GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

TRABAJO FIN DE GRADO

Estudio de las características de los consumidores y predicción de la demanda en comercios online

Autor: Javier Pérez de Vargas Belmonte

Director: Álvaro Jesús López López

Madrid

Julio de 2020

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título *"Estudio de las características de los consumidores y predicción de la demanda en comercios online"* en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el

curso académico 2019/2020 es de mi autoría, original e inédito y

no ha sido presentado con anterioridad a otros efectos. El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada

de otros documentos está debidamente referenciada.

Fdo.: Javier Pérez de Vargas Belmonte Fecha: 20/07/ 2020

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: Álvaro Jesús López López Fecha: 23/ 07/ 20

GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

TRABAJO FIN DE GRADO

Estudio de las características de los consumidores y predicción de la demanda en comercios online

Autor: Javier Pérez de Vargas Belmonte

Director: Álvaro Jesús López López

Madrid

Julio de 2020

ESTUDIO DE LAS CARACTERÍSTICAS DE LOS CONSUMIDORES Y PREDICCIÓN DE LA DEMANDA EN COMERCIOS ONLINE

Autor: **Pérez de Vargas Belmonte, Javier**

Director: López López, Álvaro Jesús

Entidad colaboradora: ICAI – Universidad Pontificia de Comillas.

RESUMEN DEL PROYECTO

Introducción

Desde que diera comienzo la tercera revolución industrial se han hecho multitud de avances gracias a Internet, no solo tecnológicos, sino también sociales y culturales. Internet ha cambiado las reglas de cómo funciona el mundo, desde la forma de interactuar entre las personas hasta los modelos de organización dentro de una empresa, pasando por la relación cliente-proveedor, caracterizada por la proliferación de los denominados *ecommerce* o comercios online.

Ahora nos vemos envueltos en las transformaciones asociadas a la cuarta revolución industrial, prosiguiendo con la búsqueda de la eficiencia que perseguían las tres revoluciones anteriores, pero esta vez caracterizada, entre otros aspectos, por el uso y tratamiento de los datos. En el ámbito del *ecommerce,* este tratamiento de los datos recibidos a través de redes sociales o del propio portal de venta es utilizado para que las empresas sean más eficiente especialmente en las áreas de producción, gestión de inventario, precios y relaciones con el cliente. Hoy en día, no sólo se trata de analizar los datos históricos de la demanda, sino de comprender e integrar en los modelos de las empresas, el impacto de otras variables importantes como el clima, las promociones comerciales, los eventos de marketing, las nuevas leyes y demás factores económicos. La capacidad de los ordenadores nos permite actualmente almacenar y procesar grandes cantidades de datos con rapidez para poder obtener estimaciones fiables de la demanda. La necesidad de estar preparado y contribuir a semejante cambio supone la motivación principal de este proyecto.

El objetivo de este trabajo es el de predecir la demanda del *ecommerce* Independence Brand, plataforma de comercio electrónico creada por el autor y auspiciada por la asociación StartComillas, así como realizar un estudio de caracterización de sus clientes mediante el estudio de sus datos históricos de ventas y la aplicación de diferentes técnicas de predicción de la demanda y el algoritmo K-Means.

El pronóstico de la demanda se entiende como el proceso de predecir ventas futuras mediante el uso de datos históricos con el objetivo de tomar decisiones comerciales en diversas áreas, desde la planificación del inventario y las necesidades de almacenamiento hasta la ejecución de promociones y el cumplimiento de las expectativas del cliente. Pretende conectar la cadena de suministro de una empresa, los hábitos de compra de sus clientes y factores externos para una estimación basada en datos de las ventas futuras esperadas.

Esta predicción de la demanda debe ir complementada con un exhaustivo estudio de las características de los consumidores, estas suponen las claves para encontrar el camino hacia la fidelización de los clientes. Una vez fidelizado un cliente, este se convierte en el embajador de la marca, promoviendo la comercialización entre conocidos, amigos y familiares. No solo retener y fidelizar a los consumidores son las claves del éxito de una compañía, también se debe aumentar y expandir esta red de consumidores, en esta línea se hace crucial conocer las características de los consumidores ya fidelizados para saber dónde lanzar nuevas campañas de marketing con el objetivo de lograr nuevos clientes.

Además, conocer las características de los consumidores también se debe entender como un paso hacia una mayor eficiencia en las operaciones, al igual que con el pronóstico de la demanda. Al direccionar todas las acciones de una compañía hacia el hábitat de sus consumidores genera grandes ahorros y garantiza una imagen más sólida y consistente de la marca dentro del mercado, ya que al conocer la conducta del consumidor se trabaja para un grupo concreto, y no para cualquier comprador.

Metodología

La metodología para llevar a cabo este trabajo ha conllevado una serie de fases enfocadas a desarrollar un modelo de predicción de la demanda y estudio de las características de los consumidores en negocios online.

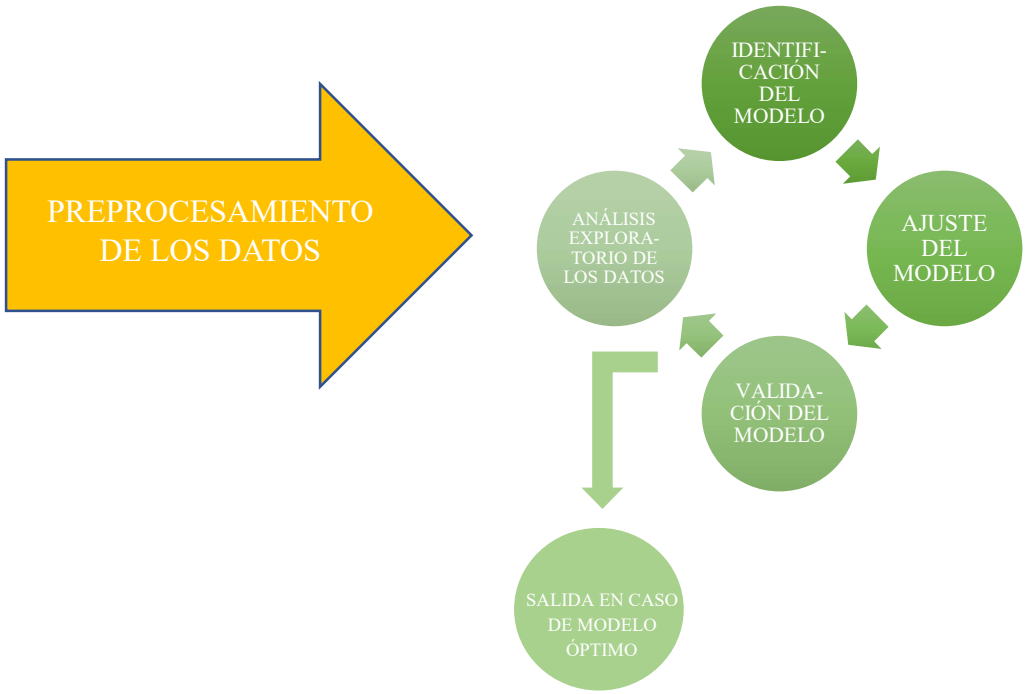

Figura 1. Representación de la metodología de desarrollo del proyecto.

La primera etapa consiste en una recolección y posterior limpieza de los datos, a fin de sentar unas buenas bases para obtener una mayor precisión en los resultados finales.

Posteriormente, se lleva a cabo el estudio de las características de los consumidores mediante la aplicación del algoritmo K-Means y, una vez realizada la caracterización de los consumidores, da comienzo el estudio de la predicción de la demanda. Para realizar esta predicción de la demanda se ha desarrollado un estudio de la estacionalidad de la demanda relacionándola con los momentos de lanzamiento de promociones. Lo ideal sería desarrollar un modelo de predicción más detallado, lo cual se deja para futuras líneas de trabajo.

Resultados

Los datos usados en este proyecto han sido extraídos de la información de ventas del *ecommerce* Independence Brand. Los datos fueron exportados desde la plataforma de creación de webs Shopify [33], la cual da la opción de exportarlos en formato *.csv*.

Valiéndome de la herramienta Excel, los datos son importados a un formato .xlsx que toda la información y variables son claramente visibles, como se muestra a continuación en la figura 2.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Name	Email	Financial Status	Paid at	Fulfillment Status	Fulfilled at	Accepts Marketing	Currency	Subtotal	Shipping	Taxes	Total	Discount Code
2	#1138	thiebaut.rafael@hotmail.com	paid	29/01/2020 0:32	fulfilled	29/01/2020 0:32	no	EUR	600	395	104	995	
3	#1138	thiebaut.rafael@hotmail.com											
4	#1137	carlos.argos2000@gmail.com	paid	24/01/2020 19:21	fulfilled	24/01/2020 19:21	yes	EUR	0	395	0	395	Tarjeta_regalo
5	#1137	carlos.argos2000@gmail.com											
6	#1137	carlos.argos2000@gmail.com											
7	#1136	javiergargon@hotmail.es	paid	24/01/2020 14:44	fulfilled	24/01/2020 14:44	no	EUR	600	395	104	995	
8	#1136	javiergargon@hotmail.es											
9	#1135	javiergargon@hotmail.es	paid	24/01/2020 14:36	fulfilled	24/01/2020 14:36	no	EUR	1900	395	0	2295	
10	#1135	javiergargon@hotmail.es											
11	#1134	ciracheb@yahoo.es	refunded	23/01/2020 15:27	fulfilled	23/01/2020 15:27	yes	EUR	3140	0	215	3140	
12	#1134	ciracheb@yahoo.es											
13	#1134	ciracheb@yahoo.es											
14	#1134	ciracheb@yahoo.es											
15	#1133	lidia.pelayo@outlook.com	paid	30/12/2019 17:37	fulfilled	30/12/2019 17:37	no	EUR	5800	0	1007	5800	
16	#1132	idecristobal@gmail.com	paid	28/12/2019 19:37	fulfilled	28/12/2019 19:37	yes	EUR	3500	0	607	3500	
17	#1132	idecristobal@gmail.com											
18	#1131	mjgarrigos69@gmail.com	paid	25/12/2019 23:06	fulfilled	25/12/2019 23:06	yes	EUR	2780	395	482	3175	
19	#1131	mjgarrigos69@gmail.com											
20	#1130	hdezglezjavier@gmail.com	paid	18/12/2019 10:14	fulfilled	18/12/2019 10:14	yes	EUR	2502	395	434	2897	INDEPENDENCEGS
21	#1130	hdezglezjavier@gmail.com											
22	#1130	hdezglezjavier@gmail.com											
23	#1129	franciscosospedra1@gmail.com	paid	17/12/2019 1:19	fulfilled	17/12/2019 1:19	yes	EUR	3500	0	278	3500	
24	#1129	franciscosospedra1@gmail.com											

Figura 2. Conjunto de datos en formato .xlsx.

Este documento cuenta con 273 filas, que representan cada uno de los productos vendidos, y 69 columnas, que representan cada una de las variables de cada producto o pedido. No todas las 69 variables nos aportan información relevante a la hora de hacer un estudio de las características de los consumidores o predecir la demanda de los próximos meses. Por ello, me deshice de las variables cuyos campos estaban vacíos, como el desglose de los impuestos (Tax 1, 2 ,3…) y "Billing Company", de las variables que no tenían variabilidad, como "Currency" (Independence Brand se ha limitado a satisfacer las necesidades del territorio nacional), y de las variables que contenían información repetida, como todas las variables del grupo "Shipping". Quedando, de esta manera, las siguientes variables:

- Name.
- Email.
- Financial Status.

- Accepts Marketing.
- Total.
- Discount Code.
- Discount Amount.
- Shipping Method.
- Created at.
- Lineitem quantity.
- Lineitem Name.
- Shipping Province.
- Payment Method.

Tras eliminar los datos nulos o erróneos, agrupar las filas por pedidos, crear nuevas variables temporales y, debido al alcance de este proyecto, solo teniendo en cuenta las variables numéricas y dejando las categóricas para futuras líneas de trabajo, quedan las siguientes variables:

- Total.
- Discount Amount.
- Timeframe.
- Month frame.
- Item quantity.

Una vez realizado el preprocesamiento y la limpieza de los datos, podremos importar los datos del documento Excel a Python, estandarizar los datos, ejecutar el algoritmo K-Means y un Análisis de Componentes Principales, y representar los resultados. De todas las técnicas de aprendizaje supervisado, finalmente se decidió utilizar el algoritmo K-Means por su popularidad, eficiencia y poca complejidad de implementación.

Una de las desventajas del método de clústering K-Menas es que necesitar especificar el número de clústeres de antemano, normalmente obtenido mediante el método del codo. Este método utiliza los valores de la inercia obtenidos tras aplicar el K-means a diferente número de clústeres, siendo la inercia la suma de las distancias al cuadrado de cada objeto del clúster a su centroide. Una vez obtenidos los valores de la inercia tras aplicar el K-means, representamos en una gráfica lineal la inercia respecto del número de clústeres. El punto en el que se observa un cambio brusco en la inercia nos indica el número óptimo de clústeres.

$$Inercia = \sum_{i=0}^{N} \|x_i - \mu\|^2$$

Figura 3. Fórmula de la inercia aplicada a clústeres.

A continuación, se muestra en la figura 4 la representación de la inercia respecto al número de clústeres para obtener el número óptimo de éstos.

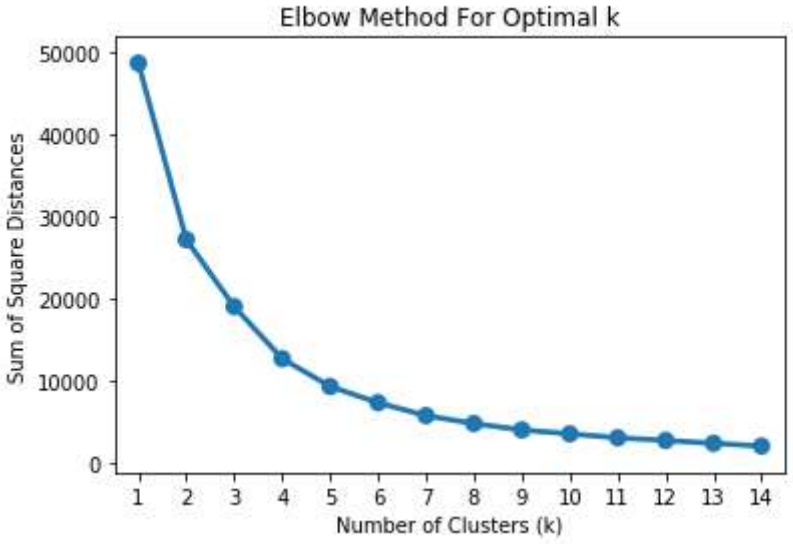

Figura 4. Representación de la inercia respecto al número de clústeres.

Atendiendo al gráfico de arriba, el codo y número óptimo de clústeres correspondería con el cuatro, sin embargo, después de hacer varias pruebas se determinan tres clústeres como el número óptimo más razonable.

Una vez ejecutado K-Means con el número de clústeres comentado anteriormente, nos encontramos con un conjunto de tres clústeres creados a partir de la información de las cinco variables numéricas que usamos, lo que nos impide su representación gráfica. Para solventar este problema recurrimos al Análisis de Componentes Principales.

El Análisis de Componentes Principales o PCA por sus siglas en inglés, permite agrupar un conjunto de variables en otro conjunto de menor dimensión que mantenga la mayor cantidad de información posible, mediante el uso de la matriz de covarianza. Después, se obtienen los autovectores y autovalores, ordenándolos de mayor a menor, resultando en tantas componentes principales como variables existentes. Sin embargo, hay componentes principales más relevantes que otras, obtenidas mediante el cálculo de la varianza explicada por cada una de ellas. En este caso, se seleccionan las dos primeras componentes principales con más varianza explicada, de manera que se pueden representar los resultados de K-Means sin perder mucha información. Debajo se muestra la representación del resultado final.

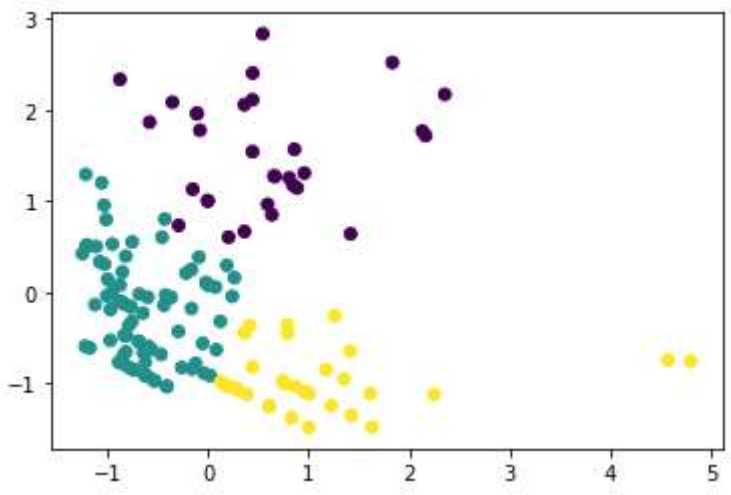

Figura 5. Representación del resultado del K-Means.

Atendiendo a los autovectores, representados en la tabla debajo, cabe destacar que hay tres variables que influyen más que el resto a la hora de asignar un clúster u otro a un pedido. Estas tres variables son el total (variable 0), el número de productos comprados (variable 4) y el capital descontado del total (variable 1), teniendo el momento del día y del mes una menor influencia (variables 2 y 3).

	0	1	2	3	4
0	0.637384	0.219558	0.147752	0.116988	0.714156
1	-0.300998	0.795697	-0.0618223	-0.507976	0.120018

Figura 6. Autovectores de las componentes principales.

De combinar la información dada por los autovectores y la representación de los clústeres, podemos observar:

1) Las transacciones recogidas en el clúster morado son influenciadas en mayor parte por el descuento.
2) Las transacciones recogidas en el clúster amarillo son influenciadas en mayor parte el total pagado y el número de productos.
3) Las transacciones recogidas en clúster verde no tienen una influencia marcada por ninguna de las variables.

Para poder caracterizar los diferentes grupos de comportamientos compradores es necesario extraer los valores medios de cada una de las cinco variables de cada clúster.

Una vez calculados, los valores obtenidos son los siguientes:

- Clúster morado:
    - Total medio: 25,76
    - Capital descontado medio: 21,88
    - Franja horaria de compra: 3,32
    - Semana del mes de compra: 2,06
    - Número medio de productos comprados: 2,65

- Clúster amarillo:
    - Total medio: 46,31
    - Capital descontado medio: 0,83
    - Franja horaria de compra: 3,48
    - Semana del mes de compra: 3,45
    - Número medio de productos comprados: 2,85

- Clúster verde:
    - Total medio: 21,03
    - Capital descontado medio: 2,03
    - Franja horaria de compra: 3,0
    - Semana del mes de compra: 2,59
    - Número medio de productos comprados: 1,21

Conclusiones

El análisis desarrollado en este trabajo brinda a Independence Brand varias oportunidades de crecimiento, sobre todo a través de estrategias de marketing. Teniendo en cuenta los valores medios de las variables estudiadas para cada clúster se pueden idear varias acciones de marketing que pueden incrementar la tasa de conversión, estas acciones son:

- *Acciones destinadas a los clientes agrupados bajo el clúster morado*: Los clientes agrupados bajo este clúster se caracterizan por ser atraídos por los descuentos, por lo que lo aparentemente más efectivo sería ofrecerles con descuentos. Los clientes agrupados bajo este clúster son interesantes, ya que son fácilmente accesibles mediante descuentos y dan movilidad de inventario.

    1) *Acción basada en total y descuento*: Basándonos en el total medio gastado de 25,76 euros y el descuento medio de 21,88 euros, podemos calcular un total gastado antes de aplicar descuento de 47,64 euros y un descuento aplicado del 46%. La promoción ideal para conseguir compras de estos clientes sería ofrecer un 50% de descuento o un 2x1 por la compra de pedidos superiores a 45 euros.
    2) *Acción basada en descuento y momento de compra*: Siguiendo con la idea de que a este grupo de clientes le mueven los descuentos y, añadiendo el factor

del momento de la compra, se puede alcanzar este grupo de clientes también. Dado lo comentado en la acción anterior y teniendo en cuenta que este grupo compra en torno a la franja horaria 3, que comprende las horas desde las 12:00 hasta las 18:00, la promoción ideal sería ofrecer un descuento para las compras realizadas durante esas horas.

3) *Acción basada en descuento y semana del mes de compra*: Esta acción es similar a la anterior, pero teniendo en cuenta la semana del mes en vez de las horas dentro del día. En este caso, el análisis indica que estos clientes suelen comprar los días comprendidos entre el 8 y el 15 de cada mes, por lo que la acción consistiría en ofrecer un descuento para las compras realizadas durante esos días.
4) *Acción basada en descuento y número de productos comprados*: Como se comentó a la hora de estudiar los autovectores, el número de productos es uno de los factores que más ha influido a la hora de la agrupación. Teniendo en cuenta el número medio de productos comprados de 2,65 por este grupo de clientes, la acción de marketing óptima sería ofrecer descuento por la compra de más de dos productos.

Estas estrategias arriba expuestas también pueden combinarse entre sí, siempre y cuando se mantenga el incentivo del descuento, pudiendo incluso lanzar una promoción que las combine todas. La promoción que intenta ser lo más precisa posible sería ofreciendo un descuento por compras realizadas entre las 12:00 y las 18:00 de los días del 8 al 15 de cada mes, por un importe superior a 45 euros y más de 2 productos.

- *Acciones destinadas a los clientes agrupados bajo el clúster amarillo*: Los clientes agrupados bajo este clúster se caracterizan por ser los que más dinero se gastan en cada pedido y no usar descuentos. Este grupo es de especial interés, ya que son los que más ingresos dan a Independence Brand, sin embargo, son difícilmente accesibles con las variables estudiadas.

1) *Acción basada en el momento de lanzamiento de nuevos productos*: Según las variables estudiadas, este grupo de clientes compra independientemente de los descuentos ofrecidos. No se conoce con exactitud lo que motiva a estos clientes a comprar, pero dado que son los que más dinero gastan y no usan descuentos, deben de representar los clientes más fieles a la marca que compran sin importar las condiciones. Partiendo de que son nuestros clientes más fieles, se entiende que son los que compran los nuevos productos Independence en cuanto salen. Dicho esto, para llegar a ellos y aprovechar su potencia de compra, lo ideal sería hacer los lanzamientos de los productos en la franja horaria 3,48 y semana del mes 3,45, correspondiendo aproximadamente a las últimas seis horas del día y última semana del mes.

Lamentablemente, falta información para poder atacar con precisión el grupo de clientes que produce más beneficios. Se deja para futuras líneas de investigación.

- *Acciones destinadas a los clientes agrupados bajo el clúster verde*: Bajo este clúster se agrupan los clientes que podrían considerarse lo menos interesantes, gastan poco, compran pocos productos y, aparentemente, no se guían por incentivos como los descuentos. Dado que el momento del día y del mes en que compran son similares a los del clúster morado, las acciones empleadas para atraer a dichos clientes también podrían atraer a los agrupados bajo este clúster.

En cuanto a la predicción de la demanda, las figuras 7 y 8 mostradas a continuación representan el historial de ventas y visitas de Independence Brand, respectivamente. En ambas es posible advertir cierta estacionalidad, localizando picos locales en los meses de noviembre-diciembre y mayo-junio, con unos pequeños repuntes los meses de septiembre y marzo. Dichos picos coinciden, principalmente, con el Black Friday y preparación de Navidad, e inicio del verano. Esta información nos es muy útil a la hora de saber cuándo lanzar un nuevo producto o una nueva acción potente en marketing.

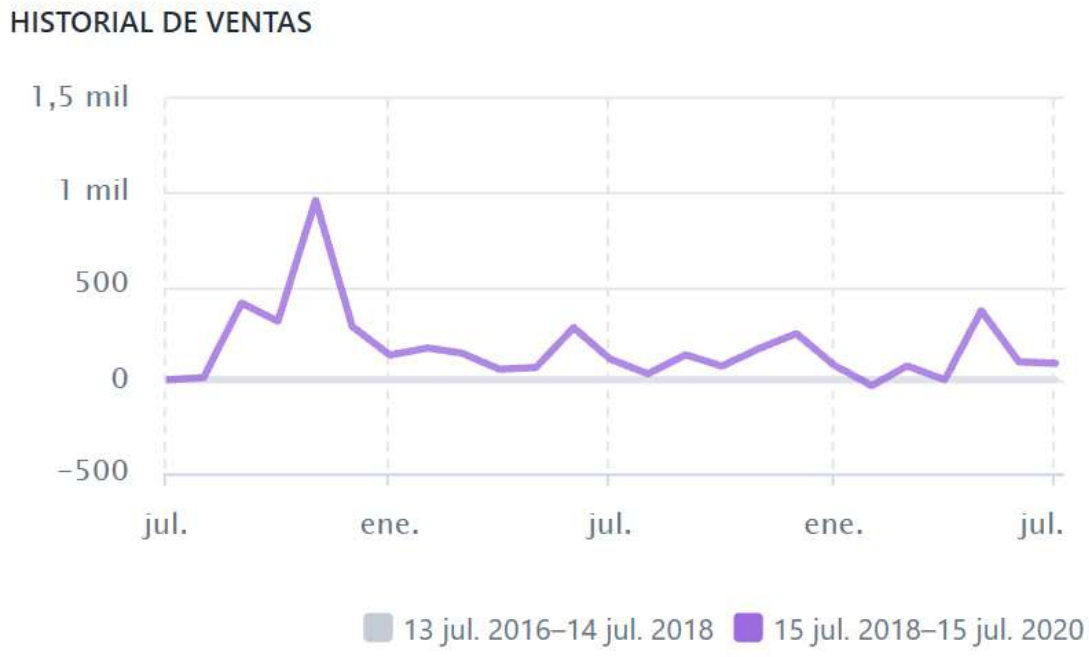

Figura 7. Histórico de las ventas de Independence Brand.

Figura 8. Histórico de las visitas a la web de Independence Brand.

Referencias

[1] Web de Independence Brand. Último acceso: 20/07/2020. https://www.independencebrand.com/

[2] Web de Shopify. Último acceso: 20/07/2020. https://www.shopify.com/

[3] Pawan Kumar Singh, Yadunath Gupta, Nilpa Jha and Aruna Rajan. 2019. "Fashion Retail Forecasting Demand for New Items".

[4]"Análisis de componentes principales" – Joaquín Amat. Último acceso: 20/07/2020. https://www.cienciadedatos.net/documentos/35_principal_component_analysis

STUDY OF THE CHARACTERISTICS OF CONSUMERS AND DEMAND FORECAST IN ECOMMERCE.

Author: **Pérez de Vargas Belmonte, Javier**

Director: López López, Álvaro Jesús

Collaborating entity: ICAI – Universidad Pontificia de Comillas.

ABSTRACT

Introduction

Since the third industrial revolution began, many advances have been made thanks to the Internet, not only technological, but also social and cultural. The Internet has changed the rules of how the world works, from the way people interact, to the organization models within a company, through the customer-supplier relationship, characterized by the proliferation of so-called e-commerce or online businesses.

Now we are involved in the transformations associated with the fourth industrial revolution, continuing with the search for efficiency pursued by the three previous revolutions, but this time characterized, among other aspects, by the use and treatment of data. In the field of e-commerce, this treatment of the data received through social networks or the sales portal itself is used to make companies more efficient, especially in the areas of production, inventory management, prices and customer relations. Today, it is not only a matter of analyzing the historical data of demand, but of understanding and integrating into the models of the companies, the impact of other important variables such as the weather, commercial promotions, marketing events, new laws and other economic factors. The capacity of computers now enables us to quickly store and process large amounts of data in order to obtain reliable estimates of demand. The need to be prepared and contribute to such a change is the main motivation for this project.

The objective of this work is to predict the demand for the Independence Brand e-commerce, an e-commerce platform created by the author and sponsored by the StartComillas association, as well as to carry out a characterization study of its clients by studying their historical sales data. and the application of different demand prediction techniques and the K-Means algorithm.

Demand forecasting is the process of predicting future sales using historical data to make business decisions in various areas, from inventory planning and warehousing needs to executing promotions and meeting expectations. the client's. It aims to connect the supply chain of a company, the purchasing habits of its customers and external factors for a data-based estimate of expected future sales.

This prediction of demand must be complemented by an exhaustive study of the characteristics of the consumers, these are the keys to finding the path to customer loyalty. Once a customer loyalty, this becomes the brand ambassador, promoting marketing among acquaintances, friends and family. Not only retaining and retaining consumers are the keys to a company's success, this consumer network must also be increased and expanded, in this line it is crucial to know the characteristics of already loyal consumers in order to know where to launch new marketing campaigns. with the aim of obtaining new clients.

In addition, knowing the characteristics of consumers should also be understood as a step towards greater efficiency in operations, as with forecasting demand. By directing all the actions of a company towards the habitat of its consumers, it generates great savings and guarantees a stronger and more consistent image of the brand within the market, since by knowing consumer behavior, you work for a specific group, and not for any buyer.

Methodology

The methodology for developing this work has involved a series of phases focused on developing a demand prediction model and studying the characteristics of consumers in online businesses.

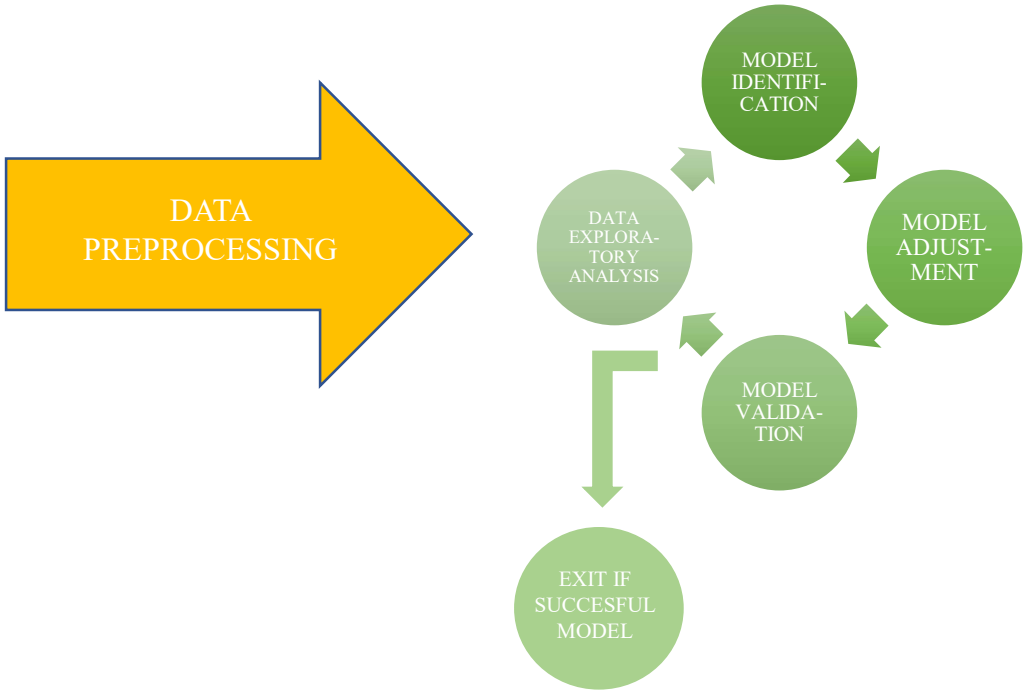

Figure 1. Representation of the methodology of the project.

The first stage consists of the collection and subsequent cleaning of the data, in order to lay a good foundation for obtaining greater precision in the final results. Subsequently, the study of consumer characteristics is carried out by applying the K-Means algorithm and, once the characterization of consumers has been carried out, the study of demand prediction begins. To make this prediction of demand, a study of the seasonality of demand has been developed, relating it to the times of launch of promotions. Ideally, a more detailed forecasting model should be developed, which is left for future lines of development.

Results

The data used in this project has been extracted from the sales information of the Independence Brand e-commerce. The data was exported from the Shopify web creation platform [33], which gives the option of exporting it in .csv format. Using the Excel tool, the data is imported into an .xlsx format that all the information and variables are clearly visible, as shown below in figure 2.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Name	Email	Financial Status	Paid at	Fulfillment Status	Fulfilled at	Accepts Marketing	Currency	Subtotal	Shipping	Taxes	Total	Discount Code
2	#1138	thiebaut.rafael@hotmail.com	paid	29/01/2020 0:32	fulfilled	29/01/2020 0:32	no	EUR	600	395	104	995	
3	#1138	thiebaut.rafael@hotmail.com											
4	#1137	carlos.argos2000@gmail.com	paid	24/01/2020 19:21	fulfilled	24/01/2020 19:21	yes	EUR	0	395	0	395	Tarjeta_regalo
5	#1137	carlos.argos2000@gmail.com											
6	#1137	carlos.argos2000@gmail.com											
7	#1136	javiergargon@hotmail.es	paid	24/01/2020 14:44	fulfilled	24/01/2020 14:44	no	EUR	600	395	104	995	
8	#1136	javiergargon@hotmail.es											
9	#1135	javiergargon@hotmail.es	paid	24/01/2020 14:36	fulfilled	24/01/2020 14:36	no	EUR	1900	395	0	2295	
10	#1135	javiergargon@hotmail.es											
11	#1134	ciracheb@yahoo.es	refunded	23/01/2020 15:27	fulfilled	23/01/2020 15:27	yes	EUR	3140	0	215	3140	
12	#1134	ciracheb@yahoo.es											
13	#1134	ciracheb@yahoo.es											
14	#1134	ciracheb@yahoo.es											
15	#1133	lidia.pelayo@outlook.com	paid	30/12/2019 17:37	fulfilled	30/12/2019 17:37	no	EUR	5800	0	1007	5800	
16	#1132	idecristobal@gmail.com	paid	28/12/2019 19:37	fulfilled	28/12/2019 19:37	yes	EUR	3500	0	607	3500	
17	#1132	idecristobal@gmail.com											
18	#1131	mjgarrigos69@gmail.com	paid	25/12/2019 23:06	fulfilled	25/12/2019 23:06	yes	EUR	2780	395	482	3175	
19	#1131	mjgarrigos69@gmail.com											
20	#1130	hdezglezjavier@gmail.com	paid	18/12/2019 10:14	fulfilled	18/12/2019 10:14	yes	EUR	2502	395	434	2897	INDEPENDENCEGS
21	#1130	hdezglezjavier@gmail.com											
22	#1130	hdezglezjavier@gmail.com											
23	#1129	franciscosospedra1@gmail.com	paid	17/12/2019 1:19	fulfilled	17/12/2019 1:19	yes	EUR	3500	0	278	3500	
24	#1129	franciscosospedra1@gmail.com											

Figure 2. Dataset in .xlsx format.

This document has 273 rows, which represent each of the products sold, and 69 columns, which represent each of the variables of each product or order. Not all 69 variables provide us with relevant information when making a study of consumer characteristics or predicting demand in the coming months. Therefore, I got rid of the variables whose fields were empty, such as the breakdown of taxes (Tax 1, 2, 3 ...) and "Billing Company", of the variables that had no variability, such as "Currency" (Independence Brand was it has limited to satisfying the needs of the national territory), and of the variables that contained repeated information, like all the variables of the "Shipping" group. Thus, the following variables remain:

- Name.
- Email.
- Financial Status.

- Accepts Marketing.
- Total.
- Discount Code.
- Discount Amount.
- Shipping Method.
- Created at.
- Lineitem quantity.
- Lineitem Name.
- Shipping Province.
- Payment Method.

After eliminating the null or erroneous data, grouping the rows by orders, creating new temporary variables and, due to the scope of this project, only taking into account the numerical variables and leaving the categorical ones for future lines of work, the following variables remain:

- Total.
- Discount Amount.
- Timeframe.
- Month frame.
- Item quantity.

Once the data has been preprocessed and cleaned, we can import the data from the Excel document into Python, standardize the data, run the K-Means algorithm and a Principal Component Analysis, and represent the results. Of all the supervised learning techniques, it was finally decided to use the K-Means algorithm due to its popularity, efficiency and low implementation complexity.

One of the disadvantages of the K-Menas clustering method is that you need to specify the number of clusters beforehand, usually obtained using the elbow method. This method uses the inertia values obtained after applying K-means to different number of clusters, with inertia being the sum of the squared distances of each object in the cluster from its centroid. Once the inertia values have been obtained after applying the K-means, we represent the inertia with respect to the number of clusters on a linear graph. The point at which a sharp change in inertia is observed indicates the optimal number of clusters.

$$Inercia = \sum_{i=0}^{N} \|x_i - \mu\|^2$$

Figure 3. Cluster inertia formula.

The representation of inertia with respect to the number of clusters is shown in Figure 4 below to obtain the optimal number of clusters.

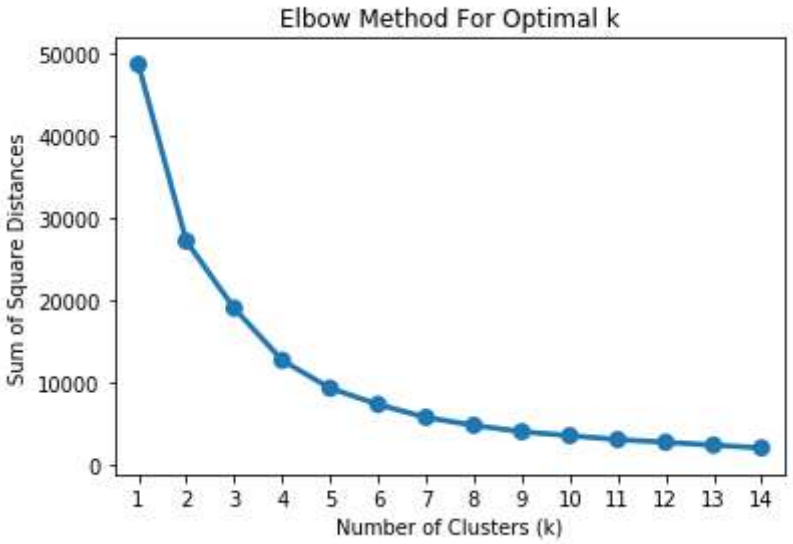

Figure 4. Inertia against number of clusters graph.

Based on the graph above, the elbow and optimal number of clusters would correspond to four, however, after doing several tests, three clusters are determined as the most reasonable optimal number.

Once K-Means has been run with the number of clusters discussed above, we find a set of three clusters created from the information of the five numerical variables that we use, which prevents us from graphing them. To solve this problem, we resort to Principal Component Analysis.

The Principal Component Analysis or PCA for its acronym in English, allows grouping a set of variables into another set of smaller dimensions that maintains as much information as possible, using the covariance matrix. Then, the eigenvectors and eigenvalues are obtained, ordering them from highest to lowest, resulting in as many main components as there are existing variables. However, there are more relevant main components than others, obtained by calculating the variance explained by each of them. In this case, the first two main components with the most explained variance are selected, so that the K-Means results can be represented without losing much information. Below is the representation of the final result.

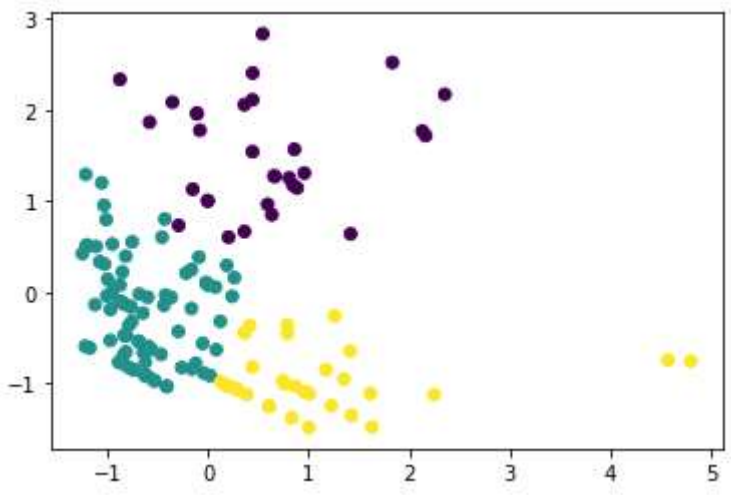

Figure 5. K-Means results graph.

Considering the eigenvectors, represented in the table below, it should be noted that there are three variables that influence more than the rest when assigning one cluster or another to an order. These three variables are the total (variable 0), the number of products purchased (variable 4) and the discounted capital of the total (variable 1), with the time of day and month having less influence (variables 2 and 3).

	0	1	2	3	4
0	0.637384	0.219558	0.147752	0.116988	0.714156
1	-0.300998	0.795697	-0.0618223	-0.507976	0.120018

Figura 6. Eigenvectors of principal components.

Combining the information given by the eigenvectors and the representation of the clusters, we can observe:

1) The transactions collected in the purple cluster are largely influenced by the discount.

2) The transactions collected in the yellow cluster are largely influenced by the total paid and the number of products.

3) The transactions collected in a green cluster do not have a marked influence by any of the variables.

In order to characterize the different groups of purchasing behaviors, it is necessary to extract the mean values of each of the five variables in each cluster. Once obtained, the values are the following ones:

- Purple cluster:
    - Average total: 25,76
    - Average discounted amount: 21,88
    - Buying time frame: 3,32
    - Purchase week: 2,06
    - Average number of products purchased: 2,65

- Yellow cluster:
    - Average total: 46,31
    - Average discounted amount: 0,83
    - Buying time frame: 3,48
    - Purchase week: 3,45
    - Average number of products purchased: 2,85

- Green cluster:
    - Average total: 21,03
    - Average discounted amount: 2,03
    - Buying time frame: 3,0
    - Purchase week: 2,59
    - Average number of products purchased: 1,21

Conclusions

The analysis developed in this work provides Independence Brand with several growth opportunities, especially through marketing strategies. Taking into account the average values of the variables studied for each cluster, several marketing actions can be devised that can increase the conversion rate, these actions are:

- *Actions aimed at customers grouped under the purple cluster*: Clients grouped under this cluster are characterized by being attracted by discounts, so apparently the most effective would be to offer them with discounts. Clients grouped under this cluster are interesting, as they are easily accessible through discounts and give inventory mobility.

1) *Action based on total and discount*: Based on the average total spent of 25.76 euros and the average discount of 21.88 euros, we can calculate a total spent before applying a discount of 47.64 euros and an applied discount of 46 %. The ideal promotion to get purchases from these customers would be to offer a 50% discount or a 2x1 for the purchase of orders over 45 euros.

2) *Action based on discount and time of purchase*: Continuing with the idea that this group of customers is driven by discounts and, adding the factor of the time of purchase, this group of customers can also be reached. Given what was discussed in the previous action and taking into account that this group buys around time slot 3, which includes the hours from 12:00 to 18:00, the ideal promotion would be to offer a discount for purchases made during those hours.

3) *Action based on discount and week of the month of purchase*: This action is similar to the previous one, but taking into account the week of the month instead of the hours within the day. In this case, the analysis indicates that these customers usually buy on the days between the 8th and 15th of each month, so the action would be to offer a discount for purchases made on those days.

4) *Action based on discount and number of products purchased*: As mentioned when studying autovectors, the number of products is one of the factors that has had the most influence when grouping. Taking into account the average number of products purchased of 2.65 by this group of customers, the optimal marketing action would be to offer a discount for the purchase of more than two products.

These strategies above can also be combined with each other, as long as the discount incentive is maintained, and may even launch a promotion that combines them all. The promotion that tries to be as accurate as possible would be offering a discount for purchases made between 12:00 and 18:00 on the days of the 8 to the 15 of each month, for an amount greater than 45 euros and more than 2 products.

- *Actions aimed at customers grouped under the yellow cluster:* Clients grouped under this cluster are characterized by being the ones that spend the most money on each order and not using discounts. This group is of special interest, since they are the ones that give the most income to the Independence Brand, however, they are hardly accessible with the variables studied.

  1) *Action based on the time of launch of new products*: According to the variables studied, this group of customers buys independently of the discounts offered. It is not known exactly what motivates these customers to buy, but since they are the ones who spend the most money and do not use discounts, they must represent the

most loyal customers to the brand who buy regardless of the conditions. Based on the fact that they are our most loyal customers, it is understood that they are the ones who buy the new Independence products as soon as they come out. That said, to reach them and take advantage of their purchasing power, the ideal would be to launch the products in the 3.48 time slot and week of the 3.45 month, corresponding approximately to the last six hours of the day and last week of the month.

Unfortunately, information is lacking to be able to precisely target the group of customers that produces the most benefits. It is left for future lines of research.

- *Actions aimed at customers grouped under the green cluster*: Clients that could be considered the least interesting are grouped under this cluster, spend little, buy few products and, apparently, are not guided by incentives such as discounts. Since the time of day and month in which they shop are similar to those in the purple cluster, the actions used to attract such customers could also attract those grouped under this cluster

Regarding demand prediction, Figures 29 and 30 below represent the sales and visit history of Independence Brand, respectively. In both of them it is possible to notice a certain seasonality, locating local peaks in the months of November-December and May-June, with a slight increase in the months of September and March. These peaks coincide, mainly, with Black Friday and Christmas preparation, and early summer. This information is very useful for us when knowing when to launch a new product or a powerful new marketing action.

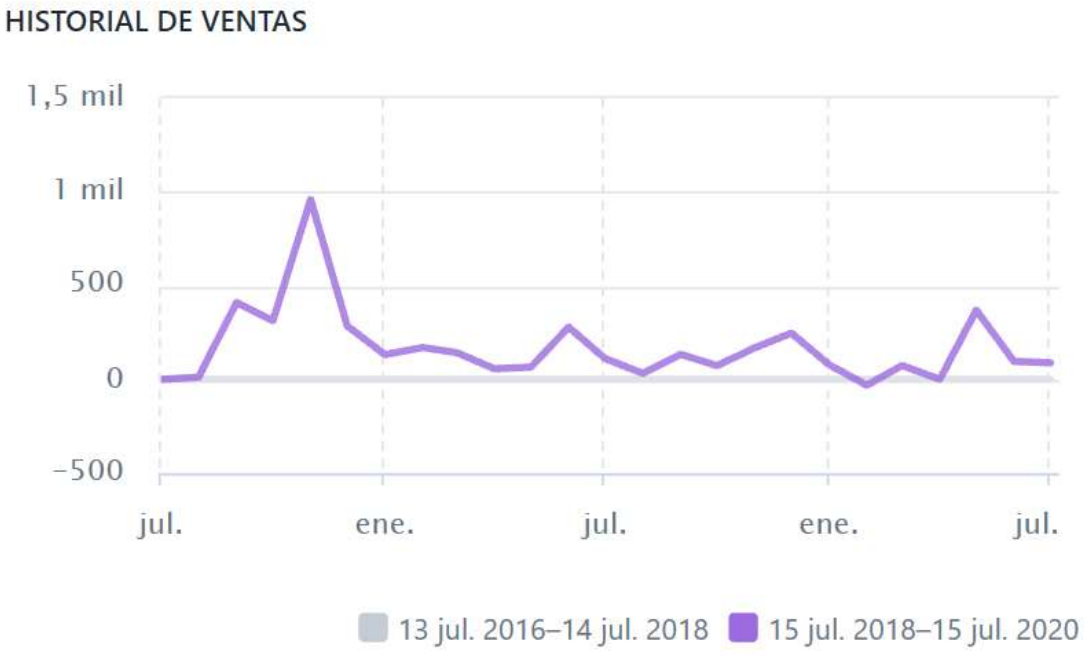

Figure 7. Sales history of Independence Brand.

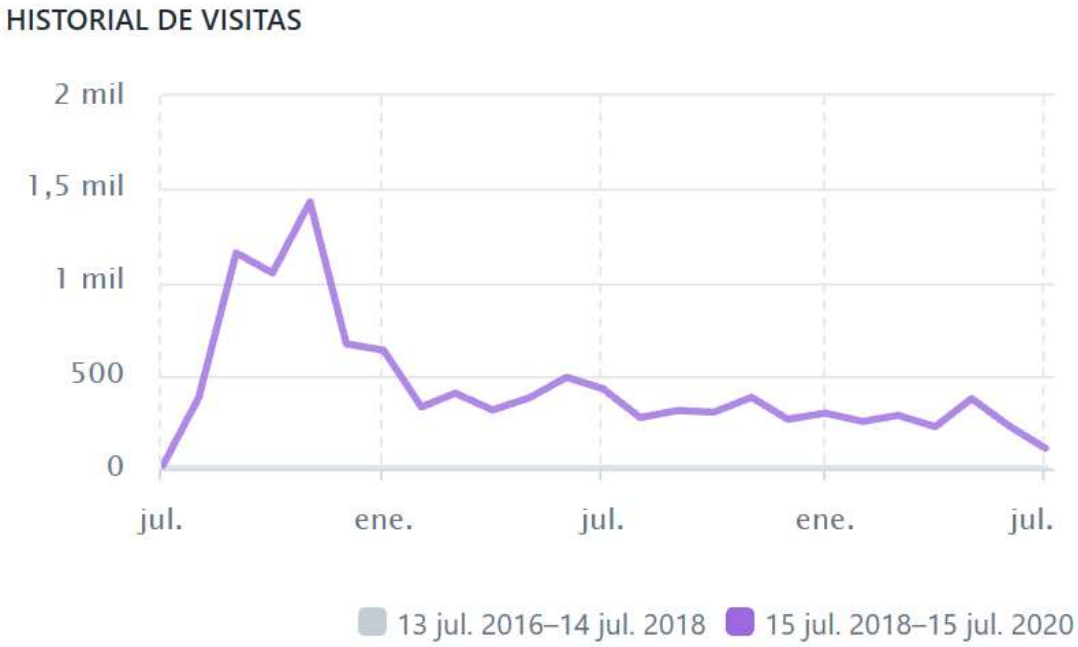

Figure 8. Website views history of Independence Brand.

References

[1] Web of Independence Brand. Last access: 20/07/2020
https://www.independencebrand.com/

[2] Web of Shopify. Las Access 20/07/2020 https://www.shopify.com/

[3] Pawan Kumar Singh, Yadunath Gupta, Nilpa Jha and Aruna Rajan. 2019. "Fashion Retail Forecasting Demand for New Items".

[4] "Análisis de componentes principales" – Joaquín Amat. Last access: 20/07/2020
https://www.cienciadedatos.net/documentos/35_principal_component_analysis

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Capítulo 1. Introducción

1.1 Motivación

Desde que diera comienzo la tercera revolución industrial se han hecho multitud de avances gracias a Internet, no solo tecnológicos, sino también sociales y culturales. Internet ha cambiado las reglas de cómo funciona el mundo, desde la forma de interactuar entre las personas hasta los modelos de organización dentro de una empresa, pasando por la relación cliente-proveedor, caracterizada por la proliferación de los denominados *ecommerce* o comercios online.

Ahora nos vemos envueltos en las transformaciones asociadas a la cuarta revolución industrial, prosiguiendo con la búsqueda de la eficiencia que perseguían las tres revoluciones anteriores, pero esta vez caracterizada, entre otros aspectos, por el uso y tratamiento de los datos. En el ámbito del *ecommerce,* este tratamiento de los datos recibidos a través de redes sociales o del propio portal de venta es utilizado para que las empresas sean más eficientes, especialmente en las áreas de producción, gestión de inventario, precios y relaciones con el cliente. Hoy en día, no sólo se trata de analizar los datos históricos de la demanda, sino de comprender e integrar en los modelos de las empresas, el impacto de otras variables importantes como el clima, las promociones comerciales, los eventos de marketing, las nuevas leyes y demás factores económicos. La capacidad de los ordendores nos permite actualmente almacenar y procesar grandes cantidades de datos con rapidez para poder obtener estimaciones fiables de la demanda. La necesidad de estar preparado y contribuir a semejante cambio supone la motivación principal de este proyecto.

1.2 Objetivos de la predicción de la demanda y del estudio de las características de los consumidores

El pronóstico de la demanda consiste en el proceso de predecir ventas futuras mediante el uso de datos históricos con el objetivo de tomar decisiones comerciales en diversas áreas, desde la planificación del inventario y las necesidades de almacenamiento hasta la ejecución de promociones y la satisfacción de las expectativas del cliente. Pretende conectar la cadena de suministro de una empresa, los hábitos de compra de sus clientes y los factores externos para alcanzar una estimación basada en datos de las ventas futuras esperadas.

Sin demanda no hay negocio y, sin una comprensión profunda de la demanda, las empresas no son capaces de tomar las decisiones correctas sobre gastos de marketing,

producción, personal y demás áreas de operación. Los objetivos principales que persigue la predicción de la demanda son:

- *Estructurar presupuestos:* la predicción de la demanda ayuda a reducir los riesgos y a tomar decisiones financieras eficientes que afectan al resultado económico, al flujo de caja, a la asignación de recursos, a las oportunidades de expansión, a la gestión del inventario, a los costos operativos, al personal y a los gastos generales. Todos los planes estratégicos y operativos se formulan en torno a la demanda prevista.

- *Planificar y programar la producción:* la predicción de la demanda permite el suministro de los productos que desean los clientes, cuando los desean. Para ello se requiere que el cumplimiento del pedido se sincronice con su comercialización antes del lanzamiento. La predicción adecuada de la demanda y la correcta gestión del inventario pueden ayudar a garantizar que una empresa no mantenga un inventario insuficiente o, por el contrario, excesivo.

- *Organización del inventario:* la predicción de la demanda puede ayudar a reducir el gasto tanto en las órdenes de compra como en su almacenamiento, ya que cuanto más inventario se tenga, más costoso será almacenarlo. Una correcta gestión de inventario implica tener suficiente producto listo para su incorporación a la cadena de producción, pero no demasiado. Seguir de cerca los niveles de inventario permite reponer y pronosticar fácilmente el inventario a lo largo del tiempo.

- *Definir una estrategia de precios:* la predicción de la demanda no se trata solo de perfeccionar el cronograma de producción de una empresa para satisfacer la demanda, sino que también debe contribuir a fijar el precio de los productos en función de la demanda. Al comprender el mercado y las oportunidades potenciales, las empresas pueden crecer, formular precios competitivos, implantar las estrategias de marketing adecuadas e invertir en su crecimiento. Si se opta por reducir los precios o poner un artículo en promoción, la demanda puede aumentar temporalmente para ese producto. Si existe una oferta limitada para una gran demanda, puede aplicarse el principio de escasez para aumentar el precio presentado como una oferta exclusiva. Sin embargo, se deben vigilar los nuevos participantes, ya que la oferta por parte de la competencia puede aumentar. [23]

Es importante resaltar que, como todas las previsiones, la predicción de la demanda no resultará precisa al 100%, pero hay pasos que se pueden seguir para mejorar los plazos de producción, aumentar la eficiencia operativa, ahorrar costes, lanzar nuevos productos y proporcionar una mejor experiencia al cliente, como se ha comentado previamente.

Esta predicción de la demanda debe complementarse con un exhaustivo estudio de las características de los consumidores. Estas resultan claves para encontrar el camino hacia la fidelización de los clientes. Una vez fidelizado un cliente, este se convierte en el embajador de la marca, promoviendo la comercialización entre conocidos, amigos y familiares. No solo retener y fidelizar a los clientes se constituyen como las claves del éxito de una compañía, también se debe aumentar y expandir esta red de compradores. En esta línea, resulta crucial conocer las características de los consumidores ya fidelizados para saber cómo y dónde lanzar nuevas campañas de marketing con el objetivo de captar nuevos clientes.

Además, conocer las peculiaridades de los consumidores también se debe entender como un paso hacia una mayor eficiencia en las operaciones, al igual que con la previsión de la demanda. Direccionar todas las acciones de marketing de una compañía hacia el entorno de sus clientes genera grandes ahorros y garantiza una imagen más sólida y consistente de la marca dentro del mercado, ya que al conocer los hábitos y criterios del comprador se trabaja para un grupo concreto, y no para cualquier consumidor.

1.3 Últimos avances en el estudio de las características de los consumidores y predicción de la demanda

Como primera conclusión a partir de una revisión de diversos artículos y referencias en internet sobre "Pronóstico de la demanda", se observa que la mayoría ellos usan métodos de predicción univariables. Los algoritmos tradicionales de predicción de la demanda están siendo influenciados en gran medida por los métodos univariables más modernos, como los métodos de suavizado exponencial [1] y los modelos ARIMA [2].

Como se puede observar en muchos casos, la predicción de la demanda en entornos de *ecommerce* no solo presenta situaciones con tendencias regulares, sino que comúnmente necesita abordar desafíos tales como tendencias de ventas irregulares, presencia de datos de ventas muy dispares y dispersos, etc. Para solventarlo, se han realizado numerosos estudios para mitigar las limitaciones de los enfoques clásicos en estas condiciones desafiantes. Esto incluye la introducción de técnicas de preprocesamiento [3], métodos de ingeniería de características [4], [5], y funciones de probabilidad modificadas [6], [7].

Como se ha comentado previamente, una limitación importante de las técnicas de predicción de la demanda univariable es que son incapaces de usar información de más de una variable de entrada. Por ello, muchos estudios basados en redes neuronales (RNN), que son reconocidos como una alternativa solvente frente los enfoques tradicionales, han estado empleando dichos métodos RNN en forma de una técnica de predicción univariable [8]-[10].

Recientemente, métodos para construir modelos globales a través de series de tiempo relacionadas han logrado buenos resultados. Trapero et al. [11] introdujo modelos de regresión en conjuntos de series de tiempo relacionadas. Estos modelos mejoran la precisión de la predicción en situaciones donde los datos históricos de ventas están limitados a una sola variable de entrada.

Chapados [15] logra buenos resultados en el ámbito de la planificación de la cadena de suministro modelando múltiples series de tiempo usando un marco bayesiano. Más recientemente, técnicas como RNN y CNN también han demostrado para ser competitivas en esta área [12] - [14], [16].

En el marco probabilístico de predicción de la demanda introducido por Flunkert, Salinas. & Gasthaus, [13], y Wen et al. [14] se intenta abordar el factor de incertidumbre del pronóstico. Los autores usan modelos RNN y LSTM para aprender de diferentes grupos de variables y proporcionar estimaciones. Además, Bandara et al. [16] desarrolla un marco de predicción basado en clústering que considera situaciones donde existan grupos de variables heterogéneas.

En la literatura referente a clústering, resulta bastante común usar el algoritmo "K-Means". Diversos proyectos difieren en la manera de preparar los datos para un algoritmo más eficiente, en [17] se usa exitosamente la metodología de "Análisis de Componentes Principales" para una reducción previa de las dimensiones y, de esta manera, encontrar unos centros de clúster iniciales. En [18] los datos son inicialmente procesados para transformarlos en positivos, después ordenados y divididos en k series iguales para posteriormente coger el valor medio de cada serie como centros iniciales. Baridam [19] propone en su trabajo una solución dinámica, de tal manera que automáticamente se determina el número apropiado de clústers, incrementando así la eficiencia total.

En particular, en cuanto a la predicción de la demanda en el sector de la moda, los autores indican que este sector presenta cierta complejidad adicional dada la naturaleza transitoria de muchos de sus factores, como son las tendencias, estampados o patrones, así como las variaciones geográficas en el consumo. En [20] prefieren abordar el problema con modelos más modernos frente a las técnicas tradicionales usadas por las corporaciones. Sin embargo, tras usar modelos LSTM y MLP (Multi Layer Perceptron) y modelos basados en árboles se dieron cuenta de que la actuación de los modelos DNN no había sido la esperada, mientras que los basados en árboles tuvieron un buen rendimiento. Otros autores [21][22] han obtenido buenos resultados mediante el uso de ANN, sin embargo, estos mismos autores comentan que, aunque se logran resultados muy precisos, el tiempo de proceso requerido puede ser una gran barrera para su aplicación en el mundo real. Esto se debe a que en estos modelos el tiempo de entrenamiento requerido aumenta bastante según la complejidad o variedad de datos, por lo que el sector de la moda quizás no es el óptimo.

1.4 Objetivos del proyecto

El objetivo de este trabajo es el de predecir la demanda del *ecommerce* Independence Brand, plataforma de comercio electrónico creada por el autor y auspiciada por la asociación StartComillas , así como realizar un estudio de caracterización de sus clientes mediante el estudio de sus datos históricos de ventas y la aplicación de diferentes técnicas de predicción de la demanda y el algoritmo K-Means.

1.5 Metodología de trabajo

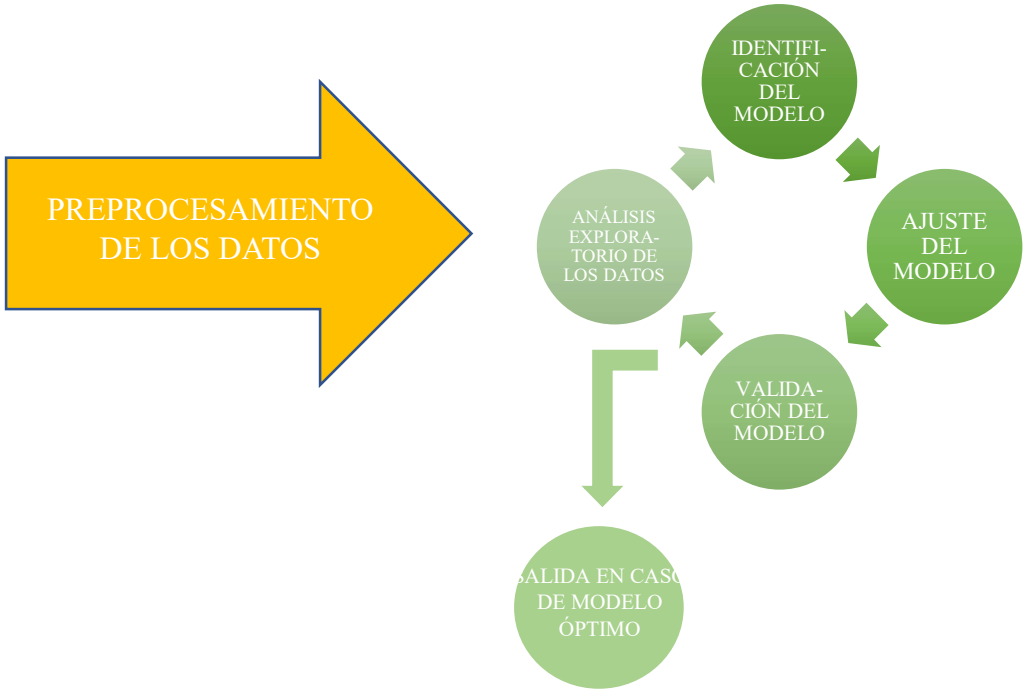

Figura 1. Representación de la metodología de desarrollo del proyecto.

La primera etapa consiste en una recolección y posterior limpieza de los datos, a fin de sentar unas buenas bases para obtener una mayor precisión en los resultados finales. Posteriormente, se lleva a cabo el estudio de las características de los consumidores mediante la aplicación del algoritmo K-Means y, una vez realizada la caracterización de los consumidores, da comienzo el estudio de la predicción de la demanda. Para realizar esta predicción de la demanda se ha desarrollado un estudio de la estacionalidad de la demanda relacionándola con los momentos de lanzamiento de promociones. Lo ideal sería desarrollar un modelo de predicción más detallado, lo cual se deja para futuras líneas de trabajo.

1.6 Recursos a emplear

Para llevar a cabo el proyecto los recursos que se usarán son:

1. El lenguaje de programación Python instalado a partir de Anaconda [31] como entorno de desarrollo integrado.
2. La plataforma de creación de páginas web Shopify [32].
3. Datos históricos de ventas del *ecommerce* Independence Brand [33].

Capítulo 2. Análisis de las diferentes técnicas

2.1 Aprendizaje no supervisado

El estudio de las características de los consumidores se lleva a cabo mediante técnicas de aprendizaje no supervisado. Este se distingue del supervisado en que no existen conocimientos a priori. El aprendizaje no supervisado es bastante común en casos en los que es necesario una reducción previa de las dimensiones de los datos a analizar y para la realización de análisis exploratorio.

En lo que se refiere a la reducción de dimensiones, aunque se han alcanzado grandes avances en términos de potencia de procesamiento y capacidad de almacenamiento en ordenadores, todavía tiene sentido mantener los conjuntos de datos lo más pequeños y sencillos posibles. La reducción de las dimensiones (número de columnas hay en el conjunto de datos) se basa en muchos de los conceptos que la teoría de la información: supone que muchos datos son redundantes y que se puede representar la mayor parte de la información con un conjunto de datos que incluya solo una fracción del contenido total. En la práctica, esto significa combinar subconjuntos de datos de maneras únicas para transmitir significado.

Para el análisis exploratorio, el aprendizaje no supervisado es realmente útil porque puede identificar automáticamente la estructura de los datos. En situaciones en las que es imposible o poco práctico para un ser humano proponer tendencias en los datos, el aprendizaje no supervisado puede proporcionar información inicial que luego puede usarse para validar hipótesis individuales. La técnica más usada es la denominada *clustering*. [28]

El clustering consiste en agrupar un conjunto de individuos de tal manera que los miembros de un mismo grupo, llamado clúster, sean similares en un sentido u otro. El análisis de grupos no es en sí un algoritmo específico, sino la tarea pendiente de solución.

Se puede hacer el agrupamiento utilizando varios algoritmos que difieren significativamente en su idea de qué constituye un grupo y cómo encontrarlos eficientemente. Las ideas clásicas de grupo incluyen distancias pequeñas entre los miembros del mismo, áreas densas del espacio de datos, intervalos o distribuciones estadísticas particulares. El agrupamiento, por tanto, puede ser formulado como un problema multi-objetivo de optimización. El algoritmo apropiado y los valores de los parámetros (incluyendo valores como la función de distancia para utilizar, un umbral de densidad o el número de grupos esperado) dependen del conjunto de datos que se analiza y el uso que se le dará a los resultados. El agrupamiento como tal no es una tarea automática, sino un proceso iterativo de minería de datos o interactivo de optimización multi-objetivo que implica prueba y fracaso. A menudo, será necesario hacer un pre-procesamiento de los datos y un ajuste de los parámetros del modelo hasta que el resultado tenga las propiedades deseadas. Existen diversos algoritmos de clustering, se debe tener en cuenta que el algoritmo más apropiado para un problema particular a menudo necesita ser escogido experimentalmente, a no ser que haya una razón matemática para preferir un modelo de grupo sobre otro. Entre las técnicas más usadas se encuentran el clústering jerárquico, K-means, clústering de mezcla Gaussiana y clústering DBSCAN. A continuación, se describe de forma resumida cada una de las citadas técnicas:

- *Clústering jerárquico*: también conocido como agrupamiento basado en conectividad, está basado en la idea principal de que los objetos más cercanos están más relacionados que los que están alejados. Estos algoritmos conectan "objetos" para formar "los grupos" basados en su distancia. Un grupo puede ser descrito, en gran parte, por la distancia máxima que se necesitó para conectar todas las partes del grupo. A distancias diferentes, se formarán grupos diferentes, los cuales pueden ser representados utilizando un dendrograma, el cual explica de donde proviene el nombre "agrupamiento jerárquico": estos algoritmos no solo proporcionan una partición del conjunto de datos, sino en cambio, proporcionan una jerarquía extensa de grupos que se fusionan con cada otro a ciertas distancias.

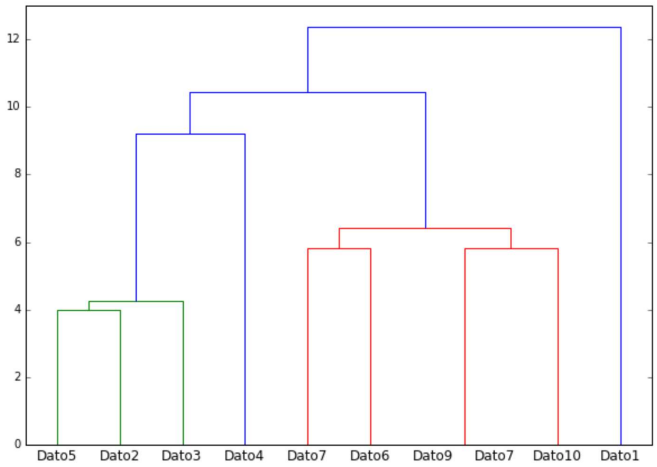

Figura 2. Representación de clústering jerárquico.

En la Figura 8 se muestra un ejemplo de dendograma. En un dendrograma, el eje "y" marca la distancia en qué los grupos se fusionan, mientras que los objetos están colocados a lo largo del eje "x", tal que los grupos se mezclan.

- *K-Means*: agrupamiento basado en centroide, los grupos están representados por un vector central, el cual puede no ser necesariamente un miembro del conjunto de datos. Cuando el número de grupos está fijado en k, k-means da una definición formal como un problema de optimización: encontrar los k centros de los grupos y asignar los objetos al centro del grupo más cercano, tal que el cuadrado de las distancias del grupo al centro están minimizadas. Aun así, sólo encuentra un óptimo local, y generalmente se ejecuta varias veces con inicializaciones aleatorias. Variaciones de k-means a menudo incluyen otras optimizaciones como: escoger el mejor resultado de varias corridas, restringir el centroide a miembros del conjunto de datos (k-medoids), escoger medianas (k-medians), escoger los centros iniciales aleatoriamente (K-Means++) o permitir una asignación de grupos difusa (Fuzzy c-means). A continuación se muestra un ejemplo de ejecución del algoritmo estándar. [30]

    1) k centroides iniciales (en este caso k=3) son generados aleatoriamente dentro de un conjunto de datos (mostrados en color).

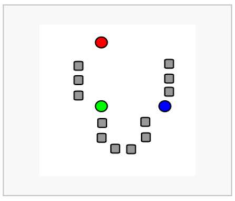

Figura 3. Paso 1 K-means.

    2) k grupos son generados asociándole el punto con la media más cercana. La partición aquí representa el diagrama de Voronoi generado por los centroides.

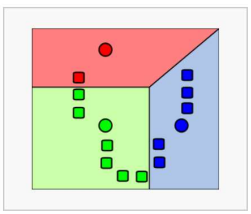

Figura 4. Paso 2 K-means.

    3) El centroide de cada uno de los k grupos se recalcula.

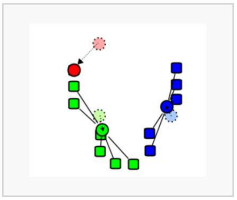

Figura 5. Paso 3 Kmeans.

4) Se repiten los pasos segundo y tercero hasta que se logre la convergencia.

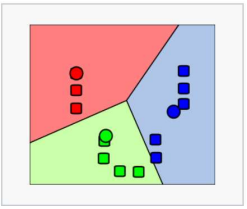

Figura 6. Paso 4 K-means.

- *Clústering de mezcla Gaussiana:* el conjunto de datos es normalmente modelado con un número fijo (para evitar el sobreajuste u overfitting) de distribuciones Gaussianas que está inicializado aleatoriamente, y cuyos parámetros son iterativamente optimizados para clasificar mejor al conjunto de datos. Esto convergerá a un óptimo local, múltiples corridas pueden producir resultados diferentes. Para obtener un agrupamiento sólido, los objetos son a menudo entonces asignados a la distribución Gaussiana con mayor probabilidad de pertenecer; para agrupamiento suave, esto no es necesario. Esta técnica produce modelos complejos para grupos que pueden capturar correlación y dependencia entre atributos. Aun así, estos algoritmos ponen una carga extra en el usuario: para muchos conjuntos de datos reales, no puede haber ningún modelo matemático definido. A continuación, se muestra un ejemplo de clústering de mezcla Gaussiana.

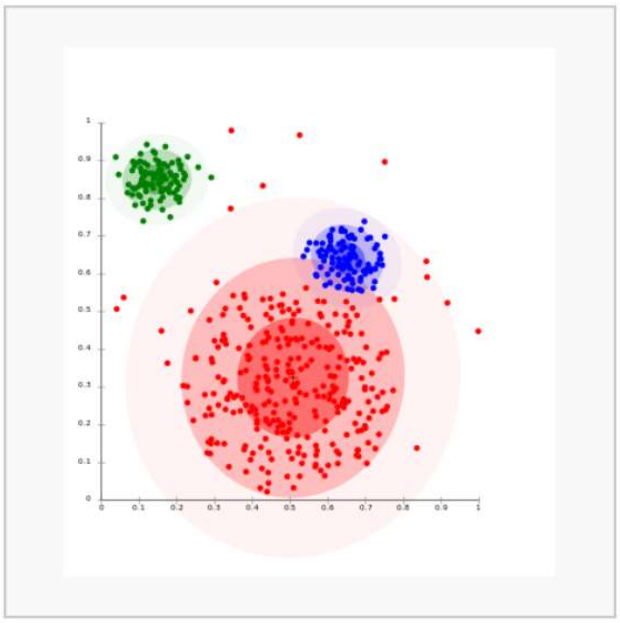

Figura 7. Representación de clústering de mezcla Gaussiana.

- *Clústering DBSCAN*: También llamado agrupamiento basado en densidad. En contraste con muchos métodos más nuevos, presenta un modelo de grupo bien definido llamado "densamente alcanzable". Similar al agrupamiento basado en conectividad, está basado en conectar puntos dentro de cierto umbral de distancia. Aun así, sólo conecta aquellos puntos que satisfacen un criterio de densidad, en la variante original definido como número mínimo de otros objetos dentro de un radio dado. Un grupo consiste en objetos densamente conectados (los cuales pueden formar un grupo de una forma arbitraria, en contraste a muchos otros métodos) más todos los objetos que están dentro del rango de estos. Otra propiedad interesante de DBSCAN es que su complejidad es bastante baja - requiere un número lineal de consultas de rango en la base de datos - y que descubrirá esencialmente los mismos resultados en cada corrida, por tanto, no hay ninguna necesidad de correrlo varias veces. A continuación, se muestra un ejemplo de clústering DBSCAN. [29]

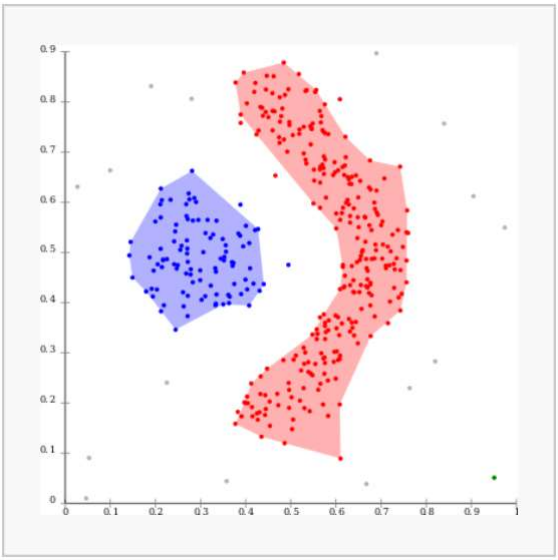

Figura 8. Representación de clustering DBSCAN.

2.2 Técnicas de predicción de la demanda

La cuestión de la predicción de la demanda puede tener dos diferentes aproximaciones: cualitativa y cuantitativamente. Los métodos cualitativos requieren la existencia de datos históricos y pueden predecir un largo periodo de tiempo. Por el contrario, los métodos cuantitativos no requieren de ningún tipo de dato previo, ya que se basan en encuestas y solo predicen el corto plazo. Los métodos cualitativos son más comunes cuando no se dispone de datos pasados o estos son inapropiados. En este trabajo, se abordarán las técnicas de predicción cuantitativas.

Uno de los primeros pasos a la hora de estudiar un problema de predicción de la demanda es la representación correcta de los datos con el objetivo de identificar tendencias o patrones. Las diferentes representaciones pueden mostrar cuatro patrones combinables entre sí: tendencia, componente estacional, componente cíclica y componente aleatoria.

- Tendencia: se observa cuando hay un claro incremento o decremento a largo plazo en los datos.

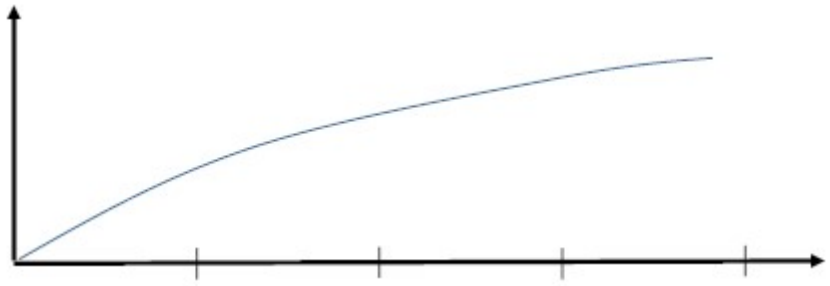

Figura 9. Gráfico de tendencia.

- Estacional: se observa cuando una secuencia de datos está fuertemente influenciada por factores estacionales, es decir, cuando se repiten datos en determinados periodos de tiempo de forma recurrente.

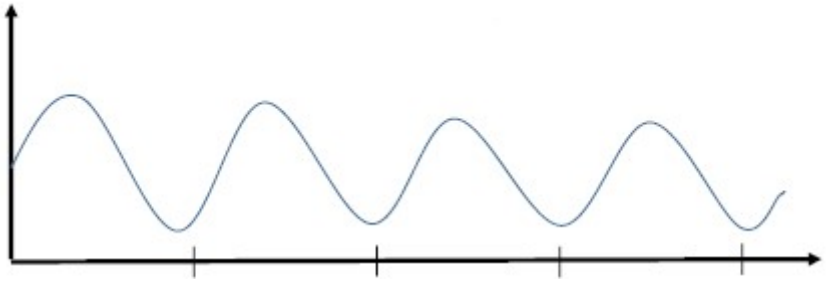

Figura 10. Gráfico de estacionalidad.

- Cíclico: se observa cuando los datos experimentan subidas y bajadas de forma cíclica sin estar influenciados por aspectos estacionales.

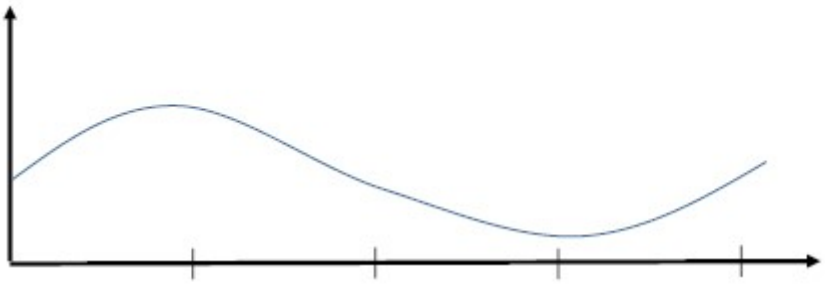

Figura 11. Gráfico de carácter cíclico.

- Aleatoria: se observa cuando los datos no muestran ningún patrón o tendencia.

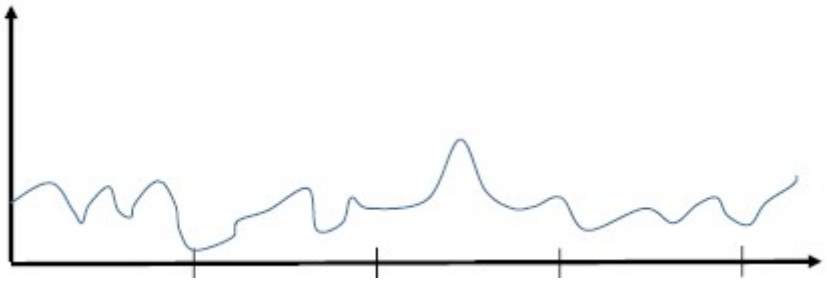

Figura 12. Gráfico de carácter aleatorio.

Una vez realizada la representación y depuración de los datos, se procede a la aplicación de una de las diferentes técnicas existentes. Estas son comúnmente divididas en univariables y multivariables. El término univariable hace referencia a una serie temporal que consiste en observaciones individuales (escalares) registradas secuencialmente en incrementos de tiempo iguales, mientras que el multivariable es una extensión del caso univariable e involucra dos o más variables de entrada. No se limita a su información pasada, sino que también incorpora el pasado de otras variables. Los procesos multivariados surgen cuando varias series de tiempo relacionadas se observan

simultáneamente a lo largo del tiempo, en lugar de observar una sola serie como en el caso univariable. Surgió de la necesidad de estudiar la interrelación entre las variables de series de tiempo. Estas relaciones a menudo se estudian mediante la consideración de las estructuras de correlación entre las series de componentes.

Entre las técnicas univariables más relevantes encontramos:

- *ARIMA*: es el acrónimo de AutoRegressive Integrated Moving Average. A partir de esta técnica, se identifican las claves del modelo que se enumeran a continuación:

  - Autorregresión (AR): emplea la relación dependiente entre una observación y cierto número de observaciones anteriores.
  - Integrado (I): uso de la diferenciación de observaciones sin procesar para hacer estacionarias las series de tiempo.
  - Media móvil (MA): utiliza la dependencia entre una observación y el promedio móvil aplicado a observaciones anteriores.

  Cada una de estas claves están explícitamente definidas por los correspondientes parámetros en el modelo. Los parámetros estándar usados son los siguientes:

  - p: número de observaciones anteriores incluidas en el modelo, también llamado orden de retraso.
  - d: número de veces que las observaciones sin procesar son diferenciadas, también llamado grado de diferenciación.
  - q: tamaño de la ventana de la media móvil, también llamado orden de media móvil.

  Con estas claves y parámetros, se desarrolla un modelo de regresión lineal que incluye el número y el tipo de términos especificados, y los datos se preparan mediante un grado de diferenciación para hacerlo estacionario, es decir, para eliminar tendencias y estructuras estacionales que afectan negativamente al modelo de regresión. [24]

- *RNN*: es el acrónimo de Recurrent Neural Network. La idea principal detrás de estas redes es hacer uso de información secuencial. En una red neuronal tradicional asumimos que todas las entradas y salidas son independientes, pero en muchas ocasiones no es posible esta suposición. Las RNNs son llamadas recurrentes porque desarrollan la misma tarea para cada elemento de la secuencia, siendo la salida dependiente de las operaciones previas. En teoría, las RNNs pueden hacer uso de la información en secuencias largas de manera arbitraria, pero en la práctica están limitadas a secuencias más cortas.

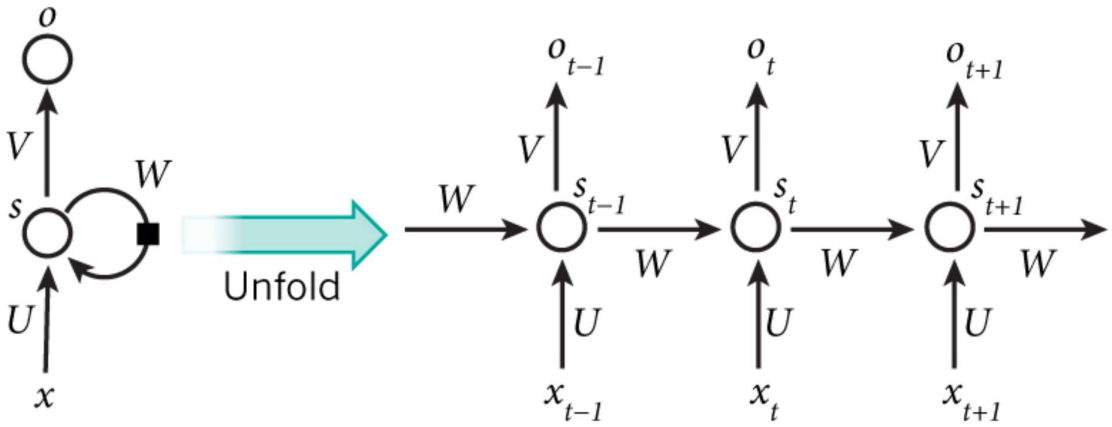

Figura 13. Metodología RNN.

El diagrama de la Figura 6 muestra una RNN desplegada en una red completa, es decir, con la secuencia completa mostrada. Las variables que gobiernan las operaciones en una RNN son las siguientes:

- x_t es la entrada para un tiempo t. Por ejemplo, x_1 podría ser un vector único que corresponde a la segunda palabra de una oración.
- s_t es el estado oculto para un tiempo t. Es la "memoria" de la red. s_t se calcula en función del estado oculto anterior y la entrada en el paso actual:

  $$s_t = f(Ux_t + Ws_{t-1})$$

  La función f es una función no lineal por lo general, como tanh o ReLU. s_{-1} se requiere para calcular el primer estado oculto, generalmente se inicializa a cero.
- o_t es la salida en el tiempo t. Por ejemplo, si quisiéramos predecir la siguiente palabra en una oración, sería un vector de probabilidades en nuestro vocabulario. [25]

  $$o_t = \text{softmax}(Vs_t)$$

- *LSTM*: es el acrónimo de Long-Short Term Memory. Es un tipo de red neuronal recurrente, con la diferencia de que LSTM presenta conexiones de retroalimentación. La unidad se llama bloque de memoria Long-Short Term porque el programa está utilizando una estructura basada en procesos de memoria a corto plazo para crear memoria a más largo plazo. Existen diversos tipos, incluyendo también algunos multivariables. [26]

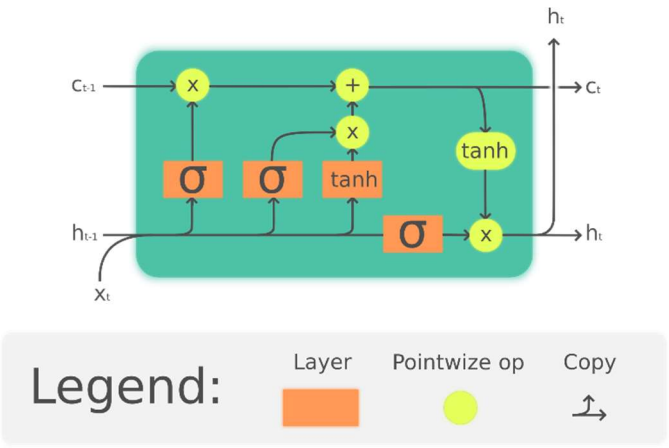

Figura 14. Metodología LSTM.

Entre las técnicas multivariables más relevantes encontramos:

- *VAR*: acrónimo de Vector AutoRegression. En un modelo VAR, cada variable es función lineal de sus datos pasados y de los de las otras variables. Por ejemplo, para dos variables:

$$y_1(t) = a_1 + w_1 11 * y_1(t-1) + w_1 12 * y_2(t-1) + e_1(t-1)$$

$$y_2(t) = a_2 + w_1 21 * y_1(t-1) + w_1 22 * y_2(t-1) + e_2(t-1)$$

Donde a_1 y a_2 son constantes, w_{11}, w_{12}, w_{21} y w_{22} son coeficientes y e_1 y e_2 son los errores.
Estos modelos multivariables son realmente útiles cuando se tiene que tratar con muchas variables dependientes entre sí. [27]

Capítulo 3. Independence Brand

Figura 15. Logo de Independence Brand.

En octubre del 2017, Javier Moreno y el autor, compañeros y amigos del ICAI, con muchas ganas de aventura y de explorar lo inexplorado, nos decidimos a lanzar Independence Brand. Los dos teníamos muchas ganas de emprender desde que habíamos empezado el grado y tras realizar varios cursos de emprendimiento y pasar a formar parte de la Asociación de Emprendedores de Comillas, StartComillas [34] nos decidimos a impulsar nuestro propio proyecto.

Independence Brand es un *ecommerce* (www.independencebrand.com) que se dedica principalmente a la comercialización de moda joven. En los últimos tiempos se nos ha considerado a los millennials (y postmillennials) una generación perdida, naciendo de allí uno de los objetivos de Independence Brand de romper con esa idea y demostrar que somos una generación con objetivos personales y compromiso social. Con esta perspectiva de poner en valor el potencial que llevamos dentro, avanzamos poniendo en práctica los valores que nos representan:

#1 Ganas de mejorar las cosas.

Somos unos inconformistas.

#2 Pasión por el emprendimiento y la innovación.

Mejoramos a partir de nuestros errores, siguiendo el modelo de Lean Startup.

#3 Sentimiento de cercanía y buen trato.

Nos gusta cuidar de nuestra comunidad Independence.

#4 Búsqueda de la aventura, abriendo nuevos horizontes.

Intentamos ir a por lo diferente, pensando de una forma disruptiva.

#5 Actitud de esfuerzo y trabajo en equipo.

Juntos conseguimos llegar más lejos.

#6 Compromiso de solidaridad con nuestro entorno.

Colaboramos con ONG para aportar nuestro granito de arena.

#7 Pasión por lo que hacemos.

Nos encanta disfrutar de nuestro trabajo cada día en Independence.

En definitiva, Independence Brand no es solo una marca de ropa, sino que representa el estilo de vida de una generación orgullosa de sí misma y que está destinada a cambiar el mundo. Nosotros, a esta generación la denominamos "Generación Independence".

Capítulo 4. Desarrollo del Proyecto

4.1 Procedencia de los datos.

Los datos usados en este proyecto han sido extraídos de la información de ventas del *ecommerce* Independence Brand. Los datos fueron exportados desde la plataforma de creación de webs Shopify [33], la cual brinda la opción de exportarlos en formato *.csv*.

4.2 Obtención y preprocesamiento de los datos

Los datos exportados desde Shopify, plataforma donde hospeda su canal de venta Independence Brand, recogen las operaciones de la marca durante los últimos dos años, en una base de datos en formato .csv con 274 registros.

	A
1	Name,Email,Financial Status,Paid at,Fulfillment Status,Fulfilled at,Accepts Marketing,Currency,Subtotal,Shipping,Taxes,Total,Discount Code,Discount Amount,Shipping Method,Created at,Lineitem quantity,Lin
2	#1138,thiebaut.rafael@hotmail.com,paid,2020-01-29 00:32:01 +0100,fulfilled,2020-01-29 00:32:05 +0100,no,EUR,6.00,3.95,1.04,9.95,"",6.00,"EnvÃo EstÃ¡ndar â€" EnvÃo gratuito a partir de 29,99â‚¬",202
3	#1138,thiebaut.rafael@hotmail.com,,,,,,,,,,,,,,2020-01-29 00:32:01 +0100,1,CALCETINES RED SEA - 39-45,6.00,6.00,"",true,true,fulfilled,,,,,,,,,,,,,,,,,,,,,,,,,Independence Brand,,,,,,,,,0.00,,,,,,,,,,,,
4	#1137,carlos.argos2000@gmail.com,paid,2020-01-24 19:21:43 +0100,fulfilled,2020-01-24 19:21:47 +0100,yes,EUR,0.00,3.95,0.00,3.95,Tarjeta_regalo,33.80,"EnvÃo EstÃ¡ndar â€" EnvÃo gratuito a partir de
5	#1137,carlos.argos2000@gmail.com,,,,,,,,,,,,,,2020-01-24 19:21:43 +0100,1,CAMISETA Creamy Cookie - XL,13.90,"","",true,true,fulfilled,,,,,,,,,,,,,,,,,,,,,,,,,Independence Brand,,,,,,,,,0.00,,,,,,,,,,,,
6	#1137,carlos.argos2000@gmail.com,,,,,,,,,,,,,,2020-01-24 19:21:43 +0100,1,CAMISETA Pistachio Spice - XXL,13.90,"","",true,true,fulfilled,,,,,,,,,,,,,,,,,,,,,,,,,Independence Brand,,,,,,,,,0.00,,,,,,,,,,,,
7	#1136,javiergargon@hotmail.es,paid,2020-01-24 14:44:12 +0100,fulfilled,2020-01-24 14:44:17 +0100,no,EUR,6.00,3.95,1.04,9.95,"",6.00,"EnvÃo EstÃ¡ndar â€" EnvÃo gratuito a partir de 29,99â‚¬",2020-0
8	#1136,javiergargon@hotmail.es,,,,,,,,,,,,,,2020-01-24 14:44:11 +0100,1,CALCETINES ADRIATIC SEA - 39-45,6.00,8.00,"",true,true,fulfilled,,,,,,,,,,,,,,,,,,,,,,,,,Independence Brand,,,,,,,,,6.00,,,,,,,,,,,,
9	#1135,javiergargon@hotmail.es,paid,2020-01-24 14:36:34 +0100,fulfilled,2020-01-24 14:36:37 +0100,no,EUR,19.00,3.95,0.00,22.95,"",19.00,"EnvÃo EstÃ¡ndar â€" EnvÃo gratuito a partir de 29,99â‚¬",202
10	#1135,javiergargon@hotmail.es,,,,,,,,,,,,,,2020-01-24 14:36:33 +0100,1,CAMISA POPELÃN BLANCA - XXL,19.00,34.95,"",true,false,fulfilled,,,,,,,,,,,,,,,,,,,,,,,,,Independence Brand,,,,,,,,,0.00,,,,,,,,,,,,
11	#1134,ciracheb@yahoo.es,refunded,2020-01-23 15:27:00 +0100,fulfilled,2020-01-23 15:27:05 +0100,yes,EUR,31.40,0.00,2.15,31.40,"",19.00,EnvÃo Gratuito,2020-01-23 15:26:59 +0100,1,Oxford - 39-45,6.
12	#1134,ciracheb@yahoo.es,,,,,,,,,,,,,,2020-01-23 15:26:59 +0100,1,CALCETINES Edinburgh (ÃšLTIMAS UNIDADES) - 39-45,6.00,8.00,"",true,true,fulfilled,,,,,,,,,,,,,,,,,,,,,,,,,Independence Brand,,,,,,,,,0.00,,,,,,,,,,,,
13	#1134,ciracheb@yahoo.es,,,,,,,,,,,,,,2020-01-23 15:26:59 +0100,1,CAMISA POPELÃN AZUL - XL,19.00,34.95,"",true,false,fulfilled,,,,,,,,,,,,,,,,,,,,,,,,,Independence Brand,,,,,,,,,19.00,,,,,,,,,,,,
14	#1134,ciracheb@yahoo.es,,,,,,,,,,,,,,2020-01-23 15:26:59 +0100,1,CAMISA POPELÃN BLANCA - XL,19.00,34.95,"",true,false,fulfilled,,,,,,,,,,,,,,,,,,,,,,,,,Independence Brand,,,,,,,,,0.00,,,,,,,,,,,,
15	#1133,lidia.pelayo@outlook.com,paid,2019-12-30 17:37:12 +0100,fulfilled,2019-12-30 17:37:18 +0100,no,EUR,58.00,0.00,10.07,58.00,"",0.00,EnvÃo Gratuito,2019-12-30 17:37:11 +0100,2,Fragancia Indep
16	#1132,idecristobal@gmail.com,paid,2019-12-28 19:37:46 +0100,fulfilled,2019-12-28 19:37:49 +0100,yes,EUR,35.00,0.00,6.07,35.00,"",0.00,EnvÃo Gratuito,2019-12-28 19:37:45 +0100,1,CALCETINES Adriat
17	#1132,idecristobal@gmail.com,,,,,,,,,,,,,,2019-12-28 19:37:45 +0100,1,Fragancia Independence MAN,29.00,"","",true,true,fulfilled,,,,,,,,,,,,,,,,,,,,,,,,,Independence Brand,,,,,,,,,0.00,,,,,,,,,,,,
18	#1131,mjgarrigos69@gmail.com,paid,2019-12-25 23:06:38 +0100,fulfilled,2019-12-25 23:06:42 +0100,yes,EUR,27.80,3.95,4.82,31.75,"",0.00,"EnvÃo EstÃ¡ndar â€" EnvÃo gratuito a partir de 29,99â‚¬",201
19	#1131,mjgarrigos69@gmail.com,,,,,,,,,,,,,,2019-12-25 23:06:38 +0100,1,CAMISETA Creamy Cookie - S,13.90,"","",true,true,fulfilled,,,,,,,,,,,,,,,,,,,,,,,,,Independence Brand,,,,,,,,,0.00,,,,,,,,,,,,
20	#1130,hdezglezjavier@gmail.com,paid,2019-12-18 10:14:28 +0100,fulfilled,2019-12-18 10:14:30 +0100,yes,EUR,25.02,3.95,4.34,28.97,INDEPENDENCEG9,16.68,"EnvÃo EstÃ¡ndar â€" EnvÃo gratuito a part
21	#1130,hdezglezjavier@gmail.com,,,,,,,,,,,,,,2019-12-18 10:14:27 +0100,1,CAMISETA Coconut Tree - L,13.90,"","",true,true,fulfilled,,,,,,,,,,,,,,,,,,,,,,,,,Independence Brand,,,,,,,,,0.00,,,,,,,,,,,,
22	#1130,hdezglezjavier@gmail.com,,,,,,,,,,,,,,2019-12-18 10:14:27 +0100,1,CAMISETA Creamy Cookie - M,13.90,"","",true,true,fulfilled,,,,,,,,,,,,,,,,,,,,,,,,,Independence Brand,,,,,,,,,0.00

Figura 16. Muestra de datos en formato .csv.

Valiéndome de la herramienta Excel, los datos son importados a un formato .xlsx en el que toda la información y variables son claramente visibles y manejables, como se muestra a continuación.

1	Name	Email	Financial Status	Paid at	Fulfillment Status	Fulfilled at	Accepts Marketing	Currency	Subtotal	Shipping	Taxes	Total	Discount Code
2	#1138	thiebaut.rafael@hotmail.com	paid	29/01/2020 0:32	fulfilled	29/01/2020 0:32	no	EUR	600	395	104	995	
3	#1138	thiebaut.rafael@hotmail.com											
4	#1137	carlos.argos2000@gmail.com	paid	24/01/2020 19:21	fulfilled	24/01/2020 19:21	yes	EUR	0	395	0	395	Tarjeta_regalo
5	#1137	carlos.argos2000@gmail.com											
6	#1137	carlos.argos2000@gmail.com											
7	#1136	javiergargon@hotmail.es	paid	24/01/2020 14:44	fulfilled	24/01/2020 14:44	no	EUR	600	395	104	995	
8	#1136	javiergargon@hotmail.es											
9	#1135	javiergargon@hotmail.es	paid	24/01/2020 14:36	fulfilled	24/01/2020 14:36	no	EUR	1900	395	0	2295	
10	#1135	javiergargon@hotmail.es											
11	#1134	ciracheb@yahoo.es	refunded	23/01/2020 15:27	fulfilled	23/01/2020 15:27	yes	EUR	3140	0	215	3140	
12	#1134	ciracheb@yahoo.es											
13	#1134	ciracheb@yahoo.es											
14	#1134	ciracheb@yahoo.es											
15	#1133	lidia.pelayo@outlook.com	paid	30/12/2019 17:37	fulfilled	30/12/2019 17:37	no	EUR	5800	0	1007	5800	
16	#1132	idecristobal@gmail.com	paid	28/12/2019 19:37	fulfilled	28/12/2019 19:37	yes	EUR	3500	0	607	3500	
17	#1132	idecristobal@gmail.com											
18	#1131	mjgarrigos69@gmail.com	paid	25/12/2019 23:06	fulfilled	25/12/2019 23:06	yes	EUR	2780	395	482	3175	
19	#1131	mjgarrigos69@gmail.com											
20	#1130	hdezglezjavier@gmail.com	paid	18/12/2019 10:14	fulfilled	18/12/2019 10:14	yes	EUR	2502	395	434	2897	INDEPENDENCEG9
21	#1130	hdezglezjavier@gmail.com											
22	#1130	hdezglezjavier@gmail.com											
23	#1129	franciscosospedra1@gmail.com	paid	17/12/2019 1:19	fulfilled	17/12/2019 1:19	yes	EUR	3500	0	278	3500	
24	#1129	franciscosospedra1@gmail.com											

Figura 17. Muestra de datos en formato .xlsx.

Este archivo cuenta con 273 registros (filas), que representan cada uno de los productos vendidos, y 69 campos (columnas), que representan cada una de las variables de cada producto o pedido. Estas variables son:

Name	Email	Financial Status
Paid at	Fulfillment Status	Accepts Marketing
Currency	Subtotal	Shipping
Taxes	Total	Discount Code
Discount Amount	Shipping Method	Created at
Lineitem quantity	Lineitem name	Lineitem Price
Lineitem Compare at Price	Lineitem sku	Lineitem Requires Shipping
Lineitem Taxable	Billing Name	Billing Street

Billing Address1	Billing Address2	Billing Company
Billing City	Billing Zip	Billing Province
Billing Country	Billing Phone	Shipping Name
Shipping Street	Shipping Address1	Shipping Address2
Shipping Company	Shipping City	Shipping Zip
Shipping Province	Shipping Country	Shipping Phone
Notes	Note Attributes	Cancelled at
Payment Method	Payment Reference	Refunded Amount
Vendor	Outstanding Balance	Employee
Location	Device ID	Id
Tags	Risk Level	Source
Lineitem discount	Tax 1 Name	Tax 1 Value
Tax 2 Value	Tax 2 Name	Tax 3 Value
Tax 4 Name	Tax 4 Value	Tax 5 Name
Tax 5 Value	Phone	Receipt Number

Tabla 1. Variables presentes en el conjunto de datos.

El procesamiento y depuración de los datos se realizó en el mismo archivo Excel, pasando por las siguientes fases.

1) Selección de variables representativas.

   No todas las 69 variables aportan información relevante a la hora de hacer un estudio de las características de los consumidores o predecir la demanda de los próximos meses. Por ello, se prescindió de las variables cuyos campos estaban vacíos, como el desglose de los impuestos (Tax 1, 2 ,3…) y "Billing Company", de las variables que no aportaban datos, como "Currency" (Independence Brand se ha limitado a satisfacer las necesidades del territorio nacional), y de las variables que contenían información repetida, como todas las variables del grupo "Shipping". Quedando, de esta manera, las siguientes 13 variables:

   - Name.
   - Email.
   - Financial Status.
   - Accepts Marketing.
   - Total.
   - Discount Code.
   - Discount Amount.
   - Shipping Method.
   - Created at.
   - Lineitem quantity.
   - Lineitem Name.
   - Shipping Province.
   - Payment Method.

2) Eliminación de datos nulos o erróneos.

   Gracias a los filtros que se pueden aplicar en Excel, se localizaron 4 filas erróneas con variables sin información. Tras investigar en Shopify la procedencia de estos pedidos, se llegó a la conclusión que son pedidos de prueba para comprobar el buen funcionamiento de la web o pedidos realizados en eventos presenciales, pero utilizando la web como canal de pago. Los pedidos de prueba fueron eliminados y los comprometidos en eventos, modificados con la información correcta.

3) Agrupación de las filas por pedidos.

   Como se muestra en la imagen inferior, cada fila corresponde a un producto comprado, siendo información menos útil que tener en cada fila el número de pedidos y con celdas vacías.

Name	Email	Financial Status	Lineitem quantit	Lineitem name	Accepts Marketi
#1137	carlos.argos2000@gmail.com	paid	1	CALCETINES RED SE	yes
#1137	carlos.argos2000@gmail.com		1	CAMISETA Creamy	
#1137	carlos.argos2000@gmail.com		1	CAMISETA Pistachic	

Figura 18. Ejemplo de la agrupación de filas por pedidos.

   Con el objetivo de obtener una mayor precisión en el análisis, se agrupa la información de todas las filas referentes a un solo pedido en una sola fila, sumando las cantidades de productos y el total de cada uno, y eliminando la columna de "Lineitem name".

4) Transformación del formato de fecha y creación de nuevas variables temporales.

   La fecha viene predeterminada en un formato de fecha del tipo dd/mm/aaaa y la hora en un formato de 24 horas. Para una mayor precisión en el análisis, se decide dividir la gran cantidad de información que posee la variable "Created at", que recoge el momento en el que se realiza la compra, en tres variables:

   - Week Day: Representa el día de la semana en el que se realizó la compra.
   - Timeframe: Representa el momento del día en el que se realizó la compra. Se divide en cuatro franjas horarias:
     - Primera: comprende las horas desde las 00:00 hasta las 6:00.
     - Segunda: comprende las horas desde las 6:00 hasta las 12:00.
     - Tercera: comprende las horas desde las 12:00 hasta las 18:00.
     - Cuarta: comprende las horas desde las 18:00 hasta las 00:00.
   - Month frame: Representa el momento del mes en el que se realizó la compra.

El mes es dividido en cuatro intervalos:

- Primero: comprende los días del 1 al 7.
- Segundo: comprende los días del 8 al 15.
- Tercero: comprende los días desde el 16 al 21.
- Cuarto: comprende los días desde el 22 al 31.

5) Agrupación de los diferentes códigos de descuento en tres categorías.
   A lo largo de estos dos años se han creado más de 200 códigos de descuento, los cuales se agrupan en tres categorías:

   - *Cambio de producto*: en esta categoría se incluyen aquellos códigos usados con el objetivo de cambiar un producto, sin tener que pagar el precio del nuevo producto.
   - *Promo*: esta categoría incluye los códigos creados de manera eventual, con el objetivo de aumentar las ventas durante festividades como el Black Friday, San Valentín o el Día del Padre.
   - *Ambassador*: bajo esta categoría se recogen los códigos entregados a nuestra red comercial y que son usados por sus respectivos clientes.

6) Creación de un nuevo dataset con variables numéricas.

   Debido al alcance de este proyecto, solo se tienen en cuenta las variables numéricas y se dejan las categóricas para futuras líneas de trabajo, quedando las siguientes variables para los análisis a desarrollar en el presente trabajo:

   - Total.
   - Discount Amount.
   - Timeframe.
   - Month frame.
   - Item quantity.

4.3 Desarrollo del código

4.3.1 Importación de librerías

Para llevar a cabo el análisis deseado es necesario importar las siguientes librerías y funciones:

```
import pandas as pd

from sklearn.cluster import KMeans

from sklearn.decomposition import PCA

from sklearn import preprocessing

from sklearn import cluster

import matplotlib.pyplot as plt

import seaborn as sns
```

Figura 19. Bloque de código para la importación de librerías y funciones.

Gracias a ello, podremos importar los datos del documento Excel, estandarizar los datos, ejecutar el algoritmo K-Means y un Análisis de Componentes Principales, así como representar los resultados. De todas las técnicas de aprendizaje supervisado, finalmente se decidió utilizar el algoritmo K-Means por su popularidad, eficiencia y baja complejidad de implementación.

4.3.2 Importación y estandarización de los datos

```
# Import excel file

df=pd.read_excel('DATOS VENTAS SHOPIFY (TRANSACCIONES).xlsx')


num_list = ['Total', 'Discount_Amount', 'Timeframe', 'Month_frame',

            'Item_quantity']


df_std = df.copy()

for i in num_list:

    df_std[i] = preprocessing.scale(df_std[i])
```

Figura 20. Bloque de código para la importación y estandarización de datos .

En la parte del código mostrada en la figura 19 se pueden observar las órdenes de importación de los datos desde el documento Excel, llamado “df”. Posteriormente, se

procede a la estandarización y centrado de los mismos con la herramienta "preprocessing" y un bucle "for", de tal manera que dicha estandarización se realiza por variables.

4.3.3 Búsqueda del número óptimo de clústers

Una de las desventajas del método de clústering K-means es que necesita especificar el número de clústeres de antemano, normalmente obtenido mediante el método del codo. Este método utiliza los valores de la inercia obtenidos tras aplicar el K-means a diferente número de clústeres, siendo la inercia la suma de las distancias al cuadrado de cada objeto del clúster a su centroide. Una vez obtenidos los valores de la inercia tras aplicar el K-means, representamos en una gráfica lineal la inercia respecto del número de clústeres. El punto en el que se observa un cambio brusco en la inercia nos indica el número óptimo de clústeres.

$$Inercia = \sum_{i=0}^{N} \|x_i - \mu\|^2$$

Figura 21. Fórmula de la inercia aplicada a clústeres.

A continuación, la figura 21 muestra la parte del código desarrollado para la búsqueda del número óptimo de clústeres y el resultado.

```
# Elbow for optimal k clusters

sum_of_sq_dist={}
for k in range(1,15):
    km = KMeans(n_clusters=k, init= 'k-means++', max_iter=1000)
    km = km.fit(df)
    sum_of_sq_dist[k] = km.inertia_

sns.pointplot(x = list(sum_of_sq_dist.keys()), y = list(sum_of_sq_dist.values()))
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Sum of Square Distances')
plt.title('Elbow Method For Optimal k')
plt.show()
```

Figura 22. Bloque de código para la optimización del número de clústeres.

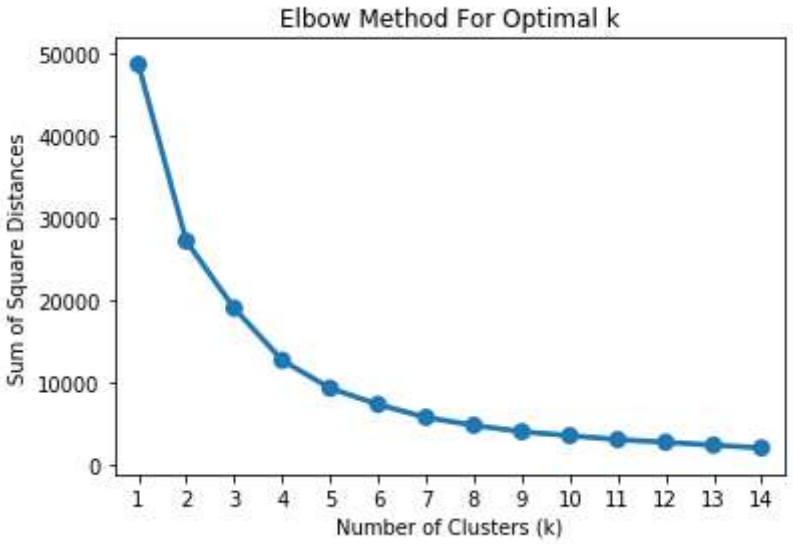

Figura 23. Representación de la inercia respecto al número de clústeres.

Atendiendo al gráfico mostrado en la figura 22, el codo y número óptimo de clústeres correspondería con el cuatro, sin embargo, después de hacer varias pruebas se determinan tres clústeres como el número óptimo más razonable.

4.3.4 Implementación de K-Means

La figura 23 muestra la parte del código en la que se aplica el K-Means, estableciendo el número de clústeres obtenido en el apartado anterior y ajustando el resto de variables para mejorar la eficiencia del algoritmo.

```
# Kmeans algorithm

km = cluster.KMeans(n_clusters=3, max_iter=300, random_state=None)

df_std['cluster'] = km.fit_predict(df_std[num_list])
```

Figura 24. Bloque de código dedicado a la aplicación de K-Means.

4.3.5 Análisis de Componentes Principales (PCA).

Una vez ejecutado K-Means, nos encontramos con un conjunto de tres clústeres creados a partir de la información de las cinco variables numéricas que usamos, lo que nos impide

su representación gráfica. Para solventar este problema recurrimos al Análisis de Componentes Principales.

El Análisis de Componentes Principales o PCA por sus siglas en inglés, permite agrupar un conjunto de variables en otro conjunto de menor dimensión que mantenga la mayor cantidad de información posible, mediante el uso de la matriz de covarianza. Después, se obtienen los autovectores y autovalores, ordenándolos de mayor a menor, resultando en tantas componentes principales como variables existentes. Sin embargo, hay componentes principales más relevantes que otras, obtenidas mediante el cálculo de la varianza explicada por cada una de ellas. En este caso, se seleccionan las dos primeras componentes principales con más varianza explicada, de manera que se pueden representar los resultados de K-Means sin perder mucha información. La figura 24 muestra la parte del código dedicada al Análisis de Componentes Principales y la representación del resultado final.

```
58 # Principal Component Analysis
59
60 pca = PCA(n_components=2, whiten=True)
61
62 pca.fit(df[num_list])
63
64 eigenvectors=pca.components_
65
66 df_std['x'] = pca.fit_transform(df_std[num_list])[:, 0]
67
68 df_std['y'] = pca.fit_transform(df_std[num_list])[:, 1]
69
70 plt.scatter(df_std['x'], df_std['y'], c=df_std['cluster'])
71
72 plt.show()
```

Figura 25. Bloque del código dedicada al Análisis de Componentes Principales.

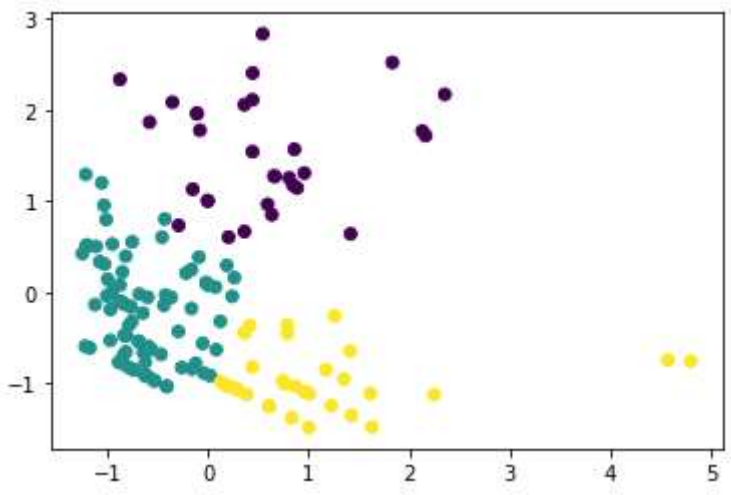

Figura 26. Representación del resultado del K-Means.

Atendiendo a los autovectores, representados en la figura 26, cabe destacar que hay tres variables que influyen más que el resto a la hora de asignar un pedido u otro a un clúster. Estas tres variables son el total (variable 0), el número de productos comprados (variable 4) y el capital descontado del total (variable 1), teniendo el momento del día y del mes una menor influencia (variables 2 y 3).

	0	1	2	3	4
0	0.637384	0.219558	0.147752	0.116988	0.714156
1	-0.300998	0.795697	-0.0618223	-0.507976	0.120018

Figura 27. Autovectores de las Components Principales.

De combinar la información dada por los autovectores y la representación de los clústeres (figura 25), podemos observar:

4) Las transacciones recogidas en el clúster morado son influenciadas en mayor parte por el descuento.
5) Las transacciones recogidas en el clúster amarillo son influenciadas en mayor parte el total pagado y el número de productos.
6) Las transacciones recogidas en el clúster verde no tienen una influencia marcada por ninguna de las variables.

4.3.6 Resultados

Para poder caracterizar los diferentes grupos de comportamientos compradores es necesario extraer los valores medios de cada una de las cinco variables de cada clúster. Con el objetivo de calcular estos valores se ha desarrollado el código mostrado en las figuras 27, 28 y 29.

```
Tot1=0
Tot2=0
Tot3=0
Disc1=0
Disc2=0
Disc3=0
T1=0
T2=0
T3=0
Mf1=0
Mf2=0
Mf3=0
Iq1=0
Iq2=0
Iq3=0
num1=0
num2=0
num3=0
```

Figura 28. Bloque del código para inicializar variables.

```
for i in range (0,148):
    if df_std.cluster[i]==0:
        Tot1=Tot1+df.Total[i]
        Disc1=Disc1+df.Discount_Amount[i]
        T1=T1+df.Timeframe[i]
        Mf1=Mf1+df.Month_frame[i]
        Iq1=Iq1+df.Item_quantity[i]
        num1=num1+1

    elif df_std.cluster[i]==1:
        Tot2=Tot2+df.Total[i]
        Disc2=Disc2+df.Discount_Amount[i]
        T2=T2+df.Timeframe[i]
        Mf2=Mf2+df.Month_frame[i]
        Iq2=Iq2+df.Item_quantity[i]
        num2=num2+1

    else:
        Tot3=Tot3+df.Total[i]
        Disc3=Disc3+df.Discount_Amount[i]
        T3=T3+df.Timeframe[i]
        Mf3=Mf3+df.Month_frame[i]
        Iq3=Iq3+df.Item_quantity[i]
        num3=num3+1
```

Figura 29. Bloque del código dedicado a la suma de las variables dentro de cada clúster.

```
121 # Medias variables CLuster 1
122
123 Tot1mean=Tot1/num1
124 Disc1mean=Disc1/num1
125 T1mean=T1/num1
126 Mf1mean=Mf1/num1
127 Iq1mean=Iq1/num1
128
129 # Medias variables CLuster 2
130
131 Tot2mean=Tot2/num2
132 Disc2mean=Disc2/num2
133 T2mean=T2/num2
134 Mf2mean=Mf2/num2
135 Iq2mean=Iq2/num2
136
137 # Medias variables CLuster 3
138
139 Tot3mean=Tot3/num3
140 Disc3mean=Disc3/num3
141 T3mean=T3/num3
142 Mf3mean=Mf3/num3
143 Iq3mean=Iq3/num3
```

Figura 30. Bloque del código dedicado al cálculo de las medias de las variables.

Una vez ejecutado el programa, los valores obtenidos son los siguientes:

- Clúster morado:
    - Total medio: 25,76
    - Capital descontado medio: 21,88
    - Franja horaria de compra: 3,32
    - Semana del mes de compra: 2,06
    - Número medio de productos comprados: 2,65

- Clúster amarillo:
    - Total medio: 46,31
    - Capital descontado medio: 0,83
    - Franja horaria de compra: 3,48
    - Semana del mes de compra: 3,45
    - Número medio de productos comprados: 2,85

- Clúster verde:
    - Total medio: 21,03
    - Capital descontado medio: 2,03
    - Franja horaria de compra: 3,0
    - Semana del mes de compra: 2,59

- Número medio de productos comprados: 1,21

4.3.7 Análisis y oportunidades

La caracterización de los consumidores ofrece un sinfín de ventajas fácilmente aprovechables por un *ecommerce.* Los principales beneficios directos son:

1. Permite afinar la precisión de las campañas de marketing.

   Al segmentar diferentes grupos de clientes, es posible centrar los esfuerzos de marketing en estos grupos específicos, lanzando mensajes individualizados para cada grupo. De esta manera, permite ajustar el mensaje que quieres transmitir para que concuerde con lo que el posible cliente está buscando, aumentando de esta manera las probabilidades de que el esfuerzo de marketing sea exitoso.

2. Incrementa las ventas.

   Como consecuencia de afinar la precisión de las campañas de marketing, se producirá un incremento en las ventas. De hecho, la segmentación de las campañas, envío de correos electrónicos y actuaciones en redes sociales puede aumentar la tasa de conversión de pedidos hasta en un 60%.

3. Aumenta la tasa de fidelización de clientes.

   La caracterización de consumidores también puede ayudar a generar una mayor fidelización de los clientes. A las personas nos encanta que nos presten atención y nos den un trato personalizado y esto se puede conseguir a través del envío de correos segmentados en función de los grupos de clientes encontrados.

Según lo comentado anteriormente, es obvio que el mayor beneficio directo de la caracterización de clientes es el afinamiento de las campañas de marketing, aumentando la tasa de conversión y, con ello, las ventas y los beneficios. Sin embargo, esta segmentación y clasificación de los intentos de conseguir nuevas ventas puede realizarse antes de la campaña de marketing, principalmente en la fase de diseño y producción. Existen tres estrategias principales:

- Marketing indiferenciado: esta estrategia se caracteriza por buscar la reducción de costes. Esto se consigue proponiendo un único producto que se considera que puede satisfacer las necesidades del conjunto de todos los consumidores. La desventaja de esta estrategia es que se pierde variedad de producto y da una

sensación de menor madurez de marca.

- Marketing diferenciado: esta estrategia es totalmente contraria a la de marketing indiferenciado, caracterizada por los altos costes destinados a producción, estudios, publicidad y distribución, aunque permite ajustarse más a los diferentes grupos de consumidores. La estrategia consiste en el lanzamiento de productos concretos para cada segmento específico logrando una gran penetración en dicho segmento.

- Marketing concentrado: esta estrategia se encuentra en un punto intermedio con respecto a las anteriores. En este caso, la compañía opta por dirigir el producto/servicio solamente a un grupo de consumidores concreto, considerado como el de mayor interés. Cabe destacar que el éxito de esta estrategia radica en el conocimiento profundo del sector hacia el que se dirige la empresa, y su inconveniente la opción de fuertes competidores o incluso de extinción del segmento.

El análisis desarrollado en este trabajo brinda a Independence Brand varias oportunidades de crecimiento, sobre todo a través de estrategias de marketing. Teniendo en cuenta los valores medios de las variables estudiadas para cada clúster se pueden idear varias acciones de marketing que pueden incrementar la tasa de conversión, estas acciones son:

- ***Acciones destinadas a los clientes agrupados bajo el clúster morado.***

  Los clientes agrupados bajo este clúster se caracterizan por ser atraídos por los descuentos, por lo que lo aparentemente más efectivo sería ofrecerles productos con descuentos. Los clientes agrupados bajo este clúster resultan de alto interés, ya que son fácilmente accesibles mediante estrategias de descuentos y otorgan movilidad de inventario. A ellos podrían dirigirse las siguientes acciones:

  1) *Acción basada en el total gastado y el descuento alcanzado*: basándonos en el total medio gastado de 25,76 euros y el descuento medio de 21,88 euros, podemos calcular un total gastado antes de aplicar descuento de 47,64 euros y un descuento aplicado del 46%. La promoción ideal para conseguir compras de estos clientes sería ofrecer un 50% de descuento o un 2x1 por la compra de pedidos superiores a 45 euros.
  2) *Acción basada en descuento y momento de compra*: siguiendo con la idea de que a este grupo de clientes le mueven los descuentos y añadiendo el factor del momento de la compra, se puede alcanzar este grupo de clientes también. Considerando lo analizado en la acción anterior y teniendo en cuenta que este grupo compra en torno a la franja horaria 3, que comprende las horas desde las 12:00 hasta las 18:00, la promoción ideal sería ofrecer un descuento para

las compras realizadas durante esas horas.

3) *Acción basada en descuento y semana del mes de compra*: esta acción es similar a la anterior, pero teniendo en cuenta la semana del mes en vez de las horas dentro del día. En este caso, el análisis indica que estos clientes suelen comprar los días comprendidos entre el 8 y el 15 de cada mes, por lo que la acción consistiría en ofrecer un descuento para las compras realizadas durante esos días.
4) *Acción basada en descuento y número de productos comprados*: como se indicó a la hora de estudiar los autovectores, el número de productos es uno de los factores que más ha influido a la hora de la agrupación. Teniendo en cuenta el número medio de productos comprados de 2,65 por este grupo de clientes, la acción de marketing óptima sería ofrecer descuento por la compra de más de dos productos.

Estas acciones arriba expuestas también pueden combinarse entre sí, siempre y cuando se mantenga el incentivo del descuento, pudiendo incluso lanzar una promoción que las incluya todas. La promoción que intentara ser lo más precisa posible consideraría ofrecer un descuento por compras realizadas entre las 12:00 y las 18:00 de los días del 8 al 15 de cada mes, por un importe superior a 45 euros y más de 2 productos.

- ***Acciones destinadas a los clientes agrupados bajo el clúster amarillo***

Los clientes agrupados bajo este clúster se caracterizan por ser los que más dinero se gastan en cada pedido y no acogerse a descuentos. Este grupo es de especial interés, ya que son los que más ingresos generan a Independence Brand, sin embargo, son difícilmente accesibles con las variables estudiadas. A ellos podrían dirigirse la siguiente acción:

1) *Acción basada en el momento de lanzamiento de nuevos productos*: según las variables estudiadas, este grupo de clientes compra independientemente de los descuentos ofrecidos. No se conoce con exactitud lo que motiva a estos clientes a comprar, pero dado que son los que más dinero gastan y no usan descuentos, deben de representar los clientes más fieles a la marca que compran sin importar las condiciones. Partiendo de que son nuestros clientes más fieles, se entiende que son los que compran los nuevos productos Independence en el momento de su lanzamiento. Dicho esto, para llegar a ellos y aprovechar su potencia de compra, lo ideal sería hacer los lanzamientos de los productos en la franja horaria 3,48 y semana del mes 3,45, correspondiendo aproximadamente a las últimas seis horas del día y última semana del mes.

Lamentablemente, por ahora no se dispone de información suficiente para poder atacar con precisión el grupo de clientes que genera más beneficios. Esta área se deja para futuras líneas de investigación.

- ***Acciones destinadas a los clientes agrupados bajo el clúster verde**.*

  Bajo este clúster se agrupan los clientes que podrían considerarse menos interesantes, gastan poco, compran pocos productos y, aparentemente, no se guían por incentivos como los descuentos. Dado que el momento del día y del mes en que compran son similares a los del clúster morado, las acciones empleadas para atraer a dichos clientes también podrían atraer a los agrupados bajo este clúster.

4.4 Predicción de la demanda

En cuanto a la predicción de la demanda de productos de Independence Brand, se ha recurrido principalmente al estudio de la estacionalidad de las ventas y las visitas al portal en internet de la marca. El fin de este estudio es el de anticiparse a estas ventas y organizar la producción y gestión del inventario, consiguiendo también una idea de cuando compran nuestros clientes.

Las figuras 30 y 31 mostradas a continuación representan el historial de ventas y visitas al portal de Independence Brand, respectivamente. En ambas es posible advertir cierta estacionalidad, localizando picos locales en los meses de noviembre-diciembre y mayo-junio, con unos pequeños repuntes los meses de septiembre y marzo. Dichos picos coinciden, principalmente, con el Black Friday, la campaña de Navidad y el inicio del verano. Esta información nos es muy útil a la hora de saber cuándo lanzar un nuevo producto o una nueva acción potente en marketing.

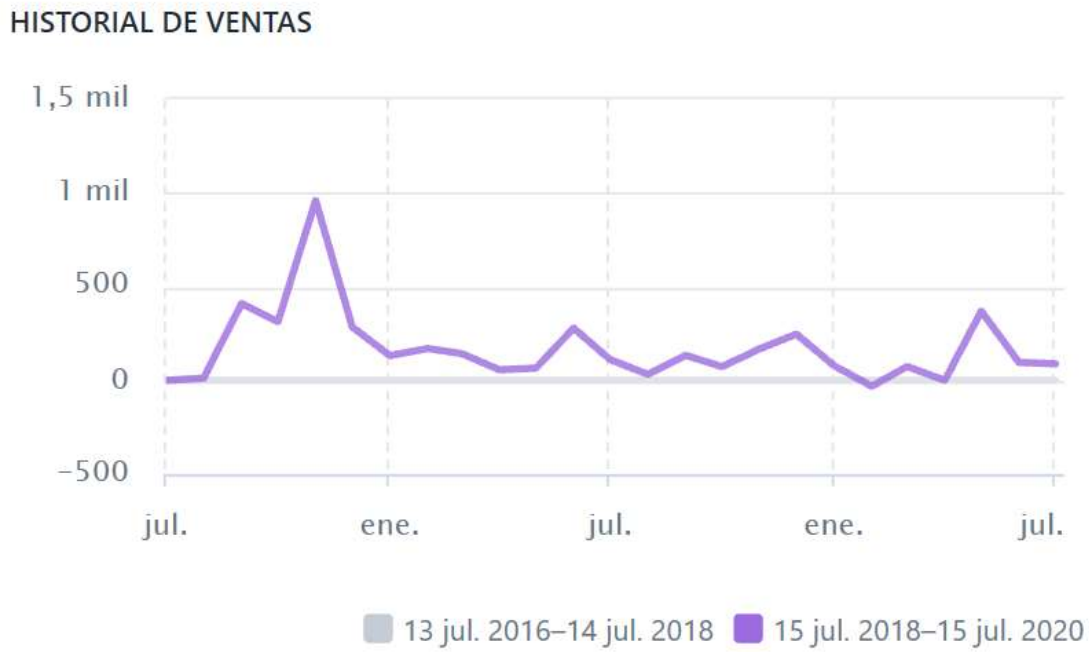

Figura 31. Histórico de las ventas de Independence Brand.

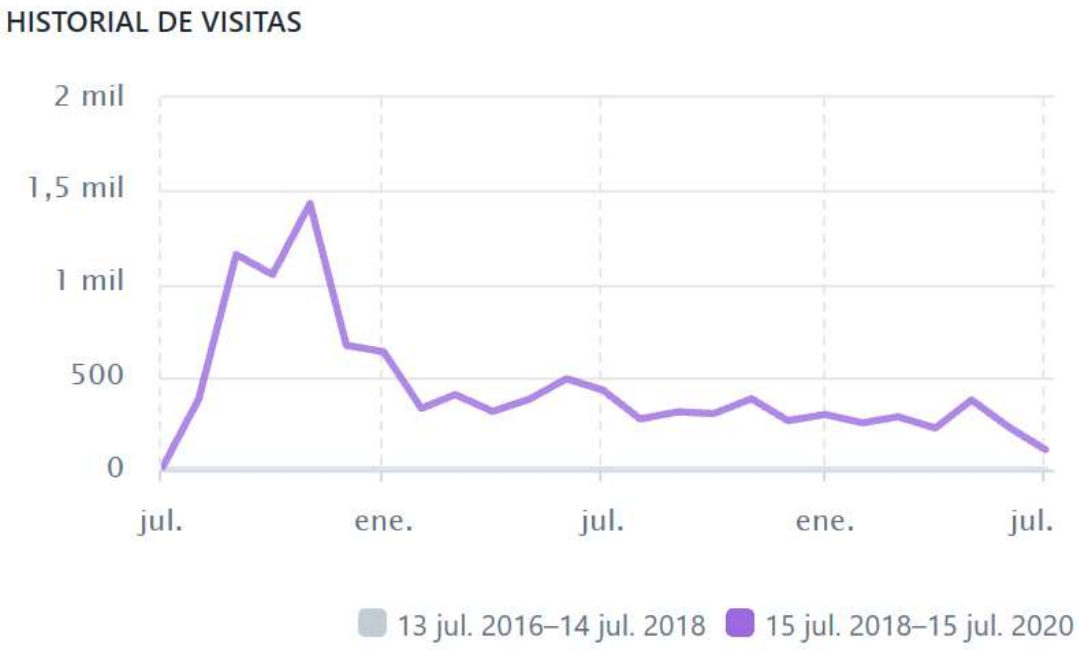

Figura 32. Histórico de las visitas a la web de Independence Brand.

Por último, una observación interesante que resulta al comparar ambos gráficos es la correlación entre ventas y visitas. Aunque resulta obvio que existe una correlación positiva, mirando el eje vertical se puede observar que las escalas son prácticamente iguales, con una relación aproximada de 0,6 euros resultantes en ventas por cada

visita a la página web. Esta información ofrece una gran oportunidad de escalabilidad, ya que el incremento en ventas vendría de la mano de un incremento en visitas con una procedencia similar a la de los dos años anteriores. En este marco, la inversión en marketing digital resultaría altamente efectiva.

Capítulo 5. Futuras líneas de trabajo

Inicialmente el alcance de este trabajo se planteó de tal manera que se pudiera cubrir un estudio detallado de la predicción de la demanda y de la caracterización de los consumidores. Aunque el estudio de la estacionalidad y el K-Means aplicado a las variables numéricas pudiera resultar suficiente para empezar a probar acciones de marketing basadas en los resultados, el objetivo es conseguir unas directrices lo más precisas posibles.

Para ello, como futuras líneas de trabajo, incorporaríamos las variables categóricas descartadas durante el preprocesamiento de los datos ("Financial Status", "Accepts Marketing", "Discount Code", "Shipping Method"y "Shipping Province") para obtener un mayor conocimiento de nuestros clientes, por ejemplo, conociendo la provincia desde la que compran para afinar aún más las acciones de marketing. Con esta información sería más fácil alcanzar a los consumidores que pertenecen al clúster amarillo, los cuales son los de más interés, y también a los pertenecientes al clúster verde que, aunque no son tan interesantes, representan el 50% de los consumidores de Independence Brand.

En cuanto al estudio de la predicción de la demanda, un análisis empleando redes neuronales (RNN) podría arrojar más luz y permitir una mayor precisión a la hora de estructurar presupuestos, planificar la producción, organizar el inventario y definir las políticas de precios.

Además, resulta muy conveniente repetir estos análisis cuando se hayan recopilado muchos más datos, ya que, aunque los resultados parecen razonables, podrían no ser preciosos debido a la poca cantidad de información empleada.

Bibliografía

[1] Hyndman, R. et al., 2008. Forecasting with Exponential Smoothing: The State Space Approach, Springer Science & Business Media.

[2] Box, G.E.P. et al., 2015. Time Series Analysis: Forecasting and Control, John Wiley & Sons.

[3] Box, G.E.P. & Cox, D.R., 1964. An Analysis of Transformations. Journal of the Royal Statistical Society. Series B, Statistical methodology, 26(2), pp.211–252.

[4] Yeo, J. et al., 2016. Browsing2purchase: Online Customer Model for Sales Forecasting in an E-Commerce Site. In Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee, pp. 133–134.

[5] Ramanathan, U., 2013. Supply chain collaboration for improved forecast accuracy of promotional sales. International Journal of Operations & Production Management.

[6] Seeger, M.W., Salinas, D. & Flunkert, V., 2016. Bayesian Intermittent Demand Forecasting for Large Inventories. In D. D. Lee et al., eds. Advances in Neural Information Processing Systems 29. Curran Associates, Inc., pp. 4646–4654.

[7] Snyder, R., Ord, J.K. & Beaumont, A., 2012. Forecasting the intermittent demand for slow-moving inventories: A modelling approach. International journal of forecasting, 28(2), pp.485–496.

[8] Zhang, G., Patuwo, B.E. & Hu, M.Y., 1998. Forecasting with artificial neural networks: The state of the art. International journal of forecasting, 14(1), pp.35–62.

[9] Yan, W., 2012. Toward automatic time-series forecasting using neural networks. IEEE transactions on neural networks and learning systems, 23(7), pp.1028–1039.

[10] Zimmermann, H.-G., Tietz, C. & Grothmann, R., 2012. Forecasting with Recurrent Neural Networks: 12 Tricks. In Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 687–707.

[11] Trapero, J.R., Kourentzes, N. & Fildes, R., 2015. On the identification of sales forecasting models in the presence of promotions. The Journal of the Operational Research Society, 66(2), pp.299–307.

[12] Borovykh, A., Bohte, S. & Oosterlee, C.W., 2017. Conditional Time Series Forecasting with Convolutional Neural Networks. arXiv [stat.ML].

[13] Flunkert, V., Salinas, D. & Gasthaus, J., 2017. DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks. arXiv [cs.AI].

[14] Wen, R. et al., 2017. A Multi-Horizon Quantile Recurrent Forecaster. arXiv [stat.ML].

[15] Chapados, N., 2014. Effective Bayesian Modeling of Groups of Related Count Time Series. In E. P. Xing & T. Jebara, eds. Proceedings of the 31st International Conference on Machine Learning. Proceedings of Machine Learning Research. Bejing, China: PMLR, pp. 1395–1403.

[16] Bandara, K., Bergmeir, C.& Smyl, S., 2017. Forecasting Across Time Series Databases Using Recurrent Neural Networks on Groups of Similar Series: A Clustering

[17] Tajunisha and Saravanan, "Performance Analysis of k-means with different initialization methods for high dimensional data" International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.4, October 2010

[18] Neha Aggarwal and Kriti Aggarwal,"A Mid- point based k –mean Clustering Algorithm for Data Mining". International Journal on Computer Science and Engineering (IJCSE) 2012.

[19] Barileé Barisi Baridam," More work on k-means Clustering algortithm: The Dimensionality Problem ". International Journal of Computer Applications (0975 – 8887)Volume 44– No.2, April 2012

[20] Pawan Kumar Singh, Yadunath Gupta, Nilpa Jha and Aruna Rajan. 2019. "Fashion Retail Forecasting Demand for New Items".

[21] Yu Y., Choi T., Hui C., (2011). "An Intelligent Fast Sales Forecasting Model for Fashion Products."

[22] Au K.F., Choi T.M., Yu Y., (2008). "Fashion retail Forecasting by evolutionary neural networks". International Journal of Production Economics, 114(2), pp. 615–630.

[23] "What is Demand Forecasting? Importance and Benefits of Forecasting Customer Demand" – Kristina Lopienski. Último acceso: 20/07/2020 https://www.shipbob.com/blog/demand-forecasting/

[24]"How to Create an ARIMA Model for Time Series Forecasting in Python" – Jason Brownlee. Último acceso: 20/07/2020 https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

[25] "Recurrent Neural Networks Tutorial, Part 1 – Introduction to RNNs" – Denny Britz. Último acceso: 20/07/2020 http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

[26] "Long Short Term Memory" . Último acceso: 20/07/2020 https://en.wikipedia.org/wiki/Long_short-term_memory

[27] "A Multivariate Time Series Guide to Forecasting and Modeling (with Python codes)" – Aishwarya Sing. Último acceso: 20/07/2020 https://www.analyticsvidhya.com/blog/2018/09/multivariate-time-series-guide-forecasting-modeling-python-codes/

[28] "Supervised vs. Unsupervised Learning" – Devin Soni. Último acceso: 20/07/2020 https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d

[29]"Análisis de grupos" . Último acceso: 20/07/2020 https://es.wikipedia.org/wiki/Análisis_de_grupos

[30] "K-Means" . Último acceso: 20/07/2020 https://es.wikipedia.org/wiki/K-medias

[31] Web de Anaconda. Último acceso: 20/07/2020. https://www.anaconda.com/

[32] Web de Shopify. Último acceso: 20/07/2020. https://www.shopify.com/

[33] Web Independence Brand. Último acceso: 20/07/2020 https://www.independencebrand.com/

Anexo A: Alineación del proyecto con los ODS y la Agenda 2030

Atendiendo a la información que aparece en la página web de la Organización de las Naciones Unidas, los Objetivos de Desarrollo Sostenible fueron adoptados por todos los Estados Miembros en 2015 como un llamado universal para poner fin a la pobreza, proteger el planeta y garantizar que todas las personas gocen de paz y prosperidad para 2030. Existen 17 objetivos diferentes, los cuales pueden clasificarse bajo tres tipos: los objetivos que persiguen mejoras económicas, los que persiguen mejoras sociales y los que persiguen mejoras en el medio ambiente.

En lo que se refiere a este trabajo, se ha detectado una fuerte alineación con los Objetivos de Desarrollo Sostenible 9, 13, 14 y 15. Estos son:

- Industria, Innovación e Infraestructura (ODS 9)
- Acción por el clima (ODS 13)
- Vida submarina (ODS 14)
- Vida de ecosistemas terrestres (ODS 15)

La motivación de este proyecto de estar preparados para la cuarta revolución industrial coincide con el noveno Objetivo de Desarrollo Sostenible, al promover la innovación dentro de la industria.

En cuanto a los Objetivos de Desarrollo Sostenibles decimotercero, decimocuarto y decimoquinto, están estrechamente relacionados con la gran contaminación causada por la industria textil, siendo la segunda mayor causa de la contaminación de las aguas. La principal causa de contaminación en las aguas es el poliéster. El poliéster es una de las fibras más populares utilizadas en la moda hoy en día, se encuentra en aproximadamente el 60% de las prendas en las tiendas minoristas, es decir, aproximadamente 21.3 millones de toneladas de poliéster. La popularidad del poliéster también sigue aumentando, ya que hubo un aumento del 157 por ciento en el consumo de ropa de poliéster entre 2000 y 2015. Éste se fabrica a partir de combustibles fósiles y no es biodegradable, por lo que produce una gran cantidad de emisiones de dióxido de carbono a la atmósfera y la presencia de microplásticos en los mares. Este dióxido de carbono es el principal agente del cambio climático y los microplásticos producen daños en la vida marina al ser ingeridos.

Figura 33. Microplásticos.

Sin embargo, no solo las fibras sintéticas contaminan, las fibras naturales también contribuyen a la contaminación a través de la contaminación agrícola siendo El algodón uno de los tejidos más comúnmente usados y más destructivos para el medio ambiente, requiere una gran cantidad de pesticidas y agua. La producción de este algodón requiere de una gran cantidad de pesticidas, que acaban vertiéndose a las aguas y dañando la vida marina. En la figura de abajo se muestra cómo los pesticidas vertidos al Mar Menor por la actividades agriculturales acabaron con la vida de miles de peces.

Figura 34. Peces muertos en el Mar Menor.

Atendiendo a los daños en la vida terrestre, la principal causa del daño por parte de la industria textil es la gran sobreproducción existente hoy en día. Según la EPA (Environmental Protection Agency), solo en 2013 se produjeron 15,1 millones de toneladas de residuos de ropa textil. Esta sobreproducción acarrea un exceso de emisiones de dióxido de carbono y los residuos de esta sobreproducción se acaban amontonando en vertederos, produciendo filtraciones a la tierra y dañando la vida terrestre.

Gracias al análisis realizado en este trabajo, las empresas textiles son capaces de producir la cantidad de productos que más se ajusta a la demanda, diseñar estos productos de la manera que más se ajusta a los gustos de los consumidores y organizar el inventario eficientemente, todo ello con el objetivo de evitar la sobreproducción y, con ello, contribuir a los Objetivos de Desarrollo Sostenible 9, 13, 14 y 15.

Referencias:

- "How polluting is the fashion industry?" – Cameron Boggon. Último acceso 23/07/2020. https://www.ekoenergy.org/how-polluting-is-the-fashion-industry/
- ONU website. Último acceso 23/07/2020. https://www.undp.org/content/undp/es/home/sustainable-development-goals.html
- Pollution in the fashion Industry. Último acceso 23/07/2020. https://en.wikipedia.org/wiki/Pollution_In_The_Fashion_Industry
- "What happens when fashion becomes Fast, Disposable and Cheap?" – Zhai Yun Tan. Último acceso 23/07/2020. https://www.npr.org/2016/04/08/473513620/what-happens-when-fashion-becomes-fast-disposable-and-cheap?t=1595588853332