Universidad Tecnológica Nacional
Facultad Regional Buenos Aires

Ciencia de Datos
Trabajo Final Grupal

Docente de Cátedra: Ing. Martin Palazzo

Docente de T. Prácticos: Ing. Nicolas Aguirre
Ing. Santiago Chas

Curso N°: I5571

Alumno: Correa, Lautaro - 155.959-0
Muñoz, Limber - 149.186-6

Grupo N°: 8

Año: 2024

Introducción y objetivos

El presente trabajo se basa en la aplicación de diferentes técnicas de ciencias de datos sobre un dataset que contiene información sobre aquellos clientes que se han suscrito a una campaña de marketing.

El objetivo consiste en generar un modelo de predicción de clientes que en el futuro se suscribirán a futuras campañas de marketing a partir de las herramientas y métodos desarrollados en la cátedra de Ciencia de Datos de la UTN FRBA.

A lo largo del presente informe se realiza un análisis exploratorio de datos buscando información relevante del dataset en cuestión y luego se pone en práctica un modelo de pipeline de Machine Learning en conjunto a un método de reducción de dimensionalidad.

Todos los desarrollos fueron realizados en Python, a través de Jupyter Notebook y Google Colab.

Descripción del dataset

El dataset bank_suscription tiene las siguientes 17 Features y 45211 filas, cada feature tiene el siguiente significado:

Variable	Significado	Tipo de Variable
Age	Edad del cliente	Float64
Job	Tipo de empleo del cliente	Object
Marital Status	Estado civil	Object
Education	Educación máxima alcanzada por el cliente	Object
Credit	Si tiene deuda de crédito o no	Object
Balance (euros)	Promedio de saldo en la cuenta en el año	Float64
Housing Loan	Si tiene seguro de hogar o no	Object
Personal loan	Si tiene préstamos o no	Object
Contact	tipo con contacto del cliente	Object
Last Contact Day	Último día de contacto con el cliente en el mes	Int64
Last Contact Month	Último mes de contacto con el cliente en el año	Object
Last Contact Duration	Duración del último contacto con el cliente medido en segundos	Float64
Campaign	Cantidad de contactos al cliente durante esta campaña, incluye el último contacto.	Int64
Pdays	Cantidad de días que pasaron desde el último contacto con el cliente de una campaña anterior. -1 significa que no hubo contacto previo	Float64
Previous	Cantidad de contactos previos a esta campaña para cada cliente	Int64
Poutcome	Performance de la campaña de marketing anterior para este cliente	Object
Subscription	**Si el cliente accede a la campania (1) o no (0).**	**Int64**

Del total de registros, un total de 5289 (13.25%) fueron los clientes que aceptaron la campaña de marketing.

Análisis exploratorio de datos

El comienzo del EDA tuvo como objetivo el análisis de la calidad de datos en cada feature.

Al observar las estadísticas descriptivas de las variables nos encontramos con los siguientes puntos de conflicto:

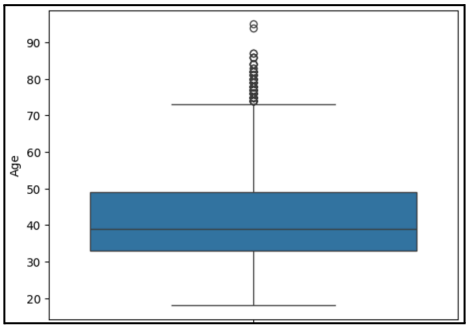
- La variable “Pdays” posee 30.685 filas (67.87%) del total de 45.211 con valor = -1, es decir, no hubo contacto previo. Con lo cual se decide no utilizar la variable en el armado del modelo.

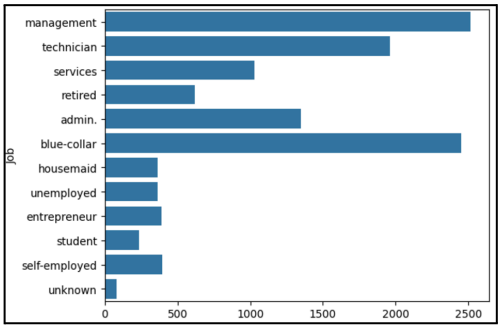
- El feature “Balance (euros)” poseía registros con inconsistencias (valor negativo en una cuenta ahorro). Por lo tanto, se eliminaron dichos registros.
- Una gran cantidad de las variables presentaron valores nulos, con lo que se optó por eliminar los registros.

Luego de efectuar las modificaciones al dataset, el tamaño del mismo fue de 11.755 registros y 15 variables.

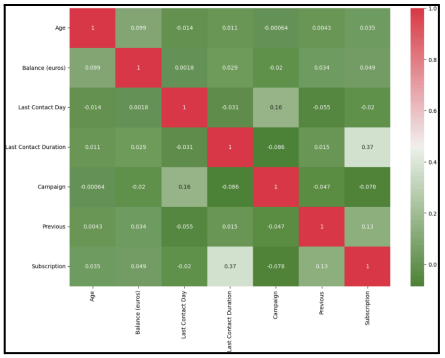
Al generar los Boxplot, se observaron algunas variables con valores Outliers. También se observa muchas features con valores "unknown", tanto en Boxplot como en Countplot.

Se optó por no eliminar más registros a fin de no disminuir la cantidad total de los mismos (11.755 registros de los 45211 iniciales).





Con respecto a la correlación entre variables , la mayor se encuentra entre "Last Contact Duration" y "Subscription" (0.37). Las demás variables poseen bajos valores.



Materiales y métodos (algoritmos utilizados)

1) Logistic Regresion con Pipeline

La Pipeline es una técnica común en el aprendizaje automático porque permite a los científicos de datos automatizar el proceso de preparación de datos, reducir el riesgo de errores y aumentar la reproducibilidad de los resultados.

En el caso desarrollado, los pasos llevados a cabo fueron los siguientes:

- Diferenciación de los tipos de variables (numéricas, categóricas).
- Definición las transformaciones para cada tipo de variable
- Combinación de los transformadores
- Creación del pipeline con clasificador de Logistic Regression.
- Separación del set de datos en conjuntos de entrenamiento y prueba.
- Entrenamiento, predicciones y evaluación del modelo.

El algoritmo de Regresión Logística se trata de una regresión lineal que es precedida por una función de activación sigmoide, generando una salida binaria.

$$p(\mathbf{x}) = \frac{1}{1 + \exp(-f(\mathbf{x}))},$$

2) Principal Component Analysis (PCA)

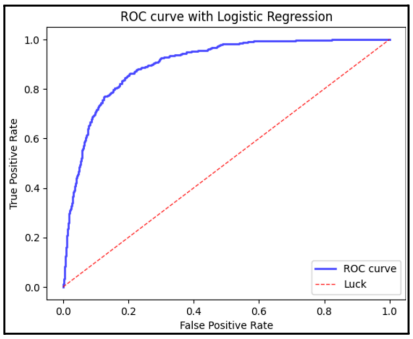
El análisis de componentes principales es un algoritmo utilizado para la reducción de dimensiones en grandes conjuntos de datos identificando componentes principales que reproducen la información original intentando disminuir las distorsiones. Consiste en la descomposición espectral de una matriz de datos a partir de autovalores y autovectores.

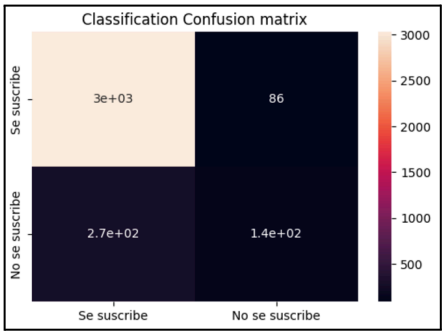
$$\Sigma = v\lambda v^{-1}$$

Tanto para el pipeline como para los modelos de Logistic Regresion (LR) y PCA, se utilizaron los modelos contenidos en la librería de ScikitLearn.

Experimentos y resultados

El modelo de Logistic Regresion a partir del pipeline obtuvo buen rendimiento (Accuracy=0.8985, AUC=0.9012).

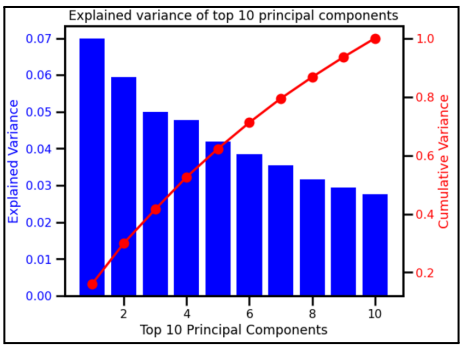


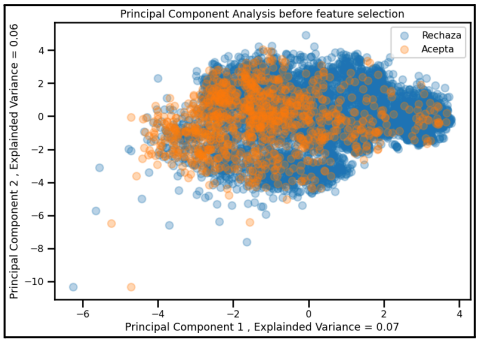


Para el desarrollo del PCA tuvimos que convertir las variables categóricas a dummies del tipo float. Observamos que luego de aplicado el algoritmo la dimensión de nuestro dataset varía de 30 a 10 dimensiones.

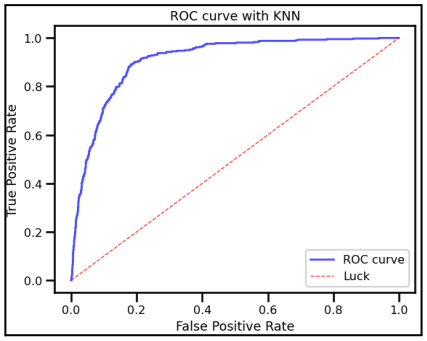
Los componentes obtenidos poseen una varianza "similar", ya que para lograr alcanzar el 80% de la varianza acumulada se requieren 8 de los 10.

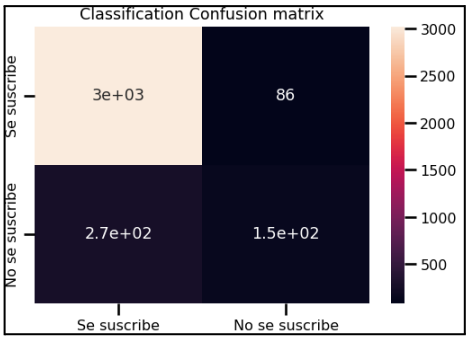
Al elegir los dos componentes de mayor varianza se obtuvo el siguiente scatter plot:





Por último, se volvió a predecir el modelo en base a los 2 principales componentes obtenidos del PCA y mediante Logistic regresion. El rendimiento fue ligeramente superior al modelo anterior (Accuracy=0.8994, AUC=0.9086).





Discusión y conclusiones

Luego de volver a predecir el problema con dimensión reducida del PCA obtuvimos resultados similares (Accuracy inicial= 0.89849730649277 vs Accuracy reducido= 0.8993478877232776).

Es decir se mantuvo la precisión, logrando una gran reducción de la complejidad del problema y aumento de la performance del modelo (cantidad de features iniciales =15 vs features final=2).

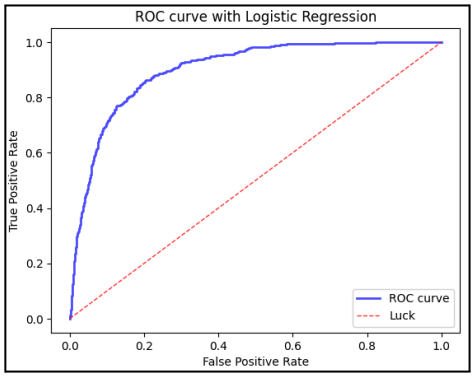
Se concluye que el utilizar métodos de reducción de la dimensionalidad proponen modelos más simples sin perder precisión en la predicción.

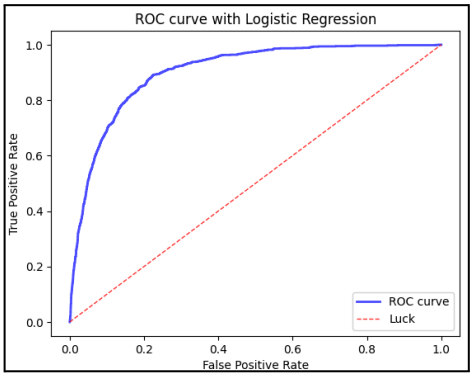
Análisis extra

El dataset utilizado para el desarrollo del modelo de predicción contenía 11.755 registros luego de haber suprimido inconsistencias y todas las filas con valores nulos (vacíos). Se propone un segundo análisis en la Jupyter notebook ESTUDIO_EXTRA completando las filas vacías de algunas features con los siguientes criterios:

Variable	**Significado**
Job	Se reemplaza el valor nulo por “unknown”
Marital Status	Se reemplaza el valor nulo por “Sigle”
Education	Se reemplaza el valor nulo por “unknown”
Credit	Se reemplaza el valor nulo por “no”
Balance (euros)	Se reemplaza el valor nulo por “0
Housing Loan	Se reemplaza el valor nulo por “no”
Personal loan	Se reemplaza el valor nulo por “no”
Pdays	Se reemplaza el valor nulo por “-1”

Al borrar los valores nulos restantes y terminar el EDA se obtuvo un dataset de 30.942 registros. El nuevo modelo obtuvo un Accuracy de 0.8950 ligeramente menor a la del primer análisis (0.8984) pero se obtuvo una mejor AUC (0.9018) como se observa en el gráfico de la derecha.





Con lo cual concluimos que en este caso la reducción de registros no tuvo un gran impacto negativo en la precisión de predicción del modelo Logistic Regression.

Referencias

- Mckinney, (2017). Python para el análisis de datos.
- James, Witten, Hastie, Tibshirani, Taylor (2023). An introduction to statistical learning with applications in Phyton.
- Murphy, Kevin P. (2012). Probabilistic Machine Learning
- Nong Ye (2003). The handbook Data Mining, Mahwa, New Jersey.
- Apuntes cátedra Ciencia de Datos UTN FRBA.