

Prueba Analítica: Solución analítica para la empresa High Garden coffee

Daniel Felipe Pérez-Grajales *a*

*a Estadístico, especialista de Inteligencia Artificial, Facultad de Ciencias,*

*Universidad Nacional de Colombia, Medellín, Colombia.* [*dfperezg@unal.edu.co*](mailto:dfperezg@unal.edu.co)

***Objetivo:***

Este reto técnico busca reconocer tus habilidades en el área. Eres libre de proponer y desarrollar la opción de solución. Ten en cuenta el contexto que se dará a continuación y Asegúrate de tomar decisiones conscientes, basadas en la analítica de datos.

**Contexto: Empresa “High Garden coffee”:**

La empresa Garden Coffee es una exportadora de café de nivel internacional. Actualmente tiene una base de datos de consumo doméstico de café que abarca datos desde 1990 hasta 2020 que incluye diferentes países y tipos de café. El área de innovación busca aprovechar los datos para lograr una ventaja competitiva en el mercado identificando tendencias, rangos de consumo, información clave y cualquier otra información que pueda resultar valiosa. Como científico de datos en jefe tu labor será utilizar el set de datos para crear una solución basada en analítica que aproveche las últimas tendencias como ML, IA generativa, LLM o cualquier otra que consideres para generar modelos y entregables de valor a la compañía.

Requisitos mínimos de solución:

• Análisis de la información

• Solución a las problemáticas de negocio usando técnicas de analítica

• Implementación y evaluación de la solución

• Presentación de los resultados

**BONUS.**

Ahora que has resuelto el problema, cuéntanos como podrías integrar las capacidades de IA generativa, LLMs o soluciones de chatbot para dar mayor valor a la solución desarrollada.

• No es obligatorio que lo implementes en arquitectura cloud, pero si quieres adelante

EXITOS!

1. ***Metodología:***

Solución a partir de las etapas de Modelamiento y a través del software Python:

0. Carga de Módulos

1. ETL

1.1. Manejo de base de datos

1.2. Creación de variables de interés

2. Métodos de imputación y transformación de variables

2.1 Análisis (EDA)

2.2 AutoML pre modelamiento

3. Configuración de pipeline (transformación, imputación, selección de variables…)

4. Definición de modelos (cross-validation)

5. Entrenamiento y selección de Modelos

5.1. Tuneo de hipérparametros

6. Interpretación de resultados (métricas ajustes, objetivo F1)

7. Marcación Fuera de muestra(producción)

1. ***Proceso ETL:***

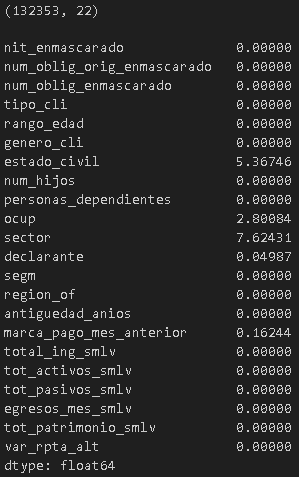
De la base ***prueba\_op\_maestra\_cuotas\_pagos\_mes\_hist\_enmascarado\_completa:*** Información que describe el comportamiento de pagos del cliente en sus obligaciones a lo largo del tiempo. Para completar la base de conocimiento para entrenar el modelo se va tomar en cuenta la variable ‘**marca\_pago’** del mes anterior a la predicción del modelo, dentro de las variables que se podrían sumar, pero tienen un costro de procesamiento, comportamiento de pago de los últimos 6 meses por cliente…

La base **prueba\_op\_master\_customer\_data\_enmascarado\_completa** Información asociada a las características generales del cliente o demográficas de forma mensual. Se identificaron variables que aportaban al perfil del cliente para aceptar una opción de pago y que no evidenciaran falta de información(NA’S):**'nit\_enmascarado','num\_oblig\_orig\_enmascarado','num\_oblig\_enmascarado','tipo\_cli','rango\_edad','genero\_cli','estado\_civil', 'num\_hijos', 'personas\_dependientes', 'ocup', 'sector', 'declarante','segm', 'region\_of',** se construyó la variables antigüedad del cliente en el mes de estudio en años: **'antiguedad\_anios',** variables continuas numéricas se transformaron a SMLV como : **'total\_ing\_smlv', 'tot\_activos\_smlv','tot\_pasivos\_smlv', 'egresos\_mes\_smlv', 'tot\_patrimonio\_smlv'.** De esta base solo se tiene en cuenta para entrenamiento 5 meses 202307,202308,202309,202310,202311 para validación fuera de muestra 202312

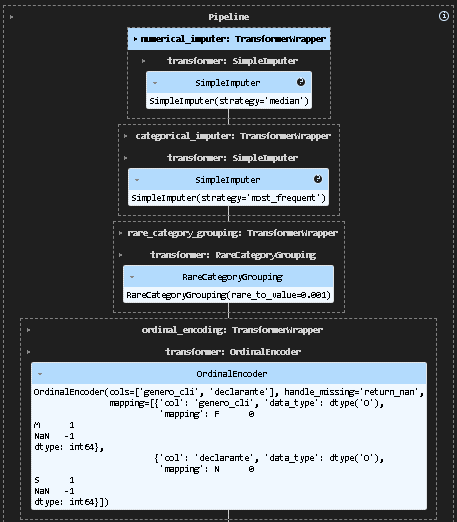
La base **prueba\_op\_base\_pivot\_var\_rpta\_alt\_enmascarado\_trtest:** Base de datos que contiene la variable respuesta y las principales características de la gestión, de resultados del pago y de características de las opciones de pago habilitadas para el cliente en el mes de gestión o de evaluación de la variable respuesta**.** se obtiene el target para la construcción del modelo predictivo ‘**var\_rpta\_alt’.**

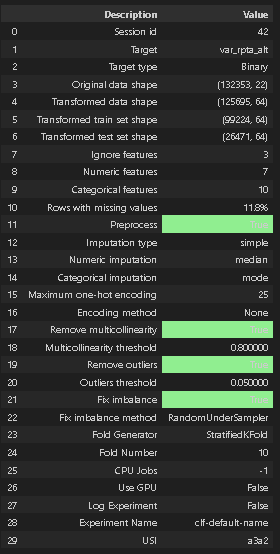
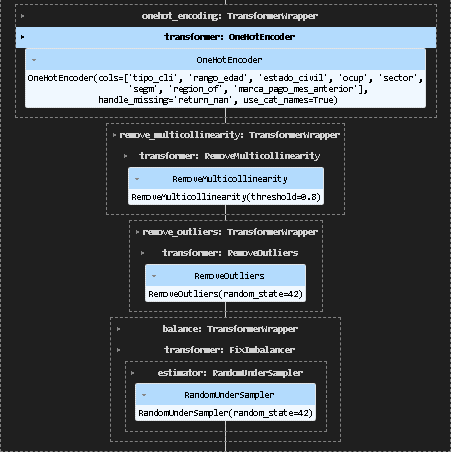
De la base **prueba\_op\_base\_pivot\_var\_rpta\_alt\_enmascarado\_oot:** Base de datos que contiene la variable respuesta y las principales características de la gestión, de resultados del pago y de características de las opciones de pago habilitadas para el cliente en el mes de gestión o de evaluación de la variable respuesta. Se obtuvo una muestra del 30% para probar el modelo fuera de muestra y marcar al final la base **resultado\_prueba.csv**

1. ***Métodos de imputación y transformación de variables:***

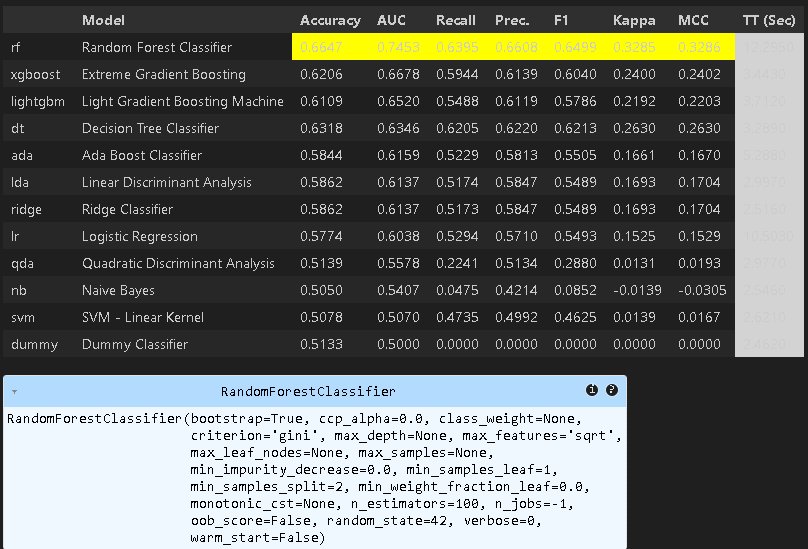
Se consolido la base de conocimiento para el entrenamiento del modelo con las siguientes variables:

1. ***Modelamiento AutoML:***

Por medio de la librería pycaret AutoML se hizo la configuración del pipeline, donde se puede configurar todo el proceso de imputación de variables con NA’s en numéricas como categóricas, transformación one hot encoding o enconder, eliminación de datos atípicos (IsolationForest,LocalOutlierFactor…), selección de variables por multicolinealidad, desbalanceo de categorías y variable objetivo 'RandomUnderSampler' .., selección de muestra para el data train y test estratificada:



1. ***Definición de modelos:***

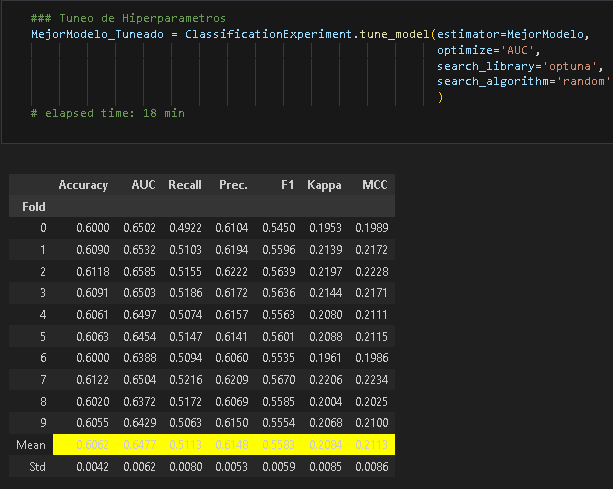
Se dieron a competir 12 modelos de Machine learning de clasificación, donde como objetivo para su selección se tomó en cuenta métricas como el AUC y el F1 Score adicionalmente para evitar temas se sobre ajustes en los modelos se configuro el pipeline con Cross-Validation estratificada en la variable objetivo y 10 lotes, donde nos entregó los siguientes resultados:

Durante este proceso tarda aproximadamente 40 min y nos entrega como mejor candidato el Random forest Classifier(RF):

Y tenemos como **F1 score 64,99%**

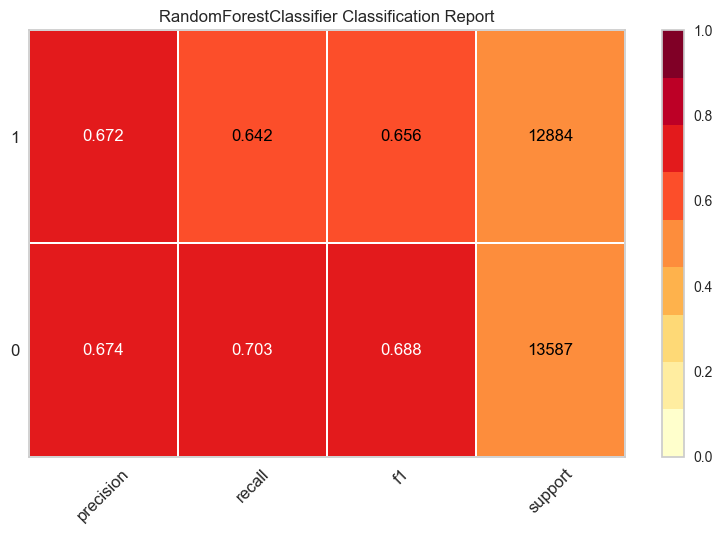
1. ***Optimización de hiperparametros del mejor modelo:***

Se intento Tunear los Hipérparametos del modelo para mejorar la métrica F1 score, para este proceso se implemente optuna donde se dispuso como objetivo encontrar los parámetros del modelo con respecto a maximizar el AUC y el F1 score:

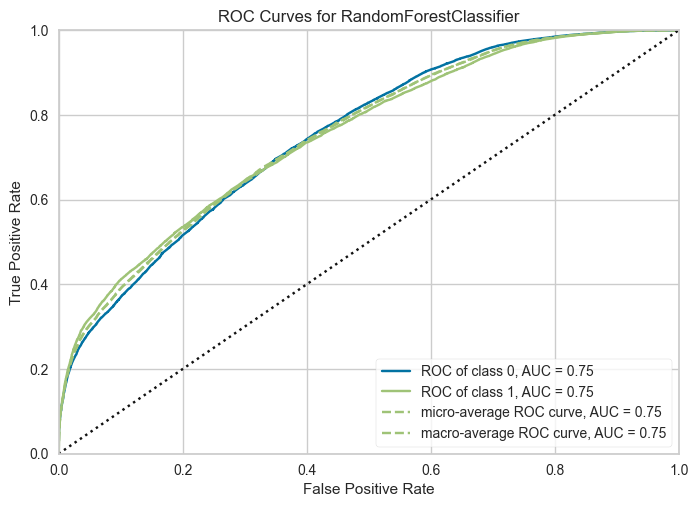


Implementado cross-validate para 10 lotes no se encontró una meja de F1 score tuneado los hiperparametros por lo que se decide continuar con el mejor modelo inicial RF.

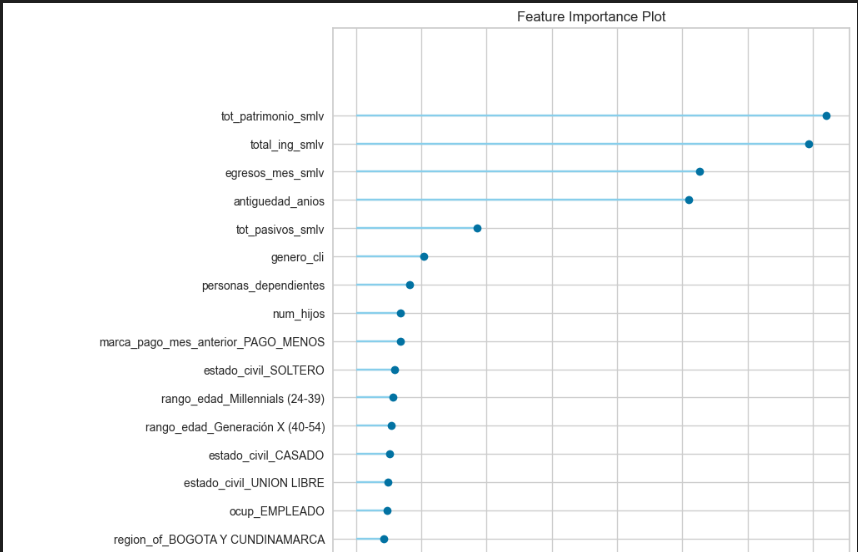
1. ***Interpretación de resultados (métricas ajustes, objetivo F1):***



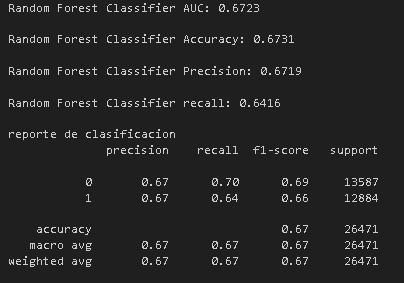
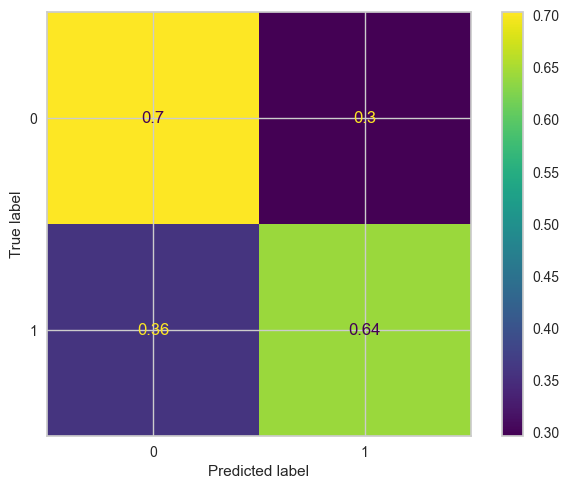
En el proceso de entrenamiento se tiene para el modelo RF un recall para predecir una aceptación de las opciones de pago del 65.6 % de los casos y un F1 score de no aceptar las opciones de pago de 68,8%



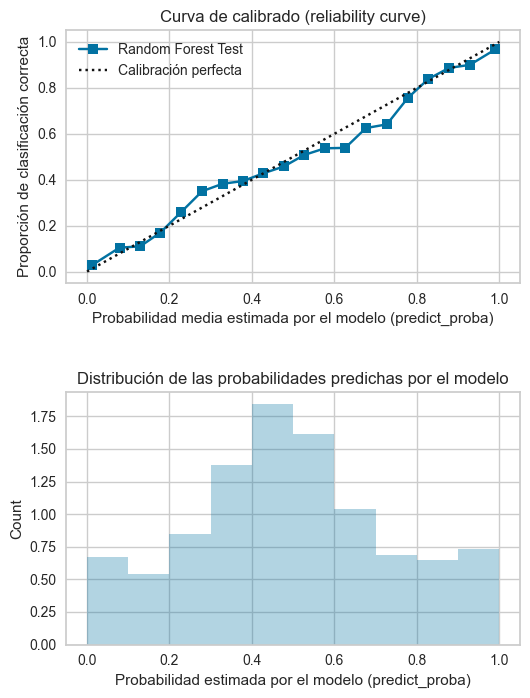
Se tiene un área bajo la curva de 75%



Las Variables muy importantes para la predicción del modelo es: el total de patrimonio en smlv, total ingresos en smlv, egresos del mes y antigüedad en años del cliente.

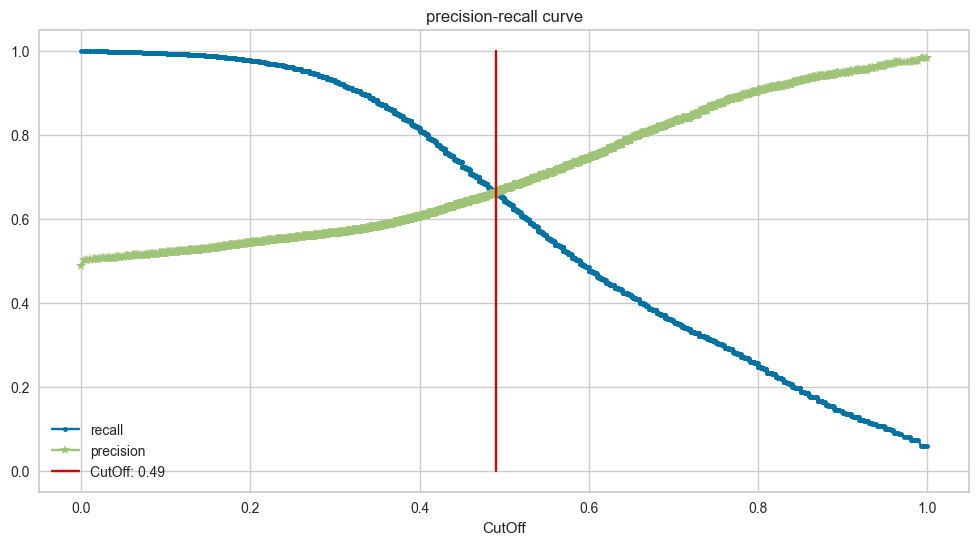
 

En un 64% se tienen casos positivos a aceptar las opciones de pago.



Distribución de la predicción probabilística por deciles nos muestra que el modelo tiende a se mas una moneda al predecir si acepta o no las opciones de pago, se debería focalizar más en los extremos.

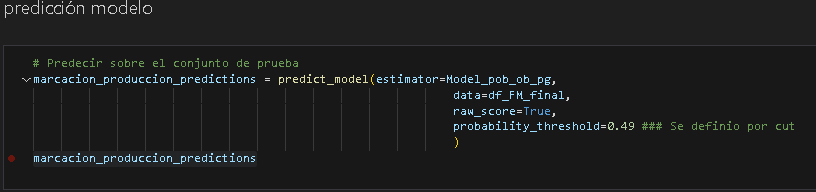
**Cutoff**: punto de corte probabilístico:

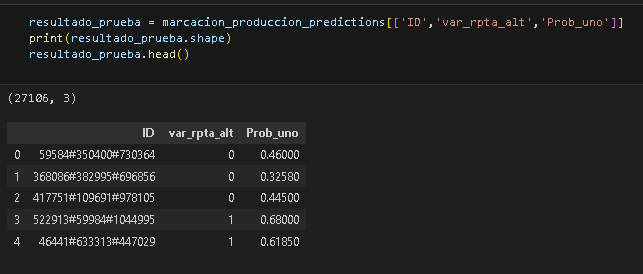


Optimización del punto de corte probabilístico donde se busca encontrar el punto de corte de la probabilidad donde maximicemos especificidad y minimizamos la sensibilidad, el mejor valor de corte que minimiza el costo de falsos positivos y falsos negativos, para este modelo es fijarlo en 0.49.

1. ***Test modelo data validación y marcación fuera de muestra:***

Se extrae el 30% de la base completa fuera de muestra donde se toma en cuenta en la predicción del modelo ese punto de corte optimo:





Se marca una muestra FM de 27.106 clientes para el mes de 202401.