Mentoría: Data Science y Marketing

Predicción de clientes propensos a adquirir un plazo fijo

Grupo 01
Santiago Seppi
Agustín Carchano
Florencia Cámara

Diplomatura en Ciencia de Datos, Aprendizaje Automático y sus Aplicaciones - 2021

Composición de la Base de datos

- Campaña de marketing desarrollada por un banco de Portugal.
- Consistió en llamadas telefónicas a los clientes del banco para que adquieran un depósito de plazo fijo (PF).
- Base de datos disponible en este link. Utilizamos la base denominada: *bank-additional-full.csv*.

Composición de la base de datos:

- **20 variables** más la variable objetivo ***"y"*** que indica si la persona contrató el PF.
- **41.188 observaciones**.
- Las observaciones van de Mayo 2008 a Noviembre 2010, la base está ordenada según lo que se indica aquí.
- Diccionario de la base de datos en este link.

Nota: algunas llamadas son hacia el mismo cliente pero en otro momento.

Variables

Sobre el cliente

Información **personal**:
- edad
- trabajo
- estado civil
- educación

Información **crediticia**:
- tiene créditos en default
- tiene crédito personal
- tiene crédito hipotecario

Sobre campañas de marketing

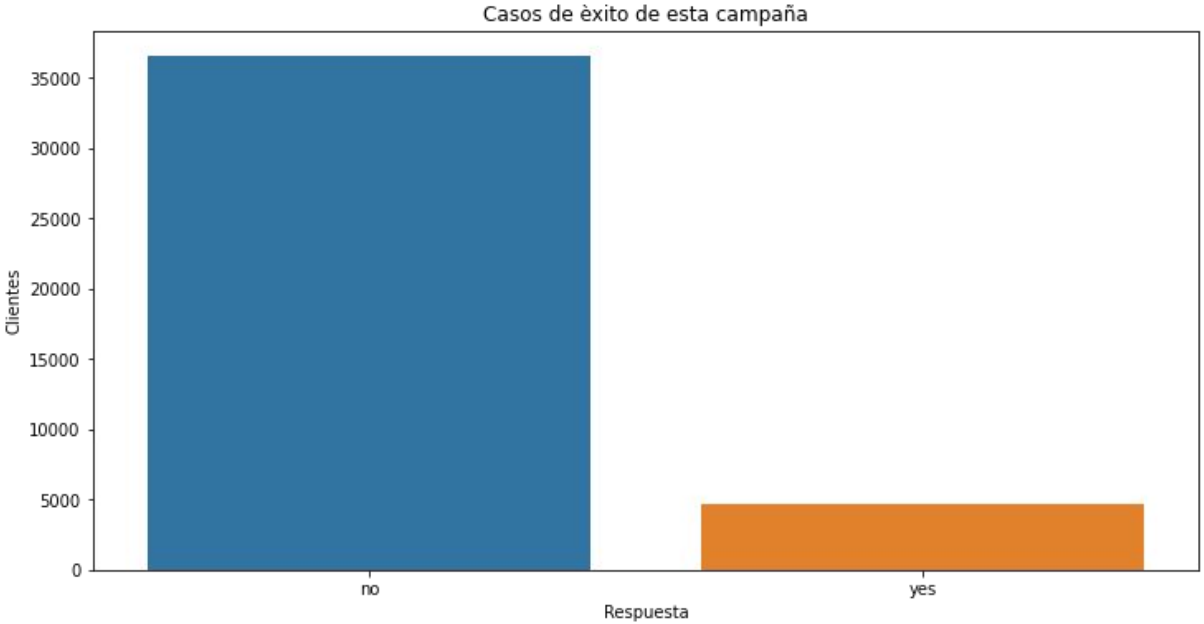
Sobre la **campaña actual**:
- medio de contacto
- mes
- día de la semana
- duración de la última llamada
- cant. veces contactado

Sobre **campañas previas**:
- cant. días desde últ. contacto
- cant. de contactos previos
- resultado campañas previas

Variable objetivo (*y*):
Contrató el PF, si o no.

Contexto
tasa de interés
índice de precios al cons.
índice de confianza cons.
var. del empleo
cant. de empleados

Análisis y visualización de datos



El **11,3% de los clientes** contactados durante la campaña de marketing **contrataron el PF** luego de ser contactados (4.640 clientes).

CASOS DE ÉXITO

Solamente el **13,67%** de los clientes alcanzados habían sido **contactados previamente** en el marco de campañas de marketing (5.695 clientes).

Análisis exploratorio de datos

Análisis descriptivo:

- Análisis y visualización de los datos: distribuciones de las variables, comparación de distribuciones de las variables entre clientes que contrataron y los que no contrataron el PF.

Análisis exploratorio permitió identificar:

- Casos nulos en las variables categóricas: categorías “desconocido”
- Valores outliers en las variables *duration* (duración de la llamada) y *campaign* (cantidad de veces que fue contactado durante la campaña)
- Inconsistencia de la variable *pdays* (cantidad de días desde el último llamado)
- Correlaciones superiores a 0.65 entre las variables de contexto (cantidad de empleados, variación de la tasa de empleo e índice de precios al consumidor)
- Falta de información respecto de la fecha

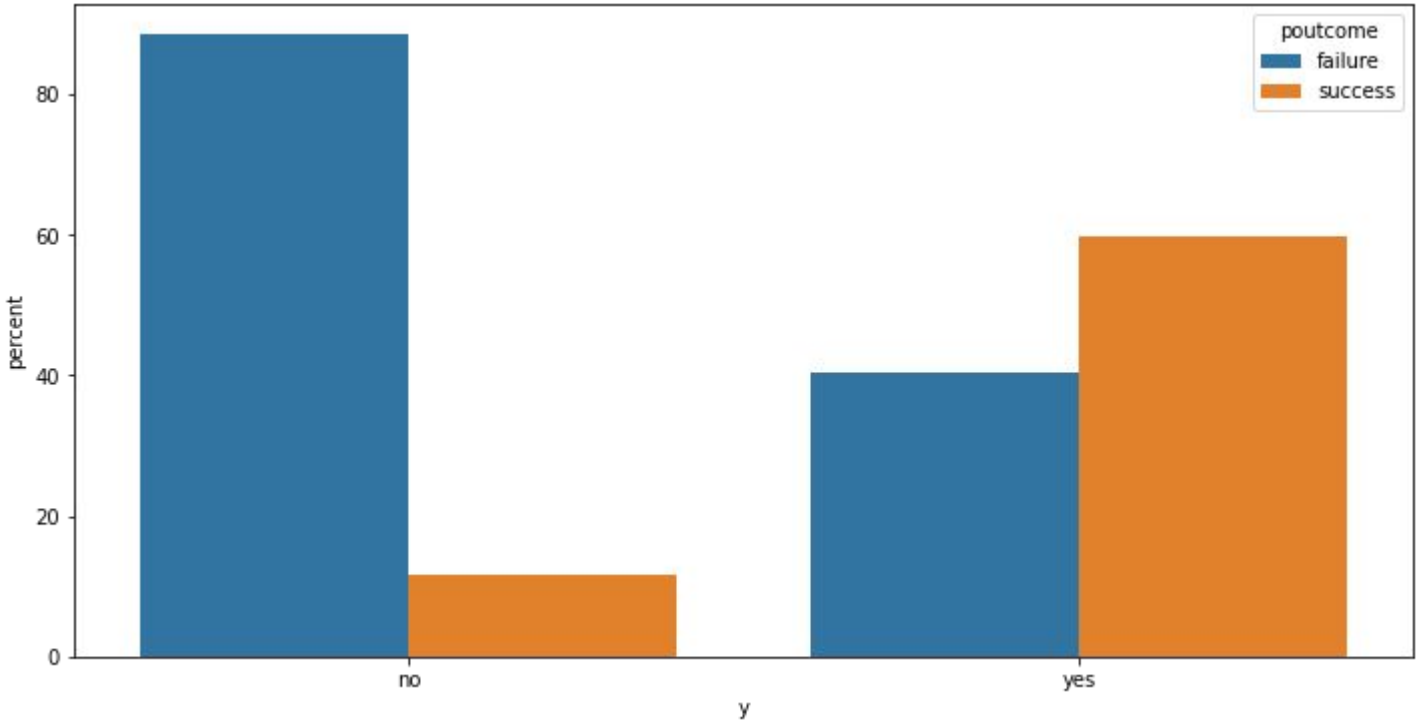
Curación

Transformaciones y modificaciones que se realizaron sobre la base de datos:

- Imputación de las las categorías “desconocido” por medio del método KNN.
- Reagrupamiento de categorías (reducción de la cardinalidad).
- Recodificación de variables *yes/no.*
- Eliminación de los valores outliers de la variable representativa de la cantidad de veces que fue contactado durante la campaña de marketing (menos del 3% de casos).
- Análisis de Componentes Principales de las variables numéricas. Se incorporaron las 4 primeras a la base de datos, explican más del 90% de la variabilidad.
- Eliminación de variables no relevantes, inconsistentes o correlacionadas.

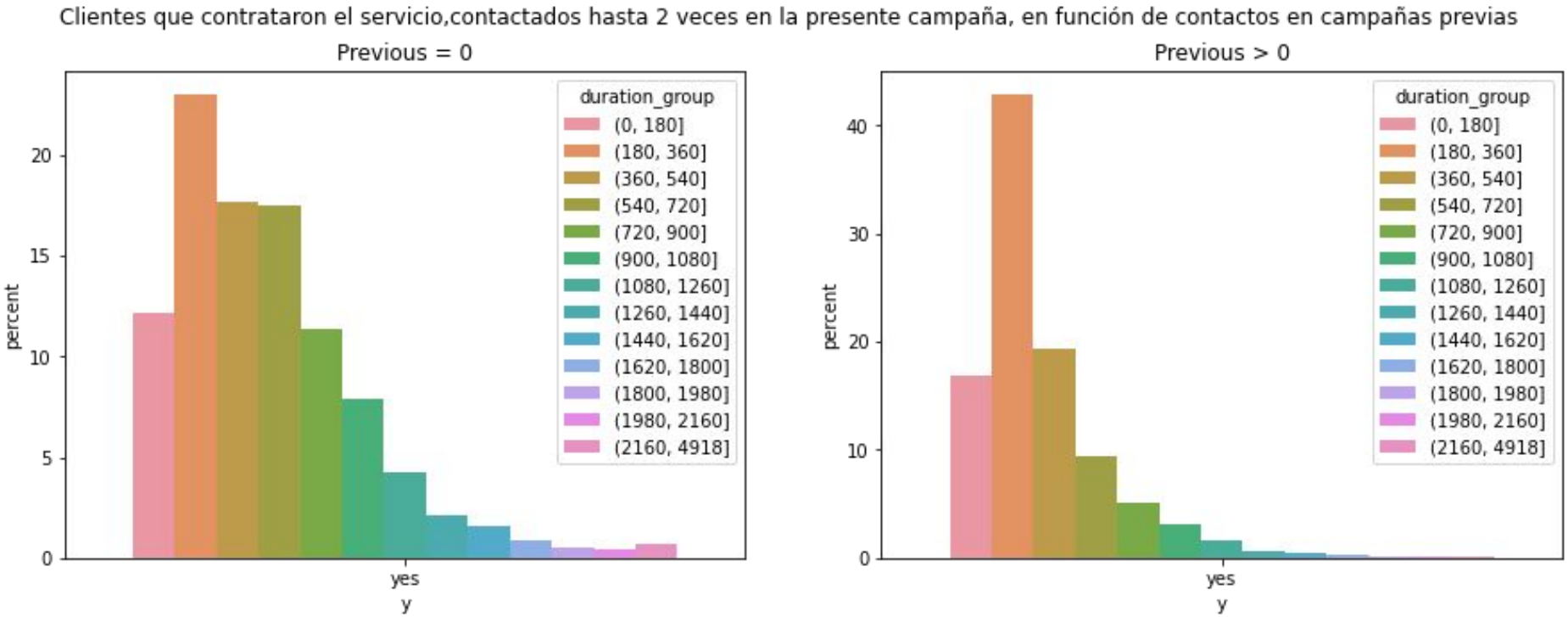
Algunos *insights* interesantes para el negocio

Resultado de la campaña actual, según si el cliente había contratado luego de campañas previas (solo clientes que habían sido previamente contactados)



Dentro de los casos de éxito de la campaña actual, **hay mayor proporción de clientes que ya habían contratado el PF previamente.**

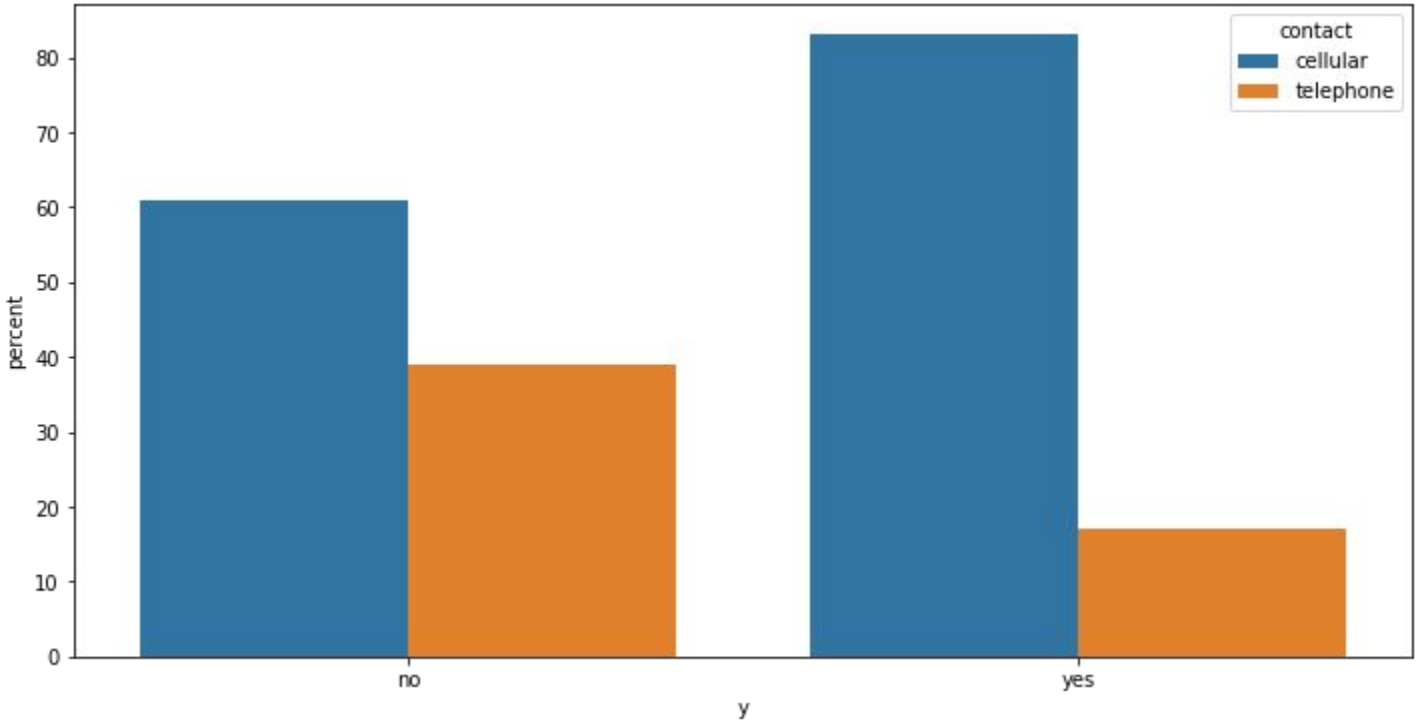
Algunos *insights* interesantes para el negocio



Cuando un cliente fue contactado previamente (previous > 0), **la contratación parece concretarse más rápido (hay un pico en las llamadas de 3 a 6 minutos)**

Algunos *insights* interesantes para el negocio

¿Influye el medio de contacto ?



Se advierte una **mayor proporción de contacto por celular en los casos exitosos**, que en los casos de no éxito.

DIVISIÓN del conjunto de datos en 3 grupos: Entrenamiento, Validación y Test.

- Debido al desbalance de los casos en la base de datos: la división se hizo incluyendo un **parámetro de estratificación** según la clase.

Conjunto de **ENTRENAMIENTO + VALIDACIÓN**:
80% de los datos

Conjunto de **TEST**:
20% de los datos

PRE-PROCESAMIENTO:

- Se emplearon pipelines y transformers sencillos.
- Se eliminaron los casos outliers de la variable “campaign” (Práctico 2).
- Exclusión de atributos que presentan alta correlación: 'emp.var.rate', 'cons.price.idx', 'nr.employed' (Práctico 2).
- Por inconsistencia: se dejó de lado la variable 'pdays' (Práctico 2).
- Se dejó de lado 'duration' para entrenar los modelos, ya que no podríamos contar con esa información para caracterizar nuevos clientes a contactar.
- Casos nulos: en algunas de las pruebas realizadas, las categorías “desconocido” de las variables categóricas fueron imputadas por medio del método “más frecuente”. En otras, se mantuvo como una categoría adicional.
- Estandarización: realizamos un escalamiento de los datos con *StandardScaler*, en casos necesarios.

MODELO BASE

- Modelo **Árbol de Decisión** para clasificación binaria (contrata o no PF).
- **Métricas**. Dado que el conjunto de datos está desbalanceado, enfocamos la atención en: **precisión**, **recall** y **F1-Score** de la clase positiva (clientes que si contratan PF). Luego, incluímos también la ROC-AUC.

Estimaciones:

1. Primera aproximación simple del modelo: sobre ajustó el conjunto de train y las métricas en el conjunto de validación no son buenas para la clase minoritaria (1 = Contrató PF). Se aplicó cross validation para tener más seguridad sobre las métricas.
2. Se optimizaron parámetros y definimos como **modelo base** al **mejor modelo encontrado del Árbol de Decisión**. Se redujo el sobreajuste de conjunto de entrenamiento respecto del primer modelo entrenado, y aumentaron los casos correctamente clasificados para la clase minoritaria.

MODELO BASE

ENTRENAMIENTO

	precision	recall	f1-score	support
0	0.96	0.86	0.91	22715
1	0.40	0.71	0.51	2945
accuracy			0.85	25660
macro avg	0.68	0.79	0.71	25660
weighted avg	0.89	0.85	0.86	25660

VALIDACIÓN

	precision	recall	f1-score	support
0	0.94	0.82	0.88	5679
1	0.31	0.61	0.41	736
accuracy			0.80	6415
macro avg	0.63	0.72	0.64	6415
weighted avg	0.87	0.80	0.83	6415

En el conjunto de validación:
- El 31% de los clientes que se predice contratarían PF, efectivamente lo hacen
- Identifica al 61% de los clientes que contratan PF
- F1 igual a 0.41.

Por medio del modelo base obtuvimos una métrica **AUC de aprox**. **0.72**, lo cual es mejor que si la clasificación fuera aleatoria.

```
ROC_AUC para datos de Train: 0.7875952742834271
ROC_AUC para datos de Val: 0.7167193732439116
```

Aprendizaje Supervisado

Análisis de diferentes modelos:

- Se entrenaron 8 modelos diferentes para clasificación binaria: 1. Decision Tree, 2. SGD Classifier, 3. Logistic regression, 4. SVM, 5. Naive Bayes, 6. Random Forest. 7. XGBoost, 8. LightGBM, y 9. Bagging.

- Se adaptó el pre-procesamiento en pipelines según las particularidades de cada modelo a entrenar (estandarizaciones/imputaciones/discretizaciones).
- Se optimizaron los hiper-parámetros de cada uno de los modelos, tanto a partir de GridSearchCV como RandomizedSearchCV.
    - Limitación: el tiempo de procesamiento para búsqueda de parámetros con GridSearchCV.

Aprendizaje Supervisado

El **MEJOR MODELO** encontrado: **XGBOOST**

Se entrenó el modelo que ofreció los mejores resultados con diferentes combinaciones de variables, de modo de identificar la mejor especificación del mismo.

MÉTRICAS CONJUNTO DE TRAIN

	precision	recall	f1-score	support
0	0.94	0.92	0.93	22715
1	0.47	0.56	0.51	2945
accuracy			0.88	25660
macro avg	0.71	0.74	0.72	25660
weighted avg	0.89	0.88	0.88	25660

MÉTRICAS CONJUNTO DE VALIDACIÓN

	precision	recall	f1-score	support
0	0.94	0.92	0.93	5679
1	0.47	0.53	0.50	736
accuracy			0.88	6415
macro avg	0.70	0.73	0.71	6415
weighted avg	0.88	0.88	0.88	6415

En el conjunto de validación:
- El 47% de los clientes que se predice contratarían PF, efectivamente lo hacen
- Identifica al 53% de los clientes que contratan PF
- F1 igual a 0.5

```
ROC_AUC para datos de Train: 0.7459556241864664
ROC_AUC para datos de Val: 0.7427901086765123
ROC_AUC para datos de Test: 0.7219738542476896
```

El mejor modelo presentó una **ROC_AUC** levemente mayor que el modelo base, pero bastante superior al rendimiento de un clasificador aleatorio

El **MEJOR MODELO** encontrado - **XGBOOST**

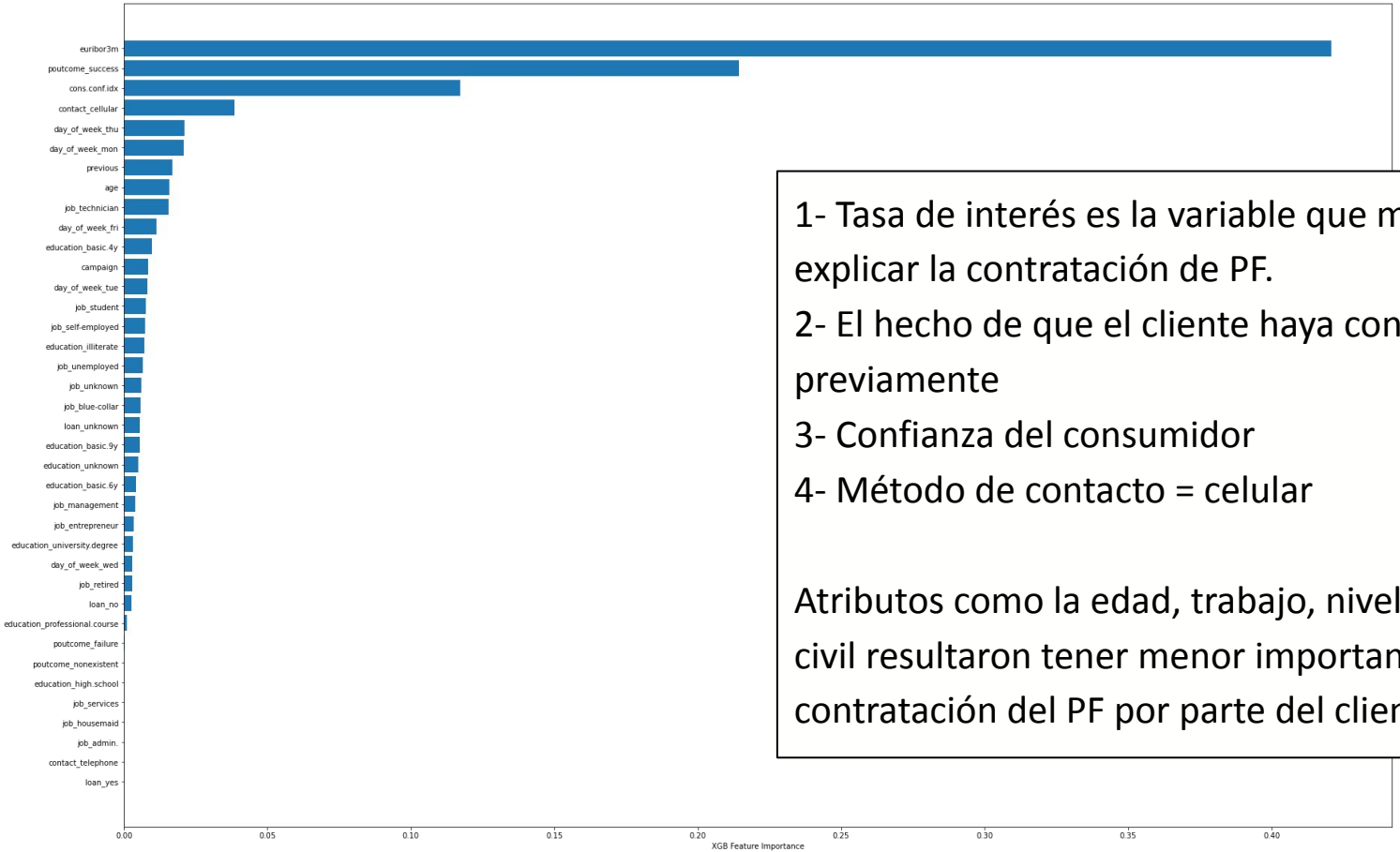
Respecto del modelo base, permitió obtener una **MEJORA** de:

- la ***precision*** en **0,16** (aunque pérdida de 0,08 en la recall).
- **F1** en **0,09.**
- **ROC_AUC** en **0,03.**

¿Qué más podríamos hacer para mejorar la performance?

- Nuevos datos
- Más información para caracterizar a los clientes que ya tenemos
- Generar nuevas variables

Importancia de las atributos en el mejor modelo



1- Tasa de interés es la variable que más contribuye a explicar la contratación de PF.

2- El hecho de que el cliente haya contratado previamente

3- Confianza del consumidor

4- Método de contacto = celular

Atributos como la edad, trabajo, nivel educativo y estado civil resultaron tener menor importancia para explicar la contratación del PF por parte del cliente.

Aprendizaje No Supervisado

Análisis de clusters:

- Se aplicaron diferentes métodos para tratar de identificar clusters de clientes
- Principal limitación: las principales características de los clientes se presentan como variables categóricas (educación, trabajo, estado civil, forma de contacto).
- Probamos diferentes estrategias para la búsqueda de clusters:
  - Kmeans (variable edad y categóricas encodeadas, de a 1 o 2), Kmodas (variables categóricas y edad categorizada), Kprototypes (variable edad numérica y categóricas)
  - Kmeans, Clusters Aglomerativos y DBSCAN (componentes principales más categóricas encodeadas).
- Se analizaron el método del codo e hiperparámetros para determinar el número de clusters óptimos.

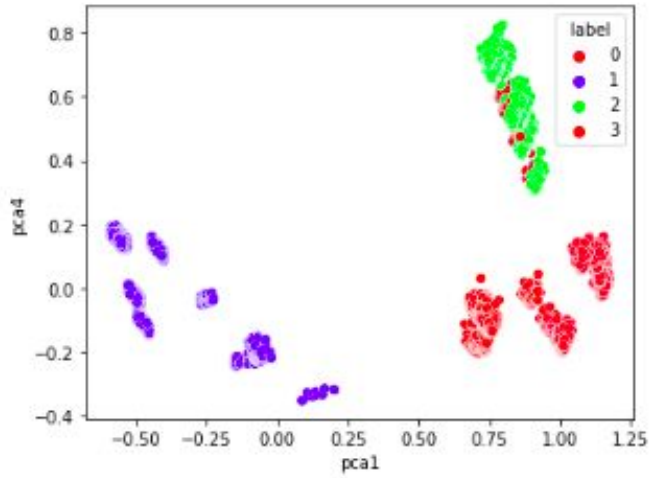
Aprendizaje No Supervisado

Análisis de clusters:

- Los métodos tienden a identificar entre 2 y 6 clusters
- 2-4 clusters por lo general no resultan suficiente para caracterizar a todos los clientes.
- En algunos casos, fue posible identificar grupos de clientes diferenciados, por ejemplo: jóvenes de entre 25-45 años con alto nivel educativo, jóvenes de entre 25-45 años con bajo nivel educativo, adultos de 45 años o más con nivel educativo medio-bajo.
- Limitaciones: no siempre fue posible obtener diferencias significativas en la propensión a contratar PF entre los grupos identificados.
- Los agrupamientos más interpretables fueron obtenidos con las variables numéricas.
- El mejor resultado obtenido a partir del análisis de clusters fue usando el método **Hierarchical Clustering**

Aprendizaje No Supervisado

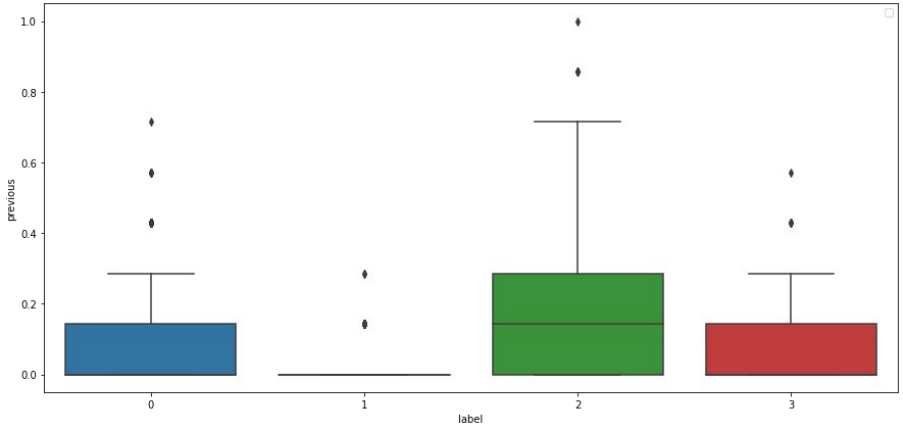
Hierarchical Clustering usando solo las PCA



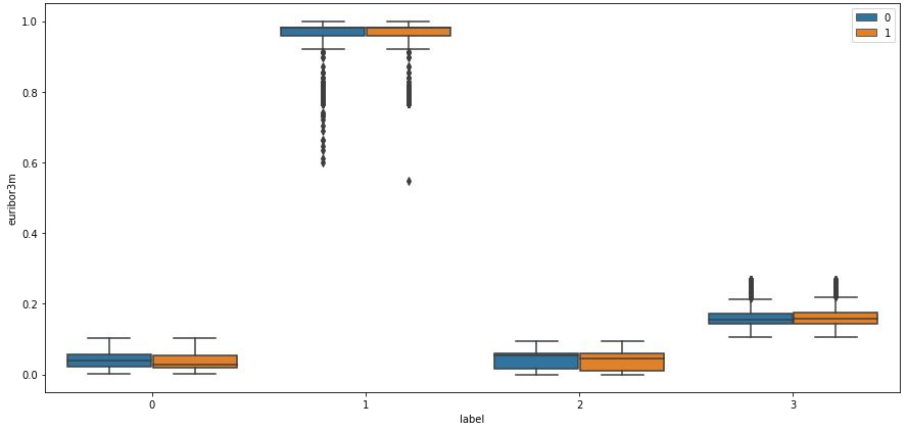
Crosstab - Cluster vs Target (y)

label	0	1	2	3	All
y					
0	1341	25090	924	7758	35113
1	873	939	1011	1220	4043
All	2214	26029	1935	8978	39156

Llamados previos (previous)

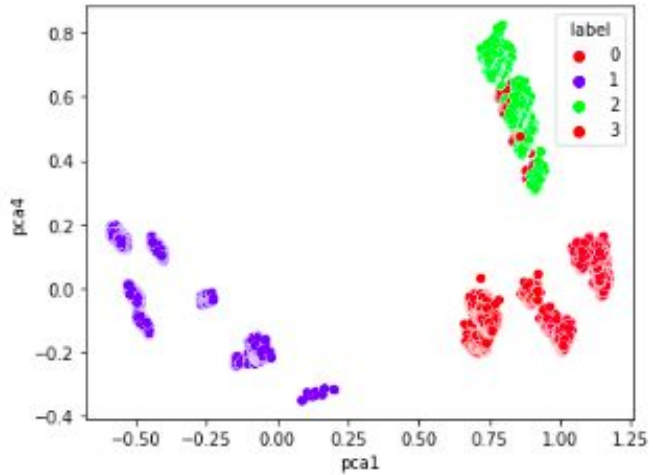


Tasa de interés (euribor3m)



Aprendizaje No Supervisado

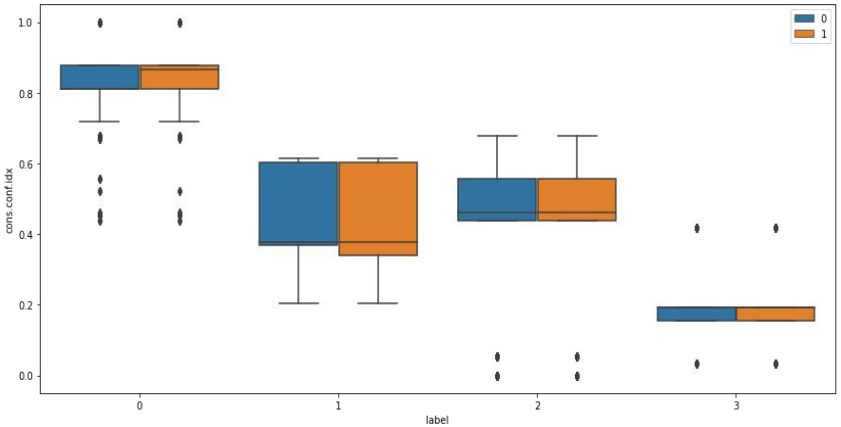
Hierarchical Clustering usando solo las PCA



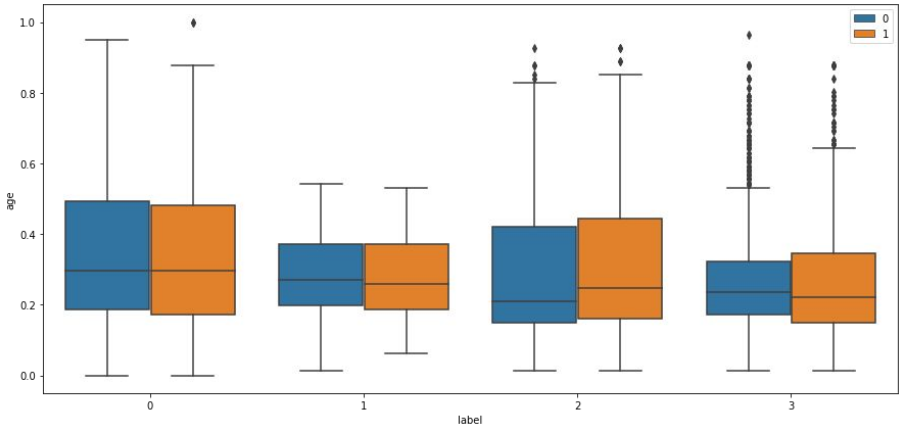
Crosstab - Cluster vs Target (y)

label	0	1	2	3	All
y					
0	1341	25090	924	7758	35113
1	873	939	1011	1220	4043
All	2214	26029	1935	8978	39156

Índice de Confianza (cons.conf.idx)



Edad (age)



Conclusión General

Principales puntos a comentar/recomendar al Banco de Portugal:

- Como lo indicaría la intuición, la **tasa de interés** es el **principal factor** que se vincula con la contratación o no de un PF por parte de un cliente.
- Adicionalmente, dentro del efecto del entorno, el **nivel de confianza de los consumidores** es otro factor que pareciera tener relevancia.
- Si bien no fue posible realizar una segmentación robusta de clientes en función de sus características personales (edad, nivel educativo, etc), se destaca la relevancia de aquellos clientes que habían sido contactados previamente en otras campañas de marketing.
- **La probabilidad de contratar PF es significativamente mayor entre aquellos clientes a los que se había contactado previamente por otra campaña.**

¡Muchas gracias!

Santiago Seppi

Agustín Carchano

Florencia Cámara

Material de mentoría disponible en: https://github.com/AgusCarchano/Mentorias-grupo1