Práctica 2: Limpieza y análisis de datos - Black Friday

Alicia Escontrela y Beatriz Figueroa Martínez

1. Descripción del dataset.

El dataset que vamos a analizar recoge datos de ventas durante el Black Friday. Contiene información sobre el comprador como: edad, estado civil, ocupación etc. Los productos los clasifica por categorías y además indica el importe de compra.

El Black Friday es un caso particular dentro de las ventas de una empresa por su corta duración y la expectación que suscita por los buenos precios y lo cercano que está a la Navidad. Las tiendas y grandes almacenes necesitan una previsión de los artículos que van a vender, así como a qué tipo de cliente lo hacen.

Con nuestro estudio, intentaremos relacionar los atributos de clientes con las categorías de producto para encontrar patrones de consumo que permitan tomar decisiones en las próximas campañas de marketing.

2. Integración y selección de los datos de interés a analizar.

En nuestro caso no fue necesaria la integración porque todos los datos estaban almacenados en solo un fichero .csv

Con respecto a la selección de datos, en la primera exploración que realizamos consideramos que los 12 atributos serán de utilidad para realizar el análisis. Por tanto, decidimos no eliminar ninguno de ellos en esta etapa.

Observamos que 11 atributos son categóricos y solo uno continuo. Al analizar los datos confirmamos que los atributos categóricos sean identificados como tal en Python para su posterior análisis.

3. Limpieza de los datos.

3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

El atributo Product_Category_2 tiene 166.986 valores nulos y el atributo Product_Category_3 tiene 373.299.

Para entender el problema un poco mejor, analizamos cuantos registros tienen valores nulos tanto en los atributos Product_Category_2 como en Product_Category_3, para lo cual obtenemos el mismo valor que el número de nulos de Product_Category_2 (166986).

Por tanto, tenemos 166986 registros de los cuales no conocemos si hay productos que pertenecen a las categorías 2 y 3.

Como no tenemos suficiente información de los productos que corresponden a cada categoría, no conocemos el significado de los niveles dentro de cada categoría y tenemos valores nulos en la categoría 2 y 3, decidimos retirar estos valores del estudio porque consideramos que no nos aportan suficiente información de interés para el análisis.

3.2. Identificación y tratamiento de valores extremos.

Como ya hemos comentado, la única medida que existe en el dataset es Purchase que indica el valor de la compra realizada. Hemos comprobado si existe algún valor que se separe sustancialmente del resto de la muestra.

En primer lugar, representamos el diagrama de cajas para la muestra completa del campo Purchase. En él podemos ver que considera **valores atípicos** los mayores de 21.384.

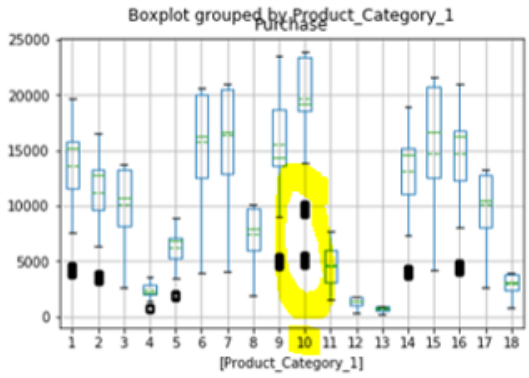
Este dato se calcula de la siguiente manera:

$X = Q_3 + 1.5 * (Q_3 - Q_1)$

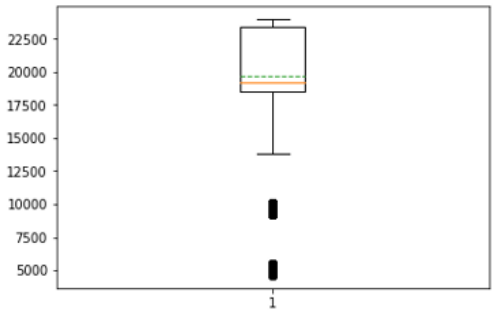
$21.384 = 12.073 + 1,5 * (12.073 - 5866)$

Si en lugar de utilizar el parámetro 1,5 utilizamos 3, obtendremos los **valores extremos** que serían los mayores de 30.694. Ningún valor de Purchase cumple esta condición así que podemos decir que los valores atípicos no son extremos. Esta afirmación unida a que son no hay señales de que haya ningún error en la recogida, nos hace pensar que no sería correcto eliminarlos de la muestra.

De todas formas, hemos profundizado un poco más en la detección de outliers y los hemos analizado según los valores que toman en otros campos.



Nos fijamos en concreto en el valor 10 de “Product_Category_1” y vemos que tiene datos atípicos más alejados del gráfico. Analizamos este caso independientemente.



Realizamos el mismo cálculo que en el caso anterior. Podemos ver que considera **valores atípicos** los menores de 11.206. Este dato se calcula de la siguiente manera:

$X = Q_1\text{-}1.5 * (Q_3 - Q_1)$

11.206= 18.545 + 1,5 * (23.438 – 18.545)

Si en lugar de utilizar el parámetro 1,5 utilizamos 3, obtendremos los **valores extremos** que serían los menores de 3.866. Ningún valor de Purchase con categoría 1 = 10 cumple esta condición así que podemos decir que los valores atípicos en este caso tampoco son extremos.

Por lo tanto, no vamos a eliminar ningún valor atípico de la muestra.

4. Análisis de los datos.

4.1. Selección de los grupos de datos.

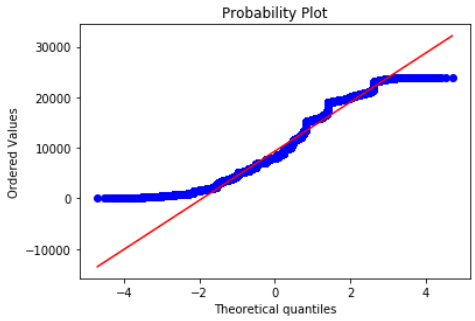
Nuestro objetivo es relacionar los atributos de clientes con las categorías de producto para encontrar patrones de consumo que permitan tomar decisiones en las próximas campañas de marketing.

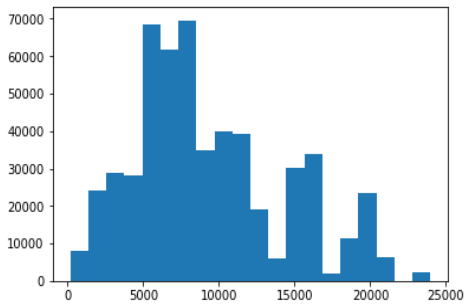
Por tanto, utilizaremos métodos de clasificación para determinar el tipo de clientes de acuerdo al tipo de consumo y a su edad, métodos de clustering para segmentar los clientes y regresión para predecir el consumo que realizarán los clientes teniendo en cuenta los datos registrados durante esta campaña y poder aplicar esta información en el diseño de nuevas campañas de marketing.

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Para realizar la comprobación de la normalidad, aplicamos el test de shapiro sobre Purchase, que es el único atributo continuo. Al realizarlo, obtenemos un p-value < 0.05. Por tanto, rechazamos la hipótesis nula y no asumimos normalidad.

Por otra parte, el qqplot y el histograma también nos indican que la distribución no parece ser normal.





En el primero observamos que el gráfico de dispersión no encaja con la línea recta que indica la normalidad.

De igual forma, el histograma no se representa como una distribución normal. Por tanto, **no asumimos normalidad** para aplicar los test estadísticos indicados a continuación.

Con respecto al análisis de la **homogeneidad en la varianza**, aplicamos el test no paramétrico de Fligner-Killen para analizar las varianzas de la selección de la variable Purchase para las diferentes categorías, encontrando que los siguientes grupos presentan varianzas similares con respecto a su nivel de Purchase:

- Personas entre 36-45 años y entre 51-55 años
- Ambos valores del estado civil .

Mientras que el resto de grupos no presentan varianzas similares.

Al no presentar distribución normal ni homocedasticidad en la mayoría de los casos, no podremos aplicar métodos de contraste de hipótesis de tipo paramétrico.

Al tratarse de variables categóricas, hemos aplicado el **Test de Chi cuadrado** para buscar dependencias entre atributos relacionados con el cliente (edad, género, ocupación...) y la categoría de producto que consume.

Para todas las variables analizadas, los valores de p son 0 o próximos a 0, por lo que existe dependencia y, por lo tanto, se ven diferencias entre las compras realizadas por cada uno de los valores de cada atributo.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos.

Clasificación:

En el histograma de Purchase observamos que puede haber 2 patrones de consumo, uno de los clientes que gastan menos de 14.000$ y otro tipo de clientes que gasta más. Por tanto, definimos una nueva variable que clasifique a los clientes como estándar si consumen menos de 14.000$ y Premium si gastan más. Esta técnica nos permitirá clasificar los clientes y por tanto enfocar este tipo de campañas a ofertas más personalizadas.

Para ellos, utilizamos 2 algoritmos y comparamos sus métricas de validación. Uno de los algoritmos que hemos utilizado es **KNN**, que es no paramétrico y permite clasificar el tipo de cliente en función de la similitud de atributos. Para verificar el k, hacemos varias pruebas obteniendo que la mayor precisión se obtiene con k=9. También aplicamos **Random Forest**, el cual construye múltiples árboles de decisión, obteniendo la clase más frecuente de cada árbol. Al comparar ambos métodos, observamos que con Random Forest obtenemos una mejor precisión.

También aplicamos este método para clasificar a los clientes por edades. En este caso obtenemos una alta precisión con Random Forest (0.98). De esta forma, podemos enfocar nuevas campañas de marketing personalizadas a clientes en los diferentes rangos de edades.

Regresión:

Intentaremos predecir el consumo realizado por los clientes, teniendo en cuenta el consumo registrado en este dataset. Esto será de utilidad para dimensionar las tiendas y el stock necesario para cubrir esta demanda para este tipo de campañas.

Para ejecutarlo, aplicamos dos métodos de ensamble como lo son **Random Forest** y **Gradient Boosting**. El primero de ellos construye múltiples árboles de decisión, obteniendo como resultado el valor de consumo promedio de los árboles individuales, realizando el entrenamiento en paralelo, mientras que el Gradient Boosting lo realiza de forma secuencial, calculando los árboles de decisión en función de los datos obtenidos del entrenamiento del árbol anterior.

Al comparar ambos modelos, observamos que el valor de r2 es bastante similar en ambos métodos, obteniendo un valor ligeramente superior al aplicar Random Forest.

https://towardsdatascience.com/basic-ensemble-learning-random-forest-adaboost-gradient-boosting-step-by-step-explained-95d49d1e2725

Clustering:

Hemos adaptado los datos de nuestro dataset para poder aplicar el método **K-means** e intentar agrupar los clientes por sus preferencias de compra.

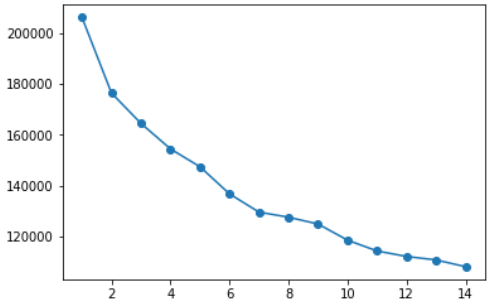
Buscamos agrupaciones de datos midiendo la distancia entre las observaciones. Las variables categóricas las transformaremos en paramétricas para poder medir distancias entre ellas utilizando el método One-Hot-Encoder.

En primer lugar, agrupamos los datos por cliente y nos quedamos con sus atributos. Con el método de codificación generamos una columna por cada posible valor del atributo y rellenamos con 0 si no le aplica y con 1 si le aplica. Hacemos los mismos con los 18 valores que puede tomar la Categoría 1 de producto.

Una vez tenemos los datos en forma paramétrica y unificados, utilizamos el método del codo para decidir el número óptimo de clústeres. Este método consiste en

representar gráficamente la suma de las distancias de cada objeto a su centroide, respecto al número de clústeres elegido. En dicha gráfica buscaremos el valor del clúster que forme el “codo”.

En nuestro caso, no se aprecia con claridad:



Para validar si el clustering es óptimo, utilizaremos el método Silhouette que compara la distancia al clúster asignado con la distancia al clúster no asignado más cercano. El valor deseado sería 1 en el que los clústeres estarían perfectamente delimitados.

Hemos ejecutado este método para el caso de K (número de clústeres) desde 1 a 13 y todos los resultados obtenidos están cerca de 0, lo que significa que los clústeres están solapados. Por este motivo, **no encontramos agrupaciones de registros significativas**.

5. Representación gráfica.

Hemos realizado gráficas de boxplot, histogramas, qqplot, gráficos de barras, los cuales se pueden encontrar junto con el código y el análisis realizado en el repositorio indicado en el punto 7.

6. Resolución del problema.

Durante la ejecución del análisis hemos encontrado algunas limitaciones en el dataset. Por ejemplo, hemos descartado valores nulos por no tener suficiente información para inferir su valor. Tampoco está clara la definición de algunos atributos como por ejemplo a qué se refieren las categorías y sus valores.

Al aplicar métodos supervisados encontramos un modelo con alta precisión que nos facilitará la adaptación de nuevas campañas de marketing para las edades. Sin embargo, hemos obtenido modelos menos precisos como lo son los modelos de regresión, lo cual nos dificultará predecir el consumo teniendo en cuenta los registros de Purchase actuales. Tampoco hemos obtenido una segmentación clara del tipo de clientes.

En general, podríamos decir que el análisis realizado nos permitiría mejorar las campañas de marketing para el nuevo Black Friday. Sin embargo, para tomar mejores decisiones necesitaríamos contar con un conjunto de datos con mayor calidad del dato, para lo cual podríamos proponer a la tienda la mejora en los sistemas para evitar la pérdida de información

y poder almacenar datos que nos permitirán un análisis más precisos para futuras campañas de marketing.

7. Código

Hemos decidido trabajar en Python porque todavía no tenemos mucha experiencia con R. El código se puede encontrar en el siguiente repositorio

https://github.com/aliescont/BlackFriday

Contribuciones	Firma
Investigación previa	AER, BFM
Redacción de las respuestas	AER, BFM
Desarrollo código	AER, BFM