Sistema de Predicción de Aprobación de Tarjetas de Crédito usando Aprendizaje Automático

Valentina C. Zapata,[1], Luis G. Sánchez[2], y Jose D. Gómez Muñetón.[3]

[1]valentina.candenaz@udea.edu.co [2]guillermo.sanchez1@udea.edu.co
[3]jose.gomez14@udea.edu.co

***Index Terms*—Modelos predictivos, Árboles de decisión, Regresión logística, Riesgo crediticio, Aprendizaje supervisado.**

I. Introduction

EN el presente proyecto se propone desarrollar un modelo de aprendizaje automático (*Machine Learning*) capaz de predecir si un solicitante será un "buen" o "mal" cliente. Este enfoque busca automatizar y optimizar el proceso de aprobación de tarjetas de crédito, permitiendo al banco reducir riesgos y mejorar la eficiencia del proceso de evaluación. La solución basada en *Machine Learning* permitirá no solo realizar predicciones, sino también adaptarse mejor a nuevas condiciones económicas o cambios en los patrones de comportamiento de los usuarios.

II. Descripción del Problema

La aprobación de tarjetas de crédito es un proceso crítico para las entidades financieras, ya que implica una evaluación cuidadosa del riesgo crediticio asociado a cada solicitante. Con el objetivo de mitigar el riesgo de impago, los bancos utilizan modelos predictivos basados en información personal, laboral y financiera de los solicitantes para decidir si aprueban o no una solicitud. Tradicionalmente, esta evaluación se realiza a través de puntuaciones crediticias construidas a partir de modelos estadísticos como la regresión logística.

El desarrollo de una solución basada en *Machine Learning* ofrece ventajas significativas sobre los métodos tradicionales, ya que estos modelos pueden captar patrones complejos y no lineales en los datos que los enfoques estadísticos convencionales podrían pasar por alto. Además, los algoritmos de *Machine Learning* tienen la capacidad de aprender y mejorar con el tiempo a medida que se dispone de nuevos datos, lo que permite una adaptación dinámica a cambios en el comportamiento de los solicitantes o en las características del perfil de los clientes objetivo.

El conjunto de datos utilizado proviene del portal Kaggle [1] y se compone de dos archivos principales: application_record.csv y credit_record.csv. El primero contiene información detallada sobre los solicitantes, incluyendo datos demográficos, laborales y financieros, mientras que el segundo recoge el comportamiento crediticio a lo largo del tiempo.

La información demográfica proporciona datos clave sobre el perfil del solicitante y permite comprender aspectos sociales y personales que pueden influir en el comportamiento crediticio a lo largo del tiempo. Por otra parte, se incluyen también variables de carácter económico y laboral, como el nivel de ingresos, el tipo de empleo, el nivel educativo y la ocupación, que ayudan a estimar la estabilidad financiera del solicitante y su capacidad para generar ingresos de manera sostenida.

Asimismo, se consideran datos sobre propiedades y activos, tales como la posesión de vivienda, automóvil y número de miembros dependientes, los cuales permiten evaluar el patrimonio y las responsabilidades económicas del solicitante. Finalmente, el historial crediticio, registrado en credit_record.csv, proporciona una visión temporal del comportamiento de pago, incluyendo retrasos y cumplimiento de pagos.

En conjunto, estas variables reflejan la solvencia económica del solicitante, lo cual es fundamental para evaluar su capacidad de pago y, por ende, el riesgo crediticio asociado a la aprobación de una tarjeta de crédito.

En total, se cuenta con 438,557 registros de información de clientes, descritos mediante 17 variables, y 1,048,575 registros en el historial de créditos. Las variables son de tipo categórico, numérico y binario. Entre las variables categóricas se incluyen el género, tipo de ingreso, nivel educativo, estado civil, tipo de vivienda y ocupación; las variables numéricas abarcan el número de hijos, ingresos totales, edad y años trabajados; mientras que las variables binarias indican características como si el solicitante posee coche o vivienda.

Durante el análisis exploratorio se identificaron valores faltantes en la variable *OCCUPATION_TYPE* (ocupación). Por ello, se propone imputar dicha variable con la categoría "Desconocido". Adicionalmente, se sugiere contrastar estos datos con la variable de días de desempleo, con el fin de imputar los valores faltantes como "Desempleado" en aquellos casos donde se registre un periodo de desempleo.

En cuanto a la codificación de variables, se aplicará codificación ordinal para aquellas que presenten una jerarquía inherente, como el nivel educativo, y codificación one-hot para el resto de las variables categóricas. Asimismo, se llevará a cabo escalamiento o normalización en las variables numéricas que lo requieran, como *AMT_INCOME_TOTAL* (ingresos anuales), con el objetivo de mejorar el rendimiento de los modelos.

Un reto importante de este dataset es la ausencia de una etiqueta explícita que indique si la solicitud fue aprobada o si el cliente fue bueno o malo. Por lo tanto, se debe construir una variable objetivo a partir de la información contenida en credit_record.csv. Una alternativa es etiquetar como "buen cliente" a aquel que no presenta atrasos significativos (por ejemplo, no más de dos meses de mora en los pagos), y como

"mal cliente" a aquellos con historial negativo recurrente.

El problema a resolver es de clasificación binaria, y se evaluarán diferentes modelos del aprendizaje automático. En primer lugar, se considerará la regresión logística debido a su facilidad de interpretación y su uso frecuente en el ámbito financiero. No obstante, este modelo puede no ser suficiente si las relaciones entre variables no son lineales. En ese caso, se recurrirá a modelos de árboles de decisión, como el *Random Forest*, que son más robustos frente a datos mixtos y no requieren un extenso preprocesamiento. Finalmente, se tendrá en cuenta el uso de modelos de gradiente como *XGBoost*, ampliamente reconocidos por su precisión en competencias y aplicaciones reales, aunque más complejos en su ajuste y explicación. La comparación entre estos modelos permitirá seleccionar aquel que logre el mejor balance entre precisión, interpretabilidad y eficiencia.

Es importante tener en cuenta el desbalance de clases que se puede presentar al construir la variable objetivo. Este fenómeno puede llevar a que los modelos predicen más frecuentemente la clase mayoritaria, afectando su desempeño real. Para mitigarlo, se planea aplicar técnicas como la reponderación de clases, el sobremuestreo de la clase minoritaria mediante métodos como SMOTE, o la utilización de métricas como el *F1-score* o el AUC-ROC, más adecuadas en escenarios desbalanceados.

III. Estado del arte

Para abordar el problema de predicción de aprobación de tarjetas de crédito, se analizó el artículo titulado *"Predicting Credit Card Approval Using Machine Learning Techniques"* de Ehsan Lotfi (2024) [1], publicado en la revista *International Journal of Applied Data Science in Engineering and Health*. En este artículo se aborda el problema de aprobación de tarjetas de crédito empleando el mismo dataset usado para este estudio.

En cuanto al paradigma de aprendizaje, se utiliza el enfoque de aprendizaje supervisado, ya que el modelo se entrena con ejemplos etiquetados que indican si la solicitud de tarjeta fue aprobada o no. Para resolver el problema, los autores comparan cuatro algoritmos de aprendizaje automático: Regresión Logística, Árbol de Decisión, *Random Forest* y *XGBoost* (*Extreme Gradient Boosting*). Cada uno de estos métodos tiene distintas características en cuanto a complejidad, capacidad de generalización e interpretación.

La metodología de validación consiste en una partición del dataset en 80% para entrenamiento y 20% para prueba. Además, se utilizaron técnicas de validación cruzada y Random Search para el ajuste de hiperparámetros del modelo *XGBoost*. Esto permitió optimizar variables como tasa de aprendizaje, profundidad máxima del árbol, número de estimadores y penalizaciones L1 y L2.

Respecto a las métricas de evaluación, se emplearon varias para medir el rendimiento del sistema: *accuracy* (exactitud), *precision* (precisión), *recall* (sensibilidad), *F1-score* (media armónica entre precisión y *recall*), y la matriz de confusión. Estas métricas permiten analizar no solo cuántas predicciones fueron correctas, sino también la calidad de las predicciones positivas, lo cual es clave en contextos donde hay desbalance de clases.

Los resultados mostraron que el modelo *XGBoost* fue el más efectivo, alcanzando un 99.04% de exactitud, 85% de *recall* y 78% de precisión sobre el conjunto de prueba. Estos valores evidencian su capacidad para identificar correctamente a los solicitantes aprobados y minimizar errores tipo I y II. El artículo concluye que los métodos de ensamble avanzados como *XGBoost* y *Random Forest* superan claramente a modelos más simples como la regresión logística, particularmente cuando se trata de capturar relaciones no lineales y patrones complejos en grandes volúmenes de datos.

Después de esto se realizó el análisis de un segundo artículo de Chang et al. (2024), titulado *"Credit Risk Prediction Using Machine Learning and Deep Learning: A Study on Credit Card Customers"* [2], el cual aborda el problema utilizando el mismo dataset. El trabajo utiliza aprendizaje supervisado, ya que se entrena un modelo para predecir si un cliente será "bueno" o "malo" (es decir, si tendrá o no problemas de pago) a partir de datos históricos etiquetados. En este caso se probaron seis algoritmos de clasificación: regresión logística, *Random Forest*, redes neuronales (ANN), *AdaBoost*, *LightGBM* y *XGBoost*.

El conjunto de datos se dividió en un 70% para entrenamiento y 30% para prueba. Se utilizó la técnica SMOTE (*Synthetic Minority Oversampling Technique*) para abordar el fuerte desbalance de clases (solo el 1.3% de los clientes eran "malos"). Aunque no se aplicó validación cruzada (K-fold), se recomienda como trabajo futuro.

Las métricas utilizadas fueron *accuracy*, *precision*, *recall*, *F1-score*, ROC-AUC y MCC (*Matthews Correlation Coefficient*). Esta última es útil para problemas con clases desbalanceadas, ya que considera todos los valores de la matriz de confusión. Un MCC de 1 indica una clasificación perfecta.

Como resultado se obtuvo que el modelo *XGBoost* fue el más preciso, con un *accuracy* del 99.3%, *precision* y *recall* de 0.993, *F1-score* de 0.993, AUC de 0.997 y MCC de 0.986. *LightGBM* obtuvo resultados similares. Las redes neuronales, en cambio, tuvieron un rendimiento inferior debido al desbalance de clases.

IV. Entrenamiento y Evaluación de los Modelos

...

A. Configuración Experimental

...

B. Resultados del Entrenamiento de Modelos

...

V. Reducción de dimensión

..

A. Selección de Características

...

B. Extracción de Características

...

REFERENCES

[1] E. Lotfi. “Predicting Credit Card Approval Using Machine Learning Techniques,” *International Journal of Advanced Development in Engineering and Sustainable Technology (IJADSEH)*, vol. 1, no. 1, pp. 18-30, 2024. Disponible: https://ijadseh.com/index.php/ijadseh/article/view/22

[2] V. Chang, S. Sivakulasingam, H. Wang, S.T. Wong, M.A. Ganatra, and J. Luo. “Credit Risk Prediction Using Machine Learning and Deep Learning: A Study on Credit Card Customers,” *Risks*, vol. 12, no. 11, Art. no. 174, Nov. 2024. Disponible: https://doi.org/10.3390/risks12110174