Calculo de probabilidad de que Clientes bancarios realicen Depósitos a Plazo

Integrantes del proyecto: Juan Cassinerio - Juan Isasi - Jose Vargas

13/05/2023

Contenido:

1. Introducción

2. Dataset y Análisis de Variables

    2.1 Objetivos

        2.1.1 Relación entre estado civil del cliente y Probabilidad de realizar depósitos a plazo

        2.1.2 Meses en que los clientes son más predispuestos a realizar depósitos a plazo

        2.1.3 Relación entre edad del cliente y duración del último contacto con el mismo con la Probabilidad que este efectué un depósito a plazo

        2.1.4 Relación entre la profesión - números de contacto durante la reciente campaña - de contactos en previas campañas con del cliente y la Probabilidad que este efectue un deposito a plazo

3. Desarrollo del Modelo

    3.1 Modelo por Regresión Lineal

    3.2 Modelo por Decision Tree

    3.3 Modelo por Random Forest

4. Conclusiones

1. **Introducción**

- ***Contexto empresarial***

Dentro de los productos y servicios que ofrecen los bancos, los depósitos a plazo representan un gran porcentaje de las fuentes de ingresos. Para poder maximizar la cantidad de depósitos a plazo realizados, se buscará identificar a los clientes (parámetros) que tengan la mayor probabilidad de subscribirse a este tipo de producto y de esa manera poder mejorar los esfuerzos de marketing en dichos clientes.

- ***Problema empresarial***

Se recopilo un conjuntos de datos con el objetivo de desarrollar modelos de **machine learning** para predecir la probabilidad de que clientes realicen depósitos a plazo en bancos. Estos datos fueron tomados de una campaña de marketing de una importante institución bancaria de Portugal.

- ***Contexto analítico***

Se nos permitió el acceso a un .cvs donde emplearemos un análisis descriptivo para poder entender de mejor forma los datos recopilados y

su relación con la variable a investigar "y" (Si el depósito a plazo fue realizado o no).

Dataset: https://www.kaggle.com/datasets/rashmiranu/banking-dataset-classification.

2. **Dataset y Análisis de Variables**

El análisis servirá para verificar si una persona será propensa a realizar un depósito a plazo. La variable **output** a predecir es **y** donde 1: se suscribió 0: no se suscribió.

Para ello se cuenta con las siguientes variables **input** :

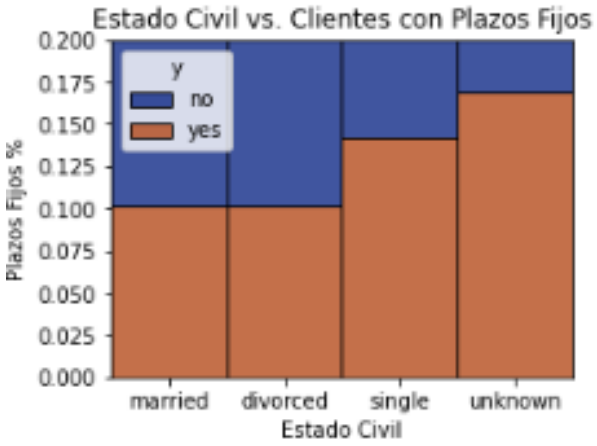
#	Variable	Concepto
1	age	edad del cliente
2	job	profesion ('admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown') - marital: estado civil
3	marital	estado civil
4	education	formacion academica ('basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5	default	si posee credito defaulteado ('no','yes','unknown')
6	housing	si posee credito inmobiliario ('no','yes','unknown')
7	loan	si posee un credito personal ('no','yes','unknown')
8	contact	metodo de contacto ('cellular','telephone')
9	month	ultimo mes de contacto realizado al cliente durante la campaña de marketing ('jan', 'feb', 'mar', ..., 'nov', 'dec')
10	day_of_week	ultimo contacto de la semana ('mon','tue','wed','thu','fri')
11	duration	duracion del ultimo contacto en dicha llamada, en segundos
12	campaign	numero de contactos realizados durante esta ultima campaña
13	pdays	numero de dias que pasaron desde la ultima vez que fue contactado (999 means client was not previously contacted)
14	previous	numero de contactos realizados a este cliente antes de la presente campaña
15	poutcome	resultado de la campaña ('failure','nonexistent','success')

2.1 **Objetivos**

Las siguientes son las preguntas a responder a partir del conocimiento de los datos:

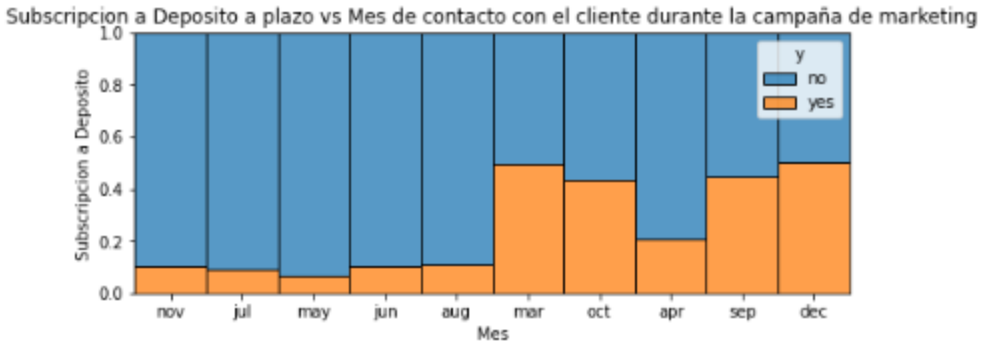
- Relación entre estado civil del cliente y Probabilidad de realizar depósitos a plazo
- Meses en que los clientes son más predispuestos a realizar depósitos a plazo
- Relación entre edad del cliente y duración del último contacto con el mismo con la Probabilidad que este realice un depósito a plazo

2.1.1 **Relación entre estado civil del cliente y Probabilidad de realizar depósitos a plazo**



Se observa que predominan los depositos a plazo en clientes Solteros con un 15%, mientras que las personas casadas muestran un 10%, un 33% menor. Igualmente cabe resaltar que se desconoce el estado civil del 17% de las muestras.

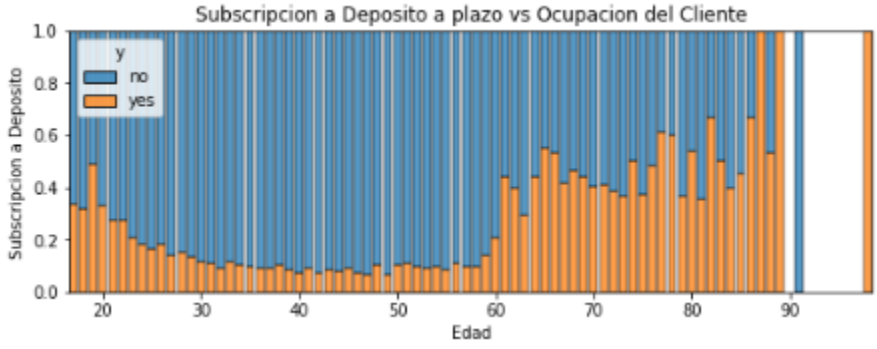
2.1.2 **Meses en que los clientes son más predispuestos a realizar depósitos a plazo**



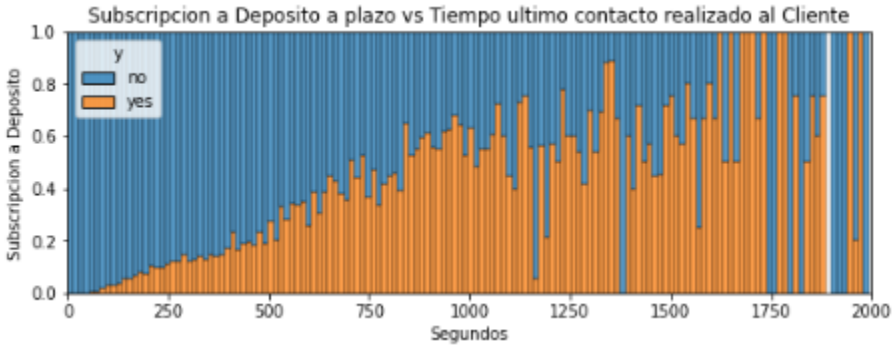
Podemos observar que los meses de Marzo, Octubre, Septiembre y Di ciembre tuvieron un ratio sobresaliente de Subscripciones a Depositos a Plazo, un 200% mayor en promedio que el resto de meses.

2.1.3 **Relación entre edad del cliente y duración del último contacto con el mismo con la Probabilidad que este efectué un depósito a plazo**

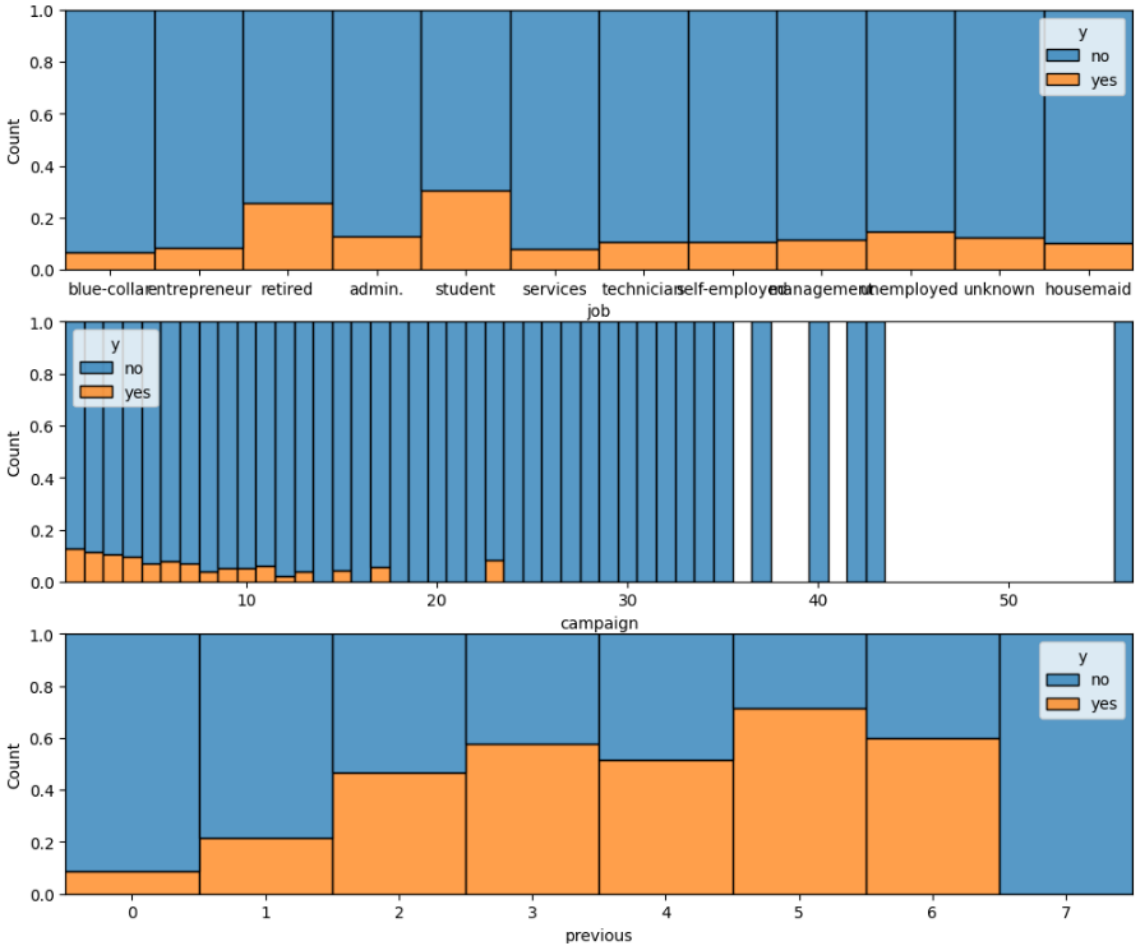
De la primera grafica podemos observar que los clientes con edad menor a 25 años y mayor a 60 años son más proclives a realizar depósitos a plazo.



De la segunda grafica se puede apreciar que las llamadas a clientes con duración menor a 100 segundos no concretaron con la realización de depósitos. Por otro lado, aquellas con una duración entre 760 segundos (12min) o más tuvieron más de un 50% de concretar en depósitos. Parece haber una correlación directa entre la ratio de suscripción y la duración de la llamada con el cliente.



2.1.4 **Relación entre la profesión - números de contacto durante la reciente campaña - de contactos en previas campañas con del cliente y la Probabilidad que este efectue un depósito a plazo**



Podemos observar que los clientes retirados o con perfil de estudiante son más proclives a realizar depósitos a plazo que las personas que trabajan en diferentes formatos.

De la segunda grafica se pude apreciar que aqueos clientes a los cuales se les hallan contactado más de 10 veces disminuye enormemente la probabilidad de que en un futuro contacto este desee continuar con la oferta.

Dela ultima y tercera grafica podemos visualizar el enorme efecto que posee el haber contactado a cliente en previas campañas, es decir, con un mayor lapso de tiempo.

El restante de variables como **"default", "education", "housing", "loan", "contact", "day_of_week", "campaign" y "pdays"** no afectan significativamente el valor de la salida (output) del modelo. Sus graficas no se incluyen para no saturar el análisis.

3. **Desarrollo del Modelo**

A continuación, realizamos 3 métodos de machine learning (**Regresión Logística, Decision Tree y Random Forest**) y evaluamos su habilidad para predecir la predisposición de clientes a realizar depósitos a plazo en relación de las variables previamente analizadas.

3.1 **Modelo por Regresión Logística**

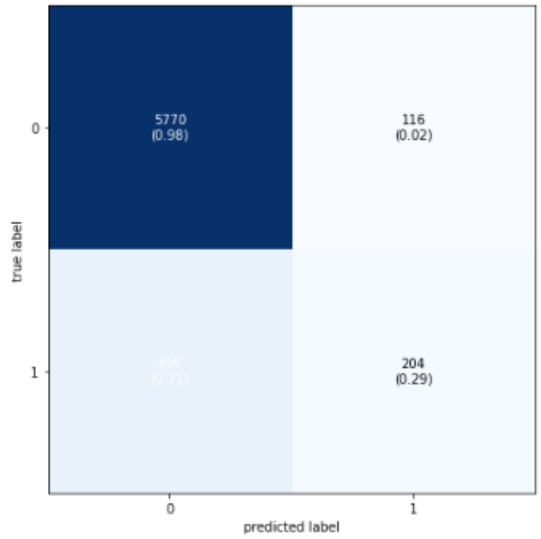
Utilizando los datos de duración de llamada, edad del cliente, el mes de contacto y estado civil del cliente realizamos un nuevo algoritmo de predicción, obteniendo un score de 91% utilizando un **Regresión Logística**.

```
accuracy_score = 0.9066626195173775
f1_score = 0.39882697947214074
recall_score = 0.2901849217638691
precision_score = 0.6375

 classification_report              precision    recall  f1-score   support

           0       0.92      0.98      0.95      5886
           1       0.64      0.29      0.40       703

    accuracy                           0.91      6589
   macro avg       0.78      0.64      0.67      6589
weighted avg       0.89      0.91      0.89      6589
```



De la matriz de confusión podemos observar la precisión del este preliminar modelo, observando los valores de falso y verdaderos negativos/positivos correspondientemente.

3.2 **Modelo por Decision Tree**

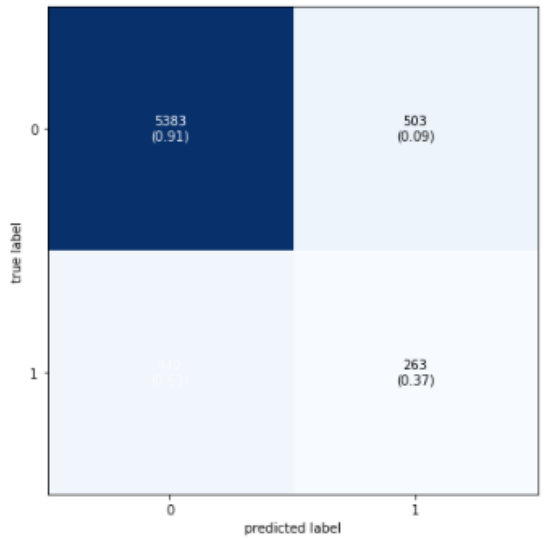
Ahora utilizamos el método **Decision Tree**, obteniendo un score de 86%, inferior en 0.04 puntos porcentuales.

```
accuracy_score = 0.8568826832599787
f1_score = 0.35806671204901297
recall_score = 0.3741109530583215
precision_score = 0.3433420365535248

 classification_report              precision    recall  f1-score   support

           0       0.92      0.91      0.92      5886
           1       0.34      0.37      0.36       703

    accuracy                           0.86      6589
   macro avg       0.63      0.64      0.64      6589
weighted avg       0.86      0.86      0.86      6589
```



3.3 **Modelo por Random Forest**

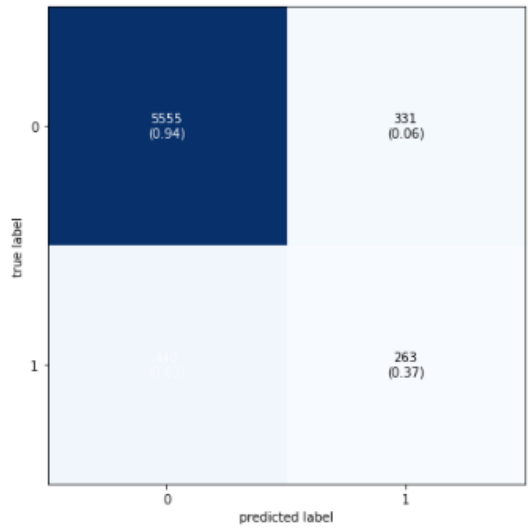
Ahora utilizamos el método **Random Forest**, obteniendo un score de 88%, inferior en 0.02 puntos porcentuales.

```
accuracy_score = 0.882986796175444
f1_score = 0.40555127216653813
recall_score = 0.3741109530583215
precision_score = 0.44276094276094274

 classification_report              precision    recall  f1-score   support

           0       0.93      0.94      0.94      5886
           1       0.44      0.37      0.41       703

    accuracy                           0.88      6589
   macro avg       0.68      0.66      0.67      6589
weighted avg       0.87      0.88      0.88      6589
```



3.4 **Modelo por Regresión Logística + Ampliación de Variables**

Dado que hemos encontrado al modelo por regresión logística como aquel que genera mejores resultados decidiremos utilizarlo para esta etapa final donde además extenderemos el número de variables con efecto apreciable en la varianza de la varíale output. Las mismas son:

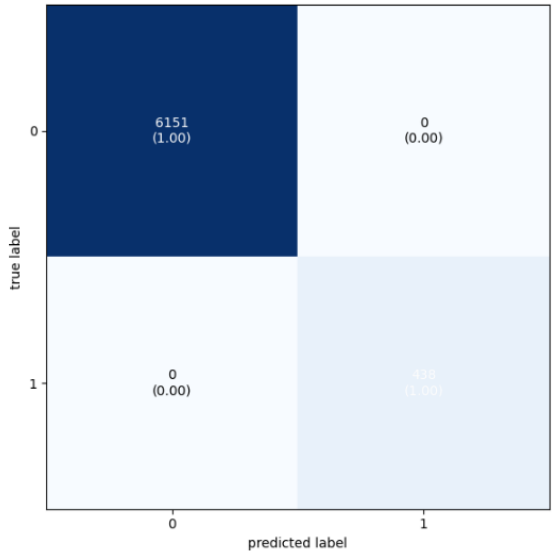
- duration
- marital
- month
- job
- previous
- campaign

```
accuracy_score = 1.0
f1_score = 1.0
recall_score = 1.0
precision_score = 1.0

 classification_report              precision    recall  f1-score   support

           0       1.00      1.00      1.00      6151
           1       1.00      1.00      1.00       438

    accuracy                           1.00      6589
   macro avg       1.00      1.00      1.00      6589
weighted avg       1.00      1.00      1.00      6589
```



Dado que hemos alcanzado un Score de 100% consideramos no necesario el incluir variables sinteticas adicionales para mejorar el modelo predictivo.

campaign: 3.710988676970976
job_sum: 3.378876737859259
duration: 2.955050459045704
age: 0.6889359688160732
month_sum: 0.5711185438270142
marital_sum: 0.45595797645724473
previous: 0.2866415398229034

Podemos observar que las variables que mayor efecto poseen en el output son **campaign, job y duration**.

4. **Conclusiones**

Los modelos implementados lograron predecir el comportamiento observado en los datos con una precisión mayor al 86%, siendo el modelo de **Regresión Logística** aquel con mayor precisión, de un **91%**.

En un nuevo intento para mejorar precisión de nuestro algoritmo aplicando regresión logística agregamos otra variables para la ronda de entrenamiento dentro de la cual obtuvimos un resultado del 100%.

Proponemos utilizar el algoritmo de regresión logística en la base de datos de bancos para seleccionar a los prospectos a los cuales se le contactará mediante una campaña telefónica exclusiva, eficientizando el tiempo y el esfuerzo del call center.