Bus Travel Time Prediction: A Comparative Study of Linear and Non-Linear Machine Learning Models

Ashwini B P¹, R Sumathi¹, H S Sudhira²
¹Department of Computer Science and Engineering, Siddaganga Institute of Technology, Tumakuru Karnataka, India-572103
²Director, Gubbi Labs, Tumakuru, Karnataka, India- 572216
ashvinibp@sit.ac.in

Abstract. Congested roads are a global problem, and increased usage of private vehicles is one of the main reasons for congestion. Public transit modes of travel are a sustainable and eco-friendly alternative for private vehicle usage, but attracting commuters towards public transit mode is a mammoth task. Commuters expect the public transit service to be reliable, and to provide a reliable service it is necessary to fine-tune the transit operations and provide well-timed necessary information to commuters. In this context, the public transit travel time is predicted in Tumakuru, a tier-2 city of Karnataka, India. As this is one of the initial studies in the city, the performance comparison of eight Machines Learning models including four linear namely, Linear Regression, Ridge Regression, Least Absolute Shrinkage and Selection Operator Regression, and Support Vector Regression; and four non-linear models namely, k-Nearest Neighbors, Regression Trees, Random Forest Regression, and Gradient Boosting Regression Trees is conducted to identify a suitable model for travel time predictions. The data logs of one month (November 2020) of the Tumakuru city service, provided by Tumakuru Smart City Limited are used for the study. The time-of-the-day (trip start time), day-of-the-week, and direction of travel are used for the prediction. Travel time for both upstream and downstream are predicted, and the results are evaluated based on the performance metrics. The results suggest that the performance of non-linear models is superior to linear models for predicting travel times, and Random Forest Regression was found to be a better model as compared to other models.

1. Introduction
Increased usage of private vehicles has led to several problems such as congested roads, increased greenhouse gas emissions, increased travel time, affecting the overall quality of living. Using public transit system is a sustainable and eco-friendly alternative for private vehicle usage, but attracting commuters towards public and mass transit modes of travel is a major challenge for the government and the related public sectors. Commuters expect a reliable [1][2] transit system, and providing such a service is not an easy task in an Indian traffic scenario. Tier-1 cities in India have local rails, metro trains, and buses for mass transit, whereas tier-2 cities have limited infrastructure and facilities, and buses are the major mass transit system used here. Though the population density of tier-2 cities is less as compared to tier-1 cities, the total population of tier-2 cities is high, as there are more than 100 such cities in India. With such a huge population using mass transit services, research on improving the transit service in these cities is of prime importance.
With the Smart City Mission being launched by the Government of India, 100 cities are identified for being transformed for providing a citizen-friendly and sustainable environment. Tumakuru is one of the cities identified under Smart City Mission, and providing reliable transit service is one of the goals of Tumakuru Smart City Limited (TSCL). The public transit buses of Tumakuru city are equipped with GPS trackers to provide necessary information to the transit operations planners and TSCL has shared the data logs for this study.

Forecasting travel times are very useful in managing the live transit operations and also the operations planners can use these forecasts to fine-tune the transit operations in the future. The main objective of this comparative analysis is to forecast bus travel time on selected routes in Tumakuru city using Machine Learning (ML) regression models. Four linear models namely, Linear Regression (LR), Ridge Regression (RR), Least Absolute Shrinkage and Selection Operator Regression (Lasso), and Support Vector Regression (SVR), and four non-linear models namely k-Nearest Neighbors (K-NN), Regression Trees (RT), Random Forest Regression (RFR), and Gradient Boosting Regression Trees (GBRT) are chosen for this study. As per the knowledge of the authors, this is the first attempt to predict the Bus Travel Time of Tumakuru city, so an empirical study is carried out to identify the suitability of various ML models. A comparative analysis is conducted using error metrics such as Mean Absolute Error (MAE), Median Absolute Error (MdAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and R squared (r² score) to identify a suitable model.

The remaining sections of the paper are presented in the following order. Existing research work in the area of bus travel time prediction of public transit buses is summarised in section 2. The method followed for the study, and each phase of the study namely, data, modeling, model evaluation, and model selection are discussed in section 3. Finally, in section 4, the insights of the study and the future scope are discussed as a conclusion.

2. Literature Review
The bus travel time prediction has been a well-researched topic in the past. Various researchers have predicted bus travel time using mathematical, statistical, and ML-based models. Among ML models linear, non-linear, and neural network-based models have been widely used. Researchers have also compared the performance of various models and derived the location-specific suitability of the model. Generally, predicting bus arrival time involves forecasting bus travel time and hence in this literature review, they go hand in hand.

Researchers in the past have used a variety of input data from different sources [3], such as Automatic Vehicle Location (AVL), manually collected, surveys, mobile phone footprints, and social media data, etc. The emerging big data technologies [4][5][6] and their applications in public transit and traffic have set a platform to provide data solutions to related problems. Among the available sources of data, Automatic Vehicle Location (AVL) [7][8] data are the most widely used as compared to others. The existing research on bus arrival time and bus travel time predictions using data logs are summarized in this section.

Simple averages of historic data are one of the basic models for bus travel time estimation and prediction, and few authors have [7][9] fine-tuned and also combined the historic averages with other models, whereas few others have compared their proposed model results with the historic averages [10][11]. Authors in [12] have presented a model-based method using real-time field data of a study route in Chennai India to predict bus arrival time. They have predicted the running travel time and delay of the bus to predict the overall arrival time of the bus at each bus stop in the selected route. And the proposed model demonstrated superior results as compared to the model deployed on the field.

Kalman filters [13] is the commonly used statistical model for bus travel time prediction. Kalman filtering algorithm estimates the required unknown variable using series of records (though noisy) priorly observed over some time. AutoRegressive Integrated Moving Average (ARIMA) [14][15]
models are the most commonly used models for time series forecasting, like forecasting bus travel time. In [11][16], authors have attempted to predict the bus arrival time at Houston, Texas using Historic averages like a model and Artificial Neural Network (ANN) model, and compared the accuracy of models to conclude that the neural network model performed better than the other models. Authors in [17] have proposed a Deep Belief Network (DBN) model to forecasts the arrival time of buses and compared its performance against the basic models such as k-NN, ANN, SVM, and random forests, and the DBN performance showed superior results.

In [18], authors have proposed a hybrid model to predict bus arrival time at Dalian, China, using Kalman Filtering and SVM. The SVM predicts the base travel time using historic data and Kalman Filtering uses the base travel time and the latest travel time information to predict bus arrival time at the next bus stop. The proposed hybrid model exhibited superior results as compared to an individual ANN model. Recently authors have used cutting-edge technologies and advanced models like Long Short Term Memory (LSTM), Recurrent Neural Networks (RNN) [19][20], etc, for bus travel time prediction.

3. Methods
The pictorial representation of the study framework is given in figure 1. The study is carried out in four main phases namely; Data, Modeling, Model Evaluation, and Model Selection. A detailed description of each phase is presented in this section.

![Comparative analysis framework](image-url)
3.1. Data

This is the initial phase of the study, in which tasks such as data collection, data pre-processing, and preliminary analysis are conducted.

3.1.1. Data Collection

Tumakuru is a tier-2 city of India located in Karnataka, a southern state of India. One of the cities being identified under the Smart city mission of India is observing overall development recently. The major commuters in the city take a private mode [21] of transport, especially two-wheelers. In 2011 through an initiative of the Directorate of Urban Land Transport [22], Government of Karnataka, India, the city bus service was started in Tumakuru city. Currently, the service has a fleet of 50 buses, all equipped with GPS tracking systems that are connected to a common control center maintained by Tumakuru Smart City Limited. The Tumakuru city service consists of 15 routes of varying lengths from 5.5 km to 15 km. The data logs of November 2020, bus-stop location information, and the schedule of the trips are provided by Tumakuru Smart City Limited for this analysis. A snapshot of the data collected is given in table 1.

| Vehicle-regno | Device-id | received_date | latitude | longitude | speed | odometer | location                      |
|---------------|-----------|---------------|----------|-----------|-------|----------|-------------------------------|
| KA-06-F-0723  | 8.69E+14  | 05-11-2020 09:32:51 | 13.34277 | 77.09883 | 5     | 118.64  | Tumakuru Bus Stop             |
| KA-06-F-0723  | 8.69E+14  | 05-11-2020 09:34:23 | 13.33834 | 77.1081  | 16    | 119.6   | Near By Mirji Complex         |
| KA-06-F-0723  | 8.69E+14  | 05-11-2020 09:34:31 | 13.33804 | 77.10837 | 24    | 119.65  | Badramma chowltry              |
| KA-06-F-0723  | 8.69E+14  | 05-11-2020 09:35:50 | 13.33241 | 77.10859 | 22    | 120.08  | Shree lakshmi complex          |
| KA-06-F-0723  | 8.69E+14  | 05-11-2020 09:36:00 | 13.3336  | 77.10857 | 24    | 120.15  | Near By 12th Cross Road        |
| KA-06-F-0723  | 8.69E+14  | 05-11-2020 09:36:03 | 13.334   | 77.10857 | 25    | 120.17  | Dr Radhakrishnan Rd           |
| KA-06-F-0723  | 8.69E+14  | 05-11-2020 09:36:10 | 13.33292 | 77.10857 | 29    | 120.22  | Dr Radhakrishnan Rd           |
| KA-06-F-0723  | 8.69E+14  | 05-11-2020 09:36:20 | 13.33222 | 77.10885 | 22    | 120.3   | Near By 12th Cross Road, SS Puram |
| KA-06-F-0723  | 8.69E+14  | 05-11-2020 09:36:23 | 13.33208 | 77.10892 | 20    | 120.32  | Upparahalli Gate Bus Stop      |

**Table 1.** Data logs(sample) of a Tumakuru city service bus

**Route Number** 205 - Tumakuru bus stand (TBS) – Goolarive (GO) (upstream and downstream) is selected for the analysis. This route is 5.5 km in length, and 15 minutes is the allotted travel time according to the schedule. The route map along with the bus-stop location is shown in figure 2.

![Figure 2. Route map of 205](image-url)
3.1.2. Data Pre-processing
The raw data logs are noisy, erroneous, and redundant, and need to be pre-processed before applying ML techniques to predict travel time. The pre-processing is done as:

Step 1: The data logs of the study route are selected from the collected data.

Step 2: The logs during the on-trip and off-trip standby are redundant, so discarded for the analysis.

Step 3: The data is split trip-wise and transformed into trip-wise aggregates, based on the start time.

Step 4: Outliers are detected using Absolute Deviation (AD) i.e., the absolute distance from the mean. Consider there are \( n \) instances in the dataset, each instance say \( x_i \) has a set of features, among which travel time \( (tt_i) \) is the target variable. The AD \( (a_i) \) of each \( tt_i \) is estimated using the equation (1), where \( \mu_{tt} \) is the mean travel time of all the instances. In this study, 20% deviation is considered as a threshold to identify outliers, and all \( x_i \) deviating more than 20% from the mean travel time is eliminated for the study using the equation (2)

\[
a_i = |tt_i - \mu_{tt}| \forall i = 1 \text{ to } n
\]

\[
\{a_i > \mu \times 0.2 \mid x_i \in \text{dataset} \}
\]

\[
\{a_i \leq \mu \times 0.2 \mid x_i \notin \text{dataset} \}
\]

By applying these steps data is prepared for analysis. Around 600 trips on all days and different trip start times are considered for this analysis. The day-of-the-week (day) and departure times (start), the direction (up-down) are the independent variables and travel time \( (tt) \) is the dependent variable. The pre-processed data is divided into the training set and the test set. 70% of the data is considered for training and 30% to test the models.

3.1.3. Preliminary analysis
A preliminary analysis is conducted to understand the correlation of the independent variables and dependent variables. A heat map using the seaborn library in python is plotted. A heatmap is a visualization tool that presents the data in a matrix with color encoding. The heat map generated is shown in figure 3, and it is observed that there is a very low correlation among the independent variables. The trip start time (start) is positively correlated with a score of 0.16, day-of-the-week (day), and the direction i.e. upstream or downstream (up-down) are negatively correlated to the dependent variable, travel time \( (tt) \) with a score of -0.056 and -0.76 respectively. It can be inferred that the departure time (start) i.e., the time-of-the day has a high impact on travel time than the day-of-the-week and the direction of the travel.

![Figure 3. Correlation matrix of independent and dependent variables](image-url)
3.2. **Modeling**

The travel time prediction for Tumakuru city is conducted for the selected route number 205 using four basic linear (LR, Lasso, RR, and SVR) and four non-linear models (k-NN, RT, RFR, and GBRT). The working and training of the models are discussed in this section.

### 3.2.1. Linear Models

Linear ML models generate a linear mathematical equation to estimate the unknown values. Linear models are relatively simple, quick, and straightforward to interpret, which can be very advantageous for suitable data and applications. A simple linear model with two independent variables is given in equation (3).

\[ y = c_0 + c_1 * z_1 + c_2 * z_2 \]  

Linear regression is a basic regression model in statistics and ML, which estimates the dependent variable using a linear combination of the independent variable(s). At times, a simple linear regressor overfits the data; to overcome this regularization has to be applied. Regularization [23] is a method to decrease the errors by fine-tuning the coefficients appropriately on the training data. There are two types of regularizations, L1 and L2. In L1, binary coefficients are applied to choose features of importance in a high dimensional dataset, Lasso regression is a classic example of the L1 regularization technique. In L2, the coefficients are dispersed, i.e., the significance of variables is considered, and insignificant variables are penalized which leads to more accurate customized final models. Ridge Regression is L2 type. Support Vector Regression fits a bar, unlike linear regression that fits a line to minimize the error between the actual and the predicted values. The errors that lie within the margins of the bars are neglected.

### 3.2.2. Non-Linear models

Non-Linear ML models generate mathematical functions that are not linear, like, higher degree polynomials, square roots, and trigonometric functions, etc. The applications where the data is not linear, the performance of linear models will be very low and in those cases, non-linear models showcase better results.

k-NN Regression is also known as a lazy learner, it is a basic non-linear model which works on the principle of similarity. It estimates the difference between the training set instances with that of the test set instances using the distance formula (quadratic equation). The most commonly used distance metric is Euclidean distance, given in equation (4).

\[ \text{distance} = \sqrt{(x2-x1)^2 + (y2-y1)^2} \]  

In equation (4), there are two instances say a and b with features x1,y1, and x2,y2 respectively, the distance between them is estimated using the given formula, the lower the distance, the higher the similarity, and vice versa. A Regression Tree also commonly known as a decision tree, is a non-linear ML model applied for predicting a continuous variable. In tree-based models, there are no mathematical equations to demonstrate the relationship among the multiple variables. In decision trees, the scheme is to construct a model that forecasts the target variable by learning through deduced decision rules, which are extracted from the feature values of the training set. Random forest regressor is one of the modern ensemble models for regression. In an ensemble model, better predictions are made as compared to a single model (RT) as it combines the results of multiple ML models. RFR builds multiple regression trees and combines the results of each tree for stable and accurate predictions.
Gradient Boosting Regression Tree is another ensemble model similar to RFR. It is a stage-wise additive tree model in which at each level a weak ML tree model is applied to minimize the error in the prediction until a threshold error (near to zero) is reached. It follows a greedy approach, i.e., at each level of the tree it chooses the best weak learner that reduced the error. All the eight ML models discussed so far are fitted to the training data set and applied to forecast the time of travel using a common test set. The experiment is conducted by implementing the Python code in the Jupyter notebook.

3.3. Model Evaluation and Comparative Analysis
The predicted travel times of each model are compared against the actual travel time in the test set. The difference in the predicted and the actual values are the errors. This difference can also be observed pictorially by plotting a scatter plot of actual vs the predicted values along with a line at 45 degrees. If the actual vs predicted values are the same, i.e., if the error is zero then the points lie on the line, if the errors are greater than zero, the extent of the distance of each point from the 45-degree line indicates the amount of error. The scatter plots of all eight models are shown in figures 4-11, where the x-axis is the actual and the y-axis is the predicted values for the travel time in seconds. It is observed that the results of the linear model are scattered while the nonlinear models are around the line. Also, the Linear regression model and regularized linear models, Lasso, Ridge, and SVR; show descent results indicating that linear models are less suitable as compared to non-linear models.
The performance of each ML model is assessed based on the performance metrics. An important aspect of using errors in the performance metrics is their ability to discriminate among model results. There are several error metrics for regression models, among them, MAE, MdAE, RMSE, and MAPE, are chosen in this study. R squared (r² score), a regression score function, known as the coefficient of determination is a commonly used evaluation metric for regression models, is used to compare the performance of ML models.

The errors estimated based on MAE, MdAE, and RMSE are measured with the same units as that of the target variable; in this study it is seconds. The graph in figure 12 shows the errors of each model. A trend line is drawn to compare the performances of the selected models, and it is observed that the performance of non-linear models is superior to that of linear models. The predictions made by RT and RFR have the least errors.
The MAPE is expressed in percentage and the MAPE values of all the models are plotted in the graph shown in figure 13. The percentage errors of the models are in the range of 0-8. The R squared (r² score) values are presented as a graph in figure 14. The results emphasize that the performance of non-linear models is better as compared to linear models.

![Mean Absolute Percentage Error (MAPE)](image1)

**Figure 13.** Mean Absolute Percentage Error (MAPE) of all the models under comparison

![r² score](image2)

**Figure 14.** r² score of all the models under comparison

3.4. Model Selection

The travel time data exhibit spatial-temporal characteristics. It depends on the spatial aspects such as route parameters (land use pattern, distance from the city center, etc.) and direction of the travel, also temporal aspects such as the time-of-the-day, the day-of-the-week, and peak-off-peak hours, etc. High variability in travel time can be observed based on the aforementioned spatial-temporal characteristics, and this variability is observed to be non-linear, hence fitting a flexible non-linear model (curve) is showing better results as compared to a linear curve or a straight line.

Overall, it can be inferred that non-linear models are suitable for predicting public transit travel time in Tumakuru city. Among the four non-linear models, especially an ensemble model, i.e., the random forest regression suits the best.
4. Conclusion and Future Scope
In this work, the performance of linear and non-linear ML models for predicting travel time are compared. Private vehicles have increased recently and this has lead to problems such as congestion, green gas emissions, and increased travel time. As a solution, public transit systems which are a sustainable alternative for private modes have to be promoted. To attract commuters towards public transit mode, the existing public transit system should be more reliable. In this context, this work aims to predict travel time in Tumakuru, a tier-2 city in India, as per the knowledge of the authors this is the first of such work for Tumakuru city. As this is an initial attempt, the performance of eight ML models including four linear and four non-linear are compared to identify a suitable model for forecasting bus travel time. Based on the error metrics such as MAE, MdAE, RMSE, MAPE, and R squared (r² score), it is concluded that the performance of the non-linear models outperformed the linear models, especially random forest regression showed the least errors of 17.99, 7.99, and 4.8 as MAE, MdAE, RMSE respectively, and highest r² score of 0.9889. Also, Random Forest Regression suits the best to be used by transit operations planners for improving the schedule and other transit operations.

This work can be further extended to predict bus arrival time at various bus stops on the routes and incorporate the results in the Passenger Information System (PIS). In PIS, ML models need to predict travel times dynamically, and models with less time complexity have to be identified. Random Forest is a complex model and is slow as compared to other models. An alternative model for an application like PIS has to be identified.

Acknowledgments
The necessary data logs of the public transit buses for this analysis are provided by Tumakuru Smart City Limited. The authors are thankful to the Director of TSCL and the technical team for their continuous support.

References
[1] Soza-Parra J, Raveau S and Muñoz J C 2021 Public transport reliability across preferences, modes, and space Transportation (Amst). p 1–25
[2] Zheng F, Li J, van Zuylen H, Liu X and Yang H 2018 Urban travel time reliability at different traffic conditions J. Intell. Transp. Syst. Technol Planning p 106–20
[3] Ashwini B P and Sumathi R 2020 Data sources for urban traffic prediction: A review on classification, comparison and technologies Proc 3rd Int. Conf. Intell. Sustain. Syst. ICISS p 628–35
[4] Chen Y, Guizani M, Zhang Y, Wang L, Cresp N and Lee G M 2019 When traffic flow prediction and wireless big data analytics meet IEEE Netw. vol. 33(3) p 161–7
[5] Zhu L, Yu F R, Wang Y, Ning B and Tang T 2018 Big Data Analytics in Action p 266–94.
[6] Liu D, Sun J and Wang S 2020 BusTime: Which is the Right Prediction Model for My Bus Arrival Time? 5th IEEE Int. Conf. Big Data Anal. ICBDA p 180–5
[7] Singla L and Bhatia P 2016 GPS based bus tracking system IEEE Int. Conf. Comput. Commun. Control IC4
[8] Mazloumi E, Currie G and Rose G 2010 Using GPS data to gain insight into public transport travel time variability J. Transp. Eng. vol 136(7) p 623–31
[9] Xu H and Ying J 2017 Bus arrival time prediction with real-time and historic data Cluster Comput. vol 20(4) p 3099–106
[10] Kumar B A, Vanajakshi L and Subramanian S C 2018 A hybrid model based method for bus travel time estimation J. Intell. Transp. Syst. Technol. Planning vol 22(5) p 390–406
[11] Jeong R and Rilett L R 2005 Prediction model of bus arrival time for real-time applications Transp. Res. Rec. p 195–204
[12] Padmanaban R P S, Divakar K, Vanajakshi L and Subramanian S C 2010 Development of a real-time bus arrival prediction system for Indian traffic conditions IET Intell. Transp. Syst. vol 4(3) p 189–200
[13] Bai C, Peng Z R, Lu Q C and Sun J 2015 Dynamic bus travel time prediction models on road with multiple bus routes Comput. Intell. Neurosci.

[14] Suwardo, Napiah M and Kamaruddin I 2009 Arima Models for Bus Travel Time Prediction J. The Inst. Eng. [Internet]. v 71(2) p 49

[15] Billings D and Jiann-Shiou Y 2006 Application of the ARIMA Models to Urban Roadway Travel Time Prediction-A Case Study Systems, Man and Cybernetics SMC’06. IEEE Int. Conf. p 2529–34

[16] Jeong R and Rilett L R 2004 Bus arrival time prediction using artificial neural network model IEEE Conf. Intell. Transp. Syst. Proceedings ITSC p 988–93

[17] Chen C, Wang H, Yuan F, Jia H and Yao B 2020 Bus travel time prediction based on deep belief network with back-propagation Neural Comput. Appl. [Internet]. vol 32(14) p10435–49

[18] Bin Yu, Zhong-Zhen Yang K C and Yu B 2010 Hybrid model for prediction of bus arrival times at next station. J. Adv. Transp. vol 44 p 193–204

[19] Islek I and Oguducu S G 2019 Use of LSTM for short-term and long-term travel time prediction CEUR Workshop Proc. p 2482

[20] Zhang H, Wu H, Sun W and Zheng B 2018 DEEPTRAVEL: A neural network based travel time estimation model with auxiliary supervision IJCAI Int. Conf. Artif. Intell. p 3655–61

[21] Directorate of Urban Land Transport, Urban Land Development, Department of Karnataka, India 2014 Comprehensive Mobility Plan for Tumkur Final Report.

[22] Directorate of Urban Land Transport, Urban Land Development, Department of Karnataka, India 2013 TUMKURU CITY BUS SERVICE Evaluation Report

[23] Bickel P J and Li B 2006 Regularization in statistics Test vol 15(2) p 271–344