Analysis of Ride-Hailing Services in NYC

BOUJIDA Nezar

21/10/2023

Abstract

The study delves into the intricacies of ride-hailing services in New York City, focusing on platforms such as Uber, Lyft, and Via. Through an analysis of trip records from the Taxi and Limousine Commission (TLC), spanning January to May 2021, the study aims to unravel the dynamics of fare pricing, demand surges, and borough classifications.

Dataset Overview

The analysis in this study is based on a comprehensive dataset obtained from Kaggle, specifically the "Uber NYC For-Hire Vehicles Trip Data 2021" dataset. This dataset provides detailed information about Uber trips in New York City during the year 2021, including trip dates, times, pickup and drop-off locations, trip distances, and fare prices.

The dataset offers insights into the operational dynamics of Uber services in NYC, enabling us to delve into various aspects such as fare pricing, demand patterns, and geographical trends. This dataset provides a robust foundation for conducting a thorough analysis of ride-hailing services in one of the world's busiest urban environments.

The complete dataset can be accessed via the following link: `https://www.kaggle.com/datasets/shuhengmo/uber-nyc-forhire-vehicles-trip-data-2021`

Contents

1 Introduction

In New York City's vibrant transportation ecosystem, ride-hailing platforms such as Uber, Lyft, and Via play a crucial role, offering a flexible alternative to traditional taxi services and enriching urban mobility. This study uses trip records from the Taxi and Limousine Commission (TLC) spanning January to May 2021, focusing on these platforms.
We aim to dissect the dynamics of fare prices, investigate the factors they depend on using predictive modeling, analyze demand surges, forecast demand trends, and categorize NYC boroughs by their unique characteristics.

2 Problem Statement and Achieved Results

In the fast-paced world of ride-sharing services, the ability to accurately predict fares and understand customer demand is pivotal for maintaining high levels of customer trust and satisfaction. Equally important is the nuanced approach to operational management, specifically in how drivers are allocated to meet this demand efficiently. When considering the diverse landscape of cities like New York, with its distinct boroughs each harboring unique demands and preferences, the challenge becomes even more complex.

To overcome thes challenges, this study introduces a specialized decision tree regressor model designed to predict fare prices accurately, leveraging features such as time, distance, and location. The model achieves an R-squared value of 0.81, indicating its ability to explain approximately 81.28% of the variance in fare prices based on the selected variables. Additionally, to forecast demand accurately, we employed the XGBoost algorithm. This approach gave an R-squared (R^2) score of 0.93, along with a Root Mean Square Error (RMSE) of 2249.67741, demonstrating the model's ability to provide reliable demand forecasts. Lastly, through classification, we successfully grouped the boroughs based on their demand, price per mile, and waiting time, offering valuable insights for strategic decision-making and resource allocation.

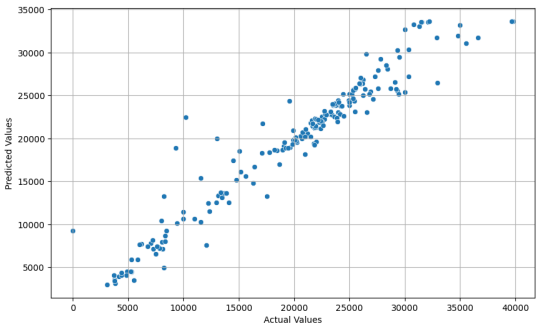

Figure 1: Actual vs. Predicted Values for the Forecasting Model

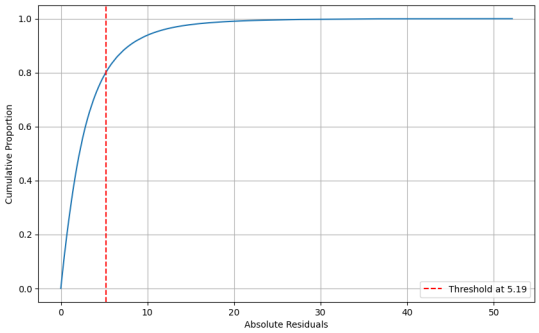

Figure 2: CDF of Absolute Residuals for the Fare Price Prediction Model

Figure 1 illustrates that the predicted values closely align with the actual data, indicating the forecasting model's accuracy. Figure 2 demonstrates that approximately 80% of the predictions have an error margin between -5.19 and 5.19, highlighting the model's precision in estimating fare prices.

3 Comparative Analysis Across Platforms

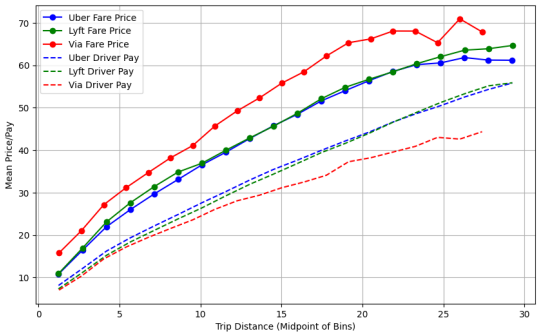

Figure 3: Mean Fare Price and Driver Pay Across Distance intervals

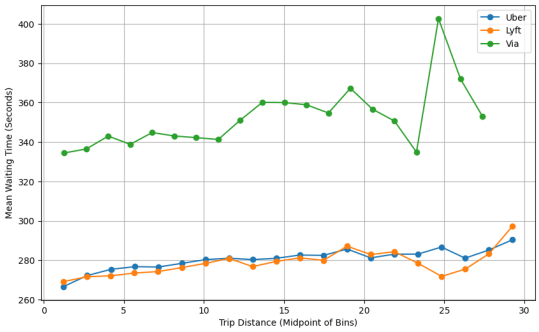

Figure 4: Waiting Time Per Distance interval

The figures presented compare the operational metrics of Uber, Lyft, and Via, with specific attention to mean fare price, driver pay, and waiting times segmented by distance bins. The distance bins partition transfer distances into equal-width intervals, facilitating a granular analysis of each platform's performance metrics.

Figure 3 depicts both the mean fare price and driver pay across various distance bins, highlighting the competitiveness of Uber and Lyft in terms of offering both higher compensation to drivers and lower fare prices to customers compared to Via. This competitive edge is a significant factor in Uber maintaining a dominant 72.5% market share, followed by Lyft with 27%, and Via lagging significantly at only 0.5%. Figure 4 focuses on the waiting times per distance bin for each platform. The analysis reveals that Via often has the highest waiting times, further exacerbating its less favorable positioning in the market. This, coupled with Via's higher fare prices and lower driver pay as illustrated in Figure 3, paints a comprehensive picture of Via offering the least desirable service among the three platforms.

4 Fare Price Prediction

4.1 Assessing Departure and Drop-Off Location Impact on Fare Price

The process of feature extraction is pivotal for accurate fare price prediction. By analyzing the probability density of fare prices per mile, we extract valuable ratios that represent distinct aspects of fare pricing dynamics. The *End Location Ratio*, highlighted in blue, offers a narrow and steep distribution, indicating limited variability and most fares clustering near a mean between 6 to 8. This implies that drop-off locations have a predictable impact on fare prices. Conversely, the *Departure Location Ratio*, in orange, has a broader distribution suggesting higher variability for pick-up locations, making it an essential feature for capturing fare deviations. Finally, the *Zones Pair Ratio*, in green, with its wide distribution, underscores the most significant variability, which is instrumental for understanding the fare structure across different travel routes. Collectively, these features are critical for developing robust fare price prediction models.

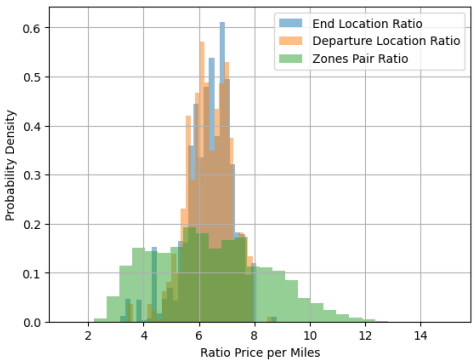

Figure 5: Probability Density Function of Ratios.

4.2 Analyzing Temporal Variations in Fare Pricing

Figure 6 delineates the temporal variations and their influence on ride-hailing fare prices. The data distinctively reveals fluctuations in the number of transport requests at various hours during weekdays and weekends. This pattern is indicative of demand variability, which in turn precipitates dynamic fare pricing. During weekends, the demand notably intensifies from 12 pm to 2 am, and this increased demand correlates with a rise in the mean fare price. Such temporal features are pivotal as they enable the formulation of predictive models that can anticipate fare adjustments based on the time of request, thereby facilitating more accurate fare estimates and better resource allocation for ride-hailing services.

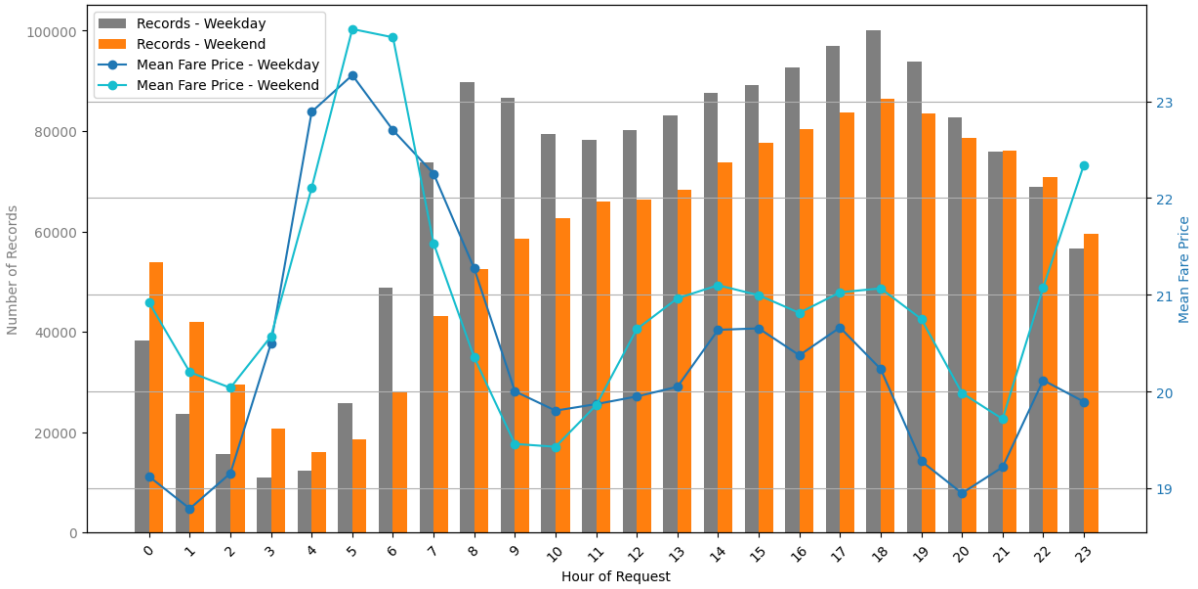

Figure 6: Hourly analysis of ride request volumes and mean fare prices during weekdays and weekends.

4.3 Model Results and Feature Importance

A Decision Tree Regressor was implemented to predict fare prices. This model was trained using features such as trip miles, hour and month of the request, week day, license number, and various ratio metrics including those for departure and end locations, as well as zone pairs. The model's performance was quantified using the Mean Squared Error (MSE) and the R-squared (R^2) metrics, giving an MSE of 27.49 and an R^2 score of 0.81, indicating a relatively high level of prediction accuracy.

The trip miles feature stands out with a staggering 90.74% importance, highlighting its critical role in fare prediction. Other features like ratio Dep location and ratio end location have a noticeably lower importance, at 2.58% and 1.33% respectively. The least influential features are related to the hvfhs license num (represents the platform uber, lyft or via), contributing very minimally to the model's predictive capability. However, when using just the trip miles feature on a decision tree regressor, the R^2 score was 0.76, so adding these features we were able to capture more information.

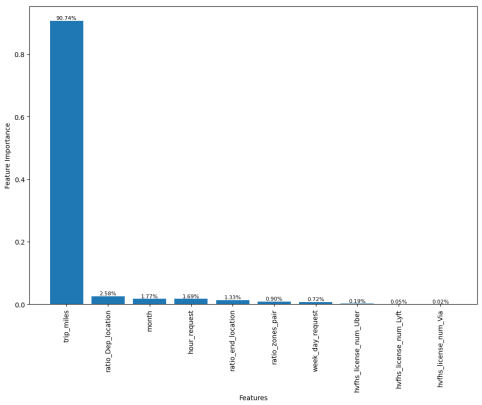

Figure 7: Feature importance as determined by the Decision Tree Regressor for fare price prediction.

5 Time Series Analysis

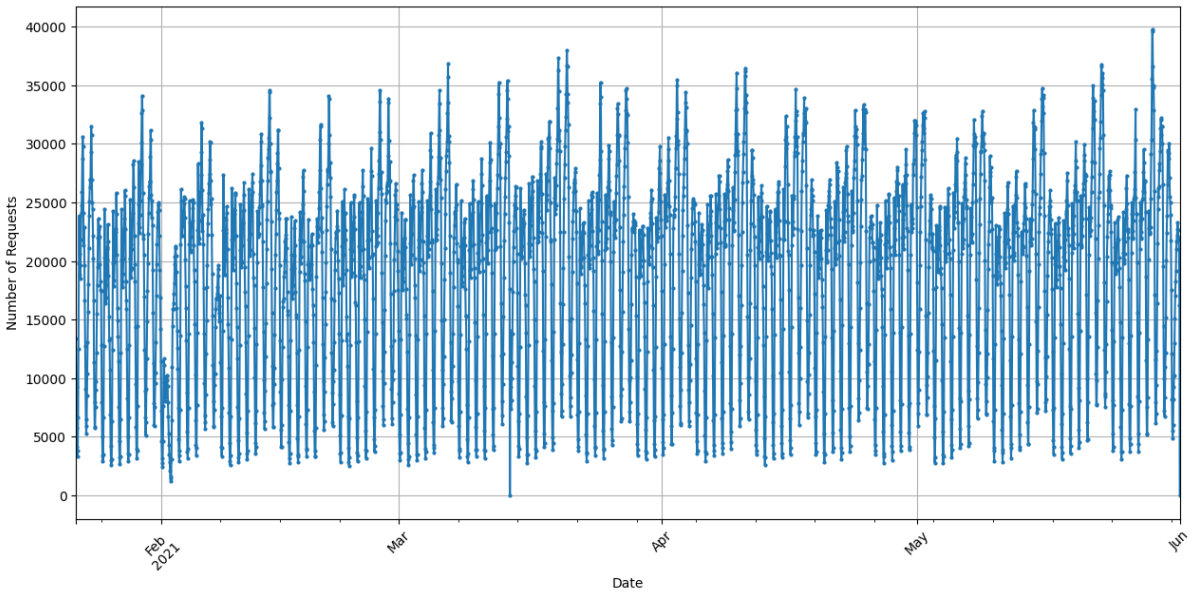

Figure 8: Time Series of the Number of Requests.

The figure above illustrates a time series graph for the number of requests from January to June. The graph exhibits substantial fluctuations, with clear peaks and troughs suggesting variability within the data.
To assess stationarity more formally, the Augmented Dickey-Fuller (ADF) test was applied to the data. The ADF test is a type of statistical test that aims to determine the presence of a unit root in a time series dataset. The presence of a unit root indicates that the series is non-stationary and has some time-dependent structure. The null hypothesis of the ADF test is that the time series has a unit root and is non-stationary.
The Augmented Dickey-Fuller (ADF) test gave an ADF Statistic of -5.526858 and a p-value of 0.000002, indicating strong evidence against the null hypothesis of a unit root. Therefore, we conclude that the time series is stationary.

5.1 Model Performance Evaluation

A gradient boosting model, specifically XGBoost, was used to forecast the number of fare requests, it was trained on a dataset split into training and validation sets, with the training set comprising data before May 24, 2021, and the validation set including data from this date onward. The model used features such as the day of the week, day of the month, hour, and lags over previous weeks, capturing both cyclic patterns and recent trends. The final model achieved a Mean Squared Error (RMSE) of approximately 2,059 , indicating a high degree of predictive accuracy.

Figure 10 displays the learning curve of the XGBoost model over 5,000 iterations. The Root Mean Squared Error (RMSE) for both the training set (blue) and the validation set (orange) is plotted against the number of boosting iterations. As observed, the RMSE decreases sharply and then gradually levels off, indicating the model's convergence. The convergence of training and validation RMSE suggests that the model has generalized well without overfitting to the training data.

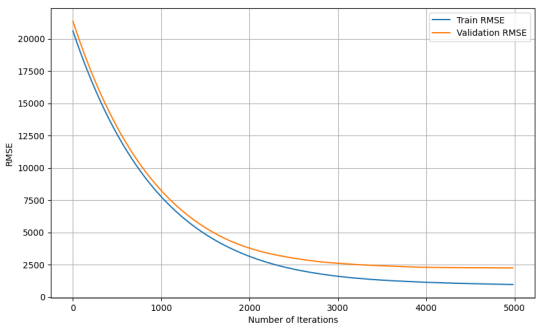

Figure 9: Training and Validation RMSE Over Iterations.

6 Boroughs Classification

6.1 Feature Extraction for Borough Classification

In predicting and classifying boroughs, we extracted features from the taxi ride dataset. For each borough, we computed the following features:
Number of Trips per Pickup Location (Departures): The count of trips originating from each borough.
Number of Trips per Drop-off Location (Arrivals): The count of trips terminating at each borough.
Mean Waiting Time per Pickup Location: The average waiting time for pickups from each borough.
Mean Ratio of Price to Miles per Pickup Location: The average ratio of fare price to distance traveled for pickups from each borough.
These features provide insights into the transportation dynamics within each borough, including demand, pricing, and service quality.

6.2 Classification Model

We applied the K-means clustering algorithm to classify the boroughs into 4 clusters. The choice of 4 clusters was determined using the Elbow method, which indicated that this number provided a meaningful partitioning of the data.
Each cluster exhibited distinct characteristics, providing insights into the different borough profiles:
Cluster 1: Characterized by areas with marginally lower demand, shorter waiting times, and reduced prices per mile. These are depicted in green.
Cluster 2: Denotes areas typically in the suburbs with lower fares and demand, illustrated in blue.
Cluster 3: Represents zones with elevated transportation demand, indicative of bustling routes and potentially higher fares, primarily situated near downtown areas. These are shown in red.
Cluster 4: Identifies regions experiencing moderate to slightly higher transportation demand and fares, coupled with shorter waiting times, often found in proximity to downtown areas.

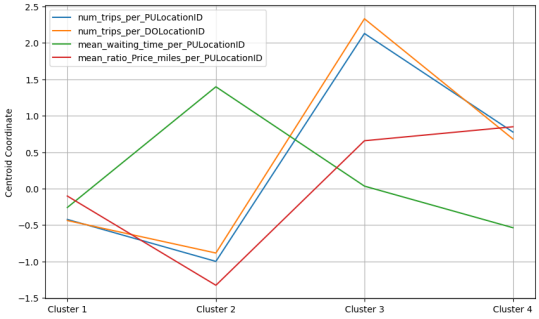

Figure 10: Centroid Coordinates by Feature.

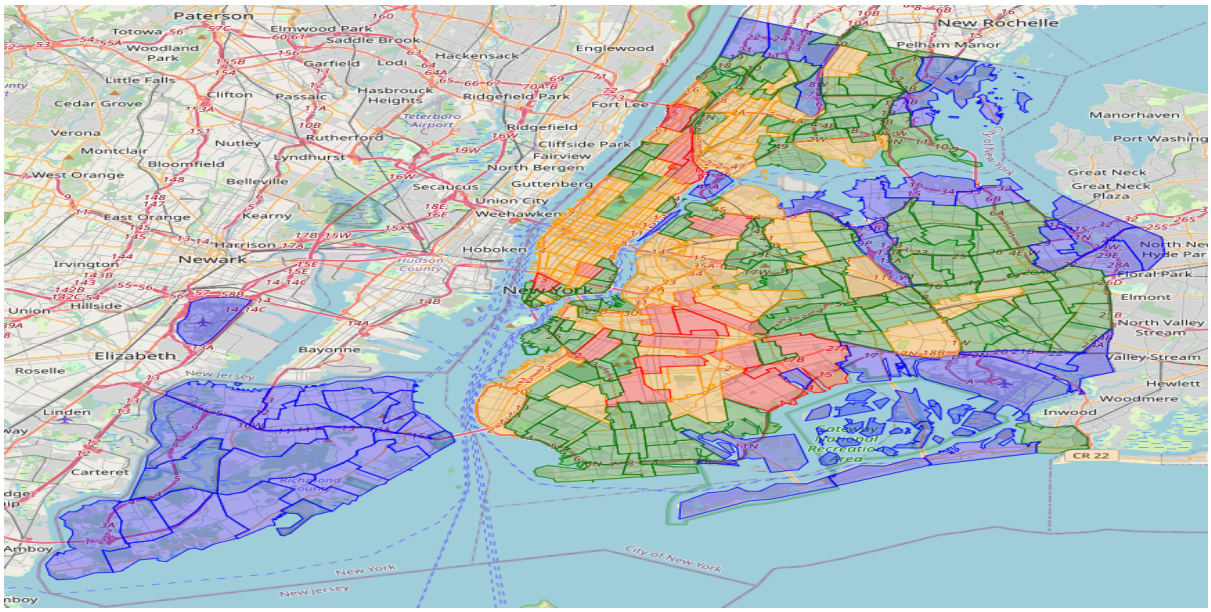

Figure 11: Map of New York City with Clusters