Experiencias Prácticas en la Medición de Riesgo Crediticio de Microempresarios utilizando Modelos de Credit Scoring

Cristián Bravo*
Sebastián Maldonado*
Richard Weber*

Resumen

Todas las instituciones financieras que ofrecen crédito a sus clientes deben abordar el problema de estimar cuánto del dinero otorgado retornará a la entidad y a qué clientes ofrecerles crédito. Sistemas de Credit Scoring *se han desarrollado de manera exitosa para determinar la probabilidad que un cierto cliente falle en devolver el crédito que le ha sido otorgado. En el presente trabajo se describen los modelos desarrollados para dos organizaciones financieras nacionales para microempresarios, ajustando los pasos del proceso KDD (*Knowledge Discovery in Databases*) a sus necesidades particulares. El documento presenta las experiencias obtenidas a partir de estos proyectos y explica en detalle como se resolvieron los problemas asociados a las características particulares de los microempresarios en Chile. La mayoría de los proyectos asociados al proceso KDD son de naturaleza estática. Sin embargo, con el paso del tiempo los modelos comienzan a perder la capacidad de explicar los fenómenos para los que fueron construidos inicialmente. Debido a los requerimientos de las entidades financieras se desarrollaron módulos para el seguimiento y la recalibración de los modelos. En particular, se proponen técnicas estadísticas con el fin de determinar cuándo los cambios en las características de la población pueden afectar el desempeño del modelo. Durante el desarrollo de las soluciones se pudo obtener un conocimiento importante sobre el comportamiento de los clientes. Algunos descubrimientos fueron sorprendentes, mientras otros confirmaron las nociones que tenían los expertos. La utilización de estos sistemas en las operaciones diarias puede reducir la tasa tanto de falsos positivos como de falsos negativos, lo que se traduce en menores costos y una mayor cobertura en los mercados respectivos.*

Palabras Clave: *Credit scoring, Regresión logística, Microempresarios.*

*Departamento Ingeniería Industrial, Universidad de Chile

1. Introducción

En el escenario actual, los avances tecnológicos han permitido un desarrollo importante en la automatización de la decisión sobre la aceptación o rechazo de una solicitud de crédito mediante modelos analíticos, evitando el otorgamiento bajo criterios ambiguos, lo que en algunos países se considera una práctica ilegal.

Los modelos analíticos requieren de información cuantitativa potencialmente útil para su construcción. Si bien la posibilidad de obtener esta información es cada vez más simple, gracias al importante aumento de la capacidad de almacenaje y la disponibilidad de mejores herramientas para el manejo de datos, el proceso de extracción de información relevante a partir de los datos disponibles sigue siendo complejo y costoso. Las técnicas utilizadas para esta tarea se engloban bajo el concepto de Minería de Datos (*data mining*).

La modelación de la falla financiera, tanto en personas como en empresas, ha sido un problema altamente estudiado en la literatura. Desde el comienzo de los años sesenta, de acuerdo a los trabajos de Beaver [3] y Altman [1], se han desarrollado modelos matemáticos y estadísticos que buscan predecir el desempeño que tendría una persona si se le otorgase crédito mediante la asignación de un puntaje estimado a partir de la información del cliente. Este problema se conoce como *Credit Scoring* [13].

Si bien los modelos de Credit Scoring han sido ampliamente estudiados en la literatura, su aplicación al segmento de los microempresarios no es directa debido a que éstos representan un grupo diferente en relación al resto de los clientes, ya sea en términos de tamaño, ingresos o estructura social.

El presente trabajo se enfoca en el problema de medición de riesgo crediticio de microempresarios mediante modelos de Credit Scoring, resumiendo los resultados y experiencias obtenidas en dos proyectos, uno para una entidad financiera gubernamental [6] y otro para una institución privada, en los últimos siete años.

La estructura de este trabajo es la siguiente: La sección 2 define el marco teórico y muestra los principales avances en la modelación de la problemática asociada a Credit Scoring, destacando el desafío que representa el segmento de los microempresarios al momento de efectuar Credit Scoring tradicional. El desarrollo del modelo propuesto, junto con algunas estrategias para el seguimiento y recalibración del modelo se presentan en la sección 3. La sección 4 presenta los principales resultados del trabajo. Finalmente, la sección 5 muestra las conclusiones del trabajo.

2. Credit Scoring Aplicado a Microempresarios

Los microempresarios en Chile representan un sector importante de la economía, los cuales presentan características especiales que los hacen diferentes a las empresas que se estudian comúnmente en Credit Scoring tradicional. Estas particularidades se describen en esta sección. Adicionalmente, este marco teórico introduce los conceptos asociados al problema de Credit Scoring, junto con la descripción del proceso dentro del cual se encuentran insertas las herramientas de clasificación. Este proceso conoce como KDD (*Knowledge Discovery in Databases*, [7]). Se coloca especial énfasis en las etapas de preparación de los datos, selección de atributos y seguimiento de los modelos que son los elementos centrales de este artículo.

2.1. Microempresarios en Chile

En Chile, un microempresario se define como una empresa muy pequeña, con un máximo de nueve trabajadores, un ingreso por ventas mensuales promedio no superior a los 200 UF o unos US$ 8.000 (para el caso de los programas de capacitación y asesoría, el umbral de ventas mensuales se disminuye a 150 UF o unos US$ 6.000) y cuentan con activos fijos menores a 500 UF o unos US$ 20.000, de acuerdo a la definición utilizada por el Fondo de Solidaridad e Inversión Social (FOSIS, www.fosis.cl). Esta definición es la más utilizada en nuestro país, puesto que muchas otras instituciones la usan como referencia.

Los microempresarios representan un pilar fundamental de la economía nacional, ya que el 81 % de las 707.634 empresas formales existentes el año 2004 pertenecen a esta categoría, porcentaje que presenta una escasa variación en la última década. Sin embargo, esto no se ve reflejado en las ventas, ya que este sector representa solamente el 3.4 % de la participación de las ventas totales de este año [5].

Si bien las microempresas consideran un conjunto de negocios con alto grado de heterogeneidad, es posible describirlas en función de ciertas características comunes [5]:

- Por lo general corresponden a negocios familiares o trabajadores auto empleados.
- Representan organizaciones con bajos rendimientos, generalmente ineficientes en el abastecimiento de materias primas, comercialización, manejo contable y financiero.
- son de gestión conservadora y salarios bajos. Como promedio la venta de las microempresas chilenas es de 456 UF anuales, o sea, US$16.000

aproximadamente, monto muy pequeño que genera inestabilidad a sus empleados y propietarios.

Si bien las microempresas se encuentran presentes en prácticamente todas las actividades económicas, su mayor representación está en aquellos sectores de menor potencialidad y mayores barreras de entrada, como es el caso del comercio. Las cifras indican que el 77 % de las microempresas se concentra en cuatro sectores económicos: comercio, servicios, transporte y agricultura [5].

El mercado de los microempresarios, usualmente apoyado financieramente por gobiernos e iniciativas de la Unión Europea, se ha transformado en un negocio atractivo para bancos y otras instituciones crediticias. Sin embargo, el mercado presenta características de riesgo únicas que no han sido abordadas por los modelos de riesgo tradicionales, surgiendo la necesidad de crear modelos ad-hoc y atrayendo el interés de tanto investigadores como corporaciones privadas. En particular, los microempresarios chilenos presentan ciertas cualidades que deben ser tenidas en cuenta al momento de desarrollar modelos de Credit Scoring, tales como:

- Los microempresarios usualmente tienen un presupuesto limitado, debido a su menor ingreso. Debido a esto, la variable ingreso, que es un candidato natural para formar parte de los modelos de riesgo, suele presentar una escasa capacidad discriminativa.
- Existe un limitado conocimiento de las variables que los caracterizan, siendo necesario un estudio detallado de sus características y necesidades, con el fin de ofrecerles crédito de manera responsable y sin caer en prácticas discriminatorias que nacen de la incertidumbre.

Debido a estas razones, las técnicas de Credit Scoring tradicionales deben ser adaptadas con el objetivo de reflejar la realidad presentada y crear las condiciones adecuadas tanto para ellos como para las mismas instituciones financieras. Esta es una preocupación tanto de los gobiernos como de las instituciones privadas, y esta experiencia busca entregar resultados aplicables para ambos.

2.2. Definición del problema

Hasta hace no mucho tiempo, la decisión de entregar créditos se basaba en el juicio humano para determinar el riesgo de no pago del postulante a crédito en base a los atributos de éste. Sin embargo, el crecimiento de la demanda por crédito ha llevado a desarrollar métodos formales y objetivos para ayudar a los proveedores del crédito a decidir a quién otorgar crédito y a quién no. Este enfoque fue introducido en los años 40 y con los años se ha desarrollado significativamente. En los años recientes, la alta competencia de la industria

financiera, los avances en la computación y el crecimiento exponencial del tamaño de las bases de datos han llevado a estos métodos a transformarse en una importante herramienta en la industria.

Credit Scoring se define formalmente como un método cuantitativo que se utiliza para predecir la probabilidad de que un aspirante a crédito o un cliente de la entidad crediticia existente deje de pagar el crédito o bien no lo haga una vez que lo reciba [11]. Su objetivo es ayudar a los proveedores de créditos a cuantificar y manejar el riesgo financiero relacionado con el otorgamiento de créditos, para así tomar decisiones de forma rápida y objetiva.

Credit Scoring tiene múltiples beneficios que incumben no sólo a las entidades crediticias, sino también a los beneficiarios del crédito. Por ejemplo, Credit Scoring ayuda a reducir la discriminación porque provee un análisis objetivo del mérito del postulante para recibir un crédito. Esto les permite a los proveedores enfocarse sólo en la información relacionada con la asignación del crédito y así evitar subjetividad. Cuando se le niega un crédito a un cliente en los Estados Unidos, la *Equal Credit Opportunity Act* exige a la institución financiera proveer las razones de por qué fue rechazado. Razones vagas o indefinidas son ilegales, por lo que variables que puedan llevar a discriminación tales como raza, sexo o religión no pueden ser incluidas en estos modelos [11].

Credit Scoring ayuda también a acelerar y a hacer más consistente el proceso de asignación de créditos, permitiendo su automatización. Esto reduce significativamente la necesidad de intervención humana y por ende los costos asociados a este proceso. Más aún, Credit Scoring puede ayudar a las instituciones financieras a determinar la tasa de interés que deben cobrar a sus clientes y para valorizar portafolios [14]. A clientes con mayor riesgo se les cobra una tasa de interés más alta. Esto ayuda a la entidad a manejar sus cuentas de manera más efectiva y provechosa en términos de utilidades.

Finalmente y gracias a los avances de la tecnología, se han desarrollado modelos para Credit Scoring más efectivos. En consecuencia, entidades crediticias utilizan esta información generada para formular mejores estrategias de cobranza y utilizar sus recursos más eficientemente. En particular, Credit Scoring ayuda a empresas aseguradoras a realizar una mejor predicción de las reclamaciones, controlar el riesgo de manera efectiva y determinar el precio de los seguros de manera adecuada. Esto les permite ofrecer mayor cobertura a más clientes a un precio equitativo, reaccionar rápido ante los cambios del mercado y obtener ventajas competitivas.

El problema principal que se aborda corresponde a definir si un cliente que presenta características $\mathbf{X}$ va a caer en una situación de falla financiera dentro de un futuro cercano y no devolverá íntegramente el crédito otorgado. Para ello, es necesario contar con características que sean relevantes para el estudio y que permitan medir el fenómeno. En particular, se busca encontrar aquel vector $\mathbf{X}$ de características tal que permita predecir la probabilidad de ocurrencia de un

fenómeno binario y, en este caso si el cliente falla en la devolución del crédito, con un margen de error razonable. Matemáticamente, podemos expresar el objetivo según la ecuación 1.

$$p(y = falla|\mathbf{X}) = f(\mathbf{X}) \tag{1}$$

Donde $p(y = falla|\mathbf{X})$ corresponde a la probabilidad que la empresa caracterizada por $\mathbf{X}$ no pueda cumplir sus compromisos financieros y $f(\mathbf{X})$ corresponde a una función matemática que aproxima la probabilidad a partir de los datos disponibles.

2.3. Proceso KDD

Se describirá a continuación el proceso KDD, el cual representa el proceso completo de extracción del conocimiento en base de datos [7]. El cumplimento de los pasos del proceso KDD permite llegar a modelos con un mejor desempeño y evita incurrir en errores de modelación, por ende será utilizado como guía para el desarrollo de este proyecto. El proceso KDD se puede aplicar usando métodos estadísticos como la regresión logística. Los pasos del proceso KDD son la consolidación de datos, el pre-procesamiento de los datos, el minado de los datos y la interpretación de los patrones encontrados, como se observa en la figura 1.

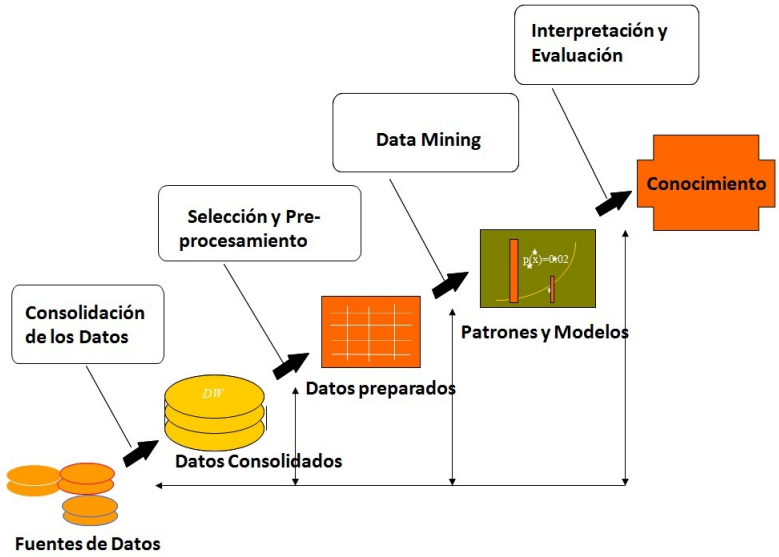

Figura 1: Proceso KDD

- *Consolidación de datos*: Para poder comenzar a analizar y extraer información útil de los datos es preciso, en primer lugar, disponer de ellos.

Esto en algunos casos puede parecer trivial, partiendo de un simple archivo de datos, sin embargo en otros, es una tarea muy compleja donde se debe resolver problemas de representación, de codificación e integración de diferentes fuentes para crear información homogénea.

- *Selección de atributos*: Para la construcción de modelos de clasificación se desea utilizar la menor cantidad de atributos posibles de manera de obtener un resultado considerado aceptable por el analista. Sin embargo, el problema radica en la elección y el número de atributos a seleccionar, debido a que esta elección determina la efectividad del modelo de discriminación construido. Este problema se conoce como *selección de atributos* y es combinatorial en el número de atributos originales [10].

- *Pre-procesamiento de datos*: El propósito fundamental de esta fase es el de manipular y transformar los datos en bruto, de manera que la información contenida en el conjunto de datos pueda ser descubierta. En esta etapa se consideran pasos como limpieza de datos ausentes o incorrectos, reducción de la información y transformación de los datos con el fin de adecuarlos al método de minería de datos.

- *Aplicación del método de minería de datos*: La aplicación de un algoritmo de aprendizaje tiene como objetivo extraer conocimiento de un conjunto de datos y modelar dicho conocimiento para su posterior aplicación en la toma de decisiones.

- *Interpretación y Evaluación*: En esta etapa se realizan distintas pruebas como análisis de sensibilidad y validación con distintas muestras para probar la robustez del modelo, así como la interpretación de los patrones minados.

En todas las etapas del proceso KDD es fundamental la cooperación con un experto del negocio como se mostrará más adelante.

2.4. Seguimiento de modelos

Una vez implementado el modelo de Credit Scoring desarrollado, la tarea siguiente, desde el punto de vista estadístico, es cuidar que el resultado obtenido mantenga su capacidad de discriminar entre los clientes que no pagan el crédito (*defaulters*) y los que sí lo hacen. Esta problemática no ha recibido mucha atención en la comunidad, aún cuando el no atenderla lleva a consecuencias graves en el uso. Se ha discutido [8, 9], por ejemplo, que la falta de actualización y mantenimiento de los modelos de riesgo de los bancos estadounidenses fueron una de las causas que precipitaron la crisis Sub-Prime de los años recientes. Debido a peticiones de varias instituciones financieras desarrollamos diferentes

enfoques de seguimiento de modelos de Credit Scoring basados en la regresión logística [4]. Para definir el problema, es posible identificar los cambios que pueden llegar a afectar de manera significativa la capacidad predictiva de un modelo:

- Capacidad discriminante de las variables: Para que una variable sea incluida en un modelo de regresión logística es necesario que esta discrimine entre las dos clases en estudio. Por discriminar se entiende el hecho que la distribución (media, desviación, etc.) de la variable sea distinta para cada una de las clases, de tal forma que a distintos valores de ella se obtengan distintas capacidades discriminantes. Este entonces corresponde a la primera condición que debe ser chequeada al momento de revisar cambios en el modelo.

- Distribución de las variables: Los supuestos básicos del modelo indican que cada una de las observaciones x_i es extraída de un conjunto X tal que se distribuye en base a una función $f(x_i)$ desconocida, pero idéntica para cada elemento. Este supuesto trae como consecuencia que los parámetros extraídos tengan aplicabilidad sólo mientras se tienen variables extraídas desde esta distribución, sin embargo, las distribuciones de las variables tienden a cambiar en el tiempo, pues la población modifica su comportamiento. Este fenómeno se observa por ejemplo en el riesgo crediticio, dónde empíricamente cada dos años se observan cambios en la población suficientes para impactar en el modelo [13].

- Capacidad discriminante del modelo en su conjunto: El cambio más drástico que puede tener una población puede volver el modelo en su conjunto no discriminante, si bien cada variable por separado puede mantener esta capacidad.

Se han desarrollado algunas aproximaciones teóricas por otros autores para identificar estos cambios en modelos de clasificación. Dentro de nuestro conocimiento, el enfoque más cercano al aquí detallado corresponde al trabajo realizado por Zeira *et al.* [15], el cual desarrolla test estadísticos para el caso general de modelos en el cual el error de validación se distribuye normal y las variables poseen un comportamiento tal que se puedan construir estadísticos a partir de sus distribuciones.

3. Metodología Propuesta y Experiencia

La construcción de los modelos se realizó siguiendo el proceso KDD. De acuerdo a esto, las experiencias más importantes se presentan siguiendo el orden señalado en este proceso (sub-sección 2.4), incluyendo el trabajo realizado para el seguimiento de los modelos.

3.1. Definición del Problema y Construcción de la base de datos

Todo proyecto parte con una definición clara de los objetivos del problema. En esta etapa es necesario definir la variable objetivo que se utilizará para clasificar, donde se consideran distintas condiciones de morosidad y se definen umbrales que separan los clientes etiquetados como "buenos" o "malos" en términos de su comportamiento crediticio. Es muy importante que este proceso se lleve a cabo en conjunto con la entidad financiera, debido a que los objetivos suelen variar. Por ejemplo, una entidad financiera estatal presenta una mayor preocupación por temas como la cobertura, a diferencia de instituciones privadas, donde la ganancia es de mayor preocupación.

Dentro de esta primera etapa se deben identificar además las fuentes de datos que son potencialmente útiles de acuerdo a los objetivos del problema y proceder a la adquisición de variables. Este proceso puede resultar complejo ya que la información relevante puede venir de diferentes fuentes. Para una institución financiera se contaban con más de 150.000 registros de créditos en un tramo de diez años, descritos por más de 100 variables. Para la segunda institución se disponían de aproximadamente 8.000 observaciones en un intervalo de tiempo de cuatro años. Sin embargo, el conjunto de atributos disponibles era de más de 650 variables. Esta entidad contaba con un sistema de riesgo traído del extranjero que no alcanzó los resultados esperados debido a que la realidad de los microempresarios difiere de manera drástica de país en país, surgiendo la necesidad de estudiar a fondo sus características en el caso particular.

Las fuentes de datos pueden ser de distinta naturaleza. A continuación se presenta una clasificación de las fuentes de datos más importantes:

- Bases de datos internas: Estas bases de datos incluyen, entre otros, la información personal del cliente, su historial crediticio con la entidad e indicadores preexistentes.
- Bases de datos externas: Muchas veces es posible obtener información de fuentes ajenas a la entidad, como la deuda en otras entidades financieras como bancos, casas comerciales o entidades privadas (DICOM).

- Variables e indicadores derivados: Más de 200 variables fueron construidas a partir de otras, tales como *ratios* de ingresos y deudas.

Considerar modelos con créditos a plazos muy diferentes puede introducir un sesgo, debido a que créditos a más largo plazo tienen asociado generalmente un monto mayor y por ende un riesgo implícito más alto, independiente de las características del cliente que lo recibe. Debido a esto, resulta importante diferenciar los clientes en distintos segmentos de riesgo y/o de acuerdo a condiciones similares. Esto último es más relevante aún cuando se cuenta con clientes antiguos para la compañía que presentan un historial de crédito, versus clientes nuevos sin información en muchas variables potencialmente útiles. A modo de ejemplo, una compañía contaba con créditos con plazos de hasta 10 años. Para esta entidad el universo se segmentó en 5 niveles distintos, de acuerdo a si los clientes eran nuevos o antiguos y en tres niveles de plazo (corto-mediano-largo), donde los dos segmentos de largo plazo se unificaron ya que presentaban características similares.

3.2. Pre-procesamiento de los Datos

Una vez con los datos provenientes de distintas fuentes consolidados en una matriz con los créditos las filas y sus atributos en las columnas, los siguientes pasos consisten en la limpieza de los datos y la selección de variables. Se desarrolla una metodología de cinco pasos con este propósito:

1. Concentración y análisis de valores perdidos: Con el fin de descartar rápidamente atributos irrelevantes, las variables muy concentradas en un único valor (en más de un 99 % de los casos) y atributos con más de un 30 % de valores perdidos fueron eliminados. La racionalidad de este segundo criterio es reducir el número de observaciones que deban ser eliminadas debido a valores perdidos.

2. Análisis univariado: Las variables fueron testeadas de manera individual si presentaban independencia con respecto a la variable objetivo. En particular, se utilizaron los tests de Kolmogorov-Smirnov para variables continuas y Chi-cuadrado para variables discretas. Si las variables estudiadas no presentaban diferencias al ser agrupadas en las dos categorías de la variable objetivo (por ejemplo, si la edad de los clientes buenos fuera estadísticamente similar en distribución a la edad de los clientes etiquetados como malos) se eliminaban del estudio.

3. Análisis Multivariado: Para poder estudiar la contribución de una variable en el método de clasificación, las variables restantes se utilizaron en un árbol de decisión sin poda, es decir, considerando todas las posibles relaciones entre variables que presentan algún tipo de comportamiento

discriminante. Las variables no incluidas en el árbol de decisión se excluyeron del estudio.

4. Limpieza y Transformación final: Las variables seleccionadas representaban un 20 % de las originales. La base de datos poseía un número pequeño de valores perdidos (menos de 1 %) que fueron eliminados. Las variables finales se transformaron para adecuarlas al modelo de clasificación. Las variables categóricas fueron agrupadas de acuerdo a criterios comunes (por ejemplo, los distintos giros de negocio se agruparon por giro primario) y finalmente fueron binarizadas. Para esta etapa de transformación de variables es esencial la comunicación con la contraparte, principalmente en la agrupación de categorías y en la construcción de indicadores que son potencialmente relevantes a priori en base a la información que maneja la entidad financiera.

3.3. Elección y Construcción del Modelo de Clasificación

El método de clasificación elegido para llevar a cabo la tarea de Credit Scoring corresponde a la regresión logística, el cual es uno de los más populares en la modelación del riesgo crediticio [13], habiendo sido utilizado con éxito en diferentes países. La regresión logística cuenta con varias ventajas en comparación con otros métodos de clasificación, tales como un buen desempeño predictivo (si bien algunos modelos avanzados de minería de datos, como las redes neuronales y Support Vector Machines, suelen presentar mejores resultados debido a la capacidad de modelar complejas funciones no lineales, esta diferencia no suele ser significativa [2]), simplicidad al momento de implementar e interpretar el modelo y robustez dado que no requiere de supuestos muy estrictos sobre los datos.

Formalmente, la regresión logística pronostica un evento dicotómico y_i en base a la información de N variables independientes $(x_1, ..., x_N)$. El método busca determinar la probabilidad de ocurrencia del evento dicotómico en función de la información contenida en las variables independientes, asumiendo una relación funcional como se muestra en la siguiente ecuación:

$$p(\mathbf{x}) = \frac{\mathbf{1}}{\mathbf{1} + \mathbf{e}^{-(\beta_\mathbf{0} + \sum_{\mathbf{i=1}}^{\mathbf{N}} \beta_\mathbf{i} \mathbf{x_i})}} \tag{2}$$

Lo anterior expresa que la probabilidad de ocurrencia del evento que se estudia (denotado por $p(\mathbf{x})$) es función de los valores de las variables independientes $\mathbf{x} = (x_1, ..., x_N)$. De esta manera, cuando se quiere ajustar un modelo de regresión logística a un conjunto de observaciones $(\mathbf{x}_i, y_i)$, $i = 1, ..., m$.

Lo más común es estimar el valor de los coeficientes $(\beta_0, ..., \beta_N)$ de acuerdo al método de máxima verosimilitud. En términos generales, el método de máxima verosimilitud encuentra los valores de los parámetros desconocidos que

maximizan la probabilidad de obtener el conjunto de datos observados. De esta manera se encuentran los estimadores de los parámetros y con ello se genera el modelo predictivo buscado.

Los estadísticos $\boldsymbol{\beta}$ se pueden interpretar como la influencia que tienen las variables en la probabilidad de que el cliente sea "malo" en términos de su comportamiento crediticio, asumiendo que $y_i = 1$ equivale a un cliente etiquetado como malo e $y_i = 0$ a uno etiquetado como bueno. Por ejemplo, si el signo de un estadístico β_i en particular es positivo quiere decir que esa variable explicativa es directamente proporcional a la probabilidad de no pagar íntegramente el crédito recibido. Si el signo es negativo, en cambio, la relación es inversa. En caso de que un estadístico β_i sea cercano a cero, la variable no afectará en la probabilidad de falla y por ende se considera irrelevante para el modelo, recomendándose su eliminación.

La pregunta que surge ahora es si, una vez encontrado los estimadores $\boldsymbol{\beta}$, ¿son éstos estadísticamente diferentes de cero? Para responder esta pregunta se construyeron test estadísticos para determinar, con un cierto nivel de significancia, si un estadístico β_i es estadísticamente diferente de cero, en base a su valor y su desviación estándar muestral asociada. Este estadígrafo se conoce como test de Wald y sigue una distribución χ^2. A partir de este test se utilizó una metodología *backward* de eliminación de atributos, que consiste en considerar todas las variables en el modelo y eliminar de manera secuencial la variable más irrelevante de acuerdo al test de Wald. Este procedimiento se repite hasta que se cuente con sólo variables relevantes en el modelo de acuerdo al test.

3.4. Metodología de Seguimiento

Una vez obtenidos los parámetros $\boldsymbol{\beta}$ a partir de los métodos anteriores, es posible definir el problema de seguimiento a ser resuelto. Obviamente, para realizar seguimiento a los modelos es necesario disponer de una nueva base de datos con créditos otorgados utilizando el modelo estadístico. Se dispone entonces de:

- Datos originales $\mathbf{X}$ y parámetros originales β_j para cada variable x_j presente en el modelo.
- Nuevo conjunto de datos $\mathbf{X'}$, asociado a nuevos casos $\mathbf{x}'_i \in \{x'_{i1}, \ldots, x'_{iJ}\}$.
- Probabilidades de *default* $p(\mathbf{x}'_i)$ calculadas con el modelo estadístico.

Uno de los puntos interesantes de este problema es que no se cuenta con las salidas reales y'_i, es decir, no se sabe si el cliente caracterizado por $\mathbf{x}'_i$ pagó el crédito o no lo realizó. Lo que se propone hacer, bajo estas condiciones, es

estimar una salida predicha $\hat{y}'_i$ para cada elemento en base a las políticas internas (punto de corte decidido) para cada compañía. Dado que lo que interesa en este caso es decidir si el modelo sigue discriminando o no, es razonable asumir que sus salidas siguen siendo relevantes y a partir de ellas realizar test acerca de la desviación obtenida.

Para realizar el seguimiento en regresión logística, una de las metodologías desarrolladas con anterioridad corresponde a la utilización de un test estadístico que permita detectar si han ocurrido cambios que sean de consideración.

Para cada parámetro β_j, la estimación de la regresión logística entrega un intervalo de confianza para el parámetro. Este intervalo corresponde al estimador de la regresión ajustado por la desviación estándar asociada al parámetro, así se tienen límites posibles para el parámetro dados por $\beta_j^{inf} = \beta_j - 2\sigma_{\beta_j}$ y $\beta_j^{sup} = \beta_j + 2\sigma_{\beta_j}$, correspondientes a los máximos y mínimos valores que puede tomar el estimador a un 95 % de confianza.

El procedimiento para probar si la nueva distribución X' se ajusta a lo que el modelo es capaz de manejar corresponde a generar un nuevo conjunto de parámetros β' a partir del conjunto de datos $\{X', y'\}$. El costo de re-entrenar una regresión logística utilizando programas computacionales actuales es muy bajo, por lo que generar este nuevo conjunto es poco costoso.

Con los nuevos parámetros β'_j y las distribuciones estándar encontradas para estos nuevos parámetros $(\sigma'_{\beta'_j})$ es posible construir un estadístico para los valores poblacionales de β'_j. Si la muestra posee un tamaño grande, se tiene que:

$$\frac{\beta'_j - \beta_{ref}}{\sigma'_{\beta'_j}} \rightsquigarrow t \tag{3}$$

Utilizando esto, es posible definir dos test estadísticos para medir si el nuevo parámetro se encuentra dentro de los intervalos de confianza anteriormente definidos.

$$\begin{array}{l} H_0 : \beta'_j = \beta_{inf} \\ H_a : \beta'_j < \beta_{inf} \end{array} \quad y \quad \begin{array}{l} H_0 : \beta'_j = \beta_{sup} \\ H_a : \beta'_j > \beta_{sup} \end{array} \tag{4}$$

Esta aplicación permite revisar si los nuevos parámetros se encuentran al interior del intervalo de confianza determinado por los parámetros antiguos, utilizando para ello la nueva estimación realizada. Se espera no rechazar las hipótesis nulas para ambos casos, dónde el valor crítico para el estadístico t con infinitos grados de libertad está dado por $1,645$ para el test unilateral para el límite superior y de $-1,645$ para el test de unilateral asociado al límite inferior. Esta aplicación debe cumplir con los siguientes requisitos:

- Se debe contar con suficientes casos en la muestra. Esto es importante por dos razones, en primer lugar, el número debe ser lo suficientemente grande para poder estimar parámetros, y en segundo lugar, la expresión (3) sólo se cumple si existe una cantidad alta de casos en muestra, es decir, el estimador t efectivamente presenta infinitos grados de libertad. En general, estos test de seguimiento se recomienda realizarlos cada tres o seis meses, de tal forma de acumular suficientes casos en muestra.
- Se deben almacenar los datos de cada caso de forma metódica. Esta es una recomendación obligatoria para cualquier aplicación real, los nuevos casos deben ser almacenados manteniendo sus variables, la probabilidad predicha y la clase seleccionada o el punto de corte utilizado para estimarla.

4. Resultados

Para mostrar los resultados del modelo se utilizarán los resultados reales de una de las instituciones en las que se aplicó la técnica para medir los índices de riesgo de las solicitudes. Esta institución entrega créditos a microempresarios dedicados a actividades agrícolas o ganaderas. Las variables utilizadas en la muestra de modelos aquí presentados se calcularon para clientes nuevos, sin historial crediticio, y para aquellos que si lo tenían. Estas corresponden a:

- Tenencia de propiedad: Quién es dueño del terreno. Se representa como una variable categórica, lo que implica que se modela con variables binarias, dejando una de las categorías como referencia. Existen cuatro clases: Propia (cat. base), Mediería (Tenen_Med), Arrendado (Tenen_Arr) y Otros(Tenen_Otro).
- Región: Región del país dónde habita el microempresario. Categorización depende de universo.
- Edad: Edad del cliente. Puede ser transformada en el logaritmo de la edad si éste aumenta la capacidad discriminante.
- Predios: Cantidad de predios que posee el microempresario. Tres categorías: Un predio (base), dos predios (Predios_Dos), más de dos predios (Predios_Mas).
- Rubro del microempresario: Categorizado según universo.
- Asociadas a créditos: Variables describiendo la situación crediticia del cliente. Se dividen en dos tipos, la cantidad de créditos en la entidad

(dos variables enteras distintas, créditos cerrados y créditos vigentes) y el plazo promedio de los créditos que ha tomado.

- Asociadas a la mora: Determinan la propensión a caer en mora de los clientes. Son tres variables, si cayó en mora en alguno de los créditos que ha tomado con la institución (Con_Mora_Ant), el porcentaje total de las cuotas que ha pagado que cayeron en mora (Porc_Mora) y el máximo de días que alguna cuota pasó en mora (Max_Mora).

- Ajustes: Si han ocurrido condonaciones, ajuste de intereses, o renegociaciones, los montos asociados a las pérdidas se almacenan en la variable Ajustes.

4.1. Aplicación del Modelo

El modelo fue aplicado a cinco universos distintos, obteniéndose parámetros y ajustes diversos para cada caso. En las tablas 1 y 2 se muestran los resultados[1] para los universos de clientes antiguos (es decir, que ya tuvieron algún otro crédito que fue pagado) con créditos de largo plazo (ALP) y de clientes nuevos con créditos de corto plazo (NCP).

Variable	β	σ_{β_j}	**P-Valor**
Tenen_Otro	0,3821	0,0966	0,0001
Tenen_Arr	0,7091	0,1805	0,0001
Tenen_Med	0,5312	0,1105	0,0000
Region_Z2	-1,0832	0,1328	0,0000
Region_Z3	-0,6011	0,1081	0,0000
Edad	-0,0053	0,003	0,0761
Predios_Dos	-1,4547	0,1029	0,0000
Predios_Mas	-2,4743	0,2232	0,0000
rubro_agric	0,0078	0,1069	0,9419
rubro_cer_prad	-0,3500	0,1169	0,0027
Constante	0,6477	0,2069	0,0017

Tabla 1: Coeficientes β, desviación estándar y significancia de los parámetros para clientes nuevos con créditos a corto plazo

En cada universo tanto los parámetros como las variables cambian, y ejemplifican la diferencia entre un score de comportamiento de pago (*behavioral scoring*) de uno que no lo es. El cambio en la cantidad de variables presentes

[1]Se han eliminado algunas variables por razones de protección de los resultados de nuestros clientes.

Variable	β	σ_{β_j}	P-Valor
Region_Z5	-0,6514	0,0612	0,0000
Region_Z6	-0,1875	0,0648	0,0040
Tenen_Med	0,2913	0,0700	0,0000
Tenen_Arr	0,8870	0,1666	0,0000
Tenen_Otro	0,4296	0,0595	0,0000
Dos_Predios	-0,5653	0,0531	0,0000
Mas_Predios	-0,9672	0,0648	0,0000
lnedad	-0,9950	0,0898	0,0000
Creditos_Cerrados	-0,0441	0,0074	0,0000
Creditos_Vigentes	-0,1509	0,0248	0,0000
Duracion_Creditos	0,0950	0,0237	0,0000
Con_Mora_Ant	17,3080	0,0593	0,0000
Porc_Mora	0,1636	0,0852	0,0550
Mora_Max	0,0019	0,0001	0,0000
Constant	16,1410	0,3695	0,0000

Tabla 2: Coeficientes β, desviación estándar y significancia de los parámetros para clientes antiguos con créditos a largo plazo

entre los distintos tipos de modelos es relevante, pues en los modelos de comportamiento tienen un peso mucho mayor los historiales de crédito del cliente y, sobre todo, las moras que haya manifestado al interior de la empresa. La conclusión que se desprende de estas tablas es que la variable principal para determinar el comportamiento de estos clientes es cuán ordenados son en sus cuentas y su propensión a desordenarse, aunque sea poco tiempo. Este efecto es aún más relevante que el ingreso del microempresario, pues la gran mayoría poseen ingresos concentrados en un pequeño intervalo de ganancias (sección 2.1)

lo que no permite diferenciar en gran manera. Esta es una diferencia importante con respecto al segmento de personas clásico, donde los indicadores de deuda y, sobre todo, las proporciones de deuda e ingreso, son variables fundamentales.

La diferenciación entre un score de comportamiento y uno que evalúa solicitudes también tiene un impacto en la capacidad de predicción del modelo, como se puede observar en la tabla 3.

El ajuste de los modelos, de todos modos, es razonable para ambos universos. El pago o no pago de un compromiso crediticio corresponde a un fenómeno social de alta complejidad, por lo que se esperan resultados con un rango entre 60-80 % de efectividad global.

Universo	No Defaulters	Defaulters
NCP	64,54 %	77,80 %
ALP	76,10 %	72,20 %

Tabla 3: Porcentaje de acierto para cada universo, por pagadores (No Defaulters) y no pagadores (Defaulters).

4.2. Seguimiento

Para el experimento de seguimiento, se dividió la muestra en créditos otorgados entre los años 2000 a 2004 y los otorgados con posterioridad de esta fecha. La institución que otorga los créditos conocía la ocurrencia de un cambio entre estos años, por lo que se esperaba que los test entregaran una diferencia significativa. Los resultados se observan en la tabla 4.

Variable	β'	$\sigma_{\beta'_j}$	β	**Lim. Inf.**	**Lim Sup.**	**t Inf.**	**t Sup**
Region_Z8	,696	,100	,794	,592	,995	1,03	-2,98
Region_Z9	,340	,089	,394	,208	,580	1,48	-2,71
Tenen_Med	-,785	,077	-,177	-,335	-,019	**-5,83**	-9,92
Tenen_Arr	-1,136	,138	-,364	-,682	-,047	**-3,30**	-7,92
Tenen_Otro	-,728	,080	-,264	-,432	-,096	**-3,71**	-7,93
Predios_Dos	1,150	,069	,151	,005	,298	16,56	**12,32**
Predios_Mas	1,845	,089	,495	,325	,665	17,01	**13,21**
Ajustes	,467	,085	,057	-,004	,117	5,54	**4,11**
Creditos_Cerrados	,132	,011	,085	,069	,100	5,75	**2,97**
Porc_Mora	-1,506	,103	-1,548	-1,795	-1,302	2,81	-1,98

Tabla 4: Resultados para el modelo de seguimiento. En negrillas aquellos cambios significativos.

Diversas variables presentan cambios significativos, destacando aquellas asociadas a los predios, pues son variables categóricas cuyo significado está unido al valor de las demás variables que forman las clases. El test detecta correctamente cambios en los intervalos asociados y, como era de esperarse, los cambios se ven reflejados para todas las categorías.

Para ejemplificar el hecho que las variables señaladas por el test sí detectan cambios relevantes, la figura 2 muestra la situación asociada a la variable Creditos_Cerrados, con una clara desviación entre ambos años.

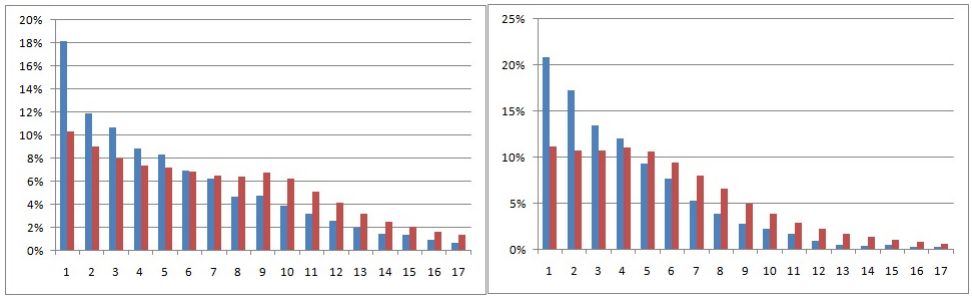

Figura 2: Cambios observados en la variable Creditos_Cerrados para los años 2000-2004 (izq.) y 2005 en adelante (der.).

5. Conclusiones

Los créditos a microempresarios corresponden a un mercado de gran importancia para Chile y Latinoamérica, pues las microempresas representan una parte significativa de las fuentes de trabajo, lo que hace que el otorgarle créditos no sea ya un negocio asociado a programas de apoyo, sino una oportunidad real y formal para mejorar las condiciones que enfrentan. A partir de esto, la necesidad por modelos de riesgo que estén adecuados a los fenómenos sociales que estas empresas enfrentas ha surgido en los últimos años.

Los microempresarios presentan características únicas que provocan que los estudios de riesgos sean asociados completamente a la realidad de los países dónde se desarrollan, impulsando la investigación científica y social como la que se presenta en este trabajo.

Desde el punto de vista de los modelos, las técnicas clásicas siguen entregando buenos resultados, pero es en la selección y construcción de variables donde se hace la diferencia con los modelos de riesgo clásicos. Situaciones como que el ingreso que posee el microempresario no sea relevante para la determinación del pago o no pago del crédito destaca como una de las razones por las que requieren estudios en profundidad. Es en las variables que muestran solidez financiera (tenencia y cantidad de bienes, por ejemplo) u orden en los pagos que realizan donde se encuentra la información que permite determinar la ocurrencia del fenómeno en estudio.

Otra necesidad importante que presentan estos modelos corresponde a realizar un seguimiento detallado del funcionamiento de éstos, pues los microempresarios están inmersos en un mercado muy volátil, siendo muy sensibles a vaivenes de la economía y presentando, por la naturaleza de su operación, un dinamismo mucho más grande que el que presentan las empresas de mayor tamaño. Así, el desarrollar herramientas que permitan determinar el momento

cuando ha ocurrido un cambio que daña la capacidad predictiva del modelo es una interrogante atractiva para los investigadores del área. El modelo aquí presentado cumple con este objetivo, siendo simple de implementar y entregando muy buenos resultados.

En cuanto a los resultados de la medición, se observan ajustes totalmente en línea con lo que se observa en las bancas de personas y de empresas, lo que avala el uso a nivel global tanto por parte de instituciones gubernamentales como privadas. A medida que aumente el interés por parte de privados para otorgar estos créditos, se hará más relevante su estudio y permitirá mejorar las condiciones que enfrentan estas empresas, sobre todo en países desarrollados.

Agradecimientos: El primer y segundo autor desean agradecer a CONICYT por las becas que permiten la realización de esta publicación. Este trabajo fue parcialmente financiado por el Instituto Sistemas Complejos de Ingeniería (ICM: P-05-004-F, CONICYT: FBO16).

Referencias

[1] Altman, E.I. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance* 23, 589-609. 1968.

[2] Baesens, B, Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. y Vanthienen, J. Benchmarking state of the art classification algorithms for credit scoring. *Journal of the Operational Research Society* 54(6):627-635, 2003.

[3] Beaver, W. H. Financial ratios as Predictors of Failure. *Journal of Accounting Research* 4, 71-111, 1966.

[4] Bravo, C., Maldonado, S. y Weber, R. Seguimiento en Modelos de Regresión Logística. *Revista de Ingeniería Industrial* Año 8, Nº 2: 31-44. 2009.

[5] Bravo, F. y Pinto, C. Modelos predictivos de la probabilidad de insolvencia en microempresas chilenas. *Contaduría Universidad de Antioquia* 53, 13-52. 2008.

[6] Coloma, P., Weber, R., Guajardo, J. y Miranda, J. Modelos analíticos para el manejo del riesgo de crédito. *Trend Management* 8: 44-51, 2006.

[7] Fayyad, U. Data mining and knowledge discovery- making sense out of data. *IEEE Expert-Intelligent Systems and Their Aplications* 11:20-25, 1996.

[8] Gerardi, K. S., Lehnert, A., Sherlund, S. M. y Willen P. S. Making Sense of the Subprime Crisis. *Public Policy Discussion Paper of the Federal Reserve* 09-1, Bank of Boston, 2009.

[9] Gerding, E. F. The Outsourcing of Financial Regulation to Risk Models and the Global Financial Crisis: Code, Crash, and Open Source *Washington Law Review*, Forthcomming, 2010.

[10] Maldonado, S. y Weber, R. A wrapper method for feature selection using Support Vector Machines. *Information Sciences* 179 (13), 2208-2217, 2009.

[11] D. Martens, B. Baesens, T. Van Gestel y J. Vanthienen. Comprehensible Credit Scoring Models using Rule Extraction from Support Vector Machines. *European Journal of Operational Research* 183(3): 1466-1476, 2006.

[12] Ohlson, J. A. Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research* 18, 109-131, 1980.

[13] Thomas, L. C. A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting* 16(2), 149-172. 2000.

[14] Yang, J. y Liu, G. The evaluation of classification models for credit scoring. *Arbeitsbericht Institut für Wirtschaftsinformatik, Georg-August-Universitat Göttingen.* 2, 2002.

[15] Zeira, G., Last, M. y Maimon, O. Segmentation on Continuous Data Streams Based on a Change Detection Methodology. En: *Advanced Techniques in Knowledge Discovery and Data Mining*, pp. 103-126, Springer. 2005.