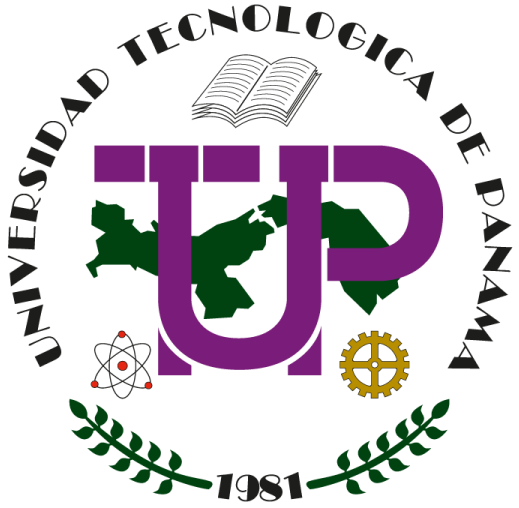

Universidad Tecnológica de Panamá

Maestría en Analítica de Datos

Materia: Modelos Predictivos

Profesor: Juan Marcos Castillo, PhD

Proyecto Final

Incumplimiento de pago de tarjeta de crédito basado en modelo predictivo

Nombre del estudiante: Xavier Alejandro Barco Macías

Cédula: E-8212984

Fecha: abril 2025

1. Introducción

En el entorno financiero actual, la gestión eficaz del riesgo crediticio es esencial para la estabilidad y rentabilidad de las instituciones bancarias. Uno de los principales desafíos que enfrentan estas entidades es el incumplimiento de pago por parte de los tarjetahabientes, lo cual puede derivar en pérdidas significativas. Ante esta problemática, el uso de herramientas de análisis de datos y modelos predictivos se presenta como una solución efectiva para anticipar comportamientos de riesgo y optimizar la toma de decisiones.

El presente proyecto tiene como objetivo desarrollar un modelo predictivo capaz de identificar, con base en datos históricos, la probabilidad de que un cliente incurra en incumplimiento de pago de su tarjeta de crédito en el próximo mes. Para ello, se utilizará la base de datos “Default of Credit Card Clients”, la cual contiene información detallada sobre el comportamiento financiero y características personales de 30,000 clientes. A través del uso de técnicas de machine learning, este estudio busca contribuir a una gestión crediticia más informada y proactiva.

2. Justificación

El motivo principal de esta investigación radica en la necesidad de fortalecer las estrategias de gestión del riesgo crediticio dentro del sector bancario, en el cual actualmente me desempeño profesionalmente. Anticipar el incumplimiento de pagos permite a las instituciones financieras no solo prevenir pérdidas económicas, sino también diseñar políticas crediticias más responsables y adaptadas al perfil del cliente.

La elección de la base de datos "Default of Credit Card Clients" se justifica tanto por su relevancia práctica como por su amplia adopción en la comunidad científica. Su estructura, enfocada en la predicción de eventos de incumplimiento a corto plazo, resulta adecuada para la implementación de modelos de clasificación binaria. Además, al tratarse de una base de datos bien documentada y utilizada en diversos estudios, permite establecer comparaciones con resultados previos y validar la efectividad del modelo propuesto.

3. Antecedentes

Diversos estudios han abordado el problema del incumplimiento de pago utilizando modelos estadísticos y de machine learning. Entre las técnicas más utilizadas se encuentran la regresión logística, los árboles de decisión, los

bosques aleatorios (Random Forest), y las redes neuronales artificiales. Estos métodos han demostrado ser eficaces en la clasificación de clientes según su nivel de riesgo crediticio.

En el caso específico de la base de datos "Default of Credit Card Clients", investigaciones previas han evaluado el desempeño de diferentes algoritmos, concluyendo que modelos como Support Vector Machines (SVM) y redes neuronales profundas ofrecen buenos resultados, aunque con mayor complejidad computacional.

Este proyecto se nutre de dichos antecedentes y busca aplicar, comparar y evaluar algunos de estos modelos dentro de un entorno práctico y realista, con miras a su futura implementación en instituciones financieras.

4. Definición del problema

El incumplimiento de pago de tarjetas de crédito representa un riesgo financiero significativo para las instituciones bancarias, impactando negativamente sus ingresos, liquidez y estabilidad. Este problema se intensifica cuando no se cuenta con herramientas adecuadas para prever cuáles clientes podrían incurrir en impago.

Actualmente, muchos procesos de evaluación crediticia se basan en modelos tradicionales que no consideran de manera dinámica el comportamiento reciente del cliente ni el análisis de grandes volúmenes de datos.

La presente investigación se plantea como solución a esta limitación, formulando el siguiente problema:

¿Es posible predecir el incumplimiento de pago de un cliente de tarjeta de crédito para el siguiente mes, utilizando técnicas de machine learning aplicadas a datos históricos financieros y demográficos?

Responder a esta pregunta permitirá desarrollar un modelo predictivo que ayude a mitigar el riesgo crediticio mediante una asignación más eficiente del crédito y la implementación de estrategias preventivas para clientes en situación de alto riesgo.

5. Análisis Predictivo

a. Determinación de la base de datos

Para el desarrollo del modelo predictivo se utilizó la base de datos 'Default of Credit Card Clients', disponible en el repositorio UCI Machine Learning. Este conjunto de datos contiene información de 30,000 titulares de tarjetas de crédito en Taiwán, recopilada por una institución financiera durante un período de seis meses. Incluye variables demográficas (edad, sexo, nivel educativo, estado civil), información financiera (límite de crédito, historial de pago, monto de pago) y una variable objetivo binaria que indica si el cliente incurrió en incumplimiento de pago en el mes siguiente.

b. Pre-procesamiento y limpieza

El preprocesamiento de los datos incluyó varias etapas esenciales. Primero, se eliminaron registros con valores nulos y se descartó la columna 'ID', ya que no aporta valor predictivo. Posteriormente, se identificaron y eliminaron valores atípicos en la variable 'LIMIT_BAL' utilizando el rango intercuartílico (IQR) como criterio. Asimismo, se consolidaron y reclasificaron valores en las variables categóricas 'EDUCATION' y 'MARRIAGE' para reducir la dispersión y facilitar su codificación. Finalmente, las variables nominales fueron transformadas mediante codificación One-Hot (OneHotEncoder), eliminando la primera categoría para evitar multicolinealidad.

c. Análisis descriptivo

Se realizaron análisis estadísticos y gráficos para comprender la distribución de las variables. Entre los hallazgos más relevantes, se observó que la mayoría de los clientes se encontraban en el rango de edad de 30 a 40 años, con una proporción ligeramente mayor de mujeres. El monto promedio de crédito (LIMIT_BAL) fue de aproximadamente 167,000 unidades monetarias, aunque con una dispersión considerable. La tasa general de incumplimiento fue cercana al 22%, lo que indica un desbalance de clases moderado que debe considerarse en la etapa de modelado.

d. Selección de variables

Se incluyeron como variables predictoras todas aquellas relacionadas con características demográficas, límites de crédito, historial de pagos y pagos recientes. La variable objetivo fue 'default payment next month', que indica si el cliente incumplió su pago en el mes siguiente. Tras realizar codificación y limpieza, se seleccionaron 23 variables independientes que fueron normalizadas o codificadas según su tipo.

e. Selección de modelos

Se entrenaron y evaluaron varios modelos de clasificación supervisada: Regresión Logística, Árbol de Decisión, Random Forest, K-Nearest Neighbors, Support Vector Machine y XGBoost. Cada modelo fue evaluado con base en las métricas de accuracy, precision, recall y F1-score. Entre ellos, el modelo XGBoost mostró el mejor desempeño general, alcanzando una precisión del 62%, un recall del 37% y un F1-score de 0.46. Adicionalmente, se aplicó ajuste de hiperparámetros mediante Grid Search para optimizar XGBoost, logrando una mejora adicional en su rendimiento. También se generó la curva ROC-AUC y se evaluó la importancia de variables, destacándose los pagos recientes y el historial de crédito como factores clave.

6. Conclusiones

El presente estudio permitió desarrollar un modelo predictivo efectivo para anticipar el incumplimiento de pago de tarjetas de crédito utilizando técnicas de machine learning. Entre los modelos evaluados, XGBoost demostró ser el más equilibrado en términos de precisión y capacidad de detección, logrando un F1-score superior a los demás algoritmos probados. El análisis de importancia de variables reveló que el comportamiento reciente de pago y el historial crediticio son los factores más determinantes en la predicción del riesgo de incumplimiento.

Además, el enfoque sistemático en el preprocesamiento, codificación y selección de variables permitió optimizar la calidad del modelo. La inclusión de técnicas de ajuste como el Grid Search contribuyó a refinar aún más el rendimiento predictivo. Estos resultados confirman que el uso de modelos avanzados puede aportar valor significativo en la gestión del riesgo crediticio.

7. Recomendaciones y futuros estudios

Como recomendaciones para futuras investigaciones, se sugiere considerar el uso de técnicas de balanceo de clases como SMOTE o el ajuste de pesos para mejorar la sensibilidad del modelo hacia la clase minoritaria. Asimismo, podría evaluarse la incorporación de variables adicionales relacionadas con comportamiento o datos externos que refuercen la capacidad de predicción.

También se recomienda explorar el uso de modelos de ensamblado, como stacking o boosting múltiple, así como el empleo de redes neuronales para comparar resultados en contextos de mayor complejidad. Por último, sería

beneficioso validar el modelo en datos reales o de otras regiones geográficas para analizar su robustez y capacidad de generalización.

8. Bibliografía

Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020

UCI Machine Learning Repository. (n.d.). Default of Credit Card Clients Dataset. Recuperado de https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

9. Anexos

1. Base de datos, link

2. Descripción de las columnas, link

3. Análisis descriptivo, link

4. Análisis predictivo, link

5. Script, link