Bank Subscription

Rodriguez Cabrera Emilce y Rodriguez Cabrera Paula
Universidad Tecnológica Nacional

1- Introducción y objetivos

El objetivo del presente informe es predecir qué clientes se suscribirán a una campaña de marketing. Para ello, analizaremos la relación entre features a través de distintas funciones, y desarrollaremos dos modelos, eligiendo el que mejor prediga.

2- Descripción del dataset

A continuación se lista las características del dataset asignado:

- Samples: 45211
- Features: 18

Las features que componen al dataset son las siguientes:

- Age (Edad del cliente)
- Job (Tipo de empleo del cliente)
- Marital status (Estado civil)
- Education (Educación máxima alcanzada por el cliente)
- Credit (Si posee deuda de crédito)
- Balance (Promedio de saldo en la cuenta en el año)
- Housing loan (Si posee un seguro de hogar)
- Persona loan (Si posee un préstamo)
- Contact (Tipo de 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)
- Previous (Cantidad de contactos previos a esta campaña para cada cliente)
- Poutcome (Performance de la campaña de marketing anterior para este cliente)
- Subscription (Si el cliente accede a la campaña)

De los features listados, 10 poseen celdas vacías. Por lo que tenemos 11% de Nans en 6 features (Age, Job, Marital Status, Education, Credit, Balance (euros)) y 17% de Nans en 4 features (Housing Loan, Personal Loan, Last Contact Duration, Pdays)

Ante esto, se decidió eliminar las muestras que tenían menos del 80% de la información, es decir, aquellas muestras que tenían más de 4 Nans. Por otro lado, los Nans restantes se reemplazaron según el tipo de dato que contenían, reemplazando para:

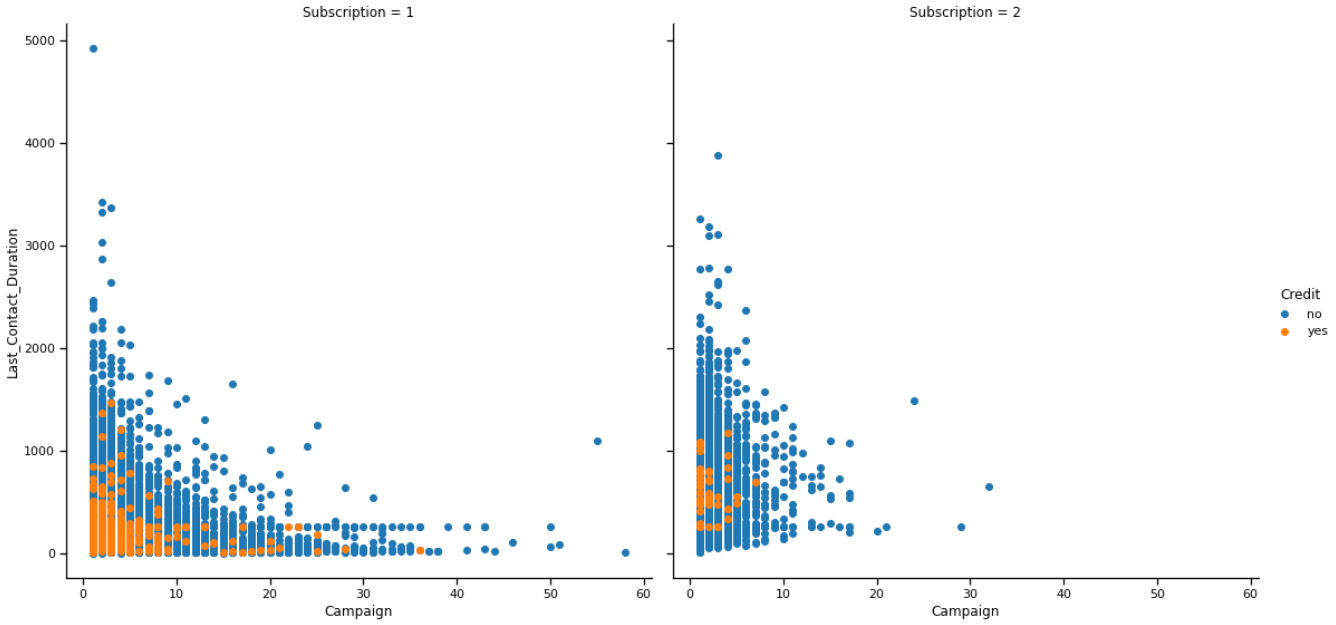
- Age, Balance (euros), Last Contact Duration y Pdays, por la media específica de cada feature.
- Job, Marital Status, Education, Credit, Housing Loan y Personal Loan, por la moda específica de cada feature.

Por último se eliminó la feature “Unnamed**”** debido a que no aportaba información. Con esto, el dataset quedó libre de Nans para avanzar con el análisis, por lo que las dimensiones finales de nuestro dataset son:

- Samples: 43650
- Feature: 17

3- EDA: análisis exploratorio de datos

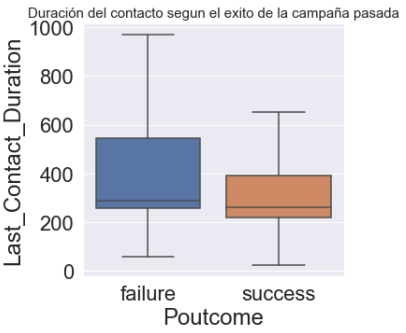
Se analizó la cantidad de contactos con el cliente y la duración de dicho contacto. Luego, se separó los que accedieron a la campaña (Subscription=2) y los que no (Subscription=1). A su vez, se segmentó según si el cliente poseia un crédito o no, obteniéndose el siguiente gráfico:



Se puede observar que los que accedieron en su mayoría, se suscribieron antes de los 10 contactos, teniendo un tiempo máximo de duración de 2000 segundos aproximadamente. En general, la duración del contacto disminuye a medida que los contactos aumentan. Además, aquellos que poseían un crédito tuvieron contactos de menor duración y tendieron a rechazar la campaña.

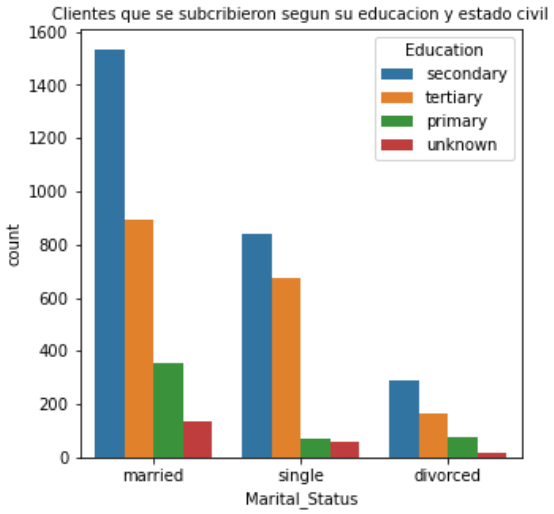
Asimismo, se realizó un boxplot, el cual permite ver la distribución que posee la duración del contacto de los que accedieron a la suscripción, donde se obtienen los siguiente datos:

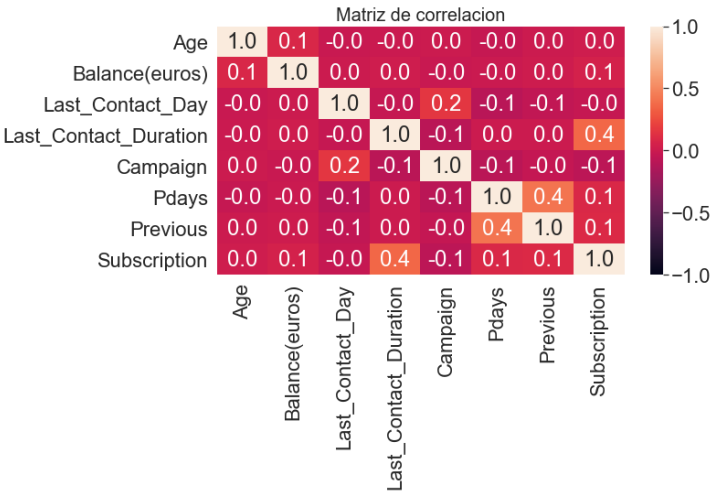
- Máximo: 1268 segundos.
- Mínimo: 8 segundos.
- Q1: 257.7 segundos.
- Mediana: 343 segundos.
- Q3: 662 segundos.



Podemos observar además que si el cliente tuvo una performance exitosa con respecto a la campaña anterior y accedió a la nueva campaña, la duración del contacto es menor respecto a aquellos clientes que tuvieron una performance fallida pasada.

Siguiendo con el análisis, en el gráfico posterior izquierdo se segmentó a los clientes que accedieron por estado civil y nivel educativo alcanzado. En consecuencia, dió como resultado que los que poseen el secundario completo sin importar el estado civil, son los que más accedieron a la campaña. Asimismo, los clientes casados accedieron en mayor medida sin importar su educación en comparación a los otros estados civiles.





Con respecto a la correlación entre features, se puede apreciar en el gráfico superior derecho una relación lineal media entre Pdays y Previous, con un valor de 0.4. Por lo que no se visualizan mayores correlaciones entre las mismas, esto no descarta otro tipo de relación no lineal.

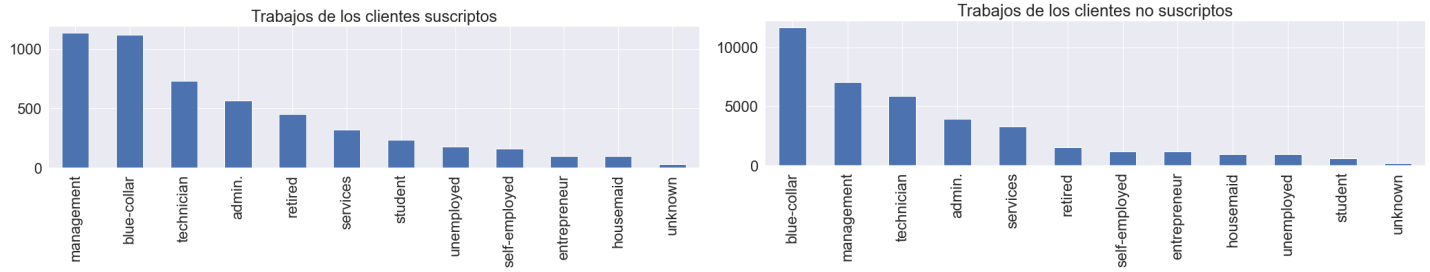
Por otro lado, se analizó la relación de los clientes que accedieron según si poseían créditos, préstamos o seguros y cómo influye en las suscripciones de manera individual:

Se observa que los seguros no difieren mucho en el acceso a la campaña. Sin embargo, que el cliente posea un crédito o un préstamo polariza la situación. Por lo que se analizó las combinaciones de estas condiciones. La probabilidad de que el cliente esté suscripto y posea:

- Ambos, es de 0.2%, teniendo solo 11 clientes del total de suscritos.
- Solo un préstamo, es de 7.4%, teniendo solo 377 clientes del total de suscritos.
- Solo un crédito, es de 0.7%, teniendo solo 36 clientes del total de suscritos.

Asimismo, también se puede apreciar que los clientes de 40 a 60 años son los que más créditos, préstamos o seguros adquirieron, según el catplot que estemos observando.

Por último, se graficaron los clientes suscritos y los no suscritos, discriminando por el trabajo. Teniendo para el primero antes mencionado, Management y obreros, mientras que para el último, se invierten teniendo primero a los obreros y luego a management.



4- Materiales y métodos

El proceso se realizó a través de Jupyter Notebook, utilizando las distintas librerías. Entre ellas podemos mencionar matplotlib.pyplot, numpy, pandas, seaborn y sklearn, entre otros.

Para poder predecir si un cliente se suscribirá o no, entrenaremos dos modelos buscando minimizar el error de clasificación. Este entrenamiento se realizó dentro de lo que se conoce como aprendizaje supervisado

Aprendizaje Supervisado

a- Train, Test y Cross Validation

Para entrenar nuestro modelo, primero se preparó el dataset. Luego, separamos en dos porciones el dataset, entre test y train y se los autoescalo. A su vez el test se separó en 5 folds, para luego iterar 90 veces. De estas 5 folds, 1 se utilizó para validar el modelo y el resto para entrenar el mismo. Esta combinación se dió de tal forma que todas las folds se utilizaron para validar una vez y no se dependiera de la suerte de la muestra determinada.

Una vez obtenido los resultados, se aseguró que el modelo no realice un sobreajuste (overfitting) y clasifique bien las muestras. Esto se realizó mediante el cálculo del promedio de todos los resultados de exactitud de todas las iteraciones.

b- Grid Search

Para facilitar la selección de los hiper-parámetros que utiliza nuestro modelo, se enlistaron los posibles valores de los mismos y se probaron las distintas combinaciones entre ellas mediante Cross Validation. Una vez que se obtuvieron los resultados, se eligió aquella combinación de hiperparametros con mayor Train Accuracy durante el cross validation. Los hiperparametros seleccionados se utilizaron para testear el modelo.

d- Métodos de clasificación

Support Vector Machine: Este método construye un hiperplano separador que maximiza el margen entre clases, donde el margen estará determinado por muestras denominadas support vector. Este método penaliza cada muestra mal clasificada con un costo (llamado C) determinado.

Logistic Regression: Es un método que se compone de una regresión lineal afectada por una función de activación, lo que genera un output binario. Este método consiste en asignarle a cada muestra clasificada una probabilidad de pertenecer a una clase determinada.

5- Experimentos y resultados

A continuación se describirán los resultados obtenidos por ambos métodos:

a- Support Vector Machine

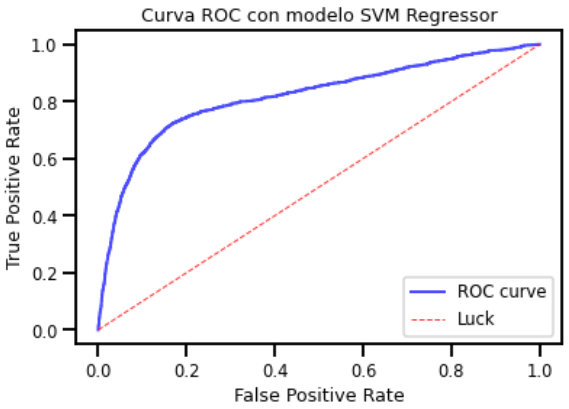
Luego de realizar el cross validation con las siguientes consideraciones:

- Kfolds = 5 folds
- Kernel Lineal y Kernel RBF
- C = [1,5,10]
- Gamma= [0.01, 0.1, 1]

Se obtuviron los siguientes resultados:

- Accuracy promedio 0.89
- Curva ROC y AUC= 0.82
- Matriz de confusión

  True Negative=11324 False Positive=225
  False Negative=1186 True Positive=360



b- Logistic Regression

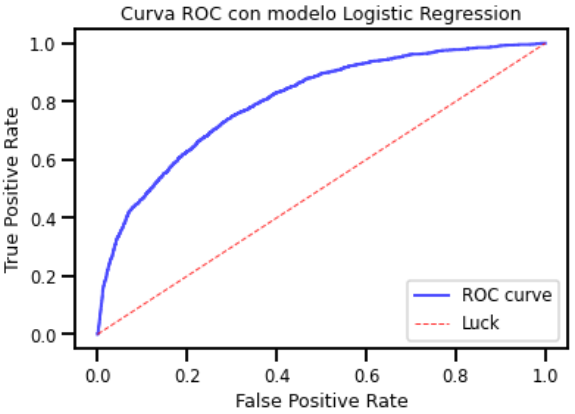
Luego de realizar el cross validation con las siguientes consideraciones:

- Penalizador L1

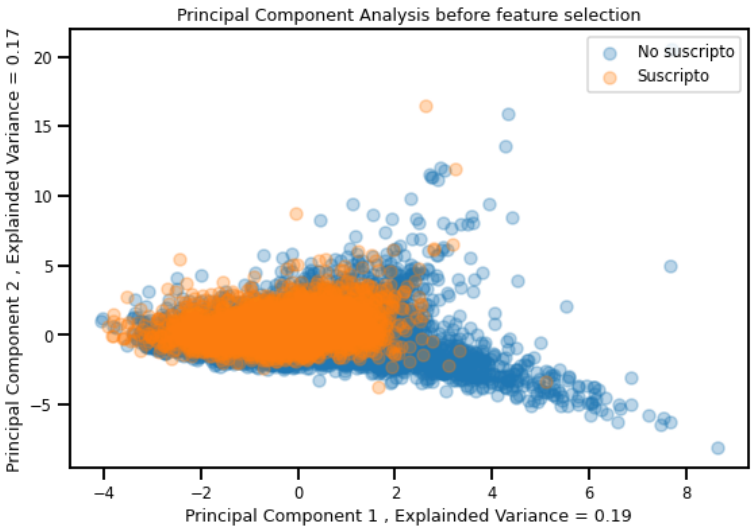
Se obtuvieron los siguientes resultados:

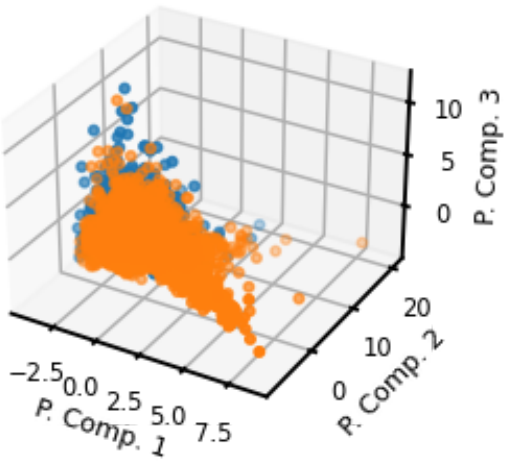
- Accuracy promedio 0.88
- Curva ROC y AUC= 0.80
- Matriz de confusión

  True Negative=11419 False Positive= 130
  False Negative=1320 True Positive=226

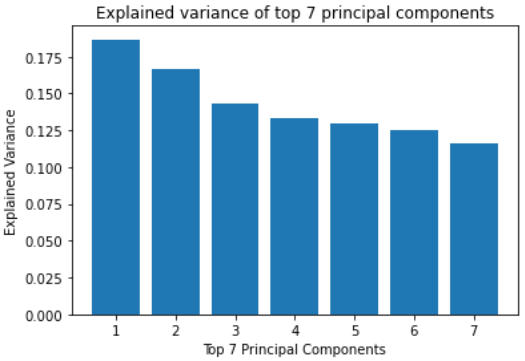


Por otro lado, se realizó una reducción de la dimensionalidad para poder visualizar los datos del dataset en dos o tres dimensiones. El resultado se puede ver a continuación:





Las nuevas features creadas a partir de las features originales, se denominan “Componentes principales”. Las mismas componen un nuevo dataset reducido, el cual se utilizó para entrenar los dos modelos antes mencionados.

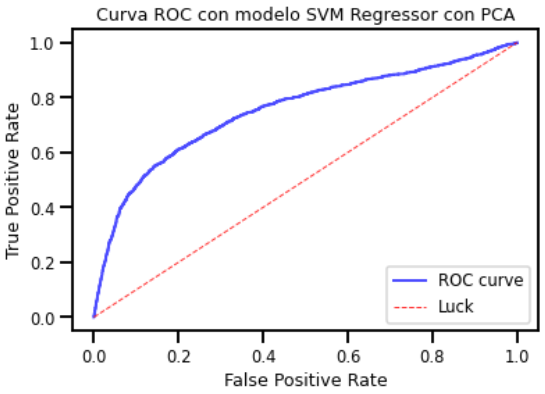


Con los mismos parámetros establecidos en el apartado anterior, se obtuvieron los siguientes resultados:

SVM

- Accuracy promedio 0.8833
- Curva ROC y AUC= 0.7527
- Matriz de confusión

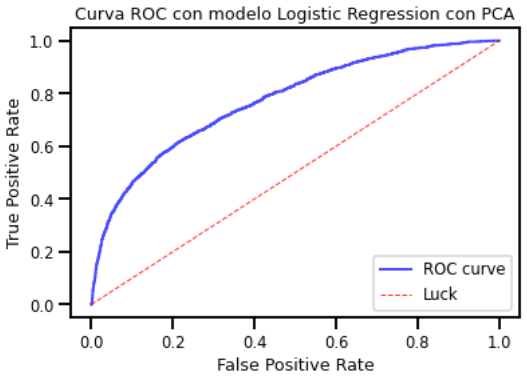
  True Negative=11333 False Positive= 216
  False Negative=1312 True Positive=234



Logistic Regression

- Accuracy promedio 0.88
- Curva ROC y AUC= 0.7732
- Matriz de confusión

  True Negative=11502 False Positive= 47
  False Negative=1527 True Positive=19



Cuadro comparativo de los modelos:

Modelo SVM

	precision	recall	f1-score	support
0	0.91	0.98	0.94	11549
1	0.62	0.23	0.34	1546
accuracy			0.89	13095
macro avg	0.76	0.61	0.64	13095
weighted avg	0.87	0.89	0.87	13095

Modelo SVM con PCA

	precision	recall	f1-score	support
0	0.90	0.98	0.94	11549
1	0.52	0.15	0.23	1546
accuracy			0.88	13095
macro avg	0.71	0.57	0.59	13095
weighted avg	0.85	0.88	0.85	13095

Modelo Logistic Regression

	precision	recall	f1-score	support
0	0.90	0.99	0.94	11549
1	0.63	0.15	0.24	1546
accuracy			0.89	13095
macro avg	0.77	0.57	0.59	13095
weighted avg	0.87	0.89	0.86	13095

Modelo Logistic regression con PCA

	precision	recall	f1-score	support
0	0.88	1.00	0.94	11549
1	0.29	0.01	0.02	1546
accuracy			0.88	13095
macro avg	0.59	0.50	0.48	13095
weighted avg	0.81	0.88	0.83	13095

6- Discusión y conclusiones

Del EDA concluimos que los clientes que poseen créditos y/o préstamos tienen una baja probabilidad de acceder a la campaña. A su vez, los clientes que accedieron a la campaña tuvieron en su mayoría menos de 10 contactos que no superan los 1268 segundos, pudiendo observarse una tendencia con la cual a medida que aumentan los contactos disminuye el tiempo de duración del mismo.

Asimismo, los clientes que habían tenido una campaña con un performance fallida, tuvieron mayor tiempo durante el contacto, con respecto a aquellos con una performance exitosa.

Entre los 2 métodos realizados, se concluyó que el modelo de SVM fue la mejor opción para predecir el dataset presente, ya que la exactitud es superior al modelo de Logistic Regression. A su vez, la curva ROC es levemente superior también en dicho modelo.

Por último, si bien el PCA facilitó la visualización de los dos grupos de clientes, no mejoró ninguno de los dos modelos.

7- Referencias (aunque sea 3 papers y/o libros).

- Python Machine Learning - Sebastián Raschka
- Python Data Science Handbook- Jake VanderPlas
- Pattern recognition and machine learning - Christopher M. Bishop