Predicción de la contratación de clientes de electricidad

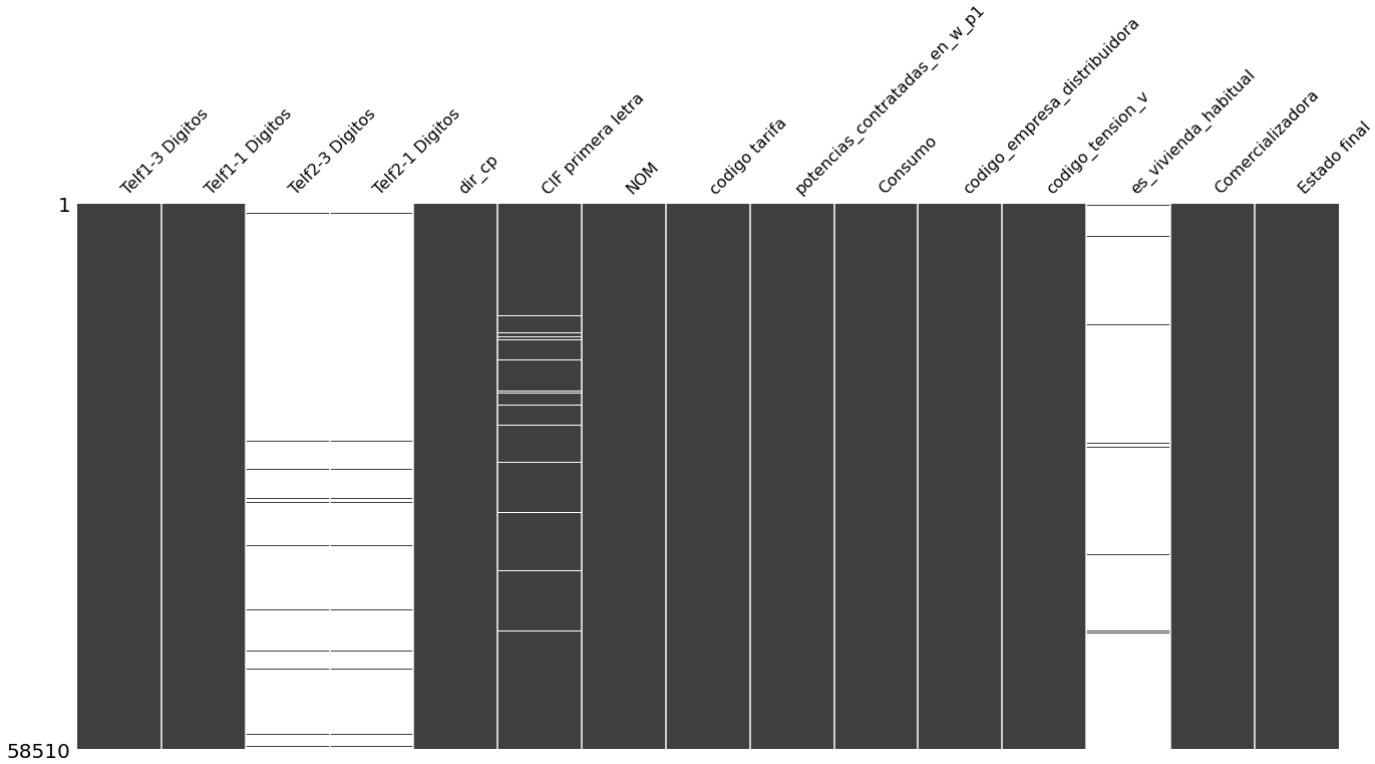
Rubén Cruz García

2/11/2020
Barcelona

Descripción del *dataset*

- Disponemos de las siguientes variables:
  - Telf1-3 Dígitos
  - Telf1-1 Dígitos
  - Telf2-3 Dígitos
  - Telf2-1 Dígitos
  - Código Postal
  - Primera Letra CIF
  - Nombre
  - Código Tarifa
  - Potencias contratadas (W)
  - Consumo anual (kW h)
  - Código empresa distribuidora
  - Código tension
  - Vivienda Habitual (Sí/No)
  - Comercializadora anterior
  - **Estado final: Variable a predecir.**

Generamos una variable → Fijo (1), móvil (2), ambos (3).



Los datos faltantes para cada variable y muestra están representados en blanco.

Número de muestras: **58510**

Contratación **media** por categoría

- **Las mayores contrataciones medias se dan para:**

- Personas con fijo y móvil (Cat. Ambos).

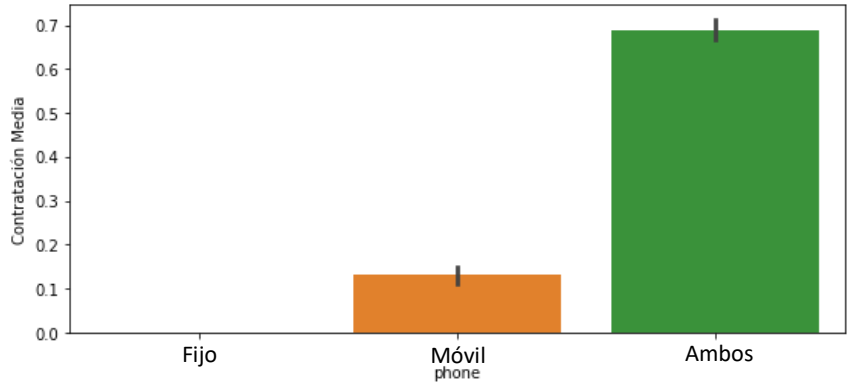
- Tarifas 3 y 6 (11 solo hay 1 persona).

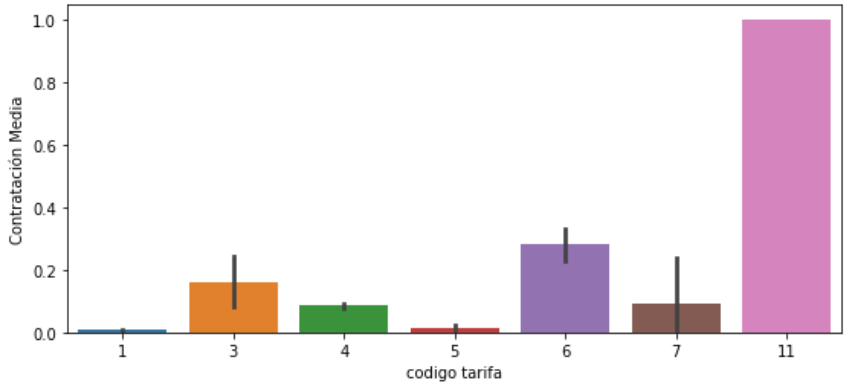
- Código tensión 10 (27 solo hay 1 persona).

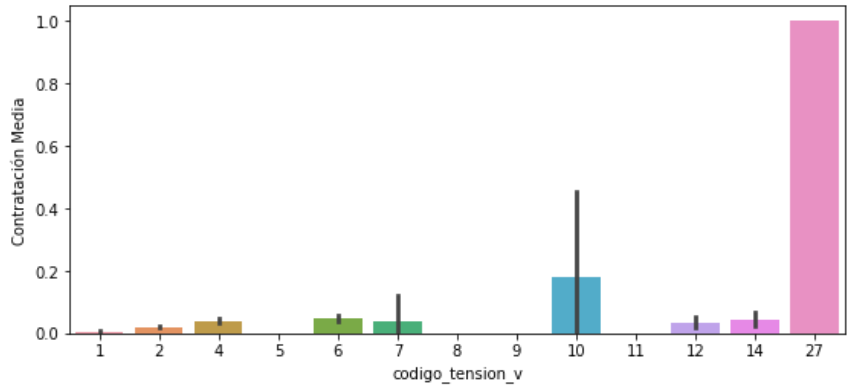
- Empresa distribuidora 27.

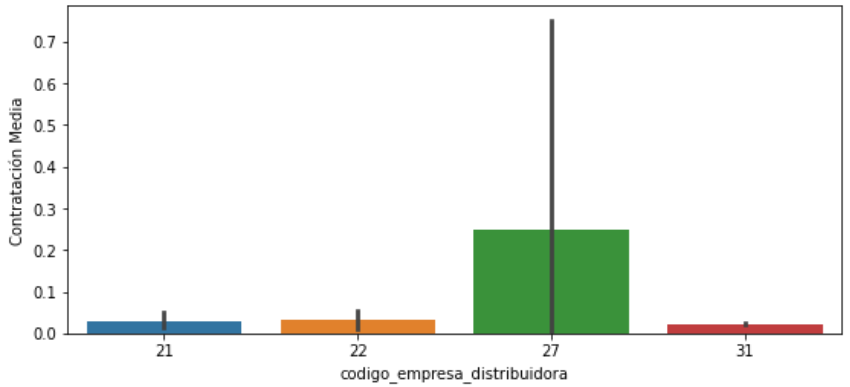
- Potencia contratada “muy alta” (6.6-40 kW).

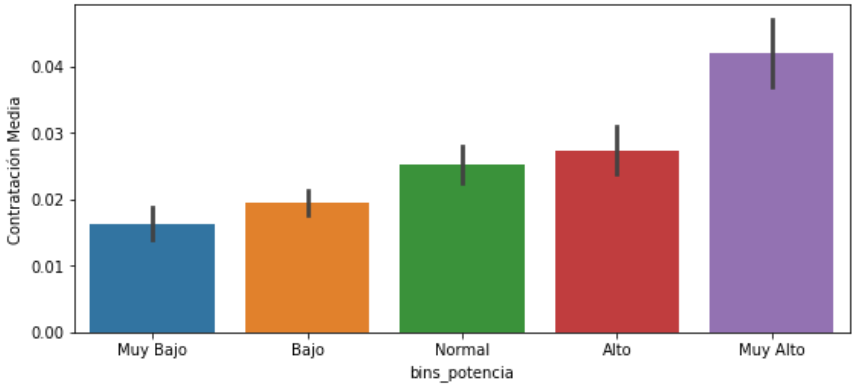
- Consumo “muy alto” (5-42 kWh anuales).

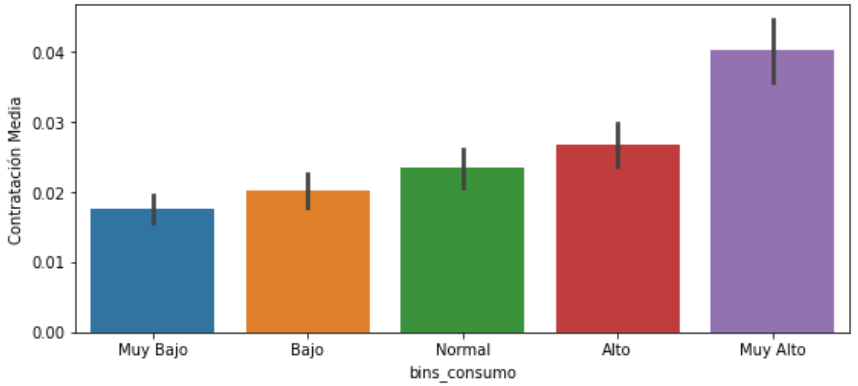








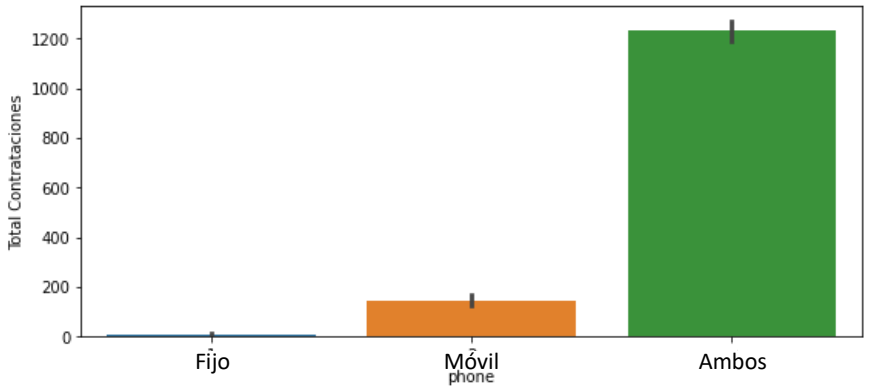


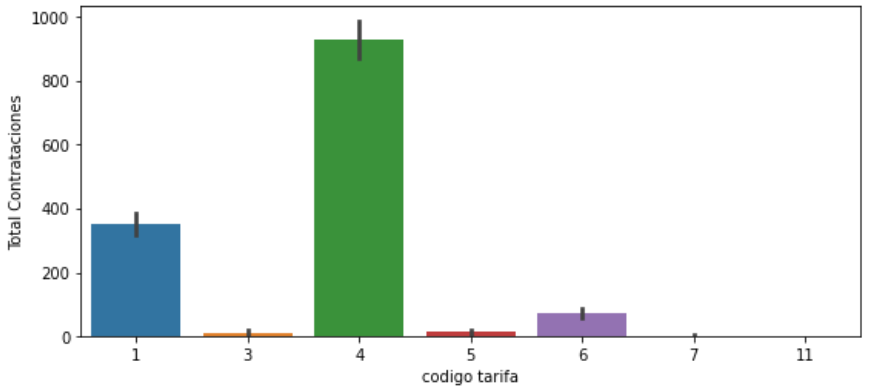


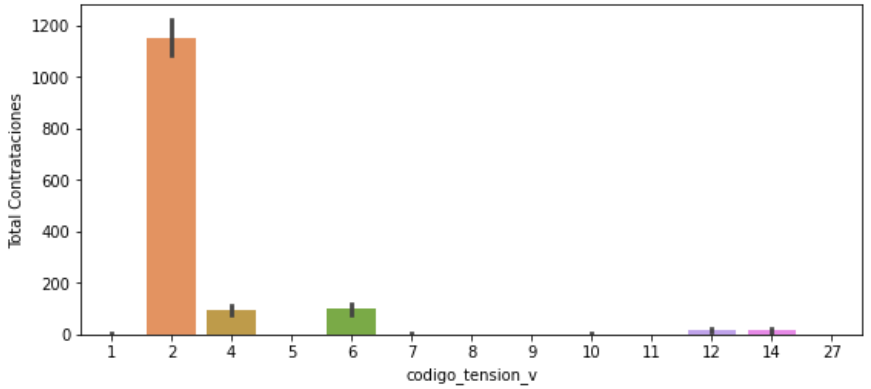
Contrataciones **totales** por categoría

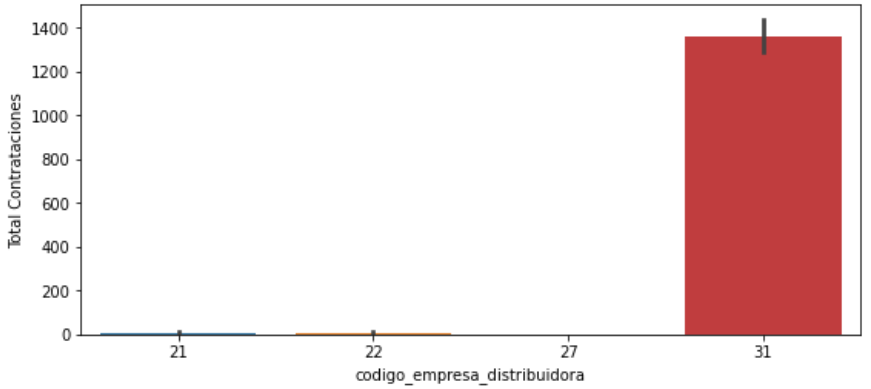
- **Las mayores contrataciones totales se dan para:**

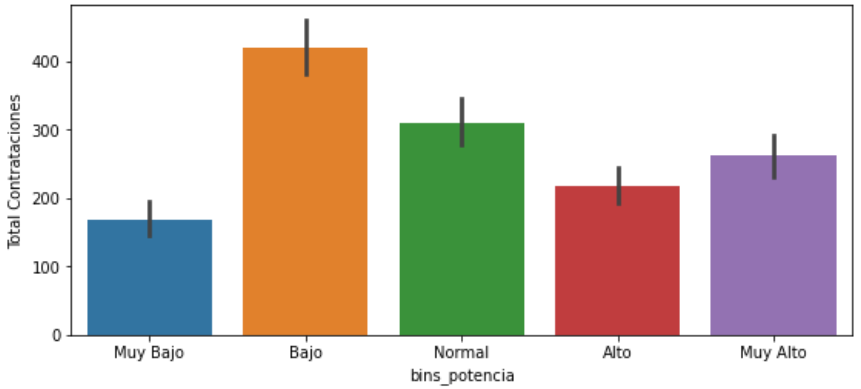
- Personas con fijo y móvil (Cat. Ambos).
- Tarifas 1 y 4.
- Código tensión 2.
- Empresa distribuidora 31.
- Potencia contratada “baja” (4.4-5.5 kW).
- No hay consumo con mayores contrataciones.
- Comercializadora anterior 91 (no mostrado).

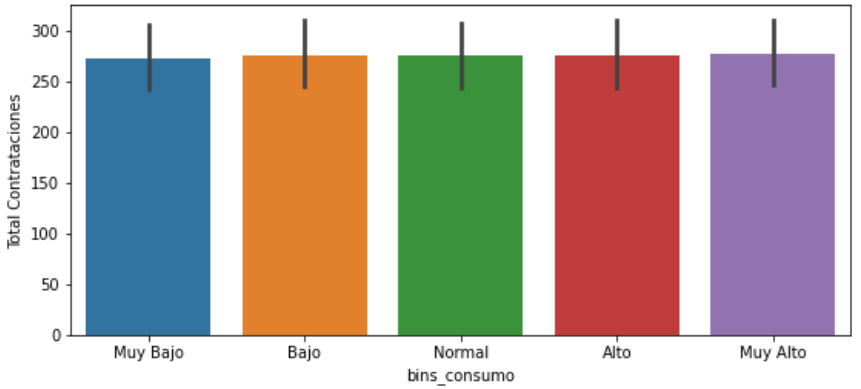










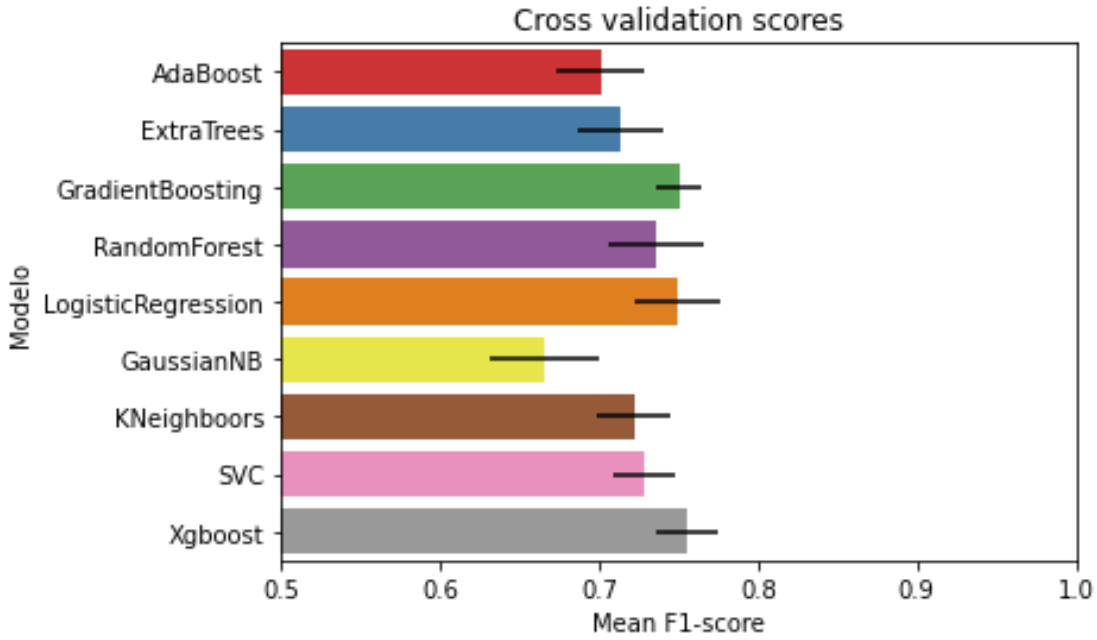


Modelos de Machine Learning

- Variables predictoras utilizadas: *Código tarifa, potencia contratada, consumo, código empresa distribuidora, comercializadora, tipo de teléfono* (variable generada).

- Inicializamos distintos algoritmos, ya que no hay uno que funcione mejor para todas las situaciones.

- Los evaluamos según su F1-Score medio (cuanto más cerca de 1, mejor).



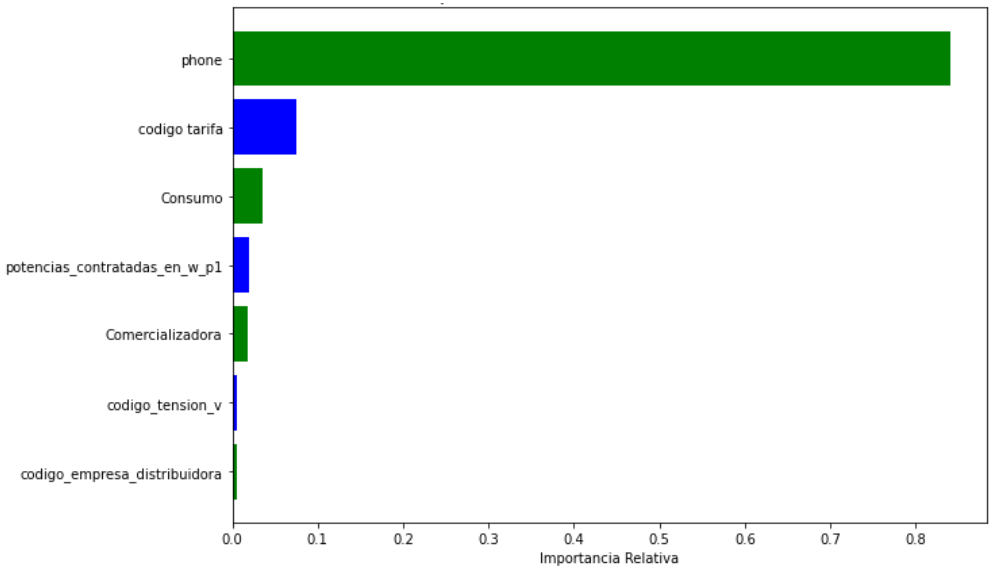
Elegimos Random Forests, Logistic Regression y XgBoost

Tuning de los modelos

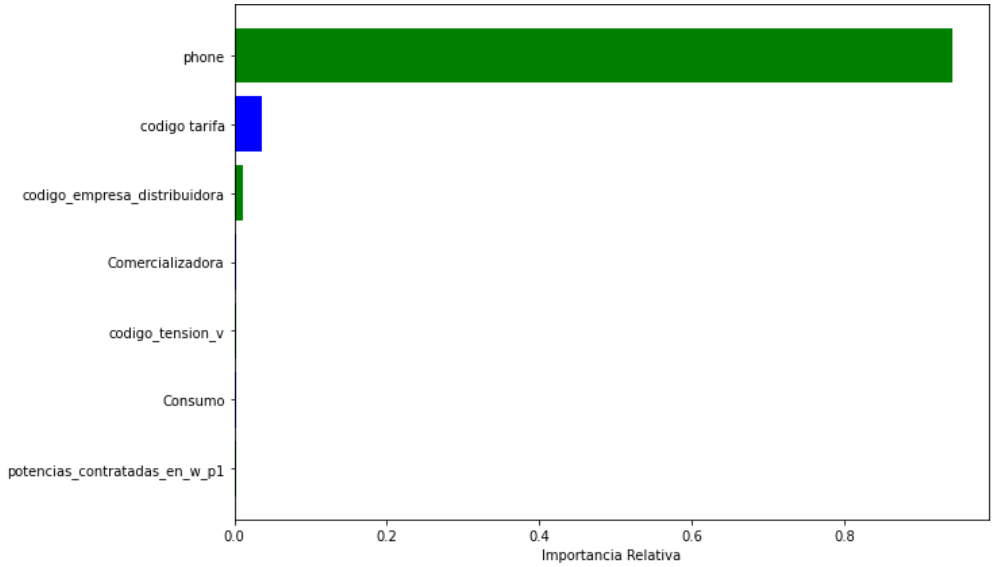
- Con *Search Grid* buscamos los mejores parámetros para los 3 algoritmos, que son:

```
Random Forests:
{'bootstrap': False, 'criterion': 'gini', 'max_depth': 8, 'max_features': 5, 'min_samples_leaf': 3, 'min_samples_split': 2, 'n_estimators': 200, 'random_state': 3}
Logistic Regression:
{'C': 0.1, 'penalty': 'l2', 'random_state': 3, 'solver': 'liblinear'}
XgBoost:
{'learning_rate': 0.01, 'max_depth': 8, 'n_estimators': 100, 'random_state': 3}
```

Importancia variables – Random Forests



Importancia variables – XgBoost



Para más detalles de la metodología y el análisis:
https://github.com/rcruzgar/prediccion_clientes/blob/main/Modelo_Clientes.ipynb

Evaluación de los modelos

Random Forests

	precision	recall	f1-score	support
1	0.73	0.71	0.72	276
0	0.99	0.99	0.99	11407
accuracy			0.99	11683
macro avg	0.86	0.85	0.86	11683
weighted avg	0.99	0.99	0.99	11683

==========

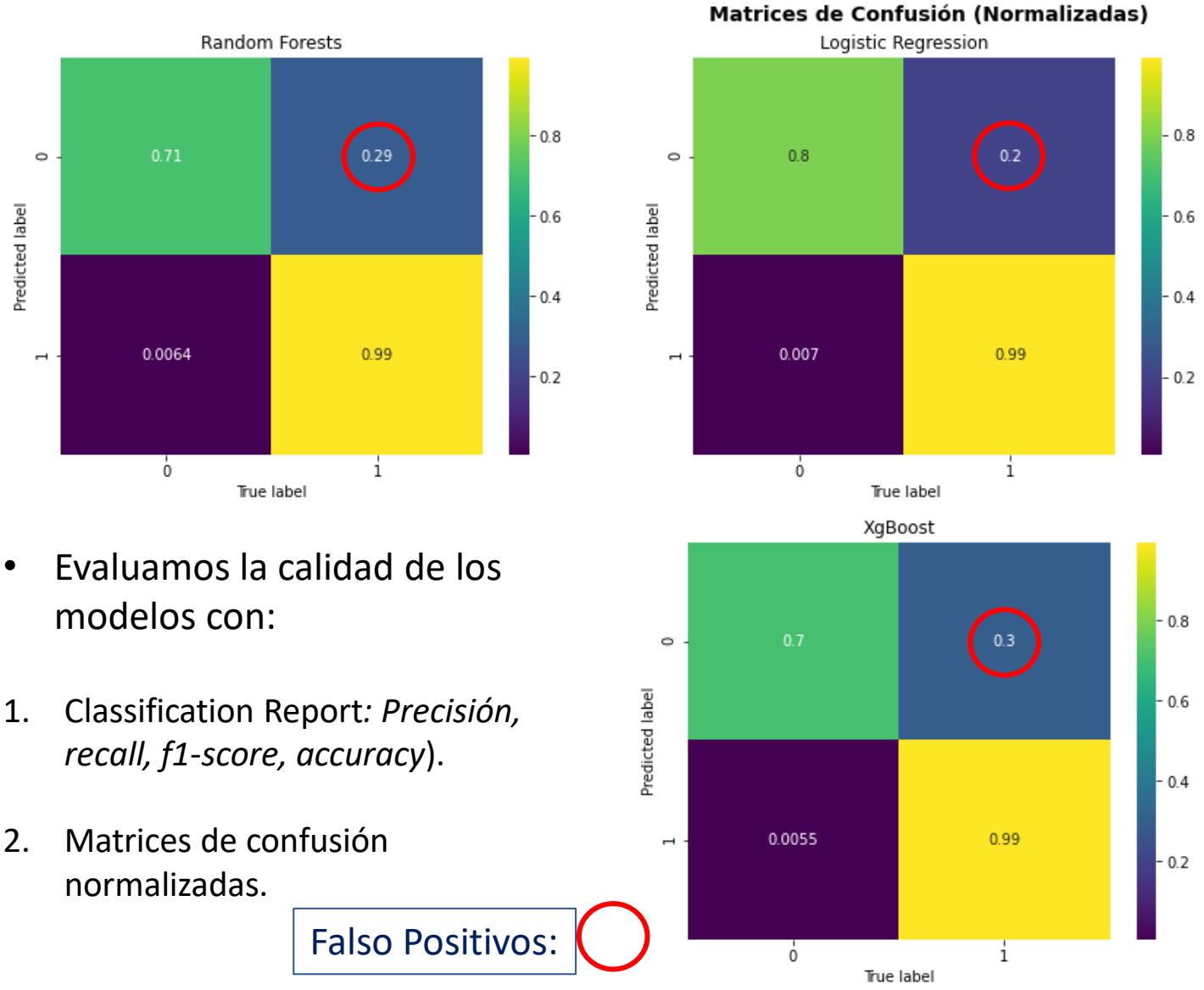
Logistic Regression

	precision	recall	f1-score	support
1	0.73	0.80	0.76	276
0	1.00	0.99	0.99	11407
accuracy			0.99	11683
macro avg	0.86	0.90	0.88	11683
weighted avg	0.99	0.99	0.99	11683

==========

XgBoost

	precision	recall	f1-score	support
1	0.75	0.70	0.73	276
0	0.99	0.99	0.99	11407
accuracy			0.99	11683
macro avg	0.87	0.85	0.86	11683
weighted avg	0.99	0.99	0.99	11683



- Evaluamos la calidad de los modelos con:

1. Classification Report*: Precisión, recall, f1-score, accuracy*).
2. Matrices de confusión normalizadas.

Falso Positivos:

Mejoras de los modelos

Tras la evaluación de los modelos, el más idóneo para predecir si un cliente potencial terminará contratando los servicios es el *Logistic Regression*. Las posibles mejoras de los modelos incluirían:

- Tuning con Feature Selection. Más variables no significa mejor modelo, sino que a menudo seleccionando las variables predictoras adecuadas se puede obtener un modelo que generalice mejor.
- Search Grid más exhaustivo.
- Probar más algoritmos.
- Eliminar previamente los *outliers* de la muestra. Igualmente, probar con otros métodos de escalado.
- Se podrían combinar los modelos mediante un *Voting Classifier*, donde se tengan en cuenta las predicciones de todos los modelos.