Predictive Modeling for Uber and Lyft Prices: A Machine Learning Approach Using Boston Dataset

Ashuthosh , Nanda Vihari,Bringesh,Vraj.

Abstract:

We built a model to predict ride-hailing prices based on 57 factors. Four machine learning algorithms were tested: Multiple Linear Regression, Polynomial Regression, Decision Trees, and Random Forests. Random Forests emerged as the winner, outperforming others in accuracy. This finding can help ride-hailing companies optimize pricing strategies and improve customer experience. Our study provides valuable insights into the complex interactions between predictor variables and ride prices.

1. Introduction:

As the ride-hailing industry continues to reshape the urban transportation landscape, understanding the intricacies of this market has become increasingly important. Boston, a city renowned for its rich history, cultural attractions, and thriving economy, presents a fascinating case study. Our research focuses on a unique dataset of ride-hailing transactions, collected from November 26, 2018, to December 18, 2018, in the New York time zone. This three-week period, spanning the winter season, offers a captivating snapshot of the market dynamics during a time of year marked by fluctuating weather conditions and holiday festivities. With 57 predictor variables at our disposal, we embark on a journey to unravel the complex factors influencing ride prices in Boston. By developing a predictive model that can accurately forecast ride-hailing prices, we aim to provide actionable insights for companies, policymakers, and customers alike, ultimately contributing to a more efficient and customer-centric ride-hailing.

2. Data:

2.1.1 About Dataset:

Our dataset comprises ride-hailing transactions from Boston, spanning a three-week period from November 26, 2018, to December 18, 2018, in the New York time zone. This dataset offers a unique glimpse into the winter season, characterized by fluctuating weather conditions and holiday festivities. With 57 predictor variables, our dataset provides a comprehensive view of the factors influencing ride prices, including temporal, spatial, and environmental factors. The dataset consists of 88986 observations, each representing a single ride-hailing transaction. The variables include information on pickup and drop-off locations, time of day, day of the week, weather conditions, and more. This rich dataset enables us to explore the complex relationships between these variables and ride prices, ultimately informing the development of a predictive model.

2.1.2 Modifications with respect to the original data:

Initially, we had 57 predictor variables, but we employed a rigorous feature selection process to identify the most relevant variables. We calculated the collinearity between each predictor variable and the response variable, and removed those with high collinearity values. This process ensured that the remaining variables were orthogonal to each other and had a minimal correlation with the response variable. After careful evaluation, we finalized 9 predictor variables that demonstrated the least collinearity with the response variable, with values close to zero. These variables include hour of the day, day of the week, month, source and destination locations, cab type, name, distance, surge multiplier, and price. Our dataset comprises [insert number] observations, with a balanced representation of both Uber (48%) and Lyft (51%) transactions.

2.1.3 Dataset's Finalized Variable Description:

1. Day - The day of the week (Monday to Sunday) when the ride was taken.

2. Month - The month (November or December) when the ride was taken.

3. Source - The pickup location of the ride, including latitude and longitude coordinates.

4. Destination - The drop-off location of the ride, including latitude and longitude coordinates.

5. Cab type - The type of vehicle used for the ride, such as UberX, Lyft, or Lyft Plus.

6. Name - The name of the ride-hailing service provider, either Uber or Lyft.

7. Price - The base fare of the ride, excluding any additional fees or surcharges.

8. Distance - The distance traveled during the ride, in miles.

9. Surge multiplier - A multiplier applied to the base fare during periods of high demand, indicating the level of surge pricing.

2.1.4 Box plots for finalized features.

To visualize the distribution of each feature and identify potential outliers, we created box plots for all 9 predictor variables. These plots revealed the median, quartiles, and extreme values for each feature, allowing us to detect skewness, outliers, and anomalies in the data.

3. Model Selection:

We employed a range of machine learning algorithms to predict cab prices in Boston. Specifically, we evaluated the performance of Multiple Linear Regression, Polynomial Regression (up to degree 9, limited by computational constraints), Decision Trees, Random Forests, and Support Vector Machines. To ensure a robust assessment of each model's predictive capabilities, we adopted a train-test split approach, allocating 70% of the data for training and 30% for testing. This systematic evaluation enabled us to identify the most accurate and reliable model for predicting cab prices in Boston.

3.1 Multiple Linear regression (MLR):

Multiple Linear Regression (MLR) is a statistical modeling technique that extends Simple Linear Regression (SLR) to predict a continuous outcome variable (y) based on multiple distinct predictor variables (x). The MLR model can be mathematically represented as:

$$y = b0 + b1*x1 + b2*x2 + \ldots + bn*xn$$

where b0 is the intercept or constant term, and b1, b2, …, bn are the regression coefficients or beta weights. These coefficients quantify the relationship between each predictor variable and the outcome variable, while controlling for the effects of all other predictors. Specifically, the coefficient bj represents the average change in y associated with a one-unit increase in xj, while holding all other predictor variables constant.

Our MLR Model and ASSUMPTIONS:

In our initial multiple linear regression model, all nine predictor variables were found to be statistically significant. However, diagnostic plots revealed a violation of the normality assumption. To address this issue, we applied a Box-Cox transformation to the response variable “price” with a lambda parameter of 0.088. This transformation successfully restored normality and homoscedasticity, as evidenced by improved diagnostic plots. Notably, the transformed model yielded an increased R-squared value of 0.943, compared to 0.931 in the original model. Our transformed model explains 94.3% of the variance in "price" with significant coefficients for most variables. The results suggest that factors like cab type, product ID, distance, and surge multiplier have a substantial impact on "price" while temporal variables have a relatively smaller effect.

3.1.2 MLR Final Model:

As the assumptions were not met and accuracy was not up to the mark, we tried boxcox analysis to check the transformation level. After the boxcox analysis we got a lambda value of 0.15. Then applied it to our response and built an model.

From this model we noted an increase in accuracy but it's been observed that one of variable distance to San Diego is not significant.

Linear Regression equation for the model is:

price^(0.088) =
1.1959 +
0.5262*I(name ==
'Black SUV') +... +
0.2206*distance +
0.9126*surge_multipl
ier + -4.568e-05*day
+ 9.612e-05*month +
-0.0002*hour

3.2 Other ML Models:

Other than multiple linear regression we also tried polynomial regression,decision trees,Random forests,Support vector machines .

3.2.1Polynomial Regression:

We attempted to fit a polynomial regression model to our data, but computational limitations prevented us from exploring degrees beyond 3. Notably, a degree 2 polynomial regression yielded an R-squared value of 0.475, indicating a moderate fit. However, the model's complexity was limited by computational constraints, precluding further exploration of higher-degree polynomials.

3.2.2Decision Tree and Random Forest:

Decision Trees are powerful but can suffer from overfitting, leading to high variance and poor generalization on unseen data. Random Forest, an ensemble learning algorithm, addresses this by combining multiple Decision Trees, each trained on random subsamples of the data. This reduces variance while maintaining low bias.

In our experiments, we implemented both a Decision Tree Regressor and a Random Forest Regressor. The Decision Tree model had bias and variance values of 0.993 and 0.943, indicating high variance and overfitting. The Random Forest model, on the other hand, achieved bias and variance values of 0.972 and 0.962, suggesting lower bias and significantly reduced variance. Therefore, the Random Forest model may generalize better on unseen data, demonstrating the benefits of ensemble learning.

3.2.3Support Vector Regressor:

Support Vector Regression (SVR) is a supervised learning algorithm used for regression tasks. It employs the kernel trick with a radial basis function (RBF) to construct a nonlinear decision boundary in the feature space. SVR aims to minimize the error between predicted and actual values while maintaining a balance between bias and variance.
In our model, we used SVR with the RBF kernel from the scikit-learn library and fitted the model using the training data (X_train, y_train). The resulting bias and variance values were 0.96 and 0.93, respectively. These values indicate that our SVR model strikes a good balance between bias and variance, suggesting decent generalization capabilities on unseen data.

4. Results and Discussions:

Our analysis revealed that the Random Forest model outperformed the other techniques, achieving an R^2 value of 0.972 and bias-variance values of (0.972, 0.962). In contrast, Multiple Linear Regression demonstrated a reasonably good fit with an R^2 value of 0.94, but had room for improvement. Polynomial Regression exhibited a subpar performance, indicated by an R^2 value of 0.47.

The Decision Tree model suffered from high variance and overfitting, as evidenced by its bias-variance values of (0.993, 0.943). Support Vector Regression, on the other hand, achieved a balance between bias and variance, with bias-variance values of (0.96, 0.93), but its overall performance was inferior to that of Random Forest.

The superior performance of the Random Forest model can be attributed to its ability to effectively balance bias and variance. This ensemble learning approach mitigates variance by combining multiple Decision Trees, resulting in a more robust and reliable model for our application. The Random Forest model's potential for accurate predictions on unseen data makes it the most suitable choice for our regression task. Based on our comprehensive evaluation, we conclude that the Random Forest model is the most optimal approach for our regression task, offering a superior balance of bias and variance, robustness, and reliability.

5. References:

https://www.kaggle.com/datasets/brllrb/uber-and-lyft-dataset-boston-ma/data

Code Appendix :

https://github.com/sainandavihari/Uber-lift-price-prediction

6. Appendix:

Fig : 1

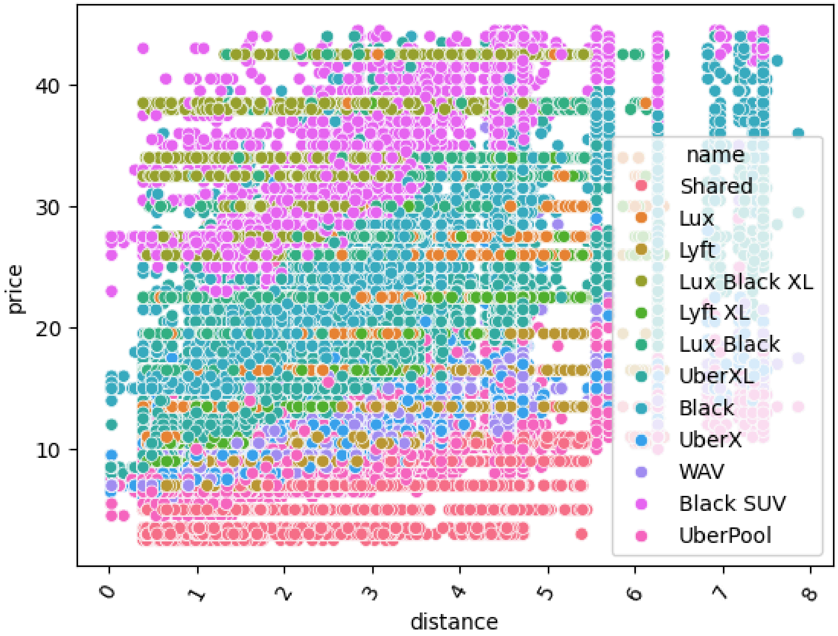


Fig 2:

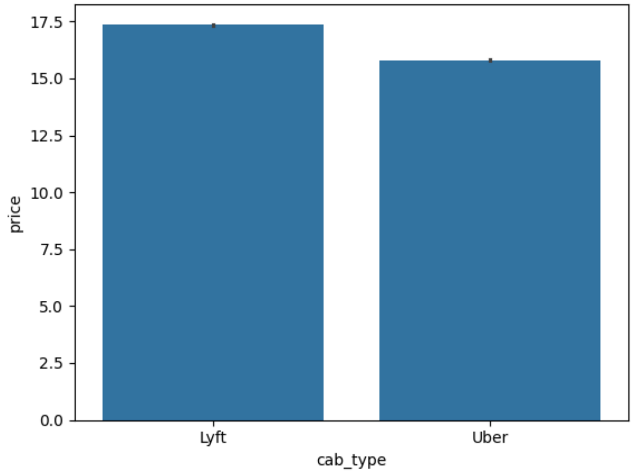


Fig 3:

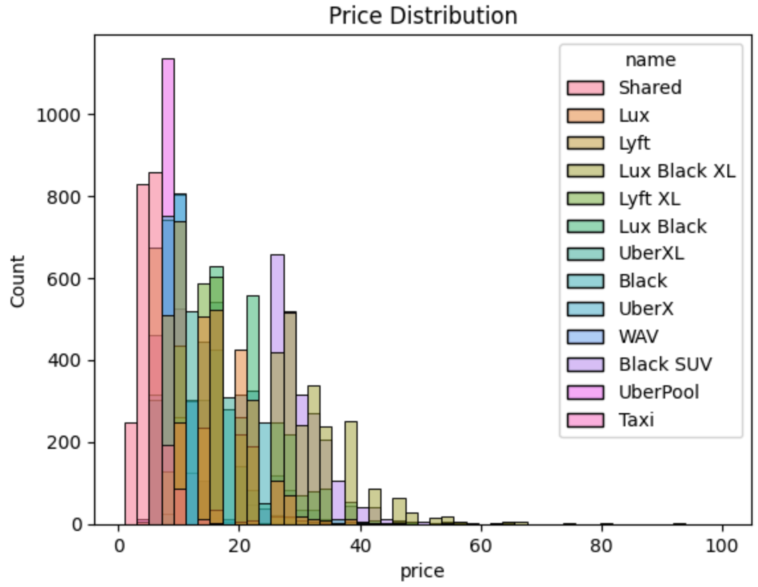


Fig:4

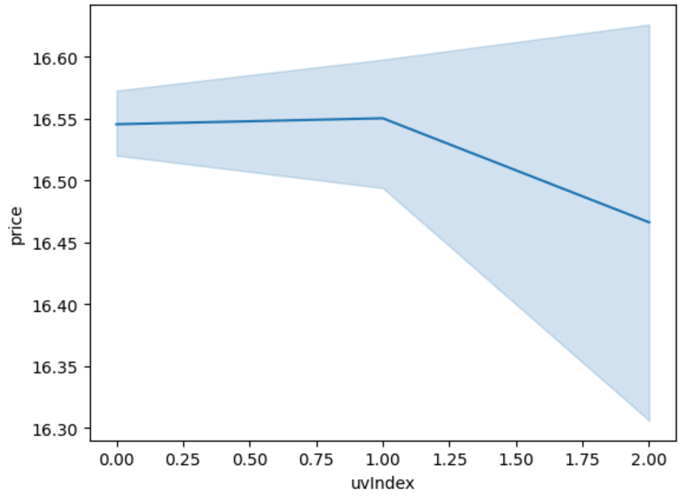
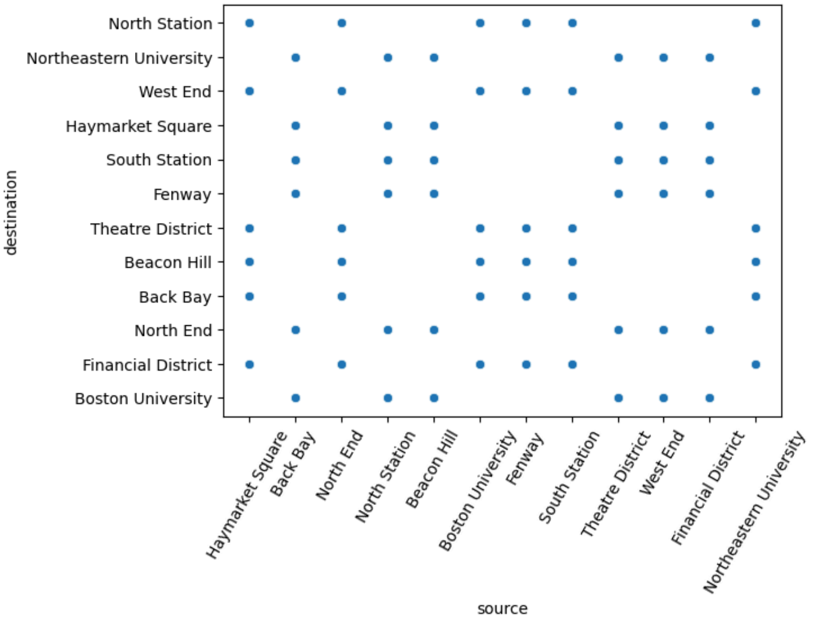


Fig: 5


North Station
Northeastern University
West End
Haymarket Square
South Station
Fenway
Theatre District
Beacon Hill
Back Bay
North End
Financial District
Boston University
destination
Haymarket Square
Back Bay
North End
North Station
Beacon Hill
Boston University
Fenway
South Station
Theatre District
West End
Financial District
Northeastern University
source