Arrival Time Prediction for Transportation System

Advait Joshi [1], Dr Arati Deshpande [2] & Dr Emmanuel Mark [3]

[1]Student; SCTR's Pune Institute of Computer Technology, (Computer Engineering), Pune, Maharashtra, India, advaitkjoshi@gmail.com
[2]Associate Professor; SCTR's Pune Institute of Computer Technology, (Computer Engineering), Pune, Maharashtra, India, ardeshpande@pict.edu
[3]Professor; SCTR's Pune Institute of Computer Technology, (Information Technology), Pune, Maharashtra, India, emmanuelm@pict.edu

Abstract:

Public transport is a preferred mode of transport in cities. People use it to travel daily. The passengers are unaware of how much time they will require to complete the journey. Machine learning techniques can be used to predict the arrival time. This research paper showcases a comparative study of machine learning algorithms that can be applied for estimating the duration required for a mode of public transport to reach the destination. It will give the passenger an idea of the travel duration. In the proposed work, a total of 7 machine-learning regression algorithms are trained and compared. The model trained on the Ada Boost algorithm has an accuracy of 96.29%.

Keywords: Regression, Machine learning, Arrival time, transportation

1. Introduction

Public transport is a system of transportation which is a chargeable service for the use of the public. Public transport plays an important role in urban settlements. It aims to reduce traffic congestion, promote environmental sustainability, and enhance accessibility. Public transport includes buses, trains, metros, taxis, bike share, rickshaws, and a few other modes. Public transport is not only cost-effective but also beneficial to the environment. Many forms of public transport, such as electric buses and trains, contribute to reduced greenhouse gas emissions and air pollution when compared to individual car use, reducing the carbon footprint manifold. It is a boon for people who cannot drive as well as for elderly people.

Public transport in Pune city comprises PMPML (Pune Mahanagar Parivahan Mahamandal Ltd) bus, Metro Rail, suburban trains, auto-rickshaws, bike sharing and cabs. Since Metro Rail is limited to a few routes, the PMPML bus is the most used mode of public transport. The daily average ridership of PMPML is 12.41 lakh (till October) [14]. The PMPML has 467 bus routes and a total of 7687 bus stops [13]. This makes PMPML one of the densest public transport networks in Pune. On certain routes, the bus runs on a special lane; known as BRT - Bus Rapid Transit. This isolates the bus lane from the traffic on the road, reducing the impact of traffic congestion.

While public transport has many advantages, it comes with a few shortcomings. The time taken to reach the destination is not fixed. The person travelling on the bus has no clear idea how much time he/she might require to reach the destination. Our research reveals that the duration depends on the time at which the passenger boards the bus. We intend to provide software solutions to this problem using a machine-learning approach. This will provide an idea of duration to the passenger and thus he/she can precisely decide at what time it will be perfect to board the bus. This will have a direct impact on the lives of lakhs of citizens who travel by bus daily. It will ultimately strengthen peoples' confidence in public transport, thus leading to an increased preference to travel by bus. When more and more citizens start preferring public transport over personal vehicles, it will directly lead to reduced traffic on roads as well as have a positive environmental impact. Our research will enhance peoples' trust in public transport and ultimately lead to a

rise in the number of people choosing public transport over a personal vehicle, hence reducing traffic congestion, and controlling emissions of greenhouse gases.

2. Literature Survey

Researchers have proposed their algorithms for arrival time prediction of three different modes of transport: Road, air, and water. The previous work can be classified based on whether the existing system uses real-time location data or is trained on historical data.

2.1 Mode of transport: Road

2.1.1 Using real-time location tracking

In [2], the study explores the traffic congestion state by leveraging the ETA from the Google Maps API, combined with factors such as weather and traffic conditions. The classification process categorizes traffic into five distinct stages, ranging from smooth flow to heavy congestion. Among the methods evaluated, the Random Forest classifier achieved the highest accuracy at 92%. Similarly, [7] presents a Real-Time Locating System that integrates Global Positioning System (GPS) and Global System for Mobile Communications (GSM) wireless technologies. The study compares predicted arrival times with actual times, providing insights into the model's accuracy.

Building upon the approach in [7], [12] focuses on a model developed using automatic vehicle location data, applied to a bus route in Texas. This study contrasts models based on historical data with deep learning-based models. Notably, instead of using an ensemble method as in [7], the researchers proposed an Artificial Neural Network (ANN) model, which yielded the highest accuracy among the tested methods. [10] introduces an ETA model specifically designed for school buses in Canada. This model incorporates data from both previous and current days, applying an operational strategy to minimize risks associated with overestimation. The model relies on a GPS-based automatic vehicle location system for real-time bus tracking and demonstrates lower prediction error levels compared to moving averages and regression methods.

LSTMs are used for the estimation of arrival time in [16]. The regularity patterns in the dataset are captured using LSTM. They have used the online learning method to keep on updating the model based on collected real-time data. Based on the conducted experiments, they have concluded that the LSTM model outperforms the SVR model.

2.1.2 Using historical data

In [1], the authors propose a classification model to replace the more traditional regression models typically used in estimating time of arrival (ETA). This approach addresses the challenge of unknown trajectories in future trips by introducing a novel Categorical Approximate Method to Estimate the Time of Arrival. The classification labels correspond to the average time of each category, with travel time approximated using the weighted average of different classes. This new method is touted for its superior performance and reduced computational power requirements. Similarly, in [11], a Support Vector Machine (SVM) neural network algorithm is utilized to train the model and forecast arrival times. The study introduces a segment-based model that predicts travel time for the subsequent segment based on the current segment's time.

Random Forest, a widely recognized algorithm in prediction models, is the focus of [4], where different variable selection methods are compared. The comparison is based on computation time, the number of variables, and the area under the receiver operating characteristic curve, providing valuable insights into the efficiency of these methods. [9] presents a model designed to predict the arrival of a bus at the same stop across different routes. The study employs a deep learning approach alongside machine learning algorithms like SVM and linear regression, with results indicating that SVM outperforms the other three models.

A novel approach is used in [15], by proposing a hybrid model. The model uses a kNN classification algorithm for pattern matching. Further, they have used the Kalman filtering technique. They have used time-series analysis. Their model depends highly on the latest entry in the time-series data. Alternatively, they have proposed exponential smoothening to omit the requirement of keeping a record of historical data.

Table 1 shows the comparative statistics of accuracy of models tested in [2], [9] and [12]

Table 1 Comparison of models tested on the transportation system by road

Reference	Model	Average accuracy
[2]	Random forest	92%
[9]	Support vector machine	93.32%
[9]	ANN	91.32%
[9]	kNN	90.73%
[9]	Linear regression	91.21%
[12]	Historical data based models	92.59%
[12]	Regression models	85.43%
[12]	ANN	96.25%

2.2 Mode of transport: Air

In [5], the focus is on predicting the estimated time of arrival (ETA) for commercial flights, with factors such as weather conditions, air traffic, and potential flight paths playing a crucial role. The study compares different models, analyzing trajectory data, meteorology data, airport data, and airspace data to identify the most accurate model. Adaptive Boosting (AdaBoost) and Gradient Boosting are applied, and the results are compared, with the researchers reporting a root mean square error (RMSE) of 4 minutes. Similarly, [8] presents a model developed for flight ETA prediction, which accounts for the statistical relationships between flight parameters, air traffic, and weather information. The model implements a Random Forest classification algorithm, observing the out-of-bag RMSE as the number of grown trees increases. Feature importance is also analyzed, revealing that the destination airport is the most significant factor among those considered.

A new approach is proposed in [3] for predicting the ETA of flights at the runway upon entering the Terminal Manoeuvring Area. The study compares the performance of various machine learning algorithms, including neural networks and ensemble methods. The authors demonstrate that their proposed stacked model outperforms individual models in terms of accuracy and stability.

2.3 Model of transport: Water

In [6], a model is proposed which makes use of an Automatic Identification System and Long-Range Identification and Tracking on ships. This historical data is rasterized. A pathfinding algorithm is developed based on which the ETA predictions are made.

3. Methodology

Estimated Time of Arrival (ETA) is an important aspect of transportation. A person who is travelling should have a clear idea about how much time will be required to reach the desired destination. As discussed in the related work, research has been done previously to accurately predict the ETA. Some of the methods demand real-time vehicle tracking. However, real-time Global Positioning System (GPS) data processing demands high computation power. Also, it demands GPS hardware to be installed in the target mode of transport, in this case, the bus. To avoid the use of demanding resources like GPS, historical data was chosen by us. Models based on historical data demand significantly less computation power to train. Also, no hardware is to be installed for it. The following methods have been used as a part of model implementation:

1. **Data collection**: The data on the time at which the bus started from the source and the number of minutes required to reach the destination was collected manually by travelling on bus number 103 of PMPML. The data includes the timestamp of boarding the bus and the duration of the journey in minutes. The bus runs on a busy route in Pune city: from Kothrud Depot to Katraj. There are a total of 35 stops on the 13 km-long route. A dataset of 100 samples has been collected for the journey from the Vanaz Metro Station stop to the Bharati Vidyapeeth Gate stop.

2. **Data pre-processing**: Due to exceptional conditions like diversion of the route due to excavation on the road, the duration showed an anomaly as compared to normal circumstances. Such anomalous data points were detected and removed after studying the plot of board time v/s duration.

3. **Splitting the data**: Split the data, 80% for training the model and 20% for testing using the scikit-learn library in Python.

4. **Model training**: Trained a model on 7 different algorithms as mentioned in Table 1. Performed hyperparameter tuning on the models using grid search.

5. **Model evaluation**: Evaluated the 7 models by calculating the Mean Squared Error (MSE), R2 Score and Mean Absolute Percentage Error (MAPE).

The flow of data is explained in the architecture diagram in Figure 1. The data is processed by removing anomalous values of duration due to exceptional circumstances, like a breakdown. The pre-processed data is used to train a supervised machine learning model. Three different models are trained on the same data to analyse which one fits the best. Hence, it is trained on linear regression, decision tree and random forest. The user inputs the time at which he is planning to board the bus. The server passes a query to the trained model. The model computes the predicted value of travel duration and returns it to the user.

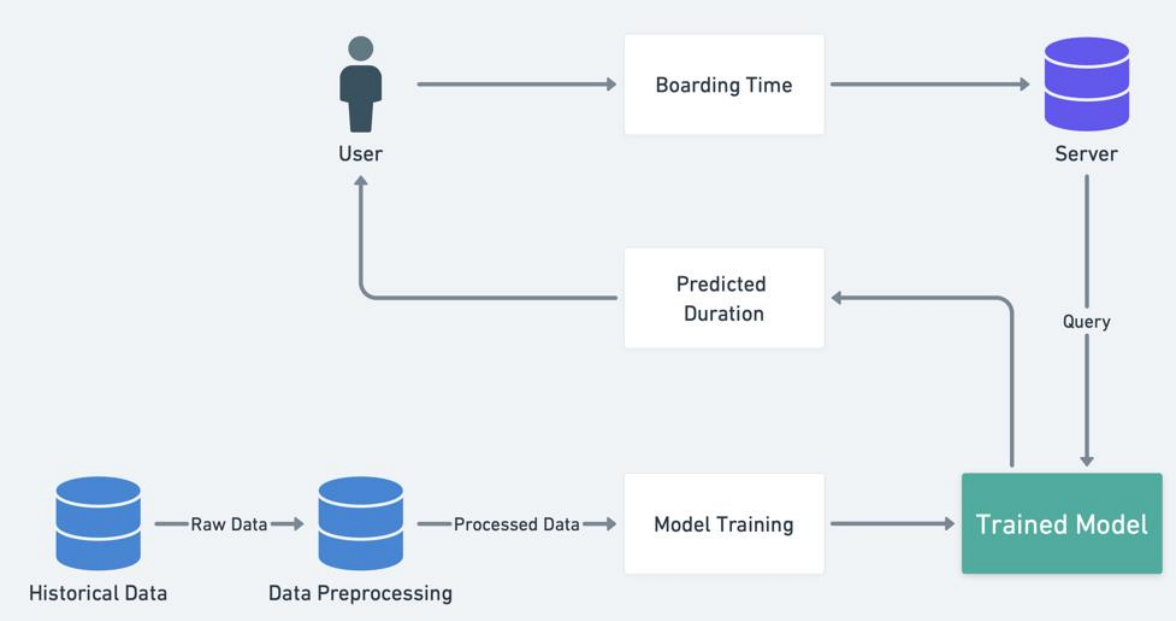

Fig. 1 Proposed Architecture

4. Results and Discussion

Table 2 shows the MSE, R2 Score and MAPE of all the 7 models.

Table 2 MSE, R2 Score and MAPE of each model

Model	Mean Squared Error	R2 Score	Mean Absolute Percentage Error
Linear Regression	15.8983903205065	0.295145256726312	0.06661971352684455
Support Vector Machine	13.4476553696827	0.403798530408155	0.06069024294442641
Decision Tree	11.4521604938272	0.492268746579091	0.05112715444818341
Random Forest	10.1000867198518	0.552212904045978	0.04580188384897036
Gradient Boost	10.1000867198518	0.527455189864827	0.04763390542108543
Ada Boost	5.072266003921091	0.8135238641472332	0.03712201500446292
K Nearest Neighbours	10.3666666666667	0.540394088669951	0.04900689887222298

Table 3 shows the values of optimal hyperparameters of the Ada Boost model.

Table 3 Hyperparameters of the AdaBoost model

Hyperparameter	**Value**
Number of estimators	10
Learning rate	0.01
Loss	Squared error

The following are the key observations:

- Linear regression does not perform well in this case. The reason is the non-linear relationship between the time stamp and the duration. In the early morning, there is less traffic on roads, so travel time is lower than average. During office hours i.e., 8 A.M. to 10 P.M. travel time is high due to peak in traffic. During noon, there is again a drop in traffic, hence leading to a reduction in travel time. It rises again in the evening i.e., 5 PM and onwards which is a usual time for people travelling from office to home.

- The Ada Boost model performs the best with an accuracy of 96.29%. It has the least MSE and MAPE among all the models.

- We observed that the model shows slightly greater deviations from the predicted value on Thursdays and Sundays; since Sunday is a holiday, there is an irregular trend in traffic patterns. Thursday is an industrial holiday, so on Thursdays, less time is required than any other day.

Figure 2 and Figure 3 show the comparison between the R2 Score and MAPE respectively of the 7 models.

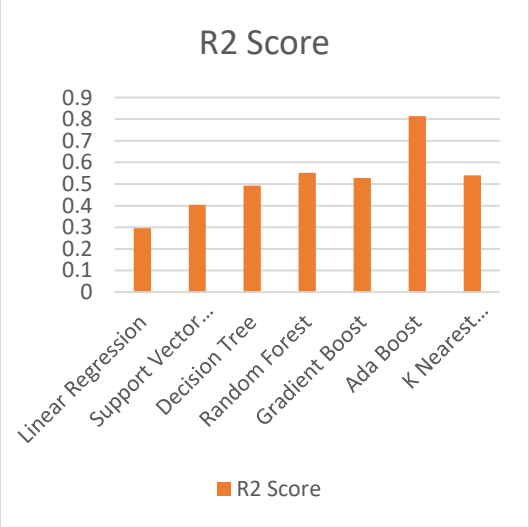

Fig. 2 Comparison of R2 Score

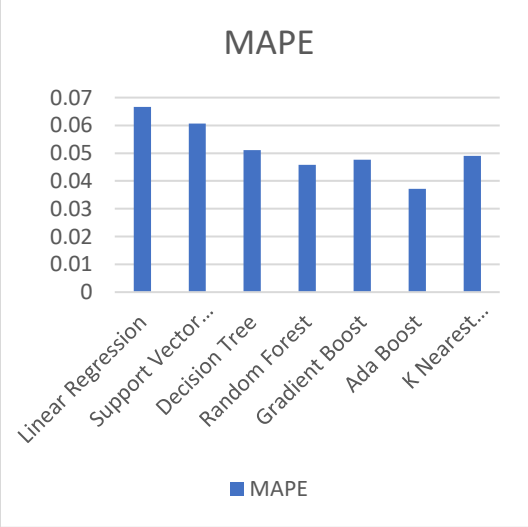

Fig. 3 Comparison of MAPE

5. Conclusion

The primary objective of this study is to comprehend the prediction of arrival time in public transport systems. Machine learning techniques prove to help predict the arrival time based on at what time the passenger has boarded the bus. Our research has demonstrated how supervised learning can predict arrival time by learning the trend and relations between the boarding time and duration required to reach the destination. Out of all the models we trained on the data, the model trained on the Ada Boost algorithm has shown a maximum R2 score of 0.8135 and an accuracy of 96.29%. This methodology can be applied to scale the model further on different routes of public transport, which might be a time-consuming process as it demands manually noting down the time stamp and duration. We can get help from people who travel daily on the same route. In the future, we plan on expanding parameters by considering the day of the week, and weather conditions, and checking for diversions on the route to further enhance predictions.

References

[1] Ye, Yongchao, et al. "Cateta: A categorical approximate approach for estimating time of arrival." *IEEE Transactions on Intelligent Transportation Systems* vol: 23.12, 24389-24400, (2022)

[2] [Zafar, Noureen, and Irfan Ul Haq. "Traffic congestion prediction based on Estimated Time of Arrival." *PloS one* vol: 15.12, e0238200, (2020)

[3] Wang, Zhengyi, Man Liang, and Daniel Delahaye. "Automated data-driven prediction on aircraft Estimated Time of Arrival." *Journal of Air Transport Management* vol: 88, 101840, (2020)

[4] Speiser, Jaime Lynn, et al. "A comparison of random forest variable selection methods for classification prediction modeling." *Expert systems with applications* vol: 134, pg.: 93-101, (2019)

[5] Ayhan, Samet, Pablo Costas, and Hanan Samet. "Predicting estimated time of arrival for commercial flights." *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018.

[6] Alessandrini, Alfredo, Fabio Mazzarella, and Michele Vespe. "Estimated time of arrival using historical vessel tracking data." *IEEE Transactions on Intelligent Transportation Systems* vol: 20.1, pg.: 7-15, (2018)

[7] Farooq, Muhammad Umar, Aamna Shakoor, and Abu Bakar Siddique. "GPS based public transport arrival time prediction." *2017 International Conference on Frontiers of Information Technology (FIT)*. IEEE, 2017.

[8] Kern, Christian Strottmann, Ivo Paixao de Medeiros, and Takashi Yoneyama. "Data-driven aircraft estimated time of arrival prediction." *2015 annual IEEE systems conference (syscon) proceedings*. IEEE, 2015.

[9] Yu, Bin, William HK Lam, and Mei Lam Tam. "Bus arrival time prediction at bus stop with multiple routes." *Transportation Research Part C: Emerging Technologies* vol: 19.6 pg.: 1157-1170, (2011)

[10] Chung, Eui-Hwan, and Amer Shalaby. "Expected time of arrival model for school bus transit using real-time global positioning system-based automatic vehicle location data." *Journal of Intelligent Transportation Systems* vol: 11.4, pg.: 157-167, (2007)

[11] Bin, Yu, Yang Zhongzhen, and Yao Baozhen. "Bus arrival time prediction using support vector machines." *Journal of Intelligent Transportation Systems* vol: 10.4, pg.: 151-158, (2006)

[12] Jeong, Ranhee, and R. Rilett. "Bus arrival time prediction using artificial neural network model." *Proceedings. The 7th international IEEE conference on intelligent transportation systems (IEEE Cat. No. 04TH8749)*. IEEE, 2004.

[13] J. Sengupta, "The Times of India," 17 November 2023. [Online]. Available:https://timesofindia.indiatimes.com/city/pune/news/dismal-uptick-pmpml-ridership-pune/articleshow/105276581.cms#. [Accessed 14 January 2024].

[14] "Moovit," [Online]. Available: https://moovitapp.com/index/en/public_transit-lines-Pune-5884-1509110. Accessed 14 January 2024].

[15] Kumar, B. Anil, Lelitha Vanajakshi, and Shankar C. Subramanian. "A hybrid model based method for bus travel time estimation." *Journal of Intelligent Transportation Systems*22.5 (2018): 390-406.

[16] Panyo, Kaisorrawat, et al. "Bus arrival time estimation for public transportation system using LSTM." *2020-5th International Conference on Information Technology (InCIT)*. IEEE, 2020.