Enhance Uber Experience by Forecasting Travel Times in SF

Kailing Ding, Yuxuan Fan, Jeffrey Wang

University of California, San Diego
La Jolla, CA, 92093, USA
Author's E-mail: k5ding@ucsd.edu, yufan@ucsd.edu, chw387@ucsd.edu,

A. Introduction

Nowadays, more and more companies in most of the industries tends to integrate big data and machine learning into their business model. Specifically, one of the most important and popular ML solution is demand forecasting. Demand forecasting allows companies to decrease their supply-demand gap in order to optimize their profitability.

In this report, our data science team will utilize Uber historical travel times data between popular tourist places and transportation stations in San Francisco to help Uber have deeper understanding of

1) how travel time would change in the future
2) and how forecasting travel times might enhance user experiences.

B. Data Cleaning/Preprocessing

Our analysis mainly uses the dataset that contains the average time uber rides took to travel between the three main BART (Bay Area Rapid Transit) stations and the three most popular attractions in San Francisco every day.

Since on certain dates, there is no one took uber rides for certain routes (a route being a trip between a specific BART station and a specific hotspot), there are values missing in our dataset.

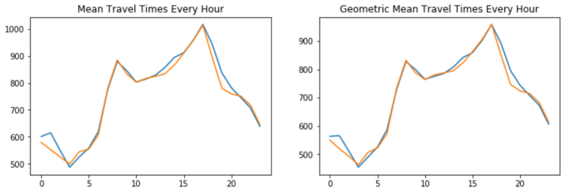

Figure 1. Comparison of Mean Travel Time and Standard Deviation Travel Time between 1st and 2nd fiscal quarter.

In Figure 1, the orange line represents mean travel times gathered from the first fiscal quarter of 2019 at any given hour of day, whilst the blue line represents the same statistic gathered from the second fiscal quarter of 2019. In essence, the figure shows that mean travel times are very similar at any given hour of the day, even when the data is gathered across different fiscal quarters. Hence, we can assume that at any given hour, the mean travel time should be very similar across different days. As such, we imputed missing travel times for any given route at that hour of day with the mean travel time from other days at the same hour and route.

C. Data visualization and Interpretation

I. Geographic relationship between BART stations and hotspots

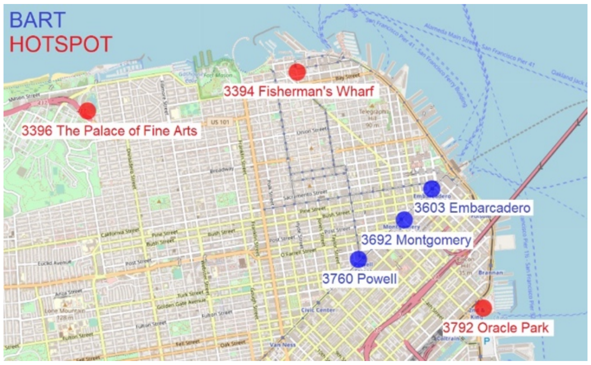

Figure 2. Map of SF downtown with location markers

We created a map using folium to visualize the distance between locations. From the graph, we can see that Oracle park is the closest hotspot to three BART stations, then Fisherman's Wharf, and the palace of fine arts is the furthest. This geological distance corresponds to the difference in travel times, in which time cost for each trip matches the geological distances.

II. Average Travel time per day of week within each month

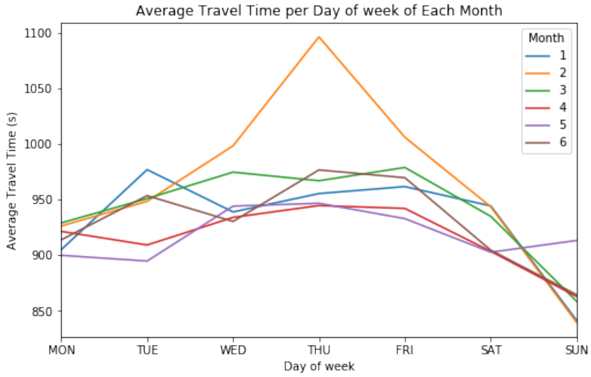

Figure 3. Average Travel Time per Day of Week within Each Month

This graph shows the amount of average travel time regarding the day of the week within each month. For each month, ride on Wednesday to Friday will take the most amount of time, and ride on Sunday will take the least amount of time. This shows a trend that rides on the weekend tends to be faster than on weekdays. Also, except for a peak on the Thursdays of February, the difference among months are not significant.

III. Average travel time per direction of travel

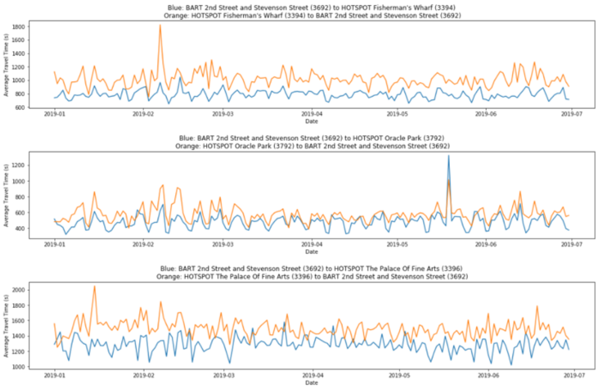

Figure 4. Sample of Average Travel Time per Direction of Travel

This graph shows the travel time difference between the two directions of the same pairs of BART stations/hotspots. From the graph, we can see there is a significant difference between trips of both directions. Specifically, the travel time from hotspots is significantly longer than travel from BART stations. Such difference has a relationship with the geological difference, in which long distances results in larger travel time differences between directions.

IV. Average travel time per hour of the day

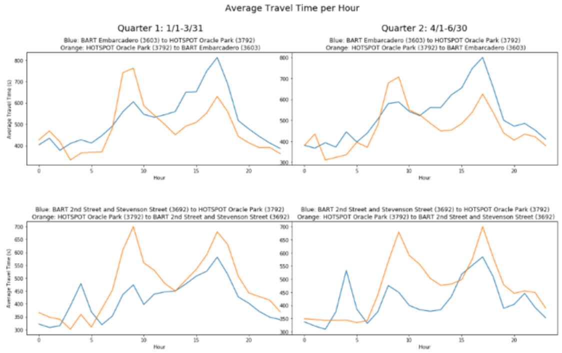

Figure 5. Sample of Average Travel Time per Hour of the Day

This graph shows the travel time difference regarding the hour of the day. Two directions of one location pair and two quarters are plotted together to spot the potential difference. Generally, the peak of travel time occurs at 8 am and 5 pm. Also, we can see for some routes, the time peak is in the morning in one direction and in the evening in another direction; while the travel time of other routes is generally longer when starting at the hotspot. On the other hand, the difference in travel time across quarters are not significant.

D. *Forecasting/Machine Learning*

Based on what we have analyzed during the process of data visualization and interpretation, our team decides to employ time-series model to forecast travel times between bart station and hotspots (6 routes in total). Before modeling, we decompose data into four unique properties of time-series data which are

- **Trend (Stationarity)**

We have conducted Augmented Dickey-Fuller test(ADF) on travel times dataset and we have achieved ADF statistics of -35.148 and P-Value of 0.0. Thus, dataset that we are using for modeling is stationary.

- **Seasonality**

From the graph below, we can clearly see that the data is seasonal; hence, we need to add a seasonality parameter into our model

- **Cyclic**

There are some spikes on the date of holidays, so we will minimize the effect of outliers as well.

- **Noise**

Based on the very bottom graph, there are some noises in the data, so we will also apply regularization and smoothing techniques when designing our model.

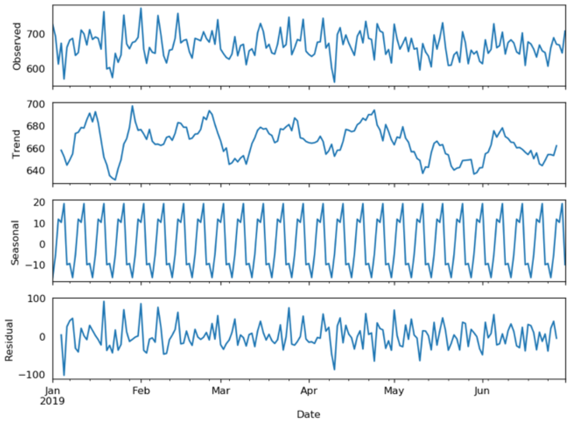

Figure 6. Decomposition of Travel times (time-series data)

Based on analysis above, we decide to choose SARIMA(Seasonal Autoregression Integrated Moving-Average) algorithm as our forecasting model, which is an univariate model. We transform our time data into Timestamp as the predictor and we leave daily travel times as our response variable. We implemented grid search to tune model's hyperparameters; in this case, parameters are p (order of autoregression), d (number of difference), q (order of moving-average) and seasonality parameter.

After tuning these parameters, we implemented a special cross validation technique designated for time-series data:

1) Divide dataset into 5 subsets
2) Put subsets into 5 different groups in orders such that training set only contains the past data, and testing set contains future data. (ex. Train: [0,1], Test:[2]; Train:[0,1,2,3], Test:[4]; etc.
3) Iterate through these train/validation groups and train and optimize SARIMA models.
4) Compare the performances of the models

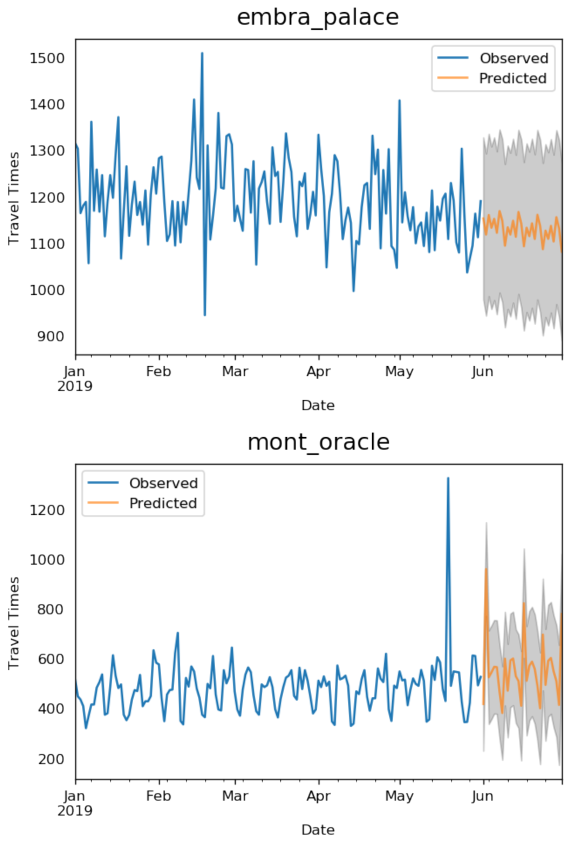

Figure 7. Prediction results of two of bart/hotspots routes

After training the model, we clearly see that as training size gets larger, the evaluation metric (mean absolute error) decreases. When we utilize the training set with the largest size, we achieved MAE value of 27.64.

We also found it interesting that the variance of our predicted travel times is quite similar as that of historical travel times and the predicted mean almost keeps the same as the previous one. We believe that this forecasting model could help Uber minimize expected time arrival.

Training Range (2019)	Testing Range (2019)	Mean Absolute Error (Test)
01/01/ – 01/31	02/01 – 02/28	610.27
01/01 – 02/28	03/01 – 03/31	56.71
01/01 – 03/31	04/01 – 04/30	37.05
01/01 – 04/30	05/01 – 05/31	33.74
01/01 – 05/31	06/01 – 06/30	27.64

Table 1. Prediction results for the future 30 days (Embarcadero bart station to Fisherman's Wharf)

E. Business Analysis and Proposal

In Figure 2.5, we see that although the time taken to travel between the hotspots and the BART stations are longer than the time taken to travel in the reverse direction in general, there are certain hours during the day where this relationship is reversed. We can assume that higher travel times are caused by higher traffic in that direction, hence we can assume that higher travel times correlate with higher demand for rides in that direction. Therefore, our recommendation is to create discounts for directions with lesser travel times or increase the price for directions with higher travel times, with specific times where the discount/price surge should be reversed.

For example, at 8 a.m. it takes longer to travel from Oracle park to Embarcadero station than from Embarcadero station to Oracle Park. At 11 a.m. however, it begins to take longer to travel from Embarcadero station to Oracle Park than from Oracle park to Embarcadero Station instead. If the higher travel time/traffic is caused by higher amount of people wishing to travel in that direction, it means demand for rides in that direction become less price elastic. Hence before 11 a.m. we can increase the price of rides from Oracle park to Embarcadero station and expect similar demand for ride requests due to the low price elasticity.

Afterwards, most drivers would likely want to return to Oracle park after dropping off their customers due to the high demand for rides starting from Oracle park regardless of whether they are making the trip with a customer or not, creating a higher supply of rides to demand. Hence, we can offer discounts for rides from Embarcadero station to Oracle park, increasing quantity demand for drivers that just dropped off customers at Embarcadero station and are hoping to return to the high demand starting point at Oracle park. This way, drivers are more likely to get requests in the direction with shorter travel time, which they would wish to drive regardless of whether a customer is present as rides starting from Oracle park are in such high demand. With this pricing strategy present, we would not only increase the revenue from price inelastic rides, but also increase quantity of rides in the reverse direction where the supply originally outstrips the demand, and therefore increase our total revenue. The discount/price increase ratio should also be proportionate to the travel time we predicted in the future, in order to reflect the corresponding demand and supply. This way at 11 a.m., the discount/price increase would be offered in the reverse direction, and therefore continue to increase our revenue when the travel time and direction relationship is reverse. As our time series model predicts travel times at any future given hour, we would be able to offer specific discount/price increase ratios at any given time in the future for the BART-hotspot routes, our model can be used in offering said price change ratios for the rideshare company to increase their revenues.

F. Conclusion

By visualizing the data, we found that travel time is affected by various features. The travel time is proportional to the geological distance, tends to be higher during weekdays than the weekend, tends to be higher for trips from a hotspot to BART station than travels of inverse direction, and two directions of some routes have peak travel time at different hours of the day. However, the travel time does not have trends across the dates, quarters or months. With our findings, we can therefore propose different price discounts/increases on rides based on predicted travel times to increase total revenue for the company.