Final Report:
San Francisco
Bike Sharing Usage Analysis

Dec. 2019

Abhishek Damera
Michael Huang
Ming Yu Chi
Moraldeepsingh Sachdeo
Sean Chuang

Introduction

Bike sharing is a trend in the major cities in the US. These shared bikes decrease pollution in the city and provides flexible short distance travel. However, according to bike sharing usage data provided by SFMTA, average trip is less than 4000 per day considering 262 stations already set up in San Francisco[1]. We would like to analyze how different factors (e.g. weather, weekend, demographic, nearby functionalities affect the usage of bike sharing). With the analysis, we hope to provide insight for bike sharing operations management and also a use case for other cities interested in adopting similar programs.

Impact

The bicycle docking station needs optimization since many a times, the station is at its peak capacity in the areas where docking is needed. The inflow in these areas such as Mall tends to be much greater than outflow since people tend to uber back causing bike docking area to be full but identifying these areas can be crucial in order to improve its efficiency. By this analysis: We plan to answer these questions: What is the average duration of a trip? Does it vary by cities? Does it vary by the day of the week or the time of the day? Does weather and demographic characteristics impact the number of trips taken?

This report summarized some of the data modeling and analysis results associated with San Francisco bike sharing data obtained from Kaggle as well as additional feature data. Our overall goal for this project is to know are the following:

1. Predict the number of station-based bike trip patterns vary by day-to-day basis
2. Understand how different factors such as weather, demographics statistics, near-by business, bus stations, parks and recreations facilities affect bike usage.

Dataset

The dataset is mainly derived from Kaggle's San Francisco bike sharing trips dataset which contains trips made from August 2013 to August 2015. We also included the additional features in the following:

Dependent Variables:

Number of arrivals of 35 San Francisco bike stations per day

Features:

Weather: daily high/low/average, heating/cooling degree day (HDD/CDD)
Statistics according to zip codes: Median household income/home value, population density, population, Number of people using bicycles

[1] SFMTA bike sharing system wide activity data: https://stats.sfmta.com/t/public/views/FordGoBike/BikeShareSystemwideActivity?iframeSizedToWindow=true&%3Aembed=y&%3AshowAppBanner=false&%3Adisplay_count=no&%3AshowVizHome=no#3'

Location-based data[2]: Number of business/bus stations/ parks within 400m (2-3 blocks) of distance of each stations
Time-based: weekday/weekend, station-based bike usage on previous day
Time series related: number of trips on the previous day

Note: Please refer to appendix I for the original data source. Appendix II for the definition for all the features.

Models

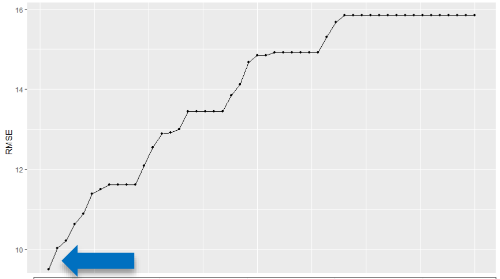
Since we will be training on the number of trips made by people based on the various categorical and numerical features, we will be using various supervised algorithms like linear regression, random forest, CART, etc. We have also implemented time series analysis into our models and look for the presence of seasonality in the trip data.

Linear Regression

After the testing and training dataset was formed in a ratio of 66.67%, The Linear Regression model was run on 34 features but based on the VIF score and linearity between columns, 13 features were dropped such as Day, Year, station, Maximum and Minimum Temperature etc. The OSR2 value obtained was 0.6838

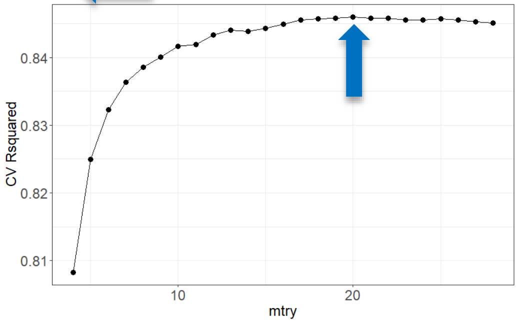
CART

The CART model was trained on the same 34 features as the linear regression did while tuning the model by cp value from 0.002 to 0.1 spaced by 0.002 with cross validation through RMSE by 5 folds. The optimal model was selected using cp value = 0.002.



Random Forest

The random forest model was also trained on the same 34 features as previous models while tuning the model on mtry from 1 to 28 with cross validation through RMSE by 5 folds. The optimal model was selected by using mytry = 20.



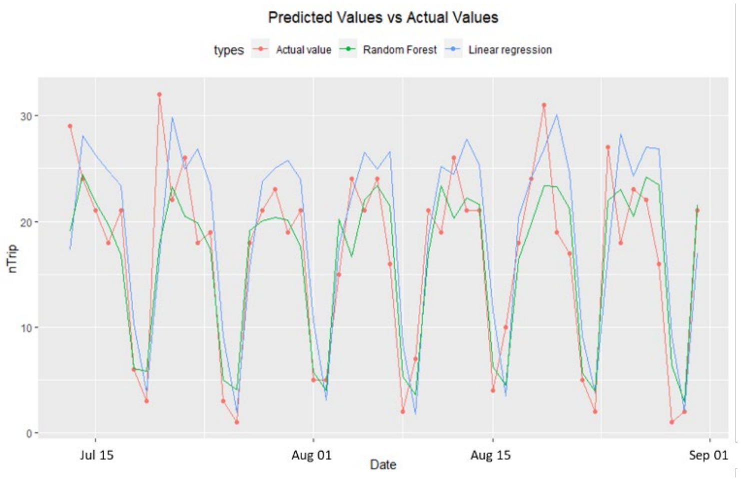
Ensemble Learning

Blend previous three models to form a composite model.

Results & Discussion

A table comparing the performance of different models is summarized below. Random forest has the best performance on predicting the number of daily bike trips

[2] Locations based data were calculated by geographic distance calculated by the data obtained from dataSF (https://datasf.org/opendata/) using additional Python code written. Note that since the distance is relatively short, we approximated by straight line distance of 400m (3-4 blocks) radius from bike stations.



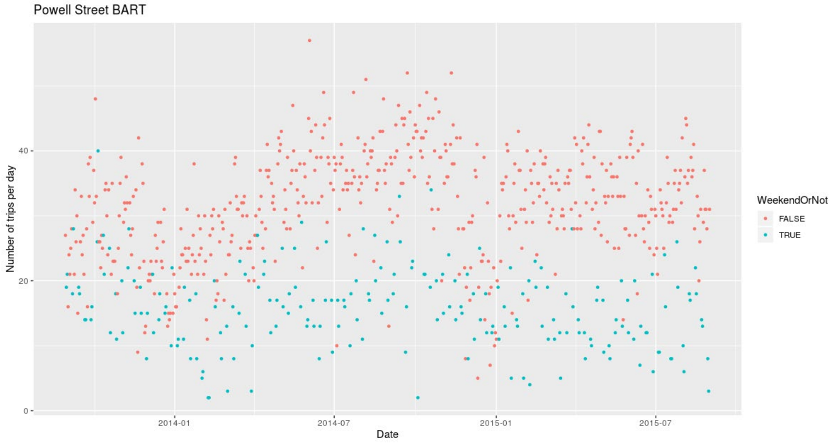
In addition, comparing the predictiveness of linear regression and random forest on number of trips from 7/13/2015 to 8/30/2015 of the stations (Broadway st. at Battery st.). Random forest provides better prediction result on weekday trips which the bike sharing system were more utilized during weekdays probably because people are using bike sharing systems as a way to commute to their final destination where bart/bus stations cannot reach directly.

A further analysis on the variable importance based on Random Forest Model, below table lists 6 most important variables which impact the number of trips made. For instance, the most important variable is the time series variable nTripYesterday.

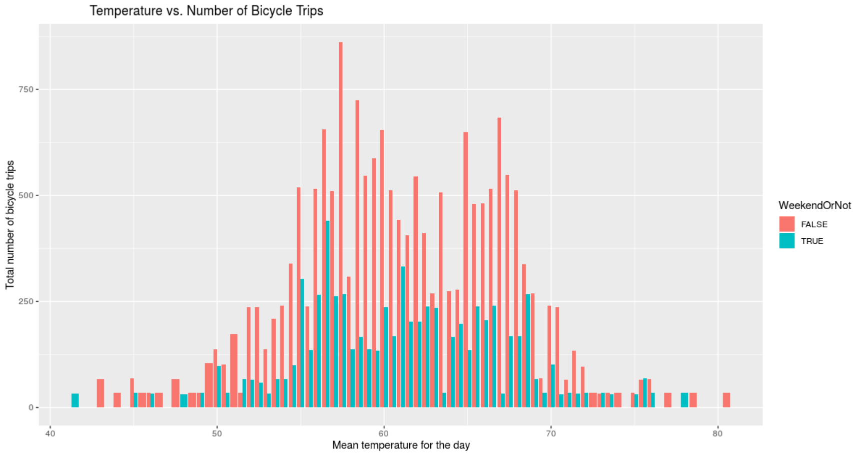
Model	OSR2
Linear Regression	0.6838
CART	0.8222
Random Forest	0.8376
Ensemble Learning	0.8143

Rank	variable	IncNodePurity
1	nTripYesterday	3100877.71
2	WeekendOrNotTRUE	1093312.57
3	Median.Household.Income	494596.31
4	Temperature.Average	373517.08
5	Temperature.Departure	240590.36
6	Business_Accommodations	185738.56

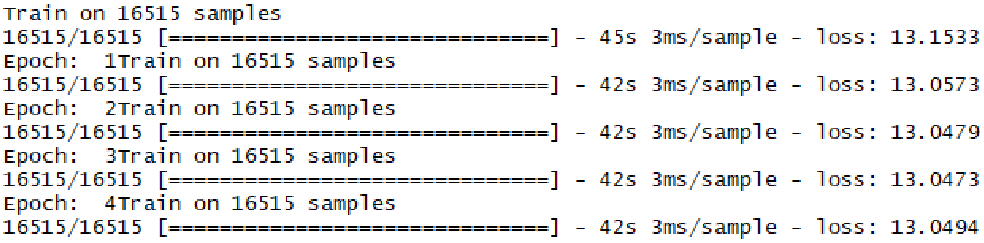
Comparison of No of trips and Day (Weekend/Weekday) (scatterplot)



As seen in the table, Weekend and Weekday ranks second in depicting the number of trips being made.



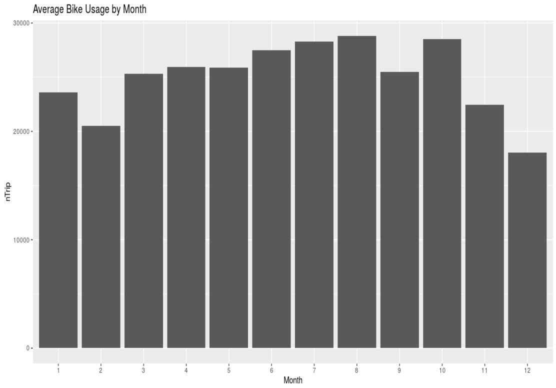
The following plot highlights 2 of the major factors: Weekend and mean temperature influencing the number of trips.

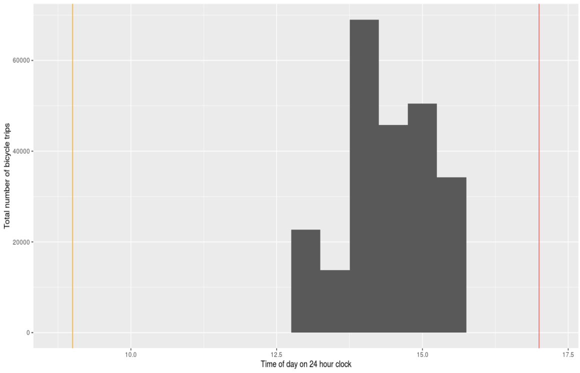


```
Train on 16515 samples
16515/16515 [==============================] - 45s 3ms/sample - loss: 13.1533
Epoch:  1Train on 16515 samples
16515/16515 [==============================] - 42s 3ms/sample - loss: 13.0573
Epoch:  2Train on 16515 samples
16515/16515 [==============================] - 42s 3ms/sample - loss: 13.0479
Epoch:  3Train on 16515 samples
16515/16515 [==============================] - 42s 3ms/sample - loss: 13.0473
Epoch:  4Train on 16515 samples
16515/16515 [==============================] - 42s 3ms/sample - loss: 13.0494
```

A basic LSTM model has been implemented with the following parameters: Batch size=3, Time Step=1 and single hidden layer. It is observed that LSTM model is overfitting the data and desired results are not obtained hence we conclude that it's not a suitable method for predicting the number of trips. Hence in order to incorporate LSTM techniques, we need to have large amount and more complex data to run to get suitable results.

Data Visualizations:





Average No of trips Monthly

No of Bicycle Trips according to Time of the Day

Appendix I - Data source:

1. San Francisco bike sharing trips
   https://www.kaggle.com/benhamner/sf-bay-area-bike-share
2. Map of Registered Business Locations
   https://data.sfgov.org/Economy-and-Community/Map-of-Registered-Business-Locations/ednt-jx6u?fbclid=IwAR3txwBNX4YjwSXVoWPLwdkf1DJdUylAEJXu0GnPP9VlqlzpXLTFE1gSTg0
3. Recreation and Parks Properties
   https://data.sfgov.org/Culture-and-Recreation/Properties/xri9-pfud
4. Bus Stops in San Francisco
   https://data.sfgov.org/Transportation/SFMTA-Transit-Stop-and-Schedule-Data-GTFS-format-/2qyp-77cq
5. Weather data in San Francisco
   https://www.wunderground.com/history/monthly/us/ca/san-francisco/KSFO/date/2014-8?fbclid=IwAR07ls7NJTqBDISO2ZrHiI-V-hB6QNuBF9Rw1yEaKY-5SMqknQheUyhch8E
6. Demographics and statistics of zip code
   https://www.unitedstateszipcodes.org/
7. US Zip Code data
   Unitedstateszipcodes.org

Appendix II - Data Dictionary

Feature	Definition
end_station_id	The end station ID that a bike trip finishes
nTrip	Total number of trips made in a specific day
Temperature.Maximum	Maximum temperature of a day
Temperature.Minimum	Minimum temperature of a day
Temperature.Average	Average temperature of a day
Temperature.Departure	Departure of the average temperature from the 30-year mean temperature for the day.
HDD (Heating Degree Day)	A measurement designed to quantify the demand for energy needed to heat a building. It is the number of degrees that a day's average temperature is below 65^{o} F (18^{o} C), which is the temperature below which buildings need to be heated.
CDD (Cooling Degree Day)	A measurement designed to quantify the demand for energy needed to cool buildings. It is the number of degrees that

	a day's average temperature is above 65° F (18° C).
dock_count	Total of docking port in a particular station
Zipcode	The zip code of the end station located at
Median.Household.Income	Median household income of the zip code that end station resides in
Median.home.value	Midian home value of the zip code that end station resides in
Population.Density	Population density of the zip code that end station resides in
Population	Population of the zip code that end station resides in
Number.of.people.using.bicycle	Number of people using bicycles of the zip code that end station resides in
total_bus_stops	Number of bus stops within 400 meters radius form the bike station
total_parks	Number of parks within 400 meters radius form the bike station
Business_Accommodations	Number of registered businesses within 400 meters radius form the bike station. Different business type is classified using North American Industry Classification System (NACIS) code description.
Business_Administrative_Support_Services	
Business_Arts_Entertainment_Recreationn	
Business_Certain_Services	
Business_Construction	
Business_Financial_Services	
Business_Food_Services	
Business_Information	
Business_Insurance	
Business_Manufacturing	
Business_Multiple	
Business_Private_Education_Health_Services	
Business_Professional_Scientific_Technical_Services	
Business_Real_Estate_Leasing	
Business_Retail_Trade	
Business_Transportation_Warehousing	
Business_Utilities	
Business_Wholesale_Trades	
WeekendOrNot	Identify the day if it is weekend or not

Appendix III – Code

(refer to the next page)