Proceso de Minería de Datos

Data Scientist
Jhoan Esteba Ruiz Borja MSc

Contenido

1 CRISP-DM 2
1.1 Comprensión del negocio (Business Understanding) 4
1.2 La comprensión de datos (Data Understanding) 4
1.3 Preparación de datos (Data Preparation) 5
1.4 Modelado (Modeling) 8
1.5 Tratamiento de datos (Data Treatment) 8
1.6 Técnicas de minería de datos (Data Mining Techniques) 8
1.7 Evaluación (Evaluation) 10
1.8 Despliegue (Deployment) 11
2 SEMMA 11
2.1 Pasos en el proceso SEMMA 12
2.1.1 Paso 1 Muestra (Sample) 12
2.1.2 Paso 2 Explorar (Explore) 13
2.1.3 Paso 3 Modificar (Modify) 13
2.1.4 Paso 4 Modelo (Model) 13
2.1.5 Paso 5 Evaluar (Assess) 13
2.1.6 Ejemplo de aplicación de proceso de minería de datos 14
3 Comparación de CRISP y SEMMA 19
3.1 Manipulación de datos (Handling Data) 20
3.2 Etapa 1. La comprensión de negocios (Business Understanding) 20
3.3 Etapa 2. La comprensión de Datos (Data Understanding) 20
3.4 Etapa 3. Preparación de los datos (Data Preparation) 21
3.5 Etapa 4. Modelado (Modeling) 21
3.6 Etapa 5. Evaluación (Evaluation) 22
3.7 Etapa 6. Despliegue (Deployment) 25
4 Resumen 26

Con el fin de llevar a cabo sistemáticamente el análisis de minería de datos, generalmente se sigue un proceso general. Existen algunos procesos estándar, dos de los cuales se describen en a continuación. Uno (CRISP) es un proceso estándar de la industria que consiste en una secuencia de pasos que suelen estar involucrados en un estudio de minería de datos. El otro (SEMMA) es específico de SAS. Aunque cada paso de cada enfoque no es necesario en cada análisis, este proceso proporciona una buena cobertura de los pasos necesarios, comenzando con la exploración de datos, la recopilación de datos, el procesamiento de datos, el análisis, las inferencias dibujadas y la implementación.

1 CRISP-DM

Existe un Proceso Estándar Interprofesional para la Minería de Datos (CRISP-DM) ampliamente utilizado por los miembros de la industria. Este modelo consta de seis fases destinadas a un proceso cíclico (ver Fig. 2.1):

- **Entendimiento del negocio:** El entendimiento del negocio incluye determinar los objetivos del negocio, evaluar la situación actual, establecer objetivos de minería de datos y desarrollar un plan de proyecto.
- **Entendimiento de los datos:** Una vez establecidos los objetivos empresariales y el plan del proyecto, la comprensión de los datos considera los requisitos de los datos. Este paso puede incluir la recolección inicial de datos, la descripción de los datos, la exploración de datos y la verificación de la calidad de los datos. La exploración de datos como la visualización de estadísticas de resumen (que incluye la visualización de variables categóricas) puede ocurrir al final de esta fase. Modelos como el análisis de conglomerados (Clustering) también se pueden aplicar durante esta fase, con la intención de identificar patrones en los datos.
- **Preparación de los datos:** Una vez identificados los recursos de datos, deben ser seleccionados, limpiados, incorporados en la forma deseada y formateados. La limpieza de los datos y la transformación de datos en la preparación del modelado de datos deben ocurrir en esta fase. La exploración de datos a mayor profundidad se puede aplicar durante esta fase, y los modelos adicionales utilizados, una vez más la oportunidad de ver los patrones basados en la comprensión del negocio.
- **Modelado:** Las herramientas de software de minería de datos como la visualización (trazar datos y establecer relaciones) y el análisis de conglomerados (Clustering) (para identificar qué variables van bien juntas) son útiles para el análisis inicial. Herramientas como la inducción de regla generalizada pueden desarrollar reglas de asociación iniciales. Una vez que se obtiene mayor comprensión de los datos (a menudo mediante el reconocimiento de patrones provocado por la visualización del modelo de salida), se pueden aplicar modelos más detallados apropiados al tipo de datos. La división de datos en conjuntos de entrenamiento y de prueba también es necesaria para modelar.
- **Evaluación:** Los resultados del modelo deben ser evaluados en el contexto de los objetivos de negocio establecidos en la primera fase (entendimiento del negocio).

Esto conducirá a la identificación de otras necesidades (a menudo a través del reconocimiento de patrones), con frecuencia volviendo a las fases anteriores de CRISP-DM. La comprensión del negocio es un procedimiento iterativo en la minería de datos, donde los resultados de varias herramientas de visualización, estadísticas y de inteligencia artificial muestran al usuario nuevas relaciones que proporcionan una comprensión más profunda de las operaciones de la organización.

- **Implementación:** La minería de datos se puede utilizar tanto para verificar las hipótesis mantenidas anteriormente, como para el descubrimiento del conocimiento (identificación de relaciones inesperadas y útiles). A través del conocimiento descubierto en las fases anteriores del proceso CRISP-DM, se pueden obtener modelos de sonido que luego se pueden aplicar a las operaciones comerciales para muchos propósitos, incluyendo predicción o identificación de situaciones clave. Estos modelos necesitan ser monitoreados por cambios en las condiciones de operación, porque lo que podría ser verdad hoy en día puede no ser cierto en un año a partir de ahora. Si ocurren cambios significativos, el modelo debe ser rehecho. También es aconsejable registrar los resultados de los proyectos de minería de datos, de modo que se disponga de pruebas documentadas para futuros estudios.

Este proceso de seis fases no es un procedimiento rígido, por-los-números. Por lo general hay una gran cantidad de retrocesos. Además, los analistas experimentados pueden no necesitar aplicar cada fase para cada estudio. Pero CRISP-DM proporciona un marco útil para la minería de datos.

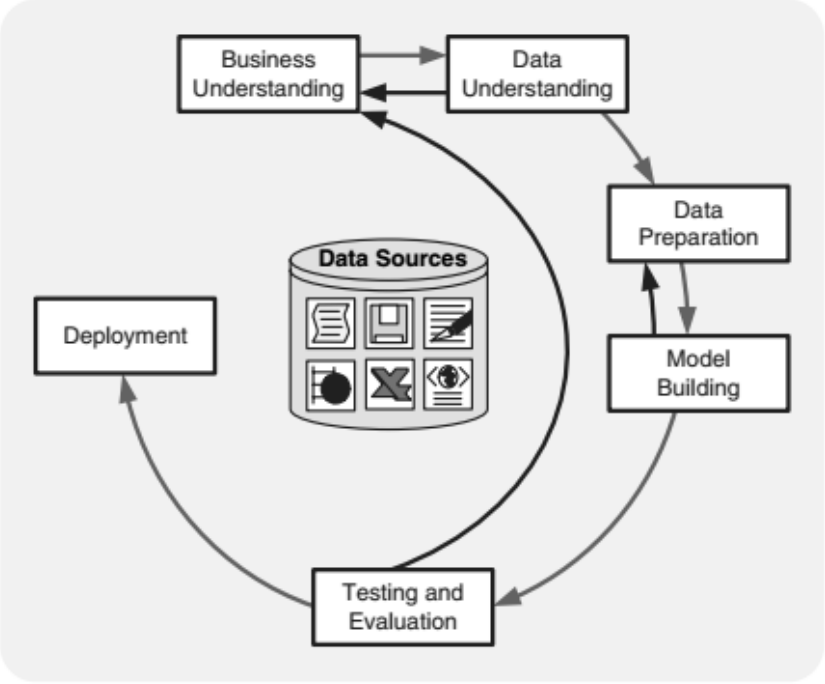

Fig. 2.1. CRISP-DM process

1.1 Comprensión del negocio (Business Understanding)

El elemento clave de un estudio de minería de datos es saber para qué sirve el estudio. Esto comienza con una necesidad gerencial de nuevos conocimientos, y una expresión del objetivo del negocio con respecto al estudio a emprenderse. Los objetivos en términos de cosas como "¿Qué tipos de clientes están interesados en cada uno de nuestros productos?" O "¿Cuáles son los perfiles típicos de nuestros clientes, y cuánto valor cada uno de ellos nos proporciona?". Luego, se debe desarrollar un plan para encontrar tal conocimiento, en términos de los responsables de recolectar datos, analizar los datos e informar. En esta etapa, debe establecerse un presupuesto para apoyar el estudio, al menos en términos preliminares.

En los modelos de segmentación de clientes, tales como la venta por catálogo de venta de Fingerhut, la identificación de un objetivo comercial significaba que identifica el tipo de cliente que se espera que produzca un retorno rentable. El mismo análisis es útil para los distribuidores de tarjetas de crédito. Para fines comerciales, las tiendas de comestibles a menudo tratan de identificar qué artículos tienden a ser comprados juntos por lo que puede ser utilizado para posicionamiento de afinidad dentro de la tienda, o para orientar inteligentemente campañas promocionales. La minería de datos tiene muchas aplicaciones de negocios útiles.

1.2 La comprensión de datos (Data Understanding)

Dado que la minería de datos está orientada a tareas, diferentes tareas empresariales requieren diferentes conjuntos de datos. La primera etapa del proceso de minería de datos es seleccionar los datos relacionados de muchas bases de datos disponibles para describir correctamente una tarea empresarial determinada. Hay al menos tres cuestiones a considerar en la selección de datos. La primera cuestión es establecer una descripción concisa y clara del problema. Por ejemplo, un proyecto minorista de minería de datos puede buscar identificar comportamientos de gasto de compradoras que compran ropa de temporada. Otro ejemplo puede tratar de identificar los patrones de bancarrota de los titulares de tarjetas de crédito. La segunda cuestión sería identificar los datos relevantes para la descripción del problema. La mayoría de los datos demográficos, transaccionales y financieros de las tarjetas de crédito podrían ser relevantes tanto para los proyectos de bancarrota de tarjetas de crédito como para los de venta al por menor. Sin embargo, los datos de género pueden estar prohibidos por ley para estos últimos, pero deben ser legales y resultar importantes para los primeros. La tercera cuestión es que las variables seleccionadas para los datos pertinentes deben ser independientes entre sí. La independencia variable significa que las variables no contienen información superpuesta. Una cuidadosa selección de variables independientes puede facilitar que los algoritmos de minería de datos descubran rápidamente patrones de conocimiento útiles.

Las fuentes de datos para la selección de datos pueden variar. Normalmente, los tipos de fuentes de datos para aplicaciones empresariales incluyen datos demográficos (tales como ingresos, educación, número de hogares y edad), datos socio gráficos (tales como afición,

membresía de clubes y entretenimiento), datos transaccionales (registros de ventas, créditos gasto de tarjetas, cheques emitidos), y así sucesivamente. El tipo de datos puede clasificarse como datos cuantitativos y cualitativos. Los datos cuantitativos se pueden medir usando valores numéricos. Puede ser discreto (como números enteros) o continuo (como números reales). Los datos cualitativos, también conocidos como datos categóricos, contienen datos nominales y ordinales. Los datos nominales tienen valores finitos no ordenados, como los datos de género que tienen dos valores: masculino y femenino. Los datos ordinales tienen valores finitos ordenados. Por ejemplo, las calificaciones crediticias de los clientes se consideran datos ordinales, ya que las calificaciones pueden ser excelentes, justas y malas. Los datos cuantitativos pueden ser fácilmente representados por algún tipo de distribución de probabilidad. Una distribución de probabilidad describe cómo se dispersan y forman los datos. Por ejemplo, los datos normalmente distribuidos son simétricos, y comúnmente se les llama en forma de campana. Los datos cualitativos pueden codificarse primero en números y luego describirse mediante distribuciones de frecuencia. Una vez que los datos relevantes se seleccionan de acuerdo con el objetivo del negocio de minería de datos, el pre-procesamiento de datos debe perseguirse.

1.3 Preparación de datos (Data Preparation)

El propósito del pre-procesamiento de datos es limpiar los datos seleccionados para una mejor calidad. Algunos datos seleccionados pueden tener diferentes formatos porque se eligen de diferentes fuentes de datos. Si los datos seleccionados son de archivos planos, mensajes de voz y texto web, deben convertirse en un formato electrónico coherente. En general, la limpieza de datos significa filtrar, agregar y rellenar los valores faltantes (imputación). Mediante el filtrado de datos, los datos seleccionados se examinan para los valores atípicos y redundancias. Los valores atípicos difieren mucho de la mayoría de los datos o datos que están claramente fuera del rango de los grupos de datos seleccionados. Por ejemplo, si el ingreso de un cliente incluido en la clase media es de $ 250,000, es un error y debe ser extraído del proyecto de minería de datos que examina los diversos aspectos de la clase media. Los valores atípicos pueden ser causados por muchas razones, como errores humanos o errores técnicos, o pueden ocurrir naturalmente en un conjunto de datos debido a eventos extremos. Supongamos que la edad de un titular de tarjeta de crédito se registra como "12." Esto es probablemente un error humano. Sin embargo, en realidad podría haber un preadolescente rico independiente con hábitos de compra importantes. Eliminar arbitrariamente este valor atípico podría descartar información valiosa.

Los datos redundantes son la misma información registrada de varias maneras diferentes. Las ventas diarias de un producto en particular son redundantes para las ventas estacionales del mismo producto, ya que podemos derivar las ventas de datos diarios o de datos estacionales. Mediante la agregación de datos, se reducen las dimensiones de los datos para obtener información agregada. Tenga en cuenta que aunque un conjunto de datos agregados tiene un volumen pequeño, la información permanecerá. Si se considera una promoción de marketing para ventas de muebles en los próximos 3 o 4 años, los datos de ventas diarias disponibles se pueden agregar como datos de ventas anuales. El tamaño de los datos de ventas se reduce

drásticamente. Al suavizar los datos, se encuentran valores faltantes de los datos seleccionados y se agregan valores nuevos o razonables. Estos valores añadidos podrían ser el número medio de la variable (media) o la moda. Un valor que falta a menudo no causa ninguna solución cuando se aplica un algoritmo de minería de datos para descubrir los patrones de conocimiento.

Los datos pueden expresarse en una serie de formas diferentes. Por ejemplo, en CLEMENTINE, se pueden utilizar los siguientes tipos de datos.

- RANGE Valores numéricos (entero, real o fecha / hora).
- FLAG Binario - Sí / No, 0/1, u otros datos con dos resultados (texto, Número entero, número real o fecha / hora).
- SET Datos con distintos valores múltiples (numérico, cadena o fecha / hora).
- TYPELESS Para otros tipos de datos.

Generalmente pensamos en los datos como números reales, como la edad en años o el ingreso anual en dólares (usaríamos RANGE en esos casos). A veces las variables ocurren como uno o de los tipos, como tener una licencia de conducir o no, o una reclamación de seguro es fraudulenta o no. Este caso podría tratarse utilizando valores numéricos reales (por ejemplo, 0 o 1). Pero es más eficiente tratarlos como variables FLAG. A menudo, es más apropiado tratar con datos categóricos, como la edad en términos del conjunto (jóvenes, de mediana edad, ancianos) o ingresos en el conjunto (bajo, medio, alto). En ese caso, podríamos agrupar los datos y asignar la categoría apropiada en términos de una cadena, usando un conjunto. La forma más completa es RANGE, pero a veces los datos no vienen en esa forma por lo que los analistas se ven obligados a utilizar los tipos SET o FLAG. A veces puede ser más preciso tratar los tipos de datos SET que los tipos de datos RANGE.

Como otro ejemplo, PolyAnalyst tiene disponibles los siguientes tipos de datos:

- Numérico (Numerical,Valores continuos)
- Entero (Integer, Valores enteros)
- SI/NO (Yes/no, Datos binarios)
- Categóricos (Category, Un conjunto finito de posibles valores)
- Fecha (Date)
- Cadena de texto (String)
- Texto (Texto)

Cada herramienta de software tendrá un esquema de datos diferente, pero los tipos primarios de datos tratados se representan en estas dos listas.

Hay muchos métodos estadísticos y herramientas de visualización que se pueden utilizar para pre-procesar los datos seleccionados. Las estadísticas comunes, como el máximo, el mínimo, la media y la moda, pueden usarse fácilmente para agregar o suavizar los datos, mientras que las gráficas de dispersión y las gráficas de caja se usan generalmente para filtrar valores atípicos. Las técnicas más avanzadas (incluyendo análisis de regresión, análisis de clúster, árbol de decisión o análisis jerárquico) se pueden aplicar en el pre-procesamiento de datos,

dependiendo de los requisitos para la calidad de los datos seleccionados. Debido a que el pre-procesamiento de datos es detallado y tedioso, requiere una gran cantidad de tiempo. En algunos casos, el pre-procesamiento de datos podría tomar más del 50% del tiempo de todo el proceso de minería de datos. Acortar el tiempo de procesamiento de datos puede reducir la mayor parte del tiempo total de computación en la minería de datos. El formato de datos simple y estándar resultante del pre-procesamiento de datos puede proporcionar un entorno de intercambio de información entre diferentes sistemas informáticos, lo que crea la flexibilidad para implementar diversos algoritmos o herramientas de minería de datos.

Como un componente importante de la preparación de datos, la transformación de datos consiste en utilizar formulaciones matemáticas simples o curvas de aprendizaje para convertir diferentes medidas de datos seleccionados y limpios en una escala numérica unificada con el propósito de analizar los datos. Muchas mediciones disponibles de las estadísticas, tales como media, mediana, moda y varianza se pueden utilizar fácilmente para transformar los datos. En términos de la representación de datos, la transformación de datos puede ser utilizada para (1) transformar de escalas numéricas a numéricas, y (2) recodificar datos categóricos a escalas numéricas. Para escalas numéricas a numéricas, podemos usar una transformación matemática para "encoger" o "ampliar" los datos dados. Una de las razones de la transformación es eliminar las diferencias en las escalas de las variables. Por ejemplo, si el atributo "salario" varía de "$20.000" a "$100.000", podemos usar la fórmula $\boldsymbol{S = (x - min)/(max - min)}$ para "reducir" cualquier valor salarial conocido, digamos $50.000 a 0.6, un número en [0.0, 1.0]. Si la media del salario se da como $45.000, y la desviación estándar se da como $15.000, los $ 50.000 pueden normalizarse como 0.33. Transformar datos del sistema métrico (por ejemplo, metro, kilómetro) a sistema inglés (por ejemplo, pie y milla) es otro ejemplo. Para escalas categóricas a numéricas, tenemos que asignar un número numérico apropiado a un valor categórico de acuerdo a las necesidades. Las variables categóricas pueden ser ordinal (como menos, moderada y fuerte) y nominal (como rojo, amarillo, azul y verde). Por ejemplo, una variable binaria {yes, no} se puede transformar en "1 = yes y 0 = no". Obsérvese que transformar un valor numérico en un valor ordinal significa transformación con orden, mientras que transformar a un valor nominal es un valor menos rígido transformación. Debemos tener cuidado de no introducir más precisión de la que está presente en los datos originales. Por ejemplo, las escalas de Likert representan a menudo información ordinal con números codificados (1-7, 1-5, y así sucesivamente). Sin embargo, estos números generalmente no implican una escala común de diferencia. Un objeto clasificado como 4 puede no ser dos veces más fuerte en alguna medida como un objeto valorado como 2. A veces, podemos aplicar valores para representar un bloque de números o un rango de variables categóricas. Por ejemplo, podemos usar "1" para representar los valores monetarios de "$0" a "$20.000", y usar "2" para "$20.001- $40.000", y así sucesivamente. Podemos usar "0001" para representar "casa de dos tiendas" y "0002" para "casa de una y media tienda". Todos los métodos "rápidos y sucios" podrían ser usados para transformar datos. No existe un procedimiento único y el único criterio es transformar los datos por conveniencia de uso durante la etapa de minería de datos.

1.4 Modelado (Modeling)

El modelado de datos es donde el software de minería de datos se utiliza para generar resultados para diversas situaciones. En primer lugar, se aplica un análisis de conglomerados y una exploración visual de los datos. Dependiendo del tipo de datos, pueden aplicarse varios modelos. Si la tarea es agrupar datos, y los grupos son dados, el análisis discriminante podría ser apropiado. Si el propósito es la estimación, la regresión es apropiada si los datos son continuos (y la regresión logística si no). Las redes neuronales podrían aplicarse para ambas tareas.

Los árboles de decisión son otra herramienta para clasificar los datos. Otras herramientas de modelado están disponibles también. El punto del software de minería de datos es permitir que el usuario trabaje con los datos para ganar comprensión. Esto es fomentado a menudo por el uso iterativo de modelos múltiples.

1.5 Tratamiento de datos (Data Treatment)

La minería de datos es esencialmente el análisis de datos estadísticos, usualmente usando enormes conjuntos de datos. El proceso estándar de minería de datos es tomar este gran conjunto de datos y dividirlo, usando una parte de los datos (el conjunto de entrenamiento) para el desarrollo del modelo (sin importar qué técnica de modelado se use), y reservando una porción de los datos (el conjunto de pruebas) para probar el modelo que se construye. En algunas aplicaciones se utiliza una tercera división de datos (conjunto de validación) para estimar parámetros a partir de los datos. El principio es que si usted construye un modelo en un determinado conjunto de datos, por supuesto, probará bastante bien. Al dividir los datos y usar parte de él para el desarrollo del modelo, y probarlo en un conjunto separado de datos, se obtiene una prueba más convincente de la precisión del modelo.

Esta idea de dividir los datos en componentes a menudo se lleva a niveles adicionales en la práctica de la minería de datos. Se pueden usar porciones adicionales de los datos para refinar el modelo.

1.6 Técnicas de minería de datos (Data Mining Techniques)

La minería de datos puede lograrse mediante la asociación (Association), la clasificación (Classification), el agrupamiento (Clustering), las predicciones (Predictions), los patrones secuenciales (Sequential Patterns) y las secuencias de tiempo similares (Similar Time Sequences).

En la **Asociación**, la relación de un elemento particular en una transacción de datos sobre otros elementos en la misma transacción se utiliza para predecir patrones. Por ejemplo, si un cliente compra una computadora portátil (***X***), entonces él o ella también compra un mause (***Y***) en el 60% de los casos. Este patrón se produce en el 5.6% de las compras de PC portátiles. Una regla de asociación en esta situación puede ser "***X*** implica ***Y***, donde 60% es el factor de confianza y 5.6% es el factor de soporte". Cuando el factor de confianza y el factor de soporte

están representados por variables lingüísticas "alta" y "baja", respectivamente, la regla de asociación se puede escribir en la forma lógica difusa, como: "donde el factor de soporte es bajo, ***X*** implica ***Y*** es alto". En el caso de muchas variables cualitativas, la asociación difusa es una técnica necesaria y prometedora en la minería de datos.

En **Clasificación**, los métodos están destinados a aprender diferentes funciones que asignan cada elemento de los datos seleccionados en uno de un conjunto predefinido de clases. Dado el conjunto de clases predefinidas, un número de atributos y un "conjunto de aprendizaje (o entrenamiento)", los métodos de clasificación pueden predecir automáticamente la clase de otros datos no clasificados del conjunto de aprendizaje. Dos problemas de investigación clave relacionados con los resultados de la clasificación son la evaluación de la clasificación errónea y el poder de predicción. Las técnicas matemáticas que se utilizan a menudo para construir métodos de clasificación son árboles binarios de decisión, redes neuronales, programación lineal y estadísticas. Mediante el uso de árboles binarios de decisión, se puede construir un modelo de inducción de árboles con un formato "Sí-No" para dividir datos en diferentes clases según sus atributos. Los modelos ajustados a los datos se pueden medir por estimación estadística o entropía de la información. Sin embargo, la clasificación obtenida de la inducción del árbol no puede producir una solución óptima donde la energía de la predicción es limitada. Mediante el uso de redes neuronales, un modelo de inducción neural puede ser construido. En este enfoque, los atributos se convierten en capas de entrada en la red neuronal mientras que las clases asociadas con los datos son capas de salida. Entre capas de entrada y capas de salida, hay un mayor número de capas ocultas procesando la exactitud de la clasificación. Aunque el modelo de inducción neural a menudo produce mejores resultados en muchos casos de minería de datos, ya que las relaciones implican relaciones no lineales complejas, la implementación de este método es difícil cuando hay un gran conjunto de atributos. En los enfoques de programación lineal, el problema de clasificación es visto como una forma especial de programa lineal. Dado un conjunto de clases y un conjunto de variables de atributos, se puede definir un límite de corte (o límite) que separa las clases. Entonces cada clase está representada por un grupo de restricciones con respecto a un límite en el programa lineal. La función objetivo en el modelo de programación lineal puede minimizar la tasa de superposición entre las clases y maximizar la distancia entre las clases. El enfoque de programación lineal resulta en una clasificación óptima. Sin embargo, el tiempo de cálculo requerido puede superar el de los enfoques estadísticos. Varios métodos estadísticos, como la regresión discriminante lineal, la regresión discriminante cuadrática y la regresión logística discriminante son muy populares y se usan comúnmente en las clasificaciones empresariales reales. A pesar de que se ha desarrollado software estadístico para manejar una gran cantidad de datos, los enfoques estadísticos tienen la desventaja de separar eficientemente los problemas multi-clases en los que se tiene que adoptar una comparación en pares (es decir, una clase frente al resto de las clases).

El análisis de **clústeres** toma datos desagrupados y utiliza técnicas automáticas para poner estos datos en grupos. El agrupamiento no se supervisa y no requiere un conjunto de aprendizaje. Comparte una base metodológica común con Clasificación. En otras palabras,

la mayoría de los modelos matemáticos mencionados anteriormente con respecto a la Clasificación pueden aplicarse también al Análisis de Cluster.

El análisis de **predicción** está relacionado con las técnicas de regresión. La idea clave del análisis predictivo es descubrir la relación entre las variables dependientes e independientes, la relación entre las variables independientes (uno frente a otro, uno frente al resto, etc.). Por ejemplo, si las ventas son una variable independiente, entonces el beneficio puede ser una variable dependiente. Utilizando datos históricos tanto de ventas como de beneficios, las técnicas de regresión lineal o no lineal pueden producir una curva de regresión ajustada que puede usarse para la predicción de beneficios en el futuro.

El análisis **secuencial de patrones** busca encontrar patrones similares en la transacción de datos durante un período de negocio. Estos patrones pueden ser utilizados por los analistas de negocios para identificar las relaciones entre los datos. Los modelos matemáticos detrás de los patrones secuenciales son reglas lógicas, lógica difusa, y así sucesivamente. Como una extensión de los patrones secuenciales, **secuencias de tiempo** similares se aplican para descubrir secuencias similares a una secuencia conocida en los períodos de negocios pasados y actuales. En la etapa de minería de datos, se pueden estudiar varias secuencias similares para identificar tendencias futuras en el desarrollo de transacciones. Este enfoque es útil para tratar con bases de datos que tienen características de series temporales.

1.7 Evaluación (Evaluation)

La etapa de interpretación de datos es muy crítica. Asimila el conocimiento de los datos minados. Dos cuestiones son esenciales. Una de ellas es cómo reconocer el valor comercial de los patrones de conocimiento descubiertos en la etapa de minería de datos. Otro problema es qué herramienta de visualización se debe utilizar para mostrar los resultados de minería de datos. Determinar el valor del negocio a partir de patrones de conocimiento descubiertos es similar a jugar "rompecabezas". Los datos minados es un rompecabezas que se debe juntar para un propósito comercial. Esta operación depende de la interacción entre analistas de datos, analistas de negocios y tomadores de decisiones (como gerentes o CEO[1]). Debido a que los analistas de datos pueden no ser plenamente conscientes del objetivo de la meta u objetivo de minería de datos, y mientras que los analistas de negocios pueden no entender los resultados de soluciones matemáticas sofisticadas, la interacción entre ellos es necesaria. Para interpretar adecuadamente los patrones de conocimiento, es importante elegir una herramienta de visualización apropiada. Muchos paquetes de visualización y herramientas están disponibles, incluyendo gráficos circulares, histogramas, diagramas de caja, diagramas de dispersión y distribuciones. Una buena interpretación conduce a decisiones empresariales productivas, mientras que el pobre análisis de interpretación puede perder información útil. Normalmente, cuanto más sencilla sea la interpretación gráfica, más fácil será para los usuarios finales.

[1] **CEO:** Director ejecutivo

1.8 Despliegue (Deployment)

Los resultados del estudio de minería de datos deben ser reportados a los patrocinadores del proyecto. El estudio de minería de datos ha descubierto nuevos conocimientos, que deben estar vinculados a los objetivos originales del proyecto de minería de datos. La administración estará entonces en condiciones de aplicar esta nueva comprensión de su entorno empresarial.

Es importante que el conocimiento obtenido de un estudio particular de minería de datos sea monitoreado para el cambio. El comportamiento del cliente cambia con el tiempo, y lo que era cierto durante el período en que los datos fueron recopilados ya puede haber cambiado. Si ocurren cambios fundamentales, el conocimiento descubierto ya no es cierto. Por lo tanto, es crítico que el dominio de interés sea monitoreado durante su período de implementación.

2 SEMMA

Para ser aplicado con éxito, la solución de minería de datos debe ser vista como un proceso en lugar de un conjunto de herramientas o técnicas. Además del CRISP-DM existe otra metodología bien conocida desarrollada por el instituto SAS, llamada SEMMA. El acrónimo SEMMA significa **muestra** (**Sample**), **explorar** (**Explore**), **modificar** (**Modify**), **modelar** (**Model**), **evaluar** (**Assess**). Comenzando con una muestra estadísticamente representativa de sus datos, SEMMA pretende facilitar la aplicación de técnicas exploratorias estadísticas y de visualización, seleccionar y transformar las variables predictivas más significativas, modelar las variables para predecir los resultados y finalmente confirmar la precisión de un modelo. Una representación gráfica de SEMMA se da en la Fig. 2.2.

Al evaluar el resultado de cada etapa en el proceso SEMMA, se puede determinar cómo modelar las nuevas preguntas planteadas por los resultados anteriores, y así proceder de nuevo a la fase de exploración para el refinamiento adicional de los datos. Es decir, como es el caso en CRISP-DM, SEMMA también impulsado por un ciclo de experimentación altamente iterativo.

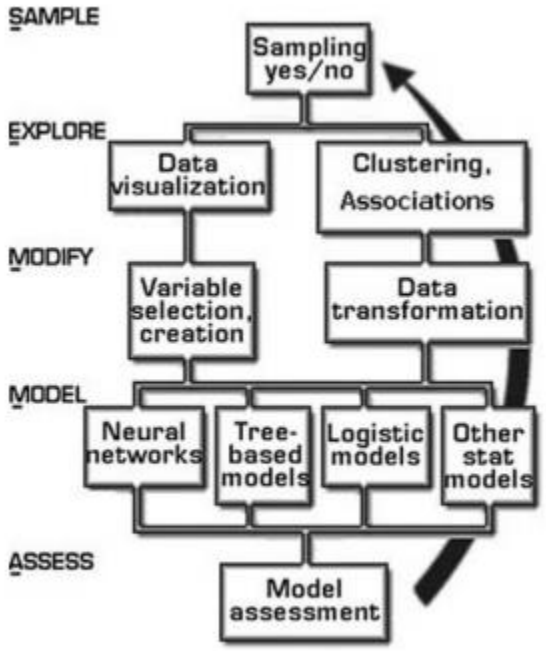

Fig. 2.2. Schematic of SEMMA (original from SAS Institute)

2.1 Pasos en el proceso SEMMA

2.1.1 Paso 1 Muestra (Sample)

Aquí es donde se extrae una porción de un conjunto de datos grande (lo suficientemente grande para contener la información significativa pero lo suficientemente pequeña como para manipular rápidamente). Para obtener un óptimo coste y rendimiento computacional, algunos (incluido el SAS Institute) aboga por una estrategia de muestreo, que aplica una muestra fiable y estadísticamente representativa de los datos completos de detalle. En el caso de conjuntos de datos muy grandes, la extracción de una muestra representativa en lugar de todo el volumen puede reducir drásticamente el tiempo de procesamiento necesario para obtener información comercial crucial. Si los patrones generales aparecen en los datos como un todo, estos serán rastreables en una muestra representativa. Si un nicho (un patrón raro) es tan pequeño que no está representado en una muestra y, sin embargo, tan importante que influye en el cuadro general, debe ser descubierto utilizando métodos exploratorios de descripción de datos. También se recomienda crear conjuntos de datos particionados para una mejor evaluación de la exactitud.

- Entrenamiento (Training) - Se utiliza para el ajuste del modelo.
- Validación (Validation) - Se utiliza para la evaluación y para prevenir el ajuste excesivo.
- Prueba (Test) - Se utiliza para obtener una evaluación honesta de lo bien que generaliza un modelo.

2.1.2 Paso 2 Explorar (Explore)

Aquí es donde el usuario buscó tendencias y anomalías imprevistas con el fin de obtener una mejor comprensión del conjunto de datos. Después de muestrear sus datos, el siguiente paso es explorarlos visualmente o numéricamente para tendencias o agrupaciones inherentes. La exploración ayuda a refinar y redirigir el proceso de descubrimiento. Si la exploración visual no revela tendencias claras, se pueden explorar los datos a través de técnicas estadísticas, incluyendo análisis factorial, análisis de correspondencia y agrupación. Por ejemplo, en la minería de datos para una campaña de correo directo, el agrupamiento puede revelar grupos de clientes con distintos patrones de pedido. Limitar el proceso de descubrimiento a cada uno de estos grupos distintos individualmente puede aumentar la probabilidad de explorar patrones más ricos que pueden no ser lo suficientemente fuertes como para ser detectados si el conjunto de datos debe procesarse juntos.

2.1.3 Paso 3 Modificar (Modify)

Aquí es donde el usuario crea, selecciona y transforma las variables sobre las cuales enfocar el proceso de construcción del modelo. Sobre la base de los descubrimientos en la fase de exploración, puede ser necesario manipular los datos para incluir información como la agrupación de clientes y subgrupos significativos o introducir nuevas variables. También puede ser necesario buscar valores atípicos y reducir el número de variables, para reducirlas a las más significativas. También puede ser necesario modificar los datos cuando cambian los datos "minados". Debido a que la minería de datos es un proceso dinámico e iterativo, puede actualizar los métodos o modelos de minería de datos cuando hay nueva información disponible.

2.1.4 Paso 4 Modelo (Model)

Aquí es donde el usuario busca una combinación de variables que predice de manera fiable un resultado deseado. Una vez que prepara sus datos, está listo para construir modelos que expliquen patrones en los datos. Técnicas de modelado de minería de datos incluyen las redes neuronales artificiales, árboles de decisión, análisis conjunto aproximado, máquinas de vectores de soporte, modelos logísticos, y otros modelos estadísticos - como el análisis de series de tiempo, el razonamiento basado en la memoria, y el análisis de componentes principales. Cada tipo de modelo tiene fortalezas particulares y es apropiado dentro de situaciones específicas de minería de datos dependiendo de los datos. Por ejemplo, las redes neuronales artificiales son muy buenos para ajustar las relaciones no lineales altamente complejas, mientras que el análisis de los conjuntos aproximados se conoce para producir resultados fiables con situaciones problemáticas inciertos e imprecisos.

2.1.5 Paso 5 Evaluar (Assess)

Aquí es donde el usuario evalúa la utilidad y la fiabilidad de los resultados del proceso de minería de datos. En este último paso del proceso de minería de datos, el usuario evalúa los modelos para estimar su rendimiento. Un medio común de evaluar un modelo es aplicarlo a

una parte del conjunto de datos puesta a un lado (y no utilizada durante el edificio del modelo) durante la etapa de muestreo. Si el modelo es válido, debería funcionar para esta muestra reservada, así como para la muestra utilizada para construir el modelo. Del mismo modo, puede probar el modelo contra datos conocidos. Por ejemplo, si sabe qué clientes de un archivo tienen altas tasas de retención y su modelo predice la retención, puede comprobar si el modelo selecciona a estos clientes con precisión. Además, las aplicaciones prácticas del modelo, como los envíos parciales en una campaña de correo directo, ayudan a probar su validez. El sitio web de minería de datos KDNuggets proporcionó los datos mostrados en la Fig. 2.3 sobre el uso relativo de las metodologías de minería de datos.

El enfoque SEMMA es completamente compatible con el enfoque CRISP. Ambos ayudan al proceso de descubrimiento del conocimiento. Una vez que los modelos se obtienen y se prueban, pueden entonces desplegarse para ganar valor con respecto a la aplicación del negocio o de la investigación.

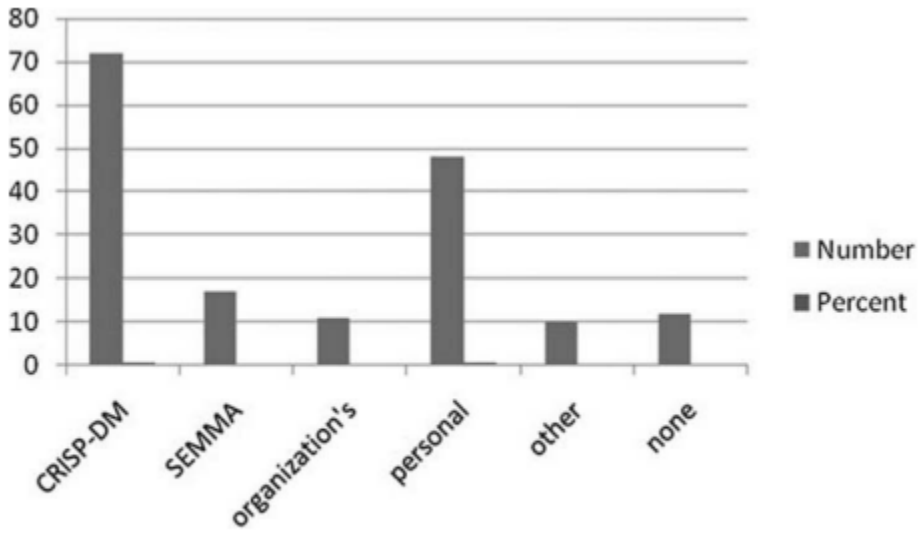

Fig. 2.3. Poll results – data mining methodology (conducted by KDNuggets.com on April 2004)

2.1.6 Ejemplo de aplicación de proceso de minería de datos

Nayak y Qiu (2005) demostraron el proceso de minería de datos en un proyecto australiano de desarrollo de software. Primero relacionaremos su proceso informado, y después compararemos esto con los marcos de CRISP y SEMMA.

El propietario del proyecto era una empresa internacional de telecomunicaciones que realizaba más de 50 proyectos de software anualmente. Se organizaron procesos de gestión de configuración de software, gestión de riesgos de software, informes de métrica de proyecto de software y gestión de informes de problemas de software. Nayak y Qiu estaban interesados en extraer los datos de los Informes de Problemas de Software. Todos los informes de problemas se recopilaron en toda la empresa (más de 40.000 informes). Para cada informe, se disponía de datos para incluir los datos que se muestran en la Tabla 2.1:

Table 2.1. Selected attributes from problem reports

Attribute	Description
Synopsis	Main issues
Responsibility	Individuals assigned
Confidentiality	Yes or no
Environment	Windows, Unix, etc.
Release note	Fixing comment
Audit trail	Process progress
Arrival date	
Close date	
Severity	Text describing the bug and impact on system
Priority	High, Medium, Low
State	Open, Active, Analysed, Suspended, Closed, Resolved, Feedback
Class	Sw-bug, Doc-bug, Change-request, Support, Mistaken, Duplicate

El proceso de minería de datos informó la definición de metas incluidas, pre-procesamiento de datos, modelado de datos y análisis de resultados.

2.1.6.1 Definición del objetivo (Goal Definition)

Se espera que la extracción de datos sea útil en dos áreas. El primero involucraba la etapa inicial de estimación y planificación de un proyecto de software, los ingenieros de la compañía tenían que estimar el número de líneas de código, el tipo de documentos a entregar y los tiempos estimados. La precisión en esta etapa mejoraría enormemente las decisiones de selección de proyectos. Poco apoyo de herramientas estaba disponible para estas actividades, y las estimaciones de estos tres atributos se basaron en la experiencia apoyada por estadísticas sobre proyectos anteriores. Por lo tanto, los proyectos que involucran nuevos tipos de trabajo eran difíciles de estimar con confianza. La segunda área de la aplicación de minería de datos se refiere al sistema de recopilación de datos, que tenía capacidad limitada de recuperación de información. Los datos fueron almacenados en archivos planos, y fue difícil reunir información relacionada con temas específicos.

2.1.6.2 Pre-procesamiento de datos (Data Pre-Processing)

Este paso consistió en la selección de atributos, limpieza de datos y transformación de datos.

Selección de campo de datos: Algunos de los datos no eran pertinentes al ejercicio de minería de datos, y se ignoró. De las variables dadas en la Tabla 2.1, Confidencialidad, Medio ambiente, nota de prensa, y de seguimiento de auditoría fueron ignoradas como no tener valor minería de datos. Sin embargo, se utilizaron durante el pre-procesamiento y el post-procesamiento para ayudar en la selección de datos y obtener una mejor comprensión de las reglas generadas. Para la estabilidad de datos, sólo se seleccionaron los informes de problemas para los valores de estado de cerrado.

Cada vez que se creaba un informe de problema, el líder del proyecto tenía que determinar cuánto tiempo tomó la corrección, cuántas personas participaron, la gravedad del impacto del cliente, el impacto en el costo y el horario y el tipo de problema. Así, los atributos enumerados a continuación fueron seleccionados como los más importantes:

- Severidad
- Prioridad
- Clase
- Fecha de llegada
- Fecha de cierre
- Responsable
- Sinopsis

Los cinco primeros atributos tenían valores fijos y el atributo responsable se convirtió en un recuento de los asignados al problema. Todos estos atributos podrían tratarse a través de herramientas convencionales de minería de datos. La sinopsis era datos de texto que requerían minería de texto. Clase fue seleccionado como el atributo objetivo, con los posibles resultados dados en la Tabla 2.2:

Limpieza de datos (Data Cleaning): La limpieza implicó la identificación de valores perdidos, inconsistentes o erróneos. Las herramientas utilizadas en este paso del proceso incluyeron herramientas gráficas para proporcionar una imagen de las distribuciones y estadísticas tales como máximos, mínimos, valores medios y sesgo. Algunas entradas eran claramente inválidas, causadas por error humano o por la evolución del sistema de notificación de problemas. Por ejemplo, con el tiempo, la entrada para el atributo Class cambió de SW-bug a sw-bug. Los errores corregibles fueron corregidos. Si no se corrigieron todos los errores detectados para un informe, ese informe se descartó del estudio.

Transformación de Datos (Data Transformation): Los atributos Llegada-Fecha y Fecha-Cierre fueron útiles en este estudio para calcular la duración. Se requería información adicional para incluir la zona horaria. El atributo responsable contenía información que identificaba cuántas personas estaban involucradas. Se creó un atributo Time-tofix multiplicando la duración multiplicada por el número de personas y luego categorizado en valores discretos de 1 día, 3 días, 7 días, 14 días, 30 días, 90 días, 180 días y 360 días (representando más de Una persona-año).

En esta aplicación, 11.000 de los 40.000 informes de problemas originales fueron dejados. Provienen de más de 120 proyectos completados durante el período 1996-2000. Se obtuvieron cuatro atributos:

Table 2.2. Class outcomes

Sw-bug	Bug from software code implementation
Doc-bug	Bug from documents directly related to the software product
Change-request	Customer enhancement request
Support	Bug from tools or documents, not the software product itself
Mistaken	Error in either software or document
Duplicate	Problem already covered in another problem report

- Tiempo para arreglar
- Clase
- Severidad
- Prioridad

La minería de texto se aplicó a 11.364 registros, de los cuales 364 no tenían valores de tiempo por lo que 11.000 se utilizaron para la clasificación convencional de minería de datos.

2.1.6.3 Modelado de datos (Data Modeling)

La minería de datos proporciona funcionalidad no proporcionada por las técnicas generales de consulta de base de datos, que no pueden tratar el gran número de registros con estructuras dimensionales altas. La minería de datos proporcionó una funcionalidad útil para responder a preguntas como el tipo de documentos de proyecto que requieren mucho tiempo de equipo de desarrollo para la reparación de errores o el impacto de varios valores de atributos de sinopsis, gravedad, prioridad y clase. Se utilizaron varias herramientas de minería de datos.

- El modelo de predicción fue útil para la evaluación del consumo de tiempo, dando estimaciones más sólidas para la estimación y planificación del proyecto.
- El análisis de enlaces fue útil para descubrir asociaciones entre valores de atributos.
- La minería de texto fue útil para analizar el campo Sinopsis.

El software de minería de datos CBA se usó para el análisis de clasificación y regla de asociación, C5 para clasificación y TextAnalyst para minería de texto. Un ejemplo de regla de clasificación fue:

IF Severity non-critical AND Priority medium
THEN Class is Document with 70.72% confidence with support value of 6.5%

There were 352 problem reports in the training data set having these conditions, but only 256 satisfied the rule's conclusion.

Another rule including time-to-fix was more stringent:

IF 21 ≤ time-to-fix ≤ 108
AND Severity non-critical AND Priority medium
THEN Class is Document with 82.70% confidence with support value of 2.7%

There were 185 problem reports in the training data set with these conditions, 153 of which satisfied the rule's conclusion.

2.1.6.4 Análisis de Resultado (Analysis of Results)

Clasificación y regla de asociación de minería: Los datos fueron estratificados utilizando el muestreo basado en la elección en lugar de muestreo aleatorio. Esto proporcionó un número igual de muestras para cada valor de campo de atributo de destino. Esto mejoró la probabilidad de obtener reglas para grupos con conteo de valores pequeños (equilibrando así los datos). Se generaron tres conjuntos diferentes de entrenamiento de diferentes tamaños. El primer conjunto de datos incluyó 1.224 informes de problemas de un proyecto de software. El segundo conjunto de datos consistió en valores igualmente distribuidos de 3.400 informes de problemas seleccionados de todos los proyectos de software. El tercer conjunto de datos constaba de 5.381 informes de problemas seleccionados de todos los proyectos.

El apoyo mínimo y la confianza se utilizaron para controlar el modelado de reglas. El apoyo mínimo es una limitación que requiere por lo menos el número de casos declarados presentes en el conjunto de entrenamiento. Un apoyo mínimo alto dará menos reglas. La confianza es la fuerza de una regla medida por la clasificación correcta de los casos. En la práctica, estos son difíciles de establecer antes del análisis, por lo que se utilizaron combinaciones de apoyo mínimo y confianza.

En esta aplicación, fue difícil para el software CBA obtener la clasificación correcta en los datos de prueba por encima del 50%. El uso de la misma densidad de casos no se encontró para dar modelos más precisos en este estudio, aunque parece un enfoque racional para más investigación. El uso de múltiples niveles de soporte tampoco fue encontrado para mejorar las tasas de error, y la minería de soporte único produjo un número menor de reglas. Sin embargo, se obtuvieron reglas útiles.

C5 también se aplicó para la clasificación de minería. C5 utilizó la validación cruzada, que divide el conjunto de datos en subconjuntos (pliegues), tratando cada pliegue como un caso de prueba y el resto como conjuntos de entrenamiento con la esperanza de encontrar un resultado mejor que un proceso de conjunto de entrenamiento único. C5 también tiene una opción de impulso, que genera y combina varios clasificadores en los esfuerzos para mejorar la precisión predictiva. Aquí C5 produjo conjuntos de reglas más grandes, con ajustes ligeramente mejores con los datos de entrenamiento, aunque a aproximadamente el mismo nivel. La validación cruzada y el impulso no producirían reglas adicionales, pero se centrarían en reglas más precisas.

Text Mining: El texto puro para el atributo Synopsis se categorizó en una serie de tipos de documentos específicos, como "SRS - missing requirements" (con SRS representando la especificación de requisito de software), "SRS - capacidad de desactivar el envío de SOH" Cambios necesarios para SCMP_2.0.0 "y así sucesivamente. Se utilizó TextAnalyst. Este producto construye una red semántica para la investigación de datos de texto. A cada elemento de la red semántica se le asigna un valor de peso y relaciones con otros elementos de la red, a los que también se le asigna un valor de peso. Los usuarios no están obligados a especificar reglas predefinidas para construir la red semántica. TextAnalyst proporcionó un árbol de red semántica que contiene las palabras o combinaciones de palabras más

importantes (conceptos), y reportó relaciones y pesos entre estos conceptos que van de 0 a 100, aproximadamente análogo a la probabilidad. La minería de texto se aplicó a 11.226 casos.

3 Comparación de CRISP y SEMMA

El caso de Nayak y Qiu demuestra un proceso de minería de datos para una aplicación específica, que incluye aspectos interesantes de los requisitos de limpieza y transformación de datos, así como una amplia variedad de tipos de datos para incluir texto. CRISP y SEMMA se crearon como marcos amplios, los cuales necesitan ser adaptados a circunstancias específicas (ver Tabla 2.3). Ahora revisaremos cómo el caso de Nayak y Qiu encaja en estos marcos.

Nayak y Qiu comenzaron con un conjunto claramente establecido de objetivos: desarrollar herramientas que utilizarían mejor la riqueza de datos en los informes de problemas de proyectos de software.

Examinaron los datos disponibles e identificaron lo que sería útil. Gran parte de la información de los informes de problemas se descartó. SEMMA incluye esfuerzos de muestreo aquí, que CRISP incluiría en la preparación de datos, y que Nayak y Qiu lograron después de la transformación de datos. Se utilizaron conjuntos de entrenamiento y prueba como parte de la aplicación de software.

Table 2.3. Comparison of methods

CRISP	SEMMA	Nayak & Qiu
Business understanding	Assumes well-defined question	Goals were defined Develop tools to better utilize problem reports
Data understanding	*S*ample *E*xplore	Looked at data in problem reports
Data preparation	*M*odify data	Data pre-processing Data cleaning Data transformation
Modeling	*M*odel	Data modeling
Evaluation	*A*ssess	Analyzing results
Deployment		

Los datos fueron limpiados, y los informes con las observaciones faltantes fueron descartados del estudio. La preparación de datos implicó la transformación de datos. En concreto, utilizaron dos atributos de informe de problemas para generar la duración del proyecto, que se transformó aún más multiplicando por el número de personas asignadas (disponibles por nombre, pero sólo se necesitaban los recuentos). La medida resultante de esfuerzo se transformó aún más en categorías que reflejaban importancia relativa sin desorden de detalle.

El modelado incluyó análisis de clasificación y regla de asociación desde la primera herramienta de software (CBA), una replicación de la clasificación con C5 y análisis de texto independiente con TextAnalyst. Nayak y Qiu generaron una variedad de modelos manipulando el soporte mínimo y los niveles de confianza en el software.

La evaluación (evaluación) fue realizada por Nayak y Qiu a través del análisis de los resultados en términos del número de reglas, así como la precisión de los modelos de clasificación aplicados al conjunto de pruebas.

CRISP aborda el despliegue de modelos de minería de datos, que está implícito en cualquier estudio. Los modelos de Nayak y Qiu fueron presumiblemente desplegados, pero eso no fue abordado en su informe.

3.1 Manipulación de datos (Handling Data)

Un estudio reciente de minería de datos en seguros aplicó un proceso de descubrimiento de conocimiento. Este proceso implicó aplicar iterativamente los pasos que cubrimos en CRISP-DM, y demostrar cómo la metodología puede funcionar en la práctica.

3.2 Etapa 1. La comprensión de negocios (Business Understanding)

Se necesitaba un modelo para predecir qué clientes serían insolventes lo suficientemente pronto como para que la empresa tomara medidas preventivas (o medidas para evitar perder buenos clientes). Esta meta incluía minimizar la clasificación errónea de clientes legítimos.

En este caso, el período de facturación fue de 2 meses. Los clientes usaron su teléfono durante 4 semanas y recibieron facturas aproximadamente una semana después. El pago se realizó un mes después de la fecha de facturación. En la industria, las empresas normalmente dan a los clientes aproximadamente 2 semanas después de la fecha de vencimiento antes de tomar acción, momento en el que el teléfono se desconectó si la factura no pagada era mayor que una cantidad establecida. Las facturas se enviaban cada mes durante otros 6 meses, período durante el cual el último cliente podía hacer arreglos de pago. Si no se recibió ningún pago al final de este período de seis meses, el saldo impago se transfirió a la categoría incobrable.

Este estudio hipotetizó que los clientes insolventes cambiarían sus hábitos de llamada y uso de teléfono durante un período crítico antes e inmediatamente después de la terminación del período de facturación. Los cambios en los hábitos de llamada, combinados con patrones de pago, fueron probados por su capacidad de proporcionar predicciones sólidas de insolvencias futuras.

3.3 Etapa 2. La comprensión de Datos (Data Understanding)

La información estática del cliente estaba disponible en los archivos del cliente. Se disponía de datos temporales sobre las facturas, los pagos y el uso. Los datos provienen de varias bases de datos, pero todas estas bases de datos eran internas a la empresa. Se creó un almacén de datos para recopilar y organizar estos datos. Los datos fueron codificados para proteger la

privacidad del cliente. Los datos incluyeron información de clientes, uso de teléfonos desde centros de conmutación, información de facturación, informes de pago por cliente, desconexiones telefónicas debido a un fallo en el pago, reconexiones telefónicas después del pago e informes de anulaciones de contratos permanentes.

Se seleccionaron datos para 100.000 clientes, que abarcaban un período de 17 meses, y fueron recolectados de una región rural / agrícola de clientes, un área de turismo semirrural y un área urbana / industrial para asegurar representaciones representativas de la base de clientes de la empresa. El almacén de datos utilizó más de 10 gigabytes de almacenamiento de datos sin procesar.

3.4 Etapa 3. Preparación de los datos (Data Preparation)

Los datos fueron sometidos a pruebas de calidad, y los datos que no fueron útiles para el estudio fueron filtrados. Los datos heterogéneos estaban interrelacionados. Como ejemplos, quedó claro que las llamadas de bajo costo tuvieron poco impacto en el estudio. Esto permitió una reducción del 50% en el volumen total de datos. El bajo porcentaje de casos fraudulentos hizo necesario limpiar los datos de valores faltantes o erróneos debido a diferentes prácticas de registro dentro de la organización ya la dispersión de fuentes de datos. Por lo tanto, era necesario cruzar los datos tales como desconexiones de teléfono. Los datos rezagados requirieron la sincronización de diferentes elementos de datos.

La sincronización de datos reveló una serie de clientes insolventes con información que faltaba que tenía que ser eliminado del conjunto de datos. Por lo tanto, era necesario reducir y proyectar datos, por lo que la información se agrupa por cuenta para facilitar la manipulación de datos y los datos de los clientes se agregan en períodos de 2 semanas. Se aplicaron estadísticas para encontrar características que fueran factores discriminantes para clientes solventes versus insolventes. Los datos incluyeron lo siguiente:

- Categoría de cuenta telefónica (23 categorías, como teléfono público, negocios, etc.).
- El monto promedio adeudado se calculó para todos los clientes solventes e insolventes. Los clientes insolventes tenían promedios significativamente más altos en todas las categorías de cuentas.
- Los cargos adicionales en las facturas fueron identificados mediante la comparación de los cargos totales por el uso del teléfono para el período en contraposición a los saldos arrastrados o compras de hardware u otros servicios. Esto también resultó ser estadísticamente significativo en las dos categorías de resultados.
- Se investigó el pago por cuotas. Sin embargo, esta variable no fue estadísticamente significativa.

3.5 Etapa 4. Modelado (Modeling)

El problema de predicción fue la clasificación, con dos clases: posiblemente disolvente (99,3% de los casos) y posiblemente insolvente (0,7% de los casos). Por lo tanto, el recuento de los casos de insolvencia fue muy pequeño en un período de facturación dado. El costo del

error varió ampliamente en las dos categorías. Esto ha sido señalado por muchos como un problema de clasificación muy difícil.

Un nuevo conjunto de datos fue creado a través de muestreo estratificado para clientes solventes, alterando la distribución de clientes a un 90% de solvente y un 10% de insolvencia. Se mantuvieron todos los casos de insolvencia, y se tuvo cuidado de mantener una representación proporcional del conjunto de datos sobre variables como la región geográfica. Se desarrolló un conjunto de datos de 2.066 casos totales.

Se estableció un período crítico para cada cuenta telefónica. Para aquellas cuentas que fueron anuladas, este período crítico fue los últimos 15 períodos de dos semanas antes de la interrupción del servicio. Para las cuentas que permanecieron activas, el período crítico se estableció como un período similar a la posible interrupción. Había seis posibles fechas de interrupción por año. Para las cuentas activas, una de estas seis fechas fue seleccionada al azar.

Para cada cuenta, las variables se definieron contando la medida apropiada para cada período de 2 semanas en el período crítico para esa observación. Al final de esta fase, se crearon nuevas variables para describir el uso del teléfono por cuenta en comparación con un promedio móvil de cuatro períodos previos de 2 semanas. En esta etapa, había 46 variables como factores discriminantes candidatos. Estas variables incluyeron 40 variables medidas como hábitos de llamada durante 15 períodos de dos semanas, así como variables relativas al tipo de cliente, independientemente de si un cliente era nuevo o no, y cuatro variables relacionadas con el pago de la factura del cliente.

Se utilizaron análisis discriminantes, árboles de decisión y algoritmos de redes neuronales para probar hipótesis sobre el conjunto de datos reducidos de 2.066 casos medidos sobre 46 variables. El análisis discriminante produjo un modelo lineal, la red neural surgió como un modelo no lineal y el árbol de decisión fue un clasificador basado en reglas.

3.6 Etapa 5. Evaluación (Evaluation)

Se realizaron experimentos para probar y comparar el rendimiento. El conjunto de datos se dividió en un conjunto de formación (alrededor de dos tercios de los 2.066 casos) y el conjunto de pruebas (los casos restantes). Los errores de clasificación se muestran comúnmente en matrices de coincidencias (llamadas matrices de confusión por algunos). Una matriz de coincidencias muestra el recuento de casos correctamente clasificados, así como el número de casos clasificados en cada categoría incorrecta. Pero en muchos estudios de minería de datos, el modelo puede ser muy bueno para clasificar una categoría, mientras que es muy pobre en la clasificación de otra categoría. El valor primario de la matriz de coincidencias es que identifica qué tipos de errores se producen. Puede ser mucho más importante evitar un tipo de error que otro. Por ejemplo, un oficial de préstamo bancario sufre mucho más de dar un préstamo a alguien que se espera que pague y no hace que cometer el error de no dar un préstamo a un solicitante que realmente habría pagado. Ambas instancias serían errores de clasificación, pero en la minería de datos, a menudo una categoría de error

es mucho más importante que otra. Las matrices de coincidencia proporcionan un medio para centrarse en qué tipos de errores suelen producir modelos particulares.

Una forma de reflejar la importancia del error relativo es a través del costo. Esta es una idea relativamente simple, que permite al usuario asignar costos relativos por tipo de error. Por ejemplo, si nuestro modelo predijo que una cuenta era insolvente, esto podría implicar una cancelación promedio de $ 200. Por otro lado, a la espera de una cuenta que en última instancia fue reembolsado podría implicar un costo de $ 10. Por lo tanto, habría una gran diferencia en el costo de los errores en este caso. Tratar un caso que resultó ser reembolsado como una cuenta muerta podría arriesgarse a la pérdida de $ 190, además de alienar al cliente (que puede o no tener implicaciones futuras de rentabilidad). Por el contrario, el tratamiento de una cuenta que nunca iba a ser reembolsado puede implicar llevar la cuenta en los libros más de lo necesario, a un costo adicional de $ 10. Aquí, una función de coste para la matriz de coincidencia podría ser:

$ 190 X (buena cuenta de cierre) + $ 10 X (manteniendo abierta la cuenta mala)

(Tenga en cuenta que usamos nuestros propios costos en dólares para propósitos de demostración, y éstos no se basaron en el caso real). Esta medida (como la tasa de clasificación correcta) se puede utilizar para comparar modelos alternativos.

Se utilizó SPSS para el análisis discriminante, incluyendo un procedimiento de selección progresiva paso a paso. El mejor modelo incluyó 17 de las 46 variables disponibles. Utilizando los mismos costos de clasificación errónea se obtuvo la matriz de coincidencias mostrada en la Tabla 2.4.

La precisión de clasificación general se obtiene dividiendo el número correcto de clasificaciones (50 + 578 = 628) por el número total de casos (718). Así, los datos de la prueba se clasificaron correctamente en 87.5. El valor de la función de coste aquí era:

$ 190 X 76 + $ 10 X 14 = $ 14,580

La alta proporción de casos de disolventes realmente clasificados como insolventes se consideró inaceptable, ya que eliminaría demasiados buenos clientes. El experimento se volvió a usar utilizando probabilidades a priori. Esta mejora de la producción significativamente, como se muestra en la matriz de coincidencias en la Tabla 2.5.

Los datos de la prueba se clasificaron correctamente en el 93,0% de los casos. Para los datos de formación, esta cifra fue del 93,6%. Los modelos suelen ajustarse a los datos de entrenamiento un poco mejor que los datos de prueba, pero eso se debe a que se basaron en datos de entrenamiento. Los datos de prueba independientes proporcionan una prueba mucho mejor. La precisión de los clientes insolventes, que es muy importante porque cuesta mucho más, disminuyó de un 78% en los datos de entrenamiento a un 56% en los datos de la prueba. El valor de la función de coste aquí era el siguiente:

$ 190 X 22 + $ 10 X 28 = $ 4,460

Table 2.4. Coincidence matrix – equal misclassification costs

Telephone bill	Model insolvent	Model solvent	
Actual Insolvent	50	14	64
Actual Solvent	76	578	654
	126	593	718

Table 2.5. Coincidence matrix – unequal misclassification costs

Telephone bill	Model insolvent	Model solvent	
Actual Insolvent	36	28	64
Actual Solvent	22	632	654
	58	660	718

Desde el punto de vista de los costos totales, se consideró que el modelo que utilizaba los costos desiguales de clasificación errónea (utilizando costos reales) era más útil.

Las 17 variables identificadas en el análisis discriminante se utilizaron para los otros dos modelos. Se emplearon los mismos conjuntos de entrenamiento y prueba. El conjunto de entrenamiento se utilizó para construir un modelo clasificador basado en reglas. La matriz de coincidencias para el conjunto de prueba se muestra en la Tabla 2.6.

Cuadro 2.6. Matriz de coincidencia - el modelo basado en reglas Factura telefónica Modelo insolvente Modelo de disolvente Real Insolvente 38 26 64 Real Solvente 8 646 654 46 672 718

Así, los datos de la prueba se clasificaron correctamente en el 95,26% de los casos. Para los datos de entrenamiento, esta cifra fue del 95,3%. El valor de la función de coste aquí

$$\$ 190 \text{ X } 8 + \$ 10 \text{ X } 26 = \$ 1{,}780$$

Esto fue una mejora con respecto al modelo de análisis discriminante.

Una serie de experimentos se llevaron a cabo con un modelo de red neuronal utilizando las mismas 17 variables y conjunto de formación. La matriz de coincidencias resultante sobre los datos del ensayo se muestra en la Tabla 2.7.

Tabla 2.7. Matriz de coincidencia - el modelo de red neural Factura telefónica Modelo insolvente Disolvente del modelo Insolvente real 24 40 64 Solvente real 11 643 654 35 683 718

Los datos de la prueba se clasificaron correctamente en el 92,9% de los casos. Para los datos de formación, esta cifra fue del 94,1%. El valor de la función de coste aquí

$$\$ 190 \text{ X } 11 + \$ 10 \text{ X } 40 = \$ 2{,}490$$

Sin embargo, estos resultados fueron inferiores a los del modelo de árboles de decisión.

El primer objetivo era maximizar la precisión de la predicción de clientes insolventes. El clasificador de árbol de decisión parecía ser el mejor en hacer eso. El segundo objetivo era minimizar la tasa de error para los clientes solventes. El modelo de red neural estuvo cerca del rendimiento del modelo de árbol de decisión. Se decidió utilizar los tres modelos caso por caso.

Table 2.6. Coincidence matrix – the rule-based model

Telephone bill	Model insolvent	Model solvent	
Actual Insolvent	38	26	64
Actual Solvent	8	646	654
	46	672	718

Table 2.7. Coincidence matrix – the neural network model

Telephone bill	Model insolvent	Model solvent	
Actual Insolvent	24	40	64
Actual Solvent	11	643	654
	35	683	718

Table 2.8. Coincidence matrix – combined models

Telephone bill	Model insolvent	Model solvent	Unclassified	Total
Actual Insolvent	19	17	28	64
Actual Solvent	1	626	27	654
	20	643	91	718

3.7 Etapa 6. Despliegue (Deployment)

Cada cliente fue examinado utilizando los tres algoritmos. Si los tres coincidieron en la clasificación, ese resultado fue adoptado. Si hubo desacuerdo en los resultados del modelo, el cliente fue clasificado como no clasificado. El uso de este esquema sobre el conjunto de ensayo produjo la matriz de coincidencias mostrada en la Tabla 2.8.

Así, los datos de la prueba se clasificaron correctamente en el 89,8% de los casos. Pero sólo un cliente realmente solvente habría sido desconectado sin más análisis. El valor de la función de coste aquí

$190 X 1 + $10 X 17 = $360

Los pasos utilizados en esta aplicación coinciden con las seis etapas que hemos presentado. La selección de datos se refiere a Aprendizaje del dominio de la aplicación y Creación de un conjunto de datos de destino. El preprocesamiento de datos implica la limpieza de datos y el preprocesamiento. La transformación de datos implica la reducción y proyección de datos. Data Mining se expandió en la aplicación anterior para incluir (1) la elección de la función de minería de datos, (2) la elección de los algoritmos de minería de datos, y (3) la minería de datos. Interpretación de los datos implica la interpretación y el uso del conocimiento descubierto.

4 Resumen

El proceso de minería de datos estándar de la industria CRISP-DM tiene seis etapas: (1) Entendimiento de negocios, (2) Entendimiento de datos, (3) Preparación de datos, (4) Modelación, (5) Evaluación y (6) Despliegue. Selección de datos y la comprensión, la preparación y la interpretación del modelo requieren trabajo en equipo entre los analistas de minería de datos y analistas de negocios, mientras que la transformación de datos y minería de datos se llevan a cabo por los analistas de minería de datos solo. Cada etapa es una preparación para la siguiente etapa.