Predictive Modeling of NYC Yellow Taxi Demand: Insights from Day, Hour, Temperature, and Location Analysis

Vanessa Gracia Tan
Student ID: 1297696
Github repo with commit

August 20, 2023

1 Introduction

Understanding and anticipating patterns across different industries such as the transportation industry has become essential for effective resource allocation, planning, and optimization in the present era of data-driven decision-making. With more data available, it is now feasible to obtain good insight and build a predictive model that can assist taxi drivers in increasing demand and improving efficiency on the road.

This study offers a thorough examination of New York City's (NYC) demand for yellow taxis. The emphasis is on a number of factors, including the pick-up time, temperature, pick-up day, and pick-up location. Our goal is to apply past data to create prediction models that can precisely predict the demand for taxis in various scenarios.

Two different Machine Learning model, **Linear Regression** and **Random Forest Regression** will be used to predict the demand for yellow taxi rides in NYC. Their evaluation will also be presented in this paper.

1.1 Datasets

The main data set is extracted from **New York City Taxi & Limousine Commission (NYCTLC) Trip Data**, which records yellow taxi data, including pick-ups and drop-offs taxi data including date and time, location, fare amount, extra, tips amount, total amount charged, etc [1]. Due to the fact that COVID-19 had severely changed the world and industry in many ways, January 2022 to December 2022 is chosen as the timeline of research. Additionally, we assume that the year 2022 is the year when everything is started to be back to normal. Therefore, this timeline will be the most relevant timeline to reflect the demand in current and future years.

In addition to this taxi data, weather data published by **Visual Crossing** [2] is used to provide more details to support the analysis as the weather is expected to correlate with the taxi demand. This data set includes date and time, temperature, precipitation, snow, wind speed, and conditions on the hourly basis.

The raw data set has 39,656,098 instances. The adjusted taxi data set (to the required period and features) has 35,624,213 instances with 7 features. The weather data set within the period range has 8760 instances with 6 features.

2 Preprocessing

This section will outline the steps taken to preprocess both taxi and weather data sets for analysis and modelling. This step is taken to ensure the data is clean, organized, and ready to be used.

2.1 Data Cleaning

This first step is identifying and handling any missing or inconsistent data. The following steps are used:

- **Update the data structure** to the most updated schema (February 2023 schema is used as reference). This is to ensure data consistency.
- **Data that are not within the date range** are removed. 56,955 data are eliminated.
- **Data with negative and extremely short trip duration** are also eliminated to meet the research and analysis requirements. Only data with trip duration above one minute is chosen as it is just logical that passenger will prefer to walk then travel for 1 minute as there is initial charge they have to pay.
- **Data with negative passenger count** are removed as this is invalid data.
- **Remove outliers** from trip distance (above 0.99 percentile and below 0.01 percentile).
- **Data with fare amount below $3.00** are removed as it is the initial charge for taxi service.
- **Data with pick-up and drop-off Location ID not in range of 1-263** are removed. 3,942,909 data are eliminated.
- **Missing value** in the weather data set for a specific date time is imputed using the mean from the previous hour and the next hour.

2.2 Feature Engineering

Creating new and transformed features to enhance the model performance for predicting taxi demand.

- **Extracting day of week** from pick-up date time in taxi data. This helps capturing the daily demand patterns.
- **Extracting hour** from pick-up date time in taxi data and weather date time.
- **Creating dummy variable** from the categorical variables such as pick-up day and pick-up borough.

2.3 Data Aggregation

The yellow taxi data set is combined with the weather data based on the pick-up date time (hourly). Furthermore, this combined data set is then inner joined with the shape file data to extract the pick-up location borough and zone.

2.4 Data Splitting

The modified and combined data sets are split into train and test subsets. The training subset is used for model training and the testing data is used to evaluate the performance of the Machine Learning model on unseen data.

- **Train-Test Split** used is 80-20 split. 80% of the data is used for training and 20% is used for testing.

In total, 4,031,885 data are removed from the raw taxi data set. The features that used in the further analysis are pick-up day, pick-up hour, pick-up location borough, pick-up zone, and temperature on that day. The other features are not included in the analysis and modelling as they are not used for predicting taxi demand. This leaves us with a total of 35,624,213 data.

3 Geospatial Visualisation

From Figure 1 and Figure 2, it can be seen that the highest taxi demand is located in Location ID 237, which is Manhattan, Upper East Side South area. Based on our analysis, this area has a total demand of 1,752,484.

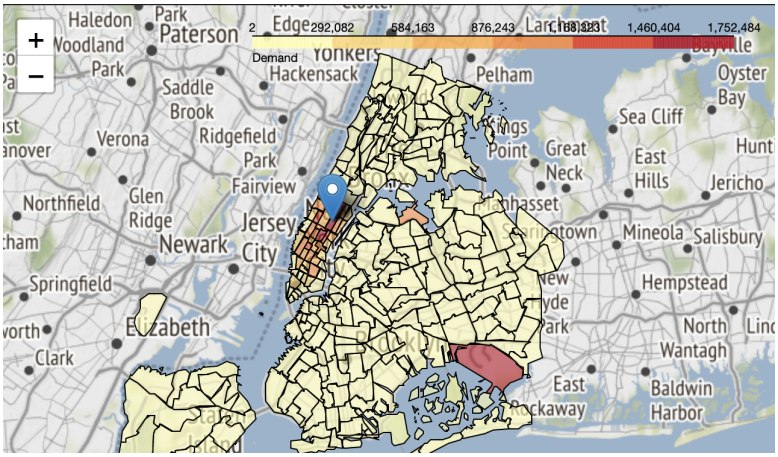

Figure 1: Zone with the Highest Demand

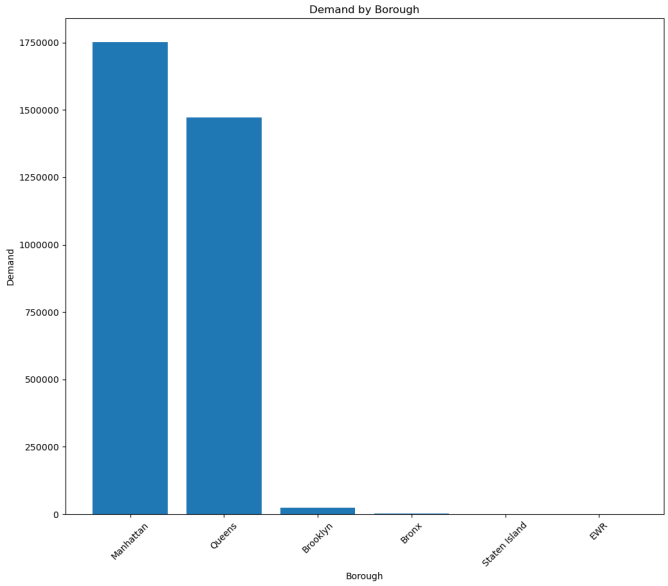

Figure 2: Distribution of Demand by Borough

4 Analysis

4.1 Day-of-Week Hourly Demand Plot

The plot presented in Figure 3 displays the daily trend in hourly demand. Across the x-axis, 24 hours (pick-up hour) are represented, while the y-axis depicts the taxi demand at each hour. The line plot provides a comprehensive view of how taxi demand is low in the morning, increases as people leave for work or school, and peaks in the evening, which is most likely the result of people returning from work or school. Furthermore, at the peak hour, it can be seen that the graph line for Sunday is not as peak as the other day, this might be due to people rather to stay at home, have rest, and get ready to start their work tomorrow. In conclusion, this figure illustrates how the time of day and hour affects the demand for taxis in NYC.

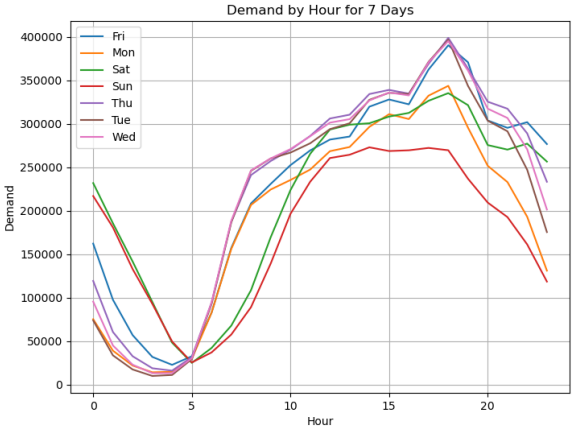

Figure 3: Day-of-Week Hourly Demand Plot

4.2 Relationship between Demand and Temperature

After plotting the data points for both demand and temperature, as y and x respectively, it can be seen from Figure 4 that temperature has a polynomial (curve) relationship with demand. As the temperature lower or higher, people are more likely to stay at home rather than going out.

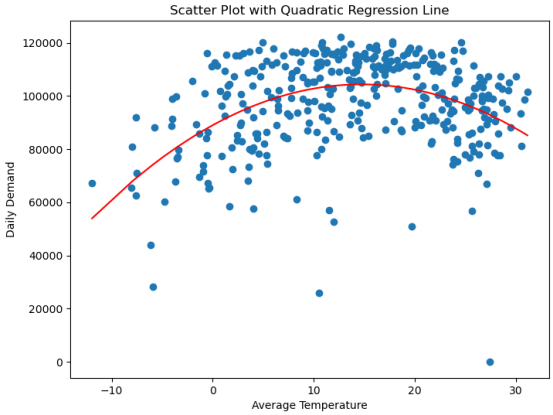

Figure 4: Temperature-Demand Relationship

5 Modelling

As numerical output is required and the attributes contain categorical and numerical variables, Linear Regression (LR) and Random Forest Regression (RFR) are used to predict the taxi demand.

5.1 Linear Regression

Linear Regression (LR) is known to establish a linear equation that captures the linear relationship between the variables. This linear equation is then used to predict the numerical output, taxi demand in this case. For our analysis, the following model is used:

$$Y_ijklm = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \xi_m + \epsilon_ijklm \tag{1}$$

where α_i is the dummy variables indicating the pick-up location, β_j is the effect of temperature on pick-up day, γ_k is the effect of temperature squared, δ_l is the effect or pick-up hour, and ξ_m is the dummy variables indicating the day of week.

The training subset is used to train this model and the coefficients are fitted to the testing subset.

5.2 Random Forest Regression

Another method used to predict this numerical output is Random Forest Regression (RFR). This method is a Machine Learning ensemble algorithm which combine multiple decision trees to make predictions. As RFR is able to capture non-linear relationships and complex relationship in data, RFR is suitable to be used for this analysis. Furthermore, it is also able to handle large data sets with no assumption regarding the mathematical relationship between the data.

This algorithm is trained using the training subset and also fitted to the testing subset.

6 Discussion

The R-squared metric is used to assess the goodness of fit of both models. As seen in Table 1, the RFR model outperforms the LR model in terms of explaining the variability in the dependent variable, with an R-squared score of 0.90 against 0.76. Furthermore, the higher R-squared value also indicates that RFR is a better model for the specific given predictive task as it can capture the underlying patterns in the data.

Model	R-squared
Linear Regression (LR)	0.76
Random Forest Regression (RFR)	0.90

Table 1: Comparison of R^2 Values for LR and RFR

7 Recommendations

The evaluation of the models indicates that RFR is better to be used to predict the taxi demand based on the pick-up day, pick-up hour, and daily temperature.

Therefore, we strongly recommend that taxi drivers increase their locations in the Manhattan, Upper East Side South area, as our analysis in Figure 1 shows that it has the highest demand. On the

other hand, Figure 3 reveals that extreme temperatures, either high or low, are associated with lower demand. Thus, when considering pick-up time and pick-up day, taxi drivers can plan their schedule accordingly and focus on peak times, such as during rush hour or after school/work hours. By following these findings, taxi drivers will be able to manage their work schedule more efficiently.

8 Conclusion

In summary, this research has evaluated two different models in order to predict the taxi demand based on the pick-up locations, daily temperature, pick-up day, and pick-up hour. By analysing past taxi data and hourly weather data, we were able to gain a better understanding of the conditions and situations that affect taxi demand. While LR model shows a relatively good R-squared metric, it can be seen that RFR model performs much better with a much higher R-squared metric.

However, future studies are highly recommended to conduct more sophisticated Machine Learning models in order to gain a much more significant and accurate predictions. This is because we believe that LR's assumption about the linear relationship between the dependent and independent variables may not be entirely accurate. Moreover, additional efforts in handling overfitting and cleaning the data sets will result in model with better performance.

References

[1] New York City Taxi and Limousine Commision. *TLC Trip Record Data*. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Accessed: 2022-08-10.

[2] Visual Crossing Corporation. *Visual Crossing Weather (2022)*. https://www.visualcrossing.com/. Accessed: 2023-08-10.