальный порог для общей точности и максимальный порог для числа используемых в модели переменных. Кроме этого, для флаговых полей назначения можно задать минимальный порог для подъема, прибыли и площади под кривой; подъем и прибыль определяются, как задано на вкладке Модель. Дополнительную информацию смотрите в разделе “Опции моделей узла автоклассификации” на стр. 65.

Дополнительно можно сконфигурировать узел для остановки выполнения после генерирования первой модели, которая удовлетворяет заданным критериям. Дополнительную информацию смотрите в разделе “Правила остановки для узла автоматического моделирования” на стр. 64.

Опции параметров узла автоклассификации

На вкладке Параметры узла автоклассификации можно предварительно сконфигурировать опции времени оценки, доступные для слепка.

Метод ансамбля. Для полей назначения, которые можно выбрать из следующих способов ансамбля:

- Голосование
- Голосование со взвешенными доверительными вероятностями
- Простое голосование со взвешенными склонностями (только для флаговых полей назначения)
- Интервалы с наибольшей доверительной вероятностью
- Усреднение простой склонности (только флаговые поля назначения)

Если голосование привязано, выбрать значение, применив. Для способов голосования можно задать, как разрешаются связи:

- **Произвольный выбор.** Одно из связанных значений выбирается случайным образом.
- **Наибольшая доверительная вероятность.** Побеждает связанное значение, предсказанное с большей доверительной вероятностью. Обратите внимание на то, что это не обязательно наивысшая доверительная вероятность среди всех предсказанных значений.
- **Простая склонность.** (только для флаговых полей назначения) Связанное значение, предсказанное с наибольшей абсолютной склонностью, где абсолютная склонность определяется следующим образом:

```
abs(0.5 - propensity) * 2
```

Узел автонумерации

Узел автонумерации оценивает и сравнивает модели для количественных полей назначения числового диапазона, используя несколько различных способов, что позволяет испытать разнообразные подходы при одном запуске моделирования. Вы можете выбрать используемые алгоритмы и экспериментировать с несколькими сочетаниями опций. Например, можно предсказать стоимость жилой недвижимости при помощи моделей нейронной сети, линейной регрессии, C&RT и CHAID, чтобы увидеть, какая работает лучше, а также попробовать различные комбинации способов пошаговой, прямой и обратной регрессии. Этот узел исследует все возможные сочетания опций, ранжирует модели-кандидаты на основании заданного вами показателя и сохраняет лучшие модели для использования при скоринге или для дальнейшего анализа. Дополнительную информацию смотрите в разделе Глава 5, “Узлы автоматического моделировани”, на стр. 63.

Пример. Муниципалитет хочет более точно оценивать реальные налоги на недвижимость и нужным образом корректировать значения по конкретным объектам недвижимости без необходимости исследовать эти объекты. Используя узел автонумерации, аналитик может сгенерировать и сравнить несколько моделей, предсказывающих стоимость недвижимости на основании типа строения, соседства, размеров и других известных факторов.

Требования. У вас должно быть только одно поле назначения (с заданной ролью **Назначение**) и один или несколько предикторов (в заданной ролью **Ввод**). Поле назначения должно быть количественным (числовой диапазон), например, *возраст* или *доход*. У входных полей может быть уровень измерения Количественный или Категориальный с тем ограничением, что некоторые входные данные могут не подходить для некоторых типов моделей. Например, модели дерева C&R могут использовать категориальные строковые поля в качестве входных, а для моделей линейной регрессии их использовать нельзя, и такие поля будут игнорироваться. Это те же требования, что и при использовании отдельных узлов моделирования. Например, модель CHAID работает одинаково при генерировании и из узла CHAID, и из узла Автонумерации.

Поля частоты и веса. Частота и вес используются для предоставления дополнительной важности некоторым записям по сравнению с остальными, так как, например, пользователь знает, что набор данных сборки недостаточно представляет раздел заполнения родительских элементов (вес) или одна запись представляет несколько идентичных наблюдений (частота). Если задано поле частоты, оно может использоваться алгоритмами дерева C&R и CHAID. Поле веса может использоваться алгоритмами C&RT, CHAID, регрессии и обобщенной линейной регрессии. Другие типы моделей будут игнорировать эти поля, но в любом случае построят свои модели. Поля частоты и веса используются только для построения моделей и не рассматриваются при проверке или скоринге моделей. Дополнительную информацию смотрите в разделе “Использование полей частоты и веса” на стр. 32.

Поддерживаемые типы моделей

К поддерживаемым типам моделей относятся нейросети, дерево C&R, CHAID, регрессия, обобщенная линейная регрессия, ближайшие соседи и SVM. Дополнительную информацию смотрите в разделе “Опции эксперта узла автонумерации” на стр. 72.

Опции моделей узла автонумерации

На вкладке Модель узла автонумерации можно задать число сохраняемых моделей и критерии для сравнения моделей.

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Создание моделей расщепления. Строит отдельную модель для каждого возможного значения входных полей, заданных как поля расщепления. Дополнительную информацию смотрите в разделе “Построение моделей расщепления” на стр. 28.

Ранжировать модели, используя. Задает критерии, используемые для сравнения моделей.

- **Корреляция.** Корреляции Пирсона между наблюдаемым и предсказанным моделью значением для каждой записи. Коэффициент корреляции - это мера линейной связи между двумя переменными, при этом более близкие к единице значения соответствуют более сильной корреляции. (Значения коэффициента корреляции изменяются в диапазоне от –1 при абсолютной отрицательной взаимосвязи до +1 при абсолютной положительной взаимосвязи). Значение 0 указывает на отсутствие линейной связи, в то время как модель с отрицательным коэффициентом корреляции будет ранжироваться на самом низком уровне).
- **Число полей.** Количество полей, используемых в модели в качестве предикторов. Выбор моделей с меньшим количеством полей может упростить подготовку данных и в некоторых случаях повысить производительность.
- **Относительная ошибка.** Относительная ошибка - это отношение дисперсии наблюдаемых значений при отклонении от предсказанных моделью к дисперсии наблюдаемых значений при отклонении от среднего. С практической точки зрения относительная ошибка показывает, насколько хорошо работает модель по сравнению с моделью **null** или **постоянный член**, когда в качестве предсказания возвращается просто среднее значение поля назначения. Для хорошей модели это значение должно быть меньше 1, обозначая лучшую точность модели по сравнению с пустой моделью. Модель с относительной ошибкой больше 1 менее точная, чем пустая модель, поэтому бесполезна. Для моделей линейной регрессии относительная ошибка равна квадрату коэффициента корреляции и не добавляет какой-либо информации. Для нелинейных моделей относительная ошибка не связана с коэффициентом корреляции и предоставляет дополнительную меру для оценки работы модели.

Ранжировать модели, применив. Если используются разделы, можно задать, как именно ранжируются модели, на основании обучающего раздела данных или испытательного раздела. Для больших наборов данных использование раздела для предварительного просмотра моделей может существенно повысить производительность.

Число используемых моделей. Задает максимальное количество моделей, которые будут перечислены в слепке моделей, создаваемым узлом. Высшие по рангу модели перечисляются в соответствии с критерием ранжирования. Увеличение этого предела позволит сравнивать результаты большего числа моделей, но может понизить производительность. Максимальное разрешенное значение - 100.

Вычислить важность предикторов. Для моделей, производящих соответствующую меру важности, можно вывести диаграмму, показывающую относительную важность каждого предиктора при оценке модели. Обычно вам потребуется направить усилия по моделированию на рассмотрение наиболее важных предикторов, отбрасывая или игнорируя наименее важные. Обратите внимание на то, что важность предикторов может увеличить время обсчета некоторых моделей, поэтому эта характеристика не рекомендуется, если вы просто хотите провести более широкое сравнение среди разных моделей. Важность становится более полезной, когда вы уже сузите свой анализ до небольшого количества моделей, которые нужно будет изучить более подробно. Дополнительную информацию смотрите в разделе “Важность предиктора” на стр. 42.

Не хранить модели, если. Задает пороговые значения для коэффициента корреляции, относительной ошибки и количества используемых полей. Модели, которым не удается удовлетворить любому из этих критериев, будут отброшены и не будут перечисляться в сводном отчете.

- **Коэффициент корреляции меньше чем.** Минимальная корреляции (по абсолютному значению) для модели, чтобы она включалась в сводный отчет.
- **Число используемых полей больше чем.** Максимальное число полей, которые можно использовать в любой модели, чтобы она включалась в сводный отчет.
- **Относительная ошибка больше чем.** Максимальная относительная ошибка для любой модели, чтобы она включалась в сводный отчет.

Дополнительно можно сконфигурировать узел для остановки выполнения после генерирования первой модели, которая удовлетворяет заданным критериям. Дополнительную информацию смотрите в разделе “Правила остановки для узла автоматического моделирования” на стр. 64.

Опции эксперта узла автонумерации

На вкладке Эксперт узла Автонумерация можно выбрать используемые алгоритмы и опции и задать правила остановки.

Выберите модели. По умолчанию для построения выбраны все модели; но если у вас есть Analytic Server (сервер аналитических служб), вы можете ограничиться теми моделями, которые можете запустить в Analytic Server (сервер аналитических служб) и задать для них либо построение моделей расщепления, либо готовность к обработке больших объемов данных.

Используемые модели. Используйте переключатели в столбце слева, чтобы выбрать типы моделей (алгоритмы) для включения в сравнение. Чем больше типов вы выберете, тем больше моделей будет создано и тем больше времени займет их обработка.

Тип модели. Перечисляет доступные алгоритмы (смотрите ниже).

Параметры модели. Для каждого типа модели можно использовать значения по умолчанию или нажать кнопку **Задать**, чтобы выбрать опции для каждого типа моделей. Конкретные опции аналогичны доступным на других узлах моделирования, однако здесь можно выбрать несколько опций или их комбинаций. Например, при использовании моделей нейросетей вместо выбора одного из шести способов обучения можно указать все из них, чтобы при одном проходе обучить шесть моделей.

Количество моделей. Содержит количество моделей, созданных для каждого алгоритма на основе текущих параметров. При комбинировании опций количество моделей может быстро возрастать, поэтому настоятельно рекомендуется особенно внимательно следить за этим значением, в частности, при использовании больших наборов данных.

Ограничить максимальное время, уходящее на построение одной модели. (Только для моделей K-средних, Коонена, двухшаговой оценки, SVM, KNN, байесовской сети и списка решений). Задает предельное максимальное время на каждую модель. Например, если некоторой конкретной модели потребуется неожиданно большое время на обучение из-за какого-то сложного взаимодействия, задержка выполнения полного прохода моделирования может быть нежелательной.

Поддерживаемые алгоритмы

Узел нейросетей использует упрощенную модель обработки информации человеческим мозгом. Нейросети работают, обсчитывая большое количество связанных между собой обрабатываемых элементов, которые представляют абстрактную версию нейронов. Нейросети - это мощные средства оценки общих функциональных зависимостей, требующие минимальных знаний статистики и математики для их обучения и применения.

Узел дерева классификации и регрессии (Classification and Regression, C&R) генерирует дерево решений, позволяющее предсказывать или классифицировать будущие наблюдения. Этот метод использует рекурсивное разделение, чтобы расщепить обучающие записи на сегменты, на каждом шаге минимизируя неоднородность, причем узел дерева считается “чистым”, если все 100% наблюдений в узле попадают в конкретную категорию поля назначения. Входные поля и поля назначения могут быть из числового диапазона или категориальными (номинальными, порядковыми или флагами); все расщепления бинарны (только две подгруппы).

Узел CHAID генерирует деревья решений, используя статистику хи-квадрат для определения оптимальных расщеплений. В отличие от узлов дерева C&R и QUEST, CHAID может генерировать не только бинарные деревья, то есть у некоторых расщеплений может быть больше двух ветвей. Входные поля и поле назначения могут быть количественными (числовой диапазон) или категориальными. Исчерпывающий CHAID - это модификация метода CHAID, при котором проделывается более тщательная работа по изучению всех возможных расщеплений для каждого предиктора, но это требует больше времени для вычислений.

Линейная регрессия - это общепринятый статистический метод обработки данных и вычисления предсказаний при подгонке прямой линии или плоскости, минимизирующих разности между предсказанными и фактическими выходными значениями.

Обобщенная линейная модель расширяет общую линейную модель, так что зависимая переменная считается линейно связанной с факторами и ковариатами через заданную функцию связи. Более того, модель допускает наличие у зависимой переменной распределения, отличающегося от нормального. Она включает в себя функциональные возможности большого количества статистических моделей, в том числе линейной регрессии, логистической регрессии, логлинейных моделей для количества данных и интервал-цензурированных моделей выживания.

Узел *k* ближайших соседей (k-Nearest Neighbor, KNN) связывает новое наблюдение с категорией или значением *k* объектов, ближайших к нему в пространстве предикторов, где *k* - это целое число. Подобные наблюдения близки друг к другу, а непохожие наблюдения, наоборот, удалены друг от друга.

Узел механизма опорных векторов (Support Vector Machine, SVM) позволяет классифицировать данные по одной или двум группам без переобучения. SVM хорошо работает с широкими наборами данных, в частности, в случае очень большого числа входных полей.

Модели линейной регрессии предсказывают значения непрерывного целевого поля на основе линейных взаимосвязей между целевым полем и одним или несколькими предикторами.

Опции параметров узла автонумерации

На вкладке Параметры узла автонумерации можно предварительно сконфигурировать опции времени оценки, доступные для слепка.

Вычислить среднеквадратичную ошибку. Если поле назначения количественное (числовой диапазон), по умолчанию запускается вычисление среднеквадратичной ошибки для определения различий между измеренными или оцененными значениями и действительными значениями, а также для демонстрации, насколько эти оценки совпали.

Узел Автокластеризация

Узел автокластеризации сравнивает и оценивает модели кластеризации, которые выявляют группы записей со сходными характеристиками. Подобно другим узлам автоматического моделирования, этот узел можно использовать для опробования нескольких сочетаний опций за один запуск моделирования. Модели можно сравнивать при помощи базовых показателей, пытаясь фильтровать и ранжировать с их использованием полезность моделей кластеризации и предоставить показатель на основе важности конкретных полей.

Модели кластеризации часто служат для обнаружения групп, которые можно использовать как входную информацию для дальнейшего анализа. Например, пусть нужно выявить целевые группы клиентов с учетом демографических характеристик, таких как доход, или с учетом услуг, которые они приобретали в прошлом. Это можно сделать, не зная заранее о группах и их характеристиках -- можно не знать, сколько групп желательно выявить, и по каким особенностям. Модели кластеризации часто называют моделями обучения без учителя, поскольку в них не используется поле назначения и они не возвращают конкретное предсказание, которое можно оценить как истинное или ложное. О ценности модели кластеризации судят по ее способности выявлять интересные группы в данных и предлагать полезные описания этих групп. Дополнительную информацию смотрите в разделе Глава 11, “Модели кластеризации”, на стр. 209.

Требования. Одно или несколько полей характеристик, представляющих интерес. В отличии от других моделей, в моделях кластеризации нет полей назначения, поскольку не делается конкретных предсказаний, которые можно оценивать как истинные или ложные. Вместо этого здесь ищутся группы наблюдений, которые могут быть связаны друг с другом. Так, в модели кластеризации нельзя предсказать, как данный клиент отнесется к некоторому предложению. Но при помощи модели кластеризации можно назначить клиентов в группы на основе склонности вести себя тем или иным образом. Поля веса и частоты не используются.

Поля нормирования. Хотя поле назначения не используется, вы все же можете указать одно или несколько полей оценки при сравнении моделей. Полезность модели кластеризации может оцениваться по тому, насколько хорошо или плохо кластеры различаются по этим полям.

Поддерживаемые типы моделей

Поддерживаемые типы моделей включают в себя двухэтапную модель, модель k-средних и модель Коонена.

Опции модели узла автоматической кластеризации

На вкладке Модель узла автокластеризации можно задать число сохраняемых моделей и критерии для сравнения моделей.

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Ранжировать модели, используя. Задает критерии, используемые для сравнения и ранжирования моделей.

- **Силуэтная мера.** Показатель одновременно сцепленности внутри кластеров и разделенности между кластерами. Дополнительную информацию смотрите ниже в разделе *Силуэтная мера ранжирования*.
- **Число кластеров.** Число кластеров в модели.
- **Размер наименьшего кластера.** Размер наименьшего кластера.
- **Размер наибольшего кластера.** Размер наибольшего кластера.
- **Наименьший кластер к наибольшему.** Отношение размера наименьшего кластера к размеру наибольшего кластера.
- **Важность.** Важность поля **Оценка** на вкладке **Поля**. Обратите внимание на то, что это можно вычислить, только если задано поле **Оценка**.

Ранжировать модели, применив. Если используются разделы, можно задать, как именно ранжируются модели, на основании обучающего набора данных или набора тестирования. Для больших наборов данных использование раздела для предварительного просмотра моделей может существенно повысить производительность.

Число сохраняемых моделей. Задает максимальное количество моделей в слепке, созданном этим узлом. Высшие по рангу модели перечисляются в соответствии с критерием ранжирования. Обратите внимание на то, что увеличение этого ограничивающего значения может понизить производительность. Максимальное разрешенное значение - 100.

Силуэтная мера ранжирования

Мера ранжирования по умолчанию, Силуэт, имеет значение по умолчанию 0, поскольку при меньших значения (т.е. отрицательных) среднее расстояние между наблюдением и другими точками того же кластера больше, чем минимальное среднее расстояние до точек в других кластерах. Таким образом, модели с отрицательной силуэтной мерой заведомо должны быть отброшены.

По сути мера ранжирования - это модифицированный силуэтный коэффициент, сочетающий в себе понятия сцепленности внутри кластера (согласно которому предпочтение отдается моделям, содержащим тесно сцепленные кластеры) и разделенности между кластерами (согласно которому предпочтение отдается моделям, содержащим в высокой степени обособленные кластеры). Средний силуэтный коэффициент - это просто среднее по всем наблюдениям для следующей формулы, вычисляемой для каждого наблюдения:

$(B - A) / \max(A, B)$

где A - расстояние от наблюдения до центроида того кластера, к которому отнесено данное наблюдение, а B - минимальное расстояние от этого наблюдения до центроидов остальных кластеров.

Силуэтный коэффициент, как и его среднее, может быть от -1 (показатель очень плохой модели) до 1 (показатель очень хорошей модели). Усреднение может выполняться как на уровне генеральной совокупности наблюдений (что дает генеральную силуэтную меру), так и на уровне кластеров (что дает силуэтную меру кластера). Для вычисления расстояний могут использоваться евклидовы расстояния.

Auto Cluster Node Expert Options

На вкладке Эксперт узла Автокластеризация можно применить разбиение (если оно доступно), выбрать алгоритм кластеризации и задать правила остановки.

Используемые модели. Используйте переключатели в столбце слева, чтобы выбрать типы моделей (алгоритмы) для включения в сравнение. Чем больше типов вы выберете, тем больше моделей будет создано и тем больше времени займет их обработка.

Тип модели. Перечисляет доступные алгоритмы (смотрите ниже).

Параметры модели. Для каждого типа модели можно использовать значения по умолчанию или нажать кнопку **Задать**, чтобы выбрать опции для каждого типа моделей. Конкретные опции аналогичны доступным на других узлах моделирования, однако здесь можно выбрать несколько опций или их комбинаций. Например, при использовании моделей нейросетей вместо выбора одного из шести способов обучения можно указать все из них, чтобы при одном проходе обучить шесть моделей.

Количество моделей. Содержит количество моделей, созданных для каждого алгоритма на основе текущих параметров. При комбинировании опций количество моделей может быстро возрастать, поэтому настоятельно рекомендуется особенно внимательно следить за этим значением, в частности, при использовании больших наборов данных.

Ограничить максимальное время, уходящее на построение одной модели. (Только для моделей K-средних, Коонена, двухшаговой оценки, SVM, KNN, байесовской сети и списка решений). Задает предельное максимальное время на каждую модель. Например, если некоторой конкретной модели потребуется неожиданно большое время на обучение из-за какого-то сложного взаимодействия, задержка выполнения полного прохода моделирования может быть нежелательной.

Поддерживаемые алгоритмы

Узел K-средних кластеризует набор данных в отдельные группы (или кластеры). Этот метод определяет фиксированное количество кластеров, итерационно распределяет записи по кластерам и настраивает центры кластеров, пока дальнейшие уточнения более не улучшают модель. Вместо попытки предсказать выходное значение *k*-средние используют процесс, называемый неконтролируемым обучением, чтобы обнаружить структуры в наборе входных полей.

Узел Коонена генерирует тип нейросети, которую можно использовать для кластеризации набора данных в отдельные группы. Когда сеть полностью обучена, похожие записи должны быть близко друг от друга на выходной карте, а отличающиеся записи должны быть сильно разделены. По количеству наблюдений, захваченных каждым нейроном в слепке модели, можно определить сильные нейроны. Это может дать представление об оправданном количестве кластеров.

Узел Двухшаговый использует метод двухшаговой кластеризации. На первом шаге проводится первый проход по данным, при котором необработанные входные данные сжимаются в управляемый набор подкластеров. На втором шаге используется способ иерархической кластеризации для все большего слияния подкластеров в крупные и еще более крупные кластеры. У двухшагового метода есть преимущество автоматической оценки оптимального числа кластеров для обучающих данных. Он может эффективно обрабатывать поля смешанных типов и большие наборы данных.

Опции отбрасывания узла автокластеризации

На вкладке Отбрасывание узла Автокластеризация можно задать автоматическое отбрасывание моделей, не отвечающих определенным критериям. Эти модели не войдут в список слепка модели.

Можно указать минимальное значение силуэтной меры, число кластеров, размеры кластеров и важность поля оценки, используемого в модели. Силуэтная мера, число кластеров и размер кластеров определяются тем, что задано в узле моделирования. Дополнительную информацию смотрите в разделе “Опции модели узла автоматической кластеризации” на стр. 74.

Дополнительно можно сконфигурировать узел для остановки выполнения после генерирования первой модели, которая удовлетворяет заданным критериям. Дополнительную информацию смотрите в разделе “Правила остановки для узла автоматического моделирования” на стр. 64.

Слепки автоматизированных моделей

Когда выполняется узел автоматического моделирования, этот узел оценивает модели-кандидаты по всем возможным комбинациям опций, ранжирует все модели-кандидаты на основании заданных вами показателей и сохраняет лучшие модели в слепке составной автоматизированной модели. Этот слепок модели на самом деле содержит набор моделей (одну или несколько), сгенерированных узлом, которые можно просматривать или выбирать для скоринга индивидуально. Для каждой модели перечисляется ее тип и время построения, а также несколько других показателей, соответствующих типу модели. Таблицу можно отсортировать по любым из этих столбцов для быстрого поиска самых интересных моделей.

- Для просмотра любого из слепков индивидуальных моделей дважды щелкните по значку слепка. Отсюда можно сгенерировать узел моделирования для этой модели на холст потока или скопировать слепок модели на палитру моделей.
- Миниизображения графиков дают быструю визуальную оценку для каждого типа модели, что подробно описано ниже. Чтобы сгенерировать полноразмерный график, дважды щелкните по миниизображению. Полноразмерный график выводит до тысячи точек, а если в наборе данных точек больше, будет использоваться выборка. (Только для диаграмм рассеяния. График генерируется повторно при каждом выводе, поэтому любые изменения в данных для пользователей, использующих результаты программы, такие как изменение случайной выборки или раздела, всякий раз будет выглядеть как перерисовка диаграммы рассеяния, если не выбрана опция **Задать начальное значение генератора псевдослучайных чисел**).
- Используйте панель инструментов, чтобы показать или скрыть конкретные столбцы на вкладке Модель или изменить столбец, используемый для сортировки таблицы. Сортировку можно изменить также, щелкнув по заголовкам столбцов).
- Используйте кнопку Удалить для окончательного удаления всех неиспользуемых моделей.
- Чтобы изменить порядок столбцов, щелкните по заголовку столбца и перетащите его в нужное положение.
- Если используются разделы, можно выбрать просмотр результатов для раздела обучения или раздела испытаний при их применимости.

Конкретные столбцы зависят от типа сравниваемых моделей, что подробно описано ниже.

Бинарные поля назначения

- Для бинарных моделей миниизображение графика показывает распределение фактических значений с наложенными предсказанными значениями, чтобы дать быстрое визуальное обозначение, сколько записей было правильно предсказано в каждой категории.
- Критерии ранжирования совпадают с опциями на узле моделирования автоклассификации. Дополнительную информацию смотрите в разделе “Опции моделей узла автоклассификации” на стр. 65.
- Для определения максимального дохода сообщается также о процентиле, в которой достигается максимум.
- Для интегрированного подъема можно с помощью панели инструментов изменить выбранную процентиль.

Номинальные поля назначения

- Для номинальных (набор) моделей миниизображение графика показывает распределение фактических значений с наложенными предсказанными значениями, чтобы дать быстрое визуальное обозначение, сколько записей было правильно предсказано в каждой категории.
- Критерии ранжирования совпадают с опциями на узле моделирования автоклассификации. Дополнительную информацию смотрите в разделе “Опции моделей узла автоклассификации” на стр. 65.

Количественные поля назначения

- Для количественных (численный диапазон) моделей график выводит предсказанные значения в зависимости от наблюдаемых для каждой модели, чтобы дать быстрое визуальное обозначение

корреляций между ними. Для хорошей модели у точек графика должна быть тенденция кластеризации вдоль диагонали, а не произвольного разброса по графику.

- Критерии ранжирования совпадают с опциями на узле моделирования автонумерации. Дополнительную информацию смотрите в разделе “Опции моделей узла автонумерации” на стр. 71.

Поля назначения кластера

- Для кластерных моделей на графике выводится количество точек в зависимости от кластера для каждой модели, чтобы дать быстрое визуальное обозначение распределения кластеров.
- Критерии ранжирования совпадают с опциями на узле моделирования автокластеризации. Дополнительную информацию смотрите в разделе “Опции модели узла автоматической кластеризации” на стр. 74.

Выбор моделей для скоринга

Столбец **Использовать?** позволяет вам выбрать модели, которые будут использоваться при скоринге.

- Только для бинарных, номинальных и числовых полей назначения вы можете выбрать несколько моделей скоринга и комбинировать оценки в слепке одной комбинированной модели. При комбинировании предсказаний от нескольких моделей можно избежать ограничений, накладываемых на отдельную модель, и часто получить в результате более высокую общую точность, чем можно бы было достигнуть в любой единичной модели.
- Для кластерных моделей одновременно можно выбрать только одну модель скоринга. По умолчанию первой выбирается самая высоко ранжированная модель.

Генерирование узлов и моделей

Вы можете сгенерировать копию слепка составной автоматизированной модели или узел автоматического моделирования, на котором она была создана. Это может быть полезно, например, если у вас нет исходного потока, в котором был построен слепок автоматизированной модели. Как вариант, можно сгенерировать слепок или узел моделирования для любой из индивидуальных моделей, перечисленных в слепке автоматизированной модели.

Слепок автоматизированного моделирования

В меню Создать выберите пункт **Модель на палитру**, чтобы добавить слепок автоматизированной модели на палитру Модели. Сгенерированную модель можно сохранить или использовать, как есть, без возвращения в поток.

Как вариант, можно выбрать в меню Создать пункт **Генерировать узел моделирования**, чтобы добавить узел моделирования на холст потока. Этот узел можно использовать для повторной оценки выбранных моделей без повторения запуска всего цикла моделирования.

Слепок индивидуального моделирования

1. В меню **Модель** дважды щелкните по нужному индивидуальному слепку. Копия этого слепка открывается в новом диалоговом окне.
2. В меню Создать нового диалогового окна выберите пункт **Модель на палитру**, чтобы добавить слепок автоматизированной модели на палитру Модели.
3. Как вариант, можно выбрать в меню Создать нового диалогового окна пункт **Генерировать узел моделирования**, чтобы добавить узел моделирования на холст потока.

Генерирование диаграмм оценки

Только для бинарных моделей. Можно сгенерировать диаграммы оценки, обеспечивающие возможность оценки и сравнения производительностей всех моделей. Диаграммы оценки недоступны для моделей, сгенерированных на узлах автонумерации или автокластеризации.

1. В столбце *Использовать?* слепка автоматизированной модели автоклассификации выберите модели, которые вы хотите оценить.
2. В меню Создать выберите пункт **Диаграммы оценки**. Появится диалоговое окно Диаграмма оценки.
3. Выберите тип диаграммы и другие нужные опции.

Графики оценки

На вкладке Модель слепка автоматизированной модели можно детализировать информацию для вывода индивидуальных графиков для каждой из показанных моделей. Для слепков автоклассификации и автонумерации на вкладке График выводится и график, и важность предикторов, отображающие результаты всех комбинированных моделей. Дополнительную информацию смотрите в разделе “Важность предиктора” на стр. 42.

Для автоклассификации показывается график распределения, а для автонумерации - мультиграфик (также известный как диаграмма рассеяния).

Глава 6. Деревья решений

Модели деревьев решений

Модели деревьев решений позволяют создать системы классификации, которые предсказывают или классифицируют будущие наблюдения на основе набора решающих правил. Если данные разделяются на интересующие вас классы (например, ссуды высокого и низкого риска, подписчики и не-подписчики, голосующие и неголосующие или типы бактерий), можно использовать существующие данные для построения правил, которые можно применять для классификации старых и новых наблюдений с максимальной точностью. Например, можно построить дерево, классифицирующее кредитные риски или намерение покупки на основании возраста и других факторов.

Такой подход, который иногда называют **выводом правила**, обладает несколькими преимуществами. Во-первых, процесс рассуждения, стоящий в основе модели, виден вам при просмотре дерева. Это важное отличие от способов типа "чёрного ящика", в которой внутреннюю логику бывает трудно понять.

Во-вторых, процесс автоматически включает в свои правила только те атрибуты, которые на самом деле имеют значение при принятии решения. Атрибуты, которые не вносят вклада в точность дерева, игнорируются. Это может дать очень полезную информацию о данных и позволяет сократить данные до релевантных полей перед тренировкой другого метода обучения, такого, как нейросеть.

Слепки модели дерева решений можно преобразовать в собрание правил IF-THEN (**набор правил**), которое во многих случаях дает информацию в более понятной форме. Презентация дерева решений особенно полезна, когда вы хотите посмотреть, как атрибуты в данных помогут выполнить **разбиение** или **расщепление** всей совокупности на подмножества, релевантные для проблемы. Презентация набора правил полезна, если вы хотите посмотреть как конкретные группы элементов соотносятся с конкретным заключением. Например, следующее правило даёт нам **профиль** для группы автомобилей, которых стоит покупать:

```
IF tested = 'yes'
AND mileage = 'low'
THEN -> 'BUY'.
```

Алгоритмы построения дерева

Для классификационного и сегментационного анализа доступны четыре алгоритма. Все эти алгоритмы выполняют в основном одно и то же - просматривают все поля из набора данных, чтобы найти то поле, которое даёт наилучшую классификацию или наилучшее предсказание путем разбиения данных на подгруппы. Этот процесс применяется рекурсивно, подгруппы разбиваются на всё меньшие и меньшие блоки до тех пор, пока дерево не будет завершено (что задается определенными критериями остановки). Поля назначения и входные поля, используемые при построении дерева, могут быть непрерывными (числовой диапазон) или категорийными, в зависимости от используемого алгоритма. Если поле назначения непрерывно, генерируется дерево регрессии; если же поле назначения категорийное, генерируется дерево классификации.

Узел дерева классификации и регрессии (Classification and Regression, C&R) генерирует дерево решений, позволяющее предсказывать или классифицировать будущие наблюдения. Этот метод использует рекурсивное разделение, чтобы расщепить обучающие записи на сегменты, на каждом шаге минимизируя неоднородность, причем узел дерева считается “чистым”, если все 100% наблюдений в узле попадают в конкретную категорию поля назначения. Входные поля и поля назначения могут быть из числового диапазона или категориальными (номинальными, порядковыми или флагами); все расщепления бинарны (только две подгруппы).

Узел CHAID генерирует деревья решений, используя статистику хи-квадрат для определения оптимальных расщеплений. В отличие от узлов дерева C&R и QUEST, CHAID может генерировать не только бинарные деревья, то есть у некоторых расщеплений может быть больше двух ветвей. Входные поля и поле назначения могут быть количественными (числовой диапазон) или категориальными. Исчерпывающий CHAID - это модификация метода CHAID, при котором проделывается более тщательная работа по изучению всех возможных расщеплений для каждого предиктора, но это требует больше времени для вычислений.

Узел QUEST предоставляет метод бинарной классификации для построения деревьев решений, разработанный для уменьшения времени обработки, требуемого для анализа больших деревьев C&R, при одновременном подавлении обнаруженного в способах деревьев классификации предпочтения входных полей, допускающих больше расщеплений. Входные поля могут быть в числовом диапазоне (количественными), но поле назначения должно быть категориальным. Все расщепления бинарные.

Узел C5.0 строит или дерево решений, или набор правил. Эта модель работает, разделяя выборку на основании значения в поле, дающего максимальный информационный выигрыш на каждом уровне. Поле назначения должно быть категориальным. Разрешено несколько разделений на подгруппы, и таких подгрупп может быть больше двух.

Общие способы анализа на основе деревьев

Ниже приведены некоторые общие применения анализа на основе деревьев:

Сегментация. Идентифицировать тех, кто, возможно, входит в определенную группу.

Стратификация. Назначить наблюдения в одну из нескольких категорий, например, в группу высокого, среднего или низкого риска.

Прогноз. Создать правила и использовать их для предсказания будущих событий. Прогнозирование может также означать попытки связать предсказываемые атрибуты со значениями некоторых непрерывных переменных.

Сокращение данных и экранирование переменных. Выбрать полезное подмножество предикторов из большого набора переменных для использования при построении формальной параметрической модели.

Идентификация взаимодействий. Идентифицировать взаимосвязи, которые принадлежат определенной подгруппе, и указать их в формальной параметрической модели.

Слияние категорий и дискретизация непрерывных переменных. Перекодировать категории предиктора групп и непрерывные переменные с минимальной потерей информации.

Интерактивный построитель деревьев

Вы можете автоматически сгенерировать модель дерева, позволив алгоритму выбрать наилучшее расщепление на каждом уровне, или же взять управление на себя и использовать интерактивный построитель дерева, при помощи ваших знаний о бизнесе уточняя или упрощая дерево перед сохранением слепка модели.

1. Создайте поток и добавьте в него один из узлов дерева решений: Дерево классификации и регрессии, CHAID или QUEST.

   Примечание: Интерактивное построение деревьев не поддерживается для деревьев C5.0.
2. Откройте узел и на вкладке Поля выберите поля назначения и предиктора, а также задайте необходимые дополнительные опции модели. Подробные инструкции смотрите в документации по каждому узлу построения дерева.
3. На панели Цели вкладки Опции построения выберите **Запустить интерактивный сеанс**.

4. Нажмите кнопку **Запуск**, чтобы запустить построитель деревьев.

Выводится текущее дерево, начиная от корневого узла. Вы можете редактировать и усекать ветви дерева по уровням и обращаться к информации по выигрышу, рискам и связанной информации до того, как сгенерировать одну или несколько моделей.

Комментарии

- Для узлов деревьев классификации и регрессии, CHAID и QUEST у любых порядковых полей, используемых в модели, должна быть числовая система хранения (не строковая). При необходимости для их преобразования может использоваться узел повторной классификации.
- (Необязательно) Можно с помощью поля разделения разделить данные на обучающую выборку и тестовую выборку.
- Помимо использования построителя деревьев, можно также сгенерировать модель непосредственно из узла моделирования, как и с другими моделями IBM SPSS Modeler. Дополнительную информацию смотрите в разделе “Непосредственное построение модели дерева” на стр. 93.

Наращивание и усечение дерева

Вкладка построителя деревьев Программа просмотра позволяет просматривать текущее дерево, начиная от корневого узла.

1. Для наращивания дерева выберите из меню:

   Дерево > Нарастить дерево

   Система построит дерево, рекурсивно выполняя расщепление каждой ветви, пока не будет достигнут один или несколько критериев остановки. В каждой точке расщепления автоматически выбирается наилучший предиктор на основе используемого метода моделирования.
2. Другой вариант - выбрать **Нарастить один уровень дерева**, чтобы добавить один уровень.
3. Чтобы добавить ветвь под определенным узлом, выберите этот узел и выберите **Нарастить ветвь**.
4. Чтобы выбрать предиктор, используемый для расщепления, выберите требуемый узел, затем выберите **Нарастить ветвь с пользовательским разделением**. Дополнительную информацию смотрите в разделе “Определение пользовательских расщеплений” на стр. 84.
5. Чтобы усечь ветвь, выберите узел и выберите **Удалить ветвь**, чтобы очистить выбранный узел.
6. Чтобы удалить из дерева нижний уровень, выберите **Удалить один уровень**.
7. Только для деревьев классификации и регрессии и деревьев QUEST: выберите **Нарастить дерево и усечь**, чтобы провести усечение на основе алгоритма стоимости-сложности, который настраивает оценку рисков, учитывая количество конечных узлов, что обычно позволяет упростить дерево. Дополнительную информацию смотрите в разделе “Узел дерева классификации и регрессии” на стр. 95.

Просмотр правил расщепления на вкладке Программа просмотра

При просмотре правил расщепления на вкладке Программа просмотра квадратные скобки означают, что соседнее значение входит в диапазон, а круглые скобки - что соседнее значение исключено из диапазона. Таким образом, выражение (23,37] означает "от 23 исключительно до 37 включительно", то есть от любого значения, превышающего 23, до 37. На вкладке Модель это же условие будет выглядеть следующим образом:

```
Age > 23 and Age <= 37
```

Прерывание наращивания дерева. Чтобы прервать операцию наращивания дерева (например, если она занимает больше времени, чем предполагалось), нажмите кнопку Остановить выполнение на панели инструментов.

Рисунок 28. Кнопка Остановить выполнение

Эта кнопка доступна только во время наращивания дерева. Она останавливает текущую операцию наращивания в текущей точке, оставив на месте все уже добавленные узлы, без сохранения изменений и закрытия окна. Построитель деревьев остается открытым, позволяя сгенерировать модель, обновить директивы или экспортировать вывод в нужном формате.

Определение пользовательских расщеплений

Диалоговое окно Определить разделение позволяет выбрать предиктор и задать условия для каждого расщепления.

1. В построителе деревьев выберите узел на вкладке Программа просмотра и выберите в меню:

   Дерево > Нарастить ветвь с пользовательским расщеплением
2. Выберите требуемый предиктор в выпадающем списке или нажмите кнопку **Предикторы** для просмотра подробной информации о каждом предикторе. Дополнительную информацию смотрите в разделе “Просмотр подробной информации о предикторах”.
3. Вы можете принять условия по умолчанию для каждого расщепления или выбрать **Пользовательские**, чтобы задать требуемые условия для расщепления.
   - Для непрерывных предикторов (числовых диапазонов) можно задать в полях **Редактировать значения диапазона** диапазон значений, которые будут попадать в каждый новый узел.
   - Для категорических предикторов можно задать в полях **Редактировать значения набора** или **Редактировать порядковые значения** конкретные значения (или диапазон значений в случае порядкового предиктора), которые будут соответствовать каждому новому узлу.
4. Выберите **Нарастить**, чтобы заново нарастить ветвь с использованием выбранного предиктора.

Обычно дерево можно расщепить с использованием любого предиктора независимо от правил остановки. Единственные исключения - когда узел является однородным (то есть 100% случаев попадают в один и тот же класс назначения, и поэтому расщеплять нечего) или когда выбранный предиктор является константой (нет критерия для расщепления).

Пропущенные значения как. Только для деревьев CHAID: если для выбранного предиктора есть пропущенные значения, при определении пользовательского расщепления у вас есть возможность назначить их определенному дочернему узлу. (Для деревьев классификации и регрессии и деревьев QUEST пропущенные значения обрабатываются с использованием суррогатов, как это определено в алгоритме). Дополнительную информацию смотрите в разделе “Подробности расщепления и суррогаты” на стр. 85.)

Просмотр подробной информации о предикторах

В диалоговом окне Выбрать предиктор выводится статистика о доступных предикторах (или, как их иногда называют, "конкурентах"), которые можно использовать для текущего разделения.

- Для CHAID и исчерпывающих CHAID, для каждого категорического предиктора выводится статистика хи-квадрат; если предиктор является числовым диапазоном, выводится статистика *F*. Статистика хи-квадрат - это показатель независимости поля назначения от поля разделения. Высокий показатель статистики хи-квадрат обычно соответствует более низкой вероятности, что означает, что вероятность независимости этих двух полей друг от друга невысока — показатель хорошего разделения. Также включены степени свободы, так как они учитывают тот факт, что тройному разделению проще иметь большую статистику и низкую вероятность, чем двойному разделению.
- Для деревьев классификации и регрессии и QUEST выводится улучшение для каждого предиктора. Чем больше улучшение, тем выше снижение неоднородности между родительским и дочерним узлами при использовании данного предиктора. (Однородный узел - это узел, в котором все случаи попадают в одну категорию назначения; чем ниже неоднородность по всему дереву, тем лучше модель подходит для данных.) Другими словами, высокое значение улучшения обычно обозначает разделение, полезное для дерева этого типа. Показатель неоднородности задается в узле построения дерева.

Подробности расщепления и суррогаты

Для просмотра подробной информации о расщеплении для любого узла можно выбрать этот узел на вкладке Программа просмотра и нажать кнопку информации о расщеплении с правой стороны панели инструментов. Появится используемое правило расщепления вместе с соответствующей статистикой. Для категорических деревьев классификации и регрессии выводится улучшение и ассоциация. Ассоциация - это мера соответствия между суррогатом и первичным полем расщепления, где “наилучший” суррогат обычно тот, который наиболее точно воспроизводит поле расщепления. Для деревьев классификации и регрессии и QUEST выводятся также все суррогаты, используемые вместо первичного предиктора.

Чтобы отредактировать разделение для выбранного узла, можно щелкнуть по значку с левой стороны панели суррогатов, чтобы открыть диалоговое окно Определить расщепление. (Для ускорения операции можно выбрать суррогат в списке перед тем, как щелкнуть по значку, чтобы выбрать его в качестве первичного поля расщепления.)

Суррогаты. Там, где они применимы, все суррогаты для поля первичного расщепления показываются для выбранного узла. Суррогаты - это альтернативные поля, используемые, если первичное значение предиктора отсутствует для данной записи. Максимальное количество суррогатов, разрешенных для данного расщепления, задается на узле построения дерева, но их фактическое количество зависит от обучающих данных. В общем случае, чем больше данных пропущено, тем больше возможных суррогатов будет использовано. Для других моделей дерева решений эта вкладка пуста.

Примечание: для включения в модель суррогаты должны быть определены на фазе обучения. Если в обучающей выборке нет пропущенных значений, никакие суррогаты определены не будут, и все записи с пропущенными значениями, встреченные при проверке или скоринге, автоматически попадут в дочерний узел с максимальным числом записей. Если предполагается, что при проверке и скоринге будут встречаться пропущенные значения, убедитесь, что и в обучающей выборке есть пропущенные значения. Для деревьев CHAID суррогаты недоступны.

Несмотря на то, что суррогаты не используются для деревьев CHAID, при определении пользовательского расщепления есть возможность назначить их определенному дочернему узлу. Дополнительную информацию смотрите в разделе “Определение пользовательских расщеплений” на стр. 84.

Настройка просмотра дерева

На вкладка построителя деревьев Программа просмотра выводится текущее дерево. По умолчанию все ветви в дереве раскрыты, но при необходимости вы можете раскрывать и сворачивать ветви, а также настраивать другие параметры.

- Щелкните по знаку минус (–) в нижнем правом углу родительского узла, чтобы скрыть все его дочерние узлы. Щелкните по знаку плюс (–) в нижнем правом углу родительского узла, чтобы показать его дочерние узлы.
- При помощи меню Просмотр или панели инструментов можно изменять ориентацию вывода (сверху вниз, слева направо или справа налево).
- Нажмите кнопку "Показать метки полей и значений" на главной панели инструментов, чтобы показать или скрыть метки полей и значений.
- Используйте кнопки с изображением увеличительного стекла для приближения или удаления дерева, или нажмите кнопку карты дерева с правой стороны панели инструментов для просмотра диаграммы дерева целиком.
- Если используется поле разделения, можно переключать окно просмотра дерева между разделом обучения и разделом тестирования (**Просмотр > Раздел**). Когда выводится тестовая выборка, дерево можно просматривать, но нельзя редактировать. (Текущий раздел выводится в полосе состояния в нижнем правом углу окна.)
- Для просмотра подробностей текущего разделения нажмите кнопку информации о расщеплении (кнопка "i" в дальнем правом конце панели инструментов). Дополнительную информацию смотрите в разделе “Подробности расщепления и суррогаты”.

- В каждом узле можно просматривать статистику, диаграммы или и то, и другое (смотрите ниже).

Просмотр статистики и диаграмм

Статистика узлов. Для категорического поля назначения таблица в каждом узле показывает число и процентную долю записей в каждой категории, а также процент всей выборки, которому соответствует этот узел. Для непрерывного поля назначения (числового диапазона) таблица показывает среднее значение, стандартное отклонение, число записей и предсказанное значение поля назначения.

Диаграммы узлов. Диаграмма для категорического поля назначения - это столбчатая диаграмма процентных долей в каждой категории поля назначения. Перед каждой строкой в таблице выводится образец цвета, соответствующий цвету, представляющему каждую из категорий поля назначения на диаграммах для этого узла. Диаграмма для непрерывного поля назначения (числового диапазона) - это гистограмма поля назначения для записей в этом узле.

Рост

На вкладке Выигрыш выводится статистика для всех конечных узлов дерева. Выигрыш показывает, насколько среднее значение или пропорция в заданном узле отличается от общего среднего значения. В общем случае, чем больше эта разница, тем полезнее дерево в качестве инструмента для принятия решений. Например, значение индекса или "роста" 148% для узла указывает, что для записей в этом узле вероятность попасть в категорию назначения примерно в полтора раза выше, чем для всего набора данных.

Для узлов деревьев классификации и регрессии и QUEST, где задано множество предотвращения сверхобучения, выводится два множества статистики:

- множество наращивания дерева - обучающая выборка за исключением множества предотвращения сверхобучения
- множество предотвращения сверхобучения

Для других интерактивных деревьев классификации и регрессии и QUEST, а также для всех интерактивных деревьев CHAID выводится только статистика множества наращивания дерева.

Вкладка Выигрыш позволяет:

- Выводить статистику по узлам, кумулятивную статистику или статистику по квантилям.
- Выводить выигрыш или прибыль.
- Переключаться между режимами просмотра таблиц и диаграмм.
- Выбирать категорию назначения (только для категорических назначений).
- Сортировать таблицу в порядке возрастания или убывания процента индекса. Если выводится статистика для нескольких разделов, сортировки всегда применяются к обучающей выборке, а не к тестовой выборке.

В общем случае выбор, сделанный в таблице выигрыша всегда обновляется в окне просмотра дерева, и наоборот. Например, если выбрать строку в таблице, соответствующий узел будет выбран и в дереве.

Классификационный выигрыш

Для деревьев классификации (деревьев с категорической переменной назначения) процент индекса выигрыша показывает, насколько сильнее пропорция выбранной категории назначения в каждом узле отличается от общей пропорции.

Статистика по узлам

В этом режиме просмотра в таблице выводится по одной строке для каждого конечного узла. Например, если общий отклик на кампанию прямой почтовой рассылки составил 10%, но 20% записей, попавших в узел X, дали положительный отклик, процент индекса для данного узла - 200%, что говорит о том, что для респондентов из этой группы вероятность покупки в два раза выше по сравнению с общим числом респондентов.

Для узлов деревьев классификации и регрессии и QUEST, где задано множество предотвращения сверхобучения, выводится два множества статистики:

- множество наращивания дерева - обучающая выборка за исключением множества предотвращения сверхобучения
- множество предотвращения сверхобучения

Для других интерактивных деревьев классификации и регрессии и QUEST, а также для всех интерактивных деревьев CHAID выводится только статистика множества наращивания дерева.

Узлы. ID текущего узла (как он выводится на вкладке Программа просмотра).

Узел: n. Общее число записей в этом узле.

Узел (%). Процент всех записей в наборе данных, попадающий в этот узел.

Выигрыш: n. Число записей с выбранной категорией назначения, попадающих в этот узел. Иными словами, сколько записей из всех записей в наборе данных, попадающих в категорию назначения, находятся в этом узле?

Выигрыш (%). Процент всех записей в категории назначения по всему набору данных, попадающий в этот узел.

Ответ (%). Процент записей в текущем узле, попадающий в категорию назначения. Отклики в этом контексте иногда называются "попаданиями".

Индекс (%). Процент отклика для текущего узла, выраженный как процентная доля от процента отклика для всего набора данных. Например, значение индекса 300% указывает, что для записей в этом узле вероятность попасть в категорию назначения в три раза выше, чем для всего набора данных.

Суммарная статистика

В режиме простора Суммарно в таблице выводится по одному узлу в строке, но статистика является кумулятивной и отсортирована в порядке возрастания или убывания процента индекса. Например, если применяется сортировка по убыванию, узел с наивысшим процентом индекса выводится первым, а статистика в последующих строках является кумулятивной для этой строки и предыдущих строк.

Кумулятивный процент индекса уменьшается строка за строкой по мере добавления узлов со все более низкими процентами отклика. Кумулятивный индекс для последней строки всегда равен 100%, так как в этой точке включен весь набор данных.

Квантили

В этом режиме просмотра каждая строка таблицы представляет квантиль, а не узел. Квантили - это квартили (четверти), квинтили (пятые), децили (десятые), вигинтили (двадцатые) или процентили (сотые). В одном квантиле могут выводиться несколько узлов, если для получения этого процента требуется несколько узлов (например, если выводятся квартили, но верхние два узла содержат менее 50% всех случаев). Остаток таблицы кумулятивен, и его можно интерпретировать таким же образом, как в кумулятивном режиме просмотра.

Прибыли классификации и возврат инвестиций

Для деревьев классификации можно также просматривать статистику выигрыша в терминах прибыли и ROI (return on investment, возврат инвестиций). В диалоговом окне Определить доходы можно задать доходы и расходы для каждой категории.

1. На вкладке Выигрыш нажмите кнопку Прибыль (помеченную как $/$) на панели инструментов, чтобы открыть диалоговое окно.

2. Введите значения доходов и расходов для каждой категории поля назначения.

Например если отправка почтой предложения каждому заказчику стоит вам 0,48 долларов, а доход от положительного отклика составляет 9,95 долларов для трехмесячной подписки, то каждый отклик *нет* стоит вам 0,48 долларов, а каждый отклик *да* приносит вам 9,47 долларов (то есть 9,95 – 0,48).

В таблице выигрыша **прибыль** вычисляется как сумма доходов минус расходы для каждой записи в конечном узле. **ROI** - это общая прибыль, поделенная на общие расходы для узла.

Комментарии

- Значения прибыли влияют только на средние значения прибыли и ROI в таблице выигрыша, как способ просмотра статистики в терминах, более применимых к вашему уровню окупаемости. Они не влияют на основную структуру дерева модели. Не надо путать прибыль со стоимостью неправильной классификации, которая указывается в узле построения дерева и факторизуется в модель как способ защиты от дорогостоящих ошибок.
- Спецификации прибыли не сохраняются между сеансами интерактивного построения дерева.

Регрессивный выигрыш

Для деревьев регресии можно выбрать просмотр по узлам, суммарный просмотр по узлам и просмотр квантилей. В таблице будут показаны средние значения. Диаграммы доступны только для квантилей.

Диаграммы выигрыша

Вместо таблиц на вкладке Выигрыш можно выводить диаграммы.

1. На вкладке Выигрыш выберите значок Квантили (третий значок слева на панели инструментов). (Диаграммы недоступны для статистики по узлам и суммарной статистики.)
2. Выберите значок Диаграммы.
3. В выпадающем списке выберите нужные единицы измерения для вывода (процентили, децили и так далее).
4. Выберите **Выигрыш**, **Отклик** или **Рост**, чтобы изменить показатель для вывода.

Диаграмма выигрыша

Диаграмма выигрыша представляет значения из столбца *Выигрыш (%)* таблицы. Выигрыш определяется как отношение числа попаданий на каждом делении шкалы к общему числу попаданий в дереве с использованием следующей формулы:

(число попаданий на инкремент / общее число попаданий) x 100%

Эта диаграмма наглядно иллюстрирует, насколько широко по дереву нужно производить поиск для набора данного процента всех попаданий. Диагональная линия показывает ожидаемый отклик для всей выборки, если бы модель не использовалась. В этом случае показатель отклика был бы константой, так как вероятность отклика для одного клиента такая же, как для другого. Для удвоения собранной информации нужно бы было опросить вдвое больше людей. Кривая линия обозначает, насколько можно повысить отклик, включив только тех клиентов, которые на основании выигрыша ранжируются в максимальных процентилях. Например, включив в рассылку верхние 50%, можно бы было собрать больше 70% положительных откликов. Чем круче кривая, тем больше выигрыш.

Диаграмма роста

Диаграмма роста представляет значения из столбца *Индекс (%)* таблицы. Эта диаграмма сравнивает процентную долю записей в каждом инкременте, представляющих собой попадания, с общей процентной долей попаданий в обучающем наборе данных, с использованием следующей формулы:

(число попаданий на инкремент / число записей на инкремент) / (общее число попаданий / общее число записей)

Диаграмма отклика

Диаграмма отклика представляет значения из столбца *Отклик (%)* таблицы. Отклик - это процент записей на инкремент, которые оказались попаданиями, с использованием следующей формулы:

(число откликов на инкремент / число записей на инкремент) x 100%

Выбор на основе выигрыша

В диалоговом окне Выбор на основе выигрыша позволяет автоматически выбрать конечные узлы с наилучшим (или наихудшим) выигрышем на основе заданного правила или порогового значения. Затем можно на основе сделанного выбора сгенерировать узел Выбор.

1. На вкладке Выигрыш выберите просмотр по узлам или суммарный просмотр, а также категорию назначения, на основе которой должен быть сделан выбор. (Выбор основан на текущем выводе таблицы и недоступен для квантилей.)
2. На вкладке Выигрыш выберите в меню:

   Изменить > Выбрать конечные узлы > Выбор на основе выигрыша

   Только выбрать. Можно выбрать соответствующие узлы *или* несоответствующие узлы — например, чтобы выбрать *все записи, кроме* верхних 100 записей.

   Сопоставление по информации выигрыша. Сопоставляет узлы на основе статистики выигрыша для текущей категории назначения, включая:

   - Узлы, где выигрыш, отклик или рост (индекс) соответствует заданному пороговому значению — например, отклик больше либо равен 50%.
   - Верхние *n* узлов на основе выигрыша для категории назначения.
   - Верхние узлы до заданного числа записей.
   - Верхние узлы до заданного процента данных обучения.
3. Нажмите кнопку **OK**, чтобы обновить выбор на вкладке Программа просмотра.
4. Чтобы создать новый узел Выбор на основе текущего выбора на вкладке Программа просмотра, выберите **Узел Выбор** в меню Генерировать. Дополнительную информацию смотрите в разделе “Генерирование узлов фильтра и выбора” на стр. 92.

Примечание: Поскольку фактически вы выбираете узлы, а не записи или проценты, не всегда можно достичь идеального соответствия критерию выбора. Система выбирает полные узлы *до* заданного уровня. Например, если выбрать верхние 12 случаев, из которых 10 находятся в первом узле, а два - во втором узле, будет выбран только первый узел.

Риски

Риски показывают вероятность неправильной классификации на любом уровне. На вкладке Риски выводится точечная оценка рисков и (для категорических выводов) таблица неправильных классификаций.

- Для числовых предсказаний риск - это объединенная оценка дисперсии в каждом из конечных узлов.
- Для категорических предсказаний риск - это пропорция неправильно классифицированных случаев с поправкой на априорные вероятности или на стоимость неправильной классификации.

Сохранение моделей деревьев и результатов

Есть несколько способов сохранения или экспорта результатов сеансов интерактивного построения деревьев, в том числе:

- Генерировать модель на основе текущего дерева (**Генерировать > Генерировать модель**).
- Сохранить директивы, использованные для наращивания текущего дерева. При следующем выполнении узла построения дерева текущее дерево будет автоматически наращено заново, включая все определенные вами пользовательские разделения.
- Экспортировать информацию модели, выигрыша и рисков. Дополнительную информацию смотрите в разделе “Экспорт информации модели, выигрыша и рисков” на стр. 92.

В построителе деревьев или в слепке моделей деревьев вы можете:

- Сгенерировать узел Фильтр или узел Выбор на основе текущего дерева. Дополнительную информацию смотрите в разделе “Генерирование узлов фильтра и выбора” на стр. 92.
- Сгенерировать слепок набора правил, представляющий структуру дерева как набор правил, которые определяют конечные ветви дерева. Дополнительную информацию смотрите в разделе “Генерирование набора правил из дерева решений” на стр. 93.
- Коме этого, только для слепков моделей деревьев, вы можете экспортировать модель в формате PMML. Дополнительную информацию смотрите в разделе “Палитра моделей” на стр. 40. Если модель содержит пользовательские расщепления, они не сохраняются в экспортируемом PMML. (Вернее, само расщепление сохраняется, но теряется информация, что оно пользовательское, а не выбранное алгоритмом).
- Сгенерировать диаграмму на основании выбранной части текущего дерева. *Примечание*: это работает только для слепков, присоединенных к другим узлам в потоке. Дополнительную информацию смотрите в разделе “Генерирование графиков” на стр. 111.

Примечание: Само интерактивное дерево невозможно сохранить. Чтобы не потерять выполненную работу, сгенерируйте модель и/или обновите директивы дерева до закрытия окна построителя деревьев.

Генерирование модели из построителя деревьев

Чтобы сгенерировать модель на основе текущего дерева, выберите в меню построителя деревьев:

Генерировать > Модель

В диалоговом окне Генерировать новую модель можно выбрать следующие опции:

Имя модели. Можно задать имя самостоятельно или автоматически генерировать имя на основе имени узла моделирования.

Создать узел на ... Можно добавить узел в требуемом положении - выберите **Холст**, **Палитра GM** или **Оба положения**.

Включить директивы дерева. Чтобы включить в сгенерированную модель директивы из текущего дерева, включите этот переключатель. Это позволит перегенерировать дерево при необходимости. Дополнительную информацию смотрите в разделе “Директивы роста деревьев”.

Директивы роста деревьев

В директивах деревьев для моделей дерева C&R, CHAID и QUEST задаются условия последовательного наращивания уровней дерева. Директивы применяются при каждом запуске интерактивного построителя деревьев из узла.

- Самое надежное использование директив - в качестве способа заново сгенерировать дерево, созданное в предыдущем интерактивном сеансе. Дополнительную информацию смотрите в разделе “Изменение директив дерева” на стр. 92. Кроме того, директивы можно редактировать вручную, хотя это требует осторожности.
- Директивы существенно зависят от структуры дерева, которую они описывают. Поэтому любое изменение в исходных данных или опциях моделирования может привести к ошибке набора директив, который до этого выполнялся успешно. Например, если на основе изменившихся данных алгоритм CHAID превращает разделение на две группы в разделение на три группы, любые директивы, основанные на прошлом разделении на две группы, завершатся неудачно.

Примечание: Если выбрать непосредственное генерирование модели (без построителя дерева), все директивы дерева игнорируются.

Редактирование директив

1. Чтобы просмотреть или отредактировать сохраненные директивы, откройте узел построения дерева и выберите панель Цель на вкладке Опции построения.

2. Выберите **Запустить интерактивный сеанс**, чтобы включить элементы управления, включите переключатель **Использовать директивы дерева** и нажмите кнопку **Директивы**.

Синтаксис директив

Директивы задают условия для роста дерева, начиная с корневого узла. Например, чтобы нарастить дерево на один уровень:

```
Grow Node Index 0 Children 1 2
```

Поскольку предикторы не заданы, алгоритм выбирает наилучшее расщепление.

Обратите внимание на то, что первое расщепление всегда должно выполняться на корневом узле (`Index 0`) и для обоих дочерних узлов должны задаваться значения индексов (в данном случае 1 и 2). Будет недопустимо задать `Grow Node Index 2 Children 3 4`, пока к корневому узлу не приращен узел 2.

Чтобы нарастить дерево:

```
Нарастить дерево
```

Чтобы наращивать и сокращать дерево (только дерево C&R):

```
Grow_And_Prune Tree
```

Чтобы задать пользовательское расщепление для непрерывного предиктора:

```
Grow Node Index 0 Children 1 2 Spliton
  ( "ОБРАЗОВАНИЕ", Interval ( NegativeInfinity, 12.5)
     Interval ( 12.5, Infinity ))
```

Чтобы расщепить по номинальному предиктору с двумя значениями:

```
Grow Node Index 2 Children 3 4 Spliton
  ( "ПОЛ", Group( "0.0" )Group( "1.0" ))
```

Чтобы расщепить по номинальному предиктору со многими значениями:

```
Grow Node Index 6 Children 7 8 Spliton
   ( "ОРГАНИЗАЦИИ", Group( "2.0","4.0" )
              Group( "0.0","1.0","3.0","6.0" ))
```

Чтобы расщепить по порядковому предиктору:

```
Grow Node Index 4 Children 5 6 Spliton
  ( "ДЕТИ", Interval ( NegativeInfinity, 1.0)
              Interval ( 1.0, Infinity ))
```

Примечание: При задании пользовательских расщеплений учитывается регистр букв в именах полей и значениях (`ОБРАЗОВАНИЕ`, `ПОЛ`, `ДЕТИ` и т.д.).

Директивы для деревьев CHAID

Директивы для деревьев CHAID особенно чувствительны к изменениям в данных или в модели, поскольку, в отличии деревьев C&R и QUEST, не ограничены использованием бинарных расщеплений. Например, следующий синтаксис кажется допустимым, но завершится неудачно, если алгоритм разобьет корневой узел более, чем на два дочерних:

```
Grow Node Index 0 Children 1 2
Grow Node Index 1 Children 3 4
```

В дереве CHAID у узла 0 может оказаться 3 или 4 дочерних узла, и тогда вторая строка синтаксиса завершится неудачно.

Использование директив в сценариях

Директивы можно также встраивать в сценарии, используя тройные кавычки.

Изменение директив дерева

Чтобы сохранить работу, сделанную в сеансе интерактивного построения дерева, можно сохранить директивы, использованные для построения текущего дерева. В отличие от сохранения слепка модели, который нельзя затем редактировать, эта возможность позволяет перегенерировать дерево в его текущем состоянии для дальнейшего редактирования.

Чтобы изменить директивы, выберите в меню построителя деревьев:

Файл > Изменить директивы

Директивы сохраняются в узле моделирования, используемом для создания дерева (Дерево классификации и регрессии, QUEST или CHAID), и их можно использовать для повторного генерирования текущего дерева. Дополнительную информацию смотрите в разделе “Директивы роста деревьев” на стр. 90.

Экспорт информации модели, выигрыша и рисков

Из построителя деревьев можно экспортировать статистику модели, выигрыша и рисков в форматах текста, HTML или изображений.

1. В окне построителя деревьев выберите вкладку или окно просмотра, которое вы хотите экспортировать.
2. Выберите в меню:

   Файл > Экспорт
3. Выберите нужный формат (**Текст**, **HTML** или **Диаграмма**) и выберите в подменю конкретные элементы, которые нужно экспортировать.

Там, где это применимо, экспорт выполняется на основе текущего выбора.

Экспорт в форматах Текст или HTML. Вы можете экспортировать статистику выигрыша или рисков для обучающего или тестового раздела (если они определены). Экспорт выполняется на основе текущего выбора на вкладке Выигрыш — например, можно выбрать статистику по узлам, суммарную статистику или статистику по квантилям.

Экспорт графики. Вы можете экспортировать текущее дерево, как оно выводится на вкладке Программа просмотра, или же экспортировать диаграммы выигрыша для обучающего или тестового раздела (если они определены). Доступные форматы - *.JPEG*, *.PNG* и *.BMP*. Для выигрыша экспорт выполняется на основе текущего выбора на вкладке Выигрыш (доступно только при просмотре диаграммы).

Генерирование узлов фильтра и выбора

В окне построителя деревьев или при просмотре слепка модели деревьев решений выберите в меню:

Генерировать > Узел фильтра

или

> Узел выбора

Узел фильтра. Генерирует узел, отфильтровывающий поля, которые не используются текущим деревом. Это быстрый способ сократить набор данных, чтобы он содержал только поля, выбранные алгоритмом в качестве важных. Если есть узел типа, расположенный выше данного узла дерева принятия решений, любые поля с ролью *Назначение* передаются дальше слепком модели Фильтр.

Узел выбора. Генерирует узел, выбирающий все записи, которые попадают в текущий узел. Для этой опции необходимо, чтобы на вкладке Программа просмотра была выбрана по меньшей мере одна ветвь дерева.

Слепок модели помещается на холст потока.

Генерирование набора правил из дерева решений

Вы можете сгенерировать слепок набора правил, представляющий структуру дерева как набор правил, которые определяют конечные ветви дерева. Часто наборы правила могут сохранять большую часть важной информации от полного дерева решений, но с менее сложной моделью. Наиболее важное отличие состоит в том, что при использовании набора правил к любой конкретной записи может применяться несколько правил или не применяться правил вовсе. Например, можно увидеть все правила, предсказывающие выход *нет*, а после них все правила, предсказывающие *да*. Если применяется несколько правил, каждое правило получает взвешенный "голос" на основании показателя доверия, связанного с этим правилом, и конечное предсказание определяется объединением взвешенных голосов правил, которые были применены к рассматриваемой записи. Если никакие правила не применялись, записи присваивается предсказание по умолчанию.

Наборы правил можно сгенерировать только из деревьев с категориальными полями назначения (а не из деревьев регрессии).

В окне построителя деревьев или при просмотре слепка модели деревьев решений выберите в меню:

Создать > Набор правил

Имя набора правил. Позволяет задать имя нового слепка модели набора правил.

Создать узел на ... Управляет положением нового слепка модели набора правил. Выберите **Холст**, **Палитра GM** или **Оба положения**.

Минимальное число экземпляров. Задайте минимальное количество экземпляров (число записей, к которым применяется правило), для сохранения в слепке модели набора правил. Правила с поддержкой меньше выбранного значения не будут включаться в новый набор правил.

Минимальная достоверность. Задайте минимальную достоверность для правил, сохраняемых в слепке модели набора правил. Правила с достоверностью меньше выбранного значения не будут включаться в новый набор правил.

Непосредственное построение модели дерева

Вместо использования интерактивного построителя деревьев вы можете построить модель дерева решений непосредственно из узла при выполнении потока. Такой подход применяется для большинства других узлов построения модели. Для моделей дерева C5.0, которые не поддерживает интерактивный построителеь деревьев, только такой метод и доступен.

1. Создайте поток и добавьте один из узлов дерева решений - дерево C&R, CHAID, QUEST или C5.0.
2. В случае дерева C&R, QUEST или CHAID выберите на панели Цель на вкладке Опции построения одну из главных целей. Выбирая Построить одно дерево, убедитесь, что для Режима задано значение **Сгенерировать модель**.

   В случае C5.0 на вкладке модели задайте **Тип вывода Дерево решений**.
3. Выберите поля назначения и предиктора и задайте дополнительные опции модели, если нужно. Конкретные указания смотрите в документации по нужному узлу построения дерева.
4. Выполните поток, чтобы сгенерировать модель.

Комментарии

- При генерировании деревьев этим методом директивы по росту деревьев игнорируются.
- В обоих методах создания деревьев решений, интерактивном и непосредственном, генерируются сходные модели. Различие в основном в степени контроля над процессом.

Узлы деревьев решений

Узлы дерева решений в IBM SPSS Modeler обеспечивают доступ к описанным выше алгоритмам построения деревьев:

- Дерево C&R
- QUEST
- CHAID
- C5.0

Дополнительную информацию смотрите в разделе “Модели деревьев решений” на стр. 81.

Алгоритмы сходны в том отношении, что все они способны построить дерево решений, рекурсивно расщепляя данные на все меньшие подгруппы. Однако есть и важные различия.

Входные поля. Входные поля (предикторы) могут быть любого из следующих типов измерений: непрерывные, категориальные, флаговые, номинальные и порядковые.

Поля назначения. Можно задать только одно поле назначения. Для дерева C&R и CHAID поле назначения может быть непрерывное, категориальное, флаговое, номинальное или порядковое. Для QUEST оно может быть категориальным, флаговым или номинальным. Для C5.0 поле назначения может быть флаговым, номинальным или ординальным.

Тип расщепления. Дерево C&R и QUEST поддерживают только бинарные расщепления (то есть каждый узел может быть расщеплен только на две ветви). В отличие от них, CHAID и C5.0 поддерживают расщепление сразу на три и более ветви.

Метод расщепления. Алгоритмы различаются по критериям принятия решений о расщеплении. Когда дерево C&R предсказывает категориальное выходное поле, используется тот или иной показатель дисперсии (по умолчанию - коэффициент Джини, но вы можете задать и другой показатель). Для непрерывных полей назначения используется метод наименьшего среднеквадратичного отклонения. CHAID использует критерий хи-квадрат; QUEST использует критерий хи-квадрат для категориальных предикторов и дисперсионный анализ для непрерывных входных полей. Для C5.0 используется показатель из теории информации, нормализованный прирост информации.

Обработка пропущенных значений. Все алгоритмы допускают пропущенные значения в полях-предикторах, хотя обрабатывают их по-разному. Дерево C&R и QUEST используют по мере необходимости суррогатные поля предсказаний для проведения записи с пропущенными значениями по дереву во время обучения. CHAID выделяет пропущенные значения в отдельную категорию и разрешает их использование в построении дерева. C5.0 использует метод дробления, при котором часть записи пропускается через каждую ветвь узла, выполняющего расщепление на основе поля с пропущенным значением.

Сокращения. Дерево C&R, QUEST и C5.0 поддерживают возможность сначала вырастить дерево полностью, а затем сократить его, удалив те расщепления нижнего уровня, которые не вносят существенного вклада в точность дерева. Однако все алгоритмы деревьев решений предусматривают возможность ограничить минимальный размер подгруппы, чтобы избегать ветвей со слишком малым числом записей данных.

Интерактивное построение деревьев. Для дерева C&R, QUEST и CHAID поддерживается возможность запустить интерактивный сеанс. В нем можно построить дерево уровень за уровнем, отредактировать расщепления и сократить дерево, перед тем как создать слепок модели. C5.0 не поддерживает интерактивную опцию.

Априорные вероятности. Дерево C&R и QUEST поддерживают задание априорных вероятностей для категорий при предсказании значения категориального выходного поля. Априорные вероятности - это оценки общей относительной частоты каждой категории назначения в совокупности, из которой извлекают обучающие данные. Другими словами, это оценки вероятности для каждого возможного значения

назначения, которые можно использовать априори, до того, как вы будете знать что-либо о предикторных значениях. CHAID и C5.0 не поддерживают задание априорных вероятностей

Наборы правил. Для моделей с категориальными полями назначения узлы дерева решений предлагают опцию создать модель в формате набора правил, что иногда удобнее для понимания, чем сложной дерево решений. Для дерева C&R Tree, QUEST и CHAID можно сгенерировать набор правил из интерактивного сеанса; для C5.0 эту опцию можно задать на узле моделирования. Кроме того, все модели деревьев решений дают возможность сгенерировать набор правил из слепка модели. Дополнительную информацию смотрите в разделе “Генерирование набора правил из дерева решений” на стр. 93.

Узел дерева классификации и регрессии

Узел дерева классификации и регрессии (Classification and Regression, C&R) - это основанный на древообразной структуре метод классификации и предсказания. Аналогично C5.0, этот метод использует рекурсивное разделение для расщепления обучающих записей на сегменты с похожими значениями выходного поля. Узел дерева C&R начинает работу с проверки входных полей, чтобы найти наилучшее расщепление, измеряемое уменьшением индекса неоднородности после расщепления. Расщепление определяет две подгруппы, каждая из которых последовательно делится на следующие две подгруппы и так далее, пока не сработает один из критериев остановки. Все расщепления бинарны (только на две подгруппы).

Отсечения

Деревья C&R предоставляют возможность сначала вырастить дерево, а затем провести усечение на основе алгоритма стоимости-сложности, который настраивает оценку рисков, учитывая количество конечных узлов. Этот метод, допускающий большой рост дерева до применения усечения на основе более сложных критериев, может привести к деревьям меньшего размера с лучшими свойствами перекрестной проверки. Увеличение количества конечных узлов в общем случае уменьшает риск для текущих (обучающих) данных, но фактический риск может увеличиться при обобщении модели на скрытые данные. Рассмотрим предельный случай. Пусть у вас есть отдельный конечный узел на каждую запись в обучающем наборе. Оценка риска составит 0%, так как каждая запись попадает на свой собственный узел, но риск неправильной классификации для невидимых данных (из набора тестирования) почти наверняка будет больше нуля. Для компенсации этого эффекта используется алгоритм стоимости-сложности.

Пример. Компания кабельного телевидения заказала маркетинговое исследование для определения, какие клиенты купят подписку на службу интерактивных новостей по кабельному каналу. Используя данные изучения, вы можете создать поток, в котором полем назначения будет намерение купить подписку, а предикторные поля будут включать в себя возраст, пол, образование, категорию дохода, количество проводимых в день у телевизора часов и количество детей. Применяя к потоку узел дерева C&R, вы сможете предсказать и классифицировать ответы для получения максимального уровня ответов для кампании по продвижению этой подписки.

Требования. Для обучения модели дерева C&R вам нужно одно или несколько *Входных* полей и ровно одно поле *Назначения*. Входные поля и поле назначения могут быть количественными (числовой диапазон) или категориальными. Поля с заданными значениями *Оба* или *Нет* игнорируются. У используемых в модели полей должны быть полностью конкретизированы типы, и у любых порядковых полей (упорядоченный набор), используемых в модели, должна быть числовая система хранения (не строковая). При необходимости может использоваться узел повторной классификации.

Достоинства. Модели дерева C&R довольно устойчивы к наличию таких проблем, как пропущенные данные или большое количество входных полей. Обычно этим моделям для оценки не требуется большое время на обучение. Кроме этого, модели C&R проще для понимания по сравнению с некоторыми другими моделями, так как у получаемых из этой модели правил есть непосредственная интерпретация. В отличие от C5.0, дерево C&R может включать в себя и количественные, и категориальные выходные поля.

Узел CHAID

CHAID (Chi-squared Automatic Interaction Detection - автоматическое обнаружение взаимодействия хи-квадрат) - это метод классификации для построения деревьев решений с использованием статистики хи-квадрат для нахождения оптимальных расщеплений.

CHAID сначала изучает таблицы сопряженности между каждым из входных полей и выходными данными, а затем проверяет важность, используя критерий независимости хи-квадрат. Если несколько из этих связей статистически значимы, CHAID выбирает наиболее значимое входное поле (наименьшее значение *p*). Если у входных данных больше двух категорий, они сравниваются, и категории, не показывающие различий выходных данных, свертываются. Это делается объединением пары категорий, показывающих наименее значимое различие. Такой процесс слияния категорий останавливается, когда все оставшиеся категории отличаются на заданном уровне тестирования. Для номинальных входных полей можно объединить все категории; для упорядоченного набора объединяются только количественные категории.

Исчерпывающий CHAID - это модификация метода CHAID, при которой проделывается более тщательная работа по изучению всех возможных расщеплений для каждого предиктора, но это требует больше времени для вычислений.

Требования. Входные поля и поле назначения могут быть количественными или категориальными; на каждом уровне узлы можно расщеплять на две или более подгрупп. У любых порядковых полей, используемых в модели, должна быть числовая система хранения (не строковая). При необходимости для их преобразования может использоваться узел повторной классификации.

Достоинства. В отличие от узлов дерева C&R и QUEST, CHAID может генерировать не только бинарные деревья, то есть у некоторых расщеплений может быть больше двух ветвей. Поэтому у данного метода есть тенденция создания более широких деревьев, чем у методов бинарного роста. CHAID работает со всеми типами входных данных, а также принимает веса наблюдений и частотные переменные.

Узел QUEST

QUEST (Quick, Unbiased, Efficient Statistical Tree), то есть быстрое, несмещенное, эффективное статистическое дерево, - это двоичный метод классификации для построения деревьев решений. Главной мотивацией при разработке этого метода было сокращение времени обработки, требующееся для больших анализов дерева C&R с многими переменными или наблюдениями. Второй целью создания QUEST было сокращение обнаруженной в методах деревьев классификации тенденции предпочтения входных полей, допускающих больше расщеплений, то есть количественных (числовой диапазон) полей или полей с большим числом категорий.

- QUEST использует основанный на критериях значимости набор правил, чтобы оценивать входные поля на узле. В целях выбора необходимо произвести по крайней мере одно испытание для каждого входного поля на узле. В отличие от дерева C&R все расщепления не проверяются, и в отличие от дерева C&R и CHAID комбинации категорий также не проверяются при оценке входного поля для выбора. Это ускоряет анализ.
- Расщепления определяются при запуске анализа квадратичного дискриминанта с использованием выбранных входных полей для групп, сформированных категориями назначения. Этот метод снова приводит к повышению производительности по сравнению с исчерпывающим поиском (дерево C&R) для определения оптимального расщепления.

Требования. Входные поля могут быть в числовом диапазоне (количественными), но поле назначения должно быть категориальным. Все расщепления бинарные. Поля веса не используются. У любых порядковых полей (упорядоченный набор), используемых в модели, должна быть числовая система хранения (не строковая). При необходимости для их преобразования может использоваться узел повторной классификации.

Достоинства. Как и в CHAID, но в отличие от дерева C&R, QUEST использует статистические критерии для решения, будет ли использоваться некоторое входное поле. В этом методе разделяются также вопросы выбора входных полей и расщепления, так как для этих задач используются разные критерии. Это отличается от CHAID, где результат статистического критерия, определяющий выбор переменных,

используется и для создания расщепления. Аналогично, в дереве C&R и для выбора входных полей, и для определения расщепления используется показатель изменения неоднородности.

Опции полей узла дерева решений

На вкладке Поля указывается, будут ли использоваться значения ролей полей, уже определенные на расположенных выше узлах, или назначение полей будет выполнено вручную.

Использовать заранее заданные роли. Эта опция применяет параметры ролей (назначений, предикторов и так далее) с восходящего узла Тип (или вкладки Типы восходящего узла источника).

Настроить назначения полей. Выберите эту опцию, если хотите назначить объекты назначения, предикторы и другие роли вручную на этом экране.

Поля. При помощи кнопок со стрелками назначьте элементы из этого списка вручную для различных полей ролей в правой части экрана. Значки указывают для каждого поля роли допустимые уровни измерения.

Нажмите кнопку **Все**, чтобы выбрать все поля в списке, или кнопку для отдельного уровня измерения, чтобы выбрать все поля с этим уровнем измерения.

Цель. Выберите одно поле в качестве назначения для предсказания.

Предикторы(входные поля). Выберите одно или несколько полей как входные поля для предсказания.

Анализируемый вес. (Только для CHAID и C&RT) Чтобы использовать некоторое поле как вес наблюдений, укажите это поле здесь. Веса наблюдений служат для компенсации различий в дисперсии разных уровней выходного поля. Дополнительную информацию смотрите в разделе “Использование полей частоты и веса” на стр. 32.

Опции построения узла дерева решений

На вкладке Параметры конструкции задаются все опции для построения модели. Можно, конечно, просто нажать кнопку **Выполнить**, чтобы построить модель со всеми опциями по умолчанию, но скорее всего вы захотите настроить конструкцию для своих собственных целей.

Здесь можно выбрать между построением новой модели и обновлением существующей. Кроме того, задается главная цель узла: построить стандартную модель, построить модель повышенной точности или стабильности, построить модель для работы с очень большими наборами данных.

Что вы хотите сделать?

Построить новую модель. (По умолчанию) При каждом выполнении потока, содержащего этот узел моделирования, строится полностью новая модель.

Продолжить обучение существующей модели. По умолчанию при каждом выполнении узла моделирования создается полностью новая модель. Если выбрана эта опция, обучение продолжается с последней модели, успешно созданной узлом. Это дает возможность скорректировать или обновить существующую модель без необходимости обращаться к исходным данным, что может выполняться значительно быстрее, так как в поток вводятся *только* новые или обновленные записи. Информация по предыдущей модели сохраняется вместе с узлом моделирования, что позволяет использовать этот вариант, даже если предыдущий nugget модели недоступен в потоке или Палитре моделей.

Примечание: Эта опция активируется, только если в качестве цели выбрать **Создать модель для очень большого набора данных**.

Какова Ваша главная цель?

- **Построить единичное дерево.** Создается единичная стандартная модель дерева решений. Стандартные модели обычно проще для интерпретации и быстрее оцениваются по сравнению с моделями, построенными при других вариантах цели.

  Режим. Задает метод, использованный для построения модели. **Сгенерировать модель** создает модель автоматически при выполнении потока. **Запустить интерактивный сеанс** открывает построитель деревьев, где можно построить дерево уровень за уровнем, отредактировать расщепления и сократить дерево, перед тем как создать слепок модели.

  Использовать директивы дерева. Выберите эту опцию, чтобы задать директивы интерактивного генерирования дерева из узла. Например, можно задать расщепления первого и второго уровня, и эти указания будут автоматически применяться при запуске построителя дерева. Кроме того, можно сохранить директивы из сеанса интерактивного построения дерева, чтобы в будущем воссоздавать такое же дерево. Дополнительную информацию смотрите в разделе “Изменение директив дерева” на стр. 92.
- **Повысить точность модели (бустинг).** Выберите эту опцию, если нужно использовать так называемый **бустинг**, специальный метод для повышения уровня точности модели. Бустинг работает, последовательно создавая несколько моделей. Первая модель строится обычным образом. Затем при построении второй модели особое внимание уделяется тем записям, которые были неправильно классифицированы первой моделью. Третья модель фокусируется на ошибках второй модели и так далее. В конечном итоге наблюдения классифицируются с применением к ним всего набора моделей и с использованием взвешенной процедуры голосования, чтобы объединить отдельные предсказания в одно общее. Бустинг может существенно повысить точность модели дерева решений, но он требует также большего времени на обучение.
- **Повысить стабильность модели (бэггинг).** Выберите эту опцию, если нужно использовать так называемый **бэггинг** (бутстреп-агрегирование), специальный метод для повышения стабильности модели и предотвращения сверхобучения. Эта опция создает несколько моделей и комбинирует их для получения более надежных предсказаний. Модели, полученные с использованием этой опции, могут занять больше времени для их построения и оценки, чем стандартные модели.
- **Создать модель для очень большого набора данных.** Выберите эту опцию при работе с базами данных, которые так велики, что их невозможно построить при помощи остальных вариантов цели. Эта опция подразделяет данные на меньшие блоки и строит модели для отдельных блоков. Затем автоматически выбираются наиболее точные модели, из которых комбинируется единый слепок модели. Если на этом экране включить переключатель **Продолжить обучение существующей модели**, можно выполнять инкрементные обновления модели. *Примечание*: Для опции очень больших наборов данных требуется соединение с сервером IBM SPSS Modeler

Узлы деревьев решений - основные опции

Здесь задаются основные опции построения дерева решений.

Алгоритм формирования дерева. (Только для CHAID) Выберите нужный тип алгоритма **CHAID**. **Исчерпывающий CHAID** - это модификация метода CHAID, при которой проделывается более тщательная работа по изучению всех возможных расщеплений для каждого предиктора, но это требует больше времени для вычислений.

Максимальное количество уровней в дереве. Задайте максимальное количество уровней под корневым узлом (то есть сколько раз будет рекурсивно расщеплена выборка). Значение по умолчанию - 5; чтобы задать другое число уровней, выберите **Пользовательское** и введите значение.

Сокращение (только для C&RT и QUEST)

Отсекать ветви, чтобы избежать переобучения. Сокращение дерева состоит в удалении расщеплений нижнего уровня, которые не вносят существенного вклада в точность дерева. Сокращение дерева упрощает его, облегчая его интерпретацию и, в некоторых случаях, улучшая обобщение. Если потребуется полное дерево без сокращения, оставьте эту опцию не выбранной.

- **Максимальное отличие риска (в стандартных ошибках).** Дает возможность задать более либеральное правило сокращения. Правило стандартной ошибки разрешают алгоритму выбрать простейшее дерево с оценкой риска, близкой (возможно, с превышением) к оценке риска минимального по риску поддерева. Это значение показывает допустимую величину различия по оценке риска между сокращенным деревом и деревом с наименьшей оценкой риска. Например, если задать 2, может быть выбрано дерево с оценкой риска, которая на (2 × стандартная ошибка) превышает оценку риска полного дерева.

Максимум суррогатов. Суррогаты - это метод обработки пропущенных значений. Для каждого расщепления в дереве алгоритм находит входные поля, наиболее сходные с выбранным полем расщепления. Эти поля назначаются **суррогатами** данного расщепления. Когда нужно классифицировать запись, у которой пропущено значение поля расщепления, для расщепления используется значение суррогатного поля. Увеличение данного значение повышает гибкость при обработке пропущенных значений, но может также увеличить расход памяти и время обучения.

Узлы деревьев решений - правила остановки

Эти опции управляют тем, как строитсяя дерево. Правила остановки определяют, когда остановить расщеплние конкретных ветвей дерева. Задайте минимальные размеры ветвей для предотвращения расщеплений, создающих очень маленькие подгруппы. **Минимальное число записей в родительской ветви** предотвратит расщепление, если количество записей на узле для расщепления (**родительском**) будет меньше, чем заданное значение. **Минимальное число записей в дочерней ветви** предотвратит расщепление, если количество записей в любой ветви после расщепления (**дочерней**) окажется меньше заданного значения.

- **Использовать процент.** Позволяет указывать размеры в форме процентной доли всех данных обучения.
- **Использовать абсолютное значение.** Позволяет задавать размеры как абсолютное количество записей.

Узлы деревьев решений - ансамбли

Данные параметры определяют поведение ансамбля, которое имеет место, когда на вкладке Цели запрашивается бэггинг, бустинг или очень большие наборы данных. Параметры, которые не применяются к выбранной цели, игнорируются.

Бэггинг и очень большие наборы данных. Это правило, которое применяется при скоринге ансамбля, чтобы объединить предсказанные значения для базовых моделей с целью вычисления значений скоринга для ансамбля.

- **Принятое по умолчанию правило объединения для категориальных полей назначения.** Предсказанные значения для ансамбля в случае категориальных целевых переменных можно объединить, используя голосование, наибольшую вероятность или наибольшую среднюю вероятность. **Голосование** позволяет выбрать категорию, которая наиболее часто имеет наибольшую вероятность в базовых моделях. **Наибольшая вероятность** позволяет выбрать категорию, которая имеет наибольшую вероятность среди всех базовых моделей. **Наибольшая средняя вероятность** позволяет выбрать категорию, которая имеет наибольшую вероятность при усреднении вероятностей категорий по базовым моделям.
- **Принятое по умолчанию правило объединения для непрерывных полей назначения.** Предсказанные значения для ансамбля в случае непрерывных целевых полей могут быть вычислены с использованием среднего значения или медианы предсказанных значений для базовых моделей.

Обратите внимание на то, что если цель состоит в повышении точности модели, выбор правила объединения игнорируется. При бустинге всегда используется взвешенное решение большинством голосов для скоринга категориальных целевых полей и взвешенная медиана для скоринга непрерывных целевых полей.

Бустинг и бэггинг. Задайте число базовых моделей для построения, когда целью является повышение точности или стабильности; для бэггинга это число бутстреп-выборок. Оно должно быть положительным целым.

Узлы дерева C&R и QUEST - стоимости и априорные вероятности

Стоимости ошибочной классификации

В некоторых контекстах определенные виды ошибок обходятся пользователю дороже других. Например, может оказаться более дорогостоящим классифицировать претендента на кредит с высоким уровнем риска, как с низким уровнем риска (один вид ошибки), чем классифицировать претендента на кредит с низким уровнем риска риском как с высокими уровнем риска (другой вид ошибки). Стоимости ошибочной классификации позволяют задать относительную важность различных видов ошибок предсказания.

Стоимости ошибочной классификации - это по существу веса, применяемые к конкретным исходам. Эти веса факторизуются в модель и могут фактически изменить предсказание (в качестве способа защиты от дорогостоящих ошибок).

За исключением моделей C5.0, стоимости ошибочной классификации при скоринге моделей не применяются, и при ранжировании или сравнении моделей во внимание не принимаются. Модель, включающая в себя стоимости, не может дать меньше ошибок, чем та, которая не ранжируется и не может ранжироваться хоть сколь-нибудь выше в единицах общей точности, но, скорее всего, она будет выполняться лучше на практике, поскольку в ней заложено предусмотренное смещение в пользу *менее дорогостоящих* ошибок.

Матрица стоимостей показывает стоимость для каждого возможного сочетания предсказанной и действительной категорий. По умолчанию для всех стоимостей ошибочной классификации задается значение 1,0. Чтобы ввести пользовательские значения стоимостей, выберите **Использовать стоимости ошибочной классификации** и введите в матрицу стоимостей нужные вам значения.

Чтобы изменить стоимость ошибочной классификации, выберите ячейку, соответствующую нужному сочетанию предсказанного и действительного значений, удалите существующее содержание ячейки и введите для нее желаемую стоимость. Стоимости не являются автоматически симметричными. Например, если для стоимости ошибочной классификации *A* как *B* задать значение 2,0, у стоимости ошибочной классификации *B* как *A* все равно будет значение по умолчанию 1,0, пока вы не измените также и его явным образом.

Априорные вероятности

Эти опции позволяют задать априорные вероятности для категорий при предсказании значения категориального выходного поля. **Априорные вероятности** - это оценки общей относительной частоты каждой категории назначения в совокупности, из которой извлекают обучающие данные. Другими словами, это оценки вероятности для каждого возможного значения назначения, которые можно использовать *априори*, до того, как вы будете знать что-либо о предикторных значениях. Есть три способа, как можно задать априорные вероятности:

- **На основе обучающих данных.** Это задано по умолчанию. Априорные вероятности основаны на относительной частоте категория в обучающих данных.
- **Равные для всех классов.** Априорные вероятности для всех категорий определены как $1/k$, где k - это количество категорий назначения.
- **Пользовательская.** Вы можете определить априорные вероятности по собственному выбору. Начальные значения априорных вероятностей заданы равными для всех классов. Вы можете заменить вероятности конкретных категорий на пользовательские значения. Чтобы скорректировать вероятность конкретной категории, выберите в таблице ячейку вероятности, соответствующую нужной категории, удалите ее содержимое и введите свое значение.

Сумма априорных вероятностей для всех категорий должна быть равна 1,0 (**условие нормировки**). Если их сумма не равна 1,0, выводится предупреждение с опцией автоматической нормализации значений. Эта автоматическая коррекция сохраняет пропорции между категориями, обеспечивая выполнение условия нормировки. Эту корректировку можно выполнить в любое время, нажав кнопку **Нормализовать**. Чтобы восстановить в таблице равные значения для всех категорий, нажмите кнопку **Уравнять**.

Корректировать вероятности, используя стоимости ошибок классификации. При помощи этой опции можно уточнять априорные вероятности с учетом стоимостей ошибок классификации (заданной на вкладке стоимостей). Таким образом вы можете заложить информацию о стоимости непосредственно в процесс

роста деревьев, использующий показатель неоднородности бинаризации. (Когда эта опция не выбрана, информация о стоимости используется только при классификации записей и оценке рисков для деревьев на основе показателя бинаризации.)

Узел CHAID - стоимости

В некоторых контекстах определенные виды ошибок обходятся пользователю дороже других. Например, может оказаться более дорогостоящим классифицировать претендента на кредит с высоким уровнем риска, как с низким уровнем риска (один вид ошибки), чем классифицировать претендента на кредит с низким уровнем риска риском как с высокими уровнем риска (другой вид ошибки). Стоимости ошибочной классификации позволяют задать относительную важность различных видов ошибок предсказания.

Стоимости ошибочной классификации - это по существу веса, применяемые к конкретным исходам. Эти веса факторизуются в модель и могут фактически изменить предсказание (в качестве способа защиты от дорогостоящих ошибок).

За исключением моделей C5.0, стоимости ошибочной классификации при скоринге моделей не применяются, и при ранжировании или сравнении моделей во внимание не принимаются. Модель, включающая в себя стоимости, не может дать меньше ошибок, чем та, которая не ранжируется и не может ранжироваться хоть сколь-нибудь выше в единицах общей точности, но, скорее всего, она будет выполняться лучше на практике, поскольку в ней заложено предусмотренное смещение в пользу *менее дорогостоящих* ошибок.

Матрица стоимостей показывает стоимость для каждого возможного сочетания предсказанной и действительной категорий. По умолчанию для всех стоимостей ошибочной классификации задается значение 1,0. Чтобы ввести пользовательские значения стоимостей, выберите **Использовать стоимости ошибочной классификации** и введите в матрицу стоимостей нужные вам значения.

Чтобы изменить стоимость ошибочной классификации, выберите ячейку, соответствующую нужному сочетанию предсказанного и действительного значений, удалите существующее содержание ячейки и введите для нее желаемую стоимость. Стоимости не являются автоматически симметричными. Например, если для стоимости ошибочной классификации *A* как *B* задать значение 2,0, у стоимости ошибочной классификации *B* как *A* все равно будет значение по умолчанию 1,0, пока вы не измените также и его явным образом.

Узел дерева классификации и регрессии - Дополнительные опции

Дополнительные опции узла служат для точной настройки процесса построения деревьев.

Минимальное изменение неоднородности. Задайте минимальное изменение неоднородности для создания нового расщепления в дереве. Понятие **неоднородности** относится к тому, в какой степени у подгрупп, определенных деревом, широкий диапазон значений выходных полей в каждой группе. Для категориальный полей назначения узел считается узел дерева считается “чистым”, если все 100% наблюдений в узле попадают в конкретную категорию поля назначения.“чистым", если все 100% наблюдений в узле попадают в конкретную категорию поля назначения. Цель построения дерева - создать подгруппы со сходными выходными значениями, другими словами, минимизировать неоднородности на каждом узле. Если наилучшее расщепление ветви уменьшает неоднородность меньше, чем на заданную величину, расщепление не производится.

Мера неоднородности для категориальных полей назначения. Для категориальных полей назначения задайте метод для измерения неоднородности дерева. (Для непрерывных полей назначения эта опция игнорируется, а как показатель неоднородности всегда используется **наименьший квадрат отклонения**.)

- **Джини** - общий показатель неоднородности, основанный на вероятностях принадлежности к категории для ветви.
- **Бинаризация** - показатель неоднородности, оптимизированный для бинарного расщепления и с большей вероятностью обеспечивающий примерно одинаковый размер ветвей расщепления.
- **Упорядоченность** добавляет то ограничение, что группируются только идущие подряд классы назначения, что применимо только к порядковым полям назначения. Если эта опция выбрана при номинальном поле назначения, то используется стандартный показатель бинаризации по умолчанию.

Множество предотвращения сверхобучения. По этому алгоритму записи внутренне разделяются на множество построения моделей и множество предотвращения сверхобучения, служащее независимым набором записей данных, используемым в целях отслеживания ошибок в ходе обучения, чтобы не допустить учета в модели случайных изменений данных. Задайте процент записей. Значение по умолчанию - 30.

Воспроизвести результаты. Задание стартового числа генератора псевдослучайных чисел позволяет воспроизвести результаты. Задайте целое число или щелкните по **Генерировать**, чтобы сгенерировать псевдослучайное целое число в диапазоне между 1 и 2147483647 включительно.

Узел QUEST - дополнительные опции

Дополнительные опции узла служат для точной настройки процесса построения деревьев.

Уровень значимости для расщепления. Задает уровень значимости (альфа) для узлов расщепления. Это значение должно быть от 0 до 1. У меньших значений есть тенденция к созданию деревьев с меньшим количеством узлов.

Множество предотвращения сверхобучения . По этому алгоритму записи внутренне разделяются на множество построения моделей и множество предотвращения сверхобучения, служащее независимым набором записей данных, используемым в целях отслеживания ошибок в ходе обучения, чтобы не допустить учета в модели случайных изменений данных. Задайте процент записей. Значение по умолчанию - 30.

Воспроизвести результаты. Задание стартового числа генератора псевдослучайных чисел позволяет воспроизвести результаты. Задайте целое число или щелкните по **Генерировать**, чтобы сгенерировать псевдослучайное целое число в диапазоне между 1 и 2147483647 включительно.

Узел CHAID - дополнительные опции

Дополнительные опции узла служат для точной настройки процесса построения деревьев.

Уровень значимости для расщепления. Задает уровень значимости (альфа) для узлов расщепления. Это значение должно быть от 0 до 1. У меньших значений есть тенденция к созданию деревьев с меньшим количеством узлов.

Уровень значимости для слияния. Задает уровень значимости (альфа) для слияния категорий. Значение должно быть больше 0 и не больше 1. Чтобы вообще не сливать категории, задайте значение 1. Для непрерывных полей назначения это означает, что число категорий для этой переменной в итоговом дереве соответствует заданному числу интервалов. Эта возможность недоступна для Исчерпывающего CHAID.

Скорректировать уровни значимости, используя метод Бонферрони. Корректирует уровни значимости при тестировании различных сочетаний категорий предиктора. При корректировке уровней учитывается число испытаний, которое находится в прямой зависимости от числа категорий и типа измерений предиктора. Обычно это выгодно, потому что помогает сдерживать частоту ложно-положительных результатов. Если эту опцию выключить, повысится способность анализа отыскивать подлинные различия, но ценой повышения частоты ложно-положительных результатов. Выключение этой опции можно рекомендовать, в частности, для малых выборок.

Допускать разбиение объединенных категорий в узле. Алгоритм CHAID пытается объединять категории, чтобы получить простейшее дерево, описывающее модель. Если эта опция выбрана, она разрешает снова расщеплять объединенные категории, если это приводит к лучшему решению.

Хи-квадрат для категориальных полей назначения. Для категориальных полей назначения можно задать метод вычисления статистического показателя хи-квадрат.

- **Пирсона.** Этот метод обеспечивает более быстрые вычисления, но для малых выборок его следует использовать с осторожностью.
- **Отношение правдоподобия.** Этот метод более устойчив, чем метод Пирсона, но вычисления занимают больше времени. Это предпочтительный метод для малых выборок. Для непрерывных полей этот метод используется всегда.

Минимальное изменение ожидаемых частот в ячейках. При оценке частот в ячейках (как для номинальной модели, так и для модели порядковых номеров эффектов строк), используется процедура итераций (эпсилон), которые сходятся к оптимальной оценке, используемой в критерии хи-квадрат при конкретном расщеплении. Эпсилон оценивает величину изменений, необходимых для продолжения итераций; если величина изменения из последней итерации меньше заданного значения, итерации прекращаются. Если возникают проблемы со сходимостью алгоритма, можете увеличить это значение или увеличить максимальное число итераций до сходимости.

Максимум итераций до сходимости. Задает максимальное число итераций, после которого они прекращаются, даже если сходимость не достигнута.

Воспроизвести результаты. Задание стартового числа генератора псевдослучайных чисел позволяет воспроизвести результаты. Задайте целое число или щелкните по **Генерировать**, чтобы сгенерировать псевдослучайное целое число в диапазоне между 1 и 2147483647 включительно.

Опции модели узла дерева решений

На вкладке Опции модели можно выбрать, задавать ли имя для модели, или сгенерировать имя автоматически. Кроме того, можно задать получение информации о важности предиктора, а также простые и скорректированные оценки склонности для флаговых полей назначения.

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Оценка модели

Вычислить важность предикторов. Для моделей, производящих соответствующую меру важности, можно вывести диаграмму, показывающую относительную важность каждого предиктора при оценке модели. Обычно вам потребуется направить усилия по моделированию на рассмотрение наиболее важных предикторов, отбрасывая или игнорируя наименее важные. Обратите внимание на то, что для некоторых моделей вычисление важности предикторов - это длительный процесс, особенно при работе с большими наборами данных, и в результате по умолчанию для некоторых моделей эта опция будет выключена. Важность предикторов недоступна для моделей списка решений. Дополнительную информацию смотрите в разделе “Важность предиктора” на стр. 42.

Оценки склонностей

Оценки склонностей можно включить на узле моделирования и на вкладке Параметры слепка модели. Эта функциональная возможность доступна только при выборе поля назначения флагового типа. Дополнительную информацию смотрите в разделе “Оценки склонностей” на стр. 35.

Вычислить простые оценки склонности. Простые оценки склонности получаются из модели на основе только обучающих данных. Если модель предсказывает значение *true* (будет отклик), склонность совпадает с P, где P - это вероятность предсказания. Если модель предсказывает значение false, склонность вычисляется как (1 – P).

- При выборе этой опции при построении модели оценки склонности будут включены в слепок модели по умолчанию. Однако вы всегда можете включить простые оценки склонности в слепке модели независимо от выбора их на узле моделирования.
- При скоринге модели простые оценки склонности будут добавлены в поле с буквами *RP*, присоединенными к стандартному префиксу. Например, если предсказания содержатся в поле с именем *$R-churn*, именем поля оценки склонности будет *$RRP-churn*.

Вычислить скорректированные оценки склонности. Простые оценки склонности основаны исключительно на оценках модели, которая может быть переобучена, что приведет к чрезмерно оптимистическим оценкам

склонности. Скорректированные оценки вносят компенсацию, изучая выполнение модели на испытательном и проверочном разделах и уточняя склонности для улучшения в соответствии с этим оценки.

- Для этого раздела требуется, чтобы в потоке присутствовало допустимое поле раздела.
- В отличие от простых оценок достоверности, скорректированные оценки склонностей нужно вычислять при построении модели; в противном случае они будут недоступны при скоринге слепка модели.
- При скоринге модели скорректированные оценки склонности будут добавлены в поле с буквами *AP*, присоединенными к стандартному префиксу. Например, если предсказания содержатся в поле с именем *$R-churn*, именем поля оценки склонности будет *$RAP-churn*. Скорректированные оценки склонности недоступны для моделей логистической регрессии.
- При вычислении скорректированных оценок склонности испытательный или проверочный раздел, используемый для вычисления, не должен быть сбалансирован. Чтобы исключить это, убедитесь, что выбрана опция **Только сбалансированные данные обучения** на любом вышележащем узле Баланс. Кроме этого, если на вышележащем уровне взята сложная выборка, это может сделать скорректированные оценки склонностей неприемлемыми.
- Скорректированные оценки склонности недоступны для моделей моделей дерева или набора правил с "бустингом". Дополнительную информацию смотрите в разделе “Усиленные модели C5.0” на стр. 111.

Основываясь на. Для вычисления скорректированных оценок склонности в потоке должно присутствовать поле раздела. Вы должны указать, какой из разделов использовать для этих вычислений - проверочный или испытательный. Для получения наилучших результатов испытательный или проверочный раздел должен включать в себя по крайней мере столько записей, сколько и раздел, использованный для обучения исходной модели.

Узел C5.0

Примечание: Эта возможность доступна в SPSS Modeler Professional и SPSS Modeler Premium.

Этот узел использует алгоритм C5.0 для построения или **дерева решений**, или **набора правил**. Модель C5.0 работает, расщепляя выборку на основе значения в поле, обеспечивающего максимальный **информационный выигрыш**. Каждая подвыборка, определенная первым расщеплением, затем расщепляется снова, обычно на основании другого поля, и этот процесс продолжается, пока на очередном шаге расщепление подвыборки станет невозможным. На последнем шаге проверяются повторно расщепления самого низкого уровне, и те из них, которые не вносят существенного вклада в качество модели, удаляются (**отсекаются**).

Примечание: Узел 5.0 может предсказывать только категориальные поля назначения. При анализе данных с использованием категориальных полей (номинальных или порядковых) этот узел с большей вероятностью совместно сгруппирует категории, чем версии C5.0 до выпуска 11.0.

C5.0 может порождать модели двух типов. **Дерево решений** - это непосредственное описание разделений, обнаруженных алгоритмом. Каждый конечный узел ("лист") описывает конкретное подмножество обучающих данных, а каждое наблюдение в обучающих данных принадлежит ровно одному конечному узлу дерева. Другими словами, для любой конкретной записи данных, представленной в дереве решений, возможно ровно одно предсказание.

Напротив, **набор правил** пытается сделать несколько предсказаний для индивидуальных записей. Наборы правил получаются из дерева решений и в некотором смысле представляют собой упрощенную или очищенную версию информации, найденной в дереве решений. Часто наборы правила могут сохранять большую часть важной информации от полного дерева решений, но с менее сложной моделью. Наборы правил работают иначе, чем деревья решений, и свойства у них другие. Наиболее важное отличие состоит в том, что при использовании набора правил к любой конкретной записи может применяться несколько правил или не применяться правил вовсе. Если применяется несколько правил, каждое правило получает взвешенный "голос" на основании показателя доверия, связанного с этим правилом, и конечное предсказание определяется объединением взвешенных голосов правил, которые были применены к рассматриваемой записи. Если никакие правила не применялись, записи присваивается предсказание по умолчанию.

Пример. Исследователь в области медицины собрал данные о наборе пациентов, пострадавших от одной болезни. Во время курса лечения каждый пациент принимал одно из пяти лекарств. Можно одновременно с другими узлами использовать модель C5.0 для определения, какое именно лекарство окажется полезным для будущего пациента с тем же заболеванием.

Требования. Для обучения модели C5.0 должно существовать одно категориальное (то есть номинальное или порядковое) поле *Назначения* и одно или несколько *Входных* полей любого типа. Поля с заданными значениями *Оба* или *Нет* игнорируются. У используемых в модели полей должны быть полностью конкретизированы типы. Можно задать также поле веса.

Достоинства. Модели C5.0 довольно устойчивы к таким проблемам, как пропуск данных или большое количество входных полей. Обычно этим моделям для оценки не требуется большое время на обучение. Кроме этого, модели C5.0 проще для понимания по сравнению с некоторыми другими моделями, так как у получаемых из этой модели правил есть непосредственная интерпретация. Модель C5.0 предлагает также мощную возможность **бустинга** для повышения точности классификации.

Примечание: Скорость построения моделей C5.0 может существенно увеличиться при включении параллельной обработки.

Опции моделей узла C5.0

Имя модели. Задайте имя модели, которая будет создана.

- **Авто.** При выборе этой опции имя модели будет сгенерировано автоматически на основании имени или имен полей назначения. Это опция по умолчанию.
- **Пользовательское.** Выберите эту опцию, чтобы самому задать имя для слепка модели, который будет создан этим узлом.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Создание моделей расщепления. Строит отдельную модель для каждого возможного значения входных полей, заданных как поля расщепления. Дополнительную информацию смотрите в разделе “Построение моделей расщепления” на стр. 28.

Тип вывода. Укажите здесь, каким вы хотите видеть полученный слепок модели - **Деревом решений** или **Набором правил**.

Символика групп. При выборе этой опции C5.0 попытается скомбинировать символические значения, у которых есть одинаковые структуры по отношению к выходному полю. Если эта опция не выбрана, C5.0 создаст дочерний узел для каждого значения символического поля, использованного для разделения родительского узла. Например, если C5.0 проводит разделение для узла *ЦВЕТ* (со значениями *КРАСНЫЙ*, *ЗЕЛЕНЫЙ* и *СИНИЙ*), по умолчанию будет создано разделение на три группы. Однако при выборе этой опции, если записи со значением *ЦВЕТ = КРАСНЫЙ* очень похожи на записи с *ЦВЕТ = СИНИЙ*, будет создано разделение на две группы, где в одной группе будут записи со значением *ЗЕЛЕНЫЙ*, а в другой - совместно записи со значением *СИНИЙ* и *КРАСНЫЙ*.

Использовать бустинг. У алгоритма C5.0 есть специальная возможность повышения степени точности, называемая **бустинг**. Она работает, последовательно создавая несколько моделей. Первая модель строится обычным образом. Затем при построении второй модели особое внимание уделяется тем записям, которые были неправильно классифицированы первой моделью. Третья модель фокусируется на ошибках второй модели и так далее. В конечном итоге наблюдения классифицируются с применением к ним всего набора моделей и с использованием взвешенной процедуры голосования, чтобы объединить отдельные предсказания в одно общее. Бустинг может существенно повысить точность модели C5.0, но он требует также большего времени на обучение. Опция **Количество попыток** позволяет управлять тем, как много

моделей будет использовано для усиленной модели (с бустингом). Эта возможность основана на исследовании Фройда и Шапиро с некоторыми собственными улучшениями, относящимися к лучшей обработке зашумленных данных.

Перекрестная проверка. Если выбрана эта опция, C5.0 будет использовать набор моделей, построенных на подмножествах обучающих данных, чтобы оценить точность модели, построенной на полном наборе данных. Это полезно, когда ваш набор данных чересчур мал для традиционного разделения на обучающий набор и набор тестирования. Модели с перекрестной проверкой отбрасываются после вычисления оценки точности. Вы можете задать **количество слоев**, то есть число моделей, используемых для перекрестной проверки. Обратите внимание на то, что в предыдущих версиях IBM SPSS Modeler построение модели и перекрестная проверка были двумя отдельными операциями. В текущей версии отдельный шаг построения модели не требуется. Построение модели и перекрестная проверка выполняются одновременно.

Режим. Для **Простого** обучения большинство параметров C5.0 задаются автоматически. **Дополнительное** обучение позволяет более явно управлять параметрами обучения.

Опции простого режима

Предпочтение. По умолчанию C5.0 будет пытаться произвести наиболее точное возможное дерево. В некоторых экземплярах это может привести к переобучению, что проявляется в виде низкой производительности при применении модели к новым данным. Выберите опцию **Обобщение**, чтобы использовать параметры алгоритма, менее восприимчивые к этой проблеме.

Примечание: Для моделей, построенных с выбранной опцией **Обобщение**, не гарантируется, что они обеспечат лучшее обобщение, чем другие модели. Если обобщение - это критическая проблема, всегда проверяйте ваши модели на имеющейся тестовой выборке.

Ожидаемый шум (%). Задайте ожидаемую долю шума или ошибочных данных в обучающем наборе.

Опции дополнительного режима

Серьезность сокращения. Определяет, насколько будут усекаться дерево решений или набор правил. Увеличьте это значение для получения меньшего, более сжатого дерева. Уменьшите это значение для получения более точного дерева. Этот параметр влияет только на локальное усечение (смотрите тему "Использовать глобальное усечение" далее).

Минимально число записей на дочернюю ветвь. Размер подгрупп можно использовать для ограничения количества разделений каждой ветви в дереве. Ветвь дерева будет разделяться только в том случае, если не менее двух результирующих подветвей будут содержать по крайней мере указанное количество записей из обучающего набора. Значение по умолчанию равно 2. Увеличьте это значение для предотвращения **переобучения** при наличии зашумленных данных.

Использовать глобальное сокращение. Деревья усекаются в две стадии. Во-первых, на стадии локального усечения, когда проверяются поддеревья и ветви сворачиваются для повышения точности модели. Во-вторых, на стадии глобального усечения дерево рассматривается в целом, и слабые ветви могут сворачиваться. Глобальное усечение выполняется по умолчанию. Чтобы пропустить стадию глобального сокращения, отмените выбор этой опции.

Атрибуты отсеивания. При выборе этой опции C5.0 до начала построения модели проверит полезность предикторов. Предикторы, которые окажутся при проверке нерелевантными, будут исключены из процесса построения модели. Эта опция может быть полезна для моделей с большим количеством предикторных полей и для предотвращения переобучения.

Примечание: Скорость построения моделей C5.0 может существенно увеличиться при включении параллельной обработки.

Слепки моделей деревьев решений

Слепки моделей дерева решений представляют структуры деревьев для предсказания конкретного выходного поля, открытые одним из узлов моделирования дерева решений (дерево C&R, CHAID, QUEST или C5.0).). Модели деревьев можно генерировать непосредственно из узла построения деревьев, а можно воспользоваться интерактивным построителем деревьев. Дополнительную информацию смотрите в разделе “Интерактивный построитель деревьев” на стр. 82.

Оценка моделей деревьев

При выполнении потока, содержащего слепок модели дерева, результат зависит от типа дерева.

- Для деревьев классификации (с категориальным полем назначения) в каждую запись данных добавляются два новых поля, содержащих предсказанное значение и достоверность. Предсказание основано на категории для конечного узла, которой чаще всего назначается запись; если большинство респондентов на данном узле отвечали *Да*, предсказание для всех записей, назначенных этому узлу, будет Да.
- Для деревьев регрессии генерируются только предсказанные значения; достоверность не назначается.
- Есть возможность для моделей CHAID, QUEST и деревьев C&R добавить новое поле, содержащее ID для узла, которому назначена запись.

Имена новых полей получаются из имени модели путем добавления префиксов. Для деревьев C&R, CHAID и QUEST префикс *$R-* добавляется к полю прогноза, *$RC-* к полю достоверности и *$RI-* к полю идентификатора узла. В случае деревьев C5.0 префикс *$C-* добавляется к полю прогноза и *$CC-* к полю достоверности. В потоке с несколькими узлами моделей деревьев *префиксы* в именах новых полей при необходимости содержат номер, отличающий данное поле от других - например, *$R1-*, *$RC1-* и *$R2-*.

Работа со слепками моделей деревьев

Есть ряд способов сохранить или экспортировать информацию, связанную с моделью.

Примечание: Многие из этих опций доступны также в окне построителя деревьев.

В построителе деревьев или в слепке моделей деревьев вы можете:

- Сгенерировать узел Фильтр или узел Выбор на основе текущего дерева. Дополнительную информацию смотрите в разделе “Генерирование узлов фильтра и выбора” на стр. 92.
- Сгенерировать слепок набора правил, представляющий структуру дерева как набор правил, которые определяют конечные ветви дерева. Дополнительную информацию смотрите в разделе “Генерирование набора правил из дерева решений” на стр. 93.
- Коме этого, только для слепков моделей деревьев, вы можете экспортировать модель в формате PMML. Дополнительную информацию смотрите в разделе “Палитра моделей” на стр. 40. Если модель содержит пользовательские расщепления, они не сохраняются в экспортируемом PMML. (Вернее, само расщепление сохраняется, но теряется информация, что оно пользовательское, а не выбранное алгоритмом).
- Сгенерировать диаграмму на основании выбранной части текущего дерева. *Примечание*: это работает только для слепков, присоединенных к другим узлам в потоке. Дополнительную информацию смотрите в разделе “Генерирование графиков” на стр. 111.
- Только для моделей C5.0 с бустингом можно выбрать **Единичное дерево решений (холст)** или **Единичное дерево решений (палитра GM)**, чтобы создать новый единый набор правил, полученный из текущего выбранного правила. Дополнительную информацию смотрите в разделе “Усиленные модели C5.0” на стр. 111.

Примечание: Хотя узел построения правил теперь заменен узлом дерева C&R, узлы деревьев решений в существующих потоках, первоначально созданные при помощи узла построения правил, все равно будут работать правильно.

Слепки моделей одного дерева

Если в качестве главной цели на узле моделирования выбрать **Построить единичное дерево**, полученный слепок модели будет содержать следующие вкладки.

Таблица 7. Вкладки в слепке единичного дерева

клавиша Tab	Описание	Дополнительная информация
Модель	Содержит правила, определяющие модель.	Дополнительную информацию смотрите в разделе “Правила модели дерева решений”.
Средство просмотра	Содержит представление дерева модели.	Дополнительную информацию смотрите в разделе “Средство просмотра модели дерева решений” на стр. 110.
Итог	Содержит информацию о полях, настройках построения и процессе оценки моделей.	Дополнительную информацию смотрите в разделе “Сводка слепков моделей / Информация” на стр. 42.
Параметры	Дает возможность задать опции достоверности и генерирования SQL при оценке модели.	Дополнительную информацию смотрите в разделе “Параметры слепков моделей набора правил и дерева решений” на стр. 110.
Аннотация	Дает возможность добавить описательные аннотации, задать пользовательское имя, добавить текст подсказки и задать ключевые слова для модели.	

Правила модели дерева решений

На вкладке Модель для слепка дерева решений выводятся определяющие модель правила. Дополнительно можно вывести график важности предикторов и третью панель с информацией о хронологии, частотах и суррогатах.

Примечание: Если выбрать опцию **Создать модель для очень больших наборов данных** на вкладке Опции сборки узла CHAID (панель Цель), на вкладке Модель выводятся только подробности правил дерева.

Правила дерева

На левой панели выводится список условий, определяющих разделение данных, обнаруженное алгоритмом, особенно ряд правил, которые можно использовать для назначения отдельных записей дочерним узлам на основе значений различных предикторов.

Деревья решений работают, рекурсивно разделяя данные на основе значений входных полей. Разделы данных называются **ветвями**. Начальная ветвь (иногда называемая **корень**) включает в себя записи всех данных. Корень расщепляется на подмножества, или **дочерние ветви**, на основании значения конкретного входного поля. Каждая дочерняя ветвь может и дальше расщепляться на подветви, которые в свою очередь делятся дальше и так далее. На самом нижнем уровне дерева находятся ветви, у которых больше нет разделений. Такие ветви называют **конечными ветвями** (или **листьями**).

Подробности правил деревьев

Браузер правил показывает входные значения, определяющие все разделения или ветви, и сводку значений выходных полей для записей в данном разделении. За общей информацией по браузеру моделей обратитесь к разделу “Просмотр слепков моделей” на стр. 41.

Для разделений на основе численных полей ветвь выводится как строка следующего вида:

```
имя_поля отношение значение [сводка]
```

где *отношение* - это численное отношение. Например, ветвь, определяемая значением больше 100 для поля *прибыль* может быть показана в следующем виде:

```
прибыль > 100 [сводка]
```

Для разделений на основе символических полей ветвь показывается строкой следующего вида:

```
имя_поля = значение [сводка] или имя_поля в [значения] [сводка]
```

где *значения* - это значения поля, определяющие ветвь. Например, ветвь, включающая в себя записи, где значениями поля *регион* могут быть *Север*, *Запад* или *Юг*, может быть представлена следующим образом:

```
регион в ["Север" "Запад" "Юг"] [сводка]
```

Для конечных ветвей предсказание также дается, и добавляется стрелка и предсказанное значение к концу условия правила. Например, конечная ветвь, определенная условием *прибыль* > 100, которое предсказывает значение *высокая* для выходного поля, может быть представлена следующим образом:

```
прибыль > 100 [Режим: высокая] ► высокая
```

Сводка для ветви определяется по-разному для символических и числовых выходных полей. Для деревьев с числовыми выходными полями сводка - это **среднее** значение для ветви, а **эффект** ветви - это разность между средним по ветви и средним по ее родительской ветви. Для деревьев с символическими выходными полями сводка - это **мода**, или наиболее частое значение, для записей в ветви.

Для полного описания ветви нужно включить условие, определяющее ветвь, а также условия, определяющее дальнейшие расщепления дерева. Например, в дереве:

```
revenue > 100
  region = "Север"
  region in ["Юг" "Восток" "Запад"]
    revenue <= 200
```

ветвь, представленная второй строкой, определяется условиями *revenue > 100* и *region = "Север"*.

Если нажать кнопку **Показать экземпляры/показатель доверия** на панели инструментов, для каждого правила будет показана также информация о числе записей, к которым применяется правило (**Экземпляры**) и о доле этих записей, для которых значение правила равно true (**Показатель доверия**).

Важность предиктора

Диаграмма, обозначающая относительную важность каждого предиктора в оцениваемой модели, может быть дополнительно также показана на вкладке Модель. Обычно вам потребуется направить усилия по моделированию на рассмотрение наиболее важных предикторов, отбрасывая или игнорируя самые не важные. Обратите внимание на то, что эта диаграмма доступна только в том случае, если перед генерированием модели на вкладке Анализ выбрана опция **Вычислять важность предикторов**. Дополнительную информацию смотрите в разделе “Важность предиктора” на стр. 42.

Дополнительная информация о модели

Если нажать кнопку **Показать панель дополнительной информации** на панели инструментов, вы увидите наверху страницы панель, показывающую подробную информацию для выбранного правила. Информационная панель содержит три вкладки.

Хронология. На этой вкладке отслеживаются условия разделения от корня вплоть до выбранного узла. Это предоставляет список условий, определяющих, когда запись назначается выбранному узлу. Записи, для которых выполняются все условия, назначаются этому узлу.

Частоты. Для моделей с символическими полями назначения на этой вкладке для каждого возможного значения назначения выводится, сколько записей назначено этому узлу (в обучающих данных) с нужным значением назначения. Выводится также график частот, выраженных в процентной доле (до трех десятичных знаков). Для моделей с числовыми полями назначения эта вкладка пустая.

Суррогаты. Там, где они применимы, все суррогаты для поля первичного расщепления показываются для выбранного узла. Суррогаты - это альтернативные поля, используемые, если первичное значение предиктора отсутствует для данной записи. Максимальное количество суррогатов, разрешенных для данного расщепления, задается на узле построения дерева, но их фактическое количество зависит от обучающих данных. В общем случае, чем больше данных пропущено, тем больше возможных суррогатов будет использовано. Для других моделей дерева решений эта вкладка пуста.

Примечание: для включения в модель суррогаты должны быть определены на фазе обучения. Если в обучающей выборке нет пропущенных значений, никакие суррогаты определены не будут, и все записи с пропущенными значениями, встреченные при проверке или скоринге, автоматически попадут в дочерний узел с максимальным числом записей. Если предполагается, что при проверке и скоринге будут встречаться пропущенные значения, убедитесь, что и в обучающей выборке есть пропущенные значения. Для деревьев CHAID суррогаты недоступны.

Средство просмотра модели дерева решений

Вкладка Средство просмотра для слепка модели дерева решений похожа на вывод в построителе дерева. Главное отличие в том, что при просмотре слепка модели вы не можете увеличивать или изменять дерево. Другие опции для просмотра и настройки вывода у обоих компонентов аналогичны. Дополнительную информацию смотрите в разделе “Настройка просмотра дерева” на стр. 85.

Примечание: Вкладка Средство просмотра не выводится в конструкции слепков моделей CHAID, если выбрана опция **Создать модель для очень больших наборов данных** на вкладке Опции построения (панель Цель).

При просмотре правил расщепления на вкладке Средство просмотра квадратные скобки означают, что граничное значение включено в диапазон, а круглые скобки указывают на то, что граничное значение исключается из диапазона. Поэтому выражение (23,37] соответствует диапазону от 23 исключительно до 37 включительно, т.е. строго больше 23 до 37. На вкладке Модель то же условие будет выведено следующим образом:

```
Age > 23 and Age <= 37
```

Параметры слепков моделей набора правил и дерева решений

На вкладке Параметры для слепка модели дерева решений или набора правил при скоринге модели можно задать опции для доверительных показателей и для генерирования SQL. Эта вкладка доступна только после добавления слепка модели в поток.

Вычислить показатели доверия. Выберите для включения доверительных показателей в операции скоринга. При скоринге моделей в базе данных исключение доверительных показателей позволяет генерировать более эффективный SQL. Для деревьев регрессии доверительные показатели не назначаются.

Примечание: Если выбрать опцию **Создать модель для очень больших наборов данных** на вкладке Опции построения (панель Способ для моделей узла CHAID), этот переключатель доступен только в слепках моделей с номинальными или флаговыми полями назначения.

Вычислить простые оценки склонности. Для моделей с флаговым полем назначения (которое возвращает предсказание Да или Нет) можно затребовать оценку склонности, которая обозначает правдоподобие правильного выходного значения, заданного для поля назначения. Это дополнение к другим предсказанным и доверительным значениям, которые можно сгенерировать при скоринге.

Примечание: Если выбрать опцию **Создать модель для очень больших наборов данных** на вкладке Опции построения (панель Способ для моделей узла CHAID), этот переключатель доступен только в слепках моделей с категориальными флаговыми полями назначения.

Вычислить скорректированные оценки склонности. Простые оценки склонности основываются только на данных обучения и могут выглядеть чрезмерно оптимистическими из-за тенденции многих моделей чрезмерно точно подгонять эти данные под модель. Корректировка склонностей направлена на

компенсацию этого эффекта путем проверки выполнения модели на испытательном и проверочном разделах. Для этой опции требуется, чтобы поле разделения было определено в потоке и скорректированные оценки склонности были включены на узле моделирования до генерирования модели.

Примечание: Скорректированные оценки склонности недоступны для моделей дерева или набора правил с "бустингом". Дополнительную информацию смотрите в разделе “Усиленные модели C5.0”.

Идентификатор правила. Для моделей CHAID, QUEST и дерева C&R эта опция добавляет поле в выводе данных скоринга, обозначающее ID для конечного узла, которому назначена каждая запись.

Примечание: При выборе этой опции генерирование SQL недоступно.

Сгенерировать SQL для данной модели. Когда используются данные, хранящиеся в базе данных, код SQL можно передавать обратно в базу данных для выполнения, что обеспечивает высокую производительность многих операций.

Выберите одну из следующих опций для указания, как выполняется генерирование SQL.

- **По умолчанию: Выполнять оценку при помощи адаптера скоринга сервера, если он установлен; в противном случае - в процессе.** При соединении с базой данных с установленным адаптером скоринга генерирует SQL при помощи этого адаптера, в противном случае генерирует SQL собственными средствами в SPSS Modeler.
- **Поддержка генерирования без пропущенных значений.** Выберите эту опцию для включения генерирования SQL без издержек на обработку пропущенных значений. При выборе этой опции для предсказания просто задается пустое значение (`$null$`), когда при скоринге наблюдения встречается пропущенное значение.

  Примечание: Эта опция недоступна для моделей CHAID. Из других типов моделей она доступна только для деревьев решений (не для наборов правил).
- **Поддержка генерирования с пропущенными значениями.** Для моделей CHAID, QUEST и дерева C&R эта опция включает полную поддержку генерирования SQL при наличии пропущенных значений. Это означает, что при генерировании SQL пропущенные значения обрабатываются, как это задано в модели. Например, дерево C&R использует правила суррогатов и максимальную дочернюю нейтрализацию.

  Примечание: Для моделей C5.0 эта опция доступна только для наборов правил (не для деревьев решений).

Усиленные модели C5.0

Примечание: Эта возможность доступна в SPSS Modeler Professional и SPSS Modeler Premium.

При создании усиленной модели C5.0 (или набора правил, или дерева решений) вы на самом деле создаете набор связанных моделей. Браузер правил модели для усиленной модели C5.0 показывает список моделей на верхнем уровне иерархии вместе с предполагаемой точностью каждой модели и общей точностью ансамбля усиленных моделей. Чтобы проверить правила или разделения для конкретной модели, выберите эту модель и раскройте ее для получения правила или ветви отдельной модели.

Вы можете выделить также конкретную модель из набора усиленных моделей и создать слепок модели с новым набором правил, содержащий только данную модель. Для создания нового набора правил из усиленной модели C5.0 выберите интересующий вас набор или дерево правил и в меню Генерировать выберите **Одно дерево решений (палитра GM)** или **Одно дерево решений (холст)**.

Генерирование графиков

Узлы дерева предоставляют много информации, однако у нее не всегда такой формат, который легко понятен пользователям из бизнеса. Чтобы представить данные в виде, в котором они легко могут быть вставлены в деловые отчеты, презентации и т.д., можно построить диаграммы для выбранных данных. Например, на вкладках Модель или Программа просмотра слепка модели или на вкладке Программа просмотра интерактивного дерева можно сгенерировать диаграмму (график) для выбранной части дерева, то есть создать диаграмму только для наблюдений в выбранном узле дерева или ветви.

Примечание: Создать диаграмму из слепка можно только в том случае, когда он присоединен к другим узлам в потоке.

Постройте диаграмму

Первый шаг - это выбор информации, которая будет показана на диаграмме (графике):

- На вкладке Модель слепка разверните список условий и правил на левой панели и выберите нужную вам позицию.
- На вкладке слепка Программа просмотра разверните список ветвей и выберите нужный вам узел.
- На вкладке интерактивного дерева Программа просмотра разверните список ветвей и выберите нужный вам узел.

  Примечание: На вкладках Программа просмотра нельзя выбирать главный узел.

Способ создания диаграммы одинаков и не зависит от того, как вы выбрали данные для представления:

1. В меню Генерировать выберите раздел **Диаграмма (из выбранного)**; как вариант, на вкладке Программа просмотра нажмите кнопку **Диаграмма (из выбранного)** в левом нижнем углу. Откроется вкладка Основная панель диаграмм.

   Примечание: при выводе панели диаграмм таким способом возможны только вкладки Основная и Подробная.
2. Используя параметры вкладки Тип и Детали, задайте элементы для вывода на диаграмме.
3. Нажмите кнопку OK, чтобы построить диаграмму.

В заголовке диаграммы будут указаны узлы или правила, выбранные для включения.

Слепки моделей для бустинга, бэггинга и очень больших наборов данных

Если в качестве главной цели на узле моделирования выбрать **Повысить точность модели (бустинг)**, **Повысить стабильность модели (бэггинг)** или **Создать модель для очень большого набора данных**, то IBM SPSS Modeler построит ансамбль из нескольких моделей. Дополнительную информацию смотрите в разделе “Модели для ансамблей” на стр. 44.

Полученный слепок модели содержит перечисленные ниже вкладки. Вкладка Модель поддерживает ряд различных представлений модели.

Таблица 8. Вкладки, доступные в слепке модели

клавиша Tab	Просмотр	Описание	Дополнительная информация
Модель	Сводка для модели	Содержит сводку о качестве ансамбля и (кроме моделей с бустингом и непрерывных полей назначения) разнородности, показатель различия между предсказаниями по разным моделям.	Дополнительную информацию смотрите в разделе “Сводка для модели” на стр. 44.
	Важность предиктора	Содержит диаграмму, показывающую относительную важность каждого предиктора (входного поля) в оценке модели.	Дополнительную информацию смотрите в разделе “Важность предикторов” на стр. 45.
	Частота встречаемости предиктора	Содержит диаграмму, показывающую относительную частоту использования каждого предиктора в наборе моделей.	Дополнительную информацию смотрите в разделе “Частота предикторов” на стр. 45.

Таблица 8. Вкладки, доступные в слепке модели (продолжение)

клавиша Tab	Просмотр	Описание	Дополнительная информация
	Точность моделей компонентов	Отрисовывает диаграмму точности предсказания по каждой модели ансамбля.	
	Детали моделей компонентов	Содержит информацию по каждой модели в ансамбле.	Дополнительную информацию смотрите в разделе “Подробности о моделях компонентов” на стр. 46.
	Информация	Содержит информацию о полях, настройках построения и процессе оценки моделей.	Дополнительную информацию смотрите в разделе “Сводка слепков моделей / Информация” на стр. 42.
Параметры		Дает возможность включить в операции оценок доверительные интервалы.	Дополнительную информацию смотрите в разделе “Параметры слепков моделей набора правил и дерева решений” на стр. 110.
Аннотация		Дает возможность добавить описательные аннотации, задать пользовательское имя, добавить текст подсказки и задать ключевые слова для модели.	

Слепки моделей правил связывания

Слепок модели Набор правил представляет правила для предсказания конкретного выходного поля, открытые узлом моделирования правил связывания (Apriori) или одним из узлов, строящих деревья (C&R Tree, CHAID, QUEST или C5.0). В случае правил связывания набор правил должен генерироваться из неуточненного слепка правил. В случае деревьев набор правил должен генерироваться из построителя деревьев, узла построения моделей C5.0 или слепка любой модели дерева. В отличие от неуточненных слепков, слепки набора правил можно ставить в потоки для генерирования прогнозов.

При выполнении потока, содержащего слепок набора правил, в поток в каждую запись данных добавляются два новых поля, содержащих предсказанное значение и достоверность. Имена новых полей получаются из имени модели путем добавления префиксов. В случае наборов правил связывания используются префиксы *$A-* для поля прогноза и *$AC-* для поля достоверности. В случае наборов правил C5.0 используются префиксы *$C-* для поля прогноза и *$CC-* для поля достоверности. В случае наборов правил C&R Tree используются префиксы *$R-* для поля прогноза и *$RC-* для поля достоверности. В потоке с несколькими слепками наборов правил, последовательно предсказывающими одно и то же выходное поле (или несколько выходных полей) *префиксы* в именах новых полей содержат номер, отличающий данное поле от других. Первый в потоке слепок набора правил использует обычные имена, после второго узла имена будут начинаться с *$A1-* и *$AC1-*, третий узел создаст имена, начинающиеся с *$A2-* и *$AC2-*, и так далее.

Как применяются правила. В отличие от остальных слепков моделей, если набор правил генерируется из правил связывания, для каждой конкретной записи может генерироваться несколько предсказаний, которые могут противоречить друг другу. Есть два метода сгенерировать предсказания из наборов правил.

Примечание: Если наборы правил генерируются из деревьев решений, то результат не зависит от метода, поскольку из дерева получаются взаимоисключающие правила.

- **Голосование.** Этот метод пытается объединить прогнозы всех правил, применимых к данной записи. Для каждой записи перебираются все правила, и из каждого применимого к данной записи правила генерируется предсказание и соответствующая достоверность. Для каждого выходного значения вычисляется сумма значений достоверности, после чего в качестве итогового предсказания выбирается

выходное значение с наибольшей суммой значений достоверности. В качестве достоверности итогового предсказания берется сумма значений достоверности для выбранного выходного значения, поделенная на число правил, сработавших для данной записи.

- **Первое совпадение.** Этот метод просто перебирает правила по порядку, и предсказание генерируется по первому правилу, которое окажется применимым к данной записи.

Использованием того или иного метода можно управлять в опциях потока.

Генерирование узлов. При помощи меню Создать вы можете создавать новые узлы на основе набора правил.

- **Узел фильтра.** Создает новый узел фильтра, чтобы отфильтровать поля, которые не используются правилами в наборе правил.
- **Узел выбора.** Создает новый узел выбора для выбора записей, к которым применимо выбранное правило. Сгенерированный узел выберет записи, для которых все антецеденты правила истинны. Для этой опции требуется выбрать правило.
- **Узел трассировки правил.** Создает новый надузел для вычисления поля, показывающего, по какому правилу сделан прогноз для той или иной записи. Если набор правил оценивается методом первого совпадения, это просто указание на первое сработавшее правило. Если набор правил оценивается методом голосования, это более сложная строка, содержащая входную информацию механизма голосования.
- **Единичное дерево решений (холст) / Единичное дерево решений (палитра GM).** Создает новый единый слепок набора правил, полученный из текущего выбранного правила. Доступна только в моделях C5.0 **с повышением значимости**. Дополнительную информацию смотрите в разделе “Усиленные модели C5.0” на стр. 111.
- **Модель для палитры.** Возвращает модель на палитру моделей. Это полезно в ситуациях, когда коллега послал вам поток, содержащий модель, но не саму модель.

Примечание: В слепке набора правил вкладки Настройки и Сводка такие же, как в моделях дерева решений.

Rule Set Model Tab

Вкладка Модель слепка набора правил содержит список правил, извлеченных из данных алгоритмом.

Правила разбиты по консеквенту (предсказанная категория) и представлены в следующем формате:

```
if антецедент_1
and антецедент_2
...
and антецедент_n
then предсказанное значение
```

где консеквент и антецедент_1 - антецедент_n представляют собой условия. Правило понимается так: "для записей, где истинны все условия антецедент_1 - антецедент_n, вероятно, что истинен консеквент." Если на панели инструментов нажать кнопку **Показать экземпляры/достоверность**, для каждого правила будет также показано число записей, к которым применимо данное правило, то есть для которых истинны антецеденты, (**Экземпляры**) и доля тех записей, для которых истинно правило в целом (**Достоверность**).

Учтите, что для наборов правил C5.0 достоверность вычисляется несколько иначе. C5.0 использует для вычисления достоверности правила следующую формулу:

```
(1 + число записей, где правило верно)
/
(2 + число записей, где истинны антецеденты правила)
```

Такой способ оценивать достоверность настроен на процесс генерирования правил из дерева решений (именно так создается набор правил в C5.0).

Импорт проектов из AnswerTree 3.0

IBM SPSS Modeler может импортировать проекты, сохраненные в AnswerTree 3.0 или or 3.1, используя диалоговое окно Файл > Открыть, как указано ниже:

1. Выберите в меню IBM SPSS Modeler:

   Файл > Открыть поток

2. Из выпадающего списка Файлы типа выберите **Файлы проекта AT (*.atp, *.ats)**.

   Каждый импортированный проект преобразуется в поток IBM SPSS Modeler со следующими узлами:

   - Один исходный узел, определяющий используемый источник данных (например, файл данных IBM SPSS Statistics или источник базы данных).
   - Для каждого дерева в проекте (их может быть несколько) создается один узел Тип, который определяет свойства для каждого поля (переменной), включая тип, роль (либо предиктор, то есть входное поле, либо предсказанное поле, то есть выходное), опции пропущенных значений и другие.
   - Для каждого дерева в проекте создается узел разделения, который разбивает данные для обучающей или контрольной выборки, и узел построения дерева, который определяет параметры генерирования дерева (это узел C&R Tree, QUEST или CHAID).

3. Чтобы просмотреть сгенерированные деревья, выполните поток.

Комментарии

- Деревья решений, сгенерированные в IBM SPSS Modeler, нельзя экспортировать в AnswerTree; импорт из AnswerTree в IBM SPSS Modeler необратим.
- Прибыли, определенные в AnswerTree, не сохраняются при импорте проекта в IBM SPSS Modeler.

Глава 7. Модели байесовских сетей

Узел Байесовская сеть

При помощи узла **Байесовская сеть** можно построить вероятностную модель, которая, опираясь и на наблюдаемые зарегистрированные свидетельства, и на практические соображениями здравого смысла, дает оценку вероятностей тех или иных исходов, привлекая атрибуты, которые на первый взгляд не имеют к этому отношения. Этот узел в основном работает с усиленными деревом наивными байесовскими сетями (Tree Augmented Naïve Bayes, TAN) и полными марковскими сетями, которые изначально используются для классификации.

Байесовские сети используются для прогнозирования в самых различных ситуациях; вот лишь некоторые примеры:

- Выбор адресата кредита с низким риском дефолта.
- Оценка времени, когда нужно выполнить техобслуживание оборудования, замену деталей или замену самого оборудования, с учетом показаний датчиков и существующих записей.
- Решение вопросов клиента через оперативные инструменты устранения неисправностей.
- Диагностика и устранение неисправностей в сетях сотовой телефонии в реальном времени.
- Оценка потенциальных рисков и выигрышей научно-исследовательских проектов с целью сосредоточить ресурсы на многообещающих направлениях.

Байесовская сеть - это графическая модель, в которой переменные из набора данных представлены **узлами**) графа, для которых имеет место вероятностная, или условная независимость друг от друга. Связи, или **дуги**, между узлами байесовской сети иногда, но не всегда отвечают причинно-следственным связям. Например, при помощи байесовской сети можно вычислить вероятность наличия у пациента некоторой болезни, зная о наличии или отсутствии определенных симптомов и имея другие важные сведения, если имеет место вероятностная независимость между показанными на графе симптомами и болезнью. Сети обладают весьма высокой устойчивостью к пропускам информации и дают наилучшие возможные предсказания исходя из наличных сведений.

Характерный пример байесовской сети построен Лауритценом и Шпигельхалтером в 1988. Он известен как модель "Азия" и представляет собой упрощенную версию сети для диагноза новых пациентов врача; связи можно считать ведущими из причины в следствие. Каждый узел представляет фасет, который может касаться особенностей пациента; например, "Курит" относится к подтвержденному курильщику, "ПосетилАзию" означает недавний визит в Азию. Вероятностные взаимосвязи показаны как соединения между узлами; так, курение повышает вероятность и бронхита, и рака легких, тогда как пожилой возраст связывается только с вероятностью рака легких. Аналогичным образом аномалии на флюорограмме легких могут вызываться либо туберкулезом, либо раком легких, тогда как шансы того, что пациент страдает от одышки (диспноэ) увеличиваются, если у него также бронхит или рак легких.

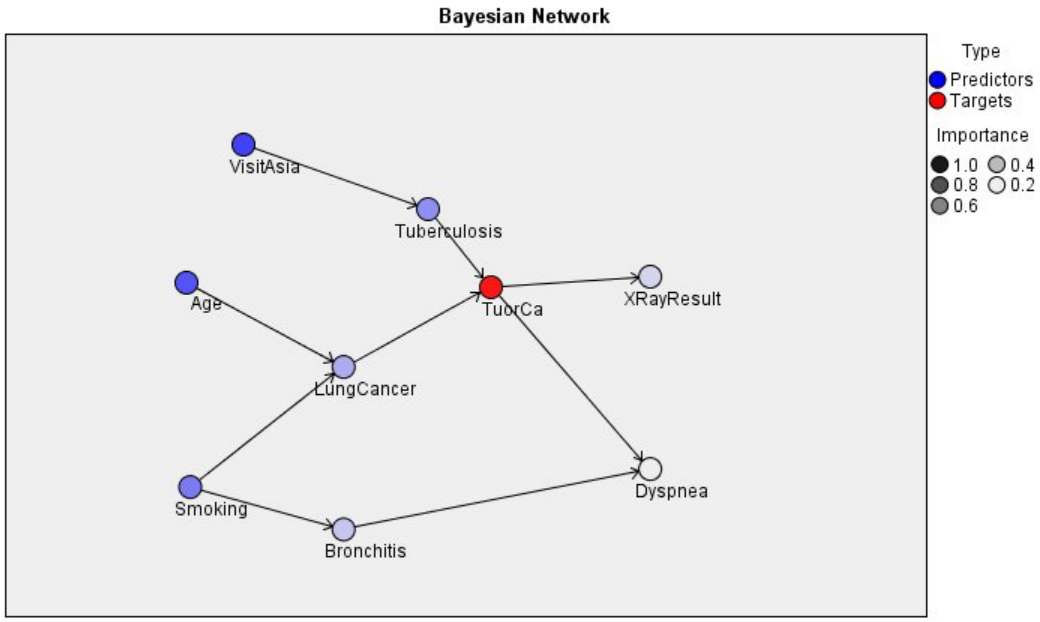


Рисунок 29. Пример сети "Азия" Лауритцена и Шпигельхалтера

Есть несколько причин использовать байесовскую сеть:

- Это помогает узнавать о причинно-следственных взаимосвязях. Благодаря этому у вас появляется возможность изучить проблемную область и предсказать последствия вмешательства.
- Сеть обеспечивает эффективный подход к предотвращению переобучения, или чрезмерной подгонки под данные.
- Сеть обеспечивает удобное, наглядное представление взаимосвязей.

Требования. Поля назначения должны быть категориальными, и их тип измерений может быть только *Номинальное*, *Порядковый номер* или *Флаг*. Входные поля могут быть любого типа. Непрерывные поля (поля числового диапазона) на входе автоматически разбиваются на поддиапазоны и преобразуются в категориальные; учтите, однако, что если распределение несимметричное, то лучшего результата можно добиться, если категоризовать поля вручную, добавив перед узлом Байесовская сеть узел Разделение на интервалы. Например, используйте узел Оптимальная категоризация, задав в качестве **управляющего** поля - поле **назначения** узла Байесовская сеть.

Пример. Аналитику банка нужна возможность прогнозировать, какие клиенты или потенциальные клиенты склонны к дефолту по погашению кредита. При помощи модели байесовской сети можно выявить характеристики клиентов, склонных к дефолту, и построить несколько различных типов моделей, чтобы выбрать из них наилучшего прогнозиста потенциальных неплательщиков.

Пример. Оператору телекоммуникационной сети нужно минимизировать число отказов от подписки (так называемое "отток клиентов") и ежемесячно обновлять модель с учетом данных истекшего месяца. При помощи модели байесовской сети можно выявить характеристики клиентов, склонных к оттоку, и продолжить обучение модели по новым данным каждый месяц.

Опции модели узла байесовской сети

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Построить модель для каждого разбиения. Строит отдельную модель для каждого возможного значения входных полей, заданных как поля расщепления. Дополнительную информацию смотрите в разделе “Построение моделей расщепления” на стр. 28.

Подмножества. При помощи этого поля можно указать поле для разделения данных на отдельные выборки для разных стадий построения модели - обучения, испытания и проверки. Сгенерировав модель при помощи одной выборки и испытав ее при помощи другой, можно получить надежный показатель качества обобщения модели на более крупные наборы данных, подобные текущим. Если при помощи узлов Тип или Раздел были определено несколько полей разделов, на вкладке Поля на каждом узле моделирования, где используется разделение, должно быть выбрано одно поле раздела. (Если представлен только один раздел, он будет использоваться автоматически при всяком включении разделения.) Кроме того, учтите, что для применения выбранного раздела в анализе на вкладке Параметры модели для узла должно быть также включено разделение. (Выключение этой опции делает возможным отключение разделения без изменения значений параметров полей.)

Разбиения. Для разбиения моделей выберите поле или поля разбиения. Это аналогично заданию для поля роли *Расщепление* на узле Тип. В качестве полей разбиения можно назначать только поля с типом измерения **Флаг**, **Номинальное**, **Порядковый номер** или **Непрерывное**. Поля, выбранные как поля разбиения, нельзя использовать в качестве полей назначения, разделов, частоты, веса или входных полей. Дополнительную информацию смотрите в разделе “Построение моделей расщепления” на стр. 28.

Продолжить обучение существующей модели. Если эта опция выбрана, результаты на вкладке Модель слепка модели обновляются при каждом запуске модели. Это полезно, например, при добавлении или изменении источника данных в существующей модели.

Примечание: Обновляться при этом может только существующая сеть; добавление или удаление узлов или соединений невозможно. Каждый раз, когда вы переобучаете модель, сеть будет иметь ту же конфигурацию; меняться могут только условные вероятности и важность предиктора. Это неважно, если новые данные в целом сходны со старыми и можно ожидать, что важными останутся те же самые поля; если же нужно проверить или изменить *состав важного* (а не просто степень важности того, что есть), нужно построить новую модель, то есть новую сеть

Тип структуры. Выберите структуру, которая будет использоваться при создании байесовской сети:

- **TAN.** Наивная байесовская модель приращиваемого дерева (Tree Augmented Naïve Bayes model, TAN) создает модель простой байесовской сети, представляющей собой усовершенствование стандартной наивной байесовской модели. Это достигается тем, что каждый предиктор может зависеть от другого предиктора, а не только от переменной назначения, что повышает точность классификации.
- **Марковское ограждение.** Выбирает в наборе данных следующий набор узлов: родители переменной назначения, дочерние узлы переменной назначения и родители дочерних узлов. По сути марковское ограждение охватывает все переменные сети, которые требуются для предсказания переменной назначения. Такой метод построения сети считается более точным, но при больших наборах данных за это приходится расплачиваться временем обработки, поскольку учитывается много переменных. Чтобы сократить объем обработки, можно на вкладке Дополнительно при помощи опций **Отбор показателей** выбрать переменные, которые существенно связаны с переменной назначения.

Включить шаг предварительной обработки для выбора возможностей. Включив этот переключатель, вы можете использовать опции **Отбор показателей** на вкладке Дополнительно.

Метод обучения параметров. Параметры байесовской сети - это условные вероятности для данного узла при известных значениях родительских узлов. Доступны два варианта оценки таблиц условной вероятности для узлов с известными значениями родительских узлов:

- **Максимум правдоподобия.** Включите этот переключатель, если используете большой набор данных. Это выбор по умолчанию.

- **Байесовская настройка для малого числа ячеек.** Для небольших наборов данных есть опасность переобучить модель, а также риск большого числа нулевых частот. Выберите эту опцию, чтобы преодолеть эти трудности путем сглаживания, уменьшающего влияние нулевых частот и ненадежных оценок.

Дополнительные опции узла байесовской сети

Дополнительные опции узла служат для точной настройки процесса построения моделей. Для доступа к экспертным опциям выберите режим **Дополнительно** на вкладке Дополнительно.

Пропущенные значения. По умолчанию IBM SPSS Modeler использует только те записи, которые содержат допустимые значения во всех полях, используемых в модели. (Эта стратегия также называется **исключение пропущенных значений целиком**.) Иногда, когда пропущенных значений много, данный подход отбрасывает непозволительно много записей, так что оставшихся данных недостаточно для генерирования хорошей модели. В таких случаях можно отменить выбор опции **Использовать только полные записи**. Тогда IBM SPSS Modeler пытается использовать для оценки модели максимум информации, включая те записи, где некоторые поля содержат пропущенные значения. (Эта стратегия также называется **попарное исключение** пропущенных значений.) Однако такое использование неполных записей иногда приводит к вычислительным проблемам при оценке моделей.

Добавить все вероятности. Задает, добавляются ли вероятности для каждой категории выходного поля к каждой записи, обрабатываемой узлом. Если эта опция не выбрана, добавляется только вероятность предсказанной категории.

Критерий независимости. Критерий независимости оценивает независимость друг от друга парных наблюдаемых значений двух переменных. Выберите тип проверки. Доступные варианты:

- **Отношение правдоподобия.** Проверяет независимость поля назначения от предиктора путем вычисления отношения между максимальными вероятностями результата при двух различных гипотезах.
- **Хи-квадрат Пирсона.** Проверяет независимость поля назначения от предиктора, используя нулевую гипотезу, согласно которой относительные частоты наступления наблюдаемых событий следуют заданному распределению частот.

Модели байесовской сети выполняют условные проверки независимости, в которых используются переменные помимо проверяемых пар. Кроме того, эти модели изучают не только взаимосвязи между полем назначения и предикторами, но и взаимосвязи между самими предикторами

Примечание: Опции проверки независимости доступны, только если на вкладке Модель выбрано **Включить шаг предварительной обработки для выбора возможностей** или **Тип структуры** марковского ограждения.

Уровень значимости. В сочетании с настройками критерия независимости дает возможность задать значение отсечения при выполнении проверок. Чем ниже это значение, тем меньше связей остается в сети; уровень по умолчанию - 0,01.

Примечание: Эта опция только доступна, только если на вкладке Модель выбрано **Включить шаг предварительной обработки для выбора возможностей** или **Тип структуры** марковского ограждения.

Максимальный размер набора настройки. Алгоритм создания структуры марковского ограждения проверяет независимость для при наборах условий все большего размера и удаляет из сети избыточные связи. Поскольку для обработки критериев с большим числом переменных условия требуется больше времени и памяти, число переменных можно ограничить. Это может быть особенно полезно при обработке данных с сильными зависимостями среди многих переменных. Имейте в виду, однако, что получающаяся сеть может содержать некоторые избыточные связи.

Укажите максимальное число переменных условия при проверке независимости. Значение по умолчанию - 5.

Примечание: Эта возможность доступна, только если на вкладке Модель выбрано **Включить шаг предварительной обработки для выбора возможностей** или **Тип структуры** марковского ограждения.

Выбор возможностей. При помощи этих опций можно ограничить число входных переменных при обработке модели, чтобы ускорить процесс построения модели. Это особенно полезно при создании структуры марковского ограждения, где число входных переменных может оказаться весьма велико; вы можете выбрать только те входные переменные, которые существенно связаны с переменной назначения.

Примечание: Опции выбора возможностей доступны, только если на вкладке Модель выбрано **Включить шаг предварительной обработки для выбора возможностей**.

- **Входные переменные, которые всегда выбраны** Открыв Средство выбора полей (кнопкой справа от текстового поля), выберите в наборе данных те поля, которые всегда будут использоваться при построении модели байесовской сети. Обратите внимание на то, что поле назначения всегда выбрано.
- **Максимальное число входных.** Задайте общее количество используемых входных полей из набора данных при построении модели байесовской сети. Наибольшее допустимое значение - общее количество входных полей в наборе данных.

  Примечание: Если в опции **Входные переменные, которые всегда выбраны** выбранных полей больше, чем **Максимальное число входных переменных**, выводится сообщение об ошибке.

Слепки моделей байесовской сети

Примечание: Если на узле моделирования на вкладке Модель выбрано **Продолжить обучение существующих параметров**, информация на вкладке Модель слепка модели обновляется каждый раз, когда модель генерируется заново.

Вкладка Модель слепок модели разбита на две панели:

Левая панель

Тип. Это представление содержит граф, изображающий взаимосвязи между узлом назначения и его важнейшими предикторами, а также взаимосвязи между предикторами. Важность каждого предиктора показана плотностью его цвета; сильный цвет отвечает важному предиктору и наоборот.

Если остановить указатель мыши над узлом, всплывает окно подсказки со значениями узлов поддиапазона.

Вы можете использовать инструменты диаграмм IBM SPSS Modeler, чтобы работать с графом, редактировать и сохранять его. Например, это можно делать для использования в других прикладных программах, таких как MS Word.

Совет: Если сеть содержит много узлов и для получения более наглядного графа нужно переместить узел, то щелкните по узлу и отбуксируйте его на новое место.

Распределение. Это представление содержит условные вероятности для каждого узла в сети в виде мини-диаграммы. Остановите указатель мыши на одной из диаграмм; появится всплывающее окно подсказки со значениями этой диаграммы.

Правая панель

Важность предиктора. Содержит диаграмму, показывающую относительную важность каждого предиктора в оценке модели. Дополнительную информацию смотрите в разделе “Важность предиктора” на стр. 42.

Условные вероятности. При выборе узла или мини-диаграммы распределения на левой панели, на правой панели выводится соответствующая таблица условных вероятностей. Эта таблица содержит значение

условной вероятности для каждого значения узла и каждого сочетания значений в его родительских узлах. Кроме того, для каждого значения записи и каждого сочетания значений в родительских узлах приводится наблюдаемое число записей.

Параметры модели байесовской сети

На вкладке Параметры слепка байесовской сети задаются опции модифицирования построенной модели. Например, можно использовать узел байесовской сети для построения нескольких различных моделей при одних и тех же данных и параметрах, а затем посмотреть, как повлияет на результаты модификация настроек на этой вкладке в различных моделях.

Примечание: Эта вкладка только доступна после того, как слепок модели добавлен в поток.

Вычислить простые оценки склонности. Для моделей с флаговым полем назначения (которое возвращает предсказание Да или Нет) можно затребовать оценку склонности, которая обозначает правдоподобие правильного выходного значения, заданного для поля назначения. Это дополнение к другим предсказанным и доверительным значениям, которые можно сгенерировать при скоринге.

Вычислить скорректированные оценки склонности. Простые оценки склонности основываются только на данных обучения и могут выглядеть чрезмерно оптимистическими из-за тенденции многих моделей чрезмерно точно подгонять эти данные под модель. Корректировка склонностей направлена на компенсацию этого эффекта путем проверки выполнения модели на испытательном и проверочном разделах. Для этой опции требуется, чтобы поле разделения было определено в потоке и скорректированные оценки склонности были включены на узле моделирования до генерирования модели.

Добавить все вероятности. Задает, добавляются ли вероятности для каждой категории выходного поля к каждой записи, обрабатываемой узлом. Если эта опция не выбрана, добавляется только вероятность предсказанной категории.

Состояние по умолчанию этого переключателя определяется соответствующим переключателем на вкладке Дополнительно узла моделирования. Дополнительную информацию смотрите в разделе “Дополнительные опции узла байесовской сети” на стр. 120.

Сводка моделей байесовской сети

На вкладке Сводка слепка модели выводится информация о самой модели (*Анализ*), об используемых в ней полях (*Поля*), значениях параметров, используемых при построении модели, (*Параметры построения*) и об обучении модели (*Сводка по обучению*).

При первом просмотре узла результаты вкладки Сводка свернуты. Чтобы увидеть нужные вам результаты, разверните соответствующие им элементы при помощи элемента управления расширением слева от них или выведите все результаты, нажав кнопку **Развернуть все**. Чтобы скрыть результаты после завершения их просмотра, используйте управляющий элемент раскрытия для сворачивания конкретных результатов, которые нужно скрыть, или нажмите кнопку **Свернуть все**, чтобы свернуть все результаты.

Анализ. Выводится информация о конкретной модели.

Поля. Список полей, используемых в качестве полей назначения и входных полей при построении модели.

Параметры компоновки. Содержит информацию об используемых при построении модели параметрах.

Сводная информация по обучению. Выводится тип модели, поток, используемый для ее создания, создавший ее пользователь, отметка времени построения модели и время, затраченное на ее построение.

Глава 8. Нейронные сети

Нейронная сеть способна аппроксимировать широкий диапазон прогнозных моделей при минимальных требованиях к структуре модели и допущениям. Форма взаимосвязей определяется в процессе обучения. Если между переменной назначения и предикторами применима линейная взаимосвязь, результаты нейронной сети должны хорошо аппроксимировать результаты модели линейной регрессии. Если лучше подходит нелинейная взаимосвязь, нейронная сеть будет автоматически аппроксимировать "правильную" структуру модели.

Обратная сторона этой гибкости состоит в том, что нейросеть нелегко интерпретировать. Если задача в том, чтобы объяснить процесс, обуславливающий взаимосвязи между переменной назначения и предикторами, лучше подобрать ту или иную традиционную статистическую модель. Однако если интерпретируемость модели не важна, хорошие предсказания можно получить при помощи нейросети.

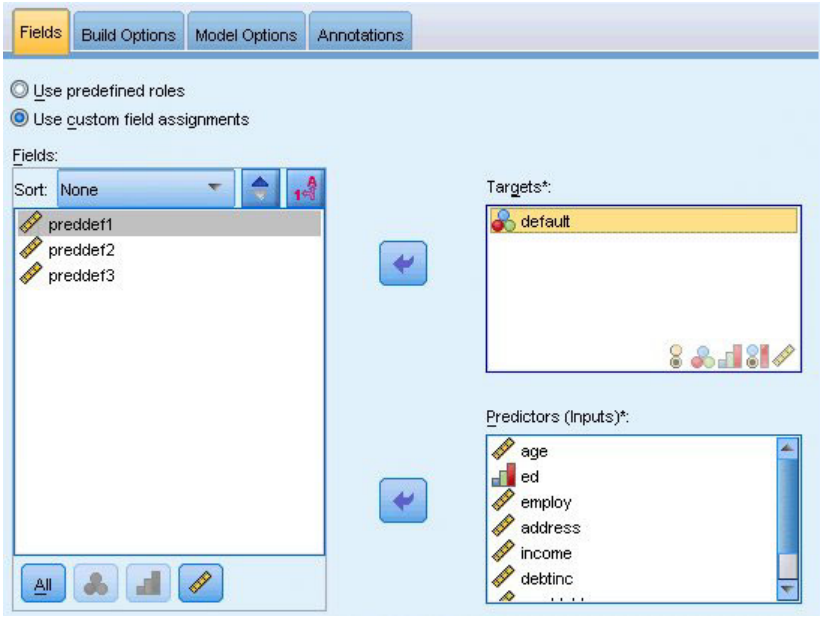


Рисунок 30. Вкладка Поля

Требования к полям. Должно быть по крайней мере одно поле назначения и одно входное поле. Поля, для которых задано Оба или Нет, игнорируются. Нет никаких ограничений на тип измерений ни для полей назначения, ни для предикторов (входных полей). Дополнительную информацию смотрите в разделе “Моделирование опций полей узла” на стр. 31.

Модель нейросетей

Нейронные сети представляют собой упрощенные модели работы нервной системы живых организмов. Базовые блоки называются **нейронами** и обычно сгруппированы в **слои**, как показано на следующем рисунке.

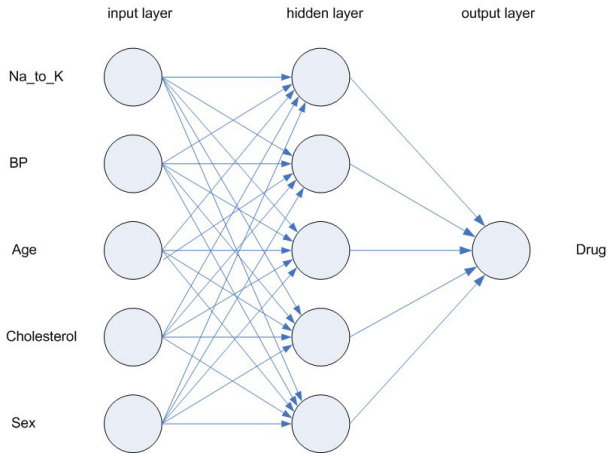


Рисунок 31. Структура нейронной сети

Нейронная сеть использует упрощенную модель обработки информации человеческим мозгом. Нейросети работают, обсчитывая большое количество связанных между собой обрабатываемых элементов, которые представляют абстрактную версию нейронов.

Обрабатывающие блоки (нейроны) сгруппированы в слои. В типичной нейросети есть три части: нейроны **входного слоя** представляют входные поля, есть один или несколько **скрытых слоев** и есть **выходной слой**, содержащий один или несколько нейронов, представляющих поля назначения. Каждому соединению между нейронами назначается та или иная сила воздействия, или **вес**. Входные данные поступают в первый слой, далее значения распространяются по слоям от каждого нейрона данного слоя в каждый нейрон следующего слоя. Окончательные результат снимается с выходного слоя.

Нейронная сеть обучается путем просмотра записей; для каждой записи нейронная сеть генерирует предсказание и, если предсказание неверно, вносит поправки в веса. Процесс повторяется большое число раз, и точность предсказаний постепенно повышается, пока не срабатывает один из критериев остановки.

Вначале все веса случайные и ответы нейронной сети на входные сигналы, скорее всего, бессмысленны. Нейронную сеть **обучают**. Примеры, для которых известны выходные значения, многократно предъявляются нейронной сети, и каждый раз ее отклик сравнивается с известным ответом. Информация от такого сравнения передается назад в нейронную сеть, постепенно изменяя веса. По мере обучения нейронная сеть начинает выдавать ответы, которые все точнее воспроизводят известные ответы. После обучения нейронную сеть применяют к будущим наблюдениям, для которых исход неизвестен.

Использование нейронных сетей совместно с унаследованными потоками

В версии 14 IBM SPSS Modeler появился новый узел нейронной сети, поддерживающий технологии бустинга и бэггинга и оптимизацию очень больших наборов данных. Существующие потоки, содержащие старый узел, по-прежнему смогут строить и оценивать модели в этом выпуске. Но такая поддержка будет удалена в будущих выпусках и рекомендуется отныне использовать новую версию.

Начиная с версии 13 поля с неизвестными значениями (то есть значениями, которые не встречались в обучающих данных) больше не приравниваются автоматически к пропущенным значениям, а оцениваются как значения `$null$`. Таким образом, если нужно оценивать поля с неизвестными значениями как непустые, используя старую (до версии 13) модель нейронной сети в версии 13 и новее, следует пометить неизвестные значения как пропущенные значения (например, при помощи узла Тип).

Обратите внимание на то, что ради совместимости унаследованные потоки, которые еще содержат старый узел, могут по-прежнему использовать пункт *Ограничить размер набора* из меню **Инструменты > Свойства потока > Опции**; эта опция применима только к сетям Коонена и узлам *k*-средних начиная с версии 14.

Целевые показатели

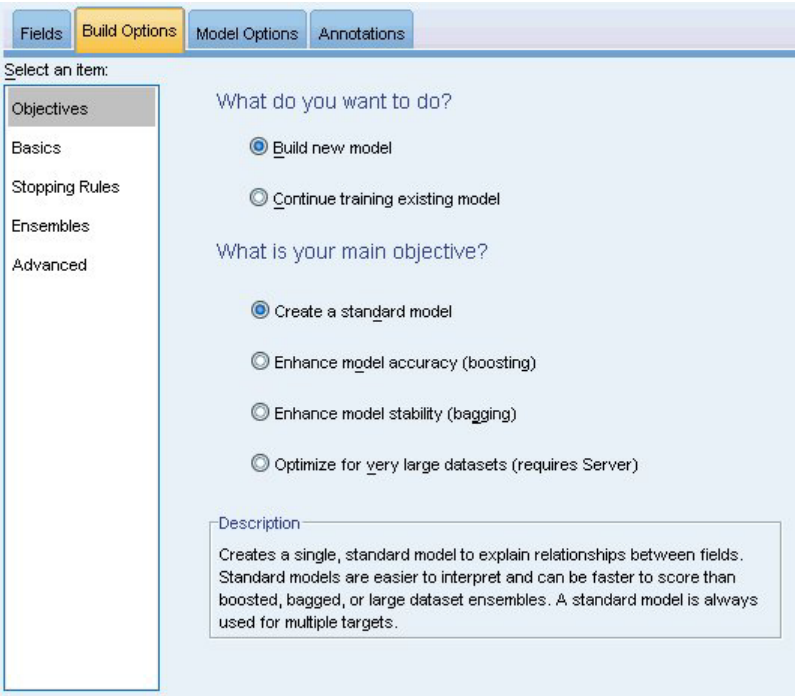


Рисунок 32. Целевые параметры

Что вы хотите сделать?

- **Построить новую модель.** Построить совершенно новую модель. Это обычная операция узла.
- **Продолжить обучение существующей модели.** Обучение продолжается с последней моделью, успешно сгенерированной узлом. Это дает возможность скорректировать или обновить существующую модель без необходимости обращаться к исходным данным, что может выполняться значительно быстрее, так как в поток вводятся только новые или обновленные записи. Информация по предыдущей модели сохраняется вместе с узлом моделирования, что позволяет использовать этот вариант, даже если предыдущий nugget модели недоступен в потоке или Палитре моделей.

  Примечание: Когда этот вариант доступен, все остальные управляющие элементы на вкладках Поля и Параметры конструкции блокируются.

Какова ваша главная цель? Выберите подходящую цель.

- **Создать стандартную модель.** Данный метод строит единичную модель для предсказания целевой переменной, используя предикторы. Вообще говоря, стандартные модели легче поддаются интерпретации и могут требовать меньше времени при скоринге, чем построенные с применением бустинга, бэггинга или ансамблей больших наборов данных.
- **Повысить точность модели (бустинг).** Данный метод строит модель ансамбля, используя бустинг, который генерирует последовательность моделей для получения более точных предсказаний. Ансамбли могут занять больше времени для их построения и скоринга, чем стандартная модель.

  Бустинг генерирует последовательность "компонентных моделей", каждая из которых строится по целому набору данных. Прежде чем строить каждую последовательную компонентную модель, записи взвешиваются на основе остатков для предшествующей компонентной модели. Наблюдениям с большими остатками придаются относительно большие веса прецедентов, с тем чтобы следующая

компонентная модель была сконцентрирована на том, чтобы хорошо предсказывать такие записи. Вместе такие компонентные модели образуют модель ансамбля. Модель ансамбля выполняет скоринг новых записей, пользуясь правилом объединения; доступные правила зависят от типа измерений целевой переменной.

- **Повысить стабильность модели (бэггинг).** Данный метод строит модель ансамбля, используя бэггинг (бутстреп-агрегирование), который генерирует множественные модели для получения более надежных предсказаний. Ансамбли могут занять больше времени для их построения и скоринга, чем стандартная модель.

  Бутстреп-агрегирование (бэггинг) формирует реплики обучающего набора данных путем выбора с возвращением из исходного набора данных. В результате создаются бутстреп-выборки исходного набора данных равного объема. Затем по каждой реплике формируется "компонентная модель". Вместе такие компонентные модели образуют модель ансамбля. Модель ансамбля выполняет скоринг новых записей, пользуясь правилом объединения; доступные правила зависят от типа измерений целевой переменной.
- **Создать модель для очень больших наборов данных (требует сервер IBM SPSS Modeler).** Данный метод строит модель ансамбля путем расщепления набора данных на отдельные блоки данных. Выберите этот вариант, если ваш набор данных слишком велик для построения моделей перечисленных выше, или для инкрементного построения модели. Данный вариант может потребовать меньше времени для построения, но больше времени для скоринга, чем стандартная модель. Для этой опции требуется соединение с сервер IBM SPSS Modeler .

Если полей назначения несколько, этот метод создает только стандартную модель, независимо от выбранной цели.

Основные параметры

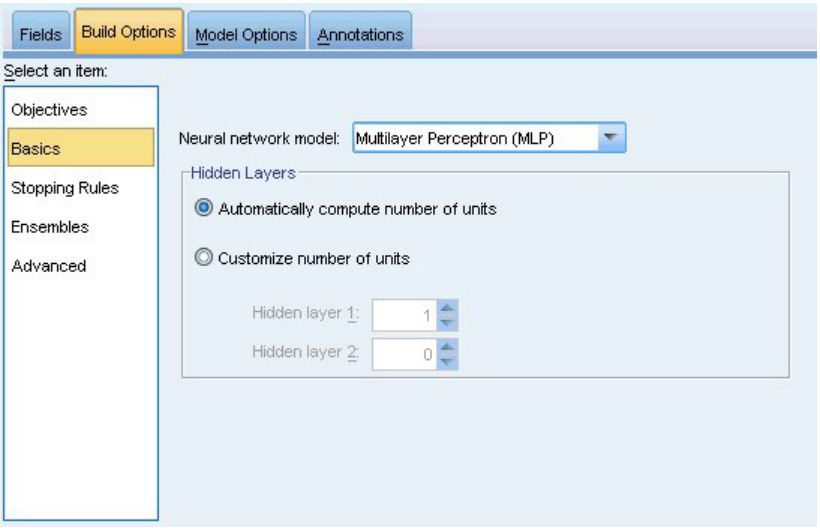


Рисунок 33. Основные параметры

Модель нейросети. Тип модели определяет способ соединения предикторов с полями назначения через один или несколько скрытых слоев нейронной сети. **Многослойный перцептрон (MLP)** создает возможность более сложных взаимосвязей, возможно, ценой больших затрат времени на обучение и оценку. **Радиальная базисная функция (RBF)** по сравнению с MLP может отличаться меньшими затратами времени на обучение и оценку, возможно, ценой меньшей мощности предсказания.

Скрытые слои. Скрытые слои нейросети содержат ненаблюдаемые обрабатывающие блоки (нейроны). Значение каждого скрытого нейрона - это некоторая функция предикторов; точная форма этой функции частично зависит от типа сети. Многослойный перцептрон может иметь один или два скрытых уровня; сеть радиальной базисной функции может иметь один скрытый слой.

- **Автоматически подсчитать число нейронов.** Эта опция строит сеть с одним скрытым слоем и вычисляет "наилучшее" число нейронов в скрытом слое.
- **Задать число нейронов.** Эта опция дает возможность задать число нейронов в каждом скрытом слое. Первый скрытый слой должен содержать как минимум один нейрон. Для второго скрытого слоя можно задать 0 нейронов, будет автоматически построен многослойный перцептрон с одним скрытым слоем.

  Примечание: Нет смысла задавать число нейронов, превышающее число непрерывных предикторов плюс суммарное число категорий во всех категориальных предикторах (флаговых, номинальных, порядковых).

Правила остановки

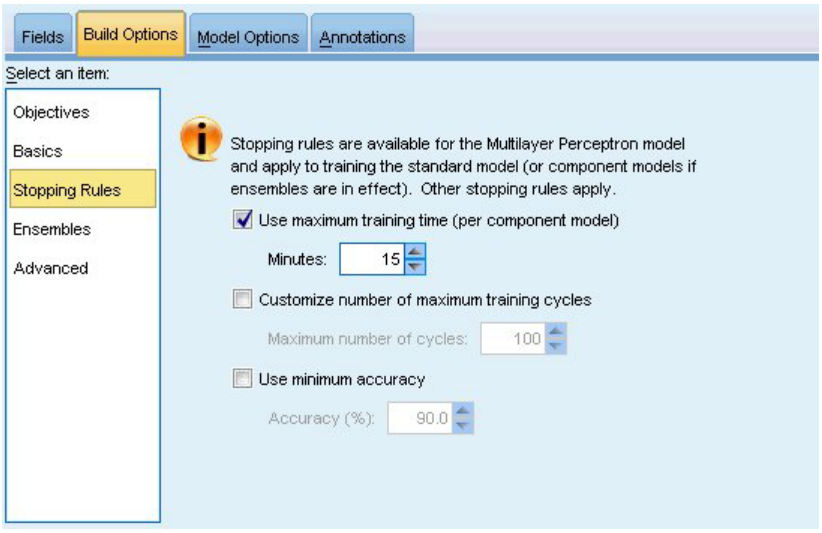


Рисунок 34. Настройка правил остановки

Это правила, которые определяют, когда останавливать обучение нейросетей многослойного перцептрона; при использовании алгоритма радиальной базовой функции эти параметры игнорируются. Обучение продолжается как минимум один цикл прохождения данных, а затем может остановиться согласно описанным ниже критериям.

Использовать максимальное время обучения (каждой модели компонентов). Определяет, нужно ли задать максимальное время выполнения алгоритма в минутах. Задайте положительное значение. При построении множественной модели это время обучения предоставляется каждому компоненту ансамбля моделей. Обратите внимание на то, что обучение может несколько выйти за заданное ограничение времени, чтобы завершить текущий цикл.

Задать максимальное число циклов обучения. Максимально допустимое число циклов обучения. Если максимальное количество циклов превышено, процесс обучения останавливается. Задайте целое число большее нуля.

Использовать минимальную точность. При помощи этой опции можно продолжать обучение до достижения заданной точности. Этого может и не произойти, но вы сможете прервать обучение в любой момент и сохранить сеть в состоянии наивысшей точности, достигнутой к этому моменту.

Кроме того, алгоритм обучения остановится, если от цикла к циклу перестанет снижаться ошибка, измеренная по набору для предотвращения сверхобучения, если относительное изменение ошибки обучения станет слишком мало или отношение текущей ошибки обучения станет мало по сравнению с начальной ошибкой.

Ансамбли

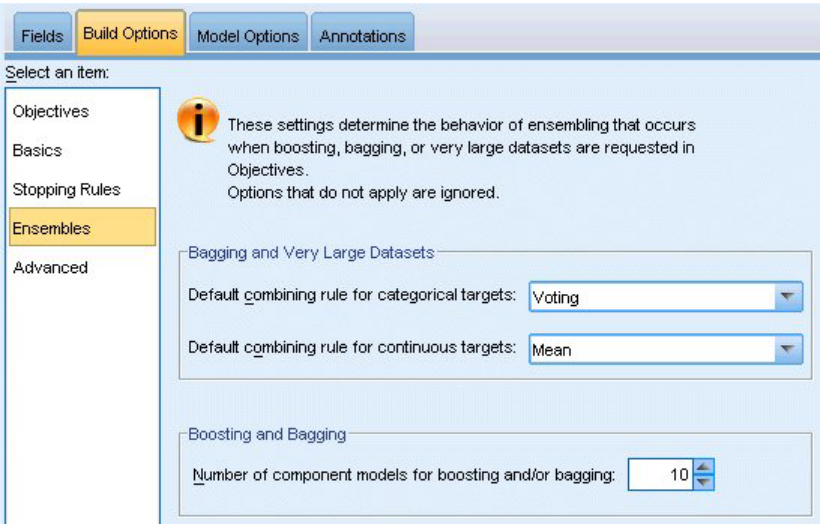


Рисунок 35. Параметры ансамблей

Данные параметры определяют поведение ансамбля, которое имеет место, когда на вкладке Цели запрашивается бэггинг, бустинг или очень большие наборы данных. Параметры, которые не применяются к выбранной цели, игнорируются.

Бэггинг и очень большие наборы данных. Это правило, которое применяется при скоринге ансамбля, чтобы объединить предсказанные значения для базовых моделей с целью вычисления значений скоринга для ансамбля.

- **Принятое по умолчанию правило объединения для категориальных полей назначения.** Предсказанные значения для ансамбля в случае категориальных целевых переменных можно объединить, используя голосование, наибольшую вероятность или наибольшую среднюю вероятность. **Голосование** позволяет выбрать категорию, которая наиболее часто имеет наибольшую вероятность в базовых моделях. **Наибольшая вероятность** позволяет выбрать категорию, которая имеет наибольшую вероятность среди всех базовых моделей. **Наибольшая средняя вероятность** позволяет выбрать категорию, которая имеет наибольшую вероятность при усреднении вероятностей категорий по базовым моделям.
- **Принятое по умолчанию правило объединения для непрерывных полей назначения.** Предсказанные значения для ансамбля в случае непрерывных целевых полей могут быть вычислены с использованием среднего значения или медианы предсказанных значений для базовых моделей.

Обратите внимание на то, что если цель состоит в повышении точности модели, выбор правила объединения игнорируется. При бустинге всегда используется взвешенное решение большинством голосов для скоринга категориальных целевых полей и взвешенная медиана для скоринга непрерывных целевых полей.

Бустинг и бэггинг. Задайте число базовых моделей для построения, когда целью является повышение точности или стабильности; для бэггинга это число бутстреп-выборок. Оно должно быть положительным целым.

Дополнительные опции

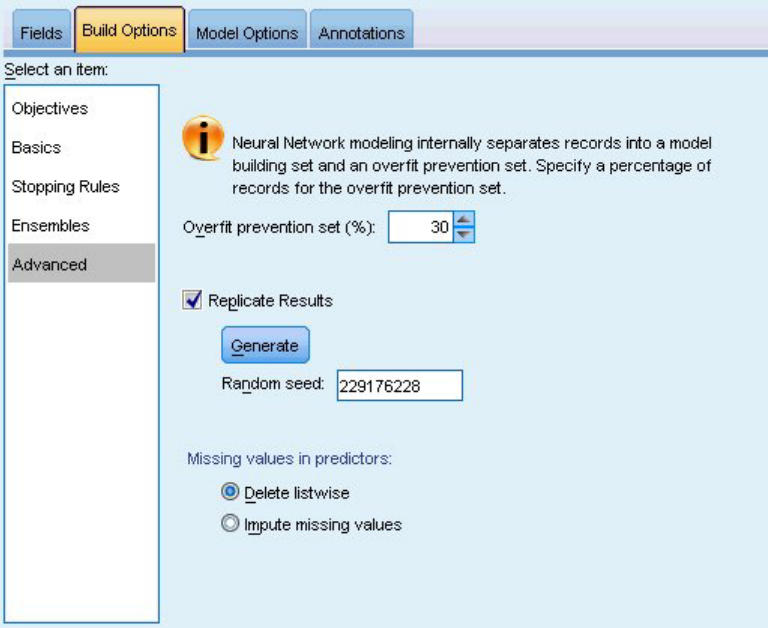


Рисунок 36. Дополнительные параметры

Дополнительные параметры включают те параметры, которые было бы неестественно отнести к одной из остальных групп параметров.

Множество предотвращения сверхобучения . По методу нейронной сети записи внутренне разделяются на набор для построения модели и набор для предотвращения сверхобучения - независимый набор записей данных для отслеживания ошибок в ходе обучения, чтобы не допустить учета в модели случайных изменений данных. Задайте процент записей. Значение по умолчанию - 30.

Воспроизвести результаты. Задание стартового числа генератора псевдослучайных чисел позволяет воспроизвести результаты. Задайте целое число или щелкните по **Генерировать**, чтобы сгенерировать псевдослучайное целое число в диапазоне между 1 и 2147483647 включительно. По умолчанию для воспроизводимости анализов используется стартовое число 229176228.

Пропущенные значения в предикторах. Задает, как обрабатывать пропущенные значения. **Удалять целиком** значит отстранять записи с пропущенными значениями предикторов от участия в построении модели. **Заполнить пропущенные значения** значит заменить пропущенные значения в предикторах и использовать соответствующие записи в анализе. В непрерывные поля подставляется среднее минимального и максимального среди наблюдавшихся значений; в категориальные поля подставляется категория, которая встречалась чаще всего. Обратите внимание на то, что записи с пропущенными значениями в любых других полях, заданных на вкладке Поля, всегда отстраняются от участия в построении модели.

Опции модели

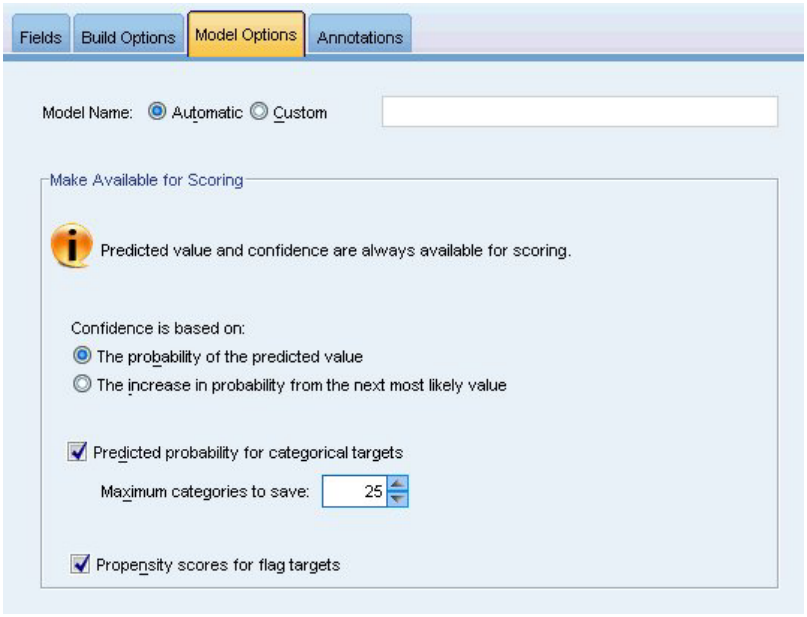


Рисунок 37. Вкладка Опции модели

Имя модели. Имя модели можно сгенерировать автоматически на основе целевых полей или задать самостоятельно. Автоматически генерируемое имя является именем целевого поля. Если полей назначения несколько, именем модели служит упорядоченный список имен полей через знак конъюнкции. Например, если поля назначения - *поле1 поле2 поле3*, то имя модели - *поле1 & поле2 & поле3*.

Сделать доступным для скоринга. При оценке модели нужно сгенерировать элементы, выбранные в этой группе. Предсказанное значение (для всех полей назначения) и достоверность (для категориальных полей назначения) вычисляются при оценке всегда. Вычисленный показатель доверия может быть основан на вероятности предсказанного значения (наивысшая предсказанная вероятность) или на разнице между наивысшей предсказанной вероятностью и вторым по величине значением предсказанной вероятности.

- **Предсказанную вероятность для категориальных целевых полей.** Будут вычислены предсказанные вероятности для категориальных полей назначения. Для каждой категории создается поле.
- **Оценки склонности для флаговых полей назначения .** Для моделей с флаговым полем назначения (которое возвращает предсказание Да или Нет) можно затребовать оценку склонности, которая обозначает правдоподобие правильного выходного значения, заданного для поля назначения. Модель находит простые оценки склонности; кроме того, если действуют разбиения, модель находит скорректированные оценки склонности на основе контрольного раздела.

Сводка для модели

Model Summary

Target	Previously defaulted
Model	Multilayer Perceptron
Stopping Rule Used	Error cannot be further decreased
Hidden Layer 1 Neurons	4

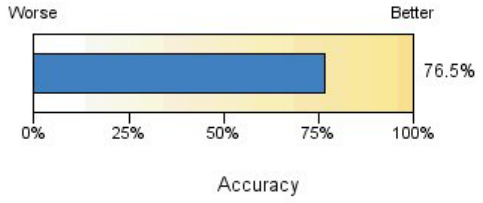


Рисунок 38. Представление Сводка для модели нейронных сетей

Представление Сводка для модели - это простая наглядная сводка о точности предсказания или классификации нейронной сети на данный момент.

Сводка для модели. Таблица содержит поле назначения, тип обученной нейронной сети, правило остановки, по которому остановилось обучение (выводится, если обучаемой сетью был многослойный перцептрон), и число нейронов в каждом скрытом слое сети.

Качество нейронной сети. Данная диаграмма показывает точность окончательной модели, представленную в форме "больше значит лучше". Для категориальной целевой переменной это просто процент записей, для которых предсказанное значение совпадает с наблюденным значением. Для непрерывной целевой переменной это 1 минус отношение средней абсолютной ошибки предсказания (среднего абсолютных значений разностей предсказанных и наблюденных значений) к диапазону предсказанных значений (разности максимального и минимального предсказанных значений).

Несколько переменных назначения. Если полей назначения несколько, то в строке **Назначение** таблицы выводятся все поля назначения. Точность, показанная на диаграмме, - это средняя точность по всем полям назначения.

Важность предикторов

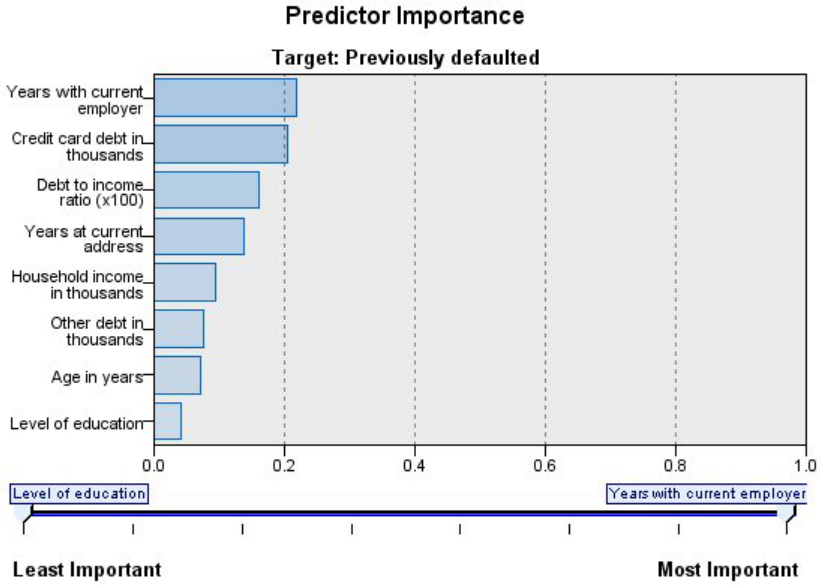


Рисунок 39. Представление Важность предикторов

Обычно при моделировании сосредотачивают внимание на наиболее важных предикторах и исключают или игнорируют наименее важные. Это помогает сделать диаграмма важности предикторов, показывая относительную важность каждого предиктора при оценке модели. Поскольку значения важности являются относительными, сумма этих значений для всех показанных предикторов равна 1,0. Важность переменных не связана с точностью модели. Она лишь связана с важностью каждого предиктора для предсказания, а не с точностью этого предсказания.

Несколько переменных назначения. При наличии нескольких целевых переменных каждая из них выводится на отдельной диаграмме, причем отдельные целевые переменные можно выбрать в в выпадающем списке **Целевая переменная**.

Предсказанные против наблюденных

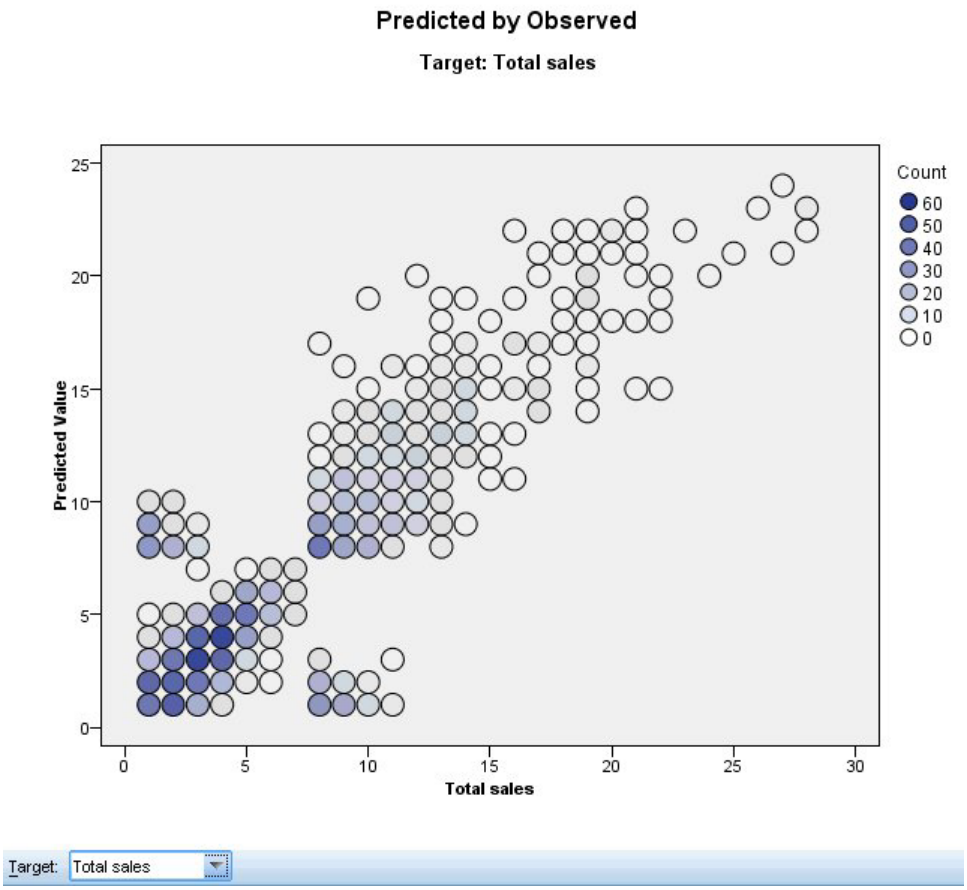


Рисунок 40. Представление Предсказанные против наблюденных

Для непрерывных целевых переменных, выводится диаграмма рассеяния с интервалами для предсказанных значений по вертикальной оси против наблюденных значений по горизонтальной оси.

Несколько переменных назначения. Если есть несколько непрерывных полей назначения, каждое из них выводится на отдельной диаграмме, причем показ и скрытие отдельных полей назначения можно выбрать в выпадающем списке **Поля назначения**.

Классификация

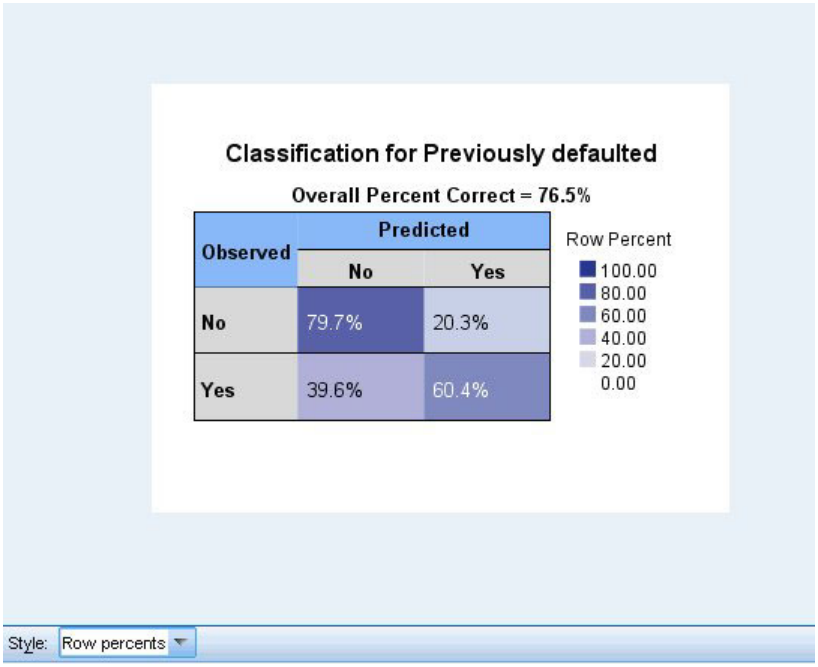


Рисунок 41. Представление Классификация, стиль процентов строк

Для категориальных целевых переменных в этой таблице выводится перекрестная классификация наблюденных и предсказанных значений целевой переменной в тепловой карте, а также общий процент правильных значений.

Стили таблиц. Существует несколько различных стилей вывода, которые можно выбрать в выпадающем списке **Стиль**.

- **Процент по строке.** Здесь выводятся проценты по строкам (количества в ячейках, выраженные в процентах от итогов по строке) в ячейках. Это задано по умолчанию.
- **Количества в ячейках,** Здесь выводятся количества в ячейках. Тени для тепловой карты по-прежнему основаны на значениях процентов по строкам.
- **Тепловая карта.** Значения для ячеек не выводятся, используется только затенение.
- **Сжатые.** Без вывода заголовков строк или столбцов или значений в ячейках. Может быть полезным, если у целевой переменной много категорий.

Пропущенные. Если в каких-либо записях для целевой переменной есть отсутствующие значения, они выводятся в строке **(Отсутствующие)** под всеми остальными действительными строками. Записи с отсутствующими значениями не участвуют в вычислении общего процента правильных.

Несколько переменных назначения. При наличии нескольких категориальных целевых переменных каждая из них выводится в отдельной таблице, причем отдельные целевые переменные можно выбрать в в выпадающем списке **Целевая переменная**.

Большие таблицы. Если у выводимой целевой переменной более 100 категорий, таблица не выводится.

Сеть

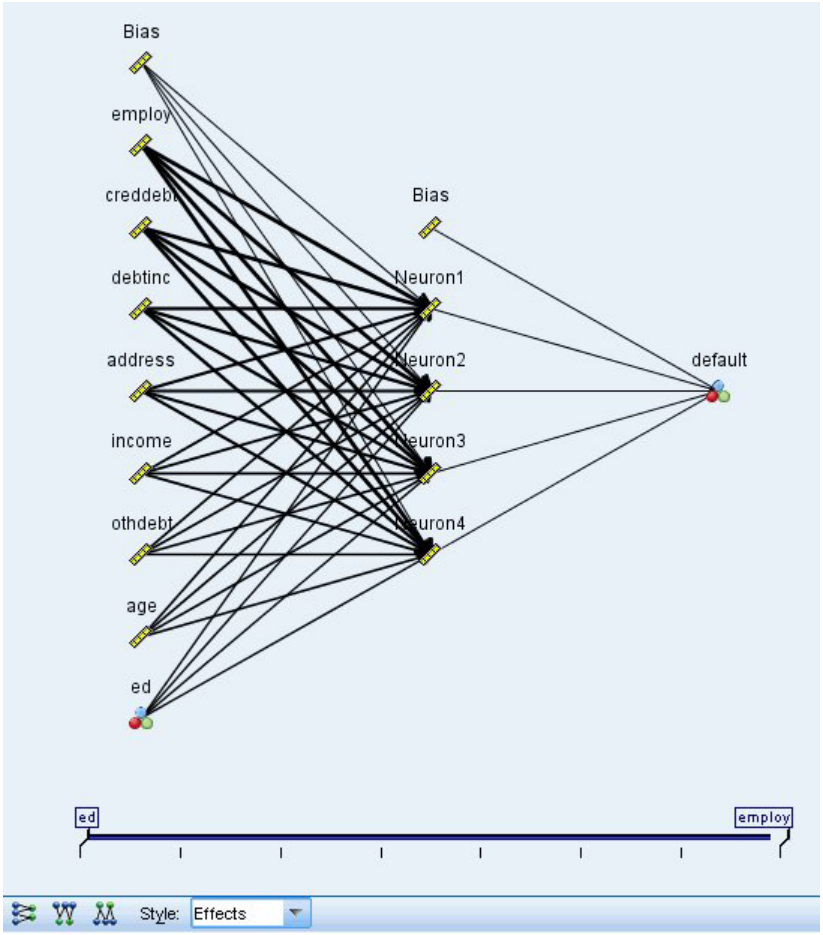


Рисунок 42. Представление сети, входные поля слева, стиль эффектов

Содержит графическое представление нейронной сети.

Стили диаграммы. Существует два различных стиля вывода, которые можно выбрать в выпадающем списке **Стиль**.

- **Эффекты.** Содержит все предикторы и поля назначения как узлы на диаграмме независимо от непрерывного или категориального типа измерений. Это задано по умолчанию.
- **Коэффициенты.** Содержит узлы нескольких индикаторов для категориальных предикторов и полей назначения. Цвет соединительных линий на диаграмме в стиле коэффициентов учитывает примерное значение веса синапса.

Ориентация диаграммы. По умолчанию сетевая диаграмма строится со входными полями слева и полями назначения справа. При помощи элементов управления на панели инструментов можно изменить ориентацию диаграммы, так что входные поля окажутся вверху и поля назначения внизу или входные поля внизу и поля назначения вверху.

Важность предикторов. Соединительные линии на диаграмме взвешиваются на основе важности предиктора, причем более толстая линяя соответствует большей важности. На панели инструментов есть ползунок важности предикторов, который управляет показом и скрытием предикторов на диаграмме сети. Это не изменяет модели, а просто позволяет сосредоточить внимание на наиболее важных предикторах.

Несколько переменных назначения. Если полей назначения несколько, на диаграмме показаны все.

Параметры

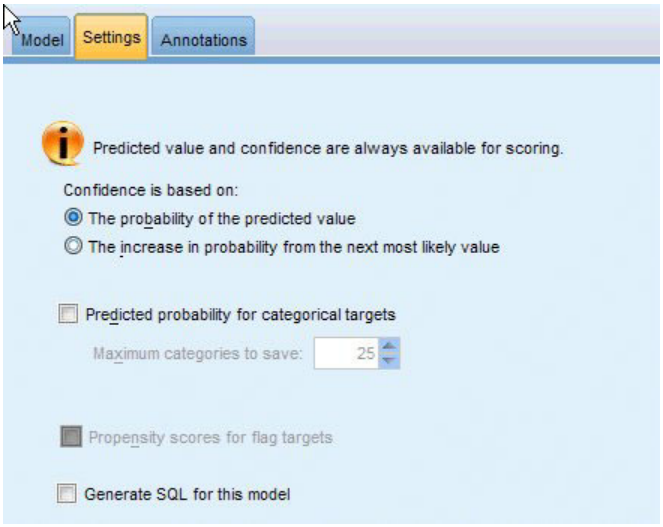


Рисунок 43. Вкладка Параметры

При оценке модели нужно сгенерировать элементы, выбранные на этой вкладке. Предсказанное значение (для всех полей назначения) и достоверность (для категориальных полей назначения) вычисляются при оценке всегда. Вычисленный показатель доверия может быть основан на вероятности предсказанного значения (наивысшая предсказанная вероятность) или на разнице между наивысшей предсказанной вероятностью и вторым по величине значением предсказанной вероятности.

- **Предсказанную вероятность для категориальных целевых полей.** Будут вычислены предсказанные вероятности для категориальных полей назначения. Для каждой категории создается поле.
- **Оценки склонности для флаговых полей назначения .** Для моделей с флаговым полем назначения (которое возвращает предсказание Да или Нет) можно затребовать оценку склонности, которая обозначает правдоподобие правильного выходного значения, заданного для поля назначения. Модель находит простые оценки склонности; кроме того, если действуют разбиения, модель находит скорректированные оценки склонности на основе контрольного раздела.

Сгенерировать SQL для данной модели. Когда используются данные, хранящиеся в базе данных, код SQL можно передавать обратно в базу данных для выполнения, что обеспечивает высокую производительность многих операций.

Оценить по преобразованию в собственный SQL. Если этот параметр выбран, то создается SQL для скоринга модели внутри приложения.

Глава 9. Список решений

Модели списка решений определяют подгруппы или **сегменты**, которые показывают более высокое или более низкое правдоподобие данных бинарных (да или нет) результатов по сравнению со всей выборкой. Например, можно бы искать клиентов с низкой вероятностью оттока или с высокой вероятностью отклика на конкретное предложение или кампанию. Средство просмотра списка решений даёт вам полный контроль над моделью, позволяя редактировать сегменты, добавлять свои собственные бизнес-правила, задавать, как оценивается каждый сегмент, а также настраивать модель многими другими способами, чтобы оптимизировать долю попаданий по всем сегментам. В результате он особенно хорошо подходит для создания списков рассылки или иного способа определения целевых записей для конкретной кампании. Вы можете также использовать несколько **задач исследования данных**, чтобы объединить подходы к моделированию - например, путём идентификации хорошо и плохо работающих сегментов в рамках одной и той же модели, и включения или исключения каждого на этапе скоринга по мере необходимости.

Сегменты, правила и условия

Модель состоит из списка сегментов, каждый из которых определяется правилом выбора соответствующих записей. Заданное правило может содержать несколько условий, например:

```
RFM_SCORE > 10 and
MONTHS_CURRENT <= 9
```

Правила применяются в порядке, указанном списком, и первое подходящее правило определяет результат для данной записи. Сами по себе правила или условия могут перекрываться, но порядок правил устраняет возможную неоднозначность. Если никакое правило не соответствует записи, запись присваивается остаточному правилу.

Полное управление скорингом

Средство просмотра списка решений позволяет просматривать, модифицировать и реорганизовывать сегменты, а также выбирать, какой сегмент следует включить или исключить для целей скоринга. Например, вы можете выбрать исключение одной группы заказчиков из будущих предложений и включение других групп и сразу увидеть, как это влияет на общий коэффициент попадания. Модели списка решений выдают оценку *Yes* для включенных сегментов и оценку *$null$* для всех прочих, включая остаточный сегмент. Такое прямое управление скорингом делает модели списка решений идеальными для генерирования списков рассылки, и они широко используются в организации отношений с клиентами, в том числе для call-центров и программ маркетинга.

Задачи исследования данных, измерения и отбора

Процесс моделирования управляется **задачами исследования данных**. Каждая задача исследования данных эффективно инициирует новый проход моделирования и возвращает новый набор альтернативных моделей, из которых можно делать выбор. Задача по умолчанию основана на ваших начальных спецификациях в узле списка решений, но вы можете определить любое количество пользовательских задач. Вы можете также применять задачи итеративно - например, запустить поиск записей с высокой вероятностью по всему набору обучения, а затем поиск записей с низкой вероятностью по остатку, чтобы отсеять плохо работающие сегменты.

Отбор данных

Вы можете определить отбор данных и пользовательские показатели модели для построения модели и её оценки. Например, можно задать отбор данных в задаче исследования данных, чтобы подогнать модель к конкретному региону, и создать пользовательский показатель, чтобы оценить, насколько хорошо эта модель

работает по всей стране. В отличие от задач исследования данных, показатели не меняют базовую модель, но они дают иной взгляд для оценки, насколько хорошо базовая модель работает.

Добавление ваших знаний бизнеса

При помощи использования тонкой настройки или расширения сегментов, идентифицированных алгоритмом, Средство просмотра списка решений даёт возможность включить ваше знание бизнеса непосредственно в модель. Вы можете редактировать сегменты, сгенерированные этой моделью, или добавлять дополнительные сегменты на основе заданных вами правил. Затем можно применять эти изменения и предварительно просматривать результаты.

Для последующего анализа динамическая связь с Excel даёт возможность экспортировать данные в Excel, где их можно использовать для создания диаграмм презентации или для вычисления пользовательских показателей, таких, как сложная прибыль или возврат инвестиций; эти показатели можно рассматривать в программе просмотра списка решений при построении модели.

Пример. Маркетинговый отдел финансовой компании хочет достичь более выгодных результатов в будущей рекламной кампании, подбирая правильные предложения для каждого из покупателей. На основе данных предыдущих кампаний вы можете использовать модель Список решений, чтобы определить характеристики клиентов, которые наиболее вероятно откликнутся положительно, и создать список рассылки на основе результатов.

Требования. Одно категориальное поле назначения с уровнем измерения типа *Флаг* или *Номинал*, содержащее двоичный вывод, который нужно предсказать (да/нет), и по крайней мере одно входное поле. Для типа поля назначения *Номинал* нужно вручную выбрать одно значение, которое будет рассматриваться как **попадание**, или **отклик**; все остальные значения объединяются вместе как **непопадание**. Может быть задано также дополнительное поле частоты. Количественные поля даты/времени игнорируются. Количественные входные данные числового диапазона автоматически делятся по интервалам с помощью алгоритма, заданного на вкладке Дополнительно узла моделирования. Для более точного управления разбиением на интервалы добавьте вышележащий узел разбиения на интервалы и используйте уже разбитое на интервалы поле с уровнем измерения *Порядковый* в качестве входного поля.

Опции модели списка решений

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Создание моделей расщепления. Строит отдельную модель для каждого возможного значения входных полей, заданных как поля расщепления. Дополнительную информацию смотрите в разделе “Построение моделей расщепления” на стр. 28.

Режим. Задает метод, использованный для построения модели.

- **Создать модель.** Автоматически генерирует модель на палитре моделей, где выполняется узел. Полученную модель можно добавлять в потоки в целях оценки, но нельзя дальше изменять.
- **Запустить интерактивный сеанс.** Открывает интерактивное (выходное) окно моделирования Средство просмотра списка решений, позволяя выбирать модели из числа нескольких альтернатив и повторно применять алгоритм с различными параметрами, эффективно увеличивая и изменяя модель. Дополнительную информацию смотрите в разделе “Средство просмотра списка решений” на стр. 141.
- **Использовать сохраненную информацию об интерактивном сеансе.** запускает интерактивный сеанс, используя ранее сохраненные параметры. Интерактивные параметры можно сохранить при помощи Средство просмотра списка решений, используя меню Генерировать (создать модель или узел моделирования) или меню Файл (изменить узел, на котором запущен сеанс).

Целевое значение. Указывает значение поля назначения, задающее выходные данные, которые вы хотите смоделировать. Например, если поле назначения churn закодировано как `0 = no` и `1 = yes`, задайте `1` для определения правил, в соответствии с которыми записи будут отнесены к вероятным для оттока клиентов.

Найти сегменты, у которых. Указывает, какие элементы нужно искать при поиске переменной назначения - с **Высокой вероятностью** или **Низкой вероятностью** появления. Нахождение и исключение таких элементов может быть полезным способом улучшения модели, в частности, когда у остатка низкая вероятность.

Максимальное число сегментов. Задает максимальное количество сегментов для возврата. Создается *N* лучших сегментов, причем самый лучший - это тот, у которого самая высокая вероятность, или, если у нескольких сегментов одинаковая вероятность, самое высокое покрытие. Минимальный разрешенный параметр - 1; максимальный параметр не определяется.

Минимальный размер сегмента. Два параметра ниже определяют минимальный размер сегмента. Приоритет у большего из этих значений. Например, если значение процентной доли оказывается больше абсолютного значения, приоритет у параметра процентной доли.

- **Как процент от предыдущего сегмента (%).** Задает минимальный размер группы как процентную долю записей. Минимальное разрешенное значение параметра - 0; максимальное - 99,9.
- **Как абсолютное значение (N).** Задает минимальный размер группы как абсолютное число записей. Минимальный разрешенный параметр - 1; максимальный параметр не определяется.

Правила сегментов.

Максимальное число атрибутов. Задает максимальное количество условий на одно правило сегмента. Минимальное разрешенное значение параметра - 1; максимальное значение не ограничивается.

- **Разрешить повторное использование атрибутов.** Если эта опция включена, каждый цикл может рассматривать все атрибуты, даже те, которые уже были использованы в предыдущих циклах. Условия для сегмента строятся по циклам, где в каждом цикле добавляется новое условие. Количество циклов определяется параметром **Максимальное число атрибутов**.

Доверительный интервал для новых условий (%). Задает доверительный уровень для проверки значимости сегмента. Этот параметр играет существенную роль для определения числа возвращаемых сегментов (если такие есть) и для определения правила число-условий-на-сегмент. Чем больше это значение, тем меньше результатов возвращается. Минимальное разрешенное значение параметра - 50; максимальное - 99,9.

Дополнительные опции узла Список решений

Дополнительные опции позволяют тонко настроить процесс построения моделей.

Метод разбиения на группы. Этот способ используется для разбиения на группы количественных полей (по равному количеству или по равной ширине).

Число интервалов. Количество интервалов, создаваемых для количественных полей. Минимальное разрешенное значение параметра - 2; максимальное значение не ограничивается.

Ширина поиска модели. Максимальное количество результатов модели на цикл, которые могут использоваться для следующего цикла. Минимальное разрешенное значение параметра - 1; максимальное значение не ограничивается.

Ширина поиска правила. Максимальное количество результатов правила на цикл, которые могут использоваться для следующего цикла. Минимальное разрешенное значение параметра - 1; максимальное значение не ограничивается.

Фактор слияния интервалов. Минимальный объем, на который должен вырасти сегмент при слиянии с соседним сегментом. Минимальное разрешенное значение параметра - 1,01; максимальное значение не ограничивается.

- **Разрешить пропущенные значения в условиях.** `True` - для разрешения критерия `IS MISSING` в правилах.
- **Отбросить промежуточные результаты.** При значении `True` возвращаются только окончательные результаты процесса поиска. Окончательный результат - это такой результат, который больше не обновляется в процессе поиска. При значении `False` возвращаются также промежуточные результаты.

  Максимальное число альтернативных вариантов. Задает количество альтернатив для вывода при запуске задачи исследования данных. Минимальное разрешенное значение параметра - 1; максимальное значение не ограничивается.

  Обратите внимание на то, что задача исследования данных возвратит только фактическое количество альтернатив до заданного максимума. Например, если задан максимум 100, но найдены только 3 альтернативы, будет показано значение 3.

Слепок модели списка решений

Модель состоит из списка **сегментов**, каждый из которых определяется **правилом**, отбирающим подходящие записи. Перед генерированием модели сегменты можно легко просмотреть и изменить, выбирая, какие из них включить в модель, а какие исключить. При использовании скоринга модели списка решений возвращают *да* для включенных сегментов и *$null$* для всего остального, в том числе для остатка. Такое прямое управление через скоринг делает модели списка решений идеальными для генерирования списков рассылки, и они широко используются в управлении связями с клиентами, в том числе в колл-центрах и в прикладных программах маркетинга.

При запуске потока, содержащего модель списка решений, узел добавляет три новых поля, в том числе оценку (*1*, что означает *да*, для включенных поле или *$null$* для исключенных полей), вероятность (коэффициент попаданий) для сегмента, в который попадает запись, и номер ID для сегмента. Имена новых полей получаются из имени выходного поля, значение которого предсказывается, добавлением префикса *$D-* для оценки, *$DP-* для вероятности и *$DI-* для ID сегмента.

Скоринг модели основан на значении назначения, заданном при построении модели. Можно вручную исключить сегменты, и для них будет задано значение *$null$*. Например, если вы запускаете поиск малых вероятностей, чтобы найти сегменты с меньшим среднего коэффициентом попадания, эти “низкие” сегменты будут оцениваться как *да*, пока вы их вручную не исключите. При необходимости значения null можно перекодировать в *нет*, используя узел Извлечение или Заполнение.

PMML

Модель списка решений можно хранить как PMML RuleSetModel с критерием выбора "первое попадание”. Однако ожидается, что у всех правил одинаковая оценка. Для разрешения изменений поля назначения или значения назначения модели с несколькими наборами правил можно сохранять в одном файле, чтобы применять по очереди, не подходящие для первой модели наблюдения передаются во вторую модель и так далее. Для обозначения такого нестандартного поведения используется имя алгоритма *DecisionList*, и только модели с наборами правил и с таким именем будут распознаваться как модели списка решений и в таком качестве оцениваться.

Параметры слепка моделей списка решений

На вкладке Параметры слепка моделей Список решений можно получить оценки склонностей и включить или выключить оптимизацию SQL. Эта вкладка доступна только после добавления в поток слепка модели.

Вычислить простые оценки склонности. Для моделей с флаговым полем назначения (которое возвращает предсказание Да или Нет) можно затребовать оценку склонности, которая обозначает правдоподобие правильного выходного значения, заданного для поля назначения. Это дополнение к другим предсказанным и доверительным значениям, которые можно сгенерировать при скоринге.

Вычислить скорректированные оценки склонности. Простые оценки склонности основываются только на данных обучения и могут выглядеть чрезмерно оптимистическими из-за тенденции многих моделей чрезмерно точно подгонять эти данные под модель. Корректировка склонностей направлена на компенсацию этого эффекта путем проверки выполнения модели на испытательном и проверочном разделах. Для этой опции требуется, чтобы поле разделения было определено в потоке и скорректированные оценки склонности были включены на узле моделирования до генерирования модели.

Оценить по преобразованию в собственный SQL. Если этот параметр выбран, то создается SQL для скоринга модели внутри приложения.

Средство просмотра списка решений

Простой в использовании и основанный на задачах графический интерфейс Средство просмотра списка решений исключает сложности из процесса построения моделей, освобождая вас от мелких деталей способов исследования данных и позволяя уделить основное внимание тем частям анализа, где требуется вмешательство пользователя, в частности, при задании целей, выборе групп назначения, анализе результатов и выборе оптимальной модели.

Панель Рабочая модель

На панели рабочей модели выводится текущая модель, в том числе задачи исследования данных и другие действия, применяемые к рабочей модели.

ID. Определяет последовательный порядок сегментов. Сегменты модели вычисляются последовательно в соответствии с их номерами ID.

Правила сегментов. Содержит имя сегмента и определенные для него условия. По умолчанию имя сегмента - это имя поля или объединенные имена полей, использованных в условии, разделенные запятыми.

Значение. Представляет собой поле, значение в котором вы хотите предсказать, используя предположение, что оно связано со значениями в других полях (предикторах).

Примечание: Через диалоговое окно “Организация показателей модели” на стр. 151 можно включать и выключать вывод на панель следующих опций.

Покрытие. Круговая диаграмма визуально представляет покрытие каждого сегмента по сравнению с полным покрытием.

Покрытие (n). Список покрытий для каждого сегмента по отношению к полному покрытию.

Частота. Список количеств попаданий по отношению к покрытию. Например, если покрытие равно 79, а частота - 50, это означает, что для данного сегмента получено 50 откликов на 79 предложений.

Вероятность. Указывает вероятность сегмента. Например, если покрытие равно 79, а частота - 50, это означает, что вероятность для сегмента равна 63,29% (50, деленное на 79).

Ошибка. Обозначает ошибку сегмента.

Данные внизу панели обозначают покрытие, частоту и вероятность всей модели.

Панель инструментов рабочей панели

Панель рабочей модели предоставляет следующие функции через свою панель инструментов.

Примечание: Некоторые функции доступны также, если щелкнуть правой кнопкой мыши по сегменту модели.

Таблица 9. Кнопки панели инструментов рабочей модели.

Кнопка панели инструментов	Описание
	Запускает диалоговое окно Генерировать новую модель, где представлены опции для создания слепка новой модели.
	Сохраняет текущее состояние интерактивного сеанса. Параметры узла моделирования Список решений изменяются на текущие, в том числе относящиеся к задачам исследования данных, снимкам моделей, выбору данных и пользовательским показателям. Для восстановления сеанса в этом состоянии включите переключатель **Использовать информацию сохраненного сеанса** на вкладке Модель узла моделирования и нажмите кнопку **Пуск**.
	Открывает диалоговое окно Организовать показатели модели. Дополнительную информацию смотрите в разделе “Организация показателей модели” на стр. 151.
	Открывает диалоговое окно Организовать выбор данных. Дополнительную информацию смотрите в разделе “Организация выбора данных” на стр. 146.
	Открывает вкладку Снимки. Дополнительную информацию смотрите в разделе “Вкладка Снимки” на стр. 143.
	Открывает вкладку Альтернативы. Дополнительную информацию смотрите в разделе “Вкладка Альтернативы”.
	Делает снимок структуры текущей модели. Снимки выводятся на вкладке Снимки и обычно используются в целях сравнения моделей.
	Запускает диалоговое окно Вставка сегментов, где представлены опции для создания новых сегментов модели.
	Запускает диалоговое окно Изменение правил сегментов, где представлены опции добавления условий для сегментов модели или изменения ранее определенных условий сегментов модели.
	Перемещает выбранный сегмент вверх по иерархии модели.
	Перемещает выбранный сегмент вниз по иерархии модели.
	Удаляет выбранный сегмент.
	Переключение между включением и невключением выбранного сегмента в модель. Если сегмент исключается, его результаты добавляются к остатку. Это отличается от удаления сегмента, так как есть опция повторной активации сегмента.

Вкладка Альтернативы

На сгенерированной при нажатии кнопки **Найти сегменты** вкладке Альтернативы выводится список всех альтернативных результатов исследования данных для выбранной модели или сегмента на панели рабочей модели.

Чтобы перевести альтернативу в состояние рабочей модели, выделите нужную альтернативу и нажмите кнопку **Загрузить**; альтернативная модель появится на панели рабочей модели.

Примечание: Вкладка Альтернативы появляется только в том случае, если вы задали **Максимальное количество альтернатив** на вкладке Эксперт узла моделирования Список решений, чтобы создать несколько альтернатив.

Каждая сгенерированная альтернатива модели выводит информацию о конкретной модели:

Имя. Все альтернативы нумеруются последовательно. Первая альтернатива обычно содержит лучшие результаты.

Цель. Указывает значение назначения. Например: 1, что эквивалентно логическому "true".

Число сегментов. Количество правил сегментов, использованных в альтернативной модели.

Покрытие. Покрытие альтернативной модели.

Част. Количество попаданий по отношению к покрытию.

Вероят. Указывает в процентах вероятность альтернативной модели.

Примечание: Альтернативные результаты не сохраняются с моделью; они действуют только в течение активного сеанса.

Вкладка Снимки

Снимок - это представление модели в конкретный момент времени. Например, может потребоваться сделать снимок модели перед загрузкой другой альтернативной модели на панель рабочей модели, если вам не хочется терять промежуточные результаты работы текущей модели. На вкладке Снимки перечисляются все снимки моделей, сделанные вручную для любого числа состояний рабочей модели.

Примечание: Снимки сохраняются вместе с моделью. Рекомендуется сделать снимок при загрузке первой модели. Этот снимок обеспечит сохранность исходной структуры модели, и вы всегда сможете вернуться к этой структуре. Сгенерированное имя снимка выводится как отметка времени, обозначающая, когда он был сделан.

Создать снимок модели

1. Выберите соответствующую модель/альтернативу для вывода на панели рабочей модели.
2. Внесите все необходимые изменения в рабочую модель.
3. Нажмите кнопку **Сделать снимок**. Новый снимок появится на вкладке Снимки.

   Имя. Имя снимка. Можно изменить имя снимка, щелкнув по нему дважды.

   Цель. Указывает значение назначения. Например: 1, что эквивалентно логическому "true".

   Число сегментов. Количество правил сегментов, использованных в модели.

   Покрытие. Покрытие модели.

   Част. Количество попаданий по отношению к покрытию.

   Вероят. Указывает в процентах вероятность модели.
4. Чтобы перевести снимок в состояние рабочей модели, выделите нужный снимок и нажмите кнопку **Загрузить**; снимок модели появится на панели рабочей модели.
5. Снимок можно удалить, нажав кнопку **Удалить**, или щелкнуть правой кнопкой мыши по снимку и выбрать в меню **Удалить**.

Работа с Средство просмотра списка решений

Модель, которая наилучшим образом предскажет отклики и поведение клиентов, строится в несколько этапов. Когда запускается Средство просмотра списка решений, рабочая модель заполняется определенными сегментами модели и показателями, и готова к тому, чтобы вы запустили задачу исследования данных, при необходимости изменили сегменты и/или показатели и сгенерировали новую модель или узел моделирования.

Можно добавлять одно или несколько правил сегментов, пока вы не разработаете удовлетворительную модель. Правила сегментов можно добавить в модель, запустив задачи исследования данных или используя функцию **Изменить правило сегмента**.

В процессе построения модели можно оценить ее производительность, испытав модель на данных измерений, визуализируя ее диаграммой или генерируя пользовательские показатели Excel.

Если есть сомнения в качестве модели, можно сгенерировать новую модель и поместить ее на холст IBM SPSS Modeler или на палитру Модель.

Задачи исследования данных

Задача исследования данных - это собрание параметров, определяющих, каким образом генерируются новые правила. Некоторые из этих параметров можно выбрать для обеспечения гибкости при простой и быстрой адаптации моделей в новых ситуациях. Задача состоит из шаблона задачи (типа), цели и выбора построения (набора данных для исследования).

В следующих разделах представлены подробности различных операций задач исследования данных:

- "Запуск задач исследования"
- "Создание и изменение задачи исследования данных"
- "Организация выбора данных" на стр. 146

Запуск задач исследования: Средство просмотра списка решений позволяет вручную добавить в модель правила сегментов, запустив задачи исследования данных или скопировав и вставив правила сегментов из одной модели в другую. Задача исследования данных содержит информацию о том, как сгенерировать новые правила сегментов (установки параметров исследования данных, таких как стратегия поиска, атрибуты поиска, ширина поиска, доверительный интервал и т.д.), а также о поведении клиентов, которое нужно предсказать, и о данных для изучения. Цель задачи исследования данных - поиск лучших возможных правил сегментов.

Чтобы сгенерировать правило сегмента модели при запуске задачи исследования данных:

1. Щелкните по строке **Остаток**. Если на панели рабочей модели уже выведены сегменты, можно выбрать также один из этих сегментов, чтобы найти дополнительные правила, основываясь на этом сегменте. После выбора остатка или сегмента используйте один из следующих способов для генерирования модели или альтернативные модели:

- В меню Инструменты выберите **Найти сегменты**.
- Щелкните правой кнопкой мыши по строке или сегменту **Остаток** и выберите **Найти сегменты**.
- Нажмите кнопку **Найти сегменты** на панели рабочей модели.

Пока задача обрабатывается, ход ее выполнения отображается в нижней части рабочего пространства и информирует, когда задача окончится. Сколько именно времени потребуется для выполнения задачи, зависит от сложности задачи исследования данных и от размера набора данных. Если в результатах есть только одна модель, она выводится на панель рабочей модели сразу после выполнения задачи; однако при наличии в результатах нескольких моделей они выводятся на вкладке Альтернативы.

Примечание: Результатом задачи могут быть или модели, или отсутствие моделей, или неудача.

Процесс поиска новых правил сегментов можно повторять до того момента, когда никакие новые правила уже не будут добавляться в модель. Это указывает на то, что все значимые группы клиентов уже найдены.

Задачу исследования данных можно запустить для любого существующего сегмента модели. Если результат задачи не тот, что вы ищите, можно выбрать запуск другой задачи исследования данных для того же сегмента. Это обеспечит нахождение дополнительных правил для выбранного сегмента. Сегменты, расположенные "ниже" выбранного (то есть добавленные в модель позже выбранного сегмента), замещаются новыми сегментами, так как каждый сегмент зависит от своих предшественников.

Создание и изменение задачи исследования данных: Задача исследования данных - это механизм поиска собрания правил, которые образуют модель данных. Вместе с критериями поиска, определенными в выбранном шаблоне, эта задача определяет также назначение (то есть фактический вопрос, который был

основанием для анализа, например, сколько покупателей могут откликнуться на почтовую рассылку), а также идентифицирует наборы данных для использования. Цель задачи исследования данных - поиск лучших возможных моделей.

Создать задачу исследования данных

Чтобы создать задачу исследования данных:

1. Выберите сегмент, в котором вы хотите исследовать дополнительные условия.
2. Нажмите кнопку **Параметры**. Откроется диалоговое окно Создать/изменить задачу исследования данных. В этом диалоговом окне предоставлены опции для определения задачи исследования данных.
3. Внесите необходимые изменения и нажмите кнопку **OK**, чтобы вернуться на панель рабочей модели. Средство просмотра списка решений использует эти параметры в качестве параметров по умолчанию для запуска каждой задачи, пока не будет выбрана альтернативная задача или альтернативные параметры.
4. Нажмите кнопку **Найти сегменты**, чтобы начать выполнение задачи исследования данных для выбранного сегмента.

Изменить задачу исследования данных

В диалоговом окне Создать/изменить задачу исследования данных представлены опции для определения новой задачи исследования данных или изменения существующей.

Большинство доступных для задачи исследования данных параметров те же, что предлагаются на узле Список решений. Исключения представлены ниже. Дополнительную информацию смотрите в разделе “Опции модели списка решений” на стр. 138.

Загрузить параметры: Если вы создали несколько задач исследования данных, выберите нужную задачу.

Создать... Щелкните здесь, чтобы создать новую задачу исследования данных на основе текущих показываемых параметров задачи.

Назначение

Поле назначения: Представляет собой поле, значение в котором вы хотите предсказать, используя предположение, что оно связано со значениями в других полях (предикторах).

Целевое значение. Указывает значение поля назначения, задающее выходные данные, которые вы хотите смоделировать. Например, если поле назначения churn закодировано как `0 = no` и `1 = yes`, задайте `1` для определения правил, в соответствии с которыми записи будут отнесены к вероятным для оттока клиентов.

Простые параметры

Максимальное число альтернативных вариантов. Задает количество альтернатив для вывода при запуске задачи исследования данных. Минимальный разрешенный параметр - 1; максимальный параметр не определяется.

Дополнительные параметры

Изменить... Открывает диалоговое окно **Изменить дополнительные параметры**, позволяющее определить расширенные настройки. Дополнительную информацию смотрите в разделе “Редактирование дополнительных параметров” на стр. 146.

Данные

Выбор построения. Предоставляет опции для задания показателя оценки, который должен будет анализироваться Средство просмотра списка решений для поиска новых правил. Перечисленные показатели оценки создаются и/или изменяются в диалоговом окне Организовать выбор данных.

Доступные поля. Предоставляет опции для вывода всех полей или для выбора вручную, какие поля выводить.

Изменить... Если выбрать опцию **Настроить**, откроется диалоговое окно **Настройка доступных полей**, в котором можно выбрать, какие поля будут доступны как атрибуты сегмента, обнаруженные задачей исследования данных. Дополнительную информацию смотрите в разделе “Настройка доступных полей”.

Редактирование дополнительных параметров: В диалоговом окне Изменение дополнительных параметров представлены следующие опции конфигурирования.

Метод разбиения на группы. Этот способ используется для разбиения на группы количественных полей (по равному количеству или по равной ширине).

Число интервалов. Количество интервалов, создаваемых для количественных полей. Минимальное разрешенное значение параметра - 2; максимальное значение не ограничивается.

Ширина поиска модели. Максимальное количество результатов модели на цикл, которые могут использоваться для следующего цикла. Минимальное разрешенное значение параметра - 1; максимальное значение не ограничивается.

Ширина поиска правила. Максимальное количество результатов правила на цикл, которые могут использоваться для следующего цикла. Минимальное разрешенное значение параметра - 1; максимальное значение не ограничивается.

Фактор слияния интервалов. Минимальный объем, на который должен вырасти сегмент при слиянии с соседним сегментом. Минимальное разрешенное значение параметра - 1,01; максимальное значение не ограничивается.

- **Разрешить пропущенные значения в условиях.** `True` - для разрешения критерия `IS MISSING` в правилах.
- **Отбросить промежуточные результаты.** При значении `True` возвращаются только окончательные результаты процесса поиска. Окончательный результат - это такой результат, который больше не обновляется в процессе поиска. При значении `False` возвращаются также промежуточные результаты.

Настройка доступных полей: В диалоговом окне Настройка доступных полей можно выбрать, какие поля будут доступны как атрибуты сегмента, обнаруженные задачей исследования данных.

Доступные. Перечисляет поля, доступные в настоящее время как атрибуты сегмента. Для удаления полей из этого списка выберите соответствующие поля и нажмите кнопку **Удалить >>**. Выбранные поля будут перемещены из списка Доступные в список Недоступные.

Недоступные. Перечисляет поля, которые недоступны как атрибуты сегмента. Чтобы включить эти поля в список доступных, выберите соответствующие поля и нажмите кнопку **<< Добавить**. Выбранные поля будут перемещены из списка Недоступные в список Доступные.

Организация выбора данных: Организуя выбор данных (наборы данных исследования), можно указать, какие показатели оценки должны использоваться Средство просмотра списка решений для поиска новых правил, и выбрать, какие наборы данных будут использоваться как основа для показателей.

Чтобы организовать выбор данных:

1. В меню Инструменты выберите пункт **Организация выбора данных** или щелкните правой кнопкой мыши по сегменту и выберите эту же опцию. Откроется диалоговое окно Организация выбора данных.

Примечание: В диалоговом окне Организация выбора данных можно также изменять и удалять существующий выбор данных.

2. Нажмите кнопку **Добавить новый выбор данных**. Запись о новом выборе данных добавится в существующую таблицу.
3. Нажмите кнопку **Имя** и введите соответствующее имя выбора.
4. Нажмите кнопку **Раздел** и выберите соответствующий тип раздела.
5. Нажмите кнопку **Условие** и выберите соответствующую опцию условия. При выборе опции **Задать** откроется диалоговое окно Задать условие выбора, предоставляющее возможность определения условий для конкретных полей.
6. Определите соответствующее условие и нажмите кнопку **OK**.

Варианты выбора данных доступны в выпадающем списке Выбор построения в диалоговом окне Создать/изменить задачу исследования данных. Этот список позволяет выбрать, какой показатель оценки будет использоваться для конкретной задачи исследования данных.

Правила сегментов

Правила сегментов модели можно найти, запустив задачу исследования данных на основе шаблона задач. Можно вручную добавить в модель правила сегментов, используя функции Вставить сегмент или Изменить правило сегмента.

Если выбрать поиск правил сегментов через исследование данных, результаты, если они будут, выводятся на вкладке Программа просмотра диалогового окна Интерактивный список. Можно быстро уточнить вашу модель, выбрав один из альтернативных результатов в диалоговом окне Альбомы моделей и нажав кнопку **Загрузить**. Так можно экспериментировать с различными результатами, пока вы не будете готовы построить модель, точно описывающую оптимальную группу назначения.

Вставка сегментов: Используя функцию вставки сегмента, можно вручную добавить к модели правила сегментов.

Чтобы добавить условие правила сегмента к модели:

1. В диалоговом окне Интерактивный список выберите положение, в котором вы хотите добавить новый сегмент. Новый сегмент будет добавлен непосредственно над выбранным сегментом.
2. В меню Изменить выберите пункт **Вставить сегмент** или щелкните правой кнопкой мыши по сегменту для доступа к выбору этой опции.

   Откроется диалоговое окно Вставить сегмент, в котором можно вставить условия правил нового сегмента.
3. Нажмите кнопку **Вставить**. Откроется диалоговое окно Вставить условие, позволяющее определить атрибуты для нового условия правила.
4. Выберите поле и оператор из раскрывающихся списков.

   Примечание: При выборе оператора **Not in** выбранное условие будет действовать как условие исключения и выводиться в диалоговом окне Вставить правило красным шрифтом. Например, когда условие `region = 'TOWN'` показано красным цветом, это означает, что `TOWN` исключен из набора результатов.
5. Введите одно или несколько значений или нажмите кнопку **Вставить значение**, чтобы открыть диалоговое окно Вставить значение. В этом диалоговом окне можно выбрать значение, определенное для заданного поля. Например, поле **состоит в браке** предоставит для выбора значения **да** и **нет**.
6. Нажмите кнопку **OK**, чтобы вернуться в диалоговое окно Вставить сегмент. Нажмите кнопку **OK** еще раз, чтобы добавить созданный сегмент в модель.

Новый сегмент будет выведен на экран в указанном положении модели.

Редактирование правил сегмента: Функциональные возможности диалогового окна Изменить правило сегмента позволяют добавлять, изменять и удалять условия правил сегментов.

Чтобы изменить условие правила сегмента:

1. Выберите сегмент модели, который вы хотите изменить.
2. В меню Правка выберите пункт **Изменить правило сегмента** или щелкните правой кнопкой мыши по правилу для доступа к выбору этой опции.

   Откроется диалоговое окно Изменить правило сегмента.
3. Выберите соответствующее условие правила и нажмите кнопку **Изменить**.

   Откроется диалоговое окно Изменить условие, позволяющее определить атрибуты для выбранного условия правила.
4. Выберите поле и оператор из выпадающих списков.

   Примечание: При выборе оператора **Not in** выбранное условие будет действовать как условие исключения и выводиться в диалоговом окне Изменить правило сегмента красным шрифтом. Например, когда условие `region = 'TOWN'` показано красным цветом, это означает, что `TOWN` исключен из набора результатов.
5. Введите одно или несколько значений или нажмите кнопку **Вставить значение**, чтобы открыть диалоговое окно Вставить значение. В этом диалоговом окне можно выбрать значение, определенное для заданного поля. Например, поле **состоит в браке** предоставит для выбора значения **да** и **нет**.
6. Нажмите кнопку **OK**, чтобы вернуться в диалоговое окно Изменить правило сегмента. Нажмите кнопку **OK** еще раз, чтобы вернуться в рабочую модель.

Выбранный сегмент будет показан с измененными условиями правил.

Удаление условий правил сегментов: **Чтобы удалить условие правил сегментов:**

1. Выберите сегмент модели, содержащий условия правил, которые вы хотите удалить.
2. В меню Изменить выберите пункт **Изменить правило сегмента** или щелкните правой кнопкой мыши по сегменту для доступа к выбору этой опции.

   Откроется диалоговое окно Изменение правила сегмента, в котором можно удалить одно или несколько условий правил сегментов.
3. Выберите соответствующее условие правил и нажмите кнопку **Удалить**.
4. Щелкните по **OK**.

Удаление одного или нескольких условий правил сегментов вызывает обновление показателей на панели рабочей модели.

Копирование сегментов: Средство просмотра списка решений предоставляет удобный способ копирования сегментов моделей. Когда вы хотите применить сегмент из одной модели к другой модели, просто скопируйте (или вырежьте) его из одной модели и вставьте из буфера памяти в другую. Вы можете скопировать также сегмент модели, показанный на панели Предварительный просмотр альтернативы и вставить его из буфера памяти в модель, показанную на панели рабочей модели. Эти функции копирования, вырезания и вставки используют системный буфер обмена для хранения или извлечения временных данных. Это означает, что условия и назначения копируются в буфере памяти. Содержимое буфера обмена не зарезервировано для использования исключительно в Средство просмотра списка решений, его можно вставлять также и в другие прикладные программы. Например, при вставке содержимого буфера памяти в текстовый редактор условия и назначение вставляются в формате XML.

Чтобы скопировать или вырезать сегменты модели:

1. Выберите сегмент модели, который вы хотите использовать в другой модели.
2. В меню Изменить выберите пункт **Копировать** (или **Вырезать**), можно также щелкнуть правой кнопкой мыши по сегменту модели и после этого выбрать **Копировать** или **Вырезать**.
3. Откройте соответствующую модель (куда будет вставлен сегмент модели).
4. Выберите один из сегментов модели и щелкните по **Вставить**.

Примечание: Вместо команд **Вырезать**, **Копировать** и **Вставить** можно использовать также комбинации клавиш: **Ctrl+X**, **Ctrl+C** и **Ctrl+V**.

Скопированный (или вырезанный) сегмент вставляется над ранее выбранным сегментом модели. Показатели вставленного сегмента и сегментов ниже вычисляются повторно.

Примечание: Обе модели в этой процедуре должны основываться на одном базовом шаблоне модели и содержать одинаковое поле назначения, в противном случае появится сообщение об ошибке.

Альтернативные модели: В тех случаях, когда существует несколько результатов, на вкладке Альтернативы выводятся результаты каждой задачи исследования данных. Каждый результат состоит из условий в выбранных данных, которые дают наиболее точное соответствие с полем назначения, а также все "достаточно хорошие" альтернативы. Общее количество выводимых альтернатив зависит от критериев поиска, использованных в процессе анализа.

Чтобы просмотреть альтернативные модели:

1. Щелкните по альтернативной модели на вкладке Альтернативы. На панель Предварительный просмотр альтернативы сегменты альтернативной модели появятся или заменят сегменты текущей модели.
2. Для работы с альтернативной моделью на панели рабочей модели выберите модель, щелкните по **Загрузить** на панели Предварительный просмотр альтернативы или щелкните правой кнопкой мыши по имени альтернативной модели на вкладке Альтернативы и выберите **Загрузить**.

Примечание: При генерировании новой модели альтернативные модели не сохраняются.

Настройка модели

Данные не статичны. Клиенты переезжают, вступают в брак, меняют место работы. Товары теряют рыночную привлекательность и устаревают.

Средство просмотра списка решений обеспечивает гибкость бизнес-пользователей для простой и быстрой адаптации моделей в новых ситуациях. Вы можете изменять модель, редактируя, удаляя, деактивируя конкретные сегменты модели или изменяя их приоритет.

Приоритизация сегментов: Ранжировать правила модели можно в любом выбранном порядке. По умолчанию сегменты модели выводятся в порядке приоритета, причем у первого сегмента самый высокий приоритет. При назначении одному или нескольким сегментам других приоритетов соответствующим образом меняется модель. Вы можете изменить нужным образом модель, перемещая сегменты на уровень более высокого или более низкого приоритета.

Для определения приоритета сегментов модели:

1. Выберите сегмент модели, которому вы хотите назначить другой приоритет.
2. Нажмите одну из кнопок со стрелками на панели инструментов рабочей модели, чтобы переместить заданный сегмент модели вверх или вниз по списку.

После приоритизации все предыдущие результаты оценок будут вычислены повторно, и будут выведены новые значения.

Удаление сегментов: Чтобы удалить один или несколько сегментов:

1. Выберите сегмент модели.
2. В меню Изменить выберите пункт **Удалить сегмент** или нажмите кнопку Удалить на панели инструментов панели рабочей модели.

Показатели будут вычислены для измененной модели повторно, а модель будет изменена соответствующим образом.

Исключение сегментов: Возможно, при поиске конкретных групп вы будете основывать бизнес-действия на выборе сегментов модели. При внедрении модели можно выбрать опцию исключения некоторых ее сегментов. Исключенные сегменты оцениваются как пустые значения. Такое исключение не означает, что сегмент не используется; смысл исключения в том, что все удовлетворяющие этому правилу записи исключаются из списка рассылки. Правило по-прежнему применимо, но иначе.

Чтобы исключить конкретные сегменты модели:

1. Выберите сегмент на панели рабочей модели.
2. Нажмите кнопку **Переключить исключение сегмента** на панели инструментов в составе панели рабочей модели. В выбранном столбце Назначение выбранного сегмента теперь появится надпись **Исключено**.

Примечание: В отличие от удаленных сегментов, исключенные сегменты остаются доступными для использования в конечной модели. Исключенные сегменты влияют на результаты диаграмм.

Изменение значения назначения: В диалоговом окне Изменить значение назначения можно изменить значение назначения для текущего поля назначения.

Снимки и результаты сеансов со значением назначения, отличным от заданного в рабочей модели, обозначаются изменением фона в таблице для соответствующей строки на желтый. Это указывает, что снимок/результат сеанса устарели.

В диалоговом окне **Создать/изменить задачу исследования данных** выводится значение назначения для текущей рабочей модели. Это значение назначения не сохраняется в задаче исследования данных. Вместо этого значение берется с панели Рабочая модель.

При продвижении сохраненной модели на панель Рабочая модель со значением назначения, отличающимся от соответствующего значения рабочей модели (например, из-за редактирования альтернативного результата или копии снимка), значение назначения сохраненной модели заменяется для совпадения со значением назначения рабочей модели (значение назначения, показанное на панели Рабочая модель, не изменяется). Показатели модели переоцениваются с новым значением назначения.

Генерирование новой модели

В диалоговом окне Генерирование новой модели представлены опции для именования модели и выбора, где будет создан новый узел.

Имя модели. Выберите **Настроить**, чтобы подправить автоматически сгенерированное имя или создать уникальное имя для узла, как оно будет показываться на холсте потока.

Создать узел на ... При выборе **Холст** новая модель будет расположена на рабочем холсте; при выборе **Палитра GM** - на палитре Модели; выбор опции **Оба положения** располагает новую модель и на рабочем холсте, и на палитре Модели.

Включить интерактивное состояние сеанса. При включении этой опции интерактивное состояние сеанса сохраняется в сгенерированной модели. Когда позже вы сгенерируете из модели узел моделирования, это состояние перенесется туда и будет использоваться для инициализации интерактивного сеанса. Независимо от выбора этой опции сама модель оценивает новые данные одинаково. Если не выбирать эту опцию, модель все равно сможет создать узел построения, но он будет более общим и будет запускать новые интерактивные сеансы, а не продолжать старые сеансы с того момента, когда они были остановлены. Если вы изменяете параметры узла, но запускаете его с сохраненным состоянием, эти измененные параметры будут игнорироваться и будут использоваться параметры сохраненного состояния.

Примечание: Стандартные показатели - это единственные показатели, которые остаются в модели. Дополнительные показатели сохраняются с интерактивным состоянием. Сгенерированная модель не представляет сохраненного интерактивного состояния задачи исследования данных. При запуске Средство просмотра списка решений выводятся параметры, исходно заданные в программе просмотра.

Дополнительную информацию смотрите в разделе “Повторное генерирование узла моделирования” на стр. 48.

Оценка модели

Успешное моделирование требует тщательной оценки модели перед ее реализацией в производственной среде. Средство просмотра списка решений предоставляет несколько статистических и бизнес-показателей, которые можно использовать для оценки воздействия модели в реальных условиях. Сюда входят диаграммы выигрышей и полная функциональная совместимость с Excel, что позволяет обсчитывать сценарии стоимость/преимущества для оценки воздействия внедрения.

Модели можно оценивать следующими способами:

- С использованием предварительно определенных показателей статистических и бизнес-моделей, доступных в Средство просмотра списка решений (вероятность, частота).
- Оценивая показатели, импортированные из Microsoft Excel.
- Визуализируя модель с использованием диаграммы выигрыша.

Организация показателей модели: Средство просмотра списка решений предоставляет опции для определения показателей, которые вычисляются и выводятся в виде столбцов. Каждый сегмент может включать в себя показатели покрытия, частоты, вероятности и ошибки по умолчанию, представленные столбцами. Вы можете создать и новые показатели, которые будут выводиться как столбцы.

Определение показателей моделей

Чтобы добавить показатель к вашей модели или определить существующий показатель:

1. В меню Инструменты выберите пункт **Организовать показатели модели** или щелкните правой кнопкой мыши и сделайте этот выбор. Откроется диалоговое окно Организовать показатели модели.
2. Нажмите копку **Добавить новый показатель модели** (справа от столбца Показать). Новый показатель будет представлен в таблице.
3. Дайте имя показателю, найдите подходящий тип, опцию вывода и выбор. Столбец Показать указывает, будет ли этот показатель выводиться для рабочей модели. Для определения существующего показателя найдите и выберите его, а также укажите, должен ли этот показатель выводиться для рабочей модели.
4. Нажмите кнопку **OK**, чтобы вернуться в рабочее пространство Средство просмотра списка решений. Если для нового показателя был отмечен столбец Показать, этот показатель будет выводиться для рабочей модели.

Пользовательские показатели в Excel

Дополнительную информацию смотрите в разделе “Оценка в Excel”.

Обновление показателей: В некоторых случаях может потребоваться повторно вычислить показатели модели, например, когда существующая модель применяется к новому набору клиентов.

Чтобы повторно вычислить (обновить) показатели модели:

В меню Изменить выберите пункт **Обновить все показатели**.

или

Нажмите клавишу F5.

Все показатели будут вычислены повторно, а для рабочей модели будут показаны новые значения.

Оценка в Excel: Средство просмотра списка решений можно интегрировать с Microsoft Excel, что позволяет использовать собственные вычисления значений и формулы прибыли непосредственно в процессе построения модели, чтобы воспроизводить сценарии стоимости/выигрыша. Связь с Excel позволяет экспортировать

данные в Excel, где их можно использовать для создания презентационных диаграмм, вычисления пользовательских показателей, таких как сложные показатели прибыли и возврата инвестиций, и просматривать все данные в Средство просмотра списка решений во время построения модели.

Примечание: Для возможности работы с электронными таблицами Excel аналитический эксперт CRM должен определить информацию конфигурации для синхронизации Средство просмотра списка решений с Microsoft Excel. Эта конфигурация содержится в файле электронной таблицы Excel и обозначает, какая информация передается из Средство просмотра списка решений в Excel и обратно.

Следующие действия возможны только в том случае, если установлен продукт MS Excel. Если Excel не установлен, опции для синхронизации моделей с Excel не выводятся.

Чтобы синхронизировать модели с MS Excel:

1. Откройте модель, запустите интерактивный сеанс и в меню Инструменты выберите пункт **Организовать показатели модели**.
2. Выберите **Да** для опции **Вычислить пользовательские показатели в Excel**. Активируется поле **Рабочая книга**, позволяющее вам выбрать предварительно сконфигурированный шаблон рабочей книги Excel.
3. Нажмите кнопку **Соединиться с Excel**. Откроется диалоговое окно Открыть, в котором можно перейти в положение предварительно сконфигурированного шаблона в вашей локальной или сетевой файловой системе.
4. Выберите соответствующий шаблон Excel и нажмите кнопку **Открыть**. Запустится выбранный шаблон Excel; используйте панель задач Windows (или нажмите сочетание клавиш Alt-Tab), чтобы вернуться назад в диалоговое окно Выбрать входные данные для пользовательских показателей.
5. Выберите соответствующие отображения между именами показателей, определенных в шаблоне Excel, и именами показателей модели и нажмите кнопку **OK**.

После установления связи Excel начинает работу со своим предварительно сконфигурированным шаблоном, где в электронной таблице выводятся правила модели. Вычисленные в Excel результаты выводятся как новые столбцы в Средство просмотра списка решений.

Примечание: показатели Excel не остаются доступными при сохранении модели, их можно использовать только в активном сеансе. Однако вы можете создать снимки, включающие в себя показатели Excel. Показатели Excel, сохраненные в представлениях снимков, подходят только для целей хронологического сравнения и не обновляются при повторном открытии. Дополнительную информацию смотрите в разделе “Вкладка Снимки” на стр. 143. Показатели Excel не будут выводиться в снимках, пока вы не установите повторное соединение с шаблоном Excel.

Конфигурирование интеграции с MS Excel: Интеграция Средство просмотра списка решений с Microsoft Excel осуществляется путем использования предварительно сконфигурированного шаблона электронной таблицы Excel. Этот шаблон состоит из трех рабочих листов:

Показатели модели. Выводит импортированные показатели Средство просмотра списка решений, пользовательские показатели Excel и итоги вычислений (определенные на рабочем листе Параметры).

Параметры. Содержит переменные для вычислений на основе импортированных показателей Средство просмотра списка решений и пользовательских показателей Excel.

Конфигурация. Содержит опции для указания, какие именно показатели импортируются из Средство просмотра списка решений, и для определения пользовательских показателей Excel.

ПРЕДУПРЕЖДЕНИЕ: Структура рабочего листа Конфигурация строго определена. **НЕ** изменяйте какие-либо ячейки в зеленой зоне.

- **Показатели из модели.** Указывает, какие показатели Средство просмотра списка решений используются для вычислений.

- **Показатели для модели.** Указывает, какие сгенерированные Excel показатели будут возвращены в Средство просмотра списка решений. Сгенерированные Excel показатели выводятся как новые столбцы показателей в Средство просмотра списка решений.

Примечание: Показатели Excel не остаются доступными для модели, когда вы генерируете новую модель; они доступны только в активном сеансе.

Изменение показателей модели: В следующих примерах объясняется, как можно несколькими способами изменить показатели модели:

- Изменение существующего показателя.
- Импорт из модели дополнительного стандартного показателя.
- Экспорт в модель дополнительного пользовательского показателя.

Изменение существующего показателя

1. Откройте шаблон и выберите рабочий лист электронной таблицы Конфигурация.
2. Измените любое **Имя** или **Описание**, выделив их и введя новый текст поверху.

   Обратите внимание на то, что если вы хотите изменить показатель, например, предложить пользователю Вероятность вместо Частоты, нужно только изменить имя и описание в **Метриках из модели**, эти изменения появятся в модели, и пользователь сможет выбрать соответствующий показатель для отображения.

Импорт из модели дополнительного стандартного показателя

1. Откройте шаблон и выберите рабочий лист электронной таблицы Конфигурация.
2. Выберите в меню:

   Инструменты > Защита > Снять защиту листа
3. Выберите ячейку A5 с желтым фоном, которая содержит слово **Конец**.
4. Выберите в меню:

   Вставить > Строки
5. Введите **Имя** и **Описание** нового показателя. Например, **Ошибка** и **Ошибка, связанная с сегментом**.
6. В ячейке C5 введите формулу **=COLUMN('Показатели модели'!N3)**.
7. В ячейке D5 введите формулу **=ROW('Показатели модели'!N3)+1**.

   Эти формулы приведут к тому, что новый показатель появится в столбце N рабочего листа Показатели модели, который в настоящее время пуст.
8. Выберите в меню:

   Инструменты > Защита > Защитить лист
9. Щелкните по **OK**.
10. Убедитесь, что у ячейки N3 рабочего листа Показатели модели появился заголовок нового столбца **Ошибка**.
11. Выберите все ячейки столбца N.
12. Выберите в меню:

    Формат > Ячейки
13. По умолчанию у всех ячеек категория числа - **Общая**. Нажмите кнопку **Процентная доля**, чтобы изменить вид вывода рисунков. Это помогает проверить ваши рисунки в Excel; кроме того, становится возможным другое использование данных, например, как выходных данных для графика.
14. Щелкните по **OK**.
15. Сохраните электронную таблицу как шаблон Excel 2003 с уникальным именем и расширением файла *.xlt* . Для простоты определения положения нового шаблона рекомендуется сохранить его в предварительно сконфигурированном для шаблонов положении в локальной или сетевой файловой системе.

Экспорт в модель дополнительного пользовательского показателя

1. Откройте шаблон, в который вы хотите добавить столбец Ошибка из предыдущего примера; выберите рабочий лист Конфигурация.
2. Выберите в меню:

   Инструменты > Защита > Снять защиту листа
3. Выберите ячейку A14 с желтым фоном, которая содержит слово **Конец**.
4. Выберите в меню:

   Вставить > Строки
5. Введите **Имя** и **Описание** нового показателя. Например, **Масштабированная ошибка** и **Масштабирование, примененное к ошибке из Excel**.
6. В ячейке C14 введите формулу **=COLUMN('Показатели модели'!O3)**.
7. В ячейке D14 введите формулу **=ROW('Показатели модели'!O3)+1**.

   Эти формулы определяют, что столбец O будет содержать новый показатель для модели.
8. Выберите рабочий лист Параметры.
9. В ячейке A17 введите описание **'- Масштабируемая ошибка**.
10. В ячейке B17 введите масштабный коэффициент **10**.
11. На рабочей странице Показатели модели введите описание **Масштабируемая ошибка** в ячейке O3 как заголовок для нового столбца.
12. В ячейке O4 введите формулу **=N4*Параметры!B17**.
13. Выберите угол ячейки O4 и перетащите эту ячейку вниз до ячейки O22, чтобы скопировать формулу в эту ячейку.
14. Выберите в меню:

    Инструменты > Защита > Защитить лист
15. Щелкните по **OK**.
16. Сохраните электронную таблицу как шаблон Excel 2003 с уникальным именем и расширением файла *.xlt* . Для простоты определения положения нового шаблона рекомендуется сохранить его в предварительно сконфигурированном для шаблонов положении в локальной или сетевой файловой системе.

При обращении к Excel с использованием этого шаблона значение Ошибка будет доступно как новый пользовательский показатель.

Визуализация моделей

Лучшая возможность понять воздействие модели - визуализировать ее. Используя диаграмму выигрыша, можно получить ценную возможность взгляда изнутри на ежедневный бизнес и технический выигрыш от применения вашей модели, изучая эффект нескольких альтернатив в реальном времени. В разделе “Диаграмма выигрыша” показаны преимущества модели перед случайно принятыми решениями; здесь же можно непосредственно сравнить несколько диаграмм для альтернативных моделей.

Диаграмма выигрыша: Диаграмма выигрыша представляет значения из столбца *% выигрыша* таблицы. Выигрыш определяется как отношение числа попаданий на каждом делении шкалы к общему числу попаданий в дереве с использованием следующей формулы:

(число попаданий на инкремент / общее число попаданий) x 100%

Диаграмма выигрыша наглядно иллюстрирует, насколько широко по дереву нужно производить поиск для набора данного процента всех попаданий. Диагональная линия показывает ожидаемый отклик для всей выборки, если модель не используется. В этом случае показатель отклика был бы константой, так как вероятность отклика для одного клиента такая же, как для другого. Для удвоения собранной информации нужно бы было опросить вдвое больше людей. Кривая линия обозначает, насколько можно повысить отклик, включив только тех клиентов, которые на основании выигрыша ранжируются в максимальных процентилях. Например, включив в рассылку верхние 50%, можно бы было собрать больше 70% положительных откликов. Чем круче кривая, тем больше выигрыш.

Чтобы просмотреть диаграмму выигрыша:

1. Откройте поток, содержащий узел Список решений, и запустите интерактивный сеанс с этого узла.
2. Перейдите на вкладку **Выигрыш**. В зависимости от того, какие разделы вы указали, можно увидеть одну или две диаграммы (две диаграммы будут показаны, например, когда и обучающий, и проверочный раздел заданы для показателей модели).

По умолчанию диаграммы выводятся по сегментам. Можно переключить диаграммы на вывод квантилей, выбрав **Квантили**, а затем выбрав соответствующий способ определения квантилей из раскрывающегося меню.

Параметры диаграммы: Возможность Параметры диаграмм предоставляет опции для выбора, какие модели и снимки будут представлены на диаграммах, для каких разделов они будут построены и выводить или нет метки сегментов.

Модели для вывода на диаграммах

Текущие модели. Позволяет выбрать, какие модели будут представлены на диаграммах. Вы можете выбрать рабочую модель или любые созданные модели снимков.

Разделы для вывода на диаграммы

Разделы для левой диаграммы. В выпадающем списке представлены опции для вывода всех определенных разделов или всех данных.

Разделы для правой диаграммы. В выпадающем списке представлены опции для вывода всех определенных разделов, всех данных или только левой диаграммы. При выборе опции **Рисовать только слева** выводится только левая диаграмма.

Отображать метки сегментов. При включении этой опции на диаграммах выводятся метки всех сегментов.

Глава 10. Статистические модели

Статистические модели используют математические уравнения для кодирования информации, извлеченной из данных. В некоторых случаях методы статистического моделирования могут очень быстро предоставить адекватные модели. Даже для задач, в которых более гибкие способы машинного обучения (такие как нейронные сети) могут в конечном счете дать лучшие результаты, можно использовать некоторые статистические модели в качестве базовых моделей предсказания, чтобы судить о характеристиках более продвинутых способов.

Доступны следующие узлы статистического моделирования.

Модели линейной регрессии предсказывают значения непрерывного целевого поля на основе линейных взаимосвязей между целевым полем и одним или несколькими предикторами.

Логистическая регрессия - это статистический метод для классификации записей на основании значений входных полей. Она аналогична линейной регрессии, но логистическая регрессия использует категориальные поля назначения вместо численных.

Узел PCA/фактора предоставляет мощные средства сокращения числа данных для уменьшения сложности ваших данных. Анализ главных компонентов (principal components analysis, PCA) находит линейные комбинации входных полей, которыми главным образом определяются изменения в целом наборе полей, где компоненты ортогональны друг другу. Факторный анализ направлен на выявление скрытых факторов, объясняющих структуру корреляций в наборе наблюдаемых полей. Цель обоих подходов - найти небольшое количество производных полей, которые эффективно суммируют информацию исходного набора входных полей.

Дискриминантный анализ делает более строгие предположения, чем логистическая регрессия, но он может быть ценной альтернативой или дополнением к анализу логистической регрессии, когда эти предположения оказываются правильными.

Обобщенная линейная модель расширяет общую линейную модель, так что зависимая переменная считается линейно связанной с факторами и ковариатами через заданную функцию связи. Более того, модель допускает наличие у зависимой переменной распределения, отличающегося от нормального. Она включает в себя функциональные возможности большого количества статистических моделей, в том числе линейной регрессии, логистической регрессии, логлинейных моделей для количества данных и интервал-цензурированных моделей выживания.

Обобщенная линейная смешанная модель (generalized linear mixed model, GLMM) обобщает линейную модель таким образом, что у значений назначения может быть отличное от нормального распределение и оно будет линейно связано с факторами и ковариатами через задаваемую функцию связи, так что наблюдения могут быть скоррелированными. Обобщенные линейные смешанные модели включают широкий набор моделей, начиная от простой линейной регрессии и кончая сложными многоуровневыми моделями для не нормально распределенных данных с повторными измерениями.

Узел регрессии Кокса позволяет построить модель дожития для данных времени-до-события в присутствии цензурируемых записей. Эта модель создает функцию дожития, которая предсказывает вероятность, что изучаемое событие произойдет в данное время (*t*) для данных значений входных переменных.

Узел линейной модели

Линейная регрессия - это обычный статистический метод для классификации записей на основании значений числовых входных полей. Линейная регрессия подгоняет прямую линию или поверхность, минимизирующую разности между предсказанными и фактическими выходными значениями.

Требования. В моделях линейной регрессии можно использовать только числовые поля. У вас должно быть только одно поле назначения (с заданной ролью *Назначение*) и один или несколько предикторов (в заданной ролью *Ввод*). Поля с заданной ролью *Обе* или *Нет* игнорируются, как и нечисловые поля. (При необходимости нечисловые поля можно перекодировать с использованием узла Производные данные).

Достоинства. Модели линейной регрессии относительно просты и дают легко интерпретируемую математическую формулу для генерирования предсказаний. Так как линейная регрессия - это давно установившаяся статистическая процедура, свойства этих моделей хорошо известны. Обычно линейные модели очень быстро обучаются. На узле линейных моделей предоставляются способы автоматического выбора полей для исключения несущественных входных полей из уравнения.

Примечание: В тех случаях, когда поле назначения категориальное, а не в количественном диапазоне, например, *да*/*нет* или *уйдет*/*не уйдет*, в качестве альтернативы можно использовать логистическую регрессию. Логистическая регрессия обеспечивает также поддержку нечисловых входных полей, исключая необходимость их перекодирования. Дополнительную информацию смотрите в разделе “Логистический узел” на стр. 165.

Линейные модели

Линейные модели предсказывают значения непрерывных целевых переменных, основываясь на взаимосвязи между целевой переменной и одним или несколькими предикторами.

Линейные модели относительно просты и дают легко интерпретируемую математическую формулу для скоринга. Свойства этих моделей хорошо понятны, и их обычно можно построить очень быстро, по сравнению с моделями других типов (такими как нейронные сети или деревья решений) на том же наборе данных.

Пример. Страховая компания с ограниченными ресурсами для исследования страховых требований домовладельцев желает построить модель для оценки стоимости требований. Применяя эту модель в центрах обслуживания, сотрудники компании могут ввести информацию от требовании, разговаривая по телефону с клиентом, и немедленно получить "ожидаемую" стоимость требования, основываясь на прошлых данных. Дополнительную информацию смотрите в разделе .

Требования к полям. Должны быть целевое и, по крайней мере, одно входное поля. По умолчанию не используются поля с предопределенными ролями Двойного назначения и Нет. Целевое поле должно быть непрерывным (количественным). Для предикторов (входов) отсутствуют ограничения на тип измерений; категориальные поля (флаговые, номинальные и порядковые) используются в модели в качестве факторов, а непрерывные поля используются как ковариаты.

Цели

Что вы хотите сделать?

- **Построить новую модель.** Построить совершенно новую модель. Это обычная операция узла.

- **Продолжить обучение существующей модели.** Обучение продолжается с последней моделью, успешно сгенерированной узлом. Это дает возможность скорректировать или обновить существующую модель без необходимости обращаться к исходным данным, что может выполняться значительно быстрее, так как в поток вводятся только новые или обновленные записи. Информация по предыдущей модели сохраняется вместе с узлом моделирования, что позволяет использовать этот вариант, даже если предыдущий слепок модели недоступен в потоке или Палитре моделей.

  Примечание: Когда этот вариант доступен, все остальные управляющие элементы на вкладках Поля и Параметры конструкции блокируются.

Какова ваша главная цель? Выберите подходящую цель.

- **Создать стандартную модель.** Данный метод строит единичную модель для предсказания целевой переменной, используя предикторы. Вообще говоря, стандартные модели легче поддаются интерпретации и могут требовать меньше времени при скоринге, чем построенные с применением бустинга, бэггинга или ансамблей больших наборов данных.
- **Повысить точность модели (бустинг).** Данный метод строит модель ансамбля, используя бустинг, который генерирует последовательность моделей для получения более точных предсказаний. Ансамбли могут занять больше времени для их построения и скоринга, чем стандартная модель.

  Бустинг генерирует последовательность "компонентных моделей", каждая из которых строится по целому набору данных. Прежде чем строить каждую последовательную компонентную модель, записи взвешиваются на основе остатков для предшествующей компонентной модели. Наблюдениям с большими остатками придаются относительно большие веса прецедентов, с тем чтобы следующая компонентная модель была сконцентрирована на том, чтобы хорошо предсказывать такие записи. Вместе такие компонентные модели образуют модель ансамбля. Модель ансамбля выполняет скоринг новых записей, пользуясь правилом объединения; доступные правила зависят от типа измерений целевой переменной.
- **Повысить стабильность модели (бэггинг).** Данный метод строит модель ансамбля, используя бэггинг (бутстреп-агрегирование), который генерирует множественные модели для получения более надежных предсказаний. Ансамбли могут занять больше времени для их построения и скоринга, чем стандартная модель.

  Бутстреп-агрегирование (бэггинг) формирует реплики обучающего набора данных путем выбора с возвращением из исходного набора данных. В результате создаются бутстреп-выборки исходного набора данных равного объема. Затем по каждой реплике формируется "компонентная модель". Вместе такие компонентные модели образуют модель ансамбля. Модель ансамбля выполняет скоринг новых записей, пользуясь правилом объединения; доступные правила зависят от типа измерений целевой переменной.
- **Создать модель для очень больших наборов данных (требует сервер IBM SPSS Modeler).** Данный метод строит модель ансамбля путем расщепления набора данных на отдельные блоки данных. Выберите этот вариант, если ваш набор данных слишком велик для построения моделей перечисленных выше, или для инкрементного построения модели. Данный вариант может потребовать меньше времени для построения, но больше времени для скоринга, чем стандартная модель. Для этой опции требуется соединение с сервер IBM SPSS Modeler .

Информацию о параметрах, связанных с бустингом, бэггингом и очень большими наборами данных, смотрите в разделе “Ансамбли” на стр. 161.

Основные параметры

Автоматически подготовить данные. Этот параметр позволяет процедуре выполнить внутренние преобразования целевой переменной и предикторов, чтобы максимизировать прогностическую силу модели. Все преобразования сохраняются вместе с моделью и применяются к новым данным при скоринге. Исходные версии преобразованных полей исключаются из модели. По умолчанию выполняются автоматические преобразования данных, описанные ниже.

- **Обработка дат и времени.** Каждый предиктор, являющейся переменной дат, преобразуется в новый непрерывный предиктор, содержащий время, прошедшее, начиная с опорной даты (1970-01-01). Каждый предиктор, являющийся переменной времени, преобразуется в новый непрерывный предиктор, содержащий время, прошедшее, начиная с опорного момента времени (00:00:00).

- **Корректировка шкалы измерений.** Непрерывные предикторы, содержащие менее 5 различных значений, преобразуются в порядковые предикторы. Порядковые предикторы, содержащие более 10 различных значений, преобразуются в непрерывные предикторы.
- **Обработка выбросов.** Значения непрерывных предикторов, которые лежат вне границ отсечения (определяемых тремя стандартными отклонениями от среднего значения), заменяются значением границы отсечения.
- **Обработка пропущенных значений.** Пропущенные значения номинальных предикторов заменяются модой обучающего разбиения. Пропущенные значения порядковых предикторов заменяются медианой обучающего разбиения. Пропущенные значения непрерывных предикторов заменяются средним значением обучающего разбиения.
- **Контролируемое объединение.** Эта операция делает модель более "экономной" путем уменьшения числа полей, обрабатываемых в связи с целевым полем. Идентифицируются подобные категории, основываясь на взаимосвязи между входным и целевым полями. Категории, которые не различаются значимо (т.е. имеющие p-значение больше 0,1), объединяются. Если все категории объединяются в одну, то исходная и полученная версии поля исключаются из модели, поскольку они не представляют ценности как предиктор.

Доверительный уровень. Это доверительный уровень, используемый при вычислении интервальных оценок коэффициентов модели, представленных на панели Коэффициенты. Задайте значение больше 0 и меньше 100. Значение по умолчанию - 95.

Подбор модели

Метод подбора модели. Выберите один из методов подбора модели (подробности ниже) или **Включить все предикторы**, когда все имеющиеся предикторы просто вводятся в модель как члены главных эффектов. По умолчанию используется **Прямой шаговый** .

Прямой шаговый отбор. Этот метод начинает работу с модели без эффектов, добавляя и удаляя эффекты по одному на каждом шаге до тех пор, пока ни один эффект нельзя будет добавить, руководствуясь критериями шагового отбора.

- **Критерии для включения/исключения.** Это статистика, используемая для определения того, следует ли эффект добавить в модель или исключить из нее. **Информационный критерий (AICC)** основывается на правдоподобии обучающего множества для данной модели и скорректирован с целью штрафовать излишне сложные модели. **F-статистики** основывается на статистическом критерии снижения модельной ошибки. **Скорректированный R-квадрат** основывается на точности подгонки для обучающего множества и скорректирован с целью штрафовать излишне сложные модели. **Критерий предотвращения переобучения (СКО)** основывается на точности подгонки (среднем квадрате ошибки или СКО) для множества предотвращения переобучения. Множество предотвращения переобучения представляет собой случайную подвыборку, содержащую приблизительно 30% наблюдений из исходного набора данных, которая не используется при обучении модели.

  Если выбран любой критерий, отличный от **F-статистики** , то на каждом шаге в модель добавляется эффект, соответствующий максимальному положительному приращению значения критерия. Все эффекты в модели, соответствующие уменьшению значения критерия, удаляются.

  Если в качестве критерия выбран **F-статистики** , то на каждом шаге в модель добавляется эффект, дающий наименьшее *p*-значение, при условии, что оно меньше порогового значения, заданного в **Включать эффекты с p-значениями, меньшими чем**. Значение по умолчанию - 0,05. Все эффекты в модели с *p*-значением, превосходящим пороговое значение, заданное в **Исключать эффекты с p-значениями, большими чем**, удаляются. Значение по умолчанию равно 0.10.
- **Задать максимальное число эффектов в окончательной модели.** По умолчанию все имеющиеся эффекты могут быть включены в модель. Как альтернатива, если шаговый алгоритм, заканчивая работу на некотором шаге, имеет заданное максимальное число эффектов в модели, то он останавливает работу, сохраняя текущий набор эффектов.
- **Задать максимальное число шагов.** Шаговый алгоритм останавливается после определенного числа шагов. По умолчанию это утроенное число имеющихся эффектов. Как альтернатива, задайте положительное целое для максимума числа шагов.

Выбор наилучших подмножеств. Проверяются "все возможные" модели или, по крайней мере, большая совокупность возможных моделей, чем при прямом пошаговом отборе, для выбора наилучших в соответствии с критерием наилучших подмножеств. **Информационный критерий (AICC)** основывается на правдоподобии обучающего множества для данной модели и скорректирован с целью штрафовать излишне сложные модели. **Скорректированный R-квадрат** основывается на точности подгонки для обучающего множества и скорректирован с целью штрафовать излишне сложные модели. **Критерий предотвращения переобучения (СКО)** основывается на точности подгонки (среднем квадрате ошибки или СКО) для множества предотвращения переобучения. Множество предотвращения переобучения представляет собой случайную подвыборку, содержащую приблизительно 30% наблюдений из исходного набора данных, которая не используется при обучении модели.

В качестве наилучшей модели выбирается модель с наибольшим значением критерия.

Примечание: Выбор наилучших подмножеств требует большего объема вычислений, чем прямой шаговый отбор. Когда выполняется выбор наилучших подмножеств в сочетании с бустингом, бэггингом или очень большими наборами данных, то для построения модели потребуется значительно больше времени, чем при построении стандартной модели с использованием прямого пошагового отбора.

Ансамбли

Данные параметры определяют поведение ансамбля, которое имеет место, когда на вкладке Цели запрашивается бэггинг, бустинг или очень большие наборы данных. Параметры, которые не применяются к выбранной цели, игнорируются.

Бэггинг и очень большие наборы данных. Это правило, которое применяется при скоринге ансамбля, чтобы объединить предсказанные значения для базовых моделей с целью вычисления значений скоринга для ансамбля.

- **Принятое по умолчанию правило объединения для непрерывных целевых полей.** Предсказанные значения для ансамбля в случае непрерывных целевых полей могут быть вычислены с использованием среднего значения или медианы предсказанных значений для базовых моделей.

Обратите внимание на то, что если цель состоит в повышении точности модели, выбор правила объединения игнорируется. При бустинге всегда используется взвешенное решение большинством голосов для скоринга категориальных целевых полей и взвешенная медиана для скоринга непрерывных целевых полей.

Бустинг и бэггинг. Задайте число базовых моделей для построения, когда целью является повышение точности или стабильности; для бэггинга это число бутстреп-выборок. Оно должно быть положительным целым.

Дополнительные параметры

Воспроизвести результаты. Задание стартового числа генератора псевдослучайных чисел позволяет воспроизвести результаты. Генератор псевдослучайных чисел используется для выбора записей, попадающих в множество предотвращения переобучения. Задайте целое число или щелкните по **Генерировать**, чтобы сгенерировать псевдослучайное целое число в диапазоне между 1 и 2147483647 включительно. Значение по умолчанию - 54752075.

Опции модели

Имя модели. Имя модели можно сгенерировать автоматически на основе целевых полей или задать самостоятельно. Автоматически генерируемое имя является именем целевого поля.

Обратите внимание на то, что при скоринге модели всегда вычисляется предсказанное значение. Имя нового поля есть имя целевого поля с префиксом *$L-*. Например, при имени целевого поля *sales*, новому полю было бы присвоено имя *$L-sales*.

Сводка для модели

{\f3 Вид Сводка для модели} {\f4 - } {\f3 это мгновенная визуальная сводка по модели и ее подгонке.}

Таблица. Данная таблица отображает некоторые установки высокого уровня для модели, включая:

- Имя назначения, указанного на вкладке Поля
- Выполнялась ли автоматическая подготовка, заданная в разделе Основные параметры
- Метод и критерий выбора модели, указанные в разделе параметров Выбор модели. Выводится также значение критерия отбора для окончательной модели и представляется в форме "меньше значит лучше".

Диаграмма. Данная диаграмма показывает точность окончательной модели, представленную в форме "больше значит лучше". Ее значение равно $100 \times$ скорректированный R^2 для окончательной модели.

Автоматическая подготовка данных

Этот вид выводит информацию о том, какие поля были исключены и как преобразованные поля были получены на этапе автоматической подготовки данных (ADP). Для каждого поля, которое было преобразовано или исключено, в таблице перечисляется имя поля, его роль в анализе и действие, совершенное на этапе ADP. Поля сортируются в алфавитном порядке имен полей по возрастанию. Возможные действия, выполняемые для каждого поля, включают:

- **Вычислить продолжительность: в месяцах** вычисляет истекшее время в месяцах, исходя из значений в поле, содержащем даты, до текущей системной даты.
- **Вычислить продолжительность: в часах** вычисляет истекшее время в часах, исходя из значений в поле, содержащем время, до текущего системного времени.
- **Сменить тип измерений с непрерывного на порядковый** преобразует непрерывные поля с менее чем 5 различных значений в порядковые поля.
- **Сменить тип измерений с порядкового на непрерывный** преобразует порядковые поля с более чем 10 различных значений в непрерывные поля.
- **Урезать выбросы** заменяет значения непрерывных предикторов, которые лежат вне границ отсечения (определяемых тремя стандартными отклонениями от среднего значения), значением границы отсечения.
- **Заменить пропущенные значения** заменяет пропущенные значения номинальных полей модой, порядковых полей медианой, а непрерывных полей средним значением.
- **Объединить категории для максимизации взаимосвязи с целевым полем** выявляет "похожие" категории предикторов на основе взаимосвязи между входными и целевой переменными. Категории, которые не различаются значимо (т.е. имеющие p-значение больше 0,05), объединяются.
- **Исключить предиктор-константу / после обработки пропущенных значений / после объединения категорий** удаляет предикторы, которые имеют единственное значение, вероятно, в результате выполнения дополнительных действий автоматической подготовки данных.

Важность предикторов

Обычно при моделировании сосредотачивают внимание на наиболее важных предикторах и исключают или игнорируют наименее важные. Это помогает сделать диаграмма важности предикторов, показывая относительную важность каждого предиктора при оценке модели. Поскольку значения важности являются относительными, сумма этих значений для всех показанных предикторов равна 1,0. Важность переменных не связана с точностью модели. Она лишь связана с важностью каждого предиктора для предсказания, а не с точностью этого предсказания.

Предсказанные против наблюденных

Выводится диаграмма рассеяния с интервалами для предсказанных значений по вертикальной оси против наблюденных значений по горизонтальной оси. В идеале точки должны лежать на прямой, проведенной под углом 45 градусов. Такое представление позволяет определить, есть ли записи, которые плохо предсказываются моделью.

Остатки

Выводится диагностическая диаграмма модельных остатков.

Стили диаграммы. Имеются различные стили вывода, которые можно выбрать в выпадающем списке **Стиль**.

- **Гистограмма.** Это диаграмма рассеяния с интервалами для стьюдентизированных остатков с наложением нормального распределения. Для линейных моделей предполагается, что остатки имеют нормальное распределение, поэтому в идеале гистограмма должна хорошо аппроксимироваться этой гладкой линией.
- **P-P диаграмма.** Это диаграмма с интервалами типа вероятность-вероятность, сравнивающая распределение стьюдентизированных остатков с нормальным распределением. Если наклон выведенных точек менее крутой, чем наклон нормальной кривой, то остатки показывают большую изменчивость, чем она должна быть для нормального распределения. Если этот наклон более крутой, то остатки показывают меньшую изменчивость, чем в случае нормального распределения. Если выведенные точки имеют форму S-образной кривой, то распределение остатков является скошенным.

Выбросы

Эта таблица выводит записи, которые оказывают чрезмерное влияние на модель, а также выводит ID записи (если это задано на вкладке Поля), значение целевого поля и расстояние Кука. Расстояние Кука - это мера того, насколько изменились бы остатки для всех записей, если конкретная запись не участвовала бы в вычислении коэффициентов модели. Большое расстояние Кука говорит о том, что исключение записи существенно изменяет коэффициенты, и должна рассматриваться как влияющая.

Влияющие записи должны быть тщательно исследованы, чтобы определить, нужно ли назначить им меньший вес при оценивании модели или урезать резко выделяющиеся значения (выбросы) до некоторого приемлемого порогового значения, или же полностью удалить влияющие записи.

Эффекты

Этот вид показывает величину каждого эффекта в модели.

Стили. Имеются различные стили вывода, которые можно выбрать в выпадающем списке **Стиль**.

- **Диаграмма.** Это диаграмма, в которой эффекты отсортированы сверху вниз по убыванию важности предикторов. Соединяющие линии на диаграмме являются взвешенными на основе значимости эффектов, с большей толщиной линии, соответствующей более значимым эффектам (меньшим *p*-значениям). При наведении указателя мыши на соединительную линию появляется всплывающая подсказка, выводящая *p*-значение и значение важности данного эффекта. Это задано по умолчанию.
- **Таблица.** Это таблица дисперсионного анализа для общих и индивидуальных эффектов модели. Индивидуальные эффекты отсортированы сверху вниз по убыванию важности предикторов. Обратите внимание на то, что по умолчанию таблица сворачивается, чтобы показать только результаты для модели в целом. Чтобы увидеть результаты для индивидуальных эффектов модели, щелкните по **Скорректированная модель** в ячейке таблице.

Важность предикторов. Есть ползунок важности предикторов, который управляет тем, какие предикторы выводятся. Это не изменяет модели, а просто позволяет сосредоточить внимание на наиболее важных предикторах. По умолчанию выводятся 10 верхних эффектов.

Значимость. Есть ползунок значимости, предоставляющий дополнительные возможности управлять тем, какие эффекты выводить, кроме тех, которые выводятся на основе значимости предикторов. Эффекты со значениями значимости, превосходящими значение ползунка, скрыты. Это не изменяет модели, а просто позволяет сосредоточить внимание на наиболее важных эффектах. По умолчанию это значение равно 1,00, так что никакие эффекты не отфильтровываются на основе значимости.

Коэффициенты

Этот вид показывает значение каждого коэффициента в модели. Обратите внимание на то, что факторы (категориальные предикторы) имеют индикаторную кодировку в модели, так что **эффекты** , содержащие факторы, обычно будут иметь несколько связанных **коэффициентов**, по одному для каждой категории, исключая категорию, соответствующую избыточному (опорному) параметру.

Стили. Имеются различные стили вывода, которые можно выбрать в выпадающем списке **Стиль**.

- **Диаграмма.** Это диаграмма, в которой сначала выводится свободный член, а затем эффекты, отсортированные сверху вниз по убыванию важности предикторов. Внутри эффектов, содержащих факторы, коэффициенты сортируются в порядке возрастания значений данных. Соединяющие линии на диаграмме раскрашены в зависимости от знака коэффициента (см. ключ диаграммы) и взвешены в зависимости от значимости коэффициента, с большей толщиной линии, соответствующей более значимым коэффициентам (меньшим *p*-значениям). При наведении указателя мыши на соединительную линию появляется всплывающая подсказка, выводящая значение коэффициента, *p*-значение для него, а также значение важности эффекта, с которым связан этот параметр. Это задано по умолчанию.
- **Таблица.** В этой таблице выводятся значения, результаты тестов на значимость и доверительные интервалы для индивидуальных коэффициентов модели. После свободного члена эффекты отсортированы сверху вниз по убыванию важности предикторов. Внутри эффектов, содержащих факторы, коэффициенты сортируются в порядке возрастания значений данных. Обратите внимание на то, что по умолчанию таблица сворачивается, чтобы вывести только коэффициент, значимость и важность для каждого параметра модели. Чтобы увидеть стандартную ошибку, *t*-статистику и доверительный интервал, щелкните по ячейке **Коэффициент** в таблице. При наведении указателя мыши на имя параметра модели в таблице появляется всплывающая подсказка, выводящая имя параметра, эффект, с которым связан этот параметр, и (для категориальных предикторов) метки значений, связанных с данным параметром модели. Это, в частности, позволяет увидеть новые категории, созданные, когда автоматическая подготовка данных привела к объединению сходных категорий категориального предиктора.

Важность предикторов. Есть ползунок важности предикторов, который управляет тем, какие предикторы выводятся. Это не изменяет модели, а просто позволяет сосредоточить внимание на наиболее важных предикторах. По умолчанию выводятся 10 верхних эффектов.

Значимость. Есть ползунок значимости, предоставляющий дополнительные возможности управлять тем, какие коэффициенты выводить, кроме тех, которые выводятся на основе значимости предикторов. Коэффициенты со значениями значимости, превосходящими значение ползунка, скрыты. Это не изменяет модели, а просто позволяет сосредоточить внимание на наиболее важных коэффициентах. По умолчанию это значение равно 1,00, так что никакие коэффициенты не отфильтровываются на основе значимости.

Оцененные средние

Это диаграммы, выводимые для значимых предикторов. На диаграмме вдоль вертикальной оси выводится оцененное по модели значение целевой переменной для каждого значения предиктора на горизонтальной оси при сохранении значений всех остальных предикторов неизменными. Это дает полезную визуализацию того, какое влияние коэффициент каждого предиктора оказывает на целевую переменную.

Примечание: если нет значимых предикторов, оцененные средние не генерируются.

Сводка по построению модели

Эта панель предоставляет некоторые детали процесса построения модели, когда в группе параметров Подбор модели сделан выбор алгоритма отбора, отличный от **Включить все предикторы** .

Прямой шаговый. Если алгоритмом отбора является прямой шаговый, то в таблице выводятся последние 10 шагов шагового алгоритма. На каждом шаге показываются значение критерия отбора и эффекты в модели. Это дает понимание того, какой вклад в модель дает каждый шаг. В каждом столбце можно сортировать строки, чтобы было легче видеть, какие эффекты содержатся в модели на каждом шаге.

Наилучшие подмножества. Если алгоритмом отбора является "наилучшие подмножества", то таблица выводит 10 лучших моделей. Для каждой модели показываются значение критерия отбора и эффекты в модели. Это позволяет проверить стабильность лучших моделей. Если для них наблюдается тенденция иметь много схожих эффектов с небольшими различиями, то наилучшей модели можно вполне доверять. Если для них наблюдается тенденция иметь сильно различающиеся эффекты, то некоторые из этих эффектов могут быть слишком схожи между собой, и их следует объединить (или один удалить). В каждом столбце можно сортировать строки, чтобы было легче видеть, какие эффекты содержатся в модели на каждом шаге.

Параметры

Обратите внимание на то, что при скоринге модели всегда вычисляется предсказанное значение. Имя нового поля есть имя целевого поля с префиксом *$L-*. Например, при имени целевого поля *sales*, новому полю было бы присвоено имя *$L-sales*.

Сгенерировать SQL для данной модели. Когда используются данные, хранящиеся в базе данных, код SQL можно передавать обратно в базу данных для выполнения, что обеспечивает высокую производительность многих операций.

Оценить по преобразованию в собственный SQL. Если этот параметр выбран, то создается SQL для скоринга модели внутри приложения.

Логистический узел

Логистическая регрессия, известная также как **номинальная регрессия**, - это статистический метод для классификации записей на основе значений входных полей. Она аналогична линейной регрессии, но логистическая регрессия использует категориальные поля назначения вместо числовых. Поддерживаются и биномиальные (для полей назначения с двумя дискретными категориями), и полиномиальные (для полей назначения с большим двух количеством категорий) модели.

При работе логистическая регрессия строит набор уравнений, связывающих значения входных полей с вероятностями, относящимися к каждой из категорий выходного поля. После того как модель сгенерирована, ее можно использовать для оценки вероятностей для новых данных. Для каждой записи вероятность принадлежности вычисляется для каждой возможной выходной категории. Категория назначения с наибольшей вероятностью назначается как предсказанное выходное значение для данной записи.

Пример биномиальной модели. Провайдер телекоммуникационных услуг озабочен количеством клиентов, теряемых из-за конкурирующих компаний. Используя данные об используемых услугах, вы можете создать биномиальную модель, чтобы предсказать, какие из клиентов склонны к переходу к другому провайдеру, и настроить предложения для сохранения максимально возможного числа клиентов. Используется биномиальная модель, так как у поля назначения два отдельных значения (например, Переходит и Остается).

Примечание: Только для биномиальных моделей. Строковые поля должны быть ограничены по длине восемью символами. При необходимости более длинные строки можно перекодировать с помощью узла переклассификации.

Пример полиномиальной модели. Провайдер связи сегментировал базу своих клиентов по шаблонам использования сервисов, категоризуя клиентов в четыре группы. Используя демографические данные для предсказания принадлежности к группам, можно создать полиномиальную модель для классификации перспективных клиентов по группам и затем настроить предложения для индивидуальных клиентов.

Требования. Одно или несколько входных полей и строго одно категориальное поле назначения с двумя или более категориями. Для биномиальной модели у поля назначения должен быть уровень измерения *Флаг*. Для полиномиальной модели у поля назначения должен быть уровень измерения *Флаг* или *Номинальное* с двумя или более категориями. Поля с заданными значениями *Оба* или *Нет* игнорируются. У используемых в модели полей должны быть полностью конкретизированы типы.

Достоинства. Модели логистической регрессии часто весьма точные. Они могут работать с символическими и числовыми входными полями. Такие модели могут давать предсказанные вероятности для всех категорий полей назначения, поэтому легко определить вторую по качеству гипотезу. Модели логистической регрессии наиболее эффективны в ситуации, когда принадлежность к группе - это истинно категориальное поле; если принадлежность к группе определяется значениями поля количественного диапазона (например, высокий IQ по сравнению с низким IQ), необходимо обратиться к линейной регрессии, чтобы воспользоваться преимуществом большей информативности полного диапазона значений. Логистические модели могут

выполнять автоматический выбор полей, хотя модели типа дерева и Выбор показателей могут делать это быстрее для больших наборов данных. И наконец, так как логистические модели хорошо известны многим аналитикам и исследователям данных, некоторые могут использовать их в качестве базовых версий для дальнейшего сравнения с другими способами.

Пи обработке больших наборов данных можно существенно повысить производительность, отключив критерий отношения правдоподобия, расширенную выходную опцию. Дополнительную информацию смотрите в разделе “Расширенный вывод для логистической регрессии” на стр. 170.

Опции моделей узла логистической регрессии

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Создание моделей расщепления. Строит отдельную модель для каждого возможного значения входных полей, заданных как поля расщепления. Дополнительную информацию смотрите в разделе “Построение моделей расщепления” на стр. 28.

Процедура. Задает, какая модель создается, биномиальная или полиномиальная. Доступные в диалоговом окне опции меняются в зависимости от типа выбранной процедуры моделирования.

- **Биномиальное.** Используется, когда поле назначения - флаговое или номинальное с двумя дискретными значениями (дихотомическими), такими как *да*/*нет*, *on*/*off*, *м*/*ж*.
- **Полиномиальное.** Используется, когда поле назначения номинальное с более чем двумя значениями. Можно задать опции **Главные эффекты**, **Полная факторная** или **Пользовательская**.

Включить в уравнение константу. Эта опция определяет, будет ли включаться в итоговые уравнения постоянный член. В большинстве случаев нужно оставить эту опцию включенной.

Биномиальные модели

Для биномиальных моделей доступны следующие способы и опции:

Метод. Задать способ, который будет использоваться при построении модели логистической регрессии.

- **Ввод.** Это способ по умолчанию, при котором все члены непосредственно вводятся в уравнение. При построении модели выбор полей не производится.
- **Прямой.** При прямом способе выбора полей модель строится последовательным продвижением вперед шаг за шагом. При этом способе начальная модель - простейшая, к ней можно добавлять только константу и отдельные члены. На каждом шаге проверяются еще не присутствующие в модели члены и определяется, насколько существенно они могут улучшить модель; лучший из членов добавляется в модель. Когда больше нет членов для добавления или лучший член-кандидат не обеспечивает достаточного улучшения модели, генерируется итоговая модель.
- **Обратный.** Обратный способ по существу противоположен прямому. При этом способе начальная модель содержит все члены в качестве предикторов, и в дальнейшем эти члены можно только удалять из модели. Члены модели, дающие малый вклад, удаляются по одному, пока ни один член не может быть удален без существенного ухудшения модели, и в результате возникает итоговая модель.

Категориальные входные поля. Перечисляет поля, которые идентифицируются как категориальные, то есть с флаговым, номинальным или порядковым уровнем измерений. Для каждого категориального поля можно задать контраст и базовую категорию.

- **Имя поля.** В этом столбце содержатся имена категориальных входных полей, и он предварительно заполнен всеми флаговыми и номинальными значениями данных. Чтобы добавить в этот столбец количественные или числовые входные поля, щелкните по значку Добавить поля справа от списка и выберите нужные поля.
- **Контраст.** Интерпретация коэффициентов регрессии для категориального поля зависит от используемых контрастов. Контраст определяет, как конфигурируется проверка гипотез для сравнения оцененных средних. Например, если известно, что у категориального поля есть подразумеваемый порядок, например, шаблон или группировка, для моделирования этого порядка можно использовать контрасты. Доступны следующие контрасты:

  Индикатор. Контрасты указывают на наличие или отсутствие принадлежности к категории. Это метод по умолчанию.

  Простая. Каждая категория предикторного поля, кроме опорной категории, сравнивается с опорной категорией.

  Дифференциальный. Каждая категория предикторного поля, кроме первой категории, сравнивается со средним эффектом предыдущих категорий. Известны также как обратные контрасты Хелмерта.

  Хелмерт. Каждая категория предикторного поля, кроме последней категории, сравнивается со средним эффектом последующих категорий.

  Повторяемый. Каждая категория предикторного поля, кроме первой категории, сравнивается с предшествующей категорией.

  Полином.. Ортогональные полиномиальные контрасты. Предполагается, что категории расположены на равных расстояниях. Полиномиальные контрасты доступны только для числовых полей.

  Отклонения. Каждая категория предикторного поля, кроме опорной категории, сравнивается с общим (суммарным) эффектом.
- **Базовая категория.** Задает, как определяется опорная категория для выбранного типа контраста. Выберите **Первая**, чтобы использовать первую категорию для входного поля (сортировка по алфавиту), или выберите **Последняя**, чтобы использовать последнюю категорию. Значение по умолчанию - Первая.

  Примечание: Это поле недоступно, если параметр контраста - Дифференциальный, Хелмерта, Повторяемый или Полиномиальный.

Оценка влияния каждого поля на общий отклик вычисляется как увеличение или уменьшение правдоподобия каждой из других категорий по отношению к опорной. Это может помочь в идентификации полей и значений, которые с большей вероятностью дадут конкретный отклик.

Базовая категория показывается при выводе как 0,0. Это связано с тем, что сравнение ее с самой собой дает пустой результат. Все другие категории показаны как уравнения относительно базовой категории. Дополнительную информацию смотрите в разделе “Подробности слепка модели логистической регрессии” на стр. 173.

Полиномиальные модели

Для полиномиальных моделей доступны следующие способы и опции:

Метод. Задать способ, который будет использоваться при построении модели логистической регрессии.

- **Ввод.** Это способ по умолчанию, при котором все члены непосредственно вводятся в уравнение. При построении модели выбор полей не производится.
- **Пошаговый.** Пошаговый способ отбора полей строит уравнение от шага к шагу, как и обозначено названием. Исходная модель - это простейшая модель без членов модели (кроме константы) в уравнении. На каждом шаге оцениваются члены, которые еще не добавлены в модель, и добавляются те из них, которые вносят наиболее существенный вклад в предсказательную силу модели. Кроме этого, уже присутствующие в модели члены оцениваются повторно для определения, можно ли удалить некоторые из них без существенного воздействия на модель. Если такие члены есть, они удаляются. Процесс

повторяется, и другие члены добавляются и/или удаляются. Когда больше нет членов, которые можно добавить для улучшения модели, и нет членов, которые можно удалить из модели без существенного ее ухудшения, генерируется окончательная модель.

- **Прямой.** Прямой способ выбора полей похож на пошаговый способ в том, что модель строится по шагам. Однако в этом способе начальная модель - простейшая, к ней можно добавлять только константу и отдельные члены. На каждом шаге еще не присутствующие в модели члены проверяются, насколько существенно они могут улучшить модель, и лучший из членов добавляется в модель. Когда больше нет членов для добавления или лучший член-кандидат не обеспечивает достаточного улучшения модели, генерируется итоговая модель.
- **Обратный.** Обратный способ по существу противоположен прямому. При этом способе начальная модель содержит все члены в качестве предикторов, и в дальнейшем эти члены можно только удалять из модели. Члены модели, дающие малый вклад, удаляются по одному, пока ни один член не может быть удален без существенного ухудшения модели, и в результате возникает итоговая модель.
- **Обратный пошаговый.** Обратный пошаговый способ существенно противоположен пошаговому. В этом способе в начальную модель включаются все возможные предикторы. На каждом шаге уже присутствующие в модели члены оцениваются для определения, можно ли удалить некоторые из них без существенного воздействия на модель. Кроме этого, повторно оцениваются и члены, ранее удаленные из модели, чтобы определить, не повысят ли лучшие из них предсказательную силу модели. Если так, данные члены возвращаются в модель. Когда больше нет членов, которые можно удалить из модели без существенного ее ухудшения, и нет членов, которые можно добавить для улучшения модели, генерируется окончательная модель.

Примечание: Автоматические способы, в том числе пошаговый, прямой и обратный пошаговый, - это очень адаптивные способы обучения, и у них есть сильная тенденции к сверхобучению данных обучения. При использовании этих способов особенно важно проверить приемлемость результирующей модели или на новых данных, или на существующей тестовой выборке, созданной на узле Разделы.

Базовая категория для назначения. Задает, как определяется опорная категория. Она используется как базовая, относительно которой оцениваются уравнения регрессии для всех остальных категорий. Выберите **Первая**, чтобы использовать первую категорию для входного поля (сортировка по алфавиту), или выберите **Последняя**, чтобы использовать последнюю категорию. Как вариант, можно выбрать опцию **Задать**, чтобы определить конкретную категорию и выбрать для нее нужное значение из списка. Доступные значения для каждого поля можно определить на узле Тип.

Часто можно задать категорию, которая представляет для вас наименьший интерес, например, продаваемый с убытком товар. Затем другие категории связываются с базовой для сравнения и идентификации, что именно повышает вероятность их принадлежности в данной категории. Это может помочь в идентификации полей и значений, которые с большей вероятностью дадут конкретный отклик.

Базовая категория показывается при выводе как 0,0. Это связано с тем, что сравнение ее с самой собой дает пустой результат. Все другие категории показаны как уравнения относительно базовой категории. Дополнительную информацию смотрите в разделе “Подробности слепка модели логистической регрессии” на стр. 173.

Тип модели. Есть три опции для определения членов в модели. Модели **Главные эффекты** включают в себя только входные поля по отдельности, и не проверяют взаимодействий (мультипликативных эффектов) между входными полями. **Полные факторные** модели включают в себя все взаимодействия, а также главные эффекты входных полей. Полные факторные модели способны лучше захватывать комплексные взаимосвязи, но их гораздо сложнее интерпретировать, а также есть риск переобучения. Из-за потенциально большого числа возможных комбинаций, способы автоматического выбора полей (отличающиеся от Ввода) отключаются для полных факторных моделей. **Пользовательские** модели включают в себя только члены (главные эффекты и взаимодействия), которые задаете вы сами. При выборе этой опции используйте список Члены модели, чтобы добавить члены в модель или удалить их оттуда.

Члены модели. При построении Пользовательской модели вам нужно явно указать члены в этой модели. Этот список показывает текущий набор членов для модели. Кнопки справа от списка Члены модели позволяют добавлять или удалять члены модели.

- Чтобы добавить члены в модель, нажмите кнопку *Добавить новые члены модели*.
- Чтобы удалить какие-то члены, выберите их и нажмите кнопку *Удалить выбранные члены модели*.

Добавление членов в модель логистической регрессии

При запросе пользовательской модели логистической регрессии можно добавить члены в модель, нажав кнопку *Добавить новые члены модели* на вкладке Модель логистической регрессии. Откроется диалоговое окно Новые члены, в котором можно задавать члены регрессии.

Тип добавляемого члена. Есть несколько способов добавить члены в модель, основываясь на выборе входных полей в списке Доступные поля.

- **Простое взаимодействие.** Вставляет члены, представляющие взаимодействие всех выбранных полей.
- **Главные эффекты.** Вставляет один член главного эффекта (само поле) для каждого выбранного поля.
- **Все двухсторонние взаимодействия.** Вставляет член двухстороннего взаимодействия (произведение входных полей) для каждой возможной пары выбранных входных полей. Например, если в списке Доступные поля выбраны входные поля *A*, *B* и *C*, этим способом можно вставить члены *A * B*, *A * C* и *B * C*.
- **Все трехсторонние взаимодействия.** Вставляет член трехстороннего взаимодействия (произведение входных полей) для всех возможных комбинаций трех одновременных выбранных полей. Например, если в списке Доступные поля выбраны входные поля *A*, *B*, *C* и *D*, этим способом можно вставить члены *A * B * C*, *A * B * D*, *A * C * D* и *B * C * D*.
- **Все четырехсторонние взаимодействия.** Вставляет член четырехстороннего взаимодействия (произведение входных полей) для всех возможных комбинаций четырех одновременных выбранных полей. Например, если в списке Доступные поля выбраны входные поля *A*, *B*, *C*, *D* и *E*, этим способом можно вставить члены *A * B * C * D*, *A * B * C * E*, *A * B * D * E*, *A * C * D * E* и *B * C * D * E*.

Доступные поля. Выводит список доступных входных полей, которые будут использоваться при конструировании членов модели.

Просмотр. Показывает члены, которые будут включены в модель после нажатия кнопки **Вставить** на основе выбранных полей и выбранных выше типов членов.

Вставить. Вставляет члены в модель (на основе текущего выбора полей и типов членов) и закрывает диалоговое окно.

Дополнительные опции узла логистической регрессии

Если вы хорошо знакомы с логистической регрессией, дополнительные опции позволят точнее настроить процесс обучения. Для доступа к дополнительным опциям выберите режим **Дополнительно** на вкладке Дополнительно.

Масштаб (только полиномиальные модели). Можно задать значение масштаба дисперсии, которое используется для коррекции оценки ковариационной матрицы параметра. **Пирсон** оценивает значение масштаба при помощи статистики хи-квадрат Пирсона. **Отклонение** оценивает значение масштаба при помощи статистики функции отклонения (хи-квадрат отношения правдоподобия). Кроме того, можно задать свое пользовательское значение масштаба. Оно должно быть положительным значением.

Добавить все вероятности. Если выбрана эта опция, вероятности для каждой категории выходного поля будут добавляться к каждой записи, обрабатываемой узлом. Если эта опция не выбрана, добавляется только вероятность предсказанной категории.

Например, в таблицу, содержащую результаты полиномиальной модели с тремя категориями, будет включено пять новых столбцов. В одном столбце будут перечислены вероятности правильно предсказанного выхода, в следующем столбце будет показана вероятность попадания или промаха этого предсказания, а в остальных трех столбцах будет показана вероятность, что предсказание каждой категории - это промах или попадание. Дополнительную информацию смотрите в разделе “Слепок логистической модели” на стр. 172.

Примечание: Для биномиальных моделей всегда выбирается эта опция.

Допуск вырожденности. Задать допуск, используемый при проверке на сингулярности.

Сходимость. Эти опции позволяют вам управлять параметрами для сходимости модели. Когда выполняется модель, настройки сходимости управляют тем, сколько раз проводятся повторные запуски с различными параметрами, чтобы увидеть, хорошо ли они подходят. Чем чаще испытываются параметры, тем ближе друг к другу результаты (то есть результаты сходятся). Дополнительную информацию смотрите в разделе “Опции сходимости логистической регрессии”.

Вывод. Эти опции позволяют затребовать дополнительные статистические данные, которые будут показаны в расширенном выводе слепка модели, построенном узлом. Дополнительную информацию смотрите в разделе “Расширенный вывод для логистической регрессии”.

Пошаговые опции. Эти опции позволяют вам управлять критериями для добавления и удаления полей при пошаговом, прямом, обратном и пошаговом обратном способах оценки. (Эта кнопка отключена, если выбран способ Ввод). Дополнительную информацию смотрите в разделе “Опции шагового отбора логистической регрессии” на стр. 171.

Опции сходимости логистической регрессии

Вы можете задать параметры сходимости для оценки моделей логистической регрессии.

Максимум итераций. Задайте максимальное количество итераций для оценки модели.

Максимальное число шагов половинного деления. Половинное деление - это используемый логистической регрессией прием при возникновении сложностей в процессе оценки. При нормальных условиях нужно использовать параметр по умолчанию.

Сходимость логарифмического правдоподобия. Итерации прекращаются, если относительное изменение Log-правдоподобия меньше этого значения. Этот критерий не используется, если значение равно 0.

Сходимость параметров. Итерации прекращаются, если абсолютное или относительное изменение в оценках параметра меньше этого значения. Этот критерий не используется, если значение равно 0.

Дельта (только для полиномиальных моделей). Можно задать значение от 0 до 1, которое будет добавляться в каждую пустую ячейку (комбинация значений входных и выходных полей). Это может помочь алгоритмам в обработке данных, когда существует много возможных комбинаций значений полей по сравнению с числом записей в данных. Значение по умолчанию - 0.

Расширенный вывод для логистической регрессии

Выберите дополнительные выходные данные, которые вы хотите использовать в расширенном выводе слепка модели регрессии. Для просмотра расширенного вывода перейдите на слепок модели и выберите вкладку **Дополнительно**. Дополнительную информацию смотрите в разделе “Расширенный вывод слепков моделей логистической регрессии” на стр. 175.

Параметры Биномиального критерия

Выберите типы выходных данных, которые будут генерироваться для модели. Дополнительную информацию смотрите в разделе “Расширенный вывод слепков моделей логистической регрессии” на стр. 175.

Вывести. Выберите, выводить ли результаты по каждому шагу или дождаться, пока будут выполнены все шаги.

ДИ для exp(B). Выберите доверительные интервалы для всех коэффициентов в выражении (показанных как бета). Задайте уровень доверительного интервала (значение по умолчанию 95%).

Диагноз остатков. Требует таблицу остатков Диагностика по наблюдениям.

- **Выбросы за пределами (среднеквадратичное отклонение).** Перечислить только остаточные наблюдения, для которых стандартизованное значение приведенной переменной не меньше заданного вами значения. Значение по умолчанию - 2.
- **Все наблюдения.** Включает все наблюдения в таблице остатков Диагностика по наблюдениям.

  Примечание: Так как при выборе этой опции перечисляются все входные записи, это может привести к появлению в отчете очень большой таблицы, в которой одна строка будет соответствовать одной записи.

Порог отсечения классификации. Эта опция позволяет определить точку отсечения для классификации наблюдений. Наблюдения с предсказанными значениями больше порога отсечения классификации, классифицируются как положительные, а с предсказанными значениями меньше порога отсечения - как отрицательные. Чтобы изменить значение по умолчанию, введите число между 0.01 и 0.99.

Полиномиальные опции

Выберите типы выходных данных, которые будут генерироваться для модели. Дополнительную информацию смотрите в разделе “Расширенный вывод слепков моделей логистической регрессии” на стр. 175.

Примечание: Выбор опции **Критерии отношений правдоподобия** существенно увеличивает время обработки, требуемое для построения модели логистической регрессии. Если для построения вашей модели требуется очень большое время, рассмотрите возможность отключения этой опции или использования вместо нее статистики Вальда и МЛ-статистики. Дополнительную информацию смотрите в разделе “Опции шагового отбора логистической регрессии”.

Хронология итераций для каждых. Выберите интервал для пошаговой печати состояния итераций в расширенном выводе.

Доверительный интервал. Доверительные интервалы для коэффициентов в уравнениях. Задайте уровень доверительного интервала (значение по умолчанию 95%).

Опции шагового отбора логистической регрессии

Эти опции позволяют вам управлять критериями для добавления и удаления полей при пошаговом, прямом, обратном и пошаговом обратном способах оценки.

Число членов в модели (только полиномиальные модели). Вы можете задать минимальное количество членов модели для обратного и пошагового обратного способа построения и максимальное число членов для прямого и пошагового способа. Если задать минимальное значение больше 0, модель будет включать в себя как минимум заданное число членов, даже если некоторые из них были удалены на основании статистических критериев. Минимальный параметр игнорируется при построении моделей прямым и пошаговым способами, а также способом Ввод. Если задать максимальное значение, некоторые члены могут быть пропущены в модели, даже если они были выбраны на основании статистических критериев. Настройка **Задать максимум** игнорируется при построении моделей обратным и пошаговым обратным способами, а также способом Ввод.

Критерий включения (только для полиномиальных моделей). Для достижения максимальной скорости обработки выберите **Оценка**. При использовании опции **Отношение правдоподобия** оценки будут несколько более устойчивыми, но потребуется больше времени для их вычисления. По умолчанию используется МЛ-статистика.

Критерий исключения. Выберите **Отношение правдоподобия** для более устойчивой модели. Для сокращения времени на построение модели попробуйте выбрать **Вальд**. Однако если вы выполнили или почти выполнили разделение данных (что можно определить на вкладке Дополнительно слепка модели), статистика Вальда становится очень ненадежной и ее не нужно использовать. По умолчанию используется статистика отношения правдоподобия. Для биномиальных моделей существует дополнительная опция **Условно**. Она обеспечивает удаление проверки на основании вероятности по статистике отношения правдоподобия, вычисленной по оценкам условных параметров.

Пороги значимости для критериев. Эта опция позволяет задать критерии выбора на основе статистической вероятности (значение *p*), связанной с каждым полем. Поля будут добавлены в модель, если только связанное значение *p* меньше, чем значение **Ввод**, и будут удалены, если только значение *p* больше, чем значение **Удаление**. Значение **Ввод** должно быть меньше, чем значение **Удаление**.

Требования для включения или исключения (только полиномиальные модели). Для некоторых прикладных программ с математической точки зрения бессмысленно добавлять члены взаимодействия в модель, пока она не содержит также членов низшего порядка для полей, входящих в члены взаимодействия. Например, нет смысла включать в модель член *A * B*, пока в нее не включены отдельно члены *A* и *B*. Эти опции позволяют определить, как такие зависимости обрабатываются при пошаговом выборе членов.

- **Иерархия для дискретных эффектов.** Эффекты высших порядков (взаимодействия, включающие больше полей) вводятся в модель, только если все эффекты порядком ниже (главные эффекты и взаимодействия, включающие меньше полей) для соответствующих полей уже есть в модели и эффекты низшего порядка не будут удаляться, если члены высшего порядка с теми же полями есть в модели. Эта опция применима только к категориальным полям.
- **Иерархия для всех эффектов.** Эта опция работает так же, как предыдущая, за тем исключением, что она применяется ко всем входным полям.
- **Ограничение распространения для всех эффектов.** Эффекты можно включать в модель, только если все эффекты, содержащиеся в данном эффекте, также включены в модель. Эта опция аналогична опции **Иерархии для всех эффектов**, но количественные поля рассматриваются немного иначе. Чтобы эффект содержал другой эффект, этот другой эффект (низшего порядка) должен включать в себя *все* количественные поля, присутствующие в содержащем эффекте (высшего порядка), а категориальные поля содержащегося эффекта должны быть подмножеством таких полей в содержащем эффекте. Например, если *A* и *B* - это категориальные поля, а *X* - это количественное поле, член *A * B * X* содержит члены *A * X* и *B * X*.
- **Нет.** Никакие взаимосвязи принудительно не устанавливаются; члены добавляются в модель и удаляются из нее независимо.

Слепок логистической модели

Слепок логистической модели представляет уравнения, оцененные узлом логистической модели. Он содержит всю информацию, собранную моделью логистической регрессии, а также информацию о структуре и производительности модели. Уравнение такого типа может быть сгенерировано также другими моделями, такими как Oracle SVM.

При запуске потока, содержащего слепок логистической модели, узел добавляет два новых поля, содержащих предсказание модели и связанную вероятность. Имена новых полей получаются из имени выходного поля, которое предсказывается, с добавлением префикса *$L-* для предсказанной категории и префикса *$LP-* для связанной вероятности. Например, если имя выходного поля *colorpref*, новые поля будут называться *$L-colorpref* и *$LP-colorpref*. Кроме этого, если вы выбрали опцию **Присоединить все вероятности** на узле логистической модели, будет добавлено дополнительное поле для каждой категории выходного поля, содержащее вероятность, принадлежащую соответствующей категории для каждой записи. Имена этих

дополнительных полей состоят из значений выходного поля и префикса *$LP-*. Например, если разрешенные значения в поле *colorpref* - это *Красный*, *Зеленый* и *Синий*, будут добавлены три новые поля: *$LP-Красный*, *$LP-Зеленый* и *$LP-Синий*.

Генерирование узла Фильтр. Меню Генерировать позволяет вам создавать новый узел Фильтр для передачи входных полей на основе результатов модели. Поля, отброшенные из модели из-за мультиколлинеарности, будут отфильтрованы сгенерированным узлом, как и поля, не использованные в модели.

Подробности слепка модели логистической регрессии

Для полиномиальных моделей на вкладке Модель слепка модели логистической регрессии выведены расщепления с уравнениями модели на левой панели и важностью предикторов на правой. Для биномиальных моделей на вкладке выводится только важность предикторов. Дополнительную информацию смотрите в разделе “Важность предиктора” на стр. 42.

Уравнения моделей

Для полиномиальных моделей на левой панели выводятся фактические уравнения, оцененные для модели логистической регрессии. Существует по одному уравнению для каждой категории в поле назначения, кроме базовой категории. Уравнения выводятся в формате дерева. Уравнение такого типа может быть сгенерировано также некоторыми другими моделями, такими как Oracle SVM.

Уравнение для. Показывает уравнения регрессии, используемые для получения вероятностей категории назначения при заданных значениях предикторов. Последняя категория поля назначения рассматривается как **базовая категория**; показанные уравнения дают отношения шансов для других категорий поля назначения относительно базовой категории для конкретного набора значений предикторов. Предсказанная вероятность для каждой категории данной структуры предикторов получается из этих значений отношения шансов.

Как вычисляются вероятности

Каждое уравнение вычисляет отношения шансов для конкретной категории назначения относительно базовой категории. **Отношения шансов**, также называемые **логитами**, - это отношения вероятности заданной категории назначения к вероятности базовой категории, к которым после вычисления применяется натуральная логарифмическая функция. Для базовой категории шансы относительно самой себя равны 1,0, то есть после логарифмирования получаем 0. Это можно представить, как подразумеваемое уравнение для базовой категории, в котором все коэффициенты равны 0.

Чтобы получить вероятность из отношения шансов для конкретной категории назначения, нужно взять значение логит, вычисленное по уравнению для этой категории, и применить следующую формулу:

$$P(\textit{группы}_i) = \exp(g_i) / \Sigma_k \exp(g_k)$$

где g - это вычисленные отношения шансов, i - индекс категории, а k изменяется от 1 до количества категорий поля назначения.

Важность предиктора

Диаграмма, обозначающая относительную важность каждого предиктора в оцениваемой модели, может быть дополнительно также показана на вкладке Модель. Обычно вам потребуется направить усилия по моделированию на рассмотрение наиболее важных предикторов, отбрасывая или игнорируя самые не важные. Обратите внимание на то, что эта диаграмма доступна только в том случае, если перед генерированием модели на вкладке Анализ выбрана опция **Вычислять важность предикторов**. Дополнительную информацию смотрите в разделе “Важность предиктора” на стр. 42.

Примечание: Для модели логистической регрессии важность предикторов может вычисляться дольше, чем для моделей других типов, и это не выбирается на вкладке Анализ по умолчанию. Выбор этой опции может понизить производительность, особенно для больших наборов данных.

Сводка слепков моделей логистической регрессии

Сводка для модели логистической регрессии выводит поля и параметры, используемые для генерирования модели. Кроме этого, если вы запускали узел Анализ, присоединенный к этому узлу моделирования, информация этого анализа также будет выводится в данном разделе. За общей информацией по браузеру моделей обратитесь к разделу “Просмотр слепков моделей” на стр. 41.

Параметры слепков моделей логистической регрессии

На вкладке Параметры в слепке модели логистической регрессии задаются опции для доверительных показателей, вероятностей, оценок склонности и генерирования SQL при скоринге моделей. Эта вкладка доступна только после того, слепок модели был добавлен в поток и вывел различные опции, зависящие от типа модели и поля назначения.

Полиномиальные модели

Для полиномиальных моделей доступны следующие опции:

Вычислить показатели доверия. Задает, вычисляются ли при скоринге показатели доверия.

Вычислить простые оценки склонности (только для флаговых полей назначения). Только для моделей с флаговым полем назначения можно затребовать простые оценки склонности, которые обозначают правдоподобие выходного значения *true*, заданного для поля назначения. Это дополнение к стандартным предсказанным и доверительным значениям. Скорректированные оценки склонности недоступны. Дополнительную информацию смотрите в разделе “Опции анализа узлов моделирования” на стр. 34.

Добавить все вероятности. Задает, добавляются ли вероятности для каждой категории выходного поля к каждой записи, обрабатываемой узлом. Если эта опция не выбрана, добавляется только вероятность предсказанной категории. Например, для номинального поля назначения с тремя категориями вывходные данные скоринга будут включать в себя столбец для каждой из трех категорий, а также четвертый столбец, обозначающий, для какой категории предсказана вероятность. Например, если вероятности для категорий *Красный*, *Зеленый* и *Синий* равны 0,6, 0,3 и 0,1 соответственно, предсказанной категорией будет *Красный* с вероятностью 0,6.

Оценить по преобразованию в собственный SQL. Если этот параметр выбран, то создается SQL для скоринга модели внутри приложения.

Примечание: Для полиномиальных моделей генерирование SQL недоступно, если была выбрана опция **Присоединить все вероятности**, а для моделей с номинальными полями назначения - если выбрано **Вычислять доверительные показатели**. Генерирование SQL с вычислениями доверительных показателей поддерживается только для полиномиальных моделей с флаговыми полями назначения. Для биномиальных моделей генерирование SQL недоступно.

Биномиальные модели

Для биномиальных полей доверительные показатели и вероятности всегда включены и нет настроек, позволяющих отключить эти опции. Для биномиальных моделей генерирование SQL недоступно. Единственный параметр, который можно изменить для биномиальных моделей, - это возможность вычисления простых оценок склонностей. Как отмечено ранее для полиномиальных моделей, это применимо только для моделей с флаговыми полями назначения. Дополнительную информацию смотрите в разделе “Опции анализа узлов моделирования” на стр. 34.

Расширенный вывод слепков моделей логистической регрессии

Расширенный вывод для логистической регрессии (другое название - **номинальная регрессия**) дает подробную информацию об оцениваемой модели и ее производительности. Большая часть содержащейся в расширенном выводе информации техническая, и требуется глубокое знание анализа логистической регрессии, чтобы правильно интерпретировать этот вывод.

Предупреждения. Обозначает любые предупреждения или потенциальные проблемы с результатами.

Сводка обработки наблюдений. Выводит количество обработанных записей с переходом на новую строку при каждом символическом поле в модели.

Сводка по шагам (необязательно). Перечисляет эффекты, добавленные в модель или удаленные из нее на каждом шаге создания модели, когда используется автоматический выбор полей.

Примечание: Показывается только для пошагового, прямого, обратного и пошагового обратного способов создания модели.

Хронология итераций (необязательно). Показывает хронологию итераций оценок параметров для каждых *n* итераций, начиная с первой, где *n* - это значение интервала печати. По умолчанию печатается каждая итерация (n=1).

Информация о подгонке моделей (полиномиальные модели). Показывает результаты испытания отношения правдоподобия вашей модели (итоговой) в сравнении с моделью, у которой коэффициенты всех параметров были равны нулю (был только свободный член).

Классификация (необязательно). Показывает матрицу предсказанных и фактических значений выходного поля с указанием их процентных долей.

Статистика критерия согласия хи-квадрат (необязательно). Показывает статистику критерия согласия хи-квадрат и Пирсона. Эти статистические данные проверяют общую подгонку модели к данным обучения.

Критерий согласия Хосмера-Лемешева (необязательно). Показывает результаты группировки наблюдений по децилям риска и сравнение наблюденной в каждой децили вероятности с её ожидаемым значением. Эта статистика согласия более устойчива, чем традиционные статистики согласия, используемые в полиномиальных моделях, особенно для моделей с непрерывными ковариатами и для исследования с выборками малого объёма.

Псевдо R-квадрат (необязательно). Выводит показатели подгонки модели по Коксу и Снеллу, Нэйджелкерку и *R*-квадрат Макфаддена. Эти статистические данные в некотором смысле аналогичны статистике *R*-квадрат в линейной регрессии.

Показатели монотонности (необязательно). Показывает число согласованных пар, рассогласованных пар и связанных пар в данных, а также процентную долю общего числа представленных пар. В этой же таблице содержатся D Сомерса, гамма Гудмана и Краскала, тау-а Кендалла и C-показатель согласия.

Информационный критерий (необязательно). Показывает информационный критерий Акаике (corrected Akaike information criterion, AIC) и байесовский информационный критерий Шварца (Schwarz's Bayesian information criterion, BIC).

Критерии отношения правдоподобия (необязательно). Показывает результаты статистической проверки, возникает ли статистическое отличие от нуля из-за влияние коэффициентов модели. Значимые входные поля - это такие поля, у которых очень низкие уровни значимости в выходных полях (помечаются как *знч.*).

Оценки параметров (необязательно). Показывает оценки коэффициентов уравнений, результаты проверки этих коэффициентов, отношения шансов, полученных из коэффициентов, помеченных как *Exp(B)*, и доверительные интервалы для коэффициентов для этих отношений шансов.

Асимптотическая ковариационная/корреляционная матрица (необязательно). Показывает асимптотические коварианты и/или корреляции оценок коэффициентов.

Наблюдаемые и предсказанные частоты (необязательно). Для каждого шаблона ковариат показывает наблюдаемые и предсказанные частоты для каждого значения выходного поля. Эта таблица может быть очень большой, особенно для моделей с числовыми входными полями. Если итоговая таблица окажется чрезмерно большой для практического применения, она пропускается или выводится предупреждение.

Узел PCA/фактора

Узел PCA/фактора предоставляет мощные средства сокращения числа данных для уменьшения сложности ваших данных. Предлагаются два похожих, но различных подхода.

- **Анализ главных компонентов (principal components analysis, PCA)** находит линейные комбинации входных полей, которыми главным образом определяются изменения в целом наборе полей, где компоненты ортогональны друг другу. PCA направлен на всю изменчивость, и совместную, и уникальную.
- Целью **факторного анализа** является выявление скрытых понятий или **факторов** , объясняющих структуру корреляций внутри набора наблюдаемых полей. Факторный анализ направлен только на совместную изменчивость. Уникальная дисперсия отдельных полей не рассматривается при оценке модели. Узел факторов/PCA предоставляет несколько вариантов факторного анализа.

Цель обоих подходов - найти небольшое количество производных полей, которые эффективно суммируют информацию исходного набора входных полей.

Требования. В факторных моделях и в моделях PCA можно использовать только числовые поля. Чтобы оценить факторный анализ или анализ PCA, нужно одно или несколько полей с заданной ролью *Входных* полей. Поля с заданной ролью *Назначение*, *Оба* или *Нет* игнорируются, как и не числовые поля.

Достоинства. Факторный анализ и анализ PCA могут эффективно уменьшить сложность ваших данных без существенного вреда для информационного содержимого. Эти способы могут помочь в построении более устойчивых моделей, выполняемых быстрее, чем было бы возможно для исходных входных полей.

Опции моделей узла PCA/факторной модели

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Способ извлечения. Задайте способ, который будет использоваться для сокращения количества данных.

- **Главные компоненты.** Это способ по умолчанию, использующий PCA для обнаружения компонентов, которые суммируют в себе сводную информацию входных полей.
- **Невзвешенный МНК.** При работе с этим инструментом факторного анализа находится набор факторов, которые наилучшим образом воспроизводят структуру взаимосвязей (корреляций) среди входных полей.
- **Обобщенный МНК.** Этот способ факторного анализа аналогичен невзвешенным наименьшим квадратам, но в нем используется взвешивание, чтобы ослабить влияние полей с большим вкладом уникальной изменчивости (отсутствующей в других полях).
- **Максимальное правдоподобие.** При этом варианте факторного анализа создаются факторные уравнения, которые с наибольшей вероятностью создадут наблюденную структуру взаимосвязей (корреляций) во

входных полях на основании предположений о форме этих взаимосвязей. В частности, этот способ предполагает, что данные обучения следуют многомерному нормальному распределению.

- **Факторизация главной оси.** Этот инструмент факторного анализа очень похож на способ главных компонентов, но он направлен только на совместную изменчивость.
- **Альфа факторизация.** При этом варианте факторного анализа поля рассматриваются как выборка из универсального множества потенциальных входных полей. При этом максимизируется статистическая надежность факторов.
- **Анализ образов.** Этот инструмент факторного анализа использует оценку данных для выделения общей изменчивости и нахождения описывающих ее факторов.

Дополнительные опции узла PCA/факторной модели

Если вы хорошо знакомы с факторным анализом или PCA, опции эксперта позволят точнее настроить процесс обучения. Для доступа к дополнительным опциям выберите режим **Дополнительно** на вкладке Дополнительно.

Пропущенные значения. По умолчанию IBM SPSS Modeler использует только те записи, которые содержат допустимые значения во всех полях, используемых в модели. (Эта стратегия также называется **исключение пропущенных значений целиком**.) Иногда, когда пропущенных значений много, данный подход отбрасывает непозволительно много записей, так что оставшихся данных недостаточно для генерирования хорошей модели. В таких случаях можно отменить выбор опции **Использовать только полные записи**. Тогда IBM SPSS Modeler пытается использовать для оценки модели максимум информации, включая те записи, где некоторые поля содержат пропущенные значения. (Эта стратегия также называется **попарное исключение** пропущенных значений.) Однако такое использование неполных записей иногда приводит к вычислительным проблемам при оценке моделей.

Поля. Укажите, какую матрицу использовать при оценке модели, корреляционную (по умолчанию) или ковариационную.

Максимум итераций до сходимости. Задайте максимальное количество итераций для оценки модели.

Извлечь факторы. Есть два способа выбора количества факторов для извлечения из входных полей.

- **Собственные значения.** При использовании этой опции будут оставлены все факторы или компоненты с собственными значениями, больше заданного значения критерия. **Собственные значения** измеряют возможность каждого фактора или компонента суммировать дисперсию в наборе входных полей. При использовании корреляционной матрицы в модели сохранятся все факторы или компоненты с собственными значениями больше заданного значения. При использовании ковариационной матрицы критерий - это заданное значение, умноженное на среднее собственное значение. При таком масштабировании смысл применения обеих матриц становится аналогичным.
- **Максимальное количество.** При этой опции будет сохранено заданное количество факторов или компонентов в убывающем порядке собственных значений. Другими словами, будут сохранены факторы или компоненты, соответствующие *n* максимальным собственным значениям, где *n* - это заданный критерий. По умолчанию критерий извлечения - это пять факторов/компонентов.

Матричный формат компонентов/факторов. Эти опции управляют форматом факторной матрицы (или матрицы компонентов для моделей PCA).

- **Сортировать значения.** Если выбрана эта опция, факторные нагрузки в выходных данных модели будут отсортированы численно.
- **Скрыть значения ниже.** Если выбрана эта опция, в матрице будут скрыты оценки ниже заданного порога, чтобы в ней было легче увидеть структуру.

Вращение. Эти опции позволяют управлять способом вращения для модели. Дополнительную информацию смотрите в разделе “Опции вращения узла PCA/факторной модели” на стр. 178.

Опции вращения узла PCA/факторной модели

Во многих случаях математическое вращение набора оставшихся факторов может увеличить их полезность и особенно возможность интерпретации. Выберите способ вращения:

- **Без вращения.** Опция по умолчанию. Вращение не используется.
- **Варимакс.** Ортогональный метод вращения, минимизирующий число полей с высокими нагрузками на каждый фактор. Этот метод упрощает интерпретацию факторов.
- **Прямой облимин.** Метод косоугольного (неортогонального) вращения. Когда **Дельта** равна 0 (по умолчанию), получаются косоугольные вращения. По мере того, как дельта отклоняется в отрицательную сторону, факторы становятся более ортогональными. Чтобы изменить задаваемое по умолчанию дельта (равное 0), введите число, меньшее или равное 0,8.
- **Квартимакс.** Ортогональный способ, минимизирующий количество факторов, нужных для объяснения каждого поля. При этом упрощается интерпретация наблюдаемых полей.
- **Эквимакс.** Способ вращения, представляющий из себя комбинацию Варимакс (упрощение факторов) и Квартимакс (упрощение полей). Минимизируется число полей с большими факторными нагрузками и число факторов, требуемых для объяснения поля.
- **Промакс.** Косоугольное вращение в предположении, что факторы могут коррелировать между собой. Вычисляется быстрее, чем прямое вращение облимин, поэтому полезно для больших наборов данных. **Каппа** управляет косоугольными направлениями решения (степенью корреляции факторов).

Слепок модели PCA/факторной модели

Слепок модели PCA/факторной модели представляет модель факторного анализа и модель анализа главных компонентов (principal component analysis, PCA), созданных узлом факторов/PCA. Они содержат всю информацию, захваченную обученной моделью, а также сведения о производительности и характеристиках модели.

Когда вы запускаете поток, содержащий модель факторных уравнений, узел добавляет новое поле для каждого фактора или компонента в модели. Имена новых полей получаются из имени модели с добавлением префикса *$F-* и суффикса *-n*, где *n* - это номер фактора или компонента. Например, если ваша модель называется *Фактор* и содержит три фактора, новые поля будут называться *$F-Фактор-1*, *$F-Фактор-2* и *$F-Фактор-3*.

Для лучшего понимания, что именно закодировала факторная модель, можно выполнить некоторый дополнительный исходящий анализ. Полезный способ просмотра результатов факторной модели - это изучение корреляций между факторами и входными полями с помощью узла Статистика. Это покажет, какие поля сильно нагружены какими факторами, и поможет обнаружить скрытые значения или интерпретации ваших факторов.

Оценить факторную модель можно также с использованием информации, доступной в расширенном выводе. Для просмотра расширенного вывода перейдите на вкладку **Дополнительно** программы просмотра слепков моделей. Расширенный вывод содержит много подробных сведений и информативен для пользователей с хорошим знанием факторного анализа или PCA. Дополнительную информацию смотрите в разделе “Расширенный вывод слепков моделей PCA/факторных моделей” на стр. 179.

Уравнения слепков PCA/факторных моделей

На вкладке Модель для слепка факторной модели выводится уравнение оценки факторов для каждого фактора. Оценки факторов или компонентов вычисляются умножением значения каждого входного поля на его коэффициент и суммированием всех произведений.

Сводка слепков PCA/факторных моделей

На вкладке Сводка для слепка факторной модели выводится количество факторов, сохраненных в факторной/PCA модели, а также дополнительная информация о полях и параметрах, использованных для генерирования модели. Дополнительную информацию смотрите в разделе “Просмотр слепков моделей” на стр. 41.

Расширенный вывод слепков моделей PCA/факторных моделей

Расширенный вывод для факторного анализа дает подробную информацию по оцениваемой модели и ее производительности. Большая часть содержащейся в расширенном выводе информации техническая, и требуется глубокое знание факторного анализа, чтобы правильно интерпретировать этот вывод.

Предупреждения. Обозначает любые предупреждения или потенциальные проблемы с результатами.

Общности. Показывает долю дисперсии каждого поля, которая объясняется факторами или компонентами. *Начальная* дает начальные общности с полным набором факторов (модель начинается с количества факторов, равного количеству входных полей), а *Извлечение* дает общности на основании оставшегося набора факторов.

Общая объясненная дисперсия. Показывает общую дисперсию, объясненную факторами в модели. *Начальные собственные значения* - показывает дисперсию, объясненную полным набором начальных факторов. *Суммы квадратов нагрузок извлечения* - показывает дисперсию, объясненную оставшимися в модели факторами. *Суммы квадратов нагрузок вращения* - показывает дисперсию, объясненную повернутыми факторами. Обратите внимание на то, что для косоугольных вращений *Суммы квадратов нагрузок вращения* показывает только суммы квадратов нагрузок, но не показывает процентные доли дисперсии.

Матрица факторов (или компонентов). Показывает корреляции между входными полями и факторами без вращения.

Матрица повернутых факторов (или компонентов). Показывает корреляции между входными полями и повернутыми факторами для ортогональных вращений.

Матрица паттерна. Показывает частные корреляции между входными полями и повернутыми факторами для косоугольных вращений.

Матрица структуры. Показывает простые корреляции между входными полями и повернутыми факторами для косоугольных вращений.

Корреляционная матрица факторов. Показывает корреляции среди факторов для косоугольных вращений.

Узел дискриминанта

При дискриминантном анализе происходит создание прогностической модели для принадлежности к группе. Данная модель строит дискриминантную функцию (или, когда групп больше двух, набор дискриминантных функций) в виде линейной комбинации предикторных переменных, обеспечивающую наилучшее разделение групп. Эти функции строятся по набору наблюдений, для которых их принадлежность к группам известна, и могут в дальнейшем применяться к новым наблюдениям с известными значениями предикторных переменных, но неизвестной групповой принадлежностью.

Пример. Телекоммуникационная компания может использовать дискриминантный анализ для классификации клиентов по группам на основании данных использования услуг. Это позволяет оценить потенциальных клиентов и сосредоточиться на тех из них, кто наиболее вероятно окажется в группах самых ценных клиентов.

Требования. Вам требуется одно или несколько входных полей и ровно одно поле назначения. Назначение должно быть категориальным полем (с уровнем измерения *Флаг* или *Номинал*) со строковой или

целочисленной системой хранения. (При необходимости систему хранения можно преобразовать, используя узел Заполнитель или узел Извлечение). Поля с заданными значениями *Оба* или *Нет* игнорируются. У используемых в модели полей должны быть полностью конкретизированы типы.

Достоинства. Дискриминантный анализ и Логистическая регрессия - это две подходящие модели классификации. Однако дискриминантный анализ использует больше предположений о входных полях, например, что у них нормальное распределение и их значения должны быть количественными, и в случае выполнения этих требований результаты могут оказаться лучше, особенно для небольших размеров выборки.

Опции моделей узла Дискриминант

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Создание моделей расщепления. Строит отдельную модель для каждого возможного значения входных полей, заданных как поля расщепления. Дополнительную информацию смотрите в разделе “Построение моделей расщепления” на стр. 28.

Метод. Доступны следующие опции для ввода в модель предикторов:

- **Ввод.** Это способ по умолчанию, при котором все члены непосредственно вводятся в уравнение. Члены, которые не вносят существенного вклада в предсказательную силу модели, не добавляются.
- **Пошаговый.** Исходная модель - это простейшая модель без членов модели (кроме константы) в уравнении. На каждом шаге оцениваются члены, которые еще не добавлены в модель, и добавляются те из них, которые вносят наиболее существенный вклад в предсказательную силу модели.

Примечание: У пошагового способа есть сильная тенденция к переобучению данных обучения. При использовании этих способов особенно важно проверить приемлемость полученной модели или на новых данных, или на существующей тестовой выборке.

Дополнительные опции узла Дискриминант

Если вы хорошо знакомы с дискриминантным анализом, дополнительные опции позволят тонко настроить процесс обучения. Для доступа к дополнительным опциям на вкладке Дополнительно задайте для **Режима** значение **Дополнительный**.

Априорные вероятности. Эта функция определяет настройку классификационных коэффициентов в соответствии с априорным знанием принадлежности к группе.

- **Все группы равны.** Предполагаются равные вероятности для всех групп, что не оказывает влияния на коэффициенты.
- **Вычислить по размерам групп.** Априорные вероятности принадлежности к группе зависят от размера наблюдаемой группы в выборке. Например, если 50% наблюдений из области анализа попадает в первую группу, 25% во вторую и 25% в третью, классификационные коэффициенты настраиваются для увеличения правдоподобия принадлежности к первой группе по отношению ко второй и третьей.

Ковариационная матрица. Вы можете выбрать один из двух способов классификации наблюдений - либо по внутригрупповой ковариационной матрице, либо по ковариационным матрицам для отдельных групп.

- *Внутри групп*. Для классификации наблюдений используется объединенная внутригрупповая ковариационная матрица.
- *Для отдельных групп*. Для классификации используются ковариационные матрицы для отдельных групп. Так как классификация производится на основе дискриминантных функций, а не на основе исходных переменных, выбор этого параметра не всегда равноценен квадратичной дискриминации.

Вывод. Эти опции позволяют затребовать дополнительные статистические данные, которые будут показаны в расширенном выводе слепка модели, построенном узлом. Дополнительную информацию смотрите в разделе “Опции вывода узла Дискриминант”.

Пошаговые опции. Эти опции позволяют вам управлять критериями для добавления и удаления полей при пошаговом способе оценки. (Эта кнопка отключена, если выбран способ Ввод). Дополнительную информацию смотрите в разделе “Опции шагового отбора узла Дискриминант” на стр. 182.

Опции вывода узла Дискриминант

Выберите дополнительные выходные данные, которые вы хотите использовать в расширенном выводе слепка модели логистической регрессии. Для просмотра расширенного вывода перейдите на слепок модели и выберите вкладку **Дополнительно**. Дополнительную информацию смотрите в разделе “Расширенный вывод слепков дискриминантных моделей” на стр. 183.

Описательные статистики. Доступны параметры: средние значения (включая стандартные отклонения), одномерный дисперсионный анализ, а также *M* -критерий Бокса.

- *Средние*. Выводятся общее и групповые средние, а также стандартные отклонения для независимых переменных.
- *Однофакторный дисперсионный анализ*. Проводит однофакторный дисперсионный анализ для проверки гипотезы о равенстве групповых средних для каждой независимой переменной.
- *M Бокса*. Критерий равенства групповых ковариационных матриц. Если p не значимо, а выборка достаточно велика, то нет достаточных свидетельств того, что матрицы различаются. Этот критерий чувствителен к отклонениям от многомерной нормальности.

Коэффициенты функции. Возможен вывод классификационных коэффициентов Фишера и нестандартизованных коэффициентов.

- *Фишера*. Коэффициенты классифицирующей функции Фишера, которые можно напрямую использовать для классификации. Для каждой группы создается отдельный набор коэффициентов, при этом наблюдение относится к группе, которой соответствует наибольшее значение дискриминантной функции (значение классифицирующей функции).
- *Нестандартизованные*. Выводит нестандартизованные коэффициенты дискриминантной функции.

Матрицы. Доступными матрицами коэффициентов для независимых переменных являются: внутригрупповая корреляционная матрица, внутригрупповая ковариационная матрица, ковариационные матрицы для отдельных групп и общая ковариационная матрица.

- *Внутригрупповая корреляция*. Выводится объединенная внутригрупповая корреляционная матрица, полученная путем усреднения ковариационных матриц отдельных групп перед вычислением корреляций.
- *Внутригрупповая ковариация*. Выводится объединенная внутригрупповая ковариационная матрица, которая может отличаться от общей ковариационной матрицы. Матрица вычисляется путем усреднения отдельных ковариационных матриц для всех групп.
- *Групповые ковариации*. Для каждой группы выводится отдельная ковариационная матрица.
- *Общая ковариация*. Выводится ковариационная матрица для всех наблюдений, как если бы они были из одной выборки.

Классификация. Следующий вывод относится к результатам классификации.

- *Поточечные результаты*. Коды для фактической группы, предсказанной группы, апостериорные вероятности и значения дискриминантной функции выводятся для каждого наблюдения.
- *Итоговая таблица*. Числа наблюдений, правильно и неправильно отнесенных к каждой из групп в дискриминантном анализе. Это иногда называют матрицей перекрестной классификации.
- *Классификация с удалением по одной точке*. Каждое наблюдение при анализе классифицируется с помощью функции, полученной по всем остальным наблюдениям, кроме данного. Используется также название "U-метод".

- *Территориальная карта*. График, на который нанесены границы, позволяющие отнести наблюдение к группе на основании значений функции. Числа соответствуют группам, по которым распределяют наблюдения. Среднее каждой группы обозначено звездочкой внутри границ этой группы. Если есть только одна дискриминантная функция, диаграмма не выводится.
- *Объединенные группы*. Строится диаграмма рассеяния значений первых двух дискриминантных функций для наблюдений из всех групп. Если есть только одна дискриминантная функция, вместо диаграммы рассеяния выводится гистограмма.
- *Для отдельных групп*. Диаграмма рассеяния значений первых двух дискриминантных функций строится для каждой группы в отдельности. Если есть только одна дискриминантная функция, вместо диаграммы рассеяния выводится гистограмма.

Пошаговый. Отчет о шагах выводит статистики для всех переменных после каждого шага; **F для попарных расстояний** выводит матрицу попарных *F*-отношений для каждой пары групп. *F*-отношения для критериев значимости расстояний Махаланобиса между группами.

Опции шагового отбора узла Дискриминант

Метод. Выберите статистику, которая будет использоваться для введения или удаления новых переменных. Возможными альтернативами являются лямбда Уилкса, необъясненная дисперсия, расстояние Махаланобиса, наименьшее *F* отношение и *V* Рао. Выбрав *V* Рао, можно задать минимальное приращение *V* , необходимое для включения переменной.

- *Лямбда Уилкса*. Метод отбора переменных в шаговом дискриминантном анализе, отбирающий переменные для ввода в уравнение на основании того, насколько они уменьшают значение "лямбда" Уилкса. На каждом шаге вводится переменная, минимизирующая это значение.
- *Необъясненная дисперсия*. На каждом шаге вводится переменная, минимизирующая сумму необъясненной изменчивости между группами.
- *Расстояние Махаланобиса*. Мера того, насколько значения наблюдений для независимых переменных отклоняются от среднего по всем наблюдениям. Большое расстояние Махаланобиса означает, что наблюдение содержит экстремальные значения в одной или более независимых переменных.
- *Наименьшее F отношение*. Метод отбора переменных в шаговом анализе, основанный на максимизации F-отношения, вычисленного по расстоянию Махаланобиса между группами.
- *V Рао*. Мера различий между групповыми средними. Также называется следом Лоули-Хотеллинга. На каждом шаге вводится та переменная, которая максимизирует прирост индекса V Рао. Выбрав этот параметр, введите минимальное значение, которое должна иметь переменная, чтобы быть включенной в анализ.

Критерии. Возможные альтернативы: **Использовать F-значение** и **Использовать вероятность F**. Введите значения для включения и удаления переменных.

- *Использовать F-значение*. Переменная вводится в модель, если ее F-значение превышает заданное значение включения, и исключается, если ее F-значение меньше значения исключения. Значение включения должно превосходить значение удаления, оба должны быть положительными. Если необходимо ввести в модель больше переменных, снизьте порог включения. Чтобы исключить из модели большее число переменных, увеличьте порог исключения.
- *Использовать вероятность F*. Переменная вводится в модель, если наблюдаемый уровень значимости ее F-значения меньше заданного порога включения, и исключается, если этот уровень значимости больше порога исключения. Порог включения должен быть меньше порога удаления, они оба должны быть положительными. Если необходимо включить в модель больше переменных, увеличьте порог включения. Чтобы исключить из модели большее число переменных, снизьте порог исключения.

Слепок дискриминантной модели

Слепки дискриминантных моделей представляют уравнения, оцененные узлами Дискриминант. Они содержат всю собранную дискриминантной моделью информацию, а также информацию о структуре и производительности модели.

При запуске потока, содержащего слепок дискриминантной модели, узел добавляет два новых поля, содержащих предсказание модели и связанную вероятность. Имена новых полей получаются из имени выходного поля, которое предсказывается, с добавлением префикса *$D-* для предсказанной категории и префикса *$DP-* для связанной вероятности. Например, если имя выходного поля *colorpref*, новые поля будут называться *$D-colorpref* и *$DP-colorpref*.

Генерирование узла Фильтр. Меню Генерировать позволяет вам создавать новый узел Фильтр для передачи входных полей на основе результатов модели.

Важность предиктора

Диаграмма, обозначающая относительную важность каждого предиктора в оцениваемой модели, может быть дополнительно также показана на вкладке Модель. Обычно вам потребуется направить усилия по моделированию на рассмотрение наиболее важных предикторов, отбрасывая или игнорируя самые не важные. Обратите внимание на то, что эта диаграмма доступна только в том случае, если перед генерированием модели на вкладке Анализ выбрана опция **Вычислять важность предикторов**. Дополнительную информацию смотрите в разделе “Важность предиктора” на стр. 42.

Расширенный вывод слепков дискриминантных моделей

Расширенный вывод для дискриминантного анализа дает подробную информацию об оцениваемой модели и ее производительности. Большая часть содержащейся в расширенном выводе информации техническая, и требуется глубокое знание дискриминантного анализа, чтобы правильно интерпретировать этот вывод. Дополнительную информацию смотрите в разделе “Опции вывода узла Дискриминант” на стр. 181.

Параметры слепков дискриминантных моделей

На вкладке Параметры слепка дискриминантной модели можно получить оценки склонности при скоринге модели. Эта вкладка доступна только для моделей с флаговыми полями назначения и только после того, как слепок модели добавлен в поток.

Вычислить простые оценки склонности. Для моделей с флаговым полем назначения (которое возвращает предсказание Да или Нет) можно затребовать оценку склонности, которая обозначает правдоподобие правильного выходного значения, заданного для поля назначения. Это дополнение к другим предсказанным и доверительным значениям, которые можно сгенерировать при скоринге.

Вычислить скорректированные оценки склонности. Простые оценки склонности основываются только на данных обучения и могут выглядеть чрезмерно оптимистическими из-за тенденции многих моделей чрезмерно точно подгонять эти данные под модель. Корректировка склонностей направлена на компенсацию этого эффекта путем проверки выполнения модели на испытательном и проверочном разделах. Для этой опции требуется, чтобы поле разделения было определено в потоке и скорректированные оценки склонности были включены на узле моделирования до генерирования модели.

Сводка слепков дискриминантных моделей

На вкладке Сводка для слепка дискриминантной модели выводятся поля и параметры, использованные для генерирования модели. Кроме этого, если вы запускали узел Анализ, присоединенный к этому узлу моделирования, информация этого анализа также будет выводится в данном разделе. За общей информацией по браузеру моделей обратитесь к разделу “Просмотр слепков моделей” на стр. 41.

Узел обобщенной линейной модели

Обобщенная линейная модель расширяет общую линейную модель, связывая зависимую переменную с факторами и ковариатами посредством задаваемой функции. Более того, модель допускает наличие у зависимой переменной распределения, отличающегося от нормального. Охватываются широко используемые статистические модели, такие как линейная регрессия, для откликов с нормальным распределением, логистические модели для двоичных данных, логлинейные модели для счетных данных, модели с дополняющим двойным логарифмированием для интервал-цензурированных данных выживания плюс многие другие статистические модели, вплоть до их очень общих редакций.

Примеры. Судоходная компания может использовать обобщенные линейные модели, чтобы подобрать нужную регрессию Пуассона для описания количества повреждений судов нескольких типов, построенных в разное время; полученная модель может помочь в определении, суда каких типов чаще всего повреждаются.

Страховая компания может использовать обобщенные линейные модели, чтобы подобрать нужную регрессию гамма для описания страховых исков о повреждении автомобилей; полученная модель может помочь в определении факторов, дающих наибольший размер исков.

Медики могут использовать обобщенные линейные модели, чтобы с помощью дополняющей лог-лог регрессии описать интервал-цензурированные данные выживания и предсказать время повторного обращения за медицинской помощью.

При работе обобщенных линейных моделей строятся уравнения, связывающие значения входных полей со значениями выходного поля. Когда модель сгенерирована, ее можно использовать для оценки значений по новым данным. Для каждой записи вероятность принадлежности вычисляется для каждой возможной выходной категории. Категория назначения с наибольшей вероятностью назначается как предсказанное выходное значение для данной записи.

Требования. Вам требуется одно или несколько входных полей и ровно одно поле назначения (у которого может быть уровень измерения *Количественный* или *Флаг*) с двумя или более категориями. У используемых в модели полей должны быть полностью конкретизированы типы.

Достоинства. Обобщенная линейная модель очень гибка, но процесс выбора структуры модели не автоматизирован, то есть требует некоторого уровня знакомства с вашими данными, что не требуется для алгоритмов "черного ящика".

Опции полей узла обобщенной линейной модели

Кроме этого, для опций поля назначения, входных полей и пользовательских разделов, обычно предлагаемых на вкладках Поля узлов моделирования (смотрите раздел “Моделирование опций полей узла” на стр. 31), узел обобщенной линейной модели обеспечивает следующие дополнительные функциональные возможности.

Использовать поле веса. Параметр шкалы является оценочным параметром модели, связанным с дисперсией ответов. Масштабные веса - это "известные" значения, которые могут изменяться от наблюдения к наблюдению. Если задана переменная масштабного веса, относящийся к дисперсии ответов масштабный коэффициент делится на значение веса для каждого наблюдения. Записи с неположительными (или пропущенными) значениями веса по шкале не используются в анализе.

Поле назначения представляет количество событий, происходящих в серии попыток. Если откликом является количество событий, встречающихся в наборе испытаний, поле назначения содержит количество событий, причем можно выбрать дополнительную переменную, содержащую количество попыток. Помимо этого, если количество испытаний одинаково для всех групп, испытания могут быть заданы при помощи фиксированного значения. Число испытаний не должно превышать числа событий для каждой записи. События должны представлять собой неотрицательные, а испытания - положительные целые числа.

Опции моделей узла Обобщенная линейная модель

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Создание моделей расщепления. Строит отдельную модель для каждого возможного значения входных полей, заданных как поля расщепления. Дополнительную информацию смотрите в разделе “Построение моделей расщепления” на стр. 28.

Тип модели. Есть две опции для типа модели, которая будет строиться. При использовании опции **Только главные эффекты** модель будет включать в себя только входные поля по отдельности, но не будет проверять взаимодействия (мультипликативные эффекты) между полями. Опция **Главные эффекты и все двусторонние взаимодействия** включает в себя все двусторонние взаимодействия, а также главные эффекты входных полей.

Смещение. Критерий смещения является "структурным" предиктором. Его коэффициент не оценивается моделью, но предполагается, что у него значение 1; поэтому значения смещения просто добавляются к значениям линейного предиктора назначения. Это особенно полезно в регрессионных моделях Пуассона, где у каждого наблюдения могут быть различные уровни влияния на исследуемое событие.

Например, при моделировании уровня автомобильных аварий для конкретных водителей есть важное различие между между водителем, побывавшим в одной аварии за три года вождения, и водителем, попавшим в аварию один раз за 25 лет. Количество аварий может быть моделировано как отклик Пуассона или отрицательный биномиальный отклик с логарифмической связью, если натуральный логарифм опыта (стажа) водителя включен как критерий смещения.

Для других типов сочетаний распределения и связи потребовались бы другие преобразования переменной смещения.

Примечание: Если используется поле смещения переменной, оно не должно использоваться также как входное поле. При необходимости задайте для поля смещения роль **Нет** на вышележащем узле источника или узле Тип.

Базовая категория для назначения флага.

Для бинарного отклика можно выбрать опорную категорию для зависимой переменной. Она может повлиять на определенный вывод (например, для оценки параметров и сохраненных значений), но не должна изменять подгонку модели. Например, если бинарный отклик принимает значения 0 и 1:

- По умолчанию указанная процедура назначает опорной категорией последнюю категорию (с наибольшим значением) или 1. В этой ситуации вероятности сохраненной модели будут оценивать шанс принятия данным наблюдением значения 0, а оценки параметров должны интерпретироваться как относящиеся к правдоподобию категории 0.
- Если в качестве опорной категории указать первую категорию (с наименьшим значением) или 0, вероятности сохраненной модели будут оценивать шанс принятия данным наблюдением значения 1.
- Если задать пользовательскую категорию и используемую переменную с определенными для нее метками, опорную категорию можно будет указать, выбрав значение в списке. Это может оказаться удобным, если (в середине определения модели) вы не вспомните в точности, как кодировалась та или иная переменная.

Включить в модель свободный член. Обычно в модель включают свободный член. Если вы предполагаете, что данные проходят через начало координат, свободный член можно исключить.

Опции эксперта узла обобщенной линейной регрессии

Если вы хорошо знакомы с обобщенными линейными моделями, опции эксперта позволят точнее настроить процесс обучения. Для доступа к дополнительным опциям на вкладке Дополнительно задайте для **Режима** значение **Дополнительный**.

Распределение и функция связи поля назначения

Распределение.

Этот вариант выбора задает распределение зависимой переменной. Возможность задать распределение, отличающееся от нормального, и нетождественную функцию связи - существенное усовершенствование обобщенной линейной модели относительно общей линейной модели. Существует множество возможных сочетаний функций связей распределения, и несколько могут оказаться приемлемыми для любого данного

набора данных, поэтому при выборе можно руководствоваться априорными теоретическими соображениями или тем, какое сочетание кажется наиболее подходящим.

- **Биномиальный.** Это распределение применимо только к переменным, представляющим бинарный отклик или количество событий.
- **Гамма.** Это распределение применимо к переменным с положительными значениями масштаба со скосом в направлении больших значений. Если значение данных меньше или равно 0 либо пропущено, соответствующее ему наблюдение в анализе не используется.
- **Обратное нормальное распределение.** Это распределение применимо к переменным с положительными значениями масштаба со скосом в направлении больших значений. Если значение данных меньше или равно 0 либо пропущено, соответствующее ему наблюдение в анализе не используется.
- **Отрицательное биномиальное распределение.** Это распределение может быть представлено в виде числа испытаний, необходимых для получения *k* успехов, и применимо к переменным с неотрицательными целыми значениями. Если значение данных не целое, меньше 0 или пропущено, соответствующее ему наблюдение в анализе не используется. Фиксированное значение вспомогательного параметра отрицательного биномиального распределения может быть любым неотрицательным числом. Если для вспомогательного параметра задано значение 0, использование этого распределения эквивалентно использованию распределения Пуассона.
- **Нормальное.** Это распределение применимо к количественным переменным, значения которых около центрального (среднего) значения принимают симметричное распределение колоколообразной формы. Зависимая переменная должна быть числовой.
- **Пуассона.** Это распределение может быть представлено в виде количества событий, произошедших в фиксированный период времени, и применимо к переменным с неотрицательными целыми значениями. Если значение данных не целое, меньше 0 или пропущено, соответствующее ему наблюдение в анализе не используется.
- **Распределение Твиди.** Это распределение применимо к переменным, которые могут быть представлены смесями Пуассона гамма-распределений; распределение считается "смешанным" в том отношении, что сочетает в себе свойства непрерывных распределений (принимает неотрицательные действительные значения) и дискретных распределений (с положительной вероятностной мерой для одного значения 0). Зависимая переменная должна быть числовой, с неотрицательными значениями данных. Если значение данных меньше нуля или пропущено, соответствующее ему наблюдение в анализе не используется. Фиксированное значение параметра распределения Твиди может быть любым числом больше единицы и меньше двух.
- **Полиномиальное.** Это распределение применимо к переменным, представляющим порядковый отклик. Зависимая переменная может быть числовой или текстовой, и у нее должно быть по крайней мере два различных допустимых значения данных.

Функции связи.

Связывающая функция служит для преобразования зависимых переменных для расчета модели. Доступны следующие функции:

- **Тождество.** $f(x)=x$. Зависимая переменная не преобразуется. Эту связь можно использовать с любым распределением.
- **Дополнительный логарифм-логарифм.** $f(x)=\log(-\log(1-x))$. Применяется только с биномиальным распределением.
- **Кумулятивное Коши.** $f(x) = \tan(\pi\ (x - 0.5))$; применяется к кумулятивной вероятности каждой категории отклика. Применяется только с полиномиальным распределением.
- **Кумулятивное дополняющее лог-лог.** $f(x)=\ln(-\ln(1-x))$; применяется к кумулятивной вероятности каждой категории отклика. Применяется только с полиномиальным распределением.
- **Кумулятивное логит.** $f(x)=\ln(x / (1-x))$; применяется к кумулятивной вероятности каждой категории отклика. Применяется только с полиномиальным распределением.
- **Кумулятивное отрицательное лог-лог.** $f(x)=-\ln(-\ln(x))$; применяется к кумулятивной вероятности каждой категории отклика. Применяется только с полиномиальным распределением.

- **Кумулятивное пробит.** $f(x)=\Phi^{-1}(x)$; применяется к кумулятивной вероятности каждой категории отклика, где Φ^{-1} - кумулятивная функция обратного стандартного нормального распределения. Применяется только с полиномиальным распределением.
- **Логарифмическая.** $f(x)=\log(x)$. Эту связь можно использовать с любым распределением.
- **Дополняющее лог.** $f(x)=\log(1-x)$. Применяется только с биномиальным распределением.
- **Логит.** $f(x)=\log(x / (1-x))$. Применяется только с биномиальным распределением.
- **Отрицательное биномиальное распределение.** $f(x)=\log(x / (x+k^{-1}))$; где k - вспомогательный параметр отрицательного биномиального распределения. Применяется только с отрицательным биномиальным распределением.
- **Отрицательный Log-log.** $f(x)=-\log(-\log(x))$. Применяется только с биномиальным распределением.
- **Степенное.** $f(x)=[(x/(1-x))^{\alpha}-1]/\alpha$; если $\alpha \neq 0$. $f(x)=\log(x)$, если $\alpha=0$. α - обязательная спецификация числа, которое должно быть действительным числом. Применяется только с биномиальным распределением.
- **Пробит.** $f(x)=\Phi^{-1}(x)$; где Φ^{-1} - кумулятивная функция обратного стандартного нормального распределения. Применяется только с биномиальным распределением.
- **Степенная.** $f(x)=x^{\alpha}$, если $\alpha \neq 0$. $f(x)=\log(x)$; если $\alpha=0$. α - обязательная спецификация числа, которое должно быть действительным числом. Эту связь можно использовать с любым распределением.

Анализ важности независимых переменных. Управляющие элементы в этой группе позволяют задать значения параметров, когда выбраны опции некоторого распределения.

- **Параметр для отрицательного биномиального распределения.** Для отрицательного биномиального распределения выберите опцию задания значения - задать самостоятельно или позволить системе предоставить оценочное значение.
- **Параметр для твиди.** Для распределения Твиди задайте для фиксированного значения число от 1,0 до 2,0.

  Оценка параметра. Элементы управления в этой группе позволяют определить методы оценок и задать начальные значения для оценок параметров.

  – **Метод.** Можно выбрать метод оценок параметров. Выберите метод Ньютона-Рафсона, метод скоринга Фишера либо гибридный метод, при котором сначала выполняются итерации скоринга Фишера, затем происходит переключение на метод Ньютона-Рафсона. Если сходимость будет получена на фазе скоринга Фишера гибридного метода до достижения максимального числа итераций Фишера, алгоритм перейдет к методу Ньютона-Рафсона.

  – **Метод масштабных коэффициентов.** Можно выбрать метод оценки масштабных коэффициентов. Метод максимального правдоподобия совместно оценивает масштабный коэффициент с эффектами модели; учтите, что если у отклика отрицательное биномиальное распределение, распределение Пуассона или биноминальное распределение, эта опция недопустима. Опции отклонения и хи-квадрата Пирсона оценивают масштабный коэффициент, исходя из значения этих статистик. Другой вариант - задать фиксированное значение масштабного коэффициента.
- **Матрица ковариаций.** Основанный на модели оценщик - это обобщенная обратная матрица Гессе с обратным знаком. Робастный оценщик (его называют также оценщиком Хубера, Уайта или сандвича) - это "исправленный" на основе модели оценщик, представляющий согласованную оценку ковариации, даже когда спецификация дисперсии и функций связи задана неверно.

Итерации. Эти опции позволяют вам управлять параметрами для сходимости модели. Дополнительную информацию смотрите в разделе “Итерации обобщенных линейных моделей” на стр. 188.

Вывод. Эти опции позволяют затребовать дополнительные статистические данные, которые будут показаны в расширенном выводе слепка модели, построенном узлом. Дополнительную информацию смотрите в разделе “Расширенный вывод для обобщенных линейных моделей” на стр. 188.

Допуск для вырожденности. Вырожденные (или необратимые) матрицы содержат линейно зависимые столбцы, которые могут вызвать серьезные ошибки алгоритма оценки. Даже почти вырожденные матрицы могут привести к неудовлетворительным результатам, поэтому процедура будет обрабатывать матрицу, определитель которой меньше допуска вырожденности. Задайте положительное значение.

Итерации обобщенных линейных моделей

Для оценки обобщенной линейной модели можно задать параметры сходимости.

Итерации. Доступны следующие параметры:

- **Максимум итераций.** Максимальное число итераций, которое будет выполнять алгоритм. Задайте неотрицательное целое число.
- **Максимум делений шага на 2.** Для каждой итерации размер шага уменьшается с коэффициентом 0,5, пока возрастает логарифм отношения правдоподобия или достигнуто максимальное количество дроблений шага. Задайте целое положительное число.
- **Проверить разделение точек данных.** Если выбрана эта опция, алгоритм выполняет проверки, обеспечивающие уникальность значений оценок параметров. Разделение имеет место, если указанная процедура может сгенерировать модель, правильно классифицирующую каждое наблюдение. Эта опция доступна для биномиальных откликов с двоичным форматом .

Критерии сходимости. Доступны следующие параметры

- **Сходимость параметров.** Если выбрана эта опция, алгоритм останавливается после итерации, в которой абсолютное или относительное изменение в оценках параметров становится меньше указанного значения (оно должно быть положительным).
- **Сходимость Log-правдоподобия.** Если выбрана эта опция, алгоритм останавливается после итерации, где абсолютное или относительное изменение в функции логарифмического правдоподобия становится меньше указанного значения (оно должно быть положительным).
- **Сходимость Гессиана.** Для спецификации Абсолютная предполагается сходимость, если статистика на основе сходимости гессиана стала меньше заданного положительного значения. Для спецификации Относительная предполагается сходимость, если указанная статистика стала меньше произведения заданного положительного значения на абсолютное значение логарифмического правдоподобия.

Расширенный вывод для обобщенных линейных моделей

Выберите дополнительные выходные данные, которые вы хотите использовать в расширенном выводе слепка обобщенной линейной модели. Для просмотра расширенного вывода перейдите на слепок модели и выберите вкладку **Дополнительно**. Дополнительную информацию смотрите в разделе “Расширенный вывод слепков обобщенных линейных моделей” на стр. 190.

Доступен следующий вывод:

- **Сводка обработки наблюдений.** Выводит число и процент наблюдений, включаемых и исключаемых из анализа и таблицы Коррелированная структура данных.
- **Описательные статистики.** Выводит описательную статистику и сводную информацию о зависимой переменной, ковариатах и факторах.
- **Информация о модели.** Выводится имя набора данных, зависимая переменная или переменные событий и испытаний, переменная смещения, переменные взвешенного значения, распределение вероятностей и функция связи.
- **Статистики согласия.** Выводится отклонение и масштабированное отклонение, хи-квадрат Пирсона и масштабированный хи-квадрат Пирсона, логарифмическое правдоподобие, информационный критерий Акаике (AIC), скорректированный информационный критерий Акаике конечной выборки (AICC), байесовский информационный критерий (BIC) и согласованный информационный критерий Акаике (CAIC).
- **Итожащие статистики для модели.** Выводятся критерии подгонки модели, включая статистику правдоподобия для универсального критерия подгонки модели и статистику для контрастов типа I или III для каждого эффекта.
- **Оценки параметров.** Выводятся оценки параметров и соответствующие им статистики критериев и доверительные интервалы. Помимо необработанных оценок параметров, можно вывести их экспоненту.
- **Ковариационная матрица для оценки параметра.** Выводится ковариационная матрица оцененных параметров.

- **Корреляционная матрица для оценки параметра.** Выводится корреляционная матрица оцененных параметров.
- **Матрицы коэффициентов контрастов (L).** Выводятся коэффициенты контрастов для эффектов по умолчанию и для оцененных маргинальных средних (если они затребованы на вкладке OM-средние).
- **Общие функции, допускающие оценку.** Выводятся матрицы для генерирования матриц коэффициентов контрастов (L).
- **История итераций.** На экран выводится хронология итераций для оценок параметров и логарифмического правдоподобия плюс на печать выводится последняя оценка вектор-градиента и матрицы Гессе. В таблице хронологии итераций выводятся оценки параметров для каждых *n* итераций, начиная с $0^{-й}$ итерации (начальные оценки), где *n* - значение интервала печати. Если затребована хронология итераций, на экран всегда выводится последняя итерация, независимо от *n*.
- **Критерий множителей Лагранжа.** Выводит статистику критерия множителей Лагранжа для оценки валидности масштабного коэффициента, вычисляемого с использованием хи-квадрат отклонений или Пирсона или задаваемого конкретным числом для распределений - нормального, гамма и обратного Гауссового. Для отрицательного биномиального распределения проверяется дополнительный фиксированный параметр.

Эффекты модели. Доступны следующие параметры:

- **Тип анализа.** Задайте тип анализа для обработки. Анализ типа I обычно приемлем для использования при наличии предварительных причин упорядочения предикторов в модели, тогда как анализ типа III более общеприменим. Вычисляется статистический критерий Вальда или отношения правдоподобия на основе выбора в группе Статистика хи-квадрат.
- **Доверительные интервалы.** Задайте доверительный уровень больше 50 и меньше 100. Интервалы Вальда основаны на допущении, что у параметров асимптотическое нормальное распределение; интервалы профильного правдоподобия более точны, но могут быть более затратны по вычислениям. Уровень допуска для интервалов профильного правдоподобия используется в качестве критериев для остановки итерационного алгоритма, вычисляющего интервалы.
- **Функция логарифма отношения правдоподобия.** Эта опция управляет форматом вывода функции логарифмического правдоподобия. Полная функция содержит дополнительный член, постоянный относительно оценок параметров, на которые он никак не влияет, и в некоторых программных продуктов исключен из вывода.

Слепок обобщенной линейной модели

Слепок обобщенной линейной модели представляет уравнения, оцененные узлом обобщенной линейной модели. Они содержат всю собранную моделью информацию, а также информацию о структуре и производительности модели.

При запуске потока, содержащего слепок обобщенной линейной модели, узел добавляет новые поля, содержимое которых зависит от природы поля назначения:

- **Флаговое поле назначения.** Добавляет поля, содержащие предсказанную категорию и связанную вероятность, а также вероятности для каждой категории. Имена первых двух полей получаются из имени выходного поля, которое предсказывается, с добавлением префикса *$G-* для предсказанной категории и префикса *$GP-* для связанной вероятности. Например, если имя выходного поля *default*, новые поля будут называться *$G-default* и *$GP-default*. Имена следующих двух дополнительных полей состоят из значений выходного поля и префикса *$GP-*. Например, если разрешенные значения поля *default* - это *Да* и *Нет*, новые поля будут называться *$GP-Да* и *$GP-Нет*.
- **Количественное поле назначения.** Добавляются поля, содержащие предсказанное среднее и среднеквадратичную ошибку.
- **Количественное поле назначения, представляющее количество событий в ряду испытаний.** Добавляются поля, содержащие предсказанное среднее и среднеквадратичную ошибку.

- **Порядковое поле назначения.** Добавляются поля, содержащие предсказанную категорию и связанную вероятность для каждого значения упорядоченного набора. Имена этих полей получаются из значения предсказываемого упорядоченного набора с префиксом *$G-* для предсказанной категории и с префиксом *$GP-* для связанной вероятности.

Генерирование узла Фильтр. Меню Генерировать позволяет вам создавать новый узел Фильтр для передачи входных полей на основе результатов модели.

Важность предиктора

Диаграмма, обозначающая относительную важность каждого предиктора в оцениваемой модели, может быть дополнительно также показана на вкладке Модель. Обычно вам потребуется направить усилия по моделированию на рассмотрение наиболее важных предикторов, отбрасывая или игнорируя самые не важные. Обратите внимание на то, что эта диаграмма доступна только в том случае, если перед генерированием модели на вкладке Анализ выбрана опция **Вычислять важность предикторов**. Дополнительную информацию смотрите в разделе “Важность предиктора” на стр. 42.

Расширенный вывод слепков обобщенных линейных моделей

Расширенный вывод для обобщенной линейной модели предоставляет подробную информацию об оцениваемой модели и ее производительности. Большая часть содержащейся в расширенном выводе информации техническая, и требуется глубокое знание этого типа анализа, чтобы правильно интерпретировать этот вывод. Дополнительную информацию смотрите в разделе “Расширенный вывод для обобщенных линейных моделей” на стр. 188.

Параметры слепков обобщенных линейных моделей

На вкладке Параметры слепка обобщенной линейной модели можно получить оценки склонности при скоринге модели. Эта вкладка доступна только для моделей с флаговыми полями назначения и только после того, как слепок модели добавлен в поток.

Вычислить простые оценки склонности. Для моделей с флаговым полем назначения (которое возвращает предсказание Да или Нет) можно затребовать оценку склонности, которая обозначает правдоподобие правильного выходного значения, заданного для поля назначения. Это дополнение к другим предсказанным и доверительным значениям, которые можно сгенерировать при скоринге.

Вычислить скорректированные оценки склонности. Простые оценки склонности основываются только на данных обучения и могут выглядеть чрезмерно оптимистическими из-за тенденции многих моделей чрезмерно точно подгонять эти данные под модель. Корректировка склонностей направлена на компенсацию этого эффекта путем проверки выполнения модели на испытательном и проверочном разделах. Для этой опции требуется, чтобы поле разделения было определено в потоке и скорректированные оценки склонности были включены на узле моделирования до генерирования модели.

Сводка слепков обобщенных линейных моделей

На вкладке Сводка для слепка обобщенной линейной модели выводятся поля и параметры, использованные для генерирования модели. Кроме этого, если вы запускали узел Анализ, присоединенный к этому узлу моделирования, информация этого анализа также будет выводится в данном разделе. За общей информацией по браузеру моделей обратитесь к разделу “Просмотр слепков моделей” на стр. 41.

Обобщенные линейные смешанные модели

Узел GLMM

Этот узел служит для создания обобщенной смешанной линейной модели (generalized linear mixed model, GLMM).

Обобщенные линейные смешанные модели

Обобщенные линейные смешанные модели обобщают линейные модели таким образом, что:

- Целевая переменная линейно связана с факторами и ковариатами посредством заданной функции связи.
- Распределение целевой переменной может отличаться от нормального.
- Наблюдения могут быть скоррелированы.

Обобщенные линейные смешанные модели включают широкий набор моделей, начиная от простой линейной регрессии и кончая сложными многоуровневыми моделями для не нормально распределенных данных с повторными измерениями.

Примеры. Районный школьный отдел может использовать обобщенную линейную смешанную модель, чтобы оценить влияние экспериментального метода преподавания на успеваемость по математике. Учащиеся из одного класса могут обнаруживать корреляцию по успеваемости, поскольку занимаются у одного преподавателя, показатели классов одной школы также могут коррелировать, поэтому мы можем включить в модель случайные эффекты на уровне школы и класса, отражающие различные источники вариации. Дополнительную информацию смотрите в разделе .

В медицинских исследованиях с помощью обобщенной линейной смешанной модели можно установить, например, может ли новый противосудорожный препарат снизить частоту эпилептических припадков. Повторные измерения на одном и том же пациенте обычно обнаруживают высокую положительную корреляцию, поэтому в данном случае подходит смешанная модель со случайными эффектами. Поле назначения (число припадков) принимает целые положительные значения, поэтому здесь можно использовать обобщенную линейную смешанную модель с распределением Пуассона и логарифмической зависимостью. Дополнительную информацию смотрите в разделе .

Администраторы служб провайдеров кабельного телевидения, телефонии и услуг интернета могут использовать обобщенную линейную смешанную модель для получения дополнительной информации о потенциальных заказчиках. Поскольку возможные ответы имеют номинальные уровни измерения, аналитики компании используют обобщенную смешанную логит-модель со случайным свободным членом для отражения корреляции между ответами на вопросы о пользовании услугами разных типов (телевидение, телефон, интернет) для конкретного респондента, заполнившего опросный лист. Дополнительную информацию смотрите в разделе .

На вкладке Структура данных можно задать структурные взаимосвязи между записями в наборе данных, если наблюдения скоррелированы. Если записи в наборе данных представляют независимые наблюдения, ничего задавать на этой вкладке не требуется.

Субъекты. Сочетание значений заданных категориальных полей должна уникальным образом определять субъекты в наборе данных. К примеру, единственного поля *ID пациента* должно быть достаточно для определения субъектов в одной больнице, но может потребоваться сочетание *ID больницы* и *ID пациента* в случае, если ID пациентов не являются уникальными в рамках всех больниц. При выборе режима повторных измерений для каждого субъекта фиксируются несколько наблюдений, поэтому каждый субъект может присутствовать в нескольких записях одного набора данных.

Субъект - это единица наблюдения, которую можно считать независимой от других субъектов. Например, результаты измерения кровяного давления у пациента в медицинском исследовании можно считать независимыми от таких же показателей у других пациентов. Определить субъекты бывает особенно важно, когда для субъектов выполняются повторные измерения, и между этими наблюдениями нужно смоделировать взаимосвязь. Например, можно ожидать, что показания кровяного давления для одного пациента при последовательных посещениях врача будут скоррелированы между собой.

Все поля, указанные как Субъекты на вкладке Структура данных, используются для определения субъектов для ковариационной структуры остатков и содержат список возможных полей для определения субъектов для ковариационных структур случайных эффектов в блоке случайных эффектов.

ОЛМ-повторные измерения. Заданные здесь поля служат для идентификации повторных наблюдений. Например, одна переменная *Неделя* может обозначать 10 недель наблюдений в медицинском исследовании, а *Месяц* и *День* могут использоваться совместно для идентификации ежедневных наблюдений в течение года.

Задать группы ковариаций с помощью. Заданные здесь категорийные поля определяют независимые наборы параметров ковариации повторяющихся эффектов, по одному для каждой категории, определяемой перекрестной классификацией полей группировки. У всех субъектов - один тип ковариации, у субъектов внутри одной ковариационной группировки значения параметров совпадают.

Тип ковариационной матрицы для повторных измерений. Задает ковариационную структуру для остатков. Доступные структуры:

- Авторегрессия первого порядка AR(1)
- Авторегрессивное скользящее среднее ARMA(1,1)
- Составная симметрия (CS)
- Диагональная
- Масштабированная единичная
- Теплицева
- Неструктурированная
- Компоненты дисперсии (VC)

Назначение: Эти параметры определяют целевую переменную, ее распределение и взаимосвязи с предикторами через функцию связи.

Цель. Целевая переменная обязательна для ввода. У нее может быть любая шкала измерения, причем шкала измерения целевой переменной определяет подходящие для нее распределения и функции связи.

- **Использовать число испытаний в качестве знаменателя.** Если целевым откликом является количество событий, встречающихся в наборе испытаний, поле назначения содержит количество событий, причем можно выбрать дополнительную переменную, содержащую количество попыток. Например, при испытаниях нового пестицида выборки муравьев подвергаются воздействию различных концентраций этого пестицида, и для каждой выборки фиксируется количество погибших муравьев, а также ее объем. В этом случае поле, содержащее число погибших муравьев, указывается как поле назначения (событий), а поле, содержащее число муравьев в каждой выборке, указывается как поле испытаний. Если число муравьев во всех выборках одинаково, число испытаний можно задать в виде постоянной величины.

  Число испытаний не должно превышать числа событий для каждой записи. События должны представлять собой неотрицательные, а испытания - положительные целые числа.
- **Задать опорную категорию.** Для категориальной целевой переменной можно выбрать опорную категорию. Это может повлиять на некоторые характеристики вывода, например, оценки параметров, но не может изменить подгонку модели. Например, если целевая переменная принимает значения 0, 1 и 2, то по умолчанию процедура сделает последнюю категорию (с наивысшим значением, то есть 2), опорной категорией. В этой ситуации оценки параметров следует интерпретировать как относящиеся к вероятности категорий 0 и 1 *по отношению* к вероятности категории 2. Если вы задали пользовательскую категорию и в целевой переменной определены метки, опорную категорию можно задать, выбрав значение в списке. Это может быть удобно, если в процессе задания модели вы забыли кодировку какого-либо поля.

Распределение целевой переменной и функция связи для линейной модели. При заданных значениях предикторов, согласно данной модели, распределение значений целевой переменной, должно соответствовать заданному виду, а сами значения должны быть связаны с предикторами заданной линейной функцией связи. Воспользуйтесь ярлыками для нескольких общих моделей или выберите параметр **Пользовательские**, если хотите использовать определенное сочетание распределения и функции связи, которого нет в списке ярлыков.

- **Линейная модель.** Задает нормальное распределение с тождественной функцией связи и применяется, когда целевую переменную можно предсказать, используя модель линейной регрессии или дисперсионного анализа.
- **Гамма-регрессия.** Задает гамма-распределение с логарифмической функцией связи и применяется, когда распределение целевой переменной содержит только положительные значения и скошено в направлении больших значений.
- **Логлинейный.** Задает распределение Пуассона с логарифмической функцией связи и применяется, когда целевая переменная представляет частоты событий, произошедших в фиксированный период времени.
- **Отрицательная биномиальная регрессия.** Задает отрицательное биномиальное распределение с логарифмической функцией связи, которую следует использовать, когда целевая переменная и знаменатель представляет собой количество попыток, требуемое для *k* успешных результатов.
- **Полиномиальная логистическая регрессия.** Задает полиномиальное распределение, которое нужно использовать, когда целевая переменная представляет собой отклик с несколькими категориями. В ней используется кумулятивная logit-модель (порядковые исходы) или обобщенную logit-модель (ответы с несколькими номинальными категориями).
- **Бинарная логистическая регрессия.** Задает биномиальное распределение с функцией связи логит и применяется, когда целевая переменная является бинарным откликом, предсказываемым логистической регрессионной моделью.
- **Бинарное пробит-распределение.** Задает биномиальное распределение с функцией связи пробит и применяется, когда целевая переменная является бинарным откликом, в основе которого лежит нормальное распределение.
- **Выживание для интервально цензурированных данных.** Задает биномиальное распределение с функцией связи Дополняющая лог-лог и используется в анализе выживания, когда некоторые наблюдения не имеют терминального события.

Распределение

Выбор в этой группе задает распределение целевой переменной. Возможность задавать распределение, отличное от нормального, и нетождественную функцию связи - существенное преимущество обобщенной линейной смешанной модели перед линейной смешанной моделью. Из множества возможных комбинаций "распределение - функция связи" некоторые могут быть пригодны для любого набора данных, поэтому выбор в таких случаях может основываться на априорных теоретических соображениях или на оценке согласия данных с моделью.

- **Биномиальное.** Это распределение подходит только для целевой переменной, представляющей бинарный отклик или число событий.
- **Гамма.** Этот вид распределения подходит для целевой переменной с положительными значениями, распределение которой скошено в сторону больших значений. Если значение равно нулю, отрицательно или отсутствует, соответствующее наблюдение не используется в анализе.
- **Обратное нормальное распределение.** Этот вид распределения подходит для целевой переменной с положительными значениями, распределение которой скошено в сторону больших значений. Если значение равно нулю, отрицательно или отсутствует, соответствующее наблюдение не используется в анализе.
- **Полиномиальное.** Это распределение подходит для целевой переменной, представляющей отклик с несколькими категориями. Вид модели зависит от шкалы измерения целевой переменной.

  Номинальной целевой переменной соответствует номинальная полиномиальная модель, в которой для каждой категории (кроме опорной) оценивается отдельный набор параметров. Оценки параметров для данного предиктора показывают взаимосвязь между этим предиктором и вероятностью каждой категории целевой переменной по отношению к опорной категории.

  Порядковой целевой переменной соответствует порядковая полиномиальная модель, в которой традиционный свободный член заменен набором **пороговых** параметров, относящихся к накопленной вероятности категорий целевой переменной.

- **Отрицательное биномиальное распределение.** Отрицательная биномиальная регрессия использует отрицательное биномиальное распределение с логарифмической функцией связи, которую следует использовать, когда целевое поле представляет собой количество событий с большой дисперсией.
- **Нормальное.** Подходит для непрерывной целевой переменной с симметричным колоколообразным распределением вокруг центрального (среднего) значения.
- **Пуассона.** Это распределение можно представить как число реализаций интересующего нас события за фиксированный период времени, оно применимо к переменным с неотрицательными целочисленными значениями. Если значение данных не целое, меньше 0 или пропущено, соответствующее ему наблюдение в анализе не используется.

Функции связи

Связывающая функция служит для преобразования назначения для расчета модели. Доступны следующие функции:

- **Тождество.** $f(x)=x$. Целевая переменная не преобразуется. Эта функция может использоваться для любых распределений, кроме полиномиального.
- **Дополнительный логарифм-логарифм.** $f(x)=\log(-\log(1-x))$. Подходит только для биномиального или полиномиального распределения.
- **Коши.** $f(x) = \tan(\pi\ (x - 0{,}5))$. Подходит только для биномиального или полиномиального распределения.
- **Логарифмическая.** $f(x)=\log(x)$. Эта функция может использоваться для любых распределений, кроме полиномиального.
- **Дополняющее лог.** $f(x)=\log(1-x)$. Подходит только для биномиального распределения.
- **Логит.** $f(x)=\log(x / (1-x))$. Подходит только для биномиального или полиномиального распределения.
- **Отрицательный Log-log.** $f(x)=-\log(-\log(x))$. Подходит только для биномиального или полиномиального распределения.
- **Пробит.** $f(x)=\Phi^{-1}(x)$, где Φ^{-1} - функция, обратная к интегральной функции стандартного нормального распределения. Подходит только для биномиального или полиномиального распределения.
- **Степенная.** $f(x)=x^{\alpha}$, если $\alpha \neq 0$. $f(x)=\log(x)$, если $\alpha=0$. α - обязательная числовая спецификация, которая должна быть действительным числом. Эта функция может использоваться для любых распределений, кроме полиномиального.

Фиксированные эффекты: Факторы фиксированных эффектов обычно представляют как поля, чьи интересующие нас значения представлены в наборе данных и могут использоваться для количественной оценки. По умолчанию поля с заранее заданной входной ролью, которые не указаны в других местах диалогового окна, вводятся в той части модели, которая связана с фиксированными эффектами. Категорийные поля (флаговые, номинальные и порядковые) используются как факторы в модели, а непрерывные поля используются как ковариаты.

Введите эффекты в модель, выбирая одно или несколько полей в исходном списке и перетаскивая их в список эффектов. Тип созданного эффекта зависит от того, после какого гиперобъекта был прекращен выбор.

- **Главные.** Отброшенные поля выводятся как отдельные главные эффекты внизу списка эффектов.
- **2-факторный.** Все возможные пары отброшенных полей выводятся как 2-факторные взаимодействия в нижней части списка эффектов.
- **3-факторный.** Все возможные триплеты отброшенных полей выводятся как 3-факторные взаимодействия в нижней части списка эффектов.
- ***.** Сочетание всех отброшенных полей выводится как одно взаимодействие в нижней части списка эффектов.

Кнопки справа от конструктора эффектов позволяют выполнять разнообразные действия.

Таблица 10. Описание кнопок конструктора эффектов.

Значок	Описание
	Удалить члены из модели фиксированных эффектов, выбрав члены для удаления и нажав кнопку Удалить.
	Изменить порядок членов в модели фиксированных эффектов, выбрав члены, которые нужно переставить, и нажав кнопку со стрелкой вверх или вниз.
	Добавить в модель вложенные члены в диалоговом окне “Добавить пользовательский член”, нажав кнопку Добавить пользовательский член.

Включить константу. Обычно в модель включают свободный член. Если вы предполагаете, что данные проходят через начало координат, свободный член можно исключить.

Добавить пользовательский член: С помощью этой процедуры можно создать вложенные члены для модели. Вложенные члены полезны при моделировании эффекта или ковариата, значения которого не взаимодействуют с уровнями другого фактора. Например, сети гастрономических магазинов требуются сведения о характере расходов покупателей в нескольких таких магазинах. Поскольку каждый покупатель посещает только один из магазинов, можно сказать, что эффект *Покупатель* **вложен в** эффект *Положение магазина*.

Дополнительно можно подключить сюда эффекты взаимодействия, например, полиномиальные члены, описывающие тот же самый ковариат, или добавить во вложенный член несколько уровней вложения.

Ограничения. Вложенные члены имеют следующие ограничения:

- Все факторы во взаимодействии должны быть уникальными. То есть, если *A* - фактор, то нельзя задать *A*A*.
- Все факторы во вложенном эффекте должны быть уникальными. То есть, если *A* - фактор, то нельзя задать *A(A)*.
- Никакие эффекты не могут быть вложенными в эффект, являющийся ковариатом. То есть, если *A* - фактор, а *X* - ковариат, то нельзя задать *A(X)*.

Построение вложенного члена

1. Выберите фактор или ковариат, вложенный в другой фактор, и нажмите кнопку со стрелкой.
2. Нажмите **(B)**.
3. Выберите фактор, в который вложен предыдущий фактор или ковариат, и затем нажмите кнопку со стрелкой.
4. Нажмите кнопку **Добавить член**.

Можно (необязательно) подключить эффекты взаимодействия, например, или добавить во вложенный член несколько уровней вложения.

Произвольные эффекты: Факторы произвольных эффектов - это поля, значения которых в файле данных можно рассматривать как случайную выборку из более обширной совокупности значений. Они помогают объяснить избыточную вариацию целевой переменной. По умолчанию, если на вкладке Структура данных выбрано несколько объектов, для каждого объекта за пределами внутреннего объекта будет создан блок произвольных эффектов. Например, если на вкладке Структура данных выбраны объекты Школа, Класс и Учащийся, автоматически будут созданы следующие блоки произвольных эффектов:

- Произвольный эффект 1: объект - школа (без эффектов, только свободный член)
- Произвольный эффект 2: объект - школа * класс (без эффектов, только свободный член)

Возможны следующие варианты работы с блоками произвольных эффектов:

1. Чтобы добавить новый блок, нажмите кнопку **Добавить блок...** Откроется диалоговое окно “Блок произвольных эффектов”.
2. Чтобы изменить существующий блок, выберите блок, который нужно изменить, и нажмите кнопку **Отредактировать блок...** Откроется диалоговое окно “Блок произвольных эффектов”.
3. Чтобы удалить один или несколько блоков, выберите блоки, которые нужно удалить, и нажмите кнопку Удалить.

Блок произвольных эффектов: Введите эффекты в модель, выбирая одно или несколько полей в исходном списке и перетаскивая их в список эффектов. Тип созданного эффекта зависит от того, после какого гиперобъекта был прекращен выбор. Категорийные поля (флаговые, номинальные и порядковые) используются как факторы в модели, а непрерывные поля используются как ковариаты.

- **Главные.** Отброшенные поля выводятся как отдельные главные эффекты внизу списка эффектов.
- **2-факторный.** Все возможные пары отброшенных полей выводятся как 2-факторные взаимодействия в нижней части списка эффектов.
- **3-факторный.** Все возможные триплеты отброшенных полей выводятся как 3-факторные взаимодействия в нижней части списка эффектов.
- ***.** Сочетание всех отброшенных полей выводится как одно взаимодействие в нижней части списка эффектов.

Кнопки справа от конструктора эффектов позволяют выполнять разнообразные действия.

Таблица 11. Описание кнопок конструктора эффектов.

Значок	Описание
	Удалить члены из модели, выбрав члены для удаления и нажав кнопку Удалить.
	Изменить порядок членов в модели, выбрав члены, которые нужно переставить, и нажав кнопку со стрелкой вверх или вниз.
	Добавить в модель вложенные члены в диалоговом окне “Добавить пользовательский член” на стр. 195, нажав кнопку Добавить пользовательский член.

Включить константу. Свободный член не включен в модель случайных эффектов по умолчанию. Если вы предполагаете, что данные проходят через начало координат, свободный член можно исключить.

Задать группы ковариаций с помощью. Заданные здесь категорийные поля определяют независимые наборы параметров ковариации случайных эффектов, по одному для каждой категории, определяемой перекрестной классификацией группирующих полей. Для каждого блока случайных эффектов можно задать отдельный набор полей группировки. Все объекты имеют один тип ковариации, у объектов внутри одной ковариационной группировки значения параметров совпадают.

Комбинация объектов. Позволяет задавать объекты случайных эффектов по заранее заданным сочетаниям объектов на вкладке Структура данных. Например, если *Школа*, *Класс* и *Студент* определены на вкладке Структура данных как объекты в указанном выше порядке, в выпадающем списке Сочетания объектов будут опции **Нет**, **Школа**, **Школа * Класс** и **Школа * Класс * Учащийся**.

Тип ковариаций случайных эффектов. Задает ковариационную структуру для остатков. Доступные структуры:

- Авторегрессия первого порядка AR(1)

- Авторегрессивное скользящее среднее ARMA(1,1)
- Составная симметрия (CS)
- Диагональная
- Масштабированная единичная
- Теплицева
- Неструктурированная
- Компоненты дисперсии (VC)

Вес и смещение: Веса в анализе. Параметр шкалы является оценочным параметром модели, связанным с дисперсией ответов. Веса в анализе - это "известные" значения, которые могут варьировать от наблюдения к наблюдению. Если задано поле веса в анализе, параметр масштаба, связанный с дисперсией ответа, будет делиться на значения веса в анализе для каждого наблюдения. Записи со значениями веса в анализе, меньшими или равными нулю (или пропущенными значениями) не используются в анализе.

Смещение. Критерий смещения является "структурным" предиктором. Его коэффициент не оценивается моделью, но предполагается, что у него значение 1; поэтому значения смещения просто добавляются к значениям линейного предиктора назначения. Это особенно полезно в регрессионных моделях Пуассона, где у каждого наблюдения могут быть различные уровни влияния на исследуемое событие.

Например, при моделировании уровня автомобильных аварий для конкретных водителей есть важное различие между между водителем, побывавшим в одной аварии за три года вождения, и водителем, попавшим в аварию один раз за 25 лет. Количество аварий может быть моделировано как отклик Пуассона или отрицательный биномиальный отклик с логарифмической связью, если натуральный логарифм опыта (стажа) водителя включен как критерий смещения.

Для других типов сочетаний распределения и связи потребовались бы другие преобразования переменной смещения.

Общие опции построения: Эти параметры задают некоторые более сложные критерии, используемые при создании модели.

Порядок сортировки. Эти элементы управления определяют порядок категорий для целевой переменной и факторов (категориальных входных данных) с целью выявления "последней" категории. Параметр порядка сортировки значений целевой переменной игнорируется, если целевая переменная не является категориальной или если в параметрах “Назначение” на стр. 192 задана пользовательская опорная категория.

Правила остановки. Вы можете задать максимальное число итераций, выполняемых алгоритмом. Этот алгоритм использует двойной итеративный процесс, который состоит из внутреннего и внешнего циклов. Заданное для максимального числа итераций значение применяется к обоим циклам. Задайте неотрицательное целое число. Значение по умолчанию - 100.

Исследование после оценивания. Эти параметры определяют, как вычисляются для просмотра некоторые из выходных данных этой модели.

- **Доверительный уровень.** Это доверительный уровень, используемый при вычислении интервальных оценок коэффициентов модели. Задайте значение больше 0 и меньше 100. Значение по умолчанию - 95.
- **Степени свободы.** Этот параметр задает, как вычисляется число степеней свободы для тестов значимости. Выберите **Фиксировано для всех тестов (Метод остатков)**, если объем выборки достаточно велик, если данные сбалансированы или если в модели используется более простой тип ковариации, например, масштабированный единичный или диагональный. Это задано по умолчанию. Выберите **Варьируется по тестам (Аппроксимация Саттертуэйта)**, если объем выборки мал, если данные не сбалансированы или если используется сложный тип ковариации, например, неструктурированный.
- **Тесты фиксированных эффектов и коэффициентов.** Это метод вычисления матрицы ковариации оценок параметров. Выберите робастную оценку, если вам кажется, что допущения модели нарушены.

Оценка: Этот алгоритм построения модели использует двойной итеративный процесс, который состоит из внутреннего и внешнего циклов. Следующие параметры применяются к внутреннему циклу.

Сходимость параметров.
Предполагается, что сходимость достигнута, если максимальное абсолютное или максимальное относительное изменение оценки параметров меньше, чем заданное значение (которое должно быть неотрицательным). Этот критерий не используется, если заданное значение равно 0.

Сходимость логарифмического правдоподобия.
Предполагается, что сходимость достигнута, если абсолютное или относительное изменение функции логарифмического правдоподобия меньше, чем заданное значение (которое должно быть неотрицательным). Этот критерий не используется, если заданное значение равно 0.

Сходимость гессиана.
Для спецификации **Абсолютная** предполагается сходимость, если статистика на основе сходимости гессиана стала меньше заданного значения. Для спецификации **Относительная** предполагается сходимость, если указанная статистика стала меньше произведения заданного значения на абсолютное значение логарифмического правдоподобия. Этот критерий не используется, если заданное значение равно 0.

Максимальное число шагов оценки Фишера.
Задайте неотрицательное целое число. Значение 0 задает метод Ньютона-Рафсона. Значение больше 0 задает использование алгоритма оценки Фишера до итерации с номером *n*, где *n* - заданное целое число, и затем - метода Ньютона-Рафсона.

Допуск для вырожденности.
Это значение используется как допуск для проверки на вырожденность. Задайте положительное значение.

Примечание: По умолчанию используется сходимость параметров, где задано максимальное **Абсолютное** изменение с допуском 1E-6. Эти параметры могут дать результаты, отличающиеся от результатов, которые получены в версиях до версии 22. Чтобы воспроизвести результаты из версий до версии 22, используйте **Относительное** изменение в качестве критерия сходимости параметров, сохранив значение допуска по умолчанию - 1E-6.

Общие: Имя модели. Имя модели можно сгенерировать автоматически на основе целевых полей или задать самостоятельно. Автоматически генерируемое имя является именем целевого поля. Если полей назначения несколько, именем модели служит упорядоченный список имен полей через знак конъюнкции. Например, если поля назначения - *поле1 поле2 поле3*, то имя модели - *поле1 & поле2 & поле3*.

Сделать доступным для скоринга. При оценке модели нужно сгенерировать элементы, выбранные в этой группе. Предсказанное значение (для всех полей назначения) и достоверность (для категориальных полей назначения) вычисляются при оценке всегда. Вычисленный показатель доверия может быть основан на вероятности предсказанного значения (наивысшая предсказанная вероятность) или на разнице между наивысшей предсказанной вероятностью и вторым по величине значением предсказанной вероятности.

- **Предсказанную вероятность для категориальных целевых полей.** Будут вычислены предсказанные вероятности для категориальных полей назначения. Для каждой категории создается поле.
- **Оценки склонности для флаговых полей назначения .** Для моделей с флаговым полем назначения (которое возвращает предсказание Да или Нет) можно затребовать оценку склонности, которая обозначает правдоподобие правильного выходного значения, заданного для поля назначения. Модель находит простые оценки склонности; кроме того, если действуют разбиения, модель находит скорректированные оценки склонности на основе контрольного раздела.

Оцененные средние: На этой вкладке можно выводить оценки маргинальных средних для уровней и взаимодействий факторов. Оценки маргинальных средних недоступны для полиномиальных моделей.

Терминология. Здесь показаны члены модели в фиксированных эффектах, которые целиком состоят из категориальных полей. Проверьте каждый член, для которого хотите получить от модели оценки маргинальных средних.

- **Тип контрастов.** Задает тип контрастов, используемый для уровней поля контрастов. Если выбрано значение **Нет**, контрасты создаваться не будут. При значении **Парные** производятся попарные сравнения для всех сочетаний уровней заданных факторов. Это единственный доступный контраст для всех взаимодействий факторов. При значении **Контрасты отклонений** каждый уровень фактора сравнивается с общим средним. При значении **Простые** контрасты каждый уровень фактора, кроме последнего, сравнивается с последним уровнем. "Последний" уровень определяется порядком сортировки для факторов, заданных в группе Параметры построения. Обратите внимание на то, что все эти типы контрастов не являются ортогональными.
- **Поле контрастов.** Задает фактор, уровни которого сравниваются с помощью выбранного типа контрастов. Если в качестве типа контрастов выбрано **Нет**, поле контрастов нельзя выбрать или это не требуется.

Количественные поля. Указанные в списке непрерывные поля извлекаются из членов в фиксированных эффектах, использующих непрерывные поля. При вычислении оценок маргинальных средних ковариаты фиксируются в соответствии с заданными значениями. Выберите среднее или задайте пользовательское значение.

Вывести оцененные средние в терминах. Задает, вычислять ли оценки маргинальных средних на основе исходного масштаба целевой переменной или на основе преобразования функцией связи. Если задано **Исходный масштаб целевой переменной**, оценки маргинальных средних будут вычисляться для целевой переменной. Учтите, что если целевая переменная задана при помощи опции события/испытания, это дает оценки маргинальных средних для соотношения события/испытания, а не для числа событий. Если задано **Преобразование функцией связи**, оценки маргинальных средних будут вычисляться для линейного предиктора.

Скорректировать на множественные сравнения, используя. При выполнении проверки гипотезы с несколькими контрастами общий уровень значимости можно отрегулировать, исходя из уровней значимости для используемых контрастов. Это позволяет выбирать метод корректировки.

- **Наименьшая значимая разность.** Этот метод не управляет полной вероятностью отклонения гипотез, некоторые линейные контрасты которых отличаются от значений пустой гипотезы.
- *Последовательный Бонферрони*. Это процедура Бонферрони последовательного отклонения, которая является значительно менее консервативной в плане отклонения отдельных гипотез, но сохраняет тот же общий уровень значимости.
- *Последовательный Шидак*. Это процедура Шидака последовательного отклонения, которая является значительно менее консервативной в плане отклонения отдельных гипотез, но сохраняющей тот же общий уровень значимости.

Метод наименьшей значимой разности менее консервативен, чем последовательный метод Шидака, который, в свою очередь, менее консервативен, чем последовательный метод Бонферрони, соответственно, метод наименьшей значимой разности будет отбрасывать, как минимум, столько же индивидуальных гипотез, как и последовательный метод Шидака, а тот, в свою очередь, как минимум, столько же индивидуальных гипотез, как последовательный метод Бонферрони.

Представление модели: По умолчанию выводится представление Сводка для модели. Чтобы посмотреть другое представление модели, выберите его из миниизображений представления.

Сводка для модели: Это представление - снимок, мгновенная визуальная сводка по модели и ее подгонке.

Таблица. Эта таблица определяет целевую переменную, распределение вероятностей и функцию связи, заданные в окне Параметры целевой переменной. Если целевая переменная определяется событиями и испытаниями, ячейка будет разделена, чтобы показать отдельно поле событий и поле испытаний или же

фиксированное количество испытаний. Кроме того, выводится скорректированный информационный критерий Акаике (corrected Akaike information criterion, AICC) конечной выборки и информационный критерий Байеса (Bayesian information criterion, BIC).

- *Акаике скорректированный.* Критерий для выбора и сравнения смешанных моделей на основе отрицательного удвоенного (ограниченного) логарифма правдоподобия. Меньшие значения указывают на лучшую модель. Критерий AICC "корректирует" информационный критерий Акаике (AIC) для малых размеров выборок. При увеличении размеров выборок критерий AICC сходится к критерию AIC.
- *Байесовский.* Критерий для выбора и сравнения моделей на основе отрицательного удвоенного логарифма правдоподобия. Меньшие значения указывают на лучшую модель. Критерий BIC также "штрафует" чрезмерно параметризованные модели, но строже, чем AIC.

Диаграмма. Если целевая переменная является категориальной, диаграмма показывает точность окончательной модели в виде процентной доли правильных классификаций.

Структура данных: Это представление содержит сводку заданной вами структуры данных и помогает проверить, что субъекты и повторные измерения заданы правильно. Информация наблюдений для первого субъекта выводится для всех полей субъектов и полей повторных измерений, а также для целевой переменной. Кроме того, выводится количество уровней для каждого поля субъекта и поля повторных измерений.

Предсказанные против наблюденных: Для непрерывных целевых переменных, включая переменные, заданные как события/испытания, выводится диаграмма рассеяния с интервалами для предсказанных значений по вертикальной оси против наблюденных значений по горизонтальной оси. В идеале точки должны лежать на прямой, проведенной под углом 45 градусов. Такое представление позволяет определить, есть ли записи, которые плохо предсказываются моделью.

Классификация: Для категориальных целевых переменных в этой таблице выводится перекрестная классификация наблюденных и предсказанных значений целевой переменной в тепловой карте, а также общий процент правильных значений.

Стили таблиц. Существует несколько различных стилей вывода, которые можно выбрать в выпадающем списке **Стиль**.

- **Процент по строке.** Здесь выводятся проценты по строкам (количества в ячейках, выраженные в процентах от итогов по строке) в ячейках. Это задано по умолчанию.
- **Количества в ячейках,** Здесь выводятся количества в ячейках. Тени для тепловой карты по-прежнему основаны на значениях процентов по строкам.
- **Тепловая карта.** Значения для ячеек не выводятся, используется только затенение.
- **Сжатый.** Без вывода заголовков строк или столбцов или значений в ячейках. Может быть полезным, если у целевой переменной много категорий.

Пропущенные. Если в каких-либо записях для целевой переменной есть отсутствующие значения, они выводятся в строке **(Отсутствующие)** под всеми остальными действительными строками. Записи с отсутствующими значениями не участвуют в вычислении общего процента правильных.

Несколько целевых переменных. При наличии нескольких категориальных целевых переменных каждая из них выводится в отдельной таблице, причем отдельные целевые переменные можно выбрать в в выпадающем списке **Целевая переменная**.

Большие таблицы. Если у выводимой целевой переменной более 100 категорий, таблица не выводится.

Фиксированные эффекты: Этот вид показывает величину каждого фиксированного эффекта в модели.

Стили. Имеются различные стили вывода, которые можно выбрать в выпадающем списке **Стиль**.

- **Диаграмма.** Это диаграмма, на которой эффекты сортируются сверху вниз в порядке, заданном параметрами в группе Фиксированные эффекты. Соединяющие линии на диаграмме являются

взвешенными на основе значимости эффектов, с большей толщиной линии, соответствующей более значимым эффектам (меньшим *p*-значениям). Это задано по умолчанию.

- **Таблица.** Это таблица дисперсионного анализа для общих и индивидуальных эффектов модели. Отдельные эффекты сортируются сверху вниз в порядке, заданном параметрами в группе Фиксированные эффекты.

Значимость. Есть ползунок значимости, который управляет тем, какие эффекты выводятся в представлении. Эффекты со значениями значимости, превосходящими значение ползунка, скрыты. Это не изменяет модели, а просто позволяет сосредоточить внимание на наиболее важных эффектах. По умолчанию это значение равно 1,00, так что никакие эффекты не отфильтровываются на основе значимости.

Фиксированные коэффициенты: Этот вид показывает значение каждого фиксированного коэффициента в модели. Обратите внимание на то, что факторы (категориальные предикторы) имеют индикаторную кодировку в модели, так что **эффекты**, содержащие факторы, обычно будут иметь несколько связанных **коэффициентов**, по одному для каждой категории, исключая категорию, соответствующую избыточному коэффициенту.

Стили. Имеются различные стили вывода, которые можно выбрать в выпадающем списке **Стиль**.

- **Диаграмма.** Это диаграмма, в которой сначала выводится свободный член, а затем эффекты, отсортированные сверху вниз в порядке, указанном параметрами в группе Фиксированные эффекты. Внутри эффектов, содержащих факторы, коэффициенты сортируются в порядке возрастания значений данных. Соединяющие линии на диаграмме являются раскрашенными и взвешенными на основе значимости коэффициентов, с большей толщиной линии, соответствующей более значимым коэффициентам (меньшим *p*-значениям). Это задано по умолчанию.
- **Таблица.** В этой таблице выводятся значения, результаты тестов на значимость и доверительные интервалы для индивидуальных коэффициентов модели. После свободного члена эффекты сортируются сверху вниз в порядке, заданном параметрами в группе Фиксированные эффекты. Внутри эффектов, содержащих факторы, коэффициенты сортируются в порядке возрастания значений данных.

Полиномиальное. Если имеет место полиномиальное распределение, выбрать категорию целевой переменной для вывода можно в выпадающем списке Полиномиальное. Порядок сортировки значений в списке определяется параметрами в окне Параметры построения.

Экспоненциальное. Выводит оценки коэффициентов экспонент и доверительные интервалы для некоторых типов моделей, включая бинарную логистическую регрессию (биномиальное распределение с функцией связи логит), номинальную логистическую регрессию (полиномиальное распределение с функцией связи логит), отрицательную биномиальную регрессию (отрицательное биномиальное распределение с логарифмической связью) и логлинейную модель (распределение Пуассона с логарифмической связью).

Значимость. Есть ползунок значимости, который управляет тем, какие коэффициенты выводятся в представлении. Коэффициенты со значениями значимости, превосходящими значение ползунка, скрыты. Это не изменяет модели, а просто позволяет сосредоточить внимание на наиболее важных коэффициентах. По умолчанию это значение равно 1,00, так что никакие коэффициенты не отфильтровываются на основе значимости.

Ковариации произвольных эффектов: В этом представлении выводится матрица ковариации произвольных эффектов (**G**).

Стили. Имеются различные стили вывода, которые можно выбрать в выпадающем списке **Стиль**.

- **Значения ковариаций.** Это тепловая карта ковариационной матрицы, на которой эффекты сортируются сверху вниз в порядке, заданном параметрами в окне Фиксированные эффекты. Цвета на коррелограмме соответствуют значениям ячеек, как показано в ключе. Это задано по умолчанию.
- **Коррелограмма.** Это тепловая карта матрицы ковариации.
- **Сжатые.** Это тепловая карта матрицы ковариации без заголовков строк и столбцов.

Блоки. Если есть несколько блоков произвольных эффектов, в выпадающем списке Блок можно выбрать блок для вывода.

Группы. Если у блока произвольных эффектов есть спецификация группы, в выпадающем списке Группа можно выбрать уровень группы для вывода.

Полиномиальное. Если имеет место полиномиальное распределение, в выпадающем списке Полиномиальное можно выбрать категорию целевой переменной для вывода. Порядок сортировки значений в списке определяется параметрами в группе Опции построения.

Параметры ковариации: Это представление выводит оценки параметров ковариации и связанные с ними статистические показатели для остатков и случайных эффектов. Это расширенные, но фундаментальные результаты, содержащие информацию о пригодности ковариационной структуры.

Итоговая таблица. Это краткий справочник по числу параметров в матрицах ковариации остатков (**R**) и случайных эффектов (**G**), рангу (числу столбцов) матриц плана фиксированных эффектов (**X**) и случайных эффектов (**Z**), а также числу объектов, которые заданы полями, определяющими структуру данных.

Таблица параметров ковариации. Для каждого параметра ковариации при выбранном эффекте отображаются его оценка, среднеквадратичная ошибка и доверительный интервал. Число выводимых параметров зависит от ковариационной структуры для этого эффекта, а для блоков случайных эффектов - от числа эффектов в этом блоке. Если видно, что недиагональные параметры незначимы, лучше использовать более простую ковариационную структуру.

Эффекты. Если есть блоки случайных эффектов, в выпадающем списке Эффект можно выбрать эффект остатков или случайный эффект для вывода. Эффект остатков всегда доступен.

Группы. Если у блока эффектов остатков или случайных эффектов есть спецификация группы, уровень группы для вывода можно выбрать в выпадающем списке Группа.

Полиномиальное. Если имеет место полиномиальное распределение, выбрать категорию целевой переменной для вывода можно в выпадающем списке Полиномиальное. Порядок сортировки значений в списке определяется параметрами в группе Параметры построения.

Оцененные средние: Значимые эффекты: Это диаграммы, показанные для десяти "наиболее значимых" фиксированных эффектов всех факторов, начиная с трехфакторных взаимодействий, далее для двухфакторных взаимодействий и, наконец, для главных эффектов. На диаграммах по вертикальной оси отложены сделанные по модели оценки значений целевой переменной для каждого значения главного эффекта (или первого эффекта, указанного во взаимодействии) на горизонтальной оси. Отдельная линия соответствует каждому значению второго эффекта, указанного для взаимодействия, а для каждого значения третьего указанного эффекта в трехфакторном взаимодействии выводится отдельная диаграмма при сохранении значений всех остальных предикторов неизменными. Это дает полезную визуализацию того, какое влияние коэффициент каждого предиктора оказывает на целевую переменную. Обратите внимание, на то, что при отсутствии значимых предикторов оценки средних не вычисляются.

Показатель доверия. Показывает верхнюю и нижнюю границы доверительного интервала для маргинальных средних, используя уровень значимости, заданный в группе Параметры построения.

Оцененные средние: Пользовательские эффекты: Это таблицы и диаграммы для затребованных пользователями фиксированных эффектов всех факторов.

Стили. Имеются различные стили вывода, которые можно выбрать в выпадающем списке **Стиль**.

- **Диаграмма.** В этом стиле выводится линейная диаграмма, где по вертикальной оси отложены сделанные по модели оценки значений целевой переменной для каждого значения главного эффекта (или первого эффекта, указанного во взаимодействии) на горизонтальной оси. Отдельная линия соответствует каждому

значению второго эффекта, указанного для взаимодействия, а для каждого значения третьего указанного эффекта в трехфакторном взаимодействии выводится отдельная диаграмма, при сохранении значений всех остальных предикторов неизменными.

Если затребованы контрасты, выводится другая диаграмма, где сравниваются уровни поля контрастов, а для взаимодействий выводится диаграмма, отражающая каждое сочетание уровней эффектов, отличное от поля контрастов. Для **парных** контрастов это сетевая диаграмма расстояний, то есть графическое представление таблицы сравнений, в котором расстояния между узлами сети соответствуют различиям между выборками. Желтые линии соответствуют статистически значимым различиям; черные линии соответствуют незначимым различиям. Наведение указателя мыши на линию в сети приведет к выводу контекстной строки со скорректированным значением значимости различия между узлами, соединенными данной линией.

Для контрастов **отклонений** выводится столбчатая диаграмма, где оцененные по модели значения целевой переменной выводятся по вертикальной оси, а значения поля контрастов - по горизонтальной оси, а для взаимодействий выводится диаграмма, отражающая каждое сочетание уровней эффектов, отличных от поля контрастов. Столбики показывают разницу между каждым из уровней поля контрастов и общим средним, изображаемым черной горизонтальной линией.

Для **простых** контрастов выводится столбчатая диаграмма, отображающая оцененные по модели значения целевой переменной по вертикальной оси и значения поля контрастов по горизонтальной оси, а для взаимодействий выводится диаграмма, отражающая каждое сочетание уровней эффектов, отличных от поля контрастов. Столбики показывают разницу между каждым из уровней поля контрастов (кроме последнего) и последним уровнем, изображаемым черной горизонтальной линией.

- **Таблица.** В этом стиле выводится таблица с оценками значений целевой переменной, сделанных по модели, стандартными ошибками и доверительными интервалами для каждой комбинации уровней полей в эффекте, при сохранении значений всех остальных предикторов неизменными.

  Если затребованы контрасты, выводится другая таблица, содержащая оценки значений, стандартные ошибки, критерии значимости и доверительные интервалы для каждого контраста. Для взаимодействий выводится отдельный набор строк, отражающий все сочетания уровней эффектов, отличные от поля контрастов. Кроме того, выводится таблица с результатами общего теста, а для взаимодействий дается отдельный общий тест для каждого сочетания уровней эффектов, отличного от поля контрастов.

Показатель доверия. Позволяет переключаться с верхней границы доверительного интервала на нижнюю для маргинальных средних, используя уровень значимости, заданный в группе Параметры построения.

Компоновка. Переключает компоновку диаграммы парного сравнения контрастов. Круговой формат диаграммы менее чувствителен к контрастам, чем сетчатый, но в нем нет пересекающихся линий.

Параметры: При оценке модели нужно сгенерировать элементы, выбранные на этой вкладке. Предсказанное значение (для всех полей назначения) и достоверность (для категориальных полей назначения) вычисляются при оценке всегда. Вычисленный показатель доверия может быть основан на вероятности предсказанного значения (наивысшая предсказанная вероятность) или на разнице между наивысшей предсказанной вероятностью и вторым по величине значением предсказанной вероятности.

- **Предсказанную вероятность для категориальных целевых полей.** Будут вычислены предсказанные вероятности для категориальных полей назначения. Для каждой категории создается поле.
- **Оценки склонности для флаговых полей назначения .** Для моделей с флаговым полем назначения (которое возвращает предсказание Да или Нет) можно затребовать оценку склонности, которая обозначает правдоподобие правильного выходного значения, заданного для поля назначения. Модель находит простые оценки склонности; кроме того, если действуют разбиения, модель находит скорректированные оценки склонности на основе контрольного раздела.

Узел Кокса

Регрессия Кокса строит прогнозную модель для данных о времени до наступления события. Эта модель генерирует функцию выживания, предсказывающую вероятность того, что исследуемое событие должно произойти в данный момент времени *t* для данных значений предикторных переменных. Форма функции выживания и коэффициенты регрессии для предикторов оцениваются из наблюдаемых групп; модель может в дальнейшем применяться к новым наблюдениям с известными значениями предикторных переменных. Имейте в виду, что информация из цензурированных субъектов, то есть субъектов, не зарегистрировавших исследуемое событие в течение времени наблюдения, вносят полезный вклад в оценку модели.

Пример. Прикладывая усилия к сокращение оттока клиентов, телекоммуникационная компания хочет, в частности, смоделировать "время текучести", чтобы определить факторы, связанные с клиентами, быстро переключающимися на услуги других компаний. Для этого определяется случайная выборка клиентов, и из базы данных достаются данные о времени использования услуг (продолжают ли эти клиенты активно использовать услуги) и данные различных демографических полей.

Требования. Вам требуется одно или несколько входных полей, ровно одно поле назначения, а также нужно задать поле времени дожития на узле Кокса. Поле назначения должно быть закодировано таким образом, чтобы значение "false" обозначало продолжающееся обслуживание, а значение "true" - что изучаемое событие произошло; у него должен быть уровень измерения *Флаг*, а хранение - строковое или целочисленное. (При необходимости систему хранения можно преобразовать, используя узел Заполнитель или узел Извлечение). Поля с заданными значениями *Оба* или *Нет* игнорируются. У используемых в модели полей должны быть полностью конкретизированы типы. Время дожития может быть любым числовым полем.

Дата и время. Поля Дата и время не могут непосредственно использоваться для определения времени дожития; если у вас есть поля Дата и время, их нужно использовать для создания поля, содержащего значения времени дожития, применяя разность между датой прихода клиента и датой наблюдения.

Анализ Каплана-Майера. Регрессию Кокса можно выполнить без входных полей. Это эквивалентно анализу Каплана-Майера.

Опции полей узла Кокса

Время дожития. Выберите числовое поле (то есть поле с уровнем измерения *Количественный*), чтобы сделать этот узел исполняемым. Время дожития обозначает продолжительность жизни записи, которая предсказывается. Например, при моделировании времени присутствия клиента для текучести клиентской базы это может быть поле, в котором записывается, как долго клиент был в контакте с организацией. Дата, когда клиент начал обслуживаться или отказался от услуг, не влияет на модель; значимой будет только длительность предоставления клиенту услуг.

Время дожития принимается как длительность без единиц измерения. Вы должны убедиться, что входные поля соответствуют времени дожития. Например, при изучении текучести по месяцам в качестве входных нужно использовать данные месячных продаж, а не годовых. Если у ваших данных есть даты начала и конца вместо продолжительности, необходимо перекодировать их в длительность выше узла Кокса.

Остальные поля в этом диалоговом окне - стандартные и используемые для всего продукта IBM SPSS Modeler. Дополнительную информацию смотрите в разделе “Моделирование опций полей узла” на стр. 31.

Опции модели узла Кокса

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Создание моделей расщепления. Строит отдельную модель для каждого возможного значения входных полей, заданных как поля расщепления. Дополнительную информацию смотрите в разделе “Построение моделей расщепления” на стр. 28.

Метод. Доступны следующие опции для ввода предикторов в модель:

- **Ввод.** Это способ по умолчанию, когда все члены прямо вводятся в модель. При построении модели выбор полей не производится.
- **Пошаговый.** Пошаговый способ отбора полей строит модель от шага к шагу, как и обозначено названием. Исходная модель - это простейшая модель без членов модели (кроме константы). На каждом шаге оцениваются члены, которые еще не добавлены в модель, и добавляются те из них, которые вносят наиболее существенный вклад в предсказательную силу модели. Кроме этого, уже присутствующие в модели члены оцениваются повторно для определения, можно ли удалить некоторые из них без существенного воздействия на модель. Если такие члены есть, они удаляются. Процесс повторяется, и другие члены добавляются и/или удаляются. Когда больше нет членов, которые можно добавить для улучшения модели, и нет членов, которые можно удалить из модели без существенного ее ухудшения, генерируется окончательная модель.
- **Обратный пошаговый.** Обратный пошаговый способ существенно противоположен пошаговому. В этом способе в начальную модель включаются все возможные предикторы. На каждом шаге уже присутствующие в модели члены оцениваются для определения, можно ли удалить некоторые из них без существенного воздействия на модель. Кроме этого, повторно оцениваются и члены, ранее удаленные из модели, чтобы определить, не повысят ли лучшие из них предсказательную силу модели. Если так, данные члены возвращаются в модель. Когда больше нет членов, которые можно удалить из модели без существенного ее ухудшения, и нет членов, которые можно добавить для улучшения модели, генерируется окончательная модель.

Примечание: Автоматические способы, в том числе пошаговый и обратный пошаговый, - это очень адаптивные способы обучения, и у них есть сильная тенденции к сверхобучению обучающих данных. При использовании этих способов особенно важно проверить приемлемость результирующей модели или на новых данных, или на существующей тестовой выборке, созданной на узле Разделы.

Группы. Задание полей групп приводит к тому, что узел вычисляет отдельные модели для каждой категории поля. Это может быть любое категориальное поле (Флаг или Номинал) со строковой или целочисленной системой хранения.

Тип модели. Есть две опции для определения членов в модели. Модели **Главные эффекты** включают в себя только входные поля по отдельности, и не проверяют взаимодействий (мультипликативных эффектов) между входными полями. **Пользовательские** модели включают в себя только члены (главные эффекты и взаимодействия), которые задаете вы сами. При выборе этой опции используйте список Члены модели, чтобы добавить члены в модель или удалить их оттуда.

Члены модели. При построении Пользовательской модели вам нужно явно указать члены в этой модели. Этот список показывает текущий набор членов для модели. Кнопки справа от списка Члены модели позволяют добавлять или удалять члены модели.

- Чтобы добавить члены в модель, нажмите кнопку *Добавить новые члены модели*.
- Чтобы удалить какие-то члены, выберите их и нажмите кнопку *Удалить выбранные члены модели*.

Добавление членов в модель регрессии Кокса

При запросе пользовательской модели можно добавить члены в модель, нажав кнопку *Добавить новые члены модели* на вкладке Модель. Появится новое диалоговое окно, в котором можно задать эти члены.

Тип добавляемого члена. Есть несколько способов добавить члены в модель, основываясь на выборе входных полей в списке Доступные поля.

- **Простое взаимодействие.** Вставляет члены, представляющие взаимодействие всех выбранных полей.
- **Главные эффекты.** Вставляет один член главного эффекта (само поле) для каждого выбранного поля.

- **Все двухсторонние взаимодействия.** Вставляет член двухстороннего взаимодействия (произведение входных полей) для каждой возможной пары выбранных входных полей. Например, если в списке Доступные поля выбраны входные поля *A*, *B* и *C*, этим способом можно вставить члены *A * B*, *A * C* и *B * C*.
- **Все трехсторонние взаимодействия.** Вставляет член трехстороннего взаимодействия (произведение входных полей) для всех возможных комбинаций трех одновременных выбранных полей. Например, если в списке Доступные поля выбраны входные поля *A*, *B*, *C* и *D*, этим способом можно вставить члены *A * B * C*, *A * B * D*, *A * C * D* и *B * C * D*.
- **Все четырехсторонние взаимодействия.** Вставляет член четырехстороннего взаимодействия (произведение входных полей) для всех возможных комбинаций четырех одновременных выбранных полей. Например, если в списке Доступные поля выбраны входные поля *A*, *B*, *C*, *D* и *E*, этим способом можно вставить члены *A * B * C * D*, *A * B * C * E*, *A * B * D * E*, *A * C * D * E* и *B * C * D * E*.

Доступные поля. Выводит список доступных входных полей, которые будут использоваться при конструировании членов модели. Обратите внимание на то, что в этот список могут включаться поля, которые нельзя использовать как входные, поэтому убедитесь, что все члены модели включают в себя только входные поля.

Просмотр. Показывает члены, которые будут включены в модель после нажатия кнопки **Вставить** на основе выбранных полей и выбранных выше типов членов.

Вставить. Вставляет члены в модель (на основе текущего выбора полей и типов членов) и закрывает диалоговое окно.

Дополнительные опции узла Кокса

Сходимость. Эти опции позволяют вам управлять параметрами для сходимости модели. Когда выполняется модель, настройки сходимости управляют тем, сколько раз проводятся повторные запуски с различными параметрами, чтобы увидеть, хорошо ли они подходят. Чем чаще испытываются параметры, тем ближе друг к другу результаты (то есть результаты сходятся). Дополнительную информацию смотрите в разделе “Критерии сходимости узла Кокса”.

Вывод. Эти опции позволяют затребовать дополнительные статистические показатели и графики, в том числе кривую дожития, которые будут выведены в расширенном выходе построенной узлом модели. Дополнительную информацию смотрите в разделе “Опции расширенного вывода узла Кокса”.

Пошаговые опции. Эти опции позволяют вам управлять критериями для добавления и удаления полей при пошаговом способе оценки. (Эта кнопка отключена, если выбран способ Ввод). Дополнительную информацию смотрите в разделе “Критерии шага узла Кокса” на стр. 207.

Критерии сходимости узла Кокса

Максимум итераций. Позволяет задать максимальное число итераций для модели, управляющее тем, как долго процедура будет выполнять поиск решения.

Сходимость логарифмического правдоподобия. Итерации прекращаются, если относительное изменение Log-правдоподобия меньше этого значения. Этот критерий не используется, если значение равно 0.

Сходимость параметров. Итерации прекращаются, если абсолютное или относительное изменение в оценках параметра меньше этого значения. Этот критерий не используется, если значение равно 0.

Опции расширенного вывода узла Кокса

Статистика. Можно получить статистики для параметров модели, включая доверительные интервалы для exp(*B*) и корреляцию оценок. Эти статистики можно получить либо по каждому, либо только по последнему шагу.

Вывести базовую функцию. Позволяет вывести на экран базовую функцию риска и кумулятивную функцию выживания, проанализированные по среднему значению ковариат.

Графики

Графики помогают оценить модель и интерпретировать результаты. Можно построить графики функций дожития, риска, логарифм минус логарифм и единица минус выживание.

- *Выживание*. Выводит кумулятивную функцию дожития (надежности) в линейном масштабе.
- *Риск*. Выводит функцию накопленного риска в линейном масштабе.
- **Log минус log.** Содержит оценку кумулятивной функции дожития после применения к ней преобразования ln(-ln).
- *Единица минус выживание*. В линейном масштабе выводится график функции, равной 1 минус функция дожития (надежности).

Нарисовать отдельную линию для каждого значения. Эта опция доступна только для категориальных полей.

Значения, используемые для графиков. Так как эти функции зависят от значений предикторов, необходимо использовать постоянные значения для предикторов при построении функций от времени. По умолчанию в качестве постоянного значения предиктора используется его среднее значение, но с помощью сетки для графика можно ввести выбранные вами значения. Для категориальных входных данных используется индикаторное кодирование, поэтому для каждой категории существует коэффициент регрессии (кроме последнего). Таким образом, у категориальных входных данных есть среднее значение для каждого контраста индикатора, равное доле наблюдений в категории, соответствующей этому контрасту индикатора.

Критерии шага узла Кокса

Критерий исключения. Выберите **Отношение правдоподобия** для более устойчивой модели. Для сокращения времени на построение модели попробуйте выбрать **Вальд**. Существует дополнительная опция **Условно**, которая предоставляет проверку на удаление, основанную на вероятности статистики отношения правдоподобия на базе оценок условных параметров.

Пороги значимости для критериев. Эта опция позволяет задать критерии выбора на основе статистической вероятности (значение *p*), связанной с каждым полем. Поля будут добавлены в модель, если только связанное значение *p* меньше, чем значение **Ввод**, и будут удалены, если только значение *p* больше, чем значение **Удаление**. Значение **Ввод** должно быть меньше, чем значение **Удаление**.

Опции параметров узла Кокса

Предсказать дожитие в будущем времени. Задает одно или несколько значений времени в будущем. Для каждой записи в каждый момент времени предсказывается дожитие, то есть доживет ли данное наблюдение по крайней мере до заданного момента от настоящего без конечного события, по одному предсказанию на каждый момент времени. Обратите внимание на то, что дожитию соответствует значение "false" в поле назначения.

- **Регулярные интервалы.** Значения времени дожития генерируются из заданного **Интервала времени** и **Количества периодов времени для оценки**. Например, если затребованы три периода времени с интервалом 2 между последовательными моментами времени, дожитие будет предсказываться в моменты времени 2, 4, 6. Каждая запись оценивается при одинаковых значениях времени.
- **Поля времени.** Времена дожития предоставляются для каждой записи в выбранном поле времени (генерируется одно поле предсказания), таким образом каждую запись можно оценить в разные моменты времени.

Прошлое время существования. Задайте время существования записи до сих пор, например, поле для длительности предоставления услуг существующему клиенту. Оценка вероятности дожития в будущем будет зависеть от прошлого времени существования.

Примечание: Значения времени дожития в будущем и времени существования в прошлом должны быть в диапазоне времени данных, использованных для обучения модели. Записи, время которых оказывается вне этого диапазона, оцениваются как пустые.

Добавить все вероятности. Задает, добавляются ли вероятности для каждой категории выходного поля к каждой записи, обрабатываемой узлом. Если эта опция не выбрана, добавляется только вероятность предсказанной категории. Вероятности вычисляются для каждого значения времени в будущем.

Вычислить кумулятивную функцию риска. Задает, добавляется ли к каждой записи значение кумулятивного риска. Кумулятивный риск вычисляются для каждого значения времени в будущем.

Слепок модели Кокса

Модели регрессии Кокса представляют собой уравнения, оцененные узлом Кокса. Они содержат всю собранную моделью информацию, а также информацию о структуре и производительности модели.

При запуске потока, содержащего сгенерированную модель регрессии Кокса, узел добавляет два новых поля, содержащих предсказание модели и связанную вероятность. Имена новых полей получаются из имени выходного поля, которое предсказывается, с добавлением префикса *$C-* для предсказанной категории и префикса *$CP-* для связанной вероятности, а также с добавлением суффикса, равного номеру интервала времени в будущем или имени поля времени, определяющего интервал времени. Например, для выходного поля с именем *churn* и двух регулярных интервалов времени в будущем, новые поля могут называться *$C-churn-1*, *$CP-churn-1*, *$C-churn-2* и *$CP-churn-2*. Если моменты времени в будущем определены в поле времени *tenure*, новыми полями будут *$C-churn_tenure* и *$CP-churn_tenure*.

Если вы выбрали опцию параметров **Присоединить все вероятности** на узле Кокса, будут добавлены два дополнительные поля для каждого момента в будущем, содержащие вероятности дожития и конечного события для каждой записи. Эти дополнительные поля называются по имени выходного поля с добавлением префикса *$CP-<значение для false>-* для вероятности дожития и *$CP-<значение для true>-* для вероятности, что событие произошло, и с добавлением суффикса, равного номеру интервала времени в будущем. Например, для выходного поля, в котором значение "false" - это 0, а значение "true" - это 1, при определении двух регулярных интервалов времени в будущем новые поля могут называться *$CP-0-1*, *$CP-1-1*, *$CP-0-2* и *$CP-1-2*. Если моменты времени в будущем определены в поле одного времени *tenure*, новыми полями будут *$CP-0-1* и *$CP-1-1*, так как есть только один интервал в будущем

Если вы выбрали опцию параметров **Вычислить кумулятивную функцию рисков** на узле Кокса, дополнительное поле будет добавлено для каждого момента времени в будущем и будет содержать кумулятивную функцию риска для каждой записи. Имена этих дополнительных полей получаются из имени выходного поля с добавлением префикса *$CH-* и суффикса, номера интервала времени в будущем или имени поля времени, определяющего интервал времени. Например, для выходного поля с именем *churn* и двух регулярных интервалов времени в будущем новые поля могут называться *$CH-churn-1* и *$CH-churn-2*. Если моменты времени в будущем определены полем времени *tenure*, именем нового поля может быть *$CH-churn-1*.

Параметры вывода регрессии Кокса

Вкладка Параметры для слепка содержит те же управляющие элементы, что и вкладка Параметры узла модели. Значения по умолчанию управляющих элементов слепка определяются значениями, заданными на узле модели. Дополнительную информацию смотрите в разделе “Опции параметров узла Кокса” на стр. 207.

Расширенный вывод регрессии Кокса

Расширенный вывод регрессии Кокса дает подробную информацию об оцененной модели и ее производительности, в том числе о кривой дожития. Большая часть содержащейся в расширенном выводе информации техническая, и требуется глубокое знание регрессии Кокса, чтобы правильно интерпретировать этот вывод.

Глава 11. Модели кластеризации

Модели кластеризации уделяют главное внимание идентификации групп сходных записей и присвоению меток записям в соответствии с группами, к которым они принадлежат. При этом не используются преимущества предварительных знаний о группах и их характеристиках. Фактически вы можете даже не знать, сколько именно групп ищется. Это отличает модели кластеризации от других приемов машинного обучения - нет предварительно определенного выходного поля (поля назначения), значение в котором предсказывалось бы моделью. Эти модели часто называют моделями **неконтролируемого обучения**, так как не существует внешнего стандарта, по которому можно было бы судить о выполнении классификации модели. У этих моделей нет *правильных* или *неправильных* ответов. Их ценность в способности захватывать интересные группировки данных и представлять полезные описания этих группировок.

Способы кластеризации основываются на измерении расстояний между записями и между кластерами. Записи назначаются кластерам таким образом, чтобы минимизировать расстояние между записями, принадлежащими одному кластеру.

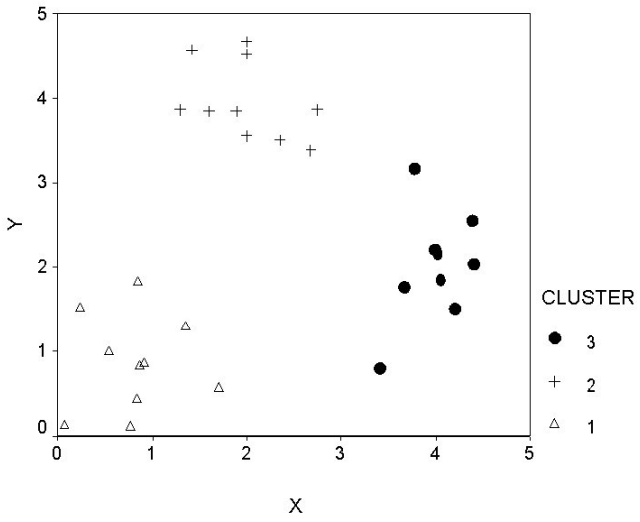


Рисунок 44. Простая модель кластеризации

Предоставляется три способа кластеризации:

Узел K-средних кластеризует набор данных в отдельные группы (или кластеры). Этот метод определяет фиксированное количество кластеров, итерационно распределяет записи по кластерам и настраивает центры кластеров, пока дальнейшие уточнения более не улучшают модель. Вместо попытки предсказать выходное значение *k*-средние используют процесс, называемый неконтролируемым обучением, чтобы обнаружить структуры в наборе входных полей.

Узел Двухшаговый использует метод двухшаговой кластеризации. На первом шаге проводится первый проход по данным, при котором необработанные входные данные сжимаются в управляемый набор подкластеров. На втором шаге используется способ иерархической кластеризации для все большего слияния подкластеров в крупные и еще более крупные кластеры. У двухшагового метода есть преимущество автоматической оценки оптимального числа кластеров для обучающих данных. Он может эффективно обрабатывать поля смешанных типов и большие наборы данных.

Узел Коонена генерирует тип нейросети, которую можно использовать для кластеризации набора данных в отдельные группы. Когда сеть полностью обучена, похожие записи должны быть близко друг от друга на выходной карте, а отличающиеся записи должны быть сильно разделены. По количеству наблюдений, захваченных каждым нейроном в слепке модели, можно определить сильные нейроны. Это может дать представление об оправданном количестве кластеров.

Модели кластеризации часто используются для создания кластеров или сегментов, которые используются в качестве входных данных при последующем анализе. Простой пример этого - сегменты рынка, используемые маркетологами для разделения рынка их продукции на однородные подгруппы. У каждого сегмента есть свои специальные характеристики, влияющие на успех нацеленных на данный сегмент маркетинговых усилий. Если для оптимизации маркетинговой стратегии используется исследование данных, можно существенно повысить значимость моделей, определив соответствующие сегменты и используя информацию об этих сегментах в ваших предсказательных моделях.

Узел Коонена

Сети Коонена - это тип нейросетей, выполняющих кластеризацию; их называют также **knet** или **самоорганизующимися картами**. Этот тип сети может использоваться для кластеризации набора данных в отдельные группы, когда вы не знаете, что эти группы представляют собой сначала. Записи группируются таким образом, чтобы они были похожи друг на друга в группе (кластере), но различались в разных группах.

Базовые элементы - это **нейроны**, причем они организованы в два уровня: **входные уровни** и **выходные уровни** (другое название - **выходная карта**). Все входные нейроны соединены со всеми выходными нейронами, и у этих соединений есть связанные с ними **сила** или **вес**. Во время обучения каждый нейрон соревнуется с другими, чтобы "выиграть" каждую запись.

Выходная карта - это решетка нейронов без соединений между ними.

Входные данные представляются входному слою, а значения распространяются на выходной слой. О выходном нейроне с самым сильным откликом говорят как о **победителе** и ответе на данный входной сигнал.

Первоначально распределение весов равномерно. Когда нейрон выигрывает запись, все веса (в том числе близлежащих нейронов, о которых говорят как о **соседях**) корректируются для лучшего соответствия структуре значений предикторов для этой записи. Так же обрабатываются все входные записи, и соответственно изменяются все веса. Этот процесс повторяется многократно, пока изменения не станут очень малыми. В процессе обучения веса нейронов сетки настраиваются так, что они образуют двумерную "карту" кластеров (отсюда происходит и термин **самоорганизующаяся карта**).

Когда сеть полностью обучена, похожие записи должны быть близко друг от друга на выходной карте, а отличающиеся записи должны быть сильно разнесены.

В отличие от большинства методов обучения в IBM SPSS Modeler, сети Коонена *не* используют поле назначения. Такой тип обучения без поля назначения называют **неконтролируемым обучением**. Вместо попытки предсказать выходное значение, сети Коонена пытаются выявить структуры в наборе входных полей. Обычно сеть Коонена дает в результате несколько нейронов, суммирующих многие наблюдения (**сильные** нейроны), и несколько нейронов, которые на самом деле не соответствуют никакому наблюдению (**слабые** нейроны). Сильные нейроны (и иногда некоторые другие нейроны, связанные с сильными в сетке) представляют возможные центры кластеров.

Другое применение сетей Коонена - это **уменьшение числа измерений** (снижение размерности). Пространственные характеристики двумерной решетки обеспечивают отображение *k* исходных предикторов на два извлеченных показателя, сохраняющих взаимосвязь сходства в исходных предикторах. В некоторых случаях это может дать выигрыш того же сорта, как при факторном или PCA-анализе.

Обратите внимание на то, что способ вычисления размера выходной решетки изменился по сравнению с предыдущими версиями IBM SPSS Modeler. Обычно новый способ создает меньше выходных слоев, которые быстрее обучаются и лучше обобщаются. Если обнаруживается, что при размере по умолчанию получаются худшие результаты, попробуйте увеличить размер выходной решетки на вкладке Эксперт. Дополнительную информацию смотрите в разделе “Дополнительные опции узлов Коонена” на стр. 212.

Требования. Для обучения сети Коонена нужно одно или несколько полей с заданной ролью *Входное*. Поля с заданными ролями *Назначение*, *Оба* или *Нет* игнорируются.

Достоинства. Для построения модели сети Коонена вам не нужны данные о принадлежности к группе. Не нужно даже знать, сколько групп надо искать. Сети Коонена начинают с большого количества нейронов, и по ходу обучения эти нейроны сжимаются в естественные кластеры данных. Для идентификации сильных нейронов можно посмотреть на количество наблюдений, захваченных каждым нейроном, что даст представление об уместном числе кластеров.

Опции моделей узла Коонена

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Продолжить обучение существующей модели. По умолчанию при каждом выполнении узла Коонена создается полностью новая сеть. Если выбрана эта опция, обучение продолжается с последней сети, успешно созданной узлом.

Показать график обратной связи. Если выбрана эта опция, при обучении выводится визуальное представление двумерного массива. Сила каждого узла представлена цветом. Красный обозначает нейрон, выигрывающий многие записи (**сильный** нейрон), а белый - это нейрон с небольшим числом выигрышей записей или вовсе без них (**слабый** нейрон). Обратная связь может не выводиться, если время построения модели относительно короткое. Обратите внимание на то, что эта возможность может затянуть обучение. Для ускорения обучения отмените выбор этой опции.

Когда остановить. По умолчанию критерий остановки останавливает обучение на основании внутренних параметров. Как критерий остановки можно задать также время. Введите время обучения сети (в минутах).

Задать начальное значение генератора псевдослучайных чисел. Если начальное значение генератора псевдослучайных чисел не задано, последовательность случайных значений, используемая для инициализации сетевых весов, будет различаться при каждом выполнении узла. Это может привести к тому, что узел будет создавать разные модели при каждом запуске, даже если параметры узла и значения данных абсолютно одинаковые. Выбрав эту опцию, вы можете задать конкретное начальное значение генератора псевдослучайных чисел, чтобы полученная модель была воспроизводимой. Конкретное начальное значение генератора псевдослучайных чисел всегда генерирует одинаковую последовательность случайных значений, и в этом случае выполнение узла всегда приводит к одной генерируемой модели.

Примечание: При использовании опции **Задать начальное значение генератора псевдослучайных чисел** для записей, читаемых из базы данных, предварительно может потребоваться узел Сортировка для подготовки выборки, чтобы обеспечить одинаковый результат при каждом выполнении узла. Это связано с тем, что начальное значение генератора псевдослучайных чисел зависит от порядка записей, который не обязательно остается постоянным в реляционной базе данных.

Примечание: Если вы хотите включить в модель номинальные (набор) поля, но есть проблемы с размером памяти при построении модели или модель строится слишком долго, рассмотрите возможность перекодирования полей больших наборов, чтобы уменьшить количество значений, или возможность использования другого поля с меньшим числом значений в качестве прокси для большого набора.

Например, если есть проблема с полем *ID_товара*, содержащим значения для индивидуальных товаров, можно рассмотреть возможность его удаления из модели и добавления вместо этого менее подробного поля *категория_товара*.

Оптимизировать. Выберите во время построения моделей предназначенные для повышения производительности опции в соответствии с конкретными потребностями.

- Выберите опцию **Скорость**, чтобы для повышения производительности алгоритм никогда не использовал сброс на диск.
- Выберите опцию **Память**, чтобы в подходящих ситуациях алгоритм использовал сброс на диск, хотя это понизит скорость. Эта опция выбирается по умолчанию.

  Примечание: При запуске в распределенном режиме этот параметр может быть перезаписан значением опции администратора из файла *options.cfg*.

Присоединить метку кластера. Выбирается по умолчанию для новых моделей, но не применяется для моделей, загруженных из более старых версий IBM SPSS Modeler, в которых создается одно поле категориальной оценки того же типа, что и на узлах K-средних и двухшаговых моделей. Это строковое поле используется узлом автокластеризации при вычислении показателей ранжирования для различных типов моделей. Дополнительную информацию смотрите в разделе “Узел Автокластеризация” на стр. 74.

Дополнительные опции узлов Коонена

Если вы хорошо знакомы с сетями Коонена, опции эксперта позволят точнее настроить процесс обучения. Для доступа к дополнительным опциям задайте на вкладке Дополнительно режим **Дополнительно**.

Ширина и длина. Задайте размер (ширину и длину) двумерной карты вывода как количество выходных нейронов по каждому измерению.

Сокращение скорости обучения. Выберите линейное или экспоненциальное сокращение скорости обучения. **Скорость обучения** - это коэффициент взвешивания, уменьшающийся со временем, так что сеть начинает с кодирования крупномасштабных показателей данных и постепенно переходит на уровень подробностей.

Фаза 1 и Фаза 2. Обучение сети Коонена разбивается на две фазы. Фаза 1 - это фаза грубой оценки, используемая для захвата крупных структур в данных. Фаза 2 - это фаза настройки, используемая для корректировки карты и моделирования более тонких особенностей данных. Для каждой карты есть три параметра:

- **Соседний.** Задает начальный размер (радиус) окрестности. Этим определяется количество "близких" нейронов, которые обновляются при выигрыше нейрон во время обучения. Во время фазы 1 размер окрестности сначала равен *Окрестности фазы 1*, а затем уменьшается до значения (*Окрестность фазы 2* + 1). Во время фазы 2 размер окрестности сначала равен *Окрестности фазы 2*, а затем уменьшается до 1,0. *Окрестность фазы 1* должна быть больше *Окрестности фазы 2*.
- **Начальная эта.** Задает начальное значение для скорости обучения **эта**. Во время фазы 1 эта начинает изменяться от значения *Начальная эта фазы 1* и уменьшается до *Начальной эты фазы 2*. Во время фазы 2 эта начинает изменяться от *Начальной эты фазы 2* и уменьшается до 0. Значение *Начальной эты фазы 1* должно быть больше *Начальной эты фазы 2*.
- **Циклы.** Задает количество циклов для каждой фазы обучения. Каждая фаза продолжается в течение заданного числа проходов по данным.

Слепки моделей Коонена

Слепки моделей Коонена содержат всю информацию, захваченную обучающей сетью Коонена, а также информацию об архитектуре сети.

При запуске потока, содержащего слепок модели Коонена, узел добавляет два новые поля, содержащие координаты *X* и *Y* элемента в сетке вывода Коонена, который сильнее всего реагирует на данную запись.

Имена новых полей получаются из имени модели с префиксами *$KX-* и *$KY-*. Например, если модель называется *Kohonen*, новые поля будут называться *$KX-Kohonen* и *$KY-Kohonen*.

Для лучшего понимания, что именно закодировала сеть Коонена, щелкните по вкладке Модель в браузере слепков моделей. Откроется средство просмотра кластеров, обеспечивающее графическое представление кластеров, полей и уровней значимости. Дополнительную информацию смотрите в разделе “Средство просмотра кластеров - Вкладка Модель” на стр. 218.

Если вы предпочитаете визуализировать кластеры в виде сетки, можно просмотреть результат сети Коонена, изобразив поля *$KX-* и *$KY-* в узле График. (В узле Plot необходимо выбрать **X-Agitation** и **Y-Agitation**, чтобы предотвратить вывод записей каждого элемента поверх друг друга). На графике можно также наложить символическое поле поля, чтобы исследовать, как сеть Коонена кластеризовала данные.

Другое мощное средство для понимания сети Коонена - это использование вывода правил методом индукции для обнаружения характеристик, различающих найденные моделью кластеры. Дополнительную информацию смотрите в разделе “Узел C5.0” на стр. 104.

За общей информацией по браузеру моделей обратитесь к разделу “Просмотр слепков моделей” на стр. 41

Сводка модели Коонена

На вкладке Сводка для слепка модели Коонена выводится информация об архитектуре или топологии сети. Длина и ширина двумерной карты показателей Коонена (выходной слой) показаны как **$KX-** *имя_модели* и **$KY-** *имя_модели*. Для входного и выходного слоев показано количество нейронов в этом слое.

Узел K-средних

Узел K-средних предоставляет метод **кластерного анализа**. Он может использоваться для кластеризации набора данных в отдельные группы, когда вы не знаете, что эти группы представляют собой в начале. В отличие от большинства способов обучения в IBM SPSS Modeler, модели K-средних *не* используют поле назначения. Этот тип обучения (без поля назначения) называется **неконтролируемым обучением**. Модель K-средних не пытается предсказать исход, а старается выявить шаблоны в наборе входных полей. Записи группируются таким образом, чтобы они были похожи друг на друга в группе (кластере), но различались в разных группах.

Модель K-средних работает посредством определения набора центров начальных кластеров. Затем он назначает каждую запись в кластер, с которым она больше всего схожа, на основе значений входных полей записи. После того, как все наблюдения назначены, центры кластеров обновляются в соответствии с новым набором назначенных в каждый кластер записей. Затем для записей снова проверяется, не следует ли их переназначить в другой кластер, и это итерационный процесс назначение-кластер продолжается либо, пока не будет достигнуто максимальное число итераций, либо когда изменению между двумя смежными итерациями не удастся превысить заданный порог.

Примечание: Полученная модель в определенной степени зависит от порядка обучающих данных. Переупорядочивание данных и повторное построение модели может привести к другой итоговой модели кластера.

Требования. Для обучения модели K-средних требуется одно или несколько полей с заданной для них ролью *Входное*. Поля с заданной ролью *Выходное*, *Двойного назначения* или *Нет* игнорируются.

Достоинства. Данные о членстве в группах для построения модели K-средних не требуются. Модель K-средних часто является самым быстрым способом кластеризации больших наборов данных.

Опции моделей узла K-средних

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Заданное число кластеров. Задайте количество кластеров, которые будут генерироваться. Значение по умолчанию - 5.

Генерировать поле расстояния. Если выбрана эта опция, слепок модели включит в себя поле, содержащее расстояние между каждой записью и центром назначенного ей кластера.

Метка кластера. Задайте формат для значений в сгенерированном поле принадлежности к кластеру. Принадлежность к кластеру может обозначаться как **Строка** с заданным **Префиксом метки** (например, `"Кластер 1"`, `"Кластер 2"` и так далее) или как **Число**.

Примечание: Если вы хотите включить в модель номинальные (набор) поля, но есть проблемы с размером памяти при построении модели или модель строится слишком долго, рассмотрите возможность перекодирования полей больших наборов, чтобы уменьшить количество значений, или использования другого поля с меньшим числом значений в качестве прокси для большого набора. Например, если есть проблема с полем *ID_товара*, содержащим значения для отдельных товаров, можно рассмотреть возможность его удаления из модели и добавления вместо этого менее подробного поля *категория_товара*.

Оптимизировать. Выберите во время построения моделей предназначенные для повышения производительности опции в соответствии с конкретными потребностями.

- Выберите опцию **Скорость**, чтобы для повышения производительности алгоритм никогда не использовал сброс на диск.
- Выберите опцию **Память**, чтобы в подходящих ситуациях алгоритм использовал сброс на диск, хотя это понизит скорость. Эта опция выбирается по умолчанию.

  Примечание: При запуске в распределенном режиме этот параметр может быть перезаписан значением опции администратора из файла *options.cfg*.

Опции эксперта узла K-средних

Если вы хорошо знакомы с кластеризацией *k*-средних, дополнительные опции позволят точнее настроить процесс обучения. Для доступа к дополнительным опциям выберите режим **Дополнительно** на вкладке Дополнительно.

Когда остановить. Задайте критерий остановки, который будет использоваться при обучении модели. Критерий остановки **По умолчанию** - это 20 итераций или изменение меньше 0,000001; выбирается событие, которое произойдет первым. Выберите **Настроить**, чтобы задать собственный критерий остановки.

- **Максимум итераций.** Эта опция позволяет остановить обучение модели после заданного числа итераций.
- **Изменить допуск.** Эта опция позволяет остановить обучение модели, когда наибольшее изменение центров кластеров для итерации окажется меньше заданного уровня.

Значение кодирования для наборов. Задайте значение от 0 до 1,0 для перекодирования полей набора как групп числовых полей. Значение по умолчанию - это корень квадратный из 0,5 (примерно 0,707107), что предоставляет подходящее взвешивание для перекодированных флаговых полей. Значения ближе к 1,0 будут присваивать полям набора больший вес, чем числовым полям.

Слепки моделей k-средних

Слепки моделей K-средних содержат всю информацию, захваченную моделью кластеризации, а также информацию об обучающих данных и процессе оценки.

При запуске потока, содержащего узел моделирования K-средних, этот узел добавляет два новых поля, содержащих информацию о принадлежности к кластеру и расстоянии от назначенного центра кластера для

данной записи. Имена новых полей получаются из имени модели с префиксами *$KM-* для принадлежности к кластеру и *$KMD-* для расстояния от центра кластера. Например, если модель называется *Kmeans*, новые поля будут называться *$KM-Kmeans* и *$KMD-Kmeans*.

Мощным средством для внутреннего понимания модели k-средних может быть изучение правил для обнаружения характеристик, различающих найденные моделью кластеры. Дополнительную информацию смотрите в разделе “Узел C5.0” на стр. 104. Можно перейти также на вкладку Модели в браузере слепок моделей, чтобы вывести средство просмотра кластеров, обеспечивающее графическое представление кластеров, полей и уровней значимости. Дополнительную информацию смотрите в разделе “Средство просмотра кластеров - Вкладка Модель” на стр. 218.

За общей информацией по браузеру моделей обратитесь к разделу “Просмотр слепков моделей” на стр. 41

Сводка моделей K-средних

На вкладке Сводка слепка модели K-средних содержится информация о данных обучения, процессе оценки и определенных моделью кластерах. Выводится количество кластеров, а также хронология итераций. Если вы выполнили узел Анализ, присоединенный к узлу моделирования, информация этого анализа будет также выведена в этом разделе.

Узел двухшаговой кластеризации

Узел двухшаговой кластеризации предоставляет одну из форм **кластерного анализа**. Он может использоваться для кластеризации набора данных в отдельные группы, когда вы не знаете, что эти группы представляют собой в начале. Как и узлы Коонена и узлы K-средних, двухшаговые модели кластеров *не* используют поле назначения. Вместо попытки предсказать выходные данные двухшаговая кластеризация пытается обнаружить структуры в наборе входных полей. Записи группируются таким образом, чтобы они были похожи друг на друга в группе (кластере), но различались в разных группах.

Двухшаговая кластеризация - это метод кластеризации в два этапа. На первом шаге проводится первый проход по данным, при котором необработанные входные данные сжимаются в управляемый набор подкластеров. На втором шаге используется способ иерархической кластеризации для все большего слияния подкластеров в крупные и еще более крупные кластеры, причем следующий проход по данным не требуется. У иерархической кластеризации есть то преимущество, что не требуется заранее выбирать нужное число кластеров. Многие способы иерархической кластеризации начинают с индивидуальных записей как начальных кластеров и рекурсивно объединяют их для получения все более крупных кластеров. Хотя такие подходы часто отказывают при работе с большим объемом данных, начальная предварительная кластеризация двухшагового метода обеспечивает быструю иерархическую кластеризацию даже для больших наборов данных.

Примечание: Полученная модель в определенной степени зависит от порядка обучающих данных. Переупорядочивание данных и повторное построение модели может привести к другой итоговой модели кластера.

Требования. Для обучения двухшаговой кластерной модели одно или несколько полей с заданным значением роли *Вход*. Поля с заданными ролями *Назначение*, *Оба* или *Нет* игнорируются. Двухшаговый алгоритм кластеризации не обрабатывает пропущенные значения. При построении модели будут игнорироваться все записи с пропусками любых входных полей.

Достоинства. Двухшаговая кластеризация может работать со смешанными типами полей и способна эффективно обрабатывать большие наборы данных. При этом есть возможность сравнивать несколько кластерных решений и выбирать лучшее, поэтому вам не нужно знать, сколько кластеров запрашивать для выходного набора. В двухшаговой кластеризации можно задать автоматическое исключение **выбросов** или особенно необычных наблюдений, которые могут испортить результаты.

Опции модели узла двухшаговой кластеризации

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Нормализовать числовые поля. По умолчанию двухшаговый процесс стандартизирует все числовые входные поля, приводя их к среднему 0 и дисперсии 1. Чтобы не масштабировать числовые поля, выключите эту опцию. На символические поля это не влияет.

Исключить выбросы. Если выбрать эту опцию, то записи, которые очевидно не впишутся ни в какой существенный кластер, автоматически исключаются из анализа. Таким наблюдениям не дадут исказить результаты.

Обнаружение выбросов происходит на шаге, предшествующем кластеризации. Когда эта опция выбрана, подкластеры с малым числом записей, по сравнению с другими подкластерами, считаются потенциальными выбросами, и дерево подкластеров строится еще раз без этих записей. Размер, ниже которого подкластеры считаются потенциальными выбросами, задается опцией **Процент**. Некоторые из таких записей - потенциальных выбросов могут быть добавлены в заново построенные кластеры, если окажутся достаточно сходны с профилями каких-либо новых подкластеров. Остальные потенциальные выбросы, которые невозможно ни с чем объединить, добавляются в кластер "шума" и исключаются из шага иерархической кластеризации.

При *оценке* данных с помощью двухшаговой модели, использующей обработку выбросов, новые наблюдения, которые отстоят по меньшей мере на пороговое расстояние от (по логарифму правдоподобия) от ближайшего существенного кластера, считаются выбросами и назначаются в кластер "шума" с именем -1.

Метка кластера. Задайте формат поля принадлежности к сгенерированному кластеру. Принадлежность к кластеру может отмечаться **Строкой** с заданным **Префиксом метки** (например, `"Кластер 1"`, `"Кластер 2"` и так далее) или как **Номер**.

Автоматически вычислять число кластеров. Двухшаговая кластеризация может быстро анализировать большие объемы кластерных решений, выбирая оптимальное число кластеров для данных обучения. Задайте испытываемый диапазон решений через **Максимальное** и **Минимальное** число кластеров. Двухшаговая кластеризация использует двухэтапный процесс для определения оптимального числа кластеров. На первом шаге выбирается верхняя граница для числа кластеров в модели с учетом изменения информационного критерия Байеса (BIC) при добавлении кластеров. На втором шаге для всех моделей, в которых кластеров меньше, чем в решении с минимальным BIC, находится изменение минимального расстояние между кластерами. Наибольшее изменение в расстоянии используется для выбора окончательной модели кластеров.

Задайте число кластеров. Если вы знаете, сколько кластеров должна содержать модель, выберите эту опцию и введите число кластеров.

Мера расстояния. Выбор в этой группе определяет, как вычисляется сходство между двумя кластерами.

- **Log-правдоподобия.** Мера правдоподобия приписывает переменным вероятностное распределение. Предполагается, что непрерывные переменные имеют нормальное распределение, а категориальные переменные - полиномиальное. Все переменные предполагаются независимыми.
- **Евклидова.** Евклидова мера является расстоянием "по прямой линии" между двумя кластерами. Она может быть использована, только когда все переменные являются непрерывными.

Критерий кластеризации. Выбор в этой группе задает способ, которым автоматический алгоритм кластеризации определяет число кластеров. Можно задать либо Байесовский информационный критерий (BIC), либо Информационный критерий Акаике (AIC).

Слепки двухшаговых моделей кластеров

Слепки двухшаговых моделей кластеров содержат всю информацию, захваченную моделью кластеризации, а также информацию об обучающих данных и процессе оценки.

При запуске потока, содержащего слепок двухшаговой модели кластера, узел добавляет новое поле, содержащее информацию о принадлежности к кластеру для данной записи. Имя нового поля получается из имени модели добавлением префикса *$T-*. Например, если модель называется *TwoStep*, новое поле будет называться *$T-TwoStep*.

Мощным средством для внутреннего понимания двухшаговой модели может быть изучение правил для обнаружения характеристик, различающих найденные моделью кластеры. Дополнительную информацию смотрите в разделе “Узел C5.0” на стр. 104. Можно щелкнуть также по вкладке Модели в браузере слепок моделей, чтобы вывести средство просмотра кластеров, обеспечивающее графическое представление кластеров, полей и уровней значимости. Дополнительную информацию смотрите в разделе “Средство просмотра кластеров - Вкладка Модель” на стр. 218.

За общей информацией по браузеру моделей обратитесь к разделу “Просмотр слепков моделей” на стр. 41

Сводка двухшаговой модели

Вкладка Сводка для слепка двухшаговой модели кластеризации содержит число найденных кластеров, а также информацию об обучающих данных, процессе оценки и применявшихся настройках построения.

Дополнительную информацию смотрите в разделе “Просмотр слепков моделей” на стр. 41.

Средство просмотра кластеров

Кластерные модели обычно используются для выявления групп (или кластеров) похожих записей путем исследования переменных, в которых сходство членов одной группы велико, а сходство представителей разных групп мало. Полученные результаты можно использовать для идентификации взаимосвязей, которые другим путем было бы трудно обнаружить. Например, с помощью кластерного анализа предпочтений покупателей, уровня доходов и покупательских привычек можно идентифицировать типы клиентов, которые с большей вероятностью откликнуться на проводимую маркетинговую кампанию.

Имеются два подхода к интерпретации выведенных результатов кластерного анализа:

- Исследовать кластеры с целью выявления уникальных особенностей отдельных кластеров. *Содержит ли один кластер всех заемщиков с высоким доходом? Содержит ли данный кластер больше записей, чем остальные?*
- Исследовать поля по кластерам, чтобы определить, как распределяются значения среди кластеров. *Определяет ли уровень образования конкретного лица принадлежность к кластеру? Определяет ли высокая кредитная оценка принадлежность к тому или иному кластеру?*

Основная и дополнительная панель Средства просмотра кластеров, а также различные виды представления моделей могут помочь получить ответы на эти вопросы.

В IBM SPSS Modeler можно сформировать следующие слепки моделей кластеров:

- Слепок модели сети Коонена
- Слепок модели k-средних
- Слепок двухэтапной кластерной модели

Чтобы получить информацию о слепках кластерной модели, щелкните правой кнопкой мыши по узлу модели и выберите **Обзор** из контекстного меню (или **Правка** для узлов в потоке). Как альтернатива при

использовании узла моделирования Автокластер, щелкните дважды по нужному слепку кластера в слепке модели Автокластер. Дополнительную информацию смотрите в разделе “Узел Автокластеризация” на стр. 74.

Средство просмотра кластеров - Вкладка Модель

Вкладка Модель для кластерных моделей графически показывает итожащие статистики и распределения для полей между кластерами; она называется **Средство просмотра кластеров**.

Примечание: Вкладка Модель недоступна для моделей, построенных в версиях IBM SPSS Modeler до версии 13.

Средство просмотра кластеров состоит из двух панелей: основной, находящейся слева, и дополнительной, находящейся справа. Имеется два основных представления:

- Сводка для модели (по умолчанию). Дополнительную информацию смотрите в разделе “Вид представления Сводка для модели”.
- Кластеры. Дополнительную информацию смотрите в разделе “Вид представления Кластеры” на стр. 219.

В дополнительной панели доступны четыре вида представления:

- Важность предикторов. Дополнительную информацию смотрите в разделе “Вид представления Важность предикторов в кластерах” на стр. 220.
- Размеры кластеров (по умолчанию). Дополнительную информацию смотрите в разделе “Вид представления Размеры кластеров” на стр. 220.
- Распределение ячеек. Дополнительную информацию смотрите в разделе “Вид представления Распределение в ячейке” на стр. 220.
- Сравнение кластеров. Дополнительную информацию смотрите в разделе “Вид представления Сравнение кластеров” на стр. 221.

Вид представления Сводка для модели

В представлении Сводка для модели показан "мгновенный снимок" или сводка для кластерной модели, включая силуэтную меру связности и разделения кластеров, с использованием затенения для индикации низкого, среднего и хорошего качества полученных результатов. "Мгновенный снимок" дает возможность быстро понять, является ли качество разбиения на кластеры низким. В этом случае, возможно, стоит вернуться к узлу моделирования, чтобы скорректировать параметры для построения модели с целью получения более приемлемых результатов.

Решение вопроса о том, являются качество разбиения на кластеры низким, средним или хорошими основывается на работе Кауфмана и Руссью (Kaufman and Rousseeuw (1990)), касающейся интерпретации кластерных структур. Показанное в сводке для модели качество разбиения считается хорошим, если согласно оценке Кауфмана и Руссью имеется обоснованное или сильное свидетельство наличия кластерной структуры в данных. Среднее качество разбиения соответствует их оценке иметь слабое свидетельство, а низкое соответствует оценке не иметь значимого свидетельства наличия кластерной структуры.

Силуэтная мера усредняет по всем записям величину (B–A) / max(A,B), где A - это расстояние от записи до центра ее кластера, а B - расстояние от записи до центра ближайшего кластера, к которому она не принадлежит. Силуэтный коэффициент, равный 1, означал бы, что все наблюдения расположены точно в центрах их кластеров. Значение –1 означало бы, что все наблюдения расположены в центрах некоторого другого кластера. Значение 0 означает, что наблюдения расположены в среднем на равных расстояниях от центра их кластера и центра ближайшего кластера.

Сводка включает таблицу, которая содержит следующую информацию:

- **Алгоритм.** Используемый алгоритм кластеризации, например, "Двухэтапный".
- **Исходные показатели.** Число полей, также называемых **входными** или **предикторами**.
- **Кластеры.** Число кластеров в решении.

Вид представления Кластеры

Представление Кластеры содержит "сетку" кластеров по показателям, которая включает имена кластеров, объемы (размеры) и профили каждого кластера.

Столбцы в сетке содержат следующую информацию:

- **Кластер.** Номера кластеров, созданных в результате работы алгоритма.
- **Метка.** Любые метки, заданные для кластеров (по умолчанию они пустые). Дважды щелкните по ячейке, чтобы ввести метку, описывающую содержимое кластера, например, "Покупатели престижных автомобилей".
- **Описание.** Описание содержимого кластеров (по умолчанию оно пустое). Дважды щелкните по ячейке, чтобы ввести описание кластера, например, "возраст 55+ лет, профессионалы, доход превосходит $100000".
- **Размер.** Размер каждого кластера в виде процента от общего размера выборки, которая использовалась для построения модели кластеризации. В каждой ячейке размера внутри сетки выводится вертикальный столбец, показывающий размер кластера в процентах, размер кластера в процентах в числовом виде и число наблюдений в кластере.
- **Элементы.** Отдельные предикторы, по умолчанию отсортированные по общей важности. Если какие-либо столбцы имеют одинаковые размеры, они выводятся в возрастающем порядке номеров кластеров.

  Общая важность показателей обозначается интенсивностью цвет фона ячейки: наиболее важный показатель является наиболее темным. Легенда над таблицей показывает соответствие между важностью и интенсивностью цвета.

Если поместить указатель мыши на ячейку, то будет выведено полное имя/метка показателя и значение важности для этой ячейки. В зависимости от типа показателя и вида представления может быть выведена дополнительная информация. Для представления Центры кластеров такая информация будет включать статистику ячейки и значение ячейки, например: “Среднее: 4,32”. Для категориальных показателей в ячейке выводится имя наиболее часто встречающейся (модальной) категории и соответствующий ей процент.

Внутри представления Кластеры можно выбрать различные способы вывода информации о кластерах:

- Транспонировать кластеры и показатели. Дополнительную информацию смотрите в разделе “Транспонировать кластеры и показатели”.
- Сортировать показатели. Дополнительную информацию смотрите в разделе “Сортировать показатели”.
- Сортировать кластеры. Дополнительную информацию смотрите в разделе “Сортировать кластеры” на стр. 220.
- Выбрать содержимое ячеек. Дополнительную информацию смотрите в разделе “Содержимое ячеек.” на стр. 220.

Транспонировать кластеры и показатели: По умолчанию, кластеры выводятся как столбцы, а показатели выводятся как строки. Чтобы поменять местами строки и столбцы в выводе, щелкните по кнопке **Транспонировать кластеры и показатели**, расположенной слева от кнопки **Сортировать показатели по**. Например, это можно сделать, чтобы реже пользоваться горизонтальной прокруткой при просмотре данных, когда выведено много кластеров.

Сортировать показатели: Кнопка **Сортировать показатели по** позволяет выбрать, как выводить ячейки показателей:

- **Общая важность.** Этот порядок сортировки задан по умолчанию. Показатели сортируются в убывающем порядке общей важности, и порядок сортировки один и тот же по всем кластерам. Если какие-либо показатели имеют совпадающие значения важности, то такие показатели перечисляются в возрастающем порядке имен показателей.
- **Важность для кластера.** Показатели сортируются по их важности для каждого кластера. Если какие-либо показатели имеют совпадающие значения важности, то такие показатели перечисляются в возрастающем порядке имен показателей. Если выбран этот вариант, порядок сортировки в кластерах обычно различается.

- **Имя.** Показатели сортируются по именам в алфавитном порядке.
- **Порядок следования в данных.** Показатели сортируются по порядку их расположения в наборе данных.

Сортировать кластеры: По умолчанию кластеры сортируются в убывающем порядке их размеров. Кнопка **Сортировать кластеры по** позволяет сортировать кластеры по именам в алфавитном порядке или, если заданы уникальные метки, в алфавитном порядке меток.

Показатели, которые имеют одну и ту же метку, сортируются по именам кластеров. Если кластеры отсортированы по метками и метки редактируются, то порядок сортировки автоматически меняется.

Содержимое ячеек.: Кнопки **Ячейки** позволяют изменить вывод содержимого ячеек для показателей и полей оценивания.

- **Центры кластеров.** По умолчанию ячейки выводят имена/метки показателей и показатель положения центра распределения для каждой комбинации кластера и показателя. Для непрерывных полей показывается среднее значение, а для категориальных полей - мода (категория, которая встречается наиболее часто) вместе с процентами по категориям.
- **Абсолютные распределения.** Показываются имена/метки показателей и абсолютные распределения показателей внутри каждого кластера. Для категориальных показателей в выводе показываются столбчатые диаграммы для категорий, упорядоченных по возрастанию значений данных. Для непрерывных полей в выводе показывается диаграмма сглаженной плотности, в которой используются конечные точки и интервалы, одинаковые для всех кластеров.

  Вывод, окрашенный в насыщенный красный цвет, показывает распределение для кластеров, тогда как бледный вывод представляет полные данные.
- **Относительные распределения.** Показываются имена/метки показателей и относительные распределения в ячейках. Вообще эти выводы подобны тем, в которых показываются абсолютные распределения, за исключением того, что на них выводятся относительные распределения.

  Вывод, окрашенный в насыщенный красный цвет, показывает распределение для кластеров, тогда как бледный вывод представляет полные данные.
- **Базовое представление.** Когда имеется много кластеров, бывает трудно увидеть все детали, не используя прокрутку. Чтобы снизить потребность в использовании прокрутки, выберите этот вид представления для вывода таблицы в более компактном виде.

Вид представления Важность предикторов в кластерах

Представление Важность предикторов показывает относительную важность каждого поля при оценивании модели.

Вид представления Размеры кластеров

Представление Размеры кластеров показывает круговую диаграмму, содержащую все кластеры. В каждом секторе показывается относительный размер каждого кластера в процентах. Поместите указатель мыши на сектор, чтобы вывести количество в этом секторе.

Ниже этой диаграммы расположена таблица, выводящая следующую информацию о размерах:

- Размер наименьшего кластера (как количество и как процент от целого).
- Размер наибольшего кластера (как количество и как процент от целого).
- Отношение размера наибольшего кластера к размеру наименьшего кластера.

Вид представления Распределение в ячейке

Представление Распределение в ячейке выводит расширенную, более детальную диаграмму распределения данных для любой ячейки показателя, выбранной в таблице в представлении Кластеры в основной панели.

Вид представления Сравнение кластеров

Представление Сравнение кластеров имеет форму сетки с показателями в строках и выбранными кластерами в столбцах. Этот вид представления помогает лучше понять, какие факторы формируют кластер. Он также позволяет увидеть различие между кластерами, не только в сравнении со всеми данными, но и в сравнении между собой.

Чтобы выбрать кластеры для вывода, щелкните по верху столбца кластера в основной панели в представлении Кластеры. Пользуйтесь клавишами Ctrl и Shift совместно с щелчком мышью для выбора или отмены выбора нескольких кластеров для сравнения.

Примечание: Можно выбрать для вывода до пяти кластеров.

Кластеры выводятся в том порядке, в котором они были выбраны, тогда как порядок полей определяется параметром **Сортировать показатели по**. При выборе **по важности для кластера** поля всегда сортируются по общей важности.

Диаграммы на заднем плане показывают общие распределения каждого показателя:

- Категориальные показатели выводятся в виде точечных диаграмм, где для указания наиболее часто встречающейся (модальной) категории в каждом кластере (по показателям) используется размер точки.
- Непрерывные показатели выводятся в виде ящичных диаграмм с усами, которые показывают общие медианы и межквартильные диапазоны.

На эти изображения заднего плана накладываются ящичные диаграммы с усами для выбранных кластеров:

- Для непрерывных показателей квадратные точечные маркеры и горизонтальные линии показывают медиану и межквартильный диапазон для каждого кластера.
- Каждый кластер представляется своим цветом, показанным в верхней части изображения.

Перемещение по средству просмотра кластеров

Средство просмотра кластеров представляет собой интерактивный вывод. Варианты:

- Выбрать поле или кластер, чтобы увидеть больше деталей.
- Сравнить кластеры, чтобы выбрать элементы, представляющие интерес.
- Видоизменить вывод.
- Транспонировать оси.
- Создать производные узлы, а также узлы фильтра и выбора, пользуясь меню Создать.

Использование панели инструментов.

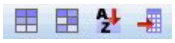
С помощью панели инструментов можно управлять выводом информации на левой и правой панелях. Пользуясь элементами управления панели инструментов, можно изменять ориентацию вывода (сверху вниз, слева направо или справа налево). Кроме того, параметрам средства просмотра можно возвратить значения, установленные по умолчанию, и открыть диалоговое окно, чтобы задать содержимое представления Кластеры в основной панели.

Возможность выбрать **Сортировать показатели по**, **Сортировать кластеры по**, **Ячейки** и **Показать** появляется, только если выбрать представление **Кластеры** в основной панели. Дополнительную информацию смотрите в разделе “Вид представления Кластеры” на стр. 219.

Таблица 12. Значки панели инструментов.

Значок	Тема
	Смотрите Транспонировать кластеры и показатели

Таблица 12. Значки панели инструментов (продолжение).

Значок	Тема
	Смотрите опцию Сортировать показатели по
	Смотрите опцию Сортировать кластеры по
	Смотрите опцию Ячейки

Формирование узлов из моделей кластеров

Меню Создать позволяет сформировать новые узлы на основе модели кластеров. Такая возможность предоставляется на вкладке Модель для построенной модели и позволяет создать узлы, основываясь либо на текущем выводе, либо на отборе (т.е. все видимые кластеры или все выбранные кластеры). Например, можно выбрать один показатель и затем создать узел фильтра, чтобы отбросить все остальные (невидимые) показатели. Созданные узлы помещаются на панель холста диаграммы несоединенными. Дополнительно можно копию слепка модели на палитре моделей. Не забудьте соединить узлы и сделать любые нужные правки перед запуском.

- **Сформировать узел моделирования.** Формирует узел моделирования на холсте потока. Это может пригодиться, например, если имеется поток, в котором вы хотите использовать эти параметры модели, но больше не имеете узла моделирования для их формирования.
- **Создать модель на Палитре.** Создает слепок на палитре моделей. Это полезно в ситуациях, когда коллега послал вам поток, содержащий модель, но не саму модель.
- **Узел фильтра.** Создает новый узел фильтра, чтобы отфильтровать поля, которые не используются моделью кластеров, и/или невидимы в текущем выводе Средства просмотра кластеров. Если имеется узел типа, расположенный выше данного узла кластера, то любые поля, имеющие роль *Целевая*, отбрасываются созданным узлом фильтра.
- **Узел фильтра (из выделенного).** Создает новый узел фильтра, чтобы отфильтровать поля, основываясь на выборе, сделанном в средстве просмотра кластеров. Щелчком мыши с нажатием клавиши Ctrl выберите несколько полей. Поля, выбранные в средстве просмотра кластеров, отбрасываются по направлению вниз, однако такое поведение можно изменить, редактируя узел фильтра перед запуском.
- **Узел выбора.** Создает новый узел выбора для выбора записей, основываясь на их принадлежности к любому из кластеров, видимых в текущем выводе средства просмотра кластеров. Условие выбора генерируется автоматически.
- **Узел выбора (из выделенного).** Создает новый узел выбора для выбора записей, основываясь на их принадлежности к кластерам, выбранным в средстве просмотра кластеров. Щелчком мыши с нажатием клавиши Ctrl выберите несколько кластеров.
- **Узел извлечения.** Создает новый узел извлечения, который порождает поле признаков, которое присваивает записям значение *Истина* или *Ложь*, основываясь на их принадлежности ко всем кластерам, видимым в средстве просмотра кластеров. Условие извлечения генерируется автоматически.
- **Узел извлечения (из выделенного).** Создает новый узел извлечения, который порождает поле признаков, основываясь на их принадлежности к кластерам, выбранным в средстве просмотра кластеров. Щелчком мыши с нажатием клавиши Ctrl выберите несколько кластеров.

В дополнение к созданию узлов также можно создавать диаграммы, пользуясь меню Создать. Дополнительную информацию смотрите в разделе “Построение диаграмм на основе моделей кластеров” на стр. 223.

Управление выводом для представления Кластеры

Чтобы получить доступ к управлению тем, что показано в представлении Кластеры в основной панели, нажмите кнопку **Показать**. Откроется диалоговое окно Показать.

Характеристики. Выбрано по умолчанию. Чтобы скрыть все входные показатели, снимите этот переключатель.

Поля для оценки. Выберите поля для оценки (поля, которые не используются для создания модели кластеров, но посылаются в средство просмотра моделей, чтобы оценить качество кластеров), которые будут выведены. По умолчанию ни одно не выводится. *Примечание*: Этот переключатель недоступен, если нет ни одного поля для оценки.

Описания кластеров. Выбрано по умолчанию. Чтобы скрыть все ячейки описания кластеров, снимите этот переключатель.

Размеры кластеров. Выбрано по умолчанию. Чтобы скрыть все ячейки размеров кластеров, снимите этот переключатель.

Максимальное число категорий. Задайте максимальное число категорий для вывода на диаграммах категориальных показателей. Значение по умолчанию равно 20.

Построение диаграмм на основе моделей кластеров

Модели кластеров предоставляют много информации, однако она не всегда имеет формат, который легко понятен пользователям из бизнеса. Чтобы представить данные в виде, в котором они легко могут быть вставлены в деловые отчеты, презентации и т.д., можно построить диаграммы для выбранных данных. Например, пользуясь средством просмотра кластеров, можно построить диаграмму для выбранного кластера, таким образом, создавая ее только для наблюдений в этом кластере.

Примечание: Создать диаграмму из средства просмотра кластеров можно только в том случае, когда слепок модели присоединен к другим узлам в потоке.

Постройте диаграмму

1. Откройте nugget модели, содержащий Средство просмотра кластеров.
2. На вкладке Модель выберите *Кластеры* в выпадающем списке **Представление**.
3. На основной панели выберите кластер или кластеры, для которых нужно построить диаграмму.
4. В меню Создать выберите **Создать диаграмму (на основе выделенного)**. Будет показана вкладка Тип диалогового окна Панель выбора диаграмм.

   Примечание: при выводе панели диаграмм таким способом возможны только вкладки Основная и Подробная.
5. Используя параметры вкладки Тип и Детали, задайте элементы для вывода на диаграмме.
6. Нажмите кнопку OK, чтобы построить диаграмму.

Заголовок диаграммы идентифицирует тип модели, а также кластер или кластеры, которые были выбраны для включения.

Глава 12. Правила связывания

Правила связывания связывают конкретное следствие (например, покупку некоторого товара) с набором условий (например, с покупкой нескольких других товаров). Например, правило

```
пиво <= консервы & замороженные_блюда (173, 17.0%, 0.84)
```

означает, что *пиво* часто встречается, если *консервы* и *замороженные_блюда* присутствуют вместе. Это правило достоверно на 84% и применимо к 17% данных, или к 173 записям. Алгоритмы правил связывания автоматически находят связи, которые вы могли бы найти и сами, используя визуализацию, такую как узел Web.

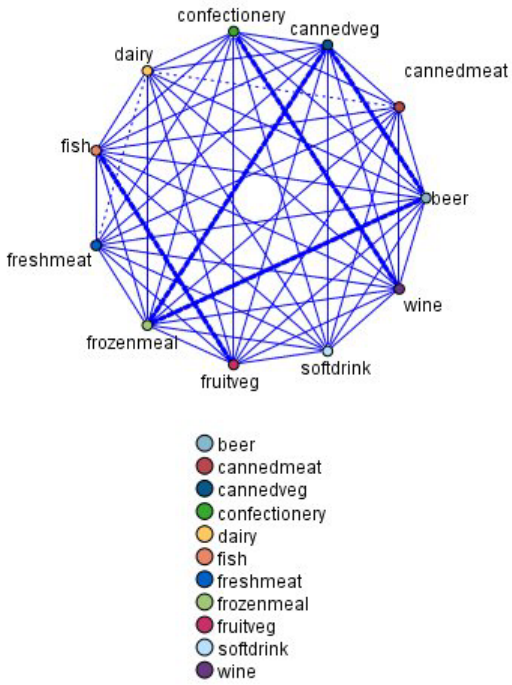


Рисунок 45. Узел Web, показывающий связи между товарами покупательской корзины

Преимущество алгоритмов правил связывания перед более стандартными алгоритмами дерева решений (C5.0 и деревья C&R Trees) состоит в том, что связывание можно произвести между *любыми* из атрибутов. Алгоритм дерева решений построит правила с возможностью одного следствия, в то время как алгоритмы связывания стараются найти несколько правил, у каждого из которых может быть отдельное следствие.

Недостаток алгоритмов связывания состоит в том, что они пытаются найти структуры в потенциально очень больших областях поиска и тем самым могут потребовать гораздо больше времени на выполнение, чем алгоритм дерева решений. Для нахождения правил эти алгоритмы используют подход **создать и проверить**, то есть сначала создаются простые правила, и они проверяются на наборе данных. Хорошие правила сохраняются, а все правила, для которых существуют некоторые ограничения, специализируются. **Специализация** - это процесс добавления в правило условий. Такие новые правила затем проверяются на данных, и в этом процессе итерационно сохраняются лучшие из наиболее интересных найденных правил. Обычно пользователь накладывает ограничение на возможное число разрешенных в правиле антецедентов, а для сокращения потенциально большого пространства поиска используются различные приемы, основанные на теории информации или эффективных схемах индексирования.

В конце обработки выводится таблица лучших правил. В отличие от дерева решений, этот набор правил нельзя использовать непосредственно для создания предсказаний, как это делается в стандартной модели (например, при использовании дерева решений или нейросети). Это связано с наличием большого числа различных возможных следствий для правил. Для преобразования правил связывания в набор правил классификации требуется дополнительный уровень преобразования. Поэтому правила связывания, созданные алгоритмами связывания, называют **неуточненными моделями**. Хотя пользователь может просматривать эти неуточненные модели, их нельзя напрямую использовать как модели классификации, пока пользователь не укажет системе сгенерировать модель классификации из неуточненной модели. Это делается в браузере через опцию меню Генерировать.

Поддерживается два алгоритма правил связывания:

Узел Априори извлекает набор правил из данных, выделяя правила с наибольшим информационным содержимым. Узел Априори предлагает пять различных способов выбора правил и использует сложные схемы индексирования для эффективной обработки больших наборов данных. Для больших задач узел Априори обычно быстрее при обучении; у него нет произвольного ограничения количества правил, которые можно сохранить, и он может обрабатывать правила с количеством предварительных условий до 32. Для узла Априори требуются категориальные входные и выходные поля, он был оптимизирован для полей такого типа и показывает с ними высокую производительность.

Узел последовательности обнаруживает правила связывания для последовательных или зависящих от времени данных. Последовательность - это список наборов элементов с тенденцией появления в предсказуемом порядке. Например, покупатель, который приобрел лезвия и лосьон после бритья, с большой вероятностью в следующий раз купит крем для бритья. Узел последовательности основан на алгоритме правил связывания CARMA, использующем эффективный двухпроходный способ обнаружения последовательностей.

Сравнение табличных и транзакционных данных

Используемые моделями правил связывания данные могут быть в транзакционном или табличном формате, как описано ниже. Это общие описания; конкретные требования могут различаться, как описывается в документации для каждого типа моделей. Обратите внимание на то, что при скоринге моделей данные для оценки должны отображать формат данных, используемых при построении модели. Модели, построенные с использованием табличных данных, можно использовать для оценки только табличных данных; модели, построенные для транзакционных данных, могут оценивать только транзакционные данные.

Транзакционный формат

У транзакционных данных есть отдельная запись для каждой транзакции или элемента. Например, если покупатель делает несколько покупок, у каждой должна быть отдельная запись с соответствующими элементами, связанными ID покупателя. Иногда в этом случае говорят о формате **данных с кассовой ленты**.

Покупатель	Покупка
1	джем
2	молоко
3	джем
3	хлеб
4	джем
4	хлеб
4	молоко

Узлы Априори, CARMA и Последовательность могут использовать транзакционные данные.

Табличные данные

У табличных данных (известных также как данные **корзины покупок** или **таблицы истинности**) есть элементы, представленные отдельными флагами, где каждый флаг обозначает наличие или отсутствие конкретного элемента. Каждая запись представляет полный набор связанных элементов. Поля флагов могут быть категориальными или числовыми, хотя у некоторых моделей могут быть конкретные требования.

Покупатель	Джем	Хлеб	Молоко
1	Т	Ж	Ж
2	Ж	Ж	Т
3	Т	Т	Ж
4	Т	Т	Т

Узлы Априори, CARMA и Последовательность могут использовать табличные данные.

Узел Априори

Узел Априори также находит в данных правила связи. Узел Априори использует пять различных методов выбора правил и сложную схему индексации для эффективной обработки больших наборов данных.

Требования. Чтобы создать набор правил Априори, требуется одно или несколько *входных* полей и одно или несколько полей *назначения*. Входные и выходные поля (с ролями *Входное*, *Назначение* или *Обе*) должны быть символическими. Поля с ролью *Нет* игнорируются. Перед выполнением узла типы полей должны быть инстанциированы. Данные могут быть в табличном или транзакционном формате. Дополнительную информацию смотрите в разделе “Сравнение табличных и транзакционных данных” на стр. 226.

Достоинства. Обычно при решении больших задач обучение узла Априори происходит быстрее. В нем нет произвольного ограничения на число сохраняемых правил и возможна обработка правил с количеством предварительных условий до 32. Узел Априори использует пять различных обучающих методов, обеспечивая большую гибкость в подборе метода анализа данных к исследуемой проблеме.

Дополнительные опции узлов Априори

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Минимальная поддержка антецедента. Можно задать критерий поддержки для выполнения правил в наборе правил. **Поддержка** указывает процент записей в обучающих данных, для которых антецеденты (часть правила, содержащая условие) истинны. (Обратите внимание на то, что это определение поддержки отличается от используемого в узлах последовательности и CARMA. Дополнительную информацию смотрите в разделе “Опции моделей узла Последовательность” на стр. 242.) Если вы используете правила, которые применяются к очень маленьким наборам данных, попробуйте увеличить значение этого параметра.

Примечание: Определение поддержки для узлов Априори основано на числе записей с антецедентами. Это отличает его от алгоритмов CARMA и последовательности, для которых определение поддержки основано на числе записей, содержащих все элементы правила (то есть и антецеденты, и консеквенты). Результаты для моделей связи содержат показатели поддержки как антецедентов, так и правил.

Минимальная достоверность правила. Вы можете также задать критерий достоверности. **Достоверность** основана на записях, для которых антецеденты правила истинны, и представляет собой процент записей, для которых консексвенты также истинны. Другими словами, это процент верных прогнозов, сделанных на основе данного правила. Правила, у которых достоверность ниже задаваемой критерием, отбрасываются. Если у вас получается слишком много правил, попробуйте увеличить значение этого параметра. Если правил получается слишком мало или не получается совсем, попробуйте уменьшить значение этого параметра.

Максимальное количество антецедентов. Вы можете задать максимальное число предварительных условий для любого правила. Это способ ограничения сложности правил. Если правила слишком сложные или слишком специальные, попробуйте уменьшить значение этого параметра. Этот параметр также оказывает большое влияние на время обучения. Если набор правил требует слишком много времени для обучения, попробуйте уменьшить значение этого параметра.

Только значения true для флагов. Если эта опция выбрана для данных в табличном (таблица истинности) формате, в полученные в итоге правила будут включены только истинные значения. Это позволяет сделать правила более понятными. Эта опция неприменима к данным в транзакционном формате. Дополнительную информацию смотрите в разделе “Сравнение табличных и транзакционных данных” на стр. 226.

Оптимизировать. Выберите во время построения моделей предназначенные для повышения производительности опции в соответствии с конкретными потребностями.

- Выберите опцию **Скорость**, чтобы для повышения производительности алгоритм никогда не использовал сброс на диск.
- Выберите опцию **Память**, чтобы в подходящих ситуациях алгоритм использовал сброс на диск, хотя это понизит скорость. Эта опция выбирается по умолчанию. *Примечание*: При запуске в распределенном режиме этот параметр может быть перезаписан значением опции администратора из файла *options.cfg*. Дополнительную информацию смотрите в разделе *сервер IBM SPSS ModelerРуководство администратора*.

Дополнительные опции узлов Априори

Для тех, кто хорошо знаком с работой узлов Априори, следующие дополнительные опции позволяют выполнять тонкую настройку процесса вывода правил. Для доступа к дополнительным опциям задайте на вкладке Дополнительно режим **Дополнительно**.

Показатель оценки. Узлы Априори поддерживает пять методов оценки потенциальных правил.

- **Достоверность правила.** Для оценки правил метод по умолчанию использует показатель достоверности (или точность) правила. Для этого показателя опция **Нижняя граница показателя оценки** отключена, поскольку при использовании вместе с опцией **Минимальная достоверность правила** на вкладке Модель она является избыточной. Дополнительную информацию смотрите в разделе “Дополнительные опции узлов Априори” на стр. 227.
- **Разность достоверности.** (Другое название - **абсолютное отличие по достоверности от предыдущего**.) Этот показатель оценки - абсолютная разность между показателем достоверности для правила и его априорной достоверностью. Эта опция устраняет смещение в случаях, когда результаты распределены неравномерно. Это препятствует сохранению "очевидных" правил. Например, может оказаться, что 80% клиентов покупают ваш самый популярный продукт. Правило, предсказывающее покупку этого популярного продукта с точностью 85%, дает вам не так много дополнительной информации, даже если считать 85%-ую точность хорошим результатом в абсолютном масштабе. Задайте нижнюю границу показателя оценки равной минимальной разнице в достоверности, для которой должны выполняться правила.
- **Отношение достоверности.** (Другое название - **отличие показателя достоверности от 1**.) Этот показатель оценки представляет собой отношение достоверности правила к априорной достоверности (или, если это отношение больше 1, из 1 вычитается величина, обратная ему). Как и разность достоверности, этот метод учитывает неравномерность распределения. Это особенно полезно при поиске правил, предсказывающих редкие события. Например, предположим, что есть редкое заболевание, которое проявляется только у 1% пациентов. Правило, которое позволяет предсказать это заболевание в 10% случаев, дает большое преимущество по сравнению со случайным угадыванием, хотя в абсолютном исчислении 10%-ная вероятность не кажется впечатляющей. Задайте нижнюю границу показателя оценки равной разнице, для которой должны выполняться правила.
- **Разность информации.** (Другое название - **отличие информации от априорной**.) Этот показатель основан на показателе **приобретенной информации**. Если вероятность определенного исхода рассматривается как логическое значение (**бит**), то приобретенная информация - это доля этого бита, поддающаяся определению на основе антецедентов. Разность информации - это различие между приобретенной информацией с учетом антецедентов и приобретенной информацией с учетом только априорной доверительной вероятности исхода. Важная особенность этого метода - учет поддержки, то есть правила,

касающиеся большего количества записей, получают предпочтение для данного уровня доверительной вероятности. Задайте нижнюю границу показателя оценки равной разности информации, для которой должны выполняться правила.

Примечание: Поскольку масштаб для этого показателя определяется несколько менее интуитивно, чем другие масштабы, для получения удовлетворительного набора правил вам, возможно, придется поэкспериментировать с различными верхними и нижними границами.

- **Нормализованный хи-квадрат.** (Другое название - **нормализованный показатель хи-квадрат**.) Это статистический показатель связи между предшествующими событиями и их последствиями. Показатель нормализован и принимает значения от 0 до 1. Он еще в большей степени зависит от поддержки, чем показатель информационной разности. Задайте нижнюю границу показателя оценки равной разности информации, для которой должны выполняться правила.

  Примечание: Как и в случае информационной разности, масштаб для этого показателя определяется несколько менее интуитивно, чем другие масштабы, поэтому для получения удовлетворительного набора правил вам, возможно, придется поэкспериментировать с различными верхними и нижними границами.

Разрешить правила без антецедентов. Выберите эту опцию, чтобы разрешить правила, которые включают только консеквент (элемент или набор элементов). Это полезно, когда вам нужно определить общие элементы или наборы элементов. Например, `cannedveg` - это правило одного элемента без антецедента, указывающее что покупка *cannedveg* часто встречается в данных. В некоторых случаях вы захотите включить такие правила просто для информации о наиболее достоверных предсказаниях. По умолчанию эта опция отключена. По договоренности поддержка антецедентов для правил без антецедентов составляет 100%, а поддержка правил совпадает с доверительной вероятностью.

Узел CARMA

Для обнаружения правил связывания в данных узел CARMA использует алгоритм обнаружения правил связывания. Правила связывания - это операторы следующего вида:

if *антецеденты* **then** *консеквенты*

Например, если покупатель в интернет-магазине приобретает карту беспроводной связи и высококлассный беспроводной маршрутизатор, он скорее всего купит и беспроводной музыкальный сервер, если его предложат. Модель CARMA извлекает из данных набор правил, не требуя, чтобы вы задавали входные или выходные поля. Это означает, что сгенерированные правила могут использоваться для более широкого набора прикладных программ. Например, можно использовать сгенерированные этим узлом правила, чтобы найти список продуктов или услуг (антецедентов), консеквент которых - это товар, который вы хотите продвигать в этом летнем сезоне. Используя IBM SPSS Modeler, вы можете определить, какие клиенты приобрели товары-антецеденты и построить маркетинговую компанию для продвижения товара-консеквента.

Требования. В отличие от узла Априори, узел CARMA не требует *Входных* полей или полей *Назначения*. Это неотъемлемая часть способа работы алгоритма, которая эквивалентна построению модели Априори, когда для всех полей задано значение *Оба*. Применяя фильтры к модели после ее построения, можно ограничить элементы, которые будут перечисляться только как антецеденты или консеквенты. Например, можно использовать браузер модели, чтобы найти список продуктов или услуг (антецедентов), консеквент которых - это товар, который вы хотите продвигать в этом летнем сезоне.

Чтобы создать набор правил CARMA, необходимо задать поле ID и одно или несколько полей содержимого. У поля ID может быть любая роль или уровень измерения. Поля с ролью *Нет* игнорируются. До выполнения узла типы полей должны быть полностью определены. Как и на узле Априори, данные могут быть в табличном или в транзакционном формате. Дополнительную информацию смотрите в разделе “Сравнение табличных и транзакционных данных” на стр. 226.

Достоинства. Узел CARMA основывается на алгоритме правил связывания CARMA. В отличие от узла Априори, узел CARMA предлагает параметры сборки для поддержки правил (поддержка относится и к антецедентам, и к консеквентам), а не только для поддержки антецедентов. CARMA допускает также

правила с несколькими консеквентами. Как и для узла Априори, сгенерированные узлом CARMA модели можно вставить в поток данных для создания предсказаний. Дополнительную информацию смотрите в разделе “Слепки моделей” на стр. 36.

Опции полей узла CARMA

Перед выполнением узла CARMA необходимо задать входные поля на вкладке Поля узла CARMA. На большинстве узлов моделирования опции вкладки Поля одинаковы, но узел CARMA содержит и некоторые уникальные опции. Все эти опции обсуждаются далее.

Использовать параметры узла Тип. Эта опция указывает узлу, что следует использовать информацию о полях из узла Тип, расположенного выше. Это опция по умолчанию.

Использовать пользовательские параметры. Эта опция указывает узлу, что следует использовать информацию о полях, заданную здесь, а не ту, что задана на расположенных выше узлах Тип. После выбора этой опции задайте поля ниже в соответствии с тем, в каком формате читаются данные, в транзакционном или в табличном.

Использовать транзакционный формат. Эта опция изменяет управляющие элементы полей на остальной части диалогового окна в зависимости от формата данных - транзакционного или табличного. Если вы используете несколько полей с транзакционными данными, предполагается, что заданные в этих полях элементы для конкретной записи принадлежат одной транзакции с единой отметкой времени. Дополнительную информацию смотрите в разделе “Сравнение табличных и транзакционных данных” на стр. 226.

Табличные данные

Если опция **Использовать транзакционный формат** не выбрана, выводятся следующие поля.

- **Поля ввода.** Выберите одно или несколько входных полей. Это аналогично заданию для поля роли *Входное* на узле Тип.
- **Подмножества.** Это поле позволяет позволяет задать поле, с помощью которого данные будут разделяться на отдельные выборки для этапов обучения, испытания и проверки при построении модели. Сгенерировав модель при помощи одной выборки и испытав ее при помощи другой, можно получить надежный показатель качества обобщения модели на более крупные наборы данных, подобные текущим. Если при помощи узлов Тип или Раздел были определено несколько полей разделов, на вкладке Поля на каждом узле моделирования, где используется разделение, должно быть выбрано одно поле раздела. (Если представлен только один раздел, он будет использоваться автоматически при всяком включении разделения.) Кроме того, учтите, что для применения выбранного раздела в анализе на вкладке Параметры модели для узла должно быть также включено разделение. (Выключение этой опции делает возможным отключение разделения без изменения значений параметров полей.)

Транзакционные данные

Если выбрана опция **Использовать транзакционный формат**, выводятся следующие поля.

- **ID.** Для транзакционных данных выберите из списка поле ID. Значения в поле ID могут быть числовыми или символическими. Каждое уникальное значение в этом поле должно обозначать конкретный объект анализа. Например, в прикладной программе Корзина покупок каждый ID может представлять одного покупателя. В прикладной программе Анализ Web-журнала каждый ID может представлять отдельный компьютер (по IP-адресу) или одного пользователя (по регистрационным данным).
- **Последовательные ID.** (только для узлов Априори и CARMA) Если ваши данные предварительно отсортированы, так что все записи с одинаковым ID сгруппированы совместно в потоке данных, выберите эту опцию для ускорения обработки. Если ваши данные предварительно не отсортированы (или вы в этом не уверены), оставьте эту опцию выключенной, и узел отсортирует данные автоматически.

  Примечание: Если данные не отсортированы, но выбирается эта опция, можно получить недопустимые результаты для вашей модели.

- **Содержимое.** Задайте поле или поля содержимого для модели. Эти поля содержат нужные элементы при моделировании связывания. Можно задать несколько полей флагов (если данные в табличном формате) или одно номинальное поле (если данные в транзакционном формате).

Опции моделей узла CARMA

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Минимальная поддержка правила (%). Вы можете задать критерий поддержки. **Поддержка правил** относится к тем ID в обучающих данных, которые содержат полное правило. (Обратите внимание на то, что это определение поддержки отличается от поддержки антецедентов, используемой на узлах Априори). Если вы хотите сосредоточиться на более общих правилах, увеличьте значение этого параметра.

Минимальная достоверность правила (%). Для удержания отдельных правил в наборе правил можно задать критерий достоверности. **Достоверность** указывает процентную долю ID, для которых сделаны правильные предсказания (среди всех ID, для которых делаются предсказания с использованием данного правила). Она вычисляется на основе обучающих данных как отношение числа ID, для которых найдено полное правило, к числу ID, для которых найдены антецеденты. Правила, у которых достоверность ниже задаваемой критерием, отбрасываются. Если вы получаете не интересующие вас или многочисленные правила, попробуйте увеличить этот параметр. Если вы получаете слишком мало правил, попробуйте уменьшить этот параметр.

Максимальный размер правила. В правиле вы можете задать максимальное число отдельных *наборов элементов* (в противоположность просто *элементам*). Если интересующие вас правила относительно короткие, можно уменьшить этот параметр, чтобы ускорить построение набора правил.

Дополнительные опции узла CARMA

Пользователям, подробно знакомым с работой узла CARMA, следующие дополнительные опции помогут точнее настроить процесс построения моделей. Для доступа к опциям эксперта на вкладке Дополнительно задайте для режима значение **Расширенный**.

Исключить правила с несколькими консеквентами. Выберите опцию исключения консеквентов с двумя следствиями, то есть консеквентов, содержащих два элемента. Например, правило `хлеб & сыр & рыба -> вино&фрукты` содержит консеквент с двумя следствиями `вино&фрукты`. По умолчанию такие правила включаются.

Задать значение усечения. Для экономии памяти используемый алгоритм CARMA при обработке периодически удаляет (**отсекает**) редкие наборы данных из своего списка потенциальных наборов данных. Выберите эту опцию для настройки частоты усечения, которая будет определяться задаваемым вами значением. Введите меньшее значение для снижения требований алгоритма к памяти (хотя потенциально это увеличивает нужное время обучения), или большее значение для ускорения обучения (но потенциально это повысит требования алгоритма к объему памяти). Значение по умолчанию - 500.

Изменение поддержки. Выберите эту опцию для повышения эффективности путем исключения нечастых наборов данных, которые могут показаться частыми при неравномерном включении. Это достигается запуском алгоритма с более высоким уровнем поддержки и последующим понижением его до уровня, заданного на вкладке Модель. Введите значение для параметра **Предполагаемое число транзакций**, чтобы задать, как быстро должен понижаться уровень поддержки.

Разрешить правила без антецедентов. Выберите эту опцию, чтобы разрешить правила, которые включают только консеквент (элемент или набор элементов). Это полезно, когда вам нужно определить общие элементы или наборы элементов. Например, `cannedveg` - это правило одного элемента без антецедента, указывающее что покупка *cannedveg* часто встречается в данных. В некоторых случаях вы захотите включить такие правила просто для информации о наиболее достоверных предсказаниях. По умолчанию эта опция выключена.

Слепки моделей правил связывания

Слепки моделей правил представляют правила, обнаруженные одним из следующих узлов моделирования правил связывания:

- Априорный анализ
- CARMA

Слепки модели содержат информацию о правилах, извлеченных из данных при построении модели.

Просмотр результатов

Можно просмотреть правила, сгенерированные моделями связывания (Apriori и CARMA) и моделями последовательностей с помощью вкладки Модель в диалоговом окне. При просмотре слепка модели вы можете вывести информацию о правилах и использовать возможности для фильтрации и сортировки результатов перед тем, как генерировать новые узлы или оценивать модель.

Скоринг модели

Можно добавлять в поток слепки уточненных моделей (Apriori, CARMA, последовательной модели) и использовать их для оценки. Дополнительную информацию смотрите в разделе “Использование слепков моделей в потоках” на стр. 47. Слепки моделей, используемые для скоринга, включают в себя дополнительную вкладку Настройки в соответствующих диалоговых окнах. Дополнительную информацию смотрите в разделе “Параметры слепков моделей правил связывания” на стр. 235.

Слепок неуточненной модели в своем исходном формате не может использоваться для скоринга. Вместо этого можно сгенерировать набор правил и уже его использовать для скоринга. Дополнительную информацию смотрите в разделе “Генерирование набора правил из слепка модели связывания” на стр. 237.

Подробности слепков моделей правил связывания

На вкладке Модель слепка модели правил связывания находится таблица с правилами, извлеченными алгоритмом. Каждая строка таблицы представляет одно правило. Первый столбец содержит консеквенты (часть правила со следствием), а следующий столбец содержит антецеденты (часть правила с условием). Последующие столбцы содержат информацию о правиле, например, показатель достоверности, поддержка и рост.

Правила связывания часто выводятся в формате, указанном в следующей таблице.

Таблица 13. Пример правила связывания

Консеквент	Антецедент
Лекарство = drugY	Пол = F BP = HIGH

Правило примера интерпретируется как *если пол = "Ж", а ПС = "ВЫСОКАЯ", то лекарство, скорее всего, будет drugY* или, другими словами, *для записей, где пол = "Ж", а ПС = "ВЫСОКАЯ", лекарство, скорее всего, будет drugY*. С помощью панели инструментов диалогового окна можно вывести на экран дополнительную информацию, например, о показателе достоверности, поддержке и экземплярах.

Меню Сортировка. Кнопка меню Сортировка на панели инструментов управляет сортировкой правил. Направление сортировки (по возрастанию или убыванию) можно изменить с помощью кнопки направления сортировки (стрелка вверх или вниз).

Для сортировки правил может использоваться:

- Поддержка

- Показатель доверия
- Поддержка правила
- Консеквент
- Рост
- Возможность внедрения

Меню Показать/Скрыть. Меню Показать/Скрыть (кнопка на панели инструментов критериев) управляет опциями вывода правил.

Рисунок 46. Кнопка Показать/Скрыть

Доступны следующие опции вывода:

- **ID правила** выводит ID правила, назначенный при построении модели. ID правила позволяет определять, какие правила применяются для данного прогноза. ID правил также позволяют добавлять в дальнейшем новую информацию о правилах, например, о возможности внедрения, продукте или антецедентах.
- **Экземпляры** выводит информацию о числе уникальных идентификаторов, к которым применяется правило, то есть тех, для которых антецеденты истинны. Например, если правило - хлеб -> сыр, то записи в обучающих данных, содержащих антецедент *хлеб*, будут именоваться **экземплярами**.
- **Поддержка** выводит поддержку антецедентов, то есть долю тех ID, для которых, согласно обучающим данным, антецеденты истинны. Например, если 50% обучающих данных включают приобретение хлеба, правило хлеб -> сыр будет иметь поддержку антецедента 50%. *Примечание*: Поддержка определяется здесь так же, как Экземпляры, но выражена в процентах.
- **Достоверность** выводит отношение поддержки правила к поддержке антецедента. Это доля ID с одним или несколькими заданными антецедентами, для которых консеквенты также истинны. Например, если в 50% обучающих данных упоминается хлеб (поддержка антецедента), но только в 20% случаев упоминается хлеб и сыр (поддержка правила), показатель достоверности для правила хлеб -> сыр будет равен Поддержка правила / Поддержка антецедента то есть, в данном случае, 40%.
- **Поддержка правила** выводит долю ID, для которых являются истинными антецеденты, консеквенты и правило в целом. Например, если в 20% обучающих данных упоминаются одновременно хлеб и сыр, поддержка для правила хлеб -> сыр составит 20%.
- **Рост** выводит отношение достоверности для правила к априорной вероятности наличия консеквента. Например, если 10% всех покупателей покупают хлеб, то у правила, предсказывающего, покупку хлеба с 20% вероятностью, будет рост 20/10 = 2. Если другое правило предсказывает покупку хлеба с 11% вероятностью, его рост будет близок к 1, что означает, что наличие антецедентов незначительно влияет на вероятность наличия консеквента. В целом правила с ростом, отличным от 1, будут информативнее правил с ростом, близким к 1.
- **Возможность внедрения** - это процент обучающих данных, удовлетворяющих условиям антецедента, но не удовлетворяющих консеквенту. В терминах покупки товара это по существу означает, какой процент пула покупателей имеет (или приобрел) антецеденты, но еще не приобрел консеквент. Статистика Возможность внедрения определяется как ((Число записей с поддержкой антецедента - Число записей с поддержкой правила) / Число записей) * 100, где *Поддержка антецедента* означает число записей, для которых верны антецеденты, а *Поддержка правила* означает число записей, для которых верны и антецеденты, и консеквент.

Кнопка Фильтр. Кнопка Фильтр (значок с воронкой) в меню раскрывает нижнюю часть диалогового окна, показывая панель, где выводятся активные фильтры правил. Фильтры используются, чтобы сузить набор правил, показанных на вкладке Модели.

Рисунок 47. Кнопка Фильтр

Чтобы создать фильтр, щелкните по значку Фильтр справа от раскрытой панели. Откроется отдельное диалоговое окно, в котором можно задать ограничения для вывода правил. Заметим, что кнопка Фильтр часто используется в сочетании с меню Генерировать, чтобы сначала фильтровать правила и затем генерировать модель, содержащую этот поднабор правил. Дополнительную информацию смотрите в разделе “Задание фильтров для правил” ниже.

Кнопка Найти правило. Кнопка Найти правило (значок с биноклем) позволяет искать правила, показанные для заданного ID правила. В смежном поле вывода указано количество правил, выведенных в данный момент на экран, из общего числа доступных правил. ID правил назначаются моделью в порядке обнаружения правил и добавляются к данным во время скоринга.

Рисунок 48. Кнопка Найти правило

Чтобы изменить порядок ID правил:

1. Изменить порядок ID правил можно в разделе IBM SPSS Modeler, предварительно отсортировав таблицу вывода правил в соответствии с нужным показателем, например, достоверностью или ростом.
2. Затем с помощью опций из меню Генерировать создайте отфильтрованную модель.
3. В диалоговом окне Отфильтрованная модель выберите **Перенумеровать правила последовательно, начиная с** и задайте начальный номер правила.

Дополнительную информацию смотрите в разделе “Генерирование фильтрованной модели” на стр. 237.

Задание фильтров для правил

По умолчанию алгоритмы правил, такие как Априори, CARMA и алгоритмы последовательности, могут генерировать большое число правил, которыми трудно оперировать. Чтобы сделать просмотр правил более ясным, а их скоринг более рациональным, имеет смысл применить правила фильтрации, позволяющие более четко представлять интересующие вас антецеденты и консеквенты. С помощью опций фильтрации на вкладке Модель программы просмотра правил можно открыть диалоговое окно для задания спецификаций фильтра.

Последующие. Выберите **Включить фильтр**, чтобы активировать опции фильтрации правил на основе добавления или исключения заданных консеквентов. Выберите **Содержит любой из**, чтобы создать фильтр, в котором правила содержат хотя бы один из указанных консеквентов. Другой вариант - выберите **Исключает**, чтобы создать фильтр, в котором указанные консеквенты исключены. Можно выбрать консеквенты с помощью значка инструмента выбора справа от поля списка. Откроется диалоговое окно со списком всех консеквентов в сгенерированных правилах.

Примечание: Консеквенты могут содержать несколько элементов. Фильтры проверят только, что консеквент содержит один из указанных элементов.

Противоположные. Выберите **Включить фильтр**, чтобы активировать опции фильтрации правил на основе добавления или исключения заданных антецедентов. Элементы можно выбрать с помощью значка инструмента выбора справа от поля списка. Откроется диалоговое окно со списком всех антецедентов в сгенерированных правилах.

- Выберите **Содержит все** чтобы задать фильтр как включающий, где все заданные антецеденты должны быть включены в правило.

- Выберите **Содержит любой из**, чтобы создать фильтр, в котором правила содержат хотя бы один из указанных антецедентов.
- Выберите **Исключает**, чтобы создать фильтр, исключающий правила, которые содержат заданный антецедент.

Показатель доверия. Выберите **Включить фильтр**, чтобы активировать опции фильтрации правил на основе уровня достоверности для правила. Нужный диапазон достоверности можно задать при помощи элементов управления **Мин** и **Макс**. При просмотре сгенерированных моделей показатель достоверности выводится в виде процентной доли. При скоринге вывода показатель достоверности выводится в виде числа от 0 до 1.

Поддержка антецедентов. Выберите **Включить фильтр**, чтобы активировать опции фильтрации правил на основе уровня поддержки антецедентов для правила. Поддержка антецедентов указывает долю обучающих данных, содержащих те же антецеденты, что и текущее правило, по аналогии с показателем популярности. Нужный диапазон фильтрации правил на основе уровня поддержки можно задать при помощи элементов управления **Мин** и **Макс**.

Рост. Выберите **Включить фильтр**, чтобы активировать опции фильтрации правил на основе показателя роста для правила. *Примечание*: Фильтрация по росту доступна только для моделей связи, построенных позднее Выпуска 8.5, или для более ранних моделей, содержащих показатель роста. Модели последовательностей не содержат этой опции.

Нажмите кнопку **OK**, чтобы применить фильтры, которые включены в этом диалоговом окне.

Построение диаграмм для правил

Узлы дерева содержат много информации, однако у нее не всегда такой формат, который легко понятен пользователям из бизнеса. Чтобы представить данные в виде, в котором они легко могут быть вставлены в деловые отчеты, презентации и т.д., можно построить диаграммы для выбранных данных. На вкладке Модель можно построить диаграмму для выбранного правила, фактически создавая ее только для наблюдений в этом правиле.

1. На вкладке Модель выберите интересующее вас правило.
2. В меню Создать выберите **Создать диаграмму (на основе выделенного)**. Откроется вкладка Основная панель диаграмм.

   Примечание: при выводе панели диаграмм таким способом возможны только вкладки Основная и Подробная.
3. Используя параметры вкладки Тип и Детали, задайте элементы для вывода на диаграмме.
4. Щелкните OK, чтобы построить диаграмму.

В заголовке диаграммы будут указаны подробности антецедентов, выбранные для включения.

Параметры слепков моделей правил связывания

Эта вкладка Параметры используется для задания опций скоринга в моделях связывания (Apriori и CARMA). Эта вкладка доступна только после добавления в поток слепка модели для нужд скоринга.

Примечание: Диалоговое окно для просмотра неуточненной модели не содержит вкладки Параметры, так как для нее нельзя выполнить скоринг. Для скоринга "неуточненной" модели надо вначале сгенерировать набор правил. Дополнительную информацию смотрите в разделе “Генерирование набора правил из слепка модели связывания” на стр. 237.

Максимальное число предсказаний. Задайте максимальное число прогнозов, включенных для каждого набора элементов корзины. Эта опция используется в сочетании с приведенным ниже критерием правила, чтобы получить наилучшие прогнозы, где *наилучшие* означает высший уровень достоверности, поддержки, роста и так далее, как указано ниже.

Критерий правила. Выберите показатель для определения силы правил. Правила сортируются в соответствии с силой критериев, выбранных здесь для получения наилучших прогнозов для набора элементов. Доступные критерии:

- Показатель доверия
- Поддержка
- Поддержка правила (Поддержка * Достоверность)
- Рост
- Возможность внедрения

Разрешить повтор предсказаний. Выберите для включения нескольких правил с одним и тем же консеквентом при скоринге. Например, выбор этой опции позволяет оценить следующие правила:

```
хлеб & сыр -> вино
сыр & фрукты -> вино
```

Выключите эту опцию, чтобы исключить повторные прогнозы при скоринге.

Примечание: Правила с несколькими консеквентами (`хлеб & сыр & фрукты -> вино & паштет`) считаются повторными прогнозами, только если для всех консеквентов (`вино & паштет`) уже были ранее сделаны прогнозы.

Игнорировать несовпавшие элементы корзины. Выберите, чтобы игнорировать наличие дополнительных элементов в наборе элементов. Например, если эта опция выбрана для корзины, содержащей элементы `[палатка & спальный мешок & чайник]`, правило `палатка & спальный мешок -> газовая_плитка` будет действовать, несмотря на присутствие в корзине дополнительного элемента (`чайник`).

При некоторых обстоятельствах требуется исключить дополнительные элементы. Например, возможно, что у клиента, приобретающего палатку, спальный мешок и чайник, может уже быть газовая плитка (на что указывает присутствие чайника). Другими словами, газовая плитка может оказаться не лучшим прогнозом. В таких случаях необходимо отменить выбор опции **Игнорировать несовпавшие элементы корзины**, чтобы антецеденты правила точно соответствовали содержимому корзины. По умолчанию несовпавшие элементы игнорируются.

Проверять отсутствие предсказаний в корзине. Выберите, чтобы консеквенты также не присутствовали в корзине. Например, если цель скоринга состоит в том, чтобы рекомендовать к приобретению предметы домашней мебели, то маловероятно, что владелец корзины, уже содержащей обеденный стол, захочет купить еще один. В таком случае необходимо выбрать эту опцию. С другой стороны, если продукты скоропортящиеся или одноразовые (такие как сыр, детское питание или салфетки), то правила, где консеквент уже есть в корзине, могут быть ценными. В последнем случае наиболее полезной могла бы быть опция **Не проверять прогнозы в корзине**, описанная ниже.

Проверять наличие предсказаний в корзине. Выберите, чтобы консеквенты также присутствовали в корзине. Этот подход полезен, когда нужно получить сведения о существующих клиентах или транзакциях. Например, вам может понадобиться выявить правила с наиболее высоким ростом и затем выяснить, какие покупатели соответствуют этим правилам.

Не проверять предсказания в корзине. Выберите, чтобы включить при скоринге все правила, независимо от присутствия или отсутствия консеквентов в корзине.

Сводка о слепках моделей правил связывания

На вкладке Сводка слепка модели правила связывания выводится число обнаруженных правил, а также минимальное и максимальное значения для поддержки, роста, достоверности и возможности внедрения правил в наборе правил.

Генерирование набора правил из слепка модели связывания

Слепки моделей связывания, такие как Априори и CARMA, можно использовать для непосредственной оценки данных, но можно также сгенерировать сначала подмножество правил, известное как **набор правил**. Наборы правил полезны в частности при работе с необновленной моделью, которую нельзя непосредственно использовать для скоринга. Дополнительную информацию смотрите в разделе “Неуточненные модели” на стр. 50.

Чтобы сгенерировать набор правил, выберите **Набор правил** в меню Создать браузера слепков моделей. Для перевода правил в набор правил можно задать следующие опции:

Имя набора правил. Позволяет задать имя нового сгенерированного узла Набор правил.

Создать узел на ... Управляет положением нового сгенерированного узла Набор правил. Выберите **Холст**, **Палитра GM** или **Оба положения**.

Поле назначения. Определяет, какое выходное поле будет использоваться для сгенерированного узла Набор правил. Выберите из списка одно выходное поле.

Минимальная поддержка. Задайте минимальную поддержку для правил, сохраняемых в сгенерированном наборе правил. Правила с поддержкой меньше выбранного значения не будут включаться в новый набор правил.

Минимальная достоверность. Задайте минимальную достоверность для правил, сохраняемых в сгенерированном наборе правил. Правила с достоверностью меньше выбранного значения не будут включаться в новый набор правил.

Значение по умолчанию. Позволяет задать значение по умолчанию для поля назначения, определенного оцененным записям, для которого нет сработавшего правила.

Генерирование фильтрованной модели

Чтобы сгенерировать фильтрованную модель из слепка модели связывания, такой как на узлах Априори, CARMA или набора правил последовательности, выберите опцию **Фильтрованная модель** в меню Создать браузера слепков моделей. При этом будет создана модель подмножества, включающая в себя только те правила, которые в настоящее время выведены в браузере. *Примечание*: Для неуточненных моделей генерировать фильтрованные модели нельзя.

Для правил фильтрования можно задать следующие опции:

Имя для новой модели. Позволяет задать имя нового узла Фильтрованная модель.

Создать узел на ... Управляет положением нового узла Фильтрованная модель. Выберите **Холст**, **Палитра GM** или **Оба положения**.

Нумерация правил. Задайте, как будут нумероваться ID правил в подмножестве правил, включенных в фильтрованную модель.

- **Оставить исходные идентификационные номера правил.** Выберите для поддержания исходной нумерации правил. По умолчанию правилам предоставляется ID, соответствующий порядку их обнаружения алгоритмом. Этот порядок может изменяться в зависимости от используемого алгоритма.
- **Перенумеровать правила в последовательном порядке, начиная с.** Выберите, чтобы назначить новые ID для фильтрованных правил. Новые ID назначаются на основании порядка сортировки, выведенного в таблице браузера правил на вкладке Модель, начиная с указанного здесь номера. Задать начальный номер для ID можно с использованием стрелок справа.

Скоринг правил связывания

Оценки, полученные при прогоне новых данных через слепок модели правил связывания, возвращаются как отдельные поля. Новые поля добавляются для каждого прогноза: *P* - прогноз, *C* - достоверность и *I* - ID правила. Организация этих выходных полей зависит от того, в каком формате, транзакционном или табличном, получены входные данные. Обзор этих форматов смотрите в разделе “Сравнение табличных и транзакционных данных” на стр. 226.

Пусть, например, выполняется скоринг данных о корзине покупателя при помощи модели, генерирующей прогнозы на основе следующих трех правил:

```
Правило_15  хлеб&вино -> мясо (достоверность 54%)
Правило_22  сыр -> фрукты (достоверность 43%)
Правило_5  хлеб&сыр -> замороженные_овощи (достоверность 24%)
```

Табличные данные. Если данные табличные, возвращается три прогноза (3 по умолчанию) в одной записи.

Таблица 14. Оценки в табличном формате.

ID	Хлеб	Вино	Сыр	P1	C1	I1	P2	C2	I2	P3	C3	I3
Фред	1	1	1	мясо	0,54	15	фрукты	0,43	22	замор_ов	0,24	5

Транзакционные данные. Если данные транзакционные, то для каждого прогноза генерируется отдельная запись. Прогнозы все равно добавляются в отдельные столбцы, но оценки возвращаются по мере вычисления. В результате возникают записи с неполными прогнозами, как показано в приведенном ниже примере вывода. В первой записи второй и третий прогнозы (P2 и P3) пусты, как и соответствующие значения достоверности и ID правила. Однако в последней записи, после возврата оценок, содержатся все три оценки.

Таблица 15. Оценки в транзакционном формате.

ID	Элемент	P1	C1	I1	P2	C2	I2	P3	C3	I3
Фред	хлеб	мясо	0,54	14	$null$	$null$	$null$	$null$	$null$	$null$
Фред	сыр	мясо	0,54	14	фрукты	0,43	22	$null$	$null$	$null$
Фред	вино	мясо	0,54	14	фрукты	0,43	22	замор_ов	0,24	5

Чтобы в отчеты или внедрение попали только полные прогнозы, отберите полные записи при помощи узла выбора.

Примечание: Имена полей в этих примерах сокращены для ясности. Настоящие имена полей результатов при использовании моделей связывания приведены в следующей таблице.

Таблица 16. Имена полей результатов для моделей связывания.

Новое поле	Пример имени поля
Прогноз	*$A-TRANSACTION_NUMBER-1*
Достоверность (или иной критерий)	*$AC-TRANSACTION_NUMBER-1*
ID правила	*$A-Rule_ID-1*

Правила с несколькими консеквентами

Алгоритм CARMA допускает правила с несколькими консеквентами, например:

```
хлеб -> вино&сыр
```

При оценке таких "двойных" правил прогнозы возвращаются в формате, который показан в следующей таблице.

Таблица 17. Оценка результатов, включая прогноз с несколькими консеквентами.

ID	Хлеб	Вино	Сыр	P1	C1	I1	P2	C2	I2	P3	C3	I3
Фред	1	1	1	мясо&овощи	0,54	16	фрукты	0,43	22	замор_ов	0,24	5

В некоторых случаях необходимо разбивать такие оценки перед внедрением. Чтобы разбить прогноз с несколькими консеквентами, потребуется синтаксический анализ поля при помощи строковых функций CLEM.

Внедрение моделей связывания

При скоринге моделей связывания прогнозы и показатели достоверности выводятся в отдельных столбцах (где *P* представляет прогноз, *C* представляет показатель достоверности, а *I* - ID правила). Нужно учесть, находятся ли входные данные в табличном или транзакционном формате. Дополнительную информацию смотрите в разделе “Скоринг правил связывания” на стр. 238.

При подготовке оценок для внедрения может оказаться, что вашей прикладной программе требуется транспонировать выходные данные в такой формат, чтобы прогнозы выводились в строках, а не столбцах (один прогноз на строку, иногда этот формат называют "данные кассовой ленты").

Транспонирование табличных оценок

Вы можете транспонировать табличные оценки для перевода их из столбцов в строки, используя комбинацию следующих действий, описанных в разделе IBM SPSS Modeler.

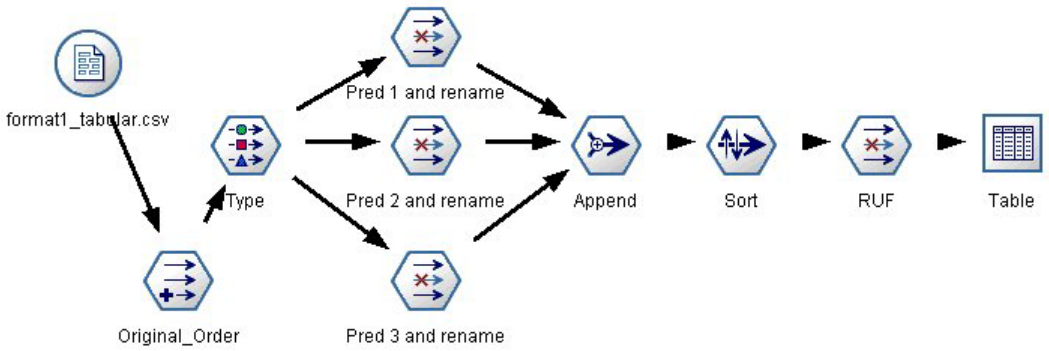


Рисунок 49. Поток примера, используемый для транспонирования табличных данных в формат кассовой ленты

1. С помощью функции `@INDEX` на узле извлечения можно установить текущий порядок прогнозов и сохранить этот показатель в новом поле, например *Исходный_порядок*.
2. Добавьте узел Тип, чтобы все поля были конкретизированы.
3. Воспользуйтесь узлом Фильтр, чтобы переименовать поля прогноза по умолчанию, показателя достоверности и ID (*P1, C1, I1*) в обычные поля, например, *Прог*, *Крит* и *ID_правила*, которые в дальнейшем будут использованы для добавления записей. Вам понадобится один узел Фильтр для каждого сгенерированного прогноза.

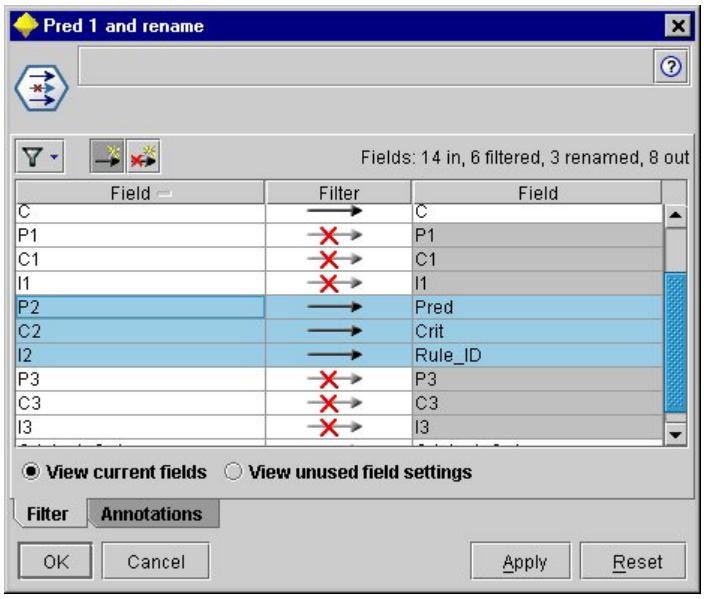


Рисунок 50. Фильтрация полей для прогнозов 1 и 3 при переименовании полей для прогноза 2.

4. Воспользуйтесь узлом Добавить, чтобы добавить значения для совместно используемых полей *Прог*, *Крит* и *ID_правила*.
5. Присоедините узел Сортировка для сортировки записей в порядке возрастания для поля *Исходный порядок* и в порядке убывания для поля *Крит*, которое используется для сортировки прогнозов по таким критериям, как достоверность, рост и поддержка.
6. Для фильтрации поля *Исходный_порядок* из вывода воспользуйтесь другим узлом Фильтр.

Теперь данные готовы к внедрению.

Транспонирование транзакционных оценок

Этот процесс подобен транспонированию табличных оценок . Например, показанный ниже поток транспонирует оценки в формат с единственным прогнозом в каждой строке, как и требуется для внедрения.

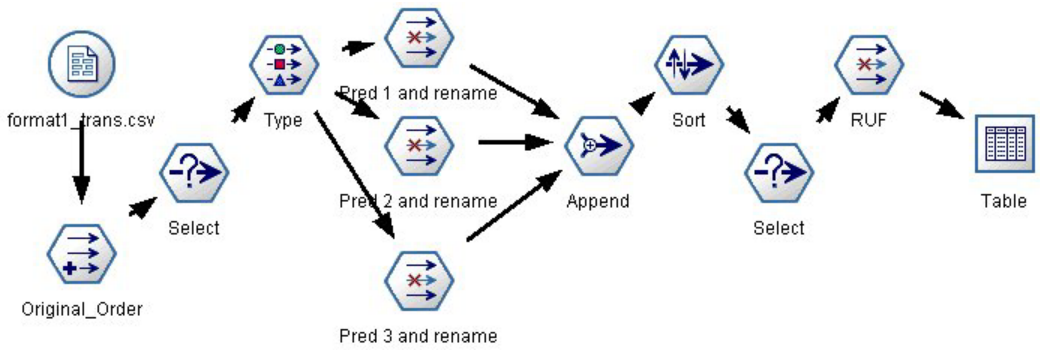


Рисунок 51. Поток примера, используемый для транспонирования транзакционных данных в формат кассовой ленты

С учетом добавления двух узлов выбора этот процесс идентичен рассмотренному ранее для табличных данных.

- Первый узел выбора используется для сравнения ID правил в соседних записях и учета только уникальных или не определенных записей. Этот узел выбора использует для выбора записей выражение CLEM: `ID /= @OFFSET(ID,-1) or @OFFSET(ID,-1) = undef`.

- Второй узел выбора используется для отбрасывания посторонних правил или правил, для которых `ID правила` имеет пустое значение. Этот узел выбора использует для отбрасывания записей следующее выражение CLEM: `not(@NULL(Rule_ID))`.

Для получения дополнительной информации по транспонированию оценок для внедрения свяжитесь со службой технической поддержки.

Узел Последовательность

Узел Последовательность обнаруживает шаблоны в последовательных данных или данных с временной ориентацией в формате `хлеб -> сыр`. Элементы последовательности - это **наборы позиций**, составляющие разовую транзакцию. Например, если человек заходит в магазин и покупает хлеб и молоко, а через несколько дней возвращается и покупает сыр, его покупательская активность может быть представлена двумя наборами товаров. Первый содержит хлеб и молоко, а второй - сыр. **Последовательность** - это список наборов товаров с тенденцией происходить в предсказуемом порядке. Узел Последовательность обнаруживает часто встречающиеся последовательности и создает узел сгенерированной модели, с помощью которого можно делать предсказания.

Требования. Для создания набора правил Последовательности необходимо определить поле ID, дополнительное поле времени и одно или несколько полей содержимого. Обратите внимание, что эти параметры должны быть внесены на вкладке Поля, их нельзя прочесть с вышележащего узла Тип. У поля ID может быть любая роль или уровень измерения. Если вы задаете поле времени, у него может быть любая роль, но система хранения должна быть числовой, даты, времени или отметки времени. Если вы не задаете поле времени, узел Последовательность будет использовать подразумеваемую оценку времени, на самом деле используя в качестве значений времени номера строк. У полей содержимого может быть любой уровень измерения и роль, но все эти поля должны быть одного типа. Если это числовые поля, они должны быть в целочисленных диапазонах, а не в числовых.

Достоинства. Узел последовательности основан на алгоритме правил связывания CARMA, использующем эффективный двухпроходный способ обнаружения последовательностей. Кроме этого, сгенерированный узлом Последовательность узел модели можно вставить в поток данных для создания предсказаний. Сгенерированный узел модели может сгенерировать также суперузел для обнаружения и учета конкретных последовательностей и для создания предсказаний на основе конкретных последовательностей.

Опции полей узла Последовательность

Перед выполнением узла Последовательность необходимо задать ID и поля содержимого на вкладке Поля узла Последовательность. Если вы хотите использовать поле времени, его также необходимо здесь указать.

Поле ID. Выберите поле ID из списка. Значения в поле ID могут быть числовыми или символическими. Каждое уникальное значение этого поля должно означать конкретную единицу анализа. Например, в прикладной программе Потребительская корзина каждый ID может представлять отдельного покупателя. В прикладной программе Анализ Web-журнала каждый ID может представлять отдельный компьютер (по IP-адресу) или одного пользователя (по регистрационным данным).

- **Последовательные ID.** Если ваши данные предварительно отсортированы, так что все записи с одинаковым ID сгруппированы совместно в потоке данных, выбор этой опции ускоряет обработку. Если ваши данные предварительно не отсортированы (или вы в этом не уверены), оставьте эту опцию выключенной, и узел Последовательность отсортирует данные автоматически.

Примечание: Если данные не отсортированы, но выбирается эта опция, можно получить недопустимые результаты для вашей модели Последовательность.

Поле времени. Если вы хотите использовать поле в данных для обозначения времени событий, выберите опцию **Использовать поле времени** и задайте поле, которое будет использоваться. Поле времени должно быть числовым, даты, времени или отметки времени. Если поле времени не задано, предполагается, что

записи будут получаться из источника данных в последовательном порядке, а номера записей будут использоваться как значения времени (первая запись происходит в момент времени "1", вторая - в момент времени "2" и так далее).

Поля содержимого. Задайте поле или поля содержимого для модели. Это поля, которые содержат интересующие вас события при моделировании последовательностей.

Узел Последовательность может обрабатывать данные или в табличном, или в транзакционном формате. Если вы используете несколько полей с транзакционными данными, предполагается, что заданные в этих полях элементы для конкретной записи принадлежат одной транзакции с единой отметкой времени. Дополнительную информацию смотрите в разделе “Сравнение табличных и транзакционных данных” на стр. 226.

Подмножества. Это поле позволяет позволяет задать поле, с помощью которого данные будут разделяться на отдельные выборки для этапов обучения, испытания и проверки при построении модели. Сгенерировав модель при помощи одной выборки и испытав ее при помощи другой, можно получить надежный показатель качества обобщения модели на более крупные наборы данных, подобные текущим. Если при помощи узлов Тип или Раздел были определено несколько полей разделов, на вкладке Поля на каждом узле моделирования, где используется разделение, должно быть выбрано одно поле раздела. (Если представлен только один раздел, он будет использоваться автоматически при всяком включении разделения.) Кроме того, учтите, что для применения выбранного раздела в анализе на вкладке Параметры модели для узла должно быть также включено разделение. (Выключение этой опции делает возможным отключение разделения без изменения значений параметров полей.)

Опции моделей узла Последовательность

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Минимальная поддержка правила (%). Вы можете задать критерий поддержки. **Поддержка правил** относится к тем ID в обучающих данных, которые содержат полную последовательность. Если вы хотите сосредоточиться на более общих последовательностях, увеличьте значение этого параметра.

Минимальная достоверность правила (%). Для удержания отдельных последовательностей в наборе последовательностей можно задать критерий достоверности. **Достоверность** указывает процентную долю ID, для которых сделаны правильные предсказания, среди всех ID, для которых делаются предсказания с использованием данного правила. Она вычисляется на основе обучающих данных как отношение числа ID, для которых найдена полная последовательность, к числу ID, для которых найдены антецеденты. Последовательности, у которых достоверность ниже задаваемой критерием, отбрасываются. Если у вас получается слишком много последовательностей вообще или неинтересных для модели последовательностей, попробуйте увеличить значение этого параметра. Если вы получаете слишком мало последовательностей, попробуйте уменьшить этот параметр.

Максимальный размер последовательности. В последовательности вы можете задать максимальное число отдельных *наборов элементов* (в противоположность просто *элементам*). Если интересующие вас последовательности относительно короткие, можно уменьшить этот параметр, чтобы ускорить построение набора последовательностей.

Прогнозы для добавления к потоку. Задайте количество предсказаний, которые будут добавлены к потоку итоговым сгенерированным узлом Модель. Дополнительную информацию смотрите в разделе “Слепки моделей последовательности” на стр. 244.

Опции эксперта узла Последовательность

Если вы хорошо знакомы с работой узла Последовательность, следующие опции эксперта позволят точнее настроить процесс построения модели. Для доступа к дополнительным опциям выберите режим **Дополнительно** на вкладке Дополнительно.

Задать максимальную длительность. Если выбрана эта опция, последовательности ограничены только теми, длительность которых (время между первым и последним наборами данных) меньше или равно заданному значению. Если вы не задали поле времени, длительность выражается в терминах строк (записей) необработанных данных. Если используемое поле времени - это поле времени, даты или отметки времени, длительность выражается в секундах. Для числовых полей длительность выражается в тех же единицах, что и эти поля.

Задать значение усечения. Для экономии памяти используемый на узле Последовательность алгоритм CARMA при обработке периодически удаляет (**отсекает**) редкие наборы данных из своего списка потенциальных наборов данных. Выберите эту опцию для настройки частоты усечения. Заданное число определяет частоту операция усечения. Введите меньшее значение для снижения требований алгоритма к памяти (хотя потенциально это увеличивает нужное время обучения), или большее значение для ускорения обучения (но потенциально это повысит требования алгоритма к объему памяти).

Задать максимальное число последовательностей в памяти. Если выбрана эта опция, алгоритм CARMA при построении модели будет ограничивать количество хранимых в своей памяти последовательностей-кандидатов заданным числом последовательностей. Выберите эту опцию, если IBM SPSS Modeler при построении моделей Последовательность использует слишком много памяти. Обратите внимание на то, что заданное здесь максимальное значение для последовательностей - это количество последовательностей-кандидатов, внутренне отслеживаемых при построении моделей. Это число должно быть больше, чем ожидаемое число последовательностей в итоговой модели.

Ограничить пропуски между наборами элементов. Эта опция позволяет задать ограничения на временные интервалы, разделяющие наборы элементов. При выборе этой опции наборы элементов с пропусками по времени, меньшими, чем **Минимальный пропуск**, или большими, чем **Максимальный пропуск**, которые вы задали, не будут рассматриваться для образования части последовательности. Используйте эту опцию, чтобы исключить учет последовательностей, которые содержат длинные интервалы времени или заключены в очень коротком промежутке времени.

Примечание: Если используемое поле времени - это поле времени, даты или отметки времени, пропуск времени выражается в секундах. Для числовых полей пропуск времени выражается в тех же единицах, что и эти поля.

Например, рассмотрим следующий список транзакций.

Таблица 18. Пример списка транзакций.

ID	Время	Содержимое
1001	1	яблоки
1001	2	хлеб
1001	5	сыр
1001	6	приправа

Если для этих данных строится модель с заданным значением 2 для пропуска времени, будут получены следующие последовательности:

```
яблоки -> сыр

яблоки -> приправа
```

```
хлеб -> сыр

хлеб -> приправа
```

Такие последовательности, как `яблоки -> хлеб`, не будут представлены, так как пропуск между `яблоки` и `хлеб` меньше минимального пропуска. Аналогично рассмотрим следующие альтернативные данные.

Таблица 19. Пример списка транзакций.

ID	Время	Содержимое
1001	1	яблоки
1001	2	хлеб
1001	5	сыр
1001	20	приправа

Если для максимального пропуска задать значение 10, не будут представлены никакие последовательности с `приправа`, так как пропуск между `сыр` и `приправа` слишком велик, чтобы рассматривать эти элементы как часть одной последовательности.

Слепки моделей последовательности

Слепки моделей последовательностей представляют последовательности, обнаруженные для конкретных выходных полей, найденных узлом Последовательность, и могут быть вставлены в потоки для генерирования предсказаний.

При запуске потока, содержащего узел Последовательность, этот узел добавляет два поля, содержащие предсказания и связанные значения достоверности для каждого предсказания, из модели последовательности в данные. По умолчанию добавляется три пары полей, содержащие три лучших предсказания (и их связанные значения правдоподобия). Вы можете изменить количество сгенерированных при построении модели предсказаний, задав опции моделей узла Последовательность во время построения, а также на вкладке Параметры после добавления слепка модели в поток. Дополнительную информацию смотрите в разделе “Параметры слепков моделей последовательностей” на стр. 247.

Имена новых полей получаются из имени модели. Имена этих полей - это *$S-последовательность-n* для поля предсказания (где *n* означает *n*-ое предсказание) и *$SC-последовательность-n* для поля достоверности. В потоке с несколькими узлами правил последовательности в ряду имена новых полей будут содержать в префиксах номера, отличающие поля друг от друга. Для первого узла Набор последовательностей в потоке будут использоваться обычные номера, для второго узла имена будут начинаться с *$S1-* и *$SC1-*, для третьего узла префиксами будут *$S2-* и *$SC2-* и так далее. Предсказания выводятся по порядку достоверности, то есть *$S-последовательность-1* содержит предсказание с самой высокой достоверностью, *$S-последовательность-2* - предсказание со следующей по порядку достоверностью и так далее. Для записей с числом доступных предсказаний меньшим, чем число требуемых предсказаний, остальные предсказания будут содержать значение `$null$`. Например, если для конкретной записи можно сделать только два предсказания, значения *$S-последовательность-3* и *$SC-последовательность-3* будут равны `$null$`.

Для каждой записи правила в модели сравниваются с набором обработанных до сих пор транзакций для текущего ID, в том числе текущая запись и все предшествующие записи с тем же ID и более ранней отметкой времени. *k* правил с самыми большими значениями достоверности, которые применимы к этому набору транзакций, используются для генерирования *k* предсказаний, где *k* - это число предсказаний, заданное на вкладке Параметры после добавления модели к потоку. (Если несколько правил предсказывают одинаковые выходные значения для набора транзакций, используется только правило с максимальной достоверностью). Дополнительную информацию смотрите в разделе “Параметры слепков моделей последовательностей” на стр. 247.

Как и для других типов моделей правил связывания, формат данных должен совпадать с форматом, использованным при построении модели последовательности. Например, модели, построенные с использованием табличных данных, можно использовать для скоринга только табличных данных. Дополнительную информацию смотрите в разделе “Скоринг правил связывания” на стр. 238.

Примечание: При скоринге данных с использованием сгенерированного узла Набор последовательностей в потоке любые параметры допуска или пропусков, выбранные при построении модели, для целей скоринга игнорируются.

Предсказания из правил последовательностей

Узел работает с записями с учетом времени (или по порядку, если для построения модели использовалось поле отметки времени). Записи должны сортироваться по полю ID и по полю отметок времен (если такие есть). Однако предсказания не привязаны к отметке времени записи, к которой они добавляются. Они просто указывают на наиболее вероятные элементы, которые могут встретиться *в некоторой точке в будущем* при данной хронологии транзакций для текущего ID вплоть до текущей записи.

Обратите внимание на то, что предсказания для каждой записи не обязательно зависят от транзакций этой записи. Если транзакции текущей записи не инициализируют конкретное правило, правила будут выбираться на основе предыдущих записей для текущего ID. Другими словами, если текущая запись не добавляет полезной прогнозирующей информации для последовательности, к текущей записи применяется предсказание от последней полезной для этого ID транзакции.

Предположим, например, что есть модель Последовательность с одним правилом

```
Джем -> Хлеб (0.66)
```

и вы применяете его для следующих записей.

Таблица 20. Примеры записей.

ID	Покупка	Прогноз
001	джем	bread
001	молоко	bread

Обратим внимание на то, что первая запись генерирует предсказание *хлеб*, как и можно было предположить. Вторая запись также содержит предсказание *хлеб*, так как не существует правила для *джем* с последующим *молоко*, поэтому транзакция *молоко* не добавляет какой-либо полезной информации, и снова применяется правило `Джем -> Хлеб`.

Генерирование новых узлов

Меню Создать позволяет сформировать новые суперузлы на основе модели последовательности.

- **Суперузел правил.** Создает суперузел, который может обнаруживать и учитывать появление последовательностей в оцениваемых данных. Если никакое правило не выбрано, эта опция отключается. Дополнительную информацию смотрите в разделе “Генерирование надузла правил из слепка модели последовательности” на стр. 247.
- **Модель для палитры .** Возвращает модель на палитру Модели. Это полезно в ситуациях, когда коллега послал вам поток, содержащий модель, но не саму модель.

Подробности слепков моделей последовательностей

На вкладке Модель для слепка модели Последовательность выводятся правила, извлеченные алгоритмом. Каждая строка в таблице представляет правило, причем за антецедентом (часть "if" в правиле) в первом столбце следует консеквент (часть "then" в правиле) во втором столбце.

Каждое правило показывается в следующем формате.

Таблица 21. Формат правил

Антецедент	Консеквент
пиво и консервы	beer
fish fish	fish

Первый пример правила интерпретируется как *для ID, у которых "пиво" и "консервы" были в одной транзакции, вероятно появление консеквента "пиво."* Второй пример правила можно интерпретировать как *для ID, у которых есть "рыба" в одной транзакции и затем "рыба" в другой, вероятно появление консеквента "рыба".* Обратите внимание, что в первом правиле *пиво* и *консервы* приобретаются одновременно, а во втором *рыба* приобретается в двух разных транзакциях.

Меню Сортировка. Кнопка меню Сортировка на панели инструментов управляет сортировкой правил. Направление сортировки (по возрастанию или убыванию) можно изменить с помощью кнопки направления сортировки (стрелка вверх или вниз).

Для сортировки правил может использоваться:

- % поддержки
- % достоверности
- % поддержки правил
- Консеквент
- Первый антецедент
- Последний антецедент
- Количество элементов (антецеденты)

Например, следующая таблица сортируется в убывающем порядке по числу элементов. Правила с несколькими элементами в наборе антецедентов предшествуют правилам с меньшим числом элементов.

Таблица 22. Правила, отсортированные по числу элементов

Антецедент	Консеквент
пиво и консервы и замороженные блюда	frozenmeal
пиво и консервы	beer
fish fish	fish
softdrink	softdrink

Меню Показать/скрыть критерии. В меню Показать/скрыть критерии кнопка со значком решетки управляет опцией для вывода правил. Доступны следующие опции вывода:

- **Экземпляры** выводит информацию о количестве уникальных ID, для которых в *полной последовательности* присутствуют и антецеденты, и консеквенты. (Обратите внимание на то, что в этом есть отличие от моделей связывания, когда количество экземпляров указывает на количество ID, для которых применимы *только* антецеденты). Например, при данном правиле хлеб -> сыр те ID в данных обучения, для которых присутствуют и *хлеб*, и *сыр*, называются **экземплярами**.
- **Поддержка** выводит процентную долю ID в данных обучения, для которых значения антецедентов - это true. Например, если 50% данных обучения включают в себя антецедент *хлеб*, поддержка для правила хлеб -> сыр составит 50%. (Как указывалось ранее, для моделей связывания поддержка *не* основана на числе экземпляров).

- **Достоверность** указывает процентную долю ID, для которых сделаны правильные предсказания, среди всех ID, для которых делаются предсказания с использованием данного правила. Она вычисляется на основе обучающих данных как отношение числа ID, для которых найдена полная последовательность, к числу ID, для которых найдены антецеденты. Например, если 50% данных обучения содержат `консервы` (обозначает поддержку антецедента), но только 20% содержат и `консервы`, и `замороженные блюда`, достоверность для правила `консервы -> замороженные блюда` будет равна `Поддержка правила / Поддержка антецедента`, то есть 40% в нашем случае.
- **Поддержка правила** для моделей Последовательность основана на экземплярах и выводит процентную долю данных обучения, для которых у полного правила, антецедентов и консеквентов значение true. Например, если в 20% данных обучения содержатся и *хлеб*, и *сыр*, поддержка для правила `хлеб -> сыр` составит 20%.

Обратите внимание, что вычисление процентной доли основано на количестве допустимых транзакций (в которых есть по крайней мере один наблюдаемый элемент или значение true), а не на полном количестве транзакций. Недопустимые транзакции, у которых нет элементов или значений true, отбрасываются при этих вычислениях.

Кнопка Фильтр. Кнопка Фильтр (значок с воронкой) в меню раскрывает нижнюю часть диалогового окна, показывая панель, где выводятся активные фильтры правил. Фильтры используются, чтобы сузить набор правил, показанных на вкладке Модели.

Рисунок 52. Кнопка Фильтр

Чтобы создать фильтр, щелкните по значку Фильтр справа от раскрытой панели. Откроется отдельное диалоговое окно, в котором можно задать ограничения для вывода правил. Заметим, что кнопка Фильтр часто используется в сочетании с меню Генерировать, чтобы сначала фильтровать правила и затем генерировать модель, содержащую этот поднабор правил. Дополнительную информацию смотрите в разделе “Задание фильтров для правил” на стр. 234 ниже.

Параметры слепков моделей последовательностей

На вкладке Параметры слепка модели Последовательность выводятся опции скоринга для этой модели. Эта вкладка доступна только после добавления модели на холст потока для скоринга.

Максимальное число предсказаний. Задайте максимальное число прогнозов, включенных для каждого набора элементов корзины. Правила с самыми большими значениями достоверности, применимые к этому набору транзакций, используются для генерирования предсказаний для записи до заданного предела.

Сводка слепков моделей последовательностей

На вкладке Сводка для слепка модели правил последовательностей выводится количество обнаруженных правил, а также минимальные и максимальные значения поддержки и достоверности в правилах. Если вы выполнили узел Анализ, присоединенный к данному узлу моделирования, информация этого анализа будет также выведена в этом разделе.

Дополнительную информацию смотрите в разделе “Просмотр слепков моделей” на стр. 41.

Генерирование надузла правил из слепка модели последовательности

Чтобы сгенерировать надузел правил на основании правила последовательности:

1. На вкладке Модель слепка модели правил последовательности щелкните по строке в таблице, чтобы выбрать нужное правило.
2. В меню браузера правил выберите:

Создать > Надузел правил

Важно: Чтобы использовать сгенерированный надузел, необходимо отсортировать данные по полю ID (и по полю Время, если такое есть), прежде чем передавать их на надузел. В несортированных данных надузел не обнаружит правильную последовательность.

Для генерирования надузла правил можно задать следующие опции:

Обнаружение. Задает, как определяются совпадения для переданных на надузел данных.

- **Только антецеденты.** Надузел будет идентифицировать совпадение всякий раз при нахождении антецедентов для выбранного правила в правильном порядке в наборе записей с одинаковым ID независимо от того, найден ли также консеквент. Обратите внимание на то, что при этом не принимается во внимание допуск отметки времени и параметры ограничения разрывов для элементов из исходного узла моделирования последовательности. Когда последний набор элементов антецедентов обнаруживается в потоке (а все другие антецеденты найдены в правильном порядке), все последовательные записи с текущим ID будут содержать выбранную ниже сводку.
- **Вся последовательность.** Надузел будет идентифицировать совпадение всякий раз при нахождении антецедентов и консеквента для выбранного правила в правильном порядке в наборе записей с одинаковым ID. При этом не принимается во внимание допуск отметки времени и параметры ограничения разрывов для элементов из исходного узла моделирования последовательности. Когда консеквент обнаруживается в потоке (а все антецеденты также найдены в правильном порядке), текущая запись и все последовательные записи с текущим ID будут содержать выбранную ниже сводку.

Вывести. Управляет, сколько именно сводок добавляется к данным в выводе надузла правил.

- **Значение консеквента для первого вхождения.** Значение, добавленное к данным, - это значение консеквента, предсказанное на основании первого появления совпадения. Значения добавляются как новое поле с именем *rule_n_consequent*, где *n* - это номер правила (на основании порядка создания надузлов правил в потоке).
- **Истинное значение для первого вхождения.** Значение, добавленное к данным, - true, если есть по крайней мере одно совпадение для данного ID, и false, если совпадений нет. Значения добавляются как новое поле с именем *rule_n_flag*.
- **Число вхождений.** Значение, добавляемое к данным, - это число совпадений для данного ID. Значения добавляются как новое поле с именем *rule_n_count*.
- **Номер правила.** Добавленное значение - это номер для выбранного правила. **Номера правил** назначаются на основании порядка, в котором надузлы добавлялись в поток. Например, первый надузел правил учитывает *правило 1*, второй надузел - *правило 2* и так далее. Эта опция окажется наиболее полезной, когда вы будете включать несколько надузлов правил в поток. Значения добавляются как новое поле с именем *rule_n_number*.
- **Включить доверительные вероятности.** При выборе этой опции в поток данных будет добавляться доверительная вероятность правил, а также выбранная сводка. Значения добавляются как новое поле с именем *rule_n_confidence*.

Глава 13. Модели временных рядов

Зачем нужно прогнозировать?

Предсказание - это прогноз значений одного или нескольких рядов на протяжении времени. Например, нужно предсказать ожидаемый спрос на линейку продуктов или услуг, чтобы выделить ресурсы для производства и распределения. Поскольку на реализацию планов требуется время, прогнозы оказываются существенным инструментом во многих процессах планирования.

В методах моделирования временных рядов предполагается, что их хронология повторяется — если не в точности, то все же достаточно близко, чтобы изучение прошлого помогало принимать лучшие решения в будущем. Чтобы предсказывать продажи в следующем году, например, можно начать с изучения продаж этого года и далее двигаться в прошлое, чтобы выявить тренды или шаблоны, если такие найдутся, по котором события развивались в прошлые годы. Но иногда выявить шаблоны оказывается непросто. Например, если ваши продажи росли несколько недель к ряду, считать ли это частью сезонного цикла или началом долговременного тренда?

Используя технологии статистического моделирования, вы можете анализировать шаблоны в прошлых данных и проецировать найденные шаблоны вперед, чтобы установить диапазон, в который скорее всего попадут будущие значения этого ряда. Результатом станут более точные прогнозы, на которых вы будете основывать свои решения.

Данные временного ряда

Временной ряд - это упорядоченное собрание измерений, выполненных через регулярные интервалы -- например, ежедневные курсы акций или еженедельные данные о продажах. Измерение может касаться любых интересующих вас показателей, и все временные ряды можно в общем случае классифицировать так:

- **Зависимый.** Ряд, который нужно предсказать.
- **Предиктор.** Ряд, который может сработать при прогнозировании ряда назначения -- например, рекламный бюджет при предсказании продаж. Предикторы можно использовать только совместно с моделями АРПСС.
- **Событие.** Специальный ряд предиктора, служащий для учета неоднократных событий - например, кампаний по увеличению сбыта.
- **Внешнее событие.** Специальный ряд предиктора, служащий для учета однократных событий в прошлом -- например, отключение энергии или забастовка сотрудников.

Интервалы могут задаваться в любых единицах времени, но они должны быть одинаковы во всем ряду измерений. Более того, если по некоторому интервалу нет измерения, должно быть признано пропущенное значение. Таким образом, число интервалов некоторого измерения (включая те, по которым значение пропущено) задает длину отрезка хронологии данных.

Характеристики временных рядов

Изучение прошлого поведения ряда поможет распознать шаблоны и улучшить прогнозирование. На графике многие ряды обладают одной или несколькими из следующих особенностей:

- Тенденции
- Сезонные и несезонные циклы
- Импульсы и ступеньки
- Выбросы

Тенденции

Тренд - это постепенный сдвиг вверх или вниз уровня ряда или тенденция значений ряда расти или уменьшаться со временем.

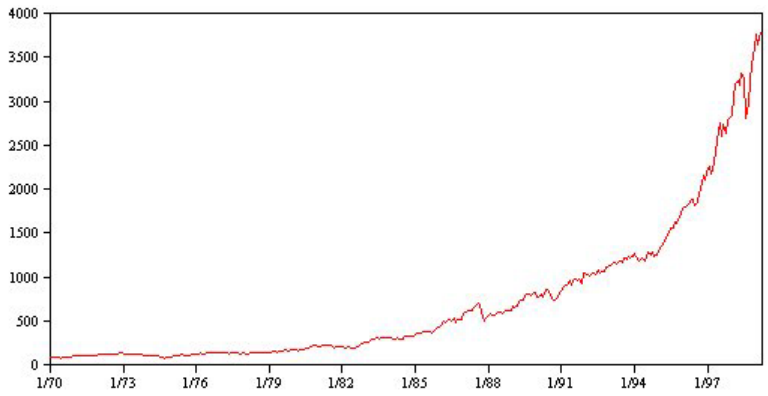


Рисунок 53. Тренд

Тренды бывают **локальные** или **глобальные**, причем в одном и том же ряду могут встречаться тренды обоих типов. Временной ряд индекса фондового рынка показывает глобальную тенденцию к росту. Локальные убывающие тренды встречаются во времена рецессии, и локальные растущие тренды встречаются во времена процветания.

Тренды бывают **линейные** и **нелинейные**. Линейные тренды представляют собой положительные или отрицательные инкременты к уровню ряда, аналогичные действию простых процентов на капитал. Нелинейные тренды часто мультипликативны, то есть их инкременты пропорциональны предыдущему значению в ряду.

Глобальные линейные тренды хорошо приближаются и прогнозируются как моделями экспоненциального сглаживания, так и моделями АРПСС. При построении моделей АРПСС те ряды, которые проявляют тренды, обычно дифференцируют, чтобы удалить влияние тренда.

Сезонные циклы

Сезонный цикл - это структура временного ряда значений с многократными предсказуемыми повторами.

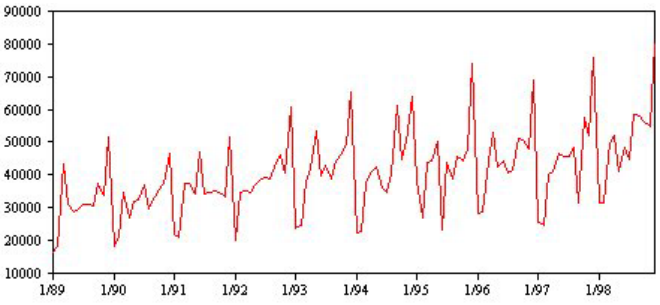


Рисунок 54. Сезонный цикл

Сезонные циклы привязаны к интервалу вашего ряда. Например, данные ежемесячных измерений часто проявляют цикличность на протяжении кварталов и лет. Данные ежемесячных измерений могут проявить выраженные квартальные циклы с понижением в первой четверти или годичные циклы с пиком в каждом декабре. Про ряды, в которых сказываются сезонные циклы, говорят, что они проявляют **сезонность**.

Сезонные шаблоны полезны для подгонки моделей и прогнозов и для захвата сезонности есть модели экспоненциального сглаживания и АРПСС.

Несезонные циклы

Несезонный цикл - это структура временного ряда значений с многократными и, возможно, непредсказуемыми повторами.

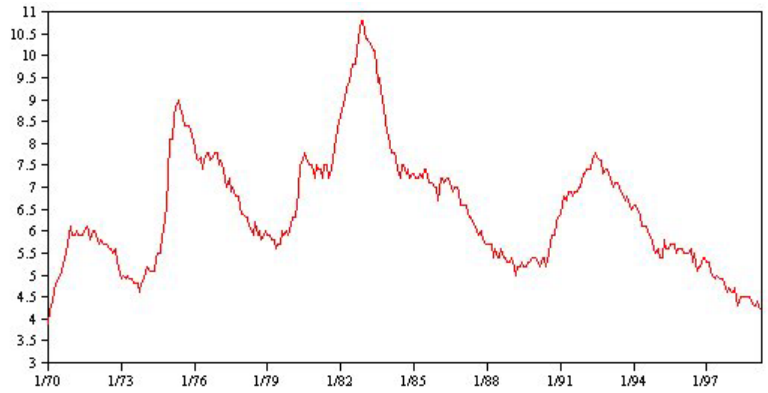


Рисунок 55. Несезонный цикл

В некоторых рядах, таких как уровень безработицы, четко прослеживаются сезонные колебания; однако период цикла варьируется и точно предсказать наступление высокого или низкого уровня трудно. Для других рядов характерны предсказуемые циклы, но они плохо согласуются с грегорианским календарем или просто имеют период больше года. Например, приливы следуют лунному календарю, международный туризм и торговля усиливаются каждые четыре года в связи с олимпиадами, а у многих религиозных праздников дата по грегорианскому календарю меняется год от года.

Несезонные циклические шаблоны трудны для моделирования и обычно увеличивают неопределенность прогноза. Например, фондовый рынок дает многочисленные примеры рядов, которые не поддаются усилиям прогнозистов. И все-таки несезонные шаблоны тоже следует учитывать, когда они встречаются. Во многих случаях удается найти модель, которая с разумной точностью подходит к данным хронологии и помогает минимизировать степень неуверенности прогнозов.

Импульсы и ступеньки

Нередко ряд содержит резкое изменение уровня. Обычно такие изменения бывают двух типов:

- Неожиданный *временный* сдвиг в уровне ряда (**импульс**)
- Неожиданный *постоянный* сдвиг в уровне ряда (**ступенька**)

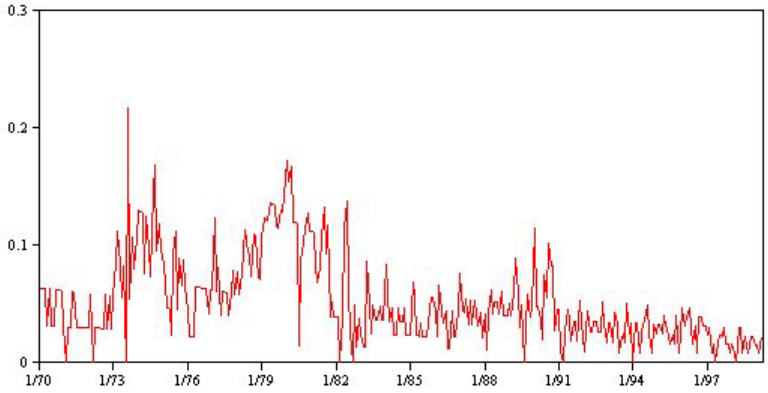


Рисунок 56. Ряд с импульсом

При наблюдении ступенек импульсов важно найти им убедительное объяснение. Модели временных рядов предназначены для учета постепенных, а не внезапных изменений. В результате они склонны недооценивать

импульсы и рушится из-за ступенек, так что модель плохо подгоняется под данные и дает ненадежные прогнозы. (В некоторых случаях сезонности также встречаются неожиданные изменения уровня, но от сезона к сезону уровень неизменен.)

Если резкое изменение можно объяснить, его можно промоделировать как **внешнее событие** или как **событие**. Например, в августе 1973 нефтяное эмбарго, введенное странами - участниками организации экспортеров нефти (ОПЕК) вызвало резкий скачок уровня инфляции, которая вернулась к нормальным уровням в последующие месяцы. Задав **точечное внешнее событие** в месяц введения эмбарго, можно улучшить подгонку модели, что, в свою очередь, улучшит прогнозы. Например, магазин может обнаружить уровень продаж много выше обычного в день, когда на все товары была введена скидка 50%. Задав акцию по привлечению покупателей с 50%-й скидкой как многократное **событие**, вы можете улучшить подгонку модели и оценить эффект от будущих промо-акций.

Выбросы

Сдвиги уровня во временных рядах, которые не удается объяснить, называются **выбросами**. Такие наблюдения не согласуются с остальным рядом и могут драматически исказить анализ, что скажется на прогностической способности модели временного ряда.

На следующем рисунке показаны различные типы выбросов, часто встречающихся во временных рядах. Синими линиями показаны ряды без выбросов. Красные линии предполагают, что данные могут следовать некоторому шаблону, если предположить наличие выбросов. Все такие выбросы классифицируются как **детерминированные**, поскольку влияют только на средние уровни рядов.

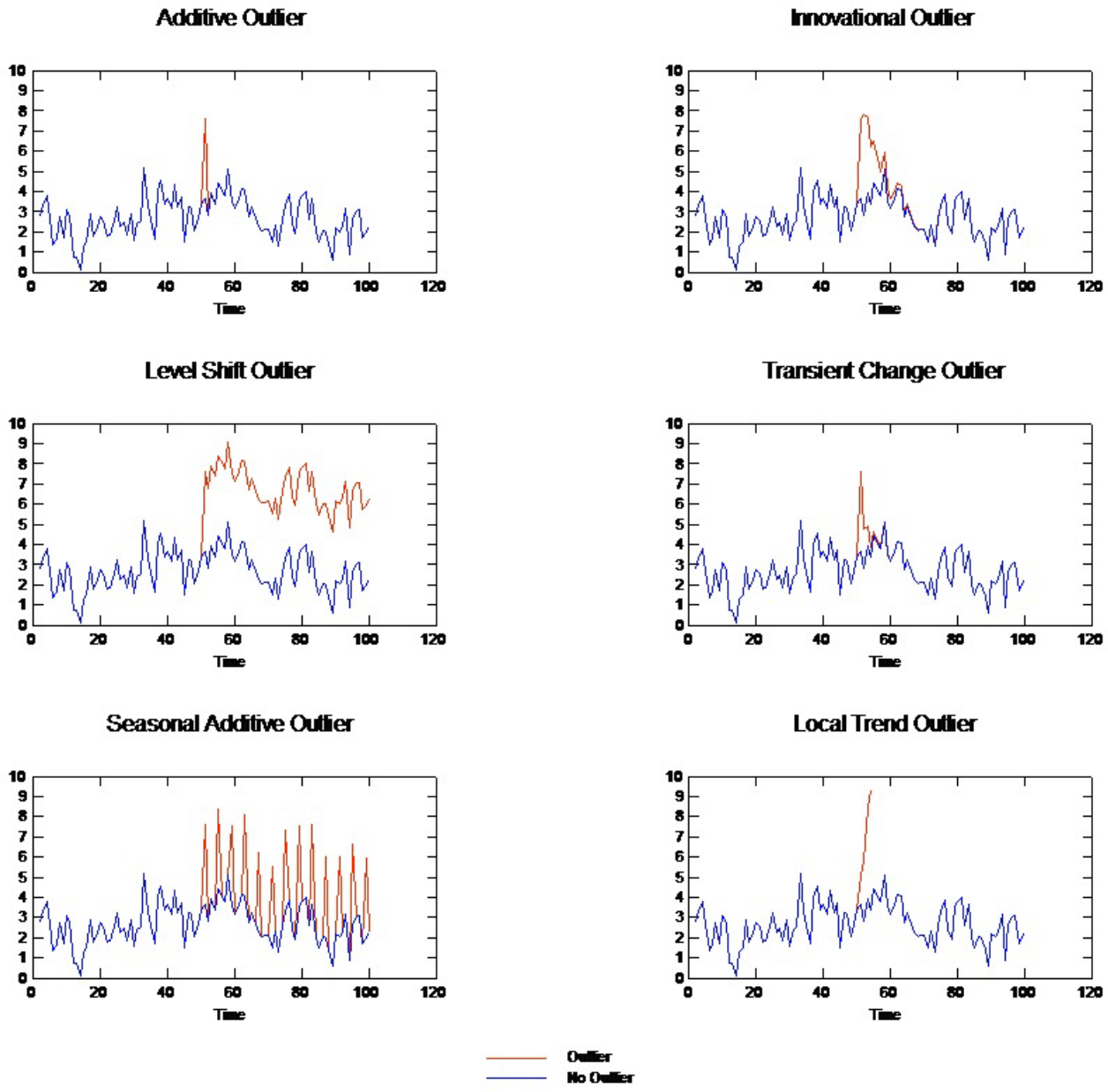


Рисунок 57. Типы выбросов

- **Аддитивный выброс.** Аддитивный выброс выглядит как неожиданно большое или маленькое значение, встречающееся только в одном наблюдении. На дальнейшие наблюдения аддитивный выброс не влияет. Идущие подряд аддитивные выбросы обычно называются **вставками аддитивных выбросов**.
- **Инновационный выброс.** Инновационный выброс характеризуется начальным воздействием и влиянием на последующие наблюдения. Влияние выбросов может расти с течением времени.
- **Выброс сдвига уровня.** При сдвиге уровня все наблюдения после выброса сдвигаются к новому уровню. В отличии от аддитивных выбросов выброс сдвига уровня влияет на все последующие наблюдения постоянным образом.
- **Выброс нестационарного изменения.** Выброс нестационарного изменения похож на выбросы сдвига уровня, но его влияние на последующие наблюдения экспоненциально затухает. Со временем ряд возвращается к своему обычному уровню.

- **Сезонный аддитивный выброс.** Сезонный аддитивный выброс выглядит как неожиданно большое или маленькое значение, встречающееся многократно через регулярные интервалы.
- **Выброс локального тренда.** Выброс локального тренда приводит к общему изменению ряда, которое вызывается систематическими выбросами, начавшимися после наступления первоначального выброса.

Обнаружение выброса во временном ряде включает в себя локализацию выбросов, выяснение их типов и величины. Цай (1988) предложил итерационную процедуру обнаружения сдвигов в уровне среднего для поиска предполагаемых детерминистских выбросов. Процесс включает в себя сравнение моделей ряда без предположения о наличии выбросов и предполагающих выбросы. Различия между моделями дают оценку, как влияет на моделирование обработка того или иного момента как выброса.

Функции автокорреляции и частной автокорреляции

Автокорреляция и частная автокорреляция - служит мерой связанности между текущими и прошлыми рядами значений и показывает, какие прошлые значения наиболее полезны для предсказания будущих значений. Опираясь на это знание можно выбрать последовательность процессов в модели АРПСС. Опишем это подробнее.

- **Функция автокорреляции (АКФ).** При лаге *k* это корреляция между рядами значений, отстоящих друг от друга на *k* интервалов.
- **Функция частной автокорреляции (ЧАКФ).** При лаге *k* это корреляция между рядами значений, отстоящих друг от друга на *k* интервалов, считая значения интервалов в промежутке.

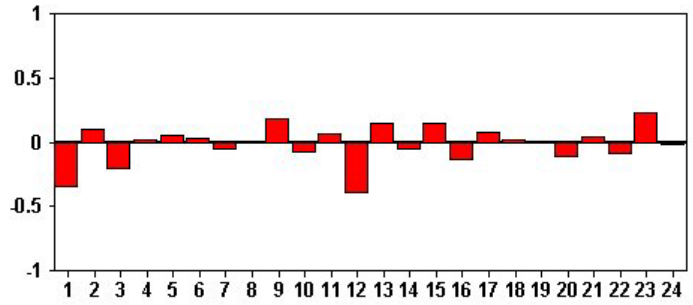


Рисунок 58. График АКФ для ряда

Ось *x* графика ACF показывает лаг, при котором вычисляется автокорреляция; ось *y* показывает значение корреляции (от –1 до 1). Например, пик при лаге 1 на графике АКФ показывает сильную корреляцию между значениям ряда и предыдущим значением, пик при лаге 2 показывает сильную корреляцию между каждым значением и значением в более ранний момент на расстоянии 2 от данного и так далее.

- Положительная корреляция при некотором лаге показывает, что большому текущему значению отвечают большие значения с этим лагом, а отрицательная корреляция показывает, что что большому текущему значению отвечают малые значения с этим лагом.
- Абсолютная величина корреляции служит мерой связанности, причем чем больше абсолютное значение, тем сильнее взаимосвязь.

Преобразования рядов

Преобразования нередко помогают стабилизировать ряд перед оценкой модели. Особенно это актуально для моделей АРПСС, в которых требуется **стационарность** ряда, для которого оценивается модель. Ряд стационарен, если его глобальный уровень (среднее) и среднее отклонение от этого уровня (дисперсия) остаются постоянными на протяжении ряда.

Хотя самые интересные ряды как раз нестационарны, метод АРПСС все же эффективен, если ряд можно сделать стационарным, применив трансформации, такие как натуральный логарифм, дифференцирование или сезонное дифференцирование.

Преобразования, стабилизирующие дисперсию. Ряды, в которых дисперсия изменяется с течением времени, нередко удается стабилизировать при помощи преобразования со взятием натурального логарифма или извлечением квадратного корня. Такие преобразования называются функциональными.

- **Нат.логарифм.** К ряду значений применяется натуральный логарифм.
- **Корень квадратный.** К ряду значений применяется квадратный корень.

Преобразования натурального логарифма и квадратного корня не применимы к ряду, содержащему отрицательные значения.

Преобразования, стабилизирующие уровень. Медленной уменьшение значений в АКФ - признак того, что каждое значения в ряду сильно скоррелировано с предыдущим. Проанализировав изменение значений в ряду, вы получите стабильный уровень.

- **Простое дифференцирование.** Для каждого значения в ряду, за исключением самого старого, вычисляется разность между этим значением и предыдущим. отсюда следует, что продифференцированный ряд на одно значение короче исходного ряда.
- **Сезонное дифференцирование.** То же, что простое дифференцирование, только разности берутся между каждым значением и значением из предыдущего сезона.

Если одновременно применяются простое или сезонное дифференцирование и преобразование логарифма или квадратного корня, первым всегда применяется преобразование, стабилизирующее дисперсию. Когда одновременно применяется и простое, и сезонное дифференцирование, результат не зависит от того, какое из этих преобразований применялось первым.

Ряды предикторов

Ряды предикторов содержат связанные данные, которые могут сработать при объяснении поведения прогнозируемого ряда. Например, интернет-магазин или магазин почтой может прогнозировать продажи исходя из количества каталогов, отправленных по почте, числа открытых телефонных линий или числа попаданий на Web-страницу компании.

Любой ряд можно назначить предиктором, если он простирается в предсказываемое будущее и содержит полные данные, без пропусков.

Соблюдайте осторожность, добавляя предикторы в модель. Добавление большого числа предикторов увеличит время, затраченное на оценку модели. Хотя добавление предикторов может улучшить способность модели приблизить данные хронологии, это не обязательно означает, что модель начнет лучше справляться с прогнозированием, так что усложнение иногда не оправдывается. В идеальном случае цель состоит в том, чтобы найти простейшую модель, которая хорошо прогнозирует.

Общая рекомендация та, что число предикторов должно быть меньше размера выборки, поделенного на 15 (то есть не больше одного предиктора на 15 наблюдений).

Предикторы с пропущенными данными. Предикторы с неполными данными или с пропусками использовать нельзя. Это требование относится как к данным хронологии, так и к будущим значениям. В некоторых случаях это ограничение можно обойти, задав для модели интервал оценки, исключающий самые старые данные при оценке моделей.

Узел моделирования временных рядов

Узел временных рядов оценивает модели экспоненциального сглаживания, одномерные и многомерные модели авторегрессии и проинтегрированного скользящего среднего, модели АРПСС (Autoregressive Integrated Moving Average, ARIMA) (или передаточных функций) для временных рядов и составляет прогнозы на основе данных временных рядов.

Экспоненциальное сглаживание - это способ предсказания, использующий взвешенные значения предыдущих наблюдений ряда для предсказания будущих значений. Само по себе экспоненциальное сглаживание не основано на теоретическом понимании данных. Оно прогнозирует одну точку во времени, корректируя ее прогнозы по мере поступления новых данных. Этот метод полезен для прогнозирования рядов, проявляющих тренд или/и сезонность. Можно выбрать одну из множества моделей экспоненциального сглаживания, отличающихся рассмотрением трендов и сезонности.

Модели **АРПСС** предоставляют более сложные методы для моделирования трендовых и сезонных компонентов, чем модели экспоненциального сглаживания, и в частности позволяют (в качестве дополнительного преимущества) включать в модель независимые (предикторные) переменные. При этом непосредственно задаются порядки авторегрессии и скользящего среднего, а также порядок исчисления разностей. Можно включить предикторные переменные и определить для любых из них или всех их передаточные функции, а также задать автоматическое обнаружение выбросов или явно задать набор выбросов.

Примечание: С практической точки зрения модели ARIMA наиболее полезны, если вы хотите включить в рассмотрение предикторы, помогающие объяснить поведение прогнозируемого ряда, например, количество отправленных по почте каталогов или число посещений Web-страницы компании. Модели экспоненциального сглаживания описывают поведение временного ряда, не пытаясь понять, с чем такое поведение связано. Например, ряд, хронологически достигающий пиковых значений каждые 12 месяцев, будет, скорее всего, продолжать так делать, даже если вы не знаете, почему.

Также доступен **Эксперт построения моделей**, который пытается автоматически идентифицировать и оценить наиболее подходящую модель АРПСС или экспоненциального сглаживания для одной или нескольких переменных назначения, устраняя тем самым необходимость определения подходящей модели методом проб и ошибок. В сомнительных ситуациях следует использовать эксперт построения моделей.

Если заданы предикторные переменные, эксперт построения моделей выбирает для включения в модели АРПСС те переменные, у которых есть статистически значимая взаимосвязь с зависимым рядом. В подходящих случаях переменные модели преобразуются с использованием исчисления разностей и/или функционального преобразования (квадратный корень или натуральный логарифм). По умолчанию эксперт построения моделей рассматривает модели экспоненциального сглаживания и все модели АРПСС и выбирает лучшую из них для каждого поля назначения. Однако эксперт построения моделей можно ограничить выбором только лучших из моделей экспоненциального сглаживания или из моделей АРПСС. Можно задать также автоматическое обнаружение выбросов.

Пример. Поставщику услуг широкополосного доступа в стране требуется аналитик для прогнозов подписок пользователей с целью предсказания использования пропускной способности. Прогнозы требуются для каждого из локальных рынков, составляющих базу подписчиков по стране. Вы можете, применив моделирование временных рядов, составить прогнозы на следующие три месяца для ряда локальных рынков.

Технические требования

Узел временных рядов отличается от остальных узлов IBM SPSS Modeler тем, что его нельзя просто вставить его в поток, а затем вызвать этот поток. Узлу временных рядов всегда должен предшествовать узел интервалов времени, задающий такую информацию, как интервал времени для использования (года, кварталы, месяцы и так далее), данные, используемые для оценки, и насколько далеко в будущее расширять прогноз (в случае его использования).

Временные ряды должны быть эквидистантны. Методам для моделирования данных временных рядов требуется универсальный интервал между измерениями, со всеми пропущенными значениями, указанными пустыми строками. Если данные еще не отвечают этому требованию, узел интервалов времени сможет преобразовать значения должным образом.

Другие моменты, которые следует отметить в связи с узлами временных рядов:

- Поля должны быть числовыми.
- Поля дат нельзя использовать в качестве входных.
- Разделы игнорируются.

Опции полей

На вкладке Поля можно задать поля для использования при построении модели. Перед построением модели необходимо указать поля, которые должны служить полями назначения и входными полями. Обычно узел временных рядов использует информацию о полях с узла типа восходящего потока. Если вы используете узел Тип для выбора входных полей и полей назначения, на этой вкладке можно ничего не менять.

Использовать параметры узла типа. Эта опция указывает узлу, что следует использовать информацию о полях из узла Тип, расположенного выше. Это опция по умолчанию.

Использовать пользовательские параметры. Эта опция указывает узлу, что следует использовать информацию о полях, заданную здесь, а не ту, что задана на расположенных выше узлах Тип. После выбора этого варианта задайте поля. Имейте в виду, что поля, хранимые как поля дат, ни как поля назначения, ни как входные поля не принимаются.

- **Поля назначения.** Выберите одно или несколько полей назначения. Это аналогично заданию для поля роли *Поле назначения* на узле Тип. У полей назначения для модели временного ряда должна быть шкала измерений *Непрерывная*. Для каждого поля назначения строится отдельная модель. Поле назначения рассматривает все заданные *Входные* поля, кроме самого себя, как возможные входные. Таким образом, одно и то же поле может быть включено в оба списка; так, поле будет использоваться как возможное входное для всех моделей, кроме одной, где оно является полем назначения.
- **Поля ввода.** Выберите входные поля. Это аналогично заданию для поля роли *Входное* на узле Тип. Входные поля для модели временного ряда должны быть числовыми.

Опции модели временных рядов

Имя модели. Задает имя модели, сгенерированной при выполнении узла.

- **Авто.** Автоматически генерирует имя модели на основании имен полей назначения или ID/имени типа модели в случаях, когда не задано назначение (например, в моделях кластеризации).
- **Пользовательская.** Позволяет создать пользовательское имя для слепка модели.

Продолжать оценку с использованием существующих моделей. Если вы уже сгенерировали модель временного ряда, выберите эту опцию, чтобы повторно использоваться параметры критериев, заданных для данной модели, и сгенерировать на палитре моделей новый узел модели вместо того, чтобы выполнять построение новой модели с начала. Этим способом можно сэкономить время, благодаря повторной оценке и созданию нового прогноза на основе тех же параметров модели, что и прежде, но с применением новых данных. Таким образом, например, если исходная модель для конкретного временного ряда была моделью линейного тренда Хольта, этот же тип модели будет использоваться для повторной оценки и прогнозирования для таких данных; система не будет снова пытаться найти лучший тип модели для новых данных. При выборе этой опции элементы управления **Метод** и **Критерии** отключаются. Дополнительную информацию смотрите в разделе “Повторная оценка и прогнозирование” на стр. 263.

Метод. Можно выбрать Эксперт построения моделей, Экспоненциальное сглаживание или АРПСС. Дополнительную информацию смотрите в разделе “Узел моделирования временных рядов” на стр. 255. Выберите **Критерии**, чтобы задать опции для выбранного метода.

- **Эксперт построения моделей.** Выберите эту опцию, чтобы использовать эксперт построения моделей, который автоматически находит наиболее подходящую модель для каждого зависимого ряда.
- **Экспоненциальное сглаживание.** Используйте эту опцию для задания пользовательской модели экспоненциального сглаживания.
- **АРПСС.** Используйте эту опцию для задания пользовательской модели АРПСС.

Информация об интервалах времени

Этот раздел диалогового окна содержит информацию о спецификациях для оценок и прогнозов, выполняемых на узле интервалов времени. Имейте в виду, что если выбрана опция **Продолжать оценку с использованием существующих моделей**, этот раздел не выводится.

В первой строке этой информации указывается, исключаются ли какие-либо записи из модели или используются как контрольные периоды.

Во второй строке предоставляется информация о всех периодах прогнозов, заданных на узле интервалов времени.

Если первая строка читается как **Не определен временной интервал**, это говорит о том, что узел интервалов времени не присоединен. Эта ситуация при попытке запустить поток приведет к ошибке; нужно будет включить с узла временных рядов восходящий поток узла интервалов времени.

Прочая информация

Ширина границ доверительного интервала (%). Доверительные интервалы вычисляются для предсказаний моделей и автокорреляций остатков. Вы можете указать любое положительное число, меньшее 100. По умолчанию используется 95%-ный доверительный интервал.

Максимальное число лагов в назначении вывода АКФ и ЧАКФ:. Вы можете задать максимальное количество лагов, показываемых в таблицах и графиках автокорреляций и частных автокорреляций.

Построить только модель для скоринга. Включите этот переключатель, чтобы сократить объем данных, хранящихся в модели. Эта операция может повысить производительность при построении моделей с очень большим числом временных рядов (с десятками тысяч). Если выбрана эта опция, вкладки Модель, Параметры и Остатки не выводятся для слепка модели временных рядов, но данные все равно можно оценить обычным способом.

Критерии эксперта построения моделей временных рядов

Тип модели. Доступны следующие параметры:

- **Все модели.** Эксперт создания моделей рассматривает и модели АРПСС, и модели экспоненциального сглаживания.
- **Только модели экспоненциального сглаживания.** Эксперт создания моделей рассматривает только модели экспоненциального сглаживания.
- **Только модели АРПСС.** Эксперт создания моделей рассматривает только модели АРПСС.

Включать в рассмотрение модели сезонности. Эта опция включается только в том случае, если для активного набора данных включена периодичность. Когда выбрана эта опция, эксперт построения моделей рассматривает и сезонные, и несезонные модели. Если эта опция не выбрана, эксперт создания моделей рассматривает только несезонные модели.

События и вмешательства. Позволяет обозначить определенные входные поля как поля событий или вмешательств. Эта опция определяет поле как содержащее данные временных рядов, затрагиваемые событиями (предсказуемыми повторяющимися ситуациями, такими как маркетинговые акции) или вмешательствами (одноразовыми инцидентами, такими как отключение электроэнергии или забастовка сотрудников). Эксперт построения моделей будет рассматривать только простую регрессию и не будет рассматривать произвольные передаточные функции для входных полей, определенных как поля событий или вмешательств.

У входных полей должна быть шкала измерений *Флаговая*, *Номинальная* или *Порядковая*, и они должны быть числовыми (например, для флагового поля должно быть указано, 1/0, а не True/False); только тогда они будут включены в этот список. Дополнительную информацию смотрите в разделе “Импульсы и ступеньки” на стр. 251.

Выбросы

Автоматически обнаруживать выбросы. По умолчанию автоматическое обнаружение выбросов не производится. Выберите эту опцию, чтобы выполнять автоматическое обнаружение выбросов, после чего выберите нужные типы выбросов. Дополнительную информацию смотрите в разделе “Выбросы” на стр. 252.

Критерии экспоненциального сглаживания временных рядов

Тип модели. Модели экспоненциального сглаживания классифицируются как сезонные или несезонные [1]. Сезонные модели доступны, только если периодичность, определенная при помощи узла интервалов времени, сезонная. Типы сезонной периодичности: периоды циклов, года, кварталы, месяцы, дни в неделю, часы в день, минуты в день и секунды в день.

- **Простой.** Эта модель подходит для рядов, в которых отсутствует тренд и сезонность. Единственный релевантный параметр такой модели предназначен для сглаживания уровня ряда. Простая модель экспоненциального сглаживания в наибольшей степени напоминает модель АРПСС с нулевым порядком авторегрессии, единичными порядками дифференцирования и скользящего среднего, и не имеющую константы.
- **Линейный тренд Хольта.** Эта модель подходит для рядов, в которых имеется линейный тренд и отсутствует сезонность. Относящиеся к ней параметры предназначены для сглаживания уровня и тренда, независимого в этой модели. Модель экспоненциального сглаживания Хольта является более общей, чем модель Брауна, но вычисление оценок для нее может занять больше времени в случае длинных рядов. Модель экспоненциального сглаживания Хольта в наибольшей степени напоминает модель АРПСС с нулевым порядком авторегрессии и двумя порядками дифференцирования и скользящего среднего.
- **Линейный тренд Брауна.** Эта модель подходит для рядов, в которых имеется линейный тренд и отсутствует сезонность. Релевантными сглаживающими параметрами для нее являются уровень и тренд, но в данной модели они предполагаются равными. Поэтому модель Брауна представляет собой частный случай модели Холта. Модель экспоненциального сглаживания Брауна в наибольшей степени напоминает модель АРПСС с нулевым порядком авторегрессии, порядками дифференцирования и скользящего среднего, равными двум, для которой коэффициент скользящего среднего для второго лага равен квадрату половины коэффициента для первого лага в квадрате.
- **Демпфированный тренд.** Эта модель подходит для рядов, в которых линейный тренд затухает , а сезонность отсутствует. Ее релевантные параметры предназначены для сглаживания уровня, тренда и скорости затухания тренда. Затухающая модель экспоненциального сглаживания в наибольшей степени напоминает модель АРПСС с единичными порядками авторегрессии и дифференцирования, имеющую порядок скользящего среднего, равный двум.
- **Простая сезонная.** Эта модель подходит для ряда, в котором нет никакого тренда, а сезонная вариация постоянна во времени. Ее релевантные параметры сглаживания - уровень и сезонная составляющая. Сезонная модель экспоненциального сглаживания в наибольшей степени напоминает модель АРПСС с нулевым порядком авторегрессии (единичными порядками несезонного и сезонного дифференцирования) и ненулевые коэффициенты скользящего среднего для лагов 1, p и p+1, где p - число периодов сезонности. Для ежемесячных данных p = 12.
- **Аддитивная Винтера.** Эта модель подходит для ряда с линейным трендом и сезонной вариацией, не меняющейся с течением времени. Ее релевантные параметры сглаживания - уровень, тренд и сезонная составляющая. Сезонная аддитивная модель Винтера в наибольшей степени напоминает модель АРПСС с нулевым порядком авторегрессии (единичными порядками несезонного и сезонного дифференцирования) и ненулевые коэффициенты скользящего среднего для лагов p, где p - число периодов сезонности. Для ежемесячных данных p = 12.

1. Gardner, E. S. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting,* 4, 1-28.

- **Мультипликативная Винтера.** Эта модель подходит для ряда с линейным трендом и сезонной вариацией, изменяющейся с величиной ряда. Ее релевантные параметры сглаживания - уровень, тренд и сезонная составляющая. Мультипликативная модель экспоненциального сглаживания Винтерса не похожа ни на одну из моделей АРПСС.

Преобразование назначения. Можно задать преобразование, выполняемое для каждой зависимой переменной перед ее моделированием. Дополнительную информацию смотрите в разделе “Преобразования рядов” на стр. 254.

- **Нет.** Преобразование не выполняется.
- **Корень квадратный.** Выполняется преобразование Квадратный корень.
- **Натуральный логарифм.** Выполняется преобразование Натуральный логарифм.

Критерии АРПСС временных рядов

Узел временных рядов позволяет построить пользовательские несезонные или сезонные модели АРПСС (другое название - модели Бокса-Дженкинса), с фиксированным набором входных переменных (предикторов) или без него[2] . Для любых или всех входных переменных можно определить передаточные функции, а также задать автоматическое обнаружение выбросов или явно задать набор выбросов.

Все задаваемые переменные явно включаются в модель. В этом состоит отличие от использования эксперта построения моделей, где входные переменные включаются в модель, только если у них есть статистически значимая взаимосвязь с переменной назначения.

Модель

На вкладке Модель можно задать структуру пользовательской модели АРПСС.

Компоненты АРПСС. В соответствующих ячейках сетки Структура введите значения для различных компонентов АРПСС вашей модели. Все значения должны быть неотрицательными целыми числами. Для авторегрессии и компонентов скользящего среднего такое значение представляет максимальный порядок. В модель будут включены все положительные меньшие порядки. Например, если задать значение 2, в модель будут включены порядки 2 и 1. Ячейки в столбце Сезонность включаются только в том случае, если для активного набора данных определена периодичность.

- **Авторегрессия (p).** Количество порядков авторегрессии в модели. Порядки авторегрессии задают, какие предыдущие значения из ряда использовались для предсказания текущих значений. Например, порядок авторегрессии 2 означает, что для предсказания текущего значения использовалось на два периода более раннее значение из ряда.
- **Разность (d).** Задает порядок исчисления разностей, применимый к ряду до оценки моделей. Вычисление разностей необходимо при наличии трендов (ряды с трендами обычно нестационарные, а моделирование АРПСС предполагает стационарность) и используется для удаления этих эффектов. Порядок исчисления разностей соответствует степени тренда ряда - разности первого порядка учитывают линейные тренды, разности второго порядка - квадратичные, и так далее.
- **Скользящее среднее (q).** Количество порядков скользящего среднего в модели. Порядки скользящего среднего задают, как отклонения от среднего значения ряда предыдущих значений используются для предсказания текущих значений. Например, порядки скользящего среднего 1 и 2 указывают, что отклонения от среднего значения ряда для каждого значения за прошлые два периода будут рассматриваться для предсказания текущих значений ряда.

Сезонные порядки. Сезонные компоненты авторегрессии, скользящего среднего и исчисления разностей играют ту же роль, что и их несезонные аналоги. Однако для сезонных порядков на текущие значения ряда влияют предыдущие значения, отделенные одним или несколькими сезонными периодами. Например, для ежемесячных данных (сезонный период 12) сезонный порядок 1 означает, что на текущее значение ряда

2. Box, G. E. P., G. M. Jenkins, and G. C. Reinsel. 1994. *Time series analysis: Forecasting and control*, 3rd ed. Englewood Cliffs, N.J.: Prentice Hall.

влияет значение ряда на 12 периодов ранее текущего. Тем самым для ежемесячных данных сезонный порядок 1 - это то же самое, что несезонный порядок 12.

Преобразование назначения. Можно задать преобразование, выполняемое для каждой переменной назначения перед ее моделированием. Дополнительную информацию смотрите в разделе “Преобразования рядов” на стр. 254.

- **Нет.** Преобразование не выполняется.
- **Корень квадратный.** Выполняется преобразование Квадратный корень.
- **Натуральный логарифм.** Выполняется преобразование Натуральный логарифм.

Включить константу в модель. Включение константы - это стандартный прием, если вы не уверены, что общее среднее значение ряда равно 0. Рекомендуется исключить константу, если применяется исчисление разностей.

Функции передачи

На вкладке Передаточные функции можно определить передаточные функции для каких-либо или всех полей. Передаточные функции позволяют задать способ использования прошлых значений этих полей для прогнозирования будущих значений ряда назначения.

Эта вкладка выводится, только если заданы входные поля (с заданной для ни ролью *Входное*) либо на узле типа, либо на вкладке Поля узла временных рядов (выберите **Использовать пользовательские параметры — Входные**).

В верхнем списке показанные входные поля. Остальная информация в этом диалоговом окне относится конкретно к выбранному входному полю в списке.

Порядки передаточной функции. В соответствующих ячейках сетки Структура введите значения для различных компонентов передаточной функции. Все значения должны быть неотрицательными целыми числами. Для компонентов числителя и знаменателя это значение представляет максимальный порядок. В модель будут включены все положительные меньшие порядки. Кроме этого, порядок 0 всегда включен для компонентов числителя. Например, если задать значение 2 для числителя, в модель будут включены порядки 2, 1 и 0. Если задать 3 для знаменателя, в модель будут включены порядки 3, 2 и 1. Ячейки в столбце Сезонность включаются только в том случае, если для активного набора данных определена периодичность.

Числитель. Порядок числителя передаточной функции задает, какие предыдущие значения из выбранного независимого ряда (предикторы) используются для предсказания значений зависимого ряда. Например, порядок числителя 1 означает, что для предсказания текущего значения каждого зависимого ряда используется значение независимого ряда на один период раньше, а также текущее значение независимого ряда.

Знаменатель. Порядок знаменателя передаточной функции задает, как отклонения от среднего значения ряда для предыдущих значений из выбранного независимого ряда (предикторов) используются для предсказания значений зависимого ряда. Например, порядок знаменателя 1 означает, что отклонения от среднего значения независимого ряда на один период в прошлом рассматриваются для предсказания текущего значения каждого зависимого ряда.

Разность. Задает порядок исчисления разностей, применимый к выбранному независимому ряду (предиктору) до оценки моделей. Вычисление разностей необходимо при наличии трендов и используется для удаления их влияния.

Сезонные порядки. Сезонные числитель, знаменатель и компоненты разностей играют ту же роль, что и их несезонные аналоги. Однако для сезонных порядков на текущие значения ряда влияют предыдущие значения, отделенные одним или несколькими сезонными периодами. Например, для ежемесячных данных

(сезонный период 12) сезонный порядок 1 означает, что на текущее значение ряда влияет значение ряда на 12 периодов ранее текущего. Тем самым для ежемесячных данных сезонный порядок 1 - это то же самое, что несезонный порядок 12.

Задержка. Определение задержки приводит к тому, что влияние входного поля задерживается на заданное число интервалов. Например, если задана задержка 5, значение входного поля в момент времени *t* не влияет на прогноз, пока не пройдет пять периодов (*t* + 5).

Преобразование. Спецификация передаточной функции для набора независимых переменных включает в себя также необязательное преобразование, которое будет выполнено с этими переменными.

- **Нет.** Преобразование не выполняется.
- **Корень квадратный.** Выполняется преобразование Квадратный корень.
- **Натуральный логарифм.** Выполняется преобразование Натуральный логарифм.

Обработка выбросов

На вкладке Выбросы предлагается ряд вариантов обработки выбросов в данных [3].

Не обнаруживать выбросы и не моделировать их. По умолчанию выбросы не обнаруживаются и не моделируются. Выберите эту опцию, чтобы отключить любое обнаружение или моделирование выбросов.

Автоматически обнаруживать выбросы. Выберите эту опцию, чтобы выполнять автоматическое обнаружение выбросов, и выберите один или несколько из показанных типов выбросов.

Типы обнаруживаемых выбросов. Выберите типы выбросы, обнаружение которых вы хотите выполнять. Поддерживаются следующие типы:

- Аддитивные (по умолчанию)
- Сдвиг уровня (по умолчанию)
- Инновационные
- Переходные
- Сезонные аддитивные
- Локальный тренд
- Аддитивная вставка

Дополнительную информацию смотрите в разделе “Выбросы” на стр. 252.

Генерирование моделей временных рядов

В этом разделе дается некоторая общая информация об определенных аспектах генерирования моделей временных рядов:

- Генерирование нескольких моделей
- Использование в прогнозировании моделей временных рядов
- Повторная оценка и прогнозирование

Генерируемый слепок модели описан в отдельной теме. Дополнительную информацию смотрите в разделе “Слепок модели временного ряда” на стр. 263.

Генерирование нескольких моделей

При моделировании временных рядов в IBM SPSS Modeler генерируется одна модель (либо АРПСС, либо экспоненциального сглаживания) для каждого поля назначения. Таким образом. если у вас несколько полей

3. Pena, D., G. C. Tiao, and R. S. Tsay, eds. 2001. *A course in time series analysis*. New York: John Wiley and Sons.

назначения, IBM SPSS Modeler генерирует несколько моделей за одну операцию, экономя время и позволяя вам сравнить значения параметров для каждой модели.

Если вы хотите сравнить модель АРПСС и модель экспоненциального сглаживания для одного и того же пол назначения, можно выполнить отдельные вызовы узла временных рядов, задавая каждый раз отличающуюся модель.

Использование в прогнозировании моделей временных рядов

В операции построения временных рядов при помощи особого ряда упорядоченных наблюдений (называемого интервалом оценки) создается модель, которую можно использовать для прогноза будущих значений ряда. Эта модель содержит информацию о промежутке времени, куда попадает указанный интервал. Для прогнозирования при помощи этой модели одна и так же информация о промежутке времени и интервале должна использоваться с одним и тем же временным рядом как для переменной назначения, так и для предикторных переменных.

Предположим, что в начале января вы захотели спрогнозировать продажи Продукта 1 за первые три месяца данного года. Вы строите модель при помощи данных о фактических продажах для Продукта 1 с января по декабрь прошлого года (который мы назовем Годом 1), задав интервал времени "Месяцы". После этого вы можете при помощи этой модели спрогнозировать продажи Продукта 1 за первые три месяца Года 2.

Фактически прогноз возможен на любое число месяцев вперед, но разумеется, чем дальше в будущее вы попытаетесь предсказать, тем меньше будет эффективность модели. Однако прогноз на первые три недели Года 2 невозможен, поскольку для построения модели использовался интервал "Месяцы". Кроме того, эту модель нет смысла использовать для предсказания продаж Продукта 2 - модель временных рядов уместна только для данных, которые использовались для ее определения.

Повторная оценка и прогнозирование

Период оценки жестко кодируется в генерируемую модель. Это означает, что при применении текущей модели к новым данным все значения вне периода оценки игнорируются. Таким образом, модель временного ряда должна оцениваться каждый раз, когда становятся доступны новые данные, в отличие от других моделей IBM SPSS Modeler (которые можно применять повторно для целей скоринга).

Продолжая предыдущий пример, предположим, что на начало апреля Года 2 у вас есть данные о фактических ежемесячных продажах с января по март. Однако если вы повторно примените модель, сгенерированную в начале января, она снова даст прогноз на период с января по март и проигнорирует данные об известных продажах за этот период.

Решением является генерирование новой модели на основе обновленных фактических данных. При допущении, что вы не изменяете параметры прогнозирования, новую модель можно использовать для прогноза на следующие три месяца, с апреля по июнь. Если все еще есть доступ к потоку, использовавшемуся для генерирования исходной модели, можно просто заменить ссылку на исходный файл в этом потоке на ссылку к файлу, содержащему обновленные данные, и вызвать поток повторно, чтобы сгенерировать новую модель. Но если все, что у вас есть - это исходная модель, сохраненная в файле, ее все равно можно использовать для генерирования узла временных рядов, который можно затем добавить в новый поток, содержащий ссылку на обновленный исходный файл. При условии, что этот новый поток предшествует узлу временных рядов с узлом интервалов времени, где задан интервал "Месяцы", при вызове этого нового потока будет сгенерирована требуемая новая модель.

Слепок модели временного ряда

Операция моделирования временных рядов создает несколько новых полей с префиксом $TS-; они показаны в следующей таблице.

Таблица 23. Новые поля, создаваемые операцией моделирования временных рядов.

Имя поля	Описание

Таблица 23. Новые поля, создаваемые операцией моделирования временных рядов (продолжение).

$TS-*имя_столбца*	Значение, прогнозируемое моделью для каждого ряда назначения.
$TSLCI-*имя_столбца*	Нижние доверительные интервалы для каждого спрогнозированного ряда.*
$TSUCI-*имя_столбца*	Верхние доверительные интервалы для каждого спрогнозированного ряда.*
$TSNR-*имя_столбца*	Значение остаточных шумов для каждого столбца данных сгенерированной модели.*
$TS-Итог	Итог по всем значениям $TS-*имя_столбца* для данной строки.
$TSLCI-Итог	Итог по всем значениям $TSLCI-*имя_столбца* для данной строки.*
$TSUCI-Итог	Итог по всем значениям $TSUCI-*имя_столбца* для данной строки.*
$TSNR-Итог	Итог по всем значениям $TSNR-*имя_столбца* для данной строки.*

* Видимость этих полей (например, в выводе с присоединенного узла таблицы) зависит от опций на вкладке Параметры слепка модели временного ряда. Дополнительную информацию смотрите в разделе “Значения параметров моделей временных рядов” на стр. 267.

Слепок модели временного ряда содержит подробности различных моделей, выбранных для каждого ряда, входного для узла построения временных рядов. Входными могут быть несколько рядов (например, с данными о линейках продуктов, регионах или их складах); для каждого ряда назначения генерируется отдельная модель. Например, если доходу в восточном регионе соответствует модель АРПСС, а западному - только модель простого скользящего среднего, каждый регион будет оценен при помощи подходящей для него модели.

В выводе по умолчанию указывается (для каждой построенной модели): тип модели, число заданных предикторов и мера согласия (по умолчанию - стационарный *R*-квадрат). Если заданы методы обработки выбросов, в отдельном столбце выводится число обнаруженных выбросов. В вывод по умолчанию также включаются столбцы для значений Q Льюнга-Бокса, степеней свободы и значимости.

Можно также выбрать расширенный вывод, содержащий следующие дополнительные столбцы:

- R-квадрат
- Среднеквадратичная ошибка (RMSE)
- Средняя абсолютная ошибка в процентах (MAPE)
- Средняя абсолютная ошибка (MAE)
- Максимальная абсолютная ошибка в процентах (MaxAPE)
- Максимальная абсолютная ошибка (MaxAE)
- Норм. BIC (Нормализованный информационный критерий Байеса)

Генерировать. Позволяет сгенерировать узел моделирования временных рядов обратно в поток или слепок модели на палитру.

- **Создать узел моделирования.** Помещает узел моделирования временных рядов в поток с параметрами, при помощи которых был создан данный набор моделей. Эта операция может оказаться полезной, например, если есть поток, в котором вы хотите использовать указанные параметры модели, но больше нет узла моделирования, с помощью которого они были сгенерированы.
- **Модель для палитры .** Помещает слепок модели, содержащий все назначения, в менеджер моделей.

Модель

Рисунок 59. Кнопки Включить все и Выключить все

Переключатели. Выберите, какие модели вы хотите использовать при скоринге. По умолчанию все переключатели включены. Кнопки **Включить все** и **Выключить все** действуют на все переключатели в одной операции.

Сортировать по. Позволяет отсортировать строки вывода по возрастанию или по убыванию для заданного столбца, выводящегося на экран. Опция "Выбранные" сортирует вывод на основе одной или нескольких строк, выбираемых при помощи переключателей. Это может пригодиться, например, для вывода на экран полей назначения с именами "Рынок_1" - "Рынок_9" перед полем "Рынок_10", поскольку при порядке сортировки по умолчанию поле "Рынок_10" выводится сразу же после поля "Рынок_1".

Просмотр. В представлении по умолчанию (простом) выводится базовый набор выходных столбцов. Опция Расширенный выводит дополнительные столбцы для меры согласия.

Число записей, используемых для оценки. Число строк в исходном файле источника данных.

Цель. Одно или несколько полей, определяемых в качестве полей назначения (с ролью *Назначение* на узле типа).

Модель. Тип модели, используемой для данного поля назначения.

Предикторы. Число предикторов (с ролью *Входной*), используемых для данного поля назначения.

Выбросы. Этот столбец выводится, только если вы затребовали (в эксперте построения моделей или критериях АРПСС) автоматическое обнаружение выбросов. Выводимое значение представляет собой число обнаруженных выбросов.

Стационарный R-квадрат . Мера, которая сравнивает стационарную часть модели с простой моделью среднего. Эта мера является более предпочтительной, чем обычный R-квадрат, когда имеется тренд или сезонная вариация. Стационарный R-квадрат может быть отрицательным с диапазоном значений от отрицательной бесконечности до 1. Отрицательные значения означают, что рассматриваемая модель хуже, чем базовая модель. Положительные значения означают, что рассматриваемая модель лучше, чем базовая модель.

R-квадрат. Мера согласия для линейной модели, также называемая коэффициентом детерминации. Равна доле изменчивости зависимой переменной, объясняемой регрессионной моделью. Принимает значения между 0 и 1. Малые значения говорят о том, что модель не адекватно описывает данные.

КСКО. Корень среднего квадрата ошибки. Корень квадратный из Среднего Квадрата Ошибки. Мера того, насколько зависимый ряд отличается от ряда его значений, предсказанных моделью, выраженная в тех же единицах, что и зависимый ряд.

СОМО. Средний относительный модуль ошибки. Мера того, насколько ряд отличается от ряда его значений, предсказанных моделью. Она не зависит от используемых единиц измерения и поэтому может использоваться для сравнения рядов с разными единицами измерения.

СМО. Средний модуль ошибки. Мера того, насколько ряд отличается от ряда его значений, предсказанных моделью. СМО представляется в исходных единицах измерения ряда.

МОМО. Максимальный относительный модуль ошибки. Наибольшая ошибка прогноза, выраженная в процентах. Эта мера полезна, чтобы представить, каким может быть наихудший прогноз.

ММО. Максимальный модуль ошибки. Наибольшая ошибка прогноза, выраженная в тех же единицах измерения, что и зависимый ряд. Как и МОМО, этот показатель полезен, чтобы представить наихудший вариант прогноза. Модуль ошибки и относительный абсолютный процент ошибки могут принимать максимальные значения в разных точках ряда, например, когда модуль ошибки для большого значения ряда

слегка превосходит модуль ошибки для малого значения ряда. В этом случае модуль ошибки достигнет максимума при большем значении ряда, а относительный модуль ошибки - при меньшем значении ряда.

Нормализованный BIC. Нормализованный информационный критерий Байеса. Обычная мера общего согласия модели, которая пытается учесть сложность модели. Это значение, основанное на среднем квадрате ошибке, включает штраф за большое число параметров при недостаточной длине ряда. Этот штраф лишает преимущества модели с большим числом параметров, позволяя с помощью данной статистики легко сравнивать разные модели для одних и тех же рядов.

Q. Статистика Q Льюнга-Бокса. Проверка случайности остаточных ошибок в данной модели.

df. Степени свободы. Число параметров модели, способных варьироваться при оценке конкретного назначения.

Знч. Значение значимости статистики Льюнга-Бокса. Значение значимости меньше 0,05 указывает, что остаточные ошибки не случайны.

Сводные статистики. Этот раздел содержит разнообразные сводные статистики для различных столбцов, включая среднее, минимальное, максимальное и процентное значения.

Параметры моделей временных рядов

Вкладка Параметры содержит подробности различных параметров, использовавшихся для построения выбранной модели.

Вывести параметры для модели. Выберите модель, подробности параметров для которой вы хотите вывести.

Цель. Имя поля назначения с ролью *Назначение*.

Модель. Тип модели, используемой для данного поля назначения.

Поле (только для моделей АРПСС). Содержит по одной записи для каждой из переменных в модели, с назначением на первом месте, за которым следуют предикторы (если они есть).

Преобразование. Указывает, какой был задан тип преобразования (если он есть) для этого поля перед построением модели.

Параметр. Параметр модели, для которого выводятся следующие подробности:

- **Лаг (только для моделей АРПСС).** Указывает лаги (если они есть), рассматриваемые в модели для данного параметра.
- **Оценка.** Оценка параметра. Это значение используется при вычислении значения прогноза и доверительных интервалов для поля назначения.
- **SE.** Среднеквадратичная ошибка оценки параметра.
- **t.** Значение оценки параметра, деленное на среднеквадратичную ошибку.
- **Знч.** Уровень значимости для оценки параметра. Значения больше 0,05 рассматриваются как статистически значимые.

Остатки моделей временных рядов

На вкладке Остатки выводится автокорреляционная функция (АКФ) и частная автокорреляционная функция (ЧАКФ) остатков (разности между ожидаемыми и фактическими значениями) для каждой построенной модели. Дополнительную информацию смотрите в разделе “Функции автокорреляции и частной автокорреляции” на стр. 254.

Вывести график для модели. Выберите модель, для которой вы хотите вывести АКФ и ЧАКФ остатков.

Сводка моделей временных рядов

На вкладке Сводка слепка модели выводится информация о самой модели (*Анализ*), об используемых в ней полях (*Поля*), значениях параметров, используемых при построении модели, (*Параметры построения*) и об обучении модели (*Сводка по обучению*).

При первом просмотре узла результаты вкладки Сводка свернуты. Чтобы увидеть нужные вам результаты, разверните соответствующие им элементы при помощи элемента управления расширением слева от них или выведите все результаты, нажав кнопку **Развернуть все**. Чтобы скрыть результаты после завершения их просмотра, используйте управляющий элемент раскрытия для сворачивания конкретных результатов, которые нужно скрыть, или нажмите кнопку **Свернуть все**, чтобы свернуть все результаты.

Анализ. Выводится информация о конкретной модели.

Поля. Список полей, используемых в качестве полей назначения и входных полей при построении модели.

Параметры компоновки. Содержит информацию об используемых при построении модели параметрах.

Сводная информация по обучению. Выводится тип модели, поток, используемый для ее создания, создавший ее пользователь, отметка времени построения модели и время, затраченное на ее построение.

Значения параметров моделей временных рядов

На кладке Параметры можно указать, какие дополнительные поля будут создаваться операцией моделирования.

Создавать новые поля для каждой оцениваемой модели. Позволяет задать новые поля, которые будут создаваться для каждой оцениваемой модели.

- **Вычислять верхнюю и нижнюю границы доверительных интервалов.** Эта опция, если она включена, создает новые поля (с префиксами по умолчанию $TSLCI- и $TSUCI-) для нижних и верхних доверительных интервалов соответственно, для каждого поля назначения, а также итоги этих значений.
- **Вычислять остаточные шумы.** Эта опция, если она включена, создает новое поле (с префиксом по умолчанию $TSNR-) для остатков модели, для каждого поля назначения, а также итог этих значений.

Глава 14. Самообучаемые модели узла ответа

Узел SLRM

Узел Самообучаемая модель откликов **Self-Learning Response Model** (SLRM) позволяет построить модель, которую можно непрерывно изменять или переоценивать при росте набора данных без необходимости повторного обучения модели с использованием всех данных. Это полезно, например, когда у вас есть несколько товаров и вы хотите определить, какой товар наиболее вероятно приобретут покупатели, если им сделать специальное предложение. Эта модель позволяет предсказать, какие из предложений будут наиболее подходить покупателям и вероятность принятия этих предложений.

Эту модель можно сначала построить с использованием небольшого набора данных со случайно сделанными предложениями и откликами на эти предложения. С ростом набора данных модель может изменяться и поэтому становиться более способной к предсказанию наиболее подходящих предложений для покупателей и вероятностей их принятия на основе таких входных полей, как возраст, пол, работа и доход. Доступные предложения можно изменять, добавляя или удаляя их в диалоговом окне узла без необходимости изменять поле назначения набора данных.

При объединении с IBM SPSS Collaboration and Deployment Services можно задать регулярные автоматические изменения модели. Без необходимости участия человека для наблюдения или каких-то действий этот процесс предоставляет гибкое и недорогое решение для организаций и прикладных программ, когда пользовательское вмешательство аналитика данных невозможно или нежелательно.

Пример. Финансовая компания хочет достичь более выгодных результатов, подбирая предложение, которое с наибольшей вероятностью будет принято каждым клиентом. Можно использовать самообучаемую модель для определения характеристик клиентов, которые наиболее вероятно откликнутся положительно, на основе предыдущих рекламных кампаний и изменять модель в текущем времени на основе самых последних откликов клиентов.

Опции полей узла SLRM

Перед выполнением узла SLRM необходимо на вкладке Поля этого узла задать и поля назначения, и поля откликов назначения.

Поле назначения. Выберите поле назначения из списка; например, номинальное (набор) поле, содержащее различные товары, которые вы хотите предложить клиентам.

Примечание: У поля назначения должна быть строковая система хранения, а не числовая.

Поле отклика назначения . Выберите поле отклика назначения из списка. Например, Принято или Отвергнуто.

Примечание: Это поле должно быть флаговым. Значение true для флага обозначает принятое предложение, а значение false - отвергнутое.

Остальные поля в этом диалоговом окне - стандартные и используемые для всего продукта IBM SPSS Modeler. Дополнительную информацию смотрите в разделе “Моделирование опций полей узла” на стр. 31.

Примечание: Если входные данные включают в себя диапазоны, которые должны использоваться как количественные (числовой диапазон) входные поля, вы должны убедиться, что метаданные включают в себя сведения и о минимальном, и о максимальном значении для каждого диапазона.

Опции моделей узла SLRM

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Продолжить обучение существующей модели. По умолчанию при каждом выполнении узла моделирования создается полностью новая модель. Если выбрана эта опция, обучение продолжается с последней модели, успешно созданной узлом. Это дает возможность скорректировать или обновить существующую модель без необходимости обращаться к исходным данным, что может выполняться значительно быстрее, так как в поток вводятся *только* новые или обновленные записи. Информация по предыдущей модели сохраняется вместе с узлом моделирования, что позволяет использовать этот вариант, даже если предыдущий nugget модели недоступен в потоке или Палитре моделей.

Значения поля назначения. По умолчанию задано **Использовать все**, и это означает, что будет построена модель, содержащая все предложения, связанные с выбранным значением поля назначения. Если вы хотите сгенерировать модель, содержащую только некоторые из предложений поля назначения, щелкните по **Задать** и используйте кнопки **Добавить**, **Изменить** и **Удалить**, чтобы добавить или исправить имена предложений, для которых нужно построить модель. Например, если вы выбираете поле назначение, в котором перечисляются все поставляемые товары, это поле можно использовать для ограничения предлагаемых продуктов только несколькими введенными здесь.

Оценка модели . Поля на этой панели независимы от модели, в которой они не влияют на скоринг. Вместо этого они позволяют вам создать визуальное представление, насколько хорошо модель будет предсказывать результаты.

Примечание: Для вывода результатов оценки в слепке модели нужно включить также переключатель **Показать анализ модели**.

- **Включить оценку модели .** Включите этот переключатель для создания графиков, показывающих предсказанную точность модели для каждого выбранного предложения.
- **Задать начальное значение рандомизации** При оценке точности модели на основании случайной процентной доли эта опция позволяет повторить те же результаты в другом сеансе. Указав начальное значение для генератора псевдослучайных чисел, вы обеспечиваете назначение тех же записей при всяком выполнении узла. Введите нужное начальное значение генератора псевдослучайных чисел. Если эта опция не выбрана, при каждом запуске узла будет генерироваться другая выборка.
- **Размер выборки моделирования .** Задайте количество записей для использования в выборке при оценке модели. Значение по умолчанию - 100.
- **Число итераций.** Эта опция позволяет остановить построение оценки модели после достижения заданного числа итераций. Задайте максимальное количество итераций; значение по умолчанию равно 20.

  Примечание: Имейте в виду, что большие размеры выборки и многочисленность итераций увеличат время построения модели.

Показать оценку модели . Выберите эту опцию, чтобы вывести графическое представление результатов в слепке модели.

Опции параметров узла SLRM

Опции параметров узла служат для точной настройки процесса построения моделей.

Максимальное число предсказаний на запись. Эта опция позволяет ограничить количество предсказаний, сделанных для каждой записи в наборе данных. Значение по умолчанию - 3.

Например, у вас может быть шесть предложений (таких как сберегательный счет, заклад, аренда машины, пенсия, кредитная карта и страховка), но вы хотели бы узнать, какие два из них наилучшие для

рекомендации; в этом случае в данном поле надо задать значение 2. Когда вы построите модель и присоедините ее к таблице, в ней будут показаны два столбца предсказаний (и связанный показатель достоверности для вероятности принимаемого предложения) на каждую запись. Предсказания можно сделать для любого из шести возможных предложений.

Уровень рандомизации. Для предотвращения любых смещений, например, для небольших или неполных наборов данных, и для одинакового рассмотрения всех потенциальных предложений, можно добавить уровень рандомизации к выбору предложений и к вероятности их включения в число рекомендуемых предложений. Рандомизация представляется как процентная доля, выраженная десятичным значением от 0,0 (нет рандомизации) до 1,0 (полностью случайный выбор). Значение по умолчанию - 0,0.

Задать начальное значение рандомизации При добавлении уровня рандомизации к выбору предложения эта опция позволяет вам дублировать те же результаты в другом сеансе. Указав начальное значение для генератора псевдослучайных чисел, вы обеспечиваете назначение тех же записей при всяком выполнении узла. Введите нужное начальное значение генератора псевдослучайных чисел. Если эта опция не выбрана, при каждом запуске узла будет генерироваться другая выборка.

Примечание: При использовании опции **Задать начальное значение генератора псевдослучайных чисел** для записей, читаемых из базы данных, предварительно может потребоваться узел Сортировка для подготовки выборки, чтобы обеспечить одинаковый результат при каждом выполнении узла. Это связано с тем, что начальное значение генератора псевдослучайных чисел зависит от порядка записей, который не обязательно остается постоянным в реляционной базе данных.

Порядок сортировки. Выберите порядок, в котором предложения будут показываться в создаваемой модели:

- **По убыванию.** Модель первыми выводит предложения с максимальной оценкой. Это предложения, у которых максимальная вероятность, что они будут приняты.
- **По возрастанию.** Модель первыми выводит предложения с минимальной оценкой. Это предложения, у которых максимальная вероятность отклония. Например, это может быть полезно при принятии решения, каких клиентов удалить из маркетинговой кампании для конкретного предложения.

Предпочтения для полей назначения. При построении модели могут быть некоторые особенности данных, которые вы хотели бы активно продвигать или удалить. Например, при построении модели, выбирающей лучшее финансовое предложение для продвижения клиенту, у вас может быть желание обеспечить включение хотя бы одного конкретного предложения для каждого клиента независимо от того, насколько хорошо оно оценивается для этого клиента.

Для включения предложения на этой панели и изменения его предпочтений нажмите кнопку **Добавить**, введите название предложения (например, Сберегательный счет или Заклад) и нажмите кнопку **OK**.

- **Значение.** Будет показано название предложения, которое вы добавили.
- **Предпочтение.** Укажите уровень предпочтения, применяемый к предложению. Предпочтение представляется как процентная доля, выраженная десятичным значением от 0,0 (нет предпочтения) до 1,0 (наиболее предпочтительно). Значение по умолчанию - 0,0.
- **Всегда включать.** Включите этот переключатель, чтобы конкретное предложение всегда включалось в предсказание.

  Примечание: Если для **Предпочтения** задано значение 0,0, параметр **Всегда включать** игнорируется.

Принимать во внимание надежность модели. Хорошо структурированная модель с богатыми данными, которая тонко подстроена через несколько повторных генераций, всегда даст более точные результаты при сравнении с новой моделью, использующей мало данных. Включите этот переключатель, чтобы использовать преимущества повышенной надежности более зрелой модели.

Слепки моделей SLRM

Примечание: Если вы задаете на вкладке Опции модели обе опции, и **Включить оценку модели**, и **Вывести анализ модели**, результаты показываются только на этой вкладке.

При запуске потока, содержащего модель SLRM, узел оценивает точность предсказаний для каждого значения поля назначения (предложения) и важность каждого используемого предиктора.

Примечание: Если на узле моделирования на вкладке Модель выбрано **Продолжить обучение существующей модели**, информация на слепке модели обновляется каждый раз, когда модель генерируется заново.

Для построения моделей с помощью IBM SPSS Modeler 12.0 или новее вкладка Модель слепка модели делится на два столбца:

Левый столбец.

- **Просмотр.** Когда у вас есть несколько предложений, выберите то из них, для которого вы хотите вывести результаты.
- **Характеристики модели .** Здесь показывается оцененная точность модели для каждого предложения. Набор тестов генерируется через имитацию.

Правый столбец.

- **Просмотр.** Выберите, какие подробности нужно вывести, **Связь с ответом** или **Важность переменной**.
- **Связь с ответом .** Выводит связь (корреляцию) каждого предиктора с переменной назначения.
- **Важность предиктора .** Обозначает относительную важность каждого предиктора при оценке модели. Обычно вам потребуется направить усилия по моделированию на рассмотрение наиболее важных предикторов, отбрасывая или игнорируя самые не важные. Эту диаграмму можно интерпретировать таким же образом, как для других моделей, которые выводят важность предикторов, хотя в случае SLRM график генерируется через имитацию по алгоритму SLRM. Это делается поочередным удалением каждого предиктора из модели и изучением, как это воздействует на точность модели. Дополнительную информацию смотрите в разделе “Важность предиктора” на стр. 42.

Параметры модели SLRM

На вкладке Параметры слепка модели SLRM задаются опции изменения построенной модели. Например, можно использовать узел модели SLRM для построения нескольких различных моделей при одних и тех же данных и параметрах, а затем использовать эту вкладку в каждой модели, чтобы посмотреть, как повлияет на результаты небольшое изменение параметров.

Примечание: Эта вкладка только доступна после того, как слепок модели добавлен в поток.

Максимальное число предсказаний на запись. Эта опция позволяет ограничить количество предсказаний, сделанных для каждой записи в наборе данных. Значение по умолчанию - 3.

Например, у вас может быть шесть предложений (таких как сберегательный счет, заклад, аренда машины, пенсия, кредитная карта и страховка), но вы хотели бы узнать, какие два из них наилучшие для рекомендации; в этом случае в данном поле надо задать значение 2. Когда вы построите модель и присоедините ее к таблице, в ней будут показаны два столбца предсказаний (и связанный показатель достоверности для вероятности принимаемого предложения) на каждую запись. Предсказания можно сделать для любого из шести возможных предложений.

Уровень рандомизации. Для предотвращения любых смещений, например, для небольших или неполных наборов данных, и для одинакового рассмотрения всех потенциальных предложений, можно добавить уровень рандомизации к выбору предложений и к вероятности их включения в число рекомендуемых предложений. Рандомизация представляется как процентная доля, выраженная десятичным значением от 0,0 (нет рандомизации) до 1,0 (полностью случайный выбор). Значение по умолчанию - 0,0.

Задать начальное значение рандомизации При добавлении уровня рандомизации к выбору предложения эта опция позволяет вам дублировать те же результаты в другом сеансе. Указав начальное значение для генератора псевдослучайных чисел, вы обеспечиваете назначение тех же записей при всяком выполнении узла. Введите нужное начальное значение генератора псевдослучайных чисел. Если эта опция не выбрана, при каждом запуске узла будет генерироваться другая выборка.

Примечание: При использовании опции **Задать начальное значение генератора псевдослучайных чисел** для записей, читаемых из базы данных, предварительно может потребоваться узел Сортировка для подготовки выборки, чтобы обеспечить одинаковый результат при каждом выполнении узла. Это связано с тем, что начальное значение генератора псевдослучайных чисел зависит от порядка записей, который не обязательно остается постоянным в реляционной базе данных.

Порядок сортировки. Выберите порядок, в котором предложения будут показываться в создаваемой модели:

- **По убыванию.** Модель первыми выводит предложения с максимальной оценкой. Это предложения, у которых максимальная вероятность, что они будут приняты.
- **По возрастанию.** Модель первыми выводит предложения с минимальной оценкой. Это предложения, у которых максимальная вероятность отклония. Например, это может быть полезно при принятии решения, каких клиентов удалить из маркетинговой кампании для конкретного предложения.

Предпочтения для полей назначения. При построении модели могут быть некоторые особенности данных, которые вы хотели бы активно продвигать или удалить. Например, при построении модели, выбирающей лучшее финансовое предложение для продвижения клиенту, у вас может быть желание обеспечить включение хотя бы одного конкретного предложения для каждого клиента независимо от того, насколько хорошо оно оценивается для этого клиента.

Для включения предложения на этой панели и изменения его предпочтений нажмите кнопку **Добавить**, введите название предложения (например, Сберегательный счет или Заклад) и нажмите кнопку **OK**.

- **Значение.** Будет показано название предложения, которое вы добавили.
- **Предпочтение.** Укажите уровень предпочтения, применяемый к предложению. Предпочтение представляется как процентная доля, выраженная десятичным значением от 0,0 (нет предпочтения) до 1,0 (наиболее предпочтительно). Значение по умолчанию - 0,0.
- **Всегда включать.** Включите этот переключатель, чтобы конкретное предложение всегда включалось в предсказание.

  Примечание: Если для **Предпочтения** задано значение 0,0, параметр **Всегда включать** игнорируется.

Принимать во внимание надежность модели. Хорошо структурированная модель с богатыми данными, которая тонко подстроена через несколько повторных генераций, всегда даст более точные результаты при сравнении с новой моделью, использующей мало данных. Включите этот переключатель, чтобы использовать преимущества повышенной надежности более зрелой модели.

Глава 15. Модели метода опорных векторов

О модели SVM

Механизм опорных векторов (Support Vector Machine, SVM) - это удобный способ классификации и регрессии, максимизирующий точность предсказаний модели без чрезмерной подгонки данных обучения. В частности, SVM подходит для анализа данных с очень большим числом предикторных полей (например, более тысячи).

SVM применяется во многих дисциплинах, в том числе при управлении взаимосвязями с клиентами (customer relationship management, CRM), при распознавании лиц и в других задачах распознавания образов, в биоинформатике, для извлечении содержания при анализе текстов, для обнаружения несанкционированного проникновения, в предсказании структуры белков и для распознавания голоса и речи.

Как работает SVM

SVM работает, отображая данные в многомерное пространство признаков, так что точки данных можно категоризовать, даже если эти данные другим способом линейно не разделяются. Когда обнаруживается разделитель между категориями, данные преобразуются таким образом, что разделитель можно было изобразить как гиперплоскость. После этого характеристики новых данных можно использовать для предсказания группы, к которой должна принадлежать новая запись.

Рассмотрим, например, следующий рисунок, на котором точки данных попадают в две разные категории.

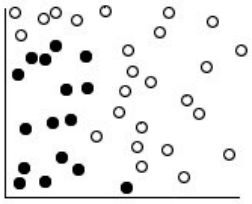


Рисунок 60. Исходный набор данных

Эти две категории можно разделить одной кривой, как показано на следующем рисунке.

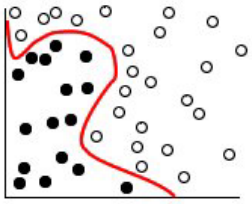


Рисунок 61. Данные с добавленным разделителем

После преобразования границу между двумя категориями можно определить гиперплоскостью, как показано на следующем рисунке.

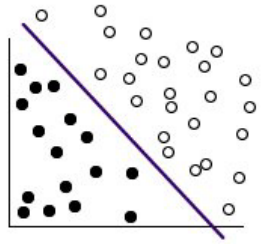


Рисунок 62. Преобразованные данные

Математическая функция, используемая для преобразования, известна как функция **ядра**. SVM в IBM SPSS Modeler поддерживает следующие типы функций ядра:

- Линейный
- Полином.
- Радиальная базисная функция (RBF)
- Сигмоид

Линейная функция ядра рекомендуется, когда проводится непосредственное линейное разделение данных. В других случаях нужно использовать какую-то из других функций. Вам потребуется экспериментировать с разными функциями, чтобы в каждом случае получить лучшую модель, так как каждая из них использует различные аргументы и параметры.

Настройка модели SVM

Кроме разделяющей линии между категориями, модель классификации SVM находит также граничные линии, определяющие пространство между двумя категориями.

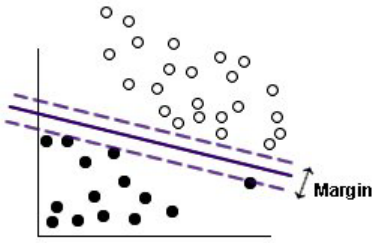


Рисунок 63. Данные, полученные с помощью предварительной модели

Точки данных, лежащие на границах, известны также как **вектора поддержки**.

Чем шире граница между двумя категориями, тем лучше модель покажет себя при предсказании категории для новых записей. В предыдущем примере граница не очень широкая, и о модели говорят, что она **переобучена**. Чтобы расширить границу, можно согласиться с небольшой ошибкой в классификации; пример этого показан на следующем рисунке.

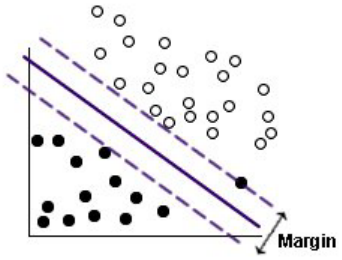


Рисунок 64. Данные, полученные с помощью улучшенной модели

В некоторых случаях линейное разделение более сложное; пример этого показан на следующем рисунке.

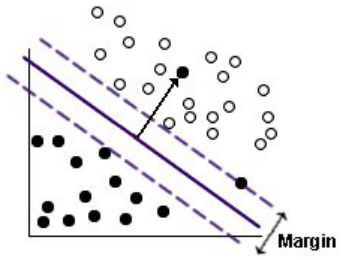


Рисунок 65. Проблема линейного разделения

В аналогичных случаях цель - найти оптимальный баланс между широкой границей и количеством неправильно классифицированных точек данных. У функции ядра есть **параметр регуляризации** (называемый C), который управляет соотношением между этими двумя значениями. Возможно, вам потребуется экспериментировать с разными значениями этого и других параметров ядра, чтобы найти наилучшую модель.

Узел SVM

Узел SVM позволяет использовать механизм опорных векторов для классификации данных. В частности, SVM полезен для использования с широкими наборами данных, то есть с такими, у которых много предикторных полей. Вы можете использовать параметры по умолчанию на узле, чтобы относительно быстро создать базовую модель, или применить параметры эксперта для экспериментирования с различными типами модели SVM.

Когда модель построена, вы можете:

- Просматривать слепок модели для вывода относительной важности входных полей для построении модели.
- Присоедините к слепку модели узел таблицы для просмотра вывода модели.

Пример. В медицинских исследованиях получен набор данных, содержащих характеристики многих образцов человеческих клеток от пациентов, для которых предполагается риск развития рака. Анализ исходных данных показал, что для здоровых и злокачественных клеток многие характеристики существенно отличаются. Медики хотят разработать модель SVM, которая сможет использовать значения характеристик аналогичных клеток в образцах от других пациентов, чтобы получить раннюю диагностику нормальности или злокачественности новых образцов.

Опции моделей узла SVM

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Создание моделей расщепления. Строит отдельную модель для каждого возможного значения входных полей, заданных как поля расщепления. Дополнительную информацию смотрите в разделе “Построение моделей расщепления” на стр. 28.

Опции эксперта узла SVM

Если вы хорошо знакомы с механизмами векторов поддержки, опции эксперта позволят точнее настроить процесс обучения. Для доступа к экспертным опциям выберите режим **Дополнительно** на вкладке Дополнительно.

Добавить все вероятности (допустимо только для категориальных полей назначения). Включение этого переключателя задает, что для каждой обработанной узлом записи выводятся вероятности для каждого возможного значения номинального или флагового поля назначения. Если эта опция не задана, для номинальных и флаговых полей назначения выводится только вероятность предсказанного значения. Настройка этого переключателя определяет состояние по умолчанию для соответствующего переключателя при выводе слепка модели.

Критерий остановки. Определяет, когда остановить алгоритм оптимизации. Диапазон значений - от 1.0E–1 до 1.0E–6, значение по умолчанию 1.0E–3. Уменьшение этого значения приводит к более точной модели, но такую модель нужно будет дольше обучать.

Параметр регуляризации (C). Управляет соотношением между максимизацией границ и минимизацией члена ошибки обучения. Обычно это значение от 1 до 10 включительно, по умолчанию используется 10. Увеличение этого значения повышает точность классификации (или сокращает ошибку регрессии) для данных обучения, но это может привести к переобучению.

Точность регрессии (эпсилон). Используется только в том случае, если уровень измерения поля назначения *Количественный*. Приводит к тому, что приемлемые ошибки принимаются только в том случае, если они меньше заданного здесь значения. Увеличение этого значения ускоряет моделирование, но ценой понижения точности.

Тип ядра. Определяет тип функции ядра, используемой для преобразования. При разных типах ядра разделитель вычисляется разными способами, поэтому рекомендуется испробовать различные опции. По умолчанию используется **RBF** (радиальная базисная функция, Radial Basis Function).

Гамма RBF. Разрешено только в том случае, если для типа ядра задано **RBF**. Обычно это значение должно лежать между 3/*k* и 6/*k*, где *k* - это количество входных полей. Например, если есть 12 входных полей, целесообразно испробовать значения от 0,25 до 0,5. Увеличение этого значения повышает точность классификации (или сокращает ошибку регрессии) для данных обучения, но это может привести также к переобучению.

Гамма. Разрешено только в том случае, если для типа ядра задано **Полиномиальное** или **Сигмоид**. Увеличение этого значения повышает точность классификации (или сокращает ошибку регрессии) для данных обучения, но это может привести также к переобучению.

Смещение. Разрешено только в том случае, если для типа ядра задано **Полиномиальное** или **Сигмоид**. Задает значение `coef0` для функции ядра. В большинстве случаев подходит значение по умолчанию 0.

Степень. Разрешено только в том случае, если для типа ядра задано **Полиномиальное**. Управляет сложностью (размерностью) пространства отображения. Обычно вы не будете использовать значение больше десяти.

Слепок модели SVM

Модель SVM создает несколько новых полей. Наиболее важное из них - это поле **$S-имя_поля**, которое показывает значение поля назначения, предсказанное моделью.

Количество и имена новых полей, созданных моделью, зависят от уровня измерения поля назначения (такое поле обозначается в следующей таблице как *имя_поля*).

Для просмотра этих полей и их значений добавьте узел Таблица в слепок модели SVM и выполните узел Таблица.

Таблица 24. Уровень измерения поля назначения 'Номинальный' или 'Флаговый'

Имя нового поля	Описание
$S-*имя_поля*	Предсказанное значение поля назначения.
$SP-*имя_поля*	Вероятность предсказанного значения.
$SP-*значение*	Вероятность для каждого возможного значения, номинального или флагового (выводится только в том случае, если включен переключатель **Присоединить все вероятности** на вкладке Параметры слепка модели).
$SRP-*значение*	(Только для флаговых полей назначения). Необработанные (SRP) и скорректированные (SAP) оценки склонности, показывающие правдоподобие выхода "true" для поля назначения. Эти оценки выводятся только в том случае, если до генерирования модели были включены соответствующие переключатели на вкладке Анализ узла моделирования SVM. Дополнительную информацию смотрите в разделе “Опции анализа узлов моделирования” на стр. 34.
$SAP-*значение*	

Таблица 25. Уровень измерения поля назначения 'Количественный'

Имя нового поля	Описание
$S-*имя_поля*	Предсказанное значение поля назначения.

Важность предиктора

Диаграмма, обозначающая относительную важность каждого предиктора в оцениваемой модели, может быть дополнительно также показана на вкладке Модель. Обычно вам потребуется направить усилия по моделированию на рассмотрение наиболее важных предикторов, отбрасывая или игнорируя самые не важные. Обратите внимание на то, что эта диаграмма доступна только в том случае, если перед генерированием модели на вкладке Анализ выбрана опция **Вычислять важность предикторов**. Дополнительную информацию смотрите в разделе “Важность предиктора” на стр. 42.

Примечание: Для модели SVM важность предикторов может вычисляться дольше, чем для моделей других типов, и поэтому не выбирается на вкладке Анализ по умолчанию. Выбор этой опции может понизить производительность, особенно для больших наборов данных.

Параметры модели SVM

На вкладке Параметры можно задавать дополнительные поля для вывода при просмотре результатов (например, выполняя узел Таблица, присоединенный к слепку). Воздействие каждой из этих опций можно увидеть, выбрав их и нажав кнопку Предварительный просмотр (прокрутите направо страницу вывода Предварительного просмотра, чтобы увидеть дополнительные поля).

Добавить все вероятности (допустимо только для категориальных полей назначения). Если включена эта опция, для каждой обрабатываемой узлом записи выводятся вероятности для каждого возможного значения номинального или флагового поля назначения. Если эта опция выключена, для номинальных и флаговых полей назначения выводятся только предсказанное значение и его вероятность.

Положение по умолчанию для этого переключателя определяется соответствующим переключателем на узле моделирования.

Вычислить простые оценки склонности. Для моделей с флаговым полем назначения (которое возвращает предсказание Да или Нет) можно затребовать оценку склонности, которая обозначает правдоподобие правильного выходного значения, заданного для поля назначения. Это дополнение к другим предсказанным и доверительным значениям, которые можно сгенерировать при скоринге.

Вычислить скорректированные оценки склонности. Простые оценки склонности основываются только на данных обучения и могут выглядеть чрезмерно оптимистическими из-за тенденции многих моделей чрезмерно точно подгонять эти данные под модель. Корректировка склонностей направлена на компенсацию этого эффекта путем проверки выполнения модели на испытательном и проверочном разделах. Для этой опции требуется, чтобы поле разделения было определено в потоке и скорректированные оценки склонности были включены на узле моделирования до генерирования модели.

Глава 16. Модели ближайших соседей

Узел KNN

Анализ ближайшего сходства представляет собой метод классификации наблюдений на основе сходства наблюдений. Этот метод машинного обучения был разработан в качестве способа распознавания структуры данных при неточном соответствии имеющих структур или наблюдений. Подобные наблюдения близки друг к другу, а непохожие наблюдения, наоборот, удалены друг от друга. Таким образом, дистанция между двумя наблюдениями является критерием их различия.

Близкие друг к другу наблюдения называются “соседи”. Когда представляется новое наблюдение, обозначенное знаком вопроса, вычисляется его расстояние от всех других наблюдений в модели. Определяется классификация наиболее похожих наблюдений (ближайшее сходство) и новое наблюдение помещается в категорию, в которой содержится наибольшее количество ближайшего сходства.

Вы можете указать количество анализируемых ближайших соседей; это значение обозначается *k*. На рисунках ниже показано, каким образом новое наблюдение будет классифицироваться с использованием двух различных значений *k*. Если *k* = 5, новое наблюдение помещается в категорию *1*, поскольку большинство ближайших соседей принадлежит категории *1*. Однако если *k* = 9, новое наблюдение помещается в категорию *0*, поскольку большинство ближайших соседей принадлежит категории *0* .

Анализ ближайшего сходства также может использоваться для вычисления значений для непрерывного целевого объекта. В этой ситуации среднее целевое значение ближайшего сходства используется для получения предсказанного значения для нового наблюдения.

Опции целей узла KNN

Вкладка Цели - это то место, где вы можете выбрать направление своих действий, или строить модель, предсказывающую значение поля назначения в ваших входных данных на основании значений их ближайших соседей, или просто найти ближайших соседей для конкретного интересующего вас наблюдения.

Анализ какого типа вы хотите выполнить?

Предсказать значение поля назначения. Выберите эту опцию, если вы хотите предсказать значение поля назначений по значениям его ближайших соседей.

Только определить ближайших соседей. Выберите эту опцию, если вы хотите только увидеть ближайших соседей конкретного входного поля.

Если вы выбираете только идентификацию ближайших соседей, остальные опции этой вкладки, относящиеся к точности и скорости, отключаются, так как они значимы только для предсказания полей назначения.

Какова ваша цель?

При предсказании поля назначения эта группа опций позволяет вам решить, какие факторы наиболее важны при предсказании поля назначения - скорость, точность или смесь обоих факторов. Как вариант, вы можете сами выбрать настройку параметров.

Если вы выбрали одну из опций - Баланс, Скорость или Точность, алгоритм предварительно выберет наиболее подходящую комбинацию параметров для этой опции. Опытным пользователям может понадобиться переопределить этот выбор; это можно сделать на разных панелях вкладки Параметры.

Сбалансировать скорость и точность. Выбрать наилучшее число соседей в узком диапазоне.

Скорость. Найти фиксированное число соседей.

Точность. Выбирает наилучшее число соседей в большом диапазоне и использует важность предикторов при вычислении расстояний.

Задать особые настройки. Выберите этот параметр, чтобы точно настроить алгоритм на вкладке Параметры.

Примечание: Размер конечной модели KNN, в отличие от большинства других моделей, линейно растет с количеством данных обучения. Если при попытке построить модель KNN появляется сообщение об ошибке "out of memory", попробуйте увеличить максимальную память системы, используемую IBM SPSS Modeler. Для этого выберите

Инструменты > Опции > Системные опции

и введите новый размер в поле **Максимальная память**. Изменения, выполненные в диалоговом окне Системные опции, вступают в силу только после перезагрузки IBM SPSS Modeler.

Параметры узла KNN

Вкладка Параметры - это то место, где вы задаете опции, специфичные для Анализа ближайших соседей. На боковом управляющем элементе слева на экране перечисляются панели, которые можно использовать для задания этих опций.

Модель

На панели Модель предоставляются опции, управляющие тем, как будет строиться модель, например, будут ли использоваться модели разделения и расщепления, будут ли преобразовываться числовые поля, чтобы все они попадали в один диапазон, и как управлять наиболее интересными для вас наблюдениями. Вы можете выбрать также пользовательское имя для модели.

Примечание: Опции **Использовать разделенные данные** и **Использовать метки наблюдений** не могут использовать одно и то же поле.

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Создание моделей расщепления. Строит отдельную модель для каждого возможного значения входных полей, заданных как поля расщепления. Дополнительную информацию смотрите в разделе “Построение моделей расщепления” на стр. 28.

Выбрать поля вручную... По умолчанию узел использует раздел и значения поля разделения (если такое есть) с узла Тип, но вы можете перезаписать здесь эти параметры. Чтобы активировать поля **Раздел** и **Разделения**, перейдите на вкладку **Поля**, выберите там опцию **Использовать пользовательские параметры** и вернитесь обратно.

- **Подмножества.** Это поле позволяет позволяет задать поле, с помощью которого данные будут разделяться на отдельные выборки для этапов обучения, испытания и проверки при построении модели. Сгенерировав модель при помощи одной выборки и испытав ее при помощи другой, можно получить надежный показатель качества обобщения модели на более крупные наборы данных, подобные текущим. Если при помощи узлов Тип или Раздел были определено несколько полей разделов, на вкладке Поля на каждом узле моделирования, где используется разделение, должно быть выбрано одно поле раздела. (Если представлен только один раздел, он будет использоваться автоматически при всяком включении разделения.) Кроме того, учтите, что для применения выбранного раздела в анализе на вкладке Параметры модели для узла должно быть также включено разделение. (Выключение этой опции делает возможным отключение разделения без изменения значений параметров полей.)

- **Разбиения.** Для разбиения моделей выберите поле или поля разбиения. Это аналогично заданию для поля роли *Расщепление* на узле Тип. В качестве полей разбиения можно выделить только поля типа **Флаг**, **Номинал** или **Порядковое**. Поля, выбранные как поля разбиения, нельзя использовать в качестве полей назначения, разделов, частоты, веса или входных полей. Дополнительную информацию смотрите в разделе “Построение моделей расщепления” на стр. 28.

Нормализовать количественные входные поля. Включите этот переключатель, чтобы нормализовать значения для непрерывных входных полей. Нормализованные показатели имеют один и тот же диапазон значений, что может повысить эффективность алгоритма оценивания. Используется скорректированная нормализация: [2*(x–min)/(max–min)]–1. Значения со скорректированной нормализацией лежат между –1 и 1.

Использовать метки наблюдений. Включите этот переключатель, чтобы был доступен выпадающий список для выбора поля, значения в котором будут использоваться как метки для идентификации интересных наблюдений на диаграмме пространства предикторов, диаграмме сходства и диаграмме квадрантов средства просмотра моделей. Для использования в качестве полей меток можно выбрать любое поле с уровнем измерения *Номинальный*, *Порядковый* или *Флаг*. Если поле здесь не выбрано, записи выводятся на диаграммах средства просмотра моделей с идентификацией ближайших соседей по номеру строки в источнике данных. Если после построения модели вы будете манипулировать данными в целом, используйте метки наблюдений, чтобы всякий раз не возвращаться к источнику данных для идентификации наблюдений при выводе.

Определить фокусную запись. Включите этот переключатель, чтобы был доступен выпадающий список для отметки особо интересных входных полей (только для флаговых полей). Если поле выбирается здесь, представляющее это поле точки первоначально выбираются в средстве просмотра моделей, когда модель построена. Выбирать здесь фокусную запись не обязательно; любая точка может временно стать фокусной записью при выборе вручную в средстве просмотра моделей.

Соседи

На панели Соседи есть набор опций, управляющих тем, как вычисляется число ближайших соседей.

Количество ближайших соседей (k). Задайте количество ближайших соседей для конкретного наблюдения. Обратите внимание на то, что использование большего числа соседей необязательно приводит к более точной модели.

Если выбрана цель предсказания поля назначения, есть две возможности:

- **Задать фиксированное k.** Используйте эту опцию, если вы хотите найти фиксированное количество ближайших соседей.
- **Автоматически выбрать k.** Другой вариант - использовать поля **Минимум** и **Максимум**, чтобы задать диапазон значений и разрешить процедуре выбрать "наилучшее" количество соседей в этом диапазоне. Метод определения числа ближайших соседей зависит от того, запрошен ли отбор показателей на вкладке Выбор возможностей.

  Если задействован отбор показателей, то он выполняется для каждого значения *k* в заданном диапазоне, и выбирается *k*, а также набор показателей, дающие наименьший процент ошибок (или наименьшую сумму квадратов ошибок, если целевая переменная является количественной).

  Если отбор показателей не задействован, для выбора “наилучшего” числа соседей используется *V*-слойная перекрестная проверка. Смотрите на панели Перекрестная проверка элемент управления назначением слоев.

Вычисление расстояний. Здесь задается метрика расстояния, используемая в качестве меры сходства наблюдений.

- **Метрика Евклида.** Расстояние между двумя наблюдениями x и y представляет собой квадратный корень из суммы квадратов разностей значений наблюдений по всем измерениям.

- **Метрика городского квартала.** Расстояние между двумя наблюдениями представляет собой сумму абсолютных разностей значений наблюдений по всем измерениям. Эта метрика также называется Манхэттенским расстоянием.

Дополнительно, если цель - это предсказание поля назначения, можно выбрать показатели взвешивания по их нормализованной важности при вычислении расстояний. Важность показателя вычисляется для предиктора как отношение процента ошибок или ошибки в виде суммы квадратов для модели с удаленным рассматриваемым предиктором к проценту ошибок или ошибке в виде суммы квадратов для полной модели. Нормализованная важность вычисляется путем деления значений важностей показателей на одно и то же число, для того чтобы их сумма равнялась 1.

При расчете расстояний взвешивать показатели значениями важности. (Выводится только в том случае, если выбрана цель предсказания поля назначения). Включите этот переключатель, чтобы при вычислении расстояний между соседями использовалась важность предикторов. Затем важность предикторов будет показана в слепке модели и использована в предсказаниях (и таким образом повлияет на скоринг). Дополнительную информацию смотрите в разделе “Важность предиктора” на стр. 42.

Предсказанные значения для количественного поля назначения. (Выводится только в том случае, если выбрана цель предсказания поля назначения). Если задано количественное (числовой диапазон) поле назначения, этим определяется способ вычисления предсказанного значения - на основе среднего значения или значения медианы ближайших соседей.

Отбор показателей

Эта панель активируется только в том случае, если выбрана цель предсказания поля назначения. Она позволяет запросить и задать опции для выбора показателей. По умолчанию при отборе показателей рассматриваются все показатели, однако можно выделить часть показателей для принудительного включения в модель.

Выполнить отбор показателей. Включите этот переключатель для доступа к опциям выбора показателей.

- **Принудительное включение.** Нажмите кнопку средства выбора полей рядом с этим переключателем и выберите один или несколько показателей для принудительного внедрения их в модель.

Критерий остановки. На каждом шаге в модель добавляется тот показатель, добавление которого в модель дает наименьшую ошибку (вычисляемую как процент ошибок для категориальной целевой переменной и как сумму квадратов ошибок для количественной целевой переменной). Отбор включением продолжается до тех пор, пока не выполнится заданное условие.

- **Остановиться, когда отобрано заданное количество показателей.** Алгоритм отбирает фиксированное число показателей в дополнение к тем, которые принудительно включаются в модель. Задайте целое положительное число. Уменьшение числа отбираемых показателей создает более компактную модель, повышая риск упустить важные показатели. Увеличение числа отбираемых показателей приведет к включению всех важных показателей, повышая риск в итоге включить показатели, которые в действительности увеличивают модельную ошибку.
- **Остановиться, когда модуль относительного изменения ошибки не превзойдет заданного минимума.** Алгоритм останавливается, когда значение модуля относительного изменения ошибки указывает на то, что модель нельзя дальше улучшить путем добавления дополнительных показателей. Задайте положительное число. При уменьшении значения минимального изменения появляется тенденция включить больше показателей, при этом возникает риск включить показатели, которые не улучшают заметно качество модели. При увеличении значения минимального изменения появляется тенденция включить меньше показателей, при этом возникает риск потерять показатели, которые важны для модели. “Оптимальное” значение минимального изменения зависит от имеющихся данных и решаемой задачи. Смотрите диаграмму значений ошибок при отборе показателей в выводе, чтобы определить, какие показатели наиболее важны. Дополнительную информацию смотрите в разделе “Журнал ошибок выбора предикторов” на стр. 289.

Перекрестная проверка

Эта панель активируется только в том случае, если выбрана цель предсказания поля назначения. Опции на этой панели управляют использованием или неиспользованием перекрестной проверки при вычислении ближайших соседей.

Кросс-проверка делит выборку на несколько подвыборок (или **сверток**). Затем формируются модели ближайшего сходства с поочередным исключением данных каждой подвыборки. Первая модель создается на основе всех наблюдений, кроме наблюдений из первого слоя выборки, вторая модель создается на основе всех наблюдений, кроме наблюдений из второго слоя выборки, и так далее. Для каждой модели оценивается ошибка путем применения модели к подвыборке, которая была исключена при ее создании. Наилучшее число ближайших соседей - это то, которое дает наименьшую среднюю ошибку по слоям.

Слои для перекрестной проверки. *V*-слойная перекрестная проверка используется для определения наилучшего числа соседей. Она недоступна совместно с отбором показателей по причинам, связанным с эффективностью работы процедуры.

- **Распределить наблюдения по слоям случайным образом.** Задайте число слоев, которое должно использоваться при перекрестной проверке. Процедура случайным образом распределяет наблюдения по слоям, пронумерованным от 1 до *V*, где V - число слоев.
- **Задать начальное значение рандомизации** При оценке точности модели на основании случайной процентной доли распределения по слоям эта опция позволяет повторить те же результаты в другом сеансе. Указав начальное значение для генератора псевдослучайных чисел, вы обеспечиваете назначение тех же записей при всяком выполнении узла. Введите нужное начальное значение генератора псевдослучайных чисел. Если эта опция не выбрана, при каждом запуске узла будет генерироваться другая выборка.
- **Для распределения наблюдений использовать поле.** Задайте числовое поле, которое относит каждое наблюдение в активном наборе данных к некоторому слою. Это поле должно быть числовым и принимать значения от 1 до *V*. Если какие-либо значения в этом диапазоне пропущены, а также если используется любое поле расщепления при действующих моделях расщепления, это приведет к ошибке.

Анализ

Панель Анализ активируется только в том случае, если выбрана цель предсказания поля назначения. Вы можете использовать эту панель для указания, будут ли включены в модель дополнительные переменные, чтобы в ней содержались:

- вероятности для каждого возможного значения поля назначения
- расстояния между наблюдением и его ближайшими соседями
- начальные и настроенные оценки склонностей (только для флаговых полей назначения)

Добавить все вероятности. Если включена эта опция, для каждой обрабатываемой узлом записи выводятся вероятности для каждого возможного значения номинального или флагового поля назначения. Если эта опция выключена, для номинальных и флаговых полей назначения выводятся только предсказанное значение и его вероятность.

Сохранить расстояния между наблюдениями и k ближайшими соседями. Для каждой фокусной записи создается по отдельной переменной для каждого из *k* ближайших соседей (из обучающей выборки) и *k* соответствующих ближайших расстояний.

Оценки склонностей

Оценки склонностей можно включить на узле моделирования и на вкладке Параметры слепка модели. Эта функциональная возможность доступна только при выборе поля назначения флагового типа. Дополнительную информацию смотрите в разделе “Оценки склонностей” на стр. 35.

Вычислить простые оценки склонности. Простые оценки склонности получаются из модели на основе только обучающих данных. Если модель предсказывает значение *true* (будет отклик), склонность совпадает с P, где P - это вероятность предсказания. Если модель предсказывает значение false, склонность вычисляется как (1 – P).

- При выборе этой опции при построении модели оценки склонности будут включены в слепок модели по умолчанию. Однако вы всегда можете включить простые оценки склонности в слепке модели независимо от выбора их на узле моделирования.
- При скоринге модели простые оценки склонности будут добавлены в поле с буквами *RP*, присоединенными к стандартному префиксу. Например, если предсказания содержатся в поле с именем *$R-churn*, именем поля оценки склонности будет *$RRP-churn*.

Вычислить скорректированные оценки склонности. Простые оценки склонности основаны исключительно на оценках модели, которая может быть переобучена, что приведет к чрезмерно оптимистическим оценкам склонности. Скорректированные оценки вносят компенсацию, изучая выполнение модели на испытательном и проверочном разделах и уточняя склонности для улучшения в соответствии с этим оценки.

- Для этого раздела требуется, чтобы в потоке присутствовало допустимое поле раздела.
- В отличие от простых оценок достоверности, скорректированные оценки склонностей нужно вычислять при построении модели; в противном случае они будут недоступны при скоринге слепка модели.
- При скоринге модели скорректированные оценки склонности будут добавлены в поле с буквами *AP*, присоединенными к стандартному префиксу. Например, если предсказания содержатся в поле с именем *$R-churn*, именем поля оценки склонности будет *$RAP-churn*. Скорректированные оценки склонности недоступны для моделей логистической регрессии.
- При вычислении скорректированных оценок склонности испытательный или проверочный раздел, используемый для вычисления, не должен быть сбалансирован. Чтобы исключить это, убедитесь, что выбрана опция **Только сбалансированные данные обучения** на любом вышележащем узле Баланс. Кроме этого, если на вышележащем уровне взята сложная выборка, это может сделать скорректированные оценки склонностей неприемлемыми.
- Скорректированные оценки склонности недоступны для моделей моделей дерева или набора правил с "бустингом". Дополнительную информацию смотрите в разделе “Усиленные модели C5.0” на стр. 111.

Слепок модели KNN

Модель KNN создает несколько новых полей, как показано в следующей таблице. Для просмотра этих полей и их значений добавьте узел Таблица к слепку модели KNN и выполните узел Таблица или нажмите кнопку Предварительный просмотр в слепке.

Таблица 26. Поля модели KNN

Имя нового поля	Описание
$KNN-*имя_поля*	Предсказанное значение поля назначения.
$KNNP-*имя_поля*	Вероятность предсказанного значения.
$KNNP-*значение*	Вероятность каждого возможного значения номинального или флагового поля. Включается только в том случае, если включен переключатель **Присоединить все вероятности** на вкладке Параметры слепка модели.
$KNN-neighbor-*n*	Имя *n*-ого ближайшего соседа фокусной записи. Включается только в том случае, если задано ненулевое значение для поля **Вывести ближайшие** на вкладке Параметры слепка модели.
$KNN-distance-*n*	Относительное расстояние от фокусной записи *n*-ого ближайшего соседа до фокусной записи. Включается только в том случае, если задано ненулевое значение для поля **Вывести ближайшие** на вкладке Параметры слепка модели.

Просмотр модели ближайшего сходства

Представление модели

Представление Модель имеет 2x-панельное окно:

- Первая панель выводит обзорное изображение модели, называемое главным видом.
- Вторая панель выводит изображение одного из двух типов:

  Дополнительное представление модели показывает дополнительную информацию о модели, но не концентрируется на самой модели.

  Связанный вид является видом, демонстрирующим один из элементов модели, когда пользователь углубляется в детали основного вида.

По умолчанию первая панель содержит пространство предикторов, а вторая - диаграмму важности предикторов. Если диаграмма важности предикторов недоступна, то есть на панели Соседи на вкладке Параметры не было выбрано **Взвешивать показатели значениями важности**, то показывается первое доступное представление из выпадающего меню Вид.

Если для какого-то представления нет доступной информации, оно пропускается в выпадающем списке Вид.

Пространство предикторов: Диаграмма пространства предикторов - это интерактивная диаграмма пространства предикторов (или подпространства, если предикторов больше трех). Каждая ось представляет предиктор в модели, а расположение точек на диаграмме показывает значения этих предикторов для наблюдений в обучающей и контрольной группах.

Ключи. Помимо значений предикторов, точки на диаграмме содержат другую информацию.

- Форма показывает, к какой группе принадлежит точка: к обучающей или к контрольной.
- Цвет/оттенок точки показывает значение целевой переменной для данного наблюдения. Различающимися цветами обозначается принадлежность к различным категориям категориальной целевой переменной. Различными оттенками обозначаются различные диапазоны значений непрерывной целевой переменной. Показанное значение для обучающей группы является наблюденным значением; для контрольной группы это предсказанное значение. Если целевая переменная не задана, этот ключ не используется.
- Более жирный контур указывает на то, что наблюдение является фокусным. Фокусные записи показываются соединенными с их *k* ближайшими соседями.

Элементы управления и интерактивность. С помощью ряда управляющих элементов, которые представлены на диаграмме, можно исследовать пространство предикторов.

- Можно выбрать подмножество предикторов, которые будут показаны на диаграмме, а также изменить соответствие между осями и предикторами.
- “Фокусные записи” - это всего лишь точки, выбранные на диаграмме пространства предикторов. Если задана переменная идентификации фокусных записей, то точки, представляющие фокусные записи, изначально будут выделены. Однако любая точка может временно стать фокусной записью, если ее выделить. Применяются “обычный” способ выделения: щелчок по точке выделяет эту точку и снимает выделение всех остальных; щелчок по точке с нажатой клавишей Ctrl добавляет ее к набору выделенных точек. Связанные виды, такие, как Диаграмма сходства, автоматически обновятся в соответствии с выбором наблюдений в пространстве предикторов.
- Можно изменить число ближайших соседей (*k*), выводимых для фокусных записей.
- Наведение указателя мыши на точку вызовет вывод строки-подсказки со значением метки наблюдения или номера, если метки наблюдений не заданы, а также наблюденного и предсказанного значений целевой переменной.
- Кнопка “Сброс” позволяет вернуть пространство предикторов в исходное состояние.

Изменение осей на диаграмме пространства предикторов: Вы можете управлять выбором показателей, которые выводятся на осях диаграммы пространства предикторов.

Чтобы изменить настройку оси:

1. Щелкните по кнопке Режим правки (значок с кисточкой) на панели слева, чтобы выбрать режим правки для пространства предикторов.
2. Измените представление на панели справа (на любое другое). Между двумя главными панелями появится панель **Показать зоны**.
3. Включите переключатель **Показать зоны**.
4. Щелкните по любой точке данных в пространстве предикторов.
5. Чтобы заменить ось на предиктор с тем же типом данных:
   - Перетащите новый предиктор через метку зоны (ту, что с маленькой кнопкой X) заменяемого предиктора.
6. Чтобы заменить ось на предиктор с другим типом данных:
   - На метке зоны заменяемого предиктора щелкните по маленькой кнопке X. Пространство предикторов переключится на двумерное представление.
   - Перетащите новый предиктор через метку зоны **Добавить измерение**.
7. Щелкните по кнопке Режим изучения (значок с наконечником стрелки) на панели слева, чтобы выйти из режима правки.

Важность предикторов: Обычно при моделировании сосредотачивают внимание на наиболее важных предикторах и исключают или игнорируют наименее важные. Это помогает сделать диаграмма важности предикторов, показывая относительную важность каждого предиктора при оценке модели. Поскольку значения важности являются относительными, сумма этих значений для всех показанных предикторов равна 1,0. Важность переменных не связана с точностью модели. Она лишь связана с важностью каждого предиктора для предсказания, а не с точностью этого предсказания.

Расстояния до ближайших соседей: Эта таблица выводит *k* ближайших соседей и расстояния до них только для фокусных записей. Она доступна, если на узле моделирования задана переменная идентификации фокусных записей, и содержит только фокальные записи, идентифицированные этой переменной.

Каждая строка

- Столбец **Фокусная запись** содержит значение переменной меток наблюдения. Если метки наблюдений не заданы, то этот столбец содержит номер наблюдения фокусной записи.
- *i*-тый столбец в группе **Ближайшие соседи** содержит значение переменной меток для *i*-того ближайшего соседа фокусной записи. Если метки наблюдений не заданы, то этот столбец содержит номер *i*-того ближайшего соседа фокусной записи.
- *i*-того столбца в группе **Наименьшие расстояния** содержит расстояние от *i*-того ближайшего соседа до фокусной записи.

Соседи: Эта диаграмма показывает фокусные наблюдения и их *k* ближайших соседей по каждому предиктору, а также целевой переменной. Она доступна, если на диаграмме пространства предикторов выбирается фокусное наблюдение.

Диаграмма соседей связана с пространством предикторов двумя способами.

- Выбранные на диаграмме пространства предикторов (фокусные) наблюдения выводятся вместе с их *k* ближайшими соседями на диаграмме сходства.
- Значение *k*, выбранное на диаграмме пространства предикторов, используется на диаграмме сходства.

Выбор предикторов. Позволяет выбрать предикторы для вывода на диаграмме сходства.

Диаграмма квадрантов: Эта диаграмма выводит фокусные наблюдения и их *k* ближайших соседей на диаграмме рассеяния (или на точечной диаграмме, в зависимости от шкалы измерений целевой переменной) с целевой переменной по оси *y* и количественным предиктором по оси *x*. Диаграмма разбита на панели по предикторам. Она доступна, если задана целевая переменная и на диаграмме пространства предикторов выбирается фокусное наблюдение.

- Для непрерывных переменных проводятся опорные линии через средние значения переменных для обучающей группы.

Выбор предикторов. Позволяет выбрать предикторы для вывода на диаграмме квадрантов.

Журнал ошибок выбора предикторов: Каждая точка на этой диаграмме по оси *y* показывает ошибку (либо долю ошибок, либо ошибку в виде суммы квадратов, в зависимости от шкалы измерений целевой переменной) для модели с предиктором, указанным на оси *x* (и всеми показателями, указанными левее по оси *x*). Эта диаграмма доступна, если заданы целевая переменная и отбор показателей.

Таблица классификации: В этой таблице выводится перекрестная классификация наблюденных и предсказанных значений целевой переменной по группам. Она доступна, если задана переменная назначения и эта переменная - категориальная (флаговая, номинальная или порядковая).

- Строка **Пропущенные** в контрольной группе содержит число наблюдений из этой группы с пропущенными значениями целевой переменной. Для опорной выборки эти наблюдения дают вклад в общий процент, но не в процент правильно классифицированных наблюдений.

Сводка ошибок: Эта таблица доступна, если задана целевая переменная. В ней выводится ошибка модели: сумма квадратов для непрерывной целевой переменной и процент ошибок ((100% – общий процент правильно классифицированных наблюдений) для категориальной целевой переменной.

Параметры модели KNN

На вкладке Параметры можно задавать дополнительные поля для вывода при просмотре результатов (например, выполняя узел Таблица, присоединенный к слепку). Воздействие каждой из этих опций можно увидеть, выбрав их и нажав кнопку Предварительный просмотр (прокрутите направо страницу вывода Предварительного просмотра, чтобы увидеть дополнительные поля).

Добавить все вероятности (допустимо только для категориальных полей назначения). Если включена эта опция, для каждой обрабатываемой узлом записи выводятся вероятности для каждого возможного значения номинального или флагового поля назначения. Если эта опция выключена, для номинальных и флаговых полей назначения выводятся только предсказанное значение и его вероятность.

Положение по умолчанию для этого переключателя определяется соответствующим переключателем на узле моделирования.

Вычислить простые оценки склонности. Для моделей с флаговым полем назначения (которое возвращает предсказание Да или Нет) можно затребовать оценку склонности, которая обозначает правдоподобие правильного выходного значения, заданного для поля назначения. Это дополнение к другим предсказанным и доверительным значениям, которые можно сгенерировать при скоринге.

Вычислить скорректированные оценки склонности. Простые оценки склонности основываются только на данных обучения и могут выглядеть чрезмерно оптимистическими из-за тенденции многих моделей чрезмерно точно подгонять эти данные под модель. Корректировка склонностей направлена на компенсацию этого эффекта путем проверки выполнения модели на испытательном и проверочном разделах. Для этой опции требуется, чтобы поле разделения было определено в потоке и скорректированные оценки склонности были включены на узле моделирования до генерирования модели.

Показать ближайшие. Если задать для этого значения *n*, где *n* - это положительное целое число, в модель включается *n* ближайших соседей фокусной записи, а также их относительные расстояния до фокусной записи.

Уведомления

Эта информация относится к продуктам и сервису, предлагаемым по всему миру.

IBM может не предоставлять в других странах продукты, услуги и аппаратные средства, описанные в данном документе. За информацией о продуктах и услугах, предоставляемых в вашей стране, обращайтесь к местному представителю IBM. Ссылки на продукты, программы или услуги IBM не означают и не предполагают, что можно использовать только указанные продукты, программы или услуги IBM. Разрешается использовать любые функционально эквивалентные продукты, программы или услуги, если при этом не нарушаются права IBM на интеллектуальную собственность. Однако ответственность за оценку и проверку работы любого продукта, программы или сервиса, не произведенного корпорацией IBM, лежит на пользователе.

IBM может располагать патентами или рассматриваемыми заявками на патенты, относящимися к предмету данного документа. Предъявление данного документа не предоставляет какую-либо лицензию на эти патенты. Вы можете послать письменный запрос о лицензии по адресу:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

По поводу лицензий, связанных с использованием наборов двухбайтных символов (DBCS), обращайтесь в отдел интеллектуальной собственности IBM в вашей стране или направьте запрос в письменной форме по адресу:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

Следующий абзац не применяется в Великобритании или в любой другой стране, где подобные заявления противоречат местным законам: INTERNATIONAL BUSINESS MACHINES CORPORATION ПРЕДСТАВЛЯЕТ ДАННУЮ ПУБЛИКАЦИЮ "КАК ЕСТЬ", БЕЗ КАКИХ-ЛИБО ГАРАНТИЙ, КАК ЯВНЫХ, ТАК И ПОДРАЗУМЕВАЕМЫХ, ВКЛЮЧАЯ, НО НЕ ОГРАНИЧИВАЯСЬ ТАКОВЫМИ, ПРЕДПОЛАГАЕМЫЕ ГАРАНТИИ СОБЛЮДЕНИЯ ЧЬИХ-ЛИБО АВТОРСКИХ ПРАВ, ВОЗМОЖНОСТИ КОММЕРЧЕСКОГО ИСПОЛЬЗОВАНИЯ ИЛИ ПРИГОДНОСТИ ДЛЯ КАКИХ-ЛИБО ЦЕЛЕЙ И СООТВЕТСТВИЯ ОПРЕДЕЛЕННОЙ ЦЕЛИ. В некоторых штатах при определенных соглашениях не допускается отказ от выраженных или подразумеваемых гарантий, поэтому данное заявление может к вам не относиться.

Эта информация может содержать технические неточности и типографские ошибки. В представленную здесь информацию периодически вносятся изменения; эти изменения будут включаться в новые издания данной публикации. Фирма IBM может в любое время без уведомления вносить изменения и усовершенствования в продукты и программы, описанные в этой публикации.

Любые приводимые здесь ссылки на web-сайты, не относящиеся к компании IBM, даются исключительно для удобства и ни в коей мере не служат целям поддержки или рекламы этих web-сайтов. Материалы этих Web-сайтов не являются частью данного продукта IBM, и вы можете использовать их только на собственную ответственность.

Любую предоставленную вами информацию IBM может использовать или распространять любым способом, какой сочтет нужным, не беря на себя никаких обязательств по отношению к вам.

Если обладателю лицензии на данную программу понадобятся сведения о возможности: (i) обмена данными между независимо разработанными программами и другими программами (включая данную) и (ii) совместного использования таких данных, он может обратиться по адресу:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
U.S.A.

Такая информация может быть доступна при соответствующих условиях и соглашениях, включая в некоторых случаях взимание платы.

Описанную в данном документе лицензионную программу и все прилагаемые к ней лицензированные материалы IBM предоставляет на основе положений Соглашения между IBM и Заказчиком, Международного Соглашения о Лицензиях на Программы IBM или любого эквивалентного соглашения между IBM и заказчиком.

Любые данные о выполнении, содержащиеся здесь, были определены в контролируемой среде. Поэтому результаты, полученные в других операционных средах, могут существенно отличаться. Некоторые измерения могли быть сделаны на системах в стадии разработки, и поэтому нет гарантии, что соответствующие показатели останутся теми же на общедоступных системах. Более того, некоторые показатели могли быть оценены путем экстраполяции. Реальные результаты могут отличаться. Пользователи этого документа должны проверить приводимые данные в их конкретной среде.

Информация о продуктах, не принадлежащих компании IBM, была получена от поставщиков этих продуктов, из их опубликованных сообщений или других общедоступных источников. Компания IBM не тестировала эти продукты и не может подтвердить правильность их работы, совместимость и другие утверждения, касающиеся продуктов, не принадлежащих компании IBM. Вопросы о возможностях этих продуктов следует направлять их поставщикам.

Все заявления, касающиеся будущих направлений деятельности или намерений корпорации IBM, подвержены изменению или отмене без предупреждения и являются не более чем выражением целей или намерений.

Эти сведения содержат примеры данных и отчетов, используемых в повседневных деловых операциях. Чтобы проиллюстрировать их настолько полно, насколько это возможно, данные примеры включают имена индивидуумов, названия компаний, брендов и продуктов. Все эти имена и названия являются вымышленными, и любое совпадения с названиями и адресами, используемыми реально действующими компаниями, является чисто случайными.

При просмотре данного электронного информационного документа фотографии и цветные иллюстрации могут не показываться.

Товарные знаки

IBM, логотип IBM, и ibm.com являются товарными знаками или зарегистрированными товарными знаками компании International Business Machines Corp., зарегистрированными во многих странах мира. Прочие наименования продуктов и услуг могут быть товарными знаками, принадлежащими IBM или другим компаниям. Текущий список товарных знаков IBM можно найти в Интернете “Copyright and trademark information” по адресу: www.ibm.com/legal/copytrade.shtml.

Intel, логотип Intel, Intel Inside, логотип Intel Inside, Intel Centrino, логотип Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium и Pentium являются товарными знаками или зарегистрированными товарными знаками компании Intel или ее дочерних компаний в Соединенных Штатах и других странах.

Linux является зарегистрированным товарным знаком Linus Torvalds в Соединенных Штатах и других странах.

Microsoft, Windows, Windows NT и логотип Windows являются товарными знаками корпорации Microsoft в Соединенных Штатах и других странах.

UNIX является зарегистрированным товарным знаком The Open Group в Соединенных Штатах и других странах.

Java и все основанные на Java товарные знаки и логотипы - товарные знаки или зарегистрированные товарные знаки Oracle и/или его филиалов.

Другие названия продуктов и услуг могут являться товарными знаками IBM или других компаний.

Глоссарий

A

AICC . Критерий для выбора и сравнения смешанных моделей на основе отрицательного удвоенного (ограниченного) логарифма правдоподобия. Меньшие значения указывают на лучшую модель. Критерий AICC "корректирует" информационный критерий Акаике (AIC) для малых размеров выборок. При увеличении размеров выборок критерий AICC сходится к критерию AIC.

M

Mean . Мера центральной тенденции. Арифметическое среднее; сумма, деленная на число наблюдений.

R

Range . Разность между наибольшим и наименьшим значениями числовой переменной; максимум минус минимум.

R-квадрат . Мера согласия для линейной модели, также называемая коэффициентом детерминации. Равна доле изменчивости зависимой переменной, объясняемой регрессионной моделью. Принимает значения между 0 и 1. Малые значения говорят о том, что модель не адекватно описывает данные.

S

Sum . Сумма или итог для всех значений по всем наблюдениям, имеющим непропущенные значения.

V

V Pao (дискриминантный анализ) . Мера различий между групповыми средними. Также называется следом Лоули-Хотеллинга. На каждом шаге вводится та переменная, которая максимизирует прирост индекса V Pao. Выбрав этот параметр, введите минимальное значение, которое должна иметь переменная, чтобы быть включенной в анализ.

А

Асимметрия . Мера асимметрии распределения. Нормальное распределение симметрично, и для него асимметрия равна 0. Распределение со значимой положительной асимметрией имеет длинный хвост справа. Распределение со значимой отрицательной асимметрией имеет длинный хвост слева. В качестве грубого правила можно сказать, что значение асимметрии, более чем вдвое превышающее ее стандартную ошибку, указывает на наличие асимметрии распределения.

В

Внутригрупповая . Для классификации наблюдений используется объединенная внутригрупповая ковариационная матрица.

Внутригрупповая ковариация . Выводится объединенная внутригрупповая ковариационная матрица, которая может отличаться от общей ковариационной матрицы. Матрица вычисляется путем усреднения отдельных ковариационных матриц для всех групп.

Внутригрупповая корреляция . Выводится объединенная внутригрупповая корреляционная матрица, полученная путем усреднения ковариационных матриц отдельных групп перед вычислением корреляций.

Г

График дожития . Выводит кумулятивную функцию дожития (надежности) в линейном масштабе.

Графики для отдельных групп. Диаграмма рассеяния значений первых двух дискриминантных функций строится для каждой группы в отдельности. Если есть только одна дискриминантная функция, вместо диаграммы рассеяния выводится гистограмма.

Графики комбинированных групп. Строится диаграмма рассеяния значений первых двух дискриминантных функций для наблюдений из всех групп. Если есть только одна дискриминантная функция, вместо диаграммы рассеяния выводится гистограмма.

Д

Дисперсия. Мера дисперсии относительно среднего значения. Равна сумме квадратов отклонений от среднего, деленной на число, на единицу меньшее числа наблюдений. Дисперсия измеряется в единицах, которые равны квадратам единиц измерения самой переменной.

Допустимо. Допустимые наблюдения не содержат ни системных, ни пользовательских пропущенных значений.

Е

Единица минус дожитие. В линейном масштабе выводится график функции, равной 1 минус функция дожития (надежности).

И

Информационный критерий Байеса (BIC). Критерий для выбора и сравнения моделей на основе отрицательного удвоенного логарифма правдоподобия. Меньшие значения указывают на лучшую модель. Критерий BIC также "штрафует" чрезмерно параметризованные модели, но строже, чем AIC.

Использование вероятности F. Переменная вводится в модель, если наблюдаемый уровень значимости ее F-значения меньше заданного порога включения, и исключается, если этот уровень значимости больше порога исключения. Порог включения должен быть меньше порога удаления, они оба должны быть положительными. Если необходимо включить в модель больше переменных, увеличьте порог включения. Чтобы исключить из модели большее число переменных, снизьте порог исключения.

Использование значения F. Переменная вводится в модель, если ее F-значение превышает заданное значение включения, и исключается, если ее F-значение меньше значения исключения. Значение включения должно превосходить значение удаления, оба должны быть положительными. Если необходимо ввести в модель больше переменных, снизьте порог включения. Чтобы исключить из модели большее число переменных, увеличьте порог исключения.

К

Классификация с удалением по одной точке. Каждое наблюдение при анализе классифицируется с помощью функции, полученной по всем остальным наблюдениям, кроме данного. Используется также название "U-метод".

Ковариация. Ненормированная мера связи между двумя переменными, равная сумме попарных произведений отклонений, деленной на N-1.

Ковариация отдельных групп. Для каждой группы выводится отдельная ковариационная матрица.

КСКО. Корень среднего квадрата ошибки. Корень квадратный из Среднего Квадрата Ошибки. Мера того, насколько зависимый ряд отличается от ряда его значений, предсказанных моделью, выраженная в тех же единицах, что и зависимый ряд.

М

М-критерий Бокса. Критерий равенства групповых ковариационных матриц. Если p не значимо, а выборка достаточно велика, то нет достаточных свидетельств того, что матрицы различаются. Этот критерий чувствителен к отклонениям от многомерной нормальности.

Максимум. Наибольшее значение числовой переменной.

Медиана. Значение, выше и ниже которого попадает по половине наблюдений, иначе 50-й процентиль. Если число наблюдений четно, медиана есть арифметическое среднее двух находящихся в середине значений, если выборку упорядочить по убыванию или по возрастанию. Медиана представляет собой меру центральной тенденции, которая нечувствительна к выбросам, в отличие от среднего значения, которое могут исказить несколько экстремально больших или малых значений.

Минимизация лямбды Уилкса. Метод отбора переменных в шаговом дискриминантном анализе, отбирающий переменные для ввода в уравнение на основании того, насколько они уменьшают значение "лямбда" Уилкса. На каждом шаге вводится переменная, минимизирующая это значение.

Минимум. Наименьшее значение числовой переменной.

MMO. Максимальный модуль ошибки. Наибольшая ошибка прогноза, выраженная в тех же единицах измерения, что и зависимый ряд. Как и MOMO, этот показатель полезен, чтобы представить наихудший вариант прогноза. Модуль ошибки и относительный абсолютный процент ошибки могут принимать максимальные значения в разных точках ряда, например, когда модуль ошибки для большого значения ряда слегка превосходит модуль ошибки для малого значения ряда. В этом случае модуль ошибки достигнет максимума при большем значении ряда, а относительный модуль ошибки - при меньшем значении ряда.

Мода. Чаще всего встречающееся значение. Если таких значений несколько, каждое из них является модой.

MOMO. Максимальный относительный модуль ошибки. Наибольшая ошибка прогноза, выраженная в процентах. Эта мера полезна, чтобы представить, каким может быть наихудший прогноз.

Н

Необъясненная изменчивость. На каждом шаге вводится переменная, минимизирующая сумму необъясненной изменчивости между группами.

Нестандартизованные. Выводит нестандартизованные коэффициенты дискриминантной функции.

Нормализованный BIC. Нормализованный информационный критерий Байеса. Обычная мера общего согласия модели, которая пытается учесть сложность модели. Это значение, основанное на среднем квадрате ошибке, включает штраф за большое число параметров при недостаточной длине ряда. Этот штраф лишает преимущества модели с большим числом параметров, позволяя с помощью данной статистики легко сравнивать разные модели для одних и тех же рядов.

О

Общая ковариационная матрица. Выводится ковариационная матрица для всех наблюдений, как если бы они были из одной выборки.

Однофакторный дисперсионный анализ. Проводит однофакторный дисперсионный анализ для проверки гипотезы о равенстве групповых средних для каждой независимой переменной.

Отдельные группы. Для классификации используются ковариационные матрицы для отдельных групп. Так как классификация производится на основе дискриминантных функций, а не на основе исходных переменных, выбор этого параметра не всегда равноценен квадратичной дискриминации.

П

Последовательный Бонферрони. Это процедура Бонферрони последовательного отклонения, которая является значительно менее консервативной в плане отклонения отдельных гипотез, но сохраняет тот же общий уровень значимости.

Последовательный Шидак. Это процедура Шидака последовательного отклонения, которая является значительно менее консервативной в плане отклонения отдельных гипотез, но сохраняющей тот же общий уровень значимости.

P

Расстояние Махаланобиса. Мера того, насколько значения наблюдений для независимых переменных отклоняются от среднего по всем наблюдениям. Большое расстояние Махаланобиса означает, что наблюдение содержит экстремальные значения в одной или более независимых переменных.

Результаты классификации. Числа наблюдений, правильно и неправильно отнесенных к каждой из групп в дискриминантном анализе. Это иногда называют матрицей перекрестной классификации.

Респонденты. Коды для фактической группы, предсказанной группы, апостериорные вероятности и значения дискриминантной функции выводятся для каждого наблюдения.

C

CMO. Средний модуль ошибки. Мера того, насколько ряд отличается от ряда его значений, предсказанных моделью. CMO представляется в исходных единицах измерения ряда.

COMO. Средний относительный модуль ошибки. Мера того, насколько ряд отличается от ряда его значений, предсказанных моделью. Она не зависит от используемых единиц измерения и поэтому может использоваться для сравнения рядов с разными единицами измерения.

Средние. Выводятся общее и групповые средние, а также стандартные отклонения для независимых переменных.

Стандартная ошибка. Мера того, насколько значение статистики критерия меняется от выборки к выборке. Это стандартное отклонение выборочного распределения статистики. Например, стандартная ошибка среднего - это стандартное отклонение выборочных средних.

Стандартная ошибка асимметрии. Отношение асимметрии к ее стандартной ошибке можно использовать как критерий нормальности (то есть, можно отвергнуть нормальность, если это отношение меньше, чем -2, или больше, чем +2). Большое положительное значение асимметрии указывает на длинный правый хвост (распределения); большое отрицательное значение - на длинный левый хвост.

Стандартная ошибка среднего. Мера того, как сильно могут отличаться значения среднего от выборки к выборке, извлекаемых из одного и того же распределения. Можно применять для грубого сравнения наблюденного среднего с гипотетическим значением (то есть можно заключить, что два значения различаются, если отношение их разности к стандартному отклонению меньше -2 или больше +2).

Стандартная ошибка эксцесса. Отношение эксцесса к его стандартной ошибке можно использовать как критерий нормальности (то есть, можно отвергнуть нормальность, если это отношение меньше, чем -2, или больше, чем +2). Большое положительное значение эксцесса указывает, что хвосты распределения длиннее, чем у нормального; отрицательное значение эксцесса указывает на более короткие хвосты (как у равномерного распределения).

стандартное отклонение. Мера разброса вокруг среднего, равная квадратному корню из дисперсии. Стандартное отклонение измеряется в тех же единицах, что и исходная переменная.

Стандартное отклонение. Мера дисперсии вокруг среднего, выраженная в тех же единицах измерения, что и наблюдения. Равна корню квадратному из дисперсии. При нормальном распределении 68% наблюдений укладываются в одно стандартное отклонение от среднего, и 95% - в два стандартных отклонения. Если, например, средний возраст равен 45 годам со стандартным отклонением 10, то 95% наблюдений должны оказаться между 25 и 65 годами при нормальном распределении.

Стационарный R-квадрат. Мера, которая сравнивает стационарную часть модели с простой моделью среднего. Эта мера является более предпочтительной, чем обычный R-квадрат, когда имеется тренд или сезонная вариация. Стационарный R-квадрат может быть отрицательным с диапазоном значений от отрицательной бесконечности до 1. Отрицательные значения означают, что рассматриваемая модель хуже, чем базовая модель. Положительные значения означают, что рассматриваемая модель лучше, чем базовая модель.

Т

Территориальная карта . График, на который нанесены границы, позволяющие отнести наблюдение к группе на основании значений функции. Числа соответствуют группам, по которым распределяют наблюдения. Среднее каждой группы обозначено звездочкой внутри границ этой группы. Если есть только одна дискриминантная функция, диаграмма не выводится.

У

Уникальные . Оцениваются все эффекты одновременно, каждый эффект корректируется по всем остальным эффектам любого вида.

Ф

Фишера . Коэффициенты классифицирующей функции Фишера, которые можно напрямую использовать для классификации. Для каждой группы создается отдельный набор коэффициентов, при этом наблюдение относится к группе, которой соответствует наибольшее значение дискриминантной функции (значение классифицирующей функции).

Функция риска . Выводит функцию накопленного риска в линейном масштабе.

Э

Эксцесс . Мера сгруппированности наблюдений вокруг центральной точки. Для нормального распределения значение эксцесса равно 0. Положительный эксцесс указывает на то, что по отношению к нормальному распределению наблюдения для таких распределений сгруппированы более плотно около центра и имеют более тонкие хвосты до экстремумов распределения, и более толстые хвосты в области экстремальных значений. Отрицательный эксцесс указывает на то, что по отношению к нормальному распределению наблюдения для таких распределений сгруппированы менее плотно около центра и имеют более толстые хвосты до экстремумов распределения, и более тонкие хвосты в области экстремальных значений.

Индекс

Г

Д

З

И

К

Л

М

Н

О

П

P

C

Т

У

Ф

Х

Ч

Э

Напечатано в Дании