CMPE 255-Team 5 Project Proposal

Team : Team 5

Project Title : SF bay area bike share analytics and predict

Team Information:

Team member Name	SJSU ID	Email ID
Jie Liu	015331121	jie.liu01@sjsu.edu
Xialu Zou	011353316	xialu.zou@sjsu.edu
Lingxiang Hu	015230631	lingxiang.hu@sjsu.edu
Chaoran Lei	015264119	chaoran.lei@sjsu.edu

Dataset : SF Bay Area Bike Share
Anonymized bike trip data from August 2013 to August 2015

Source: https://www.kaggle.com/benhamner/sf-bay-area-bike-share

(As per your suggestion will use a subset of the data with total values greater than 10M)

Dataset size	The size of the dataset is 2.07GB which consists of four input files- station.csv, status.csv, trips.csv and weather.csv.
Dataset type	Text data, numerical data

Input Files	No. of columns	No. of Rows	Features
station.csv	7	70(different stations)	Contains data that represents a station where users can pick up or return bikes.
status.csv	4	1047143(time stamps)	Data about the number of bikes and docks available for given station and minute
trips.csv	11	361559(different trips)	Data about individual bike trips

weather.csv	24	733 (dates)	Data about the weather on a specific day for certain zip codes

Project description:
This project is focused on an exploratory analysis of SF Bay Area Bike, the data was collected from 2013 to 2018. Relying on the analysis of data, we will explore how the weather impacts bike trips and how the bike trip patterns vary by time of day and the day of the week. Furthermore, we will dig into the data to see what's the relationship between bike stations, subscribers and dates.

Through the above research, we will try to predict the possible number of trips in the future in certain weather conditions in the SF bay area. From this prediction, we may get some business insight such as how to increase bike trips on weekends; how to attract more subscribers for bike trip service; how to improve the services or how to target specific locations of customers.

In this project, we will use python to implement the whole process. Packages including numpy, scipy, pandas, matplotlib, sklearn, xgboost, etc. for data dimensionality reduction, clustering, classification, regression and plot analysis will be applied.

Methodology:
Preprocessing: Data Cleaning, Removing null values, Filtering unnecessary columns.
Technique: Linear Regression (LR), Random Forest Model (RF), Decision Tree Classifier, XGBoost
Optimization: Perform if required
Evaluation Metrics: Accuracy, Precision, Recall, F1 Score