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. Para esto 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 otros medios de transporte. 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.
Este proyecto se pueden encontrar en el siguiente Github

Los datos datos en crudo y utilizados para las lineas de trabajo se encuentran en la siguiente carpeta. Video sobre el proyecto en este link.

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.

[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/.

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

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** a la **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 C. 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

[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

viaje. Una vez llegados a este punto pudimos observar el siguiente problema: 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.

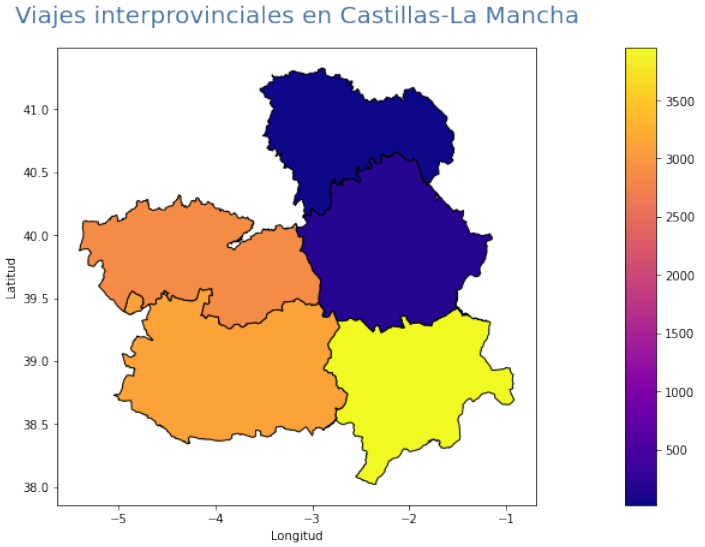
- Preprocesado de **coordenadas**: el dataset que contiene las coordenadas ha sido modificado de manera que solo contiene las columnas NAME_1 (comunidad autónoma), NAME_2 (municipio) y geometry (geolocalización del fecha). Este dataset se ha almacenado en el archivo **geolocalizaciones.geojson** que 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.

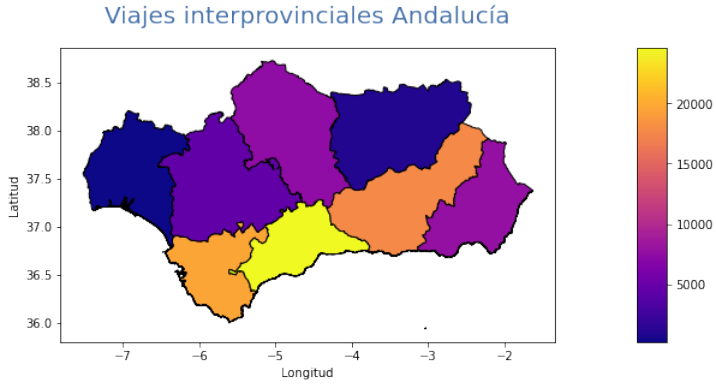
Una vez realizado el preprocesado en archivos **.py** se han realizado tres archivos **colab** para extraer una serie de tarjetas de datos y/o mostrar datos obtenidos:

- **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.
- **Agrupación_Fechas**: En este colab simplemente indicamos los días no festivos y festivos y preparamos las tarjetas de datos que utilizaremos en las lineas de trabajo.

[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
06/01/2017	[Tarancón, Illescas, Torrijos, Albacete, La Ro...
14/04/2017	[Tarancón, Illescas, Torrijos, Albacete, La Ro...
01/05/2017	[Tarancón, Illescas, Torrijos, Albacete, La Ro...
15/08/2017	[Tarancón, Illescas, Torrijos, Albacete, La Ro...
12/10/2017	[Tarancón, Illescas, Torrijos, Albacete, La Ro...

6.1 Castilla-La Mancha

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

6.2 Andalucía

DIA	PAIS	ORIGEN	DESTINO	...
2019-08-28	es	Algeciras	Granada	...
2019-08-28	es	Algeciras	Sevilla	...
2019-08-28	es	Algeciras	Jerez de la Frontera	...
2019-08-28	es	Algeciras	Marbella	...
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	...
2017-01	54015.0	12130	...
2017-02	28905.0	6550	...
2017-03	66568.0	15968	...
2017-04	27302.0	6128	...
2017-05	72446.0	19176	...

7 Líneas de Trabajo

En este punto tenemos varias lineas de trabajo abiertas:

- **Primera**: Primero se realizará una clasificación binaria para poder inferir si se realizarán viajes en una determinado día, con un origen y un destino.

- **Segunda**: Se realizará una regresión para poder predecir el número de viajes/viajeros que se realizarán en un día en específico, esto se encuentra planteado de la siguiente manera, desde el punto de vista empresarial, si yo tengo un viaje de X a Y, un determinado día, con J plazas ofertadas, L plazas ya reservadas, un determinado número de ofertantes, ¿cuantos viajes se van a realmente realizar?.

- **Tercera**: Como tercera línea tenemos una clasificación multiobjetivo, el objetivo, como en las líneas anteriores, sería predecir que ruta se dará si tenemos el número de viajes que podríamos realizar en ese mes, las plazas que necesitariamos, etc.

- **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.

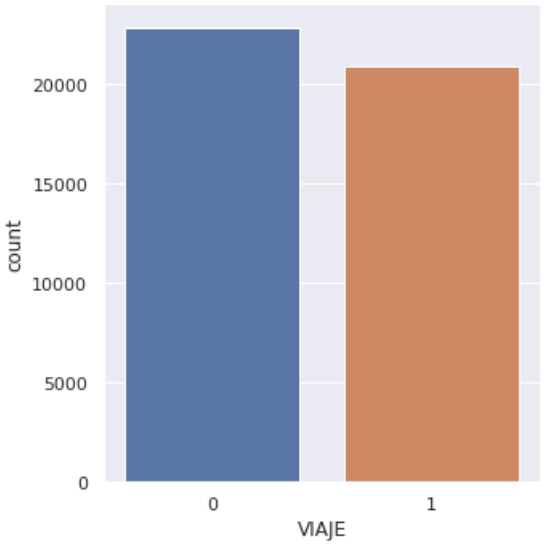
7.1 Clasificación binaria

Para predecir si va a haber viaje o no en el día elegido será necesario hacer una clasificación binaria, para ello, vamos a tener que preparar los dataframes de Andalucía y el de Castilla-La Mancha. Primero se van a codificar las variables no numéricas (ORIGEN y DESTINO) después, se va a modificar la columna DIA, la cual va a pasar a ser un número entre 1 y 365 que representará el día del año. Esta columna nos ayudará a la hora de añadir la columna VIAJE, la cual tiene un valor de 0 o 1 y representará si hay viaje en ese día o no lo hay (0 no hay viaje, 1 si hay viaje). Como resultado obtendremos una tabla como la siguiente:

DIA	AÑO	SEMANA	...	VIAJE
305	2017	44	...	1
305	2017	44	...	1
305	2017	44	...	1
...	...	...	...	...
298	2019	43	...	0
298	2019	43	...	0
298	2019	43	...	0

7.1.1 Balanceo

En Castilla-La Mancha. Hay un total de 22842 rutas sin viaje y 20862 viajes:



7.1.2 Random Forest

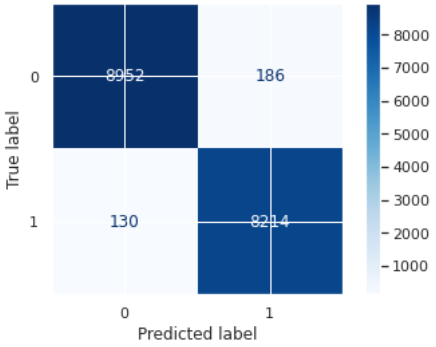
Para poder predecir, haremos un modelo utilizando Random Forest. Los resultados del modelo son:

METRICS	VALUE
MAE	0.018075735041757236
MSE	0.018075735041757236
R^2	0.9275868541647004

	Precision	Recall	F1-score	Support
0	0.99	0.98	0.98	9138
1	0.98	0.98	0.98	8344
accuracy			0.98	17482
macro avg	0.98	0.98	0.98	17482
weighted avg	0.98	0.98	0.98	17482

Feature	Importance
AÑO	0.085521
MES	0.004178
SEMANA	0.054039
ORIGEN	0.442571
DESTINO	0.248581
INTERPROVINCIAL	0.165110

Confussion Matrix Castilla-La Mancha



Como podemos ver es un modelo bastante bueno.

7.1.3 Predicción para el Usuario

Gracias a los modelos que hemos creado podemos predecir si en un día en específico habrá viaje o no de un municipio a otro. Por lo que el usuario podrá facilitar al sistema estos datos y él le dará el resultado.

Datos de entrada:

- **Origen**. Municipio desde el que parte el viaje. Para el ejemplo: Manzanares
- **Destino**. Municipio al cual se desea llegar. Para el ejemplo: Ciudad Real
- **Fecha**. Cuándo se realizará el viaje (día, mes y año). Para el ejemplo: 2019-03-04

Con estos datos de entrada se obtiene el día, el mes, el año y si es interprovincial o no el viaje y se predice:

¿Habrá un viaje a Ciudad Real con origen en Manzanares el 2019-03-04?
Respuesta: Si

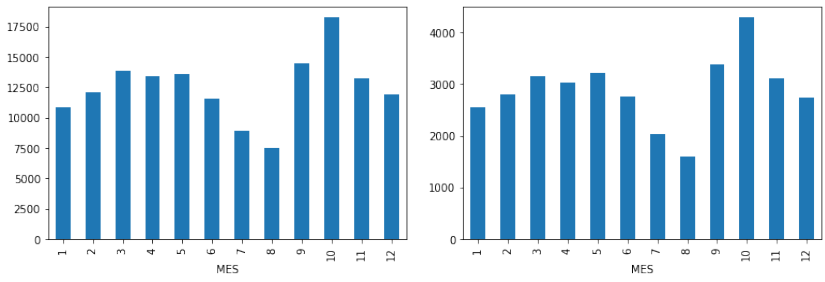
7.2 Regresión

Ahora es el turno de predecir si a Blablacar debería publicitar más un viaje en específico o no. Para ello vamos a utilizar Regresión. Con esto se pretende ayudar a la empresa a la hora de tomar decisiones como podrían ser decisiones publicitarias sobre viajes como poner anuncios sobre un determinado viaje, etc.
Para ello, como viene siendo costumbre, empezaremos por codificar las variables categóricas, es decir, vamos a transformar las variables no numéricas, como origen, destino... en numéricas.

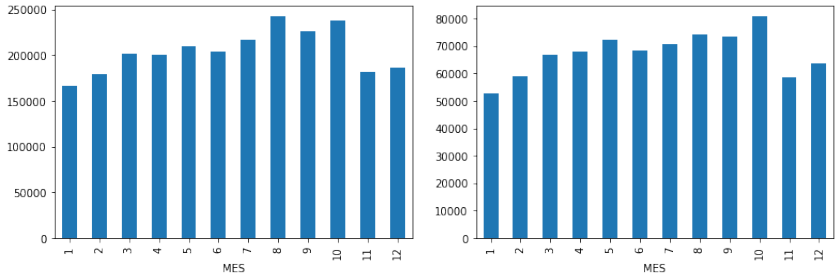
7.2.1 Distribución de los datos

Cuando las tengamos codificadas comprobaremos cómo están distribuidos los datos por meses:

- **Castilla-La Mancha**. Viajes ofertados y viajes confirmados.



- **Andalucía**. Viajes ofertados y viajes confirmados.



Como se puede observar, los datos no se encuentran desbalanceados, por lo que no será necesaria la utilización de técnicas de balanceo.

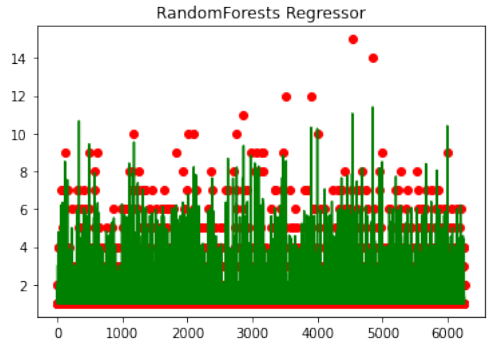
7.2.2 Random Forest

Se ha realizado un modelo base con pequeños cambios para conseguir mejoras en cada modelo. Hemos decidido incluir en el modelo las siguientes características: año, semana, mes, asientos ofertados, viajes ofertados, destino, origen, interprovincial, festivo, asientos ofertados y asientos confirmados.
Los asientos confirmados se han añadido debido a que como ya conocemos todos los datos de cada ruta diaria, para saber si esta ruta se va a realizar o no, parece necesario saber cuántas plazas se encuentran reservadas por el momento.
Modelo de Casilla-La Mancha:

- **Relevancia de características**

Atributos	Decision tree
AÑO	0.000811
SEMANA	0.002931
MES	0.000803
ASIENTOS_OFERTADOS	0.008279
ASIENTOS_CONFIRMADOS	0.897639
VIAJES_OFERTADOS	0.043028
OFERTANTES	0.040500
ORIGEN	0.002343
DESTINO	0.002706
INTERPROVINCIAL	0.000518
FESTIVO	0.000441

- **Errores**:
  - **MAE**: 0.22493928742610642
  - **MAPE**: 0.115896819591292
  - **MSE**: 0.21382761223837674
  - **R^2**: 0.8427781340342695
- **Representación gráfica**:



Como resultado obtenemos unas métricas del modelo bajas, hay poco ruido en los datos y el modelo parece ser bastante bueno.

7.2.3 Predicción

A continuación vamos a ver qué predice el modelo dándole los siguientes datos:

- **Origen**. Albacete.
- **Destino**. Guadalajara.
- **Fecha**. 2018-02-13.
- **Asientos ofertados**. 11.
- **Asientos confirmados**. 4.
- **Viajes ofertados**. 4.
- **Rutas disponibles**. 4.
- **Festivo**. No.

Con estos datos obtenemos la siguiente predicción:

Viajes que se van a confirmar: 2.05

Corresponde a un 51%, ¿se debe dar publicidad a estas rutas? Si

Cuando los viajes que se han predicho correspondan a menos del 65% de los viajes ofertados que le hemos introducido significa que se prevé que van a haber pocos por lo que se debe publicitar.

7.3 Clasificación multiobjetivo

Para predecir la ruta, va a ser necesaria hacer una clasificación multiobjetivo, para ello, primero vamos a tener que codificar las variables no numéricas. Cuando lo tengamos podemos comenzar con la clasificación multiobjetivo. Para ello utilizaremos la clase Multioutput de sklearn. El propósito de esta clase es ampliar los estimadores para poder estimar una serie de funciones objetivo (f1,f2,f3...,fn) que se entrenan en una única matriz de predicción X para predecir una serie de respuestas (y1,y2,y3...,yn). Se trata de una estrategia sencilla para ampliar los clasificadores que no admiten de forma nativa la clasificación multiobjetivo.

7.3.1 Predicción

Para la predicción, va a ser necesario que el usuario facilite los siguientes datos:

- **Mes**. Mes en el que se realizará el viaje. Para el ejemplo 3, marzo.
- **Año**. Año en el que se realizará el viaje. Para el ejemplo 2022.
- **Interprovincial**. Indica si el viaje es dentro de la misma provincia o si no lo es. En nuestro caso es interprovincial.
- **Número de asientos**. Número de asientos que el usuario necesitará. Para el ejemplo 4.
- **Número de viajes**. Número de viajes que se realizarían. Para el ejemplo 5.

Con estos datos de entrada se obtiene el día, el mes, el año y si es interprovincial o no el viaje y se predice:

Origen: Albacete, Destino: Alcázar y Festivo: no

7.4 Transporte público VS Blablacar

7.4.1 Relación por pasajeros

Las gráficas con las que han realizado el informe se encuentra aquí C.1. A finales de 2017 se aprecia una aumento en el uso de Blablacar notable. El número de viajeros en autobús es siempre mayor con respecto a los de Blablacar, excepto los meses de agosto de 2018 y 2019. El metro domina a Blablacar hasta finales de 2017. Luego, éste prácticamente le dobla en el número de pasajeros. Con el ferrocarril pasa lo mismo que en el caso anterior; pero mucho más notable. El Ave apenas se aprecia con respecto al Blablacar. Todo mes de agosto, como se puede ver en la gráfica donde se compara con el total, las personas usan menos el transporte público. Gracias a Blablacar, se puede suplir esa carencia, pues los usuarios siguen usando este transporte.

7.4.2 Relación porcentual

En esta sección se va a ver la relación que hay entre los distintos transportes con el Blablacar.

	Autobus	Metro	Ferrocarrol	AVE
count	33.000000	33.000000	33.000000	33.000000
mean	0.601008	1.210785	2.398038	68.950785
std	0.386182	0.812804	1.561002	42.858322
min	0.032040	0.063230	0.126179	3.444632
25%	0.116486	0.236439	0.460376	14.032680
50%	0.776086	1.483904	3.090173	92.354167
75%	0.855498	1.665929	3.350276	101.094153
max	1.246860	2.936332	5.329435	122.355217

- Los pasajeros del autobús son un poco más del 20% con respecto a los del Blablacar.
- A su vez, los del Blablacar son casi un 50% más que los del metro.
- Sin embargo, los del ferrocarril son 1/3 con respecto a los del Blablacar.
- La diferencia entre el AVE y Blablacar se aprecia en la gráfica anterior.

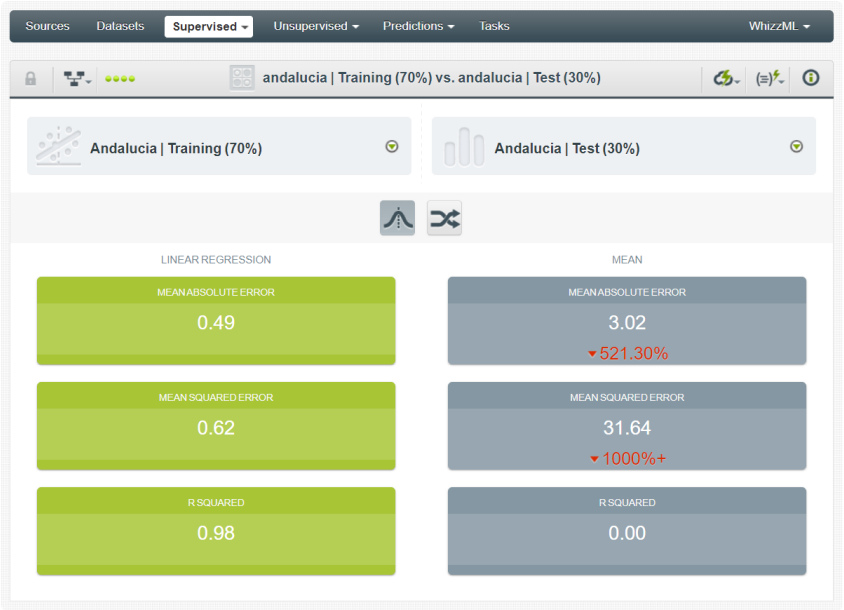
Autobús VS Blablacar
Como hay tantos viajes de autobús, vamos a hacer una comparativa con los de Blabacar para ver si hay meses en los que tienen un número de pasajeros similar. Para ello vamos a ver los meses que el Blablacar tienen menos de un 10% de diferencia. A continuación se refleja el resultado:

Fecha	Autobus
2018-07	0.987680
2018-08	1.170268
2018-12	0.930368
2019-04	1.005263
2019-07	0.974157
2019-08	1.246860

Se aprecia que los meses de julio y agosto Blablacar tiene casi o más usuarios que el autobús. Si queremos que aumente el uso de Blablacar en los otros 10 meses habría que potenciarlo.
La representación gráfica se encuentra aquí C.2

8 BigML y Andalucia

Debido al gran set de datos obtenidos para realizar las lineas de trabajo de Andalucía, se ha decidido comentar su implementación en colab e incluir los resultados obtenidos a la hora de realizar su regresión en BigML.



Appendix A Datos Blablacar Carpooling

En este anexo se ha incluido una tabla con las diferentes variables con su correspondiente descripción 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

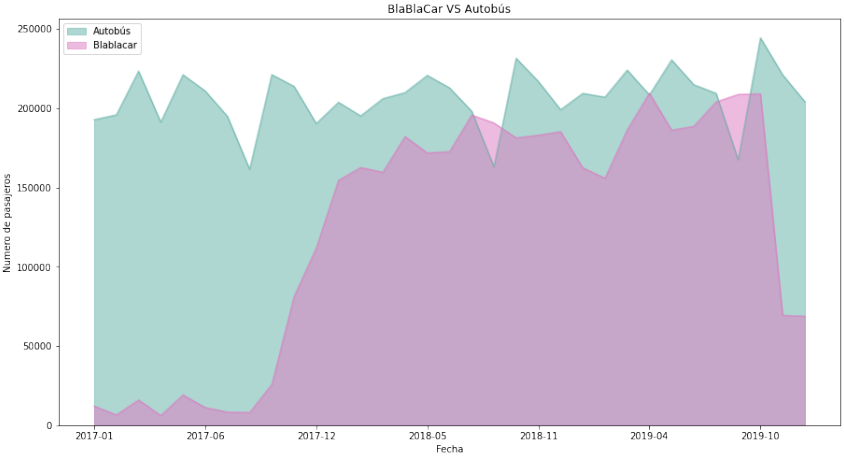
Características	Descripción	Valores
Total de viajeros	Nº viajes totales	int
Transporte urbano	Nº viajes en transporte urbano	int
Urbano por metro	Nº viajes en metro	int
Transporte urbano regular por autobús	Nº viajes en autobús	int
Transporte interurbano regular	Nº viajes transporte interurbano	int
Interurbano 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

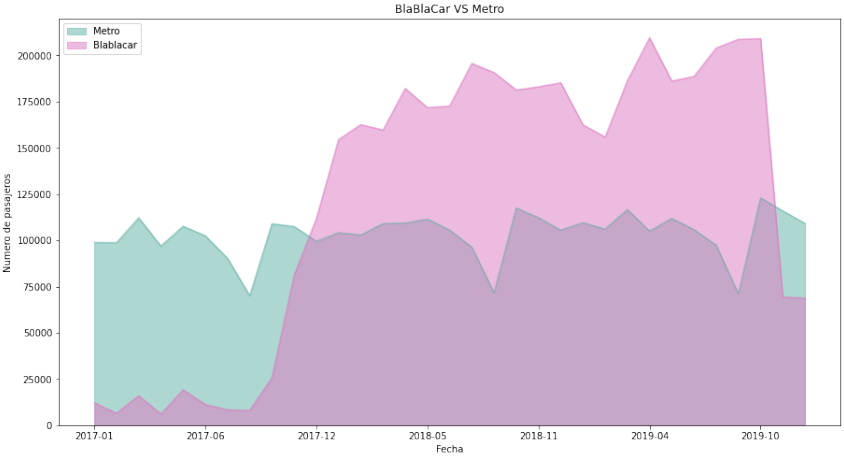
B.5 Geolocalizaciones Provinciales

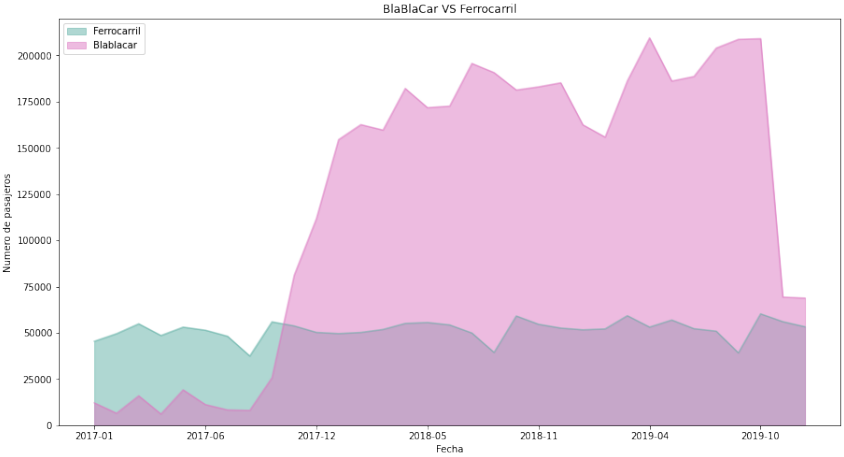
Características	Descripción	Valores
NAME_1	Nombre de la Comunidad Autónoma	String - Comunidad Autónoma española
NAME_2	Nombre de la provincia	String - Provincia española
CC_2	Identificador único de provincia	Int - [00-99]
NAT2018	Tasa bruta de natalidad, es decir, la relación entre el número de nacimientos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

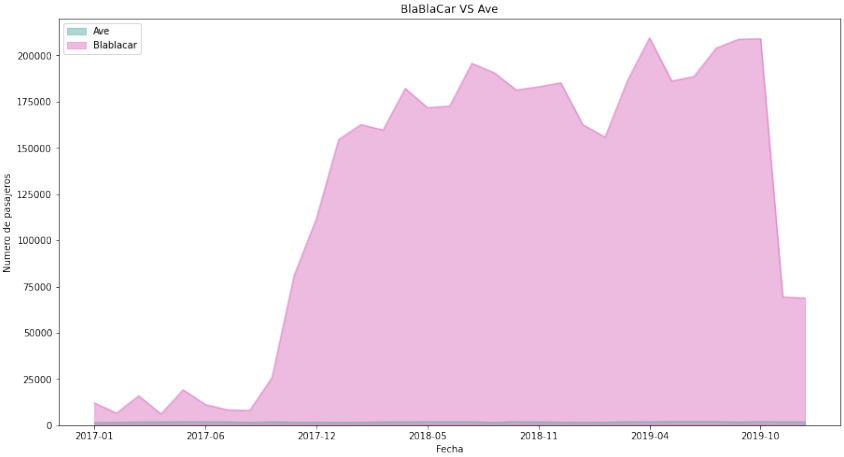
Appendix C Gráficas de trenes

C.1 Pasajeros

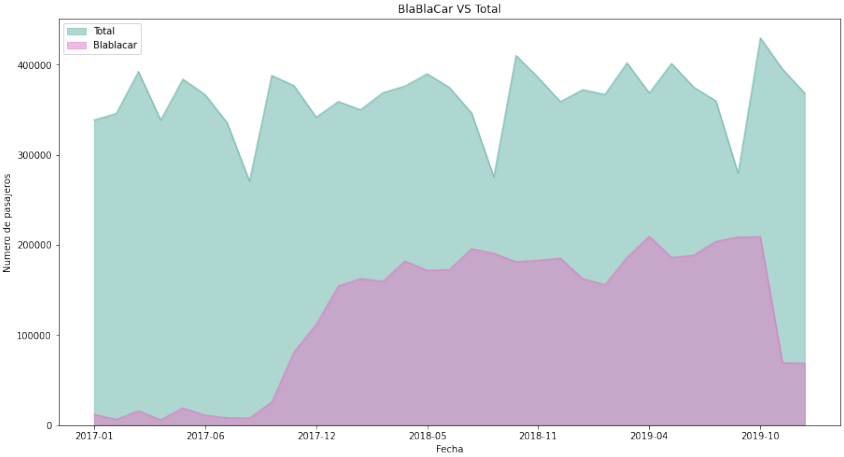




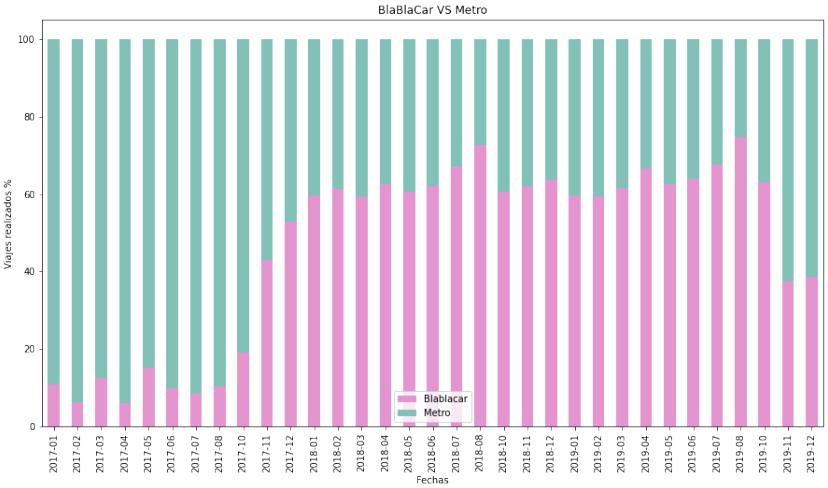




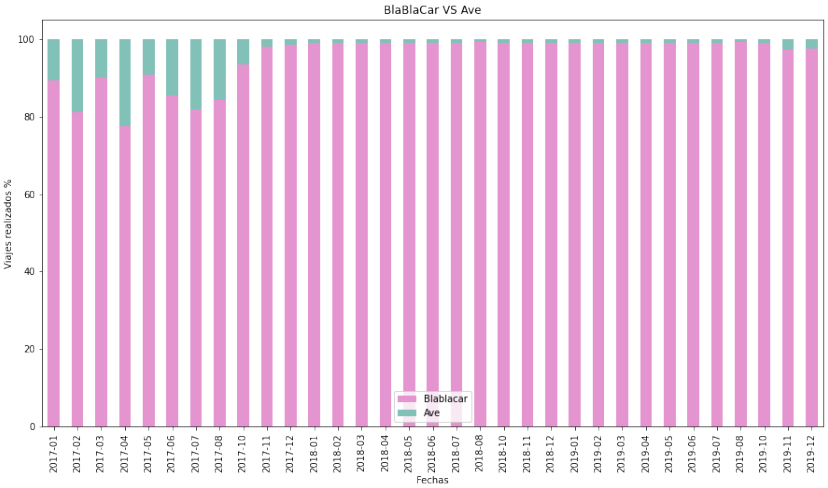
C.2 Autobus vs Blablacar


BlaBlaCar VS Total
Total
Blablacar
Numero de pasajeros
0
100000
200000
300000
400000
2017-01
2017-06
2017-12
2018-05
2018-11
2019-04
2019-10
Fecha


BlaBlaCar VS Autobús
Viajes realizados %
0
20
40
60
80
100
Blablacar
Autobús
2017-01
2017-02
2017-03
2017-04
2017-05
2017-06
2017-07
2017-08
2017-10
2017-11
2017-12
2018-01
2018-02
2018-03
2018-04
2018-05
2018-06
2018-07
2018-08
2018-10
2018-11
2018-12
2019-01
2019-02
2019-03
2019-04
2019-05
2019-06
2019-07
2019-08
2019-10
2019-11
2019-12
Fechas


BlaBlaCar VS Metro
Viajes realizados %
Fechas
Blablacar
Metro


BlaBlaCar VS Ferrocarril
Viajes realizados %
Fechas
Blablacar
Ferrocarril


BlaBlaCar VS Ave
Viajes realizados %
Fechas
Blablacar
Ave


BlaBlaCar VS Total
Viajes realizados %
Fechas
Blablacar
Total