

7 Uso de *machine learning* para cuantificar el riesgo de crédito

Sergio Edwin Torrico Salamanca

INTRODUCCIÓN

De manera general, la ciencia y la gestión de riesgos junto con todos los campos de gestión y administración están atravesando un período de transición por cambios de tecnología vistos desde tres pilares: i. crecimiento exponencial de la disponibilidad de la información, ii. avances y acceso a tecnologías de modelación de datos (métodos e infraestructura), y iii. disrupción de la inteligencia artificial que multiplica la capacidad de los analistas para explotar los puntos i y ii.

El presente documento tiene como objetivo presentar el uso de herramientas de *machine learning* (pilar ii) para explotar la disponibilidad de información en el formato de paneles que se tiene en Bolivia para el sistema bancario (pilar i). Finalmente, la construcción del presente análisis ha sido realizada con el apoyo de un motor de inteligencia artificial (pilar iii) que ha incrementado significativamente la productividad en tiempo y esfuerzo para la consulta de referencias metodológicas.

El objetivo general del presente estudio es exponer una aplicación de la cuantificación el riesgo de crédito agregado del sistema bancario boliviano mediante la adaptación de técnicas de *machine learning*. Los objetivos específicos son:

- i. Diseñar una herramienta de medición del riesgo de crédito para pérdidas esperadas en un sistema bancario mediante el uso de *machine learning*.
- ii. Cuantificar el riesgo de crédito mediante la medida de pérdida esperada en el sistema bancario boliviano mediante el uso de *machine learning*.

MARCO TEÓRICO Y CONCEPTUAL

Pérdida esperada

Los principales enfoques de cálculo de la pérdida esperada (PE) están basados en autores como Merton (1974), quien introdujo un modelo en el que la valoración de la deuda se entrelaza intrínsecamente con la estructura de capital de una empresa y con la volatilidad de sus activos, que es la introducción a la gestión de riesgos y su impacto en la valoración.

Por otro lado, Jorion (2006), en su análisis de la pérdida esperada, introdujo la perspectiva de no considerar solamente las pérdidas individuales, sino también entender cómo se interrelacionan estas pérdidas potenciales a nivel de cartera. Esto a través de la interrelación en el comportamiento de las pérdidas individuales o correlación, lo que posteriormente lleva al análisis de estrategias de diversificación y la gestión del riesgo crediticio en una cartera.

Ya en el marco de la normativa prudencial internacional, las directrices del Comité de Basilea plantean distintos métodos que han evolucionado en su concepción y sofisticación a lo largo de las últimas décadas:

- Método estándar: define ponderaciones de riesgo fijas, permite a aquellos bancos con perfiles de riesgo simples tener un cálculo de la pérdida esperada mediante parámetros previamente analizados y establecidos (Comité de Basilea 2004).

- Método de calificación interna (IRB): para bancos con perfiles de riesgo más sofisticados y con capacidad de calcular internamente los parámetros que determinan la probabilidad de incumplimiento y la pérdida esperada (Comité de Basilea 2006).

Es necesario establecer una de las debilidades de los métodos internos, que es explorado en estudios como el de Gordy (2003) donde se identifica que si bien este método permite flexibilidad a las instituciones financieras, genera poca homogeneidad en la forma de cálculo en los sistemas financieros. Para este punto, en el presente estudio se considera que si bien las instituciones financieras cuentan con información suficiente para generar modelos eficientes al tener el detalle del comportamiento de cada operación crediticia y sus características (aspecto que no está disponible en el mercado de forma consolidada), también es necesario contar con modelos alternativos sistémicos que permitan medir de manera similar a las entidades de tal manera que se tengan puntos de referencia que permitan mejorar la gestión, este es el caso de las metodologías como el *macro credit scoring*.

Estudios contemporáneos

Las investigaciones recientes han puesto de manifiesto la necesidad de adaptar y mejorar constantemente los modelos de PE para reflejar las realidades cambiantes del mercado:

Breeden, Ingram y Thomas (2008) destacan la importancia de un mantenimiento riguroso y actualizaciones periódicas de los modelos utilizados para calcular la pérdida esperada ya que la dinámica del entorno lleva a la obsolescencia a modelos estáticos. Bluhm, Overbeck y Wagner (2016) plantean que la elección de un modelo apropiado debe basarse en la identificación de características específicas de la cartera y también del entorno.

Con relación a la debilidad de trabajar con modelos no homogéneos, las calificadoras son objeto de análisis por parte de Altman y Rijken (2004) que evalúan las diferencias entre calificadoras cuyos resultados pueden mostrar estimaciones divergentes.

En cuanto a la necesidad de evolución de estas metodologías, autores como Van Deventer, Imai y Mesler (2004) plantean la necesidad de integrar modelos avanzados a los cálculos para mejorar la precisión mediante el uso de cópulas que permiten capturar de mejor manera la correlación entre créditos de las instituciones financieras.

Desarrollo actual

El crecimiento de la digitalización en los sistemas financieros ha generado una necesidad sobre los sistemas tecnológicos y su capacidad de procesamiento (*machine learning*); por otro lado, se han generado desafíos de recopilación de datos (*big data*).

De manera cada vez más marcada y recurrente se plantea la necesidad de incorporar como parte del conjunto de herramientas cuantitativas dos pilares:

- Modelos de aprendizaje automático-*machine learning*: por su capacidad de procesar volúmenes significativos de información al alcance de usuarios que no pertenecen al campo de TI y que tienen capacidad de ejecutarse en máquinas semi-estándar.
- Consumo de datos e integración de datos no tradicionales: además de la necesidad de dar un salto a información no tradicional, se tiene aún en el mundo la tarea en avance de recopilar toda la información que se genera al interior de las instituciones financieras.

Random forest

El algoritmo *random forest*, propuesto por Breiman en 2001, se constituye en uno de los métodos cuantitativos de aprendizaje

automático con mayor uso, está clasificado como un método de ensamble.

En enfoque de este algoritmo se basa en la combinación de varios árboles de decisión para crear un “bosque” (de ahí el nombre en inglés); luego, este “bosque”, mediante técnicas de votación, puede producir resultados robustos frente a las metodologías cuantitativas actuales (Breiman 2001).

Componentes

Se describen los componentes de la técnica cuantitativa de random forest:

Bagging: también denominado *bootstrap aggregating*, es el proceso mediante el cual se crean varias muestras de un conjunto de datos, utilizando la técnica de muestreo con reemplazo para luego poder construir árboles a partir de cada muestra generada. Si se combinan (ensamblan) las estimaciones de los árboles creados en problemas como el que trata el presente artículo en los cuales se necesita clasificar variables, mediante un algoritmo pueden “votar” para definir la estimación más eficiente y con menor sensibilidad a la varianza de los datos de entrenamiento (Breiman 1996).

Feature randomness: otra característica de este método, además del *bagging*, es la introducción de una capa adicional de aleatoriedad en el proceso ya que se muestrean características. Para cada división de un nodo (al analizar una variable) durante la creación de los árboles, este algoritmo puede no considerar todas las características disponibles, solo un subconjunto aleatorio de ellas. Esta característica permite que los árboles construidos no estén correlacionados entre sí y, por tanto, se puede asumir que el bosque combinado tiene una varianza reducida (Ho 1998).

Out-of-bag (OOB) error estimation: este método tiene la capacidad de estimar el error de predicción o, en este caso, de clasificación a partir del uso de observaciones del conjunto de datos que no

han sido utilizados en el entrenamiento del modelo. Estas son las observaciones que se denominan *out-of-bag* (Breiman 2001).

Importancia de las variables: si bien no se cuenta con una ecuación como en los métodos econométricos tradicionales, se puede determinar una medida de la importancia de las variables explicativas. Esto se logra con medidas como el *mean decrease Gini* (Strobl et al. 2007).

METODOLOGÍA

Se presenta la metodología del presente artículo con base en los distintos tipos de análisis: nivel de profundización, objeto de la investigación, grado de manipulación, tipo de información empleada, tipo de inferencia y período.

Este es un análisis aplicado ya que busca de forma empírica cuantificar un fenómeno abstracto y no se limita a un análisis teórico y del comportamiento. Se cataloga como una investigación no experimental, porque observa el fenómeno de la materialización del riesgo de crédito en su contexto habitual y sin manipular variables. Se clasifica como una investigación cuantitativa porque se utilizan procedimientos fundamentalmente basados en la medición con el uso de técnicas de *machine learning*. Es una investigación tanto longitudinal como transversal ya que se trabaja con una base de panel de datos de bancos a través del tiempo, el período es enero de 2005 a abril de 2023.

Metodología cuantitativa: se trabaja con un algoritmo comúnmente utilizado en el campo de *machine learning* denominado *random forest*, cuyos autores seminales son Breiman y Cutler (2007).

Siguiendo la metodología planteada por Torrico (2021) para el cálculo del *macro credit scoring* y para determinar la pérdida esperada del sistema bancario analizado a nivel agregado, se trabajó con una función en la que la pérdida esperada se calcula a través del perfil financiero de las entidades financieras de la siguiente forma:

Pérdida esperada ($E[L]$) = f (perfil financiero de las entidades financieras) (1)

En esta metodología, en lugar de construir medidas de riesgo de crédito utilizando el comportamiento de los clientes de las entidades financieras, se asume que el nivel agregado de riesgo de crédito puede ser modelado por sus propios indicadores financieros que reflejan el perfil financiero y de riesgo.

La notación tradicional de la pérdida esperada es la siguiente:

$$E[L] = PD \times LGD \times EAD \quad (2)$$

Donde:

EL = pérdida esperada

PD = probabilidad de incumplimiento

EAD = exposición al incumplimiento

LGD = pérdida dado el incumplimiento

Aplicación de la metodología propuesta

Datos analizados

La base de datos utilizada se compone de 3712 observaciones en una estructura de datos de panel de 16 bancos, en el rango de tiempo desde enero de 2005 a abril de 2023, con una periodicidad mensual. Esta es la información de 18 años basada en series históricas construidas a partir de las publicaciones del ente regulador que es la Autoridad de Supervisión del Sistema Financiero (ASFI).

Para la construcción del modelo se separa de forma aleatoria la base de la siguiente forma:

- Base de entrenamiento: 2784 observaciones (75 %).
- Base de prueba: 928 observaciones (25 %).