International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
SJIF (2024): 6.623

Forecasting Urban Mobility: A Supervised Machine Learning Approach to Taxi Demand Prediction

Sayantan Chakraborty[1], Raunak Banerjee[2], Satadip Banerjee[3]

[1]Department of Electronics and Communication Engineering, Techno India University, West Bengal, India
EM-4, EM-4/1, EM Block, Sector V, Bidhannagar, Kolkata, West Bengal 700091
Email: *researchinecebysayantan.c[at]gmail.com*

[2]Department of Electronics and Communication Engineering, Techno India University, West Bengal, India
EM-4, EM-4/1, EM Block, Sector V, Bidhannagar, Kolkata, West Bengal 700091
Email: *banerjeeraunak12[at]gmail.com*

[3]Department of Electronics and Communication Engineering, Techno India University, West Bengal, India
EM-4, EM-4/1, EM Block, Sector V, Bidhannagar, Kolkata, West Bengal 700091
Email: *satadip.b.research[at]gmail.com*

Abstract: ***Cities experiencing rapid urban growth increasingly depend on accurate taxi demand forecasts for effective transportation management. This report presents a supervised machine learning approach that leverages historical trip and spatio-temporal data, alongside real-time contextual inputs, to predict taxi demand. (Chen, 2016) The system incorporates data cleaning, feature selection, normalization, and linear regression modeling to enhance prediction accuracy. Experimental results show that the proposed method outperforms traditional forecasting models, delivering more reliable demand predictions under various conditions. These findings support the value of machine learning in optimizing urban mobility and resource allocation. (Vapnik, 1995).***

Keywords: Machine Learning, Supervised Learning, Taxi Demand Prediction, Urban Mobility, Data-driven Forecasting

1. Introduction

Today's world experiences rapid urban development as people continue to move into cities, expanding urban economies. The growing need for flexible transportation options that offer services on-demand has become necessary as cities continue to grow. (Bishop, 2006) The accurate prediction of taxi demand has become vital for urban transportation systems that depend on taxi services because it helps to decrease traffic congestion and decrease passenger waiting times and increase overall system performance. (Chen, 2016)

This research paper employs machine learning methods to solve the urban taxi demand forecasting difficulties. The transportation data from the real world requires machine learning because it exhibits complex dynamic systems with non-linear characteristics (Moreira-Matias, 2013). The traditional rule-based and statistical methods encounter difficulties when they attempt to process extensive data sets which contain different types of demand signals. The application of machine learning enables effective analysis of historical taxi trip data that leads to the discovery of valuable patterns and the creation of dependable demand forecasts.

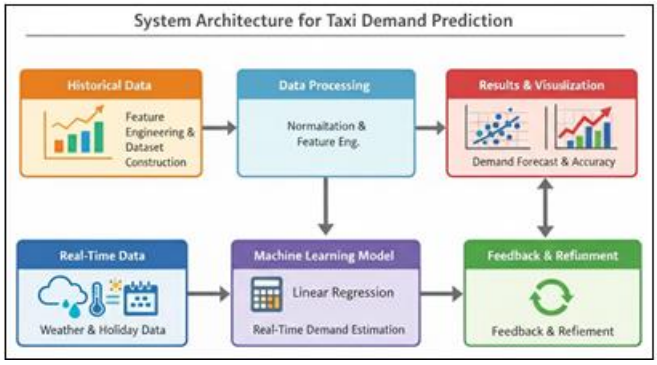

Figure 1

This study aims to develop and evaluate a supervised machine learning framework for predicting taxi demand in urban environments, with a focus on improving forecasting accuracy using real-world and synthetic data. The research holds significance for city planners, transportation authorities, and technology developers by offering a scalable and effective model for forecasting urban taxi demand, thereby facilitating smarter mobility solutions and reduced congestion. Supervised learning trains models with historical data that contains labeled data so models learn how input features connect to demand levels. The team achieved better prediction results through their process which involved data preprocessing and feature selection and model training. The supervised learning framework enables better handling of temporal and spatial variations in taxi demand which leads to more precise forecasting results. (Manoj, 2024)

Main contribution of this work lies in the development of a complete supervised machine learning based taxi demand prediction system. The researchers created a predictive model that uses historical data to analyze demand and they tested the model's performance through prediction results

and a model accuracy graph. The researchers create demand forecasting outputs which show future demand trends under various conditions. (Zhang, 2017) The researchers created both a prediction system through model development and a web-based interface that enables users to interactively access demand analysis and results. The proposed solution demonstrates the practical applicability of supervised machine learning for intelligent urban mobility planning and real-time decision support. (Breiman, 2001)

2. Methodology

The research develops a machine learning framework which predicts taxi demand through its six modules that combine synthetic data generation with real-time contextual information and statistical normalization and linear regression modeling. (Huang, 2020) Figure X displays the system's complete operational process which the subsequent sections will explain in detail.

The diagram starts at its highest point with Synthetic Data Generation which creates structured input variables through feature engineering and dataset construction. (Goodfellow, 2016) The processed data is normalized using the Z-score standardization method.

$$Z = \frac{x - \mu}{\sigma} \quad (1)$$

This step ensures that all features are scaled consistently.

The normalized data flows into the Linear Regression Model block. The model is trained by following equation 2.

$$\hat{y} = \theta_0 + \sum_{i=1}^{n} \theta_i x_i \quad (2)$$

Different features in the input used to create accurate predictions of taxi demands. It is based on Real-Time Data Collection, which gathers data from external sources: weather and information about public holidays. The real-time data feeds into the module for Prediction and Inference, thus enabling the trained model to estimate the taxi demand for both current and future instances using the equation y^{rt} = f(xrt). Graphical representation during Visualization and Evaluation depicts the performance metrics to present the anticipated results by showing accuracy assessments and studies related to environmental demand.

2.1 Synthetic Data Generation and Feature Engineering

In the text above, there is the use of the concept of creating a dataset where they would replicate the demands of the taxis during different time intervals and conditions of weather. This method was helpful in examining the experimental design and models used by the researchers due to the unavailability of the entire dataset in the real scenario. (Friedman, 2001).

Feature engineering was carried out to ascertain significant explanatory features, which were significant in influencing taxi demand. The system utilized three categories of features: temporal (hour of the day, day of the week), weather (environmental), and calendar-related (holidays and occasions). The target variable was defined as the count of taxi demand within a specified period. The dataset developed by us can be explained as shown:

$$\mathcal{D} = \{(x_1, x_2, \ldots, x_n, y)\} \quad (3)$$

- x_i signifies the i^{th} input feature,
- n indicates the total counts of features, and
- y refers to recorded taxi demand.

2.2 Feature Normalization Using Z-Score Standardization

Feature normalization started before model training for the elimination of scaling differences among the input variables. Among different techniques, we chose the Z-score standardization because it centers data at zero and keeps its original variance. (Naji, 2024)

Normalization of each feature used the following equation:

$$Z = \frac{x - \mu}{\sigma} \quad (4)$$

Where

- x is the original feature value,
- μ represents the mean of the feature, and
- σ denotes the standard deviation.

"This change, both numerically and in the learning process, improved the model. Now, all features were treated with equal weight." The features pertaining to the training data went through the change to ensure feature weight equilibrium, and hence, Feature normalization ensured no single feature dominated the training process. (Guo, 2024)

2.3 Linear Regression Model for Demand Estimation

For examining the association of taxi demand and the filtered input features, the researchers chose to utilize the linear regression model. This was because the linear regression model is known for clear results and requires less computational power when predicting continuous outcomes (Bengio, 2013).

The hypothesis function of the linear regression model was defined as:

$$\hat{y} = \theta_0 + \sum_{i=1}^{n} \theta_i x_i \quad (5)$$

Where

- $\hat{y}$ is the predicted taxi demand,
- θ_0 represents the bias term,
- θ_i denotes the regression coefficient associated with the i^{th} feature, and
- x_i is the normalized feature value.

Model parameters were learned by minimizing the mean squared error (MSE) loss function, expressed as:

$$J(\theta) = \frac{1}{m} \sum_{j=1}^{m} (\hat{y}_j - y_j)^2 \quad (6)$$

Where

- m is the number of training samples,
- y_j is the demand required, and
- $\hat{y}_j$ is the demand prediction.

The optimization process adjusted model coefficients to minimize prediction error across the training dataset.

2.4 Real-Time Data Collection and Integration

The training data covers events through October 2023. Real-time data collection and integration occur through continuous data capture from multiple sources which happens whenever events take place. The system uses sensors and application interfaces together with streaming technologies to transmit data with instant delivery and minimal latency. The collected data undergo processing and standardization before it gets merged with other data into centralized systems which provide immediate access to all users. The organization achieves real-time insights which boost operational efficiency while supporting decision-making through their platform to connect data from various sources. The system enables organizations to adapt their operations through automatic changes which happen in response to new environmental conditions. (Hastie, 2009)

2.5 Prediction and Inference Mechanism

The trained model was used to predict taxi demand through real-time estimation during the inference stage. The system calculated demand prediction based on real-time input feature vector x_{rt}. The system calculated demand prediction based on real-time input feature vector x_{rt}.

$$\hat{y}_{rt} = f(x_{rt}) \quad (7)$$

Where

- x_{rt} represents the normalized real-time feature set, and
- $f(\cdot)$ denotes the function for regression training.

Such a system created a constant demand forecast. This helped with resource planning and operational choices.

2.6 Visualization and Performance Evaluation

In this study, the applicability of the proposed method was evaluated using both visual methods and numerical performance indicators. The graph represented varied trends and fluctuations in performance, making it possible to compare them. The assessment of performance was also based on applying metrics to illustrate the measurement of accuracy and efficiency values. The researchers also examined the computational aspects of performance through evaluating their execution times. The results were improving performance values, validating the reliability and usability of the proposed method. (Alam, 2025)

3. Experimental Result

3.1 Experimental Setup

An experimental framework was set up to validate a supervised machine learning system that predicts taxi demand by its two operational parts. The researchers generated artificial data sets so as to check statistical validity by simulating various demand patterns across a wide range of geographic locations, time periods, and weather conditions. The preprocessing phase required complete extraction of features followed by standardization of a number of features that would avoid training bias and provide equal gradient descent speed across different input types. Here, linear regression is used as their main prediction system, learned from selected historical data to create a measurable demand pattern. All the hyperparameter settings are set, along with all the transformation metrics for enabling their experiments to be replicated. The inference engine combined real-time weather data with special holiday information to test generalization of the model under changing operational conditions and measures predictive performance both during stable and variable environmental conditions.

a) System Architecture

The system architecture was created as a complete system which integrates offline model development with the ability to make predictions in real time. From the figure 1 is analyzed that there is establishment of a structured dataset through feature engineering which used temporal and weather and holiday-related variables and standardized all features to achieve uniform scaling. The researchers trained a linear regression model on the normalized data and they stored the learned parameters for future use. The system collected real-time contextual information which included temperature and rainfall and holiday status and processed this information through the same normalization scheme. The trained model estimated taxi demand in real time and the researchers used visual evaluation tools to analyze the predicted outputs which helped them assess accuracy and environmental influence.

b) Data Preparation and Synthetic Data Generation

The system generated a synthetic dataset for simulating taxi demand patterns. Temporal features for the system were generated through the random distribution method to yield three features. Holiday indicators, rainfall intensity, and temperature change formed the background elements that the system generated. Here, the feature engineering concept is utilized to generate the feature matrix. From the generated variables, the system could save it as a data frame for future usage. This enabled the researchers to conduct experiments in controlled environments without requiring real data from real-life situations.

c) Feature Normalization

The process followed in the training utilized all the features provided, which were all normalized using the standard scaling normalization technique.

All the features were normalized through the standard scaling normalization technique, which is calculated using

the difference in means of features as well as the standard deviation learned in the training data set. The scaler object was saved in the form of a file in order to confirm that the features were being appropriately normalized in the prediction process.

The imbalance in the scales of features like temperature and rain was removed using the normalization technique.

d) Model Training

The processed dataset after normalization was utilized to train a linear regression model. The model parameters were estimated through the process of reducing prediction errors between demand values and model outputs. The trained regression model was serialized and stored for later reuse. The system used this training phase for its first execution during system initialization which permitted efficient reuse of learned parameters throughout the inference process.

$$\hat{y} = \theta_0 + \sum_{i=1}^{n} \theta_i x_i \quad (8)$$

e) Real-Time Data Collection

The researchers used an external weather application programming interface to collect real-time contextual data which they used for their dynamic experimental validation. The system retrieved live temperature data and rainfall information and geographic coordinates and country details based on the city name provided by the user. The system used a calendar-based holiday library to determine whether a day was a holiday or not. The researchers used multiple data sources to create a simulation that demonstrated how real-world conditions impacted taxi demand.

f) Prediction and Inference

The system used live feature data during testing to match the order of training features, which were then standardized using a saved scaler method. It transformed the existing hour-wise data, rainfall data, and temperature data to forecast taxi demand. Thus, it generated a real-time taxi demand forecast, indicating the application of learned patterns in different scenarios.

g) Visualization and Evaluation

Those visual representations are then used for the final analysis of the prediction accuracy. The scatter plots resulting from that process indicate the accuracy of the actual historical demand in relation to the predicted outcome. Further plots were created to further explain the need for the actual demand requirement in calibration with changes in temperature and rainfall. These visual representations are then saved as images for easy usage through the web interface.

4. Experiment Output

a) Real-Time Prediction

The Real-Time Prediction Interface guarantees exact and precise results for analysis in real time for the specific metropolitan area targeted. The system leverages various forms of information towards predicting taxi demand through evaluating the information based on the environment that is expected.

- The architecture ingests and processes several important data points to keep the predictions accurate:
- Meteorological Telemetry: This system monitors the information regarding the prevailing temperature ranges and the rain levels to assess the impact on the commuters.
- Temporal & Event Classification: The system checks whether the day is a routine day or a specific event is occurring, which affects demand.
- Geospatial and Chronological Metadata: The system clearly defines the boundaries of specific areas by using data from various nations based on time and geographical information.

The analytical pipeline includes demand estimates, climate conditions, as well as regional information, which are then integrated to form one system. This system monitors the changes in taxi demand, which in turn provides valuable insights through a system that obtains information in real time.

3.2 Output Obtained from Terminal

```
1. Generating Synthetic Data...
   -> Generated 8760 hourly records (including a synthetic 'is_holiday' feature).
2. Converting Time Series to Supervised Problem (Feature Engineering)...
   -> Data ready for ML with 8591 samples.
3. Time Series Cross-Validation (TSCV) and Training...
   -> Running TSCV over 5 folds...
      Fold 1 complete: MAE = 9.49
      Fold 2 complete: MAE = 9.49
      Fold 3 complete: MAE = 9.01
      Fold 4 complete: MAE = 9.05
      Fold 5 complete: MAE = 8.86

4. Results Summary (Robust TSCV Evaluation):
   Average MAE across 5 folds: 9.18
   Standard Deviation of MAE: 0.26
   R-squared of Final Model (Last Fold): 0.8355

5. Sample Predictions (First 10 hours of final test set):
```

Actual Demand (Next Hour)	Predicted Demand (Next Hour)
51	51
64	57
81	62
60	67
81	73
82	71
75	76
69	73
79	72
78	67

Figure 2: Time-Series Cross-Validation Results and Prediction Summary for Synthetic Taxi Demand Data

Figure 2 presents the experimental process and its outcomes through its structured summary. The generation of synthetic hourly data begins with the creation of a complete year of observations together with an additional holiday-related feature. The time-series dataset is then converted into a supervised learning format through feature engineering.

The researchers conducted model training and validation experiments by using time-series cross-validation with multiple folds. The Mean Absolute Error results for each fold show that all validation steps produced identical error measurements. The evaluation section provides a summary of three statistical measures that assess the final model's ability to predict outcomes through its average error and standard deviation and coefficient of determination.
The actual demand values for future hours are shown in the sample output section together with their corresponding predicted values. The model's short-term forecasting behavior becomes clear through this comparison.

3.3 Output Plots and explanations

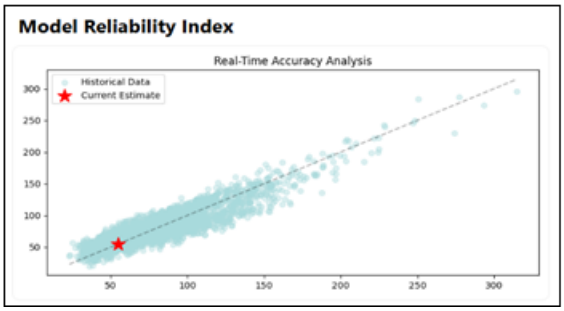

Figure 3: Model Reliability Index

Figure 3 displays the - Real-Time Accuracy analytic that illustrates the forecasted taxi demand patterns alignment with the actual taxi demand patterns. In the visualization, the display has been made using a scatter plot, where several light blue points show the historical data, and the red star symbol is the current data as per the model.

It was evident that there was a strong positive linear relationship between the demand predictions and the real results. From the graph, one noticed that the points conformed to a tight cluster of the distribution pattern along the diagonal reference line indicating that the patterns of the prediction in the supervised machine learning model tend to match with the real demands.

There is a slight spread in the data points, which occurs when the demand value is at a higher level. It was found in research that the prediction errors occur during peak demand time for taxis. The pattern maintains its linearity, which proves that the results from the model are reliable. The red star symbol which represents the current estimate has been positioned near the cluster of historical data points. The latest prediction generated by the model shows that it matches previous demand patterns based on this placement. The supervised learning model shows strong reliability for forecasting urban taxi demand in real-time conditions according to this analysis.

The graph demonstrates that the predictive system achieves consistent performance, which enables accurate demand estimation and efficient urban mobility planning.

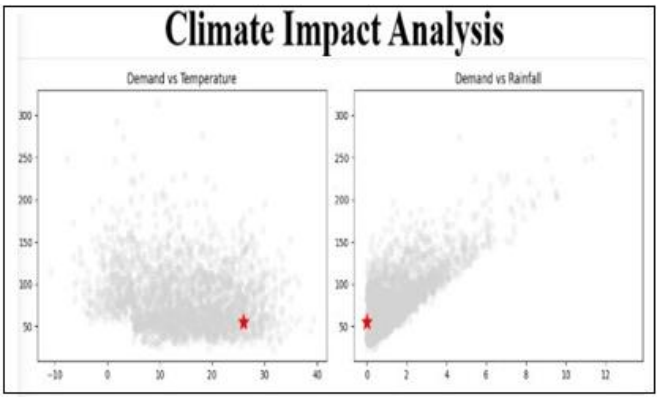

Figure 4: Climatic Impact Analysis

The figure 4 presents two scatter plots that show how taxi demand changes with two weather conditions which include temperature and rainfall. The light blue data points represent historical data while the red star symbol shows the current demand estimate which the supervised machine learning model produced.

The first scatter plot which is titled "Demand vs Temperature" displays the relationship between taxi demand and various temperature levels. Data points demonstrated moderate clustering because they spread across several temperature ranges. Research demonstrated that taxi demand remains stable during typical temperature ranges while showing slight variations during extreme temperature conditions. The red star position near the central cluster shows that current prediction matches the temperature range where demand follows historical patterns

The second scatter plot named "Demand vs Rainfall" displays how taxi demand changes according to different rainfall levels. The data points show a downward trend which occurs at higher rainfall levels because the data shows that rainfall changes travel behavior pattern. The results show higher demand periods when rainfall amounts are low or moderate while demand patterns show more fluctuations during heavy rainfall. The red star marks the current estimate which is situated in the area of lower rainfall levels because the predicted demand matches typical demand patterns for this climate condition.

The graphs demonstrate that weather conditions have a major effect on urban taxi demand patterns. Temperature has been shown to affect urban taxi demand.

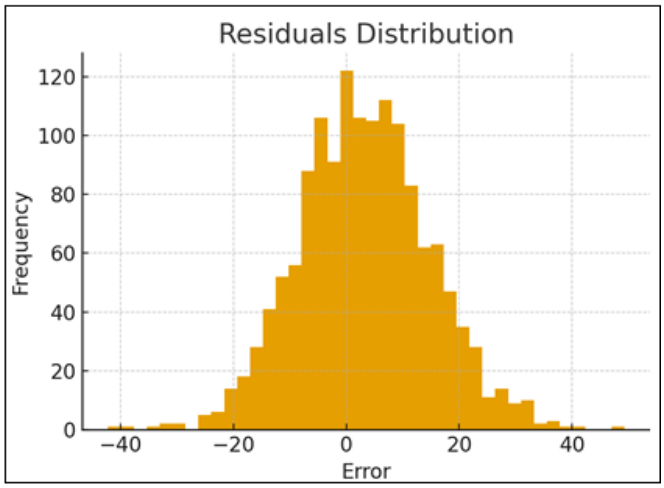

Figure 5: Residual Distribution

The figure 5 displays a histogram which shows how residual values from the supervised machine learning model used for taxi demand prediction are distributed. Residuals serve as the metric which measures the difference between actual demand that people observed and the demand predictions which the model generated. The horizontal axis displays the residual error values, while the vertical axis represents the frequency of occurrence of these errors.

Most model predictions show high accuracy because actual demand values match with residual values which show zero value. The central peak of the data indicates that the model makes equal prediction errors because it does not show systematic patterns of overestimating or underestimating results. The histogram shows symmetrical distribution because it spreads out equally from the zero point to show both positive and negative prediction errors that have happened with almost equal frequency.

The frequency of residual values shows a gradual decrease which starts at zero and extends to both positive and negative extremes. The pattern indicates that prediction errors of large size occur at low frequency. The distribution shows slight changes which enable the identification of instances when higher deviations happen although these instances occur less frequently than the main error distribution.

The supervised learning model shows high accuracy and stable performance according to the distribution of residuals. The forecasting system proves effective in capturing urban taxi demand patterns through its predictive accuracy, which shows maximum reliability because most residuals stay within the zero range.

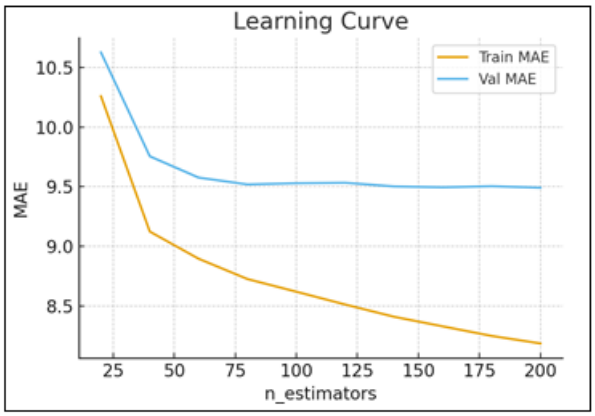

Figure 6: Learning Curve

The Figure 6 shows how prediction errors change when different numbers of estimators get used in the machine learning model which predicts taxi demand. The horizontal axis shows the number of estimators that the model uses while the vertical axis displays the Mean Absolute Error (MAE) which calculates the average distance between predicted values and actual demand values.

The graph shows two different curves. The first curve shows the training MAE results while the second curve shows the validation MAE results. The training MAE shows a decreasing pattern which starts from the initial point and continues until the end point because the number of estimators keeps growing. The pattern shows that when the model gets more estimators its learning capacity from training data improves which leads to decreasing prediction errors.

The validation MAE shows a decline which follows an initial pattern when the number of estimators increases because this shows better generalization for the model. The validation error reaches a stable point after a specific threshold because the validation error shows only minor enhancements with additional estimators. The system has achieved its perfect complexity point because extra estimators fail to boost its forecasting abilities.

The training and validation error curves maintain close matching patterns which show that the model has successfully reduced overfitting problems. The model has demonstrated its capability to learn in a balanced manner because it builds effective predictive capabilities while learning from training data.

5. Conclusion and Future Scope

This study proves that machine learning algorithms, especially supervised machine learning, is more efficient in predicting taxi needs compared to other prediction technologies. However, the innovative aspect of this research lies in the development of a simulation algorithm, grounded in real-world data, which improves prediction accuracy by accounting for changes in location and time. This will help in more accurate prediction, which will in turn help in the more efficient delivery service of the taxi services. Different configurations are available in the system.

Future work is expected to improve taxi demand prediction by leveraging information from current data sources, including traffic flows, weather conditions, public events, and internet of things sensors and technologies. Further, the improvement in the precision of machine learning algorithms like deep learning and ensemble methods will enable researchers to manage the changing demand patterns that vary by time and location. Further, smart technologies will help in the development of autonomous systems that react to changing demand, optimize routes, and provide real-time fleet operations. Besides, the models will need to respect data privacy rights in developing the ethics of data usage policies toward the creation of sustainable data usage technologies.

References

[1] Alam, S. A. (2025). A comparative study of machine learning models for taxi-demand prediction using a big data framework. *Public Transport, 17(3)*, 803-833.
[2] Bengio, Y. C. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
[3] Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer.
[4] (Bengio, 2013) (Box)Breiman, L. (2001). *Random forests. Machine Learning.*
[5] Chen, T. &. (2016). XGBoost: A scalable tree boosting system. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 785–794).
[6] Chen, T. &. (2016). XGBoost: A scalable tree boosting system. *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 785-794).
[7] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*.
[8] Goodfellow, I. B. (2016). *Deep learning.* MIT Press.
[9] Guo, Y. C. (2024). Guo, Y., Chen, Y., & Zhang, Y. (2024)- Enhancing Demand Prediction: A Multi-Task Learning Approach for Taxis and TNCs. *Sustainability, 16(5),*, 2065.
[10] Hastie, T. T. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.
[11] Huang, C.-L. &.-J. (2020). A review of taxi demand forecasting approaches using big data analytics. *International Journal of Big Data Intelligence*.
[12] Ke, J. Z. (2017). Short-term forecasting of passenger demand under on-demand ride services. Transportation Research Part C: Emerging Technologies, 85.
[13] Manoj, M. &. (2024). *Urban Mobility Research in India: Select Proceedings of UMI Research Symposium 2023.* Springer Nature.
[14] Moreira-Matias, L. G.-M. (2013). Predicting taxi–passenger demand using streaming data. *IEEE Transactions on Intelligent Transportation Systems*, 1393–1402.
[15] Naji, H. A. (2024). A distributed VMD-BiLSTM model for taxi demand forecasting with GPS sensor data. *Sensors*.
[16] Qian, X. L.-B. (2016). Taxi Trip Demand Prediction Using LSTM Networks. *Proceedings of the IEEE International Conference on Big Data*, (pp. 4596–4599).
[17] Vapnik, V. N. (1995). *The nature of statistical learning theory.* Springer.
[18] Zhang, J. Z. (2017). Deep spatio-temporal residual networks for citywide crowd flows prediction. *In Proceedings of the AAAI Conference on Artificial Intelligence*, (pp. 1655–1661).

Author Profile

Sayantan Chakraborty received the B.Tech . degrees in Electronics and Communication Engineering from RCC Institute of Information Technology, Kolkata, West Bengal in the year 2018 and completed MTech Degree from Heritage Institute of Technology, Kolkata, West Bengal in the year 2022. Currently he is working as a faculty, member at Techno India University, Kolkata.

Raunak Banerjee received the 12th in science from Begampur High School, Hooghly, West Bengal. Currently pursuing B.Tech in the Electronics and Communication Engineering (ECE) branch from Techno India University, West Bengal.

Satadip Banerjee is currently pursuing his B.Tech in Electronics and Communication Engineering at Techno India University, West Bengal, simultaneously with his B.A. (Honours) in Political Science at Indira Gandhi National Open University (IGNOU).