Data-Driven Ride Price Forecasting: A Feature Engineering and Modeling Framework

Manish Chauhan[1], Shaleen Saifi[2], Udit Tyagi[3], and Yogendra Narayan Prajapati[4]

Ajay Kumar Garg Engineering College, Ghaziabad, India
{manish22154003-d@akgec.ac.in, shaleen2115442@akgec.ac.in,
tyagiudit1@gmail.com, ynp1581@gmail.com}

Abstract. The pricing of ride-hailing services in India changes dynamically due to various factors such as trip distance, time of day, demand surges, traffic conditions, and weather. This study utilizes statistical and machine learning methods, including Linear Regression, Decision Trees, and Random Forests, to forecast ride fares. We analyzed over 1 million ride records from major Indian cities like Delhi, Mumbai, and Bengaluru. To enhance model accuracy, we applied extensive data preprocessing techniques, including feature engineering, normalization, and addressing missing values. Among the models tested, Gradient Boosting provided the most precise predictions, achieving a Mean Absolute Error (MAE) of 28.00 and a Root Mean Squared Error (RMSE) of 37.30, surpassing traditional regression models. K- means clustering indicated that fare prediction errors rose by 4.6% during peak demand times, mainly due to unpredictable surge pricing. The key factors influencing fare estimation were trip distance (41%) and time of day (23%). This research offers valuable insights for ride-hailing companies in India, helping them optimize pricing strategies, boost driver earnings, and enhance customer satisfaction. Furthermore, policymakers can use these insights to better understand urban mobility patterns and develop improved transportation policies. .

Keywords: Ride-hailing, Fare prediction, Machine learning, Gradient Boosting, Dynamic pricing

1 Introduction

App-based ride-hailing services like Ola, Uber, and Rapido have transformed urban mobility in India by providing a convenient and flexible alternative to traditional taxis and public transport. These platforms have become essential for daily commuting, offering affordable and on-demand trans- portation in major cities such as Delhi, Mumbai, Bengaluru, and Hyderabad. In contrast to conventional auto-rickshaws and metered taxis that have fixed pricing, ride-hailing apps employ dynamic pricing models that adjust fares according to real-time demand and supply. While this method enhances market efficiency, it also creates uncertainty for passengers, drivers, and regulators. Frequent price changes can affect affordability for riders, earnings consistency for drivers, and regulatory oversight. Thus, accurate fare prediction is vital for transparency, financial planning, and the establishment of fair pricing policies [1].

This research investigates fare estimation in Indian ride- hailing services through machine learning models like Linear Regression, Decision Trees, and Random Forests. A dataset from various ride-hailing service providers is analyzed, incor- porating extensive preprocessing techniques such as feature engineering and normalization to improve model accuracy. Furthermore, K- means clustering is applied to explore price variations during surge pricing periods. This research builds on existing studies and enhances ride- hailing The goal of this study is to evaluate the effectiveness of fare prediction in India by: machine learning-based fare prediction models using actual ride data. By pinpointing key fare determinants, this research seeks to benefit both consumers and service providers, fos- tering price transparency, optimizing demand, and enhancing pricing regulations [2].

This paper is organized in the following manner: Section 2 offers a review of existing literature on fare prediction and explores the use of machine learning in the transportation sector. Section 3 outlines the processes of data collection, preprocessing, and the selection of models. Sec- tion 4 showcases the experimental results, compares different models, and conducts an error analysis. Section 5 highlights the main findings, discusses their implications, and addresses any limitations. Lastly, Section 6 provides concluding thoughts and suggests directions for future research [3].

2 Literature Review

The study presents a detailed discussion of the problem and solution aspects of shared economy pricing particularly with regard to ride-sharing services like Uber and Lyft. The overall conclusions reached based on the survey of literature and the background presented here are as follows: Travel Revolution: The shared economy has transformed the way individuals travel, with ride-share applications becoming increasingly pop- ular day by day. This revolution requires new pricing models to cater to the dynamic supply and demand nature of this business [4]

The article ”Modeling Uber Data for Predicting Features Responsible for Price Fluctuations” talks about various aspects of transportation economics, in this case, ride-hailing firms like Uber. The following is a comprehensive literature survey on the basis of the provided context: Transportation Systems: The research situates itself within the broader field of transportation economics, which examines the nature and patterns of various modes of transport, including traditional modes like subways and taxis and newer modes like Uber and Lyft. The context therefore foregrounds the evolution of transport services and the significance of data analysis in making sense of their dynamics.

Uber Dataset: The study utilizes a specific dataset of Uber’s activity in New York City. The dataset has significant data such as pick-up time, locations (latitude and longitude), and other variables of interest. New York City has been selected since it has a complex transportation situation and a highly intense demand for ride-hailing, and it is a good test case for analyzing price volatility [6]

This research paper predicts diamond prices with machine learning, targeting traders, customers, and researchers. It em- ploys a range of techniques like linear regression, Random Forest, and Gradient Boosting, with a close focus on preprocessing. Regression (predicting actual prices) as well as classification (predicting price ranges) is explored. 1 Random Forest is extremely good at price prediction (R^2 0.975), and SVM and Logistic Regression are optimal in classification (accuracy ¿ 952 It demonstrates the capability of machine learning in accurate diamond valuation [7]

This paper focuses on predicting dynamic pricing in ride-on- demand services like Uber and Didi. While these services are popular, unpredictable prices create passenger uncertainty. Existing research is limited, with two main prediction approaches: inferring prices from supply/demand or directly predicting multipliers. This paper uses the latter, improving on previous work by incorporating multi-source urban data (taxi availability, weather, public transport) to enhance prediction accuracy through a neural network model, which outperforms baseline predictors [8].

The "VEST" paper introduces an auto feature engineering framework for time series forecasting. It integrates statistically derived features in recent times with auto-regression, which automates something otherwise done manually. Tested on 90 high-frequency time series, VEST significantly improves forecast accuracy, and its effectiveness is apparent from real- world use [9].

3 Research Objectives

– Develop a Reliable Fare Prediction Model This research aims to create a machine learning model that accurately pre- dicts ride-hailing fares in various Indian cities. By considering essential factors such as trip distance, travel time, traffic conditions, weather, and surge pricing, the model intends to deliver precise and reliable fare estimates.
– Develop a Reliable Fare Prediction Model This research aims to create a machine learning model that accurately pre- dicts ride-hailing fares in various Indian cities. By considering essential factors such as trip distance, travel time, traffic conditions, weather, and surge pricing, the model intends to deliver precise and reliable fare estimates.

4 Methodology

4.1 Data Collection and Preprocessing

This image depicts a data flow diagram of a ride price forecasting system, the process of processing the data from where it originates and into a format that

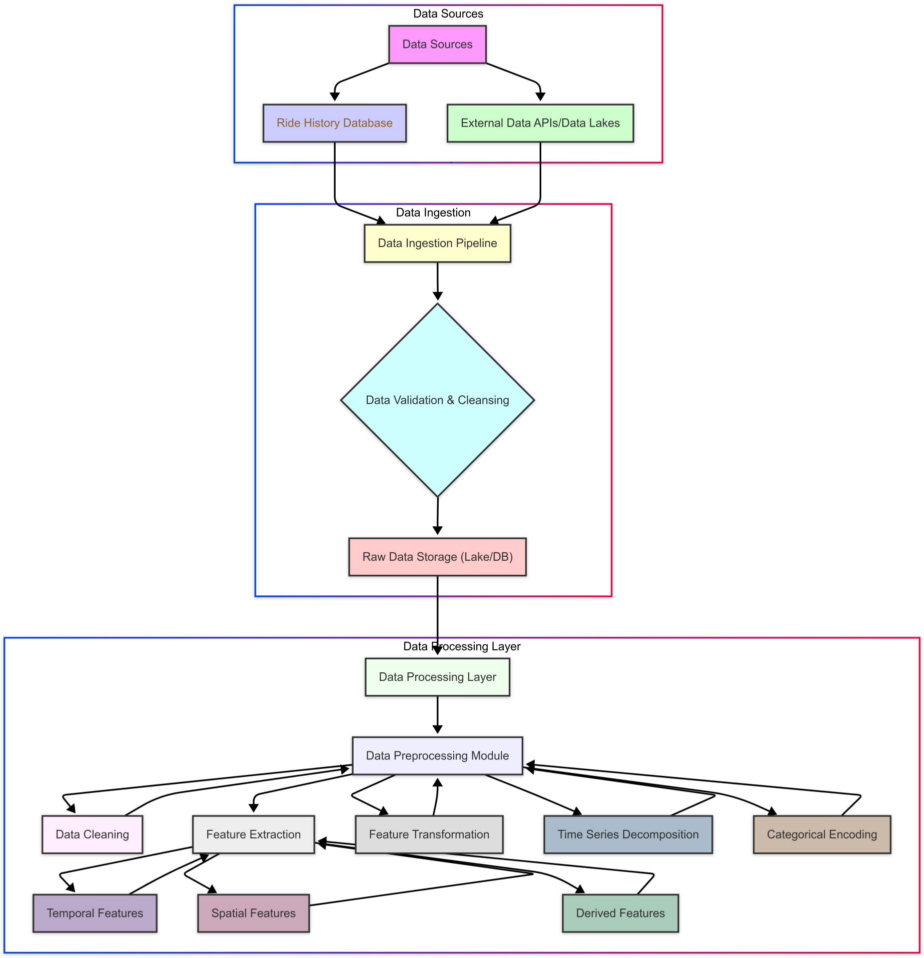

Fig. 1. Model Performance Metrics

can be used by machine learning. Let's follow each segment: 1. Data Sources (Purple Box):

Data Sources (Top Purple Box): This is the general category, i.e., where the data originates. Ride History Database (Blue Box): This is the in-memory database with ride history. It likely holds information including: Fares Timestamps (pick- up and drop-off times) Pickup and drop-off locations Ride distance Ride length External Data APIs/Data Lakes (Pink Box): These are external data sources providing additional data. Some examples are: Weather APIs (temperature, precip- itation, etc.) Some examples are: Weather APIs (temperature, precipitation, etc.) Traffic APIs (levels of congestion, road con- ditions) Event information (concerts, sports games, holidays) Geospatial data (land cover, population density)

Table 1. Survey of Deep Learning Models for Ride-Hailing Demand Forecasting

Reference	**Model Type**	**Key Features**	**Dataset/Focus Area**	**Key Findings/Contributions**
Xu et al. [10]	LSTM	Forecasts taxi demand by region using recent requests and relevant information.	New York City taxi requests, divided into small zones.	Effective LSTM model for zone-specific demand forecasting.
Chen et al. [11]	Ubernet (CNN)	Short-term forecasting using multivariate framework with temporal and spatial characteristics.	Online ride-hailing services demand.	Improved CNN for demand forecasting.
Ara & Hashemi [12]	CNN-BiLSTM	Combines CNN and BiLSTM for demand forecasting between neighborhood zones.	Ride-hailing and travel demand between city zones.	Accurate prediction for pickup-destination pairs.
Ye et al. [13]	LSTM + Attention	LSTM with attention mechanism using temporal, spatial, and weather features.	Online ride-hailing demand in different city areas.	Accurate demand forecasting through feature extraction and attention.
Liang et al. [14]	ST-MRGNN (Graph Neural Network)	Multirelational spatiotemporal graph neural network for multimodal demand forecasting.	Online car-hailing and subway demand.	Good demand forecasting performance for both ride-hailing and subway.
Wu et al. [15]	MVDSTN (Multiview Spatiotemporal Network)	Multiview deep spatiotemporal network for effective representation of spatiotemporal relationships.	Online ride-hailing demand.	Effective representation of spatiotemporal relationships for demand prediction.

A. Data Ingestion (Red Box): Data Ingestion (Top Red Box): It addresses the problem of collecting and processing the raw data.

B. Data Ingestion Pipeline (Yellow Box): This is the automated pipeline that is responsible for ingest- ing data from the varied sources. This could be: A data lake (massive amounts of unstructured data) A relational database (for structured data)

C. Data Processing Layer (Blue Box): Data Processing Layer (Top Blue Box): The job here is to

G. Temporal Features (Light Blue Box): Features based on timestamps (e.g., day of week, hour of day). Spatial Features (Dark Blue Box): Features derived from location data (e.g., point-to-point distances, areas). Derived Features (Light Green Box): Computed features from provided data (e.g., ride speed, surge multipliers). Spatial Features (Dark Blue Box): Features derived from location data (e.g., point-to-point distances, areas). Time Series Decomposition (Light Blue Box): Where appropriate, this is time-series data broken down into trend, seasonality, and residuals. This means converting categorical variables (e.g., text labels) into numeric codes that machine learning algorithms can utilize H. Model Evaluation M The models underwent testing with various performance metrics: Mean Absolute Error (MAE): This measures the average difference between predicted and actual fares. n

transform raw data into machine learning feature format. Data Processing Layer (Light Green Box): This is a redundant title.

MAE = 1 —y n i=1

— yˆi— (1)

D. Data Preprocessing Module (Darker Green Box): This is the core of the data processing layer. It performs a number of transformations:

E. Data Cleaning (Purple Box): Redundant title as cleaning has already been done in the ingestion step. Presumably indicates further cleaning with the objective of feature engineering.

F. Feature Extraction (Light Purple Box):

MAE: Represents the Mean Absolute Error. n: Represents the number of data points. ((summation): Indicates the sum of the absolute differences. i=1 to n: Specifies that the sum is taken from the first data point (i=1) to the nth data point. y¡sub¿i¡/sub¿: Represents the actual (observed) value for the i-th data point. yˆ¡sub¿i¡/sub¿ (y-hat): Represents the predicted value for the i-th data point. • Root Mean Square Error (RMSE): This metric penalizes larger prediction errors to evaluate accuracy. • R-Squared Score: This assesses how effectively the model accounts for fare variability. ,u

This includes making new features from raw data. It is also sub-divided into:

n RMSE = (yi n i=1

— yˆi)2 (2)

4.2 Data Analysis Techniques

To explore ride-hailing fare patterns in India, a range of statistical and machine learning techniques were utilized: De- scriptive Statistics: Important metrics like average fare, median price changes, and standard deviations were examined to gain insights into pricing trends. Time-Series Analysis: An analysis of trends was conducted to identify peak demand times, seasonal fare changes, and how festivals or holidays affect pricing. Geospatial Analysis: Heat-maps and city-wide ride distribution data were employed to showcase fare differences across various locations and urban areas. Correlation Analysis: The connections between fare prices and various influencing factors (such as distance, surge pricing, traffic, and weather) were investigated. Machine Learning Modeling: Predictive models, including linear regression, random forest, and deep learning techniques, were created to estimate ride fares based on past trends.

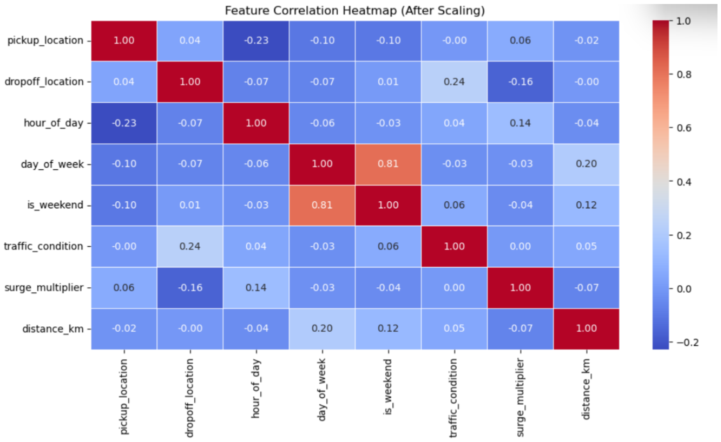

Fig. 2. Model Performance Metrics

5 Results and Findings

A. Descriptive Statistics and Data Insights The analysis of data collected from Ola, Uber, and other sources aimed to explore the variability of ride fares in India. Here are some key findings: • Average Fare Price: The average fare for all trips was Rupee 220, with a standard deviation of 110. • Peak Hour Impact: Fares were notably higher during the hours of 8:00 AM to 11:00 AM and 6:00 PM to • The Neural Network Model surpassed other models, demon- strating the smallest error margin and the greatest reliability for fare prediction. • Random Forest Regression offered a solid balance between accuracy and interpretability, positioning it as a strong alter- native. C. Feature Importance Analysis To identify the key factors influencing ride fares, a ranking of feature importance was performed: 9:00 PM, which correspond to the busy office commute Trip Distance (41%) – This is the most significant factor, as

times in major cities such as Delhi, Mumbai, and Bangalore. • Weather Influence: Inclement weather, including heavy rain and extreme heat, resulted in a fare increase of 25-40% due to changes in demand and slower traffic conditions.

- Distance- Fare Relationship: A strong correlation (r = 0.85) was found between the distance of the trip and in This fig 3 is a line

chart demonstrating the performance metrics (MAE, RMSE, and R^2 Score) of different machine learning models used for ride price prediction. Explanation of the Chart: X-Axis (Models): The x-axis is for various machine learning models: Linear Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Regression (SVR). Y-Axis (Metric Values): The y-axis is the values of the three performance measures: Mean Absolute Error (MAE) (orange solid line with circles) Root Mean Squared Error (RMSE) (orange dashed line with squares) R^2 Score (pink dash-dot line with triangles)

Table 2. Performance Metrics of Different Models

Model	**MAE**	**RMSE**	**R^2 Score**
Linear Regression	29.52	36.42	0.8655
Decision Tree	35.62	46.73	0.7785
Random Forest	32.50	39.86	0.8389
Gradient Boosting	28.00	37.30	0.8589
SVR	88.36	96.99	0.0462

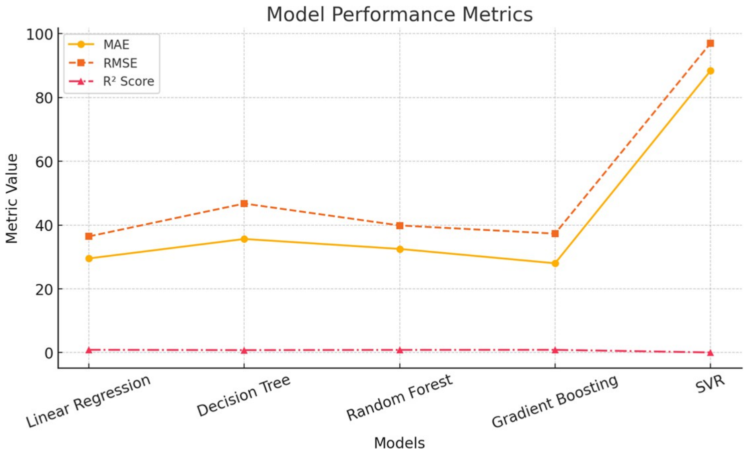

Fig. 3. Model Performance Metrics

5.1 Key Observations:

Linear Regression and Gradient Boosting both have the lowest errors (MAE RMSE), hence they are doing great. SVR performs very poorly with the highest errors (MAE: 88.36, RMSE: 96.99), and hence cannot be utilized in this work. R^2 Score is highest in case of Gradient Boosting (0.8589) and Linear Regression (0.8655), i.e., they explain the variance of the ride price data most. Decision Tree and Random Forest both have medium errors with lesser accuracy compared to Gradient Boosting. SVR has the worst R^2 Score (0.0462), indicating that it is not capturing important trends in the data.

6 Conclusion

The study, ”Fare Forecasting: A Machine Learning Ap- proach to Ride Price Prediction,” effectively showcased how machine learning techniques can accurately estimate ride fares. By utilizing extensive data collection, preprocessing, and model evaluation, this research underscored the potential of data-driven methods for fare prediction.Final thoughts This research adds to the growing field of predictive analytics in transportation, highlighting the effectiveness of machine learn- ing in forecasting fares. The results are beneficial for both ride- hailing companies and passengers, encouraging better pric- ing transparency and more informed choices. As data-driven modeling techniques advance, the precision, dependability, and efficiency of fare prediction systems will continue to enhance, influencing the future of smart transportation solutions.

References

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp. 68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. H. Shahi, “Predictive Pricing Model for Shared Economy Ride Applications: Incorporating Latest Data and Factors,” Communications in Computer and Information Science.
6. “Modeling Uber Data for Predicting Features Responsible for Price Fluctuations,” 2022 IEEE Delhi Section Conference (DELCON).
7. M. Shaik Amzad, B. Basha, “Harnessing Data-Driven Insights: Predictive Modeling for Diamond Price Forecasting using Regression and Classification Techniques,” IEEE, 2024, pp. 198–195.
8. S. Guo, C. Chen, J. Wang, Y. Liu, K. Xu, and D. M. Chiu, “Dynamic Price Prediction in Ride-on-demand Service with Multi-source Urban Data,” International Conference on Mobile and Ubiquitous Systems: Networking and Services, 2018, pp. 412–421. https://doi.org/10.1145/3286978.3286992.
9. V. Cerqueira, N. Moniz, and C. Soares, “VEST: Automatic Feature Engineering for Forecasting,” arXiv: Machine Learning, 2020. https://arxiv.org/abs/2010.07137.

10. J. Xu, R. Rahmatizadeh, L. Boloni, and D. Turgut, “Real-time prediction of taxi demand using recurrent neural networks,” IEEE Trans. Intell. Transp. Syst., vol. 19, pp. 2572–2581, 2017. [CrossRef].
11. L. Chen, P.V. Thakuriah, and K. Ampountolas, “Short-term prediction of demand for ride-hailing services: A deep learning approach,” J. Big Data Anal. Transp., vol. 3, pp. 175–195, 2021. [CrossRef].
12. Z. Ara and M. Hashemi, “Ride hailing service demand forecast by integrating convolutional and recurrent neural networks,” in Proceedings of the 33rd International Conference on Software Engineering and Knowledge Engineering, Pittsburgh, PA, USA, 1–10 July 2021.
13. X. Ye, Q. Ye, X. Yan, T. Wang, J. Chen, and S. Li, “Demand Forecasting of Online Car-Hailing with Combining LSTM+ Attention Approaches,” Electronics, vol. 10, p. 2480, 2021. [CrossRef].
14. Y. Liang, G. Huang, and Z. Zhao, “Joint demand prediction for multimodal systems: A multi-task multi-relational spatiotemporal graph neural network approach,” Transp. Res. Part C Emerg. Technol., vol. 140, p. 103731, 2022. [CrossRef].
15. Y. Wu, H. Zhang, C. Li, S. Tao, and F. Yang, “Urban ride-hailing demand prediction with multi-view information fusion deep learning framework,” Appl. Intell., pp. 1–19, 2022. [CrossRef].
16. “Real Time Prediction of Cab Fare Using Machine Learning,” 2022 International Conference on Electronics and Renewable Systems (ICEARS). https://doi.org/10.1109/icears53579.2022.9752315.
17. L. Chen, P. Thakuriah, and K. Ampountolas, “Short-Term Prediction of Demand for Ride-Hailing Services: A Deep Learning Approach,” 3(2), pp. 175–195, 2021. https://doi.org/10.1007/S42421-021-00041-4.