Demand Prediction For Bike Sharing Systems

Ahmed ElSabbagh

2022-11-26

Project Description

Introduction

- Bike Sharing Systems (BSS) have a very widely used method of transportation for many people worldwide.
- They have several advantages:
  - Non-polluting.
  - Circumvents traffic congestion.
  - Excellent for last-mile connections.
  - Have a convenient payment system.

Motivation

- Bikes are parked and rented from fixed automated bike locations.
- BSSs have a finite number of bikes and a finite number of station spaces.
- Not all stations have an equal amount of demand at any given time of the day.
- Bikes need to be relocated from low demand to high demand stations.
- Companies need to have estimate the demand and available bikes at each station.

Objectives

The objective is to create multivariate regression models capable of estimating the following

- **Demand prediction:** Number of bikes needed at each individual station.
- **Availability prediction:** Number of bikes available at each individual station.

Literature Review

Data Description

Lyft/Baywheels

The data was provided by Lyft, the owner of the BSS company Baywheels (formally GoBike) operating in San Francisco Bay Area.

- The data is available as far as 2017 before Lyft's acquisition of GoBike.
- The data used however is limited to 2021 and 2022 because:
  - The disruption caused 2020 COVID pandemic.
  - Potential business growth gap between 2019 and 2021.

Weather Data

- Weather data is very relevant in determining the demand.
- The weather data for San Francisco Bay Area is provided by Meteostat through a Python API.

Dataset

Size:

- **Original Number of Trips:** 4053524 trips .
- **Final Number of Trips:** 3463728 trips. (subject to change)
- **Number Of Stations**: 532
- **Period:** From January 2021 to September (as far as when I am finished) 2022 (2023?)

Relevant Columns:

- Station Name (start/end)
- Time (start/end)
- Coordinates (start/end)
- Ride ID

Cleaning Process

Cleaning Process

- **Station Standardization:**
  - Stations in a particular street don't always have the same coordinates.
  - Many stations don't have a standardized name or ID, making them unidentifiable.
  - **Solution To Reduce Data Loss**:
    - Use a single coordinate for any identifiable stations.
    - Approximate the closest standard station to any trip.
    - If the closest station is less than 500 meters away, keep the trip, otherwise, drop.

Cleaning Process

- **Same Station Trip:**
  - Several trips usually take place in several minutes, with the end station being the same as start stations.
  - This could be a result of users trying out the system or changed minds.
  - To prevent redundancy, any same station trip with duration less than 4 minutes will be removed.

Cleaning Process

- **Clustering**:
  - Helps reduce weather data size by:
    - Approximating areas closest to each other.
    - Approximating the weather conditions for each cluster.

Cleaning Process

Demand and Availability

Definitions

- Demand: Number of trips starting at a particular station.
- Availability: Number of trips ending at a particular station.
- "Availability" does not indicate actual number of bikes available, but it indicates how many bikes finished their trips there, which when compared with demand should provide a good idea about the available bikes.
- Therefore, demand will be the main focus, while availability will be taken as secondary.

Demand

Demand and Availability

Demand and Availability

Demand Spread

Demand Spread

Demand Spread