Analysis of Taxi Trip Pricing at Chicago

--- Final Project of Applied Data Science Capstone by IBM

Version Mar 26, 2020

Chao Wang
Stuart School of Business
Illinois Institute of Technology
Chicago, United States
e-mail: cwang118@hawk.iit.edu

https://orcid.org/0000-0002-6096-7550

***Abstract*—The market of vehicle for hire rideshare has been gradually dominated by companies like Uber and Lyft over the decade, one major difference between taxi and the two vehicle for hire giants is, the inconvenience for customers to find the trip planner and price estimator, this study intend to give more insights in this market using the public data of taxi trip over the past two years. And it suggests the taxi trip prices can be estimated by using the information in public and there is room for the taxi companies to do better prediction with traffic and geographic information.**

Keywords—vehicle for hire market, random forest, linear regression, geographic information analysis.

I. Introduction

The research problem and background Taxi is one of the most important local traveling mode, however, when we are planning a trip riding a Uber or Lyft vehicle, we can see an estimated price if we try navigation APPs like Google Map. This is a project motivated by the lack of taxi trip cost estimation supported by APPs. Also, how do taxi drivers maximize their probability to get more business? The fact of this market of rideshare is dominated by Uber and Lyft by 2018, the market share in reimbursement of taxi decreased to 5.2% by Q2 of 2018 [1]. This project is trying to figure out the patterns of taxi trips origins and destinations in a day.

In year 2018, Kaggle has posted similar topic of Machine Learning competition --- New York City Taxi Fare Prediction, by using similar data set with coordinate information and additional descriptive information of date, time and number of passengers, the posted baseline RMSE (Root Mean Square Error) of prediction is at $5 - $8. The study is attempting to use Machine Learning predictive models to predict taxi trip cost and achieve a more accurate predictive model.

Also, by using folium, a geographic information Python library, it is possible to visualize the coordinates to neighborhood and address on the map of Chicago, more cluster analysis is promising in this path in the further analysis.

The remaining of this report has been organized as follows: Section II. Data and Pre-Processing, Section III. Descriptive Analysis, Section VI. Modeling, and Section V. Conclusions.

II. Data and Pre-Processing

A. Data

Data description Similar research with Taxi data in New York City is available in Kaggle hosted in partnership with Google Cloud and Coursera 2 years ago, this study is going to use a similar data from Chicago with more metrics including not only geographic coordinates, time and travel costs, but also payment method and taxi company available from Chicago Data Portal*. The raw data set file contains the data of taxi trips in Chicago over the last decade, and it is over 70 GB large.

B. Pre-Processing

The raw data is overwhelming for data processing in most software, the first step of the data pre-processing of this project is to read the data through SAS then split into annual sub samples. All the rest analysis was done by Python, stratified split sample of year 2018 as Train Set and Validate Set, sample 2019 as Test Set, assuming limited annual effect between these 2 years.

The raw data set includes 23 variables, including date and time, tripID, taxiID, pickup and drop-off communities, coordinates, trip time in seconds and pricing variables (fare, tips, toll and extra), payment method and taxi company. I have transferred Weekday, Month, Hour Bracket from the start & end time variable, they have been suggested to be candidate fixed effects in the models; Then neighborhood code of Chicago has been matched with the community area list from Wikipedia [2], the absolute difference between Longitude and Latitude of pickup and drop-off has been calculated.

III. Descriptive Analysis

[Insert Table 1 here]

Table 1 tells the descriptive of key price and time variables. The subsample 2018 includes 20,732,088 rows, with a rich volume of records, the final sample with missing values have been deleted and 11,633,106 records remains in the sample, eventually a 60/40 stratified split on variables of Weekday, Month and Hour Bracket has been done for Training and Validate Sets, data in 2019 has been cleaned to be the Test Set in the same rule. Outliers of travel time and cost have been removed following Breemen (2018) Kaggle competition project in New York City Taxi Fare Prediction [3], extreme values in special cases make our model complicated, therefore records total cost over $ 50 or travel time over 1,000 seconds have been truncated.

A. Distribution of All Key Variables

This research is focused on the geographic effects and demographic information on the prediction of taxi trip prices, though payment methods and taxi companies box-plot have also been made and there are noticeable differences among the categories, these two effects are not been considered in the following analysis.

[Insert Fig 1-5 here]

From Fig 1. The histogram of Taxi Trip Costs, The Total cost of trips has a peak of distribution of around $7-$8, the

* Raw data resource comes from Chicago Data Portal: https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew

mode of Fare is at around $5-$6. Fig 2-4 tells the fixed effects of month, weekday and hour bracket on trip fare, they suggest a potential effect of weekday and hour bracket of a day but not for month effect. Generally, the trip fare on weekend varies more than weekday's and in early morning of a typical day, the trip fare is relatively higher and varies higher. Furthermore, as we can verify this inference from Figure 5. The heat map table of Trip Fare and Trip Time suggests taxi trips are more frequent in daytime on weekdays and over night during weekend.

B. *Correlation Information*

So far, the variables available to predicting taxi trip fare in the following modeling steps are trip distance, trip time and fixed effects of weekday and hour bracket. Also, from Table II the correlation matrix suggests Trip Fare is an amount more representative than Trip Total which includes noise of other expenses of tips, toll, etc. which is more difficult to predict. It is highly correlated with the time (Trip Seconds) variable at correlation coefficient of 0.76. But it is common sense that, unlike trip distance can be estimated from the traffic information and geographic information of routes, the time variable is never given when predicting the trip cost, one example will be the google map trip planner function, distance will be suggested as long as the routes been recommended even if you are using offline map, but the time variable will be updated only if you are connecting the internet, then it estimates the time by the routes and real-time or historical traffic information. Therefore both of cost variable and time variable can be prediction variables in a taxi (or any other vehicle for hire market) trip project, but they inappropriate to be used to predict each other. To predict the price, a feature group of distance, absolute difference of coordinate and fix effects of weekday and hour bracket will be considered.

[Insert Table 2 here]

IV. Modeling

Based on the features of the input variables, in order to predict the trip prices, we are using Linear Regression and Random Forest for modeling. The weekday and hour bracket fixed effects will be one-hot encoded to set of binary variables to increase the predicting accuracy.

A. *Linear Regression*

The first model in handling this prediction model is linear regression, since we are taking fixed effects into account in model features, we will plug in them using different variable encodings. The input variables are continuous variables of trip distance, absolute difference of longitude, absolute difference of latitude, then without fixed effects variables, finally with them numerical or one-hot encoded.

On the left hand side of Table III, we can see the model fit of RMSE by plugging data in the three linear regression models without fixed effect set (Weekday, Hour Bracket, Pickup Region and Dropoff Region), it has the worst model of fit among three linear regression models, it has the similar statistics of RMSE with the second model by transferring all fixed effects to numerical variables, both of them have RMSE of 3.08 in predicting Taxi Fare price in 2019. Model 3 with One-hot transferred fixed effect variables have a much better performance with 2.92 in predicting validation set and 3.02 in predicting the test set of 2019 data sample. As

B. *Random Forest*

Random forest is known as ensemble learning method for classification, regression, it is constructed on a multitude of decision trees when training and outputting the class as the mode of the classes when work as classification or mean prediction when work as regression of the individual trees. One benefit from it is to avoid the typical overfitting problem caused by Decision Tree. [4]

By using same feature selections with linear regression models, we have generated 3 random forest models to predict the taxi trip prices as shown in Table III. Generally random forest provides models with better predictions, the RMSE of Model 4-6 is around 2.4 in predicting validation set, and it is between 2.5 to 2.6 in predicting test set. At here, the model without fixed effect set does the best prediction of 2.37 for validation set and 2.51 for test set, it is the best model fit suggested from the model listed.

V. Geographic Insights

The coordinates of taxi trips' start and destination locations can be used for further geographical analysis, as we can see from Figure 7, by using the second best predicting model (Model 6), the distribution of prediction error has been listed, to visualize the trip density in Chicago of a typical day, we select the accurately predicted records (selected as RMSE <= 2.5 and RMSE >= -2.5), then visualize selected weekday and time period on map of Chicago by using folium library in Python.

Figure 9-11 represented the heat map of taxi trip start and destination places on two selected time period, Monday morning rush hour 6am to 9 am and Saturday evening 6pm to 10pm.

As we can see, although Fig 9 and 10 are having similar heat color on the map, the differences are still apparent, the large residential neighborhoods in near north side Chicago and south loop are more frequently to be taxi trip origins. Also, the downtown and west loop business area are more often to be taxi trip destinations.

Quite differently, Figure 10 and 11 suggest a larger scale of bustling area. Figure 10, the heat map of trip origins spread from downtown Chicago to the entire near north side and most of west loop area. Figure 11, in addition tells the target of Saturday night taxi trips also includes near south side and north east residential neighborhoods. The main difference from Monday morning is heat map is, its bright north leg, the shopping and entertainment center of Chicago the north Michigan Ave --- Magnificent Mile.

The information suggests demand of taxi service not only in downtown Chicago in rush hour and weekend, but also in major residential neighborhoods in rush hour. The taxi service demanded in weekend entertainment area and target to residential neighborhoods also suggests the potential demand in early weekend hours in the afternoon.

VI. Coclusions and Future Study

As we talked about in the beginning of this paper, the purpose of this research is to figure out the taxi trip characteristics in Chicago of 2018 and 2019, by analyzing the geographic features, demographic information and construct pricing model of it.

First of all, we have found the features of taxi trips by different features, the differences of weekday effect and the hour bracket of day effect both affect the taxi trips. The other effects also exist in trip pricing like trip start/end area, companies and payment methods.

We have further completed the model construct by using both linear regression and random forest based on three versions of features, there most noticeable effects --- weekday, hour bracket and trip start/end area, the application of machine learning to this project suggests the taxi trip cost is predictable even with limited information.

Eventually, with the visualized frequency of taxi trip origins and destinations from the heat maps, we conclude that demand of taxi service varies by weekday work hours and weekend hangout time at residential areas, entertainment area and business area.

More analysis in taxi price and time prediction with historical/real-time traffic information and other available features are promising, it will assist taxi companies to attract customers with more transparent prices while they do their own trip plans. An artificial intelligence central B2C system to convenient consumers do travel plan, making appointment and e-payment, as well allow taxi drivers to track the market movement, receive orders and payment may rescue this industry from losing more market share in competing with the dominating vehicle for hire giants.

References

[1] W. Richter, Uber and Lyft are gaining even more market share over taxis and rentals, Wolf Street, July 30, 2018. https://www.businessinsider.com/uber-lyft-are-gaining-even-more-market-share-over-taxis-and-rentals-2018-7

[2] Community areas in Chicago, wikipedia, https://en.wikipedia.org/wiki/Community_areas_in_Chicago

[3] A. Breemen, New York City Taxi Fare Prediction Playground Competition, July 27, 2018. https://www.kaggle.com/breemen/nyc-taxi-fare-data-exploration

[4] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. (2nd ed.). Feburary 2009. Springer.

TABLE I. DESCRIPTIVE STATISTICS OF KEY TRIP VARIABLES

	Trip_Seconds	*Trip_Miles*	*Fare*	*Tips*	*Tolls*	*Extras*	*Trip_Total*
Count	15595190	20731390	20730070	20730790	16530330	20730560	20729840
Mean	466.22	3.67	13.66	1.72	0.00	1.10	16.60
Standard Deviation	242.09	5.87	13.98	3.00	0.42	3.80	17.61
Min	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	300.00	0.68	6.00	0.00	0.00	0.00	7.25
50%	449.00	1.30	8.00	0.00	0.00	0.00	9.75
75%	649.00	3.40	14.00	2.00	0.00	1.00	16.25
Max	999.00	978.50	999.99	800.00	907.00	999.99	999.99

Fig. 1. HISTOGRAM OF TAXI COSTS (FARE, TIPS, TOLLS, EXTRAS AND TOTAL COST)

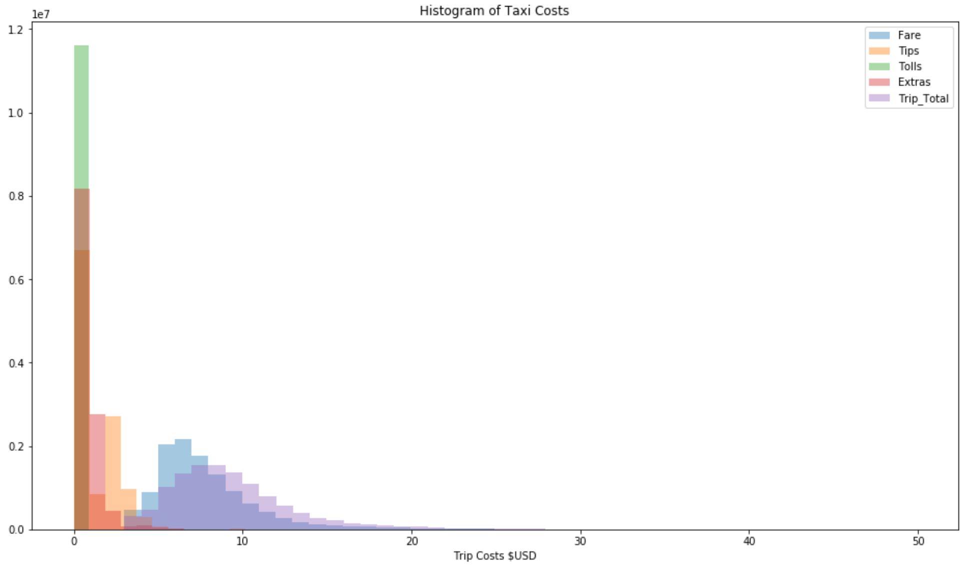


Fig. 2. BOX-PLOT OF TAXI FARE BY MONTH

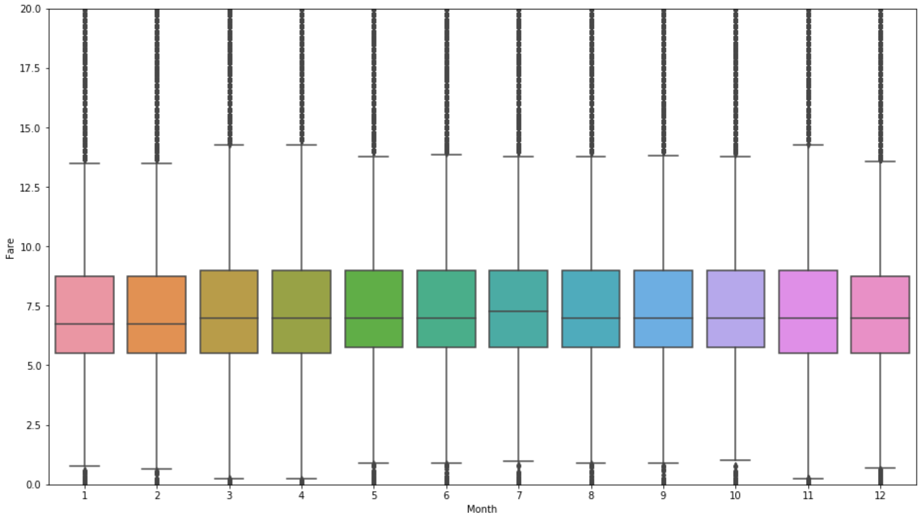


Fig. 3. Box-Plot of Taxi Fare by Weekday

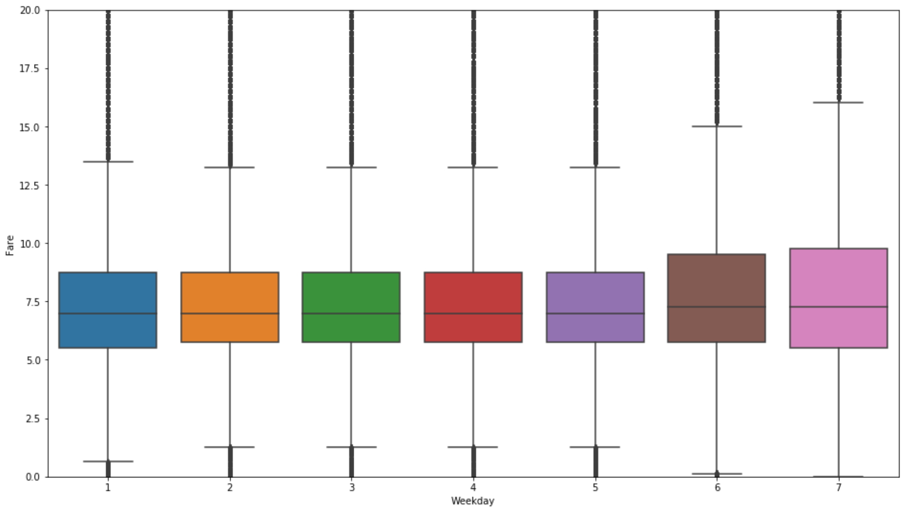


Fig. 4. Box-Plot of Taxi Fare by Hour Bracket

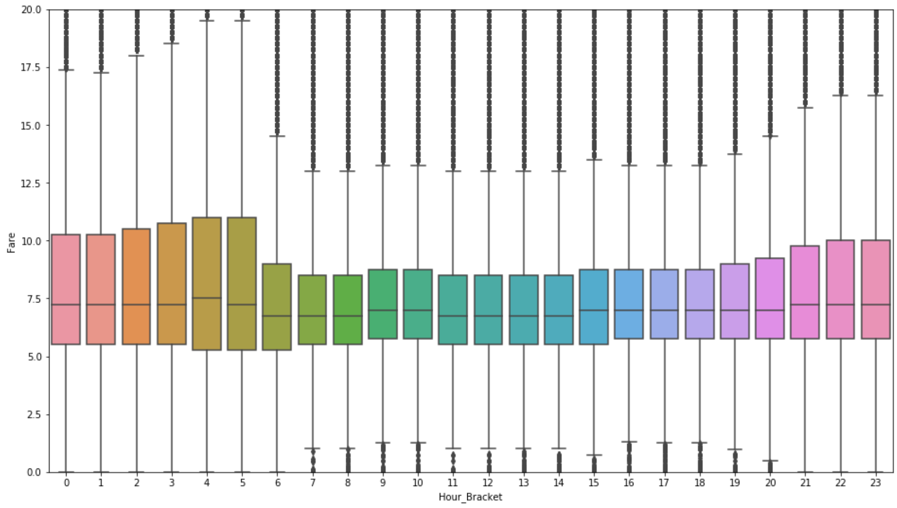


Fig. 5. Heat Map Table Trip Frequency by Weekday and Hour Bracket

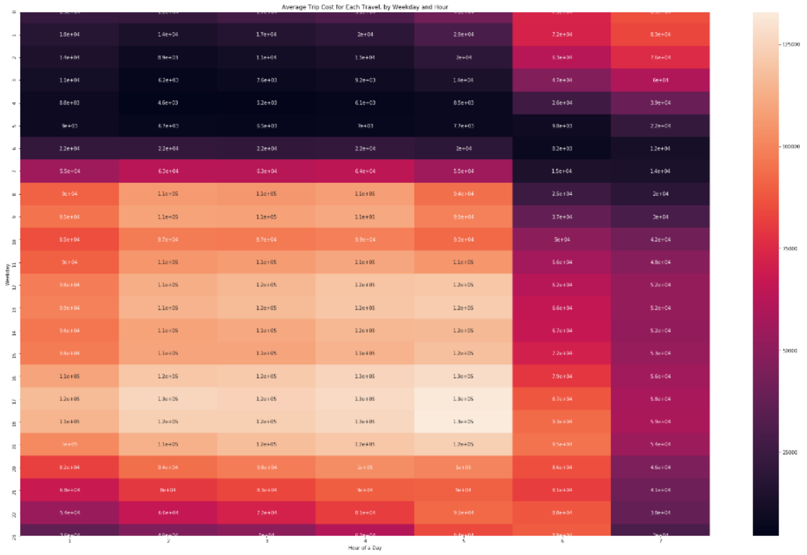


TABLE II. CORRELATION OF TRIP DISTANCE, TIME AND COSTS

	Trip_Miles	***Trip_Seconds***	***Fare***	***Trip_Total***
Trip_Miles	1	0.381498	0.535487	0.476306
Trip_Seconds	0.381498	1	0.757537	0.642592
Fare	0.535487	0.757537	1	0.863613
Trip_Total	0.476306	0.642592	0.863613	1

TABLE III. TABLE OF ALL PREDICTIVE MODEL FIT

	Linear Regression			***Random Forest***		
Parameter`	***Model 1***	***Model 2***	***Model 3***	***Model 4***	***Model 5***	***Model 6***
Distance	Numerical	Numerical	Numerical	Numerical	Numerical	Numerical
Abs_Diff_Lon	Numerical	Numerical	Numerical	Numerical	Numerical	Numerical
Abs_Diff_Lat	Numerical	Numerical	Numerical	Numerical	Numerical	Numerical
Weekday		Numerical	Binary		Numerical	Binary
Hour Bracket		Numerical	Binary		Numerical	Binary
Pickup_Region		Numerical	Binary		Numerical	Binary
Dropoff_Region		Numerical	Binary		Numerical	Binary
RMSE of Validation Set	2.98	2.97	2.92	2.37	2.43	2.43
RMSE of Test Set	3.08	3.08	3.02	2.51	2.58	2.57

Fig. 6. HISTOGRAM OF ACTUAL & PREDICTED PRICE (MODEL 1-3)

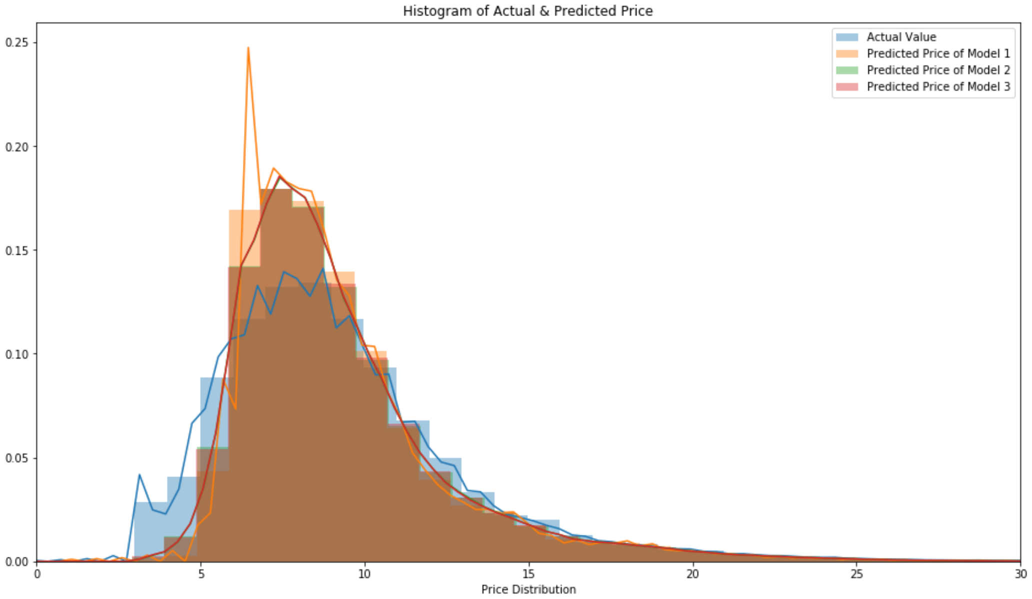


Fig. 7. Histogram of Predicted Error (Model 6)

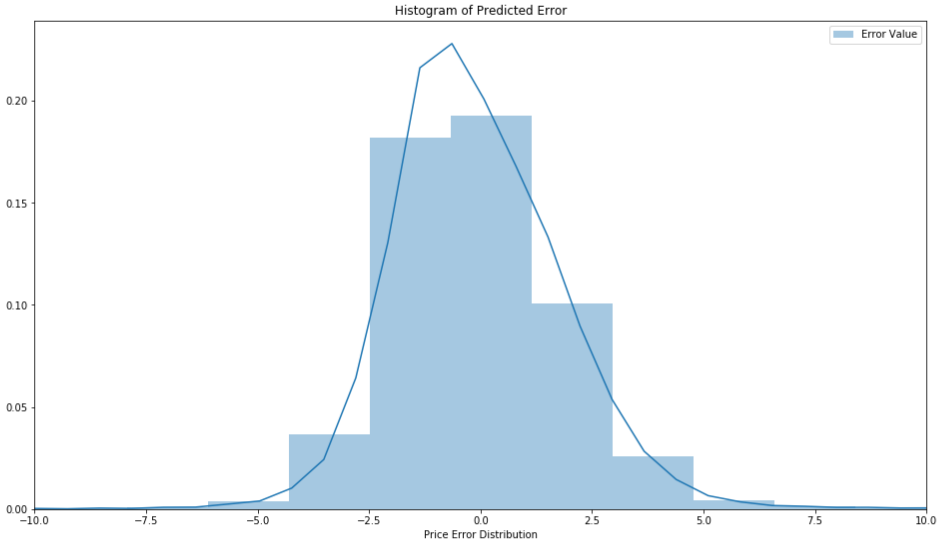


Fig. 8. Heat Map of Trip Origin Monday Morning

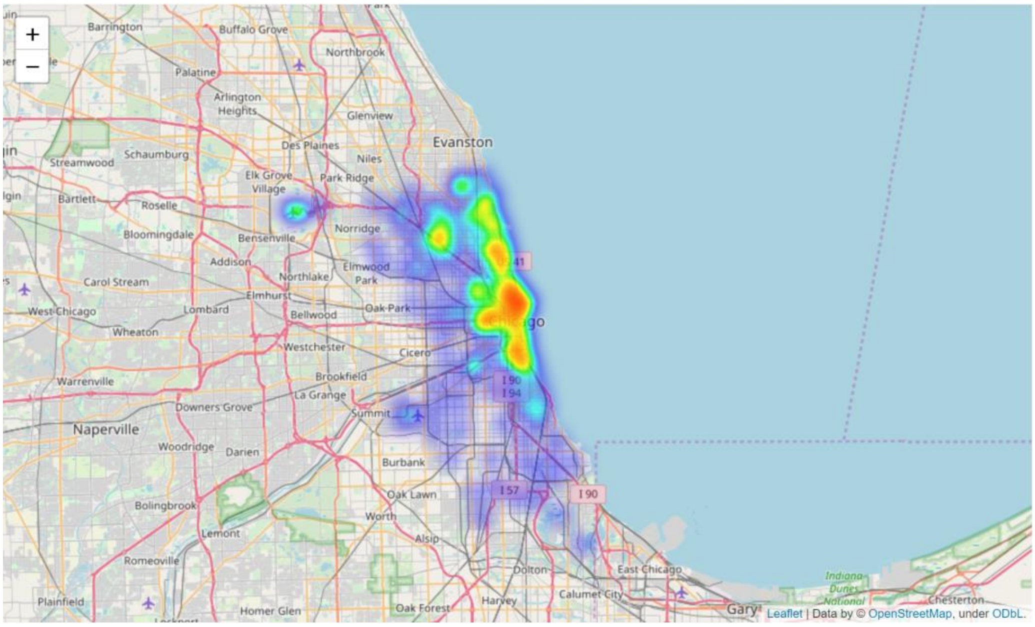


Fig. 9. HEAT MAP OF TRIP TARGET MONDAY MORNING

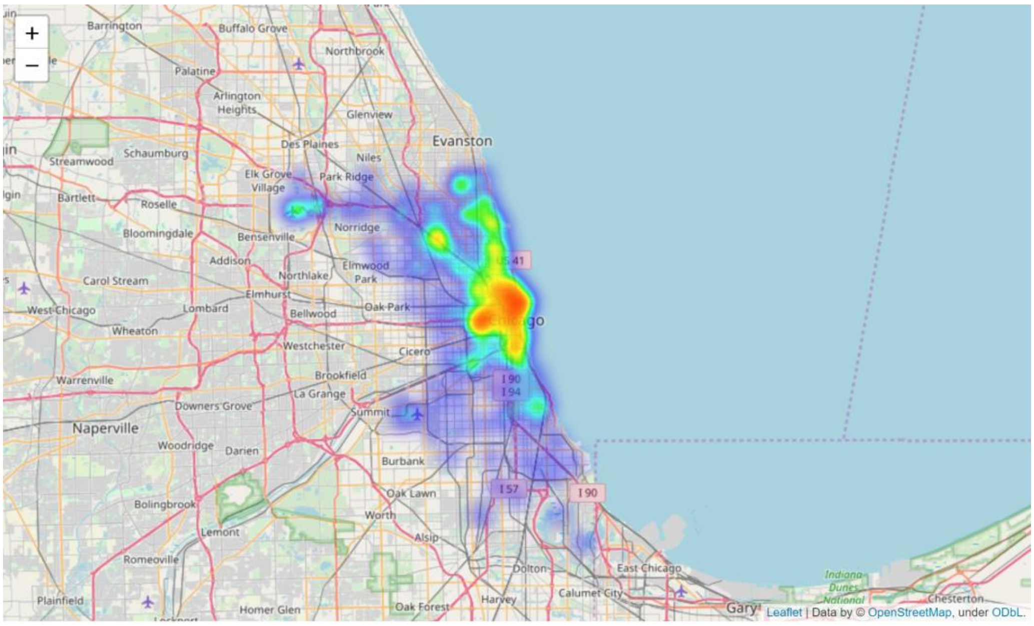


Fig. 10. HEAT MAP OF TRIP ORIGN SATUARDAY EVENING

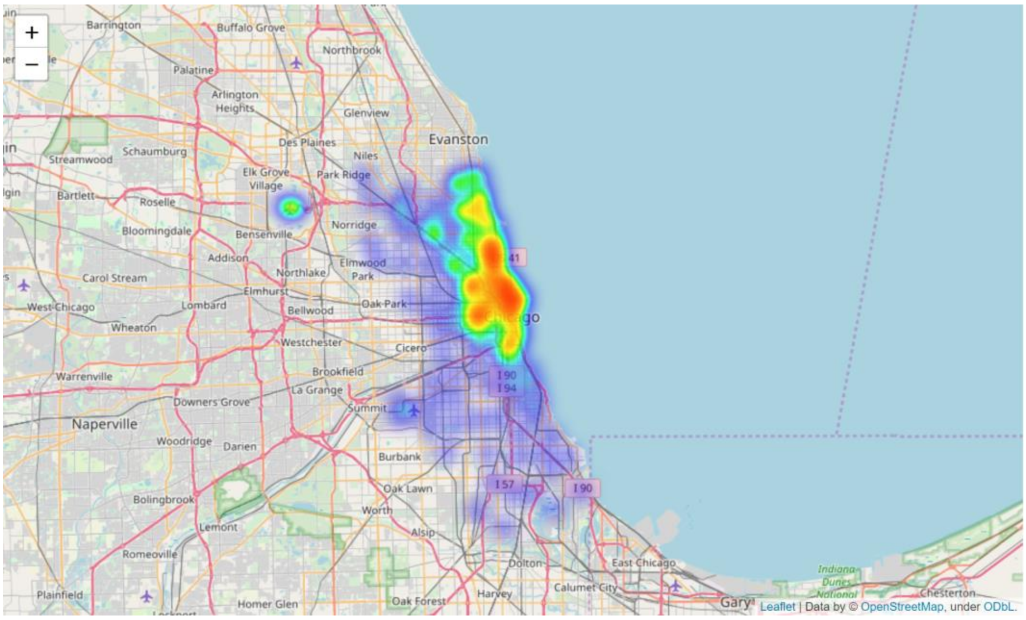
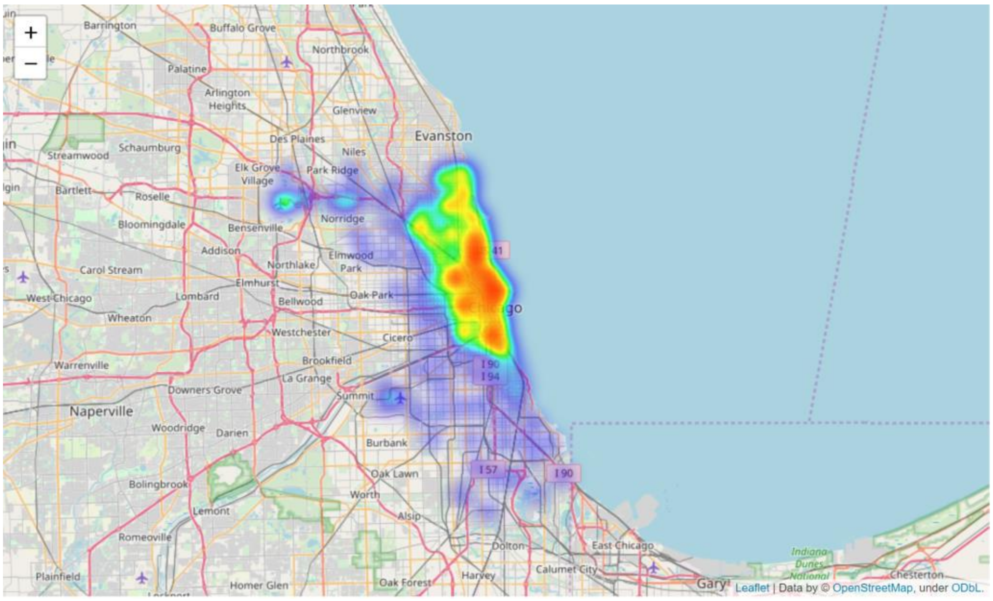


Fig. 11. HEAT MAP OF TRIP TARGET SATUARDAY EVENING


+
−
Barrington
Buffalo Grove
Northbrook
Palatine
Arlington Heights
Glenview
Evanston
Des Plaines
Niles
Schaumburg
Streamwood
Elk Grove Village
Park Ridge
Bartlett
Roselle
Norridge
Bloomingdale
Bensenville
Addison
Elmwood Park
Northlake
Carol Stream
Elmhurst
Oak Park
West Chicago
Lombard
Bellwood
Wheaton
Westchester
Cicero
Brookfield
Warrenville
La Grange
Downers Grove
Summit
Naperville
Woodridge
Darien
Burbank
Oak Lawn
Bolingbrook
Worth
Lemont
Alsip
Romeoville
Dolton
East Chicago
Harvey
Calumet City
Plainfield
Homer Glen
Oak Forest
Gary
Indiana Dunes National
Chesterton
I 90
I 94
I 57
I 90
Leaflet | Data by © OpenStreetMap, under ODbL.