Spatiotemporal Analysis of Ford GoBike Data in San Francisco Bay Area

Steven Takeshita, Kim Sha, Sami Belkadi, Shayan Monshizadeh
Department of Computer Science, Princeton University

Introduction

Motivation

- Ford GoBike is a bike-share service that operates in San Francisco, East Bay and San Jose. In order to improve their service, Ford GoBike records information about each trip and publishes it on their website on a monthly basis [1].
- Through analysis of the data set, a variety of trends and visualizations can be leveraged to gather insights into how the service is being used.
- Insights could guide decisions on how to advertise to a target demographic, select ideal locations for new stations and anticipate demand depending on the time of day and the weather.

Related Work

- Regression analysis by Venketeswaran et. al found that logistic regression could accurately predict whether or not a given user was a subscriber or customer, mainly based on if the day was a weekday or weekend [4].
- Another study was able to cluster bike stations in France, under the V´elib' service, and found clusters that could inform decisions for future stations [3].

Ford GoBike Data Set

- Ford GoBike data is published monthly on the Ford GoBike website for public use. The data set contains user and trip information:
  - Trip Information: Trip Duration (in seconds); Start/End Date and Time; Bike ID
  - Station Information: Start/End Station IDs, Names, and Coordinates
  - User Information: Type (Subscriber / Customer); Year of Birth; Gender
- OpenWeatherMap weather data can be organized hourly for every day of 2017:
  - Temperature (Kelvin) – converted to Fahrenheit
  - Weather Description (i.e. "sky is clear", "mist", "broken clouds", "light rain", "haze")

Methodology

Data Cleaning and Pre-processing

- The birth year of a user is input by the user and therefore susceptible to mistakes. We removed trips whose user birth year was less than 1918, meaning the user was 100 years old, which removes .7% of the total trips.
- Some of the trip durations are extremely long or short, with the longest trip lasting 23 hours and the shortest trip lasting one minute. We removed the outliers where the trip lasts longer than three hours or below three minutes.
- 17% of user gender and user birth year data is missing. In order to impute these values, we replaced 'NaN' entries with the other values at the frequency that we have seen them in the data, thereby maintaining the relative frequency of both the member birth year and member gender.

Feature Engineering

- The feature space was expanded by using the date and time of a trip to determine if a trip occurred during a weekday or weekend. This led to simple analysis that could be used to recognize a distinct difference in subscriber and customer use of the service depending on day of the week and time of year.

Cluster Analysis (K-Means and LDA)

- Both methods' hyperparameters were selected using cross validation, with silhouette score for K-means and log likelihood for LDA as the scoring method.

Regression Analysis

- Ordinary least squares, ridge, and LASSO regression used with the optimal hyperparameter selected using cross validation.

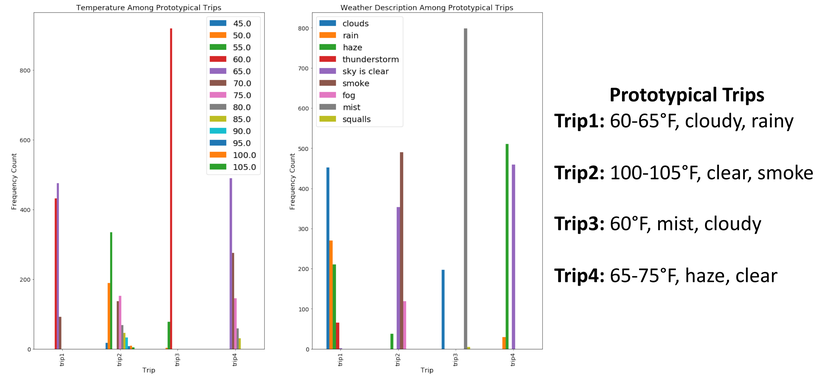
Evaluation Metrics

- Mean squared error measured accuracy for regression and silhouette score and log likelihood again measured goodness of fit for clustering

Results & Conclusions

LDA Clusters

- Distinct weather clusters were discernible from the LDA analysis.



Prototypical Trips

Trip1: 60-65°F, cloudy, rainy

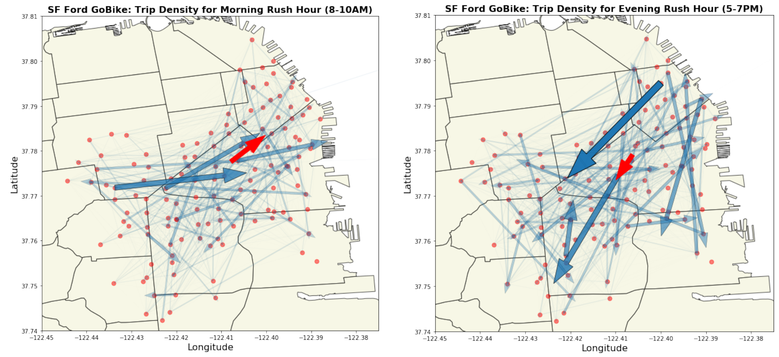
Trip2: 100-105°F, clear, smoke

Trip3: 60°F, mist, cloudy

Trip4: 65-75°F, haze, clear

K-means Clusters

- K-means clusters were used to evaluate trips during morning and evening rush hours. The density of unique trips across SF are plotted:



- Red arrows correspond to clustered trips – one at 9AM that runs for 11.4 minutes; one at 5PM that runs for 10.5 minutes.

Age Prediction

Regressor	Hyperparameter	Mean Squared Error
Ordinary Least Squares	N/A	103.34
LASSO	$\alpha = .01$	102.82
Ridge	$\alpha = 0.5$	103.34

Feature	Coefficient
start_station_latitude	-5.12
end_station_longitude	3.13
user_type_Customer	2.80
member_gender_Female	1.38
is_weekday	-0.80

References

[1] Ford Corporation. System data. Ford GoBike, 2017.
[2] Etienne Cme and Latifa Oukhellou. Model-based count series clustering for bike sharing system usage mining: A case study with the vlib system of paris. ACM Transactions on Intelligent Systems and Technology (TIST), 5, 10 2014.
[3] Yunlong Feng, Roberta Costa Affonso, and Marc Zolghadri. Analysis of bike sharing system by clustering: the vlib case.IFAC-PapersOnLine, 50(1):12422 – 12427, 2017. 20th IFAC World Congress.
[4] Aishwarya Venketeswaran, Brenton Hsu, Sucheta Banerjee, and Wiseley Wu. Midterm project: Analysis of the bayarea bike share data. 2018.

Acknowledgements
Special thanks to Matt Myers, Jonathan Lu and Barbara Engelhardt of Princeton's Department of Computer Science for their support and guidance throughout this project.