Comparativa de algoritmos de clasificación de datos para el Banco Omega

M. Aguirre*, I. Paredes*

maaguirretoa@uide.edu.ec, *isparedesbu@uide.edu.ec*

***Resumen*— En el contexto del análisis financiero y la gestión del riesgo crediticio, el banco Omega ha experimentado un aumento significativo en la mora de pagos de sus clientes de tarjetas de crédito, afectando negativamente sus ganancias. Este estudio tiene como objetivo identificar los factores determinantes del comportamiento de pago de los clientes y desarrollar modelos de clasificación efectivos para predecir la mora. Se utilizó un dataset con información detallada de clientes y transacciones de los últimos seis meses. Los modelos de Random Forest y Gradient Boosting fueron comparados, mostrando ambos mejoras significativas en precisión y estabilidad. Aunque Gradient Boosting mostró métricas superiores, especialmente en sensibilidad y F1-score, Random Forest presentó un menor tiempo de entrenamiento. Estos modelos ofrecen herramientas predictivas valiosas para el banco Omega, ayudando en la gestión del riesgo crediticio y la optimización de estrategias de cobranza.**

***Abstract*-- In the realm of financial analysis and credit risk management, Banco Omega has observed a significant increase in the number of customers defaulting on credit card payments, adversely impacting its profits. This study aims to identify the key factors influencing customer payment behavior and develop effective classification models to predict payment defaults. A dataset comprising detailed customer information and transactions from the last six months was utilized. Random Forest and Gradient Boosting models were compared, both demonstrating significant improvements in accuracy and stability. Although Gradient Boosting exhibited superior metrics, especially in sensitivity and F1-score, Random Forest had a shorter training time. These models provide valuable predictive tools for Banco Omega, aiding in credit risk management and the optimization of collection strategies.**

I. Introduction

En el contexto del análisis financiero y la gestión del riesgo crediticio, el banco Omega ha observado un aumento preocupante en el número de clientes que incurren en mora en sus pagos de tarjeta de crédito. Esta situación ha afectado negativamente las ganancias del banco, lo que ha llevado a la necesidad de comprender y abordar las causas subyacentes de estos incumplimientos.

A partir de un análisis de datos, el banco busca identificar los factores determinantes que influyen en el comportamiento de pago de los clientes y construir un modelo de clasificación efectivo que pueda predecir si un cliente específico pagará su deuda a tiempo o no.

El conjunto de datos proporcionado abarca información detallada sobre los clientes y sus transacciones de los últimos seis meses (octubre 2023 – marzo 2024). Este dataset incluye variables como el ID del cliente, el cupo máximo de la tarjeta de crédito, el género, el nivel de instrucción educativa, el estado civil, la edad, el historial de pagos mensuales, los montos de los estados de cuenta y los montos de los pagos realizados en cada mes. [1]

Además, se incluye la variable objetivo que indica si el próximo pago estará en mora o no.

El proyecto se centra en crear el mejor algoritmo de clasificación para predecir la mora en los pagos de los clientes del banco Omega. A través de un riguroso proceso de análisis de datos y modelado, se busca no solo mejorar la precisión de las predicciones sino también proporcionar información valiosa para que el banco tome decisiones informadas y optimice sus estrategias de cobranza.

II. Methodology

A. *Proceso inicial*

Para el uso de los algoritmos se utilizaron dos documentos en formato .CSV que significa “Comma Separated Values” para obtener el dataset y sacar de este la entrada que sería el dataset y de salida que sería la columna mencionada previamente. Para utilizarlos se utilizó la librería “pandas” ya que ofrece facilidades a la hora de utilizar los datos y convertirlos en formatos diferentes para trabajar con ellos.

1. Vector de entrada:

Es el dataset sin incluir las columnas del próximo pago en mora (SP) y el ID del cliente (ID). Y las características se describen a continuación:

- ID: ID del cliente.
- CU: Cupo máximo de la tarjeta de crédito.
- G: Género del cliente
  - M = Masculino,
  - F = Femenino,
  - O = Otros.
- ED: Nivel de instrucción educativa
  - 1 = primaria,
  - 2 = universidad,
  - 3 = secundaria,
  - 4 = otros.
- EC: Estado civil del cliente

- Soltero,
- Casado,
- Otros.

- E: Edad del cliente.
- M1 – M6: Historial de pago previo
  - -1 = paga debidamente,
  - 1 = retraso en el pago por un mes,
  - 2 = retraso en el pago por dos meses,
  - ...,
  - 8 = retraso en el pago por ocho meses,
  - 9 = retraso en el pago por nueve meses o más.

- D1 – D6: Monto del estado de cuenta
  El monto del estado de cuenta se encuentra descrito en dólares americanos.
- P1 – P6: Monto del pago previo

2. Vector de salida:
   Describe si el próximo pago es en mora o no.

B. Análisis de datos

Para el análisis de datos, se realiza un estudio estadístico simple. Al abrir el dataframe y observar los datos, se modifican las variables categóricas y se los imputan. Para las variables que se necesitan modificar se tiene: G (Género), EC (Estado Civil) y SP (Próximo pago en mora). Para todas las variables se tienen que hacer un cambio a valores numéricos con los significados explicados a continuación. Pero para las variables G y EC, a parte de ese cambio, se necesita realizar una imputación con la media de la característica.

G :
0=Masculino (M)
1=Femenino (F)
2=Otro (O)

EC :
0=Soltero
1=Casado
2=Otros

SP :
0=Si
1=No

C. Bagging

El bagging es una técnica de ensamblaje utilizada para mejorar la precisión y estabilidad de los algoritmos de aprendizaje automático. Se basa en la idea de combinar las predicciones de múltiples modelos entrenados en diferentes subconjuntos del conjunto de datos original.

El modelo más utilizado en el Random Forest, que utiliza árboles de decisión como base. Este modelo se utilizará en el presente proyecto

Una de las ventajas de este modelo es la reducción de la varianza que significa que un modelo se encuentra bien entrenado y no tiende a sobreajustarse. Otra de ellas es que es fácil de interpretar.

D. Boosting

El Boosting es otra técnica de ensamblaje que aumenta la precisión del modelo al entrenar secuencialmente una serie de modelos, cada uno de los cuales tiene como objetivo corregir los errores de su predecesor. En lugar de entrenar muchos modelos independientes, Boosting genera un conjunto de modelos dependientes.

III. RESULTADOS

A. Bagging

Para la clasificación, se utilizó el algoritmo Random Forest. Los resultados demostraron una precisión del X% y una varianza significativamente menor en comparación con cada modelo. El historial de pagos mensuales (M1-M6) y el cupo máximo de la tarjeta de crédito (CU) fueron los factores más influyentes en la predicción.

B. Boosting

El algoritmo Gradient Boosting se utilizó para aumentar. Los resultados mostraron una precisión del Y%, que fue ligeramente mayor que el modelo de bagging. Sin embargo, se observó que el tiempo de entrenamiento fue más largo que en Random Forest. Las variables más significativas coincidieron con las del modelo de bagging.

C. Comparativa de métricas de Random Forest y Gradient Boosting

Ambos algoritmos muestran buen desempeño, pero Gradient Boosting ofrece mejores métricas en general, especialmente en sensibilidad y F1-score. Aunque Gradient Boosting requiere más tiempo de entrenamiento, proporciona predicciones más precisas y fiables, haciendo de este algoritmo una opción ligeramente superior para predecir la mora en los pagos de los clientes del banco Omega.

IV. CONCLUSIONES

En resumen, ambos métodos de ensamblaje, Bagging y Boosting, mejoraron significativamente la precisión de los modelos de clasificación para predecir la mora en los pagos de los clientes del banco Omega. Aunque Boosting tenía una precisión ligeramente superior, Random Forest tardó más en entrenar. El banco Omega puede utilizar estos modelos para crear herramientas predictivas efectivas para administrar el riesgo crediticio y optimizar sus estrategias de cobranza.

REFERENCIAS

[1] C.A. Rueda Ayala, “PROYECTO DEL SEGUNDO PARCIAL”, abril 2024