Modelos de Predicción de Duración de Viajes BA Ecobici

Sofia Sapienza, Milagros Bochor

[1]Universidad Tecnológica Nacional, Facultad Regional Buenos Aires

Introduction

Partiendo de data sets que contienen información de cada uno de los viajes realizados utilizando el sistema de Ecobicis, y agregando a los mismos información relevante sobre el usuario o viajante, buscamos analizar algunos de los aspectos más relevantes del servicio de Ecobicis, tales como la evolución de la cantidad de viajes a lo largo del tiempo, las principales estaciones y la distribución de la duración de los viajes, entre otras. Una vez analizado el contexto de los datos, armamos un modelo que intenta predecir la duración de un viaje en Ecobici basado en las condiciones en las que se realizó el mismo.

Datasets

RECORRIDOS — 2018 | 2019 | 2020

- Estaciones origen y destino: latitud y longitud
- ID usuario
- Duración

USUARIOS

- Género
- Edad

CALENDARIO

- Dia Semana
- Dia laboral

ESTACIONES

- Barrio

Métodos

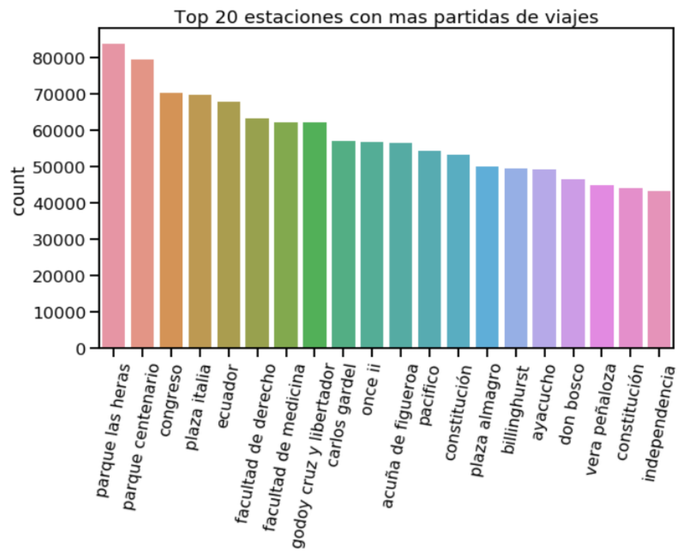
Para abordar la problemática de predicar la duración de un recorrido decidimos optar por modelos de machine learning de aprendizaje supervisado. Este tipo de aprendizaje está enfocado para problemáticas en donde se conoce a las features y a las etiquetas, pero estas últimas son continuas. En nuestro caso, la duración del recorrido es una etiqueta de carácter continuo. Utilizamos:

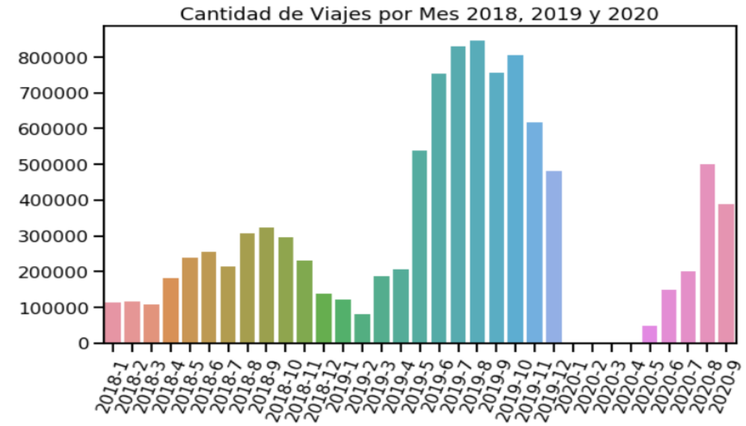
Features:

- **REGRESIÓN LINEAL** — LINEALES | POLINOMIALES
- **RIDGE REGRESSION** — LINEALES | POLINOMIALES
- **SVR** — LINEALES | POLINOMIALES

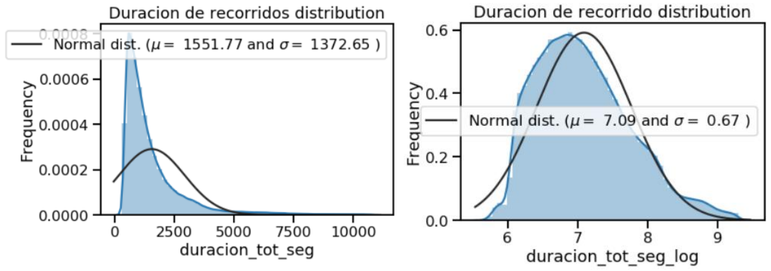
Analysis Exploratorio de datos

Son muchas estaciones! Vamos a ver cuáles forman parte del ranking #20 con más partidas.



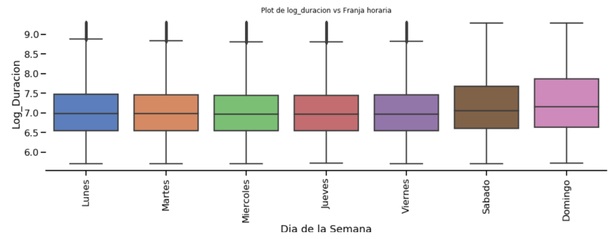


Partiendo del insight de que el data set de 2019 tenía más del doble de datos que el del 2018, nos propusimos visualizar la evolución de la cantidad de viajes a lo largo de los años



Llevamos la duración a una escala logarítmica para ajustarla a una distribución normal

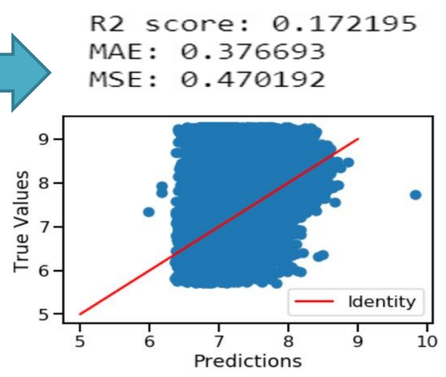
Cómo varía la duración según el día de la semana en el que se produjo el viaje?



Usamos un Boxplot. Los viajes entre semana tienen la misma mediana. Los viajes de fin de semana suelen ser más largos, sobre todo los viajes de los domingos.

Resultados

El mejor resultado que obtuvimos fue con un R2 de 0,172 utilizando una regresión lineal con features polinomiales grado 4



Conclusiones

Hemos concluido que las variables combinadas de la forma que lo hicimos no nos resuelven la problemática. Probamos con diferentes hiperparámetros para poder mejorar el ajuste de los modelos pero no encontramos una razón por la cual estos no pudieron predecir la variable continua con un valor aceptable. El máximo R2 logrado no superó el 20%, por lo que, dejamos sentadas las bases para retomar nuevamente este desafío con otras features y otras estrategias.