Informe Técnico T.P. Final

UTN FRBA - Ciencia de Datos - Curso I5521 - Grupo n°5 - Año 2025

Integrantes:

Leandro Omelanchuc - Legajo: 172880-5
Franco Vanella - Legajo: 166394-0

Índice:

1. Introducción

El objetivo de este trabajo práctico es implementar las técnicas adquiridas en la cursada, relacionadas a el preprocesamiento de datos, análisis exploratorio de datos y sus gráficos , evaluaciones estadísticas y la implementación de modelos de aprendizaje supervisados o no supervisados de generación de redes neuronales para resolver un problema simulado entregado por los profesores de la cátedra al grupo

2. Problema Planteado

El problema plantea que un banco solicitó nuestra colaboración para intentar predecir lo mejor posible qué clientes de su cartera se suscribirán a productos ofrecidos dentro de una campaña de marketing. El mismo banco sería quien nos proporcionará un conjunto de datos que contiene información de unos 45.211 clientes, incluyendo 17 variables que describen distintas características de cada uno.

3. Objetivos del Trabajo

Se solicita utilizar técnicas de EDA y luego implementar un pipeline de machine learning para generar un modelo que permita predecir en base de los datos proporcionados las características de los clientes más adecuados para suscribirse al producto ofrecido por el banco.

4. Descripción del Dataset

El dataset original provisto incluye un total de 45.211 líneas por cada uno de sus clientes y 17 columnas por cada una de las características de los mismos.
En las columnas se incluyen:

- age : Edad del cliente
- job : Tipo de empleo del cliente
- marital status : Estado civil
- education : Educación máxima alcanzada por el cliente
- Credit : Si tiene deuda de crédito o no
- balance : Promedio de saldo en la cuenta en el año
- housing loan : Si tiene préstamos hipotecarios o no
- Persona loan : Si tiene préstamos o no
- contact : tipo con contacto del cliente
- Last Contact Day : Ultimo dia de contacto con el cliente en el mes
- Last Contact Month : Último mes de contacto con el cliente en el año
- Last Contact Duration : Duración del último contacto con el cliente medido en segundos
- campaign : Cantidad de contactos al cliente durante esta campaña, incluye el último contacto.
- pdays : Cantidad de días que pasaron del último contacto con el cliente de una campaña anterior. -1 significa que no hubo contacto previo
- previous : Cantidad de contactos previos a esta campaña para cada cliente
- poutcome : Performance de la campaña de marketing anterior para este cliente

5. Metodología, herramientas y librerías utilizadas

Herramientas:

- Python
- Jupyter notebooks
- Anaconda navigator
- Google Collaboratory

Librerías utilizadas para EDA y preprocesamiento de datos:

- Pandas
- Numpy
- Seaborn
- Matplotlib

Librerías utilizadas para Modelación:

- Scikit learn
- Pytorch

Metodología:

En primer lugar se realiza la importación de las librerías de EDA para la manipulación de los datos dados en la etapa de preprocesamiento.
Luego, se procede a evaluar la información presente en el Data Frame provisto, validando que datos están ausentes o si son inconsistentes para luego ser reemplazados con datos en formatos válidos
una vez realizado esto se procede al feature engineering para la generación de columnas dummies para reemplazar las variables categóricas, para el caso de las variable numéricas se estandariza y normalizan los valores para obtener los resultados más significativos y facilitar la generación de los modelos, tal como es recomendado en el Articulo numero 2
(Ver seccion de referencias al final del documento)
Con los datos ya procesados se pasa a realizar el análisis de los mismos a través de la generación de gráficos, mediciones estadísticas y análisis de la correlación de los mismos para sacar conclusiones iniciales a utilizar en la modelación de los mismos.

6. Preprocesamiento de Datos

Para el preprocesamiento de datos, primero se validan los datos incluidos en el dataset, lo cual consiste en verificar:

- La cantidad de datos no nulos, NaN's y datos inconsistentes para cada una de las variables.
- Generación de un nuevo set de datos a ser modificado
- modificar las variables características faltantes asumiendo los valores negativos (completar la tabla con "no")

- En caso de ser de las variables de “job”, “matrial_status” o “education”, completamos con los valores de la moda obtenida
- para las variables numéricas como la edad y el balance, se completan con el valor correspondiente a la mediana de cada variable.

Como último paso, luego de esta modificación, validamos el estado final del nuevo dataset y al final de EDA, se genera una copia de backup del mismo.

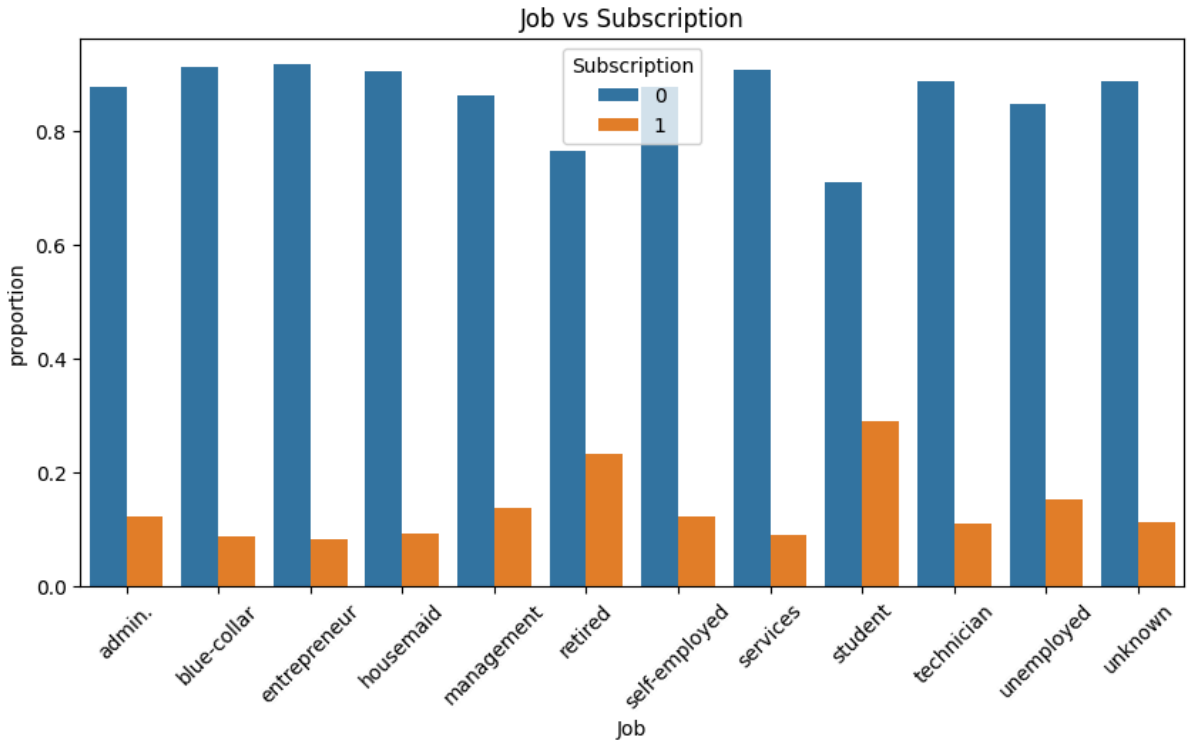
7. Análisis Exploratorio de Datos (EDA)

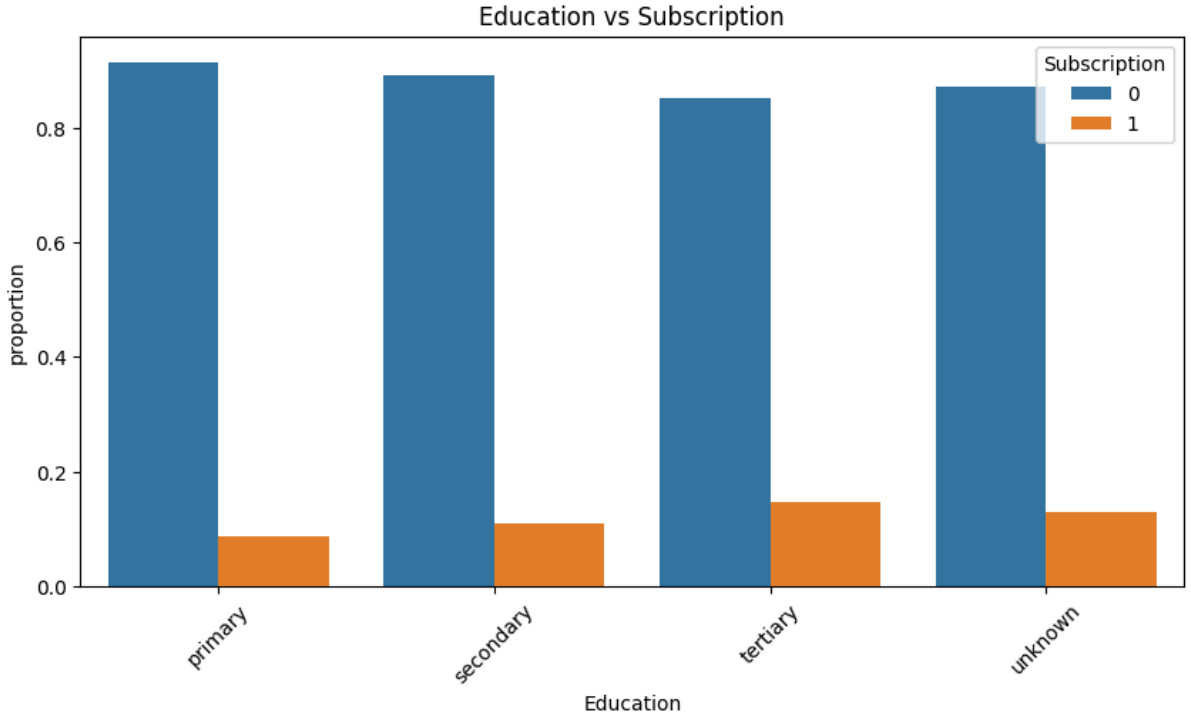
Ya modificado el dataset, validamos los siguientes valores del mismo:
- se verifica los valores que toman cada una de las variables
- se chequea la cantidad de apariciones de cada uno de los valores posibles para las variables

Luego a través de gráficos se analiza:
- La distribución de la variable “subscription” con un countplot
- Con un histograma, la distribución de las variables numéricas
- Se realiza un boxplot para identificar los cuartiles de apariciones de las variables numéricas
- la distribución de los mismo valores pero comparados segun el valor de la variable subscription (para variables numéricas con un kdeplot y para las variables categóricas con gráfico de barras)

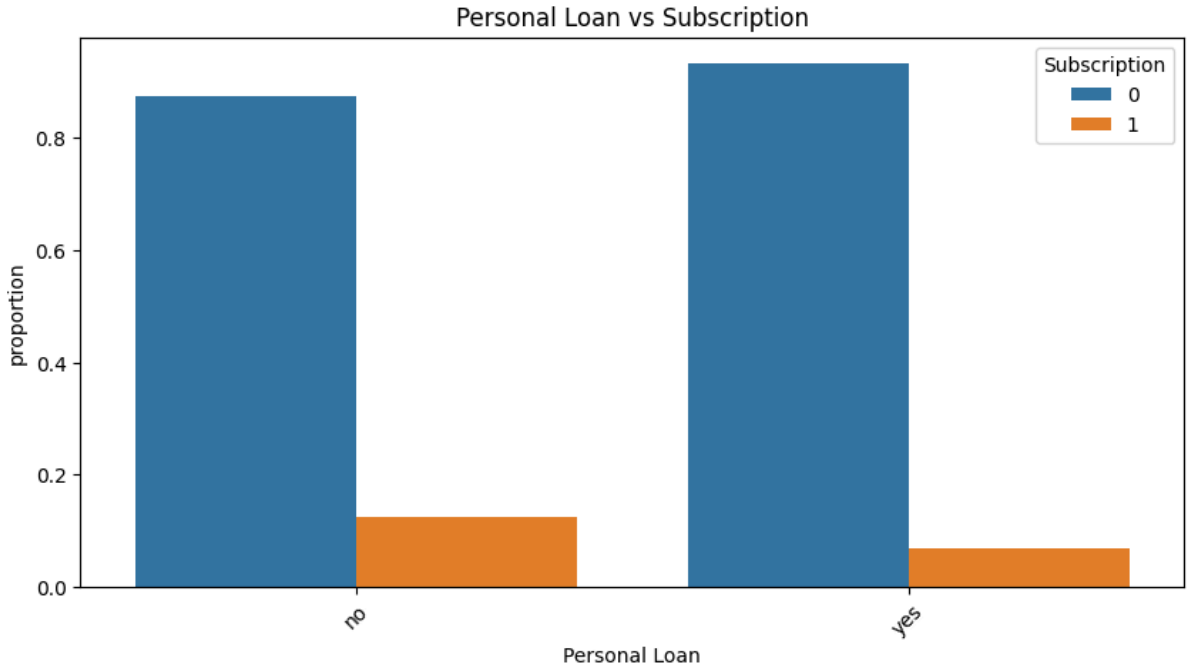
Finalizado el preprocesamiento de los datos otorgados, evaluamos los parámetros estadísticos de los datos. En esta etapa evaluamos la correlación lineal entre aquellas variables numéricas.
Para esto generamos un nuevo set que solo incluya estas variables y las comparamos con en un heatmap.
Para las variables categóricas, generamos un nuevo set de variables dummies o reagrupamos los valores dados en subcategorías para mejorar el procesamiento posterior.

8. Distribuciones y Visualizaciones

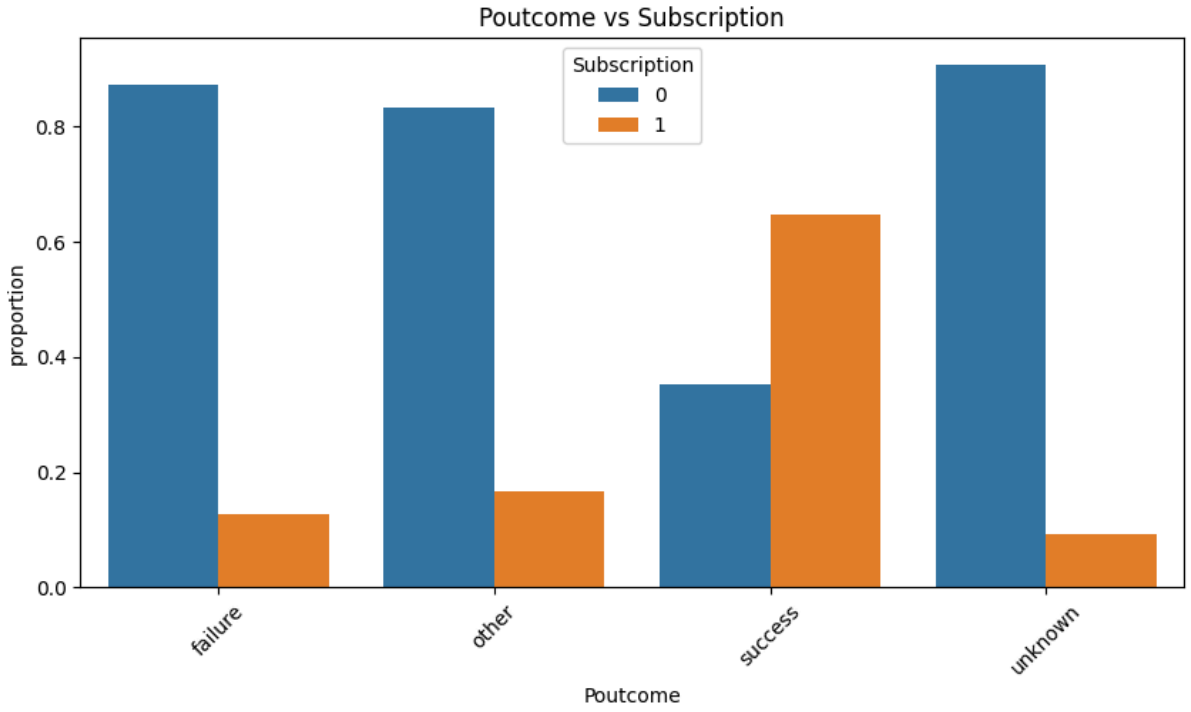




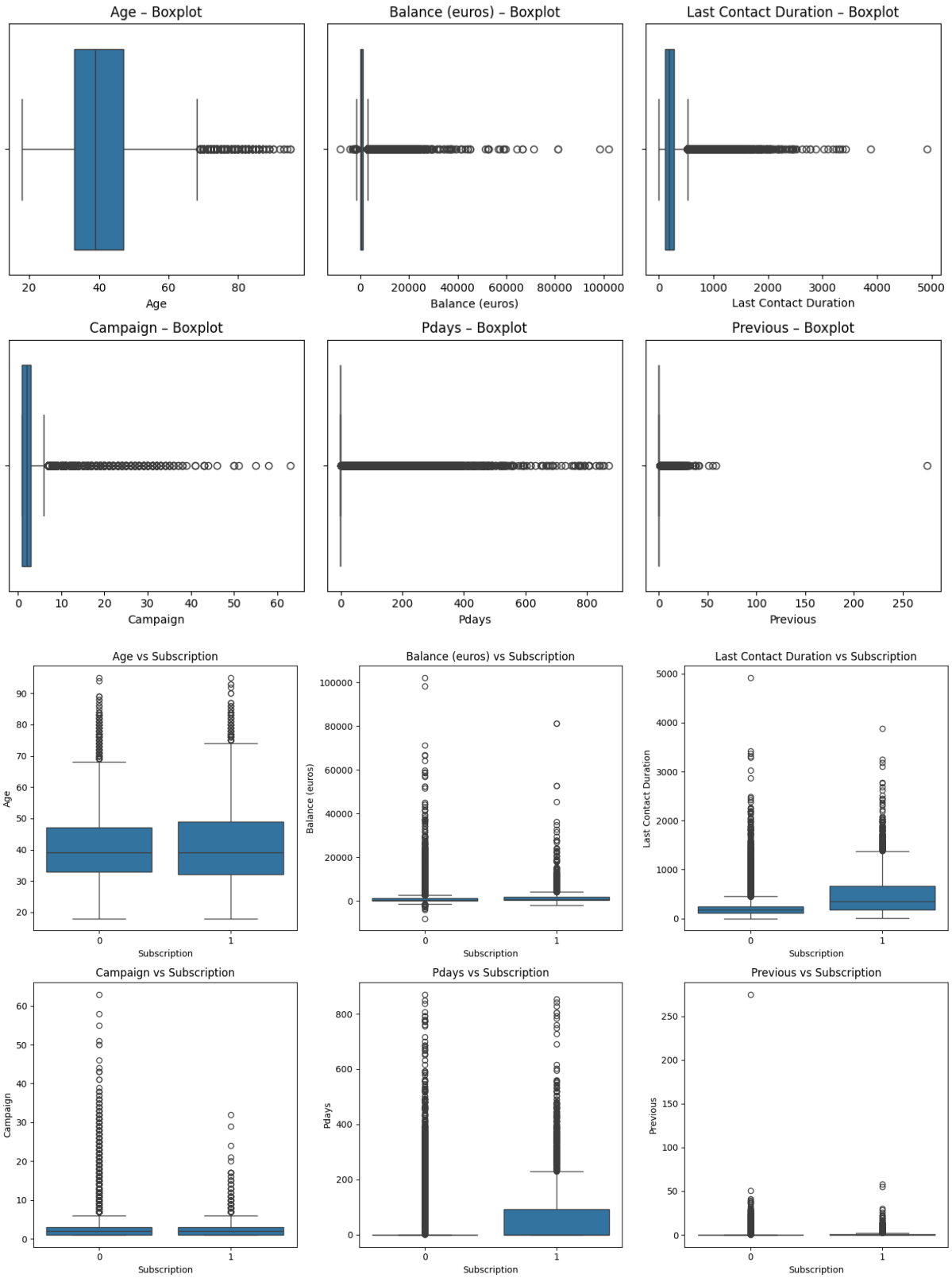
De los gráficos de barras se destaca que las personas que más suelen a la suscribirse son los estudiantes y retirados, a su vez, las personas con estudios terciarios tienen mayor proporción de suscripción



Aquellas personas sin un préstamo personal tienen mayor proporción en suscripciones, lo mismo para aquellos sin seguro para el hogar.



También, la Variable “Poutcome” tiene alto impacto para disparar la probabilidad de suscribirse.

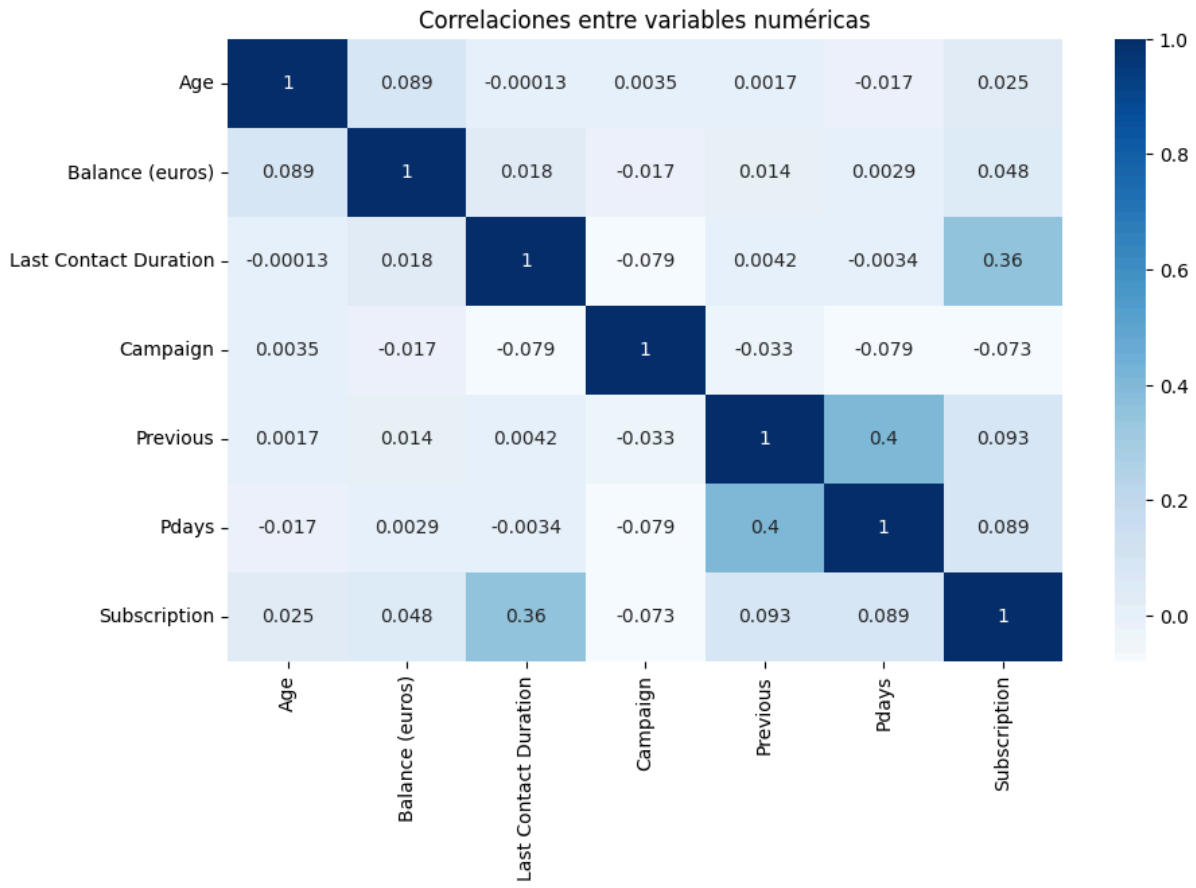


Observando los gráficos comparativos, en principio la variable más predictiva es la "Last Contact Duration" donde las llamadas de mayor duración aumentan la probabilidad de suscribirse. Sin embargo esta variable genera data leakage ya que se conoce después del contacto por lo que en campañas reales no debería usarla para predecir antes de llamar.

Se destaca que los clientes con Balance más alto tiende a suscribirse, pero tiene muchos outliers.
Los que fueron contactados recientemente, tienen mayor tasa de suscripción, pero la relación es ruidosa ya que hay muchos con valor -1 (Pdays). La variable Previous aporta poca información.
Cuanto más contactos le hicieron al cliente, puede haber una menor probabilidad de subscripción (Campaign)

9. Análisis de Correlaciones

Al calcular la correlación lineal entre las variables, se denota que la variable más correlacionada con la suscripción suele ser “duration”, pero tiene un problema de data leakage (se utiliza solo después de saber si aceptó). Igualmente, sirve como señal exploratoria a futuro.
los parámetros “Pdays” y “Previous”, a diferencia de los análisis anteriores, tienen una correlación más baja que lo esperado



10. Modelado, División de datos y Resultados

Con las observaciones de los pasos anteriores, pasamos a modelar el dataset para su aprendizaje, para esto importamos librerías de sklearn tanto para un último preprocessing y ajuste de variables dummies. Para este paso seteamos un tamaño de test data del 20% de los datos totales y seteamos un random state de valor 42.
Se decidió usar como modelos de clasificación 3 opciones:

a. Logistic Regression
b. Decision tree
c. Random forest

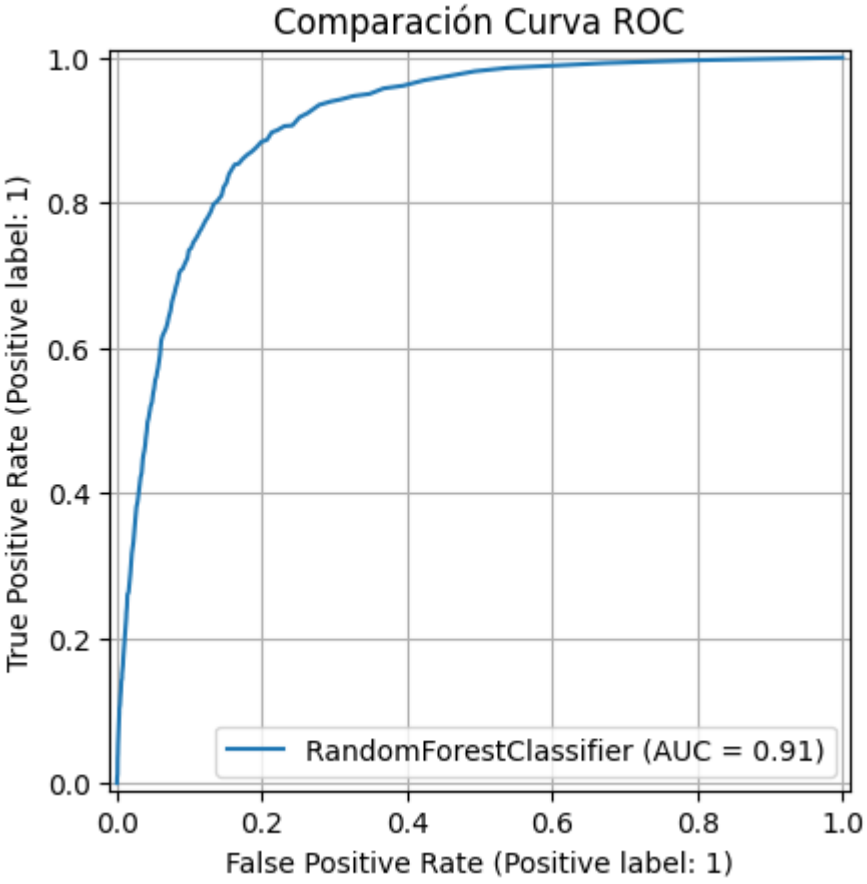
Ya entrenados estos modelos verificamos los resultados de los mismos para definir cuál es el modelo más preciso a través del gráfico de la curva ROC entre ellos y otras métricas como el F1 - Score y Recall guiándonos con las referencias del Artículo número 1(ver sección de referencias al final del documento).

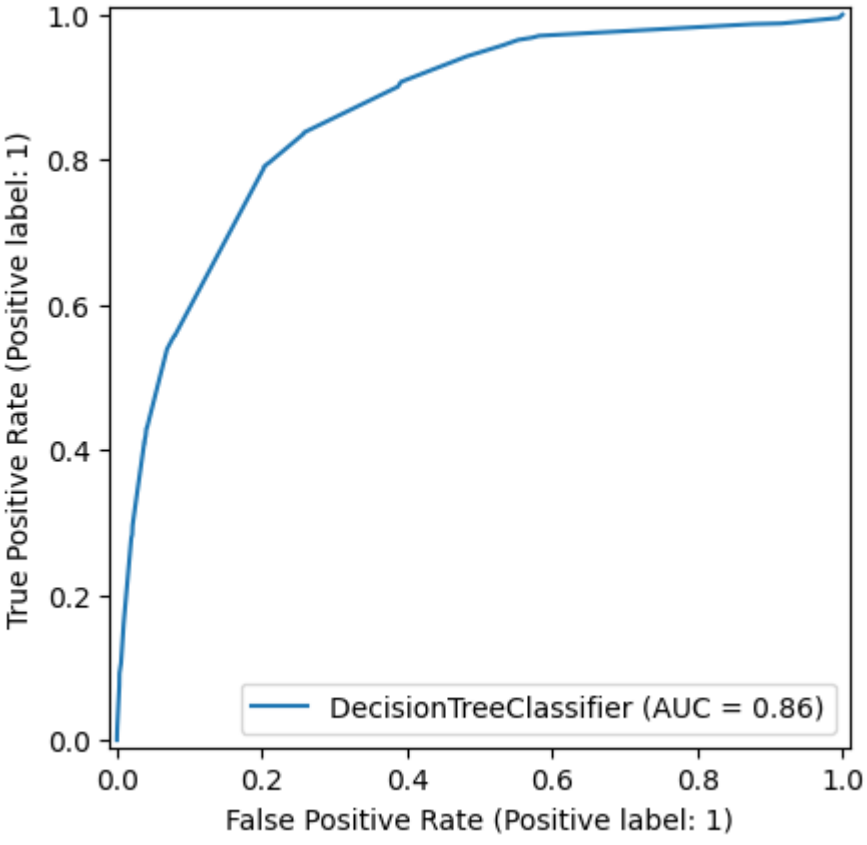
11. Resultados Modelado

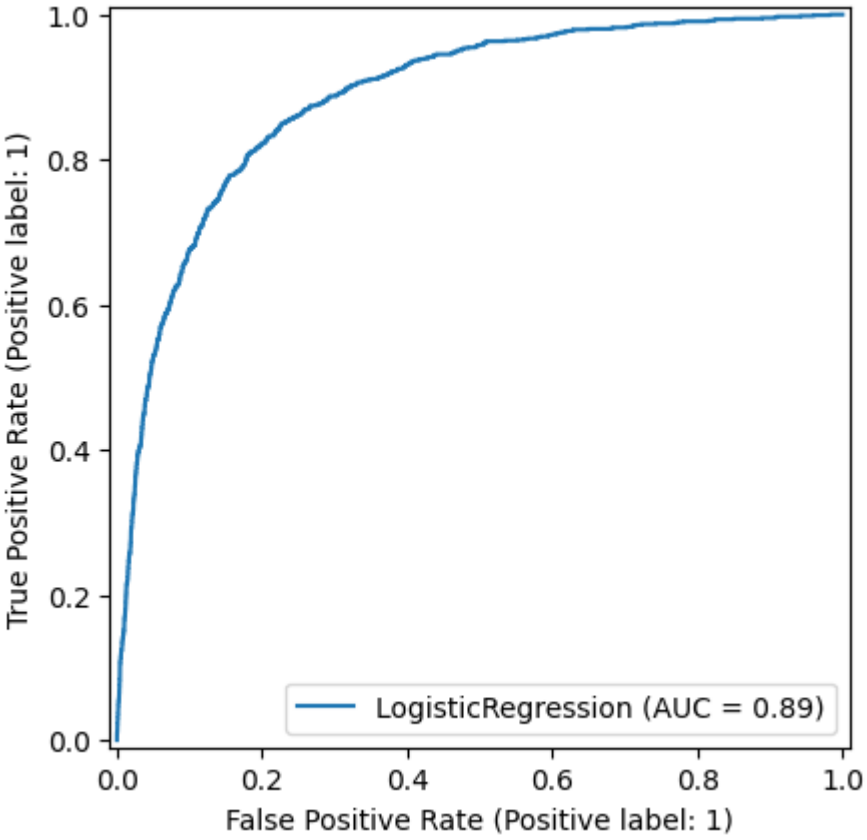
Logistic Regression:
AUC-ROC medio = 0.8836 | std = 0.0064
F1-score medio = 0.5034 | std = 0.0080
Recall medio = 0.7954 | std = 0.0130

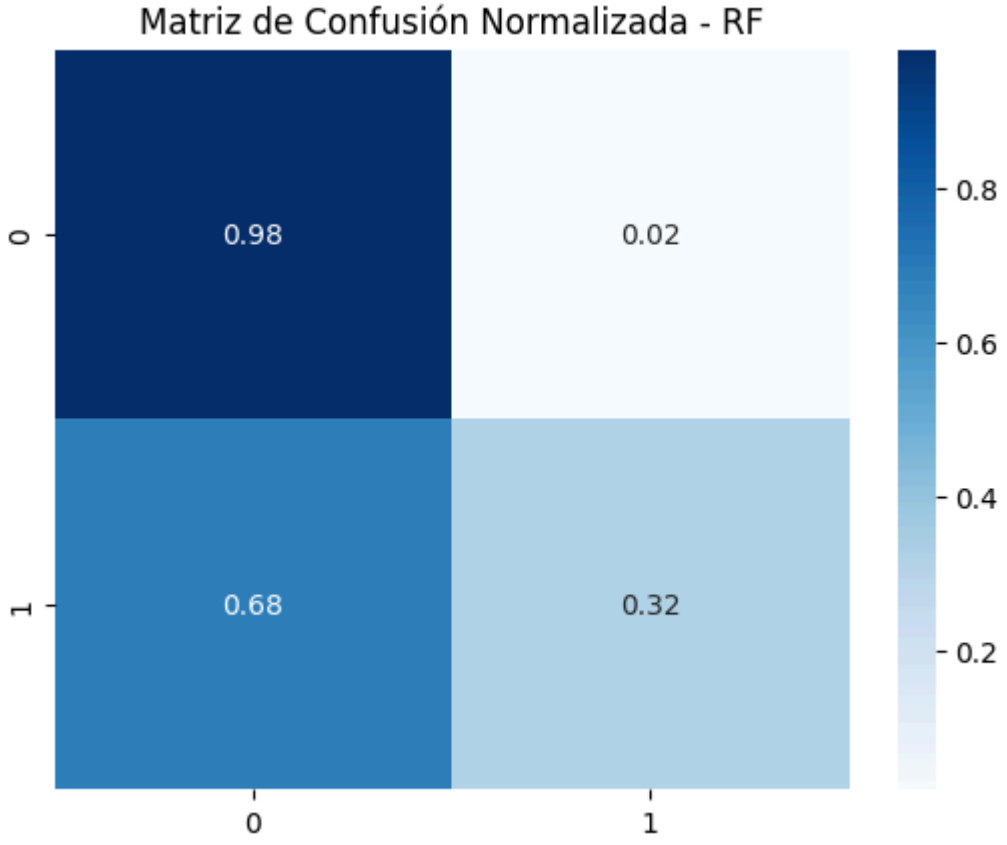
Decision Tree:
AUC-ROC medio = 0.8580 | std = 0.0059
F1-score medio = 0.4617 | std = 0.0062
Recall medio = 0.7879 | std = 0.0085

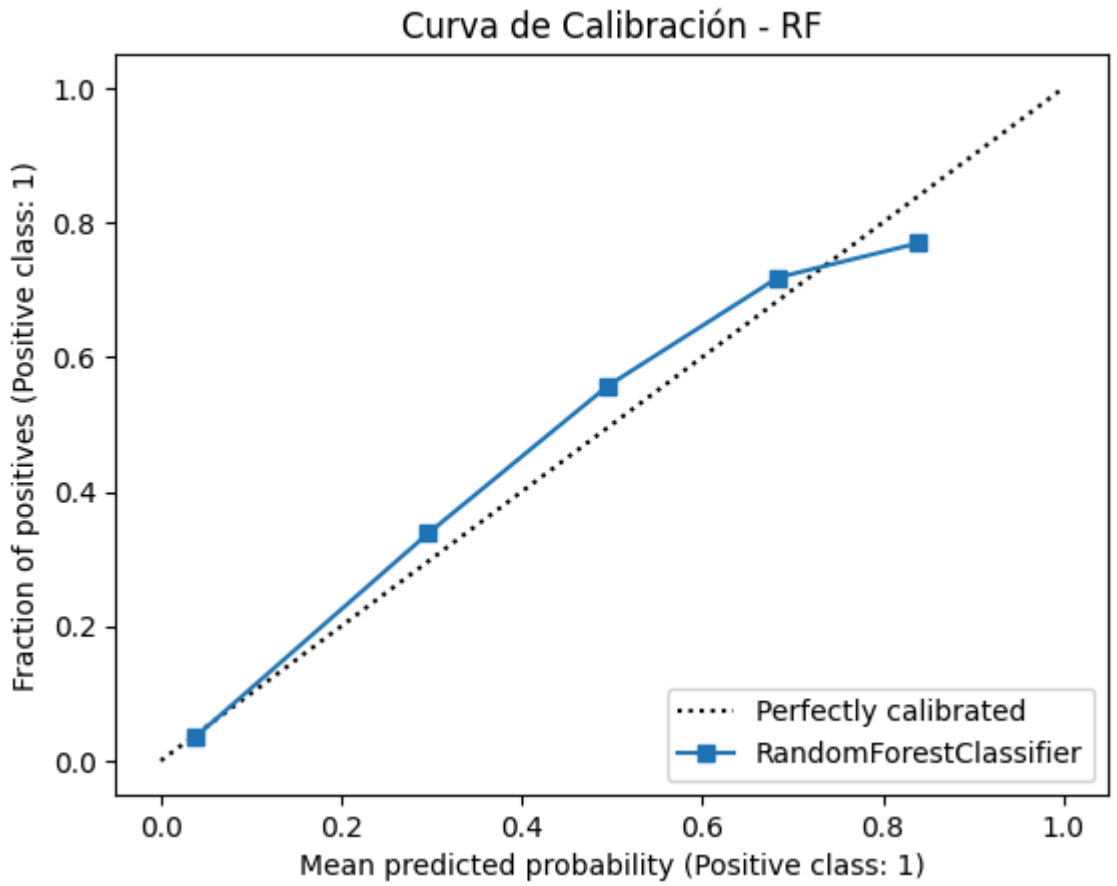
Random Forest:
AUC-ROC medio = 0.9081 | std = 0.0044
F1-score medio = 0.4306 | std = 0.0054
Recall medio = 0.3188 | std = 0.0057


Comparación Curva ROC
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
RandomForestClassifier (AUC = 0.91)
0.0
0.2
0.4
0.6
0.8
1.0


True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
DecisionTreeClassifier (AUC = 0.86)
0.0
0.2
0.4
0.6
0.8
1.0


True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
LogisticRegression (AUC = 0.89)
0.0
0.2
0.4
0.6
0.8
1.0

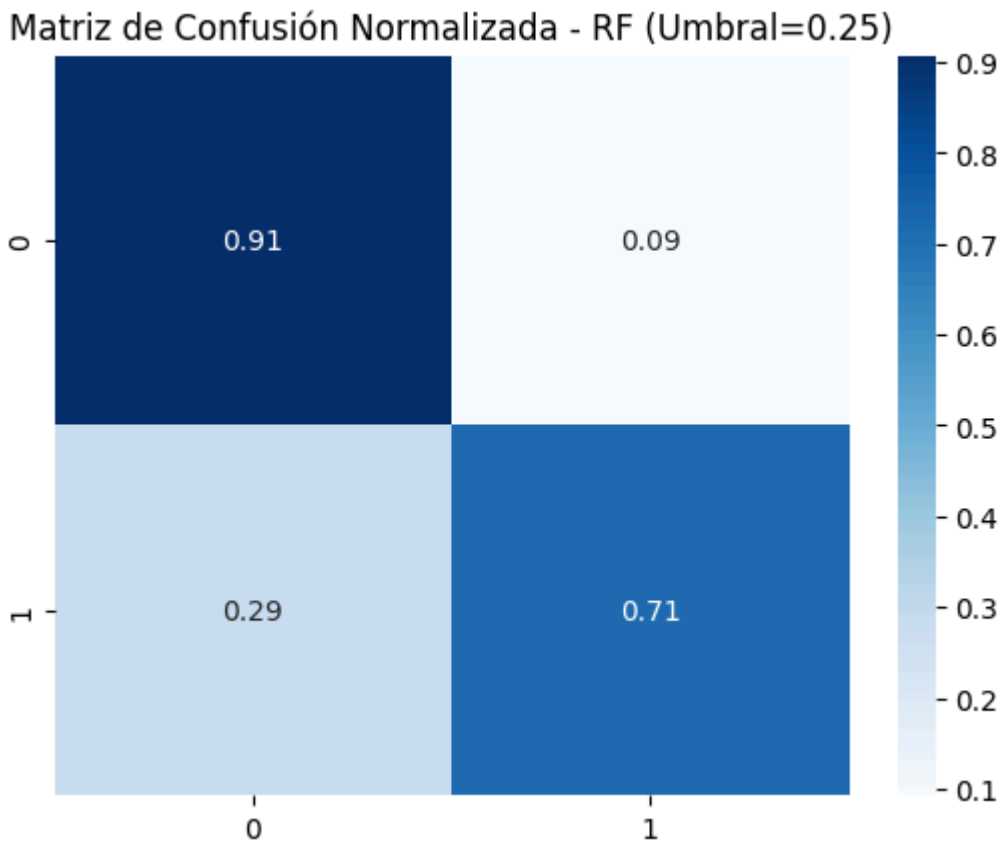

Matriz de Confusión Normalizada - RF
0.98
0.02
0.68
0.32
0
1
0.8
0.6
0.4
0.2



El mejor modelo fue el Random Forest ya que presenta el AUC- ROC promedio más alto y un desvío estándar bajo estable. Observando la matriz de confusión se puede concluir lo siguiente:

a) El 98% de los clientes que no se suscriben fueron detectados correctamente.
b) Solo el 2% de los clientes que no se suscriben fueron mal clasificados
c) El modelo está perdiendo el 68% de los clientes que sí se suscriben
d) El 32% de los que se suscriben fueron categorizados correctamente.

Por lo que el modelo predice muy bien la clase mayoritaria (0 - no suscriptos) pero falla en detectar la clase minoritaria (1- suscriptos) ya que el dataset es muy desbalanceado incluso aplicando class_weight= balanced
Si el banco se propone detectar aquellos clientes que se suscriben se debe aumentar el valor de recall de la clase 1, ya que un 68% de clientes no se detectan. Para ello ajustamos el umbral de decisión de 0.5 a 0,3 o 0,25. arriesgándose a obtener mayor número de falsos positivos (lo cual en términos de marketing es aceptable), siguiendo las recomendaciones del Artículo Número 3 anexado (ver sección de referencias al final del documento).

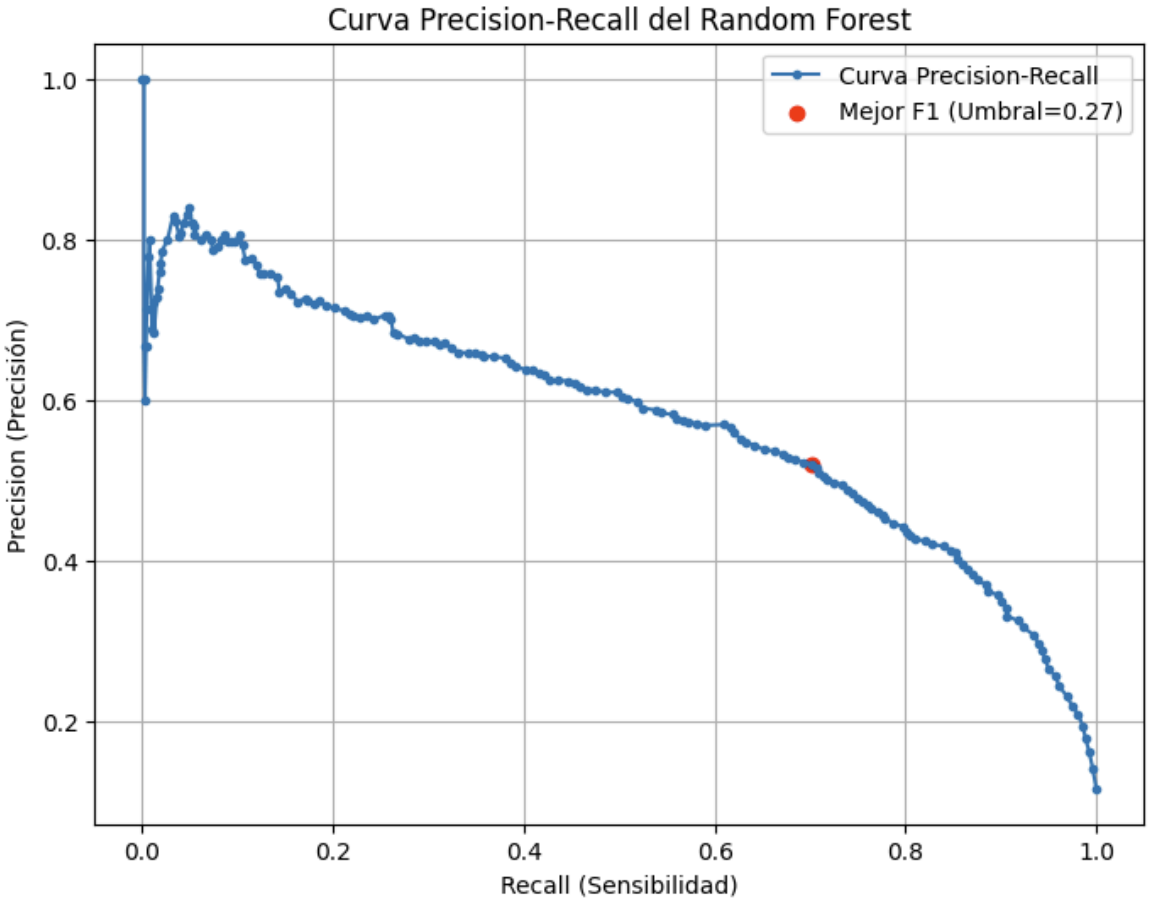


Con el ajuste del umbral de decisión de 0,5 a 0,25, se ve una mejora significativa, hemos mejorado la capacidad del modelo de detectar la clase suscriptores a costa de una menor precisión, lo cual suele ser aceptable en campañas de marketing donde la identificación de clientes potenciales es una prioridad.

El recall de suscriptores (clase 1) aumentó de 0,32 a 0,71 y el modelo ahora identifica mucho mejor a los suscriptores reales (casi el doble de la medición anterior).
La precisión de suscriptores (clase 1) baja de 0,67 a 0,54. Esta es una contrapartida esperada; al ser más agresivos en la identificación de suscriptores, también generamos más falsos positivos.
f1 score de suscriptores mejora de 0,43 a 0,59, lo que muestra un mejor equilibrio entre precisión y recordatorio para la clase minoritaria.
Se mantuvo la precisión general en rangos estables de 0,90.
AUC-ROC se mantiene en 0,9122, ya que el AUC-ROC es una métrica independiente del umbral de clasificación.

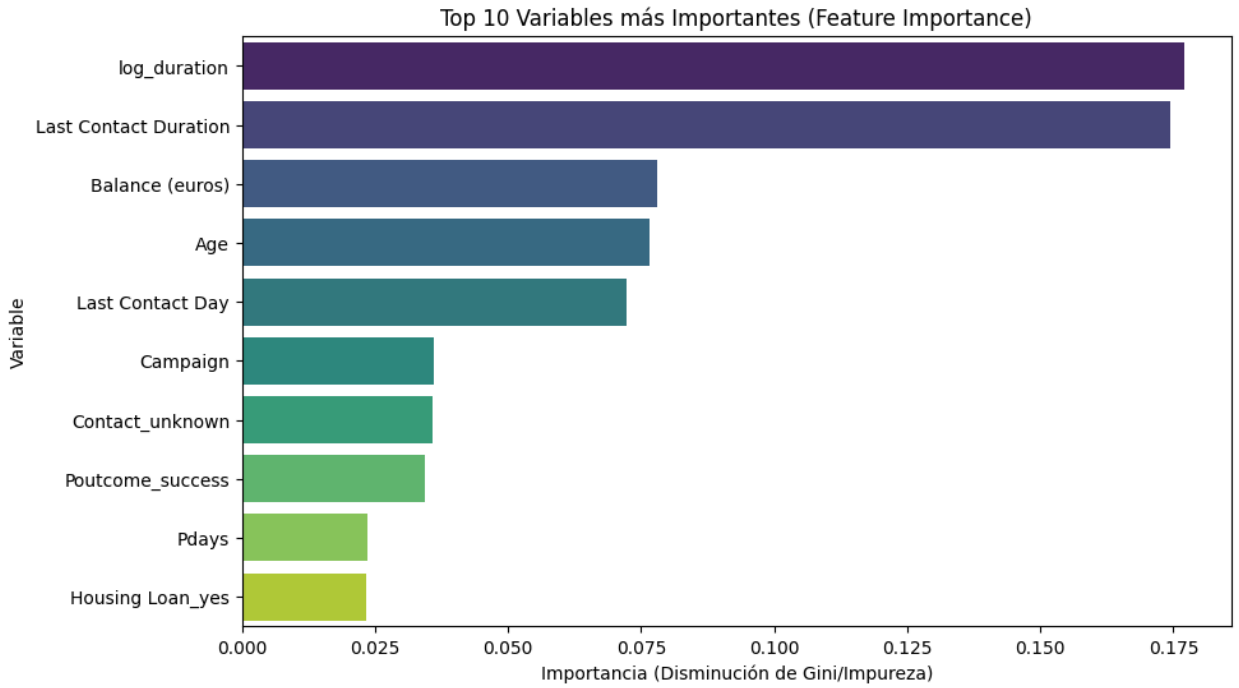
También se realizó el cálculo del threshold más óptimo, que maximice el F1 - score ya que es una métrica importante para modelos de clasificación binaria que proporciona el equilibrio entre entre la Precisión y el Recall o Sensibilidad.

```
Mejor Umbral (Max F1): 0.2650
F1-Score en ese umbral: 0.5980
Precisión en ese umbral: 0.5207
Recall en ese umbral: 0.7023
```



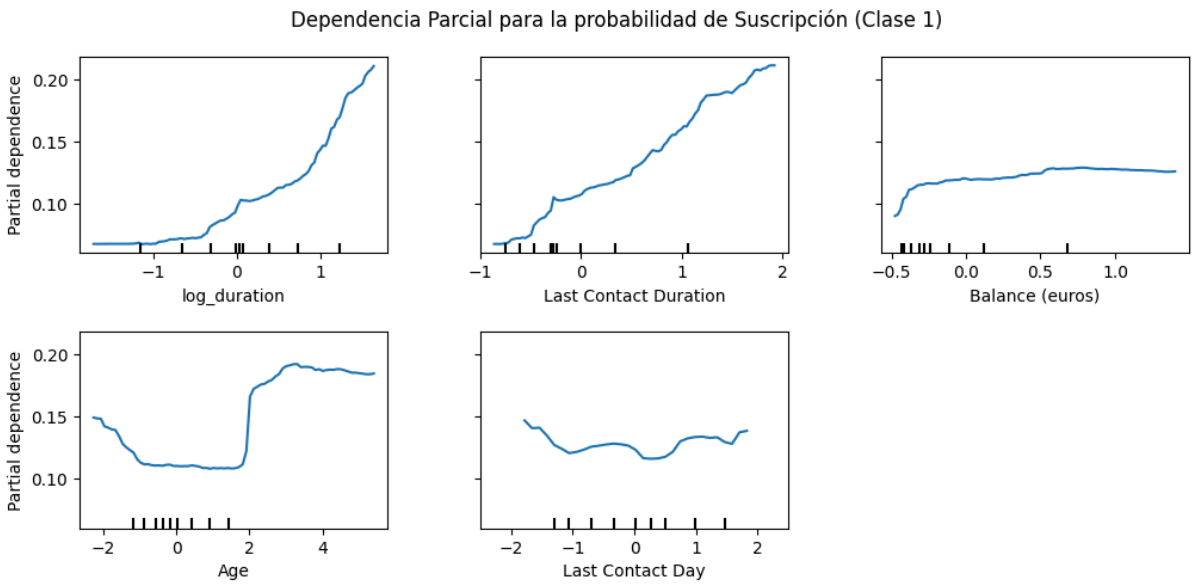
Este nos indica que el threshold más óptimo es 0,265, muy próximo al escogido el cuál es de 0,25. Y además las Métricas que arroja este threshold son muy similares al umbral de 0,25.

Finalmente se buscó obtener aquellas features con mayor peso para nuestro modelo de aprendizaje, es decir las más importantes para predecir si un cliente se suscribe o no a través del cálculo de Impureza. Esto es importante ya que es una forma útil para realizar una segmentación de clientes en un futuro análisis y reconocer aquellos clientes con mayor probabilidad de acceder a la suscripción.
A continuación se muestran las 10 variables más importantes para nuestro modelo:



La variable más importante para predecir o con mayor carácter post predictivo es “Last Contact Duration” ó su similar “log_duration” con el algoritmo aplicado para tener menor disparidad de valores. Cabe destacar que - como se mencionó anteriormente - esta variable puede generar data leakage ya que es información que puede no conocerse en el momento de la predicción, pero sí en el entrenamiento. Captura un efecto de la conversación, pero no es una característica del cliente. Sin embargo, es parte central del dataset y es muy útil para modelar y entender los factores asociados a la suscripción. Además la variable captura información relevante del comportamiento de un cliente durante la interacción, lo cual aporta valor analítico adicional.
Luego, con menor peso le siguen las variables Balance, Age y Last Contact Day

Ahora veremos el efecto de cada una de las variables a través del Partial Dependence Display para la probabilidad de suscripción.



Con el este gráfico, vemos que la dependencia en las variables de duración aumentan su peso en la decisión de suscripción a medida que aumenta sus valores, para el caso del balance, mantiene su peso si importar el valor en la mayoría de sus casos, para la edad, aumentan cuanto más cercano a los valores extremos es, y el último contacto se mantiene relativamente constante en los mismos valores

12. Conclusiones del Trabajo

Con la información obtenida, podemos concluir que aquellos clientes con mayor probabilidad de ser suscriptores a un producto ofrecido en una campaña de marketing del banco son aquellos que fueron contactados por una cantidad mayor de tiempo, el resto de los ítems más influyentes dentro de los parámetros dados son el rango de edad, el balance que tiene en su cuenta y la frecuencia de contacto con el cliente.

También notamos cómo varía la conclusión dependiendo de los métodos utilizados y el tipo de información que contiene cada feature, se demuestra cómo a medida que se utilizan métodos de análisis más complejos, las features que mayor correlación tenían en el EDA, no necesariamente demuestran lo mismo cuando se utilizan los modelos de machine learning.

13. Referencias

Artículo n°1: Guia de uso Curva AUC- ROC para Machine Learning
https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

Artículo n°2: Implementación de feature scaling para Machine Learning:
https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

Artículo n°3: Cómo manipular clases desbalanceadas en machine learning: https://www.geeksforgeeks.org/machine-learning/how-to-handle-imbalanced-classes-in-machine-learning/