Análisis del servicio BiciMAD en El Retiro

An approach to the public bike-sharing sytem in Madrid

Víctor Lavandeira Murat

Master en *Data Science*, KSchool

Noviembre de 2018

Introducción

Los servicios de bicicletas públicas compartidas gozan de gran popularidad en todo el mundo al resultar una solución de movilidad sostenible, económica y saludable. Además, tanto habitantes como turistas pueden darle un uso recreativo, utilizando las bicicletas para recorrer la ciudad.

En el caso de Madrid, BiciMAD presta este servicio desde el año 2014 y, a partir de abril de 2017, publica datos de libre acceso con información detallada sobre cada préstamo de bicicleta. Por cada trayecto se proporcionan datos básicos como la fecha y hora, el origen, destino y la duración del viaje. Además, estas bicicletas eléctricas equipan un dispositivo GPS que genera un track a razón de un punto cada 60 segundos aproximadamente. Desgraciadamente, en sus inicios la disponibilidad de datos GPS era muy pobre o inexistente pero la situación ha mejorado en los últimos tiempos.

Con esta premisa se ha trabajado sobre los datos de un mes completo (agosto 2018) con buena tasa de tracks GPS (un 82% de los trayectos, un total de unos 300.000, disponen de él) para acotar una zona de la ciudad y conocer los hábitos y necesidades de sus usuarios: el parque del Retiro de Madrid. Al tratarse de un lugar de esparcimiento, el principal objetivo era averiguar cuánto tráfico recreativo y cuánto tráfico de movilidad pasa dentro del parque, a qué horas y en qué días. Después de un análisis descriptivo, tanto en lo cuantitativo como sobre el mapa, se han ensayado técnicas de *clustering* para identificar estos casos de uso, agrupar los trayectos y observar los hábitos de cada grupo.

También se ha realizado un ejercicio de predicción aprovechando los datos obtenidos, si bien es cierto que la ausencia de períodos largos y variados con datos disponibles impide abordar la predicción de series temporales prolongadas, por el momento.

Datos de entrada

El principal juego de datos es el compendio de trayectos de agosto de 2018 detallados, codificados en formato JSON y de libre acceso en el portal de datos abiertos de la EMT. También me he servido del listado actualizado de estaciones de BiciMAD con sus identificadores, nombres y localizaciones. En cuanto al juego de datos principal, su estructura funcional es la siguiente:

- Identificador único del trayecto

- Código diario del usuario. Para una misma fecha, todos los trayectos de un mismo usuario están agrupados bajo este código.
- Identificador de la estación de origen.
- Identificador de la estación de destino.
- Número de anclaje del que se desengancha la bicicleta (este dato no nos aporta nada)
- Número de anclaje en el que se engancha la bicicleta (este dato no nos aporta nada)
- Franja horaria (hora en punto) en la que se realiza el préstamo de la bicicleta, en formato ISO-8601
- Duración en segundos del trayecto.
- Track GPS.
    - Cuando está disponible, se trata de una sucesión de puntos expresada en formato GeoJSON.
    - Cada feature contiene un punto con sus coordenadas y dos variables adicionales: velocidad instantánea y dirección del callejero aproximada.
- Tipo de usuario.
    - Abonado anual.
    - Ocasional.
    - Empleado de BiciMAD (mantenimiento).
- Franja de edad del usuario, de forma categórica en 6 grupos distintos.
- Código postal informado por el usuario, a menudo indisponible.

Herramientas y metodología

El proyecto se ha realizado utilizando Python tanto para la limpieza y análisis más general de los datos como para el tratamiento de los datos geoespaciales. Las técnicas de *machine learning* aplicadas también se ha realizado en este lenguaje. Todo se ha recopilado en Jupyter notebooks, suficientemente explicados, que deben estar listos para ejecutarse directamente.

Por último, además de las visualizaciones y conclusiones que se aportan en los notebooks, se ha utilizado Tableau par generar un *dashboard* y dos mapas interactivos, todo agrupado en una pequeña historia que permite visualizar los datos más relevantes.

Estos son los notebooks que se encuentran en el repositorio con su descripción más fundamental. Se deberían ejecutar de forma secuencial pues el siguiente toma como entrada un fichero generado por el anterior.

- **1-Carga y limpieza de datos.ipynb**
    - Se importan los datos iniciales, se realiza limpieza y transformaciones y se hace una exploración mínima. Los datos quedan preparados para la siguiente etapa.

- **2-Tratamiento geoespacial.ipynb**
    - Recibe como entrada el fichero *Clean_201808_Usage_Bicimad.json* generado en la etapa anterior. Si ocurre algún problema he depositado estos ficheros intermedios en la carpeta *cheats*.
    - En este extenso notebook se delimita el perímetro del Retiro y se desarrollan los mecanismos para filtrar los trayectos que han entrado al parque. Además se eliminan outliers geográficos y se crean nuevas features a partir de los datos geoespaciales, que serán de utilidad en el clustering.

- **3-Clustering.ipynb**
    - Recibe como entrada el fichero *Clean_Geo_201808_Usage_Bicimad.json* generado en la etapa anterior. Si ocurre algún problema he depositado estos ficheros intermedios en la carpeta *cheats*.
    - En este notebook nos centramos en aplicar clustering. Hemos creído obtener algo interesante utilizando métodos GMM.

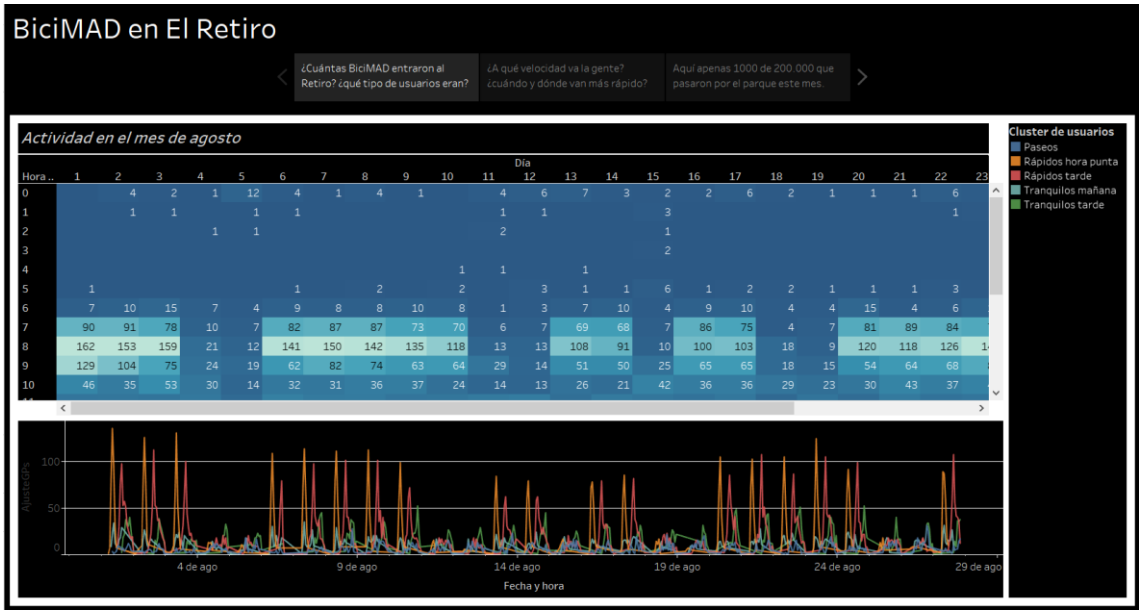
- **4-Prediccion.ipynb**
    - Recibe como entrada el fichero *Clean_Cluster_201808_Usage_Bicimad.json* generado en la etapa anterior. Si ocurre algún problema he depositado estos ficheros intermedios en la carpeta *cheats*.
    - A partir de la repartición en clusters aquí se aplican algunos modelos sencillos de regresión para predecir el tráfico de bicicletas por cada cluster.

Las conclusiones pueden encontrarse bastante detalladas en cada uno de los notebooks.

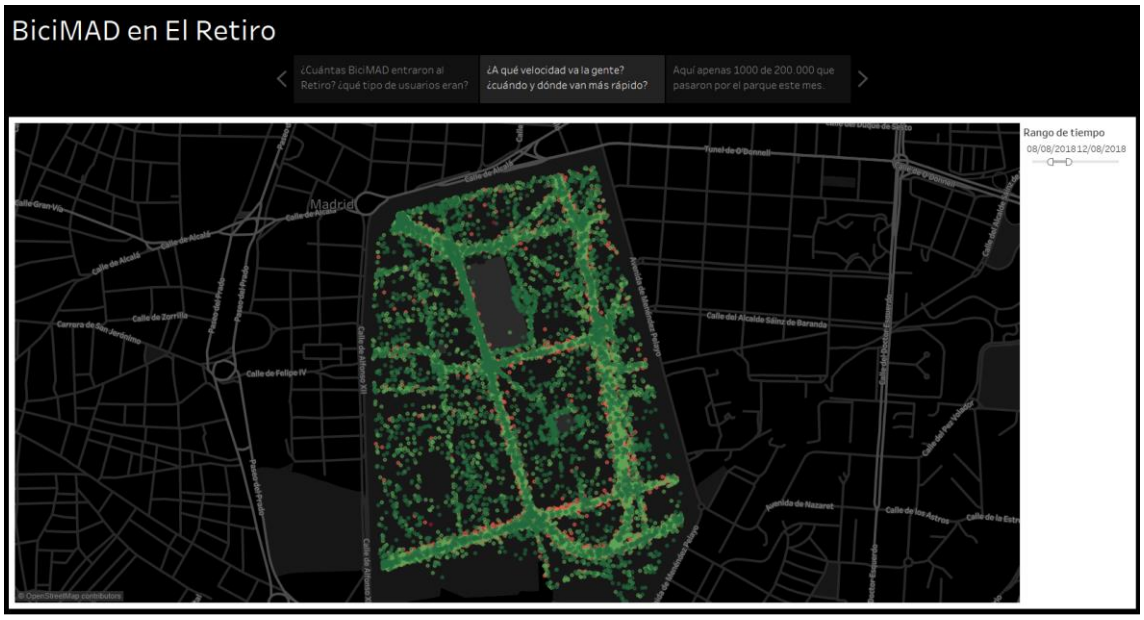
Tableau

Dashboard con dos zonas independientes: heatmap con el detalle día por día y hora por hora. Pinchando en él se puede filtrar el gráfico de abajo que detalla el número de trayecto por cada uno de los 4 clusters “horarios” que hemos definido (segunda aplicación de GMM)

En la parte derecha hay también un filtro para resaltar un cluster u otro



Mapa interactivo que contiene todos lo puntos que han caído dentro del Retiro en agosto. Se puede filtrar con el slider a la derecha para saber exactamente qué tráfico ha habido cada hora, en qué parte del parque y a qué velocidad.



Y por último otro mapa interactivo donde podemos filtrar a la derecha por 3 clusters independientes del horario (primera aplicación de GMM). En este caso, por motivos de rendimiento, se plotea una muestra de 1000 trayectos que han pasado por el Retiro. Al pinchar en uno de ellos se ve resaltado en su totalidad y la velocidad está codificada en el color.