A Research on New York City TLC Yellow Taxi

Zhi Hern Tom
School of Mathematics and Statistics
The University of Melbourne
Student ID: 1068268

August 13, 2021

Abstract

In New York City, yellow and green taxicabs are widely recognisable symbols of the city. With millions of TLC Yellow Taxi trip records in 2018, this project aims to discover potential factors that can affect profitability of a taxi driver.

1 An Overview on Sample Data

1.1 TLC Yellow Taxi Data

The Yellow Taxi data used in this report was obtained from the Taxi and Limousine Commission (TLC) official website[1]. Traditionally, Passengers could hail a Yellow taxi by signaling to a driver who is on duty, but now they can also use an e-hail app to hail a taxi. Yellow taxis are the only vehicles permitted to respond to a street hail from a passenger in all five boroughs in New York.
In order to ensure the datasets used in this report is sufficiently "large" to support the research goal, a total of six months of data from 2018 (approximately 4.5 GB) were downloaded. The raw data consists of more than 40 million instances with 19 attributes. It includes fields capturing pick-up and drop-off dates and times, pick-up and drop-off locations, trip distances, fare amount, types of rate, types of payment, and driver-reported passenger counts. A usage guide was provided[2] to verify the data integrity. Considering the limited computational power of local computer, Apache Spark was used to make the full utilization of the datasets became possible. As the trip data was not officially created by the TLC, their accuracy was not guaranteed and therefore any pre-processing step would be extremely important.

1.1.1 Trip Duration

The "trip_duration_min" (trip duration in minutes) attribute was derived by "tpep_dropoff_datetime" (pickup time) - "tpep_pickup_datetime" (dropoff time) attributes. As shown in Figure 2, there are several negative values lie in the plot, which is not valid in real life and should be removed. Outliers cleaning for this attribute would be done by the mean and standard deviation method.

1.1.2 Fare amount

Some values in the "fare_amount" attribute are found to be negative, and should be eliminated as well. In Figure 3, the distribution of fare amount is right-skewed with several data points lying above 20000 minutes (around 300 hours). Also, outliers cleaning cleaning for this attribute was done using the same way as the "trip_duration_min".

1.1.3 Seasons

In order to determines whether the trip was during the summer or the winter, "is_summer" feature was derived from the "tpep_dropoff_datetime" attribute. The attribute value would be 1 if the trip was during the summer (June, July and August) and vice versa.

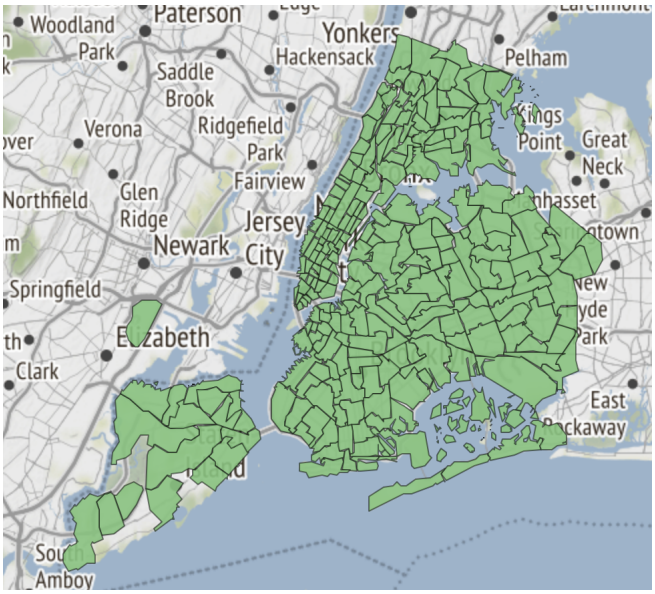

Figure 1: Research Taxi Zone in New York (Green-filled area)

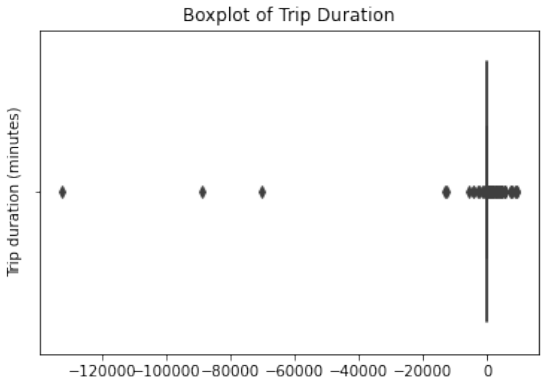

Figure 2: Boxplot of Trip Duration (minutes)

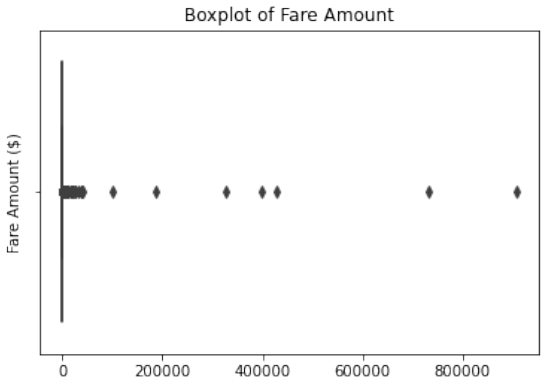

Figure 3: Boxplot of Fare Amount ($)

1.2 NOAA Weather Data (External Data)

The New York City weather data was provided by the National Oceanic and Atmospheric (NOAA)[3]. This dataset provides several metrics including air temperature, snowfall, percipitation, sunshine and average wind speed.

1.2.1 Average Temperature

The "TAVE" (average temperature) attribute was missing and thus it was manually calculated by taking the difference from the "TMAX" (maximum temparature) and "TMIN" (minimum temparature)

and divided by two.

2 Data Pre-processing

2.1 Yellow Taxi Trip Records (.csv)

1. The data types of all attributes were converted correspondingly.
2. The "trip_duration_min" (trip duration in minutes) attribute was created by subtracting "tpep_dropoff_datetime" attribute from "tpep_pickup_datetime" attribute.
3. "pickup_hour" and "pickup_day" attributes were derived from the "tpep_dropoff_datetime" attribute.
4. The "total_surcharge" feature was calculated by summing "extra", "mta_tax", "tolls_amount" and "improvement_surcharge". We then dropped those features once the "total_surcharge" attribute was created.
5. Instances containing other than payment type 1 and 2 were removed as payment type 3, 4, 5 and 6 only took up a very small portion of the whole dataset.
6. For all numerical attributes, the mean and standard deviation method was used to detect and remove the outliers. Rows containing negative values or zero were also removed. For example, trip duration and fare amount have no reason to be negative or zero.
7. "Store_and_fwd_flag" and "VendorID" attributes were removed from the dataset as they did not disclose much information for this research.
8. The clean data were then stored using .parquet format.

2.2 NOAA Weather Data (.csv)

1. Some of the columns only contain null values and thus had been removed from the dataframe.
2. All of the attributes were renamed so they are more human readable.
3. The clean weather data were then stored with .csv format.

2.3 Final Results

According to Figure 4, after pre-processing the data, it is certain that the trip duration follows a nice right-skewed distribution, where most of the data points lie on 20 minutes. It might because most of the trips were inside the CBD. In Figure 5, most of the data points lie on 25$ with several points lie on 50$. One of the possible reasons is that some of the passengers took a long trip to airports so the fare amount would be relatively higher than a short trip.

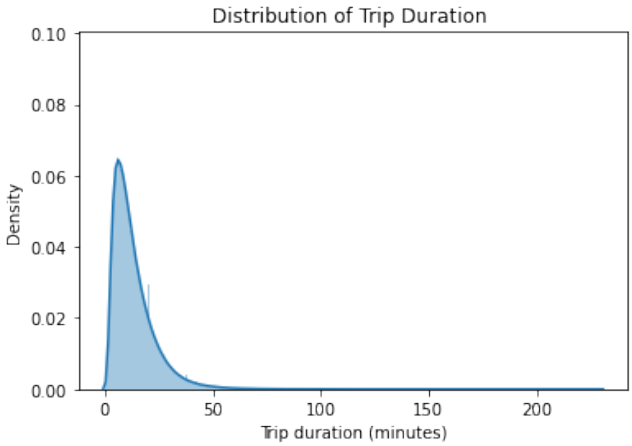

Figure 4: Distribution of Trip Duration (minutes)

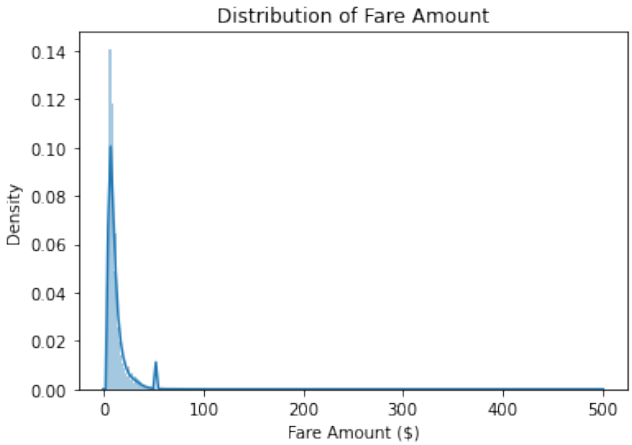

Figure 5: Distribution of Fare Amount ($)

3 Exploratory Data Analysis

3.1 An Overview on Trip Count

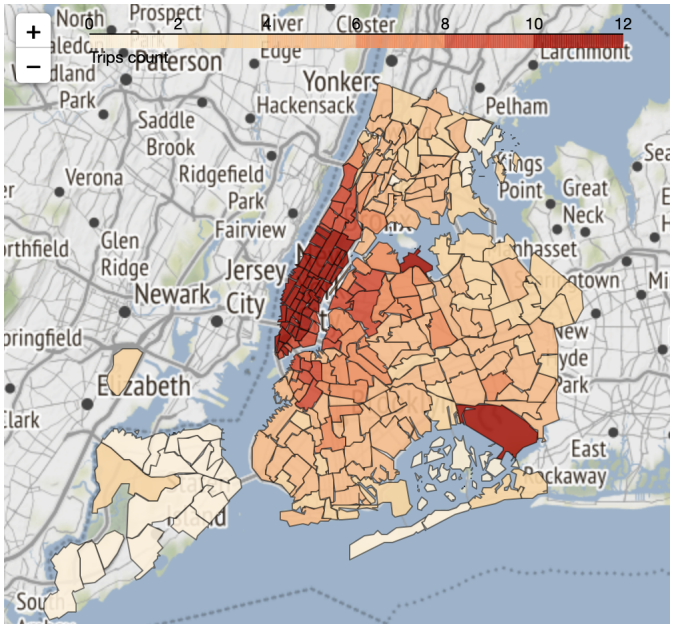

Figure 6: Total Trip Count in each Zone

In Figure 6, it is certain that most of the trips were initiated inside the Manhattan CBD, South Queens and North Queens. A brief explanation might be that Manhattan CBD is the most busy region in New York city, whilst the Southwest Airlines and John F. Kennedy International Airport located at North Queens and South Queens respectively are reasonably to have a lot of taxis hailed because of their high passenger traffic.

3.2 Time as a Factor Affecting Trip Count

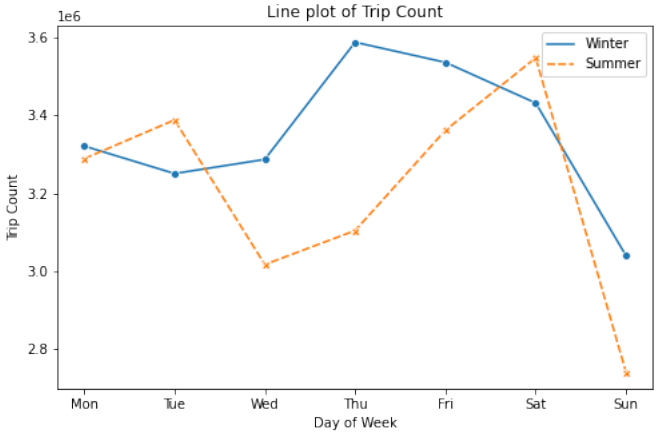

Figure 7: Summer vs Winter Trip Count

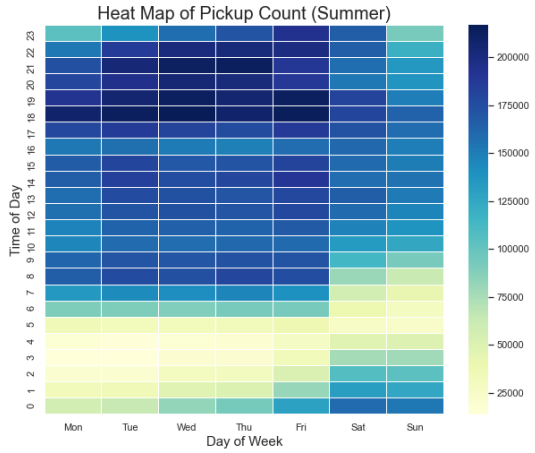

Figure 8: Heat Map of Pickup Count (Summer)

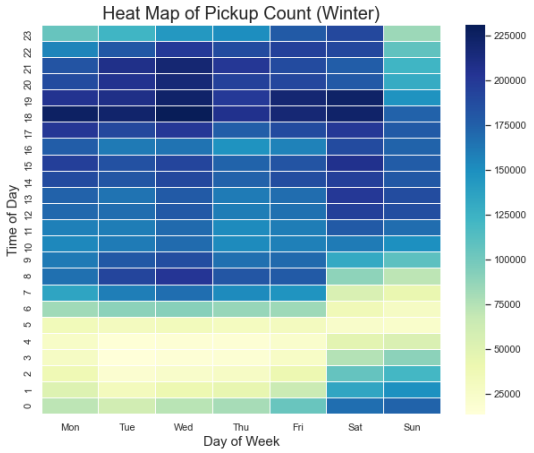

Figure 9: Heat Map of Pickup Count (Winter)

As a matter of fact, trip frequency is one of the major factors affecting profitability. As shown in Figure 7, it is clear that trip count during the summer was significantly lower than during the winter. But if we look further into the details, we found that there were more trips hailed on Saturday during the Winter than the Summer, otherwise the pattern of both heatmaps was broadly the same. According to Figure 8 and Figure 9, most of the taxis were hailed during working days especially between 5 p.m and 10 p.m One of the reasons might be that most of the white-collar workers in New York City get off their work at 5 p.m.

3.3 Taxi Zone as a Factor Affecting Profitability

After taking trip duration and trip distance into account, we derived zone profitability using the formula below:

$$\textbf{Zone Profitability} = \textbf{log}[\frac{\textbf{1}}{\textbf{2}} * [\frac{\textbf{Total Fare \& Tip}}{\textbf{Total Distance}} + \frac{\textbf{Total Fare \& Tip}}{\textbf{Total Duration}}] * \textbf{Trip Frequency per Day}]$$

The result would be the fare and tip amount gained in a taxi zone on average, given the frequency of taxi trip per day. The higher the result, the better the taxi zone quality is. In Figure 10, it had shown that the Manhattan area, and the two airports located at South Queen and North Queens have the highest taxi zone quality with approximately 1100 profitability units. After taking account of fare amount and tip amount, it was not surprised that some of the areas in Queens and Brooklyn performed well even though they did not have a high taxi frequency as shown in Figure 6.

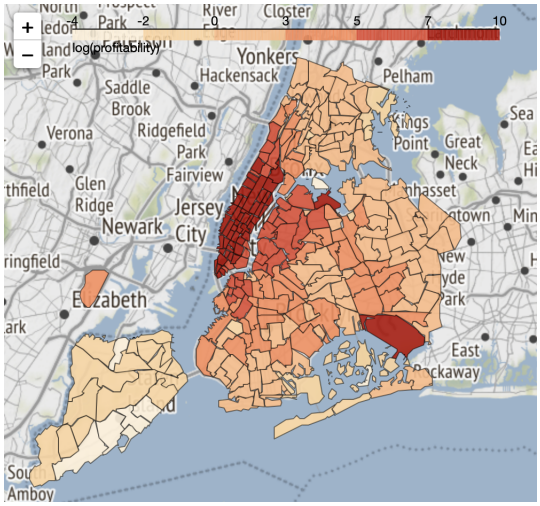

Figure 10: Zone Profitability

3.4 Attribute Analysis

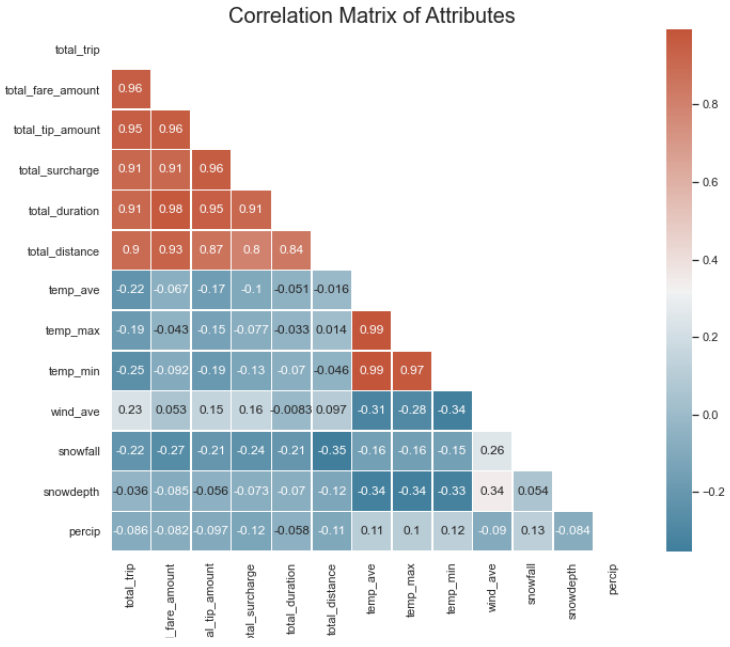

Figure 11: Attribute Correlation

To see whether the NOAA weather datasets have an effect on the original data features, a correlation for all continuous features was computed and plotted in a correlation matrix. In Figure 11, "temp_max", "temp_min" and "temp_ave" share a negative correlation of around -0.22 with "total_trip", and around -0.16 with "total_tip_amount". Apart from that, "wind_ave" has a -0.23, 0.15, and 0.16 correlation with "total_trip", "total_tip_amount", and "total_surcharge" respectively. Also, "snow_fall" has a significant negative correlation of -0.35 with "total-distance".

4 Machine Learning Modelling

4.1 Trip Profitability

The profitability of trips was calculated as:

$$\textbf{Zone Profitability} = \log[\frac{\textbf{1}}{\textbf{2}} * [\frac{\textbf{Total Fare \& Tip}}{\textbf{Total Distance}} + \frac{\textbf{Total Fare \& Tip}}{\textbf{Total Duration}}]]$$

To reduce the skewness of our original data, we log-transformed the data so the statistical analysis results from the data become more valid. We then standardised the data to make sure that it is internally consistent.

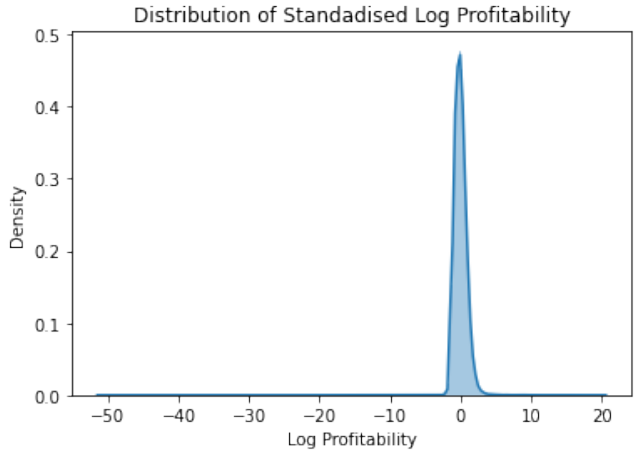

Figure 12: Distribution of log-transformed profitability (standardised)

```
count    2.295417e+06
mean    -3.461598e-13
std      1.000000e+00
min     -5.137784e+01
25%     -6.310252e-01
50%     -8.458101e-02
75%      5.088595e-01
max      2.045588e+01
Name: profitability, dtype: float64
```

Figure 13: Summary of log-transformed profitability (standardised)

Because predicting the exact profitability of every single trip might be challenging, we labelled each trip with 3 levels of quality:

1. Level -1: When the profitability was below the mean - 0.5 standard deviation of the distribution, which means the trip is not worth to be accepted by a driver.
2. Level 0: When the profitability was between the mean +- 0.5 standard deviation of the distribution, which means the trip is neutral.
3. Level 1: When the profitability was above the mean + 0.5 standard deviation of the distribution, the trip worth a lot to be accepted by a driver.

4.1.1 Test Dataset

The test dataset was subsampled from the 2019 dataset. Because the log transformation might not work for negative and zero values, they had been replaced by a very small positive number (0.00000001). Also, we had replace infinite values by a very large number (100000000).

4.2 Predicting Profitability

Considering the limited power of local computer, 5% of the dataset was subsampled to fit the model. In this project, logistic regression was considered to predict the quality of trips. Before fitting the model, some of the columns with boolean type contained NaN value and had been replaced with zero. Furthermore, features that used to derive the labels were also removed to avoid overfitting.

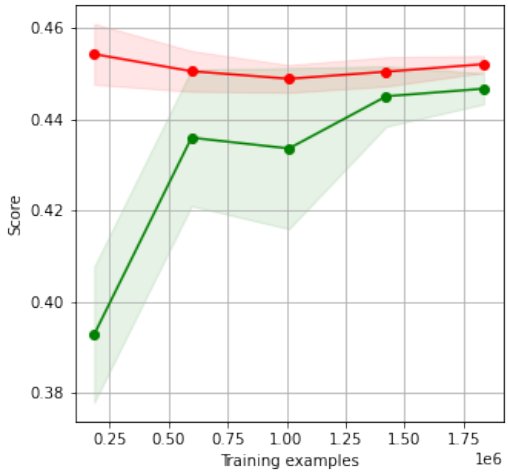

Figure 14: Learning Plot of Logistic Regression

To avoid overfitting, cross validation was used when plotting the learning curve. As Shown in Figure 14, the model did not perform very well as both training and testing accuracy were below 50%.

4.3 Final Results

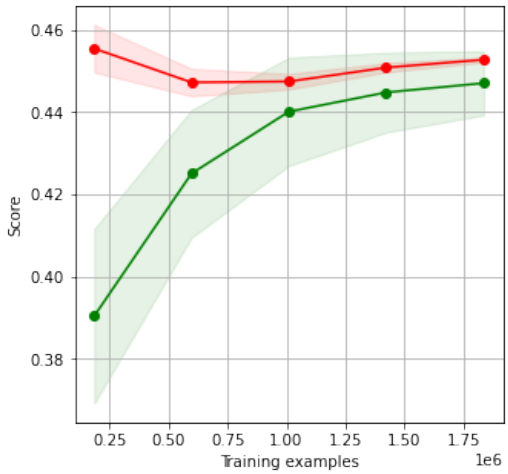

Figure 15: Learning Plot of Logistic Regression (after feature selection)

To improve the current model, feature selection using ANOVA F-value was applied and only the top 10 best features were retained. In Figure 15, after filtering 8 features, the learning curve is much smoother. However, the overall accuracy is still below 50%. The reason behind this might be that we had dropped too many features and the label was too complex. For example, profitability was derived from 4 features and we had dropped all features that were used to derive the label. Another possible reason might be that the sample size was too small and thus it did not provide sufficient information to the model when fitting it.

4.4 Error Analysis

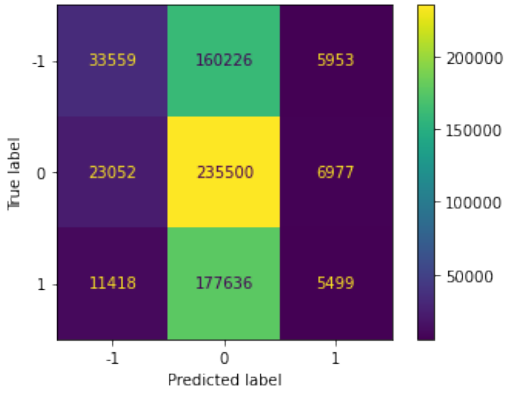

Figure 16: Confusion Matrix of Logistic Regression Model

	precision	recall	f1-score	support
Level -1	0.49	0.17	0.25	199738
Level 0	0.41	0.89	0.56	265529
Level 1	0.30	0.03	0.05	194553
accuracy			0.42	659820
macro avg	0.40	0.36	0.29	659820
weighted avg	0.40	0.42	0.32	659820

Figure 17: Classification Metrics Report

As shown in Figure 16 and Figure 17, we can see that our model had performed very well on predicting level 0 trips. However, it has a worse performance on predicting level 0 and level -1. The macro average recall and precision are 40% and 36% respectively and its weight average recall and precision are slightly higher, which is 40% and 42% respectively.

5 Recommendations

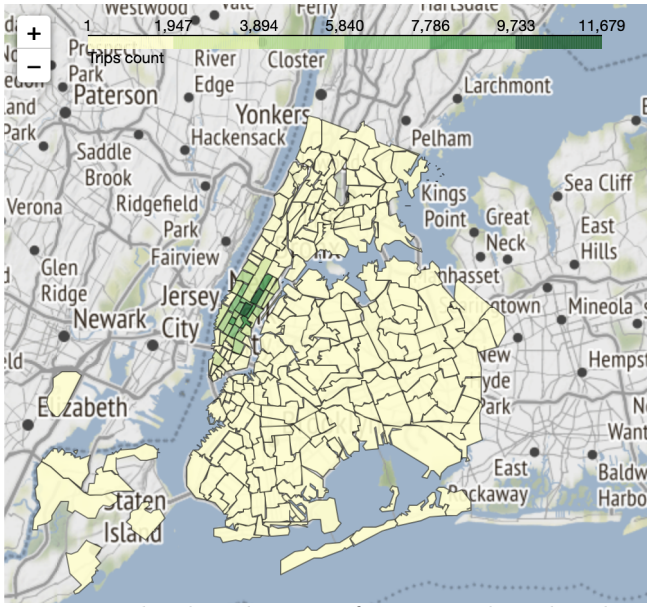

Figure 18: The distribution of trips predicted as level 1

Drivers are recommended to seek for passengers within the Manhattan area as the drivers would get the largest possible fare amount and tip amount per minute and per mile. However, because of the low performance of the model, drivers should not 100% rely on this prediction result.

6 Conclusion

This research found that the most profitable trips were mostly in the Manhattan area. Drivers within that area would get higher income than in other areas. Since this study was only based on one year dataset and had not consider other external factors, and thus it cannot guarantee that the result of this study will still be valid in the future years.

References

[1] Taxi and Limousine Commission (TLC) Trip Record Data
https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[2] TLC Trip Records User Guide
https://www1.nyc.gov/assets/tlc/downloads/pdf/trip_record_user_guide.pdf

[3] National Oceanic and Atmospheric Administration, U.S. Department of Commerce
https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094728/detail