APPLIED MACHINE LEARNING

BUAN 6341.003 – S25

Prof: Ziyi Cao

Group Project Report

Group Members:

Sri Vamsi Kota – skk24001@utdallas.edu

Amruth Pai Thukaram – axt230147@utdallas.edu

Prajwal Rudresh – pxr240019@utdallas.edu

Gautam Pai – gxp240001@utdallas.edu

Kushagra Jain – kxj230017@utdallas.edu

Vinayak Jaiwant Mooliyil - vxm230084@utdallas.edu

Table of Contents

Executive Summary

This project applies machine learning to predict ride-sharing fares for Uber and Lyft with a focus on real-world usability and strategic insight. Rather than just modelling prices, the team prioritized accuracy, scalability, and interpretability—testing multiple models and tuning them to find the best performers.

Ensemble methods like Random Forest and XGBoost stood out for their ability to capture complex relationships in pricing data, outperforming linear approaches. The project also highlights how thoughtful preprocessing—such as encoding and scaling—can dramatically improve model performance.

What sets this work apart is its practical relevance. The team translated technical results into meaningful insights, like how fare patterns shift by day or how pricing strategies differ by service tier. Their approach offers a clear, replicable framework for using machine learning not only in transportation but in other industries facing dynamic pricing challenges.

Introduction

The objective of this project is to develop a predictive model that compares the cost-effectiveness of Uber and Lyft in specific regions based on factors such as weather (rainfall, temperature), rush hours, and historical pricing data. Additionally, we aim to forecast the number of customers choosing either service under varying conditions, providing valuable insights for both ride-sharing companies and users seeking optimal transportation options. **The Dataset has been sourced from Kaggle.**

Data Summary

The Uber and Lyft Cab Prices Dataset offers insights into ride-sharing fare trends by analysing various factors that influence pricing. It includes ride details from two leading ride hailing services, Uber and Lyft, capturing key elements such as fare amounts, ride distances, timestamps, service categories, and surge pricing multipliers. Additionally, the dataset incorporates weather data, allowing an examination of how conditions like temperature and precipitation impact ride costs. Since ride sharing companies utilize dynamic pricing models, fares fluctuate based on factors like demand, availability, and external conditions. Gaining a deeper understanding of these pricing mechanisms can help riders make informed decisions to minimize costs, while also enabling data analysts to explore pricing trends and their underlying drivers.

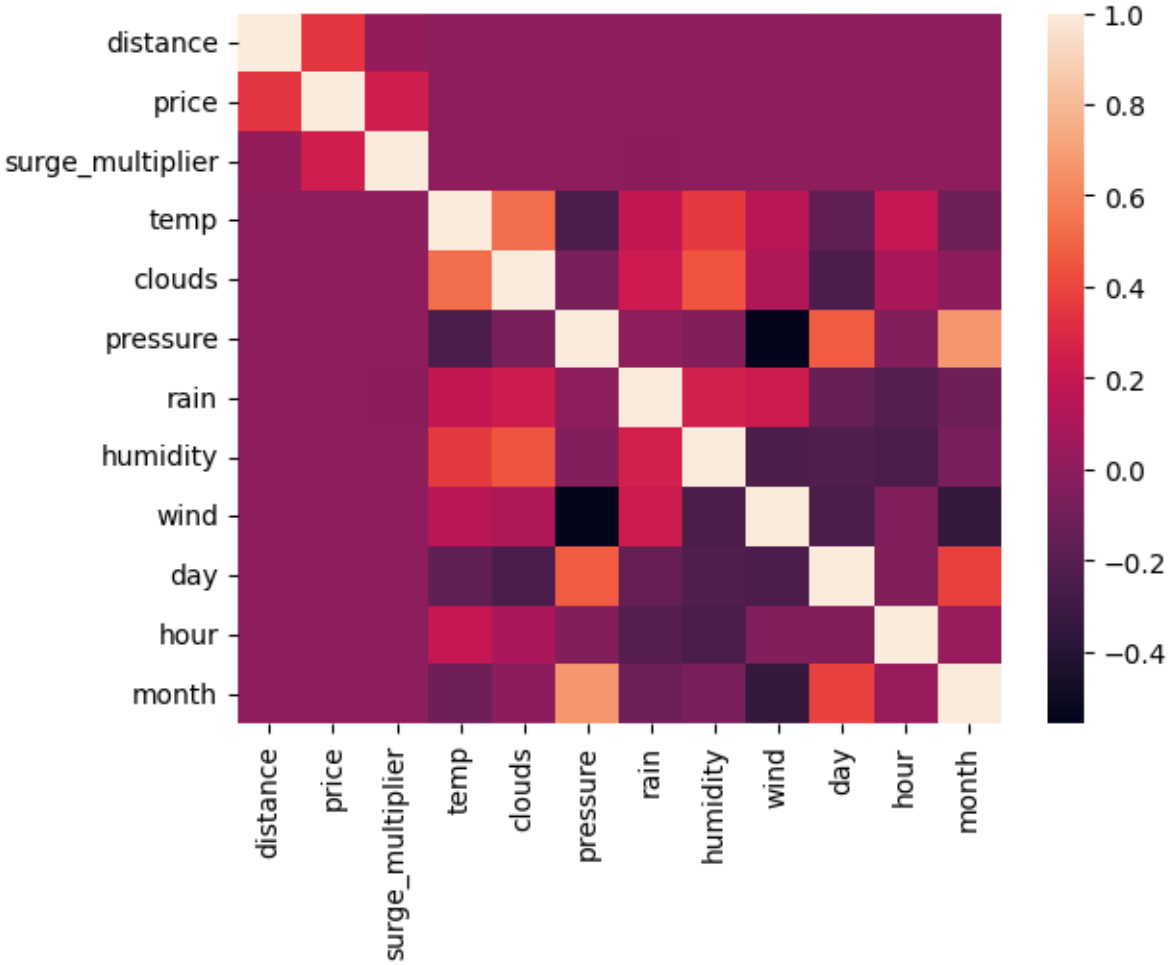
Data Description

The dataset contains Numerical variables as well as Categorical Variables. Since we're evaluating the effects of factors such as weather, distance, time of the day and days we retain only these for further analysis.

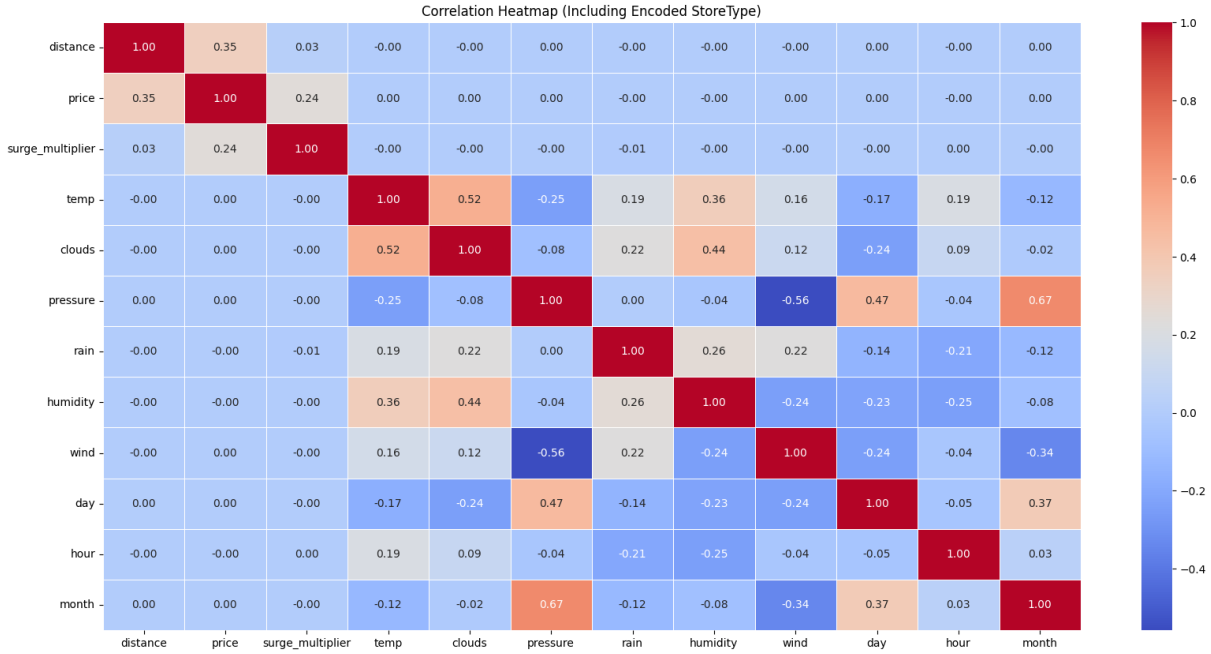
Numerical Variables	Categorical Variables
Distance	Cab-Type
Cloud	Destination
Date Time	Source
Humidity	Name
Pressure	Location
Rain	
Surge Multiplier	
Temperature	

Exploratory Data Analysis

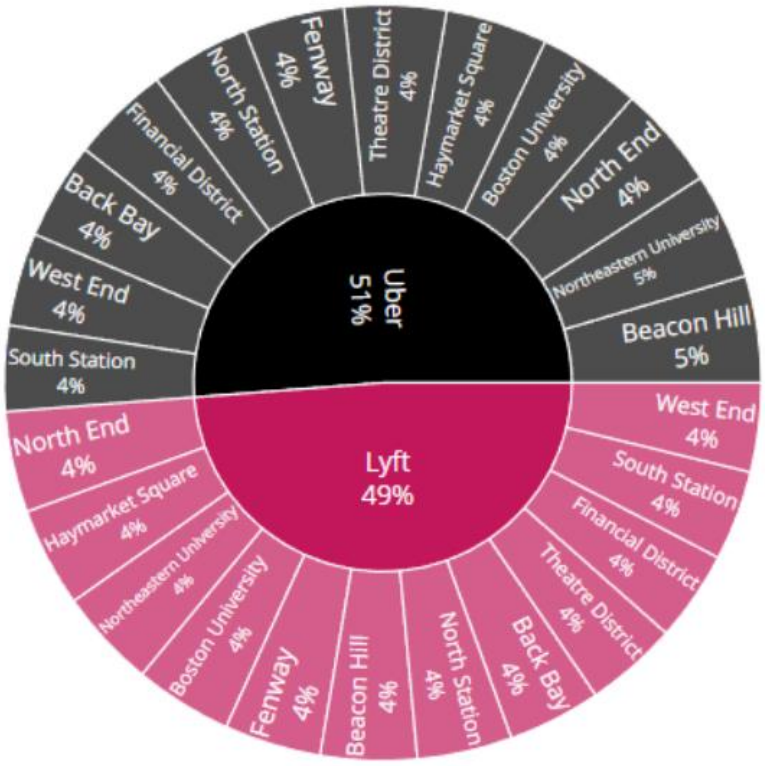
First, we used the Correlation Matrix to find what all factors affect the Cab prices. Here we can see that Distance and Surge Multiplier has significant impact on Cab price, and all other factors has very little effect on prices.



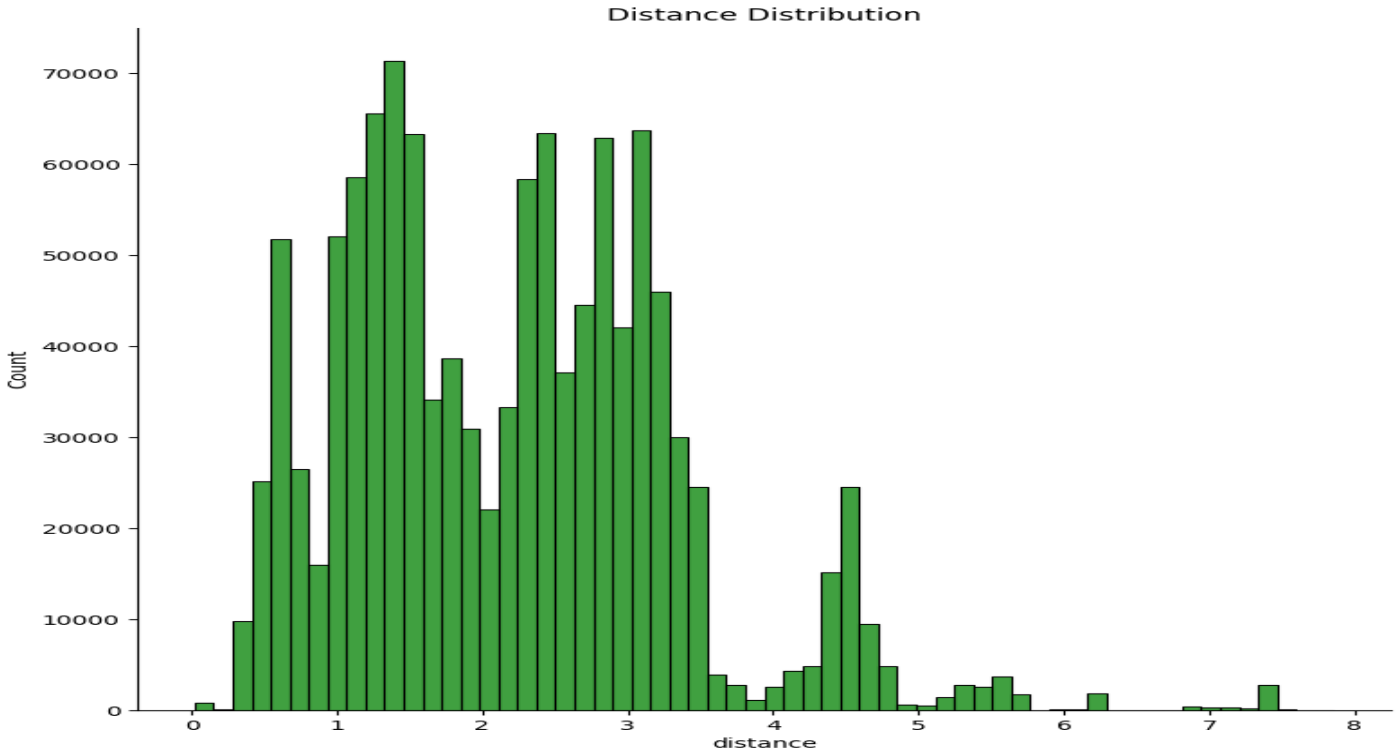
For better analysis as to how much does each factor correlate to cab price



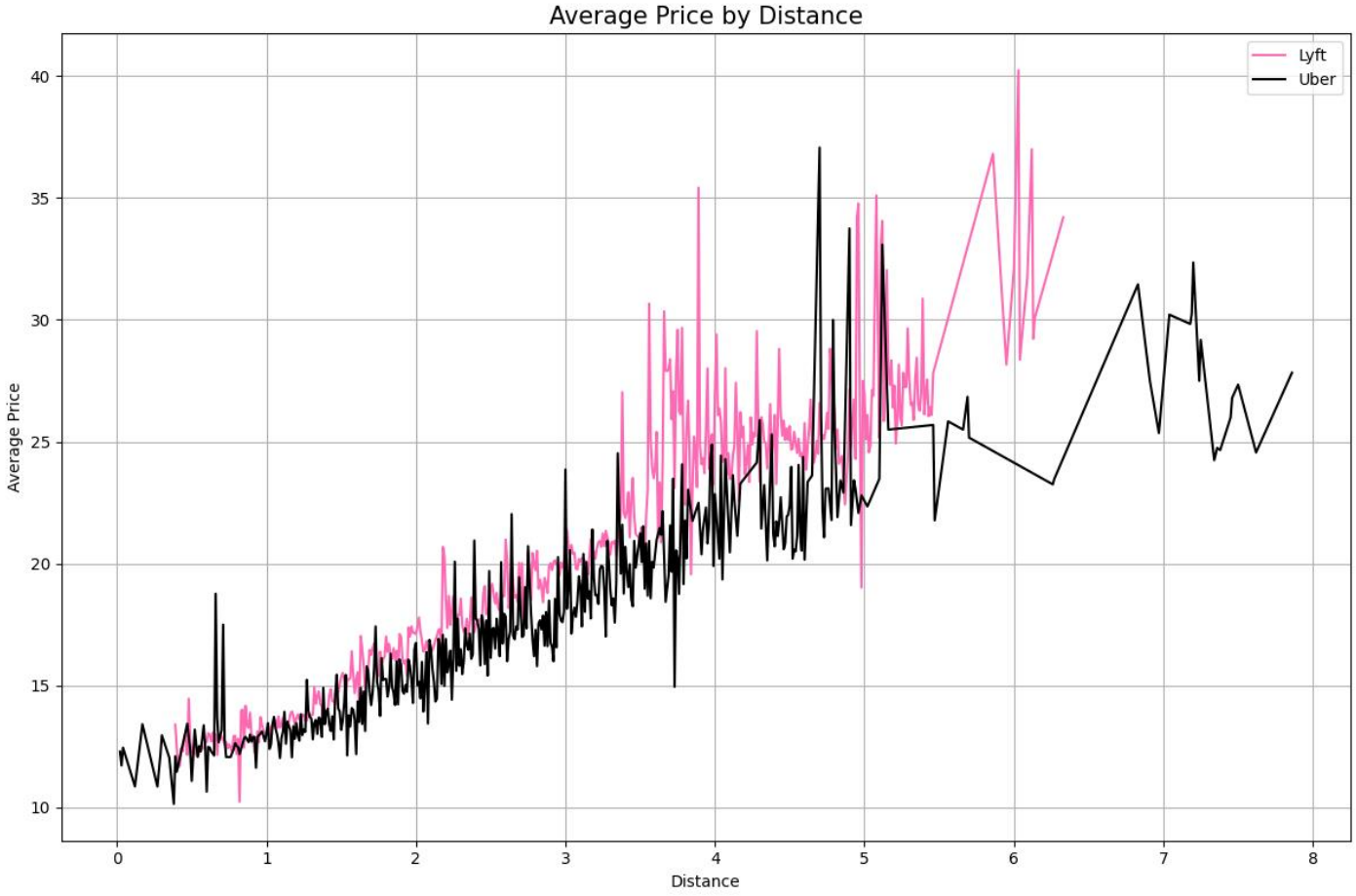
The total number of cab booking are relatively similar in both cases.



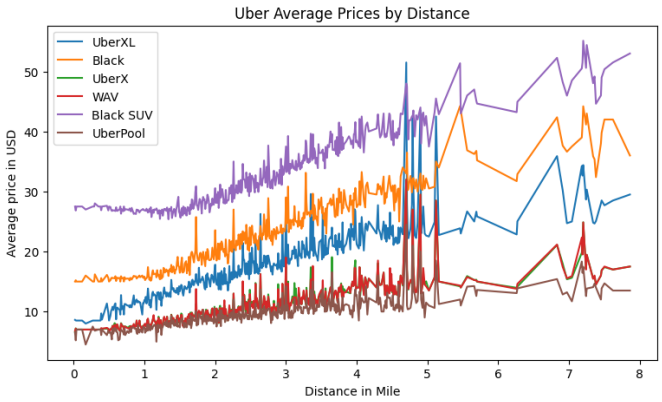
The overall Price Distribution relative to distance distribution is shown in below diagram

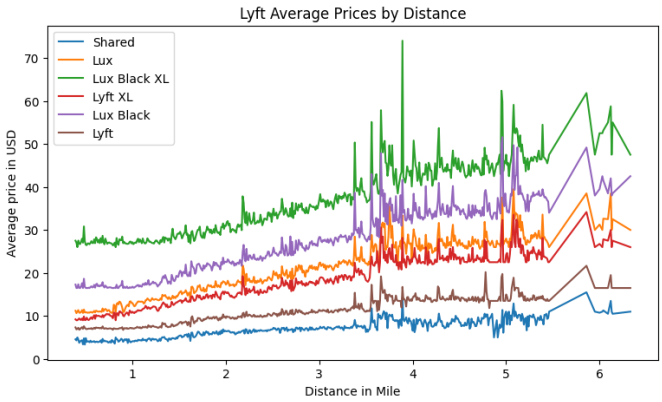


Looking further into how each service differs in pricing, we get the below diagram. While both have a linear trend, Lyft is more steeper meaning Lyft is costlier when booking for longer distance than Uber.



Drilling down further, in each cab types offered by the two services we get the following diagram.



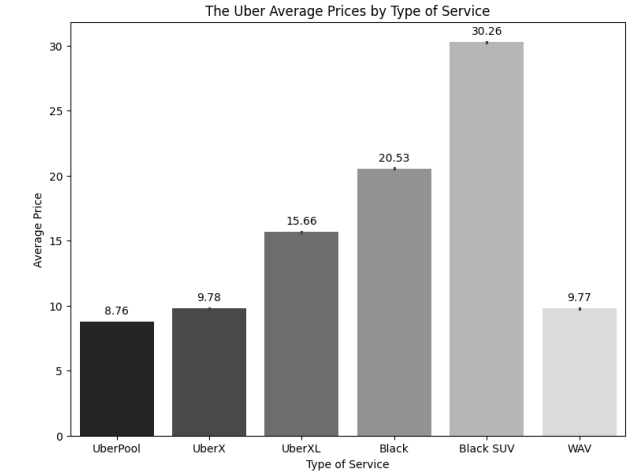


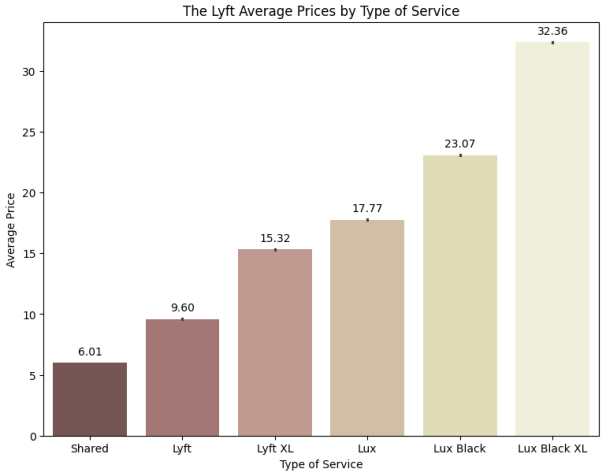
The average price comparison of cab type in both Uber and Lyft are shown below.

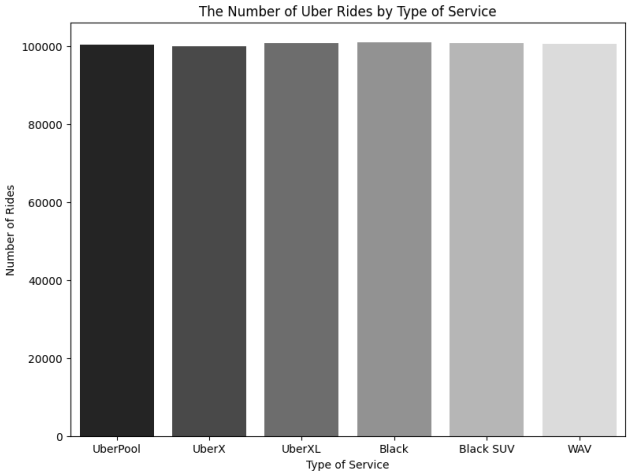
From the above and below diagram, we can understand that the services range from Uber Pool (Economy) to Black SUV (Luxury Premium) in Uber. Shared (Economy) to Lux Black XL (Luxury Premium) in Lyft.

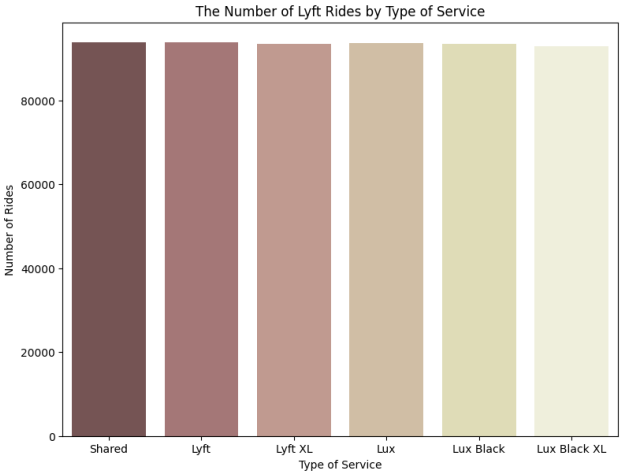
In Economy section, Lyft (Shared) is cheaper on average than Uber (Uber Pool).

For Premium Service, Uber (Black SUV) is cheaper than Lyft (Lux Black XL).

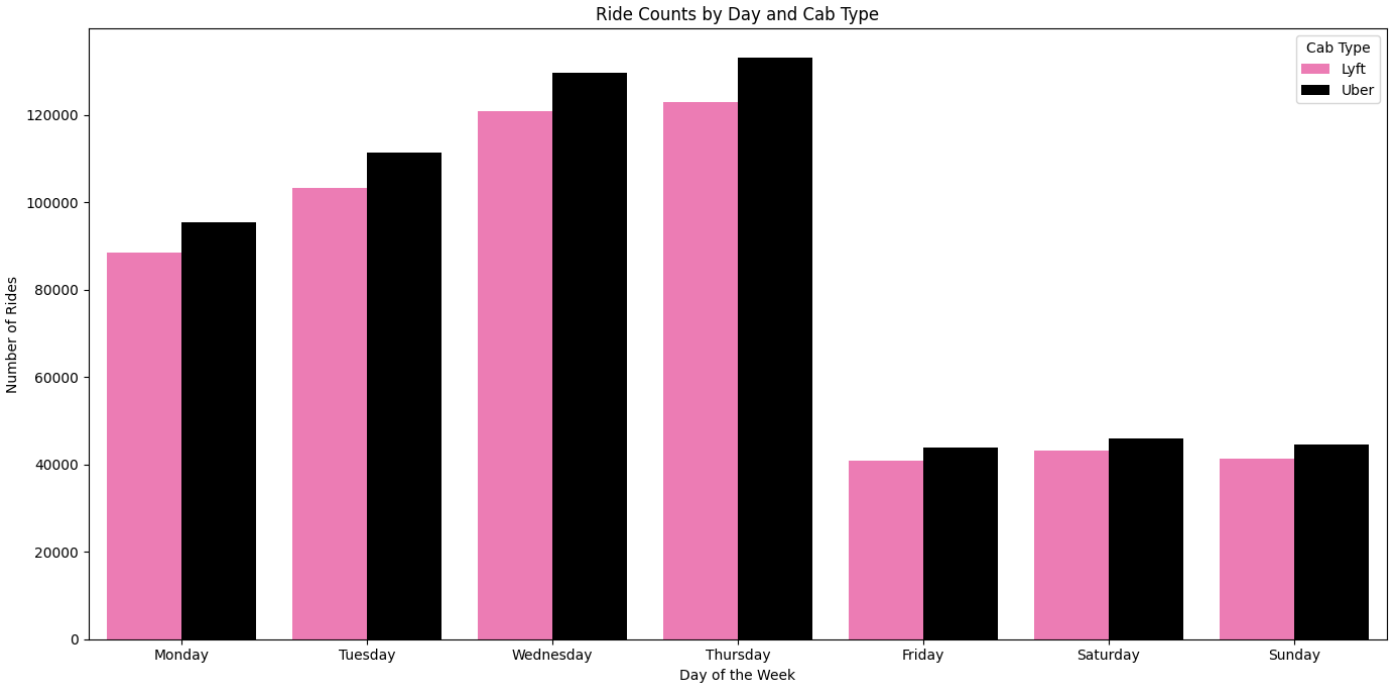








Further analysis on what days there are more bookings, we get to know that Weekdays has more booking rise each passing day, and it has a sharp fall on weekends.



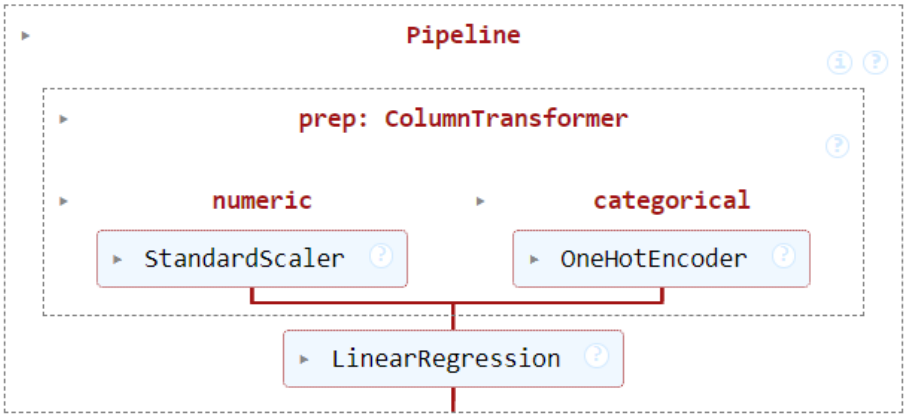
Class Variable Proportions

1. Cab Type:
    a. Uber - 56%
    b. Lyft - 44%
2. Destination:
    a. Financial District - 8%
    b. Theatre District - 8%
    c. Other - 83%
3. Source:
    a. Financial District – 8%
    b. Theatre District - 8%
    c. Other - 83%
4. Cab Type:
    a. Uber XL – 8%
    b. WAV - 8%
5. Surge Multiplier:
    a. 1.0-1.4 Surcharge - 96.97%
    b. 1.24 – 1.28 Surcharge – 1.59%

Preprocessing the Data

Here we have performed the following preprocessing steps on the dataset.

1. One Hot Encoding: Categorical columns are replaced with One Hot Encoded column subset.
2. Standard Scaling: All numeric columns are scaled down such that the mean is zero and majority of the values are withing one Standard Deviation.
3. Imputing Null Values: Removed rows that had null values.
4. We've used Column Transformer to perform the preprocessing; all pre-processed models use the same architecture.



Machine Learning Models

The data required employment of multiple Machine Learning Models for which the following have employed, to predict the cab price. To predict the accuracy of the model we're using multiple metrics to compare the models.

Models Used:

1. Linear Regression
2. Lasso Regression
3. Random Forest
4. XG Boost

Accuracy Metrics Used:

1. Mean Absolute Error (MAE)
2. Mean Squared Error (MSE)
3. R-squared Score

Our first prediction model is the linear regression model. The following is the summary of our data:

Dependent Variable: Price

Independent Variables: ['distance', 'cab_type', 'destination', 'source', 'surge_multiplier', 'name', 'date_time', 'temp', 'location', 'clouds', 'pressure', 'rain', 'humidity', 'wind', 'day', 'hour', 'month']

MODEL PERFORMANCE

1. Linear Regression

	Without Preprocessing	With Preprocessing
MSE	71.863	10.09
MAE	6.949	1.753
R^2 Score	0.172	0.883

The above table indicates that the Model Accuracy significantly increased after the preprocessing. Before preprocessing the model is able to explain only 17.2% of the variance in the dataset, whereas after preprocessing the model is able to explain 88.3% of variance. This model acts as the base model for further analysis and comparison.

2. Lasso Regression

	Results
MSE	6.176
MAE	1.752
R^2 Score	0.929

This is a hyperparameter tuned version of Linear Regression Model. Here we used GridSearchCV to find the best value for alpha. The alpha range is between 0.001 to 100 in the multiples of 10. The best alpha value is 0.001.

3. Random Forest Regression

This is an ensemble technique, which works based on majority votes for classification and average value for Regression also called Bagging Technique.

To hyper parameter tune the Random Forest Model, we have used GridSearchCV to find the best parameters and below are the results for the same.

	Hyper Parameter Range	Best Value
Max Depth	None,10,20	20
Min Sample Leaf	1,2	2
Min Sample Split	2,5	5
N-Estimators	50,100	100

For the best parameters that we have found below are the metrics for the model

	Results
MSE	2.367
MAE	0.965
R^2 Score	0.973

This is a significant improvement from Linear and Lasso Regression model. Model is able to explain 97.3% of variance in the dataset. Also, the MSE, MAE metrics are significantly improved compared to Linear Regression model and moderately better than Lasso Regression model.

4. XG Boost Regression

This is another ensemble technique, which works on the principle of Boosting method. All predictors are stubs of decision tree with max depth of 1. The losses of previous decision tree stub are fed into next decision tree stub.

	Hyper Parameter Range	Best Value
N-Estimators	100,150,200	200
Learning Rate	0.05,0.1,0.2	0.2
Max Depth	4,6,8	8

For the combination of best hyper parameters, below are the model metrics.

	Results
MSE	2.397
MAE	1.018
R^2 Score	0.972

This is significantly better than Linear Regression model. Model is able to explain 97.2% of variance in the dataset. However, Random Forest model performs marginally better when all the metrics are considered.

We can conclude that both Ensemble techniques can be used to predict the price and expect to achieve the correct price with near accuracy to actual price.

Key Takeaways

1. Although weather plays a role in Cab pricing it is not significant when compared with other metrics such as Day of the week, distance.
2. The pricing follows a linear trend when compared with distance. For longer travel, prefer public transport for cost effectiveness.
3. The prices are marginally more when compared to weekend prices. Number of booking rise in weekdays and see a sudden fall on weekends.
4. Ensemble techniques (Both bagging and boosting) are ideal for pricing models.

Learnings from Data Mining:

1. Data Preprocessing is crucial:

   The performance of models significantly improved after applying preprocessing techniques like one-hot encoding, standard scaling, and null value handling. For example, linear regression's R^2 score jumped from 0.172 to 0.883 after preprocessing.

2. Correlation Doesn't Always Imply Importance:
   While features like distance and surge multiplier had strong correlations with price, others like weather had minimal impact—emphasizing the need to validate feature importance empirically.
3. Model Choice Matters:
   Ensemble models such as Random Forest and XGBoost outperformed simpler models like linear and Lasso regression, with Random Forest achieving $R^2 = 0.973$, indicating it explained 97.3% of the variance in pricing.
4. Hyperparameter Tuning Pays Off:
   Using GridSearchCV to tune models (especially Random Forest and XGBoost) significantly improved performance, illustrating the value of rigorous model selection.
5. EDA Helps Discover Business-Relevant Patterns:
   Insights like Lyft being more expensive over long distances than Uber, and surge pricing being relatively rare (only ~3% of data), demonstrate how EDA can drive real-world understanding.

Analysis & Insights for Business Managers

1. Surge pricing is rare (~3%), so fine-tuning pricing models around core features like distance and time is more valuable than weather-based adjustments.
2. Lyft charges more than Uber for longer distances, suggesting an opportunity for Uber to market long-distance trips more aggressively or Lyft to reassess pricing structure.
3. Demand rises during weekdays and drops on weekends, so targeted promotions or driver incentives could be designed to balance this demand curve.
4. Analysis revealed that Uber's premium options (like Black SUV) are cheaper than Lyft's (Lux Black XL), offering a competitive advantage Uber can market more effectively.
5. For internal use, adopting ensemble models like Random Forest in real-time fare estimation systems can enhance price accuracy and customer satisfaction.

References:

Project Repository: Link

Data Source: Link