

JULIAN MESTRALET
CURSO DATA SCIENCE - CODERHOUSE

Proyecto Final

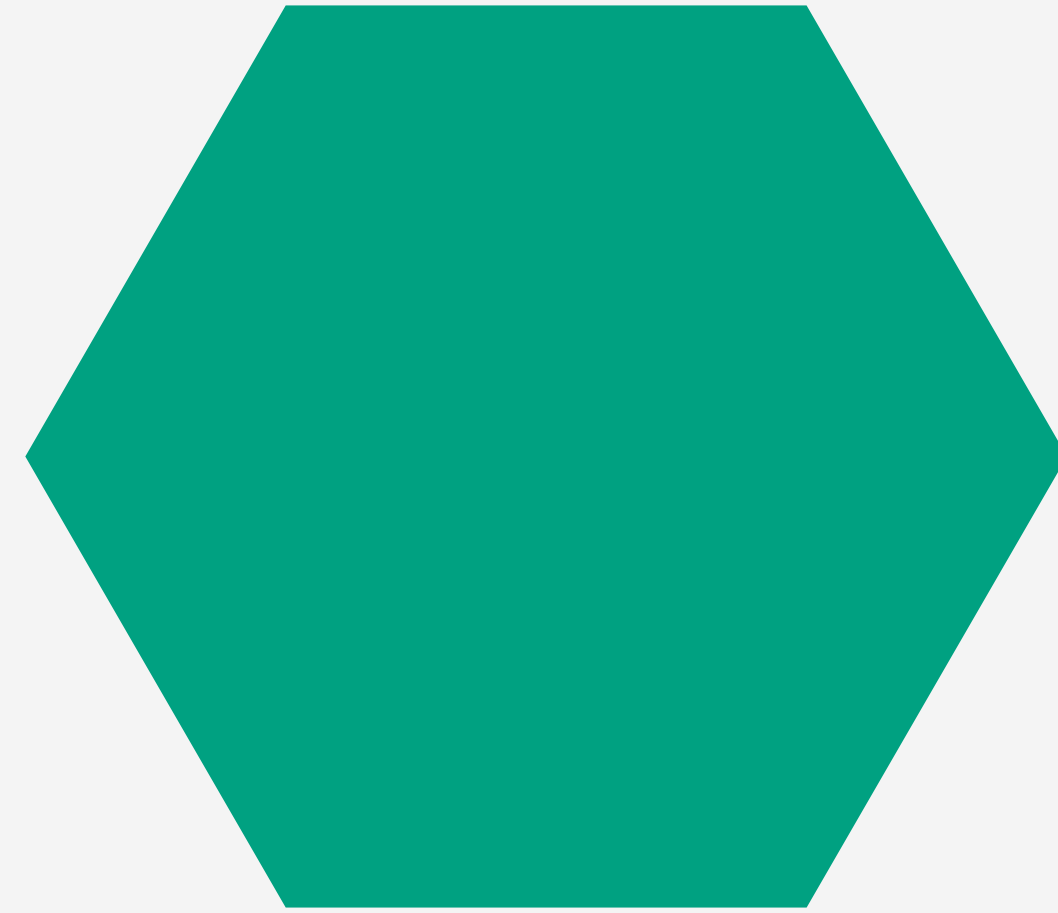
Modelo de clasificación de fuga potencial de
clientes



Introducción

Este proyecto se enfocó en el desafío de la predicción de la fuga de clientes dentro de una cartera de clientes.

El objetivo principal fue identificar características y patrones que permitan predecir con alta probabilidad qué clientes están en riesgo de abandonar la empresa.



El dataset



Se utilizó un dataset que contiene 26 columnas con información detallada de los clientes, incluyendo datos demográficos, financieros, transaccionales y de interacción con el servicio.

Las **columnas clave** incluyen:

- Datos del Cliente: Edad, Género, Tipo de Cuenta, Saldo de Cuenta, Score Crediticio, Es Empleado, Ingreso Anual, Estado Civil, Región.
- Datos Transaccionales: Fecha de Transacción, Monto de Transacción, Tipo de Transacción, Número de Transacciones, Tendencia de Actividad en Cuenta, Cambio en Saldo de Cuenta.
- Interacciones y Satisfacción: Interacciones de Servicio al Cliente, Quejas Recientes, Puntuación de Satisfacción del Cliente.
- Variable Objetivo: Estado del Cliente (Fugado/Activo).

Hipotesis



El proyecto se propuso validar la siguiente hipótesis alternativa:

"Los datos y variables existentes son suficientes para identificar a grupos de clientes con posibilidad de fuga, con más de un 60% de probabilidad del modelo."

Esta hipótesis se contrastó con la hipótesis nula, que postulaba la insuficiencia de los datos para lograr dicho nivel de predicción.

Transformaciones

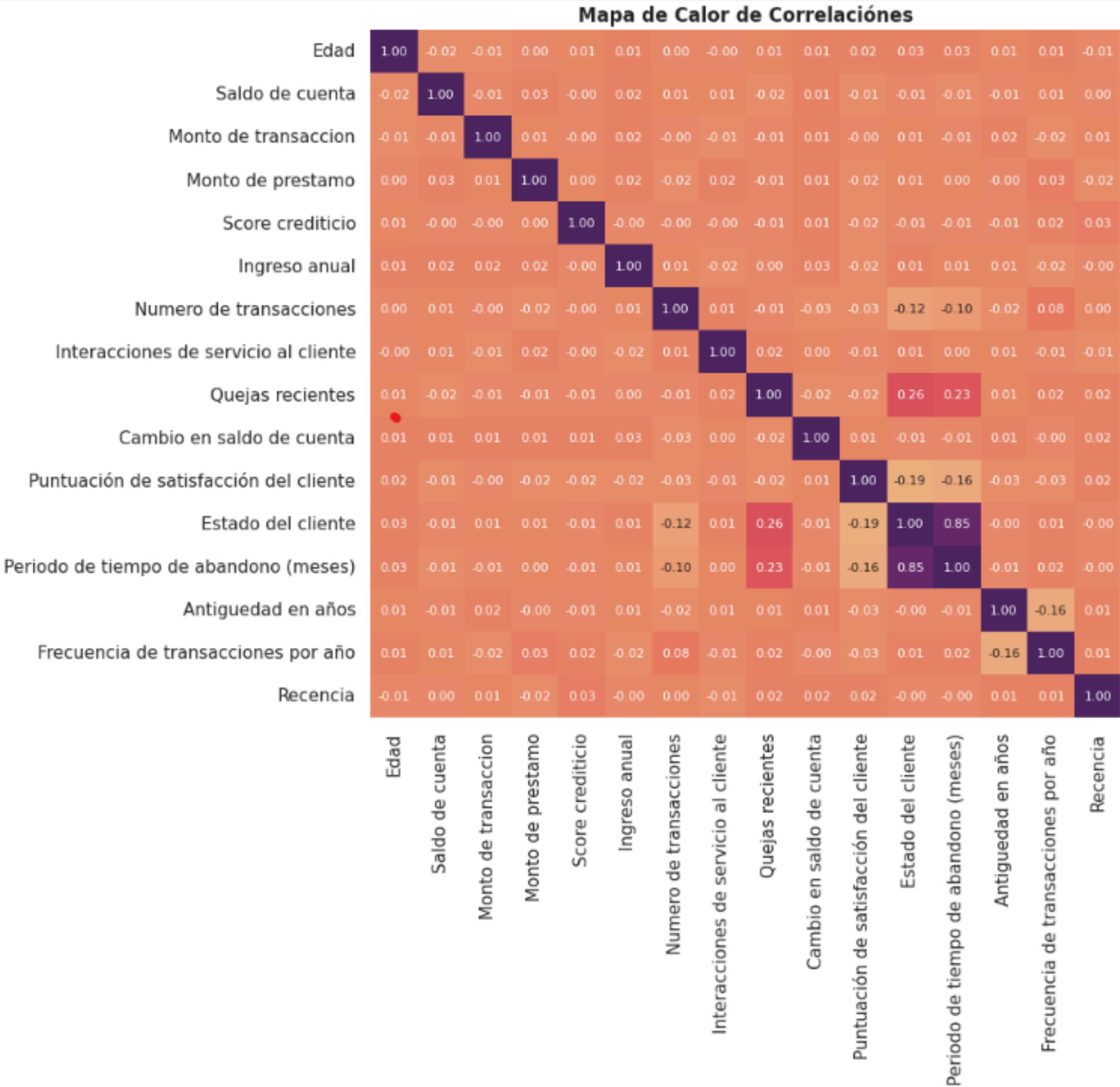


El procesamiento de datos incluyó:

- Traducción de columnas: Los nombres de las columnas se tradujeron al español para facilitar el análisis.
- Manejo de valores nulos: Se identificó y eliminó la columna 'Tipo de Préstamo' debido a un alto porcentaje de valores nulos (76%), superando un umbral del 30%.
- Transformación de tipos de datos
- Creación de nuevas variables: Se generaron nuevas características relevantes como 'Antigüedad en años', 'Frecuencia de transacciones por año' y 'Recencia' (tiempo desde la última transacción),etc.

Exploración de datos

El análisis exploratorio se centró en las variables numéricas continuas, examinando sus estadísticos descriptivos y visualizando su distribución mediante histogramas, diferenciando entre clientes activos y fugados. Además se identificaron correlaciones respecto a la variable objetivo.



Eleccion del modelo



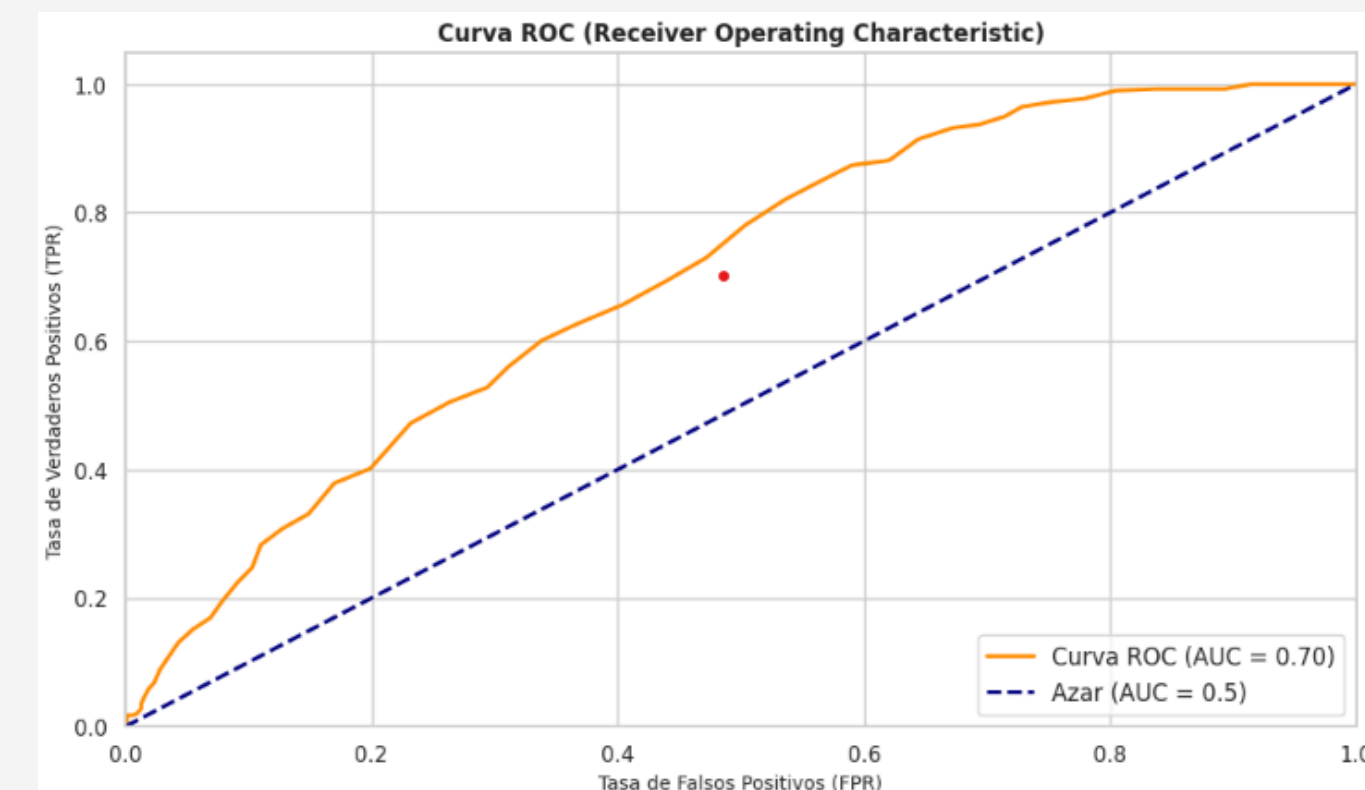
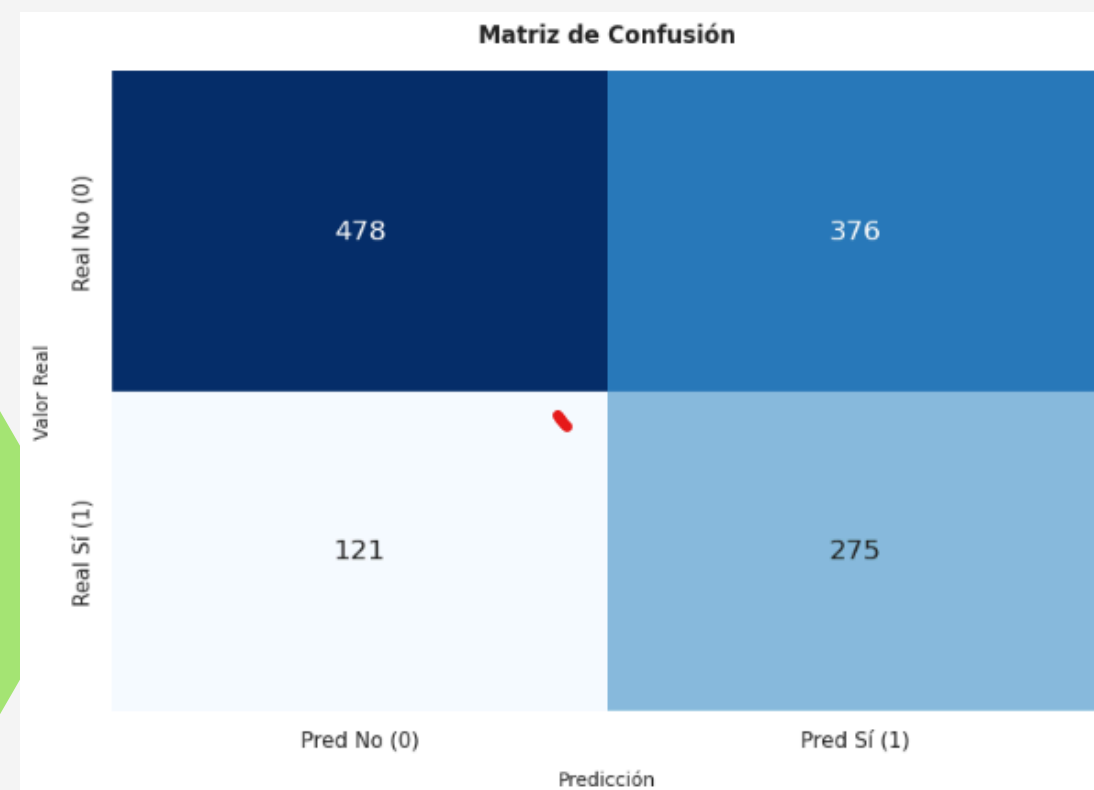
Se utilizó el modelo Random Forest, ya que este permite:

- Manejar datasets desbalanceados: Dada la menor proporción de clientes "Fugados" en comparación con los "Activos", fue crucial seleccionar un modelo que pudiera desempeñarse eficazmente en este escenario, ajustando el umbral de probabilidad para optimizar el desempeño, especialmente el recall.
- Identificar la importancia de las características
- Evaluar el modelo fácilmente: Se evaluó el modelo utilizando métricas como el recall (para asegurar la detección de la mayoría de los casos de fuga) y el AUC (para evaluar la capacidad general de discriminación entre clases).


La selección se orientó a maximizar la identificación de clientes en riesgo de fuga, incluso si esto implicaba un ligero aumento en falsos positivos, priorizando la detección temprana para intervenciones proactivas.

Resultados

- El modelo logró un recall de 0.69 en la marca "Fugado", lo que indica una buena capacidad para identificar a la mayoría de los clientes en riesgo de fuga, aunque con algunos falsos positivos.
- El AUC de 0.6971 confirma que el modelo puede diferenciar razonablemente entre clientes fugados y activos.
- El modelo es capaz de predecir con más de un 60% de probabilidad los casos de fuga.
- Las tres características más importantes identificadas por el modelo para predecir la fuga son:
 - a. Quejas recientes de los clientes.
 - b. Cantidad de transacciones al año.
 - c. Número de transacciones totales.



Validación de Hipótesis



Se logra rechazar la Hipótesis Nula y se valida la Hipótesis Alternativa. Esto significa que los datos y variables existentes son, suficientes para identificar a grupos de clientes con posibilidad de fuga, superando el umbral del 60% de probabilidad del modelo.

Este proyecto demuestra la viabilidad de utilizar técnicas de análisis de datos para la predicción de la fuga de clientes.

La identificación temprana de clientes permite a la empresa implementar estrategias de retención proactivas, lo que es crucial para evitar la disminución de la cartera de clientes y optimizar los esfuerzos de servicio al cliente.