Bike Share System in SF Bay Area

Author: Owen Hsu

Project Overview

- **The Problem Area:**
  - How to enhance the efficiency and accessibility of the bike share system in the SF Bay Area?
- **The Users:**
  - Commuters, tourists, and residents.

The Big Idea

- **Machine Learning Model:**
  - To predict usage patterns, so that we can optimize bike availability through strategic redistribution, ensuring they are accessible where and when they are most needed.

- **Approach**
  - Demand prediction by analyze historical bike usage data
  - User behavior analysis, such as preferred routes and times of usage, to optimize bike availability and placement.

Potential Impact

- Better user experience and increased usage by 10%
- Environmental benefits such as the reduction in CO2 emissions by 5%
- Cost savings on transportation by 5% (fuel and parking fees)
- Positive effects on public health by promoting physical activity

Dataset Overview

- **Original Datasets:**
  - Station, Status, and Weather datasets
  - 71,984,434 rows; 7 columns, 4 columns, and 24 columns respectively
- **Merged, Filtered, and Cleaned Dataset:**
  - 7,337,194 rows; 15 columns (13 numerical and 2 datetime)
  - Station ID, Date/Time, Available Bikes, Available Docks, Total Docks, and Weather Info
- **Data Quality:**
  - Preliminary EDA:
    - Data Distributions and Detecting Collinearity
  - Dropped 2 columns with more than 20% missing values in Weather dataset

Feature Processing and Baseline Modeling

- **Data Processing:**
  - Handling Date / Time: Extract features such as day of the week, month, or hour
  - Holiday Features: Create binary features indicating whether a day is a holiday or not.
  - Geographical differences and weather features.
  - Data Transformation: Convert categorical variables into numerical ones
  - Feature Scaling (if needed)
- **Baseline Modeling:**
  - Define Target Variable
  - Logistic Regression
  - Model Evaluation: Train the model with the training set and evaluate the performance with the testing set using metrics such as R-squared.

Thank you