NOTA TÉCNICA

Superintendencia de Bancos e Instituciones Financieras Chile

Comparación de algoritmos de clasificación para el incumplimiento crediticio. Aplicación al sistema bancario chileno

Miguel Biron L.
Víctor Medina O.

Nº 03/18 - Diciembre 2018

Superintendencia de Bancos e Instituciones Financieras

La serie de Notas Técnicas es una publicación de la Superintendencia de Bancos e Instituciones Financieras de Chile (SBIF), cuyo objetivo es aportar con artículos breves al debate de temas relevantes para estabilidad financiera y regulación bancaria. Si bien estas notas cuentan con la revisión editorial de la SBIF, los análisis y conclusiones en ellos contenidos son de exclusiva responsabilidad de sus autores.

The Technical Notes Series is a publication of the Chilean Superintendency of Banks and Financial Institutions (SBIF), whose purpose is to contribute with short articles to the discussion of issues relevant to financial stability and banking regulation. Although these notes have the editorial revision of the SBIF, the analysis and conclusions set forth are the responsibility of the authors and do not necessarily reflect the views of the SBIF.

Superintendencia de Bancos e Instituciones Financieras de Chile (SBIF)
Moneda 1123, Santiago, Chile

Editor: Dirección de Estudios SBIF

Nota Técnica N°03/18
Technical Note N°03/18

Comparación de algoritmos de clasificación para el incumplimiento crediticio: Aplicación al sistema bancario chileno*

Miguel Biron L.**
Víctor Medina O. ***

Departamento de Investigación y Riesgos, Dirección de Estudios
Superintendencia de Bancos e Instituciones Financieras

Diciembre 2018

RESUMEN

Este trabajo compara la capacidad predictiva de ocho algoritmos para la clasificación de deudores según su probabilidad de incumplimiento. La finalidad es evaluar los beneficios de modelos más complejos y los tradeoff asociados a este tipo de elección. Estos modelos son herramientas de apoyo importantes para una adecuada supervisión bancaria, coherente con el enfoque de riesgos de la SBIF. La población de estudio se concentró en los créditos comerciales otorgados a personas naturales, que representan aproximadamente el 16% del total de esta cartera. Los algoritmos estudiados fueron regresión logística, regresión logística penalizada, k-Nearest Neighbor, support vector machine, clasificador Bayesiano, redes neuronales artificiales, gradient boosting machines y ensambles heterogéneos. En términos de resultados, se encuentra que modelos más complejos aportan mayor discriminancia y estimaciones de probabilidades más precisas, en desmedro de la interpretabilidad del modelo. Además se evalúa la conveniencia de muestras balanceadas, encontrando mejores resultados en la mayoría de los casos.

ABSTRACT

This note compares the predictive capacity of eight algorithms for the classification of debtors according to their probability of default. The purpose is to evaluate the benefits of considering complex models and the tradeoffs associated with this choice. These models are important tools for an adequate banking supervision, consistent with the risk approach of SBIF. The study population are commercial credits granted to natural persons, which represent approximately 16% of this portfolio. The algorithms studied were logistic regression, penalized logistic regression, k-Nearest Neighbor, support vector machine, Bayesian classifier, artificial neural networks, gradient boosting machines and heterogeneous ensembles. In terms of results, it is found that more complex models provide better classification performance and more precise estimates of probabilities, to the detriment of the interpretability of the model. In addition, the convenience of balanced samples is evaluated, finding better results in most cases.

* Este trabajo fue elaborado mientras los autores eran funcionarios del Departamento de Investigación y Riesgos, Dirección de Estudios,SBIF. Las opiniones del estudio, errores y omisiones son de exclusiva responsabilidad de los autores y no reflejan necesariamente la visión de la institución. Se agradecen los aportes, comentarios y sugerencias de Gabriela Covarrubias, Luis Figueroa, Jaime Forteza, Alfredo Pistelli, Carlos Pulgar, Nancy Silva y de un árbitro anónimo.

**miguel.biron@stat.ubc.ca

***victor.medina@ed.ac.uk

1. Introducción

El riesgo de crédito, entendido como la posibilidad de que un prestatario no pueda cumplir con sus obligaciones contractuales de pago, es una de las principales fuentes de insolvencia en el sistema bancario (BCBS 1997). Las instituciones y el regulador, conscientes de la importancia de este riesgo, destinan grandes esfuerzos a mantener niveles de provisiones adecuados para hacerle frente y resguardar, en definitiva, la estabilidad del sistema.

Actualmente, la banca chilena se encuentra en un proceso de adopción de métodos estándares para el cómputo de provisiones. En el contexto de regulación de riesgos bancarios, un método estándar establece requerimientos mínimos de provisión.

Un método estándar para el cómputo de provisiones debe cumplir al menos cuatro características básicas: (i) debe ser de sencilla implementación, (ii) debe generar niveles conservadores de provisión, (iii) debe generar incentivos para que las instituciones desarrollen una adecuada gestión interna y (iv) debe considerar una medición del riesgo de crédito en el largo plazo o *through-the-cycle* (Forteza, Medina, y Pulgar 2017). Tomando en cuenta estas características, los métodos estándar no necesariamente tienen la mejor capacidad predictiva, pues no es su objetivo.

Los modelos internos de provisiones, por su parte, ponen mayor énfasis en la predicción. Un modelo predictivo de riesgo de crédito con un enfoque de pérdida esperada, requiere estimar tres variables aleatorias: (i) la exposición en el momento del incumplimiento (*exposure at default, EAD*), (ii) las pérdidas generadas por el incumplimiento (*loss given default, LGD*) y (iii) la probabilidad de incumplimiento (*probability of default, PD*). Este trabajo se concentra en el modelamiento de esta última variable con el objetivo de construir un modelo predictivo del incumplimiento de pago en créditos comerciales de personas naturales con giro en el sistema bancario chileno. Esta categoría de créditos representaba aproximadamente el 16 % del monto total de créditos comerciales a diciembre 2016 y fue escogida por tres razones. Desde un punto de vista estadístico, existe un número importante de observaciones, requisito necesario para la estimación de estos modelos. Por otro lado, representa un grupo de créditos poco estructurados en comparación a las carteras de consumo y vivienda, por lo que esfuerzos por indagar y entender el comportamiento de pago de este segmento resulta importante. Por último, constituye parte importante del segmento Pyme, cuya relevancia para la economía está bien documentada.

Si el evento de incumplimiento considera un horizonte de 12 meses, entonces para los créditos actuales no se tendrá información factual del desempeño que tengan en los próximos 12 meses[1], y por lo tanto, la construcción de un modelo predictivo toma relevancia.

Para la estimación de la PD, el enfoque predominante en la industria es el de clasificación. Este enfoque intenta predecir el evento de incumplimiento para un horizonte de tiempo fijo (variable y), a través de un conjunto de variables predictivas (variables $\mathbf{x}$). Otros enfoques, como el análisis de supervivencia, se concentran en predecir cuándo ocurrirá el evento de incumplimiento. El enfoque considerado aquí corresponde al de clasificación pero esta elección responde a los objetivos particulares de este trabajo y no a un análisis de superioridad de este enfoque sobre otros.

Desde la perspectiva del supervisor existen beneficios de contar con una predicción precisa del incumplimiento, pues fortalece la supervisión basada en riesgo a través de la identificación de deudores con diferentes niveles de riesgo, y otorga una visión interna y prospectiva sobre el comportamiento crediticio de los deudores, complemento relevante al enfoque de los modelos están-

[1] Lo que en la literatura se denomina *verification latency* (Hofer 2015).

dar. Ambos beneficios robustecen la tarea de resguardar la estabilidad financiera del sistema y proteger a los depositantes.

Desde un punto de vista académico, este trabajo brinda una comparación de algoritmos de clasificación avanzados, permitiendo responder la pregunta de si, en términos de predicción, es valioso usar algoritmos complejos.

El documento se estructura de la siguiente forma. En la Sección 2 se revisan los principales trabajos en relación a la comparación de algoritmos de clasificación en el ámbito de riesgo de crédito. En la Sección 3, se describen brevemente los algoritmos utilizados en la comparación. Luego, en la Sección 4 se presentan los resultados referidos al caso chileno y, finalmente, en la Sección 5 se concluye.

2. Literatura relacionada

Yeh y Lien (2009) realizan la comparación de 6 algoritmos de clasificación para deudores en Taiwan. Los algoritmos analizados son: *k-Nearest Neighbors*, regresión logística, análisis discriminante, *naive Bayesian classifier*, redes neuronales artificiales y árboles de clasificación. Los autores encuentran que las redes neuronales dan los mejores resultados. En tanto, Bellotti y Crook (2009) establece que *support vector machines* tienen mejor desempeño que métodos tradicionales como regresiones logísticas y análisis discriminante.

Lessmann et al. (2015) realiza una exhaustiva revisión de los nuevos algoritmos de *credit scoring* usados en la industria y academia, comparando 41 clasificadores aplicados en 8 bases de datos diferentes. Los autores encuentran, al igual que Yeh y Lien (2009), que las redes neuronales artificiales tienen el mejor desempeño predictivo dentro de los clasificadores individuales[2], mientras que, a nivel global, los modelos creados por un conjunto de clasificadores (*ensembles models*), obtienen las mejores predicciones. Se afirma que desde el año 2003, periodo correspondiente a la primera revisión realizada por Baesens et al. (2003), hasta el 2015 han habido importantes avances en el estudio de algoritmos de clasificación, incluyendo nuevos métodos de estimaciones, medidas de desempeño y técnicas más confiables para comparar clasificadores.

En línea con los resultados anteriores, el reciente trabajo de Fitzpatrick y Mues (2016) compara regresiones logísticas penalizadas, *gradient boosting machines* (GBM) y *random forest*. Los algoritmos se aplicaron a portafolios hipotecarios del sistema financiero irlandés, donde el mejor desempeño lo obtuvo GBM.

3. Algoritmos considerados

Este trabajo compara ocho algoritmos de clasificación: regresión logística (GLM), regresión logística penalizada (GLMNET), *k-Nearest Neighbors* (kNN), *support vector machine* (SVM), clasificador Bayesiano (NB), redes neuronales artificiales (NNET), *gradient boosting machines* (GBM) y un ensamble heterogéneo creado entre los clasificadores anteriores.

[2]El término individual se utiliza para diferenciarlos de los ensambles de modelos o *ensembles models*. Estos últimos pueden ser del tipo homogéneo, que consideran las mismas funciones bases, o heterogéneos, que mezclan funciones de distinto tipo, por ejemplo, árboles de decisión junto con regresiones logísticas.

El objetivo de cada modelo es generar una predicción para una variable dicotómica y que representa si un individuo incumple, en cuyo caso toma el valor 1, y que en caso contrario toma el valor 0 (o -1, siendo explícitamente especificado cuando corresponda). Esta predicción se realiza considerando n observaciones en la base de construcción y p variables explicativas representadas por el vector $\mathbf{x_i}$ para la observación i. Las variables consideradas pueden ser categóricas o continuas.

A continuación se describe cada algoritmo.

Regresión logística

Este clasificador es ampliamente utilizado en modelos de *scoring* crediticio y, por tanto, resulta un *benchmark* natural de comparación. La predicción para la observación i queda especificada como

$$p(y_i|\beta_0, \boldsymbol{\beta}, \boldsymbol{x_i}) = Bernoulli(y_i|\sigma(\beta_0 + \boldsymbol{\beta}^t \boldsymbol{x_i}))$$

donde $\sigma(x) = (1 + e^{-x})^{-1}$, $\boldsymbol{\beta}$ es el vector de coeficientes que acompaña a cada variable y β_0 corresponde al intercepto. La estimación se realiza a través de máxima verosimilitud (ver Friedman, Hastie, y Tibshirani (2001) para mayor detalles).

Regresión logística penalizada

Este enfoque sigue el procedimiento descrito por Zou y Hastie (2005), el cual incluye las penalizaciones l_1 (*lasso*) y l_2 (*ridge*) de manera conjunta (*elastic-net*).

La predicción tiene la misma estructura que el algoritmo anterior, sin embargo, la estimación de los coeficientes cambia y está determinada por

$$\min_{(\beta_0, \boldsymbol{\beta})} \left\{ - \left[\frac{1}{n} \sum_{i=1}^{n} y_i(\beta_0 + \boldsymbol{\beta}^t \boldsymbol{x_i}) - \ln(1 + e^{\beta_0 + \boldsymbol{\beta}^t \boldsymbol{x_i}}) \right] + \lambda \left[\sum_{j=1}^{p} (1 - \alpha)\beta_j^2 + \alpha|\beta_j| \right] \right\}$$

donde λ es el multiplicador de Lagrange y α el coeficiente que determina el nivel de importancia que se le otorga a cada penalización. Cada una de ellas tiene un propósito distinto. Por un lado, *lasso* realiza la selección de variables dejando sus coeficientes igual a cero en la medida que λ va creciendo. Por otro, *ridge* restringe los valores de los coeficientes correlacionados asintóticamente a cero cuando λ crece. De esta forma se seleccionan variables significativas, corrigiendo los coeficientes por el grado de correlación entre ellas. Tanto α como λ son determinados a través de validación cruzada (ver Friedman, Hastie, y Tibshirani (2001) para mayor detalles).

k-Nearest Neighbor

A diferencia de los clasificadores anteriores, este modelo es no paramétrico y la asignación de una nueva observación queda definida por los k vecinos "más cercanos" de la base de construcción. En consecuencia, alguna medida de distancia es necesaria. La más utilizada es la distancia de Minkowski que se reduce a la Euclidiana[3] cuando el orden es dos y es la utilizada en el presente trabajo.

[3]La línea recta que separa dos puntos.

La especificación utilizada corresponde a *kNN* ponderado, el cual considera que dentro de las k observaciones más cercanas, la ponderación se distribuye inversamente a la distancia. Para determinar los pesos existen variados esquemas, que toman una distancia (previamente estandarizada) y la transforman en un peso. El esquema utilizado acá es el sugerido por Samworth y others (2012).

Para una nueva observación $(y, \boldsymbol{x})$, se encuentran los $k+1$ vecinos más cercanos de acuerdo a la función distancia $d(\boldsymbol{x}, \boldsymbol{x_i})$ escogida. Posteriormente, el vecino $k+1$ es utilizado para estandarizar los k vecinos más cercanos a través

$$D_i = D(\boldsymbol{x}, \boldsymbol{x_i}) = \frac{d(\boldsymbol{x}, \boldsymbol{x_i})}{d(\boldsymbol{x}, \boldsymbol{x_{k+1}})}$$

donde $\boldsymbol{x_i}$ denota al i-ésimo vecino más cercano. Luego, con las distancias estandarizadas D_i se encuentran los pesos w_i a través del esquema comentado arriba. Finalmente, la probabilidad de incumplir se estima como

$$p(y = 1 | \boldsymbol{x}, \boldsymbol{\Omega}) = \frac{\sum_{i=1}^{k} w_i I(y_i = 1)}{\sum_{i=1}^{k} w_i}$$

donde $\boldsymbol{\Omega}$ corresponde al conjunto de observaciones $(y_j, \boldsymbol{x}_j)$ de la base de construcción. El valor de k se determina a través de validación cruzada (mayor detalles en Hechenbichler y Schliep (2004)).

Support vector machine (SVM)

Este algoritmo encuentra el hiperplano de separación óptimo entre dos clases a través de la maximización de la distancia entre los puntos más próximos de cada clase (estos puntos reciben el nombre de vectores de soporte). La virtud de esta metodología reside en que cuando no es posible encontrar una separación entre las clases, entonces las observaciones son proyectadas dentro de un espacio de mayor dimensión donde sí pueden separarse. Esto último se logra a través de técnicas de *Kernel*[4].

Para utilizar este algoritmo re-asignamos los valores de la variable y a 1 y -1 (en vez de 0), a modo de seguir la notación usual en la literatura. El problema de optimización que se resuelve y que corresponde al problema dual (mayor detalles en Schölkopf y Smola (2002)), es el siguiente

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \left\{ \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j K(\boldsymbol{x_i}, \boldsymbol{x_j}) \right\} \\ \text{sujeto a} \quad & \sum_{i=1}^{N} y_i \alpha_i = 0 \\ & 0 \leq \alpha_i \leq C, \; i = 1, \ldots, N. \end{aligned}$$

donde α_i corresponde al multiplicador de Lagrange para la observación i de la base de construcción y $K(\cdot)$ es el *Kernel* escogido[5].

[4] En la literatura se entiende como *Sparse Kernel Machines*.

[5] Los *Kernel* más utilizados en SVM son lineal, polinómico y *radial basis function* (RBF). Este último es el considerado en este trabajo.

Una vez estimado los parámetros a través del problema de optimización descrito, la predicción $\hat{y}$ para una nueva observación $\mathbf{x}$ se realiza a través de la siguiente expresión

$$\hat{y} = signo\left(b + \sum_{i=1}^{N} \alpha_i y_i K(\boldsymbol{x_i}, \boldsymbol{x})\right)$$

donde b corresponde al intercepto. Como se aprecia $\hat{y}$ está definido por la función signo, por lo tanto, el clasificador debe modificarse si se quieren obtener probabilidades. Platt y others (1999) propuso un mecanismo para obtenerlas y es el que aquí se utiliza para compararlas con los demás clasificadores.

Clasificador Bayesiano

Este clasificador estima la probabilidad de incumplir para valores de $\mathbf{x}$ dados, utilizando el teorema de Bayes. La particularidad de este modelo reside en el supuesto de independencia condicional entre los diferentes atributos. En términos matemáticos, la probabilidad de incumplir se especifica de la siguiente forma

$$p(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y)p(y)}{p(\boldsymbol{x})}$$

y haciendo uso de la independencia condicional de los atributos, se traduce a

$$p(y|\boldsymbol{x}) = \frac{p(y)\prod_{i=1}^{p} p(x_i|y)}{p(\boldsymbol{x})}$$

El término $p(y)$ se estima como la proporción de cada clase en la base de construcción. Por otra parte, cada distribución condicional $p(x_i|y)$ se estima utilizando *Kernel density estimation* con funciones normales y criterios de selección de ancho de bandas sugeridos por Silverman (1986).

Redes neuronales artificiales

Existe una gran variedad de especificaciones de redes neuronales y la que se utilizó en este trabajo corresponde a la llamada "Perceptrón de Multicapas". Con el objetivo de presentar una descripción breve, a continuación se describe el caso para una red neuronal de dos capas. La generalización del método considerando más capas[6], es análoga.

Definiendo M como el número de "neuronas" en la primera capa, el parámetro a_m asociado a la neurona $m \in M$ es una combinación lineal entre los p atributos de la forma

$$a_m = w_{m0}^{(1)} + \sum_{i=1}^{p} w_{mi}^{(1)} x_i$$

El superíndice (1) hace alusión a la primera capa de la red. Luego, cada parámetro a_m es transformado por una función de activación $h(\cdot)$ que usualmente corresponde a la función sigmoidal definida por $\sigma(x) = (1 + e^{-x})^{-1}$. Denominando $z_m = h(a_m)$, se realizan K nuevas combinaciones lineales correspondiente a la segunda capa, de la forma

$$\tilde{a_k} = w_{k0}^{(2)} + \sum_{m=1}^{M} w_{km}^{(2)} z_m$$

[6]La terminología para referirse al número de capas que aquí se utiliza, corresponde al número de capas de los pesos.

donde $k = 1, ..., K$. Como estamos considerando sólo dos capas para ilustrar el procedimiento, entonces K representa el número de *outputs* y, por otra parte, como estamos interesados en una clasificación binaria, entonces K se puede reducir a un sólo valor. Juntando los pasos anteriores, la red completa queda determinada por

$$\hat{y}(\boldsymbol{x}, \boldsymbol{w}) = \sigma\left(w_{k0}^{(2)} + \sum_{m=1}^{M} w_{km}^{(2)} h\left(w_{m0}^{(1)} + \sum_{i=1}^{p} w_{mi}^{(1)} x_i\right)\right)$$

La estimación de los coeficientes **w** se realiza minimizando la función de pérdida definida de la siguiente manera[7]

$$J(\boldsymbol{w}) = -\frac{1}{n}\sum_{i=1}^{n}\{y_i \ln \hat{y}(\boldsymbol{x}_i, \boldsymbol{w}) + (1 - y_i)\ln(1 - \hat{y}(\boldsymbol{x}_i, \boldsymbol{w}))\}$$

El número de capas utilizado se estimó a través de validación cruzada. Para una descripción más detallada, ver Friedman, Hastie, y Tibshirani (2001).

Gradient boosting machines

El término se acuña en el trabajo de Friedman (2001) y corresponde, en su versión más sencilla, a un conjunto (o ensamble) de árboles de decisión del tipo CART[8]. Luego, para la observación i la especificación puede denotarse como

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\boldsymbol{x}_i)$$

con $f_k \in \mathcal{F}$, donde $\mathcal{F}$ representa el espacio de todos los árboles CART y K el número de árboles escogido.

La estimación de y_i queda determinada como la composición de K árboles que se obtienen a través de la minimización de la función objetivo

$$L = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} R(f_k)$$

El primer término del lado derecho corresponde a la función de pérdida que para el caso binario se transforma en la función de pérdida logística ilustrada anteriormente. El segundo término representa la penalización que controla, por una parte, la complejidad del modelo para evitar sobreajustes en el número de "hojas" de los árboles, y por otra parte, correlaciones en los coeficientes a través de penalizaciones del tipo *ridge*.

El mecanismo de ajuste sigue un procedimiento iterativo que en cada etapa agrega un árbol que explique el error residual del modelo anterior. Por lo tanto, los primero árboles concentran la mayor importancia en términos explicativos. Mayor descripción acerca de este procedimiento y su implementación puede encontrarse en Chen y Guestrin (2016).

[7] La formulación se puede generalizar incorporando penalizaciones como las mencionadas anteriormente con el fin de evitar sobreajuste.

[8] El nombre se deriva de *Classification and Regression Trees* introducidos por Breiman et al. (1984).

Ensamble heterogéneo

Cuando el objetivo es la predicción por sobre la interpretación, la información que entrega cada modelo puede ser valiosa y complementaria, ya que uno puede ser capaz de explicar satisfactoriamente el comportamiento de los deudores para un segmento específico de la población, donde otro no lo logra, y vice versa. Este principio es el que está detrás de los ensambles heterogéneos[9]. Por lo tanto, ¿cómo podemos extraer la información de cada modelo y crear uno con mejor desempeño?

Variados son los procedimientos para lograrlo. El enfoque utilizado en el presente trabajo corresponde a *stacking* que combina las predicciones de todos los modelos, estimando los coeficientes de cada una a través de validación cruzada. Específicamente, el ajuste sigue el procedimiento de Caruana et al. (2004) que usa una selección *forward stepwise* para ir eligiendo aquellos modelos que van maximizando el desempeño. Mayor descripción de ajustes para ensambles de modelos se puede encontrar en Murphy (2012).

4. Aplicación al caso chileno

4.1. Parámetros de los modelos

Los modelos utilizados, a excepción de la regresión logística, requieren híper-parámetros que se deben definir con anterioridad al ajuste[10]. El desempeño de cada modelo es sensible a esta elección, sin embargo, en la mayoría de los casos el espacio de posibles valores para evaluar es muy grande o no acotado. Las estrategias de búsqueda de híper-parámetros óptimos más utilizadas son a través de grillas que incluyan una combinación finita y factible de valores, junto con búsquedas manuales. La desventaja de estos dos enfoques reside, principalmente en ineficiencias computacionales. No obstante, Bergstra y Bengio (2012) muestran que una búsqueda aleatoria resulta en una estrategia teórica y empíricamente más eficiente y, por lo tanto, fue la estrategia utilizada.

Para cada modelo se encontró aleatoriamente un espacio de valores candidatos de híper-parámetros de tamaño igual a 15 y se escogió el elemento del espacio que optimizaba el ajuste a través de validación cruzada utilizando el método *k-fold* con 10 divisiones en la base de construcción. En breve, lo que realiza este método de validación cruzada es que para cada una de las combinaciones de híper-parámetros en el espacio de valores candidatos, se estima el modelo excluyendo uno de los 10 cortes y se calcula el error de predicción en el corte excluido. Se realiza este procedimiento para los otros 9 cortes. De los 10 errores obtenidos en los cortes excluidos se calcula el promedio. Luego, se escoge la combinación de híper-parámetros que logró obtener el menor error promedio.

[9] *Gradient boosting machine* cumple el mismo principio, sin embargo, las funciones bases consideradas son del mismo tipo (CART) y, por lo tanto, se le conoce como un ensamble homogéneo.

[10] La diferencia entre híper-parámetros y parámetros del modelo, es que estos últimos se estiman en el ajuste mismo del modelo. En cambio, los híper-parámetros se establecen con anterioridad al ajuste. Por ejemplo, un híper-parámetro en una red neuronal sería el número de capas que esta tiene y los parámetros, los coeficientes que acompañan cada capa.

4.2. Datos

La base de datos utilizada se construyó utilizando ocho fuentes de información, entre ellas archivos normativos pertenecientes al Manual de Sistemas de Información de la SBIF. La consolidación incluyó los créditos comerciales de personas naturales con giro en 18 bancos del sistema. Esta categoría de créditos representaba aproximadamente el 16 % del monto total de los créditos comerciales al cierre de diciembre de 2016.

La muestra utilizada incluye información al cierre de cada mes, desde julio 2009 hasta diciembre 2016. El evento de incumplimiento, se mide a nivel de banco, y responde a la pregunta de si el crédito del deudor que presenta morosidad menor o igual a 90 días, transitó a mora de más de 90 días en cualquier momento transcurrido en los próximos 12 meses. Por lo tanto, la predicción se realiza para aquellos deudores que presentan sus créditos con morosidad igual o menor a 90 días, ya que los que están con morosidad mayor a 90 días tienen, por construcción, probabilidad igual a 1.

Una vez pre-procesada la data de construcción (limpieza, estandarización o categorización de variables, cuando corresponda), cada observación presentaba 128 atributos, incluidos factores relativos al deudor (como ingreso o ventas sobre deudas), al crédito (como nivel de mora actual, de los últimos 3, 6 y 9 meses), y a los demás créditos (como montos en leasing, factoring, entre otros).

Como el objetivo de este trabajo es crear un modelo predictivo y no de inferencia estadística, no hay garantías que un alto grado explicativo implicará un alto grado predictivo. Por lo tanto, la base completa se separa en dos, un 70 % corresponde a lo que denominamos base de construcción y el 30 % restante a la base de evaluación. La primera es la que utilizamos para estimar todos los parámetros de los modelos (incluidas las validaciones cruzadas para los híper-parámetros) y la segunda para presentar el desempeño de cada uno.

A la base de construcción se aplicó un tratamiento intermedio para balancear las clases. Este tratamiento se hizo con dos objetivos: (i) reducir el impacto negativo en el ajuste por el gran desbalance presente[11], y (ii) obtener una muestra que resulte en menor costo computacional (Ganganwar 2012). Esta tarea se llevó a cabo seleccionando aleatoriamente un subconjunto de la clase mayoritaria, de tal forma que las clases finales tuvieran la misma proporción (*down-sampling*)[12]. Sin embargo, aún después del balanceo existían 248.662 observaciones y resultaban en un costo computacional muy alto. Por lo tanto, se utilizó una muestra aleatoria simple correspondiente al 15 % de esas observaciones, es decir, 18.650 observaciones por clase.

En la literatura no se encontró evidencia categórica acerca de la necesidad de balancear las clases, por lo que se experimentó tomando una muestra de 37.300 observaciones sin balancear (*random-sampling*) con el propósito de comparar los ajustes en esta muestra versus los de la muestra *down-sampling*.

[11] Las clases representan a los deudores que cumplen y los que no. Al ser estos últimos muy menor en proporción a los primeros, algunos algoritmos no logran extraer diferencias de comportamiento substanciales y por tanto se hace necesario el balance de clases.

[12] Los métodos de balanceo con mayor consenso en la literatura en torno a su desempeño son los catalogados como "híbridos" (Chawla et al. 2002). La elección del método en el presente trabajo se rigió por limitaciones computacionales, principalmente.

4.3. Resultados

Los resultados que se presentan a continuación corresponden a los obtenidos en la muestra de evaluación. Las dimensiones de comparación entre los modelos fue discriminancia y grado de calibración de las probabilidades estimadas.

Para la discriminancia se utilizó el área bajo la curva ROC (*AUCROC*) y *H-Measure* (Hand 2009). La primera, ampliamente utilizada, se construye a través del área que abarca la curva de sensibilidad versus 1-especificidad para diferentes cortes. Un valor de *AUCROC* cercano a 1 representa buen desempeño y uno cercano a 0.5 es equivalente a un modelo donde las asignaciones de clases son aleatorias. *H-Measure*, por otra parte, es una medida alternativa y recientemente propuesta por D. J. Hand (Hand 2009) que logra sobrellevar algunas desventajas del *AUCROC*, como su falta de coherencia en términos del costo de clasificación errónea entre un algoritmo y otro, y que obtiene resultados inconsistentes cuando las curvas ROC se cruzan. Al igual que el *AUCROC*, a mayor valor de *H-Measure*, mejor es el desempeño en la separación de clases.

La calibración de las probabilidades, por su parte, fue medida a través de *Brier score*[13] e ilustrada en gráficos de calibración. Los gráficos de calibración se utilizan para ilustrar el grado de semejanza que existe entre las probabilidades estimadas y los incumplimientos que realmente sucedieron.

En relación a la discriminancia de los modelos, el Cuadro 1 muestra las medidas obtenidas en la base de evaluación. Cabe mencionar que las columnas *Down-sampling* y *Random-sampling* hacen referencia a la bases utilizadas para estimar los modelos y no a cambios en la base de evaluación.

Cuadro 1: Comparación de métricas de discriminancia.

Modelo	*Down-sampling*		*Random-sampling*	
	AUCROC	H	AUCROC	H
Ensamble	0.873	0.463	0.857	0.440
GBM	0.871	0.461	0.853	0.433
GLM	0.839	0.401	0.829	0.392
GLMNET	0.837	0.397	0.833	0.394
kNN	0.830	0.374	0.793	0.329
NB	0.818	0.368	0.784	0.342
NNET	0.856	0.432	0.830	0.403
SVM	0.849	0.431	0.775	0.368

Fuente: elaboración propia con datos SBIF.

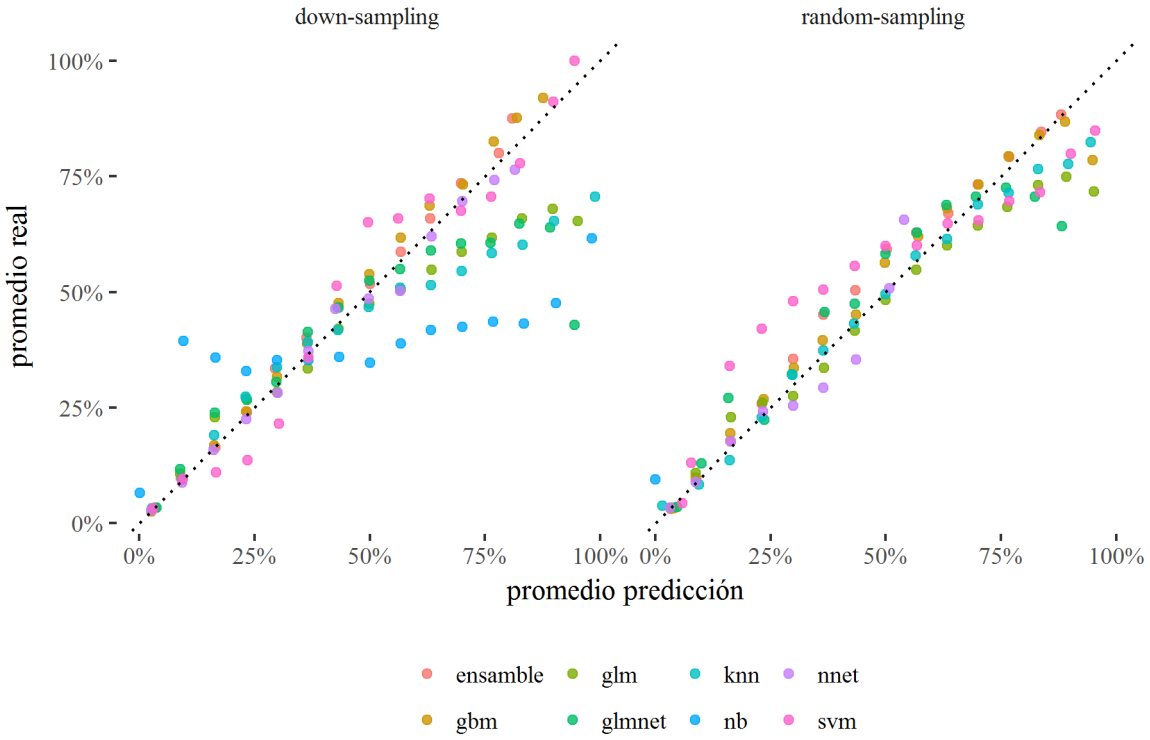
Comparativamente, el ensamble homogéneo (GBM) logra diferenciar ambas poblaciones de mejor manera en relación a los demás clasificadores. No obstante, el ensamble heterogéneo consigue mejorar ambas métricas aún más. Mejores desempeños se observan cuando los modelos fueron estimados utilizando la muestra balanceada (*Down-sampling*).

Con respecto a la comparación entre los modelos, se debe mencionar que se realizó una re-calibración de las probabilidades arrojadas por los modelos estimados con la muestra balanceada[14]. Cualquier tipo de muestreo de balance corresponde a un sesgo de selección inducido

[13] *Brier score* mide el error cuadrático medio entre la clase y la probabilidad estimada.

[14] En el caso del muestreo no balanceado, no es necesario hacer una re-calibración porque las proporciones empíricas de cada clase ya están presentes.

Figura 1: Gráficos de calibración para cada modelo y tipo de muestreo.



Fuente: elaboración propia con datos SBIF.

y, por lo tanto, conociendo las proporciones originales de desbalance se pueden "corregir" analíticamente las probabilidades estimadas a través de técnicas de re-calibración (Dal Pozzolo et al. 2015). Cabe aclarar que las proporciones de clases utilizadas son las que se presentaban en la base de construcción y no en la de evaluación.

La Figura 1 muestra el gráfico de calibración para cada modelo y muestreo. Como los incumplimientos son de naturaleza binaria, se divide el espacio de probabilidades en intervalos[15] y se calcula, por una parte y por cada intervalo, el promedio de las probabilidades estimadas de todos los créditos que pertenecen al intervalo, y por otra parte, la tasa de incumplimiento efectiva de los mismos créditos. Por lo tanto, lo que se esperaría es que las probabilidades estimadas de un grupo de créditos, en promedio, estén cercanas al porcentaje de créditos que incumplieron en ese mismo grupo. Es decir, que en el gráfico los puntos estén cercanos a la diagonal.

Se aprecia que en ambos muestreos los modelos que logran permanecer más cercanos a los valores reales son *GBM* y ensamble heterogéneo. Cabe mencionar que en el caso del modelo *NB* estimado con el muestreo aleatorio, se observa sólo un punto ya que el modelo le asigna a todos los créditos probabilidades cercanas a cero. Esto corresponde a un ejemplo donde el orden de probabilidades es relativamente bueno, como se vio con el AUCROC, sin embargo, el nivel de probabilidades no es correcto.

El Cuadro 2 muestra los resultados de *Brier score* que, como se mencionó anteriormente, mide el error cuadrático medio entre el evento de incumplimiento y la probabilidad asignada. Por lo tanto,

[15]Por efectos de visualización, se utilizaron 15 intervalos.

mientras menor es este indicador más cercano se está del valor real. Los resultados confirman lo ilustrado en la Figura 1, es decir, el mejor desempeño es obtenido por el ensamble heterogéneo ajustado con la muestra balanceada. La regresión logística penalizada (GLMNET) y no penalizada (GLM) son los únicos dos modelos que presentaron probabilidades más precisas cuando fueron ajustados con la muestra aleatoria.

Cuadro 2: *Brier score* para cada modelo y tipo de muestreo.

Modelo	*Down-sampling*	*Random-sampling*
Ensamble	0.0587	0.0602
GBM	0.0589	0.0609
GLM	0.0666	0.0656
GLMNET	0.0669	0.0664
kNN	0.0687	0.0691
NB	0.0836	0.0940
NNET	0.0621	0.0652
SVM	0.0648	0.0660

Fuente: elaboración propia con datos SBIF.

5. Conclusiones

El presente trabajo compara el desempeño de ocho algoritmos predictivos del incumplimiento crediticio, con información de personas naturales con créditos comerciales en Chile. Las métricas de comparación utilizadas fueron *AUCROC*, *H-Measure* y *Brier score*. Las dos primeras miden el poder discriminante de los modelos, es decir, qué tan buenos son en separar a los deudores que cumplen de los que no. La tercera métrica mide la cercanía entre la probabilidad estimada y el evento de incumplimiento.

Frente a la escasa literatura acerca de la elección de muestras balanceadas versus no balanceadas, los modelos se estimaron bajo estos dos escenarios. En términos generales, los resultados obtenidos en la muestra de evaluación apoyaron la elección de muestras con la misma proporción de clases. La excepción a esta tendencia se obtuvo para la regresión logística, tanto penalizada como no penalizada.

El mejor desempeño, tanto en discriminancia como en calibración de probabilidades estimadas, lo obtuvo el ensamble heterogéneo, que logró combinar óptimamente las predicciones de los otros modelos.

Las interrogantes que se desprenden en el diseño de esta herramienta son variadas y podrían constituir focos de futuras investigaciones. El procedimiento de búsqueda de híper-parámetros es una de ellas. Algunos enfoques recientes son la optimización Bayesiana o el propuesto por Yogatama y Mann (2014), que presenta un método rápido, efectivo y automático de búsqueda de híper-parámetros. Otra interrogante es cómo incorporar una visión integral del riesgo del deudor con todos sus créditos, desprendiéndose de la lógica de análisis por carteras, de tal forma de rescatar eventuales efectos cruzados que servirían como señales prematuras de contagio entre los productos de un mismo deudor.

6. Referencias

Baesens, B., T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, y J. Vanthienen (2003), «Benchmarking state-of-the-art classification algorithms for credit scoring», *Journal of the operational research society* 54 (6): 627-635.

BCBS (1997), *Report of the Central European Working Group: A Response to the Core Principles for Effective Banking Supervision*. BIS.

Bellotti, T., y J Crook (2009), «Support vector machines for credit scoring and discovery of significant features», *Expert Systems with Applications* 36 (2): 3302-3308.

Bergstra, J., y Y. Bengio (2012), «Random search for hyper-parameter optimization», *Journal of Machine Learning Research* 13 (Feb): 281-305.

Breiman, L., J. Friedman, C. Stone, y R. Olshen (1984), *Classification and regression trees*. CRC press.

Caruana, R., A. Niculescu-Mizil, G. Crew, y A. Ksikes (2004), «Ensemble selection from libraries of models», En *Proceedings of the twenty-first international conference on Machine learning*, 18. ACM.

Chawla, N., K. Bowyer, L. Hall, y W. Kegelmeyer (2002), «SMOTE: synthetic minority oversampling technique», *Journal of artificial intelligence research* 16: 321-357.

Chen, T., y C. Guestrin (2016), «Xgboost: A scalable tree boosting system», En *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785-794. ACM.

Dal Pozzolo, A., O. Caelen, R. Johnson, y G. Bontempi (2015), «Calibrating probability with undersampling for unbalanced classification», En *Computational Intelligence, 2015 IEEE Symposium Series on*, 159-166. IEEE.

Fitzpatrick, T., y C. Mues (2016), «An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market», *European Journal of Operational Research* 249 (2): 427-439.

Forteza, J., V. Medina, y C. Pulgar (2017), *Marco General Para El Diseño De Métodos Estándares De Provisiones Por Riesgo De Crédito*. Santiago. Superintendencia de Bancos e Instituciones Financieras.

Friedman, J. (2001), «Greedy function approximation: a gradient boosting machine», *Annals of statistics*: 1189-1232.

Friedman, J., T. Hastie, y R. Tibshirani (2001), *The elements of statistical learning*. Vol. 1. Springer series in statistics New York.

Ganganwar, V. (2012), «An overview of classification algorithms for imbalanced datasets», *International Journal of Emerging Technology and Advanced Engineering* 2 (4): 42-47.

Hand, D. J. (2009), «Measuring classifier performance: a coherent alternative to the area under the ROC curve», *Machine learning* 77 (1): 103-123.

Hechenbichler, K., y K. Schliep (2004), «Weighted k-nearest-neighbor techniques and ordinal classification» 399. http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-1769-9.

Hofer, V. (2015), «Adapting a classification rule to local and global shift when only unlabelled data are available», *European Journal of Operational Research* 243 (1): 177-189.

Lessmann, S., B. Baesens, H. Seow, y L. Thomas (2015), «Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research», *European Journal of Operational Research*

247 (1): 124-136.

Murphy, K. P. (2012), *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Platt, J., y others (1999), «Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods», *Advances in large margin classifiers* 10 (3): 61-74.

Samworth, R., y others (2012), «Optimal weighted nearest neighbour classifiers», *The Annals of Statistics* 40 (5): 2733-2763.

Schölkopf, B., y A. Smola (2002), *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Silverman, B. (1986), *Density estimation for statistics and data analysis*. Vol. 26. CRC press.

Yeh, I., y C. Lien (2009), «The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients», *Expert Systems with Applications* 36 (2): 2473-2480.

Yogatama, D., y G. Mann (2014), «Efficient transfer learning method for automatic hyperparameter tuning», En *Artificial Intelligence and Statistics*, 1077-1085.

Zou, H., y T. Hastie (2005), «Regularization and variable selection via the elastic net», *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2): 301-320.

Superintendencia
de Bancos
e Instituciones
Financieras
Chile