Blablacar CarPooling

María Blanco González-Mohíno, José Alberto Seco Sánchez Camacho and Pablo Velasco Crespo

Contributing authors: Maria.Blanco4@alu.uclm.es; JoseAlberto.Seco@alu.uclm.es; Pablo.Velasco2@alu.uclm.es;

Abstract

La aplicación Blablacar ha ofrecido datos del uso de su aplicación, nosotros utilizaremos estos datos para la realización de un estudio. El objetivo de este estudio es obtener la capacidad de poder anticiparse sober los comportamientos futuros. ¿Qué beneficios se podrían obtener de este proyecto? Para responder a esta cuestión analizaremos los movimientos que se producen dentro de España en días festivos y también compararemos los movimientos hechos con la aplicación de Blablacar con los movimientos hechos en tren. Los beneficios son varios: facilitar la movilidad dentro del país o recducir la contaminación, entre otros.

1 Introducción

El reto **Cajamar Carpooling** nos presenta los datos obtenidos de la aplicación **Blablacar**, app utilizada por los usuarios para realizar viajes a bajo coste en vehículos de particulares.

2 Datos

Los datos compartidos nos ofrecen los viajes realizados en la Península Ibérica, es decir, viajes entre las ciudades situadas en **España** o **Portugal**, pudiendo los viajes ser realizados dentro de una misma ciudad, entre ciudades del mismo país y entre ciudades de diferente paises.

Estos datos datan de fechas entre el **01/12/2017** y el **30/11/2019**, aproximadamente 2 años.

2.1 Volumen de datos

Estos datos[1] se nos presentan en forma de **txt**, con un total de 11 columnas:

día, país, origen, destino, imp_km, asientos_ofertados, asientos_confirmados, viajes_ofertados, viajes_confirmados, ofertantes, ofertantes_nuevos.

Los datos tienen un total de **7945002** filas, **900000** trayectos y un peso de **715,0 MB**.

2.2 Trabajos anteriores realizados

Los ganadores[2] se marcaron 3 objetivos principales desde un supuesto de compra del trabajo realizado por la empresa blablacar; ellos pretendían **incrementar el negocio**, poder llegar a un **mayor número de usuarios**, y justificar la **reducción de la huella de carbono** del usuario si se usase blablacar como forma de transporte.

Para llegar a estos **objetivos** enriquecieron sus datos con las poblaciones por municipio extraidas de datos proporcionados por el **INE**.

Desde el punto de vista del desarrollo usaron **R** con la ayuda de la API **Shiny**, que les permitió visualizar los datos de una manera más interactiva, para el traspaso de estas rutas a un mapa se apoyaron en **Leatlet** y **Plotly**, además de utilizar **Igraph** para el cómputo de rutas óptimas (ya que cuenta con una función para ello).

Se encontraron diferentes errores a la hora de **cartografiar** los diferentes municipios a lo que encontraron solución con una API llamada **CartoCiudad**, en la que se les devolvía las coordenadas de los municipios, las cuales almacenaron en **texto plano** en varios archivos que subirían a **Github** con el objetivo de utilizarlo a modo de "**base de datos**".

Una vez encontradas las **APIs** a usar empezaron a desarrollar sus objetivo para atraer nueva clientela calcularon las rutas **menos probables** y calcularon **rutas alternativas cortas con más probabilidad**. También calcularon la probabilidad de que se de una ruta para un día futuro en función de datos pasados, si la probabilidad es pequeña, se busca otra ruta con más probabilidad de que se de, esto se hizo en función de los kilómetros de trayecto y el tiempo de ruta.

Para rebajar la **huella de carbono** calcularon la demanda de los trayectos, para más tarde calcular la contaminación y promocionar el ahorro ambiental.

[1] En el apéndice A podemos encontrar una tabla con una descripción de cada columna y su intervalo de valores correspondiente.

[2] Podemos encontrar el trabajo de los ganadores del reto en la página datmen.shinyapps.io/Datmen/.

A la hora de realizar los diferentes trazados de mapas estos se dividieron por provincias, en cada provincia se muestra la **penetración** y **cobertura** de trayectos **oferta/demanda**, además de mostrar los **trayectos intraprovinciales**.

Estas ideas nos han resultado especialmente útiles a la hora de desarrollar nuestro problema por dos razones en especial, la **primera** se asemeja al problema planteado en relación a los **días festivos**, ya que nosotros plantearemos un objetivo similar y la **segunda** tiene relación con el **factor inteligente**, ya que a priori puede parecer difícil definir un factor que actue de esta manera, su trabajo referente a la probabilidad de rutas futuras nos dió una idea clara sobre este requisito.

2.2.1 Problemas que encontraron en los datos

Estos problemas se han encontrado por parte de los ganadores del reto y por nosotros.

- Viajes ofertados sin plazas. De el total de los viajes estos representan un 9.3159% aproximadamente del total, 1057077 viajes se ofertaron sin ninguna plaza.
- Viajes ofertados pero sin personas suficientes para realizar el viaje. Se encontraron viajes sin ninguna persona que aceptase la oferta, son un total de 9935352 clo que supone un 87.559% del total de viajes
- Datos negativos.

3 Objetivos

Para el desarrollo de este reto nos hemos marcado un objetivo general y otro espefícico.

Objetivo principal

- Como **principal** objetivo vamos a extraer los viajes realizados en la población **Española**, estos viajes junto con los días **festivos** a nivel nacional, de autonomía y por provincia nos servirán para extraer conclusiones sobre los desplazamientos realizados en vacaciones y poder inferir sobre **comportamientos sociales futuros**, como poder ofertar más viajes a una determinada ciudad en una festividad. También necesitaremos los diferentes municipios que se encuentran en cada una de las provincias ya que serán necesarios para comprobar los desplazamientos interprovinciales y extraprovinciales. Aquí también reside nuestro factor **inteligente**.
  Primero realizaremos la estimación por las provincias de Castilla-La Mancha y Andalucía.

Objetivo específico

- Realizar la **comparativa viajeros trenes/blablacar**, usando distintitos medios de transporte, por lo tanto, podríamos discutir si se estan ofertando unos recursos no utilizados. Para este objetivo parcial necesitaremos los trenes y demás medios ofertados junto con los viajes blablacar.

4 Enriquecimiento de datos

Para poder desarrollar nuestros objetivos hemos buscado fuentes de datos externas a las ofrecidas, para poder conseguir nuestro reto de comparativas trenes/blablacar utilizaremos los datos ofrecidos por el **OTLE** (*Observatoriode Transporte y Logística en España*)[3].

Para efectuar nuestro **objetivo** con relación a los **días festivos** hemos tenido que realizar una recopilación de datos de diversas fuentes que introduciremos en un archivo **json** con las diferentes **comunidades autónomas** y los **días festivos** de estas. En este punto obtenemos varios conjuntos de datos, tres datasets con las festividades a nivel autonómico por cada uno de los años que recoje nuestro dataset de blablacar original y dos dataset con las festividades a nivel municipal de las comunidades autonomas elegidas (CLM y Andalucía).

Una vez encontrados los días festivos necesitamos encontrar la relación **municipio/autonomía** para poder identificar si en un municipio a la hora de realizar un viaje este sea por razones de **festividad** u **otras razones**. Para ello obtuvimos un **XLS** que nos ofrece **información** sobre los **municipios españoles**. De aquí podremos extraer la provincia y comunidad autónoma a la que pertenece cada municipio.

Para poder observar la **densidad** de los viajes **interprovinciales** de una manera más **gráfica**, hemos utilizado un **geojson** que contiene **información geográfica de España**. Su utilidad reside en poder **dibujar** las comunidades autónomas en las que nos centraremos posteriormente (**Andalucía** y **Castilla-La Mancha**).

Para **eliminar** todos los viajes que estuvieran relacionados con **Portugal**, utilizamos un **XLS** con información sobre los municipios de Portugal.

[3] Podemos encontrar el trabajo de los ganadores del reto en la página https://apps.fomento.gob.es/BDOTLE/

5 Preproceso y transformación

Todos los archivos que se nombran en este apartado se pueden encontrar pulsando https://github.com/JoseAlbertoSeco/BlablacarCarpooling_DataMining.git. En este apartado expondremos desde el **preprocesado** y

transformación de los datos de blablacar y los usados como **enriquecimiento**. Daremos una **visión general** de los puntos seguidos y la **organización** que nos encontramos en los archivos del proyecto. Lo primero

a lo que nos enfrentamos cuando recibimos el dataframe fue realizar la **exploración** y sacar algunos **problemas** y **conclusiones** que exponemos en el apartado 6.1 (*Problemas que encontraron en los datos*), esto se encuentra recopilado en el archivo **problems_analysis.ipynb**. A continuación, pasamos

a preprocesar el archivo de **DATOS_BLABLACAR.txt**[4], este dataset se preprocesó en un archivo local debido al gran espacio que ocupa, podemos encontrar este preprocesado en el archivo **preprocessing_blablacar.py** en el que sólo se toman los viajes que tienen más de 0 plazas confirmadas, y más de 0 ofertadas. También obtenemos sólo los viajes realizados en territorio español, este archivo genera un **.csv** llamado **blablacar_basic.csv**[5] que podemos encontrar en la carpeta de datos procesados.

Como siguiente paso de **preprocesado** nos ocupamos de los datos utilizados como **enriquecimiento**. Estos archivos se encuentran expuestos en el Anexo B. Este proceso también se realizó de manera local, ya que el dataset extraido una vez realizado el preprocesado del Blablacar también resultaba bastante pesado.

Datos procesados en el archivo **preprocessing_external_data.py**:

- Preprocesado de **calendarios**: obtuvimos 3 datasets diferentes con los festivos generales de cada comunidad por cada año (2017, 2018 y 2019), en este paso los unificamos e indicamos cuando es 'Laborable' o festivo, la indicación de festivo viene dada directamente con la fecha de ese día. Este dataset se encuentra almacenado en la carpeta de datos preprocesados con el nombre de **calendario.csv**.
- Preprocesado de **municipios**: utilizamos estos municipios para poder crear el dataset **ccaa.csv** que encontramos en la carpeta de datos procesados, en él se encuentra el nombre del municipio, la provincia a la que pertenece y su comunidad autónoma. Nos servirá para obtener las tarjetas de datos, pudiendo incluir las provincias, comunidades y si es interprovincial o no de cada viaje. Una vez llegados a este punto pudimos observar el siguiente problema:

[4] Podemos encontrar este archivo en el enlace: https://drive.google.com/file/d/1X3OAsvt03Rv9cEcW0KOcrA6ZjwBIV94Q/view?usp=sharing

[5] Podemos encontrar este archivo en el enlace: https://drive.google.com/file/d/1XYfVdHCcOCy-p40fjcKi0b6N6x6z7awh/view?usp=sharing

la información obtenida sobre el dataset original nos indicaba que habia ciudades portuguesas, en este momento nos dimos cuenta que había ciudades pertenecientes a diferente países de la Unión Europea como Francia, o Italia. Este dataset resulta de gran ayuda a la hora de extraer esos viajes que se encuentran fuera de nuestras fronteras; estos paises extraidos se encuentran recogidos en el documento **ciudades_no_españolas_extraidas.txt** en la carpeta de datos intermedios.

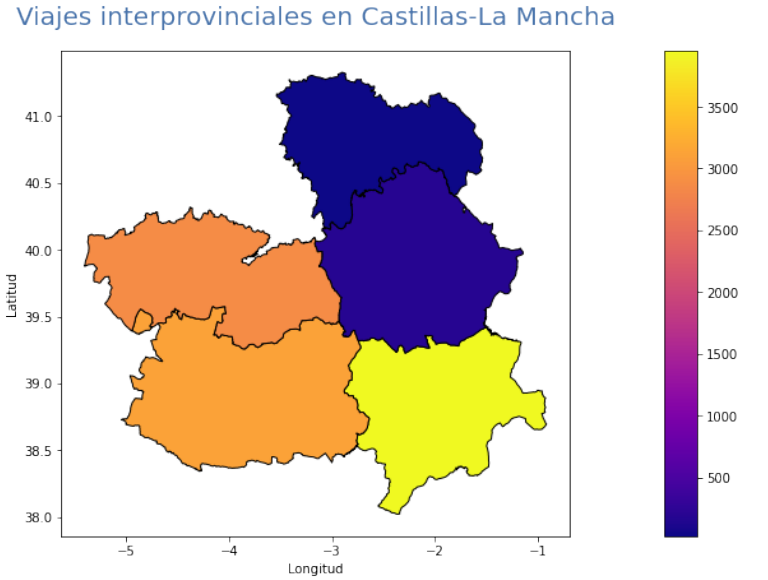
- Preprocesado de **coordenadas**: el dataset que contiene las coordenadas ha sido modificado de manera que solo contiene las columnas NAME_1 (comunidad autónoma), NAME_2 (municipio) y geometry (geolocalización del fecha). Este dataset se ha almacenado en el archivo **geolocalizaciones.geojson** que podemos encontrar en la carpeta de datos preprocesados.
- Preprocesado de **trenes**: este dataset contiene los viajeros totales de diversos medios de transpote realizados por meses, estos se han incluido en la carpeta de datos intermedios con el nombre de *trenes.csv* dado que en un cuaderno **colab** incluido en la carpeta *notebook* realizamos una tarjeta de datos que exponemos en los siguientes puntos.

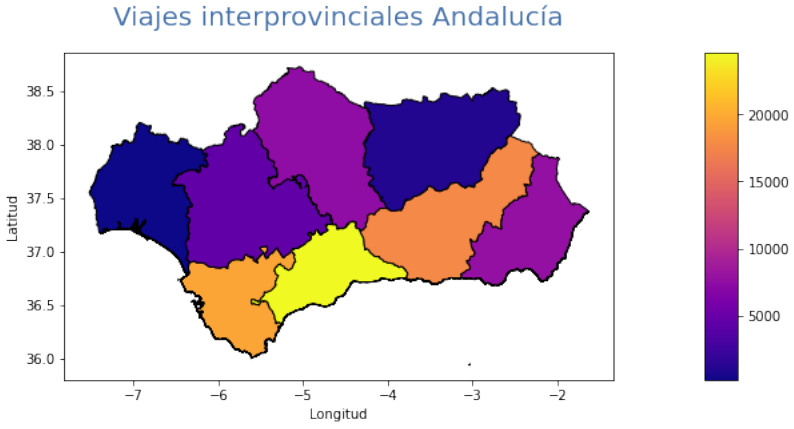
Una vez realizado el preprocesado en archivos **.py** se han realizado tres archivos **colab** para extraer una serie de tarjetas de datos y/o mostrar datos obtenidos:

- **GeolocalizacionProvincial**: es este cuaderno se han obtenido diferentes dataset, los dos primeros, los que utilizaremos como **tarjetas de datos**, son aquellos que centran los datos del blablacar en las comunidades autonomas de **Castilla-La Mancha** y **Andalucía**, en este punto se han añadido tres columnas que por razones de complejidad no se han podido añadir en los archivos **.py**. Estas columnas son las provincias de origen y destino de cada viaje, así como si el viaje es interprovincial o no en una columna extra. También hemos unificado los festivos de cada comunidad autónoma con sus respectivos municipios, obteniendo así dos datasets extras *df_totalFestivosCLM.csv* y *df_totalFestivosAndalucia.csv*. Por último, en este cuaderno se han obtenido otros dos dataset para poder hacer una **comparativa** cuando se realice el algoritmo de cluster elegido. Estos datasets son similares a las tarjetas de datos pero con la pecularidad de que se encuentran únicamente los viajes que se realizan cada uno de los días festivos y no en un espacio temporal.

- **GráficasTrasPreprocesado**: en este cuaderno se han estudiado los viajes interprovinciales usando los dataset de *df_AndaluciaLocalizado* y *df_CLMLocalizado* (tarjetas de datos). Esta información se ha mostrado en un mapa. Para ello, se ha utilizado el archivo *geolocalizaciones.geojson*, que gracias a la columna **geometry** nos permite plasmar los viajes interprovinciales totales realizados en las provincias de las comunidades de Andalucía y

Castilla-La Mancha sin ninguna restricción, es decir, viajes totales realizados durante todo el espacio temporal que ocupa el dataset.
A continuación, se muestran estos viajes en el mapa, indicamos por color el número de **viajes realizados por provincia**.





- **TrenesPreprocesado**[6]: archivo en el que unificamos nuestro dataset de trenes con el obtenido en blablacar, asi formamos una tarjeta de datos preparada para cumplir nuestro objetivo secundario.

[6]Por razones de tamaño el archivo resultante se ha almacenado en el siguiente enlace: https://drive.google.com/drive/u/0/folders/10ElkZ_vYOs5R0q4pKDEeIf4bphgYK-ED

6 Tarjeta de datos

Como nos hemos centrado en **Andalucía** y **Castilla-La Mancha**, hemos obtenido **2 tarjetas de datos**, las cuales contienen la siguiente información sobre los viajes: día, país, origen, destino, asientos ofertados, asientos confirmados, viajes confirmados, ofertantes, ofertantes nuevos, comunidad autónoma origen, comunidad autónoma destino, provincia origen, provincia destino y si el viaje es interprovincial o no.

Para poder obtener conclusiones a partir de los datos de las tarjetas, vamos a utilizar dataframes con **todas las festividades** (nacionales, autonómicas y locales) de ambas comunidades autónomas. Estos dataframes son de la forma:

	Festivos	Municipio
0	06/01/2017	[Tarancón, Illescas, Torrijos, Albacete, La Ro...
1	14/04/2017	[Tarancón, Illescas, Torrijos, Albacete, La Ro...
2	01/05/2017	[Tarancón, Illescas, Torrijos, Albacete, La Ro...
3	15/08/2017	[Tarancón, Illescas, Torrijos, Albacete, La Ro...
4	12/10/2017	[Tarancón, Illescas, Torrijos, Albacete, La Ro...

6.1 Castilla-La Mancha

	DIA	PAIS	ORIGEN	DESTINO	...
628445	2017-01-11	es	Villarrobledo	Albacete	...
713453	2017-08-12	es	Villarrobledo	Toledo	...
628469	2017-12-25	es	Villarrobledo	Albacete	...
725795	2017-12-25	es	Villarrobledo	Ciudad Real	...
628511	2018-03-29	es	Villarrobledo	Albacete	...

6.2 Andalucía

	DIA	PAIS	ORIGEN	DESTINO	...
286748	2019-08-28	es	Algeciras	Granada	...
369219	2019-08-28	es	Algeciras	Sevilla	...
455593	2019-08-28	es	Algeciras	Jerez de la Frontera	...
522130	2019-08-28	es	Algeciras	Marbella	...
552584	2019-08-28	es	Algeciras	San Fernando	...

6.3 Trenes

Por otro lado y para cumplir con nuestro objetivo secundario hemos creado una tarjeta de datos con los datos del blablacar y los datos de los trenes que nos será útil para la puesta en marcha de este objetivo.

	DIA	ASIENTOS_OFERTADOS	ASIENTOS_CONFIRMADOS	...
0	2017-01	54015.0	12130	...
1	2017-02	28905.0	6550	...
2	2017-03	66568.0	15968	...
3	2017-04	27302.0	6128	...
4	2017-05	72446.0	19176	...

7 Líneas de Trabajo

En este punto tenemos varias lineas de trabajo abiertas, aunque antes de empezar a realizar cualquiera de ellas agruparemos por fechas los diferentes viajes, es decir si los días festivos son días como el 01/01/2017 y 03/01/2017 también se contará el día 2 como día festivo. Se ha tomado la decisión de dejar 7 días para considerar una semana festiva.

Una vez esto realizado expandiremos la fecha en mes, semana y tipo (si es festivo o no).

A continuación se seguirán varias lineas de trabajo:

- **Primera**: Primero se realizará una clasificación binaria para poder inferir si se realizarán viajes en una determinada semana (o día si expandiesemos la fecha a día).

- **Segunda**: Se realizará una regresión para poder predecir el número de viajes/viajeros que se realizarán en una semana o día.

- **Tercera**: Como tercera línea tenemos una clasificación multiclase, para realizar el etiquetado es posible que tengamos que utilizar algún algoritmo de cluster, el objetivo, como en las líneas anteriores, sería predecir el número de viajes a realizar.

- **Cuarta**: como cuarta línea de trabajo, ya que no se han encontrado fechas concretas de realización de viajes de autobuses y trenes se realizará una simple comparación de viajes mensuales con cada uno de los medios de transporte.

Appendix A Datos Blablacar Carpooling

En este anexo se ha incluido una tabla con las diferentes variables con su correspondiente descripcion e intervalo de valores. Esta tabla corresponde a los datos que nos han sido entregados.

Característica	Descripción	Valores
Día	Variable fecha dd/mm/aaaa	01/12/2017-30/11/2019
País	Donde se ha dado de alta la ruta	String - [ES,PT]
Origen	Ciudad de origen de la ruta	String - Ciudad ES or Ciudad PT
Destino	Ciudad de destino de la ruta	String - Ciudad ES or Ciudad PT
Imp_km	Importe medio por kilómetro y pasajero de los viajes realizados	float64
Asientos_ofertados	Nº total de plazas ofertadas (sin conductor)	float64
Asientos_confirmados	Nº total de plazas finalmente ocupadas (sin conductor)	int64
Viajes_ofertados	Nº de viajes ofertados	int64
Viajes_confirmados	Nº de viajes realizados	int64
Ofertantes	Nº de conductores distintos que han ofrecido la ruta	int64
Ofertantes_nuevos	Nº de nuevos ofertantes (primera vez que ofrecen un servicio)	int64

Appendix B Datos de refuerzo

B.1 Festivos

Características	Descripción	Valores
Fiesta Nacional	Días festivos comunes a todas las comunidades autónomas	String - ["1 de enero"-"31 de diciembre"]
Comunidad Autónoma X	Días festivos exclusivos de X comunidad autónoma	String - ["1 de enero"-"31 de diciembre"]

B.2 Municipios

Características	Descripción	Valores
Codine	Código utilizado para hacer referencia a un municipio en el INE	int - [1001-52001]
Municipio	Nombre de un municipio español	String - Nombre del municipio
Autonomía	Comunidad Autónoma a la que pertenece el municipio	String - Comunidad Autonoma española

B.3 Portugal

Características	Descripción	Valores
Ciudad	Nombre de ciudad portuguesa	String
Municipios (concelhos)	Condado al que pertenece la ciudad	String
Distrito	Distrito al que pertenece la ciudad	String
Población	Población	int
Ciudad desde...	Año o fecha desde que la ciudad es considerada como tal	String

B.4 Viajes trenes

Características	Descripción	Valores
Total de viajeros	Nº viajes totales	int
Transporte urbano	Nº viajes en transporte urbano	int
Urbano por metro	Nº viajes en metro	int
Transporte urbano regular por autobús	Nº viajes en autobús	int
Transporte interurbano regular	Nº viajes transporte interurbano	int
Interurbano por autobús regular	Nº viajes autobús regular interurbano	int
Transporte interurbano regular por autobús: Cercanías	Nº viajes en autobús cercanías	int
Transporte interurbano regular por autobús: Media distancia	Nº viajes en autobús distancia media	int
Transporte interurbano regular por autobús: Larga distancia	Nº viajes en autobús distancia larga	int
Interurbano por ferrocarril	Nº viajes ferrocarril interurbano	int
Ferrocarril: Cercanías	Nº viajes en tren cercanías	int
Ferrocarril: Media distancia	Nº viajes en tren distancia media	int
Ferrocarril: Larga distancia	Nº viajes en tren distancia larga	int
AVE	Nº viajes AVE	int
Resto ferrocarril larga distancia	Nº viajes en tren resto distancia larga	int
Interurbano Aéreo (interior)	Nº viajes en avión interurbano	int
Aéreo: Peninsular	Nº viajes en avión península	int
Aéreo: Peninsula-Resto Territorio	Nº viajes en avión fuera península	int
Aéreo: Interinsular	Nº viajes en avión entre islas	int
Interurbano Marítimo (cabotaje)	Nº viajes en barco interurbano	int
Transporte especial y discrecional	Nº viajes especiales y discrecionales	int
Transporte especial	Nº viajes especiales	int
Transporte especial escolar	Nº viajes especiales escolares	int
Transporte especial laboral	Nº viajes especiales laborales	int
Transporte Discrecional	Nº viajes discrecionales	int
DÍA	Dia en la que se realizaron los viajes	String

B.5 Geolocalizaciones Provinciales

Características	Descripción	Valores
NAME_1	Nombre de la Comunidad Autónoma	String - Comunidad Autónoma española
NAME_2	Nombre de la provincia	String - Provincia española
CC_2	Identificador único de provincia	Int - [00-99]
NAT2018	Tasa bruta de natalidad, es decir, la relación entre el número de nacimientos ocurridos y la población existente.	float64
Geometry	Lista de tipo MultiPoligon con coordenadas X e Y	Multipoligon