Predicting Shared Bike Usage in San Francisco

Springboard Capstone Project

J Slaga

As cities are growing more congested, more people are looking toward shared mobility services over car ownership for their transportation needs. Approximately 60% of trips in the US are five miles or less. It makes sense that in dense urban areas, shared e-scooters and bikes have become popular alternative modes of transportation. They allow people to quickly get through these short distances while avoiding heavy traffic. For commuters, these vehicles can be useful to supplement public transportation by providing an efficient means to get to and from major transportation hubs such as trains or ferry stations. This can increase ridership on public transportation while simultaneously reducing the city's traffic congestion. However, in order for this to happen there must be enough bikes to meet demand. Without this, opportunities for potential revenue and increased ridership are lost.

The purpose of this analysis is to create a data driven model to help rideshare companies increase revenue while also providing a better service to the bike share community.

SF Bay Area Bike Shares provides data on all trips taken on their shared bike service as well as data on stations' bike availability and capacity, and weather data for each day. By analyzing this data, I can create a predictive model of bike usage throughout the city based on patterns of usage on different days of the week and different weather patterns. Bike share companies can use these models to better plan expansion throughout the city to provide a better service to their customer base and also increase their annual revenue. The model would also be useful in their day to day operations for optimizing bike reallocation at various stations. Knowing where most bikes end up in addition to when they are most needed, trucks that transport bikes back to stations could be more efficiently dispatched thus maximizing the available bikes for rent and increasing potential usage.

In addition, bike share companies could provide input to city planners. The analysis of this data and projections from the model would be useful to city planners in understanding bike routes and usage for expanding bike lanes across the city. Additionally, these routes can be seen as an integral part of the overall public transportation system as they can provide the important first- and last- mile solution (the trips to and from the central transportation hubs).

Predicting Shared Bike Usage in San Francisco

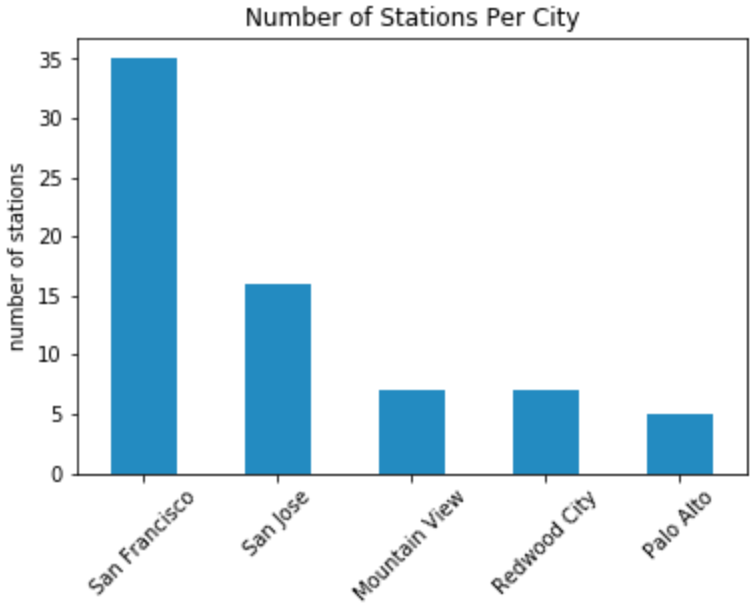
Data:

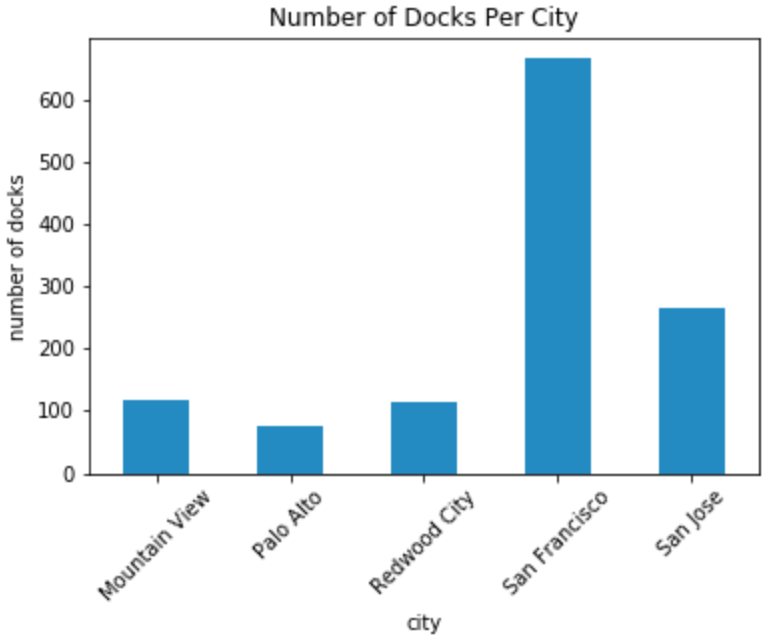
Data was found on Kaggle: https://www.kaggle.com/benhamner/sf-bay-area-bike-share. Original data is released regularly on http://www.bayareabikeshare.com/open-data. There are four files available: station.csv, status.csv, trip.csv, and weather.csv. I will be building a predictive model based on weather, trips, and station data.

station.csv	trip.csv	weather.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	start_date start_station_name start_station_id end_date end_station_name end_station_id bike_id subscription_type zip_code	mean_temperature_f min_temperature_f max_dew_point_f mean_dew_point_f min_dew_point_f max_humidity mean_humidity min_humidity max_sea_level_pressure_inches mean_sea_level_pressure_inches min_sea_level_pressure_inches max_visibility_miles mean_visibility_miles min_visibility_miles max_wind_Speed_mph mean_wind_speed_mph max_gust_speed_mph precipitation_inches cloud_cover events wind_dir_degrees zip_code

Predicting Shared Bike Usage in San Francisco

The station.csv and the trip.csv for this project are fairly clean; the weather.csv had some missing data points to address. The trip and station datasets contain data from multiple cities in the Bay Area, but for the purpose of this analysis I filtered the data to only include trips and unique stations in San Francisco (a majority of the data).





Predicting Shared Bike Usage in San Francisco

The main cleaning of the trip dataset was in removing outliers, specifically trips which seemed to be errors. The mean trip duration of 17.117811 minutes was greater than 75% of the data, and the std was 380.557814 (over 6 hours). The max duration is 4797 hours. Clearly, outliers are strongly affecting the data. Rather than replacing the values with the median value for duration, these trips will not be included in the clean dataset as these trips are most likely errors in usage. I kept data falling within the 98th percentile. The updated mean of 10.65, std 9.414, and a max duration of 102.63 are far more reasonable.

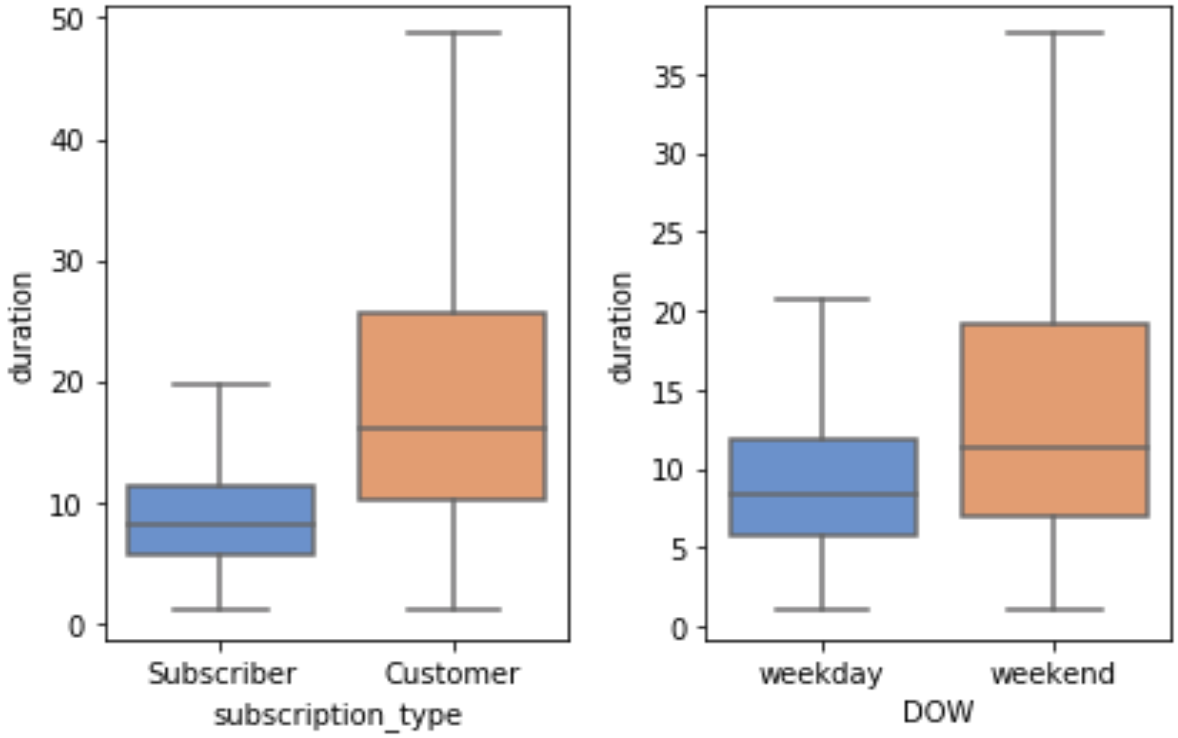
As I was driven by the analysis of commuter patterns, differentiating between weekday and weekend was deemed necessary. Thus a column was added to the trip dataset to distinguish weekday and weekend trips. A column for identifying subscription versus single-use customers already existed. We can then examine the possible correlation between subscriber/customer and weekday/weekend.

The weather dataset had the most missing data, primarily in the ‘events’ column (~85% missing). However, this was expected as this column only recorded significant weather events such as storms. Missing data was then filled with ‘Normal’ as these days were without extreme weather events. Missing precipitation data was filled in with the mean precipitation over the entire column. Data for the San Francisco zip code was extracted and kept for creating our model. Finally, variables that were most relevant for creating a predictive model were kept so that they could be merged with the trip dataset for analysis.

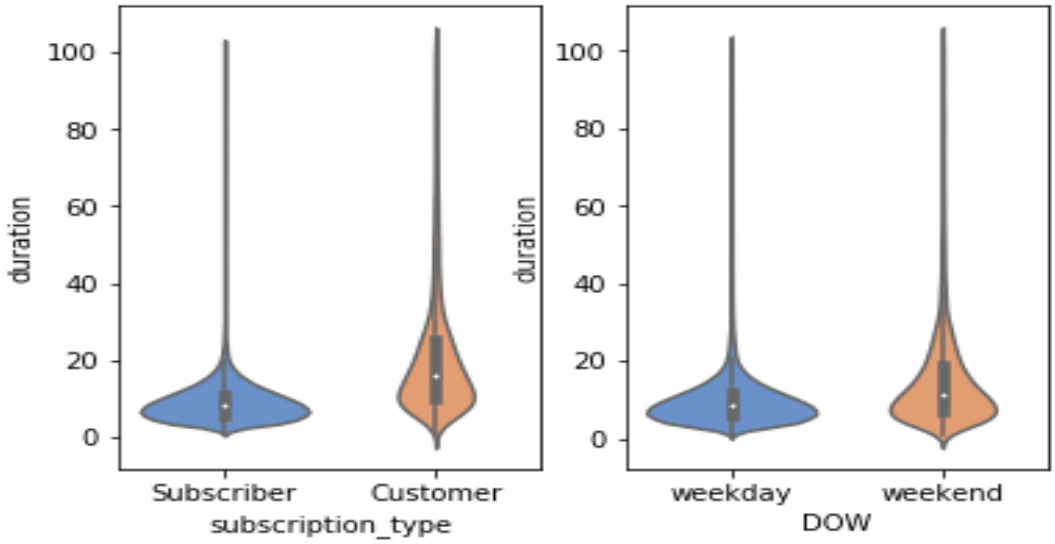
Exploratory Analysis:

I divided data by subscription type (monthly subscriber or single use customer) and by day of week (weekday or weekend) to examine possible differences in ride usage. These categories potentially provide insight into patterns of commuters or casual riders.

Predicting Shared Bike Usage in San Francisco

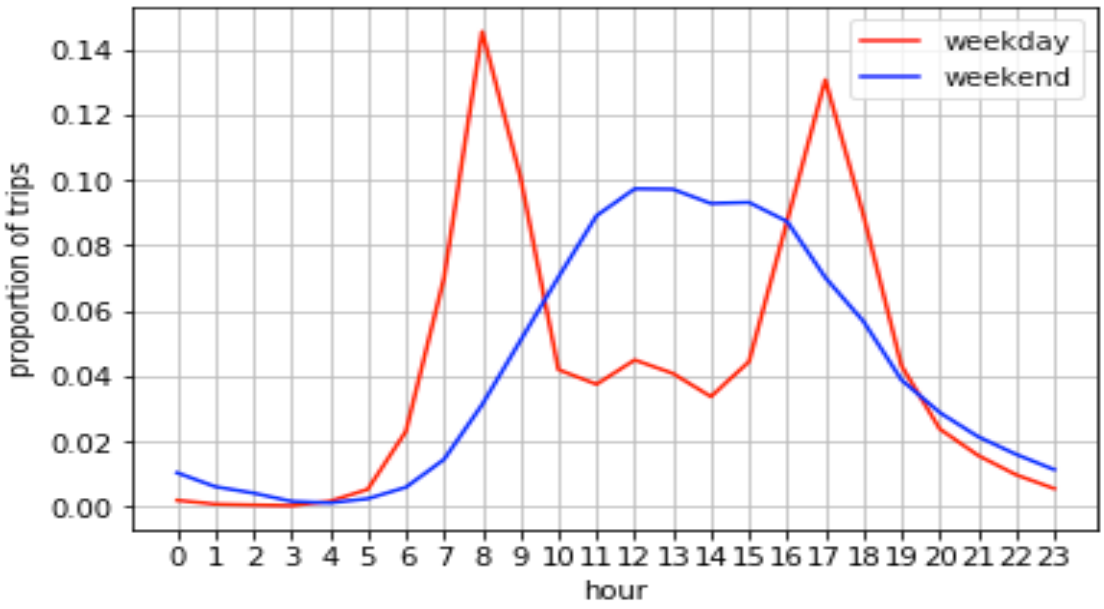


Most subscriber trips have a duration less than 20 minutes while the customer trips have a larger range and tend to a longer duration. Similar patterns can be seen in comparing the weekday versus weekend durations. It may be that weekday trips tend to be commuters (and thus subscribers) whereas weekend trips (and one off customers) would be sightseeing and more leisurely rides.



Predicting Shared Bike Usage in San Francisco

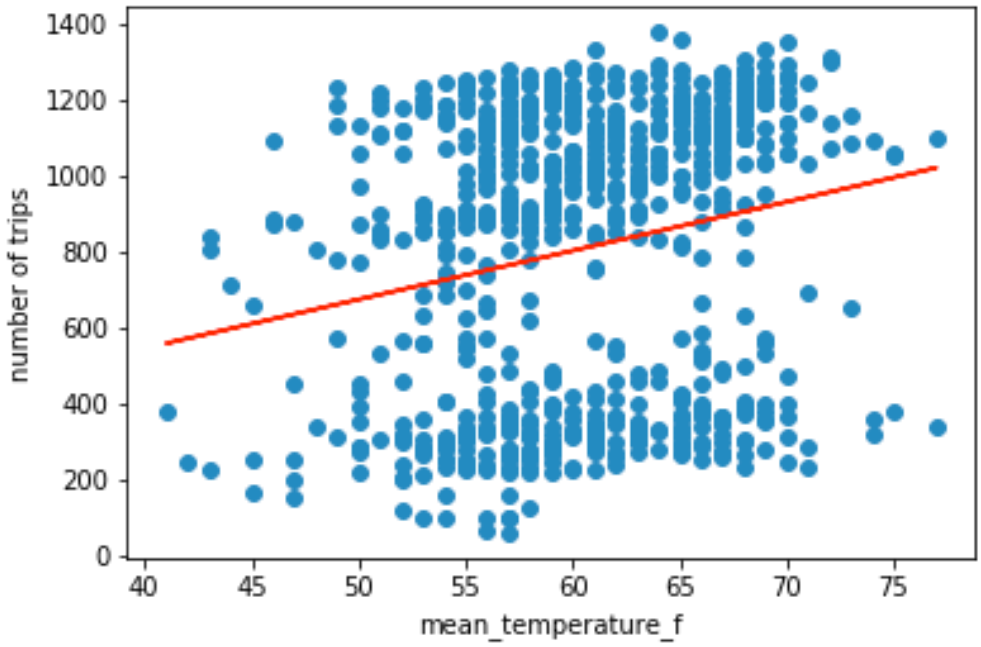
When I applied Bootstrapping Hypothesis Testing using the mean as the testing statistic, it returned a p-value of 1. Thus the Null hypothesis cannot be rejected and there is no statistical difference between the trip duration of subscribers and customers.



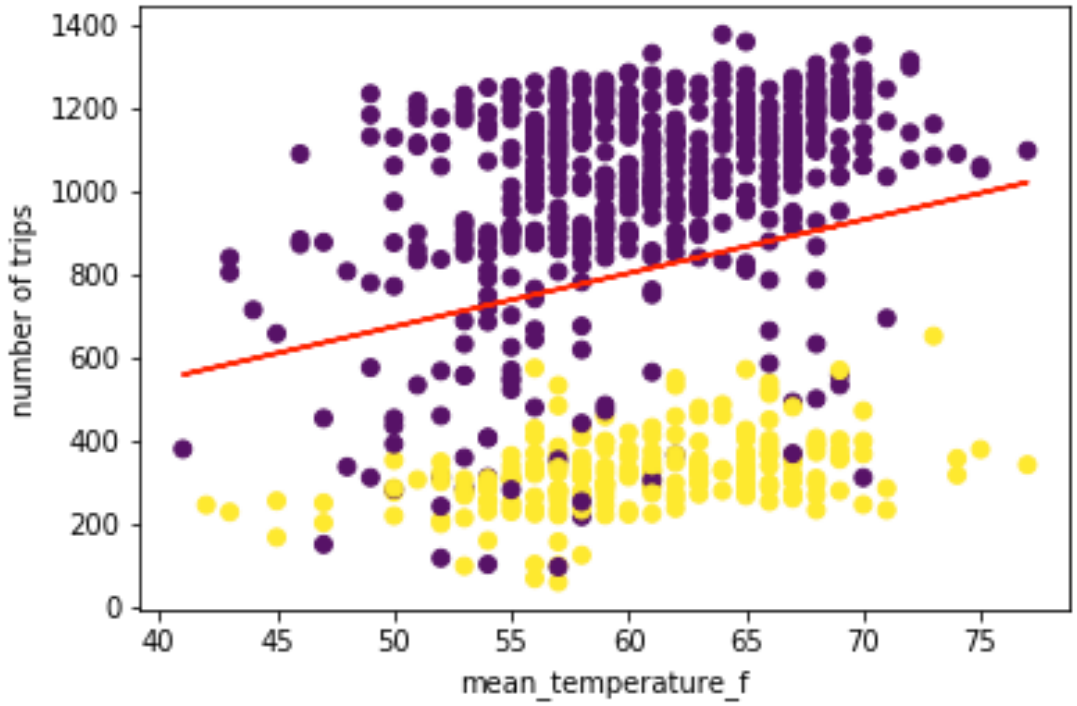
However, examining the time series of the trip data when split between weekdays and weekends revealed distinct usage patterns. The clear spikes at morning and evening commute times and a lesser spike at lunchtime support the observations that weekday usage is tied to commuting. Weekend usage, on the other hand, has a gentle curve peaking midday and gradually decreasing through the afternoon as one would expect from more leisurely rides and sightseeing.

I then turned my focus to the correlation between weather and trips. The weather dataset contains data for maximum, minimum, and mean variables. I started by examining mean temperature plotted against the number of daily trips. Mean daily temperature appears to have a slight positive correlation, and there appears to be two distinct clusters.

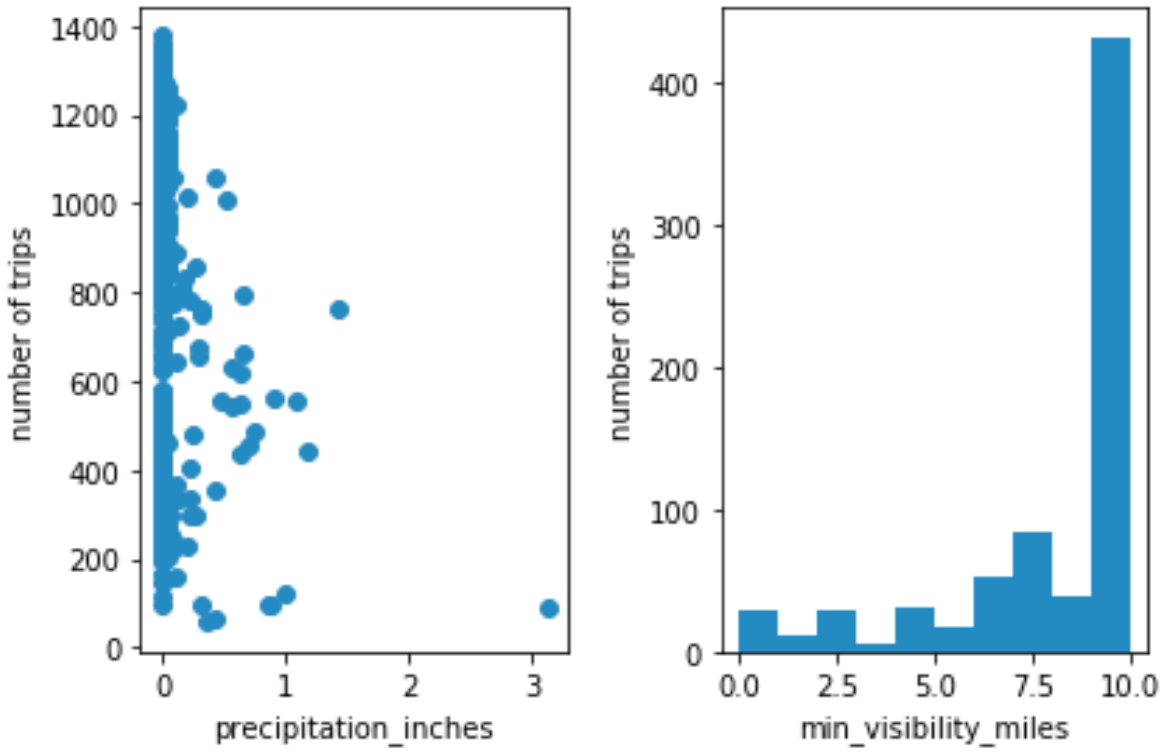
Predicting Shared Bike Usage in San Francisco



As initial data exploration revealed that trips per day were higher on weekdays than weekends, I suspected that the clusters were linked to this. Further analysis showed this to be true - the clusters were correlated to weekday (purple) and weekend (yellow) usage - again demonstrating differences between weekday and weekend usage.



The next weather variables I explored were precipitation in inches and minimum visibility in miles. There is not a lot of range in values for precipitation data (many days have zero precipitation), thus it's highly concentrated around zero. I plotted trips against minimum visibility as lower minimum visibility is indicative of denser fog. As one would suspect, there are fewer trips when there is lower visibility / dense fog. Poor weather seems correlated to a lower number of rides.



Creating Our Models:

Once the data was cleaned up and the initial exploration was complete, an in-depth data analysis was performed using machine learning. Based on the variables explored, I'll explore a set of predictive models and test to see which yields the best results. Data was split into training and test sets at 75/25. I began with a Linear Regression model. After training the model on the data, I calculated the coefficients and p-values for each independent variable.

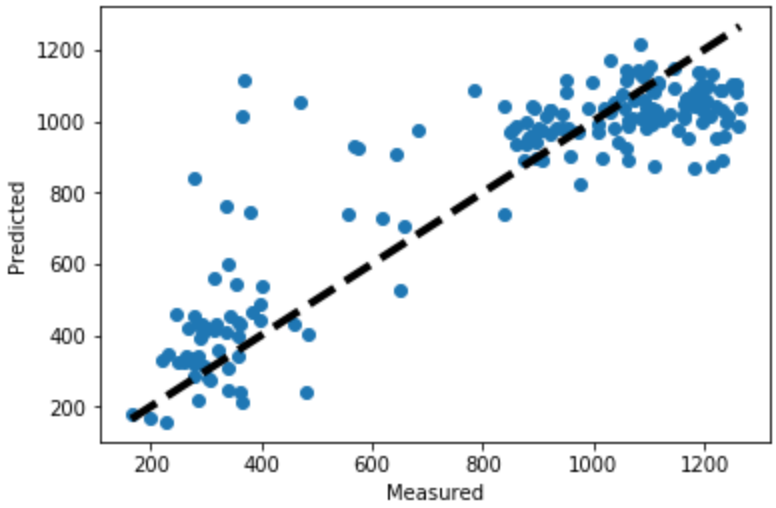
Predicting Shared Bike Usage in San Francisco

	Attribute	Co-efficient	pvalues
0	weekday	676.900292	1.000000e+00
1	max_temperature_f	420.020979	9.992079e-01
2	min_temperature_f	-16.876437	9.993661e-01
3	mean_dew_point_f	91.943932	6.160493e-01
4	max_humidity	125.294426	9.987384e-01
5	min_humidity	-30.049946	5.487181e-170
6	max_visibility_miles	33.442972	1.000000e+00
7	min_visibility_miles	58.915323	9.294101e-03
8	max_wind_Speed_mph	-69.826023	8.321388e-31
9	precipitation_inches	-912.042074	1.119223e-02
10	cloud_cover	109.198369	2.581823e-08
11	wind_dir_degrees	80.326301	0.000000e+00
12	Fog	5.430753	9.837501e-01
13	Fog-Rain	-111.845127	1.780023e-04
14	Rain	-40.243679	9.579429e-01

Weekday/weekend, maximum daily temperature, maximum humidity, precipitation in inches, cloud cover, and Fog-Rain have the most significant effect on rides per day. The Linear Regression model produced a Root Mean Squared Error of 166.82133026743972 and an R^2 of 0.7777183308707974 - the model is fairly strong.

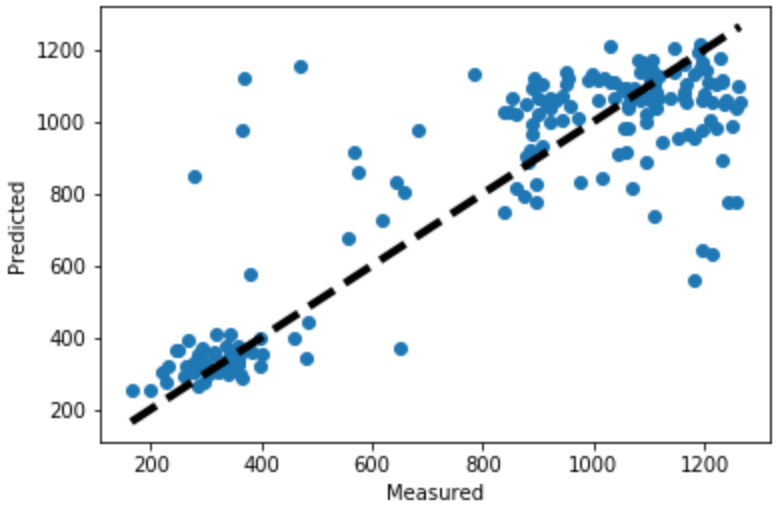
Predicting Shared Bike Usage in San Francisco

Linear Regression



The dataset was also used to train a Random Forest Regressor. This model produced a Mean Absolute Error of 121.46 and Mean Absolute Percentage Error of 18.0257

Random Forest Regressor



Predicting Shared Bike Usage in San Francisco

Conclusions

Exploration of the dataset confirmed many common sense expectations of weather's effects on bike usage. Poor weather conditions yielded less trips taken. While this is useful for forecasting bike share usage, weather is not a variable that can be controlled. For increasing revenue, more interesting were the distinct differences between weekdays and weekends and subscribers and customers. The distinct patterns of usage suggest a strong commuter ridership. This is the group that should be the focus for increasing revenue.

This can be achieved using two approaches. First, this group should be targeted for subscription services. Commuters could be offered incentives to be a monthly subscriber or this could be offered as a benefit from their employers as part of their commuter benefits package.

Second, further in depth study should focus on examining the station usage of commuters to increase bike availability. During peak hours unavailable bikes translate to lost rides and therefore lost revenue. Maximizing the available bikes at each station during commutes as well as adding stations in heavily used areas servicing public transit hubs will allow for increasing the customer base. This optimizes the integration of bike sharing into the overall transit system by providing the solution to the commuters' first leg / last leg dilemma.