Evaluation of Machine Learning Models for Bus Travel Time Prediction

M P V P Medawatte[1], H.L.K Perera[2], A.B Jayasinghe[3], K A S N Sumathipala[4]

Abstract

Accurate real-time bus arrival information is essential for an efficient public transport network, as it significantly impacts passenger experience, system reliability, reduced waiting times, dwell times, and operational efficiency. In Sri Lanka's public transport system, current bus arrival time computations primarily rely on static data, neglecting real-time information and critical factors influencing travel times. This highlights the need to identify the unique variables affecting bus arrival times within the Sri Lankan context and to develop robust prediction models that account for these influences. While traditional methods such as Historical Average Models, Regression, Time Series Analysis, and Kalman Filtering have been used in previous research for short-term travel time predictions, Machine Learning (ML) approaches have proven to deliver superior accuracy. ML models are regarded as the most effective for heterogeneous, lane-less traffic conditions with varying traffic volumes, such as those found in Sri

Lanka. ML techniques excel in processing large, high-quality datasets and provide accurate predictions by accounting for all relevant variables influencing travel times.

Although research has been conducted on developing various basic ML models for travel time prediction, there is a noticeable gap in studies comparing these models to determine the most suitable one for the Sri Lankan context. A Long- Short Term Memory (LSTM) neural network is a deep learning model that is capable of handling long-term dependencies. In the context of bus travel time prediction, LSTMs can leverage historical traffic and travel data to capture temporal patterns and fluctuations that influence travel times. By evaluating LSTMs against basic machine learning models, this study seeks to explore the advantages of applying deep learning techniques to transportation forecasting, ultimately contributing to more accurate and efficient predictive systems in transit planning. The ML models selected in this study include two basic traditional models K- Nearest Neighbours (KNN) and Support Vector Regression (SVR) and four advanced models that utilize ensemble techniques and advanced optimization as Random Forest Regression (RFR), Ada Boost, XG Boost and Gradient Boosting Machine (GBM). The performance of these models was compared with the LSTM model to identify the gap in their accuracies.

GPS data for the Moratuwa to Colombo bus route (Route 100) was gathered over 30 days for this study, covering weekdays, weekends, and public holidays, using five GPS

devices at various times throughout the day. This data set included information from 335 bus trips, totalling over 17,500 GPS data points. The route has 53 bus stops, and data filtering was applied based on GPS coordinates relative to each stop to calculate the travel times between them. To enhance accuracy, the route was divided into segments according to the number of bus stops, allowing for more precise travel time predictions by capturing speed variations in each segment. A literature review helped identify key factors influencing bus travel time. Feature analysis was conducted to identify the importance of each feature in the model. By doing this feature selection we could separate the most effective features and select them for the model instead of making it more complex. Those were identified as the road section, day of the week, hour of the day, availability of bus lanes, travel distance, weather conditions, and the number of signalized intersections and number of signalized crossings. Data collection accounted for all these variables. After data cleaning and preprocessing, they were fed into the models. In this study, the performance of various machine learning models was compared based on two key metrics: Mean Absolute Error (MAE) and R-Squared (R^2). Among the models tested, when analysing the results it was identified that the LSTM model stands out as the best performer by a significant margin. The LSTM model achieved a Mean Absolute Error of 3.826, which is much lower than the other models. The LSTM model achieved a significantly lower MAE of 3.826, outperforming all other models by a

wide margin. The next closest model, SVR, recorded an MAE of 17.717. Other models, including RFR with an MAE of 17.283, AdaBoost with 17.458, XGBoost with 17.446, and GBM with 17.397, all had higher MAE values. KNN Regression showed the highest error, with an MAE of 18.145. This indicates that LSTM makes far fewer errors in its predictions, showing greater accuracy. Furthermore, the R^2 for LSTM is 0.983, meaning it explains 98.3% of the variation in the data, which is considerably higher than the other models. The SVR model had an R^2 value of 0.618, while RFR achieved 0.651. KNN Regression recorded an R^2 of 0.616, AdaBoost had 0.642, XGBoost showed an R^2 of 0.640, and GBM had 0.647. This demonstrates that LSTM not only predicts more accurately but also provides a better fit to the data. Therefore, the study concludes that the LSTM model outperforms all other ML models discussed in both MAE and R^2 and is 78.5% more accurate than the next closest model, SVR, making it the most reliable and effective model for this task.

Keywords: *Machine Learning, Bus travel time Long-Short-term Memory Model*

Acknowledgment: The authors would like to acknowledge the ERASMUS+ LBS2ITS project for funding this research

1. Graduate Research Assistant, Department of Civil Engineering, University of Moratuwa. viduniprамodya@gmail.com

2. Senior Lecturer, Department of Civil Engineering, University of Moratuwa. loshakap@uom.lk

3. Senior Lecturer, Department of Town & Country Planning, University of Moratuwa. amilabj@uom.lk

4. Senior Lecturer, Department of Computational Mathematics, University of Moratuwa. sagaras@uom.lk