Proyecto II: Análisis campaña de marketing

1. Objetivo

El objetivo de este proyecto es responder a dos preguntas, ***¿Qué tipo de clientes gastarán más en nuestra tienda?*** y ***¿Qué tipo de clientes es más probable que vuelva a comprar?***, desde su última compra

2. Extracción de datos

Estos datos los vamos a obtener de un Dataset preestablecido, de la página "Kaggle"; [Marketing Campaign | Kaggle]

3. Preparación de datos

i. Limpieza de datos

Realizamos un vistazo general de los datos para poder seleccionar las columnas adecuadas para este proyecto. Buscamos valores que no nos interesen, por ejemplo, clientes que sean menores de edad. Seguidamente, buscamos valores nulos de modo a eliminar las filas correspondientes. En nuestro dataset, tendremos una cantidad muy dispersa de datos en muchas de las categorías, es decir, "outliers". Los eliminaremos ayudándonos de gráficos cómo "boxplot", de la librería "seaborn".

ii. Normalización y Dummy indexing

Antes de terminar, tendremos que usar los dummy índices, es decir, reducir las variables categóricas a índices binarios para poder hacer una predicción más limpia. Por último, normalizamos el dataset.

4. Exploración de datos

i. Gráficos Simples

En esta sección, ya tenemos los datos filtrados y procesados, podemos empezar a visualizarlos con gráficos simples, como un "boxplot" o "scatterplot".

ii. Exploración no gráfica

Eventualmente podremos hacer uso de funciones predeterminadas como "describe()" de modo a poder realizar una exploración no-gráfica del dataset

5. Modelización

i. Modelos y variables

Con los datos preparados, podemos iniciar el proceso de modelización. Dado nuestro dataset, elegimos el"target" para la variable a predecir y los "features" como aprendizaje para el modelo. Recalcar, que este proyecto lo abordaremos desde un punto de vista "Supervisado", es decir, proporcionamos al modelo toda la información necesaria. Ya sabemos qué queremos obtener. También podríamos abordar el modelo con redes neuronales.

ii. Ejecución

Dada la naturaleza de nuestro problema, usaremos modelos de regresión. Elegiremos un set de varios tipos y se tendrá en cuenta los que mejores resultados arrojen.

iii. Diagnóstico y comparación

Tras ejecutar los modelos y seleccionar el más óptimo, podremos hacer uso de funciones como "cross_val_score" para saber cuál es la mejor manera de particionar nuestro set "train" y "test" o bien "GridSearchCV" para encontrar los mejores hiperparámetros que se ajustan al dataset de estudio.

5. Presentación

i. Presentación

El modelo ya se ha elegido, mejorado y preformado. Podemos mostrar los resultados en una tabla o gráfico. Lo importante será comparar los datos predichos con los reales.

Resumen del Proyecto

Nuestro proyecto es una campaña de Marketing que se compone de varias columnas dándonos información sobre las compras, inscripciones, medios de compra que se han realizado, según un conjunto de clientes.

Year_Birth	Education	Marital_Status	Income	Dt_Customer	Recency	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases
1957	Graduation	Single	58138.0	2012-09-04	58	3	8	10	4
1954	Graduation	Single	46344.0	2014-03-08	38	2	1	1	2
1965	Graduation	Together	71613.0	2013-08-21	26	1	8	2	10
1984	Graduation	Together	26646.0	2014-02-10	26	2	2	0	4
1981	PhD	Married	58293.0	2014-01-19	94	5	5	3	6

Nuestro objetivo será responder dos preguntas que nos podrán orientar sobre cómo enfocar la campaña de marketing del año siguiente. ***¿Qué tipo de clientes gastarán más en nuestra tienda?*** y ***¿Qué tipo de clientes es más probable que vuelva a comprar?***

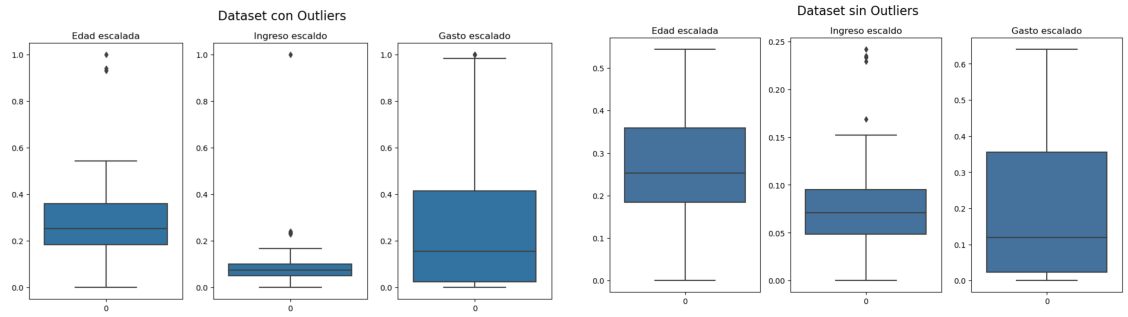
Comenzamos eligiendo las columnas que nos interesan, para resolver la primera pregunta, de manera a obtener un Subdataset.

	Year_Birth	Education	Marital_Status	Income	total_gastado
0	1957	Graduation	Single	58138.0	1617
1	1954	Graduation	Single	46344.0	27
2	1965	Graduation	Together	71613.0	776
3	1984	Graduation	Together	26646.0	53
4	1981	PhD	Married	58293.0	422

Una vez elegido, seguimos filtrando, normalizando, convirtiendo las categorías en dummy índices, etc.

Year_Birth	Income	total_gastado	Edu_2n Cycle	Edu_Basic	Edu_Graduation	Edu_Master	Edu_PhD	Marital_Absurd	Marital_Alone	Marital_Divorced	I
66	4844.833333	67.375000	0	0	1	0	0	0	0	0	
69	3862.000000	1.125000	0	0	1	0	0	0	0	0	
58	5967.750000	32.333333	0	0	1	0	0	0	0	0	
39	2220.500000	2.208333	0	0	1	0	0	0	0	0	
42	4857.750000	17.583333	0	0	0	0	1	0	0	0	

Hacemos uso de la librería “seaborn” y “matplotlib” para poder graficar un diagrama de cajas. Esto nos ayudará a ver si hay “outliers”.



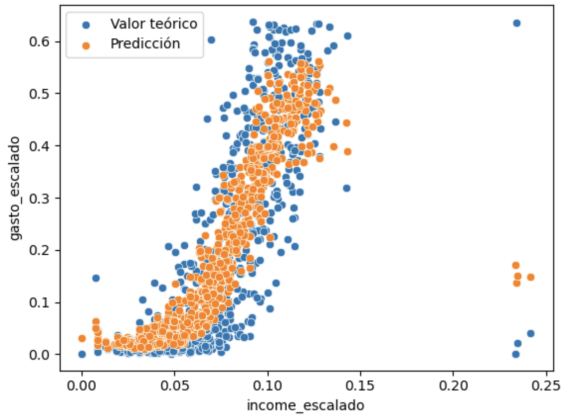
Cómo se puede apreciar, había una gran cantidad de dispersión que modificaría considerablemente el resultado de las predicciones. A pesar de reducir, al máximo, esta dispersión, no se ha podido hacer más debido a que las variables elegidas contienen, en su mayoría, bastantes datos dispersos. Si tratáramos de eliminarlos, nos quedamos con una escasa cantidad de datos.

Tras someter los datos a varios modelos, hemos obtenido los siguientes resultados:

```
Train score Random: 0.9597779268271132
Test score Random: 0.7048151690384266

_____________________________
Train score SVR: 0.5438011501495581
Test score SVR: 0.4779057852087679

_____________________________
Train score Linear: 0.6465020557838839
Test score Linear: 0.5459553725445085

_____________________________
Train score DTR: 1.0
Test score DTR: 0.4952481082071073

_____________________________
Train score Ridge: 0.5058147972427138
Test score Ridge: 0.44421591801742
```

Concluimos con que el modelo de regresión “RandomForestRegressor” es el más adecuado para este dataset. Destacamos la perfección del ajuste del modelo de regresión “DecisionTreeRegressor”, en el modo entrenamiento, sin embargo, el resultado en el “test” es demasiado malo. Podemos atribuir esta disparidad a la cantidad de datos que componen el dataset.

Pasamos a representar los valores teóricos frente a la predicción.



La predicción no se ve del todo mal, con un 70%, sin embargo, destacamos los valores dispersos que, sin duda, torpedean el modelo.

Volvemos a re-escalar el dataset para poder sacar nuestras conclusiones. Creamos una muestra de datos que sea del 10 % de longitud del dataset original y filtramos esta muestra de manera a obtener, únicamente, las filas que presenten un gasto mayor a la media (media de gastos de la propia muestra).

	Edu_2n Cycle	Edu_Basic	Edu_Graduation	Edu_Master	Edu_PhD	Marital_Absurd	Marital_Alone	Marital_Divorced	Marital_Married	Marital_Single	Ma
count	64.000000	64.0	64.000000	64.000000	64.000000	64.000000	64.0	64.000000	64.000000	64.000000	
mean	0.125000	0.0	0.484375	0.187500	0.203125	0.015625	0.0	0.109375	0.421875	0.125000	
std	0.333333	0.0	0.503706	0.393398	0.405505	0.125000	0.0	0.314576	0.497763	0.333333	
min	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	
25%	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	
50%	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	
75%	0.000000	0.0	1.000000	0.000000	0.000000	0.000000	0.0	0.000000	1.000000	0.000000	
max	1.000000	0.0	1.000000	1.000000	1.000000	1.000000	0.0	1.000000	1.000000	1.000000	

Como queremos saber qué tipo de clientes es el que gasta más, nos centraremos en su estado civil y en el nivel educativo que posea. Además, como vemos en la imagen anterior, aquellas columnas que tengan una media igual a 0, no nos interesan. Finalmente, creamos un bucle que recorra todo el dataset, de modo que si se topa, en cada columna, con un 1 (el dummy índice) añadirá el gasto, de esa fila, a una lista.

Una vez recorrido todo el dataset, se realiza una suma total de los gastos para cada categoría. Se concluye:

```
Los gastos generados por la columna Edu_2n Cycle son 203

Los gastos generados por la columna Edu_Graduation son 1081

Los gastos generados por la columna Edu_Master son 1394

Los gastos generados por la columna Edu_PhD son 1758

Los gastos generados por la columna Marital_Divorced son 1946

Los gastos generados por la columna Marital_Married son 2683

Los gastos generados por la columna Marital_Single son 2905

Los gastos generados por la columna Marital_Together son 3436

Los gastos generados por la columna Marital_Widow son 3484
```

Los **clientes que más van a consumir** en nuestra tienda son los **viudos** y los que tengan **estudios de postgrado**.

Seguimos con la segunda parte del proyecto. Respondamos a la pregunta ***¿Qué tipo de clientes es más probable que vuelva a comprar?.***

Volvemos con un Subdataset muy similar al anterior

	Year_Birth	Education	Marital_Status	Income	Recency	total_gastado
0	1957	Graduation	Single	58138.0	58	1617
1	1954	Graduation	Single	46344.0	38	27
2	1965	Graduation	Together	71613.0	26	776
3	1984	Graduation	Together	26646.0	26	53
4	1981	PhD	Married	58293.0	94	422

Esta vez tendremos una columna adicional, "Recency". Es la columna que nos indica cuántos días han pasado desde la última compra. Es decir, nos interesa el mínimo valor posible. Después de filtrar, normalizar, etc…, el nuevo dataset, filtramos las edades objetivo que permitirán responder a nuestra pregunta.

Nuestro público objetivo será una clientela menor de 60 años. Esta vez, realizaremos una búsqueda exploratoria de datos mediante comandos como "groupby", "value_counts" y "idxmin/idxmax".

Agrupando el nivel de educación, el estado civil y la edad junto a los gastos y realizando las medias de cada uno, obtenemos:

	Recency	Recency	Recency
2n Cycle	7.740741	0.000000	0.000000
Basic	7.625000	0.000000	0.000000
Graduation	7.009524	0.000000	0.000000
Master	7.239130	0.000000	0.000000
PhD	6.777778	0.000000	0.000000
Alone	0.000000	12.000000	0.000000
Divorced	0.000000	7.526316	0.000000
Married	0.000000	6.979798	0.000000
Single	0.000000	6.333333	0.000000
Together	0.000000	7.771930	0.000000
Widow	0.000000	13.000000	0.000000
YOLO	0.000000	3.000000	0.000000
29	0.000000	0.000000	11.000000
30	0.000000	0.000000	12.000000
31	0.000000	0.000000	10.500000
33	0.000000	0.000000	4.500000
34	0.000000	0.000000	7.200000
35	0.000000	0.000000	11.600000

En la primera columna tendremos los datos relativos al nivel de estudio, la segunda al estado civil y la última a la edad. Concluimos con que **el cliente que menos espera en volver a comprar** es el que tiene un nivel de **estudio de postgrado**, un **estado civil Yolo** y de **33 años**.