La Minería de Datos en la Industria Financiera: Un Nuevo Enfoque de Investigación de Mercados

Manuel Barberena y Viterbo Berberena
Pearson

Introducción

Como escribimos hace algún tiempo en esta revista, la minería de datos se percibe más como ciencia ficción que como una realidad, y peor aún, se piensa que estas metodologías y herramientas de investigación de mercados en bases de datos no están relacionadas con la actividad de las agencias en la actualidad. Los que tienen este punto de vista están renunciando a una gran pérdida de oportunidad, que lamentarán en el futuro. La tendencia en este siglo es que la investigación de mercados en bases de datos irá desplazando a la investigación de mercados clásica. Sí es bueno señalar que nunca la sustituirá completamente, pero le robará terreno.

Instituciones universitarias de gran prestigio como el ITESM (Instituto Tecnológico y de Estudios Superiores de Monterrey) y el ITAM (Instituto Tecnológico Autónomo de México) ya están preparando a sus egresados de la carrera de mercadotecnia y de los cursos de educación continua (maestrías y diplomados) en estas novedosas técnicas de investigación de mercados. Agencias de investigación nacionales como Pearson y otras, también se han sumado a este empeño y están desarrollando áreas de investigación, modestamente, contra todas las dificultades que presenta este tipo de proyectos en México. En este esfuerzo está participando, de forma destacada, el SAS Institute de México proporcionando oportunamente herramientas de muy alta calidad.

En esta ocasión, queremos presentar la aplicación de la minería de datos a la industria financiera, en el caso específico de un trabajo típico de *credit scoring*. Trataremos de presentar en forma breve el potencial de la minería de datos y además demostrar la viabilidad de este tipo de proyecto en México, lo que resultaría en una disminución de los costos para las empresas que contratan estos servicios a transnacionales extranjeras.

Comenzamos con lo que llamamos en el argot financiero *credit scoring*. Este no es más que el proceso de cálculo de puntajes de riesgo a las solicitudes de crédito o a cuentas ya existentes y tiene tres aplicaciones fundamentales:

a) *Application scoring* (para la aceptación o rechazo de las solicitudes de crédito).

(b) *Behavioral scoring* (para predecir la probabilidad de incumplimiento de los clientes que ya han sido aceptados).

(c) *Collection scoring* (para estimar el monto probable de deuda que el prestador puede esperar recuperar).

En *application scoring* el *credit scoring* se usa para optimizar la tasa de aprobación de las solicitudes. Permite a las organizaciones elegir un punto de corte óptimo de aceptación, de tal forma que se gana participación en el mercado, mientras se mantiene la máxima rentabilidad. Los puntajes obtenidos en los clientes y los prospectos son esenciales para la personalización de los productos de crédito.

El *behavorial scoring* de los clientes existentes se usa para la detección temprana de cuentas de alto riesgo y permite a las organizaciones realizar acciones enfocadas a estos objetivos (por ejemplo, la reestructuración de la deuda). También forma las bases para cálculos más precisos del riesgo crediticio de todos los consumidores.

En el *collection score*, los recursos utilizados en el cobro de las deudas se pueden optimizar enfocando las actividades a objetivos concretos de recaudación (con alto puntaje de recaudación). También se usa para determinar el valor preciso del libro de deudas antes que éste sea traspasado a una empresa recaudadora.

En algunas situaciones es apropiado comprar, a proveedores externos, modelos de crédito genéricos, listos para usar, o tener modelos de créditos para propósitos específicos desarrollados por consultores externos. Sin embargo, mantener una práctica de

construcción de modelos *in-house* ofrece ciertas ventajas:

(a) Ganancias a partir de economías de escala cuando se necesita construir muchos modelos específicos para una gran cantidad de segmentos.

(b) Consolidar una base de datos flexible y reutilizable, generar conocimientos y habilidades por sí mismo, de forma que le sea fácil a la organización ser consistente en la interpretación de los resultados de los modelos y los reportes y en la propia metodología de modelación.

(c) Verificar la precisión y analizar las fortalezas y debilidades de los modelos de crédito adquiridos.

(d) Reducir el acceso de extraños a información estratégica y retener las ventajas competitivas con la creación de las mejores prácticas de la compañía.

Para cada modelo en particular es importante evaluar su capacidad predictiva, o sea, la precisión de los puntajes que éste otorga a las solicitudes y las consecuencias de las decisiones de rechazar/aceptar que sugiere. Se usan una variedad de medidas de calidad relevantes para el negocio como:

(a) Las curvas de concentración.

(b La curva de estrategia.

(c) Las curvas de ganancia.

El mejor modelo se determinará por el propósito para el que se usará el mismo y la estructura del conjunto de datos con el que se validará.

El objetivo de este trabajo es la obtención de un modelo del sistema de crédito al consumidor en la industria financiera, mediante el uso de la minería de datos como herramienta metodológica y de cálculo, para la obtención de altos niveles de precisión y confiabilidad. Se realiza la aplicación del modelo en el caso de una Institución Financiera Mexicana.

Metodología

La metodología del proceso de modelación del sistema de crédito (*credit scoring*) con la herramienta de minería de datos consta de las siguientes etapas fundamentales:

(a) La carga de la base de datos.

(b) El agrupamiento interactivo.

(c) El ajuste del modelo de regresión logística.

(d) El cálculo de los puntajes para cada atributo.

(e) El análisis de los puntajes.

(f) La inferencia de los rechazos.

(g) La modelación en la base completa.

La explicación de la metodología propuesta se realiza sobre la base de su aplicación en el caso de una organización financiera nacional, que por motivos de discreción en lo sucesivo llamaremos Grupo Financiero Mexicano (GFM).

Preparación de la base de datos.

En esta fase inicial se procede a un estudio detallado de las funciones de distribución de cada una de las variables y al tamizado inicial de las mismas, con el propósito de disminuir en lo posible la magnitud del problema. Se creó la variable objetivo a partir de “semanas de atraso en los pagos”. Se cargaron los datos para el espacio de trabajo de la herramienta de minería de datos y que en este caso se utilizó el *Enterprise Miner de SAS*. En la gráfica No. 1 se presenta el diagrama de flujo del modelo.

El agrupamiento interactivo.

El objetivo de esta etapa es la clasificación, que no es más que el proceso automático y/o interactivo de redistribución y agrupamiento de las variables de intervalos, ordinales y nominales con el propósito de:

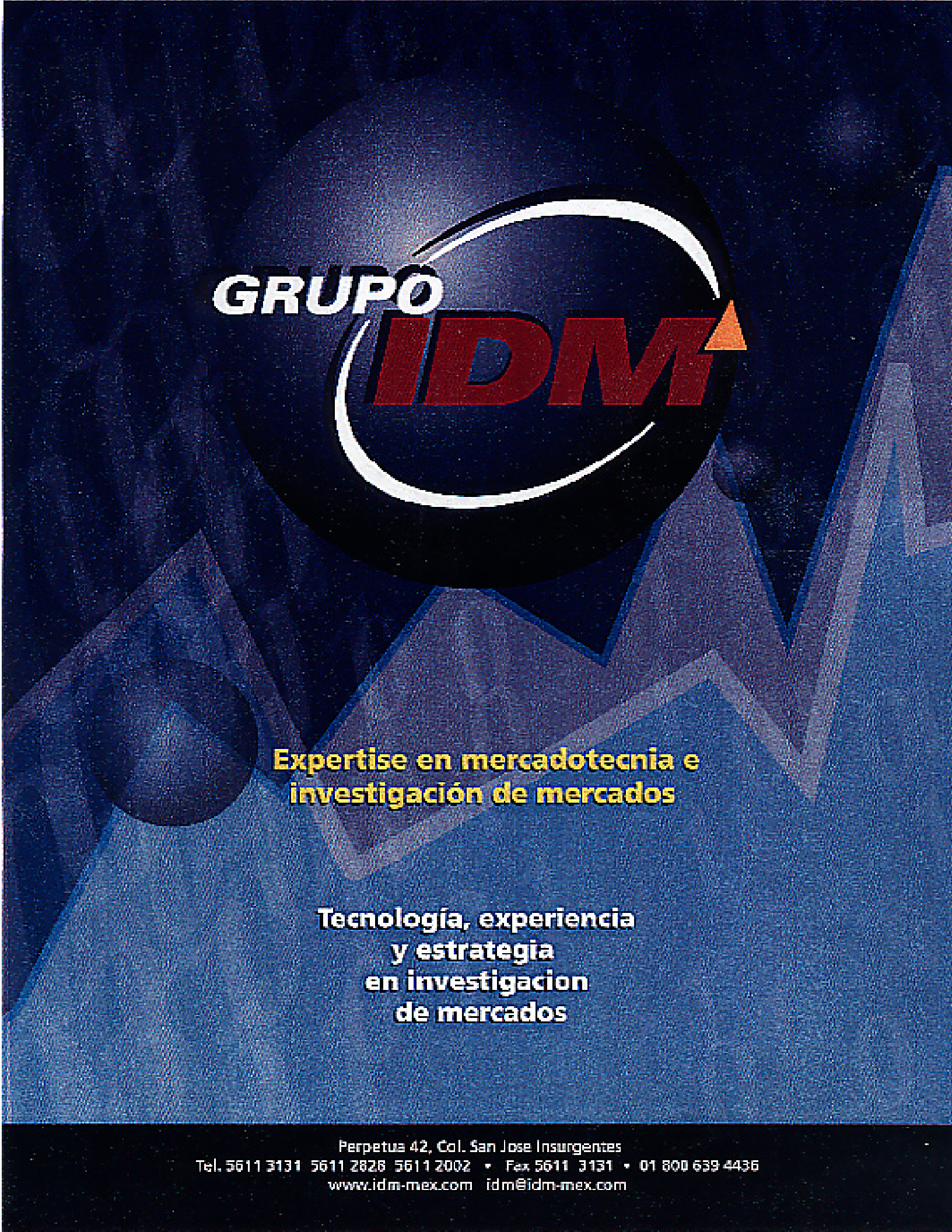
(a) Manejar el número de atributos (nuevos niveles de la variable) por característica (la variable).

(b) Mejorar el poder predictivo de la característica.

(c) Seleccionar los predictores (características con poder de predicción).

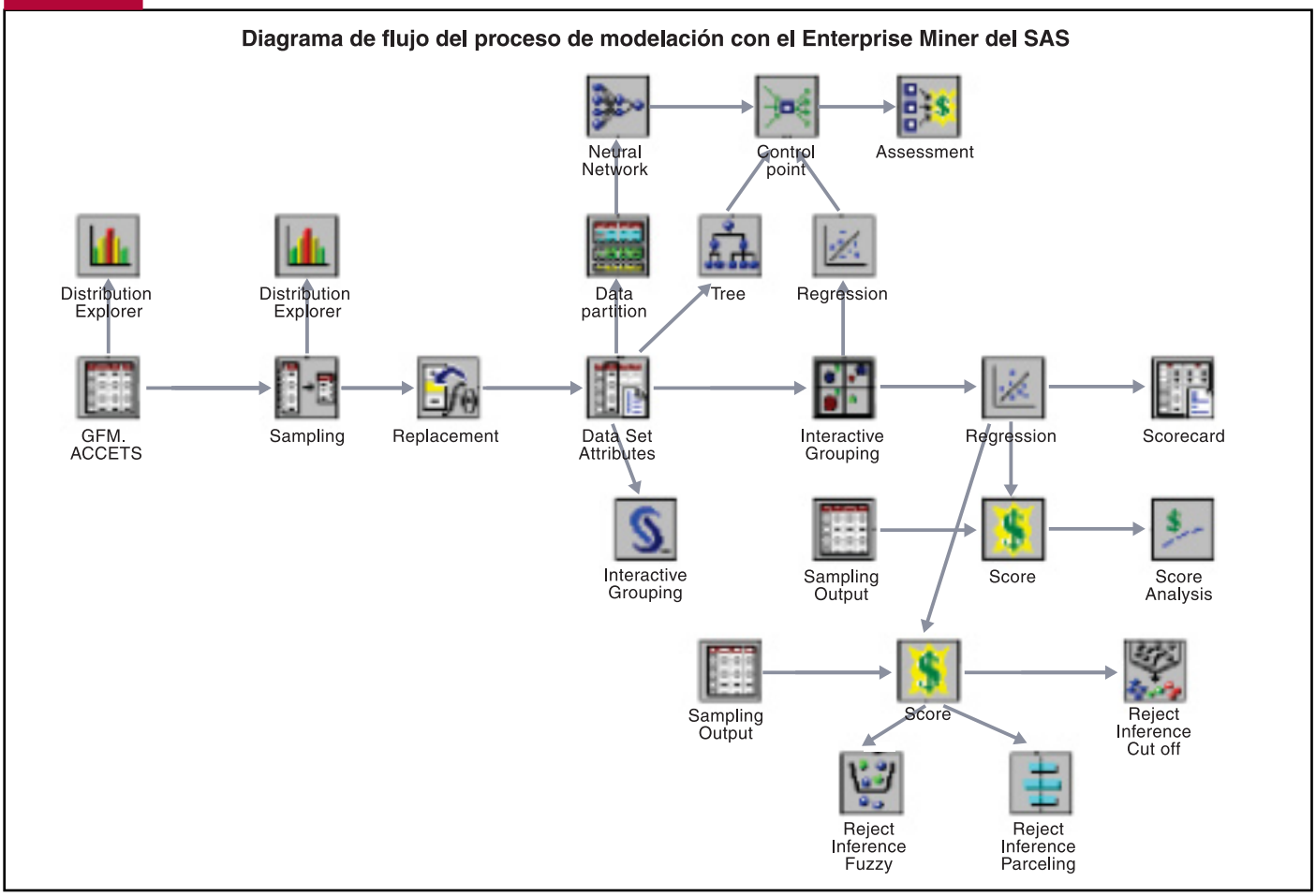
(d) Crear las variables WOE (*weights of evidence*) para lograr que los puntajes del *scorecard* varíen suavemente (de forma monótona) o linealmente a través de los atributos.

La cantidad de puntos que representa el valor del atributo en el *scorecard* se determina por dos factores: el riesgo de un atributo con relación a otros atributos de la misma característica (se determina por los *weights of evidence*) y la contribución relativa de la característica al puntaje total (se establece por los coeficientes de regresión logística).


GRUPO
IDM
Expertise en mercadotecnia e
investigación de mercados
Tecnología, experiencia
y estrategia
en investigacion
de mercados
Perpetua 42, Col. San Jose Insurgentes
Tel. 5611 3131 5611 2828 5611 2002 • Fax 5611 3131 • 01 800 639 4436
www.idm-mex.com idm@idm-mex.com

Gráfica 1

Diagrama de flujo del proceso de modelación con el Enterprise Miner del SAS



El WOE de un atributo se define como el logaritmo de la relación entre la proporción de buenos (*good)* y la de malos (*bad*) en el atributo:

$$WOE = \log \frac{p^{good}_{attribute}}{p^{bad}_{attribute}} \qquad p^{good}_{attribute} = \frac{\#\ goods\ attribute}{\#\ goods\ charact.}$$

$$p^{bad}_{attribute} = \frac{\#\ bads\ attribute}{\#\ bads\ charact.}$$

Valores grandes negativos de la variable corresponden a un alto riesgo y alto valores positivos se relacionan con bajo riesgo. Partiendo de que la cantidad de puntos en el *scorecard* es proporcional a la del WOE, el proceso de clasificación determina cuánto vale un atributo con respecto a los otros dentro de una misma característica.

Después que la clasificación ha definido los atributos de las características hay que determinar el poder predictivo de las mismas; es decir, su capacidad para separar los altos riesgos de los bajos riesgos. Esto se realiza con la ayuda de una medida llamada *Information Value* (IV). El IV es la suma ponderada de las WOE de los atributos. El factor de ponderación es la diferencia entre la proporción de buenos y la de malos en el atributo respectivo:

$$IV = \sum \left(p^{good}_{attribute} - p^{bad}_{attribute} \right) \times WOE_{attribute}$$

El valor de IV debe ser mayor que .02 para que una característica sea considerada para incluirla en el *scorecard*. IV por debajo de .1 se consideran débiles, menores a .3 medios y menores a .5 fuertes. Si el IV es mayor de .5, la característica está sobrepredicha (*overpredicting*), lo que significa que de alguna forma está trivialmente relacionada con la variable objetivo (target).

El ajuste del modelo de regresión logística.

Una vez que ha sido cuantificado el riesgo relativo de los atributos en una misma característica, un análisis de regresión logística determina el peso relativo de las características unas con otras.

El cálculo de los puntajes para cada atributo.

Aquí el WOE de cada atributo es multiplicado por el coeficiente de regresión de su característica para obtener el puntaje del *scorecard* del atributo. El puntaje total del solicitante es proporcional al logaritmo de la probabilidad de su razón *good/bad* estimada. A continuación, se presenta la ecuación para el cálculo de los puntajes del *scorecard* para cada solicitante. Los puntajes están en escala lineal de enteros conforme a las normas de esta industria.

$$score = \log (odds) * factor + offset =$$

$$(-\sum_{i=1}^{n} (woe_i * ß_i) + a) * factor + offset =$$

$$(-\sum_{i=1}^{n} (woe_i * ß_i + \frac{a}{n})) * factor + offset =$$

$$\sum_{i=1}^{n} (- (woe_i * ß_i + \frac{a}{n}) * factor + \frac{offset}{n})$$

Se tomó una escala de puntajes tal que el valor de 600 corresponde a una relación *good/bad* de 50/1 y que un incremento en el puntaje de 20 unidades coincide con el doble de la relación g*ood/bad*. Para la obtención de la regla de escalamiento que transforme los puntajes de cada atributo se usan las ecuaciones:

$$600 = \log (50) * factor + offset$$

$$620 = \log (100) * factor + offset$$

$$factor = 20 / \log (2)$$

$$offset = 600 - factor * \log (50)$$

El *scorecard* resultante es una tabla, opcionalmente en formato HTML, una parte de la cual, se muestra en la tabla 1. Se aprecia como los puntajes de las características cubren diferentes rangos.

Tabla 1 **Parte del scorecard obtenido para el Grupo Financiero Mexicano**

Characteristic Name	Attribute	Scorecard Points
Edad	.-> 27	56
Edad	27 -> 30	70
Edad	30 -> 31	61
Edad	31 -> 38	60
Edad	38 -> 44	60
Edad	44 ->.	70
Estado	Chiapas, Coahuila, Jalisco, Nayarit, Nuevo León, Sinaloa, Sonora	49
Estado	Chihuahua, Guerrero, San Luis Potosí	55
Estado	Aguascalientes, Distrito Federal	56
Estado	Guanajuato, Michoacán, Puebla, Querétaro, Yucatán Zacatecas	67
Estado	Hidalgo, México, Tamaulipas	70
Estado	Baja California, Campeche, Durango, Morelos, Oaxaca, Quintana Roo,Tabasco, Tlaxcala, Veracruz	80
Tipo Trabajo	“0” . “8”	56
Tipo Trabajo	“1” . “2”	74
Tipo Depto.	.-> 4	76
Tipo Depto.	4 ->.	-102

El análisis de los puntajes.

La función de esta etapa es:

(a) Visualizar la distribución de varios puntajes con relación a estadígrafos.

(b) Ayudar a la determinación del puntaje de corte óptimo a través de la creación de una tabla de ganancia, las curvas de equilibrio y el análisis de las características.

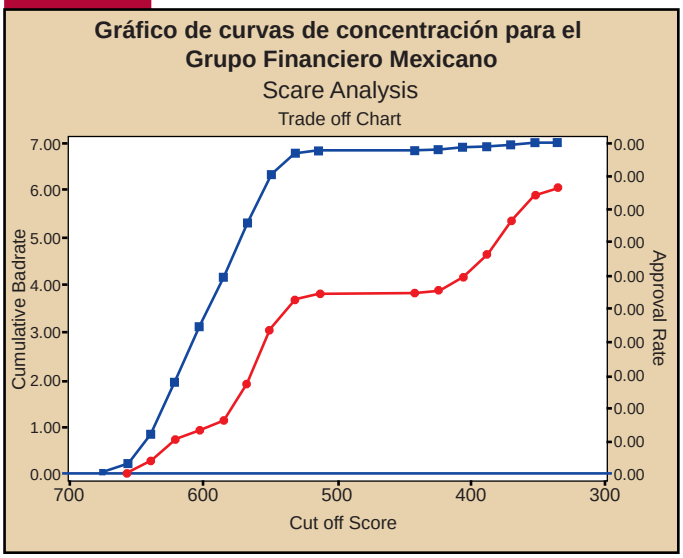
(c) Sacar los códigos de programación de la aplicación.

En la gráfica 2 se muestra una de entre las muchas tablas y gráficos que da esta etapa del proceso y que permite el análisis de los puntajes de riesgo obtenidos.

La inferencia de los rechazos.

El objetivo de este paso es llevar a cabo la inferencia *good/bad* en una muestra de solicitudes denegadas

Gráfica 2



y crear una muestra de solicitud *"through the door"* con indicadores de desempeño. Esto generalmente se hace en dos etapas:

Clasificación: Donde a las solicitudes denegadas se les asigna una clasificación *good or bad*.

Integración: Donde las solicitudes ya clasificadas se agregan a una muestra de *goob/bad* conocidos para crear una muestra unificada de todos los *good/bad (augmented data set)*.

La modelación en la base completa.

A partir de la base de datos integrada (*augmented data set*), que contiene la base original (*Accepts*); es decir, la que contiene todas las solicitudes aceptadas, más la base de datos creada por el nodo *Reject Inference* sobre la base de todas las solicitudes denegadas, se realiza de nuevo todo el proceso de modelación. El objetivo de este paso es eliminar, como se indicó anteriormente, el sesgo que se produce al predecir valores de malos y buenos a partir de la base de solicitudes aceptadas sin tener en cuenta los puntajes que daría este modelo en la base de datos de solicitudes que fueron denegadas con anterioridad.

Resultados y Discusión

Como se aprecia en la tabla 1, se obtuvo un *scorecard* de alta confiabilidad y precisión que tiene en el mercado un alto valor para las instituciones financieras. Los valores de riesgo obtenidos son óptimos dado esto por los criterios de optimización del modelo.

Por motivos de confidencialidad se ha hecho más énfasis en la parte metodológica que en la práctica; sin embargo, como señalamos anteriormente, el valor práctico es enorme. Basta señalar que un trabajo de este tipo, que es realizado por empresas consultoras transnacionales, puede tener un costo de entre uno y cinco millones de dólares americanos.

Con el objetivo de comparar la bondad de ajuste del modelo de regresión, se cálculo un árbol de decisión y una red neuronal con la arquitectura de un *Multilayer Perceptron*. En la parte superior de la gráfica 1, se pueden apreciar el *Nodo Neural Network* y el *Nodo Tree.*

Analizando el desempeño de todos los modelos se constató que los modelos de árbol y de red neuronal no sobrepasan el desempeño del *scorecard* de regresión para este caso en particular.

Conclusiones

La creación del modelo de *credit scoring* con una herramienta de minería de datos permite una serie de beneficios:

(a) El analista accede a las herramientas del minero a través de una interfase gráfica para crear los diagramas de flujo de proceso que sirven de estructura a las actividades de análisis.

(b) Los nodos que forman el diagrama de flujo del proceso están diseñados, de forma tal que el analista puede interactuar con los datos y el modelo.

(c) Emplear a fondo su *expertise*, usando el *software* como un volante y no como piloto automático.

(d) Es ideal para probar nuevas ideas y experimentar con nuevos modelos de una manera eficiente y controlada. Esto incluye, por ejemplo, la creación y comparación de varios *scorecards* de regresión, modelos de árboles y redes neuronales.

Con el minero de datos es posible crear una variedad de modelos como los *scorecards* de regresión, árboles de decisiones o redes neuronales. En la evaluación del modelo más adecuado para alcanzar la meta deseada se deben considerar criterios tales como:

(a) Facilidad en la aplicación del modelo.

(b) Facilidad para entenderlo.

(c) Facilidad para justificarlo.

Suscripción Anual

DATOS DIAGNOSTICOS TENDENCIAS

Por favor llene esta forma con letra clara y tinta negra y envíela por fax al: 01(55)52.54.42.10 esto nos permitirá integrarlo a nuestra base de datos y así poder brindarle un mejor servicio, no sólo en cuanto a nuestro boletín sino también por lo que se refiere a diversos eventos que la AMAI organiza cada año.

_____
Nombre Apellido paterno Apellido materno

Empresa donde trabaja: _____

Departamento o área: _____ Cargo: _____

Calle: _____

No.: _____ Colonia: _____

Ciudad: _____ Estado: _____

Municipio/Delegación: _____ C.P. _____

Tel(s): _____ Extensión: _____

Fax: _____ E-mail: _____

LABORO en:

- ☐ Agencia AMAI
- ☐ Agencia de Investigación de Mercado
- ☐ Depto. de Mercadotecnia
- ☐ Universidad/ Sector Académico
- ☐ Depto. de Investigación de Mercado
- ☐ Depto. Comercial/Ventas
- ☐ Independiente
- ☐ Otro _____

No lo deje para mañana,
su respuesta **ES INDISPENSABLE,** aun si lo que hace es confirmar sus datos.