*Article*

# 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:** Logistic firms produce an extensive amount of data that can be used to optimize many processes in logistics, one of which is travel time. They can use travel time information as input or auxiliary data for such tasks as dynamic navigation, infrastructure planning, congestion control, and accident detection. Travel-time prediction (TTP) refers to the prediction of current or future travel time. In recent years, a variety of learning-based TTP methods that consider the influence of both spatial and temporal features have been proposed. 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. Moreover, there is also a lack of research on explaining travel time predictions made by black-box prediction 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 have applied XAI (Explainable Artificial Intelligence) methods (SHAP and LIME) to understand and interpret models' predictions. The prediction accuracy and reliability for all models are compared. We observed that ensemble learning methods, *i.e.,* XGBoost and LightGBM, have the most consistent performance over different data sets. XAI methods can adequately explain how various features affect travel time.

**Keywords:** Travel time prediction; machine learning; deep neural networks; explainable AI; XAI; spatio-temporal; hybrid models; ensemble learning; and LSTM

**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:

## 1. Introduction

Travel time refers to the time for a vehicle to reach a destination. Precise prediction of travel time is crucial in developing Intelligent Transportation Systems (ITS). It leads to strong route planning and emergency services, decreasing fuel consumption, traffic congestion, and environmental pollution, preventing the delay of public transport [1]. Effective transportation planning helps businesses to cut costs and increase their market competitiveness. The growth of online retail sales has increased the demand for express delivery services. The logistics industry is continuously trying to improve performance by reducing costs and optimizing operations [2]. Customer satisfaction is one of the most crucial priorities in the logistic sector and is influenced by the reliability of delivery time [3][4].

Forecasting models play an important role in the development of various artificial intelligent tasks such as fuzzy systems [5], natural language processing [6–9], expert systems [10], and computer vision [11][12]. Numerous researches have been conducted on forecasting arrival times of vehicles in recent years. These studies support a range of different methods from statistical [13] and traditional machine learning models [14] to advanced neural network-based models [15,16].

This paper focuses on travel time prediction (TTP) in the logistic industry for improving the delivery time. Transport Management systems (TMS) have an extensive amount of useful information that is used for developing a reliable approach for predicting the arrival/delivery times with the help of machine learning models [1,17,18]. However, there are a few studies that applied and compared various data-driven TTP methods. Furthermore, while machine learning models achieve high performance, there is an absence of studies that focus on explainability. It is difficult to trust predictions made by a model without assessing the relationships embedded in the model. Model transparency is of great importance for transport planning agencies. It sheds light on the inner workings of trained models and explains individual predictions which are made by a model. As a result, it is easy for analysts to gain trust in models.

In this paper, we develop and compare three different types of machine learning models, namely classical machine learning, neural networks, and hybrid models, with ten different methods on three different real case data sets for predicting the train arrival time. We mainly focus on the influence of spatial and temporal information about the travels on the travel time. Moreover, by applying some XAI methods (SHAP and LIME) [19], we inspect the rationale behind models' predictions. The results of the experiments show that ensemble tree-based learning models, namely XGBoost and LightGBM, outperform other models over different data sets. 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 application performance? And which features/s have the greatest impact on the performance of a given trip?

The rest of the paper is organized as follows: In Section 2, the related works are reviewed; Section 3 presents the proposed approach in detail; Section 4 is dedicated to the details of the data sets, experiments, and results. Section 5 discusses the findings, and finally, Section 6 concludes the paper.

## 2. Travel Time Prediction Methods

### *2.1. An Overview*

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, routes taken by the drivers [1,17,18]. 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 [1,17,18], including traditional regression models [20–22], ensemble learning [23,24], deep neural networks [25–31], and hybrid models [32]. Several studies showed that the spatio-temporal information about the travels strongly impacts travel time [5,33,34].

### *2.2. Comparative Analysis of Travel Time Prediction Methods*

In this section, we present the studies that empirically compare different TTP methods. For predicting travel times for short horizons on the selected freeway corridors, Qiu and Fan [18] compared four different machine learning algorithms, namely decision trees (DT), random forest (RF), extreme gradient boosting (XGBoost), and long short-term memory neural network (LSTM). The RF model was the best performer across different prediction horizons. For predicting short-term travel times, Liu *et al.* [35] evaluated the LSTM model for 16 settings of hyper-parameters for a single data set. The LSTM models performed better for the narrow sliding windows and longer prediction horizons. The study used the linear models, namely linear regression, Ridge and Lasso regression, and ARIMA, as the baseline models. Goudarzi [36] applied windowed nearest neighbor, linear regression, and conventional neural networks (CNN) to predict travel times for short horizons. The neural network model provided the best results.

Compared with the existing studies, which only consider a few learning methods, this paper empirically evaluates ten different models (two ensemble methods, three neural networks, four hybrid models, and one linear regression model) over three different data sets. Furthermore, we apply the XAI (Explainable Artificial Intelligence) methods to TT prediction models to understand and discuss the importance of various features, prediction explainability, and model explainability. Explainable AI refers to methods and techniques in applying artificial intelligence technology (AI) such that human experts can understand the results of the AI based solution. XAI methods allow human users to comprehend and trust the results and output created by complex black-box machine learning algorithms [19].

## 3. Materials and Methods

This section presents our methodology for travel time prediction and explanation, which is depicted in Fig. 1. We used the CRISP-DM methodology (Cross-Industry Standard Process for Data Mining) [37].

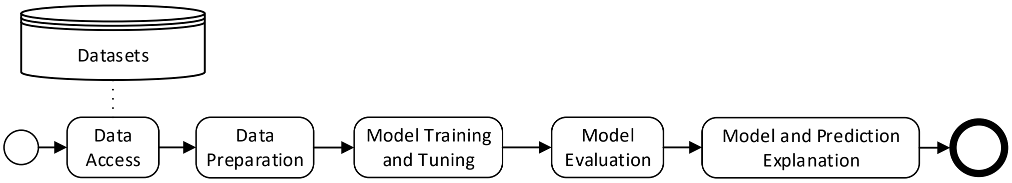


**Figure 1.** Methodology

### *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 TT predictions of orders to optimize other internal processes of logistics firms like 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 include 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 data sets.

[1] https://nextupsoftware.com/

[2] https://dot.ca.gov/programs/traffic-operations/mpr/pems-source

| | |
|---|---|
| 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

| | |
|---|---|
| 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

### *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 are transformed and split accordingly. For example, timestamp features are split into multiple features like 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 right format for analysis and machine learning tasks.

### *3.3. Machine Learning Method Section*

In this study, we use several regression algorithms from the data-driven TT prediction literature. They can be broadly categorized into ensemble learning, deep neural network, and hybrid. As the baseline model, we selected a linear regression, Linear SVMR (Support Vector Machine Regression). Each learning method was chosen for evaluation due to their prevalence in TT predictive analytics [1,17,25,33,35].

- **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 [38]. Many ensemble learning methods are available and this study considers two widely used gradient boosting methods: 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 [39].
- **Deep Neural Network Models.** Neural networks are one of the most popular machine learning techniques [40]. 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 [25–31]. 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 [41]. We use the traditional LSTM and its extension, namely bidirectional LSTM, which combines a forward and a backward pass of operations, allowing for considering past instances and future ones.

| | |
|---|---|
| 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

- **Hybrid Models.** Following the TT prediction literature [35,42], 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 [43]. We use the Linear SVMR model as the baseline since several TTP prediction studies use linear regression models as baselines [35,36].

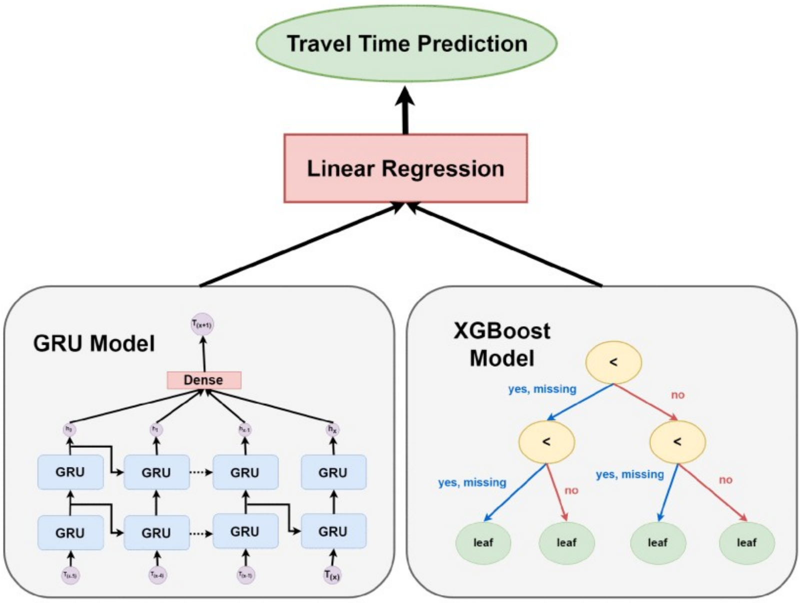


**Figure 2.** Architecture of Hybrid Models [42]

### 3.4. Model Selection

We used a *grid search* on the models' hyperparameters through a *k-folds* (k=10) cross-validation to select the best model. 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 [44]. It randomly partitions the data into 10 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 10 times, using each time a different fold as the test set. Then, the average model performance over 10 runs was calculated. We could not use this strategy for deep natural networks and hybrid models as they were computationally expensive. Therefore, we manually calibrated the regression models, and we applied *hold-out validation* [44]. We split the data set into three sets (60% training, 20% validation, and 20% test).

### 3.5. Model Evaluation

For the evaluation of the trained model, we used the standard machine learning 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.

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

For validating the best performing models and comparing their performance across data sets, we followed the recommendations by Demšar [45]. After establishing the statistical differences among the implemented machine learning models by applying the Friedman test [46], we used the pairwise post-hoc analysis [47]. In this analysis, the average rank is replaced by Wilcoxon signed-rank test with Holm's alpha correction. 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 [48]. In the critical difference diagram, a thick connecting line represents similar models grouped together, which means the difference in performance between those models is not significant.

### *3.6. Model Explanation*

Explainable AI (XAI) refers to the techniques in artificial intelligence that help humans understand and interpret the predictions made by ML models [19]. The explanations provided by XAI methods aim to give trip planners and other stakeholders insights by showing contributions of different features in TT prediction.

There are two main types of explanations for ML models – *global* and *local* [19]. 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) [49] and Local Interpretable Model-agnostic Explanations (LIME) [50]. SHAP calculates the contribution of each feature for a particular prediction by using Shapley values based on cooperative game theory. LIME attempts to understand the model by perturbing the input of data samples and understanding how predictions change. SHAP

| **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 of 10 Models across Three Data Sets: Evaluation Metrics

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 [51] and predictive business process analytics [52].

## 4. Results

### *4.1. Research Questions*

We set the following research questions for the empirical evaluation of the methods for travel time prediction and explanation:

**RQ1** - Which predictive model performs best in predicting travel time on a particular data set with different features?

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

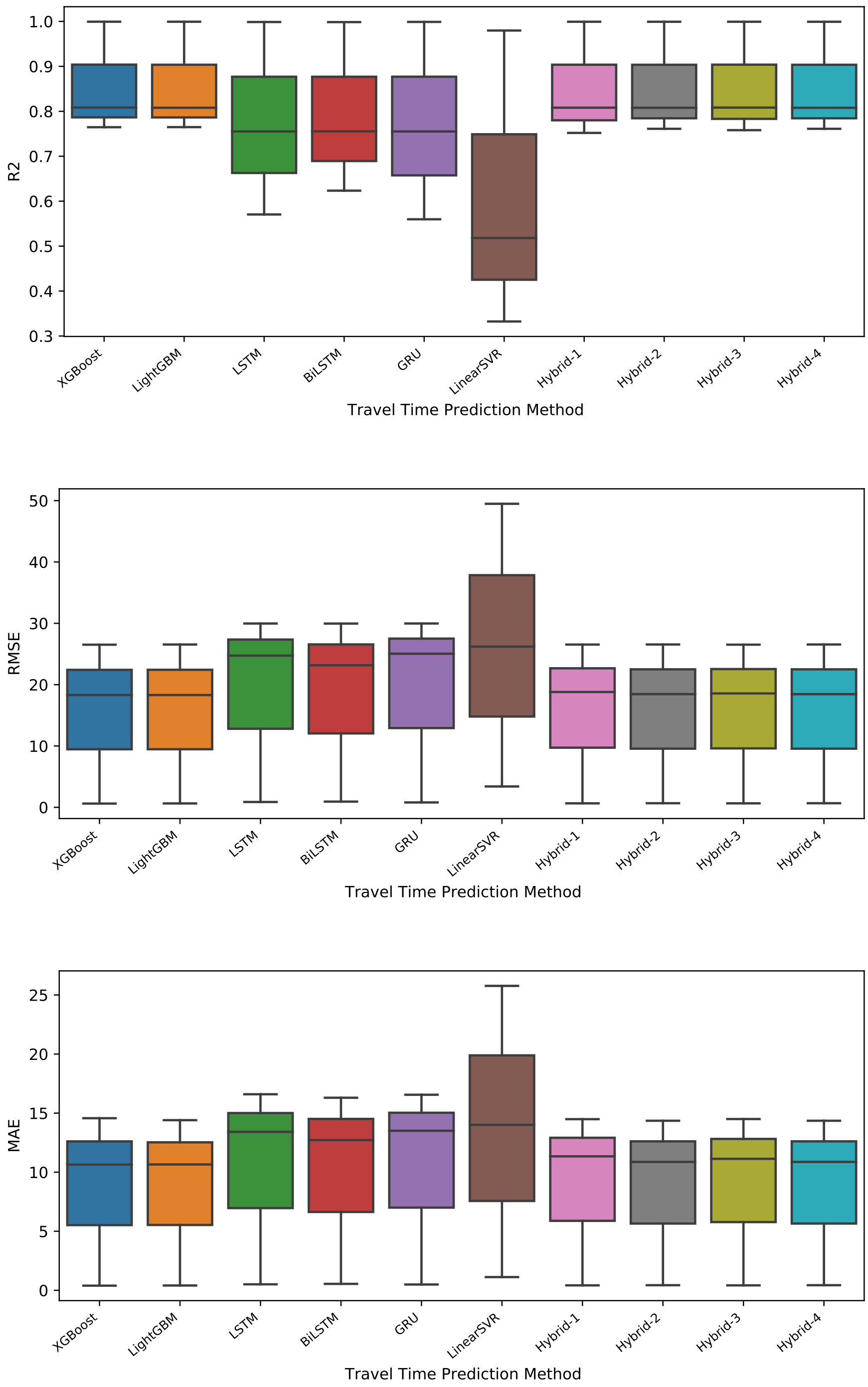
### *4.2. RQ1: Comparison of TTP Methods*

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 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), Hybrid-4 (LightGBM+LSTM+LR). While LinearSVMR is performing worst among all the methods with significantly high RMSE and MAE scores, which is considered better if lower, and very low $R^2$, which is considered better when higher. 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 of the best performing model among the combination of models.

We can observe a similar pattern of model performance for NextUp-2 data set, but the difference of performance between LinearSVMR and other models is not as big as it was for NextUp-1 data set. This 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 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 close to other models, which is a good sign, meaning that the models generalize the data well.

Figure 3 illustrates the variation in performance of each model across different data sets using the box plots for the $R^2$, RMSE, and MAE metrics. Figure 4 depicts the

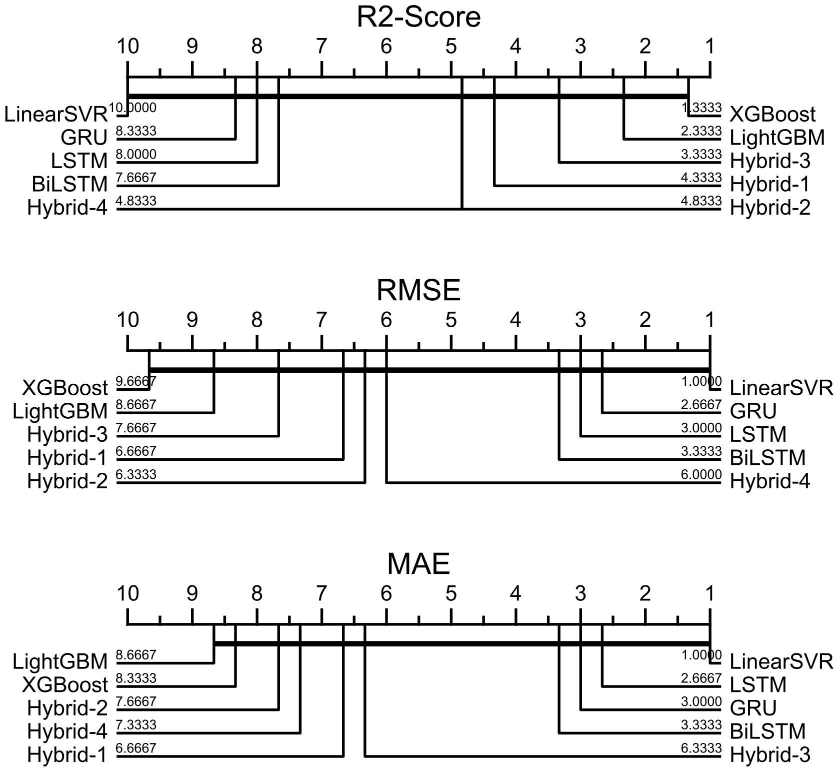


**(a)** R2

**(b)** RMSE

**(c)** MAE

**Figure 3.** Box Plots Depicting $R^2$, RMSE, and MAE Metrics for Each TTP Method



**(a)** R2

**(b)** RMSE

**(c)** MAE

**Figure 4.** CD Diagrams Comparing the Performance of the TTP Methods across the Three Data Sets

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).

### *4.3. RQ2: Comparison of TTP Explanation Methods*

This section evaluates the ability of different XAI to provide 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.

### *4.4. Global Explanations*

The global explanations provided by SHAP can be visualized using different plots. Fig. 5 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 and different

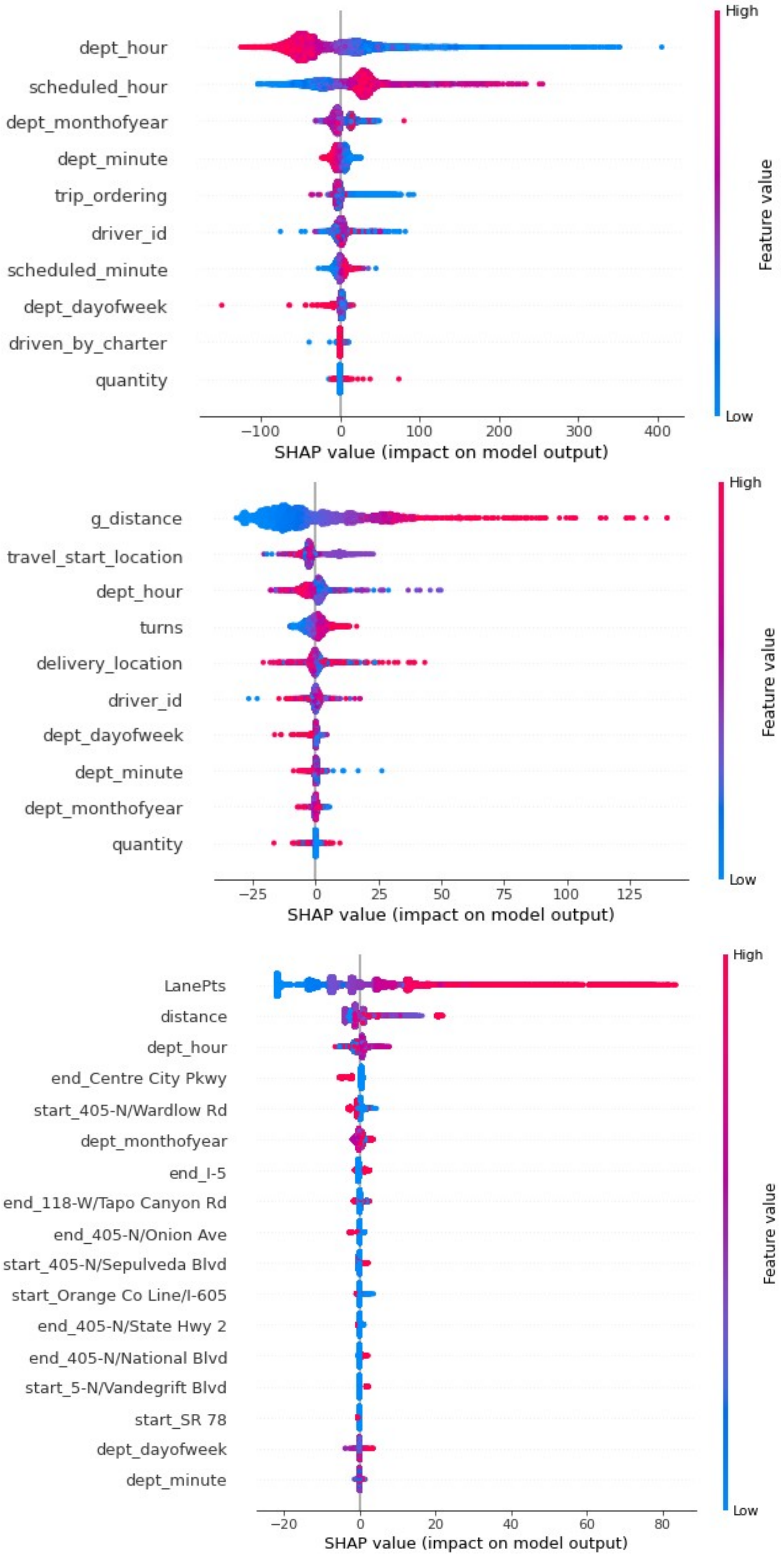


**Figure 5.** Global Explanations with SHAP for the XGBoost Model and Data Sets NextUp-1, NextUp2, and PeMS

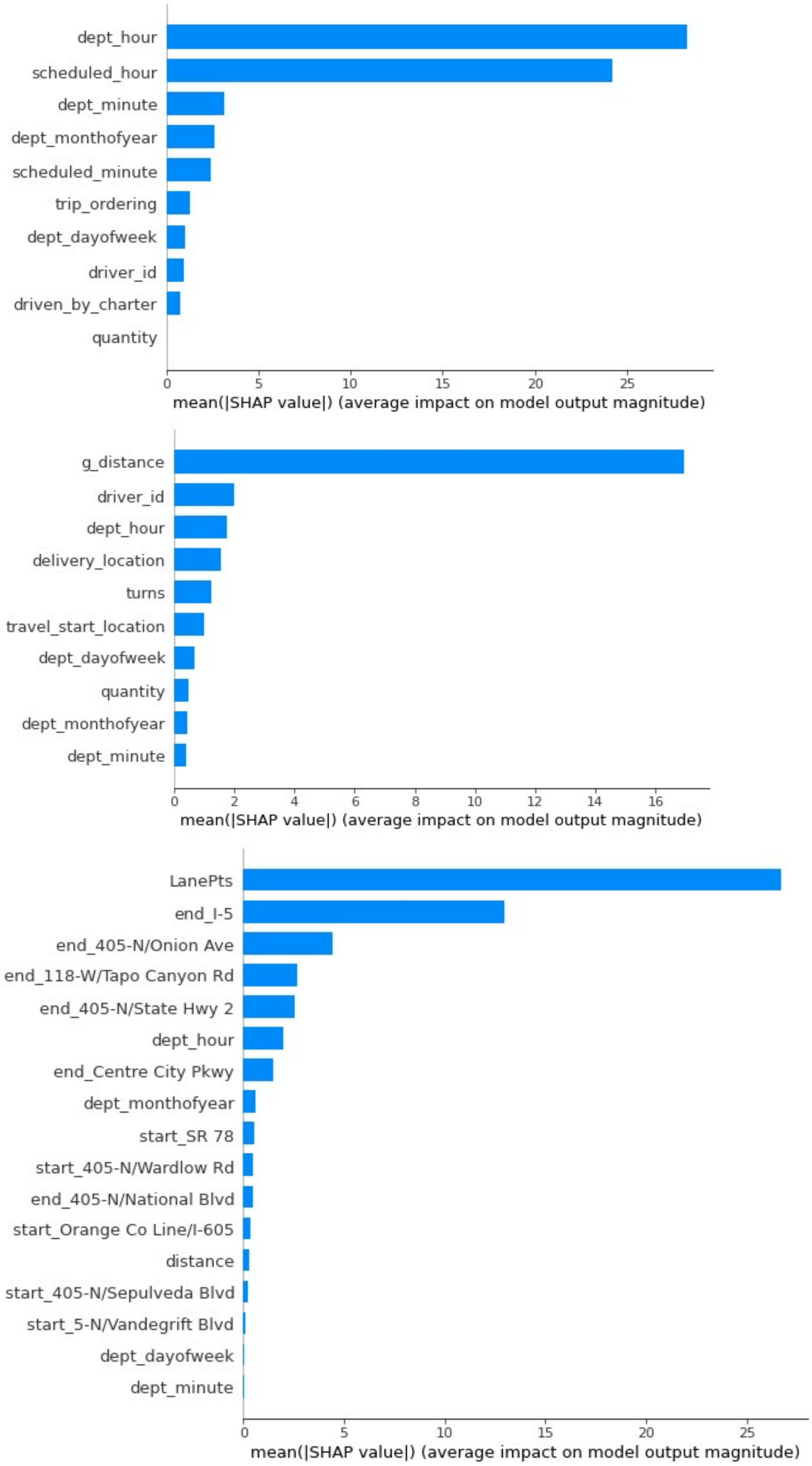


**Figure 6.** Overall Impact of Features on Model Outcomes, Summary Plots for the LSTM model and the Data Sets NextUp-1, NextUp2, and PeMS

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

It is also possible to visualize the importance of each feature by summing the absolute Shapley values of each feature across all samples. Figure 6 shows the importance of the temporal, spatial, and general properties of travels by summing the absolute Shapley values of each feature for all samples. We used the LSTM model. According to this figure, most temporal properties ( *e.g.*, *dept_hour* and scheduled_hour) and spatial properties ( *e.g.*, *distance* and *LanePts*) are the most important features. However, at the other extreme, some temporal features such as *dept_minute* and *dept_dayofweek* have the least importance and influence on the TT prediction. When comparing Fig. 6 and Fig. 5, we can observe the differences in the importance given to the features by different learning algorithms. For example, XGBoost considered *dept_hour* is more important than *driver_id*, which is in contrast to LSTM's behavior. We can made similar observations regarding the features *end_i-d*, *dept_hour*, and *distance*.

### *4.5. Local Explanations*

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

Figure 7 shows three instances of using the waterfall plot for the TT predictions made by the XGBoost model. The positive effect of the feature pushes the prediction higher from the base value, as shown in red, and the negative effect pushes the prediction lower, as shown in blue. In instance 1, we can see that the feature *scheduled_hour* pushes the prediction 35.73 minutes lower, while the feature *dept_hour* pushes the prediction 34.71 minutes higher. For instance 3, the feature *dept_hour* has a negative effect of 52.56 minutes. The force plot in Fig 8 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 method 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 local model around the unit of the predicted value to provide an explanation, while SHAP decomposes the prediction into contribution by each feature and the SHAP values for each feature adds up to the final prediction which is not the case with LIME.

Figure 9 depicts the explanations for three travel time prediction instances of the XGBoost model. The blue color shows a negative contribution to travel time, *i.e.*, increasing time, while the orange color shows a positive contribution. For a given prediction, the plot shows the essential features, their relative importance/contributions, and their actual values. 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.

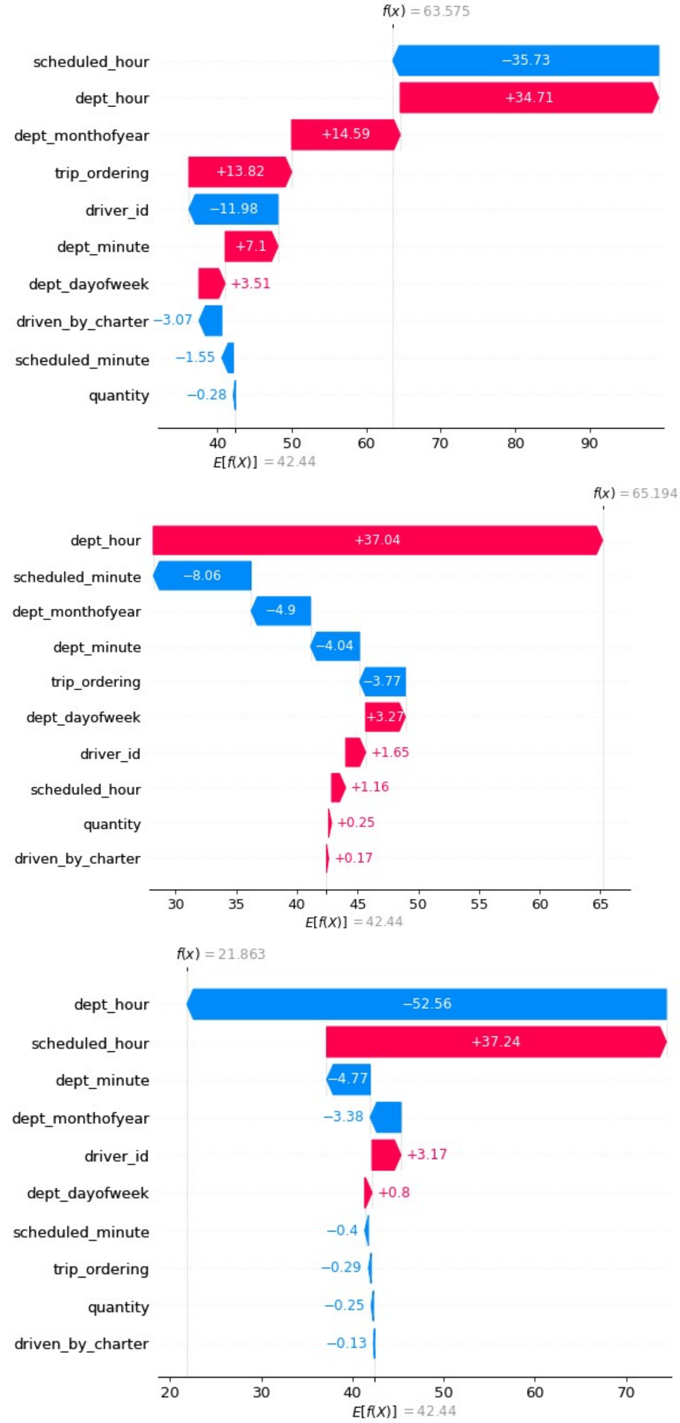


**Figure 7.** SHAP Waterfall Plots for Explaining Three Prediction Instances

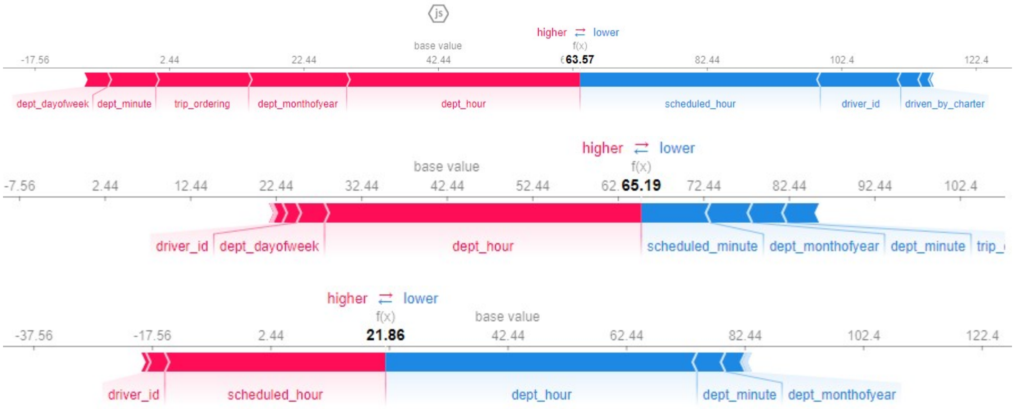


**Figure 8.** SHAP Force Plots for Explaining Three Prediction Instances

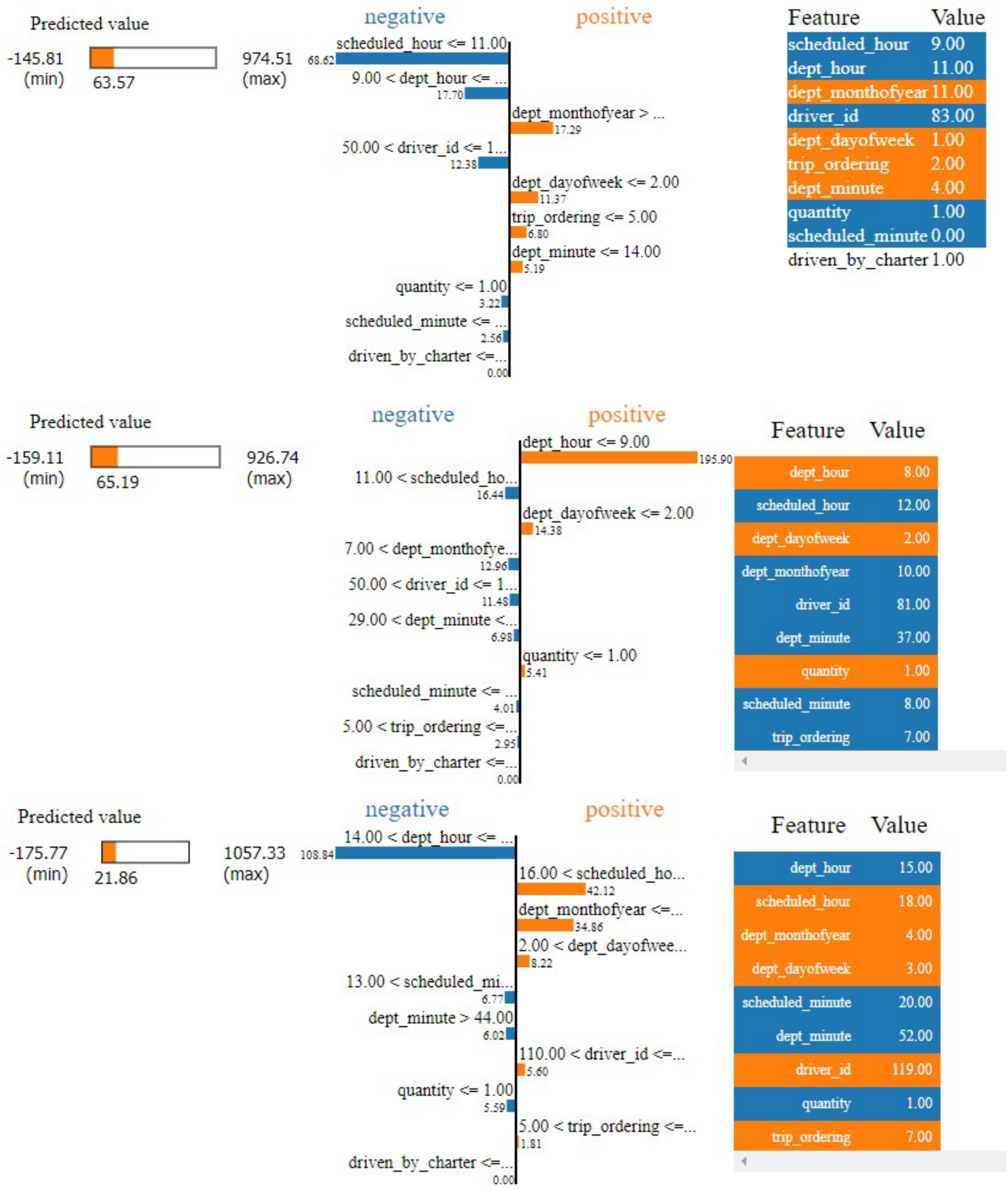


**Figure 9.** LIME feature Contribution Plots for Explaining Three Prediction Instances

## 5. Discussion

Our findings demonstrate that the temporal and spatial features and their correlations can significantly affect travel time. The data-driven models combined with the corresponding explainable methods can help understand what factors affect travel time to what extent and how to predict travel time effectively. 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 baseline model, *i.e.*, linear SVMR, showed the 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. [32]. 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 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.

The 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 the field of travel time prediction as it can help the user understand why a particular prediction is made, especially in the field of logistics where the planners need to monitor the orders, and they want to know why a particular 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. 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

This paper has two main contributions in evaluating TTP models and explaining TTP that can help to predict and explain travel time in the logistics domain. First, we implemented various TTP methods, *i.e.*, ensemble learning, neural networks, support vector machine regression, and hybrid models that combine ensemble learning methods and neural networks. Second, we compared these predictive models with various spatiotemporal features over three different data sets. The experiment results show that the ensemble method such as LightGBM and XGBoost outperform the neural network models. Finally, we assessed the ability of SHAP and LIME XAI methods to explain the travel time predictions made by black-box models. XAI methods help to understand whether the particular predictions are based on intuitively reasonable relationships embedded in the model. There is almost no study available of explainable AI methods for interpreting the TTP output of black-box models such as neural networks and ensemble learning methods.

For future work, we plan to consider the data quality issues in travel time prediction and the importance of other features that affect travel time, such as customer type, road type, area type and driver profile data. Additionally, through a user study, we will investigate the usefulness of XAI in interpreting TTP prediction models.

**Author Contributions:** Conceptualization, experiment, implementation and evaluation I.A; conceptualization; 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.

**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. Aranko, J. Developing the last mile of a parcel delivery service concept for consumers **2013**.
3. 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.
4. 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.
5. 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.
6. 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.
7. Reshadat, V.; Faili, H. A new open information extraction system using sentence difficulty estimation. *Computing and Informatics* **2019**, *38*, 986–1008.
8. 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.
9. Reshadat, V.; Hourali, M.; Faili, H. Confidence Measure Estimation for Open Information Extraction. *Information Systems & Telecommunication* **2018**, p. 1.
10. Wentworth, J. Expert systems in transportation. Technical report, AAAI Technical Report WS-93-04, 1993.
11. 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.
12. Vyborny, C.J.; Giger, M.L. Computer vision and artificial intelligence in mammography. *AJR. American journal of roentgenology* **1994**, *162*, 699–708.
13. 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.
14. 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.
15. 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.
16. 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.
17. 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.
18. 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.
19. Molnar, C. *Interpretable Machine Learning*; Lulu. com, 2020.
20. Zhang, Y.; Haghani, A. A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies* **2015**, *58*, 308–324.
21. 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.
22. 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.
23. Chen, Z.; Fan, W. A Freeway Travel Time Prediction Method Based on an XGBoost Model. *Sustainability* **2021**, *13*. doi:10.3390/su13158577.
24. 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.
25. Ran, X.; Shan, Z.; Fang, Y.; Lin, C. An LSTM-based method with attention mechanism for travel time prediction. *Sensors* **2019**, *19*, 861.

26. 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.
27. 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.
28. 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.
29. 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.
30. 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.
31. 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.
32. 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.
33. 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/MITS.2019.2919615.
34. 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.
35. 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.
36. 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.
37. 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**.
38. 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.
39. 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.
40. 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.
41. 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.
42. 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.
43. 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.
44. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An introduction to statistical learning*; Vol. 112, Springer, 2013.
45. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research* **2006**, *7*, 1–30.
46. 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.
47. 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.
48. 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.
49. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. Advances in neural information processing systems, 2017, pp. 4765–4774.
50. 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.
51. Xie, Y.; Pongsakornsathien, N.; Gardi, A.; Sabatini, R. Explanation of Machine-Learning Solutions in Air-Traffic Management. *Aerospace* **2021**, *8*. doi:10.3390/aerospace8080224.
52. 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.