Análisis preliminar para la construcción de un modelo de ventas en una empresa comercial textil a través de minería de datos
(Octubre de 2012)

Christian Orlando Zapata Vásquez
chorzavas@gmail.com

Juan Sebastián Jaramillo
jsebas980@gmail.com

***Resumen*.** La minería de datos es un proceso de extracción de información útil para toda compañía, la cual puede ser utilizada para predecir comportamientos en el mercado y permitir la toma de decisiones confiables para que los empresarios mejoren los niveles de venta de sus productos. Adoptando CRISP–DM como metodología formal de minería de datos, se adelantan los procesos de comprensión del negocio, entendimiento y adecuación de los datos, y estudio de correlación de variables; los cuales permiten orientar el trabajo, familiarizarse con el comportamiento de las ventas de la empresa y determinar la calidad de la información base para la construcción del modelo.

***Palabras claves*.** Minería de datos, modelación de las ventas, calidad de la información.

Abstract. Data mining is a process of extracting useful information for the entire company, which can be used to predict behavior on the market to allow the making of reliable decisions so businessmen improve the levels of sale of its products. Adopting CRISP - DM as formal data mining methodology is informed about business comprehension, data understanding and formatting, and correlations of variables processes; in order to focus the work, familiarize with sales behavior and determine the quality of the basic information to model construction.

Keywords: Datamining, sales modeling, information quality.

1. INTRODUCCIÓN

El presente artículo pretende servir de referencia en la realización de proyectos de minería de datos, considerando algunos aspectos que se deben tener en cuenta a la hora de validar las fuentes de datos, identificar las características de almacenamiento y determinar la relación entre las variables independientes y la variable dependiente.[1]

2. INFORMACIÓN Y CONOCIMIENTO, EL OBJETIVO DE LA MINERÍA DE DATOS

Para toda compañía es fundamental la información. Disponer de la información necesaria en el momento preciso, ofrece ventajas competitivas. De ahí la frase célebre "Quien tiene la información es quien posee el poder"[2] pero, ¿Qué es la información?. La información es un conjunto de datos, que permiten construir un mensaje o idea y que le da sentido a un tema [1]. Con base en la definición anterior, por ejemplo, tener almacenado 100 registros y mostrárselos a alguien externo como a una empresa estos no le darán alguna idea o mensaje, pero para la persona que conozca del negocio podrá descifrar el mensaje y darse

[1] En el contexto de minería de datos, la variable dependiente representa la variable de interés o de estudio (demanda del producto o nivel de ventas), mientras que las variables independientes corresponden a las variables explicativas del modelo, a partir de las cuales se permite predecir el comportamiento.

[2] David Hume (n. 7 de mayo de 1711, Edimburgo † 25 de agosto de 1776)

cuentas de cosas como cual fue el cliente que mas compró, el mes en que más se vendió, el producto más cotizado, entre otros. Generalmente, las empresas procesan informes acerca del pasado, es decir, información de cuanto se vendió, cuanto ganó la empresa, cuanto perdió. Pero nunca se hacen informes a futuro, solo en algunos casos extrapolaciones de ventas que no necesariamente esclarecen el movimiento a futuro, es en este apartado donde el conocimiento juega un papel fundamental. Pero vuelve a surgir otra pregunta ¿Qué es el conocimiento?. El conocimiento es el conjunto de información almacenada mediante la experiencia o el aprendizaje [2]. Teniendo claros estos conceptos, se puede hablar que la minería de datos busca interpretar grandes cantidades de datos y encontrar relaciones o patrones y de esta manera extraer información implícita, que permita brindar un conocimiento útil a las compañías.

3. MINERÍA DE DATOS

El concepto de la minería de datos es la extracción de información implícita almacenada en una base de datos [3]. Sin embargo, más allá de esta definición la minería de datos es un conjunto de herramientas que juntas logran darle sentido a esta definición. La minería, vista desde un punto de vista integrado, es la combinación entre estadística, inteligencia artificial y manejo de bases de datos [4].

La minería de datos puede contribuir significativamente en las aplicaciones de administración empresarial basada en la relación con el cliente. Sin embargo, no es la única área en la cual se puede implementar este tipo de modelos. Entre otros, por ejemplo:

Patrones de fuga: un uso habitual, es el de la detección de patrones de fuga. En muchas industrias como la banca o las telecomunicaciones, existe un comprensible interés en detectar aquellos clientes que puedan estar pensando en cancelar sus contratos, para posiblemente pasarse a la competencia.

Fraudes: un caso análogo es el de la detección de transacciones de lavado de dinero o de fraude en el uso de tarjetas de crédito o de servicios de telefonía móvil e incluso, en la relación de los contribuyentes con la recaudación de impuestos. Generalmente, estas operaciones fraudulentas o ilegales suelen seguir patrones característicos que permiten, con cierto grado de probabilidad, distinguirlas de las legítimas y desarrollar así mecanismos para tomar medidas rápidas frente a cualquier situación.

Recursos humanos: La minería de datos también puede ser útil para los departamentos de recursos humanos, en la identificación de las características de los empleados de mayor éxito. La información obtenida puede ayudar a la contratación de personal, centrándose en los esfuerzos de sus empleados y los resultados obtenidos.

Banca: La banca almacena la información de cuentas, tarjetas de crédito y transacciones. Esta información sirve de complemento, a la información de compra de productos de las personas en tiendas y supermercados, a través de la banca se pueden establecer comportamientos de compra para ciertos tipos de productos.

Medicina: En este grupo se almacena la información referente a pacientes tales como enfermedades pasadas, tratamientos efectuados, pruebas realizadas, evolución de enfermedades y antecedentes familiares patógenos. Se pueden emplear técnicas de minería de datos con esta información, para identificar Asociación de síntomas y clasificación diferencial de patologías, Identificación de terapias médicas satisfactorias para diferentes enfermedades, estudio de factores (genéticos, precedentes, hábitos alimenticios,...) de riesgo para la salud en distintas patologías, estudios epidemiológicos, identificación de terapias médicas e identificación de tratamientos erróneos.

En el caso de la empresa comercial textil, al no estar previstas las necesidades de los clientes, la toma de decisiones se efectúa bajo un alto nivel de incertidumbre cuyo resultado puede ser la retención de grandes cantidades de inventario, pérdida de dinero y disminución de clientes potenciales [5]. En este escenario, la minería de datos puede ayudar a organizar las ventas, mejorar la eficiencia operacional, controlar los costos; resultando en un adecuado manejo de proveedores y aumento de la rentabilidad.

4. ETAPAS DE UN PROYECTO DE MINERÍA DE DATOS

4.1 Definición del Problema, La respuesta a una pregunta mal formulada queda comprometida desde el inicio del proceso. Uno de los mayores desafíos de los analistas de sistemas, es descubrir lo que el usuario realmente quiere. Además, el ambiente en grandes corporaciones, involucra el relacionamiento con diferentes comunidades dentro de la misma empresa. Adicionalmente, hay que tener en cuenta el ambiente de software y hardware de la empresa. Al considerar el proceso de definición del problema asociado al comportamiento de las ventas, resulta de interés la información suministrada por diferentes áreas de la empresa comercializadora textil:

- Entrevistas con la alta dirección de la empresa (usuario final)
- Entrevistas con el responsable de la administración de los datos en el área de sistemas de la empresa
- Reunión con el encargado del departamento de mercado
- Compilar la documentación

4.2 Adquisición y Evaluación de los Datos. Se parte del principio que los datos son la fuente predominante para la obtención de información, por lo tanto, esta etapa al igual que la anterior, constituyen los pilares del proceso de "Minería de Datos". Las tareas en esta fase del proyecto serían las siguientes:

- Adquirir datos
- Formatear datos
- Crear ambiente y herramientas
- Validar adquisición y formato
- Crear muestras (aleatorias) de trabajo
- Partición de los datos (análisis, calibración, validación y prueba)

En la empresa comercializadora textil, dependiendo de la complejidad de los sistemas informáticos, así como la experiencia y habilidad de las personas encargadas, estas actividades pueden resultar complejas y de un alto consumo de recursos, por lo que se debe destinar un tiempo adecuado para su realización.

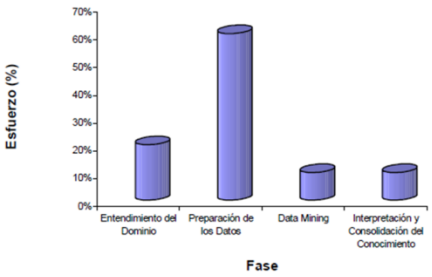


Figura 1. Esfuerzo requerido por cada fase del proceso de descubrimiento de conocimiento (KDD)

4.3 Extracción de características. Contribuyen en la solución del problema en discusión. Atributos (variables) que no se alteran, se pueden omitir dentro del análisis. De la misma forma, atributos fuertemente dependientes pueden ser reducidos. La meta de esta etapa es producir un conjunto de datos (*data set*) representativo, reproducible y confiable.

En la empresa comercializadora textil, aunque se cuenta con un gran número de registros, el número de variables independientes es reducido, lo que se debe considerar en el desarrollo del modelo.

4.4 Prototipo y desarrollo del modelo. Desarrollar el modelo de datos e implementar la solución para verificar el cumplimiento de las necesidades y objetivos planteados. Se realizan las siguientes actividades:

- Desarrollar hipótesis y plan de prueba
- Prototipaje
- Desarrollar modelos descriptivos y/o predictivos
- Evaluar el modelo de manera cuantitativa y cualitativa
- Despliegue del modelo (implementación, masificación y capacitación)
- Entregar el producto final

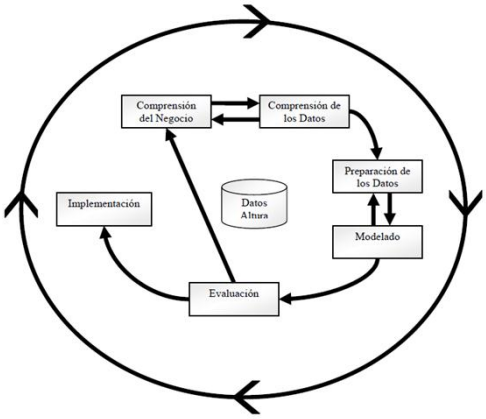
En principio, el objetivo es explorar varias técnicas de minería de datos, en especial aquellas que brinden claridad y facilidad de interpretación al modelo, como por ejemplo reglas de inducción y árboles de decisión. Esto no limita el uso de otras técnicas y el desarrollo de modelos asociados a referencias específicas; esto dependerá de un proceso de evaluación y ajustes iterativos sobre el modelo.

4.5 Evaluación del retorno sobre la inversión. Esta fase debe ser conducida por la gerencia de la empresa para evaluar si los cambios consecuentes al proyecto representaron efectivamente una ganancia material" [6].

5. Metodología CRISP-DM

La metodología CRISP-DM está descrita en términos de un modelo de proceso jerárquico Figura 2. Consiste en un conjunto de tareas desarrolladas en cuatro niveles de abstracción que van de lo general a lo específico: fase, tarea genérica, tarea especializada e instancia de procesos.

Figura 2. Fases de la metodología CRISP-DM



5.1 Comprensión del negocio. Consiste en un Análisis y definición de los objetivos y requerimientos desde una perspectiva no técnica.

El proceso de comprensión del negocio en la comercializadora textil, se realizó con un enfoque *top-down*, partiendo desde la alta gerencia, pasando por las funciones de mercadeo, hasta la exploración de los datos. Esto permite:

- Establecer los objetivos del negocio (Contexto inicial, objetivos, criterios de éxito).
- Evaluar la situación (Inventario de recursos, requerimientos, supuestos, terminología del negocio, etc).
- Establecer los objetivos de la minería de datos (objetivos y criterios de éxito).
- Generar el plan del proyecto (metodologías, herramientas y técnicas).

En la ronda de entrevistas, se contó con un formato de preguntas preestablecidas, evitando la improvisación frente al cliente y optimizando el uso de los recursos. Para esto fue necesario realizar un estudio previo de la información registrada en los manuales y procedimientos de la empresa.

5.2 Comprensión de los datos. Esta fase ayuda a familiarizarse con los datos teniendo presente los objetivos del negocio así:

- Recopilación inicial de datos
- Descripción de los datos
- Exploración de los datos
- Verificación de calidad de datos

En el análisis preliminar, se emplean hojas de cálculo que permitan apreciar de manera práctica el comportamiento de ventas de la empresa comercializadora textil, la cual se caracteriza por un elevado número de referencias, con ventas al por menor distribuidas en diferentes almacenes. Dicha situación, debe ser considerada durante la etapa de modelación a fin de ir refinando los resultados obtenidos.

5.3 Preparación de los datos. Es una de las fases más importantes para la elaboración del modelo de minería de datos. Se analiza toda la información para obtener un conjunto de datos con un formato adecuado.

En la empresa comercializadora textil, implica que se deben tener conocimientos y habilidades en el manejo de bases de datos, operaciones con registros y campos, herramientas de extracción y calidad de datos. Este conocimiento, permite realizar técnicamente las actividades descritas en la metodología:

- Selección de los datos
- Limpieza de datos
- Construcción de datos
- Integración de datos
- Formateo de datos

5.4 Modelado. Se aplican las diferentes técnicas de minería de datos a los *dataset* obtenidos en la preparación de los datos, buscando obtener un modelo inicial que permita alcanzar los objetivos planteados. En esta fase se tienen en cuenta:

- Selección de la técnica de modelado
- Diseño de la evaluación
- Construcción del modelo
- Evaluación del modelo

5.5 Evaluación. Serán analizados y evaluados los modelos generados en la fase anterior, para determinar si son útiles a las necesidades del negocio. Incluye las siguientes actividades:

- Evaluación de resultados
- Revisión del proceso
- Establecimiento de los pasos siguientes o acciones

5.6 Despliegue. Explotar la utilidad de los modelos, integrándolos en las tareas de toma de decisiones de la organización. Se contemplan las siguientes actividades:

- Planificación de despliegue
- Planificación de la monitorización y del mantenimiento
- Generación de informe final
- Revisión del proyecto. [7]

6. DIAGNÓSTICO DE CALIDAD DE DATOS

El diagnóstico de calidad de datos es una actividad del modelo de CRISP – DM, que abarca las fases de Comprensión de los datos y Preparación de los datos, ilustrando el nivel de dependencia entre ambas fases y el grado de esfuerzo requerido que garantice el entendimiento, la integridad y confiabilidad de datos, previo a la construcción del modelo; incrementando la confianza en los niveles de predicción del modelo. Es por este motivo que en esta parte de la metodología, se debe tener mucho cuidado y mantener una retroalimentación permanente con los interesados en la empresa, especialmente, con las personas responsables de la administración de los datos.

6.1 Buenas prácticas en la comprensión de los datos.

Al igual que en la comprensión del negocio, en la comprensión de los datos se busca entender el modelo Entidad-relación de la base de datos, de ser posible, analizar el diccionario de datos para hacer una valoración inicial de la información relevante y como procesarla. En caso de no tener acceso a la documentación de la base de datos, es recomendable identificar las inquietudes relacionadas con los datos y realizar reuniones con el encargado de informática para resolverlas. Se debe poner especial atención a la interpretación de los datos, presencia de datos extremos, campos vacíos, datos perdidos y valores específicos. Todo lo anterior, ayuda a orientar el objetivo del modelo y a visualizar posibles alternativas frente a las técnicas de minería de datos

6.2 Validación del estado del arte de los registros

Básicamente la realización del diagnóstico de calidad de datos se resume a una auditoría de los datos de ventas almacenados en la base de datos. Esta auditoría se enfoca en los datos y no en la administración y desempeño de la base de datos, la cual considera otros aspectos como tiempos de respuesta, tareas programadas, procedimientos almacenados, copias de respaldo y seguridad; actividades que se encuentran fuera del propósito fundamental de la minería de datos. Por este motivo, la validación se centra en los datos, incluyendo:

- Integridad de dominio: validaciones sobre el formato de los campos, valores nulos, valores atípicos, valores encontrados diferentes a lo establecido (o lo esperado). Por ejemplo, valores en cero y negativos para algunos campos, rangos de valores.

- Integridad referencial: pruebas sobre la relación entre campos y registros de una misma tabla. Por ejemplo, operaciones con campos e identificación de registros duplicados a través de claves primarias.

- Integridad relacional: restricciones de relación entre las tablas a través de sus claves foráneas y relaciones con tablas maestras.

El conjunto de pruebas se realiza detallando los resultados a nivel de tablas, campos y registros, generando la documentación de todas las validaciones (positivas y negativas) en un formato preestablecido que facilita el registro, análisis, trazabilidad, retroalimentación del usuario y la generación de acciones correctivas a nivel gerencial y operativo. El detalle correspondiente con las inconsistencias, se suministra usualmente como anexos en hojas de

cálculo o archivos planos. La Figura 3, ilustra el formato básico de calidad.

Figura 3. Formato de calidad de datos

[Nombre del campo]

Descripción de la Prueba	Valores esperados	Valores encontrados	Observaciones	Nivel de criticidad	Acción correctiva

Aunque no necesariamente corresponde a una inconsistencia, es importante identificar los campos con valores preestablecidos o sin variaciones, a fin de excluirlos en el análisis del modelo.

Se recomienda que luego de realizar esta validación se retroalimente al cliente para que conozca todo lo encontrado y así se programe en conjunto una limpieza de datos y establecer si es necesario aplicar algún filtro a la información. El mantener una comunicación abierta con el encargado del área de sistemas, en torno a los resultados de las pruebas, permite enfocar y resolver las inquietudes más fácilmente e identificar si se requiere información adicional de la empresa. El proyecto de minería de datos, es el resultado en un proceso iterativo de construcción colaborativa con la empresa.

7. CORRELACIÓN DE VARIABLES

El análisis permite identificar la existencia de una correlación lineal significativa estadísticamente, entre las variables independientes o variables explicativas del modelo y la variable dependiente o variable de interés.

Inicialmente, se hace un filtro de acuerdo a los resultados del diagnóstico de calidad de datos. Por ejemplo, se pueden descartar las columnas que presentan en todos sus registros el mismo valor y se filtran los campos o variables que presenten las siguientes inconsistencias en más del 50% de los registros:

- Valores en cero
- Valores nulos
- Valores negativos
- Registros vacíos

Posteriormente se lleva en una tabla el número original de registros, número de registros después del filtro y un porcentaje de datos empleado, lo mismo para el número de columnas.

En el proceso de correlación se define la variable de interés según el modelo a realizar. En este caso, se trata de un modelo de estimación de las ventas en una empresa comercializadora textil, por lo que la variable de interés es la cantidad solicitada de cada producto o referencia.

Para cada variable explicativa se realizan las siguientes actividades:

- Determinar el tipo de variable: categórica o numérica.
- Definir la escala: si es de razón, nominal, ordinal o intervalo.
- Identificar el gráfico y la prueba estadística apropiada de acuerdo al tipo y escala de la variable.
    - Se emplean usualmente gráficos de dispersión o cajas de bigotes
    - Se emplean las pruebas de correlación de Pearson, Spearman, y el análisis de varianza de una vía (Anova)

Pruebas estadísticas. Los gráficos nos permiten observar de una manera general la posible relación entre una variable independiente y la variable dependiente. Las pruebas estadísticas nos permiten determinar con un nivel de confianza establecido, la correlación (en este caso, lineal) entre dos variables.

En las pruebas de correlación y en general estadísticas, se introduce el concepto de Nivel de Significancia, que se define como la probabilidad de tomar la decisión de rechazar la hipótesis nula (existe una correlación significativa entre dos variables) cuando ésta es verdadera (decisión conocida como error de tipo I, o "falso positivo").

La decisión se toma a menudo utilizando el valor P (o p-valor): si el valor P es inferior al nivel de significancia, entonces la hipótesis nula es rechazada, aceptando la hipótesis alternativa (no existe una correlación significativa entre dos variables). Cuanto menor sea el valor P, más significativo será el resultado. Para el caso actual, se estableció un nivel de significancia del 5%.

El análisis de correlación se realiza con ayuda de software estadístico. Entre los más reconocidos del mercado está el SPSS y se puede utilizar la versión de prueba para llevar a cabo el análisis anteriormente mencionado.

8. CONCLUSIONES

EL concepto de minería de datos ofrece una visión a futuro que permite a la empresa comercializadora textil tomar decisiones estratégicas de mercado, gestionando mejor la demanda, ahorrando costos, maximizando utilidades y permitiéndole tener un constante flujo de mercancías.

El análisis de los datos para la elaboración de un modelo de minería es un proceso indispensable, en el cual se busca obtener una comprensión general del negocio a través de los datos y garantizar la calidad de la información insumo del modelo. El obviar algo en esta fase del proyecto, se traduce en un bajo nivel de confianza sobre los resultados y por ende se compromete la viabilidad del proyecto.

La evaluación del nivel de relación entre dos variables, desde el punto de vista de la modelación, es simplemente un análisis preliminar de las variables y que puede servir de referencia dentro del análisis del modelo.

Si al ejecutar la evaluación se encuentra que no existe correlación alguna, ello no indica que la variable independiente no se tendrá en cuenta para el desarrollo del modelo, debido a que la probabilidad o análisis de predicción lo determina finalmente la técnica de modelación.

9. RECOMENDACIONES

- Es importante desarrollar un proyecto de minería de datos, soportado en metodologías formales que orienten las actividades y permitan satisfacer las necesidades del negocio.

- Para la elaboración de un modelo de minería de datos es necesario contar con una fuente de datos consistente, sin ambigüedades en su contenido y lo más representativa de la realidad del negocio; de forma que se puedan obtener los resultados esperados.

- En los proyectos de Minería de datos se debe mantener una comunicación permanente con el cliente para retroalimentar los resultados del diagnóstico de calidad, efectuar los ajustes correspondientes y orientar los análisis.

10. REFERENCIAS

[1] Definición de Información, NET: http://definicion.de/informacion/

[2] Definición de Conocimiento, NET: http://definicion.de/conocimiento/

[3] G. Piatesky Shapiro y W.J. Frawley.Knowledge Discovery in Databases.Cambridge, MA:AAA/MIT Press, 1991.

[4] SAYAD Saed, An Introduction to Data Mining, NET: http://www.saedsayad.com/data_mining.htm. Abril 2012

[5] Berry, M. y Linoff, G., Data Mining Techniques for Marketing, Sales, and Customer Support. John Wiley, NY, 1997.

[6] Bello Peña D; Modelización de datos un enfoque práctico, primera edición, España, Editorial Lulu.com, marzo de 2009

[7] Oppel, A; Fundamentos de base de datos (Traductor MARTINEZ SARMIENTO Miguel Angel), primera edición en español, Mexico D.F, Editorial McGRAW-HILL INTERAMERICANA EDITORES S.A de C.V, 2009

C.V.: **Cristian Orlando Zapata Vásquez**; Tecnólogo de Sistemas, Estudiante de Ingeniería de Sistemas, Facultad de Ingenierías, Institución Universitaria de Envigado, Colombia.

C.V.: **Juan Sebastián Jaramillo**; Tecnólogo de Sistemas, Estudiante de Ingeniería de Sistemas, Facultad de Ingenierías, Institución Universitaria de Envigado, Colombia.