Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA

UNIVERSIDAD AUTÓNOMA METROPOLITANA – IZTAPALAPA
DIVISIÓN DE CIENCIAS BÁSICAS E INGENIERÍA

DISEÑO DE UN MODELO AJUSTADO DE COMPORTAMIENTO PARA RIESGO CREDITICIO

Tesis que presenta
Javier Sotelo Chávez
Para obtener el grado de
Maestro en Ciencias Matemáticas Aplicadas e Industriales

Asesora: Blanca Rosa Pérez Salvador

Jurado Calificador:

Presidente: Dr. Carlos Cuevas Covarrubias
Secretaria: Dra. Blanca Rosa Pérez Salvador
Vocal: Dr. Gabriel Núñez Antonio

México, D.F., Julio 2014

Índice general

Capítulo 1

Introducción

En la industria de los servicios financieros es de suma importancia saber elegir de manera precisa los clientes a los cuales se les otorgará un crédito, para contar con una cartera sana y eficiente, es por esto que se ha vuelto fundamental tener herramientas que determinen adecuadamente quiénes son clientes potenciales para la empresa y cuáles pueden representar un problema. De esta forma es necesario medir el riesgo que conlleva cada cliente y a partir de esto generar una cartera que minimice el riesgo asumido por la empresa. Dicho riesgo se ve reflejado como una pérdida económica, es por ello que al reducir el riesgo admitido, incrementan las utilidades del negocio.

Dentro del sector crediticio se consideran tres grandes grupos de productos: créditos personales, créditos hipotecarios y créditos revolventes. Estos últimos se refieren a las tarjetas de crédito. En este trabajo nos enfocaremos en los créditos revolventes, sin embargo, cabe mencionar que la metodología que se va a utilizar también es aplicable al resto de los productos.

Existe una disciplina en la Teoría de Riesgo llamada *Credit Scoring* la cual clasifica a los clientes en buenos o malos. Los modelos de *Credit Scoring* tratan de pronosticar el potencial incumplimiento los clientes a corto plazo a partir de un análisis de su información crediticia. En el caso en el que se trate de una persona física el tipo de información que se examina es principalmente datos demográficos y datos históricos crediticios, por ejemplo: edad, género, ingresos,

situación laboral, estado civil, situación en el buró de crédito, historial crediticio, índices de morosidad, etc. Mientras que si se trata de una persona moral se analisas varriables como sector de la industria, tamaño de facturación anual, nivel de apalancamiento, líneas de crédito, etc. En este trabajo nos enfocaremos es modelos de *Credit Scoring* diseñados para personas físicas.

La finalidad de los modelos de *credit scoring* es estimar la probabilidad de incumplimiento de pago por parte del cliente a partir de la información disponible. Sin embargo, para hacer más fácil la lectura del resultado, dicha probabilidad es trasladada a un puntaje el cual ayuda a clasificar de manera más ágil a los clientes. Cabe mencionar que el método asocia el mismo riesgo a diferentes clientes con las mismas características (variables). De esta forma, se reduce de manera significativa el tiempo necesario para determinar si una entidad financiera debe aceptar, mantener o rechazar un cliente. Cuando hablamos de rechazar un cliente, lo que sucede es que se rechaza una solicitud en particular relacionada a cierto cliente, pues bien, puede ser que se nos niegue un crédito hipotecario pero que al mismo tiempo se cuente con una línea de crédito con la misma institución. Además, al basarse únicamente en información histórica del cliente, el modelo permite tener una mayor objetividad durante el proceso de evaluación de riesgo de crédito ya que elimina el sesgo que se genera al evaluar la información vía un analista.

Si bien el modelo no pronostica de manera exacta el comportamiento de un cliente, sí es capaz de estimar el comportamiento promedio de un grupo de individuos que tengan características similares. No obstante, el desempeño del modelo está sujeto a la actualización oportuna de los parámetros, es decir, de la información de los de clientes, para que este se encuentre bien calibrado y sus resultados sean óptimos.

Algunas de las técnicas estadísticas utilizadas para desarrollar un modelo de *credit scoring* son: regresión lineal múltiple, regresión logística, árboles de decisión, redes neuronales, entre otras. A partir de estas técnicas se construye una tabla de calificaciones la cual asigna distintos puntajes en diferentes rangos de una variable predictiva y dependiendo del puntaje asignado a cada rango se estima una probabilidad de aceptación o rechazo de un cliente.

Para determinar cuáles son los clientes buenos y malos la empresa debe considerar toda la información que tenga disponible, tanto la propia como la de otras

instituciones, y basar su decisión para clasificarlos con respecto a su estrategia de negocio.

Si utlizamos el 100 % de la base de los clientes que se tienen para construir un modelo de *credit scoring* se corre el riesgo de que el modelo resultante quede sobreajustado, es decir, si en la población se tienen algunos clientes atípicos cuyos resultados sean significativamente distitnos a los del resto de la base, el modelo puede sesgar su respuesta hacia el comportamiento de estos clientes. Para evitar este problema una práctica común es utilizar una muestra aleatoria de los clientes equivalente al 70 % llamada **base de entrenamiento**. Con esta muestra se construye el modelo y posteriormente se evalúa su efectividad con el 30 % restante llamada **base de validación**, véase [Girault 2007].

Debido a que la cartera de clientes se encuentra en constante actualización con el paso del tiempo, es necesario monitorear y evaluar continuamente el modelo para comprobar que se encuentra dentro de los máde clasificación adecuados. Dichos ajustes suelen hacerse de manera anual para mantener actualizada la información y conservar la efectividad del modelo.

El propósito de este trabajo es estudiar diversas disciplinas y técnicas estadísticas utilizadas en la construcción de un modelo de *credit scoring*. Se revisarán distintos conceptos generales de estadística y probabilidad usados en el *credit scoring*. Posteriormente se tomará una cartera de clientes de un banco para construir un modelo utilizando las técnicas estadísticas descritas en el documento y obtener conclusiones a partir de los resultados obtenidos.

En el capítulo 2 se da una explicación sobre las etapas del ciclo del riesgo y los créditos al consumo. Esto es relevante porque el modelo de *Credit Scoring* que se construye está enfocado a una etapa en particular del ciclo de riesgo. Más adelante se construye un modelo utilizando sobre una cartera de clientes compuesta por créditos al consumo.

En el tercer capítulo se describe el modelo individual de riesgo, el modelo colectivo de riesgo y la aproximación de Liapunov para estimar la función de distribución de la pérdida esperada de una cartera de clientes.

En el capítulo 4 se desarrollan algunas de las técnicas utilizadas para construir un modelo de *Credit Scoring,* principalmente la regresión logística, así como

algunas pruebas estadísticas que se utilizan para calibrar un modelo construido con este método.

El objetivo principal de esta tesis se desarrolla en el quinto capítulo. Se trata de un método para construir un modelo ajustado de *Credit Scoring* al cual llamamos Mejor selección. Se realiza el planteamiento del modelo y se explican algunos indicadores que servirán para verificar la efectividad del modelo construido. Para comprobrar su efectividad se contrastan los resultados de un modelo construido con regresión logística y otro bajo la metodología propuesta.

Por último en capítulo 6 se expone una aplicación de los modelos de Comportamiento Crediticio. Para esto se utiliza una base de datos de clientes reales con la cual se construye el modelo y se interpretan los resultados que se obtienen. Junto los los índices de validación, se comparan los resultados del modelo con la realidad para verificar la capacidad discriminatoria del modelo y comprobar su efectividad.

Capítulo 2

Riesgo de Crédito

Para las instituciones financieras es de gran importancia tener las herramientas adecuadas para ser capaces de determinar los criterios bajo los cuales se determinará cómo se otorgan los créditos y a quienes se les otorga. En el mismo sentido es importante que tengan criterios claros para clasificar a los clientes de la institución a partir de su comportamiento crediticio en el mercado. Para esto es necesario poder medir el riesgo que conlleva cada cliente para poder tomar las decisiones adecuadas, desde ampliar las oportunidades de crecimiento de un cliente, hasta cancelar su crédito debido al alto riesgo que este conlleva, para que de esta manera se minimice la pérdida económica de la institución asociada a la cartera de crédito.

Riesgo y Función de Utilidad

La presencia de riesgo supone que las consecuencias que se derivan de cada alternativa disponible no se conocen de antemano, sino que dependen de la ocurrencia de sucesos aleatorios fuera del control de la empresa.

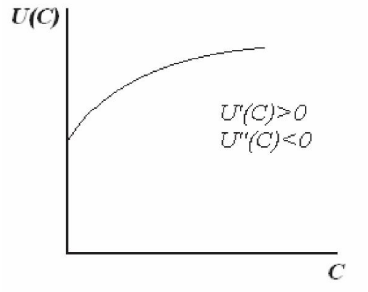
Según [Sarmiento y Velez 2007], dentro de la teoría económica el **riesgo** se interpreta como el peligro de pérdida al cual se enfrenta la empresa ante la incertidumbre sobre el porvenir de la actividad económica en la que invierte. Cuando

se comparan dos situaciones distintas, aquella que conlleva un mayor riesgo, también debe tener relacionado un mayor beneficio, sin embargo, la relación que existe entre estas dos variables no es lineal por lo cual es necesario hacer un análisis del costo-beneficio de los posibles eventos para determinar adecuadamente el nivel de riesgo que se desea aceptar.

Una **función de utilidad** es una función diseñada para reflajar tanto la satisfacción como la preocupación que se tiene ante los cambios en la cantidad de riesgo que se toma. Este concepto indica que tanto riesgo está dispuesto a aceptar un individuo a cambio de cierto beneficio.

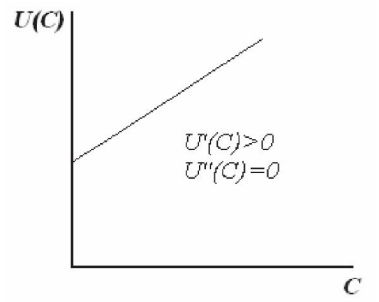
En el caso de las carteras de crédito este nivel de satisfacción suele medirse con la recompensa monetaria que recibe la institución tras haber otorgado un préstamo, una tarjeta o cualquier otro producto de crédito. Las funciones de utilidad se pueden clasificar en tres grandes grupos de acuerdo a las características de consumo crediticio (C) que practique el usuario:

- Utilidad marginal positiva y decreciente: **Aversión al riesgo**.



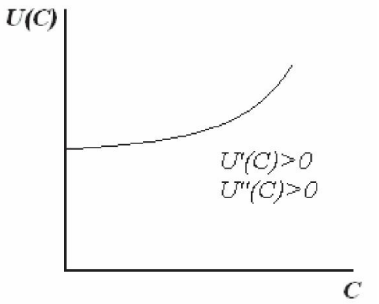
Características: La primer derivada de la función de utilidad es positiva, por lo que la función de utilidad es creciente en el consumo (C), además la segunda derivada es negativa por lo que una unidad extra de consumo genera una mayor utilidad cuando se encuentra en niveles bajos de consumo que cuando está en niveles altos. Este tipo de funciones de utilidad son funciones cóncavas.

- Utilidad marginal positiva y constante: **Neutral al riesgo**.



Características: La primer derivada de la función de utilidad es positiva, por lo que la función de utilidad es creciente en el consumo (C), además la segunda derivada es cero por lo cual la utilidad generada es la misma tanto en niveles bajos como en niveles altos. Estas funciones de utilidad son funciones lineales.

- Utilidad marginal positiva y creciente: **Amante al riesgo**.



Características: La primer derivada de la función de utilidad es positiva, por lo que la función es creciente en el consumo (C), además la segunda derivada es positiva por lo cual la utilidad generada es mayor en niveles altos de consumo que en niveles bajos. Estas funciones de utilidad son funciones convexas.

En [Elizondo 2004] se describe el riesgo crediticio como la posible pérdida económica que asume un agente económico a consecuencia del incumplimiento de las obligaciones contractuales que corresponden a las contrapartes con las que se relaciona, es decir, cuando se presenta la falta de pago de un cliente. En este caso el daño ocasionado es de carácter económico y afecta directamente al capital de la institución, es por esto que la medición adecuada del riesgo crediticio adquiere un carácter fundamental para mantener una cartera de crédito con riesgo controlado.

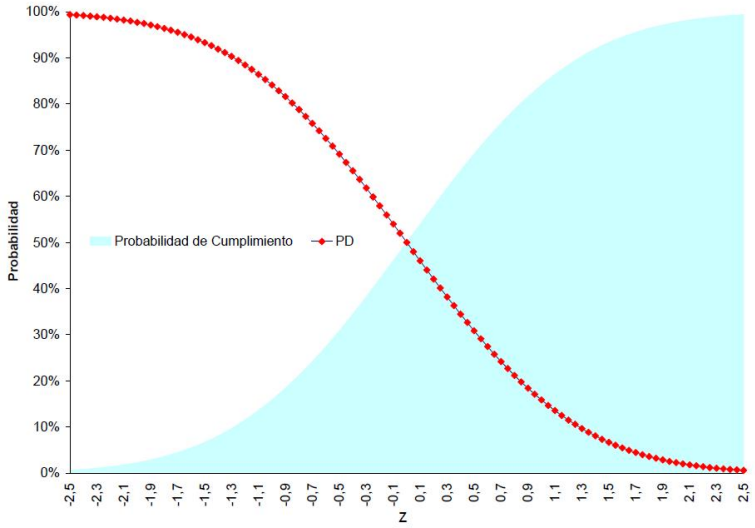
Durante la construcción de modelos de *credit scoring* las instituciones financieras deben precisar un punto de corte que determina cuáles solicitudes serán aceptadas y cuáles rechazadas de acuerdo a su puntuación obtenida. Este punto de corte lo define la empresa a partir de sus metas y expectativas. Por ejemplo, se puede definir un punto de corte "alto" el cual implicará tener una cartera con poco riesgo pero pequeña, lo cual se traduciría en obtener poca rentabilidad, por otro lado, un punto de corte "bajo" determinaría una cartera mucho más grande pero con mayor riesgo, incrementando así la cartera vencida y las pérdidas. En la siguiente tabla se muestra el comportamiento que tendría una institución según sus políticas de riesgo. Fuente: [Girault 2007].

<table>
<tr><th>Score</th><th>Banco Conservador
- minimiza riesgo -</th><th>Banco Estándar</th><th>Banco Agresivo
- maximiza colocaciones -</th></tr>
<tr><td rowspan="5">1000
riesgo
bajo

riesgo
alto
0</td><td>Acepta automáticamente</td><td rowspan="2">Acepta automáticamente</td><td rowspan="3">Acepta automáticamente</td></tr>
<tr><td>Revisión</td></tr>
<tr><td rowspan="3">Rechaza automáticamente</td><td>Revisión</td></tr>
<tr><td rowspan="2">Rechaza automáticamente</td><td>Revisión</td></tr>
<tr><td>Rechaza automáticamente</td></tr>
</table>

En la práctica, cuando una solicitud obtiene un puntaje inferior al corte esta es rechazada automáticamente, mientras que si el puntaje es superior puede pasar directamente al otorgamiento del producto de crédito o bien a un segundo análisis como por ejemplo un estudio socioeconómico. Estos resultados pueden ser desarrollados con información de la propia empresa, información externa o bien pueden llevarse a cabo por un consultor externo.

A la probabilidad que se tiene de que un cliente no realice su pago se le conoce como **Probabilidad de default.** Los modelos de *credit scoring* traducen el riesgo de los clientes en un puntaje que describe la probabilidad de default, de manera que mientras mayor sea el riesgo que conlleva un cliente mayor será su probabilidad de default y por lo tanto se le asigna un puntaje menor. Usulmente se utiliza una medida en la cual a menor puntaje mayor riesgo. La siguiente gráfica muestra la relación que existe entre la probabilidad de default y el score obtenido por un cliente. La parte sobreada muestra la **probabilidad de cumplimiento** que representa el complemento o lo contrario de la probabilidad de default. Fuente: Gutiérrez Girault (2007).



2.1. Ciclo del Riesgo

Los modelos de *credit scoring* son utilizados principalmente por las instituciones financieras para resolver solicitudes de crédito; en este caso se les conoce como modelos reactivos o *application scoring*. También son empleados para administrar el portafolio de créditos, en este caso se conocen como **modelos de Comportamiento** o *behavioural scoring*, y son utilizados principalmente para administrar límites de tarjetas y cuentas corrientes, analizar la rentabilidad de los clientes, ofrecer nuevos productos, monitorear el riesgo y detectar posibles problemas de cobranza. Por último, los modelos desarrollados para calcular la

probabilidad de recuperar un cliente son conocidos como colection scoring. Se utilizan para determinar el valor de la deuda y decidir que técnicas de recaudación utilizar con cada cliente.

El ciclo de riesgo consiste en tres etapas las cuales describen los distintos momentos o estatus de un cliente como receptor de un crédito. Estas etapas son:

Originación: La primera etapa del ciclo corresponde al otorgamiento del crédito. En ella se decide si se le dará o no el crédito a un cliente en particular.

Administración: En esta etapa se ve involucrado el análisis de la cartera. Se desarrollan indicadores como índice de morosidad y pago promedio para determinar qué tan bueno o malo ha sido el comportamiento crediticio de los clientes. A partir de este análisis se les asigna una calificación mediante la cual se resuelve qué tipos de estrategias se aplicarán para cada cliente. Aquellos con buena calificación podrán ser acreedores de un incremento en su línea de crédito o bien al otorgamiento de mejores tasas de interés, mientras que los clientes que obtengan bajas puntuaciones podrían tener mayores tasas de interés, decrementos en sus límites de crédito o hasta perder el mismo. También es importante detectar oportunamente aquellas cuentas que tengan un alto riesgo para poder definir como serán tratadas, ya sea mediante reestructuras del crédito, mayores plazos para pago, etc.

Recuperacíon: Por último, se encuentra la etapa en la cual se pretende recuperar a algunos clientes que dejaron de pagar. Para esto se requiere utilizar algunas técnicas de recaudación. El objetivo es detectar aquellos clientes que pueden pagar su deuda; pero, que por alguna circunstancia ajena dejaron de cumplir con su obligación, de igual manera se pretende incorporarlos a la cartera de clientes vigentes.

Los modelos lineales se han dejado de utilizar debido a que tienen algunas desventajas técnicas, mientras que los modelos probit, logit y la regresión logística poseen una mayor capacidad para clasificar a los clientes adecuadamente. Los modelos no paramétricos, los arboles de decisión, las redes neuronales y los algoritmos genéticos se caracterizan por tener un mejor desempeño cuando se desconoce la relación funcional entre la probabilidad de default y el puntaje.

Las variables que se eligen para el modelo dependen principalmente de la naturaleza del problema, por ejemplo, cuando se trata de un modelo para una cartera

(donde se analizan individuos principalmente) se utilizan variables socioeconómicas, mientras que en el caso de modelos corporativos (análisis de empresas) se utilizan variables de estados contables, proyección de flujos, etc.

En los modelos las variables más comunes son: edad, estad civil, cantidad de personas dependientes, tiempo de permanencia en el domicilio y empleo actual, nivel de escolaridad, si es propietario de la vivienda, egresos mensuales, ingresos mensuales, tipo de ocupación, si cuenta con tarjetas de crédito, cuentas de ahorro, consultas en buró de crédito y calificación. Dichas variables suelen clasificarse en positivas y negativas, es decir, aquellas que contengan información que pueda reducir el puntaje tales como incumplimientos y atrasos de pagos se clasifican con información negativa, mientras que la información positiva se conforma de los pagos a tiempo, montos de préstamos anteriores y tasas de interés.

La experiencia del mercado ha mostrado que aquellos modelos que incluyen ambos tipos de información son mejores que los que sólo utilizan información negativa, incrementando sustancialmente el desempeño de los modelos y así la calidad de las carteras de crédito de las instituciones financieras.

En este trabajo nos enfocaremos en la etapa de la administración. Se estudiarán modelos de *credit scoring*, se analizarán sus ventajas y desventajas, así como su funcionamiento, las hipótesis necesarias para aplicarlos y las características de los resultados. Además se presentarán algunos ejemplos con datos de una cartera de crédito real para poder hacer comparaciones y generar conclusiones sobre los resultados obtenidos.

A continuación se hace una breve descripción de las características de los créditos al consumo y cómo funcionan en la actualidad puesto que es el principal producto crediticio para el que se utiliza el *credit scoring*.

2.2. Crédito al consumo

Es el tipo de financiamiento otorgado principalmente por instituciones bancarias (a través de tarjetas de crédito) o establecimientos comerciales mediante el cual una persona puede adquirir bienes y servicios a crédito. Se otorgan con pocos requisitos y no se considera el historial crediticio de los clientes.

2.2.1. Tarjeta de crédito

Las tarjetas de crédito son instrumentos de identificación, que puede ser de plástico con una banda magnética, con un microchip o con un número en relieve. Las tarjetas de crédito son emitidas por un banco, por una entidad financiera o por grandes tiendas y almacenes; y con ellas autorizan a las personas a cuyo favor fueron emitidas a utilizarla como medio de pago en los negocios adheridos al sistema, mediante su firma y la exhibición de la tarjeta. Las tarjetas de crédito es una forma de financiamiento, por lo tanto, el usuario tiene la obligación de devolver el importe dispuesto y de pagar los intereses, comisiones bancarias y gastos pactados.

Las tarjeta de crédito consisten en una pieza de plástico, cuyas dimensiones y características generales han adquirido absoluta uniformidad, por virtualidad del uso y de la necesidad técnica. El tamaño de la mayoría de las tarjetas de crédito es de 85.60 mm por 53.98 mm (33/8 Pulgada por 21/8 Pulgada) y cumple la norma ISO/IEC 7810 ID-1 que indica el estandar internacional para las tarjetas de identificación electrónica tipo Visa.

Cada tarjeta contiene las identificaciones de la entidad emisora y del afiliado autorizado para emplearla; así como el periodo temporal durante el cual la tarejta mantendrá su vigencia. Contiene también la firma del portador legítimo y un sector con asientos electrónicos perceptibles mediante instrumentos adecuados. Estos asientos identifican esa particular tarjeta y habilitan al portador para disponer del crédito que conlleva el presentarla.

Las tarjetas de crédito aparecieron en los comienzos del siglo XX en los Estados Unidos, en concreto; la idea surgió dentro de las oficinas del Chase Manhattan Bank, a manos de su director, bajo la modalidad de tarjeta profesional, se consideró su uso mayoritario alrededor de la década de los años 1940 y tomó difusión desde la mitad del mismo siglo.

La difusión internacional fue producto del empleo en otras naciones de las tarjetas emitidas en aquel país, y del establecimiento local de sucursales de las emisoras durante las quinta y sexta décadas del siglo XX.

Entre las tarjetas de crédito más conocidas del mercado están: Visa, American Express, MasterCard, Diners Club, JCB, Discover, Cabal, entre otras.

Los usuarios tienen límites con respecto a la cantidad que pueden utilizar de acuerdo a la política de riesgos existente en cada momento y a las características personales y de solvencia económica de cada usuario. Generalmente no se requiere abonar la cantidad total cada mes. En lugar de esto, el saldo (o "revolvente") acumula un interés. Se puede hacer sólo un pago mínimo así como pagar intereses sobre el saldo pendiente. Si se paga el saldo total, no se pagan intereses.

La mayor ventaja de las tarjetas de crédito radica en la flexibilidad que le da al usuario, quien puede pagar sus saldos por completo en su fecha límite mensual o puede pagar sólo una parte. El contrato de la tarjeta establece el pago mínimo y determina los cargos de financiamiento para el saldo pendiente. Las tarjetas de crédito también se pueden usar en los cajeros automáticos o en un banco para servirse de un adelanto de efectivo, aunque a diferencia de las tarjetas de débito, se cobra un interés por la disposición, comisión y, en algunos países, un impuesto porque se trata de un préstamo.

2.2.2. Estatus de los clientes en los pagos

Cuando un cliente compra un artículo con su tarjeta de crédito, la institución crediticia le concede generalmente un mes para liquidar su adeudo sin cobrarle intereses. El cliente puede cubrir una cantidad que va del mínimo establecido por el banco hasta el saldo total antes de la fecha de límite de pago, y el monto del adeudo quedará como saldo. En este esquema el cliente paga anticipadamente los intereses. La empresa que otorga la tarjeta de crédito establece los intereses mensuales que se aplicarán por falta de pago del adeudo (intereses moratorios), el pago mínimo, la fecha de corte y el límite de pago.

Fecha límite de pago. Es la fecha establecida por la institución para pagar los adeudos sin que se cobren intereses moratorios, ni cargos por cobranza.

Fecha de Corte. Es un día de cada mes en el cual la institución financiera revisa el estado de cuenta del cliente en particular. También se le conoce como fecha de facturación, en este día se calcula el saldo entre la fecha

de corte actual y la fecha de corte inmediata anterior. Este intervalo de tiempo lo llamaremos periodo de corte, dependiendo del mes suelen ser de treinta días, entre estas fechas no se aplican intereses moratorios.

Pago mínimo. Es el porcentaje mínimo aplicado al monto del adeudo, que debe cubrirse antes de la fecha límite de pago.

2.2.3. Posibles condiciones de un cliente de acuerdo a su estatus de pago

Cuando un cliente paga su adeudo entre la fecha de corte y la fecha límite de pago, no se le cobran intereses de mora sobre su saldo, este periodo resulta ser de alrededor de 23 días, según la institución de crédito, en estas condiciones y se considera un cliente al corriente.

Cuando el cliente paga el mínimo entre la fecha de corte y la fecha límite de pago, el monto neto de su deuda se carga al siguiente periodo. El cliente aún no está al corriente pero tampoco se le considera moroso.

Si el cliente no cubre su deuda, ni tampoco realiza el pago mínimo a la fecha límite de pago, se le empieza a localizar para recordarle su adeudo e incurre en gastos de cobranza. Esto se realiza entre la fecha límite de pago y la próxima fecha de corte, aproximadamente una semana y se le conoce como tiempo preventivo. Los gastos de cobranza se cargarán al siguiente periodo de corte. Si el cliente paga la totalidad o el mínimo en prevent pasa a tomar el estatus current. No se le aplicarán intereses moratorios, pero no evitará el cobro del call center para localizarlo.

Si durante el periodo de corte se paga menos del mínimo, no se considera este pago como cumplimiento de la obligación, por lo que avanza a un pago vencido. Se le cargan intereses de moratoria sobre todo el monto y toma el estatus de cliente moroso. En la fecha de facturación se calcula el nuevo saldo y se empiezan a contar los días de retraso. Se envía al grupo de clientes que tienen de 1 a 29 días de moratoria, llamado “Bucket 1” o canasta B1. Esta información es enviada al buró de crédito. Las mismas acciones son aplicadas a un cliente que no realizó pago alguno. Si se paga más del mínimo y menos del total del saldo facturado se cargan intereses revolventes.

La cartera de crédito se clasifica en las siguientes categorías:

Calificación	Tiempo en mora
Canasta 0 (B0)	0 días (al corriente)
Canasta 1 (B1)	1 a 29 días
Canasta 2 (B2)	30 a 59 días
Canasta 3 (B3)	60 a 89 días
Canasta 4 (B4)	90 a 119 días
Canasta 5 (B5)	120 a 149 días
Canasta 6 (B6)	150 a 179 días
Canasta 7 (B7)	Mayor o igual a 180 días

Por ejemplo, supongamos que un individuo se le otorga una tarjeta de crédito departamental el 2 de enero (ver figura). Su fecha límite de pago es el día 18 de cada mes, con facturación los días 25 de cada mes. El 5 de enero realiza compras con un monto de $5000 .

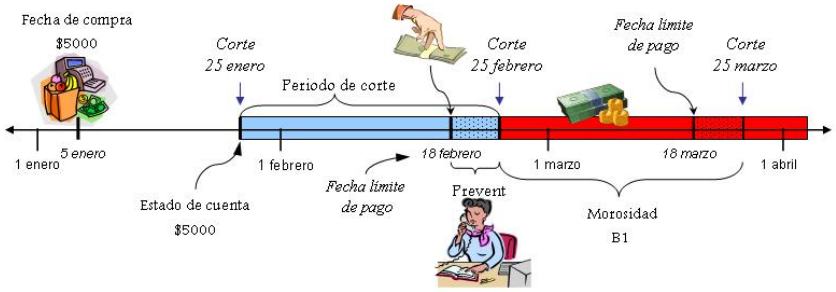


Figura 1.3: Estructura general en los eventos y tiempos asociados a una tarjeta de crédito.

Su primera factura se emitirá el 25 de enero con un saldo de $5000. Y tendrá como fecha límite de pago el 18 de febrero, sin que se le haga algún cargo adicional. En el tiempo prevent del 18 y 25 de febrero se le aplicarán técnicas de cobranza. Los gastos de cobranza se facturarán hasta el 25 de marzo. Si el cliente

no paga el mínimo, a partir del 26 de febrero se contarán los días de moratoria. Del 26 de febrero al 25 de marzo se encuentra en la canasta B1 (bucket 1), del 25 de marzo al 25 de abril la canasta B2 y así sucesivamente hasta B6. Es común considerar que después de B6 se cae en perdida o write off, esto es, cuando un cliente tiene más de 180 días de mora.

Cuando se paga con tarjeta en el comercio, el cobrador suele pedir una identificación (identificación personal, permiso de conducir, etc.) y exige la firma del pagaré o voucher para acreditar que se es propietario de la tarjeta. Existen algunas excepciones donde no se solicita firmar el recibo, a éste sistema se le denomina "autorizado sin firma" y suele utilizarse en comercios con grandes aglomeraciones de gente, como lo son cines, restaurantes de comida rápida y otros lugares similares. En algunos países se solicita el ingreso de un NIP para autorizar las compras de manera presencial.

En caso de uso fraudulento hay que dar aviso a la entidad financiera o tienda que le da la tarjeta pidiendo que anule el cargo y seguir los trámites de cada institución. El emisor de la tarjeta debe demostrar que la compra ha sido hecha por el propietario.

Las compras con tarjeta de crédito pueden tener diversos seguros sobre el saldo financiado lo cual protege al propietario de un mal uso de la misma, dándole así una mayor seguridad para utilizar este producto crediticio.

Capítulo 3

Modelos de Riesgo Crediticio

Un crédito es el otorgamiento de un bien por parte de un inversor en favor de un deudor para que haga uso de ese bien con el compromiso de regresarlo en un tiempo pactado de antemano junto con una ganancia como pago por el uso del bien. Cuando un inversor otorga un crédito a un deudor existe la posibilidad de que ocurra una pérdida si el deudor no cumple plenamente con las obligaciones financieras acordadas en el contrato, en relación al tiempo, a la forma, o a la cantidad a pagar. Sin embargo, en términos generales, también se puede definir como riesgo de crédito la disminución del valor de los activos debido al deterioro de la calidad crediticia de la contrapartida, incluso en el caso en que la contrapartida cumpla totalmente con lo acordado. Por lo tanto, la calidad del riesgo puede estar determinada tanto por la probabilidad de que se produzca el incumplimiento del contrato, como por la reducción de las garantías. En este capítulo se describen el Modelo Individual de Riesgo y el Modelo Colectivo de Riesgo, véase [Rincón 2012].

3.1. Modelo individual de riesgo

Las instituciones que otorgan créditos buscan hacerlo a personas cuya probabilidad de incumplimiento sea pequeña. Se sabe que la probabilidad de incumplimiento de los deudores es diferente de una persona a otra. En este sentido,

asumiendo que las créditos son independientes, el riesgo de la cartera es la suma de los riesgos individuales.

Sea p_i la probabilidad de que el i-esimo cliente deje de cubrir sus obligaciones con la entidad financiera, y q_i la probabilidad que las cumpla, de esta manera se satisface la relación $p_i + q_i = 1$. Se define la variable aleatoria

$$J_i = \begin{cases} 1 & \text{si el cliente } i \text{ cae en mora} \\ 0 & \text{si el cliente } i \text{ cumple con los pagos} \end{cases}$$

La variable J_i es Bernoulli con parámetro p_i.

A su vez, el monto de la perdida sufrida por el incumplimiento del cliente i se modela con la variable aleatoria C_i, la cual depende la variable M_i que denota el monto otorgado en el crédito ($0 \leq C_i \leq M_i$), por lo que la variable aleatoria que determina la perdida sufrida por el otorgamiento de un crédito a la persona i está dado por

$$X_i = J_i C_i$$

El modelo individual de riesgo indica que el monto total sufrido por incumplimiento en una cartera compuesta por n clientes es igual a:

$$S = \sum_{i=1}^{n} X_i$$

donde $J_1, \ldots, J_n, C_1, \ldots, C_n$ son variables aleatorias independientes con $C_i > 0$ y J_i con distribución $Ber(p_j)$. La variable aleatoria S es el monto que afronta la empresa crediticia por el incumplimiento de los clientes durante el periodo. Además se cumple que lo siguiente:

1. $E(S) = \sum_{i=1}^{n} p_i E(C_i)$
2. $Var(S) = \sum_{i=1}^{n} p_i \left[Var(C_i) + q_i E^2(C_i) \right]$

Demostración:

1. Dado que las variables son independientes entonces,

$$E(S) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} E(J_i C_i) \sum_{i=1}^{n} E(J_i) E(C_i) = \sum_{i=1}^{n} p_i E(C_i).$$

2. Primero calculemos la $Var(X_i)$

$$\begin{aligned} Var(X_i) = Var(J_i C_i) &= E\left(J_i^2 C_i^2\right) - E^2(J_i C_i) \\ &= p_i E\left(C_i^2\right) - p_i^2 E^2(C_i) \\ &= p_i\left[Var(C_i) + E^2(C_i)\right] - p_i^2 E^2(C_i) \\ &= p_i Var(C_i) + p_i q_i E^2(C_i) \end{aligned}$$

por lo tanto

$$Var(S) = \sum_{i=1}^{n} Var(X_i) = \sum_{i=1}^{n} Var(J_i C_i) = p_i\left[Var(C_i) + q_i E^2(C_i)\right] \quad \blacksquare$$

Es importante conocer una función de distribución para el monto de la perdida sufrida S. Suponiendo que la probabilidad de incumplimiento es igual para cada cliente y con un número de clientes n grande se puede utilizar el teorema central del límite para aproximar la distribución de S mediante la distribución normal de la siguiente manera:

$$P(S \leq z) = P\left(\frac{S - E(S)}{\sqrt{Var(S)}} \leq \frac{z - E(S)}{\sqrt{Var(S)}}\right) \approx \Phi\left(\frac{z - E(S)}{\sqrt{Var(S)}}\right).$$

Esta aproximación puede ser adecuada para ciertos riesgos pero tiene la desventaja de que asigna una probabilidad positiva al intervalo $(-\infty, 0)$, lo cual no es consistente con el hecho de que $S \geq 0$; sin embargo, dado que la distribución $N\left(\mu, \sigma^2\right)$ se concentra principalmente en el intervalo $(\mu - 4\sigma, \mu + 4\sigma)$, cuando la esperanza y la varianza de S son tales que $E(S) - 4\sqrt{Var(S)} \geq 0$, la probabilidad asignada a la parte negativa del eje es sumamente pequeña.

3.2. Modelo Colectivo de Riesgo

Considere una cartera de créditos de tamaño no determinado. Sea N la variable aleatoria que denota el número de créditos con incumplimiento y sean las variables positivas $Y_1, \ldots, Y_N$ los montos de estos incumplimientos. El número de incumplimientos y el monto de estos son variables aleatorias independientes entre sí y además comparten la misma distribución de probabilidad.

El modelo colectivo de riesgo indica que el monto total sufrido por incumplimiento en la cartera compuesta es igual a:

$$S = \sum_{i=1}^{N} Y_i$$

y la esperanza y la varianza de S son:

1. $E(S) = E(N) E(Y)$
2. $Var(S) = Var(N) E^2(Y) + Var(Y) E(N)$

Demostración:

1. Hay que condicionar primero sobre el valor de N y posteriormente utilizar la hipótesis de independencia,

$$\begin{aligned} E(S) &= \sum_{n=0}^{\infty} E\left(\sum_{i=1}^{N} Y_i \,|\, N = n\right) P(N = n) \\ &= \sum_{n=0}^{\infty} E\left(\sum_{i=1}^{n} Y_i \,|\, N = n\right) P(N = n) \\ &= \sum_{n=0}^{\infty} nE(Y_i) P(N = n) \\ &= E(N) E(Y). \end{aligned}$$

2. Primero calculemos $E\left(S^2\right)$ condicionando sobre el valor de N,

$$E\left(S^2\right) = \sum_{n=0}^{\infty} E\left(\left(\sum_{i=1}^{N} Y_i\right)^2 \mid N = n\right) P\left(N = n\right)$$

$$= \sum_{n=0}^{\infty} E\left(\left(\sum_{i=1}^{n} Y_i\right)^2 \mid N = n\right) P\left(N = n\right)$$

$$= \sum_{n=0}^{\infty} E\left(\left(\sum_{i=1}^{n} Y_i\right)^2\right) P\left(N = n\right)$$

$$= \sum_{n=0}^{\infty} \left[\sum_{i=1}^{n} E\left(Y_i^2\right) + \sum_{\substack{i,k=1 \\ i \neq k}}^{n} E\left(Y_i\right) E\left(Y_k\right)\right] P\left(N = n\right)$$

Asumiendo que las variables Y_i tienen distribuciones identicas,

$$= \sum_{n=0}^{\infty} nE\left(Y^2\right) P\left(N = n\right) + = \sum_{n=0}^{\infty} n\left(n-1\right) E^2\left(Y\right) P\left(N = n\right)$$

$$= E\left(N\right) E\left(Y^2\right) + E\left(N\left(N-1\right)\right) E^2\left(Y\right).$$

Ahora podemos calcular la varianza de S,

$$Var\left(S\right) = E\left(S^2\right) - E^2\left(S\right)$$

$$= E\left(N\right) E\left(Y^2\right) + E\left(N\left(N-1\right)\right) E^2\left(Y\right) - E^2\left(N\right) E^2\left(Y\right)$$

$$= E\left(N\right)\left[E\left(Y^2\right) - E^2\left(Y\right)\right] + \left[E\left(N^2\right) - E^2\left(N\right)\right] E^2\left(Y\right)$$

$$= E\left(N\right) Var\left(Y\right) + Var\left(N\right) E^2\left(Y\right) \quad \blacksquare.$$

Con este resultado general podemos obtener distintos resultados dependiendo de como se distribuya la variable aleatoria N. Sin pérdida de generalidad consideremos que $\mu = E\left(Y\right)$ y $\mu_2 = E\left(Y^2\right)$. Por ejemplo,

- Si $N \sim Bin(n,\, p) \Rightarrow E\left(S\right) = np\mu$ y $Var(S) = np\left(\mu_2 - p\mu^2\right)$.
- Si $N \sim Poisson(\lambda) \Rightarrow E\left(S\right) = \lambda\mu$ y $Var(S) = \lambda\mu_2$.

Adicionalmente, podemos calcular la función de distribución de S de la siguiente manera.

Sea G la función de distribución de la variable Y y $\mu = E(Y)$, entonces la función de distribución de la variables aleatoria S es igual a:

$$F(x) = \sum_{n=0}^{\infty} G^{*n}(x) P(N = n)$$

donde $G^{*n} = P(Y_1 + \cdots + Y_n \leq x)$.

Demostración:

$$F(x) = \sum_{n=0}^{\infty} P(S \leq x \,|\, N = n) P(N = n)$$

$$= P(S \leq x \,|\, N = 0) P(N = 0) + \sum_{n=1}^{\infty} P(Y_1 + \cdots + Y_n \leq x) P(N = n)$$

$$= G^{*0} P(N = 0) + \sum_{n=1}^{\infty} G^{*n}(x) P(N = n)$$

$$= \sum_{n=0}^{\infty} G^{*n}(x) P(N = n) \quad \blacksquare$$

Este modelo cuenta con la desventaja de suponer que la pérdida ésperada de cada incumplimiento tiene la misma distribución para todos, lo cual no necesariamente ocurre. Es por esto que surge la necesidad de buscar otra aproximación que nos permita obtener un resultado donde la distribución de los incumplimientos no sea la misma para todos.

3.3. Aproximación de Liapunov

La forma más conocida del teorema central del límite supone que si W_1, W_2, ..., W_n son variables aleatorias independientes e identicamente distribuidas, con media

μ y varianza σ^2, entonces la variable aleatoria $\frac{\sqrt{n}(\bar{W}-\mu)}{\sigma}$ converge en distribución a una normal estándar cuando $n \to \infty$. Sin embargo, en el caso de una cartera de deudores, ni los créditos son por la misma cantidad, ni la probabilidad de incumplimiento de cada cliente es la misma, por lo que no se satisface que los sumando sean igualmente distribuidos, de esta forma el teorema central del límite no se aplica. Afortunadamente, existen otras versiones más generales del teorema central del límite, y una de ellas es la propuesta por Liapunov (1900-1901), véase [Rényi 1976].

Teorema central del límite (Liapunov): Sean $W_1, W_2, \ldots, W_n$ variables aleatorias independientes cuyos tres primeros momentos existen. La media y la varianza de W_k son $m_k = E(W_k)$, $\sigma_k^2 = V(W_k) < \infty$; el momento centrado de tercer orden y el momento centrado absoluto de tercer orden son $a_k = E(W_k - m_k)^3$, $b_k = E\left(|W_k - m_k|^3\right)$, respectivamente. Se definen las variables

$$s_n = \sqrt{\sum_{k=1}^{n} \sigma_k^2} \text{ y } B_n = \sqrt[3]{\sum_{k=1}^{n} b_k}$$

Si se verifica la siguiente condición (condición de Liapunov)

$$\lim_{n\to\infty} \frac{B_n}{s_n} = 0$$

entonces

$$\lim_{n\to\infty} F_n(x) = \Phi(x)$$

donde F_n es la función de distribución de

$$S_n^* = \frac{\sum_{k=1}^{n} (W_k - m_k)}{s_n} \quad \blacksquare$$

Proposición: El monto de la perdida esperada $S = \sum_{i=1}^{n} W_i$, satisface los supuestos del teorema central del límite de Liapunov. Consideremos que cada crédito tiene un monto otorgado equivalente a M_i.

Demostración

- Los tres primeros momentos de X_k existen.
  Dado que $0 \leq W_i \leq M_i$, entonces sus primeros tres momentos son $E(W_i) \leq M_i$, $E\left(W_i^2\right) \leq M_i^2$ y $E\left(W_i^3\right) \leq M_i^3$es decir, son finitos.

- S satisface la propiedad de Liapunov.
  En el caso de un crédito existen números M y σ_0 tales que $M \geq M_i$ y $\sigma_0 \leq V(W_i)$ para toda $i = 1, 2, \ldots, n$, esto se justifica con el hecho de que a nadie le prestan más de una cantidad fija y conocida de dinero, (el número M y la varianza siempre es positiva, solo pedimos que sea mayor a un número positivo fijo).

  Dado M y σ_0, se tiene que $b_i \leq M_i^3 \leq M^3$ y en consecuencia $B_n \leq \sqrt[3]{n}M$, por otro lado se tiene que $s_n \geq \sqrt{n}\sigma_0$, y por lo tanto se satisface la relación

$$\lim_{n\to\infty} \frac{B_n}{s_n} \leq \lim_{n\to\infty} \frac{\sqrt[3]{n}M}{\sqrt{n}\sigma_0} = \frac{M}{\sigma_0} \lim_{n\to\infty} n^{-1/6} = 0.$$

  Por lo tanto, si la cartera de clientes es suficientemente grande podemos asumir que la variable $(S_n - E(S_n)) / \sqrt{V(S_n)}$ se distribuye aproximadamente normal con media 0 y varianza 1, véase [Rényi 1976] ■

Capítulo 4

Métodos de Clasificación

El proceso de otorgamiento de un crédito se reduce a elegir entre aceptar o rechazar la solicitud de un cliente. El objetivo del *credit scoring* es ayudar a tomar esta decisión determinando cuál es la regla adecuada que se debe aplicar para tomar esta decisión a partir de los resultados obtenidos por otros clientes en el pasado. Como sólo se debe elegir entre aceptar y rechazar una solicitud, se trata de clasificar a los clientes en dos grupos, buenos y malos. Se consideran **buenos** clientes aquellos que tienen un comportamiento crediticio aceptable según la empresa y no presentarán incumplimientos mientras que los **malos** clientes son aquellos que se atrasarán en sus pagos y la empresa habría preferido rechazar su solicitud de crédito.

Es importante notar que exite un sesgo en este enfoque debido a que la información crediticia que se tiene sólo muestra el comportamiento de aquellos clientes que fueron aceptados, de tal forma que la muestra es representativa para aquellos que fueron aceptados y no para aquellos que solicitaron un crédito.

Sea $X = (X_1, X_2, \ldots, X_p)$ el conjunto de p variables aleatorias que describen la información disponible de un cliente que solicita un crédito. A estas variables aleatorias se les conoce como **características**. Cuando se cuenta con el valor de cada una de estas variables se denota $x = (x_1, x_2, \ldots, x_p)$ y se les llama **atributos**. Por ejemplo, una caracterísitca es su estado civil, mientras que el atributo asociado puede ser casado, soltero, viudo, divorciado, etc.

A continuación se muestran algunas técnicas utilizadas para la construcción de un modelo de *credit scoring*.

4.1. Regresión Lineal

En este caso, se trata de encontrar la mejor combinación lineal de las características que cumpla lo siguiente:

$$p_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon_i$$

donde$(X_1, X_2, \ldots, X_k)$ son las variables explicativas o independientes y $(\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$ son los regresores o parámetros que miden la influencia que tiene cada variable explicativa sobre la varible dependiente p_i la cual indica la probabilidad de default que tiene el $i-ésimo$ cliente.

Los supuestos que se deben cumplir al utlizar una regresión lineal son los siguientes:

1. El valor esperado de la perturbación o error ϵ debe ser cero, i.e. $E(\epsilon_i) = 0$.
2. Se debe cumplir la característica de la Homocedasticidad, i.e. $var(\epsilon_i) = \sigma^2$ para todo i.
3. No correlación entre los errores, i.e. $cov(\epsilon_i, \epsilon_j) = 0$ para toda $i \neq j$.
4. El error se distribuye como una normal con media cero y varianza σ^2, i.e. $\epsilon \sim N(0, \sigma^2)$.

Sin pérdida de generalidad podemos separar el total de clientes n en 2 grupos n_G y n_B tal que $n_G + n_B = n$. Si suponemos que los primeros n_G clientes de la muestra son buenos, entonces tenemos que $p_i = 0$ para $i = 1, \ldots, n_G$. De manera recíproca, si denominamos al resto de la muestra n_B como los clientes malos, entonces $p_i = 1$ para $i = n_G + 1, \ldots, n_G + n_B$.

En la regresión lineal se eligen los coeficientes que minimizan el error cuadrático medio de la ecuación anterior. Esto es equivalente a minimizar:

$$\sum_{i=1}^{n_G}\left(1-\sum_{j=0}^{k}\beta_j x_{ij}\right)^2+\sum_{i=n_G+1}^{n_G+n_B}\left(\sum_{j=0}^{k}\beta_j x_{ij}\right)^2$$

Dado que la variable dependiente p_i es una densidad de probabilidad, cualquier estimación mayor a uno o menor a cero sería un resultado erroneo. La regresión lineal no es capaz de limitar la estimación a un cierto intervalo, por lo cual no es apropiado utilizar esta metodología para la construcción de modelos de *credit scoring.*

4.2. Regresión Logística

Sea $\mathbf{x} = (X_1, X_2, \ldots, X_k)$ el vector de variables que contienen la información crediticia del cliente y p_i la variable que deseamos pronosticar. A las variables contenidas en $\mathbf{x}$ les llamaremos **variables independientes** o **predictivas** y a p_i **variable dependiente** o **de respuesta**. Además contamos con la variable binaria y que indica si a qué grupo pertenece, en este caso buenos o malos clientes.

A diferencia del modelo de regresión lineal, el modelo de regresión logística no requiere de los supuestos de normalidad de los errores de observación ni tampoco necesita que las variables predictivas sean continuas. De esta forma, el modelo de regresión logística es útil para predecir variables con valores discretos, categóricos, ordinales o no ordinales.

Las variables predictivas tampoco presentan restricciones, pudiendo ser estas discretas, continuas, cualitativas o cuantitativas. La capacidad predictiva del modelo se valora comparando cómo clasifica a un grupo de individuos a partir de sus características entre buenos y malos vs. su verdadero status observado.

La hipótesis fundamental del enfoque logístico para discriminar es que el cociente de log-verosimilitud es linear, es decir

$$\ln\left\{\frac{L\left(\mathbf{x}\,|\,y_i=0\right)}{L\left(\mathbf{x}\,|\,y_i=1\right)}\right\}=\beta_0+\beta^T\mathbf{x}_i$$

donde $\beta^T = (\beta_1, \ldots, \beta_k)$ corresponde a los coeficientes de las variables predictivas.

Con esto podemos calcular las probabilidades posteriori de la forma

$$Pr\left(y_i = 0 \,|\, \mathbf{x_i}\right) = \frac{e^{\beta_0 + \ln\, \alpha + \beta^T \mathbf{x_i}}}{1 + e^{\beta_0 + \ln\, \alpha + \beta^T \mathbf{x_i}}}$$

donde $\alpha = \frac{\Pi_0}{\Pi_1}$ y Π_s es la proporción clientes que de la cartera tales que $y = s$ para $s = 1, 2$. Una vez que β_0, β^T y α son estimadas la regla de decisión es muy simple ya que depende solo de la función lineal $\beta_0 + \ln\, \alpha + \beta^T \mathbf{x}$, ver [Krishnaiah 1987].

4.2.1. Estimación de Parámetros de la Regresión Logística

Usualmente la estimación de parámetros de un modelo logit se realiza utilizando el método de máxima verosimilitud. Como la variable predictiva y_i toma los valores 1 con probabilidad p_i y 0 con probabilidad $1 - p_i$, esta tiene una distribución de probabilidad Bernoulli. Entonces

$$P\left(y_i\right) = p_i^{y_i} \left(1 - p_i\right)^{1 - y_i} \quad \text{con } \; y_i = 0, 1.$$

La función de verosimilitud para una muestra aleatoria de n datos se calcula de la siguiente forma

$$MV\left(y_1, \ldots, y_n\right) = P\left(y_1, \ldots, y_n\right) = \prod_{i=1}^{n} p_i^{y_i} \left(1 - p_i\right)^{1 - y_i}$$

aplicando logaritmo natural se tiene

$$Log\, MV\left(\mathbf{y}\right) = Log\, P\left(\mathbf{y}\right) = \sum_{i=1}^{n} y_i log\, p_i + \sum_{i=1}^{n} \left(1 - y_i\right) log\, \left(1 - p_i\right)$$

quedando la función de log verosimilitud como sigue

$$Log\, P\left(\mathbf{y}\right) = \sum_{i=1}^{n} y_i log\, \left(\frac{p_i}{1 - p_i}\right) + \sum_{i=1}^{n} log\, \left(1 - p_i\right).$$

Si consideramos $\beta^T = (\beta_0, \ldots, \beta_p)$ y $\mathbf{x}_i^T = (1, x_{i1}, \ldots, x_{ip})$ para escribir el modelo de la forma

$$log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i^T\beta.$$

Sustituyendo esto en la función de log verosimilitud obtenemos la función en términos de los parámetros β dada por

$$L(\beta) = \sum_{i=1}^{n} y_i \mathbf{x}_i^T\beta - \sum_{i=1}^{n} log\left(1 + e^{\mathbf{x}_i^T\beta}\right).$$

Para obtener los estimadores β de máxima verosimilitud hay que derivar la función $L(\beta)$ con respecto de cada uno de los parámetros β_j con $j = 1, 2, \ldots, p$ e igualamos a cero. En términos matriciales esto es

$$\begin{matrix} \frac{\partial L(\beta)}{\partial \beta_0} \\ \frac{\partial L(\beta)}{\partial \beta_1} \\ \vdots \\ \frac{\partial L(\beta)}{\partial \beta_j} \\ \vdots \\ \frac{\partial L(\beta)}{\partial \beta_p} \end{matrix} = \begin{bmatrix} \sum_{i=1}^n y_i (1) \\ \sum_{i=1}^n y_i x_{i1} \\ \vdots \\ \sum_{i=1}^n y_i x_{ij} \\ \vdots \\ \sum_{i=1}^n y_i x_{ip} \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^n (1)\left(\frac{e^{\mathbf{x}_i^T\beta}}{1+e^{\mathbf{x}_i^T\beta}}\right) \\ \sum_{i=1}^n x_{i1}\left(\frac{e^{\mathbf{x}_i^T\beta}}{1+e^{\mathbf{x}_i^T\beta}}\right) \\ \vdots \\ \sum_{i=1}^n x_{ij}\left(\frac{e^{\mathbf{x}_i^T\beta}}{1+e^{\mathbf{x}_i^T\beta}}\right) \\ \vdots \\ \sum_{i=1}^n x_{ip}\left(\frac{e^{\mathbf{x}_i^T\beta}}{1+e^{\mathbf{x}_i^T\beta}}\right) \end{bmatrix}$$

cada una de estas derivadas están expresadas en un vector columna de la forma

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \left(\frac{e^{\mathbf{x}_i^T\beta}}{1 + e^{\mathbf{x}_i^T\beta}}\right)$$

igualando al vector cero y despejando se tiene que

$$\sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} x_i \left(\frac{e^{\mathbf{x}_i^T\beta}}{1 + e^{\mathbf{x}_i^T\beta}}\right) = \sum_{i=1}^{n} x_i p_i.$$

Si $\hat{\beta}$ es el vector de parámetros que resuelve el sistema de ecuaciones matricial expuesto anteriormente, calculamos p_i en términos de esos estimadores y de aquí

se obtiene una estimación para y_i, tal que $\hat{y_i} = \hat{p_i}$. De esta manera se tiene el siguiente resultado

$$\sum_{i=1}^{n} y_i x_{ij} = \sum_{i=1}^{n} x_{ij} \hat{y_i}$$

y

$$\sum_{i=1}^{n} x_{ij} e_i = \sum_{i=1}^{n} x_{ij} \left(y_i - \hat{y_i} \right) = 0$$

donde e_i representa los residuos del modelo, los cuales deben ser ortogonales al espacio de observaciones $\mathbf{x}$, de manera similar a como se hacer en la regresión lineal utilizando los mínimos cuadrados. No obstante, podemos observar que el sistema de ecuaciones matricial es no lineal por lo cual se utilizan métodos numéricos como por ejemplo *Fisher Scoring* o *Newton-Raphson* para resolverlo.

4.3. Pruebas estadísticas al modelo Logístico

Una de las características deseables de los modelos utilizados es que sus estimadores tengan capacidad discriminatoria. Para medir la capacidad discriminatoria se aplican diferentes técnicas de prueba que se explican a continuación.

4.3.1. Devianza

Se trata de una generalización de la idea de utilizar la suma de los cuadrados de los residuales en mínimos cuadrados para casos donde se utiliza la máxima verosimilitud.

Considere la función $D\left(\beta\right) = -2\log\left(\beta\right)$. Ha esta función se le conoce como devianza, ver [Thomas 2000].

Si desglosamos la función del modelo logístico en la ecuación anterior se tiene lo siguiente

$$D\left(\beta\right) = 2\sum_{i=1}^{n} \left[\log\left(1 + e^{\mathbf{x_i^T}\beta}\right) - y_i \mathbf{x}_i^T \beta \right]$$

lo cual, en término de las probabilidades es equivalente a

$$D(\beta) = 2\sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

Esta función nos da una medida de la desviación máxima del modelo.

4.3.2. Estadístico de Wald

Se utiliza para saber si una variable es significativa en el modelo. Se trata de una prueba de hipótesis donde se contrasta la hipótesis nula

$$H_0 : \beta_i = 0$$

vs.

$$H_1 : \beta_i \neq 0$$

El estadístico de prueba se define como

$$w_j = \frac{\hat{\beta}_i}{s\left(\hat{\beta}_i\right)}, \text{ para } j = 1, \ldots, n.$$

donde la función s denota el error estándar y w_j se distribuye como $t - student$ con $n - p - 1$ grados de libertad bajo el supuesto de que H_0 es cierta; si $n \to \infty$ entonces w_j tiende en distribución a una normal estándar. En caso de que w_j tenga un valor alejado de cero se entiende que existe evidencia para afirmar que H_0 es falsa, por lo tanto la región crítica de la prueba es de la forma $|w_j| > z_\alpha/2$, para un nivel de significancia adecuado. Por lo tanto, si el verdadero valor del parámetro β_j es cero entonces la variable x_j debe excluirse.

4.3.3. Cociente de Verosimilitudes

Se utiliza para determinar si un conjunto de variables son significativas en el modelo. Suponga que se tienen n variables explicativas. Sin pérdida de generalidad se puede suponer que las variables a prueba son las últimas s. De esta forma se confrontan dos modelos; el primero incluye todas las variables, esto es

$$w_1 = \beta_0 + \beta_1 x_1 + \cdots + \beta_{n-s} x_{n-s} + \cdots + \beta_n x_n$$

y el segundo modelo incluye incluye solo las primeras $n - s$ variables

$$w_1 = \beta_0 + \beta_1 x_1 + \cdots + \beta_{n-s} x_{n-s}$$

Se contrasta la hipótesis nula H_0 de que las variables x_i con $i = n-s+1, \ldots, n$ no influyen significativamente en el modelo contra la hipótesis alternativa H_1 de que si influyen, de esta forma se tiene lo siguiente:

$$H_0 : \beta_{n-s+1} = 0 \, , \cdots , \, \beta_n = 0$$

vs.

$$H_0 : \beta_{n-s+1} \neq 0 \, \text{ó} \, \cdots \, \text{ó} \, \beta_n \neq 0$$

Para evidenciar que H_0 es falsa se utiliza la región crítica que surge del cociente de verosimilitud

$$\frac{\text{máx}\, L\left(H_0\right)}{\text{máx}\, L\left(H_1\right)} < \lambda$$

donde $L(H_0)$ y $L(H_1)$ son la función de log-verosimilitud de cada modelo. En términos de la desviación esto es

$$\chi^2_s = D\left(H_0\right) - D\left(H_1\right).$$

Si H_0 es cierta entonces el estadístico sigue una distribución χ^2_s con s grados de libertad y para un α dada se cumple que la región crítica $\chi^2_s > \chi^2_\alpha$. Nótese que cuando $s = 1$ sólo se está evaluando un coeficiente del modelo, lo cual es equivalente a la prueba de Wald.

4.3.4. Coeficiente de determinación

El coeficiente de determinación denomiado R^2 es un estadístico cuyo principal propósito es predecir futuros resultados o probar una hipótesis. El coeficiente determina la calidad del modelo para replicar los resultados y la proporción de variación de los resultados que puede explicarse por el modelo se utiliza para estimar la influencia de todas las variables en el modelo de manera global.

Un caso particular es verificar el modelo que no incluye variables de predicción y contiene únicamente β_0 contra el modelo que si incluye las variables de predicción. Esto es calcular el estadístico de la siguiente manera:

$$R^2 = 1 - \frac{D\left(\hat{\beta}\right)}{D\left(\hat{\beta}_0\right)}.$$

El estadístico R^2 arroja un valor entre cero y uno, donde el uno denota un ajuste perfecto a la variable dependiente, i.e. las variables explicativas definen perfectamente el comportamiento de y, la variable dependiente; mientras que un valor de cero en el estadístico indica que las variables no son capaces de predecir el valor de y, esto debido a que la desviación esperada para el modelo que incluye todas las variables es igual a la desviación del modelo que no las incluye, i.e. $D\left(\hat{\beta}\right) = D\left(\hat{\beta}_0\right)$.

Las pruebas estadísticas que utilizan estimadores de máxima verosimilitud en los modelos evalúan en gran medida la desviación general del modelo; si al añadir una variable al modelo este no mejora la verosimilitud o no disminuye la desviación de forma apreciable, entonces esta variable no se incluye en la ecuación.

Capítulo 5

Método de la Mejor Selección

En este capítulo se aborda el problema de la construcción del modelo de *credit scoring* con un efoque de análisis discriminante. EL objetivo es separar las poblaciones de clientes entre buenos y malos a traves de la segmentación de la cartera en grupos según sus características. El total de grupos que se obtienen depende del número de variables empleadas para la construcción del modelo. Las variables desciptivas nos ayudan a identificar la población en distintos grupos a los que posteriormente se evaluarán y se les determinará la tendencia que tienen a ser de alto o bajo riesgo. Con esto seremos capaces

5.1. Planteamiento del Modelo

Sea $X = (X_1, X_2, \ldots, X_p)$ el conjunto de p variables aleatorias que describen la información disponible de un cliente que solicita un crédito. A estas variables aleatorias se les conoce como *características*. Cuando se cuenta con el valor de cada una de estas variables se denota $x = (x_1, x_2, \ldots, x_p)$ y se les llama *atributos*. Por ejemplo, una caracterísitca es su estado civil, mientras que el atributo asociado puede ser casado, soltero, viudo, divorciado, etc.

Suponga que A es el conjunto de todos los posibles valores que las variables $X = (X_1, X_2, \ldots, X_p)$ pueden tomar, es decir todas las posibles maneras en

las que la solicitud de crédito puede ser llenada. El objetivo es encontrar la regla adecuada para separar el conjunto A en dos subgrupos A_G y A_B de tal forma que las solicitudes cuyas respuestas se encuentren en A_G ("Goods") sean clasificados como buenos y por lo tanto sean aceptados, mientras que aquellas solicitudes cuyas respuestas se encuentren en A_B ("Bads") serán clasificados como malos y por lo tanto serán rechazados, minimizando así el riesgo de la empresa y el costo esperado asociado. El costo corresponde a los dos tipos de error que se pueden cometer al clasificar a los clientes. Uno puede clasificar un cliente bueno como malo y rechazar su solicitud. En este caso la utilidad que se iba a generar por el cliente se pierde. Supongamos que la utilidad esperada de un cliente es L. El segundo error es clasificar un cliente malo como bueno y por lo tanto aceptar su solicitud. De esta forma se generará una deuda cuando dicho cliente caiga en default. Supongamos que la deuda esperada de un cliente es D.

Si se considera que es más grave para la institucion crediticia aceptar clientes potencialmente malos que rechazar clientes potencialmente buenos, entonces se debe asegurar que la probabilidad de aceptar clientes malos sea menor o igual a un numero pequeño α, previamente seleccionado, esto es, se pide que $P(A_G \,|\, B) \leq \alpha$, y se busca que $P(A_B \,|\, B)$ sea lo mas pequeño posible.

El siguiente teorema identica al mejor conjunto $A_G \subset \mathfrak{X}$, en el sentido que $P(A_G \,|\, B)$ tiene un valor fijo α pequeño, y $P(A_B \,|\, B)$ es lo menor posible, o equivalentemente, que su complemento $P(A_G \,|\, B) = 1 - P(A_B \,|\, B)$ sea lo mas grande posible.

Teorema: Dado el conjunto

$$A_G = \left\{ x \in \mathfrak{X} \,|\, \frac{P\,(X = x \,|\, B)}{P\,(X = x \,|\, G)} \leq \lambda_\alpha \right\} \text{ tal que } P\,(A_G \,|\, B) = \alpha,$$

entonces $P\,(A_G \,|\, G) > P\,(C \,|\, G)$ para todo $C \subset \mathfrak{X}$ que satisface $P\,(C \,|\, B) = \alpha$.

Demostración:

Los conjuntos A_G y C se pueden escribir como :

$$A_G = (A_G \cap C) \cup (A_G \cap C^c)$$

y

$$C = (A_G \cap C) \cup (A_G^c \cap C)$$

de aqui se sigue que

$$P(A_G \mid B) = P(A_G \cap C \mid B) + P(A_G \cap C^c \mid B)$$

y

$$P(C \mid B) = P(A_G \cap C \mid B) + P(A_G^c \cap C \mid B)$$

y como $P(A_G \mid B) = P(C \mid B) = \alpha$, entonces de las dos ecuaciones anteriores se sigue que $P(A_G \cap C^c \mid B) = P(A_G^c \cap C \mid B)$.

Cuando $x \in (A_G \cap C^c) \subset A_G$por la definición de A_G se sigue que $P(x \mid B) \leq \lambda P(x \mid G)$, y cuando $x \in (A_G^c \cap C) \subset A_G$por la definición de A_G^c se sigue que $P(x \mid B) > \lambda P(x \mid G)$; por lo tanto

$$\lambda P(A_G^c \cap C \mid G) < P(A_G^c \cap C \mid B) = P(A_G \cap C^c \mid B) \leq \lambda P(A_G \cap C^c \mid G).$$

De aqui se sigue que $P(C \mid G) < P(A_G \mid G)$. ■

Consideremos que en una población cerrada la proporción de solicitudes buenas es p_G y la proporción de solicitudes malas es p_B. Además supongamos que el número de características es finito y que cada una de ellas tiene un número finito de atributos, de tal forma que el conjunto A es finito. Esto es equivalente a decir que existe un número finito de formas para llenar una solicitud. Sea $p(x|\mathrm{G})$ la probabilidad de que un cliente bueno tenga los atributos x. Esta probabilidad condicional se representa como

$$p(x|G) = \frac{Prob(\text{El cliente es bueno y tiene los atributos } x)}{Prob(\text{El cliente es bueno})}.$$

De manera similar se define $p(x|B)$ como la probabilidad de que en cliente malo tenga los atributos x.

Si ahora se define $q(G|x)$ como la probabilidad de que un cliente con los atributos

x sea bueno, entonces la probabilidad condicional se representa como

$$q\left(G|x\right)=\frac{Prob\left(\text{El cliente tiene los atributos } x \text{ y es bueno}\right)}{Prob\left(\text{El cliente tiene los atributos } x\right)},$$

y si $p\left(x\right)=Prob\left(\text{El cliente tiene los atributos } x\right)$, entonces las probabilidades anteriores se pueden reordenar de la siguiente forma

$$Prob\left(\text{El cliente tiene los atributos } x \text{ y es bueno}\right)=q\left(G|x\right)p\left(x\right)=p\left(x|G\right)p_G.$$

Utilizando el teorema de Bayes se tiene que

$$q\left(G|x\right)=\frac{p\left(x|G\right)p_G}{p\left(x\right)}.$$

Un resultado similar se obtiene para $q\left(B|x\right)$, la probabilidad de que un cliente con los atributos x sea malo de la siguiente manera

$$q\left(B|x\right)=\frac{p\left(x|B\right)p_B}{p\left(x\right)}.$$

De las dos fórmulas anteriores se obtiene el resultado

$$\frac{q\left(G|x\right)}{q\left(B|x\right)}=\frac{p\left(x|G\right)p_G}{p\left(x|B\right)p_B}.$$

Utilizando los resultados anteriores podemos calcular el costo esperado de aceptar aquellos clientes con atributos en A_G y rechazar aquellos en A_B como

$$L\sum_{x\in A_B}p\left(x|G\right)p_G+D\sum_{x\in A_G}p\left(x|B\right)p_B$$

$$=L\sum_{x\in A_B}q\left(G|x\right)p\left(x\right)+D\sum_{x\in A_G}q\left(B|x\right)p\left(x\right).$$

Si se desea minimizar el costo esperado se debe considerar el costo de los dos escenarios. Cuando se clasifica a $x\in A_G$ entonces habrá un costo si el cliente es malo, en ese caso el costo esperado es $Dp\left(x|B\right)p_B$. Si se clasifica a $x\in A_B$

entonces habrá un costo si el cliente es bueno el cual es $Lp\left(x|G\right)p_G$. Por lo tanto uno debe clasificar a x en A_G si $Dp\left(x|B\right)p_B \leq Lp\left(x|G\right)p_G$. De esta forma la regla de decisión para minimizar el costo esperado esta dada por

$$A_G = \left\{x \mid Dp\left(x|B\right)p_B \leq Lp\left(x|G\right)p_G\right\}$$

$$= \left\{x \mid \frac{D}{L} \leq \frac{p\left(x|G\right)p_G}{p\left(x|B\right)p_B}\right\} = \left\{x \mid \frac{D}{L} \leq \frac{q\left(G|x\right)}{p\left(B|x\right)}\right\}.$$

La crítica más común hacia este resultado es que depende de los costos D y L los cuales son desconocidos. En lugar de minimizar el costo esperado uno podría buscar minimizar la probabilidad de cometer un tipo de error mientras se mantiene el otro a un nivel aceptable. Esto es minimizar la probabilidad de caer en default mientras se mantiene fijo el porcentaje de solicitudes aceptadas. La última condicíon en esquivalente a fijar la probabilidad de rechazar solicitudes buenas a un cierto nivel.

Supongamos que se desea que el porcentaje de solicitudes aceptadas sea a. Entonces A_G debe satisfacer lo siguiente

$$\sum_{x \in A_G} p\left(x|G\right)p_G + \sum_{x \in A_G} p\left(x|B\right)p_B = a$$

minimizando la tasa de default

$$\sum_{x \in A_G} p\left(x|B\right)p_B.$$

Si definimos $b(x) = p\left(x|\mathrm{B}\right)p_B$ para cada $x \in \mathrm{A}$, entonces se quiere obtener el conjunto A_G tal que se minimize

$$\sum_{x \in A_G} b\left(x\right) = \sum_{x \in A_G} \left(\frac{b\left(x\right)}{p\left(x\right)}\right) p\left(x\right)$$

sujeto a

$$\sum_{x \in A_G} p\left(x\right) = a.$$

La solución a este problema se da al eleguir al conjunto de atributos x tales que $\frac{b(x)}{p(x)} \leq c$, donde c se elige de manera que la suma de las $p\left(x\right)$ satisfagan la

condición igual a a. Entonces

$$A_G = \left\{ x \,|\, \frac{b\,(x)}{p\,(x)} \leq c \right\} = \{ x \,|\, q\,(B|x) \leq c \} = \left\{ x \,|\, \frac{1-c}{c} \leq \frac{p\,(x|G)\,p_G}{p\,(x|B)\,p_B} \right\}.$$

De esta forma se puede observar que el resultado es el mismo que se obtuvo anteriormente, ver [Edelman 2002].

5.2. Validación del modelo de comportamiento

Para validar la eficiencia del método de clasificación se utilizan los datos donde se conoce cuál es su situación real y a qué grupo pertenecen. Se clasifican dichos datos con el modelo generado y se contabiliza cuantos de ellos han sido clasificados adecuadamente. A continuación se describen los indicadores utilizados para validar el modelo que se va a construir.

5.2.1. Índice de Gini

Es uno de los métodos más utilizados para medir la desigualdad entre dos poblaciones. Se desprende a partir de la Curva de Lorenz, el cual es un gráfico utilizado para representar la distribución relativa de una variable en un dominio determinado, ver [Medina 2001].

Se tiene a G y F como las funciones de distribución teóricas asociadas a los clientes buenos y malos respectivamente, donde x es el puntaje o score. La curva de Lorenz de las funciones G y F es el conjunto del producto cartesiano dado por

$$\mathcal{L}(F,\, G) = \{ (u,\, v) \,|\, u = F(x) \,\mathrm{y}\, v = G(x)\,;\, \mathrm{con}\, x \in \mathbb{R} \}\,.$$

Si el puntaje para buenos es mayor que el puntaje para malos (como se desea que suceda por construcción del modelo), la curva de Lorenz tendrá una forma convexa. Claramente, si $F(x) = G(x)$ entonces $\mathcal{L}(F,\, G)$ describe la recta $u = v$ con $u \in (0,\, 1)$. Mientras más se separe $\mathcal{L}$ de la recta $u = v$, mayor será la

diferencia de $F(x)$ y $G(x)$. Llamémosle A al área que se encuentra entre la identidad y la curva de Lorenz. El índice de Gini se obtiene de la siguiente manera:

Definición (Índice de Gini): equivale a la razón entre el área A y el triángulo delimitado por la identidad, el eje horizontal u y la recta $u = 1$. Notemos que el triángulo mencionado en la definición del índice de Gini tiene un área de 0.5, por lo tanto el índice de Gini se puede escribir como $\frac{A}{0.5} = 2A$.

Cuando se desconocen las funciones de distribución G y F pero se cuenta con una muestra aleatoria de cada una de estas dos distribuciones empíricas de tamaño n_1 y n_2 respectivamente, se puede estimar la curva de Lorenz para poder calcular el índice de Gini. Para esto primero se tiene una partición de $\mathbb{R}$ dada por $x_0 \leq x_1 \leq x_2 \leq \cdots \leq x_k$, posteriormente se obtienen los estimadores de G y F en los puntos x_i como sigue:

$$\hat{F}(x_i) = \frac{\#\text{ de elementos en la muestra 1 menores o iguales que } x_i}{n_1}$$

y

$$\hat{G}(x_i) = \frac{\#\text{ de elementos en la muestra 2 menores o iguales que } x_i}{n_2}$$

La estimación de la curva de Lorenz de $F(x)$ y $G(x)$ es igual a la unión de los segmentos de recta que unen los puntos $\left(\hat{F}(x_{i-1}) = \hat{G}(x_{i-1})\right)$ y $\left(\hat{F}(x_i) = \hat{G}(x_i)\right)$. El área por debajo de la curva de Lorenz estimada para un intervalo tiene la forma de un trapecio y la calculamos como

$$A_i = \frac{\left(\hat{F}_i - \hat{F}_{i-1}\right)\left(\hat{G}_i - \hat{G}_{i-1}\right)}{2}$$

El área total por debajo de la curva de Lorenz estimada es

$$A = \sum_{i=2}^{k} A_i$$

El índice de Gini estimado se calcula como, ver [Medina 2001].

$$Gini = \frac{0.5 - A}{0.5} = 1 - 2A.$$

El índice de Gini denota un número entre 0 y 1. Si queremos contrastar dos modelos de clasificación distintos, aquel que tenga un índice de Gini mayor tendrá una mejor clasificación de los clientes.

5.2.2. Prueba de Kolmogorov-Smirnov

Es una prueba no paramétrica utilizada en bondad de ajuste. Si se desea probar que dos muestras independientes provienen de la misma distribución utilizamos la prueba Kolmogorov-Smirnov (K-S). El estadístico de prueba se calcula como la máxima diferencia absoluta entre sus distribuciones empíricas, entonces se busca detectar las discrepancias existentes entre las frecuencias relativas acumuladas de las dos muestras de estudio. Estas diferencias están determinadas no solo por las medias sino también por la dispersión, simetría o la oblicuidad. La prueba se construye sobre las hipótesis nula y alternativa como sigue:

H_0 : Las distribuciones poblacionales son iguales.

H_1 : Las distribuciones poblacionales son diferentes.

Para realizar esta prueba se requiere tener dos muestras de una variable aleatoria continua o de escala ordinal. Con los datos agrupados en k categorías o intervalos se calculan las frecuencias relativas acumuladas $\hat{F}_i$ y $\hat{G}_i$ con $i = 1, 2, \ldots, k$ que corresponden a las dos muestras de tamaño n_1 y n_2 respectivamente. Posteriormente se calculan las diferencias de las frecuencias relativas acumuladas. El estadístico esta dado como la máxima diferencia de las distribuciones de frecuencias relativas acumuladas

$$D_{max} = \max_{1 \leq i \leq k} \mid \hat{F}_i - \hat{G}_i \mid$$

Se selecciona aquel intervalo de clase que tenga la mayor desviación absoluta D. EL valor crítico se calcula como

$$D_{crítico} = 100K\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

donde K es el valor obtenido de la tabla K-S con $n_1 + n_2 - 2$ grados de libertad y un nivel de significancia dado. Si la desviación observada es menor que la desviación crítica tabulada entonces se acepta H_0, es decir que los datos observados no presentan diferencia significativa entre las dos poblaciones (buenos y malos). La función de distribución no discrimina las poblaciones, es la misma para ambas. Se rechaza H_0 si $D_{max} > D_{crítico}$, la distribución no es la misma para cada población, la prueba indica que hay discriminación entre las poblaciones.

5.2.3. Divergencia

Estadístico que mide la diferencia entre las medias de dos distribuciones estandarizadas utilizando las varianzas, el cual se calcula con la siguiente expresión:

$$Divergencia = \frac{2\left(\mu_1 - \mu_2\right)^2}{\sigma_1^2 + \sigma_2^2}.$$

Al construir un modelo logístico con la finalidad de clasificar dos poblaciones, es deseable que los dos grupos se encuentren estadísticamente bien separados, es decir, que la diferencia entre sus medias sea importante. Si contamos con una divergencia pequeña quiere decir que la distribución de las poblaciones es muy parecida, y por lo tanto será muy difícil diferenciar un grupo del otro. Un modelo que clasifique adecuadamente a los clientes deberá tener una divergencia mayor, ver [Simbaqueba 2004].

5.2.4. Sensibilidad y Especificidad

La sensibilidad nos indica la capacidad de nuestro modelo para dar como casos malos los casos realmente malos; proporción de malos correctamente identificados. Es decir,la sensibilidad caracteriza la capacidad del modelo para detectar alto riesgo en clientes malos. Se define como

$$Sensibilidad = \frac{VP}{VP + FN}$$

donde VP se refiere a Verdaderos Positivos y FN a Falsos Negativos.

La especificidad nos indica la capacidad de nuestro modelo para dar como casos buenos los casos realmente buenos; proporción de buenos correctamente identificados. Es decir, la especificidad caracteriza la capacidad del modelo para detectar bajo riesgo en clientes buenos.

$$Especificidad = \frac{VN}{VN + FP}$$

donde VN se refiere a Verdaderos Negativos y FP a Falsos Positivos.

No obstante, cada prueba tendrá mayor o menos porcentaje tanto en especificidad como en sensibilidad dependiendo de su punto de corte. Lo ideal sería que no se crucen los resultados y solamente obtuviésemos verdaderos positivos y verdaderos negativos pero no es así. Por eso esimportante conocer la especificidad y la sensibilidad del modelo. Referencia [Juez 1997].

Capítulo 6

Construcción del Modelo

En este capítulo se muestra la construcción de un modelo de comportamiento utilizando el método de la regresión logística y el de mejor selecćon. Para el armado de los modelos se utilizó una base de datos que contiene la información crediticia de varios clientes proporcionada por una institución financiera.

Los modelos de *credit scoring* se utilizan para evaluar el riesgo de crédito, es decir, para estimar una probabilidad de incumplimiento (probabilidad de default). Dicha probabilidad indica qué tan susceptible a caer en cartera vencida es un crédito en particular y se calcula en función de la información obtenida por las variables descriptivas seleccionadas tales como el historial crediticio, ingresos, egresos, etc. El resultado de esta evaluación se utiliza para decidir si se debe otorgar un crédito, aumentar o disminuir la línea de un crédito ya existente, o bien cancelar la cuenta de algún deudor. Los modelos de *credit scoring* enfocados en la etapa de la administración se conocen como **modelos de comportamiento** o *behavioral scoring*.

Los modelos de *credit scoring* asignan al evaluado un puntaje que determina el perfil de riesgo que el crédito tiene. El objetivo es obtener una estimación de la probabilidad de incumplimiento del deudor a partir de la calificación asignada, o bien, estimarla en función de la tasa de incumplimiento histórica observada por los integrantes de cada grupo.

6.1. Criterios de Clasificación

En los modelos de comportamiento las variables involucradas corresponden a características que describan el uso y los pagos que realice el cliente con el crédito. Dicha información se obtiene de la cartera de clientes a los cuales se les da seguimiento durante un periodo determinado el cual se conoce como **periodo de desempeño**, y tomando como base un **punto de observación**, es decir, una fecha en particular para definir el periodo. Posteriormente se considera otra fecha, por ejemplo 12 meses después, llamada **punto de resultados** y se clasifican los clientes de la muestra en *buenos* y *malos* según su estatus en ese momento. En la práctica se consideran como *malos* clientes aquellos que tengan dos meses de retraso.

Las variables consideradas incluyen datos como la deuda total y varios promedios del balance, por ejemplo, el total pagado en el último mes, en los últimos seis meses, etc., así como la cantidad de incrementos realizados y el uso que se le dió a la cuenta durante periodos de observación similares. Otras variables importantes son aquellas que se refieren al estatus de la cuenta en ese momento tales como el número de veces que excedió su límite de crédito, cuantas avisos han sido enviados, y cuánto tiempo ha pasado desde su último pago. Para construir el modelo de comportamiento es importante seleccionar adecuadamente las variables que se van a utilizar para que contengan información relevante que nos ayude a definir el riesgo del cliente.

Otros usos que se le dan a los modelos de comportamiento son: estimar si el usuario seguirá utilizando su linea de crédito, si cancelará su cuenta y cambiará de prestamista, cuando es buen momento para ofrecerle productos en venta, etc. Hay que tomar en cuenta que para cada uno de los diferentes usos es necesario recopilar información histórica relacionada al tema para poder construir el modelo.

6.2. Segmentación de la Cartera

Para realizar el behavior score se debe segmenta a la población según distintas características para obtener mejores resultados. A continuación se enlistan los

grupos que pertenecen a la cartera vigente de la cual nos interesa conocer su comportamiento en el futuro.

Current Old Clean: Son aquellos clientes que al momento de la observación se encuentran al corriente, cuentan con una antigüedad mayor a seis meses y nunca se han atrasado en sus pagos.

Current Old Dirty: Son aquellos clientes que al momento de la observación se encuentran al corriente, cuentan con una antigüedad mayor a seis meses, sin embargo ya se han atrasado en algún momento.

Current New Clean: Son aquellos clientes que al momento de la observación se encuentran al corriente, cuentan con una antigüedad menor a seis meses y nunca se han atrasado en sus pagos.

Current New Dirty: Son aquellos clientes que al momento de la observación se encuentran al corriente, cuentan con una antigüedad menor a seis meses, sin embargo ya se han atrasado en algún momento.

La información con la que se cuenta para hacer el modelo incluye clientes que están dentro de los primeros dos grupos, es decir, aquellos que actualmente se encuentran al corriente y cuya antigüedad es mayor a seis meses. Esto indica que nos enfocaremos que el segmento de la cartera que se encuentra al corriente y con al menos un semestre de operación, sin embargo, estas características no los exenta de caer en cartera vencida en algún momento del tiempo, lo cual es justamente lo que se modelará con el behavior score.

La base corresponde a los segmentos denominados *Current Old Clean* y *Current Old Dirty* de usuarios de tarjeta de crédito al 30 de septiembre del 2011, i.e. se trata de clientes cuya antigüedad en el banco es mayor a seis meses y que durante el mes de observación se encuentran al corriente con la característica de que algunos nunca se han atrasado en sus pagos (Clean) y el complemento si ha tenido algun atraso en su historial (Dirty). El tipo de variables que se toman en cuenta para la construcción del modelo son principalmente de historial crediticio por ejemplo, meses de antigüedad en el banco, meses desde el último uso, saldo total al término del mes, máximo saldo histórico, límite de crédito, número de veces que el cliente ha estado en mora, nivel máximo de morosidad, canasta en el que se encuentra el cliente en cada mes, entre muchas otras; también se combinan

algunos datos de mercado que puedan aportar información adicional como el puntaje del Buró de Crédito o el saldo total de todas sus tarjetas registradas, sin embargo estos datos no suelen ser relevantes porque corresponden a una institución y una metodología distinta lo cual hace que no sea seguro que lo que reporten coincida con el objetivo de la empresa.

Estas personas ya pasaron por un primer filtro de Scoring durante la originación del crédito que es cuando se les aceptó como clientes de la empresa, dicho Score trabaja principalmente con los datos demográficos de los usuarios, estas variables ya no serán consideradas dentro del modelo de comportamiento debido a que ya no aportan información relevante al modelo.

Por último se cuenta con el estatus de los clientes al 29 de febrero del 2012, cinco meses (120 días) después de la fecha de observación de la información, equivalente al tiempo necesario para enviar un crédito a cartera vencida y clasificarlo como un mal cliente por su falta de pago. Con todas estas variables el objetivo del Behavioral Scoring es clasificar a los clientes entre clientes buenos y clientes malos a partir de la información crediticia que se tiene, discriminando entre aquellos que si realizan sus pagos de la manera adecuada (clientes buenos) y aquellos que registran retrasos (clientes malos).

6.3. Análisis de la información

La base de datos utilizada cuenta con 47,385 registros. La información se considera hasta el 30 de septiembre del 2011 (fecha de observación) en este momento todos los clientes se encuentran al corriente, i.e. canasta 0, además se cuenta con el estatus de los clientes al 29 de febrero del 2012 que se distribuyen en 38,334 y 9,051 clientes buenos y clientes malos respectivamente. Un crédito es considerado como bueno si se mantiene dentro de las canastas 0 o 1 (máximo un periodo de retraso); mientras que se le considera como malo si alcanza la canasta 2 en adelante. Para construir el modelo se divide la base en 2 partes. Primero se toma una muestra aleatoria del 70 % (33,170) llamada de **entrenamiento** con la cual se construye y se calibra el modelo, mientras que el 30 % (14,215) restante se le llama de **validación** y se utiliza para evaluar los resultados del modelo obtenido. A continuación se muestra una tabla con las características o variables utilizadas para la contrucción del modelo:

ID	Variable	Descripción
1	Segmento	población de cartera al corriente
2	MOB 2	months on book; antigüedad en el banco
3	MFD	months first use date
4	MUD	meses desde ultima disp de efectivo
5	saldo_201103	saldo a la fecha
6	saldo_201104	saldo a la fecha
7	saldo_201105	saldo a la fecha
8	saldo_201106	saldo a la fecha
9	saldo_201107	saldo a la fecha
10	saldo_201108	saldo a la fecha
11	saldo_sep11	saldo a la fecha
12	actv_t-6	actividad en los últimos 6 meses
13	actv_t-5	actividad en los últimos 5 meses
14	actv_t-4	actividad en los últimos 4 meses
15	actv_t-3	actividad en los últimos 3 meses
16	actv_t-2	actividad en los últimos 2 meses
17	actv_t-1	actividad en los últimos 1 meses
18	actv_t-0	actividad en los últimos 0 meses
19	Max saldo	maximo saldo hist
20	CR_LIM	limite de credito
21	NUM_PD_30	numero de veces en toda su historia que el cliente a estado en 30 dias de mora
22	NUM_PD_60	numero de veces en toda su historia que el cliente a estado en 60 dias de mora
23	NUM_PD_90	numero de veces en toda su historia que el cliente a estado en 90 dias de mora
24	NUM_PD_120	numero de veces en toda su historia que el cliente a estado en 120 dias de mora
25	NUM_PD_150	numero de veces en toda su historia que el cliente a estado en 150 dias de mora
26	NUM_PD_180	numero de veces en toda su historia que el cliente a estado en 180 dias de mora
27	amt_incr	monto de incremento en limite de credito
28	mora_201103	bucket del mes en observacion
29	mora_201104	bucket del mes en observacion
30	mora_201105	bucket del mes en observacion
31	mora_201106	bucket del mes en observacion
32	mora_201107	bucket del mes en observacion
33	mora_201108	bucket del mes en observacion
34	cr_lim_201103	limite de credito al mes de observacion
35	cr_lim_201104	limite de credito al mes de observacion
36	cr_lim_201105	limite de credito al mes de observacion
37	cr_lim_201106	limite de credito al mes de observacion
38	cr_lim_201107	limite de credito al mes de observacion
39	cr_lim_201108	limite de credito al mes de observacion
40	NUM_CTASBC	numero de tarjetas del cliente en el sistema financiero (inlcuye otros bancos)
41	SUMSALDOS_BC	suma de saldos de todas sus tarjetas
42	MOP_CPRODBC	nivel de mora actual en las tarjetas bancarias ajenas
43	MOP_ACTBC	nivel de mora actual en las tarjetas departamentarias ajenas
44	MOP_12M	maxima mora historica LTM ajenas
45	SUMLINEAS_BC	suma de limites de creditos de todas sus tarjetas
46	MAX_LINBC	maxima linea de credito en todas sus tarjetas
47	MAX_ANITGBC	maxima antigüedad en buro de credito
48	PCT_USOBC	porcentaje de utilización de tarjetas de credito en sistema finaciero
49	consultas_12	numero de consultas realizadas a su reporte de buro de credito LTM
50	MAXPCT_USOBC	maximo porcentaje uso historico de sus tarjetas de credito
51	consultas_6	numero de consultas realizadas a su reporte de buro de credito ultimos 6 meses
52	BCSCORE	score de buro de credito
53	mslcl	meses desde ultimo incremento de linea
54	MFP	meses desde su primer pago
55	Cash_Disp	Disposición de efectivo 1: Si, 0:No
56	Months_Cash_Disp	Meses desde la disposición de efectivo
57	Months_last_Pur	Meses desde la ultima compra
58	Months_last_Use	Meses desde el ultimo uso
59	Months_open	Meses desde apertura
60	Months_First_Use	Meses desde el primer uso
61	Months_last_CrLim	Meses desde el ultimo cambio de limite de credito

Estas características son cosideradas como variables independientes y servirán para modelar el comportamiento de la variable dependiente **FLAG** la cual indica si un cliente es bueno o malo. La variable FLAG es igual a 1 uno cuando el cliente tiene 2 o más meses de retraso y por lo tanto debe ser clasificado como un mal cliente, mientras que la variable sera igual a 0 si el cliente está al corriente o tiene a lo más un periodo de retraso, esto es que se debe clasificar como bueno.

Es importante realizar una depuración a la base de datos de información para asegurarnos que la información que se tiene es correcta. Por ejemplo, en algunos casos podemos tener registros vacios que indican la ausencia de información de ese cliente relacionada a esa variable, sin embargo, algunos programas estadísticos podrían confundir esta ausencia de información con un cero.

Posteriormente hay que clasificar los atributos relacionados a cada variable por categorías, por ejemplo la característica $MOB\,2$ indica los meses de antigüedad de un cliente en la empresa, de manera que $MOB\,2 \in \mathbb{N}$. Se crean categorías que describan el comportamiento de los clientes por prupos, cada categoría seleccionada muestra una clara diferencia entre ellas con respecto a la variable dependiente FLAG. Si dos categorías muestran un comportamiento similar entonces es mejor combinarlos en una sola y reducir el número de categorias en observación. En el caso de la variable $MOB\,2$ las categorías creadas son: {0-36, 37-48, 49-60, 60+}. Se repite el mismo procedimiento con cada una de las variables y y se clasifica la información según las categorías generadas.

La siguiente tabla muestra las categorías establecidas para cada variable:

ID	Variable	Categorías					
1	Segmento	COCA	CODA				
2	MOB 2	0-36	36-48	48-60	60+		
3	MFD	0-36	36-48	48-60	60+		
4	MUD	0	1	2	3+		
5	saldo_201103	-0	0-5000	5001+			
6	saldo_201104	-0	0-5000	5001+			
7	saldo_201105	-0	0-5000	5001+			
8	saldo_201106	-0	0-5000	5001+			
9	saldo_201107	-0	0-5000	5001+			
10	saldo_201108	-0	0-5000	5001+			
11	saldo_sep11	-0	0-5000	5001+			
12	actv_t-6	0	1				
13	actv_t-5	0	1-2				
14	actv_t-4	0-1	2-3				
15	actv_t-3	0-2	3-4				
16	actv_t-2	0-1	2-5				
17	actv_t-1	0	1-2	3-6			
18	actv_t-0	0	1	2	3-5	6-7	
19	Max saldo	-0	0-10000	10001+			
20	CR_LIM	0-5000	5000-10000	10000-20000	20000-50000	50001+	
21	NUM_PD_30	0	1	2	3-5	6-10	11+
22	NUM_PD_60	0	1	2+			
23	NUM_PD_90	0	1+				
24	NUM_PD_120	0	1+				
25	NUM_PD_150	0	1+				
26	NUM_PD_180	0	1+				
27	amt_incr	0	0-10000	10001+			
28	mora_201103	0	1	2	3+		
29	mora_201104	0	1	2	3+		
30	mora_201105	0	1	2	3+		
31	mora_201106	0	1	2	3+		
32	mora_201107	0	1	2	3+		
33	mora_201108	0	1	2+			
34	cr_lim_201103	0-10000	10000-15000	15001+			
35	cr_lim_201104	0-10000	10000-15000	15001+			
36	cr_lim_201105	0-10000	10000-15000	15001+			
37	cr_lim_201106	0-10000	10000-15000	15001+			
38	cr_lim_201107	0-10000	10000-15000	15001+			
39	cr_lim_201108	0-10000	10000-15000	15001+			
40	NUM_CTASBC	0-1	1-6	7+			
41	SUMSALDOS_BC	0	1+				
42	MOP_CPRODBC	0-1	2	3+			
43	MOP_ACTBC	0-1	2+				
44	MOP_12M	0-1	2	3+			
45	SUMLINEAS_BC	0-1000	1000-50000	50001+			
46	MAX_LINBC	0-1000	1000-50000	50001+			
47	MAX_ANITGBC	0-1	2+				
48	PCT_USOBC	0-80	81+				
49	consultas_12	0-2	3+				
50	MAXPCT_USOBC	0-80	81+				
51	consultas_6	0	1-7	8+			
52	BCSCORE	200-	200-550	551+			
53	mslcl	0-10	10-30	30-60	61+		
54	MFP	0-10	10-60	61+			
55	Cash_Disp	0	1				
56	Months_Cash_Disp	1-12	13-24	25-36	37+		
57	Months_last_Pur	0-12	12-24	24-36	37+		
58	Months_last_Use	0-12	12-24	24-36	37+		
59	Months_open	0-36	36+				
60	Months_First_Use	0-36	36-60	61+			
61	Months_last_CrLim	0-12	12-36	37+			

6.4. Modelo de Regresión Logística

Para construir el modelo se hace uso de la base de entrenamiento. Utilizando una regresión logística se utilizó el paquete estadístico SAS. Con el podemos importar los datos que desamos analizar y ejecutar el modelo de regresión logística. En el anexo A se muestran los resultados arrojados por SAS al ejecutar la regresión logística con todas las variables en observación.

Con los resultados de la regresión listos es posible calcular para cada registro lo siguiente:

$$P = Pr\left[Flag = 1 \,|\, \mathbf{x}\right]$$

donde $\mathbf{x}$ es el vector de variables consideradas dentro del modelo y la variable $Flag$ indica si el clientes es bueno ($Flag$ =0) o malo ($Flag$ =1).

Posteriormente se procede a ordenar los registros por la variable P de menor a mayor y se divide la base en n grupos, en este caso se hizo una partición de 20 grupos (711 registros por grupo aproximadamente). Contabilizamos el total de clientes malos por grupo y acumulamos este resultado desde el último grupo, que es el que tiene mayor probabilidad de ser malo, hasta el primero. La siguiente tabla muestra el resultado de esta acción con los datos:

ID	Min P	Max P	# registros	# Flag=1	Acum Flag=1	% Acum Flag=1
1	0.0000	0.0027	710	5	2,734	100.0%
2	0.0027	0.0046	711	7	2,729	99.8%
3	0.0046	0.0064	709	9	2,722	99.6%
4	0.0064	0.0083	711	12	2,713	99.2%
5	0.0083	0.0103	708	17	2,701	98.8%
6	0.0103	0.0128	715	15	2,684	98.2%
7	0.0128	0.0156	713	18	2,669	97.6%
8	0.0156	0.0191	707	15	2,651	97.0%
9	0.0191	0.0233	712	20	2,636	96.4%
10	0.0233	0.0293	711	27	2,616	95.7%
11	0.0293	0.0375	710	35	2,589	94.7%
12	0.0375	0.0501	713	29	2,554	93.4%
13	0.0501	0.0717	710	45	2,525	92.4%
14	0.0717	0.1102	711	52	2,480	90.7%
15	0.1102	0.1850	710	79	2,428	88.8%
16	0.1851	0.3393	711	156	2,349	85.9%
17	0.3394	0.6145	711	333	2,193	80.2%
18	0.6146	0.8271	711	530	1,860	68.0%
19	0.8271	0.9562	711	631	1,330	48.6%
20	0.9563	1.0000	710	699	699	25.6%

Si deseamos que el error de clasificar un cliente malo como bueno sea aproximadamente 10 %, esto es:

$$Pr\left[M_{reg} = 0 \,|\, Flag = 1\right] = 0.1$$

donde M_{reg} es la clasificación que le asigna el Modelo de regresión al cliente.

Se sekeccionaron los grupos que van del 15 al 20 debido a que deseamos que el 90 % de los clientes malos sean clasificacdos adecuadamente. Esto quiere decir que segùn nuestro modelo de regresión logística todos aquellos registros donde se cumplan que $P < 0.1102$ serán clasificados como buenos y aquellos donde $P \geq 0.1102$ serán clasificados como malos. Esto nos garantiza que el 90 % del total de clientes malos que se encuentran dentro de la base serán clasificados efectivamente como malos, y solo el 10 % restante se confundirán con clientes buenos.

Una vez seleccionado el punto de corte se procede a construir las **matrices de confusión** con la base de validación, las cuales indican los resultados obtenidos con la clasificación del modelo. A continuación se muestran y explican las matrices construidas para este ejercicio:

M1		REAL		
MODELO	TOTALES	0	1	Total
	0	9,645	306	9,951
	1	1,836	2,428	4,264
	Total	11,481	2,734	14,215

M2		REAL		
MODELO	% Totales	0	1	Total
	0	68%	2%	70%
	1	13%	17%	30%
	Total	81%	19%	100%

M3		REAL		
MODELO	Pr[Flag | M]	0	1	Total
	0	97%	3%	100%
	1	43%	57%	100%
	Total	81%	19%	100%

M4		REAL		
MODELO	Pr[M | Flag]	0	1	Total
	0	84%	11%	70%
	1	16%	89%	30%
	Total	100%	100%	100%

La matriz M1 muestra cómo se distribuye la población total en las distinas opciones que son:

- $M1_{00}$: Aquellos que el modelo clasifica como buenos y en realidad son buenos.
- $M1_{01}$: Aquellos que el modelo clasifica como buenos y en realidad son malos.
- $M1_{10}$: Aquellos que el modelo clasifica como malos y en realidad son buenos.
- $M1_{11}$: Aquellos que el modelo clasifica como malos y en realidad son malos.

Esto quiere decir que el modelo clasificará como buenos un total de 9,951 clientes (70 %), los cuales en realidad se distribuyen en 9,645 buenos (68 %) y 306 malos (2 %); mientras que clasificará como malos a 4,264 clientes (30 %) de los cuales 1,836 (13 %) en realidad se comportaron como buenos y 2,428 (17 %) tuvieron un mal comportamiento. La matriz M2 muestra los porcentajes de esta selección de la cartera.

Para medir la precisión general del modelo calculamos

$$aciertos = \frac{M1_{00} + M1_{11}}{M1_{00} + M1_{01} + M1_{10} + M1_{11}}$$

en este caso los aciertos totales corresponden a 85 % de la base, lo cual implica que el 15 % restante fue clasificado erróneamente.

La matriz M3 evalúa la distribución de los clientes según los resultados del modelo, esto es $Pr\left[Flag \,|\, Modelo\right]$:

- $M3_{00} : Pr\left[Flag = 0 \,|\, Modelo = 0\right] = 0.97$, i.e. probabilidad de que el crédito sea bueno dado que el modelo lo clasifica como bueno.
- $M3_{01} : Pr\left[Flag = 0 \,|\, Modelo = 1\right] = 0.43$, i.e. probabilidad de que el crédito sea bueno dado que el modelo lo clasifica como malo.
- $M3_{10} : Pr\left[Flag = 1 \,|\, Modelo = 0\right] = 0.03$, i.e. probabilidad de que el crédito sea malo dado que el modelo lo clasifica como bueno.

- $M3_{11} : Pr\left[Flag = 1 \,|\, Modelo = 1\right] = 0.57$, i.e. probabilidad de que el crédito sea malo dado que el modelo lo clasifica como malo.

La matriz M4 evalúa los resultados del modelo según la distribución de los clientes, esto es $Pr\left[Modelo \,|\, Flag\right]$:

- $M4_{00} : Pr\left[Modelo = 0 \,|\, Flag = 0\right] = 0.84$, i.e. probabilidad de que el crédito sea clasificado por el modelo como bueno dado que es bueno.
- $M4_{01} : Pr\left[Modelo = 0 \,|\, Flag = 1\right] = 0.11$, i.e. probabilidad de que el crédito sea clasificado por el modelo como bueno dado que es malo.
- $M4_{10} : Pr\left[Modelo = 1 \,|\, Flag = 0\right] = 0.16$, i.e. probabilidad de que el crédito sea clasificado por el modelo como malo dado que es bueno.
- $M4_{11} : Pr\left[Modelo = 1 \,|\, Flag = 1\right] = 0.89$, i.e. probabilidad de que el crédito sea clasificado por el modelo como malo dado que es malo.

Con los resultados de M3 podemos evaluar la efectividad del modelo con los clientes clasificados como buenos. Observemos que $M3_{00}$ nos indica cuántos de los clientes asignados como buenos son en realidad buenos, esta variable describe que tan "limpia" o "sana" será nuestra cartera vigente y de qué tamaño será la cartera vencida. En este caso solo el 3 % de la cartera está compuesta por clientes morosos. Si asumimos que todos los créditos son independientes y además tienen la misma pérdida esperada D, con este dato calculamos la pérdida esperada de la cartera y el VaR de la empresa utilizando la versión del Teorema Tentral del Límite de Liapunov. Por otro lado si observamos los valores de $M3_{10}$ y $M3_{11}$ notaremos que son muy parecidos, quiere decir que de todos los clientes que el modelo considera malos casi la mitad son clientes buenos (43 % de los clientes clasificados como malos en realidad son buenos).

La matriz M4 muestra la efectividad del modelo para detectar adecuadamente los clientes. $M4_{00}$ muestra que el 84 % de los clientes buenos fueron identificados por el modelo, mientras que $M4_{11}$ indica que del total de clientes malos que existen en la población el 89 % fue clasificado adecuadamente.

Los errores que podemos cometer al clasificar nuestra cartera de clientes los podemos ver reflejados en $M4_{01}$ y $M4_{10}$. Para poder calibrar el modelo usualmente

se fija una de estas dos variables en un nivel bajo y se trabaja con el resto de las variables. En este caso se buscó que $M4_{01} = 0.1$ Esto es resolver el problema:

$$Minimizar \;\; M4_{01}$$

$$sujeto\, a \;\; M4_{10} = 0.1$$

Si deseamos ser más exigentes con la detección de clientes malos podemos reducir el valor de esta variable, por ejemplo $M4_{01} = 0.05$, esta acción tendría como resultado un menor número de clientes clasificados como buenos.

Con estas variables evaluadas podemos comparar los resultados de este modelo con otros y decidir si es mejor a partir del error que se cometa en cada uno de ellos. Para que sean resultados comparables se buscará obtener un valor menor o igual en las variables $M4_{01}$ y $M4_{10}$.

6.5. Modelo de Mejor Selección

Para construir el modelo utilizando la metodología de la Mejor Selección se utilizó la misma base de entrenamiento utilizada en el modelo de regresión. Lo primero que debemos hacer es elegir las características o variables que se emplearán en la construcción del mismo. En este caso se utilizarán las mismas variables que seleccionó el modelo de regresión logística para poder hacer los resultados comparables.

Si $X = (X_1, X_2, \ldots, X_p)$ es el conjunto de características que se va a utilizar en la construcción debemos enlistar todos los posibles valores de X para poder clasificarlos. A cada una de estas **clases** se le asigna un ID para poder identificarlo. De esta forma cualquier cliente que observemos tendrá atributos imputables a una cierta clase que nos permitirá clasificarlo de manera inmediata según los datos que alimentaron el modelo.

Si $x = (x_1, x_2, \ldots, x_p)$ es un escenario de los atributos se calcula

$$p = Pr\left[x \in A_0 \,|\, Flag = 0\right]$$

$$q = Pr\left[x \in A_1 \,|\, Flag = 1\right]$$

$$z = \frac{p}{q} = \frac{Pr\left[x \in A_0 \,|\, Flag = 0\right]}{Pr\left[x \in A_1 \,|\, Flag = 1\right]}$$

donde A_0 es el conjunto de clientes buenos y A_1 es el conjunto de clientes malos. De esta forma calculamos el valor z para todos los distintos escenarios que pueda alcanzar x y los ordenamos de menor a mayor con respecto a z.

A continuación se muestra un estracto de la clasificación con las probabilidades y el cocientes ya calculados y ordenados:

CLASE	Bueno	Malo	P(x|B)	P(x|M)	P(x|M) / P(x|B)
591	269	1	0.0070	0.0001	0.0157
2884	215	1	0.0056	0.0001	0.0197
1975	190	1	0.0050	0.0001	0.0223
725	162	1	0.0042	0.0001	0.0261
744	162	1	0.0042	0.0001	0.0261
2136	136	1	0.0035	0.0001	0.0311
2015	109	1	0.0028	0.0001	0.0389
1970	97	1	0.0025	0.0001	0.0437
2151	92	1	0.0024	0.0001	0.0460
2885	87	1	0.0023	0.0001	0.0487
2138	86	1	0.0022	0.0001	0.0492
821	82	1	0.0021	0.0001	0.0517
2319	77	1	0.0020	0.0001	0.0550
2179	73	1	0.0019	0.0001	0.0580
2188	73	1	0.0019	0.0001	0.0580
861	72	1	0.0019	0.0001	0.0588
1968	72	1	0.0019	0.0001	0.0588
2027	144	2	0.0038	0.0002	0.0588
2149	144	2	0.0038	0.0002	0.0588
615	70	1	0.0018	0.0001	0.0605

Por último hay que establecer un punto de corte c de manera que aquellas clases donde se cumpla $z \leq c$ se clasificarán como buenas y el resto serán clasificadas como malas. En esta ocación se buscó el valor de $c = 0.6$ de manera que la probabilidad de clasificar un cliente malo como bueno fuera 0.1; A continuación se muestran las matrices de confusión del modelo utilizando la base de validación:

	N1	Real		
MODELO	TOTALES	0	1	TOTAL
	0	10,522	288	10,810
	1	959	2,446	3,405
	TOTAL	11,481	2,734	14,215

	N2	Real		
MODELO	% Totales	0	1	TOTAL
	0	74%	2%	76%
	1	7%	17%	24%
	TOTAL	81%	19%	100%

	N3	Real		
MODELO	Pr[Flag | M]	0	1	TOTAL
	0	97%	3%	100%
	1	28%	72%	100%
	TOTAL	81%	19%	100%

	N4	Real		
MODELO	Pr[M | Flag]	0	1	TOTAL
	0	92%	11%	76%
	1	8%	89%	24%
	TOTAL	100%	100%	100%

El modelo clasifica como buenos un total de 10,810 clientes 76 %), los cuales se en realidad se distribuyen en 10,522 buenos (74 %) y 288 malos (2 %); mientras que clasifica como malos a 3,405 clientes (24 %) de los cuales 959 (7 %) en realidad tuvieron un buen comportamiento y 2,446 (17 %) fueron malos. La matriz N2 muestra los porcentajes de esta selección de la cartera.

Los aciertos totales corresponden a 91 % de la base, lo cual implica que el 10 % restante fue clasificado erroneamente.

Con los resultados de N3 podemos evaluar la efectividad del modelo con los clientes clasificados como buenos. $N3_{00}$ nos indica cuántos de los clientes asignados como buenos son en realidad buenos. En este caso sólo el 3 % de la cartera está compuesta por clientes morosos. Note que los valores de $N3_{10}$ y $N3_{11}$ son muy diferentes. Esto quiere decir que de la población clasificada como mala el 28 % de ellos en realidad eran clientes buenos y el resto fue clasificado adecuadamente, lo cual indica que el modelo discrimina correctamente dos tercios del total de clientes malos.

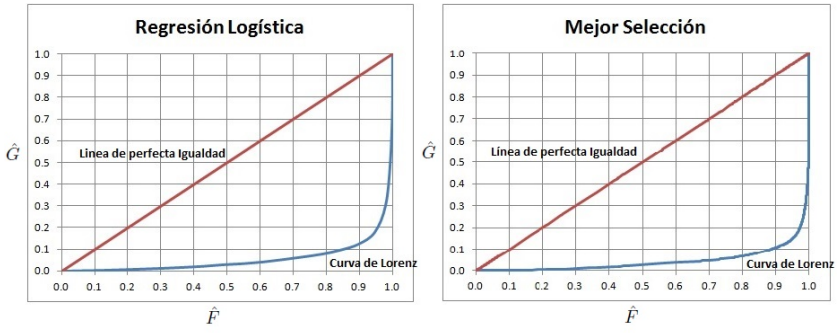
La matriz N4 muestra la efectividad del modelo para detectar adecuadamente los clientes. $N4_{00}$ muestra que el 92 % de los clientes buenos fueron identificados

por el modelo, mientras que $N4_{11}$ indica que del total de clientes malos que existen en la población el 89 % fue clasificado adecuadamente.

6.6. Análisis de Resultados

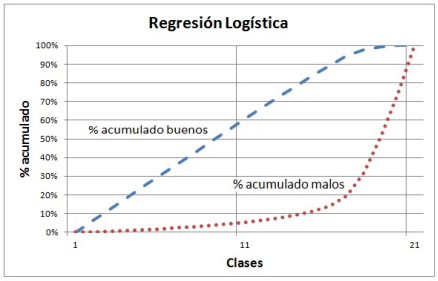
Si constrastamos las matrices de confusión de ambos métodos podemos observar las diferencias y decidir cuál de los métodos es mejor, sin embargo, contamos con algunos estadísticos de validación descritos en el capítulo anterior los cuales nos ayudan a tomar una decisión basada en indicadores.

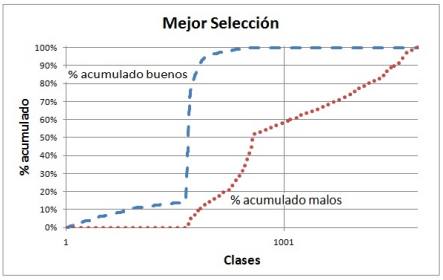
En primer lugar se cuenta con el índice de Gini. la siguiente gráfica muestra el resultado obtenido en ambos modelos para este indicador.



Si bien los resultados son muy similares en ambos modelos, existe una pequeña mejora en el modelo de mejor selección que alcanza un nivel de 92 % vs. el 87 % de la regresión.

El segundo indicador descrito fue la prueba K-S la cual tiene el siguiente comportamiento:





En esta prueba se busca la máxima distancia entre las distribuciones de buenos y malos. La regresión alcanza una diferencia máxima de 0.7542 mientras que en la mejor selección es de 0.8164 .

El indicador de la divergencia que combina media y varianza de la clasificación de las poblaciones. La divergencia calculada para el modelo logístico es de 0.9109 mientras que el cálculo del modelo de mejor selección es de 0.9437, por lo cual la divergencia apunta que la capacidad para identificar y clasificar adecuadamente a los clientes es mejor en el modelo de mejor selección que en el modelo logit.

Por último se calcula la sensibilidad y la especificidad de ambos modelos. En el modelo logístico se alcanza una sensibilidad de 88.8 % vs. 89.5 % del modelo de mejor selección, miesntras que la especificidad es de 84.0 % vs. 91.6 % respectivamente.

Capítulo 7

Conclusiones

Para construir un modelo de *Behavioral Score* es necesario contar con características que describan el comportamiento crediticio de los clientes. Si bien la técnica utilizada tradicionalmente en las instituciones financieras ofrece un ajuste bueno en la discriminación de clientes, es posible manejar la información de diferente manera logrando que se mejore el resultado antes obtenido. Si se trata de una cartera compuesta de millones de dolares un cambio del 3 % puede significar la diferencia entre generar utilidades o pérdidas para la empresa, por lo cual cualquier mejora en el proceso de selección es bienvenido.

El principal objetivo del modelo propuesto de La Mejor Selección es proporcionar un procedimiento más eficaz y exacto a la institución para que sea capaz de mejorar su clasificicación de una manera rápida y sin que esto implique un esfuerzo mucho mayor en cuanto a tiempo y recursos. Con este modelo se exhibe una herramienta adicional que se puede aplicar a los modelos ya existentes para mejorar los resultados obtenidos y asi incrementar las ganancias generadas o bien disminuir el riesgo o la pérdida que conlleve una cartera.

Las variables que se eligen para la construcción del modelo deben ser seleccionadas cuidadosamente ya que si no se utilizan las adecuadas no obtendremos información relevante que sea capaz de predecir el comportamiento de los clientes.

Una vez que se cuenta con la información apropiada para ajustar el modelo se debe depurar la base para evitar introducir errores así como seleccionar adecuadamente las variables descriptivas evitando sobreajustar los parámetros y un sesgo al resultado. Para esto es importante dividir nuestra población en dos grupos distribuidos adecuadamente, uno para la construcción del modelo y otro para la validación del mismo.

El modelo de La mejor Selección es una herramienta adicional fácil de construir y en poco tiempo la cual mejora los resultados obtenidos en la construcción de un modelo de Bebavioral Score. Esperamos que la aplicación de esta herramienta en las instituciones financieras de resultados satisfactorios que comprueben lo que la teoría desarrollada nos afirma.

Durante la elaboración del trabajo se encontraron varias áreas de oportunidad para continuar la investigación, por ejemplo, respecto a la selección de variables se propone que una prueba de independencia entre las variables y el comportamiento de los clientes podría seleccionar las variables adecuadas, otra alternativa es utilizar la prueba no parámtrica de Mann Whitney. El estudio de estas alternativas quedan pendientes para trabajos futuros.

Bibliografía

[Edelman 2002] Edelman, David B., "Credit Scoring and its applications". Society for Industrial and Applied Mathematics, Philadelphia, 2002.

[Elizondo 2004] Elizondo, Alan, "Medición Integral del Riesgo de Crédito". Editorial Limusa, 2004.

[Girault 2007] Gutiérrez Girault, Matías Alfredo, "Modelos de Credit Scoring, Qué, Cómo, Cuándo y Para Qué", Octubre de 2007.

[Groot 1988] de Groot, Morris H. "Probabilidad y Estadística". Addison-Wesley Iberoamericana, Wilmingon, Delaware, 1988.

[Hand and Heley 1997] Hand, D. J. and Henley, W. E., "Statistical Classification Methods in Consumer Credit Scoring: a Review". Royal Statistical Society, 160 (1997), Part 3, pgs. 523-541.

[Juez 1997] Juez, Pedro M., "Probabilidad y estadística en medicina". Ediciones Díaz de Santos, S.A., 1997.

[Krishnaiah 1987] Krishnaiah, P.R., "Handbook of Statistics 2". North-Holland, 1987.

[Medina 2001] Medina Fernando, "Consideraciones sobre el índice de Gini para medir la concentración del ingreso".

Estudios estadísticos y prospectivos, serie 9. Publicación de las Naciones Unidas. Santiago de Chile, Marzo 2001.

[Mester 1997] Mester, Loretta J., "What´s the Point of Credit Scoring?". Federal Reserve Bank of Philadelphia, September/October 1997, pgs. 3-16.

[Rényi 1976] Rényi, Alfréd, "Cálculo de Probabiliades". Editorial Reverté, 1976.

[Rincón 2012] Rincón, Luis, "Introducción a la Teoría del Riesgo". Facultad de Ciencias UNAM. Agosto 2012.

[Sarmiento y Velez 2007] Sarmiento Lotero, R., Vélez Molano,R., "Teoría del Riesgo en Mercados Financieros: Una visión Teórica". Cuadernos Latinoamericanos de Administración; Vol II No.4, Enero – Junio de 2007.

[Simbaqueba 2004] Simbaqueba Lilian. "¿Qué es el scoring? Una visión práctica de la gestión del riesgo de crédito". Instituto del Riesgo Financiero, Bogotá, 2004.

[Srinivasan and kim 1987] Srinivasan, V. and Kim, Y. H., "Credit Granting: A Comparative Analysis of Classification Procedures". The Journal of Finance, vol. XLII, N° 3, July 1987.

[Thomas 2000] Thomas, L. C., "A Survey of credit and behavioral scoring: forecasting financial risk of lending to consumers". International Journal of Forecasting, 16 (2000), pgs. 149-172.

Anexo A

Resultados de la regresión logística

Procedimiento LOGISTIC

Información del modelo	
Conjunto de datos	WORK.SORTTEMPTABLESORTED
Variable de respuesta	Flag 0 - 1
Número de niveles de respuesta	2
Modelo	logit binario
Técnica de optimización	Puntuación de Fisher

Número de observaciones leí	33170
Número de observaciones usa	33170

Perfil de respuesta		
Valor ordenado	**Flag 0 - 1**	**Frecuencia total**
1	0	26853
2	1	6317

La probabilidad modelada es Flag 0 - 1='1'.

Procedimiento de selección stepwise

Información de nivel de clase						
Clase	**Valor**	**Diseño de variables**				
Segmento	**COCA**	1				
	CODA	-1				
MOB 2	**0-36**	1	0	0		
	37-48	0	1	0		
	49-60	0	0	1		
	60+	-1	-1	-1		
MFD	**0-36**	1	0	0		
	37-48	0	1	0		
	49-60	0	0	1		
	60+	-1	-1	-1		
MUD	**0**	1	0	0		
	1	0	1	0		
	2	0	0	1		
	3+	-1	-1	-1		
saldo_201103	**-0**	1	0	0		
	1-5000	0	1	0		
	5001+	0	0	1		
	NULL	-1	-1	-1		
saldo_201104	**-0**	1	0	0		
	1-5000	0	1	0		
	5001+	0	0	1		
	NULL	-1	-1	-1		
saldo_201105	**-0**	1	0	0		
	1-5000	0	1	0		

Resultados de la regresión logística

Procedimiento LOGISTIC

Procedimiento de selección stepwise

Información de nivel de clase						
Clase	**Valor**	**Diseño de variables**				
	5001+	0	0	1		
	NULL	-1	-1	-1		
saldo_201106	**-0**	1	0	0		
	1-5000	0	1	0		
	5001+	0	0	1		
	NULL	-1	-1	-1		
saldo_201107	**-0**	1	0	0		
	1-5000	0	1	0		
	5001+	0	0	1		
	NULL	-1	-1	-1		
saldo_201108	**-0**	1	0	0		
	1-5000	0	1	0		
	5001+	0	0	1		
	NULL	-1	-1	-1		
saldo_sep11	**-0**	1	0			
	1-5000	0	1			
	5001+	-1	-1			
actv_t-6	**0**	1				
	1	-1				
actv_t-5	**0**	1				
	1-2	-1				
actv_t-4	**0-1**	1				
	2-3	-1				
actv_t-3	**0-2**	1				
	3-4	-1				
actv_t-2	**0-1**	1				
	2-5	-1				
actv_t-1	**0**	1	0			
	1-2	0	1			
	3-6	-1	-1			
actv_t-0	**0**	1	0	0	0	
	1	0	1	0	0	
	2	0	0	1	0	
	3-5	0	0	0	1	
	6-7	-1	-1	-1	-1	
Max saldo	**-0**	1	0			
	1-10000	0	1			
	10001+	-1	-1			
CR_LIM	**0-5000**	1	0	0	0	
	10001-20000	0	1	0	0	
	20001-50000	0	0	1	0	

Resultados de la regresión logística

Procedimiento LOGISTIC

Procedimiento de selección stepwise

Información de nivel de clase						
Clase	**Valor**	**Diseño de variables**				
	50001+	0	0	0	1	
	5001-10000	-1	-1	-1	-1	
NUM_PD_30	**0**	1	0	0	0	0
	1	0	1	0	0	0
	11+	0	0	1	0	0
	2	0	0	0	1	0
	3-5	0	0	0	0	1
	6-10	-1	-1	-1	-1	-1
NUM_PD_60	**0**	1	0			
	1	0	1			
	2+	-1	-1			
NUM_PD_90	**0**	1				
	1+	-1				
NUM_PD_120	**0**	1				
	1+	-1				
cr_lim_201103	**0-10000**	1	0	0		
	10001-15000	0	1	0		
	15001+	0	0	1		
	NULL	-1	-1	-1		
cr_lim_201104	**0-10000**	1	0	0		
	10001-15000	0	1	0		
	15001+	0	0	1		
	NULL	-1	-1	-1		
cr_lim_201105	**0-10000**	1	0	0		
	10001-15000	0	1	0		
	15001+	0	0	1		
	NULL	-1	-1	-1		
cr_lim_201106	**0-10000**	1	0	0		
	10001-15000	0	1	0		
	15001+	0	0	1		
	NULL	-1	-1	-1		
cr_lim_201107	**0-10000**	1	0	0		
	10001-15000	0	1	0		
	15001+	0	0	1		
	NULL	-1	-1	-1		
cr_lim_201108	**0-10000**	1	0	0		
	10001-15000	0	1	0		
	15001+	0	0	1		
	NULL	-1	-1	-1		
NUM_CTASBC	**0-1**	1	0	0		
	2-6	0	1	0		

Resultados de la regresión logística

Procedimiento LOGISTIC

Procedimiento de selección stepwise

Información de nivel de clase						
Clase	Valor	Diseño de variables				
	7+	0	0	1		
	NULL	-1	-1	-1		
SUMSALDOS_BC	0	1	0			
	1+	0	1			
	NULL	-1	-1			
MOP_CPRODBC	0-1	1	0	0		
	2	0	1	0		
	3+	0	0	1		
	NULL	-1	-1	-1		
MOP_ACTBC	0-1	1	0			
	2+	0	1			
	NULL	-1	-1			
MOP_12M	0-1	1	0	0		
	2	0	1	0		
	3+	0	0	1		
	NULL	-1	-1	-1		
SUMLINEAS_BC	0-1000	1	0	0		
	1001-50000	0	1	0		
	50001+	0	0	1		
	NULL	-1	-1	-1		
MAX_LINBC	0-1000	1	0	0		
	1001-50000	0	1	0		
	50001+	0	0	1		
	NULL	-1	-1	-1		
MAX_ANITGBC	0-1	1	0			
	2+	0	1			
	NULL	-1	-1			
PCT_USOBC	0-80	1	0			
	81+	0	1			
	NULL	-1	-1			
consultas_12	0-2	1	0			
	3+	0	1			
	NULL	-1	-1			
MAXPCT_USOBC	0-80	1	0			
	81+	0	1			
	NULL	-1	-1			
consultas_6	0	1	0	0		
	1-7	0	1	0		
	8+	0	0	1		
	NULL	-1	-1	-1		
BCSCORE	-200	1	0	0		

Resultados de la regresión logística

Procedimiento LOGISTIC

Procedimiento de selección stepwise

Información de nivel de clase						
Clase	**Valor**	**Diseño de variables**				
	200-550	0	1	0		
	551+	0	0	1		
	NULL	-1	-1	-1		
mslcl	**0-10**	1	0	0	0	
	11-30	0	1	0	0	
	31-60	0	0	1	0	
	61+	0	0	0	1	
	NULL	-1	-1	-1	-1	
MFP	**0-10**	1	0			
	11-60	0	1			
	61+	-1	-1			
Cash_Disp	**0**	1				
	1	-1				
Months_Cash_Disp	**1-12**	1	0	0	0	
	13-24	0	1	0	0	
	25-36	0	0	1	0	
	37+	0	0	0	1	
	NULL	-1	-1	-1	-1	
Months_last_Pur	**0-12**	1	0	0	0	
	13-24	0	1	0	0	
	25-36	0	0	1	0	
	37+	0	0	0	1	
	NULL	-1	-1	-1	-1	
Months_Last_Use	**0-12**	1	0	0		
	13-24	0	1	0		
	25-36	0	0	1		
	37+	-1	-1	-1		
Months_open	**0-36**	1				
	36+	-1				
Months_fisrt_use	**0-36**	1	0			
	37-60	0	1			
	61+	-1	-1			
Months_last_CrLim	**0-12**	1	0	0		
	13-36	0	1	0		
	37+	0	0	1		
	NULL	-1	-1	-1		
NUM_PD_150	**0**	1				
	1+	-1				
NUM_PD_180	**0**	1				
	1+	-1				
amt_incr	**0**	1	0			

Resultados de la regresión logística

Procedimiento LOGISTIC

Procedimiento de selección stepwise

Información de nivel de clase						
Clase	**Valor**	**Diseño de variables**				
	1-10000	0	1			
	10001+	-1	-1			
mora_201103	**0**	1	0	0	0	
	1	0	1	0	0	
	2	0	0	1	0	
	3+	0	0	0	1	
	NULL	-1	-1	-1	-1	
mora_201104	**0**	1	0	0	0	
	1	0	1	0	0	
	2	0	0	1	0	
	3+	0	0	0	1	
	NULL	-1	-1	-1	-1	
mora_201105	**0**	1	0	0	0	
	1	0	1	0	0	
	2	0	0	1	0	
	3+	0	0	0	1	
	NULL	-1	-1	-1	-1	
mora_201106	**0**	1	0	0	0	
	1	0	1	0	0	
	2	0	0	1	0	
	3+	0	0	0	1	
	NULL	-1	-1	-1	-1	
mora_201107	**0**	1	0	0	0	
	1	0	1	0	0	
	2	0	0	1	0	
	3+	0	0	0	1	
	NULL	-1	-1	-1	-1	
mora_201108	**0**	1	0	0		
	1	0	1	0		
	2+	0	0	1		
	NULL	-1	-1	-1		

Paso 0. Término independiente introducido:

Estado de convergencia del modelo
Criterio de convergencia (GCONV=1E-8) satisfecho.

-2 LOG L	=	32298.619

Test residual de chi-cuadrado		
Chi-cuadrado	**DF**	**Pr > ChiSq**
19280.7177	123	<.0001

Resultados de la regresión logística

Procedimiento LOGISTIC

Paso 1. Efecto NUM_PD_60 introducido:

Estado de convergencia del modelo
Criterio de convergencia (GCONV=1E-8) satisfecho.

Estadístico de ajuste del modelo		
Criterio	Sólo términos independientes	Términos independientes y Variables adicionales
AIC	32300.619	19536.217
SC	32309.029	19561.445
-2 LOG L	32298.619	19530.217

Probar hipótesis nula global: BETA=0			
Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	12768.4023	2	<.0001
Puntuación	14350.6994	2	<.0001
Wald	8620.1085	2	<.0001

Test residual de chi-cuadrado		
Chi-cuadrado	DF	Pr > ChiSq
5825.8654	121	<.0001

Note: No effects for the model in Step 1 are removed

Paso 2. Efecto Months_Last_Use introducido:

Estado de convergencia del modelo
Criterio de convergencia (GCONV=1E-8) satisfecho.

Estadístico de ajuste del modelo		
Criterio	Sólo términos independientes	Términos independientes y Variables adicionales
AIC	32300.619	16121.931
SC	32309.029	16172.388
-2 LOG L	32298.619	16109.931

Probar hipótesis nula global: BETA=0			
Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	16188.6880	5	<.0001
Puntuación	17126.3858	5	<.0001
Wald	8344.0800	5	<.0001

Test residual de chi-cuadrado		
Chi-cuadrado	DF	Pr > ChiSq
2562.3314	118	<.0001

Resultados de la regresión logística

Procedimiento LOGISTIC

Paso 3. Efecto NUM_PD_120 introducido:

Estado de convergencia del modelo
Criterio de convergencia (GCONV=1E-8) satisfecho.

Estadístico de ajuste del modelo		
Criterio	Sólo términos independientes	Términos independientes y Variables adicionales
AIC	32300.619	15488.064
SC	32309.029	15546.930
-2 LOG L	32298.619	15474.064

Probar hipótesis nula global: BETA=0			
Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	16824.5555	6	<.0001
Puntuación	17805.2424	6	<.0001
Wald	7691.2648	6	<.0001

Test residual de chi-cuadrado		
Chi-cuadrado	DF	Pr > ChiSq
2173.7993	117	<.0001

Note: No effects for the model in Step 3 are removed.

Paso 4. Efecto saldo_201103 introducido:

Estado de convergencia del modelo
Criterio de convergencia (GCONV=1E-8) satisfecho.

Estadístico de ajuste del modelo		
Criterio	Sólo términos independientes	Términos independientes y Variables adicionales
AIC	32300.619	14925.720
SC	32309.029	15009.814
-2 LOG L	32298.619	14905.720

Probar hipótesis nula global: BETA=0			
Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	17392.8994	9	<.0001
Puntuación	18142.4836	9	<.0001
Wald	7637.7443	9	<.0001

Test residual de chi-cuadrado		
Chi-cuadrado	DF	Pr > ChiSq
1753.4139	114	<.0001

Resultados de la regresión logística

Procedimiento LOGISTIC

Paso 5. Efecto MOP_CPRODBC introducido:

Estado de convergencia del modelo
Criterio de convergencia (GCONV=1E-8) satisfecho.

Estadístico de ajuste del modelo		
Criterio	Sólo términos independientes	Términos independientes y Variables adicionales
AIC	32300.619	14569.777
SC	32309.029	14679.099
-2 LOG L	32298.619	14543.777

Probar hipótesis nula global: BETA=0			
Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	17754.8423	12	<.0001
Puntuación	18318.7879	12	<.0001
Wald	7523.5349	12	<.0001

Test residual de chi-cuadrado		
Chi-cuadrado	DF	Pr > ChiSq
1417.7332	111	<.0001

Note: No effects for the model in Step 5 are removed.

Paso 6. Efecto mora_201108 introducido:

Estado de convergencia del modelo
Criterio de convergencia (GCONV=1E-8) satisfecho.

Estadístico de ajuste del modelo		
Criterio	Sólo términos independientes	Términos independientes y Variables adicionales
AIC	32300.619	14128.610
SC	32309.029	14263.160
-2 LOG L	32298.619	14096.610

Probar hipótesis nula global: BETA=0			
Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	18202.0097	15	<.0001
Puntuación	18577.6634	15	<.0001
Wald	7283.8621	15	<.0001

Test residual de chi-cuadrado		
Chi-cuadrado	DF	Pr > ChiSq
1103.7374	108	<.0001

Resultados de la regresión logística

Procedimiento LOGISTIC

Paso 7. Efecto Months_last_Pur introducido:

Estado de convergencia del modelo
Criterio de convergencia (GCONV=1E-8) satisfecho.

Estadístico de ajuste del modelo		
Criterio	Sólo términos independientes	Términos independientes y Variables adicionales
AIC	32300.619	13947.750
SC	32309.029	14115.938
-2 LOG L	32298.619	13907.750

Probar hipótesis nula global: BETA=0			
Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	18390.8695	19	<.0001
Puntuación	18757.6465	19	<.0001
Wald	7242.0609	19	<.0001

Test residual de chi-cuadrado		
Chi-cuadrado	DF	Pr > ChiSq
916.9448	104	<.0001

Note: No effects for the model in Step 7 are removed.

Paso 8. Efecto mslcl introducido:

Estado de convergencia del modelo
Criterio de convergencia (GCONV=1E-8) satisfecho.

Estadístico de ajuste del modelo		
Criterio	Sólo términos independientes	Términos independientes y Variables adicionales
AIC	32300.619	13838.021
SC	32309.029	14039.847
-2 LOG L	32298.619	13790.021

Probar hipótesis nula global: BETA=0			
Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	18508.5981	23	<.0001
Puntuación	18774.7901	23	<.0001
Wald	7113.7743	23	<.0001

Test residual de chi-cuadrado		
Chi-cuadrado	DF	Pr > ChiSq
809.7923	100	<.0001

Resultados de la regresión logística

Procedimiento LOGISTIC

Paso 9. Efecto Months_fisrt_use introducido:

Estado de convergencia del modelo
Criterio de convergencia (GCONV=1E-8) satisfecho.

Estadístico de ajuste del modelo		
Criterio	Sólo términos independientes	Términos independientes y Variables adicionales
AIC	32300.619	13744.373
SC	32309.029	13963.018
-2 LOG L	32298.619	13692.373

Probar hipótesis nula global: BETA=0			
Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	18606.2463	25	<.0001
Puntuación	18844.9230	25	<.0001
Wald	7106.0032	25	<.0001

Test residual de chi-cuadrado		
Chi-cuadrado	DF	Pr > ChiSq
719.4946	98	<.0001

Note: No effects for the model in Step 9 are removed

Paso 10. Efecto NUM_PD_90 introducido:

Estado de convergencia del modelo
Criterio de convergencia (GCONV=1E-8) satisfecho.

Estadístico de ajuste del modelo		
Criterio	Sólo términos independientes	Términos independientes y Variables adicionales
AIC	32300.619	13682.237
SC	32309.029	13909.291
-2 LOG L	32298.619	13628.237

Probar hipótesis nula global: BETA=0			
Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	18670.3822	26	<.0001
Puntuación	18955.7760	26	<.0001
Wald	7092.4566	26	<.0001

Test residual de chi-cuadrado		
Chi-cuadrado	DF	Pr > ChiSq
660.1906	97	<.0001

Resultados de la regresión logística

Procedimiento LOGISTIC

Paso 11. Efecto Months_Cash_Disp introducido:

Estado de convergencia del modelo
Criterio de convergencia (GCONV=1E-8) satisfecho.

Estadístico de ajuste del modelo		
Criterio	Sólo términos independientes	Términos independientes y Variables adicionales
AIC	32300.619	13635.216
SC	32309.029	13895.908
-2 LOG L	32298.619	13573.216

Probar hipótesis nula global: BETA=0			
Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	18725.4033	30	<.0001
Puntuación	18976.7958	30	<.0001
Wald	7064.9561	30	<.0001

Test residual de chi-cuadrado		
Chi-cuadrado	DF	Pr > ChiSq
608.4551	93	<.0001

Note: No effects for the model in Step 11 are removed.

Paso 12. Efecto amt_incr introducido:

Estado de convergencia del modelo
Criterio de convergencia (GCONV=1E-8) satisfecho.

Estadístico de ajuste del modelo		
Criterio	Sólo términos independientes	Términos independientes y Variables adicionales
AIC	32300.619	13594.775
SC	32309.029	13872.285
-2 LOG L	32298.619	13528.775

Probar hipótesis nula global: BETA=0			
Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	18769.8446	32	<.0001
Puntuación	19018.3477	32	<.0001
Wald	7047.2526	32	<.0001

Test residual de chi-cuadrado		
Chi-cuadrado	DF	Pr > ChiSq
567.0646	91	<.0001

Resultados de la regresión logística

Procedimiento LOGISTIC

Paso 13. Efecto NUM_PD_150 introducido:

Estado de convergencia del modelo
Criterio de convergencia (GCONV=1E-8) satisfecho.

Estadístico de ajuste del modelo		
Criterio	Sólo términos independientes	Términos independientes y Variables adicionales
AIC	32300.619	13555.119
SC	32309.029	13841.039
-2 LOG L	32298.619	13487.119

Probar hipótesis nula global: BETA=0			
Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	18811.5005	33	<.0001
Puntuación	19024.0987	33	<.0001
Wald	6943.0793	33	<.0001

Test residual de chi-cuadrado		
Chi-cuadrado	DF	Pr > ChiSq
539.1787	90	<.0001

Note: No effects for the model in Step 13 are removed.

Paso 14. Efecto MOP_12M introducido:

Estado de convergencia del modelo
Criterio de convergencia (GCONV=1E-8) satisfecho.

Estadístico de ajuste del modelo		
Criterio	Sólo términos independientes	Términos independientes y Variables adicionales
AIC	32300.619	13519.180
SC	32309.029	13821.919
-2 LOG L	32298.619	13447.180

Probar hipótesis nula global: BETA=0			
Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	18851.4390	35	<.0001
Puntuación	19040.9205	35	<.0001
Wald	6928.5168	35	<.0001

Test residual de chi-cuadrado		
Chi-cuadrado	DF	Pr > ChiSq
500.0874	88	<.0001

Resultados de la regresión logística

Procedimiento LOGISTIC

Paso 15. Efecto NUM_PD_30 introducido:

Estado de convergencia del modelo
Criterio de convergencia (GCONV=1E-8) satisfecho.

Estadístico de ajuste del modelo		
Criterio	Sólo términos independientes	Términos independientes y Variables adicionales
AIC	32300.619	13481.477
SC	32309.029	13826.263
-2 LOG L	32298.619	13399.477

Probar hipótesis nula global: BETA=0			
Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	18899.1424	40	<.0001
Puntuación	19073.1071	40	<.0001
Wald	6865.4476	40	<.0001

Test residual de chi-cuadrado		
Chi-cuadrado	DF	Pr > ChiSq
452.9433	83	<.0001

Paso 16. Efecto mora_201108 eliminado:

Estado de convergencia del modelo
Criterio de convergencia (GCONV=1E-8) satisfecho.

Estadístico de ajuste del modelo		
Criterio	Sólo términos independientes	Términos independientes y Variables adicionales
AIC	32300.619	13878.070
SC	32309.029	14197.627
-2 LOG L	32298.619	13802.070

Probar hipótesis nula global: BETA=0			
Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	18496.5495	37	<.0001
Puntuación	18853.6831	37	<.0001
Wald	7064.7421	37	<.0001

Test residual de chi-cuadrado		
Chi-cuadrado	DF	Pr > ChiSq
747.8724	86	<.0001

Note: No effects for the model in Step 16 are removed.

Resultados de la regresión logística

Procedimiento LOGISTIC

Paso 17. Efecto mora_201108 introducido:

Estado de convergencia del modelo
Criterio de convergencia (GCONV=1E-8) satisfecho.

Estadístico de ajuste del modelo		
Criterio	Sólo términos independientes	Términos independientes y Variables adicionales
AIC	32300.619	13481.477
SC	32309.029	13826.263
-2 LOG L	32298.619	13399.477

Probar hipótesis nula global: BETA=0			
Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	18899.1424	40	<.0001
Puntuación	19073.1071	40	<.0001
Wald	6865.4476	40	<.0001

Test residual de chi-cuadrado		
Chi-cuadrado	DF	Pr > ChiSq
452.9433	83	<.0001

Paso 18. Efecto mora_201108 eliminado:

Estado de convergencia del modelo
Criterio de convergencia (GCONV=1E-8) satisfecho.

Estadístico de ajuste del modelo		
Criterio	Sólo términos independientes	Términos independientes y Variables adicionales
AIC	32300.619	13878.070
SC	32309.029	14197.627
-2 LOG L	32298.619	13802.070

Probar hipótesis nula global: BETA=0			
Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	18496.5495	37	<.0001
Puntuación	18853.6831	37	<.0001
Wald	7064.7421	37	<.0001

Test residual de chi-cuadrado		
Chi-cuadrado	DF	Pr > ChiSq
747.8724	86	<.0001

Note: No effects for the model in Step 18 are removed.

Resultados de la regresión logística

Procedimiento LOGISTIC

Note: Model building terminates because the last effect entered is removed by the Wald statistic criterion.

Resumen de selección de paso a paso

Paso	Efecto Introducido	Efecto Eliminado	DF	Número en	Chi-cuadrado de puntuación	Chi-cuadrado de Wald	Pr > ChiSq
1	NUM_PD_60		2	1	14350.6994		<.0001
2	Months_Last_Use		3	2	3756.8317		<.0001
3	NUM_PD_120		1	3	516.2465		<.0001
4	saldo_201103		3	4	515.2847		<.0001
5	MOP_CPRODBC		3	5	365.2951		<.0001
6	mora_201108		3	6	342.7764		<.0001
7	Months_last_Pur		4	7	199.2550		<.0001
8	mslcl		4	8	106.9733		<.0001
9	Months_fisrt_use		2	9	97.3373		<.0001
10	NUM_PD_90		1	10	63.8354		<.0001
11	Months_Cash_Disp		4	11	54.9502		<.0001
12	amt_incr		2	12	43.5640		<.0001
13	NUM_PD_150		1	13	36.5186		<.0001
14	MOP_12M		2	14	40.2576		<.0001
15	NUM_PD_30		5	15	45.9078		<.0001
16		mora_201108	3	14		6.1953	0.1025
17	mora_201108		3	15	308.5748		<.0001
18		mora_201108	3	14		6.1953	0.1025

Tipo 3 Análisis de efectos

Efecto	DF	Chi-cuadrado de Wald	Pr > ChiSq
saldo_201103	3	350.7271	<.0001
NUM_PD_30	5	47.2275	<.0001
NUM_PD_60	2	1299.0172	<.0001
NUM_PD_90	1	91.8000	<.0001
NUM_PD_120	1	93.4881	<.0001
MOP_CPRODBC	3	127.1136	<.0001
MOP_12M	2	39.8479	<.0001
mslcl	4	148.6475	<.0001
Months_Cash_Disp	4	47.7179	<.0001
Months_last_Pur	4	123.6641	<.0001
Months_Last_Use	3	133.2095	<.0001
Months_fisrt_use	2	92.5776	<.0001
NUM_PD_150	1	29.9355	<.0001
amt_incr	2	52.6716	<.0001

Resultados de la regresión logística

Procedimiento LOGISTIC

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

MOP_12M3+ =
MOP_CPRODBC0-1 + MOP_CPRODBC2 + MOP_CPRODBC3+ - MOP_12M0-1 - MOP_12M2

Análisis del estimador de máxima verosimilitud

Parámetro		DF	Estimador	Error estándar	Chi-cuadrado de Wald	Pr > ChiSq
Intercept		1	0.9264	0.1930	23.0465	<.0001
saldo_201103	**-0**	1	-0.9366	0.0989	89.6122	<.0001
saldo_201103	**1-5000**	1	0.2937	0.0655	20.1170	<.0001
saldo_201103	**5001+**	1	0.9054	0.0649	194.8669	<.0001
NUM_PD_30	**0**	1	-0.4964	0.0864	32.9866	<.0001
NUM_PD_30	**1**	1	-0.0539	0.0709	0.5771	0.4475
NUM_PD_30	**11+**	1	0.2967	0.0536	30.5874	<.0001
NUM_PD_30	**2**	1	0.0199	0.0706	0.0797	0.7777
NUM_PD_30	**3-5**	1	0.1156	0.0481	5.7685	0.0163
NUM_PD_60	**0**	1	-1.3826	0.0384	1298.2644	<.0001
NUM_PD_60	**1**	1	0.4074	0.0332	150.7136	<.0001
NUM_PD_90	**0**	1	-0.3506	0.0366	91.8000	<.0001
NUM_PD_120	**0**	1	-0.7690	0.0795	93.4881	<.0001
MOP_CPRODBC	**0-1**	1	-0.1880	0.0808	5.4098	0.0200
MOP_CPRODBC	**2**	1	0.2288	0.0909	6.3424	0.0118
MOP_CPRODBC	**3+**	1	0.4541	0.0583	60.6256	<.0001
MOP_12M	**0-1**	1	-0.5272	0.0849	38.5784	<.0001
MOP_12M	**2**	1	-0.3321	0.0981	11.4607	0.0007
MOP_12M	**3+**	0	0	.	.	.
mslcl	**0-10**	1	-0.9127	0.0822	123.2933	<.0001
mslcl	**11-30**	1	0.1129	0.0502	5.0536	0.0246
mslcl	**31-60**	1	-0.0289	0.0528	0.2987	0.5847
mslcl	**61+**	1	0.1755	0.0976	3.2325	0.0722
Months_Cash_Disp	**1-12**	1	0.1759	0.0656	7.1852	0.0074
Months_Cash_Disp	**13-24**	1	-0.1112	0.0660	2.8402	0.0919
Months_Cash_Disp	**25-36**	1	0.1977	0.0643	9.4639	0.0021
Months_Cash_Disp	**37+**	1	0.00197	0.0456	0.0019	0.9655
Months_last_Pur	**0-12**	1	-0.3687	0.1196	9.5062	0.0020
Months_last_Pur	**13-24**	1	-0.1462	0.0963	2.3044	0.1290
Months_last_Pur	**25-36**	1	-0.2349	0.0903	6.7719	0.0093
Months_last_Pur	**37+**	1	0.8015	0.0743	116.3357	<.0001
Months_Last_Use	**0-12**	1	-1.2894	0.1177	120.0927	<.0001

Resultados de la regresión logística

Procedimiento LOGISTIC

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

Análisis del estimador de máxima verosimilitud						
Parámetro		DF	Estimador	Error estándar	Chi-cuadrado de Wald	Pr > ChiSq
Months_Last_Use	13-24	1	0.1103	0.0938	1.3817	0.2398
Months_Last_Use	25-36	1	0.7813	0.0896	76.1100	<.0001
Months_fisrt_use	0-36	1	0.3279	0.0423	60.0224	<.0001
Months_fisrt_use	37-60	1	0.0200	0.0331	0.3643	0.5461
NUM_PD_150	0	1	-1.0453	0.1910	29.9355	<.0001
amt_incr	0	1	-0.4604	0.0646	50.8011	<.0001
amt_incr	1-10000	1	0.1811	0.0465	15.1565	<.0001

Estimadores de cocientes de disparidad;			
Efecto	Estimador del punto	95% Wald Límites de confianza	
saldo_201103 -0 vs NULL	0.510	0.320	0.812
saldo_201103 1-5000 vs NULL	1.744	1.148	2.650
saldo_201103 5001+ vs NULL	3.215	2.121	4.874
NUM_PD_30 0 vs 6-10	0.541	0.434	0.675
NUM_PD_30 1 vs 6-10	0.842	0.700	1.014
NUM_PD_30 11+ vs 6-10	1.196	1.057	1.352
NUM_PD_30 2 vs 6-10	0.907	0.756	1.088
NUM_PD_30 3-5 vs 6-10	0.998	0.878	1.134
NUM_PD_60 0 vs 2+	0.095	0.082	0.109
NUM_PD_60 1 vs 2+	0.567	0.500	0.642
NUM_PD_90 0 vs 1+	0.496	0.430	0.572
NUM_PD_120 0 vs 1+	0.215	0.157	0.293
MOP_CPRODBC 0-1 vs NULL	1.359	0.835	2.214
MOP_CPRODBC 2 vs NULL	2.062	1.253	3.393
MOP_CPRODBC 3+ vs NULL	2.583	1.689	3.950
MOP_12M 0-1 vs NULL	0.250	0.154	0.406
MOP_12M 2 vs NULL	0.304	0.182	0.507
mslcl 0-10 vs NULL	0.209	0.161	0.271
mslcl 11-30 vs NULL	0.583	0.490	0.693
mslcl 31-60 vs NULL	0.506	0.414	0.617
mslcl 61+ vs NULL	0.620	0.463	0.832
Months_Cash_Disp 1-12 vs NULL	1.553	1.309	1.843
Months_Cash_Disp 13-24 vs NULL	1.166	0.980	1.386
Months_Cash_Disp 25-36 vs NULL	1.588	1.341	1.879
Months_Cash_Disp 37+ vs NULL	1.305	1.160	1.469
Months_last_Pur 0-12 vs NULL	0.728	0.501	1.059
Months_last_Pur 13-24 vs NULL	0.910	0.653	1.267

Resultados de la regresión logística

Procedimiento LOGISTIC

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

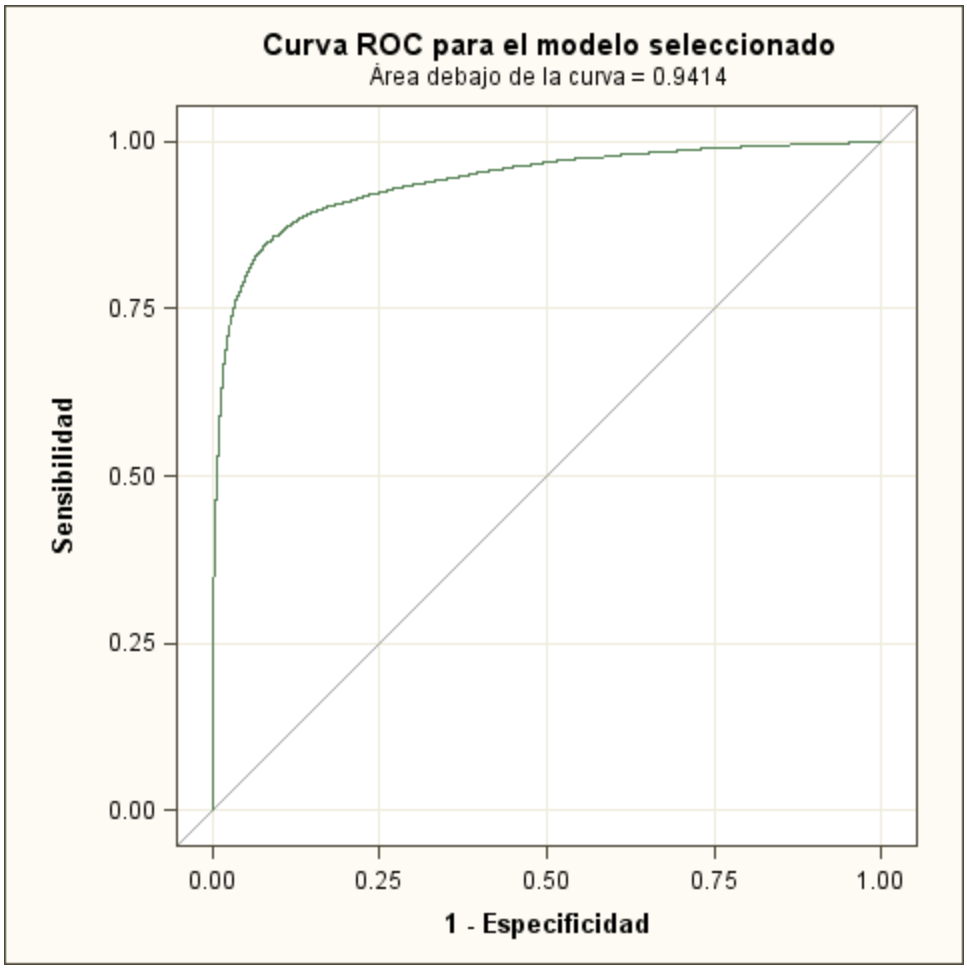
Estimadores de cocientes de disparidad;			
Efecto	Estimador del punto	95% Wald Límites de confianza	
Months_last_Pur 25-36 vs NULL	0.833	0.609	1.138
Months_last_Pur 37+ vs NULL	2.347	1.805	3.053
Months_Last_Use 0-12 vs 37+	0.185	0.131	0.261
Months_Last_Use 13-24 vs 37+	0.750	0.561	1.004
Months_Last_Use 25-36 vs 37+	1.467	1.144	1.882
Months_fisrt_use 0-36 vs 61+	1.966	1.706	2.265
Months_fisrt_use 37-60 vs 61+	1.445	1.294	1.612
NUM_PD_150 0 vs 1+	0.124	0.058	0.261
amt_incr 0 vs 10001+	0.477	0.377	0.604
amt_incr 1-10000 vs 10001+	0.906	0.757	1.085

Asociación de probabilidades predichas y respuestas observadas			
Concordancia de porcentaje	94.1	D de Somers	0.883
Discordancia de porcentaje	5.9	Gamma	0.883
Porcentaje ligado	0.0	Tau-a	0.272
Pares	169630401	c	0.941

Resultados de la regresión logística

Procedimiento LOGISTIC

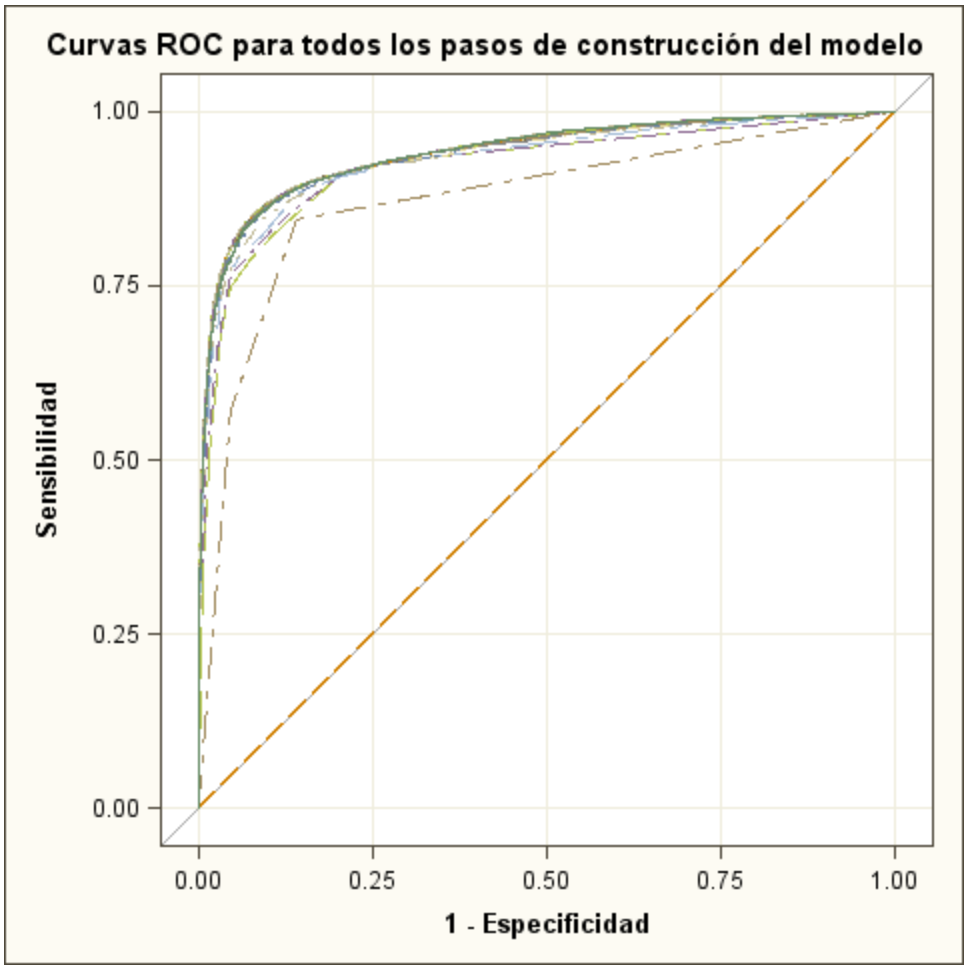
Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.



Resultados de la regresión logística

Procedimiento LOGISTIC

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.



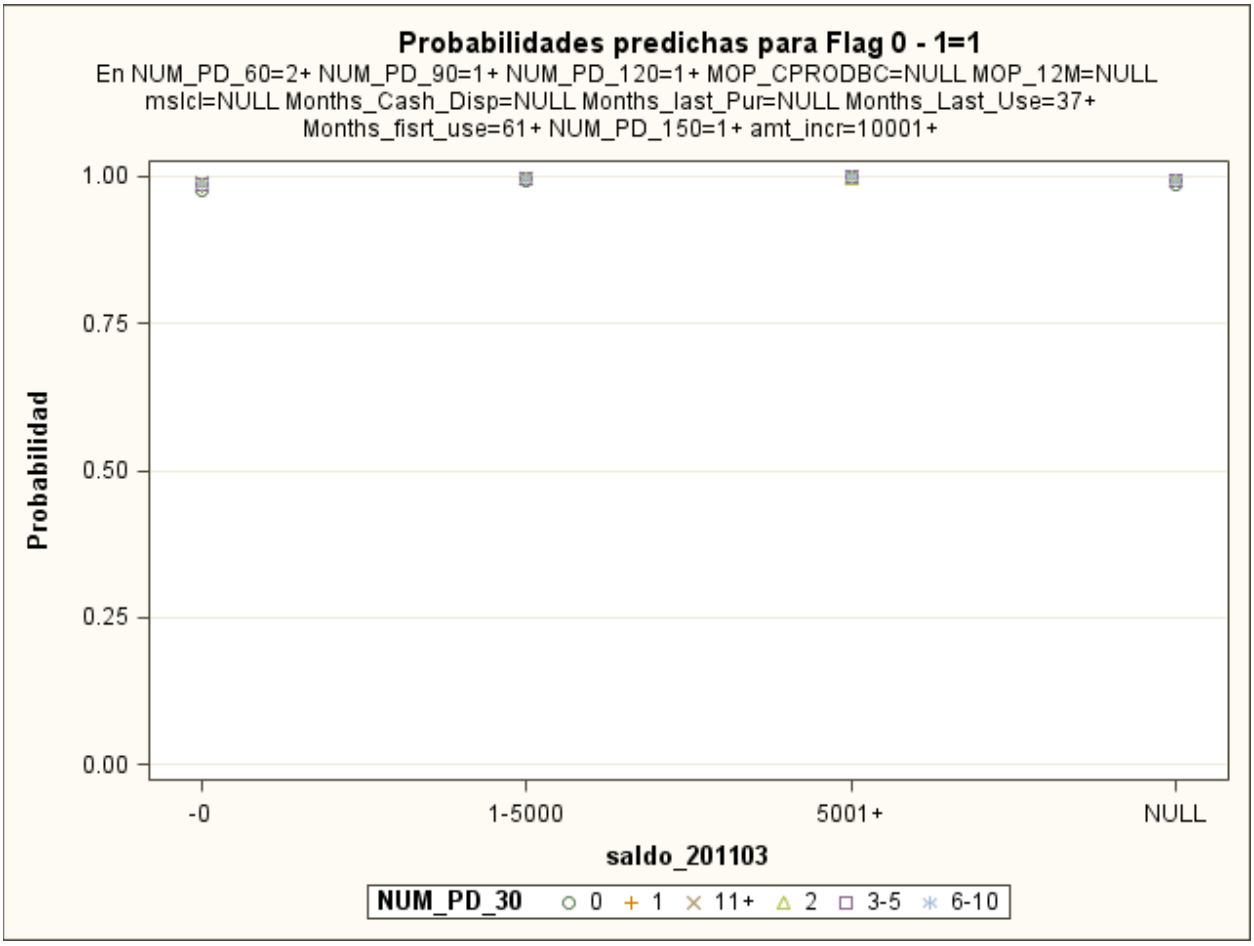
Partición para el test de Hosmer y Lemeshow					
		Flag 0 - 1 = 1		Flag 0 - 1 = 0	
Grupo	Total	Observado	Esperado	Observado	Esperado
1	3313	27	10.92	3286	3302.08
2	3320	31	23.83	3289	3296.17
3	3319	56	37.16	3263	3281.84
4	3317	69	54.38	3248	3262.62
5	3317	117	81.08	3200	3235.92
6	3317	139	129.36	3178	3187.64
7	3317	206	253.07	3111	3063.93
8	3317	544	658.16	2773	2658.84
9	3317	1986	1964.62	1331	1352.38
10	3316	3142	3104.42	174	211.58

Resultados de la regresión logística

Procedimiento LOGISTIC

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

Test de bondad de ajuste de Hosmer y Lemeshow		
Chi-cuadrado	DF	Pr > ChiSq
98.5511	8	<.0001



Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA

UNIVERSIDAD AUTÓNOMA METROPOLITANA – IZTAPALAPA
DIVISIÓN DE CIENCIAS BÁSICAS E INGENIERÍA

DISEÑO DE UN MODELO AJUSTADO DE COMPORTAMIENTO PARA RIESGO CREDITICIO

Tesis que presenta
Javier Sotelo Chávez

Para obtener el grado de
Maestro en Ciencias Matemáticas Aplicadas e Industriales

Asesora: Blanca Rosa Pérez Salvador

Jurado Calificador:

Presidente: Dr. Carlos Cuevas Covarrubias
Secretaria: Dra. Blanca Rosa Pérez Salvador
Vocal: Dr. Gabriel Núñez Antonio

México, D.F., Julio 2014