Desarrollo de un Score de Riesgo Financiero con XGBoost para Evaluación de Crédito Fintech

Juan David Correa Restrepo
C.C. 8357527
jdcorrear@eafit.edu.co

Director
Santiago Ortiz
sortiza2@eafit.edu.co

Maestría en Ciencias de los Datos y Analítica
Escuela de Ciencias Aplicadas e Ingeniería
Universidad EAFIT, Medellín, Colombia

Resumen

El score de crédito es utilizado por las entidades financieras para evaluar el riesgo de sus clientes basado en la probabilidad de caer en mora, informando decisiones de aprobación de crédito. Este proyecto desarrolla un modelo XGBoost para estimar el score de crédito del perfil típico de clientes de una empresa Fintech, utilizando datos sociodemográficos, de empleabilidad y comportamiento financiero. El entrenamiento y validación del modelo involucraron la creación de diversas bases de datos mediante la eliminación de outliers y la reducción de dimensionalidad con PCA Robusta (ROBPCA). Los hiperparámetros se optimizaron utilizando una simulación Montecarlo de 6,000 muestras, evitando el sobreajuste y mejorando el rendimiento. A pesar de desafíos como la dependencia de la calidad de los datos y posibles sesgos, el modelo predice eficazmente el default, especialmente en scores bajos. El proyecto concluye con recomendaciones para la mejora continua, como la integración de nuevas fuentes de datos y el uso de técnicas avanzadas de machine learning, para aumentar la precisión y robustez del modelo. El modelo propuesto puede adaptarse al mercado Fintech y a la población objetivo, facilitando una correcta segmentación de clientes y manteniendo niveles de riesgo aceptables para la organización.

Palabras Clave: Score de crédito, Reducción de dimensionalidad, Eliminación de outliers, Simulación Montecarlo, Estadístico Kolmogorov-Smirnov, Apetito de riesgo..

1. Introducción

El sector financiero y la banca tradicional enfrentan retos significativos debido a la coyuntura económica. Según la Organización Internacional del Trabajo (OIT), el año 2023 experimentó un deterioro económico generalizado, influenciado por factores como las presiones inflacionarias y las tensiones geopolíticas mundiales, que han afectado el crecimiento de muchos países. Mientras tanto, las tasas de desempleo han disminuido, y se observa en países de la Organización para la Cooperación y el Desarrollo Económico (OCDE) una tasa de 4.8 % al cierre del año (Cormann and Scarpetta, ????). Sin embargo, países como Colombia, Chile y Costa Rica continúan siendo un desafío para esta organización, manteniendo los mayores niveles de desempleo, como es el caso de Colombia que cerró con una tasa de 10.2 %, según el Departamento Administrativo Nacional de Estadística (2023) (DANE).

En contraste, la inflación experimentada en los últimos años, posterior a la pandemia, ha resultado en una disminución real de los salarios y del poder adquisitivo de los hogares. Esto ha llevado a muchas familias a buscar opciones para cumplir con sus obligaciones financieras, incrementando la demanda de créditos. En cuanto a la situación del crédito en el país, según Asobancaria en su Reporte de Perspectivas Económicas para 2024, el 2023 dejó un balance de contracción de crédito debido al aumento en las tasas de interés y las estrictas condiciones financieras de hogares y empresas (Malagón González et al., 2024), lo que se tradujo en una caída del 8.7 % real anual en los desembolsos, el resultado más bajo desde mediados de 2021. Además, señala que el sector financiero continúa expuesto a riesgos elevados debido a las proyecciones económicas y la necesidad de impulsar la estabilidad económica del país a través de la profundización financiera.

El reporte del Banco de la República sobre la situación del crédito en Colombia para el cierre de 2023 que, incluye tanto las entidades financieras como aquellas no vigiladas por la Superintendencia Financiera de Colombia, indica una disminución en la demanda de créditos y un aumento en las exigencias para la otorgación de primeros créditos, especialmente en el último trimestre del año (Rodríguez-Novoa et al., 2023). Las entidades de crédito ahora consideran criterios como el flujo de caja proyectado, el historial crediticio, los ingresos recientes del deudor y la relación patrimonio-deuda, directamente asociados con la capacidad de pago y la actividad económica de la persona. Desde la perspectiva de los deudores, la contracción del crédito se debe a la percepción de altas tasas de interés y las dificultades en el acceso al crédito debido a las exigencias de las organizaciones. Esto ha llevado a algunas personas a recurrir a alternativas de crédito no formal, conocidas como “gota a gota”, que aunque accesibles, vienen con altos intereses y prácticas de cobro cuestionables, representando un riesgo significativo tanto para la seguridad personal como para la estabilidad financiera de los individuos.

Toda esta situación ha representado una oportunidad para el surgimiento de opciones de crédito desde el sector Fintech, que busca democratizar el acceso a los servicios financieros mediante tecnologías innovadoras que permiten evaluar de manera más completa y justa el perfil crediticio de individuos tradicionalmente desatendidos. Las Fintech o financieras

tecnológicas corresponden a startups que tienen como estrategia competir con productos o servicios de banca tradicional. Según el Banco Interamericano de Desarrollo (2022), en su informe Fintech en América Latina y el Caribe , el ecosistema Fintech ha crecido considerablemente, especialmente en países como Brasil, México y Colombia. La medición realizada por Finnovista en el primer cuatrimestre de 2023 registró 369 Fintech en Colombia, con 90 nuevas compañías desde 2021 (Finnovista, 2023).

Estas compañías han ganado terreno rápidamente, consolidándose como una alternativa viable frente a los métodos tradicionales de las entidades crediticias, promoviendo la inclusión financiera. Según el Banco Mundial (2022), la inclusión financiera "se refiere al acceso que tienen las personas y las empresas a diversos productos y servicios financieros útiles y asequibles que atienden sus necesidades —transacciones, pagos, ahorro, crédito y seguros — y que se prestan de manera responsable y sostenible". Esto motiva a las Fintech a orientarse a públicos como jóvenes y mujeres. Sin embargo, este público, frecuentemente catalogado como de "alto riesgo", ha resultado en un aumento en el deterioro de las carteras de crédito de estas organizaciones.

Si vemos la evaluación de riesgo desde la metodología tradicional, esta se basa en identificar la pérdida esperada del cliente (Elizondo and Lopez, 1999), tomando como base la aplicación de filtros duros que explican de forma univariada el incumplimiento de las obligaciones. En muchos casos, estos filtros duros poseen criterios subjetivos de evaluación y asignación. Según Sepúlveda et al. (2012), el análisis de riesgo presenta gran volatilidad a la hora afrontan las decisiones de crédito de la organización por lo que es vital conocer, identificar y medir la evolución del mismo y generar planes y medidas anticipadas para prevenir el deterioro de la cartera

La industria Fintech, motivada por la innovación, ha desarrollado nuevas metodologías de evaluación del riesgo de crédito, como el uso de big data con datos no tradicionales y la aplicación de metodologías de machine learning. Estas técnicas han demostrado ser efectivas al discriminar poblaciones perfiladas como de alto riesgo, permitiendo el desarrollo de nuevas metodologías de evaluación. Según Crock et al. (2007), se pueden categorizar tres tipos de modelos para la evaluación de riesgo de score: los que estiman la probabilidad de default (PD), aquellos que miden la exposición al default (EAD) y los que calculan la pérdida esperada (PE). En nuestro caso, utilizamos la probabilidad de default para entrenar nuestro modelo de aprendizaje supervisado.

Este trabajo tiene como objetivo diseñar un modelo predictivo para estimar el score crediticio de los usuarios que solicitan créditos en una empresa Fintech. Este modelo se desarrollará para adaptarse a las características únicas y condiciones específicas de este segmento de la población, utilizando un enfoque metodológico riguroso y técnicas avanzadas de ciencia de datos, proporcionando una herramienta valiosa para la toma de decisiones crediticias basadas en información más completa, promoviendo la inclusión financiera y fortaleciendo la estabilidad del sector financiero en su conjunto. El modelo predictivo se basará en el análisis de diversas fuentes de información obtenidas del ecosistema digital, utilizan-

do herramientas tecnológicas para el análisis del sistema crediticio, incluyendo variables demográficas, socioeconómicas, de empleo e historial crediticio. Estas fuentes de datos digitales son un activo valioso para comprender mejor a los clientes y ofrecer alternativas para hacer visibles a las personas, considerándolas en el ámbito del crédito (Costa et al., 2015).

A día de hoy existen innumerables metodologías de estimación del score de riesgo de crédito basado en metodologías de machine learning, entre estos los mas utilizados son las redes neuronales (West, 2000), los arboles de decisión (Lee and Scolari, 2010) y las máquinas de soporte vectorial (Min and Lee, 2005; Huang et al., 2007). A pesar que se han popularizado el uso de estas técnicas, muchas empresas de la industria persisten en la utilización de metodos estadísticos tradicionales debido a su facilidad de implementación y su precisión (Lessmann et al., 2015).

Para la selección del modelo, se optó por XGBOOST como herramienta principal, dado que esta metodología demuestra una robustez y eficiencia significativas en el manejo de grandes volúmenes de información, características cruciales para los datos procesados en la industria Fintech. Adicionalmente, su capacidad para optimizar tanto la velocidad de computación como la precisión del modelo permite una mayor flexibilidad en la configuración de hiperparámetros, facilitando un ajuste más eficaz y rápido. La implementación de este modelo en un entorno productivo resulta notablemente simplificada gracias a su amplia documentación, y el seguimiento y la calibración se desarrollan y actualizan con mayor facilidad

El documento está estructurado de la siguiente manera: En la Sección 2, se presenta un resumen detallado de la metodología utilizada para entrenar el modelo, incluyendo una descripción de los datos utilizados, las técnicas de preprocesamiento aplicadas, la selección y justificación de los algoritmos de aprendizaje automático elegidos, así como los criterios de evaluación y validación del modelo. En la Sección 3, se resumen los resultados obtenidos del entrenamiento del modelo, discutiendo las métricas de rendimiento como el estimador KS y el F1-Score. Además, se presentarán gráficos y tablas que ilustran el comportamiento del modelo durante el proceso de entrenamiento y validación. Finalmente, en la Sección 4, se exponen las conclusiones más relevantes derivadas del análisis de los resultados, incluyendo comentarios finales sobre las limitaciones del trabajo, posibles mejoras futuras y la aplicabilidad práctica de los resultados obtenidos en contextos reales.

2. Metodología

Para el diseño del modelo predictivo de score de crédito de clientes del sector fintech, se tuvieron en cuenta un conjunto de materiales y métodos que se resumen a continuación y que fueron explorados para el diseño y entrenamiento del modelo, como se muestra en la Figura 1. El tratamiento de estos datos crudos, a través de procesos no paramétricos robustos, busca identificar cuál de estos métodos presenta los mejores resultados. El objetivo es lograr una solución más eficiente al problema a resolver, que finalmente conduzca

a un modelo adaptable a la industria, capaz de predecir el comportamiento de pago de los clientes a corto, mediano y largo plazo.

Figura 1: Análisis detallado de los datos - Fuente: Construcción propia

Al desarrollar cada uno de los métodos, se utilizará validación cruzada para garantizar la robustez del modelo mediante la validación de métricas de rendimiento como la Matriz de Confusión y el Estadístico Kolmogorov-Smirnov (KS). El objetivo es asegurar que el modelo sea teóricamente sólido y también aplicable en ambientes productivos, alineado con las necesidades específicas del sector fintech. Para esto, se considera una muestra poblacional $X_{n,p}$ que proviene de la población con crédito originado con muestras totalmente independientes, además sea $Y \in \mathbb{R}$ con $Y_{n,1} = (y_1)$ es la variable dependiente, donde p es la dimensión y n el tamaño de la muestra para $i = 1, \ldots, n$.

2.1. Transformación y exploración de los datos

La exploración de los datos habilita la comprensión inicial de su estructura y calidad, sirviendo como punto de partida para la selección y aplicación de técnicas de modelado dentro del estudio, y permite la identificación de posibles anomalías o valores atípicos. Para la exploración de los datos en este estudio, se emplea un enfoque de estadística descriptiva (Field and Field, 2012), lo que facilita una comprensión detallada de la naturaleza y calidad de los datos disponibles. Este enfoque permite una visión univariada de las variables, examinando su distribución, tendencias centrales y dispersión.

Los resultados de las estadísticas descriptivas permitieron obtener muestras de entrenamiento balanceadas de las variables dependientes en relación con la variable de default (Mora90@9). Se reagruparon las categorías basándose en una metodología de frecuencias para las variables categóricas y en rangos según distribución para las variables continuas.

Además, se realizaron análisis de frecuencia para comprender la ocurrencia de diferentes categorías en las variables categóricas, y se calcularon medidas de tendencia central y dispersión para las variables numéricas, ayudando a identificar posibles sesgos o patrones en los datos. En esta etapa, se realizaron procesos de estandarización y normalización de los datos (James et al., 2017), llevándolos a la misma escala para facilitar el entrenamiento del modelo y evitar sesgos debidos a la escala. La estandarización garantiza que todas las variables contribuyan de manera equitativa al modelo, mejorando así su desempeño e interpretabilidad (Becker et al., 1988). Cabe anotar que las bases estandarizadas y normalizadas no serán utilizadas para entrenar el modelo de árbol de decisión, sino los modelos con los cuales se comparará su rendimiento, como la regresión logística (RegLog).

Estandarización y Normalización de Datos

La estandarización de los datos es un procedimiento comúnmente utilizado escalar los datos de tal forma que estos cumplan con las caracteristicas de $\bar{X} = 0$ y $\sigma = 1$. representado con la siguiente formula.

$$x^i_{std} = \frac{x_i - \mu_x}{\sigma_x}$$

La normalización de los datos es un procedimiento comúnmente utilizado escalar los datos de tal forma que estos se escalen en valores entre 0 y 1, donde el valor mínimo será cero y el valor máximo será 1. Se calcula con la siguiente formula.

$$x^i_{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

donde x_{min} y x_{max} representan los valores mínimos y máximos de la variable a normalizar.

2.2. Detección de Observaciones Atípicas Multivariantes

La identificación de muestras que no cumplen con los parámetros propios de la población en estudio, también conocidas como outliers, tiene como fin minimizar los datos que podrían originar ruido en el modelo y hacerlo ineficiente, lo cual podría resultar en métricas de desempeño pobres. Como se señala, "la mayoría de los conjuntos de datos pueden contener una serie de observaciones o muestras atípicas que no parecen pertenecer al patrón de variabilidad producido por las demás observaciones" (Johnson and Wichern, 2002).

Para la detección de datos atípicos, se plantea el uso de la metodología Local Outlier Factor (LOF) con el objetivo de mejorar la predicción del modelo al excluir observaciones que puedan distorsionar la relación entre las variables y el comportamiento de pago de los clientes. Este enfoque aumenta la robustez del modelo y lo hace más generalizable y eficiente al implementarse en entornos de producción en el sector de las Fintech. El uso de LOF permite contrastar patrones representativos dentro de la población objetivo, optimizando así la confiabilidad de las predicciones del modelo.o.

Local Outlier Factor (LOF)

El *Local Outlier Factor*, con sus siglas LOF es una metodología de identificación de puntos atípicos en un set de datos, el cual identifica y ajusta las variaciones de las densidades locales. Para cada muestra multivariada, se representa un punto de datos X_i, se tiene un $D^k(X_i)$ y su distancia al k-vecino más cercano de X, y $L_k(X_i)$, que es el conjunto de puntos situados a una distancia k-vecino más cercanos de X. Observe que $L_k(X_i)$ suele contener una cantidad k de puntos, pero a veces contiene más de k puntos debido a las muestras en los k-vecino más próximos. Entonces, la distancia de alcanzabilidad $R_k(\overline{X}, \overline{Y})$ del objeto X con respecto a Y se define como el máximo de $dist(X, Y)$ y los k-vecinos más cercanos de Y:

$$R_k(X, Y) = \text{máx}\{\text{dist}(X, Y), D^k(Y)\}$$

La distancia de alcanzabilidad no es simétrica entre X y Y. Intuitivamente, cuando Y se encuentra en una región densa y la distancia entre X y Y es grande, la distancia de alcanzabilidad de X con respecto a ella es igual a la distancia verdadera $dist(X, Y)$. Por otro lado, cuando las distancias entre X y Y son pequeñas, la distancia de alcanzabilidad se suaviza mediante los k-vecinos más cercanos de Y. Cuanto mayor sea el valor de k, mayor será el suavizado. En consecuencia, las distancias de accesibilidad con respecto a diferentes puntos también serán más similares.

Por tanto, la distancia de accesibilidad media $AR_k(X)$ del punto de datos X se define como la media de sus distancias de accesibilidad a todos los objetos de su vecindario $L_k(X)$. La función *MEAN* representa simplemente la media de un conjunto de valores. Los valores LOF se pueden expresar de forma más sencilla e intuitiva en términos de la distancia de alcanzabilidad media $AR_k(\mathbf{X})$. El factor atípico local es entonces simplemente igual a la relación media de $AR_k(\mathbf{X})$ con los valores correspondientes de todos los puntos de los k-vecinos más cercanos de $\mathbf{X}$:

$$LOF_k(\mathbf{X}) = MEAN_{\mathbf{Y} \in L_k(\mathbf{X})} \left(\frac{AR_k(\mathbf{X})}{AR_k(\mathbf{Y})} \right)$$

$$LOF_k(\mathbf{X}) = \frac{AR_k(\mathbf{X})}{MEAN_{\mathbf{Y} \in L_k(\mathbf{X})} \left(AR_k(\mathbf{Y}) \right)}$$

$$LOF_k(\mathbf{X}) = \frac{AR_k(\mathbf{X})}{\frac{1}{|L_k(\mathbf{X})|} \sum_{\mathbf{Y} \in L_k(\mathbf{X})} AR_k(\mathbf{Y})}$$

El uso de ratios de distancia en la definición garantiza que el comportamiento de la distancia local se tenga bien en cuenta. Como resultado, los valores de LOF para los objetos de un clúster son a menudo cercanos a 1 cuando los puntos de datos del clúster están homogéneamente distribuidos. Por ejemplo, en la parte izquierda de la Figura 2, se muestra un conjunto de datos bidimensional que contiene un clúster gaussiano de baja densidad con 200 muestras y tres grandes clústeres con 500 muestras cada uno: uno gaussiano denso y los otros dos clústeres uniformes de diferentes densidades. Además, contiene un par de valores atípicos.

En la parte derecha de la Figura 2 se representa el LOF (Breunig et al., 2000). Los valores de LOF de los puntos de datos de ambos clústeres serán próximos a 1, aunque las densidades de los dos clústeres sean diferentes. Por otro lado, los valores de LOF de los dos puntos periféricos serán mucho más altos, ya que se calcularán en términos de las relaciones con las distancias medias de alcanzabilidad de los vecinos. Las puntuaciones LOF también pueden verse como la distancia de alcanzabilidad normalizada de un punto, donde el factor de normalización es la media armónica de las distancias de alcanzabilidad en su localidad. Por ejemplo, la ecuación siguiente puede reescribirse como sigue:

$$LOF_k(X) = \frac{AR_k(X)}{HMEAN_{Y \in L_k(X)} AR_k(Y)}$$

Aquí, HMEAN denota la media armónica de las distancias de alcanzabilidad de todos los puntos de su localidad. En principio, podríamos utilizar cualquier tipo de media en el denominador (Aggarwal, 2017).

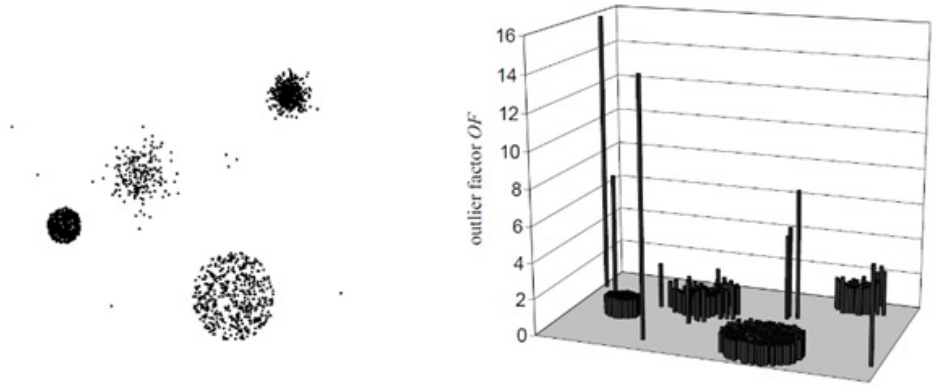

Figura 2: Outlier - Fuente: Charu C. Aggarwal. Outlier Analysis (Aggarwal, 2017)

2.3. Reducción de Dimensionalidad

Uno de los aspectos más importantes cuando se tiene información es la reducción de dimensionalidad, ya que permite identificar variables representativas o construir nuevas variables como resultado de la operatividad matemática de otras variables.

Robust PCA (ROBPCA)

ROBPCA se presenta como un método robusto de análisis de componentes principales (PCA), resolviendo un problema fundamental del PCA clásico, que se basa en la matriz de covarianza empírica de los datos y es muy sensible a las observaciones atípicas. Hasta la fecha, se han desarrollado dos enfoques robustos: el primero se basa en los vectores propios de una matriz de dispersión robusta, como el determinante de covarianza mínimo o un S-estimador, y se limita a datos de dimensiones relativamente bajas. El segundo enfoque se basa en la búsqueda de proyección y puede manejar datos de alta dimensión. Desde un enfoque práctico, se utiliza una metodología que combina las ideas de búsqueda de proyección con una estimación robusta de la matriz de dispersión. ROBPCA produce estimaciones más precisas en conjuntos de datos no contaminados y estimaciones más robustas en datos contaminados. Además, el ROBPCA puede calcularse rápidamente y es capaz de detectar situaciones de ajuste exacto.

Suponemos que los datos originales se almacenan en una matriz de datos np donde $X = X_{(n,p)}$, donde n indica el número de objetos y p el número original de variables. El método ROBPCA consta de tres pasos principales. En primer lugar, los datos se preprocesan de forma que los datos transformados se encuentren en un subespacio cuya dimensión

sea como máximo $n-1$. A continuación, se construye una matriz de covarianza preliminar S_0 y se utiliza para seleccionar el número de componentes k que se retendrán en la secuela, obteniendo un subespacio k-dimensional que se ajusta bien a los datos. Entonces, los puntos de datos se proyectan en este subespacio, donde se estiman de forma robusta su ubicación y la matriz de dispersión, a partir de la cual se calculan sus k valores propios no nulos $l_1, \ldots, l_k$. Los vectores propios correspondientes son los k componentes principales robustos.

En el espacio original de dimensión p , estos k componentes abarcan un subespacio k-dimensional. Formalmente, si se escriben los vectores propios (columna) uno junto a otro, se obtiene la matriz pk donde $P_{(p,k)}$, con columnas ortogonales. La estimación de la ubicación se denota mediante el vector columna u de p-variables y se denomina centro robusto. Las puntuaciones son las entradas de la matriz nk

$$T_{n,k} = (X_{n,p} - 1_n\hat{\mu}')P_{p,k},$$

donde 1_n es el vector columna con todos los n componentes iguales a 1. Además, los k componentes principales robustos generan una matriz S de dispersión robusta pp de rango k dada por:

$$S = P_{p,k}L_{k,k}p'_{p,k};$$

siendo $L_{k,k}$ la matriz diagonal con los valores propios $l_1, \ldots, l_k$.

Al igual que el PCA clásico, el método ROBPCA es equivariante de localización y ortogonal. Es decir, cuando se aplica un desplazamiento y/o una transformación ortogonal (por ejemplo, una rotación o una reflexión) a los datos, el centro robusto también se desplaza, y las cargas se rotan en consecuencia. Por lo tanto, las puntuaciones no cambian bajo este tipo de transformación. Dejemos que $A_{p,p}$ defina una transformación ortogonal; así A es de rango completo y $A' = A^{-1}$, y $\hat{\mu}_x$ y $P_{p,k}$ son el centro ROBPCA y la matriz de carga para el $X_{n,p}$ original. Entonces el centro ROBPCA y las cargas para los datos transformados $XA' + 1_n v'$ son iguales a $X\hat{\mu}_x + v$. En consecuencia, las puntuaciones siguen siendo las mismas bajo estas transformaciones.

$$T(XA' + 1_n v') = (XA' + 1_n v' - 1_n(A\hat{\mu}_x + v)')AP$$

$$T(XA' + 1_n v') = (X - 1_n\hat{\mu}'_x)PP$$

$$T(XA' + 1_n v') = T(X)$$

Aunque estas propiedades parecen muy naturales para un método de PCA, no las comparten algunos otros estimadores robustos de PCA, como el remuestreo por medias y el menor medio volumen de Egan y Morgan (1998) (Hubert and J.Rousseeuw, 2005). ROBPCA es especialmente valioso en contextos donde tanto la presencia de datos atípicos como la alta dimensionalidad son preocupaciones, como en genómica y finanzas, permitiendo descubrimientos más claros y relevantes.

2.4. Aprendizaje por Computadora

El aprendizaje por computador, o machine learning, ha cobrado un papel fundamental en la industria financiera. Se ha convertido en una de las principales herramientas para predecir el comportamiento de pago de las obligaciones crediticias de los clientes, permitiendo tomar decisiones basadas en las variables que estas entidades tienen sobre sus clientes, como la aprobación o el rechazo de una solicitud de crédito. Esta rama de la inteligencia artificial se adapta a la naturaleza de los datos y al tipo de comportamiento que se desea predecir, permitiendo, de acuerdo con el modelo que mejor se ajuste, obtener la mayor precisión posible en las predicciones. Cada uno de estos modelos tiene características particulares que pueden adaptarse a diferentes aspectos del análisis de riesgo crediticio, dependiendo de la naturaleza de los datos y los objetivos específicos del análisis.

Extreme Gradient Boosting (XGBoost)

Este modelo es ampliamente utilizado en el sector financiero para modelar el score de crédito debido a su capacidad para manejar grandes conjuntos de datos con alta dimensionalidad y proporcionar resultados predictivos robustos y precisos. La capacidad de XGBoost para trabajar con variables categóricas y continuas lo hace ideal para conjuntos de datos financieros que combinan diferentes tipos de datos (por ejemplo, edad, ingresos, historial de crédito). Este algoritmo tiene en cuenta la dispersión de los datos y un esbozo cuantílico ponderado para el aprendizaje aproximado de árboles (Chen and Guestrin, 2016).

Mejora el Gradient Boosting Machine tradicional mediante el uso de regularización (ℓ_1 y ℓ_2) en la función de costo para controlar el sobreajuste, una técnica de optimización basada en gradientes estocásticos y un algoritmo de particionamiento de árboles basado en histogramas para manejar grandes volúmenes de datos de manera eficiente. El modelo XGBoost puede expresarse como:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F},$$

donde y_i es la predicción para la i-ésima instancia, $\varphi(x_i)$ es la suma de las predicciones de k arboles de decisión, y $\mathcal{F}$ es el espacio de todos los árboles posibles. El aprendizaje se realiza minimizando la siguiente función de pérdida regularizada:

$$\mathcal{L}(\phi) = \sum_i L(y_i, \hat{y}_i) + \sum_k \Omega(f_k),$$

donde l es una función de pérdida diferenciable convexa y ω es un término de regulación que penaliza la complejidad del modelo.

2.5. Evaluación

La evaluación del modelo nos permitirá validar la calidad del procedimiento y determinar si puede hacer predicciones confiables basadas en el aprendizaje de los datos de entrada

o si, por el contrario, se está presentando sobreajuste. Por tal motivo, probar el modelo con nuevos datos es importante, ya que esto llevará al modelo a predecir comportamientos en escenarios totalmente desconocidos. En el escenario de evaluación, podremos llevar a los modelos entrenados a competir entre ellos con el fin de identificar el mejor desempeño mediante un *Champion Challenger*, pudiendo ajustar los parámetros y tomar decisiones sobre nuevas características a incluir para hacerlos más eficientes.

Dependiendo del sector productivo del modelo y el apetito de riesgo, con la evaluación del modelo se pueden aceptar máximos y mínimos de desempeño. En el caso del score de crédito, esto implica identificar los puntos de corte partiendo de un default esperado. Cada una de las métricas calculadas para el desempeño ofrece una perspectiva única sobre el modelo y, en conjunto, pueden proporcionar una evaluación comprensiva que ayudará a optimizar y mejorarlo.

Matriz de Confusión

Una de las principales herramientas utilizadas para evaluar el rendimiento de los modelos de clasificación es la matriz de confusión, ya que muestra de manera clara y sencilla la precisión del modelo implementado. Estos datos se presentan de forma gráfica mediante una tabla que segmenta las predicciones en cuatro categorías: verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN), como se muestra en el Cuadro 1.

- **Verdaderos Positivos (VP):** Escenario positivo con predicción acertada.
- **Verdaderos Negativos (VN):** Escenario negativo con predicción acertada.
- **Falsos Positivos (FP):** Escenario negativo con predicción positiva - Error tipo 1.
- **Falsos Negativos (FN):** Escenario positivo con predicción negativa - Error tipo 2.

	Clase Predicha	
Clase Real	Verdadero Positivo (VP)	Falso Positivo (FP)
	Falso Negativo (FN)	Verdadero Negativo (VN)

Cuadro 1: Matriz de Confusión

Permiten calcular varias métricas de evaluación, como precisión, recall, especificidad y la puntuación F1, entre otros, ofreciendo una visión comprensiva del rendimiento del modelo en términos de su capacidad para clasificar correctamente las clases en cuestión.

Precisión (Pre): La proporción de predicciones correctas (tanto positivas como negativas) entre el total de casos examinados.

$$\text{Pre} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

Recall (Rec): La proporción de casos positivos que fueron correctamente identificados por el modelo.

$$\text{Rec} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

Accuracy (Acc): La proporción de casos de acierto (positivos y negativos) respecto a la cantidad real.

$$\text{Acc} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

F1-Score (F1): El promedio armónico de la precisión y la sensibilidad, que proporciona una medida única de la calidad del modelo.

$$\text{F1} = 2\frac{\text{Pre} \times \text{Rec}}{\text{Pre} + \text{Rec}}$$

Estadístico Kolmogorov-Smirnov (KS)

El estadístico KS es una medida no paramétrica utilizada para determinar si dos muestras de datos provienen de la misma distribución. En la industria financiera y fintech, el valor de KS se ha convertido en una herramienta estándar para evaluar la confiabilidad de los modelos de score de crédito(Thomas et al., 2002).

El cálculo de KS se obtiene de la máxima distancia vertical entre las funciones de distribución acumulada de dos grupos independientes y se expresa mediante la siguiente fórmula:

$$D_{n,m} = \sup_x \left| \hat{F}_{1,n}(x) - \hat{F}_{2,m}(x) \right|,$$

donde $\hat{F}_{1,n}$ y $\hat{F}_{2,n}$ son las funciones de distribución acumulada empíricas de las dos muestras independientes y el $\sup_x$ indica el supremo para todos los puntos x

Esta métrica es comúnmente utilizada en la industria del crédito para evaluar la capacidad de los modelos de score de crédito para distinguir entre clientes que entrarán en default y los que no. Un valor alto de KS indica una buena separación de las distribuciones de los grupos, lo que implica un modelo más efectivo (Rodríguez Benavides and Perrotini Hernández, 2019).

3. Resultados y Discusión

Durante esta sección, abarcaremos en detalle cada una de las etapas de desarrollo del modelo, mostrando los resultados, analizándolos y concluyendo respecto a los mismos. Como punto de partida, utilizaremos el conjunto de datos con las 11 variables seleccionadas y las 30,521 muestras. Asimismo, durante el proceso probaremos otras métricas no especificadas en el diseño original con el objetivo de identificar y realizar un benchmark entre las propuestas y otras comúnmente utilizadas. Finalmente, buscaremos los parámetros de

entrenamiento óptimos para el modelo diseñado.

Los datos que se utilizarán como punto de partida para desarrollar todo el proceso de transformación (atípicos y reducción de dimensionalidad) incluyen dos variables numéricas: edad y salario, cuya función de distribución se puede ver en la Figura 3, con una correlación es 15,84 % Además, se incluyen 9 variables categóricas detalladas en su tabla de frecuencia en la Figura 5. Estos datos serán utilizados como la base de referencia para la construcción de las bases sin outliers y componentes principales en pro de la optimización del modelo.

3.1. Caso de Estudio

Los datos utilizados provienen de las bases internas de los clientes de una fintech originadora de crédito. Durante el proceso de colocación del crédito, se consultan y almacenan datos provenientes de diferentes fuentes, entre las que se incluyen datos del dispositivo utilizado para solicitar el crédito (fingerprint), datos sociodemográficos del solicitante, datos del empleo, historial de crédito y comportamiento de pago, llegando a un total de más de 2000 datos únicos por cliente y más de 850,000 registros.

Mediante procesamiento y análisis estadístico para garantizar la calidad y utilidad de los datos, se seleccionaron 31,880 muestras que cumplen con las siguientes condiciones:

- Solo clientes desembolsados.
- Clientes desembolsados entre el 2020-01-01 y el 2023-12-31.
- Solo registros completos con las variables seleccionadas para el modelo.

Algunos de los criterios utilizados para descartar variables del conjunto de datos original fueron: incompletitud de los datos, baja varianza, alta correlación con otras variables del conjunto de datos y baja asociación con la variable respuesta, entre otros. El conjunto de datos resultante utilizado para el entrenamiento del modelo se centró en variables sociodemográficas, comportamiento de pago y de empleo (Cuadro 2), las cuales posteriormente fueron procesadas para obtener nuevas variables que pudieran ser representativas con el objetivo de optimizar el modelo.

El tratamiento de las variables categóricas seleccionadas se centró en convertir las variables dicotómicas en variables indicadoras $I_n = 0, 1$. Por ejemplo, sexo (masculino, femenino) y tipo de ocupación (empleado, independiente). Además, se asignaron valores cuantitativos a variables categóricas que pueden tratarse como ordinales, como estrato y nivel educativo. Para el tratamiento de las variables categóricas no ordenables, se definió utilizar la metodología one-hot encoding para convertirlas en varias variables indicadoras con valores 1 y 0; por ejemplo, ciudad de residencia, banco de cuenta de ahorro, uso del crédito, entre otras.

La variable explicativa del modelo para cada muestra es una variable dicotómica que especifica si el cliente cayó en default (1) o no cayó en default (0). Para nuestro caso, utilizaremos varias métricas de default, entre las cuales tenemos `M1@1(EED), M30@3(ED),`

Variable	Descripción	Tipo
genero	Genero del cliente	Categórica
edad	Edad	Numérica
tipoEmpleo	Tipo de contrato laborar	Categórica
periodicidadPago	Frecuencia en la cual recibe salario	Categórica
personasCargo	Número de personas que dependen económicamente	Categórica
nivelEstudios	Nivel educativo máximo adquirido	Categórica
tipoVivienda	Tipo de vivienda donde reside	Categórica
permanenciaVivienda	Tiempo de antigüedad en el actual domicilio	Categórica
estadoCivil	Estado civil	Categórica
salario	Salario devengado por actividad laboral	Numérica
metodoPago	Forma en la cual recibe el salario	Categórica
EED	Mora 1 en el primer mes del crédito	dicotómica
ED	Mora 30 en los primeros 3 meses del crédito	Dicotómica
D	Mora 60 en los primeros 6 meses del crédito	Dicotómica
LD	Mora 60 en los primeros 9 meses del crédito	Dicotómica
ELD	Mora 90 en los primeros 9 mese del crédito	Dicotómica

Cuadro 2: Variables seleccionadas para el entrenamiento

`M606(D), M60@9(LD) y M90@9(ELD)`. Estas métricas indican que el cliente tuvo `M##` días de mora en el último día del mes en el que debía cancelar la cuota durante los primeros `@#` meses de vida del crédito.

Dentro de los procesos de transformación y exploración de los datos, se realizó la normalización y estandarización de los datos, obteniendo dos bases de datos: una con los datos normalizados y otra con los datos estandarizados. Ambas serán comparadas para identificar cuál de ellas proporciona mejores resultados en el entrenamiento del modelo. Con estos conjuntos de datos, se llevaron a cabo procesos de detección de observaciones atípicas multivariantes mediante la metodología LOF y la reducción de dimensionalidad mediante la metodología ROBPCA, obteniendo así un conjunto de datos limpios.

Con este conjunto de datos, se procedió a la fase de modelado, donde se utilizaron dos metodologías altamente conocidas y empleadas en el ámbito financiero para predecir comportamientos de pagos de los clientes: Regresión Logística Elastic-Net y XGBoost. Para comparar los resultados de los dos modelos, se realizó una validación con la matriz de confusión y su métrica F1, así como con el estadístico KS, con el fin de identificar y validar la capacidad predictiva y la robustez de ambos ante cambios de muestras. Se seleccionará como mejor modelo aquel que obtenga mejores métricas de predicción y estabilidad frente a la variación de los datos.

Una vez definida la base a trabajar, se crearon otras dos bases de datos: una con los datos normalizados y otra con los datos estandarizados. Se compararon los resultados de un primer entrenamiento para determinar si existe algún tipo de optimización, tanto en la

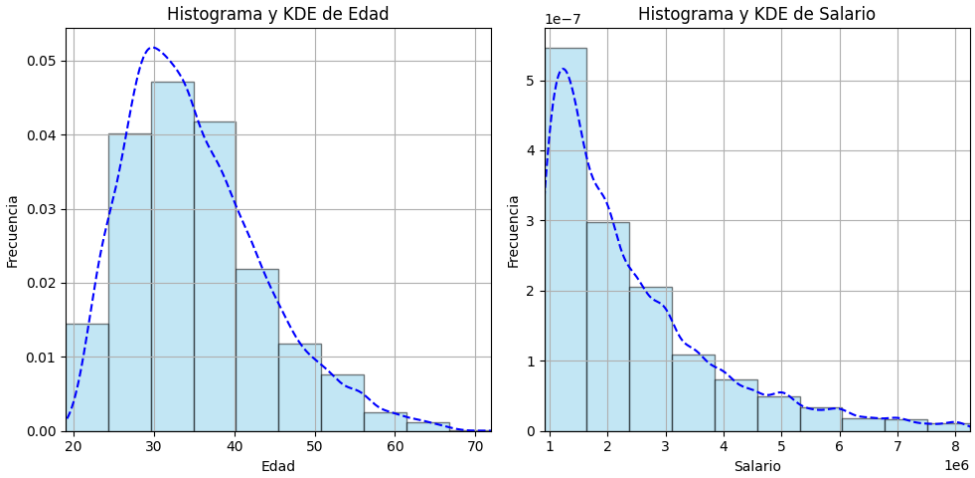

Figura 3: Distribución de las variables numéricas

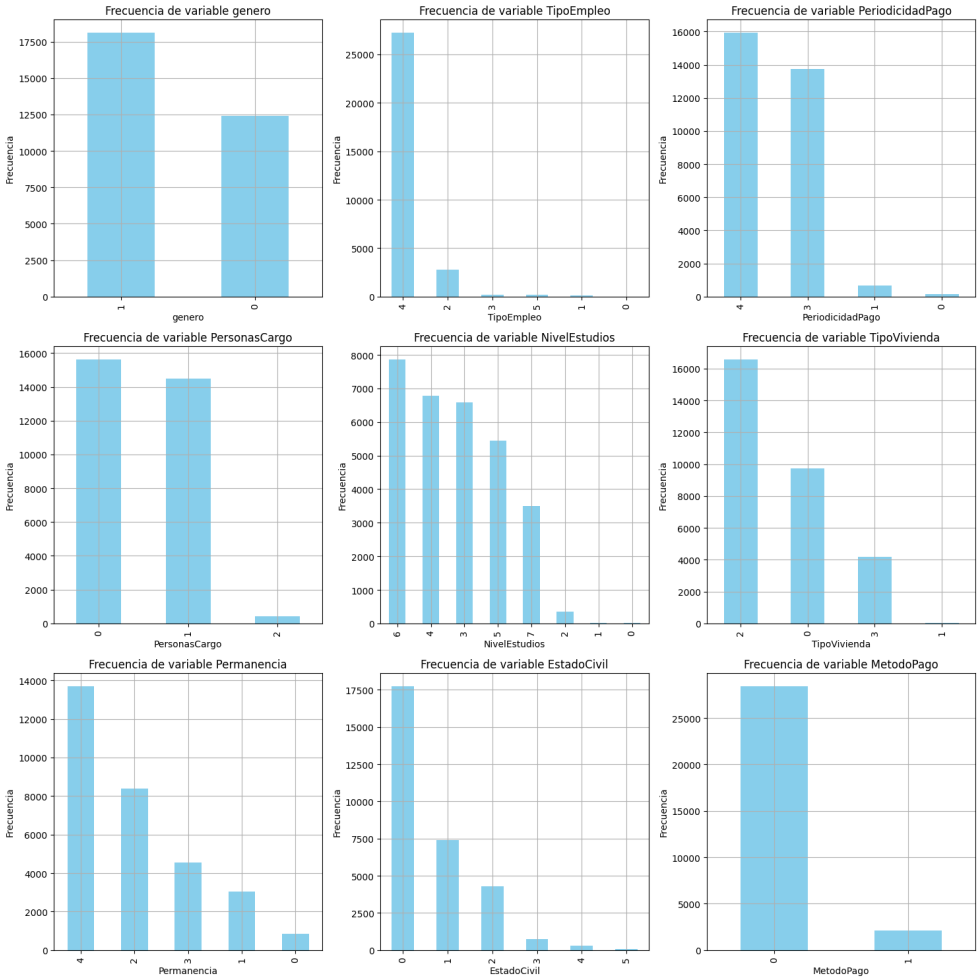

Figura 4: Distribución de las variables categóricas

calidad del modelo como en el tiempo de entrenamiento, al transformar los datos respecto al entrenamiento del modelo de árbol de decisión XGBoost. Las bases resultantes son: dfModelo para la base original, dfModeloStd para los datos estandarizados y dfModeloNorm para los datos normalizados. Como resultado, y validando la teoría de los árboles de decisión que expresa que estos no son afectados por la escala, podemos ver el resultado en la Figura 5

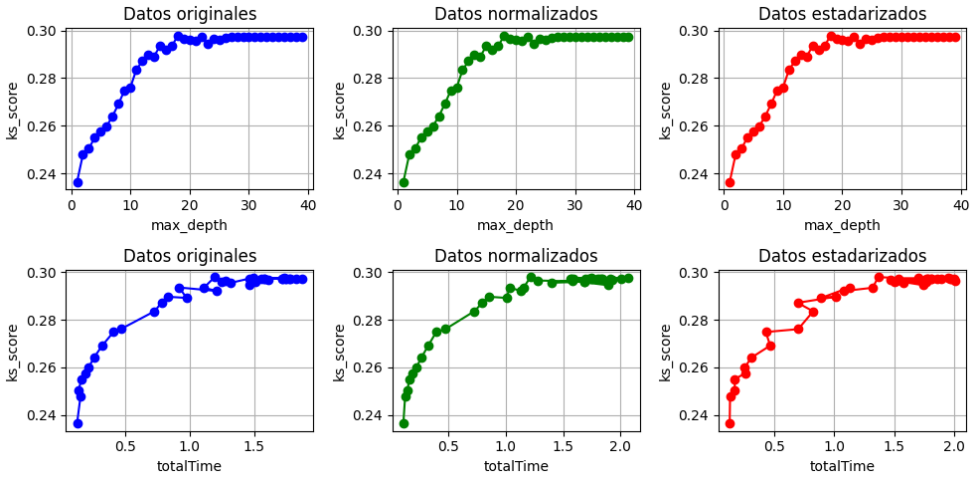

Figura 5: Resultado de entrenamiento del modelo con dfModelo, dfModeloNorm y dfModeloStd

Como resultado se puede observar que no hay diferencias entre el valor KS ni en tiempo al entrenar el modelo con la base de datos original, la base normalizada y la base estandarizada. Por tanto para desarrollo de las etapas posteriores solo utilizaremos las base sin escalar denominada dfModelo. Una vez definida la base a entrenar se definió la variable target para el entrenamiento, ya que contamos con 5 variables objetivos denominadas EED, ED, D, LD y ELD, descritas anteriormente. Desarrollamos el mismo procedimiento anteriormente cambiando solo el parámetro de entrenamiento Max_depth y comparando los valores KS del entrenamiento para encontrar cual métrica se adapta mejor a las variables de entrenamiento. Obteniendo el resultado de la Figura 6. Por tanto se define la variable EED como variable objetivo de nuestro entrenamiento, la cual representa un default de Mora 90 en los primeros 9 meses de vida del crédito.

Los resultados de aplicar la metodología de identificación y eliminación de outliers a la base de datos y la variable objetivo definida para el entrenamiento bajo la técnica de LOF con una selección de atípicos del 5 % identificaron 1,526 muestras, como era de esperarse. Asimismo, se hizo un proceso de comparación con otras metodologías de identificación de outliers como Isolation Forest (IF), One-Class SVM (SVM) y Elliptic Envelope (EE). El resultado de la comparación de cada método con las variables encontradas en común se puede ver en el Cuadro 3, obteniendo al final cinco nuevas bases de datos llamadas dfModeloLOF, dfModeloIF, dfModeloSVM y dfModeloEE. Además, se crearon dos bases adicionales, dfModeloALL y dfModeloANY, con las siguientes características:

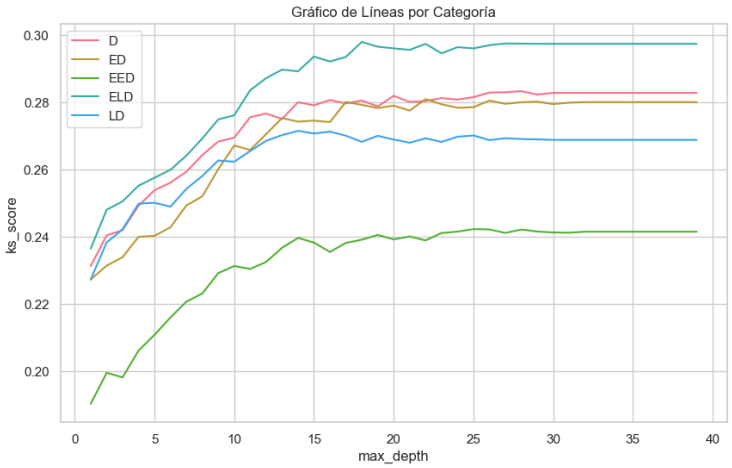

Figura 6: Resultado de entrenamiento del modelo con diferentes variables respuesta

- dfModeloALL: Base sin las muestras clasificadas como outliers por todos los métodos.
- dfModeloANY: Base sin las muestras clasificadas como outliers por al menos un método.

Método	LOF	IF	SVM	EE
LOF	1526	64	655	68
IF	64	1526	549	879
SVM	655	549	9717	569
EE	68	879	569	1527

Cuadro 3: Cantidad de variables Outliers compartidas entre cada método

De la misma forma que se entrenó preliminarmente un modelo para identificar si existe alguna diferencia significativa en el valor del KS con diferentes variables respuesta, en esta ocasión entrenaremos el modelo con las bases resultantes de aplicar los diferentes métodos de eliminación de outliers respecto a la base original, utilizando las variables respuesta EED y ELD, y variando el parámetro max_depth. El resultado se observa en la Figura 7, y se concluye que el modelo que mejor comportamiento obtuvo fue con el método de eliminación de outliers LOF y SVM para la variable respuesta EED, y con el método SVM para la variable respuesta ELD. Por tanto, se opta por utilizar estos dos métodos para continuar con las siguientes etapas:

Continuando con el proceso, procedemos a implementar la técnica de reducción de dimensionalidad Robust PCA (ROBPCA) en nuestro conjunto de datos dfModelo. Realizamos un análisis utilizando 2, 3, 4, 5 y 6 componentes principales y visualizamos el impacto en el desempeño entrenando los modelos de referencia XGBoost y usando la métrica KS para diferentes valores del parámetro max_depth, como se observa en la Figura 8. Finalmente, compararemos estos resultados con los entrenamientos realizados en las mismas bases que

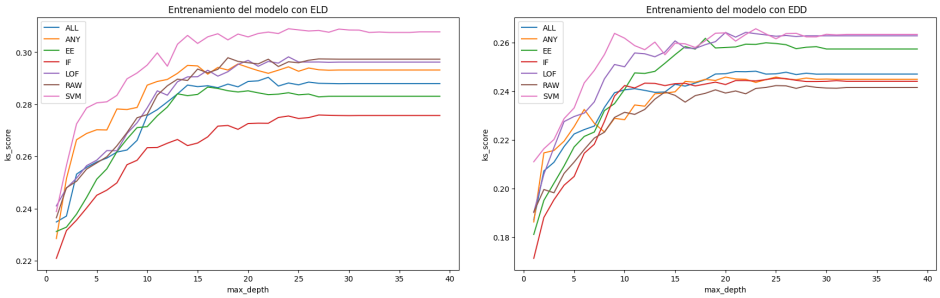

Figura 7: Resultado de entrenamiento del modelo con dfModelo, dfModeloNorm y dfModeloStd

no pasaron por reducción de dimensionalidad.

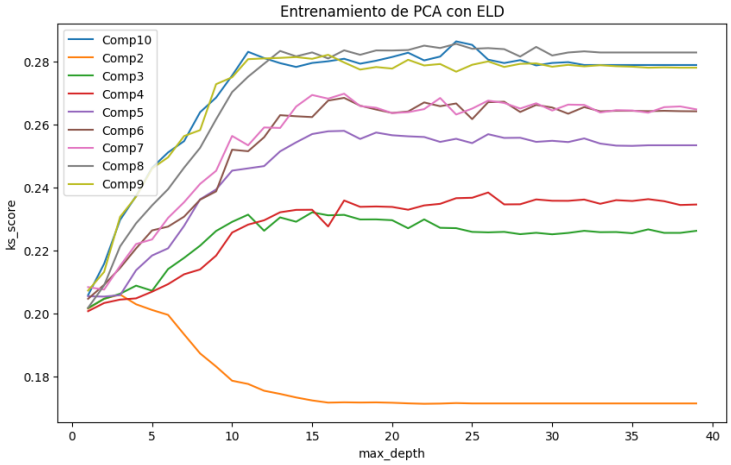

Figura 8: Resultado de entrenamiento del modelo de 2 a 10 PCA

El mejor comportamiento del modelo se dio con 8 componentes, en el cual se obtuvo un KS máximo de 0.28 con un max_depth de 12. Basado en este resultado, y teniendo en cuenta que las variables de entrada al modelo son 11 y con un KS máximo de 0.31, y que los tiempos de entrenamiento son similares, se descarta el uso de ROBPCA para el entrenamiento del modelo final. En busca de optimizar el entrenamiento del modelo, se decide utilizar la base de datos dfModeloLOF, que corresponde a la base con eliminación de atípicos por LOF, y la variable respuesta ELD, que corresponde a Mora 90 en 9 meses. Mediante un proceso iterativo, se buscarán los parámetros, descritos en el Cuadro 4, que maximicen el estimador KS y la métrica F1.

Una vez finalizado el proceso de entrenamiento, se generó una muestra aleatoria de 65,000 parámetros en los rangos descritos en el Cuadro 4 y se analizó el comportamiento del resultado con el fin de identificar el intervalo de confianza en el cual dicha combinación de parámetros maximizará el estimador KS y el F1-Score (Figura 9).

parámetro	Rango
n_estimators	10-100
max_depth	5-35
learning_rate	1-10
subsample	0.1-1
colsample_bytree	0.1-1

Cuadro 4: Cantidad de variables Outliers compartidas entre cada método

n_estimators: Hace referencia a la cantidad de árboles o estimadores a construir durante el entrenamiento. Cuanto más alto sea este valor, más complejo será el modelo y más tiempo durará el entrenamiento. Es importante notar que un número elevado de árboles podría llevar al sobreajuste del modelo.

max_depth: Este parámetro controla la profundidad máxima de cada árbol, limitando la cantidad de nodos que se pueden crear desde el nodo raíz hasta las hojas. Controlar este valor ayuda a equilibrar la complejidad del modelo y buscar un equilibrio entre el sesgo y la varianza.

learning_rate: La tasa de aprendizaje controla la contribución de cada árbol al modelo final. Cuanto más bajo sea este valor, más árboles se requerirán para la optimización, lo que podría llevar a tiempos de entrenamiento mayores y a posibles sobreajustes.

subsample: Especifica la proporción de instancias de entrenamiento que se utilizarán para construir cada árbol. Un valor menor reduce la correlación entre los árboles y puede ayudar a prevenir el sobreajuste.

colsample_bytree: Corresponde a la proporción de características (columnas) que se utilizarán para construir cada árbol. Un valor menor reduce la correlación entre los árboles y ayuda a prevenir el sobreajuste.

El resultado mostró que el valor máximo de KS tiende a 0.31 El resultado mostró que el valor máximo de $F1 - Score$ tiende a 0.51. Los parámetros con valores mínimos para cumplir estas condiciones se especifican en el Cuadro 5

parámetro	Valor
n_estimators	60
max_depth	30
learning_rate	0.03
subsample	0.8
colsample_bytree	0.7

Cuadro 5: Parámetros del entrenamiento para un máximo de KS y F1

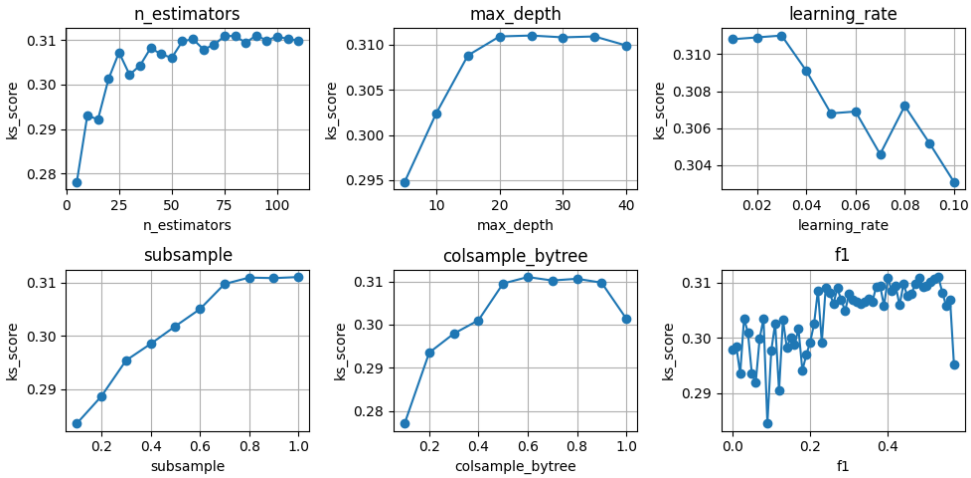

Figura 9: Obtimización de los parámetros para maximo ks y f1-score

Para validar la estabilidad del modelo con los parámetros establecidos en el Cuadro 5, se decidió realizar la prueba con 1,000 muestras aleatorias, utilizando el 50 % de los datos para entrenamiento y el 50 % de los datos para pruebas, como se observa en la Figura 10. Los resultados mostraron estabilidad en los valores de KS y F1-Score, según se puede observar en el Cuadro 6. Esto indica que el modelo mantiene su estabilidad independientemente del grupo de datos utilizado para entrenamiento y pruebas al momento de entrenar el modelo con los parámetros propuestos. En el Cuadro 7 podemos ver la importancia que tiene cada variable en la estimación del score. Los resultados muestran que la variable que más pesa en la estimación del score es el género. La distribución del valor por mil o score de los clientes que se defaultearon o tocaron ELD `(M90@9)` y de los clientes que tuvieron buen comportamiento se puede ver en la Figura 11.

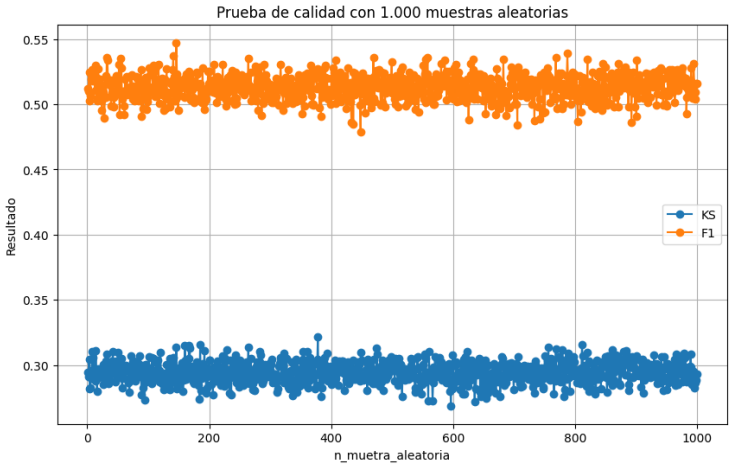

Figura 10: Prueba aleatoria con diferentes bases de entrenamiento y prueba

Medida	KS	F1
Media	0.294442	0.513145
Mediana	0.294723	0.513580
Moda	0.268790	0.500000
DesvStd	0.007813	0.009286
Mínimo	0.268790	0.478764
Máximo	0.321513	0.547327

Cuadro 6: Medidas estadísticas para las variables KS y F1

Variable	Importancia
genero	36.26 %
Edad	5.78 %
TipoEmpleo	7.74 %
PeriodicidadPago	5.03 %
PersonasCargo	5.42 %
NivelEstudios	7.15 %
TipoVivienda	5.17 %
Permanencia	5.71 %
EstadoCivil	5.50 %
salario	8.77 %
MetodoPago	7.47 %

Cuadro 7: Importancia de las variables de entrenamiento

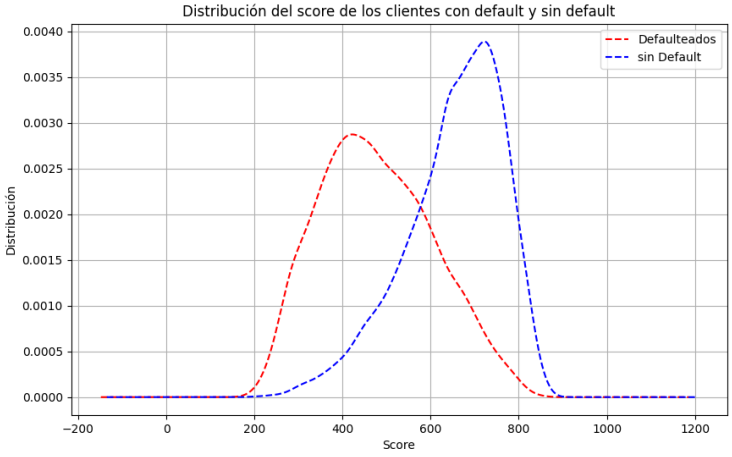

Figura 11: Distribución del score de los clientes con y sin default y

Las compañías de financiamiento cuentan con un apetito de riesgo, en el cual definen qué porcentaje de su cartera puede caer en default. Basado en la Figura 11, podríamos estimar la probabilidad de impago de un cliente a partir del score obtenido. En la Figura12, se muestra que para un escenario de apetito de riesgo del 20 %, el punto de corte recomendado para aprobar a los clientes debe ser como mínimo de 700.

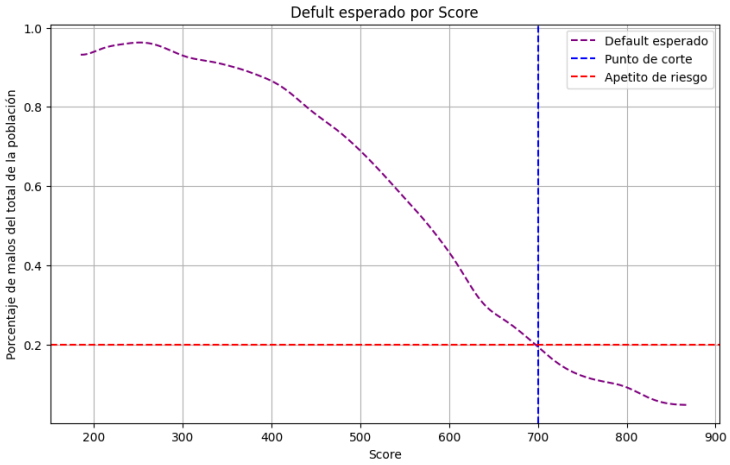

Figura 12: Curva Lift: Distribución del score (porMil) de los clientes con y sin default

4. Conclusión

El uso de modelos como el XGBoost para el diseño e implementación de modelos de score de crédito ha demostrado dar resultados positivos que aportan a mantener los niveles de riesgo dentro de los apetitos de la organización. Este trabajo permitió diseñar un modelo tipo XGBoost para estimar el score de crédito de una población propia del sector Fintech utilizando diferentes tipos de variables, entre ellas sociodemográficas, variables de empleo e información de obligaciones financieras, logrando obtener finalmente 11 variables explicativas y una variable respuesta correspondiente al *Default* Mora 90 en 9 meses. Los resultados obtenidos permitieron identificar la probabilidad de caer en default partiendo del *score* obtenido, y con ello la compañía de financiamiento tiene la posibilidad de adaptar el punto de corte de los clientes dependiendo de su apetito de riesgo.

Para el entrenamiento del modelo propuesto se construyeron bases de datos que fueron comparadas entre ellas en un primer entrenamiento. Estas bases correspondieron a muestras de una base de datos original e incluyeron 5 bases de datos con datos eliminados de *outliers* mediante diferentes metodologías, y 8 bases de datos con reducción de dimensionalidad mediante modelos robustos PCA (ROBPCA) desde 2 hasta 8 componentes. Una vez estas bases se pusieron a competir mediante un entrenamiento con los mismos parámetros, logramos obtener una base de datos óptima para el entrenamiento, la cual fue la base con eliminación de *outliers* mediante la metodología LOF, logrando un mayor nivel de KS y F1-Score.

Para definir los parámetros óptimos, se utilizó una simulación Montecarlo de 6,000 muestras, logrando maximizar su resultado y prevenir el sobreajuste. Esta muestra fue suficiente para estimar los parámetros óptimos de entrenamiento. Bajo la metodología de

entrenamiento propuesta, se llegó a un modelo que, aunque presenta valores del estimador KS relativamente bajos, se espera que bajo el escenario de apetito de riesgo propuesto, el porcentaje de aprobación ronde el 15 %, porcentajes normales si se tiene en cuenta que naturalmente es un público de riesgo que atiende la industria Fintech.

Como caso particular del modelo, la variable que más peso tiene es el género, inclinándose a generar mejor score en mujeres que en hombres. Esto puede indicar que tanto hombres como mujeres son dos grupos poblacionales que se comportan de manera diferente respecto al default, dependiendo de las variables explicativas. Una posible forma de mejorar la estimación y segmentación del modelo es generar dos entrenamientos, uno para hombres y otro para mujeres. Este tipo de modelos, que utilizan información proveniente de diferentes fuentes, son altamente sensibles a fallas debido a la caída de alguno de los proveedores de información. Por tanto, se deben tomar medidas que mitiguen el riesgo frente a datos faltantes por esta causa.

En ambientes productivos, este tipo de modelos son reforzados por filtros duros que, en muchas ocasiones, son definiciones de negocio y en otros casos obedecen a condiciones de los clientes obtenidas de otras fuentes no consideradas dentro del modelo. Estas fuentes no siempre son consultadas y se utilizan como refuerzo de la decisión. Por ejemplo, clientes desempleados, con cuentas embargadas o con procesos judiciales activos. Aunque un modelo puede segmentarlos con un score alto, puede ser necesario una no aprobación debido a condiciones de riesgo no especificadas en el modelo original y que también pueden ser consecuencia de decisiones dentro de la estrategia de la organización.

Aunque la variable respuesta se seleccionó partiendo de la probabilidad que tiene el modelo para segmentarla, también existen otras metodologías para seleccionar la variable respuesta. Entre ellas está el utilizar la curva de deterioro por altura de mora y considerar el momento en que la cosecha se estabilice en su deterioro. En tal caso, el modelo predecirá la probabilidad de deteriorarse antes de que la cosecha madure a una altura de mora determinada.

Más allá de los avances logrados, durante el estudio se identificaron algunas limitaciones, las cuales se pueden resumir en la alta sensibilidad del modelo a la calidad de los datos y la disponibilidad de los mismos, pues la ejecución del modelo depende de la disponibilidad de los datos provenientes de diferentes fuentes externas, y la caída de alguno de estos servicios haría inoperante el modelo. Otra limitante importante a tener en cuenta es la relevancia de las variables utilizadas, pues los cambios macroeconómicos y sociales pueden acelerar la descalibración del modelo, lo cual podría requerir ajustes continuos.

Frente a las diferentes situaciones y limitaciones que pueda tener el modelo, se dan algunas recomendaciones que permitirán garantizar su confiabilidad, las cuales se resumen en monitorear permanentemente los datos para garantizar la calidad de las fuentes de datos externas y revisar periódicamente las variables para asegurar su relevancia y ajuste frente a las condiciones actuales del mercado. Por último, como recomendación para trabajos

futuros de investigación en el tema, se propone desarrollar modelos segmentados por género y evaluar otros criterios relevantes para mejorar la precisión y equidad en la evaluación de los clientes, incorporar nuevas fuentes de datos no tradicionales que aporten a la robustez del modelo y aplicar otras técnicas de *machine learning* que permitan probar y comparar el rendimiento del modelo.

Referencias

Aggarwal, C. C. (2017), *Outlier Analysis*, Springer.

Banco Interamericano de Desarrollo (2022), "Fintech en América Latina y el Caribe: un ecosistema consolidado para la recuperación," *BID Invest*, 1, 1 – 10.

Banco Mundial (2022), "La inclusión financiera es un elemento facilitador clave para reducir la pobreza y promover la prosperidad," *Banco Mundial*, 1, 1 – 3.

Becker, R., Chambers, J., and Wilks, A. (1988), *The New S Language: A Programming Environment for Data Analysis and Graphics*, Wadsworth Brooks/Cole Advanced Books Software.

Breunig, M., Kriegel, H.-P., Ng, R., and Sander, J. (2000), "LOF: Identifying Density-based Local Outliers," *ACM SIGMOD Conference*, 1, 9 – 10.

Chen, T. and Guestrin, C. (2016), "XGBoost: A Scalable Tree Boosting System," *The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 978-1-4503-4232-2, 1 – 10.

Cormann, M. and Scarpetta, S. (????), "OECD Employment Outlook 2023, Artificial Intelligence and the Labour Market," .

Costa, A., Deb, A., and Kubzansky, M. (2015), "Big Data, Small Credit: The Digital Revolution and Its Impact on Emerging Market Consumers," *MIT Press*, 10, 49 – 80.

Crock, J., Edelman, D., and Thomas, L. (2007), "Recent development in consumer credit risk assessment," *European Journal of Operational Reserch*, 3, 1447 – 1465.

Departamento Administrativo Nacional de Estadística (2023), "Indicadores de mercado laboral, Diciembre y año total 2023," *DANE*, COM-070-PDT-001-f-001, 1 – 10.

Elizondo, A. and Lopez, C. (1999), "El riesgo de crédito en México: Una evaluación de modelos recientes para cuantificarlo," *Gaceta de economía*, 4, 51 – 74.

Field, Andy, J. M. and Field, Z. (2012), *Discovering Statistics Using R*, SAGE Publications Ltd.

Finnovista (2023), "El número de empresas Fintech locales en Colombia asciende hasta las 369 para este primer cuatrimestre de 2023," *Finnovista*, 1, 5 – 7.

Huang, C., Chen, M., and Wang, C. (2007), "Credit Scoring with a data mining approach based on support vector machines," ., 1.

Hubert, M. and J.Rousseeuw, P. (2005), "ROBPCA. A New Approach to Robust Principal Component Analysis," *American Statistical Association and the American Society for Quality*, 47, 49 – 80.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2017), *Introduction to Statistical Learning: with Applications in R*, Springer Texts in Statistics.

Johnson, R. A. and Wichern, D. W. (2002), *Applied Multivariate Statistical Analysis*, Prentice Hall.

Lee, Y. and Scolari, A. (2010), "Toward High-Performance Perdiction Serving System," ., 1, 15 – 25.

Lessmann, S., Baesens, B andSeow, H., and Thomas, L. (2015), "Benchmarking state of the art classification algorithms for credit scoring," *European Journal of Operational*, 247, 124 – 136.

Malagón González, J., , Vera Sandoval, A., and Montoya Moreno, G. (2024), "Perspectivas crediticias para 2024: un año de lenta recuperación," *ASOBANCARIA*, 1409, 6 – 8.

Min, J. and Lee, Y. (2005), "Bankruptcy prediction using support vector machine with optimal choice of kerner function parameters," *Expert System with Applications*, 28, 603 – 614.

Rodríguez Benavides, D. and Perrotini Hernández, I. (2019), "Las correlaciones dinámicas de contagio financiero: Estados Unidos y América Latina," *Revista Mexicana de Economía y Finanzas*, 14, 151–168.

Rodríguez-Novoa, D., Najar, A. R., and Gómez, A. J. F. S. (2023), "Reporte de la situación de crédito en Colombia," *Banco de la republica*, 1, 27 – 29.

Sepúlveda, C., Reina, W., and Gutíerrez, J. (2012), "Estimación del riesgo de crédito en empresas del sector real en Colombia," *Estudios Gerenciales*, 28, 169 – 190.

Thomas, L. C., Edelman, D. B., and Crook, J. N. (2002), *Credit Scoring and Its Applications*, Society for Industrial and Applied Mathematics SIAM.

West, D. (2000), "Neural network credit scoring models," *Computer and Operation Reserch*, 27, 1131 – 1151.