Forecasting Seasonal Taxi Demand in NYC

Do Nhat Anh Ha
Student ID: 1194034
Github repo with commit

August 25, 2024

1 Introduction

The evolution of technologies in the past decades created dynamic shifts in different industries. For instance, in the transportation landscape, ride-sharing services such as Uber and Lyft are likely to be one of the best options that come to one's mind when moving between destinations within a city, especially one that is similar to New York City. However, the presence of traditional taxi services still remains prevalent despite it no longer being among the top options.

The aim of this report is to predict the monthly demand for taxi rides in New York City (NYC) and how the season affects it. Two Machine Learning methods will be employed to carry out such task. This report adopts the point of view of different audiences, such as the ordinary citizen, with curiosity if the demand for taxis is still high, or taxi companies, with the aim of understanding and projecting demand trends.

In this report, we will refer to both the iconic yellow taxi and boro taxi as **taxi**. Furthermore, **monthly taxi trips** refers to the total number of trip records for that particular month, regardless of the region.

1.1 Dataset

The main dataset used to conduct the analysis in this paper, **TLC Taxi Trip Record Data**, is collected from the Taxi & Limousine Commission (TLC). The data includes all trip records taken by taxis, containing information about the trip such as drop-off and pickup time and venue, number of customers, and many more. As the main focus of this paper is to predict monthly taxi demands, we will use the most recent data available to us, starting from March 2023 to May 2024.

Use of external datasets was employed to aid the research, which includes the **Unemployment Rate: Balance of New York, state less New York City (URNY)**, published by the U.S. Bureau of Labor Statistics (BLS), and the US National Center for Environmental Information's **Integrated Surface Data (IES)**, collected at LaGuardia Airport.

The URNY dataset provides information regarding the unemployment rate, calculated each month, across all 5 boroughs of NYC. We believe that the unemployment rate holds some correlation with taxi demands as a high rate may reduce the need to commute between work destinations using taxis as well as the finances to afford the service. Consequently, we expect that weather conditions, from the IES data, also have an impact on the monthly demand.

2 Preprocessing

Preprocessing was carried out on all datasets to ensure they followed their expected structure. They were also transformed to be merged for the analysis and modeling in Sections 3 and 4. This section will provide information on the inconsistencies found in the datasets and how they were processed

2.1 Data Wrangling

2.1.1 TLC Trip Record Data

With the data size of 50,061,504 rows (combined), we found numerous issues regarding inconsistencies with our research requirements and the data dictionaries. Despite only a subset of features being used during analysis and model fitting, we will be using all features during cleaning to ensure data quality. The following steps are taken to clean the data:

- **Entries with vendor IDs different from 2 valid vendors** stated in the data dictionary were filtered out of the data.
- To make sure that entries are inside the designated research period, **records that fall outside of the chosen date range** were filtered.
- **Null and zero passenger count** was found in a few records and removed as invalid entries.
- Logically, the drop-off time can't be earlier than the pick-up time; however, **Negative or unreasonably short trip distance** were found and discarded. We presume that in reality, customers would choose to walk rather than get a taxi if the distance is short, thus we only included trips that were further than or equal to 0.5 miles. (Tan, 2022)
- There are only 6 predefined numeric code options for payment and rate code ID. Therefore, **trips with invalid codes for payment type and rating** were filtered out.
- We also found **invalid records of storage options of trip memories**. These are filtered to only retain 2 binary codes acting as the data flag.
- As we want to only look at demand in NYC, **records outside of the research region** (Location ID of 1-263) were discarded.
- **Records with negative amount for fare, tips, and other fees-related columns** were removed from the data entirely. For fare amount, we only kept records with a value of at least $3.00 because according to the TLC website, this is the initial fare for every trip.
- Modified Winsorizing method was used to discard trips with **extreme values in trip distance, trip duration, and fare-related fee**. This is an effective statistical method for retaining values within a user-defined percentile range. We adopt common practices and keep the values between the 1-99th percentile. The method involving using IQR was not favored as it tends to remove more values when there is more data.

In the end, **39,989,422** entries are valid and ready to use for our modeling and analysis as 10,072,082 ($\approx 20.12\%$) were discarded. As mentioned above, only a subset of the features was used for analysis and modeling, including date, location, and total fare. We also applied a log transformation on the total fare as it was heavily skewed.

2.1.2 Unemployment Rate: Balance of New York, state less New York City

The overall downloaded data is very small, consisting of 5 features which only 2 are useful to us. The features are:

- Date,
- Unemployment Rate

We also filtered the data from March 2023 to May 2024 to match our research period.

2.1.3 NCEI Integrated Surface Dataset

The preprocessing of the Integrated Surface Dataset was a bit challenging as the entries of main crucial features contain multiple values separated with commas, each holding different information related to the feature. Additionally, the values were scaled with factors associated with each feature. We were able to extract and unscale the values using the data dictionary provided. This dataset contains more than 90 features where most aren't mandatory fields, according to the dictionary. Thus, most of the fields were discarded only the following features remained:

- Date and Time
- Dew point
- Atmospheric Pressure
- Temperature
- Wind Speed

Similar to other datasets, we retain entries from March 2023 to May 2024 to match our research period. Following Tan's (2022) techniques, we impute the missing value of entries with the data from the previous hour. (p.3)

2.2 Feature Engineering and Data Aggregation

As this paper aims to predict monthly taxi demands, we extracted the months from date-related features across all datasets into a separate column.

For the TLC Trip Record Data, we grouped the data entries by pickup location ID, season, and month, counting the total number of trips and calculating the average of the log of the total amount. We also engineered a feature that indicates seasons of the year as we speculate that season might have a temporal effect on taxi demand.

A similar approach was used on the Integrated Surface Dataset. We grouped the data by month and computed the average of the remaining features, representing the monthly average of weather-related metrics.

As the URNY data is recorded monthly, there was no need to aggregate it. Ultimately, we joined the dataset on the months.

3 Data Analysis

3.1 Pickup Demands Distribition

As we suspect that the pickup location might have some geographic link with taxi demands, we analyse the distribution of monthly taxi demand across all locations. From Figure 1, Manhattan, LaGuardia Aiport, and John F. Kennedy International Airport are areas with high trip demands.

NYC is considered a bustling city with many business visitors as well as tourists coming every year. According to the website Road Genius's New York Tourism Statistic, in 2023, New York welcomed approximately 62.2 million visitors, more than 6 million and 30 million visitors in 2022 and 2021,

respectively. Indeed, not all visitors came to NYC via plane but it does provide adequate insights on why the airports are dense with taxi demands.

Many of NYC's tourist attractions such as Time Square, Empire State, Central Park, Broadway, and Grand Central Terminal all lie within Manhattan. Thus, it could possible explain why Manhattan also has high taxi demands, as it's a safe assumption that tourists would rely transportation mode such as taxi as they don't have their own vehicle.

In Staten Island, trip demands are less prevalent when compared with other boroughs. This might be due to the geographical location of the borough being quite isolated from the others.

Thus, we believe our speculation of geographical link between locations and taxi demands is valid.

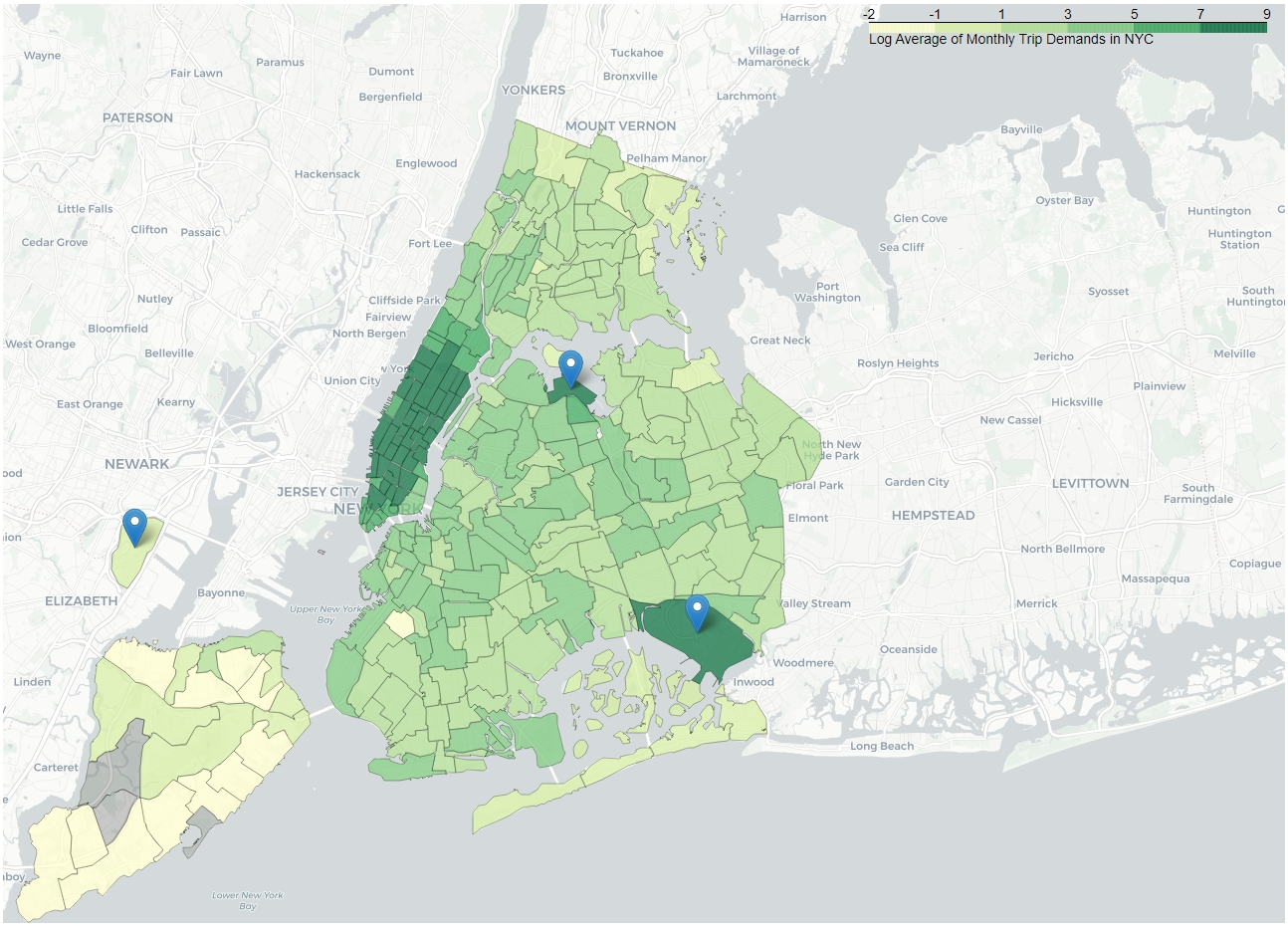

Figure 1: Distribution of Average Monthly Taxi Demands in NYC

3.2 Monthly Trips Demand Over Time

Upon inspection of monthly trip demands trend, from Figure 2, it is possible to see some trend happening. However, it's not backed up by strong evidence since the scope of our data is quite general, and we haven't collected data on monthly taxi demands from previous year. Thus, we couldn't compare if the trend of this year resembles the previous one. Though, there is still a potential indication of seasonal patterns. For instance, taxi demands declines rapidly from December. This is expected as Christmas and New Year's Eve occur during this time period, thus, most people don't go out or have to work (including taxi drivers), decreasing the need for transportation.

As previously mentioned, we believe that the unemployment rate has a socio-economical link with taxi rides. It is shown in Figure 2 that the unemployment rate was on an increasing trend, starting from April 2023. After the next month, there was a huge drop in taxi demand until it went up again in October 2023. Thus, this enhances our belief that there was a response to taxi demands when the unemployment rate went up.

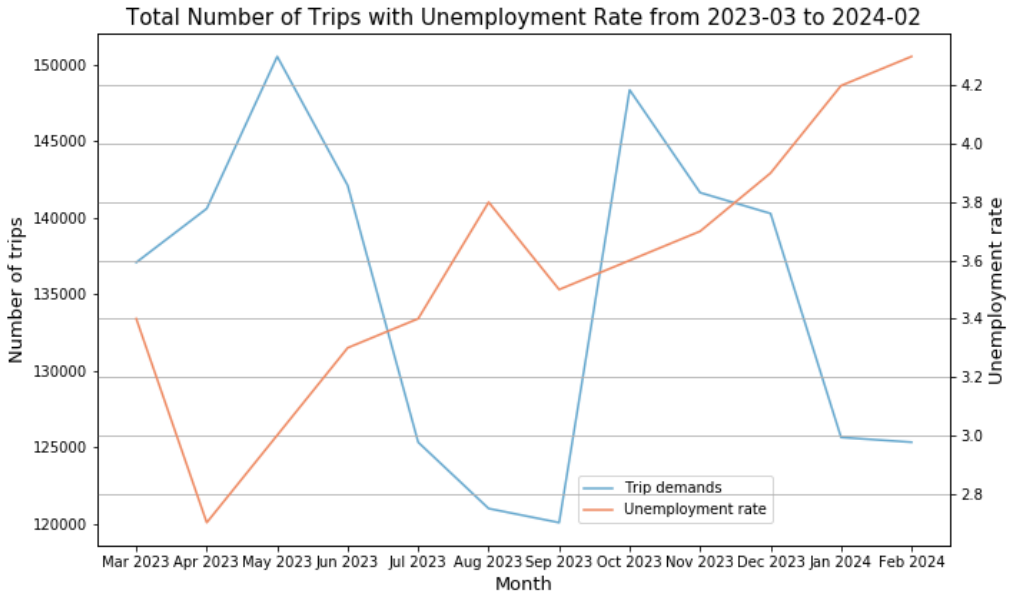

Figure 2: Total Monthly Demand

4 Modelling

We will use 2 regression models, *Random Forest Regression (RFR)* and *Generalised Linear Model*, to predict monthly taxi demand and evaluate their performance.

Due to our research interest, our data was reduced from 39,989,422 to simply **2,914** rows after data aggregation. Since only 8 features were used to predict taxi demands, our data is lacking in the feature's space complexity and the reduction in the number of entries further decreases the complexity. Thus, we decided to use all of the train data (March 2023 to Feb 2024) on all models for training and validation. In the same manner, we used all of the test data (March 2024 to May 2024) for predictions and evaluation.

4.1 Generalised Linear Model

Upon plotting the trip demands (response variable) against other variables (predictors), we notice in most plots, the data points tend to form a pillar-like structure. We expected such a situation as the way we aggregated our data led to reductions in our data's granularity. Despite the data isn't very linear, we want to use GLM as a baseline model when comparing the model's performance. Furthermore, our response variable is a discrete number, encouraging the use of a GLM from the Poisson family. We will assume that the response variable follows Poisson distribution and there are no interactions between levels of factors.

Let Y be the taxi demands and $Y \sim \text{Po}(\lambda)$. Thus:

$$\lambda = \exp(\theta)$$

and

$$g(\mu) = g(E(Y)) = \theta = X^T\beta$$

where X is the design matrix, containing the values of our features, and β contains the main effects of season as well as the parameters of numerical features

4.2 Random Forest Regression

The Random Forest Regressor (RFR) is a supervised learning algorithm that combines multiple decision trees, each constructed from bootstrapped samples of the training data, to perform regression. (Tan, 2022). This method can handle both non-linearity in the data as well as categorical variables, which is beneficial for us due to the pillar-like structure of our data points.

Since we don't have a complex dataset, we think we can use this to our advantage and create many decision trees without worrying about over-fitting too much. Having more decision trees would help improve our model's accuracy. Furthermore, since we have location ID (1-263) as a categorical feature, we can use the hyperparameter **number of bins** to help us capture the influence of this feature. As the main goal is to predict future taxi demands, we are not concerned with the interpretability of the model.

A 3-fold cross-validated grid search was performed to identify good values for our hyperparameter. Due to restricted computational power, we only fine-tuned 3 hyperparameters: numbers of bins, number of trees, and tree's depth.

5 Results and Discussion

We will use the common evaluation metrics for regression models, **Root Mean Squared Error (RMSE)**, to assess and compare the models' performance. The RMSE is the squared-rooted average difference between the predicted and actual values. It also has the same unit as the response variable which is good for interpretability. Additionally, we will use R^2 to measure the amount of variability in the data that was explained by the models. Table 1 shows the performance of each model.

RMSE		R^2	
GLM	**RFR**	**GLM**	**RFR**
2935.21	18620.28	0.9879	0.5129

Table 1: Performance of models on test data

From Table 1, it is clear that GLM was far superior to the RFR in both the RMSE and R^2. The RMSE of GLM shows that it has fewer prediction errors than RFR. Also, an R^2 of 0.9879 indicates that a substantial amount of variation in the test data was explained by the GLM. However, this could potentially be an indicator of overfitting. Nonetheless, Figure 3 shows that the predicted demands for GLM aren't that close to the actual demands. Thus, we should interpret the R^2 metrics with caution. If GLM was used to further project future demands in June July, we expect the RMSE will increase due to potential overfitting.

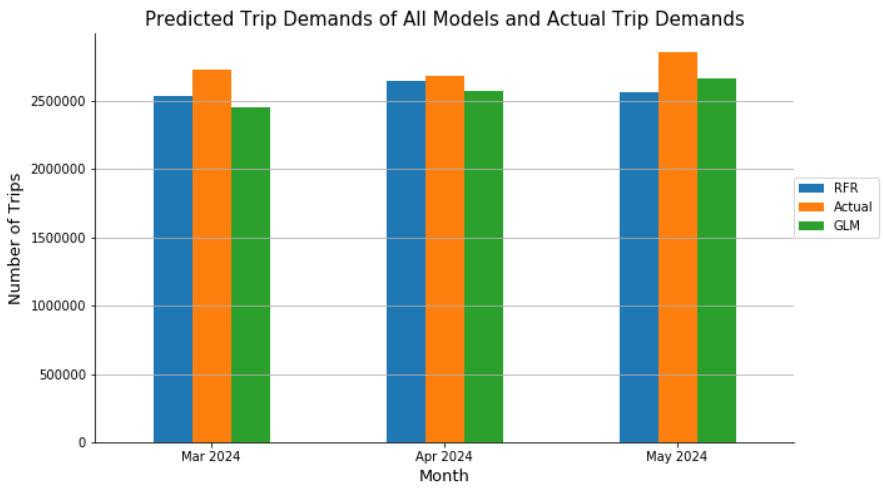

Figure 3: Predicted Demands of Model and Actual Demands

Despite being bested by GLM in performance metrics, RFR's performance should not be underestimated. For instance, in Figure 3, RFR's predicted demands for March and April 2024 were closer to the actual demands than those of GLM. The occurrence of high RMSE might be due to the big difference in predicted and actual demands in March and May 2024. Nonetheless, RFR's predictions are rather impressive, considering that only half of the variation in the data was captured by the model. We speculate that RFR did a better job in capturing the seasonal as well as the temporal effects on taxi demands. If RFR were used to predict June and July demands, we expect it to perform better than GLM.

6 Recommendations

Based on our model evaluation, we can see that GLM provides more consistent predictions on taxi demands than RFR. Though, it may be susceptible to overfitting. On the other hand, RFR generalises quite well upon new data which might be more beneficial in the long run.

These models are quite simple, thus we recommended that taxi companies should consider adopting these models within their internal system to have their own forecast on taxi demands. Of course, they shouldn't use the same models that were presented in this paper. Instead, it's recommended to improve and build on top of these models for stronger predictive power in the long run. In the end, we have only forecasted the 3 months and we expected to see the performance of the model to deteriorate when used to project demands further into the future.

As for the ordinary citizens of New York, perhaps traditional taxis are still going quite well!

7 Conclusion

In this report, we adopted 2 Machine Learning models to predict monthly taxi demands in NYC based on previous data. Some external data, socio-economic and weather data, were used to supplement the analysis and modeling. RFR didn't capture many of the variations in the data but generalised quite well, and GLM provided consistent predictions and captured most of the variation in the data.

For areas of improvement for this research, it's recommended to explore more external datasets to increase the number of features for prediction. Furthermore, a wider range of data should be collected as it is believed to provide more information and better predictive power. It is also advised that the model's hyperparameter tuning should be improved, assuming that computational power is resourceful.

[1, 2, 3, 4, 5, 6, 7, 8].

References

[1] New York City Taxi and Limousine commission. *TLC Trip Record Data*. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Accessed: 2024-08-15.

[2] U.S. Bureau of Labor Statistic. *Unemployment Rate: Balance of New York, state less Nyew York City (URNY)*. https://data.bls.gov/dataViewer/view/timeseries/LAUBS360000000000003. Accessed: 2024-08-15.

[3] National Centers for Environmental Information. *Integrated Surface Dataset*. https://www.ncei.noaa.gov/access/search/data-search/global-hourly?startDate=2023-03-01T00:00:00&endDate=2024-05-31T23:59:59&bbox=40.963,-74.257,40.463,-73.757&pageNum=1. Accessed: 2024-08-15.

[4] Road Genius. *New York Tourism Statistic*. https://roadgenius.com/statistics/tourism/usa/new-york/. Accessed: 2024-08-24.

[5] AnalytixLabs. *Random Forest Regression - How it Helps in Predictive Analysis*. https://medium.com/@byanalytixlabs/random-forest-regression-how-it-helps-in-predictive-analytics-01c31897c1d4. 2023-12-26.

[6] Heejung Shim. *Exponential Families - Semester 2 2024 MAST30027 Lecture Notes*.

[7] Heejung Shim. *General Linear Model 1 - Semester 2 2024 MAST30027 Lecture Notes*.

[8] Ming Hui Tan. *Predicting Hourly Demand for Taxi Rides in NYC*. Accesed: 2024-08-15.