# *Руководство по CRISP-DM для IBM SPSS Modeler*

**Примечание**

Прежде чем использовать эту информацию и продукт, описанный в ней, прочтите сведения в разделе “Уведомления” на стр. 39.

**Информация о продукте**

Это издание применимо к версии 18, выпуск 0, модификация 0 IBM SPSS Modeler и ко всем последующим версиям и модификациям до тех пор, пока в новых изданиях не будет указано иное.

# Содержание

# Предисловие

IBM® SPSS Modeler - это IBM Corp. инструментальная среда масштаба предприятия для анализа данных. 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. Введение в CRISP-DM

## Обзор справки CRISP-DM

CRISP-DM (Cross-Industry Standard Process for Data Mining - межотраслевой стандартный процесс для исследования данных) - это проверенный в промышленности инструмент содействия усилиям по исследованию данных.

- **Методологически** он включает в себя описание типичных фаз проекта, задач в составе каждой из этих фаз и объяснение взаимосвязей между существующими задачами.
- Как **модель процесса** CRISP-DM предоставляет обзор жизненного цикла исследования данных.

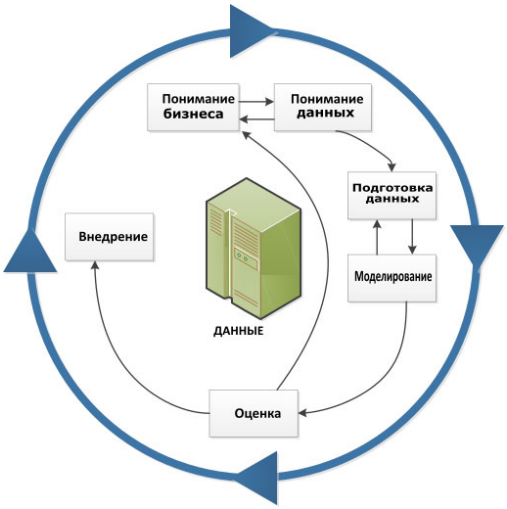


*Рисунок 1. Жизненный цикл исследования данных*

Модель жизненного цикла состоит из шести фаз, а стрелки обозначают наиболее важные и частые зависимости между фазами. Последовательность этих фаз не определена строго. На самом деле, в большинстве проектов приходится возвращаться к предыдущим этапам, а затем снова двигаться вперед.

Гибкую модель CRISP-DM можно легко настраивать. Например, если ваша организация должна обнаруживать факты отмывания денег, вполне вероятно, что вам надо будет проанализировать большой объем данных, не ставя никакой конкретной цели моделирования. Вместо моделирования ваша работа будет направлена на исследование данных и выявление подозрительных структур финансовых данных. CRISP-DM позволяет создавать модель анализа данных, соответствующую вашим определенным потребностям.

В такой ситуации фазы моделирования, оценки и внедрения могут оказаться не столь значимыми, как фазы понимания и подготовки. Однако для целей долгосрочного планирования и будущего исследования данных важно рассмотреть некоторые вопросы, возникающие на всех фазах проекта.

# CRISP-DM в IBM SPSS Modeler

Существует два аспекта использования методологии CRISP-DM в IBM SPSS Modeler для обеспечения уникальной поддержки эффективного анализа данных.

- Инструмент проекта CRISP-DM помогает организовать потоки проекта, выходные данные и аннотации в соответствии с фазами типичного проекта исследования данных. В любой момент при выполнении проекта вы можете создать отчеты, основанные на замечаниях для потоков и фаз CRISP-DM.
- Справка для CRISP-DM сопровождает вас по всему процессу выполнения проекта исследования данных. Справочная система включает в себя списки задач для каждого шага, а также примеры, как работает CRISP-DM в реальном мире. Обратиться к справке CRISP-DM можно, выбрав раздел **Справка CRISP-DM** в главном окне меню Справка.

## Инструмент проектов CRISP-DM

Инструмент проектов CRISP-DM обеспечивает структурный подход к исследованию данных, который может помочь в обеспечении успеха вашего проекта. По существу это расширение стандартного инструмента проектов IBM SPSS Modeler. При работе вы можете переключаться между представлением CRISP-DM и стандартным представлением Классы, чтобы просматривать потоки и выходные данные, организованные по типам или по фазам CRISP-DM.

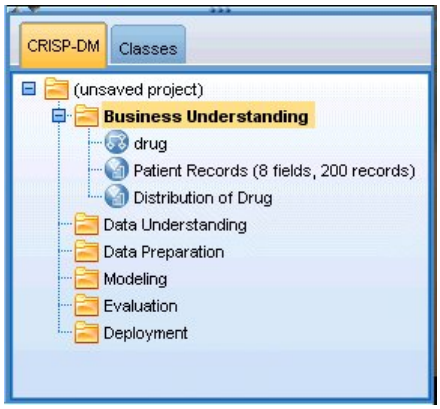


*Рисунок 2. Инструмент проектов CRISP-DM*

Используя представление CRISP-DM инструмента проектов, вы можете:

- Организовывать потоки и выходные данные в соответствии с фазами исследования данных.
- На каждой фазе вставлять замечания о целях вашей организации.
- Создавать пользовательские подсказки для каждой фазе.
- Вставлять замечания о выводах, сделанных на основе конкретной диаграммы или модели.
- Генерировать или изменять отчет HTML для распространения в команде проекта.\

## Справка для CRISP-DM

IBM SPSS Modeler предлагает оперативное руководство для модели стороннего процесса CRISP-DM. Это руководство организовано по фазам процесса и предоставляет следующую поддержку:

- Обзор и список задач для каждой фазы CRISP-DM
- Справку по созданию отчетов при завершении отдельных этапов работы
- Примеры из реальной практики, иллюстрирующие использование CRISP-DM командой проекта для прояснения способов исследования данных
- Ссылки на дополнительные ресурсы по CRISP-DM

Чтобы обратиться к справке по CRISP-DM, в главном окне меню Справка выберите раздел **Справка по CRISP-DM**.

## Дополнительные ресурсы

В дополнение к поддержке для CRISP-DM в IBM SPSS Modeler есть несколько возможностей расширить ваше понимание процессов анализа данных.

- Прочтите руководство пользователя CRISP-DM, созданное консорциумом по CRISP-DM и поставляемое с этим выпуском.
- Прочтите книгу *Data Mining with Confidence* (Исследование данных и достоверность), copyright 2002 by SPSS Inc., ISBN 1-56827-287-1.

# Глава 2. Понимание бизнес-целей

## Обзор изучения бизнеса

Даже не начав работать в IBM SPSS Modeler, нужно уделить время на выяснение выгод, которые ваша организация ожидает получить от исследования данных. Постарайтесь привлечь к обсуждениям этого вопроса как можно больше ключевых лиц и запишите результаты. На окончательном этапе этой фазы CRISP-DM обсуждается, как составить план проекта на основе собранной здесь информации.

Описанное исследование может показаться бесполезным, но это не так. Выяснение бизнес-причин трудозатрат на исследование данных помогает обеспечить взаимопонимание всех членов команды перед расходованием представляющих ценность ресурсов.

## Определение бизнес-целей

Ваша первая задача - попытаться привнести как можно больше уникальной информации в бизнес-цели для исследования данных. Это может оказаться не так легко, как кажется, но позднее вы сможете свести риск к минимуму, внеся ясность в проблемы, цели и ресурсы.

Методика CRISP-DM обеспечивает организованный способ выполнения этой задачи.

Список задач

- Начните со сбора базовой информации о состоянии дел на текущий момент.
- Запишите конкретные бизнес-цели, на которых решили остановиться ответственные за принятие основных решений.
- Согласуйте критерии, на основе которых определяется успешность исследования данных с точки зрения бизнеса.

## Пример розничного Интернет-магазина - Определение бизнес-целей

Сценарий поиска в Web с использованием CRISP-DM

С переходом на продажу через Web все большего числа компаний давно торгующий розничный интернет-магазин компьютеров и электроники сталкивается с растущей конкуренцией со стороны более новых сайтов. Столкнувшись с реальностью, что интернет-магазины возникают настолько же быстро (если не быстрее!), насколько покупатели уходят в Интернет, компания должна изыскать способы сохранения рентабельности, несмотря на растущие затраты на привлечение покупателей. Одно из предлагаемых решений - развитие существующих взаимоотношений с клиентами для максимизации значимости каждого существующего покупателя компании.

Таким образом, изучение дополняется следующими целями:

- Усовершенствование сопутствующих продаж посредством составления улучшенных рекомендаций.
- Повышение лояльности покупателей благодаря более персонализированному обслуживанию.

Ориентировочно изучение будет оценено как успешное, если:

- Сопутствующие продажи увеличатся на 10%.
- Покупатели будут тратить больше времени и просматривать больше страниц на сайте за одно посещение.
- Изучение закончится в срок и в пределах сметы.

## Составление исходной картины бизнеса

Осознание состояния дел организации поможет понять следующие аспекты вашей работы:

- Доступные ресурсы (сотрудники и материалы)
- Проблемы
- Цели

Потребуется провести небольшое исследование текущего состояния дел, чтобы найти реальные ответы на вопросы, которые могут повлиять на конечный результат проекта исследования данных.

Задача 1 - Определение организационной структуры

- Разработайте организационные схемы, иллюстрирующие организационные единицы, отделы и проектные группы. Обязательно включите в них имена и обязанности руководителей.
- Идентифицируйте ключевых сотрудников в организации.
- Выявите внутреннего инвестора, который обеспечит финансовую поддержку и/или опыт работ в нужной сфере знаний.
- Выясните, существует ли организационный комитет, и получите список его членов.
- Определите бизнес-подразделения, которые будут затронуты проектом исследования данных.

Задача 2 - Описание проблемной области.

- Выявите проблемную область, например, расширение рынка сбыта, развитие обслуживания клиентов или привлечение новых покупателей.
- Опишите проблему в общих чертах.
- Классифицируйте предварительные требования к проекту. Какие побудительные факторы заложены в основу проекта? Используется ли уже исследование данных для бизнеса?
- Проверьте состояние проекта исследования данных в рабочей группе. Был ли утвержден объем работ, или исследование данных нужно "прорекламировать" как ключевое техническое решение для рабочей группы?
- В случае необходимости подготовьте для организации информационные наглядные материалы.

Задача 3 - Описание текущего решения

- Опишите все решения, используемые в настоящий момент для устранения экономических проблем.
- Опишите преимущества и недостатки текущего решения. Кроме того, выясните уровень восприятия этого решения в организации.

## Определение бизнес-целей

Здесь ваши наработки обретают конкретность. На основе результатов исследований и совещаний нужно сформулировать первоочередную задачу, согласованную с инвесторами проекта и другими заинтересованными в результатах бизнес-подразделениями. В конечном счете эта цель будет преобразована из некоего расплывчатого понятия, типа "сокращение оттока покупателей", в конкретные задачи исследования данных, которые будут направлять использование методов анализа.

Список задач

Обязательно запишите замечания по следующим пунктам для включения их позднее в план проекта. Не забывайте про поддержание конкретности целей.

- Опишите проблемы, которые вы хотите решить при помощи исследования данных.
- Определите все вопросы бизнеса с максимально возможной точностью.
- Определите все остальные бизнес-требования (такие как лояльность существующих покупателей при повышении потенциальных возможностей сопутствующих продаж).
- Определите ожидаемые преимущества в бизнес-терминах (например, сокращение оттока высокоценных покупателей на 10%).

## Критерии успеха в бизнесе

Далекие цели могут быть ясными, но как понять, где вы сейчас? Перед дальнейшим продвижением важно понять природу успеха в бизнесе для вашего проекта исследования данных. Критерии успеха могут принадлежать к одной из двух категорий:

- **Объективные.** Эти критерии могут быть совсем простыми, например, конкретное повышение точности аудита или заранее заданное сокращение оттока заказчиков.
- **Субъективные.** Субъективные критерии, такие как "обнаружить кластеры эффективной обработки", более сложны для работы с ними, но можно заранее договориться, кто будет принимать окончательное решение об успехе.

Список задач

- Насколько возможно точно задокументируйте критерии успеха для этого проекта.
- Убедитесь, что для каждой бизнес-цели существует соответствующий ей критерий успеха.
- Подберите арбитров для субъективной оценки успешности. По возможности сделайте заметки об их ожиданиях.

---

## Оценка ситуации

Теперь, когда цель ясно определена, пора оценить ваше положение на данный момент. Этот шаг включает в себя получение ответов на вопросы, такие как:

- Какие виды данных доступны для анализа?
- Если ли у вас необходимый для выполнения проекта персонал?
- Каковы самые высокие причастные к делу факторы риска?
- Есть ли у вас для каждого риска план на случай непредвиденных обстоятельств?

## Пример розничного Интернет-магазина - Оценка ситуации

Сценарий поиска в Web с использованием CRISP-DM

Это первая попытка розничного интернет-магазина исследовать данные в Web, и компания решила обратиться к специалисту по исследованию данных, чтобы тот помог начать работу. Одна из первоочередных задач, с которыми сталкивается консультант - это оценка ресурсов компании по исследованию данных.

**Персонал.** Ясно, что существуют штатные специалисты по управлению журналами серверов и базами данных продуктов и покупок, но с небольшим опытом работы с хранилищами данных и очисткой данных для анализа. Поэтому специалист по базам данных тоже может быть консультантом. Поскольку компания надеется, что результаты изучения станут частью непрерывного процесса исследования данных в Web, руководители должны также учитывать, станут ли постоянными какие-либо рабочие места, созданные во время текущих исследований.

**Данные.** Поскольку компания существует давно, у нее большой объем данных веб-журналов и сведений о покупках через Интернет. Фактически для этого начального изучения компания ограничивает анализ покупателями, "зарегистрированными" на сайте. В случае успеха программу можно расширить.

**Риски.** Помимо денежных расходов на консультантов и времени, затраченного сотрудниками на изучение, никакого существенного непосредственного риска в этом проекте не существует. Однако время всегда важно, поэтому данный начальный проект планируется на один отчетный квартальный период.

Кроме того, поскольку на данный момент нет большого притока избыточной денежной наличности, крайне важно, чтобы изучение начиналось в пределах финансовой сметы. Если уложиться в сроки или в бюджет не получается, руководители компании предложат сократить объем работ по проекту.

## Перечень ресурсов

Точный перечень ресурсов - это необходимая часть проекта. Вы избавите себя от потерь времени и головной боли, если реально оцените нужные аппаратные средства, источники данных и проблемы персонала.

Задача 1 - Изучить ресурсы аппаратных средств

- Какие аппаратные средства вам нужно поддерживать?

Задача 2 - Идентифицировать источники данных и склады знаний

- Какие источники данных доступны для исследования данных? Примите во внимание типы и форматы данных.
- Как хранятся данные? Есть ли у вас доступ в реальном времени к хранилищам данных и к работающим базам данных?
- Планируете ли вы приобрести внешние данные, например, демографическую информацию?
- Есть ли проблемы с защитой, мешающие доступу к нужным данным?

Задача 3 - Идентифицировать ресурсы персонала

- Есть ли у вас доступ к экспертам по бизнес-вопросам и по обработке данных?
- Определили ли вы администраторов баз данных и другой технический персонал, который может потребоваться?

Если вы ответили на все эти вопросы, включите список контактных лиц и ресурсов в отчет по этой фазе.

## Требования, предположения и ограничения

Ваши усилия с большей вероятностью принесут пользу, если правильно оценить обязательства для проекта. Все возможные проблемы нужно представить возможно более ясно, чтобы исключить затруднения в дальнейшем.

Задача 1 - Определить требования

Основное требование - это бизнес-цель, обсуждавшаяся ранее, но имейте в виду и следующее:

- Существуют ли юридические ограничения или ограничения безопасности для данных и результатов проекта?
- Назначен ли кто-то следить за требованиями при планировании проекта?
- Есть ли требования к внедрению результатов (например, помещение в базу данных публикаций в Web или считывание оценок в базу данных)?

Задача 2 - Уточнить предположения

- Существуют ли экономические факторы, которые могут повлиять на проект (например, оплата за консультации или конкурирующие продукты)?
- Есть ли какие-либо предположения о качестве данных?
- Каким видится результат для спонсоров или менеджеров проекта? Другими словами, хотят ли они разбираться с моделью или им нужен только результат?

Задача 3 - Проверить ограничения

- Есть ли у вас все пароли для доступа к данным?
- Проверили ли вы юридические ограничения на использование данных?
- Все ли финансовые ограничения описаны в бюджете проекта?

## Риски и непредвиденные обстоятельства

Весьма разумно рассмотреть возможные риски по ходу выполнения проекта. Это могут быть риски следующих типов:

- Планирование - Что будет, если проект займет больше времени, чем хотелось бы?
- Финансы - Что будет, если спонсор столкнется с финансовыми проблемами?
- Данные - Что будет, если данные окажутся недостаточного качества или полноты?
- Результаты - Что будет, если начальные результаты окажутся менее показательными, чем ожидается?

После рассмотрения разнообразных рисков переходите к плану действий при непредвиденных ситуациях, чтобы предотвратить возможные неприятности.

Список задач

- Документируйте каждый возможный риск.
- Документируйте план действий в непредвиденных обстоятельствах для каждого риска.

## Терминология

Чтобы убедиться, что бизнес-команда и группа исследования данных "говорят на одном языке", надо рассмотреть возможность составления глоссария технических терминов и жаргонных слов, которые нуждаются в пояснениях. Например, если у термина "отток заказчиков" в вашем бизнесе есть конкретное и уникальное значение, его стоит явным образом определить в интересах всей команды. Аналогично, команде в целом может быть выгодно прояснение использования диаграммы выигрышей.

Список задач

- Сохраните список терминов и жаргонизмов, которые могут ввести членов команды в заблуждение. Включите сюда и бизнес-терминологию, и термины анализа данных.
- Рассмотрите возможность публикации этого списка во внутренней сети или в другой документации проекта.

## Анализ затрат и результатов

Этот шаг отвечает на вопрос: **"Что мы имеем в итоге?"** В составе окончательной оценки нужно обязательно сравнить затраты на проект с потенциальными преимуществами его успешного выполнения.

Список задач

Включите в анализ сметную стоимость, которую обуславливают:

- Сбор данных и все внешние используемые данные
- Внедрение результатов
- Операционные затраты

Затем оцените выгоду, которую дают:

- Достигнутая основная цель
- Дополнительные аналитические наработки, полученные в результате исследования данных
- Возможные преимущества из-за лучшего понимания данных

---

# Определение целей исследования данных

Теперь, когда бизнес-цель ясна, пора перевести ее в реальность исследования данных. Например бизнес-цель "сокращение оттока заказчиков" можно перевести в цель исследования данных, куда входит:

- Выявление высокоценных покупателей на основе свежих данных о закупках

- Построение модели по доступным данным о покупателях с целью предсказания вероятности оттока для каждого покупателя
- Присвоение каждому покупателю ранга на основе предрасположенности к оттоку и его ценности

Эти цели исследования данных (если они достигнуты) в дальнейшем могут использоваться в бизнесе для сокращения оттока среди значимых покупателей.

Как видите, для эффективного исследования данных бизнес и технология должны работать в тесном взаимодействии. Читайте дальше конкретные советы по определению целей исследования данных.

## Цели исследования данных

Работая с бизнес-аналитиками и аналитиками данных над определением технического решения для бизнес-проблемы, не забывайте придерживаться конкретных понятий.

Список задач

- Опишите **тип** проблемы исследования данных, такой как кластеризация, предсказание или классификация.
- Запишите специализированные цели в конкретных единицах времени, например, предсказания с трехмесячной правомерностью.
- По возможности задавайте для ожидаемых результатов фактические числовые значения, например, полученные оценки оттока для 80% существующих покупателей.

## Пример розничного Интернет-магазина - Цели исследования данных

Сценарий поиска в Web с использованием CRISP-DM

Розничный интернет-магазин с помощью своего консультанта по исследованию данных смог перевести бизнес-цели компании в термины исследования данных. В этом квартале нужно выполнить следующие целевые задачи начального изучения данных:

- При помощи хронологической информации о предыдущих покупках сгенерировать модель, которая связывает "сопутствующие" товары. При просмотре пользователями описания товара предоставьте им ссылки на другие товары в связанной группе (**анализ потребительской корзины**).
- При помощи веб-журналов определите, что пытаются найти покупатели различных типов, а затем перестройте сайт так, чтобы выделить эти позиции. Каждый отдельный "тип" покупателей увидит особую главную страницу для сайта (**профилирование**).
- При помощи веб-журналов попробуйте предсказать, куда человек перейдет дальше, исходя из того, откуда они пришли и где они уже побывали на вашем сайте (**анализ последовательностей**).

## Критерии успешности изучения данных

Успешность нужно также определить в технических терминах, чтобы поддерживать соответствие плану трудозатрат на исследование данных. Исходя из определенной ранее цели исследования данных, сформулируйте эталонные тесты успешности. Предоставляемые IBM SPSS Modeler инструменты, такие как узел Оценка и узел Анализ, помогают проанализировать точность и правомерность результатов.

Список задач

- Опишите методы оценки модели (например, ее точности, производительности и так далее).
- Определите эталонные тесты успешности. Задайте конкретные числовые показатели.
- Определите как можно качественнее субъективные параметры оценки и задайте независимого оценщика успешности.
- Решите, следует ли успешное внедрение результатов модели считать частью понятия успешности исследования данных. Теперь начинайте планирование внедрения.

## Создание плана проекта

Теперь вы готовы создать план для проекта исследования данных. Основой для этого плана будут ответы на заданные вам ранее вопросы и сформулированные цели бизнеса и исследования данных.

## Написание плана проекта

План проекта - это главный документ при всей вашей работе по исследованию данных. При хорошей подготовке он может информировать всех связанных с проектом о его целях, ресурсах, рисках и расписании всех фаз исследования данных. Вам может потребоваться опубликовать этот план во внутренней сети компании вместе со всей документацией, собранной в течение этой фазы.

Список задач

При создании плана убедитесь, что в нем есть ответы на следующие вопросы:

- Обсуждали ли вы задачи проекта и предлагаемый план с кем-либо еще?
- Включены ли оценки времени для всех задач и фаз?
- Включены ли в план трудозатраты и ресурсы для внедрения результатов или бизнес-решения?
- Отмечены ли в плане точки принятия решений и запросы обзоров?
- Отметили ли вы фазы, для которых обычно требуется несколько итераций, например, моделирование?

## Образец плана проекта

Обзорный план для изучения показан в таблице ниже.

*Таблица 1. Обзор образца плана проекта*

| **Фаза** | **Время** | **Ресурсы** | **Риски** |
|---|---|---|---|
| Понимание бизнес-целей | 1 неделя | Все аналитики | Изменение экономической ситуации |
| Начальное изучение данных | 3 недели | Все аналитики | Проблемы с данными, технологические проблемы |
| Подготовка данных | 5 недель | Консультант по исследованию данных, часть времени - аналитики по базам данных | Проблемы с данными, технологические проблемы |
| Моделирование | 2 недели | Консультант по исследованию данных, часть времени - аналитики по базам данных | Технологические проблемы, невозможность подобрать адекватную модель |
| Оценка | 1 неделя | Все аналитики | Изменения экономической ситуации, невозможность реализовать результаты |
| Внедрение | 1 неделя | Консультант по исследованию данных, часть времени - аналитики по базам данных | Изменения экономической ситуации, невозможность реализовать результаты |

## Инструменты и способы оценки

Так как вы уже выбрали IBM SPSS Modeler в качестве инструмента обеспечения успеха анализа данных, можно использовать этот шаг для исследования, какие способы исследования данных наиболее соответствуют потребностям вашего бизнеса. IBM SPSS Modeler предлагает полный набор инструментов для каждой фазы анализа данных. Чтобы решить, когда какие способы использовать, обратитесь к разделу моделирования в оперативной справке.

## Готовы к следующему шагу?

Перед исследованием данных и началом работы с IBM SPSS Modeler обязательно ответьте на следующие вопросы.

С точки зрения бизнеса:

- Какую пользу ваша бизнес-группа надеется извлечь из этого проекта?
- Каким образом вы определите успешное завершение наших исследований?
- Есть ли у вас финансовая смета и ресурсы, необходимые для достижения наших целей?
- Есть ли у вас доступ к данным, необходимым для этого проекта?
- Обсудили ли вы и ваша команда связанные с этим проектом риски и непредвиденные обстоятельства?
- Целесообразен ли этот проект по результатам вашего анализа рентабельности?

Ответив на приведенные выше вопросы, сделали ли вы из ответов выводы для цели исследования данных?

С точки зрения исследования данных:

- Насколько исследование данных может помочь приблизиться к целям вашего бизнеса?
- Если ли у вас представление о том, какие методы исследования данных могут дать оптимальные результаты?
- Каким образом вы узнаете, когда полученные результаты окажутся достаточно точны или эффективны? (*Задали ли мы показатель успеха исследования данных?*)
- Каким образом будут внедрены результаты моделирования? Учли ли вы внедрение в плане проекта?
- Содержит ли план проекта все фазы CRISP-DM?
- Указаны ли в плане риски и зависимости?

Если вы можете ответить на вышеприведенные вопросы "Да", вы готовы к более подробному рассмотрению данных.

# Глава 3. Начальное изучение данных

## Обзор изучения данных

В фазу начального изучения данных CRISP-DM входит тщательное рассмотрение данных, доступных для исследования. Этот шаг критичен для предотвращения неожиданных проблем на протяжении следующей фазы (подготовки данных), обычно представляющей собой самую продолжительную часть проекта.

В фазу начального понимания данных входит оценка данных и их исследование с применением таблиц и графиков, которые можно подготовить в IBM SPSS Modeler при помощи инструмента проектирования CRISP-DM. Это позволяет определить качество данных и описать результаты этих шагов в документации проекта.

## Сбор исходных данных

На данном этапе в CRISP-DM вы готовы обратиться к данным и занести их в IBM SPSS Modeler. Данные поступают из целого ряда источников, таких как:

- **Существующие данные.** В их состав входят самые разнообразные данные, например, транзакционные данные, данные опросов, веб-журналы и так далее. Выясните, достаточно ли существующие данные соответствуют вашим потребностям.
- **Приобретенные данные.** Использует ли ваша организация вспомогательные данные, например, демографические? Если нет, обдумайте, могут ли они потребоваться.
- **Дополнительные данные.** Если вышеупомянутые источники не соответствуют вашим потребностям, может потребоваться провести опросы или начать дополнительное отслеживание для пополнения существующих складов данных.

Список задач

Просмотрите данные в IBM SPSS Modeler и обдумайте следующие вопросы. Обязательно запишите замечания по полученным результатам. Дополнительную информацию смотрите в разделе “Написание отчета о сборе данных” на стр. 14.

- Какие атрибуты (столбцы) из базы данных выглядят наиболее многообещающими?
- Какие атрибуты кажутся ненужными и могут быть исключены?
- Достаточно ли данных для построения обобщаемых заключений или точных предсказаний?
- Не слишком ли много атрибутов для выбранного вами метода моделирования?
- Выполняете ли вы слияние различных источников данных? Если да, то существуют ли области, которые могут представлять собой проблему при слиянии?
- Обдумали ли вы, как обрабатывать пропущенные значения в каждом из используемых источников данных?

## Пример розничного Интернет-магазина - начальный сбор данных

Сценарий поиска в Web с использованием CRISP-DM

В этом примере розничный интернет-магазин использует несколько важных источников данных, в том числе:

**Веб-журналы.** Журналы непосредственного доступа содержат всю информацию о путях и способах навигации покупателей по сайту. Ссылки на файлы изображений и другие неинформативные записи в веб-журналах в процессе подготовки данных потребуется удалить.

**Данные покупок.** При передаче покупателем заказа сохраняется вся относящаяся к этому заказу информация. Заказы в базе данных покупок должны быть отображены на соответствующие сеансы в веб-журналах.

**База данных продуктов.** Атрибуты продуктов могут быть полезны при определении "связанных" продуктов. Информация о продуктах должна быть отображена на соответствующие заказы.

**База данных покупателей.** Эта база данных содержит добавочную информацию от зарегистрированных покупателей. Эти записи далеко не полные, поскольку многие покупатели не заполняют анкеты. Информация о покупателях должна быть отображена на соответствующие покупки и сеансы в веб-журналах.

В настоящее время у компании нет планов на приобретение внешних баз данных и денежные затраты на проведение опросов, поскольку ее аналитики заняты данными, которые есть у компании на текущий момент. Однако в какой-то момент компания может решить рассмотреть расширенное внедрение результатов, и тогда приобретение дополнительных демографических данных для незарегистрированных покупателей может стать весьма полезным. Может также оказаться полезной демографическая информация для понимания отличия базы покупателей розничного интернет-магазина от среднего покупателя через Интернет.

## Написание отчета о сборе данных

Используя полученный на предыдущем шаге материал, можно начать писать отчет о сборе данных. После завершения этот отчет можно разместить на сайте проекта или распространить его по команде. Его можно объединить также с отчетами, которые будут подготовлены на следующих этапах, в том числе с описанием данных, исследований и с результатами проверки качества. Эти отчеты помогут в вашей работе на фазе подготовки данных.

---

## Описание данных

Существует множество способов описания данных, но большинство из них ориентированы на описание количества данных (объема доступных данных) и их состояния. В списке ниже указаны некоторые ключевые характеристики, к которым обращаются при описании данных.

- **Объем данных.** Большинству методов моделирования свойственны преимущества и недостатки, связанные с размером данных. Большие наборы данных могут создавать более точные модели, но при этом растет и время обработки. Продумайте, можно ли использовать подмножество данных. При создании примечаний к заключительному отчету обязательно включите в них статистику размеров для всех наборов данных и не забудьте учесть число записей и число полей (атрибутов) при описании данных.
- **Типы значений.** Данные могут поступать в различных форматах - **числовых**, **категориальных** (строковых) или **логических** (да/нет). Внимание, уделенное типу значений, может предотвратить проблемы при моделировании в дальнейшем.
- **Схемы кодирования.** Значения в базе данных часто являются представлениями характеристик, таких как пол покупателя или тип продукта. Например, для одного набора данных могут использоваться значения *М* и *Ж*, представляющие *мужчин* и *женщин*, а для другого - числовые значения *1* и *2*. Укажите все конфликтующие схемы в отчете о данных.

С этим знанием в руках теперь вы готовы составить отчет с описанием данных и совместно использовать полученные результаты с более широкой аудиторией.

## Пример розничного Интернет-магазина - Описание данных

Сценарий поиска в Web с использованием CRISP-DM

Существует множество записей и атрибутов для обработки в прикладной программе Web-исследования. Даже при том, что розничный интернет-магазин, проводящий этот проект исследования данных, ограничил начальное изучение приблизительно 30 тысячами покупателей, зарегистрированных на сайте, веб-журналы все равно содержат миллионы записей.

Большинство типов значений в этих источниках данных - символические, будь то типы, представляющие даты и времени, посещаемые страницы или отвечающих на вопросы с готовыми вариантами ответа из анкета регистрации. При помощи некоторых из этих переменных будут созданы новые числовые переменные, такие как число посещенных веб-страниц и время, проведенное покупателем на сайте. Некоторые существующие числовые переменные в источнике данных включают в себя номер каждого заказанного продукта, сумму, потраченную при покупке, а также спецификации веса и размера продукта из базы данных продуктов.

В схемах кодирования для различных источников данных существует небольшое перекрытие, поскольку источники данных содержат самые разные атрибуты. Едиными будут только "ключевые" переменные, такие как ID покупателей и коды продуктов. У этих переменных должны быть одинаковые схемы кодирования для разных источников данных; в противном случае слияние источников данных окажется невозможным. С целью перекодирования этих ключевых полей для возможности слияния потребуется некоторая дополнительная подготовка данных.

## Написание отчета с описанием данных

Для эффективного продолжения проекта исследования данных обдумайте важность составления отчета с точным описанием данных по следующим показателям:

Количество данных

- Каков формат данных?
- Определите метод, используемый для захвата данных (например, ODBC).
- Насколько велика база данных (в строках и столбцах)?

Качество данных

- Входят ли в состав данных характеристики, относящиеся к вашему бизнес-вопросу?
- Какие представлены типы данных (символические, числовые и так далее)?
- Вычислили ли вы базовую статистику для ключевых атрибутов? Какую уникальную информацию она дала по вашему бизнес-вопросу?
- Можете ли вы установить приоритеты для нужных атрибутов? Если нет, доступны ли вам бизнес-аналитики, способные предоставить дополнительную аналитическую картину?

---

# Исследование данных

Эта страница CRISP-DM используется для исследования данных при помощи таблиц, диаграмм и других инструментов визуализации, доступных в IBM SPSS Modeler. Такие способы анализа могут помочь решить целевые задачи исследования данных, сформулированные на фазе Понимание бизнес-целей. Они могут также помочь сформулировать гипотезы и оформить задачи по преобразованию данных, выполняемые во время подготовки данных.

## Пример розничного Интернет-магазина - Исследование данных

Сценарий поиска в Web с использованием CRISP-DM

CRISP-DM предлагает на настоящий момент провести начальное исследование, но, как обнаружил наш Интернет-магазин, исследование данных по веб-журналам необработанных данных трудно, если вообще возможно. Как правило, данные веб-журналов сначала должны пройти обработку на фазе подготовки данных, при которой будут получены данные, годные для достоверного исследования. Это отклонение от CRISP-DM подчеркивает тот факт, что указанный процесс может и должен быть настроен для ваших конкретных потребностей исследования данных. Цикличность CRISP-DM предполагает переход исследователей данных между фазами в обоих направлениях.

Хотя веб-журналы и должны быть обработаны перед исследованием, но есть и другие доступные розничному интернет-магазину источники данных, более пригодные для исследования. Использование базы

данных покупок для исследования выявляет содержательные сводки о покупателях, такие как: сколько товаров они приобрели за одну закупку, и откуда они появились? Сводки базы данных покупателей покажут распределение ответов по позициям в анкете регистрации.

Кроме того, исследование полезно и для поиска ошибок в данных. Хотя большинство источников данных и сгенерировано автоматически, но информация в базе данных продуктов вводилась вручную. Некоторые оперативные сводки перечисленных измерений продуктов помогут обнаружить опечатки типа "119-дюймовый" монитор (вместо "19-дюймовый").

## Написание отчета об исследовании данных

По мере создания диаграмм и выполнения статистик по доступным данным начните построение гипотез о том, как данные могут отвечать специализированным и бизнес-целям.

Список задач

Запишите замечания о ваших наработках для включения в отчет об исследовании данных. Обязательно ответьте на следующие вопросы:

- Какие виды гипотез о данных вы построили?
- Какие атрибуты кажутся подходящими для дальнейшего анализа?
- Выявили ли ваши исследования новые характеристики данных?
- Как эти исследования изменили начальную гипотезу?
- Можете ли вы выявить конкретные подмножества данных для дальнейшего использования?
- Просмотрите еще раз цели исследования данных. Изменило ли их это исследование данных?

---

## Проверка качества данных

Данные редко бывают идеальными. Фактически, в большинстве данных содержатся ошибки кодирования, пропущенные значения или несогласованности другого типа, которые иногда сильно затрудняют анализ. Один из способов исключить возможные проколы - провести подробный анализ качества доступных данных, прежде чем переходить к моделированию.

Инструменты составления отчетов в IBM SPSS Modeler (такие как Аудит данных, Таблица и другие узлы выходных данных) могут помочь найти проблемы следующего типа:

- **Отсутствующие данные**, в том числе значения, представленные пробельными символами, а также закодированные как отсутствие ответа (например, *$-$*, *?* или *999*).
- **Ошибки данных** - обычно это опечатки при вводе данных.
- **Ошибки измерений** включают в себя данные, которые введены правильно, но основаны на неправильной схеме измерений.
- **Несогласованность кодирования** обычно связана с использованием нестандартных единиц измерения или несогласованности значений, например, использование и буквы *M*, и слова *муж.* для указания гендерной принадлежности.
- **Плохие метаданные** включают в себя несогласованности между очевидным значение поля и названием поля или определением.

Не забудьте записать обнаруженные проблемы качества. Дополнительную информацию смотрите в разделе “Написание отчета о качестве данных” на стр. 17.

## Пример розничного Интернет-магазина - Проверка качества данных

Сценарий поиска в Web с использованием CRISP-DM

Проверка качества данных часто выполняется в течение процессов описания и исследования данных. Некоторые проблемы, встречающиеся у розничных Интернет-магазинов:

**Отсутствующие данные.** Известные отсутствующие данные - это, например, незаполненные анкеты от некоторых зарегистрированных пользователей. Без дополнительной информации из таких анкет этих пользователей, возможно, придется исключить из построения некоторых последующих моделей.

**Ошибки данных.** Большинство источников данных генерируется автоматически, и о них можно не беспокоиться. Опечатки в базе данных продуктов можно найти в процессе исследования.

**Ошибки измерений.** Главный потенциальный источник ошибок изменения - это анкеты. Если какой-либо пункт сформулирован неправильно или непонятно, магазин не получит информацию, на которую надеялся. Но и в этом случае нужно воспользоваться процессом исследования и уделить особое внимание пунктам с необычным распределением ответов.

## Написание отчета о качестве данных

На основе исследования и проверки качества данных сейчас вы готовы составить отчет, который будет руководить проведением следующей фазы CRISP-DM. Дополнительную информацию смотрите в разделе “Проверка качества данных” на стр. 16.

Список задач

Как было выяснено ранее, существует несколько типов проблем качества данных. Перед переходом к следующему шагу обдумайте приведенные ниже вопросы вопросы качества и план для решения. Запишите все ответы в отчете о качестве данных.

- Определены ли пропущенные атрибуты и пустые поля? Если они определены, есть ли смысловое содержание за такими пропущенными значениями?
- Нет ли разночтений написания, которые могли бы привести к проблемам в дальнейших операциях слияния или преобразования?
- Исследовали ли вы отклонения, чтобы выяснить, представляют ли они собой "шум" или необычные явления, которые стоит анализировать в дальнейшем?
- Провели ли вы проверку правдоподобия значений? Запишите замечания о явно выраженных противоречиях (например, о подростках с высоким уровнем доходов).
- Рассмотрели ли вы исключение данных, не влияющих на ваши гипотезы?
- Данные хранятся в плоских файлах? Если да, то согласованы ли между ними разделители? Каждая ли запись содержит одинаковое число полей?

---

## Готовы к следующему шагу?

Перед подготовкой данных для моделирования в IBM SPSS Modeler рассмотрим следующие моменты:

Насколько хорошо вы понимаете данные?

- Ясно ли определены и оценены источники данных? Известны ли вам какие-либо проблемы или ограничения?
- Определены ли ключевые атрибуты по доступным данным?
- Помогли ли эти атрибуты сформулировать гипотезы?
- Записали ли вы размеры всех источников данных?
- Можете ли вы использовать поднабор данных, где это целесообразно?
- Вычислена ли базовая статистика для каждого исследуемого атрибута? Выявлена ли значимая информация?
- Использовали ли вы исследовательскую графику для привнесения дополнительной уникальной информации в ключевые атрибуты? Перепрофилировали ли эти аналитические наработки ваши гипотезы?
- Каковы проблемы качества данных для этого проекта? Есть ли у вас план для устранения этих проблем?
- Ясны ли шаги по подготовке данных? Например, знаете ли вы, слияние каких источников данных следует выполнить и какие атрибуты следует отфильтровать или отобрать?

Теперь, когда вы вооружены пониманием и бизнеса и данных, настало время подготовить данные для моделирования при помощи IBM SPSS Modeler.

# Глава 4. Подготовка данных

## Обзор подготовки данных

Подготовка данных - один из важнейших аспектов исследования данных и часто занимающий много времени. Фактически оценено, что подготовка данных, как правило, занимает 50-70% времени и трудозатрат проекта. Потратив соответствующую энергию на более ранние фазы Понимание бизнес-целей и Начальное изучение данных, эти расходы можно минимизировать, но все равно потребуется большой объем работ по подготовке и компоновке данных для исследования.

В зависимости от организации и ее целей при подготовке данных обычно выполняются следующие задачи:

- Слияние наборов и/или записей данных
- Выбор поднабора данных примера
- Агрегирование записей
- Получение новых атрибутов
- Сортировка данных для моделирования
- Удаление или замена пробельных или пропущенных значений
- Разбиение на наборы данных обучения и тестирования

## Выделение данных

На основании произведенного на предыдущей фазе CRISP-DM начального сбора данных вы готовы начать выделение данных, соответствующих целям анализа данных. В общем случае есть два способа выделения данных:

- **Выделение элементов (строк)** предполагает решения, какие счета, продукты или каких заказчиков включать.
- **Выделение атрибутов или характеристик (столбцы)** предполагает решения об использовании характеристик, таких как объем транзакций или доход домовладений.

## Пример розничного Интернет-магазина - Отбор данных

Сценарий поиска в Web с использованием CRISP-DM

Многие решения розничного интернет-магазина о том, какие данные отбирать, уже были приняты на предыдущих фазах процесса исследования данных.

**Отбор элементов.** Начальное изучение будет ограничено (приблизительно) 30 тысячами покупателей, зарегистрировавшимися на сайте, поэтому нужно сконфигурировать фильтры для исключения покупок и веб-журналов незарегистрированных покупателей. Другие фильтры следует установить для удаления вызовов файлов изображений и других неинформативных записей в веб-журналах.

**Отбор атрибутов.** База данных покупок будет содержать конфиденциальную информацию о покупателях розничного интернет-магазина, поэтому важно отфильтровать такие атрибуты, как имя покупателя, его адрес, телефон и номера кредитных карточек.

## Включение или исключение данных

Поскольку вы остановились на включении или исключении поднаборов данных, обязательно запишите обоснование целесообразности после принятых решений.

Вопросы к рассмотрению

- Относится ли данный атрибут к целям исследования данных?
- Нарушает ли качество конкретного набора данных или атрибута достоверность ваших результатов?
- Можно ли восстановить ценность таких данных?
- Существуют ли какие-нибудь ограничения на использование конкретных полей, таких как *пол* или *раса*?

Отличаются ли ваши решения на этой фазе от гипотез, сформулированных на фазе начального изучения данных? Если да, обязательно запишите вашу мотивировку в отчете проекта.

## Очистка данных

В очистку данных входит тщательное рассмотрение проблем с данными, выбранными для включения в анализ. В IBM SPSS Modeler есть несколько способов очистки данных при помощи узлов операций с записями и полями.

*Таблица 2. Очистка данных*

| **Проблема с данными** | **Возможное решение** |
|---|---|
| Пропущенные данные | Исключите строки или характеристики. Либо заполните пробелы оценочными значениями. |
| Ошибки данных | Используйте логику для обнаружения ошибок и замены вручную. Либо исключите характеристики. |
| Несогласованность кодирования | Остановитесь на одной схеме кодирования, после чего преобразуйте и замените значения. |
| Пропущенные или неверные метаданные | Изучите подозрительные поля и отыщите правильное смысловое значение. |

Подготовленный на фазе начального изучения данных Отчет о качестве данных содержит подробности о типах проблем, касающихся конкретно используемых данных. Его можно использовать в качестве отправной точки для обработки данных в IBM SPSS Modeler.

## Пример розничного Интернет-магазина - Очистка данных

Сценарий поиска в Web с использованием CRISP-DM

Розничный интернет-магазин использует процесс очистки данных для устранения проблем, указанных в отчете о качестве данных.

**Пропущенные данные.** Покупателей, не заполнивших электронную анкету, в дальнейшем, возможно, придется исключить из некоторых моделей. Этим покупателям можно было бы предложить заполнить анкету повторно, но на это уйдут время и деньги, которые розничный интернет-магазин не может себе позволить потратить. Розничный интернет-магазин может смоделировать различия в покупках между покупателями, ответившими и не ответившими на анкету. Если покупательские привычки у этих двух наборов покупателей будут похожи, о незаполненных анкетах можно меньше беспокоиться.

**Ошибки данных.** Здесь могут быть исправлены ошибки, обнаруженные в процессе исследования. Тем не менее, по большей части на сайте контролируется правильность ввода данных перед передачей заполненной страницы покупателем во внутреннюю базу данных.

**Ошибки измерения.** Плохо сформулированные пункты в анкете могут сильно повлиять на качество данных. Как и в случае незаполненных анкет, это трудноразрешимая проблема, поскольку может не оказаться доступного времени или денег, чтобы собрать ответы на новый заменяющий вопрос. Лучшее решение для проблематичных пунктов, возможно, будет следующим: вернуться к процессу выбора и отфильтровать эти пункты из дальнейшего анализа.

## Написание отчета об очистке данных

Написание отчета об очистке данных важно для отслеживания изменений данных. Будущим проектам исследования данных будут полезны подробности вашей работы, доступные в письменной форме.

Список задач

При написании этого отчета рекомендуется рассмотреть следующие вопросы:

- Какого типа шумы встречаются в данных?
- Какие подходы вы использовали для удаления шума? Какие способы оказались успешны?
- Существуют ли наблюдения или атрибуты, которые невозможно исправить? Не забудьте записать, какие данные были исключены из-за проблем с шумом.

---

## Построение новых данных

Часто встречается ситуация, когда требуется построить новые данные. Например, может оказаться полезным создать новый столбец, помечающий флагом приобретение продленной гарантии для каждой покупки. Это новое поле *purchased_warranty* можно легко сгенерировать при помощи узла Задать как флаг в IBM SPSS Modeler.

Новые данные можно построить двумя способами:

- Получив атрибуты (столбцы или характеристики)
- Сгенерировав записи (строки)

В IBM SPSS Modeler предлагается множество способов построения данных при помощи предусмотренных там узлов операций с записями и полями.

## Пример розничного Интернет-магазина - Построение данных

Сценарий поиска в Web с использованием CRISP-DM

При обработке веб-журналов может быть создано множество новых атрибутов. Розничный интернет-магазин хочет, чтобы для событий, записываемых в журналах, создавались отметки времени, определялись посетители и сеансы и отмечалась посещаемая страница и представляемый событием тип операций. Некоторые из этих переменных будут использоваться для создания дополнительных атрибутов, таких как промежуток времени между событиями в сеансе.

В результате слияния или реструктуризации данных иного типа могут быть созданы дополнительные атрибуты. Например, при "свертывании" веб-журналов типа одно событие на строку таким образом, чтобы каждая строка представляла собой сеанс, будут созданы новые атрибуты с записью общего числа действий, суммарного затраченного времени и итогового количества покупок, сделанных во время сеанса. При слиянии веб-журналов с базой данных покупателей так, чтобы каждая строка представляла собой покупателя, будут созданы новые атрибуты с записью числа сеансов, общего числа действий, суммарного затраченного времени и итогового количества покупок, сделанных каждым покупателем.

После построения новых данных розничный интернет-магазин выполняет процесс исследования, позволяющий убедиться в правильности выполнения построения данных.

## Производные атрибуты

В IBM SPSS Modeler можно получить новые атрибуты при помощи следующих узлов операций с полями:

- При помощи узла **Получить** можно создать новые поля из существующих.
- При помощи узла **Задать как флаг** можно создать поле признаков.

Список задач

- При получении атрибутов надо учитывать требования к данным для моделирования. Ожидается ли для алгоритма моделирования конкретный тип данных, например, числовой? Если это так, выполните необходимые преобразования.
- Требуется ли перед моделированием нормализация данных?
- Можно ли построить пропущенные атрибуты методом агрегации, усреднения или индукции?
- Исходя из ваших базовых знаний, существуют ли такие важные факты (например, время, потраченное на сайте), которые могут быть получены из существующих полей?

## Интеграция данных

Нередко существует несколько источников данных для одного набора бизнес-вопросов. Например, у вас может быть доступ и к данным по ипотечным ссудам, и к приобретенным демографическим данным для той же группы клиентов. Если эти данные содержат уникальные идентификаторы (например, номера социального обеспечения), их можно слить в IBM SPSS Modeler, используя это ключевое поле.

Есть два основных способа интеграции данных:

- **Слияние** данных подразумевает объединение двух наборов данных с аналогичными записями, но с разными атрибутами. Данные сливаются с использованием одинакового ключевого идентификатора для каждой записи (такого как ID заказчика). У полученного набора данных увеличивается число столбцов или характеристик.
- **Присоединение** данных представляет из себя объединение двух или более наборов данных с аналогичными атрибутами, но разными записями. Данные интегрируются на основе одинаковых полей (таких как название продукта или продолжительность контракта).

## Пример розничного Интернет-магазина - Интеграция данных

Сценарий поиска в Web с использованием CRISP-DM

При наличии нескольких источников данных у розничных продавцов через Интернет есть много разных способов интеграции данных:

- **Добавление атрибутов заказчиков и продуктов к данным событий.** Чтобы смоделировать события Web-журнала с использованием атрибутов из других баз данных, любой ID заказчика, номер продукта или номер заказа на покупку, связанные с каждым из событий, должны быть правильно определены, а соответствующие атрибуты - слиты в обрабатываемые Web-журналы. Обратите внимание на то, что в слитом файле информация о заказчике и продукте повторяется всякий раз, когда заказчик или продукт связывается с событием.
- **Добавление покупки и информации Web-журнала к данным заказчика.** Чтобы смоделировать значимость заказчика для вас, надо собирать информацию о его покупках и сеансах в соответствующих базах данных, суммировать и сливать ее с базой данных заказчиков. Эта процедура включает в себя создание новых атрибутов, как обсуждалось при описании процесса конструирования данных.

После объединения баз данных розничный Интернет-магазин проводит исследование, чтобы убедиться в правильности выполнения слияния данных.

## Задачи интеграции

Интеграция данных может оказаться весьма сложной задачей, если вы не уделили достаточного внимания исследованиям по первоначальному изучению своих данных. Обдумайте, какие элементы и атрибуты наиболее важны для целей исследования данных, и приступайте к их интеграции.

Список задач

- Используя узлы Слияние или Добавление в IBM SPSS Modeler, интегрируйте наборы данных, которые вы считаете полезными для моделирования.
- Прежде чем перейти к моделированию, обдумайте сохранение полученных выходных данных.

- После слияния данные можно упростить, используя **агрегирование** значений. Агрегирование означает, что новые значения вычисляются при суммировании информации от нескольких записей и/или таблиц.
- Вам может потребоваться также сгенерировать новые записи (такие как средние вычеты из налогов с учетом возврата налогов за несколько лет).

## Форматирование данных

В качестве завершающего шага перед построением модели полезно проверить, требуют ли некоторые способы определенного формата или упорядочивания данных. Например, часто бывает, что алгоритму последовательности требуется предварительная сортировка данных перед запуском модели. Даже если модель сама может выполнить такую сортировку, можно сократить время обработки, используя перед моделированием узел Сортировка.

Список задач

При форматировании данных рассмотрите следующие вопросы:

- Какие модели вы планируете использовать?
- Требуется ли для этих моделей определенный формат или порядок данных?

Если рекомендуются изменения, инструменты обработки в IBM SPSS Modeler могут обеспечить необходимые преобразования данных.

## Готовы к моделированию?

Перед построением моделей в IBM SPSS Modeler проверьте, ответили ли вы на следующие вопросы.

- Все ли данные доступны из IBM SPSS Modeler?
- Можно ли на основании вашего начального исследования и понимания выделить значимые подмножества данных?
- Эффективно ли вы очистили данные или удалили невосстановимые элементы? Документируйте в итоговом отчете все решения.
- Правильно ли интегрированы разные наборы данных? Были ли с объединением данных проблемы, которые нужно документировать?
- Изучили ли вы требования инструментов моделирования, которые собираетесь использовать?
- Существуют ли какие-то проблемы форматирования, которые нужно решить до моделирования? Это касается и обязательных требований форматирования, и задач, которые могут сократить время моделирования.

Если вы смогли ответить на все поставленные вопросы, вы готовы приступить к основной части анализа данных - моделированию.

# Глава 5. Моделирование

## Обзор моделирования

В этом месте трудная выполненная работа начинает приносить свои плоды. Данные, на которые вы потратили время, поступают в инструменты анализа в IBM SPSS Modeler, а результаты начинают проливать свет на бизнес-проблему, поставленную в разделе Понимание бизнес-целей.

Моделирование обычно производится в несколько итераций. Чаще всего исследователи данных запускают несколько моделей с параметрами по умолчанию, а затем настраивают параметры точнее или возвращаются к этапу подготовки данных для выполнения преобразований, требуемых выбранной моделью. В редком случае на вопрос по исследованию данных для организации удается удовлетворительно ответить, используя одну модель и за одно выполнение. Это именно то, что делает исследование данных столь интересным - существует много способов взглянуть на данную проблему, а IBM SPSS Modeler предлагает широкий выбор инструментов для помощи в этом.

## Выбор способов моделирования

Хотя у вас уже могут быть идеи, какого типа моделирование наиболее соответствует целям вашей организации, теперь время принять окончательное решение, что именно использовать. Определение наиболее подходящей модели обычно основывается на следующих соображениях:

- **Типы данных, доступных для анализа.** Например, предназначены ли нужные поля для категориальных (символических) переменных?
- **Ваши цели анализа данных.** Хотите ли вы просто лучше понять устройство складов данных транзакций и найти в них интересующие вас структуры покупок? Или вам нужно произвести индексацию оценок, например, склонность принять условия по умолчания для кредита на образование?
- **Конкретные требования к моделированию.** Требуется ли модели данные конкретного типа или определенного объема? Нужна ли вам модель с легко представимыми результатами?

Более подробную информацию о типах моделей в IBM SPSS Modeler и их требованиях смотрите в документации IBM SPSS Modeler или в оперативной справке.

## Пример розничного Интернет-магазина - Способы моделирования

Используемые розничным Интернет-магазином способы моделирования определяются целями анализа данных компании:

**Улучшенные рекомендации.** Это простейший случай, включающий в себя кластеризацию заказов на покупку для определения, какие именно товары чаще всего приобретаются совместно. Для обогащения результатов можно добавить пользовательские данные и даже записи о посещениях. Для такого типа моделирования подходит двухшаговый способ кластеризации или сеть Коонена. Позже кластеры можно профилировать при помощи набора правил C5.0, чтобы определить, какие рекомендации наиболее подходят в любой момент при посещении Интернет-магазина заказчиком.

**Улучшенная навигация по сайту.** В данном случае розничный продавец должен обратить особое внимание на то, какие страницы часто используются, но требуют от пользователя нескольких переходов между страницами, чтобы найти нужную. Это влечет за собой применение алгоритма упорядочивания к Web-журналам, чтобы сгенерировать "уникальные пути", по которым клиенты проходят на сайте, а затем специальный поиск сеансов с большим количеством посещенных страниц без каких-либо действий (или с действиями после этого долгого пути). Позже, при более глубоком анализе, можно использовать способы кластеризации для идентификации различных "типов" посещений и посетителей, а содержимое сайта можно организовать и представить в соответствии с типами.

## Выбор правильных методов моделирования

В IBM SPSS Modeler доступно много способов моделирования. Часто исследователи данных используют несколько подходов к проблеме с разных сторон.

Список задач

Принимая решение, какие модели использовать, изучите вопрос, влияют ли на ваш выбор следующие проблемы:

- Требуется ли для модели, чтобы данные были разделены на обучающие и проверочные наборы?
- Достаточно ли у вас данных, чтобы получить надежные результаты для данной модели?
- Требуется ли модели определенный уровень качества данных? Можете ли вы обеспечить такой уровень с текущими данными?
- Соответствует ли тип ваших данных конкретной модели? Если нет, можно ли произвести необходимые преобразования при помощи узла Преобразование данных?

Более подробную информацию о типах моделей в IBM SPSS Modeler и о требованиях к ним смотрите в документации IBM SPSS Modeler или в оперативной справке.

## Предположения моделирования

Уточняя выбор инструментов моделирования, не забудьте записать, почему выбрано то или иное решение. Документируйте все предположения о данных, а также любую обработку данных в целях удовлетворить требованиям модели.

Например, и для узла Логистическая регрессия, и для узла Нейросеть требуется, чтобы перед выполнением были полностью **конкретизированы** типы данных (типы данных известны). Это означает, что вам нужно добавить в поток узел Тип и выполнить его для прохода по данным перед построением и выполнением модели. Аналогично, прогнозирующие модели, такие как C5.0, могут выиграть от перебалансировки данных при предсказании правил для редких событий. Делая предсказания такого типа, вы часто можете получить лучшие результаты, вставив в поток узел Баланс и подавая на модель более сбалансированное подмножество данных.

Не забудьте документировать решения такого типа.

---

## Генерирование проверочной структуры

В качестве завершающего шага перед фактическим построением модели нужно снова рассмотреть вопрос, как будут проверяться результаты модели. Генерирование полноценной проверочной структуры состоит из двух частей:

- Описание критерия "добротности" модели
- Определение данных, на которых эти критерии будут проверяться

**Добротность** модели можно измерить несколькими способами. Для контролируемых моделей, таких как C5.0 и дерево C&R, для измерения добротности обычно оценивают уровень ошибки конкретной модели. Для неконтролируемых моделей, таких как кластерные сети Коонена, эти измерения могут включать в себя такие критерии, как простота интерпретации и внедрения или требуемое время обработки.

Не забывайте, что построение модели - это итеративный процесс. Это означает, что обычно вы проверяете результаты нескольких моделей и лишь затем решаете, какие из них использовать и внедрить.

## Описание структуры проверки

Структура проверки - это описание шагов, которые будут выполнены для проведения тестирования созданных моделей. Так как моделирование - это итерационный процесс, важно знать, когда прекратить настройку параметров и попробовать другой метод или модель.

Список задач

При создании структуры проверки рассмотрите следующие вопросы:

- Какие данные будут использоваться для тестирования моделей? Разделили ли вы данные на наборы для обучения и проверки? (Это общепринятый подход, используемый в моделировании).
- Как можно измерить успех контролируемых моделей (таких как C5.0)?
- Как можно измерить успех неконтролируемых моделей (таких как кластерные сети Коонена)?
- Сколько раз вы хотите перезапускать модель со скорректированными параметрами, прежде чем попытаетесь использовать другой тип модели?

## Пример розничного Интернет-магазина - Структура проверки

Сценарий поиска в Web с использованием CRISP-DM

Критерии, по которым оцениваются модели, зависят от самих рассматриваемых моделей и от целей анализа данных:

**Улучшенные рекомендации.** Пока улучшенные рекомендации не будут в реальном времени представлены клиентам, нет абсолютно объективного способа их оценить. Однако розничному Интернет-магазину может потребоваться, чтобы правила, генерирующие рекомендации, были не очень сложными, чтобы их можно было понять с деловой точки зрения. Одновременно правила должны быть и сравнительно сложными, чтобы генерировать различные рекомендации для разных клиентов и сеансов.

**Улучшенная навигация по сайту.** Получив сведения о посещаемых заказчиками страницах Web-сайта, розничный Интернет-магазин сможет объективно оценить изменение дизайна сайта для упрощения доступа к важным страницам. Однако как и в случае с рекомендациями, сложно заранее оценить, насколько успешно клиенты приспособятся к реорганизованному сайту. Если позволяют время и деньги, можно заказать проверку удобства сайта для пользователей.

---

# Построение модели

На данный момент вы должны быть достаточно подготовлены для построения моделей, на рассмотрение которых было потрачено столь много времени. Найдите время и место, чтобы испытать ряд различных моделей перед тем, как делать окончательные выводы. Большинство исследователей данных обычно строят несколько моделей и сравнивают результаты перед тем, как их внедрить или интегрировать.

Для слежения за продвижением работ по целому ряду моделей обязательно сохраняйте замечания об используемых для каждой модели значениях параметров и данных. Они помогут вам обсудить результаты с другими и при необходимости пройти обратно по своим действиям. По завершении процесса построения моделей у вас будет три вида информации для использования в решениях по исследованию данных:

- **Значения параметров**, в том числе ваши замечания о параметрах, генерирующих оптимальные результаты.
- Полученные реальные **модели**.
- **Описания результатов моделей**, в том числе проблемы с производительностью и данными при обработке модели и исследовании ее результатов.

## Пример розничного Интернет-магазина - построение модели

Сценарий поиска в Web с использованием CRISP-DM

**Улучшенные рекомендации.** Генерируются кластеризации для изменяющихся уровней интеграции данных, начиная всего лишь с базы данных покупок, с последующим включением связанной информации о покупателях и сеансах. Кластеризации для каждого уровня интеграции генерируются с различными значениями параметров, задаваемыми для алгоритмов двухэтапной кластеризации и кластеризации сети Коонена. Для каждой из этих кластеризаций генерируется несколько наборов правил C5.0 с различными значениями параметров.

**Улучшенная навигация по сайту.** При помощи узла моделирования последовательностей генерируются пути покупателей. Этот алгоритм разрешает спецификацию критерия минимальной поддержки, помогающего сосредоточиться на чаще всего используемых путях покупателей. Испытываются различные значения параметров.

## Значения параметров

В большинстве способов моделирования есть множество параметров, которые можно настроить для управления процессом моделирования. Например, деревом решений может управлять, настраивая высоту дерева, расщепления и некоторые другие параметры. В большинстве случаев модель строят, используя первоначально опции по умолчанию, а затем уточняя параметры при последующих сеансах.

После определения параметров, обеспечивающих наиболее точные результаты, не забудьте сохранить поток и сгенерированные узлы моделей. Кроме этого, составленные замечания об оптимальных параметрах помогут, когда вы решите автоматизировать модель или перестроить ее с новыми данными.

## Запуск моделей

Запуск моделей в IBM SPSS Modeler выполняется просто. После вставки узла Модель в поток и изменения параметров просто запустите модель для получения просматриваемых результатов. Результаты появятся в навигаторе Сгенерированные модели в правой части рабочего пространства. Для просмотра результатов можно щелкнуть правой кнопкой мыши по модели. Для большинства сгенерированных моделей их можно вставить в поток для дальнейшей оценки и внедрения результатов. Модели можно также сохранить в IBM SPSS Modeler для упрощения повторного использования.

## Описание модели

При проверке результатов модели не забудьте записать замечания о вашем опыте моделирования. Эти комментарии можно сохранить с самой моделью, используя диалоговое поле аннотаций узлов или инструмент проекта.

### Список задач

Для каждой модели запишите следующую информацию:

- Можно ли извлечь значимые выводы из этой модели?
- Открывает ли эта модель какие-то новые аспекты или неожиданные структуры данных?
- Были ли проблемы с исполнением этой модели? Насколько разумно время обработки?
- Были ли у модели сложности с проблемами качества данных, например, с большим количеством пропущенных значений?
- Были ли какие-то несогласованности в вычислениях, которые следует отметить?

---

## Оценка модели

Теперь, когда у вас есть набор начальных моделей, рассмотрите их детальнее, чтобы определить, какие из них достаточно точны или эффективны в качестве окончательной модели. Термин "окончательная" может означать несколько вещей, таких как "готова к внедрению" или "демонстрирующая привлекаемые схемы". Сверка с созданным ранее планом тестирования может помочь выполнить эту оценку с точки зрения вашей организации.

## Всесторонняя оценка модели

Для каждой рассматриваемой модели имеет смысл провести методическую оценку на основе критериев, сгенерированных в вашем плане тестирования. Именно здесь вы можете добавить сгенерированную модель в поток и использовать диаграммы оценки или узлы анализа для анализа эффективности результатов. Необходимо рассмотреть также, насколько результаты соответствуют здравому смыслу и не слишком ли они упрощают ваши бизнес-цели (например, не обнаружены ли последовательности покупок типа вино > вино > вино).

После проведения этих оценок ранжируйте модели по порядку на основе и объективных (точность модели), и субъективных (простота использования или интерпретации результатов) критериев.

Список задач

- Используя инструменты исследования данных в IBM SPSS Modeler, такие как диаграммы оценок, узлы анализа или диаграммы перекрестной проверки, оцените результаты вашей модели.
- Проведите изучение результатов на основании своего понимания бизнес-проблемы. Проконсультируйтесь с аналитиком данных или с другими экспертами, которые могут вникнуть в суть конкретных результатов.
- Посмотрите, легко ли внедряются результаты модели. Нужно ли вашей организации внедрить результаты через Web или отослать их обратно в хранилище данных?
- Проанализируйте влияние результатов на ваши критерии успешности. Отвечают ли они целям, установленным на этапе понимания бизнеса?

Если вы готовы решить все перечисленные проблемы и верите, что текущие модели соответствуют вашим целям, пора переходить к более подробным оценкам моделей и к окончательному внедрению. В противном случае примите во внимание обнаруженные недостатки и повторно запустите модели со скорректированными значениями параметров.

## Пример розничного Интернет-магазина - оценка модели

Сценарий поиска в Web с использованием CRISP-DM

**Усовершенствованные рекомендации.** И сеть Коонена, и двухэтапная кластеризация дают правдоподобные результаты, и розничному интернет-магазину трудно выбрать между этими двумя методами. На данный момент компания надеется использовать и то, и другое, принимая рекомендации, относительно которых эти методы согласны друг с другом, и изучая более подробно ситуации, в которых они отличаются. С небольшими трудозатратами и примененным знанием дела розничный интернет-магазин может разработать дополнительные правила устранения противоречий между этими двумя методами.

Розничный интернет-магазин находит также, что результаты, включающие информацию о сеансах, на удивление хороши. Есть основание предположить, что рекомендации можно было бы связать с навигацией по сайтам. Набор правил, определяющий, куда вероятнее всего перейдет покупатель далее, можно использовать в реальном времени для непосредственного влияния на содержимое сайтов по мере их открытия покупателем в браузере.

**Усовершенствованная навигация по сайту.** Модель последовательностей гарантирует розничному интернет-магазину высокий доверительный уровень, с которым могут быть предсказаны некоторые пути покупателя, генерируя результаты, предполагающие контролируемое количество изменений, вносимых в макет сайта.

## Отслеживание измененных параметров

На основе полученных при оценке моделей знаний время взглянуть на модели под другим углом зрения. У вас есть две возможности:

- Настроить параметры существующих моделей.
- Выбрать другую модель для решения вашей проблемы с анализом данных.

В обоих случаях вы возвращаетесь к задаче построения моделей и проводите итерации, пока результаты не будут вас удовлетворять. Не беспокойтесь о повторении шагов. Это абсолютно общая практика для исследователей данных - оценивать и перезапускать модели несколько раз, прежде чем найдется модель, соответствующая потребностям. Это хороший аргумент в пользу построения нескольких моделей сразу и сравнения их результатов до корректировки параметров для каждой из моделей.

## Готовы к следующему шагу?

Прежде чем перейти к окончательной оценке модели, проверьте, была ли ваша начальная оценка в должной степени подробной.

Список задач

- Способны ли вы понять результаты моделей?
- Осмысленны ли результаты моделей с логической точки зрения? Есть ли явные несоответствия, которые требуют дальнейшего исследования?
- Как на ваш взгляд, отвечают ли результаты на бизнес-вопросы вашей организации?
- Использовали ли вы узлы анализа и диаграммы роста и выигрыша для сравнения и оценки точности модели?
- Исследовали ли вы несколько моделей и сравнивали ли результаты?
- Можно ли внедрить результаты вашей модели?

Если результаты вашего моделирования данных выглядят точными и достоверными, пора произвести более подробную оценку перед окончательным внедрением.

# Глава 6. Оценка

## Обзор оценки

На данный момент завершена преобладающая часть проекта исследования данных. Кроме того, на фазе Моделирование вы определили, что построенные модели технически верны и эффективны в соответствии с определенными ранее **критериями успешности исследования данных**.

Однако перед продолжением следует оценить результаты проведенных исследований при помощи **критериев успешности бизнеса**, установленных в самом начале проекта. Это основной момент, гарантирующий, что ваша организация сможет воспользоваться полученными вами результатами. При исследовании данных генерируются результаты двух типов:

- Окончательные **модели**, выбираемые на предыдущей фазе CRISP-DM.
- Любые заключения или выводы, получаемые как из самих моделей, так и из процесса исследования данных. Их называют **наработками**.

## Оценка результатов

На этой странице вы формализуете оценку соответствия результатов проекта критериям успешности бизнеса. Этот шаг требует четкого понимания сформулированных бизнес-целей для гарантированного включения ответственных за принятие основных решений в оценку проекта.

Список задач

Во-первых, нужно записать оценку соответствия результатов исследований данных критериям успешности бизнеса. Рассмотрите в отчете следующие вопросы:

- Сформулированы ли ваши результаты четко и в удобопредставимой форме?
- Есть ли особо оригинальные или уникальные наработки, на которых следует заострить внимание?
- Можете ли вы ранжировать модели и наработки по степени их применимости к целям бизнеса?
- Насколько хорошо в общих чертах эти результаты отвечают бизнес-целям вашей организации?
- Какие дополнительные вопросы подняли ваши результаты? Как эти вопросы можно было бы выразить в бизнес-терминах?

После оценки результатов составьте список утвержденных моделей для включения в заключительный отчет. Этот список должен содержать модели, удовлетворяющие и целям исследования данных, и бизнес-целям вашей организации.

## Пример розничного Интернет-магазина - Оценка результатов

Сценарий поиска в Web с использованием CRISP-DM

Общими результатами первого опыта применения розничным интернет-магазином исследования данных достаточно легко обмениваться с точки зрения бизнеса: изучение дало (на что и надеялись) улучшенные рекомендации по продуктам и усовершенствованный макет сайта. Усовершенствованный макет сайта основывается на просматриваемых покупателями последовательностях, показывающих ресурсы сайта, которые нужны покупателям, но доступны им только за несколько переходов по страницам. Обоснование улучшения рекомендаций по продуктам труднее, поскольку возможно усложнение решающих правил. Чтобы составить заключительный отчет, аналитики попытаются выявить некоторые общие тенденции в наборах правил, которые можно проще объяснить.

**Ранжирование моделей.** Поскольку некоторые из начальных моделей, похоже, должны составлять бизнес-значение, ранжирование в их группе основывалось на статистических критериях, простоте интерпретации и многоплановости. Таким образом, моделирование дало различные рекомендации для различных ситуаций.

**Новые вопросы.** Самый важный вопрос, полученный в результате изучения: как розничный Интернет-магазин может узнать больше о своих покупателях? Информация в базе данных покупателей играет важную роль в формировании кластеров для рекомендаций. Пока для составления рекомендаций для покупателей, информация о которых отсутствует, доступны специальные правила, эти рекомендации по характеру будут более общими, чем рекомендации, которые можно составить для зарегистрированных покупателей.

## Обзор процесса

Эффективные методологии обычно включают в себя некоторое время на изучение успехов и слабых мест только что завершенного процесса. Исследование данных - это не исключение. Часть CRISP-DM - это обучение на собственном опыте, чтобы будущие проекты исследований данных были более эффективны.

Список задач

Во-первых, нужно собрать вместе все операции и решения для каждой фазы, в том числе шаги по подготовке данных, построение моделей и т.д. Затем для каждой фазы рассмотрите следующие вопросы и сделайте предложения по улучшению:

- Внесла ли эта стадия свой вклад в конечные результаты?
- Есть ли способы оптимизировать или улучшить конкретную стадию или операцию?
- Были ли какие-то ошибки или неудачи на этой фазе? Как их можно избежать в следующий раз?
- Встретились ли тупиковые ситуации, например, модели, оказавшиеся бесполезными? Есть ли способы предвидеть такие тупиковые ситуации, чтобы тратить усилия более продуктивно?
- Были ли какие-то неожиданности в этой фазе (и хорошие, и плохие)? Оценивая задним числом, есть ли простой способ предсказать их появление?
- Существуют ли альтернативные решения или стратегии, которые могли использоваться на данной фазе? Внесите эти альтернативы в замечания для будущих проектов исследования данных.

## Пример розничного Интернет-магазина - Обзорный отчет

Сценарий поиска в Web с использованием CRISP-DM

В результате обзора процесса начального проекта исследования данных розничный Интернет-магазин высоко оценил взаимосвязь между шагами в процессе. Первоначально не собираясь просматривать процесс CRISP-DM в "обратном порядке", теперь этот продавец понял, что циклическая природа этого процесса увеличивает его силу. Обзор процесса привел его к пониманию, что:

- Когда что-то необычное появляется на некоторой фазе процесса CRISP-DM, всегда можно вернуться к процессу исследования.
- Подготовка данных, особенно Web-журналов, требует терпения, так как может занять весьма длительное время.
- Очень важно в любой момент помнить о бизнес-целях, так как при готовности данных для анализа очень просто начать моделирование, упустив картину в целом.
- Когда фаза моделирования завершена, понимание бизнес-целей становится даже более важным, особенно для решения о том, как реализовать результаты и определить, какие дальнейшие исследования необходимы.

## Определение следующих действий

К настоящему моменту вы получили результаты, оценили свой опыт в исследовании данных и теперь можете задать себе вопрос: **Что дальше?** Эта фаза поможет вам ответить на такой вопрос в свете ваших бизнес-целей для исследования данных. По существу, в данный момент у вас есть выбор из двух вариантов:

- **Продолжить фазу внедрения.** Следующая фаза поможет вам включить результаты моделей в бизнес-процесс и создать окончательный отчет. Даже если ваши усилия по исследованию данных оказались безрезультатны, необходимо использовать стадию внедрения CRISP-DM для создания окончательного отчета и его представления спонсору проекта.
- **Вернуться назад и уточнить или заменить модели.** Если оказывается, что ваши результаты уже практически оптимальны, рассмотрите возможность следующего раунда моделирования. Можно воспользоваться приобретенным на этой фазе опытом и применить его для уточнения моделей и получения лучших результатов.

Ваше решение в данный момент должно основываться на точности и соответствии действительности результатов моделирования. Если ваши результаты соответствуют целям анализа данных и бизнеса, вы готовы для фазы внедрения. При любом решении не забудьте подробно документировать, как именно и почему такое решение было принято.

## Пример розничного Интернет-магазина - Следующие шаги

Сценарий поиска в Web с использованием CRISP-DM

Розничный Интернет-магазин вполне уверен в точности и обоснованности результатов проекта и переходит к фазе внедрения.

В то же время команда проекта готова вернуться назад и подправить некоторые модели для включения прогнозирования. В это время они ожидают получения окончательных отчетов и принятия решения о дальнейших действиях от руководства.

# Глава 7. Внедрение

## Обзор внедрения

Внедрение - это процесс использования новых аналитических наработок для внесения усовершенствований в организации. Оно может означать формальную интеграцию, такую как реализация модели IBM SPSS Modeler, генерирующей оценки оттока, которые затем считываются в хранилище данных. В другом варианте внедрение может означать использование аналитических наработок по исследованию данных для выявления изменения в вашей организации. Скажем, вы обнаружили настораживающие шаблоны в данных, указывающие на сдвиг в поведении покупателей возрастом более 30 лет. Эти результаты могут и не быть формально интегрированы в применяемые информационные системы, но они безусловно будут полезны для планирования и принятия решений о рынках сбыта.

В общем случае в фазу внедрения CRISP-DM входят операции двух типов:

- Планирование и мониторинг внедрения результатов
- Выполнение задач по подведению итогов, таких как составление заключительного отчета и проведение проверки проекта

В зависимости от требований организации может потребоваться выполнить один или оба этих шага.

## Планирование внедрения

Понятно ваше стремление как можно быстрее воспользоваться плодами усилий по исследованию данных, но сначала уделите некоторое время планированию удобного и исчерпывающего внедрения результатов.

Список задач

- Первый шаг - это составление сводки ваших результатов, и по самой модели, и по обнаруженным особенностям. Это поможет определить, какие модели можно интегрировать в ваши системы баз данных, и какие обнаруженные особенности следует показать коллегам.
- Для каждой внедряемой модели составьте пошаговый план внедрения и интеграции с вашими системами. Записывайте все технические подробности, такие как требования базы данных к выходным данным модели. Например, ваша система может требовать, чтобы выходные данные моделирования внедрялись в формате с символом табуляции в качестве разделителя.
- Для каждого последующего обнаруженного факта создайте план распространения этой информации среди сотрудников, ответственных за создание стратегии.
- Существуют ли альтернативные планы внедрения для обоих типов результатов, о которых следует упомянуть?
- Рассмотрите вопрос, как будет происходить мониторинг внедрения. Например, как будет обновляться модель, внедряемая с использованием IBM SPSS Modeler Solution Publisher? Как именно вы поймете, что модель больше не применима?
- Определите все проблемы внедрения и план действий на случай непредвиденных обстоятельств. Например, ответственные за принятие решений могут затребовать дополнительную информацию о результатах моделирования и дополнительные технические подробности.

## Пример розничного Интернет-магазина - планирование внедрения

Сценарий поиска в Web с использованием CRISP-DM

Для успешного внедрения результатов исследования данных розничного интернет-магазина требуется, чтобы правильная информация дошла до нужных людей.

**Ответственные за принятие решений.** Ответственные за принятие решений должны быть извещены о рекомендациях и предлагаемых изменениях сайта с краткими объяснениями о том, как смогут помочь эти изменения. При условии, что они примут результаты изучения, должны быть уведомлены сотрудники, вносящие изменения.

**Веб-разработчики.** Сотрудники, обслуживающие сайт, будут должны включить новые рекомендации в организацию содержимого сайта. Сообщите им о том, какие изменения *возможны* в результате будущих исследований, чтобы они могли заложить фундамент наработок уже сейчас. Проведенная подготовка команды к построению сайта "по ходу дела", исходя из анализа последовательностей в реальном времени, может оказаться полезной в дальнейшем.

**Специалисты по базам данных.** Сотрудников, обслуживающих базы данных покупателей, покупок и продуктов, нужно держать в ведении о том, как как используется информация из баз данных и какие атрибуты могут быть добавлены в базы данных в будущих проектах.

Прежде всего, команда проекта должна поддерживать связь с каждой из этих групп для согласования внедрения результатов и планирования будущих проектов.

## Планирование мониторинга и обслуживания

При полномасштабном внедрении и интеграции результатов моделирования исследование данных может выполнять регулярно. Например, если модель внедряется для предсказания последовательности покупок в покупательской корзине на интернет-сайте, этой модели скорее всего потребуется регулярная оценка для обеспечения ее эффективности и постоянных усовершенствований. Аналогично, модель, внедренная для удержания заказчика среди ценных заказчиков, по-видимому, нужно будет немного подправить после достижения конкретного уровня удержания. Эту модель можно изменить и использовать повторно для сохранения заказчиков на более низком, но все-таки обеспечивающем прибыль, уровне пирамиды значимости.

Список задач

Подготовьте замечания по следующим вопросам и не забудьте вставить их в итоговый отчет.

- Для каждой модели или обнаруженной особенности, какие факторы или дополнительные влияющие обстоятельства (например, рыночная стоимость или сезонные изменения) нужно отслеживать?
- Как можно измерять и отслеживать точность модели?
- Как вы определите момент, когда модель "устареет"? Предоставьте конкретные пороги для измерения точности, ожидаемые изменения данных и так далее.
- Что произойдет, когда модель устареет? Можно ли будет просто повторно построить модель с новыми данными или провести небольшую корректировку? Или же изменения могут быть настолько существенными, что потребуется новый проект исследования данных?
- Может ли эта модель использоваться для аналогичных бизнес-вопросов, когда она уже не будет использоваться по назначению? Именно на этом этапе хорошая документация становится критически важной для оценки бизнес-целей каждого проекта исследования данных.

## Пример розничных Интернет-торговли - Мониторинг и обслуживание

Сценарий поиска в Web с использованием CRISP-DM

Первоочередная задача для мониторинга - определить, работают ли улучшенные рекомендации и новый сайт организации. Например, могут ли пользователя по более коротким маршрутам перейти на страницы, которые они ищут? Увеличились ли совместные покупки рекомендуемых элементов? После нескольких недель мониторинга розничный Интернет-магазин сможет определить успешность исследования.

Включение новых зарегистрированных пользователей можно обрабатывать автоматически. Когда заказчики регистрируются на сайте, к информации о них могут применяться текущие наборы правил для определения, какие рекомендации можно предоставить.

Определение, когда именно надо изменять наборы правил для определения рекомендаций - непростая задача. Изменение наборов правил - это не автоматический процесс, так как создание кластеров требует человеческого участия в отношении приемлемости данного кластерного решения.

Так как будущие проекты генерируют более сложные модели, необходимость в мониторинге и его объем будут только возрастать. Где возможно, массовые задачи мониторинга нужно автоматизировать для получения регулярных запланированных отчетов, доступных для проверки и изучения. Другое возможное решение для компании - создание моделей, обеспечивающих быстрые оперативные предсказания на лету. Это требует более сложной работы по сравнению с проектом первичного исследования данных.

## Составление итогового отчета

Написание заключительного отчета не только увязывает невыясненные вопросы, но может также использоваться и для обнародования ваших результатов. Хотя это и выглядит необязательным, но результаты важно представить различным людям, которые в них заинтересованы. Сюда могут быть включены и технические администраторы, ответственные за реализацию результатов моделирования, и инвесторы в сфере маркетинга и менеджмента.

Список задач

Сначала рассмотрим аудиторию вашего отчета. Входят ли в ее состав технические разработчики или менеджеры с рыночной ориентацией? Если потребности аудиторий в корне отличаются, может потребоваться создать для каждой из них отдельные отчеты. В любом случае отчет должен содержать преобладающую часть следующих пунктов:

- Доскональное описание бизнес-проблемы
- Процесс, используемый для проведения исследования данных
- Затраты на проект
- Замечания к отклонениям от исходного плана проекта
- Сводка результатов исследования данных (включая и модели, и наработки)
- Обзор предлагаемого плана внедрения
- Рекомендации по дальнейшей работе исследования данных, включая интересные направления, обнаруженные при исследовании и моделировании

## Подготовка итоговой презентации

Помимо отчета по проекту, может также потребоваться представить наработки по проекту команде инвесторов или смежным отделам. В этом случае почти всю одну и ту же информацию можно будет взять из отчета, но представить ее с более широкой точки зрения. Для этого типа представления можно без труда экспортировать диаграммы и графики в IBM SPSS Modeler.

## Пример розничного Интернет-магазина - Итоговый отчет

Сценарий поиска в Web с использованием CRISP-DM

Самое большое отклонение от исходного плана проекта состоит также в направлении будущей работы по исследованию данных. Исходный план ставил задачи, как побудить покупателей тратить больше времени и просматривать побольше страниц на сайте за одно посещение.

Как выясняется, наличие удовлетворенных покупателей - это не просто вопрос удержания их на сайте. Гистограммы потраченного за сеанс времени с учетом того, привело ли оно к покупке, выявили, что

промежутки времени большинства сеансов, завершившихся покупками, попадают в диапазон между промежутками времени сеансов для двух кластеров безрезультатных сеансов.

Теперь, когда это уже известно, должен быть поставлен вопрос, что делают покупатели, проводящие на сайте много времени, ничего не покупая: просматривают его содержимое или просто не могут найти то, что они ищут. Следующий шаг подразумевает выявление способа предоставления нужных им товаров, чтобы побудить их делать покупки.

## Проведение итогового обзора проекта

Это заключительный шаг методологии CRISP-DM; он предоставляет вам возможность сформулировать окончательные впечатления и обобщить извлеченные в процессе анализа данных уроки.

Список задач

Необходимо провести короткие интервью со всеми вовлеченными в процесс анализа данных. Во время этих интервью нужно задать следующие вопросы:

- Каково ваше общее впечатление о проекте?
- Чему вы научились во время этого процесса, и в части анализа данных в целом, и конкретно при использовании доступных данных?
- Какие части проекта прошли успешно? В чем возникали сложности? Была ли доступна информация, которая могла бы помочь в устранении сложностей?

После внедрения результатов анализа данных можно было бы также взять интервью у тех, для кого эти результаты интересны, в частности, у заказчиков и бизнес-партнеров. Цель этого опроса - определить, насколько ценен проект и предоставил ли он выгоды, ради которых был начат.

Результаты этих интервью можно объединить с вашими собственными впечатлениями о проекте в окончательном отчете, который должен фокусироваться на уроках, извлеченных из опыта исследования ваших складов данных.

## Пример розничного Интернет-магазина - Итоговый обзор

Сценарий поиска в Web с использованием CRISP-DM

**Интервью с участниками проекта.** Розничный интернет-магазин выясняет, что участники проекта, теснейшим образом связанные с проводимым изучением от начала и до конца, большей частью довольны результатами и с нетерпением ждут будущих проектов. Группа, работающая с базами данных, выглядит сдержанно оптимистичной; они хоть и оценивают практическую ценность изучения, но отмечают дополнительные косвенные затраты на ресурсы баз данных. По ходу изучения был доступен консультант, но в перспективе, по мере расширения области действия проекта, будет необходим другой сотрудник, выделенный специально для техобслуживания баз данных.

**Интервью с покупателями.** На настоящий момент уже была налажена почти бесперебойная обратная связь с покупателями. Одна плохо продуманная проблема состояла в том, что изменение макета сайта затронуло постоянных покупателей. По прошествии нескольких лет зарегистрированные покупатели имели некоторые ожидания относительно организации сайта. Отзывы от зарегистрированных пользователей не настолько положительны, как от незарегистрированных, и некоторые были сильно недовольны реорганизацией. Розничный интернет-магазин должен оставаться в курсе этой проблемы и очень внимательно подходить к пониманию вопроса о том, привлечет ли указанное изменение новых покупателей без риска потерять существующих.

# Уведомления

Эта информация относится к продуктам и сервису, предлагаемым в США. Этот материал может быть доступен от IBM на других языках. Однако для его получения может понадобиться приобрести продукт или версию продукта на нужном языке.

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, MD-NC119*
*Armonk, NY 10504-1785*
*US*

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.*
*19-21, Nihonbashi-Hakozakicho, Chuo-ku*
*Tokyo 103-8510, Japan*

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 странах для ряда сделок не допускается отказ от явных или предполагаемых гарантий; в таком случае данное положение к вам не относится.

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 ссылки в этой публикации на сайты, не принадлежащие IBM, приведены только для удобства и никоим образом не означают их поддержки. Материалы на этих сайтах не входят в число материалов по данному продукту IBM, и весь риск пользования этими сайтами несете вы сами.

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 Director of Licensing*
*IBM Corporation*
*North Castle Drive, MD-NC119*
*Armonk, NY 10504-1785*
*US*

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.

Данные производительности и примеры клиентов представлены только для иллюстрации. Фактическая производительность зависит от конкретной конфигурации и условий работы.

Информация о продуктах других компаний (не IBM) получена от поставщиков этих продуктов, из их опубликованных объявлений или из иных общедоступных источников. IBM не производила тестирование этих продуктов и никак не может подтвердить информацию о их точности работы и совместимости, а также прочие заявления относительно продуктов других компаний (не IBM). Вопросы о возможностях продуктов других компаний (не IBM) следует направлять поставщикам этих продуктов.

Все утверждения о будущих планах и намерениях IBM могут быть изменены или отменены без уведомлений, и описывают исключительно цели фирмы.

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 являются вымышленными и любое их сходство с реальными именами и адресами предприятий является случайным.

## 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 смотрите на веб-сайте "Copyright and trademark information" (Информация об авторских правах и товарных знаках) по адресу www.ibm.com/legal/copytrade.shtml.

Adobe, логотип Adobe, PostScript и логотип PostScript являются либо зарегистрированными товарными знаками, либо товарными знаками корпорации Adobe Systems в Соединенных Штатах и/или других странах.

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.

## Правила и условия для документации продукта

Разрешения для использования этих публикаций предоставляются на следующих условиях.

### Применимость

Данные правила и условия являются дополнением к правилам использования для сайта IBM.

### Персональное использование

Вы можете воспроизводить эти публикации для персонального некоммерческого использования при условии сохранения всех замечаний о правах собственности. Вам запрещается распространять эти публикации, полностью или по частям, демонстрировать их или создавать из них производные продукты без явного на то согласия от IBM.

### Коммерческое использование

Вам предоставляется право воспроизводить эти публикации исключительно в пределах своего предприятия при условии, что будут воспроизведены все замечания об авторских правах. За пределами вашего предприятия вам запрещается распространять эти публикации, полностью или по частям, демонстрировать их или создавать из них производные продукты без явного на то согласия от IBM.

### Права

За исключением прав, явным образом предоставляемых настоящим разрешением, никаких иных разрешений, лицензий и прав, ни явных, ни подразумеваемых, в отношении публикаций и любой содержащейся в них информации, данных, программ или иной интеллектуальной собственности, не предоставляется.

IBM оставляет за собой право отозвать разрешения, предоставленные этим документом, если, по мнению IBM, использование публикаций наносит ущерб IBM или, как это установлено IBM, вышеприведенные инструкции не соблюдаются должным образом.

Запрещается загружать, экспортировать или реэкспортировать эту информацию, если при этом не будут полностью соблюдаться все применимые законы и постановления, включая все законы и постановления США, касающиеся экспорта.

IBM НЕ ДАЕТ НИКАКИХ ГАРАНТИЙ ОТНОСИТЕЛЬНО СОДЕРЖАНИЯ ЭТИХ ПУБЛИКАЦИЙ. ПУБЛИКАЦИИ ПРЕДСТАВЛЯЮТСЯ "КАК ЕСТЬ", БЕЗ КАКИХ-ЛИБО ГАРАНТИЙ, ЯВНЫХ ИЛИ ПОДРАЗУМЕВАЕМЫХ, ВКЛЮЧАЯ (НО НЕ ОГРАНИЧИВАЯСЬ ТАКОВЫМИ) ПРЕДПОЛАГАЕМЫЕ ГАРАНТИИ ОТСУТСТВИЯ НАРУШЕНИЙ, КОММЕРЧЕСКОЙ ПРИГОДНОСТИ ИЛИ СООТВЕТСТВИЯ ОПРЕДЕЛЕННОЙ ЦЕЛИ.

# Индекс

## П



## Р



## С



## Т



## У



## Ф



## Ц



## Ч



## Ш

Напечатано в Дании