Manual CRISP-DM de IBM SPSS Modeler

Nota

Antes de utilizar esta información y el producto al que da soporte, lea la información que figura en el apartado "Avisos" en la página 39.

Información del producto

Esta edición se aplica a la versión 17, release 1, modificación 0 de IBM(r) SPSS(r) y a todos los releases y modificaciones subsiguientes hasta que se indique lo contrario en nuevas ediciones.

Contenido

Prefacio

IBM® SPSS Modeler es el conjunto de programas de minería de datos de IBM Corp. orientado a las empresas. SPSS Modeler ayuda a las organizaciones a mejorar la relación con sus clientes y los ciudadanos a través de la comprensión profunda de los datos. Las organizaciones utilizan la comprensión que les ofrece SPSS Modeler para retener a los clientes más rentables, identificar las oportunidades de venta cruzada, atraer a nuevos clientes, detectar el fraude, reducir el riesgo y mejorar la prestación de servicios del gobierno.

La interfaz visual de SPSS Modeler invita a los usuarios a aplicar su experiencia empresarial específica, lo que deriva en modelos proactivos más eficaces y la reducción del tiempo necesario para encontrar soluciones. SPSS Modeler ofrece muchas técnicas de modelado tales como predicciones, clasificaciones, segmentación y algoritmos de detección de asociaciones. Una vez que se crean los modelos, IBM SPSS Modeler Solution Publisher permite su distribución en toda la empresa a los encargados de tomar las decisiones o a una base de datos.

Acerca de IBM Business Analytics

El software IBM Business Analytics ofrece información completa, coherente y precisa en la que confían los encargados de la toma de decisiones para mejorar el rendimiento comercial. Un conjunto integral de inteligencia empresarial, análisis predictivo, rendimiento financiero y gestión de estrategias y aplicaciones de análisis que ofrece una perspectiva clara, inmediata e interactiva del rendimiento actual y la capacidad de predecir resultados futuros. En combinación con extensas soluciones sectoriales, prácticas probadas y servicios profesionales, las organizaciones de cualquier tamaño pueden conseguir el máximo de productividad, automatizar las decisiones de forma fiable y alcanzar mejores resultados.

Como parte de esta familia, el software de análisis predictivo de IBM SPSS ayuda a las organizaciones a predecir eventos futuros y actuar proactivamente según esa información para lograr mejores resultados comerciales. Los clientes comerciales, gubernamentales y académicos de todo el mundo confían en la tecnología de IBM SPSS como ventaja ante la competencia para atraer, retener y hacer crecer a los clientes, reduciendo al mismo tiempo el fraude y el riesgo. Al incorporar el software IBM SPSS en las operaciones diarias, las organizaciones se convierten en empresas predictivas, capaces de dirigir y automatizar las decisiones para alcanzar los objetivos empresariales y lograr una ventaja considerable sobre la competencia. Para obtener más información o contactar con un representante, visite http://www.ibm.com/spss.

Asistencia técnica

Hay asistencia técnica disponible para los clientes de mantenimiento. Los clientes podrán ponerse en contacto con el servicio de asistencia técnica si desean recibir ayuda sobre la utilización de los productos de IBM Corp. o sobre la instalación en los entornos de hardware admitidos. Para ponerse en contacto con el servicio de asistencia técnica, consulte el sitio web de IBM Corp. en http://www.ibm.com/support. Tenga a mano su acuerdo de asistencia y esté preparado para identificarse a sí mismo y a su organización al solicitar ayuda.

Capítulo 1. Introducción al CRISP-DM

Conceptos básicos de ayuda de CRISP-DM

CRISP-DM, que son las siglas de Cross-Industry Standard Process for Data Mining, es un método probado para orientar sus trabajos de minería de datos.

- Como **metodología**, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas.
- Como **modelo de proceso**, CRISP-DM ofrece un resumen del ciclo vital de minería de datos.

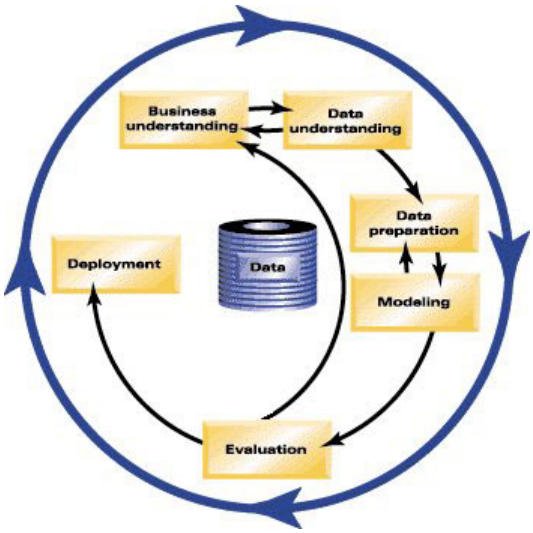

Figura 1. Ciclo de vida de minería de datos

El ciclo vital del modelo contiene seis fases con flechas que indican las dependencias más importantes y frecuentes entre fases. La secuencia de las fases no es estricta. De hecho, la mayoría de los proyectos avanzan y retroceden entre fases si es necesario.

El modelo de CRISP-DM es flexible y se pueden personalizar fácilmente. Por ejemplo, si su organización intenta detectar actividades de blanqueo de dinero, es probable que necesite realizar una criba de grandes cantidades de datos sin un objetivo de modelado específico. En lugar de realizar el modelado, su trabajo se centrará en explorar y visualizar datos para descubrir patrones sospechosos en datos financieros. CRISP-DM permite crear un modelo de minería de datos que se adapte a sus necesidades concretas.

En tal situación, las fases de modelado, evaluación y despliegue pueden ser menos relevantes que las fases de preparación y comprensión de datos. Sin embargo, es muy importante considerar algunas cuestiones que surgen durante fases posteriores para la planificación a largo plazo y objetivos futuros de minería de datos.

CRISP-DM en IBM SPSS Modeler

IBM SPSS Modeler incorpora la metodología CRISP-DM de dos formas para proporcionar una compatibilidad exclusiva para una minería de datos efectiva.

- La herramienta de proyectos de CRISP-DM le ayuda a organizar rutas de proyectos, resultados y anotaciones, de acuerdo con las fases de un proyecto normal de minería de datos. Puede realizar informes en cualquier momento durante el proyecto en función de las notas de las rutas y fases de CRISP-DM.

- La ayuda de CRISP-DM le guía por el proceso de realización de un proyecto de minería de datos. El sistema de ayuda incluye una lista de tareas para cada paso y ejemplos de cómo funciona CRISP-DM en aplicaciones reales. Puede acceder a la ayuda de CRISP-DM seleccionando **Ayuda de CRISP-DM** en la ventana principal del menú Ayuda.

Herramienta de proyectos de CRISP-DM

La herramienta de proyectos de CRISP-DM proporciona un método estructurado de minería de datos que puede ayudarle a asegurar el rendimiento de su proyecto. Se trata esencialmente de una extensión de la herramienta de proyectos estándar IBM SPSS Modeler. De hecho, puede alternar entre la vista de CRISP-DM y la vista Clases estándar para ver las rutas y los resultados organizados por el tipo o la fase de CRISP-DM.

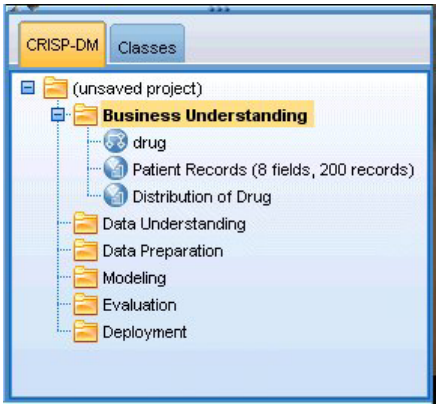

Figura 2. Herramienta de proyectos de CRISP-DM

Si utiliza la vista de CRISP-DM de la herramienta de proyectos, podrá:

- Organizar las rutas y resultados de un proyecto en función de las fases de minería de datos.
- Registrar los objetivos de su organización de cada fase.
- Crear información personalizada sobre herramientas para cada fase.
- Registrar la conclusiones obtenidas de un modelo o gráfico concreto.
- Generar un informe HTML o actualizar la distribución del equipo de proyecto.

Ayuda sobre CRISP-DM.

IBM SPSS Modeler ofrece una guía en línea para el modelo de proceso independiente de CRISP-DM. Esta guía está organizada por fases de proyecto y proporciona la siguiente información:

- Un resumen y lista de tareas de cada fase de CRISP-DM
- Ayuda sobre la elaboración de informes de diferentes hitos
- Ejemplos reales que ilustran cómo el equipo de proyecto pueden utilizar CRISP-DM para ilustrar la minería de datos
- Enlaces a recursos adicionales de CRISP-DM

Puede acceder a la ayuda de CRISP-DM seleccionando **Ayuda de CRISP-DM** en la ventana principal del menú Ayuda.

Recursos adicionales

Además de la ayuda de IBM SPSS Modeler para CRISP-DM, existen varias formas de ampliar su conocimiento de los procesos de minería de datos.

- Visite el sitio Web de CRISP-DM en www.crisp-dm.org.
- Consulte el manual de CRISP-DM, creado por CRISP-DM Consortium y proporcionado con esta versión.

- Consulte el manual *Data Mining with Confidence*, copyright 2002, SPSS Inc., ISBN 1-56827-287-1.

Capítulo 2. Comprensión del negocio

Conceptos básicos sobre comprensión comercial

Incluso antes de utilizar IBM SPSS Modeler, debe dedicar tiempo a explorar las expectativas de su organización con respecto a la minería de datos. Intente implicar a la mayor cantidad de personas que sea posible en estas discusiones y documente los resultados. El paso final de la fase de CRISP-DM trata de cómo producir un plan de proyecto utilizando la información que se contiene en esta documentación.

Aunque este estudio pueda parecer prescindible, no lo es. Conozca las razones comerciales para que sus esfuerzos en minería de datos aseguren que todos los usuarios están de acuerdo antes de asignar recursos.

Determinación de los objetivos comerciales

Su primera tarea es obtener la máxima información posible de los objetivos comerciales de la minería de datos. Es posible que esta tarea no sea tan fácil como parece, pero puede reducir los futuros riesgos clarificando los problemas, objetivos y recursos.

La metodología CRISP-DM proporciona una forma estructurada de alcanzar estos objetivos.

Lista de tareas

- Comience a recopilar información acerca de la situación comercial actual.
- Registre los objetivos comerciales específicos que decidan los gerentes.
- Consensúe los criterios que se utilizarán para determinar el rendimiento del proceso de minería de datos desde una perspectiva comercial.

Ejemplo de venta en línea: búsqueda de los objetivos comerciales

Un ejemplo de minería Web utilizando CRISP-DM

A medida que las empresas realizan la transición para realizar las ventas en la Web, una reputada empresa de informática/electrónica está experimentando un crecimiento en la competencia de nuevos sitios. Al afrontar la realidad de que las tiendas Web están aflorando tan rápido (o más) que los clientes que utilizan la Web, la empresa debe encontrar fórmulas para mantener la rentabilidad, sin aumentar el coste de la adquisición de clientes. Una solución propuesta es cultivar las relaciones de los clientes actuales con objeto de aumentar al máximo el valor de cada uno de los clientes actuales.

Además se ha encargado un estudio con los siguientes objetivos:

- Mejorar las ventas cruzadas realizando mejores recomendaciones.
- Fomentar la lealtad de los clientes con un servicio más personalizado.

Provisionalmente, el estudio se considerará un éxito si:

- Las ventas cruzadas se aumentan en un 10%.
- Los clientes pasan más tiempo y ven más páginas en el sitio por cada visita.
- El estudio se completa dentro del plazo y del presupuesto.

Compilación de la información de la empresa

La comprensión de la situación comercial de su organización le ayudará a conocer su trabajo en términos de:

- Recursos disponibles (personal y material)

- Problemas
- Objetivos

Deberá examinar su situación comercial para encontrar respuestas reales a preguntas que puedan tener un impacto en el resultado del proyecto de minería de datos.

Tarea 1: Determinar la estructura de la organización
- Desarrolle gráficos de la organización para ilustrar divisiones corporativas, departamentos y grupos de proyectos. Asegúrese de incluir nombres y responsabilidades de los directivos.
- Identifique individuos clave de la organización.
- Identifique un patrocinador interno que proporcionará apoyo financiero y/o la experiencia en el dominio.
- Determine si existe un comité de dirección y procure una lista de miembros.
- Identifique las unidades comerciales que se verán afectadas por el proyecto de minería de datos.

Tarea 2: Describir el área problemática
- Identifique el área problemática, como marketing, atención al cliente o desarrollo comercial.
- Describa el problema de forma general.
- Describa los requisitos previos del proyecto. ¿Cuáles son las motivaciones que contiene el proyecto? ¿La empresa utiliza ya la minería de datos?
- Compruebe el estado del proyecto de minería de datos en el grupo comercial. ¿Se ha aprobado el esfuerzo o es necesario "publicitarlo" como tecnología clave para el grupo comercial?
- Si fuera necesario, prepare presentaciones informativas sobre minería de datos para su organización.

Tarea 3: Describir la solución actual
- Describa las soluciones actuales para resolver el problema comercial.
- Describa las ventajas y desventajas de la solución actual. Además, describa el nivel de aceptación de la solución en la organización.

Definición de los objetivos comerciales

Es el punto en el que las soluciones se especifican. Como resultado de sus investigaciones y reuniones, debe crear un objetivo principal concreto acordado por los patrocinadores del proyecto y otras unidades comerciales que se vean afectadas por los resultados. Este objetivo se puede traducir de forma eventual de algo tan nebuloso como "reducir la tasa de abandono de clientes" a objetivos específicos de minería de datos que dirigirán el análisis.

Lista de tareas

Asegúrese de que toma notas de los siguientes puntos para incorporarlas posteriormente a su plan de proyecto. Recuerde que los objetivos deben ser realistas.
- Describa el problema que desea resolver mediante la minería de datos.
- Especifique todas las cuestiones comerciales de la forma más específica posible.
- Determine otros requisitos comerciales (como no perder los clientes actuales a la vez que aumenta las oportunidades de ventas cruzadas).
- Especifique los beneficios esperados en términos comerciales (como reducir la tasa de abandono entre los clientes de mayor valor en un 10 %).

Criterios de rendimiento comercial

El objetivo puede ser claro, pero ¿lo reconocerá llegado el momento? Es importante definir la naturaleza del rendimiento comercial del proyecto de minería de datos antes de continuar. Los criterios de rendimiento se dividen en dos categorías:

- **Objetivo.** Estos criterios pueden ser tan simples como un aumento específico en la precisión de las auditorías o una reducción acordada de abandono de clientes.
- **Subjetivo.** Los criterios subjetivos como "descubrir clústeres de tratamientos efectivos" son más difíciles de precisar, pero puede acordar quién toma la decisión final.

Lista de tareas

- De la forma más precisa que sea posible, documente los criterios de rendimiento del proyecto.
- Asegúrese de que cada objetivo comercial tiene un criterio de rendimiento relacionado.
- Alinee los criterios de las mediciones subjetivas del rendimiento. Si es posible, documente las expectativas.

Evaluación de la situación

Ahora que ha definido un objetivo comercial claro, es hora de realizar una evaluación de su situación actual. Este paso implica cuestiones como:

- ¿Qué tipos de datos están disponibles para el análisis?
- ¿Dispone del personal necesario para completar el proyecto?
- ¿Cuáles son los principales factores de riesgo?
- ¿Dispone de planes de contingencia para cada factor de riesgo?

Ejemplo de venta en línea: evaluación de la situación

Un ejemplo de minería Web utilizando CRISP-DM

Es el primer intento en minería Web del comerciante y la empresa ha decidido consultar a un especialista de minería de datos que le ayudará a iniciar el proceso. Una de las primeras tareas que encara el asesor es valorar los recursos de minería de datos de la empresa.

Personal. Es evidente que hay experiencia con la gestión de registros de servidor y bases de datos de producto y compras, pero poca experiencia en almacenamiento de datos y limpieza de datos para análisis. Además, también se puede consultar a un especialista en base de datos. Como la empresa espera que los resultados del estudio formarán parte de un proceso continuado de minería Web, la gestión también debe considerar si alguna de las posiciones creadas durante las actividades actuales serán permanentes.

Datos. Como se trata de una empresa consolidada, existen multitud de datos de registro Web y de compras para incorporar. De hecho, para el estudio inicial, la empresa restringirá el análisis a los clientes que se hayan "registrado" en el sitio. Si es satisfactorio, el programa se puede ampliar.

Riesgos. Además de los desembolsos monetarios en asesores y el tiempo que han dedicado los empleados al estudio, no existen otros riesgos inmediatos en el proyecto. Sin embargo, el tiempo es siempre importante, por lo que el proyecto inicial se programa para un único trimestre.

Además, no se dispone de liquidez extra por el momento, por lo que es esencial que el estudio se complete dentro del presupuesto. Si alguno de estos objetivos se pone en peligro, los directivos comerciales han sugerido que el ámbito del proyecto se debe reducir.

Inventario de recursos

Es indispensable que mantenga un inventario actualizado de los recursos. Puede ahorrar mucho tiempo y disgustos si valora el hardware, los orígenes de datos y los problemas personales.

Tarea 1: Valorar los recursos de hardware

- ¿Qué tipo de hardware necesita?

Tarea 2: Identificar orígenes de datos y almacenes de conocimientos

- ¿Qué tipo de orígenes de datos están disponibles en la minería de datos? Documente los tipos y formatos de datos.
- ¿Cómo se almacenan los datos? ¿Dispone de acceso directo a los almacenes de datos o a las bases de datos operativas?
- ¿Ha planificado adquirir datos externos, como información demográfica?
- ¿Existen problemas de seguridad que evitan el acceso a los datos necesarios?

Tarea 3: Identificar recursos personales

- ¿Dispone de acceso a empresas y expertos de datos?
- ¿Ha identificado administradores de base de datos y otro personal de apoyo que pueda identificar?

Una vez haya respondido a estas preguntas, incluya una lista de contactos y recursos del informe de fase.

Requisitos, supuestos y restricciones

Es más probable que sus esfuerzos tengan recompensa si valora correctamente las responsabilidades del proyecto. Si valora estos problemas de la forma más explícita posible ayudará a identificar problemas futuros.

Tarea 1: Determinar requisitos

El requisito fundamental es el objetivo comercial mencionado con anterioridad, pero tenga en cuenta los siguientes elementos:

- ¿Existen restricciones legales y de seguridad sobre los datos o resultados del proyecto?
- ¿Todos los usuarios están alineados con los requisitos de planificación del proyecto?
- ¿Existen requisitos sobre el despliegue de los resultados (por ejemplo, publicación en la Web o lectura de resultados en una base de datos)?

Tarea 2: Describir los supuestos

- ¿Existen factores económicos que pueden afectar al proyecto (por ejemplo, honorarios de asesoría o productos de la competencia)?
- ¿Existen supuestos de calidad de datos?
- ¿Cómo espera el patrocinador/equipo de dirección del proyecto ver los resultados? En otras palabras, ¿pretenden comprender el modelo o únicamente visualizar los resultados?

Tarea 3: Comprobar las restricciones

- ¿Dispone de todas las contraseñas necesarias para acceder a los datos?
- ¿Ha comprobado todas las restricciones legales sobre el uso de los datos?
- ¿Las restricciones financieras están incluidas en el presupuesto del proyecto?

Riesgos y contingencias

También es buena idea considerar posibles riesgos que puedan surgir en el curso del proyecto. Entre los tipos de riesgos se incluyen:

- Programación (¿Qué sucede si el proyecto dura más de lo programado?)
- Financieros (¿Qué sucede si el patrocinador del proyecto detecta problemas presupuestarios?)
- Datos (¿Qué sucede si los datos son de escasa calidad o cobertura?)
- Resultados (¿Qué sucede si los resultados iniciales son menos dramáticos que los esperados?)

Una vez haya considerado los diferentes riesgos, elabore un plan de contingencia para ayudar a detectar los problemas.

Lista de tareas

- Documente cada uno de los posibles riesgos.
- Elabore un plan de contingencia para cada uno de los riesgos.

Terminología

Para asegurar que los equipos comerciales y de minería de datos "hablan el mismo idioma", debe considerar elaborar un glosario de los términos técnicos y expresiones que necesiten clarificación. Por ejemplo, si "abandono" tiene un significado particular y exclusivo para su empresa, es mejor especificarlo para el beneficio del equipo. Del mismo modo, el equipo puede beneficiarse de la explicación del uso de un gráfico de ganancias.

Lista de tareas

- Mantenga una lista actualizada de términos o vocablos que sean confusos para miembros del equipo. Incluya datos comerciales y terminología de minería de datos.
- Considere publicar la lista en la intranet o en otra documentación del proyecto.

Análisis de costes/beneficios

En este paso se responde a la pregunta **¿Cuáles son sus prioridades?** Como parte final de la evaluación, es esencial comparar los costes del proyecto con los beneficios potenciales del rendimiento.

Lista de tareas

Incluya los costes estimados del análisis de:

- Recopilación de datos y cualquier dato externo utilizado
- Despliegue de los resultados
- Costes operativos

A continuación, considere los beneficios de:

- El objetivo principal completado
- Conocimientos adicionales adquiridos de la exploración de los datos
- Posibles beneficios de una mejor comprensión de los datos

Determinación de los objetivos de minería de datos

Ahora que el objetivo comercial ha quedado claro, es hora de traducirlo en una realidad de minería de datos. Por ejemplo, el objetivo comercial para "reducir el abandono" se puede traducir en un objetivo de minería de datos que incluye:

- Identificación de clientes de mayor valor en función de los datos de compra recientes
- Creación de un modelo utilizando datos disponibles de clientes para predecir las posibilidades de abandono de cada cliente
- Asignación de un rango a cada cliente basado en las posibilidades de abandono y valor del cliente

Estos objetivos de minería de datos, si se cumplen, se pueden utilizar por la empresa para reducir el abandono entre los clientes de mayor valor.

Como puede comprobar, la tecnología y la empresa deben ir de la mano para que el proyecto de minería de datos sea efectivo. Consulte las sugerencias específicas sobre cómo determinar los objetivos de minería de datos.

Objetivos de minería de datos

Cuando trabaje con analistas comerciales y de datos para definir una solución técnica al problema comercial, recuerde que los objetivos deben ser concretos.

Lista de tareas

- Describa el **tipo** de problema de minería datos, como clúster, predicción o clasificación.
- Documente objetivos técnicos utilizando unidades específicas de tiempo, como predicciones con una validez de tres meses.
- Si es posible, proporcione datos reales para resultados deseados, como producir resultados de abandono para el 80 % de los clientes actuales.

Ejemplo de venta en línea: Objetivos de minería de datos

Un ejemplo de minería Web utilizando CRISP-DM

Con la ayuda del asesor en minería de datos, el comerciante ha podido traducir los objetivos comerciales de la empresa en términos de minería de datos. Los objetivos del estudio inicial que se deben cumplir en este trimestre son:

- Utilice información histórica acerca de compras anteriores para generar un modelo que enlace elementos "relacionados". Si los usuarios consultan la descripción de un elemento, proporcione enlaces a otros elementos del grupo relacionado (**análisis de la cesta de compra**).
- Utilice registros Web para determinar que intentan buscar los diferentes clientes y rediseñe el sitio para resaltar estos elementos. Cada "tipo" de cliente diferente verá una página de inicio diferente del sitio (**perfil del sitio**).
- Utilice registros Web para predecir la siguiente persona, teniendo en cuenta su procedencia y si ha estado en su sitio (**análisis de secuencias**).

Criterios de rendimiento de minería de datos

El rendimiento también se puede definir en términos técnicos para mantener actualizados sus esfuerzos de minería de datos. Utilice el objetivo de minería de datos determinado anteriormente para formular puntos de referencia del rendimiento. IBM SPSS Modeler proporciona herramientas como el nodo de evaluación y el nodo de análisis, que le ayudarán a analizar la precisión y la validez de los resultados.

Lista de tareas

- Describa los métodos de evaluación de modelos (por ejemplo, precisión, rendimiento, etc.).
- Defina los puntos de referencia para evaluar el rendimiento. Proporcione datos numéricos concretos.
- Defina las mediciones subjetivas lo mejor posible y determine los criterios de rendimiento.
- Considere si el despliegue satisfactorio de los resultados del modelo forma parte de su rendimiento de minería de datos. Comience ahora a planificar el despliegue.

Producción de un plan de proyecto

En este punto, ya está listo para producir un plan para el proyecto de minería de datos. Las cuestiones que haya planteado hasta el momento y los objetivos comerciales y de minería de datos que haya formulado formarán la base de este plan.

Escritura del plan de proyecto

El plan de proyecto es el documento principal del trabajo de minería de datos. Si se elabora correctamente, puede informar a todos los usuarios relacionados con los objetivos, recursos, riesgos del proyecto y programar todas las fases de minería de datos. Es posible que desee publicar el plan, así como la documentación recopilada en esta fase en la intranet de la empresa.

Lista de tareas

Cuando cree el plan, asegúrese de que puede responder a las siguientes preguntas:

- ¿Ha discutido las tareas del proyecto y ha propuesto el plan con todos los usuarios implicados?
- ¿Se incluyen estimaciones de tiempo de todas las fases o tareas?
- ¿Ha incluido los esfuerzos y recursos necesarios para desplegar los resultados o soluciones comerciales?
- ¿Se resaltan los puntos de decisión y solicitudes de revisión en el plan?
- ¿Ha destacado las fases en las que suelen ocurrir iteraciones, como modelado?

Plan de proyecto de muestra

El plan resumido del estudio es como el que se muestra en la tabla siguiente.

Tabla 1. Visión general del plan de proyecto de ejemplo

Fase	Hora	Recursos	Riesgos
Comprensión del negocio	1 semana	Todos los analistas	Cambio económico
Comprensión de los datos	3 semanas	Todos los analistas	Problemas de datos, problemas tecnológicos
Preparación de los datos	5 semanas	Asesor de minería de datos, tiempo de análisis de base de datos	Problemas de datos, problemas tecnológicos
Modelado	2 semanas	Asesor de minería de datos, tiempo de análisis de base de datos	Problemas de tecnología, incapacidad para encontrar un modelo adecuado
Evaluación	1 semana	Todos los analistas	Cambio económico, incapacidad para implementar resultados
Despliegue	1 semana	Asesor de minería de datos, tiempo de análisis de base de datos	Cambio económico, incapacidad para implementar resultados

Evaluación de herramientas y técnicas

Como ya ha seleccionado utilizar IBM SPSS Modeler como herramienta de rendimiento de minería de datos, puede utilizar este paso para buscar las técnicas de minería de datos más adecuadas para sus necesidades comerciales.IBM SPSS Modeler ofrece una amplia gama de herramientas para cada fase de minería de datos. Para decidir cuándo debe utilizar las diferentes técnicas, consulte la sección de modelado de la ayuda en línea.

¿Está listo para el siguiente paso?

Antes de explorar los datos y comenzar a trabajar en IBM SPSS Modeler, asegúrese de que ha respondido a las siguientes preguntas.

Desde una perspectiva comercial:

- ¿Qué espera obtener de este proyecto?
- ¿Cómo define la finalización de los trabajos?
- ¿Dispone de la dotación presupuestaria y de los recursos necesarios para completar los objetivos?
- ¿Dispone de acceso a todos los datos necesarios para el proyecto?
- ¿Ha tratado con su equipo los riesgos y contingencias asociadas con el proyecto?
- ¿Los resultados del análisis de costes/beneficios hacen que el proyecto sea viable?

Una vez haya respondido a las preguntas anteriores, ¿ha podido traducir las respuestas en objetivos de minería de datos?

Desde una perspectiva de minería de datos:

- ¿En qué forma puede ayudarle la minería de datos a cumplir sus objetivos comerciales?
- ¿Sabe qué técnicas de minería de datos producen los mejores resultados?
- ¿Cómo puede saber que sus resultados son precisos o efectivos? (*¿Hemos definido una medición del rendimiento de la minería de datos?*)
- ¿Cómo se desplegarán los resultados de modelado? ¿Ha considerado el despliegue en su plan de proyecto?
- ¿El plan de proyecto incluye todas las fases de CRISP-DM?
- ¿Los riesgos y dependencias se incluyen en el plan?

Si puede responder afirmativamente a todas las preguntas anteriores, está listo para observar los datos más de cerca.

Capítulo 3. Comprensión de los datos

Conceptos básicos sobre comprensión de datos

La fase de comprensión de datos de CRISP-DM implica estudiar más de cerca los datos disponibles de minería. Este paso es esencial para evitar problemas inesperados durante la siguiente fase (preparación de datos) que suele ser la fase más larga de un proyecto.

La comprensión de datos implica acceder a los datos y explorarlos con la ayuda de tablas y gráficos que se pueden organizar en IBM SPSS Modeler utilizando la herramienta de proyectos CRISP-DM. De esta forma podrá determinar la calidad de los datos y describir los resultados de estos pasos en la documentación del proyecto.

Recopilación de datos iniciales

En este punto en CRISP-DM, puede acceder a los datos e incluirlos en IBM SPSS Modeler. Los datos provienen de diversos orígenes como:

- **Datos existentes.** Incluye una amplia variedad de datos, como datos transaccionales, datos de encuesta, registros Web, etc. Tenga en cuenta si los datos existentes son suficientes para adaptarse a sus necesidades.
- **Datos adquiridos.** ¿Su organización utiliza datos adicionales, como datos demográficos? Si no los utiliza, considere si son necesarios.
- **Datos adicionales.** Si los orígenes anteriores no satisfacen sus necesidades, es posible que necesite realizar encuestas o realizar seguimientos adicionales para servir de complemento a los almacenes de datos actuales.

Lista de tareas

Observe los datos de IBM SPSS Modeler y considere las siguientes cuestiones. Asegúrese de tomar notas sobre sus hallazgos. Consulte el tema “Escritura de un informe de recopilación de datos” en la página 14 para obtener más información.

- ¿Qué atributos (columnas) de la base de datos parecen más prometedores?
- ¿Qué atributos no parecen relevantes y se pueden excluir?
- ¿Existen datos suficientes para obtener conclusiones generales o realizar predicciones precisas?
- ¿Dispone de atributos suficientes para su método de modelado?
- ¿Está fusionando varios orígenes de datos? En caso afirmativo, ¿existen áreas que puedan plantear problemas al fusionar?
- ¿Ha considerado cómo se gestionan los valores perdidos en cada origen de datos?

Ejemplo de venta en línea: recopilación inicial de datos

Un ejemplo de minería Web utilizando CRISP-DM

El comerciante de este ejemplo utiliza varios orígenes de datos de importancia, incluyendo:

Registros Web. Los registros de acceso brutos contienen toda la información de cómo los clientes navegan por el sitio Web. Es necesario eliminar referencias a archivos de imágenes y entradas no informativas en los registros Web como parte del proceso de preparación de datos.

Adquisición de datos. Si un cliente envía un pedido, se guarda toda la información relativa a ese pedido. Los pedidos de la base de datos de adquisiciones se deben correlacionar a las sesiones correspondientes en los registros Web.

Base de datos de productos. Los atributos de productos pueden ser de gran utilidad cuando determine productos "relacionados". Es necesario correlacionar la información de productos a los pedidos correspondientes.

Base de datos de clientes. Esta base de datos contiene información adicional recopilada de clientes registrados. Los registros no son completos ya que muchos clientes no completan los cuestionarios. Es necesario correlacionar la información de los clientes a las adquisiciones y sesiones correspondientes en los registros Web.

En este momento, la empresa no tiene planes de adquirir bases de datos externas o realizar inversiones para realizar encuestas, porque sus analistas están ocupados gestionando los datos de que disponen actualmente. Sin embargo, en algún momento pueden considerar un despliegue ampliado de los resultados de minería de datos, en cuyo caso la adquisición de datos demográficos adicionales de clientes sin registrar puede ser muy útil. También es muy útil disponer de información demográfica para comprobar las diferencias entre la base de clientes del comerciante y del comprador Web medio.

Escritura de un informe de recopilación de datos

Si utiliza el material recopilado en el paso anterior, puede comenzar a escribir el informe de recopilación de datos. Una vez completado, el informe se puede añadir al sitio Web del proyecto o distribuir al equipo. También se puede combinar con los informes preparados en los pasos siguientes; descripción de datos, exploración y verificación de la calidad. Estos informes le guiarán por la fase de preparación de los datos.

Descripción de los datos

Existen muchas formas de describir datos, pero la mayoría de datos se centra en la cantidad y calidad de los datos; la cantidad de datos disponible y el estado de los datos. A continuación se incluyen algunas características clave para describir datos.

- **Cantidad de datos.** En la mayoría de técnicas de modelado, los tamaños de datos tienen un equilibrio relacionado. Los grandes conjuntos de datos pueden producir modelos más precisos, pero también pueden aumentar el tiempo de procesamiento. Considere utilizar un subconjunto de datos. Cuando tome notas para el informe final, asegúrese de incluir estadísticos de tamaños para todos los conjuntos de datos y recuerde tener en cuenta tanto el número de registros como los campos (atributos) cuando describa los datos.
- **Tipos de valores.** Los datos pueden incluir una variedad de formatos, como **numérico**, **categórico** (cadena) o **Booleano** (verdadero/falso). Si presta atención al tipo de valor puede evitar posteriores problemas durante la fase de modelado.
- **Esquemas de codificación.** Con frecuencia, los valores de la base de datos son representaciones de características como género o tipo de producto. Por ejemplo, un conjunto de datos puede utilizar *M* y *F* para representar *masculino* y *femenino*, mientras que otro puede utilizar los valores numéricos *1* y *2*. Registre los esquemas incoherentes en el informe de datos.

Teniendo en cuenta este conocimiento, ahora está listo para escribir el informe de descripción de datos y compartir sus descubrimientos con más usuarios.

Ejemplo de venta en línea: descripción de los datos

Un ejemplo de minería Web utilizando CRISP-DM

Existen multitud de registros y atributos para procesar en una aplicación de minería Web. Aunque el comerciante que realice el proyecto de minería de datos haya limitado el estudio inicial a unos 30.000 clientes aproximadamente, que se hayan registrado en el sitio, aún quedan millones de registros en los registros Web.

La mayoría de tipos de valor de estos orígenes de datos son simbólicos, ya sean fechas y horas, accesos de páginas Web o respuestas a preguntas de opciones múltiples del cuestionario de registro. Algunas de estas variables se utilizarán para crear nuevas variables numéricas, como el número de páginas Web visitadas y el tiempo que se ha permanecido en el sitio Web. Las pocas variables numéricas existentes en los conjuntos de datos incluyen el número de cada producto solicitado, la cantidad gastada durante una compra y las especificaciones de ponderación y dimensiones de la base de datos del producto.

Los esquemas de codificación de los diferentes orígenes de datos se solapan muy poco, porque los orígenes de datos contienen atributos muy diferentes. Las únicas variables que se solapan son "claves", como las ID de clientes y códigos de productos. Estas variables deben tener esquemas de codificación idénticos desde un origen de los datos a otro; de otro modo será imposible fundir los orígenes de datos. Deberá realizar una preparación adicional de los datos para volver a codificar estos campos clave para fusionar.

Escritura de un informe de descripción de datos

Para completar de forma efectiva su proyecto de minería de datos, considere el valor de producir un informe de descripción de datos precisos utilizando las siguientes medidas:

Cantidad de datos

- ¿Cuál es el formato de los datos?
- Identifique el método utilizado para capturar los datos, por ejemplo, ODBC.
- ¿Qué dimensiones tiene la base de datos (en números de filas y columnas)?

Calidad de datos

- ¿Incluyen los datos características relativas a la cuestión comercial?
- ¿Qué tipos de datos están presentes (simbólicos, numéricos, etc.)?
- ¿Ha realizado una cálculo básico de los estadísticos de los atributos clave? ¿Qué conocimientos ha proporcionado este proceso en cuestiones comerciales?
- ¿Puede priorizar los atributos relevantes? En caso negativo, ¿los analistas empresariales pueden proporcionar más conocimientos?

Exploración de datos

Utilice esta fase de CRISP-DM para explorar los datos con las tablas, gráficos y otras herramientas de visualización disponibles en IBM SPSS Modeler. Estos análisis pueden ayudarle a describir los objetivos de minería de datos generados durante la fase de comprensión comercial. También pueden ayudarle a formular hipótesis y dar forma a las tareas de transformación de datos que tienen lugar durante la preparación de los datos.

Ejemplo de venta en línea: exploración de los datos

Un ejemplo de minería Web utilizando CRISP-DM

Aunque CRISP-DM sugiere realizar una exploración inicial llegados a este punto, este proceso es difícil, si no imposible, en registros Web brutos, tal y como nuestro comerciante ha descubierto. Normalmente, los datos de registros Web se deben procesar en primer lugar en la fase de preparación de datos para producir los datos que se pueden explorar de forma significativa. Esta desviación de CRISP-DM subraya

el hecho de que el proceso se puede y debe personalizar de acuerdo con sus necesidades de minería de datos concretas. CRISP-DM es un proceso cíclico y los analistas de datos suelen retroceder y avanzar entre fases.

Aunque los registros Web se deben procesar antes de la exploración, el resto de orígenes disponibles para el comerciante son más fáciles de explorar. Si utiliza la base de datos de compras para la exploración revela información interesante acerca de los clientes, como las cantidades gastadas, la cantidad de artículos que compran en cada visita y su procedencia. La información de la base de datos de los clientes muestra la distribución de las respuestas de los artículos del cuestionario de registro.

La exploración también es de gran utilidad para buscar errores en los datos. Mientras que la mayoría de los orígenes de datos se generan automáticamente, la información de la base de datos de productos se introduce manualmente. Algunos resúmenes rápidos de dimensiones de productos ayudarán a descubrir errores tipográficos, como monitor de "119 pulgadas" (en lugar de monitor de "19 pulgadas").

Escritura de un informe de exploración de datos

Cuando cree gráficos y ejecute estadísticos sobre los datos disponibles, formule hipótesis acerca de cómo los datos pueden adaptarse a los objetivos técnicos y comerciales.

Lista de tareas

Registre sus descubrimientos para su inclusión en el informe de exploración de datos. Compruebe que responde a las cuestiones siguientes:

- ¿Qué tipo de hipótesis sobre los datos ha formulado?
- ¿Qué atributos parecen ser prometedores de cara a futuros análisis?
- ¿Ha realizado exploraciones que revelen nuevas características de los datos?
- ¿En qué forma han cambiado estas exploraciones su hipótesis inicial?
- ¿Puede identificar subconjuntos concretos de datos para un uso posterior?
- Vuelva a comprobar sus objetivos de minería de datos. ¿Esta exploración ha modificado sus objetivos?

Verificación de calidad de datos

Los datos no suelen ser perfectos. De hecho, la mayoría de los datos contienen errores de codificación, valores perdidos u otro tipo de incoherencias que hacen que los análisis resulten difíciles en algunas ocasiones. Una forma de evitar posibles problemas es realizar un análisis de calidad de los datos disponibles antes de proceder al modelado.

Las herramientas de informes de IBM SPSS Modeler (como el nodo Auditoría de datos, el nodo Tabla y otros nodos de resultados) pueden ayudarle en la búsqueda de los siguientes tipos de problemas:

- **Los datos perdidos** incluyen valores vacíos o codificados como sin respuesta (como *$null$*, *?* o *999*).
- **Los errores de datos** suelen ser errores tipográficos cometidos al introducir los datos.
- **Los errores de mediciones** incluyen datos que se introducen correctamente, pero se basan en un esquema de mediciones incorrecto.
- **Las incoherencias de codificación** suelen incluir unidades no estándar de medida o valores incoherentes, como el uso de *M* y *masculino* para expresar el género.
- **Los metadatos erróneos** incluyen errores entre el significado aparente de un campo incluido en un nombre o definición de campo.

Asegúrese de anotar estos problemas de calidad. Consulte el tema “Escritura de un informe de calidad de datos” en la página 17 para obtener más información.

Ejemplo de venta en línea: verificación de la calidad de los datos

Un ejemplo de minería Web utilizando CRISP-DM

La verificación de la calidad de los datos se suele realizar durante el curso de los procesos de descripción y exploración. Algunos de los problemas detectados por el comerciante suelen incluir:

Datos perdidos. Los datos perdidos conocidos incluyen cuestionarios sin responder por parte de algunos usuarios registrados. Sin la información extra que proporciona este cuestionario, estos clientes se pueden omitir en algunos de los siguientes modelos.

Errores de datos. La mayoría de los orígenes de datos se generan automáticamente, por lo que no es un problema grave. Los errores tipográficos de la base de datos de producto se pueden detectar durante el proceso de exploración.

Errores de mediciones. El origen principal de los errores de mediciones es el cuestionario. Si alguno de los elementos no está cumplimentado correctamente, es posible que no proporcione la información que el comerciante espera obtener. De nuevo, durante el proceso de exploración, es importante prestar una especial atención a los elementos que tienen una distribución inusual de las respuestas.

Escritura de un informe de calidad de datos

En función de los resultados de su exploración y verificación de calidad de datos, ya está listo para preparar un informe que le guiará por la siguiente fase de CRISP-DM. Consulte el tema “Verificación de calidad de datos” en la página 16 para obtener más información.

Lista de tareas

Tal y como se ha mencionado anteriormente, existen diferentes tipos de problemas de calidad de datos. Antes de pasar a la fase siguiente, tenga en cuenta los siguientes problemas de calidad y planifique una solución. Registre todas sus respuestas en el informe de calidad de datos.

- ¿Ha identificado atributos perdidos y campos vacíos? Si es así, ¿los valores perdidos tienen significado?
- ¿Existen faltas de ortografía que puedan causar problemas en futuras fusiones o transformaciones?
- ¿Ha detectado desviaciones para determinar si son "ruido" o fenómenos que merecen un análisis en profundidad?
- ¿Ha realizado una comprobación correcta de los valores? Registre todos los conflictos aparentes (como adolescentes con ingresos muy elevados).
- ¿Ha considerado excluir los datos que no tengan ninguna influencia en sus hipótesis?
- ¿Los datos están almacenados en archivos planos? Si es así, ¿los delimitadores mantienen la coherencia entre los archivos? ¿Contiene cada registro el mismo número de campos?

¿Está listo para el siguiente paso?

Antes de preparar los datos para su modelado en IBM SPSS Modeler, considere los siguientes aspectos:

¿Cuál es su nivel de comprensión de los datos?

- ¿Ha identificado y accedido correctamente a todos los orígenes de datos? ¿Ha tenido algún problema o restricción de algún tipo?
- ¿Ha identificado atributos clave de los datos disponibles?
- ¿Le han ayudado estos atributos a formular hipótesis?
- ¿Ha detectado el tamaño de todos los orígenes de datos?
- ¿Puede utilizar un subconjunto de datos cuando lo estime conveniente?
- ¿Ha calculado los estadísticos básicos de cada atributo de su interés? ¿Ha obtenido información de interés?

- ¿Ha utilizado gráficos de exploración para obtener atributos clave? ¿Este conocimiento ha reformulado alguna de sus hipótesis?
- ¿Cuáles fueron los problemas de calidad de datos del proyecto? ¿Tiene una planificación para resolver estos problemas?
- ¿Las fases de preparación de los datos son claras? Por ejemplo, ¿sabe qué orígenes de datos debe fusionar y los atributos que debe filtrar o seleccionar?

Ahora que ha completado la comprensión comercial y de datos, es hora de utilizar IBM SPSS Modeler para preparar sus datos para el modelado.

Capítulo 4. Preparación de los datos

Conceptos básicos sobre preparación de datos

La preparación de datos es uno de los aspectos más importantes y con frecuencia que más tiempo exigen en la minería de datos. De hecho, se estima que la preparación de datos suele llevar el 50-70 % del tiempo y esfuerzo de un proyecto. Dedicar los esfuerzos adecuados a las primeras fases de comprensión comercial y comprensión de datos puede reducir al mínimo los gastos indirectos relacionados, pero aún deberá dedicar una buena cantidad de esfuerzo para preparar y empaquetar los datos para la minería.

Dependiendo de su organización y sus objetivos, la preparación de datos suele implicar las tareas siguientes:

- Fusión de conjuntos y/o registros de datos
- Selección de una muestra de un subconjunto de datos
- Agregación de registros
- Derivación de nuevos atributos
- Clasificación de los datos para el modelado
- Eliminación o sustitución de valores en blanco o perdidos
- División en conjuntos de datos de prueba y entrenamiento

Selección de datos

En función de la recopilación de datos inicial realizada en la fase CRISP-DM anterior, ahora puede comenzar a seleccionar los datos relevantes a sus objetivos de minería de datos. De forma general, existen dos formas de seleccionar datos:

- **Selección de elementos (filas)** implica la toma de decisiones como las cuentas, productos o clientes que se van a incluir.
- **Selección de atributos o características (columnas)** implica la toma de decisiones sobre el uso de características como la cantidad de las transacciones o los ingresos por hogar.

Ejemplo de venta en línea: selección de datos

Un ejemplo de minería Web utilizando CRISP-DM

Muchas de las decisiones del comerciante acerca de los datos que va a seleccionar ya se han tomado en fases anteriores del proceso de minería de datos.

Selección de elementos. El estudio inicial se limitará a los (aproximadamente) 30.000 clientes registrados en el sitio, por lo que es necesario configurar los filtros para que excluyan las compras y los registros Web de clientes sin registrar. Otros filtros se deben configurar para eliminar llamadas a archivos de imágenes y otras entradas no informativas de los registros Web.

Selección de atributos. La base de datos de adquisiciones contendrá información confidencial de los clientes del comerciante, por lo que es importante filtrar los atributos como nombre del cliente, dirección, número de teléfono y números de tarjeta de crédito.

Inclusión o exclusión de datos

Cuando decide los subconjuntos de datos que se van a incluir o excluir, asegúrese de que documenta los motivos de sus decisiones.

Cuestiones que debe tener en cuenta

- ¿Existe un atributo relacionado con sus objetivos de minería de datos?
- ¿La calidad de un conjunto o atributo de datos concreto excluye la validez de los resultados?
- ¿Puede recuperar estos datos?
- ¿Existen limitaciones acerca del uso de campos concretos como *género* o *raza*?

¿Existen decisiones que son diferentes de las hipótesis formuladas en la fase de comprensión de datos? Si es así, asegúrese de que documenta las razones en el informe de proyectos.

Limpieza de datos

La limpieza de datos implica observar más de cerca los problemas en los datos que ha seleccionado incluir en el análisis. Existen diferentes formas de limpiar los datos utilizando los nodos de registro y de operaciones con campos de IBM SPSS Modeler.

Tabla 2. Limpieza de datos

Problema de datos	**Solución posible**
Datos perdidos	Excluya las filas o características. O cumpliméntelas con un valor estimado.
Errores de datos	Utilice recursos lógicos para descubrir errores manuales y corríjalos. O, excluya las características.
Incoherencias de codificación	Decida un esquema de codificación simple y convierta y sustituya los valores.
Metadatos perdidos o erróneos	Examine manualmente los campos sospechosos y compruebe el significado correcto.

El informe de calidad de datos preparado durante la fase de comprensión de datos contiene detalles sobre los tipos de problemas concretos de sus datos. Puede utilizarlo como punto de inicio para la manipulación de datos en IBM SPSS Modeler.

Ejemplo de venta en línea: limpieza de los datos

Un ejemplo de minería Web utilizando CRISP-DM

El comerciante utiliza el proceso de limpieza de datos para solucionar los problemas detectados en el informe de calidad de datos.

Datos perdidos. Es posible que los clientes que no han completado el cuestionario en línea se tengan que omitir de los modelos posteriores. Es necesario volver a pedir a estos clientes que cumplimenten el cuestionario, pero esta solución requiere dedicar tiempo y dinero que es posible que el comerciante no pueda invertir. Lo que el comerciante puede hacer es modelar las diferencias de compras entre los clientes que responden y los que no responden el cuestionario. Si estos dos conjuntos de clientes tienen hábitos de compra similares, los cuestionarios que faltan son menos preocupantes.

Errores de datos. Los errores detectados durante el proceso de exploración se pueden corregir en esta fase. La mayoría de las veces, sin embargo, la entrada correcta de datos se realiza en el sitio Web antes de que un cliente envíe una página a la base de datos de back-end.

Errores de mediciones. Los elementos incorrectos del cuestionario pueden afectar en gran medida a la calidad de los datos. Al igual que los cuestionarios perdidos, se trata de un problema difícil, porque es posible que no se disponga del tiempo o recursos disponibles para recopilar las respuestas a una nueva pregunta. Para elementos problemáticos, la mejor solución puede ser volver al proceso de selección y filtrar estos elementos de análisis posteriores.

Escritura de un informe de limpieza de datos

Registrar sus actividades de limpieza de datos es esencial para registrar las modificaciones de los datos. Los futuros proyectos de minería de datos se beneficiarán de los detalles del trabajo disponible.

Lista de tareas

Es una excelente idea considerar las siguientes cuestiones cuando genere el informe:

- ¿Qué tipos de ruido se han producido en los datos?
- ¿Qué métodos utiliza para eliminar el ruido? ¿Qué técnicas han demostrado ser eficaces?
- ¿Existen casos o atributos que no se pueden recuperar? Asegúrese de registrar los datos que se han excluido por causas del ruido.

Construcción de nuevos datos

Con frecuencia, necesitará construir nuevos datos. Por ejemplo, puede ser de gran utilidad crear una nueva columna con la adquisición de una garantía ampliada en cada transacción. Este nuevo campo, *garantía_adquirida*, se puede generar fácilmente utilizando un nodo Marcas en IBM SPSS Modeler.

Existen dos formas de construir nuevos datos:

- Derivación de atributos (columnas o características)
- Generación de registros (filas)

IBM SPSS Modeler ofrece múltiples formas de construir datos utilizando sus nodos de registro y operaciones con campos.

Ejemplo de venta en línea: construcción de los datos

Un ejemplo de minería Web utilizando CRISP-DM

El procesamiento de registros Web puede crear múltiples registros nuevos. En los casos registrados, el comerciante puede preferir crear marcas de tiempo, identificar visitantes y sesiones y registrar la página a la que se ha accedido y el tipo de actividad del evento. Algunas de estas variables se utilizarán para crear más atributos, como los tiempos entre eventos en una sesión.

Se pueden crear más atributos como resultado de una fusión u otra reestructuración de los datos. Por ejemplo, si los registros Web de evento por fila se "acumulan" de forma que cada fila sea una sesión, se crearán nuevos atributos que registran el número total de acciones, el tiempo total empleado y el número total de compras realizadas durante la sesión. Si los registros Web se fusionan con la base de datos del cliente de forma que cada fila es un nuevo cliente, se crearán los nuevos atributos que registran el número de sesiones, el número total de acciones, el tiempo total empleado y el número de compras totales realizadas por cada cliente.

Después de la construcción de los nuevos datos, el comerciante debe realizar un proceso de exploración para comprobar que la creación de los datos se ha realizado correctamente.

Derivación de atributos

En IBM SPSS Modeler, puede utilizar los siguientes nodos de operaciones con campos para derivar nuevos atributos:

- Cree nuevos campos derivados de los actuales mediante un **nodo Derivar**.
- Cree un campo de marca mediante un **nodo Marcar**.

Lista de tareas

- Tenga en cuenta los requisitos de datos de modelado cuando derive atributos. ¿El algoritmo de modelado espera un tipo de datos concreto, como datos numéricos? En caso afirmativo, realice las transformaciones necesarias.
- ¿Necesita normalizar los datos antes de proceder con el modelado?
- ¿Se pueden construir los atributos que faltan mediante la agregación, media o inducción?
- En función de sus conocimientos, ¿existen hechos importantes (como la cantidad de tiempo en el sitio Web) que se puedan derivar de los campos existentes?

Integración de datos

No es raro disponer de varios orígenes de datos para el mismo conjunto de cuestiones comerciales. Por ejemplo, puede tener acceso a los datos de un crédito hipotecario, así como a los datos demográficos para el mismo conjunto de clientes. Si estos conjuntos de datos contienen el mismo identificador exclusivo (como el número de seguridad social), puede fusionarlos en IBM SPSS Modeler utilizando este campo de clave.

Existen dos métodos básicos para integrar los datos:

- La **fusión** de datos implica unir dos conjuntos de datos con registros similares, pero con atributos diferentes. Los datos se fusionan utilizando el mismo identificador clave en cada registro (como el ID de usuario). Los datos resultantes aumentan las columnas o las características.
- La **adición** de datos implica integrar dos o más conjuntos de datos con atributos similares, pero con registros diferentes. Los datos se integran en función de los campos similares (como el nombre de producto o la longitud del contrato).

Ejemplo de venta en línea: integración de datos

Un ejemplo de minería Web utilizando CRISP-DM

Con múltiples orígenes de datos, existen diferentes formas en las que el comerciante puede integrar los datos:

- **Adición de atributos de clientes y productos a datos de eventos.** Para modelar eventos de registros Web que utilicen atributos de otras bases de datos, cualquier ID de cliente, número de producto y número de orden de compra asociado con cada evento se debe identificar correctamente y los atributos correspondientes se deben fusionar con los registros Web procesados. Tenga en cuenta que el archivo replica la información del cliente y del producto cada vez que un cliente o producto se asocia con un evento.
- **Adición de información de compra y registro Web a los datos del cliente.** Para modelar el valor de un cliente, su información de compras y de sesión se debe extraer de las bases de datos adecuadas, totalizadas y fusionadas con la base de datos de clientes. Este método implica la creación de nuevos atributos, tal y como se explica en el proceso de construcción de datos.

Después de integrar las bases de datos, el comerciante debe realizar un proceso de exploración para comprobar que la fusión de los datos se ha realizado correctamente.

Tareas de integración

La interpretación de los datos puede ser un proceso complejo si no ha invertido la cantidad de tiempo adecuada en el desarrollo y comprensión de los datos. Dedique algún tiempo para reflexionar acerca de los elementos y atributos que parecen más relevantes a los objetivos de minería de datos y a continuación comience a integrar los datos.

Lista de tareas

- Utilizando los nodos Fundir o Añadir en IBM SPSS Modeler, integre los conjuntos de datos que considere útiles para el proceso de modelado.
- Considere guardar los resultados antes de proceder al proceso de modelado.

- Tras la fusión, los datos se pueden simplificar **agregando** valores. La agregación implica que los nuevos valores se calcularán resumiendo la información de diferentes registros y/o tablas.
- Es posible que tenga que generar nuevos registros (como el descuento medio de varios años de devoluciones de impuestos combinadas).

Formato de datos

Como paso final antes de la generación del modelo, es muy útil comprobar si algunas técnicas requieren aplicar un formato concreto o la clasificación de los datos. Por ejemplo, no es extraño que un algoritmo de secuencia requiera que los datos estén clasificados de forma previa antes de ejecutar el modelo. Incluso si el modelo puede ejecutar la clasificación de forma automática, puede ahorrar tiempo si utiliza un nodo Ordenar antes del modelado.

Lista de tareas

Considere las siguientes cuestiones cuando aplique formato a los datos:

- ¿Qué modelos ha planeado utilizar?
- ¿Estos modelos requieren un formato de datos o una clasificación concreta?

Si los cambios son recomendables, las herramientas de procesamiento de IBM SPSS Modeler pueden ayudarle a aplicar la manipulación necesaria de los datos.

¿Está listo para comenzar el proceso de modelado?

Antes de construir modelos en IBM SPSS Modeler, asegúrese de contestar las siguientes cuestiones.

- ¿Puede acceder a todos los datos de IBM SPSS Modeler?
- En función de los procesos de exploración y comprensión iniciales, ¿ha podido seleccionar los subconjuntos relevantes de datos?
- ¿Ha limpiado los datos de forma efectiva o eliminado los elementos que no se pueden guardar? Registre todas las decisiones en el informe final.
- ¿Los diferentes conjuntos de datos se integran correctamente? ¿Existen problemas de fusión que se deben mencionar?
- ¿Conoce las herramientas de modelado necesarias que ha planificado utilizar?
- ¿Existen problemas de formato que deba solucionar antes del proceso de modelado? Se incluyen problemas de formato y tareas que pueden reducir el tiempo de modelado.

Si puede responder las cuestiones anteriores, está listo para pasar al proceso clave de modelado de minería de datos.

Capítulo 5. Modelado

Conceptos básicos sobre modelado

Este es el punto donde todo el duro trabajo anterior comienza a tener sentido. Los datos que ha preparado se incorporan a las herramientas analíticas de IBM SPSS Modeler y los resultados comenzarán a arrojar algo de luz al problema planteado en Comprensión del negocio.

El modelado se suele ejecutar en múltiples iteraciones. Normalmente, los analistas de datos ejecutan varios modelos utilizando los parámetros predeterminados y ajustan los parámetros o vuelven a la fase de preparación de datos para las manipulaciones necesarias por su modelo. Es extraño que las cuestiones relativas a la minería datos de una empresa se solucionen satisfactoriamente con un modelo y ejecución únicos. Esto es lo que hace la minería de datos tan interesante; existen muchas formas para resolver un problema concreto y IBM SPSS Modeler ofrece una amplia variedad de herramientas para ello.

Selección de técnicas de modelado

Aunque pueda tener algunos conocimientos acerca de los tipos de modelado que sean los más adecuados para las necesidades de su organización, es el momento de tomar la decisión de los tipos de modelado que se van a utilizar. La determinación del modelado más adecuado se basará en las siguientes consideraciones:

- **Los tipos de datos disponibles para la minería.** Por ejemplo, ¿los campos de interés son categóricos (simbólicos)?
- **Sus objetivos de minería de datos.** ¿Sólo quiere tener un mejor conocimiento de los almacenes de datos transaccionales y descubrir patrones de compras interesantes? ¿Necesita producir una puntuación indicando, por ejemplo, las posibilidades de impago de un préstamo a un estudiante?
- **Requisitos específicos de modelado.** ¿Necesita el modelo un tipo o un tamaño de datos concreto? ¿Necesita un modelo con unos resultados fácilmente presentables?

Si desea obtener más información sobre los tipos de modelado en IBM SPSS Modeler y sus requisitos, consulte la documentación de IBM SPSS Modeler o la ayuda en pantalla.

Ejemplo de venta en línea: técnicas de modelado

Las técnicas de modelado utilizadas por el comerciante se basan en los objetivos de minería de datos de la empresa:

Recomendaciones mejoradas. En su forma más simple, supone agrupar en clústeres los pedidos de compra para determinar los productos que se adquieren conjuntamente con más frecuencia. Se pueden añadir datos de clientes e incluso registros de visita, para obtener unos resultados más completos. Las técnicas de agrupación en clústeres en dos pasos o red de Kohonen son las más adecuadas para este tipo de modelado. Posteriormente, los clústeres se pueden perfilar utilizando un conjunto de reglas C5.0 para determinar las recomendaciones más adecuadas en cualquier punto durante una visita del cliente.

Navegación mejorada por el sitio. Por ahora, el comerciante se centrará en identificar las páginas que se utilizan con más frecuencia pero que requieren que el usuario realice varias operaciones para llegar a ellas. Implica aplicar un algoritmo de secuencia a los registros Web para generar las "rutas exclusivas" que los clientes utilizan en la Web y busca específicamente sesiones con multitud de visitas sin (o antes) de realizar una acción. Posteriormente, en un análisis con mayor profundidad, se pueden utilizar técnicas de clúster para identificar diferentes "tipos" de visitas y visitantes y el contenido del sitio se puede organizar y presentar según su tipo.

Selección de las técnicas de modelado correctas

Existen muchas técnicas de modelado disponibles en IBM SPSS Modeler. Con frecuencia, los analistas de datos utilizan más de un método para solucionar un problema desde diferentes puntos de vista.

Lista de tareas

Cuando decida el modelo(s) que va a utilizar, tenga en cuenta los siguientes aspectos que pueden afectar a sus opciones:

- ¿Requiere el modelo que los datos se dividan en conjuntos de entrenamiento y prueba?
- ¿Dispone de datos suficientes para producir resultados fiables para un modelo concreto?
- ¿Requiere el modelo un cierto nivel de calidad de datos? ¿Puede alcanzar este nivel con los datos que dispone?
- ¿Son sus datos el tipo correcto para un modelo concreto? En caso contrario, ¿puede realizar las conversiones necesarias utilizando nodos de manipulación de datos?

Si desea obtener más información sobre los tipos de modelado en IBM SPSS Modeler y sus requisitos, consulte la documentación de IBM SPSS Modeler o la ayuda en pantalla.

Modelado de supuestos

A medida que limite sus herramientas de modelado, registre el proceso de toma de decisiones. Documente cualquier supuesto de datos y modificación realizada para cumplir los requisitos del modelo.

Por ejemplo, tanto los nodos Regresión logística como Red neuronal requieren los tipos de datos que se **instancien** completamente (tipos de datos conocidos) antes de su ejecución. Significa que necesitará añadir un nodo Tipo a la ruta y ejecutarla para ejecutar los datos antes de crear y ejecutar un modelo. De igual forma, los modelos predictivos, como C5.0, pueden beneficiarse del reequilibrado de datos al predecir reglas de eventos raros. Cuando realice este tipo de predicciones, puede obtener mejores resultados insertando un nodo Equilibrar en la ruta e introduciendo el subconjunto más equilibrado en el modelo.

Asegúrese de documentar este tipo de decisiones.

Generación de un diseño de comprobación

Como paso final antes de generar el modelo, debe volver a tener en cuenta cómo se comprobarán los resultados del modelo. Existen dos partes para generar un diseño de comprobación global:

- Descripción de los criterios de "bondad" de un modelo
- Definición de los datos en los que se comprobarán estos criterios

La **bondad** de un modelo se puede medir de varias formas. Para modelos supervisados, como C5.0 y C&R Tree, las mediciones de bondad suelen calcular la tasa de error de un modelo concreto. Para modelos sin supervisión, como redes de clústeres de Kohonen, las mediciones pueden incluir criterios como facilidad de interpretación, despliegue o el tiempo de procesamiento necesario.

Recuerde, la generación de modelos es un proceso iterativo. Significa que normalmente comprobará los resultados de varios modelos antes de decidir los que se usarán y los que se desplegarán.

Escritura de un diseño de comprobación

El diseño de la prueba es una descripción de pasos que debe realizar para comprobar los modelos producidos. Como el modelado es un proceso iterativo, es importante saber cuándo debe dejar de ajustar parámetros e intentar otro método o modelo.

Lista de tareas

Cuando cree un diseño de comprobación, considere las siguientes cuestiones:

- ¿Qué datos se utilizarán para comprobar los modelos? ¿Ha particionado los datos en conjuntos de entrenamiento/prueba? (Es un método muy utilizado en el modelado.)
- ¿Cómo puede medir el rendimiento de modelos supervisados (como C5.0)?
- ¿Cómo puede medir el rendimiento de modelos sin supervisar (como redes de clústeres de Kohonen)?
- ¿Cuántas veces piensa volver a ejecutar un modelo con los valores ajustados antes de intentar otro tipo de modelo?

Ejemplo de venta en línea: diseño de comprobación

Un ejemplo de minería Web utilizando CRISP-DM

Los criterios por los que se evaluarán los modelos dependerán de los modelos que se considerarán y de los objetivos de minería de datos:

Recomendaciones mejoradas. Hasta que se presenten las recomendaciones mejoradas a los clientes activos, no existe un método puramente objetivo de evaluarlas. Sin embargo, el comerciante puede exigir que las reglas que generen las recomendaciones sean lo suficientemente simples para que tengan sentido desde una perspectiva comercial. Del mismo modo, las reglas deben ser lo suficientemente complejas para generar recomendaciones diferentes para clientes y sesiones diferentes.

Navegación mejorada por el sitio. Conociendo las páginas a las que acceden los clientes en el sitio Web, el comerciante puede evaluar de forma objetiva el diseño del sitio actualizado en términos de facilidad de uso a las páginas más importantes. Sin embargo, al igual que con las recomendaciones, es difícil evaluar de forma anticipada cómo se ajustarán los usuarios al sitio reorganizado. Si lo permiten los plazos y el presupuesto, se pueden realizar comprobaciones de la facilidad de uso.

Generación de los modelos

En este punto, debe tener la preparación suficiente para generar los modelos que haya considerado. Tómese el tiempo necesario para experimentar con diferentes modelos antes de llegar a conclusiones definitivas. La mayoría de analistas de datos suelen generar varios modelos y comparar los resultados antes de desplegarlos o integrarlos.

Para poder registrar su progreso con una amplia variedad de modelos, asegúrese de registrar los ajustes y datos utilizados para cada modelo. De esta forma podrá analizar los resultados con otras personas y comprobar sus pasos si fuera necesario. Al final del proceso de generación de modelos dispondrá de tres tipos de información que puede utilizar en la toma de decisiones de minería de datos:

- **Configuración de parámetros** incluye las notas que ha tomado sobre los parámetros que producen los mejores resultados.
- Los **modelos** reales producidos.
- **Descripciones de resultados de modelos**, incluyendo problemas de datos y rendimiento que hayan ocurrido durante la ejecución del modelo y exploración de los resultados.

Ejemplo de venta en línea: generación de modelos

Un ejemplo de minería Web utilizando CRISP-DM

Recomendaciones mejoradas. Las agrupaciones en clúster se producen por diferentes niveles de integración de datos, comenzando por la compra de la base de datos e incluyendo información relacionada con el cliente y la sesión. Para cada nivel de integración, los clústeres se producen en función de la diferente configuración de los parámetros para los algoritmos de dos pasos y la red de Kohonen. Para cada uno de estas agrupaciones en clúster, se generan algunos conjuntos de reglas de C5.0 con una configuración de parámetros diferente.

Navegación mejorada por el sitio. El nodo de modelado Secuencia se utiliza para generar rutas de clientes. El algoritmo permite especificar unos criterios de soporte mínimos, lo cual es de gran utilidad para centrarse en las rutas más comunes de los clientes. Se comprueban diferentes configuraciones de los parámetros.

Configuración de parámetros

La mayoría de técnicas de modelado tienen diferentes parámetros o configuraciones que se pueden ajustar para controlar el proceso de modelado. Por ejemplo, los árboles de decisión se pueden controlar ajustando la profundidad del árbol, divisiones y otros ajustes. Normalmente, la mayoría de usuarios genera un modelo utilizando las opciones predeterminadas y refina los parámetros en subsiguientes sesiones.

Una vez haya determinado los parámetros que producen los resultados más precisos, asegúrese de que guarda la ruta y los nodos de modelo generados. Además, si toma notas de los ajustes óptimos, le pueden servir de gran ayuda si decide automatizar o volver a generar el modelo con nuevos datos.

Ejecución de los modelos

En IBM SPSS Modeler, la ejecución de modelos es una tarea sencilla. Una vez haya introducido el nodo de modelo en la ruta y editado los parámetros, sólo tiene ejecutar el modelo para producir resultados visibles. Los resultados aparecen en el navegador Modelos generados en la parte derecha del espacio de trabajo. Puede pulsar con el botón derecho del ratón en un modelo y comprobar los resultados. En la mayoría de modelos, puede introducir el modelo generado en la ruta para evaluar posteriormente y desplegar los resultados. Los modelos se pueden guardar en IBM SPSS Modeler para poder volver a utilizarlos con facilidad.

Descripción de modelo

Cuando examine los resultados de un modelo, asegúrese de tomar notas del proceso de modelado. Puede guardar las notas con el propio modelo utilizando el cuadro de diálogo de anotaciones del nodo o la herramienta de proyectos.

Lista de tareas

En cada modelo, registre información como:

- ¿Puede llegar a conclusiones significativas a partir de este modelo?
- ¿Revela este modelo nuevas oportunidades o patrones alternativos?
- ¿El modelo presenta problemas de ejecución? ¿Fue razonable el tiempo de procesamiento?
- ¿El modelo presenta problemas de calidad de datos, como un alto número de valores perdidos?
- ¿Existen incoherencias de cálculos que se deben mencionar?

Evaluación del modelo

Ahora que ha definido un conjunto de modelos iniciales, obsérvelos detenidamente para determinar cuáles son los más precisos o eficaces para considerarse finales. Finales puede significar varias cosas, como "listo para desplegar" o "ilustra patrones interesantes". Si consulta el plan de pruebas que ha creado previamente, puede ayudarle a crear esta evaluación desde el punto de vista de su organización.

Evaluación global del modelo

Para cada modelo que se va a considerar, es una buena idea crear un método de evaluación basado en los criterios generados en su plan de pruebas. En este punto puede añadir el modelo generado a la ruta y utilizar diagramas de evaluación o nodos de análisis para analizar la efectividad de los resultados. También debe tener en cuenta si los resultados tienen un sentido lógico o si son demasiado simples para sus objetivos comerciales (por ejemplo, una secuencia que revele compras como vino > vino > vino).

Una vez haya realizado la evaluación, clasifique los modelos en función de criterios objetivos (precisión del modelo) y subjetivos (facilidad de uso o interpretación de los resultados).

Lista de tareas

- Utilice las herramientas de minería de datos de IBM SPSS Modeler, como diagramas de evaluación, nodos de análisis o gráficos de validación cruzada para evaluar los resultados de su modelo.
- Revise los resultados en función de su conocimiento del problema comercial. Consulte con analistas de datos u otros expertos que puedan dar una opinión relevante de resultados concretos.
- Considere si los resultados de un modelo son fácilmente desplegables. ¿Su organización requiere que los resultados se desplieguen en la Web o que se envíen al almacén de datos?
- Analice el impacto de los resultados según sus criterios de rendimiento. ¿Cumplen los objetivos establecidos durante la fase de comprensión comercial?

Si ha podido resolver los problemas satisfactoriamente y cree que los modelos actuales cumplen sus objetivos, es hora de pasar a realizar una evaluación más profunda de los modelos y el despliegue final. De otro modo, utilice los conocimientos adquiridos y vuelva a ejecutar los modelos con los parámetros de configuración ajustados.

Ejemplo de venta en línea: evaluación de modelos

Un ejemplo de minería Web utilizando CRISP-DM

Recomendaciones mejoradas. Una de las redes de Kohonen y una agrupación en clústeres en dos pasos ofrecen resultados razonables y el comerciante tiene dificultades a la hora de decidirse entre ambos. En ocasiones, la empresa espera poder utilizar ambos, aceptando las recomendaciones que sugieren ambas técnicas y estudiando en detalle sus diferencias. Con un poco esfuerzo y un mínimo conocimiento comercial, el comerciante puede desarrollar más reglas para reconciliar las diferencias entre ambas técnicas.

El comerciante también encuentra que los resultados que incluye la sesión de información son sorprendentemente útiles. Existen pruebas que sugieren que las recomendaciones se pueden utilizar para mejorar la navegación por el sitio. Se puede utilizar en tiempo real un conjunto de reglas que defina el próximo destino del cliente de forma que afecte al contenido del sitio a medida que el cliente navega.

Navegación mejorada por el sitio. El modelo de secuencia proporciona al comerciante un alto nivel de confianza en la predicción de ciertas rutas de clientes, lo que produce resultados que sugieren un número gestionable de cambios para el diseño del sitio.

Seguimiento de los parámetros revisados

En función de los conocimientos adquiridos durante la evaluación del modelo, es hora de volver a observar los modelos. Existen dos opciones:

- Ajuste de los parámetros de los modelos existentes.
- Selección de un modelo diferente para solucionar problemas de minería de datos.

En ambos casos, volverá a la tarea de generación de modelos y el proceso se repetirá hasta que los resultados sean satisfactorios. No se preocupe si tiene que repetir este paso. Es muy común que los analistas de datos tengan que evaluar y volver a ejecutar los modelos varias veces antes de encontrar el modelo que mejor se ajuste a sus necesidades. Es un excelente argumento para generar varios modelos de una vez y comparar los resultados antes de ajustar los parámetros de cada modelo.

¿Está listo para el siguiente paso?

Antes de pasar a la evaluación final de los modelos, considere si su evaluación inicial era lo suficientemente precisa.

Lista de tareas

- ¿Puede comprender los resultados de los modelos?
- ¿Los resultados del modelo tienen sentido desde una perspectiva meramente lógica? ¿Existen incoherencias aparentes que necesiten una mayor exploración?
- Desde el inicio, ¿los resultados parecen resolver los problemas de su organización?
- ¿Ha utilizado nodos de análisis y gráficos de ganancias o elevaciones para comparar y evaluar la precisión de los modelos?
- ¿Ha explorado más de un tipo de modelo y comparado los resultados?
- ¿Se pueden desplegar los resultados del modelo?

Si los resultados del modelado de datos parecen precisos y relevantes, es hora de realizar una evaluación más profunda antes del despliegue final.

Capítulo 6. Evaluación

Conceptos básicos sobre evaluación

En este punto, habrá completado la mayor parte de su proyecto de minería de datos. También habrá determinado, en la fase de modelado, que los modelos son técnicamente correctos y efectivos en función de los **criterios de rendimiento de minería de datos** que ha definido previamente.

Sin embargo, antes de continuar, debe evaluar los resultados de sus esfuerzos utilizando los **criterios de rendimiento comercial** establecidos en el inicio del proyecto. Es la clave para asegurar que su organización pueda utilizar los resultados que ha obtenido. La minería de datos produce dos tipos de resultados:

- Los **modelos** finales seleccionados en la fase anterior de CRISP-DM.
- Las conclusiones o interferencias obtenidas de los modelos y del proceso de minería de datos. Reciben el nombre de **descubrimientos**.

Evaluación de los resultados

En esta etapa, formalizará su evaluación en función de si los resultados del proyecto cumplen los criterios del rendimiento comercial. Este paso requiere una clara comprensión de los objetivos comerciales, por lo que debe estar seguro de incluir factores de toma de decisiones en la evaluación del proyecto.

Lista de tareas

En primer lugar, debe registrar su evaluación, indicando si los resultados de minería de datos cumplen sus criterios de rendimiento comercial. Considere las siguientes cuestiones en su informe:

- ¿Sus resultados se expresan con claridad y de forma que se puedan presentar con facilidad?
- ¿Ha realizado descubrimientos especiales o particularmente relevantes que deba resaltar?
- ¿Puede evaluar los modelos y descubrimientos en función de su capacidad de poderse aplicar a los objetivos comerciales?
- En general, ¿en qué medida estos resultados se adaptan a los objetivos comerciales de su organización?
- ¿Qué cuestiones adicionales generan los resultados? ¿Cómo puede formular estas cuestiones en términos comerciales?

Una vez haya evaluado los resultados, realice una lista de los modelos aprobados para incluirlos en el informe final. Esta lista debe incluir los modelos que cumplan los requisitos de minería de datos y los objetivos comerciales de su organización.

Ejemplo de venta en línea: evaluación de resultados

Un ejemplo de minería Web utilizando CRISP-DM

Los resultados globales de la primera experiencia del comerciante de minería de datos son muy fáciles de comunicar desde una perspectiva comercial: el estudio refleja recomendaciones de mejora de producto y un diseño mejorado del sitio. El diseño mejorado del sitio se basa en las secuencias de navegación del cliente, que muestran las características del sitio que los clientes desean pero que requieren varios pasos. La prueba de que las recomendaciones de producto son de mejora es más difícil de comunicar, porque las reglas de decisión se pueden complicar. Para producir el informe final, los analistas intentarán identificar algunas tendencias generales en los conjuntos de reglas que se pueden explicar con mayor facilidad.

Clasificación de los modelos. Como varios de los modelos iniciales parecían tener sentido comercial, la clasificación dentro de ese grupo se basa en criterios estadísticos, fáciles de interpretar y de gran diversidad. Además, el modelo ofreció diferentes recomendaciones para diferentes situaciones.

Nuevas cuestiones. La cuestión más importante que surge del estudio es, ¿cómo puede el comerciante tener un mayor conocimiento de sus clientes? La información en la base de datos de clientes desarrolla un importante rol en la formación de clústeres de recomendaciones. Mientras existen reglas especiales para realizar recomendaciones a los clientes cuya información falta, las recomendaciones son más generales en comparación con las recomendaciones a los clientes registrados.

Proceso de revisión

Las metodologías eficaces suelen incluir tiempo para reflexionar sobre los aciertos y errores del proceso que se acaba de completar. La minería de datos no es muy diferente. Una parte fundamental de CRISP-DM es aprender de su propia experiencia para que sus proyectos de minería de datos sean más efectivos.

Lista de tareas

En primer lugar, debe resumir las actividades y decisiones de cada fase, incluyendo pasos de preparación de datos, generación de modelos, etc. Además, en cada fase, debe considerar las cuestiones y realizar sugerencias para la mejora:

- ¿Esta fase ha contribuido al valor de sus resultados finales?
- ¿Existen formas de simplificar o mejorar esta fase u operación particular?
- ¿Cuáles fueron los fallos o errores cometidos en esa fase? ¿Cómo se pueden evitar la próxima vez?
- ¿Hay callejones sin salida, como modelos específicos que no ofrecen ningún resultado? ¿Existen formas de predecir esos callejones sin salida de forma que los esfuerzos se puedan dirigir con más productividad?
- ¿Se han producido sorpresas (buenas y malas) en esta fase? A posteriori, ¿existe alguna forma de predecir esas instancias?
- ¿Existen decisiones alternativas o estrategias que se puedan utilizar en una fase concreta? Registre esas alternativas para utilizarlas en proyectos de minería de datos futuros.

Ejemplo de venta en línea: informe de revisión

Un ejemplo de minería Web utilizando CRISP-DM

Como resultado de revisar el proceso del proyecto de minería de datos inicial, el comerciante ha desarrollado un mayor aprecio por las interrelaciones entre los pasos de un proceso. Inicialmente reacio a "revisar" el proceso CRISP-DM, el comerciante percibe que la naturaleza cíclica del proceso aumenta su potencialidad. La revisión del proceso también lleva al comerciante a comprender los siguientes conceptos:

- Cuando se produce un suceso inesperado en una fase diferente al CRISP-DM, se vuelve siempre al proceso de exploración.
- La preparación de los datos, especialmente de registros Web, requiere paciencia, ya que puede llevar mucho tiempo.
- Es esencial mantenerse centrado en el problema comercial, porque una vez que los datos están preparados para el análisis, es muy fácil iniciar la construcción de modelos sin depender de la imagen mayor.
- Una vez concluida la fase de modelado, la comprensión comercial es más importante a la hora de decidir la importancia de aplicar los resultados y determinar los sucesivos estudios que se garantizan.

Determinación de los pasos siguientes

Por ahora ha obtenido unos resultados, ha evaluado sus experiencia de minería de datos y se debe estar preguntando, **¿qué viene a continuación?** Esta fase le ayuda a responder esa pregunta en términos de objetivos comerciales de minería de datos. Básicamente, llegados a este punto dispone de dos opciones:

- **Continuar con la fase de despliegue.** La siguiente fase le ayudará a incorporar los resultados del modelo a su proceso comercial y producir un informe final. Incluso si sus esfuerzos invertidos en la minería de datos no han sido satisfactorios, debe utilizar la fase de despliegue de CRISP-DM para crear un informe final para su distribución al patrocinador del proyecto.
- **Volver y refinar o sustituir los modelos.** Si encuentra que los resultados son casi óptimos, pero no lo suficiente, considere otro tipo de modelado. Puede utilizar sus conocimientos adquiridos en esta fase para refinar los modelos y producir mejores resultados.

En este punto, su decisión incluye la precisión y relevancia de los resultados de modelado. Si los resultados se adaptan a sus objetivos comerciales de minería de datos, puede pasar a la fase de despliegue. Con independencia de la decisión que tome, asegúrese de registrar el proceso de evaluación.

Ejemplo de venta en línea: siguientes pasos

Un ejemplo de minería Web utilizando CRISP-DM

El comerciante confía en la precisión y relevancia de los resultados del proyecto y continúa con la fase de despliegue.

Al mismo tiempo, el equipo de proyecto también está listo para volver y aumentar algunos de los modelos que van a incluir técnicas predictivas. En este punto, esperan los resultados de los informes finales y luz verde de los responsables en la toma de decisiones.

Capítulo 7. Despliegue

Conceptos básicos sobre despliegue

El despliegue es el proceso que consiste en utilizar sus nuevos conocimientos para implementar las mejoras en su organización. Puede significar una integración formal como la aplicación del modelo de IBM SPSS Modeler que genera puntuaciones de abandono de clientes que se leen en un almacén de datos. Además, el despliegue puede significar que utilice los conocimientos adquiridos en minería de datos para aplicar modificaciones en su organización. Por ejemplo, es posible que descubra patrones de alarma en sus datos que indican un cambio en el comportamiento de los clientes de más de 30 años. Es posible que estos resultados no se integren formalmente en sus sistemas de información, pero serán de gran utilidad para la planificación y toma de decisiones de marketing.

En general, la fase de despliegue de CRISP-DM incluye dos tipos de actividades:

- Planificación y control del despliegue de los resultados
- Finalización de tareas de presentación como la producción de un informe final y la revisión de un proyecto

Dependiendo de las necesidades de su organización, es posible que necesite completar una o varias fases.

Planificación de despliegue

Aunque pueda estar ansioso por compartir el fruto de sus esfuerzos en minería de datos, dedique un tiempo a planificar un despliegue completo y preciso de los resultados.

Lista de tareas

- El primer paso es resumir los resultados; modelos y descubrimientos. Este método le ayudará a determinar los modelos que se pueden integrar en sus sistemas de base de datos y los descubrimientos que se presentarán a sus colegas.
- En cada modelo desplegable, cree una planificación paso a paso para el despliegue e integración con sus sistemas. Registre los detalles técnicos como requisitos de base de datos para los resultados del modelo. Por ejemplo, es posible que su sistema requiera que los resultados del modelado se desplieguen en formato delimitado por tabulaciones.
- Para cada descubrimiento, cree un plan para difundir la información a los estrategas de la organización.
- ¿Dispone de planes de despliegue alternativos para ambos tipos de resultados que se deben documentar?
- Considere cómo controlará el despliegue. Por ejemplo, ¿cómo se actualizará un modelo desplegado utilizando IBM SPSS Modeler Solution Publisher? ¿Cómo decidirá el momento en que el modelo ya no es aplicable?
- Identifique los problemas de despliegue y realice un plan de contingencia. Por ejemplo, es posible que los gerentes de la organización quieran más información sobre los resultados del modelado y que proporcione más detalles técnicos.

Ejemplo de venta en línea: planificación del despliegue

Un ejemplo de minería Web utilizando CRISP-DM

Un despliegue satisfactorio de los resultados de minería de datos del comerciante requiere que la información correcta llegue a las personas adecuadas.

Gerentes. Los gerentes necesitan recibir informaciones sobre las recomendaciones y modificaciones propuestas en el sitio y breves explicaciones acerca de cómo afectarán estos cambios. Asumiendo que aceptarán los resultados del estudio, es necesario notificar a las personas que implementarán esas modificaciones.

Desarrolladores Web. Los responsables del mantenimiento del sitio Web deberán incorporar las nuevas recomendaciones y el contenido del sitio de la organización. Debe informar de los cambios que *pueden* ocurrir por causas de futuros estudios, de forma que puedan comenzar los trabajos en breve. Mantener al equipo preparado para que puedan realizar las construcciones del sitio de forma inmediata en función de los análisis de secuencias en tiempo real puede ser de gran ayuda posteriormente.

Expertos de bases de datos. Las personas que mantienen las bases de datos de clientes, compras y productos deben saber cómo se utiliza la información de las bases de datos y conocer qué atributos se pueden añadir a las bases de datos en proyectos futuros.

Sobre todo, el equipo de proyecto necesita obtener la información de todos estos grupos para coordinar el despliegue de los resultados y planificación de proyectos futuros.

Planificación del control y del mantenimiento

En un despliegue e integración completos de los resultados de modelado, su trabajo de minería de datos puede ser continuado. Por ejemplo, si un modelo se despliega para predecir las consecuencias de las compras en línea, es probable que este modelo se tenga que evaluar periódicamente para asegurar su eficacia y realizar mejoras continuas. Del mismo modo, un modelo desplegado para aumentar la retención de los clientes más importantes se deberá modificar una vez se ha alcanzado un nivel concreto de retención. El modelo se puede modificar y reutilizar para retener clientes de un nivel inferior, pero que siguen teniendo un nivel de rentabilidad en la pirámide de valores.

Lista de tareas

Registre los siguientes elementos y asegúrese de que los incluye en el informe final.

- En cada modelo o descubrimiento, ¿qué factores o influencias (como valor de mercado o variaciones estacionales) necesita controlar?
- ¿Cómo se puede medir y controlar la validez y precisión de cada modelo?
- ¿Cómo se determina que un modelo ha "expirado"? Proporcione detalles sobre umbrales de precisión o modificaciones esperadas en los datos, etc.
- ¿Qué ocurre cuando un modelo expira? ¿Puede reconstruir el modelo con nuevos datos o tiene que realizar algunas modificaciones? ¿O por contra, las modificaciones son tantas que se requiere un nuevo proyecto de minería de datos?
- ¿Se puede utilizar este modelo para problemas comerciales similares una vez expirado? En este punto se hace indispensable disponer de documentación suficiente para valorar el propósito comercial de cada proyecto de minería de datos.

Ejemplo de venta en línea: control y mantenimiento

Un ejemplo de minería Web utilizando CRISP-DM

La tarea inmediata del control es determinar si la nueva organización del sitio y las recomendaciones mejoradas funcionan correctamente. Es decir, ¿los usuarios pueden tomar rutas directas a las páginas que buscan? ¿Han aumentado las ventas cruzadas de los artículos recomendados? Transcurridas algunas semanas de control, el comerciante podrá determinar el rendimiento del estudio.

Lo que sí se puede gestionar automáticamente es la inclusión de los nuevos usuarios registrados. Cuando los clientes se registran en el sitio, los conjuntos de reglas actuales se pueden aplicar a su información para determinar las recomendaciones que se pueden dar.

Decidir cuándo debe actualizar los conjuntos de reglas para determinar las recomendaciones es una tarea más difícil. La actualización de los conjuntos de reglas no es un proceso automático porque la creación de clústeres requiere la intervención humana para definir la idoneidad de una solución de clúster concreta.

Como los proyectos futuros generarán modelos de mayor complejidad, también aumentará la necesidad de control. Cuando sea posible, la tarea de control debe ser automática con informes programados que se puedan revisar, de forma regular. Además, la creación de modelos que proporcionan predicciones de forma instantánea puede ser una dirección que la empresa desee tomar. Requiere un equipo más sofisticado que el primer proyecto de minería de datos.

Creación de un informe final

La escritura de un informe final no sólo resuelve los cabos sueltos de la documentación previa, sino que también se puede utilizar para comunicar los resultados. Aunque pueda parecer una tarea sencilla, es importante presentar los resultados a las diferentes personas relacionadas con los resultados. Se pueden incluir a los administradores técnicos, que son responsables de la aplicación de los resultados de modelado, así como el departamento de marketing y gestión, encargado de tomar las decisiones en función de los resultados obtenidos.

Lista de tareas

En primer lugar, tenga en cuenta los receptores del informe. ¿Son desarrolladores técnicos o gestores de ventas? Es posible que necesite crear informes independientes para cada tipo de destinatario, si sus necesidades son diferentes. En cada caso, el informe debe incluir los siguientes elementos:

- Una descripción detallada del problema original
- El procedimiento utilizado para realizar el proyecto de minería de datos
- El coste del proyecto
- Comentarios sobre las desviaciones del plan del proyecto original
- Un resumen de los resultados de minería de datos, incluyendo los modelos y los descubrimientos
- Un resumen del plan propuesto para el despliegue
- Recomendaciones para futuros proyectos de minería de datos, incluyendo reglas interesantes descubiertas durante los procesos de exploración y modelado

Preparación de una presentación final

Además del informe de proyecto, es posible que también necesite presentar los descubrimientos a un equipo de patrocinadores o departamentos relacionados. Si es el caso, puede utilizar parte de la información del informe, pero presentarlo desde una perspectiva más amplia. Los gráficos de IBM SPSS Modeler se pueden exportar fácilmente para este tipo de presentaciones.

Ejemplo de venta en línea: informe final

Un ejemplo de minería Web utilizando CRISP-DM

La principal desviación del plan del proyecto original también es una regla interesante para proyectos de minería de datos posteriores. El plan original se utilizó para detectar la forma en la que los clientes permanecen más tiempo y acceden a más páginas en cada visita.

Como se ha demostrado, un cliente feliz no es una cuestión de mantenerlo más tiempo en línea. Las distribuciones de frecuencia del tiempo empleado en cada sesión de las sesiones que acabaron en compra, descubrieron que los tiempos de sesión de la mayoría de sesiones que han terminado en compra es inferior al tiempo de sesión de dos clústeres de sesiones que no han terminado en compra.

Una vez que se conoce este dato, el problema es descubrir si estos clientes que pasan gran cantidad de tiempo en el sitio sin realizar compras se limitan a navegar o no terminan de encontrar lo que están buscando. El siguiente paso es determinar cómo proporcionarles lo que buscan con objeto de potenciar las compras.

Revisión final del proyecto

Es el paso final del método CRISP-DM y le ofrece una oportunidad de formular sus impresiones finales e incorporar los conocimientos adquiridos durante el proceso de minería de datos.

Lista de tareas

Debe realizar una breve entrevista con las personas implicadas en el proceso de minería de datos. Entre las cuestiones que debe tener en cuenta durante las entrevistas que realice se incluyen:

- ¿Cuál es su impresión global del proyecto?
- ¿Qué conocimientos ha adquirido durante el proceso de minería de datos en general y los datos disponibles?
- ¿Qué partes del proyecto han funcionado correctamente? ¿Dónde han surgido las dificultades? ¿Existe algún tipo de información que le podría haber evitado confusiones?

Tras el despliegue de los resultados de minería de datos, también debería entrevistarse con las personas afectadas por los resultados, como clientes o socios comerciales. Su objetivo es determinar si el proyecto ha sido fructífero y si ha obtenido los resultados esperados.

Los resultados de estas entrevistas se pueden resumir junto con sus propias impresiones en un informe final que debe contener los conocimientos adquiridos de la experiencia de minería de sus almacenes de datos.

Ejemplo de venta en línea: revisión final

Un ejemplo de minería Web utilizando CRISP-DM

Entrevistas a miembros del proyecto. El comerciante encuentra que los miembros del proyecto más íntimamente relacionados con el estudio de inicio a fin son en su mayoría entusiastas con los resultados y esperan poder implicarse en proyectos futuros. El grupo de la base de datos tiene un optimismo contenido y aunque que aprecian la utilidad del estudio, señalan la carga añadida que suponen los recursos de la base de datos. Durante el estudio, se dispuso de la ayuda de un asesor, pero en el futuro se necesitará un empleado dedicado al mantenimiento de la base de datos a medida que se amplíe el proyecto.

Entrevistas de clientes. Los comentarios de los clientes han sido muy positivos hasta el momento. Uno de los temas que no tuvo una buena previsión fue el impacto de las modificaciones en el diseño del sitio en los clientes habituales. Tras algunos años, los clientes registrados desarrollan determinados hábitos con respecto a la organización del sitio. Los comentarios de los usuarios registrados no son tan positivos como los de los usuarios no registrados y a algunos de ellos no les gustan los cambios en absoluto. El comerciante necesita conocer este problema y considerar si un cambio atraerá a la cantidad de nuevos clientes suficiente como para arriesgarse a perder algunos de los clientes actuales.

Avisos

Esta información se ha desarrollado para los productos y servicios ofrecidos en todo el mundo.

Es posible que IBM no ofrezca los productos, servicios o características que se tratan en este documento en otros países. Póngase en contacto con el representante local de IBM para obtener información sobre los productos y los servicios actualmente disponibles en su zona. Las referencias a productos, programas o servicios de IBM no pretenden establecer ni implicar que sólo puedan utilizarse productos, programas o servicios de IBM. En su lugar se puede utilizar cualquier producto, programa o servicio funcionalmente equivalente que no infrinja ningún derecho de propiedad intelectual de IBM. Sin embargo, es responsabilidad del usuario evaluar y comprobar el funcionamiento de todo producto, programa o servicio que no sea de IBM.

IBM puede tener patentes o solicitudes de patente en tramitación que cubran la materia descrita en este documento. Este documento no le otorga ninguna licencia para estas patentes. Puede enviar preguntas acerca de las licencias, por escrito, a:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
Estados Unidos

Para consultas sobre licencias respecto a la información de doble byte (DBCS), póngase en contacto con el Departamento de propiedad intelectual de IBM de su país o envíe sus consultas, por escrito, a:

Intellectual Property Licensing
Ley de propiedad intelectual y legal
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

El párrafo siguiente no se aplica al Reino Unido ni a ningún otro país donde estas disposiciones sean incompatibles con la legislación local: INTERNATIONAL BUSINESS MACHINES CORPORATION PROPORCIONA ESTA PUBLICACIÓN "TAL CUAL" SIN GARANTÍA DE NINGÚN TIPO, NI EXPLÍCITAS NI IMPLÍCITAS, INCLUYENDO, PERO SIN LIMITARSE A, LAS GARANTÍAS IMPLÍCITAS DE NO VULNERACIÓN, COMERCIALIZACIÓN O IDONEIDAD PARA UN FIN DETERMINADO. Algunos estados no permiten la renuncia a expresar o a garantías implícitas en determinadas transacciones , por lo tanto , esta declaración no se aplique a usted.

Esta información puede incluir imprecisiones técnicas o errores tipográficos. Periódicamente, se efectúan cambios en la información aquí y estos cambios se incorporarán en nuevas ediciones de la publicación. IBM puede realizar en cualquier momento mejoras o cambios en los productos o programas descritos en esta publicación sin previo aviso.

Cualquier referencia a sitios Web que no sean de IBM en esta información solamente es ofrecida por comodidad y de ningún modo sirve como aprobación de esos sitios Web. Los materiales que se encuentran en los mencionados sitios web no forman parte de los materiales para este producto de IBM y el usuario los utiliza por su cuenta y riesgo.

IBM puede utilizar o distribuir la información que el usuario le suministre en el modo que considere apropiado sin incurrir en ninguna obligación con el usuario.

Los titulares de licencias de este programa que deseen tener información sobre el mismo con el fin de permitir: (i) el intercambio de información entre programas creados independientemente y otros programas (incluido este) y (ii) el uso mutuo de la información que se ha intercambiado, deberán ponerse en contacto con:

Tel. 901 100 400
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
Estados Unidos

Esta información estará disponible, bajo las condiciones adecuadas, incluyendo en algunos casos el pago de una cuota.

El programa bajo licencia que se describe en este documento y todo el material bajo licencia disponible los proporciona IBM bajo los términos de las Condiciones Generales de IBM, Acuerdo Internacional de Programas Bajo Licencia de IBM o cualquier acuerdo equivalente entre las partes.

Cualquier dato de rendimiento mencionado aquí ha sido determinado en un entorno controlado. Por lo tanto, los resultados obtenidos en otros entornos operativos pueden variar de forma significativa. Es posible que algunas mediciones se hayan realizado en sistemas en desarrollo y no existe ninguna garantía de que estas medidas sean las mismas en los sistemas comerciales. Además, es posible que algunas mediciones hayan sido estimadas a través de extrapolación. Los resultados reales pueden variar. Los usuarios de este documento deben consultar los datos que corresponden a su entorno específico.

Se ha obtenido información acerca de productos que no son de IBM de los proveedores de esos productos, de sus publicaciones anunciadas o de otros orígenes disponibles públicamente. IBM no ha probado estos productos y no puede confirmar la precisión del rendimiento, la compatibilidad o cualquier otra reclamación relacionada con productos que no son de IBM. Las preguntas acerca de las aptitudes de productos que no sean de IBM deben dirigirse a los proveedores de dichos productos.

Todas las declaraciones sobre el futuro del rumbo y la intención de IBM están sujetas a cambio o retirada sin previo aviso y representan únicamente metas y objetivos.

Esta información contiene ejemplos de datos e informes utilizados en operaciones comerciales diarias. Para ilustrarlos lo máximo posible, los ejemplos incluyen los nombres de las personas, empresas, marcas y productos. Todos esos nombres son ficticios y cualquier parecido con los nombres y direcciones utilizados por una empresa real es pura coincidencia.

Si está viendo esta información en copia electrónica, es posible que las fotografías y las ilustraciones en color no aparezcan.

Marcas comerciales

IBM, el logotipo de IBM e ibm.com son marcas registradas o marcas comerciales de International Business Machines Corp., registradas en muchas jurisdicciones en todo el mundo. Puede que otros productos o nombres de servicio sean marcas registradas de IBM u otras compañías. Hay disponible una lista actual de marcas registradas de IBM en la Web en "Copyright and trademark information" (www.ibm.com/legal/copytrade.shtml).

Intel, el logotipo de Intel, Intel Inside, el logotipo de Intel Inside, Intel Centrino, el logotipo de Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium y Pentium son marcas comerciales o marcas registradas de Intel Corporation o sus filiales en Estados Unidos y otros países.

Linux es una marca registrada de Linus Torvalds en Estados Unidos, otros países o ambos.

Microsoft, Windows, Windows NT, y el logotipo de Windows son marcas comerciales de Microsoft Corporation en Estados Unidos, otros países o ambos.

UNIX es una marca registrada de The Open Group en Estados Unidos y otros países.

Java y todas las marcas comerciales y logotipos basados en Java son marcas comerciales o marcas registradas de Oracle y/o sus asociados.

Otros productos y nombres de servicio pueden ser marcas comerciales de IBM u otras empresas.

Índice

N

O

P

R

S

T

V

Impreso en España