Transport and Communications Science Journal, Vol. 77, Issue 4 (05/2026), 571-582

Transport and Communications Science Journal

URBAN TRAFFIC FLOW PREDICTION AND TRAFFIC STATE IDENTIFICATION USING CATBOOST WITH SHAP BASED ANALYSIS

Pham Thi Ly, Nguyen Duc Du, Nguyen Thi Hong Hoa*

University of Transport and Communications, No 3 Cau Giay Street, Hanoi, Vietnam

ARTICLE INFO

TYPE: Research Article
Received: 15/03/2026
Revised: 29/04/2026
Accepted: 30/04/2026
Published online: 15/05/2026
https://doi.org/10.47869/tcsj.77.4.17

* *Corresponding author*
Email: hoanth@utc.edu.vn; Tel: 0982108957

Abstract. Traffic flow forecasting is a critical component in intelligent transportation systems, supporting traffic management, reducing congestion, and improving the operational efficiency of urban road networks. However, this is a challenging problem due to the temporal variability of traffic data and the influence of numerous complex factors. In this study, we propose a traffic flow forecasting method based on the CatBoost algorithm to effectively exploit tabular traffic data collected from traffic sensors. The dataset consists of 2,976 records containing temporal information and vehicle counts across four categories (cars, motorcycles, buses, and trucks). In addition to the original features, the study constructs supplementary temporal and time-series features, including Hour, DayOfWeek, IsWeekend, Total_lag1, and Total_roll3, to enhance the model's ability to capture traffic flow variation trends. Based on this, two independent machine learning tasks are established: (i) total traffic flow forecasting as a regression problem, and (ii) traffic condition classification into four levels. Experimental results demonstrate that the proposed model achieves strong predictive performance. Furthermore, feature importance analysis using the SHAP method reveals that vehicle count-related variables, particularly CarCount and BusCount, have a significant impact on prediction outcomes. The study demonstrates that CatBoost is an effective approach for traffic flow forecasting with tabular data and holds strong potential for application in intelligent traffic management systems.

Keywords: traffic prediction, catboost, machine learning, time series analysis, intelligent transportation systems, SHAP.

1. INTRODUCTION

Traffic flow prediction is a crucial component of intelligent transportation systems (ITS), as it supports traffic management, reduces congestion, and improves the efficiency of infrastructure utilization. Due to its nonlinear and non-stationary characteristics, as well as its strong dependence on spatio-temporal factors, this problem has attracted significant attention from the research community over the past several decades [1], [2]. Recent survey studies have highlighted a clear evolution from traditional statistical models to machine learning, deep learning, and graph-based spatio-temporal models [1], [2]. However, the selection of an appropriate model remains highly dependent on the characteristics of the input data and the requirements of real-world deployment.

Early studies on traffic flow prediction mainly relied on parametric time-series models such as ARIMA and seasonal variations [3], [4]. These models offer advantages including simple structures, ease of interpretation, and low computational cost, making them suitable for short-term forecasting under relatively stable traffic conditions. However, traditional statistical models assume linearity and stationarity in the data, which limits their applicability to real-world traffic datasets that are often highly dynamic, noisy, and influenced by various exogenous factors [1], [3]. These limitations have motivated a shift toward more flexible machine learning approaches.

Machine learning, especially ensemble methods based on decision trees, has shown good performance in modeling complex nonlinear relationships. The theoretical foundation of boosting was established early on [5] and later developed into powerful algorithms such as XGBoost [6] and LightGBM [7].

In the transportation field, ensemble machine learning models have been applied to short-term traffic flow prediction and have shown competitive results compared with traditional models [8]. These methods work well with tabular data, which makes them suitable for traffic datasets stored in formats such as CSV that contain temporal, statistical, and contextual features. CatBoost was proposed to address some limitations of traditional boosting methods, especially the bias that occurs when handling categorical variables. It uses ordered boosting and built-in categorical feature encoding to solve this problem [9]. These features help CatBoost achieve stable performance and reduce overfitting, particularly when working with real-world datasets that have heterogeneous structures.

Some recent studies have applied CatBoost to traffic flow prediction and have reported promising results compared with other machine learning models [10]. In addition, combining CatBoost with explainable machine learning techniques helps clarify the role of temporal and contextual features, thereby improving the reliability of the model in intelligent transportation applications [11]. At the same time, deep learning models such as LSTM, BiLSTM, and CNN–LSTM have been widely studied to capture long-term dependencies in traffic time series [12], [13], [14]. These models have shown strong performance in many scenarios, especially when the data have clear sequential structures.

Recently, graph-based and attention-based spatio-temporal deep learning models, such as GCN, Transformer, and Graph ODE, have achieved impressive results in traffic prediction on large-scale networks [15], [16], [17], [18]–[22]. However, these methods often require complex data structures, high computational cost, and are difficult to deploy in real-world environments, especially when the input data are only available in tabular formats such as CSV.

Although deep learning models can achieve high accuracy, many studies indicate that not all traffic prediction problems require complex architectures [1], [2]. In the case of traffic data stored in CSV format, where appropriate temporal and statistical features can be designed, ensemble machine learning models such as CatBoost can achieve competitive performance while maintaining computational efficiency and interpretability. Therefore, this study focuses on applying CatBoost to traffic flow prediction using CSV-based data, aiming to evaluate the potential of this approach in practical scenarios and to provide additional empirical evidence for lightweight yet effective machine learning methods.

2. MATHEMATICAL MODEL OF THE CATBOOST ALGORITHM FOR TRAFFIC FLOW PREDICTION

2.1. Dataset description and predictions tasks

In this study, two learning models are trained in parallel, including a regression model and a classification model. The regression model predicts the total number of vehicles at the current time based on historical observations and the input feature vector at time t. The classification model, on the other hand, identifies the traffic situation according to vehicle density levels. Both models are trained using the same dataset but with different target variables corresponding to the regression and classification tasks, enabling simultaneous traffic flow prediction and traffic state identification.

The dataset used in this study was collected from traffic sensors and consists of 2,976 records stored in CSV format, sampled at regular time intervals. The dataset provides information for each time slot of the day, including temporal attributes (date and hour) and the number of different vehicle types, namely CarCount, BikeCount, BusCount, and TruckCount. In addition, the dataset includes the total traffic flow (Total) and a traffic situation variable (Traffic Situation). The Total variable is used as the target for the regression task, while the traffic situation variable is used as the target for the classification task, with four labels: 1–Low, 2–Normal, 3–High, and 4–Very High. For objective evaluation, the dataset is divided into a training set (80%) and a testing set (20%).

The input features for traffic prediction are constructed from the original data and grouped into three categories: original, temporal, and historical features. The original features include CarCount, BikeCount, BusCount, and TruckCount, which represent the number of each vehicle type. Temporal features are derived from the Time and Date attributes, including Hour (0–23), DayOfWeek (0–6), and IsWeekend to capture weekly traffic patterns. Historical features are constructed from past values of the Total variable, including Total_lag1 (previous time step) and Total_roll3 (three-step moving average).

2.2. Traffic flow forecasting process

In this paper, the CatBoost algorithm is employed to develop a traffic flow prediction model. CatBoost is a machine learning algorithm based on gradient boosting over decision trees, designed to efficiently handle tabular datasets. The proposed traffic flow prediction framework consists of four main stages (as illustrated in Figure 1): (i) preprocessing the traffic data stored in CSV format, (ii) constructing features from the original data, (iii) training the CatBoost model, and (iv) predicting traffic flow on the dataset.

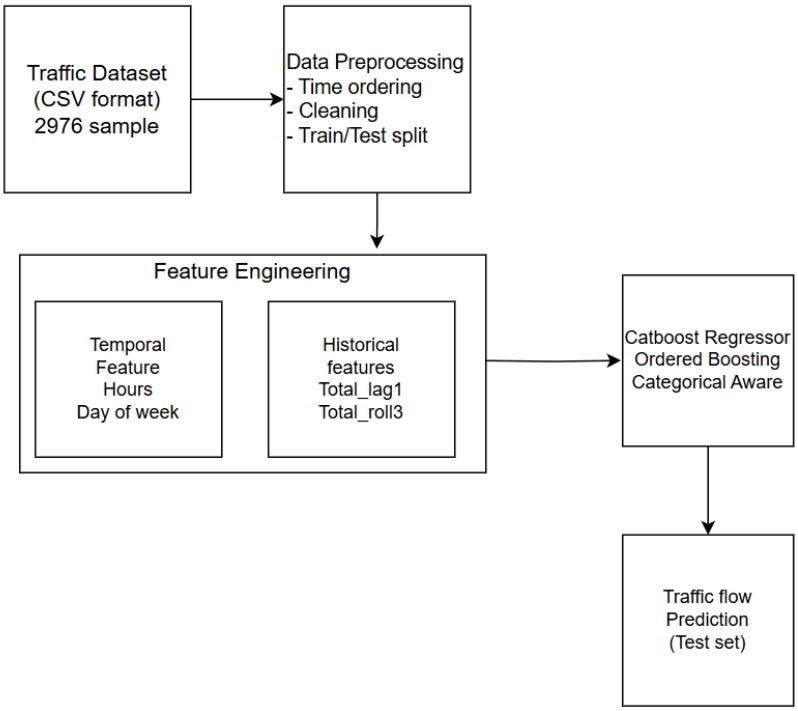

Figure 1. Overall architecture of the proposed CatBoost-based traffic flow prediction method.

Data Preprocessing Stage: This is the first stage of the traffic flow prediction process. The procedure begins with the raw traffic dataset stored in CSV format, which contains 2,976 records with attributes related to time and vehicle counts. To maintain the integrity of the time-series data, the records are arranged in chronological order. Data cleaning steps are then performed to remove missing or invalid values. After preprocessing, the dataset is divided into training and testing sets using a time-based splitting strategy in order to avoid information leakage during model training.

Feature Construction from the Original Data: This stage is an important part of the proposed method. Besides the original traffic features such as the numbers of cars, bikes, buses, and trucks, this study also creates temporal and historical features to capture periodic patterns and short-term dependencies in traffic flow. Temporal features such as Hour, DayOfWeek, and IsWeekend are extracted from the time information to represent daily and weekly traffic variations. Historical features Total_lag1 and Total_roll3 are generated from the total traffic flow observed in previous time steps. These features help the CatBoost model better capture the characteristics of traffic flow and improve prediction performance.

CatBoost Model Training: The constructed feature set is used to train a CatBoost regression model. The CatBoost is selected due to its ability to efficiently handle tabular data with heterogeneous features, reduce overfitting through the ordered boosting mechanism, and provide stable performance on small and medium-sized datasets. These characteristics make CatBoost particularly suitable for short-term traffic flow prediction based on time-series data.

Traffic Flow Prediction on the Test Dataset: After training, the CatBoost model is applied to the test dataset to generate traffic flow predictions. The predicted values are evaluated using standard regression metrics to measure the accuracy and effectiveness of the proposed approach.

2.3. Mathematical Model for CatBoost-Based Traffic Flow Prediction

Consider a traffic dataset collected from sensors at regular time intervals, consisting of T=2976 observations. The dataset can be represented as follows:

$$D = \{(X_t, Y_t, C_t)\}_{t=1}^{T} \quad (1)$$

Where t is the time index$t = 1,2, \dots, T$); T=2976;$X_t \in \mathbb{R}^d$$Y_t \in \mathbb{R}$ represents the total traffic flow to be predicted; D is the entire dataset; (x_t,y_t,c_t) is a data sample at time t; .The traffic condition variable$C_t \in \{1,2,3,4\}$

The total traffic flow is defined as:

$$Y_t = CarCount_t + BikeCount_t + BusCount_t + TruckCount_t \quad (2)$$

The feature vector X_t at time t consists of three groups of features:

$$X_t = \left[X_t^{(raw)}, X_t^{(temp)}, X_t^{(hist)}\right]$$

where $X_t^{(raw)}$ denotes the original features, $X_t^{(Temp)}$ represents temporal features, and $X_t^{(hist)}$ denotes historical features.

$$X_t^{(raw)} = [CarCount_t, BikeCount_t, BusCount_t, TruckCount_t] \quad (3)$$

$$X_t^{(temp)} = [Hour_t, DayOfWeek_t, IsWeekend_t] \quad (4)$$

$$X_t^{(hist)} = \left[Y_{t-1}, \frac{1}{3}(Y_{t-1} + Y_{t-2} + Y_{t-3})\right] \quad (5)$$

The total traffic flow is the target variable of the regression model. The objective is to learn a mapping:

$$f_r: \mathbb{R}^d \to \mathbb{R} \quad (6)$$

such that the predicted traffic volume at time t is given by $\hat{Y}_t = f_r(X_t)$ and the squared loss function is minimized:

$$\mathcal{L}_{reg} = \sum_{t \in \mathcal{T}_{\text{train}}} (Y_t - \hat{Y}_t)^2 \quad (7)$$

$\mathcal{T}_{\text{train}}$ denotes the dataset used for training the model, accounting for 80% of the data, while the remaining 20% is used as the testing set$\mathcal{T}_{\text{test}}$. This data partition ensures that the model only utilizes past information to predict future values, which is consistent with the time-series nature of traffic data**.**

For the objective of the problem, which is to classify traffic conditions, the following mapping needs to be learned:

$$f_c: \mathbb{R}^d \to \mathbb{R}^4 \quad (8)$$

where the probability vector is defined as: $\hat{p}_t = f_c(X_t)$ with

$$\hat{p}_t = [P(c_t = 1|x_t), P(c_t = 2|x_t), P(c_t = 3|x_t), P(c_t = 4|x_t)] \quad (9)$$

The predicted label is determined as:

$$\hat{c}_t = argmax_{j \in \{1,2,3,4\}} \hat{p}_{t,j} \quad (10)$$

The Categorical Cross-Entropy (CCE) loss function is employed to optimize the multi-class traffic condition classification model:

$$\mathcal{L}_{CCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} Y_{i,k}\log\left(p_{i,k}\right) \quad (11)$$

where N denotes the number of training samples; $K = 4$ represents the number of traffic condition classes (Low, Normal, High, Heavy); $Y_{i,k} \in \{0,1\}$ is a one-hot indicator variable that equals *1* if sample i belongs to class k and 0 otherwise; and $p_{i,k}$ denotes the predicted probability that sample i belongs to class k. The probability $p_{i,k}$ is computed using the Softmax function applied to the output of the CatBoost model, satisfying $0 \leq p_{i,k} \leq 1$ and $\sum_{k=1}^{K} p_{i,k} = 1$.

The mathematical formulation of the CatBoost prediction function $f(x)$ can be expressed as a weighted sum of K decision trees:

$$f(x) = \sum_{k=1}^{K} \alpha_k h_k(x) \quad (12)$$

where
$h_k(x)$ denotes the k-th decision tree,
α_k represents the weight associated with the k-th tree, and
K is the total number of trees in the model.

The decision trees are trained sequentially according to the Gradient Boosting principle, where at iteration k, the tree h_k is optimized to approximate the negative gradient of the loss function.

The training objective is to learn an optimal function $f^*(x)$ such that:

$$f_r^* = \arg\min_{f_r} \mathcal{L}_{reg} \text{ or } f_c^* = \arg\min_{f_c} \mathcal{L}_{CCE}$$

The prediction performance is evaluated using the error metrics MAE, RMSE, and R^2.

3. EXPERIMENTS AND EVALUATION

Traffic prediction experiments were conducted using the CatBoost model on a dataset consisting of 2,976 records stored in CSV format and nine input features: Time, Date, Day of Week, Carcount, Bikecount, Buscount, Truckcount, Total, and Situation. The evaluation results of the model are presented as follows:

3.1. Model Performance Evaluation

In this study, the CatBoostRegressor model was configured with *iterations = 500*, *depth = 6*, and *learning_rate = 0.1*. The evaluation results on the test set (20% of the dataset) are summarized as follows: (i) the regression model (CatBoost Regressor) predicts the total vehicle traffic volume within each time interval; and (ii) the classification model (CatBoost Classifier) identifies the level of traffic congestion ("low", "normal", "high", and "heavy"). The experimental results demonstrate that both models achieve high and stable performance, reflecting their strong capability to capture the underlying data structure and the relationships among the input features, as shown in Table 1.

Table 1. Performance evaluation results of the CatBoost classification and regression models.

CatBoost Regression Model		**CatBoost Classifier Model**	
MAE	0.6158	Accuracy	0.9681
RMSE	0.8673	Precision	0.9680
R^2	0.9936	Recall	0.9681
		F1-score	0.9678

As shown in Table 1, the experimental results on the test set (20% of the dataset) demonstrate outstanding predictive performance, with a Mean Absolute Error (MAE) of 0.6158 and a Root Mean Square Error (RMSE) of 0.8673. Notably, the coefficient of determination reaches $R^2 = 0.9936$, indicating that the model is able to explain 99.36% of the variance in the observed traffic data. These metrics not only confirm the high accuracy of the algorithm in capturing nonlinear relationships between temporal factors and vehicle traffic volume, but also highlight the stability of the model, as extreme prediction errors are effectively minimized. Such performance satisfies the practical requirements for traffic monitoring and analytical applications.

For the traffic condition classification task, the CatBoost model achieves *Accuracy = 0.9681, Precision = 0.9680, Recall = 0.9681, and F1-score = 0.9678.* In the multi-class classification of traffic congestion levels, the CatBoostClassifier demonstrates superior performance across all four categories (Low, Normal, High, and Heavy). The experimental results reveal strong consistency among the evaluation metrics. In particular, the F1-score of 96.78% indicates an optimal balance between precision and recall, confirming the model's capability to accurately identify different traffic states without introducing bias toward dominant classes. The stability of these statistical indicators further demonstrates the algorithm's effectiveness in separating the feature space, thereby providing a reliable foundation for real-time traffic congestion warning systems with minimal prediction errors.

Overall, the results from both regression and classification experiments indicate that CatBoost is highly suitable for traffic flow data analysis, addressing both prediction and classification tasks effectively. The model achieves low prediction errors, high explanatory power, and strong stability. The consistency among performance metrics confirms that the model does not suffer from overfitting and exhibits strong generalization capability across temporally diverse traffic data. Owing to its high accuracy and ability to capture complex nonlinear relationships, CatBoost shows significant potential for practical deployment in real-time traffic flow prediction systems.

To further evaluate the robustness of the proposed model, a time-series cross-validation strategy is employed. Unlike standard k-fold cross-validation, this approach preserves the temporal order of the data and avoids information leakage. The average results across multiple folds are reported in Table 2 and Table 3:

Table 2. Evaluation using Time-Series Cross-Validation for total traffic flow prediction with CatBoost.

FOLD	**MAE**	**RMSE**	**R2**
Fold 1	2.62	3.82	0.9955
Fold 2	1.58	2.27	0.9986
Fold 3	1.38	2.21	0.9987
Fold 4	1.72	3.64	0.9966
Fold 5	1.30	2.02	0.9988
Avg of 5 fold	1.72	2.79	0.9976

Table 3. Evaluation using Time-Series Cross-Validation for traffic condition prediction with CatBoost.

Traffic condition	Pricision	Recall	F1-Score	Support
Heavy	0.97	0.98	0.97	137
High	0.97	0.88	0.92	64
Low	0.95	0.95	0.95	60
Normal	0.97	0.99	0.98	334

While the time-based split reflects real-world forecasting scenarios, cross-validation is further conducted to assess the stability of the model performance. The results show that the proposed CatBoost model maintains consistent performance across different data partitions.

3.2. Evaluation Using the Confusion Matrix

To comprehensively evaluate the performance of the traffic flow classification model, this study employs the confusion matrix as a detailed analytical tool in addition to aggregate metrics such as Accuracy, Precision, Recall, and F1-score. The confusion matrix provides a visual representation of the model's correct and incorrect predictions for each class, thereby offering important insights into the model's behavior across different traffic states ("low", "normal", "high", and "heavy"). The obtained results are illustrated in Figure 2

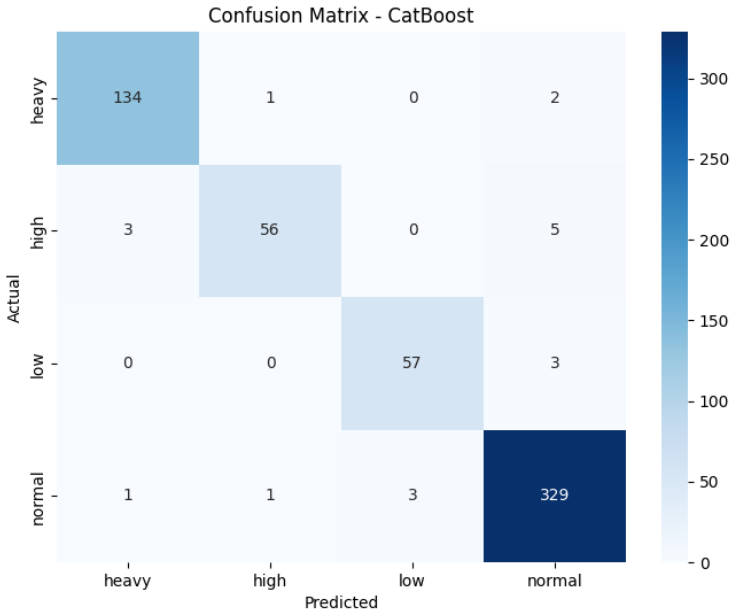

Figure 2. Confusion Matrix of the CatBoost Model.

As illustrated in Figure 2, the confusion matrix of the CatBoost model demonstrates high and stable classification performance across all four traffic state classes. Specifically, the *normal* class achieves the best result, with 329 out of 334 samples correctly predicted, while only five samples are misclassified into other classes (1 heavy, 1 high, and 3 low), indicating that the model is highly effective in identifying normal traffic conditions. For the *heavy* class, the model correctly classifies 134 out of 137 samples, with only three minor misclassifications into the *high* and *normal* classes. The *high* class achieves 56 correct predictions out of 64 samples, where eight samples are mainly misclassified into *normal* (five samples) and *heavy*

(three samples), reflecting the overlap between higher traffic intensity levels. Similarly, the *low* class shows favorable performance, with 57 out of 60 samples correctly identified, and only three samples misclassified as *normal*. Overall, most classification errors occur between adjacent traffic levels, which is consistent with the continuous nature of traffic flow data and further confirms the reliability of the model in practical applications.

3.3. Evaluation Based on the Correlation Between Actual and Predicted Values

Figure 3 illustrates the correlation between the actual traffic volume (Actual Total) and the predicted values (Predicted Total) generated by the CatBoost Regressor on the test dataset. The ideal reference line is given by $y = x$.

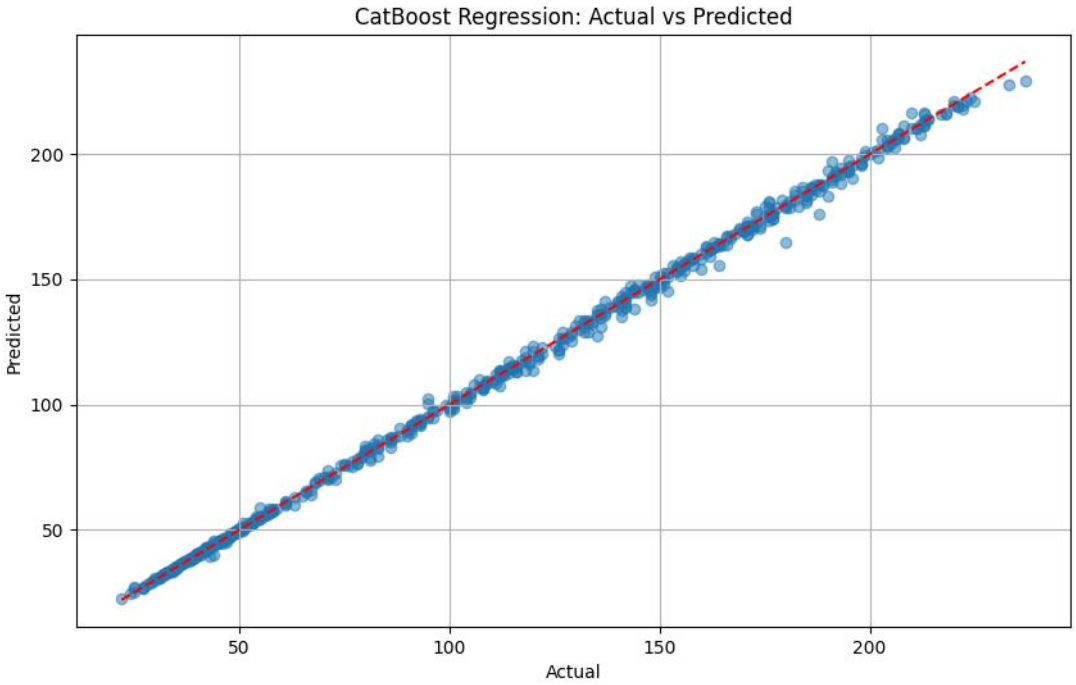

Figure 3. Relationship between actual and predicted values of the CatBoost model.

The predicted data points generated by the CatBoost model are densely distributed and closely aligned with the reference line $y = x$, indicating that the model effectively captures the relationship between the input features and the target variable. Most prediction errors fall within a range of approximately ±5 to 10 units across the entire value domain. In the high-traffic region (200–260), the data points remain closely aligned with the ideal reference line, while in the low-traffic region (<50), the prediction errors are minimal. No significant outliers are observed, suggesting that the model does not suffer from distributional bias and maintains high stability in practical applications.

To further analyze the contribution of each feature to the prediction results, this study employs the SHAP (SHapley Additive exPlanations) method to quantify the influence of input variables on the output of the CatBoost model. The SHAP summary plot illustrates the relative importance of features, ranked according to their mean absolute contribution values.

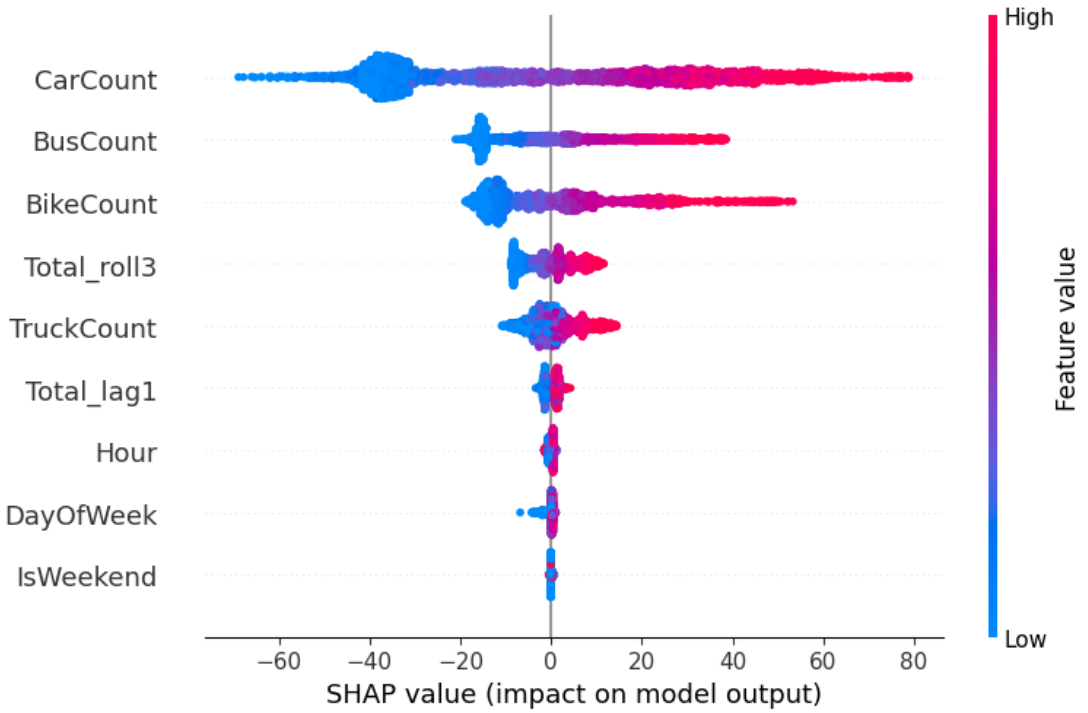

Figure 4. SHAP Plot.

As shown in Figure 4, CarCount is the feature with the greatest influence on the predicted values. Observations with higher CarCount values correspond to large positive SHAP values, indicating that an increase in the number of passenger cars significantly raises the predicted total traffic volume. This finding is consistent with real-world urban traffic conditions, where private vehicles typically constitute a substantial proportion of the overall traffic flow.

Following CarCount, BusCount and BikeCount also exhibit notable influence and tend to contribute positively as their values increase. This indicates that the model successfully captures the direct relationship between individual vehicle components and the total traffic volume, thereby ensuring the physical plausibility of the prediction results. Notably, the historical feature Total_roll3 shows a greater impact than Total_lag1, suggesting that the use of a rolling average over the three most recent time steps enables the model to capture short-term trends more effectively than relying solely on a one-step lag value. This result highlights the important role of historical feature engineering in improving time-series prediction performance.

In contrast, cyclical temporal features such as Hour, DayOfWeek, and IsWeekend exhibit relatively lower influence, with their SHAP values concentrated around zero. This suggests that temporal factors play a supportive rather than dominant role in the current prediction model. Nevertheless, the presence of these features still contributes to stabilizing the predictions by providing information about periodic traffic patterns. Overall, the SHAP analysis demonstrates that the CatBoost model not only achieves high predictive accuracy but also maintains strong interpretability. The ranking of feature importance aligns well with domain intuition and the real characteristics of traffic data, thereby reinforcing the reliability of the proposed approach.

4. CONCLUSION

This study proposes a CatBoost-based approach for traffic flow prediction and traffic state classification. Experimental results show that the CatBoost Regressor achieves high prediction accuracy with MAE = 0.6158, RMSE = 0.8673, and $R^2 = 0.9936$, indicating strong capability in modeling the relationship between temporal factors and traffic volume. For the classification task, the CatBoost Classifier obtains high performance with Accuracy = 0.9681 and F1-score = 0.9678, demonstrating reliable identification of four traffic states (low, normal, high, and heavy). Furthermore, SHAP analysis confirms that vehicle-related features such as CarCount, BusCount, and BikeCount play the most significant role in the prediction results. Overall, the findings indicate that CatBoost provides a highly accurate, stable, and interpretable solution for traffic flow prediction and congestion classification, showing strong potential for real-time intelligent transportation applications.

REFERENCES

[1]. Y. Zheng, R. Jiang, X. Song, D. Yin, Z. Wang, R. Shibasaki, A Comprehensive Survey on Traffic Prediction, IEEE Transactions on Knowledge and Data Engineering, 2025 (Early Access). https://doi: 10.1109/TKDE.2025.3461234

[2]. R. Jiang, D. Yin, Z. Wang, Y. Wang, J. Deng, H. Liu, Z. Cai, J. Deng, X. Song, R. Shibasaki, "DL-Traff: Survey and Benchmark of Deep Learning Models for Urban Traffic Prediction, in Proc. 30th ACM Int. Conf. on Information and Knowledge Management (CIKM), (2021) 4515–4525. https://doi: 10.1145/3459637.3482000

[3]. B. L. Smith, B. M. Williams, R. K. Oswald, Comparison of Parametric and Nonparametric Models for Traffic Flow Forecasting, Transportation Research Part C: Emerging Technologies, 10 (2002) 303–321. https://doi: 10.1016/S0968-090X(02)00023-8

[4]. B. M. Williams, L. A. Hoel, Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results, Journal of Transportation Engineering, 129 (2003) 664–672. https://doi: 10.1061/(ASCE)0733-947X(2003)129:6(664)

[5]. Y. Freund, R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences, 55 (1997) 119–139. https:// doi: 10.1006/jcss.1997.1504

[6]. T. Chen and C. Guestrin, XGBoost: A Scalable Tree Boosting System, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) (2016) 785–794. https://doi: 10.1145/2939672.2939785

[7]. G. Ke et al., LightGBM: A Highly Efficient Gradient Boosting Decision Tree, in Advances in Neural Information Processing Systems (NeurIPS), 30 (2017) 3146–3154.

[8]. H. Lv, W. Duan, Y. Wang, C. Hua, J. Li, Short-Term Traffic Flow Prediction Based on Ensemble Machine Learning Strategies, IEEE Access, 7 (2019) 160140–160150. https://doi: 10.1109/ACCESS.2019.2951815

[9]. L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, A. Gulin, CatBoost: unbiased boosting with categorical features, in Advances in Neural Information Processing Systems (NeurIPS), 31(2018) 6638–6648.

[10]. A. A. Alghamdi, A Study on the Traffic Flow Prediction through CatBoost Algorithm, International Journal of Advanced Computer Science and Applications, 13 (2022)123-132.

[11]. X. Zhang, Traffic flow prediction based on explainable machine learning, Highlights in Science, Engineering and Technology, 56 (2023) 56–65. https://doi: 10.54097/hset.v56i.10620

[12]. X. Ma, Z. Tao, Y. Wang, H. Yu, Y. Wang, Long short-term memory neural network for traffic speed prediction using remote microwave sensor data, Transportation Research Part C: Emerging Technologies, 54 (2015) 187-197.

[13]. Y. Lv, Y. Duan, W. Kang, Z. Li, F.-Y. Wang, Traffic Flow Prediction with Big Data: A Deep

Learning Approach, IEEE Transactions on Intelligent Transportation Systems, 16 (2015) 865–873. https://doi: 10.1109/TITS.2014.2346413
[14]. J. Jin, F. Chen, Y. Zhang, Traffic Flow Forecasting Based on Hybrid Deep Learning Framework, IEEE Access, 7 (2019) 82502–82513. https://doi: 10.1109/ACCESS.2019.2922667
[15]. L. Bai, L. Yao, C. K. M. Lee, S. R. K. A. Yeung, X. Zhang, Y. Wang, Adaptive Graph Convolutional Recurrent Network for Traffic Forecasting, in Advances in Neural Information Processing Systems (NeurIPS), 33 (2020) 17830-17842.
[16]. J. Guo, Z. Xie, Y. Qin, L. Jia, Y. Wang, Short-term abnormal passenger flow detection with deep learning method for subway stations, Transportation Research Part C: Emerging Technologies, 136 (2022) 103556. https://doi: 10.1016/j.trc.2022.103556
[17]. J. Jiang, C. Han, W. X. Zhao, J. Wang, PDFormer: Propagation Delay-Aware Dynamic Long-Range Transformer for Traffic Flow Prediction, in Proceedings of the AAAI Conference on Artificial Intelligence, 37 (2023) 4365-4373.
[18]. Y. Zhang, Z. Chen, J. Li, Spatio-Temporal Graph ODE Networks for Traffic Flow Forecasting, in Proceedings of the AAAI Conference on Artificial Intelligence, 36 (2022) 4086-4094.
[19]. J. Zhou et al., Spatio-Temporal Identity: A Simple yet Effective Baseline for Multivariate Time Series Forecasting, in Advances in Neural Information Processing Systems (NeurIPS), 36 (2023).
[20]. B. Yu, H. Yin, and Z. Zhu, Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting, in Proceedings of the AAAI Conference on Artificial Intelligence, (2018) 3634-3640.
[21]. H. Liu, Z. Dong, R. Jiang, X. Song, I. W. Tsang, STAEformer: Spatio-temporal adaptive embedding makes vanilla transformer SOTA for traffic forecasting, in Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM), Birmingham, U.K., Oct. 2023, pp. 1–10. https://doi: 10.1145/3583780.3615160
[22]. C. Song, Y. Lin, S. Liu, X. Hu, and Z. Wang, Spatial-temporal synchronous graph convolutional networks: A new framework for spatio-temporal network data forecasting, in AAAI, 35 (2021) 11806-11814.