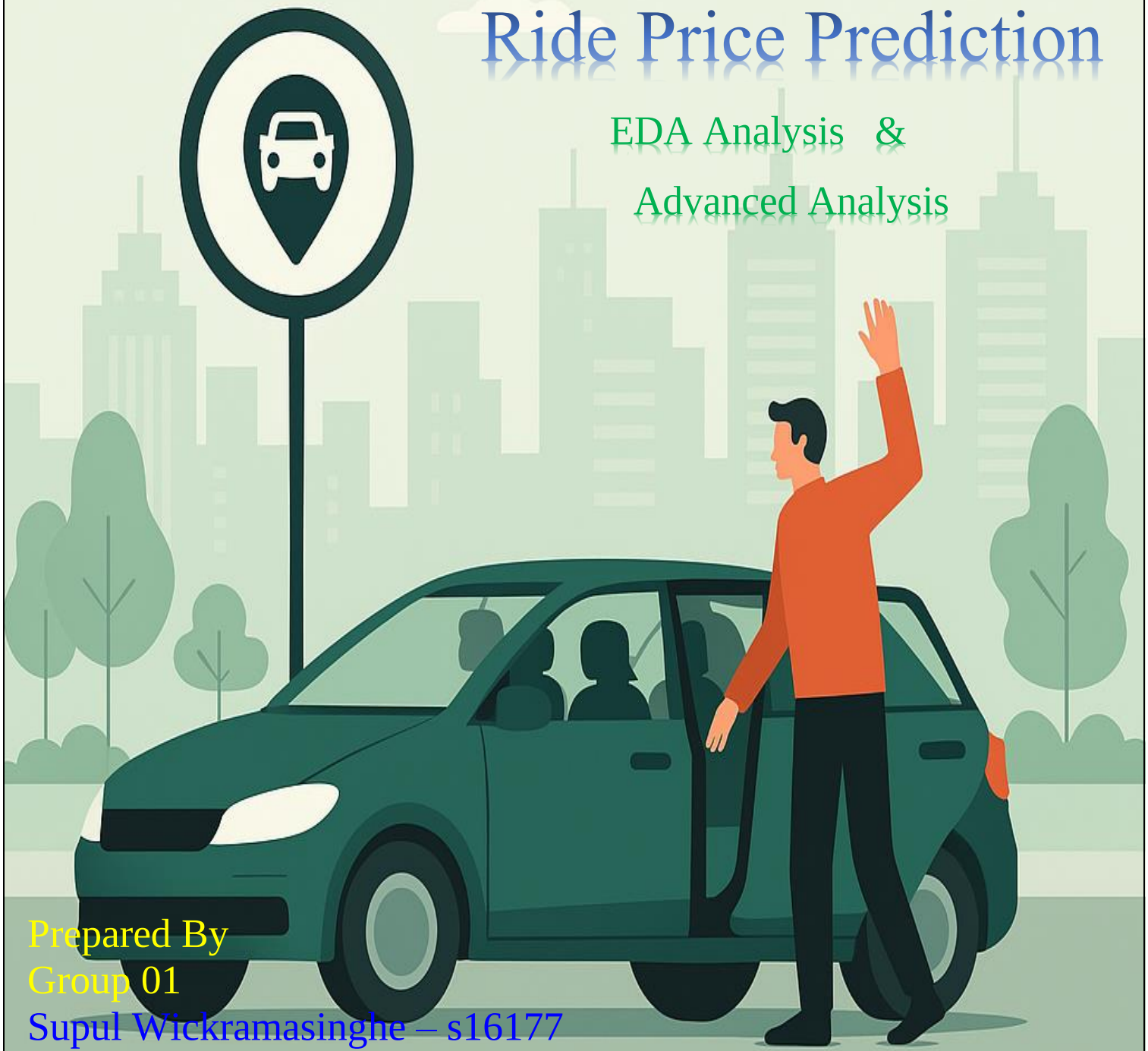
Dynamic Ride Price Prediction

EDA Analysis & Advanced Analysis



Prepared By
Group 01
Supul Wickramasinghe – s16177
Thenuka Yatawra – s16383
Vayani Kavindya –s16322
Sanduni Fonseka – s16026

ABSTRACT

This report presents a data-driven methodology for developing a dynamic pricing model tailored for the ride-sharing industry. Leveraging historical ride data, we apply advanced machine learning algorithms to forecast optimal fares in response to real-time market dynamics. The analysis is grounded on a rich dataset obtained from Kaggle, encompassing various pricing-related features. Our findings offer valuable insights into fare optimization strategies, aiming to enhance operational efficiency and pricing transparency. This research holds practical significance for ride-sharing platforms, benefiting stakeholders including customers, drivers, business strategists, and academic researchers.

Contents

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Dynamic pricing is a strategy where prices are adjusted in real-time based on market demand, supply conditions, competitor pricing, customer behavior, and other external factors. Commonly used in industries such as airlines, hospitality, e-commerce, and ride-sharing, this approach allows businesses to optimize revenue and resource utilization. In the context of ride-sharing, dynamic pricing helps balance rider demand and driver availability—ensuring better service efficiency and fair compensation. With the rise of data availability and machine learning technologies, dynamic pricing models have become more accurate and responsive, enabling companies to make smarter, real-time pricing decisions. This report delves into the complex landscape of dynamic pricing in the ride-sharing industry, aiming to develop a predictive model that captures the multifaceted elements influencing fare determination. By exploring historical patterns and market behavior, the study seeks to provide a deeper understanding of the mechanisms driving pricing strategies and their impact on stakeholders across the ecosystem.

DESCRIPTION OF THE QUESTION

The global ride-sharing market has experienced significant growth in recent years, driven by increasing demand for convenient, affordable, and technology-driven transportation solutions. A key factor in this expansion has been the widespread adoption of mobile app-based ride-sharing services, which allow users to seamlessly book rides from nearby drivers at competitive rates. These services often prove to be a more cost-effective alternative to traditional taxi services, further fueling their popularity(Source: https://www.fortunebusinessinsights.com/ride-sharing-market-103336). Dynamic pricing has emerged as a standard practice in the ride-sharing industry, first introduced by Duke University and widely adopted by companies like Uber and Lyft since 2015. This pricing model divides metropolitan areas into smaller hubs, where fares are adjusted in real-time based on demand fluctuations. By doing so, it ensures optimal pricing for both passengers and drivers, balancing affordability and profitability. (Source: https://www.fuqua.duke.edu/duke-fuqua-insights/algorithms-behind-pricing-your-ride). In light of these developments, a ride-sharing company aims to enhance its pricing strategy by incorporating a

dynamic pricing approach that adjusts fares based on real-time market conditions. Currently, the company determines fares solely based on ride duration, which limits its ability to respond to changing demand and other influencing factors. To address this, the company plans to leverage data-driven techniques to analyze historical data and develop a predictive model that adaptively sets prices in response to dynamic market conditions.

Objectives:

1. **Identify Key Factors**: Determine the key factors that influence dynamic pricing, uncover hidden patterns, and analyze the relationships between these variables to understand their impact on ride fares.
2. **Develop a Predictive Model**: Build a machine learning-based dynamic pricing model to predict and optimize ride fares in real-time, ensuring responsiveness to market conditions.

This project aims to create a robust and adaptive pricing mechanism that enables the ride-sharing platform to efficiently respond to market dynamics while enhancing service quality for both riders and drivers.

DESCRIPTION OF THE DATASET

The dynamic pricing dataset from Kaggle has 1000 observations and ten variables, four of which are categorical. The ride sharing app is popular and operates on a hub and spoke arrangement, with each location connected to a larger metropolitan area. The main response variable, 'Historical_Cost_of_Ride' refers to the fare determined simply based on ride duration. New variables 'Adjusted_Cost' have been added to improve fare computations by taking into account elements outside ride duration.

Table 1:Description of variables

Variable	Data Type	Description
Number_of_Riders	Numerical-Discrete	Number of riders available in a specific area.
Number_of_Drivers	Numerical-Discrete	Number of drivers available in a specific area.
Location_Category	Categorical-Nominal	Category of the ride location.
Customer_Loyalty_Status	Categorical-Ordinal	Category of the customer loyalty status.
Number_of_Past_Rides	Numerical-Discrete	Number of past rides taken by the customer.
Average_Ratings	Numerical-Continuous	Average ratings given to driver in that specific location by riders from that location.
Time_of_Booking	Categorical-Nominal	Time of the day ride was booked.
Vehicle_Type	Categorical-Ordinal	Type of the vehicle requested for the ride.

Expected_Ride_Duration	Numerical-Discrete	Expected duration of the ride.
Historical_Cost_of_ride	Numerical-Continuous	Cost of the ride borne by the rider, including any penalties incurred.
adjusted_ride_cost	Numerical-Continuous	Historical cost adjusted for number of riders and no of drivers. (Newly defined variable)
Profit_percentage	Numerical-Continuous	Change in profit percentage when the company shifts to dynamic pricing strategy from static pricing strategy. (Newly defined variable)

FEATURE ENGINEERING

As we mentioned earlier, we defined a variable for our data set.

1. **adjusted_ride_cost** :- The originally present variable in the dataset " Historical_Cost_of_Ride" only depends on the predictor variable "Expected_Ride_Duration". So, by further referencing we identified that variable as Static cost for each ride in the dataset. For implementation of dynamic pricing strategy, we created a new variable "Adjusted_Cost" by combining "Historical_Cost_of_Ride" with demand and supply levels. It will capture high-demand periods and low-supply scenarios to increase prices, while low-demand periods and high-supply situations will lead to price reductions. This newly defined variable acts as response variable for modelling the dynamic price. (formulas used for creating new variable will be on appendix).
2. **Profit_percentage**:- Change in profit percentage when the company shifts to dynamic pricing strategy from static pricing strategy. This variable is used as the response variable for modelling change in profit percentage.

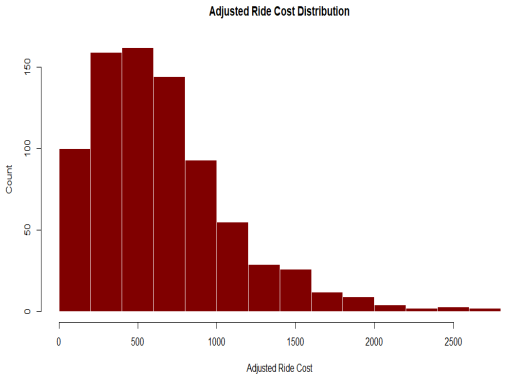
$$\text{Profit percentage} = \frac{\text{adjusted ride cost} - \text{Historical cost of ride}}{\text{Historical cost of ride}} * 100$$

DATA PRE – PROCESSING

- ✓ The dataset was checked for duplicates and missing values. There was no duplicate or missing values.
  - ✓ The dataset split into training and test sets. The training dataset contains 800 observations.
  - ✓ Checked for outliers and there were not many significant outliers; so we decided to keep outliers.
  - ✓ We remove the historical ride cost variable from the predictor space since it is directly used to produce the new response variable.

IMPORTANT RESULTS OF DESCRIPTIVE ANALYSIS

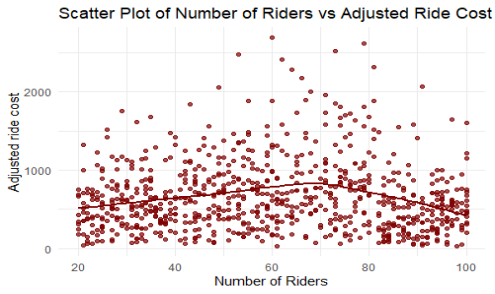
Distribution of Main Response Variable: Adjusted Ride Cost



The adjusted ride cost variable exhibits a left-skewed distribution with a higher frequency of lower cost values, as the majority of customers tend to schedule rides at lower prices, resulting in a mean of 655.43 exceeding the median value of 573.37.

Figure 1:Distribution of adjusted ride cost

Relationship of Predictor Variables with Adjusted Cost



The analysis reveals several key findings. The graph of the number of riders versus adjusted ride cost does not display the upward trend expected of a supply curve but, a clear demand curve emerges between adjusted cost and the number of drivers, indicating a decrease in price with decreasing demand and available drivers.

Figure 2: Scatter Plot of Number of ride Vs. adjusted ride cost

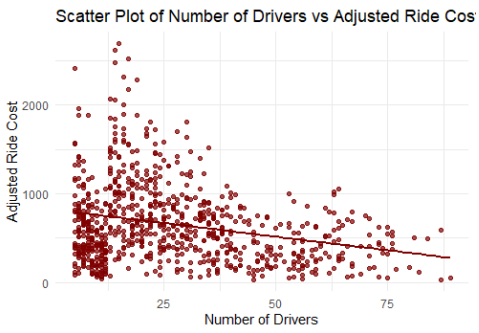
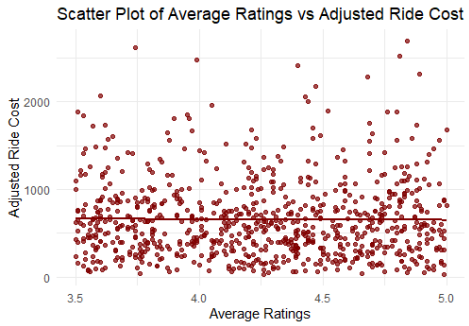


Figure 3: Scatter Plot of Number of ride Vs. adjusted ride cost

And also the analysis confirms the expected positive relationship between ride duration and fare, yet average ratings do not seem to directly affect the fare.

Figure 4:Scatter Plot of average ratings Vs. adjusted ride cost



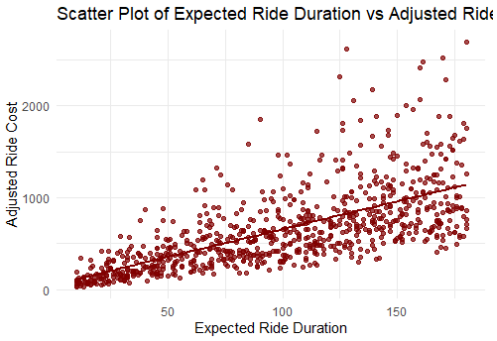


Figure 5:Scatter Plot of Expected ride duration Vs. adjusted ride cost

The bar plots comparing average adjusted ride cost across categorical predictors highlight key variations. "Premium" vehicle types and "Morning/Night" bookings show higher average costs. "Location_Category" has subtle differences, with "Rural" potentially slightly higher. "Customer_Loyalty_Status" shows similar average costs across groups.

Figure 6:The bar plot comparing adjusted ride cost across categorical variables

A comparison of historical cost and adjusted ride cost reveals a strong, direct linear relationship, which is expected given that the adjusted ride cost is derived from the historical ride cost.

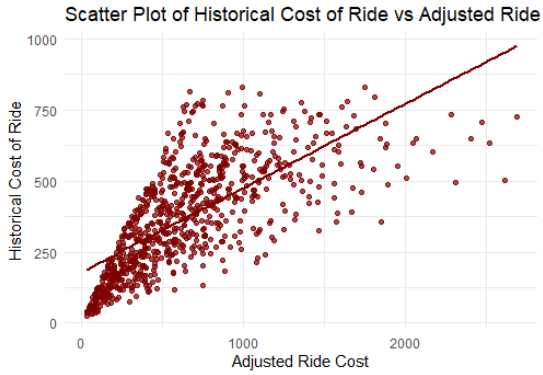
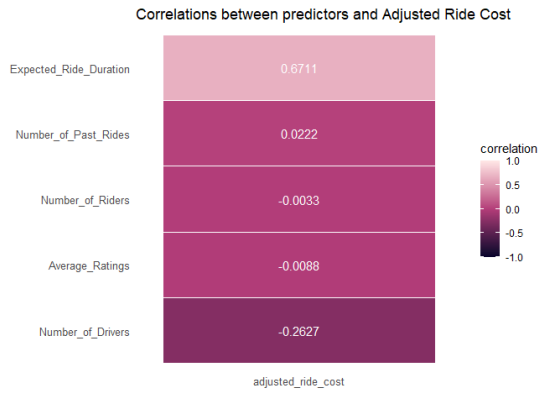


Figure 7:Scatter Plot of historical cost of ride Vs. adjusted ride cost

Figure 8:Correlation between predictors and adjusted ride cost

The correlation plot indicates that adjusted ride cost is strongly positively correlated with expected ride duration (longer rides, higher cost) and weakly negatively correlated with the number of drivers (more drivers, slightly lower cost). The correlations with the number of past rides, number of riders, and average ratings are negligible, suggesting they have little direct linear impact on the adjusted ride cost this confirms the findings of the scatter plots.



Correlation between predictor variables

The predictor correlation heat map shows a moderately strong positive correlation (0.62) between the Number of Riders and Drivers, indicating potential multicollinearity due to the close relationship between demand and supply. Other predictor correlations are generally very weak.

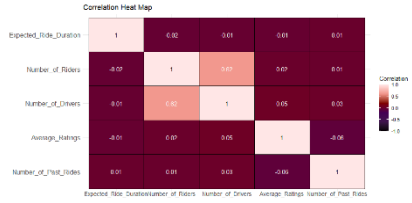


Figure 9:Correlation heat map

Cluster Analysis

Given that our dataset comprises numerical and categorical variables, we employed Factor Analysis of Mixed Data (FAMD) to reduce dimensionality and explore underlying patterns.

And also the following Scree Plot of FAMD reveals that the first two components explain 27.9% of the variance.

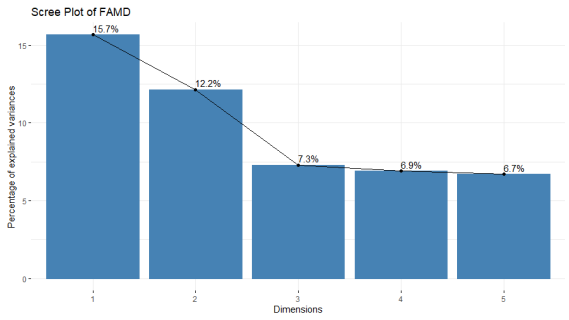


Figure 10:Scree Plot of FAMD

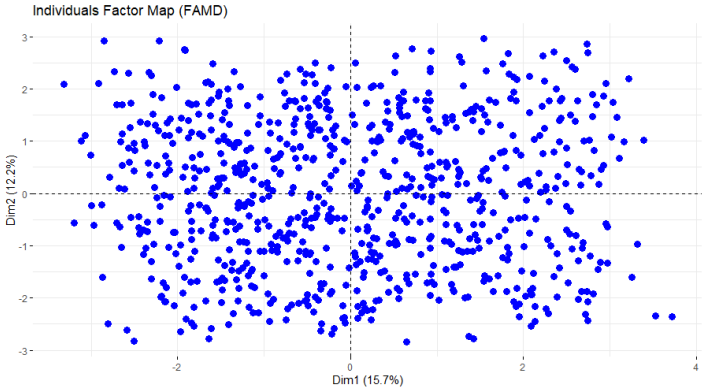


Figure 11:Individual factor map(FAMD)

The Individuals Factor Map (Figure 11), which projects our data points onto the first two principal components, and it shows no clear clustering.

This absence of separation indicates that distinct group structures are not visible within the primary dimensions of the mixed data.

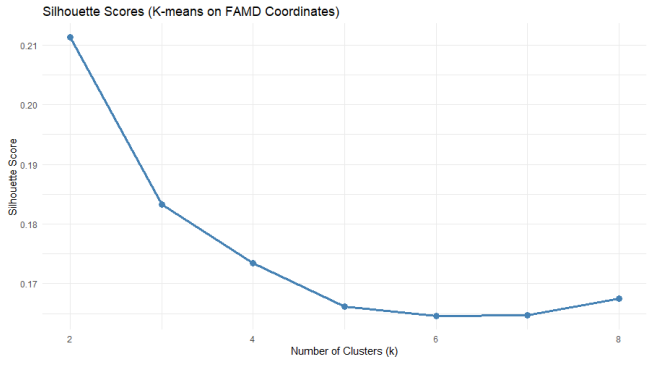


Figure 12:Silhouette Scores

To further investigate this, we performed K-means clustering on the coordinates derived from the FAMD and evaluated the cluster quality using silhouette scores. The resulting Silhouette Score plot shows consistently low average silhouette scores across different numbers of hypothesized clusters, indicating that the individuals do not form well-defined, groups in the FAMD-reduced space.

Therefore, based on the visual inspection of the uniformly distributed points in the Individuals Factor Map and the poor cluster quality metrics obtained from the FAMD coordinates, we concluded that there is no strong evidence of cluster structures within our dataset.

IMPORTANT RESULTS OF ADVANCED ANALYSIS

In the advanced analysis phase, our objective was to develop a machine learning models to predict the adjusted ride cost. During data preprocessing, no significant outliers were detected, allowing us to proceed with the full dataset intact. Cluster analysis was conducted, and results indicated the presence of only a single cluster across all data points. Therefore, we opted to build the model using the entire dataset. To prepare the data for analysis, numerical features were standardized, and categorical variables both nominal and ordinal were appropriately

encoded using One Hot Encoding. This ensured that all variables were suitably transformed for the modeling process.

Multiple linear regression (MLR) – Best subset selection

Table 2:Evaluation Matrix of MLR

	RMSE	R^2
Training	290.88	0.5669
Test	296.27	0.5756

To develop the multiple linear regression model, we employed best subset selection to identify the optimal combination of predictors. The final model selected includes four variables: *Number of Riders*, *Number of Drivers*, *Expected Ride Duration*, and *Vehicle Type (Premium)*. This selection was based on minimizing the prediction error and improving model interpretability. The coefficients of the model suggest that an increase in the number of riders and the use of premium vehicles positively influence the adjusted ride cost, while an increase in the number of drivers has a negative effect.

Furthermore, to check the validity of the given model we conducted a residual analysis.

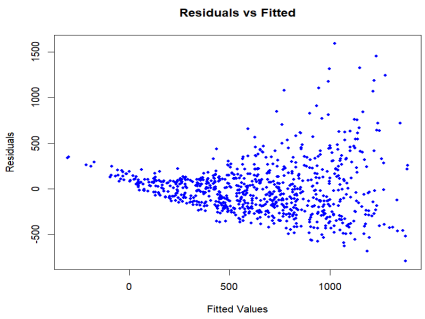
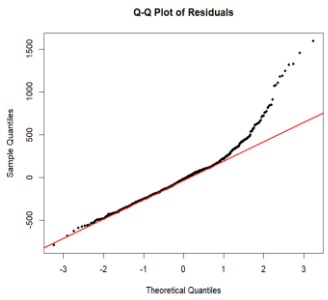


Figure 13:Residual Vs. Fitted values plot

1. Linearity
   he correlation plot underscores small associations between predictors except “expected ride duration” and the response variable, indicating a limited strength of linearity in the model.
2. Independence & Homoscedasticity



The residual vs. predicted plot reveals a lack of random scattering around the zero-center line, accompanied by a corn shape. This implies a violation of both homoscedasticity and independence of residuals.

Figure 14:Q-Q Plot of Residuals

3. Multicollinearity

   Since the all the variable doesn't show VIF values greater that 10 we can conclude that multicollinearity doesn't exit.

4. Multivariate Normality

   Q-Q plot (Figure 14) indicates the departure from normal distribution assumptions in the model's residuals.

Regularization Methods

Following the multiple linear regression (MLR) analysis, the implementation of regularization techniques is recommended. These methods, such as Ridge and Lasso regression, by introducing a penalty for model complexity, regularization enhances the robustness and generalizability of the model, leading to more reliable predictions on unseen data.

Table 3:Evaluation Matrix of Regularization methods

Model	Training		Test	
	RMSE	**R^2**	**RMSE**	**R^2**
Ridge	291.40	0.5654	298.76	0.5684
Lasso	290.94	0.5668	297.44	0.5722
Elastic Net	290.88	0.5669	297.64	0.5717

Regularization techniques—Lasso, Ridge, and Elastic Net regression—were employed to refine the prediction of adjusted ride cost and address limitations of the MLR model. Lasso regression, optimized at $\lambda = 6.54$, effectively identified key predictors by shrinking less important coefficients to zero, highlighting the influence of ride duration, number of riders, Premium vehicle type, and customer category. Similarly, the Elastic Net model (optimal $\lambda = 9.90$) confirmed these drivers while additionally distinguishing cost differences across urban and rural locations. Ridge regression, while retaining all predictors, emphasized similar trends by penalizing large coefficients and reducing model complexity. Collectively, these methods enhanced model interpretability and robustness by mitigating overfitting and multicollinearity, ultimately leading to more reliable insights into the factors affecting ride costs.

Tree-Based Methods

Tree-based methods are a class of non-linear models used for both regression and classification tasks. These models work by recursively partitioning the data into subsets based on feature values, forming a decision tree structure. One of the key advantages of tree-based methods is their ability to capture complex relationships and interactions between variables without requiring explicit specification, unlike traditional linear models. They are also intuitive and easy to interpret, making them useful for understanding the underlying structure of the data. Common tree-based algorithms include Decision Trees, Random Forests, and Gradient Boosted Trees, each

offering different strengths in terms of accuracy, interpretability, and resistance to overfitting. These methods provide a powerful alternative to linear models, particularly when the relationship between predictors and the response variable is highly non-linear or involves complex interaction.

Random Forest

Random Forest is a machine learning algorithm that leverages multiple regression trees as its base learning model.

The key assumption behind Random Forest is that each tree will make different mistakes, so aggregating the results of multiple trees will lead to a more accurate model than relying on a single tree. By training multiple decision tree regressors on various sub-samples of the dataset and using averaging, Random Forest improves predictive accuracy and helps control overfitting. Upon training the model and evaluating its performance, we observe that the training R^2 value is 0.9435, which is considerably higher than the testing R^2 value of 0.8312. This discrepancy suggests overfitting, as the model performs better on the training data. However, comparing the RMSE values, the Random Forest model achieves a lower RMSE on the training set (116.72) compared to the testing set (180.20), indicating that while the model generalizes reasonably well, there is room for further improvement in handling unseen data.

Table 4:Evaluation Matrix of Random Forest

	RMSE	R^2
Training	116.72	0.9435
Test	180.20	0.8312

XG BOOST

XG Boost is a powerful and widely used open-source machine learning algorithm that builds better models by combining decision trees with gradient boosting. Although it's a boosting method, it also incorporates some elements of bagging by training multiple decision trees and combining their outputs. This approach allows XG Boost to learn more efficiently than many other algorithms, especially when working with datasets that contain a large number of features.

Because of its strong performance often surpassing that of the Random Forest Classifier XG Boost was also applied to the Dynamic Pricing Dataset in this analysis. After training the model and tuning its parameters, it was observed that the test R^2 score improved. Here we tuned parameter using grid search cross validation.

Table 5:Evaluation Matrix of XG Boost

	RMSE	R^2
Training	99.09	0.9497
Test	156.87	0.8810

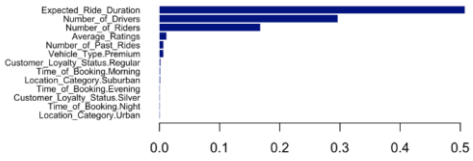


Figure 15:Feature Importance Plot of XG Boost

The variable importance plot showed that 'Expected ride duration', 'Number of riders', and 'Number of drivers' were the most influential variables in the model. The remaining variables contributed very little. Based on this, we tried improving the model by removing the less important variables and tuning the parameters. However, this actually led to a drop in model performance, as reflected by a lower test R^2 value. As a result, we decided to keep all the predictors in the model. The XG Boost model with all predictors was ultimately chosen as the best-performing model.

ISSUED ENCOUNTED AND PROPOSED SOLUTIONS

Throughout our advanced analysis, we encountered several challenges that required strategic problem-solving. Fortunately, the pre-processing stage proceeded smoothly, with no significant data quality or formatting issues. However, during the modeling phase, we observed violations of key assumptions underlying the Multiple Linear Regression (MLR) model, such as non-linearity. To address these, we conducted thorough diagnostic checks, including residual analysis recognizing the limitations of MLR, we implemented regularization techniques Lasso, Ridge, and Elastic Net using cross-validation to fine-tune hyper parameters and reduce model complexity. This

helped improve model generalizability and robustness. Additionally, during our application of tree-based methods, we found that certain variables consistently exhibited zero feature importance. To enhance model efficiency and predictive performance, these variables were excluded, and the models were refitted accordingly. This iterative troubleshooting approach, involving critical evaluations and refinements at each stage, reflects our commitment to developing accurate, interpretable, and dependable models despite the analytical hurdles encountered.

DISSCUSION AND CONCLUSIONS

Table 6:Summary of all Evaluation Matrices

Model	Training		Test	
	RMSE	R^2	RMSE	R^2
MLR	290.88	0.5669	296.27	0.5756
Ridge	291.40	0.5654	298.76	0.5684
Lasso	290.94	0.5668	297.44	0.5722
Elastic Net	290.88	0.5669	297.64	0.5717
Random Forest	116.72	0.9435	180.20	0.8312
XG Boost	99.0859	0.9497	156.8695	0.8810

As uncovered during the stepwise Advanced Analysis, XGBoost Regressor is the best model for predicting adjusted cost for dynamic pricing strategy with a relatively high-Test R^2 and lowest testing RMSE among all

algorithms, along-side a minimal difference between train and test R^2 implying that control of over-fitting. Note that for predicting adjusted cost XG Boost model with all predictors is the best model.

Hence, the data-products to predict adjusted cost will be developed containing a back-end feature where the input variables of the factors affecting to adjusted cost will be analyzed using a relevant Hyper-parameter tuned XG Boost Regressor to give an output of the adjusted cost.

REFERENCE

1. https://www.fortunebusinessinsights.com/ride-sharing-market-103336
2. https://www.fuqua.duke.edu/duke-fuqua-insights/algorithms-behind-pricing-your-ride
3. GitHub - vasupradha2003/Dynamic-Pricing-for-Ride-Sharing-Services: This project implements a dynamic pricing model for ride-sharing services using machine learning, specifically Gradient Boosting, to predict the price of a ride based on various factors such as distance, demand, time of day, and weather conditions.
4. IJSRET_V10_issue6_542.pdf
5. Elastic Net Regression in R Programming | GeeksforGeeks
6. Regularization in R Tutorial: Ridge, Lasso & Elastic Net Regression | DataCamp
7. Decision Trees, Random Forests, and Overfitting – Machine Learning for Biologists
8. https://www.datacamp.com/tutorial/multiple-linear-regression-r-tutorial
9. XGBoost Documentation — xgboost 3.0.0 documentation
10. https://www.youtube.com/watch?v=33fGfuleXw0
11. https://www.youtube.com/watch?v=OtD8wVaFm6E
12. Chat GPT

APENDIX

1. Link for dataset : Ride Dynamic Pricing
2. R codes: https://drive.google.com/drive/folders/19jRQ_XowZl1esgLyeib1j2vGZqz1PwwZ
3. Python Colab links:
   https://colab.research.google.com/drive/1D5988akOANldtK_8KTVaK0DEyrl2uWCs
   Ride Prediction-Random forest.ipynb - Colab