Forecasting Trips of San Francisco's Bike Share Program

Dec 08, 2018

Team K

PROJECT OVERVIEW

Challenge:

How many trips can be expected to start from any given station?

Does the station require rebalancing of bikes to meet the demand?

Goal: Forecast with reasonable accuracy, the number of trips that will be made from a station at any given date and hour.

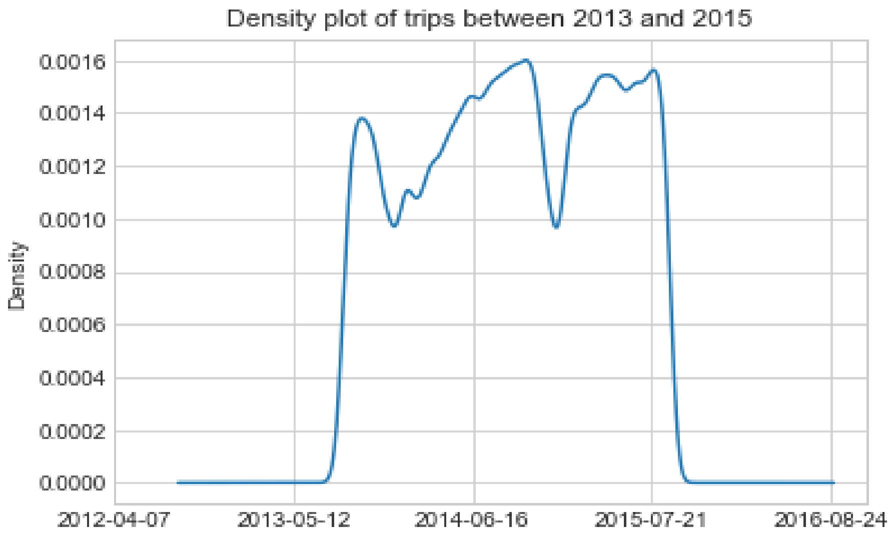
With the forecast, we can then evaluate whether the station will require rebalancing.

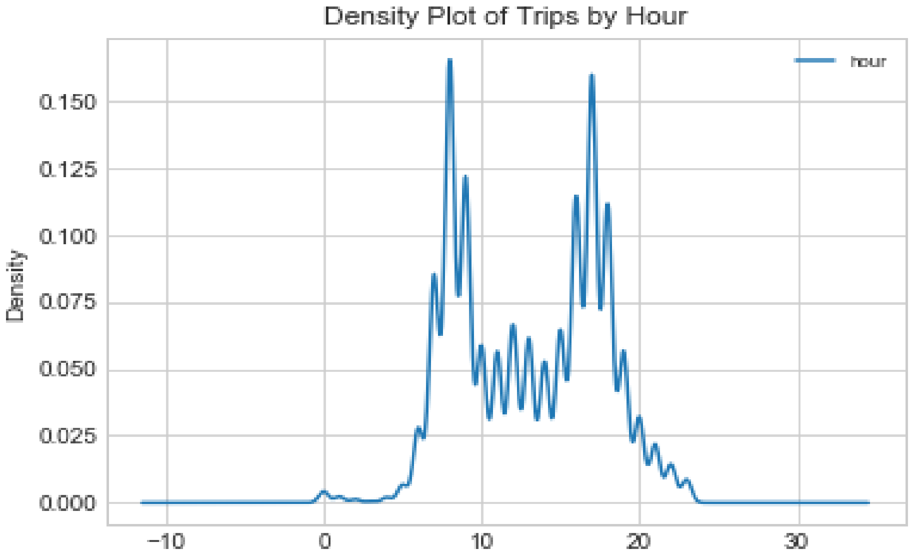
DATA OVERVIEW

We'll use the publicly available data of the Bay Area bike share (source):

- **trips:** trip-level records, includes date, start/end time, start/end station ID, start/end station name, bike ID, rider subscription type, and trip duration
- **station:** metadata for each station (n = 70) Contains data that represents a station where users can pickup or return bikes.
- **status:** minute-by-minute update of the number of bikes and number of docks available for each of the 70 stations
- **weather:** zip-level daily weather patterns for the SF Bay Area

Distinct Pattern Observed in Number of Trips Taken



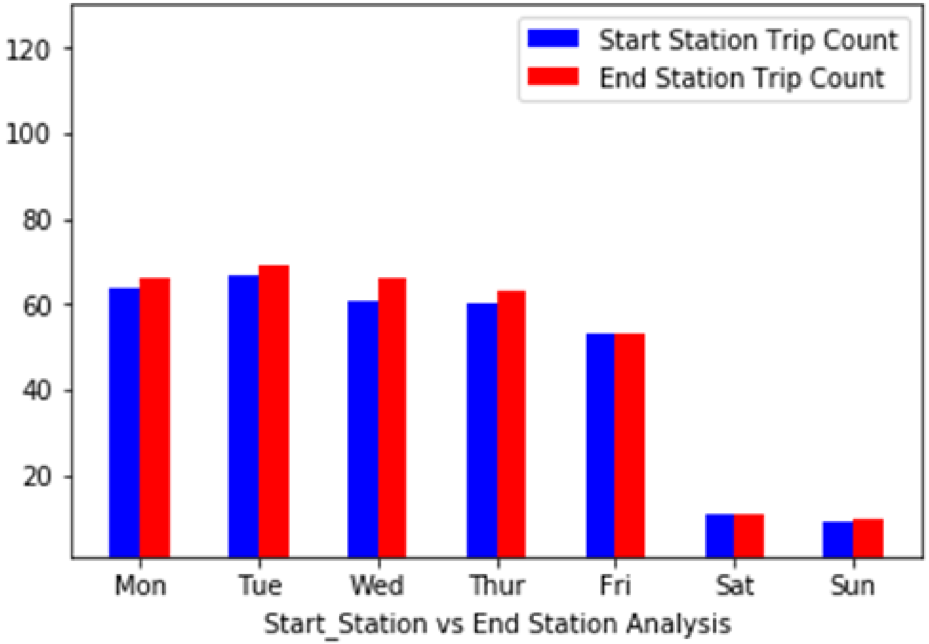


Finding: Busiest hours are from 7am to 10am and 5pm to 8pm. These hours represent almost 70% of the trips taken during the day

Need for Rebalancing of Bikes

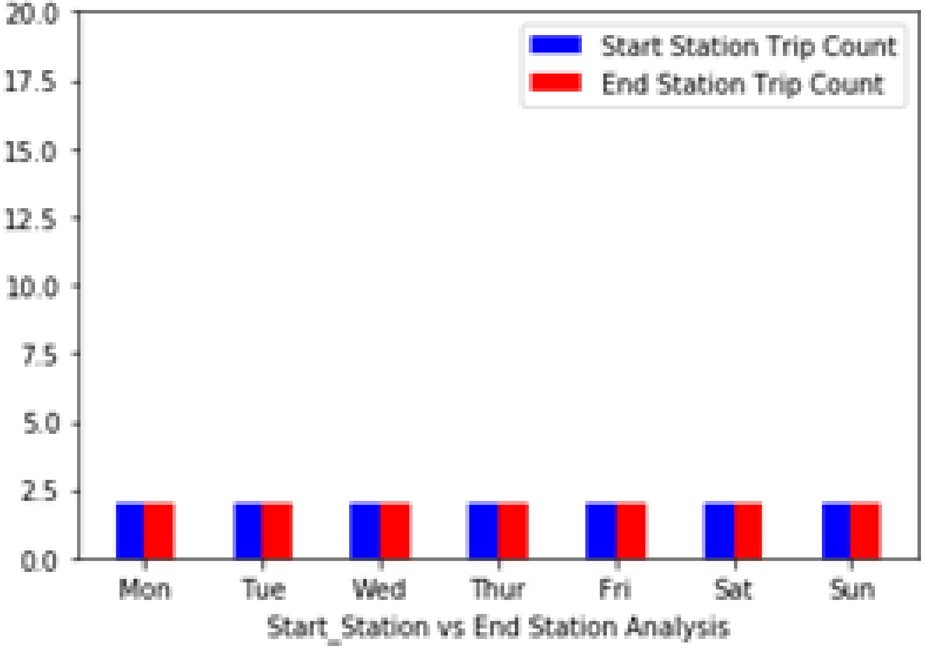
Station No:69

SF 330 at Townsend



Station No 16

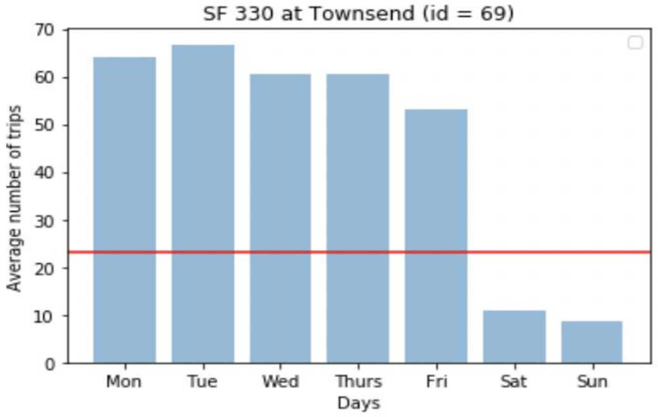
SJSU San Savlador at 9th CalTrain

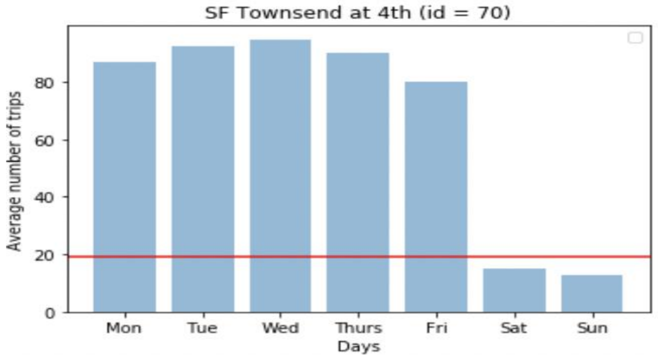


Finding: At busy stations, the number of trips ending are higher than trips starting. This suggests that there is a need for rebalancing of bikes at busy stations.

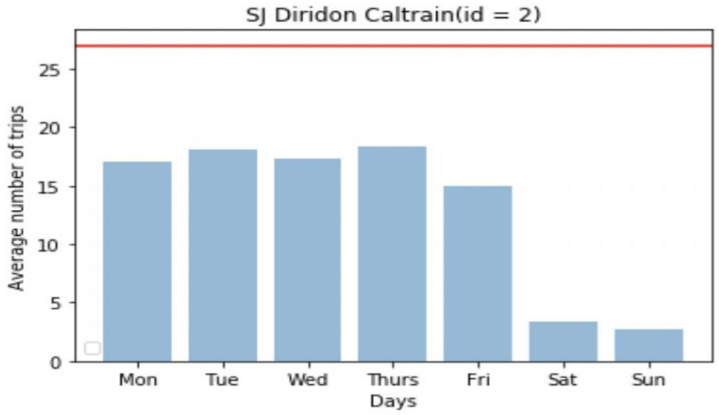
Dock Capacity is Not Optimized

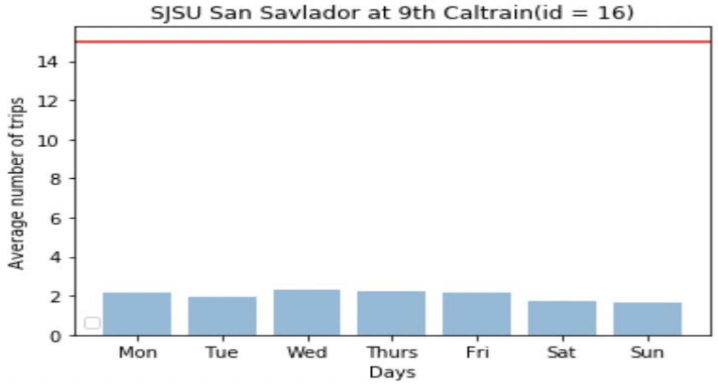
Busy Stations





Sparse Stations

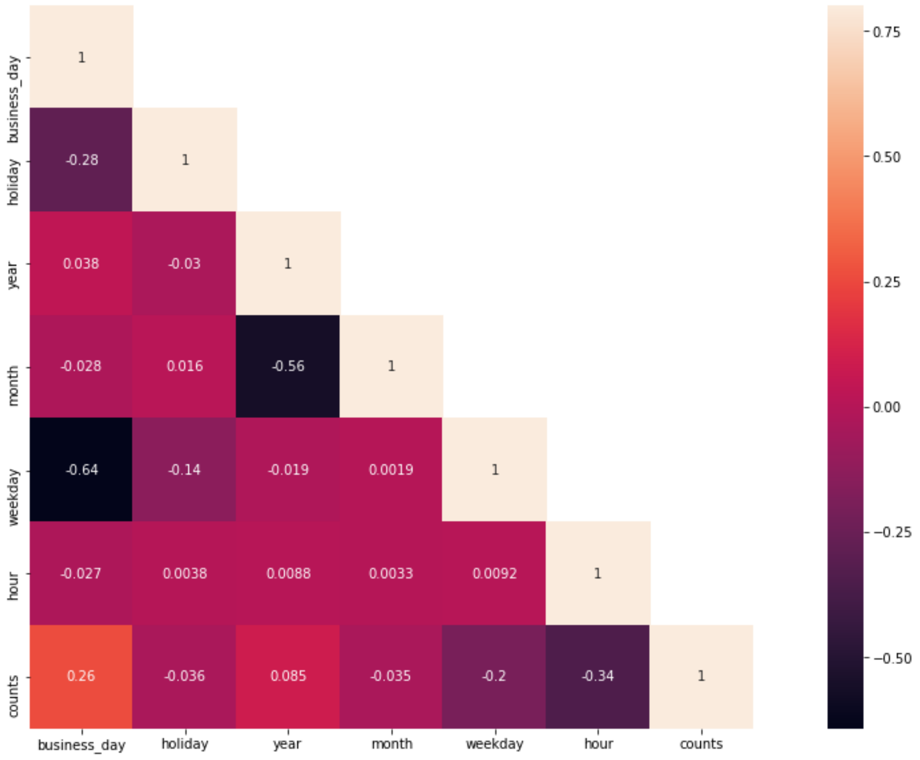


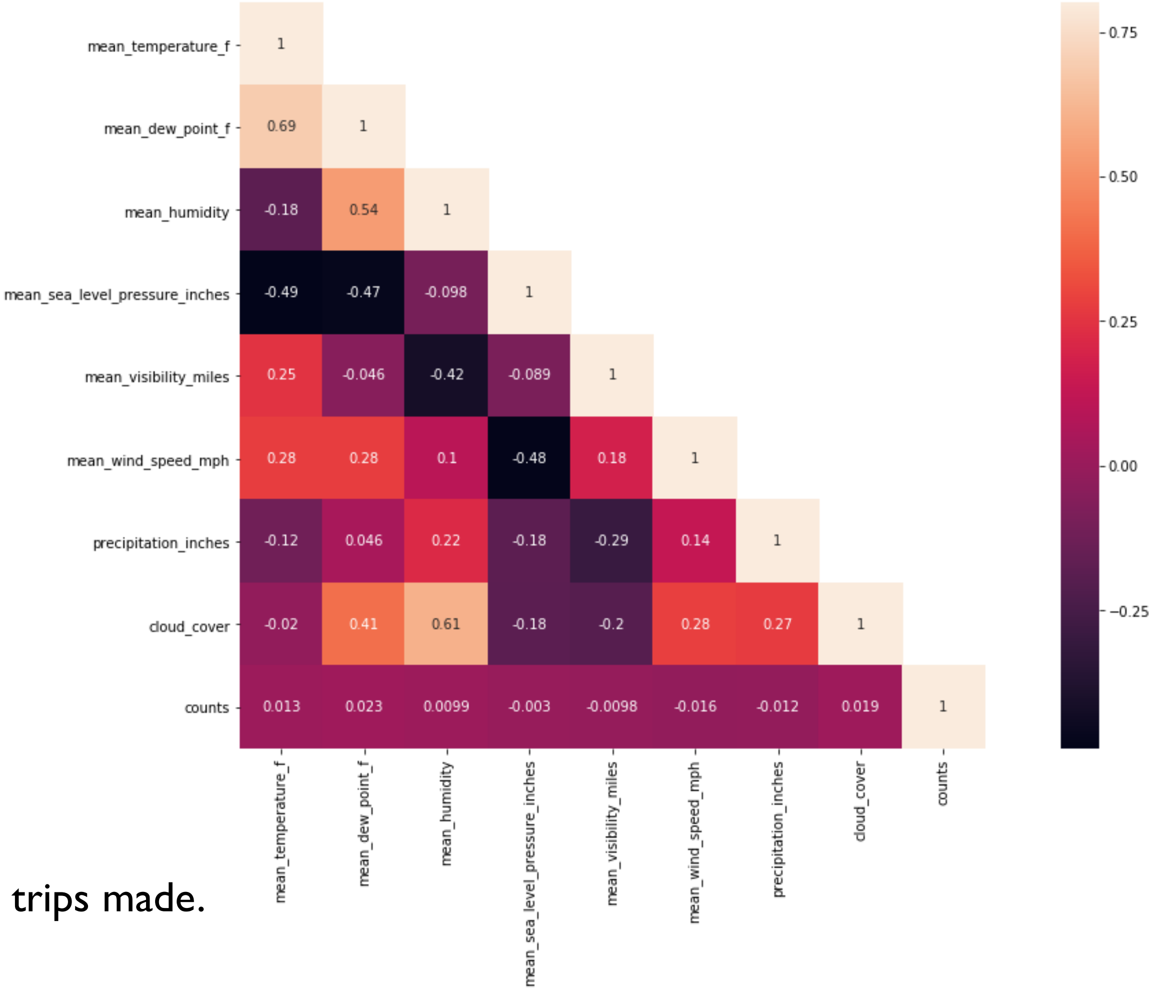


Finding:

- Sparse stations are taking up unnecessary space with their capacity
- The dock capacity at busy stations could be increased to meet the latent demand
- This confirms our hypothesis that the business does not know how many trips can be expected

Aggregated Weather Data Does Not Influence Trips





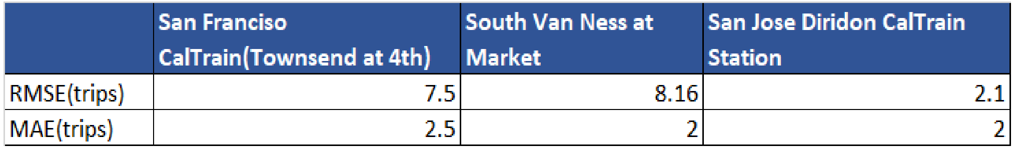
Finding: Attributes of the day have a strong correlation with the trips made.

Specifically:

- Whether it’s a working day
- Hour of the day
- Day of the week

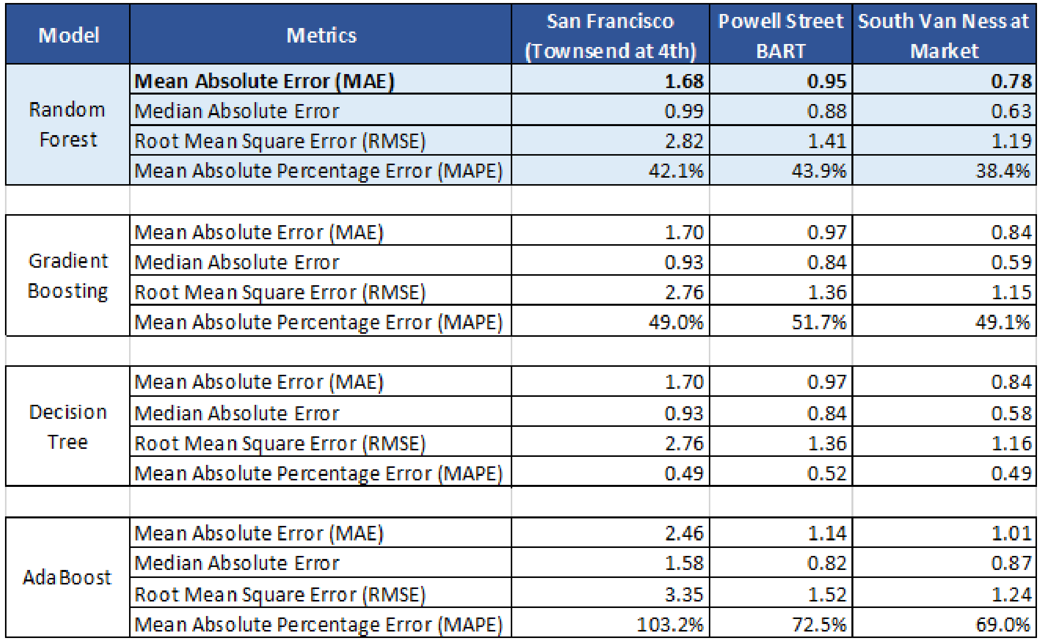
Random Forest Regression Minimized our Error in Prediction

Default Model Output



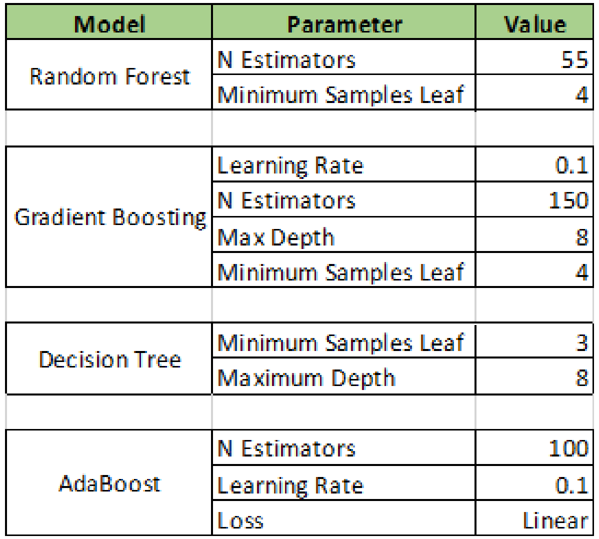
	San Franciso CalTrain(Townsend at 4th)	South Van Ness at Market	San Jose Diridon CalTrain Station
RMSE(trips)	7.5	8.16	2.1
MAE(trips)	2.5	2	2

Regression Model Output



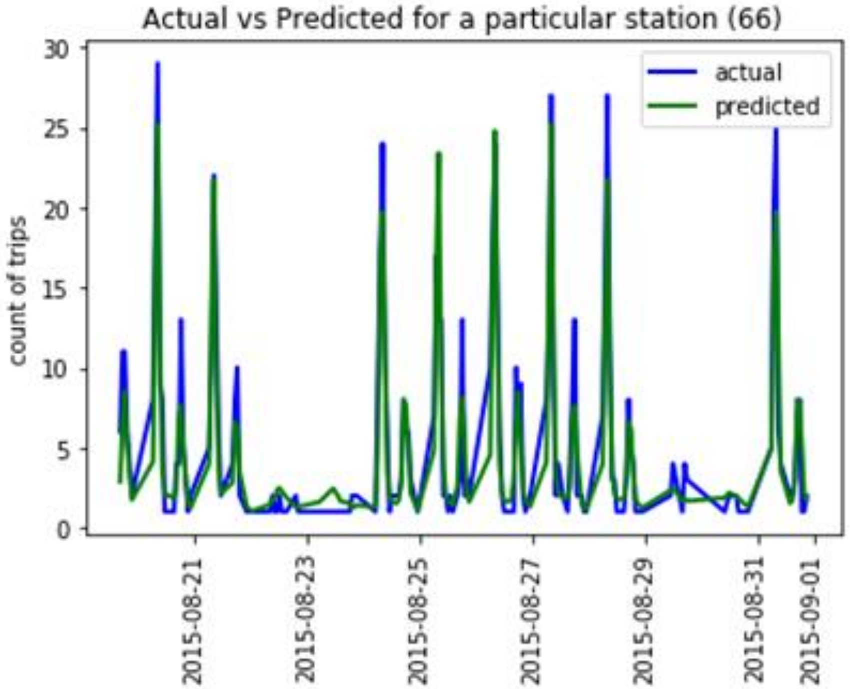
Model	Metrics	San Francisco (Townsend at 4th)	Powell Street BART	South Van Ness at Market
Random Forest	**Mean Absolute Error (MAE)**	**1.68**	**0.95**	**0.78**
	Median Absolute Error	0.99	0.88	0.63
	Root Mean Square Error (RMSE)	2.82	1.41	1.19
	Mean Absolute Percentage Error (MAPE)	42.1%	43.9%	38.4%
Gradient Boosting	Mean Absolute Error (MAE)	1.70	0.97	0.84
	Median Absolute Error	0.93	0.84	0.59
	Root Mean Square Error (RMSE)	2.76	1.36	1.15
	Mean Absolute Percentage Error (MAPE)	49.0%	51.7%	49.1%
Decision Tree	Mean Absolute Error (MAE)	1.70	0.97	0.84
	Median Absolute Error	0.93	0.84	0.58
	Root Mean Square Error (RMSE)	2.76	1.36	1.16
	Mean Absolute Percentage Error (MAPE)	0.49	0.52	0.49
AdaBoost	Mean Absolute Error (MAE)	2.46	1.14	1.01
	Median Absolute Error	1.58	0.82	0.87
	Root Mean Square Error (RMSE)	3.35	1.52	1.24
	Mean Absolute Percentage Error (MAPE)	103.2%	72.5%	69.0%

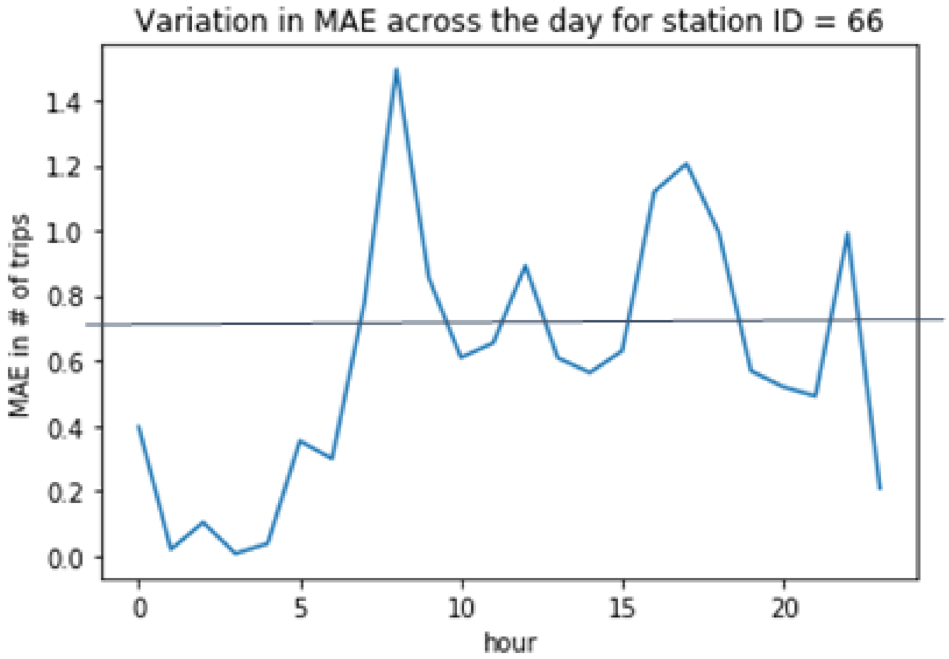
Regression Model Parameters



Model	Parameter	Value
Random Forest	N Estimators	55
	Minimum Samples Leaf	4
Gradient Boosting	Learning Rate	0.1
	N Estimators	150
	Max Depth	8
	Minimum Samples Leaf	4
Decision Tree	Minimum Samples Leaf	3
	Maximum Depth	8
AdaBoost	N Estimators	100
	Learning Rate	0.1
	Loss	Linear

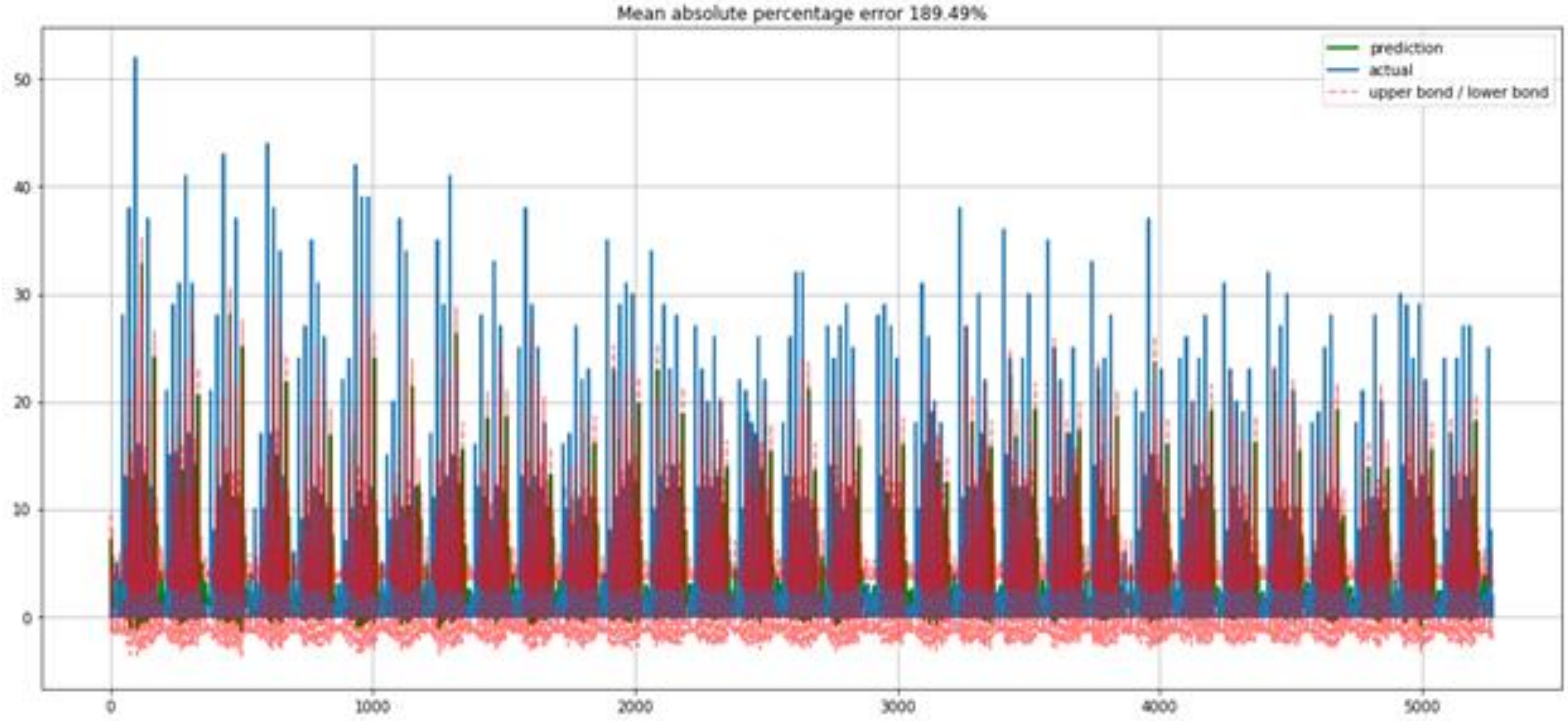
Random Forest Regression Results


Actual vs Predicted for a particular station (66)

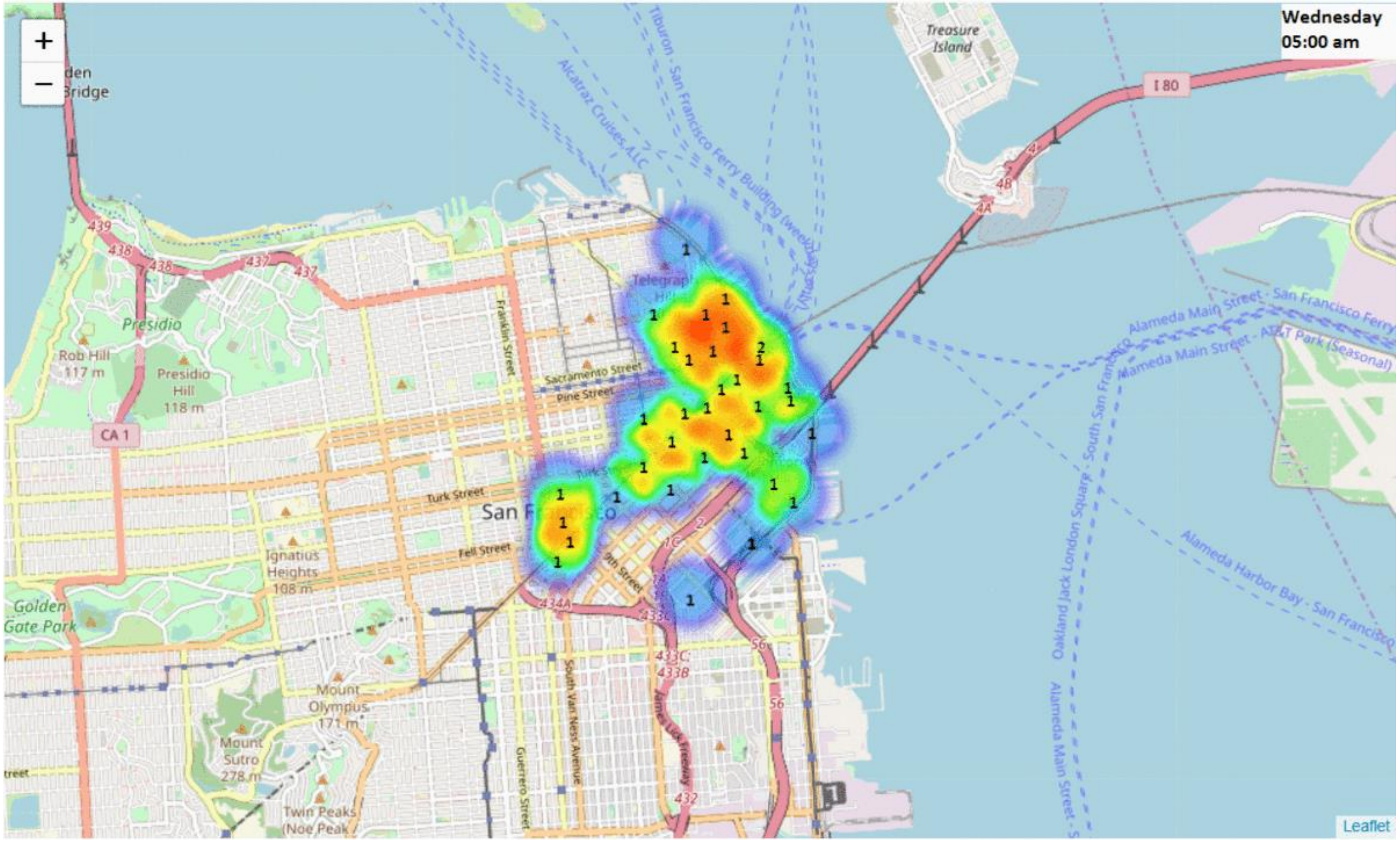

Variation in MAE across the day for station ID = 66

Finding: The error in prediction is not consistent across the hours of the day. The prediction is more accurate in wee hours of the day as compared to later in the evening.

Attempt at Time Series Modelling did not give accurate results



Forecast Visualization in San Francisco on Wednesdays from 5:00 to 22:00



Wednesday
05:00 am
+
−
Treasure Island
I 80
Alcatraz Cruises, LLC
Tiburon - San Francisco Ferry Building
Presidio
Rob Hill 117 m
Presidio Hill 118 m
CA 1
Franklin Street
Sacramento Street
Pine Street
Telegraph Hill
Turk Street
San Francisco
Fell Street
Ignatius Heights 108 m
9th Street
Golden Gate Park
Mount Olympus 171 m
Mount Sutro 278 m
Twin Peaks (Noe Peak)
South Van Ness Avenue
Guerrero Street
James Lick Freeway
Alameda Main Street - San Francisco Ferry
Alameda Main Street - AT&T Park (Seasonal)
South San Francisco
Oakland Jack London Square
Alameda Main Street
Alameda Harbor Bay - San Francisco
Leaflet