UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

DESARROLLO Y EVALUACIÓN DE METODOLOGÍAS PARA LA APLICACIÓN DE REGRESIONES LOGÍSTICAS EN MODELOS DE COMPORTAMIENTO BAJO SUPUESTO DE INDEPENDENCIA

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

MIGUEL BIRON LATTES

UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

DESARROLLO Y EVALUACIÓN DE METODOLOGÍAS PARA LA APLICACIÓN DE REGRESIONES LOGÍSTICAS EN MODELOS DE COMPORTAMIENTO BAJO SUPUESTO DE INDEPENDENCIA

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

MIGUEL BIRON LATTES

PROFESOR GUÍA:
JOSÉ MIGUEL CRUZ.

MIEMBROS DE LA COMISIÓN:
CRISTIÁN BRAVO.
ROGER LOWICK-RUSSEL

SANTIAGO DE CHILE
ABRIL 2012

RESUMEN DE LA MEMORIA PARA OPTAR AL
TÍTULO DE INGENIERO CIVIL INDUSTRIAL
POR: MIGUEL IGNACIO BIRON LATTES
FECHA: 17/04/2012
PROF. GUIA: JOSÉ MIGUEL CRUZ

DESARROLLO Y EVALUACIÓN DE METODOLOGÍAS PARA LA APLICACIÓN DE REGRESIONES LOGÍSTICAS EN MODELOS DE COMPORTAMIENTO BAJO SUPUESTO DE INDEPENDENCIA

El presente documento tiene por objetivo desarrollar y evaluar una metodología de construcción de regresiones logísticas para *scorings* de comportamiento, que se haga cargo del supuesto de independencia de las observaciones inherente al método de estimación de máxima verosimilitud.

Las regresiones logísticas, debido a su facilidad de interpretación y a sus buen desempeño, son ampliamente utilizadas para la estimación de modelos de probabilidad de incumplimiento en la industria financiera, los que a su vez sirven múltiples objetivos: desde la originación de créditos, pasando por la provisión de deuda, hasta la pre aprobación de créditos y cupos de líneas y tarjetas. Es por esta amplia utilización que se considera necesario estudiar si el no cumplimiento de supuestos teóricos de construcción puede afectar la calidad de los *scorings* creados.

Se generaron cuatro mecanismos de selección de datos que aseguran la independencia de observaciones para ser comparados contra el método que utiliza todas las observaciones de los clientes (algoritmo base), los que posteriormente fueron implementados en una base de datos de una cartera de consumo de una institución financiera, en el marco de la metodología KDD de minería de datos.

Los resultados muestran que los modelos implementados tienen un buen poder de discriminación, llegando a superar el 74% de KS en la base de validación. Sin embargo, ninguno de los métodos propuestos logra superar el desempeño del algoritmo base, lo que posiblemente se debe a que los métodos de selección de datos reducen la disponibilidad de observaciones para el entrenamiento, lo que a su vez disminuye la posibilidad de poder construir modelos más complejos (mayor cantidad de variables) que finalmente entreguen un mejor desempeño.

A mi familia, por estar en las malas y en las más malas

A los amigos verdaderos, con los que se puede contar siempre

A los compañeros de CL Group, por sus consejos, apoyo y buena onda

A mis profesores guía y co-guía, por su excelente disposición y paciencia

Como Cerati solía decir, "gracias totales."

Índice general

Índice de figuras

Índice de tablas

Capítulo 1: Introducción

"I sincerely believe, with you, that banking institutions are more dangerous than standing armies..."

-Thomas Jefferson, carta a John Taylor, 1816.

La industria bancaria ha sufrido importantes transformaciones en los últimos años, debido en gran parte a la innovación en productos, marketing, y el manejo del riesgo. Estas transformaciones han sido particularmente importantes en el sector minorista (o *retail banking*). Este segmento es el que típicamente constituye las fuentes más estables de ingreso en la industria bancaria global (10).

La banca minorista atiende tanto a pequeñas empresas, como a consumidores, e incluye las líneas de negocios de depósitos y préstamos. Los créditos *retail* consisten principalmente en:

- Créditos hipotecarios: enfocados a financiar la compra de vivienda o de otros bienes. Están respaldados por propiedades residenciales.
- Créditos en cuotas: incluyen líneas de crédito, tarjetas de crédito, créditos automotrices. Están respaldados por bienes, propiedades y patrimonio en general.
- Créditos *revolving*: no tienen asociado un número fijo de cuotas. No están garantizados.
- Créditos a pequeñas empresas: están respaldados por los activos del negocio y/o por el patrimonio de los dueños.

A todo contrato crediticio se asocian dos conceptos fundamentales: la calidad crediticia del deudor, y el riesgo de crédito. El primero se refiere a la capacidad y disposición de la contraparte a cumplir sus obligaciones de pago (1). Investigaciones recientes se han enfocado en modelar la relación entre la disposición a pagar de los clientes y las características propias de los créditos solicitados (8). Por su parte, el riesgo de crédito se define como el impacto económico que potencialmente podría tener para la institución financiera, cualquier cambio en la calidad crediticia de un

deudor (1). Una definición alternativa pero similar, propone que el riesgo de crédito es el que se desprende de la posibilidad de que un cliente incurra en un incumplimiento, lo cual podría resultar en una importante pérdida para el banco (3). El Comité de Basilea de Supervisión Bancaria (BCBS, por sus siglas en inglés) es un foro internacional que provee cooperación en temas de supervisión bancaria. Este comité ha señalado que la exposición al riesgo de crédito es, por lejos, la más importante para los bancos a nivel mundial, por lo que las instituciones financieras debiesen identificar, medir y controlar de manera especial esta fuente de riesgo (4). Otros autores también confirman la importancia del riesgo de crédito sobre otras fuentes (3).

Ahora bien, el riesgo de crédito generado por la banca minorista es significativo, pero posee una dinámica muy diferente al que se deriva de las líneas de negocio orientadas a grandes empresas. La característica fundamental de los préstamos minoristas es que estos se presentan en exposiciones individuales bajas, por lo que el incumplimiento de un deudor nunca es suficiente para amenazar la estabilidad de un banco. Otra característica importante es que los clientes minoristas tienden a ser independientes financieramente de otros clientes de su mismo tamaño. Los puntos anteriores implican que la estimación de las pérdidas que pueden hacer los bancos en este tipo de carteras es considerablemente mejor. Y la herramienta que, por lejos, ha concentrado la mayor atención tanto de la industria como de la academia para medir riesgo de crédito en carteras *retail* corresponde al *credit scoring* (10).

1.1. Definición de c*redit scoring*

El *credit scoring* es una técnica de gestión del riesgo de crédito que analiza el riesgo de un deudor (3). Otro autor lo define como el conjunto de modelos de decisión, y sus técnicas subyacentes, que asisten a las instituciones bancarias en la entrega de crédito (31). Por otro lado, (1) lo define como el uso de modelos estadísticos para transformar datos relevantes en medidas numéricas que guíen las decisiones de crédito.

Es usual exigir las siguientes características a un *scoring* de riesgo (11):

1. Exactitud: bajas tasas de error.
2. Parsimonia: uso de un número reducido de variables explicativas.
3. No-trivial: que produzca resultados interesantes.
4. Factible: que se calcule en una cantidad razonable de tiempo.
5. Transparencia e interpretabilidad: que provea una capacidad de entendimiento en las relaciones y tendencias ocultas en los datos.

En particular, un *scoring* debe ser altamente discriminante (3); esto significa que debe asignar puntajes bajos a deudores poco riesgosos, y puntajes altos a deudores más riesgosos (o viceversa, depende de la convención de signo). Se ha mostrado que *credit scorings* con mejor poder discriminación pueden llevar a desempeños económicos significativamente mejores (6). En la misma línea, (31) muestra las diversas aplicaciones y beneficios que los *scorings* de crédito pueden ofrecer para las instituciones financieras. Algunas de ellas son:

- Preselección de clientes
- Pre aprobación de créditos
- Detección de fraude
- *Pricing* basado en riesgo
- Manejo de límites de crédito y autorización de transacciones
- *Scoring* de cobranza y litigación
- Cálculo de provisiones por riesgo de crédito

Otros beneficios asociados a los modelos de predicción de pérdida en general son (32):

- Analizar el pasado
    - Estudiar el impacto de campañas de marketing.
    - Estudiar el impacto de políticas crediticias y campañas de cobranza.
    - Diagnosticar tendencias del portafolio y sus desviaciones de ella.
- Predecir el futuro
    - Predecir la morosidad y castigos de nuevos clientes.
    - Pronosticar las pérdidas agregadas del portafolio.

    - Medir el impacto de cambios en los ambientes económico e industrial.
- Planificación
    - Cálculo de provisiones por pérdida económica.
    - Evaluar el impacto de nuevas campañas de marketing y cobranza.
- Cumplimiento de requerimientos regulatorios
    - Basilea II.

Existen, fundamentalmente, dos tipos de *credit scoring*: los que se utilizan para clientes que nunca han tenido relación con la institución financiera, denominados *application scorings*, y los que se utilizan para clientes con los cuales se cuenta con una historia de comportamiento dentro de la institución, denominados *behavioral scorings*. Los primeros cumplen la función de pre seleccionar clientes, mientras que los segundo pueden ser utilizados en más ámbitos, como por ejemplo en la pre aprobación de créditos, en *scorings* de cobranza, y en el cálculo de provisiones. Por esta mayor gama de aplicación es que en este trabajo se tratará exclusivamente con *behavioral scorings*.

Es necesario también hacer la diferencia entre *credit scoring* y *credit ratings*. Los primeros son populares en carteras de alto volumen de operaciones con bajas colocaciones promedio (*retail*), mientras que los segundos son utilizados para préstamos a grandes compañías, gobiernos y otros, y se basan en un porcentaje mucho más alto de apreciaciones subjetivas (criterio experto) (1).

En este trabajo se dará énfasis a la importancia que tienen los modelos de probabilidad de default para la gestión del riesgo crediticio por parte de las instituciones financieras, particularmente en el cálculo de requerimientos capital, capital económico y de provisiones por riesgo de crédito.

Un marco de trabajo que guía la regulación del riesgo (en particular, el de crédito) en bancos a lo largo de muchos países, es el propuesto por BCBS. Este comité publicó en 2004 el *Basel II Framework*, con el fin de asegurar la convergencia en las regulaciones de los bancos por parte de los supervisores en cada país (5). El marco de trabajo ahí propuesto se basa en tres pilares: requerimientos de capital mínimo,

revisión de los supervisores, y disciplina del mercado. Según (23), la gran diferencia con el documento antecesor de Basilea II, es que éste aumenta considerablemente las posibilidades de los bancos para utilizar sus propias medidas internas de riesgo como insumos para el cálculo de requerimientos de capital.

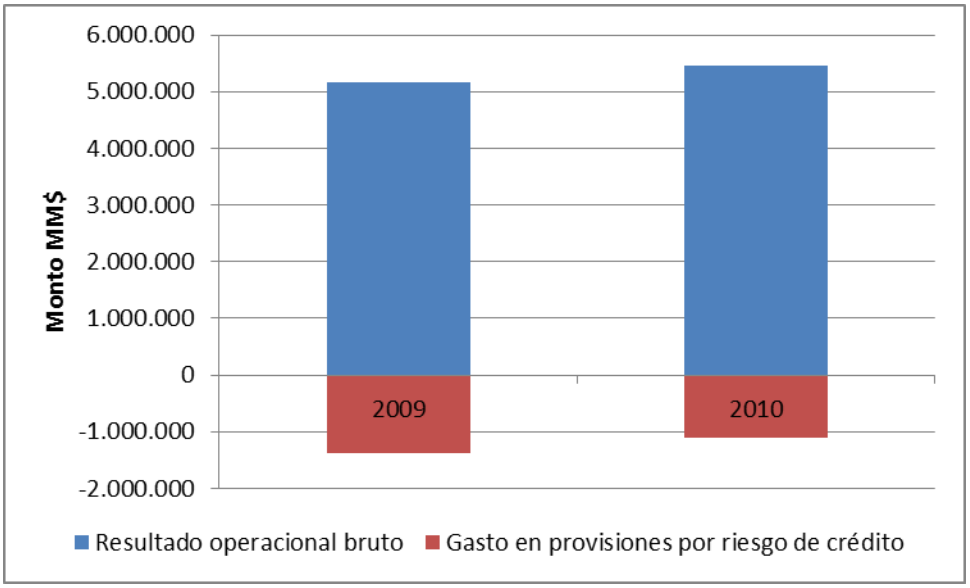
En línea con lo anterior, la Superintendencia de Bancos e Instituciones Financieras (SBIF) es, en Chile, el ente regulador encargado de "supervisar las empresas bancarias y otras instituciones financieras, en resguardo de los depositantes u otros acreedores y del interés público" (25). En su Compendio de Normas Contables, la SBIF propone para la evaluación del riesgo de crédito de carteras grupales (caracterizadas por un alto número de operaciones, bajos montos y mayoritariamente compuestas por personas naturales o empresas pequeñas), un método en el que los bancos segmentan a sus deudores en grupos homogéneos, asociando a cada uno una probabilidad de incumplimiento y un porcentaje de recuperación, "en un análisis histórico fundamentado" (26).

La provisión por riesgo de crédito es una característica de los sistemas de contabilidad y gestión, y de los mecanismos de control impuestos por los bancos centrales y otras entidades regulatorias. Una provisión corresponde a la estimación de los costos esperados que provocarán los incumplimientos de la cartera de un banco, a cierto plazo definido (31) (10). La mayoría de las técnicas de estimación de provisiones se basan tanto en el uso de información histórica, como en resultados de *behavioral scorings* (1). Estos modelos permiten a las instituciones financieras tomar mejores decisiones en la gestión de sus clientes, al proyectar su comportamiento en el futuro (30).

Así como existen marcos regulatorios en los países que controlan a los bancos para que mantengan niveles adecuados de capital y provisiones, existen a la vez incentivos para disminuir estas cantidades, y por lo tanto, de buscar modelos con mejores capacidades predictivas, ya que cualquier desviación excesiva entre lo proyectado y lo realmente ocurrido, implica costos directos e indirectos para el banco.

El gráfico siguiente ilustra el resultado operacional bruto de la industria chilena y su gasto en provisiones para los años 2009 y 2010:

Figura 1. Resultados operacionales brutos y gastos en provisiones de la banca chilena.



Fuente: Elaboración propia en base a datos de la SBIF.

Se aprecia que los gastos en provisiones por riesgo de crédito representan cerca del 20% del resultado operacional bruto. En este sentido, una sobreestimación excesiva de este gasto conlleva una disminución innecesaria en los beneficios percibidos por los accionistas de los bancos.

En resumen, las razones expuestas anteriormente son las que motivan al autor de este trabajo a elaborar una metodología con un fuerte sustento cuantitativo, que permita a analistas de entidades financieras y regulatorias desarrollar modelos que apoyen de manera confiable y robusta la gestión del riesgo de crédito.

1.2. Título de la memoria

"Desarrollo y evaluación de una metodología de muestreo para modelos estadísticos de probabilidad de incumplimiento."

1.3. Objetivos

1.3.1. Objetivo General

Desarrollar y evaluar una metodología de muestreo para modelos estadísticos de probabilidad de incumplimiento, compatible con el supuesto de independencia de las observaciones.

1.3.2. Objetivos Específicos

I. Revisión bibliográfica para establecer el estado del arte en el área.
II. Diseño de metodologías de muestreo.
III. Preparación una base de datos para la evaluación de las metodologías.
IV. Evaluación del desempeño de las metodologías utilizando modelos estadísticos.

1.4. Metodología

La estructura de este trabajo está diseñada a partir de los objetivos específicos anteriormente planteados. Así, la primera etapa de este proyecto se enfoca en realizar un estudio acabado de la literatura científica que guarda relación con estimación estadística de pérdidas por riesgo de crédito, *credit scoring*, y metodologías de muestreo.

A continuación, el trabajo se dirige a diseñar metodologías de muestreo de datos, proponiendo distintos algoritmos que sean aplicables al diseño de modelos de probabilidad de incumplimiento en *scorings* de comportamiento.

Una vez diseñado los algoritmos de muestreo a evaluar, el tercer paso constituye la preparación de una base de datos de comportamiento de una institución financiera, sobre la cual se puedan construir modelos de probabilidad de incumplimiento. La finalidad de esto es poder evaluar los algoritmos diseñados.

Por último, con el conjunto de datos preparados, se procederá a evaluar la metodología diseñada, implementando y validando modelos de *credit scoring* sobre muestras elaboradas usando los algoritmos propuestos.

1.5. Alcances

Los resultados de esta memoria podrán ser utilizados como guía para cualquier analista con la intención de desarrollar modelos de probabilidad de incumplimiento en el ámbito de *scorings* de comportamiento.

La metodología desarrollada en este trabajo se enfoca en estimar probabilidades de incumplimiento por individuo. No es el objetivo de este trabajo incursionar en temas de estimación de pérdidas netas.

La memoria utilizará datos de una cartera de créditos *retail*. Las conclusiones que se desprendan de este trabajo estarán estrictamente asociadas a este hecho y no serán extrapolables a otro tipo de carteras.

Capítulo 2: Marco conceptual

"If I have seen further, it is by standing on the shoulders of giants."

-Isaac Newton, carta a Robert Hooke (1676).

2.1. El riesgo de crédito y su relación con modelos de PD

El riesgo de crédito corresponde a la distribución de las pérdidas financieras debido a cambios inesperados en la calidad crediticia de una contraparte en un contrato financiero (14). Los autores en (7) dividen la medición del riesgo crediticio en dos cantidades: la pérdida esperada y la pérdida no esperada. A continuación se definen ambas medidas, siguiendo a los autores citados anteriormente.

2.1.1. Pérdida esperada.

La exposición que las instituciones financieras tienen hacia el riesgo de no pago de sus clientes, sugieren la necesidad por parte de estas de una protección ante pérdidas en la forma de "seguros". La idea básica de los seguros es transversal a todas las industrias. En salud, por ejemplo, los costos de unos pocos pacientes enfermos son cubiertos por la suma de los ingresos que obtienen las firmas desde todos sus clientes. Así, la tarifa que un hombre de 30 años paga a su compañía de seguros, refleja los costos esperados que la firma estima incurrir con respecto a los clientes de estas características. Para los créditos bancarios, es posible hacer el mismo argumento: si se cobran primas por riesgo adecuadas a cada uno de los clientes del banco, creando una cuenta llamada "pérdidas esperadas", entonces la firma tendrá un colchón que cubrirá las pérdidas originadas por los préstamos que caigan en incumplimiento (7).

Es posible definir por cada deudor una "variable de deuda" como la siguiente:

$$\bar{L} = L \times EAD \times LGD \qquad \text{con} \qquad L = 1_D, \quad \mathbb{P}(D) = PD$$

donde *D* corresponde al evento de incumplimiento, y $\mathbb{P}(D)$ la probabilidad de dicho evento; 1_D se refiere a la función indicadora, la cual es igual a 1 si ocurre el evento D, e igual a 0 en cualquier otro caso; PD es la probabilidad de incumplimiento, EAD es el valor esperado de la exposición al momento del incumplimiento, y LGD corresponde al valor esperado de la fracción no recuperada de la exposición. Dado que el valor esperado de una variable Bernoulli es su probabilidad de ocurrencia, la pérdida esperada será igual a:

$$PE = \mathbb{E}(\bar{L}) = \mathbb{P}(D) \times EAD \times LGD = PD \times EAD \times LGD$$

Como puede apreciarse, la probabilidad de default es un elemento central al momento de intentar estimar la pérdida esperada asociada a un deudor. Más adelante se destinará una sección a la revisión de los enfoques existentes para la estimación de esta cantidad.

2.1.2. Capital económico y la pérdida no esperada.

Mantener un colchón de capital igual al monto de pérdidas esperadas de un portafolio no es suficiente. Los bancos debiesen guardar dinero para cubrir adicionalmente las pérdidas no esperadas que excedan al promedio de las pérdidas experimentadas en el pasado. En este sentido, se define la cantidad "capital económico" (EC) como forma de cuantificar el capital en riesgo. Dado un nivel de confianza α, el capital económico se define como el α-quantil de la distribución de pérdida, menos la pérdida esperada del portafolio (7). En otras palabras,

$$EC_\alpha = q_\alpha - PE_{PF}$$

donde q_α es el α-quantil de la distribución de pérdida del portafolio $\bar{L}_{PF} = \sum_{i=1}^{m} \bar{L}_i$; es decir,

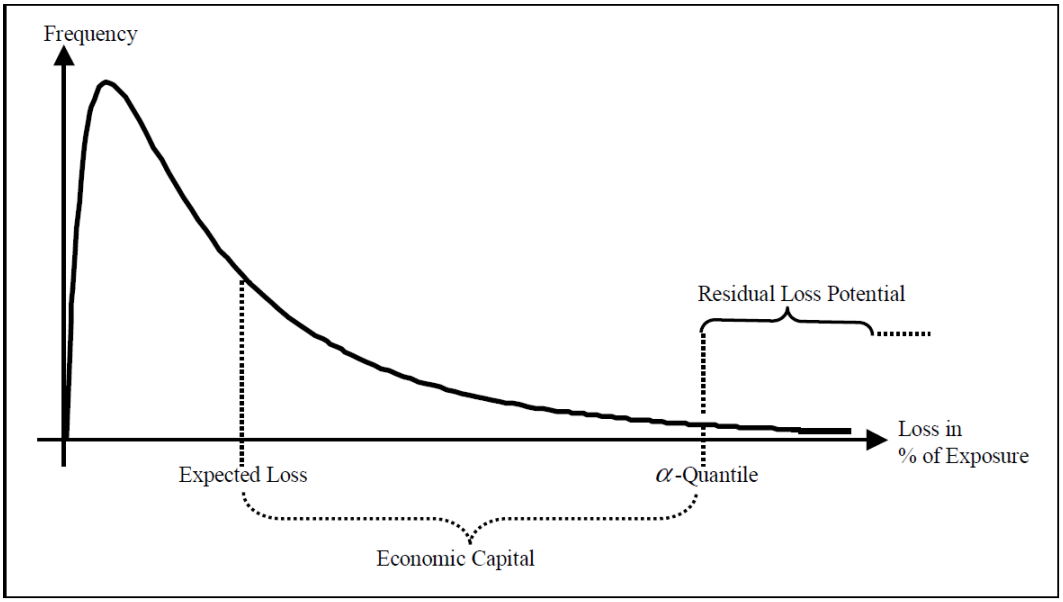
$$q_\alpha = \inf\{q > 0 \mid \mathbb{P}[\bar{L}_{PF} \leq q] \geq \alpha\}$$

Por ejemplo, si el nivel de confianza se fija en $\alpha = 99{,}98\%$, entonces el capital económico será suficiente (en promedio) para cubrir las pérdidas no esperadas en 9.998 de 10.000 años, si se utiliza un horizonte de planificación anual. La desventaja

es que, al igual, uno debiese esperar que en 2 de los 10.000 años se incurra en pérdidas mayores a las que se podrían cubrir con EL y EC. Sin embargo, hoy en día la mayoría de los bancos utiliza esta definición de capital económico.

El gráfico siguiente resume los conceptos mostrados anteriormente:

Figura 2. Distribución de pérdida del portafolio.



Fuente: Extraído desde (7).

El documento de Basilea II (5) propone dos grandes líneas de cálculo del capital económico:

1. *Standardized Approach*: el banco utiliza ponderadores de riesgo dados por el regulador de su país.
2. *Internal Rating-Based Aprroach (IRB)*: el banco utiliza sus propias medidas de riesgo para generar ponderadores de capital:
    a. *Foundation IRB*: el banco utiliza sus propias medidas de Probabilidad de Default (PI).
    b. *Advanced IRB*: el banco utiliza sus propias estimaciones de PI, severidad de pérdida (LGD) y exposición al incumplimiento (EAD).

Para ambos métodos internos, Basilea II propone funcionas que utilizan cómo inputs las medidas de PI, EAD y LGD para el cálculo de ponderadores de capital por conceptos de riesgo crediticio, según la banca en estudio. Estas funciones pueden encontrarse en (5).

2.2. Enfoques de estimación de la probabilidad de default.

Varios métodos alternativos a los de probabilidad de incumplimiento se han utilizado tradicionalmente para estimar las pérdidas de una cartera de créditos, entre los que se encuentran los métodos conocidos como “incondicionales” (32). Estos son:

- Net flow rates
- Curvas vintage de pérdida
- Distribuciones de score

La característica principal de estos tres enfoques es que todos se basan exclusivamente en el comportamiento pasado agregado de la cartera. Por lo tanto, la crítica que usualmente se hace a estos enfoques es que, a diferencia de las metodologías que incorporan información del entorno y de los clientes, no son capaces de capturar cambios en la cartera si es que ésta no se comporta exactamente igual que en el pasado (datos con los que se calibró el método) (32).

Los autores de (11) plantean dos enfoques de estimación que son comúnmente utilizados en la industria financiera. Uno de ellos corresponde a los modelos estructurales de riesgo de crédito, los cuales se basan en la modelación de la probabilidad de incumplimiento de manera causal en función de la información de mercado de las firmas. Estos modelos no serán tratados en profundidad, pues no forman parte de los alcances de este trabajo. Los lectores interesados pueden referirse a (14) o (7). El segundo enfoque corresponde al de los *credit scorings*, los cuales sí serán materia de estudio en esta memoria. La Tabla 1 muestra un resumen con las características de ambas metodologías.

Tabla 1. Comparación de enfoques de medición de riesgo de crédito.

	Modelos estructurales	**Credit scoring**
Captura de señales del mercado.	Sí	No
Robusto frente a "burbujas".	No	Sí
Extensible a empresas no listadas en bolsa	No	Sí

Fuente: elaboración propia.

Si bien los modelos estructurales tienen la ventaja de capturar señales de mercado de manera más directa (evaluación con enfoque *mark-to-market*), son difícilmente aplicables a empresas que no transan en bolsa ni menos a clientes particulares, pues se basan en información del patrimonio que no está disponible; o que si está, puede no ser totalmente fidedigna, debiendo ser "adaptada", con el grado de discrecionalidad que eso implica. En estos casos, por lo tanto, las instituciones financieras recurren a los *credit scoring* para evaluar el riesgo de su cartera (11). Esta es la razón de que este trabajo se enfoque en la evaluación de este tipo de metodologías.

Por su parte, los modelos de *credit scoring* poseen sus propios beneficios y limitaciones. Según (27), algunos de los beneficios de esta metodología son:

- Provee de criterios estándares para predecir el comportamiento de pago del deudor, dado que el modelo provee un análisis objetivo. Se enfoca sólo en información relacionada con el riesgo crediticio y evita la subjetividad de quienes realicen el proceso de evaluación.
- Aumenta la velocidad de análisis de la cartera: permite la automatización del cálculo de provisiones y cuantificación del riesgo asociado a cada crédito vigente.
- Aporta información para gestionar de forma anticipada futuras renegociaciones o acciones que eviten el no pago de algún crédito vigente.

- Está completamente alineado con los nuevos requerimientos de BIS-II, en el sentido que son herramientas que otorgan argumentos estadísticos y robustos para la toma de decisiones y gestión de carteras.

Por otro lado, algunas de las limitaciones del *credit scoring* son las siguientes:

- Se construyen con un sesgo de selección, ya que previamente son rechazados los clientes que se cree no tendrán capacidad de pago.
- Los patrones de comportamiento cambian a través del tiempo.
- La omisión de variables o atributos importantes en el modelo.
- Es necesario que los individuos tengan toda la información utilizada por el modelo de *credit scoring* desarrollado, antes que un puntaje o probabilidad pueda ser calculado.
- Posible sobre utilización cuando se confía en la tecnología y se reduce el uso del juicio experto.

Los modelos de *credit scoring* se han desarrollado en los últimos años bajo el área de la minería de datos (11). Existen tres metodologías importantes dentro de las cuales es posible desarrollar un proyecto de minería de datos[1]. Éstas son KDD, CRISP-DM y SEMMA. La metodología en la que se enmarcará el trabajo de esta memoria se basará en el proceso KDD. Esta fue elegida porque es el estándar de la literatura de minería de datos, mientras que los otros dos son impulsados por empresas del área. Los detalles de esta pueden encontrarse en la sección 2.4.

Los métodos de clasificación aplicados en modelos de *credit scoring* provienen de la rama de reconocimiento estadístico de patrones. En la siguiente sección se introducirá esta disciplina, así como también las técnicas que serán utilizadas en el presente trabajo.

[1] Información obtenida en entrevista con Richard Weber, académico del DII, Universidad de Chile.

2.3. Clasificación mediante reconocimiento estadístico de patrones.

La clasificación estadística de patrones trata el problema de identificar la clase a la que pertenece una observación particular de una población de datos, para la cual su etiqueta de clase es desconocida. Este procedimiento se realiza mediante métodos estadísticos construidos sobre un conjunto de observaciones denominados “base de entrenamiento”, para las cuales sí se conoce su etiqueta de clase (33).

El modelo básico sobre el cual se desarrollan todos los métodos de clasificación de patrones, asume que un patrón es un vector de **x** dimensión *p*, cuyas componentes corresponden a medidas de atributos de un objeto. Así, los atributos son variables especificadas por el investigador, y que parecen ser importantes para el proceso de clasificación. Se asumen que existen *C* grupos o clases, identificados por $\omega_1 \dots \omega_C$, y asociado a cada patrón **x** existe una etiqueta *y*, que indica la clase a la cual pertenece el objeto; si $y = c$, entonces el patrón pertenece a la clase ω_c (33). Por último, se supone que uno cuenta con un conjunto de observaciones $\{(\mathbf{x}_\mathrm{i}, y_i), i = 1 \dots n\}$ (el conjunto de entrenamiento), el cual se utilizará para diseñar un clasificador, que más tarde podrá ser utilizado para estimar la pertenencia a cada clase para un patrón desconocido **x**.

Ejemplos de aplicaciones del reconocimiento estadístico de patrones pueden encontrarse en las áreas de reconocimiento del habla, diagnóstico de pacientes, predicción climática, reconocimiento de caracteres (33), marketing, inversiones, detección de fraudes (13), entre otras.

En las secciones siguientes se asumirá que se está ante un problema de discriminación binaria; es decir, en el que sólo existen dos etiquetas para la variable *y*. En el caso de un modelo de probabilidad este es el caso interesante, pues interesa predecir si un cliente entrará o no en default en un lapso de tiempo dado.

A continuación se describirán tres métodos del aprendizaje estadístico para la clasificación de patrones: regresión logística, árboles de clasificación y redes neuronales.

2.3.1. Árboles de clasificación.

Un árbol de decisión es un método de clasificación no paramétrico, cuya estructura es semejante a un diagrama de flujo, en donde cada nodo interno denota un test en cierto atributo, cada rama representa un resultado de ese test, y cada nodo hoja (o terminal) conlleva una etiqueta de clase. La inducción en árboles de decisión es el aprendizaje a través de estas estructuras, a partir de datos de entrenamiento marcados con etiquetas de su clase correspondiente (15). La popularidad de estos métodos se debe a que la construcción de estos clasificadores no requiere conocimiento del dominio particular desde donde provienen los datos, ni de ajuste de parámetros. Esto los hace particularmente atractivos para análisis exploratorio de datos. Más aún, estos métodos pueden manejar una alta dimensionalidad, y su estructura de representación es intuitiva y fácil de asimilar.

Algunos algoritmos populares que implementan el concepto de árboles de clasificación son:

- ID3: utiliza el criterio de “ganancia de información” para seleccionar atributos.
- C4.5: utiliza el criterio de “razón de ganancia” para seleccionar atributos.
- CART: utiliza el criterio del “índice Gini” para seleccionar atributos.
- CHAID: utiliza el criterio “chi-cuadrado” para seleccionar atributos.

Para esta memoria será de particular interés el algoritmo CHAID. A continuación se explicará en más detalle el funcionamiento de esta herramienta.

2.3.1.1. CHAID (Chi-square Automated Interaction Detector)

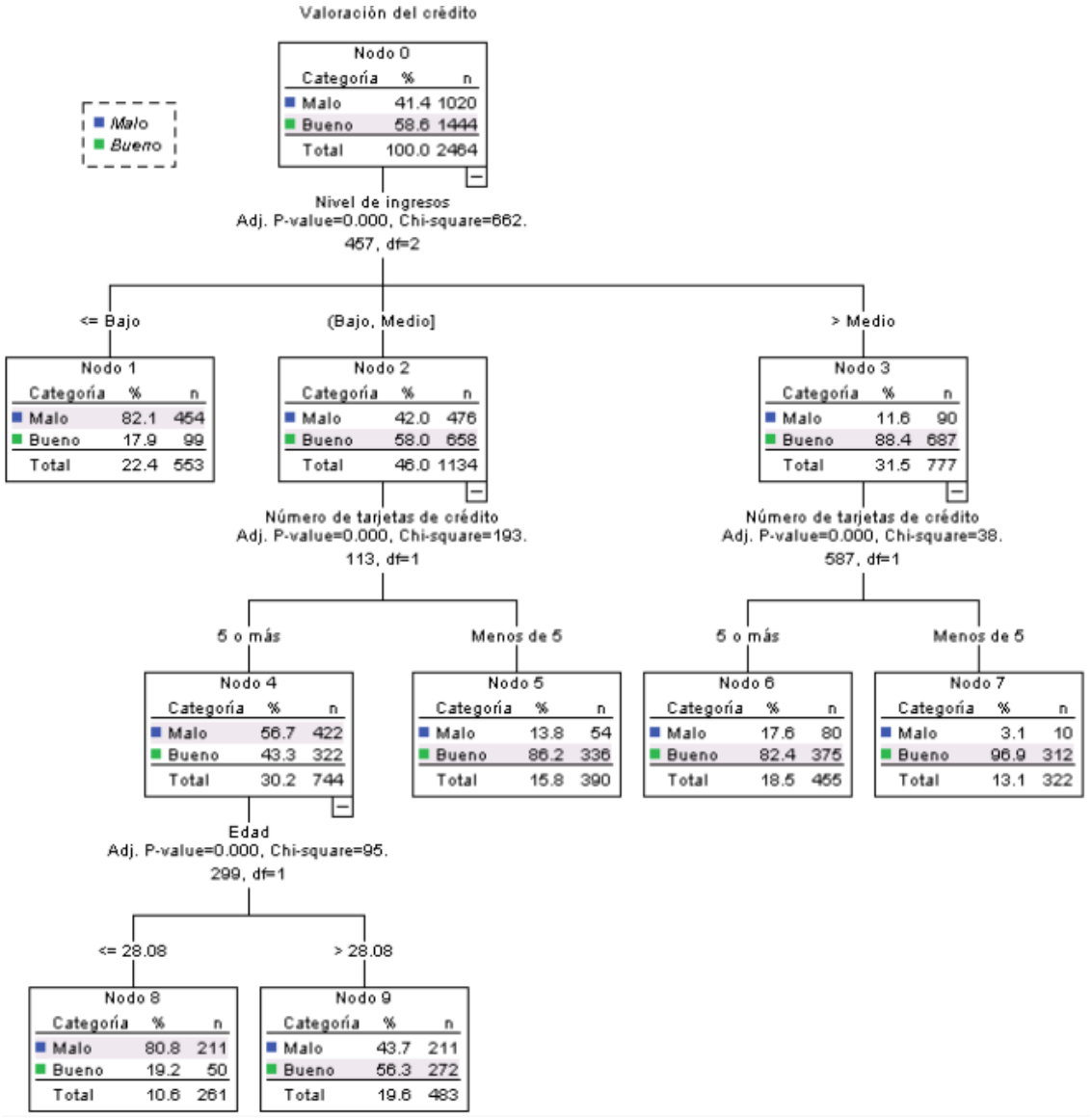
Básicamente, la técnica CHAID divide el conjunto de datos en sub conjuntos que son mutuamente excluyentes y exhaustivos, y que describen de mejor manera el comportamiento de la variable dependiente (20). Los subconjuntos se construyen utilizando un grupo pequeño del total de predictores disponibles. Los predictores seleccionados pueden ser usados más tarde en análisis subsiguientes, en la predicción de la variable dependiente, o en lugar del total de predictores en próximos recolecciones de datos.

El algoritmo bajo el cual se implementa CHAID, que es iterativo en esencia, procede de la siguiente manera. En primer lugar, se identifica la mejor partición por cada predictor disponible. Luego, los predictores se comparan y se escoge al mejor. Los datos se subdividen de acuerdo a la partición del atributo seleccionado, y cada una de las particiones vuelve a someterse al proceso descrito anteriormente, hasta que no se encuentren más particiones significativas.

La Figura 3 muestra una posible salida del algoritmo en el software SPSS. Se aprecia que la variable dependiente corresponde a la “Valoración del crédito” de una persona. El predictor que mejor explicaría el comportamiento de dicha variable sería “Nivel de ingresos”, obteniéndose que la valoración aumenta con los ingresos de la persona. En segundo lugar, se encontraría “Número de tarjetas de crédito”, mostrando que a mayor número de tarjetas, mayor es la valoración que esa persona tendría por el crédito. Finalmente, en tercer lugar aparece “Edad”, donde se observa que los jóvenes tienden a valorar menos el crédito.

Debido a la importancia que tiene el test chi-cuadrado en el funcionamiento del algoritmo CHAID, se expondrán a continuación sus detalles y los supuestos sobre los que se sustenta.

Figura 3. Ejemplo de salida del algoritmo CHAID en SPSS.



Fuente: Extraído desde (28).

2.3.1.2. Test chi-cuadrado de independencia

En el contexto del estudio de variables aleatorias categóricas, el test chi-cuadrado puede utilizarse para medir el grado de asociación de dos variables.

Supóngase que se cuenta con una tabla de contingencia, en donde un conjunto de observaciones aleatorias de una población son clasificadas en *R* filas y *C* columnas, en

donde estos dos últimos números corresponden a la cantidad de categorías disponibles en cada una de las variables aleatorias en estudio. Para $i = 1,..,R$ y $j = 1,\dots,C$ se definirá p_{ij} como la probabilidad de que un individuo seleccionado al azar de la población sea clasificado en la fila *i* y columna *j*. Más aún, sean

$$p_{i+} = \sum_{j=1}^{C} p_{ij}$$

$$p_{+j} = \sum_{i=1}^{R} p_{ij}$$

las distribuciones marginales de cada una de las variables aleatorias.

Supóngase ahora que se tiene una muestra de tamaño *N*, y sea N_{ij} el número de observaciones que caen en la posición (*i,j*) en la tabla de contingencia. Sean además N_{i+} el número total de observaciones clasificados en *i*, y N_{+j} el número total de observaciones en la fila *j*. En base a estas observaciones, se testearán las siguientes hipótesis:

$$H_0: p_{ij} = p_{i+}p_{+j} \quad \forall i = 1,\dots,R\ j = 1,\dots,C$$

H_1: la hipótesis nula no es cierta.

La hipótesis nula asume que las variables en estudio son independientes. Sea $\hat{E}_{ij}$ el estimador de máxima verosimilitud, cuando H_0 es cierta, del número esperado de observaciones que serán clasificados en la casilla (*i,j*) de la tabla de contingencia. En esta situación, el estadístico

$$Q = \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{\left(N_{ij} - \hat{E}_{ij}\right)^2}{\hat{E}_{ij}}$$

tiene una distribución de probabilidad asintótica χ^2 de $(R-1)(C-1)$ grados de libertad. Ahora bien, para calcular el estimador de máxima verosimilitud, considere que bajo la hipótesis nula, la probabilidad de que una observación caiga en la casilla (*i,j*) es simplemente Np_{ij}. Además, se tendría que $p_{ij} = p_{i+}p_{+j}$. Por último, dado que p_{i+}

puede entenderse como el parámetro de una distribución Bernoulli, en el que la variable es 1 si el individuo toma el valor *i* en la primera variable, y 0 en otro caso, entonces el estimador de máxima verosimilitud de este número es simplemente la proporción de observaciones que cumplen con esta condición (12) (lo mismo ocurre con p_{+j}). Es necesario recalcar que tal estimador será consistente, sólo en el caso en que las observaciones sean efectivamente independientes entre sí. Con esto, se tiene que

$$\hat{E}_{ij} = N\left(\frac{N_{i+}}{N}\right)\left(\frac{N_{+j}}{N}\right) = \frac{N_{i+}N_{+j}}{N}$$

La discusión presentada indica que la hipótesis nula debe ser rechazada cuando *Q* es mayor que un umbral dado un cierto nivel de confianza (12).

2.3.2. Regresión logística.

Se supondrá que es del interés del investigador modelar el comportamiento de una variable *y* dicotómica, la cual puede tomar sólo valores en {0,1}. Dado un vector de datos **x**, el método de regresión logística asume la siguiente forma funcional para la probabilidad de que *y* tome el valor 1 (19):

$$p(y = 1|\mathbf{x}) = \Lambda(\boldsymbol{\beta}^T\mathbf{x}) = \frac{1}{1 + \exp(-\boldsymbol{\beta}^T\mathbf{x})}$$

donde $\Lambda(\cdot)$ corresponde a la función de distribución (acumulada) logística y $\boldsymbol{\beta}$ es un vector de parámetros que deben ser ajustados en función de los datos. Usualmente se supone que el vector **x** contiene un término constante o intercepto.

Si se estuviera en presencia de una regresión lineal, el método preferido para la calibración de parámetros corresponde a mínimos cuadrados ordinarios (OLS, por sus siglas en inglés), el cual se basa en el principio de minimizar la suma de las desviaciones cuadráticas entre el valor predicho y el valor observado de *y* para un set de datos. Este estimador suele tener propiedades estadísticas deseables. Desgraciadamente, cuando se aplica el estimador OLS a un modelo de variable

dicotómica, estas propiedades se pierden, debido a que los supuestos básicos de éste método no se satisfacen (19).

El método general que entrega la estimación OLS en el ámbito de regresión lineal es conocido como máxima verosimilitud (ML, por sus siglas en inglés). El método de ML entrega parámetros que maximizan la probabilidad de obtener la data observada a través de realizaciones del modelo que se desea calibrar (19).

Ahora bien, suponiendo que se cuenta con *N* pares de observaciones $(\mathbf{x}_i, y_i)$, con $i = 1 \dots N$, la probabilidad de observar el dato *i* según el modelo y dado su vector de atributos $\mathbf{x}_i$, es igual a $\Lambda(\boldsymbol{\beta}^T \mathbf{x_i})$ para $y_i = 1$, y es igual a $1 - \Lambda(\boldsymbol{\beta}^T \mathbf{x_i})$ para $y_i = 0$. De forma más compacta, esto puede ser escrito como

$$L_i(\boldsymbol{\beta}) = \Lambda(\boldsymbol{\beta}^T \mathbf{x_i})^{y_i} \left(1 - \Lambda(\boldsymbol{\beta}^T \mathbf{x_i})\right)^{1-y_i}$$

Para encontrar la función de verosimilitud total, debe encontrarse una expresión para la probabilidad conjunta de observar la data recolectada utilizando realizaciones del modelo. En otras palabras, se debe poder escribir la cantidad

$$L(\boldsymbol{\beta}) = p(y_1 = \omega_1, y_2 = \omega_2, \dots, y_N = \omega_N | \mathbf{X})$$

donde ω_i corresponde a la realización observada del dato *i* (es decir, un número en {0,1}), y la matriz $\mathbf{X} = (\mathbf{x}_1\ \mathbf{x_2} \dots\ \mathbf{x}_N)$ contiene los atributos observables para los *N* datos.

El supuesto usualmente utilizado para hacer más tratable la expresión anterior, es que las *N* observaciones son independientes. En otras palabras, se asume que el conocer el resultado de una observación no entrega información útil para estimar el resultado de otra observación. En tal caso, la función de verosimilitud puede escribirse como

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} L_i(\boldsymbol{\beta})$$

Debido a que es matemáticamente más simple trabajar con sumatorias que con multiplicaciones, suele trabajarse con la log-verosimilitud $l(\boldsymbol{\beta}) = \log\big(L(\boldsymbol{\beta})\big)$. Como el logaritmo es una función monótona creciente, se cumple que

$$\underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, l(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, L(\boldsymbol{\beta})$$

Con todo lo anterior, el problema de optimización (irrestricto) a resolver es el siguiente (19):

$$\max_{\boldsymbol{\beta}} l(\boldsymbol{\beta}) = \max_{\boldsymbol{\beta}} \sum_{i=1}^{N} y_i \log(\Lambda(\boldsymbol{\beta}^T \mathbf{x_i})) + (1 - y_i)\log(1 - \Lambda(\boldsymbol{\beta}^T \mathbf{x_i}))$$

Muchas veces, la estructura de los datos que se poseen no cumple con el supuesto de independencia que entrega la forma cerrada para la verosimilitud antes descrita. Uno de estos casos ocurre cuando se está en presencia de una base de datos de panel. En la siguiente sección se explora en métodos de regresión logística diseñados para estas estructuras de datos. Se seguirá a Wooldridge (34).

2.3.2.1. Regresión logística con datos de panel

Un panel de datos de elección discreta consta de observaciones sobre *N* individuos que se siguen a lo largo de *T* periodos. En este caso, se cuenta para cada individuo *i* con un vector $\mathbf{y}_i = (y_{i1} \dots y_{iT})^T$ que contiene las decisiones en cada momento del tiempo, y una matriz $\mathbf{X}_i = (\mathbf{x}_{i1} \dots \mathbf{x}_{iT})^T$ que contiene sus atributos medibles para cada instante del tiempo.

El caso más simple de estimación de una regresión logística con datos de panel ocurre cuando se asume que

$$p(y_{it} = 1|\mathbf{x}_{it}, y_{it-1}, \mathbf{x}_{it-1}, y_{it-2} \dots) = p(y_{it} = 1|\mathbf{x}_{it})$$

Esto significa que para estimar la probabilidad en el instante *t*, no aporta información conocer las realizaciones de la variable *y* en otros instantes de tiempo, así como tampoco aporta información conocer los atributos del individuo en otros instantes

de tiempo. Esto supuesto se conoce como “completitud dinámica” (34). En este caso, la estimación del vector de parámetros del modelo logístico pasa a ser un ejercicio de “estimación agrupada” o “*pooled estimation*”. En otras palabras, el conjunto de datos puede tratarse de la misma manera que se haría con una base de datos de corte transversal de tamaño *NT*.

El modelo logístico de efectos no observados constituye otra forma de tomar en cuenta la estructura de datos de panel. Dos son los supuestos principales de este enfoque:

$$p(y_{it} = 1|\mathbf{x}_i, c_i) = p(y_{it} = 1|\mathbf{x}_{it}, c_i) = \Lambda\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + c_i\right)$$
$$y_{i1}, \dots, y_{iT} \text{ son independientes condicionales en } (\mathbf{x}_i, c_i)$$

donde c_i es el efecto no observado. La primera ecuación de la primera línea, indica que $\mathbf{x}_{it}$ es estrictamente exógeno condicional en c_i. En otras palabras, una vez que se conocen las características idiosincráticas del individuo, $\mathbf{x}_{it}$ contiene toda la información necesaria de $\mathbf{x}_i$ para estimar la probabilidad del resultado en *t* para el individuo *i*. La segunda ecuación de la primera línea sólo está utilizando el supuesto básico del modelo logístico, sólo que ahora el argumento de la función incluye al efecto no observable de forma aditiva. Usando lo anterior la función de verosimilitud para el individuo *i* queda (los detalles del desarrollo que deriva en esta fórmula pueden encontrarse en (34)):

$$l_i(\boldsymbol{\beta}) = \log\left\{\exp\left(\sum_{t=1}^{T} y_{it}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it}\right)\left[\sum_{\mathbf{a}\in R_i}\exp\left(\sum_{t=1}^{T} a_t\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it}\right)\right]^{-1}\right\}$$

donde R_i es el subconjunto de $\mathbb{R}^T$ definido como $\{\mathbf{a} \in \mathbb{R}^T : a_t \in \{0,1\}$ y $\sum_{t=1}^{T} a_t = n_i\}$, y donde $n_i \equiv \sum_{t=1}^{T} y_{it}$. Sumando las log-verosimilitudes anteriores sobre todo *i* y resolviendo el problema de optimización, se obtiene un estimador de máxima verosimilitud consistente de $\boldsymbol{\beta}$.

Por último, el modelo logístico dinámico de efectos no observables se basa en el siguiente supuesto

$$p(y_{it} = 1|y_{it-1}, \dots, y_{i0}, \mathbf{z}_i, c_i) = \Lambda(\boldsymbol{\delta}^T \mathbf{z}_{it} + p y_{it-1} + c_i)$$

donde $\mathbf{z}_{it}$ es un vector de variables explicativas contemporáneas y $\mathbf{z}_i = (\mathbf{z}_{i1}, \dots, \mathbf{z}_{iT})$. En base a este supuesto, es claro ver que este modelo, al igual que los anteriores, asume una exogeneidad estricta, ya que sólo $\mathbf{z}_{it}$ aparece en el lado derecho de la igualdad. Sin embargo, este modelo permite que la realización en *t* dependa tanto de la realización en el periodo anterior como de los efectos no observados del individuo. Los detalles de la estimación de este modelo están en (34).

2.3.3. Redes neuronales.

Una red neuronal es un conjunto de unidades (o neuronas) conectadas, donde a cada conexión se asocia un peso. Durante la fase de aprendizaje, la red aprende mediante el ajuste de los pesos de las conexiones, de manera de predecir correctamente la clase de los individuos de la base de entrenamiento (15). Existen muchos tipos de redes neuronales, pero en este trabajo se enfocará la atención en los "*Multilayer Perceptrons*" o MLP, pues son las redes neuronales más usadas para tareas de clasificación (2).

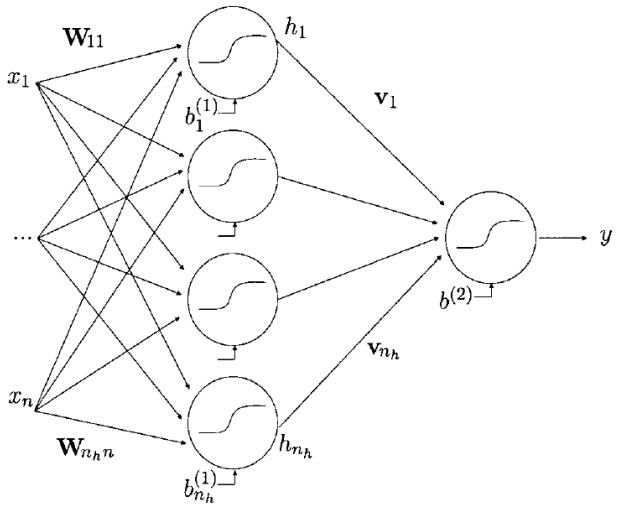
Un MLP se compone típicamente de una capa de entrada, una o más capas ocultas, y una capa de salida. Cada capa puede consistir de varias neuronas. Las neuronas procesan sus datos de entrada, y generan una salida que se transmite a las neuronas de la capa siguiente (2).

La salida de la neurona *i* de la capa escondida se calcula de la siguiente manera:

$$h_i = f^{(1)}\left(b_i^{(1)} + \sum_{j=1}^{n} W_{ij} x_j\right)$$

donde h_i es la salida; $f^{(1)}$ es la función de transferencia de la capa oculta; $b_i^{(1)}$ es el término de sesgo correspondiente a la neurona *i* de la capa oculta; W_{ij} es el peso del arco formado entre la neurona *i* y la entrada *j*; x_j corresponde al valor de la entrada *j*, y *n* es el número de señales de entrada.

Figura 4. Arquitectura de un MLP con una capa escondida.



Fuente: Extraída desde (2).

De manera similar, la señal de la capa de salida se calcula como

$$y = f^{(2)}\left(b_i^{(2)} + \sum_{j=1}^{m} v_j x_j\right)$$

Los términos de sesgo juegan un papel similar al intercepto de una regresión lineal. Las funciones de transferencia permiten modelar relaciones no lineales en los datos. Para problemas de clasificación, es conveniente utilizar la función de transferencia logística (2):

$$f(x) = \frac{1}{1 + \exp(-x)}$$

Esta función tiene una salida limitada al intervalo [0,1]. Esto permite que *y* sea interpretable como una probabilidad condicional, o a posteriori (2). Los parámetros son estimados a través del algoritmo "*Backpropagation*".

Se ha demostrado que los MLP estiman probabilidades Bayesianas (a posteriori), cuando se utilizan las medidas de error cuadrático o de *cross-entropy* (24). Ahora bien, para obtener estimadores muestrales consistentes de ambos errores, es necesario que

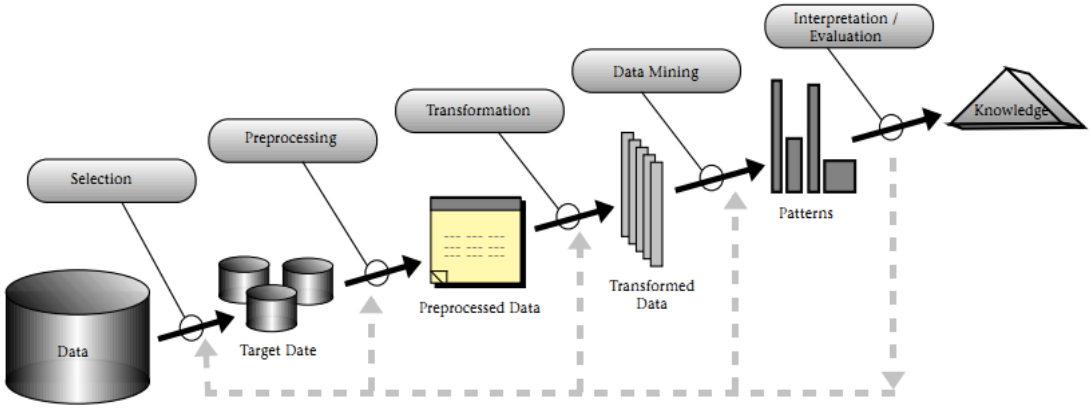
la muestra con que se entrena el MLP consista de observaciones independientes e idénticamente distribuidas.

Una desventaja de las redes neuronales aparece al intentar interpretar su resultado. El conocimiento adquirido en la forma de una red conectada por pesos es difícil de interpretar. Este factor ha motivado la investigación enfocada en extraer conocimiento embebido en una red neuronal entrenada, y en representar este conocimiento de manera simbólica (15). De hecho, este modelo no es aconsejable bajo los lineamientos de BIS-II, por lo que se implementación en este trabajo se reduce meramente a un objetivo de *benchmark*.

2.4. Metodología KDD.

Un proyecto de diseño y calibración de una herramienta de reconocimiento estadístico de patrones, puede situarse en el marco de trabajo para la minería de datos conocido como "*Knowledge Discovery in Databases*" o KDD. El KDD "es el proceso no-trivial de identificar patrones en los datos que sean válidos, novedosos, potencialmente útiles y entendibles" (13). El KDD ha sido utilizado en un amplio espectro de disciplinas. En el ámbito de los negocios, se conocen aplicaciones en marketing, finanzas, detección de fraudes, manufactura, telecomunicaciones y otros. La Figura 5 muestra un resumen del proceso KDD.

Figura 5. Resumen de los pasos que componen el proceso KDD.



Fuente: Extraído desde (13).

Las descripciones de cada una de las etapas del proceso según (13) son:

1. Selección de variables: se escoge el set de datos sobre el cual se pretende extraer conocimiento.
2. Limpieza de la data y pre procesamiento: algunas de las operaciones realizadas en esta etapa son la eliminación de ruido en los datos y uso de estrategias de manejo de *outliers*[2].
3. Transformación de la data: los objetivos de esta etapa son dos: en primer lugar, aplicar operaciones matemáticas a los datos de manera de adecuarlos a los requerimientos de los modelos de minería de datos que se utilizarán. Y en segundo lugar, aplicar técnicas matemáticas para maximizar la cantidad de conocimiento que se extrae de los atributos originales.
4. Aplicación de un método de minería de datos: en el caso particular de esta memoria, se aplicarán modelos de clasificación para estimar probabilidades de incumplimiento.
5. Interpretación y evaluación: corresponde al último paso del proceso, pero puede también llevar a realizar nuevamente cualquiera de los anteriores puntos, debido a una evaluación insatisfactoria.

Es importante notar que, tal como se menciona en el paso número cinco, el proceso KDD es iterativo, y puede implicar ciclos entre cualquier subconjunto de pasos.

2.5. Evaluación de modelos de *credit scoring*.

Según (33), existen dos dimensiones que deben necesariamente evaluarse a la hora de determinar el nivel de desempeño de un modelo de *credit scoring*. La primera dimensión corresponde a la capacidad de discriminación que posea el modelo, o qué tan bien clasifica datos previamente desconocidos. La segunda corresponde a la capacidad de predicción o confiabilidad, y mide la precisión con que la herramienta estima las probabilidades a posteriori de pertenencia a cada clase.

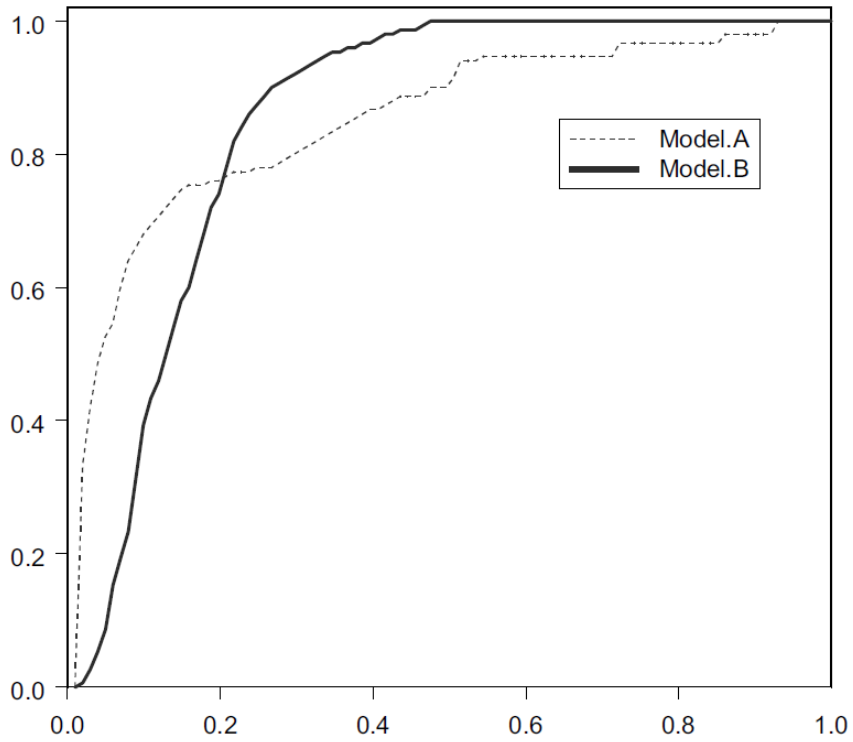
[2] Valores fuera de rango.

Una medida conocida de evaluación de la capacidad de discriminación de un modelo corresponde al área bajo la curva ROC, o AUC. En la siguiente sección se introducirá esta metodología. Otra medida ampliamente utilizada en el ámbito del *credit scoring* corresponde al estadístico de Kolmogorov-Smirnov, el cual se presentará más adelante.

2.5.1. Curva ROC y AUC.

La curva ROC (del inglés "*Receiver Operating Characteristic*"), fue introducida como medio de evaluación del desempeño de una regla de discriminación de dos clases, a la vez que provee un método visual que permite escoger un umbral de decisión apropiado. Esta curva es un gráfico de la tasa de "verdaderos positivos" en el eje vertical, contra la tasa de "falsos positivos" en el eje horizontal, para distintos umbrales de decisión (33). A continuación se muestra un gráfico de eta curva para dos modelos distintos:

Figura 6. Curva ROC para dos modelos distintos.



Fuente: extraído desde (29).

Si los costos por errar en la clasificación de cada una de las clases fuesen conocidos, entonces podría computarse el umbral óptimo que minimiza este error. La solución con costo mínimo es la cual cumple con que la pendiente de la curva ROC es igual a $\lambda_{21}p(\omega_2)/\lambda_{12}p(\omega_1)$, donde λ_{21} (λ_{12}) corresponde al costo de clasificar mal un objeto de la clase 2 (1), y $p(\omega_2)$ ($p(\omega_1)$) es la probabilidad a priori de pertenencia a la clase 2 (1) (en otras palabras, la proporción en la muestra da cada una de las clases). Para distintos valores de los parámetros anteriores, el error mínimo ocurrirá en un punto diferente de la curva ROC. Sin embargo, en la práctica es difícil conocer exactamente los valores para los costos por mala clasificación. Una medida de separación que ignora estos parámetros corresponde al área bajo la curva ROC o AUC. Ésta provee una medida única que no asume nada en relación a los costos de clasificación. Un AUC=0,5 indica que el modelo entrega la misma información que un clasificador aleatorio; un AUC=1,0 indica discriminación perfecta.

La desventaja de utilizar el enfoque AUC, es que si dos curvas se cruzan (tal como en el gráfico anterior), entonces no es cierto que el modelo con mayor AUC será preferido frente a otro. De hecho, dos curvas ROC pueden tener el mismo valor para AUC, pero a la vez mostrar formas muy distintas. Dependiendo de la aplicación, un modelo podrá ser favorecido por sobre el otro.

2.5.2. Test de Kolmogorov-Smirnov para dos muestras

El test Kolmogorov-Smirnov (K-S en adelante) para dos muestras tiene por objetivo testear la hipótesis de que dos muestras aleatorias provienen de la misma distribución teórica (12). Ahora bien, en el ámbito del *credit scoring*, el test K-S se utiliza para describir cuan lejanas están las características de dos poblaciones. En particular, si uno tiene un sistema de *scoring* que a cada miembro de la población le asigna un puntaje, entonces uno podría utilizar estas medidas para describir cuan diferentes son los puntajes de los “buenos” y los “malos” (clientes que cumplieron con sus obligaciones crediticias y clientes que no lo hicieron) (31).

Considere una situación en la cual se tiene una muestra de *m* observaciones $X_1, \dots, X_m$ de una distribución para la cual su función de distribución acumulada (o c.d.f.

por sus siglas en inglés) $F(x)$ es desconocida, y por otro lado, se tiene otra muestra aleatoria (independiente de la anterior) de *n* observaciones $Y_1, \dots, Y_n$ de una distribución cuya c.d.f. $G(x)$ también es desconocida. Se asumirá que tanto $F(x)$ como $G(x)$ son funciones continuas, y que se desea probar la hipótesis de que ambas funciones son idénticas, sin especificar su forma funcional (12). Por tanto, se intentará probar las siguientes hipótesis:

$$H_0: F(x) = G(x) \quad \forall -\infty < x < \infty$$
$$H_1: H_0 \text{ no es cierta.}$$

Sean $F_m(x)$ y $G_n(x)$ las c.d.f. muestrales (o empíricas) calculadas a partir de los valores $X_1, \dots, X_m$ y $Y_1, \dots, Y_n$, respectivamente. Considere el estadístico D_{mn}, definido como:

$$D_{mn} = \sup_{-\infty<x<\infty} |F_m(x) - G_n(x)|$$

Cuando la hipótesis nula es cierta, el lema de Glivenko-Cantelli indica que (12):

$$D_{mn} \xrightarrow{p} 0 \text{ cuando } m \to \infty \text{ y } n \to \infty$$

El estadístico K-S de dos muestras permite obtener una distribución asintótica para D_{mn}, lo que a su vez permite construir un test aproximado. Para cada valor de $t > 0$, sea

$$H(t) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2t^2}$$

Si la hipótesis nula es cierta, entonces

$$\lim_{m\to\infty, n\to\infty} \mathbb{P}\left[\left(\frac{mn}{m+n}\right)^{1/2} D_{mn} \le t\right] = H(t)$$

El test construido a partir del estadístico anterior se conoce como el test de Kolmogorov-Smirnov para dos muestras (12).

2.5.3. Suficiencia de las medidas de clasificación

En el ámbito de *credit scoring*, las medidas de clasificación son insuficientes (11). La pregunta no es cuan bien una compañía especifica se comporta en comparación con otros, sino cuál es la probabilidad de que esta compañía incumpla en un horizonte dado. Esto no significa que las medidas de clasificación sean innecesarias, pues entregan información sobre los errores Tipo I y Tipo II en que se incurre. Esta información resulta útil para, por ejemplo, definir niveles de corte o umbrales para clasificación. Pero por otro lado, la capacidad de clasificación es insuficiente, en el sentido de que un modelo muy preciso a la hora de ordenar individuos, puede resultar muy impreciso en la predicción de probabilidades. A esta medida se le conoce como calibración del modelo.

Ahora bien, en el caso en que se cuente con una regresión logística con un buen nivel de discriminancia, es directo corregir sus parámetros de manera de obtener además una regresión calibrada a los datos. Una forma de hacer esto es corregir el coeficiente de intercepto, de manera que, en promedio, la probabilidad estimada por la regresión sea igual a la observada en el set de datos donde interesa aplicarla, usando la siguiente fórmula (21):

$$\beta_0' = \beta_0 - \ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right]$$

donde $\bar{y}$ corresponde al nivel promedio de transiciones a incumplimiento observada en el conjunto de datos con el que se calibró la regresión, y τ a la proporción de incumplimientos en el conjunto donde se aplicará la regresión. Debido a que este procedimiento es estándar, no se profundizará en la calibración de las regresiones implementadas en este trabajo.

2.6. Teoría de diseño de muestreo

Esta sección tiene por objetivo exponer, a grandes rasgos, la teoría de diseño de muestras, de manera de poder adaptarla a la situación particular a la que apunta esta

memoria: el desarrollo de un *scoring* de comportamiento. En adelante se seguirá a (22).

2.6.1. Conceptos básicos

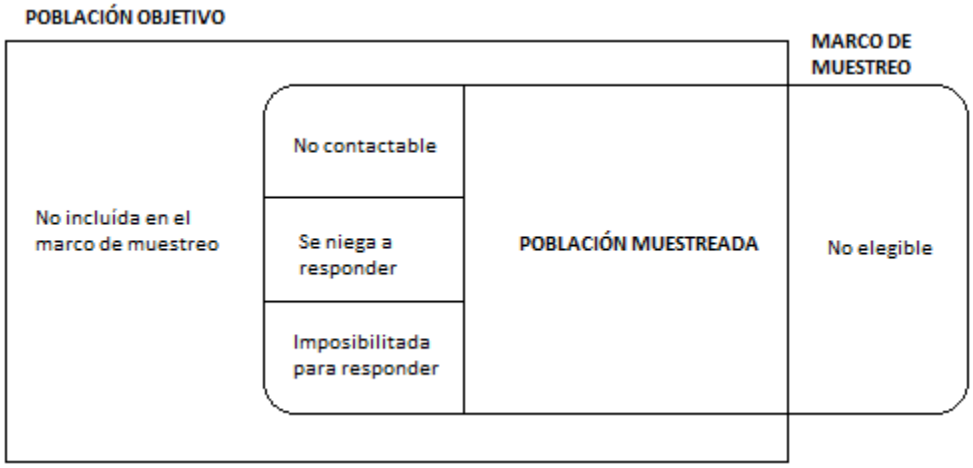
Las definiciones imprescindibles de manejar en el contexto de muestreo, son las siguientes:

- Unidad de observación: objeto que será medido.
- Población objetivo: colección completa de individuos a estudiar.
- Muestra: subconjunto de la población objetivo.
- Población muestreada: observaciones que podrían haber salido en la muestra.
- Unidad de muestreo: observación que puede ser incluida en la muestra. Por ejemplo, si se quiere estudiar individuos de un país, la unidad de muestreo podrían ser los hogares de ese país, y por lo tanto, las unidades de observación serían los individuos viviendo en hogares.
- Marco de muestreo: una lista de unidades de muestreo de la población desde la cual se obtendrá una muestra.

Es claro que, en un muestreo ideal, la población muestreada será idéntica en sus características a la población objetivo, aunque esto muy pocas veces es posible de lograr. La Figura 7 muestra un ejemplo gráfico de algunos de los conceptos definidos.

Existen, esencialmente, dos clases de errores que son susceptibles de ocurrir a la hora de muestrear: error muestral, y error no muestral. El error no muestral se compone, básicamente, de dos tipos; sesgo de selección y error de medición. Se espera que un muestreo tenga el menor sesgo de selección posible. El sesgo de selección ocurre cuando alguna parte de la población objetivo no está presente en la población muestreada. O, de manera más general, cuando algunas unidades de la población se muestrean a una tasa distinta a la que al investigador le interesaría. Un alto nivel de sesgo puede ir en detrimento de la calidad de los estimadores muestrales obtenidos.

Figura 7. Ejemplo gráfico de conceptos de muestreo.



Fuente: Traducción propia de contenido de (22).

Cuando el valor de una medición difiere del valor real, se incurre en error de medición. En otras palabras, éste ocurre cuando la medición tiene a diferir del verdadero valor, en alguna dirección. Este tipo de error es particularmente importante en el caso de las encuestas.

Por último, el error muestral es el que se deriva de medir sólo una parte de la población objetivo. Es claro que, aunque se controlaran perfectamente los errores no muestrales, si se tomara otra población muestreada, los resultados del estudio serían distintos.

2.6.2. Muestreos probabilísticos.

Un muestreo probabilístico es un mecanismo en el cual cada unidad de la población tiene una probabilidad de selección conocida, y donde se utiliza un generador de números aleatorios para escoger qué unidades específicas serán incluidas en la muestra. Si un muestreo probabilístico se implementa correctamente, será posible hacer inferencia de una población arbitrariamente grande. A continuación se definen cuatro modelos básicos de muestreo probabilístico:

- Muestreo aleatorio simple (MAS): se selecciona una muestra de tamaño *n*, asignándole la misma probabilidad a cada subconjunto de tamaño *n* de la

población. Los MAS son la base de cualquier diseño de muestreo más complejo.

- Muestreo aleatorio estratificado: en este caso, la población se divide en sub grupos denominados “estratos”. Luego, se realiza un MAS en cada estrato, independiente de cualquier otro estrato. Los elementos de cada estrato tienden a ser más similares que cualquier par de elementos escogidos al azar de la población completa, por lo que la precisión tiende a aumentar.
- Muestreo de *clusters*: se utiliza cuando no se cuenta con una lista de todos elementos de la población que se quiere estudiar, pero sí está disponible una lista con conjuntos de estos individuos. Estos conjuntos se denominan *clusters*. Por lo tanto, en este caso se muestrean en primer lugar los *clusters* mediante MAS, para luego seleccionar todos o algunos de los miembros dentro de cada *cluster*.
- Muestreo sistemático: es un proceso que se inicia con un elemento de la población tomado al azar de una lista. Luego, cada k-ésimo siguiente en la lista es escogido para entrar en la muestra.

Dada la importancia del MAS y del muestreo estratificado, en la siguiente sección se entrará en detalle en estas dos modalidades.

2.6.3. MAS y muestreo estratificado.

El MAS es la forma más básica de muestreo y provee un marco teórico fundamental para otros métodos más complejos. Existen dos maneras de llevar a cabo un MAS: con reemplazo, en el cual cada unidad puede ser incluida más de una vez, y sin reemplazo, en el que todas las unidades en la muestra son distintas.

Un MAS con reemplazo de tamaño n en una población de N unidades es equivalente a realizar n MAS de tamaño 1. Cada unidad se selecciona al azar con probabilidad $1/N$. En el caso de poblaciones finitas, sin embargo, muestrear el mismo caso más de una vez no agrega información. Por lo tanto, a veces es preferible muestrear sin reemplazo, de manera que la muestra no contenga duplicados. Un MAS sin reemplazo de tamaño o n se selecciona de manera que cada sub conjunto de

tamaño *n* en la población tenga la misma probabilidad de ser seleccionado. Dado que existen $\binom{N}{n}$ sub conjuntos con *n* elementos, entonces, la probabilidad de seleccionar una muestra S_n de tal tamaño cualquiera es

$$P(S_n) = \frac{1}{\binom{N}{n}}$$

En consecuencia, la probabilidad de que cada individuo aparezca en la muestra es

$$\pi_i = \frac{n}{N}$$

Los MAS son usualmente fáciles de diseñar y analizar. Sin embargo, pueden no ser los más apropiados en las siguientes situaciones:

1. cuando no se maneja una lista con los individuos de la población objetivo, o
2. cuando se cuenta con información adicional que puede ser usada para diseñar un esquema más costo-eficiente. Si se maneja previamente información sobre las características de subgrupos de la población, entonces ésta se puede usar para crear estratos y llevar a cabo MAS en cada uno de ellos.

Si a uno se le presenta el segundo caso, entonces es preferible utilizar un muestreo aleatorio estratificado.

2.6.4. Muestreo estratificado

El muestreo estratificado se utiliza en situaciones en donde se asume que las variables que se intentan medir tienen distinto comportamiento en estratos distintos de la población. Los estratos son una partición del conjunto población; esto es, son no vacíos, no se solapan y cubren todo el espacio. Esto implica que cada unidad de la población pertenece a exactamente un estrato. El muestreo estratificado consiste en realizar MAS en cada estrato, y luego agrupar los datos para obtener los estimadores de la población.

Las razones para utilizar un muestreo estratificado son varias. Entre ellas se encuentran las siguientes

1. Se requiere que todos los estratos estén representados en la muestra.
2. Se necesita precisión adicional en algún sub conjunto de la población.
3. Puede resultar en menores costos para el estudio.
4. El muestreo estratificado entrega estimadores poblacionales más preciso (menor varianza).

Existen dos maneras de determinar el tamaño muestral de cada estrato: asignación proporcional y asignación óptima. En la asignación proporcional, el número de unidades muestreadas en cada estrato es proporcional al tamaño del estrato. Esto implica que la probabilidad de incluir al objeto *j* del estrato *h* es $\pi_{hj} = n_h/N_h$, y es igual para todos los estratos ($= n/N$), en donde $n = \sum_H n_h$ es el tamaño de la muestra, $N = \sum_H N_h$ es el tamaño de la población, y *H* es la partición de la población (conjunto de estratos).

La asignación proporcional es útil cuando la varianza dentro de cada estrato es más o menos constante a lo largo de estos. En los casos en que la varianza dentro de los estratos varía considerablemente, la asignación óptima puede resultar en costos menores. El objetivo es ganar la mayor cantidad de información al menor costo posible. Esto implica minimizar el costo del muestreo para un nivel de varianza dado, o equivalentemente, minimizar la varianza para un nivel de costo dado. Aquí se estudiará el primer caso. Se supone que se cuenta con el valor del costo de muestrear un dato en cada estrato c_h y el costo de realizar el estudio c_0. Entonces, el costo total será

$$C = c_0 + \sum_{h \in H} c_h n_h$$

Es posible probar que la asignación óptima en este caso de n_h es proporcional a

$$\frac{N_h S_h}{\sqrt{c_h}}$$

donde S_h corresponde a la varianza dentro del estrato *h* de la variable que se quiere observar. Por lo tanto, bajo este criterio, se deberán tomar muestras más grande en estratos de mayor tamaño, con varianza mayor y donde el costo de muestrear sea bajo.

La asignación Neyman es un caso particular del anterior, en que todos los costos son iguales. En tal caso $n_h \propto N_h S_h$. Por lo tanto, los estratos de mayor tamaño y con mayor varianza serán los que tendrá una participación mayor.

Capítulo 3: Diseño de metodología de muestreo

"It is utterly implausible that a mathematical formula should make the future known to us, and those who think it can, would once have believed in witchcraft."

-Jacob Bernoulli, en *Ars Conjectandi* (1713).

3.1. Supuestos teóricos de los modelos estadísticos escogidos

El Marco Conceptual de este trabajo consistió en describir en extenso las características de cuatro métodos de clasificación estadística que podrían ser potencialmente incluidos en un modelo de probabilidad de incumplimiento. Así, podrá notarse que tanto la Regresión Logística como los árboles CHAID asumen ciertos supuestos para llevar a cabo inferencia estadística. Específicamente, se tratará el supuesto de independencia de las observaciones que utilizan ambos métodos, y de cómo éste no se cumple directamente en las bases de datos comúnmente utilizadas para calibrar modelos de probabilidad de incumplimiento. Este análisis se inicia, por tanto, con una descripción de estas bases de datos.

3.1.1. Estructuras de datos utilizadas para modelos de *scoring* de comportamiento

Los datos comúnmente utilizados para desarrollar modelos de *scoring* de comportamiento en instituciones financieras, corresponden bases de “cierre de mes”. Estas bases de datos se construyen a partir de la consolidación de datos de bases operacionales, que incluye un nivel de detalle mayor. Tal como su nombre lo dice, las bases de cierre de mes cuentan con observaciones mensuales para cada cliente activo en el banco. Las variables más importantes corresponden al Saldo Insoluto del cliente y los días de mora que el cliente tenía al cierre de mes. Un ejemplo de estas bases puede verse en Tabla 2. Nótese que el formato usualmente utilizado para las fechas

corresponde a “aaaa-mm”. Este ejemplo, muestra el estado de dos clientes en dos meses. Puede verse que uno de sus clientes ha dejado de cancelar sus obligaciones, y que en un mes más entrará en default, según la definición antes acordada en este trabajo.

Tabla 2. Ejemplo de base de datos de cierre de mes.

Id Cliente	Fecha	Saldo insoluto	Días de mora	Otras variables
001	200702	1.000.000	31	…
001	200703	1.000.000	60	…
002	200702	500.000	0	…
002	200703	450.000	0	…

Fuente: elaboración propia.

Para el efecto del desarrollo de modelos de probabilidad de incumplimiento, estas bases suelen cruzarse con datos de distintas fuentes para obtener otras variables potencialmente interesantes, tales como:

- Deuda y comportamiento de pago del cliente en el sistema
- Información de EE.FF. (en el caso de clientes empresas)
- Datos demográficos (clientes personas naturales)
- Características del producto contratado
- Variables macroeconómicas.

Más adelante, cuando se caracterice la base de datos sobre la que se construirá y validará la metodología de este trabajo, podrá verse un set de variables que típicamente se encuentra en el desarrollo de un modelo de P.I.

Ahora se verá cómo se construye una marca de deterioro (90 días de mora o más) para un cliente particular, que cambia a través del tiempo, y que considera un horizonte de 12 meses. La Tabla 3 muestra la evolución de un cliente, con su saldo de deuda insoluto y sus días de mora, que en enero de 2007 pide un crédito por MM$ 1.000.000, paga tres cuotas mensuales, y luego entra en mora. La última observación de la base de datos corresponde a diciembre de 2008. Las dos columnas de la derecha se construyen “ex post”; es decir, una vez observado el comportamiento del cliente en el

tiempo. En particular, la última columna corresponde a la marca de incumplimiento. Esta es la variable que se intentará modelar, de manera de generar un modelo de probabilidad de incumplimiento. La marca es

- igual a 1, si el cliente entra en default en los próximos 12 meses, y
- cero en otro caso.

A partir de esta definición, se desprende que para poder calcular la marca, se deben tener por lo menos 12 meses de observación hacia adelante. Por lo tanto, y tal como se ve en el ejemplo (la marca para enero de 2008 en adelante no tiene valor), será necesario desechar los últimos 12 meses de la base de datos, pues si se utilizaran se estaría cayendo en un error de datos censurados. En otras palabras, no se tiene suficiente información para decir qué ocurrirá con los clientes en fechas posteriores.

En el ejemplo, el cliente incumple al cuarto mes de entregado el crédito. Por lo tanto, la marca de incumplimiento capturará esto desde el primer mes que aparece (ya que cuatro meses caen dentro de un rango de observación de doce meses). Es por esto que la marca es igual a uno desde el primer momento.

3.1.1. Consideraciones con respecto a supuestos de independencia de las observaciones.

Tal como se explicitó en el marco teórico, tanto la regresión logística como los árboles de clasificación asumen que las observaciones usadas para calibrar los modelos provienen de datos independientes entre sí.

Se mostró en la sección Marco Conceptual, que los árboles basados en la técnica CHAID realizan inferencia estadística mediante el uso de *tests* de independencia basados en distribuciones chi-cuadrado. En particular, se vio que para derivar la distribución del estadístico *Q*, era necesario utilizar el estimador de máxima verosimilitud de la cantidad de observaciones en una celda (i,j) ($\hat{E}_{ij}$) de la tabla de contingencia. Se mostró que este estimador es consistente sólo en el caso de que las observaciones de cada celda sean independientes unas de otras. Por tanto, esto indica

que frente a un problema con este último supuesto, la inferencia estadística obtenida a través del estadístico *Q* no será válida, lo que a su vez implica que los árboles CHAID no pueden usarse directamente sobre todo el set de datos de la base de cierre de mes, sin incurrir en errores de estimación.

Tabla 3. Ejemplo de comportamiento de un cliente en el tiempo.

Id	Fecha	Deuda	Días de mora	En default?	Transita a default en 12 meses?
001	200701	1.000.000	0	No	1
001	200702	950.000	0	No	1
001	200703	900.000	0	No	1
001	200704	850.000	0	No	1
001	200705	850.000	30	No	1
001	200706	850.000	60	No	1
001	200707	850.000	90	Sí	1
001	200708	850.000	120	Sí	1
001	200709	850.000	150	Sí	1
001	200710	850.000	180	Sí	1
001	200711	850.000	210	Sí	1
001	200712	850.000	240	Sí	1
001	200801	850.000	270	Sí	---
001	200802	850.000	300	Sí	---
001	200803	850.000	330	Sí	---
001	200804	850.000	360	Sí	---
001	200805	850.000	390	Sí	---
001	200806	850.000	420	Sí	---
001	200807	850.000	450	Sí	---
001	200808	850.000	480	Sí	---
001	200809	850.000	510	Sí	---
001	200810	850.000	540	Sí	---
001	200811	850.000	570	Sí	---
001	200812	850.000	600	Sí	---

Fuente: elaboración propia.

Ahora bien, con respecto a la regresión logística aplicada a bases de corte transversal (no en bases tipo panel), vista en la sección 2.3.1, también necesita basarse en el supuesto de independencia de las observaciones, de modo de poder calibrar el vector de parámetros $\boldsymbol{\beta}$ utilizando el método de máxima verosimilitud. Así, al entrenar un modelo de este tipo ignorando el cumplimiento del supuesto antes mencionado, resultará en una estimación de parámetros inconsistente.

Por otro lado, las regresiones logísticas sobre datos de panel intentan hacerse cargo de cierto tipo de correlación serial (en el tiempo) en los datos, al reconocer que éstos provienen de una observación a lo largo del tiempo de individuos distintos. Sin embargo, se mostrará más adelante por qué las técnicas de regresión logística para datos de panel tampoco son adecuadas para el tipo de datos en una base de cierre de mes.

A continuación se mostrará por qué al utilizar todos los pares (Id, fecha) de la base anteriormente descrita se incurre en violaciones al supuesto de independencia. Sin embargo, se puede adelantar que el problema fundamental que enfrenta la estimación de regresiones logísticas de manera directa en este ambiente, es el de suponer que todos los datos provienen de una distribución única y estable, que es independiente del tiempo. En otras palabras, el supuesto de que se cuenta con pares $(y_j, \mathbf{x}_j)$, para los cuales la probabilidad de incumplimiento se puede modelar como (19):

$$p\left(y_j = 1 \middle| \mathbf{X}, \mathbf{Y}\right) = p(y_j = 1|\mathbf{x}_j)$$

donde $\mathbf{X}, \mathbf{Y}$ corresponden al conjunto completo de información disponible, tanto para la variable dependiente como para las independientes, no se cumple. Esto ocurre por no reconocer que las observaciones j provienen realmente de realizaciones de individuos i, a través del tiempo t, las cuales pueden ser, eventualmente, dependientes entre sí. Haciendo este cambio de variable, la ecuación anterior se transforma en

$$p(y_{it} = 1|\mathbf{X}, \mathbf{Y}) = p(y_{it} = 1|\mathbf{x}_{it})$$

En otras palabras, se impone que la probabilidad de incumplimiento de un individuo en un instante del tiempo cualquiera, no depende de lo que ese mismo individuo haga

en otro momento. Lo que se argumentará en los siguientes párrafos es que éste es un supuesto demasiado restrictivo en el ámbito de la construcción de *scorings* de comportamiento.

En efecto, un analista querría utilizar todas las observaciones disponibles en la base de datos para calibrar sus modelos. Esto, ya que un cliente presenta características dinámicas en el tiempo, y que podrían explicar de cierta manera su propensión a caer en default en el futuro. Por ejemplo, a medida que un cliente paga sus obligaciones, la proporción pagada del crédito cambia, partiendo de un 0% y llegando a un 100%, si cancela todo el crédito. Esta variable podría, en principio, explicar la propensión a caer en default de un cliente.

Sin embargo, considerar todas las observaciones de un cliente causa problemas de independencia. En efecto, considérese un cliente que nunca cae en default a lo largo de relación con el banco. Debido a que la mayor parte de la población paga sus créditos, este caso es el más frecuente en la base de datos. La Tabla 4 muestra un caso similar al mostrado anteriormente, pero en que el cliente sí paga correctamente sus obligaciones. Se mostrará que incluso sin recurrir a la variable de días de mora, uno puede deducir el valor para la variable de marca de incumplimiento en 12 meses para un gran número de fechas, conocidos el valor de ésta en dos puntos.

En efecto, observe que se sabe que en enero de 2007 y de 2008, el cliente tiene una marca de default en 12 meses igual a cero. Que en enero de 2007 sea igual a cero, implica que en los próximos doce meses (desde febrero de 2007 a enero de 2008) la variable días de mora puede tomar sólo valores menores a 90. Algo similar ocurre al conocer que la marca de transición en enero de 2008 es igual a cero: se puede deducir que los días de mora entre febrero de 2008 y enero de 2009 son menores que 90. En resumen, sabiendo que la marca es igual a cero en enero de 2007 y en enero de 2008, se sabe de inmediato que los días de mora entre febrero de 2008 y enero de 2009 son todos menores a 90. Por lo tanto, de esto se concluye que todas las marcas de default entre febrero de 2007 y diciembre de 2007 tendrán que ser igual a cero.

Tabla 4. Ejemplo de comportamiento de un cliente que paga su crédito.

Id	Fecha	Deuda	Días de mora	En default?	Transita a default en 12 meses?
001	200701	1.000.000	0	No	0
001	200702	950.000	0	No	?
001	200703	900.000	0	No	?
001	200704	850.000	0	No	?
001	200705	800.000	0	No	?
001	200706	750.000	0	No	?
001	200707	700.000	0	No	?
001	200708	650.000	0	No	?
001	200709	600.000	0	No	?
001	200710	550.000	0	No	?
001	200711	500.000	0	No	?
001	200712	450.000	0	No	?
001	200801	400.000	0	No	0

Fuente: elaboración propia.

El argumento anterior muestra que si se utilizaran todos los pares (cliente, fecha) en la base de datos para calibrar modelos como los escogidos en este trabajo, se estaría violando el supuesto de independencia de las observaciones. Esto ya que conociendo el valor de la marca de incumplimiento en dos fechas separadas por 12 meses, el valor de la variable queda completamente determinado dentro de ese lapso de tiempo.

Tal como se mencionó anteriormente, en el apartado siguiente se mostrará por qué no corresponde aplicar directamente los modelos de regresión logística para datos de panel (estudiados en la sección Marco Conceptual) a las estructuras de datos en cuestión.

3.1.1.1. Dificultades en la aplicación de modelos logísticos para datos de panel

El caso más simple de estimación de regresiones logísticas con datos panel corresponde a la "estimación agrupada". Esta considera que los datos utilizados cumplen con

$$p(y_{it} = 1|\mathbf{x}_{it}, y_{it-1}, \mathbf{x}_{it-1}, y_{it-2} \dots) = p(y_{it} = 1|\mathbf{x}_{it})$$

Aquí se mantiene la terminología utilizada en el marco teórico. La ecuación anterior implica que para conocer la probabilidad de que el individuo "decida" caer a default en los 12 meses siguientes a *t*, las observaciones pasadas de la variable de transición a deterioro (y_{it-1}, y_{it-2}, etc.) no aportan información, y que toda la información útil puede resumirse en el vector de datos $\mathbf{x}_{it}$. Este supuesto no se cumple en contexto de bases de cierre de mes, por la siguiente razón: si se sabe que $y_{it-1} = 0$ (nuevamente, el caso más probable), entonces se sabe que los siguientes 12 meses presentarán días de mora estrictamente menores a 90. Por lo tanto, al tratar de estimar y_{it}, conociendo y_{it-1}, sólo faltará determinar qué ocurrirá con los días de mora en *t+12*, ya que el resto de los meses estarán determinados. Esto significa que y_{it-1} sí aporta información, y que por lo tanto el supuesto representado en la ecuación anterior no es válido en este tipo de estructura de datos.

Ahora bien, el modelo logístico de efectos no observados asume el siguiente supuesto

$$p(y_{it} = 1|\mathbf{x}_i, c_i) = p(y_{it} = 1|\mathbf{x}_{it}, c_i)$$

Esto indica que sólo la información en *t* es suficiente para estimar la probabilidad de ocurrencia de $y_{it} = 1$, por lo que sólo $\mathbf{x}_{it}$ es suficiente y no toda la matriz $\mathbf{x}_i$. Vale la pena recalcar que $\mathbf{x}_i$ contiene toda la información observada para el individuo *i* en todos los periodos de observación (por lo tanto, contiene la información en *t*, anterior a *t* y posterior a este periodo). Pero ya se ha mostrado que conocer los días de mora en que incurrirá el individuo en el futuro aporta información con respecto a la estimación de la probabilidad buscada, lo que se deriva directamente de la forma de construir la

variable y_{it} a partir de los días de mora futuros. Consecuentemente, se concluye que no es posible tampoco utilizar este tipo de regresión logística para datos de panel.

Finalmente, el modelo logístico dinámico de efectos no observables asume la siguiente afirmación

$$p(y_{it} = 1|y_{it-1}, \dots, y_{i0}, \mathbf{z}_i, c_i) = \Lambda(\boldsymbol{\delta}^T \mathbf{z}_{it} + p y_{it-1} + c_i)$$

Si bien este modelo reconoce el aporte que ofrece a la estimación utilizar los rezagos de y_{it} (evidenciado por la presencia del término y_{it-1} en la ecuación), se asume nuevamente que la información futura no aporta a la estimación, ya que sólo aparece $\mathbf{z}_{it}$, y no la matriz completa de información del individuo $\mathbf{z}_i$. Es decir, el modelo muestra el mismo problema que en el caso del modelo logístico de efectos no observados.

Expuestos los problemas asociados con la estimación tanto de árboles CHAID como de regresiones logísticas utilizando bases de datos de cierre de mes, se presentan en la próxima sección tres métodos alternativos de selección de datos que permiten el cumplimiento de los supuestos de independencia en las observaciones utilizadas para estimar los modelos.

3.2. Métodos de selección de datos para asegurar la independencia de las observaciones[3]

Como se expuso anteriormente, al considerar todas las observaciones para un cliente a lo largo de la base de datos para calibrar un modelo de regresión logística se incurre en errores metodológicos. Más aún, según lo expuesto en (23), el supuesto de independencia es comúnmente violado en aplicaciones de regresiones logísticas a datos de comportamiento crediticio. Por otro lado, este mismo documento señala la relevancia de validar los supuestos teóricos de un modelo de probabilidad de incumplimiento.

[3] Se agradecen los aportes de Gabriel Soto, Victor Medina y Cristián Urbina en el diseño de estos algoritmos.

En vista de lo anterior, se propone evaluar el desempeño de tres métodos que asegure la independencia de las observaciones, basado en la extracción de *m* subconjuntos de datos de la base original. Estos subconjuntos tendrán a lo más una observación por cada cliente, lo que indica que se cumplirá con los supuestos de independencia de las observaciones, por lo que podrá utilizarse directamente la Estimación de Máxima Verosimilitud de los parámetros de la regresión. Esto significa que se tendrán *m* estimaciones del vector $\boldsymbol{\beta}$, las que se denominarán $\{\widehat{\boldsymbol{\beta}}_i\}_{1\leq i\leq m}$. Finalmente, se propone utilizar como estimador de $\boldsymbol{\beta}$ el promedio de los $\widehat{\boldsymbol{\beta}}_i$ sobre los *m* sets de datos. En otras palabras:

$$\widehat{\boldsymbol{\beta}} = \frac{1}{m}\sum_{i=1}^{m}\widehat{\boldsymbol{\beta}}_i$$

Considerando que los subconjuntos de datos se generarán aleatoriamente siguiendo un mismo algoritmo y a partir de una misma base de datos, y que el método de estimación de parámetros que se aplicará en cada uno de ellos será idéntico, se puede postular que los valores $\hat{\beta}_{ki}$ (el parámetro estimado en la muestra *i* asociado al atributo *k*) serán variables aleatorias independientes e idénticamente distribuidas, para cada atributo *k*. Por tanto, se tendrá que

$$\hat{\beta}_k = \frac{1}{m}\sum_{i=1}^{m}\hat{\beta}_{ki}$$

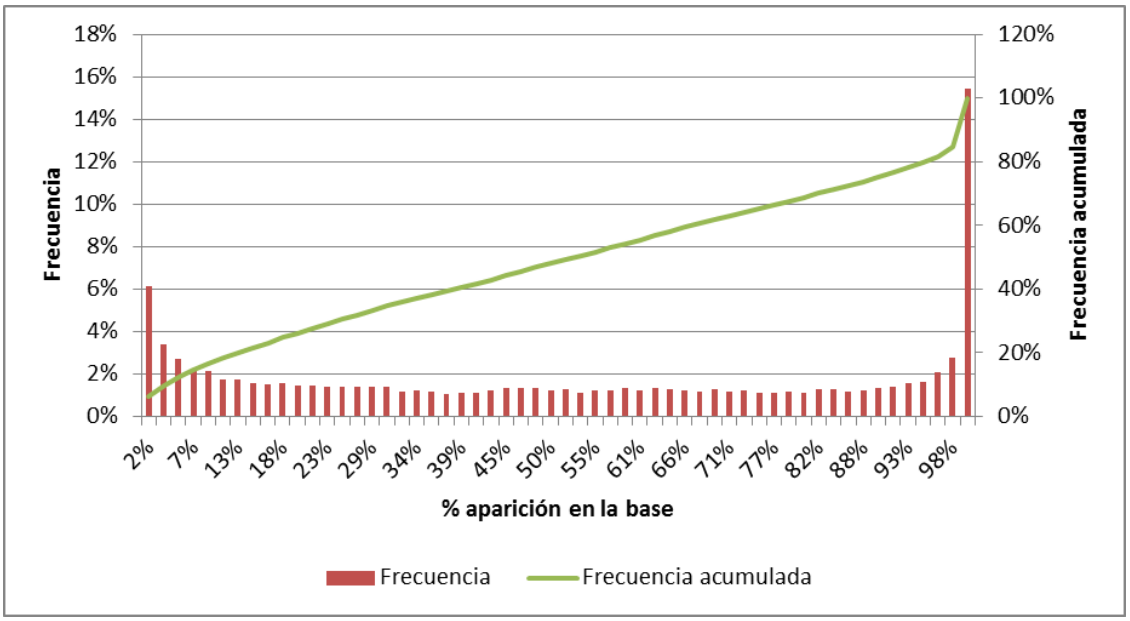
según el teorema del límite central (12), tendrá una distribución asintóticamente normal, con una media estimada $\hat{\beta}_k$, por lo que es insesgado, y varianza estimada igual a

$$\widehat{Var}(\hat{\beta}_k) = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\beta}_{ki} - \hat{\beta}_k\right)^2$$

En el marco de la teoría de muestreo presentada en la sección 2.6, la construcción de *credit scorings* se caracteriza por tener costos de muestreo iguales, cercanos a cero

(extracción de información desde una base de datos), y por tener distinta varianza a través de distintos sub grupos de la base de datos, ya sea a través de las distintas fechas, o a través de la cantidad de veces que los clientes aparecen. Además, se sabe que la mayor cantidad de observaciones en la base de datos está asociada a clientes que aparecen en todos los meses en estudio.

Figura 8. Distribución de "veces en la base".



Fuente: elaboración propia.

Debido a que el costo de obtener datos de una base es igual a cero, se tendrá que el problema se reduce a escoger qué observación de cada cliente se utilizará, de todas las disponibles, pero siempre asegurando que sea a lo más una. Los tres algoritmos propuestos tienen, cada uno, formas distintas de escoger y agrupar esas observaciones.

A continuación se entregan los detalles de cada algoritmo de extracción de datos. Estos asumen que se parte con una base de datos con observaciones de N clientes en T instantes de tiempo para k atributos distintos. Los algoritmos 1 y 2 se programaron en Visual Basic para Excel 2007. Este programa recibe como input una matriz de tamaño N x T. En la celda (c,t), la matriz posee un 1 si el clientes c es observado en el periodo t y un 0 en otro caso. Esta matriz se genera a su vez con una consulta SQL.

Por otro lado, la salida de este programa es una matriz de tamaño *N* x *m*, donde en la celda (*c,i*), la matriz tiene una fecha (en formato aaaa-mm) si el cliente *c* fue seleccionado para la muestra *i*, y un 0 en otro caso. Utilizando esta última matriz será posible generar *m* tablas distintas correspondientes a los *m* sub conjuntos de datos propuestos.

3.2.1. Algoritmo 1

1. Escoger un número de subconjuntos de datos *m*.
2. Para cada cliente *c* en la base de datos:
    a. Generar vector $\mathbf{F}_c$ con las fechas en que el cliente *c* es observado, de largo n_c.
    b. Para cada subconjunto *i=1…m*:
        i. Generar un número aleatorio *t* entre 1 y n_c.
        ii. Incluir la observación (c, F_{ct}) de la base en el subconjunto de datos *i*.
        iii. Avanzar al siguiente subconjunto de datos. Si *i=m*, terminar.
3. Escribir salida del algoritmo y terminar.

Se puede apreciar que en este marco, la probabilidad de que una observación (*c,t*) aparezca en un subconjunto de datos es $\pi_{ct} = 1/T_c$, donde T_c corresponde al número de meses que el cliente *c* aparece en la base. Esta probabilidad es mínima para las observaciones asociadas a clientes con $T_c = T$, y es máxima para observaciones de clientes con $T_c = 1$.

3.2.2. Algoritmo 2

1. Escoger un número de subconjuntos de datos *m*.
2. Para cada cliente *c* en la base de datos:
    a. Para cada subconjunto *i=1…m*:
        i. Generar un número aleatorio *t* entre 1 y *T*.

ii. Si el cliente es observado en *t*, incluir la observación (*c,t*) de la base en el subconjunto de datos *i*. Si no, no incluir al cliente en el subconjunto *i*.
iii. Avanzar al siguiente subconjunto de datos. Si *i*=*m*, terminar.

3. Escribir salida del algoritmo y terminar.

En este caso, $\pi_{ct} = 1/T$, constante e igual para todas las observaciones.

3.2.3. Algoritmo 3

Este algoritmo propone utilizar cada mes de la base datos como sub conjuntos de datos donde luego se entrenarán los modelos. Por lo tanto, en este caso $m = T$. En este enfoque, todas las observaciones de la base de datos son utilizadas.

3.2.4. Algoritmo 4

En este caso, se parte de una base más amplia, con la hipótesis de que, para cada cliente, todas las observaciones que tienen más de 12 meses de separación son independientes. Esto abre la posibilidad de desechar menos datos que en los métodos anteriores. Así, un método basado en este enfoque debiera resolver el siguiente problema de optimización: sea *F* el conjunto de fechas en donde un cliente existe, $\mathcal{P}(F)$ el conjunto de partes de F, y sea $h \in \mathbb{N}$:

$$\max_{S \in \mathcal{P}(F)} Card(S) \quad s.a \quad d(i,j) \geq h \; \forall i > j \in S$$

La solución de este problema entregará un $S \subseteq F$, tal que todos los elementos dentro de él están a una distancia mayor o igual a *h*. La solución a este problema existe, pues el conjunto factible es no vacío, ya que un *singleton* siempre es solución. Este problema puede ser expresado como un problema de programación lineal entera, de la siguiente forma

$$\max_{\{x_i\}_{i\in F}} \sum_{i\in F} x_i$$

$$s.a \quad M\left(2 - x_i - x_j\right) \geq h - d_{ij} \quad \forall i > j \in F$$

$$x_{ij} \in \{0,1\}$$

donde *M* es un número suficientemente grande. La solución de este PPL de una manera eficiente, de modo de poder aplicarlo a cada uno de los clientes de una base de datos, asegurando, además, que se tenga un solución única, no está dentro de los alcances de este proyecto, por lo que se deja planteado como un posible desarrollo futuro.

Otra solución factible consiste en tomar un punto *i* cualquiera, y luego incorporar todos los puntos *j* de *F* que cumplan la condición

$$d(i,j)\%h = 0$$

donde "%" es el operador módulo. Esto asegura que todos elementos que cumplan esta condición estén a una distancia mayor a *h* entre ellos. Sin embargo, no es óptimo, pues descarta algunos puntos en casos en que los clientes no aparecen a lo largo de toda la base de datos. Por ejemplo, supóngase que se fija $h = 12$ y un cliente existe entre el mes 1 y 13 (incluidos), luego desaparece, volviendo a aparecer en el mes 26 para después desaparecer en el mes 30. Si se parte del mes 13, el algoritmo seleccionará el subconjunto {1,13}, descartando todos los demás datos, siendo que aún podría incluir cualquier observación entre los meses 26 y 30 (incluidos).

La situación anterior es común en bases de datos de tarjetas de crédito, en el que el uso de ésta es esporádico, por lo que es frecuente observar que a clientes que mantienen saldo insoluto igual a cero, hasta que realizan una compra, pagan las obligaciones que esta genera, y luego vuelven a mostrar saldo cero. Sin perjuicio de lo anterior, esta solución sería una buena aproximación a lo que podría ser el óptimo, si la base estuviera dominada por clientes que aparecen a lo largo de toda la ventana de observación, y además es simple de implementar.

En base a lo anterior, se implementará un algoritmo que se inicia con un set de datos construido a partir del algoritmo 1 (que por lo tanto, contiene exactamente una

observación por cada cliente, aleatoria), sobre el cual se agregan observaciones adicionales, asegurando que éstas cumplan con la condición antes mencionada.

3.2.5. Algoritmo Base

Se referirá al método de calibrar una regresión logística utilizando todo el conjunto de datos de entrenamiento como Algoritmo Base. Este será el método contra el que se compararán el resto de los algoritmos propuestos.

3.2.6. Características de los métodos propuestos

En primer lugar, se aprecia que ya que el algoritmo 1 toma con probabilidad 100% las observaciones de clientes con una aparición, se tiene que el algoritmo está diseñado asumiendo que la varianza en este tipo de clientes es mayor (en el marco de muestreo estratificado con asignación óptima). De hecho, la probabilidad de inclusión de una observación es decreciente con el número de veces que aparece el cliente asociado en la base, por lo que se estaría asumiendo que la varianza también tiene el mismo comportamiento. El tamaño esperado de estos subconjuntos de datos será

$$n = \sum_{h \in H} N_h \pi_h = \sum_{h=1}^{T} \frac{N_h}{h}$$

en donde la partición se escoge como la cantidad de veces que aparece el cliente en la base. Esto resulta en *T* estratos, de tamaño N_h cada uno.

En segundo lugar, se observa que el algoritmo 2 asigna una probabilidad de selección constante para cada unidad de la base de datos, igual al inverso del número de fechas disponibles en la base. Utilizando el enfoque de MAS, se puede hacer un símil con la probabilidad de escoger una observación en ese marco ($\pi_i = n/N$):

$$\pi_{ct} = \frac{\bar{n}}{N} = \frac{1}{\frac{N}{\bar{n}}} = \frac{1}{\frac{T\bar{n}}{\bar{n}}} = \frac{1}{T}$$

donde $\bar{n}$ corresponde al número promedio de observaciones por fecha. Con esta interpretación, se estaría asumiendo que el algoritmo 2 corresponde a tomar un MAS de tamaño $\bar{n}$, modificado para que cada aparezca a lo más una observación por cliente.

La cantidad de subconjuntos de datos que se extraerán para el caso de los algoritmos 1 y 2 serán 50. Los demás métodos no se basan en extracciones repetitivas al azar. El algoritmo 3 queda limitado por el número de fechas observables. Finalmente, tanto para el algoritmo 4 como para el Base se hará sólo una extracción de datos.

Capítulo 4: Implementación de modelos

"In God we trust, all others bring data."

-William Edwards Deming[4].

4.1. Elección de métodos de clasificación a utilizar

Con respecto al método que se utilizará para desarrollar el modelo de probabilidades de incumplimiento, se construyó una tabla comparativa con los principales atributos de los procedimientos más utilizados (en base a (31), y (16)).

La consultora CLGroup recomienda que los modelos de probabilidad de incumplimiento sean fáciles de comprender, de manera que los clientes (instituciones financieras) puedan tener un acercamiento real a la solución propuesta. Por otro lado, se comentó que tanto los modelos de regresión logística, como los de árboles de clasificación, son estándares en la industria chilena en este ámbito, siendo ya avalados en el pasado por los organismos supervisores (SBIF). Además, poseen un amplio reconocimiento dentro de las gerencias de bancos. Es por estas razones que se decide incluir estas herramientas en el desarrollo de la metodología de este trabajo. En particular, la regresión logística se utilizará para crear el modelo de probabilidad de incumplimiento, mientras que el árbol se utilizará para el análisis univariado y la tramificación de variables. Se escogerá específicamente el algoritmo de construcción CHAID, pues este algoritmo se encuentra ya programado en el software SPSS, el cual es, según la consultora CLGroup, ampliamente utilizado en el mundo financiero.

Si bien es cierto que la literatura muestra un gran desempeño en términos de clasificación para métodos como los SVM o las redes neuronales (2), no es menos cierta la reconocida dificultad asociada a ellos para poder representar el conocimiento adquirido por éstos, especialmente en el caso de las redes neuronales (15). De hecho, ninguno de estos dos métodos es compatible con los requerimientos de BIS-II.

[4] Los autores de (18) indican que, irónicamente, no les fue posible encontrar evidencia de que efectivamente Deming haya sido el que enunció tal frase.

Sin perjuicio de lo anterior, se incluirá la red neuronal a modo de *benchmark* para la regresión logística.

Tabla 5. Comparación entre métodos de *credit scoring*.

	Test estadísticos de parámetros	Manejo de restricciones	Interacciones entre atributos	Actualización continua	Facilidad de Comprensión	Abundante uso
Análisis discriminante			Sí			Sí
Regresión lineal	Sí				Sí	Sí
Regresión probit	Sí				Sí	
Regresión logística	Sí				Sí	Sí
Programación lineal		Sí				Sí
Redes neuronales			Sí			Sí
Árboles de clasificación			Sí		Sí	Sí
K-Vecinos más cercanos				Sí	Sí	Sí
Sistemas expertos			Sí		Sí	
Algoritmos genéticos			Sí			

Fuente: elaboración propia.

4.2. Construcción de modelos de *scoring*

Para la implementación de los modelos, se utilizó la base de datos de cierre de mes de la cartera de consumo de una institución bancaria chilena. Esta cuenta con observaciones de las operaciones de sus clientes entre enero de 2006 y agosto de 2010. Algunas de las variables incluidas en esta base son:

- Id del cliente
- Id de la operación
- Fecha de proceso
- Fecha de otorgamiento de la operación
- Deuda actual
- Deuda original
- Días de mora
- Fecha de nacimiento del cliente
- Tipo de producto
- Naturaleza del cliente
- Estado civil
- Protestos del sistema
- Marca de renegociado
- Garantías
- Cuotas totales
- Cuotas pagadas
- Valor de la cuota
- Cupos y montos utilizados de tarjetas y líneas de crédito.

Esta base de datos se consolidó por cliente según lineamientos de BIS-II, pues el riesgo que interesa medir es del individuo y no de cada operación. Además, se crearon variables a partir de las disponibles, principalmente a través de ratios y de rezagos.

4.2.1. Filtro de datos

Para la construcción de los modelos, se aplicó una política de eliminación de datos, que se describe a continuación:

- Registros con deuda activa inferior a $2.000.
- Registros con fecha igual o posterior a septiembre de 2009, pues no se cuenta con 12 meses de datos para observar su incumplimiento.

4.2.2. Pre selección y tramificación de variables

El proceso se inicia con la pre selección de las variables a considerar. El objetivo es encontrar un listado abordable de inputs con el mejor poder predictivo posible. Basados en la literatura y en la experiencia del Banco en torno al conocimiento de la cartera estudiada, y a la vez limitados por los datos disponibles, se escogió el conjunto de variables a evaluar.

Luego se seleccionaron los factores que presentaron mayor poder de descripción univariada, mediante un ranking por valor del estadístico chi-cuadrado que entrega el algoritmo CHAID mediante la segmentación óptima de variables para explicar la transición a deterioro en 12 meses.

La anterior segmentación se implementa para obtener tramos óptimos (según criterio CHAID), tanto de variables continuas como categóricas. Estos tramos se denominarán *dummies*. Algunos de los beneficios del uso de *dummies* son (31):

- captura comportamientos no lineales,
- robusto frente a *outliers*,
- permite correlacionar variables categóricas,
- se enfoca en el poder predictivo más que en su poder descriptivo.

Utilizando el ranking chi-cuadrado es posible realizar un filtro inicial de eliminación de *dummies* con alta correlación. El algoritmo para realizar este filtro es el siguiente:

1. Generar matriz de correlaciones entre las *dummies*.
2. Para cada variable *i* en el ranking univariado,
    a. Identificar el conjunto de *dummies* D_i asociadas a la variable i.
    b. Si existen correlaciones mayores a 50% dentro del conjunto D_i, eliminar las *dummies* con mayor número de conflictos con otras *dummies*, hasta que no existan correlaciones mayores a 50% dentro del conjunto D_i.
    c. Eliminar las *dummies* cuya correlación con algún elemento de D_i sea mayor a un 50%.
    d. Avanzar en el ranking.
3. Terminar.

Los resultados del análisis univariado, del procedimiento de tramificación y del filtro de correlaciones pueden encontrarse en las secciones 7.1, 7.2 y 7.3, respectivamente. La siguiente fase consiste en el proceso que involucra combinar los inputs transformados en un modelo multivariado.

4.2.3. Selección de variables

Una vez que se cuenta con los 50 sub conjuntos de datos en donde se sabe que cada cliente aparece exactamente una vez, es posible llevar a cabo los métodos de selección de variables.

El siguiente paso consiste en llevar a cabo selecciones de datos utilizando algoritmos de selección de variables del estilo *wrapper* implementados sobre regresiones logísticas. La consultora propuso trabajar con el algoritmo *forward selection* con criterio de razón de verosimilitud (implementado en SPSS). La ventaja de este algoritmo es que proporciona conjuntos de variables más estables (18).

El procedimiento anterior genera 50 sets de *dummies* filtradas. El set final de *dummies* a considerar se obtuvo imponiendo que cada una hubiese sido seleccionada en al menos 90% de los sets de datos, con el fin de asegurar consistencia.

4.2.4. Análisis multivariado

Teniendo el sets de *dummies* final, se procede a calibrar el modelo con el 70% de los datos en los 50 sets, lo que entrega 50 vectores de parámetros. Se calculan la media y la desviación estándar de cada parámetro, y se genera un intervalo de confianza al 95%. Si dicho intervalo contiene al cero, la variable se descarta. Además de la estimación sobre las 50 muestras, el modelo se calibra también tanto en el primer como en el último mes. Esto sirve para identificar variables cuya asociación con la variable dependiente no sea constante en el tiempo, por lo que se descartarán las *dummies* cuyo signo cambie en ambas fecha, ya que se busca construir modelos estables en el tiempo. Eliminadas todas las variables conflictivas, se vuelve a estudiar la estabilidad de las *dummies*, hasta que ninguna de las condiciones sea violada.

Los arreglos de variables seleccionadas para cada método se encuentran disponibles en la sección 7.4.

4.2.5. Validación

El proceso descrito en el párrafo anterior termina con un modelo multivariado de probabilidad de incumplimiento que estable a través del tiempo y a través de distinto conjuntos seleccionados aleatoriamente de la base de datos.

Finalmente, se estiman estadísticos de poder de discriminancia con el 30% de los clientes no utilizados para el entrenamiento del modelo, de manera de obtener una validación fuera de la muestra. El cálculo tanto del estadístico KS como AUC se realiza a través del tiempo, pues de esta forma las observaciones son independientes (un mismo cliente sólo aparece una vez por mes), y además permite observar una evolución en el comportamiento del modelo.

4.3. Resultados

El cuadro a continuación muestra para cada método de selección de datos, la cantidad de sub conjuntos que se utilizó, el tamaño promedio de estos, la proporción de observaciones de incumplimiento, y la cantidad de variables seleccionadas utilizando cada uno, a través del procedimiento descrito en 4.2.4.

Tabla 6. Características de los conjuntos de datos de entrenamiento.

Método	N° cjtos.	Tamaño promedio	Proporción malos prom.	N° variables
Algoritmo 1	50	19.063	2,55%	12
Algoritmo 2	50	10.738	2,01%	5
Algoritmo 3	50	10.733	2,05%	4
Algoritmo 4	1	57.038	2,08%	25
Base	1	407.844	2,07%	39

Fuente: elaboración propia.

En primer lugar, es importante notar la gran diferencia entre el número de variables seleccionadas a lo largo de los 5 métodos estudiados. Esto se debe, probablemente, al considerable mayor tamaño de las muestras de los algoritmos 4 y base, lo que permite

la estimación de modelos más complejos de manera robusta. En segundo lugar, se puede apreciar que el algoritmo 4, que se construye a partir de una de las muestras del algoritmo 1, logra casi triplicar el tamaño de ella, agregando observaciones que cumplen los supuestos de independencia. Finalmente, se puede ver que los sets de datos construidos con el método 1 tienen una proporción de malos mayor a la observada en la base.

La tabla siguiente muestra el promedio de los estadísticos KS y AUC a lo largo del tiempo, para la base de validación. En ella se puede apreciar que, en primer lugar, el método que arroja un mayor KS promedio corresponde al Base. Sin embargo, el Algoritmo 4 tiene una diferencia de apenas 0,5%, pero entrega la desviación estándar más baja de todas. Con respecto al AUC, en este caso el método base obtiene tanto el mejor promedio como la menor desviación, aunque las diferencias con el algoritmo 4 sigan siendo pequeñas.

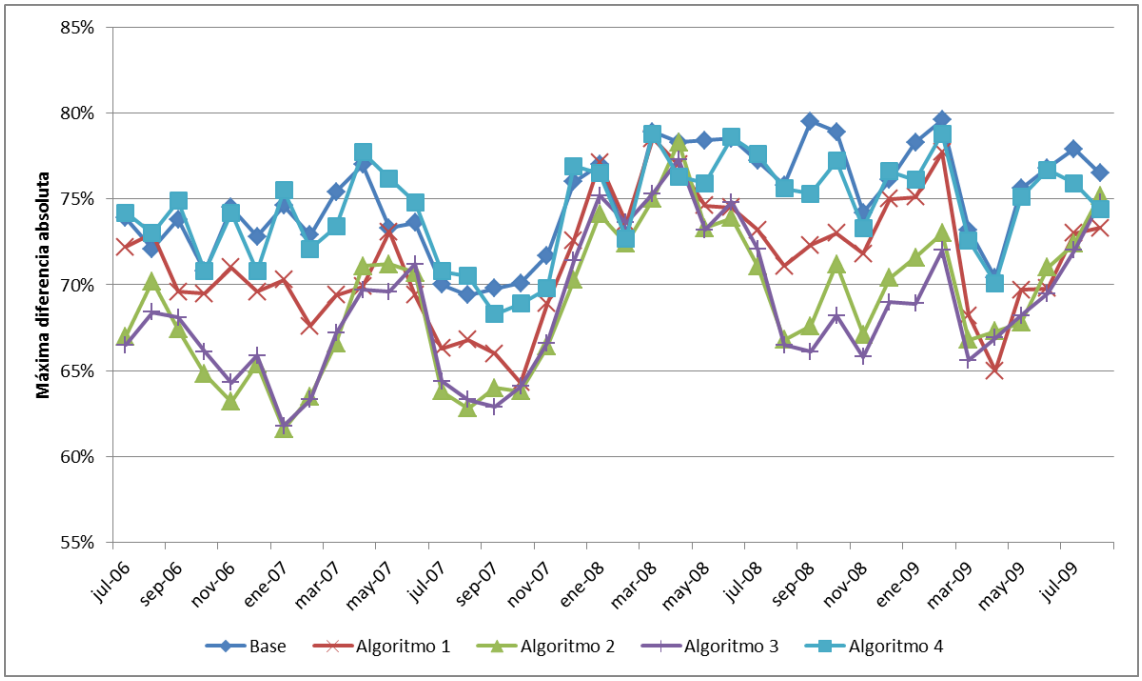
Tabla 7. Resumen estadísticos de discriminancia en set de validación.

	KS		AUC	
Método	**Promedio**	**Desv. Est.**	**Promedio**	**Desv. Est.**
Algoritmo 1	71,39%	3,49%	91,58%	1,78%
Algoritmo 2	68,95%	4,06%	87,92%	1,71%
Algoritmo 3	68,68%	3,98%	87,00%	2,07%
Algoritmo 4	74,39%	2,85%	93,11%	1,27%
Base	74,89%	3,02%	93,55%	1,20%

Fuente: elaboración propia.

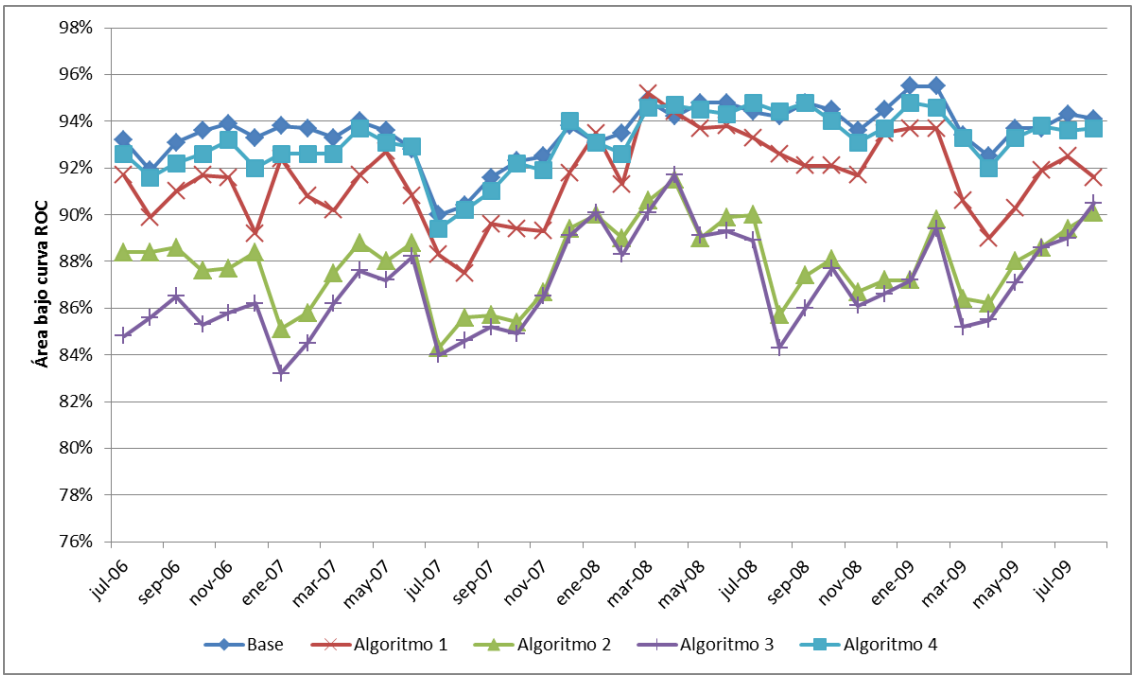
La Figura 9 muestra la evolución del KS a lo largo del tiempo para el set de validación. Aquí también se aprecia con claridad que los peores resultados se obtienen con los Algoritmos 1, 2 y 3; de hecho, los dos últimos muestran una evolución bastante similar. La Figura 10, por otro lado, muestra la evolución del área bajo la curva ROC (AUC) para cada mes de observación. Aquí también es posible apreciar la superioridad de los algoritmos 4 y base frente al resto de los métodos.

Figura 9. Evolución KS en base de validación.



Fuente: elaboración propia.

Figura 10. Evolución AUC en base de validación.



Fuente: elaboración propia.

Para descartar que las diferencias en los desempeños de los dos métodos que mejor se comportan se deba a una mala especificación funcional de modelo, se entrenaron y validaron redes neuronales para los algoritmos 4 y base, de manera de estudiar su desempeño en ese marco. Para este efecto se utiliza el paquete "*Neural Net"* del software *Rapidminer,* con su configuración por defecto. Las variables de entrenamiento que se utilizan son las mismas que se usan para calibrar la regresión logística final de cada método. Los resultados se muestran a continuación:

Tabla 8. Resultados de redes neuronales en set de validación.

	KS		AUC	
Método	**Promedio**	**Desv. Est.**	**Promedio**	**Desv. Est.**
Algoritmo 4	69,46%	2,92%	91,16%	1,47%
Base	71,08%	3,04%	91,66%	1,39%

Fuente: elaboración propia.

Es posible apreciar que bajo ambos estadísticos, el método superior es el Base con respecto al promedio. En desviación, gana en AUC, pero pierde en KS (aunque por poco). Es decir, se mantiene prácticamente el mismo comportamiento que con la regresión logística.

Descartadas ya otras opciones para explicar la diferencia en los desempeños de los métodos, queda estudiar las diferencias en los sets de datos creados a través de cada mecanismo. La Tabla 6 muestra que la mayor diferencia entre las muestras creadas con cada mecanismo tiene relación con el tamaño promedio de estos. De hecho, cruzando esta información con la de la Tabla 7, se observará una relación entre desempeño, tanto en KS como en AUC, con el tamaño promedio de las muestras generadas con cada método: a mayor tamaño, mejor desempeño. Una explicación probable para este efecto es que al contar con más observaciones, es posible estimar modelos más complejos de manera robusta. Visto de otra forma, si se cuenta con pocos datos, es difícil que se pueda estimar un modelo muy complejo (con muchos parámetros) de manera robusta.

Capítulo 5: Conclusiones

> *"I understand that my work may have enormous effects on society and the economy, many of them beyond my comprehension."*
>
> Derman, E. y Wilmott, P., *Financial Modelers' Manifesto* (2009)[5].

El objetivo principal de este proyecto fue diseñar y evaluar la conveniencia de una metodología que permitiese hacerse cargo del cumplimiento del supuesto de independencia de las observaciones en la estimación de regresiones logísticas para *credit scorings* de comportamiento. Para lograr esto, se hizo, en primer lugar, un estudio del estado del arte en el área, corroborándose que el problema antes mencionado está descrito en la literatura, pero que, a la vez, no se ha estudiado suficientemente como para haberse propuesto metodologías de solución (23).

En segundo lugar, se diseñaron 4 métodos de selección de datos que generan muestreos con distintas características, pero con observaciones independientes, controlándose las veces y las condiciones que deben cumplir los datos asociados para cada cliente de manera de ser incluidos en una muestra. Estos métodos se compararan en su desempeño contra la alternativa tradicional de utilizar todas las observaciones disponibles para la construcción de modelos.

En tercer lugar, se preparó una base de datos de una cartera *retail* perteneciente a una institución bancaria, para la implementación de los métodos diseñados a través de regresiones logísticas. Para esto, se realizó un filtro de datos, se tramificaron las variables disponibles y se filtraron algunas *dummies* creadas por tener correlaciones sobre 50%.

Finalmente, se implementaron regresiones logísticas utilizando cada uno de los métodos de selección de datos propuestos. Los resultados de estos cálculos arrojan que la mayor diferencia entre los sets creados por cada mecanismo recae en el tamaño promedio de éstos. Esto, a su vez, provoca que el algoritmo de selección de variables

[5] Ver http://www.wilmott.com/blogs/paul/index.cfm/2009/1/8/Financial-Modelers-Manifesto.

(*forward selection*) incluya más variables en los sets de datos de mayor tamaño, pues un mayor número de observaciones permite la estimación de modelos más complejos (con más parámetros). Lo anterior se refleja en que los mejores desempeños tanto en KS como en AUC, y tanto en promedio (mayor) como en desviación estándar (menor), se obtengan con los algoritmos que generan sets de datos más grandes; esto es, los métodos 4 y base.

La conclusión anterior se mantiene al comparar el desempeño de los dos algoritmos con mejores resultados (4 y base) en la construcción de redes neuronales. El método con el set de mayor cantidad de observaciones (base) es el que logra mejores resultados.

Se concluye, finalmente, que la adecuación para cumplir el supuesto de independencia de las regresiones logísticas, y de en general cualquier mecanismo de clasificación estadística de patrones, entrega pocos beneficios, y que además estos no son concluyentes. Es más, la restricción excesiva del número de observaciones a considerar puede conllevar consecuencias indeseadas, como un bajo poder de discriminación y una alta variabilidad en las predicciones, debido a la omisión de variables relevantes. Por otro lado, el algoritmo 4 logra prácticamente el mismo desempeño que el algoritmo base; sin embargo, el primero cuenta con 14 *dummies* menos que el segundo, por lo que lo convierte un modelo preferible en cuanto a su parsimonia.

Sin perjuicio de lo anterior, el marco de construcción de regresiones logísticas implementado (filtro de correlaciones, análisis univariado y multivariado), entregó resultados de discriminancia bastante positivos, encontrándose KS superiores al 74% en las muestras de validación, con una relativamente alta estabilidad a lo largo de la ventana de observación disponible. Esto constituye un subproducto relevante de este trabajo, dado que existe una demanda por modelos cada vez más precisos y que permitan automatizar más partes de la cadena productiva en instituciones financieras.

En resumen, el cumplimiento del supuesto de independencia no asegura que los modelos estimados tengan mejores desempeños. Sí permite tener modelos con bajo número de variables, manteniendo un desempeño comparable a la situación base,

cuando se utiliza la mayor cantidad de datos posible (por ejemplo, con el algoritmo 4). Así, estos métodos de selección de datos deberán ser utilizados sólo en caso de que se quiera mejorar la parsimonia de los modelos, o en el caso de que, por ejemplo, el ente regulador bancario exija el cumplimiento del supuesto de independencia.

5.1. Desarrollos futuros

Si bien es cierto que la separación entre datos de clientes para entrenamiento y para validación tiene por objetivo remover completamente la dependencia entre ambos conjuntos, de manera de obtener estimadores fidedignos de la capacidad de discriminación de los modelos, es posible que permanezca una interdependencia temporal no controlada. En otras palabras, aunque los clientes de ambas bases sean completamente distintos, puede que de todas maneras muestren un comportamiento correlacionado a través del tiempo. La forma de controlar este efecto necesariamente pasa por eliminar meses completos de datos de la base de entrenamiento, de manera de generar una separación temporal con la base de validación. Sin embargo, la cantidad de periodos disponibles (38 meses) no permite realizar este tipo de estudios, ya que haber trabajado con menos de 3 años para la calibración va en contra de los lineamientos de BISII, que promueven la inclusión de al menos un ciclo económico completo en la construcción de este tipo de modelos. Queda propuesto como un desarrollo futuro un estudio que incorpore esta separación entre bases de entrenamiento y validación.

Otro posible desarrollo posterior de este proyecto, podría consistir en la utilización del Algoritmo 4 en un contexto de estimación al estilo *bootstrap*, realizando varios muestreos para obtener distintos sets de datos, tal como se hace con los algoritmos 1 y 2, lo que resultaría en un mejor aprovechamiento de los datos en la base.

Otros trabajos futuros podrían basarse en la solución eficiente del PPL entero planteado en 3.2.4, llevada a cabo simultáneamente para un conjunto de clientes arbitrariamente grande, maximizando la cantidad de datos que se utilice a la vez que se asegura la independencia.

Una posible solución al excesivo número de variables que toma el algoritmo base podría aumentar el nivel de significancia exigido al mecanismo de *forward selection*, para así obtener un sub conjunto de variables de menor tamaño.

Capítulo 6: Bibliografía

[1] ANDERSON, Raymond. The Credit Scoring Toolkit. 1ª ed. New York, Oxford University Press Inc., 2007. 731p.

[2] BAESENS, Bart "et al". Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, 54(6):627-635, 2003.

[3] BAESENS, Bart y VAN GESTEL, Tony. Credit Risk Management: Basic Concepts. 1ª ed. USA, Oxford University Press, 2009. 500p.

[4] BCBS. Principles for the Management of Credit Risk. Basilea, Suiza, Bank for International Settlements, 2000.

[5] BCBS. International Convergence on Capital Measures and Standards. Comprehensive version. Basilea, Suiza, Bank for International Settlements, 2006.

[6] BLÖCHLINGER, Andreas y LEIPPOLD, Markus. Economic Benefit of Powerful Credit Scoring. Working Paper No. 216. National Centre of Competence in Research. Suiza, 2005. 42p.

[7] BLUHM, Christian, OVERBECK, Ludger y WAGNER, Christoph. An introduction to credit risk modeling. 1ª ed. Boca Raton, Florida, Chapman & Hall/CRC, 2002. 297p.

[8] BRAVO, Cristián, THOMAS, Lyn y WEBER, Richard. Strategic Clustering: A Semi-Supervised Method for Improving Credit Scoring by Differentiating Defaulters. Management Science.

[9] BURGES, Christopher. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2(2):121-167, 1998.

[10] CROUHY, Michel, GALAI, Dan y MARK, Robert M. The Essentials of Risk Management. 1ª ed. [S.l.], McGraw-Hill, 2005.

[11] DE SERVIGNY, Arnaud y RENAULT, Olivier. Measuring and managing credit risk. 1ª ed. New York, McGraw-Hill Companies, Inc., 2004. 466p.

[12] DEGROOT, Morris H. y SCHERVISH, Mark J. Probability and Statistics. 4ª ed. Boston, Massachusetts, Addison Wesley, 2011. 912p.

[13] FAYYAD, Usama, PIATETSKY-SHAPIRO, Gregory y SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3):37-54, 1996.

[14] GIESECKE, Kay. Credit risk modeling and valuation: an introduction. Paper of School of Operations Research and Industrial Engineering. Cornell University. New York, 2004. 67p.

[15] HAN, Jiawei y KAMBER, Micheline. Data Mining: Concepts and Techniques. 2ª ed. San Francisco, Morgan Kaufmann, 2005. 800p.

[16] HAND, D.J. y HENLEY, W.E. Statistical classification methods in consumer credit scoring: a review. Journal of the Royal Statistical Society, 160(3):523-541, 1997.

[17] HÄRDLE, Wolfgang, MORO, Rouslan y SCHÄFER, Dorothea. Estimating Probabilities of Default with Support Vector Machines. Discussion Paper Series 2: Banking and Financial Studies. Deutsche Bundesbank, Research Centre. Berlin, 2007. 21p.

[18] HASTIE, Trevor, TIBSHIRANI, Robert y FRIEDMAN, Jerome. The Elements of Statistical Learning. 2ª ed. California, Springer, 2009. 746p.

[19] HOSMER, David W. y LEMESHOW, Stanley. Applied Logistic Regression. 2ª ed. Newark, NJ, Wiley, 2000.

[20] KASS, Gordon. An exploratory technique for investigating large quantities of categorical data. Journal of the Royal Statistical Society, 29(2):119-127, 1980.

[21] KING, G. y ZENG, L. Logistic Regression in Rare Events Data. Political Analysis, 9(2):137-163, 2001.

[22] LOHR, Sharon L. Sampling: Design and Analysis. 2ª ed. Boston, Cengage Learning, 596p.

[23] MEDEMA, L., KONING, H. R. y LENSINK, R. A practical approach to validating a PD model. Journal of Banking & Finance, 33(4):701-708, abr. 2009.

[24] RICHARD, Michael D. y LIPPMAN, Richard P. Neural network classifiers estimate Bayesian a posteriori probabilities. Neural Computation, 3(4):461-483, 1991.

[25] SBIF. ¿Qué es SBIF? Sitio Web SBIF. [en línea] <http://www.sbif.cl>. [consulta: 24 Mayo 2011]

[26] SBIF. Compendio de Normas Contables. Sitio Web Sbif. [en línea] <http://www.sbif.cl/>. [consulta: 24 Mayo 2011]

[27] SOTO, Gabriel. Heurística de Estimación de Pérdidas Financieras Extremas. Tesis (Magister en Gestión de Operaciones). Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas. Santiago, Chile, 2007. 123p.

[28] SPSS INC. SPSS Classification Trees™ 13.0. Irlanda, SPSS Inc., 2004.

[29] STEIN, Roger. Benchmarking default prediction models: pitfalls and remedies in model validation. Journal of Risk Model Validation, 1(1):77-113, 2007.

[30] THOMAS, Lyn. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. Journal of Forecasting, 16(2):149-172, 2000.

[31] THOMAS, Lyn, EDELMAN, David y CROOK, Jonathan. Credit Scoring and Its Applications. 1ª ed. Philadelphia, SIAM, 2002. 248p.

[32] VIJAY, S. D. Loss forecasting for consumer loan portfolios. Credit Scoring and Credit Conference IX, Edinburgh: [s.n.]. 8 de septiembre de 2005.

[33] WEBB, Andrew. Statistical Pattern Recognition. 2ª ed. West Sussex, John Wiley & Sons, Ltd., 2002. 534p.

[34] WOOLDRIDGE, Jeffrey M. Econometric Analysis of Cross Section and Panel Data. 1ª ed. Cambridge, The MIT Press, 2001. 752p.

Capítulo 7: Anexos

7.1. Análisis univariado

Tabla 9. Ranking chi-cuadrado de variables.

Variable	Descripción	Chi-cuadrado	Variable	Descripción	Chi-cuadrado
max_dias_mora	Días de mora	55.631	avance_cuotas_avg	Cuotas pagadas (prom)	3.344
cant_prot	Protestos	38.267	cuotas_tot_orig_avg	Cuotas totales (prom.)	3.318
veces_mora_dura_6m	Veces en mora >30 6 meses	35.629	cupo_disp_tc	Cupo disponible t.c.	3.266
max_mora_6m	Max. Mora 6 meses	32.997	ratio_deuda_tc	Deuda tarjeta / total	3.238
tasa_uso_disponible	% uso línea y tarjeta	15.047	mnt_orig_min	Monto otorgado (min.)	3.052
tasa_uso_lc	% uso línea	12.235	ratio_deuda_lc	Deuda línea / total	3.021
renegociado	Renegociado	11.158	val_cuota_total	Valor de la cuota total	2.847
cant_lc	N° operaciones línea	11.086	max_deu_6m	Máxima deuda 6m	2.758
leverage	Cuota / renta	9.443	nro_op	N° operaciones	2.609
tasa_uso_tc	% uso tarjeta	6.568	avance_deuda_max	% deuda pagada (máx.)	2.546
cupo_disp_lc	Cupo disponible línea	6.376	mnt_ut_lc	Utilizado línea	2.515
crecimiento_deuda_6m	% crecimiento deuda 6m	5.855	avance_deuda_avg	% deuda pagada (prom.)	2.454
RELACION_PAGO	Relación de pago 6m	5.228	cuotas_tot_orig_min	Cuotas totales (min.)	2.301
avance_cuotas_min	Cuotas pagadas (mínima)	5.060	mnt_deuda_cred	Deuda crédito cons.	2.207
renta	Renta estimada	4.741	ratio_cuota_deuda	Cuota / deuda	2.011
mnt_orig_sum	Monto otorgado (total)	4.531	cant_cred	N° operaciones cred.	1.908
veces_sube_deu_6m	Veces sube deuda 6m	4.453	avance_deuda_min	% deuda pagada (min.)	1.773
mnt_deuda_lc	Deuda línea	4.424	mnt_ut_tc	Utilizado t.c.	1.733
deuda_6m_antes	Deuda 6 meses antes	4.381	segmento	Segmento	1.512
contingente_aprobado	Aprobado línea+t.c.	4.251	mnt_deuda_tc	Deuda tarjeta	1.005
mnt_deuda_total	Deuda total	4.186	cant_tc	N° operaciones t.c.	776
mnt_orig_max	Monto otorgado (máx.)	4.166	edad_cl	Edad	535
avance_cuotas_max	Cuotas pagadas (máxima)	3.971	cuota_hip	Cuota hipotecario	328
antig_op_min	Antig. Operación (min.)	3.957	sub_segmento	Sub segmento	36
antig_op_max	Antig. Operación (máx.)	3.776	est_civ	Estado civil	30
cuotas_tot_orig_max	Cuotas totales (máx.)	3.558	tipo_cl	Naturaleza del cliente	27
antig_op_avg	Antig. Operación (prom.)	3.542			

Fuente: elaboración propia.

7.2. Tramificación de variables

Se muestra a continuación el código de SPSS para la tramificación de las variables involucradas en la implementación de los modelos.

```
COMPUTE dummy_deuda_tramo_1 = (mnt_deuda_total<=82662).
COMPUTE dummy_deuda_tramo_2 = (mnt_deuda_total>82662 & mnt_deuda_total<=216098).
COMPUTE dummy_deuda_tramo_3 = (mnt_deuda_total>216098 & mnt_deuda_total<=447262).
COMPUTE dummy_deuda_tramo_4 = (mnt_deuda_total>447262 & mnt_deuda_total<=792812).
COMPUTE dummy_deuda_tramo_5 = (mnt_deuda_total>792812 & mnt_deuda_total<=1310646).
```

```
COMPUTE dummy_deuda_tramo_6 = (mnt_deuda_total>1310646 & mnt_deuda_total<=7259659).
COMPUTE dummy_deuda_cred_tramo_1 = (mnt_deuda_cred<=0).
COMPUTE dummy_deuda_cred_tramo_2 = (mnt_deuda_cred>0 & mnt_deuda_cred<=1517347).
COMPUTE dummy_deuda_cred_tramo_3 = (mnt_deuda_cred>1517347 & mnt_deuda_cred<=4918287).
COMPUTE dummy_deuda_lc_tramo_1 = (mnt_deuda_lc<=132).
COMPUTE dummy_deuda_lc_tramo_2 = (mnt_deuda_lc>132 & mnt_deuda_lc<=20835).
COMPUTE dummy_deuda_lc_tramo_3 = (mnt_deuda_lc>20835 & mnt_deuda_lc<=115369).
COMPUTE dummy_deuda_lc_tramo_4 = (mnt_deuda_lc>115369 & mnt_deuda_lc<=439343).
COMPUTE dummy_deuda_lc_tramo_5 = (mnt_deuda_lc>439343 & mnt_deuda_lc<=987851).
COMPUTE dummy_deuda_tc_tramo_1 = (mnt_deuda_tc<=0).
COMPUTE dummy_deuda_tc_tramo_2 = (mnt_deuda_tc>0 & mnt_deuda_lc<=44433).
COMPUTE dummy_deuda_tc_tramo_3 = (mnt_deuda_tc>44433 & mnt_deuda_lc<=152716).
COMPUTE dummy_deuda_tc_tramo_4 = (mnt_deuda_tc>152716 & mnt_deuda_lc<=330896).
COMPUTE dummy_deuda_tc_tramo_5 = (mnt_deuda_tc>330896 & mnt_deuda_lc<=585009).
COMPUTE dummy_deuda_tc_tramo_6 = (mnt_deuda_tc>585009 & mnt_deuda_lc<=1094991).
COMPUTE dummy_mnt_orig_tramo_1 = (mnt_orig_sum<=87530).
COMPUTE dummy_mnt_orig_tramo_2 = (mnt_orig_sum>87530 & mnt_orig_sum<=234210).
COMPUTE dummy_mnt_orig_tramo_3 = (mnt_orig_sum>234210 & mnt_orig_sum<=485368).
COMPUTE dummy_mnt_orig_tramo_4 = (mnt_orig_sum>485368 & mnt_orig_sum<=890872).
COMPUTE dummy_mnt_orig_tramo_5 = (mnt_orig_sum>890872 & mnt_orig_sum<=1550995).
COMPUTE dummy_mnt_orig_tramo_6 = (mnt_orig_sum>1550995 & mnt_orig_sum<=2841887).
COMPUTE dummy_mnt_orig_tramo_7 = (mnt_orig_sum>2841887 & mnt_orig_sum<=5000959).
COMPUTE dummy_mnt_orig_tramo_8 = (mnt_orig_sum>5000959 & mnt_orig_sum<=9436540).
COMPUTE dummy_tiene_mora = (max_dias_mora>0).
COMPUTE dummy_tiene_cred = (cant_cred>0).
COMPUTE dummy_cant_lc_tramo_1 = (cant_lc<=0).
COMPUTE dummy_cant_lc_tramo_2 = (cant_lc>0 & cant_lc<=1).
COMPUTE dummy_cant_lc_tramo_3 = (cant_lc>1 & cant_lc<=2).
COMPUTE dummy_cant_tc_tramo_1 = (cant_tc<=0).
COMPUTE dummy_cant_tc_tramo_2 = (cant_tc>0 & cant_tc<=1).
COMPUTE dummy_cl_soltero_separado = (est_civ='Soltero' | est_civ='Separad').
COMPUTE dummy_tiene_prot = (cant_prot>0).
COMPUTE dummy_renegociado = renegociado.
COMPUTE dummy_cuotas_max_tramo_1 = (cuotas_tot_orig_max<=1).
COMPUTE dummy_cuotas_max_tramo_2 = (cuotas_tot_orig_max>1 & cuotas_tot_orig_max<=23).
COMPUTE dummy_cuotas_max_tramo_3 = (cuotas_tot_orig_max>23 & cuotas_tot_orig_max<=36).
COMPUTE dummy_val_cuota_tramo_1 = (val_cuota_total<=19057).
COMPUTE dummy_val_cuota_tramo_2 = (val_cuota_total>19057 & val_cuota_total<=67626).
COMPUTE dummy_val_cuota_tramo_3 = (val_cuota_total>67626 & val_cuota_total<=146040).
COMPUTE dummy_val_cuota_tramo_4 = (val_cuota_total>146040 & val_cuota_total<=263384).
COMPUTE dummy_val_cuota_tramo_5 = (val_cuota_total>263384 & val_cuota_total<=443763).
COMPUTE dummy_val_cuota_tramo_6 = (val_cuota_total>443763 & val_cuota_total<=708581).
COMPUTE dummy_val_cuota_tramo_7 = (val_cuota_total>708581 & val_cuota_total<=1101372).
COMPUTE dummy_val_cuota_tramo_8 = (val_cuota_total>1101372 & val_cuota_total<=1872201).
COMPUTE dummy_tasa_uso_disponible_tramo_1 = (tasa_uso_disponible<=0).
COMPUTE dummy_tasa_uso_disponible_tramo_2 = (tasa_uso_disponible>0 &
tasa_uso_disponible<=0.048).
COMPUTE dummy_tasa_uso_disponible_tramo_3 = (tasa_uso_disponible>0.048 &
tasa_uso_disponible<=0.116).
COMPUTE dummy_tasa_uso_disponible_tramo_4 = (tasa_uso_disponible>0.116 &
tasa_uso_disponible<=0.242).
COMPUTE dummy_tasa_uso_disponible_tramo_5 = (tasa_uso_disponible>0.242 &
tasa_uso_disponible<=0.432).
COMPUTE dummy_tasa_uso_disponible_tramo_6 = (tasa_uso_disponible>0.432 &
tasa_uso_disponible<=0.708).
COMPUTE dummy_tasa_uso_disponible_tramo_7 = (tasa_uso_disponible>0.708 &
tasa_uso_disponible<=0.956).
COMPUTE dummy_avance_cuotas_min_cero = (avance_cuotas_min<=0).
```

```
COMPUTE dummy_antig_op_min_cero = (antig_op_min<=0).
COMPUTE dummy_edad_tramo_1 = (edad_cl<=29.397).
COMPUTE dummy_edad_tramo_2 = (edad_cl>29.397 & edad_cl<=35.816).
COMPUTE dummy_edad_tramo_3 = (edad_cl>35.816 & edad_cl<=45.147).
COMPUTE dummy_edad_tramo_4 = (edad_cl>45.147 & edad_cl<=53.314).
COMPUTE dummy_edad_tramo_5 = (edad_cl>53.314 & edad_cl<=59.815).
COMPUTE dummy_nro_op_tramo_1 = (nro_op<=1).
COMPUTE dummy_nro_op_tramo_2 = (nro_op>1 & nro_op<=2).
COMPUTE dummy_nro_op_tramo_3 = (nro_op>2 & nro_op<=3).
COMPUTE dummy_nro_op_tramo_4 = (nro_op>3 & nro_op<=4).
COMPUTE dummy_nro_op_tramo_5 = (nro_op>4 & nro_op<=5).
COMPUTE dummy_tuvo_mora_6m = (max_mora_6m>0).
COMPUTE dummy_veces_sube_deu_6m_tramo_1=(veces_sube_deu_6m<=0).
COMPUTE dummy_veces_sube_deu_6m_tramo_2=(veces_sube_deu_6m>0 & veces_sube_deu_6m<=1).
COMPUTE dummy_veces_sube_deu_6m_tramo_3=(veces_sube_deu_6m>1 & veces_sube_deu_6m<=2).
COMPUTE dummy_veces_sube_deu_6m_tramo_4=(veces_sube_deu_6m>2 & veces_sube_deu_6m<=3).
COMPUTE dummy_renta_tramo_1 = (renta<=467379).
COMPUTE dummy_renta_tramo_2 = (renta>467379 & renta<=981052).
COMPUTE dummy_renta_tramo_3 = (renta>981052 & renta<=1888476).
COMPUTE dummy_renta_tramo_4 = (renta>1888476 & renta<=2366777).
COMPUTE dummy_renta_tramo_5 = (renta>2366777 & renta<=3861193).
COMPUTE dummy_renta_tramo_6 = (renta>3861193 & renta<=4412659).
COMPUTE dummy_renta_tramo_7 = (renta>4412659 & renta<=5999396).
COMPUTE dummy_seg_banca_pref_emp = (segmento='Banca Preferente' | segmento='Empresas').
COMPUTE dummy_seg_banca_pers_sininf = (segmento='Banca Personas' | segmento='SIN INF.').
COMPUTE dummy_rel_pago_tramo_1 = (RELACION_PAGO<=0.111).
COMPUTE dummy_rel_pago_tramo_2 = (RELACION_PAGO>0.111 & RELACION_PAGO<=0.348).
COMPUTE dummy_rel_pago_tramo_3 = (RELACION_PAGO>0.348 & RELACION_PAGO<=0.561).
COMPUTE dummy_rel_pago_tramo_4 = (RELACION_PAGO>0.561 & RELACION_PAGO<=0.723).
COMPUTE dummy_rel_pago_tramo_5 = (RELACION_PAGO>0.723 & RELACION_PAGO<=0.842).
COMPUTE dummy_rel_pago_tramo_6 = (RELACION_PAGO>0.842 & RELACION_PAGO<=0.932).
COMPUTE dummy_rel_pago_tramo_7 = (RELACION_PAGO>0.932 & RELACION_PAGO<=1.002).
COMPUTE dummy_rel_pago_tramo_8 = (RELACION_PAGO>1.002 & RELACION_PAGO<=1.035).
COMPUTE dummy_rel_pago_tramo_9 = (RELACION_PAGO>1.035 & RELACION_PAGO<=1.498).
COMPUTE dummy_crec_deu_6m_tramo_1 = (crecimiento_deuda_6m<=-0.695).
COMPUTE dummy_crec_deu_6m_tramo_2 = (crecimiento_deuda_6m>-0.695 &
crecimiento_deuda_6m<=-0.343).
COMPUTE dummy_crec_deu_6m_tramo_3 = (crecimiento_deuda_6m>-0.343 &
crecimiento_deuda_6m<=-0.145).
COMPUTE dummy_crec_deu_6m_tramo_4 = (crecimiento_deuda_6m>-0.145 &
crecimiento_deuda_6m<=-0.000).
COMPUTE dummy_crec_deu_6m_tramo_5 = (crecimiento_deuda_6m>-0.000 &
crecimiento_deuda_6m<=0.185).
COMPUTE dummy_crec_deu_6m_tramo_6 = (crecimiento_deuda_6m>0.185 &
crecimiento_deuda_6m<=0.710).
COMPUTE dummy_crec_deu_6m_tramo_7 = (crecimiento_deuda_6m>0.710 &
crecimiento_deuda_6m<=3.010).
COMPUTE dummy_crec_deu_6m_tramo_8 = (crecimiento_deuda_6m>3.010 &
crecimiento_deuda_6m<=64.391).
COMPUTE dummy_crec_deu_6m_tramo_9 = (crecimiento_deuda_6m>64.391 &
crecimiento_deuda_6m<=38794).
COMPUTE dummy_leverage_tramo_1 = (leverage<=0.012).
COMPUTE dummy_leverage_tramo_2 = (leverage>0.012 & leverage<=0.047).
COMPUTE dummy_leverage_tramo_3 = (leverage>0.047 & leverage<=0.410).
COMPUTE dummy_leverage_tramo_4 = (leverage>0.410 & leverage<=0.629).
COMPUTE dummy_leverage_tramo_5 = (leverage>0.629 & leverage<=1.020).
COMPUTE dummy_leverage_tramo_6 = (leverage>1.020 & leverage<=2.111).
COMPUTE dummy_ratio_deu_lc_tramo_1 = (ratio_deuda_lc<=0.000).
```

```
COMPUTE dummy_ratio_deu_lc_tramo_2 = (ratio_deuda_lc>0.000 & ratio_deuda_lc<=0.011).
COMPUTE dummy_ratio_deu_lc_tramo_3 = (ratio_deuda_lc>0.011 & ratio_deuda_lc<=0.186).
COMPUTE dummy_ratio_deu_lc_tramo_4 = (ratio_deuda_lc>0.186 & ratio_deuda_lc<=0.561).
COMPUTE dummy_ratio_deu_lc_tramo_5 = (ratio_deuda_lc>0.561 & ratio_deuda_lc<=0.978).
COMPUTE dummy_ratio_deu_tc_tramo_1 = (ratio_deuda_tc<=0.000).
COMPUTE dummy_ratio_deu_tc_tramo_2 = (ratio_deuda_tc>0.000 & ratio_deuda_tc<=0.082).
COMPUTE dummy_ratio_deu_tc_tramo_3 = (ratio_deuda_tc>0.082 & ratio_deuda_tc<=0.769).
COMPUTE dummy_ratio_deu_tc_tramo_4 = (ratio_deuda_tc>0.769 & ratio_deuda_tc<=0.999).
COMPUTE dummy_cont_aprob_tramo_1 = (contingente_aprobado<=299869).
COMPUTE dummy_cont_aprob_tramo_2 = (contingente_aprobado>299869 &
contingente_aprobado<=1393750).
COMPUTE dummy_cont_aprob_tramo_3 = (contingente_aprobado>1393750 &
contingente_aprobado<=1999968).
COMPUTE dummy_cont_aprob_tramo_4 = (contingente_aprobado>1999968 &
contingente_aprobado<=3661800).
COMPUTE dummy_cont_aprob_tramo_5 = (contingente_aprobado>3661800 &
contingente_aprobado<=4999998).
COMPUTE dummy_cont_aprob_tramo_6 = (contingente_aprobado>4999998 &
contingente_aprobado<=6999841).
COMPUTE dummy_cont_aprob_tramo_7 = (contingente_aprobado>6999841 &
contingente_aprobado<=10475179).
COMPUTE dummy_tiene_cred_hip = (cuota_hip>0).


EXECUTE.
```

7.3. Filtro de correlaciones

Tabla 10. *Dummies* filtradas por correlaciones

Dummy eliminada	Por
dummy_deuda_tramo_1	dummy_leverage_tramo_1
dummy_deuda_tramo_2	dummy_mnt_orig_tramo_2
dummy_deuda_tramo_3	dummy_mnt_orig_tramo_3
dummy_deuda_tramo_4	dummy_mnt_orig_tramo_4
dummy_deuda_tramo_5	dummy_mnt_orig_tramo_5
dummy_deuda_cred_tramo_1	dummy_cuotas_max_tramo_3
dummy_deuda_lc_tramo_1	dummy_cant_lc_tramo_1
dummy_deuda_lc_tramo_2	dummy_cant_lc_tramo_2
dummy_deuda_tc_tramo_1	dummy_ratio_deu_tc_tramo_1
dummy_deuda_tc_tramo_2	dummy_cant_lc_tramo_1
dummy_deuda_tc_tramo_3	dummy_ratio_deu_tc_tramo_1
dummy_deuda_tc_tramo_4	dummy_deuda_tc_tramo_6
dummy_deuda_tc_tramo_5	dummy_deuda_tc_tramo_6
dummy_mnt_orig_tramo_1	dummy_leverage_tramo_1
dummy_tiene_cred	dummy_cuotas_max_tramo_3
dummy_cant_lc_tramo_3	dummy_cant_lc_tramo_1
dummy_cant_tc_tramo_1	dummy_ratio_deu_tc_tramo_1
dummy_cuotas_max_tramo_1	dummy_cuotas_max_tramo_2
dummy_val_cuota_tramo_1	dummy_leverage_tramo_1
dummy_val_cuota_tramo_5	dummy_mnt_orig_tramo_3
dummy_val_cuota_tramo_6	dummy_mnt_orig_tramo_4
dummy_antig_op_min_cero	dummy_avance_cuotas_min_cero
dummy_nro_op_tramo_1	dummy_tasa_uso_disponible_tramo_1
dummy_seg_banca_pers_sininf	dummy_seg_banca_pref_emp
dummy_rel_pago_tramo_1	dummy_crec_deu_6m_tramo_1
dummy_ratio_deu_lc_tramo_1	dummy_cant_lc_tramo_1
dummy_ratio_deu_lc_tramo_2	dummy_cant_lc_tramo_2
dummy_cont_aprob_tramo_1	dummy_tasa_uso_disponible_tramo_1

Fuente: elaboración propia.

7.4. Variables seleccionadas e inferencia estadística

7.4.1. Algoritmo 1

Tabla 11. Variables seleccionadas - Algoritmo 1

Variable	Beta	Desv. Est.	I.C.(95%)-	I.C.(95%)+	200607	200908
dummy_tiene_mora	1,906	0,179	1,556	2,256	1,388	1,477
dummy_tiene_prot	2,073	0,107	1,864	2,282	1,571	2,286
dummy_tuvo_mora_6m	1,391	0,173	1,053	1,730	1,383	1,649
dummy_renta_tramo_1	1,010	0,099	0,816	1,203	1,746	0,501
dummy_renta_tramo_2	0,592	0,128	0,340	0,843	1,041	0,393
dummy_rel_pago_tramo_7	1,053	0,158	0,744	1,363	0,858	1,002
dummy_rel_pago_tramo_8	1,432	0,150	1,138	1,725	0,861	2,165
dummy_rel_pago_tramo_9	0,775	0,180	0,422	1,127	0,327	0,775
dummy_cant_lc_tramo_1	-1,621	0,241	-2,093	-1,148	-1,024	-1,629
dummy_avance_cuotas_min_cero	-1,112	0,298	-1,695	-0,529	-0,504	-2,021
dummy_renegociado	0,765	0,181	0,410	1,120	0,793	0,440
dummy_nro_op_tramo_2	-1,533	0,156	-1,838	-1,228	-1,285	-1,037
Constant	-3,761	0,338	-4,422	-3,099	-4,438	-3,720

Fuente: elaboración propia.

7.4.2. Algoritmo 2

Tabla 12. Variables seleccionadas - Algoritmo 2

Variable	Beta	Desv. Est.	I.C.(95%)-	I.C.(95%)+	200607	200908
dummy_tiene_mora	1,995	0,267	1,472	2,518	1,743	2,100
dummy_tiene_prot	2,257	0,160	1,944	2,570	1,863	2,510
dummy_tuvo_mora_6m	1,549	0,214	1,129	1,970	1,533	1,824
dummy_rel_pago_tramo_8	1,030	0,199	0,639	1,420	0,523	1,701
dummy_renta_tramo_1	0,669	0,151	0,373	0,965	1,389	0,406
Constant	-5,243	0,109	-5,457	-5,029	-4,983	-5,876

Fuente: elaboración propia.

7.4.3. Algoritmo 3

Tabla 13. Variables seleccionadas - Algoritmo 3

Variable	Beta	Desv. Est.	I.C.(95%)-	I.C.(95%)+	200607	200908
dummy_tiene_mora	2,057	0,246	1,575	2,539	1,935	2,125
dummy_tiene_prot	2,287	0,346	1,609	2,965	2,110	2,526
dummy_tuvo_mora_6m	1,547	0,291	0,976	2,118	1,572	1,812
dummy_rel_pago_tramo_8	1,030	0,352	0,339	1,720	0,699	1,707
Constant	-5,135	0,254	-5,633	-4,637	-4,690	-5,829

Fuente: elaboración propia.

7.4.4. Algoritmo 4

Tabla 14. Variables seleccionadas - Algoritmo 4

Variable	Beta	Desv. Est.	I.C.(95%)-	I.C.(95%)+	200607	200908
dummy_tiene_mora	1,885	0,096	1,697	2,073	1,199	1,540
dummy_cant_lc_tramo_1	-0,754	0,108	-0,966	-0,542	-0,625	-0,538
dummy_tiene_prot	1,825	0,079	1,670	1,980	1,348	2,002
dummy_renegociado	0,749	0,148	0,459	1,039	0,716	0,713
dummy_val_cuota_tramo_2	-0,596	0,186	-0,961	-0,231	-0,655	-0,954
dummy_tasa_uso_disponible_tramo_2	-1,008	0,188	-1,376	-0,640	-1,697	-2,781
dummy_tasa_uso_disponible_tramo_3	-1,785	0,385	-2,540	-1,030	-1,205	-2,108
dummy_tasa_uso_disponible_tramo_4	-1,465	0,299	-2,051	-0,879	-1,028	-1,394
dummy_tasa_uso_disponible_tramo_5	-0,373	0,163	-0,692	-0,054	-1,001	-1,132
dummy_tasa_uso_disponible_tramo_6	-0,419	0,147	-0,707	-0,131	-0,540	-0,591
dummy_nro_op_tramo_2	-0,911	0,110	-1,127	-0,695	-0,982	-0,647
dummy_tuvo_mora_6m	1,077	0,096	0,889	1,265	1,156	1,441
dummy_veces_sube_deu_6m_tramo_1	0,363	0,081	0,204	0,522	0,157	0,519
dummy_renta_tramo_1	1,452	0,139	1,180	1,724	3,744	0,747
dummy_renta_tramo_2	1,178	0,149	0,886	1,470	3,111	0,542
dummy_renta_tramo_3	0,847	0,139	0,575	1,119	2,630	0,579
dummy_renta_tramo_4	0,790	0,165	0,467	1,113	2,402	0,171
dummy_renta_tramo_5	0,577	0,149	0,285	0,869	1,983	0,182
dummy_rel_pago_tramo_6	0,511	0,113	0,290	0,732	0,617	0,757
dummy_rel_pago_tramo_7	0,956	0,105	0,750	1,162	0,867	1,030
dummy_rel_pago_tramo_8	1,502	0,114	1,279	1,725	0,972	2,205
dummy_rel_pago_tramo_9	1,136	0,130	0,881	1,391	0,522	0,837
dummy_crec_deu_6m_tramo_8	-0,682	0,217	-1,107	-0,257	-0,750	-0,284
dummy_crec_deu_6m_tramo_9	-1,018	0,260	-1,528	-0,508	-1,716	-1,065
dummy_tiene_cred_hip	-0,276	0,134	-0,539	-0,013	-0,609	-0,355
Constant	-5,503	0,147	-5,791	-5,215	-6,658	-5,738

Fuente: elaboración propia.

7.4.5. Algoritmo Base

Tabla 15. Variables seleccionadas - Algoritmo Base

Variable	Beta	Desv. Est.	I.C.(95%)-	I.C.(95%)+	200607	200908
dummy_tiene_mora	1,580	0,036	1,509	1,651	1,228	1,613
dummy_cant_lc_tramo_1	-0,675	0,041	-0,755	-0,595	-0,618	-0,500
dummy_tiene_prot	1,818	0,028	1,763	1,873	1,297	1,995
dummy_renegociado	0,539	0,056	0,429	0,649	0,719	0,731
dummy_cuotas_max_tramo_2	-0,147	0,036	-0,218	-0,076	-0,046	-0,252
dummy_val_cuota_tramo_2	-0,477	0,076	-0,626	-0,328	-0,621	-0,860
dummy_val_cuota_tramo_7	-0,144	0,044	-0,230	-0,058	-0,466	-0,111
dummy_tasa_uso_disponible_tramo_2	-0,752	0,072	-0,893	-0,611	-1,466	-2,394
dummy_tasa_uso_disponible_tramo_3	-1,637	0,132	-1,896	-1,378	-1,146	-2,024
dummy_tasa_uso_disponible_tramo_4	-0,857	0,085	-1,024	-0,690	-0,968	-1,355
dummy_tasa_uso_disponible_tramo_5	-0,708	0,066	-0,837	-0,579	-0,967	-1,090
dummy_tasa_uso_disponible_tramo_6	-0,542	0,052	-0,644	-0,440	-0,566	-0,604
dummy_edad_tramo_4	0,185	0,031	0,124	0,246	0,136	0,276
dummy_nro_op_tramo_2	-0,985	0,041	-1,065	-0,905	-0,966	-0,690
dummy_tuvo_mora_6m	1,123	0,034	1,056	1,190	1,080	1,390
dummy_veces_sube_deu_6m_tramo_3	-0,323	0,036	-0,394	-0,252	-0,234	-0,428
dummy_veces_sube_deu_6m_tramo_4	-0,319	0,045	-0,407	-0,231	-0,330	-0,577
dummy_renta_tramo_1	1,171	0,060	1,053	1,289	17,299	0,644
dummy_renta_tramo_2	1,035	0,060	0,917	1,153	16,787	0,498
dummy_renta_tramo_3	0,890	0,055	0,782	0,998	16,415	0,549
dummy_renta_tramo_4	0,752	0,063	0,629	0,875	16,181	0,170
dummy_renta_tramo_5	0,603	0,057	0,491	0,715	15,810	0,220
dummy_renta_tramo_7	0,177	0,093	-0,005	0,359	15,309	0,307
dummy_rel_pago_tramo_2	-0,266	0,086	-0,435	-0,097	-0,899	-0,703
dummy_rel_pago_tramo_6	0,329	0,041	0,249	0,409	0,455	0,606
dummy_rel_pago_tramo_7	0,664	0,039	0,588	0,740	0,711	0,886
dummy_rel_pago_tramo_8	1,212	0,042	1,130	1,294	0,821	2,080
dummy_rel_pago_tramo_9	0,659	0,050	0,561	0,757	0,389	0,838
dummy_crec_deu_6m_tramo_7	-0,259	0,054	-0,365	-0,153	-0,056	-1,025
dummy_crec_deu_6m_tramo_8	-0,577	0,074	-0,722	-0,432	-0,698	-0,356
dummy_crec_deu_6m_tramo_9	-0,607	0,097	-0,797	-0,417	-1,540	-1,112
dummy_leverage_tramo_1	-1,818	0,149	-2,110	-1,526	-15,692	-15,179
dummy_leverage_tramo_2	-0,703	0,087	-0,874	-0,532	-0,571	-0,534
dummy_leverage_tramo_3	-0,473	0,042	-0,555	-0,391	-0,300	-0,161
dummy_leverage_tramo_4	-0,359	0,052	-0,461	-0,257	-0,407	-0,218
dummy_leverage_tramo_5	-0,225	0,043	-0,309	-0,141	-0,357	-0,317
dummy_cont_aprob_tramo_6	-0,174	0,053	-0,278	-0,070	-0,193	-0,380
dummy_cont_aprob_tramo_7	-0,163	0,058	-0,277	-0,049	-1,460	-0,444
dummy_tiene_cred_hip	-0,463	0,049	-0,559	-0,367	-0,696	-0,445
Constant	-4,613	0,066	-4,742	-4,484	-19,798	-5,005

Fuente: elaboración propia.