Uber/Lyft Ride Price Prediction

Jinjia Zhang

Data Science Institute, Brown University

1. Introduction

Ride-sharing services like Uber and Lyft have completely changed how we think about getting around, providing flexible and cost-effective options compared to traditional taxis. While these services have become very popular, their pricing models could be more precise. Prices fluctuate based on several factors, including the ride's distance, the type of ride, the time, and even the weather. Understanding and anticipating these prices is essential for riders looking to save money and drivers and businesses that want to maximize their earnings.

In recent years, several studies have applied machine learning techniques to predict ride-sharing prices, aiming to enhance both user experience and operational efficiency. For instance, a study published in the International Journal of Research Publication and Reviews developed a machine learning-based approach to forecast ride requests and predict prices for Ola, a popular ride-sharing service in India. The researchers utilized historical ride data and time-distance factors, and they applied algorithms such as decision trees, random forests, and gradient boosting to build their predictive models. Their findings demonstrated convincing results in model accuracy and efficiency, which led to better resource allocation and improved customer satisfaction [1].

For this project, we used a dataset collected on Kaggle with 693,071 records from Boston, Massachusetts, gathered over two months in late 2018 [2]. It includes 57 attributes covering ride-specific details (like distance and ride type), time factors (such as the hour and day), and weather conditions. Previous research on predicting ride prices usually focused on linear and tree-based models, showing that distance and surge pricing are two of the most significant factors. In this project, we will take that knowledge further by applying various machine learning models to predict ride prices, and we will also look at ways to make these models more straightforward to understand and improve their performance.

To show the main features of our dataset intuitively, here is a variable table including the essential variables:

Variable Name	Description
id	Unique identifier for each record.
timestamp	Timestamp of the data recording.
hour / day / month	Time and date components.
datetime / timezone	Full date-time and timezone.
source / destination	Pickup and drop-off locations.

cab_type / product_id / name	Cab service details.
price / distance	Fare and distance.
surge_multiplier	Pricing adjustments during peak times.
latitude / longitude	Location coordinates.
temperature / apparentTemperature	Weather conditions.
long_summary / icon	Weather description and icon.
precipIntensity / probability / humidity	Precipitation and humidity.
windSpeed / gust / bearing	Wind details.
cloudCover / uvIndex / ozone	Atmospheric conditions.
sunriseTime / sunsetTime	Time of sunrise and sunset.
moonPhase	Phase of the moon.

Table 1: Variable Table of our Dataset [3]

2. Exploratory Data Analysis

A detailed exploration of the dataset revealed several key insights. The first notable finding was the distinction between Uber and Lyft ride prices. From these hexbin plots, it is easy to observe that Uber rides have higher median prices and more significant variability than Lyft rides. Premium ride types, such as UberBlack and LyftLux, further amplified this difference, as evidenced by box plots showcasing their pricing distributions.

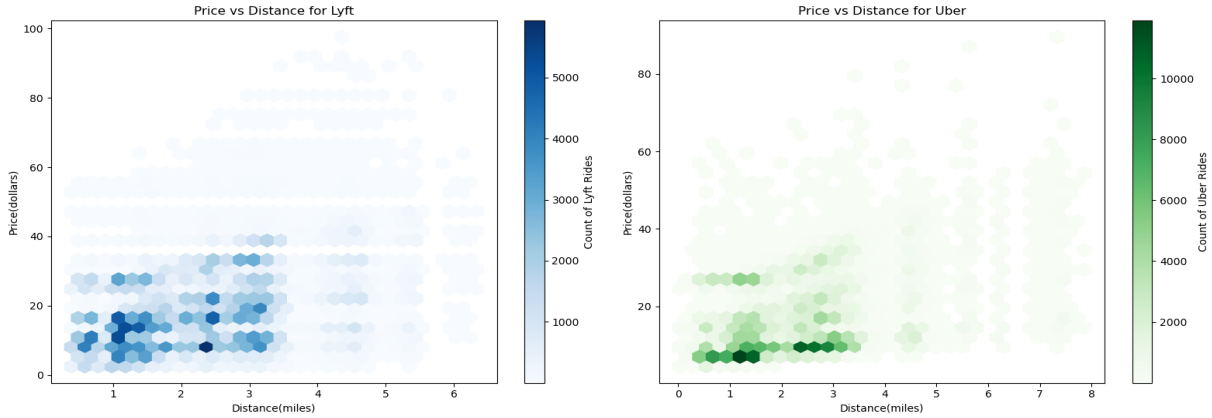

Figure 1: Hexbin Plots of Price vs Distance for Uber and Lyft

Temporal analysis revealed predictable patterns in ride demand and pricing. We used histograms of ride volume by hour and by ride type to see if there is a pattern between ride volume and ride types/time slots. Note that the histograms are not stacked ones, but in an overlapping form. The result indicated peaks during morning (8–10 AM) and evening (5–7 PM) commute hours. Correspondingly, prices increased during these times, likely due to heightened demand and surge pricing. Meanwhile, Uber has more ride requests than Lyft at any time, possibly due to the customers' preferences.

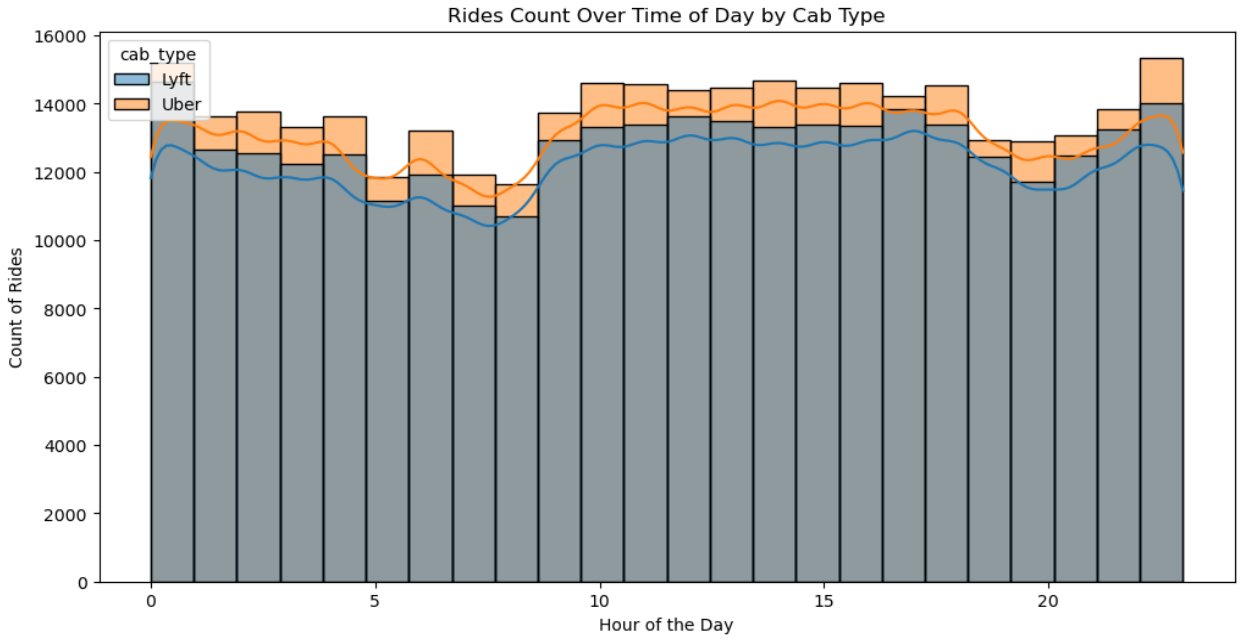

Figure 2: Histograms of Rides Count Over Time of Day by Cab Type

We also analyzed the weather conditions with a tree map, but the results showed that their impact on ride prices was secondary to ride-specific variables. For instance, while wind speed and precipitation contributed to price variations, distance and ride type weakened these effects. This finding provides insight into the fact that environmental conditions play a lesser role in determining dynamic pricing, even if they are important for user comfort.

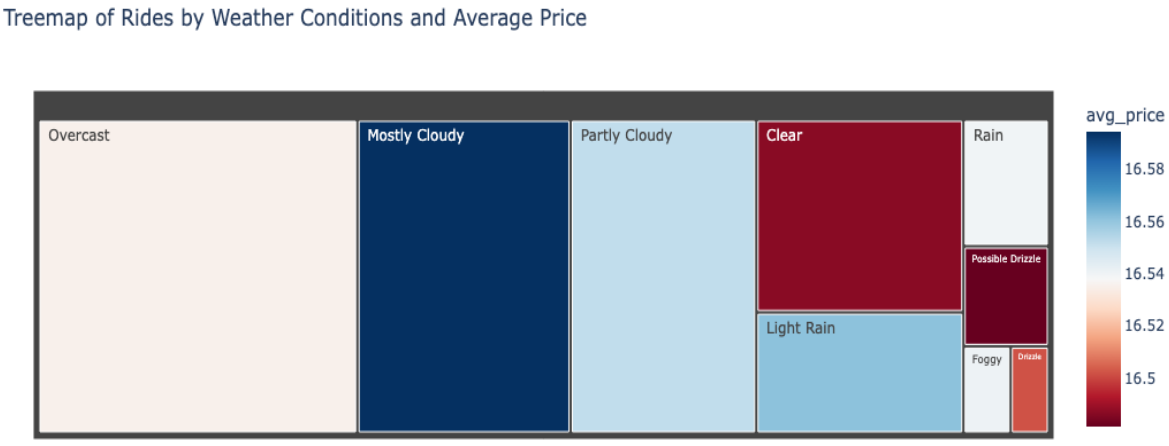

Figure 3: Treemap of Rides by Weather Conditions and Average Price

3. Methods

We first split our dataset into 80% other and 20% test sets. Then, 3-fold cross-validation was applied to the other set during hyperparameter tuning to ensure robust evaluation and minimize overfitting. The other set was split into 75% training and 25% validation sets. Therefore, our ultimate splitting result will be a 60% training set, 20% validation, and 20% test set. Since our dataset only contains 55095 rows of missing values in the target variable (price), we dropped those instances. During data preprocessing steps, categorical variables were one-hot encoded, while continuous features were scaled using StandardScaler, ensuring uniform input ranges across models. These preprocessing steps were integrated into a machine learning pipeline to facilitate seamless model training and evaluation.

After preprocessing data, we defined the machine learning models and hyperparameters needed in GridSearchCV. We not only tried linear models such as Lasso, Ridge, and ElasticNet, but also explored tree-based models like random forest and XGBoost. To evaluate the performances of those models, we used both root mean squared error (RMSE) and R2 score to ensure both accuracy and explainability of the models. We trained and tested our model using 10 different random states and took the average value of RMSE and R2 scores from those 10 different random states to address uncertainties from data splitting and model variability. To summarize the models and hyperparameters we used in GridSearchCV, here is a table for reference.

Model	Hyperparameters
Lasso	'lasso__alpha': [0.01, 0.1, 1]
Ridge	'ridge__alpha': [0.01, 0.1, 1]
ElasticNet	'elasticnet__alpha': [0.01, 0.1, 1], 'elasticnet__l1_ratio': [0.1, 0.5, 0.9]
Random Forest	'randomforestregressor__n_estimators': [50, 100, 200], 'randomforestregressor__max_depth': [None, 10, 20]
XGBoost	'xgbregressor__n_estimators': [50, 100, 200], 'xgbregressor__learning_rate': [0.01, 0.1, 1], 'xgbregressor__max_depth': [3, 5, 7]

Table 2: ML Models and Hyperparameters

To show the performances of each model, we displayed the model results in the following table, and we used bar plots to visualize the RMSE and R2 scores. The baseline model, predicting the mean price for all rides, achieved an RMSE of 9.15. All machine learning models significantly improved upon this baseline. Among the models we used, XGBoost achieved the best

performance with an RMSE of 1.73 and an R^2 score of 0.96, indicating its superior ability to capture complex relationships within the data. Random Forest followed closely with an RMSE of 1.86 and R^2 of 0.96, while linear models like Lasso, Ridge, and ElasticNet lagged behind due to their limited capacity to model non-linear interactions.

Model	**Mean RMSE**	**Std RMSE**	**Mean R^2**	**Std R^2**
Lasso	2.40	0.07	0.93	0.004
Ridge	2.40	0.07	0.93	0.004
ElasticNet	2.40	0.07	0.93	0.004
Random Forest	1.86	0.07	0.96	0.003
XGBoost	**1.73**	0.09	**0.96**	0.004

Table 3: Model Results—Evaluation Metrics

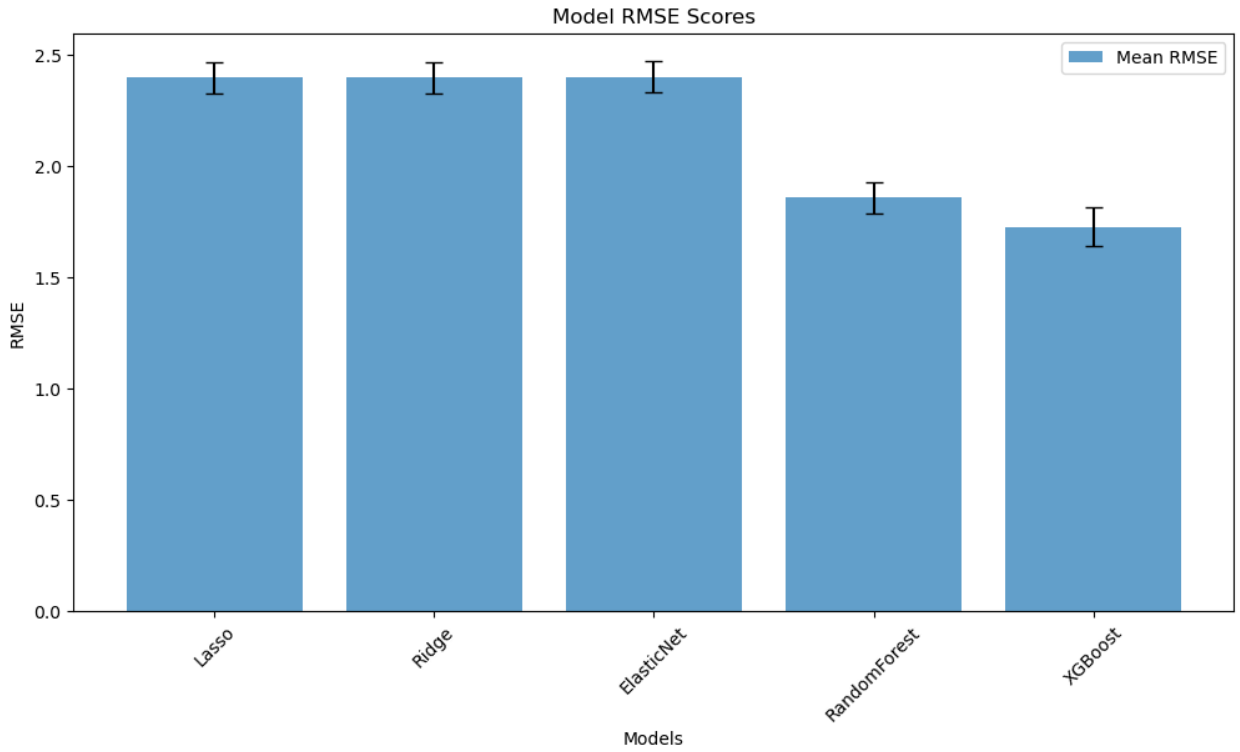

Figure 4: Model RMSE Scores

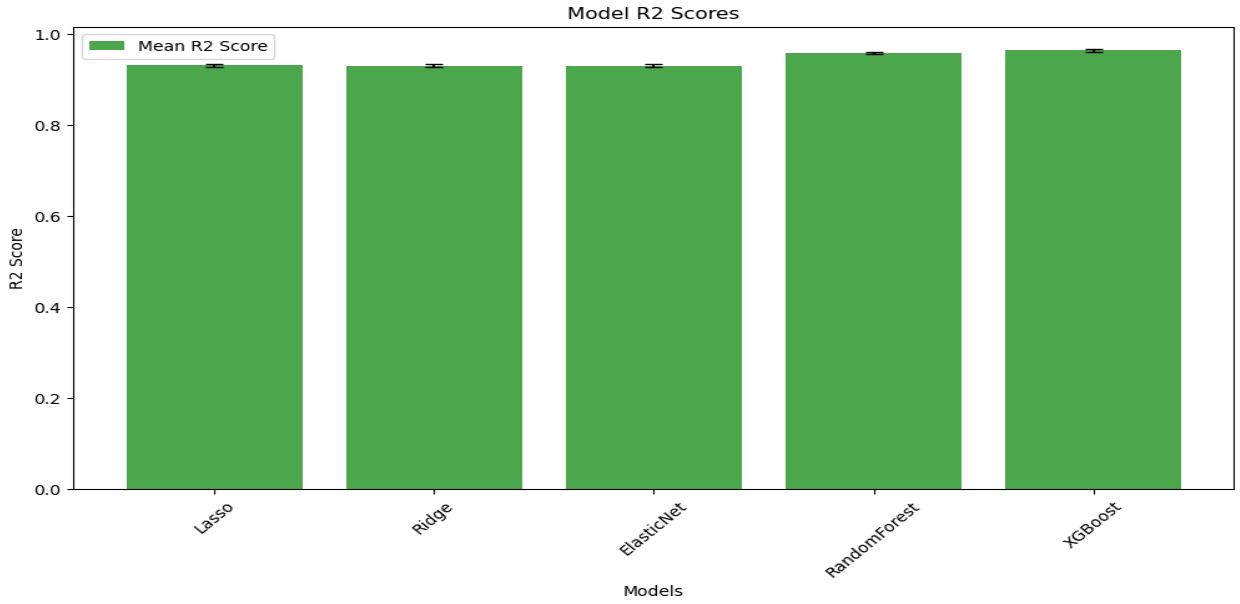

Figure 5: Model R2 Scores

Apart from the evaluation metrics, we also dived into the feature importance to find the key factors influencing the ride price. We chose to analyze the feature importance by using the importance types in the XGBoost model. The following figures show the top 10 most important features for each metric in the XGBoost model.

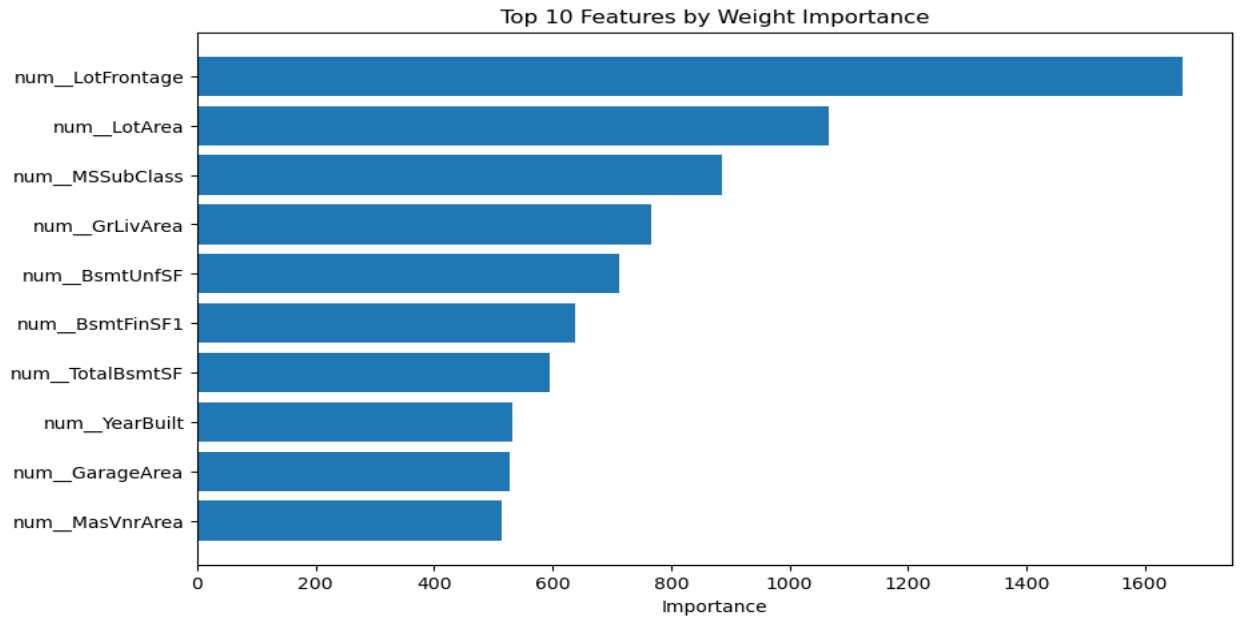

Figure 6: Top 10 Features by Weight Importance

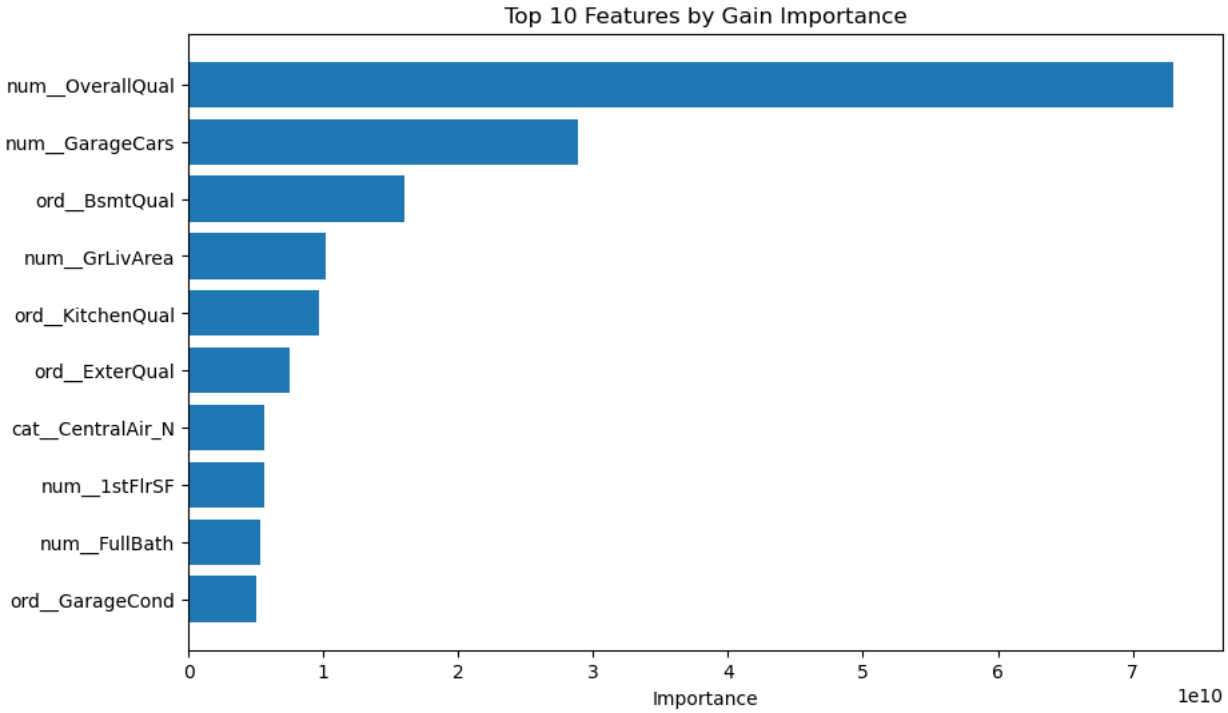

Figure 7: Top 10 Features by Gain Importance

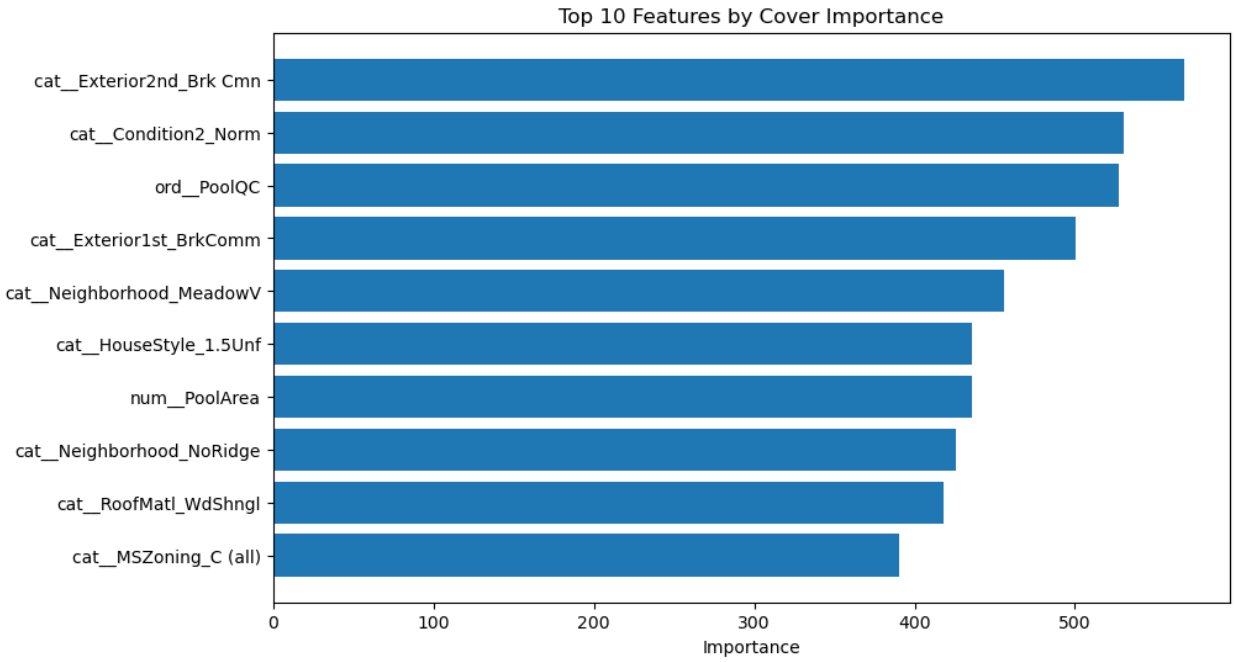

Figure 8: Top 10 Features by Cover Importance

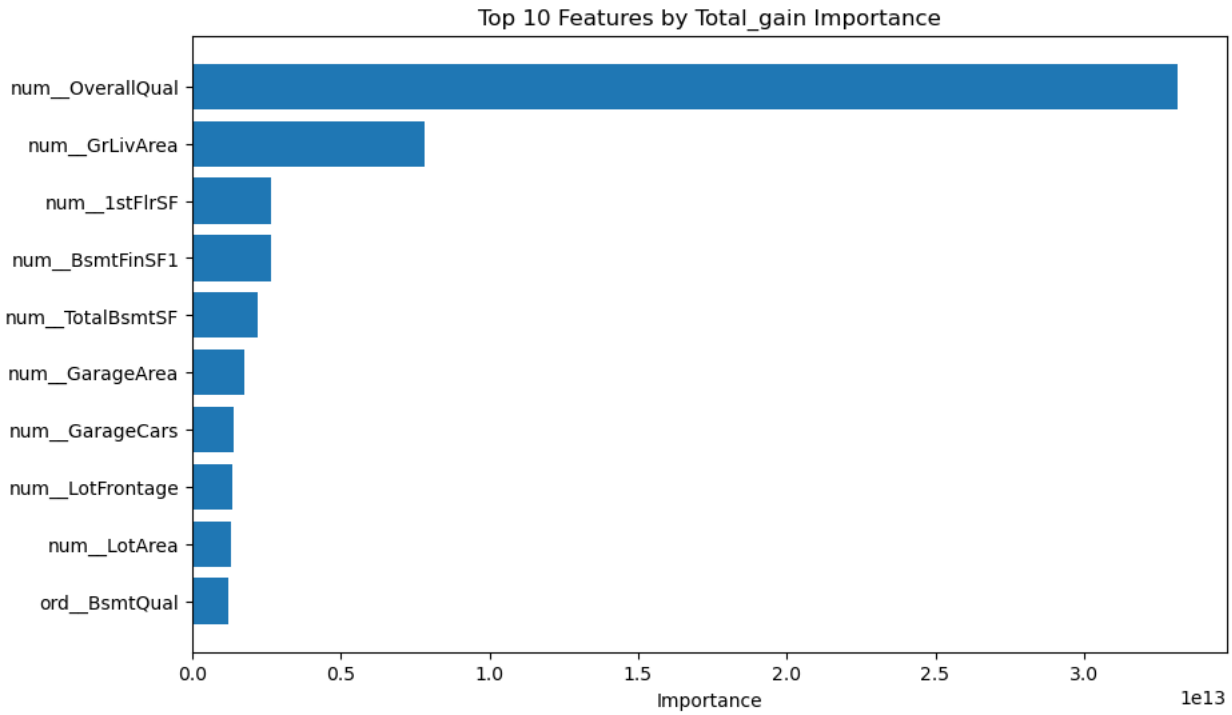

Figure 9: Top 10 Features by Total Gain Importance

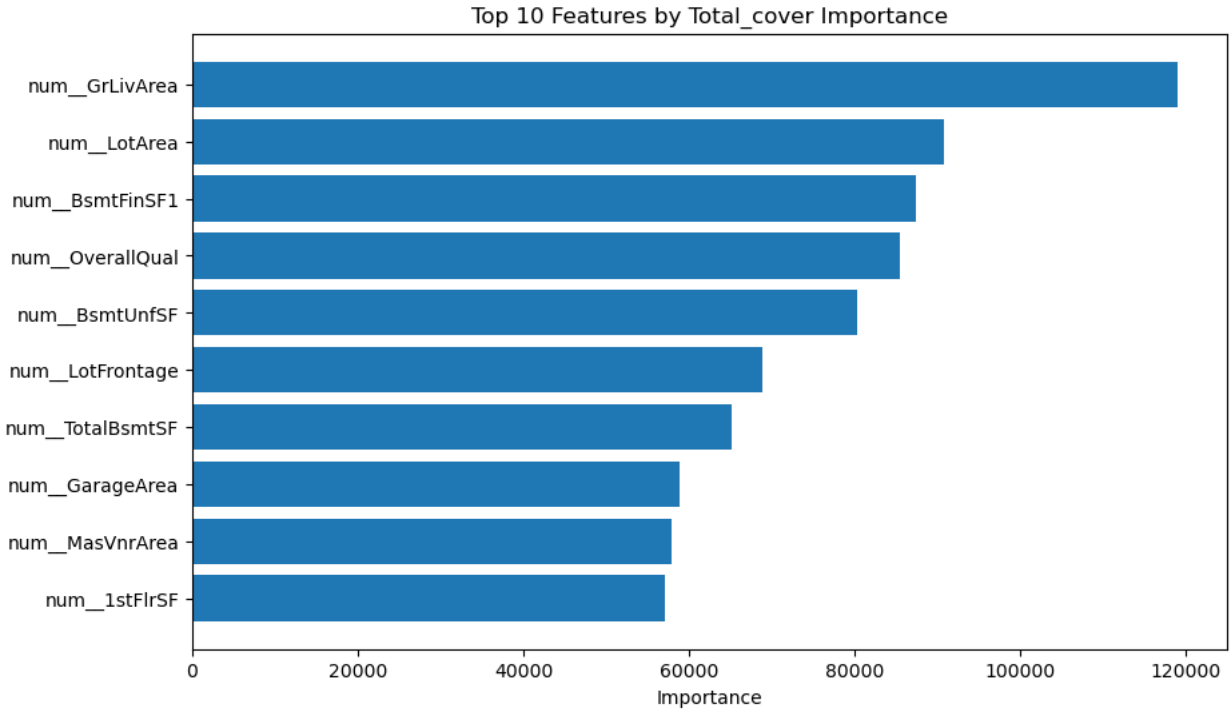

Figure 10: Top 10 Features by Total Cover Importance

The combination of the above results demonstrated that distance, surge multiplier, and ride type as the most influential factors. Distance is the most influential factor, contributing significantly to the ride price. The surge multiplier ranks second, reflecting its direct impact on pricing during peak periods. Ride type is the third important factor, indicating that premium services like UberBlack or LyftLux are priced higher than standard options.

To prove the above conclusion, we also explored the importance of SHAP features. First, we calculated the top 10 most important features by global SHAP importance. From the plot, it is easy to observe that distance and ride types are still key factors in predicting ride prices. Interestingly, while weather-related features such as temperature and wind speed were included in the models, their importance was relatively low. This suggests that ride-specific and temporal factors, such as time of day and ride distance, overshadow the influence of weather conditions on pricing. For instance, the effect of wind speed on prices was minimal compared to the direct impact of surge multipliers or ride distance.

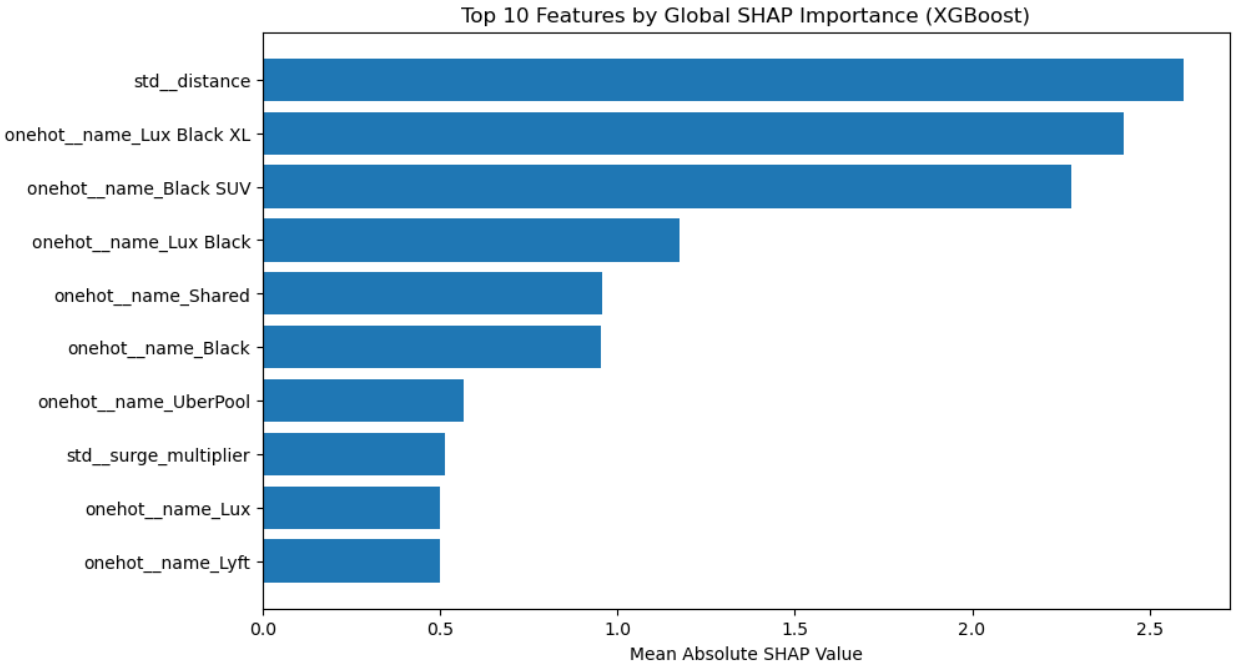

Figure 11: Top 10 Features by Global SHAP Importance

Additionally, we also tried local SHAP importance for different data indices, which allows a deeper exploration of individual predictions, highlighting how specific features contributed to deviations from the baseline (average price). Below are the three force plots for data indices 0, 100, and 200. Overall, Distance consistently contributes positively to the predicted price across all three indices, confirming its critical role in determining ride fares. Again, ride types still has a significant impact. Premium options like "Black SUV" and "UberXL" increase prices, while common options reduce prices. Similar to the global results, weather features such as wind speed make minor contributions but are not as influential as ride-specific factors.

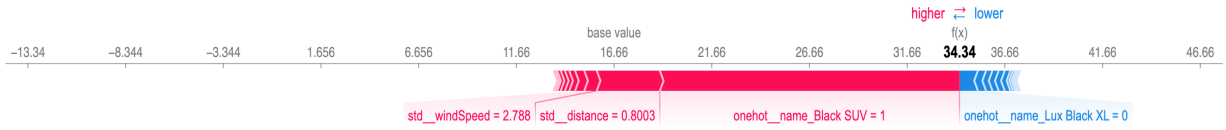

Figure 12: SHAP Local Value for Indice=0

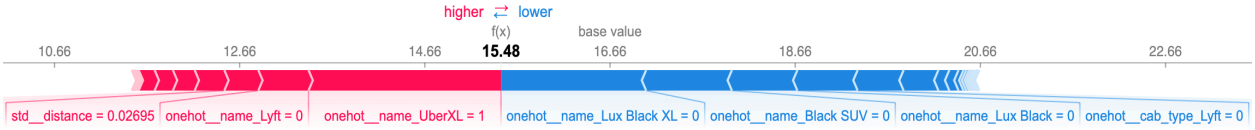

Figure 13: SHAP Local Value for Indice=100

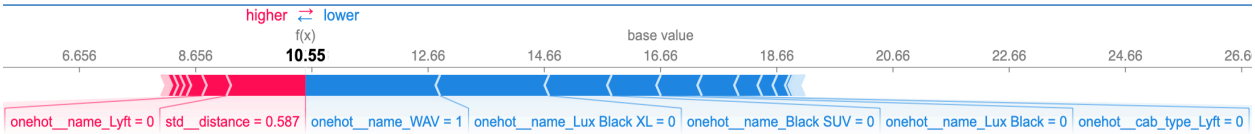

Figure 14: SHAP Local Value for Indice=200

4. Outlook

While our current model captures individual feature effects, interaction terms like distance * surge_multiplier or time_of_day * surge_multiplier could better capture non-linear relationships and synergies between features. For instance, surge pricing might have a more pronounced effect during peak hours, and these interaction terms could enhance accuracy.

Another limitation of this project is that the current machine learning pipeline relies on pre-existing features without significant feature engineering. Additional transformations, such as clustering pickup/drop-off locations to create “zones” or extracting peak hours dynamically, could reveal spatial and temporal patterns in pricing.

Meanwhile, in the future, we may try to collect real-time or historical traffic data, which could provide insights into congestion levels. The real-time or historical traffic data likely impact pricing, especially during peak hours. Additionally, geospatial data can also be helpful in predicting ride prices. Beyond latitude and longitude, features such as road networks, pickup/drop-off neighborhood types (commercial, residential), and zone-based demand aggregations could enhance spatial modeling.

References

[1] https://ijrpr.com/uploads/V4ISSUE12/IJRPR20241.pdf

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

[3] https://github.com/vummanenidilip/Uber-and-Lyft-Dataset-Boston-MA

GitHub Repo

https://github.com/EricZhangJr/Data1030Project