Recomendaciones de moda personalizadas de H&M

Por: Miguel López Vélez, María Fernanda Villarreal Teherán.

En este proyecto se busca predecir qué artículos comprará cada cliente en el período de 7 días inmediatamente después de que finalicen los datos de capacitación. Los clientes que no hayan realizado ninguna compra durante ese tiempo quedan excluidos de la puntuación.

Para lograr esto lo primero que se realizó fue una lectura de los archivos que nos entregaban, en este caso artículos, clientes y transacciones, teniendo en cuenta que el archivo más relevante para nosotros es el de transacciones ya que nos interesa conocer a los clientes más frecuentes.

Las transacciones cuentan con datos como ID del cliente, ID del artículo, precio y canales de venta a los cuales posteriormente se les graficó el histograma obteniendo las imágenes de la figura 1.

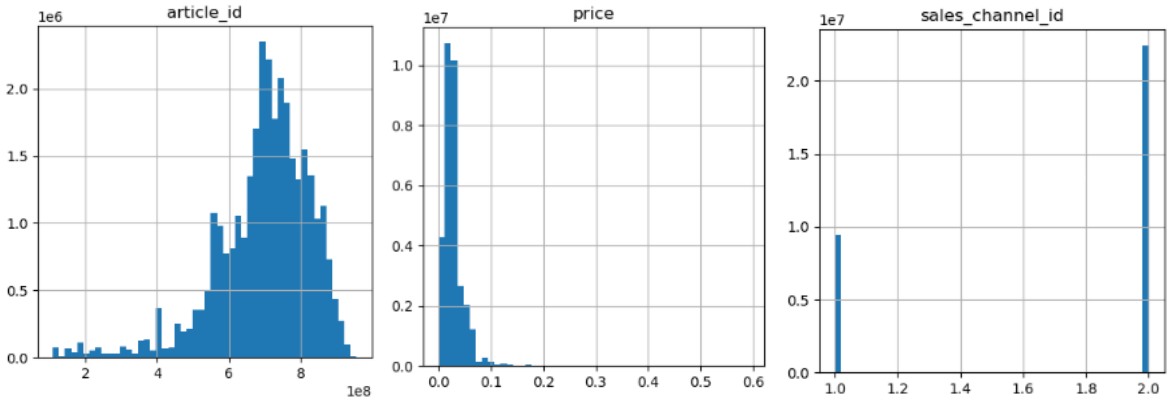

Figura 1. Histogramas de los datos de las transacciones.

Teniendo en cuenta el histograma y el dato encontrado de simetría de -1.26 se puede concluir que hay una distribución asimétrica hacia la derecha, también se hallaron valores de la moda para conocer el valor que aparece con mayor frecuencia en cada conjunto de datos, dando como resultado la imagen de la figura 2.

	t_dat	customer_id	article_id	price	sales_channel_id
0	2019-09-28	be1981ab818cf4ef6765b2ecaea7a2cbf14ccd6e8a7ee9...	706016001	0.016932	2

Figura 2. Moda de cada conjunto de datos.

Posteriormente, se buscan los 10 valores más representativos en las categorías: artículos más vendidos, clientes que más compran, fechas de mayores ventas, tipo de producto que más se fabrica y colores, obteniendo las gráficas mostradas a continuación.

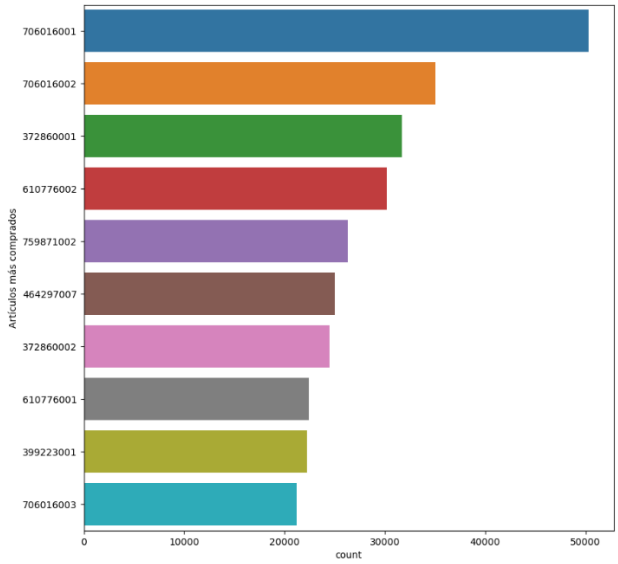

Figura 3. Artículos más vendidos

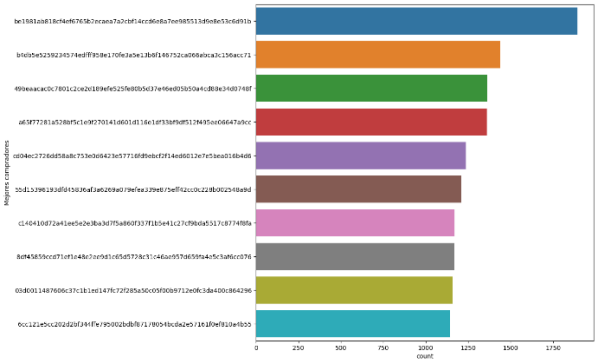

Figura 4. Mejores compradores

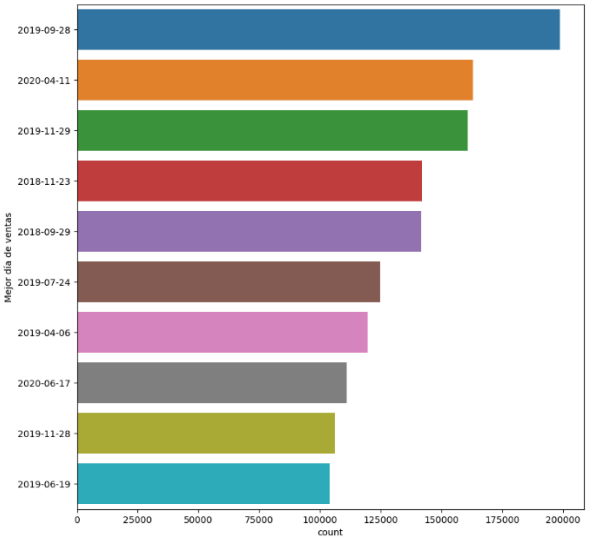

Figura 5. Fechas de mayores ventas

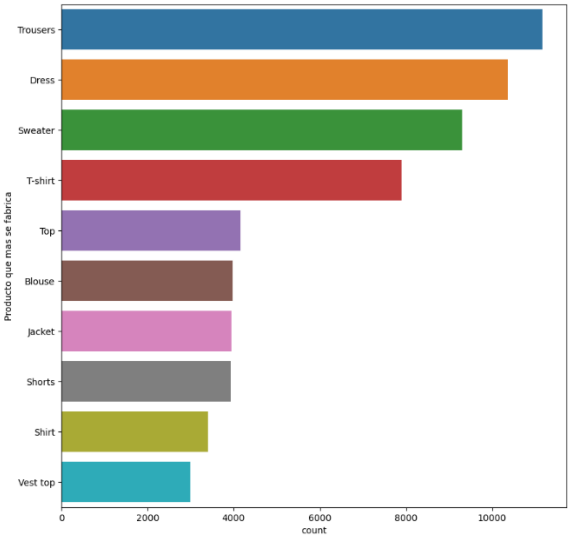

Figura 6. Tipo de producto que más se fabrica

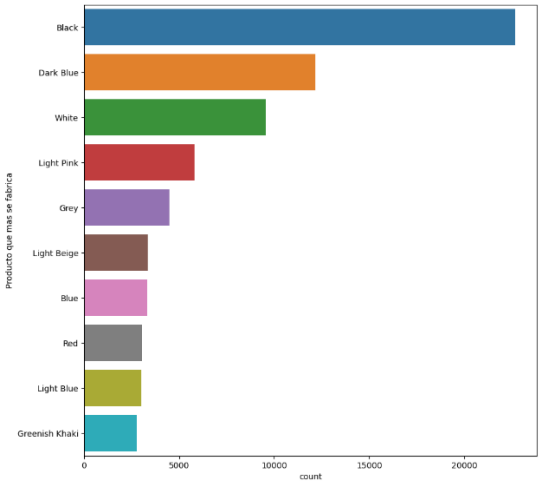

Figura 7. Colores que más se utilizan

Para poder saber si existen datos faltantes primero se debe analizar el dataset de los archivos de ventas, clientes y artículos, en este caso se crearon artificialmente el 5% de datos faltantes en columnas de interés, para lograr esto primero se desorganiza el dataset, luego se elimina el 5% de cada variable con la función “sample()” para finalmente restablecer el índice del DataFrame con la función “reset_index()”.

Para agrupar los datos y eliminar las variables primero se hace una correlación entre las compras y los artículos de cada cliente para generar un dataframe, que resuma los artículos comprados por cliente, sin embargo, al ser un dataset tan grande se reducen el número de transacciones.

Cuando se tenían los datos faltantes se procedió a rellenar estos campos vacíos, en el caso del artículo y el “fashion news” se hizo con la moda y los datos de “FN” y “active” con 0, mientras que, la edad se rellenó con el promedio.

Lo siguiente que se realizó fue un filtrado de datos pensando en cuáles variables son las que más podrían darnos información sobre las posibles compras futuras, entre las variables se filtraron, el tipo de artículo, el color y la sección a la que pertenecen, teniendo en cuenta que estas variables también podrían arrojar información sobre el género del cliente.

Finalmente se utiliza la función “LabelEncoder” la cual codifica etiquetas de una característica categóricas en valores numéricos entre 0 y el número de clases menos 1 y por último, se separan los datos del test y de train para ser utilizados próximamente.

Se crea una nueva columna de recompra utilizando los valores repetidos basándonos en los clientes y los artículos comprados, para luego realizar una matriz de correlación como se puede ver a continuación.

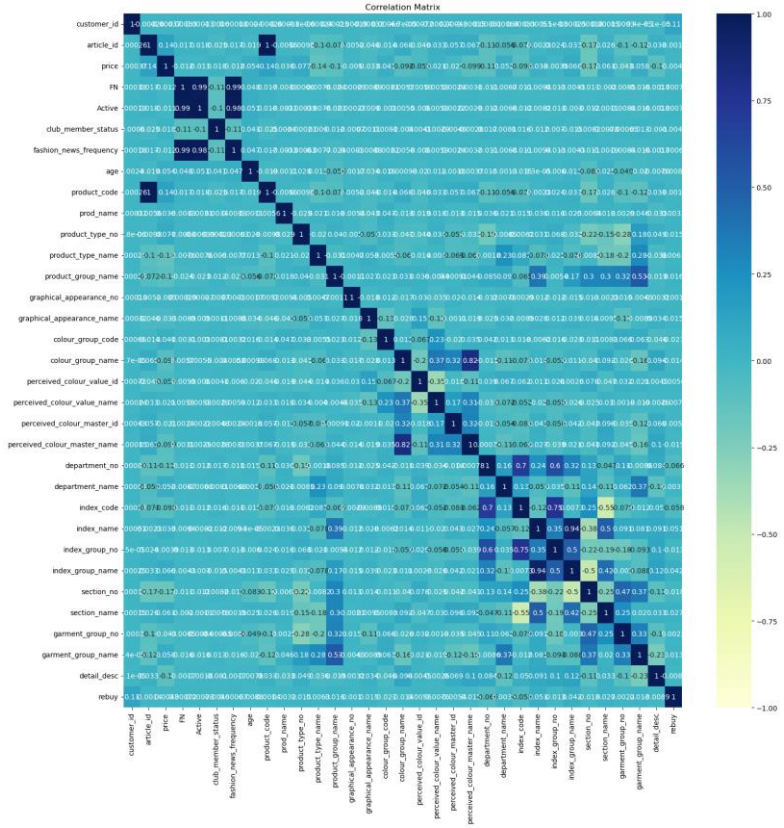

Figura 8. Matriz de correlación para rebuy

En este caso no se puede identificar bien los parámetros, pero los mejores valores se pueden observar con respecto a la siguiente tabla:

	rebuy
rebuy	1.000000
customer_id	0.111928
index_name	0.051096
index_group_name	0.041857
section_name	0.027317
product_group_name	0.016342
graphical_appearance_name	0.015439
product_type_no	0.015132
garment_group_name	0.013455
product_type_name	0.006256
perceived_colour_value_id	0.005583
department_name	0.003455
prod_name	0.003209
garment_group_no	0.002234
graphical_appearance_no	0.001035
Active	0.000782
FN	0.000717
fashion_news_frequency	0.000673
article_id	-0.001419

Tabla 1. Valores de correlación para recompra

Así tomando las mejores 11 columnas se exportan los datos a un cvs para su futuro uso en la generación de modelos.