

---

---

# Bank Marketing

[Link al repositorio del proyecto.](#)

---

# Objetivo

Los datos están relacionados con campañas de marketing directo (llamadas telefónicas) de una institución bancaria portuguesa. El objetivo de la clasificación es predecir si el cliente suscribirá o no a un depósito a plazo (variable  $y$ ).



# Índice

→ **Análisis exploratorio de datos:**

Visualización de datos e insights.

→ **Métricas de predicciones:**

Comparación de resultados entre modelos de clasificación.

→ **Próximos pasos:**

Siguientes posibles acciones para la mejora de las métricas..

# Análisis exploratorio de datos



## Obs:

Veremos las visualizaciones e insights **más destacados**, los cuales surgieron como resultado del EDA.

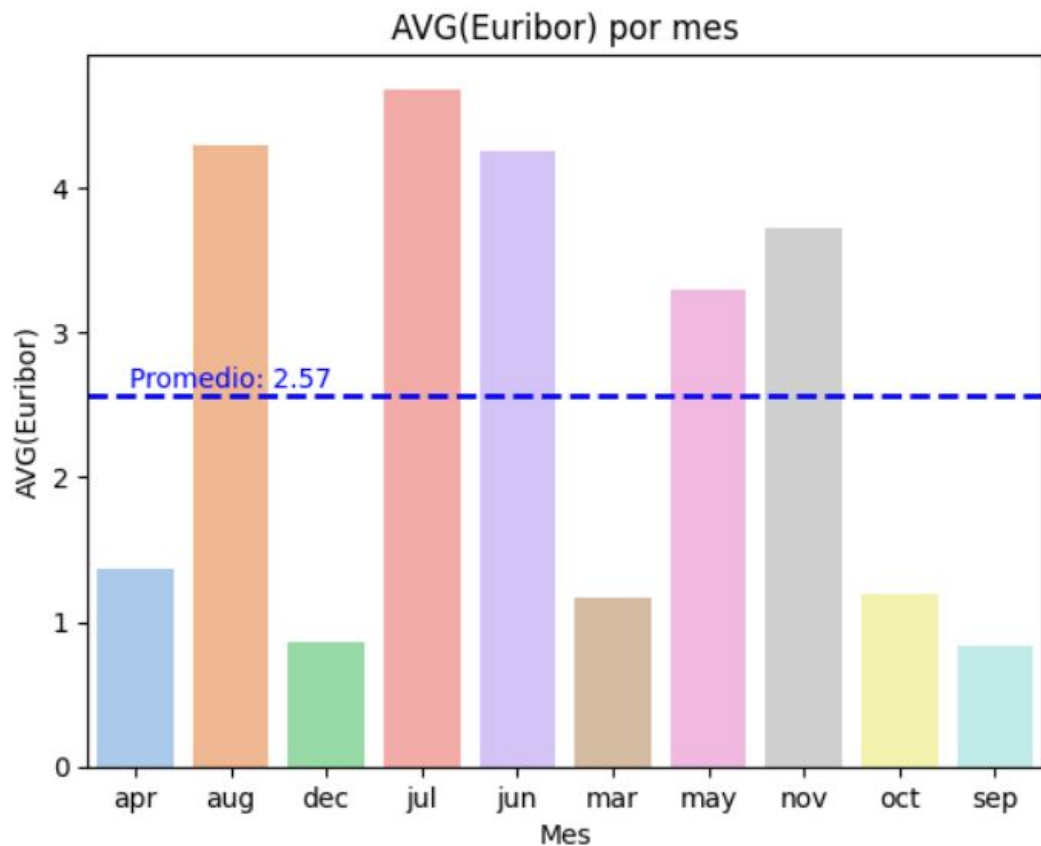
---

# EDA

A teal-colored speech bubble with a white border and a small tail pointing downwards and to the left.

Datos continuos

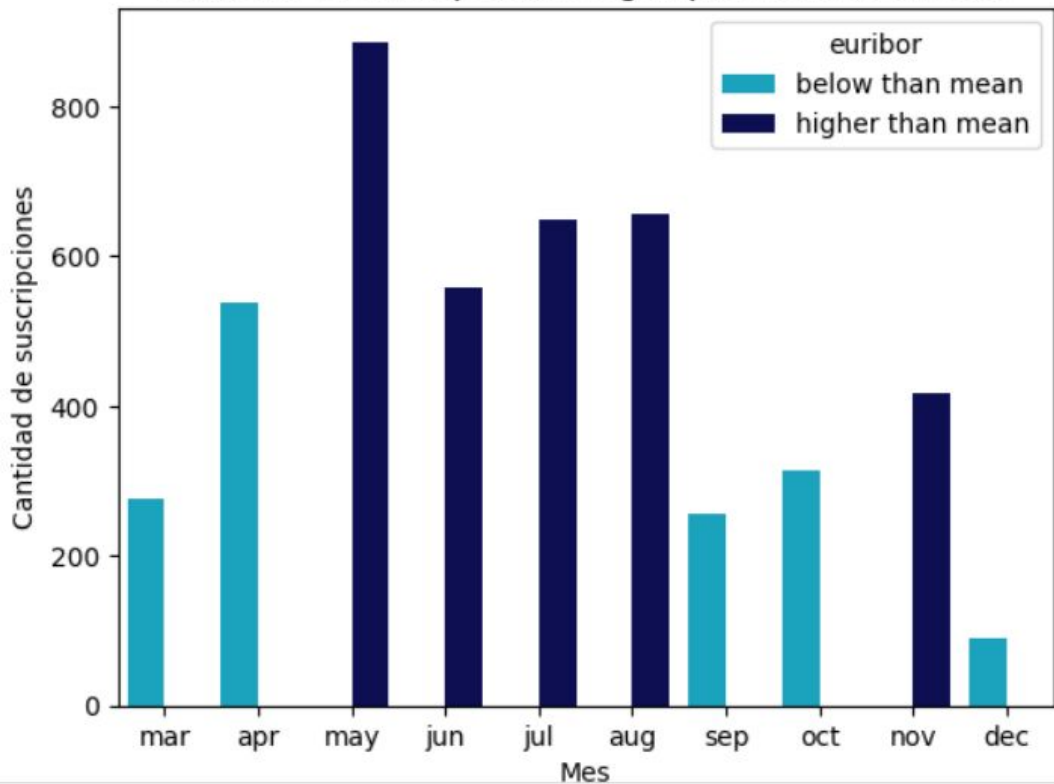
# Euribor3m



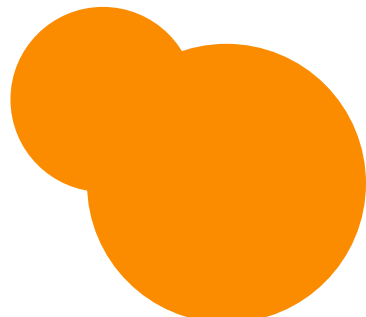
- ◆ Hay ciertos meses que poseen un promedio de “euribor3m” **menor** a la media y otros donde el promedio es **mayor** a la media.

# Cantidad suscripciones y promedio de euribor3m por mes

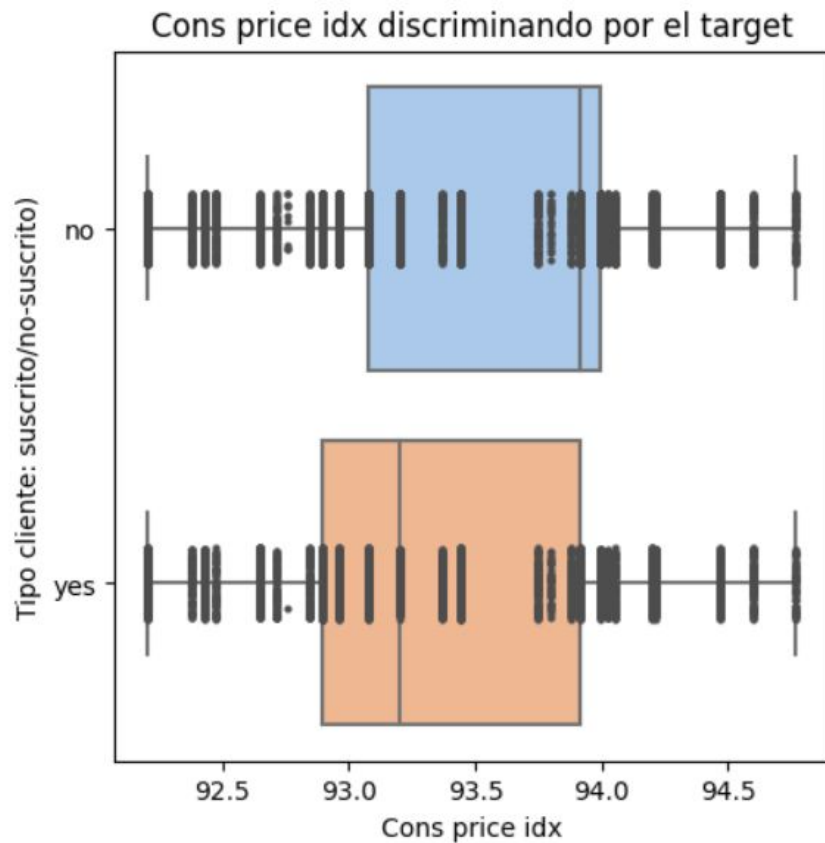
Cantidad de suscripciones segun promedio de euribor



- ◆ Los meses que tienen un euribor mayor a la media tuvieron más cantidad de suscripciones.
  - Tiene sentido pensar que al haber una tasa mayor de interes haya más cantidad de suscripciones.



# Cons price idx

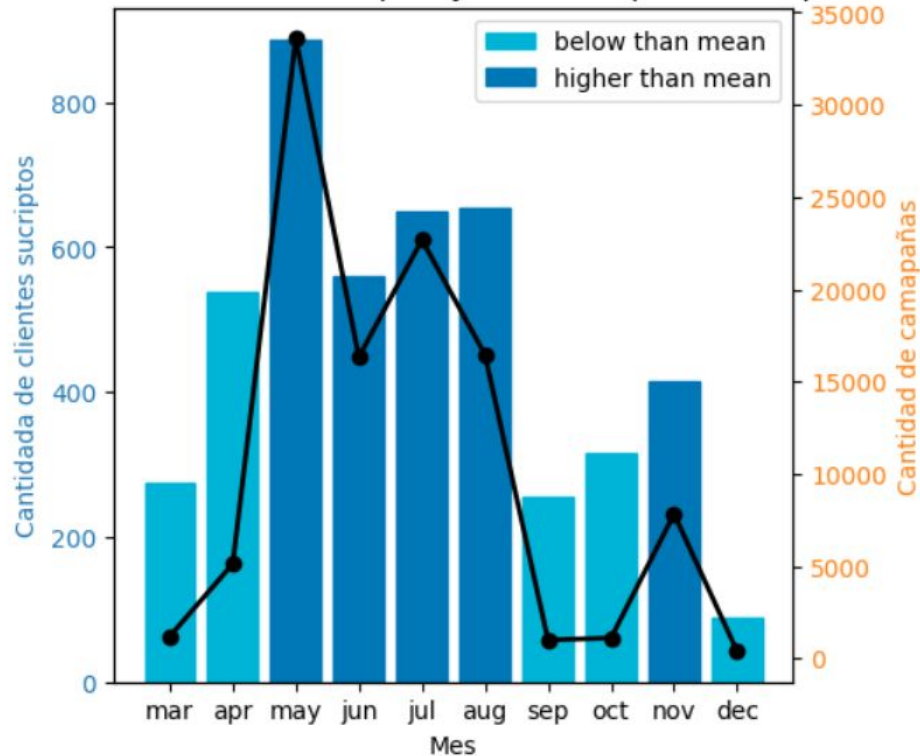


- ◆ Hay una diferencia en la media de dicha variable en clientes suscritos vs clientes no-suscritos.



# Cantidad de campañas discriminando según euribor3m

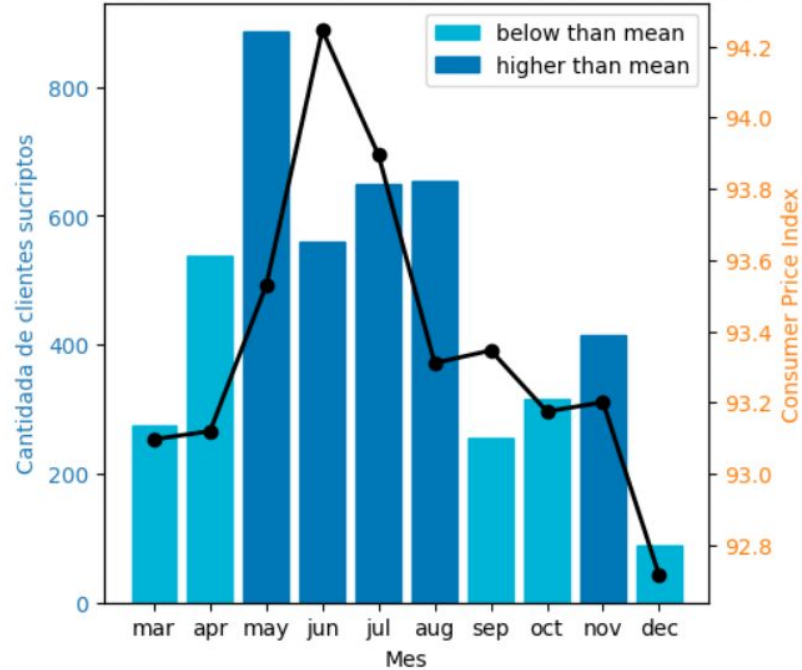
Cantidad de clientes suscriptos y consumer price index por mes



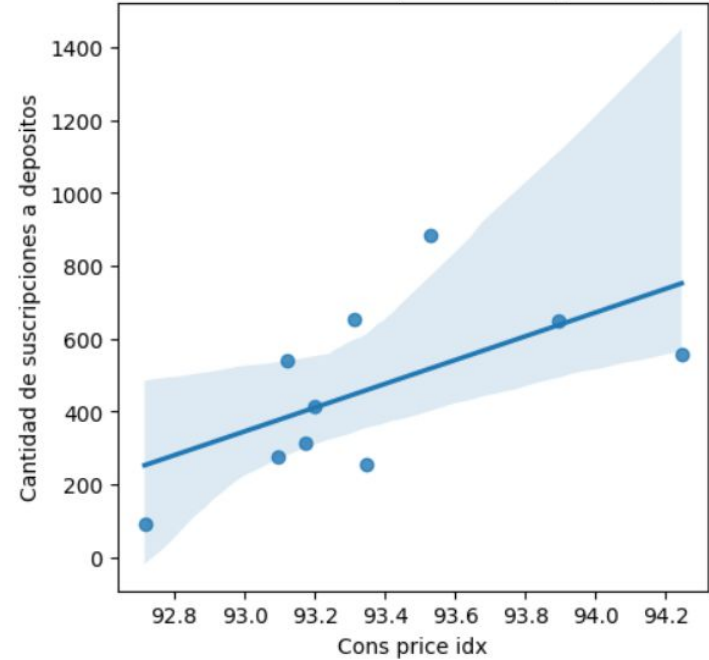
- ◆ Los meses que más clientes suscriptos hubo se correlaciona con la cantidad de campañas que se hicieron.

# Cons price idx discriminando por euribor3m

Cantidad de clientes suscritos y consumer price index por mes



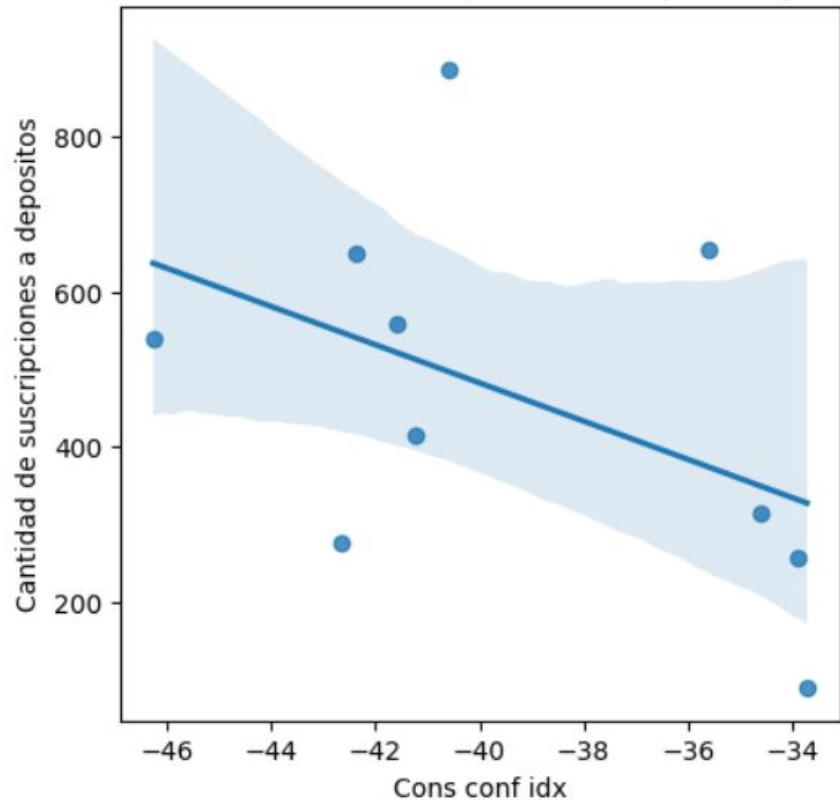
Relacion entre Cantidad de suscripciones a depositos y cons price idx



- ❖ Los meses con más suscripciones tuvieron un valor más alto de consumer price index, pareciera haber cierta relación lineal entre la cantidad de suscripciones y el valor de cons price idx.

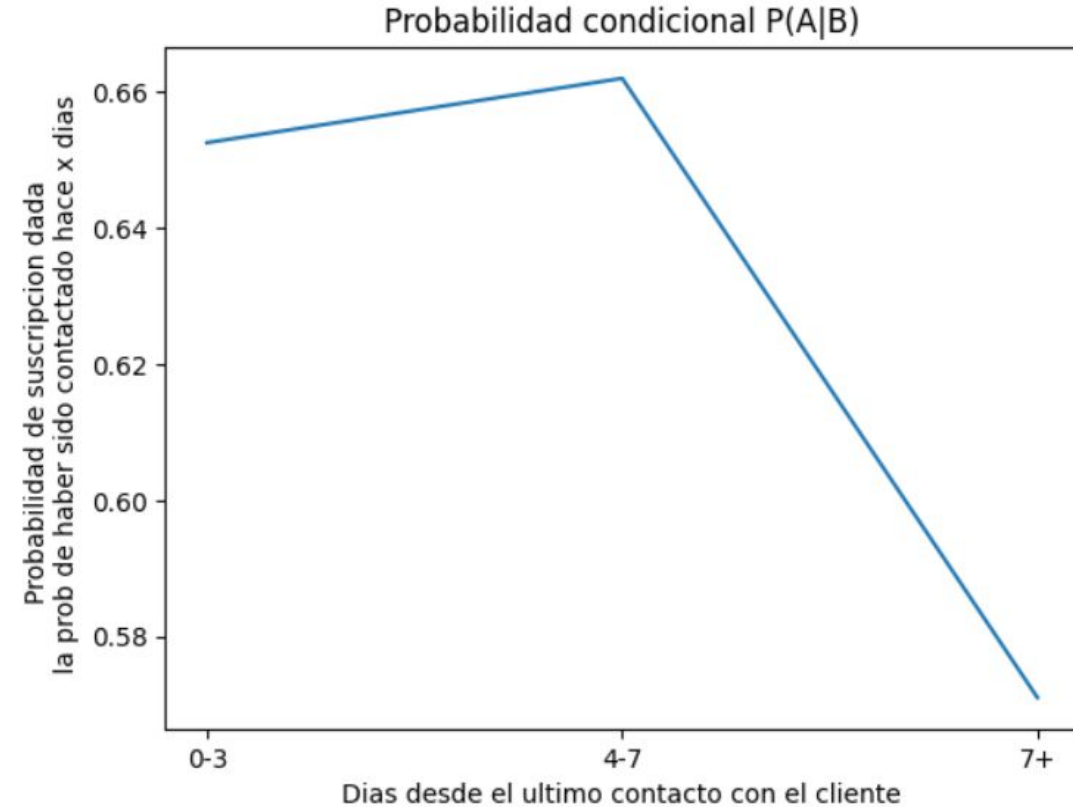
# Cons conf idx

Relacion entre Cantidad de suscripciones a depositos y cons conf idx



◆ Se aprecia cierta relación negativa entre cons conf idx y la cantidad de clientes suscriptos.

# Pdays

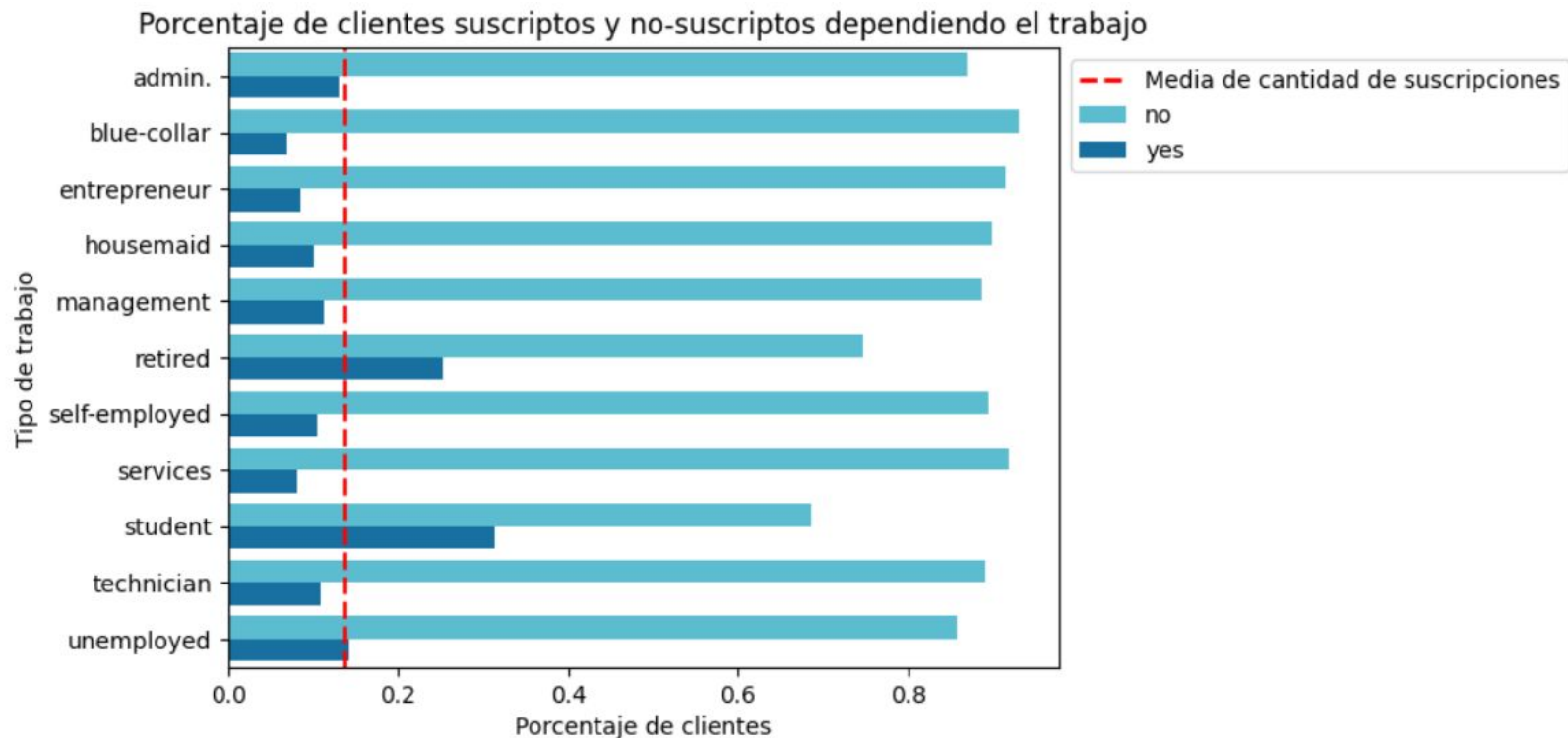


- ❖ La probabilidad condicional de suscripción es mayor cuando el cliente tuvo un contacto recientemente.

# EDA

Datos categóricos

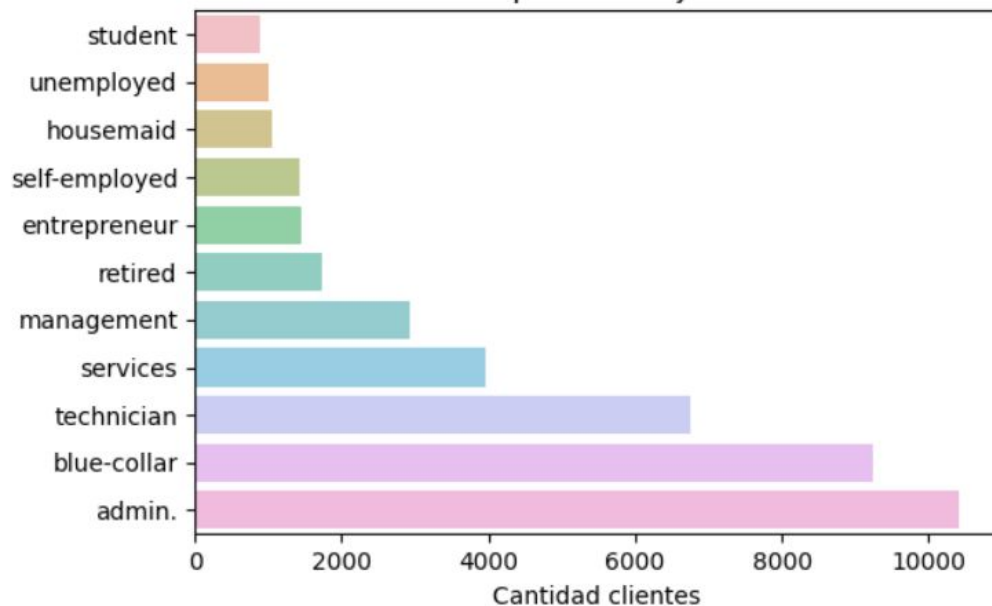
# Job y porcentaje de no-subscribers/subscribers



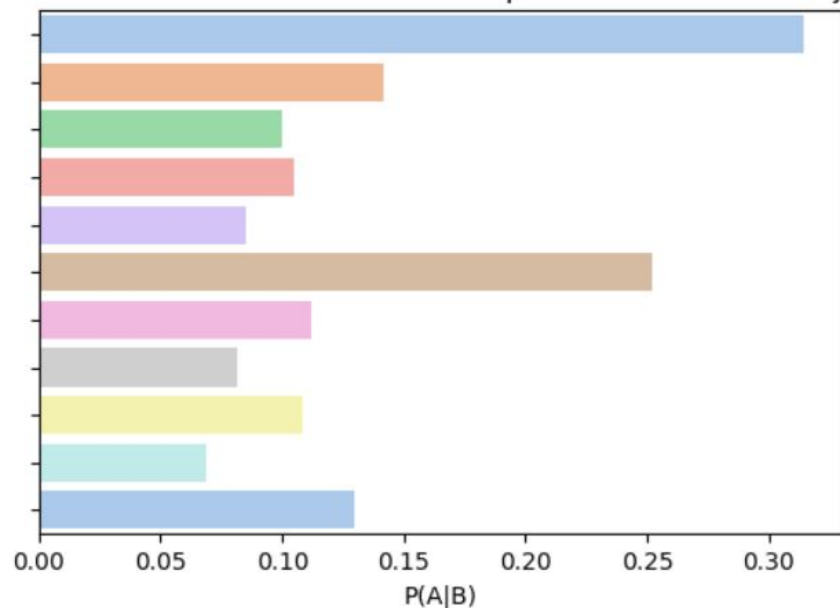
- ❖ Existen 2 labels que presentan una mayor cantidad de clientes suscritos (student y retired).

# Job y probabilidad condicional de suscripción

Distribucion del tipo de trabajos en los clientes



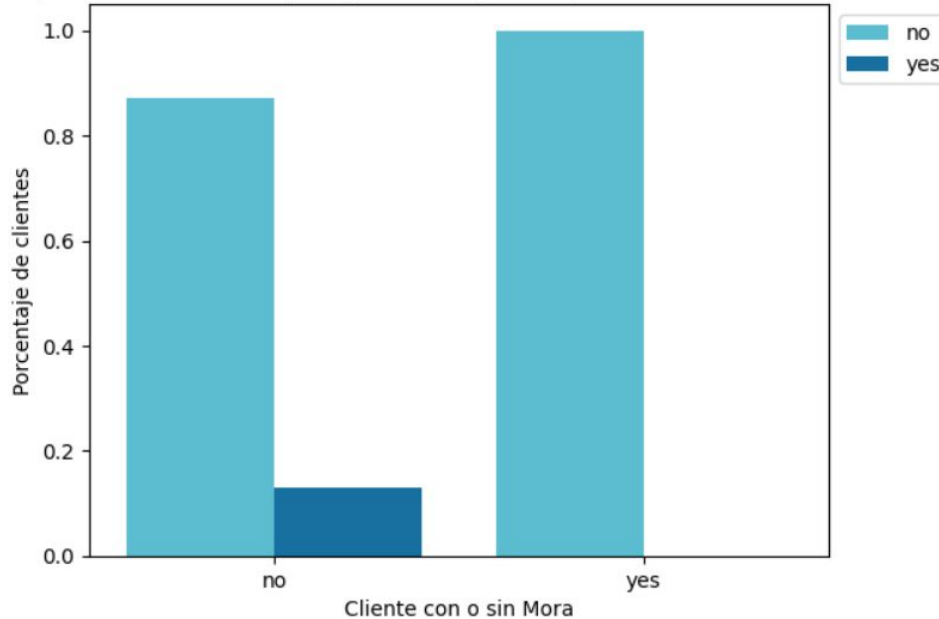
Probabilidad condicional de suscripcion dado dicho trabajo



- ❖ Pareciera ser que la probabilidad condicional de suscripción depende del trabajo del cliente.

# Default (clientes moroso vs no-moroso)

Porcentaje de clientes suscriptos y no-suscriptos dependiendo si el cliente es moroso

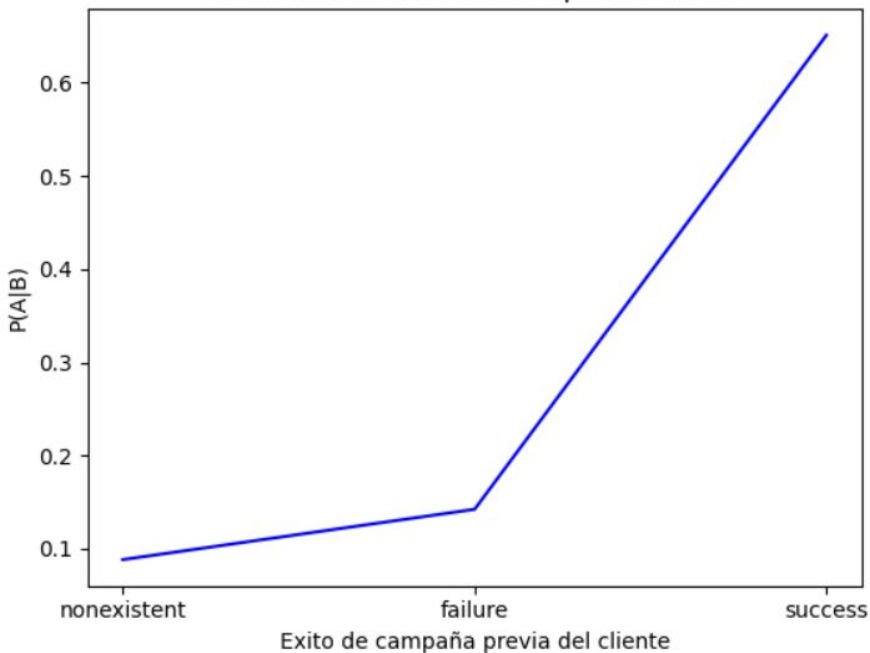


- ❖ Los clientes que presentan deudas **no se suscriben a los depósitos.**



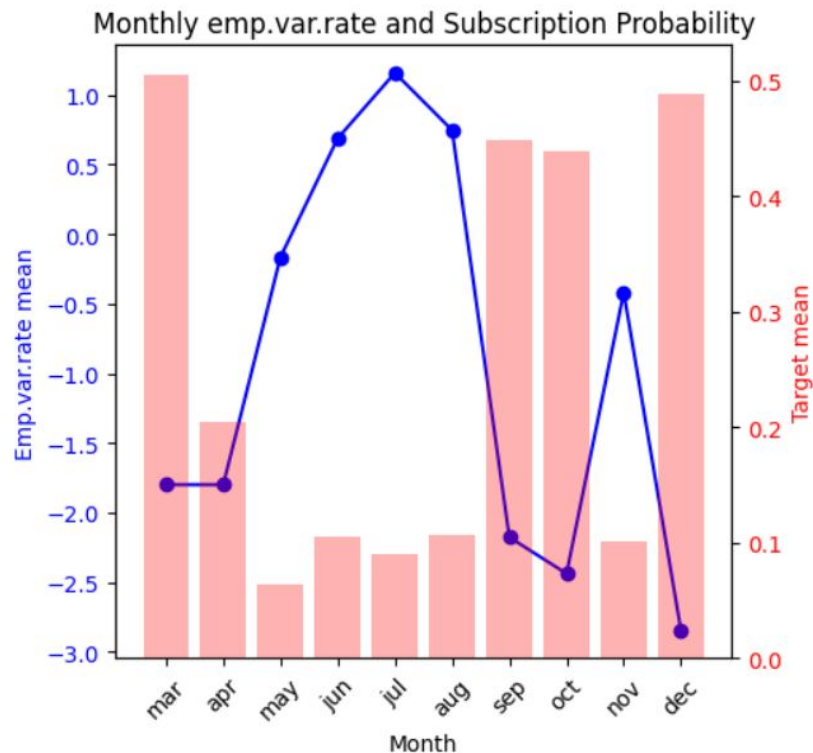
# Poutcome

Probabilidad condicional de suscribirse a deposito  
dada el resultado de camapaña anterior

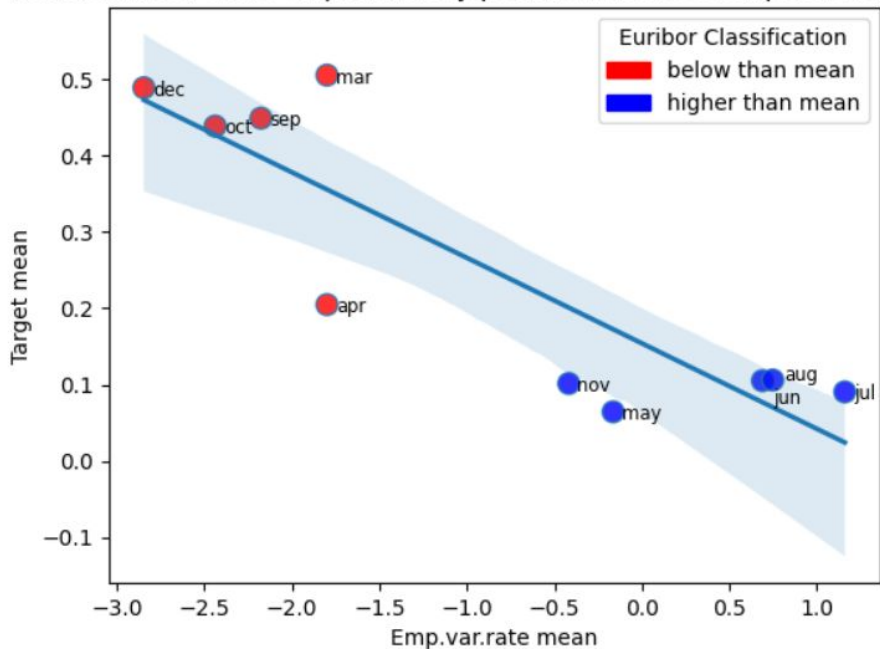


- ❖ La **probabilidad condicional** de suscripción **es** considerablemente **mayor dependiendo si el cliente se suscribio** a depósitos **en campañas anteriores.**

# Emp var rate



Relacion lineal entre emp.var.rate y probabilidad de suscripcion a deposito



- ❖ Vemos una relación lineal negativa entre la cantidad de suscripciones y el valor promedio por mes de emp var rate.

# Modelos de clasificación



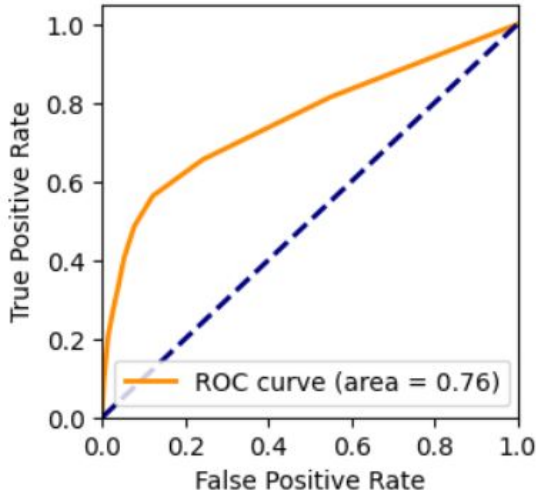
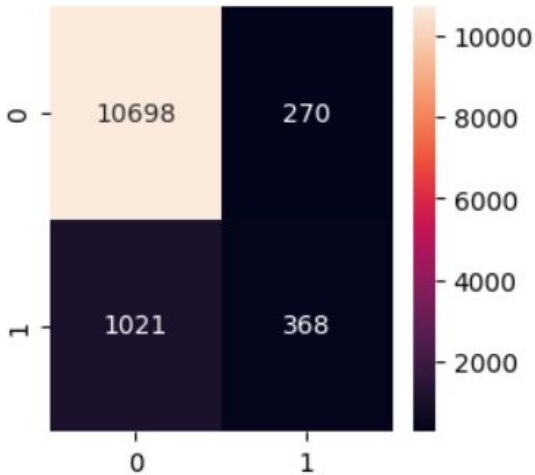
## Obs:

Veremos los distintos  
modelos de clasificación  
conjunto sus métricas.

# KNN.

	precision	recall	f1-score	support
0	0.91	0.98	0.94	10968
1	0.58	0.26	0.36	1389
accuracy			0.90	12357
macro avg	0.74	0.62	0.65	12357
weighted avg	0.88	0.90	0.88	12357

Precisión (KNeighbors): 0.8955248037549567

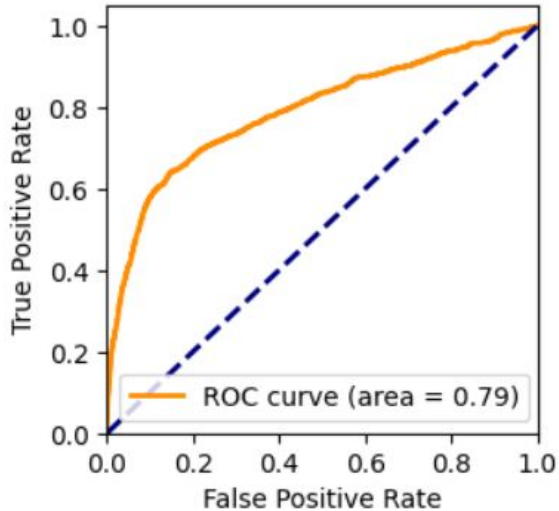
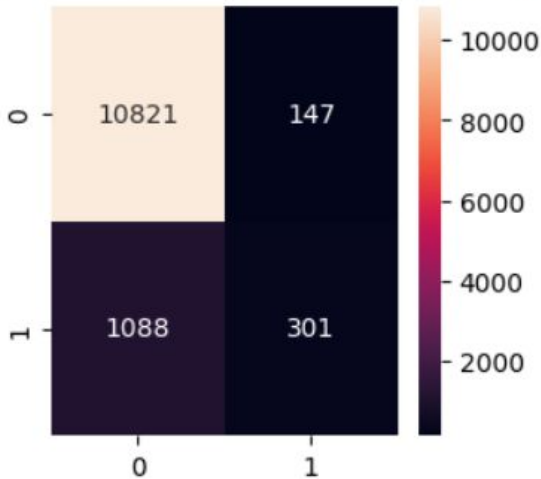


❖ Vemos muy buenas métricas a la hora de predecir los clientes que **NO** se suscriben a los depósitos.

# RF.

	precision	recall	f1-score	support
0	0.91	0.99	0.95	10968
1	0.67	0.22	0.33	1389
accuracy			0.90	12357
macro avg	0.79	0.60	0.64	12357
weighted avg	0.88	0.90	0.88	12357

Precisión (KNeighbors): 0.9000566480537348

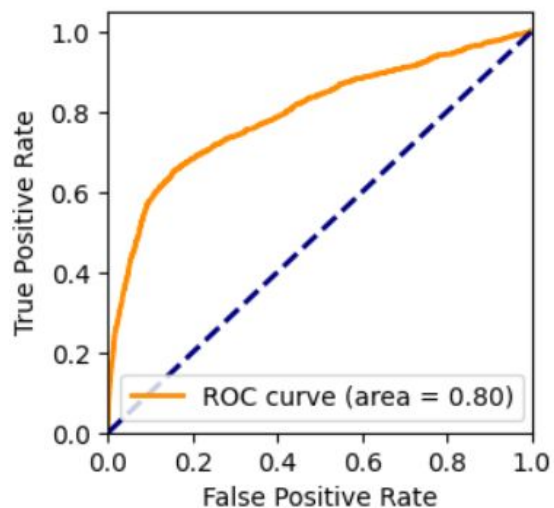
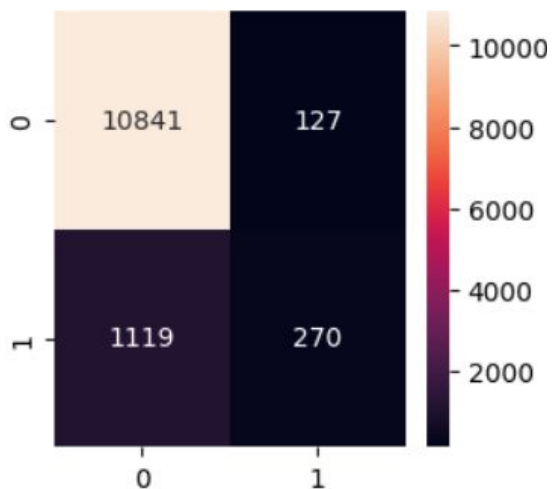


❖ Con random forest parece mejorar la métrica ROC AUC con respecto a KNN.

# XGB.

	precision	recall	f1-score	support
0	0.91	0.99	0.95	10968
1	0.68	0.19	0.30	1389
accuracy			0.90	12357
macro avg	0.79	0.59	0.62	12357
weighted avg	0.88	0.90	0.87	12357

Precisión (KNeighbors): 0.899166464352189





## ● Sobre sampleo

Ninguno de los 3 modelos de clasificación logró predecir (con buenos resultados) los clientes suscriptos a depósitos.

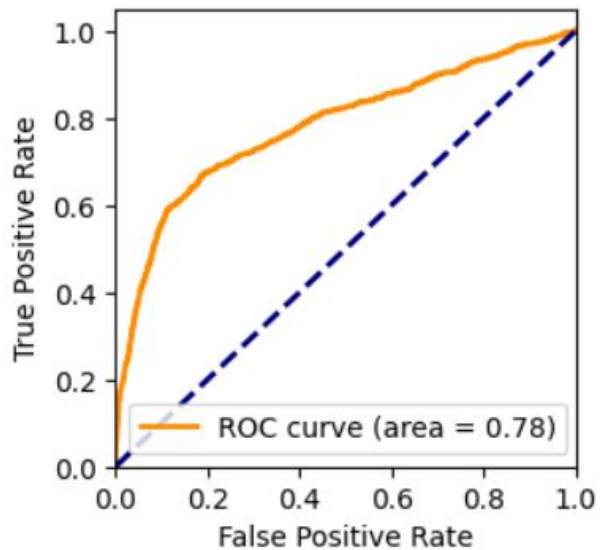
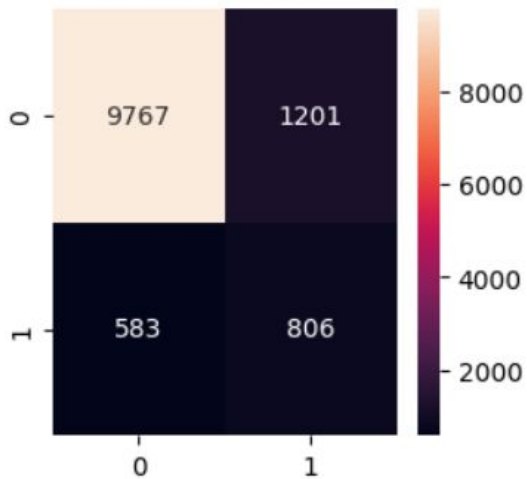
Esto puede deberse a que **hay considerablemente más clientes no-suscriptos (90%) vs 10% de clientes suscriptos**, por ende nuestros modelos aprenden muy bien a predecir la clase mayoritaria.

Optamos por hacer un **sobre-sampleo en la clase minoritaria** y luego obtener nuevas métricas.

# RF.

	precision	recall	f1-score	support
0	0.94	0.89	0.92	10968
1	0.40	0.58	0.47	1389
accuracy			0.86	12357
macro avg	0.67	0.74	0.70	12357
weighted avg	0.88	0.86	0.87	12357

Precisión (KNeighbors): 0.8556283887675002

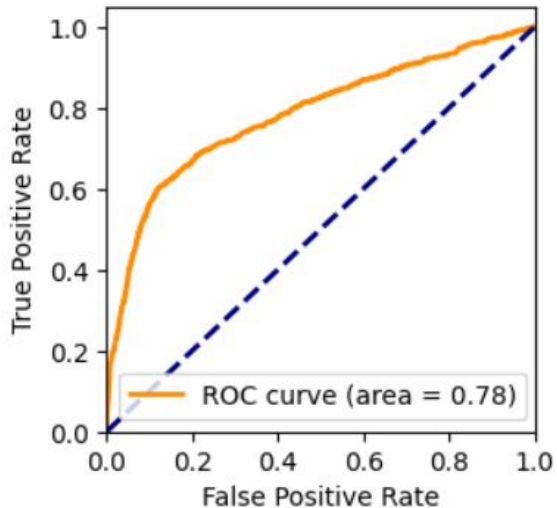
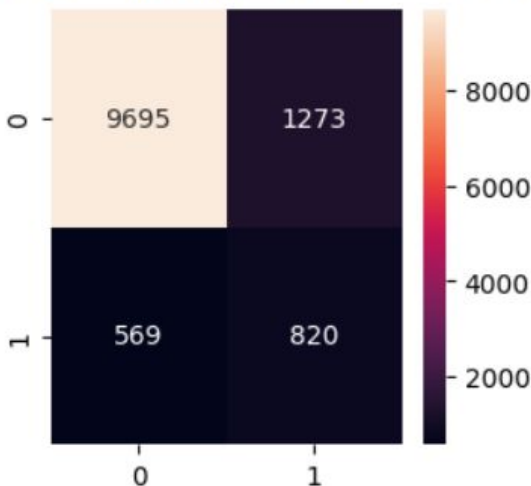






	precision	recall	f1-score	support
0	0.94	0.88	0.91	10968
1	0.39	0.59	0.47	1389
accuracy			0.85	12357
macro avg	0.67	0.74	0.69	12357
weighted avg	0.88	0.85	0.86	12357

Precisión (KNeighbors): 0.8509346928866229



## — Resumen

El modelo analizado **es eficaz para predecir clientes que no se suscriben a depósitos**, con una **precisión de 0.91 y un recall de 0.99**. Sin embargo, su **rendimiento disminuye** al predecir **clientes que sí se suscriben**, posiblemente debido a que solo el 10% de las observaciones corresponde a clientes suscriptos.

Para mejorar esto, se realizó un sobre-muestreo en la clase minoritaria (clientes suscriptos), lo que aumenta el recall pero reduce la precisión.

**Dependiendo de los objetivos de la empresa, el modelo puede ser ajustado:**

- Para campañas de marketing dirigidas a clientes con baja probabilidad de suscripción, los modelos de Random Forest y XGBoost, sin sobre-muestreo, son muy efectivos.
- En cambio, para retener clientes con alta probabilidad de suscribirse, los mismos modelos con sobre-muestreo muestran resultados prometedores. Ajustar el umbral de decisión puede aumentar aún más el número de clientes suscriptos identificados (mejorando el recall) a costa de reducir la precisión.

# Próximos **pasos**

## **Analizar FN**

Analizar los clientes que fueron erróneamente predichos como falsos (no-sub) en busca de algun patron que lleve a entender el motivo de que nuestro modelo cometa dicho error.

## **Analizar FP**

Similarmente, analizar los clientes que fueron erróneamente predichos como verdaderos (sub) en busca de algun patron que lleve a entender el motivo de que nuestro modelo cometa dicho error.

## **Incorporación de nuevos features.**

Implementar más técnicas de ingeniería de características con el objetivo de generar nuevos features que aporten información de clientes suscriptos.