Proyecto de Tesis

Crédito al Consumo: La Estadística aplicada a un problema de Riesgo Crediticio

---

Alumna:
Soraida Nieto Murillo

Asesores:
Dra. Blanca R. Pérez Salvador
Act. Fernando Soriano Flores

18 de mayo de 2010

Índice general

Introducción **1**

1. Tarjetas de Crédito **5**
1.1. Créditos 5
1.1.1. Tipos de créditos 6
1.2. Tarjetas de crédito 7
1.2.1. Funcionamiento de las tarjetas de crédito 7
1.2.2. Clasificación de las tarjetas de crédito 8
1.3. Crédito al consumo 10
1.3.1. Ciclo de riesgo 11
1.4. Estatus de los clientes en los pagos 12
1.4.1. Posibles condiciones de un cliente 12
1.4.2. Definición de buenos y malos clientes 14
1.5. Estados de los clientes 15

2. Herramientas de Estadística y Probabilidad **17**
2.1. Cadenas de Markov: matrices de transición 17
2.1.1. Cadenas de Markov 18
2.1.2. Matrices de transición 20
2.1.3. Probabilidades de transición en n pasos 22
2.1.4. Estados absorbentes 25
2.2. Regresión logística 28
2.2.1. El modelo de regresión logística 28
2.2.2. Estimación del modelo logit usando MV 32
2.3. Pruebas estadísticas al modelo logito 34
2.3.1. Deviance 34
2.3.2. Estadístico de Wald 35
2.3.3. Estadístico R^2 36

2.3.4. Residuos de Pearson 37
2.3.5. Criterios para elegir el mejor modelo 37
2.3.6. Validación del método de clasificación 38
2.4. Prueba de diferencias de dos poblaciones 39
2.4.1. Índice de Gini 39
2.4.2. Divergencia 43
2.4.3. Test de Kolmogorov-Smirnov 44

3. Credit Scoring **47**
3.1. ¿Qué es el credit scoring? 47
3.1.1. Tipos de score 47
3.1.2. Tipos de modelos 49
3.2. La scorecard 49
3.3. Determinación de clientes buenos y malos 52
3.3.1. ¿Qué es un cliente bueno y un cliente malo? 52
3.3.2. Ventana de muestreo 53
3.3.3. Proceso para determinar a los clientes buenos y a los clientes malos 54
3.4. Obtención de la función de clasificación 57
3.4.1. Definición de parámetros 58
3.4.2. Pesos de Evidencia, WOE 60
3.4.3. Pruebas sobre la clasificación de atributos 61
3.4.4. El modelo logístico 64
3.4.5. Construcción de la *Scorecard* 64
3.5. Determinación del punto de corte o *Cut Off* 66

4. Una Aplicación del Credit Scoring **67**
4.1. Archivos requeridos 67
4.2. Limpieza de la base de datos 69
4.2.1. Posibles errores en los campos de datos 69
4.2.2. Campos vacíos 70
4.3. Bueno o malo en nuestra base de datos 71
4.3.1. Descripción 71
4.3.2. Determinación de cliente bueno o malo 72
4.4. Selección de variables que explican el modelo 73
4.4.1. Clasificación fina 74
4.4.2. Clasificación dura 75
4.5. Regresión logística 76

4.6. Scorecard 80
4.6.1. Validación de la Scorecard 81
4.6.2. Determinación del punto de corte o *Cut Off* 85
4.7. Scorecard sobre un segmento de Universitarios 87
4.8. Conclusiones 89

A. Resultados de la regresión logística **91**

Bibliografía **93**

Introducción

Una de las necesidades más importantes de las instituciones crediticias es tener criterios confiables para determinar a quienes deben otorgar el crédito y en que medida hacerlo. En este sentido, es importante medir el riesgo que se corre cuando se otorga un crédito y es de sumo interés para las empresas reducir lo más posible este riesgo cuando se adquieren nuevos clientes con lo que se reducirá la pérdida económica debida a una mala decisión.

Segun Enrique Castillo Sánchez Mejorada presidente de la Asociación de Bancos de México uno de los principales instrumentos de crédito son las tarjetas de crédito, en el país existen aproximadamente 40 millones de tarjetas de crédito, de las cuales 26.5 millones son emitidas por bancos y el resto por otro tipo de establecimientos, como las cadenas comerciales. A pesar del gran número de tarjetas emitidas, alrededor del 60 % de los mexicanos no poseen una tarjeta de crédito, esto hace a México un mercado potencial para este producto. Se considera que la economía de una persona es sana cuando dedica menos del 35 % del ingreso al pago de sus deudas, por lo que si alguien requiere más del 35 % de sus ingresos para el pago de deudas de tarjetas, esa persona tiene un factor de gran peso que seguramente la llevará a dejar de pagar [ver González (2008)].

Igualmente, Castillo, indica que la cifra de cartera vencida en tarjetas de crédito, es ahora equivalente a 9.9 % del total prestado y va en aumento, ésta cifra indica el enorme problema que es para las instituciones de crédito aprobar un préstamo a personas que no van a pagarles. En este sentido es importante medir el riesgo que se corre cuando se otorga un crédito y es de sumo interés para las empresas lograr minimizar este riesgo [ver González (2008)].

El scoring experto o estadístico es una herramienta que sirve para discriminar los buenos prospectos de los malos prospectos. Se trata de una metodología que pronostica el riesgo futuro por el incumplimiento de pagos en una ventana de tiempo determinada. Se basa en el análisis de dos tipos de datos referente a los clientes. Datos demográficos como pueden ser edad, sexo, ingresos, situación

laboral, y datos de buró de crédito que pueden ser el número de tarjetas de crédito en mora, su historial crediticio y su comportamiento en cuanto a la morosidad de pagos.

El modelo para discriminar a los buenos clientes generalmente consiste en una fórmula con parámetro desconocido que se pueden estimar con los datos de la institución objetivo; es decir, basándose en la experiencia propia; sin embargo, también puede construirse con información de otras instituciones; es decir con experiencia externa. Podremos entonces estimar la probabilidad de que el préstamo, si se desembolsara se clasifique como "malo" o "bueno" de acuerdo a la definición de clasificación del prestamista. A partir de las probabilidades se genera un puntaje (*score*) que se le asocia a las variables predictivas para indicar un nivel de riesgo. El método asocia el mismo riesgo a clientes con la mismas características (variables), y su utilización reduce el tiempo para determinar si se concede un préstamo, ya que una vez implementado en el sistema de la institución, es instantáneo en comparación con el tiempo en días que le llevaría a un analista de créditos tomar la decisión.

El modelo *scoring* da un marco para la decisión final en el otorgamiento o rechazo de la solicitud de créditos. Esto permite tener mayor objetividad en el proceso de evaluación del riesgo de crédito. Aunque el modelo no determina con exactitud el comportamiento de un cliente en particular, debe poder estimar el comportamiento promedio de individuos que cumplen características similares. Si esto se satisface puede ser considerado un buen modelo. La eficiencia del método depende de factores como la actualización de los datos, la representación significativa de "buenos" y "malos" clientes; el examen y ajuste periódico del modelo para incorporar los cambios del contexto económico del país, cambios que se reflejan en la relación de los atributos (valores de la variable) y el comportamiento de pago. Estos cambios deben ser explicados por el modelo.

Algunas de las técnicas estadísticas utilizadas para desarrollar un modelo *scoring* son los modelos de análisis multivariado, regresión lineal múltiple, regresión logística, árboles de decisión o redes neuronales, entre otras. A través de estas metodologías podemos construir una *scorecard*, donde se asignan puntajes en diferentes rangos de una variable predictiva. Esta tabla de puntajes estima una probabilidad de aceptación o rechazo de un cliente. Para construir esta tabla de puntajes se utilizan técnicas estadísticas, aunque la *scorecard* definitiva debe contener información propia de la empresa, que no necesariamente tiene una justificación científica; como por ejemplo estas pueden ser controladas por leyes o políticas de la empresa o del país.

Las empresas crediticias deciden el mínimo puntaje (*cut off*) para aceptar un

nuevo cliente, que suele ser en base a sus metas corporativas. También definen las características para clasificar a los nuevos clientes en buenos y malos, normalmente se basará en su experiencia comercial. Para determinar el mínimo puntaje que separará los clientes aceptados y rechazados se evaluan variables como tasas de aprobación de las solicitudes y mora deseada por la institución. Por ejemplo, una institución puede tener la estrategia de aceptar pocos clientes con un bajo riesgo, pero implica poca rentabilidad. Puede también aceptar muchos clientes de alto riesgo que implica mayor cartera vencida que se traducirá en perdida (*write off*), por tanto poca rentabilidad. El encontrar el punto óptimo para discriminar entre buenos y malos clientes no es tarea fácil; sin embargo, la experiencia y la profesionalización del personal, así como el conocimiento integral de la institución permitirá hacer la discriminación de manera más efectiva.

La población donde se obtuvo la muestra con la que se construye el modelo será comparada con la población actual para determinar si existe evidencia de que las poblaciones están idénticamente distribuidas; para tal fin se pueden utilizar pruebas como el *índice de estabilidad* de la población, la prueba de *Kolmogorov Smirnov (K-S)* o la prueba *Ji-cuadrada*. En esta etapa de desarrollo se analiza con una prueba de hipótesis sobre el modelo de regresión logística si las variables extraídas de los registros de los clientes son significativas. Una proporción de la muestra se utiliza para estimar el modelo para discriminar, ésta es llamada muestra de desarrollo. El resto de la muestra se utiliza para validar la efectividad del modelo, a esta parte se le llama muestra de validación. La validación del modelo también puede hacerse con una muestra de clientes recientes.

Después de construir el modelo pasa por un proceso de validación el cual puede hacerse por ejemplo utilizando la R múltiple, la R^2 ajustada o el error estándar si se trata de una regresión lineal, entre otros. Para medir la eficacia del *scorecard* se utiliza el *índice de Gini*, la *divergencia*, el *estadístico K-S* que mide la diferencia entre la media de las distribuciones de buenos y malos; comúnmente se usa el estadístico *k-s* para saber que tanto difieren las poblaciones de buenos y malos.

La información del buró de crédito permite analizar el comportamiento de los clientes con otras instituciones y nos permite precisar mejor el modelo. Las áreas o departamentos del negocio que puedan influir en el modelo deberán participar o aportar información para que el modelo final represente a la empresa y éste no tenga que ser alterado en un futuro por esta falta de información.

El modelo requiere de monitoreo conforme pasa el tiempo. Debe ser evaluado generalmente cada año para comprobar que se encuentra dentro de los márgenes de clasificación con respecto de la nueva población. Para ello se hace uso de técnicas estadísticas y de probabilidad. De tal modo que puedan verse éstas variaciones

para hacer los ajustes pertinentes. El modelo tiene que ser calibrado para que siga siendo efectivo, si con la calibración no queda entonces se desecha el modelo y se debe proponer otro nuevo.

Finalmente, diremos que los tipos de créditos existentes son diversos como son: créditos personales, créditos hipotecarios y créditos revolventes (tarjetas de crédito). En este trabajo se abordará éste último tipo de créditos, pero es importante aclarar que puede utilizarse la misma metodología para estudiar los tres tipos de crédito. El propósito de este trabajo es hacer uso de una colección de técnicas que se utilizan en el *Credit Scoring* y que usan las empresas consultoras para generar una scorecard; esto es, el puntaje (*score*) que se le asigna a las características de los clientes en base a sus datos demográficos y de buró de crédito. Haremos una comparación de la scorecard que obtengamos con una generada por una empresa consultora sobre un segmento de población. Ya que las empresas consultoras sólo entregan resultados como cajas negras, sin indicar que desarrollos llevan a cabo, y además tienen costos elevados para quienes consultan estos servicios.

La tesis está estructurada en cuatro capítulos. En el primer se hace una revisión de las tarjetas de crédito; se estudian, entre otros conceptos, sus características, su utilización, asi como el comportamiento de los clientes cuando son poseedores de un crédito. En el capítulo dos se revisan los conceptos generales de estadística y de probabilidad que son utilizados en el *credit scoring*. El desarrollo general de la técnica credit scoring utilizada para este trabajo se encuentran en el capítulo tres. En el último capítulo se reporta un caso práctico; trataremos una base con datos de tarjetahabientes, a partir de esta base se obtiene el modelo estadístico que nos permite diseñar una scorecard para una población dada. Se cuenta con la *scorecard* diseñada con los mismos datos por una empresa consultora, se desconocen los procedimientos exactos con los que obtuvieron los valores reportados, una de las acciones que se hacen en este capítulo es comparar estadísticamente, sus reultados con los obtenidos por nosotros.

Capítulo 1

Tarjetas de Crédito

La tarjeta bancaria de crédito ha seguido un proceso evolutivo que se remonta hacia el año de 1914, cuando se dio un gran auge en Estados Unidos, país donde fue creada. Las instituciones financieras y las necesidades del mercado fueron dándole el formato que tienen el día de hoy.

En México las tarjetas de crédito empezaron su desarrollo partir de 1956, sin reglamentos específicos aplicables. Es hasta 1967 que la Secretaría de Hacienda y Crédito Público dictó un reglamento para tarjetas de crédito bancarias aplicable exclusivamente a instituciones de crédito que son bancos de depósito. Este reglamento no limitó a otras instituciones que no tenían carácter de institución de crédito para promover su difusión en el país. Actualmente se intenta legislar reglamentos para proteger a los consumidores, debido a la gran demanda de las tarjetas de crédito y al aumento en la cantidad de clientes que dejan de pagar (cartera vencida).

Según Iñigo Ocejo Rojo, delegado de la Condusef en Puebla [ver Apr (2008)], en México existen aproximadamente un 25 % de usuarios de tarjetas de créditos. Pero también se sabe que de ese porcentaje sólo el 8 % cumplen puntualmente con el pago de sus préstamos, 9 % corresponde a cartera vencida, 50 % hace los pagos mínimos y cerca de 30 % trabajan con planes de 6 meses sin intereses.

1.1. Créditos

Entenderemos como crédito al compromiso pactado entre una persona o institución que otorga capacidad de compra por adelantado al deudor, que también puede ser una persona física o moral. El crédito permite realizar ventas y satisfacer la necesidad de compra de los consumidores, conforme a su capacidad de pago. Las

condiciones del convenio que permiten el acuerdo comercial pueden ser flexibles y negociables en cuanto a plazos, montos, tipo de interés, etc.; con la finalidad de concluir en buenos términos el compromiso de crédito adquirido.

A la parte otorgante de crédito la llamaremos institución crediticia y a la parte deudora como sujetos de crédito o simplemente clientes. Los posibles clientes son aquellos que reúnen requisitos para que se les otorgue un crédito, en efectivo o venta de un artículo con pagos en cantidades más pequeñas en un plazo de tiempo determinado. Estos requisitos son propios de la política de cada empresa otorgante, aunque existen variables comunes, se está en busca de encontrar mejores maneras de determinarlas.

En esta dinámica comercial se busca aumentar el volumen de ventas por parte de las instituciones involucradas en la concesión del crédito, con el propósito de aumentar sus ingresos y rentabilidad. Proveen de dinero efectivo o electrónico a personas físicas o morales para adquirir artículos o servicios con facilidades de pago, cuando éstos no cuentan con el capital propio disponible para tal transacción.

Los clientes una vez que han sido aceptados deben dejar garantías reales o prendarias. Los pagos son convenidos a través de documentos por cobrar como facturas, letras, pagarés, etc. Estos documentos deben contener la tasa de interés pactada, monto de crédito, plazos y modalidad de pago.

Cuando los clientes no cumplen con los compromisos adquiridos, la institución crediticia los cataloga como clientes morosos o malos clientes, según sus políticas de cobro. Y ésta información es enviada al buró de crédito, institución que guarda la información del comportamiento de pago del cliente en todos sus créditos. La información se envía mensualmente y registra hasta 24 meses.

Las instituciones que se dedican a otorgar créditos son los bancos y las instituciones financieras, así como también empresas comerciales, industriales, de servicios, etc. No importando la empresa o la institución que otorga el crédito, en todos los casos se requiere disminuir el riesgo de moratoria de pagos.

1.1.1. Tipos de créditos

Créditos de Consumo o Créditos comerciales. Son aquellos créditos otorgados por empresas para la adquisición de bienes o servicios de uso personal en plazos determinados.

Créditos Empresariales. Cuando una empresa requiere materia prima, insumos, servicios, etc. solicita el bien o servicio a otras empresas a crédito para continuar su actividad empresarial, realizando convenios para cubrir el adeudo en un futuro.

Créditos Bancarios. Son los otorgados por entidades bancarias o empresas del

sistema financiero a personas físicas o morales que necesitan recursos para financiar sus actividades. Como la adquisición de bienes, servicios, pagar deudas, etc.

Dentro de los productos financieros que ofrecen las instituciones crediticias se encuentran las tarjetas de crédito.

1.2. Tarjetas de crédito

Es un pequeño crédito provisto a través de una tarjeta de plástico personalizada por lo cual también se le llama dinero plástico. La tarjeta de crédito sustituye al dinero en efectivo. Son utilizadas cuando las personas se ven en la necesidad de adquirir productos, servicios, cancelar deudas y no cuentan con efectivo a la mano. Las tarjetas de crédito sirven para adquirir productos y aplazar sus pagos durante

Figura 1.1: Tarjeta de crédito

varios meses. Las compras realizadas durante un periodo se acumulan en un saldo mensual. Estos bienes anticipados pueden ser pagados al final del mes, a plazos, o un saldo mínimo. Se puede gastar hasta un límite concedido y el crédito se repone automáticamente una vez se ha pagado la deuda de la tarjeta. El periodo durante el cual se pueden realizar cargos en comercios y establecimientos con la tarjeta sin que se cobren intereses se le llama periodo de gracia.

1.2.1. Funcionamiento de las tarjetas de crédito

Son utilizadas en cajeros automáticos y medios electrónicos en los comercios. También es utilizado como requisito para adquirir algún bien o servicio, ya que provee información del tarjetahabiente. Los agentes que intervienen en éste proceso económico están dados por

a) La institución de crédito o entidad emisora. Quien proporciona la tarjeta. Además realiza el financiamiento de la compra mediante el crédito. Pueden ser empresas comerciales que emiten sus propias tarjetas de crédito a sus clientes o entidades financieras como los bancos.

b) El usuario o tarjetahabiente. Puede ser una persona física o moral, autorizada por las entes emisoras después de un adecuado examen crediticio para hacer uso de la tarjeta.

c) Comercio receptor o proveedor. Es quien acepta la tarjeta como medio de pago.

d) El banco con el que el comerciante tiene pactado el servicio de la tarjeta. Es quien proporciona una terminal, se le llama banco adquiriente. Por el uso de su terminal cobra un porcentaje del valor de la compra al comercio receptor.

1.2.2. Clasificación de las tarjetas de crédito

Se pueden clasificar por su naturaleza, fórmulas de pago, acceso a esos créditos y su objetivo final. Entre ellas encontramos

a) Por su alcance:

- Las tarjetas *locales* aplican solo en el país de origen.
- Las tarjetas *internacionales* pueden ocuparse para hacer compras en cualquier país del mundo; las más aceptadas en México, son Visa, Mastercard y American Express.

b) Por su emisor:

- Las tarjetas de crédito *departamentales* son aquellas que proporcionan los establecimientos comerciales a su clientela para otorgarles productos o servicios que ellos ofrecen, es también un elemento del crédito al consumo, que trataremos más adelante.
- Se llaman *bancarias* a las que proporcionan los bancos para la compra de bienes o servicios a cargo de terceros, aunque también provee de manera directa dinero efectivo. Este es un tipo de crédito inmediato que otorgan los comercios, sin que exista necesariamente alguna relación entre el tarjetahabiente y el establecimiento afiliado.

c) Por su fórmula de pago y acceso a los créditos:

- Tarjeta de crédito *clásica* o *estándar*. El monto total del crédito adquirido se cobra generalmente después de un mes vencido. Si no se sobrepasa el límite y se paga el capital endeudado en el periodo de gracia establecidos por el ente emisor no se cobran intereses. En otras palabras, se obtiene una financiación a 30 días sin costo alguno. En caso de que no se pague el saldo completo se comenzarán a cobrar intereses como parte del préstamo. El costo anual por contar con el servicio de la tarjeta de crédito es bajo, la línea de crédito generalmente es baja y dado que este nivel de producto es mas riesgoso sus intereses ordinarios son altos. Estos porcentajes tan elevados servirán para pagar el porcentaje de los clientes que no pagarán.
- Tarjetas *oro*. Similarmente a las tarjetas clásicas cuenta con límites de crédito alto medio, con anualidades altas e intereses altos que oscilan alrededor de los 55 %. Suelen tener incluidos una serie de servicios adicionales y preferencias. Se les otorga a clientes con ingresos altos, pero aun no hay garantía de que se trate de un buen cliente. Para acceder a esta línea de crédito se requiere aproximadamente de dos años de historia crediticia en la institución, tiempo en que el cliente demuestre ser redituable para la empresa.
- Tarjetas *platinum*. Están destinadas a los llamados clientes VIP, a aquellos que hagan un uso muy frecuente de su tarjeta y que fueron usuarios de las líneas de crédito anteriores; por lo que pertenece a un número selecto de usuarios dentro del mercado total de tarjetahabientes. Su anualidad es alta, límite de crédito alto e intereses bajos que varían en un 25 %. Este producto puede ser otorgado después de tres años de ser usuario cumplido de los dos productos anteriormente descritos.

Las tarjetas de crédito manejan tres tipos de interés:

a) Tasa revolvente. Es la tasa de interés anual por hacer uso de la tarjeta y está en función de la TIIE (Tasa de Interés Interbancaria de Equilibrio).

b) Tasa de interés fija. No cambia con el tiempo, depende del plazo del cliente.

c) Tasa cero. Significa cero porcentaje de interés aplicado sobre la deuda adquirida durante un periodo determinado.

1.3. Crédito al consumo

El financiamiento otorgado por establecimientos comerciales o de autoservicio sobre *bienes de consumo duradero* [1] a plazos o en "pagos chiquito", se le llama *crédito al consumo*. En este tipo de crédito se pueden adquirir bienes y servicios personales, y no necesariamente se hace uso de una tarjeta, depende del establecimiento que provee el financiamiento.

Se otorgan éste tipo de créditos comúnmente con pocos requisitos y no se considera el historial crediticio del candidato. En algunas cadenas solo se pide comprobante de domicilio y la identificación oficial para otorgar el préstamo; por esta razón la tasa de interés suele ser alta. Con los bancos también se accede a este tipo de productos por medio de las tarjetas de crédito, solo que como ya mencionamos requieren generalmente de un tercero; que es quien provee el bien a consumir. Como requisitos algunas instituciones piden cuenta de nómina y si es necesario se puede completar el requisito con la nómina de su cónyuge para que obtengan el préstamo.

Según cifras del Banco de México (Banxico), el crédito al consumo ha crecido en los último años en el país, en solo siete años aumentó en un 212.5 % [ver Moreno (2008)]. Obtener un crédito al consumo generado básicamente por establecimientos comerciales, suele ser sencillo, y utilizarlo de manera poco adecuada puede incurrir en morosidad o en deudores de la cartera vencida, similar como deberle al banco. De manera similar a los créditos bancarios, se paga comisiones de apertura e intereses moratorios si no se liquida la deuda antes de la fecha de corte, o se cubre el pago mínimo requerido.

Según informes de la Comisión Nacional Bancaria de Valores (CNBV) y del Banco de México (BdeM), la cartera vencida en el crédito al consumo sufrió un incremento de 50.3 por ciento con respecto al monto registrado en diciembre de 2007. Al término del año pasado el índice de morosidad, el cual determina la proporción de la cartera vencida respecto de la total del crédito al consumo, se elevó a 8.75 por ciento, desde 8.49 en noviembre anterior [ver Lino (2009)].

El segmento de la cartera vencida correspondiente a tarjetas de crédito elevó de 10.32 a 10.51 por ciento su índice de morosidad entre noviembre y diciembre de 2008. Aun cuando los bancos reducen el saldo de los créditos al consumo, el financiamiento bancario ha mostrado elevaciones en la cartera vencida. Esto significa

[1] Bienes de consumo duradero son las mercancías que tienen una vida útil mayor a un año y que son demandadas por los agentes económicos: familias, empresas y gobierno, para su funcionamiento y/o manutención, tales como casa habitación, automóviles y enseres domésticos, entre otros.

que los bancos prestan menos, pero los deudores se retrasan más en sus pagos. En los créditos personales el índice de morosidad aumentó de 5.93 a 6.53 por ciento de enero a diciembre del año pasado [ver Zúñiga y Rodríguez (2009)].

1.3.1. Ciclo de riesgo

Hablando muy general del crédito al consumo, se puede describir desde el punto de vista de la administración de riesgo un ciclo en el cual participa todo cliente. Este se describe como sigue (véase figura 1.2).

Figura 1.2: Ciclo de riesgo

- *Originación.* La intención en este punto es otorgar crédito a un cliente por primera vez en la institución.
- *Administración.* La intención en esta parte de ciclo es premiar a los clientes que se están "portando bien" (incrementos de límite de crédito) y castigar a los que se están "portando mal" (decrementos de límite de crédito); en ambos casos se pueden combinar estrategias de incrementos o decrementos de tasas de interés, la reestructuración de deudas, etc. Aquí se busca la detección temprana de cuentas de alto riesgo y poder realizar acciones tempranas de corrección.
- *Recuperación.* En esta parte del ciclo de riesgo se pretende recuperar a todos aquellos clientes que dejaron de pagar. Se aplican actividades de recaudación

a clientes con un alto puntaje de *score* según la empresa y determinar los clientes no recuperables para hacer el traspaso a una empresa recaudadora y así recuperar parte del capital perdido.

1.4. Estatus de los clientes en los pagos

Cuando un cliente compra un artículo con su tarjeta de crédito, la institución crediticia le concede un periodo de un mes generalmente para liquidar su adeudo sin cobrarle intereses. En teoría se espera que el cliente cuente con el dinero para pagar su adquisición y se mantenga al corriente. Si el cliente no cuenta con el capital suficiente para cubrir el monto total puede cubrir un cantidad que va del mínimo establecido por el banco hasta el saldo total antes de la fecha de límite de pago, y quedará como saldo el monto del adeudo. 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.

a) 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.

b) Fecha de corte. Es un día de cada mes en que la institución financiera reviza el estado de cuenta del cliente en particular. También se le conoce como fecha de facturación, donde 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.

c) Pago mínimo. Es el porcentaje mínimo aplicado al monto del adeudo, que debe cubrirse antes de la fecha límite de pago. Se considera como el pago de intereses anticipados. El precio por no pagar su cuenta en tiempo y forma.

1.4.1. Posibles condiciones de un cliente

Cuando paga su adeudo entre la fecha de corte y la fecha límite de pago, no se le cobran intereses de mora sobre su saldo, que resulta ser alrededor de 23 días, según la institución de crédito y se considera un cliente al corriente *"current"*.

Si 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 aun 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. Ésto se realiza entre la fecha límite de pago y la próxima fecha de corte, aproximadamente una semana y se le conoce como tiempo *"prevent"*. 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 correspondiente. No se le aplicarán intereses moratorios, pero no evitará el cobro de *"call center"* por 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 $B1$. Ésta información es enviada al buró de crédito. Las mismas acciones son aplicadas a un cliente que no realizo algún pago. Si se paga más del mínimo y menos del total del saldo facturado se cargan intereses revolventes.

Así que la cartera de crédito se clasifica en las siguientes categorías:

Calificación	**Tiempo en mora**
Bucket 0 (B_0)	0 días (al corriente)
Bucket 1 (B_1)	1 a 29 días
Bucket 2 (B_2)	30 a 59 días
Bucket 3 (B_3)	60 a 89 días
Bucket 4 (B_4)	90 a 119 días
Bucket 5 (B_5)	120 a 149 días
Bucket 6 (B_6)	150 a 179 días
Bucket 7 (B_7)	Mayor o igual a 180 días

Para comprender el proceso de la historia de crédito de un cliente analizamos sus inicios con el siguiente ejemplo.

Ejemplo 1.4.1. *Supongamos que un individuo se le otorga una tarjeta de crédito departamental el 2 de enero (ver figura 1.3). Decide que su fecha límite de pago sea el día 18 de cada mes, con facturación los días 25. El 5 de enero realiza compras con un monto de $5000 .*

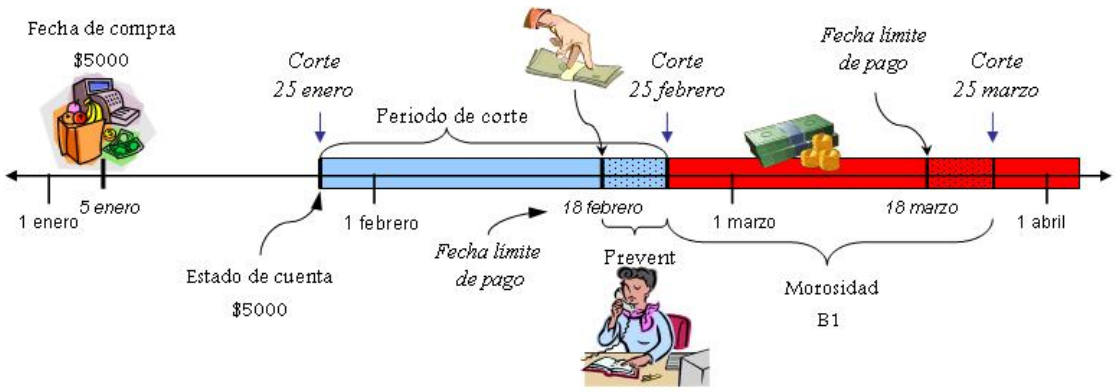


Figura 1.3: Estructura general en los eventos y tiempos asociados a una tarjeta de crédito.

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 facturaran 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 la canasta* $B1$*, del 25 de marzo al 25 de abril la canasta* $B2$ *y asi 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.*

1.4.2. Definición de buenos y malos clientes

Primero tendremos que definir el concepto de buenos o malos clientes y como asignar ésta clasificación. Este concepto depende de la experiencia de la compañía, de la cartera en mora, del proceso de cobranza, de la población, etc. También se puede tomar de la experiencia de otras empresas para construir la definición. Conforme avanza el tiempo se va identificando el comportamiento de pago y se asigna en alguno de los siguientes casos.

1. Se considera buen cliente a aquellos individuos que

   a) Pagan el monto de su deuda en el periodo de gracia (entre los límites de pago o de corte).

   b) Cuando no cuenta con el capital para pagar la totalidad de la deuda pero pagan al menos el mínimo requerido por la empresa acreedora.

c) Liquidan su adeudo en no más del tiempo determinado por la empresa. Generalmente como máximo a los tres meses o que no pasen de B_3.

d) Si el cliente oscila entre B_1 y *current*. Este es un caso conveniente para la institución dado que genera mayores ganancias que los otros casos.

2. Un cliente intermedio es aquel deudor que no podemos clasificar como bueno o malo. Debido a que tiene adeudos de capital e intereses acumulados y se requiere de más tiempo para ver la tendencia de su comportamiento.

3. Un mal cliente se refiere a aquel deudor que causa pérdidas económicas a la compañía. Este tipo de clientes no pagaron su cuenta, aun después de aplicarles técnicas de cobranzas. Comúnmente es considerado mal cliente después de 90 días de mora (B_4). Por experiencia se sabe que cuando ya está en esta etapa o estado es difícil que regrese al estado B_0 de buen cliente o *current*.

1.5. Estados de los clientes

Cuando se otorga un crédito se espera que sean cubiertos en un plazo determinado, de tal modo que se generen ganancias. Un mal cliente primero fue considerado bueno, después pasa a un tiempo de moratoria. Finalmente pasa al grupo de perdida (*write off*), un estado del que ya no saldrá, donde se considera que el adeudo ya es incobrable.

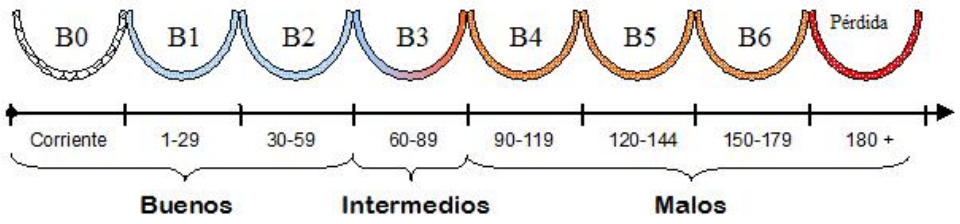


Figura 1.4: Canastas. Posibles estados en los que puede caer un cliente.

Los estados de cuenta se realizan en la fecha de corte de cada mes y se determina el estado de los clientes. Los clientes pasan de un estado a otro dependiendo de los pagos que realicen. Si los clientes pagan toda su deuda se colocan en la canasta

current y ya no se moverán a otro estado. Si se encuentran en algún estado y pagan el mínimo permanecen en el estado en que se encuentran. Un cliente puede retornar una o más canastas hacia atrás, cuando reestructura su deuda, dependiendo del pago que haga.

Ejemplo 1.5.1. *La figura 1.4 nos muestra una clasificación conforme a los tiempos de mora. La primera canasta B0 corresponde a deuda pagada (estado 0), de 1 a 29 días de mora, cubeta B1 (estado 1) y así sucesivamente hasta la canasta B6 (estado 6). A partir de los 180 días ya está en* write off. *Entre 1 y 59 días de retraso se clasifica como bueno, entre 60 y 89 días un cliente que no se puede catalogar como bueno o malo y más de 90 días un mal cliente.*

Capítulo 2

Herramientas de Estadística y Probabilidad

En éste capítulo estudiaremos algunas técnicas que se utilizan para obtener el credit *scoring*. Comenzaremos con las cadenas de Markov y las matrices de transición [ver Randolph (1995)].

2.1. Cadenas de Markov: matrices de transición

Una tarea importante en el proceso de *credit scoring* es clasificar a los clientes de acuerdo a su estatus moratorio, esto es de acuerdo al número de días que se encuentre en mora se puede clasificar a los clientes en buenos, malos o indeterminados. Queremos hacer predicciones sobre el comportamiento futuro de los clientes y para este fin es conveniente conocer la probabilidad de que un cliente pase de un *bucket* (*canasta*) a otro a través del tiempo y para ello son útiles las matrices de transición. Los clientes pasan entre *buckets* (Figura 1.4) con cierta probabilidad; esto es, la distribución de los clientes en las diferentes canastas cambia a través del tiempo. Es deseable estimar cuántos días de mora podemos considerar permisible para que un cliente bueno, pase a ser un cliente no definido y luego se convierta en un cliente malo, es decir; cómo evoluciona su estatus en el tiempo de mora. Calcular la probabilidad con la que un cliente puede llegar a perdida es de suma importancia.

2.1.1. Cadenas de Markov

Cuando estudiamos la probabilidad de que un cliente pase de una canasta a otra a través del tiempo, estamos ante un proceso estocástico. Un proceso estocástico se define como una familia $\{X(t), t \in T\}$ de variables aleatorias; donde t representa los valores del tiempo, tomados de un conjunto T que pueden ser discretos o continuos. Por ejemplo el número de clientes en un banco resulta ser un proceso estocástico de tiempo continuo [ver Hillier (1997)].

El tipo de procesos estocásticos que analizaremos para obtener el *credit scoring* son aquellos con variable en tiempo discreto. Estos se representan con índices enteros, de la forma $\{X_1, X_2, \ldots\}$ o simplemente X_t. Los posibles valores que puede tomar la variable $X(t)$ se llaman categorías o *estados*. El conjunto de estados se denota con la letra S y es un conjunto finito. Los estados son eventos mutuamente excluyentes. La variable X_t puede tomar valores cualitativos o cuantitativos, que pueden ser representados mediante etiquetas numéricas 0, 1, ..., M sin perder generalidad. Por ejemplo

1. X_t representa la marca del producto que prefiere un consumidor en la semana t.

2. X_t representa la canasta de moratoria en que se encuentra un tarjetahabiente en el tiempo t. En este ejemplo los estados son las canastas B_0, B_1,...,B_6 y B_7.

Un cliente puede estar en uno de esos estados en un determinado mes. Podríamos considerar a X_0 como el estado del cliente cuando se le otorga el crédito, aunque aquí el único valor posible es *current* o B_0. Para $t = 1$, X_1 representa los estados en los que el cliente se encuentra al terminar el primer mes, esto es B_0 o B_1. Al final del segundo mes $t = 2$, X_2 tiene como posibles estados B_0, B_1 y B_2; y así sucesivamente. Únicamente a partir de X_7 el sistema toma cualquiera de los estados B_i, $i = 0, 1..., 7$.

Las cadenas de Markov son procesos estocásticos que tienen la propiedad de que para cualquier lapso de tiempo, la probabilidad condicional de transición de un estado a otro no depende de los eventos ocurridos en el pasado, solo depende del estado actual del proceso, se dice que el sistema no tiene memoria del pasado.

Definición 2.1.1. *Un proceso estocástico con tiempo discreto t y el espacio de estados, S, en el conjunto de enteros no negativos, es una cadena de Markov, si satisface la propiedad Markoviana*

$$P\{X_{t+1} = j \mid X_0 = x_0, X_1 = x_1, \ldots, X_{t-1} = x_{t-1}, X_t = i\}$$

$$= P\{X_{t+1} = j \mid X_t = i\}, \tag{2.1}$$

para cualquier valor de t=0, 1, ... y cualquier par de estados i, j, los valores x_0, x_1, ..., $x_{t-1} \in \{0, 1, \ldots, M\}$. *Las probabilidades condicionales*

$$P\{X_{t+1} = j \mid X_t = i\}$$

se llaman probabilidades de transición.

Esto significa que la probabilidad de que estemos en el estado $X_{t+1} = j$ solo depende del estado anterior $X_t = i$. El estado presente $X_t = i$ contiene toda la información sobre la evolución del pasado del proceso y es suficiente para determinar la distribución del proceso en el futuro.

Definición 2.1.2. *Una cadena de Markov es de tiempo homogéneo o estacionaria si, para cada i,j* $\in S$,

$$P\{X_{t+n} = j \mid X_t = i\} = P\{X_n = j \mid X_0 = i\}, \;\; t = 0, 1, \ldots, \;\; n \geq 0.$$

Dado que esta probabilidad depende únicamente de la diferencia entre t y $t + n$, *entonces las probabilidades de transición estacionaria de n pasos se denotan por*

$$p_{ij}^{(n)} = P\{X_{t+n} = j \mid X_t = i\}$$

Los valores de n representan el número de pasos o unidades en el tiempo entre un estado y otro. De lo anterior $p_{ij}^{(n)}$ denota la probabilidad de estar en el estado j después de n unidades de tiempo partiendo del estado i, en otras palabras, se pasa del estado i al estado j en determinado periodo de tiempo fijo. Para $n = 1$ tenemos

$$P\{X_{t+1} = j \mid X_t = i\} = P\{X_1 = j \mid X_0 = i\}.$$

Las probabilidades de transición estacionarias de un paso se denotan por

$$p_{ij} = P\{X_{t+1} = j \mid X_t = i\}.$$

Para interpretar esto consideramos el ejemplo de preferencia de marcas, la probabilidad que un consumidor prefiera en la siguiente semana la marca j, dado que en la presente semana compró la marca i se conserva para cualesquiera dos semanas consecutivas.

2.1.2. Matrices de transición

Para una cadena de Markov con probabilidades p_{ij} estacionaria tendremos las siguientes propiedades:

a) Las probabilidades de transición son no negativas $p_{ij} \geq 0$ para toda i, j.

b) La suma de las probabilidades de transición es uno, para toda i, en cualquier número de pasos $n = 0, 1, ...$ y M el número de estados.

$$\sum_{j=0}^{M} p_{ij}^{(n)} = 1$$

c) Se puede escribir las probabilidades de transición para $n= 0, 1, 2, ...$ pasos en forma de tabla

$\mathbf{P}^{(n)}=$

Estado	0	1	$\cdots$	M
0	$p_{00}^{(n)}$	$p_{01}^{(n)}$	$\cdots$	$p_{0M}^{(n)}$
1	$p_{10}^{(n)}$	$p_{11}^{(n)}$		$p_{1M}^{(n)}$
$\vdots$			$\vdots$	
M	$p_{M0}^{(n)}$	$p_{M1}^{(n)}$	$\cdots$	$p_{MM}^{(n)}$

o en forma matricial, llamada *matriz de transición*

$$\mathbf{P}^{(n)} = \begin{bmatrix} p_{00}^{(n)} & p_{01}^{(n)} & \cdots & p_{0M}^{(n)} \\ p_{10}^{(n)} & p_{11}^{(n)} & \cdots & p_{1M}^{(n)} \\ & \vdots & & \\ p_{M0}^{(n)} & p_{M1}^{(n)} & \cdots & p_{MM}^{(n)} \end{bmatrix}$$

Para una matriz de transición de un paso escribimos las probabilidades sin el superíndice. Retomando el ejemplo de preferencia de un producto en diferentes marcas tenemos el siguiente ejemplo.

Ejemplo 2.1.3. *Supongamos que el comportamiento de los consumidores se puede modelar con una cadena de Markov. La Tabla 2.1 muestra a 250 consumidores en un periodo de dos semanas. Compiten tres marcas para el producto de consumo, la elección se ha hecho en la semana 7 y el cambio de marca en la semana 8. Además supongamos que los datos simulan el comportamiento de la población.*

			Marca en la semana 8	
Marca en la semana 7	1	2	3	Total
1	72	4	4	80
2	12	102	6	120
3	2	6	42	50
Total	86	112	52	250

Cuadro 2.1: Consumidores y sus preferencias de la semana 7 a la 8

De los 250 consumidores en la semana siete, 80 adquirieron la marca uno, 120 la marca dos y 50 la marca tres. En el primer renglón que corresponde al estado uno en la semana siete, del total de 80 consumidores 72 adquirieron la misma marca en la semana ocho, 4 prefirieron la marca dos y 4 optaron por la marca tres. En la columna del estado uno, se tiene que 12 clientes de la marca dos y 2 de la tres se pasaron a la marca uno. La marca uno tuvo una pérdida de 8 clientes pero gano conquistando 14 de las otras marcas. La matriz de transición se puede representar con una matriz de tres estados. Las probabilidades de transición se estiman con los datos de la tabla 2.1, de acuerdo a la fórmula $\hat{p}_{ij} = \frac{n_{ij}}{n_i}$, para todo i, j donde n_{ij} es el número de consumidores que escogieron la marca i en la semana siete y cambiaron a la marca j en la semana ocho. El denominador n_i es el número de consumidores de la marca i al comienzo del periodo. La matriz de transición estimada es

$$\widehat{\mathbf{P}} = \begin{bmatrix} \hat{p}_{11} & \hat{p}_{12} & \hat{p}_{13} \\ \hat{p}_{21} & \hat{p}_{22} & \hat{p}_{23} \\ \hat{p}_{31} & \hat{p}_{32} & \hat{p}_{33} \end{bmatrix} = \begin{bmatrix} \frac{72}{80} & \frac{4}{80} & \frac{4}{80} \\ \frac{12}{120} & \frac{102}{120} & \frac{6}{120} \\ \frac{2}{50} & \frac{6}{50} & \frac{42}{50} \end{bmatrix}$$

$$\widehat{\mathbf{P}} = \begin{bmatrix} 0.90 & 0.05 & 0.05 \\ 0.10 & 0.85 & 0.05 \\ 0.04 & 0.12 & 0.84 \end{bmatrix} \tag{2.2}$$

La entrada $p_{12} = P\{X_8 = 1 | X_7 = 2\} = 0.05$ la entenderemos como la probabilidad que un cliente que compró la marca dos en la semana siete, prefiera la marca uno en la semana ocho; es decir, se estima que el 10 % haga esta transición. La suma de las entradas de cada renglón suman uno. Las columnas contienen las probabilidades que un cliente haga su transición a la marca j, la segunda columna nos da la proporción de clientes en cada marca que se pasan a la marca dos. Los elementos de la diagonal representan la probabilidad de que no haya cambio de

marcas, se conoce como *poder de retención*, los valores de $i \neq j$ miden el *poder de atracción* de la marca j.

Ejemplo 2.1.4. *En una tienda departamental el estado de los clientes es considerado según el número de pagos vencidos (PV) de un mes a otro. Los clientes realizan sus pagos mes con mes y se clasifican al corriente con cero pagos vencidos (estado 0 PV), con un pago vencido (1 PV), dos pagos vencidos (2 PV), tres pagos vencidos (3 PV) y cuatro o más pagos vencidos (4 PV). La matriz representa la transición de los clientes en cuanto al número de pagos vencidos en el transcurso de un semestre.*

Estado	*0 PV*	*1 PV*	*2 PV*	*3 PV*	*4 PV*
0 PV	*0.898*	*0.047*	*0.014*	*0.012*	*0.029*
1 PV	*0.520*	*0.120*	*0.070*	*0.050*	*0.240*
2 PV	*0.287*	*0.099*	*0.079*	*0.060*	*0.475*
3 PV	*0.212*	*0.050*	*0.051*	*0.061*	*0.626*
4 PV	*0.170*	*0.020*	*0.030*	*0.040*	*0.740*

Observamos que los clientes que al inicio del semestre tienen cero pagos vencidos la probabilidad que pasen a cuatro o más pagos vencidos es pequeña de 0.29. Mientras que los que al inicio del semestre están en cuatro o más pagos vencidos tienen una probabilidad alta de 0.74 que sigan en ese estado.

2.1.3. Probabilidades de transición en n pasos

Ahora se va a encontrar una expresión para la probabilidad de transición en n pasos $P(X_n = j|X_0 = i)$. El proceso se hace por inducción, se va a comenzar con $n = 2$. Se tiene que $P(X_2 = j|X_0 = i)$ por definición de probabilidad condicional, se sigue que

$$P(X_2 = j|X_0 = i) = \frac{P(X_2 = j, X_0 = i)}{P(X_0 = i)}. \tag{2.3}$$

Ahora se considera el numerador de esta expresión, por el teorema de probabilidad total

$$\begin{aligned} P(X_2 = j, X_0 = i) &= \sum_{k=0}^{M} P(X_2 = j, (X_1 = k, X_0 = i)) \\ &= \sum_{k=0}^{M} P(X_2 = j|X_1 = k, X_0 = i)P(X_1 = k, X_0 = i), \end{aligned}$$

y por la propiedad de falta de memoria, se sigue que

$$P(X_2 = j, X_0 = i) = \sum_{k=0}^{M} P(X_2 = j|X_1 = k)P(X_1 = k, X_0 = i).$$

Sustituyendo esta expresión en (3.3) se sigue que

$$\begin{aligned} P(X_2 = j|X_0 = i) &= \frac{\sum_{k=0}^{M} P(X_2 = j|X_1 = k)P(X_1 = k, X_0 = i)}{P(X_0 = i)} \\ &= \sum_{k=0}^{M} P(X_2 = j|X_1 = k)P(X_1 = k|X_0 = i). \end{aligned}$$

Por la propiedad de estacionalidad, se tiene que

$$P(X_2 = j, X_0 = i) = \sum_{k=0}^{M} P(X_1 = j|X_0 = k)P(X_1 = k|X_0 = i), \tag{2.4}$$

esta expresión se puede escribir como

$$p_{ij}^{(2)} = \sum_{k=0}^{M} p_{ik}p_{kj},$$

que es el elemento i, j de la matriz $\mathbf{P}^2$, por lo tanto $\mathbf{P}^{(2)} = \mathbf{P}^2$. Ahora consideremos que se satisface $\mathbf{P}^{(n-1)} = \mathbf{P}^{n-1}$ y se va a probar que esto implica que $\mathbf{P}^{(n)} = \mathbf{P}^n$. Entonces,

$$p_{ij}^{(n)} = P(X_n = j|X_0 = i) = \frac{P(X_n = j, X_0 = i)}{P(X_0 = i)}, \tag{2.5}$$

considerando el numerador de la expresión (3.5)

$$\begin{aligned} P(X_n = j, X_0 = i) &= \sum_{k=0}^{M} P(X_n = j, (X_1 = k, X_0 = i)) \\ &= \sum_{k=0}^{M} P(X_n = j|X_1 = k)P(X_1 = k, X_0 = i). \end{aligned}$$

De aquí se sigue que

$$\begin{aligned} p_{ij}^{(n)} &= \frac{\sum_{k=0}^{M} P(X_n = j|X_1 = k)P(X_1 = k, X_0 = i))}{P_0 = i} \\ &= \sum_{k=0}^{M} P(X_{n-1} = j|X_0 = k)P(X_1 = k|X_0 = i), \end{aligned}$$

expresión que se puede escribir como

$$p_{ij}^{(n)} = \sum_{k=0}^{M} p_{ik} p_{kj}^{(n-1)}.$$

Este término es el elemento i, j de la matriz $\mathbf{P}\mathbf{P}^{(n-1)}$, y por la hipótesis de inducción se sigue que

$$\mathbf{P}^{(n)} = \mathbf{P}\mathbf{P}^{n-1} = \mathbf{P}^n.$$

Esto significa que la matriz de probabilidades de transición después de n pasos, se calcula como n veces el producto de la matriz de probabilidades de transición.

Del ejemplo 2.1.3 calculamos $\mathbf{P}^2$, dado por

$$\begin{aligned} \mathbf{P}^2 &= \begin{bmatrix} 0.90 & 0.05 & 0.05 \\ 0.10 & 0.85 & 0.05 \\ 0.04 & 0.12 & 0.84 \end{bmatrix} \begin{bmatrix} 0.90 & 0.05 & 0.05 \\ 0.10 & 0.85 & 0.05 \\ 0.04 & 0.12 & 0.84 \end{bmatrix} \\ &= \begin{bmatrix} 0.818 & 0.0935 & 0.0895 \\ 0.177 & 0.7335 & 0.0895 \\ 0.0816 & 0.2048 & 0.7136 \end{bmatrix}. \end{aligned} \tag{2.6}$$

Así mismo podemos calcular

$$\begin{aligned} \mathbf{P}^3 &= \mathbf{P}^2 \cdot \mathbf{P} \\ &= \begin{bmatrix} 0.818 & 0.0935 & 0.0895 \\ 0.177 & 0.7335 & 0.0895 \\ 0.0816 & 0.2048 & 0.7136 \end{bmatrix} \begin{bmatrix} 0.90 & 0.05 & 0.05 \\ 0.10 & 0.85 & 0.05 \\ 0.04 & 0.12 & 0.84 \end{bmatrix} \\ &= \begin{bmatrix} 0.7482 & 0.1311 & 0.1207 \\ 0.2362 & 0.6431 & 0.1207 \\ 0.1225 & 0.2638 & 0.6137 \end{bmatrix} \end{aligned}$$

Sea $\Pi_t^{(0)} = (\pi_{00}, \pi_{10}, \ldots, \pi_{M0})$ el vector de las probabilidades no condicionadas, esto es $\Pi_t(i) = P(X_t = i)$ tal que

$$\begin{aligned} \Pi_j^{(n)} &= P(X_n = j) = \sum_{k=0}^{M} P(X_n = j, X_0 = k) \\ &= \sum_{k=0}^{M} P(X_n = j | X_0 = k) P(X_0 = k) \\ &= \sum_{k=0}^{M} p_{ik}^{(n)} \Pi_k^{(0)}, \end{aligned}$$

entonces

$$\Pi^{(n)} = \mathbf{P}^n \Pi^{(0)}.$$

Si $\Pi^{(0)} = (0.30, 0.38, 0.32)$ entonces $\Pi^{(3)}$ es

$$\begin{aligned} \Pi^{(3)} &= \begin{bmatrix} 0.30 & 0.38 & 0.32 \end{bmatrix} \begin{bmatrix} 0.7482 & 0.1311 & 0.1207 \\ 0.2362 & 0.6431 & 0.1207 \\ 0.1225 & 0.2638 & 0.6137 \end{bmatrix} \\ &= \begin{bmatrix} 0.3234 & 0.3681 & 0.2785 \end{bmatrix}. \end{aligned}$$

Observamos que las entradas $(i,\ j)$ de la matriz de transición en dos pasos $\mathbf{P}^2$ como se muestra en (3.6) se calcularon de la forma

$$p_{ij}^{(2)} = \sum_{k=0}^{M} p_{ik} p_{kj}.$$

Como ya vimos $\mathbf{P}^{(n)} = \mathbf{P}^n$ y por propiedades de multiplicación de matrices

$$\mathbf{P}^n = \mathbf{P}^m \mathbf{P}^{(n-m)}, \quad 0 \leq m \leq n,$$

lo escribimos en términos de las entradas,

$$p_{ij}^{(n)} = \sum_{k=0}^{M} p_{ik}^{(m)} p_{kj}^{(n-m)}, \quad 0 \leq m \leq n. \tag{2.7}$$

Las ecuaciones (2.7) nos dicen que si comenzamos en el estado i después de m pasos se estará en el estado k y después al estado j en $n-m$ pasos. Estas ecuaciones son llamadas las *ecuaciones de Chapman-Kolmogorov*.

2.1.4. Estados absorbentes

Definición 2.1.5. *Se denota por f_{ij} a la probabilidad que estando en el estado i pase alguna vez el estado j. Así f_{ii} es la probabilidad de que estando en el estado i alguna vez regrese al estado i. Cuando $f_{ii} = 1$ el estado i se llama* ***recurrente****, quiere decir que regresará a sí mismo consecutivamente después de un cierto número de pasos; si $f_{ii} < 1$ se denomina* ***estado transitorio***

Un caso especial de estados recurrente son los estados absorbentes; un estado absorbente es aquel del que no se puede salir. Si i es un estado absorbente, entonces

$f_{ii} = p_{ii} = 1$ y $f_{ik} = 0$ para toda $k \neq i$. Así que, si una cadena de Markov contiene algún estado absorbente, la matriz de transición contendrá un uno en la posición ii de la diagonal y en el resto de la fila correspondientes ceros. Si la cadena contiene únicamente estados transitorios y absorbentes, se denomina cadena de Markov absorbente. También es posible calcular el tiempo esperado de recurrencia de un estado (o tiempo de primera pasada); esto es, el número de transiciones o pasos hasta que el proceso regresa al estado i de donde partió.

Por ejemplo un consumidor que tiene como preferencia la marca uno en una semana y cambia su preferencia a la marca dos en las semanas siguientes, es conveniente saber con que probabilidad regresará nuevamente a comprar la marca uno.

La experiencia basada en datos indica que un cliente que ha caído en $PV7$ es poco probable que salga de ese estado, continuamente seguirá en mala deuda; podemos pensar que se trata de un estado absorbente.

La matriz del ejemplo 3.1.4 muestra la probabilidades de transición de un paso $p_{00} = 0.899$ y $p_{44} = 0.740$ bastante cercanas a uno, lo que indica que es altamente probable que los que están al corriente sigan así y los que caen en perdida continúen como mala deuda.

Si i es un estado absorbente, y $k \neq i$, la probabilidad que del estado k se llegue alguna vez al estado i, se llama *probabilidad de absorción* del estado k (f_{ki}).

Si el estado i es recurrente e $i \neq k$ tal que $f_{kk} = 1$ y $f_{ik} = 0$. Las probabilidades f_{ij} se encuentran resolviendo el sistema de ecuaciones lineales

$$f_{ij} = \sum_{k=0}^{M} p_{ik} f_{kj}, \quad \text{para } i,\, j = 0,1,...,M$$

El sistema se escribe como

$$\begin{bmatrix} f_{0j} \\ f_{1j} \\ \vdots \\ f_{Mj} \end{bmatrix} = \begin{bmatrix} p_{0j} & p_{01} & \cdots & p_{0M} \\ p_{1j} & p_{11} & \cdots & p_{1M} \\ \vdots & & & \\ p_{Mj} & p_{M0} & \cdots & p_{MM} \end{bmatrix} \begin{bmatrix} f_{0j} \\ f_{1j} \\ \vdots \\ f_{Mj} \end{bmatrix}$$

Ejemplo 2.1.6. *Dependiendo de su situación los estados de cuenta de clientes con tarjetas de crédito en un banco se clasifican al corriente quienes pagaron su saldo (estado 0), quienes tienen de 1 a 30 días de retraso (estado 1), los que tienen de 31 a 60 días (estado 2) y quienes tienen 61 días o mas (estado 3). La siguiente matriz muestra las probabilidades de transición estimados con los datos de una institución financiera con dos estados absorbentes.*

Estados	*0*	*1*	*2*	*3*	*4*
0	*1*	*0*	*0*	*0*	*0*
1	*0.520*	*0.120*	*0.070*	*0.050*	*0.240*
2	*0.287*	*0.099*	*0.079*	*0.060*	*0.475*
3	*0.212*	*0.050*	*0.051*	*0.061*	*0.626*
4	*0*	*0*	*0*	*0*	*1*

De la matriz anterior se tiene que $f_{04} = 0$ y $f_{44} = 1$, probabilidades que se mantienen seguidamente.

Es de interés determinar la probabilidad con la que un cliente acaba por ser un mal pagador. Podemos obtener f_{14}, f_{24} y f_{34}. Las ecuaciones que deben resolverse son:

$$\begin{aligned} f_{14} &= p_{10}f_{04} + p_{11}f_{14} + p_{12}f_{24} + p_{13}f_{34} + p_{14}f_{44} \\ f_{24} &= p_{20}f_{04} + p_{21}f_{14} + p_{22}f_{24} + p_{23}f_{34} + p_{24}f_{44} \\ f_{34} &= p_{30}f_{04} + p_{31}f_{14} + p_{32}f_{24} + p_{33}f_{34} + p_{34}f_{44} \end{aligned}$$

sustituyendo los valores conocidos nos queda

$$\begin{aligned} f_{14} &= p_{11}f_{14} + p_{12}f_{24} + p_{13}f_{34} + p_{14} \\ f_{24} &= p_{21}f_{14} + p_{22}f_{24} + p_{23}f_{34} + p_{24} \\ f_{34} &= p_{31}f_{14} + p_{32}f_{24} + p_{33}f_{34} + p_{34} \end{aligned}$$

así que,

$$\begin{aligned} p_{14} &= (1 - p_{11})f_{14} - p_{12}f_{24} - p_{13}f_{34} \\ p_{24} &= -p_{21}f_{14} + (1 - p_{22})f_{24} - p_{23}f_{34} \\ p_{34} &= -p_{31}f_{14} - p_{32}f_{24} + (1 - p_{33})f_{34}. \end{aligned}$$

El sistema en forma de matriz nos queda

$$\begin{bmatrix} p_{14} \\ p_{24} \\ p_{34} \end{bmatrix} = \begin{bmatrix} 1 - p_{11} & -p_{12} & -p_{13} \\ -p_{21} & 1 - p_{22} & -p_{23} \\ -p_{31} & -p_{32} & 1 - p_{33} \end{bmatrix} \begin{bmatrix} f_{14} \\ f_{24} \\ f_{34} \end{bmatrix}$$

El sistema que debemos resolver tiene la forma

$$\begin{bmatrix} 0.240 \\ 0.475 \\ 0.626 \end{bmatrix} = \begin{bmatrix} 0.480 & -0.120 & -0.070 \\ -0.287 & 0.901 & -0.079 \\ -0.212 & -0.050 & 0.949 \end{bmatrix} \begin{bmatrix} f_{14} \\ f_{24} \\ f_{34} \end{bmatrix}$$

utilizando la matriz inversa obtenemos

$$\begin{aligned} f_{14} &= 0.338281 \\ f_{24} &= 0.702656 \\ f_{34} &= 0.772232 \end{aligned}$$

2.2. Regresión logística

Si se desea discriminar a los solicitantes de crédito en buenos y malos, la regresión logística es un buen método de clasificación que se utiliza comúnmente en el *credit scoring*. El modelo de regresión logística no requiere de los supuestos de la regresión lineal, como son el supuesto de normalidad de los errores de observación y el supuesto que las variables involucradas sean continuas. En este sentido, la regresión logística, se aplica tanto a datos que son gausaianos como a datos que no lo son, y por lo tanto, el modelo de regresión logística es útil cuando la variable de respuesta $\mathbf{x}$ no está distribuida normalmente y tanto las variables predictoras como de respuesta tienen valores discretos, categóricos, ordinales o no ordinales. La regresión lineal no es aplicable a este tipo de variables, dado que la variable respuesta y sólo presenta dos valores. La capacidad predictiva del modelo logístico se valora mediante la comparación entre el grupo de pertenencia observado y el pronosticado por el modelo. El modelo debe ser capaz de clasificar a los individuos en cada uno de los dos grupos: buenos o malos, basado en las variables que definen las características de los individuos.

Esta clasificación está a cargo de una distribución de probabilidad que separa a la población en dos grupos, la separación está basada en un punto de corte preestablecido en el rango de cero a uno. La probabilidad sirve para estimar el valor de y que dependiendo de su valor asignara al individuo a un grupo; por ejemplo se espera que para buenos clientes se obtengan valores muy cercanos a uno y para malos clientes valores cercanos a cero.

2.2.1. El modelo de regresión logística

En las variables explicativas, no se establece ninguna restricción, pudiendo ser cualitativas o cuantitativas tanto continuas como discretas y categóricas, con dos o mas valores. La variable respuesta y tendrá los valores cero y uno, podemos definir $y = 1$ si se trata de un buen cliente y $y = 0$ si se trata de un mal cliente. Con la regresión logística se modela la probabilidad de que y sea igual a uno,

dado los valores observados de las variables predictoras contenidas en el vector $\mathbf{x}$, $P(y = 1|\mathbf{x})$. Para discriminar a un individuo se estima su probabilidad con $\widehat{P}(y = 1|\mathbf{x})$ y si por ejemplo, $\widehat{P}(y = 1|\mathbf{x}) > 0.5$ se clasifica al cliente como bueno y si $\widehat{P}(y = 1|\mathbf{x}) < 0.5$ se clasifica como malo.

Para justificar el modelo de regresión logística, consideremos una muestra de n datos donde $\mathbf{x}_i$ es el vector de variables explicativas binarias de la forma

$$\mathbf{x}_i^T = [x_{i1}, x_{i2}, \ldots, x_{ip}], \quad i = 1, 2, ..., n$$

que tiene asociada una variable de respuesta binaria dependiente y_i que toma el valor de $y = 0$, si el cliente i es malo, y $y = 1$ si el cleinte i es bueno, [ver Peña (2002)]. Sea $P(y = 1|\mathbf{x}_i) = p_i$ la probabilidad de que $y = 1$, dado el vector $\mathbf{x}_i$ de datos observados. Se define una relación entre p_i y un modelo lineal mediante una función monótona y creciente g, llamada función *link*

$$g(p_i) = \beta_0 + \boldsymbol{\beta}_1^T \mathbf{x}_i,$$

tal que

$$\boldsymbol{\beta}_1^T = [\beta_1, \beta_2, \ldots, \beta_p],$$

vector de parámetros de coeficientes de las variables explicativas del modelo y β_0 la ordenada al origen. La función *link* que se aplica se conoce como la transformación logito y es el logaritmo del cociente de probabilidades de p_i y $(1 - p_i)$

$$g(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \boldsymbol{\beta}_1^T \mathbf{x}_i. \tag{2.8}$$

El modelo en términos de $g(p_i)$ puede escribirse como $g(p_i) = \beta_0 + \boldsymbol{\beta}_1^T \mathbf{x} + \varepsilon$, con ε variable aleatoria tal que $E(\varepsilon) = 0$ y $V(\varepsilon) = \sigma^2$. La función de distribución logística dada por la transformación inversa de g se escribe como

$$p_i = \frac{e^{\beta_0 + \boldsymbol{\beta}_1^T \mathbf{x}_i}}{1 + e^{\beta_0 + \boldsymbol{\beta}_1^T \mathbf{x}_i}} \tag{2.9}$$

que satisface $0 \leq p_i \leq 1$. Así

$$1 - p_i = \frac{1}{1 + e^{\beta_0 + \boldsymbol{\beta}_1^T \mathbf{x}_i}}. \tag{2.10}$$

Los coeficientes del modelo logístico sirven para calcular un parámetro de cuantificación de riesgo conocido como *odds ratio*. El *odds* asociado a un evento es el cociente entre la probabilidad de que ocurra con la probabilidad de que no ocurra.

$$odss = \frac{p_i}{1 - p_i}. \tag{2.11}$$

Un caso particular es cuando todas las variables explicativas pueden ser representadas de tal modo que todas sean binarias independientes. A cada variable se le asocia una probabilidad según sea la población a la que pertenecen. Sean $P(\mathbf{x}_i|y = 1)$ y $P\mathbf{x}_i|y = 0)$ las probabilidades del vector $\mathbf{x}_i$ dado que el individuo pertenece a la población uno y dos respectivamente. Bajo el supuesto que las coordenadas de $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ son binarias e independientes, se tiene que

$$P(\mathbf{x}_i|y = 1) = P[(x_{i1}, x_{i2}, \ldots, x_{ip})|y = 1] =$$

$$= \prod_{j=1}^{p} P(x_{ij}|y = 1) = \prod_{j=1}^{p} p_{1j}^{x_{ij}}(1 - p_{1j})^{1-x_{ij}}$$

y

$$P(\mathbf{x}_i|y = 0) = \prod_{j=1}^{p} p_{2j}^{x_{ij}}(1 - p_{2j})^{1-x_{ij}}. \tag{2.12}$$

Suponiendo que las probabilidades a priori son las mismas, es decir

$$P(y = 0) = P(y = 1)$$

y considerando que nuestro modelo es construido sobre la misma cantidad de datos para las dos poblaciones. Se tiene que la probabilidad condicionada esta dada por

$$P(y = 1|\mathbf{x}_i) = \frac{P(\mathbf{x}_i|y = 1)P(y = 1)}{P(\mathbf{x}_i)} = \frac{P(y = 1)}{P(\mathbf{x}_i)} \prod_{j=1}^{p} p_{1j}^{x_{ij}}(1 - p_{1j})^{1-x_{ij}}$$

y para $1 - P(y = 1|\mathbf{x}_i) = P(y = 0|\mathbf{x}_i)$

$$P(y = 0|\mathbf{x}_i) = \frac{P(\mathbf{x}_i|y = 0)P(y = 0)}{P(\mathbf{x}_i)} = \frac{P(y = 0)}{P(\mathbf{x}_i)} \prod_{j=1}^{p} p_{2j}^{x_{ij}}(1 - p_{2j})^{1-x_{ij}}. \tag{2.13}$$

Calculamos la distribución logística

$$\begin{aligned}\frac{P(y=1|\mathbf{x}_i)}{1-P(y=1|\mathbf{x}_i)} &= \frac{\frac{P(y=1)}{P(\mathbf{x}_i)}\prod_{j=1}^{p} p_{1j}^{x_{ij}}(1-p_{1j})^{1-x_{ij}}}{\frac{P(y=0)}{P(\mathbf{x}_i)}\prod_{j=1}^{p} p_{2j}^{x_{ij}}(1-p_{2j})^{1-x_{ij}}} \\ &= \prod_{j=1}^{p}\left(\frac{p_{1j}}{p_{2j}}\right)^{x_{ij}}\left(\frac{1-p_{1j}}{1-p_{2j}}\right)^{1-x_{ij}}\end{aligned}$$

calculamos la transformación función logito,

$$\begin{aligned}g_i(\mathbf{x}_i) &= \log\frac{P(y=1|\mathbf{x}_i)}{1-P(y=1|\mathbf{x}_i)} \\ &= \sum_{j=1}^{p} x_{ij}\log\left(\frac{p_{1j}}{p_{2j}}\right)+\sum_{j=1}^{p}(1-x_{ij})\log\left(\frac{1-p_{1j}}{1-p_{2j}}\right) \\ &= \sum_{j=1}^{p}\left[\log\left(\frac{p_{1j}}{p_{2j}}\right)-\log\left(\frac{1-p_{1j}}{1-p_{2j}}\right)\right]x_{ij}+\sum_{j=1}^{p}\log\left(\frac{1-p_{1j}}{1-p_{2j}}\right)\end{aligned}$$

$$g_i(\mathbf{x}_i) = \sum_{j=1}^{p}\log\left[\frac{p_{1j}(1-p_{1j})}{p_{2j}(1-p_{2j})}\right]x_{ij}+\sum_{j=1}^{p}\log\left(\frac{1-p_{1j}}{1-p_{2j}}\right). \tag{2.14}$$

Se observa que $g_i(\mathbf{x}_i)$ es una función lineal que coincide con la ecuación (2.8), donde

$$\beta_0 = \sum_{j=1}^{p}\log\left(\frac{1-p_{1j}}{1-p_{2j}}\right)$$

y

$$\boldsymbol{\beta}_1^T = \left[\log\frac{p_{11}(1-p_{11})}{p_{21}(1-p_{21})},\ldots,\log\frac{p_{1p}(1-p_{1p})}{p_{2p}(1-p_{2p})}\right]$$

para el modelo $g_i(\mathbf{x}_i)=\beta_0+\beta_1^T\mathbf{x}_i$. El parámetro β_0 nos da la ordenada al origen, y $\beta_1=(\beta_1,\ldots,\beta_p)^T$ es el vector de pendientes. Así

$$\log\left(\frac{p_i}{1-p_i}\right)=\beta_0+\beta_1x_{i1}+\ldots+\beta_px_{ip} \tag{2.15}$$

y β_j con $j=0,1,\ldots,p$, nos sirve para analizar la cantidad de cambio del *ratio* de probabilidades cuando se incrementa una variable de predicción x_j en una unidad

$$OR=\exp(\beta_j).$$

El modelo de regresión $g(\mathbf{x}_i) = \beta_0 + \boldsymbol{\beta}_1^T \mathbf{x}_i$ es una transformación lineal en los parámetros del modelo por lo que podemos utilizar algunas técnicas aplicadas en el análisis de regresión lineal, como por ejemplo la selección de variables mediante el proceso de *backward* y *forward*.

2.2.2. Estimación del modelo logit usando MV

Para estimar los parámetros del modelo logístico se utiliza el método de máxima verosimilitud (MV). Como y_i toma dos valores, 0 con probabilidad p_i y 1 con probabilidad $1 - p_i$, tiene como distribución de probabilidad una Bernoulli.

$$P(y_i) = p_i^{y_i}(1 - p_i)^{(1-y_i)} \quad y_i = 0, 1.$$

La función MV para una muestra aleatoria de n datos (x_i, y_i) se calcula como

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

aplicando logaritmos

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

y la función log verosimilitud se escribe como

$$\log P(\mathbf{y}) = \sum_{i=1}^{n} y_i \log \left(\frac{p_i}{1 - p_i} \right) + \sum_{i=1}^{n} \log(1 - p_i). \tag{2.16}$$

Consideremos $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_p)$ y $\mathbf{x}_i^T = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$ para escribir el modelo de la forma

$$\log \left(\frac{p_i}{1 - p_i} \right) = \mathbf{x}_i^T \boldsymbol{\beta}. \tag{2.17}$$

Ahora la ecuación (2.17) la sustituimos en la ecuación (2.16). De aquí, obtenemos la función de verosimilitud en logaritmos en términos de los parámetros $\boldsymbol{\beta}$ dada por

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \mathbf{x}_i^T \boldsymbol{\beta} - \sum_{i=1}^{n} \log \left(1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}} \right). \tag{2.18}$$

Para obtener los estimadores $\boldsymbol{\beta}$ de máxima verosimilitud derivamos $L(\boldsymbol{\beta})$ con respecto de cada uno de los parámetros β_j con j= 1, 2, ..., p e igualamos a cero. En términos de matrices

$$\begin{bmatrix} \frac{\partial L(\beta)}{\partial \beta_0} \\ \frac{\partial L(\beta)}{\partial \beta_1} \\ \vdots \\ \frac{\partial L(\beta)}{\partial \beta_j} \\ \vdots \\ \frac{\partial L(\beta)}{\partial \beta_p} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i(1) \\ \sum_{i=1}^n y_i x_{i1} \\ \vdots \\ \sum_{i=1}^n y_i x_{ij} \\ \vdots \\ \sum_{i=1}^n y_i x_{ip} \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^n (1) \left(\frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right) \\ \sum_{i=1}^n x_{i1} \left(\frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right) \\ \vdots \\ \sum_{i=1}^n x_{ij} \left(\frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right) \\ \vdots \\ \sum_{i=1}^n x_{ip} \left(\frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right) \end{bmatrix} \tag{2.19}$$

cada una de estas derivadas se experesan en un vector columna de la forma

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n y_i \mathbf{x}_i - \sum_{i=1}^n \mathbf{x}_i \left(\frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right), \tag{2.20}$$

igualando (2.19) al vector cero

$$\sum_{i=1}^n y_i \mathbf{x}_i = \sum_{i=1}^n \mathbf{x}_i \left(\frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right) = \sum_{i=1}^n \mathbf{x}_i p_i. \tag{2.21}$$

Si $\hat{\boldsymbol{\beta}}$ es el vector de parámetros que cumple el sistema (2.19), calculamos p_i en términos de esos estimadores y de aquí se obtiene una estimación para y_i, tal que $\hat{y}_i = \hat{p}_i$, así

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

de aquí

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

donde e_i representa los residuos del modelo y deben ser ortogonales al espacio de observaciones $\mathbf{x}$, esto es similar que en la regresión estándar (mínimos cuadrados).

Observamos que el sistema de ecuaciones (2.19) no es lineal en los parámetros $\boldsymbol{\beta}$ y para obtener los estimadores MV es común que se utilice el método de *Newton-Raphson*

2.3. Pruebas estadísticas al modelo logito

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 a continuación veremos.

2.3.1. Deviance

La función $D(\beta) = -2L(\beta)$ se le conoce como la *desviación* o *deviance* [ver Thomas (1997)]

$$D(\beta) = 2\sum_{i=1}^{n}\left[\log\left(1+e^{\mathbf{x}_i^T\beta}\right) - y_i\mathbf{x}_i^T\beta\right], \tag{2.22}$$

y en término de las probabilidades

$$D(\beta) = -2\sum_{i=1}^{n}\left[y_i\log p_i + (1-y_i)\log(1-p_i)\right] \tag{2.23}$$

nos dan una medida de la desviación máxima del modelo.

Ejemplo 2.3.1. *Para el modelo más simple que solo contiene a β_0 estimamos su desviación a partir de n datos, donde m es el número de elementos en la muestra tales que $y = 1$ y $n-m$ con $y = 0$. Supongamos que la probabilidad p_i es igual para todas las observaciones con $y = 1$ y $1 - p_i$ para $y = 0$, esto es, estimamos $\hat{p}_i = p$.*

Sustituyendo en la expresión de la ecuación (2.23).

$$\begin{aligned} D(\beta_0) &= -2\sum_{i=1}^{n}\left[y_i\log p + (1-y_i)\log(1-p)\right] \\ &= -2\left[\log p\sum_{i=1}^{n}y_i + \log(1-p)\sum_{i=1}^{n}(1-y_i)\right] \\ &= -2\left[\log p\cdot(m) + \log(1-p)\cdot(n-m)\right] \end{aligned}$$

entonces

$$D(\beta_0) = -2m\log p - 2(n-m)\log(1-p). \tag{2.24}$$

Así $D_0 = D(\beta_0)$ se considera como el valor inicial de la desviación del modelo.

2.3.2. Estadístico de Wald

Para determinar si una variable debe ser incluida en un modelo porque tiene un peso significativo se aplica la prueba de *estadístico de Wald*. La prueba resulta de contrastar la hipótesis nula

$$H_0 : \beta_i = 0$$

contra la alternativa

$$H_1 : \beta_i \neq 0$$

con un estadístico de prueba definido como

$$w_j = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)},$$

que bajo el supuesto que H_0 es cierto sigue una distribución t con $n-p-1$ grados de libertad y para muestras grandes se distribuye como una normal estándar. Se entiende que si w_i es un valor alejado de cero se tendrá evidencia que H_0 es falsa, por lo tanto la región crítica de la prueba es de la forma $|w_j| > t_\alpha/2$, para un nivel de significación adecuado. Entendemos que si el verdadero valor del parámetro β_i es cero la variable x_i debe excluirse. Otra manera equivalente de escribir la región crítica es usando el $p-value$ donde $p = P(t > |w_j|)$, el $p-value$ es reportado por la mayoría de los paquetes estadísticos. La región crítica es de la forma $p < \alpha$, α nivel de significancia adecuado.

Comparando modelos

Suponga que se tiene k variables explicativas, x_1, x_2, x_3,...,x_k y se desea saber si ellas son significativas o no sin pérdida de generalidad se puede suponer que las variables a prueba son las últimas, x_{k-s+1}, x_{k-s+2},..., x_k, $s < k < n$. de esta manera se esta confrontando dos modelos. El primer modelo que incluye todas las variables

$$w_1 = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-s} x + \ldots + \beta_{k-1} x_{k-1} + \beta_k x_k$$

el segundo modelo incluye solo las $k-s$ primeras variables

$$w_2 = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_{k-s}.$$

Se realiza la prueba para contrastar la hipótesis nula H_0, de que las variables x_i con $i = k-s+1, \ldots, k-1, k$ no influyen significativamente en el modelo contra

la alternativa H_1 que dice que si influyen; esto es

$$\begin{aligned} H_0: &\quad \beta_{k-s+1} = 0 \ y \ \beta_{k-s+2} = 0 \ y \ \ldots \ y \ \beta_k = 0 \\ H_1: &\quad \beta_{k_s+1} \neq 0 \text{ ó } \beta_{k-s+2} \neq 0 \text{ ó } \ldots \text{ ó } \beta_k \neq 0 \end{aligned}$$

Para encontrar evidencia de que H_0 es falsa se usa la región crítica que surge del cociente de verosimilitud

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

de esta relación se obtiene él estadístico

$$\chi^2_s = 2L(H_1) - 2L(H_2)$$

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

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

Si H_0 es cierta el estadístico sigue una distribución χ^2_s con s grados de libertad, para un α dada la región crítica es $\chi^2_s > \chi^2_\alpha$. Esto nos da una medida de mejora entre un modelo y el otro. Nótese que cuando $s = 1$, únicamente se ésta probando un coeficiente del modelo de regresión y entonces se estaría en el mismo caso que la prueba de Wald, por lo que, la prueba de razón de verosimilitud, en este caso, es alternativa a la prueba de Wald.

2.3.3. Estadístico R^2

El estadístico R^2 sirve para estimar de manera global la influencia de todas las variables en el modelo Un caso particular es verificar el modelo que no incluye variables de predicción y contiene únicamente β_0 contra el modelo que las incluye. Esto es, con el cálculo de

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

Los valores extremos para R^2 son cero y uno. El valor $R^2 = 1$ se obtiene cuando el modelo tiene un ajuste perfecto esto es, si las observaciones $y = 1$ tienen probabilidad $p_i = 1$ y $y = 0$ probabilidad $p_i = 0$, las variables explican completamente el comportamiento de y. El valor $R^2 = 0$ se obtiene cuando las variables no influyen en el modelo, no pueden predecir los valores de y, porque la desviación esperada

para el modelo que inluye todas las variables es igual a la desviación del modelo que no las incluye, o sea que $D(\hat{\beta}) = D(\hat{\beta}_0)$.

Las pruebas con estimadores de máxima verosimilitud para un modelo nos da una medida de cuán compatible es éste con los datos realmente observados. Si al añadir o quitar una variable al modelo no mejora la verosimilitud o no disminuye la *desviación* de forma apreciable, en sentido estadístico, ésta variable no se incluye en la ecuación.

2.3.4. Residuos de Pearson

Para hacer un contraste global del modelo logit podemos utilizar los *resi-duos de Pearson*. Las pruebas de hipótesis que se contrastan son:

H_0 : El modelo es adecuado

H_1 : El modelo no es adecuado

Los residuos del modelo logit están definidos como

$$e_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}, \qquad 0 < \hat{p}_i < 1.$$

Si el modelo es adecuado e_i tiene media cero y varianza uno, de aquí se construye el estadístico de prueba que se distribuye asintóticamente como:

$$\chi^2_c = \sum_{i=1}^{n} e_i^2$$

que sigue una distribución χ^2 con $n - (p+1)$ grados de libertad con p variables.

2.3.5. Criterios para elegir el mejor modelo

Un buen modelo debe satisfacer dos condiciones, la primera es que tenga una fuerte capacidad predictora y la segunda es que la estimación de los parámetros tenga una alta precisión. Una condición adicional es que el modelo sea lo más sencillo posible, esto es que contenga el mínimo de variables explicativas y que satisfaga las dos condiciones anteriores. En este sentido se pregunta uno sí todas las variables explicativas son necesarias para construir el modelo, o si alguna de

ellas puede ser excluida. Para determinar que variables podrán ser excluidas se realiza una prueba de hipótesis.

Entre otros se pueden usar dos métodos automatizados de selección de variables: el *forward* que consiste en ir incluyendo variable cada vez y probar en cada paso si hay mejoras en el modelo o el *backward* que consiste en iniciar con todas las variables y se va excluyendo una variable en cada paso; de igual manera se va probando si no existe desmejora en el modelo. Estos dos métodos automatizados tienen la desventaja de que una variable que es incluida (excluida) ya no vuelve a salir (entrar) del modelo, esto podría restringir las opciones de seleccionar un mejor modelo, en este sentido, una alternativa es permitir incluir o excluir, variables que ya entraron o salieron, usando criterios estadísticos como el de correlaciones parciales. En todo caso, se pueden obtener todos los posibles modelos y escoger entre ellos, el mejor, aunque este método exaustivo consume más tiempo.

2.3.6. Validación del método de clasificación

Para validar la eficacia del método de clasificación, se utilizan datos de clientes de la misma población los cuales se conoce a que población pertenece, pero diferentes a los utilizados para estimar el modelo. Se clasifican a este conjunto de clientes mediante el modelo estimado y luego se cuentan cuantos de ellos quedan bien clasificados y cuantos quedan mal clasificados. Bajo el supuesto que el método de clasificación es adecuado, se esperaría que todos los clientes fueran bien clasificados. En este sentido se tiene los siguientes términos:

O_{11} = número de clientes buenos clasificados como buenos.
O_{12} = numero de clientes buenos clasificados como malos.
O_{21} = número de clientes malos clasificados como buenos.
O_{22} = número de clientes malos clasificados como malos.

	Clasificados	
Realidad	*Buenos*	*Malos*
Buenos	O_{11}	O_{12}
Malos	O_{21}	O_{22}

Cuadro 2.2: Tabla de éxitos.

Mientras que se considera como valores esperados los siguientes:

E_{11} = número esperado de clientes buenos clasificados como buenos, y es igual al número de clientes buenos en la muestra de validación.

E_{12} = número esperado de clientes buenos clasificados como malos, es igual a cero.

E_{21} = número de clientes malos clasificados como buenos, también es igual a 0, y

E_{22} = número esperado de clientes malos clasificados como malos, es igual al número de clientes malos en la muestra de validación.

Usando la estadística

$$\chi_c^2 = \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

se puede validar el metodo de clasificación. Mientras más grande es el valor de esta estadística, es menor la capacidad clasificatoria del método.

Algunas consideraciones

El proceso de seleccionar el conjunto de variables explicativas significantes es muy importante, pues el incluir variables poco significativas o con información redundante (colinealidad) puede distorsionar la capacidad predictiva de las funciones discriminantes estimadas. Por otro lado, no incluir variables explicativas altamente significativas da una pobre estimación de los parámetros.

2.4. Prueba de diferencias de dos poblaciones

Una vez que se ha calculado el *score* con la fórmula estimada por las matrices de transición o la regresión logística, se pretende determinar si estos valores calculados en la muestra identifican bien a que grupo pertenecen. Mientras mayor sea la diferencia de los puntajes *score* de los grupos mayor será la capacidad discriminante del modelo usado. Entre las técnicas para determinar la diferencia entre los puntajes del *score* en los grupos de buenos y malos clientes están: el índice de Gini, la divergencia y la prueba de Kolmogorov Smirnov. En ésta sección trataremos el índice de Gini.

2.4.1. Índice de Gini

En 1960 se propuso medir la desigualdad en la salud a partir de la curva de Lorenz, el índice de Gini se deriva de esta. El índice de Gini es uno de los más

utilizados para medir la desigualdad entre dos poblaciones [1]. En el caso que nos ocupa se utiliza para medir la desigualdad de las poblaciones de buenos y malos clientes. Teóricamente la curva de Lorenz de las funciones de distribución $F(x)$ y $G(x)$ es el subconjunto del producto cartesiano dado por

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

Definimos a F y G como la funciones de distribución teóricas asociadas a los clientes malos y buenos respectivamente, donde x es el puntaje de *score*. Si el puntaje de *score* para buenos es mayor que el puntaje *score* para malos, la curva de Lorenz de F y G es cóncava hacia arriba como en la figura 2.2. Se ve que sí $F(x) = G(x)$ entonces $\mathcal{L}(F,G)$ describe la recta $u = v$ con $u \in (0,1)$ entre las distribuciones F y G.

Por lo tanto mientras $\mathcal{L}$ se separe más de la recta $v = u$, mayor será la diferencia entre $F(x)$ y $G(x)$. Por esta razón, el área A que se encuentra entre la identidad y la curva de Lorenz es una medida de desigualdad entre las distribuciones F y G.

El índice de Gini resulta de la razón entre el área A y el área del triangulo delimitado por la identidad, el eje horizontal u y la recta $u = 1$

Índice de Gini con observaciones agrupadas

Cuando se desconocen las funciones de distribución $F(x)$ y $G(x)$, pero se cuenta con una muestra aleatoria de cada una de esta dos distribuciones empíricas de tamaño n_1 y n_2 respectivamente se puede estimar la curva de Lorenz y por lo tanto el índice de Gini. Para hacer esto primero se define una partición de $\mathbb{R}$ dada por $x_0 \leq x_1 \leq x_2 \leq \ldots \leq x_k$, luego se obtiene los estimadores de F y G en los puntos x_i de la siguiente manera

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

y

$$\hat{G}(x_i) = \frac{\#\text{ de elementos en la muestra } 2 \leq x_i}{n_2}.$$

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 $(\hat{F}(x_{i-1}), \hat{G}(x_{i-1}))$ y $(\hat{F}(x_i), \hat{G}(x_i))$. El

[1] http://es.wikipedia.org/wiki/Coeficiente_de_Gini
http://www.asinorum.com/el-indice-de-gini/309/

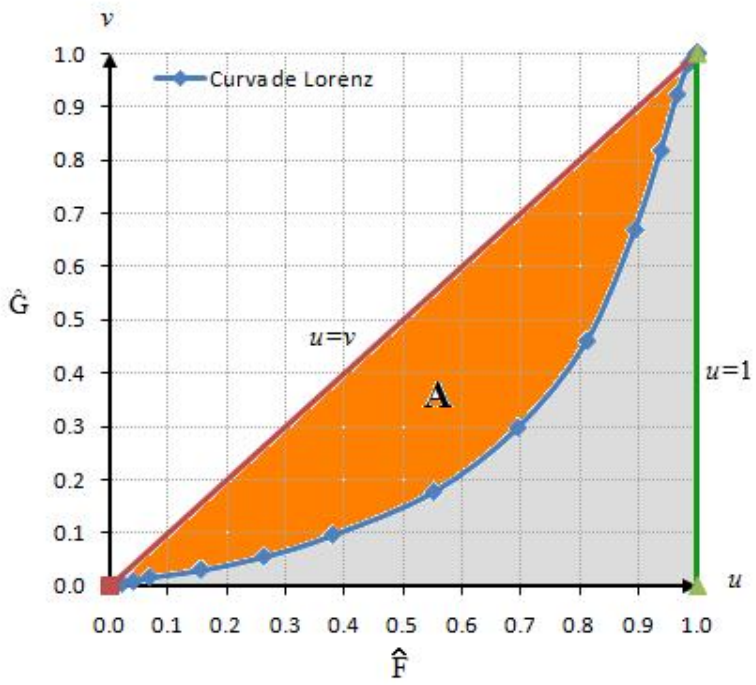


Figura 2.1: El Índice de Gini y la curva de lorenz.

área por debajo de la curva de Lorenz estimada para un intervalo tiene la forma de un trapecio y la calculamos como

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

El área total por debajo de la curva de Lorenz estimada es $(\hat{F}(x_i), \hat{G}(x_i))$

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

El índice de Gini (1914) estimado se calcula como [ver Medina (2001)]

$$Gini = \frac{1/2 - A}{1/2}$$

Ejemplo 2.4.1. *Sea x la variable que mide el puntaje de* score *de los clientes en una institución de crédito para los cuales se le asocian la función de distribución $F(x)$ y $G(x)$ para el número de clientes malos y buenos respectivamente según el puntaje de* score*. Deseamos calcular el índice de Gini para comparar el porcentaje de cuentas buenas y malas para los mismos puntajes de* score*.*

x_i	Malos	Buenos	$\hat{F}_i$	$\hat{G}_i$	$\hat{F}_i - \hat{G}_i$	A_i
410	45	0	0.019938	0.000000	0.019938	0.000000
430	87	5	0.058485	0.000117	0.058368	0.000002
450	134	45	0.117856	0.001172	0.116684	0.000038
470	252	76	0.229508	0.002952	0.226556	0.000230
490	320	132	0.371289	0.006046	0.365244	0.000638
510	389	289	0.543642	0.012818	0.530824	0.001626
530	321	789	0.685866	0.031306	0.654561	0.003138
550	246	1873	0.794860	0.075194	0.719666	0.005804
570	176	2543	0.872840	0.134783	0.738057	0.008187
590	113	3765	0.922907	0.223006	0.699901	0.008957
610	76	5469	0.956580	0.351158	0.605422	0.009667
630	54	7654	0.980505	0.530509	0.449996	0.010547
650	21	8844	0.989809	0.737745	0.252065	0.005900
670	14	5639	0.996012	0.869880	0.126132	0.004986
690	5	3193	0.998228	0.944700	0.053528	0.002010
710	3	1786	0.999557	0.986550	0.013007	0.001284
730	1	345	1.000000	0.994634	0.005366	0.000439
750	0	189	1.000000	0.999063	0.000937	0.000000
770	0	32	1.000000	0.999813	0.000187	0.000000
790	0	3	1.000000	0.999883	0.000117	0.000000
810	0	5	1.000000	1.000000	0.000000	0.000000
Sumas	42676	2257				**0.063452**

Cuadro 2.3: Frecuencias relativas acumuladas de buenos y malos

En la tabla 2.3 se muestran los intervalos en orden creciente para la variable *score* para buenos y malos clientes. La cuarta y quinta columna registran los porcentajes acumulados correspondientes a las distribuciones para el número de buenos y malos clientes en cada intervalo. La última columna contiene el área que corre-

sponde a cada intervalo. Así que

$$A = \sum_{i=2}^{k} A_i = 0.06345.$$

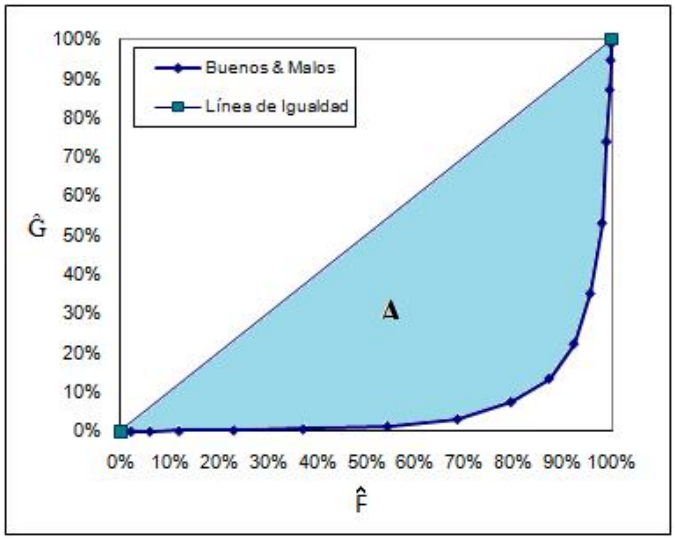


Figura 2.2: La identidad y la curva de Lorenz para obtener el Índice de Gini. El área entre estas dos gráficas es cercana a 0.5, se obtiene un Gini mas cercano a uno.

El índice de Gini es

$$Gini = \frac{0.5 - A}{0.5} = 0.8731.$$

Obtenemos 87.3 % para el índice de Gini. La curva de Lorenz la graficamos en términos de procentajes como se ve en la figura 2.2.

2.4.2. Divergencia

La divergencia mide la diferencia entre las medias de dos distribuciones estandarizadas usando las varianzas y tiene la siguiente expresión

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

Cuando construimos un modelo logístico que clasifica dos poblaciones, se espera que los dos grupos estén estadísticamente bien separados; esto es, la diferencia entre

sus medias sea importante. Entre más pequeña la divergencia nos estará diciendo que la distribución de cada población es parecida y no sabremos diferenciar un grupo del otro, es decir para un mismo puntaje de *score* tendremos cantidades similares de buenos y malos. La divergencia debe ser mayor de 0.95 para tener poblaciones estadísticamente separadas [ver Simbaqueba (2004)].

Ejemplo 2.4.2. *Queremos analizar las distribuciones de buenos y malos para un puntaje de* score*. Calculamos la divergencia de las medias de las distribuciones de buenos y malos según el puntaje obtenido para las muestras de las poblaciones que se muestran en la tabla 2.3*

Calculamos la media y desviación estándar para los datos. Las medias encontradas son $\overline{x}_b$ =631.97 y $\overline{x}_m$=519.24 para buenos y malos respectivamente con varianzas de s_b^2=2107.85 y s_m^2=2812.24.

Muestra	Tamaño de la muestra	Media aritmética	Desviación estándar
Buenos	42676	631.97	45.91
Malos	2257	519.24	53.03

Así que el valor de la divergencia es 5.17, el cual es mucho mayor a 0.95, esto índica que hay una separación importante entre las poblaciones de buenos y malos; esto es claramente visible en la grafica de la figura 2.3

2.4.3. Test de Kolmogorov-Smirnov

Una de las pruebas no paramétricas para la bondad de ajuste es el test de Kolmogorov-Smirnov. Si se desea probar que dos muestras independientes provienen de la misma distribución utilizamos la prueba de Kolmogorov-Smirnov tambien conocida como la prueba *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.

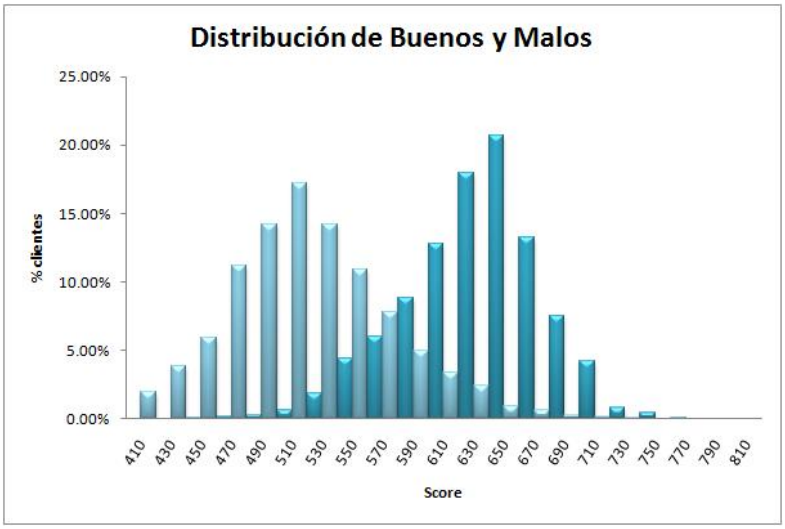


Figura 2.3: Distribución de buenos y malos

Para esta prueba se requiere tener dos muestras de una variable aleatoria continua, o al menos 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. Calculamos entonces las diferencias de las frecuencias relativas acumuladas. El estadístico esta dado como la máxima diferencia de las distribuciones de frecuencias relativas acumuladas

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

Se selecciona aquel intervalo de clase que tenga mayor desviación absoluta D. El valor crítico es calculado como

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

donde n_1 y n_2 son los tamaños de las muestras y K es valor obtenido de tabla de Kolmogorov-Smirnov con $n_1 + n_2 - 2$ grados de libertad a un nivel de significancia dado. Si la la desviación observada es menor que la desviación crítica tabulada 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 $D_{max} > D_{critico}$, la distribución no es la misma para cada población, la prueba marca que hay discriminación entre la dos poblaciones.

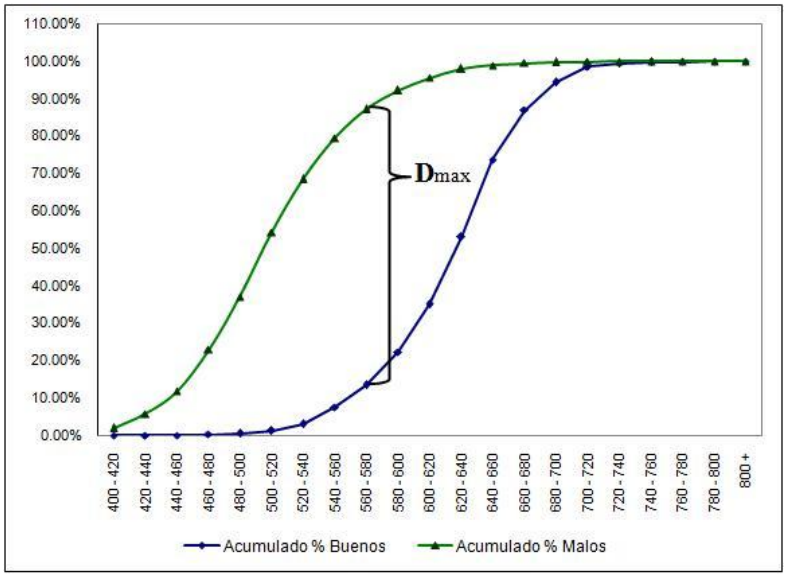


Figura 2.4: Distribucion acumulada de buenos y malos

Ejemplo 2.4.3. *Con los datos del cuadro 2.3 realizamos la prueba K-S para determinar si la distribución del* score *de las cuentas buenas y las cuentas malas es la misma.*

La prueba la construimos sobre las hipótesis:

H_0 = La distribución del score para las cuentas buenas y cuentas malas son iguales.

H_1 = La distribución del score para las cuentas buenas y cuentas malas no son iguales.

Los datos de la tabla 2.3 dan una diferencia máxima igual a 0.7381. Por otro lado, obtenemos un valor $D_{critico} = 0.0138$ con un nivel de significancia de $\alpha = 0.05$ lo que significa que $D_{max} > D_{critico}$ por lo que se rechaza la hipótesis nula, existe evidencia suficiente para rechazar que las dos distribuciones de buenos y malos son iguales. En términos de porcentajes obtenemos un *K-S* de 87.3 y lo alcanza a los 570 puntos de *score* como se observa en la figura 2.4.

Capítulo 3

Credit Scoring

En este capítulo presentaremos algunas técnicas que se utilizan en el *credit scoring*. Algunas ideas ya fueron mencionadas en los capítulos anteriores. La terminología *scoring* o *credit scoring* la manejaremos indistintamente. El método aquí presentado es uno de los existentes, pues existen varias formas de obtener el *credit scoring*, dependiendo del país y leyes que rigen. Algunos bancos, por ejemplo, construyen sus propios *scorecards*, otros solo las adaptan de las hechas o construidas por sus matrices fuera del país.

3.1. ¿Qué es el credit scoring?

El *credit scoring* es una exitosa colección de técnicas estadísticas que se han utilizado para otorgar créditos en la industria del crédito [ver Simbaqueba (2004)]. Ha sido utilizado por más de 40 años permitiendo el crecimiento de consumidores de crédito, crecimiento que ha sido propiciado por el uso de la computadora lo que permitió el avance de la estadística por el manejo de grandes cantidades de datos. Existen diversas compañías a nivel internacional que hacen uso del *credit scoring* para ofrecer servicios de análisis de riesgo en el crédito. Es una técnica que es bastante utilizada y además rentable, dado que una pequeña mejora en el desempeño puede significar un incremento en las ganancias de los prestamistas debido al volumen de préstamos realizados usando *scoring*.

3.1.1. Tipos de score

La colección de técnicas que conforman el *scoring* tiene como propósito principal generar un puntaje de riesgo a las solicitudes de crédito o a cuentas ya existentes.

Como ya vimos en el ciclo de riesgo (figura 1.2) existen tres etapas en el proceso de otorgar un crédito y dependiendo en qué parte del ciclo estemos trabajando se calcula uno de los siguientes puntajes de *score*:

- *Aquisition Score* o *Score de originación*. En el departamento de Originación se utiliza éste puntaje para la aceptación o rechazo de las solicitudes de crédito. Los tipos de variables utilizadas son demográficas y de buró de crédito. Este puntaje estima la probabilidad de incumplimiento de pago de un posible cliente y de esta manera se decide si se acepta o rechaza como posible consumidor de crédito, es decir da una estimación de su puntualidad de pago en el futuro, de tal manera que se optimice la tasa de aprobación de las solicitudes. Permite a la organización decidir el puntaje mínimo óptimo de aceptación en conjunto con otros departamentos. También permite definir productos de crédito personalizados y realizar actividades de mercadeo para aumentar el número de clientes con características deseables y cumplir con las metas corporativas. Con esta técnica finalmente se hace clasificación de los individuos en buenos y malos clientes.

- *Behavior score* o *Score de comportamiento*. Es utilizado en la etapa de administración del ciclo de riesgo. Predice la probabilidad de incumplimiento de los clientes que ya son objeto de crédito en la institución. Se utilizan las variables de comportamiento de las cuentas dentro de la propia institución. Permite dar seguimiento al comportamiento de los clientes lo que permitirá al departamento de cobranzas emplear técnicas para que un cliente siga siendo rentable para la empresa.

- *Colection score*. Puntaje que se calcula en la parte de recuperación de cuentas para estimar la probabilidad de recuperar a un cliente. Las variables utilizadas resultan de la combinación de variables de comportamiento y buró de crédito. Es posible determinar el valor preciso del libro de deudas antes de hacer el traspaso a una empresa recaudadora.

En éste proyecto se consideró solamente el *score* de originación (*Aquisition Score*) aunque es importante mencionar que las técnicas usadas pueden ser emuladas en los demás tipos de *score*. Esta idea también puede ser utilizada en otros medios donde se necesite este tipo de clasificación binaria.

3.1.2. Tipos de modelos

Modelo experto

Es cuando el modelo se construye con información de otras instituciones; esto es, está listo para usar. Se trata de modelos de crédito genéricos que se compran a consultores externos, que son adaptaciones de modelos hechos en otras matrices, etc. Esto es común en instituciones que apenas se inician como prestamistas dado que no tiene historial de sus clientes. Estos modelos con similitudes de población se utilizan en primera instancia aunque no es lo más conveniente.

Modelo estadístico

Son modelos que se construyen con información propia, son conocidos como modelos *in-house*. Tiene como beneficio que se pueden construir modelos específicos para distintos segmentos de la población. Se maneja información propia flexible y manejable. Se adquiere conocimiento y experiencia propia sobre la población, y también habilidad en el diseño e interpretación de los resultados. Se conserva la confidencialidad de la información.

3.2. La scorecard

Una *scorecard* es una tabla que contiene los puntajes asignados a cada atributo de cada una de las variables usadas [ver Thomas *et al* (2002)]. El puntaje determina, por ejemplo, la probabilidad de pago de la deuda para un cliente cuando se le otorgue una tarjeta de crédito. Así que a mayores puntajes corresponden a una mayor probabilidad de pago. La compañía prestamista es la que define finalmente la probabilidad mínima de pago para determinar cuando un cliente es considerado bueno; esto es, el puntaje de separación entre clientes buenos y malos; esto es, a partir de que puntaje un cliente se hará acreedor de un crédito. Este puntaje de corte llamado *cut off* que veremos más adelante es indicado en principio por los analistas de *credit score* pero se verá influido por las decisiones gerenciales o se basará en las metas corporativas de la propia institución.

Ejemplo 3.2.1. *Consideremos una* scorecard *simple con cinco variables o atributos: edad, estado civil, antigüedad en el empleo, sexo y nivel de estudios como se muestra en la tabla que sigue.*

Un individuo con 37 años de edad, soltero, con 5 años de antigüedad en su empleo, masculino y profesionista tendrá un puntaje (score) en base a esta tabla

Características	**Atributos**	**Score**
Edad	Menor a 24 años	-40
	4 - 30 años	-28
	31 - 40 años	10
	Mayor a 40 años	30
Estado Civil	Casado	12
	Soltero(a)	0
	Otro	-60
Antigüedad Empleo	0 - 1 años	-5
	2 - 5 años	4
	6 - 10 años	10
	Mayor a 10 años	15
Sexo	Masculino	-10
	Femenino	8
Nivel Estudios	Superior	-15
	Medio	3
	Básica	20
	Profesionista	37

Cuadro 3.1: El puntaje *score* en una *scorecard*

de $41 = 10 + 0 + 4 - 10 + 37$. Este puntaje es sólo una parte del total que se utiliza para tomar la decisión. A esta cantidad se le suma otra cantidad obtenida mediante un modelo estadístico. Obsérvese que un individuo con mayor antigüedad en su empleo no necesariamente tiene mayor probabilidad de obtener el préstamo, un individuo obtiene su puntaje en base a los valores de todas sus variables explicativas. Obsérvese que la score card no dice que a mayor atigüedad mayor puntaje. Un individuo con menos antigüedad puede tener un mayor puntaje; por ejemplo, considere el caso de un cliente de más de 40 años, casado, con un año en su empleo, de sexo femenino y de nivel medio en sus estudios obtiene $48 = 30 + 12 - 5 + 8 + 3$ puntos.

Finalmente la premisa más importante de un score es que, a mayor puntaje de score menos riesgoso es el cliente y viceversa.

Son varios los pasos a seguir para obtener la *scorecard* y muchas las metodologías que se pueden utilizar [ver Barberena (2002)]. A continuación listamos de forma general los pasos que seguimos para encontrar nuestra *scorecard*:

Conformar la base de datos. Este proceso se inicia con el vaciado de la información contenida en las solicitudes de los clientes en un archivo electrónico con un formato definido. También se construye una base que contiene el comportamiento de la mora de los clientes registrados en la base de solicitudes. Luego los datos se depuran excluyendo las variables con exceso de campos sin respuesta o respuestas múltiples, etc.

Agrupar los datos contenidos en la base. Una vez que se tiene la base de solicitudes limpia se procede a formar intervalos de clase o grupos de clase para cada característica (variable), estas clases también se les llama atributos de la característica. Los atributos se forman tomando en cuenta su proporción de buenos y malos a través de una medida conocida como *WOE*.

Determinar los clientes buenos y malos. En este paso utilizamos la base de datos que registra el comportamiento de la moratoria mensual o pagos vencidos de los clientes. Construimos con estos datos una matriz de transición de un paso de seis meses. Auxiliándonos de esta matriz se determina cuales de los clientes son buenos, cuales son malos y cuales son indeterminados en nuestra base de datos. En este punto, se forma una nueva base con los datos de los clientes que se clasificaron como clientes buenos y clientes malos (variable dependiente) con sus respectivas características (variables explicativas) extraídas de la base de solicitudes.

Determinar una función de clasificación. Una vez que se tiene la base completa, con ella se procede a hacer una selección de las características que tienen mayor valor de predicción global. Para hacer esta selección utilizamos el *Valor de Información* que es una función de la proporción de buenos y malos clientes en los atributos de cada característica. Formamos una base que contiene buenos y malos, y las características con mayor poder de predicción. De esta base se genera una nueva base, de la siguiente manera:

Una columna para la variable dependiente con los valores de 1, si el cliente fue clasificado como bueno y 0 si el cliente fue clasificado como malo. Una columna para cada característica, con los valores del WOE del atributo correspondiente para cada cliente.

Obtenemos una base que contiene únicamente los datos de los registros de los clientes buenos y malos de características con suficiente poder de predicción, reportando el valor del WOE correspondiente al atributo del cliente. Finalmente con esta base obtenemos el modelo de predicción para los nuevos clientes, para ello utilizamos la regresión logística.

Elaborar la *scorecard*. Los puntajes asignados a los atributos de cada característica se calculan en función de las estimaciones de los parámetros de la función de clasificación obtenida con la regresión logística y el *WOE* de los atributos. Se hace una calibración de estos puntajes de acuerdo a criterios propios de la empresa.

Medir la eficiencia de la *scorecard*. Utilizamos métodos estadísticos como el índice de Gini y la prueba de Kolmogorov-Smirnov para determinar que tan bien clasifica nuestro modelo de predicción.

Establecer el punto de corte. Se determina el punto de corte (*cut off*) que separará a las solcitudes nuevas en aceptados o rechazadas. Para ello calculamos el porcentaje de rechazo y la moratoria asociada para un *score* dado.

En el capítulo cinco se presentará una aplicación utilizando datos reales de una empresa que permenece en el anonimato y sus datos son manejados con la debida confidencialidad. En las secciones siguientes se explicara con detalle los puntos anteriores.

3.3. Determinación de clientes buenos y malos

3.3.1. ¿Qué es un cliente bueno y un cliente malo?

Para los efectos de la *scorecard* los clientes se clasifican en buenos y malos. Los clientes buenos son los que pagan sus mensualidades a tiempo o permanecen en mora poco tiempo. La determinación de clientes buenos y clientes malos se hace en base a varios factores entre los que se encuentran:

- El comportamiento de los clientes. Es la información contenida en el archivo de comportamiento que contiene los pagos vencidos de los clientes.

- El proceso de cobranza. Esta acción puede reducir el número de pagos vencidos de los clientes y puede en consecuencia afectar la probabilidad de que un cliente llegue a convertirse en un mal cliente. Si un cliente cayó en morosidad y por la labor del área de cobranza regresa a ser buen cliente, quiere decir que no era tan mal cliente después de todo.

- Las metas corporativas de la institución de crédito. Una institución de crédito determina el número de pagos vencidos utilizados para determinar el corte entre buenos y malos clientes. Finalmente se busca descartar esos clientes que

causan perdidas a la empresa y que no debieron ser aceptados para obtener del crédito.

Con todos estos factores se determina en que momento un cliente se considera bueno, malo o indeterminado. Existen diferentes técnicas estadísticas para determinar a los clientes buenos y a los clientes malos de una base de datos, en esta parte nosotros explicaremos como usar una matriz de transición para llagar a este objetivo. Los datos consisten en el seguimiento de clientes durante un periodo de tiempo que se le conoce como ventana de muestreo.

3.3.2. Ventana de muestreo

El periodo de exposición o ventana de muestreo es un periodo de tiempo en el que se observa el comportamiento de las cuentas a partir de su apertura. Conforme avanza la edad de las cuentas la tasa de moratoria va variando y se espera que en un momento determinado se estabilice, esto significa que a partir de ese momento ya podemos clasificar con una variación mínima a un cliente como bueno o malo. En esta etapa la cantidad de cuentas indeterminadas se ha reducido significativamente ya que estas no figuran en el modelo, de modo que al quitarlas se reduce nuestra base de datos. Para este propósito es útil un grafico que contenga la edad de la cuenta contra la tasa de malos, como el de la figura 3.1, en el podemos observar el comportamiento en mora de los clientes a partir de su alta.

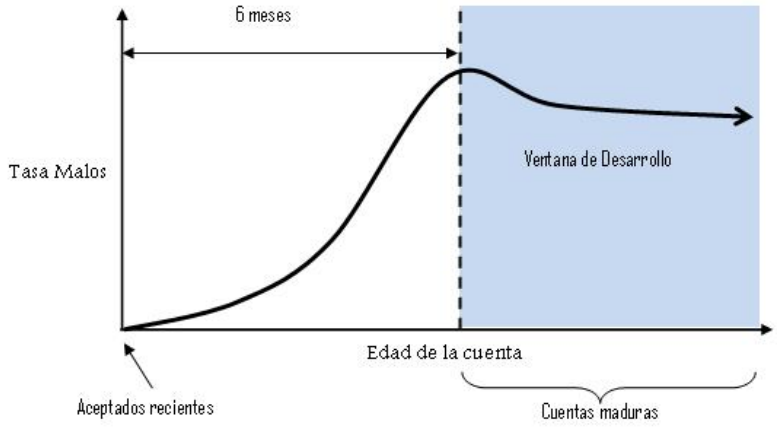


Figura 3.1: Ventana de muestreo. Periodo de observación del comportamiento de la tasa de mora en las cuentas a partir de su alta

Cuando la tasa de morosidad se ha estabilizado se dice que las cuentas llegan a su madurez de comportamiento. No se debe perder de vista que la muestra tomada debe representar a la actual población; es decir, si se toman muestras muy lejanas en el tiempo, las características pueden no representar a las actuales y si tomamos muy cercanas reducimos el tamaño de la muestra porque el periodo de exposición se reduce y por tanto en ella el número de cuentas observables. Usualmente se observan en un intervalo de tiempo de entre 12 y 18 meses (anteriores a la fecha en que se hace el estudio), para el periodo de exposición. Esto permitirá determinar cuanto tiempo se requiere para que las cuentas adquieran su madures (ver gráfica 3.1). Pueden pasar hasta 15 meses para que se logre. Después de este intervalo de tiempo, empieza un periodo de tiempo con comportamiento estabilizado de la tasa de morosos conocida como *ventana de desarrollo* como se muestra en la figura 3.1. La curva empieza a decaer cuando la cuentas se estabilizan y son quitados los clientes que son considerados malos.

Una vez que se conoce el tiempo que se requiere para llegar a la estabilización de las cuentas y la ventana de desarrollo se debe establecer que cuentas están contenidas en este periodo, para que formen parte de la muestra que se usará para construir los modelos estadísticos, en consecuencia requiere determinar el periodo de tiempo donde estarán incluidas las fechas de alta de estas cuentas.

3.3.3. Proceso para determinar a los clientes buenos y a los clientes malos

Para determinar a los clientes buenos y a los clientes malos se construye una matriz de transición para analizar la moratoria de los clientes. Con esta matriz de transición veremos el comportamiento de las cuentas en la institución después de un periodo de tiempo, generalmente de seis meses. Aquí se identifican los estados que se utilizaran como marca para determinar si un cliente es bueno, indeterminado o malo. Las cuentas se analizan en un periodo de tiempo a partir de la fecha que se aperturan. Se debe correr un tiempo a partir del estado inicial para analizar su comportamiento.

Los estados de la matriz de transición se definen en función del número de pagos vencidos, y una vez que se tienen definidos los estados se estiman las probabilidades de transición, $P(X_2 = j|X_1 = i)$, con los datos de nuestra base. Esta matriz de transición en el contexto del *credit scoring* se le conoce como *Roll Rate*.

Independientemente del momento en que se apertura una cuenta el periodo de observación o ventana de muestreo se contabiliza como primer mes, segundo mes, tercer mes, etc. a partir del momento de apertura, esto es, debemos “alinear” todas

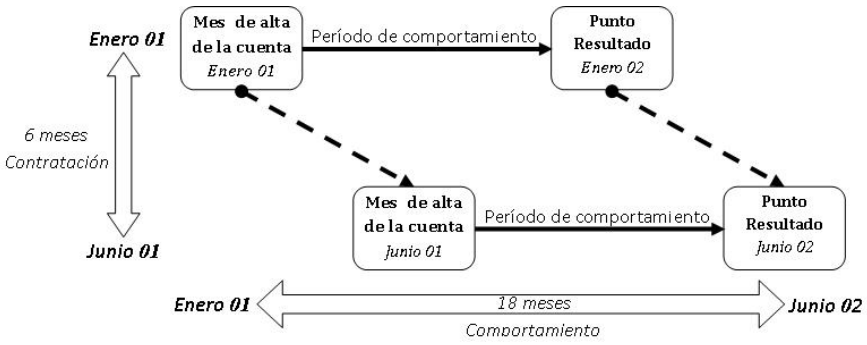


Figura 3.2: Cuentas consideradas en un periodo de contratación y un periodo de comportamiento

las cuentas a un punto inicial igual a cero (figura 3.2)

Todas las cuentas parten ahora de un mismo punto y los estados de la matriz de transición se definen de acuerdo al comportamiento de las cuenta durante cada seis meses (figura 3.3), y el estados de la cuenta al final de los seis meses.

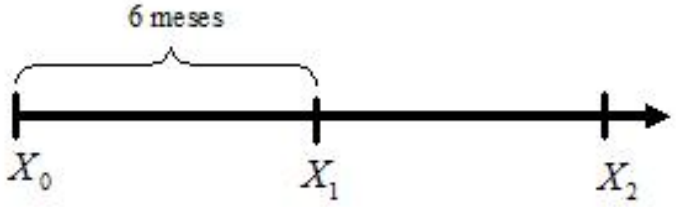


Figura 3.3: Estados de la matriz de transición en periodos de seis meses.

Los estados deben ser una partición de las cuentas; esto es, una cuenta debe pertenecer exclusivamente a un estado. Considerando la situación de la cuenta al final de los seis meses, los estados se muestran en el cuadro 3.2

Al incorporar la información de las cuentas durante los seis meses obtenemos un refinamiento de estos estados, por ejemplo, el estado $PV0$ (*current*) lo particionamos en cinco nuevos estados que se reportan en la tabla 3.3.

El estado $PV1$ se particiona en cuatro nuevos estados que se encuentran en la tabla 3.4.

Nótese que en la tabla anterior no aparece el estado $PV01$ dado que no tendría

Estados	Descripción al final de los 6 meses
PV0	Al corriente
PV1	1 pago vencidos
PV2	2 pagos vencidos
PV3	3 pagos vencidos
PV4	4 o más pagos vencidos

Cuadro 3.2: Posibles estados al final de 6 meses

Estados	**Descripción al final de los 6 meses**
PV00	al corriente y máximo 0 pagos vencidos durante los 6 meses
PV01	al corriente y máximo 1 pagos vencidos durante los 6 meses
PV02	al corriente y máximo 2 pagos vencidos durante los 6 meses
PV03	al corriente y máximo 3 pagos vencidos durante los 6 meses
PV04	al corriente y 4 o más pagos vencidos durante los 6 meses

Cuadro 3.3: Posibles estados con 0 PV en los primeros 6 meses

sentido de hablar de clientes con un pago vencido y en los seis meses anteriores cero pagos vencidos como máximo. Esta misma idea se utiliza para construir la siguiente partición para los demás estados. Así, el último estado sería $PV44$ como único elemento para 4 o más pagos vencidos en el sexto mes y como máximo 4 o más pagos vencidos. Así construimos una partición de 15 estados. El estado $PV44$ es un comportamiento indeseado por las intituciones de crédito, los clientes que están en este estado se pueden identificar como malos "graves". El malo "grave" es típicamente definido basándose en castigo o *write off* o mora mayor a 90 días.

Ahora, buscamos encontrar que estados en X_1 conducen al estado no deseado, $PV44$, en X_2. Los elementos de la matriz de transición (*roll rate*) son las estimaciones de las probabilidades de estar en el estado j en los primeros seis meses a pasar al estado i en los seis meses siguientes

$$\widehat{P}(X_2 = i | X_1 = j) =$$

$$= \frac{\text{Número de cuentas que estaban en el estado } j \text{ y pasaron al estado } i}{\text{Número de cuentas que estaban en el estado } j}.$$

Una vez que se tenga la matriz de transición correspondiente debe hacerse un análisis para determinar que estados tienen una alta probabilidad de pasar al

Estados	Descripción al final de los 6 meses
PV11	1 pago vencido y máximo 1 pagos vencidos durante los 6 meses
PV12	1 pago vencido y máximo 2 pagos vencidos durante los 6 meses
PV13	1 pago vencido y máximo 3 pagos vencidos durante los 6 meses
PV14	1 pago vencido y 4 o más pagos vencidos durante los 6 meses

Cuadro 3.4: Posibles estados con 1 PV en los primeros 6 meses

estado $PV44$ o en el siguiente periodo. La matriz estimada es como la presentada en el ejemplo 2.1.3.

La idea es estimar la probabilidad de caer en $PV44$, esto es:

$$P(X_2 = PV44 \mid X_1 = i) \quad \text{donde} \quad i = PV00, PV01, ..., PV44$$

y proceder de la siguiente manera

- Si $\widehat{P}(X_2 = PV44 \mid X_1 = i) < a$, las cuentas que están en el estado i se consideran cuentas buenas.
- Si $\widehat{P}(X_2 = PV44 \mid X_1 = i) > b$, las cuentas que están en el estado i se consideran cuentas malas.
- Si $a \leq \widehat{P}(X_2 = PV44 \mid X_1 = i) \leq b$, las cuentas que están en el estado i se consideran cuentas indeterminadas.

Los números a y b mencionados arriba satisfacen la relación $0 < a < b < 1$ y la institución crediticia es quien decide el valor de estos números.

3.4. Obtención de la función de clasificación

Una vez que se tiene identificados a los clientes buenos y a los clientes malos se procede a estimar una función para clasificar a los nuevos clientes y así poder determinar si se les otorga o no el crédito que están solicitando. La función de clasificación depende de las variables que están en el archivo de solicitud y de buró de crédito que no fueron excluidas de la base de datos. Debido a la naturaleza de las variables con las que se va a calcular la función clasificadora se elige a la regresión logística para estimar a la fución clasificadora.

3.4.1. Definición de parámetros

Para la característica (variable explicativa) i se obtiene el número de cuentas buenas y el número de cuentas malas que se denotan como b_i y m_i respectivamente. Para el atributo (valor) j de la característica i se obtiene el número de cuentas buenas y el número de cuentas malas y se denota por b_{ij} y m_{ij} respectivamente, de esta manera se satisfacen las relaciones

$$b_i = b_{i1} + b_{i2} + b_{i3} + \ldots + b_{in_i}$$

y

$$m_i = m_{i1} + m_{i2} + m_{i3} + \ldots + m_{in_i},$$

donde n_i es el número de atributos para la característica i Con estos valores se calculan los siguientes términos:

- La distribución de buenos en la característica i

$$Pb_{ij} = \frac{\text{no. de buenos en el atributo } j \text{ de la característica } i}{\text{no. de buenos en la característica } i} = \frac{b_{ij}}{b_i}$$

- La distribución de malos en la característica i

$$Pm_{ij} = \frac{\text{no. de malos en el atributo } j \text{ de la característica } i}{\text{no. de malos en la característica } i} = \frac{m_{ij}}{m_i}.$$

- La distribución de atributos de la característica i

$$Pc_{ij} = \frac{b_{ij} + m_{ij}}{b_i + m_i}.$$

- El *Bad Rate* que corresponde a la proporción de malos respecto del total de casos en el atributo j de la característica i

$$M_{ij} = \frac{m_{ij}}{b_{ij} + m_{ij}}.$$

- El *Good:Bad odds* que corresponde a la proporción de buenos respecto de los malos, término que también se conoce como la probabilidad de Buenos/Malos

$$O_{ij} = \frac{b_{ij}}{m_{ij}}.$$

Para ejemplificar esta etapa en el procedimiento presentamos una tabla con los cálculos correspondientes a la variable o característica "Tipo de Universidad".

Atributo	m_{ij}	b_{ij}	$m_{ij} + b_{ij}$	Pm_{ij}	Pb_{ij}	Pc_{ij}	M_{ij}	O_{ij}
NULL	1	25	26	0.0002	0.0003	0.0003	0.0385	25:1
Pública	2498	47835	50333	0.4991	0.6685	0.6575	0.0496	19.1:1
Privada Tipo I	273	3710	3983	0.0545	0.0519	0.0520	0.0685	13.6:1
Privada Tipo II	353	4030	4383	0.0705	0.0563	0.0573	0.0805	11.4:1
Privada Tipo III	784	7427	8211	0.1566	0.1038	0.1073	0.0955	9.5:1
Privada Tipo IV	1096	8525	9621	0.2190	0.1191	0.1257	0.1139	7.8:1
Total	5005	71552	76557	1.0	1.0	1.0	0.0654	14.3:1

Cuadro 3.5: Característica "Tipo de Universidad" y sus atributos

Estos datos corresponden a una muestra de 76557 cuentas de solicitantes de una institución de crédito. En el primer renglón de la tabla 3.7 se encuentra el atributo "NULL" que contiene 26 datos sin presencia significativa. Si vemos los datos como porcentajes, en el segundo renglón aparece la información del atributo "Pública" con un 65.7 % del total de la muestra (ver columna $100 \cdot Pc_{ij}$), este mismo atributo tiene un 49.9 % en la distribución de malos y un 66.9 % en la distribución de buenos respectivamente. Es interesante observar que los clientes que estudian en una universidad privada son más proclives a caer en mora que los clientes que estudian en universidades públicas, como se ve en la columna de "M_{ij}" donde los datos presentan un orden creciente en el *Bad Rate*. Los clientes de las universidades "Privada Tipo IV" tienen mayor proporción de malos, esto es $343/4383 = 0.114$ que corresponde al 11.4 % del total de casos en ésta clase. En consecuencia la columna O_{ij} muestra una probabilidad decreciente.

Para aclarar los resultados en la columna *Good:Bad odds* (O_{ij}) mencionamos que en esta columna se muestra el número de buenos por cada malo.

Una vez que tenemos las características con las que se va a efectuar la clasificación de buenos y malos se procede a formar grupos de atributos que tengan semejante proporción de buenos y malos, esto es, igual o semejante O_{ij} o M_{ij}. A esta agrupación se le conoce como "Clasificación dura".

También se deben tomar en cuenta consideraciones operacionales y de negocios tales como políticas de la corporación ya establecidas para hacer los grupos en los atributos, así como grupos con sentido lógico.

La clasificación dura consiste en juntar atributos con proporción semejantes de buenos y malos para tener un número más pequeño de atributos en cada carac-

terística. Esto se requiere para generar una *scorecard* con la forma que se muestra en el cuadro 3.1 donde las característica tanto continuas como discretas están agrupadas en un número reducido de atributos, estos atributos son intervalos, clases o grupos de variables.

Para ejemplificar como se forman los grupos consideramos los datos de la característica "Tipo de Universidad" vistos anteriormente (cuadro 3.7). En esta tabla se tiene el *Good:Bad odds*, en la última columna de la tabla ordenados de mayor a menor. Agrupamos el atributo "NULL" con el atributo "Pública" a que sus *odds* están contiguos en el ordenamiento, y "NULL" tiene únicamente 26 casos (menos del 5 % del total), los atributos Privada Tipo I y II los agrupamos en una clase dado que tienen odss cercanos y contiguos, así mismo con los atributos Privada Tipo III y IV se forma un nuevo grupo al final tres grupos.

Con esta agrupación se pretende que la diferencia entre el *Goog:Bad odds* de los diferentes atributos de una misma característica sean significativamente diferentes. Para medir esta diferencia se utilizan algunos estadísticos como la χ^2 y el *Valor de Información (IV)* que se verá en la siguiente subsección.

3.4.2. Pesos de Evidencia, WOE

El poder de predicción en cada atributo o grupo de atributos se calcula con los *Pesos de Evidencia* (*Weigth of Evidence*) que se denotan como WOE, que es una medida entre la diferencia de las proporciones de buenos y malos en cada atributo. La definición del WOE es

$$\begin{aligned} WOE_{ij} &= 100 \cdot \ln\left(\frac{\text{Distribución de buenos en el atributo } j \text{ de la característica } i}{\text{Distribución de malos en el atributo } j \text{ de la característica } i}\right) \\ &= 100 \cdot \ln\left(\frac{Pb_{ij}}{Pm_{ij}}\right) \\ &= 100 \cdot \ln\left(\frac{b_{ij} \cdot m_i}{m_{ij} \cdot b_i}\right). \end{aligned}$$

Cuando se obtienen valores negativos del WOE significa que se tienen proporciones altas de malos sobre los buenos. De ésta medida se toman en cuenta algunas consideraciones:

- El WOE_{ij} varía dependiendo de la forma en que se agrupan los atributos. Se acostumbra a ordenar de manera creciente al WOE.
- Para que el WOE_{ij} este definido, ninguna de las clases debe estar formada únicamente por buenos o por malos.

Atributo	*Malo*	*Bueno*	*Total*	Pm_{ij}	Pb_{ij}	woe_{ij}
Pública, NULL	2499	47860	50359	0.4993	0.6689	29.240
Privada Tipo I y II	626	7740	8366	0.1251	0.1082	-14.518
Privada Tipo III y IV	1880	15952	17832	0.3756	0.2229	-52.167
Total	5005	71552	76557	1.0	1.0	

Cuadro 3.6: WOE de los atributos para "Tipo de Universidad"

- Se sugiere no tener más de 8 clases y para que cada clase sea significativa debe contener al menos un 5 % de los datos analizados.

Esto permite identificar datos *outliers* y clases raras, además de identificar comportamientos y adquirir conocimiento acerca del portafolio.

Ejemplo 3.4.1. *Consideremos los datos de la siguiente tabla (cuadro 3.6) que muestran el WOE para cada atributo de "Tipo de Universidad". Obsérvese que en la columnas Pm_{ij} y Pb_{ij} se muestran las distribuciones respectivas de malos y buenos. El WOE del atributo "Pública-NULL" es:*

$$100 \ln\left(\frac{0.6689}{0.4993}\right) = 29.240$$

3.4.3. Pruebas sobre la clasificación de atributos

Estadístico χ^2

El número esperado de buenos y malos en el atributo j de la característica i bajo el supuesto que en cada atributo de la misma característica, la respectiva proporción de buenos y malos es igual que la proporción en el total de la característica, se define por

$$\hat{b}_{ij} = \frac{(b_{ij} + m_{ij})b_i}{b_i + m_i} \quad \text{y} \quad \hat{m}_{ij} = \frac{(b_{ij} + m_{ij})m_i}{b_i + m_i}.$$

Entonces el estadístico χ^2_c para la característica i esta dado por

$$\chi^2_c = \sum_{j=1}^{n_i} \left(\frac{(b_{ij} - \hat{b}_{ij})^2}{\hat{b}_{ij}} + \frac{(m_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \right) \sim \chi^2_{k-1} \tag{3.1}$$

que será "pequeño" si $b_{ij} \simeq \hat{b}_{ij}$ y $m_{ij} \simeq \hat{m}_{ij}$, y será "grande" si no se da la semejanza. De esta manera, χ^2_c se usa para determinar cuan diferentes están los *odds* en

cada clase. Entre más grande resulte el estadístico éste refleja mayores diferencias en los *odds*, por lo que al comparar dos formas de agrupar tomamos el más alto como una mejor forma de dividir la característica y de manera absoluta la probabilidad $P(\chi^2 > \chi_c^2)$ es una estimación de la probabilidad que se equivoque uno al afirmar que existe diferencia significativa de los *Good:Bad odds* de la característica.

Ejemplo 3.4.2. *Con los datos del atributo para "Tipo de Universidad" de la tabla 3.5 se efectuaran tres formas diferentes de establecer los grupos o clases. La primera clasificación (Clasificación A) corresponde a la vista anteriormente, esto es; Grupo 1 "NULL" y "Pública", Grupo 2 "Privadas Tipo I" y "Privadas Tipo II" y grupo 3 "Privadas Tipo III" y "Privadas Tipo VI". Los datos para esta agrupación se presentan en la siguiente tabla (cuadro 3.7). El cálculo del estadístico* χ_c^2 *es*

Atributos	m_{ij}	b_{ij}	$b_{ij} + m_{ij}$	$\hat{b}_{ij}$	$\hat{m}_{ij}$
Pública, NULL	2499	47860	50359	47067	3292
Privada Tipo I y II	626	7740	8366	7819	547
Privada Tipo III y IV	1880	15952	17832	16666	1166
Total	5005	71552	76557	71552	5005

Cuadro 3.7: Datos para obtener χ^2 en una forma de agrupar los atributos de "Tipo de Universidad"

$$\begin{aligned}\chi_c^2 &= \frac{(47860-47067)^2}{47067} + \frac{(2499-3292)^2}{32292} + \frac{(7740-7819)^2}{7819} \\ &\quad + \frac{(626-547)^2}{547} + \frac{(15952-16666)^2}{16666} + \frac{(1880-1166)^2}{1166} \\ &= 204.5 + 49.7 + 519.1 \\ &= 773.3\end{aligned}$$

La segunda forma de clasificar (Clasificación B) es "NULL" con "Pública" (47860 buenos y 2499 malos), Privada tipo I, II y III (15167 buenos y 1410 malos), y Privada Tipo IV (8525 buenos y 1096 malos) obtenemos

$$\begin{aligned}\chi_c^2 &= \frac{(47860-47067)^2}{47067} + \frac{(2499-3292)^2}{3292} + \frac{(15167-15493)^2}{15493} \\ &\quad + \frac{(1410-1084)^2}{1084} + \frac{(8525-8992)^2}{8992} + \frac{(1096-629)^2}{629} \\ &= 680.6\end{aligned}$$

Con los cálculos anteriores podemos concluir que la mejor manera de agrupar los atributos corresponde al primer cálculo, obtenemos valor mayor de χ_c^2 para el primer grupo (NULL-Pública, Privada I-II y Privada III-IV).

Valor de Información, IV

El *Valor de Información* (IV) es una medida de entropía que aparece en la teoría de información [ver Siddiqi (2006)] y se define por

$$IV = \sum_i \left(\frac{b_{ij}}{b_i} - \frac{m_{ij}}{m_i} \right) \ln \left(\frac{b_{ij} m_i}{m_{ij} b_i} \right). \tag{3.2}$$

Los valores que puede tomar el estadístico IV son no negativos, y es cero cuando $\frac{b_{ij}}{b_i} = \frac{m_{ij}}{m_i}$ lo que equivale a que $b_{ij} = \hat{b}_{ij}$ ó $m_{ij} = \hat{m}_{ij}$, como se decuce directamente de la definición.

El IV es una medida del poder de predicción global de la característica.

Siddiqi (2006) considera que una característica con un IV

- menor a 0.02 es impredictivo
- entre 0.02 y 0.1 es de predicción débil
- entre 0.1 y 0.3 es de predicción medio
- más de 0.3 es de predicción fuerte.

Segun Siddiqi (2006) indica que con las características con IV por debajo de 2 % deben ser excluidas del modelo. Cuando se obtiene un IV mayor de 0.5 se dice que la característica está sobre prediciendo. En este trabajo se siguió el mismo critério.

Con el IV podemos medir el poder de predicción de dos maneras diferentes de agrupar los atributos de una característica como se ve en el siguiente ejemplo.

Ejemplo 3.4.3. *Retomamos los datos de la característica "Tipo de Universidad" que se registran en el cuadro 3.7 y calculamos el IV para las dos clasificaciones dadas anteriormente.*

Clasificación A:

$$\begin{aligned} IV &= (0.669 - 0.499) \ln(\frac{0.669}{0.499}) + (0.108 - 0.125) \ln(\frac{0.108}{0.125}) \\ &+ (0.222 - 0.376) \ln(\frac{0.222}{0.376}) \\ &= 0.132, \end{aligned}$$

Clasificación B:

$$\begin{aligned} IV &= (0.669 - 0.499)\ln(\frac{0.669}{0.499}) + (0.212 - 0.282)\ln(\frac{0.212}{0.282}) \\ &+ (0.119 - 0.219)\ln(\frac{0.119}{0.219}) \\ &= 0.130, \end{aligned}$$

Se obtiene una mayor diferencia en el primera clasificación por lo que nuevamente, los datos indican que es la mejor opción para agrupar las clases.

3.4.4. El modelo logístico

Antes de generar el modelo de clasificación debemos elegir las características que seran consideradas en la regresión logística, para ello utilizamos el *Valor de Información*. Aquellas características con un IV menor a 0.02 no son consideradas en esta etapa.

El modelo de regresión logística esta dado por

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

donde β_0, β_1, $\beta_2, \ldots,$ β_k son parámetros desconocidos y x_1, $x_2, \ldots,$ x_k son variables explicativas cuyos valores numéricos están en función de la proporción de buenos y malos en cada atributo. El rango de valores de la variable explicativa x_i asociada a la característica i es igual a

$$x_i = woe_{i1},\ woe_{i2}, \ldots, woe_{in_i},$$

donde

$$woe_{ij} = \ln\left(\frac{b_{ij} \cdot m_i}{m_{ij} \cdot b_i}\right),$$

tal que n_i es el número de atributos de la característica i y $j = 1, 2, \ldots, n_i$.

De este proceso obtenemos los valores β_i para el siguiente paso.

3.4.5. Construcción de la *Scorecard*

La scorcard se construye con los estimadores de los parámetros de la regresión logística. Los puntajes del score son resultado de un reescalamiento y una traslación del modelo logístico, dado por la ecuación

$$\text{Score} = \text{Offset} + \text{Factor} \cdot \ln(\text{odds}),$$

donde *offset* es un término de traslación (o compensación) y *Factor* es un término de reescalamiento. *Offset* y *Factor* deben satisfacer condiciones impuestas por la empresa de crédito. Este procedimiento permite la estandarización del score para que diferentes scorcards puedan ser comparadas. Los valores de la *scorecard* son resultado de una transformación de los coeficientes β_i del modelo de regresión logística. La transfromación tiene como finalidad obtener valores enteros para cada atributo j de la característica i en un rango considerado adecuado por la empresa crediticia. Se acostumbra a calibrar la *scorecard* de tal manera que cada cierto incremento en el puntaje P_0, se obtenga el doble de la relación *good/bad*. Para obtener los valores de *Offset* y *Factor* se resuelve el siguiente sistema de ecuaciones

$$\begin{aligned} \text{Score} &= \text{Offset} + \text{Factor} \cdot \ln(\text{odds}) \\ \text{Score} + \text{P}_0 &= \text{Offset} + \text{Factor} \cdot \ln(2 \cdot \text{odds}), \end{aligned}$$

de aquí obtenemos

$$\begin{aligned} \text{Factor} &= \frac{P_0}{ln(2)}; \\ \text{Offset} &= \text{Score} - \text{Factor} \cdot \ln(\text{Odds}) \end{aligned}$$

Por ejemplo si consideramos que un *Odds* de 1:1 equivale a 600 puntos en la *scorecard*, que los *odds* se duplican cada $P_0 = 80$ puntos en la *scorecard*; es decir, que 680 puntos equivalen a un *odds* de 2:1, a los 760 puntos equivalen a 4:1 y así sucesivamente. Entonces los valores de *Factor* y *Offset* quedan como:

$$\begin{aligned} \text{Factor} &= \frac{80}{ln(2)} = 115.4156; \\ \text{Offset} &= 600 - 115.4156 \cdot \ln(1) = 600 \end{aligned}$$

Con esto se obtiene la función de score dada por

$$Score = 600 + 115.416 \cdot \ln(odds).$$

Estos puntajes están en escala lineal y la relación del modelo de regresión logística con los WOE está dada por

$$\begin{aligned} score &= \text{Offset} + \text{Factor} \cdot \ln(\text{odds}) \\ &= \text{Offset} + \text{Factor} \cdot \left(\hat{\beta}_0 + \sum_{\text{i}=1}^{\text{p}} \hat{\beta}_{\text{i}} \cdot \text{woe}_{\text{ij}} \right) \\ &= \text{Offset} + \text{Factor} \cdot \hat{\beta}_0 + \text{Factor} \cdot \left(\sum_{\text{j}=1}^{\text{p}} \hat{\beta}_{\text{j}} \cdot \text{woe}_{\text{ij}} \right) \end{aligned}$$

De ésta última ecuación se puede ver que los puntajes en la scoreard se pueden descomponer en un puntaje inicial dado por

$$\text{Offset} + \text{Factor} \cdot \hat{\beta}_0,$$

y el puntaje asociado al atributo j de la característica i dado por:

$$Factor \cdot \hat{\beta}_i \cdot woe_{ij}.$$

Es claro que los puntajes de una *scorecard* depende de los parámetros de traslación y reescalamiento que se utilicen

3.5. Determinación del punto de corte o *Cut Off*

Cuando se tiene los datos de un nuevo solicitante, se calcula su score y con el resultado se decide si se le otorga o no el crédito.

Si $score > a$ se otorga el crédito, en caso contrario si $score \leq a$ se rechaza la solicitud. El punto "a" se conoce como punto de corte o *Cut Off* y es importante determinarlo para optimizar la decisión. En esta sección presentamos dos maneras de estimar el punto e corte.

La primera forma es cuando la desigualdad

$$\frac{e^{\hat{\beta}_0+\hat{\boldsymbol{\beta}}_1^T\mathbf{x}}}{1+e^{\hat{\beta}_0+\hat{\boldsymbol{\beta}}_1^T\mathbf{x}}} = \hat{p} > \frac{1}{2},$$

de aquí se sigue que

$$\Rightarrow e^{\hat{\beta}_0+\hat{\boldsymbol{\beta}}_1^T\mathbf{x}} > \frac{1}{2}\left(1+e^{\hat{\beta}_0+\hat{\boldsymbol{\beta}}_1^T\mathbf{x}}\right)$$

$$\frac{1}{2}e^{\hat{\beta}_0+\hat{\boldsymbol{\beta}}_1^T\mathbf{x}} > \frac{1}{2} \Rightarrow e^{\hat{\beta}_0+\hat{\boldsymbol{\beta}}_1^T\mathbf{x}} > 1$$

$$\Rightarrow \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1^T\mathbf{x} > 0$$

$$\Rightarrow a = \textit{Offset}.$$

La segunda forma de obtener el punto de corte es calcular el score para todas las cuentas de la base, estos *score's* se ordenan y el valor de a satisface la ecuación

$$\frac{no.\ de\ score > a}{no.\ de\ score} = un\ porcentaje.$$

El porcentaje es un valor seleccionado por la empresa.

Capítulo 4

Una Aplicación del Credit Scoring

Para la aplicación de la técnica *credit scoring* utilizamos una base de datos de tarjetas de crédito que por confidencialidad cambiaremos los nombres de algunas variables, fechas y demás. Llamaremos a la fuente "Empresa de Crédito" para referirnos a la institución donde se generaron los datos.

Para tener un modelo de originación que clasifique los clientes en buenos o malos se requiere tener una base de datos apropiada que contenga las variables independientes y la variable clase. Para obtener la base de datos apropiada se sigue un proceso que tiene un alto costo de tiempo, ya que se requiere limpiar, crear nuevas variables, relacionar datos, etc. Para estos efectos utilizamos Excel y Acces.

La base de datos con la que trabajamos es un segmento de la población total crediticia de la "Empresa de Crédito". En particular nuestros objetos de estudios son estudiantes universitarios que los hacen diferentes al resto de los clientes. Utilizamos dos bases, una que contiene los registros de las solicitudes "Autorizadas" y otra que contiene el número de pagos vencidos (PV) de estos clientes en los meses siguientes a su contratación. Cabe mencionar que en clientes de población abierta se utiliza una tercera base con los datos de buró de crédito y esta base no existe información de la población de estudiantes universitarios.

4.1. Archivos requeridos

Para hacer los análisis adecuados a las variables o construir clasificadores eficientes y confiables es necesario conformar la base de datos con una estructura específica. La base de datos debe ser funcional y apropiada para la mayoría de los software estadísticos. La mayoría de la paquetería o software requiere que los datos

se encuentren en archivos de texto en forma de matriz en la cual cada columna corresponde a una variable, tanto de respuesta como variable predictora.

El formato de los archivos deben poseer un identificador único del cliente que permita relacionar sus datos contenidos en los diferentes archivos. Los archivos deben cubrir algunos requisitos como:

- Deben ser archivos planos.
- Deben contener un registro por cliente.
- Deben estar en formato ASCII.
- Deben contener formatos de campos, y descripciones de códigos entre otros.

La base de datos final se construye con la información de los siguientes archivos:

Datos de solicitud

Es un archivo de solicitudes recolectadas en la ventana de tiempo de estudio, que contiene los datos recolectados en las solicitudes que llenan los interesados en obtener una tarjeta de crédito al momento de la originación. Este archivo contiene la información de todas las solicitudes ingresadas tanto de los clientes que fueron aceptados como de los clientes que fueron recha-zados y cuyas solicitudes tengan al menos 12 meses después que se origino el crédito.

Datos de buró de crédito

En el caso de México, Buró Nacional de Crédito es la principal institución que maneja el historial de crédito de los mexicanos, aunque hay otras instituciones encargadas de manejar información de este tipo. Manejan datos del sistema financiero tales como: número de tarjetas de crédito que maneja el cliente, su historial de mora, etc. de aquellos solicitantes que tienen o han tenido un crédito hipotecario, de tarjetas de crédito, que realizan pagos fijos de servicios, etc. Este archivo suele generarse con consultas a una base externa. La consulta de la base de datos de buró tiene un costo para quien la consulta.

Datos de comportamiento

Esta base contiene las cuentas autorizadas en un periodo determinado y el comportamiento de pago mensual, esto puede consistir en registros mensuales de los pagos vencidos después de las cuentas contratadas después de su primera compra.

Fechas	Descripción
1° enero 2001 al 31 de diciembre 2001	Ventana de muestreo temporal de contratación de las cuentas. A partir de este periodo se tomara una muestra para construir el modelo de originación.
1° enero 2001 al 30 de junio 2002	Ventana temporal donde se observará el número de PV en cada mes de cada cuenta.

Cuadro 4.1: Primer periodo de observación para posteriormente delimitar la ventana de tiempo.

Los datos de comportamiento de pago permitirán calcular la variable respuesta para la construcción del modelo que será la marca de cliente bueno o malo.

Es importante hacer un análisis de la muestra de datos que se utilizará para el desarrollo de los modelos. Podemos considerar las cuentas dadas de alta en un año para observar su mora en una ventana de tiempo temporal hasta por 18 meses, esto es una solicitud recibida en un mes dado deberá tener consecutivamente mes a mes el número de PV hasta 12 meses, como se ve en la figura 3.2.

La gráfica 3.2 dice que tomemos 6 meses de contratación, podemos ampliar este periodo hasta por un año (ventana temporal) para observar el comportamiento de la moratoria.

Un primer acercamiento para definir las ventanas de tiempo se muestra en el cuadro 4.1 para tener una idea global del comportamiento de la cuentas.

4.2. Limpieza de la base de datos

Es importante que antes de cualquier análisis se limpie la base de datos, ya que en el proceso llenado de solicitudes y de captura de datos se genera una serie de errores que deben ser eliminados o corregidos. Los datos que utilizamos para esta aplicación no habían sido limpiados por lo que el primer paso que dimos fue limpiarlos, este proceso se realizo básicamente con Excel.

4.2.1. Posibles errores en los campos de datos

Cuando se genera una base de datos a partir de la captura de información de los formularios, se pueden introducir diferentes tipos de errores, entre los errores que se pueden generar en esta etapa están:

- Datos fuera de rango, alguien menor de 18 años o mayor de 70 años no es objeto de crédito por lo tanto datos con estos valores debieron ser mal capturados.

- Campos con valores inconsistentes. Es cuando no hay concordancia entre los valores de dos o más campos.

- Datos falsos.

- Campos vacíos o nulos.

La primera acción tendiente a corregir este tipo de errores es acudir a los formularios originales para revisar si el error detectado es debido a una mala captura. Cuando la base se genera automáticamente al obtener la información del cliente se reducen los errores de captura. La base de datos debe reflejar la información de las solicitudes de crédito al momento de ser recibidas, por lo que debe ser sometida a procesos de calidad y consistencia para detectar eventuales errores en la información. Se pueden aplicar pruebas estadísticas para datos atípicos que permita encontrar inconsistencias en los datos.

4.2.2. Campos vacíos

El valor "NULL" ó "vacío" significa lo mismo. Los registros que están totalmente nulos son quitados de la base. Por ejemplo alguien que no llenó ningún campo en la solicitud o únicamente llenó el campo correspondiente a su nombre, se elimina del estudio. Los registros que solo tienen algún campo nulo o vacío pero en los demas campos si tienen información si se consideran para el análisis. Se puede estimar el valor correspondiente a un campo con valor nulo o vacío considerando la combinación de los atributos cuyos campos están llenos, de esta manera se puede deducir un posible valor para ser asignado al campo sin respuesta. Por ejemplo, considere un registro que tiene el valor nulo en el campo "estado civil", y en el campo "edad" dice que tiene 18 años, en el campo "nivel de estudios" dice medio, en el campo "número de dependientes" dice 0 esto hace deducir que un valor adecuado para el campo nulo estado civil es "soltero". Finalmente es decisión de quien hace el análisis como debe considerar estos valores faltantes que han sido corregidos. En particular en el tratamiento de nuestra base de datos existían variables que tenían valor "NULL" para diferentes clientes por lo tanto fueron eliminadas. Después de esta depuración a los campos que continúan con valores vacios los consideramos como un atributo más de la característica.

4.3. Bueno o malo en nuestra base de datos

4.3.1. Descripción

Nosotros consideramos que un cliente con 4 o más pagos vencidos es un cliente con el estatus de malo (muy malo) y son quienes se considera que tienen una alta probabilidad de seguir siendo malos; esto es, no volverán a ser clientes buenos son malos indeseable por la institución.

Para determinar la ventana de tiempo requerido en nuestro análisis construimos una grafica con el porcentaje acumulado de clientes con tres o más meses de mora a través del tiempo. En condiciones ideales se esperaría que el comportamiento de esta grafica se estabilizara al final de la ventana de tiempo e incluso disminuira como ya se mencionó en el capítulo tres.

En la gráfica que construimos para nuestro datos (figura 4.1) observamos que esta estabilizacion y decrecimiento no aparece, solo se ve como un indicio de este comportamiento en el futuro. La razon de que esto no se observe es que no tenemos una muestra con información en un tiempo suficientemente largo.

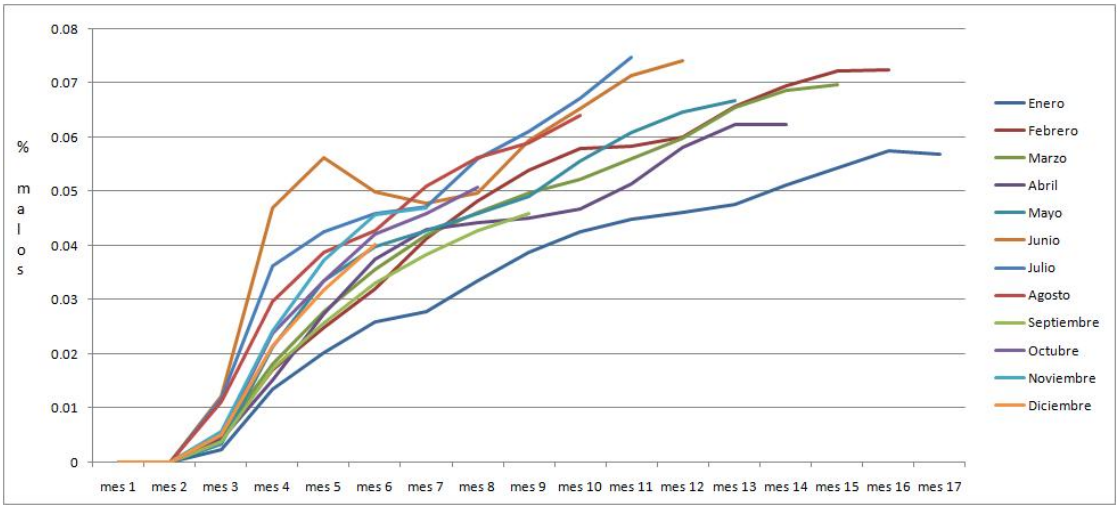


Figura 4.1: Tasas de morosidad por cada mes de contratación en base a clientes morosos con tres o más PV vencidos.

El porcentaje acumulado de clientes con 3 meses o más de mora en contratación de cada mes se registra en la figura 4.1. Con una ventana de contratación de 6 meses y con un tiempo de 12 meses de comportamiento para cada una, se requiere un periodo de 18 meses de observación en los datos de comportamiento. Por ejemplo,

Fechas	Descripción
Enero 2001 a julio 2001	Ventana final de contratación de las cuentas. A partir de este periodo se tomara una muestra para construir el modelo de originación
Enero 2001 a junio 2002	Ventana final de observación del comportamiento de la moratoria en las cuentas.

Cuadro 4.2: Ventana de aplicación a considerada para el desarrollo del modelo.

las cuentas que se abren en enero de 2001 se observan hasta enero de 2002 y las cuentas que se abren en junio de 2001 se observan hasta junio de 2002, de enero de 2001 a junio de 2002 son 18 meses. Una práctica común seguida por las empresas consultoras es observar las cuentas que se abren en el séptimo mes, aun cuando para estas cuentas no se tenga el seguimiento de 12 meses sino solo de 11 meses. El mes doce se estima con el comportamiento del mes once bajo el supuesto que dos meses contiguos no tiene una diferencia significativa en su comportamiento.

En nuestra aplicación efectuamos esta acción, así en los primeros 6 meses se abrieron 74270 y al incluir el siguiente mes tenemos siete meses con 81029 registros de cuentas autorizadas.

4.3.2. Determinación de cliente bueno o malo

Una vez que se tiene la matriz *roll rate* como se muestra en la tabla de la figura 4.2 se procede a determinar cuales de los clientes se clasificarán como buenos y cuales se clasificarán como malos.

Se considera que los malos clientes son los que caerán en alguna ocasión en *write off*. Entonces se utiliza la probabilidad de transición de cualquier estado a $PV44$; esto es, $P(X_2 = PV44 \mid X_1 = i)$, donde $i = PV00, \ldots, PV44$, probabilidad que se encuentra en la ultima columna del *roll rate* (figura 4.2) $P(X_2 = PV44 \mid X_1 = i) \leq 0.14$ los clientes que están en el estado i se consideran buenos y el valor asociado de la variable y será igual a uno. Si $P(X_2 = PV44 \mid X_1 = i) > 0.42$ entonces los clientes que están en el estado i se consideran malos y el valor de y asociado tomara el valor de cero. Los clientes que están en un estado i tal que $0.14 < P(X_2 = PV44 \mid X_1 = i) < 0.42$ se consideran indeterminados y se excluyen en la generación del modelo de clasificación. Este corte de clasificación (0.14, 0.42) es utilizado por la empresa crediticia.

Los estados con pocos datos se excluyen para la definición de buenos y malos clientes por falta de representatividad (ver figura 4.3). Así obtenemos la separación

EPV6 \ EPV12	PV00	PV01	PV02	PV03	PV04	PV11	PV12	PV13	PV14	PV22	PV23	PV24	PV33	PV34	PV44	Total
PV00	87	6	1	0	0	2	0	0	0	1	0	0	1	0	**1**	79
PV01	47	21	5	2	2	8	1	1	0	4	0	0	2	0	**5**	8
PV02	34	17	6	3	5	7	2	1	1	6	1	0	4	0	**13**	2
PV03	33	15	6	4	5	6	2	1	0	6	1	0	5	0	16	1
PV04	34	9	5	3	6	3	2	1	1	5	1	0	4	0	26	0
PV11	23	17	11	4	7	7	3	1	2	4	1	1	4	0	**14**	4
PV12	8	6	11	9	13	3	3	1	3	6	3	2	7	0	25	0
PV13	13	4	10	6	14	1	1	1	6	5	4	2	5	1	28	0
PV14	5	2	2	5	19	7	7	5	7	2	2	2	10	0	24	0
PV22	9	6	2	10	15	4	1	2	3	4	1	3	4	2	34	2
PV23	0	0	0	5	20	0	0	0	5	0	5	0	0	20	45	0
PV24	0	20	0	20	0	0	0	0	0	0	0	0	0	0	60	0
PV33	4	3	2	1	25	2	0	0	5	2	0	6	2	5	42	1
PV34	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0
PV44	5	2	1	1	19	1	0	0	1	1	0	1	2	3	64	3
Total	75	8	2	1	2	3	0	0	0	2	0	0	1	0	5	100

Figura 4.2: *Roll Rate* con los estados de la tabla 4.3 obtenida con las 81029 cuentas autorizadas.

de estados formados por clientes malos o clientes buenos (cuadro 4.4) conocida como *Good Bad Flag*.

Después de quitar los indeterminados nos queda una base de 76557 datos para generar los modelos de originación. La figura 4.3 muestra como el total de las solicitudes de universitarios se distribuyeron, primero en aceptados y rechazados y luego los aceptados en clientes buenos, clientes malos y clientes indeterminados.

Para probar que tan buena es la clasificación de los clientes en buenos, malos e indeterminados se estima la matriz de transición cuyos estados son esta clasificación. El cuadro 4.5 muestra los resultados encontrados.

Es de interés notar en esta matriz que los clientes clasificados como indeterminados tienen la misma probabilidad de pasar a cualquiera de los tres estados, por lo que consideramos que su exclusión del modelo no afecta a los resultados finales.

4.4. Selección de variables que explican el modelo

Una vez que se eliminó de la base de datos las variables con un excesivo número de campos vacíos, también se eliminan aquellas variables con rango de valores muy

Estados	Descripción	Marca
PV00	Al corriente y máximo 0 pagos vencidos durante los 6 meses	Bueno
PV01	Al corriente y máximo 1 pagos vencidos durante los 6 meses	Bueno
PV02	Al corriente y máximo 2 pagos vencidos durante los 6 meses	Bueno
PV03	Al corriente y máximo 3 pagos vencidos durante los 6 meses	Indeterminado
PV04	Al corriente y 4 o más pagos vencidos durante los 6 meses	Indeterminado
PV11	1 pago vencido y máximo 1 pagos vencidos durante los 6 meses	Bueno
PV12	1 pago vencido y máximo 2 pagos vencidos durante los 6 meses	Indeterminado
PV13	1 pago vencido y máximo 3 pagos vencidos durante los 6 meses	Indeterminado
PV14	1 pago vencido y 4 o más pagos vencidos durante los 6 meses	Indeterminado
PV22	2 pagos vencidos y máximo 2 pagos vencidos durante los 6 meses	Indeterminado
PV23	2 pagos vencidos y máximo 3 pagos vencidos durante los 6 meses	Indeterminado
PV24	2 pagos vencidos y 4 o más pagos vencidos durante los 6 meses	Indeterminado
PV33	3 pagos vencidos y máximo 3 pagos vencidos durante los 6 meses	Malo
PV34	3 pagos vencidos y 4 o más pagos vencidos durante los 6 meses	Malo
PV44	4 pagos vencidos y 4 o más pagos vencidos durante los 6 meses	Malo

Cuadro 4.3: Posibles estados de un clientes y su marca de clase asociado.

Definición	Estados
Bueno	Actualmente al día y peor situación en los últimos seis meses menor o igual a 2 PV. Actualmente 1 PV y peor situación en los últimos seis meses igual a 1 PV
Malo	Actualmente con 3 o más PV

Cuadro 4.4: *Good Bad Flag.* Definición de buenos(1) y malos(0) para las cuentas.

amplio y con información irrelevante para el estudio como el nombre de la "colonia de residencia". De igual manera se eliminan las variables binarias con uno de los atributos relativamente escasos (menos del 5 %). Después de este proceso se procede a generar variables que no venían en las solicitudes; por ejemplo, en la base calculamos la variable edad, porque en las solicitudes encontramos la fecha de evaluación del cliente y su fecha de nacimiento. La edad del cliente en el momento de la contratación se obtiene con la diferencia entre las dos fechas, el dato se redondea al entero menor.

4.4.1. Clasificación fina

La base depurada resultó tener 17 variables o características las cuales se listan en el cuadro 4.6, a cada una de ellas se les calculó su *Valor de Información* y se excluyeron para el estudio las características que tuvieron un IV menor a 0.02, estas son las variables que no tienen fuerza suficiente para diferenciar a los buenos y a los malos clientes. Como resultado de ello se obtienen 10 variables (aplicamos redondeo a dos dígitos después del punto).

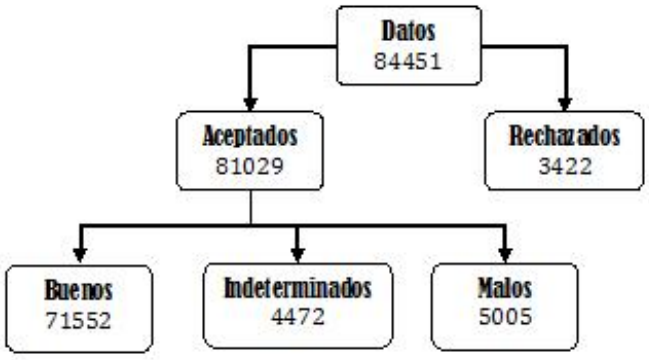


Figura 4.3: Población de Datos

	Bueno	**Indet**	**Malo**	Total	% B	% I	% M	Total
Bueno	70360	2708	2166	75234	93	4	3	93
Indet	903	891	910	2704	33	33	33	3
Malo	289	873	1929	3091	9	28	62	4
Total	71552	4472	5005	81029	88	6	6	100

Cuadro 4.5: Cantidad y porentajes asociados a una matriz de transición con tres estados (buenos, indeterminados y malos), se obtiene del *Roll Rate* anterior.

4.4.2. Clasificación dura

Las variables que quedan se reagrupan haciendo que los atributos con semejante *Bad Rate* queden en una misma clase. Por ejemplo en la tabla se muestran el cuadro 4.7 resultados obtenidos para la variable Edad.

El porcentaje de malos en las categorías 18, 19 y 24 años son semejantes (ver gráfica 4.4) y por lo tanto las agrupamos en una misma categoría; 20 y 21 forman otra categoría y finalmente 22, 23 y mayor igual a 25 la última categoría (ver gráfica 4.5 y cuadro 4.8). Para el caso de las variables nominales ordenamos los atributos de menor a mayor conforme al *Bad Rate* y de acuerdo a esta medida las agrupamos en un número más pequeño de segmentos. En la figura 4.6 se observa la segmentación de los estados de la república conforme al porcentaje de malos. Los estados de la república que no tienen presencia se incluyen en el valor más pequeño de los atributos. Se agupan de acuerdo a la segmentación del eje y. La forma final en que se organizaron los atributos se pueden observar en la tabla 4.9 que son los que se utilizan para construir el modelo.

No.	Características	IV
1	Empleo	0.018947709
2	Edad	0.001092976
3	Sexo	0.001429264
4	Tpo_Prop	0.024264078
5	Ant_Domi	0.023178872
6	Edo_Dom	0.062491986
7	Telefono	0.023528895
8	Ingre_Men	0.024937380
9	Tel_Empleo	0.003476962
10	Ant_Empleo	0.030786553
11	Dist_Pobla	0.025685096
12	Ahorro	0.003495179
13	Empresa	0.005210494
14	Tpo_Empleo	0.008913526
15	Seguro	0.020695453
16	Auto	0.001393343
17	Tpo_Univ	0.131690553

Cuadro 4.6: Valor de Información de 17 características.

Después de hacer los grupos de atributos se calcula su WOE respectivo que son los valores de las variables que su utilizarán en la regresión logística.

En la base de datos encontramos datos atípicos (*outlayer*) o datos erróneos que se consideran como un atributo más, al igual que los valores NULL, blanco; esto es, se agrupan en base a su *Bad Rate* como ya mencionamos en el capítulo anterior.

4.5. Regresión logística sobre una muestra de entrenamiento

Con el número de atributos ya reducido de las diez variables que se eligieron se realiza la regresión logística. La variable dependiente es $y_i = 0$ si el cliente i se clasificó como malo y $y_i = 1$ si el cliente se clasificó como bueno. Las variables explicativas son los WOEs de las características no excluidas esta información se encuentra en la tabla 4.9.

Edad	Malos	Buenos	Total	% Total	% Malos	Bueno/Malo
18	1120	15703	16823	22.0	6.7	14:1
19	1220	16653	17873	23.3	6.8	13.7: 1
20	921	13337	14258	18.6	6.5	14.5:1
21	705	10262	10967	14.3	6.4	14.6:1
22	465	7158	7623	10.0	6.1	15.4:1
23	298	4484	4782	6.2	6.2	15.0:1
24	178	2449	2627	3.4	6.8	13.8:1
25	96	1501	1597	2.1	6.0	15.6:1
26	2	2	4	0.0	50.0	1:1
29		1	1	0.0	0.0	
33		1	1	0.0	0.0	
34		1	1	0.0	0.0	
Total	5005	71552	76557	100	6.5	14.3:1

Cuadro 4.7: Característica númerica "Edad" y sus valores.

Edad	Malos	Buenos	Total	% Malos
18-19, 24	2518	34805	37323	6.7
20-21	1626	23599	25225	6.4
22-23, $\geq$ 25	861	13148	14009	6.1
Total	5005	71552	76557	6.5

Cuadro 4.8: Discretización de la característica "Edad". Se agruparon de acuerdo al *Bad Rate*.

Es una práctica común partir la muestra en dos submuestras, la primera es conocida como muestra de entrenamiento y se utiliza para estimar el modelo, la segunda se conoce como muestra de validación, en este trabajo se considera un 80 % para la muestra de entrenamiento y un 20 % para la muestra de validación, porque son los porcentages utilizados en Siddiqi (2006). La partición de la muestra en dos se hizo aleatoriamente utilizando los recursos de Excel.

Con la muestra de entrenamiento se calcularon los estimadores de los parámetros β del modelo de regresión logística. La ecuación del modelo ajustado para $p_i = P(y_i = 1|\mathbf{w}_i)$ donde $\mathbf{w}_i = (woe_{i1}, woe_{i2}, \ldots, woe_{i10})$ es

$$P(y_i = 1|\mathbf{w}_i) = \frac{e^{\beta_0 + \boldsymbol{\beta}_1^T \mathbf{w}_i}}{1 + e^{\beta_0 + \boldsymbol{\beta}_1^T \mathbf{w}_i}} \tag{4.1}$$

La salida (cuadro 4.10) muestra los resultados de ajustar el modelo de regresión

Característica	Atributos	WOE
Empleo	Empleo Padres	0.0733210
	Otros	-0.2588296
Tpo_Prop	Con Padres	0.0767066
	Pagandola, NULL, Familiares	-0.2752925
	Propia, Rentada, Hipotecada, Blanco	-0.3359534
Ant_Domi	0 a 4 años	-0.2398026
	5 a 13 años	-0.0005565
	14 a 21 años	0.1072457
	Mayor a 22 años	0.2193569
Edo_Dom	Blanco, Bc, Chi, Chs, Zac,Ags, Qro, Col, Nay, Slp	0.0373762
	DF, Qroo, Yuc, NL, Mor, Mich, Cam, Tlax, Mex	0.0034292
	Dgo, Chih, Hgo, Chis, Ver,Coah, NULL.	0.0163755
	Pue, Sin, Tab	0.0201072
Teléfono	0	0.7293345
	1	-0.0323236
Ingre_Men	Mayor a $4500 y menor igual a $10500	0.1482371
	Menor igual a $4500, mayor a $10500 y menor igual $16500	-0.0124525
	Mayor a $16500 y menor a $25500	-0.2003873
	Mayor a $25500 y Otros	-0.3287890
Ant_Empleo	1 a 4	-0.2634768
	0 años	-0.1738270
	5 a 7 años	-0.0449420
	Mayor igual a 8 años, vacío.	0.1469863
Dis_Pobla	A) Menos de 15 min, E) Mas de 2 horas	-0.2117328
	B) Entre 15 y 30 min	-0.0407594
	C) Entre 30 y 60 min	0.1754699
	D) Entre 1 y 2 hrs, NULL	0.2531586
Seguro	NULL, Escolar, Auto, Casa	0.1864414
	Gastos Médicos Mayores, No, Otros, Vida	-0.1111941
Tpo_Uni	NULL, Pública	0.2924024
	Privada con colegiatura mayor a $5000, de $3001 a $5000	-0.1451804
	Privada de $1500 a $3000, Privada menor $1500	-0.5216746

Cuadro 4.9: Características selecionadas y su WOE.

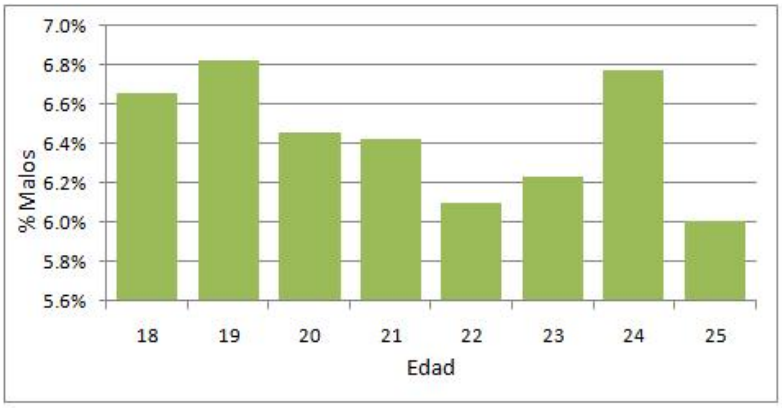


Figura 4.4: *Bad Rate* para los valores de la característica "edad".

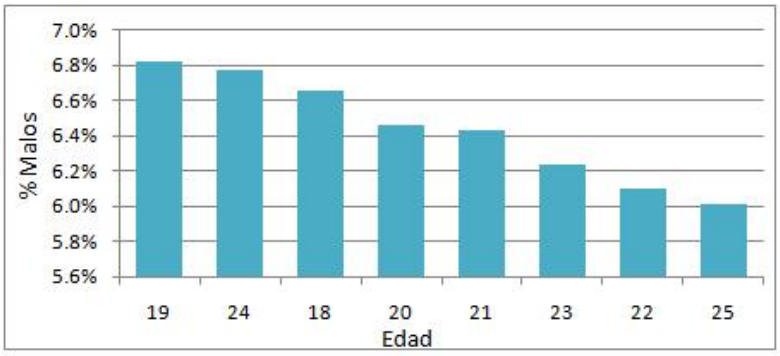


Figura 4.5: Edad ordenada conforme a su *Bad Rate*.

logística. para describir la relación entre la variable dependiente y y las diez variables independientes $\mathbf{w}_i$, la ecuación estimada es

$$\begin{aligned}\log \frac{\hat{p}_i}{1-\hat{p}_i} &= 3.05002 + 0.414015Empleo - 0.455576Tpo\_Prop \\ &\quad 0.376746Ant\_Domi - 0.951115Edo\_Dom - \\ &\quad 0.549708Telefono - 0.824063Ingre\_Men + \\ &\quad 0.990193Ant\_Empleo + 0.61051Dis\_Pobla - \\ &\quad 0.286074Seguro - 0.866912Tpo\_Uni.\end{aligned}$$

En el cuadro 4.10 se muestran los resultados que se obtienen al hacer la regresión

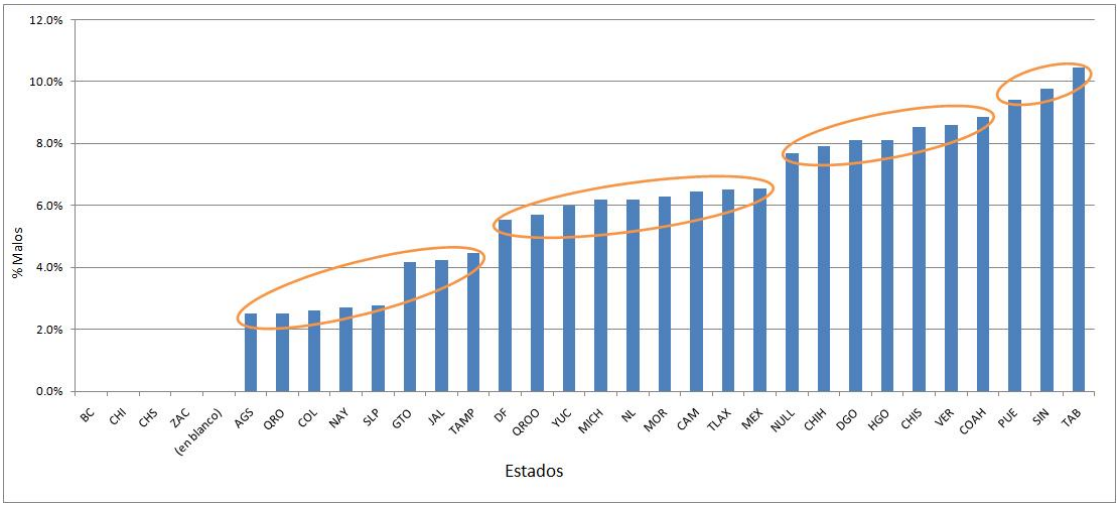


Figura 4.6: "Estados de residencia"

logística. El modelo obtenido determina un clasificador para estimar el valor de y tal que $\hat{y} = \hat{p}$.

Para determinar si el modelo de regresión logística predice bien se analiza el *p-valor*. En la tabla de análisis de varianza el *p-valor* es inferior a 0.01 por lo que se puede concluir que existe una relación funcional de los clientes buenos y malos explicada por el modelo ajustado. El *p-valor* correspondiente a cada característica es pequeño, excepto el correspondiente a la característica "Seguro" que es igual a 0.01670; sin embargo, aun este valor es significativo si se considera un nivel de significancia igual a 0.05. Dependiendo del nivel de confianza que se desee utilizar, podemos o no excluir la característica "Seguro" del modelo. Para nuestro modelo decidimos conservar las diez variables ya descritas.

4.6. Scorecard

Para tener una mejor lectura del score asociado a cada cliente, se procede a reescalar los valores del modelo logístico.

$$Score = 400 + \frac{80}{\ln(2)} \ln(odds)$$

donde

$$\log(odds) = \beta_0 + \Sigma_i^n \beta_i woe_{ij}.$$

Variables	Coeficiente	Error Est. Estimado	Ratio de Prob.	Chi-Cudrado	Gl	P-Value
Constante	3.05002	0.089221				
Empleo	0.41402	0.132438	1.51288	9.67748	1	0.00190
Tpo_Prop	0.45558	0.115936	1.57708	15.1799	1	0.00010
Ant_Dom	0.37675	0.121545	1.45753	9.54167	1	0.00200
Edo_Dom	0.95112	0.059625	2.58859	267.794	1	0.00000
Telefono	-0.54971	0.124781	0.57712	21.7482	1	0.00000
Ingre_Men	0.82406	0.105835	2.27974	59.4863	1	0.00000
Ant_Empleo	0.99019	0.105772	2.69175	85.2159	1	0.00000
Dis_Pobla	0.61051	0.105642	1.84137	33.4956	1	0.00000
Seguro	0.28607	0.119893	1.33119	5.72807	1	0.01670
Tpo_Uni	0.86691	0.0449287	2.37955	360.265	1	0.00000

Cuadro 4.10: Resultados de la regresión logística.

El score queda determinado con la ecuación

$$Score = Offset + Factor \cdot \log(\text{odds}),$$

los términos *Offset* (factor de compensación) y Factor deben satisfacer la condición que en la relación de buenos y malos de 1:1, $Ofsset = 400$ y $Factor$ es tal que el $Odds$ se duplica cada 80 puntos. De esta manera $Offset = 400$ y $Factor = 80/\ln(2)$ quedando la ecuación para el score.

Finalmente se puede asignar un score para el atributo j de la característica i como el sumando correspondiente del score general, esto es:

$$score_{ij} = \frac{80}{\ln(2)}\hat{\beta}_i woe_{ij}.$$

Los valores del *score* ij se muestran en la tabla 4.11. Así a cada cliente se le asigna la suma de los $score_{ij}$ correspondientes más el término constante dado por

$$Constante = 400 + 80/\ln(2)\hat{\beta}_0$$

Los puntajes son redondeados. De esta manera, el score mínimo que se puede obtener es de 486 y el máximo de 920. El cuadro 4.11 muestra la *scorecard* obtenida.

4.6.1. Validación de la Scorecard

Para hacer la validación del modelo de clasificación de buenos y malos se calcula la distribución de los clientes de acuerdo al score asignado. Para ello consideramos

Característica	**Atributos**	**Score**
Constante		752
Empleo	Empleo Padres	4
	Anterior, Otros, Conyugue, Actual, NULL	-12
Tpo_Prop	Con Padres	4
	Pagandola, NULL, Familiares	-14
	Propia, Rentada, Hipotecada, Blanco	-18
Ant_Domi	0 - 4 años	-10
	5 - 13 años	0
	14 - 21 años	5
	mayor de 22 años	10
Edo_Dom	Blanco, Bc, Chi, Chs, Zac,Ags, Qro, Col, Nay, Slp	64
	DF, Qroo, Yuc, NL, Mor, Mich, Cam, Tlax, Mex	9
	Dgo, Chih, Hgo, Chis, Ver, Coah, NULL	-30
	Pue, Sin, Tab	-47
Teléfono	No tiene (0)	-46
	Si tiene (1)	2
Ingre_Men	Mayor a $4500 y menor igual a $ 10500	14
	Menor igual a $4500, mayor a $ 10500 y menor igual a $16500	-1
	Mayor a $16500 y menor a $25500	-19
	Mayor $25500 y Otros	-31
Ant_Empleo	1 a 4	-30
	0 años	-20
	5 a 7 años	-5
	Mayor igual a 8 años, vacío	17
Dis_Pobla	Menos de 15 min, Mas de 2 horas	-15
	Entre 15 y 30 min	-3
	Entre 30 y 60 min	12
	Entre 1 y 2 hrs, NULL	18
Seguro	NULL, Escolar, Auto, Casa	6
	Gastos Médicos Mayores, No, Otros, Vida	-4
Tpo_Uni	NULL, Pública	29
	Privada colegiatura mayor a $5000, Privada de $3001 a $5000	-15
	Privada de $1500 a $3000, Privada menor a $1500	-52

Cuadro 4.11: *Scorecard* para el segmento "Universitarios".

Score	Malos	Buenos	Total	% M	% B	% AM	% AB	B/M
502 - 525	4	10	14	0.1	0.0	0	0	2.5
526 - 550	26	117	143	0.7	0.2	1	0	4.5
551 - 575	84	434	518	2.1	0.8	3	1	5.2
576 - 600	182	990	1172	4.6	1.7	7	3	5.4
601 - 625	308	2102	2410	7.7	3.7	15	6	6.8
626 - 650	486	3652	4138	12.2	6.4	27	13	7.5
651 - 675	564	5325	5889	14.2	9.3	42	22	9.4
676 - 700	581	6861	7442	14.6	12.0	56	34	11.8
701 - 725	563	8246	8809	14.2	14.4	70	48	14.6
726 - 750	464	8286	8750	11.7	14.5	82	63	17.9
751 - 775	349	8165	8514	8.8	14.3	91	77	23.4
776 - 800	236	7029	7265	5.9	12.3	97	89	29.8
801 - 825	89	3769	3858	2.2	6.6	99	96	42.3
826 - 850	28	1655	1683	0.7	2.9	100	99	59.1
851 - 875	12	542	554	0.3	0.9	100	100	45.2
876 - +	1	86	87	0.0	0.2	100	100	86.0
Total	3977	57269	61246	100.0	100.0			14.4

Cuadro 4.12: Distribución de la *scorecard* con la muestra de desarrollo.

intervalos de longitud igual a 25 puntos de score para obtener un número considerable de clases (ver cuadro 4.12).

En la tabla 4.12 se registra la distribución de buenos y malos, la distribución acumulada de cuentas buenas (%AB) y malas (%AM) en porcentajes según puntaje de score para la muestra de validación. La figura 4.7 muestra la distribución de Buenos y Malos.

La grafica de la figura 4.8 registra un *Gini*= 30.62 y la grafica de la figura 4.9 un *K-S* = 22.16 para la muestra de desarrollo de 61246 cuentas. Este mismo proceso se hizo para la muestra de validación con 15311 cuentas.

Los indicadores obtenidos para la muestra de desarrollo y de validación se registran en la siguiente tabla.

Indicador	*Desarrollo*	*Validación*
Gini	30.62	31.75
K-S	22.16	23.01

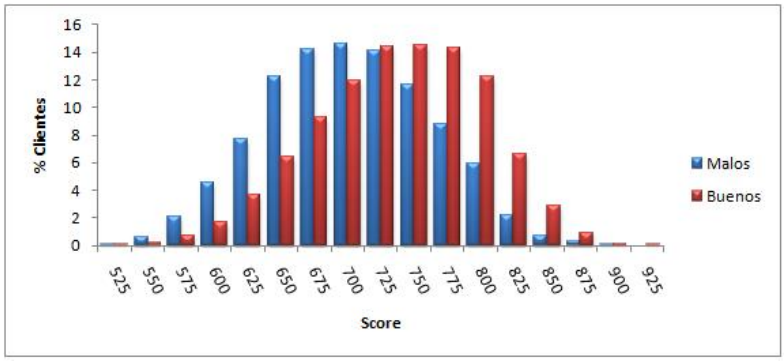


Figura 4.7: Dsitribución de Buenos y Malos para la muestra de desarrollo

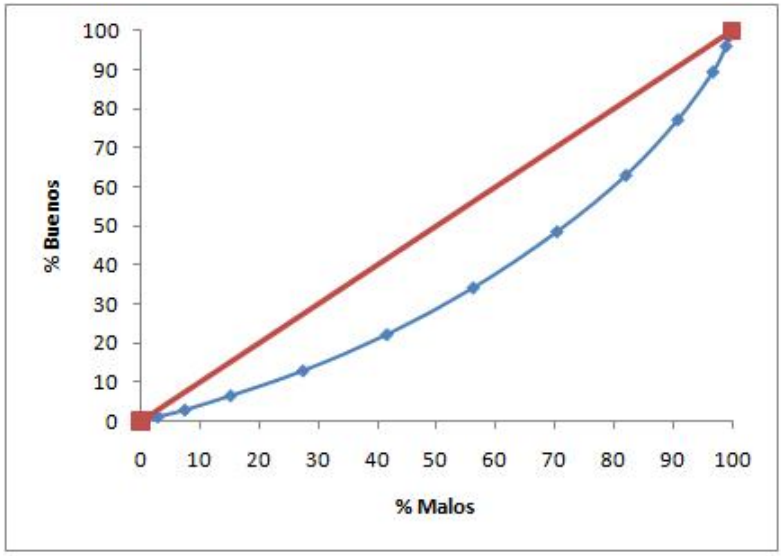


Figura 4.8: Gini = 30.6 para la muestra de desarrollo

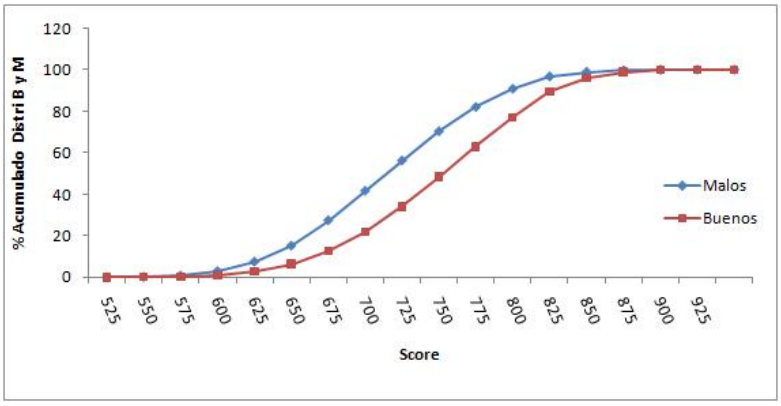


Figura 4.9: K-S = 22.16 para la muestra de desarrollo.

La empresa consultora que hizo los análisis a la empresa de crédito donde se generaron los datos reporta resultados de estos índices para el modelo de desarrollo del segmento Universitarios con una misma calibración en el puntaje de score

Indicador	*Desarrollo*
Gini	30.04
K-S	22.34

Se puede ver que nuestros resultados son bastante cercanos a los reportados por esta empresa consultora lo que indica que estamos bien a pesar de que desconocemos los pasos que ellos siguieron por ser tratados confidencialmente.

4.6.2. Determinación del punto de corte o *Cut Off*

El *Cut Off* es un valor de score por medio del cual la empresa crediticia va a determinar si se le otorga o no el crédito al cliente solicitante, si el score del cliente es mayor o igual al *Cut Off* el crédito es aprobado, si es menor el crédito es rechazado.

Para determinar el punto de corte *Cut Off* hacemos un análisis con todos los clientes aceptados, buenos, malos e indeterminados. A todos los clientes aceptados le calculamos su score. Después se hace una tabla que contiene la información de cuantos malos hay por arriba de un *score* dado.

Se procuro que los valores del score asignando a cada renglón de la tabla contengan aproximadamente el mismo número de datos de la muestra, la tabla 4.13 presenta el acumulado por arriba del número de clientes con un score mayor o igual

	Acumulados			
Score Final	Malos Aceptados	Total Aceptado	% Mora Aceptado	% Rechazo
619	5005	81029	6.2	0.0
644	4345	76024	5.7	6.2
662	3791	71043	5.3	12.3
677	3299	65814	5.0	18.8
690	2866	60731	4.7	25.1
701	2473	55812	4.4	31.1
712	2164	51081	4.2	37.0
723	1854	46084	4.0	43.1
734	1524	40794	3.7	49.7
744	1267	35736	3.5	55.9
756	1004	30823	3.3	62.0
767	789	25963	3.0	68.0
769	585	20945	2.8	74.2
790	401	15995	2.5	80.3
806	241	10604	2.3	86.9
High	100	5319	1.9	93.4

Cuadro 4.13: Porcentajes de mora y rechazo para diferentes puntos de corte.

que el correspondiente al renglón determinado. Se elige el *cut off* de acuerdo al máximo número de malos que se desea aprobar para darle un crédito.

Por ejemplo en la tabla 4.13, en la primera columna y primer renglón se tiene el número de clientes malos que tienen un score mayor o igual a 619, en la primera columna segundo renglón se tiene el número de clientes malos con un score mayor o igual a 644 y así para los demás elementos de la primer columna. La tercera columna indica el porcentaje correspondiente de la primera columna, si por ejemplo elegimos como punto de corte 677 implicaría que aceptamos como valor adecuado el 5 % de clientes que caerán en mora y un 18 % de clientes rechazados.

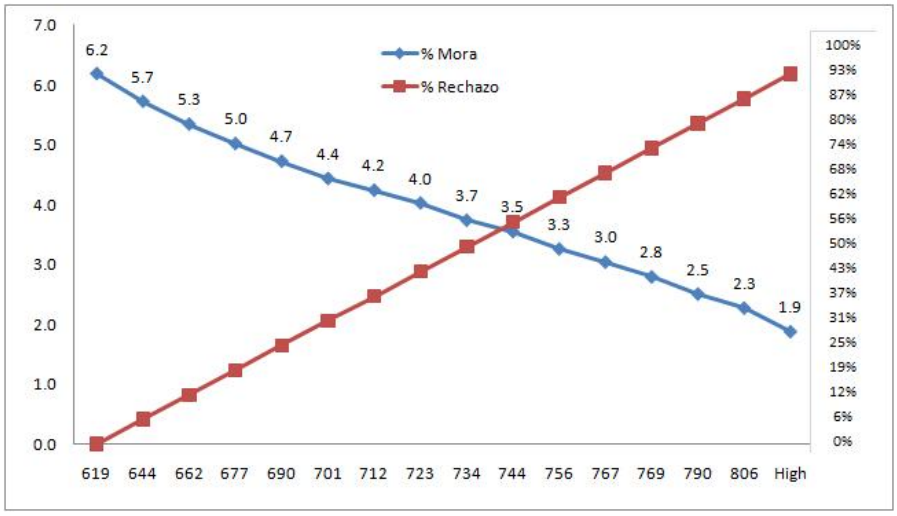


Figura 4.10: Porcentaje de mora (eje izquierdo) y porcentaje de rechazos correspondiente (eje derecho)

La figura 4.10 representa las gráficas de la proporción de mora y la proporción de rechazados correspondiente para un score dado.

4.7. Scorecard sobre un segmento de Universitarios

De la base que hemos venido manejando tomamos un segmento de la población de estudiantes universitarios que no tienen empleo propio, esto es, el empleo registrado es de sus padres u otros. La base que se obtiene consta de 58,530 cuentas buenas y 3,821 cuentas malas. De aquí se toma una muestra de 49,873 registros

para la muestra de entrenamiento. La *scorecard* obtenida se muestra en el cuadro 4.14. Estos puntajes se usan para calcular el *score* de un cliente para decidir si se le otorga o no el crédito solicitado.

Característica	**Atributos**	**Score**
Constante		715
Sexo	Femenino, Vacío.	1
	Masculino	-1
Tpo_Prop	Con Padres	3
	Vacío, Con Familia, Familiares, Hipotecada, Pagandola, Propia, Rentada.	-12
Ant_Domi	≤ 3	-18
	4 - 14	1
	≥ 15	8
Edo_Dom	Chih, Chis, Coah, Hgo, Pue, Sin, Tab, Ver.	60
	Ags, BC, Cam, Chi, Chs, Coa, DF, Dgo, Gro, Mex, Mich, Mor, Nay, NL, Oax, Qr, Qro, QRoo, Slp, Son, Tamp, Tlax, Yuc, Zac.	11
	Col, Gto, Jal	-35
Ant_Empleo	≤ 4	-29
	5 - 14	-4
	≥ 15	14
Tpo_Empleo	NULL, Comisionista, Empleado, Honorarios, Jubilado, Propio sin Empleado.	4
	Propio $<$ 3 Empleados, Propio $>$ 3 Empleados.	-27
Tpo_Uni	Privada $>$ 5000, Privada 3001 a 5000, Privada 1500 a 3000, Privada $<$ 1500.	-51
	Pública, NULL.	35

Cuadro 4.14: *Scorecard* para "Universitarios sin empleo"

Finalmente se realizó la validación de la *scorecard* al calcular el índice de Gini y la estadística *K-S*. Los resultados se muestran en la siguiente tabla

Indicador	*Desarrollo*	*Validación*
Gini	28.53	28.51
K-S	21.65	19.99

Es interesante ver que los resultados obtenidos son muy cercanos a los reportados por la empresa consultora que se presentan en la siguietne tabla:

Indicador	*Desarrollo*	*Validación*
Gini	29.66	26.95
K-S	22.57	22.2

El modelo debe ser monitoreado por la institución crediticia para asegurar un funcionamiento óptimo y hacer los ajustes de acuerdo a los objetivos de la empresa, se considera que el monitoreo debe hacerse cada tres meses como mínimo para verificar la capacidad de predicción del mismo.

4.8. Conclusiones

Generar una *scorecard* es una combinación de ciencia y arte (Simbaqueba, 2004); este es un trabajo que debe ser realizado en equipo de todos los implicados en el proceso de generación e implementación.

La realización de esta tesis, no se hizo en equipo, pero si se utilizó resultados obtenidos por muchas personas, las cuales aparecen en la bibliografía. De todos los capítulos que conforman la tesis, el más interesante es el cinco, pues en el se presenta un problema práctico, con datos reales, correspondientes al sector universitario de los clientes de una empresa crediticia, y en su desarrollo se aplicaron los resultados estudiados en los capítulos anteriores. Entre las particularidades que se revisaron están:

- El primer paso en el proceso del *credit scoring* es la depuración y preparación de la base de datos. La limpieza de la base de datos es un proceso largo y tedioso que se aligera con el uso de software adecuado para ello, en este caso nos auxiliamos de EXCEL y ACCESS de Microsoft.

- Resulto conveniente utilizar siete meses de contratación para obtener una muestra de mayor tamaño, ya que la muestra se reduce en el proceso de limpieza, en la selección de buenos y malos porque se quitan las cuentas que se clasificaron como indeterminadas y se quita también la muestra de validación.

- De acuerdo a los resultados obtenidos al estimar la matriz de transición, se decidió que los clientes buenos son aquellos que al final de los seis meses estaban al día en sus pagos y como máximo tuvieron dos pagos vencidos durante el periodo de seis meses. También son buenos clientes los que al final de los seis meses tienen un pago vencido y durante los seis meses tuvieron como máximo un pago vencido, este resultado es acorde con lo que se esperaba, pues los buenos clientes deben ser los que pagan y no entran en mora muchas veces. Los malos clientes son los que tienen tres o más pagos vencidos al final de los seis meses.

- Se puede utilizar las propiedad de las matrices de transición para discriminar a los buenos clientes de los malos clientes. Si se obtiene las potencias de la matriz de transición, se puede estimar la probabilidad que de cualquier estado se caiga en cartera vencida en dos, tres o más pasos, la decisión de cuales pueden ser buenos o malos clientes se haría de manera semejante a lo realizado en esta tesis, esto también queda para trabajos futuros.

- En general, consideramos que los resultados obtenidos son adecuados para ser utilizados en la práctica por una empresa crediticia, no es el procedimiento óptimo, pero es adecuado de acuerdo a los resultados obtenidos y se tiene una mejor posición para intentar encontrar un procedimiento óptimo.

Apéndice A

Modelo de regresión logística estimado (máxima probabilidad) para el segmento universitarios utilizando STATGRAPHICS

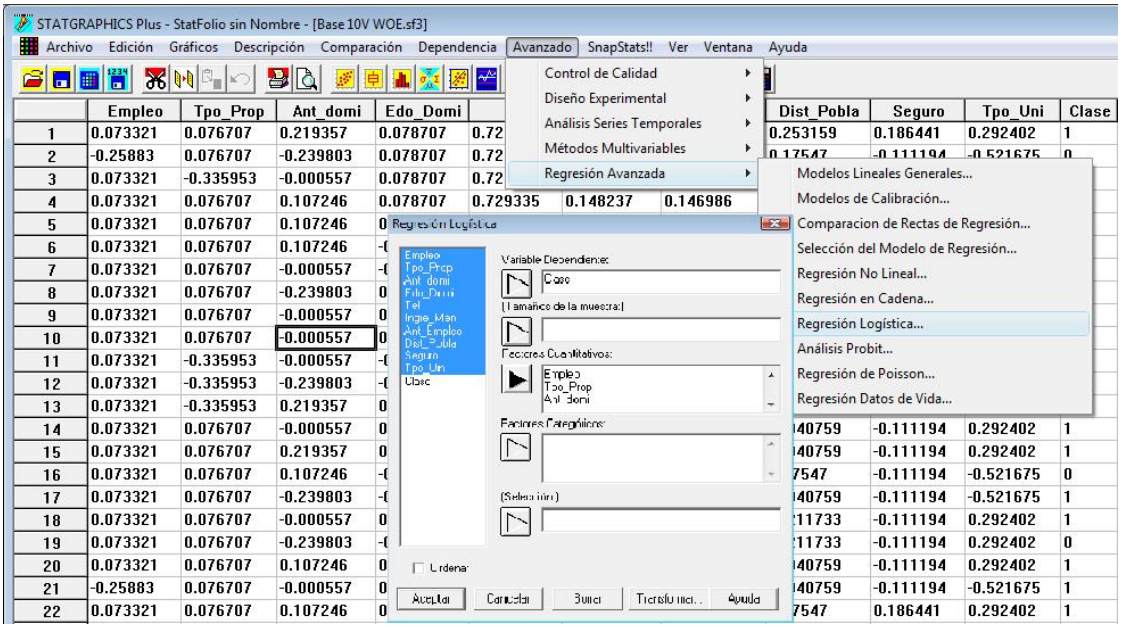


Figura A.1: Ventanas de acceso a Statgraphics para la regresión logística.

	Parámetro Estimado	Error Estándar Estimado	Ratio de Probabilidad
CONSTANTE	3.05002	0.0892196	
Empleo	0.414015	0.132438	1.51288
Tpo_Prop	0.455576	0.115936	1.57708
Ant_domi	0.376746	0.121545	1.45753
Edo_Domi	0.951115	0.0596249	2.58859
Tel	-0.549708	0.124781	0.577119
Ingre_Men	0.824063	0.105835	2.27974
Ant_Empleo	0.990193	0.105772	2.69175
Dist_Pobla	0.61051	0.105642	1.84137
Seguro	0.286074	0.119893	1.33119
Tpo_Uni	0.866912	0.0449287	2.37955

Análisis de Desviación			
Fuente	Desviación	G.L.	P-Valor
Modelo	1113.43	10	0.0000
Residuos	28325.7	61235	1.0000
Total(corr.)	29439.7	61245	

Porcentaje de desviación explicado por el modelo = 3.78213

Porcentaje ajustado = 3.7074

Test de Ratio de Probabilidad			
Factores	Chi-Cuadrado	G.l.	P-Valor
Empleo	9.67748	1	0.0019
Tpo_Prop	15.1799	1	0.0001
Ant_domi	9.54167	1	0.0020
Edo_Domi	267.794	1	0.0000
Tel	21.7482	1	0.0000
Ingre_Men	59.4863	1	0.0000
Ant_Empleo	85.2159	1	0.0000
Dist_Pobla	33.4956	1	0.0000
Seguro	5.72807	1	0.0167
Tpo_Uni	360.265	1	0.0000

Análisis de Residuos	
Estimación	Validación
n	61246
MSE	0.00436846
MAE	0.146796
ME	0.000385067

El StatAdvisor

La salida muestra el resultado de ajustar un modelo de regresión logístico para describir la relación entre Clase y 10 variable(s) independientes. La ecuación del modelo ajustado es

$$\texttt{Clase} = \texttt{exp(eta)/(1+exp(eta))}$$

donde

$$\begin{aligned}\texttt{eta} &= 3.05002 + 0.414015 * \texttt{Empleo} + 0.455576 * \texttt{Tpo\_Prop} + \\ &+ 0.376746 * \texttt{Ant\_domi} + 0.951115 * \texttt{Edo\_Domi} - 0.549708 * \texttt{Tel} \\ &+ 0.824063 * \texttt{Ingre\_Men} + 0.990193 * \texttt{Ant\_Empleo} + 0.61051 * \texttt{Dist\_Pobla} \\ &+ 0.286074 * \texttt{Seguro} + 0.866912 * \texttt{Tpo\_Uni}.\end{aligned}$$

Dado que el p-valor para el modelo en la tabla del Análisis de la Varianza es inferior a 0.01, hay una relación estadísticamente significativa entre las variables al 99 % de nivel de confianza. Además, el p-valor para los residuos es mayor o igual a 0.10, indicando que el modelo no es significativamente peor que el mejor modelo posible para estos datos al 90 % de nivel de confianza o superior. La ventana también muestra que el porcentaje e desviación en Clase explicado por el modelo es igual a 3.78213. El porcentaje ajustado más adecuado para comparar modelos con diferentes números de variables independientes, es 3.7074 %. Determinando si el modelo puede simplificarse, observe que el p-valor más alto para los test de proporción de probabilidad es 0.0167, perteneciendo a Seguro. Dado que el p-valor es inferior a 0.05, ese término es estadísticamente significativo al 95 % de nivel de confianza. Por consiguiente, probablemente no querrá eliminar ninguna variable del modelo.

Bibliografía

[1] Apr, *"Usuarios de tarjetas de crédito, sólo 8 % paga puntual."*. El Economista, D. F., 9 febrero 2009.

http://eleconomista.com.mx/notas-online/finanzas/2009/02/09/usuarios-tarjetas-credito-solo-8-paga-puntual

[2] Barberena Manuel, Barberena Viterboo. *La mineria de Datos en la Industria Financiera: Un nuevo enfoque de Investigación de Mercados*. AMAI, No. 31, Año 9, México, Febrero de 2002.

[3] González Roberto, *"Cada día caen en cartera vencida unos 3 mil 305 préstamos al consumo"*, La jornada, Finanzas, D. F., 11 de diciembre de 2008.

http://www.jornada.unam.mx/2008/12/11/index.php?section-=economia&article=029n1eco

[4] Hillier Frederick S., Lieberman Gerald J. *Introducción a la Investigación de Operaciones*. McGraw-Hill, New York, 1997.

[5] Lino Arturo. *"Baja 2.68 % el crédito al consumo"*. El Sol de México, D.F., 27 de enero de 2009.

http://www.oem.com.mx/elsoldemexico/notas/n1022683.htm

[6] Medina Fernando, *Consideraciones sobre el índice de Gini para medir la concentración del ingreso*. Estudios estadísticos y prospectivos, serie 9. Públicación de las Naciones Unidas. Santiago de Chile, marzo 2001.

[7] Moreno, Tania M. *"Crédito al consumo conquista a México"*. cnnexpansion, D. F., viernes 28 de marzo de 2008.

http://www.cnnexpansion.com/midinero/2008/03/28/credito-al-consumo-2018conquista2019-a-mexico-1

[8] Peña Daniel. *Análisis de datos Multivariados*. McGraw Hill, Madrid, 2002.

[9] Randolph Nelson. *Probability, Stochastics and Queing Theory. The Mathematics of computer Performance Modeling*. Springer-Verlag, New York, 1995.

[10] Siddiqi, Naeem. *Credit Risk Scorecards: developing and implementing intelligent credit scoring*. John Wiley & Sons, New Jersey, 2006.

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

[12] Thomas Lyn, Edelman David and Crook Jonathan. *Credit Scoring and its applications*. SIAM, Philadelphia, 2002.

[13] Thomas P. Ryan. *Modern Regresión Methods*. John Wiley and Sons, Wiley Series in Probability and Statistics, New York, 1997.

[14] Zúñiga Antonio y Rodríguez Israel. *"Creció más de 50 % la cartera vencida del crédito al consumo"*. La jornada, D. F., 27 de enero de 2009.

http://www.jornada.unam.mx/2009/01/27/index.php?section=economia-&article=020n1eco

Crédito al Consumo:
La Estadística Aplicada a un Problema de Riesgo Crediticio

Que presenta:

Soraida Nieto Murillo

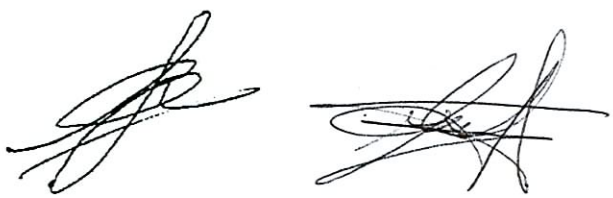
Para obtener el grado de:

Maestra en Ciencias (Matemáticas Aplicadas e Industriales)

Asesores:

Dra. Blanca R. Pérez Salvador

Act. J. Fernando Soriano Flores



20 de Mayo de 2010

Casa abierta al tiempo

UNIVERSIDAD AUTONOMA METROPOLITANA

UNIDAD IZTAPALAPA División de ciencias Básicas e ingenierías