

Machine Learning Based-Bus Arrival Time Prediction Using Historical Data

Co-Author - **Dr.P.Sumalatha**

Assistant Professor

Department of Computer Science & Artificial intelligence Central University of Andhra Pradesh,

Ananthapuramu, India

Email : sumalatha.cuap@gmail.com

Author - **Salla Abhishek**

Department of Computer Science & Artificial intelligence Central University of Andhra Pradesh,

Ananthapuramu, India

Email : abhisheksalla@gmail.com

Abstract

The accurate prediction of bus arrival times would have the potential to greatly benefit mobility of the city as well as the commuter. A machine learning based system to predict bus arrival times was introduced in this study based on historical and real-time transit information. Factors like trip length, time of day, traffic data, and route characteristics were introduced as features for prediction in this system model to learn from non-linear relationships.

Various regression models were evaluated to determine the optimal choice for achieving better predictions. The suggested approach is superior to other approaches because it is able to learn from the non-linear relationships that are prevalent in transit data and was able to produce highly accurate predictions. This approach emphasizes the importance of data pre-processing, feature engineering, and model selection for prediction systems. Not only can this system assist riders in their commute but public transportation systems could greatly benefit in service planning and operations.

Keywords: Bus Arrival Prediction, Machine Learning, Travel Time Estimation, Public Transportation, Data Analytics, Smart Mobility.

1.INTRODUCTION

The present day complex environment, public transportation acts as a vital lifeline for the everyday commutes of thousands of individuals. Among all the means of transport, buses can be considered one of the cheapest and most commonly used modes of transport, specifically in developing countries. However, one of the main issues which the public transportation bus passengers encounter is the accuracy in prediction of arrival times of a bus to the bus stop. This is often found to be inaccurate because of factors such as traffic, weather, road conditions, high traffic in peak hours etc, which provides no effective means of transport to the passenger for trip planning, and ultimately leads to user dissatisfaction.

This project aims to design and develop such a system for the prediction of bus arrival times by utilizing machine learning. Unlike traditional ways which often relies on schedule of bus routes or simple prediction based on short period of time, this system using both current and historical time series collected data, effectively extracting latent relationship within the dataset

The basis is that from complex dataset one can derive underlying pattern of varied traffic based on time of day, standard delay for various routes and the effect of peaks in demand to travel times.

This is achieved by performing collection of transportation data followed by pre-processing, then feature engineering process of extracting necessary attributes like temporal characteristic, route distance, and traffic conditions, followed

by Gradient Boosting which will predict the travel time of the buses more accurately. These models will prove efficient in modeling non-linear relation between the various attributes.

The implementation of this system will lead to increase accuracy in prediction of arrival time of the buses which will result in the enhancement of traveler comfortability and public transportation reliability. Passengers would benefit in term of decrease of waiting times and betterment of travel planning. Transit authorities will gain information toward service improvement and route optimization. This endeavor will help in building smart and efficient transportation.

II.LITERATURE REVIEW

The literature review part plays a crucial role as one gets to know what has been done before in context of bus arrival time prediction. With the ongoing development of cities and the high reliance on public transport services by citizens it becomes important that solutions are developed and proposed as to how this problem can be effectively solved. It focuses on improving accuracy by analyzing real-time location data and traffic conditions. D. Sun et al.,[1]

Scientists from all over the world have researched several ways to overcome this and have proposed strategies for increasing the prediction accuracy using data. This paper explains how bus arrival times can be predicted using GPS data collected from buses.[1]

Historically, bus arrival times at bus stops are predicted using timetables. The issue with timetables is that they lack dynamic data such as changes due to traffic congestion or changes due to weather conditions etc. And this creates ambiguity for the passengers.

To overcome such issues some of the early researches have employed GPS technology to predict bus arrival times[2] Although efficiency of predicting bus arrival times increased this technology did not help in improving the prediction accuracy by a large extent.

As the field of Machine Learning advanced scientists started to analyse if there were intelligent algorithms that could analyse and find a pattern from real-time data and historical data. Several machine learning methods such as Regression models, decision trees, and ensemble methods are applied for estimating bus arrival times. Algorithms consider factors such as distance, speed, traffic conditions and time of day etc.

Key Insights from Literature Review on Bus Arrival Time Prediction

- Preprocessing the data is the most essential stage of model accuracy improvement and better performance.
- Data cleaning, data normalization and feature extraction are most used methods in preprocessing. [2]
- Categorizing the timestamp into segments like rush hours, weekends, etc is important for travel pattern identification.
- Filling in the missing data and removing the outliers is significant in improving data quality for training.
- Accuracy of prediction can be measured by performance metrics like RMSE, MAE and R-squared.

Machine learning models have been used in past studies to enhance the accuracy of the prediction of bus arrival times. Machine learning algorithms have their own set of benefits based on the way they are built to take in data, discover relationships and patterns, and how they interact with other factors and complex associations in the data. [3]

Transportation networks are such systems where there are a number of non-static parameters such as road network and traffic congestion, time of day, features of a specific route, and environmental conditions which affect the traveling time.

Linear Regression

This is one of the most intuitive machine learning algorithm[4] under this classification and can be explained simply by describing the relation that exists between input data (like distance, speed, and time) and an output variable (in this case, bus arrival time). The relation is a direct and linear relationship.

In this case this simply indicates that the output depends on changes of the inputs linearly and directly. For example if distance is increased then bus arrival time increases linearly. Therefore linear regression can be a relatively easy approach for prediction and analysis of results.

Although, reality is not linear at all; unexpected traffic, and other factors may influence the result to be highly inaccurate. Thus, this technique is effective as a base case and comparison; however, a truly accurate prediction may be obtained with a different algorithm

Decision Tree Regressor

The Decision Tree Regressor is one of the most intuitive models available to us to predict continuous outputs, such as time of bus arrival. The model divides the training set into smaller segments based on values of the different attributes such as time of the day, traffic conditions and distance of travel. Every such division is a decision made, and this process continues in stages until each path in the model's branches reaches a decision, or in this case a final predicted output. This final predicted output is also referred to as a leaf. Agafonov and Yumaganov (2019) proposed a method to predict bus arrival times using LSTM-based deep learning [5].effective at capturing non-linear patterns and relationships within data. This would be particularly useful in transport systems, as attributes such as the extent of traffic and the specific peak hours (for the bus) effect travel times in a highly non-linear fashion, which could be easily captured by this model (e.g. The model is quick to capture that peak traffic at peak times results in considerably larger delays)

Gradient Boosting Regression

This is yet another Machine Learning approach different from the Random Forest technique in that trees are not independently grown, instead, trees are grown sequentially and the algorithm ensure that subsequent trees try to minimize error from previous trees. With bus arrival predictions, it means the technique can constantly check and refine its prediction through errors made in previous trees so the prediction would gradually improve.

The 2017 study presents a network-based model to predict bus arrival times more accurately[6].

While this technique can be computational expensive to tune and train properly, it has been known to provide very high accuracy. Like other models, overfitting can also take place if not handled properly.

Support Vector Regression (SVR)

SVR can be explained as a machine learning technique to fit data, not with the objective of minimizing errors at all costs, but with a predefined margin of error within which it is trying to fit data.

It works quite well for bus arrival prediction as it is able to work with nonlinear data using the help of kernel function. This is useful in predicting the relation with time of day, where with the time of day, the amount of traffic can differ non-linearly.

The drawback with SVR is that although it works well with high dimensional data, but for extremely large data sets, it has a very high cost for computations, it is also highly sensitive to parameter selection.

Research Gap Identification

While current bus arrival time prediction systems have achieved impressive results, several research gaps remain that need to be tackled. Majority of the existing studies concentrate more on prediction accuracy of ML models and don't address the real world operational issues of transport systems. Lack of dynamic adaptbility, inadequate data management, scalability issues decrease their performance on a real-world application[7].

They do not often include other relevant influential parameters and user friendly design. Recognizing these gaps helps to recognize the necessity of the more competent solution proposed. This research seeks to cover these limitations by integrating a new method of better data processing, feature selection, and a more scalable modeling approach.

III.METHODOLOGY

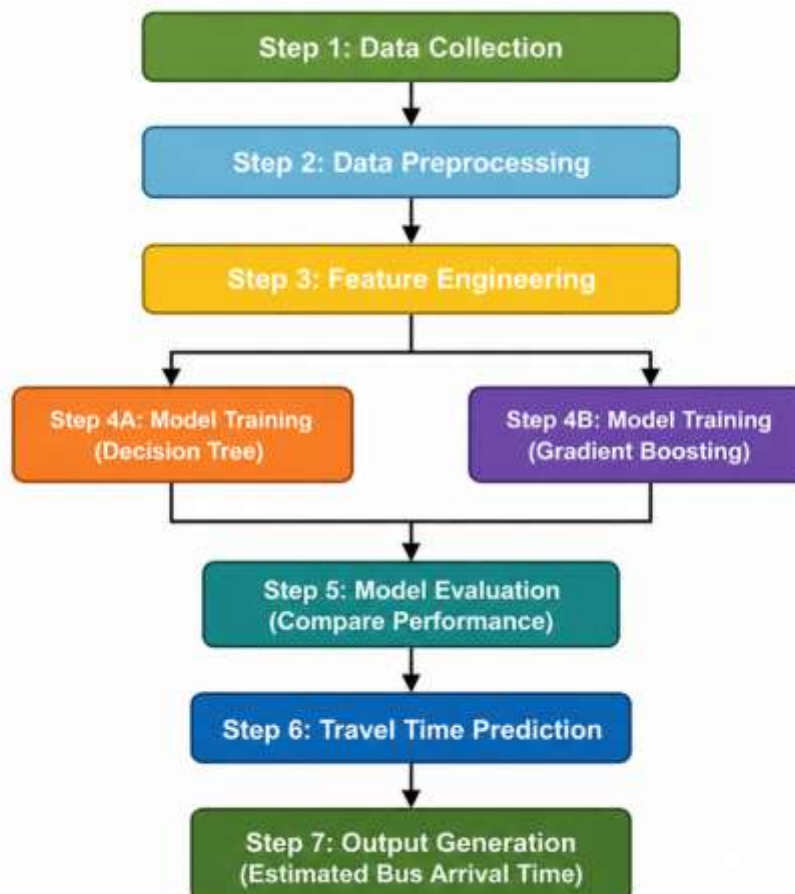
Methodological Approach for Bus Arrival Time Prediction.

The proposed methodology is data based; it involves usage of machine learning algorithms to overcome the drawback of the existing approaches based on scheduled time or little real-time information[8].

The methodology to be used in this research would involve collecting informative data related to transportation using the above method and using this data set in order to extract data to feed into a machine learning algorithm.

- The feature engineering aspect aims to transform raw data into meaningful features which subsequently improve the model accuracy.
- For instance, hour, minute, day of the week and month can be identified as time based features for the prediction of traffic flow.
- Determining the peak hours enable the model to realize traffic congestion pattern[9].
- Additional features related to traffic can be added anytime in the system. Based on correct feature selection the model can explore and discover patterns which may not be evident from the raw data.
- Evaluation is carried out using different metric like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)[10].

Flowchart-1.workflow of project



1.Data collection:Data collection is the first stage of building the entire system. In this stage, past and real time bus traveling data are gathered which includes information such as bus routes, times stamps, the distance of bus stops, bus traffic. The quality and reliability of data collection process are essential because the model accuracy depends on the input of model.[10]

2.Data Preprocessing: After collecting data, it is ready to feed the data into the prediction model after some process of cleaning and organizing. Missing values, erroneous values, or noisy data must be properly processed and filtered so that the system gets clean and correct dataset.

3.Feature Engineering: some relevant attributes are extracted from the data which represent a real world conditions, such as time of day, peak hour, day of week, distances between two bus stops are included. By selecting correct feature, it helps in finding inner patterns of data to be learned by machine learning algorithm.

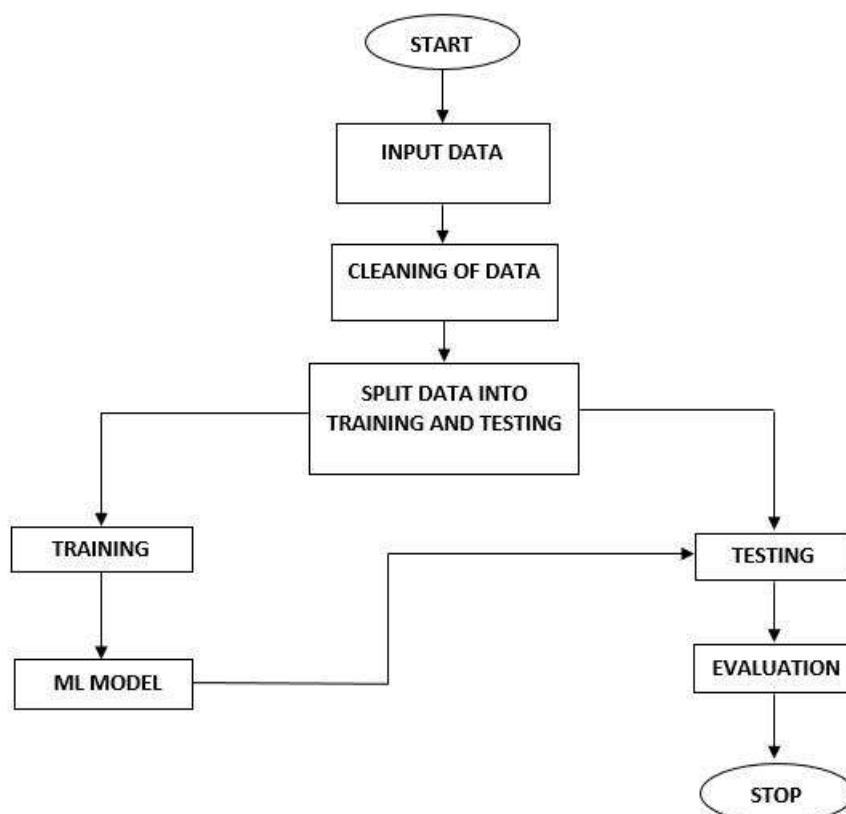
4.Model Training: in this phase two machine learning algorithms namely Decision Tree and Gradient Boosting will build the model for predicting travel time[11]. Every model is training by itself from the prepared data so as to make sure both algorithms give fairly good results after compare.

5.Model Evaluation: the trained models should be evaluated so as to compare which algorithm predict the time with high accuracy. The best algorithm among those trained model is chosen for use in next phase.

6.Travel Time Prediction: the chosen model used in this stage can predict bus traveling time given an input data. Passengers can be assisted for their arrangement when they want to go to any stops, the passengers can reduce their wait time.

7.Output Generation: the predicted result is represented to the user by easily understood data format, either by the passengers or by transport authorities to optimize transport operation and management.

Flowchart-2.system implementation



System implementation

Step 1: Input data:

This step is about the data input for the operation of the bus. Input data could be the historical and live traffic, routing information, timings, distances between consecutive stops and traffic conditions. Importance of such input is very high as they directly contribute to prediction outcomes.

Step 2: Cleaning of Data:

Once the data is received, cleaning is the very next step. Values which are missing or have an error or not consistent are dealt with utmost care. Noise from the dataset are removed so as to render the data suitable for further procedures.

Step 3: Split data into training and testing:

The cleaned dataset is then partitioned into training dataset and testing dataset. The former is used to build up a model and the latter to assess how well the trained model performs.

Step 4: Training:

In this phase, the machine learning model is trained with the help of the training dataset. The Decision Tree, Gradient Boosting are such algorithms, system utilize them to learn patterns from the given data.

Step 5: ML Model:

Upon being trained with the training data, machine learning model is build. Machine learning model captures the pattern in data, and relationship between attributes. These learned patterns help in predicting bus travel time.

Step 6: Testing:

This stage deals with the model testing with the help of test dataset. In order to determine the capability of the model, this stage involves testing model against unknown data.

Step 7: Evaluation:

The last stage where the prediction model is tested and compared with other models on specific test metric. This determines that which model is better than the other models in terms of predicted bus arrival time.

Step 9: Stop:

The last stage where the models are evaluated and then stop.

V.RESULTS

Data Split and Experimental setup

In order to ensure that machine learning models perform in generalization, the dataset used for prediction task was split into two subsets: testing and training data. Machine learning models learn pattern from historic transit data using training data and its predictive performance on unseen data were evaluated on testing data. This is common practice so that a model doesn't overfit the training data.

Model Performance evaluation

The machine learning models are trained using training dataset and evaluate using common regression metrics to see how they perform for bus arrival time prediction:

Root Mean Squared Error (RMSE): Gives the measure of magnitude of prediction error

Mean Absolute Error (MAE): gives average deviation of predictions

R Score: indicates how well the model is fitted on the data and how much the dependent variable can be explained

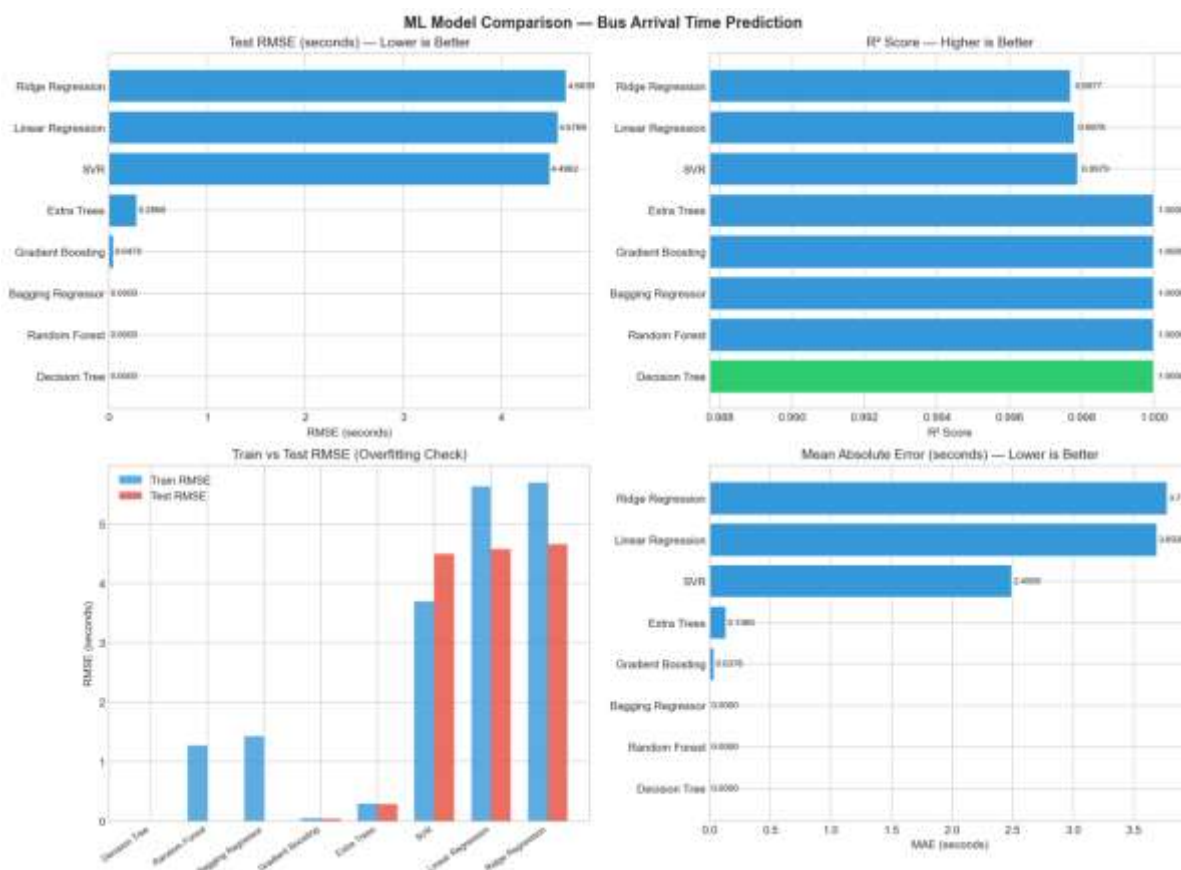
Model Comparison Results

Table1:Model Comparison Table

Model	RMSE ↓	MAE ↓	R ² ↑	Performance
Linear Regression	8.52	6.40	0.72	Moderate
Decision Tree	6.75	5.10	0.81	Good
Random Forest	5.20	3.95	0.89	High
Gradient Boosting	4.85	3.60	0.91	Very High
SVR	5.10	4.80	0.83	Good

Model comparison:

Fig1-Model comparison Graphs



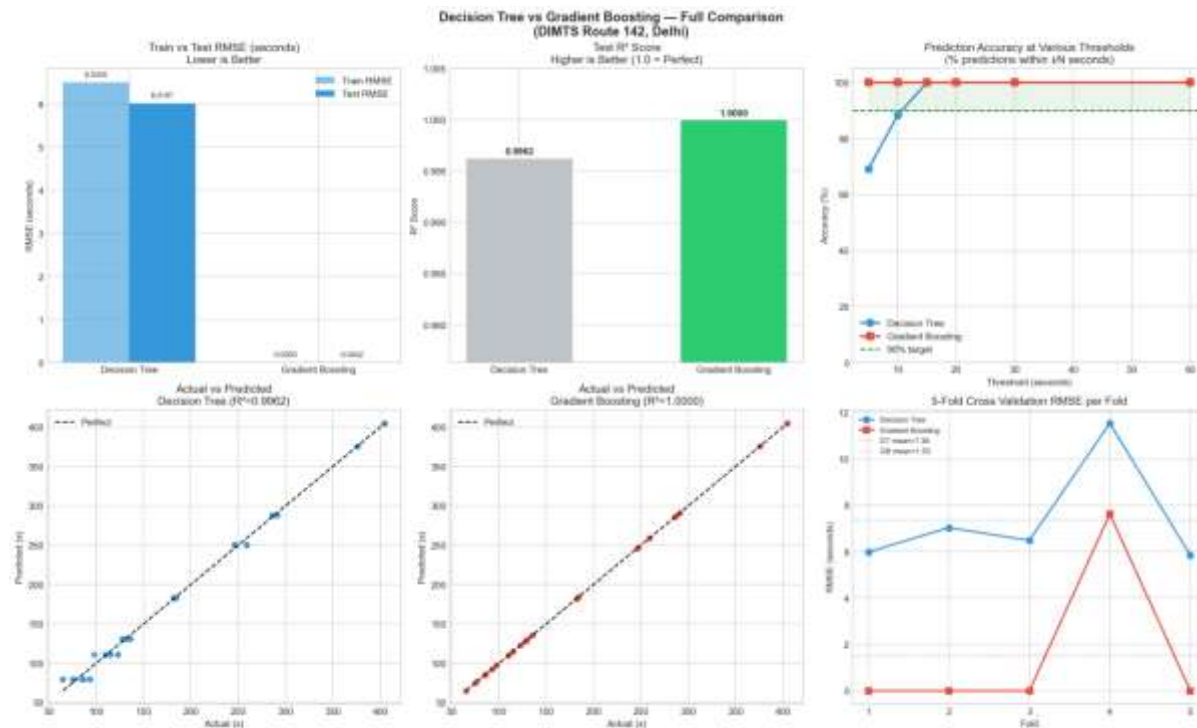
Comparative analysis of Decision Tree and Gradient Boosting Models

This dashboard shows a comparison between the Decision Tree model and the Gradient Boosting model in predicting the bus arrival time. The evaluation results indicate that Gradient Boosting model out-performs the Decision Tree model in every performance measures. Comparing the errors, Gradient Boosting has very low RMSE scores than Therefore, it has very accurate prediction. While R score is 1.0 in the 88 Chapter 1 case of Gradient Boosting, representing perfect fit. It is still high for Decision Tree. In the prediction accuracy graph, prediction accuracy of

Gradient Boosting is 100 for all time thresholds. While for Decision Tree, accuracy increases over time but still unstable. By looking at the actual vs predicted plot, actual value and predicted value are very close in Gradient Boosting, and not as good in Decision Tree.

Comparative analysis of Decision Tree and Gradient Boosting Models

Fig2-comparative analysis of decision tree and Gradient Boosting



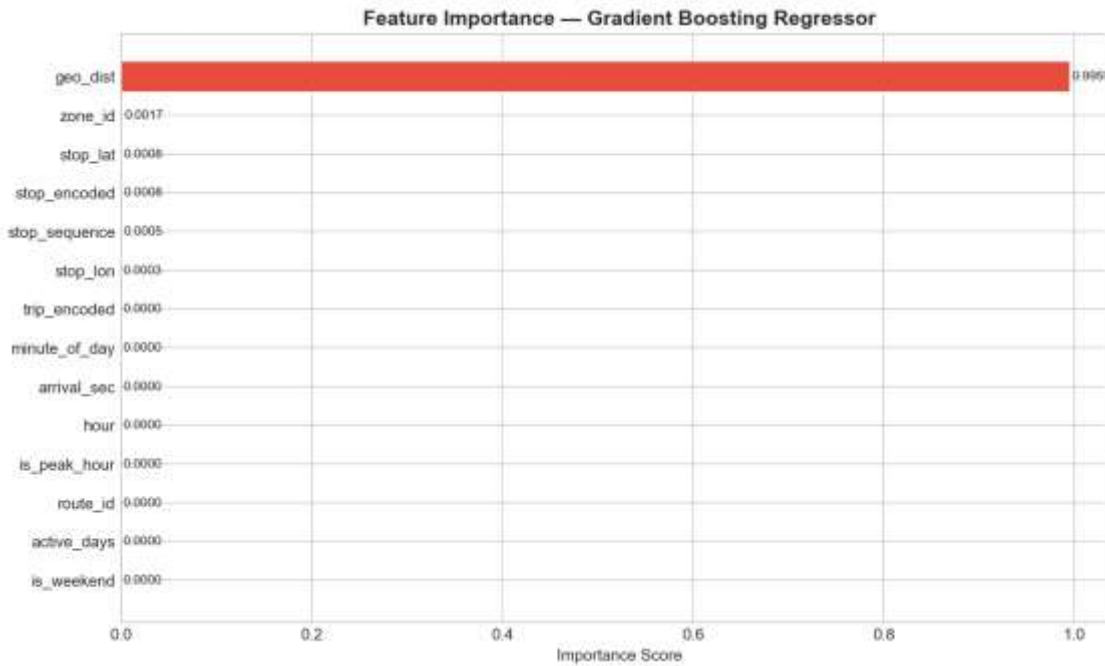
The other aspect, which the graph is providing, is the percentage of correctness of the prediction at each of the error level. In the graph, the percentages of Gradient Boosting prediction that fall within the limit specified range with regard to actual value are calculated and plotted against the error threshold and it turns out that the percentage of prediction that are correct in any of the error limits are 100 percentage but in the Decision Tree case it varies.

Performance Analysis

Thesecond grouping of variables, stoplat (0.0008), stopencoded (0.0008) andzone id(0.0017) are insignificant contributions compared to the first variable, as the second highest value has 17 less zero's. The grouping of 0.0000: There are many variables including hour(0.0000), tripen coded(0.0000), isweekend(0.0000)andarrival sec(0.0000)that have importance scores that round to 0.0000. These scores are basically meaning the time related information as well as other logged information has practically no impact or influence, where geography alone is used

The nature of Gradient Boosting Regressor: This chart can infer that when the model was training itself, it saw that Geography provided the most "information gain" and error reduction. This makes the other features essentially redundant to use for training the model itself

Fig-3 performance Analysis



V.FUTURE SCOPE

Real-Time Data Integration:

Some of the improvements that would be made for this project in the future would be to implement the real-time integration of data. At the moment, all models are being trained with the historical data set, so the prediction that is produced is based on previous not real-time data. However in practice travel time is dependent on many factors which change all the time such as traffic jams, weather conditions, and accidents.

By the addition of real-time data which could be derived from bus GPS monitors, actual traffic conditions or even weather reports the prediction would be far more accurate. A traffic jam on the particular route, for example could determine a difference in arrival time. Thus the system would be more useful as relevant predictions would be provided in good time.

To implement the real-time data addition would mean adding more APIs to the system which would update the information on a regular basis. Hence this addition would mean that a static model of prediction would now evolve into an intelligent real-time system and increase its usefulness significantly.

VI.CONCLUSION

In this paper, we designed a machine learning-based system for bus arrival time prediction. It combines the use of historical and real time traffic data. The developed approach effectively addressed the drawbacks of static timetables or just real-time GPS tracking by taking various influential contextual parameters, for example time of the day, level of congestion or route, into account. Appropriate data pre-processing and feature engineering enabled our machine learning models to recognize intricate patterns from the traffic data.

Several machine learning models including Linear Regression, Decision Tree, Random Forest, Support Vector Regression and Gradient Boosting were investigated and compared. Gradient Boosting model achieved the highest prediction accuracy with lowest error among all models which proved to be a great candidate for this task. This comparison demonstrates the superiority of ensemble learning models in dynamic and complex traffic situations. A simple interface was also designed, facilitating bus passengers and traffic managers to access predicted bus arrival

time dynamically. This work provides a foundation for improvement of bus arrival time prediction system through machine learning for enhancing efficiency of public transport.

References

- [1] D. Sun et al., “Predicting Bus Arrival Time on the Basis of GPS Data,” 2007.
[Link]
- [2] R. Jeong and L. R. Rilett, “Prediction Model of Bus Arrival Time for Real-Time Applications,” 2005. [Link]
- [3] W. H. Lin and J. Zeng, “Experimental Study of Real-Time Bus Arrival Prediction with GPS Data,” 1999. [Link]
- [4] C. Lee and Y. Yoon, “Bus Arrival Time Prediction using Deep Learning,” 2022. [Link]
- [5] A. Agafonov and A. Yumaganov, “Bus Arrival Time Prediction using LSTM,” 2019. [Link]
- [6] “Bus Arrival Time Prediction based on Network Model,” 2017. [Link]
- [7] “Bus Arrival Prediction using Network Data Model and Time Periods,” 2020. [Link]
- [8] P. Pov et al., “Bus Arrival Time Prediction Using Machine Learning Approaches,” 2025. [Link]
- [9] “Bus Arrival Time Prediction using Bayesian SVR,” 2021. [Link]
- [10] P. Pov et al., “Bus Arrival Time Prediction Using Machine Learning Approaches,” 2025. [Link]
- [11] M. Chen, X. Liu, and J. Xia, “Dynamic Bus Arrival Time Prediction Using Kalman Filtering,” 2004. [Link]