Statistical Analysis of Uber and Lyft

Avishek Saha and Noman Mohammad

2023-03-23

Dataset Introduction and Hypotheses

The data for this project was collected and uploaded to Kaggle by the user RaviMunde, and is licensed under CC0: Public Domain, which allows for unrestricted use and distribution of the data. The dataset includes simulated ride prices for Uber and Lyft, as well as weather conditions corresponding to the source destination for the cab rides. The data was collected for the interval from 2018-11-25 to 2018-12-18, and was obtained through custom application queries collected using the respective API's of Uber and Lyft.

The data collection process involved building a custom application in Scala to query data at regular intervals of 5 minutes for cab ride estimates and 1 hour for weather data, in order to collect as much data as possible without unnecessary redundancy. The collected data was then saved to DynamoDB, and finally stored into two separate CSV files: cab_rides.csv and weather.csv. To perform our analysis, we merged the two datasets based on the time_stamp variable to create a new dataset containing all the variables from both datasets. We opted to remove the 'id', and 'product_id' variables due to redundancy. After cleaning the merged data set, we are left with 87125 observations. The merged dataset and its description can be seen in the table below:

Variable	Description	Data Type
distance	Distance traveled from the pickup location to the drop off location in miles.	Float
cab_type	Represents the type of ride sharing service used, such as Uber or Lyft.	String
destination	Drop off location of the ride.	String
source	Pickup location of the ride.	String
price	Cost of the ride in US dollars.	Float
surge_multiplier	How much the ride price was multiplied based on current ride demand. Default surge_multiplier is 1.	Numeric
name	Type of cab used, such as Lyft XL or UberX.	String
time_stamp_date	Date and time of the ride	Long

Variable	Description	Data Type
temp	Temperature in Fahrenheit at the time the weather was recorded.	String
clouds	Percentage of cloud cover in the sky at the time the weather was recorded.	Float
pressure	Air pressure in millibars at the time the weather was recorded.	Float
rain	Amount of rain in inches for the last hour at the time the weather was recorded.	Float
humidity	Percentage of humidity in the air at the time the weather was recorded.	Float
wind	Wind speed in miles per hour at the time the weather was recorded.	Float

Our project primarily aims to investigate the accuracy of predicting ride fares based on various factors, such as the ride-hailing service provider, time of day, cab type, and weather conditions. The goal is to unravel the intricate interactions among these factors and understand their impact on the ride fare. In order to address our hypothesis, we employed both parametric and non-parametric models to delve into the underlying patterns and relationships among the variables in our dataset. Through this approach, we were able to identify which variables significantly influence ride fares and how they contribute to fare prediction.

Regression

To build an accurate and effective linear regression model, we included categorical variables such as cab type and name of the cab as predictor variables. However, we know that categorical variables cannot be directly included in regression models as they are not numerical in nature. To overcome this, we created dummy variables for these categorical variables. By converting categorical variables into binary variables, we were able to represent the presence or absence of a particular category in the variable, and include it in the regression model as a predictor variable. This approach allowed us to investigate the relationship between the cab 'name' variable and price and identify which variables were most important in predicting the outcome of interest.

Multicollinearity

Before evaluating the performance of our models, it is essential to examine the presence of multicollinearity among the predictor variables. As we know, multicollinearity usually occurs when two or more predictor variables are highly correlated, which can lead to unreliable and unstable estimates of the regression results. To assess multicollinearity we computed the correlation matrix for all of our numeric predictor variables. The correlation graph for the numeric predictor variables can be seen below in Figure 1:

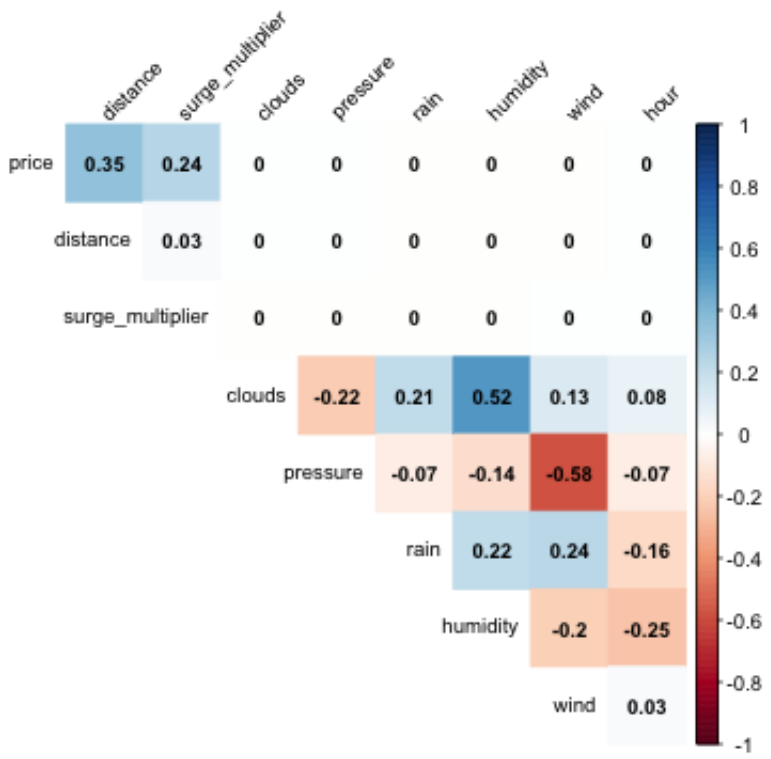

Figure 1: Correlation heatmap. Positive correlations are shown in blue, while negative correlations are shown in red, with the intensity of the color indicating the strength of the correlation.

As we can see from the correlation plot, we can observe that most of the predictor variables have low correlation values, indicating little to no multicollinearity. However, we can see that there are a few notable exceptions. Clouds and humidity have a correlation of 0.515, suggesting a moderate positive relationship between these two variables along with pressure and wind having a correlation of -0.576, indicating a moderate negative relationship.These correlations, although significant, did not impact our final model as the weather-related variables were not included in the model after the variable selection process. The variable selection methods deemed these variables as not significant in predicting the cab prices. This indicates that multicollinearity is not a major concern for our analysis, as the final model does not include the correlated variables.

Parametric Regression

We decided to fit a linear model on our data to see how well it can predict the price variable. Before doing so we also performed a 70/30 split on our data into both training and testing sets, so we can further compute the MSE for model comparisons.

The linear regression model output shows the results of our variable selection process for predicting the price of different types of cabs based on various features, such as distance, surge multiplier, and the name and type of cab. We used multiple variable selection methods, including backward selection based on AIC, backward selection based on BIC, and stepwise selection based on AIC. The final model presented here includes 13 significant predictors, including distance, surge multiplier, and various dummy variables representing the different cab names and types. These variables were found to be consistent across all three variable selection methods, indicating their importance in predicting cab prices. The overall model had an impressively high R-squared value of 0.9272, indicating that roughly 93% percent of the variance in price can be explained by our predictor variables.

In light of the findings for multicollinearity, our linear model remains unaffected by the highlighted concerns. Now that we have developed our linear regression model and performed variable selection, it is important to assess the assumptions of the model and determine whether or not they hold. In particular, we will need to examine the residual plots to check for normality and homoscedasticity.

Upon examining the diagnostic plots, we observed that the assumptions of homoscedasticity and normality were potentially violated. In Figure 2 above, we noticed a significant concentration of residuals around the center of the plot, with a discernible trend in the residuals towards one end of the plot. This suggested the presence of heteroscedasticity in the model, which could potentially result in biased and unreliable estimates. Additionally, the Q-Q plot revealed that the distribution of the residuals deviated significantly from a normal distribution, particularly in the extremes. This violation of normality assumptions could potentially impact the accuracy and reliability of the model. For reference we also computed the testing MSE for our reduced linear model using the testing set and got a result of 6.338.

Moving forward, lets try and fit our data on some models learnt in class. A generalized linear model (GLM) may be an appropriate fit for this dataset because our response variable, price, is continuous and possibly not normally distributed, as indicated by the residual plots

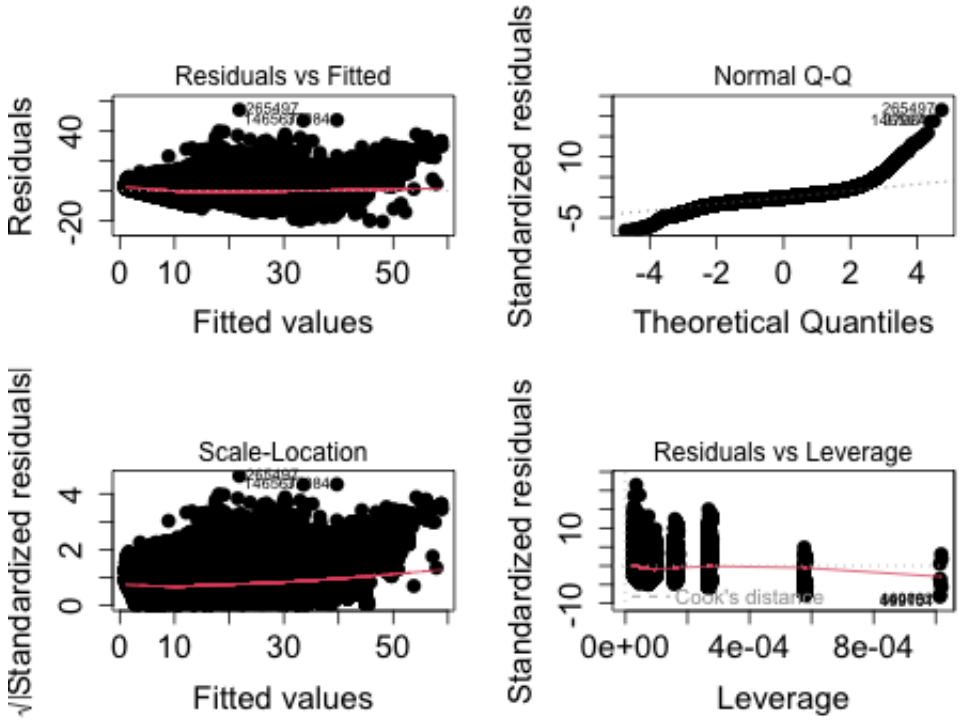

Figure 2: Diagnostic plots for linear model after variable selection. Notice a discernible pattern in the Residuals vs Fitted plot along with deiviations in thr Q-Q plot tails.

in Figure 2. GLMs allow us to model the relationship between the response variable and the predictors using different types of distributions such as Gaussian, Poisson, or Binomial, and a link function that connects the expected value of the response variable to the linear predictors.

In this case, we will use the Gaussian distribution with a log link function. This choice is due to the positive nature of the price variable and the potential presence of heteroscedasticity in the residuals. The log link function can help stabilize the variance and better capture the relationship between the predictors and the response variable. As a result, in Figure 3 we can see that the diagnostics do not have much improvement as the residual are now actually more concentrated in the middle.Most of the issues we saw in our linear model are still quite apparent. We again computed the MSE and now see it has improved, computing at 4.420.

Moving forward, from our understanding during exploratory data analysis we saw that the price variable appears to have a somewhat right-skewed distribution. The density plot in Figure 4 shows multiple peaks concentrated on the left side of the graph, indicating that most of the observations are clustered around lower prices, with relatively few high-priced observations. This observation is further supported by a skewness value of 1.046, which is substantially greater than 0, confirming the right-skewed nature of the distribution. Therefore, in light of the issues with the Gaussian distribution, we will explore another option for modeling our data. Given that the price variable is positive and continuous, a Gamma distribution with a log link function could be a better fit for our dataset. The Gamma dis-

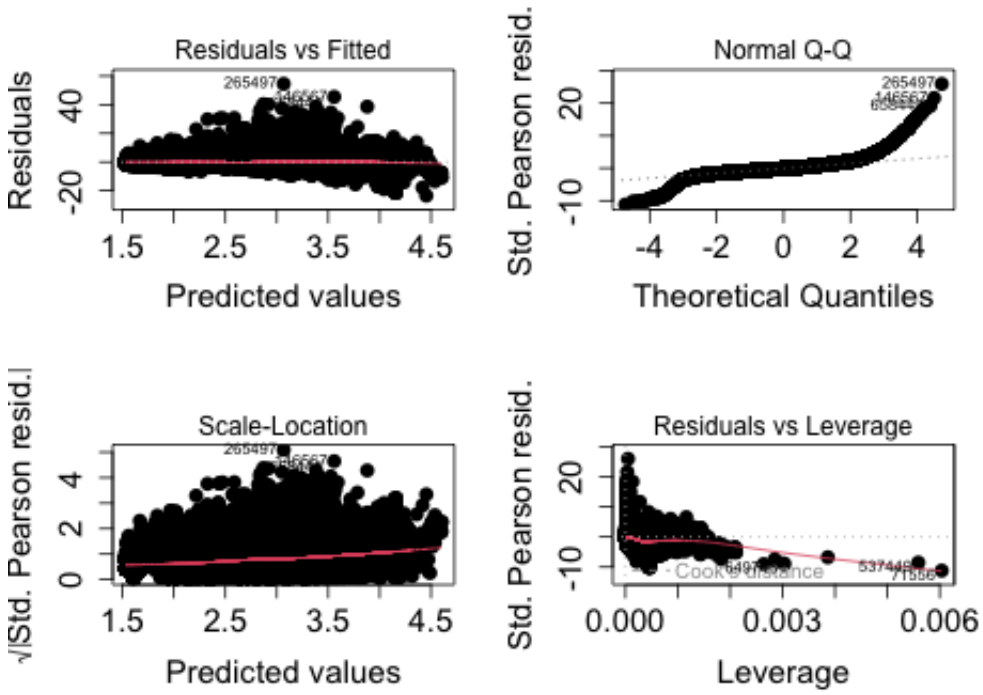

Figure 3: Diagnostic plots for GLM with gaussian family and link = log. We can clearly see some violations as the residuals are mainly clustered in the center. There is also siginificant deviation on both tails in the Q-Q plot.

tribution is particularly suitable for modeling positive continuous variables that may exhibit a right-skewed distribution, such as our price variable.

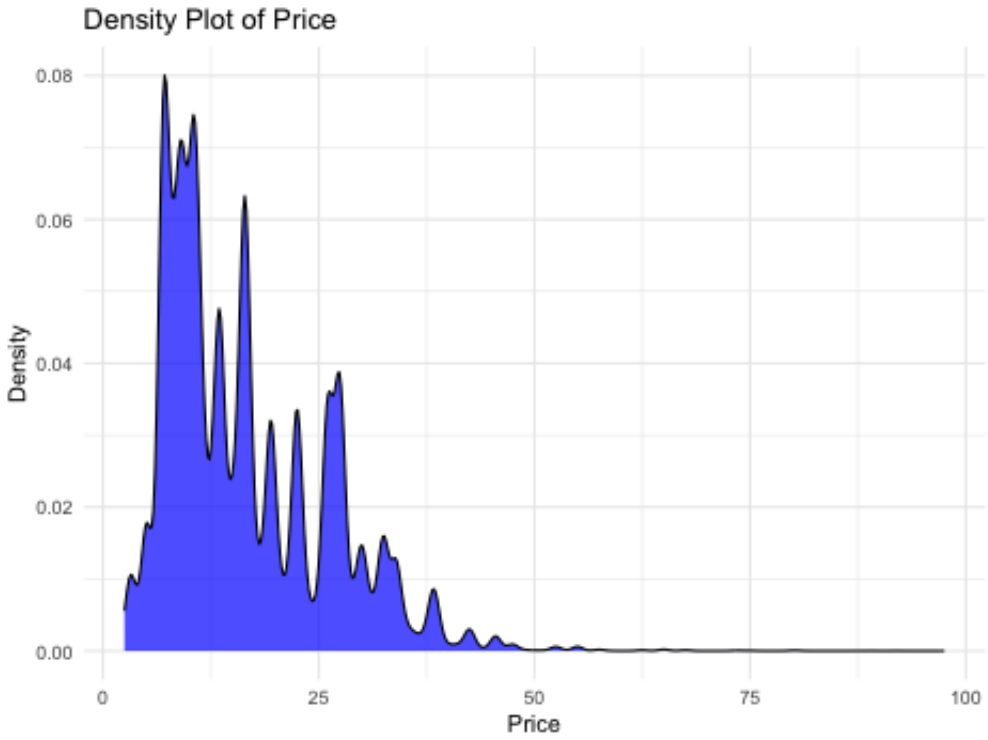

Figure 4: Density plot of the price variable, revealing a right-skewed distribution with multiple peaks concentrated on the left side of the graph.

As we can see in Figure 5 After fitting the Gamma GLM with a log link function, we noticed improvements in the residual diagnostics. The QQ-plot of residuals for the Gamma GLM shows less deviation in the bottom tail, indicating an improvement in the normality assumption compared to previous models. However, despite these improvements, the mean squared error for the Gamma GLM still comes out to be slightly higher at 4.874 when compared to the Gaussian GLM with log link model. While the Gamma GLM does show some improvement in terms of residuals, it may not necessarily provide the best predictive performance based on the MSE values.

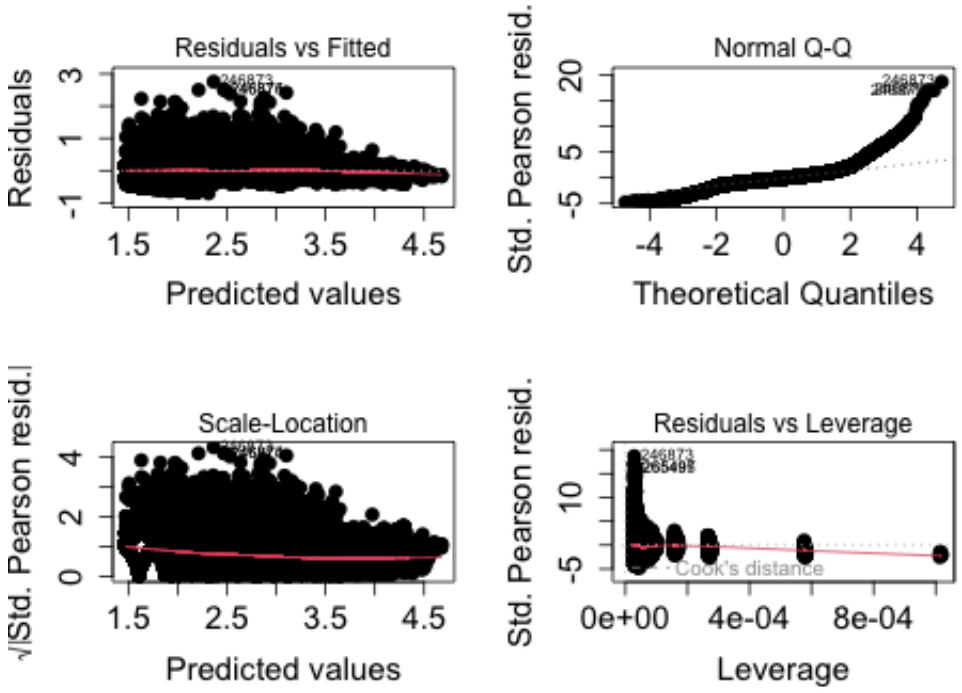

Figure 5: Diagnostic plots for a Gamma distributed Generalized Linear Model (GLM) fitted to the data. The plot shows a noticeable improvement in the residuals vs fitted values plot, along with much less deviation in the bottom tail of the QQ-plot, indicating an improvement in the normality assumption.

Non-parametric Regression

To explore some other models, we aimed to assess whether our linear regression model was capturing the complex relationships between the predictor variables and cab price. To achieve this, we initially attempted to evaluate the fit of the model using a non-parametric test called npcmstest. We know that the npcmstest compares the performance of a parametric model, in our case the linear regression model, to a non-parametric model. We ran it using the predictor variables selected in our variable selection such as, distance, surge_multiplier, cab_lyft, nameBlack_SUV, nameLux, nameLux_Black, nameLux_Black_XL, nameLyft, nameLyft_XL, nameUberPool, nameUberX, and nameUberXL along with the response variable price. Unfortunantely, after running our code for around 13 hours without any results, we decided to move on to exploring other modeling approaches, including Random Forest.

Overall, the purpose of wanting to do a npcmstest was to see if our parametric model was misspecified and whether it could capture the complex relationships that existed in the data.

Although our linear regression model provided a relatively high R-squared value, indicating a good fit to the data, it may not fully capture the complexity of the relationships between the predictor variables and cab price. To address this limitation, we decided to fit a Random Forest model, which is a powerful non-parametric model capable of handling non-linear relationships.

The primary objective of the first part of our hypothesis was to attempt to predict the cab price, and in doing so, we evaluated the predictive power of different models including Random Forest. As a model that doesn't necessarily prioritize interpretability but rather focuses on accuracy we chose Random Forest as a potential model for our analysis. However, despite our initial expectations, the Random Forest model had a higher MSE of 9.68 compared to other models such as linear regression, GLM Gaussian (log link), and GLM Gamma (log link). Although it lacked in terms of answering our hypothesis, we still find value in the insights provided by the Random Forest model, particularly in highlighting the complexity of the relationships between the predictor variables and cab prices.

Conclusion

Key Findings: The Role of Car Type and Distance in Ride Fare Prediction

Our project aimed to investigate the accuracy of predicting ride fares based on various factors, with the ultimate goal of unraveling the intricate interactions among these factors and understanding their impact on ride fares.Our analysis revealed that car type alone has a considerable impact on ride fares with 67.35% of the price variation being explained by this factor. This may make sense, as luxury cars typically have higher fares than economy cars due to differences in service quality, comfort, and amenities, no matter the weather conditions or distance travelled since this is shows to be the main driving force behind the price variation. Conversely, we found that distance only accounted for 11.87% of the variation in ride prices. This surprising result suggests that other factors may play a more

significant role in determining fares. Our study highlights the importance of developing a more comprehensive model incorporating additional variables for accurate fare prediction.

Addressing Our Hypothesis with Modeling Techniques

To address our hypothesis, we employed both parametric and non-parametric models to delve into the underlying patterns and relationships among the variables in our dataset. We started with linear regression and then explored more complex models, such as generalized linear models with Gaussian and Gamma distributions, to better capture potential non-linear relationships and address the issues observed in the residual diagnostics. Moreover, we also attempted non-parametric models, like the random forest, to improve our predictions. Although, the random forest model did have a higher overall mean squared error when compared to other models. Ultimately, our analysis revealed that the generalized linear model with a Gaussian distribution and a log link function best predicted ride fares in our dataset. This model yielded the lowest MSE of all the models that we fit on our data.

Potential Limitations and the Importance of Real-World Data

Although our analysis has yielded interesting results, it is essential to recognize the potential limitations of the simulated data and how these limitations might have affected our findings. As highlighted in our proposal, some key issues include the inaccurate representation of demand and supply, and the inadequate capture of human behavior, which could lead to discrepancies in our analysis. For example, the simulated data may not provide an accurate representation of the actual demand and supply for rides in the studied area. Consequently, the relationships and patterns observed in the data might not accurately reflect the real-world situation. This limitation could lead to biased results, particularly if certain car types or specific times of day are over- or under-represented in the data, which might affect the correlations we have observed between car type, distance, and ride fares.

In addition, human behavior plays a strong role in determining the demand for ride-sharing services. Factors such as commuting patterns, differences in travel patterns between weekdays and weekends, and responses to weather conditions all have an important impact on ride demand. The simulated data may not accurately capture these nuanced aspects of human behavior, which could lead to discrepancies in our analysis. For instance, the unexpected low explanatory power of distance might be attributed to unrealistic travel patterns in the simulated data that do not adequately represent actual rider behavior.

All in all, our analysis has uncovered intriguing findings, but it is crucial to recognize the potential limitations of the simulated data and their impact on our results. By combining our insights with real-world ride-sharing data, we could refine our understanding of the relationships between all of the variables. This fusion of simulated and actual data would allow for the development of more reliable price prediction models.