International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 01 | Jan – 2026 DOI: 10.55041/ISJEM05374
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Taxi Fare and Ride Demand Prediction Using Machine Learning Techniques

D. NANDHINI., MCA
(Assistant Professor, Master of Computer Applications)
S. VASANTHAN., MCA
Christ College of Engineering and Technology
Moolakulam, Oulgaret Municipality, Puducherry – 605010.

***Abstract*—**Urban transportation systems generate vast amounts of trip-related data through taxi and ride-hailing services. Accurate prediction of taxi fares and ride demand is essential for improving passenger satisfaction, optimizing driver allocation, and enhancing operational efficiency. Traditional fare estimation methods rely on fixed tariff rules and simple distance–time calculations, which fail to capture complex real-world factors such as traffic congestion, travel patterns, and temporal demand variations.

This paper proposes an intelligent Taxi Fare and Ride Prediction System using machine learning techniques to provide accurate and data-driven fare estimation and ride demand analysis. The system analyzes historical taxi trip data and key influencing features such as pickup and drop-off locations, trip distance, travel time, passenger count, traffic conditions, weather factors, and surge pricing. Data preprocessing techniques including data cleaning, normalization, and feature selection are applied to improve prediction accuracy. Supervised machine learning algorithms such as Linear Regression, Decision Tree Regression, Random Forest, and Gradient Boosting models are trained and evaluated using standard performance metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score.

The proposed system demonstrates reliable prediction performance and efficient processing, making it suitable for real-world transportation analytics. The results confirm that machine learning-based models significantly outperform traditional rule-based approaches and provide a scalable solution for intelligent transportation systems.

Keywords—Taxi Fare Prediction, Ride Demand Prediction, Machine Learning, Regression Models, Intelligent Transportation Systems, Predictive Analytics.

I. INTRODUCTION

Urban mobility has undergone a significant transformation with the rapid growth of taxi and ride-hailing platforms. These services produce massive volumes of trip-related data, offering valuable opportunities for data-driven analysis and intelligent decision-making in transportation systems [1], [2]. Accurate estimation of taxi fares and ride demand plays a critical role in ensuring transparency for passengers, effective pricing strategies for service providers, and optimized resource utilization for drivers.

Conventional taxi fare calculation methods are primarily based on predefined tariff structures, distance traveled, and waiting time. While these approaches are simple to implement, they do not account for dynamic factors such as traffic congestion, peak-hour demand, weather conditions, and location-specific travel patterns [3]. As a result, traditional systems often produce inaccurate fare estimates, leading to customer dissatisfaction and inefficient operational planning.

Machine learning (ML) offers a powerful alternative by enabling systems to learn complex patterns from historical data and generate accurate predictions for unseen scenarios [4]. By analyzing large-scale taxi trip datasets containing spatial, temporal, and behavioral attributes, ML models can capture nonlinear relationships between influencing factors and fare outcomes [5].

Recent studies have shown that regression-based and ensemble learning techniques are particularly effective in modeling transportation-related prediction problems [6], [7].

This research presents a machine learning–based Taxi Fare and Ride Prediction System designed to overcome the limitations of traditional rule-based methods. The system integrates structured data preprocessing, feature engineering, and supervised learning algorithms to accurately estimate taxi fares and analyze ride demand trends. The proposed approach aims to support intelligent transportation planning and contribute to the advancement of smart mobility solutions.

II. RELATED WORK

Enhancements to the proposed system may include the integration of Several studies have investigated the use of machine learning algorithms for predictions related to taxi fare estimation as well as ride demands. In the early days, research on predictions has involved statistical models based on linear regression with the aim of estimating the fare based on the distance and time variables [8].

Later works have integrated decision tree-based approaches and ensemble methods, including Random Forest and Gradient Boost, for enhanced accuracy of predictions in [9], [10]. They show excellent performance for capturing interaction between features and coping with potential issues of overfitting or underfitting. Research on time series and demand predictions has further utilized ride sharing data for estimating demands for ride sharing at different locations and time ranges in [11].

Deep learning methods based on neural networks or recurrent neural architectures have also been explored for large-scale ride demand forecast tasks [12]. Despite their strong performance capabilities, these models also require significant computational capabilities as well as large amounts of data, making them unsuitable for small-scale or academic applications. However, there exists a requirement for an integrated, efficient, and scalable system for performing both fare estimation tasks and ride prediction by leveraging machine learning algorithms. This proposed system serves as an answer for such a requirement by being an effective platform for integrated evaluation and visualization.

III. PROPOSED METHODOLOGY

The Taxi Fare and Ride Prediction System is a structured process for machine learning that aims to improve the predictions made from historical data on taxi rides by providing accurate estimates for fare as well as ride demands. These processes are a series of steps such as data collection, processing, variable selection, model development, prediction, and assessment.

A. Data Collection

The dataset used for this study consists of historical taxi ride records collected from publicly available transportation datasets [13]. Each record includes essential trip-related attributes such as pickup location, drop-off location, trip distance, travel time, passenger count, fare amount, traffic level, weather conditions, and surge pricing indicators. These attributes represent both spatial and temporal factors that significantly influence taxi fare and ride behavior.

The collected dataset serves as the foundation for training and evaluating machine learning models. Data integrity checks are performed to ensure consistency and completeness before further processing.

B. Data Preprocessing

Raw taxi datasets often contain missing values, duplicate records, and inconsistent formats, which can negatively impact model accuracy. To address these issues, comprehensive data preprocessing techniques are applied [14].

The preprocessing stage includes:

- Removal of duplicate and invalid records
- Handling missing values using statistical imputation methods
- Encoding categorical attributes such as locations and time categories
- Normalization and scaling of numerical features such as distance and duration
- These steps ensure that the dataset is clean, standardized, and suitable for machine learning model training.

C. Feature Selection and Engineering

Feature selection plays a crucial role in improving prediction accuracy and reducing computational complexity. Relevant features that directly influence taxi fares and ride demand are selected based on correlation analysis and domain knowledge [15].

Key features used in the system include:

- Trip distance
- Travel time
- Pickup and drop-off locations
- Passenger count
- Traffic conditions
- Weather indicators
- Surge multiplier

Feature engineering techniques are applied to extract meaningful representations from raw data, enabling models to better capture hidden relationships within the dataset.

D. Machine Learning Models

The system employs supervised regression-based machine learning models to predict taxi fares and ride demand. Multiple algorithms are implemented and compared to identify the most effective prediction approach [16].

The primary models used include:

Linear Regression:

Linear Regression is used as a baseline model due to its simplicity and interpretability. It models the linear relationship between input features and fare values [17].

Decision Tree Regression:

Decision Tree Regression captures nonlinear relationships by recursively splitting data based on feature values. This model improves prediction accuracy over linear models by handling complex interactions [18].

Random Forest Regression:

Random Forest is an ensemble learning technique that combines multiple decision trees to produce robust predictions. It reduces overfitting and improves generalization performance [19].

Gradient Boosting Models:

Gradient Boosting algorithms iteratively improve prediction accuracy by correcting errors made by previous models. These models are effective in handling complex datasets with high variance [20].

E. Prediction and Evaluation

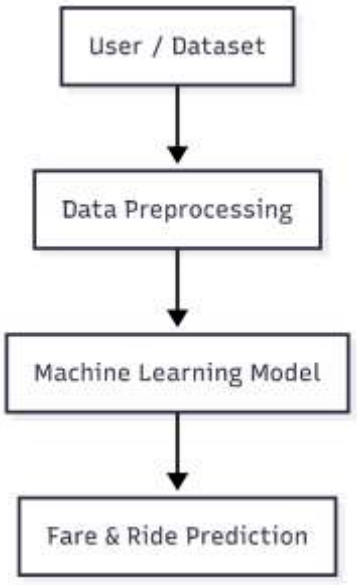
After training, the models are used to predict taxi fares and ride demand for unseen data. Model performance is evaluated using standard regression metrics [21]:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- R^2 Score

These metrics provide quantitative measures of prediction accuracy and model reliability. The best-performing model is selected based on overall evaluation results and computational efficiency.

IV. SYSTEM ARCHITECTURE

The proposed Taxi Fare and Ride Prediction System is designed using a modular and layered architecture to ensure clarity, scalability, and ease of maintenance. The architecture follows a data-driven workflow that transforms raw historical taxi trip data into accurate fare and ride demand predictions through a sequence of well-defined components.



At a high level, the system architecture consists of five major layers: Data Input Layer, Preprocessing Layer, Machine Learning Layer, Storage Layer, and Output & Visualization Layer. Each layer performs a specific function and interacts with adjacent layers through structured data flow.

V. IMPLEMENTATION

The proposed Taxi Fare and Ride Prediction System is implemented using Python and standard machine learning libraries to ensure efficiency and reliability. Historical taxi trip data is processed using Pandas and NumPy for data cleaning, normalization, and feature transformation. Supervised regression models such as Linear Regression, Decision Tree Regression, Random Forest, and Gradient Boosting are trained using the Scikit-learn framework. The trained models are then utilized to predict taxi fares and ride demand for unseen data. Model performance is evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score, while visualization tools such as Matplotlib are used to present prediction results and error analysis. The modular implementation ensures fast execution, easy maintenance, and scalability for future enhancements.

VI. RESULTS AND DISCUSSION

The performance of the proposed Taxi Fare and Ride Prediction System was evaluated using historical taxi trip datasets after applying preprocessing and feature selection techniques. The dataset was divided into training and testing sets to assess model generalization capability. Multiple regression-based machine learning models were trained and compared to identify the most effective approach for fare and ride prediction.

The experimental results indicate that Linear Regression provides a simple and interpretable baseline but shows limited performance when handling nonlinear relationships between trip attributes and fare values. Decision Tree Regression improves prediction accuracy by capturing nonlinear patterns; however, it is prone to overfitting when used independently. Ensemble-based models such as Random Forest and Gradient Boosting demonstrate superior performance by reducing variance and improving generalization.

Evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score confirm that ensemble models achieve lower prediction error and higher accuracy compared to traditional models. Features such as trip distance, travel time, traffic level, and surge multiplier were observed to have a significant impact on fare prediction. The results validate that machine learning–based approaches effectively model complex fare patterns and outperform conventional rule-based estimation methods. Overall, the system provides reliable, consistent, and accurate predictions suitable for transportation analytics and decision support.

VII. CONCLUSION

This paper presented an intelligent Taxi Fare and Ride Prediction System using machine learning techniques to estimate taxi fares and analyze ride demand based on historical transportation data. The proposed system overcomes the limitations of traditional rule-based pricing models by learning complex relationships between multiple influencing factors such as distance, time, location, traffic conditions, and passenger count.

Through effective data preprocessing, feature selection, and supervised learning algorithms, the system achieves accurate and efficient predictions. Experimental results demonstrate that ensemble machine learning models significantly improve prediction accuracy and reliability. The modular design and lightweight implementation make the system suitable for academic research as well as real-world transportation

applications. The study confirms that machine learning provides a scalable and data-driven solution for intelligent fare estimation and ride analysis.

VIII. FUTURE WORK

Future enhancements to the proposed system may include the integration of real-time traffic data, weather conditions, and GPS-based routing information to enable dynamic fare prediction. Advanced machine learning and deep learning models such as Neural Networks, LSTM, and hybrid ensemble techniques can be explored to further improve accuracy on large-scale datasets. Deployment of the system as a web-based or mobile application, along with cloud integration, would enhance accessibility and scalability. Additionally, extending the system to support multi-city datasets and real-time ride-hailing platforms can broaden its practical applicability.

IX. REFERENCES

[1] T. M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.

[2] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, Springer, 2013.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[4] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, OTexts, 2018.

[5] M. Rahman and M. Hasan, "Predicting Taxi Fare Using Machine Learning Algorithms," *Int. J. Computer Applications*, 2019.

[6] L. Moreira-Matias et al., "Predicting Taxi–Passenger Demand Using Streaming Data," *IEEE Trans. Intelligent Transportation Systems*, 2013.

[7] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *JMLR*, 2011.

[8] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, 1986.

[9] L. Breiman, "Random Forests," *Machine Learning*, 2001.

[10] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *KDD*, 2016.

[11] Y. Wang et al., "Deep Learning for Ride Demand Prediction," *IEEE Access*, 2019.

[12] H. Yin et al., "Urban Traffic Prediction Using Big Data," *IEEE Trans. Big Data*, 2017.

[13] Kaggle Contributors, "Taxi Fare Prediction Datasets," Kaggle, 2024.

[14] W. McKinney, *Python for Data Analysis*, O'Reilly, 2017.

[15] J. VanderPlas, *Python Data Science Handbook*, O'Reilly, 2016.

[16] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

[17] A. Géron, *Hands-On Machine Learning*, O'Reilly, 2019.

[18] S. Raschka and V. Mirjalili, *Python Machine Learning*, Packt, 2017.

[19] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning*, Cambridge Univ. Press, 2014.

[20] IEEE ITS Society, "Intelligent Transportation Systems Overview," IEEE, 2022.

[21] Python Software Foundation, "Python Documentation," 2024.

[22] Scikit-learn Documentation, https://scikit-learn.org