ISSN 1668-5792

Ideas innovadoras para una mejor práctica de negocios

Volumen V, Marzo de 2007

El Portafolio de Productos en el Mercado de Consumo Masivo - Parte II

Demián Siburi **Pág. 3**

Data Mining y Generación de Valor en la Firma

Ezequiel Sapoznik **Pág. 10**

Gestión por Competencias e Integración de Sordos en Empresas Privadas

Cristina Minolli **Pág. 14**

La Naturaleza de las Primas de Control Empresario ¿Opciones Reales?

Santiago Fidalgo y José Pablo Dapena **Pág. 24**

DATA MINING Y GENERACIÓN DE VALOR EN LA FIRMA

Por Ezequiel Sapoznik

INTRODUCIÓN

En forma permanente las empresas agregan millones de datos a sus bases de información. La disminución en los costos de almacenamiento y la mayor velocidad de procesamiento de las computadoras lo permiten. La capacidad de los discos se duplica año a año o inclusive en un lapso menor. El primer disco creado por IBM en la mitad de la década de los años cincuenta tenía una capacidad de almacenamiento de 5 megabyte y un costo de 35,000 dólares (o podía ser alquilado por 7000 dólares anuales por megabyte). Desde ese momento hubo una evolución no solo en la capacidad sino también en el tamaño. De acuerdo a diversos estudios, en el año 2010 una computadora hogareña va a contar con discos de 200 terabytes. Esto implica una importante cantidad de información.

"En forma permanente las empresas agregan millones de datos a sus bases de información……. El Data Mining automatiza la detección de patrones relevantes dentro de una base de datos."

Para hacer una comparación, el contenido completo de la Biblioteca del Congreso de los Estados Unidos, considerada una de las mayores del mundo, puede ser almacenado en aproximadamente 20 terabytes. Llevado al ámbito de las empresas, es sencillo darse cuenta de la cantidad enorme de información que cualquier industria puede tener sobre sus clientes y que puede no estar aprovechando. Justamente como esta información crece de manera exponencial, es clave la habilidad de acceder a la información que se necesita y en el momento que se requiere. Es posible generar consultas en SQL (Structured Query Language) y armar agregaciones multidimensionales para contestar preguntas como cuánto me ingreso por tal o cuál campaña de marketing ó cuantos clientes tiene cada sucursal de un banco ó cuantos clientes de un supermercado hicieron compras con su tarjeta de fidelidad.

Si bien esto es muy valioso para las empresas, estos análisis no permiten predecir la cancelación de un contrato en forma anticipada o la probabilidad de compra de un producto o cuándo un cliente va a cometer fraude en una tarjeta de crédito. Sin embargo es factible usar Minería de Datos (o Data Mining) para poder armar modelos que permitan contestar esas preguntas.

Existen muchas definiciones de Data Mining en innumerables publicaciones, sin embargo hay una que por su sencillez permite comprender la esencia de esta herramienta:

El Data Mining automatiza la detección de patrones relevantes dentro de una base de datos[1].

Ahora bien, vamos a ver un ejemplo de la vida real para entender de qué se trata el Data Mining.

Supongamos que usted va siempre a comer al mismo lugar al mediodía durante un año. Usted sabe que entre las 13 y 15 hrs. el lugar estará lleno. Sabe además que el restaurante sirve un plato cada día de la semana y el menú se repite todas las semanas. Si usted quiere almorzar tranquilo, sin hacer cola, en una mesa que usted pueda elegir y un plato diferente al del menú del día seguramente irá antes de las 13 horas (o después de las 15) o buscará otro lugar para variar la comida. Siendo obvio, este ejemplo sirve para entender como los datos pueden ser transformados en información para hechos accionables: se almacena información histórica que luego va a ser usada para estimar que pasará en el futuro.

El campo de aplicación del Data Mining es extenso. Se aplica tanto en empresas como así también en medicina o economía. Publicaciones de Harvard dieron a conocer recientemente que diversos centros de investigación médica, entre

[1] Berason *et. al*. 2000

los que se encuentra el Harvard Medical School y el Boston Medical Center, identificaron un método para predecir el riesgo de derrames cerebrales usando Data Mining integrando la información genética del paciente con su información clínica[2].

USO DEL DATAMINING EN EMPRESAS

El DataMining sirve para resolver ciertos problemas de negocios que pueden ser ejemplificados en la siguiente lista (no siendo la misma exhaustiva):

- ¿Cuáles son los clientes que tienen mayor probabilidad de irse a la competencia?

- ¿Cuál es la probabilidad de que un cliente cometa fraude?

- ¿Cuáles son los productos que tienen mayor afinidad en un supermercado?

- ¿Cuál es la publicidad que mejor se ajusta a determinado segmento de clientes?

- ¿Cómo me conviene segmentar mi cartera?

Existen diversas técnicas de Data Mining que pueden ser usadas en forma independiente o en forma conjunta para armar modelos predictivos que permitan contestar una pregunta de negocio. En cualquier proyecto de DataMining la definición del problema de negocio es el elemento clave para el éxito del modelo. A continuación se detallarán tres de las técnicas más comunes que pueden ser encontradas en el software comercial que se encuentran en el mercado y ejemplos de su aplicación de negocio.

Clustering

Segmentación es el término en español que más se acerca a la técnica de Clustering. Es usada para agrupar elementos de acuerdo a atributos en común que dichos elementos tengan y como no existe una variable que tenga que predecirse, no es necesario hacer distinción entre variables dependientes y variables independientes.

El Clustering resuelve problemas de clasificación de elementos. Su objetivo es distribuir casos que pueden ser personas, objetos, eventos, etc, en grupos donde el grado de asociación sea potente entre elementos del mismo grupo y débil entre elementos de diferentes clusters. El clustering es una herramienta de descubrimiento. Revela asociaciones y estructura de la información que no es evidente a simple vista, pero una vez descubierta es extremadamente útil. Los resultados de un análisis de clusters contribuye a la definición de un esquema formal de clasificación, como por ejemplo una taxonomía para clientes o productos en común. Sin importar en que industria usted se encuentre en algún momento va a necesitar resolver un problema de clasificación. Usar esta metodología es un excelente puntapié inicial para la aplicación futura de otras técnicas ya que permitiría trabajar con conjuntos homogéneos.

Ejemplo de aplicación de negocio:

1) Una entidad financiera que quiere realizar una segmentación de su cartera de clientes para determinar los clientes con mayor aporte de valor.

2) Una empresa de fidelización que quiere conocer características comunes de los clientes que viven en una zona geográfica para ofrecer premios acordes al perfil de cada grupo.

3) Una empresa de teléfonos celulares desea clasificar sus clientes en función de las variables de consumo, características técnicas de los aparatos celulares que tienen, la antigüedad en la empresa y la edad.

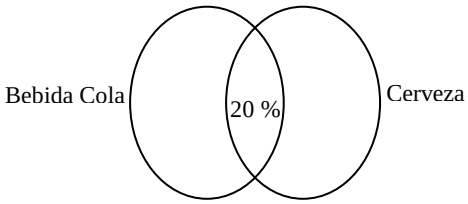
Asociación

Esta es una de las técnicas que más aplicación directa se le puede encontrar principalmente cuando se trata de industrias donde la oferta de productos es abundante. Está técnica aplicada ampliamente en supermercados y entidades financieras es también llamada análisis de canasta de productos. La base de esta técnica es descubrir relaciones entre los productos de tal forma que me permita identificar como fin último cuál es la

[2] http://www.researchmatters.harvard.edu/story.php?article_id=894

probabilidad de que un cliente adquiera un producto X dado que adquirió otro producto o un conjunto de productos. En esta metodología entran en juego 3 conceptos fundamentales que son:

1) Soporte: es la probabilidad que una transacción contenga los elementos X e Z u otros (por ejemplo que en todos los tickets de compra de un supermercado la participación de determinada bebida cola sea del 35% y que la probabilidad de que exista la bebida cola y una marca de cerveza sea del 20%). Esto se calcula normalmente como



2) Confianza (también llamado probabilidad en algunos textos): Es la probabilidad condicional de ocurrencia

$$Prob\,(B/A) - Soporte\,(A,B)\,/\,Soporte(A)\ .$$

Esto se lee como la probabilidad de que ocurra B dado que ocurrió A. En este caso A puede ser uno o más elementos (por ejemplo cual es la probabilidad de que se compre una tarjeta de crédito (B) dado que compró un préstamo y una cuenta corriente (ambos elementos forman A).

3) Lift (o Importancia): es el cociente entre la confianza y la confianza esperada. De esta manera este indicador demuestra en cuenta aumenta la probabilidad de la "consecuencia" si aplico el modelo de asociación.

¿Cómo se utilizan los tres conceptos?

Si la base de datos de un supermercado tiene 100,000 transacciones de las cuales 2,000 incluyen los elementos X y Z (que constituirían el elemento A) y 800 de las 2,000 incluyen al elemento B, la regla de asociación "Si X y Z son comprados entonces B es comprado en la misma operación" tiene un soporte de 800 transacciones (o 0.8% = 800/100,000) y una confianza de 40% (800/2000). Supongamos que la cantidad de compras de B en la base completa es de 5,000. Entonces la confianza ó probabilidad de B es 5% (5,000/100,000). De esta manera el lift aplicando el modelo es de 8 (40% / 5%). Esto significa que aplicando el modelo la probabilidad de compra de B aumenta en 8 veces por lo que al supermercado le conviene colocar los productos X y Z cerca del producto B para aumentar su probabilidad de compra.

Ejemplo de aplicación de negocio:

1) Un supermercado quiere conocer la probabilidad de compra de marca de cerveza dado que el cliente compró una determinada marca de pizza, con el objetivo de mejorar la organización de los productos en las góndolas y maximizar el beneficio

2) Una empresa de software quiere conocer que producto conviene ofrecer a una empresa sabiendo qué producto adquirió previamente.

3) Una empresa que tiene una cadena de videoclub y quiere enviar piezas de mailing personalizadas con ofertas de las películas que más probabilidad de compra tiene cada cliente.

Regresión

La regresión es una técnica que permite determinar la probabilidad de un suceso (variable dependiente) de acuerdo a los valores de una o más variables que actúan como predictoras. Es una de las técnicas más antiguas y más usada incluso antes de la existencia de computadoras que permitieran el desarrollo de las técnicas de Data Mining. La regresión es usada principalmente para establecer la probabilidad de un suceso (variable predictiva) dada la ocurrencia de una o más variables predictoras. En este sentido siempre es importante realizar un análisis de correlación de las variables para determinar el grado de relación de cada una de las variables predictoras con la variable a predecir. De esta manera un modelo puede iniciarse con más de cien variables y finalmente queda determinado por quince ya que las restantes ochenta y cinco no tienen una fuerte relación con la variable a predecir.

Normalmente son utilizadas dos formas de regresión que son el algoritmo de Regresión Lineal y el de Regresión Logística. La regresión lineal se expresa de la siguiente manera:

$$y = b_1x_1 + b_2x_2 + \ldots + b_nx_n + c$$

donde y es la variable a predecir (variable dependiente), b son los coeficientes de regresión de las correspondientes x (variables independientes) y c es la constante.

Por ejemplo:

Probabilidad de fraude (y) = 0.98 por días sin pagar + 0.56 por gastos realizados + 0.35 por cantidad de transacciones + 1.4

Esta ecuación nos dice que la probabilidad de fraude en por ejemplo una tarjeta de crédito esta determinada de manera positiva (ya que los coeficientes son positivos) por los días sin pagar (mientras mas días sin pagar mas probabilidad), por los gastos realizados por el cliente o consumidor (mayor nivel de gastos mayor probabilidad de fraude) y por la cantidad de transacciones realizadas (mientras mayor sea mayor es la probabilidad de fraude).

La regresión logística involucra en su cálculo la utilización de elementos de logaritmos, pero con igual objetivo que el anterior.

Ejemplo de aplicación de negocio:

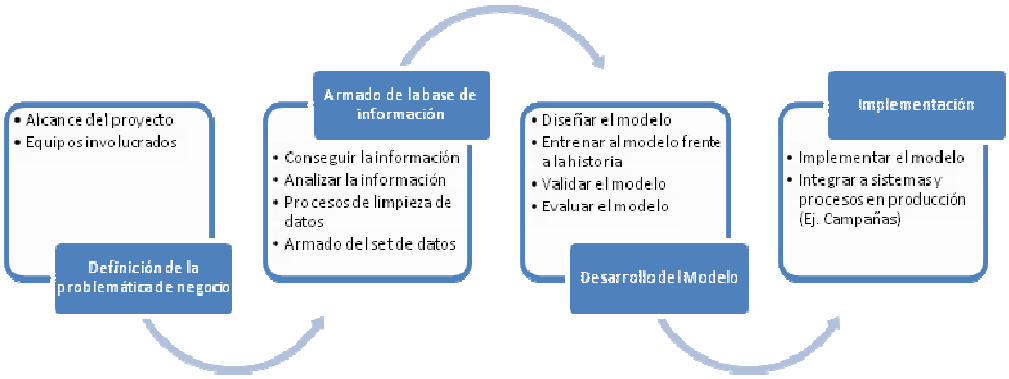
1) Una empresa que brinda servicios de internet desea predecir cuales son los clientes que mayor probabilidad de abandonar la empresa para realizar acciones preventivas.

2) Un supermercado quiere medir la fidelidad de los clientes y determinar cuál es la probabilidad de que un cliente compre una determinada canasta de productos.

3) Un banco quiere conocer cuál es el ciclo de vida de un cliente.

PROCESO DE IMPLEMENTACIÓN DE UN PROYECTO DE DATA MINING

En el siguiente gráfico se ejemplifica como es la metodología de implementación de un Proyecto de Data Mining



Vemos que los pasos en el proceso de implementación son cuatro:

- Definición de la problemática de negocios
- Armado de la base de datos o información
- Desarrollo del modelo
- Implementación

BENEFICIOS DEL DATA MINING

Al contrario de muchos proyectos cuyos objetivos son maximizar la eficiencia comercial de una empresa, los proyectos de Data Mining son de corto plazo con alto impacto. Asumiendo que la base de datos con la que cuenta la empresa tiene una buena calidad, un proyecto de Data Mining no puede extenderse más de dos meses con implementación inmediata. Adicionalmente es importante destacar que normalmente sus costos son bajos en relación a otras implementaciones.

Los resultados tienen visibilidad prácticamente en

forma inmediata. Una vez desarrollado el modelo, se puede aplicar a la realidad (por ejemplo en una acción de Marketing) en un grupo de clientes comparando luego contra otro grupo donde no se aplicó el modelo para determinar el aumento real de la tasa de éxito con el uso del Data Mining a través de la comparación entre las dos.

CONSEJOS PARA LA IMPLEMENTACIÓN DE PROYECTOS DE DATA MINING

1) Si es su primer proyecto, es recomendable contratar consultoría externa. De esta manera podrá incorporar mucho conocimiento de personas que ya lo tienen en un plazo muy corto.

"….es importante destacar que normalmente sus costos son bajos en relación a otras implementaciones. Los resultados tienen visibilidad prácticamente en forma inmediata."

2) Las personas involucradas en el negocio deben participar. El proyecto de Data Mining no sirve si sólo participan los expertos en estadística o sólo el área de sistemas.

3) Sea concreto en el problema de negocio. Una pregunta concreta a resolver (¿Cuál es la probabilidad de que un cliente compre el producto X?) permite armar modelos concretos y más enfocados.

4) No quiera "reinventar la rueda". Ajústese a la problemática de negocio definida en la fase inicial del proyecto.

5) Su base de datos es la materia prima en este tipo de proyectos. Conocerla y realizar trabajos permanentes de calidad de datos (Data Quality) es condición casi imprescindible para obtener buenos resultados. Recuerde que la degradación de una base de datos es constante (clientes que se mudan, que cambian de estado civil, o que tienen número de celular nuevo).

6) No todos los receptores del resultado del modelo son estadísticos. Busque ejemplos de negocio aplicado al modelo para presentarlo. De esta manera logrará mejores resultados y apoyo para proyectos futuros.

7) Actualice el modelo. Estos modelos interactúan con seres humanos quienes tienen comportamientos que cambian con el tiempo y por ende tienden a descalibrarse cada X cantidad de meses, donde X va a depender de las características de cada modelo, pero usualmente una revisión semestral será de mucha utilidad.

CONCLUSIÓN

El uso del Data Mining no implica certeza. Se basa en técnicas estadísticas para identificar patrones de comportamiento, similitudes entre conjuntos y probabilidades de ocurrencia. Estas técnicas no buscan predecir el futuro, sino que brindan soporte cuantitativo al proceso de toma de decisiones al otorgar una base de rigurosidad al mismo, ayudando a maximizar la eficiencia de cualquier industria bajando costos de Marketing, direccionando mejor los recursos y obteniendo resultados medibles a corto plazo. Constituye una nueva y poderosa metodología de análisis que puede ayudar a las empresas a maximizar el uso de su información.

No sea temeroso de implementar estos modelos: No es estadística, es negocio.

Bibliografía

- Berson A., Smith S. y Thearling K. *Building Data Mining Applications for CRM*. Ed. McGrawHill, 2000.
- Tang Z. y MacLennan J. *Data Mining with SQL Serve.* Ed. Wiley, 2005.
- Website "Research matters at Harvard University": http://www.researchmatters.harvard.edu.