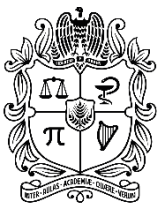


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Pronóstico de la pérdida crediticia esperada de los clientes con mayor nivel de riesgo de un banco por medio de modelos paramétricos y no paramétricos.

Brandon López Avendaño
Estudiante de Maestría en Ciencias-Estadística

Nelfi Gertrudis González Álvarez
Profesora asociada
Escuela de estadística

Universidad Nacional de Colombia

Escuela de estadística

Medellín, Colombia

2023

Pronóstico de la pérdida crediticia esperada de los clientes con mayor nivel de riesgo de un banco por medio de modelos paramétricos y no paramétricos.

Brandon López Avendaño

Trabajo de grado presentado como requisito parcial para optar al título de:

Magister en Ciencias-Estadística

Directora:

Doctora en Ciencias Estadísticas Nelfi Gertrudis Gonzáles Álvarez

Universidad Nacional de Colombia

Escuela de estadística

Medellín, Colombia

2023

Resumen

La pérdida crediticia esperada (ECL) permite establecer bajo la normatividad IFRS 9 el nivel de provisión y el cálculo de reservas esperadas de una entidad financiera, donde a mayor riesgo percibido, existirá un mayor nivel de provisión en los balances del banco. Se ha encontrado en la literatura que, por medio de indicadores macroeconómicos, información transaccional y sectorial, índices financieros y medidas de riesgo, es posible prever la pérdida crediticia esperada en diferentes periodos de tiempo, por lo tanto, en el presente trabajo se proponen 437 variables que han resultado ser significativas en diferentes estudios, a las cuales, se les realizó una reducción de dimensionalidad y selección de variables, resultando 10 de éstas las que mejor explican la ECL. Adicionalmente, se propusieron modelos paramétricos y no paramétricos como: Regresión Lineal Múltiple, Lasso, Ridge, Bosques Aleatorios, entre otros para pronosticar la pérdida crediticia esperada, siendo el modelo Extremely Randomized Trees (Extra Trees) el que mejor desempeño tuvo en las medidas MSE, MAE y coeficiente de determinación, con valores de 0.0078, 0.0564 y 0.9199, respectivamente. Se encontró que gran parte de los predictores presentaban relaciones no lineales con la variable respuesta que el modelo era capaz de capturar, y por medio de los valores de SHAP (Shapley Additive Explanation) se pudo evidenciar que las relaciones de las variables independientes con la ECL guardaban sentido con la teoría económica.

Palabras claves: pérdida crediticia esperada, ECL, provisión, entidades financieras, riesgo de default, Extremely Randomized Trees, Extra Trees.

Abstract

Forecasting expected credit loss of high-risk bank clients using parametric and non-parametric models

Expected credit loss (ECL) enables financial institutions to determine the provision level and calculate expected reserves in accordance with IFRS 9 regulations. Higher perceived risk corresponds to higher provision levels recorded in the bank's balance sheets. Extensive research has shown that by utilizing macroeconomic indicators, transactional and sectorial information, financial ratios, and risk measures, it is possible to forecast the expected credit loss across different time periods. In this study, a set of 437 variables, identified as significant in previous research, underwent dimensionality reduction and variable selection procedures, resulting in the identification of 10 key predictors that best explain the ECL.

Moreover, a range of parametric and non-parametric models, including Multiple Linear Regression, Lasso, Ridge, Random Forests, among others, were evaluated for their ability to forecast the expected credit loss. Among these models, the Extremely Randomized Trees (Extra Trees) model demonstrated superior performance in terms of MSE, MAE, and coefficient of determination, with values of 0.0078, 0.0564 and 0.9199, respectively. Notably, the analysis revealed that a significant number of predictors exhibited non-linear relationships with the response variable, which the Extra Trees model effectively captured. By employing SHAP values (Shapley Additive Explanation), the relationships between the independent variables and ECL were found to align with the economic theory.

Keywords: Expected Credit Loss, ECL, provision, financial institutions, default risk, Extremely Randomized Trees, Extra Trees.

Contenido

Lista de tablas

Lista de figuras

Lista de ecuaciones

Glosario

- **Agregados monetarios M1 y M2:** Según el Banco de la República de Colombia (2022) corresponde al conjunto de pasivos, emitidos por un agente del sistema financiero, y demandado por alguno de los sectores de la economía que cumple alguna de las funciones de la demanda de dinero. Esto es, medio de cambio, unidad de cuenta y, depósito de valor. El **M1** se define como las cuentas corrientes más el efectivo en poder del público. Las cuentas corrientes son emitidas exclusivamente por los bancos comerciales, mientras que el efectivo es emitido por el Banco de la República. El **M2** incluye a M1 más los cuasidineros. Los cuasidineros están compuestos por los depósitos de ahorro (bajo cualquiera de sus modalidades: con certificado, ordinario o indexado), y los certificados de depósito a término en todos los plazos y condiciones de indexación.

- **Altman´s Z-score:** indicador propuesto por Altman (1968) que mide la probabilidad de que una empresa entre en bancarrota. Un valor de Z superior a 2.99 significa que la empresa no tiene señal alguna de quiebra, un valor entre 1.81 y 2.99 es que existe una probabilidad de bancarrota en un corto plazo, y menor a 1.81 un riesgo inminente.

- **Colateral:** En el mundo financiero, el colateral es un activo que un prestatario promete como garantía de un préstamo (Chen, 2022).

- **Default:** El riesgo de Default o de no-pago, es ampliamente usado en la economía para definir personas jurídicas o naturales que incumplieron con el pago oportuno hacia sus acreedores, Chen (2022) hace énfasis es que es incumplimiento en realizar pagos sea de intereses o de capital sea en deudas bancarias o activos financieros.

- **DLTV:** Es la relación entre el préstamo con el valor de la garantía al momento donde se incurrió en default, se puede definir como el LTV cuando se llega a mora.

Esta definición fue propuesta como variable significativa en el estudio realizado por Leow & Mues (2012).

- **EBIT:** La utilidad EBIT es un indicador financiero que permite medir el desempeño de la empresa sin descontar los pagos de intereses o de impuestos, es decir, la utilidad EBIT no se ve afectada por la estructura de capital ni por la carga impositiva. Corresponde al acrónimo de los términos en inglés *Earnings Before Interest Taxes*, el cual en español corresponde a la utilidad antes de intereses y de impuestos o UAII, una forma de hallar este indicador es partiendo de la utilidad EBITDA y restar las depreciaciones y amortizaciones (Ross et al., 2010).

- **EBITDA:** La utilidad EBITDA es un indicador financiero que mide la generación de flujos provenientes de la operación. Corresponde al acrónimo de los términos en inglés *Earnings Before Interest Taxes Depreciation and Amortization*, el cual en español corresponde a la utilidad antes de intereses, impuestos, depreciaciones y amortizaciones. Es una medida muy conocida en el sector financiero dado que mide la utilidad de una empresa sin considerar su estructura financiera y desconociendo las partidas contables como la depreciación que no es una salida real de efectivo (Hayes, 2022).

- **EEFF:** Estados Financieros de una empresa, entre ellos se encuentra el Estado de Resultados Integral, Estado de Situación Financiera, Estado de Flujo de Efectivo, entre otros.

- **Haircut:** Se refiere al descuento que se le debe hacer a la garantía para reflejar un valor más aproximado al de mercado, es una tasa que varía de acuerdo al colateral. Según el Banco Central Europeo, un Haircut se refiere a una reducción aplicada sobre el valor de un activo (ECB, 2023)

- **KTNO:** El Capital de Trabajo Neto Operativo**,** es un indicador financiero que relata la actividad de una empresa y está asociado directamente a la liquidez, Gitman & Zutter (2012) lo definen como “La diferencia entre el cambio en los activos corrientes y el cambio en los pasivos corrientes” donde **KT** hace referencia al capital

de trabajo de una empresa y está ligado con los rubros catalogados como activos corrientes que pueden volverse efectivo en un plazo menor a un año. **N** hace referencia a neto e indica aquellos pasivos que se requieren para generar la actividad económica y deben ser descontados, y por último la letra **O** indica que sólo se tomará en cuenta aquellos rubros asociados a la operación. Las fórmulas plateadas por el CFI (Corporate Finance Institute) son activos corrientes (sin considerar el efectivo) – pasivos corrientes (descontando la deuda) o, cuentas por cobrar + inventario – cuentas por pagar (Vipond Tim, 2022).

- **LTV:** Mide la relación entre el préstamo con el valor de la garantía. Corresponde al acrónimo de los términos en inglés *Loan-to-value.* planteado por Leow & Mues (2012).

- **Repos:** Según el Banco de la República de Colombia (2022) Es una operación donde se vende un activo (como títulos financieros) a cambio de una suma de dinero, con el pacto de recomprarlo en una fecha posterior. En ese sentido, es similar a un préstamo de dinero con una garantía (el activo). Las operaciones repo son el principal mecanismo mediante el cual el Banco de la República suministra liquidez a la economía, con el objetivo de ajustar la oferta de dinero para garantizar que las tasas, con las que se prestan dinero las entidades financieras entre sí, se acerquen a la tasa de intervención del Banco de la República.

- **ROA:** Mide la rentabilidad de los activos de la empresa sobre la generación de utilidades, calculando la eficiencia de los activos fijos para convertir flujos netos. Corresponde al acrónimo de los términos en inglés *Return over assets y* su fórmula es Utilidad neta/ activos totales (Ross et al., 2010).

- **ROE:** Mide la generación de utilidad neta que tiene la empresa sobre el capital pagado por los socios, corresponde al acrónimo de los términos en inglés *Return over Equity y* su fórmula es Utilidad neta/ Promedio del capital de los accionistas, según Fernando (2022). Por otro lado Ross et al. (2010) plantean el ROE como Utilidad neta/Capital contable total.

- **Rotación de inventarios:** Este indicador permite conocer el número de veces que rotan los inventarios de una empresa en un periodo especifico (Ross et al., 2010). Entre mayor sea este número mejor es el manejo de inventarios en la compañía, dado que una rotación muy baja querrá decir que la empresa no vende sus inventarios con tanta frecuencia y que requiere financiación de capital de trabajo, su fórmula es costo de la mercancía vendida /promedio de inventarios (Fernando, 2022).

- **Tasa de recuperación:** Es el porcentaje del valor que se recupera al momento del incumplimiento, siendo el valor restante de la pérdida al momento de default (Bastos, 2010).

- **WK o KT:** El capital de trabajo es un indicador que mide la liquidez de la empresa para solventar los compromisos de corto plazo y de la eficiencia operacional. Corresponde al acrónimo de los términos en inglés *Working Capital* que en español se traduce a capital de trabajo, se halla descontando los pasivos corrientes de los activos corrientes (Robles, 2012).

Introducción

De acuerdo con el Comité de Supervisión Bancaria de Basilea que es el organismo encargado a nivel mundial de la regulación prudencial de los bancos y, específicamente, de su solvencia, el riesgo de crédito se puede definir como la posibilidad de que un prestatario bancario no cumpla con sus obligaciones de acuerdo con los términos acordados. Este comité plantea que los bancos deben gestionar el riesgo de crédito inherente a toda la cartera, así como el riesgo en créditos a los diferentes instrumentos financieros que existen en su portafolio (BCBS, 2000).

En la gestión del riesgo de crédito se debe medir y mitigar la probabilidad del incumplimiento de sus obligaciones en el marco de las cinco “c”, a saber: el carácter, que mide la credibilidad y reputación del prestatario, la capacidad, que indica la solvencia para cumplir con sus obligaciones, el capital, que relata la estructura financiera de la empresa, el colateral, que mitiga el riesgo considerando el activo que se deja como garantía, y las condiciones, las cuales indican las circunstancias y propósitos del crédito (Peterdy, 2023).

ECL corresponde al acrónimo de las palabras en inglés “Expected Credit Loss”, las cuales en español hacen referencia a las pérdidas crediticias esperadas. Este término es definido por Antonsson (2018) como el determinante en el establecimiento de niveles de provisiones y el cálculo de reservas para pérdidas crediticias esperadas e inesperadas, como parte del cumplimiento de los requisitos regulatorios de capital. KPMG (2017) la define como las diferencias entre los flujos de efectivo adeudados según el contrato y los flujos de efectivo que una entidad espera recibir.

De acuerdo al marco de regulación IFRS 9, se requiere que una institución financiera reconozca inmediatamente una pérdida crediticia esperada (ECL) de 12 meses de un activo financiero en la primera fecha de presentación de sus EEFF después del desembolso del crédito, y como consecuencia del desembolso, genere una provisión para cubrir la ECL (Temim, 2016).

En el presente trabajo se analizará la ECL o pérdida crediticia esperada, para aquellos clientes de la entidad financiera que tienen un saldo significativo y un nivel de riesgo mayor. Este grupo de clientes son los que generan los mayores movimientos de constitución o liberación de provisión cada semestre, por lo que pronosticar los valores de la ECL es de gran importancia para la toma de decisiones. Este trabajo tiene el propósito de brindar a la entidad financiera un modelo que capture las relaciones entre la variable respuesta y los predictores buscando que guarden sentido con la teoría económica, además, generar un pronóstico de pérdida crediticia esperada para el grupo de clientes que permita identificar previamente la provisión estimada para los cierres contables. La metodología que se empleará para pronosticar la pérdida crediticia esperada, consistirá en proponer modelos paramétricos y no paramétricos. Para esto, primero se realizará una selección de variables independientes con el fin de identificar las relaciones existentes entre cada predictor y la ECL, y sólo conservar aquellas variables que generen un mayor impacto en los modelos, buscando tener un modelo parsimonioso. La información y alcance del modelo, se limita a este grupo específico de clientes, por lo que no se puede extrapolar a todo el portafolio de la entidad financiera.

Planteamiento del problema

Luego de la crisis financiera de 2008 el mundo entero presenció un suceso que años atrás parecía inconcebible: la quiebra de múltiples entidades financieras. Un sector que ha sido percibido a lo largo del tiempo como inelástico mostró en dicho año cómo la mala gestión del riesgo puede generar un efecto en cadena hacia toda la economía, poniendo en evidencia la gran importancia de evaluar adecuadamente los riesgos asociados a cada instrumento financiero, los problemas del enfoque tradicional de evaluar el riesgo por el tamaño y no por el desempeño, la falta de regulación que se tenía en la industria y su desajuste en la liquidez (Apergis et al., 2019)

Diez años más tarde, salió en vigor la nueva normatividad que entraría a regir al sector financiero con el gran objetivo de medir adecuadamente el riesgo. La normatividad contable IFRS 9 (International Financial Reporting Standard 9) publicado por la IASB (International Accounting Standards Board) brinda los lineamientos para que las entidades bancarias reflejen contablemente los instrumentos financieros, sus amortizaciones, provisiones y todos los rubros asociados a sus EEFF, conllevando a replantear el apetito de riesgo, estrategias de portafolio, políticas internas y todos los aspectos del ciclo de crédito (Filippo et al., 2017).

La normatividad IFRS 9 establece que cada entidad debe clasificar sus créditos en tres “Stages” donde el Stage 1 son aquellas obligaciones para las que se considera un riesgo de impago muy bajo, el Stage 2 es un riesgo medio y el Stage 3 es cuando existe un nivel de riesgo alto. La gran novedad de la IFRS 9 es que permite calcular la provisión de los créditos en un horizonte de tiempo futuro, es decir, que la provisión estará reflejando la incapacidad de pago futura, y no la actual. Con ello la entidad bancaria estará reflejando en una mejor medida el riesgo de los productos crediticios, estableciendo que para los créditos ubicados en el primer Stage, la pérdida esperada será a 12 meses y para los otros dos niveles, se calculará en la vida del crédito, siendo más alta en cada nivel. Adicionalmente, la provisión estará reflejando el riesgo asociado al sector donde opere el cliente, el tiempo del crédito donde se asume que a mayor plazo existirá un mayor riesgo,

los respaldos que permitan mitigar el riesgo de impago y la calificación del cliente (Filippo et al., 2017).

No obstante, aunque la normatividad IFRS 9 ayuda a mitigar el riesgo mediante sus lineamientos internacionales y permite reflejar la probabilidad de impago a 12 meses o en la vida del crédito, existen entidades bancarias que carecen de modelos predictivos para determinar la pérdida esperada de los clientes que afectan en una mayor proporción su balance, debido a que el gran movimiento de la provisión se explica en aquellos clientes que tienen una alta deuda con el banco y que su probabilidad de impago es demasiado alta, afectando significativamente la cartera vencida, sus castigos, y con ello las utilidades generadas en cada periodo financiero.

En el presente trabajo la variable respuesta corresponde a la pérdida crediticia esperada (ECL) en términos porcentuales, donde se pretende pronosticar la ECL de los clientes que tienen un mayor nivel de riesgo percibido por el banco y que su deuda con la entidad financiera sea significativa. Para ello, se pretende evaluar la incorporación de diferentes tipos de variables que han sido propuestas por diferentes autores para pronosticar la pérdida esperada de los clientes, planteando variables relacionadas a las garantías que se tengan como respaldo de su deuda, las cuales son fuertes predictores según lo planteado por Bastos (2010) y por Bandyopadhyay (2022) como el valor de garantía, los costos de comercialización, los costos de adjudicación, la valoración de acuerdo al tipo de inmueble, los tiempos asociados a la recuperación del activo. Adicionalmente, se buscará incorporar variables que fueron propuestas por Leow & Mues (2012) como el LTV, el Haircut relacionado al inmueble y el DLTV que fueron estadísticamente significativos en su análisis.

En el planteamiento del modelo también se evaluará la posibilidad de incorporar diferentes variables financieras, sectoriales e indicadores económicos, entre aquellas que fueron reportadas como variables significativas en las investigaciones de algunos autores, como por ejemplo, el EBIT, los intereses pagados y el WK, que han sido fuertes predictores según el estudio realizado por Härdle & Prastyo (2013), asimismo, indicadores como el EBITDA, los pasivos, el margen de utilidad neta y el KTNO, que han sido propuestos por Wang (2011), el rendimiento de los bonos, el promedio de la hora salarial de los empleados

del sector privado y el índice de confianza del consumidor investigado por Xia et al. (2021), la variación del PIB y la tasa bancaria que se encontraron estadísticamente significativas en el trabajo de Antonsson (2018), las tasas de retorno de la bolsa, el estado de ánimo en el mercado de valores y las tasas de interés que fueron investigados en una economía emergente por Mazibaş & Tuna (2017), el índice de precios para vivienda, el crecimiento del PIB y las condiciones del préstamo que fueron propuestas por Luong & Scheule (2022). Adicionalmente, se buscará incluir variables transaccionales, que relatan el comportamiento de entradas y de salidas de flujos, siendo significativas para Khandani et al., (2010) y variables macroeconómicas que según Taghiyeh et al. (2021) ayudaron a pronosticar la pérdida esperada, tales como permisos de construcción, los agregados monetarios M1 y M2, índices de gestores de compras, horas semanales trabajadas por empleados manufactureros y la tasa de desempleo.

Objetivos

Objetivo principal

- Proponer modelos paramétricos y no paramétricos para la estimación de la pérdida crediticia esperada de los clientes que cuentan con un nivel de exposición significativo y representan un alto nivel de riesgo para el banco.

Objetivos secundarios

- Proponer variables de tipo financiero, económico, transaccional, sectorial y de riesgo para pronosticar la pérdida crediticia de los clientes.

- Revisar los modelos y métodos de selección de variables que han sido propuestos desde la literatura para el pronóstico de la pérdida crediticia esperada.

- Determinar el mejor modelo que pronostique la pérdida esperada del grupo de clientes seleccionado y verificar que las relaciones de las variables independientes guarden sentido con la teoría económica.

1. Capítulo 1

Antecedentes

La revisión del estado del arte que se presenta en este capítulo fue realizada mediante la búsqueda en las siguientes bases de datos multidisciplinares: Science Direct, Wiley Online Library y SSRN de Elsevier, y, además, se utilizó el buscador abierto Google Scholar. La identificación de términos de búsqueda se hizo con base en el tema y el objetivo del estudio. Por lo tanto, se emplearon los términos: *expected credit loss, ECL, bank exposures, parametric models and credit risk, non-parametric models and credit risk, machine learning models and credit risk, ECL-IFRS9, default forecasting*, *bankruptcy*. Se aplicaron dos criterios de inclusión: 1) artículos económicos y 2) artículos de investigación, criterio que se aplicó verificando que los trabajos tuvieran la estructura propia de los artículos de investigación (introducción, métodos, resultados, discusión). También se aplicaron dos criterios de exclusión: 1) se excluyeron artículos que no emplearan indicadores financieros o macroeconómicos asociados a entidades bancarias y 2) artículos de revisión donde no se utilizaron modelos estadísticos.

Los estudios revisados van desde el año 1968 hasta el 2022, siendo la mayoría del 2010 en adelante. Se encontraron algunas problemáticas ampliamente estudiadas, las cuales han sido abordadas por diversos autores desde diferentes escenarios, entre las cuales se destacan: la tasa de default, el riesgo de incumplimiento, la bancarrota corporativa, el riesgo crediticio, la pérdida dado el incumplimiento (LGD), la pérdida crediticia esperada

(ECL), la tasa de recuperación, la calificación crediticia y la tasa de cancelación de las tarjetas de crédito.

Se evidencia que en la mayoría de los estudios revisados se tuvieron en cuenta variables macroeconómicas e indicadores financieros, tales como el PIB, el EBITDA, las tasas de interés, los índices de apalancamiento, los indicadores de rentabilidad, la tasa de desempleo, entre otras, para observar sus efectos o contribuciones sobre la variable respuesta; en casi todos los casos se reporta que estas variables, en especial las financieras, resultan ser buenas predictoras. En algunos trabajos también incluyeron variables de comportamiento de los consumidores como el saldo de la cuenta y los gastos, variables relacionadas con el ciclo de vida del cliente como lo mencionan Luong & Scheule (2022) y Rubaszek & Serwa (2014), las variables asociadas a las garantías propuestas por Bastos (2010), Bandyopadhyay (2022) y Leow & Mues (2012), y variables transaccionales estudiadas por Khandani et al., (2010).

Se observa que en algunos estudios se emplearon métodos para la selección de variables tales como: el Algoritmo WEKA (evaluador de subconjuntos de funciones basado en correlaciones), modelos de regularización como Lasso, Ridge y Elastic Net, eliminación de predictores por medio de la multicolinealidad entre las variables, regresión paso a paso hacia adelante y hacia atrás (Stepwise), el Test T, modelos de machine Learning (MLM), análisis factorial y análisis de componentes principales (PCA).

En lo que respecta a los modelos se encontró que los más comúnmente empleados son los modelos de Machine Learning (MLM) como las máquinas de soporte vectorial (SVM), los bosques aleatorios (RF), modelos de Boosting, entre otros, y frecuentemente se compara su eficacia con modelos tradicionales como el análisis discriminante múltiple (MDA), la regresión logística (LR), los mínimos cuadrados ordinarios (OLS) y el modelo de regresión de Cox. Resulta interesante que gran parte de los estudios que propusieron modelos de máquinas de soporte vectorial fueron aquellos en lo que se buscaba examinar la bancarrota corporativa.

A continuación, se presenta una tabla resumen con la información de los autores, la variable respuesta, los modelos empleados, algunos predictores, el método de selección

de variables y el principal resultado encontrado en la revisión del estado del arte. Es importante aclarar que, dado que existe una gran cantidad de modelos evaluados en los antecedentes, las abreviaciones usadas fueron las siguientes: Adaptive Boosting (AdaBoost), Algoritmo K-vecinos más cercanos (KNN), Análisis de Componentes Principales (PCA), Análisis Discriminante (DA), Análisis Discriminante Lineal (LDA), Análisis Discriminante Múltiple (MDA), Árboles de clasificación y regresión (CART), Árboles de decisión (DT), Árboles de regresión (RT), Bosques Aleatorios (RF), Gradient Boosting Machine (GBM), Heterogeneous Stacking Ensemble (HSE), Máquinas de Soporte Vectorial (SVM), Mínimos Cuadrados Ordinarios (OLS), Mínimos Cuadrados Ordinarios Dinámicos (DOLS), Mínimos Cuadrados Ordinarios totalmente modificados (FMOLS), Modelo Autorregresivo con rezagos (ARDL), Modelos de ciclo de vida (LCM), Modelos de Machine Learning (MLM), Redes Neuronales Artificiales (ANN) y Regresión Logística (LR).

Tabla 1-1: Resumen de los antecedentes considerando los autores, modelos, variables y resultados

Autor y año	Variable respuesta	Modelo planteado	Variables	Método selección de variables	Resultado
Altman (1968)	Bancarrota corporativa	Análisis discriminante múltiple (MDA)	Indicadores financieros como Capital de trabajo/Activos totales	1) Popularidad en la literatura 2) Relevancia en el estudio	El MDA tiene la capacidad de predecir la probabilidad de que una empresa entre en bancarrota con suficiente anticipación para permitir la toma de decisiones.
Antonsson (2018)	Frecuencia de default	Regresión lineal múltiple por medio de un estimador OLS.	Indicadores macroeconómicos como el PIB y el índice de precios de la vivienda (HPI)	No se realiza. Se toman los indicadores encontrados en la literatura.	Mediante MLM se encontró que el PIB y la tasa repo (RR) con varios meses de retraso, son variables significativas para explicar los cambios en la frecuencia del default de los créditos.
Bandyopadhyay (2022)	Pérdida dado el incumplimiento (LGD)	Modelo de regresión multifactorial tobit	Información relacionada a los sectores económicos de clientes y sus productos financieros	Variables independientes con p-valor menor a 5%.	El modelo para la predicción de la LGD es útil para los bancos, ya que les permite evaluar el riesgo de una línea de crédito y mejora la toma de decisiones.
Barboza et al., (2017)	Bancarrota corporativa	MLM como SVM, Bagging, Boosting y RF, vs Modelos	Variables financieras relacionadas a la liquidez rentabilidad y productividad	Las variables propuestas y seleccionadas,	Los MLM planteados tienen más precisión con relación a los modelos

Autor y año	Variable respuesta	Modelo planteado	Variables	Método selección de variables	Resultado
		tradicionales como DA, LR y ANN	como crecimiento de los activos y aumento de ventas	se basaron en la literatura.	tradicionales, el mejor modelo fue el de RF con un grado de acierto del 87%
Bastos (2010)	Tasa de recuperación de créditos	Árboles de regresión (RT), Regresiones de respuesta fraccional y Regresión logística	Variables explicativas como tamaño del préstamo, sector comercial y tasa de préstamo	No se realiza. Todas las variables de la base de datos son involucras a los modelos.	Los árboles de regresión mostraron mejores resultados en el corto plazo, y los modelos de regresión de respuesta fraccional tuvieron un mejor grado de ajuste en el largo plazo, usando el MSE y MAE como medidas de desempeño.
Bellotti & Crook (2013)	Calificación crediticia	Modelos discretos de supervivencia	Variables de comportamiento que varían en el tiempo (BV), variables macroeconómicas (MV) y variables de aplicación (AV)	De acuerdo al nivel se significancia estadística.	Las variables BV como el saldo de la cuenta, pagos electrónicos y el número de transacciones por mes son significativas y buenos predictores. Y las tasas de interés bancarias y de desempleo afectan significativamente el riesgo.
Danenas & Garsva (2015)	Riesgo crediticio	SVM, Regresión Logística (LR) y RBF (Radial Basis Function) Network Classifier	Indicadores financieros como cuentas por pagar/ventas	Algoritmo WEKA	El pronóstico de la bancarrota de las empresas tiende a estar bien explicado por medio de SVM, dado que fue el modelo con mayor grado de accuracy en comparación con los otros modelos.
Figlewski et al. (2012)	Default y calificación crediticia	Modelo de regresión de Cox	Variables macroeconómicas y variables relacionadas con la dirección de la economía, como la tasa de desempleo y el PIB	Regresión paso a paso hacia atrás con criterio AIC	Las transiciones al incumplimiento a partir de una calificación de grado especulativo son las que generan un mayor default corporativo.
Giesecke et al. (2011)	Tasa de default del mercado de bonos	Modelo de regímenes cambiantes	Variables financieras y macroeconómicas como el PIB	El análisis de los autores es de 1866 a 2008, por lo que se encontraron pocas variables. No se realiza reducción de predictores.	Los rendimientos de las acciones, la volatilidad de los rendimientos de las acciones y los cambios en el PIB son fuertes predictores de las tasas de incumplimiento.

Autor y año	Variable respuesta	Modelo planteado	Variables	Método selección de variables	Resultado
Härdle & Prastyo (2013)	Riesgo de default	Regresión Logística (LR)	Indicadores financieros relacionados a: rentabilidad, apalancamiento, estructura financiera, liquidez, actividad y dinamismo	Modelos de regularización, Lasso y red elástica	El enfoque de regularización en el modelo LR es capaz de estimar y seleccionar simultáneamente los predictores predeterminados con una alta precisión. Los predictores seleccionados por Lasso en algunos conjuntos de datos son significativamente más pequeños que los seleccionados por la red elástica.
Heo & Yang (2014)	Pronóstico de quiebra	AdaBoost, ANN, SVM, DT y regresión lineal múltiple	Indicadores financieros como (EBIT/total activos) y (EBT/capital).	No se realiza selección. Se incorporan todos los predictores al modelo.	El modelo AdaBoost mostró la mejor capacidad predictiva entre los modelos, especialmente para las empresas de gran tamaño, en la industria de la construcción. Teniendo los menores errores de tipo 1 y tipo 2.
Hiau et al. (2008)	Bancarrota corporativa	MDA, LR, Modelo de regresión de Cox	Índices de apalancamiento. índices de rentabilidad. índices de flujo de efectivo. Tamaño. Crecimiento	Regresión progresiva hacia adelante	La regresión de Cox es la que mejor grado de predicción obtuvo en sus datos con un 94.9% de aciertos (predicciones correctas sobre observaciones), seguido del MDA y por último LR. El índice de apalancamiento es un predictor importante en todos los modelos
Jiménez & Mencia (2009)	Pérdidas crediticias de un portafolio agregado	Modelo dinámico	Variables macroeconómicas como el PIB, y tasas de interés	Involucran variables con sus rezagos, sólo se dejan aquellas con un nivel de significancia mejor a 5%	Existen diferentes niveles de riesgo asociados a cada sector económico. Se evidenció un mayor riesgo de impago para los sectores hipotecario y de construcción.
Jondeau & Khalilzadeh (2022)	Pérdida crediticia estresada (SEL)	Modelo dinámico	Variables financieras como Crecimiento de la producción industrial y Crecimiento del empleo	No se realiza una selección. Se incorporan todos los predictores al modelo.	SEL brinda información de la inyección de capital necesaria para los bancos en periodos de crisis, y es útil para dar una predicción de la pérdida de capital de los bancos comerciales para escenarios adversos.

Autor y año	Variable respuesta	Modelo planteado	Variables	Método selección de variables	Resultado
Khandani et al., (2010)	Riesgo de crédito al consumidor	Árboles de clasificación y regresión generalizados (CART)	Información transaccional de los clientes como el número de transacciones. E información brindada por los burós de crédito	Sólo se eliminan los predictores que contienen registros con datos vacíos o faltantes.	Los pronósticos por medio de CART son considerablemente más adaptables y pueden captar la dinámica de los ciclos crediticios cambiantes, y los niveles absolutos de las tasas de incumplimiento.
Leow & Crook (2014)	Probabilidad de mora e incumplimiento	Modelos de intensidad basados en Modelos de supervivencia	Variables de comportamiento como los gastos y reembolsos. Y variables de aplicación como las características de la cuenta y del deudor	Se tomaron todas las variables y sus rezagos, no hay reducción de predictores.	Las variables comportamentales de los clientes tienen una incidencia directa en el riesgo de incumplimiento, y la mayoría de las variables de aplicación afectan el riesgo de mora
Leow & Mues (2012)	Pérdida dado el incumplimiento (LGD)	Modelo de Probabilidad de Recuperación (Probability of Repossession Model) y el Modelo de Descuento (Haircut Model)	Variables categóricas que brindan un rango como la relación del valor de la propiedad de acuerdo a la región, y variables cuantitativas como el valor del crédito sobre el colateral (LTV) y el valor porcentual de descuento del valor de la garantía (el Haircut)	Regresión logística con un método de selección hacia atrás y un test de Wald	Para los créditos hipotecarios las variables LTV, el Haircut y DLTV son fuertes predictores y estadísticamente significativos de acuerdo al modelo OLS. La combinación de los dos modelos propuestos permite obtener mayor R^2, evaluando en las modelos medidas como el MSE y MAE.
Liu & Xu (2003)	Riesgo de crédito al consumidor	Modelo vectorial autorregresivo	Indicadores económicos como la tasa de desempleo (UR), PIB y el índice de confianza del consumidor (CCI)	Regresión paso a paso hacia atrás.	Existe una relación inversa entre la estabilidad económica del país y la tasa de cancelación de tarjetas de crédito. Algunas variables que más inciden en la predicción de dicha tasa de cancelación son UR y CCI.
Luong & Scheule (2022)	Riesgo de crédito hipotecario	Modelo hibrido, combinando Modelos de ciclo de vida (LCM) con Modelos forward	Variables relacionadas con el ciclo de vida del cliente como la edad. Y variables financieras como la relación deuda-ingreso (DTI) y el PIB	Basado en plausibilidad económica y significancia estadística en pruebas univariadas y multivariadas.	Los modelos híbridos brindaron los mejores resultados, seguidos por los modelos forward y los LCM en predicciones de múltiples periodos futuros. El PIB, el índice de precios para vivienda y algunos factores externos, son significativos para pronosticar el riesgo de crédito

Autor y año	Variable respuesta	Modelo planteado	Variables	Método selección de variables	Resultado
Chen (2011)	Bancarrota corporativa	Modelos de inteligencia artificial como C5.0 (algoritmo basado en árboles), CART, SVM, vs Modelos tradicionales como LDA, LR	Indicadores financieros, indicadores no financieros y variables macroeconómicas, como el índice de apalancamiento y el índice de endeudamiento	Análisis factorial y PCA con una técnica de rotación (VARIMAX)	Las razones financieras tuvieron un efecto mayor en el desempeño de las predicciones que las razones no financieras y los índices macroeconómicos. Las SVM realizaron un mejor pronóstico en comparación con otros modelos, teniendo mejor ajuste en el largo plazo de acuerdo a la medida F, la cual combina precisión y recuperación.
Marinelli et al. (2022)	Riesgo crediticio	Regresión lineal	Variables relacionadas con la complejidad del banco. Y variables relacionadas a la diversificación de los ingresos por comisiones.	Análisis factorial	Los bancos más complejos se comprometen a tomar menos riesgos y también cobran tasas de interés más altas a los prestatarios más riesgosos, tanto a nivel nacional como internacional.
Mazibaş & Tuna (2017)	Préstamos al consumidor y tarjetas de crédito	OLS, Modelo OLS dinámico (DOLS) y el modelo OLS completamente modificado (FMOLS)	Indicadores macroeconómicos como el PIB, tasas de interés y el estado de ánimo en el mercado de valores	Pruebas de causalidad de Granger y el test de Wald.	El PIB, las tasas de retorno de la bolsa, el estado de ánimo en el mercado de valores y las tasas de interés son los que mejor predicen el aumento de los créditos en los productos investigados
Rubaszek & Serwa (2014)	Créditos en los hogares	Modelo de ciclo de vida (LCM)	Indicadores macroeconómicos como el PIB y tasas de interés. Y variables asociadas al ciclo de vida como el consumo	Nivel de significancia estadística.	El valor del crédito de los hogares está negativamente correlacionado con la persistencia de la productividad individual.
Taghiyeh et al. (2021)	Tasa de cancelación de las tarjetas de crédito	Gradient Boosting Machine (GBM), Random Forest (RF), Lasso y Ridge	Variables macroeconómicas como la tasa de desempleo y las horas semanales trabajadas	Regresión Lasso	Seis indicadores macroeconómicos y sus rezagos óptimos, son los que mejor grado de predicción generan al observar el MSE y un R^2=0.77 por medio del modelo GBM. Entre ellos los permisos de construcción, el desempleo inicial, los reclamos por seguros y el agregado monetario M1.
Telg et al. (2022)	Riesgo crediticio (rating y riesgo de	Modelo dinámico de transiciones de calificación.	Variables relacionadas con la calificación crediticia como grado de inversión. Y variables macroeconómicas	No se realiza. Se incluyen todas los predictores al modelo.	Usando un modelo dinámico de transiciones de calificación, los componentes no observables son muy útiles para capturar grandes

Autor y año	Variable respuesta	Modelo planteado	Variables	Método selección de variables	Resultado
	incumplimiento)		como el crecimiento anual de la producción industrial		oscilaciones por el riesgo de impago, en periodos de alto estrés financiero
Wang (2011)	Tasas de default corporativas	Modelo logístico mixto, Modelo de fragilidad de Cox	Variables macroeconómicas. Índices contables e indicadores de mercado. Como EBITDA, indicadores de liquidez y el tamaño de la firma	Test T y regresión progresiva hacia adelante y hacia atrás, basados en una regresión logística dinámica	Los dos modelos presentan una buena habilidad de predicción. El EBITDA, los indicadores de liquidez, el tamaño de la firma, la utilidad neta y el default histórico son las variables más significativas.
Xia et al. (2021)	Pérdida dado el incumplimiento (LGD)	Modelo HSE (Heterogeneous Stacking Ensemble), Regresión lineal, Ridge y otros modelos no paramétricos.	Variables macroeconómicas como la tasa de desempleo. Y datos relacionados con préstamos entre pares (P2P)	1) Selección predictores relacionados a la economía de EEUU. 2) Eliminación de predictores con multicolinealidad mediante el VIF. 3) valores de significancia en la regresión lineal.	Variables macroeconómicas como la tasa a tres meses de los bonos del tesoro, la tasa de desempleo y el promedio de la hora salarial de los empleados del sector privado son buenos predictores para estimar la LGD. Donde Modelo HSE fue el que obtuvo mejores medidas de desempeño en cuanto a MASE, MSE y R^2.
Yeh et al., (2014)	Bancarrota corporativa	Random Forest (RF), SVM y ANN	Índices financieros. Indicadores no financieros. Indicadores relacionados con el capital intelectual (IC) como ventas/empleados	Importancia de las variables por medio del modelo Bosques aleatorios	Un modelo híbrido entre RF y teoría de conjuntos aproximados brindó mejores resultados de predicción en comparación con otros modelos y utilizó menos variables.
Zhang & Chen (2021)	Riesgo de default	Algoritmo Xgboost, KNN, LR, SVM, DT	Indicadores financieros como retorno de los activos, y EBITDA/Ingresos Totales	Se eligen todos los predictores de acuerdo la literatura.	Incluso en conjuntos de datos muy desequilibrados, XGboost obtiene mejores resultados que los algoritmos de clasificación tradicionales.

2. Capítulo 2

Modelos y selección de variables

2.1 Modelos

La pérdida crediticia esperada fue estudiada por múltiples autores en la revisión del estado del arte, donde la gran mayoría de los autores planteaba modelos paramétricos y no paramétricos en sus estudios, comparando el desempeño de estos modelos en el conjunto de datos de entrenamiento y de prueba, determinando la calidad en el ajuste y en la predicción, respectivamente. En el presente trabajo se tiene un gran cumulo de predictores que no presentan fuertes relaciones lineales con la variable respuesta, por lo que no es fácil establecer relaciones paramétricas entre las variables independientes y la pérdida crediticia esperada, aun así, se plantea el modelo de regresión lineal múltiple como un modelo de línea base, aunque se sepa desde un principio que tal vez no es la mejor opción en este caso en particular. Por otro lado, se plantean modelos no paramétricos que permiten simplificar la modelación en el sentido que no obliga a establecer la forma funcional en la que deberían entrar las variables predictoras al modelo.

2.1.1 Regresión lineal múltiple

La regresión lineal múltiple es un método estadístico que se utiliza para modelar la relación entre dos o más variables. Este modelo se basa en la suposición de que la variable respuesta es una combinación lineal de las variables independientes más un término de

error aleatorio (Maulud & Abdulazeez, 2020). La ecuación de la regresión lineal múltiple se define como:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon, \quad (2\text{-}1)$$

donde y representa la variable de respuesta, β_0 es el intercepto, β_1, β_2,..., β_p son los coeficientes de regresión para las variables predictoras x_1, x_2,..., x_p, y ε es el término de error aleatorio que sigue una distribución normal con media cero y varianza constante (Maulud & Abdulazeez, 2020).

Estimación por mínimos cuadrados ordinarios

La estimación por mínimos cuadrados ordinarios, o OLS por sus siglas en inglés Ordinary Least Square, es un método estadístico que se utiliza para estimar los parámetros desconocidos en un modelo de regresión lineal. El objetivo de la regresión lineal múltiple es estimar los valores óptimos de los coeficientes de regresión que minimizan la suma de los cuadrados de los errores de ajuste, esto se logra mediante el método OLS, que encuentra los valores de los coeficientes que maximizan la variabilidad explicada en la variable dependiente (Maulud & Abdulazeez, 2020).

Métodos de estimación regularizada

La regularización es una técnica utilizada para reducir el sobreajuste del modelo que permite reducir el conjunto de predictores, reducir el error en la predicción y mejorar su interpretabilidad (Fonti & Belitser, 2017). Cabe resaltar que evitar el sobreajuste en un modelo es muy importante debido a que cuando un modelo se ajusta demasiado a los datos de entrenamiento, pierde la capacidad de generalización para el conjunto de prueba, teniendo un bajo nivel de predicción en los datos nuevos. La regularización implica la adición de un término de penalización en la función de pérdida del modelo que reduce los coeficientes de las variables predictoras, lo que disminuye la complejidad del modelo y, por lo tanto, la probabilidad de sobreajuste (Hastie et al., 2001).

Los métodos de estimación regularizada incorporan términos de penalización para controlar los coeficientes del modelo, pero todos los métodos que se explican a continuación parten de una regresión lineal múltiple.

Regresión Lasso

Autores como Härdle & Prastyo (2013) y Taghiyeh et al. (2021) propusieron en sus investigaciones la regresión Lasso para la selección de variables. Lasso es la abreviación de Least Absolute Shrinkage and Selection Operator, siendo un modelo que impone una restricción a la suma de los valores absolutos de los parámetros, donde la suma debe ser menor que un límite superior establecido. Lasso utiliza un proceso de regularización donde penaliza los coeficientes de las variables de regresión reduciendo algunas de ellas a cero, y aquellas variables con un coeficiente diferente de cero después del proceso de regularización se seleccionan para formar parte del modelo, minimizando así el error en la predicción (Fonti & Belitser, 2017).

Es importante señalar que este modelo tiene un parámetro (λ) que controla la fuerza de la penalización, el cual brindará resultados del modelo muy diferentes si el valor es cercano a cero, o si es muy grande. Un parámetro igual a cero, será equivalente a realizar una Regresión por Mínimos Cuadrados Ordinarios (OLS), pero si el parámetro es suficientemente grande, los coeficientes en el proceso de regularización serán obligados a ser iguales a cero y se tendrá una drástica reducción de dimensionalidad (Fonti & Belitser, 2017). Matemáticamente, es un modelo lineal que tiene como objetivo minimizar la siguiente función:

$$\frac{1}{2 \cdot \mathrm{n}_{observaciones}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 + \lambda \cdot ||\boldsymbol{w}||_1. \tag{2-2}$$

En esta fórmula, $\mathrm{n}_{observaciones}$ representa el número de observaciones en los datos, $\boldsymbol{y}$ corresponde al vector de la variable respuesta, $\boldsymbol{X}$ es la matriz de variables independientes, $\boldsymbol{w}$ es el vector de coeficientes y λ es el parámetro de regularización (Pedregosa et al., 2011). El primer término de la fórmula representa la suma de los cuadrados de los errores, mientras que el segundo término es el término de regularización L1 que penaliza los

coeficientes no relevantes, lo que promueve la selección de variables, reduce la complejidad del modelo y ayuda a prevenir el sobreajuste (Giraud, 2021).

El modelo Lasso tiene como ventaja su buen nivel de predicción, dado que al regularizar las variables y eliminar las que no aportan al modelo, reduce significativamente la varianza sin aumentar el sesgo. Adicionalmente, brinda una alta interpretabilidad al tener un menor número de coeficientes que están relacionados a la variable respuesta, disminuyendo el sobreajuste (Fonti & Belitser, 2017).

Regresión Ridge

La regresión Ridge fue propuesta por Hoerl & Kennard (1970) como una generalización de la regresión de mínimos cuadrados ordinarios, pero incluyendo un parámetro de regularización para la optimización del problema, también conocido como L2. El parámetro controla la fuerza de la regularización y con ello mejora la robustez del modelo y su generalización (Dupré la Tour et al., 2022). Los coeficientes de la regresión Ridge minimizan la penalización de la sumatoria de los errores al cuadrado, el cual tiene como objetivo minimizar la siguiente función:

$$||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 + \alpha \cdot ||\boldsymbol{w}||_2^2. \quad (2\text{-}3)$$

En la fórmula anterior $\boldsymbol{y}$ corresponde al vector de la variable respuesta, $\boldsymbol{X}$ la matriz o el conjunto de datos de los predictores, $\boldsymbol{w}$ el vector de coeficientes y α es el parámetro de regularización. La primera parte de la fórmula representa la suma de los cuadrados de los errores, mientras que el segundo término corresponde a la regularización Ridge (L2) que penaliza los coeficientes grandes (Pedregosa et al., 2011). Es importante mencionar que por medio de esta regresión se simplifica el modelo al reducir los pesos de los coeficientes que carecen de importancia, promoviendo la estabilidad y evitando el sobreajuste. Un valor de α igual a cero se obtendrá los mismos resultados de una regresión de mínimos cuadrados ordinarios (Wang et al., 2023).

Diferencia entre Ridge y Lasso

La regresión Ridge y la regresión Lasso son dos técnicas de regularización que se utilizan en problemas de regresión cuando hay un gran número de predictores y se quiere evitar el sobreajuste y la multicolinealidad. La principal diferencia entre ambas técnicas es la forma en que se aplican las penalizaciones en los coeficientes de la regresión (Melkumova & Shatskikh, 2017).

La regresión Ridge agrega una penalización a la suma de los cuadrados de los coeficientes de la regresión (L2), que se multiplica por un parámetro de regularización α. La regresión Lasso, por otro lado, agrega una penalización a la suma de los valores absolutos de los coeficientes de la regresión (L1), que, análogamente a la regresión Ridge, se multiplica por un parámetro de regularización λ. La diferencia fundamental entre ambas técnicas radica en la forma en que se aplican las penalizaciones en los coeficientes de la regresión, la penalización L2 hace que los coeficientes sean pequeños, pero no necesariamente cero. En cambio, la penalización L1 hace que algunos coeficientes sean iguales a cero, lo que implica que algunas variables no sean consideradas en el modelo (Pereira et al., 2016).

2.1.2 Métodos de ensamble

Son técnicas que combinan múltiples modelos individuales para formar un modelo más robusto y preciso, como es el caso de los Bosques Aleatorios que se basan en la combinación de árboles de decisión con el fin de mejorar la capacidad de generalización y la precisión de las predicciones en comparación con un único modelo (Breiman, 2001). En el presente trabajo se propondrán los siguientes métodos de ensamble: Bosques Aleatorios, Extremely randomized trees y Extreme Gradient Boosting.

Bosques Aleatorios (RF)

Los Bosques Aleatorios o Random Forest, son un tipo de modelo de aprendizaje automático utilizado para realizar predicciones numéricas en función de un conjunto de variables de entrada. Los bosques aleatorios se construyen mediante la división recursiva del conjunto de datos de entrenamiento en subconjuntos más pequeños en función de los

valores de las variables de entrada (Loh, 2011). En cada división se elige la variable de entrada que mejor separa los datos en subconjuntos y se establece un umbral de corte para esa variable. Una vez que se han creado las divisiones, se calcula la salida numérica promedio en cada hoja del árbol, la cual, es la combinación de árboles de predicción donde cada árbol depende del valor de un vector aleatorio elegido independientemente, pero con la misma distribución de todos los árboles en el bosque (Breiman, 2001).

Árboles Extremadamente Aleatorizados (Extra Trees)

Extremely randomized trees, conocido como Extra Trees es un modelo de aprendizaje automático de ensamble que se basa en la construcción de múltiples árboles de decisión. A diferencia de otros modelos de ensamble, el modelo Extra Trees se construye utilizando bootstrap y divide los nodos eligiendo puntos de corte completamente al azar utilizando toda la muestra de aprendizaje. El modelo es robusto y resistente al sobreajuste, ya que utiliza un conjunto de árboles no correlacionados e independientes para producir un modelo menos propenso a seleccionar variables irrelevantes con una alta eficiencia computacional, siendo una gran alternativa para problemas con alta dimensionalidad (Geurts et al., 2006).

La ventaja de los árboles de regresión es que son fácilmente interpretables y pueden proporcionar información sobre la relación entre las variables de entrada y la variable de salida, tienen la facilidad de manejar datos no lineales y faltantes, y son menos susceptibles al sobreajuste que algunos modelos más complejos (Hastie et al., 2001). Adicionalmente, según lo planteado por Breiman (2001) los bosques aleatorios son robustos para datos atípicos, brindando información acerca del error y de la importancia de cada variable, siendo muy útiles para la selección de variables.

Diferencias entre el modelo Extra Trees y Bosques Aleatorios

Es importante resaltar que ambos modelos son de ensamble y están basados en árboles de regresión, pero tienen diferencias en su implementación y elaboración. Una de las principales diferencias es cómo se construyen los árboles. En los bosques aleatorios se utiliza un bootstrap para crear diferentes subconjuntos de datos de entrenamiento, y para cada subconjunto, se construye un árbol de decisión utilizando el algoritmo de selección

de características más importante (Breiman, 2001). Mientras en el modelo Extra Trees, los árboles se construyen utilizando bootstrap como técnica de remuestreo, pero se usa un umbral aleatorio de corte para cada árbol para seleccionar la mejor variable en cada nodo (Genuer et al., 2010). Debido a esta aleatoriedad en la construcción de árboles, Genuer et al. (2010) plantean que el modelo Extra Trees puede ser más rápido en cuanto a la eficiencia computacional y menos propenso al sobreajuste que el modelo de bosques aleatorios.

Refuerzo de Gradientes Extremo (XGBoost)

Extreme Gradient Boosting o también conocido como XGBoost, es un algoritmo de aprendizaje automático basado en árboles de decisión que se utiliza para problemas de clasificación y regresión, donde se resalta su gran capacidad para manejar conjuntos de datos grandes y complejos, su poder de predicción y su eficiencia en términos de tiempo de entrenamiento (Rory et al., 2018).

En este modelo, los árboles de decisión se construyen de forma secuencial donde cada árbol se ajusta a los residuos generados por los árboles anteriores, mejorando su precisión. Además, se utiliza una combinación de regularización L1 y L2 para controlar la complejidad del modelo para prevenir el sobreajuste, incorporando términos de penalización en la función de pérdida que permite restringir los coeficientes (Chen & Guestrin, 2016). Matemáticamente se puede expresar así:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\boldsymbol{x_i}), f_k \in \mathcal{F}. \tag{2-4}$$

En esta fórmula, $\hat{y}_i$ representa la predicción del modelo para cada conjunto de predictores $\boldsymbol{x_i}$. La predicción se obtiene sumando las predicciones de todos los árboles en el ensamble, donde K es el número total de árboles y f_k representa un árbol de decisión en el espacio de funciones $\mathcal{F}$ (Nalluri et al., 2020).

Cada árbol f_k en XGBoost se construye de manera secuencial y se ajusta a los residuos de los árboles anteriores para mejorar la precisión del modelo. Los árboles se construyen

utilizando una técnica de boosting que optimiza una función de pérdida específica, como la pérdida cuadrática para problemas de regresión (Chen & Guestrin, 2016).

Importancia de las variables

A continuación, se explica cómo se calcula la importancia de los predictores y su significado dentro de los modelos Bosques aleatorios (RF), Árboles Extremadamente Aleatorizados (Extra Trees) y Refuerzo de Gradientes Extremo (XGBoost).

En los modelos RF y Extra Trees

En estos modelos se evalúa la relevancia de las variables de entrada en el modelo mediante la importancia relativa que tiene cada variable independiente en su capacidad de predicción de la variable respuesta (Loh, 2011). La importancia se puede interpretar como una medida de la contribución de cada variable al poder predictivo del modelo, aunque no proporcione información sobre la dirección de la relación entre la variable independiente y la dependiente, aquellos predictores con una importancia más alta se consideran más relevantes para la predicción del resultado. Adicionalmente, es útil para simplificar el modelo, dado que si se busca reducir la dimensionalidad se puede elegir un umbral y dejar aquellas variables cuya importancia sea mayor al umbral establecido (Louppe, 2014). Esta medida se calcula mediante la evaluación de la disminución de la impureza de Gini o la ganancia de información obtenida por cada variable en todas las divisiones del árbol, cuanto menor sea el índice de Gini, se considera que dicha variable es más importante para el modelo dado que reduce significativamente la impureza de los nodos del árbol (Breiman, 2001). Matemáticamente el índice de Gini, se puede expresar como:

$$Gini = 1 - \sum_{i=1}^{C} p_i^2, \tag{2-5}$$

donde $Gini$ es el índice de Gini para un nodo dado el cual oscila entre 0 y 1, C es el número de categorías, y p_i es la proporción de observaciones en el nodo que pertenecen a la clase i. El índice de Gini se calcula sumando el cuadrado de las proporciones de cada clase en el nodo y luego restándolo de 1, cuanto más cercano a 0 sea el valor del índice de Gini, más puro es el nodo (Hastie et al., 2009). la impureza se refiere a qué tan mezcladas están las clases en un nodo, donde un mayor valor del índice Gini lleva a una menor separación

de clases, siendo el índice de Gini la medida que indica la impureza que tiene cada nodo (Pedregosa et al., 2011). Según la documentación de la función de Scikit Learn en Python tomada de Pedregosa et al. (2011) la importancia se halla así:

$$n_{ij} = w_j C_j - w_{\text{left(j)}} C_{\text{left(j)}} - w_{\text{right(j)}} C_{\text{right(j)}}, \tag{2-6}$$

donde:

n_{ij} Es la importancia de la variable X_i.

w_j : Se refiere al número ponderado de muestras en el nodo j.

C_j : Es la impureza del nodo j.

$w_{\text{left(j)}}$: Se refiere al número ponderado de muestras en el nodo hijo izquierdo de j.

$C_{\text{left(j)}}$: La impureza del nodo hijo izquierdo de j.

$w_{\text{right(j)}}$: El número ponderado de muestras en el nodo hijo derecho de j.

$C_{\text{right(j)}}$: La impureza del nodo hijo derecho de j.

Considerando que n_{ij} es un valor no negativo, luego se halla la importancia de la variable independiente:

$$fi_i = \frac{\sum_{j:\text{nodo } j \text{ se divide en la variable } i} n_{ij}}{\sum_{j \in \text{ todos los nodos}} n_{ij}}. \tag{2-7}$$

En esta fórmula, fi_i representa la importancia de la variable para cada i. n_{ij} es la importancia de la variable i en el nodo j. Por último, la suma se realiza sobre todos los nodos donde se realiza la división de variables. Luego de realizar el cálculo anterior lo que se hace es sumar el valor de fi_i sobre todas las variables, es decir $\sum_{i \in \text{ todas las variables}} fi_i$, para que los valores de la importancia se encuentren entre 0 y 1 (Pedregosa et al., 2011).

En el modelo XGBoost

El modelo XGBoost cuantifica la importancia de una variable en un árbol de decisión mediante la mejora empírica en el error cuadrático que se logra al dividir los nodos del árbol utilizando esa variable. La suma se realiza sobre los nodos no terminales del árbol,

y el resultado es una medida de la contribución de cada variable al rendimiento predictivo del modelo (Hastie et al., 2009). Matemáticamente puede expresarse como:

$$I_j^2(T) = \sum_{j=1}^{j-1} \hat{\delta}_t^2 \times 1(v_t = j). \tag{2-8}$$

En esta fórmula, $I_j^2(T)$ representa la importancia al cuadrado de la variable j en el árbol T, $\hat{\delta}_t^2$ es la mejora empírica en el error cuadrático como resultado de la división en el nodo t, y v_t es la variable de división asociada con el nodo t. La suma se realiza sobre los nodos no terminales t del árbol T y $1(v_t = j)$ es una función indicadora que toma el valor 1 si la variable de división del nodo t es igual a j y 0 de lo contrario (Friedman, 2001). La expresión $\hat{\delta}_t^2$ se refiere a la mejora del error cuadrático en el conjunto de entrenamiento, hallando la impureza antes de la división en el nodo t, y luego de la división en el nodo t. Un mayor valor de $\hat{\delta}_t^2$ significa una mejora en la reducción del error cuadrático al considerar esa división en el nodo t.

La importancia se puede interpretar como la contribución que tiene cada variable en el modelo, donde las variables con mayor importancia tienen un mayor impacto en la capacidad predictiva del modelo. Es decir, las variables con mayor ganancia de impureza tienen un mayor efecto en el modelo XGBoost.

Interpretabilidad de los resultados de los métodos de ensamble usando SHAP Values

Para los modelos como la regresión lineal ordinaria, o regularizada como Ridge o Lasso, los coeficientes del modelo y su signo brindan un fácil entendimiento de cómo está afectando cada variable predictora a la variable repuesta. Pero, en modelos no paramétricos, o también llamados de “caja negra”, la interpretación se torna más compleja dado el funcionamiento de los modelos. Por lo tanto, para generar un entendimiento de los modelos de ensamble, se usarán los SHAP values para observar cómo están interactuando las variables predictoras en la pérdida crediticia esperada.

Los SHAP values, Shapley Additive Explanation, son una técnica basada en teoría de juegos que se utiliza para asignar la contribución de cada variable a la predicción de un modelo de aprendizaje automático. En esta teoría, se asigna una contribución justa a cada jugador en una coalición, teniendo en cuenta todas las posibles combinaciones de jugadores. En términos matemáticos, se pueden definir como la contribución promedio de una variable en todas las posibles combinaciones de variables (Lundberg & Lee, 2017), a continuación, la fórmula de los SHAP values:

$$\boldsymbol{\phi}_i = \sum_{\boldsymbol{S}\subseteq \boldsymbol{F}\setminus\{i\}} \frac{|\boldsymbol{S}|!\,(|\boldsymbol{F}| - |\boldsymbol{S}| - 1)!}{|\boldsymbol{F}|!} \left(f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)\right). \tag{2-9}$$

En esta fórmula, $\boldsymbol{\phi}_i$ representa el valor de Shapley para la variable i. $|\boldsymbol{S}|$ denota el número de variables predictoras sin incluir la variable i, tomadas del conjunto total de variables independientes representado como $\boldsymbol{S}$, $|\boldsymbol{F}|$ es el número total de predictores en el conjunto de datos, $f_{S\cup\{i\}}(x_{S\cup\{i\}})$ es la predicción del modelo cuando se incluyen todas las variables en $\boldsymbol{S}$ junto con la variable i, y $f_S(x_S)$ es la predicción del modelo solo con las variables de $\boldsymbol{S}$, es decir, sin incluir la variable i (Hastie et al., 2009).

Los valores de SHAP se calculan tomando la diferencia entre las predicciones cuando se incluye la variable i y cuando no se incluye la variable i, brindando una medida de la contribución de dicha variable en la predicción en comparación con todas las demás combinaciones de variables posibles. Adicionalmente, generan una interpretación global al observar la relación que tienen las variables independientes en la predicción, donde las variables pueden contribuir positiva o negativamente en el resultado final, pero este análisis también permite una interpretación local, ya que genera información de cómo cada variable independiente afecta a cada predicción individual (Meng et al., 2021).

2.2 Métodos de selección de variables

A continuación, se describen brevemente los métodos utilizados en este trabajo para la selección de variables predictoras con miras a la obtención de un conjunto reducido que contribuya al mejor ajuste y pronostico en cada uno de los modelos propuestos.

2.2.1 Selección de variables por medio de la importancia de los modelos

Los Bosques Aleatorios y Extra Trees son métodos se ensamble que ofrecen beneficios para la selección de variables, debido a que proporcionan una medida de la importancia de las variables utilizadas en el modelo, capturan relaciones no lineales entre la variable respuesta y los predictores, son capaces de manejar conjuntos de datos con una alta dimensionalidad, y son robustos para manejar datos atípicos (Loh, 2011). Por medio de la importancia que tiene cada variable independiente, la cual está en un rango de 0 a 1, se identifican cuáles son los predictores que tienen una mayor relevancia dentro del modelo y que se deben considerar como un buen predictor, y por tanto, permite identificar cuáles se deberían descartar del conjunto de variables independientes dada su baja importancia.

2.2.2 Regresión paso a paso

La selección de variables secuenciales, también conocida como Stepwise o SFS por sus siglas en inglés Sequential Feature Selector, es una técnica de selección de variables donde se reduce el espacio original de variables de dimensión n a un subespacio de variables de dimensión k. Este método construye conjuntos de variables de manera iterativa buscando mejorar la calidad del ajuste y puede emplearse en tres formas: hacia adelante (forward), hacia atrás (backward) y en ambas direcciones (stepwise). El modo hacia atrás, parte del espacio original con dimensión n hasta llegar al subespacio con dimensión k, optimizando la función objetivo en cada iteración. Por otro lado, la selección de variables hacia adelante parte de un espacio vacío añadiendo una variable por iteración que contenga la mayor calidad de ajuste, hasta llegar a la dimensión k (Bemister-Buffington et al., 2020). En el presente trabajo se plantea una regresión paso a paso hacia atrás usando el criterio AIC para la inclusión y/o exclusión de variables.

Eliminación recursiva de variables

La eliminación recursiva de variables o RFE por sus siglas en inglés Recursive Feature Elimination, es un método de selección de variables que brinda un subconjunto de variables para la construcción de modelos. El objetivo de este método es seleccionar variables considerando subconjuntos cada vez más pequeños. Para ello, el estimador se entrena con un conjunto de variables iniciales y aquellas que menos contribuyan al modelo

se eliminan del conjunto, este procedimiento se repite iterativamente hasta que llega a un número final de variables (Pedregosa et al., 2011).

Al seleccionar solo las variables más importantes, se logra llegar a el subconjunto de predictores que maximizan el rendimiento, mejoran la precisión y la capacidad de generalización del modelo. Este método al ser recursivo, no deja por fuera variables que pueden tener una importancia diferente cuando se evalúa sobre un subconjunto diferente de variables durante el proceso de eliminación, este proceso de selección consiste en únicamente tomar las primeras variables con mayor medida de importancia para el modelo (Granitto et al., 2006).

Entradas:

Conjunto de entrenamiento: $\mathbf{T}$

Conjunto inicial de predictores: $\mathbf{F} = \{f_1, \dots, f_p\}$

Salidas:

Ranking final de predictores: $\mathbf{R}$

Procedimiento:

Repetir para i en $\{1, \dots, p\}$

- Entrenar el modelo en el conjunto de entrenamiento $\mathbf{T}$ utilizando los predictores en $\mathbf{F}$.
- Evaluar la importancia de las variables en $\mathbf{F}$ utilizando alguna medida de importancia M obteniendo $M(\mathbf{T}, \mathbf{F})$. En el presente trabajo se usa la medida MSE.
- Ordenar las variables en $\mathbf{F}$ en orden decreciente de importancia.
- $f^* \leftarrow$ última variable del ordenamiento en $\boldsymbol{F}$
- Eliminar la variable menos importante, f^*, de $\mathbf{F}$.
- Guardar la variable eliminada en el ranking final: $\mathbf{R}(p - i + 1) \leftarrow f^*$
- $\boldsymbol{F} \leftarrow \boldsymbol{F} - f^*$

Figura 2-1: Selección recursiva de variables basado en el gráfico presentado por Granitto et al., (2006).

Los autores Granitto et al. (2006) explican cómo funciona el método RFE, el cual parte del entramiento de un modelo especificado, y se evalúa la importancia de cada variable en función de su capacidad para mejorar la precisión del modelo. Esta evaluación se puede realizar utilizando diferentes técnicas, como la importancia, el coeficiente de determinación, error cuadrático medio (MSE), y otros. Posteriormente, se elimina las características con menor importancia y se forma un nuevo conjunto de datos, así iterativamente, hasta llegar al conjunto de variables final (Guyon et al., 2002).

Diferencias con la regresión paso a paso

Ambos métodos, regresión paso a paso y RFE, son algoritmos de selección de variables, pero la eliminación recursiva de variables ajusta un modelo inicial con todas las variables y luego se van eliminando las menos importantes de forma iterativa, hasta alcanzar el número deseado de variables o un umbral de precisión predefinido (Granitto et al., 2006). Mientras la regresión paso a paso, construye de manera secuencial un subconjunto de variables, donde en cada iteración se evalúan todas las características no seleccionadas y se agrega la que tenga el mejor rendimiento según una determinada métrica (Bemister-Buffington et al., 2020). En otras palabras, si se compara una regresión paso a paso hacia atrás con una eliminación recursiva, la diferencia radicaría en el enfoque de búsqueda y en el ordenamiento de los predictores, pero ambos parten de un conjunto completo de predictores y eliminan variables en cada paso o iteración. Por último, la eliminación recursiva de variables al ser un enfoque de eliminación iterativa, puede ser más costoso computacionalmente que un método secuencial (Pedregosa et al., 2011).

3. Capítulo 3

Variables y datos

En el presente Capítulo se define la variable respuesta y las variables predictoras relacionadas a la transaccionalidad del cliente, a su nivel de riesgo, a los respaldos que se tengan como garantía de su deuda, a indicadores sectoriales, a variables financieras que han sido significativas en la revisión del estado del arte y, por último, a indicadores macroeconómicos propuestos por diferentes autores, que mejor han explicado la predicción de la pérdida crediticia esperada pero adaptadas al contexto económico colombiano. Posteriormente, se realiza una selección de variables mediante los métodos propuestos por los autores Taghiyeh et al. (2021), Wang (2011) y Härdle & Prastyo (2013), para obtener modelos parsimoniosos y evitar tanto como sea posible un alta multicolinealidad entre las variables explicativas. Adicionalmente, se explica la variable respuesta y los datos que están siendo usados para el pronóstico de la pérdida crediticia esperada.

3.1 Variable respuesta

La variable respuesta es la pérdida crediticia esperada de los clientes que cuentan con un nivel de exposición significativo y representan un alto nivel de riesgo para el banco. Esta variable es una proporción, es decir, es un porcentaje que está en función del monto adeudado por la contraparte, indicando cuánto la entidad financiera debe reflejar en su cartera para proteger su estabilidad económica ante un evento de impago, tomando valores entre 0 y 1 (Antonsson, 2018).

Dado que la variable respuesta se encuentran en un rango de 0 a 1, no se puede presumir normalidad ya que se encuentra restringida en un intervalo, sin embargo, se busca por medio de análisis gráficos observar el comportamiento de la variable respuesta. A continuación, se muestra la distribución de la variable:

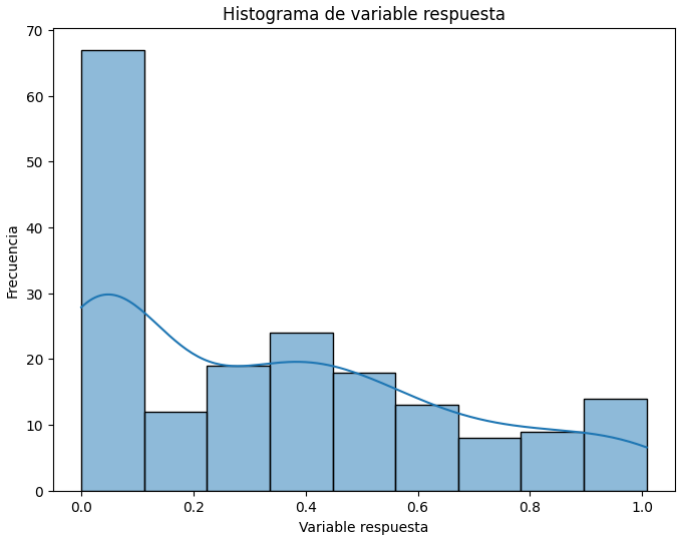


Figura 3-1: Histograma de la variable respuesta

En el histograma de la variable respuesta se observa que se tiene una concentración de valores en porcentajes entre 0 a 0.1, evidenciando que la mayor parte de los clientes cuentan con un nivel de pérdida crediticia esperada relativamente bajo. Además, se tiene una distribución muy asimétrica en la muestra de datos seleccionada. Por ello, la normalidad no será un supuesto que se va a cumplir en los modelos lineales, los cuales son modelos de base para los métodos de regularización. Es importante señalar que, en la revisión del estado del arte, diferentes autores plantearon modelos lineales en sus análisis, por lo cual, también en este trabajo se plantean dichos modelos con el fin de determinar qué resultados se obtienen.

3.2 Variables explicativas

Se seleccionaron un total de 437 variables para la predicción de la pérdida crediticia esperada, distribuidas en los siguientes grupos: variables macroeconómicas, de riesgo,

financieras, transaccionales, descriptivas y sectoriales. Cada variable se identificó de acuerdo al grupo al que pertenece, tal como se muestra en la siguiente tabla:

Tabla 3-1: Clasificación de variables por grupos

Grupo de variables	Inicial grupo	Número de variables
Variables de riesgo	RG	106
Variables financieras	FIN	154
Variables transaccionales	TRX	29
Variables macroeconómicas	MEC	117
Variables descriptivas y sectoriales	DyS	31

Como se puede observar en la tabla anterior, se tienen 437 variables entre los grupos de riesgo, financiero, transaccional, macroeconómico y descriptivas. Teniendo un mayor número de variables las financieras, macroeconómicas y de riesgo.

3.2.1 Variables macroeconómicas

Se evalúa el impacto que tienen las variables macroeconómicas como el IPC, la TRM, el PIB, y otras, con sus respectivos rezagos en la variable respuesta, incluyendo en el análisis hasta 6 meses de rezago en cada indicador macroeconómico, tal como lo plantearon Taghiyeh et al. (2021) en su investigación, quienes identifican el valor del rezago en estas variables predictoras más significativo para la variable respuesta. Las variables macroeconómicas consideradas en este trabajo, se eligieron considerando los estudios de Liu & Xu (2003), Giesecke et al. (2011), Antonsson (2018), Xia et al. (2021), Luong & Scheule (2022), entre otros autores. En este grupo se encontró un total de 117 variables, incluyendo la tasa de inflación mensual, el PIB y la tasa de cambio (Ver Tabla 8-1).

3.2.2 Variables de riesgo

En este grupo se encuentran las variables que indican el grado de riesgo que percibe la entidad financiera hacia el cliente, tales como la calificación interna, los días de incumplimiento, el Stage por IFRS9, entre otras. Adicionalmente, se consideran los montos adeudados por el cliente hacia la entidad financiera y sus garantías, donde se espera que los clientes con un mayor nivel de riesgo para la entidad financiera presenten una mayor

pérdida crediticia esperada. Es importante resaltar que existe un total de 106 variables en este grupo (Ver Tabla 8-2).

3.2.3 Variables financieras

Los estados financieros contienen una gran cantidad de información relacionada a la liquidez, rentabilidad y viabilidad económica de los clientes, los cuales tienen un impacto directo en la percepción de riesgo que genera y, por ende, en su pérdida crediticia esperada. Autores como Härdle & Prastyo (2013), Wang (2011), Zhang & Chen (2021), Chen (2011), Yeh et al. (2014), entre otros, propusieron indicadores financieros en su estudios, las cuales resultaron de gran importancia en sus análisis. Dada la gran cantidad de cuentas contables e indicadores financieros, este es el grupo con mayor número de variables, teniendo un total de 154, entre ellas el margen EBITDA, el KTNO, ROE, ROA, etc. (Ver Tabla 8-3).

3.2.4 Variables transaccionales

Khandani et al., (2010) propusieron diferentes variables relacionadas a la transaccionalidad de los clientes, dentro de este grupo se encuentran las variables que describen el comportamiento de flujos de dinero entrantes y salientes de cada cliente. Se espera que, ante un aumento significativo de flujos entrantes, estos se vean reflejados en un mayor nivel de ingresos en los próximos periodos, caso contrario, si fuese un aumento significativo de flujos salientes, estos podrían asociarse con un mayor costo de ventas y/o gastos operacionales. Cabe resaltar que los montos transaccionales están ligados al saldo que tenga el cliente con la entidad financiera. En este grupo se identificaron un total de 29 variables (Ver Tabla 8-4).

3.2.5 Variables descriptivas y sectoriales

Se analizan las variables descriptivas propias del cliente y las relacionadas a su actividad económica, tales como los meses que lleva el cliente vinculado con la entidad financiera y su sector económico. Según lo planteado por Jiménez & Mencía (2009) existen sectores con mayores niveles de riesgo en comparación a otros cuyo comportamiento puede ser más estable, lo que se traduce en mayores o menores pérdidas crediticias esperadas. En

este grupo hay un total de 31 variables, siendo en su mayoría variables indicativas de sectores económicos (Ver Tabla 8-5).

3.3 Datos y frecuencia de observación de las variables

A continuación, se brinda información relacionada a los datos que se usan para el pronosticado de la pérdida crediticia esperada, haciendo énfasis en la cantidad de registros que se tiene y la frecuencia de las variables explicativas.

3.3.1 Datos

La constitución o liberación de pérdida crediticia esperada usualmente se explica por un grupo limitado de clientes que se evalúan una vez por semestre, realizando un análisis individual de sus estados financieros para determinar su capacidad de pago por medio de proyección de flujos de efectivo considerando variables de riesgo, macroeconómicas, sectoriales, transaccionales y financieras. Los clientes que se encuentran bajo este proceso, son el conjunto de unidades sobre las que se busca predecir la ECL con el fin de prever movimientos contables.

La información disponible al momento de realizar el presente trabajo corresponde al periodo comprendido entre junio de 2019 hasta diciembre de 2022, y se tienen en promedio 24 clientes por semestre, para un total de 194 datos, lo cual genera un reto para la modelación y ajuste, considerando la gran cantidad de variables que se tienen planteadas.

Se debe hacer claridad en que, de los 194 datos, se eliminaron 10 registros ya que presentaban errores en sus valores. Lo anterior debido a que los estados financieros se mandan en formato PDF a la entidad financiera, y un grupo de personas se encarga de pasarlos a formatos numéricos, por lo cual, al ser un proceso manual pueden existir errores asociados a la digitación o en la confusión de cuentas contables. Luego de eliminar esos registros la base quedó con 184 datos.

Por otro lado, en la base se tienen variables discretas ordinales, las cuales no se plantean como Dummys para evitar que se pierda el valor ordinal de las variables al trabajarlas como

cualitativas, por ello se dejan como variables tipo numéricas. Las variables cualitativas donde no importa el orden, sí se trabajan como Dummys.

3.3.2 Frecuencia de las variables

La información de la variable respuesta, la cual hace referencia a la pérdida crediticia esperada, es tomada de forma semestral, es decir, se toman los valores registrados en junio y diciembre. Esto se debe a que los clientes que se encuentran en el proceso de análisis individual de la ECL se deben revisar una vez por semestre, pero, dado que es un proceso detallado y manual, se reparten los clientes en los diferentes meses que componen cada semestre, por ello, al tomar sólo los valores semestrales, se garantiza que todos los clientes hayan sido analizados.

Con respecto a las variables independientes, las variables relacionadas al riesgo y las descriptivas, todas se encuentran en la misma periodicidad que la variable respuesta, es decir, a corte de junio y diciembre. Pero también se toma información rezagada, para identificar si en periodos anteriores, existieron movimientos que podrían afectar la pérdida crediticia esperada. En cuanto a la información financiera, en cada semestre se toman los estados financieros más recientes que se encontraban disponibles en la fecha de análisis, incorporando también información promedio y rezagada con estados financieros anteriores, buscando capturar los deterioros o mejoras que han tenido a lo largo del tiempo con respecto a su desempeño financiero. Por último, para las variables macroeconómicas que se encuentran publicadas en diferentes frecuencias, por ejemplo, frecuencia diaria, se toma la información del cierre de mes. Y se incorporan variables hasta con 6 meses de rezago para capturar los movimientos económicos y que pueden afectar la variable respuesta, como lo realizado por Taghiyeh et al. (2021).

4. Capítulo 4

Aplicación y comparación de las metodologías en la selección de variables

Dado el gran conjunto inicial de predictores potenciales identificados que se tiene, se aplica un proceso en tres etapas descrito en la Figura 4-1, primero se realiza una reducción de la dimensionalidad de la base inicial, esto con el fin de acotar los predictores y poder realizar un análisis más detallado posteriormente al conjunto de variables independientes. En un segundo paso se aplican métodos de selección de variables, los cuales, al tener un enfoque iterativo, son costosos computacionalmente y no se pueden realizar directamente sobre el conjunto de datos inicial. Finalmente, comparando los resultados de los diferentes métodos, se define el conjunto final de predictores. A continuación, una descripción del proceso de selección de variables:

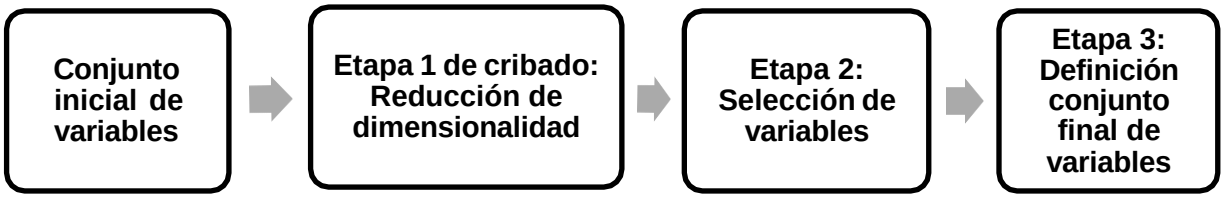


Figura 4-1: Metodología empleada para la reducción de dimensionalidad y selección de variables.

4.1 Transformación

La transformación de variables es un primer paso para la reducción de dimensionalidad y selección de variables, planteándose en los métodos de regularización Lasso y Ridge, para evitar que la escala de las variables independientes afecte la penalización de los coeficientes y con ello el resultado del modelo. Entre los predictores se encuentran variables cuantitativas, cualitativas ordinales y cualitativas nominales, por lo que no se puede plantear una sola transformación para todo el conjunto de variables independientes. Para las variables continuas, se plantea una transformación Z-score, también conocida como estandarización o normalización, la cual genera una nueva variable cuya media es igual a 0 y una varianza igual a 1 (Pedregosa et al., 2011) . A continuación, la ecuación de la transformación Z-score:

$$Z = \frac{x - \bar{x}}{S_x}. \tag{4-1}$$

En la anterior ecuación se tiene a Z que es el valor transformado o Z-score, x es el valor original de la variable, $\bar{x}$ es la media muestral de la variable independiente y S_x corresponde a la desviación estándar muestral del predictor (Pedregosa et al., 2011). Para las variables cualitativas ordinales se plantea una transformación min-max, la cual consiste en tomar cada valor de la variable independiente, restarle el valor mínimo de dicha variable y luego dividirlo entre la diferencia entre el valor máximo y el valor mínimo. A continuación, la ecuación de la estandarización por medio de min-max:

$$x_{\text{std}} = \frac{x - \min(x)}{\max(x) - \min(x)}, \tag{4-2}$$

donde x_{std} es el valor estandarizado de la variable x, $\min(x)$ es el valor mínimo en el conjunto de datos y $\max(x)$ es el valor máximo de dicho predictor. Por medio de esta transformación se respeta el orden de las variables, pero estando un rango de 0 y1.

Por último, las variables cualitativas nominales no se transforman dado que son variables que se trabajan como Dummys.

4.2 Reducción de dimensionalidad

Dada la gran cantidad de variables que se tienen en el análisis y un número limitado de observaciones, se utilizan diferentes métodos de selección de variables para disminuir la dimensionalidad de la base de datos. Para ello, se empieza eliminando las variables que no aportan al modelo por medio de diferentes métodos de selección de variables, buscando reducir significativamente el número de predictores de la base de datos. En una etapa posterior, se realiza nuevamente la selección de variables, pero teniendo como punto de partida una base de predictores más reducida y sin tanto ruido.

Es importante señalar que para los procedimientos de reducción de dimensionalidad que brindan como resultado un valor de importancia para todos los predictores, como es el caso de Random Forest y Extra Tree, se debe establecer un umbral de corte (Loh, 2011). También en la regresión Ridge se usa un umbral dado que su penalización L2 hace que los coeficientes sean pequeños, pero no iguales a cero (Pereira et al., 2016). Para la regresión Lasso no hay necesidad de establecer un umbral dado que las variables con un coeficiente igual a 0 no deben ser consideradas.

El umbral que se define consiste en comparar el coeficiente de la importancia de los predictores en los modelos de Bosques Aleatorios, Extra Trees y Ridge con el coeficiente obtenido al tomar una variable aleatoria normal con media 0 y varianza 1, como variable independiente en cada modelo, con el fin de determinar si el predictor debe ser considerado o no, en el conjunto de datos. La razón de usar una variable aleatoria se debe a que proporciona una referencia neutral que no está influenciada por el modelo en sí, y al compararla contra los predictores, se evalúa si el coeficiente de la importancia tiene un valor mayor diferente al obtenido a la v.a, dejando sólo las variables independientes que tienen una mayor relevancia en cada modelo (Stoppiglia et al., 2003).

Sobre el conjunto de datos, se plantea una muestra de entrenamiento del 75% de los registros, esto con el fin de que se tenga una cantidad suficiente de datos para ajustar los modelos de manera efectiva y dejar un 25% para el conjunto de prueba, que permita evaluar el rendimiento de los modelos. La idea de particionar el conjunto de datos es para

evitar el sobreajuste, para evaluar la capacidad de generalización del modelo y medir su desempeño en datos no observados.

El conjunto de datos inicial es de 437 variables, siendo las variables que pertenecen a los grupos de riesgo, financiero, transaccional, macroeconómico y descriptivas. Importante señalar que la variable aleatoria normal, sólo se usa para compararla contra las demás variables independientes.

4.2.1 Selección en los bosques aleatorios

Los bosques aleatorios, definidos en el Capítulo 2, brindan la importancia relativa de cada variable independiente para predecir la variable respuesta, donde a mayor importancia, mayor relevancia para el modelo. Dado que la importancia en los bosques aleatorios genera un valor para todos los predictores del conjunto de datos, se compara la importancia de cada predictor con una variable aleatoria normal para determinar un umbral de selección, y sólo aquellos predictores con un valor de importancia superior, serán seleccionados por este método. Es importante señalar que ninguna variable con una importancia menor a una variable aleatoria normal, debe ser considerada como predictor.

Luego de realizar lo anterior, se observa que de las 437 variables sólo 163 tienen una importancia superior a la obtenida por una variable aleatoria normal.

Tabla 4-1: Variables más importantes obtenidas por Bosques Aleatorios

Orden	Variable	Valor
1	RG6	0.269
2	RG21	0.083
3	RG1	0.043
4	RG16	0.038
5	RG25	0.024
6	RG100	0.022
7	FIN120	0.019
8	RG4	0.018
9	RG23	0.013
10	FIN136	0.011

En la tabla anterior, se observan las 10 variables que presentan una mayor importancia por medio de los bosques aleatorios, siendo en su mayoría variables de riesgo y financieras.

4.2.2 Extremely randomized trees

Por medio del algoritmo ExtraTreesRegressor, el cual se explicó en el capítulo 2, se obtuvo la importancia de cada variable calculada como la disminución promedio de la impureza medida mediante el índice de Gini. Con este modelo se obtienen 413 variables con una importancia mayor a 0 y sólo 133 variables que aportan más que una variable aleatoria normal.

Tabla 4-2: Variables con mayor importancia ExtraTreesRegressor

Orden	Variable	Valor
1	RG6	0.369
2	RG100	0.041
3	RG1	0.033
4	RG2	0.020
5	RG102	0.016
6	FIN77	0.015
7	RG21	0.013
8	FIN93	0.013
9	DyS2	0.012
10	FIN76	0.011

En la tabla anterior se listan las variables que resultaron más importantes en este método, y a diferencia del modelo anterior, en este último todas las variables con un mayor nivel de importancia se encuentran en el grupo de riesgo.

4.2.3 Lasso

Por medio de Lasso y su algoritmo de penalización L1 se reduce significativamente el conjunto de variables. Se logró pasar de 437 variables a 38, siendo estas 38 variables las que obtuvieron un coeficiente mayor a 0. Es importante señalar que, para la implementación de Lasso, se realiza la transformación de acuerdo al tipo de variable que sea el predictor, de este modo se evitan algunas variables que por su escala de medida resultan con coeficientes numéricamente muy grandes y con una mayor relevancia en comparación con otras variables. Adicionalmente, se realiza una validación cruzada con 5 particiones aleatorias sobre la regresión con el fin de generar un resultado más robusto en el conjunto de entrenamiento.

Tabla 4-3: Variables con mayor coeficiente modelo Lasso

Orden	Variable	Coeficiente
1	MEC84	0.066
2	FIN76	-0.063
3	MEC85	0.054
4	FIN39	-0.036
5	DyS19	0.034
6	FIN78	0.034
7	RG13	0.033
8	FIN107	-0.029
9	RG84	-0.025
10	DyS12	-0.022

En la tabla anterior se observa que en las primeras 10 variables con mayor coeficiente son en su mayoría aquellas relacionadas a indicadores macroeconómicos y variables financieras, seguidas de las de riesgos y, por último, las transaccionales. Es importante señalar que se ordena la tabla anterior de acuerdo al valor del coeficiente en términos absolutos.

4.2.4 Ridge

La regresión Ridge, definida en el Capítulo 2, también es usada para la selección de variables mediante su penalizador L2. Para esta metodología, tal como se hizo en Lasso, las variables se transformaron considerando si la variable independiente era continua u ordinal, proponiendo una validación cruzada con 5 particiones aleatorias en el conjunto de entrenamiento.

Por este método, se obtuvieron 415 variables con coeficientes diferentes a 0, pero sólo 32 de estas tuvieron un coeficiente mayor al de una variable aleatoria normal. A continuación se adjunta los predictores con mayor coeficiente.

Tabla 4-4: Variables con mayor coeficiente modelo Ridge

Orden	Variable	Coeficiente
1	MEC84	0.043
2	RG15	0.023
3	FIN76	-0.022
4	RG13	0.017
5	RG14	0.017
6	MEC85	0.016
7	FIN75	-0.016
8	FIN95	0.016
9	RG3	0.015
10	FIN123	0.015

En la tabla anterior, se encuentran las 10 variables con mayor coeficiente de acuerdo a la regresión Ridge, teniendo variables macroeconómicas, de riesgo, financieras y transaccionales. Cabe señalar, que el ordenamiento de los coeficientes se hizo considerando su valor absoluto.

4.2.5 Variables usadas en el proceso de la pérdida crediticia esperada

Es importante señalar que en el proceso para hallar la pérdida crediticia esperada en el grupo de clientes que tienen una mayor exposición con la entidad financiera, se involucran algunas variables que brindan información acerca de la viabilidad financiera que tiene cada cliente, el riesgo percibido y su transaccionalidad. Por lo tanto, es importante tener en cuenta dentro de la reducción de dimensionalidad, aquellos predictores que la entidad financiera usa siempre dentro de sus análisis de flujos de caja y percepción del riesgo.

4.2.6 Resultados

Considerando los cuatro métodos de selección anteriores, se toman sólo aquellas variables que fueron seleccionadas en mínimo tres de los cuatro métodos de selección propuestos, y aquellas que se involucran dentro del proceso de la entidad financiera para hallar la pérdida crediticia esperada, de este modo se logra reducir el conjunto inicial de 437 variables a un conjunto de 58 variables predictoras (Ver Tabla 8-6) A continuación, se

presentan las variables que fueron seleccionadas en los 4 métodos usados para reducir la dimensionalidad, siendo Bosques Aleatorios, Extra Tree, Ridge y Lasso.

Tabla 4-5: Variables seleccionadas en los 4 métodos de reducción de dimensionalidad, considerando el conjunto de 58 predictores.

Variable	Frecuencia	Ridge	Forest	Extratree	Lasso
RG10	4	X	X	X	X
FIN109	4	X	X	X	X
FIN93	4	X	X	X	X
FIN91	4	X	X	X	X
RG23	4	X	X	X	X
FIN4	4	X	X	X	X
FIN64	4	X	X	X	X

En la tabla superior se encuentran las variables que fueron seleccionadas en los 4 métodos propuestos para la reducción de dimensionalidad, donde se observa que en su mayoría son variables de riesgo y financieras.

El procedimiento desarrollado en esta sección puede considerarse como una primera etapa en la reducción del espacio de las variables explicativas. Sin embargo, todavía se tiene una dimensionalidad alta por lo que se aplicará una segunda fase de selección sobre el subconjunto de las 58 variables resultantes en esta etapa preliminar de reducción. Veremos esto en la próxima sección.

4.3 Selección de variables

Luego de reducir considerablemente el subconjunto de variables preseleccionadas, se procede con una segunda etapa de selección de variables mediante los métodos secuenciales explicados en el Capítulo 2.

4.3.1 Selección secuencial hacia atrás usando AIC

La selección secuencial hacia atrás, también conocida como backward feature selection es un método de selección de variables que puede realizarse usando el criterio de información de Akaike (AIC). Con este método se empieza con un modelo de mínimos cuadrados ordinarios (OLS) que incluye todas las variables del subconjunto de predictores

y luego iterativamente se elimina una característica a la vez basándose en el criterio AIC, el cual, es un criterio estadístico utilizado para seleccionar modelos que logren un buen equilibrio entre el ajuste del modelo y su complejidad. Esto se debe a que el criterio AIC se define como la suma de dos términos, el primero es el estadístico de -log verosimilitud y el segundo, una medida de penalización dado el número de parámetros, el objetivo del AIC es encontrar el modelo que minimice dicha suma (Akaike, 1974).

Por medio de este método de selección, el modelo OLS resultante contiene sólo 20 variables de las 58 que fueron consideradas, brindando un R^2_{adj} de 0.904. A continuación, se observan las variables que fueron seleccionadas:

Tabla 4-6: Las 20 variables que fueron seleccionadas con la selección secuencial hacia atrás y el criterio AIC

Orden	Variable	Coeficiente
1	RG6	0.094
2	RG19	8.150E-12
3	FIN91	-0.016
4	FIN140	-0.006
5	RG93	-0.002
6	FIN93	-2.162E-04
7	FIN84	0.013
8	TRX19	-0.002
9	FIN57	-0.001
10	RG17	1.746E-11
11	FIN125	-0.004
12	FIN78	-0.010
13	FIN109	-0.134
14	FIN76	0.004
15	FIN134	-2.910E-13
16	FIN46	-0.002
17	RG18	9.071E-12
18	RG98	0.091
19	RG21	0.027
20	RG13	-0.389

En la tabla anterior se muestran las 20 variables que fueron seleccionadas con el método de regresión paso a paso hacia atrás por medio del criterio de información AIC, evidenciando que gran parte de las variables corresponden al grupo de riesgo y al financiero.

4.3.2 Eliminación recursiva de variables: Bosques de regresión

Considerando que el método de regresión paso a paso parte de un modelo de mínimos cuadrados ordinarios (OLS), como segundo método de selección se realiza una eliminación recursiva de variables (explicada en el Capítulo 2) con validación cruzada y un modelo de Bosques Aleatorios, el cual permite capturar relaciones no lineales entre las variables explicativas con la variable respuesta (Breiman, 2001). A continuación, se muestra la importancia de las variables:

Tabla 4-7: Variables seleccionadas con mayor importancia por medio de RandomForest

Orden	Variable	Importancia	Importancia acumulada
1	RG6	0.281	0.281
2	RG21	0.091	0.372
3	RG16	0.053	0.424
4	RG1	0.049	0.474
5	RG100	0.047	0.520
6	RG10	0.046	0.566
7	RG25	0.030	0.597
8	FIN120	0.024	0.620
9	RG23	0.021	0.641
10	TRX14	0.019	0.660
11	FIN76	0.017	0.677
12	FIN92	0.016	0.692
13	FIN70	0.014	0.707
14	FIN93	0.014	0.721
15	FIN57	0.014	0.735
16	FIN39	0.011	0.746
17	FIN105	0.011	0.757
18	FIN81	0.011	0.768
19	FIN46	0.010	0.778
20	FIN140	0.010	0.788
21	FIN68	0.010	0.798

En la tabla anterior se observan las 21 variables explican el 80% de la importancia total del modelo. Es importante señalar que debido a la normalización que usa el algoritmo, la importancia total es la sumatoria de la importancia de cada variable independiente, siendo igual a 1. Por ello, en busca de obtener un modelo parsimonioso, se decide sólo tomar las variables que explican el 80% de la importancia total, debido a que como se explicó

previamente, este modelo brinda valores de importancia para todas las variables, por lo que se debe establecer un umbral de corte.

Por medio de este método se obtienen 21 variables de las 58 variables preseleccionadas en el procedimiento descrito en la sección de reducción de dimensionalidad.

4.3.3 Selección final de las variables

Luego de realizar la selección de variables, se dejan aquellas que fueron elegidas en alguno de los dos métodos anteriores, para un total de 34 variables, las cuales son buenas predictoras para alguno de los modelos. Se obtienen por medio del primer método de selección variables por medio de un modelo OLS con criterio AIC. Por otro lado, se tienen las variables que explican un 80% de la importancia total del modelo Bosques Aleatorios, modelo que captura relaciones no lineales entre la variable predictora y la variable respuesta.

4.4 Análisis de las variables independientes y variable respuesta.

Considerando que se inició con un conjunto de 437 predictores, pasando a 58 luego de la reducción de dimensionalidad, y llegando a 34 variables en el conjunto final de predictores, se procede a realizar un análisis de las variables independientes y de la variable respuesta. De las 34 variables que se tienen, 27 son de tipo continuo, 5 cualitativas ordinales y 2 cualitativas nominales.

4.4.1 Análisis variables continuas: multicolinealidad

Luego de tener las 34 variables seleccionadas, y teniendo 27 de ellas continuas para el modelo predictivo de la pérdida crediticia esperada, se hacen los análisis de multicolinealidad para este grupo de variables. Uno de los métodos empleados es el análisis de componentes principales, éste agrupa variables con alta correlación en componentes principales, los cuales también pueden usarse para reducir la dimensionalidad del espacio de variables predictoras.

De acuerdo al análisis de componentes principales realizado sobre las 27 variables, se observa que el número de componentes que serán necesarios para explicar la mayor variabilidad lo proporciona el mismo número de variables explicativas, y la varianza explicada en los primeros componentes es muy pequeña. Adicionalmente, si observamos de forma gráfica su varianza explicada, es difícil determinar un punto de inflexión, concluyendo que no hay dependencias lineales muy fuertes dentro de las variables analizadas.

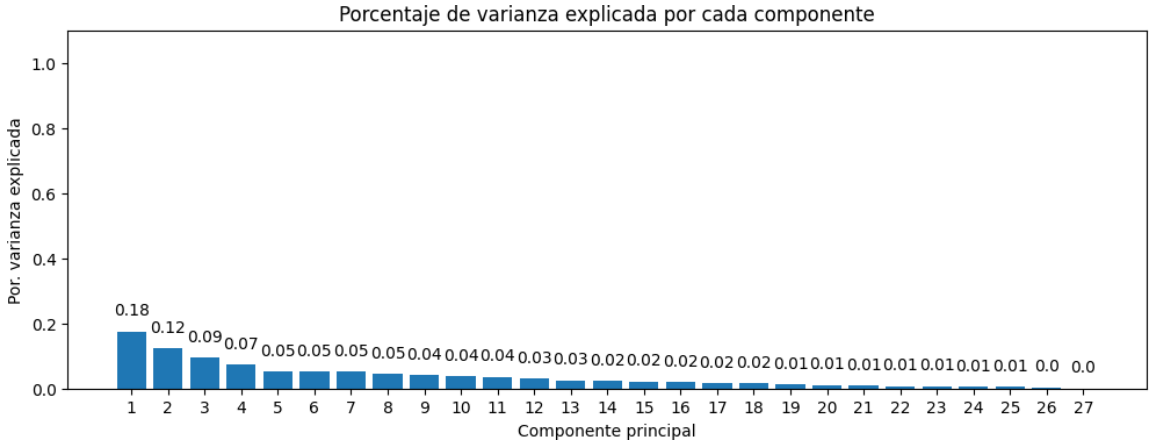


Figura 4-2: Porcentaje de varianza explicada por componentes principales

Otra técnica de diagnóstico de multicolinealidad aplicada es mediante el factor de inflación de la varianza o VIF, por sus siglas en inglés. El VIF mide la severidad de multicolinealidad en una regresión lineal múltiple, refiriéndose a la proporción de cuánto se ve afectada la varianza de un coeficiente de regresión debido a la multicolinealidad Cheng et al. (2022). La fórmula de esta medida es:

$$VIF_j = \frac{1}{1 - {R^2}_j} \tag{4-3}$$

${R^2}_j$ representa el coeficiente de determinación obtenido al ajustar un modelo de regresión de una variable $x_j$en función de las demás variables independientes. Un VIF_j alto indica un alta multicolinealidad, lo que implica que la variable predictora está altamente correlacionada con otras variables independientes. Según Cheng et al. (2022) un VIF mayor a 10 indica multicolinealidad.

Tabla 4-8: Factores de inflación de varianza (VIF) obtenidos en el grupo de variables independientes

Orden	Variable	VIF
1	FIN84	33.672
2	FIN78	31.366

En la tabla anterior, se observa que sólo existen dos variables con un VIF mayor a 10. Siendo FIN78 una variable que brinda la relación entre pasivos y activos, y FIN84 es la misma variable, pero tomando un promedio de 24 meses, es decir, indica el promedio de los pasivos y activos de los últimos cortes de información financiera disponible. Dado que ambas variables parten del mismo indicador financiero, sólo se deja FIN78 que es la relación entre pasivos y activos en la fecha de análisis.

Por medio de los análisis de multicolinealidad se observa que no hay una alta asociación lineal entre las variables independientes cuantitativas, y por medio del valor de los VIF se elimina FIN84 del conjunto de predictores, pasando de 34 a 33 variables independientes, y de 27 variables continuas a 26.

4.4.2 Correlaciones entre las variables predictoras y la variable respuesta.

Considerando las 26 variables independientes continuas, se procede a analizar la correlación Pearson entre cada variable independiente cuantitativa con la variable respuesta, con el objetivo de analizar la fortaleza de las relaciones lineales entre las variables independientes y la pérdida crediticia esperada.

En la Tabla 4-9 se exhiben las diez variables con mayor correlación con la pérdida crediticia esperada, estando ordenadas en términos absolutos, observando que no hay valores cercanos a 1 en ningún predictor.

Tabla 4-9: Top 10 variables independientes continuas con mayor correlación con la variable respuesta

Variables	Correlación
FIN81	0.352
FIN68	-0.342
FIN78	0.332
FIN57	-0.285
TRX14	-0.271
FIN70	-0.269
RG16	0.264
RG19	0.245
RG18	0.218
FIN105	-0.167

En la tabla anterior se puede observar que las variables continuas no presentan una alta asociación lineal con la variable respuesta, siendo todas menores a 0.5 en términos absolutos. No obstante, si se observa los signos de los coeficientes, se determina que las relaciones presentan el signo correcto, es decir, la primera variable que se relaciona al nivel de endeudamiento del cliente, brinda una correlación positiva, donde a mayor endeudamiento una mayor pérdida crediticia esperada. La segunda variable, asociada al rendimiento financiero, presenta un signo negativo, lo que indica que un mejor rendimiento en términos económicos, estará ligado a una disminución de la variable respuesta.

Considerando que la correlación de Pearson se analizó para las variables continuas independientes, se procede a hallar el coeficiente de correlación de Kendall's Tau-b para las variables ordinales, dado que este indicador permite medir la correlación cuando una variable es ordinal y la otra continua (Khamis, 2008). La medida de Kendall's Tau-b se basa en la comparación de los rangos de los valores de las variables en lugar de utilizar los valores exactos, basándose en la proporción de concordancias y discordancias en los pares de observaciones, la concordancia se refiere a los pares de observaciones que tienen el mismo orden relativo en ambas variables, mientras que la discordancia se refiere a los pares de observaciones que tienen un orden relativo diferente. Esta medida brinda valores entre -1 a 1, donde 1 indica una correlación perfectamente positiva, -1 indica una correlación perfectamente negativa y 0 indica la ausencia de correlación (Kendall, 1948).

A continuación, se brinda el coeficiente de correlación de Kendall's Tau-b para las variables ordinales con la variable respuesta.

Tabla 4-10:Correlación de Kendall's Tau-b de las variables independientes ordinales con la variable respuesta

Variables	Correlación
RG6	0.626
RG23	0.511
RG21	0.494
RG25	0.486
RG1	0.447

En la tabla anterior se observa que las variables ordinales son asociadas al riesgo del cliente, como la calificación interna y externa, las cuales tienen una correlación positiva alta con la pérdida crediticia esperada. Esto es coherente dado que un cliente con un mayor nivel de deterioro percibido por la entidad financiera debería tener una mayor pérdida crediticia esperada.

Ahora, se analiza la correlación de las variables nominales, para ello se plantea el coeficiente de correlación punto-biserial, que permite evaluar la relación entre una variable binaria y una variable continua, donde un valor cercano a -1 indica una relación negativa fuerte entre la variable nominal y la variable continua, mientras que un valor cercano a 1 indica una relación positiva fuerte (Khamis, 2008). A continuación, la correlación punto-biserial de las variables nominales con la variable respuesta:

Tabla 4-11: Correlación de punto-biserial de las variables independientes nominales con la variable respuesta

Variables	Correlación
RG100	0.62
RG98	0.46

En la tabla anterior se observa que las dos variables nominales presentan una correlación positiva con la variable respuesta. Ambas variables indican un aumento del riesgo del cliente, brindando información acerca del incumplimiento que ha tenido en el pago de sus

obligaciones, por ende, un cliente que ha presentado default genera un mayor valor de la variable respuesta.

4.4.3 Conclusiones del análisis de variables

Luego de realizar la selección de variables y reducir considerablemente el número de predictores, se evidencia que no hay una fuerte multicolinealidad entre las variables explicativas cuantitativas. Además, cuando se analiza la correlación de Pearson entre las variables independientes cuantitativas con la variable respuesta se observa que los valores no son cercanos a 1 en la mayoría de casos, por lo que no se encuentran fuertes dependencias lineales. Por otro lado, al analizar la correlación por medio del Kendall's Tau-b para variables ordinales y la correlación punto-biserial para las nominales, se observan coeficientes mayores en comparación con las variables cuantitativas.

5. Capítulo 5

Evaluación modelos reducidos y selección del modelo final

En esta sección se realiza la evaluación de los modelos que fueron explicados en el Capítulo 2 del presente trabajo, tomando el conjunto de variables que fue seleccionado en el Capítulo anterior, donde se evidenció que no se observaba una dependencia lineal clara entre los predictores cuantitativos y la pérdida crediticia esperada.

Además, se consideran y comparan modelos paramétricos y no paramétricos tales como: Ridge, Lasso y Regresión Lineal Múltiple Ordinario, propuestos por Taghiyeh et al. (2021) y modelos de machine learning como Xgboost y RandomForest presentados en los estudios de Heo & Yang (2014), Chen, 2011), Yeh et al. (2014) y otros autores, con el fin de predecir la pérdida crediticia esperada del grupo de clientes que cuentan con una alta exposición y un nivel de riesgo significativo. El objetivo es establecer un modelo predictivo basado en el mejor subconjunto de predictores entre el grupo de variables propuesto en el Capítulo anterior.

Considerando las 33 variables que fueron seleccionadas se particiona la base en un conjunto de entrenamiento equivalente al 75% de los datos y un conjunto de prueba de 25%. Con ello, se busca que los modelos no caigan en sobreajuste y tengamos una muestra que el algoritmo desconozca para garantizar que las predicciones sean

consistentes. Es importante señalar que la partición de los datos se da de forma aleatoria, dado que las variables rezagadas están como predictores y no sujetas al tiempo, es decir, si se tiene una variable en un periodo anterior, dicha variable estará rezagada como una variable independiente en el conjunto de datos, tal como lo realizaron Taghiyeh et al. (2021) en su estudio. Adicionalmente, la pérdida crediticia esperada del grupo de clientes con mayor exposición y riesgo percibido parte de un análisis de variables independientes que no están sujetas al tiempo, por ende, se establece una partición aleatoria. La precisión en la calidad del ajuste y de la predicción se mide mediante el MSE, MAE y el coeficiente de determinación.

5.1 Regresión Lineal Ordinaria

Se plantea como primer modelo una regresión lineal múltiple, con el fin de observar si por medio de este modelo se obtiene un buen ajuste dada su fácil interpretabilidad. Al plantear el modelo se obtiene un $R^2_{aj} = 0.74$ con 14 variables independientes con un p-valor menor a 0.05. A continuación, se observa la tabla con los resultados:

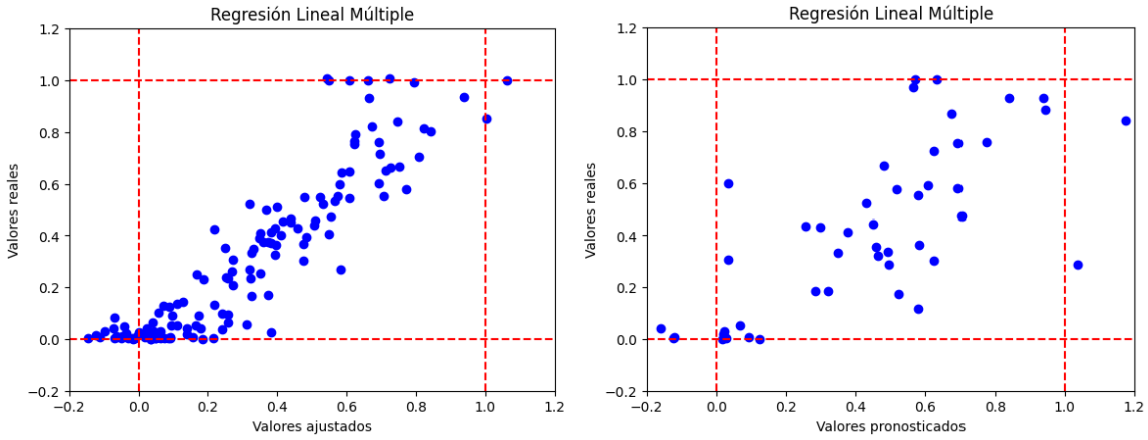
Tabla 5-1: Variables predictoras con un p-valor menor a 0.05 obtenidas por una Regresión Lineal Múltiple

Variables	Coeficiente	P-valor
FIN91	-0.012	0.000
RG93	-0.002	0.000
FIN140	-0.005	0.001
FIN134	-3.912E-13	0.002
FIN125	-0.006	0.002
RG19	6.612E-12	0.002
FIN109	-0.170	0.004
RG6	0.072	0.005
FIN76	0.003	0.007
FIN46	-0.002	0.010
RG18	9.395E-12	0.017
TRX19	-0.002	0.020
RG17	1.335E-11	0.040
RG98	0.084	0.045

En la tabla anterior podemos observar las 14 variables que tienen un p-valor menor a 0.05 considerando el conjunto de los 33 predictores, donde se encuentra que gran parte de las

variables independientes pertenecen al grupo financiero y de riesgo, contando con un solo predictor del grupo transaccional. Por otro lado, al observar el signo del coeficiente estimado de cada variable independiente, se denota que los predictores pertenecientes al grupo de riesgo, tienden en su gran mayoría, a aumentar la variable respuesta, mientras los relacionados al grupo financiero, tienden a disminuirla.

Al realizar la predicción por medio de la Regresión Lineal Múltiple en el conjunto de datos de prueba, se obtiene un coeficiente de determinación predicho de 0.4204, lo que indica que no existe una buena calidad de predicción en el modelo de regresión, Ahora, se anexa la gráfica de las respuestas observadas en comparación de las respuestas estimadas por medio de la regresión:



(a) (b)

Figura 5-1: (a) Valores reales vs Valores ajustados (b) Valores reales vs Valores pronosticados por medio de la Regresión Lineal Múltiple

En la gráfica anterior se encuentran dos figuras, la (a) corresponde a los valores observados vs ajustados en el conjunto de entrenamiento, y la (b) a los valores observados vs pronosticados en el conjunto de prueba. En la (a) podemos observar que hay un gran número de valores ajustados negativos y otros superiores a 1, pero en comparación a la figura (b) hay una nube de puntos menos dispersa. Por otro lado, en (b) se observa que la nube de puntos está muy distante entre los valores reales a los que se obtienen en la predicción de la regresión, por ello, el R^2 de predicción muestra la poca capacidad del

modelo para pronosticar los datos fuera de la muestra de ajuste. A continuación, se anexan otras medidas obtenidas en el conjunto de entrenamiento y en el de prueba:

Tabla 5-2: Resultados de la Regresión Lineal Múltiple en el conjunto de datos de entrenamiento y de prueba

Modelo	Datos	MSE	MAE	R^2
Regresión Lineal Múltiple	Entrenamiento	0.0183	0.0994	0.8056

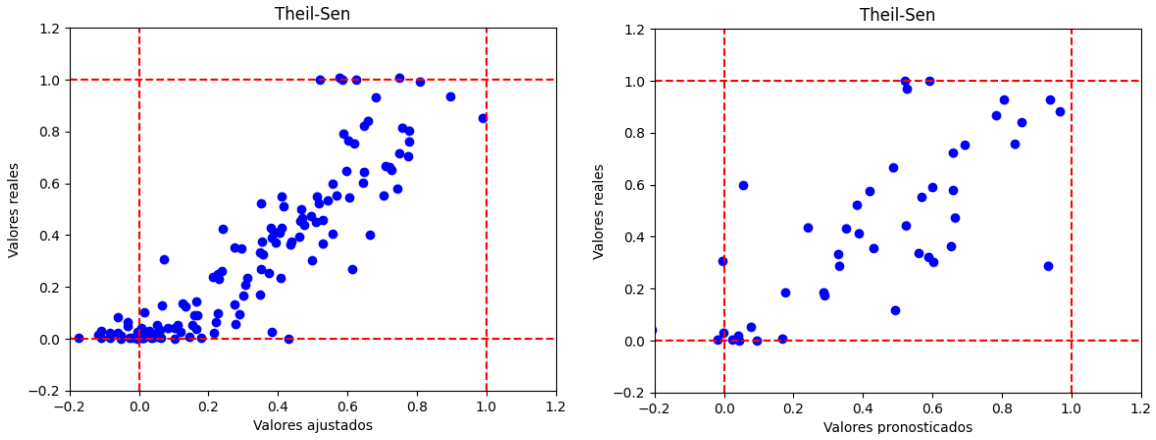
Modelo	Datos	MSE	MAE	R^2
Regresión Lineal Múltiple	Prueba	0.0577	0.1803	0.4204

En la tabla anterior se encuentran los resultados de calidad de ajuste en los datos de entrenamiento y de predicción en los datos de prueba, teniendo como medidas el MSE, el cual representa la magnitud promedio de los errores al cuadrado, el MAE siendo el promedio de los errores absolutos de las predicciones de la regresión y los valores reales, y el coeficiente de determinación (R^2) que indica la proporción de varianza de la variable respuesta explicada por el modelo.

En el conjunto de entrenamiento se observa un coeficiente de determinación de 0.8, un MSE de 0.01 y un MAE menor a 0.1, los cuales están bastante alejados de los resultados en el conjunto de prueba, dado que, en este último, se observa un MAE de 0.18, un MSE de 0.05 y un R^2 de 0.42. Lo que permite concluir que la regresión lineal múltiple no es el modelo indicado para la predicción de la pérdida crediticia esperada, dado los errores de predicción en el conjunto de prueba. Esta situación también podría estar siendo influenciada por los datos atípicos, generando que la regresión no se ajuste correctamente, por lo tanto, en la siguiente sección se plantean métodos robustos.

5.2 Regresión lineal robusta

Considerando que las variables predictoras continuas cuentan con distribuciones muestrales de colas pesadas, se plantean dos métodos robustos buscando mejorar el ajuste del modelo. El primero es el método Theil-Sen el cual es un estimador robusto de regresión que se usa para conjuntos de datos que contienen datos atípicos y es resistente a desviaciones altas (Theil, 1949). Por medio de este método se obtiene lo siguiente:



(a) (b)

Figura 5-2: (a) Valores reales vs Valores ajustados (b) Valores reales vs Valores pronosticados por medio del modelo Theil-Sen

En la figura anterior se observa que en el gráfico (a) los valores ajustados en el conjunto de entrenamiento están por fuera del rango esperado de la variable respuesta, situación análoga en los valores pronosticados en el conjunto de prueba (b). Adicionalmente las nubes de puntos en ambos conjuntos no mejoran en comparación a la regresión ordinaria, a continuación, las medidas de ajuste y de predicción:

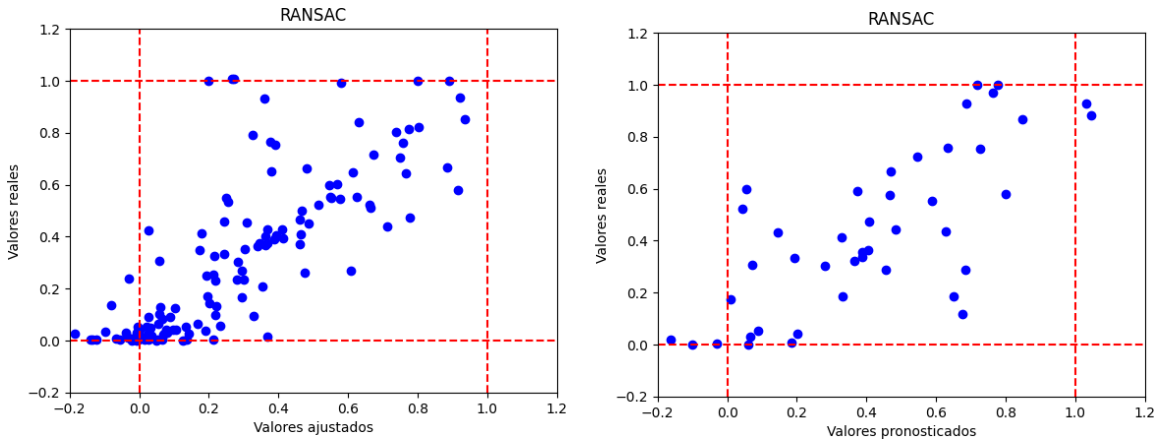
Tabla 5-3:Resultados del modelo Theil-Sen en el conjunto de datos de entrenamiento y de prueba

Modelo	Datos	MSE	MAE	R^2
Theil-Sen	Entrenamiento	0.0296	0.1088	0.6858

Modelo	Datos	MSE	MAE	R^2
Theil-Sen	Prueba	0.0562	0.1796	0.4359

En la tabla anterior se observa que en el conjunto de entrenamiento se obtienen mayores errores y un menor R^2 en comparación a la regresión lineal múltiple, pasando de un coeficiente de determinación de 0.8056 en la regresión ordinaria a 0.6858 por medio del método Theil-Sen. Por otro lado, en el conjunto de prueba se obtienen medidas muy cercanas con respecto a la regresión lineal múltiple.

El otro método que se plantea es el Random Sample Consensus o RANSAC por sus siglas en inglés, éste es un método robusto de regresión que se utiliza para ajustar un modelo a un conjunto de datos que contiene valores atípicos y utiliza los puntos que se consideran valores representativos o no atípicos (Fischler & Bolles, 1981). A continuación, se exhiben los resultados de las medidas obtenidas y el gráfico de los valores reales en comparación a los obtenidos por medio de la función RANSACRegressor de la librería de scikit learn (Pedregosa et al., 2011).



(a) (b)

Figura 5-3: (a) Valores reales vs Valores ajustados (b) Valores reales vs Valores pronosticados por medio del método de ajuste RANSAC

En la figura anterior se observa que tanto el gráfico (a) como en el (b) hay valores ajustados y pronosticados, respectivamente, que se encuentran por fuera del rango de la variable respuesta, situación que también se presentó en la regresión ordinaria y en el estimador robusto de regresión RANSAC. A continuación, las medidas MSE, MAE y coeficiente de determinación en el conjunto de datos de prueba y entrenamiento.

Tabla 5-4: Resultados de RANSAC en el conjunto de datos de entrenamiento y de prueba

Modelo	Datos	MSE	MAE	R^2
RANSAC	Entrenamiento	0.0520	0.1384	0.4472

Modelo	Datos	MSE	MAE	R^2
RANSAC	Prueba	0.0643	0.1906	0.3546

Por medio de RANSAC se obtienen menores valores en el coeficiente de determinación y mayores errores tanto en el conjunto de datos de prueba como en el conjunto de datos de entrenamiento. Es por ello, que considerando los métodos robustos de regresión RANSAC y Theil-Sen no se obtiene una mejora en comparación a la regresión ordinaria. Para más detalles de los métodos robustos Theil-Sen y RANSAC se puede profundizar en los estudios de Theil (1949) y Fischler & Bolles (1981).

Por último, es importante señalar que al tener un R^2 de 0.8056 en el conjunto de entrenamiento en la regresión lineal ordinaria, comparado con un R^2 de 0.6858 y 0.4472 en Theil-Sen y RANSAC, respectivamente, se nota la presencia de observaciones atípicas que son bastante influyentes en el ajuste del modelo y que sesgan la apreciación de su desempeño en la mayor parte de los datos. Los métodos robustos reducen la influencia sobre el ajuste de esas pocas observaciones atípicas y permiten una mejor valoración de la calidad del ajuste sobre la mayor parte de las observaciones. Por ello, cuando los métodos de estimación robusta disminuyen el efecto que tienen esas observaciones alejadas, se denota que el ajuste que brinda la relación lineal ordinaria no es tan bueno, dada la influencia de esos datos atípicos.

5.3 Regresión Ridge

Considerando la definición de la regresión Ridge brindada en el Capítulo 2, se procede a ajustar este modelo usando una validación cruzada de 5 particiones aleatorias y una transformación de variables según lo explicado en la Sección 4.1.

En la Tabla 5-5 se observan las 10 variables que tuvieron un mayor coeficiente en términos absolutos por medio del modelo Ridge, donde los predictores que tienen una mayor relevancia dentro del modelo pertenecen a los grupos de riesgo y financiero, situación similar a la regresión lineal múltiple.

Tabla 5-5: Top 10 variables con mayor coeficiente absoluto obtenido por el modelo Ridge

Orden	Variable	Coeficiente	Coeficiente absoluto
1	FIN76	0.070	0.070
2	RG13	0.064	0.064
3	RG17	-0.059	0.059
4	FIN39	-0.059	0.059
5	RG19	-0.057	0.057
6	RG93	0.052	0.052
7	FIN46	-0.050	0.050
8	RG10	0.045	0.045
9	RG18	-0.044	0.044
10	FIN57	-0.044	0.044

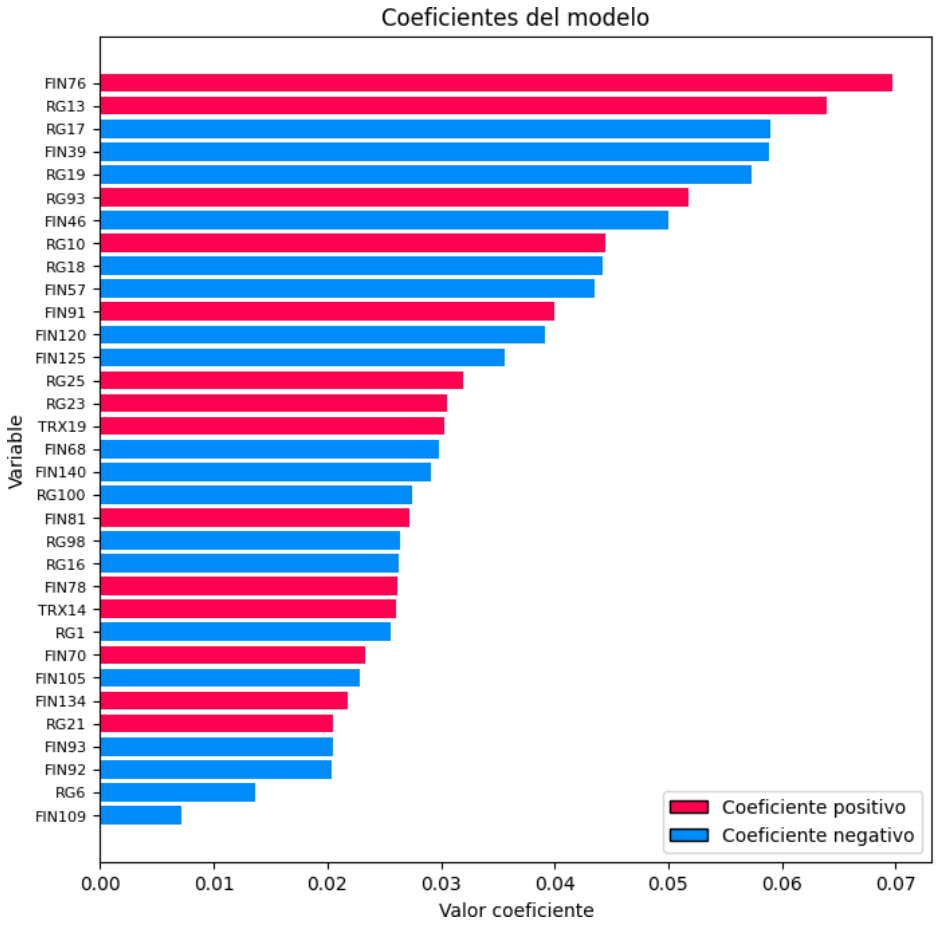
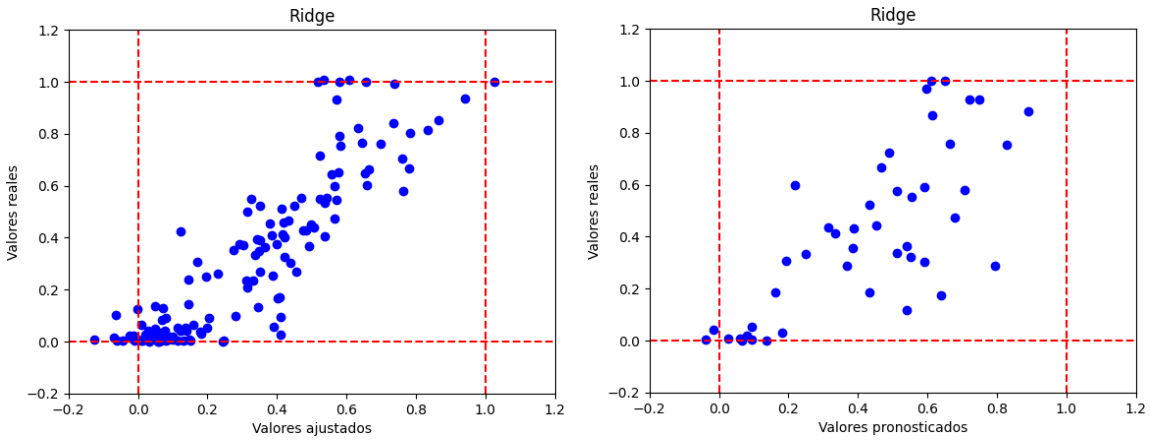


Figura 5-4: Coeficientes de Ridge ordenados por medio de su valor absoluto

En la figura anterior se puede observar que el número de variables que aportan positivamente a la variable respuesta es similar al número de predictores con signo negativo, teniendo 14 con un coeficientes positivos y 19 negativos. A continuación, se procede a realizar la predicción en el conjunto de datos de prueba, para obtener las métricas de desempeño y graficar tanto los valores de ajuste como de predicción.



(a) (b)

Figura 5-5: (a) Valores reales vs Valores ajustados (b) Valores reales vs Valores pronosticados por medio del modelo Ridge

En la figura anterior se puede evidenciar que se tienen valores negativos y mayores a 1 tanto en los valores ajustados en el conjunto de entrenamiento como en los valores pronosticados en el conjunto de prueba, situación que también se presentó en la regresión ordinaria y los métodos de estimación robustos. A continuación, las medidas de desempeño.

Tabla 5-6: Resultados del modelo Ridge en el conjunto de entrenamiento y de prueba

Modelo	Datos	MSE	MAE	R^2
Ridge	Entrenamiento	0.0213	0.1050	0.7735

Modelo	Datos	MSE	MAE	R^2
Ridge	Prueba	0.0461	0.1642	0.5372

Al observar las medidas de desempeño se denota que en el conjunto de entrenamiento se obtienen valores similares a los de la regresión ordinaria, tanto en el coeficiente de determinación como en los errores. Sin embargo, se obtienen mejores medidas en el conjunto de prueba, teniendo menores errores y un coeficiente de determinación mayor en la predicción del modelo, explicando de una mejor forma la proporción de variabilidad en comparación a la regresión lineal múltiple y los métodos robustos.

5.4 Regresión Lasso

Este modelo, explicado en el Capítulo 2, se plantea considerando una validación cruzada con 5 particiones aleatorias en el conjunto de entrenamiento, buscando mejorar el ajuste del modelo. Y, del mismo modo como se hizo con el modelo Ridge, se transformaron las variables independientes de acuerdo a si el predictor es continuo, ordinal o nominal. A continuación, se presentan los coeficientes obtenidos:

Tabla 5-7: Top 10 variables con mayor coeficiente obtenido por el modelo Lasso

Orden	Variable	Coeficiente	Coeficiente absoluto
1	FIN76	0.171	0.171
2	RG13	0.063	0.063
3	FIN39	-0.057	0.057
4	FIN57	-0.054	0.054
5	RG17	-0.051	0.051
6	FIN46	-0.049	0.049
7	RG18	-0.047	0.047
8	RG19	-0.047	0.047
9	RG93	0.046	0.046
10	RG10	0.043	0.043

En la tabla anterior se tienen las diez variables independientes con mayor coeficiente obtenido por el modelo Lasso, las cuales coinciden en las diez variables con mayor coeficiente en el modelo Ridge, aunque en diferente orden. Con ello, se puede concluir que, aunque se tengan valores diferentes en los coeficientes de los predictores, las variables independientes con mayor relevancia coinciden en ambos modelos de regularización. Es importante mencionar, que por medio de Lasso se obtienen 8 variables

independientes con un valor de 0 en su coeficiente estimado. Se procede a realizar la gráfica de los coeficientes obtenidos por el modelo Lasso:

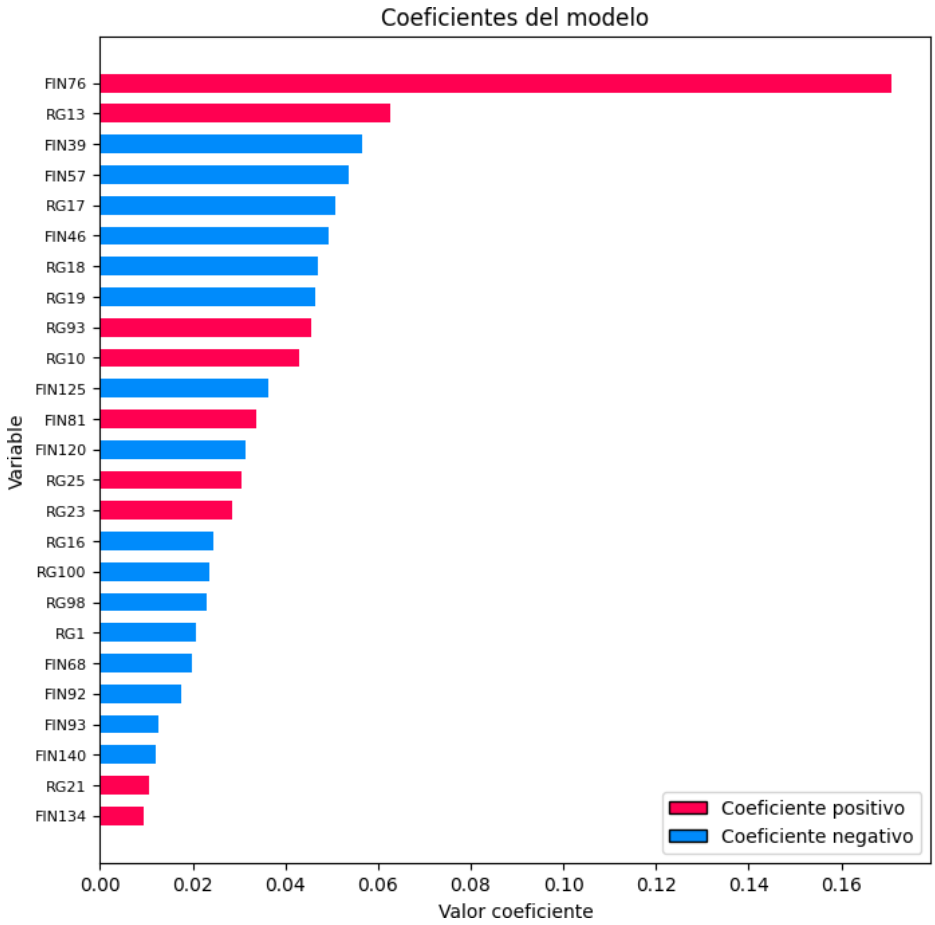
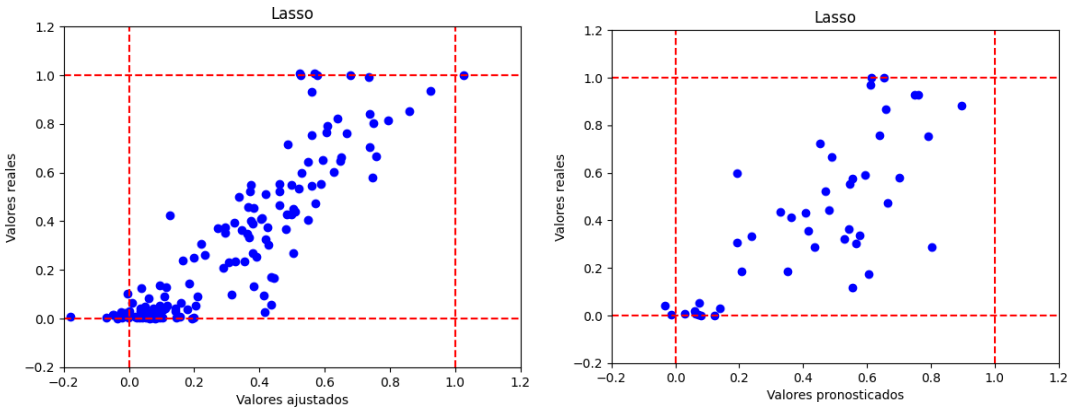


Figura 5-6: Coeficientes de Lasso ordenados por medio de su valor absoluto

Se puede observar en el gráfico anterior que la variable FIN76 aparece con el mayor coeficiente tanto en Ridge como en Lasso, pero en esta última, se observa un coeficiente positivo mucho mayor y mucho más distante en comparación a los otros predictores. Además, por medio de Lasso se obtienen 9 variables predictoras con coeficiente positivo y 16 con negativo. A continuación, se realiza el pronóstico de los datos:



(a) (b)

Figura 5-7: (a) Valores reales vs Valores ajustados (b) Valores reales vs Valores pronosticados por medio del modelo Lasso

En el gráfico anterior, se puede observar que se obtienen valores ajustados (a) negativos y sólo un valor superior a uno. Por otro lado, en los valores pronosticados (b) se generan predicciones negativas, pero no mayores a 1. Se procede a hallar las medidas de desempeño.

Tabla 5-8: Resultados del modelo Lasso en el conjunto de entrenamiento y de prueba

Modelo	Datos	MSE	MAE	R^2
Lasso	Entrenamiento	0.0224	0.1077	0.7623

Modelo	Datos	MSE	MAE	R^2
Lasso	Prueba	0.0455	0.1604	0.5426

En la tabla anterior podemos evidenciar que el modelo Lasso genera un menor coeficiente de determinación en comparación al modelo Ridge y mayores errores en el MSE y en el MAE en el conjunto de entrenamiento, pero siendo valores muy similares en ambos métodos de regularización. No obstante, por medio de Lasso se genera un mayor R^2en el conjunto de prueba y menores errores, pasando de un coeficiente de determinación en Ridge de 0.537 a un valor de 0.5426 en Lasso, siendo valores muy cercanos. En comparación a la regresión lineal múltiple vemos mejores valores en las métricas de desempeño en el conjunto de prueba, pero, aun así, la variabilidad explicada del modelo y su capacidad de ajuste y pronóstico es regular.

5.5 Extreme Gradient Boosting (XGBoost)

Este modelo fue explicado en el Capítulo 2 del presente trabajo. Considerando que por medio de los modelos lineales entre la variable respuesta y los predictores no se obtuvieron unos resultados prometedores, se plantea el modelo de Gradient Boosting Regression o XGBoost el cual puede capturar relaciones no lineales entre variables. Por medio de este modelo se puede obtener los coeficientes de importancia, los cuales se refieren a la frecuencia con la que usan las variables para dividir los árboles del algoritmo XGBoost, cuanto más frecuentes, más importantes (Lundberg et al., 2018). A continuación, se anexan los 10 predictores con mayor coeficiente de importancia.

Tabla 5-9: Top 10 variables con mayor coeficiente de importancia obtenidos por XGBoost

Orden	Variable	Coeficiente
1	RG6	0.290
2	FIN70	0.122
3	FIN76	0.068
4	RG1	0.058
5	FIN92	0.046
6	TRX14	0.040
7	RG21	0.039
8	RG16	0.038
9	FIN120	0.037
10	FIN140	0.036

En la tabla anterior se observa que las variables con mayor coeficiente de importancia son aquellas relacionadas a al grupo de riesgo y financiero, teniendo un solo predictor del grupo transaccional.

En la Figura 5-8, es más fácil observar la magnitud que tiene la primera variable en comparación a las demás, la cual tiene un coeficiente muy superior a los obtenidos en las demás variables independientes. Por otro lado, aproximadamente las primeras diez variables, tienen valores de importancia considerables para el modelo dado el pico en los gráficos, y después de ese punto, los coeficientes de importancia son cercanos a 0.

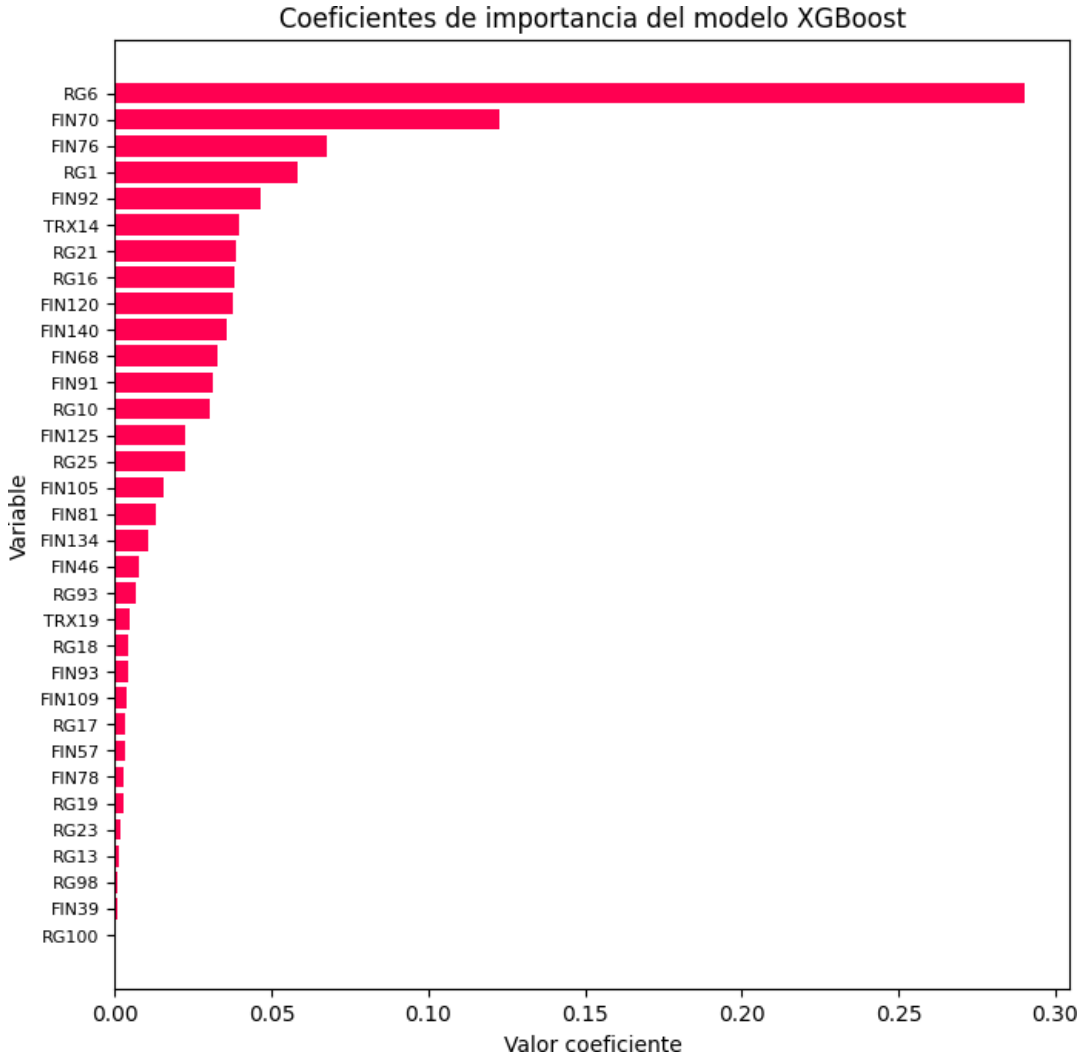
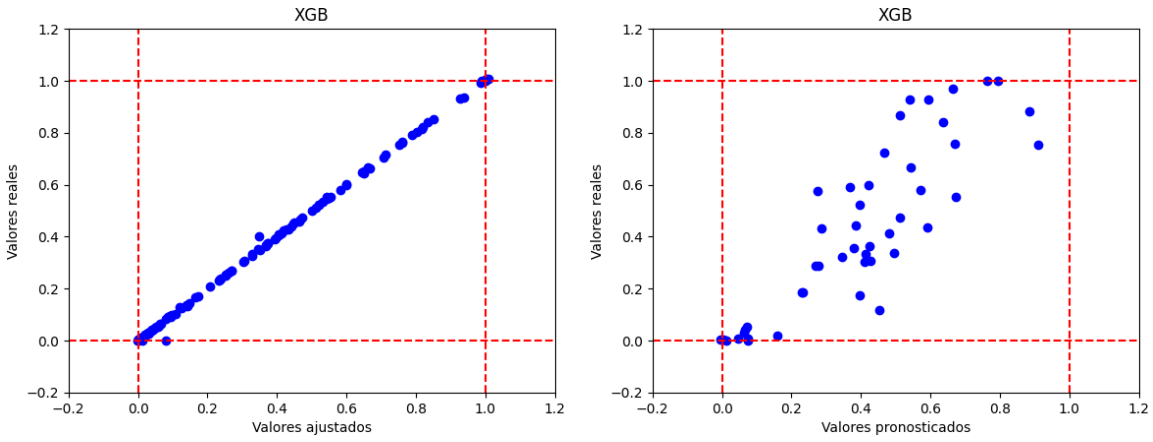


Figura 5-8: Variables predictoras con su coeficiente de importancia obtenidos por XGBoost

Considerando el modelo XGBoost, y la importancia de los predictores, se procede a realizar el pronóstico en el conjunto de prueba y a analizar los valores en el conjunto de entrenamiento. En la Figura 5-9 podemos observar que, en comparación a los modelos lineales planteados previamente, por medio del modelo XGBoost se tienen valores en el rango de la variable respuesta, estando acotados entre 0 y 1. Adicionalmente, se observa en (a) un sobreajuste en los datos de entrenamiento.



(a) (b)

Figura 5-9: (a) Valores reales vs Valores ajustados (b) Valores reales vs Valores pronosticados por medio del modelo XGBoost

Tabla 5-10: Resultados del modelo XGBoost en el conjunto de entrenamiento y de prueba

Modelo	Datos	MSE	MAE	R^2
XGB	Entrenamiento	0.0002	0.0062	0.9981

Modelo	Datos	MSE	MAE	R^2
XGB	Prueba	0.0307	0.1287	0.6922

Los resultados obtenidos la Tabla 5-10 se evidencia que en el conjunto de entrenamiento generan errores cuadráticos medios y absolutos muy pequeños en comparación a los modelos lineales, y un coeficiente de determinación cercano a 1. Por otro lado, en el conjunto de prueba se tiene un coeficiente de determinación de 0.7 aproximadamente, lo cual genera un mejor grado de predicción si lo comparamos con la regresión lineal múltiple o los modelos de regularización. Además, por medio de XGBoost se obtiene el menor MSE y MAE que se ha producido hasta el momento, pero generando un sobreajuste en el conjunto de entrenamiento.

5.6 Bosques Aleatorios

Este modelo fue explicado en el Capítulo 2 del presente trabajo. Los bosques aleatorios fueron usados tanto en la reducción de dimensionalidad como en la selección de variables

dado que permiten capturar relaciones no lineales entre los predictores y la variable respuesta, además de su gran capacidad de predicción (Breiman, 2001). A continuación, se anexan las variables con mayor importancia del modelo:

Tabla 5-11: Top 10 variables con mayor importancia obtenido por el modelo de Random Forest

Orden	Variable	Importancia	Impor. Acum.
1	RG6	0.282	0.282
2	RG21	0.096	0.378
3	RG16	0.057	0.436
4	RG1	0.052	0.488
5	RG10	0.050	0.538
6	RG100	0.047	0.584
7	RG25	0.030	0.615
8	FIN120	0.030	0.645
9	TRX14	0.024	0.669
10	FIN70	0.023	0.691

En la tabla anterior se evidencia que las diez variables con mayor importancia del modelo son aquellas relacionadas a medidas de riesgo y a variables financieras, en especial aquellas relacionadas con su calificación y el riesgo percibido por la entidad financiera. Por otro lado, se observa que las diez variables explican cerca del 70% de la importancia del modelo, dado que, como se mencionó en las secciones pasadas, la sumatoria de las importancias de las variables independientes da como resultado 1.

En la Figura 5-10: Variables predictoras con su coeficiente de importancia obtenidos por el modelo de Bosques AleatoriosFigura 5-10 es posible apreciar la primera variable es la que genera la mayor importancia de todos los predictores y que, como se vio con el modelo de XGBoost, las primeras diez variables son las que generan un mayor coeficiente de importancia acumulada. Ahora, se procede a realizar el pronóstico en el conjunto de prueba.

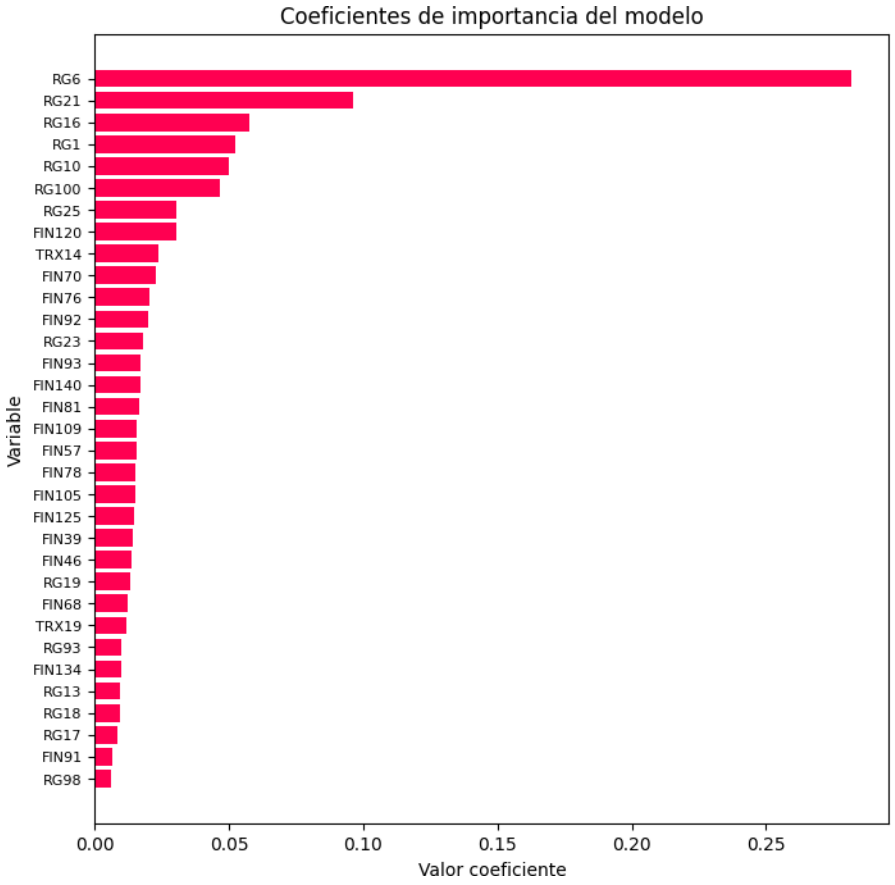
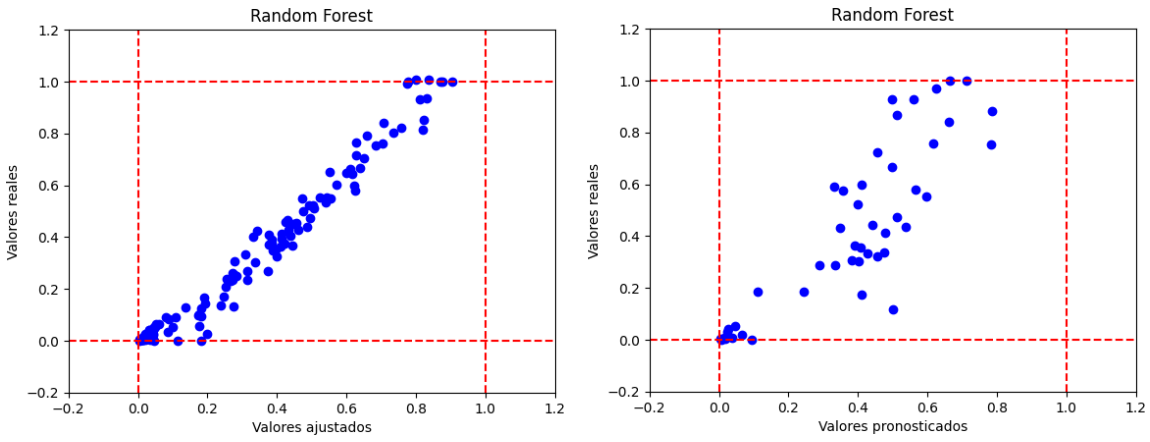


Figura 5-10: Variables predictoras con su coeficiente de importancia obtenidos por el modelo de Bosques Aleatorios



(a) (b)

Figura 5-11: (a) Valores reales vs Valores ajustados (b) Valores reales vs Valores pronosticados por medio del modelo Bosques Aleatorios

En la gráfica anterior se puede observar que los valores que se pronostican por medio de bosques aleatorios se encuentran entre 0 y 1, lo cual, en comparación a la regresión o los métodos de regularización, genera una ganancia en la predicción. En comparación al modelo XGBoost, cuando se observan los valores de ajuste (a) en el modelo de Bosques aleatorios no se genera una nube de puntos tan ajustada, teniendo un menor sobreajuste. A continuación, las medidas de desempeño.

Tabla 5-12: Resultados del modelo Bosques Aleatorios en el conjunto de entrenamiento y de prueba

Modelo	Datos	MSE	MAE	R^2
Random Forest	Entrenamiento	0.0041	0.0419	0.9562

Modelo	Datos	MSE	MAE	R^2
Random Forest	Prueba	0.0302	0.1253	0.6967

Por medio del modelo de Bosques Aleatorios, se obtiene mayores valores en el error cuadrático medio y absoluto en el conjunto de entrenamiento, y un coeficiente de determinación un poco menor en comparación al modelo XGBoost pero sin incurrir en sobreajuste como este último. En el conjunto de prueba se obtienen valores levemente menores en las medidas del error, y un coeficiente de determinación similar en los Bosques aleatorios con respecto al XGBoost, siendo de 0.6922 y de 0.6967, respectivamente.

5.7 Extra Trees Regressor (Extra Trees)

Este modelo, el cual fue explicado en el Capítulo 2 del presente trabajo, se basa en los árboles de regresión para su implementación, usando una partición recursiva y una división de nodos de forma aleatoria, lo que genera robustez y gran poder predictivo (Geurts et al., 2006).

En la Tabla 5-13 se observa que las 10 variables con mayor importancia generan aproximadamente el 72% de la importancia acumulada en el modelo, importancia que, al igual que en los Bosques Aleatorios, se normaliza para que la sumatoria de los coeficientes de importancia de los predictores sea igual a 1.

Tabla 5-13: Top 10 variables con mayor importancia obtenido por el modelo de Extra Trees

Orden	Variable	Importancia	Impor. Acum.
1	RG6	0.392	0.392
2	RG100	0.071	0.464
3	RG1	0.050	0.513
4	FIN76	0.049	0.562
5	FIN93	0.040	0.602
6	TRX14	0.028	0.630
7	RG10	0.023	0.654
8	RG21	0.022	0.676
9	RG93	0.021	0.697
10	FIN57	0.021	0.717

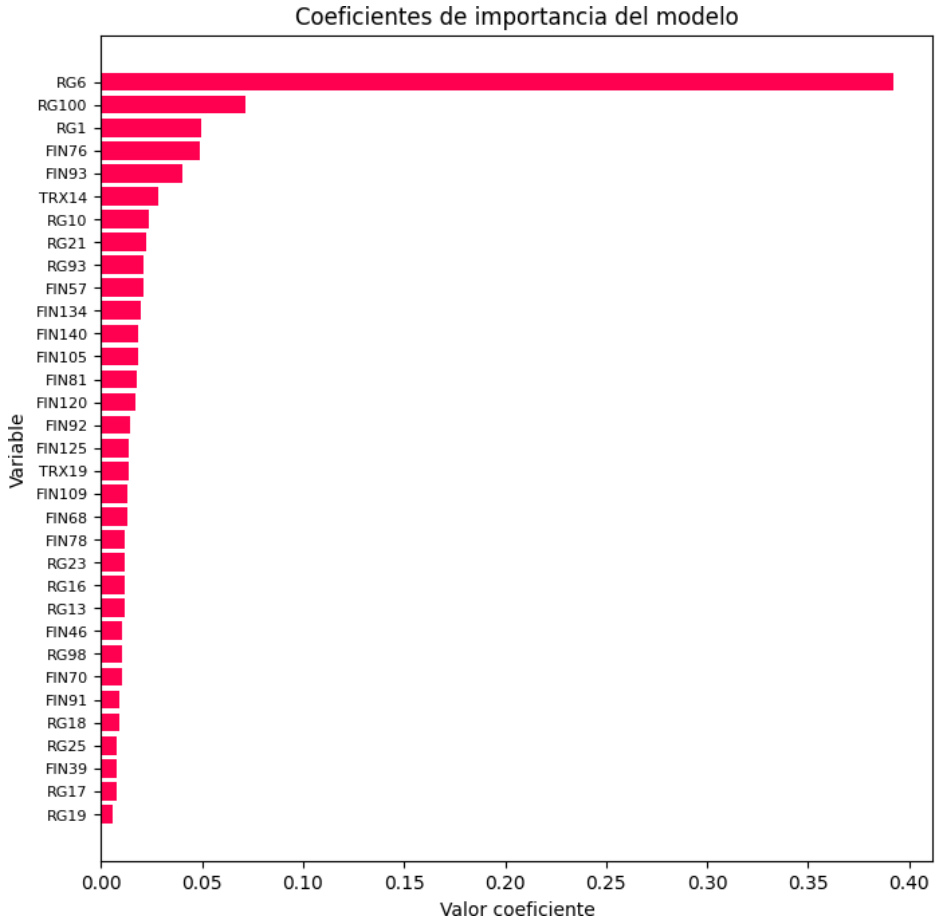
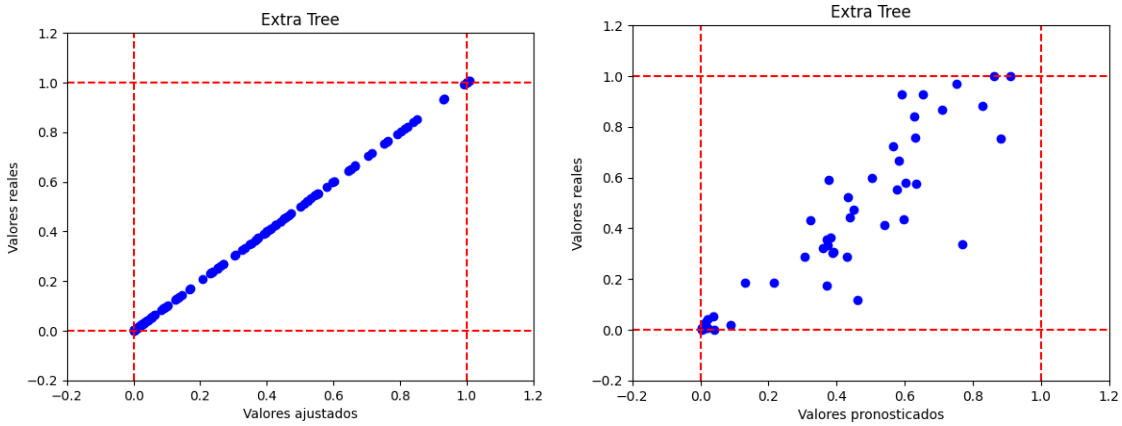


Figura 5-12: Variables predictoras con su coeficiente de importancia obtenidos por Extra Trees

En el gráfico se observa que, similar a lo obtenido en el modelo XGBoost y en el modelo de Bosques Aleatorios, la primera variable es la que genera la mayor importancia para el modelo, siendo considerablemente alta en comparación a los demás predictores. También se evidencia que, a diferencia de los otros modelos, los coeficientes de importancia del modelo Extra Trees no están tan cercanos a 0. Ahora, se procede a pronosticar la variable respuesta en el conjunto de prueba.



(a) (b)

Figura 5-13: (a) Valores reales vs Valores ajustados (b) Valores reales vs Valores pronosticados por medio del modelo Extra Trees

En los valores ajustados (a) se observa la misma situación que en el modelo XGBoost, donde los valores en el conjunto de entrenamiento presentan sobreajuste, por otro lado, en el conjunto de prueba (b), se observa que por medio de Extra Trees, al igual que en los modelos de Bosques Aleatorios y de XGBoost los valores pronosticados se encuentran entre 0 y 1. A continuación, las medidas de desempeño.

Tabla 5-14: Resultados del modelo Extra Trees en el conjunto de entrenamiento y de prueba

Modelo	Datos	MSE	MAE	R^2
Extra Trees	Entrenamiento	0.0000	0.0000	1.0000

Modelo	Datos	MSE	MAE	R^2
Extra Trees	Prueba	0.0205	0.0998	0.7945

Cuando se analizan las medidas de desempeño en el conjunto de entrenamiento, se observa sobreajuste, situación que también ocurrió en el modelo XGBoost pero en menores proporciones. En el conjunto de prueba, se obtiene un coeficiente de determinación de 0.7945, generando un buen grado de predicción en el pronóstico de la perdida crediticia esperada y permitiendo explicar la variabilidad en la muestra de prueba en una buena proporción. Con respecto a los errores cuadráticos y absolutos, se genera una mejora considerable en comparación con los otros modelos expuestos, donde el MSE es de 0.0205 y el MAE de 0.0998, siendo el modelo con menores errores cuadráticos y absolutos.

El modelo Extra Trees es el que obtiene mejores medidas de desempeño en el conjunto de entrenamiento y de prueba. Por ende, este el modelo que será empleado en el próximo Capítulo, pero buscando reducir el sobreajuste en el conjunto de entrenamiento.

6. Capítulo 6

Simplificación del mejor modelo e interpretación

Considerando que el modelo Extra Trees fue el modelo que obtuvo un mejor desempeño en la base de datos tanto de entrenamiento como de prueba, es el modelo que se plantea para la predicción de la pérdida crediticia esperada. Con este modelo se procede a continuación a una simplificación en busca de parsimonia y evaluar interpretabilidad del modelo final.

6.1 Depuración adicional de variables

En el análisis de la importancia de los predictores del modelo Extra Trees, se observó que un grupo específico de variables independientes explicaba en su gran mayoría la importancia total en el modelo, para ser específicos, 10 predictores generaban el 72% de la importancia total del modelo. Con el fin de reducir el sobreajuste en el conjunto de datos de entrenamiento y en busca de un modelo más parsimonioso, se plantea un menor número de variables independientes para el modelo. Para ello, se establece un umbral de corte relacionado a la importancia total en el modelo, dejando sólo los 10 predictores que explican el 72% de la importancia total. Es importante aclarar, que el umbral es necesario considerando que el modelo Extra Trees, al igual que los Bosques Aleatorios y el modelo XGBoost, brindan un coeficiente de importancia a cada uno de los 33 predictores.

Una vez definido el umbral de corte de la importancia total, el cual corresponde a los predictores que explican el 72% de la importancia total en el modelo, se pasa de una base de 33 variables independientes a un total de 10, dejando así, sólo los predictores que mayor importancia tienen en el modelo y en la capacidad de predicción. A continuación, se anexa la gráfica con la importancia de las variables independientes del modelo Extra Trees, el umbral de corte de 72% de la importancia total y la evolución de la importancia acumulada en el modelo.

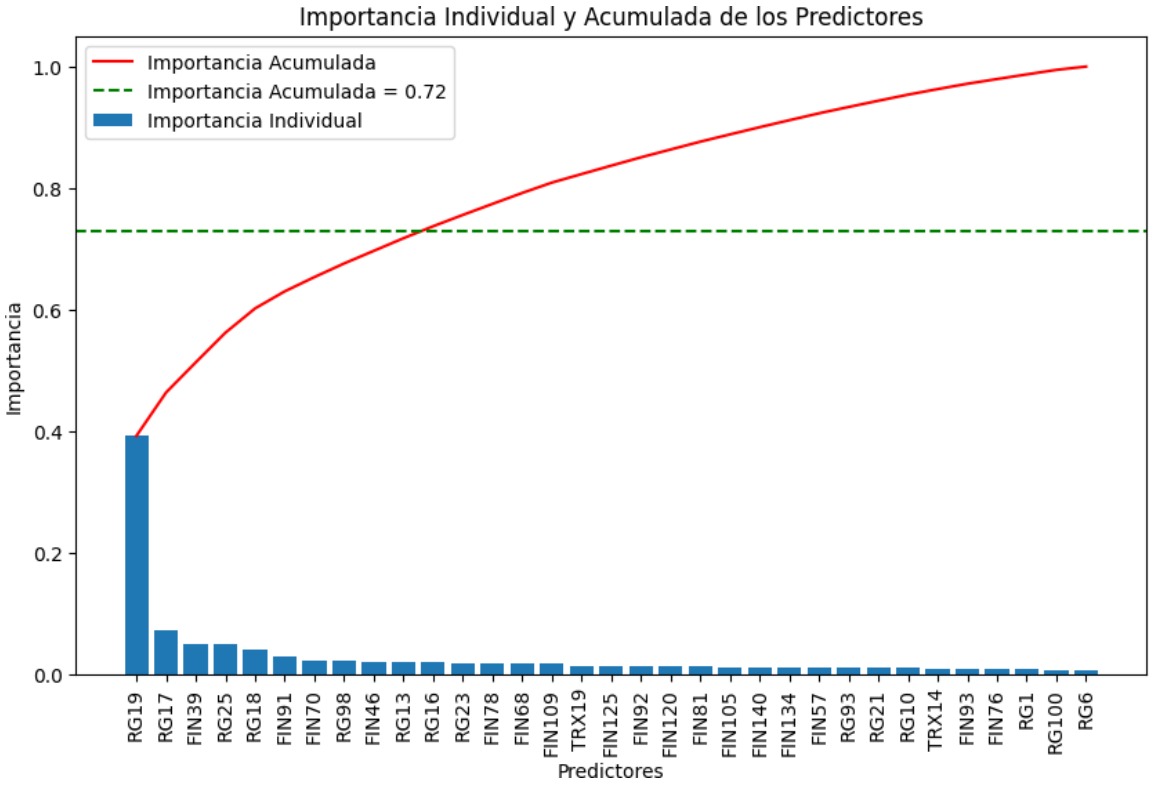


Figura 6-1: Variables predictoras con su coeficiente de importancia obtenidos por Extra Trees de forma individual y acumulada.

En el gráfico anterior se ilustra cómo las primeras variables generan la mayor importancia en el modelo y cómo va decayendo el coeficiente en las variables con menos peso dentro del modelo, evidenciando el umbral de corte del 72% sobre la importancia acumulada, la cual la generen 10 predictores.

Considerando el conjunto de 10 variables independientes que generan el 72% de la importancia total, se particiona de forma aleatoria la base de datos en un conjunto de entrenamiento equivalente al 75% de los datos y un conjunto de prueba de 25%. Buscando en esta sección un menor sobreajuste y un modelo más parsimonioso. Luego de la partición de los datos en el conjunto de entrenamiento y de prueba, se plantea nuevamente el modelo Extra Trees con las 10 variables, a continuación, los predictores con mayor importancia.

Tabla 6-1: Coeficiente de importancia individual y acumulado de las variables del modelo Extra Trees.

Orden	Variable	Importancia	Impor. Acum.
1	RG6	0.467	0.467
2	RG1	0.087	0.554
3	FIN76	0.084	0.639
4	FIN93	0.058	0.697
5	RG100	0.057	0.754
6	RG93	0.053	0.806
7	TRX14	0.052	0.858
8	RG21	0.050	0.907
9	FIN57	0.049	0.957
10	RG10	0.043	1.000

En la tabla anterior se observan los 10 predictores con su coeficiente de importancia para el modelo Extra Trees, teniendo en su gran mayoría variables asociadas al riesgo y a indicadores financieros. Adicionalmente, se observa que luego de pasar de 33 predictores a 10, la variable RG6 genera un mayor coeficiente de importancia al considerar este conjunto reducido de predictores, y las demás variables independientes generan una importancia mayor.

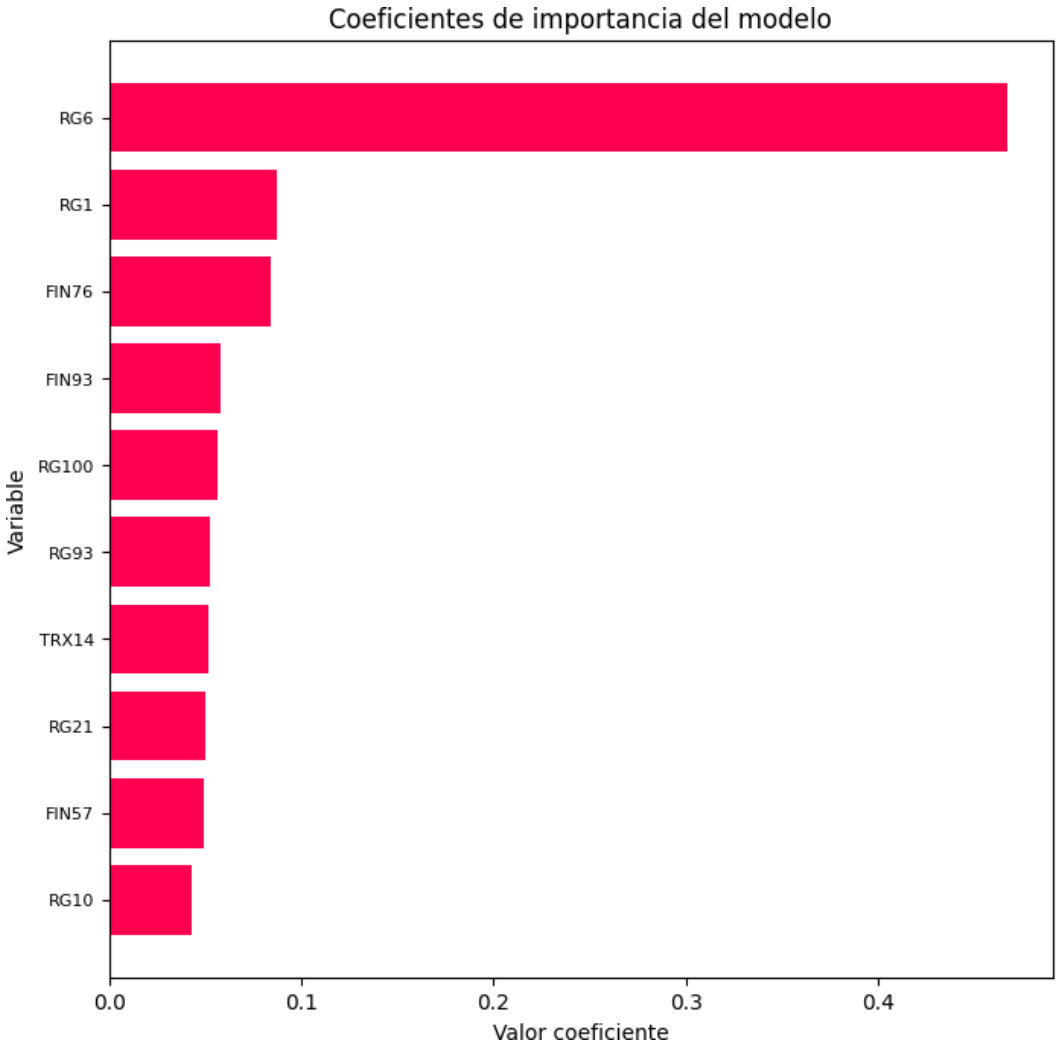
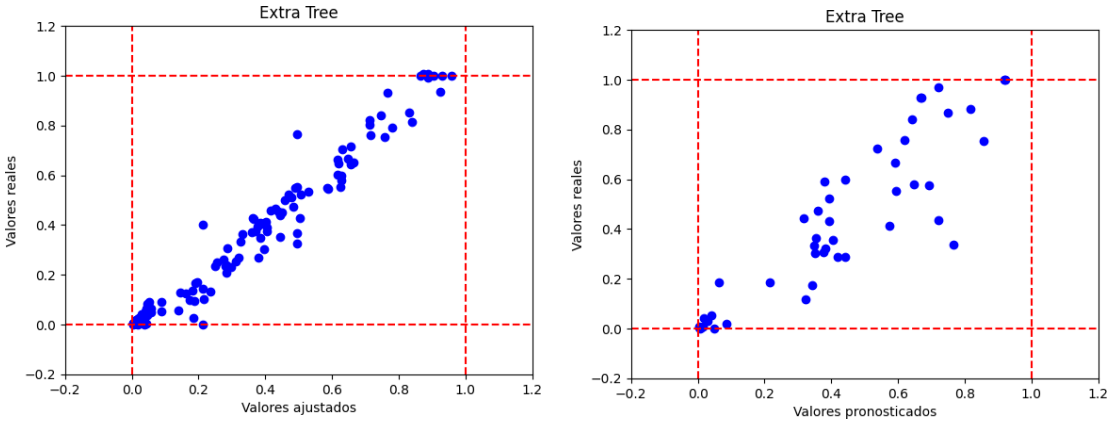


Figura 6-2:Variables predictoras con su coeficiente de importancia obtenidos por Extra Trees.

En la figura anterior se observa nuevamente que la variable RG6 es la que obtiene la mayor importancia del modelo, y la importancia de las otras variables está en un rango similar, teniendo coeficientes de importancia que oscilan entre 0.04 y 0.05. A continuación, se procede a realizar la predicción del modelo considerando el conjunto de predictores con 10 variables.



(a) (b)

Figura 6-3: (a) Valores reales vs Valores ajustados (b) Valores reales vs Valores pronosticados por medio del modelo Extra Trees en el conjunto de 10 predictores

En el gráfico anterior se puede observar que los valores ajustados (a) ya no presentan tanto sobreajuste como se tenía en el modelo Extra Trees con 33 predictores, teniendo una nube de puntos un poco más dispersa. Con respecto a los valores pronosticados (b) éstos siguen ocurriendo entre 0 y 1. Ahora, se procede a hallar las medidas de desempeño.

Tabla 6-2:Resultados del modelo Extra Trees luego de la reducción de predictores en el conjunto de entrenamiento y de prueba

Modelo	Datos	MSE	MAE	R^2
Extra Trees	Entrenamiento	0.0038	0.0395	0.9592

Modelo	Datos	MSE	MAE	R^2
Extra Trees	Prueba	0.0199	0.1073	0.8007

En cuanto a las métricas de desempeño, se evidencia un menor sobreajuste en el conjunto de entrenamiento, obteniendo un R^2 menor a 1 y errores cuadráticos y absolutos mayores a 0. Por lo que pasar de 33 variables predictoras a 10, permitió aumentar la capacidad de generalización del modelo y disminuir el sobreajuste. Por otro lado, al analizar el conjunto de datos de prueba, se observa una leve mejora en el coeficiente de determinación y menores errores, aunque los resultados en general son muy similares a los obtenidos previamente. En otras palabras, se obtiene un menor sobreajuste y mayor capacidad de generalización del modelo y buenos resultados en la calidad de predicción.

6.2 Resultados del mejor modelo simplificado

Una vez elegido el mejor modelo y el conjunto final de variables predictoras, se hallan las medidas de desempeño en todo el conjunto de datos, a continuación, los resultados.

Tabla 6-3: Resultados del modelo simplificado Extra Trees en el conjunto de datos completo considerando 10 predictores.

Modelo	Datos	MSE	MAE	R^2
Extra Tree	Conjunto de datos completo	0.0078	0.0564	0.9199

En la tabla anterior podemos observar un buen ajuste del modelo en la totalidad de las observaciones, teniendo un R^2 de 0.9199, un MSE de 0.0078 y un MAE de 0.0564, siendo resultados bastante alentadores dado que el modelo Extra Trees con 10 predictores permite explicar la variabilidad de la pérdida crediticia esperada con errores cuadráticos medios y absolutos considerablemente bajos.

Al comparar los resultados del modelo en todo el conjunto de datos a los obtenidos en la Sección anterior, se evidencia un leve aumento en las medidas de error en el conjunto de entrenamiento, pasando de un MSE de 0.0038 y un MAE de 0.0395 en los datos de entrenamiento, a un MSE de 0.0078 y un MAE de 0.0564 en el conjunto datos completo. No obstante, el modelo en el conjunto de datos completo mejora considerablemente al compararlo contra el conjunto de prueba, pasando de un R^2 de 0.8007 en los datos de prueba a un R^2 de 0.9199 con todas las observaciones. Se procede a graficar el ajuste en el conjunto de datos completo.

En la Figura 6-4 al igual que en las medidas de desempeño, se evidencia que no hay sobreajuste dada la nube de puntos dispersa y al tener medidas de error mayores a 0. Por último, se observa que se siguen generando valores entre 0 y 1, y que todavía se generan valores ajustados muy alejados a la variable respuesta observada.

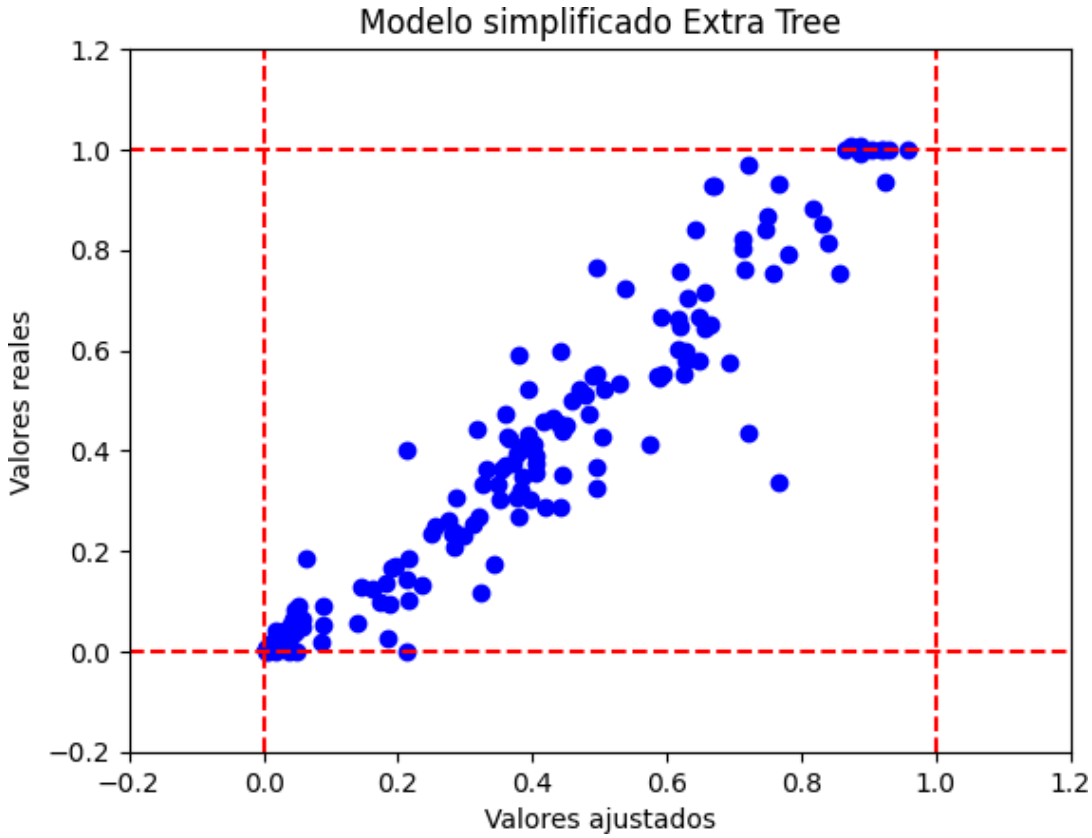


Figura 6-4: Valores reales vs Valores ajustados por medio del modelo simplificado Extra Trees con 10 predictores

Se procede a realizar un análisis de los residuales del modelo, con el fin de encontrar las mayores diferencias entre los valores ajustados y los valores reales. En la Figura 6-5 se puede observar una carencia en el ajuste considerando un conjunto de predictores reducido de 10 variables. No obstante, en busca de parsimonia y de no incurrir en sobreajuste como se evidenció en el modelo con los 33 predictores, no se incorporan variables adicionales que podrían aumentar el ajuste.

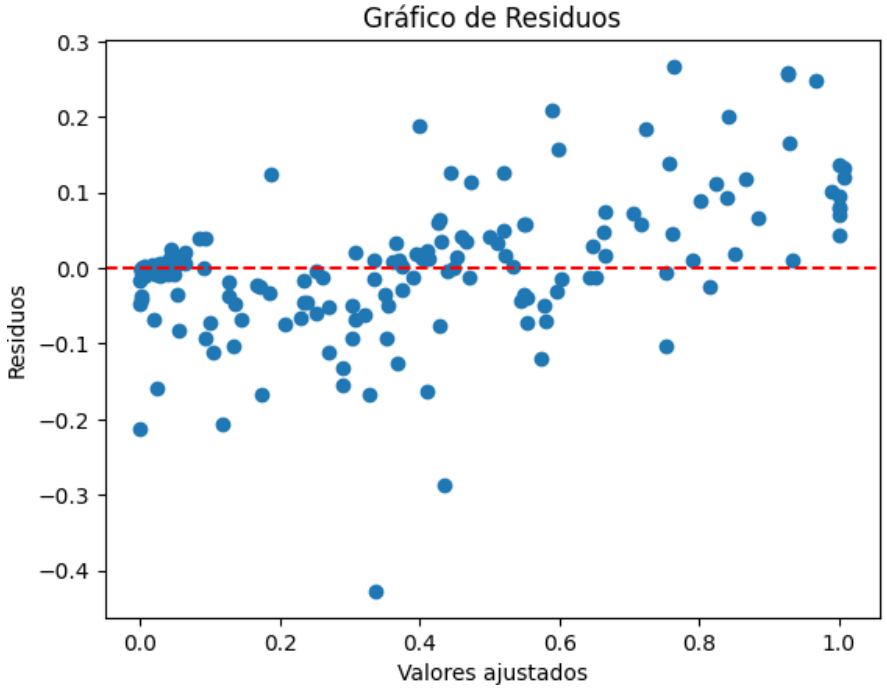


Figura 6-5: Gráfico de dispersión de residuos en el conjunto completo de datos considerando el modelo simplificado de Extra Trees con 10 predictores.

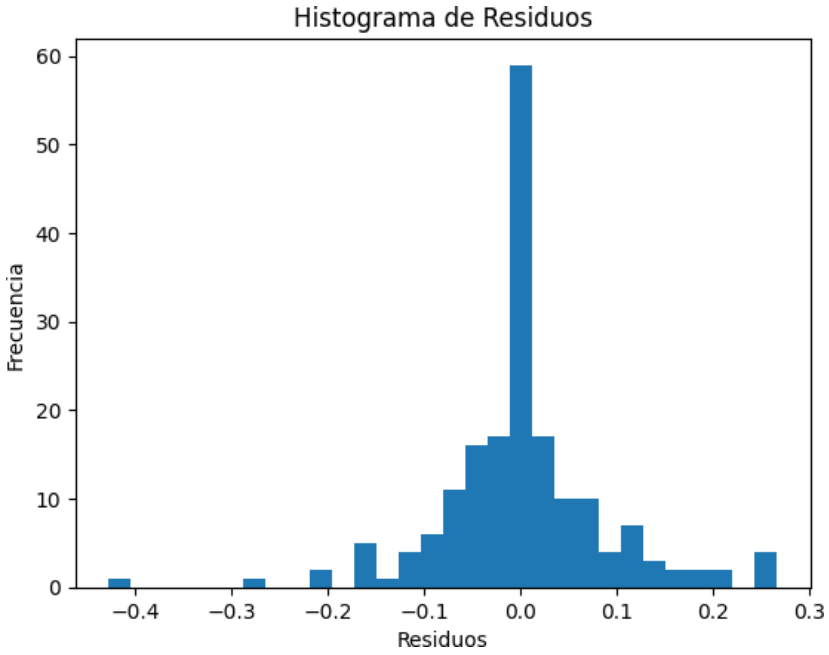


Figura 6-6: Histograma de residuos en el conjunto completo de datos considerando el modelo simplificado de Extra Trees con 10 predictores.

A partir de los gráficos anteriores se aprecia que gran parte de los residuales se encuentran sobre 0, para ser específicos, el 63% se encuentran entre -0.05 y 0.05, y el 80% entre -0.1 y 0.1. Adicionalmente, se tiene valores ajustados que se encuentran por encima o por debajo de 20%, lo cuales estarían afectando las medidas de error en el conjunto de datos completo, esta situación también se observa en la Figura 6-4 donde el modelo simplificado brinda valores ajustados muy cercanos en la mayoría de las observaciones. Pero, cuando se hace énfasis en algunas observaciones puntuales, se evidencian valores ajustados muy alejados a la variable respuesta observada; por lo que muy probablemente esos registros con grandes diferencias están generando los mayores errores absolutos.

Para comprobar lo anterior, se decide encontrar en cuáles registros es dónde se genera una mayor diferencia entre la pérdida crediticia esperada y la ajustada, para así hallar nuevamente las medidas de desempeño en un ejercicio de análisis.

Se encuentra que 5 registros de los 184 que componen el conjunto de datos completo son los que generan mayores errores de ajuste, teniendo diferencias absolutas entre 0.2 a 0.4, siendo valores altos considerando que la variable respuesta es una proporción. Continuando con el ejercicio de análisis, se halla nuevamente las medidas de desempeño, pero sin tener en cuenta esos 5 registros. Esto con el único fin de observar si esa pequeña muestra de clientes son los que afectan en una mayor proporción el MSE, MAE y coeficiente de determinación de predicción.

Tabla 6-4: Ejercicio de análisis para observar la incidencia de algunos registros en las medidas de desempeño

Modelo	Datos	MSE	MAE	R^2
Extra Trees	Conjunto sin los 5 registros atípicos	0.0054	0.0496	0.9432

En la tabla anterior, se tiene las medidas de desempeño considerando el conjunto de datos completo sin los 5 registros, con el fin de comparar estos resultados contra las medidas de desempeño de la Tabla 6-3Tabla 6-2 la cual tiene las medidas de MSE, MAE y R^2 para todo el conjunto de datos. En la tabla anterior se observa que sí hay una diferencia notoria en las medidas de desempeño cuando se hallan sin los 5 registros mencionados,

aumentando el coeficiente de determinación en 0.0233, disminuyendo el MAE en 0.0058 y mejorando el MSE en 0.0024.

Se puede evidenciar que son pocos los registros los cuales hacen que el desempeño del modelo no sea el mejor, dado que al tener un grupo de clientes con un riesgo alto percibido por la entidad financiera y con particularidades económicas, financieras y transaccionales, dichos casos puntuales afectan la capacidad de generalización del modelo. No obstante, esos registros deben permanecer dado que hacen parte de la muestra de clientes que se pretende pronosticar.

Es importante señalar que se plantearon transformaciones de la variable respuesta buscando mejorar el ajuste y la predicción de los datos, como logaritmo o raíz cuadrada. Sin embargo, no se mejoró sustancialmente la predicción y lo más importante, como se pretende analizar el resultado de la variable respuesta y su interpretabilidad con los predictores, se considera más adecuado no realizar transformaciones de la variable respuesta, ni de las variables independientes, haciendo claridad en que en los predictores hay valores negativos y ceros que imposibilitan ciertas transformaciones de datos.

6.3 Interpretación

Dado que el modelo final no fue un modelo paramétrico, una forma de interpretar los resultados es por medio de los Shapley values, los cuales fueron explicados en el Capítulo 2 del presente trabajo. Estos valores permiten evidenciar qué predictores son más relevantes y cómo afectan la variable respuesta.

Para la interpretación se presentan varias figuras buscando entender las relaciones que están generando las variables predictoras hacia la pérdida crediticia esperada por medio del modelo ExtraTree Regressor considerando sólo los 10 predictores finales en el conjunto de datos completo. A continuación, se adjuntan los gráficos.

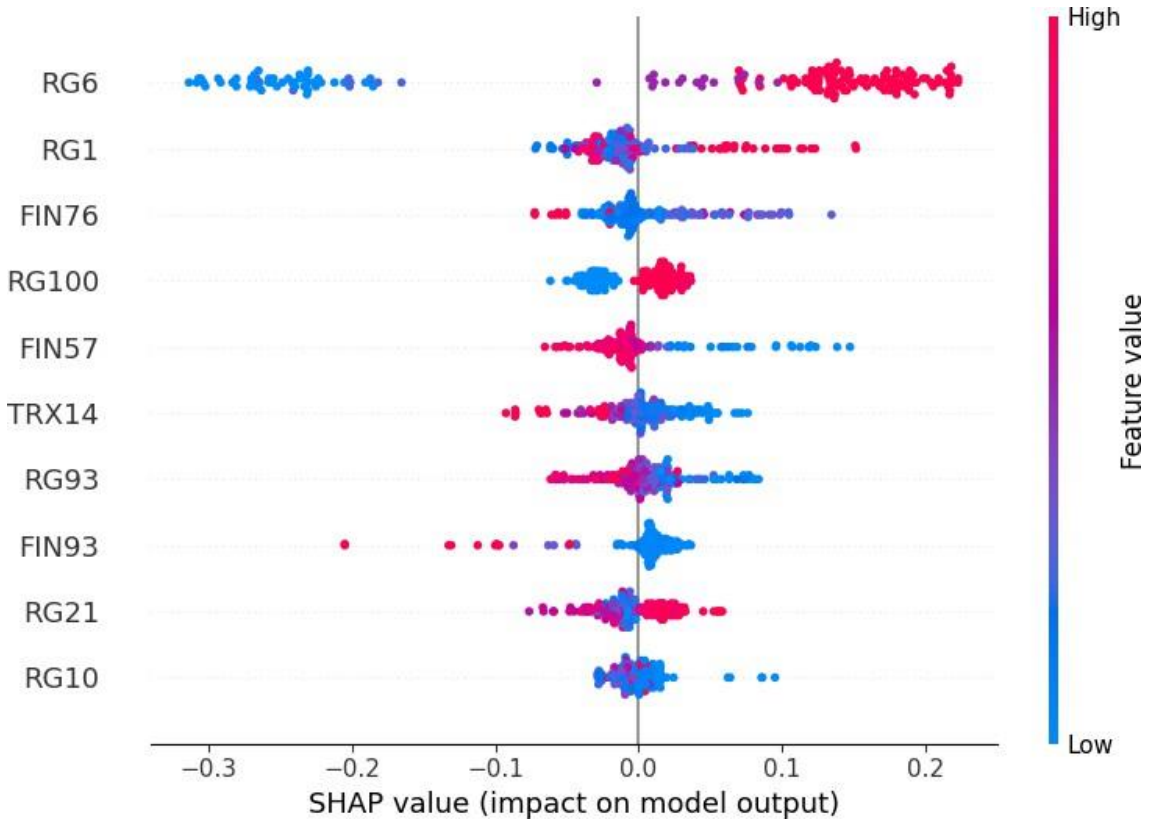


Figura 6-7: Valores de SHAP considerando todas las variables predictoras hacia la variable independiente.

En el eje de la izquierda se encuentran listadas las 10 variables independientes finales, en el eje de la derecha se indica el color que toma las variables independientes en función de su aumento o disminución, así, los valores altos de los predictores están de color rojo, mientras los valores bajos están de color azul. Por último, en el eje inferior se encuentra el impacto sobre la variable respuesta, la cual está en un rango de -0.3 a 0.2. Es importante resaltar que los colores permiten analizar la relación del predictor con la pérdida crediticia esperada, dado que, ante un valor mayor del predictor, es decir, color rojo, se puede observar si está afectando positiva o negativamente a la variable respuesta. Este análisis es equivalente cuando el predictor tiene valores bajos (color azul).

La primera variable predictora en el grupo final seleccionado es RG6, siendo la variable independiente que tiene un mayor coeficiente de importancia de acuerdo al modelo Extra Trees. Este predictor tiene valores rojos en el eje de la derecha del gráfico, y azules en el

eje de la izquierda, lo que significa que valores altos de RG6 (los puntos rojos) generan un aumento en la pérdida crediticia esperada. Y los puntos azules, que denotan valores pequeños de RG6, generan una disminución en la variable respuesta. Esta relación tiene gran coherencia dado que RG6 es una variable de calificación, la cual entre más alta sea indica un mayor nivel de riesgo, que debe trasladarse a mayor pérdida crediticia esperada, mientras que una menor calificación, genera una menor pérdida crediticia esperada.

Continuando con las variables del grupo de riesgo, se observa que RG100 también está generando relaciones que guardan sentido con la variable respuesta. Cuando se analiza la variable RG100, los valores rojos están ubicados en el eje de la derecha de la variable respuesta, lo que indica que, a mayores valores, existe una mayor pérdida crediticia esperada. Esta variable indica si el cliente ha entrado en un estado de incumplimiento o no, por lo que toma el valor de 1 si el cliente entró en default, y toma el valor de 0 si no ha estado en incumplimiento. Esto significa que los clientes que han estado en default generan un mayor impacto en la variable respuesta. El análisis de RG100 es similar al de la variable RG1 y RG21, las cuales indican la calificación al momento del desembolso, y la calificación en fecha de análisis, respectivamente. En ambas variables se observan puntos rojos en el eje de la derecha de la pérdida crediticia esperada, donde a mayor número en su calificación, un mayor valor de la variable respuesta.

Ahora bien, si observamos las variables financieras encontramos a FIN57 la cual mide el nivel de rentabilidad sobre el patrimonio, e indica que los valores de baja rentabilidad aumentan la pérdida crediticia esperada, y que los valores altos, la disminuyen. Esto cumple con la teoría económica donde se espera que los clientes más rentables tengan una menor pérdida crediticia esperada. Por otro lado, la variable FIN93 que brinda información relacionada a la rotación de la cartera, genera una mayor pérdida crediticia esperada si se tiene una rotación baja, mientras una rotación alta de cartera, asociada a mayores ventas y mejores políticas de cobranza, genera un menor valor de la variable respuesta. Por último, la variable FIN76 correspondiente a un promedio de 12 meses, es más difícil de generalizar dado que no se sabe con certeza si un valor mayor o menor, conduce a aumentar o disminuir la variable respuesta, por lo que se debe analizar caso a caso.

La única variable transaccional en el grupo final de predictoras es TRX14, ésta variable genera una diminución en la variable respuesta al tener valores altos, y un aumento en la pérdida crediticia esperada al tener valores bajos. Esto guarda coherencia con la naturaleza de la variable, dado que TRX14 mide la relación de entradas transaccionales, las cuales están ligadas a un aumento en ventas sobre el gasto financiero, por lo que menores flujos de ingresos o un aumento de su endeudamiento brinda un valor menor, lo cual está ligado a un mayor deterioro. Por otro lado, un valor alto de TRX14 estaría reflejando una mayor dinámica en ventas o una disminución en su endeudamiento financiero, estando ligado a una menor pérdida crediticia esperada.

En general, se observa que las relaciones del modelo son correctas considerando la teoría económica y la lógica del negocio. Adicionalmente, por medio de los SHAP values también se puede analizar la importancia de las variables. En el siguiente gráfico se pueden evidenciar los predictores más importantes del modelo.

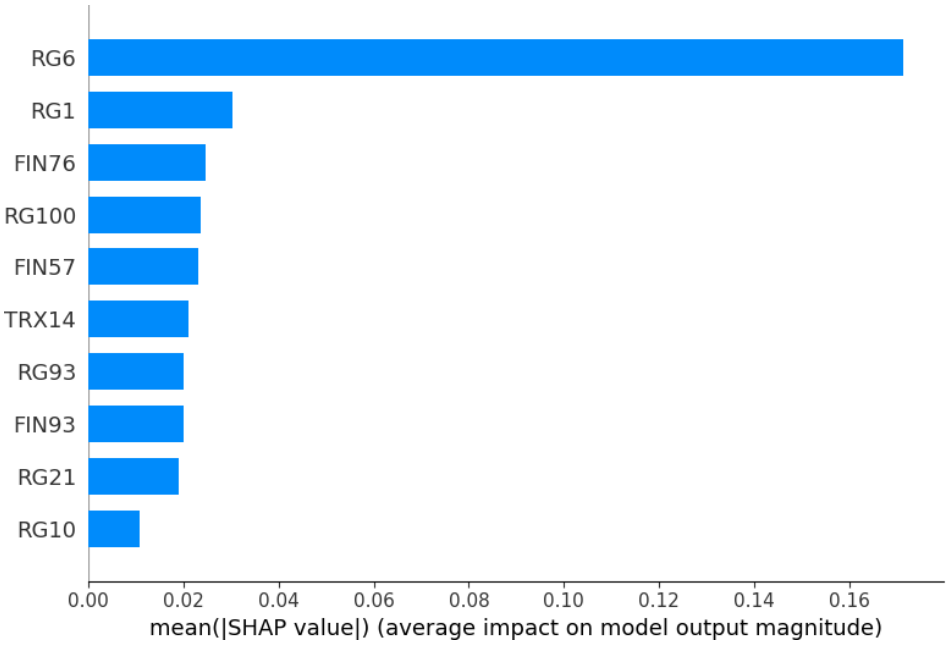
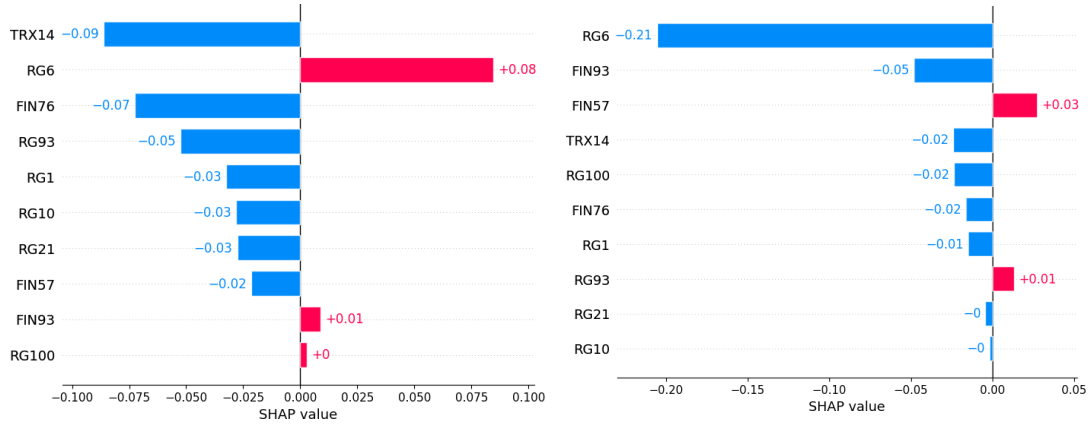


Figura 6-8: Importancia relativa de los predictores en la variable respuesta de acuerdo a los valores de SHAP

En el gráfico anterior se evidencia la importancia que tiene cada predictor en el impacto promedio del ajuste, observando que la variable RG6, es la que mayor impacto genera. Por otro lado, si se compara el orden de los predictores de esta gráfica con la Figura 6-2, se observa que las tres primeras variables independientes: RG6, RG1 y FIN76 permanecen como los predictores más relevantes en ambas figuras, mientras los demás predictores se encuentran en diferente orden. La diferencia en el ordenamiento de las variables independientes se debe a que la Figura 6-2 mide la importancia de las variables explicativas en el modelo Extra Trees, siendo una medida relacionada a la disminución de impureza, mientras el gráfico anterior, relacionado a los SHAP values, brinda el efecto que tiene cada predictor en el ajuste de cada observación.

Otra ganancia de los SHAP values es que permiten generar las relaciones e importancias relativas para cada observación. Por lo que se pueden generar gráficos para cada cliente y así observar la interacción de la variable respuesta con los predictores de manera individual, dado que, como se muestra en la Figura 6-7, cada punto rojo o azul que impacta la pérdida crediticia esperada, es una observación. A continuación, se muestran los gráficos de las primeras 4 predicciones.



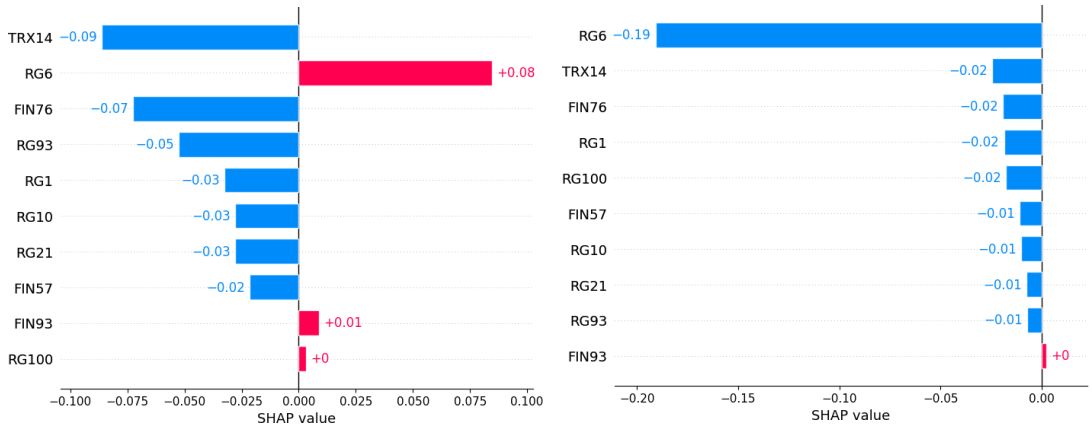


Figura 6-9: Gráficos de SHAP Values para las primeras cuatro observaciones del conjunto de prueba.

En estos gráficos se evidencia cómo cada predicción se ve afectada por las variables independientes, donde de acuerdo a los valores que tenga ese cliente, le podrán afectar en mayor o menor medida los predictores. Es importante señalar que la interpretabilidad se debe realizar considerando la Figura 6-7, la cual brinda un panorama completo.

6.4 Consistencia

En los resultados del presente trabajo, se realizó la ejecución del código en diez ocasiones considerando diferentes conjuntos de datos de entrenamiento y de prueba que fueron elegidos de forma aleatoria guardando la relación de 75% de registros para entrenamiento de 25% para prueba. Los resultados no generaron diferencias significativas, y se muestran los valores obtenidos en la última ejecución.

7. Capítulo 7

Conclusiones y Recomendaciones

7.1 Conclusiones

Los clientes con mayor exposición en la entidad financiera son los causantes de que se generen los movimientos más abruptos en cuanto a constitución o liberación de provisión. Esto se debe a que son estudiados bajo un proceso de análisis individual, por lo tanto, se pueden tener cambios importantes en su pérdida crediticia esperada considerando la nueva información disponible en cada semestre, siendo un proceso reciente que sólo cuenta con historia desde el año 2019, lo que brinda un conjunto de datos con 184 registros. Las variables independientes propuestas en el presente trabajo partieron de los estudios de Wang (2011), Zhang & Chen (2021), Chen (2011), Yeh et al. (2014), entre otros autores, quienes plantearon variables financieras, sectoriales, transaccionales y económicas, y con base en ello, se identificaron en este trabajo un total de 437 variables. Cabe resaltar, que tal como lo plantearon Taghiyeh et al. (2021) se involucraron variables rezagadas en el conjunto de datos.

Para realizar un pronóstico de la pérdida crediticia esperada primero se realizó una reducción de la dimensionalidad de la base de datos y una selección de variables, con el fin de proponer un modelo parsimonioso y que explicara los movimientos de la variable respuesta. En el trabajo se evidenció que no existían relaciones lineales entre la gran

mayoría de predictores con la pérdida crediticia esperada, lo que generó altos valores en las medidas de error MSE y MAE, y un coeficiente de determinación considerablemente bajo en los modelos de regularización, en la regresión lineal múltiple y en los métodos robustos.

Cuando se analizaron los resultados obtenidos por modelos que no parten de relaciones lineales, como los bosques aleatorios, XGBoost y Extra Trees, se observó que los resultados mejoraban considerablemente en comparación a los modelos paramétricos. Estos modelos fueron propuestos por varios autores, los cuales plantearon modelos no paramétricos dadas las relaciones no lineales que se tenían entre la variable respuesta y los predictores, y la gran capacidad de predicción de estos modelos.

El modelo que mejor resultados generó en el conjunto de datos observados en este trabajo fue el de Extremely Randomized Trees (Extra Treess) el cual obtuvo un MSE de 0.0078 y un R^2 de 0.9199 en el conjunto completo de datos, a excepción de 5 clientes puntuales donde se obtuvieron valores muy diferentes a los observados. Estos 5 valores generaron un cambio en los resultados obtenidos, debido a que aumentaron las medidas de error y disminuyeron el coeficiente de determinación.

Se encontró que las variables que mejor predicen la pérdida crediticia esperada son las relacionadas al nivel de riesgo percibido por la entidad financiera hacia el cliente, las variables financieras y las variables transaccionales. Siendo RG6, RG1, RG100, RG93 FIN76, FIN93 Y TRX14 los predictores que más importancia tuvieron en el modelo Extra Trees.

Por último, por medio de los SHAP values se interpretaron los resultados de las predicciones del modelo Extra Trees, en los que se observó que las direcciones de las relaciones entre variables predictoras relacionadas al riesgo y a sus estados financieros, con la respuesta, guardan sentido con la teoría económica, donde las variables que indican mayor rentabilidad, liquidez y mejor desempeño financiero de los clientes impactan inversamente a la pérdida crediticia esperada de la entidad financiera, y las variables relacionadas al riesgo del cliente tienen una relación directa, donde a mayor riesgo percibido mayor será la pérdida crediticia esperada de la entidad financiera.

En general, los resultados obtenidos son bastante alentadores considerando el conjunto de datos tan limitado que se tiene debido a la información disponible, donde se logró pasar de 437 variables a 10 predictores para pronosticar la pérdida crediticia esperada. Adicionalmente, se consiguió interpretar los resultados obtenidos de un modelo no paramétrico como lo es el modelo Extra Trees, evidenciando que las relaciones entre la variable respuesta y los predictores guardan sentido con la teoría económica.

7.2 Recomendaciones

Se debe seguir almacenando información relacionada a los clientes con mayor pérdida crediticia esperada para que en un futuro se tenga una base de datos con mayor número de registros, dado que como se evidenció en el presente trabajo el conjunto de datos se encontraba influenciado por 5 registros afectando las medidas de MSE, MAE y R^2. Adicionalmente, cuando se tenga una mayor historia, se deben volver a proponer modelos no paramétricos que permitan capturar esas relaciones no lineales que se evidenciaron a lo largo del trabajo. Además, si es necesario, realizar el análisis aparte de los clientes con situaciones muy atípicas para que no afecten los resultados de la mayoría.

8. Capítulo 8

Anexos

Tabla 8-1: Variables pertenecientes al módulo macroeconómico

NOMBRE	TIPO VARIABLE	DESCRIPCIÓN
MEC1	Continua	Producto interno bruto
MEC2	Continua	IPC
MEC3	Continua	Tasa interbancaria
MEC4	Continua	Tasa desempleo
MEC5	Continua	TRM
MEC6	Continua	Exportaciones
MEC7	Continua	Importaciones
MEC8	Continua	Balance cuenta corriente
MEC9	Continua	Balance fiscal
MEC10	Continua	PIB un mes anterior
MEC11	Continua	PIB dos meses anteriores
MEC12	Continua	PIB tres meses anteriores
MEC13	Continua	PIB cuatro meses anteriores
MEC14	Continua	PIB cinco meses anteriores
MEC15	Continua	PIB seis meses anteriores
MEC16	Continua	IPC un mes anterior
MEC17	Continua	IPC dos meses anteriores
MEC18	Continua	IPC tres meses anteriores
MEC19	Continua	IPC cuatro meses anteriores
MEC20	Continua	IPC cinco meses anteriores
MEC21	Continua	IPC seis meses anteriores
MEC22	Continua	Tasa interbancaria un mes antes
MEC23	Continua	Tasa interbancaria dos meses antes

NOMBRE	TIPO VARIABLE	DESCRIPCIÓN
MEC24	Continua	Tasa interbancaria tres meses antes
MEC25	Continua	Tasa interbancaria cuatro meses antes
MEC26	Continua	Tasa interbancaria cinco meses antes
MEC27	Continua	Tasa interbancaria seis meses antes
MEC28	Continua	Tasa desempleo un mes antes
MEC29	Continua	Tasa desempleo dos meses antes
MEC30	Continua	Tasa desempleo tres meses antes
MEC31	Continua	Tasa desempleo cuatro meses antes
MEC32	Continua	Tasa desempleo cinco meses antes
MEC33	Continua	Tasa desempleo seis meses antes
MEC34	Continua	TRM un mes antes
MEC35	Continua	TRM dos meses antes
MEC36	Continua	TRM tres meses antes
MEC37	Continua	TRM cuatro meses antes
MEC38	Continua	TRM cinco meses antes
MEC39	Continua	TRM seis meses antes
MEC40	Continua	Exportaciones un mes antes
MEC41	Continua	Exportaciones dos meses antes
MEC42	Continua	Exportaciones tres meses antes
MEC43	Continua	Exportaciones cuatro meses antes
MEC44	Continua	Exportaciones cinco meses antes
MEC45	Continua	Exportaciones seis meses antes
MEC46	Continua	Importaciones un mes antes
MEC47	Continua	Importaciones dos meses antes
MEC48	Continua	Importaciones tres meses antes
MEC49	Continua	Importaciones cuatro meses antes
MEC50	Continua	Importaciones cinco meses antes
MEC51	Continua	Importaciones seis meses antes
MEC52	Continua	Balance cuenta corriente un mes antes
MEC53	Continua	Balance cuenta corriente dos meses antes
MEC54	Continua	Balance cuenta corriente tres meses antes
MEC55	Continua	Balance cuenta corriente cuatro meses antes
MEC56	Continua	Balance cuenta corriente cinco meses antes
MEC57	Continua	Balance cuenta corriente seis meses antes
MEC58	Continua	Balance cuenta fiscal un mes antes
MEC59	Continua	Balance cuenta fiscal dos meses antes
MEC60	Continua	Balance cuenta fiscal tres meses antes
MEC61	Continua	Balance cuenta fiscal cuatro meses antes
MEC62	Continua	Balance cuenta fiscal cinco meses antes
MEC63	Continua	Balance cuenta fiscal seis meses antes
MEC64	Continua	PIB pronosticado un mes después

NOMBRE	TIPO VARIABLE	DESCRIPCIÓN
MEC65	Continua	PIB pronosticado dos meses después
MEC66	Continua	PIB pronosticado tres meses después
MEC67	Continua	PIB pronosticado cuatro meses después
MEC68	Continua	PIB pronosticado cinco meses después
MEC69	Continua	PIB pronosticado seis meses después
MEC70	Continua	IPC pronosticado un mes después
MEC71	Continua	IPC pronosticado dos meses después
MEC72	Continua	IPC pronosticado tres meses después
MEC73	Continua	IPC pronosticado cuatro meses después
MEC74	Continua	IPC pronosticado cinco meses después
MEC75	Continua	IPC pronosticado seis meses después
MEC76	Continua	Tasa interbancaria pronosticado un mes después
MEC77	Continua	Tasa interbancaria pronosticado dos meses después
MEC78	Continua	Tasa interbancaria pronosticado tres meses después
MEC79	Continua	Tasa interbancaria pronosticado cuatro meses después
MEC80	Continua	Tasa interbancaria pronosticado cinco meses después
MEC81	Continua	Tasa interbancaria pronosticado seis meses después
MEC82	Continua	Tasa desempleo pronosticado un mes después
MEC83	Continua	Tasa desempleo pronosticado dos meses después
MEC84	Continua	Tasa desempleo pronosticado tres meses después
MEC85	Continua	Tasa desempleo pronosticado cuatro meses después
MEC86	Continua	Tasa desempleo pronosticado cinco meses después
MEC87	Continua	Tasa desempleo pronosticado seis meses después
MEC88	Continua	TRM pronosticado un mes después
MEC89	Continua	TRM pronosticado dos meses después
MEC90	Continua	TRM pronosticado tres meses después
MEC91	Continua	TRM pronosticado cuatro meses después
MEC92	Continua	TRM pronosticado cinco meses después
MEC93	Continua	TRM pronosticado seis meses después
MEC94	Continua	Exportaciones pronosticado un mes después
MEC95	Continua	Exportaciones pronosticado dos meses después
MEC96	Continua	Exportaciones pronosticado tres meses después
MEC97	Continua	Exportaciones pronosticado cuatro meses después
MEC98	Continua	Exportaciones pronosticado cinco meses después
MEC99	Continua	Exportaciones pronosticado seis meses después
MEC100	Continua	Importaciones pronosticado un mes después
MEC101	Continua	Importaciones pronosticado dos meses después
MEC102	Continua	Importaciones pronosticado tres meses después
MEC103	Continua	Importaciones pronosticado cuatro meses después
MEC104	Continua	Importaciones pronosticado cinco meses después

NOMBRE	TIPO VARIABLE	DESCRIPCIÓN
MEC105	Continua	Importaciones pronosticado seis meses después
MEC106	Continua	Balance cuenta corriente pronosticado un mes después
MEC107	Continua	Balance cuenta corriente pronosticado dos meses después
MEC108	Continua	Balance cuenta corriente pronosticado tres meses después
MEC109	Continua	Balance cuenta corriente pronosticado cuatro meses después
MEC110	Continua	Balance cuenta corriente pronosticado cinco meses después
MEC111	Continua	Balance cuenta corriente pronosticado seis meses después
MEC112	Continua	Balance cuenta fiscal pronosticado un mes después
MEC113	Continua	Balance cuenta fiscal pronosticado dos meses después
MEC114	Continua	Balance cuenta fiscal pronosticado tres meses después
MEC115	Continua	Balance cuenta fiscal pronosticado cuatro meses después
MEC116	Continua	Balance cuenta fiscal pronosticado cinco meses después
MEC117	Continua	Balance cuenta fiscal pronosticado seis meses después

Tabla 8-2: Variables pertenecientes al módulo de riesgo

NOMBRE	TIPO VARIABLE	DESCRIPCIÓN	Rango
RG1	Cualitativa ordinal	Calificación del cliente al momento del desembolso	La calificación es un valor entero que se da entre 1 a 11, siendo 1 un menor riesgo y 11 el mayor.
RG2	Cualitativa ordinal	Stage según IFRS9	El Stage es un valor entero que se da entre 1 a 3, siendo 1 un menor riesgo y 3 el mayor.
RG3	Continua	Meses de incumplimiento	No aplica
RG4	Continua	Días de incumplimiento	No aplica
RG5	Continua	Probabilidad de default	No aplica
RG6	Cualitativa ordinal	Calificación externa	La calificación externa es un valor entero que se da entre 1 a 5, siendo 1 un menor riesgo y 5 el mayor.
RG7	Continua	Exposición mes anterior	No aplica
RG8	Continua	Exposición dos meses antes	No aplica
RG9	Continua	Exposición tres meses antes	No aplica
RG10	Continua	Exposición seis meses antes	No aplica
RG11	Continua	Exposición doce meses antes	No aplica
RG12	Continua	Variación exposición mes anterior	No aplica
RG13	Continua	Variación exposición dos meses antes	No aplica
RG14	Continua	Variación exposición tres meses antes	No aplica

NOMBRE	TIPO VARIABLE	DESCRIPCIÓN	Rango
RG15	Continua	Variación exposición cuatro meses antes	No aplica
RG16	Continua	Variación gasto de provisión anterior	No aplica
RG17	Continua	Variación gasto de provisión dos meses antes	No aplica
RG18	Continua	Variación gasto de provisión tres meses antes	No aplica
RG19	Continua	Variación gasto de provisión seis meses antes	No aplica
RG20	Continua	Variación gasto de provisión doce meses antes	No aplica
RG21	Cualitativa ordinal	Calificación interna del cliente	La calificación es un valor entero que se da entre 1 a 11, siendo 1 un menor riesgo y 11 el mayor.
RG22	Cualitativa ordinal	Calificación interna del cliente un mes antes	La calificación es un valor entero que se da entre 1 a 11, siendo 1 un menor riesgo y 11 el mayor.
RG23	Cualitativa ordinal	Calificación interna del cliente dos meses antes	La calificación es un valor entero que se da entre 1 a 11, siendo 1 un menor riesgo y 11 el mayor.
RG24	Cualitativa ordinal	Calificación interna del cliente tres meses antes	La calificación es un valor entero que se da entre 1 a 11, siendo 1 un menor riesgo y 11 el mayor.
RG25	Cualitativa ordinal	Calificación interna del cliente seis meses antes	La calificación es un valor entero que se da entre 1 a 11, siendo 1 un menor riesgo y 11 el mayor.
RG26	Cualitativa ordinal	Calificación interna del cliente doce meses antes	La calificación es un valor entero que se da entre 1 a 11, siendo 1 un menor riesgo y 11 el mayor.
RG27	Continua	Número de días de mora	No aplica
RG28	Continua	Número de días de mora un mes antes	No aplica
RG29	Continua	Número de días de mora dos meses antes	No aplica
RG30	Continua	Número de días de mora tres meses antes	No aplica
RG31	Continua	Número de días de mora seis meses antes	No aplica

NOMBRE	TIPO VARIABLE	DESCRIPCIÓN	Rango
RG32	Continua	Número de días de mora doce meses antes	No aplica
RG33	Continua	Saldo de capital un mes antes	No aplica
RG34	Continua	Saldo de capital dos meses antes	No aplica
RG35	Continua	Saldo de capital tres meses antes	No aplica
RG36	Continua	Saldo de capital seis meses antes	No aplica
RG37	Continua	Saldo de capital doce meses antes	No aplica
RG38	Continua	Castigos sobre el capital	No aplica
RG39	Continua	Castigos sobre el capital un mes antes	No aplica
RG40	Continua	Castigos sobre el capital dos meses antes	No aplica
RG41	Continua	Castigos sobre el capital tres meses antes	No aplica
RG42	Continua	Castigos sobre el capital seis meses antes	No aplica
RG43	Continua	Castigos sobre el capital doce meses antes	No aplica
RG44	Continua	Cartera vencida	No aplica
RG45	Continua	Cartera vencida un mes antes	No aplica
RG46	Continua	Cartera vencida dos meses antes	No aplica
RG47	Continua	Cartera vencida tres meses antes	No aplica
RG48	Continua	Cartera vencida doce meses antes	No aplica
RG49	Continua	Índice de cartera vencida	No aplica
RG50	Continua	Índice de cartera vencida un mes antes	No aplica
RG51	Continua	Índice de cartera vencida dos meses antes	No aplica
RG52	Continua	Índice de cartera vencida tres meses antes	No aplica
RG53	Continua	Índice de cartera vencida seis meses antes	No aplica
RG54	Continua	Índice de cartera vencida doce meses antes	No aplica
RG55	Continua	Total cartera vencida mayor a 90 días	No aplica
RG56	Continua	Total cartera vencida mayor a 90 días un mes antes	No aplica
RG57	Continua	Total cartera vencida mayor a 90 días dos meses antes	No aplica

NOMBRE	TIPO VARIABLE	DESCRIPCIÓN	Rango
RG58	Continua	Total cartera vencida mayor a 90 días tres meses antes	No aplica
RG59	Continua	Total cartera vencida mayor a 90 días seis meses antes	No aplica
RG60	Continua	Total cartera vencida mayor a 90 días doce meses antes	No aplica
RG61	Continua	Índice de cartera vencida mayor a 90 días	No aplica
RG62	Continua	Índice de cartera vencida mayor a 90 días un mes antes	No aplica
RG63	Continua	Índice de cartera vencida mayor a 90 días dos meses antes	No aplica
RG64	Continua	Índice de cartera vencida mayor a 90 días tres meses antes	No aplica
RG65	Continua	Índice de cartera vencida mayor a 90 días seis meses antes	No aplica
RG66	Continua	Índice de cartera vencida mayor a 90 días doce meses antes	No aplica
RG67	Continua	Valor cubierto por garantías	No aplica
RG68	Continua	Valor cubierto por garantías un mes antes	No aplica
RG69	Continua	Valor cubierto por garantías dos meses antes	No aplica
RG70	Continua	Valor cubierto por garantías tres meses antes	No aplica
RG71	Continua	Valor cubierto por garantías seis meses antes	No aplica
RG72	Continua	Valor cubierto por garantías doce meses antes	No aplica
RG73	Continua	Exposición total que se encuentra en estado mayor o igual a Stage 2	No aplica
RG74	Continua	Exposición total que se encuentra en estado mayor o igual a Stage 2 un mes antes	No aplica
RG75	Continua	Exposición total que se encuentra en estado mayor o igual a Stage 2 dos meses antes	No aplica
RG76	Continua	Exposición total que se encuentra en estado mayor o igual a Stage 2 tres meses antes	No aplica

NOMBRE	TIPO VARIABLE	DESCRIPCIÓN	Rango
RG77	Continua	Exposición total que se encuentra en estado mayor o igual a Stage 2 seis meses antes	No aplica
RG78	Continua	Exposición total que se encuentra en estado mayor o igual a Stage 2 doce meses antes	No aplica
RG79	Continua	Variación de exposición que se encuentra en estado mayor o igual a Stage 2	No aplica
RG80	Continua	Variación de exposición que se encuentra en estado mayor o igual a Stage 2 un mes antes	No aplica
RG81	Continua	Variación de exposición que se encuentra en estado mayor o igual a Stage 2 dos meses antes	No aplica
RG82	Continua	Variación de exposición que se encuentra en estado mayor o igual a Stage 2 tres meses antes	No aplica
RG83	Continua	Variación de exposición que se encuentra en estado mayor o igual a Stage 2 seis meses antes	No aplica
RG84	Continua	Variación de exposición que se encuentra en estado mayor o igual a Stage 2 doce meses antes	No aplica
RG85	Continua	Tasa de interés promedio ponderada	No aplica
RG86	Continua	Tasa de descuento IFSR9	No aplica
RG87	Continua	Tasa de descuento norma local	No aplica
RG88	Continua	Tasa de descuento norma local un mes antes	No aplica
RG89	Continua	Tasa de descuento norma local dos meses antes	No aplica
RG90	Continua	Tasa de descuento norma local tres meses antes	No aplica
RG91	Continua	Vida media promedio ponderada de las obligaciones	No aplica
RG92	Continua	Vida media promedio ponderada de las obligaciones un mes antes	No aplica
RG93	Continua	Vida media promedio ponderada de las obligaciones dos meses antes	No aplica

NOMBRE	TIPO VARIABLE	DESCRIPCIÓN	Rango
RG94	Continua	Vida media promedio ponderada de las obligaciones tres meses antes	No aplica
RG95	Continua	Valor esperado de default en 12 meses	No aplica
RG96	Continua	LGD a 12 meses	No aplica
RG97	Continua	EAD a 12 meses	No aplica
RG98	Cualitativa nominal	Indica si sus obligaciones tienen cura o no	No aplica
RG99	Cualitativa nominal	Indica si tiene obligaciones en Stage 2 o Stage 3	No aplica
RG100	Cualitativa nominal	Indica si ha tenido default	No aplica
RG101	Cualitativa nominal	Indica si tiene obligaciones reestructuradas	No aplica
RG102	Cualitativa ordinal	Nivel de riesgo percibido	El nivel de riesgo es un valor entero que se da entre 0 a 4, siendo 0 un menor riesgo percibido muy bajo y 4 implica materialización del riesgo.
RG103	Cualitativa ordinal	Nivel de riesgo percibido un mes anterior	El nivel de riesgo es un valor entero que se da entre 0 a 4, siendo 0 un menor riesgo percibido muy bajo y 4 implica materialización del riesgo.
RG104	Cualitativa nominal	Indica si no hay un nivel de riesgo materializable, aunque antes lo tenía	No aplica
RG105	Cualitativa nominal	Indica si hay un nivel de riesgo materializable, aunque antes no lo tenía	No aplica
RG106	Cualitativa ordinal	Variación en el nivel de riesgo	Variación del nivel de riesgo percibido. Un mayor valor, implica un mayor riesgo percibido en la fecha de análisis

Tabla 8-3: Variables pertenecientes al módulo financiero

NOMBRE	TIPO VARIABLE	DESCRIPCIÓN
FIN1	Continua	Margen EBITDA
FIN2	Continua	Margen EBITDA tres meses antes
FIN3	Continua	Margen EBITDA seis meses antes
FIN4	Continua	Margen EBITDA doce meses antes

NOMBRE	TIPO VARIABLE	DESCRIPCIÓN
FIN5	Continua	Margen EBITDA veinticuatro meses antes
FIN6	Continua	Promedio Margen EBITDA de los últimos doce meses
FIN7	Continua	Promedio Margen EBITDA de los últimos veinticuatro meses
FIN8	Continua	Variación de los ingresos operacionales
FIN9	Continua	Variación de los ingresos operacionales tres meses antes
FIN10	Continua	Variación de los ingresos operacionales seis meses antes
FIN11	Continua	Variación de los ingresos operacionales doce meses antes
FIN12	Continua	Variación de los ingresos operacionales veinticuatro meses antes
FIN13	Continua	Promedio de la variación de las ventas doce meses
FIN14	Continua	Promedio de la variación de las ventas veinticuatro meses
FIN15	Continua	KTNO
FIN16	Continua	KTNO tres meses antes
FIN17	Continua	KTNO seis meses antes
FIN18	Continua	KTNO doce meses antes
FIN19	Continua	KTNO veinticuatro meses antes
FIN20	Continua	Promedio KTNO doce meses antes
FIN21	Continua	Promedio KTNO veinticuatro meses antes
FIN22	Continua	Diferencia entre pasivos y activos
FIN23	Continua	Diferencia entre pasivos y activos tres meses antes
FIN24	Continua	Diferencia entre pasivos y activos seis meses antes
FIN25	Continua	Diferencia entre pasivos y activos doce meses antes
FIN26	Continua	Diferencia entre pasivos y activos veinticuatro meses antes
FIN27	Continua	Promedio de la diferencia entre pasivos y activos de los últimos doce meses
FIN28	Continua	Promedio de la diferencia entre pasivos y activos de los últimos veinticuatro meses
FIN29	Continua	EBITDA sobre intereses
FIN30	Continua	EBITDA sobre intereses tres meses antes
FIN31	Continua	EBITDA sobre intereses seis meses antes
FIN32	Continua	EBITDA sobre intereses doce meses antes
FIN33	Continua	EBITDA sobre intereses veinticuatro meses antes
FIN34	Continua	Promedio EBITDA sobre intereses doce meses
FIN35	Continua	Promedio EBITDA sobre intereses veinticuatro meses
FIN36	Continua	Deuda financiera sobre EBITDA
FIN37	Continua	Deuda financiera sobre EBITDA tres meses antes
FIN38	Continua	Deuda financiera sobre EBITDA seis meses antes
FIN39	Continua	Deuda financiera sobre EBITDA doce meses antes
FIN40	Continua	Deuda financiera sobre EBITDA veinticuatro meses antes
FIN41	Continua	Promedio deuda financiera sobre EBITDA de los últimos doce meses

NOMBRE	TIPO VARIABLE	DESCRIPCIÓN
FIN42	Continua	Promedio deuda financiera sobre EBITDA de los últimos veinticuatro meses
FIN43	Continua	PKTNO
FIN44	Continua	PKTNO tres meses antes
FIN45	Continua	PKTNO seis meses antes
FIN46	Continua	PKTNO doce meses antes
FIN47	Continua	PKTNO veinticuatro meses antes
FIN48	Continua	Promedio PKTNO doce meses
FIN49	Continua	Promedio PKTNO veinticuatro meses
FIN50	Continua	Margen neto
FIN51	Continua	Margen neto tres meses antes
FIN52	Continua	Margen neto seis meses antes
FIN53	Continua	Margen neto doce meses antes
FIN54	Continua	Margen neto veinticuatro meses antes
FIN55	Continua	Promedio margen neto de los últimos doce meses
FIN56	Continua	Promedio margen neto de los últimos veinticuatro meses
FIN57	Continua	ROE
FIN58	Continua	ROE tres meses antes
FIN59	Continua	ROE seis meses antes
FIN60	Continua	ROE doce meses antes
FIN61	Continua	ROE veinticuatro meses antes
FIN62	Continua	Promedio ROE de los últimos doce meses
FIN63	Continua	Promedio ROE de los últimos veinticuatro meses
FIN64	Continua	ROA
FIN65	Continua	ROA tres meses antes
FIN66	Continua	ROA seis meses antes
FIN67	Continua	ROA doce meses antes
FIN68	Continua	ROA veinticuatro meses antes
FIN69	Continua	Promedio ROA de los últimos doce meses
FIN70	Continua	Promedio ROA de los últimos veinticuatro meses
FIN71	Continua	Margen bruto
FIN72	Continua	Margen bruto tres meses antes
FIN73	Continua	Margen bruto seis meses antes
FIN74	Continua	Margen bruto doce meses antes
FIN75	Continua	Margen bruto veinticuatro meses antes
FIN76	Continua	Promedio margen bruto doce meses
FIN77	Continua	Promedio margen bruto veinticuatro meses
FIN78	Continua	Pasivos sobre activos
FIN79	Continua	Pasivos sobre activos tres meses antes
FIN80	Continua	Pasivos sobre activos seis meses antes

NOMBRE	TIPO VARIABLE	DESCRIPCIÓN
FIN81	Continua	Pasivos sobre activos doce meses antes
FIN82	Continua	Pasivos sobre activos veinticuatro meses antes
FIN83	Continua	Promedio pasivos sobre activos últimos doce meses
FIN84	Continua	Promedio pasivos sobre activos últimos veinticuatro meses
FIN85	Continua	Apalancamiento
FIN86	Continua	Apalancamiento tres meses antes
FIN87	Continua	Apalancamiento seis meses antes
FIN88	Continua	Apalancamiento doce meses antes
FIN89	Continua	Apalancamiento veinticuatro meses antes
FIN90	Continua	Promedio apalancamiento últimos doce meses
FIN91	Continua	Promedio apalancamiento últimos veinticuatro meses
FIN92	Continua	Número de días que tarda la empresa en recuperar su cartera.
FIN93	Continua	Diferencia entre los dos últimos estados financieros tomando la diferencia del número de días que tarda la empresa en recuperar su cartera.
FIN94	Continua	Número de días que tarda la empresa en realizar el pago a sus proveedores.
FIN95	Continua	Diferencia entre los dos últimos estados financieros tomando la diferencia del número de días que tarda la empresa en realizar el pago a sus proveedores.
FIN96	Continua	Número de días que tarda la empresa en vender su inventario.
FIN97	Continua	Diferencia entre los dos últimos estados financieros tomando la diferencia del número de días que tarda la empresa en vender su inventario.
FIN98	Continua	Diferencia entre los dos últimos estados financieros del PKTNO
FIN99	Continua	KTNO sobre ventas
FIN100	Continua	Ciclo financiero
FIN101	Continua	Diferencia entre los dos últimos estados financieros tomando la diferencia del ciclo financiero
FIN102	Continua	Diferencia entre los dos últimos estados financieros del KTNO
FIN103	Continua	Diferencia entre los dos últimos estados financieros del KTNO sobre ventas

NOMBRE	TIPO VARIABLE	DESCRIPCIÓN
FIN104	Continua	Activos fijos netos (entendido como propiedad, planta y equipo descontando la depreciación) sobre ventas
FIN105	Continua	Activos corrientes sobre pasivos corrientes
FIN106	Continua	Diferencia entre los dos últimos estados financieros de activos corrientes sobre pasivos corrientes
FIN107	Continua	Activos sobre deuda financiera
FIN108	Continua	Diferencia entre los dos últimos estados financieros de activos sobre deuda financiera
FIN109	Continua	Inventarios sobre deuda financiera
FIN110	Continua	Diferencia entre los dos últimos estados financieros de los inventarios sobre la deuda financiera
FIN111	Continua	Utilidad operacional sobre gasto financiero
FIN112	Continua	Diferencia entre los dos últimos estados financieros de la utilidad operacional sobre el gasto financiero
FIN113	Continua	Activo corriente sobre deuda financiera
FIN114	Continua	Diferencia entre los dos últimos estados financieros del margen EBITDA sobre intereses
FIN115	Continua	Cuentas por cobrar sobre deuda financiera
FIN116	Continua	Diferencia entre los dos últimos estados financieros de las cuentas por cobrar sobre deuda financiera
FIN117	Continua	Efectivo sobre deuda financiera
FIN118	Continua	Diferencia entre los dos últimos estados financieros del efectivo sobre deuda financiera
FIN119	Continua	Gasto financiero (Intereses) sobre deuda financiera

NOMBRE	TIPO VARIABLE	DESCRIPCIÓN
FIN120	Continua	Diferencia entre los dos últimos estados financieros del gasto financiero (Intereses) sobre deuda financiera
FIN121	Continua	Diferencia entre los dos últimos estados financieros del pasivo corriente sobre los activos
FIN122	Continua	Deuda financiera sobre el patrimonio
FIN123	Continua	Diferencia entre los dos últimos estados financieros de la deuda financiera sobre el patrimonio
FIN124	Continua	Diferencia entre los dos últimos estados financieros de los pasivos sobre el patrimonio
FIN125	Continua	Deuda financiera corto plazo sobre activos
FIN126	Continua	Diferencia entre los dos últimos estados financieros de la deuda financiera de corto plazo sobre activos
FIN127	Continua	Deuda financiera largo plazo sobre activos
FIN128	Continua	Diferencia entre los dos últimos estados financieros de la deuda financiera de largo plazo sobre activos
FIN129	Continua	Deuda financiera sobre ventas
FIN130	Continua	Diferencia entre los dos últimos estados financieros de la deuda financiera sobre ventas
FIN131	Continua	Diferencia entre los dos últimos estados financieros de la deuda financiera sobre EBITDA
FIN132	Continua	Cuentas por pagar vinculados económicos sobre activos
FIN133	Continua	Diferencia entre los dos últimos estados financieros de las cuentas por pagar vinculados económicos sobre activos

NOMBRE	TIPO VARIABLE	DESCRIPCIÓN
FIN134	Continua	Ventas anuales considerando el deflactor del PIB
FIN135	Continua	Diferencia entre los dos últimos estados financieros del efectivo
FIN136	Continua	Capital suscrito sobre patrimonio
FIN137	Continua	Diferencia entre los dos últimos estados financieros del capital suscrito sobre patrimonio
FIN138	Continua	Diferencia entre los dos últimos estados financieros del patrimonio
FIN139	Continua	Activos sobre patrimonio
FIN140	Continua	Capital suscrito sobre activos
FIN141	Continua	Diferencia entre los dos últimos estados financieros del capital suscrito sobre activos
FIN142	Continua	Diferencia entre los dos últimos estados financieros de las utilidades retenidas
FIN143	Continua	Diferencia entre los dos últimos estados financieros de la utilidad operacional
FIN144	Continua	Diferencia entre los dos últimos estados financieros de la utilidad EBITDA
FIN145	Continua	Gastos de administración y ventas sobre ventas
FIN146	Continua	Diferencia entre los dos últimos estados financieros de los gastos de administración y ventas sobre ventas
FIN147	Continua	Diferencia entre los dos últimos estados financieros del margen bruto
FIN148	Continua	Diferencia entre los dos últimos estados financieros del margen EBITDA

NOMBRE	TIPO VARIABLE	DESCRIPCIÓN
FIN149	Continua	Diferencia entre los dos últimos estados financieros del margen neto
FIN150	Continua	Diferencia entre los dos últimos estados financieros del ROA
FIN151	Continua	Diferencia entre los dos últimos estados financieros del ROE
FIN152	Continua	Utilidad anterior sobre patrimonio
FIN153	Continua	Diferencia entre los dos últimos estados financieros de la utilidad anterior sobre patrimonio
FIN154	Continua	Utilidad bruta descontando gastos de administración sobre el pasivo

Tabla 8-4: Variables pertenecientes al módulo transaccional

NOMBRE	TIPO VARIABLE	DESCRIPCIÓN
TRX1	Continua	Promedio 12 meses de las entradas transaccionales sobre las ventas
TRX2	Continua	Promedio 12 meses de las entradas menos salidas transaccionales sobre el gasto financiero
TRX3	Continua	Promedio 6 meses del saldo de cartera sobre entradas transaccionales
TRX4	Continua	Diferencia del promedio de 6 y 12 meses del saldo de cartera sobre las entradas transaccionales
TRX5	Continua	Promedio 6 meses de pagos de nómina sobre ventas.
TRX6	Continua	Promedio 12 meses de la suma de cuentas de ahorro y corrientes sobre el total de la deuda financiera.
TRX7	Continua	Diferencia del promedio de 6 y 12 meses de la suma de cuentas de ahorro y corrientes sobre el total de la deuda financiera.
TRX8	Continua	Promedio 12 meses de los saldos de cuenta de ahorro y corriente sobre el saldo de cartera
TRX9	Continua	Promedio 6 meses de los saldos de cuenta de ahorro y corriente sobre el saldo de cartera
TRX10	Continua	Diferencia del promedio de 6 y 12 meses de la suma de cuentas de ahorro y corrientes sobre el total de cartera
TRX11	Continua	Promedio 12 meses de sobregiro sobre deuda financiera
TRX12	Continua	Diferencia del promedio de 6 y 12 meses de sobregiro sobre deuda financiera
TRX13	Continua	Diferencia entre el promedio de 6 y 12 meses de las entradas transaccionales

NOMBRE	TIPO VARIABLE	DESCRIPCIÓN
TRX14	Continua	Promedio 12 meses de las entradas transaccionales sobre el gasto financiero
TRX15	Continua	Diferencia del promedio de 6 y 12 meses de las entradas menos las salidas transaccionales
TRX16	Continua	Promedio 12 meses de las entradas menos las salidas transaccionales
TRX17	Continua	Promedio 6 meses de las entradas menos las salidas transaccionales
TRX18	Continua	Diferencia del promedio de 6 y 12 meses del total de pagos de nómina
TRX19	Continua	Diferencia del promedio de 6 y 12 meses del total de pagos a proveedores
TRX20	Continua	Diferencia del promedio de 6 y 12 meses del total de salidas transaccionales
TRX21	Continua	Variación del promedio de los últimos 6 y 12 meses del valor total de salidas transaccionales
TRX22	Continua	Promedio 12 meses de las salidas transaccionales sobre las ventas
TRX23	Continua	Promedio 12 meses del saldo de cartera sobre el monto de entradas transaccionales
TRX24	Continua	Promedio 12 meses del saldo de cartera sobre las entradas transaccionales descontando las salidas transaccionales
TRX25	Continua	Promedio 6 meses del saldo de cartera sobre las entradas transaccionales descontando las salidas transaccionales
TRX26	Continua	Diferencia del promedio 6 y 12 meses de la deuda financiera sobre las entradas transaccionales descontando las salidas transaccionales
TRX27	Continua	Diferencia del promedio 6 y 12 meses del conteo de pagos realizados a proveedores
TRX28	Continua	Diferencia 6 y 12 meses del conteo de pagos realizados por seguridad social
TRX29	Continua	Diferencia 6 y 12 meses del valor del monto de los pagos realizados a la seguridad social

Tabla 8-5: Variables pertenecientes al módulo descriptivo y sectorial

NOMBRE	TIPO VARIABLE	DESCRIPCIÓN
DyS1	Continua	Número de meses desde la vinculación
DyS2	Cualitativa nominal	Cliente gerenciado
DyS3	Cualitativa nominal	Cliente sector agropecuario
DyS4	Cualitativa nominal	Cliente sector comercio
DyS5	Cualitativa nominal	Cliente sector edificaciones
DyS6	Cualitativa nominal	Cliente sector infraestructura
DyS7	Cualitativa nominal	Cliente sector manufactura
DyS8	Cualitativa nominal	Cliente sector servicios

NOMBRE	TIPO VARIABLE	DESCRIPCIÓN
DyS9	Cualitativa nominal	Cliente sector recursos naturales
DyS10	Cualitativa nominal	Cliente servicios sector público
DyS11	Cualitativa nominal	Cliente servicios sector personas
DyS12	Cualitativa nominal	Cliente servicios sector comunicaciones
DyS13	Cualitativa nominal	Cliente servicios sector hoteles
DyS14	Cualitativa nominal	Cliente servicios sector social
DyS15	Cualitativa nominal	Cliente servicios sector empresas
DyS16	Cualitativa nominal	Cliente servicios sector financiero
DyS17	Cualitativa nominal	Cliente servicios sector transporte
DyS18	Cualitativa nominal	Cliente servicios sector no financiero
DyS19	Cualitativa nominal	Cliente servicios sector turismo
DyS20	Cualitativa nominal	Cliente servicios sector social financiero
DyS21	Cualitativa nominal	Cliente servicios sector inmobiliario
DyS22	Cualitativa nominal	Cliente servicios sector banca
DyS23	Cualitativa nominal	Cliente servicios sector transporte público
DyS24	Cualitativa nominal	Cliente servicios sector social otros
DyS25	Cualitativa nominal	Cliente servicios sector no financieros otros
DyS26	Cualitativa nominal	Cliente servicios sector financiero otros
DyS27	Cualitativa nominal	Cliente servicios sector transporte otros
DyS28	Cualitativa nominal	Cliente servicios agrícola
DyS29	Cualitativa nominal	Cliente sector pecuario
DyS30	Cualitativa nominal	Cliente servicios agrícola otros
DyS31	Cualitativa nominal	Cliente sector pecuario otros

Tabla 8-6: Variables que tuvieron una frecuencia de 3 o 4 en los métodos de reducción de dimensionalidad y/o que se encuentran dentro del proceso para hallar la ECL.

Variable	Frecuencia Modelos	Ridge	Forest	Extratree	Lasso	Variables proceso
RG10	4	X	X	X	X	
FIN76	4	X	X	X	X	
FIN125	4	X	X	X	X	
FIN70	4	X	X	X	X	
RG16	4	X	X	X	X	X
FIN109	4	X	X	X	X	X
RG13	4	X	X	X	X	X
FIN95	3	X	X		X	
FIN123	3	X	X		X	
FIN105	3	X	X	X		
FIN39	3	X	X		X	
FIN130	3	X	X		X	
FIN93	3	X	X	X		X
FIN92	3	X	X	X		X

Variable	Frecuencia Modelos	Ridge	Forest	Extratree	Lasso	Variables proceso
FIN61	3		X	X	X	X
FIN91	3	X		X	X	
FIN78	3		X	X	X	
RG23	3	X	X	X		
FIN75	3	X	X	X		
FIN4	3	X	X	X		
RG15	3	X	X		X	
RG88	3	X	X	X		X
RG100	2		X	X		X
RG98	2		X	X		X
FIN16	2		X	X		X
FIN53	2		X	X		X
FIN46	2		X	X		X
FIN22	2		X	X		X
RG1	2		X	X		X
FIN88	2		X	X		X
FIN140	2		X	X		X
TRX1	2		X	X		X
TRX14	2		X	X		X
TRX17	2		X	X		X
TRX19	2		X	X		X
TRX24	2		X	X		X
FIN127	2		X	X		X
FIN81	2		X	X		X
FIN84	2		X	X		X
FIN85	2		X	X		X
FIN120	2		X	X		X
FIN57	2		X	X		X
RG6	2		X	X		X
RG18	2		X	X		X
RG20	2		X	X		X
RG21	2		X	X		X
RG25	2		X	X		X
RG26	2		X	X		X
RG91	2		X	X		X
RG89	2		X	X		X
DyS2	1			X		X
RG19	1		X			X
RG17	1		X			X
RG101	1			X		X
FIN134	1			X		X
FIN68	1			X		X

Variable	Frecuencia Modelos	Ridge	Forest	Extratree	Lasso	Variables proceso
RG93	1			X		X
RG5	1			X		X

Bibliografía

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, *23*(4), 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Antonsson, H. (2018). *Macroeconomic factors in Probability of Default A study applied to a Swedish credit portfolio* [KTH Royal Institute of Technology]. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1264976&dswid=3197

Apergis, E., Apergis, I., & Apergis, N. (2019). A new macro stress testing approach for financial realignment in the Eurozone. *Journal of International Financial Markets, Institutions and Money*, *61*(4), 52–80. https://doi.org/10.1016/j.intfin.2019.02.002

Banco de la República de Colombia. (2022). *Sectorización Monetaria y Económica*. https://www.banrep.gov.co/sites/default/files/paginas/sectormon.pdf

Bandyopadhyay, A. (2022). Loan level loss given default (LGD) study of Indian banks. *IIMB Management Review*, *34*(2), 168–177. https://doi.org/10.1016/J.IIMB.2022.06.003

Banerjee, R., & Venkateshwaran, S. (2017, July). *Demystifying Expected Credit Loss (ECL)*. KPMG. https://assets.kpmg/content/dam/kpmg/in/pdf/2017/07/Demystifying-Expected-Credit-Loss.pdf

Bastos, J. A. (2010). Forecasting bank loans loss-given-default. *Journal of Banking & Finance*, *34*(10), 2510–2517. https://doi.org/10.1016/J.JBANKFIN.2010.04.011

BCBS. (2000). Principles for the Management of Credit Risk. *Basel Committee on Banking Supervision*. https://www.bis.org/publ/bcbs75.htm

Bemister-Buffington, J., Wolf, A. J., Raschka, S., & Kuhn, L. A. (2020). Machine Learning to Identify Flexibility Signatures of Class A GPCR Inhibition. *Biomolecules*, *10*(3). https://doi.org/10.3390/biom10030454

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cheng, J., Sun, J., Yao, K., Xu, M., & Cao, Y. (2022). A variable selection method based on mutual information and variance inflation factor. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *268*(6), 1–7. https://doi.org/https://doi.org/10.1016/j.saa.2021.120652

Chen, J. (2022, September 6). *Default: What It Means, What Happens When You Default, Examples*. Investopedia. https://www.investopedia.com/terms/d/default2.asp

Chen, M. (2011). Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches. *Computers & Mathematics with Applications*, *62*(12), 4514–4524. https://doi.org/10.1016/J.CAMWA.2011.10.030

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Dupré la Tour, T., Eickenberg, M., Nunez-Elizalde, A. O., & Gallant, J. L. (2022). Feature-space selection with banded ridge regression. *NeuroImage*, *264*(19), 1–19. https://doi.org/10.1016/J.NEUROIMAGE.2022.119728

ECB. (2023, June 30). *What are haircuts?* European Central Bank. https://www.ecb.europa.eu/ecb/educational/explainers/tell-me-more/html/haircuts.en.html

Fernando, J. (2023a, March 27). *Inventory Turnover Ratio: What It Is, How It Works, and Formula*. Investopedia. https://www.investopedia.com/terms/i/inventoryturnover.asp

Fernando, J. (2023b, May 24). *Return on Equity (ROE) Calculation and What It Means*. Investopedia. https://www.investopedia.com/terms/r/returnonequity.asp

Filippo, M., Alfonso, N., Theodore, P., Enrico, R., & Gerhard, S. (2017). IFRS 9: A silent revolution in banks' business models. *McKinsey*. https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/ifrs-9-a-silent-revolution-in-banks-business-models

Fischler, M. A., & Bolles, R. C. (1981). Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM*, *24*(6), 381–395. https://doi.org/10.1145/358669.358692

Fonti, V., & Belitser, E. (2017). Feature selection using lasso. In *VU Amsterdam research paper in business analytics*. Vrije Universitetit Amsterdam.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189 – 1232. https://doi.org/10.1214/aos/1013203451

Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236. https://doi.org/10.1016/J.PATREC.2010.03.014

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Giesecke, K., Longstaff, F. A., Schaefer, S., & Strebulaev, I. (2011). Corporate bond default risk: A 150-year perspective. *Journal of Financial Economics*, *102*(2), 233–250. https://doi.org/10.1016/J.JFINECO.2011.01.011

Giraud, C. (2021). *Introduction to High-Dimensional Statistics* (CRC Press, Ed.; 2nd, illustrated ed.). https://www.imo.universite-paris-saclay.fr/~christophe.giraud/Orsay/Bookv3.pdf

Gitman, L. J., & Zutter, C. J. (2012). *Principios de Administración financiera* (12th ed.). Pearson Educación de México, S.A. de C.V. https://economicas.unsa.edu.ar/afinan/informacion_general/book/pcipios-adm-finan-12edi-gitman.pdf

Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, *83*(2), 83–90. https://doi.org/10.1016/J.CHEMOLAB.2006.01.007

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, *46*(1), 389–422. https://doi.org/10.1023/A:1012487302797

Härdle, W. K., & Prastyo, D. D. (2013). Default Risk Calculation based on Predictor Selection for the Southeast Asian Industry. *SSRN Electronic Journal*, *SFB 649*(Discussion Paper 2013-037), 1–24. https://doi.org/10.2139/ssrn.2892650

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning* (2nd ed.). Springer New York Inc.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*. https://doi.org/10.1007/978-0-387-84858-7

Hayes, A. (2022, August 10). *EBITDA: Meaning, Formula, and History*. Investopedia. https://www.investopedia.com/terms/e/ebitda.asp

Heo, J., & Yang, J. Y. (2014). AdaBoost based bankruptcy forecasting of Korean construction companies. *Applied Soft Computing*, *24*(13), 494–499. https://doi.org/10.1016/J.ASOC.2014.08.009

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Jiménez, G., & Mencía, J. (2009). Modelling the distribution of credit losses with observable and latent factors. *Journal of Empirical Finance*, *16*(2), 235–253. https://doi.org/10.1016/j.jempfin.2008.10.003

Kendall, M. G. (1948). Rank correlation methods. In *Rank correlation methods*. Griffin & Co. https://doi.org/10.1017/S0020268100013019

Khamis, H. (2008). Measures of Association: How to Choose? *Journal of Diagnostic Medical Sonography*, *24*(3), 155–162. https://doi.org/10.1177/8756479308317006

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767–2787. https://doi.org/10.1016/J.JBANKFIN.2010.06.001

Leow, M., & Mues, C. (2012). Predicting loss given default (LGD) for residential mortgage loans: A two-stage model and empirical evidence for UK bank data. *International Journal of Forecasting*, *28*(1), 183–195. https://doi.org/10.1016/J.IJFORECAST.2011.01.010

Liu, J., & Xu, X. E. (2003). The Predictive Power of Economic Indicators in Consumer Credit Risk Management. *The Rma Journal*, *86*, 40–45. https://acortar.link/7x2mfu

Loh, W.-Y. (2011). Classification and Regression Trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 14–23. https://doi.org/10.1002/widm.8

Louppe, G. (2014). *Understanding Random Forests: From Theory to Practice* [Université de Liège]. https://doi.org/10.48550/arXiv.1407.7502

Lundberg, S., Erion, G., & Lee, S.-I. (2018). Consistent Individualized Feature Attribution for Tree Ensembles. *Arxiv*. https://doi.org/10.48550/arXiv.1802.03888

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. https://www.researchgate.net/publication/317062430

Luong, T. M., & Scheule, H. (2022). Benchmarking forecast approaches for mortgage credit risk for forward periods. *European Journal of Operational Research*, *299*(2), 750–767. https://doi.org/10.1016/j.ejor.2021.09.026

Maulud, D., & Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, *1*(4), 140–147. https://doi.org/10.38094/jastt1457

Mazibaş, M., & Tuna, Y. (2017). Understanding the Recent Growth in Consumer Loans and Credit Cards in Emerging Markets: Evidence from Turkey. *Emerging Markets Finance and Trade*, *53*(10), 2333–2346. https://doi.org/10.1080/1540496X.2016.1196895

Melkumova, L. E., & Shatskikh, S. Y. (2017). Comparing Ridge and LASSO estimators for data analysis. *Procedia Engineering*, *201*(31), 746–755. https://doi.org/10.1016/J.PROENG.2017.09.615

Meng, Y., Yang, N., Qian, Z., & Zhang, G. (2021). What Makes an Online Review More Helpful: An Interpretation Framework Using XGBoost and SHAP Values. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*(3), 466–490. https://doi.org/10.3390/jtaer16030029

Nalluri, M., Pentela, M., & Eluri, N. R. (2020). A Scalable Tree Boosting System: XG Boost. *Int. J. Res. Stud. Sci. Eng. Technol*, *7*(12), 36–51. https://doi.org/10.22259/2349-476X.0712005

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. http://scikit-learn.sourceforge.net.

Pereira, J. M., Basto, M., & Silva, A. F. da. (2016). The Logistic Lasso and Ridge Regression in Predicting Corporate Failure. *Procedia Economics and Finance*, *39*(5), 634–641. https://doi.org/10.1016/S2212-5671(16)30310-0

Peterdy, K. (2023, June 14). *Credit Risk*. CFI. https://corporatefinanceinstitute.com/resources/knowledge/finance/credit-risk/

Robles, C. L. (2012). *Fundamentos de administración financiera* (M. E. Buendía, Ed.; 1st ed.). Red Tercer Milenio S.C. http://biblioteca.udgvirtual.udg.mx/jspui/handle/123456789/3175

Rory, M., Andrey, A., Thejaswi, R., & Eibe, F. (2018). XGBoost: Scalable GPU Accelerated Learning. *Cornell University*. https://doi.org/10.48550/arXiv.1806.11248

Ross, S. A., Westerfiel, R. W., & Jordan, B. D. (2010). *Fundamentos de finanzas corporativas* (9th ed.). McGraw-Hill/IInteramericana Editores, S.A.de C.V. https://www.mheducation.com.co/fundamentos-de-finanzas-corporativas-9781456291136-col-group

Rubaszek, M., & Serwa, D. (2014). Determinants of credit to households: An approach using the life-cycle model. *Economic Systems*, *38*(4), 572–587. https://doi.org/10.1016/J.ECOSYS.2014.05.004

Stoppiglia, H., Dreyfus, G., Dubois, R., & Oussar, Y. (2003). Ranking a Random Feature For Variable And Feature Selection. *Journal of Machine Learning Research*, *3*, 1399–1414. https://doi.org/10.1162/153244303322753733

Taghiyeh, S., Lengacher, D. C., & Handfield, R. B. (2021). Loss rate forecasting framework based on macroeconomic changes: Application to US credit card industry. *Expert Systems with Applications*, *165*(3), 113954. https://doi.org/10.1016/J.ESWA.2020.113954

Temim, J. (2016, November). The IFRS 9 Impairment Model and its Interaction with the Basel Framework. *Moody's Analytics. Risk Perspectives*. https://acortar.link/3HQHP5

Theil, H. (1949). A Rank-Invariant Method of Linear and Polynomial Regression Analysis. In *Advanced Studies in Theoretical and Applied Econometrics* (Vol. 23). Springer, Dordrecht. https://doi.org/10.1007/978-94-011-2546-8_20

Vipond Tim. (2022, June). *Net Working Capital*. CFI. https://corporatefinanceinstitute.com/resources/knowledge/finance/what-is-net-working-capital/

Wang, X., Wang, X., Ma, B., Li, Q., Wang, C., & Shi, Y. (2023). High-performance reversible data hiding based on ridge regression prediction algorithm. *Signal Processing*, *204*(3), 108818. https://doi.org/10.1016/J.SIGPRO.2022.108818

Wang, Y. (2011). *Corporate Default Prediction: Models, Drivers and Measurements* [Doctoral thesis, University of Exeter]. http://hdl.handle.net/10036/3457

Xia, Y., Zhao, J., He, L., Li, Y., & Yang, X. (2021). Forecasting loss given default for peer-to-peer loans via heterogeneous stacking ensemble approach.

International Journal of Forecasting, *37*(4), 1590–1613. https://doi.org/10.1016/J.IJFORECAST.2021.03.002

Yeh, C. C., Chi, D. J., & Lin, Y. R. (2014). Going-concern prediction using hybrid random forests and rough set approach. *Information Sciences*, *254*(1), 98–110. https://doi.org/10.1016/J.INS.2013.07.011

Zhang, Y., & Chen, L. (2021). A Study on Forecasting the Default Risk of Bond Based on XGboost Algorithm and Over-Sampling Method. *Theoretical Economics Letters*, *11*(2), 258–267. https://doi.org/10.4236/tel.2021.112019