Uber&Lyft cab price prediction

Yijia Xue
Data Science Initiative, Brown University
Github: https://github.com/Ikea-179/Uber_Lyft_price_prediction

Introduction

Problem Description

This research project conducted a comparative analysis of Uber and Lyft rideshare services in Boston, Massachusetts, utilizing a substantial dataset comprising 693,071 rideshare instances, covering aspects such as ride details, temporal elements, and meteorological conditions.

Motivation

The impetus for this investigation stemmed from its pertinence to our experiences as university students residing in the Greater Boston area. The frequent utilization of Uber and Lyft services for urban transportation prompted an interest in comprehending the dynamics of their pricing models. This understanding is crucial in making informed choices about which rideshare service to employ under varying conditions.

Target Variable and Features

The uber and lyft price dataset is a prediction dataset[1]. The target variable is 'price' which stands for the price for each ride. There are 603071 data points and 18 feature columns. The explanation table of data columns is listed below:

Weather part	Ride part
Temp: Temperature in F	Distance: The Distance Between Source and Destination
Location: Location Name	Cab_type: Uber or Lyft
Visibility: Visibility in miles	Time_stamp: Epoch Time When Data was Queried
Temperature: Temperature in F	Destination: Destination of the Ride
Precipitation: Rain in Inches for the Last Hr.	Source: The Starting Point of the Ride
Time_stamp: Epoch Time When Row Data was Collected	Price: Price Estimate for the Ride in USD
Humidity: Humidity in %	Surge_multiplier: The Multiplier by Which Price was Increased, Default 1
Wind: Wind Speed in MPH	Id: Unique Identifier
	Product_id: Uber/Lyft Identifier for Cab-type
	Product_id: Uber/Lyft Identifier for Cab-type

Table 1. Variable Explanation table

Previous Work

Recent research has focused on the analysis of Uber and Lyft cab price data to enhance machine learning models used in ride-hailing price prediction. These studies aim to provide a robust framework for accurately forecasting trip costs, thereby aiding users in making cost-effective transportation decisions. One investigation within this domain revealed that the linear machine learning predictive model is capable of efficiently handling large datasets of ride information, achieving R square score of 85.4%[2]. Another study indicated that the ensemble algorithms could achieve about 93% accuracy in price prediction, demonstrating superior performance compared to the linear model[3].

EDA

Each row in the dataset represents a ride-share order which contains all the related information such as distance, cab company(uber/lyft), time stamp, cab type and weather information etc. The data almost contains no duplicate and missing value. However, for only target variable 'price', there are about 7 percent value are missing since no prices were associated with rows with the variable cab type Taxi. The visualizations of continuous features are shown below. We can find that most features are in nearly normal or uniform distribution. For visibility and percipintensity, the rides occur during various levels. Most rides appear to occur with little to no precipitation and around 10 miles visibility, as indicated by the taller bars on the left/right. The red line shows the median value with each feature during rides. For the target variable price, the data likely skews to the right, indicating that most rides fall within a lower price range, with fewer rides being more expensive. The red line denotes the median price of rides.

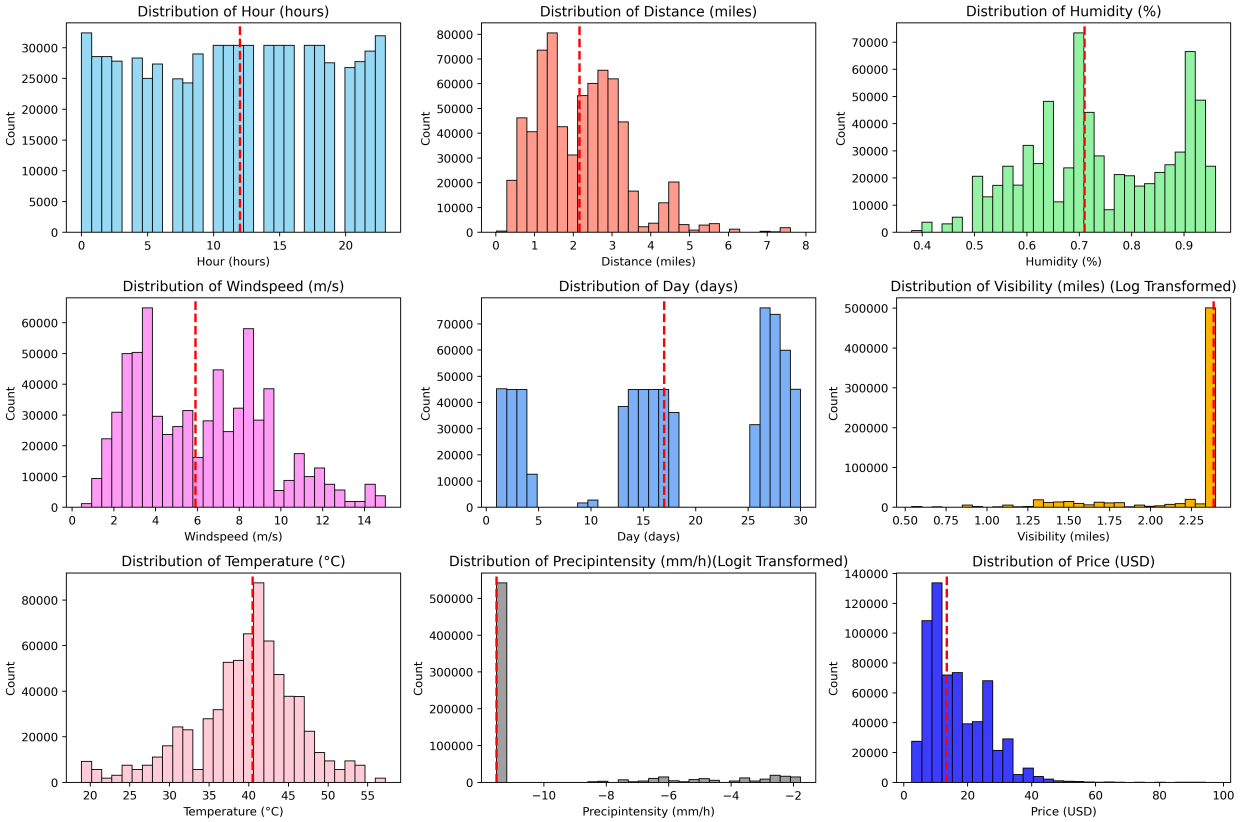

Figure 1. Visualization of the distribution of the continuous variables

From the dataset, we can get the geographical information. Using folium heatmap tools along with the longitude and latitude information, the visualization of the cab ride source position is shown below.

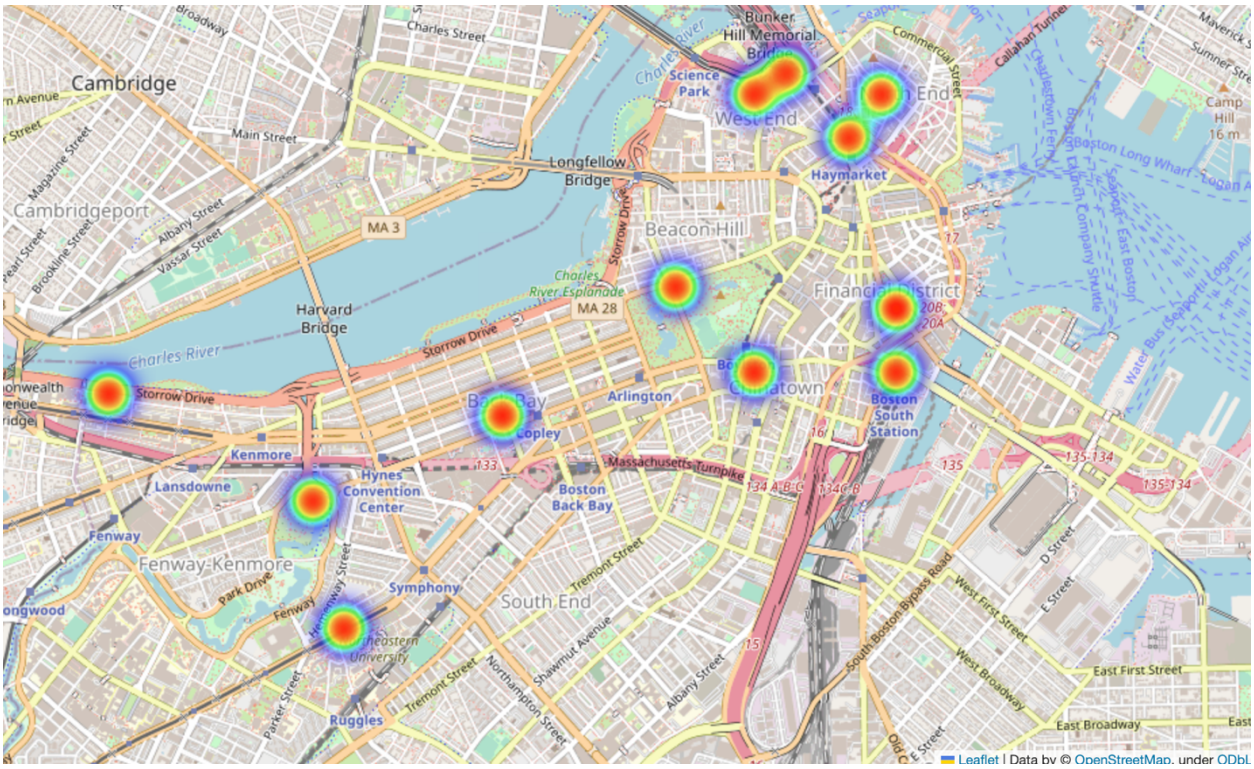

Figure 2. Visualization of the cab ride source position

The analysis reveals that all rides in the dataset are concentrated in the North Boston Area, with a relatively uniform distribution across various locations within this region. This pattern suggests a deliberate selection approach by the data contributor, who appears to have methodically chosen ride orders from these primary locations to ensure a balanced representation in the dataset.

We want to get some information about other features with target variables, below heatmap visualizes median ride prices to various destinations in Boston at different times of the day. Darker shades indicate higher prices, while lighter shades represent lower prices.

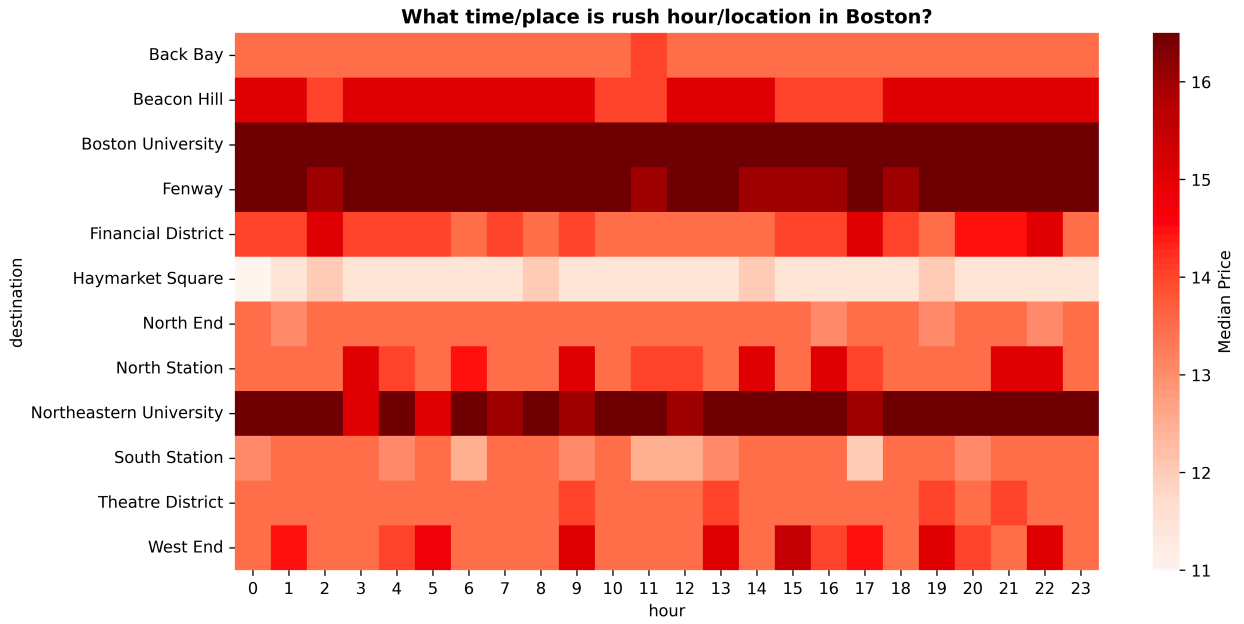

Figure 3. Heatmap between location/hour and median ride price

2 key observations are that: 1. There are notable periods during the day where median prices are higher, particularly in the early morning and late evening hours, which may correspond to typical rush hours when demand for rides is high. 2. Certain destinations show consistently higher median prices throughout the day, suggesting that factors such as distance, demand, or location desirability may influence the cost of rides to these areas.

The plot below is a boxplot comparison between Uber and Lyft prices across five categorized groups of cab types ranging from shared to luxury SUV options. Both companies offer a variety of car types to cater to different customer preferences and needs. This project categorized the ride product into the group listed on the visualization:

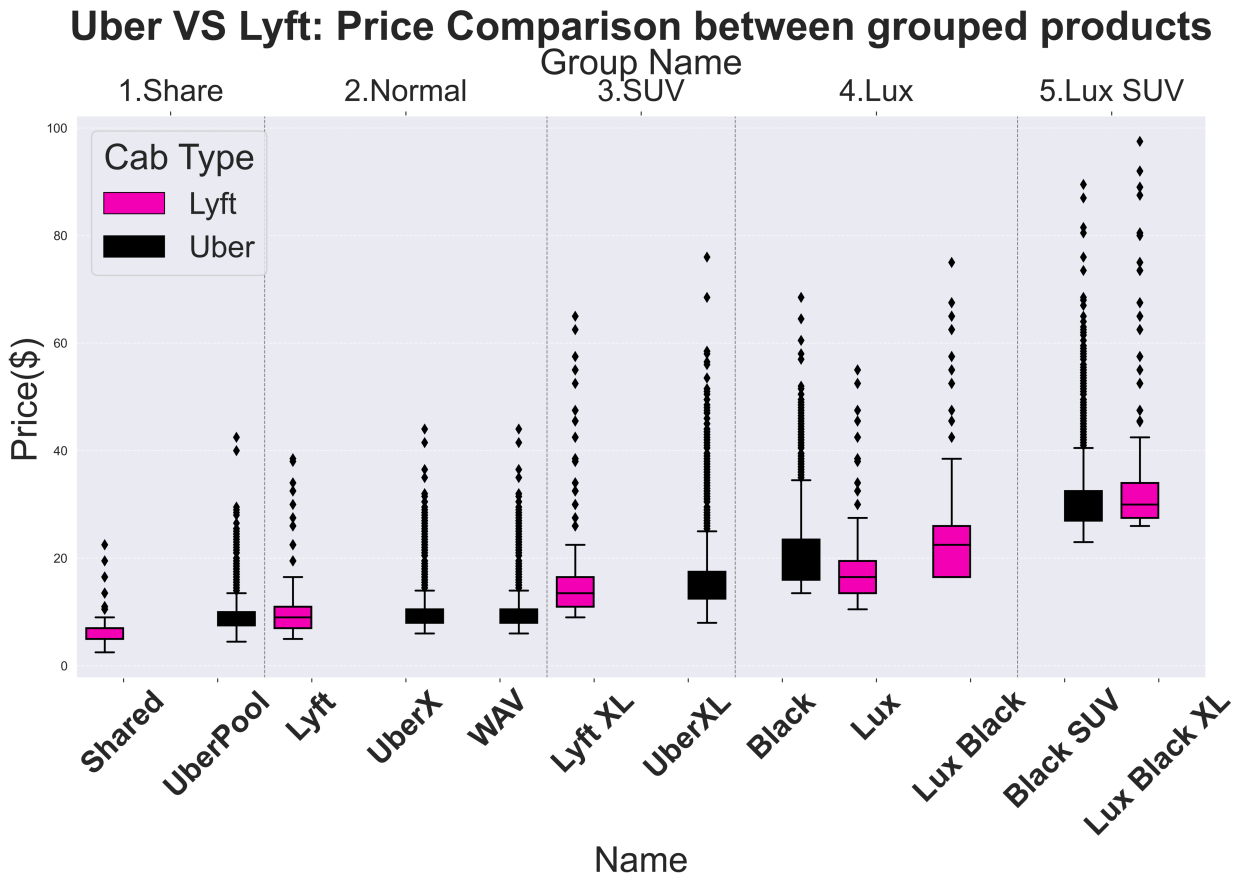

Figure 4. Boxplot compares the relationship between different product prices in uber and lyft

Shared and Normal Rides: For the most economical options, 'Share' and 'Normal', Lyft tends to be cheaper than Uber. This is reflected in the boxplot where Lyft (in pink) has lower median prices than Uber (in black) for both 'UberPool/Shared' and 'UberX/Lyft' categories. When it comes to luxury options, Lyft's 'Lux' and 'Lux SUV' services appear to be more expensive than Uber's 'Black' and 'Black SUV' offerings. This is evident in the higher median prices and upper quartiles for Lyft in the 'Lux' and 'Lux SUV' categories, diverging from the pricing pattern observed in the lower-tier service categories.

Distance is a key determinant in pricing for both Uber and Lyft, we want to find out whether two companies will have different pricing strategies on distance. Thus, a scatter plot was made to compare the pattern distance.

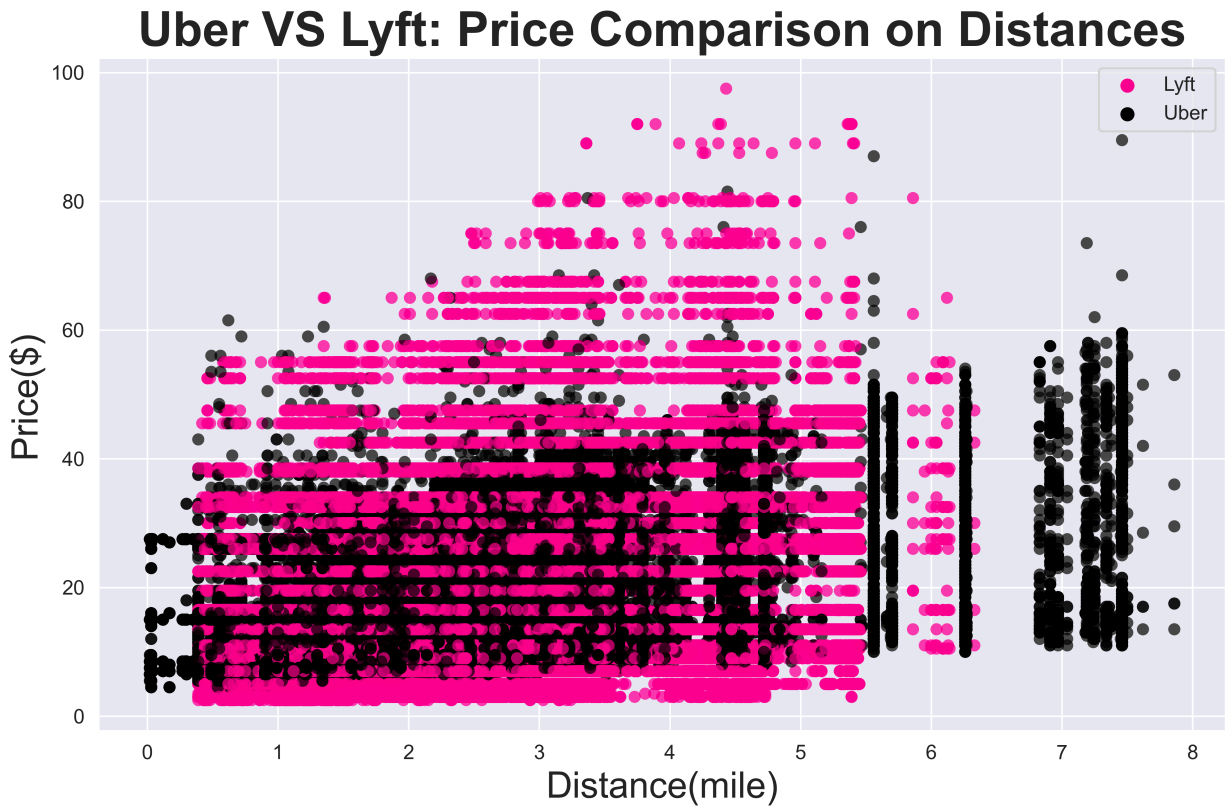

Figure 5. Scatterplot compares the relationship between distance and price in both company

The dataset reveals that Lyft's distance data exhibits a narrower range than Uber's. Additionally, Lyft's app displays a wider range in price estimations, suggesting there may be a more significant discrepancy between estimated and actual prices when compared to Uber. This points to the existence of other influential factors in Lyft's pricing mechanism, particularly for its high-end services.

Feature Engineering

New features like 'day_of_week', and 'morning/evening/night' were derived from the datetime variable, capturing the temporal dynamics that significantly impact ride prices. This process of feature creation was instrumental in enriching our dataset, allowing us to explore the intricate relationships between time-based factors and pricing.

Methods

data split

The dataset, comprising independent and identically distributed rows of ride information, was split into 60% training, 20% validation, and 20% test sets. Initially, a train-test split allocated 20% of the

data for testing and the rest for training and validation. Then, a 4-fold KFold split further divided this into 60% for training and 20% for validation, facilitating 4-fold cross-validation.

data preprocessing

For missing data, this dataset only has missing values in the target variable, a direct drop was applied to solve the issue. A pipeline is created using one hot encoder for categorical columns and Standard Scaler for continuous columns. During feature preprocessing process, 10 categorical columns and 11 numerical columns were transformed into 105 features.

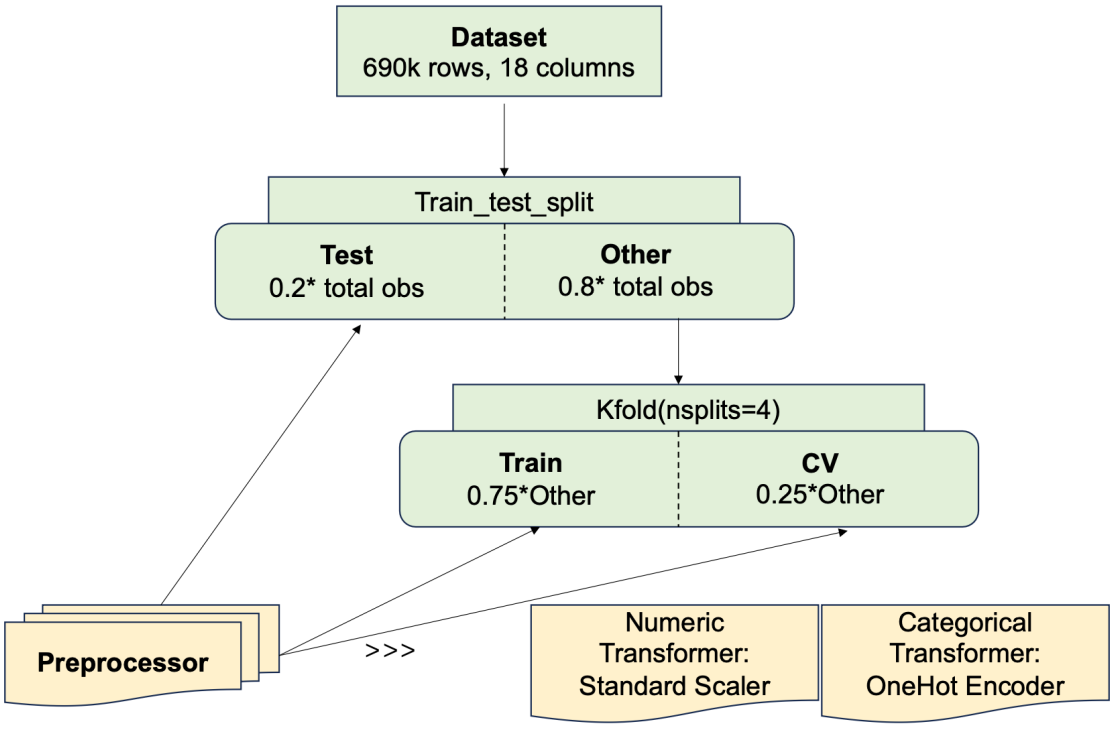

Figure 6. Flowchart of the data splitting and preprocessing process

Models and parameters parameter choosing

Ridge regression, Elastic Net regression, Random Forest regression and XGBoost regression were applied with hyperparameter tuning using grid search techniques. Hyperparameters tuning grid is shown in the below table.

Algorithms	**Hyperparameter tuning grids**
Ridge Regression	alpha: [1e-2, 3e-2, 6e-2, 1.6e-,1 4e-1, 1e0, 2.51e+0,6.31e+0, 1.585e+1, 3.981e+1, 1e+2]
Elastic Net	alpha: [1e-2, 3e-2, 6e-2, 1.6e-,1 4e-1, 1e0, 2.51e+0,6.31e+0, 1.585e+1, 3.981e+1, 1e+2]

	l1_ratio: [0.01,0.11,0.21,0.31,0.41,0.5,0.6,0.7, 0.8,0.9, 1]
Random Forest Regression	n_estimators: [100,200] max_depth; [None,1, 3, 10, 30, 100] max_features: [0.25, 0.5, 0.75, 1.0]
XGBoost Regression	n_estimators: [100, 200, 300] max_depth: [3, 6, 10] learning_rate: [0.01, 0.1, 0.2] colsample_bytree: [0.3, 0.7]

Table 2. Hyperparameter tuning grids

To address uncertainties from data splitting and model variability, training and testing were conducted across 10 random states and getting the same results for each state. For this regression problem, Root Mean Squared Error (RMSE) and R-squared (R2) were chosen as evaluation metrics. RMSE measures prediction error magnitude, while R2 indicates the proportion of variance explained by the model, both apt for assessing performance in this continuous outcome prediction.

Result

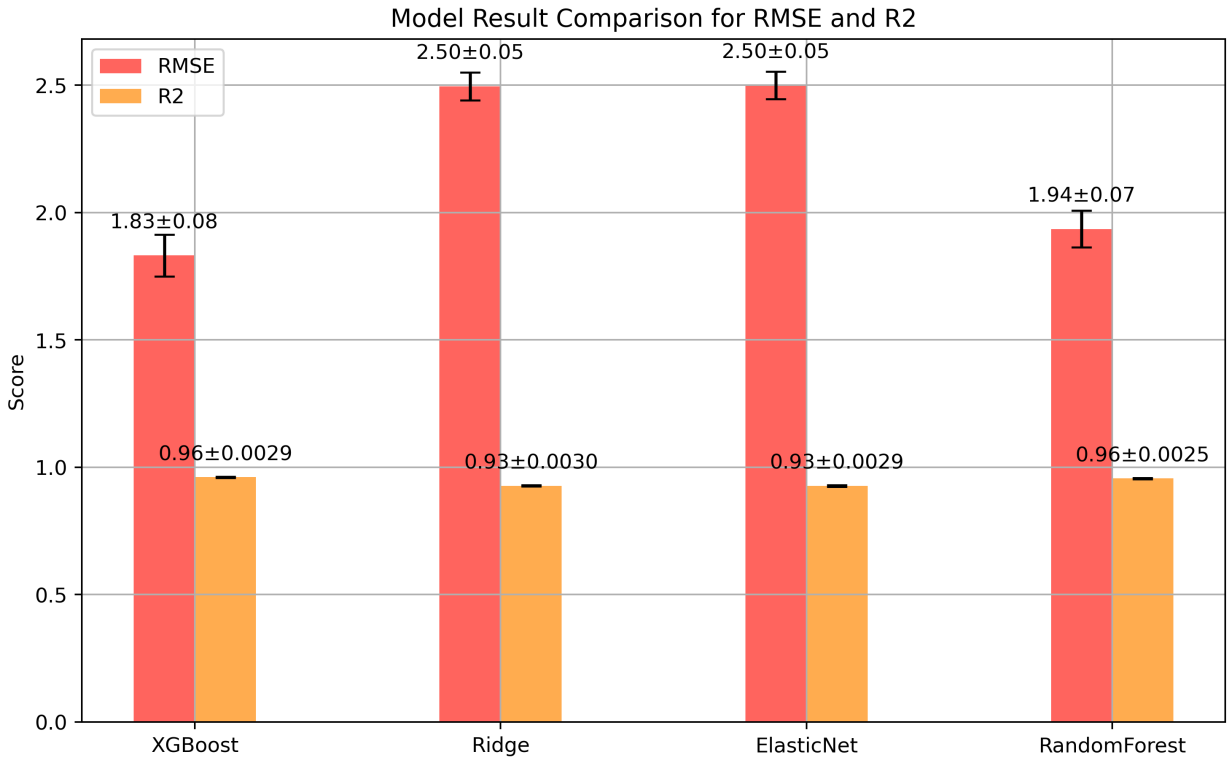

Figure 7. Evaluation Result comparison between 4 implemented models

The baseline RMSE of the test set is 9.32 and the baseline R2 score is -0.0007 They are measured by letting prediction equal to the mean of y_true for all points in the test set and calculating the metric. Among the 4 implemented models, xgboost appears to be the best-performing one since it has both the highest R2 score and lowest RMSE. Moreover, the consistency of its performance is underscored by the small standard deviation in its RMSE across different runs, suggesting that the model is robust and reliable.

Regression model	R2(mean/std)	RMSE(mean/std)
Baseline	-0.0007	9.32
Ridge	0.93±0.0030	2.50±0.05
Elastic Net	0.93±0.0029	2.50±0.05
Random Forest	0.96±0.0025	1.94±0.07
XGBoost	0.96±0.0029	1.83±0.08

Table 3. Evaluation metrics comparison between baseline and 4 implemented models

7 different feature importance scores were calculated, shown in below figures.

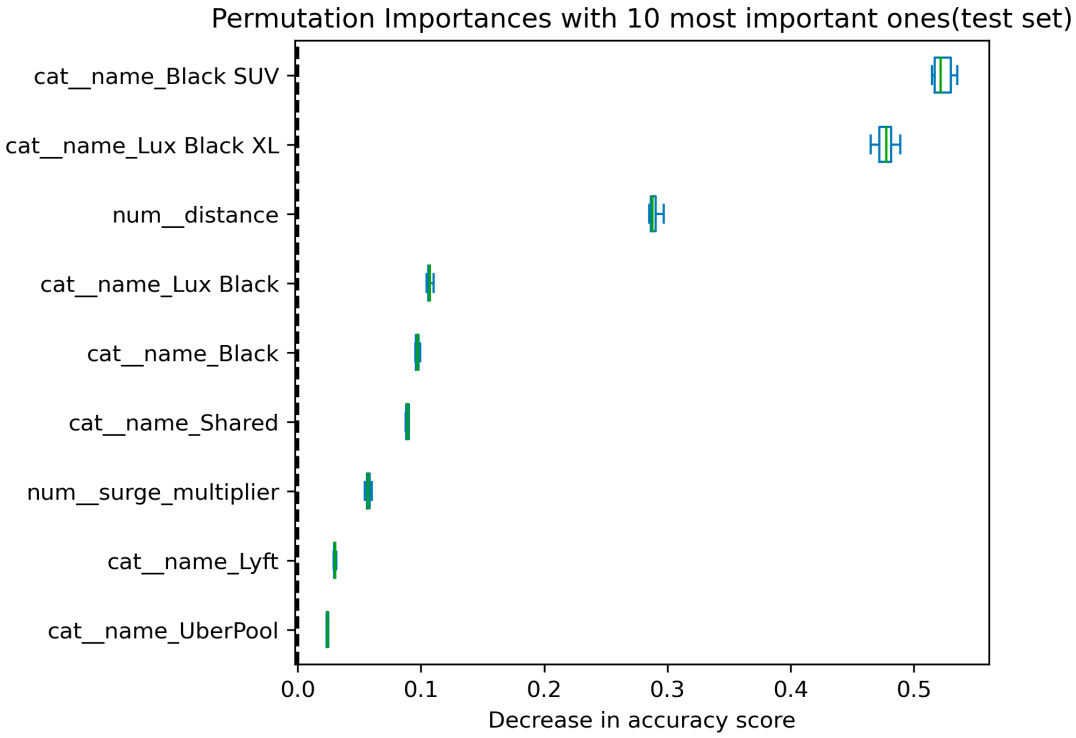

Figure 8. Top 10 most important features as per permutation feature importance.

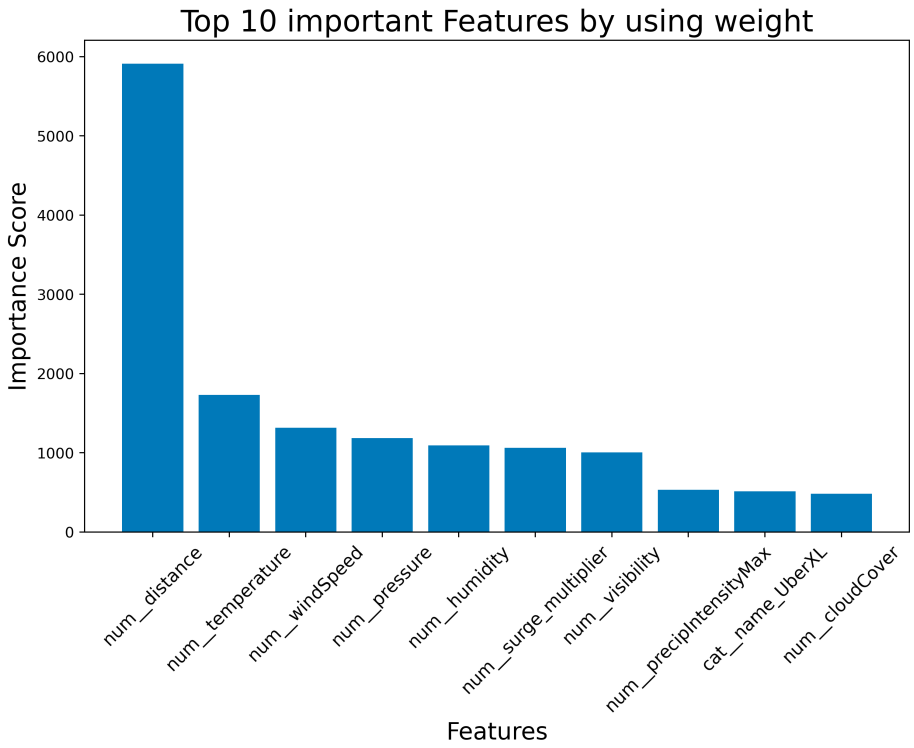

Figure 9. Top 10 most important features as per XGBoost relative importance measure-weight

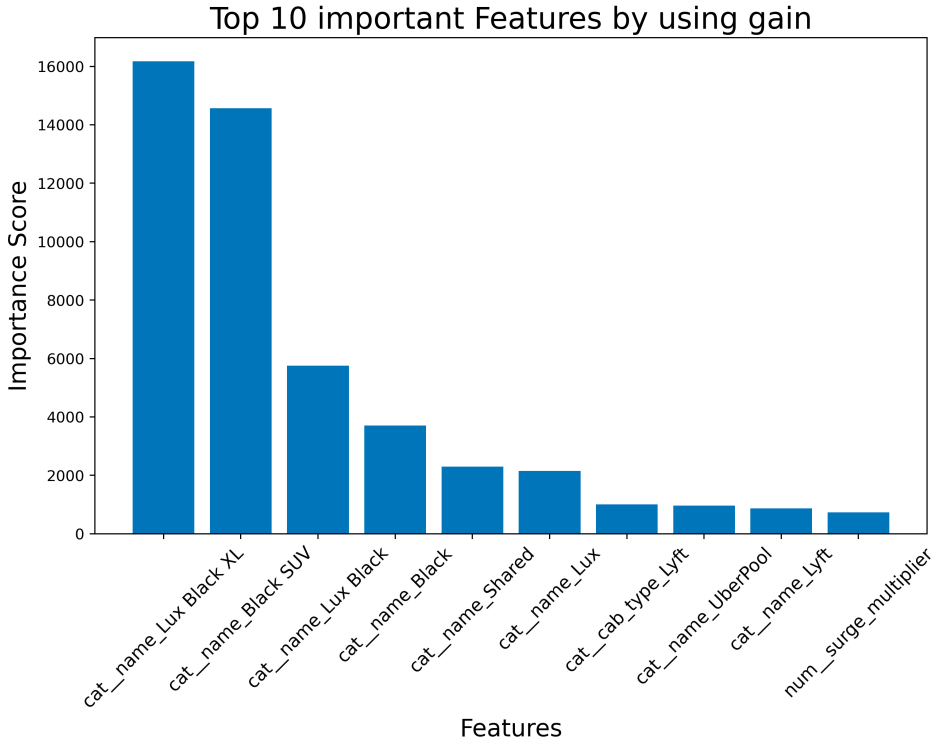

Figure 10. Top 10 most important features as per XGBoost relative importance measure-gain

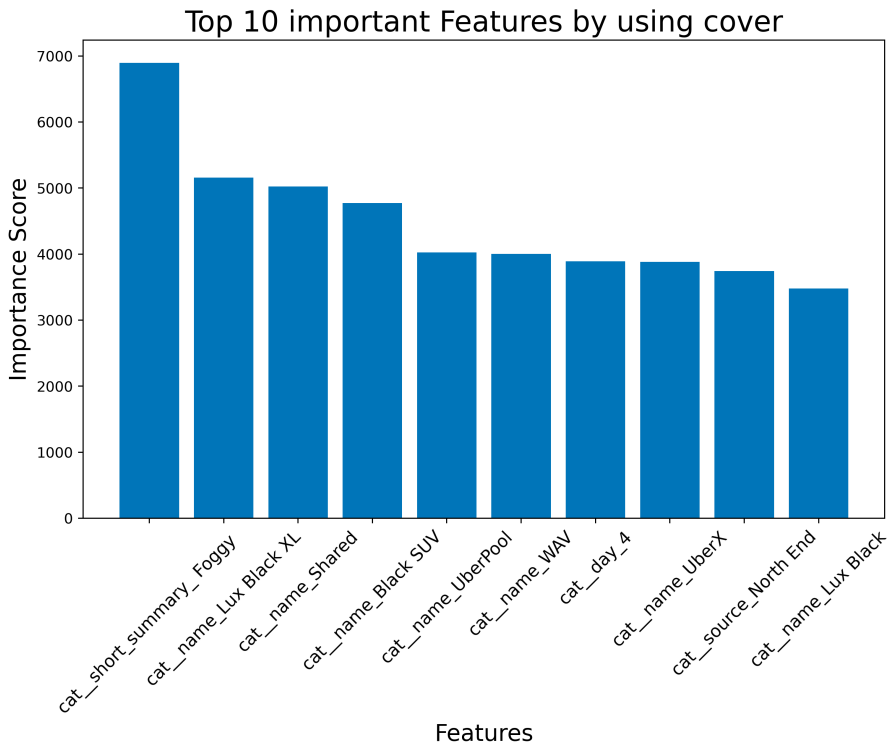

Figure 11. Top 10 most important features as per XGBoost relative importance measure-cover

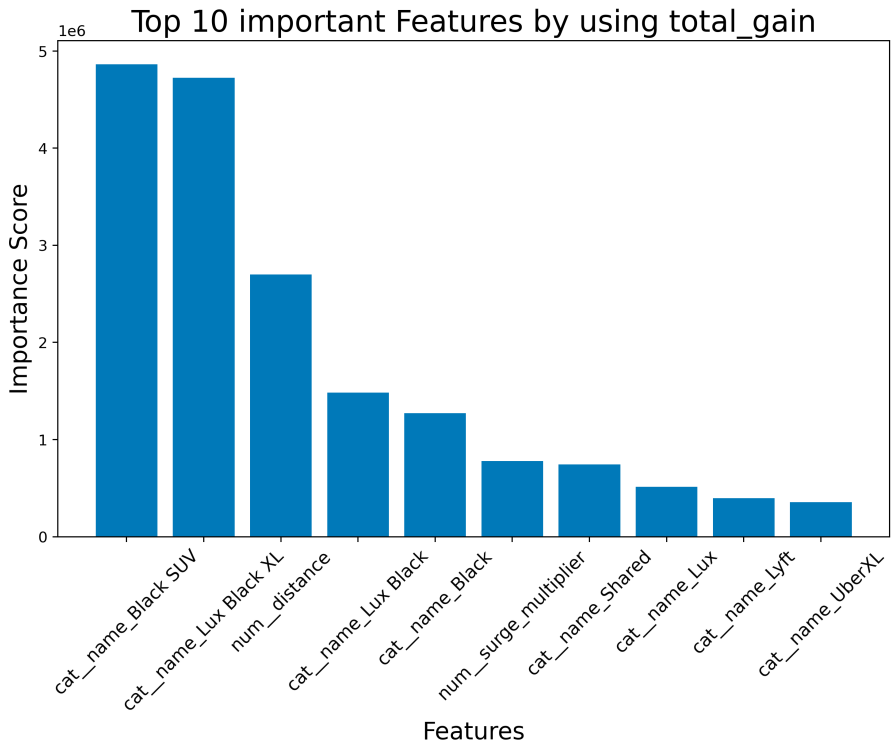

Figure 12. Top 10 most important features as per XGBoost relative importance measure-total_gain

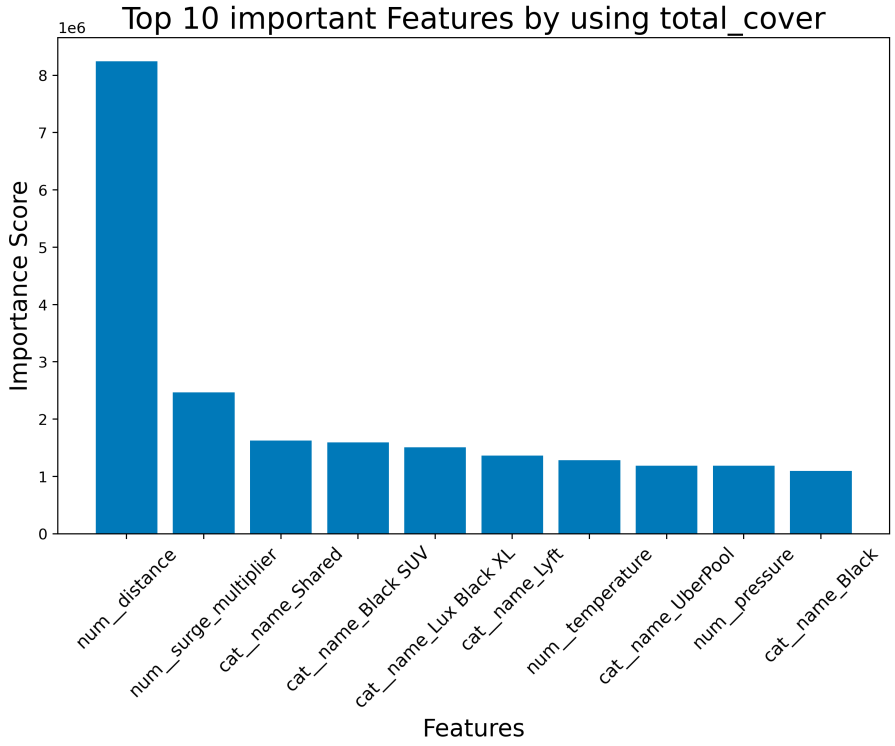

Figure 13. Top 10 most important features as per XGBoost relative importance measure-total_cover.

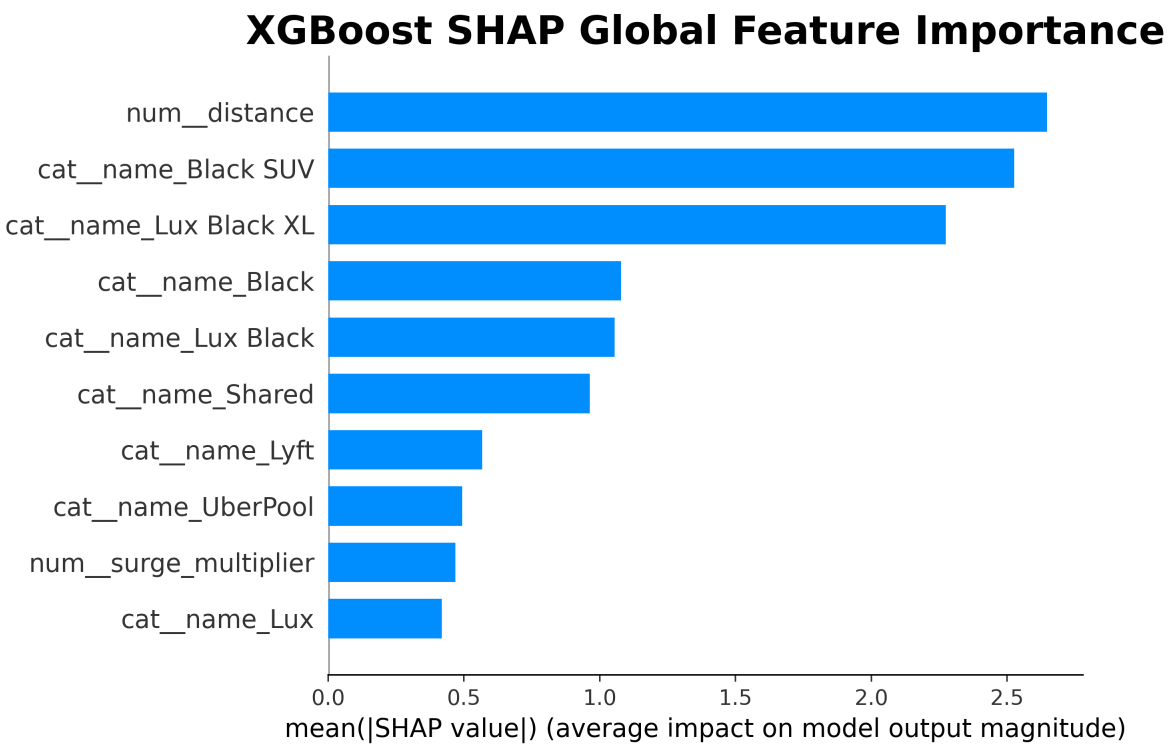

Figure 14. XGBoost SHAP global feature importance

In the permutation importance plot, cat_name_Black SUV and cat_name_Lux Black XL stand out as the most significant features, indicating that altering these features' values has the most substantial impact on model predictions. Similarly, in the weight-based importance, num_distance is the most influential feature, implying it is used most frequently to split data in trees. For the gain-based importance, cat_name_Lux Black XL and cat_name_Black SUV again emerge as top features, suggesting that splits on these features contribute to significant gains in model performance. The cover-based importance plot also highlights num_distance as the most impactful, suggesting this feature covers a significant number of data points when used for splitting. For the SHAP global importance, it seems to restate all the important features, including num_distance and luxury category.

Collectively, these plots suggest that while num_distance is a critical feature in terms of frequency and data coverage in the model, the categories related to the type of service (Black SUV, Lux Black XL) are highly impactful in terms of model performance improvement when they're involved in a split. This indicates that both the type of ride and the distance of the trip are key determinants in predicting the target variable.

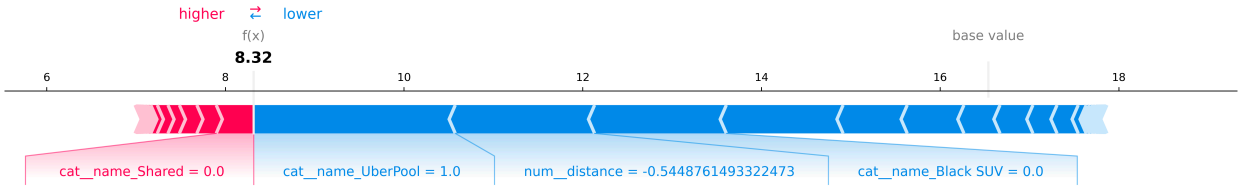

Figure 15. Shap local value for instance 0

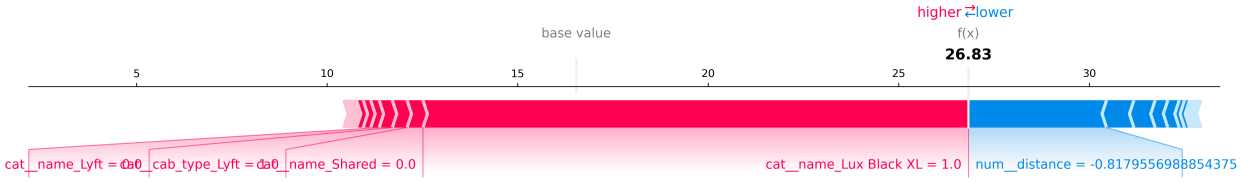

Figure 16. Shap local value for instance 100

The above plots show the shap local feature importance values of 2 different instances. It is interesting that in Figure.15, cat_uberpool plays more importance than the black suv category and is the most influential factor in affecting the predicted value for this specific instance, outweighing the other features that are contributing negatively. This is consistence with the low predicted price. However, in the high price order, the luxury black xl feature is more influential than others.

Outlook

One of the primary limitations is the static nature of the dataset, which does not account for real-time variables such as current traffic conditions and driver availability. Incorporating these dynamic elements could significantly enhance the accuracy and applicability of the predictive models. Future work could focus on integrating real-time data streams, such as traffic flow and weather updates, to develop more sophisticated and responsive pricing models. Another promising direction for future research is the deployment of these models in a real-time predictive analytics framework. This would enable ride-sharing companies to dynamically adjust their pricing strategies based on current data, leading to more efficient operations and improved customer satisfaction. Additionally, testing further models such as K-Nearest Neighbors (KNN) regression and models with reduced features could potentially enhance the evaluation metrics.

References

[1]Uber and Lyft dataset Boston, MA. Kaggle. Retrieved from https://www.kaggle.com/datasets/brllrb/uber-and-lyft-dataset-boston-ma
[2]https://medium.com/@rouhinadey98/prediction-of-prices-for-cab-rides-on-uber-and-lyft-f50dc7f28f80
[3] https://nycdatascience.com/blog/meetup/uber-vs-lyft-price-prediction-machine-learning-model/

Github repository

https://github.com/lkea-179/Uber_Lyft_price_prediction