Estudio de la venta de bebidas alcohólicas en plataforma de e- commerce a través de la aplicación de aprendizaje no supervisado

Lucas Chicco, Conrado Ochoa y Agustín Velurtas
Catedra de Ciencia de Datos – UTN FRBA – Cluster AI

Abstract

El presente informe consiste en el estudio y análisis de la venta bebidas alcohólicas, principalmente cervezas, y productos relacionados a las mismas a través de una plataforma de comercio electrónico.

El objetivo será comprender en gran medida el comportamiento de compra/venta dentro de dicha plataforma de Ecommerce y el hallazgo de patrones y comportamientos en los registros de del portal a través de la aplicación de algoritmos de aprendizaje no supervisado sobre los datos.

Se iniciara de un set de datos con el registro de todas las compras efectuadas en un período determinado en dicha plataforma, al cual se le hará un tratamiento previo que tendrá como finalidad adecuar el set de datos al análisis extrayendo la información que no sea relevante.

Finalmente, se aplicará un modelo de aprendizaje no supervisado con el fin de determinar grupos similares entre las muestras, a los cuales finalmente se les aplicará un último análisis exploratorio para obtener conclusiones.

2 DESCRIPCION DEL DATASET

El set de datos utilizado es de carácter privado, y consiste en compras efectuadas a través de un sitio web de comercio electrónico. Cada registro (de ahora en más: "sample") del mismo representa una operación de compra, incluyendo en cada caso campos (de ahora en más: "features") con información detallada sobre distintos aspectos de ella (económicos, transaccionales, relacionados al producto, y relacionados a las características del usuario). Se han contabilizado un total de 359388 samples con 144 features cada una.

Sin embargo, durante el primer análisis se descubre que la gran mayoría de las features no cuenta con información (principalmente las asociadas al enriquecimiento de datos a partir de la información personal) o muestran información repetida, por lo que se procede a quitar las mismas quedándonos únicamente con aquellas que poseen información relevante sobre la operación de compra.

3 ANALISIS EXPLORATORIO DE DATOS (EDA)

El objetivo de dicho análisis consiste en comprender el comportamiento de las compras en Ecommerce, lo cual implica saber qué productos son los más adquiridos, a que rubro y marca pertenecen, la estacionalidad de los mismos, y cantidad de reincidencias en la plataforma por usuario, entre otros. Para ello se procede a llevar a cabo un proceso de limpieza del dataset en el que se eliminarán aquellas features que contengan valores nulos o una cantidad de valores nulos que justifique su extracción y aquellas que no aporten información relevante para el análisis posterior. De esta manera se extraen 114 features, quedando un dataset de 27 features, con la posibilidad de quitar otras a lo largo del análisis.

En segundo lugar, se analizará la existencia de outliers dentro de distintas features, los cuales también se procederá a quitarlos debido a que corresponden a compras "de prueba" frente al lanzamiento de nuevos productos o campañas.

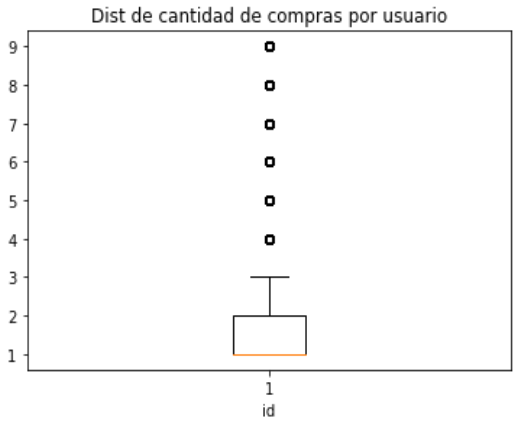
Luego, se procede a ordenar el dataset por fecha de compra, con lo cual puede observarse que el primer registro corresponde a febrero de 2016, y el último a agosto de 2018. Dado que 2017 es

el único año integro con registros se realizará un filtrado del dataset para trabajar únicamente con registros de dicho año. Esto se debe a que en 2016 se comenzó a llevar registro de las compras realizadas, haciendo que los datos no sean representativos en comparación con los de 2017 y 2018. Luego, la extracción de los registros correspondientes a 2018 se debe a que, como el último registro es de agosto de dicho año (en lugar de ser diciembre de modo que sea año completo), se desvirtuarían los resultados a la hora de evaluar estacionalidad.

Una vez definido el dataset para 2017, se realiza un filtrado por país, de los cuales se analizarán los samples correspondientes a las compras en Argentina. Adicionalmente, se extraen nuevos outliers encontrados. Luego del filtrado y de las extracciones quedará un dataset de 57761 samples.

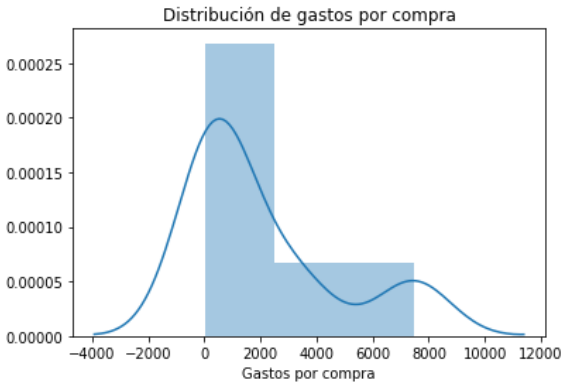
Ya se posee un dataset definido con el que se desarrollarán los distintos análisis. Se procede a analizar los productos más vendidos, pudiendo hallar que, de los primeros 20 registros del ranking, solo 3 corresponden únicamente a la categoría cervezas. A su vez se puede observar, también, que, de las cervezas adquiridas la gran mayoría son de carácter importadas, por lo cual se procede a listar todas las marcas que abarca el dataset y a reemplazar a aquellas que sean importadas con el label "importada", lo cual permitirá segmentar de una mejor manera para el análisis. De esta manera, las labels correspondiente a las marcas son: Stella Artois, Patagonia Brewing Co., Corona, BevyBar, Cervecería y Maltería Quilmes, e Importada.

El siguiente paso consiste en analizar la cantidad de compras por usuario (id) para apariciones menores a 10, pudiendo observar lo siguiente:



La mediana de las compras es de una unidad, pudiendo ser hasta dos ó tres y, en muy pocos casos, mayor. Esto significa que **la mayoría de los usuarios no reincide a efectuar segundas compras en la plataforma de ecommerce.**.

Luego, analizamos la distribución de los gastos por compra:



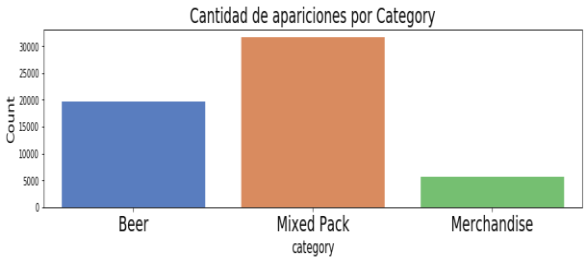
std: 714.26 **min**: 2.2517 **25%:** 190 **50%:** 650.176 **75%:** 955 **max**: 15732

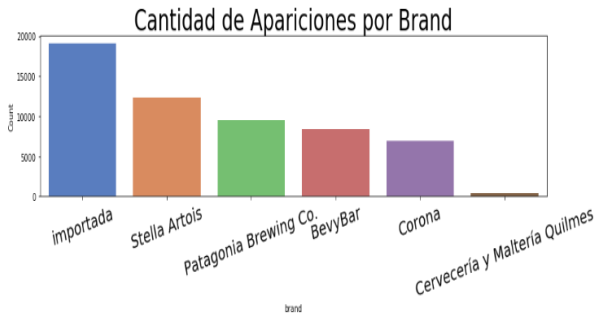
Observando las compras por producto ("name"), se puede verificar que los productos más comprados son "Pack Conservadora Corona" y "Copas Stella Artois", lo que se corresponde con las grandes campañas publicitarias efectuadas por la empresa durante los últimos años.

El EDA continúa con el análisis de las marcas más adquiridas, así también como las categorías más adquiridas. Para esta última variable, lo más adquirido son Mixed Packs (cerveza + merchandiseing), seguido por cerveza y merchandising. También, pero con muy baja cantidad, están las categorías Wine, Spirits, Unknown, y Home Breweing, las cuales

representan menos del 1,8% del total, de modo que se procede a extraerlas y trabajar únicamente con las tres primeras. De esta manera, puede observarse lo siguiente en la Figura A:

Lo más adquirido en el Ecommerce son Mixed Pack (55.6%), seguido por Beer (34.58%) y Merchandise (9.8%).



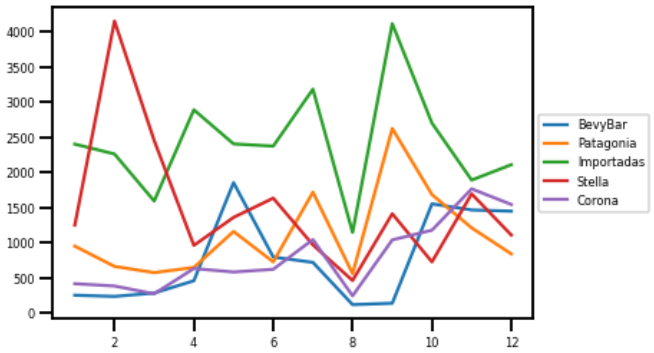


Luego, en la Figura B se ve que lo más adquirido en el Ecommerce son marcas importadas (34.6%), seguido por Stella Artois (21.31%), Patagonia (16.6%), BevyBar (14.6%), Corona (12%) y Quilmes (0.82%).

Luego, nos facilitaremos de tablas pivot para analizar qué rubros son más adquiridos dentro de cada marca. Dicho análisis se realizará mensualmente lo cual, además, permitirá ver la evolución de las ventas. De la tabla mencionada podemos observar los siguientes puntos más relevantes:

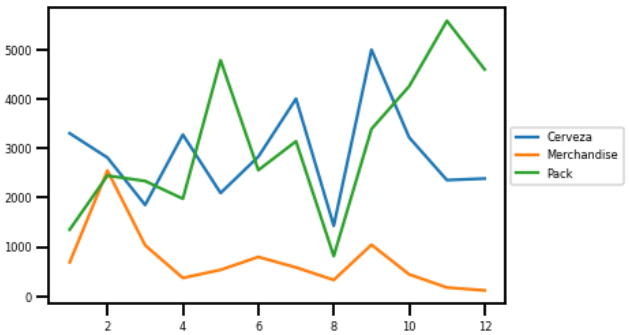
- De las marcas importadas, se compra, en su gran mayoría, cerveza.
- De las marcas nacionales, se compra, en su gran mayoría, mixed pack y merchandising.
- De Patagonia se adquiere una gran proporción de Mixed Packs, seguido por cerveza, y una escasa participación de Merchandising. Caso similar ocurre con Stella Artois, aunque este sí adquiere una significativa proporción de merchandising en lugar de cerveza.

A continuación, se analizará la evolución de las ventas tanto a nivel general como por distintos nichos. El objetivo es verificar la existencia (o no) de estacionalidad y poder comparar con registros de otros años. Quedará para un análisis posterior fuera de este informe la estacionalidad definitiva. Observaremos primero lo sucedido con las marcas:



A nivel general, podemos afirmar que las cantidades son similares en la última etapa del año. Por otro lado, Corona y BevyBar presentan una evolución similar entre sí. Stella Artois cuenta con un nivel elevado de ventas en la primera época del año, lo cual se conoce (de forma externa a este reporte) que responde a una importante campaña de publicidad.

Vemos lo ocurrido para las diferentes categorías:



La categoría “Merchandise” tiene un comportamiento similar al de Stella en el gráfico anterior. Esto puede ser explicado por lo encontrado en las tablas pivot. Caso similar sucede con Cerveza e Importadas.

4 MODELO DE APRENDIZAJE

Una vez finalizado el EDA se procede a aplicar un aprendizaje no supervisado (Clustering). Se aplicarán tres Clusterizaciones: un primero para todo el dataset, el segundo para los productos

nacionales (productos *core* del negocio) y un tercero para los productos importados.

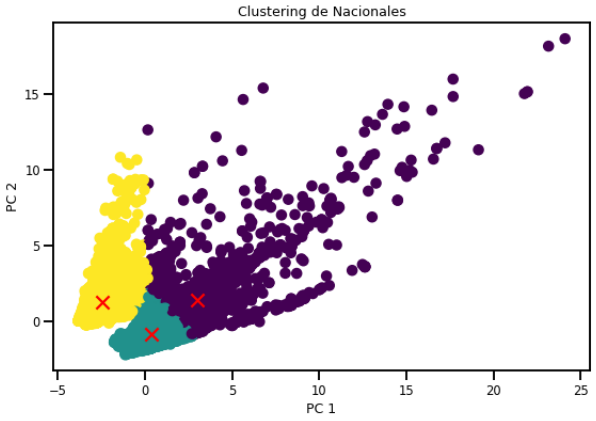
Para ello se generan, previamente, dummies para las features de Category y Brand, las cuales serán relevantes para este analisis. (Para el caso de las importadas, se procede a retirar aquellas que hayan tenido menos de 60 adquisiciones ya que la variedad de marcas es muy elevada). Además, no se introducirán al algoritmo algunas variables que no aportarían información o dificultarían su aplicación, por lo que se las remueve temporalmente del dataset.

En primer paso se procede a una reducción de la dimensionalidad de cada dataset vía PCA, un método de combinación lineal de variables que permite obtener nuevas features que expliquen en forma más adecuada la variabilidad de los datos.

Posteriormente, se aplica el método de aprendizaje no supervisado "K-means" clustering, que a partir de una cantidad K de clusters definida como hiperparametro, calcula la distancia de cada sample al centroide de cada uno para determinar a cuál de ellos pertenece.

Se observa que utilizando este método, los resultados obtenidos para los dataset generales y de importadas no manifiestan una representación de la varianza ni un silhouette score (validación de los resultados de clusterización a partir de las distancias *intra* y *entre* clusters) alto.

Los mejores resultados (90% de explicación de varianza y silhouette score=0,6) se obtienen para la aplicación del método en el dataset que solo incluye marcas nacionales "core" del negocio, por lo que se decide poner foco en ellas y continuar con su analisis exploratorio.



5 RESULTADOS DEL CLUSTER

Luego de reincorporar dos features claves como 'id' y 'name', comparar (mediante boxplots y distintas visualizaciones) las features de cada uno de los 3 clusters obtenidos, se pudo apreciar que el método de aprendizaje no supervisado nos permitió encontrar características destacadas para cada uno de ellos:

0 → El **cluster 0** está compuesto por **productos extraordinarios, de alto valor**, cuya cantidad de compras muestra que **los usuarios de este cluster solo compraron por única vez**, y otros en caso diferente se pueden considerar como outliers. A su vez, **el 55% de los samples consisten en el producto "Conservadora Corona",** el producto que al principio del análisis detectamos como **el más vendido de todos**, lo cual también permite observar una amplia predominancia de la marca "Corona" por sobre el resto.

1 → El **cluster 1** está compuesto principalmente por distintos **packs** tanto de **copas como de cervezas**, los cuales tienen un valor medio monetario inferior al cluster 0 y superior al cluster 2. En él, **no se aprecian grandes distinciones entre marcas**, sino que las 3 marcas que ofrecen este tipo de producto (Stella Artois, Patagonia, Bevybar) se encuentran presentes en proporciones similares. Este clúster **representa** además ampliamente **el de mayor volumen de compras** (23438 operaciones V.S. 9570 y 4654 operaciones) en el año 2017.

2→ El **cluster 2** está compuesto por **productos individuales, principalmente copas y cervezas, incluyendo la gran mayoría de los productos de merchandising** que comercializa el portal. Este, a su vez, presenta **el precio de venta más bajo de los 3 clusters encontrados.**

Esto nos permitió además validar que **la mejor representación se obtiene con 3 clusters**, ya que empíricamente encontramos características en común dentro de cada uno y muy diferentes entre ellos. Se encuentra además que, en la mayoría de los casos, **cada producto** ('name') **se encuentra agrupado en un clúster diferente**, aun cuando **esta feature no fue incluida dentro del algoritmo** de K-means clustering.

Respecto a las reincidencias de compras en la plataforma, los clusters 1 y 2, tienen una cantidad de operaciones de compra (mean y boxplot) muy similar, con hasta 2 o 3 compras por usuario, mientras que el cluster 0, no evidencia reincidencias. Esto nos permite obtener como conclusión que **la venta de la Conservadora marca "Corona", el producto más vendido y al cual se le invierte un mayor monto de publicidad en las actividades de marketing de la empresa, no genera una fidelización del cliente sobre la plataforma de e-commerce, ya que estos no vuelven a comprar en ella.**

6 DISCUSION Y CONCLUSIONES

Luego de los distintos análisis llevados a cabo se han podido observar ciertos comportamientos en la compra de bienes dentro de la plataforma empleada. Por un lado, el e-commerce para bebidas alcohólicas es ampliamente utilizado en la adquisición de cervezas importadas. Las mismas no son comercializadas en los típicos supermercados de barrio en los cuales se compran las cervezas más comunes y de carácter nacional, sino que sus puntos de venta son limitados y no siempre la variedad es amplia, por lo que su consumo se da mayormente en este tipo de plataformas. Caso similar podemos observar con lo que refiere a merchandising y mixed packs, los cuales se adquieren principalmente para las marcas de cervezas Premium nacionales y de consumo corriente. Por otro lado, para el año analizado, se ha visto que la gran mayoría de los usuarios no han efectuado nuevas compras, es decir, solo han llevado a cabo una. En cuanto al modelo de aprendizaje no supervisado empleado (Clusters) se han podido observar tres grupos de clientes bien diferenciados dentro de la adquisición de productos de marca nacional, que se explican principalmente por el monto de la compra realizada, es decir, un grupo que adquiere productos de alto valor y solo efectúa una compra y de una marca en particular, un segundo grupo que realiza un gasto medio pero que su consumo es más variado, y un tercer grupo cuyos gastos son los más bajos y pueden encontrarse en más de una compra.

A efectos del alcance de dicho trabajo, quedan excluidos del mismo: análisis profundos de estacionalidad (con disponibilidad de datos de otros años) y la aplicación de algoritmos que permitan predecir compras; análisis de clusters para compras de cervezas importadas; e incidencia de campañas de publicidad sobre las ventas.

7 REFERENCIAS

1. Python Data Science Handbook - Jake VanderPlas: https://tanthiamhuat.files.wordpress.com/2018/04/pythondatasciencehandbook.pdf

2. Machine Learning Yearning - Andrew NG https://d2wvfoqc9gyqzf.cloudfront.net/content/uploads/2018/09/Ng-MLY01-13.pdf

3 .http://www.Stackoverflow.com

4. https://github.com/clusterai