



Travel Time Prediction and Explanation with Spatio-temporal Features: A Comparative Study

Irfan Ahmed ^{1,2}, Indika Kumara ^{1,2,*} , Vahideh Reshadat ³, A. S. M. Kayes ^{4,*} , Willem Jan Van Den Heuvel ^{1,2} and Damian Tamburri ^{1,3}

¹ Jheronimus Academy of Data Science, Sint Janssingel 92, 5211 DA 's-Hertogenbosch, Netherlands

² Tilburg University, Warandelaan 2, 5037 AB Tilburg, Netherlands

³ Eindhoven University of Technology, 5612 AZ Eindhoven, Netherlands

⁴ Department of Computer Science and Information Technology, La Trobe University, Plenty Road, Bundoora, Victoria 3086, Australia

* Correspondence: i.p.k.weerasinghadewage@tilburguniversity.edu (I.K.); a.kayes@latrobe.edu.au (A.S.M.K.)

Abstract: Travel time information is used as input or auxiliary data for tasks such as dynamic navigation, infrastructure planning, congestion control, and accident detection. Various data-driven Travel Time Prediction (TTP) methods have been proposed in recent years. One of the most challenging tasks in TTP is developing and selecting the most appropriate prediction algorithm. The existing studies that empirically compare different TTP models only use a few models with specific features. Moreover, there is a lack of research on explaining TTPs made by black-box models. Such explanations can help to tune and apply TTP methods successfully. To fill these gaps in the current TTP literature, using three data sets, we compare three types of TTP methods (ensemble tree-based learning, deep neural networks, and hybrid models) and ten different prediction algorithms overall. Furthermore, we apply XAI (Explainable Artificial Intelligence) methods (SHAP and LIME) to understand and interpret models' predictions. The prediction accuracy and reliability for all models are evaluated and compared. We observed that the ensemble learning methods, *i.e.*, XGBoost and LightGBM, are the best performing models over the three data sets, and XAI methods can adequately explain how various spatial and temporal features influence travel time.

Keywords: Travel Time Prediction; Spatio-temporal; XGBoost; LightGBM; LSTM; Hybrid Models; Explainable AI; XAI; SHAP and LIME

Citation: Ahmed, I.; Kumara, I.; Reshadat, V. Travel Time Prediction and Explanation with Spatio-temporal Features: A Comparative Study. *Journal Not Specified* **2021**, *1*, 0. <https://doi.org/>

Received:

Accepted:

Published:

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Submitted to *Journal Not Specified* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Travel time refers to the time for a vehicle to reach a destination. Precise prediction of travel time leads to strong route planning and emergency services, preventing the delay of public transport, decreasing fuel consumption, traffic congestion, and environmental pollution [1–3]. The growth of online retail sales has increased the demand for express delivery services. Short-term TTP (Travel Time Prediction) is crucial for delivering goods to the customers, and the reliability of delivery time influences customer satisfaction, which is one of the most critical priorities in the logistic sector [4,5].

This paper focuses on TTP in the logistic industry. TTP studies have developed a range of different data-driven methods from statistical [6] and traditional machine learning models [7] to advanced neural network-based models [8,9]. However, a few studies applied and compared various data-driven TTP methods. In particular, there is a lack of studies on the impact of spatial and temporal travel features on the accuracy of different types of TTP models. Understanding such impact is of great significance for improving the performance level of travel planning services [10,11]. Furthermore, while data-driven TTP models achieve high performance, studies that

focus on their explainability are absent. Identifying the key parameters of trained models and the rationale behind them make sophisticated AI models understandable to human analysts. It sheds light on the inner workings of trained models and explains individual predictions which a model makes. As a result, it is easy to rely on the predictions since the meaningful explanations help analysts gain trust in them.

To address the above-mentioned limitations of the TTP literature, we set the following two research questions:

RQ1 - To what extent can data-driven methods be applied for predicting travel time using spatiotemporal features?

RQ2 - To what extent can XAI methods be applied for explaining travel time predictions?

To answer RQ1, we developed ten TTP models using the learning algorithms from three different categories of data-driven methods, namely classical machine learning, neural networks, and hybrid models. Then, we empirically compared them using three different real case data sets with spatiotemporal features. The experiments showed that temporal and spatial characteristics of journeys could significantly affect travel time. Furthermore, we observed that ensemble learning models, deep neural network models, and hybrid models could predict travel time with reasonable accuracy (R^2 of 0.83, MAE of 9.07, and RMSE of 16.27, on average). To answer RQ2, we used Explainable AI (XAI) techniques, which can provide human interpretable explanations for the models and their predictions created by machine learning algorithms [12]. In particular, we applied the two widely used XAI methods, namely SHAP [13] and LIME [14], to the TTP models. XAI methods enable users to extract plausible answers and explanations for questions such as: why specific characteristics (features) of travels are considered important by the model? How does each feature influence travel time? And which feature(s) have the most significant impact on the duration of a given trip?

The rest of the paper is organized as follows. Section 2 describes the background and related works. In Section 3, we present three data sets and our approach to developing and comparing different TTP models and applying XAI techniques to TTP models. Section 4 presents the results of our empirical studies. Section 5 discusses the answers to the research questions and implications and threats to the validity of our findings. Finally, Section 6 concludes the paper and outlines future works.

2. Travel Time Prediction Methods

2.1. An Overview

Forecasting models play an important role in the development of various artificial intelligent tasks such as fuzzy systems [15], natural language processing [16–19], expert systems [20], and computer vision [21,22]. Travel time prediction (TTP) is one of the essential but uncertain components for logistics platforms. It is challenging and requires complex traffic or data-driven models to learn complex patterns in various data sources such as weather, driver profiles, road conditions, and routes taken by the drivers [1–3]. Various studies applied many different techniques and data variables for travel time prediction. According to [1], travel time prediction methods are classified into two main categories: model-based methods and data-driven methods. Model-based methods build models based on traffic variables such as vehicle speed, traffic density, and traffic flow to predict travel time and traffic conditions over time. In contrast, data-driven methods learn hidden linear and non-linear patterns in the travel time data. In this paper, we consider data-driven methods. The TTP literature have studied many different learning algorithms [1–3], including traditional regression models [23–25], ensemble learning [26,27], deep neural networks [28–34], and hybrid models [35]. Several studies showed that the spatiotemporal information about the travels strongly impacts travel time [10,15,36].

84 2.2. Related Work: Comparative Analysis of Travel Time Prediction Methods

85 In this section, we present the studies that empirically compare different TTP
 86 methods. For predicting travel times for short horizons on the selected freeway corridors,
 87 Qiu and Fan [3] compared four different machine learning algorithms, namely decision
 88 trees (DT), random forest (RF), extreme gradient boosting (XGBoost), and long short-
 89 term memory neural network (LSTM). The data set is collected and processed from the
 90 Regional Integrated Transportation Information System (RITIS), and the predictions are
 91 based on short intervals (ranging from 15 to 60 min). The RF model is the best performer
 92 across different prediction horizons. For predicting short-term travel times, Liu *et al.* [37]
 93 evaluated the LSTM model for 16 settings of hyper-parameters for a single data set. The
 94 study has used the linear models, namely linear regression, Ridge and Lasso regression,
 95 and ARIMA, as the baseline models. The LSTM models performed better for the narrow
 96 sliding windows and longer prediction horizons.

97 Goudarzi [38] applied windowed nearest neighbor, linear regression, and Conventional
 98 Neural Networks (CNN) to predict travel times for short horizons. The neural
 99 network model provides the best results. Google Maps is used to collect the data set
 100 for travel time in this study. Another study in [11], developed a travel time prediction
 101 method using CNN to extract essential features to improve traffic information prediction
 102 performance. The travel time records of highways and alternative roads are collected
 103 and used as the evaluation data set. The results show a low mean absolute error and,
 104 therefore, an improvement in travel time prediction accuracy.

105 Adewale and Hadachi [39] built two neural network models for predicting travel
 106 time of busy routes using origin-destination travel time matrix. The data set is derived
 107 from a historical GPS data set. Experiments demonstrate that although LSTM is more
 108 susceptible to noise as time increases, both Multi-Layer Perceptron (MLP) and Long
 109 Short Term Model(LSTM) achieve good results.

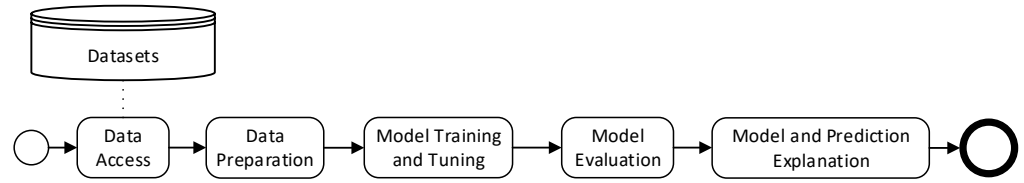
110 A study in [40] utilizes a set of Monte Carlo experiments and heat maps produced
 111 by the Layer-wise Relevance Propagation (LRP) approach for explaining the particular
 112 predictions of neural networks in the analysis of travel options. The results indicate that
 113 LRP helps gain trust in a trained ANN model, and analysts can employ it for general or
 114 travel demand analysis purposes.

115 In [41], XGBoost is applied for predicting the freeway travel time. The relative
 116 impact of each variable in the model is explained. Optimized modeling results of
 117 XGBoost are assessed and compared with the gradient boosting model. The results
 118 demonstrate that the XGBoost travel time prediction model considerably enhances the
 119 performance and efficiency. XGBoost algorithm is also utilized to develop models for
 120 different domains such as estimating the hydrogen solubility in hydrocarbons [42] and
 121 predicting flooding susceptibility [43].

122 Compared with the existing studies, which only consider a few learning methods,
 123 this paper empirically evaluates ten different models (two ensemble methods, three
 124 neural networks, four hybrid models, and one linear regression model) over three
 125 different data sets. Furthermore, we apply the XAI (Explainable Artificial Intelligence)
 126 methods to travel time prediction models to understand and discuss the importance
 127 of various features, prediction explainability, and model explainability. Explainable AI
 128 refers to methods and techniques in applying artificial intelligence technology (AI) such
 129 that human experts can understand the results of the AI-based solution. XAI methods
 130 allow human users to comprehend and trust the results and output created by complex
 131 black-box machine learning algorithms [12].

132 3. Materials and Methods

133 This section presents our methodology for travel time prediction and explanation,
 134 which is depicted in Figure 1. We used the CRISP-DM methodology (Cross-Industry
 135 Standard Process for Data Mining) [44].

**Figure 1.** Methodology

Total Number of Trips	14135
Average Number of Trips per Month	643
Average Number of Trips per Day	24
Average Number of Stops per Trip	10

Table 1: Summary of NextUp-1 Data Set

3.1. Data Understanding

This study uses three different travel time data sets. The first two data sets, namely NextUp-1 and NextUp-2, are from a logistics software company (NextUp Software ¹), which has many logistics companies as their customers who use the software to plan and monitor orders. NextUp aims to provide more reliable delivery time prediction of orders to optimize the internal processes of the logistics firms, such as trip planning and resource planning, making the life of a trip planner easier while improving customer satisfaction. The third data set, namely PeMS, was obtained from Caltrans Performance Measurement System ², which includes the data from the freeway system across all major metropolitan areas of California.

NextUp-1 and NextUp-2 data sets consist of temporal information (e.g., departure time and scheduled order delivery time) and general information (e.g., such as order information and driver information) about travels. Additionally, NextUp-2 data set also includes spatial information, e.g., travel start location and delivery location. PeMS data set contains temporal, spatial, and general information. Compared with NextUp data sets, PeMS data set include additional information such as distance and lane numbers. Tables 1, 2, and 3 summarize the three data sets. We used the domain experts at the NextUp company to validate its two data sets in terms of the essential travel information and the representativeness of the data points.

3.2. Data Preparation

After collecting data, we performed the exploratory data analysis and cleaned the raw data as necessary. In particular, duplicate data points and outliers were identified and removed, and missing values were filled using interpolation. After completing the data cleaning steps, the features were transformed and split accordingly. For example, the timestamp features were split into multiple features using units hour, minute, day of the week, and month of the year. Finally, we transformed some of the attributes in the data sets to ensure the data are in the proper format for analysis and machine learning tasks.

Total Number of Trips	5272
Average Number of Trips per Month	195
Average Number of Trips per Day	7
Average Number of Stops per Trip	7

Table 2: Summary of NextUp-1 Data Set

¹ <https://nextupsoftware.com/>

² <https://dot.ca.gov/programs/traffic-operations/mpr/pems-source>

Total Number of Travels	211392
Average Number of Travels per Route	30198
Average Number of Travels per Month	35232
Average Number of Travels per Day	1155

Table 3: Summary for PeMS Data Set

3.3. *Model Training and Tuning*

This study uses several regression algorithms from the data-driven TT prediction literature to build TTP models that can predict travel time using spatiotemporal features. They can be broadly categorized into ensemble learning, deep neural network, and hybrid. We selected a linear regression, Linear SVMR (Support Vector Machine Regression), as the baseline model. Each learning method was chosen for evaluation due to their prevalence in TT predictive analytics [1,2,28,36,37].

- Ensemble Learning Models.** Ensemble learning enhances the prediction performance of one model by training multiple models simultaneously and combining their predictive power to achieve the best performance possible [45]. Many ensemble learning methods are available, and this study considers two widely used gradient boosting methods [46,47]: XGBoost (eXtreme Gradient Boosting) and LightGBM (Light Gradient Boosting Machine). Boosting models consist of a sequence of regression trees, where every successive tree tries to correct the previous tree's mistakes. Hence, increasing the prediction accuracy of the overall model [48]. XGBoost applies level-wise (horizontal) tree growth, whereas LightGBM applies leaf-wise (vertical) tree growth. Compared with XGBOOST, LightGBM is computationally less expensive and has better prediction accuracy [46]. We used the standard hyper-parameters of the two learning algorithms [42,43,47]: `learning_rate`, `colsample_bytree`, `n_estimators`, and `max_depth` for XGBoost, and `learning_rate`, `bagging_frequency`, `n_estimators`, and `max_dept` for LightGBM.
- Deep Neural Network Models.** Neural networks are one of the most popular machine learning techniques [49]. They are represented as layered organizations of neurons with connections to other neurons, mimicking how biological neurons signal to one another. Neural networks can be used for travel time prediction as they can learn non-linear relations among variables [28–34]. This study uses long short-term memory (LSTM) and gated recurrent units (GRU) techniques of neural networks as they are more suitable for long sequence data [50]. We use the traditional LSTM and its extension, namely bidirectional LSTM, which combines a forward and a backward pass of operations, enabling considering past instances and future ones.
- Hybrid Models.** Following the TT prediction literature [37,51], we selected multiple hybrid models by combining one deep learning model and one ensemble learning model in combination with a linear model for final prediction. Figure 2 shows the architecture of the hybrid model used. In this architecture, two different types of machine learning models are combined, and then the output of those two models is passed through a linear regression model to get the final result. In this study, four hybrid models are considered: a GRU model in combination with LightGBM, a GRU model in combination with XGBoost, an LSTM model in combination with LightGBM, and an LSTM model in combination with XGBoost.
- Linear SVMR.** Support vector machine regression (SVMR) is based on statistical learning theory and can improve the ability of generalization by seeking the minimum structural risk [52]. We use the Linear SVMR model as the baseline since several TTP prediction studies use linear regression models as baselines [37,38].

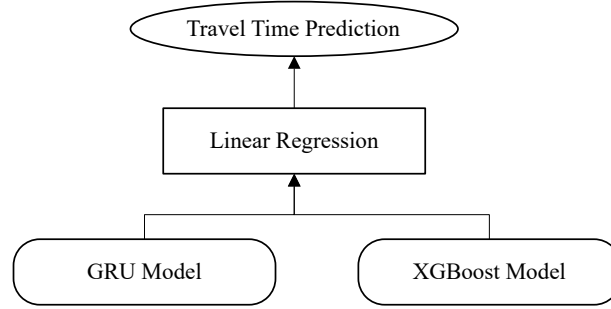


Figure 2. Architecture of Hybrid Models [51]

We tuned the models using a *grid search* on the models' hyperparameters through a *k-folds* cross-validation. Grid search is an exhaustive search algorithm through a manually-specified subset of parameters, while k-folds cross-validation is a widely used validation method that ensures that every observation from the data set has the chance of appearing in the training and test set [53]. We used 10-folds to partition the data randomly into ten folds of equal size. A single fold is used as the test set, while the remaining ones are used as the training set. The process was repeated ten times, using a different fold as the test set. Then, the model performance was reported using the mean achieved over the ten runs. We could not use this strategy for deep natural networks and hybrid models as it was computationally expensive. Therefore, we manually calibrated the regression models, and we applied *hold-out validation* [53]. We split the data set into three sets using the ratio of 3:1:1 (60% training, 20% validation, and 20% test), which is a commonly used data partition rule in machine learning model selection tasks [54].

3.4. Model Evaluation

For the evaluation of the trained models, we used the standard performance measurement metrics for regression problems:

- **R^2 Score.** It is a statistical measure that determines the proportion of variance in the dependent variable that can be explained by one or more independent variables in a regression model. R^2 score indicates how well the trained model fits the data. The score lies between 0 and 1, where a score of 0 means that the model does not capture any pattern in the data, and the predictions will be random. On the other hand, if the score is 1, the model perfectly fits the data and generalizes very well.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$$

- **RMSE.** Root mean square error or deviation is a measurement of the difference between model prediction and actual value. The deviations in predicted values from actual values are known as residual. It is calculated over the test set and is also known as prediction error. RMSE is always positive, and 0 is considered a perfect fit on the data, which is impossible to achieve in practice.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2}$$

- **MAE.** Mean absolute error is the mean of the absolute errors, differences between predicted and actual values. It indicates how big of an error we can expect from the prediction on average.

$$\text{Mean Absolute Error} = \frac{1}{n} \sum_{j=1}^n |\hat{y}_j - y_j|$$

For validating the best performing models and comparing their performance across data sets, we followed the recommendations by Demšar [55]. After establishing the statistical differences among the implemented machine learning models by applying the Friedman test [56], we used the pairwise post-hoc analysis [57]. In this analysis, the average rank is replaced by Wilcoxon signed-rank test with Holm's alpha correction. Finally, the results for model comparison are plotted using critical difference diagrams (CD-diagrams) for RMSE, MAE, and R^2 for all three data sets and machine learning models [58]. In the critical difference diagram, a thick connecting line represents similar models grouped, which means the difference in performance between those models is insignificant.

3.5. Model Explanation

Explainable AI (XAI) refers to the techniques in artificial intelligence that help humans understand and interpret the predictions made by ML models [12]. The explanations provided by XAI methods aim to give trip planners and other stakeholders insights by showing contributions of different features in travel time prediction.

There are two main types of explanations for ML models – *global* and *local* [12]. Global explanations provide an overview of the trained model as a whole and how each input variable contributes, either positively or negatively, to the prediction. As a result, one can readily understand how different features in the ML model can affect the prediction. Local explanations refer to the explanation provided for an individual prediction; they can explain why an individual instance has been assigned a specific outcome from the trained model.

In this study, we selected the two most popular XAI methods that are model-agnostic and can provide local and global explanations: SHapley Additive exPlanations (SHAP) [13] and Local Interpretable Model-agnostic Explanations (LIME) [14]. SHAP computes the contribution of each feature for a particular prediction by using Shapley values based on cooperative game theory. LIME tries to understand the model by perturbing the input of data samples and understanding how predictions change. SHAP provides mathematical guarantees for the accuracy and consistency of explanations. Several studies have comparatively analyzed the efficacy of LIME and SHAP methods to explain the prediction models used by different domains, *e.g.*, air traffic management [59] and predictive business process analytics [60].

4. Results

This section presents the results of our empirical studies for answering the two research questions of this paper.

4.1. RQ1: Comparison of TTP Methods

Table 4 summarizes the performance of the selected TTP methods on the three data sets used in the study. For NextUp-1 data set, by looking at the RMSE, MAE, and R^2 score of models, we can observe that XGBoost, LightGBM, and Hybrid models have similar performance, with XGBoost and LightGBM being the best-performing models. The hybrid models are Hybrid-1 (XGBoost+GRU+LR), Hybrid-2 (LightGBM+GRU+LR), Hybrid-3 (XGBoost+LSTM+LR), and Hybrid-4 (LightGBM+LSTM+LR). **LinearSVM performs worst among all the methods with significantly high RMSE and MAE scores and very low R^2 .** Moreover, the hybrid models do not improve the prediction accuracy significantly compared with the individual ensemble learning models. The reason behind it can be that the output of the ensemble method and the neural network do not have a linear relation with the actual output. **Hence, the performance remains the same as the best performing model among the combined models.**

Data Set / Model	NextUp-1			NextUp-2			PeMS		
	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
XGBoost	26.51	14.57	0.8083	18.31	10.64	0.7646	0.61	0.39	0.9993
LightGBM	26.54	14.40	0.8079	18.30	10.65	0.7647	0.64	0.41	0.9992
LSTM	29.97	16.59	0.7551	24.74	13.41	0.5704	0.87	0.51	0.9987
BiLSTM	29.96	16.30	0.7553	23.16	12.72	0.6234	0.93	0.55	0.9985
GRU	29.97	16.555	0.7550	25.04	13.51	0.5597	0.80	0.49	0.9989
LinearSVMR	49.48	25.77	0.3323	26.20	14.01	0.5180	3.40	1.12	0.9797
Hybrid-1	26.53	14.49	0.8080	18.80	11.34	0.7519	0.65	0.42	0.9993
Hybrid-2	26.55	14.35	0.8078	18.45	10.87	0.7611	0.67	0.43	0.9992
Hybrid-3	26.52	14.50	0.8082	18.56	11.13	0.7580	0.65	0.42	0.9993
Hybrid-4	26.55	14.35	0.8078	18.45	10.87	0.7611	0.67	0.43	0.9992

Table 4: Performance Metrics of 10 Models across Three Data Sets: Evaluation Metrics

We can observe a similar pattern of model performance for the NextUp-2 data set. However, the differences in the performance between LinearSVMR and other models are not as significant as the NextUp-1 data set. This behavior may be due to the differences in features available in two data sets; NextUp-1 contains only the temporal features while NextUp-2 includes both temporal and spatial features.

All the chosen machine learning models, including LinearSVMR, perform significantly better for the PeMS data set, which consists of travel time data of seven freeways having both temporal and spatial features, with 288 travel time observations each day for six months. While LinearSVMR is still the worst-performing as RMSE and MAE are significantly higher than other models, but R^2 score is similar to other models.

Figure 3 depicts the result of the statistical analysis we conducted on all the considered TTP models. We can see how on average, LightGBM and XGBoost were the best algorithms over the three data sets. The results also show that the differences between the performance of the predictors are not statistically significant; there is a thick line connecting the predictors in the CD diagram.

4.2. RQ2: Comparison of TTP Explanation Methods

This section evaluates the ability of the XAI techniques to provide meaningful explanations for different TTP models and their outcomes. Although SHAP and LIME can be applied to all machine learning methods as they are model-agnostic, due to brevity of space, and to avoid repetition, we provide the explanations for two best performing models only.

4.3. Global Explanations

The global explanations provided by SHAP can be visualized using different plots. Figure 4 shows the popular and informative dot plots for the XGBoost model and all three data sets. In this plot, the Y-axis indicates the TT predictors, ordered by importance; for example, *dept_hour* (departure hour) is the most important feature, and *quantity* is the least important feature for NextUp-1 data set. The X-axis represents the Shapley values. A positive Shapley value means that the corresponding feature has a positive influence on the predicted performance metric, *i.e.*, it increases the travel time. In contrast, a negative Shapley value indicates the associated feature has a negative impact on the predicted value, *i.e.*, it decreases the travel time. The color of each dot represents the value of the corresponding feature in the data set, *i.e.*, reddish colors represent higher values while bluish colors denote lower values.

We can observe that temporal and spatial information about travels significantly affects travel time. Please consider that the three data sets have similar as well as different features (see Section 3.1). According to the plots for NextUp1 and NextUp2, late departure decreases travel time, and early departure increases travel time. The delivery scheduled hour has the opposite effect. As the travel distance increases, the travel time also increases (from the plots for NextUp2 and PeMS). The Shapley values for the feature

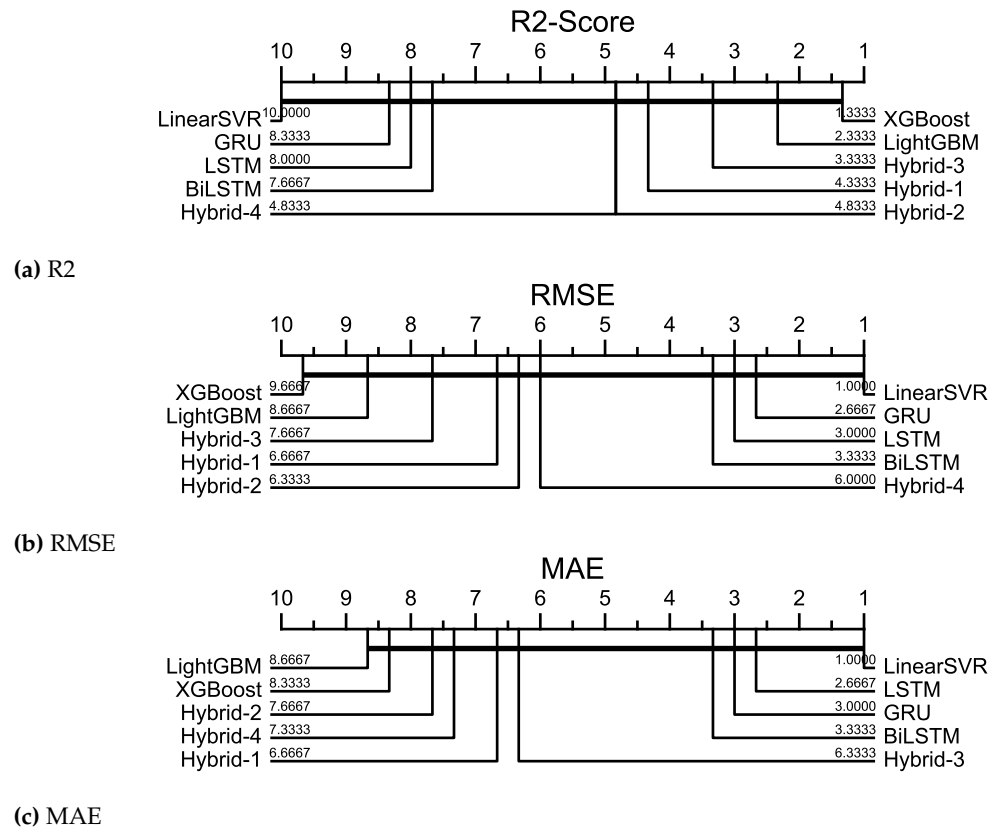


Figure 3. CD Diagrams Comparing the Performance of TTP Methods across the Three Data Sets

325 *LanePts* indicate that a vehicle traveling in a faster lane will take less time and vice versa.
 326 The Shapley values for the features *travel_start_location* and *delivery_location*) indicate
 327 that different locations can take more or less travel time.

328 It is also possible to visualize the importance of each feature by summing the
 329 absolute Shapley values of each feature across all samples. Figure 5 shows the importance
 330 of the temporal, spatial, and general properties of travels by summing the absolute
 331 Shapley values of each feature for all samples. We used the LSTM model for this plot.
 332 According to this figure, temporal properties, e.g., *dept_hour* and *scheduled_hour*, and
 333 spatial properties, e.g., *distance* and *LanePts*, are the most important features. However,
 334 at the other extreme, some temporal features such as *dept_minute* and *dept_dayofweek*
 335 have the least influence on travel time. When comparing Figure 5 and Figure 4, we can
 336 observe the differences in the importance given to the features by different learning
 337 algorithms. For example, XGBoost considered *dept_hour* is more important than *driver_id*,
 338 which is in contrast to LSTM's behavior. We can made similar observations regarding
 339 the features *end_i-d*, *dept_hour*, and *distance*.

340 4.4. Local Explanations

341 We can provide the local explanations, i.e., explanations for individual predictions,
 342 using SHAP and LIME. To visualize them with SHAP, we can employ the waterfall plot
 343 and the force plot. With these plots, we can see each feature's positive and negative
 344 contributions for a single prediction, which helps us understand which features are
 345 essential and whether a given feature affects the travel time positively or negatively.

346 Figure 6 shows three instances of using the waterfall plot for the TT predictions
 347 made by the XGBoost model. The positive effect of the feature pushes the prediction
 348 higher from the base value, as shown in red, and the negative effect pushes the prediction
 349 lower, as shown in blue. In instance 1, we can see that the feature *scheduled_hour* pushes
 350 the prediction 35.73 minutes lower, while the feature *dept_hour* pushes the prediction

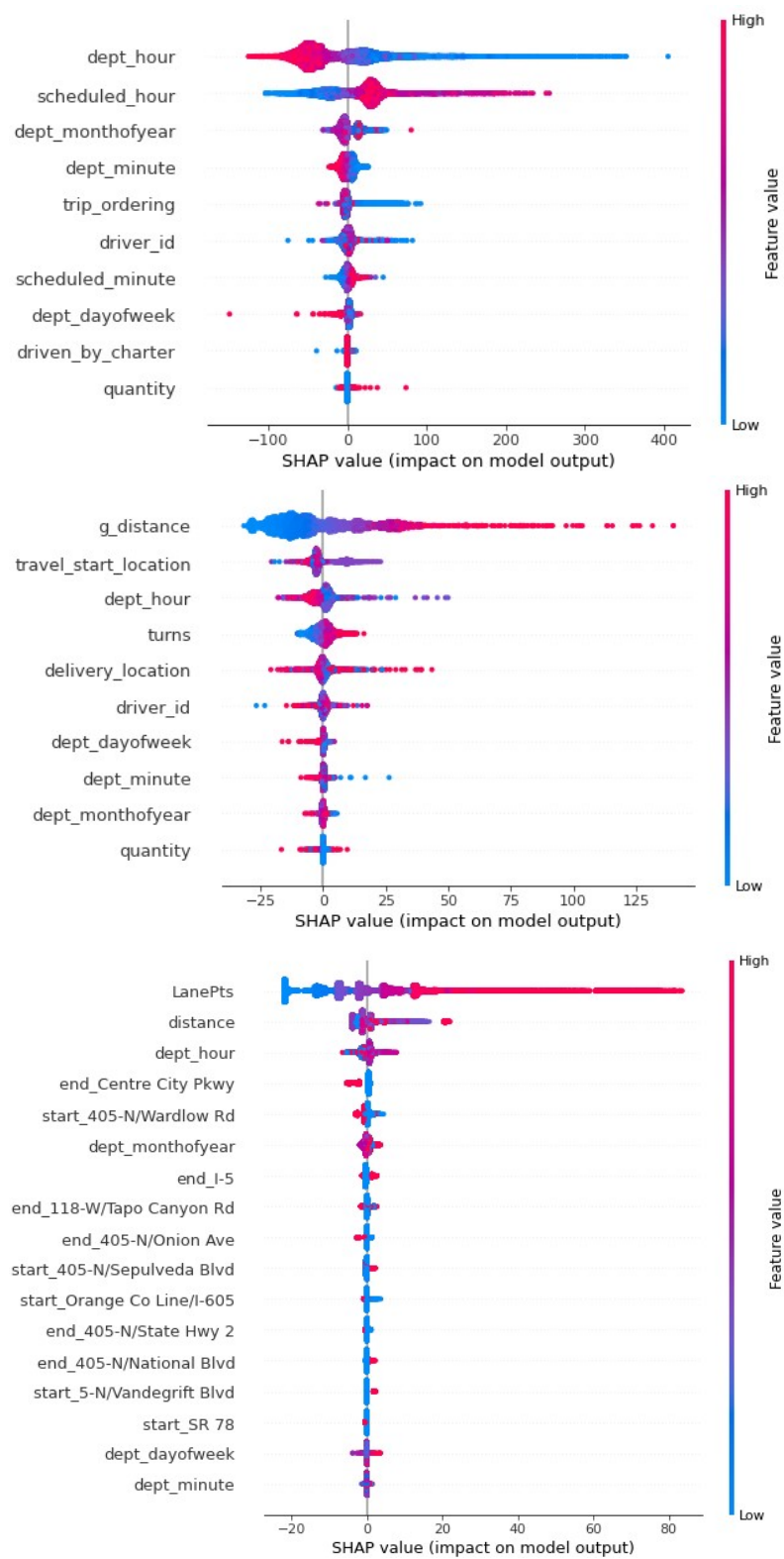


Figure 4. Global Explanations with SHAP for the XGBoost Model and Data Sets NextUp-1, NextUp2, and PeMS

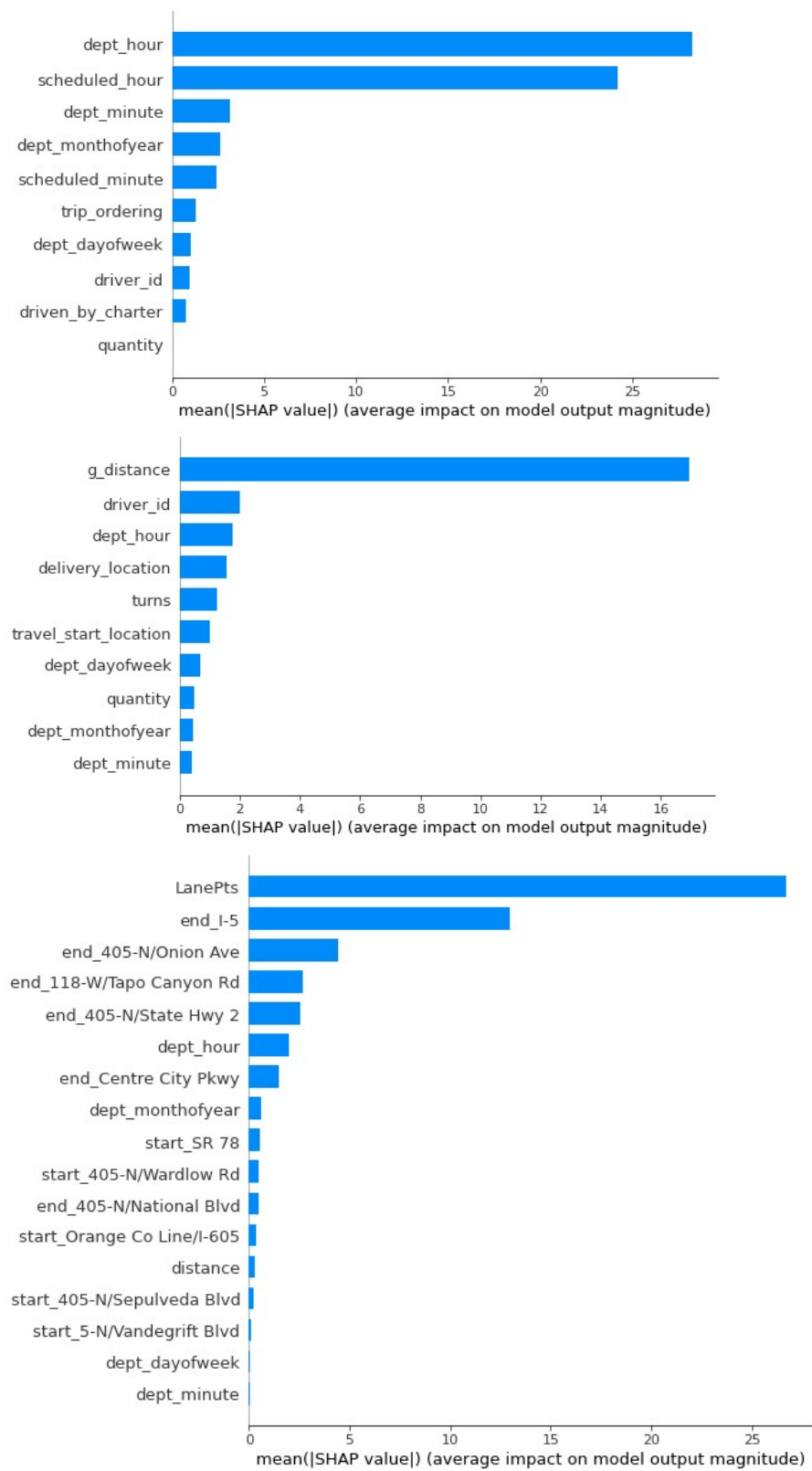


Figure 5. Overall Impact of Features on Model Outcomes, Summary Plots for the LSTM model and the Data Sets NextUp-1, NextUp2, and PeMS

34.71 minutes higher. In instance 3, the feature *dept_hour* has a negative effect of 52.56 minutes. The force plot in Figure 7 provides a similar representation of the explanations but without exact numerical contributions. Each feature pushes the prediction lower or higher depending on the importance of that feature in that particular prediction. We can observe that *scheduled_hour* and *dept_hour* are comparably more critical than any other features for the three arbitrary prediction instances considered.

LIME can also provide local explanations for model outputs, explaining the features' negative and positive contributions on individual prediction samples. However, LIME works very differently than SHAP values. LIME is inherently local as it creates a sparse linear local model around the predicted value to explain the prediction, while SHAP decomposes the prediction into contributions by each feature, and the SHAP values for each feature adds up to the final prediction.

Figure 8 depicts the explanations for three travel time prediction instances of the XGBoost model using the LIME method. The blue color shows a negative contribution to travel time, *i.e.*, increasing time, while the orange color shows a positive contribution. The plot shows the essential features, relative importance/contributions, and actual values for a given prediction. As an example, consider instance 2. The departure hour is 8 and caused an increase in the travel time. On the other hand, the scheduled hour is 12, and reduced the travel time.

5. Discussion

In this section, we first discuss the answers to the research questions of this work. Next, we outline the potential threats to external, construct, and internal validity [61] that may apply to our study.

5.1. Summary of Answers to Research Questions

With the first research question, we systematically assessed data-driven methods' ability to predict travel time accurately using spatiotemporal attributes of travels. Among the learning algorithms used, ensemble learning models, deep neural network models, and hybrid models that combined those two types of models could predict travel time with reasonable accuracy for all three data sets (R^2 of 0.83, MAE of 9.07, and RMSE of 16.27, on average). The two ensemble models, *i.e.*, XGBoost and LightSGM, were the best performing models. On the other hand, the baseline model, *i.e.*, linear SVMR, showed poor performance for two NextUp data sets and but achieved a comparable performance for PeMS data set, which is the largest data set we used in this study. Moreover, according to the statistical analysis results that we conducted using the Wilcoxon-Holm post hoc test on all the considered prediction models, there are no statistically significant differences among the performance of the models in terms of the evaluation metrics.

Hybrid models only show little or no improvement over the individual ensemble learning models XGBoost and LightSGM. However, they gained a noticeable improvement over the deep neural network models LSTM, BiLSTM, and GRU (9.7% increase in R^2 and 17.01% decrease in RSME). These findings are aligned with those of Ting *et al.* [35]. Moreover, the hybrid models are almost two times slower than LightSGM and XGBoost models concerning training time and execution time. Similarly, as the complexity of neural network architecture and the number of hidden layers increases, training a neural network model becomes computationally very expensive. In this study, we trained neural networks with a maximum of two hidden layers because of a longer training time. Thus, researchers and practitioners should be careful when selecting machine learning models for travel time prediction. The traditional machine learning models via hyperparameter tuning should be evaluated before exploring more complex models.

With the second research question, we investigated the ability of XAI methods to provide rational explanations for the predictions made by the data-driven models. Our findings from applying the two most popular XAI methods, *i.e.*, SHAP and LIME, to

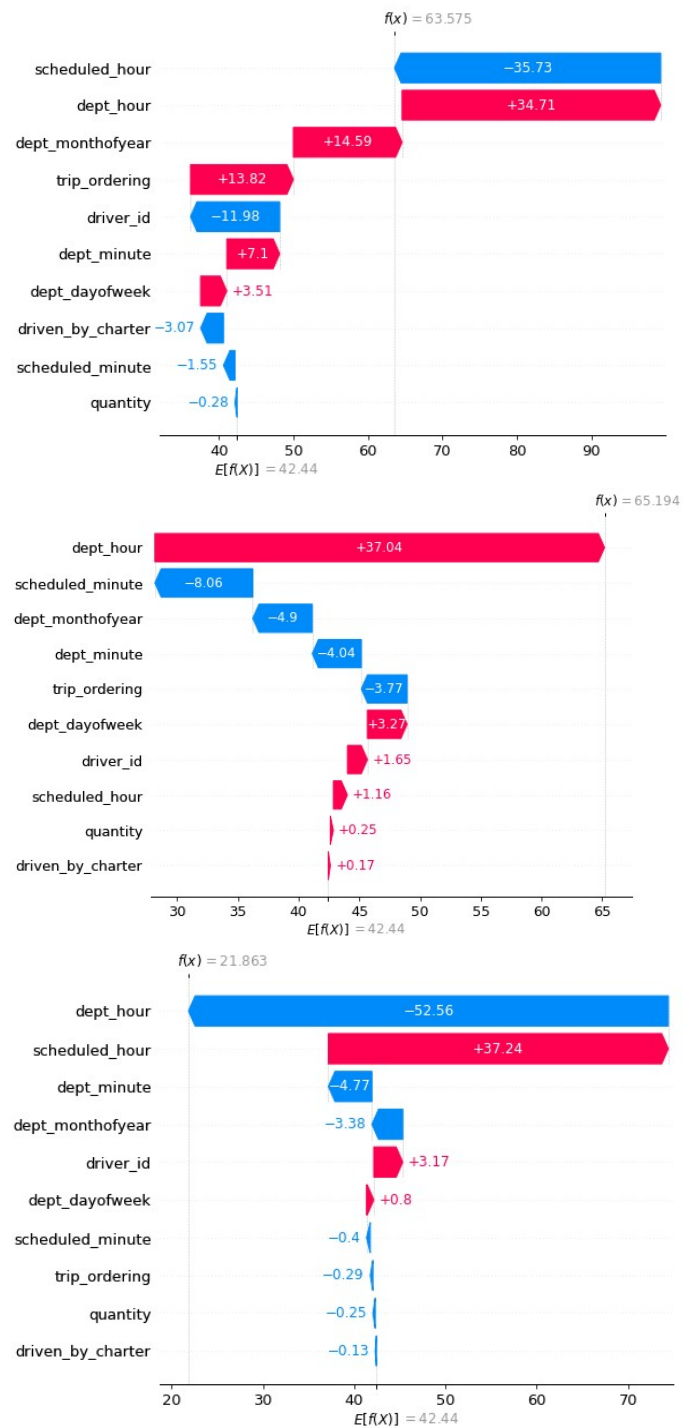


Figure 6. SHAP Waterfall Plots for Explaining Three Prediction Instances

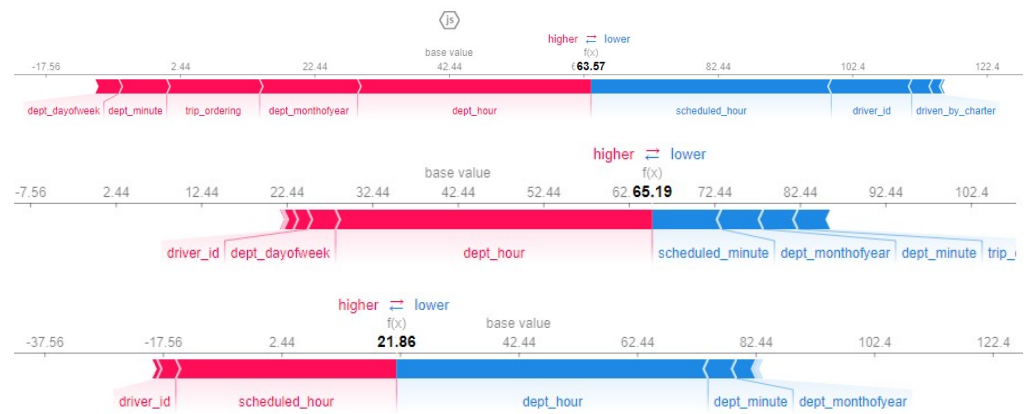


Figure 7. SHAP Force Plots for Explaining Three Prediction Instances

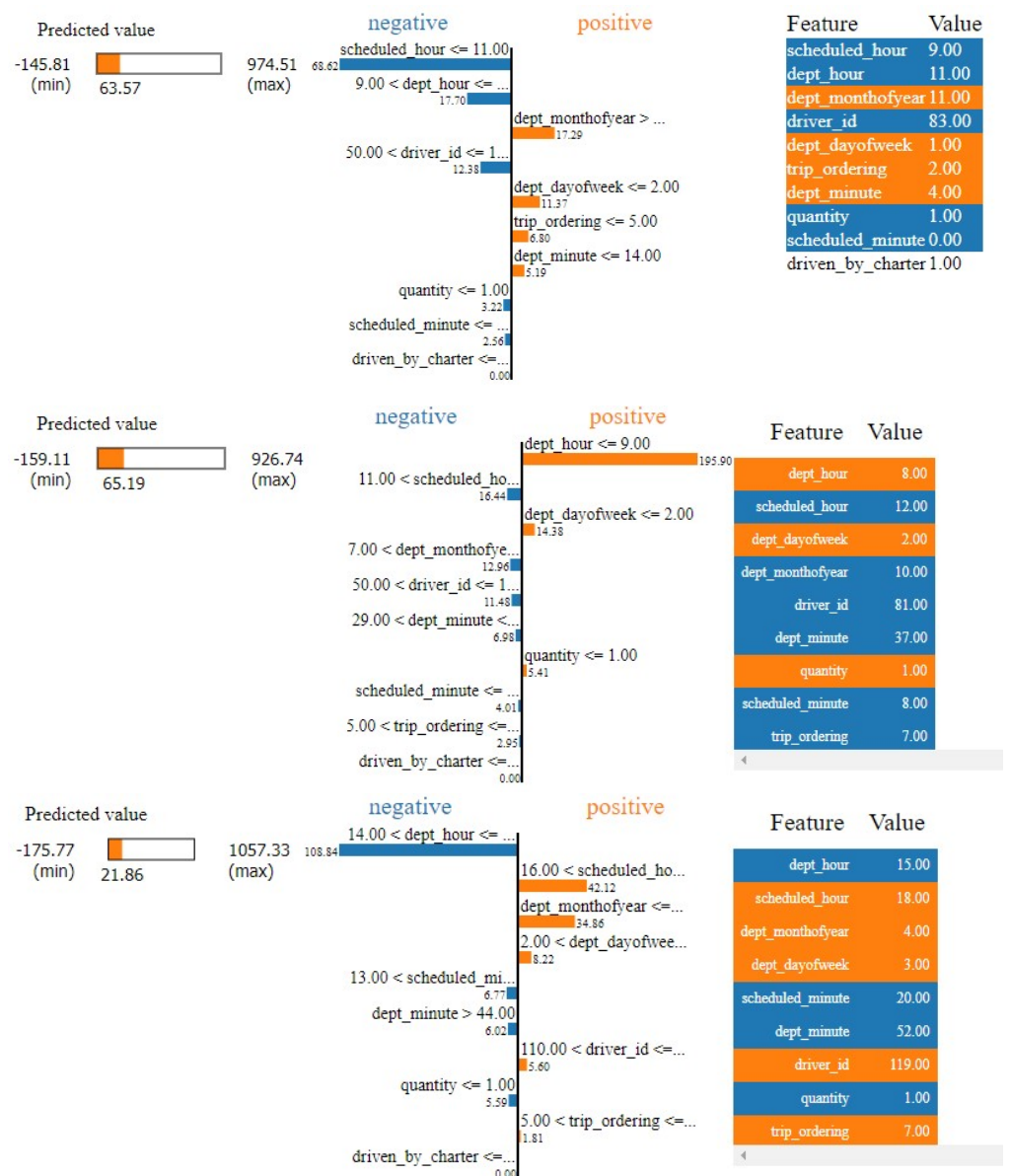


Figure 8. LIME feature Contribution Plots for Explaining Three Prediction Instances

our ten data-driven models demonstrated that XAI methods could generate intuitive explanations for TTPs. The global and local explanations provided by the XAI methods were plausible and aligned with the common knowledge about the impact of the temporal and spatial features on travel time. Thus, the XAI methods can potentially play a significant role in travel time prediction as they can help the user understand why a particular prediction is made. For example, the planners in a logistic company need to monitor the orders and know why a specific package was delivered late. In such cases, explainable AI methods can help understand the prediction outputs by explaining the contribution of each feature for every single prediction.

5.2. Threats to Validity

The data sets from the NextUp company may not accurately represent the actual characteristics of the travels. We partially mitigated this threat by using the domain experts at NextUp to validate the quality of the collected data sets. We plan to further reduce this threat by gathering more travel time data from NextUp's customers who use the transport management software developed by NextUp.

The features used to build the machine learning models could influence the accuracy of the travel time predictions. We partially mitigated this threat by training the models using the features that are considered necessary by the domain experts at NextUp. As a follow-up study, with new data sets, we plan to evaluate the impact of the travel information that was absent in the data sets we used, such as customer type, area type, weather, and driver profile.

Model-based methods such as queuing theory and cell transmission model can also be employed to build TTP models [1]. They can use traffic variables such as the speed of a vehicle, traffic density, and traffic flow to predict travel time and traffic conditions over time. Within the scope of our study, we only considered the data-driven methods, which is the most popular approach according to the TTP literature [1–3]. We plan to extend our empirical study to evaluate model-based methods over multiple data sets with spatiotemporal features.

The explanations provided by XAI methods were plausible and aligned with the common knowledge about the impact of spatiotemporal features on travel duration. However, a separate empirical study is necessary to generalize and validate the usefulness of XAI for addressing the interpretability of ML-based travel time prediction models and validating the outcomes by conducting a survey with users, which is part of our research agenda.

6. Conclusions and Future Work

In this paper, we investigated two key research issues related to travel time prediction (TTP): 1) data-driven TTP methods' ability to predict travel time by considering the spatiotemporal characteristics of journeys, and 2) XAI's methods' ability to explain decisions made by the TTP models rationally. First, we implemented various TTP methods, *i.e.*, ensemble learning, neural networks, linear support vector machine regression (LinearSVMR), and hybrid models that combine ensemble learning methods and neural networks. Second, we compared these predictive models over three different data sets that include the data points from various spatiotemporal features. Compared with the baseline LinearSVMR model, the other models showed consistently good performance over the three data sets (R^2 of 0.83, MAE of 9.07, and RMSE of 16.27, average). Our findings indicate that practitioners can build effective TTP models with the careful selection and application of data-driven methods.

Finally, we assessed the ability of SHAP and LIME XAI methods to explain the travel time predictions made by black-box TTP models. Our results demonstrated that XAI methods could help practitioners to understand what factors affect travel time to what extent and how to predict travel time effectively. In particular, the explanations showed that the temporal and spatial features and their correlations could significantly

455 affect travel time. The travel planners can use such explanations to diagnose delayed
 456 trips or order deliveries. They can also identify the specific characteristics of the journeys
 457 that determine travel time in general.

458 For future work, we plan to investigate the data quality issues in travel time
 459 prediction and the importance of other features that affect travel time, such as customer
 460 type, road type, area type, and driver profile data. Furthermore, the TTP literature
 461 includes several promising model-based techniques [1], and thus, we also aim to expand
 462 the empirical study presented in this paper by comparing the data-driven methods
 463 with the model-based methods. Finally, through a user study, we will investigate the
 464 usefulness of XAI in interpreting TTP models.

Author Contributions: Conceptualization, experiment, implementation and evaluation I.A; supervised the research; writing - original draft, editing, and review, I.K, and V.R.; writing-review and editing, A.S.M.K, W.J.V.D.H, and D.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. We would like to thank the Google research and education programme for Google Cloud Platform (GCP) credits.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bai, M.; Lin, Y.; Ma, M.; Wang, P. Travel-Time Prediction Methods: A Review. *Smart Computing and Communication*; Qiu, M., Ed.; Springer International Publishing: Cham, 2018; pp. 67–77.
2. Oh, S.; Byon, Y.J.; Jang, K.; Yeo, H. Short-term Travel-time Prediction on Highway: A Review of the Data-driven Approach. *Transport Reviews* **2015**, *35*, 4–32, [<https://doi.org/10.1080/01441647.2014.992496>]. doi:10.1080/01441647.2014.992496.
3. Qiu, B.; Fan, W.D. Machine Learning Based Short-Term Travel Time Prediction: Numerical Results and Comparative Analyses. *Sustainability* **2021**, *13*. doi:10.3390/su13137454.
4. Teresa, G.; Evangelos, G. Importance of logistics services attributes influencing customer satisfaction. 2015 4th International Conference on Advanced Logistics and Transport (ICALT), 2015, pp. 53–58. doi:10.1109/ICAdLT.2015.7136590.
5. Li, S.; Ragu-Nathan, B.; Ragu-Nathan, T.; Rao, S.S. The impact of supply chain management practices on competitive advantage and organizational performance. *Omega* **2006**, *34*, 107–124.
6. Tang, J.; Zheng, L.; Han, C.; Yin, W.; Zhang, Y.; Zou, Y.; Huang, H. Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review. *Analytic methods in accident research* **2020**, *27*, 100123.
7. Cheng, J.; Li, G.; Chen, X. Research on travel time prediction model of freeway based on gradient boosting decision tree. *IEEE access* **2018**, *7*, 7466–7480.
8. Abdollahi, M.; Khaleghi, T.; Yang, K. An integrated feature learning approach using deep learning for travel time prediction. *Expert Systems with Applications* **2020**, *139*, 112864.
9. Petersen, N.C.; Rodrigues, F.; Pereira, F.C. Multi-output bus travel time prediction with convolutional LSTM neural network. *Expert Systems with Applications* **2019**, *120*, 426–435.
10. Zhao, J.; Qu, Q.; Zhang, F.; Xu, C.; Liu, S. Spatio-Temporal Analysis of Passenger Travel Patterns in Massive Smart Card Data. *IEEE Transactions on Intelligent Transportation Systems* **2017**, *18*, 3135–3146. doi:10.1109/TITS.2017.2679179.
11. Chen, C.H. Temporal-Spatial Feature Extraction Based on Convolutional Neural Networks for Travel Time Prediction. *arXiv preprint arXiv:2111.00149* **2021**.
12. Molnar, C. *Interpretable Machine Learning*; Lulu. com, 2020.
13. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 2017, pp. 4765–4774.
14. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.
15. Khetarpaul, S.; Gupta, S.; Malhotra, S.; Subramaniam, L.V. Bus arrival time prediction using a modified amalgamation of fuzzy clustering and neural network on spatio-temporal data. *Australasian Database Conference*. Springer, 2015, pp. 142–154.
16. RESHADAT, V.; HOORALI, M.; FAILI, H. A hybrid method for open information extraction based on shallow and deep linguistic analysis. *Interdisciplinary Information Sciences* **2016**, *22*, 87–100.
17. Reshadat, V.; Faili, H. A new open information extraction system using sentence difficulty estimation. *Computing and Informatics* **2019**, *38*, 986–1008.
18. Reshadat, V.; Feizi-Derakhshi, M.R. Studying of semantic similarity methods in ontology. *Research Journal of Applied Sciences, Engineering and Technology* **2012**, *4*, 1815–1821.
19. Reshadat, V.; Hourali, M.; Faili, H. Confidence Measure Estimation for Open Information Extraction. *Information Systems & Telecommunication* **2018**, *p. 1*.
20. Wentworth, J. Expert systems in transportation. Technical report, AAAI Technical Report WS-93-04, 1993.

21. Kakani, V.; Nguyen, V.H.; Kumar, B.P.; Kim, H.; Pasupuleti, V.R. A critical review on computer vision and artificial intelligence in food industry. *Journal of Agriculture and Food Research* **2020**, *2*, 100033.
22. Vyborny, C.J.; Giger, M.L. Computer vision and artificial intelligence in mammography. *AJR. American journal of roentgenology* **1994**, *162*, 699–708.
23. Zhang, Y.; Haghani, A. A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies* **2015**, *58*, 308–324.
24. Zahid, M.; Chen, Y.; Jamal, A.; Mamadou, C.Z. Freeway Short-Term Travel Speed Prediction Based on Data Collection Time-Horizons: A Fast Forest Quantile Regression Approach. *Sustainability* **2020**, *12*. doi:10.3390/su12020646.
25. Cristóbal, T.; Padrón, G.; Quesada-Arencibia, A.; Alayón, F.; de Blasio, G.; García, C.R. Bus Travel Time Prediction Model Based on Profile Similarity. *Sensors* **2019**, *19*. doi:10.3390/s19132869.
26. Chen, Z.; Fan, W. A Freeway Travel Time Prediction Method Based on an XGBoost Model. *Sustainability* **2021**, *13*. doi:10.3390/su13158577.
27. Zhang, F.; Zhu, X.; Hu, T.; Guo, W.; Chen, C.; Liu, L. Urban Link Travel Time Prediction Based on a Gradient Boosting Method Considering Spatiotemporal Correlations. *ISPRS International Journal of Geo-Information* **2016**, *5*. doi:10.3390/ijgi5110201.
28. Ran, X.; Shan, Z.; Fang, Y.; Lin, C. An LSTM-based method with attention mechanism for travel time prediction. *Sensors* **2019**, *19*, 861.
29. Wang, Z.; Fu, K.; Ye, J. Learning to estimate the travel time. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 858–866.
30. Li, X.; Wang, H.; Sun, P.; Zu, H. Spatiotemporal Features—Extracted Travel Time Prediction Leveraging Deep-Learning-Enabled Graph Convolutional Neural Network Model. *Sustainability* **2021**, *13*. doi:10.3390/su13031253.
31. Yuan, Y.; Shao, C.; Cao, Z.; He, Z.; Zhu, C.; Wang, Y.; Jang, V. Bus Dynamic Travel Time Prediction: Using a Deep Feature Extraction Framework Based on RNN and DNN. *Electronics* **2020**, *9*. doi:10.3390/electronics9111876.
32. Wu, J.; Wu, Q.; Shen, J.; Cai, C. Towards Attention-Based Convolutional Long Short-Term Memory for Travel Time Prediction of Bus Journeys. *Sensors* **2020**, *20*. doi:10.3390/s20123354.
33. Ran, X.; Shan, Z.; Fang, Y.; Lin, C. A Convolution Component-Based Method with Attention Mechanism for Travel-Time Prediction. *Sensors* **2019**, *19*. doi:10.3390/s19092063.
34. Ran, X.; Shan, Z.; Fang, Y.; Lin, C. An LSTM-Based Method with Attention Mechanism for Travel Time Prediction. *Sensors* **2019**, *19*. doi:10.3390/s19040861.
35. Ting, P.Y.; Wada, T.; Chiu, Y.L.; Sun, M.T.; Sakai, K.; Ku, W.S.; Jeng, A.A.K.; Hwu, J.S. Freeway Travel Time Prediction Using Deep Hybrid Model—Taking Sun Yat-Sen Freeway as an Example. *IEEE Transactions on Vehicular Technology* **2020**, *69*, 8257–8266.
36. Yang, S.; Qian, S. Understanding and Predicting Travel Time with Spatio-Temporal Features of Network Traffic Flow, Weather and Incidents. *IEEE Intelligent Transportation Systems Magazine* **2019**, *11*, 12–28. doi:10.1109/ITS.2019.2919615.
37. Liu, Y.; Wang, Y.; Yang, X.; Zhang, L. Short-term travel time prediction by deep learning: A comparison of different LSTM-DNN models. 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), 2017, pp. 1–8. doi:10.1109/ITSC.2017.8317886.
38. Goudarzi, F. Travel Time Prediction: Comparison of Machine Learning Algorithms in a Case Study. 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 2018, pp. 1404–1407. doi:10.1109/HPCC/SmartCity/DSS.2018.00232.
39. Adewale, A.E.; Hadachi, A. Neural Networks Model for Travel Time Prediction Based on ODTravel Time Matrix. *arXiv preprint arXiv:2004.04030* **2020**.
40. Fiosina, J. Explainable Federated Learning for Taxi Travel Time Prediction. VEHITS, 2021, pp. 670–677.
41. Fan, W.D.; Chen, Z.; others. Predicting Travel Time on Freeway Corridors: Machine Learning Approach. Technical report, University of North Carolina at Charlotte. Center for Advanced Multimodal . . . , 2020.
42. Mohammadi, M.R.; Hadavimoghaddam, F.; Pourmahdi, M.; Atashrouz, S.; Munir, M.T.; Hemmati-Sarapardeh, A.; Mosavi, A.H.; Mohaddespour, A. Modeling hydrogen solubility in hydrocarbons using extreme gradient boosting and equations of state. *Scientific reports* **2021**, *11*, 1–20.
43. Janizadeh, S.; Vafakhah, M.; Kapelan, Z.; Mobarghaee Dinan, N. Hybrid XGboost model with various Bayesian hyperparameter optimization algorithms for flood hazard susceptibility modeling. *Geocarto International* **2021**, pp. 1–20.
44. Martínez-Plumed, F.; Contreras-Ochando, L.; Ferri, C.; Orallo, J.H.; Kull, M.; Lachiche, N.; Quintana, M.J.R.; Flach, P.A. CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering* **2019**.
45. Sagi, O.; Rokach, L. Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery* **2018**, *8*, e1249, [<https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1249>]. doi:https://doi.org/10.1002/widm.1249.
46. Liang, W.; Luo, S.; Zhao, G.; Wu, H. Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and LightGBM Algorithms. *Mathematics* **2020**, *8*. doi:10.3390/math8050765.
47. Mei, Z.; Xiang, F.; Zhen-hui, L. Short-Term Traffic Flow Prediction Based on Combination Model of Xgboost-Lightgbm. 2018 International Conference on Sensor Networks and Signal Processing (SNSP), 2018, pp. 322–327. doi:10.1109/SNSP.2018.00069.

48. Zhang, Y.; Haghani, A. A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies* **2015**, *58*, 308–324. Big Data in Transportation and Traffic Engineering, doi:<https://doi.org/10.1016/j.trc.2015.02.019>.
49. Pouyanfar, S.; Sadiq, S.; Yan, Y.; Tian, H.; Tao, Y.; Reyes, M.P.; Shyu, M.L.; Chen, S.C.; Iyengar, S.S. A Survey on Deep Learning: Algorithms, Techniques, and Applications. *ACM Comput. Surv.* **2018**, *51*. doi:10.1145/3234150.
50. Liu, Y.; Wang, Y.; Yang, X.; Zhang, L. Short-term travel time prediction by deep learning: A comparison of different LSTM-DNN models. 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2017, pp. 1–8.
51. Ting, P.Y.; Wada, T.; Chiu, Y.L.; Sun, M.T.; Sakai, K.; Ku, W.S.; Jeng, A.A.K.; Hwu, J.S. Freeway Travel Time Prediction Using Deep Hybrid Model – Taking Sun Yat-Sen Freeway as an Example. *IEEE Transactions on Vehicular Technology* **2020**, *69*, 8257–8266. doi:10.1109/TVT.2020.2999358.
52. Zhang, J.; Liao, Y.; Wang, S.; Han, J. Study on driving decision-making mechanism of autonomous vehicle based on an optimized support vector machine regression. *Applied Sciences* **2018**, *8*, 13.
53. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An introduction to statistical learning*; Vol. 112, Springer, 2013.
54. Lever, J.; Krzywinski, M.; Altman, N. Model selection and overfitting. *Nature Methods* **2016**, *13*, 703–704. doi:10.1038/nmeth.3968.
55. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research* **2006**, *7*, 1–30.
56. Zimmerman, D.W.; Zumbo, B.D. Relative power of the Wilcoxon test, the Friedman test, and repeated-measures ANOVA on ranks. *The Journal of Experimental Education* **1993**, *62*, 75–86.
57. Benavoli, A.; Corani, G.; Mangili, F. Should we really use post-hoc tests based on mean-ranks? *The Journal of Machine Learning Research* **2016**, *17*, 152–161.
58. Fawaz, H.I.; Forestier, G.; Weber, J.; Idoumghar, L.; Muller, P.A. Deep learning for time series classification: a review. *Data mining and knowledge discovery* **2019**, *33*, 917–963.
59. Xie, Y.; Pongsakornsathien, N.; Gardi, A.; Sabatini, R. Explanation of Machine-Learning Solutions in Air-Traffic Management. *Aerospace* **2021**, *8*. doi:10.3390/aerospace8080224.
60. Velmurugan, M.; Ouyang, C.; Moreira, C.; Sindhgatta, R. Evaluating Fidelity of Explainable Methods for Predictive Process Analytics. *Intelligent Information Systems*; Nurcan, S.; Korthaus, A., Eds.; Springer International Publishing: Cham, 2021; pp. 64–72.
61. Wohlin, C.; Runeson, P.; Höst, M.; Ohlsson, M.C.; Regnell, B.; Wesslén, A. *Experimentation in software engineering*; Springer Science & Business Media, 2012.