TAXI DEMAND PREDICTION

[1] T Manasa, [2] Telugu Millikarjun, [3] Seepathi Mahalaxmi, [4] Vempati Rohith Charan
Department of Computer Engineering(Software Engineering)
Siddhartha Institute of Technology & Sciences, Narapally
thirumanasa@siddhartha.org.in, 23tq1a5660@siddhartha.co.in, 23tq1a5610@siddhartha.co.in, 23tq1a5653@siddhartha.co.in,

Abstract

Predictive analytics plays a significant role in forecasting future events by analyzing historical data and identifying meaningful patterns. In recent years, the growth of advanced technologies in Machine Learning and Big Data has greatly enhanced the accuracy and efficiency of predictive models. These techniques are widely used across various industries to make informed decisions and optimize resource planning.

One such important application is taxi fare prediction, which helps estimate the cost of a ride based on several influencing factors. The fare of a taxi ride is not fixed and depends on multiple variables such as distance traveled, time taken, traffic conditions, pickup and drop locations, weather conditions, and time of the day. Accurately predicting taxi fares can benefit both customers and service providers by ensuring transparency, better pricing strategies, and improved customer satisfaction.

This project focuses on developing a machine learning-based model to predict taxi fares within a city. The system uses historical trip data to analyze patterns and relationships between different features affecting the fare. Various preprocessing techniques are applied to clean and prepare the dataset, followed by feature engineering to extract meaningful insights. Machine learning algorithms such as Linear Regression, Decision Tree, and Random Forest are implemented to build predictive models.

The performance of the models is evaluated using appropriate metrics, and the best-performing model is selected for fare prediction. The project also includes data visualization techniques to better understand trends and patterns in taxi rides. Overall, this work demonstrates how predictive analytics and machine learning can be effectively used to estimate taxi fares accurately, making transportation systems more efficient and user-friendly.

I. Introduction

Predicting taxi fares is a challenging task due to the dynamic nature of urban transportation systems. The cost of a taxi ride is influenced by multiple factors such as distance traveled, travel time, traffic conditions, weather, demand fluctuations, and pickup and drop-off locations. These factors are highly variable and often unpredictable, making it difficult for service providers to estimate fares accurately in advance. Without proper prediction mechanisms, customers may face uncertainty in pricing, while taxi companies may struggle with inefficient pricing strategies and revenue management.
Currently, many taxi services rely on basic fare calculation methods that consider only distance and time, without fully accounting for real-world complexities such as

traffic congestion or peak-hour demand. This can result in inconsistent pricing, reduced customer satisfaction, and operational inefficiencies.
The main problem addressed in this project is to develop a predictive model that can accurately estimate taxi fares using historical trip data. By applying machine learning techniques to features such as trip distance, duration, pickup time, passenger count, and location data, the system aims to identify patterns that influence fare pricing. Algorithms such as Linear Regression, Decision Tree, and Random Forest can be used to build models that predict the expected fare for a given trip.
The goal of this project is to create an efficient and reliable fare prediction system that enhances pricing transparency, improves customer experience, and supports better decision-making for taxi service providers.

II. Literature Survey

Taxi demand prediction has become a crucial research area in intelligent transportation systems due to its importance in optimizing fleet management, reducing passenger waiting time, and improving urban mobility. Early studies primarily focused on traditional statistical approaches such as moving averages, linear regression, and time series models. These methods were simple and computationally efficient but had limitations in capturing the complex and dynamic nature of taxi demand, especially in large urban environments.

To address these limitations, researchers began adopting machine learning techniques such as Decision Trees, Random Forest, and Support Vector Machines. These models demonstrated improved performance by capturing nonlinear relationships between input variables such as time, location, and weather conditions. Comparative studies have shown that regression-based machine learning models and ensemble techniques can achieve high prediction accuracy, with some models reaching around 90% performance depending on the dataset and configuration.

With the growth of large-scale datasets and urban mobility data, advanced algorithms such as XGBoost and Gradient Boosting have been widely applied. These models are capable of handling high-dimensional data and improving prediction accuracy through ensemble learning. Studies using these techniques reported significant reductions in prediction error and improved performance compared to traditional methods.

Recent research has shifted toward deep learning approaches to better capture spatiotemporal dependencies in taxi demand data. Models such as Artificial Neural Networks (ANN), Long Short-Term Memory (LSTM), Recurrent Neural Networks (RNN), and Graph Neural Networks (GNN) have shown superior performance in modeling both temporal trends and spatial relationships.

III. System Analysis

The system focuses on predicting taxi demand by analyzing various influencing factors such as time, location, weather conditions, traffic patterns, and historical trip data. Taxi demand is highly dynamic and varies across different regions and time intervals, making accurate prediction a complex task. The system requires processing large-scale spatiotemporal data to capture both temporal trends and spatial

dependencies. Data preprocessing is essential to handle missing values, noise, and inconsistencies in the dataset. The analysis also emphasizes the importance of feature selection, such as peak hours, holidays, and weather conditions, which significantly impact demand. The system must support scalability to handle real-time data and large urban datasets. Additionally, model selection and evaluation are critical to ensure prediction accuracy. Overall, the system aims to provide an intelligent and data-driven solution to optimize taxi allocation and improve transportation efficiency.

Existing System

The existing systems for taxi demand prediction mainly rely on traditional statistical methods such as historical averaging, linear regression, and time series models like ARIMA. These approaches use past demand patterns to estimate future demand. While they are simple and easy to implement, they assume linear relationships and fail to capture the complex spatiotemporal dynamics of taxi demand. The existing systems often do not consider multiple influencing factors such as weather conditions, traffic congestion, or special events. They also struggle to handle large-scale datasets and real-time data streams efficiently. As a result, these models provide limited accuracy, especially during peak hours or unexpected demand fluctuations. The reliance on static models makes them less adaptable to changing urban environments.

Disadvantages of Existing System

- Inability to capture complex spatiotemporal relationships
- Limited accuracy during peak hours and sudden demand changes
- Assumes linear relationships between variables
- Poor adaptability to real-time data updates
- Inefficient handling of large-scale datasets
- Ignores important external factors like weather and events

Proposed System

The proposed system introduces a machine learning and deep learning-based approach to accurately predict taxi demand. It utilizes advanced algorithms such as Random Forest, Gradient Boosting, and Long Short-Term Memory (LSTM) networks to model complex relationships in the data. The system begins with collecting historical trip data along with external factors such as weather, time, and location. Data preprocessing techniques are applied to clean and transform the dataset, followed by feature engineering to extract meaningful patterns. The model is trained using both spatial and temporal features to capture demand variations across different regions and time periods. Evaluation metrics such as MAE, RMSE, and R^2 score are used to assess performance.

Advantages of Proposed System

- High prediction accuracy using advanced ML/DL models
- Captures both spatial and temporal dependencies
- Adaptable to real-time and dynamic data
- Efficient handling of large-scale datasets
- Considers multiple influencing factors (weather, traffic, events)

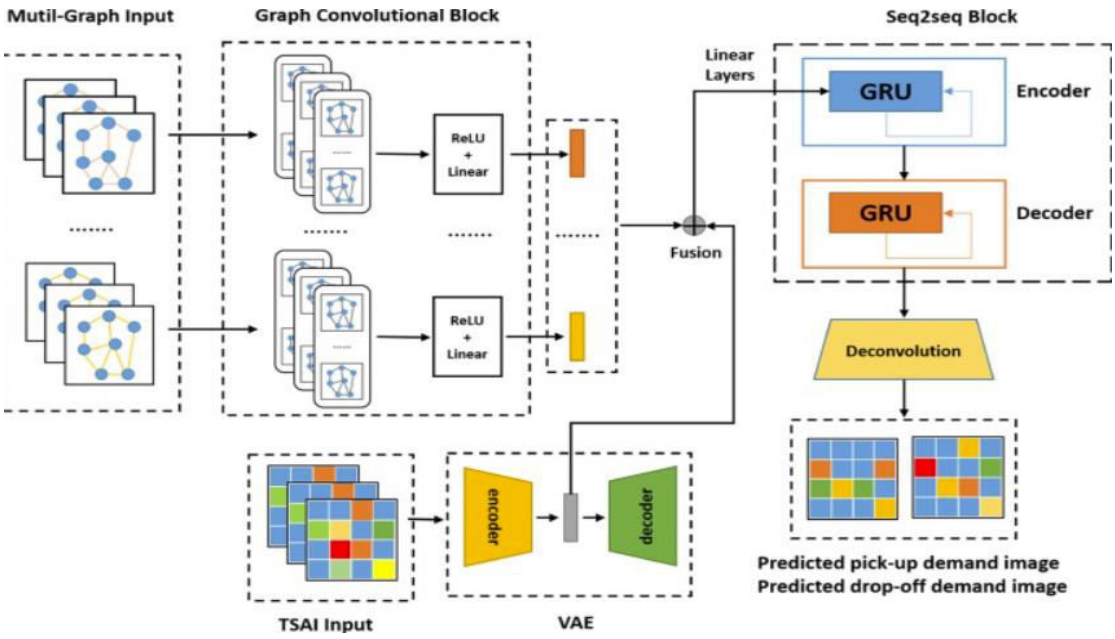
IV. Methodology

The proposed system for taxi demand prediction follows a structured machine learning and deep learning pipeline. Initially, historical taxi trip data is collected along with external factors such as date, time, location, weather conditions, traffic patterns, and special events. This data serves as the foundation for understanding demand patterns across different regions and time intervals.

The next step involves data preprocessing, where missing values are handled, outliers are removed, and the data is transformed into a suitable format. Categorical variables such as location zones and weather conditions are encoded, and feature scaling is applied to improve model performance.

Following preprocessing, exploratory data analysis (EDA) is conducted to identify trends such as peak hours, high-demand locations, and seasonal variations. This helps in understanding the underlying patterns in taxi demand.

System Architecture

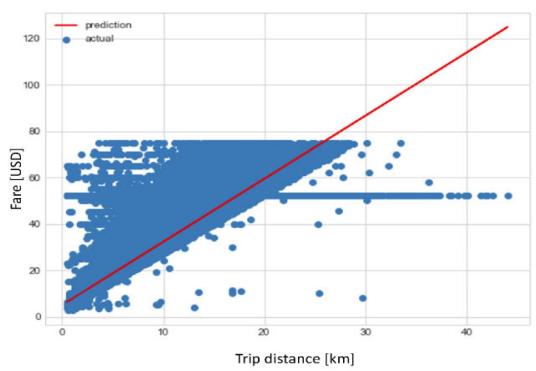


The system architecture for taxi demand prediction is designed as a comprehensive machine learning pipeline that processes large-scale spatiotemporal data to generate accurate demand forecasts. The process begins with data collection from multiple sources, including historical taxi trip records, weather data, time information, and location-based data. This data is then passed to the preprocessing layer, where missing values are handled, noise is removed, and categorical variables such as locations and weather conditions are encoded into numerical formats.

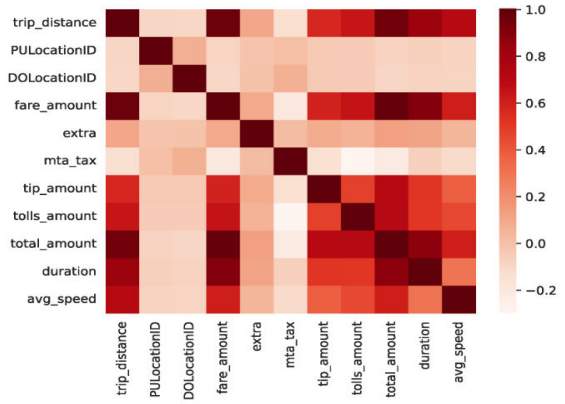
After preprocessing, the data undergoes exploratory data analysis to identify patterns such as peak demand hours, high-traffic zones, and seasonal trends. The feature engineering module extracts meaningful features, including time-based attributes (hour, day, weekend), spatial information (pickup and drop locations), and external

factors (weather, holidays). These refined features are then fed into the model training layer, where advanced machine learning and deep learning models such as Random Forest, Gradient Boosting, and Long Short-Term Memory (LSTM) networks are used to learn complex temporal and spatial relationships.

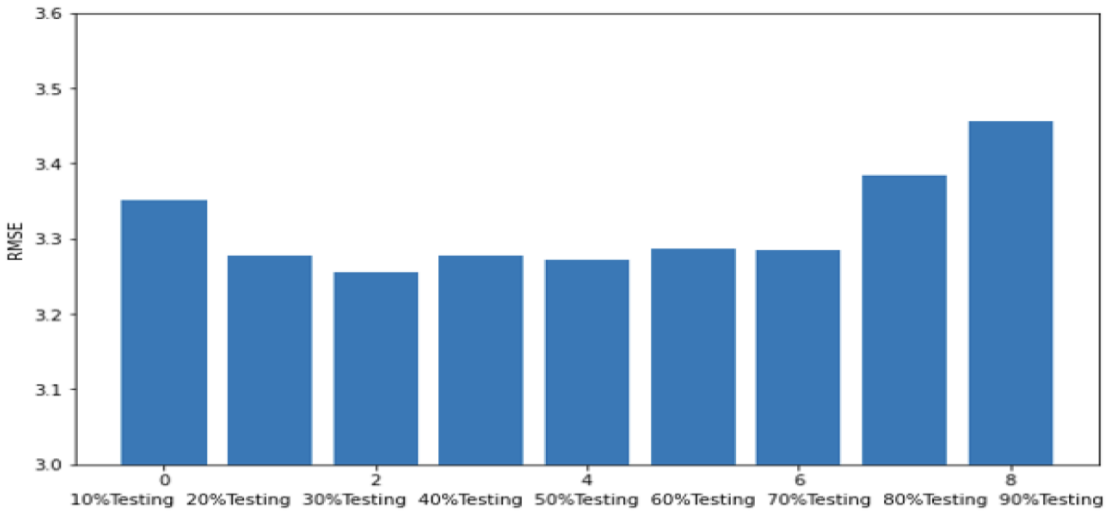
V. Result and Output

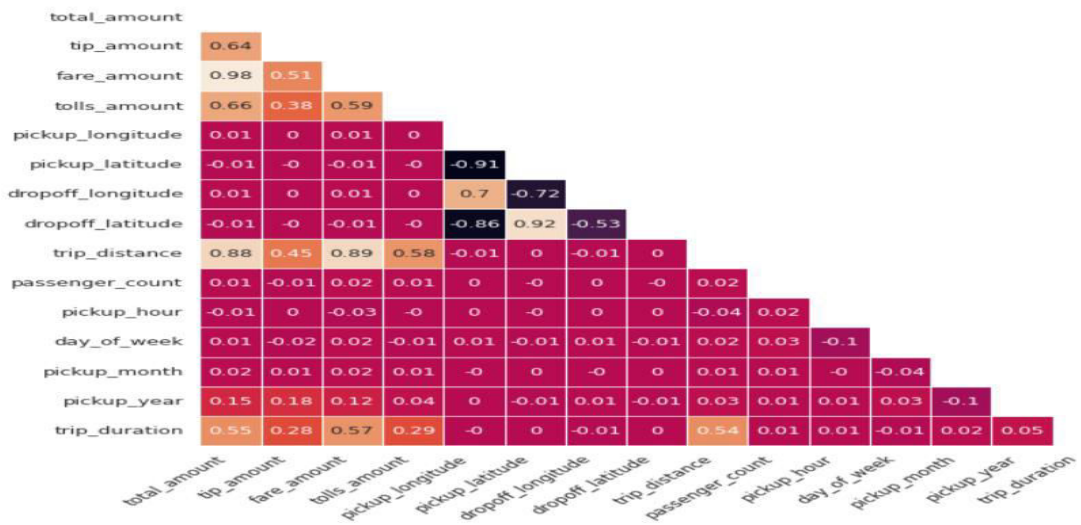


(a)



(b)





VI. Conclusion

In conclusion, the Student Performance Prediction project successfully demonstrates the potential of machine learning in analyzing academic and socio-economic data to generate meaningful and actionable insights. By applying models such as Linear Regression and Random Forest, the system effectively identifies patterns that influence student success, achieving reliable prediction accuracy. The study highlights that previous academic performance (G1, G2) is the most significant factor, while elements like study time, absences, and social conditions also play an important supporting role.

The project emphasizes the importance of transitioning from traditional reactive approaches to proactive, data-driven strategies in education. It provides educators with an early warning system to identify students who may require additional support, enables students to understand and improve their performance, and assists institutions in enhancing overall academic outcomes and graduation rates.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, “Real-Time Object Detection in Drone Surveillance Using YOLOv5,” in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, “Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks,” in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S.Manaswini, “Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology,” in

Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, “An advanced movie recommender using collaborative filtering and sentiment analysis,” International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, “Automatic crop recommendation system using LightGBM and decision tree machine learning models,” Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.

[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, “Smart agriculture through IoT and machine learning for analyzing carbon footprints,” in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.

[11] Ravi Kumar Banoth, B. V. Ramana Murthy, “Soil image classification using transfer learning approach: MobileNetV2 with CNN,” SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.