Bay Area Bike Prediction

Team Members:

1. Anshul Shandilya
2. Girish Bisane
3. Parth Saraiya
4. Stavan Patel

Github Link: stavan30/cmpe255_final_project (github.com)

Introduction

- The goal for this project is to use machine learning approaches to forecast bike demand for certain cities.

- What we have:
  - We have previous data for bike trips
  - The weather for dates of all the trips
  - Data about stations and cities

- What do we do with it?
  - Use ML and Data mining techniques to use the obtained data, manipulate it and predict the number of trips for a specified day.

Dataset used

Bay area bike share dataset was used (source) - from Kaggle

About Bay Area Bike Share:

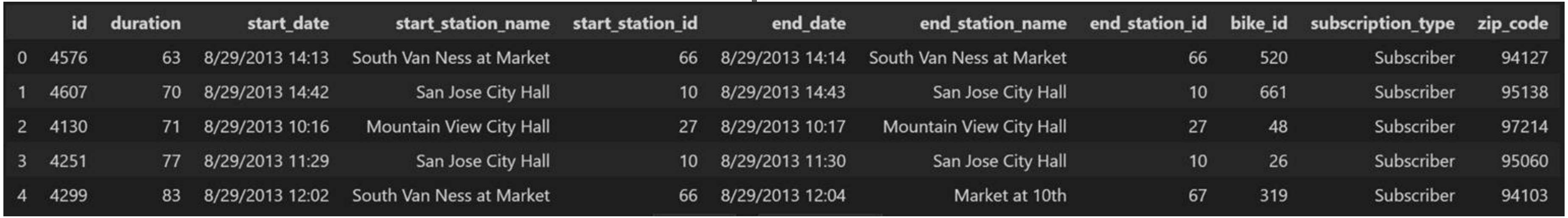
- The Bay Area bike Share enables quick, easy, and affordable bike trips around the San Francisco Bay Area.
- They also make regular open releases of the dataset we used

Dataset used

Individual csv files in the dataset:

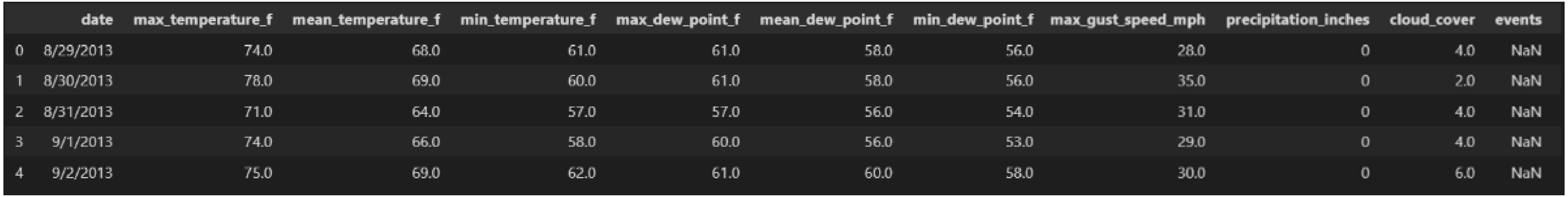
- station.csv - Contains data the represents a station where users can pickup or return bikes
- Status.csv (**not used for our problem**)
- trips.csv - Data about individual bike trips
- weather.csv - Data about the weather on a specific day for certain zip codes.

trips.csv



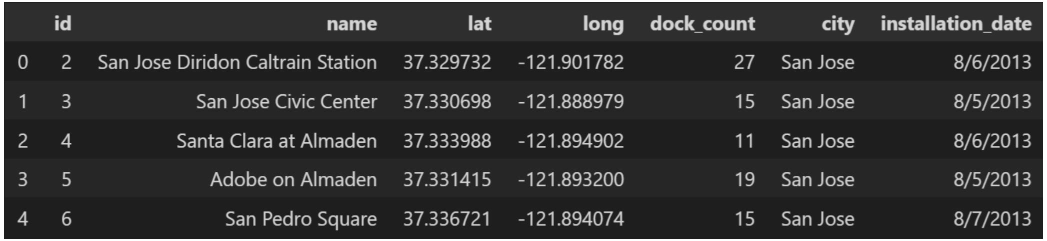
	id	duration	start_date	start_station_name	start_station_id	end_date	end_station_name	end_station_id	bike_id	subscription_type	zip_code
0	4576	63	8/29/2013 14:13	South Van Ness at Market	66	8/29/2013 14:14	South Van Ness at Market	66	520	Subscriber	94127
1	4607	70	8/29/2013 14:42	San Jose City Hall	10	8/29/2013 14:43	San Jose City Hall	10	661	Subscriber	95138
2	4130	71	8/29/2013 10:16	Mountain View City Hall	27	8/29/2013 10:17	Mountain View City Hall	27	48	Subscriber	97214
3	4251	77	8/29/2013 11:29	San Jose City Hall	10	8/29/2013 11:30	San Jose City Hall	10	26	Subscriber	95060
4	4299	83	8/29/2013 12:02	South Van Ness at Market	66	8/29/2013 12:04	Market at 10th	67	319	Subscriber	94103

weather.csv



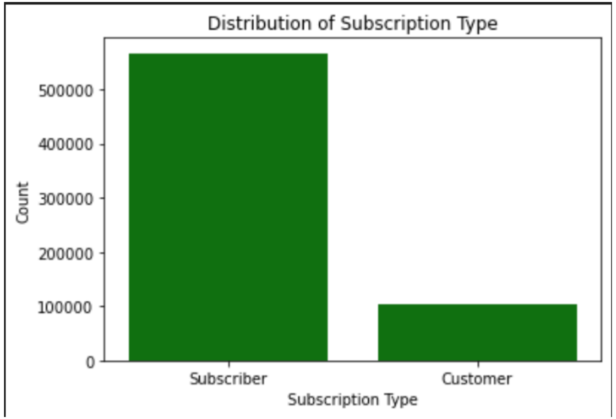
	date	max_temperature_f	mean_temperature_f	min_temperature_f	max_dew_point_f	mean_dew_point_f	min_dew_point_f	max_gust_speed_mph	precipitation_inches	cloud_cover	events
0	8/29/2013	74.0	68.0	61.0	61.0	58.0	56.0	28.0	0	4.0	NaN
1	8/30/2013	78.0	69.0	60.0	61.0	58.0	56.0	35.0	0	2.0	NaN
2	8/31/2013	71.0	64.0	57.0	57.0	56.0	54.0	31.0	0	4.0	NaN
3	9/1/2013	74.0	66.0	58.0	60.0	56.0	53.0	29.0	0	4.0	NaN
4	9/2/2013	75.0	69.0	62.0	61.0	60.0	58.0	30.0	0	6.0	NaN

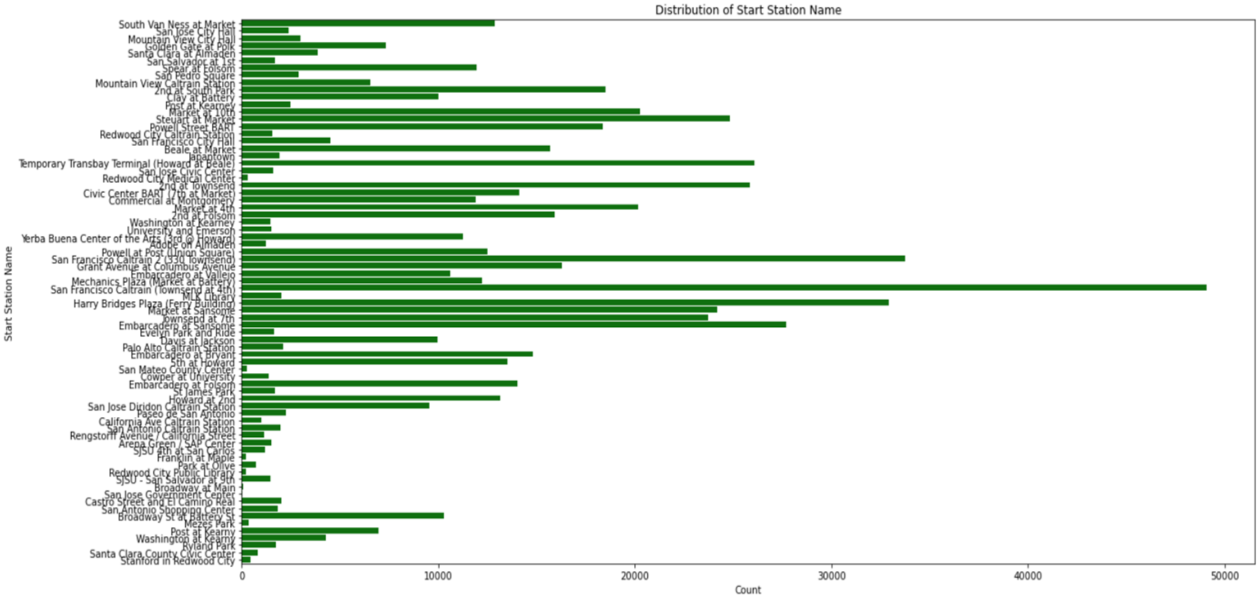
station.csv



	id	name	lat	long	dock_count	city	installation_date
0	2	San Jose Diridon Caltrain Station	37.329732	-121.901782	27	San Jose	8/6/2013
1	3	San Jose Civic Center	37.330698	-121.888979	15	San Jose	8/5/2013
2	4	Santa Clara at Almaden	37.333988	-121.894902	11	San Jose	8/6/2013
3	5	Adobe on Almaden	37.331415	-121.893200	19	San Jose	8/5/2013
4	6	San Pedro Square	37.336721	-121.894074	15	San Jose	8/7/2013

Visualizing the Data





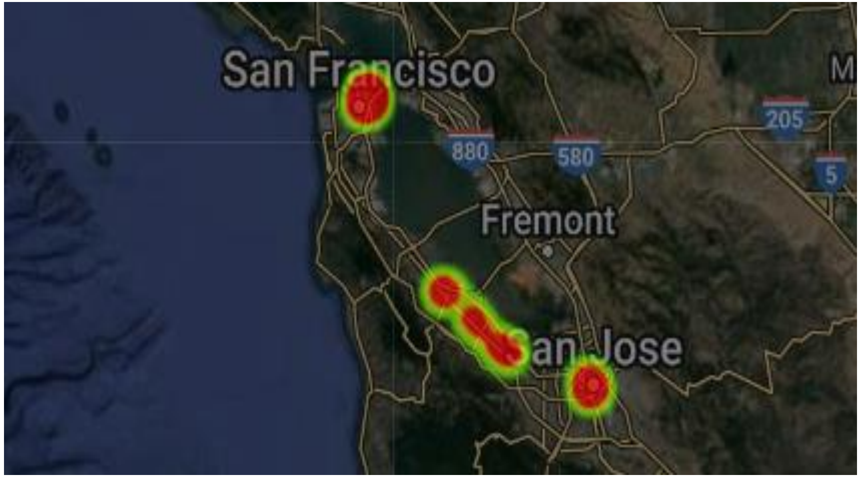
Distribution of subscription type tells us:

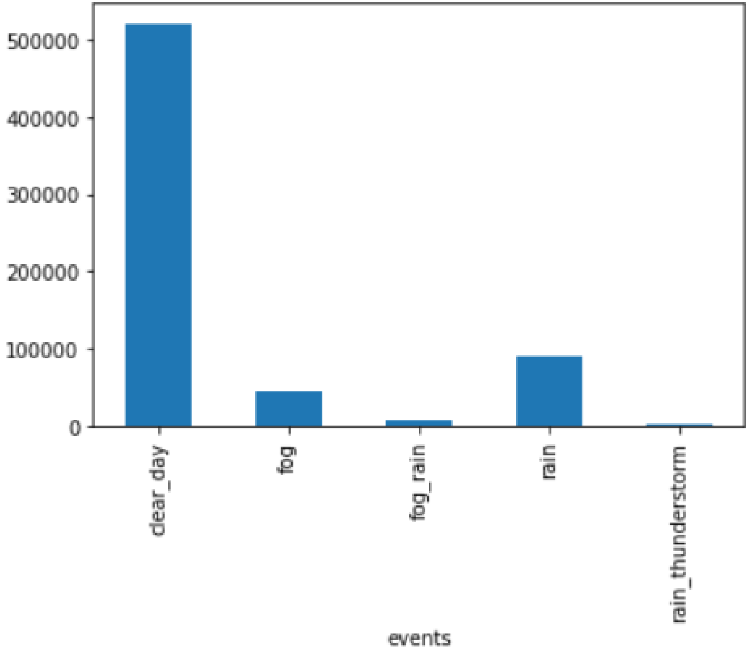
- Most bike rides are from subscribers

Distribution of Stations from where trips started tells us:

- Most Trips are from San Francisco
- Least Trips are from San Jose Govt. Center

Visualizing the Data





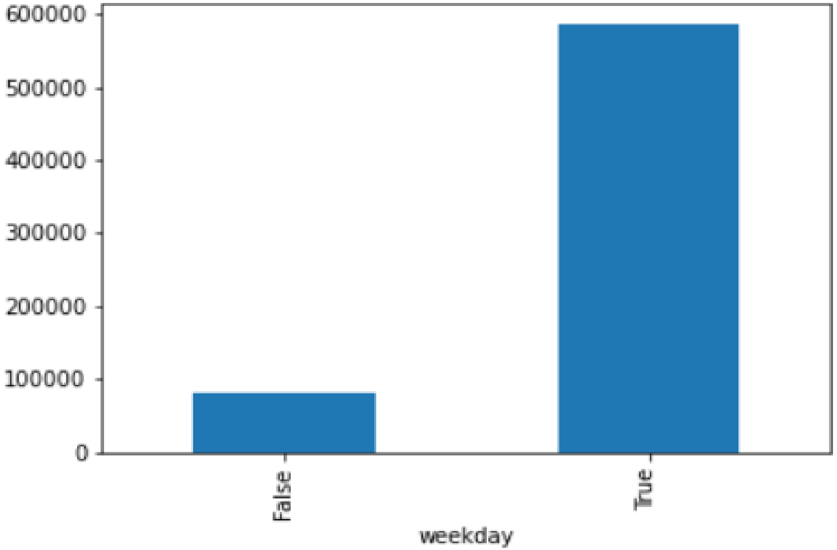
Heatmaps of the stations available

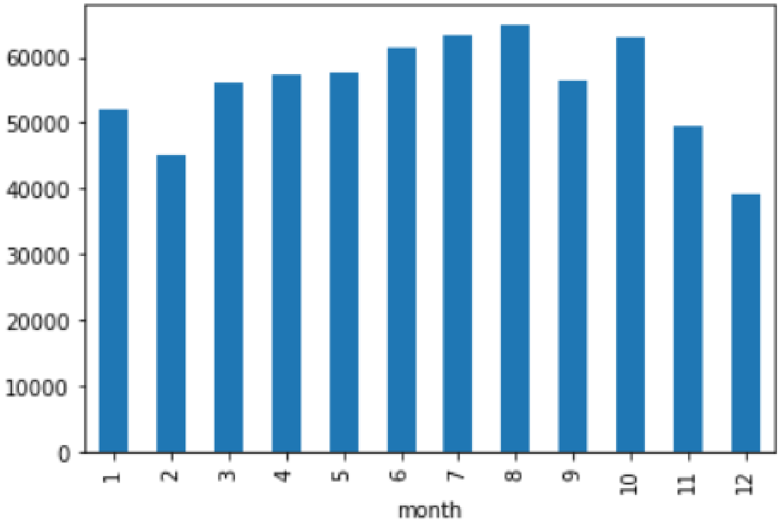
- Tells us the frequency of stations per city

Distribution of trips on types of days

- You obviously wouldn’t bike in a thunderstorm

Visualizing the Data





Distribution of trip counts for weekend or weekday

- This tells us that people might be using bikes for work or errands.

Distribution of trips on months

- This tells us there is almost a consistent demand all year round.

Data Preparation for training

Performed **Feature selection** and **Creation** to create the training data.

Trips.csv and station.csv

- Selected ‘date’ and ‘trip count’ for each date along with ‘station ID’ and in from station.csv
- Determined if ‘date’ was weekend, business day, weekday or a holiday and added as features
- Labeled months for ‘date’ and added as feature
- Created 5 different df’s, for each city

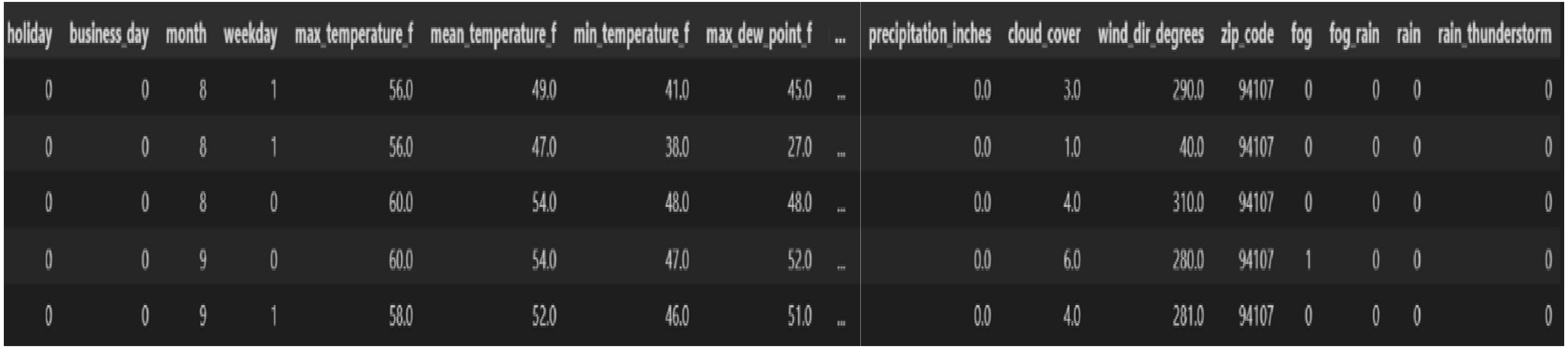
Weather.csv

- Weather given for **San Francisco, Mountain View, Palo Alto, San Jose and Redwood** (all extracted into individual df’s.)
- Handled NaN values in ‘max_gust_speed’ for all cities
- Selected all features for this file
- Found multiple non-float and non-int values, converted them

Making the Training Dataset

Making the training dataset

- Joined trips.csv (count for each day) and weather.csv on ‘date’ column for each city to obtain correlated data containing weather and day information for each date a trip was taken
- All columns not a float or int were dropped



holiday	business_day	month	weekday	max_temperature_f	mean_temperature_f	min_temperature_f	max_dew_point_f	...	precipitation_inches	cloud_cover	wind_dir_degrees	zip_code	fog	fog_rain	rain	rain_thunderstorm
0	0	8	1	56.0	49.0	41.0	45.0	...	0.0	3.0	290.0	94107	0	0	0	0
0	0	8	1	56.0	47.0	38.0	27.0	...	0.0	1.0	40.0	94107	0	0	0	0
0	0	8	0	60.0	54.0	48.0	48.0	...	0.0	4.0	310.0	94107	0	0	0	0
0	0	9	0	60.0	54.0	47.0	52.0	...	0.0	6.0	280.0	94107	1	0	0	0
0	0	9	1	58.0	52.0	46.0	51.0	...	0.0	4.0	281.0	94107	0	0	0	0

Pre-processing and Algorithm selection

Pre-processing:

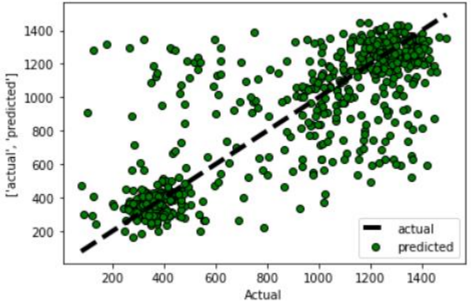
- Data was scaled using Min-Max Scalar

Algorithms Tried on the dataset:

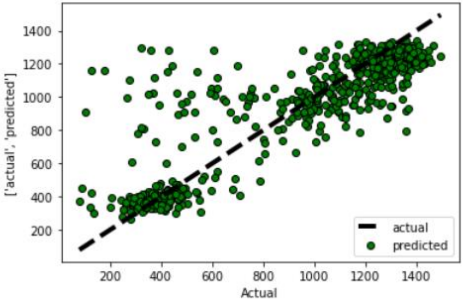
- Random Forest Regression
- ExtraTrees Regression
- XGBoost
- KNN
- Lasso Linear Regression
- Gradient Boosting

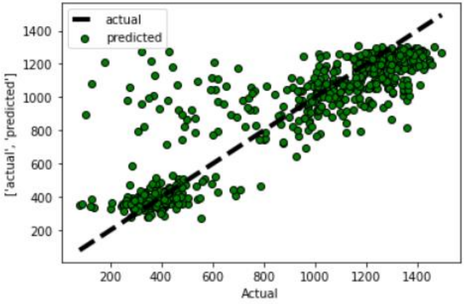
Model training

Random Forest Regressor

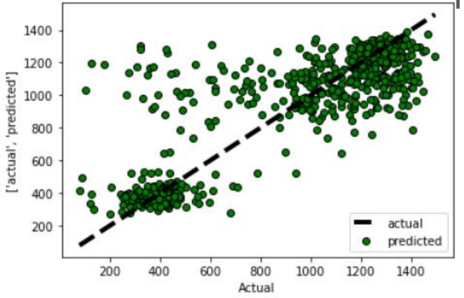


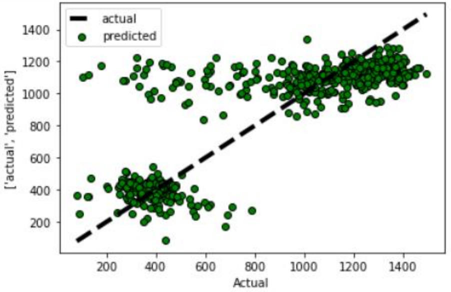
Extra Trees



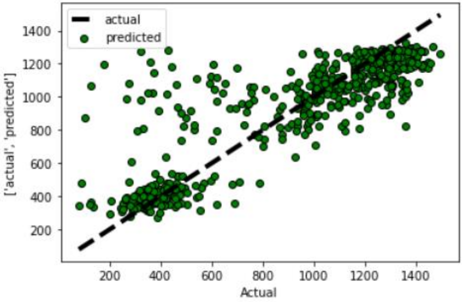


KNN

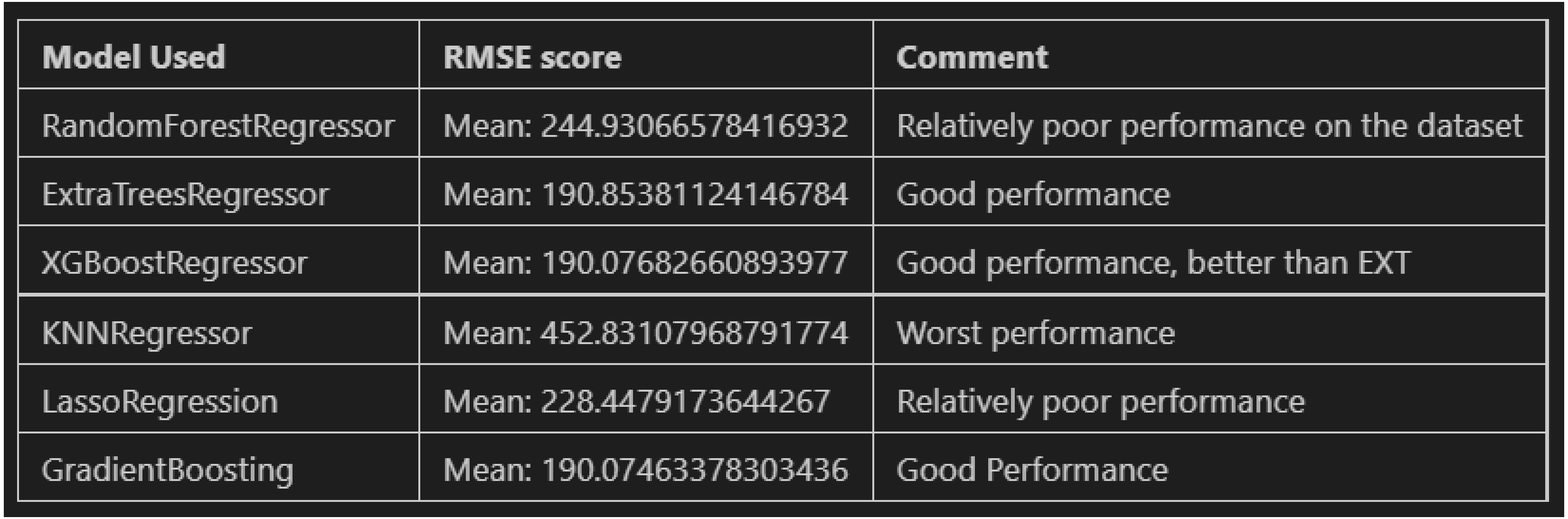




Lasso



Comparisons of all Models

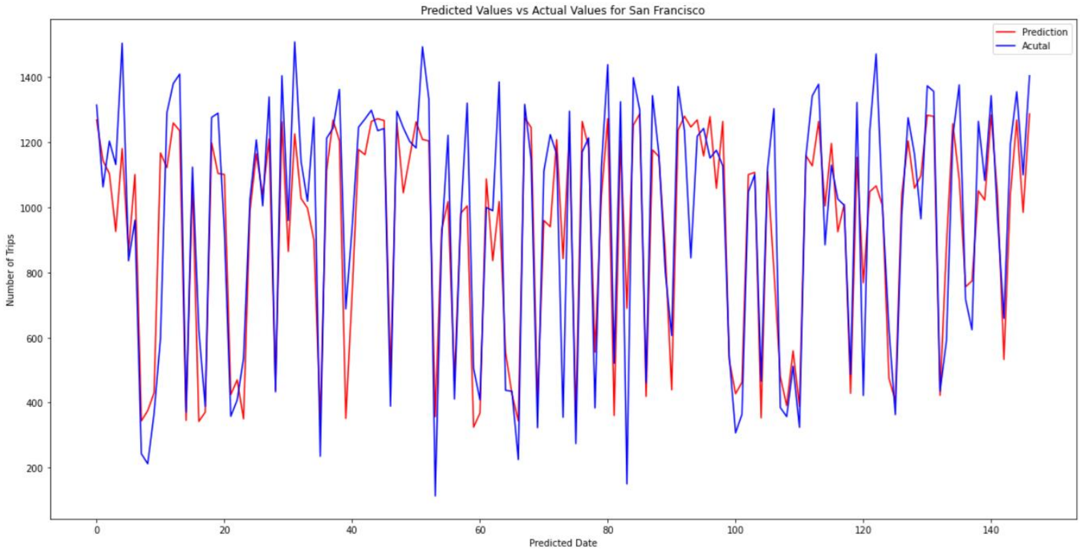


Model Used	RMSE score	Comment
RandomForestRegressor	Mean: 244.93066578416932	Relatively poor performance on the dataset
ExtraTreesRegressor	Mean: 190.85381124146784	Good performance
XGBoostRegressor	Mean: 190.07682660893977	Good performance, better than EXT
KNNRegressor	Mean: 452.83107968791774	Worst performance
LassoRegression	Mean: 228.4479173644267	Relatively poor performance
GradientBoosting	Mean: 190.07463378303436	Good Performance

Selected Gradient Boosting Regression for our problem with n_estimators=50 (param tuning)

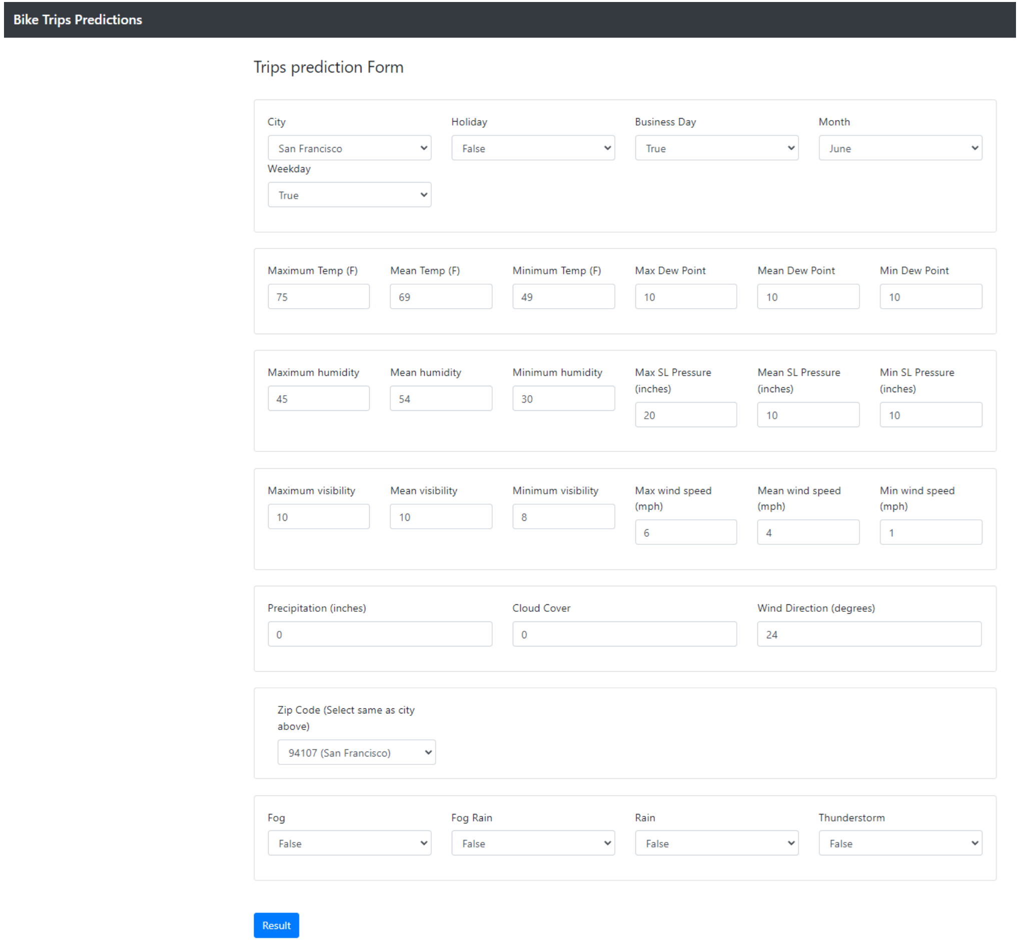
Results

- Selected Gradient Boosting Regressor for our prediction problem.
- Created separate models for each city (**San Francisco, Mountain View, San Jose and Redwood**)



Application

Created web application for predictions.



Bike Trips Predictions

Trips prediction Form

City	Holiday	Business Day	Month
San Francisco	False	True	June
Weekday			
True			

Maximum Temp (F)	Mean Temp (F)	Minimum Temp (F)	Max Dew Point	Mean Dew Point	Min Dew Point
75	69	49	10	10	10

Maximum humidity	Mean humidity	Minimum humidity	Max SL Pressure (inches)	Mean SL Pressure (inches)	Min SL Pressure (inches)
45	54	30	20	10	10

Maximum visibility	Mean visibility	Minimum visibility	Max wind speed (mph)	Mean wind speed (mph)	Min wind speed (mph)
10	10	8	6	4	1

Precipitation (inches)	Cloud Cover	Wind Direction (degrees)
0	0	24

Zip Code (Select same as city above)

94107 (San Francisco)

Fog	Fog Rain	Rain	Thunderstorm
False	False	False	False

Thank You.