METODOLOGÍA PARA UN SCORING DE CLIENTES SIN REFERENCIAS CREDITICIAS

Osvaldo Espin-García[1]
Carlos Vladimir Rodríguez-Caballero[2]

Espin-García, O. y Rodríguez-Caballero. C. (2013). Metodología para un scoring de clientes sin referencias crediticias. *Cuadernos de Economía*, 32(59), XX-XX.

Las decisiones de otorgamiento de crédito son cruciales en la administración de riesgos. Las instituciones financieras han desarrollado y usado modelos de *credit scoring* para estandarizar y automatizar las decisiones de crédito, sin embargo, no es común encontrar metodologías para aplicarlos a clientes sin referencias crediticias, es decir clientes que carecen de información en los burós nacionales de

[1]Teaching Assistant (TA) y estudiante de maestría en bioestadística en el departamento de estadística y ciencias actuariales de la Universidad de Waterloo. E-mail: oespinga@uwaterloo.ca Dirección de correspondencia: 395 Hazel St. Apt 3, Waterloo, (Ontario, Canada), N2L3P7.

[2]Magister en economía con especialidad en econometría. Profesor Adjunto de la Facultad de Ciencias de la Universidad Nacional Autónoma de México. E-mail: vladimir.rodriguez@ciencias.unam.mx. Dirección correspondencia: Facultad de Ciencias. Universidad 3000 Circuito Exterior S/N C.P. 04510. Ciudad Universitaria (México DF, México).

Al iniciar la elaboración del presente artículo, Carlos Vladimir Rodríguez-Caballero se encontraba realizando una estancia de investigación en la WSB-National Louis University en Nowy Sacz (Polonia), bajo la supervisión de Jacek Leskow, a quien le agradece la hospitalidad. Igualmente, el autor manifiesta su agradecimiento a la Universidad de Guanajuato y al CONACYT, por el financiamiento otorgado.

Los autores agradecen a José Manuel de Caso Pando por sus valiosas aportaciones al momento de desarrollar el modelo presentado y a dos árbitros anónimos por sus apreciadas sugerencias que permitieron mejorar la calidad del trabajo.

Este artículo fue recibido el 24 de noviembre de 2011, la nueva versión el 25 de marzo de 2012 y su publicación aprobada el 15 de marzo de 2012.

crédito. En este trabajo se presenta una metodología general para construir un modelo sencillo de *credit scoring* enfocado justamente a esa población, la cual ha venido tomando una mayor importancia en el sector crediticio latinoamericano. Se usa la información sociodemográfica proveniente de las solicitudes de crédito de una pequeña institución bancaria mexicana para ejemplificar la metodología.

Palabras clave: Scorecard, CHAID, logit, administración de riesgos, crédito.
JEL: C14, C18, C25, G21.

Espin-García, O. and Rodríguez-Caballero. C. (2013). Scoring method for clients without bank references. *Cuadernos de Economía*, 32(59), XX-XX.

The credit grant decisions are crucial for risk management. Financial institutions have developed and used credit scoring models for automating and standardizing credit granting. However, in the literature it is not common to find a methodology to be applied to clients without previous credit experience, in other words those who lack information in the national credit bureaus. In this paper a basic methodology to build a scorecard model is presented, considering that the Latin-American banks have been incremented the credit policies in favor of this kind of population. We use demographic information for an objective population from a small Mexican bank to illustrate the methodology.

Keywords: score card, CHAID, logit, risk administration, credit.
JEL: C14, C18, C25, G21.

Espin-García, O. et Rodríguez-Caballero. C. (2013). Méthode de notation pour les clients sans références de crédit. *Cuadernos de Economía*, 32(59), XX-XX.

Les décisions de crédit sont cruciales pour la gestion des risques. Les institutions financières ont développé et utilisé des modèles de notation de crédit dans le but de standardiser et automatiser les décisions de crédit, cependant, ce n'est pas fréquent de trouver des méthodologies à appliquer aux clients sans références de crédit, à savoir les clients qui n'ont pas d'informations sur les bureaux de crédit nationaux. Cet article présent une méthode générale pour construire un modèle simple de notation de crédit portait précisément cette population, laquelle a gagné de plus en plus importance dans le secteur du crédit dans l´Amérique latine. On utilise les renseignements relatifs aux caractéristiques sociodémographiques des demandes de crédit auprès d'une banque mexicaine petite pour illustrer la méthodologie.

Mots clés : tableau de bord, CHAID, logit, gestion des risques, crédit.
JEL : C14, C18, C25, G21.

INTRODUCCIÓN

El mercado de créditos al consumo ha crecido rápidamente en las últimas dos décadas. De acuerdo con el *Federal Reserve Board's Statistical Release on Consumer Credit*[3], el total del *crédito revolvente*[4] al consumo en los Estados Unidos fue de USD$ 792,5 billones para julio de 2011. Lo interesante es que dicho crecimiento se encuentra también en Latinoamérica, lo cual obedece a que diversos bancos comerciales extranjeros tienden a emplear las mismas estrategias masivas de crédito.

Por otro lado, diversas instituciones financieras, entre ellas algunos bancos comerciales, han incrementado sus nichos de mercado a favor de la inclusión de clientes que carecen de información crediticia en los burós nacionales de crédito, la cual, a finales de la década pasada fue una importante política a seguir, tanto en instituciones grandes como en pequeñas. A diferencia de las grandes compañías, usualmente, las pequeñas no consiguen tener el suficiente capital para solicitar a las consultoras financieras con experiencia internacional, algún modelo de *originación* que esté enfocado a este tipo de clientes.

Por lo anterior, y dado el continuo crecimiento del mercado de crédito al consumo, la eficiente toma de decisiones es cada vez más importante, tanto en aspectos sociales (eficiencia) como privados (rentabilidad). Frente a ello, existe un creciente interés en modelar de manera más oportuna el riesgo y se han usado modelos estadísticos como los de *credit scoring o scorecard*, cuyo objetivo es el de identificar la probabilidad de impago de un grupo de clientes con características similares.

Kiefer y Larson (2006) proveen un resumen de cuestiones conceptuales y estadísticas que surgen durante el desarrollo de un modelo de *credit scoring*. Bierman y Hausman (1970), Dirickx y Wakeman (1976), Srinivasan y Kim (1987), Thomas, Crook y Edelman, (1992), entre otros, han usado distintas técnicas matemáticas y estadísticas para su diseño. Sin embargo, aunque existen avances substanciales en este corpus científico, aun no se tiene una metodología que haya sido internacionalmente aceptada como una práctica a seguir.

En este trabajo se desarrollará una metodología para elaborar un modelo que logre predecir el comportamiento de impago en función de la información sociodemográfica del cliente. La institución financiera, debería usar este predictor, mediante un valor crítico, en la toma de decisión para otorgar o no el crédito. La información recopilada de la solicitud de crédito permite, mediante técnicas estadísticas

[3] La información se encuentra disponible en: http://www.federalreserve.gov/releases/g19/Current/

[4] La Comisión Nacional Bancaria y de Valores de México define un crédito al consumo revolvente como "la característica contractual de la apertura de crédito, que da derecho al acreditado a realizar pagos, parciales o totales, de las disposiciones que previamente hubiere hecho, quedando facultado, mientras el contrato no concluya, para disponer en la forma pactada del saldo que resulte a su favor" (Esta y otras definiciones de términos muy específicos en el argot crediticio y financiero pueden ser consultados libremente en el glosario de términos de http://portafoliodeinformacion.cnbv.gob.mx/Paginas/default.aspx

de árboles de decisión y de regresión logística, calibrar dicho modelo, siendo el objetivo primordial del mismo el asignar un puntaje *(score)* a cada cliente, de acuerdo con sus características. Lo anterior para determinar la probabilidad de impago de los solicitantes y poder coadyuvar al área de control de riesgos para establecer si se otorga o no el crédito, sobretodo en pequeñas instituciones financieras que no puedan incluir en sus costos los modelos de las famosas consultoras[5].

Como ejercicio empírico, utilizando datos provenientes de un pequeño banco mexicano, se analiza el segmento poblacional que carece de información crediticia en alguno de los dos burós nacionales de crédito en México. El modelo de *scoring* genérico que emane de este trabajo, se desarrollará tomando en cuenta la información socio-demográfica. Para ello, se segmentará a la población usando árboles de decisión estadística y prosiguiendo con el cálculo del puntaje de la probabilidad de impago, usando regresión logística.

El artículo está dividido en cuatro secciones: En la primera de ellas se explican los elementos para el desarrollo de la metodología propuesta en la cual se detallan las variables que habrán de usarse a lo largo del artículo. La segunda sección trata de las herramientas microeconométricas que se usan en la metodología, se explica a detalle la construcción de árboles de decisión y los fundamentos teóricos de los modelos *logit*. La sección tres presenta la aplicación de la metodología propuesta con datos de un banco mexicano. Finalmente, la cuarta y última sección presenta las conclusiones del trabajo. Elementos para el desarrollo de la metodología propuesta

MODELOS SCORING CONVENCIONALES

Un *scorecard* clasifica a la población objetivo dentro de dos o más grupos usando diversas técnicas estadísticas. Es común encontrar, por un lado, propuestas que emplean métodos econométricos de variables dependientes limitadas, como lo son los modelos logit, probit y logísticos; y por otro lado, los métodos de clasificación estadística.

Los modelos *scoring* que una institución financiera desarrolla, pueden distinguirse en cuanto al tiempo de cobertura o ventana de información. Lo más común es tomar un diseño muestral en forma de sección cruzada, considerando una selección de crédito al consumo en el tiempo t, para posteriormente seguir su comportamiento de pago sobre k periodos en el futuro. Igualmente, este tipo de modelos se desarrollan para predecir el comportamiento del cliente sobre el intervalo $[t; t+k]$, como una función de las características observadas en el tiempo t.

En contraste, el estudio dinámico del comportamiento de la calidad de crédito requiere observaciones en múltiples periodos de tiempo, para un conjunto fijo que

[5]Fair Isaac Corporation (FICO) es tal vez la más conocida entre todas ellas (Fair Issac, 2004), sus metodologías y sus modelos de *scoring* de comportamiento han sido utilizados en una gran cantidad de bancos comerciales.

haya sido muestreado en un tiempo t base, es decir, un diseño longitudinal o de datos de panel. En ambos casos, los datos tienen que ser extraídos con suficiente detalle, para permitir el seguimiento del comportamiento crediticio, determinar cómo podría ser su saldo, incremento de línea, nivel de revolvencia, entre muchos otros aspectos, por cada cliente en cada instante del tiempo.

Obtención de información

Para el desarrollo del modelo de *scoring* para clientes sin referencias crediticias, se deben formar bases de datos usando la información perteneciente a la solicitud de crédito que enmarcará el perfil socio-demográfico del cliente. Sobre el total de clientes, el primer filtro a realizarse debería permitir depurar aquellos que tengan un estatus de inactividad dentro del portafolio de la institución financiera.

Por otro lado, cabe señalar que en la información socio-demográfica proveniente de las solicitudes de crédito, es altamente probable encontrar falsedad en la información, por lo que se recomienda considerar ciertos casos como anormales para evitar que traigan complicaciones en los análisis posteriores, ejemplo de ello serían ingresos demasiados altos, número de dependientes económicos elevado, entre otros. Cada institución bancaria debería elegir los rangos a considerar entre dichas variables.

Clasificación de clientes

Para la clasificación de los clientes es común encontrar que las instituciones bancarias usen formas convencionales. Una vez definida la ventana de observación (podrían considerarse 24 meses), se procede a determinar el tipo de cliente según al número de pagos vencidos que hayan presentado en dicho periodo. En la práctica bancaria, sobre todo en aquellos departamentos de riesgo que se rigen con estándares internacionales o bien bajo lineamientos de sus matrices corporativas, se suele tomar el caso especial de definir a la población como: clientes que hayan tenido un máximo de un pago vencido como cliente bueno, el que haya tenido un máximo de dos como indeterminado y aquellos con 3 o más pagos vencidos como malos. La decisión de incorporar a los clientes indeterminados o no, obedece a motivos de aumentar la población buena o mala.

A pesar de que la forma anterior ha sido ampliamente usada por instituciones financieras o burós de créditos, existen otras propuestas capaces de clasificar a los clientes de acuerdo con formas no convencionales, ejemplo de ello es la desarrollada por Karlis y Rahmouni (2007), usando modelos de mezclas *poisson* o la ejemplificada con datos mexicanos de Rodríguez-Caballero (2011) desde la propuesta original de Dellaportas, Karlis y Xekalaki (1997).

Variables para desarrollar el *scoring*

Existe una enorme diversidad de variables que se pueden considerar para el desarrollo de un modelo *scoring*. La información socio-demográfica podría incluir variables cualitativas como el estado civil, la educación, el tipo de vivienda, entre muchas otras, y cuantitativas como el ingreso, la edad, la capacidad de pago declarada, entre otras.

A partir de un análisis exhaustivo realizado por los autores, se han encontrado que las variables mostradas en el Cuadro 1, han resultado ser eficientes en la predicción de la probabilidad de incumplimiento. Un análisis descriptivo de estas variables siempre será necesario para poder identificar sesgos e irregularidades entre los datos, así como su estructura, con el fin de encontrar posibles cortes poblacionales, siempre teniendo en cuenta una visión correcta del negocio.

Lo anterior, debe tomarse en cuenta sobre todo para evitar tener inferencias estadísticamente correctas, pero de escasa o nula relevancia. Un ejemplo sencillo de ello podría ser que al correr un modelo predictivo se obtenga un resultado final en donde la población más joven, y con muy altos ingresos, es aquella que presenta el menor nivel de riesgo de toda la población bajo estudio. Si bien estadísticamente pudiera ser correcto al momento de correr una regresión, por ejemplo, es muy probable que la población que se encuentra en este segmento sean solo unas cuantas personas, o en el peor de los casos, se tenga presencia de errores de medición.

Una condición importante que deben mostrar dichas variables es que no presenten una alta correlación significativa[6], ello logrará, como se comentará más adelante, satisfacer una mejor segmentación en la población[7]. Cabe señalar que las variables de capacidad de pago calculada (cpc) y la razón, capacidad de pago declarada entre ingreso $(cpd/ingreso)$, del Cuadro 1, son variables calculadas, cuya finalidad es hacer comparaciones e incluir indicadores que pudieran reflejar de una forma más adecuada el comportamiento de los clientes.

Inicialmente, se deben desarrollar grupos de segmentación para encontrar una eficiente separación entre las poblaciones mencionadas usando técnicas multivariadas, específicamente, se pueden usar técnicas de árboles de decisión, para posteriormente utilizar algún tipo de modelo lineal que sirva para puntear los resultados que de la primera parte emanan. En la sección siguiente se explican dichas técnicas estadísticas.

[6] En el presente trabajo no se muestran los estudios de correlación correspondientes, no obstante, están disponibles para aquel lector interesado.

[7] Para el trabajo actual, los autores encontraron que al eliminar duplas de variables con correlación por encima al 75 %, se favorecía la correcta construcción de los árboles de decisión que se explican más adelante en el trabajo.

CUADRO 1.
VARIABLES PARA ELABORAR UN MODELO *SCORING* DE TARJETA DE CRÉDITO

Grupo	Variables
Variables cualitativas	Sexo Estado Educación Tipo de vivienda Profesión Estado civil
Variables cuantitativas	Ingreso Edad Capacidad de pago declarada Capacidad de pago calculada Capacidad de pago declarada /ingreso Número de dependientes económicos Tiempo empleo actual Tiempo empleo anterior Tiempo en vivienda actual Tiempo en vivienda anterior Número de autos

Fuente: elaboración propia.

HERRAMIENTAS MICROECONOMÉTRICAS

Un modelo *scoring* construye una segmentación que pueda ser usada para clasificar a los clientes dentro de dos o más grupos distintos. Existen varias técnicas analíticas que han sido discutidas en la literatura y después implementadas en la industria para el desarrollo de los modelos *scoring*, los cuales pueden estar basados en regresión multivariada (Orgler, 1971), regresión de variable dependiente limitada como modelos logit (Wiginton, 1980), probit y tobit, (Henley, 1995).

A pesar de que los modelos de variable dependiente limitada constituyen un mejor mecanismo que aquellos de regresión lineal, debido a que usualmente la variable dependiente es discreta (clientes buenos y malos), Henley (1995) encontró que la regresión logística no fue mucho mejor que la lineal. Ello se atribuye a que una gran cantidad de créditos se encuentran entre los cuantiles 0,2 y 0,8, en los cuales ambas distribuciones son prácticamente idénticas. No obstante, al ser el comportamiento de las colas justamente la preocupación en el riesgo de crédito, es común seguir realizando el modelaje vía regresiones logísticas.

Por otro lado, existen técnicas de análisis estadístico multivariado como el análisis discriminante, árboles de decisión, análisis factorial, clúster, entre otros (Girault, 2007). Por su parte, Eisenbeis (1977, 1978) presenta una crítica severa al momento de usar el análisis discriminante en estudios de negocios, economía y finanzas. Estos modelos han sido demeritados debido a que algunos artículos han sobre estresado los resultados (Hand *et al.*, 1997). Igualmente, se han empleado otras

técnicas no paramétricas como redes neuronales (Yobas, Crook y Ross, 2000; Desai, Convay, Crook y Overstreet, 1997), métodos de programación lineal (Boyle, Crook, Hamilton y Thomas, 1992) y más recientemente el análisis de sobrevivencia (Andreeva, 2005).

En Srinivasan y Kim (1987) se puede encontrar un estudio acerca de la eficiencia relativa al estudiar los modelos de *credit scoring*, bajo enfoques paramétricos y no paramétricos. Dicho artículo presenta evidencia en favor de usar los métodos de clasificación recursiva (árboles de decisión). También se puede consultar el trabajo de Thomas (2000), para una inspección de las técnicas estadísticas y de optimización, usadas al construir modelos de otorgamiento de créditos.

Por su parte, Baesens (2003) realiza un estudio cuidadoso a partir de ocho bases de datos usando ocho métodos diferentes y 17 modelos, para evaluar la precisión del scoring: dos modelos lineales, regresión logística, programación lineal, cuatro diferentes variantes de máquinas de soporte vectorial, cuatro diferentes árboles de decisión, dos variantes de la técnica de vecinos más cercanos *(nearest neighbours)*, redes neuronales y dos técnicas bayesianas de segmentación. Baesens confirmó que la eficiencia del *scoring* más pobre, estadísticamente hablando, se consiguió con el método de *Naïve Bayes*, mientras que los mejores resultados se lograron al emplear regresiones logísticas, redes neuronales o árboles de clasificación.

Finalmente, Crook, Edelman y Thomas (2007) hacen un compendio acerca de los resultados encontrados en la literatura especializada, con respecto a las técnicas usadas a lo largo de un par de décadas. En la tabla 2 de dicho artículo, se muestra que la tendencia a modelar el *scoring* a través de técnicas de segmentación, como los árboles de decisión, ha prevalecido con el tiempo, a pesar del avance considerable de la literatura con respecto al aprendizaje artificial (algoritmos genéticos, redes neuronales, programación genética y máquinas de soporte vectorial).

De hecho, se puede encontrar en dicha tabla, que la precisión de los árboles de decisión estadística es igual o incluso algunas veces mejor, que aquellos que emplean tecnologías más avanzadas. Lo mismo sucede con los métodos econométricos de variables dependientes limitadas, los cuales han sido aplicados en la industria bancaria debido, posiblemente, a su sencilla conceptualización y a que ambas técnicas se encuentran disponibles, casi en cualquier paquete estadístico.

En el presente trabajo se emplea un híbrido de ambas técnicas, usando las variables descritas en la sección anterior. Este modelo resultará de combinar las técnicas de análisis multivariado con regresión logística, y tiene como objetivo predecir el comportamiento de la población buena y mala basándose en sus experiencias de pago e incumplimiento.

La principal diferencia entre el uso de una técnica u otra, radica en que unos se encargan de modelar vía criterios de divergencia entre los tipos de cliente (análisis multivariados y programación lineal), la selección de la mejor combinación de factores y el peso de los mismos para el desarrollo del modelo de clasificación.

Mientras que otros métodos (regresión y redes neuronales) emplean criterios de minimización del error, los cuales están adecuados para construir los modelos predictivos.

Técnica CHAID y sus extensiones

Los árboles de decisión en el desarrollo de un *scoring* son usados como herramienta para el cálculo de los momios de incumplimiento (*odds*, disparidad o razón de oportunidades) y representa un método efectivo para la estimación. Un mismo modelo permite diferentes usos, como mantenimiento de clientes considerados como buenos (probabilidades bajas de incumplimiento), cobranza proactiva y discriminada por nivel de riesgo para los clientes calificados como malos o con probabilidades altas de llegar a incumplimiento. Cuando en la administración de riesgos se busca qué perfil socio-demográfico pertenece a un nivel determinado de riesgo, se construyen una serie de tablas que permiten ver la asociación existente entre variables. Escobar-Mercado (1992) comenta al respecto:

> No se trata de cruzar cada pregunta con el resto, sino de seleccionar una serie de hipótesis plausibles con el conocimiento previo, teórico o empírico, de la realidad que se está investigando, y, de acuerdo con ellas, realizar los cruces que pongan a prueba las conjeturas. Una manera de facilitar la tarea de selección de variables relevantes en la explicación de la contestación a una pregunta dada es la técnica del análisis de segmentación, que proporciona además una descripción de las diferencias que los distintos grupos de una muestra pueden presentar en un determinado rasgo[...] En su uso, se distinguen, por un lado, una variable cuya distribución se desea explicar y, por el otro, un conjunto de variables, nominales u ordinales, con estatus de independientes. Estas reciben el nombre de predictoras y tienen la finalidad de conformar grupos que sean muy distintos entre sí en la variable o variables dependientes (Escobar-Mercado, 1992, 2).

El análisis de segmentación debe ser utilizado, primordialmente, con fines exploratorios y su ideología consiste en buscar exhaustivamente las mejores asociaciones de las variables explicativas con la dependiente. Seleccionar automáticamente las mejores variables predictivas permite hallar grupos distintos para diversas características. De este modo, las muestras quedan fragmentadas en distintos tipos de personas u objetos, cuya descripción constituye un objetivo adicional de esta técnica (Escobar-Mercado, 1998).

Este tipo de análisis se ha usado, fundamentalmente, para estudiar variables dependientes cuantitativas, utilizando el algoritmo presentado por Morgan y Sonquist (1963), de manera frecuente. No obstante, aquí se emplea una derivación de esta técnica que se distingue por utilizar el estadístico χ^2 para seleccionar las mejores variables predictivas[8].

[8] El estadístico de prueba puede desarrollarse tanto con el enfoque de Pearson $[\chi^2 = \sum\sum \frac{(f_{ij}-f^*_{ij})^2}{(f^2_{ij})}]$ como con el cociente de verosimilitud $[L^2 = 2\sum\sum f_{ij} ln \frac{f_{ij}}{f_{ij^*}}]$.

Escobar-Mercado (1998) recomienda seguir los siguientes pasos lógicos para realizar esta tarea:

1. Preparación de las variables. Selección de variable dependiente y elección de posibles variables predictivas. Dicho artículo comenta que es preferible trabajar con menos de 10 variables.
2. Agrupación de las categorías de las variables independientes en el caso de que tengan un perfil similar al de la variable dependiente.
3. Primera segmentación, que consiste en la selección de la variable que mejor prediga la variable dependiente.
4. Segunda segmentación. Para cada segmento formado en el paso anterior, se debe buscar entre las variables cuyos valores han sido previamente agrupados de la misma forma que en el paso 2, la que tenga mayor poder predictivo.
5. Sucesivas segmentaciones. Se procede de forma similar al paso anterior en cada grupo formado por la segmentación previa.

Hay varios procedimientos para llevar a cabo la segmentación. A continuación se presenta con mayor detalle el algoritmo llamado CHAID *(Chi-squared Automatic Interaction Detection)*. Esta técnica, desarrollada fue por Cellard, Labbe y Cox (1967); Bouroche y Tennenhaus (1972); Kass (1980) y Magdison (1993) –quien finalmente lo adaptó para el programa computacional SPSS–, tiene como distintiva de otros algoritmos binarios, que se pueden formar segmentos con más de dos categorías al mismo tiempo. Al igual que otras prácticas de segmentación, las operaciones elementales que realiza son:

- La agrupación de las categorías de las variables predictivas.
- La comparación de efectos entre distintas variables.
- La finalización del proceso de segmentación.

El algoritmo CHAID es tal vez el más conocido y usado, sin embargo, existen otro tipos de segmentaciones como el C&RT de Breiman, Friedman, Olshen y Stone, (1984), el CHAID Exhaustivo de Biggs, De Ville y Suen (1991) y el QUEST de Loh y Shih (1998). Eso sin mencionar las técnicas bayesianas de aprendizaje como el *Naïve Bayes* de Duda y Hart (1973), y redes neuronales bayesianas de Pearl (1988). Todas ellas se encuentran programadas en sistemas informáticos como SPSS, SAS, entre otros. A continuación se explica la mecánica general de este tipo de técnicas estadísticas.

Reducción de las categorías más discriminantes

En esta etapa se seleccionarán las categorías de las variables predictivas que discriminen de mejor forma a la variable dependiente. Se trata de reducir la complejidad de la segmentación original sin incurrir en una pérdida de información. La reducción se logra de acuerdo con las características de las variables predictivas: nominales, ordinales, ordinales con valores perdidos, y cuantitativas.

El funcionamiento de formación de grupos de categorías homogéneas se basa en el estadístico χ^2. De acuerdo con Escobar-Mercado (1998), los pasos son los siguientes:

1. Se forman todos los pares posibles de categorías. Esto dependerá de la opción que se haya preferido dar a un determinado predictor.
2. Para cada posible par se calcula el estadístico χ^2 correspondiente a su cruce con la variable dependiente[9]. El par con χ^2 más bajo, siempre que no sea significativo, formará una nueva categoría de dos valores fusionados. La condición de que no sea significativo es muy importante, ya que en el caso de que lo fuese, indicaría que las dos categorías que se pretenden fusionar no lo pueden hacer, ya que son heterogéneas entre sí al considerar los valores de la variable dependiente; y el objetivo es justo lo contrario, asimilar categorías con comportamiento semejante.
3. Si se ha fusionado un determinado par de categorías, se procede a realizar nuevas fusiones de los valores del predictor, pero esta vez con una categoría menos, pues dos de las antiguas han sido reducidas a una sola.
4. El proceso se acaba cuando ya no pueden realizarse más fusiones, porque los χ^2 ofrecen resultados significativos.

Las segmentaciones binarias suelen ahorrar una gran cantidad de cálculos. Esto implica que se busque la mejor combinación de predictores que conduzcan a sólo dos grupos, para que finalmente las posibilidades de agrupación sean reducidas. Por ende, el χ^2 mayor de todas las posibles combinaciones grupales (con $k = 2$) será seleccionado.

El CHAID exhaustivo de Biggs *et al.* (1991) fue propuesto justamente para que la fusión continua de pares de valores fuera reducido, hasta que sólo quedara una dicotomía de valores.

Selección de variables predictivas

Teniendo las categorías más discriminantes, se debería continuar con la selección de aquellas variables que resulten ser las más predictivas. Para ello, se compara el χ^2 correspondiente de cada categoría; sin embargo, será conveniente modificar

[9]El estadístico χ^2 que se utiliza en este trabajo es aquel que sirve para probar la independencia de dos variables entre sí bajo H_0. En este caso, una observación consiste en los valores de ambas y está localizada en una entrada de la tabla de contingencia respectiva. El estadístico de contraste se forma con $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}}$, en donde $E_{i,j} = \frac{\sum_{k=1}^{c} O_{i,k}\sum_{k=1}^{r} O_{k,j}}{N}$ es la frecuencia teórica para la cual N es el tamaño de muestra y se hace referencia a una tabla de contingencia formada con r filas y c columnas. Si $\chi^2 > \chi^2_{(r-1)(c-1)}$ se rechazará la hipótesis nula de independencia.

la significación de cada predictor con el ajuste de Bonferroni[10], debido a que la probabilidad de obtención de un resultado significativo tiende a ser mayor con la proliferación de pruebas estadísticas. Lo mismo se repetirá para cada uno de los grupos formados por la primera segmentación.

Se puede conducir a interpretaciones precipitadas si el proceso de segmentación no es examinado con paciencia en cada fase. Finalmente, será muy útil estudiar el comportamiento de un cruce, por simple que sea, entre la variable dependiente y alguna otra compuesta, ello para analizar la capacidad predictiva de la segmentación realizada.

Regla de paro

Se debe determinar una regla que sirva para parar el algoritmo, con el propósito de evitar la formación de grupos terminales sin ninguna validez estadística, es decir, que de no detener el algoritmo se podrían tener nodos finales para cada elemento de la variable dependiente en cuestión.

Normalmente, en programas como SPSS es común encontrar cuatro tipos de filtros o reglas de paro: significancia, asociación, tamaño y nivel. Los primeros son los más utilizados en la técnica CHAID y consisten, básicamente, en no permitir segmentaciones que no sean estadísticamente significativas, como en el caso expuesto con anterioridad. Los segundos cumplen una función semejante y es común aplicar en los programas estadísticos los siguientes coeficientes de asociación: ϕ, V de Cramer, C de Pearson u otros[11]. La diferencia primordial entre ambos procedimientos radica en que al preocuparse por la asociación entre variables, no se es sensible al número de casos, a diferencia de aquellos de significancia.

Por otro lado, los filtros de tamaño evitarán la formación de grupos demasiado pequeños, dado el problema que supone la generalización en estos casos. Finalmente, con los filtros de nivel[12] se especifica *a priori* el nivel máximo de segmentación.

De esta forma, después de haber explicado de manera detallada la metodología de un análisis de segmentación tipo CHAID, es posible entender que la función

[10] Existen diferentes formas de ajustar los valores p en las comparaciones múltiples. La idea general que subyace es ser más exigentes con el valor estándar de p, usualmente por debajo de $0,05$, en función del número total de comparaciones hechas para justificar que las diferencias sean estadísticamente significativas. Uno de los métodos más conocidos es el de Bonferroni, en el cual el valor p ajustado es calculado tras multiplicar el p valor por $B = \binom{I-1}{r-1}$, con originalmente I categorías, las cuales son reducidas a r después del proceso de fusión, en el caso de CHAID y con variables predictivas ordinales, o bien por $B = \frac{I(I-1)}{2}$, en el caso *Exhaustive CHAID* con variables predictivas ordinales, como en el caso del árbol de decisión primario.

[11] El coeficiente ϕ, la V de Cramer y el coeficiente de contingencia o C de Pearson son medidas de asociación en escala nominal. En una tabla de contingencia de dimensión $p \times q$, se tiene que $0 < \phi < A$, donde 0 representará independencia total y $A = \sqrt{min\{p-1, q-1\}}$ una asociación perfecta. A su vez la V de Cramer estará comprendida entre 0 (independencia) y 1 (asociación) y la C de Pearson entre 0 (independencia) y $B = \sqrt{\frac{min\{p-1,q-1\}}{min\{p-1,q-1\}+1}}$ (asociación).

[12] Por nivel se debe entender a cada una de los cortes del árbol.

clasificadora del análisis de segmentación permite configurar una serie de grupos que se distinguen por su comportamiento con respecto a una variable dependiente determinada. La especificación de las características de los grupos terminales formados por esta técnica es un excelente medio para describir grupos heterogéneos de la muestra.

En este trabajo se ha seguido la metodología planteada. Cada paso y cada filtro usado fue llevado a cabo mediante las instrucciones pre programadas es el sistema estadístico SPSS, ajustando ciertos valores con el software SAS en su librería de minería de datos.

Regresión logística

Dentro de los enfoques econométricos, los modelos de probabilidad lineal han caído en desuso por sus desventajas técnicas, en tanto que los modelos *probit, logit* y demás son superiores al análisis discriminante, ya que proveen para cada deudor una probabilidad de impago. A pesar de que los modelos de variable dependiente limitada son, en teoría, herramientas econométricas más apropiadas que la regresión lineal, esta arroja estimaciones similares a las de los anteriores cuando sus probabilidades estimadas se ubican entre el 20 % y el 80 %. No obstante, un *scoring* de riesgo deberá utilizar el primer tipo de modelos econométricos porque la importancia radica en alguna de las colas de la distribución condicional de la variable dependiente, es decir, del tipo de cliente.

Cuando al plantear un modelo econométrico, la variable dependiente toma valores discretos, se emplean modelos de regresión discreta. El caso más simple se da cuando es binaria y toma los valores de 0 ó 1, y se puede estimar con distintos enfoques como el modelo de probabilidad lineal, análisis discriminante o los modelos de tipo probit y logit.

La regresión logística es un modelo lineal general, en el cual las variables respuesta $Y_1, Y_2, \ldots, Y_n$ son independientes y $Y_i \sim Bernoulli(\pi_i)$. π_i se asume que está relacionado a x_i por

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta X \tag{1}$$

El lado izquierdo de la ecuación [1] es el logaritmo de las razones de probabilidad u *odds* para Y_i. El modelo asume que estos log-*odds* (o logit) son una función del predictor de x. El término $log\left(\frac{\pi}{1-\pi}\right)$ es el parámetro natural de esta familia exponencial, y en la ecuación [1], la función de enlace $g(\pi) = log\left(\frac{\pi}{1-\pi}\right)$ es usada.

La ecuación [1] puede ser reescrita como

$$\pi_i = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}} \tag{2}$$

De donde es posible ver a π_i como una probabilidad, naturalmente $0 < \pi_i < 1$.

Al igual que en un modelo de regresión lineal simple, cuando β es igual a cero, en un modelo de regresión logística, si $\pi(-\alpha/\beta) = 1/2$, no hay ninguna relación entre π y x. Por otro lado, β es el cambio en los log-*odds* correspondientes al incremento de una unidad en x.

En el contexto de los modelos de *credit scoring* se puede asociar βx_i a la calidad crediticia del individuo (variable latente o no observada), mientras que Y_i es definida mediante una variable binaria, donde será 1 si el cliente es identificado como malo o cero cuando sea clasificado como bueno. La calidad crediticia del individuo se supone como el resultado de una función lineal en sus parámetros y X contiene la información específica de los deudores. Las estimaciones de los parámetros se realizan mediante máxima verosimilitud y tras haberlas obtenido, la variable Y_i será el *score* o calificación crediticia del cliente, la cual representará la probabilidad de incumplimiento del mismo.

Habiendo definido el *score* de crédito, cambios en Y_i implicarán modificaciones en la probabilidad de incumplimiento (PD, por sus siglas en inglés) del individuo. La relación entre *score* y riesgo no es lineal, por lo que para valores del *score* muy bajos, un aumento en el mismo produce una rápida subida en la probabilidad de cumplimiento y una rápida disminución de la PD, mientras que para valores del *score* altos, una mejora en el mismo hace que la probabilidad de cumplimiento aumente poco y genera una leve caída en el riesgo. En otras palabras, cuanto mayor sea el *score*, menor será la caída en el riesgo derivada de un aumento en el primero.

Es importante mencionar que las estimaciones $\hat{\beta}_i$ no tienen una interpretación directa como en mínimos cuadrados ordinarios, ya que solo representan el efecto que un cambio en x_i tiene sobre el *score* del individuo, a la vez que su signo muestra si la relación con la PD es directa o inversa. Sin embargo, para cuantificar el efecto de x_i sobre la PD se debe computar su efecto marginal.

APLICACIÓN

Desarrollo del modelo

Para el desarrollo del modelo de *scorecard* para clientes sin referencias crediticias, se utiliza una base de datos correspondiente a una pequeña institución bancaria en México, con fecha de corte al 15 de octubre de 2007. Se seleccionan únicamente aquellos registros que cumplen con los filtros explicados en la sección correspondiente.

Dichos filtros lograron depurar la base de estudio a 4.064 registros, debido a que los demás tenían un estatus de inactivos (cuentas que por sus características de pagos y saldos no han tenido actividad en el producto bancario). De estas sólo fueron

clasificables (como bueno, malo o indeterminado) 2.674, los 1.390 restantes no contaban con información para poderlos asignar dentro de alguna categoría (su histórico de pagos no es el suficiente para definir al cliente) o bien fueron depurados tras la aplicación de ciertos filtros. Finalmente, la población buena es de 1.938, correspondiente al 85,3 % de la población total, mientras que la población definida como mala pagadora (333) corresponde al 14,7 %. Cabe mencionar que la población indeterminada se eliminó del estudio debido a la poca relevancia estadística dentro de los árboles probados.

El plan comercial del banco está enfocado a un sector poblacional en específico, por ello se toma la suposición de datos falsos para aquellos registros que indicaron ingresos superiores a los $20.000 en moneda nacional mexicana. Asimismo, se consideraron como atípicos aquellos casos cuyo número de dependientes económicos rebasara los 4 integrantes, debido a su no significancia estadística, en el sentido de que este segmento poblacional es menor al 1 % de la población total en la muestra estudiada.

La base se diseñó con el objetivo de permitir que el panel sea balanceado, es decir, que en cada instante en el tiempo cuente con exactamente la misma información. De esta forma, se define el tipo de cliente (bueno o malo) de acuerdo con el histórico de pagos del cliente, explicados con anterioridad.

Para la realización del árbol final fueron utilizadas todas las variables consideradas como predictivas para el modelo y se fueron descartando paulatinamente conforme iban presentando o no significancia al momento de generar particiones de la información (variables que no segregaban la información se iban sustrayendo), esto para seleccionar aquellas que mostraran mayor predictibilidad al momento de crear particiones en los árboles (ramas).

Para la construcción de los árboles se analizaron los dos grupos de variables, obteniendo una agrupación natural y haciendo esta distinción debido a las características intrínsecas a cada grupo de variables. Fueron probados diversos métodos (algoritmos) para la elaboración del árbol seleccionando, seleccionado aquel que mejor particionara la información (separación coherente y estadísticamente significativa). Para este fin fueron ajustados algunos parámetros dentro del programa, como es el caso de la utilización de la variable cpd/ingreso como variable de influencia, dada su construcción como un ponderador (idealmente entre 0 y 1), que ajusta las demás características con respecto a este índice.

Los diferentes árboles obtenidos, así como los códigos respectivos, están documentados en un archivo que el lector interesado puede solicitar, esto para analizar la evolución de los mismos y las variaciones en los parámetros que llevaron a la construcción final.

Para desarrollar el modelo *scoring* se estudian todas las variables y grupos descritos en el Cuadro 1, encontrando que los grupos y las variables más predictivas son las mencionadas en el Anexo 1. Los grupos fueron identificados mediante la utilización de técnicas de árboles de decisión explicadas en la sección de herramientas

microeconométricas. En el mismo anexo se presentan los árboles modelados. La Gráfica A1, corresponde al árbol de decisión primario y la Gráfica A2 al árbol de decisión secundario, los cuales se explican a continuación:

1. El árbol de decisión primario es utilizado para realizar la regresión logística y así generar el *score* correspondiente a cada conjunto de características.

2. El árbol de decisión secundario se utiliza para, una vez dado el punto de corte del *score*, seleccionar adicionalmente otras ramas del árbol primario para aumentar el nivel de aceptación con las mejores características. Esto debe realizarse identificando las tres mejores ramas de este árbol y conjuntamente con el primario, considerar las ramas que condensarán a más malos que buenos.

El árbol de decisión primario está formado por las variables: ingreso, edad y capacidad de pago declarada (CP_DEC). Mientras que el secundario ha sido formado usando las variables de sexo (SEX), nivel máximo de estudios (ESTUDIOS), estado civil (EDO_CIVIL), número de dependientes económicos (NUM_DEP).

Después de analizar el comportamiento de las técnicas de segmentación CHAID, Exhaustive CHAID, Quest y C&RT, se encontró que la mejor de ellas terminó siendo el Exhaustive CHAID. Usando la variable de influencia mencionada con anterioridad, pueden existir diversos mecanismos para la selección del mejor árbol, por ejemplo, el sentido de negocio, las políticas de crédito internas de la institución financiera, los valores estadísticos χ^2 encontrados, entre otros. En el trabajo actual se desarrolló la selección de aquellas segmentaciones que hicieran un mejor sentido del negocio con respecto a las políticas vigentes de otorgamiento de crédito de la institución precedente.

Como ya se ha comentado, el objetivo de realizar una regresión logística radica en la cuantificación de las variables segmentadas mediante los árboles calculados, esto para asignar a cada persona un puntaje *(score)*, el cual indique de manera rápida y clara su nivel de riesgo.

Cada grupo de análisis da lugar a distintas variables independientes, específicamente, cada rama final se vuelve una variable. Cada una de ellas será dicotómica, esto debido a que un cliente no puede estar presente en más de una rama en cada árbol a la vez. La variable dependiente en las cuatro regresiones logísticas siempre será la misma y representa el tipo de cliente, 0 y 1, bueno o malo, respectivamente.

El árbol primario es el que define inicialmente el *score* de riesgo, por lo que se deben considerar las ramas que arroja, 12 en el caso actual. Dichas variables son las independientes en el modelo logit empleado, donde $x_i = 1$, si y sólo si, esa observación cumple con las características descritas en la i-ésima rama del árbol de decisión primario $(1 \leq i \leq 12)$. Para x_k, $k \neq i$ su valor correspondiente es cero.

En el Anexo 2 se presentan los resultados referentes a la regresión logística. Como se aprecia en los resultados del Cuadro 2, todas las variables son significativas al 5 %, inclusive al 1 %. Una vez estimados los parámetros se procede a realizar la transformación logística, la cual modela la función de probabilidad de la muestra (Ecuación 2). Una vez obtenida la PD, se debe realizar un re-escalamiento de las probabilidades a puntos *score*. Posteriormente, se debe ajustar mediante una transformación del tipo

$$Score = \alpha * log\left(\frac{1-(\beta_1 * TPD - \beta_2)}{\beta_1 * TPD - \beta_2}\right) + \eta \tag{3}$$

En esta ecuación los parámetros $\alpha, \beta_1, \beta_2 y \eta$, son calculados para expresar el *score*, con respecto a un rango específico. Cada institución debe definir dicho rango.

Una vez que se obtiene un *score* para cada cliente se procede a realizar un análisis distribucional de la población, para determinar un punto de corte del modelo, es decir, determinar un *score* a partir del cual se aceptarán las solicitudes, rechazando todas aquellas que se encuentren por debajo de este punto de corte. Para esto se realiza lo siguiente:

1. Se hace un histograma de los *scores* segmentando el tipo de cliente.
2. Se grafica el kernel de la distribución de dichos *scores*, comparando las gráficas de buenos y malos, para visualizar la acumulación de los clientes con respecto a su score.
3. Se calculan los percentiles de los *scores* para cada uno de los tipos de cliente.

Se debe determinar un punto de corte que represente una probabilidad de incumplimiento. También es indispensable analizar los errores tipo I (clientes buenos rechazados) y II (malos aceptados) y el porcentaje de aceptación total. Para el modelo desarrollado de obtuvieron los siguientes puntos de corte (Cuadro 2).

CUADRO 2.
PUNTOS DE CORTE

Corte 1

	P(Incumplimiento)		**Aceptados**	**Rechazados**
>=650	14,20 %	Buenos	985	1.000
			49,62 %	50,38 %
		Malos	118	223
			34,60 %	65,40 %
		Indeterminados	142	206
			40,80 %	59,20 %

Error tipo I: 50,38 %; Error tipo II: 34,60 %; % aceptación: 45,56 %.

Corte 2

	P(Incumplimiento)		Aceptados	Rechazados
>=665	12,80 %	Buenos	648	1.337
			32,64 %	67,36 %
		Malos	65	276
			19,06 %	80,94 %
		Indeterminados	93	255
			26,72 %	73,28 %

Error tipo I: 67,36 %; Error tipo II: 19,06 %; % aceptación: 30,14 %.

Corte 3

	P(Incumplimiento)		Aceptados	Rechazados
>=690	9,10 %	Buenos	468	1.517
			23,58 %	76,42 %
		Malos	39	302
			11,44 %	88,56 %
		Indeterminados	75	273
			21,55 %	78,45 %

Error tipo I: 76,42 %; Error tipo II: 11,44 %; % aceptación: 21,77 %.

Corte 4

	P(Incumplimiento)		Aceptados	Rechazados
>=691	8,90 %	Buenos	346	1.639
			17,43 %	82,57 %
		Malos	27	314
			7,92 %	92,08 %
		Indeterminados	57	291
			16,38 %	83,62 %

Error tipo I: 82,57 %; Error tipo II: 7,92 %; % aceptación: 16,08 %.

Fuente: elaboración propia.

En el ejercicio actual se opta por considerar a una institución bancaria que tenga una alta aversión al riesgo, por lo que se considera un punto de corte mayor o igual a 691, que representa una probabilidad de incumplimiento a 6 meses de 8,90 %. Si bien dicha probabilidad de incumplimiento posee un valor muy alto, la selección del punto de corte está en función de las políticas de negocio de la institución financiera precedente, la cual consistía en que los modelos de originación de crédito no deberían soportar tasas de aceptación por debajo de 15 %. En realidad, la elección del punto de corte sopesaría la dupla PD-Nivel de aceptación, la primera impuesta por la propia área de riesgos y, la segunda por la estrategia de negocios aceptada por el consejo directivo.

Lo deseable siempre será seleccionar más clientes cumplidos (buenos), pero sin aceptar clientes malos (error tipo II). Para esto, se debe utilizar el árbol de decisión secundario. El problema en cuestión radica en que de manera general resulta complicado segmentar idealmente a la población sin referencias crediticias, debido a que no presentan información que suponga un comportamiento de pago y la sociodemográfica puede resultar no ser tan robusta, pese a tener un tamaño muestral grande.

Implementación del árbol secundario y modelo final

La aceptación presentada radica, principalmente, en las restricciones del modelo y al punto de corte seleccionado. Para aumentar el nivel de aceptación total, pero sin descuidar el error tipo II, se procede a hacer uso del árbol de decisión secundario.

Dado que por el punto de corte definido únicamente se aceptan clientes que hayan provenido de las 3 mejores ramas del árbol de decisión primario (ADP_Rama03, ADP_Rama08, ADP_Rama09), se consideran las 3 siguientes mejores ramas (ADP _Rama05, ADP_Rama06, ADP_Rama12) y algunas características adicionales para tomar la decisión de aceptarlas o rechazarlas. Estas características están asociadas con las 4 mejores ramas del árbol de decisión secundario (ADA_Rama01, ADA_Rama03, ADA_Rama06, ADA_Rama07).

No fueron utilizadas todas las relaciones de las tres ramas del árbol de decisión primario con las cuatro del árbol de decisión secundario. Hubo una selección de cuáles serían las relaciones que se tomarían en cuenta, esto para aminorar el error tipo II y aumentar la aceptación. Las características adicionales que fueron consideradas están contenidas en el Cuadro 3.

CUADRO 3.
CARACTERÍSTICAS ADICIONALES CONSIDERADAS

Relaciones entre ramas

	ADP_Rama05	**ADP_Rama06**	**ADP_Rama12**
ADA_Rama01			X
ADA_Rama03	X	X	X
ADA_Rama06	X	X	
ADA_Rama07		X	

Fuente: elaboración propia.

Realizado lo anterior se consideran los nuevos aprobados y rechazados. Los resultados se encuentran en el Cuadro 4.

CUADRO 4.
RESULTADOS DEL ARBOL DE DECISIÓN

Análisis final-Punto de corte		**Aceptados**	**Rechazados**
>= 691+ nodos finales seleccionados	Buenos	415	1.570
		20,91 %	79,09 %
	Malos	32	309
		9,38 %	90,62 %
	Indeterminados	65	283
		18,68 %	81,32 %

Error tipo I: 79,09 %; Error tipo II: 9,38 %; % Aceptación: 19,62 %.

Fuente: elaboración propia.

Se observa un incremento de 1,47 % en el error tipo II, pero de igual forma, hay un aumento de 3,54 % en el porcentaje de aceptación total, por lo que puede considerarse más adecuado este modelo. Una de las ventajas de utilizar este método radica en que se considera un segundo filtro (árbol de decisión secundario), para reforzar la decisión de aceptar a un cliente. El modelo *scoring* final se encuentra en el Anexo 3, para observar el proceso de originación resultante tras la metodología presentada.

Se requiere mencionar que cada institución financiera debe elegir su punto de corte sobre el árbol de decisión primario, conforme a las políticas de crédito que se tengan. De manera semejante, se tomará la decisión acerca de si es recomendable o no modelar el árbol de decisión secundario, con el fin de mejorar la calidad crediticia de originación contra la parsimonia del proceso.

Finalmente, aunque se ha probado en distintas poblaciones que las variables presentadas resultan ser altamente predictivas, también es cierto que pudiesen no serlo, es más, pudiesen no existir, por lo que se recomienda ampliamente tomar previsiones al respecto.

CONCLUSIONES

El artículo ha revisado a *grosso modo* la literatura correspondiente al desarrollo de los llamados *credit scoring*, con el objetivo de introducir al lector en la utilización y finalidad de dichos modelos en la administración de riesgos, aunque vale la pena mencionar que la misma metodología pudiera ser extendida hacia fines más comerciales como sería la elaboración de modelos de propensión de consumo, en los cuales la población objetivo está en función no de sus costumbres de pago sino de su propensión de un consumo específico.

Por otro lado, se ha presentado una metodología asaz sencilla con la que, primordialmente pequeñas empresas, pueden generar modelos confiables para originar clientes que no tengan experiencia crediticia. Para mostrar dicha sistemática, utilizando información de un pequeño banco mexicano, se emplearon árboles de decisión, por ser una herramienta efectiva para la predicción de probabilidades de incumplimiento, no sólo a nivel de capacidad de discriminación y estabilidad a través del tiempo, sino como una herramienta de fácil entendimiento que permite potencializar sus usos y servir además de predicción, para la planeación de estrategias comerciales de venta de servicios, estrategias de cobranza, entre muchas otras. Por otro lado, se usaron técnicas microeconométricas, específicamente el modelo logit, para calcular los *odds* por el grupo poblacional carente de información crediticia.

Del presente trabajo se pueden extender las siguientes ideas que serán parte de trabajos futuros tanto dentro de un marco teórico como aplicado. Por una parte es posible estudiar si la metodología presentada a lo largo de este trabajo permite

estimaciones consistentes en los parámetros del modelo logit, este punto es realmente muy importante ya que es bien sabido la presencia de heteroscedasticidad en los datos microeconómicos. Por otra parte la metodología presentada puede ser ampliada tras considerar una población que si posea información crediticia en algún buró de crédito.

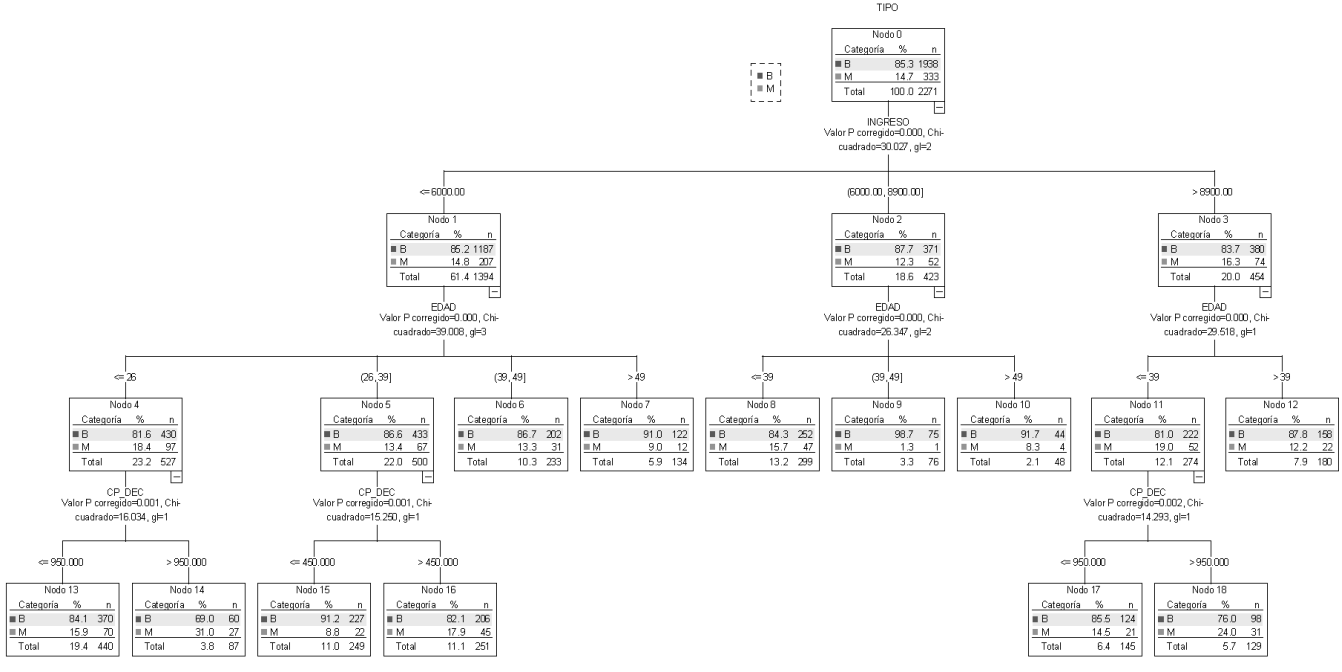
REFERENCIAS BIBLIOGRÁFICAS

1. Andreeva, G. (2005). European generic scoring models using survival analysis. *Journal of the Operational research Society, 57*(10), 1180-1187.
2. Baesens, B. (2003). *Developing Intelligent Systems for Credit scoring using Machine Learning Techniques*. (Tesis doctoral), Katholieke Universiteit Leuven, LIRIS, Louvain, Bel.
3. Bierman, H. y Hausman, W. H. (1970). The credit granting decision. *Management Science, 16*(8), 519-532.
4. Biggs, D., De Ville, B. y Suen, E. (1991). A Method of Choosing Multiway Partitions for Classification and Decision Trees. *Journal of Applied Statistics, 18*(1), 49-62.
5. Bouroche, J. y Tennenhaus, M. (1972). Some segmentation methods. *Metra*, 7, 407-418.
6. Boyle M., Crook J.N., Hamilton R. y Thomas L.C. (1992). *Methods for credit scoring applied to slow payers in Credit scoring and Credit Control*. Oxford: Oxford University Press.
7. Breiman, L., Friedman, J., Olshen, R. y Stone, C. (1984). *Classification and Regression Trees*. Belmont: Wadsworth.
8. Cellard, I., Labbe, B. y Cox, G. (1967). Le programme Elisèe. Presentation et Applicaction. *Metra*, 3, 511-519.
9. Crook, J.N., Edelman, D.B. y Thomas, L.C. (2007). Recent development in consumer credit risk assessment. *European Journal of Operational Research, 183*(3), 1447-1465.
10. Dellaportas, P., Karlis, D. y Xekalaki, E. (1997). *Bayesian analysis of finite poisson mixtures*. Manuscript.
11. Desai V.S., Convay D.G., Crook J.N. y Overstreet G.A. (1997). Credit scoring models in the credit union environment using neural networks and genetic algorithms. IMA J. *Mathematics applied in Business and Industry*, 8, 323-346.
12. Dirickx, Y. y Wakeman, L. (1976). An extension of the Bierman-Hausman model for credit granting. *Management Science, 22*(11), 1229-1237.
13. Duda, R.O. y Hart. P.E. (1973). *Pattern classification and scene analysis*. Nueva York: John Wiley and Sons.
14. Escobar-Mercado, R.M. (1992). *El análisis de segmentación: Concepto y aplicaciones* (Estudios del Centro de Estudios Avanzados en Ciencias Sociales, 1992/31). Madrid: Estudios del Centro de Estudios Avanzados en Ciencias Sociales.
15. Escobar-Mercado, R.M. (1998). Las aplicaciones del análisis de segmentación: El procedimiento CHAID. *Empiria, Revista de Metodología de Ciencias Sociales*, 1, 13-49.

16. Eisenbeis, R.A. (1977). Pitfalls in the application of discriminant analysis in business, finance and economics. *The Journal of Finance, 32*(3), 875-900.
17. Eisenbeis, R.A. (1978). Problems in applying discriminant analysis in credit scoring models. *The Journal of Banking & Finance, 2*(3), 205-219.
18. Fair Issac, C. (2004). *Understanding your credit scoring*. Recuperado de: http://www.myfico.com/downloads/files/myfico_uyfs_booklet.pdf
19. Girault, M.A.G. (2007). *Modelos de Credit Scoring -Qué, Cómo, Cuándo y Para Qué* (MPRA Paper, University Library of Munich). Berlín: MPRA.
20. Hand, D.J. y Henley, W.E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 160*(3), 523-541.
21. Henley W.E. (1995). *Statistical aspects of credit scoring*. Open University Press.
22. Karlis, D. y Rahmouni, M. (2007). Analysis of defaulters' behaviour using the Poisson-mixture approach. *Journal of Management Mathematics, 18*(3), 297-311.
23. Kass, G. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics, 29*(2), 119-127.
24. Kiefer, N. M. y Larson, C. E. (2006). Specification and informational issues in credit scoring. International *Journal of Statistics and Management Systems*, 1, 152-178.
25. Loh, W. y Shih, Y. (1998). Split Selections methods for classification trees. *Statistica Sinica*, 7, 815-840.
26. Magdison, J. y SPSS Inc. (1993). *SPSS for Windows CHAID release 6.0*. SPSS inc.
27. Morgan, J., y Sonquist, J. (1963). Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association, 58*(302), 415-434.
28. Orgler, Y.E. (1971). Evaluation of bank consumer loans with credit scoring models. *Journal of Bank Research, 2*(1), 31-37.
29. Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* USA: Morgan Kaufmann.
30. Rodríguez-Caballero, C.V. (2011). *La inferencia bayesiana en la administración de riesgos*. México DF: Administración de riesgos.
31. Srinivasan, V. y Kim, Y. (1987). The Bierman-Hausman credit granting model: A note. *Management Science, 33*(10), 1361-1362.
32. Srinivasan, V. y Kim, Y. (1987). Credit granting: A comparative analysis of classification procedures. *Journal of Finance, 42*(3), 665-681.
33. Thomas, L. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting, 16*(2), 149-172.
34. Thomas, L., Crook, J. y Edelman, D. (1992). *Credit scoring and credit control*. Oxford: Oxford University Press.
35. Wiginton, J.C. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior. *Journal of Financial and Quantitative Analysis, 15*(3), 757- 770.
36. Yobas M.B., Crook J.N. y Ross P. (2000). Credit scoring using neural and evolutionary techniques. IMA *Journal of Management Mathematics, 11*(2), 111-125.

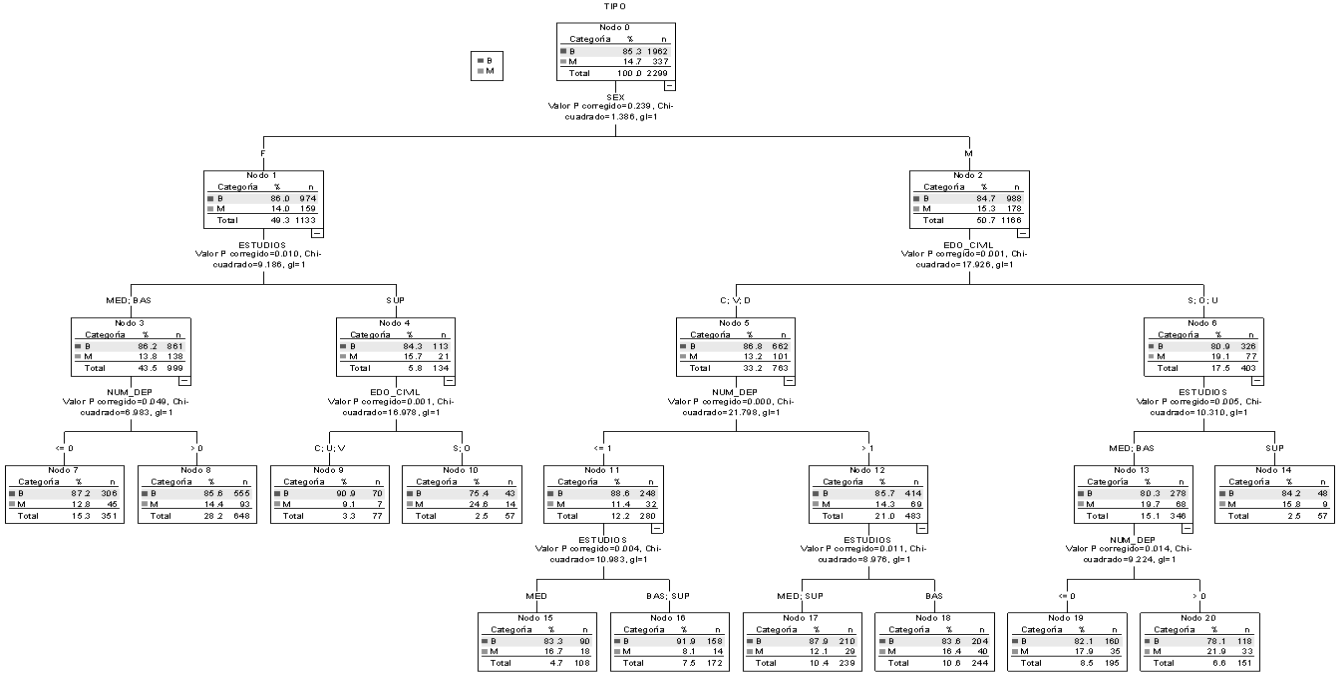
ANEXO 1. ÁRBOLES DE DECISIÓN

GRÁFICA A1.
Árbol de decisión primario construido mediante la técnica Exhaustive CHAID (variable de influencia - capacidad de cliente declarada / ingresos)



Fuente: elaboración propia.

GRÁFICA A2.
Árbol de decisión secundario construido mediante la técnica Exhaustive CHAID usando como variable de influencia la capacidad de cliente declarada / ingresos



Fuente: elaboración propia.

ANEXO 2. REGRESIÓN LOGÍSTICA

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	0	8	1

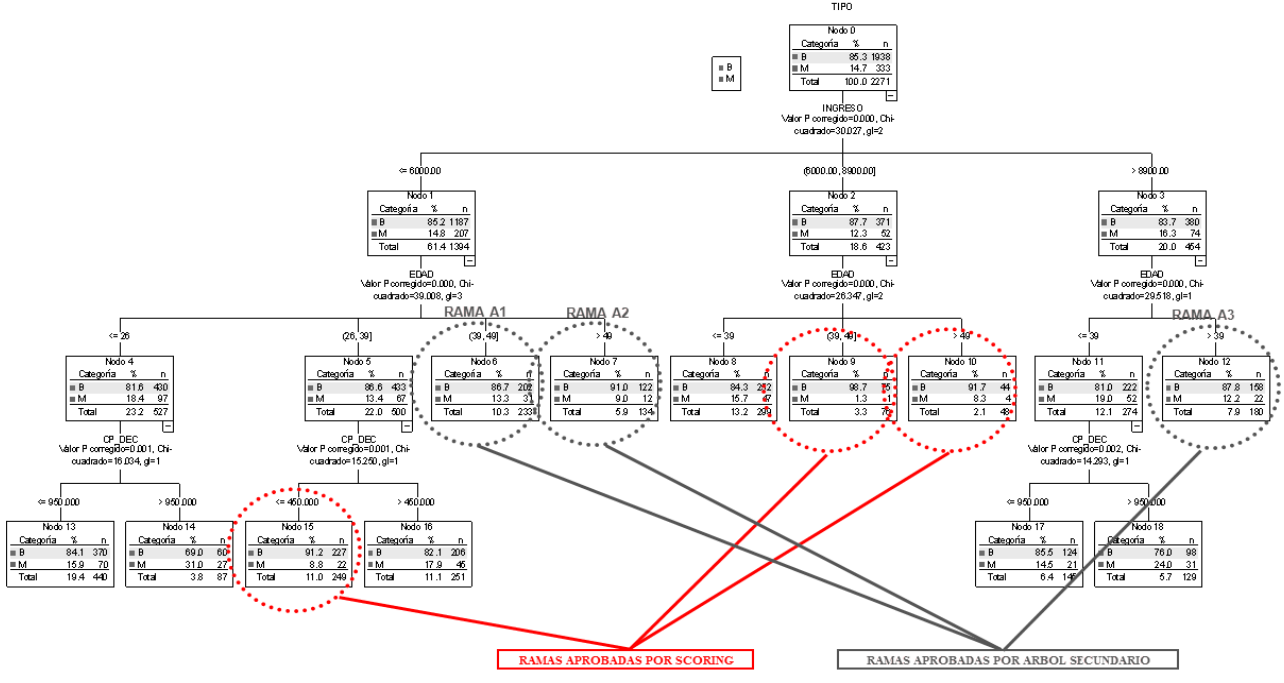
Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95.0% para EXP(B)	
		Inferior	Superior	Inferior	Superior	Inferior	Superior	Inferior	Superior
Paso 1(a)	x1	1.665	0.13	163.185	1	0.0000	5.286	4.094	6.824
	x2	0.799	0.232	11.873	1	0.0010	2.222	1.411	3.5
	x3	2.334	0.223	109.249	1	0.0000	10.318	6.661	15.983
	x4	1.521	0.165	85.465	1	0.0000	4.578	3.316	6.32
	x5	1.874	0.193	94.412	1	0.0000	6.516	4.465	9.51
	x6	2.319	0.303	58.76	1	0.0000	10.167	5.619	18.395
	x7	1.679	0.159	111.705	1	0.0000	5.362	3.927	7.321
	x8	4.317	1.007	18.395	1	0.0000	75	10.428	539.41
	x9	2.398	0.522	21.083	1	0.0000	11	3.952	30.614
	x10	1.814	0.23	62.265	1	0.0000	6.136	3.91	9.63
	x11	1.192	0.196	37.068	1	0.0000	3.294	2.244	4.835
	x12	1.935	0.21	85.051	1	0.0000	6.923	4.589	10.444

a Variable(s) introducida(s) en el paso 1: x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12.

Fuente: elaboración propia.

ANEXO 2. ACEPTACIÓN FINAL DEL MODELO SCORING Y METODOLOGÍA PLANTEADA



Fuente: elaboración propia.