PROCESO ANALÍTICO PARA LA PRUEBA DE PROPENSIÓN A LA ACEPTACIÓN DE OPCIONES DE PAGO

Yeison Andrés Correa Castaño
Bancolombia
2025

Introducción

Bancolombia se enfrenta al desafío de gestionar una cartera financiera con clientes con obligaciones morosas. En esta prueba se busca desarrollar un modelo de pronóstico para predecir la probabilidad de que un cliente en mora acepte una opción de pago preaprobada en el próximo mes. Este documento detalla el proceso para implementar y monitorear un modelo predictivo capaz de pronosticar la aceptación de opciones de pago ofertadas a estos clientes.

Descripción de las bases de datos

- **(data_rpta) Prueba_op_base_pivot_var_rpta_alt_enmascarado_trtest:** contiene la variable respuesta y características relacionadas con la gestión y las opciones de pago habilitadas.
- **(data_prob): Prueba_op_probabilidad_oblig_base_hist_enmascarado_completa:** incluye probabilidades generadas por modelos existentes (alerta temprana, auto cura, y propensión de pago).
- **(data_customer) Prueba_op_master_customer_data_enmascarado_completa:** contiene información demográfica de los clientes.
- **(data_payments)Prueba_op_maestra_cuotas_pagos_mes_hist_enmascarado_completa:** describe el comportamiento histórico de pagos de los clientes. Proceso analítico

Análisis exploratorio EDA

1. Análisis inicial de las bases de datos:

El análisis exploratorio de datos es una etapa fundamental en el proceso de análisis de datos. Su objetivo es entender la naturaleza y distribución de los datos antes de realizar cualquier modelado o inferencia estadística. En este análisis exploratorio se realiza un estudio inicial sobre las características del dataset, su origen, tamaño, limpieza de valores nulos (si los hay),

correlaciones y gráficas de distribuciones, lo que permite conocer de una manera más clara la información contenida por el dataset.

- Se realizó la lectura de las bases utilizando python (pandas) y se verificó su estructura.
- Se identificaron las claves primarias y relaciones entre tablas (p. Ej., nit_enmascarado).
- Se agregaron los índices por cada base para no eliminar datos, se observó que no tuvieran repetidos en cada subconjunto de datos ejemplo: nit_enmascarado, num_oblig_orig_enmascarado, num_oblig_enmascarado, fecha_var_rpta_alt.

- Para la base demográfica de clientes se creó la información asociada a las características generales del cliente de forma mensual, es decir, si un cliente solo aparece en un mes, se ampliará la información de los demás meses, para poder cruzar con la fecha de la base de respuesta o gestión.

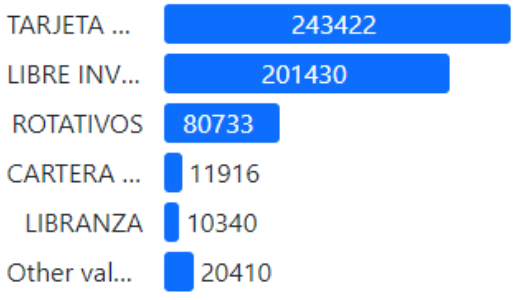
- Para la base de los pagos, se creó una base vectorizada donde se tiene por nit del cliente, nit de la obligación y la fecha, para obtener la información del cliente en el tiempo, así podemos ver la información histórica de los clientes.

2. Limpieza y transformación de datos

Creación de nuevas variables:

Una vez combinadas las cuatro bases de datos, se crearon variables adicionales basadas en conocimientos y relaciones detectadas en los datos.

- **Variable producto_n:** Se seleccionan solo los tipos de producto financieros más relevantes que posee el cliente. Esta variable puede incluir diferentes tipos de productos como créditos hipotecarios, créditos de consumo, tarjetas de crédito.

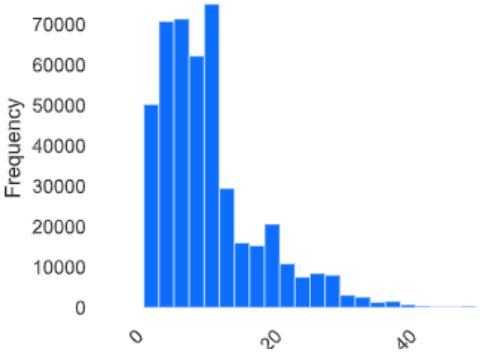


Distinct	10
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%

- **Variable dependientes:** Esta variable representa si el cliente tiene dependientes económicos o no. Anteriormente era una variable continua que indicaba cuantas personas dependen financieramente del cliente.

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%

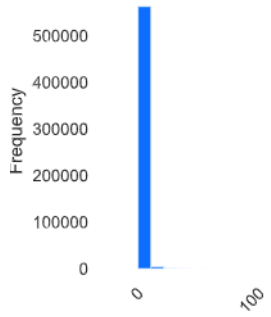
- **Variable antigüedad:** Refleja la duración de la relación del cliente con el banco, la medida está dada en años.



Minimum	1
5-th percentile	3
Q1	5
median	9
Q3	13
95-th percentile	27

Standard deviation	7.993138817
Coefficient of variation (CV)	0.7491094771
Kurtosis	24.06127322
Mean	10.67018782
Median Absolute Deviation (MAD)	4

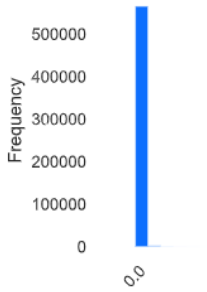
- **Variable relación pagos:** Representa la frecuencia y puntualidad en los pagos del cliente en relación con sus obligaciones. Indica la proporción de pagos que se han realizado en el tiempo.



Minimum	0
5-th percentile	0.1631443433
Q1	0.6988775607
median	1.376906561
Q3	2.365671338
95-th percentile	5.417759265

Standard deviation	3.285534358
Coefficient of variation (CV)	1.675704923
Kurtosis	8087.251049
Mean	1.960687895
Median Absolute Deviation (MAD)	0.7802644713

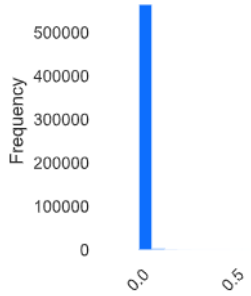
- **Variable promedio pagos:** Calcula el pago promedio realizado por el cliente en el tiempo.



Minimum	0
5-th percentile	38484.10417
Q1	217913.2083
median	555942.5
Q3	1384853.583
95-th percentile	5571792.771

Standard deviation	4786930.715
Coefficient of variation (CV)	3.157110892
Kurtosis	7364.284601
Mean	1516237.75
Median Absolute Deviation (MAD)	421557.3333

- **Variable promedio_valor_cuota_mora:** Indica el promedio del valor de las cuotas que el cliente ha dejado de pagar, acumulando mora.



Minimum	0
5-th percentile	65848.60417
Q1	240903.6458
median	451758.5833
Q3	854809.7292
95-th percentile	2441516.646

Standard deviation	1769618.955
Coefficient of variation (CV)	2.225153638
Kurtosis	4904.673864
Mean	795279.4468
Median Absolute Deviation (MAD)	258714.5417

Eliminación de columnas:

- Se eliminan las columnas que contienen un solo valor, ya no son significativas o no tienen ningún aporte al análisis.
- Se eliminan las columnas tipo fechas, ids, etc.
- Se eliminan columnas que contienen otras o están repetidas.
- Se eliminan los vectores ya que se crearon variables con estas.
- Se eliminan las variables con correlación > 0.85

Análisis de características relevantes:

Se identificaron características que tienen mayor influencia en la variable objetivo usando análisis de correlación. La correlación entre las variables numéricas y la variable respuesta var_rpta_alt. Se calcula con el coeficiente de correlación de Pearson.

```
var_rpta_alt            1.000000
promesas_cumplidas      0.309341
porc_pago_cuota         0.149906
prob_propension         0.141272
porc_pago_mes           0.123206
pago_mes                0.119021
prob_auto_cura          0.118420
cant_alter_posibles     0.116634
pago_cuota              0.076250
relacion_pagos          0.069360
promedio_pagos          0.056331
saldo_capital           0.023142
```

Distribución de la variable objetivo (var_rpta_alt).

Value	Count	Frequency (%)
0	295483	52.0%
1	272768	48.0%

El gráfico indica que la base de datos está relativamente balanceada, con una ligera preponderancia del valor 0.

Visualización de datos

Para este parte se realizó un informe con la ayuda de **ProfileReport** donde me muestra variable por variable usando gráficos para entender las distribuciones y relaciones entre ellas.

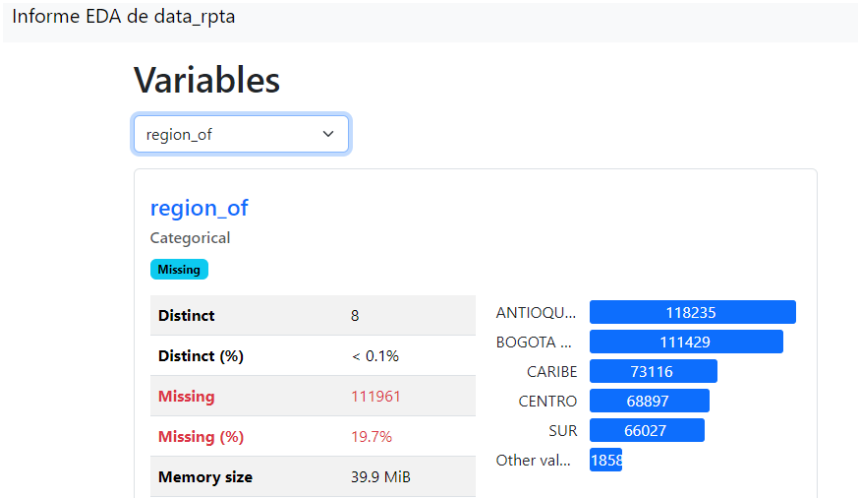
Dataset statistics

Number of variables	57
Number of observations	568251
Missing cells	4692418
Missing cells (%)	14.5%

Variable types

Numeric	28
Categorical	29

En el informe se pueden seleccionar una a una las variables y ver todas sus estadísticas, ejemplo:



Identificación de valores faltantes:

- Valores numéricos: imputación por o media.
- Valores categóricos: imputación con la moda.

Codificación de variables categóricas:

Se realizó la codificación de variables categóricas usando One-Hot Encoding.

Normalización:

Escalado de variables numéricas mediante StandardScaler. Se usa la normalización y la estandarizamos de datos numéricos para asegurar la coherencia en el análisis.

Selección y entrenamiento del modelo.

Una vez se tiene un conocimiento preliminar del comportamiento de los datos, sus principales estadísticas, distribuciones, se procede a realizar un modelamiento de datos. Para esto se debe crear una variable respuesta **var_rpta_alt** como variable objetivo y las otras variables como variables dependientes.

1. **División del conjunto de datos.**

Se dividieron los datos en conjuntos de entrenamiento y prueba con un ratio del 70:30.

2. Modelos probados

Se probaron varios algoritmos predictivos

- **Regresión logística**: modelo base.

```
Model: Logistic Regression
Cross-validated F1 Score: 0.81
Accuracy: 0.8079436401604918
F1 Score: 0.8046570230629531
Classification Report:
              precision    recall  f1-score   support

           0       0.76      0.92      0.83     88769
           1       0.89      0.68      0.77     81707

    accuracy                           0.81    170476
   macro avg       0.82      0.80      0.80    170476
weighted avg       0.82      0.81      0.80    170476
```

- **Decision Tree**.

```
Model: Decision Tree
Cross-validated F1 Score: 0.84
Accuracy: 0.8426699359440625
F1 Score: 0.8426909874759753
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.85      0.85     88769
           1       0.83      0.84      0.84     81707

    accuracy                           0.84    170476
   macro avg       0.84      0.84      0.84    170476
weighted avg       0.84      0.84      0.84    170476
```

- **Random forest:** modelo interpretable con capacidad de manejar no linealidades.

```
Model: Random Forest
Cross-validated F1 Score: 0.90
Accuracy: 0.899194021445834
F1 Score: 0.8988960081288838
Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.94      0.91     88769
           1       0.93      0.86      0.89     81707

    accuracy                           0.90    170476
   macro avg       0.90      0.90      0.90    170476
weighted avg       0.90      0.90      0.90    170476
```

- **Gradient boosting (xgboost):** modelo avanzado para optimizar el f1 score.

```
Model: XGBoost
Cross-validated F1 Score: 0.89
Accuracy: 0.8871981979868134
F1 Score: 0.8868078767750006
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.93      0.90     88769
           1       0.92      0.84      0.88     81707

    accuracy                           0.89    170476
   macro avg       0.89      0.89      0.89    170476
weighted avg       0.89      0.89      0.89    170476
```

Validación cruzada:

Evaluación del f1 score en cada iteración.

El mejor modelo es: RandomForest con una Cross-validated F1 Score de 0.90

Ajuste de hiperparámetros:

Se realizó una búsqueda de hiperparámetros utilizando **GridSearchCV** para optimizar el rendimiento del modelo.

Entrenamiento del modelo:

Se plantean varios modelos de clasificación y se hacen algunos ajustes de hiper parámetros para los modelos. Al final de cada modelo, se presentan las métricas al evaluar los datos del subconjunto de prueba.

Entrenamos cada modelo con los datos preprocesados y evaluamos su rendimiento utilizando métricas como el F1 Score.

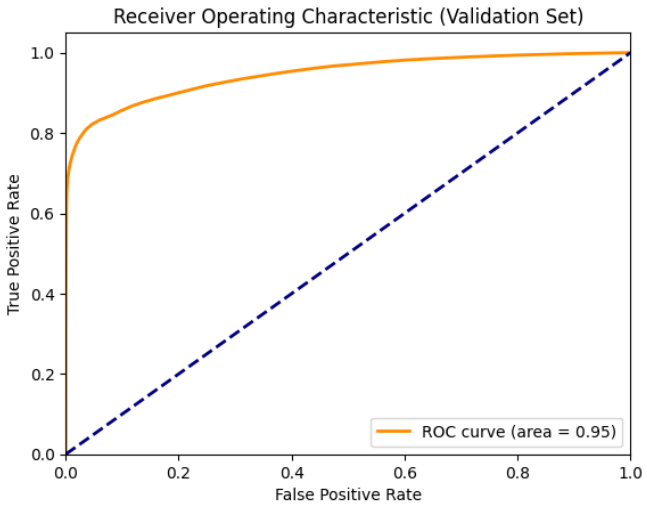
3. Evaluación del Modelo.

En esta sección se resumen los resultados obtenidos por todos los modelos de Machine Learning y se selecciona el modelo que presente mejores resultados según los requisitos del problema.

Seleccionar el mejor modelo:

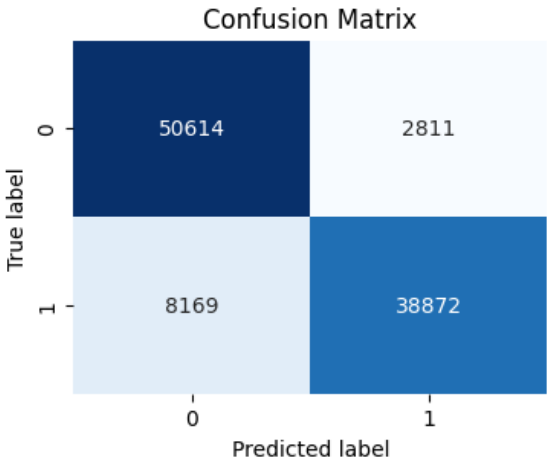
- Basamos la selección en el F1 Score del conjunto de prueba.

Curva ROC y AUC: Se genera la curva ROC y calculamos el área bajo la curva (AUC) para evaluar el desempeño del modelo.



Con un área bajo la curva (AUC) de 0.95, el modelo demuestra un alto nivel de precisión y capacidad discriminatoria, lo que indica que es efectivo para predecir la aceptación de opciones de pago por parte de los clientes.

Matriz de confusión: Se obtiene la matriz de confusión para visualizar el desempeño del modelo en términos de verdaderos positivos/negativos y falsos positivos/negativos.

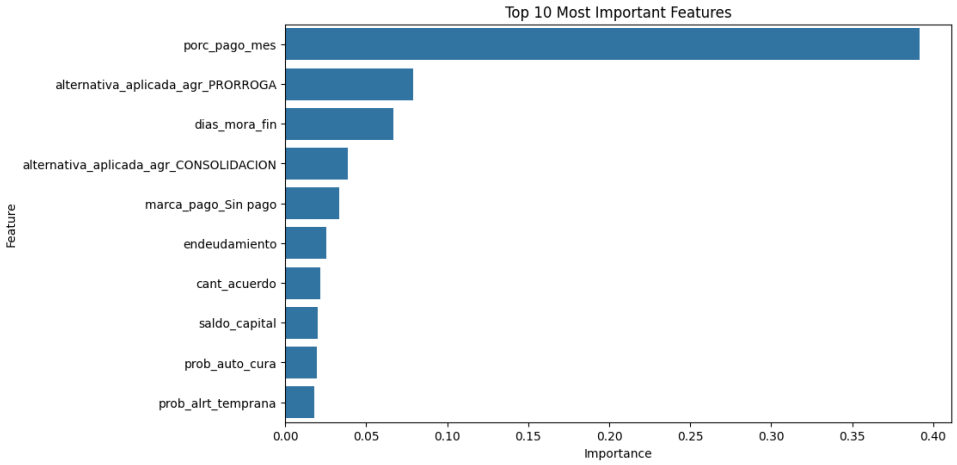


La matriz de confusión nos ofrece una visión detallada y cuantitativa del rendimiento del modelo de clasificación. La alta proporción de verdaderos positivos y negativos sugiere un alto nivel de precisión y fiabilidad en las predicciones del modelo. Las bajas proporciones de falsos positivos y negativos reflejan que el modelo es efectivo en la clasificación de instancias, minimizando errores de predicción.

El modelo de clasificación evaluado muestra un rendimiento robusto y eficiente, lo cual es esencial para la detección y gestión oportuna de opciones de pago entre los clientes en mora.

Importancia de características:

Se identifican las características más importantes para el modelo seleccionado y se presenta los resultados en gráficos ordenados.



Inferencia y generación de resultados.

Predicciones: Cálculo de prob_uno (probabilidad de aceptación) y asignación de var_rpta_alt con un umbral > 0.5.

Formato de salida: Generación del archivo csv en el formato requerido.

Recomendaciones para mejora de datos.

- Historial de llamadas o interacciones con los clientes, podemos sacar datos como a que hora se le marco al cliente, cual asesor lo llamo, que día de la semana, etc.
- Mejorar la calidad de los datos socioeconómicos como ingresos o nivel de educación: estas bases de datos contienen muchos datos nulos que podrían completarse con información de buros.
- Variables macroeconómicas: como tasas de interés.

Factibilidad:

Recolección de estos datos puede implicar costos adicionales y regulaciones de privacidad.

Conclusiones

- Se desarrolló un modelo robusto y escalable que cumple con los objetivos planteados, optimizando el F1 Score. La solución incluye recomendaciones claras para su implementación y estrategias para mejorar continuamente su desempeño.

- El modelo que se escogió fue un random forest ya que tiene mayor F1 Score, aunque si es por temas de cómputo e interpretabilidad de las variables en torno al modelo se podría elegir la regresión logística.

- El modelo de predicción de ha mostrado un rendimiento sobresaliente no solo con la métrica F1 Score si no con alta precisión, recall y AUC, lo que permite al banco implementar estrategias de retención personalizadas para propensión a la aceptación de opciones pago.

Diseño teórico del sistema para disponibilizar los resultados del modelo analítico.

El objetivo es diseñar un sistema escalable y eficiente que permita consumir los resultados del modelo analítico de manera sencilla desde servicios externos como páginas web, aplicaciones móviles, u otros sistemas bancarios.

A continuación, se describe la solución teórica.

1. Infraestructura en la Nube

- **Servicios en la Nube**: Utilizaremos AWS o Azure para asegurar que nuestros servicios sean escalables, confiables y siempre estén disponibles.
- **Almacenamiento**: Usamos MongoDB (NoSQL) para almacenar los resultados del modelo y PostgreSQL (SQL) para información transaccional.

2. Componentes del Sistema MLOps

Preprocesamiento y Entrenamiento de Modelos:

- **Pipeline de Datos**: Usamos Apache Airflow o Kubeflow Pipelines para orquestar el flujo de trabajo de datos y el entrenamiento del modelo.
- **Almacenamiento de Modelos**: MLflow para guardar, versionar y gestionar los modelos entrenados.
- **Repositorio de Modelos**: Los modelos entrenados se almacenan en el MLflow Model Registry**.**

Servicio de Almacenamiento de Resultados:

- **Qué Hace**: Almacena las predicciones en una base de datos.
- **Herramientas**: Implementación con Node.js o Django.
- **Endpoint**: /store, que recibe y guarda los resultados.

Servicio de Consulta de Resultados:

- **Qué Hace**: Permite que servicios externos consulten los resultados almacenados.
- **Herramientas**: Utilizamos Express, FastAPI o Spring Boot.
- **Endpoint**: /results, donde se proporcionan los resultados basados en criterios de búsqueda.

3. Integración con Servicios Externos

- **Páginas Web y Aplicaciones Móviles**:

  Acceso a Servicios: Consumimos los resultados del modelo a través de una API REST.

  Interfaz de Usuario: Diseño de UI intuitivas y responsivas para la interacción del usuario final.

4. Monitoreo y Mantenimiento

- **Mantenimiento y Actualizaciones**:

  Pipeline CI/CD: Implementamos un pipeline de Integración y Despliegue Continuo con herramientas como GitHub

  Entorno de Pruebas: Validamos nuevas versiones y actualizaciones en un entorno de pruebas antes de desplegarlas en producción.

Flujo de Trabajo General:

1. **Entrenamiento y Preprocesamiento**: Automatiza la preparación y entrenamiento del modelo a través de pipelines.
2. **Almacenamiento del Modelo**: Los modelos se almacenan, versionan y gestionan en MLflow.
3. **Inferencia en Tiempo Real**: Los servicios externos pueden realizar predicciones en tiempo real utilizando los modelos desplegados.
4. **Almacenamiento y Consulta de Resultados**: Los resultados se almacenan y pueden ser consultados fácilmente por otros servicios.
5. **Monitoreo y Mantenimiento**: Monitoreamos el sistema en tiempo real y realizamos actualizaciones continuas para garantizar su eficiencia.