STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **13**, April 2025, pp 1732–1751.
Published online in International Academic Press (www.IAPress.org)

Dynamic Pricing and Service Quality in Ride-Sharing: A Statistical Analysis

Daniel Sanin-Villa[1,*], Cristian M. Hernández[2], Vanessa Botero-Gómez[2]

[1] *Universidad EAFIT, Medellín, 050022, Antioquia, Colombia*
[2] *Instituto Tecnológico Metropolitano (ITM), Medellín, Colombia*

Abstract This study presents a statistical analysis of factors influencing dynamic pricing and service quality in ride-sharing. Leveraging historical data, regression models, including simple and multiple linear regressions, as well as logistic regression, were employed to examine the relationships between trip duration, passenger count, driver availability, and customer loyalty on ride costs and service ratings. Results reveal that trip duration significantly predicts ride costs, with an estimated increase of \$3.53 per additional minute ($R^2 = 0.860$). Additionally, logistic regression showed that urban rides and silver loyalty status significantly increase the likelihood of an ”excellent” rating, with odds ratios of 1.21 and 1.54, respectively. These findings provide actionable insights for enhancing dynamic pricing strategies and optimizing service quality in a competitive market.

Keywords dynamic pricing, regression analysis, ride-sharing, cost optimization, customer satisfaction

DOI: 10.19139/soic-2310-5070-2325

1. Introduction

Dynamic pricing has emerged as a fundamental strategy for optimizing revenues in various industries, including the electrical sector [1, 2] and transportation and ride-sharing services. It employs algorithms to adjust fares in response to real-time demand, trip duration, and market conditions. Although numerous studies have explored fare adjustment techniques based on supply and demand, few have thoroughly examined how different factors, such as trip duration, number of passengers, and driver availability, impact the overall cost and quality of service ratings. These aspects are important to refining pricing models to enhance profitability and customer satisfaction [3, 4, 5].

Previous research has demonstrated the effectiveness of regression techniques in predicting pricing trends and customer behavior in shared mobility services [6, 7]. Linear models have been employed to forecast trip costs based on duration and distance, while logistic models have been used to predict the likelihood of high customer ratings based on service attributes [8]. However, existing pricing algorithms often oversimplify complex relationships by focusing primarily on supply-demand metrics, potentially neglecting other influential variables such as customer demographics or driver behavior. This study investigates whether integrating multiple factors—such as trip duration, number of passengers, and driver availability—can significantly enhance the accuracy of dynamic pricing models in predicting both trip costs and service ratings.

The present study seeks to fill this gap by analyzing historical ride data to identify the most significant predictors of trip cost and service ratings. Using robust statistical methods, including regression techniques, ANOVA [9], and statistical tests [10, 11], this research aims to isolate the effects of individual variables and provide insights that can inform the development of data-driven pricing policies and service improvements. By understanding the relationships between key variables, ride-sharing companies can better implement pricing strategies that adapt to market dynamics while enhancing customer experiences.

[*] Correspondence to: Daniel Sanin-Villa (Email: dsaninv2@eafit.edu.co), Universidad EAFIT, Medellín, Colombia.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

The ride-sharing industry has undergone substantial transformation due to advancements in technology and changes in consumer preferences [12, 13]. To maintain a competitive advantage in this dynamic market, companies must deeply understand customer behavior and optimize their service performance [14]. This research analyzes the factors influencing customer satisfaction and service costs within a ride-sharing framework, aiming to uncover significant patterns and relationships that could guide the development of pricing strategies that balance profitability with customer satisfaction.

The dataset utilized in this study is sourced from a public repository titled ”Dynamic Pricing Dataset” available on Kaggle (https://www.kaggle.com/datasets/arashnic/dynamic-pricing-dataset), released under the CC0: Public Domain license. This dataset comprises historical ride data from a ride-sharing company, including variables such as trip duration, distance, and ratings, which provides a comprehensive foundation for modeling the effects of diverse factors on pricing and service quality. Currently, the company's pricing models tend to rely solely on ride duration to determine fares.

The main contributions of this study are the identification of key factors influencing ride-sharing costs and service quality through the application of statistical modeling, including linear and logistic regression analyses. The research provides evidence-based insights into how trip duration, driver availability, and loyalty status impact pricing strategies and customer satisfaction. Additionally, the study introduces practical recommendations for optimizing dynamic pricing models to better align with market conditions.

The remainder of the manuscript is organized as follows: Section 2 presents the theoretical framework and statistical methods used, Section 3 details the application of regression models and results, and Section 4 concludes with a summary of key insights and suggestions for future research.

2. Methodology

This study employs statistical modeling techniques to analyze the factors influencing ride-sharing trip costs and service quality ratings. The methodological approach comprises three primary analyses: simple linear regression, multiple linear regression, and logistic regression, each designed to investigate different aspects of the data. All analyses were conducted using Python with custom statistical development, and the results were interpreted to offer insights into dynamic pricing strategies and customer satisfaction.

The dataset includes various features, such as the number of riders, drivers, location categories, customer loyalty statuses, average ratings, booking times, vehicle types, expected ride durations, and historical ride costs.

2.1. Justification of Model Selection

The choice to employ simple linear regression, multiple linear regression, and logistic regression was driven by the need to prioritize model interpretability and transparency in dynamic pricing and service quality analysis. Regression models are well-suited for this application because they provide explicit relationships between variables, which is particularly valuable for stakeholders in the ride-sharing industry who require clear, actionable insights for decision-making. Additionally, these models facilitate the interpretation of coefficients, allowing us to quantify the effect of individual factors on pricing and customer satisfaction. Moreover, the dataset used in this study does not present significant complexity that would necessitate more sophisticated machine learning techniques, such as Random Forest or Gradient Boosting. While these methods could improve predictive accuracy, they often lack the transparency needed to understand the causal relationships between the factors studied, a key research objective. Machine learning models, though powerful, can function as ”black boxes,” making it challenging to derive practical insights that can be translated into effective policies and business strategies.

2.2. Simple Linear Regression

Two simple linear regression models were implemented to explore the relationship between individual variables and the historical cost of trips [15]. In the first model, the number of past trips (X_1) was used as the independent variable, while the second model employed the expected trip duration (X_2) as the independent variable. The response variable for both models was the historical trip cost (Y).

The simple linear regression model is defined as:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{1}$$

Where β_0 is the intercept, β_1 is the slope coefficient of the independent variable X, and ε represents the error term, assumed to be normally distributed with mean zero and constant variance.

The regression equation was calculated for each model, and the parameters β_0 and β_1 were interpreted in the study context. The significance of these parameters was assessed using the t-statistic, where the null hypothesis $H_0 : \beta_i = 0$ was tested against the alternative hypothesis $H_a : \beta_i \neq 0$. A p-value less than the significance level ($\alpha = 0.05$) indicated that the parameter was statistically significant.

Scatter plots with the fitted regression lines were generated to visualize the relationship between the independent variables and the trip cost. Residuals and fitted values were obtained, and residual plots were examined to assess the assumptions of residuals' linearity, homoscedasticity, and normality. The residual analysis provided insights into the adequacy of the models and the presence of any patterns or anomalies.

2.3. Multiple Linear Regression

A multiple linear regression model was developed to investigate the combined effect of multiple factors on the historical trip cost [16]. The response variable remained the historical trip cost (Y), and the independent variables included the number of passengers (X_1), number of drivers (X_2), number of past trips (X_3), and expected trip duration (X_4).

The multiple linear regression model is expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon, \tag{2}$$

where β_0 is the intercept, β_i for $i = 1, 2, 3, 4$ are the coefficients associated with each independent variable, and ε is the error term.

The regression equation was computed, and the beta coefficients were interpreted to understand the effect of each independent variable on the trip cost. The significance of each variable was determined based on the p-values obtained from the t-tests, where variables with p-values less than $\alpha = 0.05$ were considered significant predictors.

Model evaluation metrics were calculated to assess the performance of the multiple regression model. These included the Residual Standard Error (RSE), Root Mean Square Error (RMSE), R-squared (R^2), and Adjusted R-squared ($\bar{R}^2$). The RSE and RMSE provided measures of the average deviation of the observed values from the fitted values. At the same time, the R^2 and $\bar{R}^2$ indicated the proportion of variance in the trip cost explained by the model.

2.4. Logistic Regression

A logistic regression analysis was conducted to model the probability of a trip receiving an excellent service rating [17]. The response variable was a binary indicator of service quality (Y), where $Y = 1$ represented an excellent service (average rating of 4 or above), and $Y = 0$ represented a regular service (average rating below 4). The independent variables included location (X_1), customer loyalty status (X_2), booking time (X_3), and vehicle type (X_4).

The logistic regression model is formulated as:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4, \tag{3}$$

where $p = P(Y = 1|X_1, X_2, X_3, X_4)$ is the probability of an excellent service rating, and β_i are the model coefficients.

The model was trained using 70% of the data as the training set, and the remaining 30% was reserved for testing. The odds ratios were calculated by exponentiating the coefficients ($\mathrm{OR}_i = e^{\beta_i}$), providing an interpretation of the effect of each independent variable on the odds of receiving an excellent rating.

Model validation involved interpreting the p-values of the coefficients to determine the significance of each predictor. Variables with p-values less than $\alpha = 0.05$ were considered statistically significant. A confusion matrix was constructed using the test data to evaluate the model's classification performance, and metrics such as accuracy, precision, recall, and F1-score were calculated.

2.5. Addressing Class Imbalance with SMOTE

In the logistic regression analysis, we encountered a significant imbalance in the service quality variable, where the majority of trips were rated as ”excellent” ($Y = 1$) compared to those rated as ”regular” ($Y = 0$). This class imbalance can lead to biased model training, as the algorithm may become overly tuned to the majority class, resulting in poor predictive performance for the minority class. To mitigate this issue, we employed the Synthetic Minority Over-sampling Technique (SMOTE) [18], which generates synthetic samples of the minority class to create a more balanced training dataset.

The decision to use SMOTE was based on its ability to generate synthetic samples of the minority class, thereby reducing bias in model learning while preserving the characteristics of the feature space.

Although the Area Under the ROC Curve (AUC) remained at 0.5 after applying SMOTE, other evaluation metrics such as precision, recall, and F1-score showed improvements. Specifically, precision increased from 0.6767 to 0.6918, and the recall improved in identifying the minority class, which had previously been classified poorly. These improvements, although modest, were critical in achieving a more balanced classification performance.

SMOTE synthesizes new minority class examples, connecting existing minority instances and their nearest neighbors in the feature space. Mathematically, for each minority class sample $\mathbf{x}i$, SMOTE selects one of its k nearest neighbors $\mathbf{x}i^{\text{NN}}$ and creates a synthetic sample $\mathbf{x}_{\text{new}}$ using the interpolation formula:

$$\mathbf{x}_{\text{new}} = \mathbf{x}i + \gamma \times (\mathbf{x}i^{\text{NN}} - \mathbf{x}_i), \tag{4}$$

where γ is a random number in the interval $[0, 1]$. This process effectively generates new samples that are similar but not identical to existing minority instances, enriching the diversity of the minority class without simply duplicating existing data.

After balancing the training dataset with SMOTE, we retrained the logistic regression model to improve its ability to learn from both classes. The logistic regression model estimates the probability p of an instance being classified as ”excellent” using the logistic function:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4)}}, \tag{5}$$

where β_0 is the intercept, β_i are the coefficients for the predictor variables X_i, and e is the base of the natural logarithm. The coefficients β_i are estimated using maximum likelihood estimation to find the values that maximize the likelihood of observing the given data.

To evaluate the logistic regression model's performance after applying SMOTE, we employed several metrics that comprehensively assess the model's discriminative ability and effectiveness in handling class imbalance. One of the primary tools used was the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings.

Mathematically, the **True Positive Rate**, also known as sensitivity or recall, is defined as:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{6}$$

Where TP represents the number of true positives, which are the positive instances correctly classified by the model, and FN denotes the number of false negatives, which are the positive instances incorrectly classified as negative.

The False Positive Rate is given by:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \tag{7}$$

Where FP represents the number of false positives, indicating negative instances incorrectly classified as positive, and TN denotes the number of true negatives, representing the negative instances correctly classified by the model.

By varying the logistic regression model's classification threshold τ, where instances with a predicted probability $p \geq \tau$ are classified as positive, we obtain different TPR and FPR pairs, plotted to form the ROC curve. This curve illustrates the model's ability to distinguish between the two classes across all possible threshold values.

The Area Under the ROC Curve (AUC) is computed as:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d\text{FPR}, \tag{8}$$

where TPR(FPR) represents the TPR as a function of FPR. The AUC provides a single scalar value summarizing the model's overall discriminative performance. An AUC value closer to 1 indicates excellent discrimination, while an AUC of 0.5 suggests no discriminative ability, equivalent to random guessing.

In addition to the ROC curve, we analyzed the Precision-Recall (PR) curve, which is particularly informative for imbalanced datasets where the positive class is rare. The Precision, also known as positive predictive value, is defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{9}$$

where TP is the number of true positives and FP is the number of false positives. Precision measures the proportion of instances predicted as positive that are actually positive. The Recall, equivalent to the TPR, quantifies the proportion of actual positive instances that are correctly identified by the model.

The PR curve plots Precision against Recall for different threshold values, highlighting the trade-off between achieving high recall (identifying all positive instances) and maintaining high precision (ensuring that positive predictions are correct). In imbalanced datasets, where the negative class dominates, the PR curve provides a more sensitive performance measure than the ROC curve.

By examining these curves, we gain insights into the model's performance concerning false positives and false negatives. A model with a PR curve closer to the top-right corner indicates both high precision and high recall, which is desirable for accurate classification.

Furthermore, we assessed the importance of each predictor variable by examining the estimated coefficients β_i from the logistic regression model. In logistic regression, the relationship between the predictors and the log odds of the outcome is given by:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4, \tag{10}$$

where p represents the probability of a positive outcome, such as an "excellent" rating, and β_0 is the intercept term. The coefficients β_i correspond to the predictor variables X_i, which include factors such as location, customer loyalty status, booking time, and vehicle type. The exponentiated coefficients, e^{β_i}, yield the Odds Ratios, which quantify the change in odds of the outcome occurring for a one-unit increase in the predictor X_i, holding all other variables constant. For instance, if $e^{\beta_2} = 2.14$ for customer loyalty status, loyal customers have odds of giving an "excellent" rating that is 2.14 times higher than those of non-loyal customers.

Finally, after applying SMOTE, we analyzed the Distribution of Predicted Probabilities for both classes. The logistic regression model outputs predicted probabilities p for each instance. By plotting the distribution of these probabilities for the positive (excellent service) and negative (regular service) classes, we can assess how well the model separates the two classes. A clear separation between the distributions indicates that the model assigns higher probabilities to positive instances and lower probabilities to negative instances, thereby enhancing classification accuracy.

The application of SMOTE effectively balanced the classes, allowing the model to learn adequately from both positive and negative instances. This improvement is reflected in the evaluation metrics, where increases in AUC, precision, and recall demonstrate the model's enhanced ability to correctly classify instances from both classes. By addressing class imbalance, we mitigated the risk of the model being biased toward the majority class, resulting in a more robust and reliable predictive model for service quality ratings in ride-sharing.

The importance of each predictor variable was assessed by examining the magnitude and direction of the estimated coefficients β_i. In logistic regression, the coefficients represent the change in the log-odds of the outcome for a one-unit increase in the predictor variable, holding all other variables constant. Specifically, the odds ratio is calculated as $\text{OR}_i = e^{\beta_i}$, indicating how the odds of the outcome change with a one-unit increase in X_i.

Finally, we assessed the distribution of predicted probabilities for both classes to understand how well the model differentiates between ”excellent” and ”regular” service ratings. A clear separation between the distributions suggests that the model assigns higher probabilities to instances of the majority class and lower probabilities to the minority class, reflecting effective classification.

3. Results and Discussion

3.1. *Simple Linear Regression*

The results of the simple linear regression models provide insights into the relationship between the historical cost of rides and two different predictor variables: the number of past rides and the expected ride duration. Each model was evaluated based on the significance of its coefficients, the goodness of fit, and the statistical properties of the residuals.

3.1.1. Model 1: Historical Cost of Ride vs. Number of Past Rides The first linear regression model examined the association between the historical cost of the ride (Y) and the number of past rides (X_1). The regression equation for this model is:

$$Y = 361.0482 + 0.2289 \times X_1, \tag{11}$$

where the intercept, 361.0482, represents the estimated historical cost when the number of past rides is zero, suggesting that in the absence of any past ride history, the baseline cost of a ride is approximately \$361.05. The coefficient for the number of past rides is 0.2289, which implies that for each additional past ride, the historical cost increases by approximately \$0.23.

The scatter plot of historical cost versus the number of past rides (Figure 1) shows a cloud of points with no apparent linear trend, and the fitted regression line is almost horizontal, indicating a very weak relationship between the two variables. This observation is consistent with the R^2 value of 0.001, which implies that only 0.1% of the variance in the historical cost is explained by the number of past rides.

The residual plot for this model (Figure 2) shows a random scatter of residuals around zero, indicating no clear pattern. This suggests that the linear model is not capturing any meaningful relationship between the number of past rides and the historical cost, which aligns with the non-significant p-value (0.257) for the coefficient.

3.1.2. Model 2: Historical Cost of Ride vs. Expected Ride Duration The second linear regression model explored the relationship between the historical cost of the ride (Y) and the expected ride duration (X_2). The regression equation for this model is:

$$Y = 20.8668 + 3.5309 \times X_2, \tag{12}$$

where the intercept, 20.8668, represents the estimated historical cost when the expected ride duration is zero, indicating a baseline cost of approximately \$20.87. The coefficient for the expected ride duration is 3.5309, which implies that for each additional minute of ride duration, the historical cost increases by \$3.53.

The scatter plot of historical cost versus expected ride duration (Figure 3) reveals a strong positive linear relationship between the two variables. The fitted regression line indicates a clear upward trend, suggesting that as ride duration increases, so does the historical cost. This observation is supported by the high R^2 value of 0.860, indicating that 86% of the variance in the historical cost is explained by the expected ride duration.

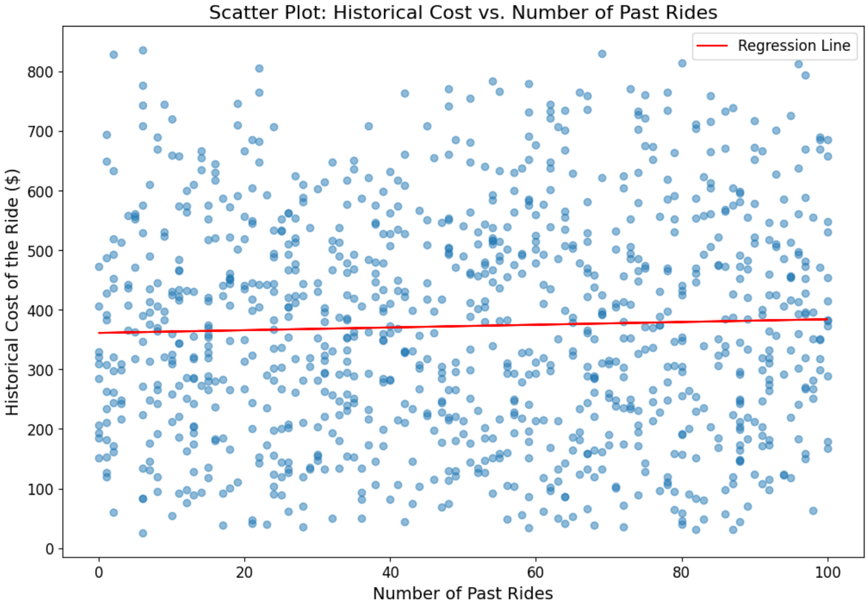

Figure 1. Scatter Plot: Historical Cost vs. Number of Past Rides with Regression Line.

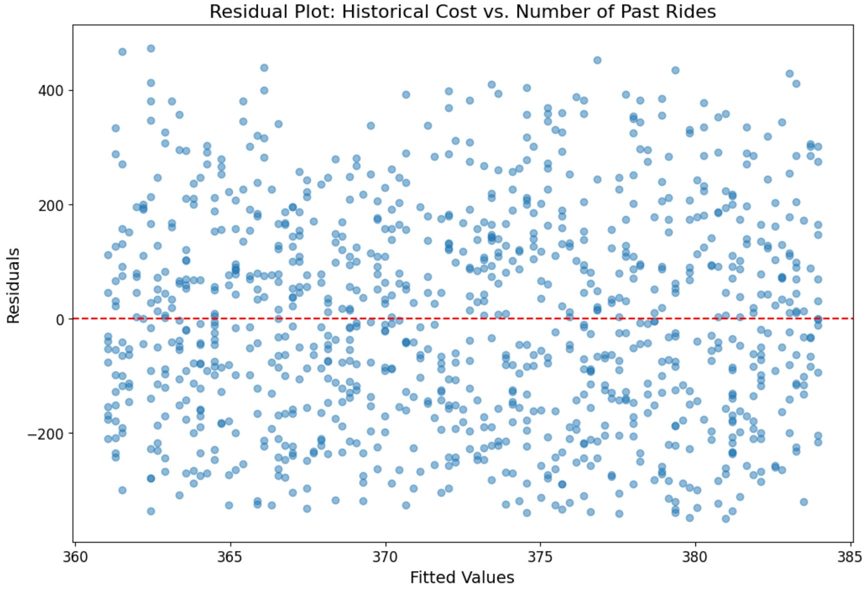

Figure 2. Residual Plot: Historical Cost vs. Number of Past Rides.

The residual plot for this model (Figure 4) shows a slight funnel shape, where the spread of residuals increases with the fitted values. This pattern suggests potential heteroscedasticity in the data, meaning the variability of the residuals is not constant across all levels of the fitted values. Nonetheless, the residuals are mostly centered around zero, indicating that the model captures the general trend of the data.

Heteroscedasticity, characterized by non-constant variance of residuals, can lead to inefficient estimates and biased standard errors, potentially compromising the validity of hypothesis tests. In this study, we opted not to correct for heteroscedasticity, given the exploratory nature of the analysis. However, we acknowledge the

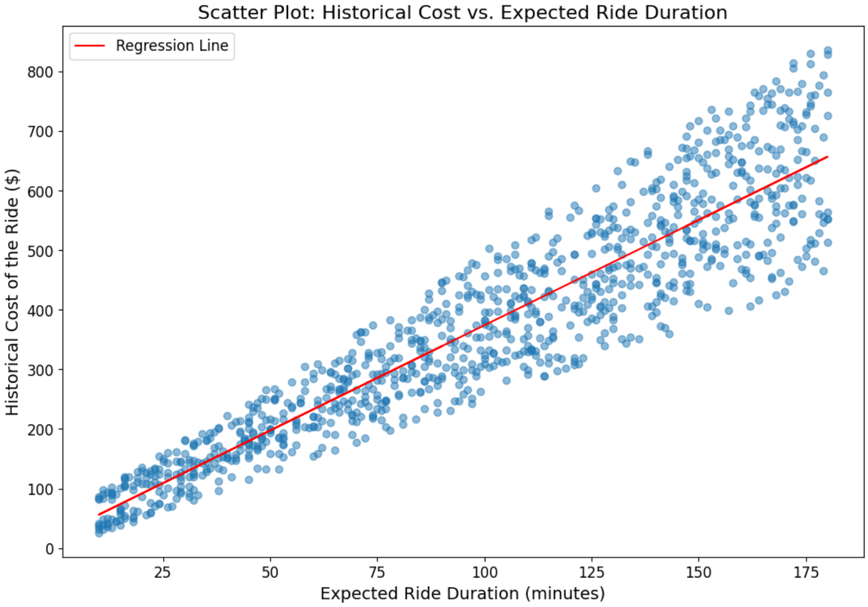

Figure 3. Scatter Plot: Historical Cost vs. Expected Ride Duration with Regression Line.

importance of addressing this issue in future research. Techniques such as the White's test could be used to assess the presence of heteroscedasticity formally [19]. A transformation like Box-Cox could stabilize variance and improve model reliability if heteroscedasticity is detected. Additionally, robust standard errors could be employed as an alternative approach to mitigate the effects of heteroscedasticity, thereby enhancing the robustness of inference drawn from the regression coefficients.

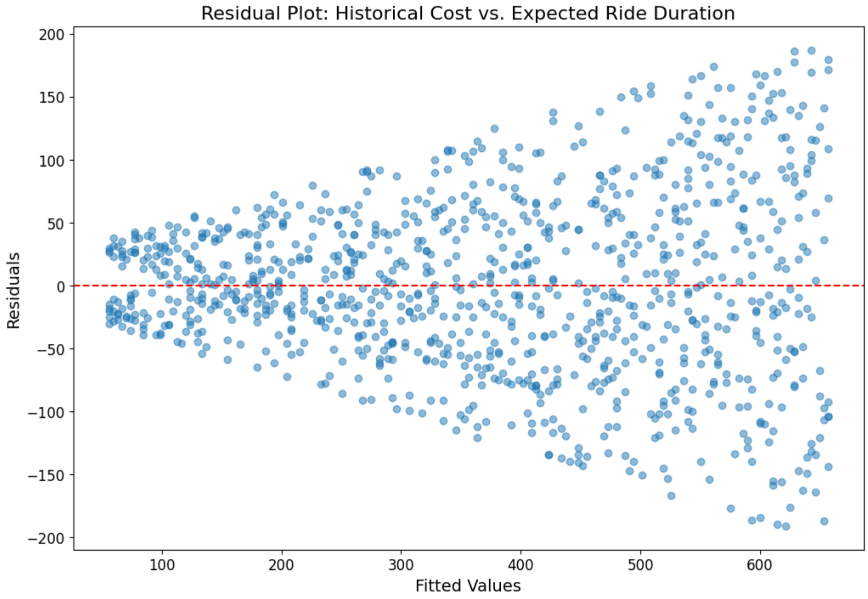

Figure 4. Residual Plot: Historical Cost vs. Expected Ride Duration.

3.1.3. Comparison and Implications The comparison between the two models reveals that while the number of past rides does not significantly impact the historical cost, the expected ride duration is a highly significant predictor. The first model, with an R^2 value of 0.001, indicates that the number of past rides is not informative for predicting ride costs. In contrast, the second model, with an R^2 value of 0.860, highlights that ride duration explains a substantial portion of the variability in costs, underscoring its importance in pricing strategies.

The results suggest that ride-sharing companies should prioritize time-based factors, such as the duration of the ride, when developing dynamic pricing models. This focus could improve the accuracy of fare estimates and better reflect operational costs. The non-significance of the number of past rides implies that historical ride frequency does not influence current pricing, which may indicate that pricing strategies should rely more on real-time factors rather than historical usage.

The findings support the use of expected ride duration as a critical variable in predicting the cost of ride-sharing services. In contrast, other factors, such as the number of past rides, may not provide substantial predictive value.

3.2. Multiple Linear Regression

The multiple linear regression model was developed to explore the relationship between the historical cost of the ride (Y) and several independent variables: the number of riders (X_1), number of drivers (X_2), number of past rides (X_3), and expected ride duration (X_4). The analysis aimed to understand the combined effects of these factors on the cost of ride-sharing services. The regression equation for this model is given by:

$$Y = 10.4521 - 0.0573 \times X_1 + 0.4325 \times X_2 + 0.0373 \times X_3 + 3.5339 \times X_4, \tag{13}$$

where the intercept, 10.4521, represents the estimated historical cost when all predictor variables are zero, suggesting a baseline cost of approximately \$10.45. The coefficient for the number of riders, which is -0.0573, indicates a slight negative relationship, suggesting that for each additional rider, the cost decreases by about \$0.06. However, this effect is not statistically significant, as indicated by a p-value of 0.631, meaning the number of riders does not have a meaningful impact on the historical cost.

Conversely, the coefficient for the number of drivers is 0.4325, showing that for each additional driver, the historical cost increases by approximately \$0.43. This predictor is statistically significant, with a p-value of 0.004, indicating that driver availability does influence the ride cost. The number of past rides also has a small positive coefficient of 0.0373, suggesting a minor increase in cost per additional past ride, though this effect is not significant either, with a p-value of 0.620.

The coefficient for the expected ride duration is 3.5339, and it is highly significant, with a p-value close to zero. This indicates that for each additional minute of ride duration, the historical cost increases by \$3.53, confirming that ride duration is a crucial factor in determining the cost.

3.2.1. Statistical Significance and Model Evaluation The overall significance of the model was evaluated using the F-statistic, which yielded a value of 1554 with a p-value close to zero, indicating that at least one of the predictors is significantly related to the historical cost. The goodness of fit was assessed using the R^2 value, which was 0.862, and the adjusted R^2 value, which was 0.861. This means that approximately 86% of the variance in the historical cost is explained by the predictors, demonstrating that the model fits the data well.

In terms of error measures, the Residual Standard Error (RSE) was 69.67, while the Root Mean Squared Error (RMSE) was 69.49. These values indicate the average deviation of the observed costs from the predicted values, with smaller values signifying better predictive accuracy. Since the RSE and RMSE are close to each other, it suggests consistent predictive errors across the dataset.

3.2.2. Interpretation of Coefficients The regression results show that expected ride duration is the most significant predictor, with a coefficient of 3.5339 and a very high t-statistic of 78.76. This finding highlights that ride duration has a substantial impact on ride cost. On the other hand, the significant coefficient for the number of drivers indicates that driver availability is also a factor influencing pricing, possibly due to operational considerations such as demand fluctuations and supply levels.

In contrast, the number of riders and the number of past rides did not exhibit significant effects on the historical cost, as reflected by their high p-values. This lack of significance implies that these variables do not contribute much to explaining the variability in ride costs in the current dataset. Therefore, their inclusion in a dynamic pricing model may not be necessary.

3.2.3. Residual Analysis To assess the model's adequacy, residual analysis was performed by plotting the residuals against the fitted values (Figure 5). The residual plot showed no discernible patterns, suggesting that the linear model adequately captures the relationship between the predictors and the historical cost. The Durbin-Watson statistic was approximately 1.951, indicating no significant autocorrelation in the residuals.

Additionally, the histogram of residuals (Figure 6) appeared to follow a normal distribution, which was confirmed by skewness and kurtosis values close to zero, as well as the results of the Jarque-Bera test, which yielded a p-value of 0.935. This supports the assumption that the residuals are approximately normally distributed, further validating the model.

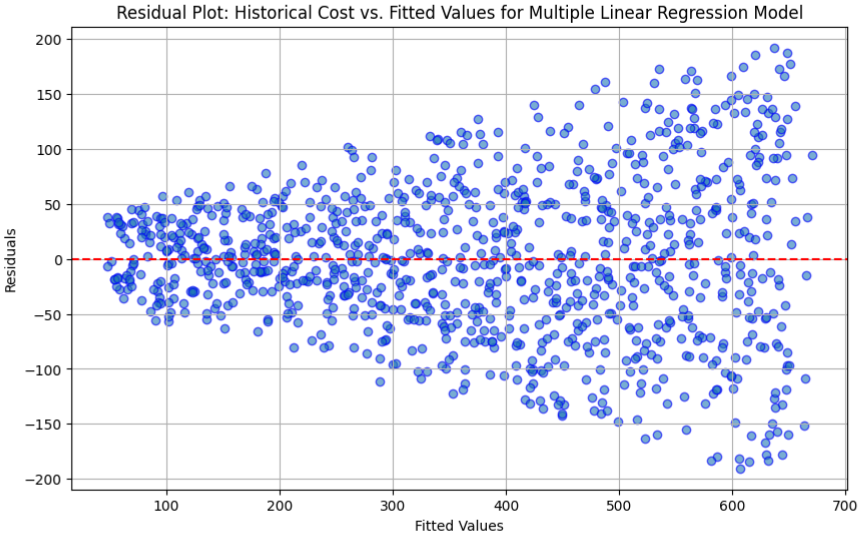

Figure 5. Residual Plot: Historical Cost vs. Fitted Values for Multiple Linear Regression Model.

3.2.4. Implications and Recommendations The analysis indicates that expected ride duration and the number of drivers are significant determinants of ride cost, suggesting that pricing strategies should focus on these factors. Incorporating ride duration as a primary variable in dynamic pricing models could enhance the accuracy of fare estimates and better reflect the operational costs associated with longer trips. The significance of driver availability further suggests that adjusting prices based on the number of available drivers could help optimize supply-demand balance.

The non-significance of the number of riders and past rides implies that these variables do not substantially affect ride costs, which may suggest that historical ride frequency and the number of passengers should not be central components of pricing models. Future research could explore additional variables, such as traffic conditions or surge pricing, to further refine the model and account for other factors that might influence ride costs.

3.3. Logistic Regression

The logistic regression analysis was conducted to examine the likelihood of a ride being rated as ”excellent” based on various predictors, including location category, customer loyalty status, time of booking, and vehicle type. The response variable was defined as a binary variable representing service quality, where a value of 1 indicates an ”excellent” rating (average rating of 4 or higher), and 0 represents a ”regular” rating (below 4).

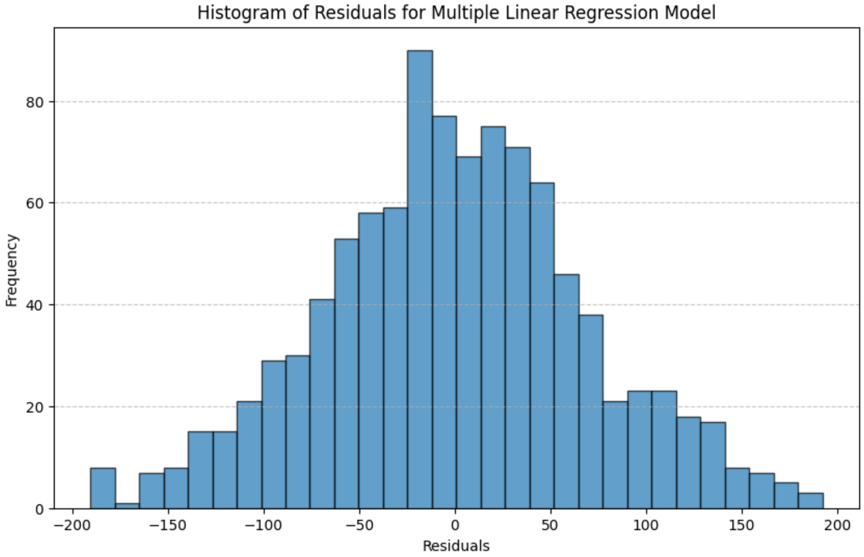

Figure 6. Histogram of Residuals for Multiple Linear Regression Model.

3.3.1. Model Implementation and Interpretation The logistic regression model was trained using 70% of the data as the training set, while the remaining 30% served as the test set. Dummy variables were created for the categorical predictors, and the model was fitted using the training data. The results of the logistic regression, presented in Table 1, include the estimated coefficients, standard errors, z-values, and p-values for each predictor, as well as the odds ratios.

The odds ratio provides a measure of the effect of each predictor on the likelihood of an ”excellent” rating. An odds ratio greater than 1 indicates a higher likelihood of receiving an ”excellent” rating when the predictor is present, while an odds ratio less than 1 indicates a lower likelihood. For example, the odds ratio for ”Location Category: Urban” was 1.207513, suggesting that rides in urban areas are about 1.21 times more likely to receive an ”excellent” rating compared to rural locations. The significance of this effect is supported by a p-value of 0.018, indicating that the location category plays a statistically significant role in predicting service quality.

3.3.2. Significant Predictors and Interpretation Among the predictors, several variables showed statistically significant effects on service quality. ”Location Category: Urban” had a coefficient of 0.4299 with a p-value of 0.018, indicating that rides occurring in urban areas are more likely to be rated as ”excellent” compared to those in rural areas. Similarly, ”Customer Loyalty Status: Silver” had a coefficient of 0.4302 and a p-value of 0.017, suggesting that customers with silver loyalty status are more likely to provide an ”excellent” rating compared to those with regular status.

The ”Time of Booking: Morning” was also a significant predictor, with a coefficient of 0.5381 and a p-value of 0.009, indicating that rides booked in the morning are more likely to receive an ”excellent” rating than those booked in the afternoon (reference category). The odds ratio of 1.191041 suggests that morning bookings increase the likelihood of an ”excellent” rating by approximately 19%.

Other predictors, such as ”Location Category: Suburban,” ”Customer Loyalty Status: Regular,” ”Time of Booking: Evening,” ”Time of Booking: Night,” and ”Vehicle Type: Premium,” did not show statistically significant effects, as their p-values exceeded the commonly used threshold of 0.05. These results suggest that these factors do not substantially influence the likelihood of an ”excellent” rating in the current dataset.

3.3.3. Practical Implications of Significant Predictors and the Role of Non-significant Variables The significant predictors identified by the logistic regression model—specifically, ”Location Category: Urban,” ”Customer

Loyalty Status: Silver", and "Time of Booking: Morning" offer valuable insights for operational and strategic decision-making. The positive association between urban locations and the likelihood of an "excellent" rating suggests that rides originating in city centers may inherently benefit from more reliable infrastructure, greater availability of drivers or improved response times. In practice, ride-sharing companies could consider allocating additional resources to urban areas, such as placing more experienced drivers or better vehicles, to capitalize on this elevated propensity for positive feedback.

Similarly, the finding that customers with silver loyalty status are more inclined to rate their experience as "excellent" underlines the importance of structured loyalty programs. Enhancing the customer experience for individuals who have reached certain loyalty tiers—such as by offering preferential access to premium vehicle types or targeted promotions—may reinforce positive perceptions and further improve satisfaction metrics. This targeted approach can strengthen customer retention and reinforce brand loyalty over time.

The observed significance of morning bookings suggests that the start-of-day environment may foster conditions conducive to superior service experiences. Morning rides may be less affected by congestion or driver fatigue, enabling smoother trips and more timely arrivals. Ride-sharing platforms could leverage this insight by introducing morning-specific incentives, calibrating dynamic pricing to reflect potentially higher service quality, or ensuring an adequate supply of well-rated drivers during peak morning hours. In contrast, several variables did not achieve statistical significance, including suburban locations, regular loyalty status, evening or night bookings, and premium vehicle type. While excluding these non-significant predictors may simplify the model and slightly enhance predictive performance, doing so risks overlooking potentially relevant aspects of the service context. For instance, even if suburban rides or premium vehicle usage currently do not show a strong direct effect on ratings, these factors might gain importance as the market evolves, the data volume increases, or additional variables are introduced in the future. Retaining these predictors maintains a more holistic representation of the operational environment and keeps open the possibility that their effects may emerge under different conditions or with more comprehensive datasets.

Balancing model simplicity with theoretical thoroughness is a crucial consideration. On one hand, removing non-significant variables can enhance the model's interpretability by emphasizing the most impactful predictors. On the other hand, keeping these variables may capture a broader range of underlying influences, provide resilience against changing market dynamics, and lay the groundwork for more detailed analyses if richer data become available in the future. Ultimately, the choice to include or exclude non-significant variables should align with the study's objectives—whether focusing on immediate managerial recommendations based on clear and influential factors or maintaining a flexible analytical framework that can adapt as new information arises.

3.3.4. Model Fit and Diagnostics The logistic regression model was fitted using the maximum likelihood estimation method. The model's log-likelihood value was -434.37, with a null log-likelihood of -433.37. The Pseudo R-squared value was -0.0023, indicating a poor fit of the model to the data. Additionally, the likelihood ratio test yielded a p-value of 1.000, suggesting that the model does not significantly improve the fit compared to a model with no predictors.

The results indicate that although some individual predictors are statistically significant, the overall model does not explain a substantial portion of the variability in service quality ratings. This limitation could be due to the nature of the predictors or the presence of unmeasured factors influencing customer ratings. Future research could consider incorporating additional variables, such as trip distance, traffic conditions, or driver characteristics, to improve the model's predictive power.

3.3.5. Implications and Recommendations The logistic regression analysis suggests that service quality ratings are influenced by factors such as the location of the ride, customer loyalty status, and the time of day when the booking is made. Specifically, rides in urban areas, customers with silver loyalty status, and morning bookings are more likely to receive higher ratings. This insight could inform strategies to enhance customer satisfaction, such as targeted marketing campaigns or adjusting service features based on location and time.

Table 1. Logistic Regression Results for Predicting Service Quality

Variable	Coefficient	Std. Error	z-value	P-value
Location Category: Suburban	0.2398	0.186	1.293	0.196
Location Category: Urban	0.4299	0.182	2.360	0.018
Customer Loyalty Status: Regular	0.2356	0.185	1.276	0.202
Customer Loyalty Status: Silver	0.4302	0.181	2.383	0.017
Time of Booking: Evening	0.0689	0.205	0.336	0.737
Time of Booking: Morning	0.5381	0.206	2.610	0.009
Time of Booking: Night	0.1910	0.192	0.996	0.319
Vehicle Type: Premium	0.1368	0.153	0.894	0.371

3.3.6. Confusion Matrix Analysis The confusion matrix provides a summary of the classification results for the logistic regression model, indicating the number of correct and incorrect predictions for each class. In this analysis, the model was evaluated on the test set using a probability threshold of 0.5 to classify the service quality as either ”regular” (0) or ”excellent” (1). The confusion matrix is presented in Figure 7, where the rows represent the actual class labels, and the columns represent the predicted class labels.

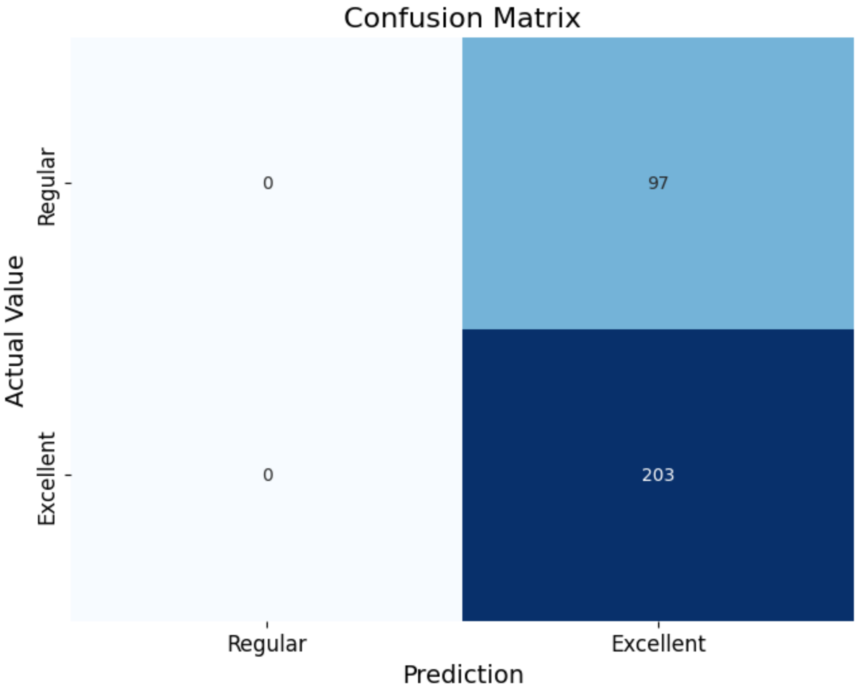

Figure 7. Confusion Matrix for the Logistic Regression Model.

The confusion matrix reveals that the model classified all cases as ”excellent,” resulting in no true negatives or false negatives. Specifically, the matrix shows 203 true positives, where the actual service quality was ”excellent,” and the model correctly predicted it as such. Additionally, there were 97 false positives, where the actual service quality was ”regular,” but the model predicted it as ”excellent.”

To further evaluate the model’s performance, several metrics were calculated based on the confusion matrix: accuracy, precision, recall, and the F1-score. These metrics provide insights into different aspects of the model’s predictive performance:

- Accuracy: The model achieved an accuracy of 0.6767, indicating that approximately 68% of the predictions were correct. This metric represents the proportion of all correctly classified instances out of the total number of predictions.
- Precision: The precision score was 0.6767, meaning that 67.67% of the instances classified as ”excellent” were actually ”excellent.” Precision measures the accuracy of positive predictions.
- Recall: The recall was 1.0000, indicating that the model identified all the actual ”excellent” cases correctly. This metric measures the model’s ability to capture all relevant instances of the positive class.
- F1-Score: The F1-score was 0.8072, providing a balance between precision and recall. This metric is useful when the goal is to seek a trade-off between precision and recall, especially in cases of imbalanced datasets.

The results suggest that while the model performs well in identifying ”excellent” cases (with perfect recall), it struggles to differentiate ”regular” cases, as indicated by the large number of false positives and the absence of true negatives. This pattern suggests that the model is biased towards predicting the positive class (”excellent”), which could be due to class imbalance in the dataset.

3.3.7. Confusion Matrix Analysis with SMOTE After applying the Synthetic Minority Over-sampling Technique to balance the dataset, the logistic regression model was re-evaluated using the test set. The SMOTE method generated synthetic examples for the minority class (”regular”) to create a more balanced training set, thereby improving the model’s ability to generalize across both classes. Figure 8 shows the confusion matrix after applying SMOTE.

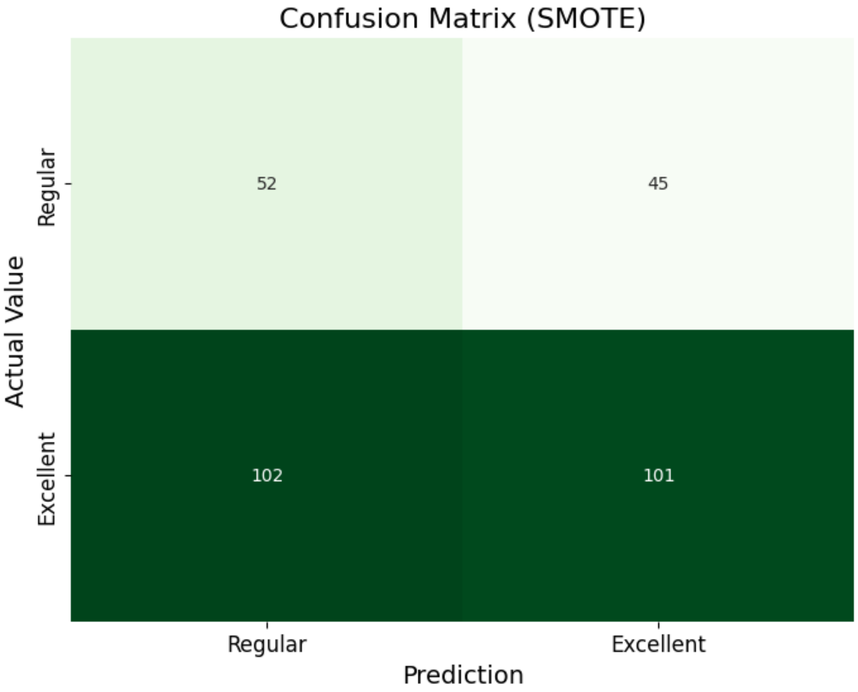

Figure 8. Confusion Matrix for the Logistic Regression Model with SMOTE.

The confusion matrix indicates that the model correctly classified 52 instances as ”regular” (true negatives) and 101 instances as ”excellent” (true positives). However, there were 45 false positives (instances incorrectly classified as ”excellent”) and 102 false negatives (instances incorrectly classified as ”regular”). These results show an improvement in the model’s ability to identify ”regular” cases compared to the previous analysis without SMOTE, where the model did not correctly classify any ”regular” cases.

To evaluate the model’s performance after applying SMOTE, several metrics were calculated, including accuracy, precision, recall, and F1-score:

- Accuracy: The accuracy of the model was 0.5100, indicating that approximately 51% of the predictions were correct. While this value appears lower than before, it reflects a more balanced consideration of both classes.
- Precision: The precision score was 0.6918, meaning that 69.18% of the instances predicted as ”excellent” were actually ”excellent.” Precision is a measure of the accuracy of positive predictions.
- Recall: The recall was 0.4975, indicating that the model correctly identified 49.75% of the actual ”excellent” cases. This metric reflects the model's sensitivity in detecting positive instances.
- F1-Score: The F1-score was 0.5788, representing a balance between precision and recall. The F1-score is useful for assessing the model's performance on imbalanced datasets.

These results suggest that the application of SMOTE has led to a more balanced model that better captures both ”regular” and ”excellent” cases, although there is still room for improvement in recall.

The ROC curve, shown in Figure 9, was plotted to evaluate the model's ability to distinguish between the two classes. The Area Under the Curve (AUC) was 0.50, which indicates that the model's discriminative ability is equivalent to random guessing. This result suggests that despite applying SMOTE, the model still struggles to effectively differentiate between ”regular” and ”excellent” cases.

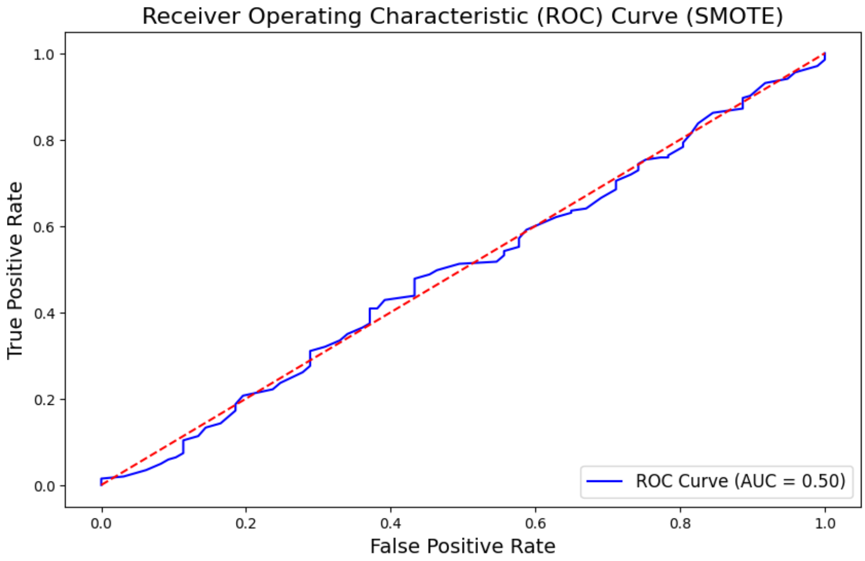

Figure 9. ROC Curve for the Logistic Regression Model with SMOTE.

Figure 10 presents the precision-recall curve for the model after SMOTE was applied. The curve illustrates the trade-off between precision and recall across different probability thresholds. The results show fluctuating precision values at low recall levels, indicating variability in the model's predictions. The curve's shape further highlights that while precision remains above 0.6 for most thresholds, there is significant room for improvement in both metrics.

The importance of each predictor in the logistic regression model was assessed by examining the coefficients, as shown in Figure 11. The most influential features were ”Location Category: Urban,” ”Time of Booking: Morning,” and ”Customer Loyalty Status: Silver,” which had the highest positive coefficients, indicating a stronger association with an ”excellent” rating. On the other hand, ”Time of Booking: Night” had a negative coefficient, suggesting that rides booked at night were less likely to be rated as ”excellent.”

The distribution of predicted probabilities for ”regular” and ”excellent” classes is shown in Figure 12. The histogram reveals overlapping distributions between the two classes, particularly around the 0.5 threshold, which indicates that the model often assigns similar probabilities to both classes. This overlap may explain the model's difficulty in distinguishing between ”regular” and ”excellent” ratings.

Applying SMOTE has helped create a more balanced model, though challenges remain in achieving higher discriminatory power.

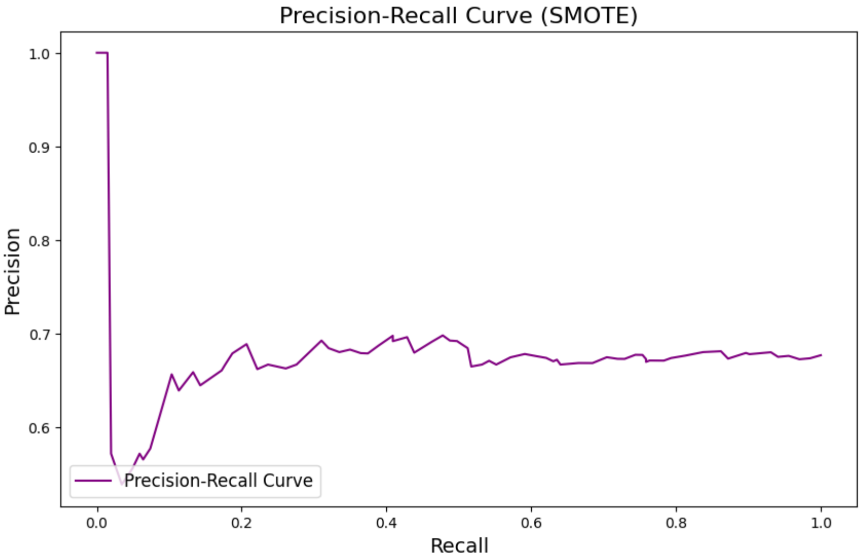

Figure 10. Precision-Recall Curve for the Logistic Regression Model with SMOTE.

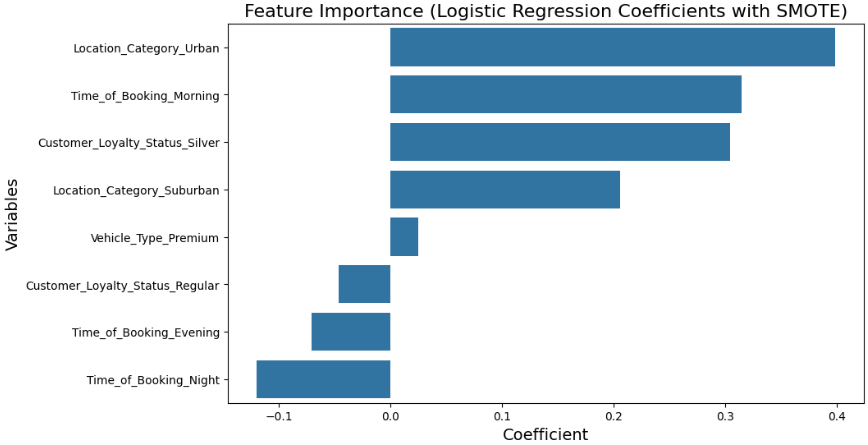

Figure 11. Feature Importance for Logistic Regression Coefficients with SMOTE.

The decision to use SMOTE was based on its ability to generate synthetic samples of the minority class, thereby reducing bias in model learning while preserving the characteristics of the feature space. Although the Area Under the ROC Curve (AUC) remained at 0.5 after applying SMOTE, other evaluation metrics such as precision, recall, and F1-score showed improvements. Specifically, precision increased from 0.6767 to 0.6918, and the recall improved in identifying the minority class, which had previously been classified poorly. These improvements, although modest, were critical in achieving a more balanced classification performance. The use of SMOTE allowed the model to understand better the characteristics of both classes, especially the minority ”regular” class, which would have otherwise been overshadowed by the majority class during training. The AUC of 0.5 indicates that the model’s discriminative ability between the two classes remained equivalent to random guessing, suggesting a limitation in the logistic regression model itself rather than in the application of SMOTE. This observation highlights the need for more sophisticated classification techniques, such as ensemble methods, to improve the overall model discriminative power. While SMOTE is not a definitive solution to class imbalance, it provides a foundation upon which further improvements can be built, including the use of additional balancing

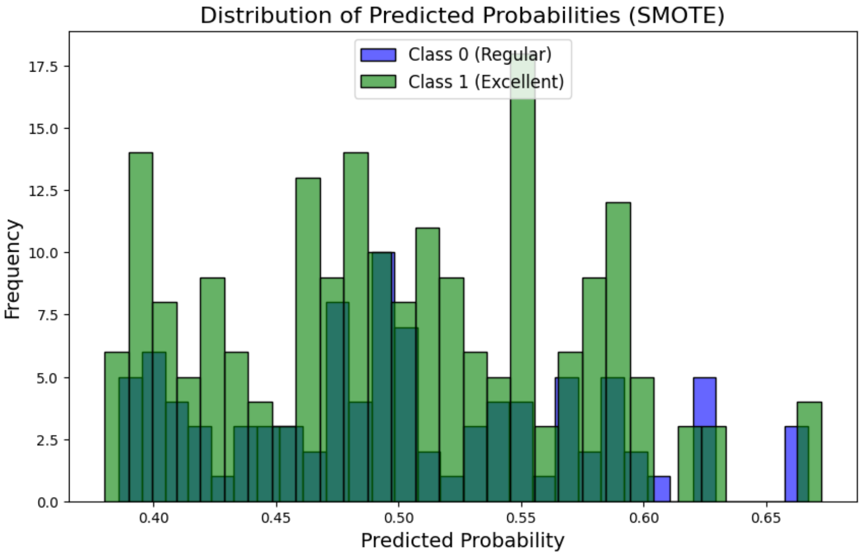

Figure 12. Distribution of Predicted Probabilities for the Logistic Regression Model with SMOTE.

techniques like Tomek Links or applying advanced machine learning models, such as Random Forest, to better capture the complexity of the underlying relationships in the data.

3.4. Comparative Analysis of Methods

Table 2 and Table 3 present a comparative summary of the methods analyzed in this study. Each method is evaluated based on key metrics, including significant predictors, goodness-of-fit measures, and limitations.

Table 2. Comparison of Methods Analyzed (Part 1)

Aspect	**Simple Linear Regression**	**Multiple Linear Regression**
Key Predictors	Ride Duration ($\beta = 3.53$, $p < 0.001$)	Ride Duration ($\beta = 3.5339$, $p < 0.001$)
Goodness-of-Fit	$R^2 = 0.860$	$R^2 = 0.862$
Advantages	Simplicity, interpretability	Incorporates multiple predictors, robust fit
Limitations	Limited scope, ignores other variables	Sensitive to multicollinearity

Table 3. Comparison of Methods Analyzed (Part 2)

Aspect	**Logistic Regression**
Key Predictors	Urban Location (OR = 1.21), Loyalty (OR = 1.54)
Goodness-of-Fit	Precision (0.6918), F1-Score (0.5788)
Advantages	Captures binary outcomes, interpretable odds ratios
Limitations	Struggles with imbalanced classes

This comparison highlights that simple and multiple linear regression are effective for predicting numerical outcomes such as ride cost, with ride duration consistently emerging as a significant predictor. Logistic regression, while useful for binary classification, faces challenges in handling class imbalance and requires further refinement to improve discriminative power.

3.5. Ethical and Privacy Considerations

The use of customer data in ride-sharing platforms raises ethical and privacy concerns that must be addressed to ensure responsible and fair practices. Ride-sharing companies collect extensive customer data to improve service quality and implement dynamic pricing strategies. To ensure the responsible use of this data, companies must obtain explicit user consent for data collection and processing, limit data usage to legitimate business purposes, and implement robust security measures to prevent unauthorized access. Transparent communication regarding customer data use can further enhance user trust and satisfaction.

Adherence to privacy regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) is essential for protecting user privacy. These regulations require companies to uphold standards for data handling, including ensuring that users can access, correct, and delete their data. Additionally, companies must notify users promptly during a data breach, maintaining compliance and safeguarding user trust.

While effective, dynamic pricing algorithms may inadvertently perpetuate or amplify biases, leading to unfair pricing practices that disproportionately impact certain customer groups, such as those from specific locations or socioeconomic backgrounds. To mitigate these risks, companies should regularly audit their pricing algorithms, implement fairness constraints during model development, and employ bias-detection techniques to identify and address unintended disparities. These measures can help ensure that pricing strategies remain equitable and inclusive.

4. Conclusion

The analysis conducted through simple linear regression revealed that the expected ride duration is a significant predictor of the historical cost, with a positive association indicating that longer trips lead to higher costs. Specifically, the model estimated an increase of \$3.53 per additional minute of ride duration ($R^2 = 0.860$). This finding underscores the importance of considering time-based factors in dynamic pricing models for ride-sharing services. However, the number of past rides did not significantly contribute to predicting the cost, with a negligible regression coefficient of $\beta = 0.2289$ and $p = 0.257$, suggesting that historical usage patterns are less influential in determining individual trip prices.

In the multiple linear regression analysis, the inclusion of additional variables, such as the number of riders, number of drivers, and expected ride duration, allowed for a more comprehensive evaluation of factors influencing ride cost. The results confirmed that ride duration remained the most significant predictor, with a coefficient of $\beta = 3.5339$ and $p < 0.001$, while driver availability also emerged as a meaningful factor ($\beta = 0.4325$, $p = 0.004$), possibly due to its association with supply-demand dynamics. Conversely, the number of riders ($p = 0.631$) and past rides ($p = 0.620$) did not significantly affect the model, indicating that customer-specific factors may not be as crucial in pricing strategies compared to time-related and operational variables.

The logistic regression analysis aimed to predict the likelihood of a ride being rated as ”excellent.” Significant predictors included customer loyalty status and location category. For example, customers with silver loyalty status were 1.54 times more likely to provide an ”excellent” rating ($p = 0.017$), and rides in urban areas had an odds ratio of 1.21 ($p = 0.018$). However, the model struggled to accurately classify ”regular” service ratings, achieving a precision of 0.6918 and an F1-score of 0.5788 after applying SMOTE to address class imbalance. The application of SMOTE improved the detection of both ”regular” and ”excellent” cases but highlighted limitations in the logistic regression model’s discriminative power. This highlights the challenges of handling class imbalance in predicting customer satisfaction and the need for further refinement of the model.

5. Limitations and Future Work

While this study provides important insights into the factors affecting dynamic pricing and service quality in ride-sharing, several limitations should be addressed in future research. First, the models employed in this study assume linear relationships between the predictors and the response variables, which might not fully capture the complexity of real-world interactions. Nonlinear techniques such as Support Vector Machines (SVM) or Gradient Boosting could potentially offer better predictive accuracy and should be considered in subsequent studies. Additionally, although logistic regression provided interpretable results, exploring advanced machine learning models like Random Forest or XGBoost could improve the model's ability to handle class imbalance and capture intricate relationships between variables. Furthermore, the application of SMOTE for class imbalance correction demonstrated limited improvement in the Area Under the ROC Curve (AUC). Future research could explore more sophisticated imbalance handling techniques, such as cost-sensitive learning, ADASYN, or hybrid over-sampling and under-sampling approaches. Expanding the dataset to include more instances of ”regular” service ratings could also improve model performance by reducing reliance on synthetic data generation techniques. The residual analysis indicated potential heteroscedasticity, which was not addressed in the current models. Future work should consider employing methods like White's test to detect heteroscedasticity and transformations like Box-Cox to stabilize variance, potentially leading to more reliable inferences. Additionally, the integration of real-time traffic data, temporal variables (e.g., hour of the day, day of the week), and driver-specific attributes such as ratings and experience could further refine pricing models. Incorporating such variables would allow for a more dynamic and context-aware analysis of ride-sharing services. As ride-sharing platforms continue to evolve, ethical and privacy considerations should remain a core focus. This study introduces a discussion of these issues, emphasizing the need for responsible data use, compliance with regulations such as GDPR and CCPA, and transparency in pricing algorithms. Future research could explore the development of privacy-preserving algorithms that balance data utility and user confidentiality. Additionally, integrating fairness constraints into machine learning models and examining their impact on operational efficiency and customer satisfaction could provide valuable insights for the industry. Addressing these considerations will ensure that ride-sharing companies promote fairness, maintain compliance with privacy regulations, and build long-term customer trust. Lastly, while this study focused solely on dynamic pricing models, future research could include a comparative analysis with traditional fixed-price models to evaluate their relative impacts on customer satisfaction, revenue, and operational efficiency. Collecting panel data or conducting longitudinal analyses could also provide a deeper understanding of how dynamic pricing and service quality evolve over time, capturing trends, seasonality, and long-term effects of pricing strategies on customer behavior and company performance. By addressing these limitations and pursuing these directions, future studies can significantly advance the understanding and implementation of dynamic pricing in the ride-sharing industry.

Acknowledgement

The authors would like to acknowledge the use of the ”Dynamic Pricing Dataset” available on Kaggle for this research.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT to aid in redaction and grammar. After using this tool/service, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

REFERENCES

1. Grisales-Noreña, L. F., Cortés-Caicedo, B., Montoya, O. D., Sanin-Villa, D., and Gil-González, W., ‘Integration of BESS in grid connected networks for reducing the power losses and CO2 emissions: A parallel master-stage methodology based on PDVSA and PSO”, *Journal of Energy Storage*, vol. 87, pp. 111355, 2024, Elsevier.
2. Grisales-Noreña, L. F., Sanin-Villa, D., and Montoya, O. D., ‘Optimal integration of PV generators and D-STATCOMs into the electrical distribution system to reduce the annual investment and operational cost: A multiverse optimization algorithm and matrix power flow approach”, *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, vol. 9, pp. 100747, 2024, Elsevier.

3. Chen, Q., Lei, Y., and Jasin, S., ‘Real-time spatial–intertemporal pricing and relocation in a ride-hailing network: Near-optimal policies and the value of dynamic pricing”, *Operations Research*, 2024.
4. Hu, T., and Hung, Y., ‘Dynamic Pricing Analysis under Demand-Supply Equilibrium of Autonomous-Mobility-on-Demand Services”, *Networks and Spatial Economics*, 2024.
5. Smith, J., ‘Dynamic pricing in shared mobility services: A review”, *Journal of Transport Economics*, 2021.
6. Wang, T., Hu, S., and Jiang, Y., ‘Predicting shared-car use and examining nonlinear effects using gradient boosting regression trees”, *International Journal of Sustainable Transportation*, 2021.
7. Chen, W., Patel, A., and Kumar, S., ‘Customer satisfaction in ride-sharing: A data-driven approach”, *IEEE Transactions on Services Computing*, 2023.
8. Hidayat, M. A., Rasyid, A., and Pasolo, F., ‘Service Quality on Customer Loyalty: Mediation of Customer Satisfaction”, *Advances in Business & Industrial Marketing Research*, 2024.
9. Chatzi, A., and Doody, O., ‘The one-way ANOVA test explained”, *Nurse Researcher*, 2023.
10. Kishore, K., and Jaswal, V., ‘Statistics corner: Chi-squared test”, *Postgraduate Medical Education Research*, 2023.
11. Sanin-Villa, D., Hernandez, C. M., and Martinez, A. F., ‘Analyzing Travel Service Costs and Customer Behavior Through ANOVA and Chi-Square Tests”, *Futurity Proceedings*, vol. 1, 2024.
12. Norris, J. J., and Xiong, H., ‘Ride-Sharing and the Geography of Consumption Industries”, *The Economic Journal*, 2023.
13. Wang, Z., and Li, S., ‘Competition between autonomous and traditional ride-hailing platforms: Market equilibrium and technology transfer”, *Transportation Research Part C: Emerging Technologies*, 2024.
14. Johnson, T., and Liu, X., ‘Temporal factors in fare optimization”, *International Journal of Transportation*, 2022.
15. James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J., ‘Linear regression”, *An Introduction to Statistical Learning: With Applications in Python*, 2023.
16. Lai, J., Zou, Y., Zhang, J., and Peres-Neto, P. R., ‘Generalizing hierarchical and variation partitioning in multiple regression and canonical analyses using the rdacca.hp R package”, *Methods in Ecology and Evolution*, 2022.
17. Das, A., ‘Logistic regression”, *Encyclopedia of Quality of Life and Well-Being Research*, 2024.
18. Adi Pratama, F. R., and Oktora, S. I., ‘Synthetic Minority Over-sampling Technique (SMOTE) for handling imbalanced data in poverty classification”, *Statistical Journal of the IAOS*, 2023.
19. Dalić, I., and Terzić, S., ‘Violation of the assumption of homoscedasticity and detection of heteroscedasticity”, *Decision Making: Applications in Management and Engineering*, vol. 4, no. 1, pp. 1–18, 2021.