Revista de investigación de sistemas e informática
16(2): 155-168 (2023)
DOI: https://doi.org/10.15381/risi.v16i2.27140

ISSN: 1815-0268 (Impreso) / ISSN: 1816-3823 (Electrónico)
Facultad de Ingeniería de Sistemas e Informática - UNMSM

Modelo de Machine Learning basado en la Maquina Potenciadora de Gradiente de Luz para predecir la probabilidad de impago en clientes de la cartera de Tarjeta de Crédito

Machine Learning model based on the Light Gradient Boosting Machine to predict the probability of default in customers of the Credit Card portfolio

Eduardo Rafael Jáuregui Romero
https://orcid.org/0009-0008-0248-2601
eduardo.jauregui1@unmsm.edu.pe
Universidad Nacional Mayor de San Marcos, Lima, Perú

Recibido: 31/10/2023 - Aceptado: 25/11/2023 - Publicado: 30/12/2023

RESUMEN

Los créditos bancarios son un medio de pago muy utilizado en los últimos tiempos, cada vez más personas acceden a productos como tarjeta de crédito, prestamos, etc. Los bancos han implementado modelos clásicos de predicción, la gran mayoría basados en regresión logística ya que permite una gran interpretabilidad de cara al negocio y efecto de las variables del modelo. El propósito de esta investigación es realizar un análisis predictivo sobre la probabilidad de impago de clientes en la cartera de tarjeta de crédito mediante un score de riesgo. El dataset utilizado es el denominado default of credit card clients Data Set proveniente de la BD de UCI Machine Learning, el enfoque es cuantitativo y la metodología es analítica descriptiva, se utilizará técnicas basada en potenciadores de gradiente para realizar la predicción, entre los algoritmos entrenados tenemos Regresión Logística con WOE, CatBoost, XGBoost y LigthGBM, además para poder suplir la falta de interpretabilidad de los algoritmos de Machine Learning se utilizará en enfoque basado en Importancia Gain y Shapley para medir el impacto de las variables en la predicción. Como resultado se obtuvo la máquina potenciadora de gradientes de luz (LightGBM) tuneada con una búsqueda bayesiana obtuvo un GINI de 57.4 lo cual mejora en +6 puntos a la Regresión Logística con Woe y en +3p a XgBoost y CatBoost. Finalmente obteniendo los valores de Gain y Shapley se suplió la falta de interpretabilidad de las variables, permitiendo una mejor toma de decisiones a la hora de evaluar a los clientes. Así mismo como trabajos futuros se pretende agregar variables no estructuradas que permitan mejorar los indicadores del Modelo.

Palabras clave: Riesgo de crédito, aprendizaje automático, análisis de datos, probabilidad de incumplimiento, tarjeta de crédito, comprobante de pago.

ABSTRACT

Bank loans are a widely used means of payment in recent times, more and more people are accessing products such as credit cards, loans, etc. Banks have implemented classic prediction models, the vast majority based on logistic regression since it allows great interpretability for the business and the effect of the model variables. The purpose of this research is to perform a predictive analysis on the probability of customer default in the credit card portfolio using a risk score. The dataset used is the so-called default of credit card clients Data Set from the UCI Machine Learning DB, the approach is quantitative and the methodology is descriptive analytics, techniques based on gradient boosters will be used to make the prediction, among the trained algorithms We have Logistic Regression with WOE, CatBoost, As a result, the light gradient enhancement machine (LightGBM) tuned with a Bayesian search was obtained, obtaining a GINI of 57.4, which improves by +6 points to the Logistic Regression with Woe and by +3p to XgBoost and CatBoost. Finally, obtaining the Gain and Shapley values made up for the lack of interpretability of the variables, allowing better decision making when evaluating clients. Likewise, as future work, it is intended to add unstructured variables that allow the Model's indicators to be improved.

Keywords: Credit risk, machine learning, data analysis, probability of default, credit card, proof of payment.

I. INTRODUCCIÓN

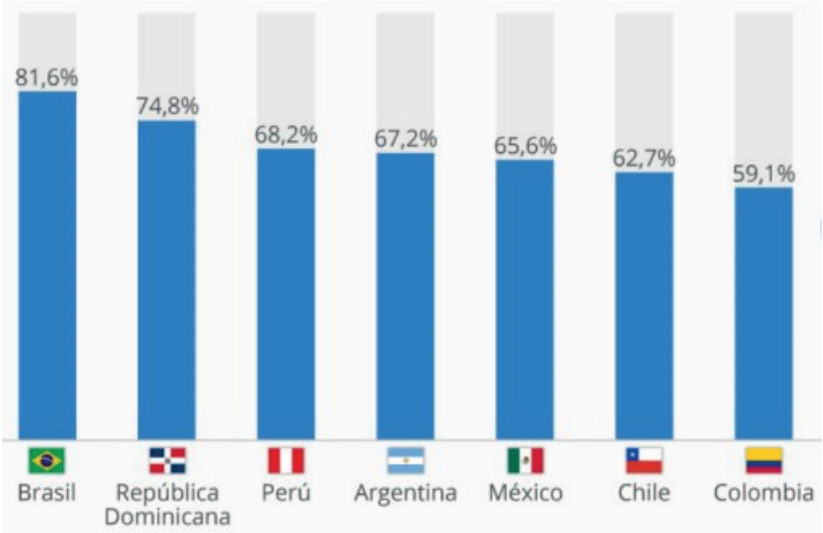
En la actualidad es muy común que las personas puedan acceder cada vez más a productos financieros tales como préstamos personales, préstamos hipotecarios, tarjeta de créditos, entre otros. Según El País el 71% de los adultos cuentan con productos financieros, + 29pp por encima en comparación de hace 10 años. La tarjeta de crédito es uno de los productos más usados, cerca del 40% de los peruanos ya usan este producto como su medio de pago principal. En los últimos años existe una tendencia en aumento del uso de la tarjeta de crédito, el Perú se ubica en el tercer puesto (68.2%) en él % de población que uso su tarjeta de crédito para pagar productos a plazo, tal como lo refleja la Figura 1.

Debido a esto se ha convertido en un requisito indispensable para los bancos poder controlar el incumplimiento de sus clientes en las diversas carteras y sobre todo en una de las más importantes como es la cartera de tarjeta de crédito.

2. DEFINICIÓN DEL PROBLEMA

El problema radica en poder detectar la probabilidad de incumplimiento de clientes en la cartera de crédito, debido a que esta cartera está en creciente aumento a nivel de clientes y también representa uno de los principales ingresos para cualquier banco. Según la Republica en el Perú aproximadamente 8.7 millones de personas tienen un crédito con alguna institución financiera. Así como aumento el número de clientes, también aumenta el riesgo

Figura 1.
Porcentaje de población que usa T.C para consumo en cuotas.



crediticio que asume el banco, cerca de 406 mil tarjetahabientes registran casos de atrasos en sus deudas, la mora se elevó especialmente en jóvenes de 18 – 24 años.

Esta es una situación alarmante ya que los bancos deben estar más atentos a la hora de evaluar tanto la admisión a la tarjeta de créditos (para no entregar tarjetas a personas con alta probabilidad de incumplimiento) como para el comportamiento (regular su cartera de clientes) para disminuir líneas de créditos o no ofrecer nuevos productos a personas en zonas de alto riesgo crediticio. Según La República en tiempos de pandemia algunas regiones del Perú llegaron a tener cifras alarmantes de deuda, por ejemplo, en Tacna el promedio de mora por persona era de S/, 1690, llegando a tener una deuda impaga total de cerca de 468 millones de soles, esto es 17% más con respecto a 2019. De hecho, en este último año se registró que en total hay aproximadamente 8.5 millones de peruanos morosos, una cifra muy por encima de la que se manejaba antes de la pandemia. Se estima que cada persona en el Perú tiene una deuda de 3668 soles, en donde Lima concentra la mayor cantidad de morosos con un total de 15.242 millones de soles. Por lo tanto es primordial que toda institución financiera pueda controlar y segmentar a sus clientes de las diferentes carteras en base a modelos de machine Learning, bajo este contexto este proyecto busca calcular la probabilidad de impago también llamado mora, default de los clientes para la cartera de tarjeta crédito, de esta manera se podrá identificar clientes con alta probabilidad de riesgo crediticio y realizar acciones específicas, de igual manera se podrá identificar clientes con baja probabilidad para ofrecerles nuevos productos o aumentar los que ya tienen disponible. Finalmente vale recalcar que este modelo puede ser extensible a otras carteras como préstamos personales, préstamos hipotecarios, etc.

3. REVISIÓN DE LITERATURA

El trabajo relacionado en el dominio del credit scoring basado en machine Learning se revisa brevemente en este análisis. El uso del Machine Learning llamo la atención de la industria bancaria al ser una herramienta basada en estadística y matemática avanzada para obtener métricas de riesgo crediticio más precisas que enfoques estadísticos tradicionales. Las técnicas tradicionales más comunes son las de regresión logística gracias a la interpretabilidad que ofrece a la hora de ver el impacto de las variables y su movimiento respecto a la predicción. Según Abdou (2019) la clasificación crediticia es ampliamente utilizada en los sectores bancarios.

En 2017, Torres y Farroñay plantearon resolver el problema del incremento en la tasa de morosidad durante el 2014, basados en la metodología CRISP-DM y algoritmos de Minería de Datos como descripción de clases se logró implementar una solución de inteligencia de negocio en donde se pudo detectar patrones generales de clientes morosos respecto a clientes normales. Así mismo se llegó a la conclusión de que la técnica de Descripción de clases permite identificar de mejor manera los patrones de comportamiento y logrando analizar los perfiles de los clientes morosos que provocaron el incremento de morosidad de 10 pp por encima del periodo anterior.

De la misma manera en 2018, Fernández con el objetivo de conocer los perfiles de clientes que son más aptos a caer en morosidad para la tarjeta de crédito desarrollo diversos modelos tales como regresión bayesiana con enlaces asimétricos clocog, power logit y scobit con una muestra de clientes del banco que tenía como target a la variable mora60 que indicaba si los clientes caen en default con más de 60 días en el transcurso del año. de esta manera se pudo llegar a la conclusión de que el algoritmo basado en regresión binaria bayesiano con enlace asimétrico cloglog fue el más apto para evaluar a estos clientes logrando situarse por cerca de 9 pp por encima vs modelos power logit y scobit.

Por otro lado, los estudios realizados por Garcia y Zhichao et al. ya utilizan enfoques basados en Machine Learning, en ambos estudios queda evidenciado que los modelos basados tanto en XgBoost como en MLP son mucho más precisos que enfoques tradicionales de regresión logística, pero también señalan una gran debilidad de estos enfoques, la cual es la falta de interpretación de las variables y la dificultad de explicar a los usuarios internos la aplicabilidad del modelo.

Por último, en 2023, Slabber et al. desarrollaron un modelo de clasificación binaria utilizando máquinas de factorización basado en una rutina de ajuste con perdida logit y máxima verosimilitud. Este algoritmo permite mediante 3 simulaciones en escenarios distintos, comparar sus métricas vs la regresión logística, Random forest. Como resultado principal se evidencia que el LFM2 obtuvo un resultado a nivel de GINI para el train sin aplicar WOE de 0.797 lo cual es superior a la regresión logística en +0.025 (0.772 de Gini), permitiendo así que el LFM2 sea el modelo con un mejor balance entre performance, estabilidad e interpretabilidad.

IV. MÉTODOLOGIA

La investigación se desarrollará utilizando la metodología CRISP-DM (Cross Industry Standard Process for Data Mining). Esta metodología proporciona un marco estructurado para el proceso de minería de datos y análisis de datos. Se utiliza ampliamente en la industria para abordar problemas complejos y extraer información valiosa de los conjuntos de datos. Además, esta metodología es más accesible y fácil de entender para personas no especialista en aprendizaje automático, tal y como lo muestra la Figura 2

Figura 2.
Fases de la metodología CRISP-DM

Según Clark (2018) al seguir la metodología CRISP-DM, se espera obtener resultados sólidos y confiables en el análisis de los datos relacionados con la probabilidad de impago en tarjetas de crédito. Esto permitirá una comprensión más profunda de los factores que influyen en el impago y facilitará la toma de decisiones informadas en la gestión de riesgos crediticios.

A. Set de Datos y Entendimiento del Negocio

El set de datos a utilizar proviene del portal UCI Machine Learning Repository , que es una recopilación de información relacionada al comportamiento de pago de clientes en una institución financiera en Taiwan. El conjunto de datos contiene información demográfica, características crediticias, genero, estado civil, nivel educativo, historial de pagos, límites de crédito y pagos realizados en los últimos 6 meses. Se está utilizando este dataset ya que contiene información de los clientes desde diferentes enfoques y no solo a nivel crediticio. Vale recalcar que este conjunto de datos tiene 30000 instancia para modelar. En la tabla 1 se detallan las variables disponibles en el conjunto de datos.

Table 1.
Variables del conjunto de datos default of credit card clients Data Set

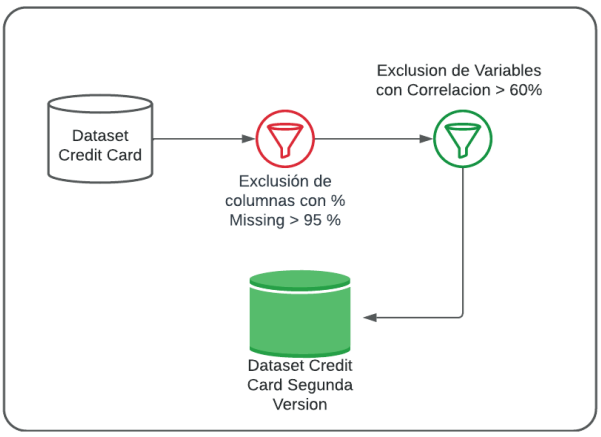
Nombre de Columna	Descripción
Monto Crédito Otorgado	Monto de Crédito cargado en dólares
genero	Genero
educación	Nivel Educativo
estado civil	Estado Civil,
Edad	Edad del cliente
ep_09 – ep_04	Estado de pagos desde 200509 hasta 200504
ec_09 – ec_04	Estado de Cuenta desde 200509 hasta 200504
mp_09 – mp04	Monto pagado desde 200509 hasta 200504
target	Default próximo mes, SI = 1 , NO = 0

Como se evidencia en la tabla este dataset contiene tanto información sociodemográfica como comportamental crediticia, lo cual nos permitirá realizar un amplio análisis exploratorio de datos. Adicionalmente este conjunto de datos cuenta con una columna que servirá como la variable objetivo, también conocido como target, la cual está representada por 2 valores, 1 = Cliente cumplió con su próximo pago y 0 = cliente no cumplió con su siguiente pago.

B. Preparación de Datos

Antes de empezar con el entrenamiento del modelo realizaremos una serie de filtros que tienen como objetivo poder obtener la mejor calidad de Datos. se realizarán los pasos tal y como lo detalla la Figura 3.

Figura 3.
Fases del Preprocesamiento de Datos



Durante el primer filtro se pudo observar que todos los atributos contaban con valores no nulos, por lo tanto, no se procedió a eliminar ninguna columna. Tal y como lo demuestra la Tabla 2, se renombraron las columnas con sus respectivos nombres, también se procedió a calcular indicadores estadísticos que nos ayuden a conocer la distribución de las variables. Se puede notar que las variables relacionadas al estado de cuenta tienen una distribución con sesgo, tal y como lo muestra la Figura 4.

En todo dataset hay variables que tienen una correlación fuerte, por lo tanto, debemos eliminar estas variables ya que serán redundantes en el modelado, a continuación, mostraremos la matriz de correlación para las variables numéricas, estamos excluyendo las variables ordinales ya que estas serán analizadas a nivel de importancia de variables. En la Figura 5 se muestra la matriz de correlación para las variables de estado de pago, estado de cuenta y monto pagado.

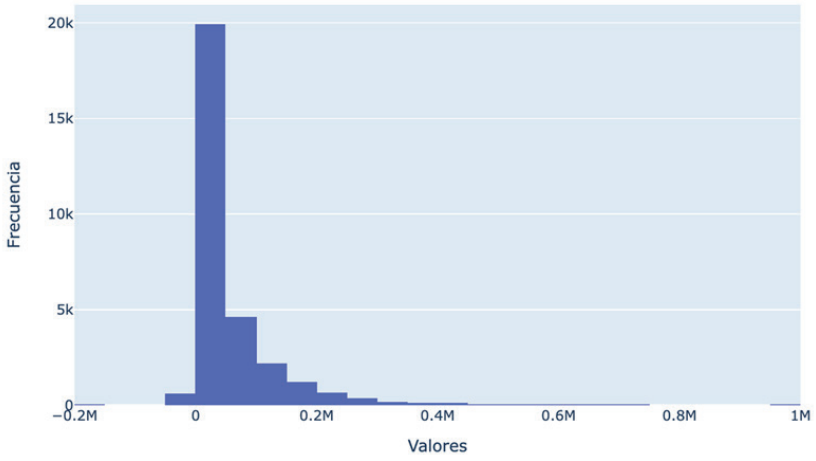
Table 2.

Análisis Univariado de las Variables del Dataset

Variables	Total Filas	Media	Mediana	std	Valores Nulos	Porcentaje de Nulos
Id_Cliente	30000	15	15	8.66	0	0%
monto_credito_otorgado	30000	167.85	140	129747	0	0%
genero	30000	1	2	0	0	0%
educacion	30000	1	2	0	0	0%
estado_civil	30000	1	2	0	0	0%
edad	30000	35	34	9	0	0%
ep_09	30000	0	0	1	0	0%
ep_08	30000	0	0	1	0	0%
ep_07	30000	0	0	1	0	0%
ep_06	30000	0	0	1	0	0%
ep_05	30000	0	0	1	0	0%
ep_04	30000	0	0	1	0	0%
ec_09	30000	51.223	22.381	73.635	0	0%
ec_08	30000	49.179	21.2	71.173	0	0%
ec_07	30000	47.013	20.088	69.349	0	0%
ec_06	30000	43.262	19.052	64.332	0	0%
ec_05	30000	40.311	18.104	60.797	0	0%
ec_04	30000	38.871	17.071	59.554	0	0%
mp_09	30000	5.663	2.1	16.563	0	0%
mp_08	30000	5.921	2.009	23.04	0	0%
mp_07	30000	5.225	1.8	17.606	0	0%
mp_06	30000	4.826	1.5	15.666	0	0%
mp_05	30000	4.799	1.5	15.278	0	0%
mp_04	30000	5.215	1.5	17.77	0	0%
default_p_m	30000	0	0	0	0	0%

Figura 4.

Histograma del Estado de Cuenta en septiembre del 2005



Se realizará un filtrado de variables, eliminando las correlaciones mayores a 60%, para eso se construye una matriz triangular y se procede a escoger el atributo más reciente en caso de contar con correlaciones múltiples. Una vez realizado este proceso logramos reducir las variables numéricas a 9 variables, tal y como lo muestra la Figura 6.

Al finalizar este procesamiento de datos, nuestro dataset cuenta con 12 variables de las 24 iniciales, se lograron eliminar variables redundantes para el modelo que no aportaran mayor precisión a la hora de ver los resultados.

C. Transformación WOE

Según Disha (2022) la técnica WOE (1) nos permitirá realizar una transformación poderosa, logrando aumentar el rendimiento de los Modelos de regresión Logística evitando el tema de outliers y

Figura 5.
Matriz de Correlación de las Variables Numéricas Originales

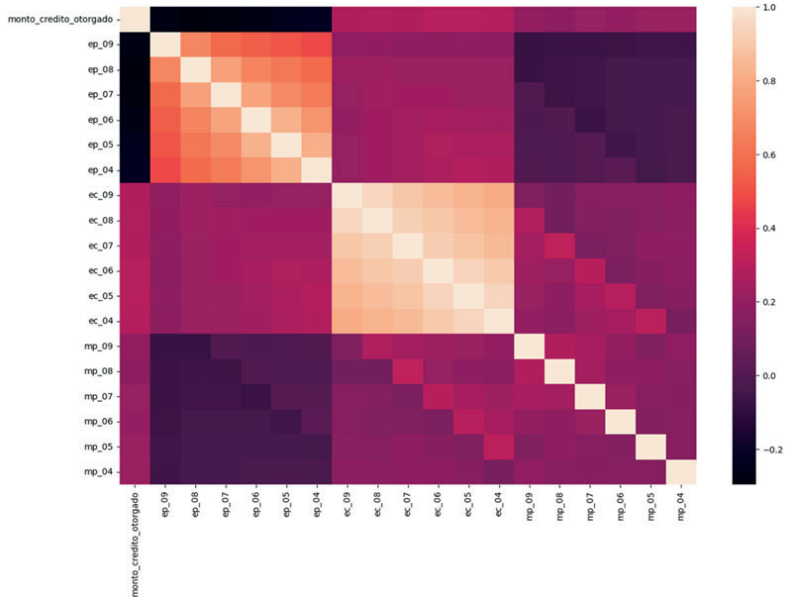
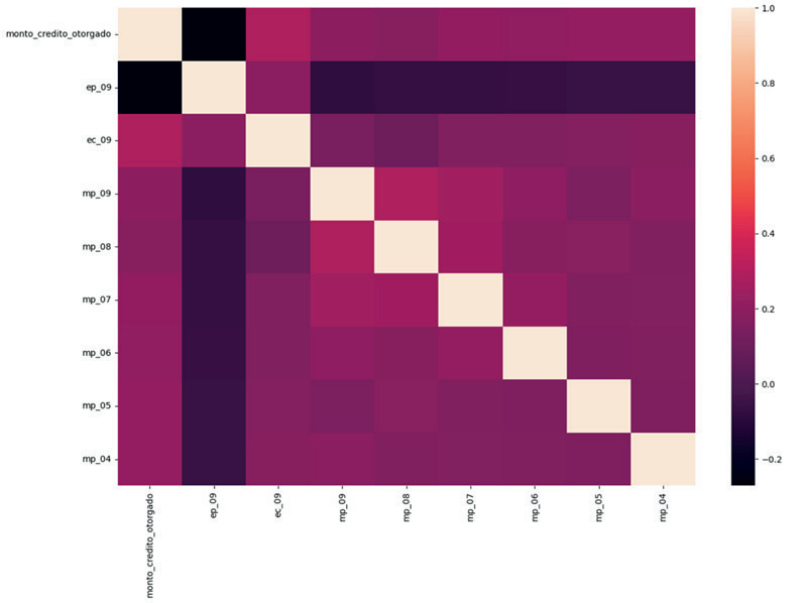


Figura 6.
Matriz de Correlación de las Variables filtradas.



distribuciones sesgadas. Para este proceso aplicaremos la siguiente formula, donde Default es no pagar el siguiente mes es decir target = 1:

$$Woe = \ln(\frac{\%Default}{\%NoDefault}) \quad (1)$$

$$IV = Woe * \sum(\%Default - \%NoDefault) \quad (2)$$

Según la ecuación (2), el I.V nos permite seleccionar variables de manera correcta, variables con valor menor a 0.02 no son buenas para la predicción, así como también las de valor superior a 0.5 . Realizamos la transformación WOE lo cual nos permitirá realizar agrupaciones de nuestras variables en función del target, de esta manera podemos analizar si se tiene un sentido de interpretación correcto para el problema. Tal y como lo demuestra la Figura 7, se realizó la transformación WOE para la variable ep_09. Este grafico nos permite interpretar la variable de una manera mucho más sencilla, vemos que a medida que aumenta el valor del ep_09 (estado del pago, mes anterior, -1 = pago a tiempo, 1 retraso de un mes … 9 = retraso de 9 o más meses) existe una mayor probabilidad de caer en default el siguiente mes.

Todos los valores del atributo que estén contenido en el BIN serán cambiados por su valor WOE a través de la fórmula de transformación de logaritmos de eventos y no eventos, tal y como lo muestra la Tabla 3.

Aquellos atributos que tengan un IV menor a 0.02 no son significativos a la hora de modelar, por lo

Figura 7.
Gráfico Bivariado WOE Estado de Pago Mes anterior (ep_09) y Target

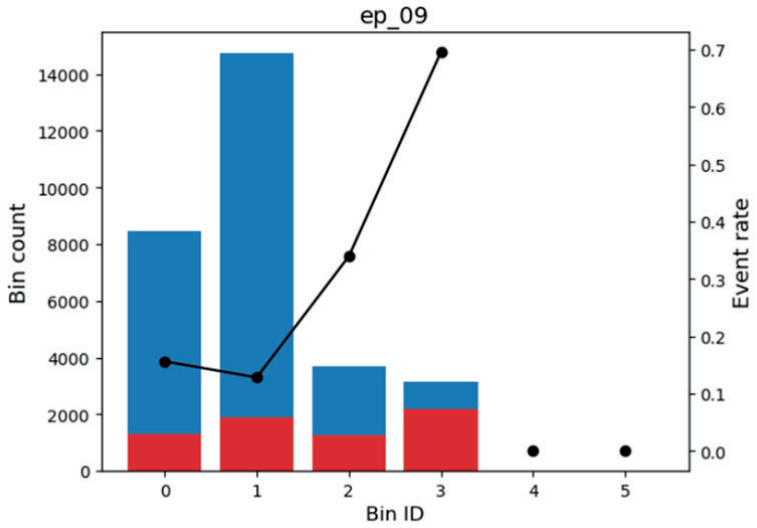


Table 3.
Transformación WOE de variable ep_09

BIN	Count	% Count	Not Event	Event	Woe	Event Rate
-inf, -0.5	8445	28%	7126	1319	0.42	15.60%
-0.5 , 0.5	14737	50%	12849	1888	0.65	12.80%
0.5 , 1.5	3688	12%	2436	1252	-0.59	33.90%
1.5 , +inf	3130	10%	953	2177	-2.08	69.60%

Table 4.
Variables con IV menor a 0.02

VAR	IV	GINI
Género	0.009	0.047
Educación	0.015	0.060
Estado Civil	0.004	0.034
ec_09	0.009	0.051

tanto, se procede a eliminarlas. Las variables para retirar son las de género, educación, estado civil y ec_09, tal y como lo refleja la Tabla 4

D. Regresión Logística con WOE

Se realizará el entrenamiento del Modelo de Regresión Logística con la Librería sklearn y el módulo linear_model. Miranda (2017) indica que este método es muy utilizado gracias a su interpretabilidad. Para este artículo el conjunto de entrenamiento se separa del conjunto de datos en una división de 70% entrenamiento y 30% como muestra fuera de tiempo (OOT). El modelo entrenado nos da como resultado una CURVA ROC de 0.76, llevado a indicador GINI es un valor de 52.35 para entrenamiento y 51.03 para OOT. Este resultado se puede apreciar en la Figura 8.

E. Algoritmos basados en Arboles

En este apartado entrenaremos 3 Modelos basados en Arboles, los modelos a entrenar serán XGBoost, CatBoost y LightGBM. Durante este entrenamiento no es necesario aplicar la transformación WOE ya que internamente los algoritmos basados en arboles establecen los cortes de las variables, sin embargo, si se utilizaran los filtros realizados en el procesamiento de datos. Según Niu (2022) los algoritmos de boosting son más versátiles con valores atípicos.

Para el entrenamiento de XGBoost se especificaron los siguientes hiper parámetros, max_dehp = 3, num_leaves = 7, objective = 'binary:logistic', Learning_raye = 0.08, drop_rate = 0.3, bagging_fraction = 0.4, feature_fraction = 0.35, etc. Se entreno el modelo con 120 árboles, dando como resultado una curva AUC para el entrenamiento de 0.78 (Gini = 56.05) y en OOT de 0.77 (Gini = 54.07).

El algoritmo de CatBoost fue entrenado con los siguientes parámetros, numero de iteraciones de 150, deph = 4, Learning_rate = 0.09, factor de regularización de 15, un sampling de 0.12, numero de iteraciones para hojas de 5, un número máximo de hojas de 16 y como función de perdida a LogLoss. Obteniendo un resultado de AUC en entrenamiento de 0.7808 (GINI de 56.16) y para la muestra fuera de tiempo un valor de 0.7725 (Gini de 54.5).

Por último, para el Modelo de LightGBM se declararon los siguientes hiper parámetro, Objective = Binary, Metric = auc, Learning_rate = 0.05, max_depth = 3, num_leaves = 6, bagging_fracction = 0.5, feature_fracction = 0.5, lamda_l1 = 20, lamda l2 = 7 y un numero de arboles = 120. Dando como resultado una curva AUC en train de 0.7815 (Gini de 56.31) y para OOT un AUC de 0.7727 (Gini 54.55). Tal y como se muestra en la Figura 9, todos los algoritmos tienen un performance similar, sin embargo, el algoritmo LightGBM es el segundo que presenta menor diferencia entre el train y oot a nivel de AUC (1.65 variación de GINI), y además es el que presenta también un mejor valor de GINI, por lo tanto, se puede decir que es el mejor algoritmo para este conjunto de datos.

Figura 8.
Curva ROC AUC para Regresión Logística con WOE

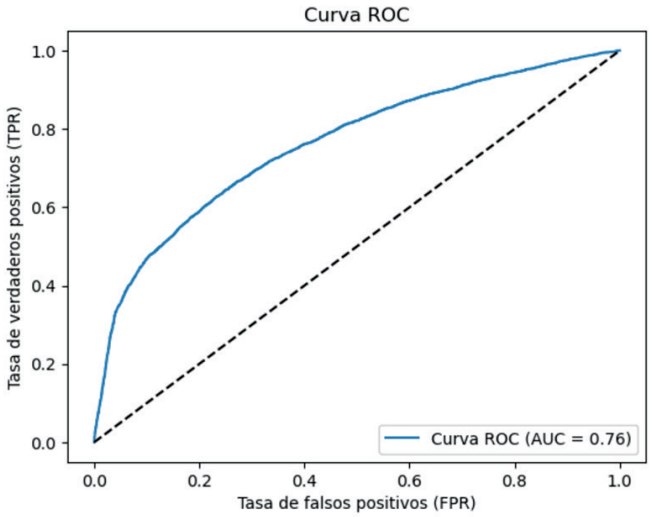


Figura 9.

Curva AUC de los Modelos basado en GB

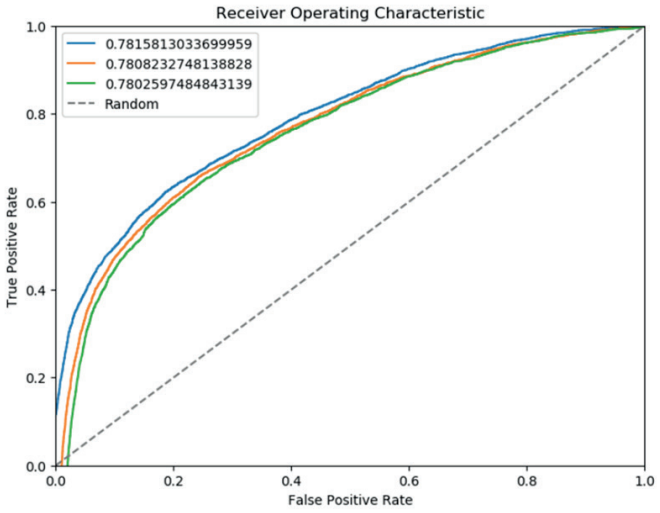


Table 5.

Hiper Parámetros obtenidos mediante búsqueda Bayesiana

Hiper Parametro	Rango Búsqueda	Valor Obtenido
learning_rate	0.07 – 0.12	0.1198
max_deph	1 – 3	3
Num_leaves	2 – 8	7
Subsample	0.1 – 0.2	0.11
colsample_bytree	0.2 – 0.3	0.23
min_child_samples	10 – 20	13
Lamdal1	1 – 10	5.41
Lamdal1	1-10	7.7
Scale_pos_weight	5	5

Las pruebas anteriores nos permitieron escoger el algoritmo más prometedor para nuestro conjunto de datos, sin embargo, podemos aun mejorar su rendimiento a través de un tuneo de Hiper parámetros, para esto utilizaremos el método basado en la búsqueda Bayesiana utilizando la librería Optuna de Python, los hiper parámetros obtenidos de detallan en la Tabla 4.

Este nuevo modelo entrenado con los parámetros especificados por la búsqueda Bayesiana nos permite incrementar la curva AUC en entrenamiento a un valor de 0.7947 (Gini = 58.94) y para OOT a un valor de 0.7867 (Gini = 57.4), de esta manera se logra una diferencia entre los 2 conjuntos de 1.54, lo cual es mucho más estable que el modelo anterior.

F. Importancia de Variables GAIN

Para ver cuanto aporta cada variable al Modelo se utilizará el enfoque de Ganancia (GAIN), que mide como cada variable aporta a la reducción de impureza (Gini Inpurity o Entropia), cuando mayor sea la reducción de impureza, se dirá que la variable aporta más a la predicción. En la Figura 10 se detalla la importancia de las variables para el Modelo de LGB Tuneado, en donde el estado de pago del mes anterior representa la más importante de todas.

G. Gráficos de Dependencia Parcial

Si bien es cierto conocemos que variables son las más importantes, no sabemos cómo interpretar el incremento o decremento de los valores , como estos se relacionan con un aumento o disminución de la probabilidad de default en el siguiente mes (target), esta es una de las grandes desventajas de implementar modelos de machine Learning a la hora de credit scoring, los modelos de regresión logística nos permiten ganar mucha interpretabilidad, cosa que no podemos aplicar directamente con un

modelo LGBM. Para superar este inconveniente vamos a utilizar el enfoque SHAP.

Según Ghorbani et al. (2019) el concepto de SHAP values está basado en la teoría de juegos y utiliza algunos conceptos para asignar de manera justa el aporte o contribución de cada característica en la predicción final . De esta forma capturamos la importancia de cada variable a la hora de dar la probabilidad, esto también nos ayuda a poder darle una interpretabilidad a las variables y superar así el inconveniente y desventaja que se tenía frente a la regresión logística. Por ejemplo, en la Figura 11 podemos ver el grafico de dependencias parciales basado en SHAP para la variable más importante.

Figura 10.
Importancia de Variables del Modelo LGM

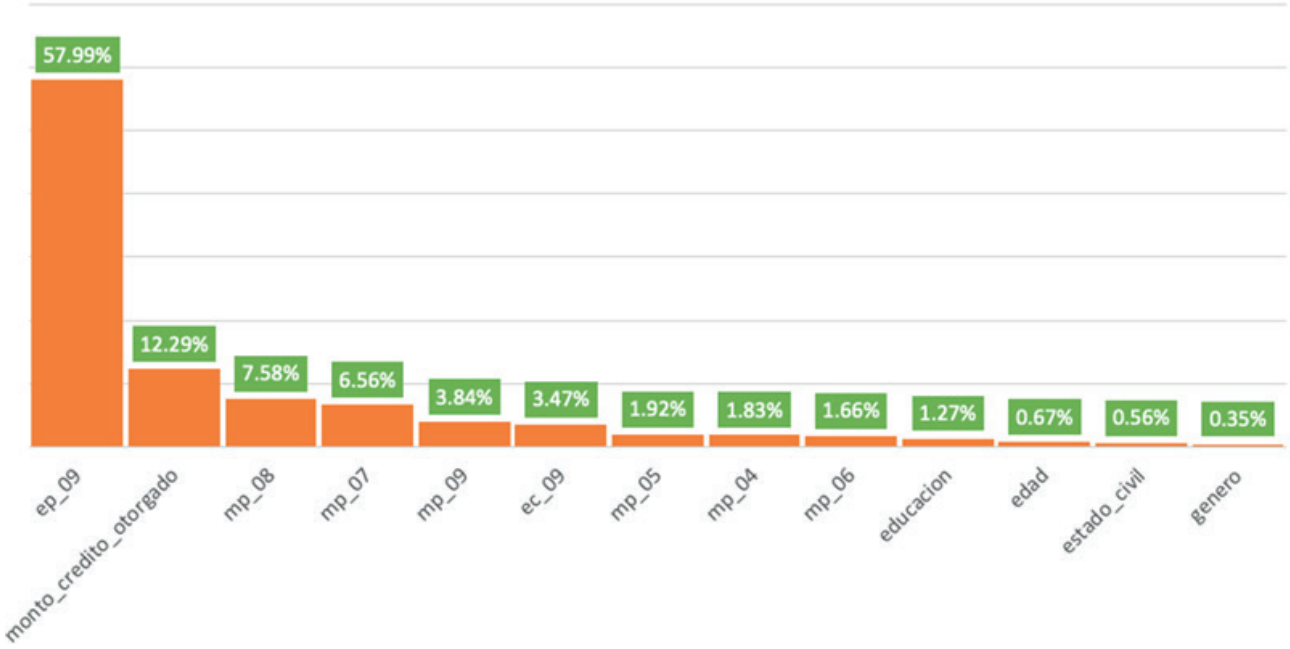
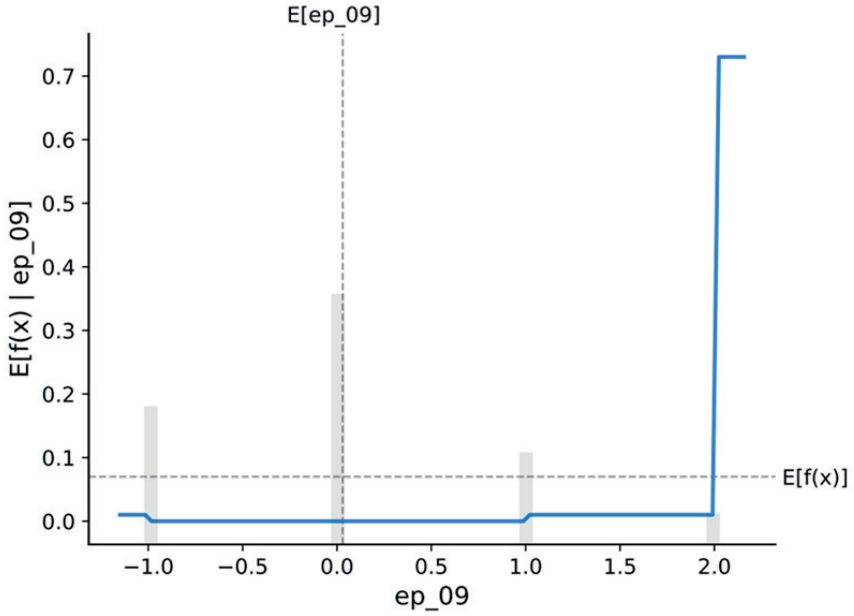


Figura 11.
Dependencia Parcial ep_09



Este grafico nos permite interpretar que un aumento de valor en la variable ep_09 logra aumentar la probabilidad, sobre todo existe un gran salto en aquellos con valores SHAP de 2. De igual manera en la figura 12 se puede apreciar el comportamiento de las diversas variables, por ejemplo, a mayor valor de crédito otorgado menor probabilidad de caer en default, a mayor cantidad de monto pagado menor probabilidad de default.

De la misma manera el Grafico de dependencia parcial para la variable ec_09 nos permite interpretar que a mayor estado de cuenta (asociado a una mayor cantidad de dinero) reduce la probabilidad de incumplir en el pago (default), tal y como lo demuestra la Figura 13

Figura 12.
Plot Sumary de SHAP

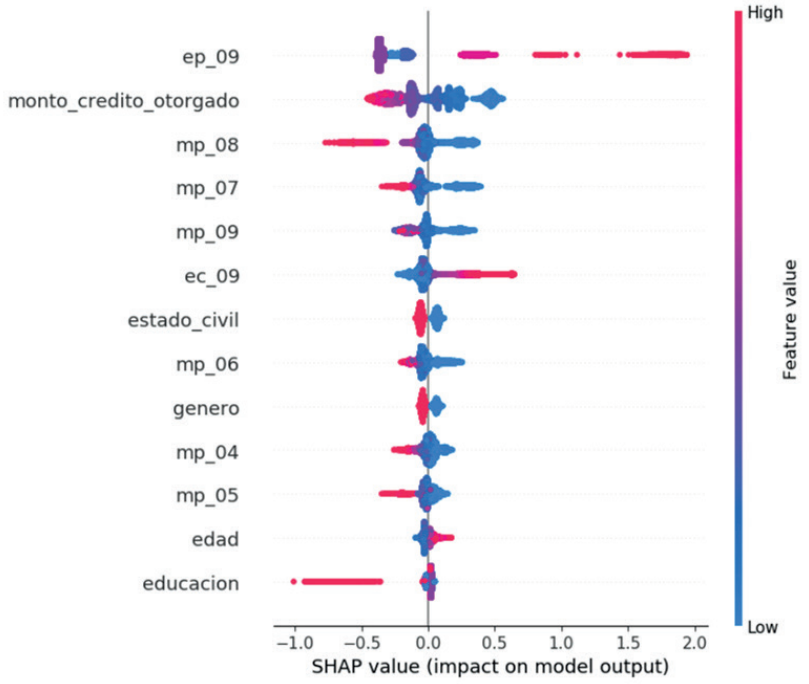
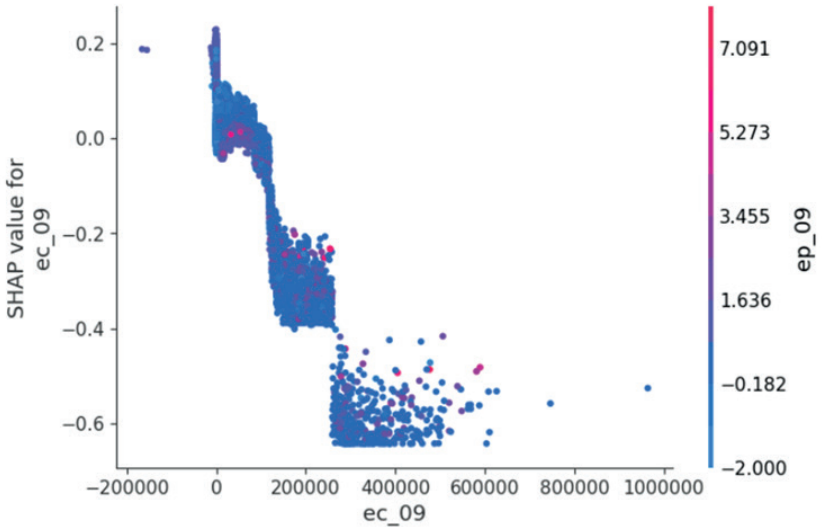


Figura 13.
Dependencia Parcial Estado de Cuenta



V. RESULTADOS

Los resultados obtenidos se reflejan en la Tabla 5, como se puede apreciar el Modelo de mejor rendimiento fue el de LightGBM, ganando tanto en Gini como en estabilidad, al presentar una menor variación entre los periodos de entrenamiento y fuera de tiempo.

Además, se puede observar que el enfoque tradicional de RL que es comúnmente aplicado tiene un valor de 51.01 de Gini. Indicadores de Gini por encima de 50% representan modelos buenos o satisfactorios para la industria bancaria. Con fin de poder tener una mejor interpretación de la predicción se procederá a crear un indicador de Score que varie de 0-1000, este puntaje mientras más se acerque a 1000 reflejará un mejor comportamiento del cliente y por lo tanto un menor riesgo crediticio. La creación de este score sigue la sigue la formula (3).

$$Score\ Crediticio =$$

$$[[((Probabilidad\ Default - 1) * -1000)]]+1 \quad (3)$$

Donde: [[x]] Función Máximo Entero

La tabla de resultados obtenida ha revelado claramente que el modelo de scoring basado en el modelo LGBM ha logrado superar a sus competidores en términos de cortes de score más efectivos. Los cortes iniciales aplicados han mostrado consistentemente una mayor proporción de clientes de calidad, es decir, aquellos con un bajo riesgo crediticio. Por otro lado, los cortes más elevados han demostrado ser altamente precisos al identificar a los clientes con mayor probabilidad de morosidad. Estos hallazgos respaldan de manera concluyente la efectividad y robustez del modelo de scoring implementado, destacando el valor de utilizar el LGBM para la evaluación y selección de clientes en procesos de otorgamiento de créditos. Según Yap et al. (2011) el score crediticio es primordial para la solicitud de prestamos. Como resultado final se puede observar que el modelo basado en Gradiente Potenciadora de Luz logra obtener un resultado muy por encima del enfoque tradicional, de esta manera se puede ganar precisión a la hora de detectar clientes morosos en el sector bancario, tal y como lo refleja la Tabla 7 y la Tabla 8.

Por último, comparando el enfoque tradicional RL con Woe vs LGBM mediante la curva de morosidad, tal y como lo demuestra la Figura 14, es apreciable la calidad del modelo tanto en buckets inferiores (mejores clientes) y bucket superiores (clientes con mayor riesgo).

Table 6.
Métricas de AUC y GINI para los modelos testeados

Modelo	# Variables	Train AUC	OOT AUC	Train Gini	OOT Gini
Rl Woe	8	0.76	0.75	52.35	51.03
Xgboost	12	0.78	0.77	56.05	54.07
Catboost	12	0.7808	0.7725	56.16	54.5
Lightgbm	12	0.7815	0.7727	56.31	54.55
Lightgbm Tunning	12	0.7947	0.7867	58.94	57.4

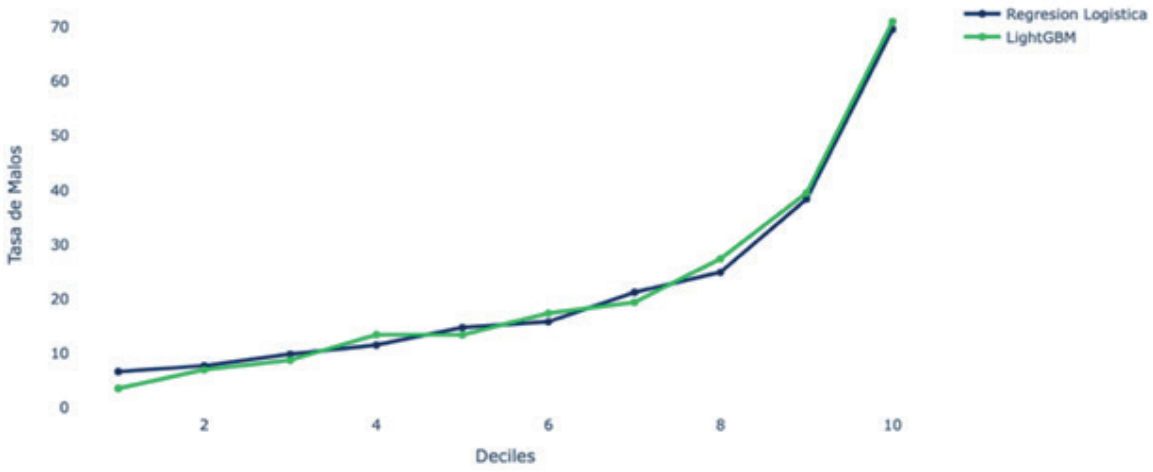
Table 7.
Cortes del Score para RL

Corte de Puntaje	Cantidad Registros	Cantidad de Malos	Tasa de Malos
(921, 961]	2902	193	6.65%
(902, 921]	3055	237	7.76%
(884, 902]	2937	290	9.87%
(865, 884]	3012	348	11.55%
(849, 865]	3028	447	14.76%
(830, 849]	2938	465	15.83%
(798, 830]	3071	652	21.23%
(688, 798]	3042	758	24.92%
(469, 688]	3012	1156	38.38%
(166, 469]	3003	2090	69.60%

Table 8.
Cortes del Score para LGBM

Corte de Puntaje	Cantidad Registros	Cantidad de Malos	Tasa de Malos
(772, 876]	2940	105	3.57%
(729, 772]	3007	211	7.02%
(696, 729]	3006	264	8.78%
(661, 696]	2991	402	13.44%
(616, 661]	3026	406	13.42%
(563, 616]	3014	524	17.39%
(498, 563]	2996	581	19.39%
(413, 498]	3013	826	27.41%
(215, 413]	3007	1187	39.47%
(76, 215]	3000	2130	71.00%

Figura 14.
Curva de Morosidad RL VS LGBM



VI. CONCLUSIONES

Este artículo presenta un enfoque propuesto para la predicción de morosidad también conocido como default o tasa de incumplimiento en los clientes de tarjeta de crédito. El enfoque se basa en una metodología basada en potenciador de gradiente de luz para realizar la predicción, además se usó un optimizador de hiper parámetros de búsqueda bayesiana para mejorar el rendimiento del modelo. Finalmente se presentó una solución a la falta de interpretabilidad de este tipo de modelos debido a su complejidad matemática y estadística, para esto se usó el enfoque Shapley y la importancia Gain. Los resultados obtenidos muestran que los algoritmos de ML acompañados de técnicas de interpretación de variables logran superar significativamente a los algoritmos estadísticos tradicionales, además de suplir el mayor inconveniente que tienen al aplicarlos, que es la falta de interpretación.

REFERENCIAS

[1] Abdou, H. A. (2009). Genetic programming for credit scoring: The case of Egyptian public sector banks. Expert systems with applications, 36(9), 11402-11417.

[2] Clark, A. (2018). La auditoría de aprendizaje automático-CRISP-DM Framework. Revista Isaca, 1, 42-47.

[3] Disha, R. A., & Waheed, S. (2022). Performance analysis of machine learning models for intrusion detection system using

Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique. Cybersecurity, 5(1), 1.

[4] Fernández Vásquez, R. F. (2018). Regresión bayesiana con enlaces asimétricos para la clasificación de clientes con propensión a caer en mora en una entidad bancaria.

[5] Garcia Nazario, A. M. (2020). Sistema de información basado en redes neuronales para la predicción de riesgo en el otorgamiento de créditos personales en una cooperativa de ahorro y crédito en el departamento de Lambayeque.

[6] Ghorbani, A., & Zou, J. (2019, mayo). Data shapley: Equitable valuation of data for machine learning. En International Conference on Machine Learning (pp. 2242-2251). PMLR.

[7] La República. (2021, 19 de marzo). Aumentan deudas morosas por créditos personales y tarjetas. Recuperado de https://larepublica.pe/economia/2021/03/19/aumentan-deudas-morosas-por-creditos-personales-y-tarjetas-lrsd/

[8] La República. (2022, 18 de agosto). Hay 8.5 millones de peruanos morosos: cifra aún por encima de prepandemia de deudas Infocorp créditos Equifax. Recuperado de https://larepublica.pe/economia/2022/08/18/hay-85-millones-de-peruanos-morosos-cifra-aun-por-encima-de-prepandemia-deudas-infocorp-creditos-equifax/

[9] MasFinanzas. (s.f.). Más del 40% de peruanos ya usa las tarjetas de crédito como su medio de pago principal. Recuperado de https://masfinanzas.com.pe/pagos-digitales/mas-del-40-de-peruanos-ya-usa-las-tarjetas-de-credito-como-su-medio-de-pago-principal/

[10] Miranda Pilco, A. (2021). Predicción del riesgo de incumplimiento en el pago de los créditos del portafolio de una entidad financiera utilizando regresión logística.

[11] Ojo Público. (s.f.). Bancos suben intereses en créditos a niveles que superan prepandemia. Recuperado de https://ojo-publico.com/sala-del-poder/bancos-suben-intereses-creditos-niveles-que-superan-prepandemia

[12] Red de Expertos. (2022, 20 de julio). Una revolución digital para la inclusión financiera. El País. Recuperado de https://elpais.com/planeta-futuro/red-de-expertos/2022-07-20/una-revolucion-digital-para-la-inclusion-financiera.html

[13] Si, Z., Niu, H., & Wang, W. (2022). Credit Risk Assessment by a Comparison Application of Two Boosting Algorithms. En Fuzzy Systems and Data Mining VIII (pp. 34-40). IOS Press.

[14] Slabber, E., Verster, T., & de Jongh, R. (2023). Some Insights about the Applicability of Logistic Factorisation Machines in Banking. Risks, 11(3), 48.

[15] Statista. (s.f.). Uso de tarjeta de crédito para pagos a plazo en Latinoamérica. Recuperado de https://es.statista.com/grafico/18346/uso-de-tarjeta-de-credito-para-pagos-a-plazo-en-latinoamerica/

[16] Torres Chero, E. R., & Farroñay Julca, J. M. (2017). Implementacion de Mineria de Datos para detectar Patrones de Comportamiento de Clientes Morosos en Empresa de Credito Crediserv EIRL–Chiclayo.

[17] Yap, B. W., Ong, S. H., & Husain, N. H. M. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. Expert Systems with Applications, 38(10), 13274-13283.

Financiamiento:

Propio

Conflictos de interés:

Los autores declaran no tener conflictos de interés.

Contribuciones de autoría:

Eduardo Rafael Jauregui Romero (El articulo completo)