Trabajo final Machine Learning III

Álvaro Villadangos del Río
Carlos de la Torre Pertierra
Elena Rodríguez Juliani
Andrea Secades Sierra

Introducción

Para realizar este trabajo, hemos escogido un dataset que muestra las transacciones de un número de clientes en una tienda de CDs llamada CDNOW. Esta tienda ha sufrido un severo decrecimiento en su volumen de negocios, como se mostrará más adelante. El problema con el que nos encontramos es que queremos saber qué tipo de estrategia de marketing debemos adoptar respecto a cada cliente que hemos tenido para maximizar el número de clientes que vuelven a efectuar compras en nuestra tienda, además del valor de estas. Para ello, hemos decidido agrupar a todos los clientes en base a varios aspectos relativos a sus compras en nuestro local para posteriormente dirigir una estrategia diferente hacia cada grupo en función de sus características concretas. Por ejemplo, si nos encontramos con un grupo de clientes que compran con cierta frecuencia en nuestra tienda, pero lo hacen por cantidades bajas, deberíamos intentar anunciarles productos más caros, pero también a retenerles ya que es importante mantener esa fidelidad que demuestran. Por otro lado, a los clientes que han dejado de venir deberíamos mostrarles nuestras novedades y nuestros mejores productos para intentar atraerlos de nuevo.

Por tanto, y en resumen, el objetivo del trabajo es dividir a nuestros clientes en grupos según ciertas variables relativas a sus compras en el local, para así saber qué tipo de publicidad le debemos hacer a cada cliente o futuro cliente. Hemos entendido que para realizar esta clasificación sería muy útil realizar una partición de las observaciones mediante el uso de clusters. Lo bueno que tiene este modelo es que, una vez creado, cualquier cliente nuevo que tengamos podrá ser también clasificado en uno de los grupos en función de las compras que realice.

Estrategia de datos

- Fuente de los datos

  https://data.world/mktg-776-ta/fader-and-hardie-datasets

El primer dataset que utilizamos no era válido para el tipo de trabajo que teníamos que hacer, por ello tuvimos que buscar otro distinto. Decidimos utilizar este que refleja las compras en una tienda de CDs llamada CDNOW:

> Fader, Peter & Hardie, Bruce. (2000). Forecasting repeat sales at CDNOW: A case study. Interfaces. 31. 10.1287/inte.31.4.94.9683.

El hecho de que hubiese un estudio hecho sobre los datos, nos permitió profundizar más fácilmente en la idea que teníamos ya que contábamos con una explicación técnica del dataset en sí.

- Descripción de los datos (nº registros, variables, significado y tipo de las variables)

Nuestro dataset cuenta con **23.570 registros**, cada uno representa el estudio de las transacciones realizadas por un mismo cliente en la tienda de CDs. No incluye missing values o ceros. A su vez cuenta con cuatro variables [0,1,2,3] .
Son todas variables numéricas, donde:

- 0 es el ID del cliente que realizó la compra, ID.
- 1 es la fecha en la que se realizó la compra, Fecha.
- 2 es el número de artículos que se compraron, Artículos.
- 3 es el valor de la compra, Valor.

- Tratamiento de datos → creación o transformación de las variables originales

Cambiamos los nombres de las variables dos veces, una antes de tratar los datos en sí y posteriormente una segunda vez para que fuera más sencillo comprenderlas y fuese evidente qué tipo de información recogían.

```
datos = pd.DataFrame({
    'ID': data[0],
    'Fecha': data[1],
    'Valor': data[3],
    'Articulos': data[2]
})

df = pd.DataFrame({
    'ID':lista_ID,
    'Value':valor,
    'Quantity':cantidad,
    'Recency':ultima_compra
})
```

Después hemos pasado al tratamiento de los datos, en sentido más estricto. Para la columna VALUE, lo que hemos hecho ha sido hallar la media de todo el gasto que había hecho ese cliente concreto (ID) en la tienda.

```
valor=[]
for i in lista_ID:
    x = datos[datos['ID']==i]
    y = x['Valor'].mean()
    valor.append(y)
len(valor)
```

En el caso de QUANTITY, para cada cliente o ID, se suma el número de compras que ha hecho en total en todo su tiempo como cliente de la tienda.

```
cantidad = []
for i in lista_ID:
    x = datos[datos['ID']==i]
    cantidad.append(sum(x['Articulos']))
len(cantidad)
```

Para RECENCY hemos calculado a partir de la fecha que aparecía en el dataset, el número de días que habían pasado desde esa fecha de última compra.

```
ultima_compra=[]
for i in lista_ID:
    x = datos[datos['ID']==i]
    ultima_compra.append(max(x['Fecha']))
len(ultima_compra)
```

Finalmente las variables serían estas:

ID: número identificador del registro o transacciones de un cliente en particular

Value: media de dinero gastado por cliente (ID).

Recency: mide la frecuencia con la que dicho cliente realiza transacciones con la empresa objeto de estudio en función de la última fecha de compra. Te indica el número de días que han pasado desde la última compra.

Quantity: cantidad de transacciones realizadas por el cliente en total. Número de artículos que el cliente ha comprado a lo largo del tiempo.

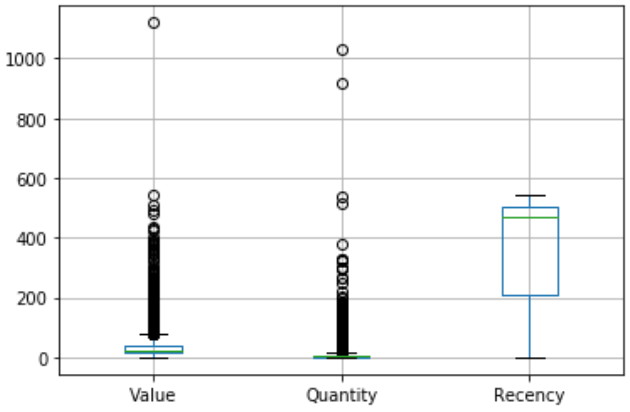
Tabla antes del procesado de datos

ID	Fecha	Valor	Artículos
1	19970101	11.77	1
2	19970112	12.00	1
2	19970112	77.00	5
3	19970102	20.76	2
3	19970330	20.76	2

Tabla después del procesado de datos

ID	Value	Quantity	Recency
1	11.770000	1	545.0
2	44.50000	6	534.0
3	26.076667	16	33.0
4	25.125000	7	200.0
5	35.055455	29	178.0

Respecto a los registros del dataset, hemos tratado de mantener el máximo de información posible. Para ello, hemos hecho un boxplot de las variables del dataset para ver cómo se distribuían. Para nuestra sorpresa nos encontramos con muchos outliers tanto en Value como en Quantity. Entonces hemos procedido a apartarlos en un dataset 'outliers' para no perder los clientes y realizar el clustering sin ellos.



- ¿Qué variables han sido incluidas en el clustering?¿Se excluye alguna variable para luego poder hacer interpretaciones de los clusters?

Tras el tratamiento de los datos, apareció una variable desconocida que decidimos eliminar, como es lógico, porque no aportaba ningún tipo de información.

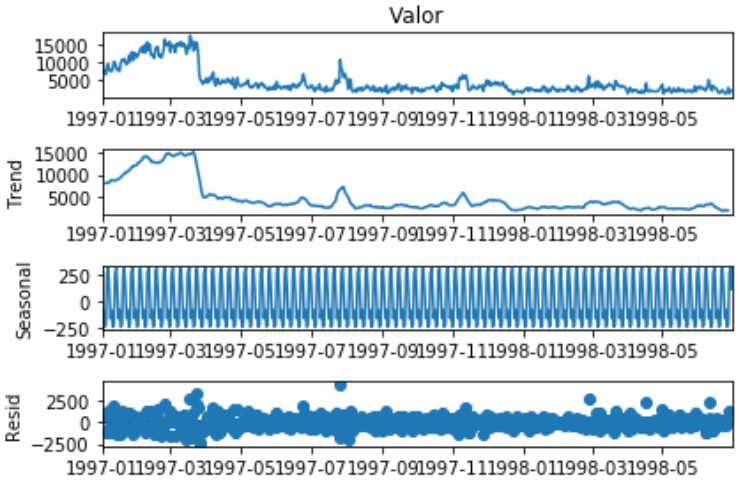
Para llevar a cabo el clustering excluimos a la variable ID, puesto que sólo identifica al cliente y posteriormente la volvimos a incluir para poder realmente asociar a cada cliente a un cluster ya que va a ser esto lo que nos permita asignarle un perfil de marketing especializado.

Hemos incluido todos los registros del dataset excepto uno, como hemos dicho antes. El gran volumen de datos que contenía nos supuso un problema inicialmente. Comenzamos tratando

de ejecutar el código en R pero fue imposible hacerlo con la totalidad de los datos. Sólo podíamos hacerlo con el 20-30% de los datos, por ello decidimos intentar hacerlo en Python y así finalmente funcionó.

Análisis

Antes de comenzar con el análisis exploratorio, visualizamos los datos. Al tratarse de una serie temporal utilizamos la función time_series() y seasonal_decompose() en Python o ts() decompose() en R para indicarselo a dichos programas y obtener cuatro gráficos separados para poder analizar la gráfica original, la tendencia, la estacionalidad y el ruido blanco o residuos. Cabe destacar respecto de la tendencia que, aunque observamos un incremento importante de las ventas al principio de la serie, en aproximadamente abril hay una caída brusca y nunca vuelve a alcanzarse el volumen de ventas inicial. Debemos recordar aquí que, debido a la gran cantidad de residuos y a que la distancia euclídea se ve muy afectada por los mismos (eleva al cuadrado las distancias), se han eliminado todos los outliers.



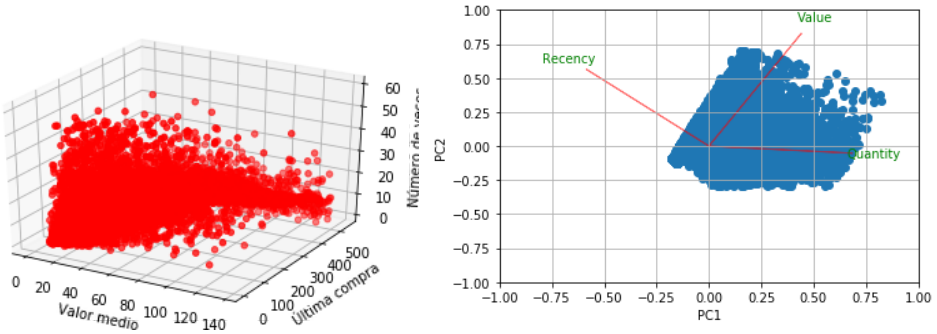
Análisis Exploratorio

El PCA (o **Análisis de las Componentes Principales**) es una técnica de extracción de variables. Consiste en la transformación de un conjunto de variables originales en un nuevo conjunto de variables más pequeño, las componentes principales, esto son combinaciones lineales de las variables de partida u originales. Estas componentes principales están incorreladas entre sí, de manera que cuanta más información aporten, mayor será su varianza. El objetivo es pasar de p dimensiones a k dimensiones ($k \leq p$), pero manteniendo la mayor cantidad de información posible.

Siguiendo con el análisis, el objetivo será mantener la mayor cantidad de información posible, lo cual se consigue a través de una mayor varianza. ¿Cómo sabremos qué vector aporta una mayor cantidad de información? Utilizamos la **Proporción de Varianza Explicada** (PVE) y la acumulada. Con la primera obtendremos el porcentaje de información que contiene cada componente; mientras que la segunda indicará la proporción de

información respecto del total, es decir, cuanta información se ha acumulado en total hasta dicha variable. En este caso concreto, observamos que con las dos primeras componentes principales captamos el 90% de la información total. No obstante, cada resaltar que únicamente con la primera ya captamos más de la mitad de la información (60%). Aunque en principio podremos reducir a dos dimensiones, realmente consideramos que para obtener mejores resultados es conveniente trabajar con las tres dimensiones iniciales. Basamos esto en que, a pesar de usar la totalidad de los datos, el programa es eficaz y rápido.

De manera que, a través de un **Scatter Plot o Gráfico de Dispersión** (figura de la izquierda) se realizará un estudio de la relación entre variables de un mismo conjunto, en este caso concreto entre tres variables: "Valor Medio", "Última Compra" y "Número de Veces". Pudiendo completar esta información con un **Biplot** (figura de la izquierda).



La interpretación de estos gráficos permitirá no solo encontrar una posible relación entre las variables (i.e. lineal positiva, cuadrática) y evaluar la fuerza de dicha relación, sino que además facilitará la posibilidad de identificar los patrones dentro de cada uno de los grupos (posibles clusters). A su vez, y como estudiaremos más adelante, con este tipo de visualización podremos distinguir los valores extremos o outliers.
Si bien es cierto que más adelante se realizará un análisis más profundo del número de posibles clusters en los que se podrán agrupar a los individuos de este conjunto de datos, es necesario entender que con este tipo de gráficos se visualizará un patrón a través de la agrupación de los puntos alrededor de una línea. En función de la distancia entre los puntos y dicha línea, la asociación entre las variables positiva o negativa, existirá dependencia entre ambas o serán independientes. No obstante, y en relación con los ya mencionados outliers, esta relación no es exclusivamente de causalidad, sino que podía, en ocasiones, darse a circunstancias externas.

Por último, los datos del conjunto representan variables totalmente distintas (por ejemplo, el valor de la compra y el número de artículos comprados), por lo tanto, para poder trabajar con ellos deberemos estandarizar los valores. ¿Qué significa esto? Desde un punto de vista teórico, esto implica la transformación de la media en 0 y la varianza en 1; en la práctica se traduce en asegurarnos de que todas las variables tienen el mismo peso. ¿Por qué introducimos la escalada de variables? Porque para poder calcular las distancias entre los

clusters y realizar un análisis preciso, necesitaremos que todas las variable tengan una misma influencia.
Relacionado con lo anteriormente expuesto acerca la distancia entre las variables, hemos calculado la matriz de distancias o pesos, en la cual se representan las distancias entre los puntos del conjunto, elegidos por pares. En su cálculo hemos utilizado la distancia euclídea.

```
Z = linkage(df_norm,'ward', metric='euclidean')
```

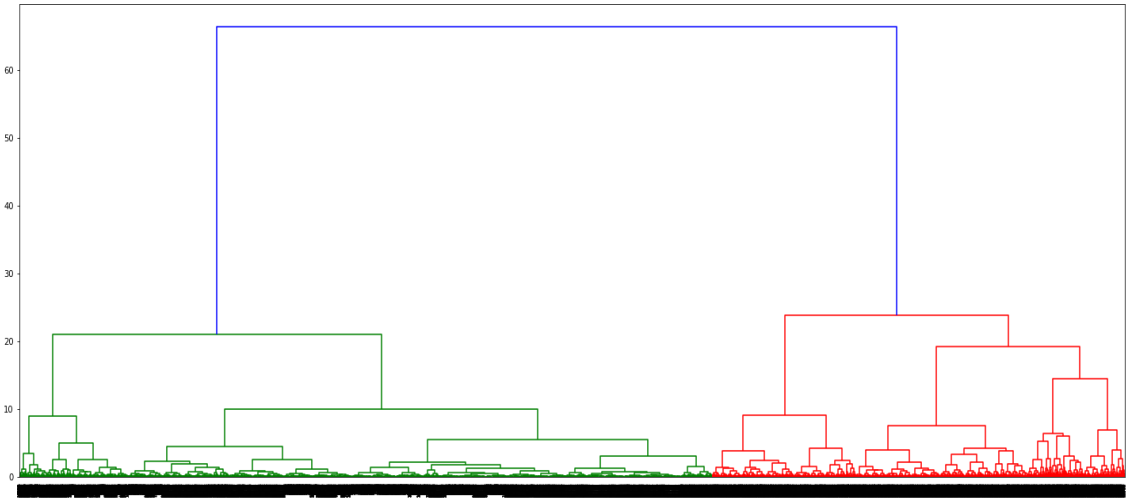
De manera que, con el estudio realizado anteriormente obtenemos la información necesaria para poder proceder al análisis de clusters.

Análisis de Cluster

Dentro del análisis de clusters, podemos utilizar, a grandes rasgos, dos técnicas: (i) partitional clustering, en el cual se especificará previamente el número de clusters (i.e. k-means); y (ii) hierarchical clustering, donde se comienzan con n clusters y se producirán uniones entre los más similares hasta que solo quede uno. No obstante, nosotros nos hemos centrado en esta segunda, en particular en el clustering aglomerativo. Este tipo de clustering jerárquico implica que el dendograma (que veremos más adelante) va a construirse de abajo hacia arriba, es decir, las observaciones irán uniéndose hasta formar un solo cluster.

Una vez establecido la técnica de clustering a utilizar, *agglomerative hierarchical clustering*, a través de la función dendrogram() en Python y hclust() en R, podremos visualizar una agrupación de variables y la distancia entre estos grupos. Obtendremos información bastante detallada de estos clusters que hemos visualizado utilizando

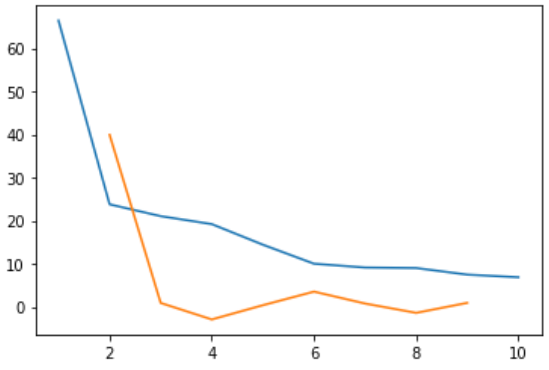
```
plt.figure(figsize=(25,10))
dendrogram(Z, leaf_rotation=90)
plt.show()
```



Con la finalidad de medir cómo de precisa es la solución de agrupación entre los clusters, en otras palabras, para calcular la correlación entre las distancias obtenidas a través del árbol y las distancias reales, hemos utilizado el **Índice de Cophenet**. El resultado de este índice será mejor (calificado de 'de alta calidad') cuanto más cercano sea de uno. En este caso concreto obtenemos un resultado de 0.77, de manera que el árbol de decisión representa las diferencias entre las observaciones de forma muy verídica y precisa.

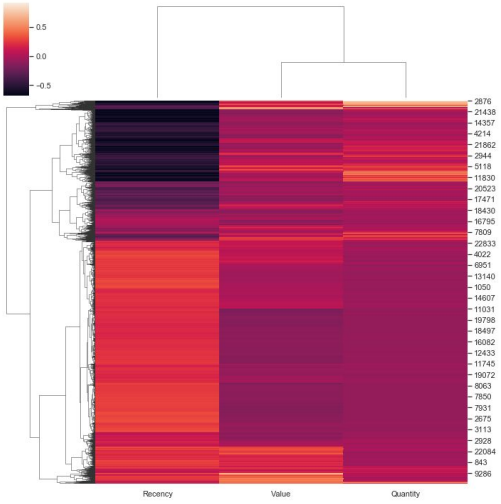
Realmente este gráfico en sí no aporta ninguna información ni separa en clusters, sino que para ello debemos lugar a la distancia entre los grupos (en el eje de la izquierda o height). De manera que para poder crear grupos o familias podemos o bien "cortar el árbol" escogiendo una altura máxima (h=15) o indicar el número de clusters que queremos crear (k=6, en nuestro caso). Realmente podría no parecer del todo apropiado separar en dos la última rama, no obstante, en el siguiente párrafo se realizará una explicación más amplia acerca este tema al analizar la regla del codo.

Además, con el Scree Plot o Gráfico de Sedimentación y siguiendo la **Regla del Codo** se respalda la conclusión. ¿Por qué decimos que respalda esta conclusión? ¿En qué consiste la regla del codo? Debemos primero hacer un pequeño inciso para explicar que con esta regla se suma la distancia entre todos los puntos de un cada cluster con su centroide. Básicamente con su representación indica que llega un punto en el que el aumento de centroides o clusters (eje de abscisas) y la disminución del valor WCSS (eje de coordenadas) formarán un codo. Para encontrar el número óptimo de clusters o valor óptimo de k, escogeremos el número de centroides en el que se produzca dicho codo. No obstante, en este caso concreto podemos diferenciar dos codos en k=2, k=4 y k=6. Aunque el cambio brusco se produce en k=2, nos quedaremos con k=6 porque es cuando dejan de producir variaciones importantes y se produce la segunda máxima aceleración.
Debemos añadir que, esta conclusión no solo se corresponde con lo observado en el dendograma, sino que además si utilizamos un número de clusters inferior correriamos el riesgo de generalizar demasiado.

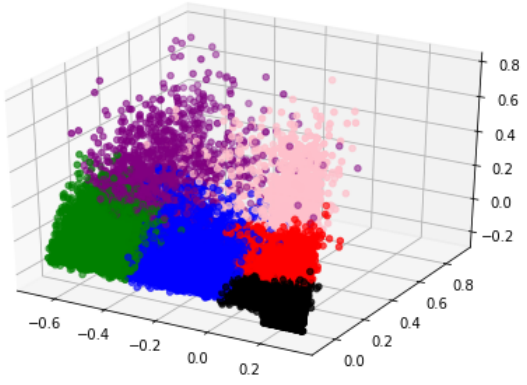


Para completar la información obtenida a través del dendograma, utilizaremos un **HeatMap** o Mapa de Calor, el cual representa una matriz de valores. Al tratarse de una gráfica en la que

se incluye un dendograma,el heatmap ordena por semejanzas las filas o columnas de la matriz indicando a su vez un código de colores dependiendo del valor de cada variable en cada posición.
Observamos que diferencia entre las tres dimensiones escogidas (“Recency”, “Value”, y “Quantity”). Procedemos ahora a la interpretación del gráfico. Cada cuadrado muestra la correlación de cada variable con el eje, que irá de -1 (ambas variables disminuyen) a 1 (ambas variables incrementan). Cuanto más cercana sea la correlación a 0, menor relación habrá entre ambas. La característica principal de los heatmaps es que, en vez de utilizar números para indicar la correlación entre las variables, crea una escala de colores cuanto más oscuro más cercano de -1 y cuanto más clarito más cercano de 1. Utilizando estos colores, los heatmaps hacen mucho más fácil el análisis de los datos y la clasificación en clusters de las variables.



Finalmente, utilizando las conclusiones alcanzadas anteriormente, hemos representado una vez más todo el conjunto de datos a través de un Scatter Plot, no obstante, añadiendo colores dependiendo del cluster al que pertenece cada uno de los clientes. En ‘*Resultados*’ clasificamos los clusters según los segmentos estudiados al que pertenece y las actuaciones que deberá tomar la tienda.




```
ax=Axes3D(fig)
color = []
for i in df_norm['clust_k']:
    if i==0:
        color.append('red',)
    elif i==1:
        color.append('green')
    elif i==2:
        color.append('blue')
    elif i==3:
        color.append('pink')
    elif i==4:
```

```
        color.append('purple')
    elif i==5:
        color.append('black')

ax.scatter(df_norm['Recency'],df_norm['Quantity'],df_norm['Value'],color=color)
plt.show()
plt.savefig("Scatter 3D K-means.jpg", bbox_inches='tight')
```

Resultados

La finalidad real de realizar este análisis no es únicamente saber el número de clusters del conjunto de datos, sino que calificar a los individuos que conforman dicho subconjunto e identificar quienes son los clientes que queremos mantener.
En primer lugar, obtenemos una tabla con la media de cada uno de los clusters de cada una de las variables utilizando la función

```
df.groupby('clust_k').mean()
```

Clust_k	Value	Quantity	Recency (en días)
0	45.150784	3.956106	490.910805
1	27.456807	8.998948	75.090431
2	28.820578	6.818303	271.689185
3	89.944946	8.715622	468.132834
4	55.428006	30.724607	82.026466
5	17.560603	1.596288	494.444909

Los **‘champions’** serán los que hayan comprado más recientemente, compran frecuentemente y gastan más dinero. Esta descripción se identifica con el Cluster 4, pues aunque el cluster 1 tenga una media de días menor, la media de las otras dos variables son bastante más pequeñas. A este grupo debemos recompensarlo y favorecerlo, pues podrán ser los compradores de productos nuevos. Además serán quienes promuevan y recomienden la tienda.

Los **‘promising’** son los clientes como los del Cluster 1, aquellos que han comprado hace poco, pero que no han gastado mucho dinero. Las campañas que mejor funcionarán con ellos serán las ofertas de pruebas gratuitas y actuaciones relacionadas con la concienciación de la marca.

Los **'customers needing attention'** se identifican con aquellos que están por encima de la media, pero no han comprado desde hace tiempo, como ocurre con los clientes del Cluster 3. La mejor forma de atraerlos es haciendo ofertas de tiempo limitado y recomendar productos basándonos en compras pasadas. Nuestra finalidad es reactivarlos. No obstante, podemos añadir que este grupo también podría identificarse con los **'can't lose them'**, que serán quienes han gastado mucho dinero, pero hace mucho tiempo. Para estos, debemos incorporar nuevos productos o renovar los medios de comunicación utilizados para promocionarlos, pues es muy importante que no les perdamos frente a la competencia. Deberemos prestarles mucha atención.

Los **'about to sleep'** se identifican con los clientes del Cluster 0, que han comprado hace mucho tiempo y gastado por debajo de la media. Realmente, han gastado bastante dinero, pero debemos reactivarlos para no perderles a través de recomendaciones de los productos más vendidos o en descuento.

Los **'hibernating'** son aquellos que han comprado hace mucho y han gastado poco dinero, como ocurre con los clientes del Cluster 2. Las actuaciones que deberá tomar la tienda es recrear la marca y ofrecerles otros productos que puedan ser relevantes y descuentos especiales.

Los **'lost'** serán quien es hayan comprado hace más tiempo y que hayan comprado y gastado las cantidades más pequeñas. El Cluster 5 son quienes más se identifican con esta caracterización. Intentaremos llamar su atención y atraerlos de nuevo con campañas publicitarias, pero no les prestaremos más atención, sino que nos centraremos en mantener los dos primeros grupos.

Conclusiones

Una vez realizado el trabajo, conviene analizar los datos y extraer las conclusiones que nos puedan ser de utilidad para el negocio, y cumplan con el fin que buscábamos. Lo primero en lo que nos hemos de fijar es en la serie temporal, que señala un claro declive de la cifra neta de negocios del local. Aquí identificamos el bajón que sufrió el negocio, y por ende su necesidad de cambiar la estrategia de marketing para volver a los niveles de actividad que tuvo. Además, es interesante el gráfico seasonal dentro de la serie temporal, ya que muestra que la gran mayoría de la actividad se desarrolla durante los fines de semana, sin apenas ventas de lunes a jueves. Esta información es también muy útil para la tienda.

Las 3 variables creadas son muy concretas y nos ayudan a identificar claramente los hábitos de cada cliente. Respecto al análisis exploratorio, merece la pena incidir en el número de clusters que hemos elegido. Pese a que la 'regla del codo' nos señalaba que el número óptimo de clusters podía ser 2, hemos entendido que es necesario dividir a los clientes en más grupos

porque así nuestra estrategia de marketing puede ser más concreta y mucho más personalizada, lo cual llevará a un mayor porcentaje de acierto. Finalmente nos decidimos por 6 clusters, ya que en esta cifra obtenemos la segunda mayor aceleración del gráfico. También es necesario tener en cuenta que había una serie de outliers de los que hemos decidido prescindir, ya que de lo contrario los resultados eran de peor calidad ya que se formaban grupos en los cuales se encontraban clientes con características más dispares entre sí, lo cual dificulta nuestra labor principal.

Finalmente, en la sección de 'Resultados' se encuentran analizados los seis clusters. Les hemos atribuido un nombre en función de sus características básicas así como el tipo de marketing óptimo que habría que hacerles. Esta es la información realmente útil para la tienda ya que, además de esto, permite identificar qué clientes son más importantes y merecen más atención. Al obtener clientes nuevos, será necesario clasificarlos también en uno de los grupos para poder así evaluarlos.

Tras este análisis de los clientes y de los grupos, solo resta mencionar el tratamiento de los outliers. Como ya se ha mencionado, por razones de optimización de resultados ha sido necesario excluirlos. Sin embargo, pensando en la finalidad del trabajo y de la tienda como negocio, no es lógico que una serie de clientes se queden sin ser asignados a un grupo ya que, como consecuencia de ello, no tendremos una estrategia determinada que dirigir hacia ellos. Es por eso que hemos decidido desarrollar un árbol de clasificación para esos clientes existentes que no encajaban en el modelo, así como en el improbable caso de que algún futuro cliente tampoco encaje. Después de validar el modelo y ver que obtiene alrededor de un 98% de precisión, hemos procedido a clasificar a los outliers en sus grupos correspondientes.

A la luz de estas conclusiones, pensamos que hemos conseguido cumplir el objetivo principal que nos propusimos al realizar el trabajo. Hemos sabido tratar los datos para extraer de ellos una información realmente útil, de las cuales hemos podido sacar consecuencias útiles. Realmente, el método en el que un negocio se anuncia es muy importante, y realizar esta publicidad de una manera más personalizada probablemente ayudaría a cualquier negocio. En un plano más personal, siempre se habla de la importancia de los datos que dejamos en internet para las grandes empresas a la hora de identificar nuestros gustos y necesidades, y por ello nos parecía buena idea enfocar nuestro trabajo en uno de las vertientes más importantes del mundo del data analysis. Por supuesto que nuestro modelo está lejos de extraer conclusiones personalizadas para cada usuario específico, ya que para ello haría falta un modelo mucho más potente, pero este trabajo ha sido una buena introducción a una de las herramientas con las que posiblemente nos encontremos en nuestro futuro laboral.