resumen

En este artículo se presenta la construcción de un modelo de pronóstico "de mejor siguiente oferta", desarrollado por medio de la utilización de técnicas de minería de datos, con la información contenida en la bodega de datos de una entidad financiera colombiana.

Un modelo de mejor siguiente oferta

Octavio A. Echeverri B. • José Abásolo Prieto.

introducción

En el contexto actual del sistema financiero colombiano, las fusiones y adquisiciones se han constituido en la mejor opción para ganar de forma rápida participación de mercado, así como en un mecanismo ideal para prepararse, en el caso de la banca nacional, a los desafíos que impondría un potencial tratado de libre comercio con los Estados Unidos. Sin embargo, para materializar tales bondades, la entidad adquiriente o fusionada debe asegurarse que explotará efectivamente su nueva base de clientes y las sinergias resultantes de la unión entre las empresas [5].

En el caso que se presenta, una entidad financiera, luego de comprar una compañía competidora, decidió implementar como uno de sus planes de bienvenida a los nuevos clientes, provenientes de la entidad adquirida, una oferta personalizada de productos propios, de acuerdo con los intereses y necesidades particulares de cada uno de ellos. Así mismo, dependiendo del éxito de la puesta en marcha de dicha campaña, se contempla la opción de extender la oferta personalizada al resto de clientes del banco.

Dado el anterior objetivo, se optó, como proyecto de corto y mediano plazo, crear un modelo de pronóstico por producto, para un conjunto de productos elegidos que, dependiendo de la historia de clientes en general que ya lo tienen y del comportamiento reciente de un cliente en particular, y de los productos que este posee hasta la fecha con el banco, permitiera predecir cuál de ellos despertaría más interés para

su adquisición. Los modelos que cumplen la función que se acaba de describir se denominan en mercadeo "de mejor siguiente oferta". Bajo esta definición, las técnicas tradicionales de minería de datos se presentaron como una alternativa ideal para la construcción de los modelos.

El objetivo de corto plazo, incluido en el alcance del proyecto de grado, fue desarrollar el modelo para uno de los productos más rentables del banco: las tarjetas de crédito para una de las franquicias más populares[1]. En el mediano plazo se planteó realizar los modelos para los productos restantes. Como metodología se utilizó la propuesta por Berry y Linoff [1]; el proceso planteado por estos autores es cíclico, toda vez que, cada paso o cada conjunto pequeño de pasos puede servir como retroalimentación para refinar pasos anteriores, como se muestra en la figura 1. En los siguientes acápites de este documento se destaca cómo fue estructurado el trabajo siguiendo cada una de estas etapas, comenzando por la traducción del problema de negocio a uno de minería hasta la construcción y la evaluación preliminar del modelo inicial. Sin embargo, para adaptar la metodología aplicada a las necesidades y restricciones particulares del proyecto, algunos pasos fueron realizados o bien en un orden diferente al sugerido o llevados a cabo de manera concomitante.

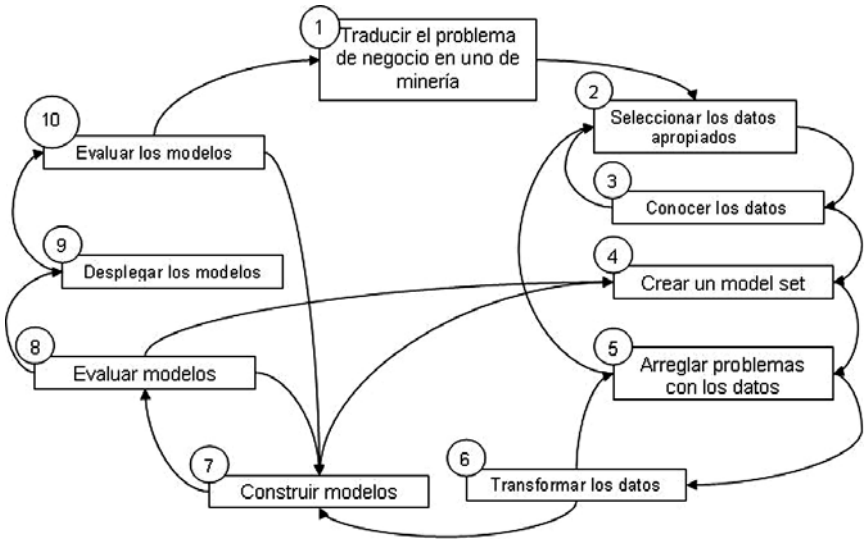

Figura 1. Ciclo metodológico de Berry y Linoff para un proyecto de minería de datos

Desarrollo del modelo

En los siguientes párrafos se muestra cómo se adelantaron los diferentes pasos de la metodología de referencia del proyecto dentro de la entidad financiera, así como el modelo inicial y los resultados obtenidos con el prototipo construido para uno de los productos ofrecidos por el Banco.

Siguiendo los parámetros de la metodología de referencia, en la traducción del problema de negocio al problema de minería de datos, se decidió construir un modelo de pronóstico que calculara las probabilidades de que un cliente adquiriera cada uno de los productos del Banco y seleccionara, al mismo tiempo, como producto "de la mejor siguiente oferta", aquel cuya probabilidad de adquisición sea la mayor. Adicionalmente, si un cliente ya tiene el producto el modelo deberá calcular una probabilidad de 0 para dicho producto, puesto que el interés del Banco en este momento es generar ventas cruzadas que aumenten la fidelidad del cliente con la organización a través del efecto de *lock-in*[2] que se produce cuando este posee un número significativo de productos financieros con la misma entidad.

En vista de que el número de productos en el Banco es muy alto, más de 90 para los orientados a las personas naturales, la primera tarea consistió en determinar en qué productos resultaría de interés calcular la probabilidad de adquisición por parte de cada uno de los clientes. Para esto se revisó la codificación de productos manejada en los sistemas de información del Banco. La codificación es jerárquica, dada por el código de empresa (del banco original, y de cada uno de los bancos que éste ha adquirido con el tiempo), el código de producto y el código de subproducto.

En el nivel de codificación más bajo subproducto, el Banco maneja 179 subproductos diferentes; de estos se seleccionaron, exclusivamente, aquellos orientados a las personas naturales, lo que implica que categorías como crédito corporativo y crédito a los constructores y subcategorías como tarjetas de crédito empresariales fueron eliminadas del análisis. Hecha esta primera clasificación, resultaron 93 subproductos que fueron considerados en el ejercicio de minería.

A aquellos subproductos considerados como diferentes al interior del Banco, pero cuyas diferencias radican fundamentalmente en aspectos de mercadeo (por segmentación, para poner un ejemplo) y no de funcionalidad, se les asignó un número igual de codificación para que sirviera como identificación más adelante en el ejercicio de minería. Así, todas las opciones de créditos de vivienda, en su conjunto, se agruparon dentro de un solo

identificador, con código 01. Utilizado este esquema, las agrupaciones resultantes fueron:

Código	Línea de negocio
01	Crédito de vivienda
02	Tarjeta de Crédito Franquicia 1
03	Tarjeta de Crédito Franquicia 2
04	Crédito de Consumo Rotativo
05	Crédito de Consumo Fijo
06	Cuenta de Ahorros
07	Cuenta Corriente
08	Certificados de Depósito
09	Fondos
10	Títulos
11	Fondo especial

Tabla 1. Codificación de productos para el ejercicio de minería

Para construir un modelo que prediga la probabilidad de que un prospecto responda positivamente a una oferta, se necesitan ejemplos del pasado, tanto de prospectos que adquirieron el producto como de los que no. La proporción de unos y otros debe ser balanceada y cubrir una ventana de tiempo amplia que tenga en cuenta efectos de estacionalidad. En el caso del proyecto que nos ocupa, dichos casos se toman de la Bodega de Datos del Banco adquiriente, con datos de sus clientes, para el producto seleccionado: para un periodo de tiempo determinado, se mira qué clientes no tenían el producto en cuestión, y se arma su "foto" en ese momento; luego, se mira cuáles adquirieron el producto dos meses después y cuáles no. Así, si el mes de octubre de 2006 hace parte del periodo seleccionado, se miran los clientes que en ese mes no tenían el producto, y de ésos cuáles lo adquirieron en diciembre de 2006. Lo mismo se hace para los otros meses del período seleccionado. La diferencia de dos meses se usa para tener en cuenta la latencia, es decir, la diferencia de tiempo entre la información disponible más reciente y el periodo para el cual se quiere hacer la predicción.

Una de las tareas más difíciles en la construcción de un modelo de predicción es determinar las variables predictoras. Para ello se necesita del conocimiento y experiencia de especialistas del negocio. También implica un trabajo iterativo de ensayo con diferentes conjuntos de variables, hasta encontrar un modelo satisfactorio. Con este propósito, se adelantó una serie de entrevistas con los gerentes de las diferentes líneas de negocio del Banco. Como resultado de este ejercicio, se obtuvieron dos conjuntos de variables: uno de 86 variables para medir el comporta-

miento de los clientes en todas las líneas de negocio y un segundo de 8 variables demográficas que sobresalieron en varias entrevistas.

Una vez se contó con la definición de las variables, se dio paso a la tarea de conocimiento de los datos, consignados en la bodega de datos del Banco, que se alimenta, fundamentalmente, de la información registrada en archivos VSAM[3] utilizados por los sistemas transaccionales ES/390 y AS/400. El cargue de la información de estos archivos a la bodega se lleva a cabo por medio de un proceso en lote nocturno mensual. En la bodega no existen reglas de validación de integridad referencial, se utilizan implícitamente las del sistema fuente. Así mismo, como la bodega importa directamente la información desde los archivos VSAM, no cuenta con llaves primarias, ni foráneas, como tampoco restricciones (*constraints*); pero si con algunos índices.

Para determinar qué datos almacenados serían de utilidad para el trabajo restante, se realizó un diagnóstico del estado de la información consignada en la bodega, luego del cual se descartaron los registros cuya calidad de la información, para los campos requeridos, era deficiente. Finalizado este paso, se concluyó que de 2.836.919 clientes con productos vigentes en el banco al 31 de Diciembre de 2.006, solo 911.194 contaban con la información en las condiciones necesarias en cuanto a la calidad de sus datos demográficos. Estos últimos fueron los que se incluyeron en las siguientes actividades del proyecto.

Por otra parte, durante la revisión se detectó que no todas las variables seleccionadas luego de las entrevistas con las líneas de negocio se encontraban como un campo de una tabla dentro de la bodega, pero podían ser calculadas a partir de los campos existentes, por lo que se decidió realizar las transformaciones sobre los datos en el momento en el que se construyó el *model set*[4] para la elaboración del modelo.

Para el proyecto de corto plazo, y por razones de restricción de acceso a los datos, se tomó como período de tiempo un mes, a sabiendas de que la ventana no era suficientemente representativa. Se especificó que la variable de respuesta sería una que tomaría valores discretos de cero o uno, a fin de señalar la adquisición (1) o no adquisición (0) de una tarjeta de crédito, dos meses después de la fecha de corte utilizada para las demás variables.

Como técnica de minería, por su habilidad para manejar muchas variables durante la fase exploratoria y por la claridad de entendimiento de los resultados (reglas de clasificación), se escogió un algoritmo que implanta *Árboles de Decisión*.

El árbol obtenido, buscando aumentar la pureza[5] en cada una de sus hojas, concentra en una sola rama buena parte de los datos, debido a la baja densidad de observaciones positivas en la muestra para la variable de respuesta. Las reglas resultantes señalan como propensos a la adquisición de una tarjeta de crédito a aquellos que cumplan alguna de las condiciones que se presentan a continuación:

1. Si su nivel de ingresos es inferior a un millón de pesos, pero tienen una cuenta corriente sin chequera y una tarjeta de crédito de otra franquicia cuya utilización no sea superior al 35%, la probabilidad calculada es de 82%.
2. Si su nivel de ingresos es superior a un millón de pesos, y poseen los mismos productos que en la primera regla, sin importar la utilización de las tarjetas de la otra franquicia, la probabilidad calculada es de 71%.
3. Si no poseen cuenta corriente y tienen un tiempo de vinculación con el banco menor a 3.5 años, con un nivel de ingresos superior a un salario mínimo y medio, no son solteros y tienen experiencia crediticia con al menos un crédito de consumo, la probabilidad calculada es igual a 81%.

Conclusiones y trabajo futuro

En este capítulo se cierra la exposición destacando algunos puntos luego del desarrollo del proyecto y señalando las tareas pendientes junto con las recomendaciones para el trabajo restante.

Como principales logros del trabajo desarrollado se destacan:

- Haber vendido al área comercial del Banco la idea de utilizar técnicas de minería de datos como parte de la estrategia de ventas cruzadas.
- Haber desarrollado un prototipo, cuyo principal aporte se encuentra en la definición de "productos" para generar modelos, identificación de variables predictoras, análisis de fuentes para verificar calidad y completitud de datos, y definición de "casos" del *Model Set*.

Se observa que aún cuando el modelo obtenido se comporta aceptablemente bien tanto para el subconjunto de entrenamiento del *Model Set* como para el de prueba, y que las reglas generadas son coherentes con la experiencia al interior del Banco, el hecho de que hayan sido construidas con un número reducido de observaciones del evento positivo de la variable de pronóstico, obliga a afirmar que estas reglas no pueden ser consideradas aún para su utilización directa en el mercado. Como paso previo, se deberá construir modelos

con un número mayor de meses considerados, por lo menos 6, para capturar parte del efecto cíclico anual del mercado y de la economía en general.

Al momento de completar y desplegar el modelo propuesto en el escenario del Banco, es importante tener en cuenta los supuestos que se realizaron en este trabajo, los cuales no se pueden conservar si se desea obtener resultados que no presenten sesgos o sobreadaptación (*over-fitting*) a los datos con los que se construyó el modelo. El principal de tales supuestos es la ausencia de ciclos en el comportamiento de adquisición de productos por parte de los clientes. Se puede pensar que si bien el ciclo de consumo de ciertos productos es superior a un año y está fuertemente relacionado con variables exógenas a las consideradas en el estudio, la inclusión de información de varios meses logrará suavizar el efecto de sobreadaptación a un conjunto de datos de un mes en particular.

Al extender el modelo para que calcule la probabilidad de adquisición de más de un tipo de producto, será necesario cuantificar el grado de canibalización[6] entre productos, especialmente los de inversión. Para esto es recomendable calcular la probabilidad condicional de adquirir cada uno de los productos dado que se ha comprado el anterior. Este tipo de análisis no hizo parte del alcance del proyecto.

Como la proporción entre valores positivos y negativos, dentro de la variable de respuesta, estará con seguridad desequilibrada en una relación cercana a 1:100, es importante aplicar alguna estrategia que permita mitigar los efectos negativos de esta diferencia en los algoritmos de árboles de decisión. Por ejemplo, asignar pesos a los casos del conjunto de entrenamiento, para incrementar la importancia de los positivos, recalculando posteriormente la probabilidad ó propensión a comprar, usando datos sin pesos.

Finalmente, pensamos que, dadas las características del banco en las que se adelantó este trabajo, orientado a la banca hipotecaria y de tarjetas y con unas políticas para aprobación de crédito consideradas como conservadoras en el mercado, el modelo obtenido no es generalizable para otras entidades del sector con un perfil diferente.

Referencias

[1] Berry, M.J.A. y Linoff, G.S., Mastering Data Mining, Wiley: New York, 2000.

[2] Berry, M.J.A. y Linoff, G.S., Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. Segunda Edición, Wiley: New York, 2002.

[3] Cabena, P., Choi, H., Kim, I., Otsuka, S., Reinschmidt, J. y Saarenvirta, G. Intelligent Miner for Data: Application's Guide. IBM Redbooks. 1999.

[4] Chidanand A., Weiss S. Data Mining With Decision Trees and Decision Rules. Future Generation Computer Systems. Noviembre de 1.997.

[5] Kollem T., Goedhart, M., Wessels, D. Valuation: Measuring and Managing the Value of Companies. Wiley. 2003.

[6] Lovelace, M., Lovelace, D., Rama, A., Sala. A. y Sokal. V. VSAM Demystified. IBM Redbooks. 2004.

[7] Shapiro, C y and. Varian H. R. Information rules: A Strategic Guide to the Network Economy. Harvard Business School Press, 1998.

[8] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

Notas de pie de página

1 Las franquicias de tarjeta de crédito existentes en el país son Visa, Master Card, American Express y Diners Club. Las últimas 2 son ofrecidas de forma exclusiva por bancos diferentes.

2 El efecto de lock-in, como lo describen Shapiro y Varian en [6], consiste en la dependencia que crea un consumidor con un proveedor de productos o servicios, al no poder cambiarlo por otro proveedor sin tener que incurrir en costos muy altos directamente asociados a este cambio.

3 Método virtual de acceso secuencial. Esta es una solución de almacenamiento de archivos, introducida por IBM para el manejo de información persistente bajo sus sistemas operativos.

4 *Model set*, en la teoría de minería de datos, consiste en el conjunto de datos utilizados para la construcción de un modelo.

5 La pureza, en este contexto, tiene que ver con el número de eventos positivos y negativos de la variable de respuesta en una misma hoja: Si la hoja del árbol está compuesta exclusivamente por registros donde la variable de respuesta toma un mismo valor, entonces la pureza es alta, si en los registros se observa un número similar de eventos positivos y de negativos, la pureza es baja.

6 Canibalización, en el ámbito de mercadeo, está relacionado a la reducción en el volumen de ventas o participación de mercado por el efecto de ofrecer otro producto del mismo productor.

autores

Octavio Echeverri Botero. Ingeniero Industrial, Universidad de los Andes. 2006. Ingeniería de Sistemas y Computación. Universidad de los Andes. 2007. Vinculado desde el año 2004 con el sector financiero. o-echeve@uniandes.edu.co

José Abásolo Prieto. Ingeniero de Sistemas y Computación. Profesor Universidad de los Andes.