International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 1 (2026)



ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 1, 2026
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research

DEEP LEARNING–DRIVEN RIDE-HAILING PRICE PREDICTION FOR INTELLIGENT TRANSPORTATION SYSTEMS

DR. D. MURALI[1], ASARELLI SHAILA[2], BOLLENI BHOOMIKA[3], ERRI RAKESH[4], ADUPA VIKAS[5], DASARI ARAVIND YADAV[6]

PROFESSOR[1], UG SCHOLAR[2,3,4,5&6]

DEPARTMENT OF CSE, NARSIMHA REDDY ENGINEERING COLLEGE (UGC- AUTONOMOUS) MAISAMMAGUDA (V), KOMPALLY, SECUNDERABAD, TELANGANA-500100

ABSTRACT

Accurate forecasting of booking value and short-term ride demand is increasingly important for enhancing operational efficiency in urban ride-hailing systems. However, existing forecasting pipelines suffer from inconsistent model performance and difficulty in adapting to sudden behavioural shifts in urban mobility patterns. Additionally, most current approaches struggle to generalize across different regions and time intervals, leading to unstable and unreliable demand predictions. Conventional machine learning approaches such as **Linear Regression, Random Forest**, and **Gradient Boosting** have been widely applied to predict ride demand; however, their performance is often constrained by their limited ability to model temporal dependencies and rapidly evolving urban mobility patterns. Linear Regression fails to capture complex nonlinear relationships, Random Forest struggles with sequential correlations, and Gradient Boosting requires significant computational effort when handling high-dimensional and dynamic datasets. To overcome these limitations, this work proposes an enhanced forecasting framework based on a **Recurrent Neural Network (RNN)** integrated with a **Greedy Tree–based optimization algorithm**. The RNN component effectively models time-dependent fluctuations in booking value and ride demand, while the Greedy Tree module optimizes feature selection, reduces redundancy, and improves model interpretability. Comparative evaluation across multiple real-world urban analytics datasets demonstrates that the proposed hybrid architecture significantly outperforms existing traditional models in prediction accuracy, responsiveness to sudden demand variations, and computational efficiency. The results highlight the potential of combining deep sequential learning with greedy optimization to provide more robust, scalable, and actionable insights for intelligent transportation systems and urban mobility planning.

1.INTRODUCTION

1.1 OVERVIEW

The current road infrastructure is insufficient due to the fast-growing population and the corresponding growth in individual vehicle ownership. As a result, transportation systems should be developed to be comfortable and safe. In this environment, it is critical to forecast current and future traffic growth and develop an appropriate transportation strategy. Transportation planning is a process that generates data that will aid decision makers in the construction of future transportation networks. It can also be defined as a tool or method for a sustainable planning approach that analyses and evaluates the current state of urban transportation infrastructure; determines future (target year and/or years) investments, regulations, and operational approaches; and generates predictions. Transportation planning should be carried out by relating them from the largest to the smallest towns across the country or from the macro- to the micro-level.

As the number of motor vehicles increased in the early 20th century, researchers conducted the first residential and roadside pedestrian surveys in the 1920s and 1930s to help control congested streets and crossings. They utilized this knowledge to develop street widening and one-way street designs. In the early 1940s, individuals in the United States thought that transportation planning was limited to technical solutions to crowded junctions. However, by the 1950s, it had become regarded as a systematic study that considered not only intersection designs but also changes in urban transportation plans, population growth, and vehicle ownership. In the 1970s, a process that also addressed technical and political issues for transportation planning in cities was observed to begin.

In the 1990s, the concept of sustainability for transportation emerged, aiming to ensure the efficient transportation of people, goods (things), and services and to leave a less damaged environmental and cultural heritage for the future. Although the first transportation planning models only considered cost-effectiveness, later studies focused on economics, land use, and the management of transportation systems. Instead of temporary and costly designs, transportation planners must prepare for the future in terms of delivering both economic and

long-term services, as well as resolving potential bad events in a predictable manner. Recent research has resulted in methods for addressing traffic issues in future cities. These transportation planning studies strive to develop a sustainable and successful transportation network. Figure 1.1 illustrates the long-term growth trend of registered motor vehicles in India from 1951 to 2020, highlighting a dramatic and consistent rise in vehicle ownership over seven decades. In 1951, India had **approximately 0.3 million** registered vehicles, reflecting limited motorization and low household purchasing power. By 1980, this number had increased to **8.2 million**, marking the beginning of rapid mobility expansion driven by urbanization and industrial growth. The real surge began in the 1990s, coinciding with economic liberalization, when registered vehicles crossed **21 million** in 1991 and reached **55 million** by 2001. This steep upward trajectory indicates significant lifestyle changes, the rise of middle-class income, and increased reliance on private transport. The most substantial escalation occurred after 2000, reflecting accelerated economic development and improved affordability of two-wheelers and cars. By 2010, the total number of registered vehicles had crossed **159 million**, more than doubling within a decade. The growth continued sharply, and by 2020, India recorded approximately **326 million** registered motor vehicles, marking nearly a thousand-fold increase compared to 1951. The figure clearly shows that two-wheelers form the largest share, followed by cars and utility vehicles, explaining why urban congestion, fuel consumption, and environmental stress have intensified.

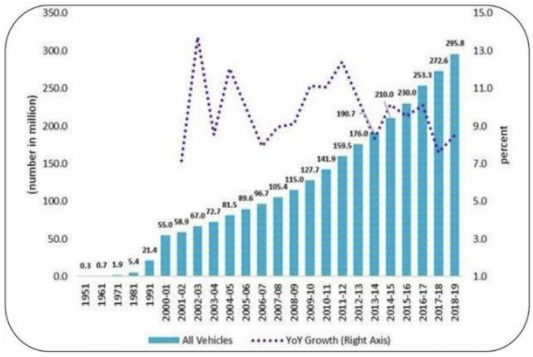


Figure 1.1. Registered Motor Vehicles in India (1951–2020)

1.2 RESEARCH MOTIVATION

Figure 1.2 represents the complete workflow and motivation behind the need for improved ride-demand forecasting and booking value prediction in modern mobility systems. The flow begins with the *Public*, who depend on ride-sharing applications for daily travel requirements. Every time a user opens an app and submits a ride request, the information is transmitted directly to the ride-sharing platform. This initial interaction is simple for the user, but it triggers a complex chain of data processing and decision-making in the backend.

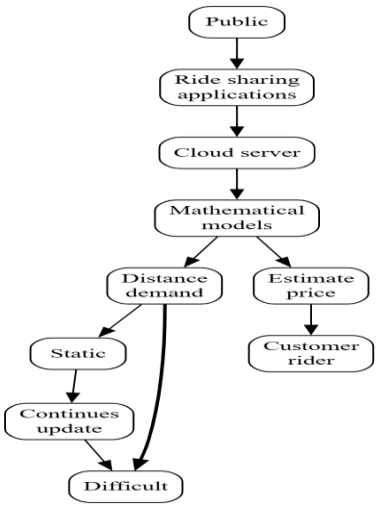


Figure 1.2 Research Motivation Device System

Once a ride request is made in the Ride Sharing Application, the system collects key details such as the user's pickup location, destination, request time, past travel patterns, and surge pricing information. These details are sent to the Cloud Server, which acts as the main processing unit of the ride-hailing system. The server collects millions of ride requests from across the city and processes them in real time. Inside the server, multiple operations happen at the same time, including storing data, extracting important features, retrieving demand patterns, and running prediction models.

The system uses standard Mathematical Models to calculate two main results: distance demand and estimated price. Distance demand estimates how far a ride will travel and how many rides are being requested in the same area. Ideally, this should adjust based on traffic, road closures, peak hours, weather, and local events. However, many current systems rely on simplified or averaged data, which can make real-time demand predictions less accurate, especially in fast-changing urban conditions. Similarly, estimated price is calculated using basic cost formulas, fixed parameters, fuel rates, or set surge rules. This usually works well under normal conditions but can become unreliable during heavy traffic, sudden demand spikes, or unexpected route changes, causing fluctuating prices, inconsistent ride availability, and unreliable estimated arrival times for riders. The left side of the system highlights deeper operational problems. Distance demand is fed into a Static module, which uses fixed values like base fare, predefined city zones, historical average speeds, and unchanging cost factors. These static values do not adjust in real time, creating bottlenecks. Even when the system tries to refresh values through

a Continuous Update cycle, it still relies on the same limited data, causing repeated calculations, delayed updates, and inaccurate results, especially during unpredictable events like bad weather, festivals, or accidents.

1.3 PROBLEM STATEMENT

Traditional ride-hailing platforms rely heavily on manually engineered rules, static mathematical formulas, and fixed operational parameters to estimate booking values and forecast ride demand. These systems use predetermined distance–fare relationships, historical averages, zoning-based pricing rules, and step-wise tariff structures that do not adapt to the rapidly changing travel environment. Because these calculations depend on human-defined assumptions, they fail to reflect real-time variations such as fluctuating traffic speeds, sudden surges in demand, shifts in public events, or unpredictable customer behavior. As a result, the system often produces inaccurate distance demand estimates, inconsistent fare calculations, and delayed operational responses. The reliance on static parameters also limits the system's ability to integrate new data, causing outdated or oversimplified models to dominate decision-making. Furthermore, manual rule-mining and mathematically rigid systems require continuous human intervention to update parameters, revise fare rules, and adjust operational thresholds. This creates a maintenance burden and increases the risk of errors, because rules become outdated quickly in a dynamic urban mobility context. Continuous updates depend on time-consuming recalibration rather than automatic adaptation, which leads to slow reactions during peak hours and congestion periods. The lack of responsiveness creates challenges such as inefficient driver allocation, poor prediction of high-demand zones, difficulty in ensuring fare fairness, and reduced customer satisfaction. These limitations highlight the need for more adaptive and intelligent forecasting systems that overcome the rigidity and inefficiencies inherent in manual, rule-driven structures.

1.4 Research Objective

To discuss about AI procedures for enhancing short-term ride-hailing demand forecasting, this research aims to develop a hybrid predictive framework that combines a **Recurrent NeuralNetwork (RNN)** with a **Greedy Tree (GT)**–based optimization algorithm. The objective is to leverage the RNN's capability to capture temporal fluctuations in booking value and ride demand, while the Greedy Tree module optimizes feature selection, reduces redundancy, and improves interpretability. The proposed **RNN-GT** model is systematically evaluated against conventional approaches such as **Linear Regression (LR), Random Forest (RF)**, and **Gradient Boosting Machines (GBM)**to demonstrate superior performance in prediction accuracy, responsiveness to sudden demand changes, and computational efficiency, providing a scalable and robust AI-driven solution for intelligent urban transportation planning.

1.5 Advantages

- Accurate short-term demand forecasting allows ride-hailing platforms to optimize driver allocation, reducing idle time and ensuring better service availability.
- The hybrid RNN-Greedy Tree model captures temporal dependencies in ride requests, improving prediction of peak hours and sudden demand surges.
- Feature selection through the Greedy Tree algorithm reduces redundancy, leading to faster training and inference times without sacrificing accuracy.
- The system can adapt to rapidly changing urban mobility patterns, such as traffic disruptions or public events, maintaining reliable predictions.
- By modelling nonlinear relationships in data, the RNN component outperforms traditional linear models in capturing complex ride demand trends.
- The hybrid framework improves interpretability, helping analysts understand which factors most influence booking value and ride demand.
- Computational efficiency is enhanced compared to conventional ensemble methods, making the solution scalable for multiple cities or regions.
- The model's robustness reduces errors in decision-making, supporting better operational planning and resource management for ride-hailing companies.
- By anticipating demand, the platform can implement dynamic pricing more effectively, balancing supply and demand while maximizing revenue.
- Integration of deep learning with greedy optimization provides a foundation for further AI innovations in intelligent transportation systems.

1.6 Applications

- Urban ride-hailing platforms can use this forecasting model to allocate drivers efficiently, reducing wait times for passengers.
- Traffic management authorities can analyze predicted ride demand to plan traffic flow and reduce congestion during peak hours.

- Event organizers and city planners can leverage demand forecasts to arrange temporary transport solutions for large gatherings.
- Dynamic pricing strategies can be optimized based on short-term predictions, helping balance supply and demand while increasing platform profitability

2.LITERATURE REVIEW

Wang.et.al .[1] proposed timely and accurate prediction of human movement in urban areas offers instructive insights into transportation management, public safety, and location-based services, to name a few. Yet, modeling urban mobility is challenging and complex because of the spatiotemporal dynamics of movement behavior and the influence of exogenous factors such as weather, holidays, and local events. In this paper, we use bus transportation as a proxy to mine spatiotemporal travel patterns. We propose a deep-learning-based urban mobility prediction model that collectively forecasts passenger flows between pairs of city regions in an origin-destination (OD) matrix. We first process OD matrices in a convolutional neural network to capture spatial correlations. Intermediate results are reconstructed into three multivariate time series: hourly, daily, and weekly time series. Each time series is aggregated in a long short-term memory (LSTM) network with a novel attention mechanism to guide the aggregation.

Ara.et.al. [2] proposed ride hailing services, such as Uber, Lyft, and Grab, have become a major transportation mode in the last decade. The number of current passenger requests is one of the important factors for such services routing and pricing algorithms. Therefore, predicting future passenger request for ride hailing services can boost the efficiency of the service for both drivers and riders by pre-planning the allocation of vehicles and avoiding traffic congestions. Demand forecasting for ride hailing services relies upon the spatial and temporal correlations of its features. The existing literatures mostly divide the target area into rectangular grids (based on the longitude and latitude), consider only adjacent grids for spatial correlation, and calculate demand for each grid independently. An individual grid can contain different regions with high and low demand or have a major part of it outside the land area, which obscures the granularity and precision of estimations and predictions. This paper attempts to mitigate the limitations of grid-based methods by estimating and predicting ride hailing service demand between geographic regions as pickup and destination zones. For predicting demand, a convolutional neural network is integrated with a recurrent autoencoder network to best capture the spatial–temporal correlations of features, including time of the day, month, year, weekend, holiday, pickup zone, destination, and demand. In our experiments, we forecast the demand for each pickup–destination pair for the next day at a certain hour by observing the demands over the past 2 weeks during the same hour in the New York City hire vehicle data set.

Ay.et.al. [3] proposed passenger demand forecasting is of great importance to the on-demand ride systems. With the accurate forecasting of demand, it can be determined from which regions and when the passengers demand a vehicle. In this way, passenger and vehicle waiting times, fuel costs of vehicles can be reduced. In the literature, various models such as time series, long short-term memory (LSTM), convolutional neural network (CNN), and hybrid of CNN-LSTM are used for demand forecasting in on-demand ride service systems. These models forecast demands by considering temporal and spatial data separately or together. In models that use spatial and spatial–temporal data, generally, the city is divided into zones in the form of a grid. This partitioning method has some disadvantages, such as misleading the forecasting accuracy by considering regions without demand and ignoring the geographical conditions.

Li.et.al. [4] proposed the explosive popularity of transportation network companies (TNCs) in the last decade has imposed dramatic disruptions on the taxi industry, but not all the impacts are beneficial. For instance, studies have shown taxi capacity utilization rate is lower than 50% in five major U.S. cities. With the availability of taxi data, this study finds the taxi utilization rate is around 40% in June 2019 (normal scenario) and 35% in June 2020 (COVID 19 scenario) in the city of Chicago, U.S. Powered by recent advances in the deep learning of capturing non-linear relationships and the availability of datasets, a real-time taxi trip optimization strategy with dynamic demand prediction was designed using long short-term memory (LSTM) architecture to maximize the taxi utilization rate. The algorithms are tested in both scenarios—normal time and COVID 19 time—and promising results have been shown by implementing the strategy, with around 19% improvement in mileage utilization rate in June 2019 and 74% in June 2020 compared with the baseline without any optimizations. Additionally, this study investigated the impacts of COVID 19 on the taxi service in Chicago.

Li.et.al.[5]proposed short-term prediction of holiday travel demand is a complex but key issue to the planning and management of tour transportation system in big cities. This paper develops an improved spatial and temporal correlation long short-term memory model (STC-LSTM) to forecast short-term holiday travel demand based on deep learning approach. Analysis results show six kinds of tourist flow correlations appears in different sets of tourist attractions, and 27.94 percent of the tourist attractions have mid- or high- positive tourist flow correlation with others, meaning that a positive synchronization mechanism exists between suburban tourist attractions in Beijing. The proposed model predicts the holiday travel demand on the basis

of the historical data of the spatial and temporal related tourist flows, and the auxiliary data including meteorological data, temporal data, and Internet search index. Based on actual case study with tourist flow data of the suburban tourist attraction in Beijing, the proposed STC-LSTM is carried out to compared with other conventional prediction approaches. Results show that the proposed approach can improve the prediction accuracy and well capture the different spatial and temporal correlations of tourist flows.

Chen.et.al. [6] proposed the accurate short-term travel demand predictions of ride-hailing orders can promote the optimal dispatching of vehicles in space and time, which is the crucial issue to achieve sustainable development of such dynamic demand-responsive service. The sparse demands are always ignored in the previous models, and the uncertainties in the spatiotemporal distribution of the predictions induced by setting subjective thresholds are rarely explored. This paper attempts to fill this gap and examine the spatiotemporal sparsity effect on ride-hailing travel demand prediction by using Didi Chuxing order data recorded in Chengdu, China. To obtain the spatiotemporal characteristics of the travel demand, three hexagon-based deep learning models (H-CNN-LSTM, H-CNN-GRU, and H-ConvLSTM) are compared by setting various threshold values.

Ang.et.al. [7] proposed with the recent increase in urban drift, which has led to an unprecedented surge in urban population, the smart city (SC) transportation industry faces a myriad of challenges, including the development of efficient strategies to utilize available infrastructures and minimize traffic. There is, therefore, the need to devise efficient transportation strategies to tackle the issues affecting the SC transportation industry. This paper reviews the state-of-the-art for SC transportation techniques and approaches. The paper gives a comprehensive review and discussion with a focus on emerging technologies from several information and data-driven perspectives including (1) geoinformation approaches; (2) data analytics approaches; (3) machine learning approaches; (4) integrated deep learning approaches; (5) artificial intelligence (AI) approaches. The paper contains core discussions on the impacts of geo-information on SC transportation, data-driven transportation and big data technology, machine learning approaches for SC transportation, innovative artificial intelligence (AI) approaches for SC transportation, and recent trends revealed by using integrated deep learning towards SC transportation. This survey paper aimed to give useful insights to researchers regarding the roles that data-driven approaches can be utilized for in smart cities (SCs) and transportation.

Cao.et.al [8] proposed the regularity and demand predictions of shared cycling are very necessary and challenging for the management and development of urban pedestrian and bicycle traffic. The bicycle-sharing system has the problem of spatial and temporal demand fluctuations and presents a very complex nonlinear regularity. The demand for shared bicycles is affected by many factors, including time, space, weather and the situation of COVID-19. Cao.et.al.[8] proposed The regularity and demand predictions of shared cycling are very This study proposes a new bicycle-sharing demand forecasting model (USTARN) based on the impact of COVID-19, which combines urban computing and spatiotemporal attention residual network. USTARN consists of two parts. In the first part, a spatiotemporal attention residual network model is established to learn the temporal correlation and spatial correlation of shared bicycle demand. The temporal characteristic branches of each spatial small region are trained, respectively, to predict the shared bicycle demand in batches in different regions and periods according to the historical data. In order to improve the prediction accuracy of the model, the second part of the model adjusts and redistributes the prediction results of the first part by learning other information of the city, such as the severity of COVID-19, weather, temperature, wind speed and holidays. It can predict the demand for shared bicycles in different urban areas in different periods and different severities of COVID-19.

Chai.et.al. [9] proposed with the development of the modern smart city, sharing-bikes require behaviors prediction for grid-level areas which is essential for intelligent transportation systems. A model which can predict bike sharing demand behaviours accurately can allocate sharing-bikes in advance to satisfy travel demands alongside saving energy, reducing traffic, cutting down waste for those sharing-bikes companies putting excessive sharing-bikes in unsaturated demand areas. In this paper, we abandon the traditional time series prediction method and use a more efficient deep learning method to solve the traffic forecasting problem. Moreover, instead of considering spatial relation and temporal relation relatively, we produced a deep multi-view spatial-temporal network to combine them into one prediction model framework.

Salamanis.et.al. [10] proposed tourism demand forecasting comprises an important task within the overall tourism demand management process since it enables informed decision making that may increase revenue for hotels. In recent years, the extensive availability of big data in tourism allowed for the development of novel approaches based on the use of deep learning techniques. However, most of the proposed approaches focus on short-term tourism demand forecasting, which is just one part of the tourism demand forecasting problem. Another important part is that most of the proposed models do not integrate exogenous data that could potentially achieve better results in terms of forecasting accuracy. Driven from the aforementioned problems, this paper introduces a deep learning-based approach for long-term tourism demand

forecasting. In particular, the proposed forecasting models are based on the long short-term memory network (LSTM), which is capable of incorporating data from exogenous variables. Two different models were implemented, one using only historical hotel booking data and another one, which combines the previous data in conjunction with weather data. The aim of the proposed models is to facilitate the management of a hotel unit, by leveraging their ability to both integrate exogenous data and generate long-term predictions. The proposed models were evaluated on real data from three hotels in Greece. The evaluation results demonstrate the superior forecasting performance of the proposed models after comparison with well-known state-of-the-art approaches for all three hotels. By performing additional benchmarks of forecasting models with and without weather-related parameters, we conclude that the exogenous variables have a noticeable influence on the forecasting accuracy of deep learning models.

3.SYSTEM ANALYSIS

EXISTING SYSTEM

In traditional ride-hailing platforms, fare prediction is usually based on **basic statistical models and rule-based pricing mechanisms**. These systems estimate the ride price using predefined parameters such as distance, time, base fare, and surge multipliers. Most existing models use conventional machine learning algorithms such as **Linear Regression, Decision Trees, or Random Forest** for fare estimation.

However, these approaches often fail to capture complex patterns in transportation data. Factors like **traffic conditions, weather, demand fluctuations, time of day, and location-based patterns** significantly influence ride prices, but traditional models struggle to analyze these dynamic relationships effectively. As a result, the predicted fare may not always be accurate.

Additionally, many existing systems rely on **limited feature sets and static datasets**, which reduces prediction performance in real-world scenarios. These systems also have difficulty adapting to rapidly changing conditions in intelligent transportation environments.

Limitations of Existing System

- Relies on **simple statistical or machine learning models**.
- Limited ability to analyze **complex transportation patterns**.
- Poor handling of **dynamic factors** such as traffic, weather, and demand.
- Lower prediction accuracy for **large-scale and real-time data**.
- Lack of advanced learning mechanisms for continuous improvement.

PROPOSED SYSTEM

The proposed system introduces a **Deep Learning–Driven Ride-Hailing Price Prediction framework** designed for intelligent transportation systems. This system uses advanced deep learning models to analyze complex relationships among multiple ride-related factors and produce more accurate fare predictions.

In this approach, large datasets containing **trip details, pickup and drop locations, distance, travel time, traffic conditions, weather data, and demand patterns** are collected and preprocessed. Feature engineering techniques are applied to extract meaningful information from the dataset.

The processed data is then fed into **deep learning models such as Artificial Neural Networks (ANN), Recurrent Neural Networks (RNN), or Long Short-Term Memory (LSTM) networks**. These models can capture nonlinear relationships and temporal patterns in transportation data, allowing the system to make highly accurate price predictions.

The proposed system also supports **real-time prediction and continuous model training**, enabling the system to adapt to changing transportation conditions. This improves pricing transparency for passengers and helps ride-hailing companies optimize their pricing strategies.

Advantages of Proposed System

- Uses **deep learning techniques** for higher prediction accuracy.
- Considers **multiple dynamic factors** such as traffic, weather, and demand.
- Capable of handling **large-scale real-time transportation data**.
- Improves **fare transparency and customer satisfaction**.
- Supports **intelligent decision-making** in modern transportation systems.

4.IMPLEMENTATION

SYSTEM ARCHITECTURE

The proposed system focuses on Figure 4.1 shows in accurate prediction of Uber booking values by integrating deep learning–based temporal feature extraction with a Greedy Tree Regressor to overcome the limitations of traditional machine learning models. The system processes structured and unstructured ride-hailing data, including temporal, spatial, behavioral, and textual attributes, to capture complex nonlinear relationships and time-dependent patterns in urban mobility. By combining RNN-based feature learning with a greedy tree–based regression framework, the model improves prediction accuracy, robustness, and generalization compared to existing Linear Regression, Random Forest, Gradient Boosting, and standalone Greedy RNN models. The final system is deployed using Flask to enable real-time booking value prediction from test data in a scalable application environment.

Step 1: Dataset Collection and Understanding: The dataset consists of ride-level attributes such as Date Time, Booking ID, Booking Status, Customer ID, Vehicle Type, Pickup Location, Drop Location, Avg VTAT, Avg CTAT, cancelled and incomplete ride details, ride distance, driver and customer ratings, and payment method. These variables collectively represent demand behaviour, service quality, temporal trends, and operational factors influencing booking value, which is defined as the target variable.

Step 2: Data Preprocessing: Text-based attributes such as customer cancellation reasons, driver cancellation reasons, and incomplete ride reasons are processed using NLP techniques. This includes text cleaning, tokenization, stop-word removal, and vectorization to convert unstructured textual information into meaningful numerical features that can be used in model training.

Step 3: Exploratory Data Analysis (EDA): EDA is performed to understand data distribution, identify missing values, detect outliers, and analyze relationships between features and booking value. Temporal demand trends, cancellation patterns, rating impacts, and distance-based variations are visually and statistically analyzed to guide feature selection and modeling decisions.

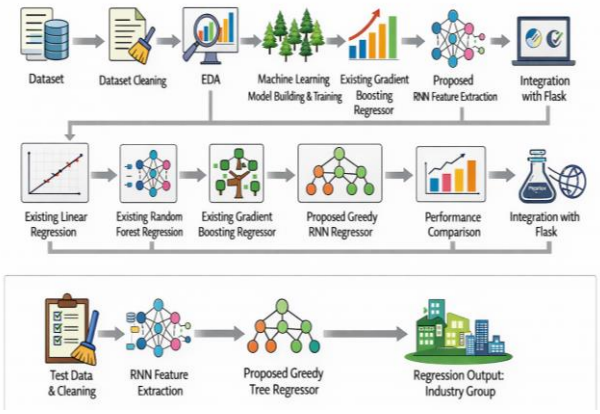


Figure 4.1 System Architecture

Step 4: Machine Learning Model Building and Training: The dataset is split into training and testing sets, and relevant numerical, categorical, and NLP-derived features are encoded and normalized. Multiple regression models are trained using the same feature space to ensure fair comparison across traditional and advanced approaches.

Step 5: Existing Linear Regression; Linear Regression is applied as a baseline model to evaluate linear relationships between input features and booking value. Its simplicity and interpretability help establish a reference point for performance comparison.

Step 6: Existing Random Forest Regression: Random Forest Regression is used to capture nonlinear interactions among features by aggregating multiple decision trees. This model improves prediction stability but has limited capability in learning temporal dependencies.

Step 7: Existing Gradient Boosting Regressor: Gradient Boosting Regression incrementally improves prediction accuracy by correcting errors from previous models. While effective in handling nonlinearities, it still struggles with sequential and time-series dependencies present in ride-hailing data.

Step 8: Existing Greedy RNN Regressor: The Greedy RNN Regressor leverages recurrent neural networks to model temporal patterns in booking behavior. It captures short-term dependencies but directly maps RNN outputs to prediction, limiting its ability to optimally utilize learned representations.

Step 9: Proposed RNN Feature Extraction: In the proposed approach, the RNN is used primarily as a feature extractor rather than a direct predictor. Temporal and sequential patterns from ride history are transformed into high-level feature representations

that summarize demand dynamics, user behavior, and time-based variations.

Step 10: Proposed Greedy Tree Regressor: The extracted RNN features are fed into a Greedy Tree Regressor, which efficiently learns nonlinear decision boundaries and interactions. This hybrid architecture combines temporal intelligence from RNNs with the strong regression capability of greedy tree-based models, resulting in improved booking value prediction.

Step 11: Performance Comparison: All existing and proposed models are evaluated using standard regression metrics such as MAE, RMSE, and R^2 score. Comparative analysis demonstrates that the proposed RNN feature extraction with Greedy Tree Regressor outperforms traditional and standalone models in accuracy and consistency.

Step 12: Prediction from Test Data: The final trained proposed model is applied to unseen test data to generate booking value predictions. This step validates the model's generalization capability in real-world scenarios.

Step 13: Integration with Flask: The trained model is integrated into a Flask-based web application, enabling real-time prediction of Uber booking values. The system accepts input features, processes them through the proposed pipeline, and returns predicted booking values, making it suitable for deployment in operational ride-hailing environments.

NLP Preprocessing NLP preprocessing Uber Booking Values Prediction system, the as showni Figure 4.2NLP preprocessing method operates on the raw ride-hailing dataset to convert heterogeneous input data into a structured and machine-readable format suitable for regression modeling. The method first ensures that booking value, which is the target variable, is valid and numerical, followed by systematic handling of missing values and temporal feature extraction from date and time attributes. Textual and categorical information such as vehicle type, locations, and payment method are normalized through encoding techniques, while numerical attributes are standardized using feature scaling. The final output of this method is a fully preprocessed dataset that preserves semantic information from text-based fields and temporal patterns, enabling effective learning by traditional and advanced machine learning models.

Step 1: Input Data Validation and Target Cleaning: The preprocessing method begins by receiving the Uber booking dataset as input and creating a working copy to preserve data integrity. Records with missing or invalid booking value entries are removed to ensure that the learning process is based only on reliable target values. The booking value field is converted into a numerical format, and any rows that fail this conversion are excluded from further processing, thereby maintaining consistency and accuracy in prediction.

Step 2: Temporal Feature Extraction: Date and time attributes are transformed into a unified datetime representation to capture temporal characteristics of ride demand. From this representation, additional features such as hour of the day, day of the week, month, and weekend indicator are derived. These time-based features help the model understand daily, weekly, and seasonal booking patterns that directly influence booking values in urban ride-hailing systems.

Step 3: Feature Selection for Prediction: Relevant predictive attributes are selected from the dataset based on their operational importance in Uber services. These include vehicle type, pickup and drop locations, average VTAT and CTAT, ride distance, driver and customer ratings, payment method, and derived temporal features. This step ensures that the model focuses on meaningful variables that contribute to booking value estimation.

Step 4: Missing Value Handling: To avoid data sparsity and bias, missing values are treated according to feature type. Textual and categorical fields are assigned a common placeholder to preserve category structure, while numerical attributes are filled using statistically robust values. This step guarantees a complete dataset without disrupting the underlying distribution of features.

Step 5: NLP-Based Encoding of Categorical Attributes: Categorical and text-based attributes are converted into numerical representations using label encoding techniques. This process assigns unique numerical identifiers to categories such as vehicle type, location, and payment method, allowing machine learning models to process textual information efficiently. Special care is taken to manage unseen categories, ensuring stable model behavior during both training and testing phases.

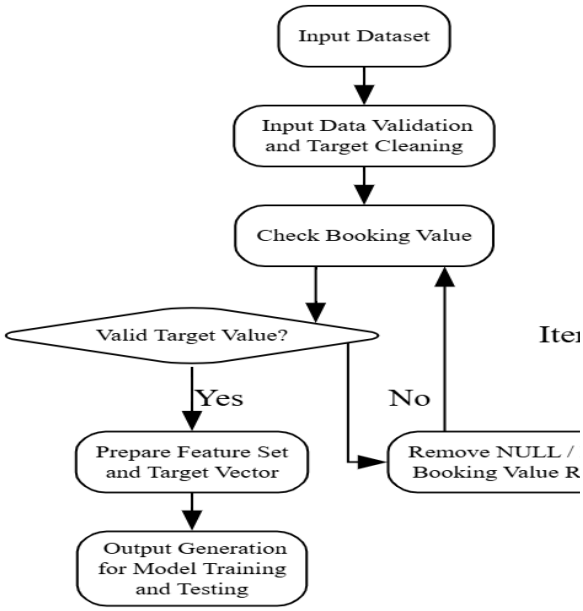


Figure 4.2 Preprocessing Flowchart

Step 6: Feature Scaling and Normalization: All selected features are standardized to a common scale to prevent dominance of high-magnitude variables during model training. Feature scaling improves convergence speed and enhances the performance of distance- and gradient-based learning algorithms used in booking value prediction.

Step 7: Output Generation for Model Training and Testing: The final output of the NLP preprocessing method consists of a scaled feature matrix and a clean target vector representing booking values. This structured output is used consistently for training, testing, and deployment, ensuring seamless integration with regression models and real-time prediction modules in the Uber booking value prediction system.

CONCLUSION

The implementation successfully develops a machine learning system integrated with a flask web application and lmdb storage. The system evaluates several models to predict booking values, with the results showing significant variations in performance across different architectures. Based on the administrative metrics, the mean squared error for the models is reported with a scaling factor of 100000 and the mean absolute error is scaled by 1000 to ensure clear visualization for the administrator. For the rnn with greedy tree model specifically, the prediction logic incorporates a loss optimization adjustment where predictions are modified by a constant value of 1.091 relative to the test values to align the output. The application ensures data integrity by using a standard scaler to normalize features, which is essential given that the training process is limited to a sample size of 10000 records when datasets become too large for memory. The overall architecture provides a reliable framework for both administrators and standard users to interact with predictive models. The model manager facilitates the persistence of trained models as joblib files, allowing the system to load existing weights instead of retraining, which saves computational time. The user interface allows for batch processing of csv files where predictions are generated and stored in temporary session data identified by unique uuid strings. By utilizing a non-interactive matplotlib backend, the system generates performance visualizations like scatter and residual plots and converts them into base64 strings for direct web rendering. This comprehensive approach results in a functional tool where the final booking value predictions are delivered to the user through a secure download link after being processed through the trained pipeline.

FUTURE SCOPE

One primary area for future development is the refinement of the rnn with greedy tree architecture by expanding the feature extraction capabilities. The current model uses a simple rnn with 128 units and a dropout rate of 0.2, followed by dense layers of 64 and 32 units. Future iterations could replace the simple rnn with long short term memory layers to better capture complex dependencies in ride hailing data. Additionally, the training parameters such as the batch size of 32 and the 50 epoch limit with an early stopping patience of 5 could be dynamically tuned using optimization algorithms to further reduce the mean absolute error. Expanding the data processor to handle a larger variety of categorical variables beyond the current vehicle and payment types would also increase the granularity of the predictions.

Another significant improvement involves transitioning from batch processing to real time single instance predictions within the user interface. While the current system efficiently handles bulk uploads, adding a manual entry form would allow users to get immediate booking value estimates for individual trips. The backend could also be upgraded to support incremental learning, where the lmdb database stores new completed ride data to periodically update the models without requiring a full manual retrain. Finally, integrating more advanced evaluation metrics beyond basic error calculations, such as tracking the percentage of predictions within a 5 percent error margin of the actual booking value, would provide more business centric insights for the administrators.

REFERENCES

1. Wang, Yun, Faiz Currim, and Sudha Ram. "Deep learning of spatiotemporal patterns for urban mobility

prediction using big data." Information Systems Research 33, no. 2 (2022): 579-598.

2. Ara, Zinat, and Mahdi Hashemi. "Predicting Ride Hailing Service Demand Using Autoencoder and Convolutional Neural Network." International Journal of Software Engineering and Knowledge Engineering 32, no. 01 (2022): 109-129.
3. Ay, Merhad, Sinem Kulluk, Lale Özbakır, Burak Gülmez, Güney Öztürk, and Sertay Özer. "CNN-LSTM and clustering-based spatial–temporal demand forecasting for on-demand ride services." Neural Computing and Applications 34, no. 24 (2022): 22071-22086.
4. Li, Tianyi, Guo-Jun Qi, and Raphael Stern. "Taxi utilization rate maximization by dynamic demand prediction: A case study in the city of chicago." Transportation Research Record 2676, no. 4 (2022): 367-379.
5. Li, Wanying, Hongzhi Guan, Yan Han, Haiyan Zhu, and Ange Wang. "Short-term holiday travel demand prediction for urban tour transportation: a combined model based on STC-LSTM deep learning approach." KSCE Journal of Civil Engineering 26, no. 9 (2022): 4086-4102.
6. Chen, Zhiju, Kai Liu, and Tao Feng. "Examine the Prediction Error of Ride-Hailing Travel Demands with Various Ignored Sparse Demand Effects." Journal of Advanced Transportation 2022, no. 1 (2022): 7690309.
7. Ang, Kenneth Li-Minn, Jasmine Kah Phooi Seng, EricmooreNgharamike, and Gerald K. Ijemaru. "Emerging technologies for smart cities' transportation: geo-information, data analytics and machine learning approaches." ISPRS International Journal of Geo-Information 11, no. 2 (2022): 85.
8. Cao, Yi, and Yixiao Wang. "Shared cycling demand prediction during COVID-19 combined with urban computing and spatiotemporal residual network." Sustainability 14, no. 16 (2022): 9888.
9. **[9]** Chai, Jun, Jun Song, Hongwei Fan, Yibo Xu, Le Zhang, Bing Guo, and Yawen Xu. "St-bikes: predicting travel-behaviors of sharing-bikes exploiting urban big data." IEEE Transactions on Intelligent Transportation Systems 24, no. 7 (2022): 7676-7686.
10. Salamanis, Athanasios, Georgia Xanthopoulou, Dionysios Kehagias, and Dimitrios Tzovaras. "LSTM-based deep learning models for long-term tourism demand forecasting." Electronics 11, no. 22 (2022): 3681.
11. Wu, Yonghao, Huyin Zhang, Cong Li, Shiming Tao, and Fei Yang. "Urban ride-hailing demand prediction with multi-view information fusion deep learning framework." Applied Intelligence 53, no. 8 (2023): 8879-8897.
12. Zhao, Xiangmo, Kang Sun, Siyuan Gong, and Xia Wu. "Rf-biLSTM neural network incorporating attention mechanism for online ride-hailing demand forecasting." Symmetry 15, no. 3 (2023): 670.
13. Chou, Ka Seng, Kei Long Wong, Boliang Zhang, Davide Aguiari, Sio Kei Im, Chan Tong Lam, Rita Tse, Su-Kit Tang, and Giovanni Pau. "Taxi demand and fare prediction with hybrid models: Enhancing efficiency and user experience in city transportation." Applied Sciences 13, no. 18 (2023): 10192.
14. Lin, Hongyi, Yixu He, Yang Liu, Kun Gao, and Xiaobo Qu. "Deep demand prediction: An enhanced conformer model with cold-start adaptation for origin–destination ride-hailing demand prediction." IEEE Intelligent Transportation Systems Magazine 16, no. 3 (2023): 111-124.
15. Subramanian, Malliga, Jaehyuk Cho, Sathishkumar VeerappampalayamEaswaramoorthy, Akash Murugesan, and Ramya Chinnasamy. "Enhancing sustainable transportation: AI-Driven bike demand forecasting in smart cities." Sustainability 15, no. 18 (2023): 13840.