Determinants of Yellow Taxi Demand in New York City

Muhammad Usman
Student ID: 1229086
Github repo with commit

August 25, 2023

1 Introduction

The Yellow Taxi has been an iconic vehicle in New York City (NYC) since **1937**. To operate legally in the city, each taxi needs a **medallion**. Before **2015**, having a medallion was like owning a treasure. In the **1970s**, they were priced at around $50,000. By **2014**, their value had soared to over a million dollars. However, the rise of ride-hailing apps like "Uber" and "Lyft" led to a significant drop in medallion prices. This shift left many taxi drivers in heavy debt. Today, there are over 13,000 yellow taxi medallions in the city, but there's a growing feeling among drivers that the iconic taxis might not have a bright future. While many drivers have switched to ride-hailing apps, some are still searching for ways to pay off the big loans they took out to buy those once-valuable medallions.[1]

Most drivers operate taxis by leasing a medallion, paying around $100 per 12-hour shift.[1] Thus, it's necessary for them to understand the factors that shape taxi demand in New York City. This paper focuses on several factors/determinants that may affect the demand. The insights will assist drivers in maximizing their earnings, considering the period of time, prevailing weather conditions, and several other factors.

2 Dataset

The main dataset chosen for this paper is **TLC Taxi Trip Record Data**, published by the **NYC Taxi & Limousine Commission**.[2] We will be using only the Yellow Taxi data from Jan-2022 to May-2023. This timeline was selected given that the dynamics of almost all industries have changed post-Covid. For this research to remain relevant, it must be based on current world dynamics.

To further aid our research goal, we will use the daily summary of weather reports taken at **JFK airport** from Jan-2022 to May-2023. This data is publicly available, courtesy of the **US National Center for Environmental Information's Integrated Surface dataset**.[3] Weather is expected to be a significant determinant in human behavior, and therefore, it might influence taxi demand.

2.1 Data Range

As previously mentioned, this paper focuses on post-Covid data. To ensure that none of the months in our selected timeline (Jan-2022 to May-2023) were still affected by Covid, we plotted a bar graph showing the total trips for each month. As per Figure 1, Jan-2022 appears to have a significantly lower number of trips compared to other months; hence, data from this month was removed. Although Feb-2022 seems to have a slightly fewer total number of trips compared to the general trend, when compared to its counterpart in Feb-2023, there isn't much difference between them. Hence, our revised

timeline for our analysis and modeling will be from **Feb-2022 to May-2023** that includes **53,378,584** instances of taxi trips.

Figure 1: Total number of trips from January-2022 to May-2023

We did not remove Jan-2022 from the **weather** data because it will be removed automatically during an inner join on 'date' with the taxi data.

3 Pre-processing

This section focuses on the techniques and business logic used to cater the missing values and existing outliers in our data.

3.1 Weather Data

This dataset was well-formatted. However, the date was of 'String' datatype, so it was cast to the 'date' type to allow for a seamless join with the taxi data. It was found that the feature **"Peak Gust Time"** was completely missing, so it was removed. After that, we were left with no missing values.

3.2 Yellow Taxi Data

3.2.1 Schema Inconsistencies

There wasn't a consistent schema across all months. Feb-2023 had a schema much closer to the data dictionary. As a result, the schema for all other months was cast to match that of Feb-2023.

3.2.2 Missing Data

Initially, we had a total of **1,725,465** records with missing values. All records with a missing **Passenger count** were removed because imputation didn't make sense in this context. However, it was found that removing just those records addressed our missing data issue for the entire dataset. Thus, it was evident that records with missing passenger counts were indeed erroneous

3.2.3 Outlier Detection

Before beginning feature selection to retain relevant features for our area of study, it was necessary to identify and address outliers across all features. Outlier detection was conducted sequentially. **Note**: The percentages reported reflect the data remaining after each detection step.

- **Trips lasting more than 5 hours** were considered as extreme cases. According to various reputable web mapping platforms, traveling from one end of NYC to the other takes about 3-4 hours. Consequently, **0.116%** of the records were removed.
- **Records with negative trip duration**, where the drop-off time was earlier than the pickup time, were presumed to result from incorrect taxi-meter entries. Thus, these **0.043%** of instances were discarded.
- **Location IDs outside the range of 1-263** were excluded. The focus of this paper is on taxi demand within NYC. Hence, any trips that began or ended outside NYC were treated as potential outliers. This led to the removal of **1.760%** of the data.
- **Instances with Passenger Count of 0 or less** were removed, suspecting that this was either due to incorrect reporting by the driver or an error in the data entry. As a result, **1.912%** of the records were eliminated.
- **Trips with distances of 0 or less** were either seen as errors in the taxi-meter that failed to record distance for those trips or perhaps were trips that were canceled before starting. **1.116%** of these instances were discarded.
- **Records with a negative money amount** could have been adjusted using imputation. However, given that they comprised just **0.656%** of our dataset, and considering the large volume of data at our disposal, it was deemed suitable to eliminate these instances.
- **Dates outside of our timeline** present in the months of Feb-2022 to May-2023, accounting for a mere **0.002%** of our data, were considered to be recording errors on the part of **TLC**.

After addressing missing values in Section 3.2.2 and conducting outlier detection in Section 3.2.3, we removed a total of **8.546%** of the data. We were left with **48,816,707** instances.

4 Feature Selection

4.1 Yellow Taxi Data

Given the paper's focus on factors influencing taxi demand, features solely related to trip expenditure were excluded. Additionally, we introduced two new features, **Day of Week** and **Hour of Day**, derived from **Date and Time**. Hence, the following features from the Yellow Taxi data were given attention:

• Pick-up Location ID • Drop-off Location ID • Date and Time • Day of Week • Hour of Day

4.2 Weather Data

From the weather data, we retained nearly all features. However, we decided to exclude the minimum and maximum temperature as they were presumed to be closely correlated with the Average Temperature. Thus, we retained following features:

• Date • Average Temperature • Average Wind Speed • Precipitation • Snowfall

5 Analysis

This section delves into analyzing taxi demand based on several factors that were retained/generated in Section 4.

5.1 Day Type

Day type often influences human outdoor activity. Factors such as work schedules, cultural practices, personal habits, recreational activities, and even some religious activities are associated with different days of the week. Therefore, the combination of these elements, particularly in relation to day type, may have a direct or indirect effect on taxi demand in NYC. As indicated in Figure 2, taxi demand remains significantly high from Tuesday through Saturday, with Thursday seeing the highest demand. This could be because many NYC employees receive their fortnightly salaries on Fridays, while others get paid on Wednesdays.[4] Thus, leading to more outdoor activities around these days. Furthermore, as depicted in Figure 3, it was anticipated that weekdays and weekends would differ in taxi demand, and this is supported by the data: the average number of trips on weekends is considerably lower than on weekdays.

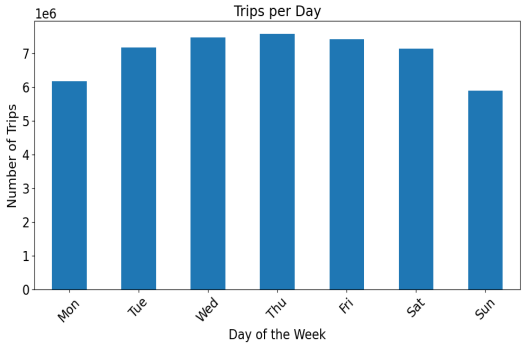

Figure 2: Total trips for each day

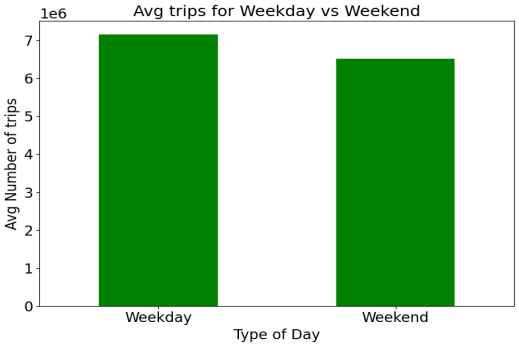

Figure 3: Avg trips for weekday vs Weekend

5.2 Hourly Trend

Having observed the effect of Day Type on taxi demand in Section 5.1, it was pertinent to delve deeper into the relationship of **Temporal Factors** with taxi demand. Consequently, the relationship between the hour of the day and the number of trips was illustrated in Figure 4. This demonstrates that taxi demand significantly varies based on the time of day. As highlighted by Figure 4, taxi drivers

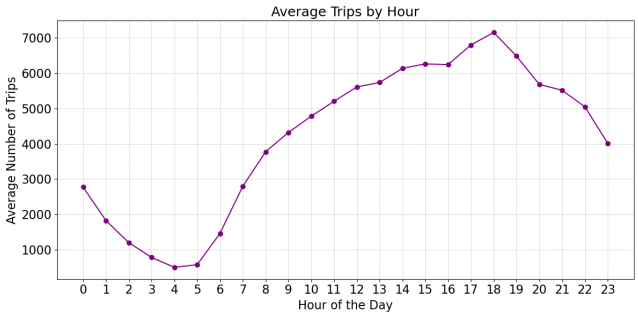

Figure 4: Average number of trips per hour on each day

should ideally remain active from **2:00 pm - 7:30 pm**, as during these hours, NYC sees an average of over **6,000** taxi trips per hour each day.

5.3 Zones and Boroughs

NYC is divided into **263** zones, encompassing a total of **6** boroughs. Some zones are airports, others are business hubs, while some are primarily residential areas. The unique characteristics associated with each zone play a pivotal role in influencing taxi demand. As depicted in Figure 5, only a selected few zones account for the majority of the high taxi demand. This was somewhat surprising given that NYC is one of the busiest cities in the world [5], yet most of its activities are concentrated in specific areas.

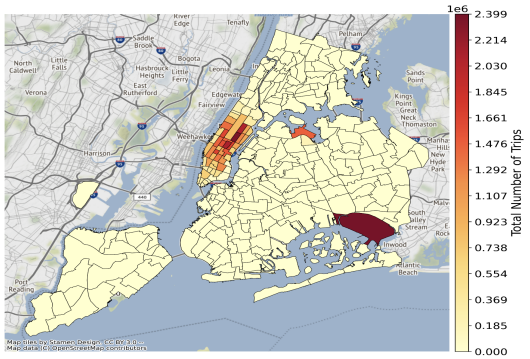

Figure 5: Total number of trips by zones

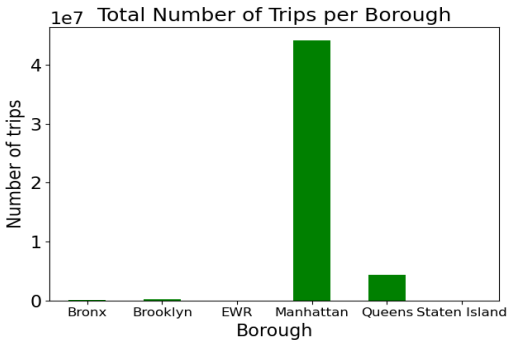

Figure 6: number of trips started in each borough

Further investigation revealed that these high-demand zones predominantly belong to two boroughs: **Manhattan** and **Queens**, as outlined in Table 1 and illustrated in Figure 6. It's noteworthy that **Queens** is highlighted mainly due to JFK airport. Apart from that, the bulk of the activities and taxi demand is centralized in **Manhattan**. Compared to these two, the demand from the other four boroughs seems to be negligible.

Total Trips	Borough	Zone
2,398,512	Queens	JFK Airport
2,374,387	Manhattan	Upper East Side South
2,111,290	Manhattan	Midtown Center
2,097,776	Manhattan	Upper East Side North
1,696,672	Manhattan	Midtown East
1,675,273	Manhattan	Penn Station
1,652,865	Manhattan	Lincoln Square East
1,586,271	Manhattan	Times Sq

Table 1: Top 8 zones with highest taxi demand

5.4 Weather

Weather significantly impacts various human activities and social behaviors.[6] Therefore, it was expected that weather elements would have a strong correlation with taxi demand. However, according to Figure 7, no such strong linear relationship is evident. All elements exhibit a correlation of less than **0.2** with total trips. Notably, of all the elements, snowfall has a negative correlation, which was anticipated.

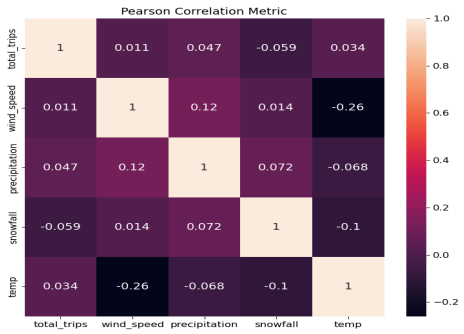

Figure 7: Correlation heat map between several weather elements and total trips

6 Modelling

We will utilize **Linear Regression (1)** as our initial approach and the more advanced **Random Forest Regressor (2)** to predict taxi demand. Analysis in Section 5 indicated that temporal factors and location types have a strong influence on taxi demand in NYC. Despite the results from Section 5.4 showing no strong linear relationship between weather elements and taxi trips, we will incorporate them in the modelling. We anticipate that in conjunction with other features, their significance might emerge. Furthermore, based on results from table 1 and figure 5, we will limit our study to only those trips that originated in **Manhattan** or at **JFK Airport**. Based on our findings thus far, we believe that it may not be as lucrative for a taxi driver to operate in other areas.

Thus, we have the following predictors to forecast total trips:

- Hour of Day
- Pick-up Location ID
- Precipitation
- Snowfall
- Average Wind Speed
- Average Temperature
- Weekday

Both models discussed in Sections 6.1 and 6.2 were trained using data spanning from **February 2022** to **January 2023**. Subsequently, these models were utilized to make predictions for the period from **February 2023** to **May 2023**. This approach resulted in a dataset split of **75%** for training and **25%** for testing.

6.1 Linear Regression (LR)

Linear regression is well-suited for datasets combining both categorical and continuous variables. In our model, we did not include any interaction terms, operating on the assumption that our categorical predictor (Weekday) are independent of continuous ones. Furthermore, interactions among continuous predictors were not considered, so as to examine their direct impact on taxi demand.

Based on Table 2, the results from the linear regression indicate that, except for average temperature, all these predictors exhibit a slight linear relationship with our response variable "total trips" at a **5%** level of significance.

	Coefficient	$P > \|t\|$
Hour of Day	3.6289	0.000
Pick-up Location	0.2076	0.000
Avg Wind Speed	0.2111	0.000
Precipitation	0.0966	0.000
Snowfall	−0.0761	0.008
Avg Temperature	0.0072	0.567
Weekday	5.6618	0.000
Intercept	−10.1892	0.000

Table 2: Linear Regression Results

6.2 Random Forest Regressor (RFR)

The Random Forest Regressor is an ensemble learning supervised algorithm suited for regression tasks. Given the limitations of linear models in capturing non-linear and complex relationships between features, we opted for this approach to predict "total trips" using the predictors outlined in Section 6. This regressor naturally handles a mix of continuous and categorical predictors. Moreover, as observed in Section 5.4, our weather elements weren't strongly correlated with taxi trips. Yet, this regressor, capable of implicitly performing feature selection and generating uncorrelated decision trees, becomes valuable. Its method of leveraging outputs from various trees and averaging them helps strike a balance in the bias-variance trade-off.[7] Unlike Linear Regression, the RFR is less interpretable. However, given our target audience of taxi drivers, they may not be particularly concerned with the model's interpretability.

To obtain the best hyperparameters, we conducted a cross-validated grid search on the training data. Through a more involved process, we identified hyperparameters that exhibited strong performance on both training and test data, while also narrowing the difference in their performance. This strategy helped us avoid overfitting.

6.3 Models Comparison

We adopted the **Root Mean Squared Error (RMSE)** as our evaluation metric for model comparison.

As indicated by Table 3, the **LR** model yielded consistent scores for both training and test data. This was anticipated because linear regression is resistant to overfitting. In contrast to the **RFR**, the **LR** model underperformed. With the **RMSE** of **LR** exceeding **75** on both training and test datasets, and the **RFR**'s **RMSE** being below **27** for both, it becomes evident that a mere linear relationship is insufficient to accurately represent the connection between taxi demand and the determinants presented in Section 6. Instead, there appears to be a complex or non-linear relationship that the **RFR** captures more effectively.

	Training	**Testing**
Linear Regression	75.808	78.504
Random Forest Regressor	22.317	26.839

Table 3: Root Mean Squared Error on training and testing data

Additionally, given that the **RFR** surpassed the **LR** in performance, we further examined how the **RFR** fared on our test data by plotting the actual versus predicted trips over the course of a day. According to Figure 8, the **RFR** slightly overestimated trips during daytime hours. However, since the predicted and actual counts were closely aligned, this deviation is not of significant concern.

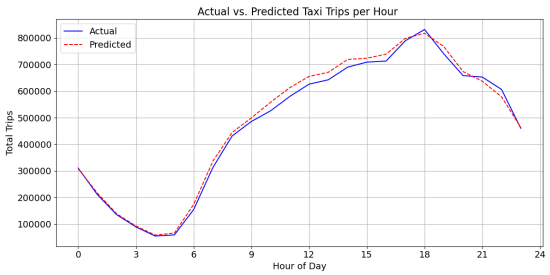

Figure 8: Actual & predicted trips for Manhattan and JFK aiprort

To delve deeper we plotted separate line graphs for **Manhattan** and **JFK airport**. Figure 9 indicates that for **Manhattan**, the trend mirrored the overall data trend presented in Figure 8. Conversely, Figure 10 for **JFK airport** is somewhat alarming, as it deviates from the overall pattern. Notably, the **RFR** appears to **underestimate** during peak hours but is accurate during off-peak periods. This discrepancy could be attributed to having less data for JFK airport compared to Manhattan, potentially compromising the **RFR**'s predictive accuracy for JFK airport.

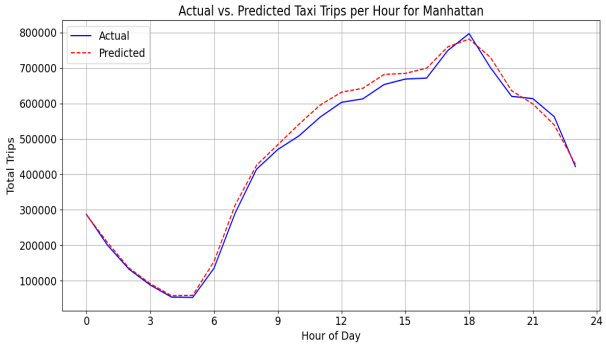

Figure 9: Manhattan's actual & predicted trips

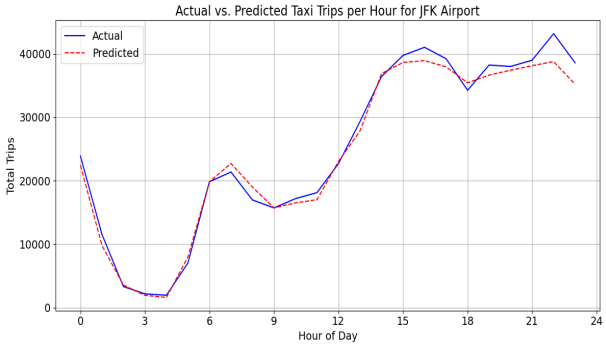

Figure 10: JFK airport actual & predicted trips

7 Recommendations & Conclusion

After studying the relationship between various potential determinants of taxi demand, we recommend drivers, rather than searching for trips randomly throughout NYC, to primarily focus on Manhattan and JFK airport. Beginning the day at JFK airport and choosing trips destined for nearby locations or Manhattan is recommended. Once in Manhattan, drivers should concentrate their efforts there throughout the day. Our findings supporting this recommendation can be referenced in Table 1, Figure 5, and Figure 6. The significance of location as a determinant is also corroborated by the **LR** output in Table 2.

For optimal returns, taxi drivers should capitalize on the peak hours from **2:00 pm to 7:30 pm**. For drivers who have leased a medallion for a 12-hour shift by paying $100 [1], a recommended timeframe would be **10:00 am to 10:00 pm**. Starting extremely early or working late into the night isn't typically necessary. This guidance is supported by data in Figure 4 and Figure 8.

Furthermore, as evidenced by Section 6, temporal factors and weather conditions significantly influence taxi demand. Typically, weekdays see a higher demand than weekends. While snowfall tends to negatively impact demand, other weather conditions generally have a positive effect on taxi services.

Transportation firms and government agencies like the **TLC** [2] would find the insights from Section 6 significant. To forecast demand, it's beneficial to use models that can capture complex and non-linear relationships. An example of such a model is the **RFR**, as detailed in Section 6.2. Besides this, neural networks and statistical models, especially Generalized Linear Models, can be considered for predictive purposes.

Although this report primarily focused on Yellow Taxis, the methodology can be adapted for other modes of transportation, including Green Taxis and ride-hailing app services.

References

[1] Cecilia Saixue Watt. *'There's no future for taxis': New York yellow cab drivers drowning in debt.* https://www.theguardian.com/us-news/2017/oct/20/new-york-yellow-cab-taxi-medallion-value-cost. Accessed: 2023-08-24.

[2] NYC Taxi & Limoussine Commission. *TLC trip record data.* https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Accessed: 2022-08-19.

[3] National Centers for Environmental Information. *Integrated Surface dataset.* https://www.ncei.noaa.gov/access/search/data-search/global-hourly. Accessed: 2022-08-20.

[4] NYC Office of Payroll Administration. *Frequently Asked Questions About Pay.* https://www.nyc.gov/site/opa/my-pay/my-pay.page. Accessed: 2022-08-20.

[5] 10MostToday. *The 10 busiest cities in the world.* https://10mosttoday.com/the-10-busiest-cities-in-the-world/. Accessed: 2022-08-25.

[6] Astha Gupta. *How weather can affect your mental state, mood and behaviour.* https://www.dailytelegraph.com.au/lifestyle/health/body-soul-daily/how-weather-can-affect-your-mental-state-mood-and-behaviour/news-story/65619a69d326f869a97b7a452c5b82f6. Accessed: 2022-08-25.

[7] Jagandeep Singh. *Random Forest: Pros and Cons.* https://medium.datadriveninvestor.com/random-forest-pros-and-cons-c1c42fb64f04. Accessed: 2022-08-25.