UNIVERSIDAD NEBRIJA

MÁSTER DE MACHINE LEARNING CON R

Interpretación local en modelos de Machine Learning para la estimación del incumplimiento crediticio bajo la regulación financiera chilena

AUTOR: Macarena del Pilar Pino Montoya

DIRECTOR: Ignacio García Vicente

FECHA: 12 de diciembre de 2021

ENTIDAD COLABORADORA:

MAXIMA formación

RESUMEN

Este trabajo de investigación se realizó con el fin de comparar distintos modelos de Machine Learning que permitieran estimar la marca de incumplimiento crediticio en una base de clientes de una entidad financiera chilena, considerando el cumplimiento de las normas establecidas por la Comisión de Mercado Financiero chilena que establece como criterio específico que cualquier clasificación de clientes debe ser detallada y se deben conocer los atributos que componen tal clasificación, dejando actualmente modelos de Machine Learning más sofisticados, no interpretables o de caja negra fuera de uso, aunque se ha demostrado que éstos mejoran el desempeño en cuanto a la precisión y la discriminación.

La población de estudio correspondió a una base con información de comportamiento entre los años 2014 y 2019, utilizando 28 atributos del cliente, los créditos, la empresa y el comportamiento que permitieron realizar el ajuste de 6 modelos de Machine Learning de clasificación: Regresión Logística, Regresión Lasso, Árboles de clasificación por partición recursiva, Extreme Gradient Boosting, Random Forest y Redes Neuronales.

El modelo seleccionado correspondió al ajustado bajo la metodología Random Forest, con ajuste de hiperparámetros mediante Grid Search y validación cruzada, que obtuvo excelentes resultados de precisión y discriminación, valores muchos más altos respecto de las otras metodologías evaluadas. Dado que este modelo cupo dentro de las técnicas de caja negra, se utilizó la metodología LIME para realizar una estimación local de los atributos que determinan la clasificación de una cantidad aleatoria de clientes, logrando cumplir con el objetivo general que establece la normativa chilena, al ser capaz de explicar de manera detallada la clasificación obtenida.

Palabras claves: Machine Learning, Random Forest, Redes Neuronales, Regresión Logística, Regresión Lasso, Extreme Gradient Boosting, Árboles de Clasificación, Incumplimiento Crediticio, Interpretación Local, LIME

ABSTRACT

This research work was made in order to compare different Machine Learning methods that would allow estimating the brand of credit default in a customer base of a Chilean financial institution considering compliance with the Chilean Financial Market Commission standards that it establishes as specific criterion that any customer classification must be detailed and the attributes that make up such classification must be known, currently leaving more sophisticated, non-interpretable or black box Machine Learning methods out of use although it has been shown that these improve performance in as for precision and discrimination.

Study population corresponded to a base with behavioral information between the years 2014 and 2019, using 28 attributes of the client, the credits, the company and the behavior that allowed the adjustment of 6 classification Machine Learning methods: Logistic Regression, Lasso Regression, Recursive Partition Classification Trees, Extreme Gradient Boosting, Random Forest and Neural Networks.

Selected model corresponded to the one adjusted under the Random Forest methodology, with hyperparameter adjustment through Grid Search and cross validation which obtained excellent precision and discrimination results, values much higher than the other methodologies evaluated. Given that this model fit within the black box techniques, the LIME methodology was used to make a local estimate of the attributes that determine the classification of a random number of clients, achieving the general objective established by the Chilean regulations by be able to explain in detail the classification obtained.

Keywords: Machine Learning, Random Forest, Neural Networks, Logistic Regression, Lasso Regression, Extreme Gradient Boosting, Classification Trees, Credit Default, Local Interpretation, LIME

AGRADECIMIENTOS

*A mis padres, **Eugenia y Carlos**, por confiar en mi talento y enseñarme a actuar con perseverancia y rigor en cada uno de los proyectos que emprendo.*

*A mi hija, **Javiera**, por la paciencia demostrada cuando no he podido estar 100% presente en pos de tener un mejor futuro para nosotras, mi futura licenciada en estadística.*

*A mi hermana y sobrina, **Verónica y Oriana**, por ser mis admiradoras y demostrarme lo orgullosas que están de cada uno de mis logros.*

*A mi novio, **Boris**, por estar incondicionalmente en todos mis procesos.*

*A **Ignacio García** y **Rosana Ferrero**, y a todo el equipo de **Máxima Formación** y la **Universidad de Nebrija** por estar constantemente impulsando el proceso de aprendizaje, por la disposición puesta en cada instancia de preguntas y por el gran conocimiento traspasado a cada uno de nosotros.*

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABLAS

CAPÍTULO 1: INTRODUCCIÓN

1.1. Antecedentes del tema

En Chile, la Comisión para Mercado Financiero (en adelante, CMF) establece que, para la estimación del incumplimiento crediticio[1], *"Cada calificación dentro de una escala de riesgo se define mediante una descripción detallada de los criterios y atributos utilizados para encasillar a los deudores. Esos criterios o atributos explican la discriminación de riesgo contenida en cada calificación."* (CMF, 2007, Capítulo B-1, p.3) [1]. Dado lo anterior, existe una limitación normativa para el uso de modelos de aprendizaje automático (en adelante, Machine Learning), pertenecientes al subgrupo de algoritmos conocidos como caja negra, los cuales mejoran el desempeño en la clasificación de clientes u operaciones; sin embargo, no permiten conocer cuáles son los atributos que influyen en el incumplimiento crediticio.

1.2. Finalidad y objetivos de la investigación

Por medio del análisis de los atributos para una base de datos de una entidad financiera chilena y el ajuste de diferentes modelos de Machine Learning que permitan determinar el incumplimiento crediticio, el ***objetivo general*** de la investigación es responder a la siguiente pregunta:

¿Es posible encontrar un modelo de Machine Learning de caja negra que mejore el desempeño respecto del uso de metodologías clásicas, pudiendo realizar la interpretación de los atributos que determinan el incumplimiento crediticio?

Para poder cumplir el objetivo general de esta investigación, consideramos los siguientes ***objetivos específicos***:

[1] Deudores que presentan operaciones con más de 90 días de morosidad en una ventana de 12 meses, más otros criterios que se definan de acuerdo con el objetivo del modelo a desarrollar.

- Disponer de una base de datos que contenga una variable de respuesta dicotómica, cuyos valores representen una marca de incumplimiento crediticio, además de atributos o variables independientes que permitan determinar cuáles influyen en el comportamiento de los clientes;
- Determinar la población de estudio, de acuerdo con la información disponible;
- Generar un análisis descriptivo de las variables, considerando medidas de tendencia central y posición, distribución, valores perdidos, valores atípicos, etc.;
- Analizar la influencia de cada uno de los atributos respecto del incumplimiento crediticio;
- Separar la base de datos en conjuntos de training y test, con el fin de entrenar y validar los resultados, minimizando el sobreajuste de los modelos;
- Ajustar diferentes modelos de machine Learning y seleccionar el modelo que entregue las mejores métricas de desempeño que se hayan definido para el estudio;
- Validar el modelo seleccionado, considerando la evaluación de las métricas de desempeño en los conjuntos de training y test definidos;
- Interpretar los atributos para una muestra de observaciones, mediante la metodología LIME (local interpretable model-agnostic explanations, por su sigla en inglés).

1.3. Hipótesis

El uso de metodología LIME, permite interpretar los atributos que influyen en el incumplimiento crediticio de clientes de una entidad financiera chilena, al utilizar modelos de Machine Learning, pertenecientes al subgrupo de algoritmos conocidos como caja negra, mejorando el desempeño y los niveles de discriminación.

1.4. Delimitaciones y alcances

Para la construcción de cualquier tipo de modelo se necesita contar con información detallada, la cual debe ser íntegra y de calidad, ya que puede asegurar que el modelo seleccionado sea robusto y estable. La base utilizada para este proyecto viene en formato tabular, por lo que no pasó por un proceso de revisión de la integridad desde un sistema producto.

Dentro de las variables disponibles para el análisis, además, no se encontraron variables de comportamiento externo, tales como deudas en otras instituciones financieras que, en Chile, entregan información rica al momento de determinar el comportamiento crediticio.

CAPÍTULO 2: MATERIAL Y MÉTODOS

2.1. Tipo de investigación

Este estudio se basó en una investigación cuantitativa, que implicó el uso de herramientas de programación y técnicas estadísticas y Machine Learning, por medio del ***lenguaje R*** y su interfaz gráfica ***RStudio***, con el fin de obtener resultados y conclusiones que permitieran responder a la pregunta planteada en el punto 1.2 de este documento. Este tipo de investigación derivó en el análisis exploratorio de los datos, ajustes de modelos de Machine Learning, tales como: Regresión Logística, Regresión Logística penalizada (Lasso), Árbol de clasificación por método de partición recursiva, Random Forest, Redes Neuronales y Extreme Gradient Boosting (XGBoost); además de la comparación, selección de modelos, cumplimiento de supuestos e interpretación de parámetros por medio de la metodología LIME.

2.2. Definición de la población

En este trabajo de investigación se consideró como población objetivo las operaciones crediticias de una entidad financiera chilena, con las siguientes características:

- Ventana de información entre los años 2014 y 2019, contemplando un ciclo económico completo[2];
- Operaciones con menos de 90 días de morosidad en el periodo de estudio;
- Selección aleatoria de periodo de estudio para aquellas operaciones que cumplieron las condiciones anteriores.

Teniendo en cuenta disponibilidad de información para el análisis, se realizó el estudio en base total de 250.675 casos, con 29 variables.

[2] Un ciclo económico considera momentos de expansión y recesión, por lo que se recomienda una ventana de 5 años de información (CMF, 2007, Capítulo B-1, p.3) [1]

2.3. Definición de la variable objetivo y variables independientes

La variable objetivo ("*default*") que se definió para este estudio, representa la marca de incumplimiento crediticio de la operación, la cual es una variable dicotómica que toma el valor 1 si la operación presenta 90 o más días de morosidad en cualquiera de los siguientes 12 meses desde el periodo de estudio; y 0, en otro caso.

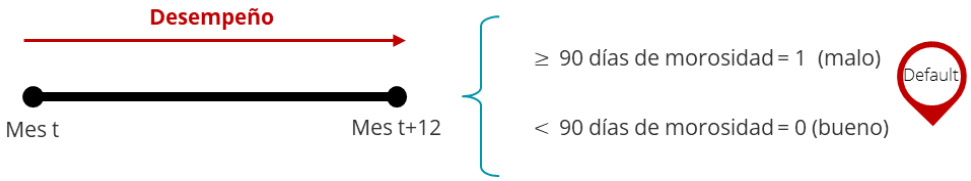


Figura 1: Esquema de construcción variable objetivo "default" (Fuente: Elaboración propia)

Dada la estructura de la variable, el problema correspondió a uno de clasificación de riesgo, por medio de la estimación de una probabilidad que permita discriminar entre un cliente bueno y uno malo.

Las variables analíticas independientes o predictoras, por su lado, permitieron determinar cuáles influyen en el incumplimiento crediticio, de acuerdo con la siguiente distribución:

Variable	Descripción	Variable	Descripción
DiasDeMora	Días de mora en el periodo de observación	NroTrabajadoresEmpresa	Cantidad de trabajadores de la empresa que representa al cliente
AntiguedadLaboral	Antigüedad laboral del cliente en meses	CantVecesMoraUlt3m	Cantidad de veces que la operación ha estado en mora en los últimos 3 meses
AntiguedadEmpresa	Antigüedad de la empresa en meses	PercentilRenta	Percentil de los ingresos del cliente respecto de la empresa donde trabaja
PorcAvanceCredito	Porcentaje de avance del crédito	PlazoCredito	Cantidad de cuotas totales del crédito
Capital	Capital solicitado	PorOpMoraEmpresa	Porcentaje de operaciones en mora que tiene la empresa
CargaFinancieraDeudor	Ratio entre los ingresos del cliente y la cuota del crédito	PorOpMoraEmpresa30d	Porcentaje de operaciones con 30 o más días de mora que tiene la empresa
CreditosTotalesEmpresa	Cantidad de créditos totales de la empresa que representa al cliente	RatioRentaUlt6m	Ratio entre los ingresos informados en los últimos 6 meses respecto de los ingresos actuales del cliente

DiasTrabajadosMes	Días trabajados por el cliente en el mes	RentaDeudor	Ingresos del deudor en el periodo de observación
EdadDeudor	Edad del cliente	MontoCuota	Monto de la cuota del crédito
NivelEndeudamiento	Ratio entre la deuda total y los ingresos del cliente	Licencia6m	Indicador si el cliente ha tenido ausencias por enfermedad en los últimos 6 meses
MaxDiasMoraUlt3m	Máximo días de mora de la operación en los últimos 3 meses	MesCompleto3m	Indicador si el cliente ha trabajado el mes completo en los últimos 3 meses
MaxMoraDeudor	Máximo días de mora en el periodo de observación	MoraUlt3m	Indicador si la operación ha tenido morosidad en los últimos 3 meses
MaxMoraDeudorUlt3m	Máximo días mora del cliente en los últimos 3 meses	EstadoCivil	Estado Civil del cliete
NrosCreditosEmpresa	Cantidad de créditos vigentes del deudor	RubroEmpresa	Rubro de la empersa donde trabaja el cliente

Tabla 1 : Descripción de variables por dimensión

2.4. Método de análisis

Una vez obtenidos los datos, se realizó un análisis descriptivo de las variables, por medio de visualizaciones, medidas de tendencia central, posición, análisis de datos perdidos y atípicos. Posteriormente, se dividió la base en conjuntos de training y test, con el objeto de validar los resultados obtenidos. Luego, se realizó la transformación de las variables, mediante la metodología WOE, un análisis de correlación y reducción de dimensionalidad. Por último, se procedió a ajustar diferentes modelos de Machine Learning, los cuales fueron comparados, seleccionando un modelo óptimo que tuviera altos niveles de precisión, discriminación y potencia.

2.4.1. Análisis descriptivo

El análisis descriptivo consideró la identificación de las variables continuas y categóricas. En el caso de las variables continuas, el análisis se realizó mediante el cálculo de las siguientes medidas de tendencia central, de posición y forma: Mínimo, Q1, mediana, media, Q3, máximo, percentiles (1,5,95,99), varianza, desviación estándar, coeficiente de variación, coeficiente de asimetría, coeficiente de kurtosis.

Adicionalmente, para cada una de las variables, independiente de su tipo, se calculó el porcentaje de valores perdidos, considerando un umbral de un 10%.

2.4.2. Muestreo

El muestreo consideró el balanceo de los casos por medio de una técnica de submuestreo aleatorio que considera la exclusión de casos de la clase mayoritaria, ya que la proporción de casos buenos y malos de la variable objetivo resultó diferente. El proceso de submuestreo consideró la totalidad de casos malos (default = 1, clase minoritaria), seleccionando aleatoriamente una cantidad de casos buenos (default = 0, clase mayoritaria), de tal manera de tener una tasa media de default de un 50%. Este proceso dio origen a la base de training. Para validar los resultados, se consideró la base completa.

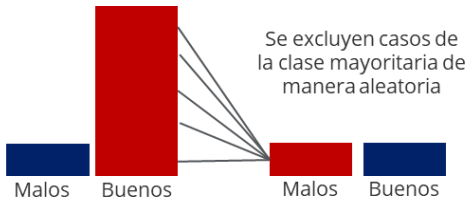


Figura 2: Esquema submuestreo aleatorio (Fuente: Elaboración propia)

2.4.3. Transformación de variables

La transformación de variables continuas y categóricas se realizó mediante una técnica de agrupamiento conocida como "Optimal Binning", la cual busca discretizar variables en grupos que permitan determinar la relación lineal entre la variable analizada y la variable objetivo. Esta técnica de agrupamiento es capaz de abordar problemas de valores atípicos, valores perdidos y la escala de los datos [2], considerando los siguientes criterios en su aplicación:

- Los valores perdidos se consideran en primera instancia como un grupo independiente;
- Los grupos generados deben seguir un orden monotónico de la tasa de default, sea creciente o decreciente;
- Los grupos generados deben tener una participación de, al menos, un 5% de casos, para que sea considerado como representativo. Esta

restricción disminuye la posibilidad de encontrar variables con la condición de varianza cercana a cero (Ferrero, Tema 3, Preprocesado de Datos, p.6), la cual también es analizada;

- Algunas variables pueden considerar cortes definidos bajo criterio experto o de negocio. Además, en caso de que los valores perdidos no alcancen un 5% de participación, se pueden agrupar con alguna categoría que presente tasas de default similares.

El resultado de este análisis derivó en una salida visual que muestra la relación entre los distintos grupos generados y las tasas de default, además de una tabla que permitió tener un resumen de la variable agrupada. Dentro de las métricas de resumen, se realizó la transformación por medio del cálculo del WOE (Weight of Evidence, por sus siglas en inglés):

$$WOE_i = ln\left(\frac{\%Buenos_i}{\%Malos_i}\right) \qquad ecu(1)$$

Donde $\%Buenos_i$ ($\%Malos_i$) corresponde a la proporción de clientes buenos (malos) del grupo i.

Una vez realizada la transformación óptima, se procedió a calcular la potencia de la variable, mediante el estadístico Information Value (IV, por sus siglas en inglés), considerando la misma definición de $\%Buenos_i$ ($\%Malos_i$) descrita anteriormente:

$$IV = \sum_{i=1}^{n} (\%Buenos_i - \%Malos_i) \times \ln\left(\frac{\%Buenos_i}{\%Malos_i}\right) \qquad ecu(2)$$

Esta metodología propone la siguiente clasificación de potencia (Siddiqui, 2006, p. 81) [3]:

Rango IV	Potencia
[0,00; 0,02)	No predictive
[0,02; 0,10)	Débil
[0,10; 0,30)	Media
[0,30; 0,50)	Fuerte
[0,50; +∞)	Posible sobreajuste, revisar.

Tabla 2 : Interpretación potencia estadística IV

De acuerdo con la definición de la Tabla 2, se excluyeron del análisis aquellas variables que presentaron valores de $IV < 2\%$, si corresponde.

2.4.4. Análisis de correlaciones

El análisis de correlaciones se realizó sobre las variables que tuvieran un valor de $IV \geq 2\%$, mediante el método de Spearman que es una medida de asociación lineal por medio de rangos, la cual se define como:

$$\rho = 1 - \frac{6 \sum_i d_i^2}{n \cdot (n^2 - 1)} \qquad ecu(3)$$

donde n es la cantidad de observaciones, y d_i es la diferencia numérica entre los rangos.

Una vez realizado el cálculo de la correlación de Spearman entre todas las variables independientes, se procedió marcar aquellas que tuvieran un valor absoluto > 0.7, considerándolas como *"correlación alta"*. La selección de las variables iniciales, entonces, siguió el siguiente proceso de selección:

- Variables sin correlación alta respecto del total de variables independientes, seleccionadas de manera automática.
- Variables con correlación alta con respecto del resto de variables independientes, se seleccionó aquella que tuviera una mayor potencia estadística IV.

2.4.5. Modelos de Machine Learning

Existen distintos Modelos de Machine Learning para resolver problemas de clasificación. En este proyecto se abordaron metodologías clásicas y metodologías de caja negra, con el objeto de comparar el desempeño en la precisión y el nivel de discriminación de clientes buenos y malos.

En algunas metodologías y con el objetivo de optimizar el desempeño del modelo, fue necesario realizar la búsqueda de hiperparámetros que son configuraciones que se utilizan en el proceso de entrenamiento, las cuales no es posible conocer a priori. Dado esto, la selección de los hiperparámetros se realiza mediante la aplicación de una metodología conocida como Grid Search, la cual consiste en una búsqueda exhaustiva sobre un subconjunto de hiperparámetros en forma de grilla, especificados previamente y guardando el resultado para cada una de las combinaciones.

2.4.5.1 Regresión Logística

La Regresión Logística es una metodología estadística que fue desarrollada por David Cox en 1958, la cual es utilizada cuando la variable de respuesta es categórica, estimando la probabilidad mediante un conjunto de variables independientes. Para realizar la estimación de la probabilidad, se utiliza la transformación *logit*, mediante la siguiente ecuación:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \qquad ecu(4)$$

donde:

- p: probabilidad del evento
- X: variable independiente
- β_0: Intercepto de la regresión

- β_k: parámetros

La transformación *logit* corresponde al logaritmo del odds, es decir, $\log\left(\frac{p_i}{1-p_i}\right)$ y es usada para linealizar la probabilidad, dejando la estimación en un rango entre 0 y 1 (Siddiqui, 2006, p.90) [3]. Esta probabilidad se utiliza para estimar los parámetros de la regresión logística β_0 a β_k, los cuales miden cuánto cambia el logit cuando las variables independientes varían en una unidad.

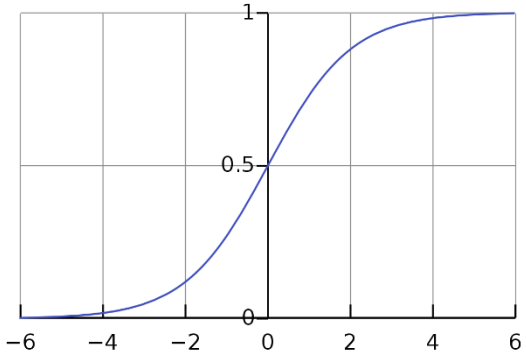


Figura 3: Función Logística (Fuente)

2.4.5.2 Regresión Logística Lasso

La Regresión Logística Lasso es un enfoque de regresión penalizado que estima los coeficientes de regresión con la restricción de que la suma de los valores absolutos de los coeficientes de regresión $\sum_{j=1}^{k}|\beta_j|$, son menores o iguales a una constante positiva, eliminando las variables no significativas, ya que las penalizaciones convierten algunos coeficientes en cero (Sun Mi Kim et al., 2018) [4].

El estimador logístico Lasso está dado por la siguiente fórmula:

$$\sum_{i=1}^{n}[-y_i(\beta_0+\beta_1X_{i1}+\cdots+\beta_kX_{ik}) \\ +log\left(1+exp(\beta_0+\beta_1X_{i1}+\cdots+\beta_kX_{ik})\right)] \qquad ecu(5)$$

sujeto a la restricción $\sum_{j=1}^{k} |\beta_j| \leq \lambda$, donde $\lambda \geq 0$ corresponde al parámetro de penalización que controla los coeficientes con valores cercanos a cero.

2.4.5.3 Árbol de clasificación

Los Árboles de Clasificación y Regresión (CART) fueron construidos por Breiman et al. (1984). Son algoritmos principalmente usados en problemas de clasificación, donde las variables independientes pueden ser tanto continuas como categóricas. El objetivo es dividir el espacio de las variables independientes en conjuntos homogéneos disjuntos, es decir, no sobrepuestos.

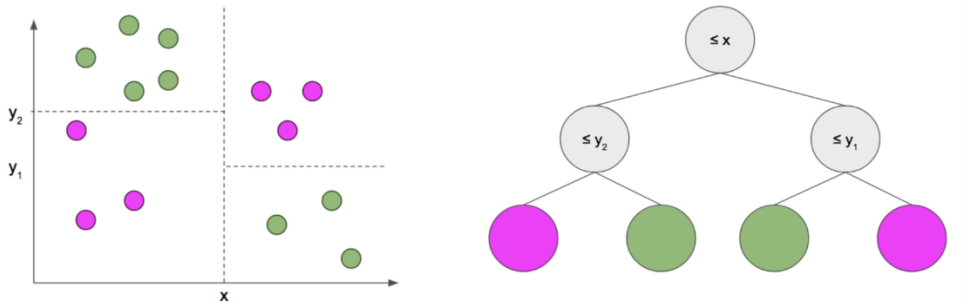


Figura 4: Esquema Árbol de clasificación (Fuente)

La construcción del árbol se realiza por medio de particiones binarias recursivas, analizando cada una de las variables y seleccionando la mejor partición sobre cada uno de los grupos. Los algoritmos más comunes para la selección de las particiones son: Índice de Gini, Chi cuadrado y Entropía. Las definiciones de cada indicador se encuentran en el Anexo 4.

2.4.5.4 Random Forest

Random Forest corresponde a una metodología que construye múltiples árboles de clasificación y/o regresión (2.4.5.3). En el caso de problemas de clasificación, para cada uno de los árboles construidos se guarda la estimación para posteriormente tomar la moda que da lugar a la predicción final.

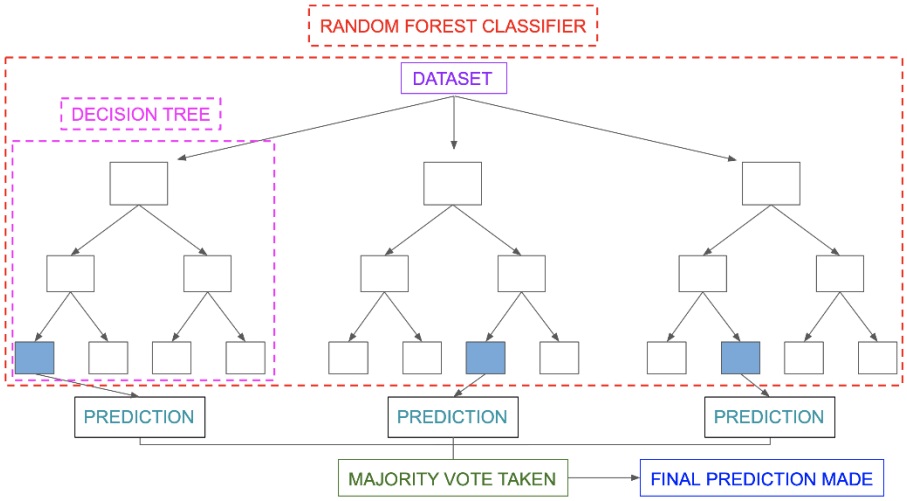


Figura 5: Esquema Random Forest (Fuente)

2.4.5.5 Redes Neuronales

Las redes neuronales son algoritmos que aprenden de los datos, inspirados en cómo funciona el cerebro (Ferrero, 2020) [5], donde cada una de las neuronas recibe entradas, las procesa y entrega una única salida. Luego, los modelos de red neuronal son un conjunto de neuronas que se agrupan en tres capas: capa de entrada, capa oculta y capa de salida. Entre las capas de entrada y salida se pueden agregar múltiples capas ocultas. Si ese es el caso, entonces, el modelo pasa a ser de aprendizaje profundo o Deep Learning.

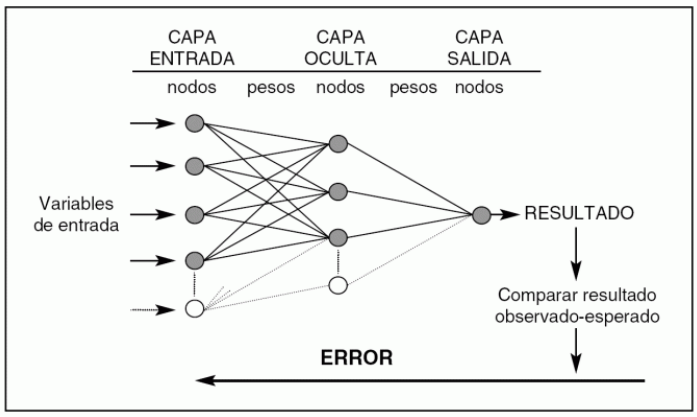


Figura 6: Esquema Red Neuronal (Fuente)

Para calcular la salida o resultado, la red neuronal utiliza pesos, que representa la conexión entre las neuronas, midiendo la influencia de cada variable en el resultado final. Esto se puede representar por medio de la siguiente fórmula:

$$y = w_1 x_1 + w_2 x_2 \ + \cdots + w_n x_n + b, \qquad ecu(6)$$

donde:

- y: probabilidad del evento
- x_i: valores de entrada
- w_i: pesos o parámetros
- b: intercepto

Lo anterior se visualiza como una regresión lineal; sin embargo, para no distorsionar el valor de la salida, se utilizan funciones de activación. Luego, la red puede ser representada como:

$$y = f(w_1 x_1 + w_2 x_2 \ + \cdots + w_n x_n + b), \qquad ecu(7)$$

donde f corresponde a la función de activación. Dentro de las funciones de activación más comunes se encuentran: Función lineal, función escalón, función sigmoide y función RELU, las cuales se describen a continuación:

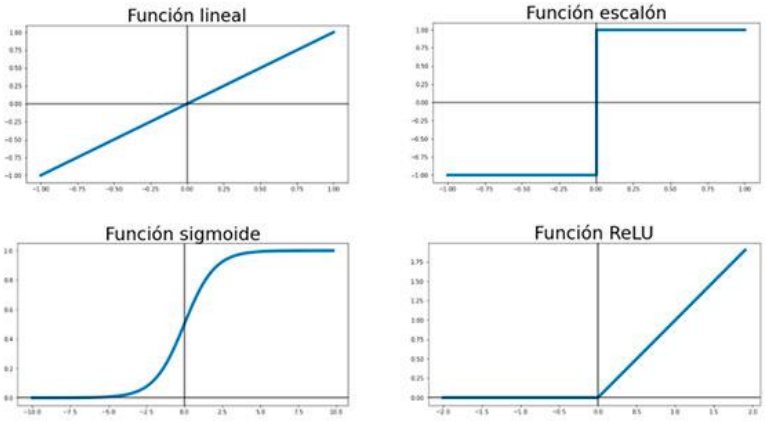


Figura 7: Ejemplos de funciones de activación para Redes Neuronales (Fuente)

- Función lineal: considera la salida igual que la entrada, representada por la función identidad. Su uso implica un comportamiento igual a una regresión lineal;
- Función escalón: comúnmente utilizada cuando la salida es categórica y el objetivo es clasificar. En la práctica no es muy utilizada;
- Función sigmoide: función que toma valores entre -1 y 1, por lo que es útil controlando los valores extremos y sirve para representar probabilidades.
- Función RELU: es una de las funciones más utilizadas, ya que se comporta como una función constante para valores negativos, mientras como una función lineal para valores positivos.

2.4.5.6 Extreme Gradient Boosting

Es un algoritmo de aprendizaje supervisado que suele aumentar la precisión frente a otros modelos de Machine Learning. Realiza un procesamiento secuencial de los datos, aplicando árboles de regresión y/o clasificación y teniendo por objetivo minimizar el error en cada iteración (boosting), por medio de una función de pérdida o coste (Ferrero, 2020) [6]. Esta función de pérdida depende de la variable de respuesta: para problemas de regresión la más comúnmente utilizada es el error cuadrado, mientras que para problemas de clasificación se utiliza la pérdida logarítmica [7].

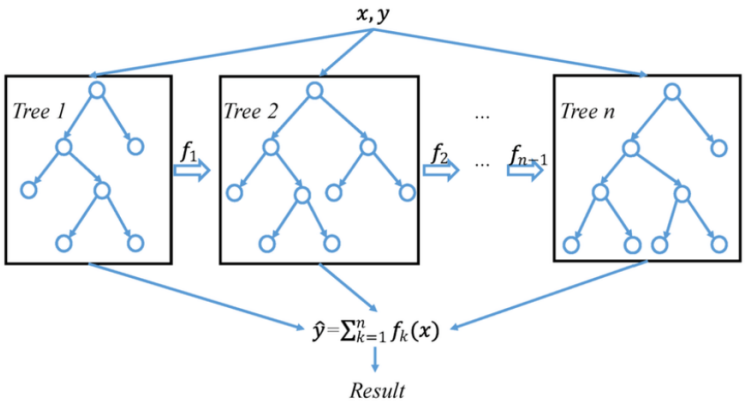


Figura 8: Esquema Extreme Gradient Boosting (Fuente)

2.4.6. Validación de modelos y métricas de desempeño

2.4.6.1 Validación

Para determinar la robustez y estabilidad de los modelos ajustados, se utilizaron dos formas de validación. En primer lugar, un análisis del desempeño en el conjunto de validación y, en algunos algoritmos, para la optimización de los parámetros, una validación cruzada mediante k-fold:

Figura 9: Estructura de validación cruzada k-fold (Fuente)

K-fold sigue el siguiente proceso [8]:

1. Divide los datos de entrenamiento en K partes iguales
2. Ajusta el modelo en k-1 partes y calcule métricas de desempeño usando el modelo ajustado en la k-ésima parte;
3. Repita k veces, usando cada subconjunto de datos como el conjunto de prueba una vez. (generalmente k = 5 ~ 20).

Nota: La estimación mediante k-fold puede resultar sobreajuste si la base que se utiliza para la estimación está ordenada.

2.4.6.1 Métricas de desempeño

Cada uno de los modelos ajustados fueron sometidos al proceso de validación, calculando en cada uno distintas métricas de desempeño:

Matriz de confusión

Una matriz de confusión C [9] es una matriz de $c \times c$, donde c corresponde a la cantidad de clases o categorías. Cada elemento C_{ij} es el número de observaciones etiquetados como clase i (valor predicho), pero que pertenece a la clase j (valor real). Luego, se definen los siguientes elementos: TP, TN, FP y FN.

- TP: Verdadero positivo, registros de clase 1 que están correctamente clasificados como clase 1;
- TN: Verdadero negativo, registros de clase 0 que están correctamente clasificados como clase 0;
- FP: Falso positivo, registros de la clase 0 que están incorrectamente clasificados como clase 1 (Error Tipo-I);
- FN: Falso negativo, registros de la clase 1 que están incorrectamente clasificados como clase 0 (Error Tipo-II).

Luego, para la comparación de modelos, se utilizó la métrica de precisión o *Accuracy* que es una de las métricas más populares de clasificación y se define como:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad ecu(8)$$

Mientras mayor sea el nivel de precisión, mejor es el desempeño del modelo.

Kolmogorov - Smirnov

La prueba de Kolmogorov - Smirnov, más conocida como KS, determina la distancia máxima entre dos funciones de probabilidad acumulada, comparando una distribución empírica versus una distribución teórica (Cieslak et al., 2007).

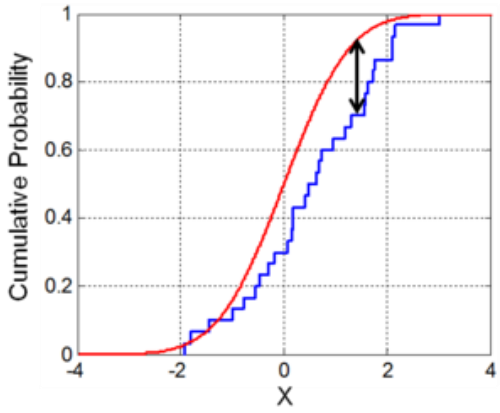


Figura 10: Ejemplo curva Kolmogorov – Smirnov (Fuente)

El contraste de la prueba se basa en las diferencias de las frecuencias relativas acumuladas, mediante la siguiente relación:

$$D_i = F_1(x_i) - F_2(x_i) \qquad ecu(9)$$

Donde $F_1(x_i)$ y $F_2(x_i)$ son la frecuencia acumulada de la variable x_i en un punto de corte determinado. Considerando las diferenciasen distintos puntos de corte, se construye el estadístico KS:

$$Z_{KS} = max(|D_i|) \qquad ecu(10)$$

Mientras mayor sea el valor de KS, mejor es el desempeño del modelo.

Área bajo la curva ROC

El área bajo la curva ROC, AUC [9] es una métrica que toma valores entre 0.5 y 1. La curva ROC es una gráfica basada en los errores de clasificación respecto de la variable de respuesta en distintos puntos de corte dados.

Para la construcción de esta curva se utilizan 2 parámetros:

- Sensibilidad (TP): Proporción de verdaderos positivos
- 1 - Especificidad (TN): Proporción de verdaderos negativos

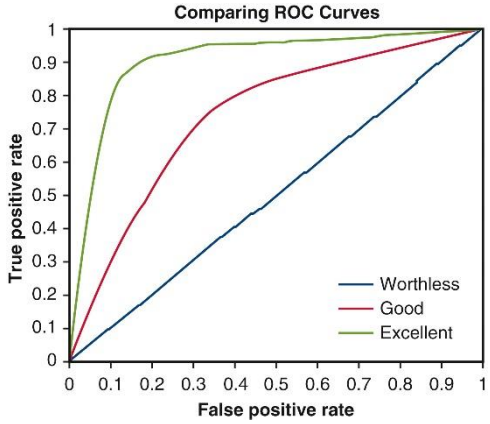


Figura 11: Ejemplo curva ROC (Fuente)

Mientras mayor sea el valor de KS, mejor es el desempeño del modelo.

2.4.7. Comparación de modelos y selección del modelo óptimo

Para la selección del modelo final se utilizó de manera combinada las métricas de desempeño Accuracy, KS y AUC definidas en el punto 2.4.6, considerando además el contraste entre las metodologías clásicas de Machine Learning y las metodologías de caja negra. Adicionalmente, con el objeto de revisar el nivel de sobreajuste de los modelos, las métricas de desempeño se calcularon tanto en el conjunto de desarrollo y de validación.

2.4.8. Importancia de los predictores e interpretación

Una vez validado el modelo final, se procedió a calcular la importancia relativa de cada predictor, es decir, medir cuál es su contribución individual en el modelo, para luego realizar la interpretación de cada uno de los parámetros, mediante la metodología LIME (local interpretable model-agnostic explanations, por su sigla en inglés), la cual corresponde a un tipo de modelos sustitutos que se entrenan para aproximarse a las predicciones de un modelo de caja negra subyacente [10], con el objeto de explicar predicciones individuales por medio de visualizaciones.

El enfoque que utiliza LIME consiste en probar lo que sucede con la predicción de un caso en particular cuando varían los datos de entrada que se ingresan al modelo de caja negra, generando un nuevo conjunto de datos con muestras con perturbaciones de datos y las predicciones obtenidas originalmente. Las perturbaciones de datos tabulados, en general, se obtienen extrayendo una distribución normal con la media y desviación estándar de la variable perturbada.

Luego, se entrena un modelo interpretable que se pondera por la proximidad de las muestras perturbadas, donde no necesariamente debe ser una buena aproximación del modelo global, sino que basta con que exista una buena aproximación local del caso particular estudiado.

La representación matemática de los modelos sustitutos locales es como sigue:

$$explicacion(x) = arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \qquad ecu(11)$$

Donde la explicación del individuo x que minimiza el error L por medio del modelo g, midiendo qué tan cerca está la explicación de la predicción respecto del modelo original f. Por otro lado, G es el conjunto de modelos que se prueban y $\Omega(g)$ es la complejidad del modelo g, por ejemplo, la cantidad de variables

utilizadas en la estimación. Por último, π_x corresponde a la proximidad que se considera para la explicación.

Luego, el proceso para interpretar un individuo es como sigue [11]:

1. Dada una observación, crear datos de características replicados con ligeras modificaciones en las variables;
2. Calcular la medida de la distancia de similitud entre la observación original y las observaciones modificadas;
3. Aplicar el modelo de aprendizaje automático seleccionado para predecir los resultados de los datos modificados;
4. Seleccionar m número de características para describir mejor los resultados previstos;
5. Ajustar un modelo simple a los datos modificados, explicando el resultado del modelo complejo con m características de los datos modificados ponderados por su similitud con la observación original;
6. Utilizar las ponderaciones de características resultantes para explicar el comportamiento local de la observación.

CAPÍTULO 3: RESULTADOS Y DISCUSIÓN

3.1. Análisis descriptivo[3]

La base utilizada correspondió a una cartera de crédito de una institución financiera chilena, considerando información de un ciclo económico de 5 años, cumpliendo con las exigencias normativas locales CMF y un total de 250.675 registros independientes y 29 variables, con la siguiente estructura:

```
str(data)
'data.frame':    250675 obs. of  29 variables:
 $ default               : int  0 0 0 0 0 0 0 1 0 1 ...
 $ DiasDeMora            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AntiguedadLaboral     : int  302 115 125 39 22 144 317 54 187 47 ...
 $ AntiguedadEmpresa     : int  99 458 64 500 365 7 219 241 268 49 ...
 $ PorcAvanceCredito     : num  0.667 0.714 0.214 0.667 0.667 ...
 $ Capital               : int  356150 2940012 408068 1222024 304224 1018825 303929 1015696 612763
1017148 ...
 $ CargaFinancieraDeudor : num  10.68 7.03 10.07 7.58 7.3 ...
 $ CreditosTotalesEmpresa: int  1 34 35 505 2064 90 2156 5 32 13 ...
 $ DiasTrabajadosMes     : int  30 30 30 30 30 30 30 30 30 30 ...
 $ EdadDeudor            : int  59 43 70 42 64 48 56 51 65 47 ...
 $ NivelEndeudamiento    : num  0.359 1.531 0.959 0.95 0.514 ...
 $ MaxDiasMoraUlt3m      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ MaxMoraDeudor         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ MaxMoraDeudorUlt3m    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ NrosCreditosEmpresa   : int  1 1 1 1 1 1 1 1 1 ...
 $ NroTrabajadoresEmpresa: int  10 66 121 972 3504 336 8080 26 73 41 ...
 $ CantVecesMoraUlt3m    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ PercentilRenta        : num  1 0.625 0.7143 0.5647 0.0812 ...
 $ PlazoCredito          : int  12 49 14 24 12 12 12 60 18 12 ...
 $ PorOpMoraEmpresa      : num  0 0 0.0833 0.0472 0.0496 ...
 $ PorOpMoraEmpresa30d   : num  0 0 0.0556 0.0335 0.0453 0.0222 0.0128 0 0.0625 0 ...
 $ RatioRentaUlt6m       : num  0.987 1.069 1.095 0.789 1.088 ...
 $ RentaDeudor           : num  356325 813028 353250 509995 210000 ...
 $ MontoCuota            : int  33377 115613 35065 67322 28781 98536 28727 31098 42301 97967 ...
 $ Licencia6m            : int  0 1 1 0 0 0 0 0 0 0 ...
 $ MesCompleto3m         : int  1 1 0 1 1 1 1 1 1 1 ...
 $ MoraUlt3m             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ EstadoCivil           : chr  "S" "S" "C" "S" ...
 $ RubroEmpresa          : chr  "COMERCIO AL POR MAYOR Y AL POR MENOR, REPARACION DE VEHICULOS
AUTOMOTORES Y MOTOCICLETAS" "TRANSPORTE Y ALMACENAMIENTO" "COMERCIO AL POR MAYOR Y AL POR MENOR,
REPARACION DE VEHICULOS    AUTOMOTORES Y MOTOCICLETAS" "ACTIVIDADES FINANCIERAS Y DE SEGUROS" ...
```

Del cuadro anterior se observa que, a excepción de *EstadoCivil* y *RubroEmpresa*, el resto de las variables son del tipo numérico; sin embargo, por la definición de las variables, *default* corresponde a la variable objetivo del tipo dicotómica y representa la tasa de incumplimiento crediticio para cada uno de los clientes en la base, mientras que las variables: *CantVecesMoraUlt3m*, *Licencia6m*, *MesCompleto3m* y *MoraUlt3m* representan categorías. Luego, la distribución de variables quedó como sigue:

Variable de respuesta	Variables continuas	Variables categóricas
1	22	6

Tabla 3 : Descripción de variables por dimensión

[3] El detalle del análisis se encuentra en el Anexo 1.

Una vez clasificadas las variables, se realizó un análisis de valores perdidos, identificando valores especiales que los representan, de acuerdo con las siguientes definiciones:

- El valor perdido de la variable *NivelEndeudamiento* está representado por el valor 22418290, que corresponde a 10 veces el valor máximo real.
- El resto de los valores perdidos están representados por el valor -99999.

Para excluir una variable del análisis, se definió como umbral un 10% de valores perdidos. A continuación, se presenta una visualización para cada variable:

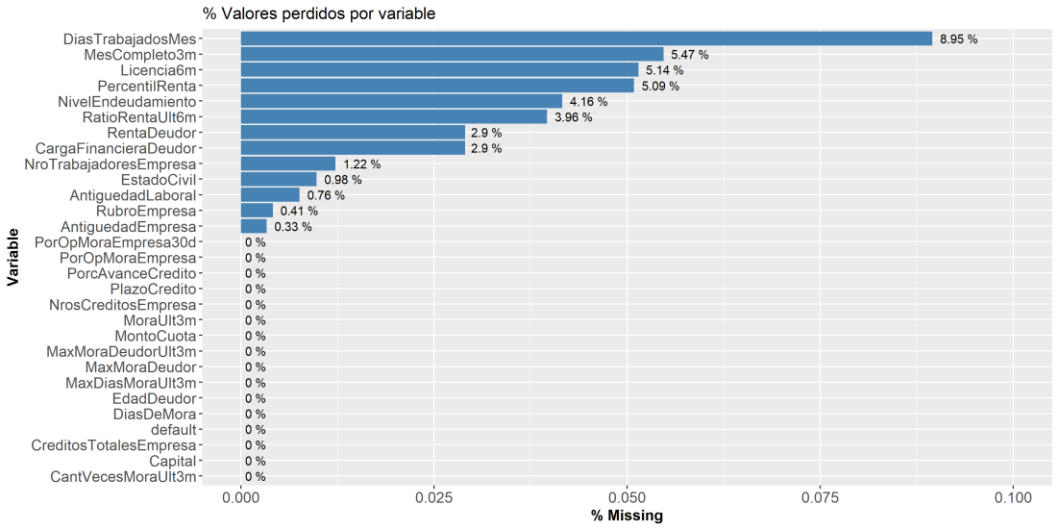


Figura 12: Resultado análisis valores perdidos base completa

La Figura 12 muestra que la variable con mayor cantidad de valores perdidos corresponde a *DiasTrabajadosMes*, con valor que asciende a 8.95%; sin embargo, ninguna supera el 10%, por lo que no hay exclusiones. Adicionalmente, se calcularon medidas de tendencia central para las variables continuas, a modo informativo, ya que las transformaciones aplicadas, tanto en variables continuas como categóricas, incorporaron el análisis de los valores perdidos y atípicos, y su relación con el incumplimiento.

3.2. Muestreo

La tasa media de incumplimiento en la base total fue un **11,48%**, lo que supuso un problema de desbalanceo. Por lo anterior, se realizó el proceso de balanceo de las clases, mediante la metodología de submuestreo, manteniendo todos los casos de clientes marcados como 1, y seleccionando a través de un muestreo aleatorio simple los clientes marcados como 0, obteniendo el conjunto de training (balanceado) y el conjunto de test (base completa), tal como se muestra a continuación:

Indicador	Training Muestreo Balanceado	Test Base Completa
Registros	57.570	250.675
Tasa incumplimiento	50%	11,48%

Tabla 4 : Resultado muestreo. Cantidad de registros en conjuntos de Training y Test

3.3. Transformación de variables[4]

La transformación de variables se realizó por medio de la metodología "Optimal Binning", generando agrupaciones que permitieron optimizar la relación de la variable estudiada respecto de la tasa de incumplimiento, controlando los valores perdidos y atípicos. Una vez seleccionada la agrupación, se calculó la potencia de la variable por medio del valor del Information Value (IV), obteniendo el siguiente resultado:

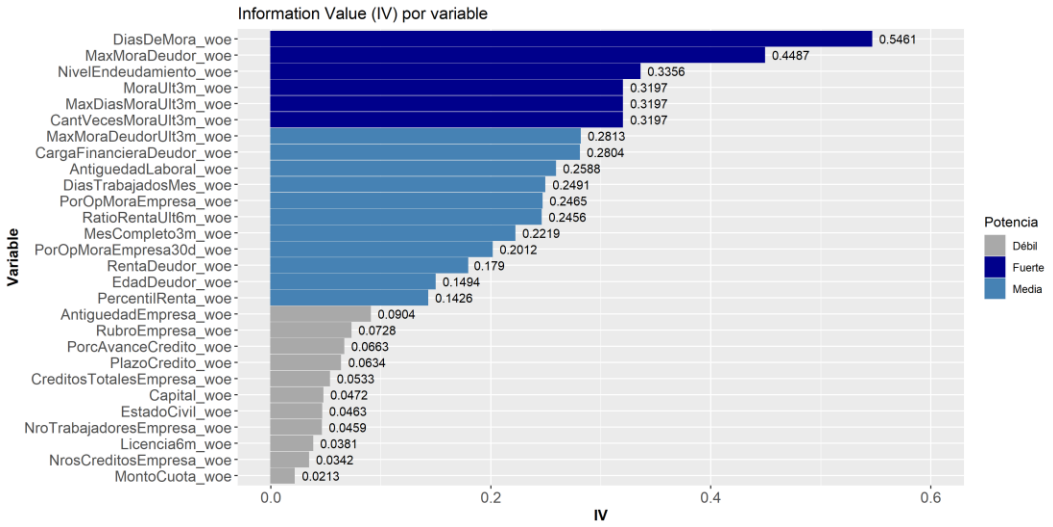


Figura 13: Information Value por variable – Base Training

[4] El detalle del análisis se encuentra en el Anexo 2.

De la visualización anterior, se desprende que:

- No existieron variables con $IV < 0.02$, por lo que no hay exclusión de variables;
- 11 variables resultaron tener potencia débil;
- 11 variables resultaron tener potencia media;
- 6 variables resultaron tener potencia fuerte. Particularmente, las variables: *MoraUlt3m*, *MaxDiasMoraUlt3m* y *CantVecesMoraUlt3m*, coincidieron en el valor de IV, debido a que la agrupación entregó el mismo resultado, por lo que las variables con potencia fuerte finalmente fueron 4.
- La variable *DiasDeMora* presentó un $IV > 0.5$, lo que podía suponer un sobreajuste; sin embargo, al ser un valor no muy alto, no generó problemas.
- Adicionalmente, el análisis de varianza cercana a cero arrojó que ninguna variable presentaba categorizaciones con baja frecuencia.

Por último, los WOEs calculados para cada una de las variables fueron asignados simultáneamente a las bases de training y test, para poder realizar el proceso de modelamiento y validación.

3.4. Análisis de correlaciones[5]

Se realizó el cálculo del coeficiente de correlación de Spearman entre las 28 variables independientes, definiendo como correlación alta un nivel $> 70\%$. De este análisis se seleccionaron aquellas variables que no tuvieran una alta correlación con ninguna otra y, para aquellas variables correlacionadas, se seleccionó aquella que tuviera una mayor potencia, considerando la relación con el incumplimiento.

De este proceso, se excluyeron 7 variables, considerando aquellas relativas a la morosidad identificadas en el proceso de transformación que entregaban exactamente la misma agrupación, más otras que se detallan a continuación:

[5] El detalle del análisis se encuentra en el Anexo 3.

A continuación, se presentan las variables con alta correlación:

Var1	Var2	corr	IV Var1	IV Var2
MaxDiasMoraUlt3m_woe	CantVecesMoraUlt3m_woe	1,0000	0,3197	0,3197
MaxDiasMoraUlt3m_woe	MoraUlt3m_woe	1,0000	0,3197	0,3197
CantVecesMoraUlt3m_woe	MoraUlt3m_woe	1,0000	0,3197	0,3197
DiasDeMora_woe	MaxMoraDeudor_woe	0,9728	0,5461	0,4487
MaxDiasMoraUlt3m_woe	MaxMoraDeudorUlt3m_woe	0,9160	0,3197	0,2813
MaxMoraDeudorUlt3m_woe	CantVecesMoraUlt3m_woe	0,9160	0,2813	0,3197
MaxMoraDeudorUlt3m_woe	MoraUlt3m_woe	0,9160	0,2813	0,3197
CreditosTotalesEmpresa_woe	NroTrabajadoresEmpresa_woe	0,8942	0,0533	0,0459
DiasDeMora_woe	MaxDiasMoraUlt3m_woe	0,8027	0,5461	0,3197
DiasDeMora_woe	CantVecesMoraUlt3m_woe	0,8027	0,5461	0,3197
DiasDeMora_woe	MoraUlt3m_woe	0,8027	0,5461	0,3197
MaxDiasMoraUlt3m_woe	MaxMoraDeudor_woe	0,7670	0,3197	0,4487
MaxMoraDeudor_woe	CantVecesMoraUlt3m_woe	0,7670	0,4487	0,3197
MaxMoraDeudor_woe	MoraUlt3m_woe	0,7670	0,4487	0,3197
PorOpMoraEmpresa_woe	PorOpMoraEmpresa30d_woe	0,7499	0,2465	0,2012
DiasDeMora_woe	MaxMoraDeudorUlt3m_woe	0,7416	0,5461	0,2813
MaxMoraDeudor_woe	MaxMoraDeudorUlt3m_woe	0,7368	0,4487	0,2813

Tabla 5 : Variables con correlación de Spearman $>$ **70%**

Luego de la exclusión, se contó con 21 variables para el ajuste de los Modelos de Machine Learning.

3.5. Modelos de Machine Learning[6]

Se ajustaron 6 modelos de Machine Learning, considerando las 21 variables con sus respectivas transformaciones WOE y con niveles de correlaciones $\leq$ **70%** entre ellas. Dentro de las metodologías de Machine Learning utilizadas que sí son interpretables: Regresión Logística, Árbol de clasificación con partición recursiva y Regressión Lasso, se ajustaron mínimamente los hiperparámetros, encontrando resultados similares de precisión y discriminación; para las metodologías de Machine Learning de caja negra: Random Forest, Extreme Gradient Boosting y Redes Neuronales, se utilizó la

[6] El detalle de los resultados se encuentra en el Anexo 5.

metodología Grid Search y validación cruzada para optimizar varios hiperparámetros en la base de training.

A continuación, se presenta un resumen del resultado obtenido para cada uno de los modelos estimados:

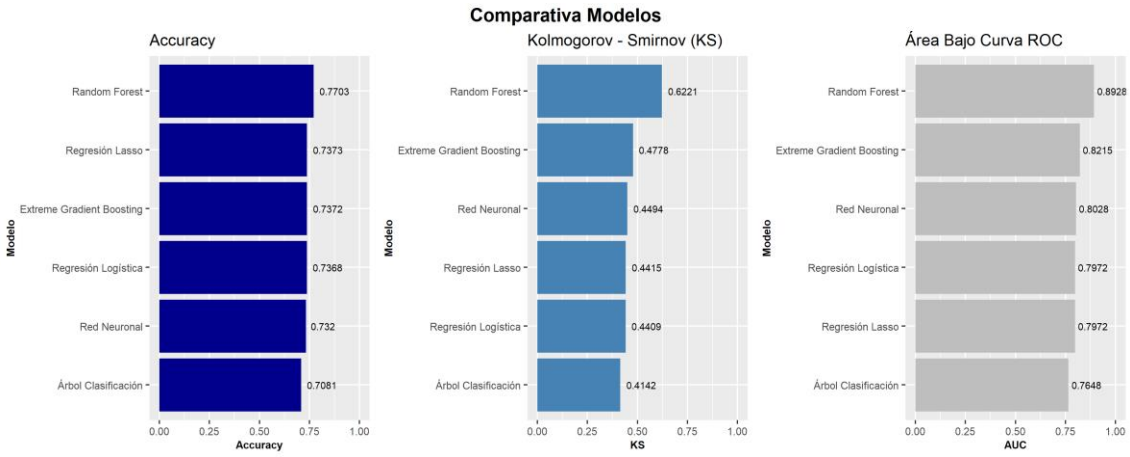


Figura 14: Métricas de desempeño para modelos de Machine Learning ajustados

La Figura 14 muestra el resultado comparativo de los 6 modelos estimados, considerando las métricas de desempeño Accuracy, KS y AUC en la base de test. El modelo con el peor desempeño fue el estimado mediante Árbol de Clasificación con partición recursiva, seguido de los modelos de Regresión Logística y Regresión Lasso, en las métricas de KS y AUC, mientras que en Accuracy el peor desempeño lo presenta la Red Neuronal que, si bien tuvo ajuste de hiperparámetros, al considerar sólo una capa oculta, no logra mejorar sustancialmente el desempeño. Lo anterior ratifica que los modelos de caja negra aumentan el desempeño de los modelos.

El modelo con el mejor desempeño correspondió a Random Forest, con métricas superiores tanto en Accuracy, KS y AUC, obteniendo un 77,03%, 62,21% y 89,28%, respectivamente. Particularmente, este modelo tiene niveles muy superiores de KS y AUC, con respecto de los otros modelos, lo que implica un buen nivel de discriminación entre clientes buenos y malos. A continuación, se muestra una visualización de las métricas obtenidas en la base de test:

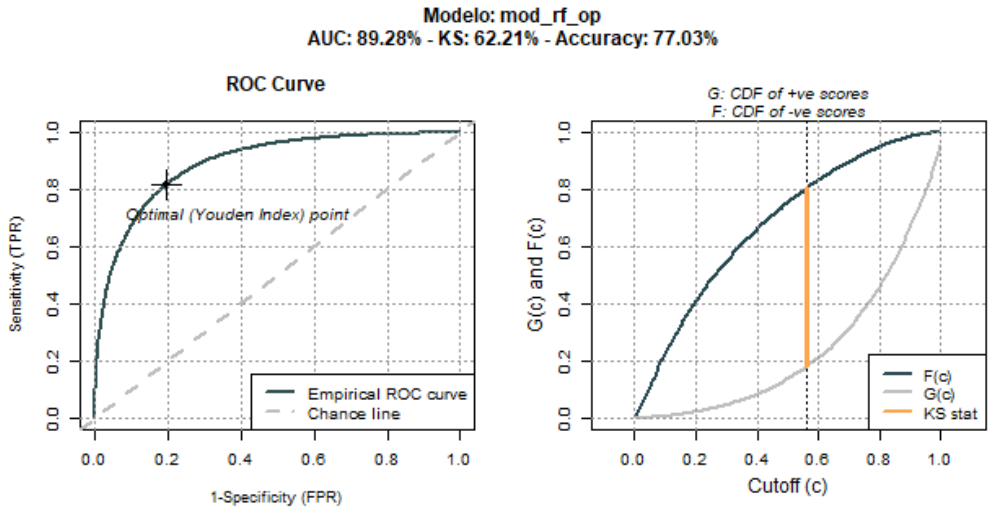


Figura 15: Métricas de desempeño Modelo Random Forest

Este resultado estuvo impulsado por la selección de los hiperparámetros en la estimación mediante Grid Search y validación cruzada que, consideró inicialmente un ajuste por defecto, con el fin de observar la cantidad de árboles óptimo donde el error de clasificación lograra una estabilidad, para luego optimizar el valor de las variables candidatas para cada división. Vale mencionar que el tiempo de ejecución en este modelamiento fue elevado y que presentó un leve sobreajuste. El resultado de estos ajustes dio como resultado 200 árboles y 3 variables:

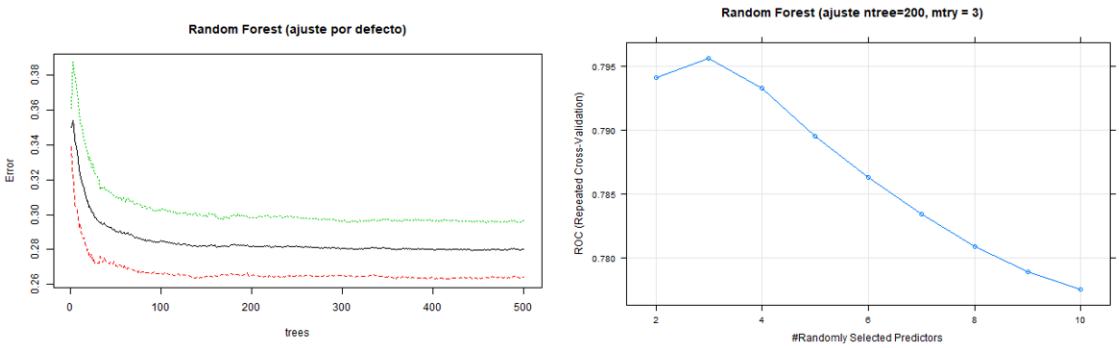


Figura 16: Desarrollo modelamiento Random Forest – Selección de hiperparámetros mediante Grid Search y Validación Cruzada

3.6. Importancia de los predictores e interpretación

El modelo Random Forest seleccionado, consideró ajuste múltiples árboles de clasificación, obteniendo el siguiente orden de importancia de las variables:

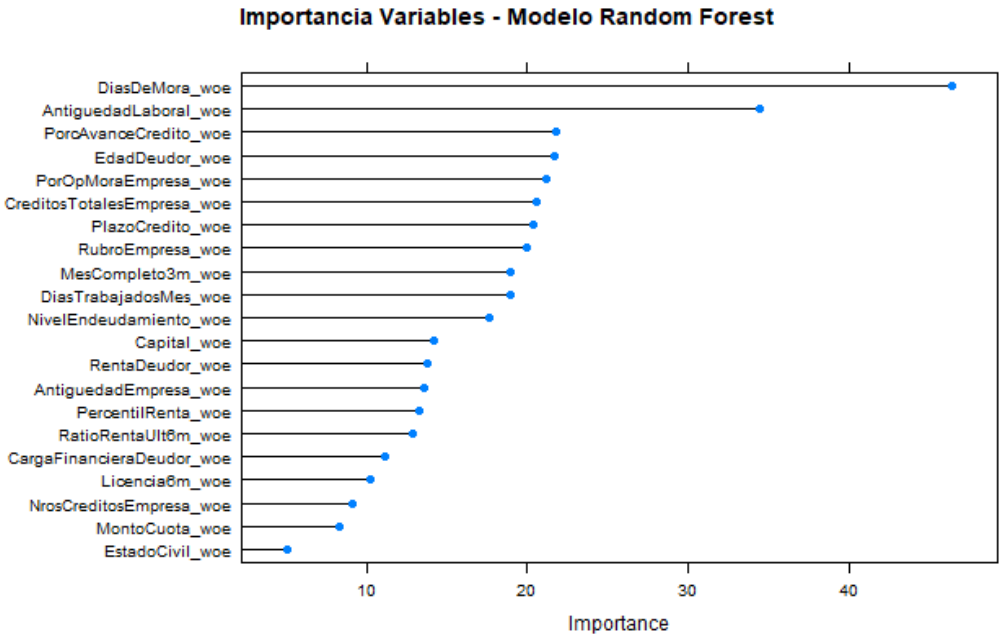


Figura 17: Importancia variables Modelo Random Forest

De acuerdo con la Figura 17, la variable más importante correspondió a los días de mora en el periodo de observación que se condice con el resultado obtenido en el cálculo de la potencia, seguida de la antigüedad laboral del cliente, el porcentaje de avance del crédito y la edad. La variable menos importante correspondió al estado civil del cliente, que tiene un nivel de importancia bastante disminuido respecto del resto, seguida del monto de cuota del crédito.

Dada la complejidad de realizar una visualización de las reglas que componen Random Forest, es que la interpretación se realizó para una cantidad limitada de clientes de la base, mediante la metodología LIME, la cual nos indicó una visualización con la probabilidad estimada de clasificación y las cinco variables más importantes que la determinan. Como la metodología supone la estimación de un modelo simple, tiene un error intrínseco, además de heredar los errores de estimación del modelo original; sin embargo, resultaron eficientes y útiles.

A continuación, se presenta el resultado de la interpretación local para 6 clientes seleccionados de manera aleatoria, donde las barras azules indican las variables que aportan a la probabilidad de manera positiva (support) o de manera negativa (contradicts):

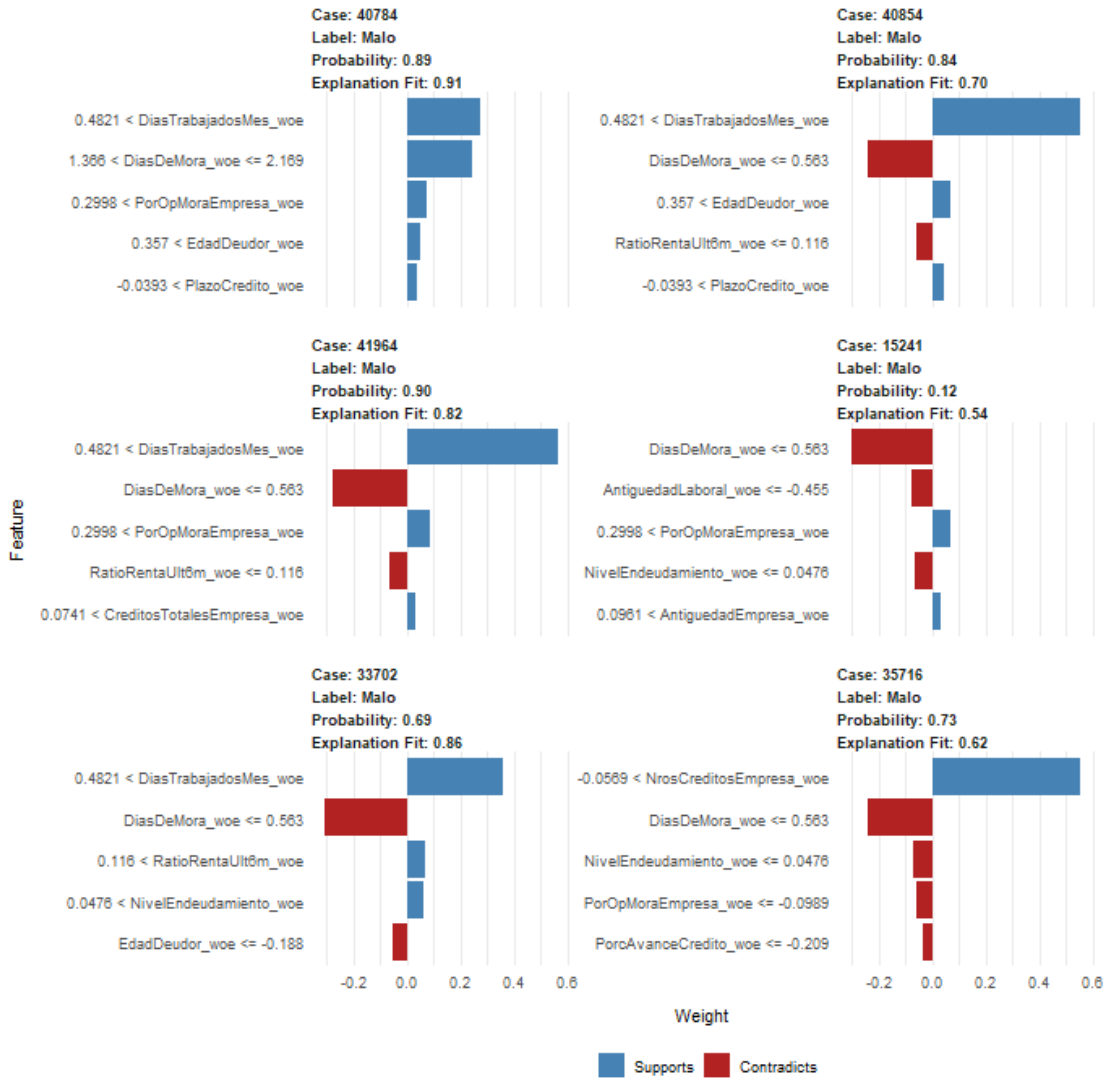


Figura 18: Visualización Resultado estimación vía LIME – 6 clientes seleccionados aleatoriamente

Analizando la Figura 18, se pudo observar que la etiqueta para la estimación es un cliente clasificado como "Malo"; luego, las probabilidades representadas por el valor de "probability" nos ayudaron a determinar la clasificación que entrega LIME a cada caso.

Por ejemplo, el caso 15241 tiene una probabilidad igual a 0.12, lo que indica que el cliente queda clasificado como "Bueno", resultado que se condice con la clasificación real; lo mismo sucede con el resto de las estimaciones. Finalmente, al observar el valor de "Explanation Fit", el cual muestra qué tan bien el modelo simplificado explica la región local, tenemos valores sobre un 50%, por lo que pudimos concluir que para estos 6 casos se logra una buena estimación. A continuación, se presenta el peso de cada variable dentro de la estimación de la probabilidad en el modelo simplificado, donde a mayor intensidad mayor efecto, sea éste positivo o negativo, predominando los días de mora que ya vimos que obtuvo la mayor importancia en el modelo.

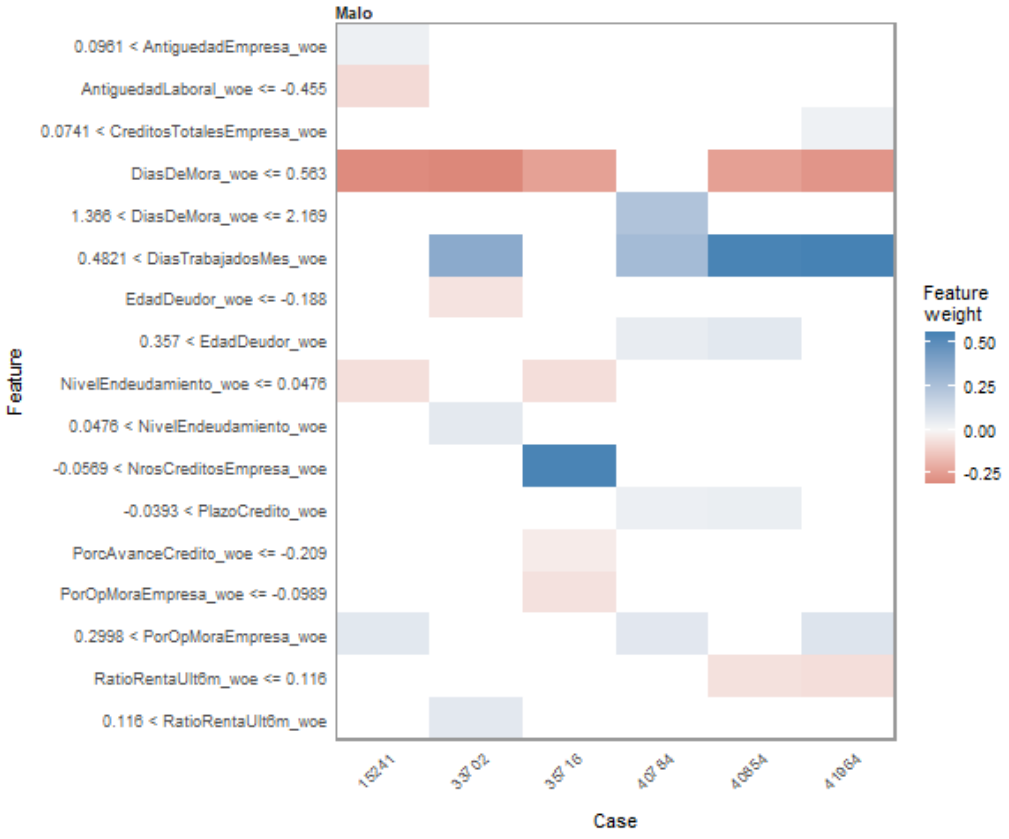


Figura 19: Visualización pesos de las variables en la estimación mediante LIME

3.7. Discusión

Tal como se planteó en el apartado 1.2 de este documento, el objetivo general de este proyecto era responder a la pregunta: ***¿Es posible encontrar un modelo de Machine Learning de caja negra que mejore el desempeño respecto del uso de metodologías clásicas, pudiendo realizar la interpretación de los atributos que determinan el incumplimiento crediticio?***, considerando la comparación de modelos de Machine Learning clásicos que no requieren ajustes complejos y son interpretables, respecto de modelos de caja negra.

El resultado de los análisis que se realizaron logró determinar que los modelos de caja negra fueron mejores en cuanto a las métricas de desempeño definidas para comparación (Accuracy, KS y AUC). Particularmente, el modelo Random Forest seleccionado obtuvo resultados muy superiores en las métricas de discriminación KS y AUC, pero similares a otros modelos considerando Accuracy. Sin embargo, el desafío que tiene el uso de estas metodologías más sofisticadas es la interpretación de los parámetros, logrando en este caso ajustar un modelo simplificado, mediante la metodología LIME, que permitió interpretar localmente algunos casos seleccionados aleatoriamente, con buenos resultados, lo que puede permitir romper el paradigma del uso de metodologías sofisticadas de Machine Learning que permitan realizar la estimación del incumplimiento crediticio en la industria financiera chilena, cumpliendo así con la normativa local que establece que los atributos y criterios que determinan la clasificación crediticia de un cliente deben ser detallados.

Finalmente y, luego de todos los análisis, se logró responder de manera satisfactoria a la pregunta de investigación, obteniendo resultados robustos, consistentes y significativos.

CAPÍTULO 4: CONCLUSIONES

Al término de esta investigación de llegó a las siguientes conclusiones:

i. La variable objetivo consistió en la marca de incumplimiento, correspondiendo a un problema de clasificación, con una tasa de incumplimiento de un 11.48%.

ii. Luego del análisis de datos perdidos, análisis descriptivos, se realizó un muestreo balanceado, utilizando la metodología de submuestreo aleatorio. La validación de los resultados se realizó sobre la base completa.

iii. Posteriormente se realizó la transformación de variables vía WOE y selección de variables iniciales por medio de correlaciones de Spearman, donde se obtuvieron 21 variables predictoras.

iv. Estas 21 variables permitieron ajustar 6 modelos de Machine Learning, considerando la optimización de hiperparámetros mediante Grid Search y validación cruzada K-fold, calculando en cada uno de los ajustes las métricas de Accuracy (precisión), KS y AUC (discriminación).

v. Para seleccionar el modelo final, se utilizó la combinación de las métricas de precisión y discriminación, seleccionando el modelo de Random Forest, que utilizó como hiperparámetros el ajuste de 200 árboles y la selección de 3 variable en cada paso.

vi. El resultado en la base de validación para las métricas de Accuracy, KS y AUC fueron un 77,03%, 62,21% y 89,28%, respectivamente, con un nivel de sobreajuste explicado porque la base de entrenamiento se encuentra ordenada, de acuerdo con la clasificación real de los clientes.

vii. La variable más importante correspondió a los días de mora en el periodo de observación, seguida de la antigüedad laboral del cliente, el porcentaje de avance del crédito y la edad. La variable menos importante correspondió al estado civil del cliente.

viii. Por medio de la metodología LIME, se pudo realizar la interpretación local de algunos casos seleccionados aleatoriamente, llegando a la misma clasificación real de los clientes y obteniendo resultados con buena estimación, logrando responder la pregunta de investigación, abriendo una oportunidad de usos de modelos de caja negra.

REFERENCIAS BIBLIOGRÁFICAsS

[1] Comisión para el Mercado Financiero (octubre, 2021). *Compendio de Normas contables*. https://www.cmfchile.cl/portal/principal/613/articles-29177_doc_pdf.pdf

[2] Navas-Palencia, G. (enero, 2020). *Optimal binning: mathematical programming formulation*. https://arxiv.org/pdf/2001.08025v1.pdf

[3] Siddiqi, N. (2006). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring.*

[4] Sun Mi, K. et al. (2018). *Logistic LASSO regression for the diagnosis of breast cancer using clinical demographic data and the BI-RADS lexicon for ultrasonography*. https://www.e-ultrasonography.org/journal/view.php?doi=10.14366/usg.16045

[5] Ferrero, R. (2020). *Introducción a las redes neuronales y Deep Learning*. Máxima Formación S.L.

[6] Ferrero, R. (2020). *Modelos Ensemble: Boosting*. Máxima Formación S.L.

[7] Brownlee, J. (2016). *A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning.* https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/

[8] Patro, R. (2021). *Cross-Validation: K Fold vs Monte Carlo: Choosing the right validation technique.* https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b

[9] Bethapudi, S. y Desaia, S. (2018). *Separation of pulsar signals from noise using supervised machine learning algorithms*. Astronomy and Computing Journal 23, 15. https://arxiv.org/abs/1704.04659

[10] Molnar, C. (2021). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* https://christophm.github.io/interpretable-ml-book/

[11] University of Cincinnati (2018). *Interpreting Machine Learning Models with the iml Package: Visualizing ML Models with LIME.* http://uc-r.github.io/lime

ANEXOS

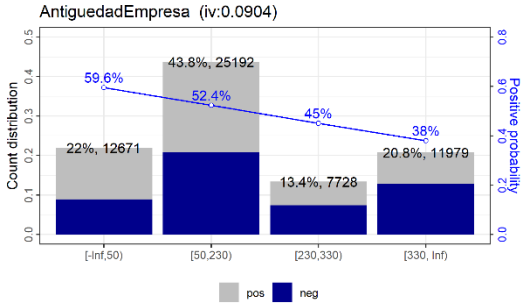
Anexo 1: Análisis descriptivo variables totales

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	1%	5%	95%	99%	sd	CV	skewness	kurtosis
DiasDeMora	0,00	0,00	0,00	2,10	0,00	82,00	0,00	0,00	20,00	52,00	10,27	490,22	5,94	37,72
AntiguedadLaboral	0,00	24,00	50,00	81,76	106,00	768,00	4,00	9,00	270,00	386,00	85,06	104,04	1,91	3,94
AntiguedadEmpresa	0,00	68,00	179,00	207,58	328,00	598,00	5,00	15,00	505,00	588,00	157,68	75,96	0,59	-0,68
PorcAvanceCredito	0,00	0,02	0,19	0,27	0,45	0,98	0,00	0,00	0,81	0,92	0,27	98,76	0,79	-0,52
Capital	215	404.189	807.533	1.197.472	1.521.224	41.337.503	7.730	30.170	3.747.234	6.570.088	1.360.938	113,65	3,74	31,79
CargaFinancieraDeudor	0,00	6,12	7,91	10,78	11,51	2.766,67	0,00	4,24	24,78	53,70	14,64	135,81	41,78	5.726,56
CreditosTotalesEmpresa	1,00	12,00	45,00	331,19	239,00	2.670,00	1,00	3,00	2.130,00	2.591,00	638,86	192,90	2,32	4,14
DiasTrabajadosMes	1,00	30,00	30,00	29,15	30,00	58,00	10,00	25,00	30,00	30,00	3,34	11,44	-5,40	32,02
EdadDeudor	18,00	34,00	43,00	42,81	52,00	118,00	21,00	25,00	61,00	68,00	11,54	26,95	0,17	-0,40
NivelEndeudamiento	0,00	0,62	1,31	109,13	2,48	2.241.829	0,04	0,16	4,96	7,65	11.441	10.483,99	137,04	21.246,85
MaxDiasMoraUlt3m	0,00	0,00	0,00	2,04	0,00	356,00	0,00	0,00	20,00	51,00	12,12	595,01	13,43	285,31
MaxMoraDeudor	0,00	0,00	0,00	3,75	0,00	7.325,00	0,00	0,00	20,00	80,00	78,24	2.086,79	61,72	4.206,35
MaxMoraDeudorUlt3m	0,00	0,00	0,00	7,88	0,00	9.517,00	0,00	0,00	21,00	82,00	123,66	1.569,72	34,52	1.461,20
NrosCreditosEmpresa	1,00	1,00	1,00	1,26	1,00	11,00	1,00	1,00	2,00	4,00	0,70	55,38	3,92	20,28
NroTrabajadoresEmpresa	1	43	163	1.172	879	14.119	4	9	7.381	8.430	2.234	190,56	2,50	5,94
PercentilRenta	0,00	0,27	0,54	0,55	0,82	1,00	0,02	0,07	1,00	1,00	0,31	56,65	-0,01	-1,27
PlazoCredito	1,00	12,00	24,00	24,86	36,00	148,00	2,00	3,00	60,00	60,00	16,64	66,92	0,78	-0,30
PorOpMoraEmpresa	0,00	0,00	0,04	0,08	0,09	1,00	0,00	0,00	0,28	0,80	0,13	176,58	4,18	21,43
PorOpMoraEmpresa30d	0,00	0,00	0,02	0,04	0,05	1,00	0,00	0,00	0,15	0,50	0,09	212,57	6,13	50,69
RatioRentaUlt6m	0,00	0,95	1,03	24,08	1,14	982.773,00	0,00	0,61	1,69	3,26	3.541,47	14.708,41	187,19	40.603,58
RentaDeudor	0	353.811	512.122	623.625	780.830	4.311.636	0	220.005	1.485.133	1.996.474	395.350	63,40	1,50	2,48
MontoCuota	38	33.816	52.721	64.471	81.070	1.282.313	3.599	10.222	155.324	240.007	49.244	76,38	2,90	21,61

Tabla 6 : Análisis descriptivo variables continuas – Base Total

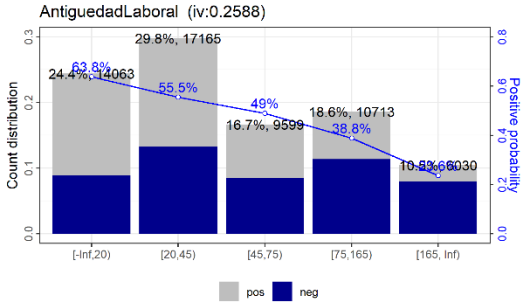
Anexo 2: Transformación de variables. WOE e IV

Variable: Antigüedad Empresa



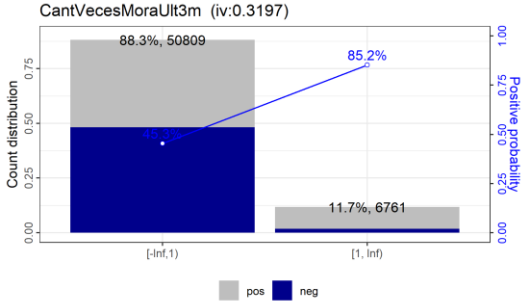
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,4878	11.979	20,80%	7.422	15,80%	4.557	25,80%	38,00%	0,0485	20,80%
-0,1989	7.728	13,40%	4.247	12,10%	3.481	14,80%	45,00%	0,0053	13,40%
0,0961	25.192	43,80%	11.991	45,90%	13.201	41,70%	52,40%	0,004	43,80%
0,3869	12.671	22,00%	5.125	26,20%	7.546	17,80%	59,60%	0,0325	22,00%

Variable: Antigüedad Laboral



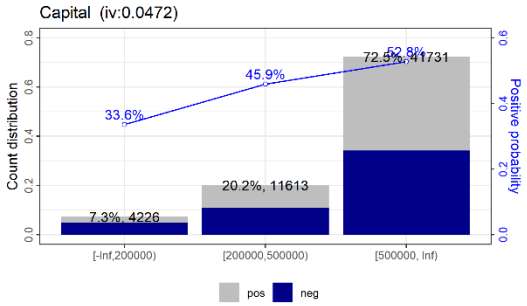
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-1,1739	6.030	10,50%	4.606	4,90%	1.424	16,00%	23,60%	0,1298	10,50%
-0,4552	10.713	18,60%	6.555	14,40%	4.158	22,80%	38,80%	0,0379	18,60%
-0,0415	9.599	16,70%	4.899	16,30%	4.700	17,00%	49,00%	0,0003	16,70%
0,2226	17.165	29,80%	7.631	33,10%	9.534	26,50%	55,50%	0,0147	29,80%
0,5657	14.063	24,40%	5.094	31,20%	8.969	17,70%	63,80%	0,0762	24,40%

Variable: Cantidad veces en mora en los últimos 3 meses



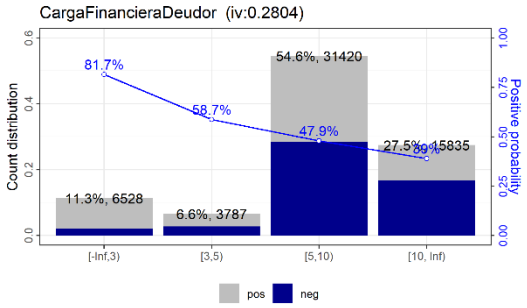
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,1877	50.809	88,30%	27.782	80,00%	23.027	96,50%	45,30%	0,031	88,30%
1,7476	6.761	11,70%	1.003	20,00%	5.758	3,50%	85,20%	0,2887	11,70%

Variable: Capital



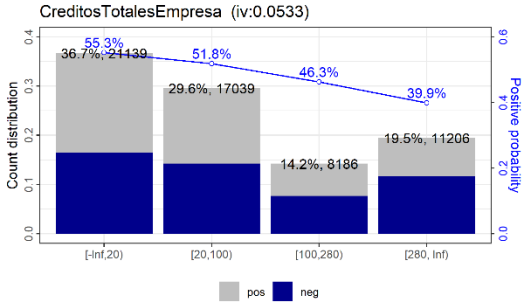
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,68	4.226	7,30%	2.805	4,90%	1.421	9,70%	33,60%	0,0327	7,30%
-0,1638	11.613	20,20%	6.281	18,50%	5.332	21,80%	45,90%	0,0054	20,20%
0,1119	41.731	72,50%	19.699	76,50%	22.032	68,40%	52,80%	0,0091	72,50%

Variable: Carga Financiera Deudor



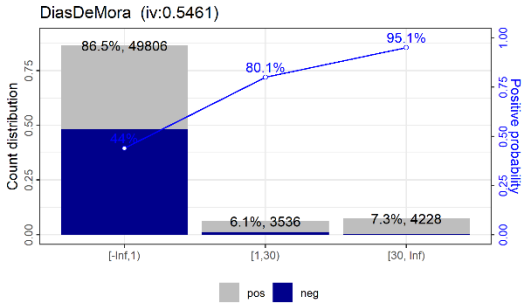
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,4485	15.835	27,50%	9.664	21,40%	6.171	33,60%	39,00%	0,0544	27,50%
-0,0825	31.420	54,60%	16.358	52,30%	15.062	56,80%	47,90%	0,0037	54,60%
0,3505	3.787	6,60%	1.565	7,70%	2.222	5,40%	58,70%	0,008	6,60%
1,4927	6.528	11,30%	1.198	18,50%	5.330	4,20%	81,60%	0,2143	11,30%

Variable: Créditos Totales Empresa



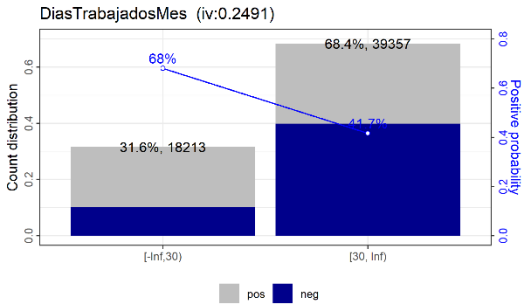
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,4082	11.206	19,50%	6.731	15,50%	4.475	23,40%	39,90%	0,032	19,50%
-0,1493	8.186	14,20%	4.398	13,20%	3.788	15,30%	46,30%	0,0032	14,20%
0,0741	17.039	29,60%	8.204	30,70%	8.835	28,50%	51,90%	0,0016	29,60%
0,2123	21.139	36,70%	9.452	40,60%	11.687	32,80%	55,30%	0,0165	36,70%

Variable: Días de Mora



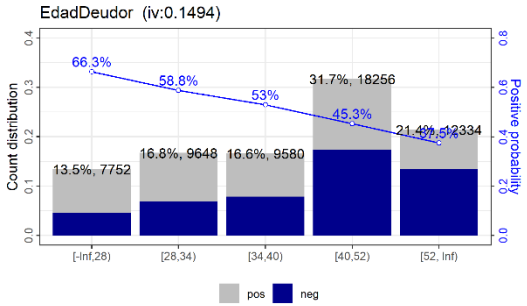
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,2397	49.806	86,50%	27.874	76,20%	21.932	96,80%	44,00%	0,0495	86,50%
1,3902	3.536	6,10%	705	9,80%	2.831	2,40%	80,10%	0,1027	6,10%
2,9717	4.228	7,30%	206	14,00%	4.022	0,70%	95,10%	0,3939	7,30%

Variable: Días trabajados en el mes



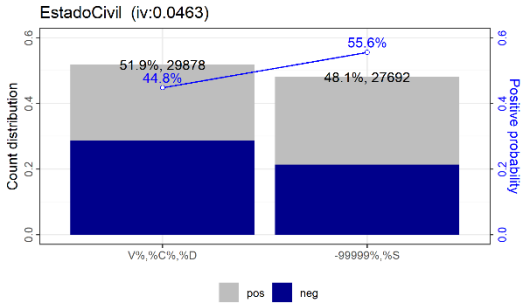
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,3369	39.357	68,40%	22.962	57,00%	16.395	79,80%	41,70%	0,0769	68,40%
0,7551	18.213	31,60%	5.823	43,00%	12.390	20,20%	68,00%	0,1723	31,60%

Variable: Edad del deudor



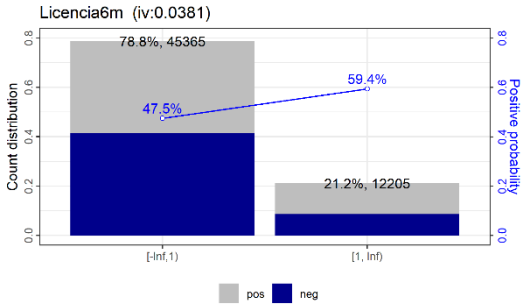
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,5123	12.334	21,40%	7.713	16,10%	4.621	26,80%	37,50%	0,055	21,40%
-0,1883	18.256	31,70%	9.985	28,70%	8.271	34,70%	45,30%	0,0112	31,70%
0,1187	9.580	16,60%	4.506	17,60%	5.074	15,70%	53,00%	0,0023	16,60%
0,3574	9.648	16,80%	3.971	19,70%	5.677	13,80%	58,80%	0,0212	16,80%
0,6781	7.752	13,50%	2.610	17,90%	5.142	9,10%	66,30%	0,0596	13,50%

Variable: Estado Civil cliente



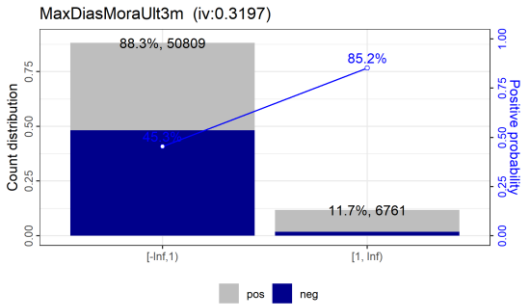
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,2074	29.878	51,90%	16.483	46,50%	13.395	57,30%	44,80%	0,0223	51,90%
0,224	27.692	48,10%	12.302	53,50%	15.390	42,70%	55,60%	0,024	48,10%

Variable: Licencia médica últimos 6 meses



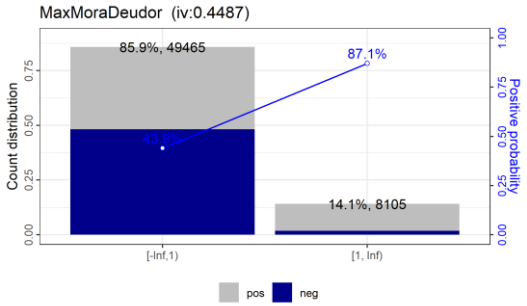
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,1008	45.365	78,80%	23.825	74,80%	21.540	82,80%	47,50%	0,008	78,80%
0,3789	12.205	21,20%	4.960	25,20%	7.245	17,20%	59,40%	0,0301	21,20%

Variable: Máximo días de mora últimos 3 meses



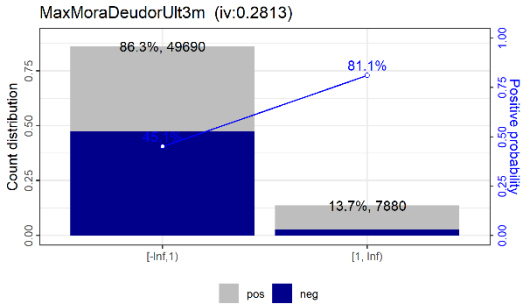
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,1877	50.809	88,30%	27.782	80,00%	23.027	96,50%	45,30%	0,031	88,30%
1,7476	6.761	11,70%	1.003	20,00%	5.758	3,50%	85,20%	0,2887	11,70%

Variable: Máxima mora deudor



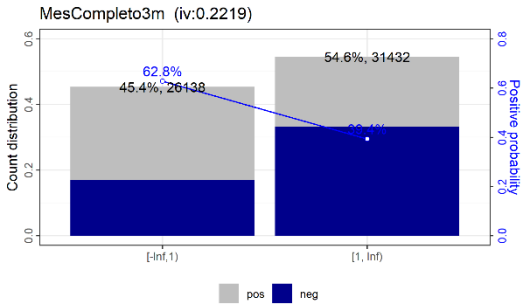
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,2441	49.465	85,90%	27.736	75,50%	21.729	96,40%	43,90%	0,0509	85,90%
1,906	8.105	14,10%	1.049	24,50%	7.056	3,60%	87,10%	0,3978	14,10%

Variable: Máxima mora deudor últimos 3 meses



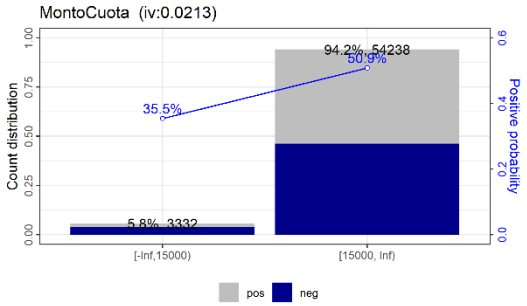
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,1978	49.690	86,30%	27.294	77,80%	22.396	94,80%	45,10%	0,0337	86,30%
1,4551	7.880	13,70%	1.491	22,20%	6.389	5,20%	81,10%	0,2476	13,70%

Variable: Indicador de trabajo mes completo



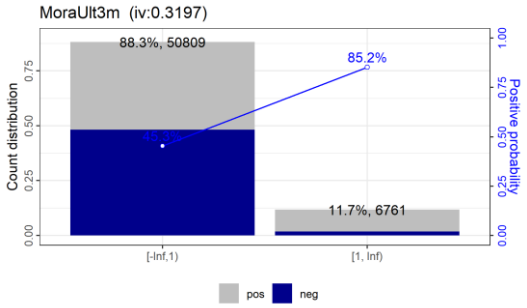
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,432	31.432	54,60%	19.059	43,00%	12.373	66,20%	39,40%	0,1003	54,60%
0,5232	26.138	45,40%	9.726	57,00%	16.412	33,80%	62,80%	0,1215	45,40%

Variable: Monto de la cuota



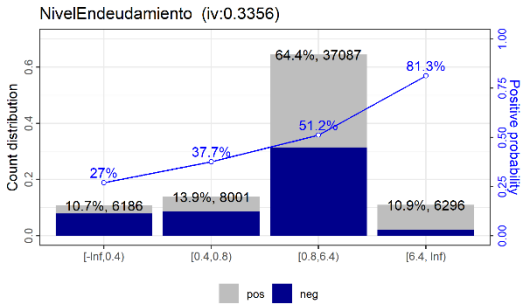
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,5983	3.332	5,80%	2.150	4,10%	1.182	7,50%	35,50%	0,0201	5,80%
0,0357	54.238	94,20%	26.635	95,90%	27.603	92,50%	50,90%	0,0012	94,20%

Variable: Mora Crédito últimos 3 meses



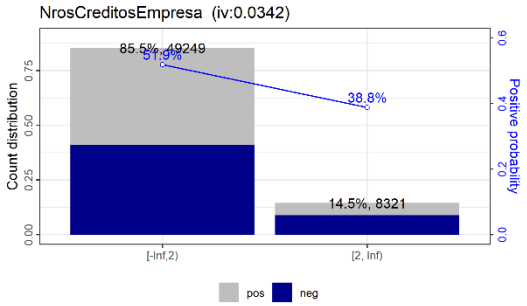
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,1877	50.809	88,30%	27.782	80,00%	23.027	96,50%	45,30%	0,031	88,30%
1,7476	6.761	11,70%	1.003	20,00%	5.758	3,50%	85,20%	0,2887	11,70%

Variable: Nivel de endeudamiento del cliente



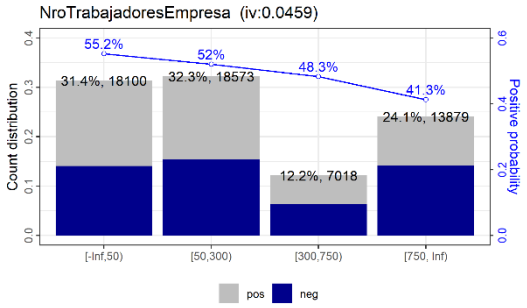
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,9956	6.186	10,70%	4.517	5,80%	1.669	15,70%	27,00%	0,0985	10,70%
-0,5041	8.001	13,90%	4.988	10,50%	3.013	17,30%	37,70%	0,0346	13,90%
0,0476	37.087	64,40%	18.102	66,00%	18.985	62,90%	51,20%	0,0015	64,40%
1,4689	6.296	10,90%	1.178	17,80%	5.118	4,10%	81,30%	0,2011	10,90%

Variable: Número de créditos empresa



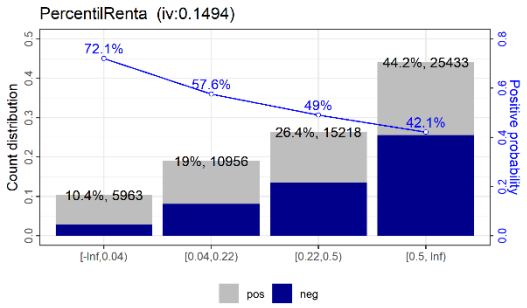
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,454	8.321	14,50%	5.089	11,20%	3.232	17,70%	38,80%	0,0293	14,50%
0,0754	49.249	85,50%	23.696	88,80%	25.553	82,30%	51,90%	0,0049	85,50%

Variable: Número trabajadores empresa



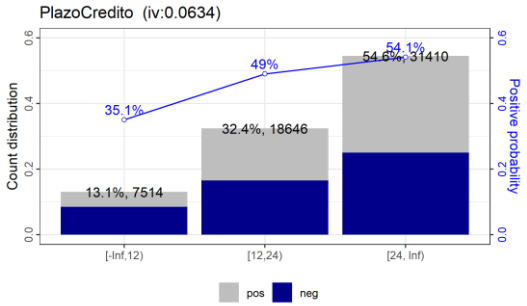
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,351	13.879	24,10%	8.145	19,90%	5.734	28,30%	41,30%	0,0294	24,10%
-0,069	7.018	12,20%	3.630	11,80%	3.388	12,60%	48,30%	0,0006	12,20%
0,082	18.573	32,30%	8.906	33,60%	9.667	30,90%	52,00%	0,0022	32,30%
0,2098	18.100	31,40%	8.104	34,70%	9.996	28,20%	55,20%	0,0138	31,40%

Variable: Percentil Renta Cliente



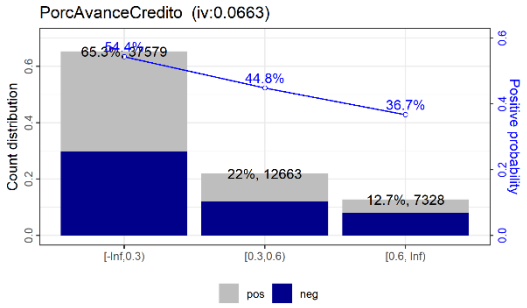
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,2801	27.450	47,70%	15.635	41,00%	11.815	54,30%	43,00%	0,0372	47,70%
-0,073	13.206	22,90%	6.844	22,10%	6.362	23,80%	48,20%	0,0012	22,90%
0,3094	10.986	19,10%	4.650	22,00%	6.336	16,20%	57,70%	0,0181	19,10%
0,9477	5.928	10,30%	1.656	14,80%	4.272	5,80%	72,10%	0,0861	10,30%

Variable: Plazo del crédito



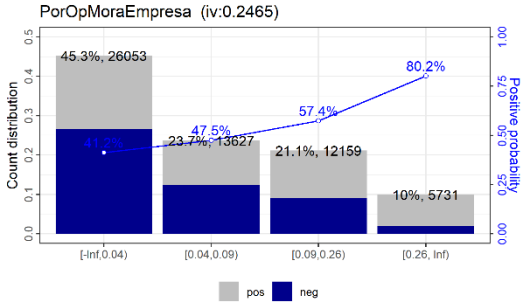
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,6149	7.514	13,10%	4.877	9,20%	2.637	16,90%	35,10%	0,0478	13,10%
-0,0393	18.646	32,40%	9.506	31,80%	9.140	33,00%	49,00%	0,0005	32,40%
0,1663	31.410	54,60%	14.402	59,10%	17.008	50,00%	54,10%	0,0151	54,60%

Variable: Porcentaje de avance del crédito



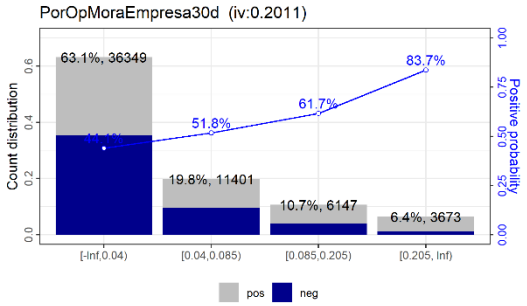
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,5453	7.328	12,70%	4.639	9,30%	2.689	16,10%	36,70%	0,0369	12,70%
-0,2088	12.663	22,00%	6.990	19,70%	5.673	24,30%	44,80%	0,0096	22,00%
0,1743	37.579	65,30%	17.156	71,00%	20.423	59,60%	54,30%	0,0198	65,30%

Variable: Porcentaje de operaciones de empresa con mora



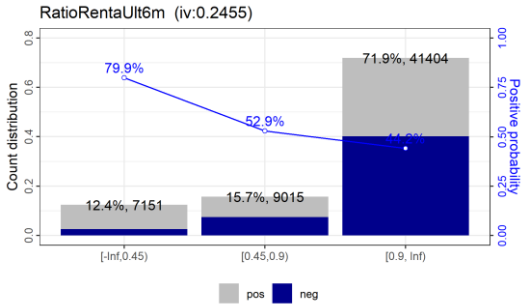
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,3586	25.799	44,80%	15.188	36,90%	10.611	52,80%	41,10%	0,057	44,80%
-0,1003	13.865	24,10%	7.280	22,90%	6.585	25,30%	47,50%	0,0024	24,10%
0,2996	12.174	21,10%	5.182	24,30%	6.992	18,00%	57,40%	0,0188	21,10%
1,3988	5.732	10,00%	1.135	16,00%	4.597	3,90%	80,20%	0,1682	10,00%

Variable: Porcentaje de operaciones de empresa con 30+ días de mora



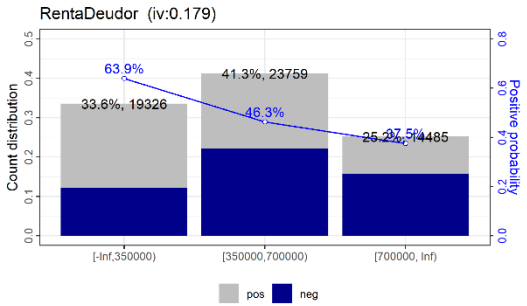
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,2397	36.178	62,80%	20.247	55,30%	15.931	70,30%	44,00%	0,0359	62,80%
0,0705	11.572	20,10%	5.582	20,80%	5.990	19,40%	51,80%	0,001	20,10%
0,475	6.147	10,70%	2.357	13,20%	3.790	8,20%	61,70%	0,0236	10,70%
1,6355	3.673	6,40%	599	10,70%	3.074	2,10%	83,70%	0,1406	6,40%

Variable: Ratio ingresos deudor últimos 6 meses



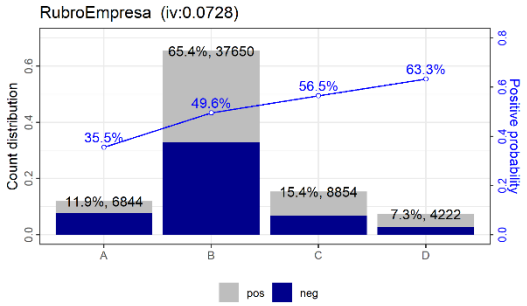
WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,2327	41.450	72,00%	23.126	63,70%	18.324	80,30%	44,20%	0,0388	72,00%
0,1181	8.969	15,60%	4.220	16,50%	4.749	14,70%	52,90%	0,0022	15,60%
1,3786	7.151	12,40%	1.439	19,80%	5.712	5,00%	79,90%	0,2046	12,40%

Variable: Ingresos mensuales deudor



WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,5087	14.485	25,20%	9.046	18,90%	5.439	31,40%	37,50%	0,0637	25,20%
-0,1482	23.759	41,30%	12.758	38,20%	11.001	44,30%	46,30%	0,009	41,30%
0,5701	19.326	33,60%	6.981	42,90%	12.345	24,30%	63,90%	0,1062	33,60%

Variable: Rubro del a empresa



WOE	Casos	Distr.Total	Buenos	Distr.Buenos	Malos	Distr.Malos	Bad.Rate	IV	Part(%)
-0,595	6.844	11,90%	4.411	8,50%	2.433	15,30%	35,50%	0,0409	11,90%
-0,0155	37.650	65,40%	18.971	64,90%	18.679	65,90%	49,60%	0,0002	65,40%
0,2599	8.854	15,40%	3.855	17,40%	4.999	13,40%	56,50%	0,0103	15,40%
0,5466	4.222	7,30%	1.548	9,30%	2.674	5,40%	63,30%	0,0214	7,30%

NOTA: La variable Rubro se caracteriza de la siguiente manera:

Categoría	Rubro	Categoría	Rubro
A	ACTIVIDADES INMOBILIARIAS	B	COMERCIO AL POR MAYOR Y AL POR MENOR, REPARACION DE VEHICULOS AUTOMOTORES Y MOTOCICLETAS
	ADMINISTRACION PUBLICA Y DEFENSA, PLANES DE SEGURIDAD SOCIAL DE AFILIACION OBLIGATORIA		EXPLOTACION DE MINAS Y CANTERAS
	ENSENANZA		INDUSTRIA MANUFACTURERA
	SUMINISTRO DE ELECTRICIDAD, GAS, VAPOR Y AIRE ACONDICIONADO		OTRAS ACTIVIDADES DE SERVICIOS
B	SIN INFORMACIÓN		SUMINISTRO DE AGUA, EVACUACION DE AGUAS RESIDUALES, GESTION DE DESECHOS Y DESCONTAMINACION
	ACTIVIDADES ARTISTICAS, DE ENTRETENIMIENTO Y RECREATIVAS	C	ACTIVIDADES PROFESIONALES, CIENTIFICAS Y TECNICAS
	ACTIVIDADES DE ALOJAMIENTO Y DE SERVICIO DE COMIDAS		INDETERMINADO
	ACTIVIDADES DE ATENCION DE LA SALUD HUMANA Y DE ASISTENCIA SOCIAL		INFORMACION Y COMUNICACIONES
	ACTIVIDADES DE ORGANIZACIONES Y ORGANOS EXTRATERRITORIALES		TRANSPORTE Y ALMACENAMIENTO
	ACTIVIDADES FINANCIERAS Y DE SEGUROS	D	ACTIVIDADES DE SERVICIOS ADMINISTRATIVOS Y DE APOYO
	AGRICULTURA, GANADERIA, SILVICULTURA Y PESCA		CONSTRUCCION

Tabla 7 : Detalle variable Rubro

Anexo 3: Análisis de correlaciones variables WOE

Correlaciones	DiasDeMora_woe	AntiguedadLaboral_woe	AntiguedadEmpresa_woe	PorcAvanceCredito_woe	Capital_woe	CargaFinancieraDeudor_woe	CreditosTotalesEmpresa_woe	DiasTrabajadosMes_woe	EdadDeudor_woe	NivelEndeudamiento_woe	MaxDiasMoraUlt3m_woe	MaxMoraDeudor_woe	MaxMoraDeudorUlt3m_woe	NrosCreditosEmpresa_woe	NroTrabajadoresEmpresa_woe	PercentilRenta_woe	PlazoCredito_woe	PorOpMoraEmpresa_woe	PorOpMoraEmpresa30d_woe	RatioRentaUlt6m_woe	RentaDeudor_woe	MontoCuota_woe	CantVecesMoraUlt3m_woe	Licencia6m_woe	MesCompleto3m_woe	MoraUlt3m_woe	EstadoCivil_woe	RubroEmpresa_woe
DiasDeMora_woe	1,000	0,060	0,116	-0,079	0,002	0,245	0,111	0,217	0,022	0,222	0,803	0,973	0,742	0,011	0,117	0,173	0,020	0,458	0,361	0,266	0,217	-0,030	0,803	-0,015	0,187	0,803	0,005	0,053
AntiguedadLaboral_woe	0,060	1,000	0,234	0,158	-0,165	-0,003	-0,023	0,098	0,344	-0,004	0,036	0,058	0,040	0,113	-0,037	0,171	-0,237	0,076	0,067	0,053	0,194	-0,015	0,036	0,025	0,130	0,036	0,188	0,141
AntiguedadEmpresa_woe	0,116	0,234	1,000	0,000	-0,046	0,049	0,233	0,083	0,044	0,030	0,112	0,116	0,111	0,068	0,220	0,025	-0,083	0,178	0,136	0,053	0,144	-0,013	0,112	-0,037	0,105	0,112	0,014	0,180
PorcAvanceCredito_woe	-0,079	0,158	0,000	1,000	0,047	0,038	-0,035	0,001	0,101	0,286	-0,120	-0,074	-0,082	-0,035	-0,030	0,004	0,080	-0,041	-0,047	-0,020	-0,064	-0,041	-0,120	0,005	0,009	-0,120	0,060	-0,019
Capital_woe	0,002	-0,165	-0,046	0,047	1,000	0,222	0,115	-0,030	-0,036	0,372	0,023	-0,009	0,007	0,235	0,122	-0,109	0,639	0,011	-0,001	0,007	-0,208	0,483	0,023	0,010	-0,025	0,023	-0,009	0,005
CargaFinancieraDeudor_woe	0,245	-0,003	0,049	0,038	0,222	1,000	0,130	0,366	0,020	0,678	0,189	0,244	0,178	-0,088	0,162	0,385	0,172	0,108	0,087	0,484	0,419	0,092	0,189	0,063	0,221	0,189	0,016	-0,022
CreditosTotalesEmpresa_woe	0,111	-0,023	0,233	-0,035	0,115	0,130	1,000	0,015	-0,063	0,088	0,115	0,112	0,112	0,063	0,894	-0,122	0,041	0,052	0,014	-0,022	0,164	0,015	0,115	-0,104	0,010	0,115	-0,029	0,121
DiasTrabajadosMes_woe	0,217	0,098	0,083	0,001	-0,030	0,366	0,015	1,000	0,120	0,323	0,157	0,213	0,147	0,026	0,031	0,470	0,002	0,131	0,116	0,468	0,357	-0,030	0,157	0,147	0,513	0,157	0,072	-0,002
EdadDeudor_woe	0,022	0,344	0,044	0,101	-0,036	0,020	-0,063	0,120	1,000	0,036	0,000	0,023	-0,001	0,052	-0,058	0,128	-0,078	0,028	0,031	0,064	0,068	-0,013	0,000	0,072	0,139	0,000	0,525	0,020
NivelEndeudamiento_woe	0,222	-0,004	0,030	0,286	0,372	0,678	0,088	0,323	0,036	1,000	0,160	0,215	0,153	-0,004	0,122	0,320	0,391	0,097	0,078	0,437	0,274	0,157	0,160	0,072	0,213	0,160	0,021	-0,028
MaxDiasMoraUlt3m_woe	0,803	0,036	0,112	-0,120	0,023	0,189	0,115	0,157	0,000	0,160	1,000	0,767	0,916	0,022	0,114	0,118	0,041	0,386	0,352	0,195	0,173	-0,006	1,000	-0,019	0,159	1,000	-0,005	0,052
MaxMoraDeudor_woe	0,973	0,058	0,116	-0,074	-0,009	0,244	0,112	0,213	0,023	0,215	0,767	1,000	0,737	-0,031	0,118	0,169	0,012	0,463	0,344	0,262	0,210	-0,049	0,767	-0,015	0,183	0,767	0,003	0,050
MaxMoraDeudorUlt3m_woe	0,742	0,040	0,111	-0,082	0,007	0,178	0,112	0,147	-0,001	0,153	0,916	0,737	1,000	-0,008	0,110	0,107	0,027	0,376	0,339	0,180	0,161	-0,023	0,916	-0,021	0,149	0,916	-0,008	0,048
NrosCreditosEmpresa_woe	0,011	0,113	0,068	-0,035	0,235	-0,088	0,063	0,026	0,052	-0,004	0,022	-0,031	-0,008	1,000	0,029	0,052	0,104	0,018	0,016	0,012	0,145	0,394	0,022	0,014	0,043	0,022	0,055	0,067
NroTrabajadoresEmpresa_woe	0,117	-0,037	0,220	-0,030	0,122	0,162	0,894	0,031	-0,058	0,122	0,114	0,118	0,110	0,029	1,000	-0,098	0,063	0,042	-0,001	-0,004	0,175	-0,002	0,114	-0,108	0,014	0,114	-0,027	0,104
PercentilRenta_woe	0,173	0,171	0,025	0,004	-0,109	0,385	-0,122	0,470	0,128	0,320	0,118	0,169	0,107	0,052	-0,098	1,000	-0,052	0,078	0,071	0,444	0,511	-0,036	0,118	0,093	0,327	0,118	0,089	-0,052
PlazoCredito_woe	0,020	-0,237	-0,083	0,080	0,639	0,172	0,041	0,002	-0,078	0,391	0,041	0,012	0,027	0,104	0,063	-0,052	1,000	0,019	0,016	0,022	-0,106	0,313	0,041	0,031	-0,002	0,041	-0,051	-0,045
PorOpMoraEmpresa_woe	0,458	0,076	0,178	-0,041	0,011	0,108	0,052	0,131	0,028	0,097	0,386	0,463	0,376	0,018	0,042	0,078	0,019	1,000	0,750	0,114	0,122	-0,016	0,386	-0,047	0,119	0,386	0,009	0,083
PorOpMoraEmpresa30d_woe	0,361	0,067	0,136	-0,047	-0,001	0,087	0,014	0,116	0,031	0,078	0,352	0,344	0,339	0,016	-0,001	0,071	0,016	0,750	1,000	0,102	0,108	-0,013	0,352	-0,032	0,114	0,352	0,010	0,067
RatioRentaUlt6m_woe	0,266	0,053	0,053	-0,020	0,007	0,484	-0,022	0,468	0,064	0,437	0,195	0,262	0,180	0,012	-0,004	0,444	0,022	0,114	0,102	1,000	0,352	-0,014	0,195	0,121	0,291	0,195	0,030	0,001
RentaDeudor_woe	0,217	0,194	0,144	-0,064	-0,208	0,419	0,164	0,357	0,068	0,274	0,173	0,210	0,161	0,145	0,175	0,511	-0,106	0,122	0,108	0,352	1,000	-0,030	0,173	0,058	0,268	0,173	0,059	0,026
MontoCuota_woe	-0,030	-0,015	-0,013	-0,041	0,483	0,092	0,015	-0,030	-0,013	0,157	-0,006	-0,049	-0,023	0,394	-0,002	-0,036	0,313	-0,016	-0,013	-0,014	-0,030	1,000	-0,006	0,021	-0,011	-0,006	0,014	0,016

Tabla 8 : Análisis de correlaciones variables WOE – Base Training

Anexo 4: Definición métricas

Para el cálculo de las métricas se trabaja en base al análisis de la variable objetivo del tipo dicotómica (éxito/fracaso), contrastando cada partición. A continuación, se muestra un ejemplo que permitirá entender los cálculos siguientes:

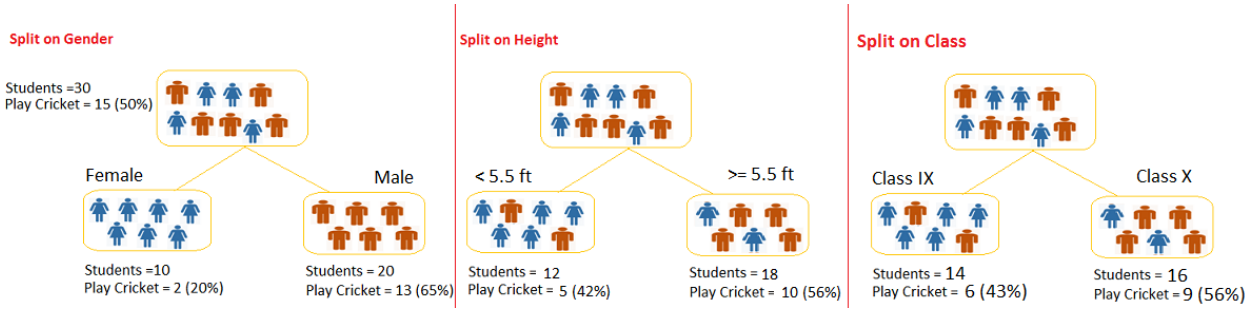


Figura 20: Ejemplo para cálculo de métricas (Fuente)

a. Índice de Gini
   - En cada subnodo se calcula la suma de cuadrados de probabilidad para éxito o fracaso, mediante la siguiente relación:

$$gini = p^2 + (1-p)^2 \qquad ecu(12)$$

   - Considerando Split on Class de la Figura 20, tenemos que:
     Gini Class IV → $(6/14)^2 + (8/14)^2 = 0.1837 + 0.3265 = 0.5102$
     Gini Class X → $(9/16)^2 + (7/16)^2 = 0.3164 + 0.1914 = 0.5078$
   - El índice de Gini corresponde al promedio ponderado de la partición, es decir:

$$índice\ de\ Gini = \left(\frac{14}{30}\right)(0.5102) + \left(\frac{16}{30}\right)(0.5078) \ = \ 0.5089$$

   - A mayor valor del índice de Gini, hay mayor homogeneidad.

b. Chi-Cuadrado
   - Cálculo del Chi-Cuadrado de cada nodo individualmente a través de la desviación de éxito/fracaso, mediante la fórmula:

$$\sqrt{\frac{(real - esperado)^2}{esperado}} \qquad ecu(13)$$

   - Considerando "Split on Gender" de la Figura 20, el análisis se puede realizar mediante una tabla de $2x2$:

	Real		Esperado		Desviación		Chi-Cuadrado	
Sexo	Juega	No Juega	Juega	No Juega	Juega	No Juega	Juega	No Juega
Mujer	6	8	7	7	-1	1	0,3780	0,3780
Hombre	9	7	8	8	1	-1	0,3536	0,3536

- El valor de Chi-Cuadrado corresponde a la suma de los valores individuales, es decir

$$Chi - Cuadrado = (0.3780) + (0.3780) + (0.3536) + (0.3536) = (1.4630)$$

- A mayor valor de Chi-Cuadrado, más alta la significancia estadística.

c. Entropía

- Mide el grado de desorganización de una agrupación, mediante la fórmula:

$$entropía = -p \cdot log_2(p) - (1-p) \cdot log_2(1-p) \qquad ecu(14)$$

- Considerando "Split on Gender" de la Figura 20, se tiene que:
  Entropía Female → $-(2/10)\log_2(2/10) - (8/10)\log_2(8/10) = 0.7219$
  Entropía Male → $-(13/20)\log_2(13/20) - (7/20)\log_2(7/20) = 0.9341$
- La entropía corresponde al promedio ponderado de la partición, es decir:

$$Entropía = \left(\frac{10}{30}\right)(0.7219) + \left(\frac{20}{30}\right)(0.9341) = 0.8634$$

Anexo 5: Resultados Modelos

En el proceso analítico, se ajustaron distintos modelos de Machine Learning, tanto iniciales, como con resultado de ajustes de hiperparámetros, los cuales se muestran a continuación:

Regresión Logística – Modelo inicial

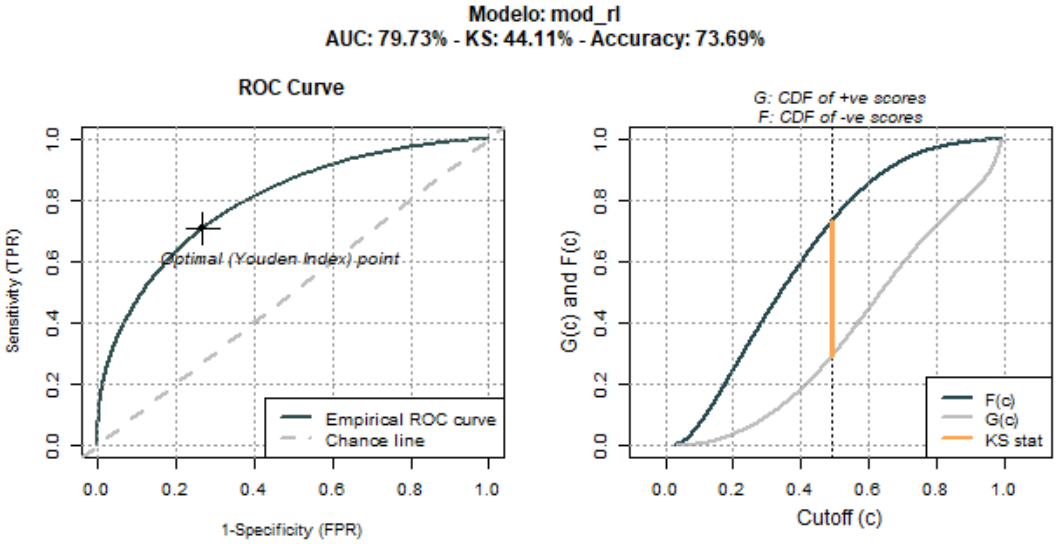


Figura 21: Desempeño Modelo Regresión Logística – Base Test

Regresión Logística – Stepwise (selección de variables significativas)

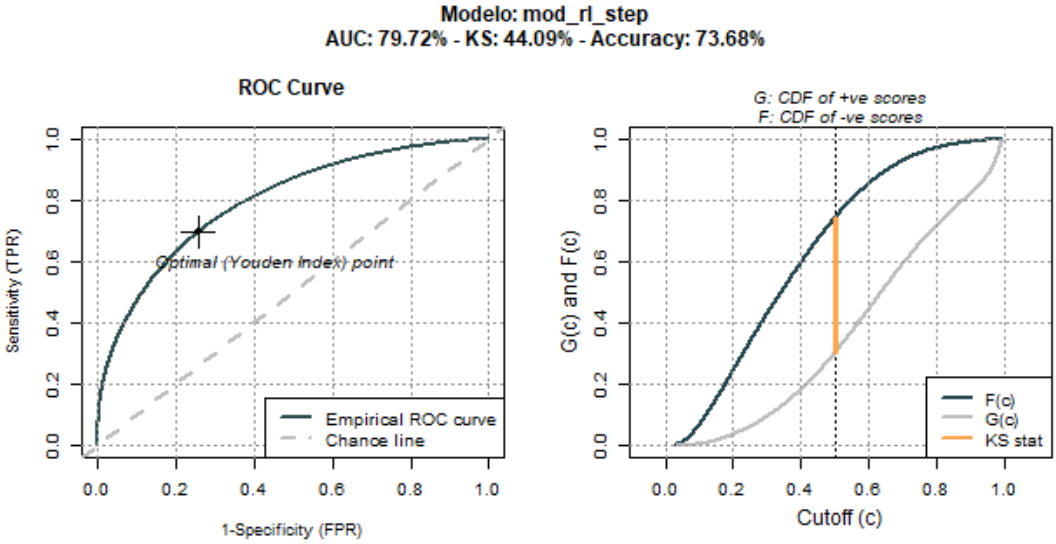


Figura 22: Desempeño Modelo Regresión Logística variables significativas – Base Test

Regresión Logística Lasso

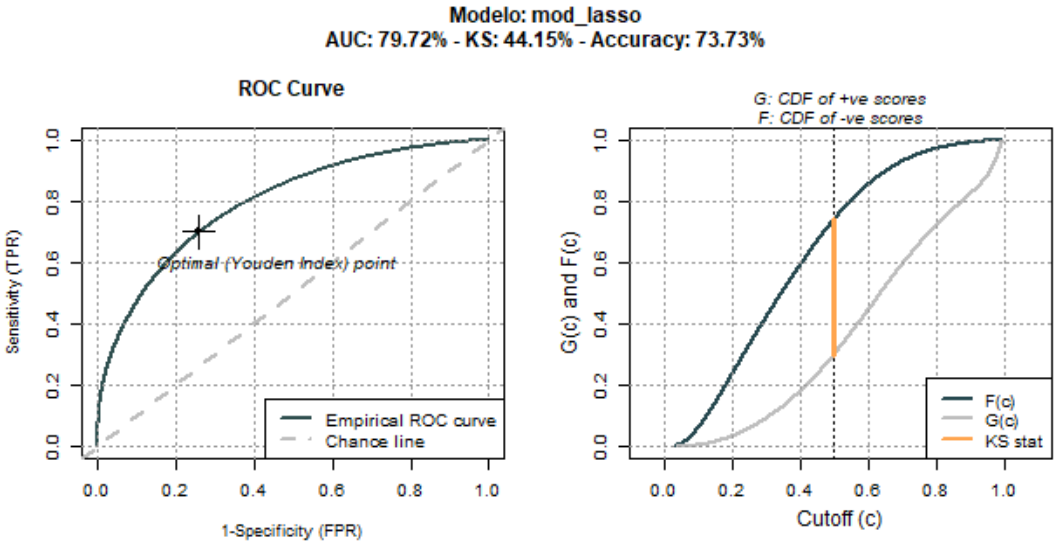


Figura 23: Desempeño Modelo Regresión Logística Lasso – Base Test

Árbol de Clasificación, partición recursiva

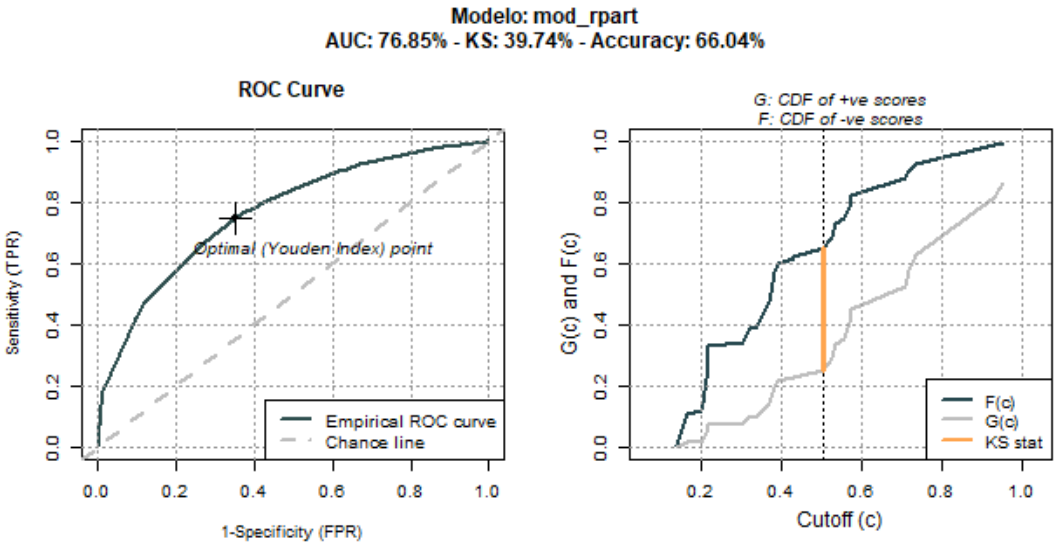


Figura 24: Desempeño Modelo Árbol Clasificación – Base Test

Árbol de Clasificación, partición recursiva optimizado

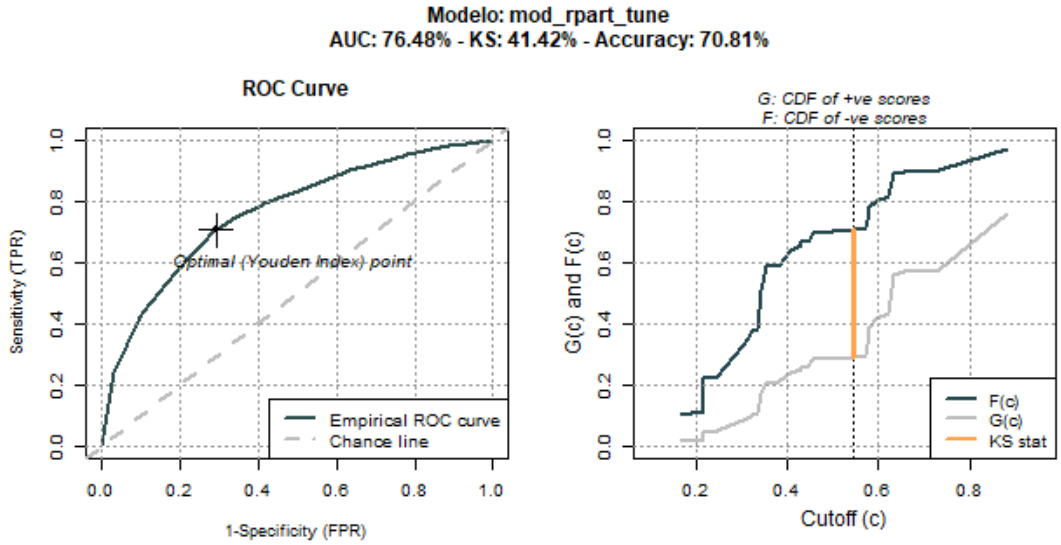


Figura 25: Desempeño Modelo Árbol Clasificación Optimizado – Base Test

Redes neuronales

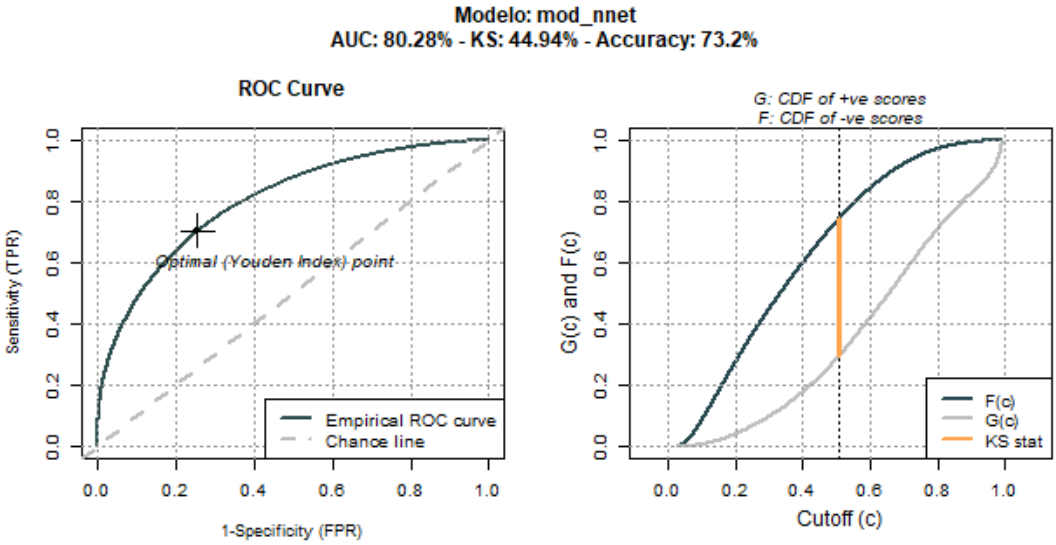


Figura 26: Desempeño Modelo Redes Neuronales – Base Test

Extreme Gradient Boosting – Versión Inicial

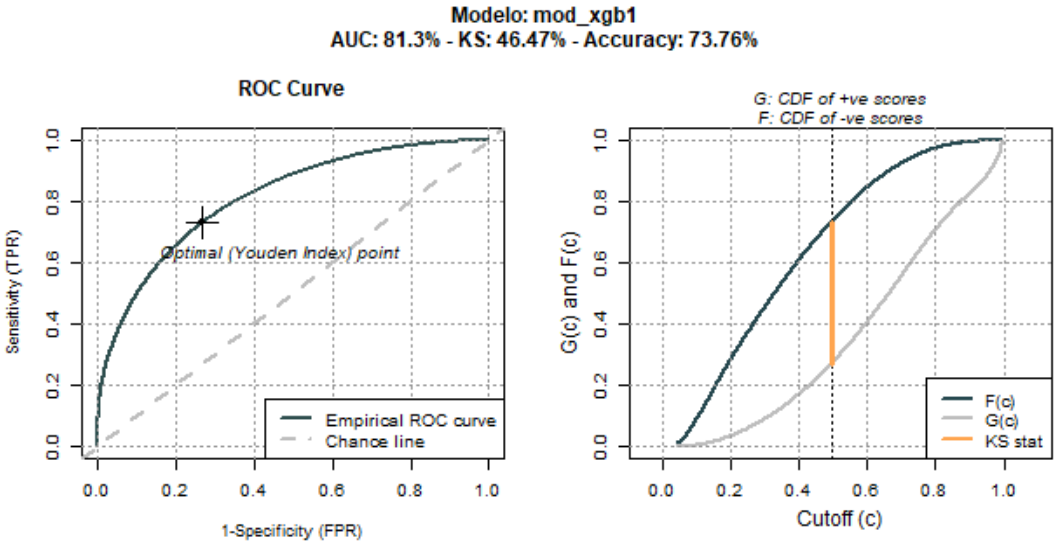


Figura 27: Desempeño Modelo XGBoost Inicial – Base Test

Extreme Gradient Boosting – Versión Optimizada

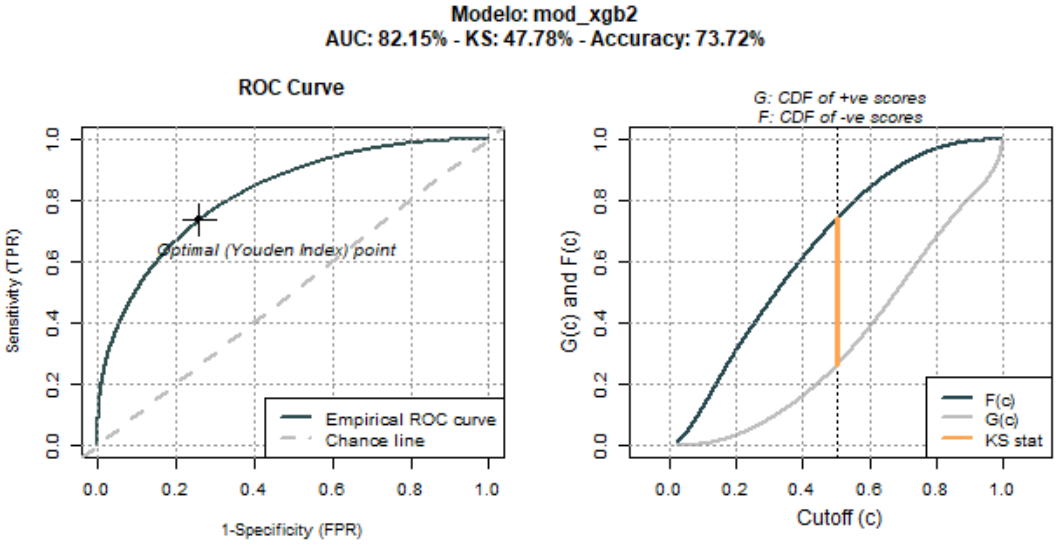


Figura 28: Desempeño Modelo XGBoost Optimizado – Base Test

Random Forest – Versión inicial

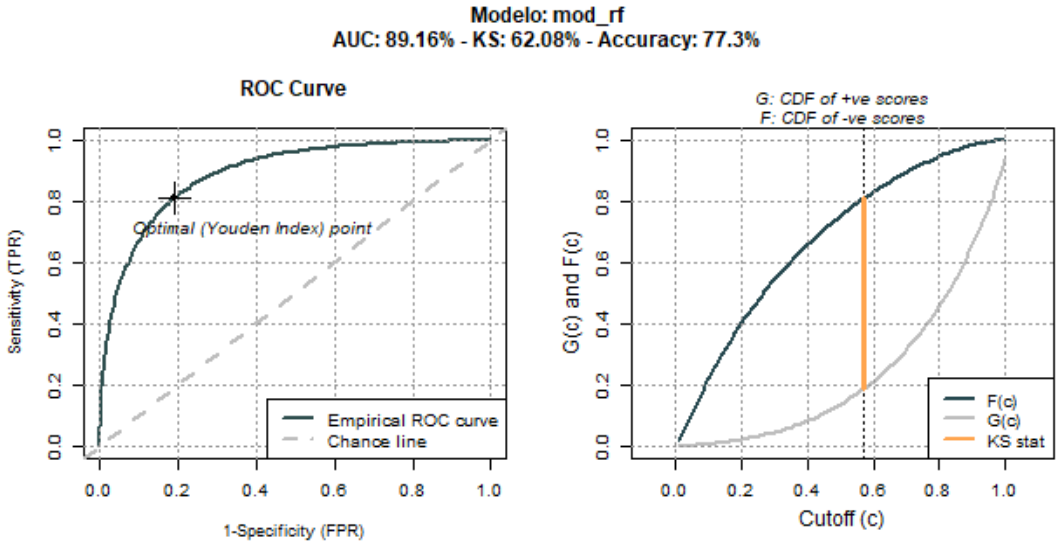


Figura 29: Desempeño Random Forest Inicial – Base Test

Random Forest – Versión optimizada

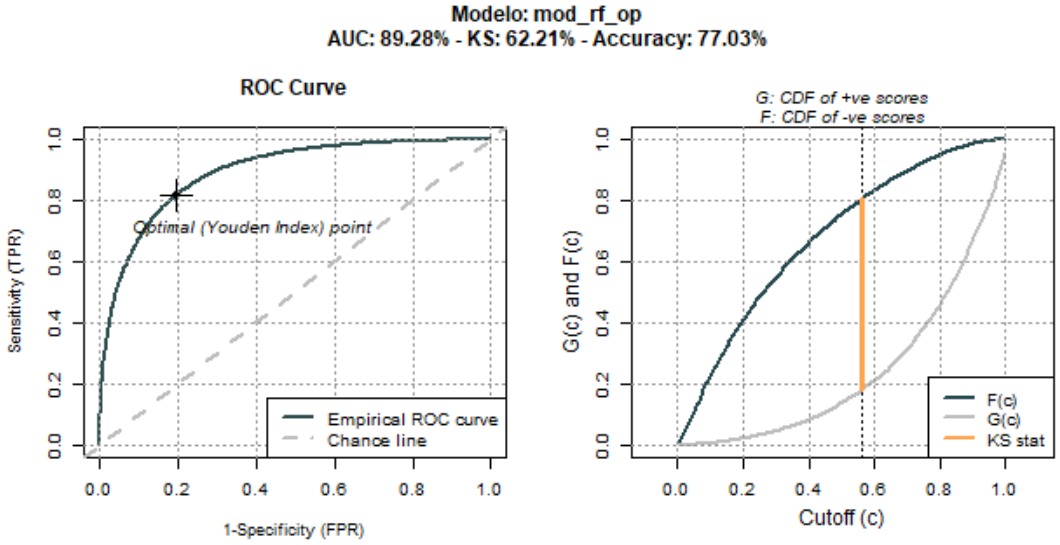


Figura 30: Desempeño Modelo Random Forest Optimizado – Base Test