Minería de Datos. Entregable 1: Definición del problema

Álvaro López, José Carlos Gualo, David Illescas

November 29, 2021

1 Introducción y descripción de los datos

En este proyecto, nos dispondremos a tratar los datos recogidos a partir de la aplicación BlaBlaCar, la cual consiste en una red social de viajes en coche que permite la compartición de vehículo a aquellas personas que desean desplazarse al mismo lugar. De esta manera, los usuarios que utilizan esta aplicación tendrían diferentes ventajas, como, por ejemplo:

- Redistribución de los gastos. Esta ventaja la tomarían los conductores, ya que, por ejemplo, un viaje que vas a realizar solo te podría salir más barato ahorrándote gastos de combustible si compartes el coche haciendo pagar una tasa a un pasajero.
- Reducción de costes por parte del pasajero. Ya que de esta manera un viaje entre dos puntos saldría más rentable al pasajero que si lo hiciera mediante otro medio de transporte.
- Eficiencia del consumo energético, debido a la compartición de recursos, ya que esta aplicación, indirectamente, permite reducir el consume de gasolina, y por tanto se emitirían menos emisiones de CO2, beneficiando así al medio ambiente.

En cuanto a los datos que se nos ofrecen, podemos observar las siguientes variables:

NOMBRE	**DESCRIPCION**
DÍA	Variable de fecha del trayecto en formato dd/mm/aaaa (entre 01/12/2017 y 30/11/2019)
PAÍS	Variable categórica, que representa el país desde donde se ha dado de alta la ruta. Puede tomar valores ES, (España), o PT, (Portugal).
ORIGEN	Variable categórica, que representa la ciudad de origen del trayecto.
DESTINO	Variable categórica, que representa la ciudad de destino del trayecto.
IMP KM	Variable numérica, que representa el coste medio por kilómetro, que un pasajero paga.
ASIENTOS OFERTADOS	Variable numérica, que representa el número de plazas libres ofertadas a viajeros.
ASIENTOS CONFIRMADOS	Variable numérica, que representa el numero de plazas que han sido ocupadas por viajeros.
VIAJES OFERTADOS	Variable numérica, que representa el número de viajes ofertados.
VIAJES CONFIRMADOS	Variable numérica, que representa el numero de viajes llevados a cabo.
OFERTANTES	Variable numérica, que representa cuantos conductores ofrecen ese trayecto. También incluye a los ofertantes nuevos.
OFERTANTES NUEVOS	Variable numérica, que representa cuantos conductores nuevos ofrecen un servicio.

2 Trabajos anteriores

En cuanto a los previos trabajos que se han realizado sobre este dataset, podemos echar un vistazo tanto al equipo ganador como al equipo subcampeón de la datathon de 2020.

`El equipo ganador` realiza un buen trabajo de visualización además de contrastar los datos proporcionados con el censo de población para obtener la cantidad porcentual de habitantes españoles que usan la aplicación de BlaBlaCar.

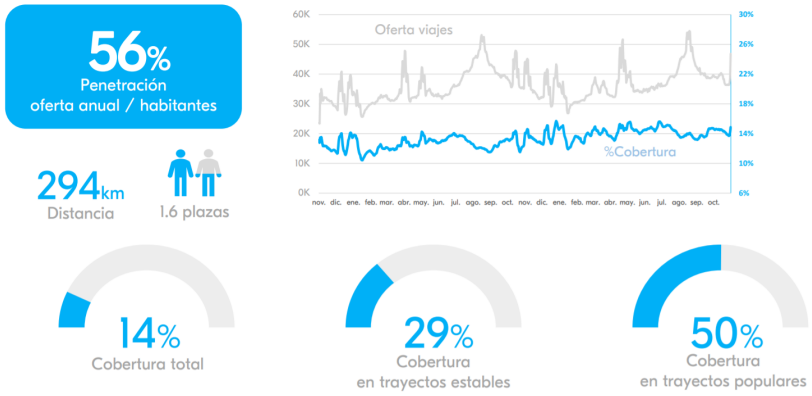

Figure 1: Visualización de los datos del equipo ganador

No obstante, el principal problema que resuelven es el caso de los viajes de un lugar a otro que no se suelen ofertar, obteniendo un conjunto de viajes con un gran porcentaje de oferta a lugares que estén cerca o de camino para poder llegar al destino inicial aunque el usuario tuviera que hacer varias paradas y/o viajes con ofertantes diferentes.

Por otro lado, `El equipo subcampeón` se enfoca en la celebración de diversos festivales musicales en la península y trata de relacionar el uso de la aplicación durante las fechas que se realizan dichos festivales en las zonas próximas para determinar si el número tanto de viajes ofertados como de asientos ocupados aumenta en relación a la media general.

3 Planteamiento de la hipótesis. Objetivos a perseguir

Entre los objetivos que se nos han ocurrido para resolver encontramos las siguientes ideas:

- Pronosticar en qué época del año va a haber un incremento masivo de ofertantes nuevos y de ofertas de viajes para evitar una posible saturación de la red.
- Hacer un análisis sobre aquellas zonas que generan más ingresos con el uso de la aplicación.
- Obtener las ciudades que obtienen un mayor incremento o disminución de población debido a los viajes que se realizan usando BlaBlarCar.

Tras debatir sobre las posibles hipótesis planteadas de las que disponíamos anteriormente, finalmente llegamos a la conclusión de que la que más nos convencía era la siguiente:

Se tratarán los datos que se nos han prestado, así como enriquecer los mismos, como veremos a continuación, para pronosticar los incrementos masivos tanto de ofertantes, como de viajes y viajeros que se pueden llegar a dar en fechas clave del año, como son los días festivos nacionales del año 2018; con el objetivo de prevenir la congestión y saturación de la red y los servidores de los que hace uso la aplicación.

4 Posibilidades de enriquecimiento de datos

Como hemos mencionado anteriormente, el enriquecimiento de datos nos va a ofrecer la oportunidad de completar la información base de la que partíamos para poder llevar a cabo nuestro objetivo de manera adecuada. Estos datos que van a enriquecer nuestra información base serán los siguientes:

- Fechas de los días festivos en Castilla-La Mancha. Para ello se han obtenido de una página oficial, datos.gob.es tres archivos en formato

Excel con los días festivos de dicha comunidad entre los años 2017 y 2019. Estos datos se han sometido a un procesado para tener un solo archivo con todas las fechas en un formato legible por la máquina.

- Datos del INE de todos los municipios de Castilla-La Mancha. Para ello se ha obtenido un archivo en formato CSV de paginas oficiales del Estado, del cual se obtiene información relativa al nombre del municipio, numero de habitantes, densidad de población, código postal, etcétera.

Estos datos serán importantes a la hora del preprocesado de los datos para la obtención del conjunto de datos sobre el cual trabajar, aunque también podrá ser usado en las otras fases restantes. Tampoco se descarta la posibilidad del uso de algún algoritmo que permita saber la distancia que se recorre en un viaje, aunque este dato tiene mayor potencial a la hora de estudiar el precio costo que le supondrá al viajero, y será de utilidad a la hora de sacar conclusiones.