Bus on-time performance prediction based on machine learning models

Anran Zheng[1], Yuran Sun[1], and Yuetong Zhang[1]

[1] *University of Florida*

Abstract

The accurate prediction of bus on-time performance has significant implications for both passengers to plan their trips and bus operations to manage bus fleets effectively. In this study, we compared and evaluated the performance of multiple machine learning models for predicting the on-time performance of buses in Miami transit system. The study uses historical on-time performance data and weather data to train and test several machine learning models, including decision tree, random forest, support vector machine, and XGBoost. The predictive performance of the models is evaluated using various statistical metrics, including mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE). The results indicate that the random forest model outperforms the other models, with the lowest MAE (1.9063), MSE (19.2031) and RMSE (4.3821). This model detects important features such as long bus headways, stops at or around the end of bus trips, and arrival time since afternoon peak hours. These are significant determinants that provide guidance for transit agencies and operators on how to ease transit delays and improve service reliability.

Keywords: Bus on-time performance prediction; machine learning; transit service reliability; random forest

1 Introduction

Public transportation, commonly referred to as public transit or mass transit, denotes the network of transportation services that are accessible to the general public. The term "public

transportation" involves a range of transportation modes, including buses, trains, subways, trams, light rail, and ferries, which are intended to convey passengers to diverse locations within a given urban or regional area (19). Bus riding was identified as the primary mode of commuting by the largest cohort of public transportation commuters, comprising 46.3 percent of the total number of commuters, according to the American Community Survey (ACS) (5). The popularization of buses makes traveling for long or short distances more affordable and helps reduce carbon emissions to combat climate change in urban and suburban regions (11). The popularization of buses offers an accessible and cost-effective means for individuals to commute to various destinations such as work, school, healthcare facilities, and shopping centers. On the other hand, the efficacy and reliability of bus transit systems are frequently impeded by a number of factors, including traffic congestion, meteorological conditions, and unanticipated events (14). When a vehicle delays, it may fail to make its remaining scheduled stops and cause a cascading reaction of delays throughout transit routes (15). When passengers take such buses, they may miss the connecting bus due to the delayed arrival and have to wait extended period for the next bus, which leads to annoyance, inconvenience, and even economic and environmental consequences in some cases.

In order to solve these issues, there is growing interest in building predictive models for bus on-time performance. Transportation organizations may better manage their resources, enhance the quality of service for customers, and lessen the effect of delays on the whole system by properly projecting when buses are likely to arrive at their destinations. Machine learning approaches has been widely applied to evaluate bus on-time performances with traffic, meteorological and population density data, etc. In this study, we aim to accurately forecast bus arrival times based on data obtained from Miami bus system and Miami weather department. Nine models have been built for the prediction and their performances were compared. We also sought to find the most important factors that influence the bus on-time performance, which could provide insights to transportation department for better services.

2 Literature Review

In general, there are several categories of studies to predict bus on-time performance: (1). Historical average (HA) approach predicts the bus on-time performance with average travel time and speed, but it is challenging to collect sufficient journey records for OD pairs with this method (10). (2). Kalman filtering (KF) approach can forecast unknown travel times based on a series of travel time records (8; 6). However, this approach is sensitive to complicated scenarios with anomalies. (3) Machine learning models generate more accurate results than the other approaches. Most common prediction models include Support vector machines (SVM), long short-term memory (LSTM), Artificial neural networks (ANN) and deep neural network model (DNN) (20; 4). One study combined LSTM and ANN to predict the arrival for both long- and short- distance circumstances (12). Using deep neural network model (DNN) and GPS data from transportation sector, the travel time prediction for Thailand buses has a mean absolute percentage error as low as 0.55 (17). Long short-term memory (LSTM) networks coupled with GPS calibration achieved better bus arrival time prediction compared to methods based on traditional time-of-arrival techniques (9).

According to the previous literature, influencing factors of bus on-time performance include driving habits, traffic conditions (congestion or emergency), passenger demand, and bus facilities (1). Previous studies have employed Automatic Vehicle Location (AVL), manually gathered questionnaires, mobile phone footprints, and social media data to meet this end (2). Big data technologies and their applications in public transportation and traffic offer a platform for data solutions to relevant challenges (7). AVL and GPS data are popular in transportation applications. GPS data such as longitude, and latitude were also widely used as input variables in bus service reliability prediction (15) (3). Apart from getting real-time information from the transportation sector, it necessitates the acquisition of weather information for better prediction. For example, shorter travel time was observed on rainy days in a study conducted in Malaysia, possibly due to less passenger demand and more skipped bus stops since fewer riders would like to travel by bus on rainy days (14). Nonetheless, another study conducted in Australia found long travel time on rainy days due to reduced vehicle

speed for safety concerns (13).

3 Methods

3.1 Data Splitting and hyperparameter tuning

The dataset was randomly partitioned into two subsets with a 7:3 ratio. The training set, which comprises 70% of the data, was used for model training, feature selection, and hyperparameter tuning. The remaining 30% of the dataset was reserved as the test set, which was exclusively used to evaluate the model's performance.

In order to optimize the model's performance, hyperparameter tuning is performed for each model. This process involves selecting the optimal combination of hyperparameters that maximizes the model's accuracy and minimizes overfitting. To fine-tune the model's hyperparameters, a grid search approach was employed in our study. This method involves systematically testing a range of hyperparameter values to find the optimal combination that yields the best performance. Moreover, we utilized 10-fold cross-validation during hyperparameter tuning. This technique splits the data into 10 equally sized subsets, trains the model on 9 subsets, and validates it on the remaining subset. By repeating this process 10 times with different subsets as validation data, we can obtain a more stable estimate of the optimal hyperparameter. To ensure consistency in model training and hyperparameter tuning, all models were trained using the same training and test sets. Additionally, the same random seed was employed for all the models.

3.2 Machine Learning Models

Best Subset The best subset is used in linear regression analysis to determine the subset of independent variables that best explain the variation in the dependent variable. This method involves testing all possible combinations of independent variables and selecting the one that yields the highest model fit statistics, such as R-squared, adjusted R-squared, and AIC (Akaike Information Criterion). The process of selecting optimal subsets may be

computationally intensive, but the method can lead to a model with high interpretability.

Gams Natural Spline Natural Spline captures the nonlinear relationships between the dependent variables and the independent variables. It is a piece-wise polynomial function that consists of multiple spline knots, which join together smoothly at their endpoints. The natural spline is a specific type of spline that enforces additional constraints, such as ensuring that the function is linear at the endpoints, which improves the stability and interpretability of the model.

Gams Smoothing Spline The smoothing spline is usually designed to fit a smooth curve with noisy or irregular data. It does not require the specification of the number or location of the knots, which makes it suitable for dealing with complex and unpredictable patterns. Smoothing spline uses a regulation parameter to control the amount of smoothing and to avoid overfitting.

Lasso Regression Lasso Regression is a form of linear regression that is particularly well-suited for analyzing high-dimensional datasets where there are more predictors than observations. By introducing a penalty term into the regression function, Lasso Regression can shrink coefficients exactly toward zero and perform automatic variable selection. The optimal amount of shrinkage is usually determined by hyperparameter tuning.

Ridge Regression Ridge Regression is a regularization technique that adds a penalty term proportional to the square of the magnitude of the coefficients to the least-squares objective function of linear regression. This penalty term, also known as the L2-norm or Euclidean norm, shrinks the coefficients toward zero. Ridge Regression is specifically suitable for dealing with multicollinearity and reducing the impact of highly correlated variables.

Decision Tree A Decision Tree (DT) is a tree-structured model constructed by recursively splitting the dataset into subsets based on the value of the input variables. Each internal node of the decision tree corresponds to the variable selected for dividing the data, each

branch denotes a decision rule, and each leaf node represents the final class label, and the samples that are classified into the class.

Random Forest A Random Forest is a collection of decision trees, each built from a randomly sampled set of data, with the feature selection for each split also being random. The final outputs are obtained by either taking the mode or the average of the predictions from all decision trees in the forest. Random Forest is widely applied for both classification and regression.

Support Vector Machine A Support Vector (SVM) can be applied to classify the data. It aims to identify the optimal decision boundary that separates the data into distinct classes. The boundary is a hyperplane determined by the support vectors, which are observations closest to the boundary. SVM is efficient in dealing with complex datasets or non-linear boundaries.

XGBoost XGBoost, a machine learning algorithm that belongs to the boosting family, is a robust and powerful model. It utilizes an ensemble of weak decision trees, each of which is trained on a subset of data, and combines their outputs to make the final prediction. XGBoost is renowned for its speed, high accuracy, and proficiency in handling missing values and feature interactions.

4 Data

The main data source of this study is the historical on-time performance data of Miami Dade transit collected through Swiftly APIs, which provide schedule-adherence information on all routes in the system (Swiftly). The data are collected during March 1st to 6th, 2023, including a total of 59028 data records with 3 bus routes that delay most frequently (route 9, 132 and 54). This dataset also includes detailed trip information such as arrival stops, vehicle ID, bus routes, direction, stop sequence, as well as status (delay or not). Besides

on-time performance data, historical hourly weather data is also collected through Visual crossing weather (visualcrossing).

We evaluated the bus on-time performance by measuring the difference between scheduled and actual arrival time (in min), which is the outcome variable in our machine learning models. This continuous variable can be either positive or negative, as positive values suggest late arrivals while negative values suggest early arrivals. In the data preprocessing stage, we visualized the multicollinearity among all the independent Variables with correlation matrix as figure 1 shows and excludes some variables with the highest correlation. The correlation of any two independent variables is smaller than 0.5. The final chosen ten independent variables and the dependent variable are shown in table 1. The descriptive statistics table of these variables is shown in table 2.

5 Results

5.1 Performance Metrics

In our evaluation of model performance, we utilized several regression metrics, including Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE). These metrics provide a quantitative assessment of how well the model predictions match the actual values.

Mean Absolute Error (MAE) Mean Absolute Error (MAE) is a commonly used metric in regression analysis to evaluate the accuracy of a model's predictions. It measures the average absolute difference between the predicted and actual values of the dependent variable. The lower the MAE, the better the model's performance.

Mean Square Error (MSE) Mean Square Error (MSE) is a useful metric in regression analysis to evaluate the accuracy of a model's predictions. It measures the average of the squared differences between the predicted and actual values of the dependent variable. The lower the MSE, the better the model's performance.

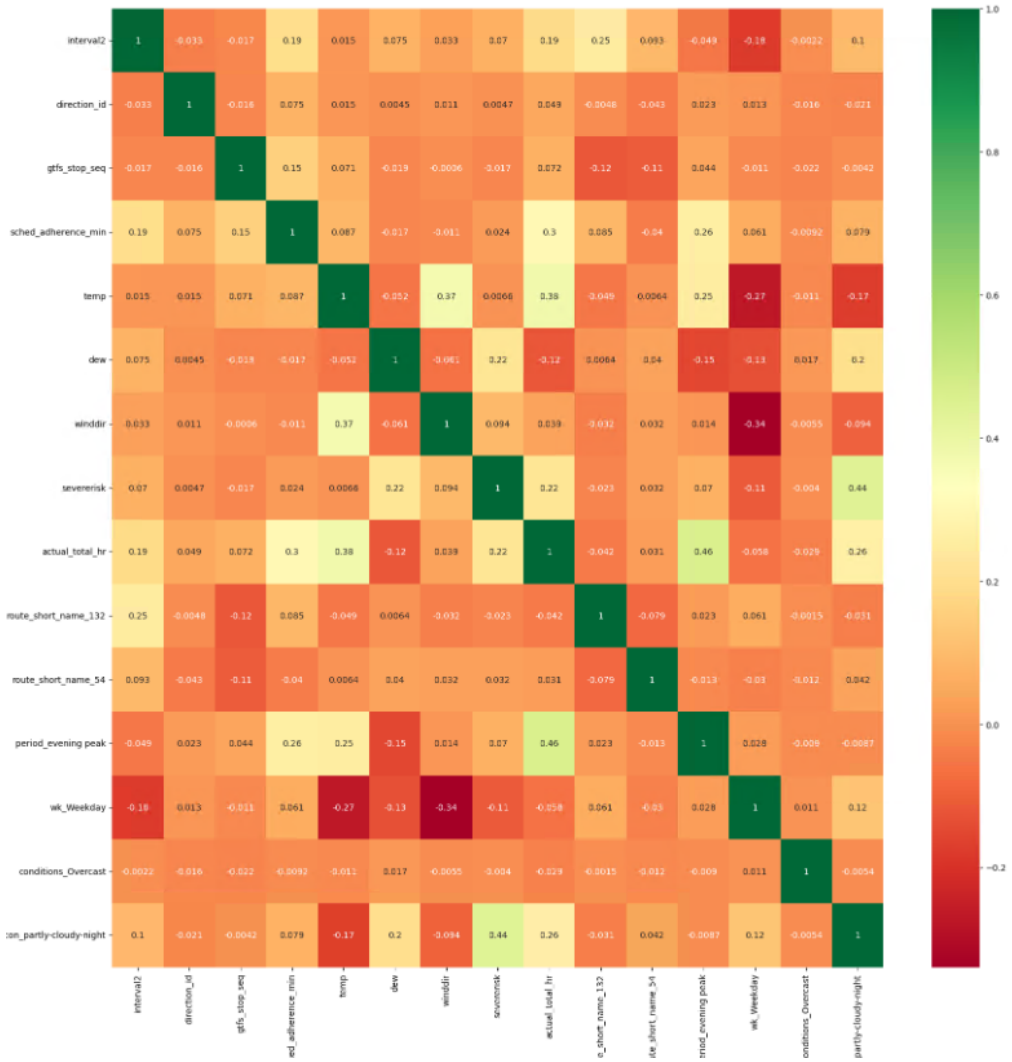

Figure 1: Correlation heatmap of all the variables

Root Mean Square Error (RMSE) Root Mean Square Error (RMSE) is a widely used metric in regression analysis to measure the accuracy of a model's predictions. It is the square root of the average of the squared differences between the predicted and actual values of the dependent variable. The lower the RMSE, the better the model's performance.

The three metrics are defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|\,, \tag{1}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{2}$$

Table 1: Descriptive profile of the outcome variable and input variables

Name	Description
Outcome variable	
sched_adherence_min	Time differences between actual and scheduled arrival time
Bus operating characteristics	
direction_id	Dummy variable equal to 1 if bus travels in one direction and 0 if opposite direction
actual_total_hr	Bus actual arrival time (duration in hour since 00:00)
interval2	Bus headway
route_short_name_54	Dummy variable equal to 1 if the Bus_id is 54 and 0 if not
route_short_name_132	Dummy variable equal to 1 if the Bus_id is 132 and 0 if not
wk_weekday	The bus operates during weekdays
period_evening peak	The bus operates during afternoon peak hour (4-7pm)
gtfs_stop_seq	Order of stops for a particular trip
Weather	
temp	Temperature
dew	Dew point
winddir	Wind direction
severerisk	Severe weather risk probability
conditions_Overcast	Dummy variable equal to 1 if the weather is overcast and 0 if not
icon_partly-cloudy-night	Dummy variable equal to 1 if the current weather is partly cloudy and the time is at night as well and 0 if not

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}, \tag{3}$$

where n is the total number of observations in the testing set, y_i is the observed value for observation i, and $\hat{y_i}$ is the predicted value for observation $i, i = 1, \cdots, n$.

5.2 Model Comparison

Table 3 displays the performance metrics of the nine models:

By comparing the MAE, MSE and RMSE, RF is significantly outperforming than other models, in terms of the lowest MAE (1.9063), MSE (19.2031) and RMSE (4.3821). This

Table 2: Descriptive statistics for all the variables

Variable	Category	%	Mean	SD	Max	Min
Outcome variable						
sched_adherence_min			10.17	11.70	69.38	-18.68
Independent variable						
interval2			33.11	24.41	193.93	0.00
direction_id	inbound(0)	50.64				
	outbound(1)	49.36				
gtfs_stop_seq			45.77	27.54	108	1
temp			79.27	5.81	87.80	63.60
dew			64.67	3.32	72.60	56.80
winddir			163.42	76.29	360	0.00
severerisk			11.18	4.70	30.0	10.0
actual_total_hr			13.85	4.99	24	0.04
route_short_name_132	No(0)	99.2				
	Yes(1)	0.8				
route_short_name_54	No(0)	57.2				
	Yes(1)	42.8				
period_evening peak	No(0)	75.79				
	Yes(1)	24.21				
wk_Weekday	No(0)	30.69				
	Yes(1)	69.31				
conditions_Overcast	No(0)	99.97				
	Yes(1)	0.03				
icon_partly-cloudy-night	No(0)	89.55				
	Yes(1)	10.45				

Table 3: Model Performance Metrics

	MAE	MSE	RMSE
Best Subset	7.9677	112.9013	10.6250
Gams Natural Spline	7.5551	101.5432	10.0763
Gams Smoothing Spline	7.5458	101.0096	10.0499
Lasso Regression	8.0165	113.6001	10.6583
Ridge regression	8.0166	113.6018	10.6584
Decision Tree	2.1026	38.8553	6.2334
Random Forest	1.9063	19.2031	4.3821
Support Vector Machine	6.2610	91.0538	9.5422
XGBoost	4.2768	41.2510	6.4227

suggests that RF model can better predict bus on-time performance in Miami transit system. Furthermore, the fact that the RF model outperforms other models such as decision tree, support vector machine, and XGBoost suggests that RF model is more effective at handling the noise and complexity of the data in the bus transit system. The results also suggest

that the random forest model is robust and stable, as it performs consistently well across different evaluation metrics. In contrast, Lasso and Ridge regression have the worst model performance with the highest MAE (8.02), MSE (113.60) and RMSE (10.66), since they assume a linear relationship between the outcome and the independent variables. However, the relationship is nonlinear and can be captured by more flexible and complex models like random forest and decision tree.

5.3 Model Interpretation

We plotted the feature importance of the final random forest model in figure 3, which shows the predictive performance of all independent variables. Two features have the most important influence at around 26% on the bus on-time performance: *interval2* and *actual_total_hr*. *gtfs_stop_seq* comes third in terms of variable importance (12%), followed by temperature, dew point and wind direction. The least important features include *conditions_Overcast* and *severerisk*.

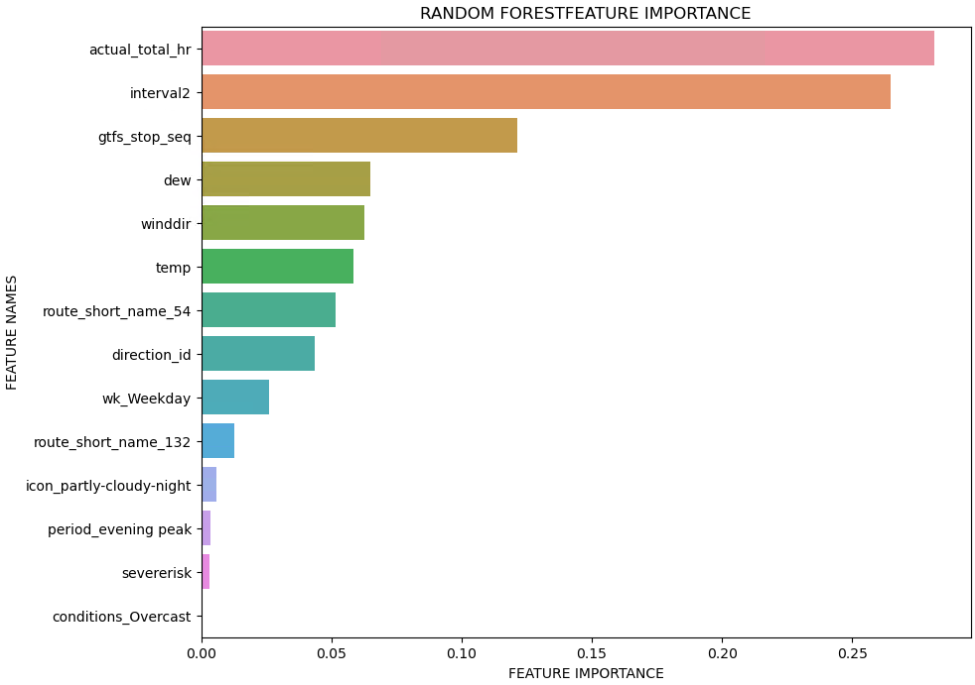

Figure 2: Variable importance by Random forest

We also generated partial dependence plots (PDPs) for some of the important independent

variables, which visualize the relationship between the outcome variable and the independent variable. According to the PDP for *interval2*, buses with 10-20 minutes headway usually arrive most punctually, while buses with longer headway (>50 minutes) are more likely to delay longer. The PDP for *gtfs_stop_seq* indicates that buses delay more seriously as they arrive at the end trip. From the PDPs of *actual_total_hr* and *wk_Weekday*, bus delay frequently happen during afternoon peak hours (4-7pm) at weekdays. The PDP plots for *temp*, *dew* and *winddir* indicate high non-linearity between bus arrival time difference and weather factors such as temperature, dew points and wind direction. Bus typically delay under the following weather conditions: temperature at 63-76 F, dew point at around 65 F, wind direction at 150-250 degrees.

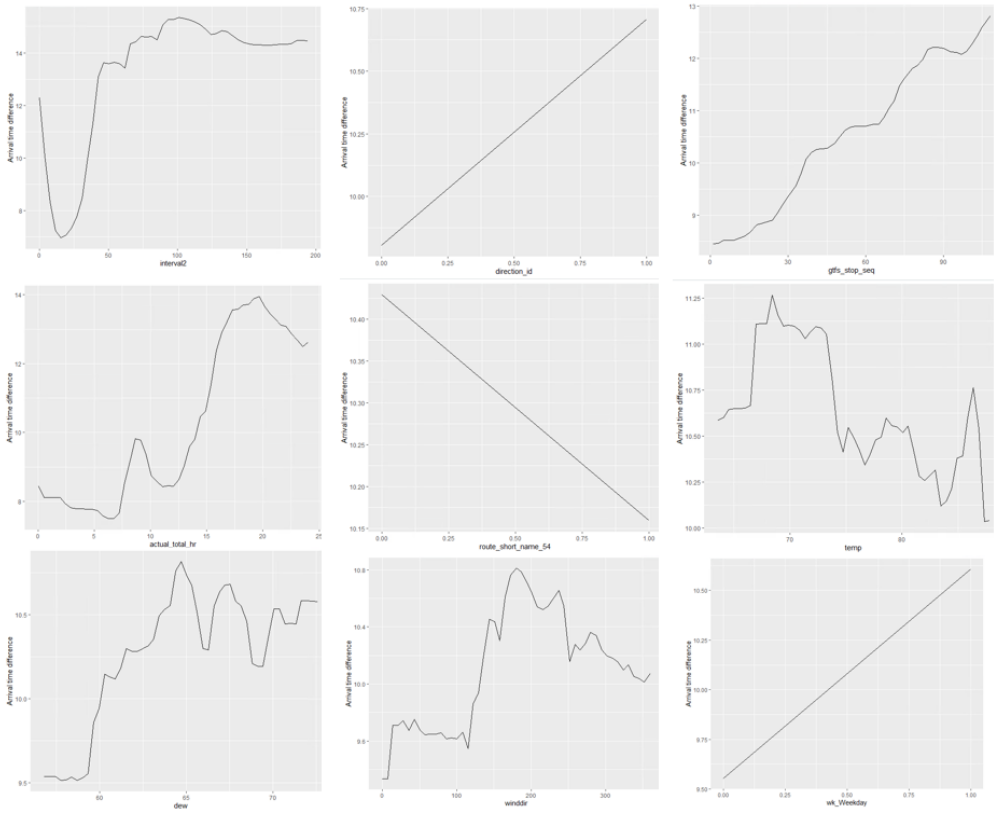

Figure 3: Partial dependence plots for some important features

Predicting on-time performance can help transit agencies to figure out what factors lead to bus delays and how to better manage their services and improve reliability. According to our model interpretation results, most important determinants of bus delays include long

bus headways, end-of-trip stops, and arrival time since afternoon peak hours. This can provide guidance on how to improve the bus on-time performance. Routing and scheduling are suggested to be optimized during peak hours to reduce passenger wait times and help keep the buses on schedule.

6 Conclusion

The purpose of our research was to predict the variance between the scheduled arrival time and the real arrival time of transit. We accomplished this by creating nine unique machine-learning models and evaluating their performance. Following a comparison of the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), we determined that the Random Forest model was the most effective at predicting the time difference.

Additionally, we employed feature importance and partial derivative plots (PDP) to aid in the interpretation of our findings. Through this process, we discovered a crucial insight that had not been extensively studied in prior research: the significance of bus headway in predicting the variance between scheduled and actual transit arrival times. This novel finding highlights the importance of considering bus headway as a critical predictor in transit time prediction models. Besides, nonlinear relationships between specific variables and the time difference are captured by PDP. This contributes to a better understanding of the complex and dynamic nature of transit systems.

Admittedly, there are some limitations to our study. One limitation is that we were only able to analyze the three worst routes due to computational constraints. This partial analysis may not provide a comprehensive understanding of the transit system's performance as a whole. Besides, in our future studies, we plan to incorporate additional variables into our prediction model to improve its accuracy and effectiveness. These variables may include passenger load, traffic conditions, and road conditions, which can significantly impact transit performance and arrival times.

Author contributions

(%)	Anran Zheng	Yuran Sun	Yuetong Zhang
Conceptualization	33	33	33
Methodology	33	33	33
Software	35	30	35
Resources	40	30	30
Data collection & investigation	40	30	30
Writing (original draft)	30	40	30
Writing (review & editing)	33	33	33
Visualization	40	30	30
Project administration	33	33	33

References

[1] Ansari Esfeh, M., Wirasinghe, S., Saidi, S., and Kattan, L. (2021). Waiting time and headway modelling for urban transit systems–a critical review and proposed approach. *Transport Reviews*, 41(2):141–163.

[2] Ashwini, B. and Sumathi, R. (2020). Data sources for urban traffic prediction: A review on classification, comparison and technologies. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pages 628–635. IEEE.

[3] Ashwini, B., Sumathi, R., and Sudhira, H. (2022). Bus travel time prediction: a comparative study of linear and non-linear machine learning models. In *Journal of Physics: Conference Series*, volume 2161, page 012053. IOP Publishing.

[4] Bai, C., Peng, Z.-R., Lu, Q.-C., and Sun, J. (2015). Dynamic bus travel time prediction models on road with multiple bus routes. *Computational intelligence and neuroscience*, 2015:63–63.

[5] Burrows, M., Burd, C., and McKenzie, B. (2021). Commuting by public transportation in the united states: 2019. *American Community Survey Reports*.

[6] Cathey, F. and Dailey, D. J. (2003). A prescription for transit arrival/departure pre-

diction using automatic vehicle location data. *Transportation Research Part C: Emerging Technologies*, 11(3-4):241–264.

[7] Chen, Y., Guizani, M., Zhang, Y., Wang, L., Crespi, N., Lee, G. M., and Wu, T. (2018). When traffic flow prediction and wireless big data analytics meet. *IEEE network*, 33(3):161–167.

[8] Chien, S. I.-J., Ding, Y., and Wei, C. (2002). Dynamic bus arrival time prediction with artificial neural networks. *Journal of transportation engineering*, 128(5):429–438.

[9] Han, Q., Liu, K., Zeng, L., He, G., Ye, L., and Li, F. (2020). A bus arrival time prediction method based on position calibration and lstm. *IEEE Access*, 8:42372–42383.

[10] He, P., Jiang, G., Lam, S.-K., and Sun, Y. (2020). Learning heterogeneous traffic patterns for travel time prediction of bus journeys. *Information Sciences*, 512:1394–1406.

[11] Kang, A. S., Jayaraman, K., Soh, K.-L., and Wong, W.-P. (2020). Tackling single-occupancy vehicles to reduce carbon emissions: Actionable model of drivers' implementation intention to try public buses. *Journal of Cleaner Production*, 260:121111.

[12] Liu, H., Xu, H., Yan, Y., Cai, Z., Sun, T., and Li, W. (2020). Bus arrival time prediction based on lstm and spatial-temporal feature vector. *IEEE Access*, 8:11917–11929.

[13] Ma, Z.-L., Ferreira, L., Mesbah, M., and Hojati, A. T. (2015). Modeling bus travel time reliability with supply and demand data from automatic vehicle location and smart card systems. *Transportation Research Record*, 2533(1):17–27.

[14] Mohamed, A. H., Adwan, I. A., Ahmeda, A. G., Hrtemih, H., and Al-MSari, H. (2021). Identification of affecting factors on the travel time reliability for bus transportation. *Knowledge-Based Engineering and Sciences*, 2(1):19–30.

[15] Park, Y., Mount, J., Liu, L., Xiao, N., and Miller, H. J. (2020). Assessing public transit performance using real-time data: spatiotemporal patterns of bus operation delays in columbus, ohio, usa. *International Journal of Geographical Information Science*, 34(2):367–392.

[Swiftly] Swiftly. Swiftly api reference. https://swiftly-inc.stoplight.io/docs/swiftly-docs/6zpcgvbu5wbb3-swiftly-api-reference.

[17] Treethidtaphat, W., Pattara-Atikom, W., and Khaimook, S. (2017). Bus arrival time prediction at any distance of bus route using deep neural network model. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 988–992.

[visualcrossing] visualcrossing. visualcrossing. https://www.visualcrossing.com/weather-data.

[19] Vuchic, V. R. (2002). Urban public transportation systems.

[20] Yu, B., Lam, W. H., and Tam, M. L. (2011). Bus arrival time prediction at bus stop with multiple routes. *Transportation Research Part C: Emerging Technologies*, 19(6):1157–1170.