DESARROLLO DEL PROYECTO

Autores:

Abraham Carrera Groba

Miguel Magaña Suanzes

TABLA DE CONTENIDO

DESARROLLO DEL PROYECTO

INTRODUCCIÓN

Tal y como se ha explicado en el documento de planificación del proyecto, se pretende aumentar el volumen de ventas del negocio averiguando patrones en el comportamiento de venta de los consumidores que puedan ser aprovechables desde el punto de vista comercial.

Para lograr encontrar estos patrones se utilizará el algoritmo Apriori. Dicho algoritmo nos proporcionará reglas a partir de las que averiguaremos cuál es el par de productos entre los que existe una cierta asociación. El algoritmo procede identificando los ítems individuales frecuentes en el dataset y extendiéndolos a conjuntos de mayor tamaño siempre y cuando esos conjuntos de datos aparezcan suficientemente soportados en dicha base de datos. El objetivo será encontrar reglas de la forma if {Producto A} then {producto B} que tengan un soporte alto para los datos de nuestro dataset.

PLAN DE DESARROLLO

Las acciones llevadas a cabo a lo largo del desarrollo del proyecto han sido, en términos generales, las siguientes:

1. Importación y exploración de las propiedades del dataset.
2. Preparación de los datos.
3. Visualización y entendimiento de los datos.
4. Análisis de las distribuciones por franjas horarias.
5. Modelado de los datos e implementación del algoritmo.
6. Evaluación.
7. Revisión de objetivos

La estructura de este documento se organizará en base a estas acciones.

PROPIEDADES DEL DATASET

El dataset está compuesto por información de 21293 transacciones de las que se nos proporciona información sobre 4 atributos:

- **Date**: variable que nos indica la fecha en la que ha tenido lugar la transacción en formato YYYY-MM-DD. El rango de fechas abarca desde el 30/10/2016 al 9/4/2017.
- **Time**: variable que nos indica la hora a la que ha tenido lugar la transacción en formato HH:MM:SS. El rango abarca desde las 01:21:05 a las 23:38:41.
- **Item**: variable que nos informa del producto adquirido en la transacción. Existen 84 productos distintos.
- **Transaction**: variable que nos indica el código de transacción.

PREPARACIÓN DE LOS DATOS

Aunque el data set no incluye directamente missing values, sí encontramos transacciones en las que el valor del artículo es NONE. Probablemente esto ocurre cuando el producto no se ha registrado correctamente o el artículo fue anulado. Estas transacciones serán eliminadas del dataset.

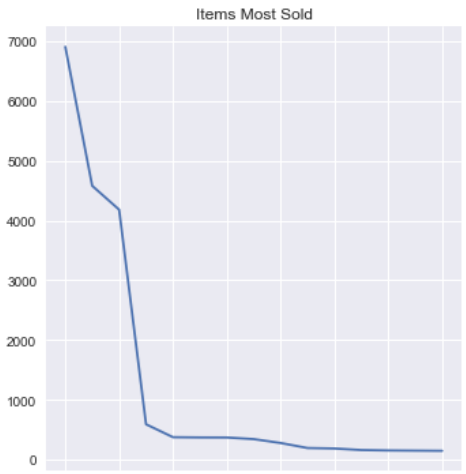
Adaptaremos además el formato de las fechas para que pueda ser manejado por Python. Se incluirá en cada fila del dataset un atributo indicando el día de la semana de forma numérica (0 para lunes y 6 para domingo).

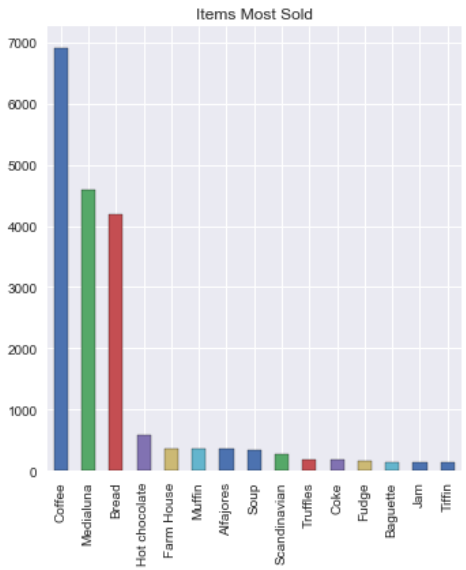
ENTENDIMIENTO DE LOS DATOS Y ANÁLISIS DE DISTRIBUCIONES

Comenzaremos analizando cuáles son los productos más vendidos en general. Obtenemos la siguiente relación:

```
Most Sold Items:

Coffee            6906
Medialuna         4580
Bread             4181
Hot chocolate      590
Farm House         374
Muffin             370
Alfajores          369
Soup               342
Scandinavian       277
Truffles           193
Coke               185
Fudge              159
Baguette           152
Jam                149
Tiffin             146
```

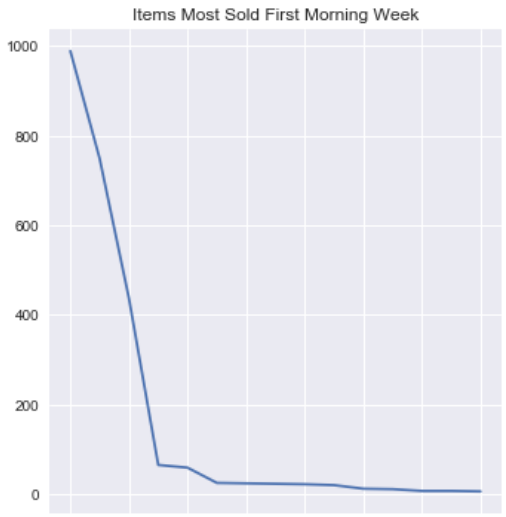


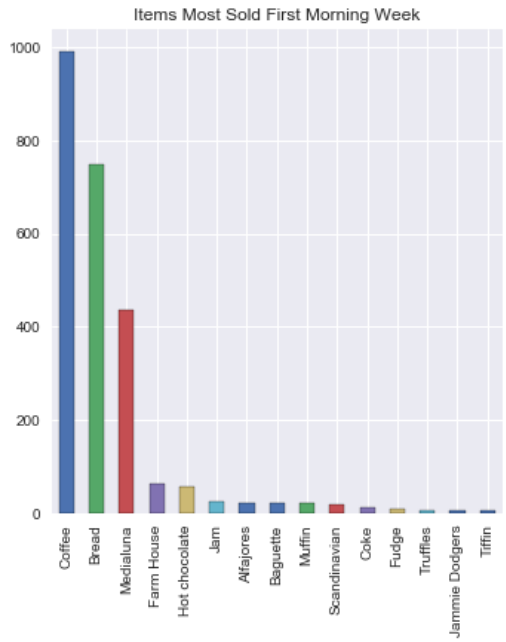


Como puede verse, existen tres productos que son los más claramente vendidos con gran diferencia: el café, las mediaslunas y el pan.

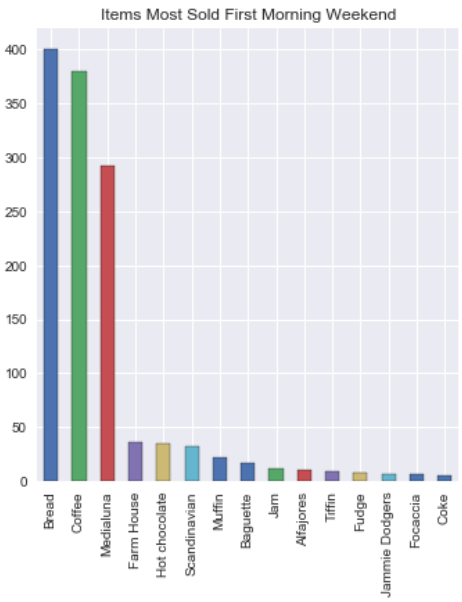
A continuación analizaremos si existen diferencias significativas al tener en cuenta distintas divisiones temporales del horario de mañana. Las subfranjas a considerar serán la primera parte de la mañana (7:30-10:00) de días de la semana, la primera parte de la mañana de días de fines de semana, la segunda parte de la mañana (10:00-12:30) de días de la semana y de días de fin de semana. Los resultados han sido los siguientes:

- Primera mitad de la mañana de días de semana

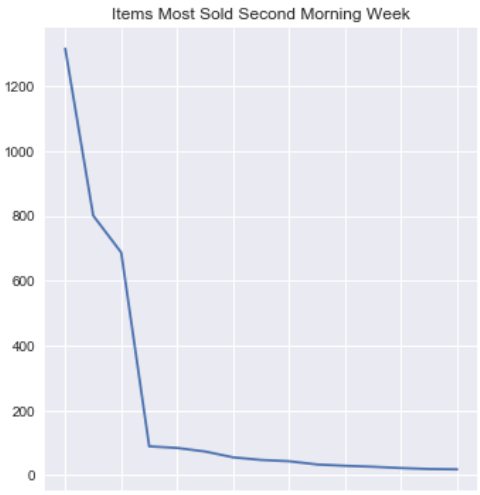


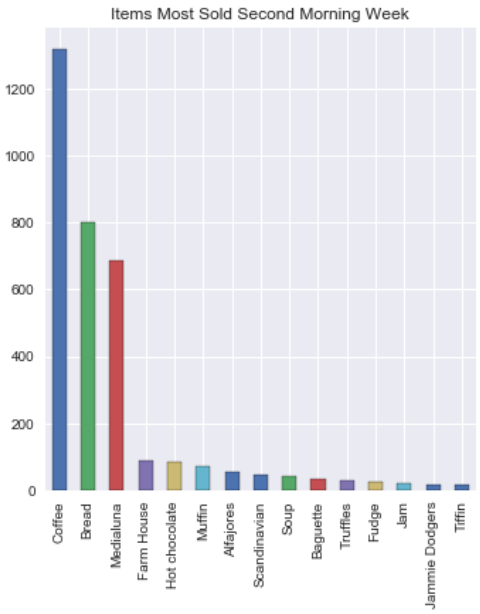


- Primera mitad de la mañana de días de fin de semana

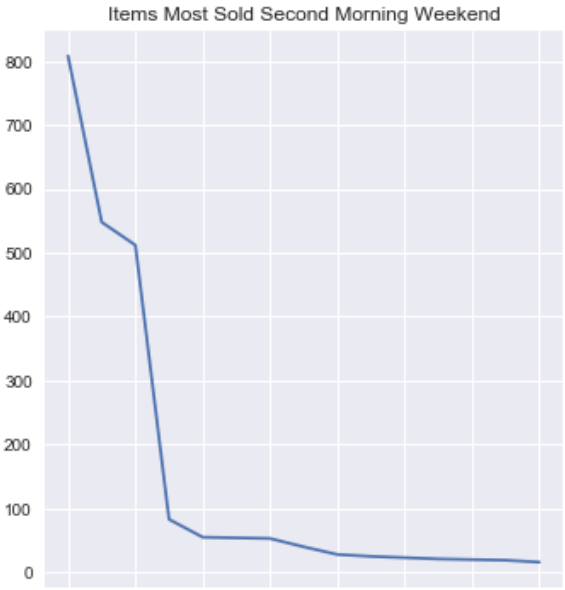


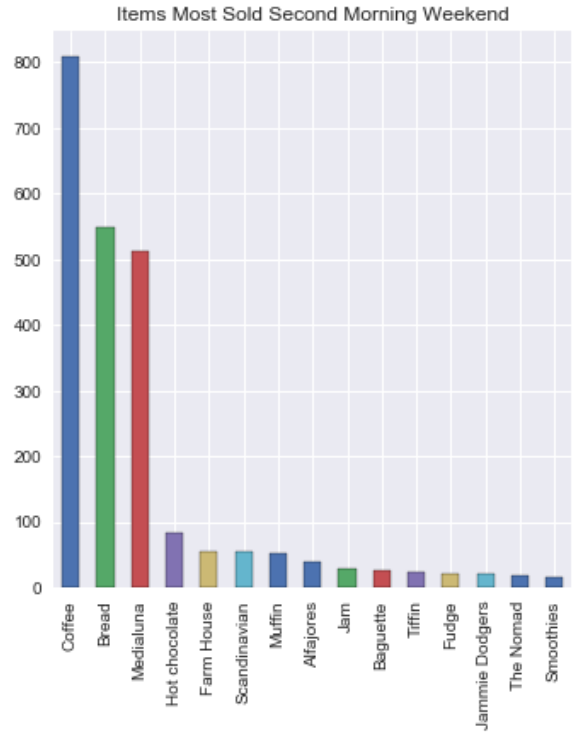
- Segunda mitad de la mañana de días de semana





- Segunda mitad de la mañana de días de fin de semana





No se observan grandes diferencias en las distribuciones entre las distintas franjas horarias consideradas. En todas el top 3 de productos más vendidos es el mismo, y a gran distancia del resto de artículos. Sin embargo, en la primera parte de la mañana de los días de la semana el producto más vendido es el café mientras que en el fin de semana es el pan. En la segunda mitad de la mañana (tanto semana como fin de semana) el top 3 coincide con el de la primera mitad durante la semana. En ninguna franja horaria el top

3 coincide enteramente con el obtenido al tener en cuenta el total de transacciones de la panadería, en el que las mediaslunas están en el segundo puesto.

ALGORITMO

Como ya se ha comentado, nuestro objetivo a nivel práctico se resume en encontrar patrones de asociación entre los artículos vendidos. Queremos ser capaces entonces de contestar a preguntas del estilo de: si alguien compra el pan ¿es probable que compre también café? Para lograrlo se ha hecho uso del Algoritmo Apriori.

Dicho algoritmo expresa las propiedades asociativas dentro de las transacciones a través de las llamadas reglas de asociación. La efectividad de estas reglas queda determinada a través de tres medidas*:* el soporte, la confianza y el *Lift*.

- **Support**: se puede considerar como el porcentaje de la cantidad total de transacciones relevantes para una asociación.

  Support(Producto) = (Transacciones que contienen al Producto) / (Transacciones totales)

- **Confidence**: corresponde a la probabilidad de que la compra de un producto involucre la compra de otro.

  Confidence (Producto 1 -> Producto 2) = (Transacciones que contienen ambos productos) / (Transacciones que contienen al Producto 1)

- **Lift**: se refiere a cómo aumentan las posibilidades de que se compre el producto 2, dado que se compra el artículo 1.

  Lift (Producto 1 -> Producto 2) = (Confidence(Producto 1 -> Producto 2)) / (Support(Producto 2))

IMPLEMENTACIÓN

La implementación del algoritmo Apriori se ha llevado a cabo utilizando el lenguaje de programación Python, más concretamente a través de los paquetes TransactionEncoder, association_rules y a priori de la librería mlxtend. En esta librería podemos encontrar múltiples herramientas para llevar a cabo tareas de ciencia de datos y machine learning.

Se aplicará el algoritmo para cada una de las franjas temporales especificadas en el apartado 5 de este documento. En primer lugar se creará una lista con los productos agrupados por transacciones. A continuación se transformarán los datos en forma de una lista de listas de Python en una matriz NumPy a través de la función fit del Transaction Encoder. Posteriormente se creará un DataFrame a través de la librería pandas con dicha matriz. Finalmente se le aplicará el algoritmo apriori a ese DataFrame especificando como parámetro el soporte mínimo de 0.01 ya que no tenemos uno preestablecido. A la hora de escoger las reglas de asociación se establece que el lift mínimo sea 1.0 porque si es menor que uno, es probable que los dos artículos no se compren juntos.

Tal y como se ha establecido, a la hora de ordenar las reglas se tomará como criterio principal el soporte, ya que estamos buscando parejas de artículos que representen un patrón de compra frecuentemente repetido en el conjunto total de transacciones.

EVALUACIÓN

Con el objetivo de asegurar la correcta implementación del algoritmo, se aplicará a un dataset de test del que previamente se conocen las reglas de asociación que se deben esperar. Dicho dataset contiene información correspondiente a 4 transacciones. En 3 de ellas el cliente ha comprado pan y café juntos y en la cuarta café y una magdalena. Se espera por tanto que la regla que aparezca como recomendada sea la de {Pan}-> {Café} con un soporte del 0.75 (se cumple para 3 de las 4 transacciones) y una confianza de 1 (siempre que se compra pan, se ha comprado café) . Efectivamente, el resultado arrojado por el algoritmo es el esperado.

RESULTADOS

Tras la aplicación del algoritmo, las reglas obtenidas como recomendadas para cada franja horaria han sido las siguientes:

FRANJA	REGLA	SOPORTE	CONFIANZA
Primera mañana semana	{Medialuna}->{Café}	0,20	0,69
Segunda mañana semana	{Medialuna}->{Café}	0,22	0,66
Primera mañana fin de semana	{Medialuna}->{Café}	0,23	0,61
Segunda mañana fin de semana	{Café}->{Medialuna}	0,25	0,44

Para todas las franjas, el conjunto de ítems implicado en las reglas es el mismo: café y medialuna. Es por esto que ese será el par de productos señalado como el que tiene una asociación más fuerte en el horario de mañana. Las cifras de soporte, aunque no son especialmente altas debido al alto número de transacciones y la variedad de las mismas, son suficientes para el contexto y más si se comparan con las cifras obtenidas para las otras reglas. La confianza es elevada en todos los casos, tal vez algo inferior en la franja de la segunda parte de la mañana de los fines de semana donde esta asociación parece darse de una forma menos intensa.

REVISIÓN DE OBJETIVOS Y CONCLUSIÓN

Tras la aplicación del algoritmo hemos obtenido una pareja de productos entre los que existe una fuerte asociación al analizar el comportamiento de consumo durante el horario de mañana: medialuna y café. Las ventas medias semanales de estos productos son respectivamente 269 mediaslunas y 406 cafés. La creación de una estrategia de venta adecuada podría llegar a elevar las ventas de mediaslunas hasta el nivel de ventas de café, lo que supondría elevar las ventas semanales en 137. Respecto al total de ventas medias semanales en horario de mañana (854 artículos vendidos) supondría un incremento de ventas del 16%, superando así el 10% marcado como objetivo con cierto margen en caso de que no se de la situación óptima de que las mediaslunas se vendan tanto como el café.

La información que le vamos a entregar al cliente como resultado del proyecto tiene por tanto el potencial necesario para lograr el incremento de ventas deseado para sanear las cuentas del negocio.