Segunda entrega proyecto final de Data Science

Integrante:

- Reinaldo Barberan

CODER HOUSE

06/05/2023

Contexto comercial.

Ha habido una disminución de ingresos en el Banco Portugués y les gustaría saber qué acciones tomar. Después de la investigación, descubrieron que la causa raíz era que sus clientes no estaban invirtiendo lo suficiente en depósitos a largo plazo. Por lo tanto, al banco le gustaría identificar a los clientes existentes que tienen una mayor probabilidad de suscribirse a un depósito a largo plazo y centrar los esfuerzos de marketing en dichos clientes.

Problema comercial.

¿Determinar cuales son los grupos de personas con mas posibilidad de suscribir (sí/no) un depósito a plazo ?

Contexto analítico.

Los datos están relacionados con campañas de marketing directo de una institución bancaria portuguesa, las cuales se basaron en llamadas telefónicas.

Se requería más de un contacto con el mismo cliente, para poder acceder si el producto (depósito bancario a plazo) estaría suscrito ('sí') o no ('no') suscrito.

Descripción de los datos

Variables

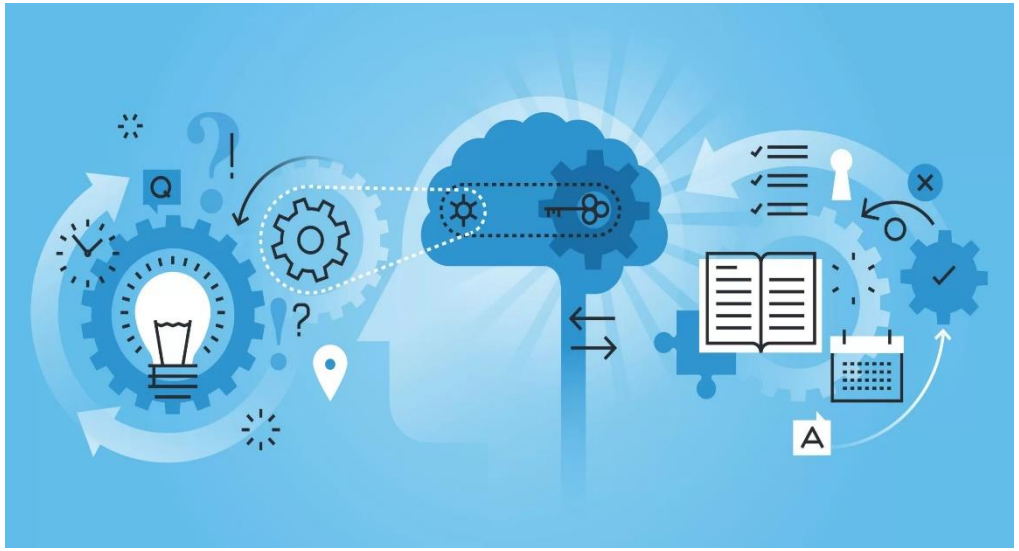
Age: Edad del grupo muestra.
Duration: Duración de llamada por cada encuestado.
Campaign: número de la campaña.
Pdays: Cantidad de días que dura la campaña.

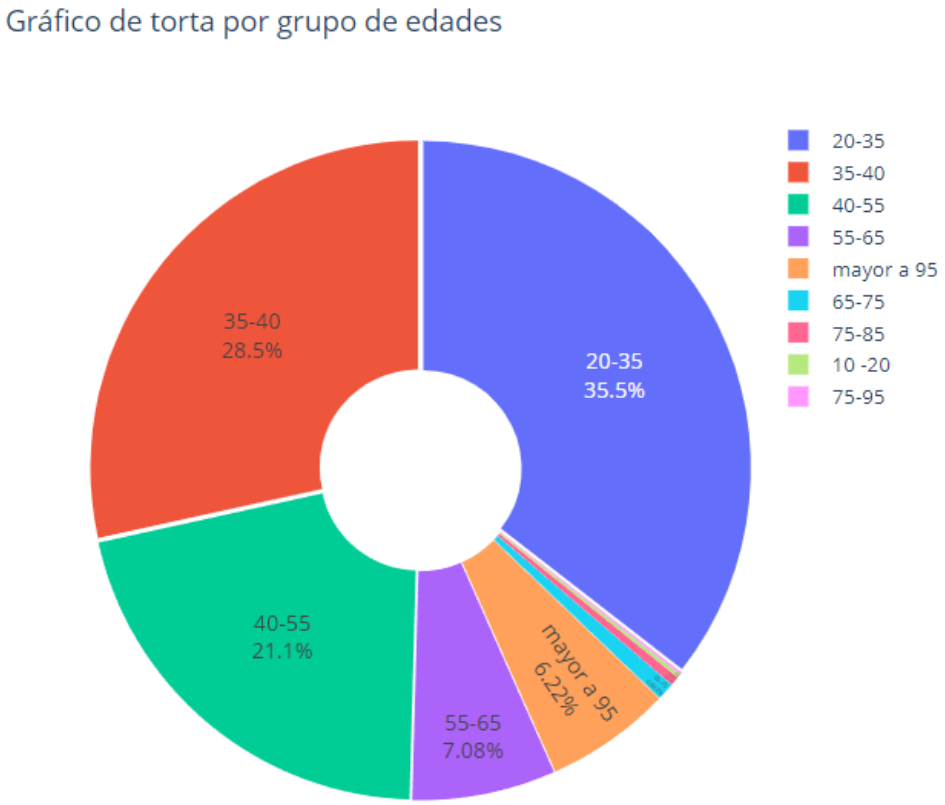
Hay dos conjuntos de datos: train.csv con todos los ejemplos (32950) y 21 entradas que incluyen la función de destino, ordenadas por fecha (de mayo de 2008 a noviembre de 2010), muy cerca de los datos analizados en [Moro et al., 2014] test.csv que son los datos de prueba que consisten en 8238 observaciones y 20 características sin la característica de destino.

Preguntas de hipótesis iniciales

1) ¿Que grupos de edad son los mas propensos a tomar un deposito a plazo. Tomando en cuenta una agrupación por rangos de edad?.

2) ¿Cual es el estado civil con mayor porcentaje en la base?.





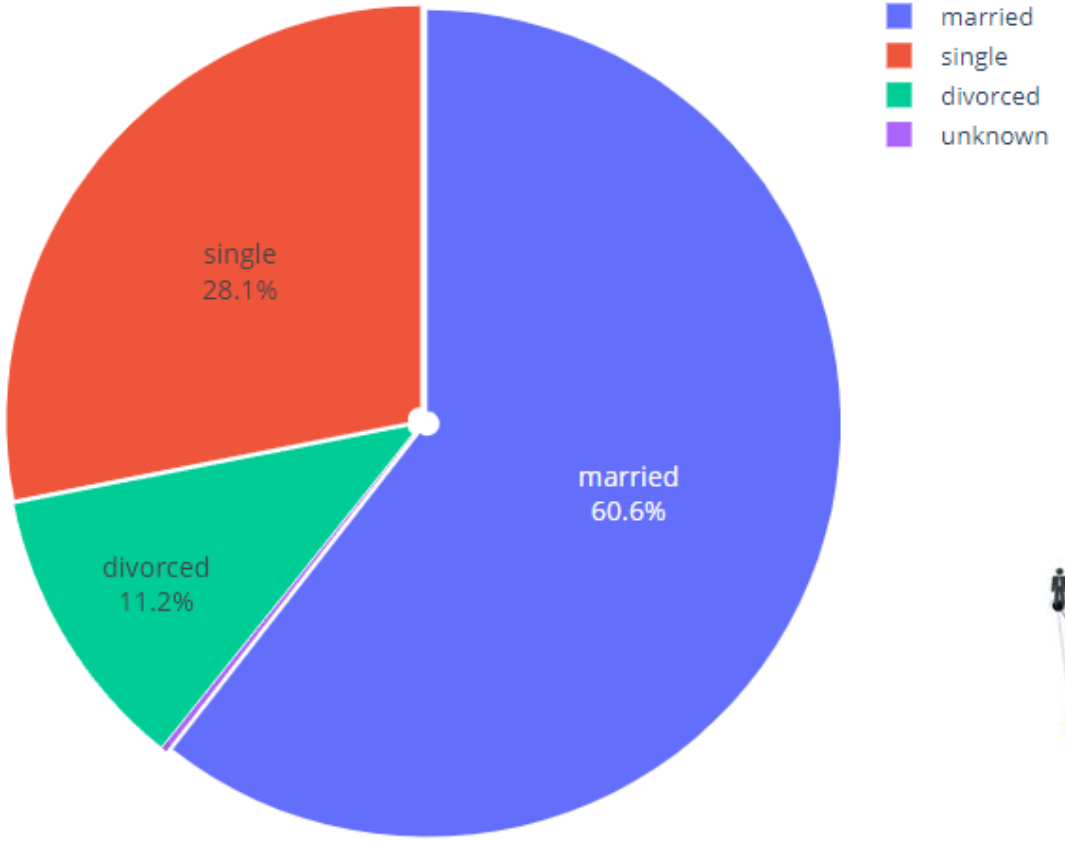
Como primer análisis se puede observar:

Que los grupos predominantes son los de las edades comprendidas entre 20-35, 35-40, y 40-55.

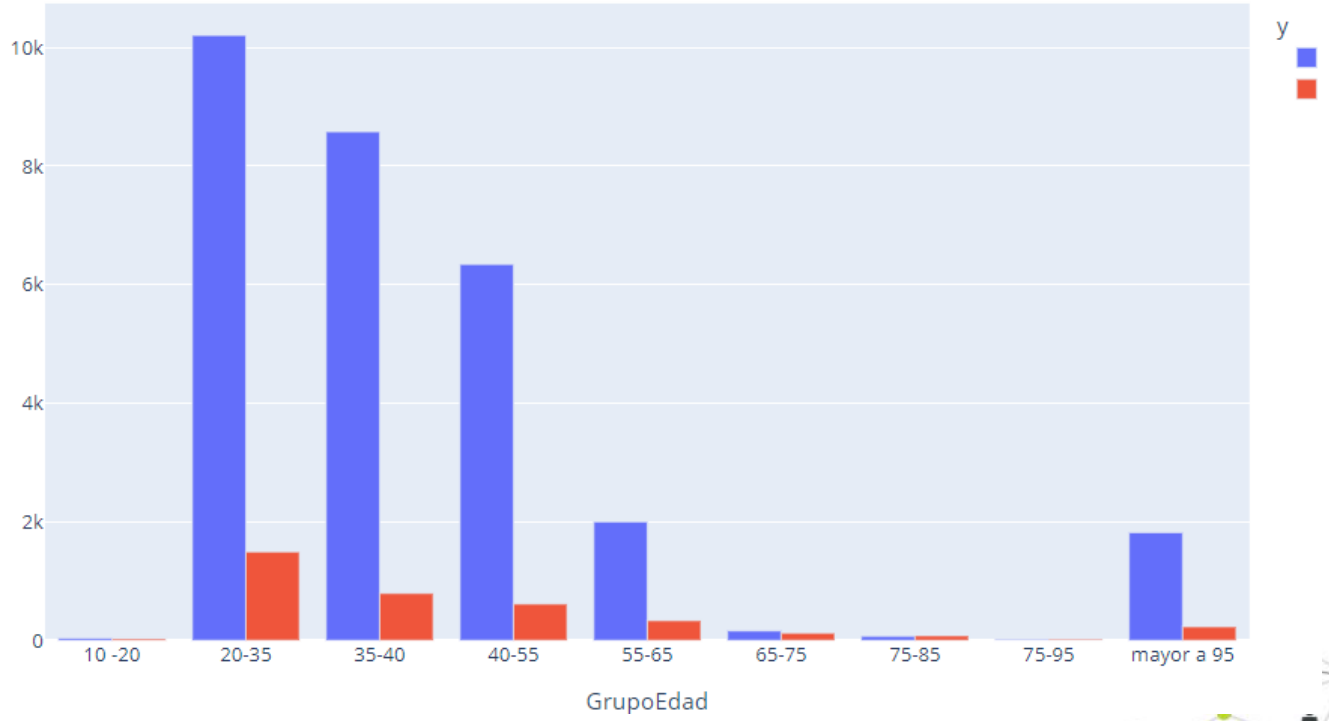
Sin embargo, ahora analizaremos los porcentajes de aceptación por los grupos de edades.

Para no descartar ningún grupo.

Estado civil



Se puede observar que dentro de la base la población de casados es la mayor con respecto a toda la base, donde le sigue los solteros y por último la población de divorciados.



En base de este resultado vamos a realizar un análisis de 3 poblaciones, por grupo de edad para determinar entre cada grupo de edades quienes tienen mas porcentaje de éxito.

Donde será enmarcado de la siguiente manera:

•Población joven edad entre 20 a 40 años.
•Población Adulto intermedio 40 a 65 años.
•Población Adulto mayor 65 a 95 años.

Resumen de hallazgos

edad	poblacion	poblacionTotal	si	no	porcentajeAprobacionPorEdad	porcentajeAprobacionTotal
Joven	17942	31615	2067	15875	11.52%	6.53%
Adulto	13414	31615	1285	12129	9.57%	4.06%
Adulto Mayor	488	31615	229	259	46.92%	0.72%

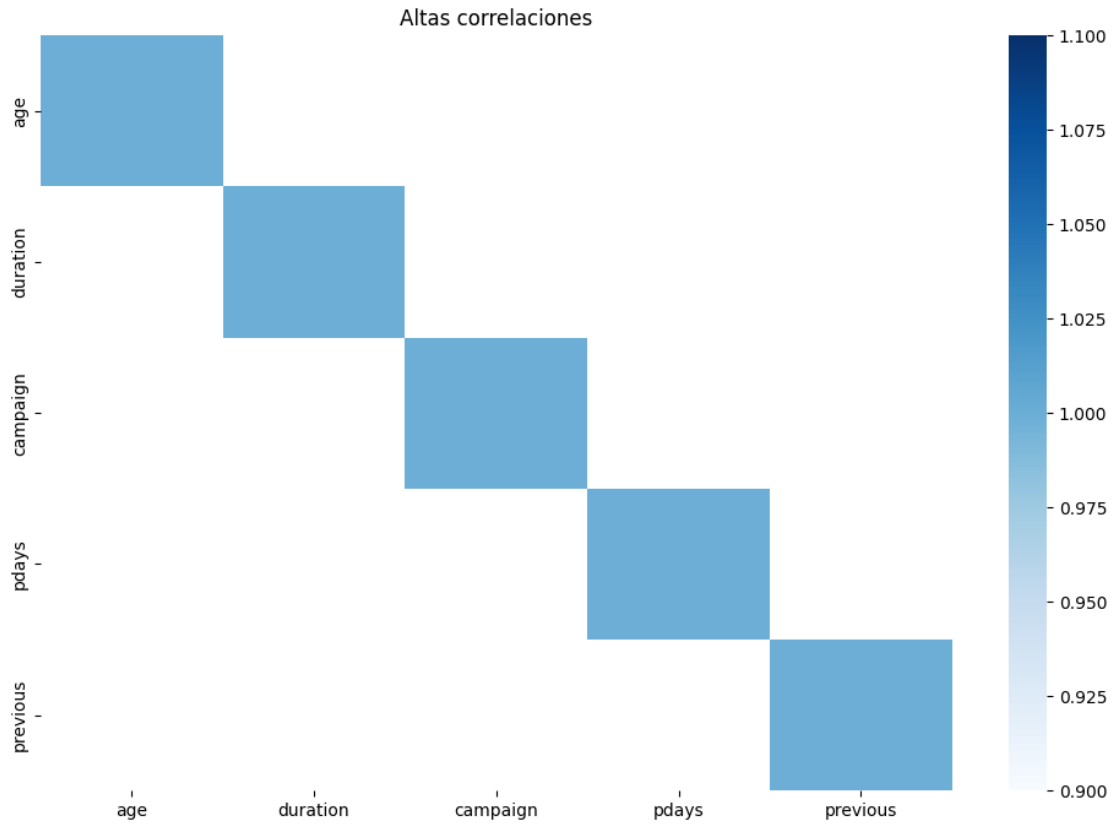
De los datos calculados se pueden obtener varios resultados:

1. Con respecto a la población Joven tenemos un porcentaje de aceptación de 6.53%, que seria mayor a los demás grupos, sin embargo, esto se debe a que su población es mas alta que los demás grupos.

2. La población Adulta tiene un porcentaje de aprobación de 4.06%, siendo el grupo con mayor aceptación.

3. Población Adulto Mayor a pesar de ser un grupo minoritario tiene un buen porcentaje de aceptación.

Matriz de correlación



Se puede observar que la edad y la duración de la llamada tiene una gran relevancia, en los casos de éxito, para la aceptación de la campaña.

Algoritmo elegido

Se realizo el test con los algoritmos de regresión logística y decisión tree, cuando se le suministro la información para su entrenamiento el que tuvo mejor resultado fue el de regresión logística con un accuracy de 0.89.

	Metodo	Accuracy	Precision	Recall	ROCAUC
0	Regresion logistica	0.895072	NaN	NaN	NaN
1	Decision Tree - label yes	0.839118	0.257019	0.208772	NaN
2	Decision Tree - label no	0.839118	0.899297	0.921308	NaN

Conclusiones

Se puede obtener varias conclusiones:

1.- Dentro de las poblaciones elegidas puesta en estudio se concluye que la población con mayor aceptación en la campaña es la joven entre un rango de edad de 20 a 40 años.

2.- La población de adulto y adulto Mayor, aunque es menor a la población joven, su nivel de aceptación de la campaña es tolerable.

3.- El algoritmo de decisión tree, es muy adecuado para este análisis y presenta una gran tasa de rendimiento.

Muchas Gracias!!!