MACHINE LEARNING-BASED TRAVEL COST PREDICTION FOR RICKSHAW TRIPS IN DHAKA CITY

Arnob Protim Roy*[1], Md. Zunaid Farouque[2], Bharga Narayan Tarafder[3], and Nasrin Jahan Shila[4]

[1] *PG Student , Bangladesh University of Engineering & Technology, Dhaka,Bangladesh , e-mail: pr.arnob@gmail.com*
[2] *PG Student , Bangladesh University of Engineering & Technology, Dhaka,Bangladesh, e-mail: zunaid.farouque@gmail.com*
[3] *PG Student , Bangladesh University of Engineering & Technology, Dhaka,Bangladesh, e-mail: argorayan0@gmail.com*
[4] *PG Student , Bangladesh University of Engineering & Technology, Dhaka,Bangladesh, e-mail: shilacebuet3124@gmail.com*

***Corresponding Author**

ABSTRACT

Cycle rickshaws constitute over 30% of daily trips in Dhaka, yet they operate within an unregulated fare system where prices are determined through opaque, informal negotiations. This practice creates significant price uncertainty for passengers and income volatility for pullers. To address this gap, the study develops a machine learning framework for predicting rickshaw fares, introducing a data-driven approach to a traditionally data-poor, informal transport sector. Using a primary survey dataset of 3,346 cleaned observations, interaction and polynomial features were engineered to capture complex, non-linear relationships between trip characteristics, including distance, number of intersections crossed, time, weather, and the negotiated fare. A range of regression models were evaluated, from linear baselines to advanced tree-based ensembles. An optimized ensemble combining Gradient Boosting and CatBoost regressors achieved the highest performance, explaining 79.5% of fare variance ($R^2 = 0.795$) with a Mean Absolute Error of 9.71 Taka. Feature importance analysis identified trip distance and the number of intersections crossed as the most influential predictors, providing quantitative evidence that route complexity, a proxy for congestion and physical effort, is a primary driver of fare determination. Overall, the proposed framework offers a robust foundation for digital tools that can provide real-time equitable fare suggestions and support evidence-based policy for integrating informal transport into modern urban planning.

Keywords: *Rickshaw Fare Prediction, Machine Learning, Informal Transport, Urban Mobility, Dhaka*

1. INTRODUCTION

In the urban landscape of Dhaka, the cycle-rickshaw is not merely a mode of transport but the veritable lifeblood of the city's mobility network, accounting for over thirty percent of daily trips (The Asian Age, 2025). These human-powered vehicles are an indispensable feature, prized for their ability to provide affordable, last-mile connectivity by navigating the narrow and congested lanes where formal public transport cannot operate (Hossain & Susilo, 2011). This door-to-door convenience makes them a cornerstone of mobility for millions, particularly for women, children, and the elderly who often perceive them as a safer and more personalized alternative to dangerously overcrowded public buses (The Asian Age, 2025). Beyond its role in transport, the rickshaw sector is a primary driver of Dhaka's informal economy, providing a vital source of livelihood for hundreds of thousands of rural migrants who find few other employment opportunities in the urban environment (Karim & Salam, 2019; Roy et al., 2024).

Despite its critical function, the rickshaw sector is deeply embedded in the informal economy, and a central manifestation of this is the complete absence of a standardized fare system (Hasan & Davila, 2018). There is no government control, metered pricing, or fixed fare chart for rickshaw journeys. Instead, fares are determined through a process of on-the-spot negotiation between the passenger and the puller, a transaction governed by a complex interplay of perceived distance, time, weather, and traffic. This unregulated environment creates a "fare anarchy" that mirrors issues seen in other transport sectors in the city, leading to significant price opacity, uncertainty, and the potential for financial exploitation of passengers, while also contributing to the volatility of rickshaw pullers (Rahman, 2024; The Asian Age, 2025).

The application of machine learning to predict fares in the formal transportation sector is well-established, with a substantial body of research dedicated to taxis and the dynamic pricing of ride-hailing services (Huang, 2023; Silveira-Santos et al., 2023). These studies typically operate in a "data-rich" environment, benefiting from large-scale, passively collected datasets that contain millions of trip records complete with precise GPS coordinates and timestamps. However, a significant research gap exists when it comes to the informal transport sector, which forms the backbone of mobility in many cities across the Global South. The very nature of informality means such systems are "data-poor"; trip data is ephemeral and not systematically recorded, making large-scale quantitative analysis exceptionally challenging (Gram-Hansen et al., 2019). Consequently, the application of predictive modeling to informal paratransit systems, and specifically to cycle-rickshaws in a megacity like Dhaka, remains a largely unexplored frontier that this research aims to address.

The primary objective of this research is to develop and evaluate a machine learning framework capable of accurately predicting rickshaw fares based on a set of easily observable, trip-specific variables. This paper makes several novel contributions to the fields of urban transport and data science. Firstly, it introduces a unique, primary dataset on informal transport fares, meticulously collected through direct physical surveys in a data-scarce environment. Secondly, it presents one of the first comparative applications of multiple machine learning models to quantitatively model and predict cycle-rickshaw fares in Dhaka. Thirdly, through feature importance analysis, it empirically identifies and ranks the key determinants of fare variability, providing quantitative evidence for the factors that govern price negotiations. The resulting framework provides a reproducible, data-driven approach to underpin fairer fare suggestions, inform mobile estimation tools, and guide evidence-based policy.

2. DATASET

2.1 Data Collection

Primary data were collected through in-person surveys of rickshaw pullers in major Dhaka zones: Polashi-Nilkhet, Mohammadpur, Farmgate, Uttara, BUET, and Mirpur.

Each record contains:

- Origin and destination (categorical; zone labels)
- Distance (km) - numeric
- Intersection_crossed - numeric
- Geographical zone types for origin/destination (commercial/residential) - categorical
- Weather (Sunny/Rainy) - categorical
- Rider_no. (integer) - numeric
- Time_of_day (Day/Night) - categorical
- AVG_TK (average fare in Taka) - numeric (target)

2.2 Data Cleaning

Given that the primary data were collected through manual in-person surveys, a process susceptible to entry errors, inconsistencies, and recall bias (Bonnel & Munizaga, 2018), a rigorous data cleaning pipeline was implemented to ensure data quality and reliability. The variables Intersection_crossed, Rider_no, and AVG_TK were converted to integer types where appropriate. Missing or invalid entries were handled through case-by-case imputation or removal, with exclusion criteria documented in the methods supplement. Categorical expansions were performed by splitting "Any time" into two entries (Day and Night), duplicating "Any Weather" for Sunny and Rainy conditions, and expanding "Any" in Rider_no to represent 1, 2, and 3 riders. Outliers in continuous variables were addressed using winsorization, capping values above the 99th percentile. Numerical features, including Distance, Intersection_crossed, and engineered polynomial or interaction terms, were standardized using the Standard Scaler, while categorical variables such as From_Zone, Destination_Zone, Weather, and Time_of_day were transformed through one-hot encoding to prepare the dataset for model training, adhering to standard data mining practices (Koukaras & Tjortjis, 2025). The initial corpus consisted of 1,138 raw observations. After cleaning and categorical expansion, the final processed dataset comprised 3,346 observations.

2.3 Feature Engineering

Feature engineering was performed to address the observed nonlinearity and heteroscedasticity between trip characteristics and fare values. It was hypothesized that interactions between variables, particularly Distance and Intersection_crossed, exert a multiplicative influence on fare, as longer trips with more intersections typically involve higher travel time and delays. To capture these complex relationships, interaction and polynomial features were introduced, including Distance × Intersection to represent the combined effect of trip length and route complexity, and squared terms ($Distance^2$ and $Intersection^2$) to model accelerating fare growth with increasing distance and intersections. Polynomial transformations up to the second degree were applied to key numerical attributes to better capture non-linear fare patterns, thereby enhancing model flexibility and predictive accuracy.

3. METHODOLOGY

3.1 Model Selection

To predict average ride cost, a diverse set of machine learning models was employed to evaluate different algorithmic approaches suited to the dataset's characteristics. The selected models spanned both linear and non-linear frameworks, enabling a comprehensive performance comparison across varying complexities.

3.1.1 Linear Models

- **Linear Regression:** Baseline model assuming linear feature-target relationships.

- **Ridge Regression:** Adds L2 regularization to reduce overfitting by penalizing large coefficients.
- **Lasso Regression:** Employs L1 regularization to perform implicit feature selection by shrinking some coefficients to zero.

3.1.2 Tree-Based and Ensemble Models

- **Random Forest Regressor:** Builds multiple decision trees and averages their predictions, capturing non-linearities robustly.
- **Gradient Boosting Regressor:** Sequentially adds trees to correct prior errors.
- **XGBoost Regressor:** Optimized gradient boosting framework offering high predictive accuracy and computational efficiency.
- **LightGBM Regressor:** Gradient boosting algorithm optimized for speed and scalability on large datasets.
- **CatBoost Regressor:** Efficient gradient boosting method with native handling of categorical variables and high robustness.
- **Ensemble (Voting Regressor):** Combines predictions from top-performing models through averaging to leverage their collective strengths.

This comprehensive set of models ensured coverage of multiple algorithmic paradigms, facilitating identification of the most effective predictive approach when integrated with engineered features.

3.2 Model Training and Evaluation

A systematic procedure was followed to ensure robust and reliable model performance assessment. After comprehensive preprocessing and feature engineering, the dataset was divided into training (80%) and testing (20%) subsets, using a fixed random state to maintain reproducibility; this separation is a fundamental practice to prevent data leakage and provide an unbiased assessment of how the models will generalize to unseen data (Huang, 2023). The initial phase involved training baseline models-Linear Regression, Ridge, Lasso, Random Forest, and Gradient Boosting-on various combinations of basic features to establish benchmark performance and identify promising feature subsets. Random Forest and Gradient Boosting, which exhibited comparatively strong initial results, underwent hyperparameter optimization using GridSearchCV with 5-fold cross-validation. A grid of parameters, including the number of estimators, learning rate, maximum depth, minimum samples split, minimum samples leaf, and maximum features, was systematically explored to identify configurations maximizing the R^2 score across validation folds.

Following this, advanced tree-based models such as XGBoost, LightGBM, and CatBoost were trained on the fully engineered feature set to capture complex, non-linear interactions within the data. The tuned Gradient Boosting model was also retrained on this feature set for consistent comparison with these advanced algorithms. To leverage the complementary strengths of top-performing models, an averaging ensemble (Voting Regressor) was constructed using the retrained tuned Gradient Boosting and CatBoost Regressors.

All models were evaluated on the unseen test set using standard regression metrics to ensure a fair comparison. Model accuracy and robustness were quantified through Mean Absolute Error (MAE), which measures the average magnitude of prediction errors; Mean Squared Error (MSE), which penalizes larger deviations and highlights sensitivity to outliers; and the coefficient of determination (R^2), representing the proportion of variance in fare values explained by the model, with $R^2 = 1$ denoting perfect prediction and $R^2 = 0$ indicating no predictive power. This structured pipeline-comprising baseline exploration, hyperparameter tuning, advanced modeling, and ensemble averaging-enabled a rigorous comparison of model performances and identification of the most suitable approach for predicting average rickshaw trip costs.

4. RESULT

Figure 4.1 shows histograms of numerical features after addressing skewness and outliers, summarizing the preprocessing impact. Before cleaning, "Distance (km)" and "Intersection crossed" were highly right skewed with long tails and extreme values. After applying winsorization (capping at the 99th percentile) and standardization, distributions became more symmetric, reducing outlier influence. Engineered terms (Distance × Intersection, Distance², Intersection²) showed smoother, balanced patterns, confirming effective scaling. Rider_no remained stable, preserving categorical balance. These adjustments enhanced data quality and model stability.

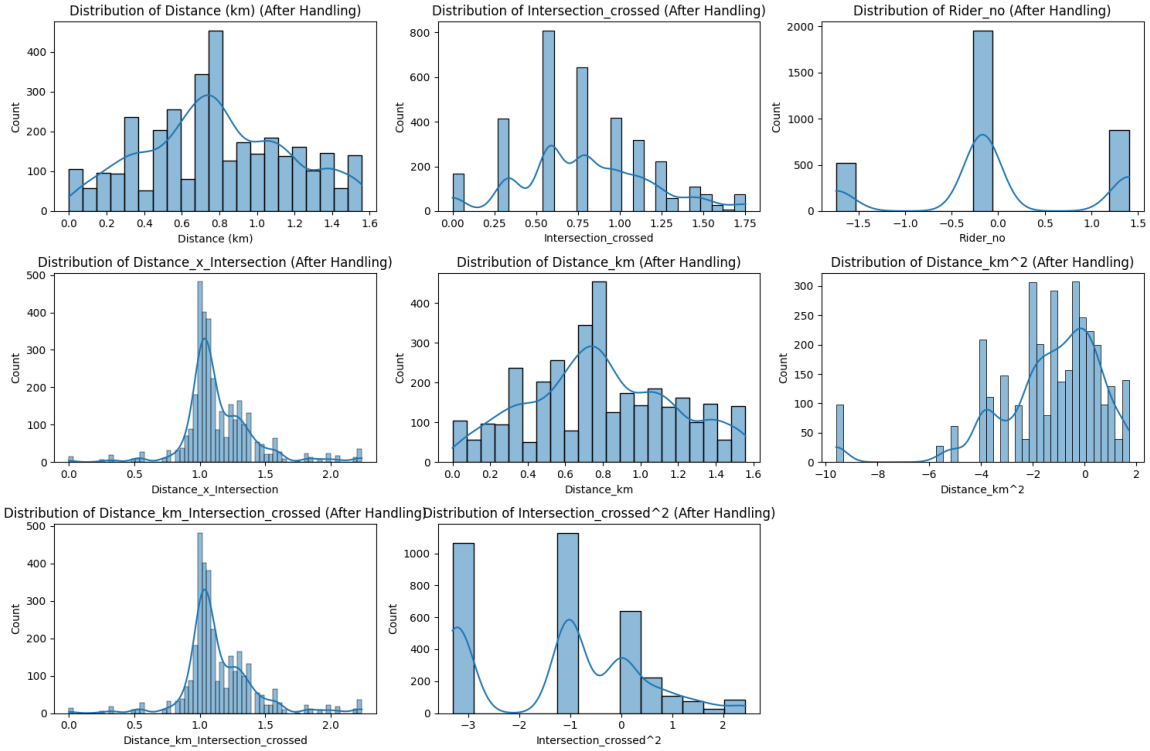

Figure 4.1: Histograms of Numerical Features Handling Skewness and Outliers

The cleaned data showed that most rides were short with few intersections, two-passenger trips were most frequent, and AVG_TK remained right-skewed-indicating common low fares with fewer high-cost trips. The cleaned dataset revealed distinct distributional patterns: most rides were short distance with few intersections, while trips with two passengers were the most frequent. The target variable (AVG_TK) also showed a right-skewed distribution, indicating that lower fares dominate with a long tail of higher-cost trips.

4.1 Distributions and Relationships of Features

Histograms of numerical features showed right-skewed distributions for Distance (km) and Intersection_crossed, indicating most rides were short with few intersections, and fewer long trips. Rider_no peaked at two riders, while AVG_TK was also right-skewed, with low fares dominant and a tail of higher costs. Scatter plots (Figure 4.2) revealed a positive linear trend between Distance and average fare (AVG_TK), confirming that longer trips cost more. Intersection_crossed also correlated positively but less linearly, suggesting higher costs for more complex routes. Box plots (Figure 4.3) show that rides during rainy conditions have slightly higher and more variable fares than those during sunny conditions. Additionally,night rides show a higher median fare and higher dispersion than day rides.

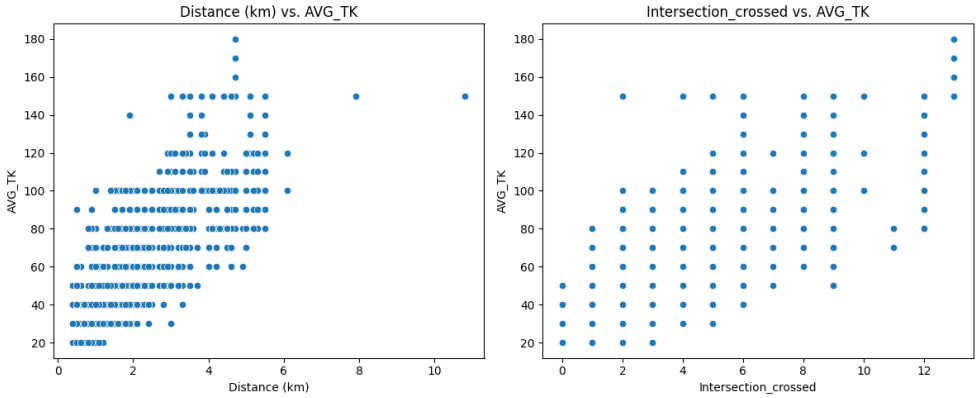

Figure 4.2: Scatter Plots Showing Relationships Between Key Numerical Variables and Average Fare

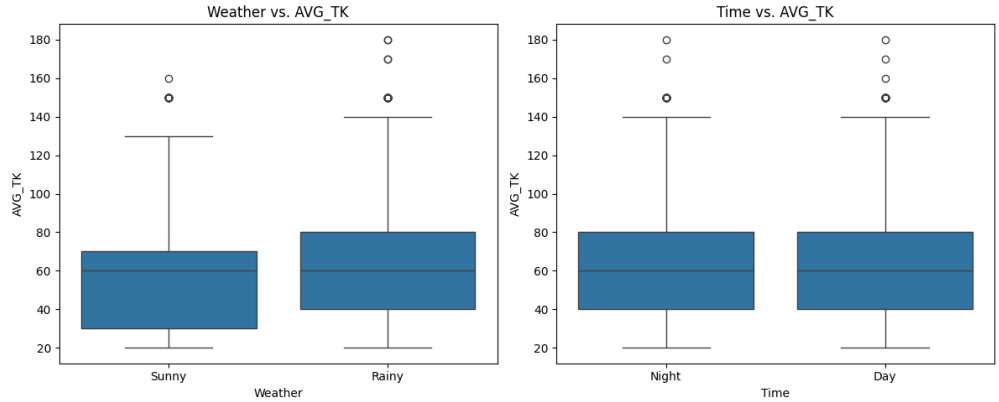

Figure 4.3: Box Plots of Average Fare across Weather and Time Conditions

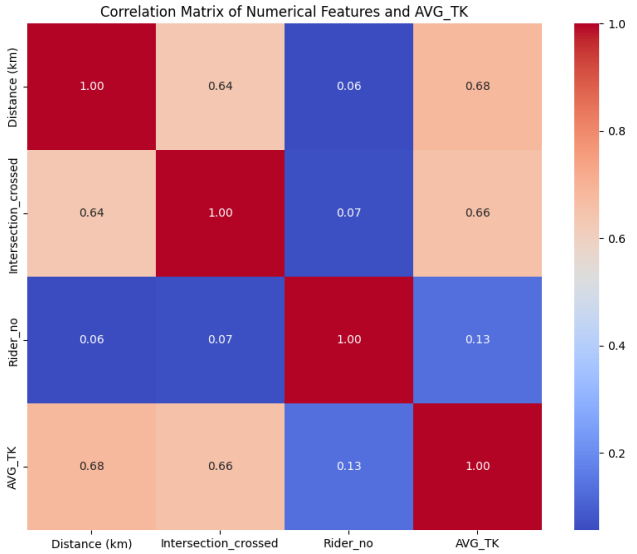

Figure 4.4: Correlation Heatmap of Numerical Features and Average Fare

The correlation heatmap (Figure 4.4) confirmed Distance as the strongest predictor of AVG_TK (r = 0.68), followed by Intersection_crossed (r = 0.64). Additionally, low inter-feature correlations indicated minimal multicollinearity.

These initial data exploration findings provided a foundational understanding of the dataset, highlighting the distributions of individual features and revealing preliminary relationships between predictor variables and the target variable, which guided the subsequent feature engineering and modeling efforts.

4.2 Model Performance Comparison

A comprehensive evaluation was conducted to compare the performance of the various regression models and feature sets employed in predicting the average ride cost. The models were assessed based on key regression metrics: R-squared (R^2), Mean Absolute Error (MAE), and Mean Squared Error (MSE). The initial exploration with basic feature combinations provided baseline performance indicators. As shown in summary table (Table 4.1), the performance varied significantly across different models and feature sets.

As summarized in Table 4.1, model performance varied considerably across algorithms and feature representations. Models trained on engineered features consistently outperformed those using only the initial basic variables. The advanced tree-based algorithms-CatBoost, LightGBM, and XGBoost-along with the ensemble model, demonstrated the highest predictive accuracy, achieving notably higher R^2 values and lower error metrics. Among all, the Ensemble model (Retrained GBR + CatBoost) achieved the best overall performance with an R^2 of 0.7953, MAE of 9.7098, and MSE of 162.2405, closely followed by the CatBoost Regressor (R^2 = 0.7924).

These results confirm that the inclusion of engineered interaction and polynomial features significantly enhanced model capability by capturing complex non-linear relationships between trip attributes and fare. While hyperparameter tuning of the Gradient Boosting Regressor on basic features provided modest improvements (R^2 = 0.7006), the impact was far less pronounced than that achieved through feature engineering and the use of advanced ensemble techniques.

Table 4.1: Summary of Top Performing Models and Feature Combinations

Model	Features	MAE	MSE	R^2
Ensemble (Retrained GBR + CatBoost)	Engineered Features	9.70	162.24	0.79
CatBoost Regressor	Engineered Features	9.71	164.53	0.79
LightGBM Regressor	Engineered Features	9.99	169.98	0.78
XGBoost Regressor	Engineered Features	9.80	181.40	0.77
Tuned GradientBoostingRegressor (Initial Features)	Distance (km), Intersection_crossed, Weather_Sunny	11.85	237.35	0.70

Overall, the comparison underscores the importance of combining robust feature design with powerful non-linear modeling algorithms to achieve accurate, data-driven fare prediction in informal urban transport systems.

4.3 Residual Analysis

Analyzing the residuals, which are the differences between the actual observed values and the values predicted by the model, is a crucial step in evaluating the model's performance beyond aggregate metrics. Residual analysis helps to identify patterns in the errors, assess the assumptions of the regression model, and understand where the model might be performing less effectively.

For the best-performing models, particularly the Ensemble model, the residuals were calculated and visualized. The scatter plot of predicted values versus residuals (Figure 4.5) shows the distribution of errors across the range of predicted average costs. Ideally, for a well-performing model with homoscedastic errors, the residuals should be randomly scattered around zero with no discernible pattern or fanning-out effect. The histogram of the residuals (Figure 4.5) illustrates the distribution of the errors themselves. For an unbiased model, the residuals should ideally be normally distributed around a mean of zero. The residual plots for the Ensemble model indicate that, following feature engineering and the use of advanced models, the residuals are relatively scattered around the zero line. While some slight fanning out might still be observed at higher predicted values (suggesting some

remaining heteroscedasticity), the pattern is significantly reduced compared to what might be expected from simpler models on less-engineered features.

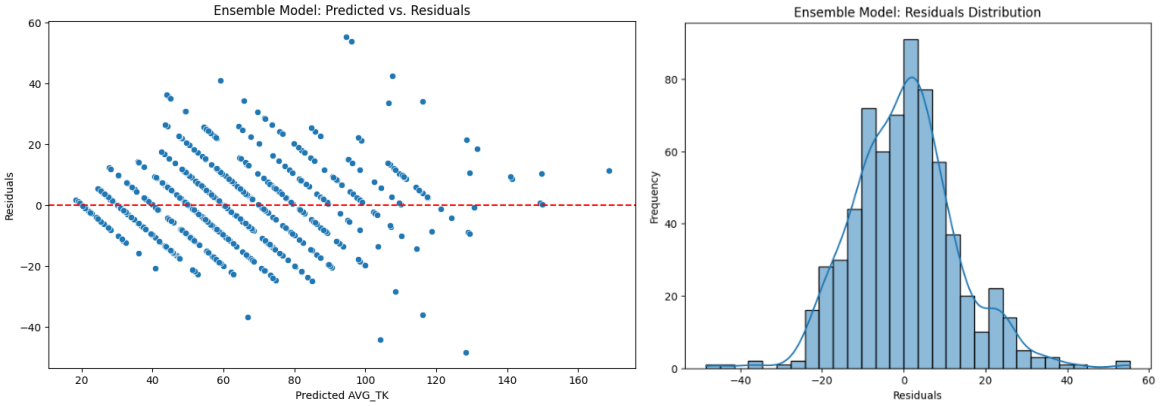

Figure 4.5: Residual Analysis of the Ensemble Model (Retrained GBR + CatBoost)

The histogram of residuals shows a distribution that is approximately centered around zero, although it may exhibit some slight skewness or deviations from perfect normality. These findings suggest that the model's errors are generally unbiased and distributed around zero, supporting the overall good performance indicated by the R^2, MAE, and MSE metrics. Further analysis or transformation might be considered if the remaining heteroscedasticity or non-normality of residuals is deemed problematic for specific applications, but for many predictive tasks, the current state represents a substantial improvement.

4.4 Feature Importance

Table 4.2: Feature Importances of CatBoost Regressor and Tuned Gradient Boosting Regressor

CatBoost Regressor		Tuned Gradient Boosting Regressor	
Feature	**Importance**	**Feature**	**Importance**
Distance (km)	35.08%	Distance (km)	44.19%
Intersection_crossed	13.56%	Intersection_crossed	18.17%
Distance_x_Intersection	10.94%	Distance_x_Intersection	10.00%
Distance_km²	9.02%	Rider_no	6.74%
Intersection_crossed²	7.26%	Distance_km²	5.04%
Rider_no	6.56%	Intersection_crossed²	3.94%
From_Zone_Residential_	3.50%	Time_Night	2.54%
Destination_Zone_Residential_	3.21%	Weather_Sunny	2.07%
Time_Night	2.34%	From_Zone_Residential_	1.81%
Weather_Sunny	2.08%	Destination_Zone_Residential_	1.57%
Distance_km_Intersection_crossed	1.72%	Distance_km_Intersection_crossed	1.22%
From_Zone_Commerical_	1.44%	From_Zone_Commerical_	0.96%
Destination_Zone_Commerical_	1.24%	Destination_Zone_Commerical_	0.73%
Destination_Zone_Educational_	1.06%	From_Zone_Educational_	0.59%
From_Zone_Educational_	0.98%	Destination_Zone_Educational_	0.44%

Understanding the relative importance of different features in the predictive models provides valuable insights into which factors have the most significant influence on the target variable.

For tree-based models such as CatBoost and Gradient Boosting, feature importance scores were calculated to quantify the contribution of each feature to the model's predictions. Analysing these scores helps identify the key drivers of average ride costs according to the models. Feature importance was specifically computed for the top-performing individual tree-based models: the CatBoost Regressor and

the retrained, tuned Gradient Boosting Regressor, both trained on the engineered feature set. The resulting feature importance values are presented in Table 4.2.

4.5 Prediction vs. Actual Plots

To visually assess the performance of the trained models and understand how well their predictions align with the true observed values of the average ride cost, scatter plots comparing the actual 'AVG_TK' values against the predicted 'AVG_TK' values were generated for the best-performing models. For a perfect model, all points would lie precisely on the diagonal line where Actual= Predicted. Deviations from this line indicate prediction errors.

These plots visually reinforce the quantitative performance metrics. The points are clustered relatively close to the diagonal line, particularly for the higher-performing models (CatBoost and the Ensemble), indicating a good level of predictive accuracy across the range of average ride costs. Some dispersion around the line is expected, representing the inherent variability in the data that the models do not fully capture. The plots allow for a qualitative assessment of the models' ability to predict both lower and higher average costs and can sometimes reveal patterns in prediction errors (e.g., systematic over- or under-prediction in certain ranges).

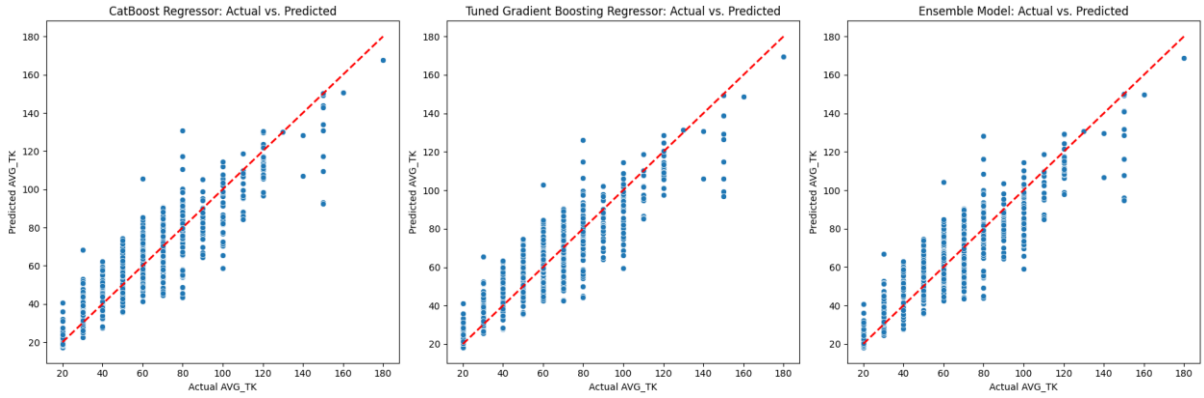

Figure 4.6: Comparison of Actual vs. Predicted Average Fare for Top Models

5. DISCUSSION

The superior performance of the Gradient Boosting and CatBoost ensemble (R^2=0.795) confirms that rickshaw fare negotiation is governed by complex, non-linear relationships. This finding aligns with established transport literature, where tree-based ensembles consistently outperform linear models in fare prediction tasks precisely because of their ability to capture intricate variable interactions (Theke, 2024). Our model, for instance, effectively learns how factors like adverse weather might amplify fares disproportionately during peak hours-a dynamic that simpler linear models would fail to detect.

The feature importance analysis empirically validates long-held assumptions about Dhaka's transport reality. While the dominance of Distance was expected, the high rank of Intersection_crossed is a critical insight. This result quantitatively confirms that route complexity and congestion-and the consequent increase in travel time and physical exertion for the puller-are primary determinants of the negotiated fare, not merely an inconvenience. It provides a data-driven quantification of the economic cost that crippling gridlock imposes on the city's informal transport sector (Haider & Papri, 2021).

Taken together, our findings confirm the feasibility of developing data-driven tools to improve transactional transparency. For policymakers, this research offers a novel lens; instead of viewing the sector as an inscrutable "black box," the model provides a quantitative deconstruction of its de facto pricing structure. This data-driven understanding, which is transferable to other paratransit systems

across the Global South, can inform future regulatory dialogue and evidence-based integration strategies (Silveira-Santos et al., 2023).

Despite its strong predictive performance, the model has limitations inherent to its static, survey-based data. Residual analysis indicates a modest increase in error variance for higher-priced trips (heteroscedasticity), suggesting the influence of unobserved dynamic variables such as real-time congestion, specific negotiation tactics, or route-level road quality. This remaining variance underscores the challenge of modeling data-poor informal systems and presents a clear opportunity for future work to incorporate more dynamic, spatiotemporal data-such as real-time traffic feeds or GPS traces from a dedicated mobile application-for even greater accuracy.

6. CONCLUSION

This study introduces a machine learning framework to model the unregulated fare system of cycle rickshaws in Dhaka, a transport sector traditionally characterized by data scarcity. Through systematic feature engineering applied to a unique primary survey dataset, an optimized ensemble of Gradient Boosting and CatBoost regressors achieved strong predictive performance, explaining 79.5% of the variance in negotiated fares. Feature importance analysis reveals a key insight: beyond trip distance, the number of intersections crossed emerges as a dominant predictor. This finding provides robust quantitative evidence that route complexity, acting as a proxy for congestion and physical effort, is a primary determinant of fare formation within this informal market.

The proposed framework establishes a practical foundation for developing real-time fare estimation tools that can improve transactional transparency for both passengers and rickshaw pullers. Moreover, by unpacking the sector's implicit pricing mechanisms, this research offers an empirical basis to support evidence-based urban transport policy and constructive regulatory dialogue. Future research should focus on integrating dynamic data streams, such as GPS-based trip traces, to capture real-time traffic variability and further enhance predictive accuracy.

7. LIMITATIONS AND FUTURE WORK

This study has several limitations related to both data and modeling. The survey-based dataset, while feasible, is self-reported and subject to recall and response biases, as accuracy depends on respondents' memory and truthful reporting. Data were collected from a limited number of zones in Dhaka and represent a static snapshot in time, restricting generalizability across seasons, temporal variations, or other cities. The dataset also lacks dynamic variables such as real-time traffic flow, route choices, and road conditions, which are known to strongly influence fare outcomes. From a modeling perspective, the trained models are static and cannot adapt to changing travel behaviour or demand patterns without retraining, limiting their direct applicability in real-world deployment scenarios. Additionally, the current feature engineering approach treats zones as independent categorical variables, which restricts spatial reasoning and prevents the model from capturing relationships between nearby locations.
Future work should prioritize app-based and sensor-driven data collection approaches to address these limitations. A mobile application for volunteer rickshaw pullers could leverage GPS tracking to record detailed trip-level information, including routes, travel times, speeds, and final negotiated fares, thereby generating a richer and more reliable dataset. Incorporating geospatial attributes such as Haversine or OpenStreetMap-based route distances, network density, points of interest, and socio-economic indicators, along with external data sources such as real-time weather and traffic APIs, would further enhance model precision. With access to spatially and temporally rich data, advanced modeling techniques such as Graph Neural Networks (GNNs) for spatial dependency learning and Long Short-Term Memory (LSTM) networks for temporal pattern capture could be explored. Finally, predicting fare ranges rather than single point estimates, through quantile regression or confidence interval-based approaches, would better reflect the inherent uncertainty of negotiation-based pricing and improve the usability and credibility of ML-driven fare estimation tools in Dhaka's informal transport context.

DECLARATION OF USE OF AI

Artificial intelligence (AI) tools were used solely for the purpose of language refinement to improve the readability and grammatical flow of the manuscript. The use of this technology was strictly limited to formatting and linguistic expression. The conceptualization, research design, data analysis, graphing, charting, and the formulation of discussion points were performed entirely by the human authors. All output generated by the AI tool was rigorously reviewed and verified by the authors to ensure accuracy, validity, and alignment with the original research findings. The authors take full responsibility for the content of the published work.

REFERENCES

Bonnel, P., & Munizaga, M. (2018). Transport survey methods—In the era of big data facing new and old challenges. *Transportation Research Procedia*, *32*, 1–15. https://doi.org/10.1016/j.trpro.2018.10.001

Gram-Hansen, B., Helber, P., Varatharajan, I., Azam, F., Coca-Castro, A., Kopackova, V., & Bilinski, P. (2019). Mapping Informal Settlements in Developing Countries using Machine Learning and Low Resolution Multi-spectral Data. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 361–368. https://doi.org/10.1145/3306618.3314253

Haider, M. Z., & Papri, R. S. (2021). Cost of traffic congestion in Dhaka Metropolitan City. *Public Transport*, *13*(2), 287–299. https://doi.org/10.1007/s12469-021-00270-4

Hasan, M. M. U., & Davila, J. (2018). The politics of (im)mobility: Rickshaw bans in Dhaka, Bangladesh. *Journal of Transport Geography*, *70*, 246–255. https://doi.org/10.1016/j.jtrangeo.2018.06.002

Hossain, M., & Susilo, Y. (2011). Rickshaw Use and Social Impacts in Dhaka, Bangladesh. *Transportation Research Record: Journal of the Transportation Research Board*, *2239*, 74–83. https://doi.org/10.3141/2239-09

Huang, H. (2023). Taxi fare prediction based on multiple machine learning models. *Applied and Computational Engineering*, *16*, 7–12. https://doi.org/10.54254/2755-2721/16/20230849

Karim, P. R., & Salam, K. A. (2019). *Organising the Informal Economy Workers: A Study of Rickshaw Pullers in Dhaka City orkers: A Study of Rickshaw Pullers in Dhaka City*.

Koukaras, P., & Tjortjis, C. (2025). Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices. *AI*, *6*(10). https://doi.org/10.3390/ai6100257

Rahman, S. (2024, January 29). *Use of metres in auto-rickshaw fading from public mind*. Prothomalo. https://en.prothomalo.com/bangladesh/city/ak08ei9ita

Roy, S., Mostafa Kamal, M., Biswas, T., Chowdhury, S., & Sani, J. (2024). Socio-Economic Condition of A Rickshaw Puller of Bangladesh- A Study on Dhaka City. *American Journal of Social Development and Entrepreneurship*, *3*, 35–41. https://doi.org/10.54536/ajsde.v3i1.3277

Silveira-Santos, T., Papanikolaou, A., Rangel, T., & Manuel Vassallo, J. (2023). Understanding and Predicting Ride-Hailing Fares in Madrid: A Combination of Supervised and Unsupervised Techniques. *Applied Sciences*, *13*(8), 5147. https://doi.org/10.3390/app13085147

The Asian Age. (2025). *How Rickshaw Pullers Keep Dhaka's Informal Transport System Moving Despite Modern Alternatives | The Asian Age Online, Bangladesh*. The Asian Age. https://dailyasianage.com/news/341258/how-rickshaw-pullers-keep-dhakas-informal-transport-system-moving-despite-modern-alternatives

Theke, P. P. B. (2024). *Estimating Taxi Fares Using Machine Learning*.