

Research Manuscript

Short-Term Highway Traffic Volume Prediction: A Case Study of the I-94 Interstate in the Saint Paul–Minneapolis Metropolitan Area Using Hybrid Machine Learning Models and Multidimensional Temporal Features

Pooyan Fallah^{*1}, Vadood Keramati²

¹Allameh Tabataba'i University, Tehran, Iran.

²Allameh Tabataba'i University, Tehran, Iran.

Received: 29/05/2026

Accepted: 05/10/2026

Abstract: This study investigates short-term highway traffic-volume prediction as a key component of Intelligent Transportation Systems (ITS), supporting traffic management, infrastructure planning, and reductions in delays and emissions. The Metro Interstate Traffic Volume dataset (48,204 records; 9 variables) was preprocessed through data-quality checks and removal of unrealistic observations. To capture temporal and weather-related effects, systematic feature engineering incorporated day/night status, working days, week position, hourly bins, seasons, lagged traffic volume, and environmental variables such as temperature, precipitation, and weather conditions. Several models were evaluated, including linear and negative binomial regression, nonparametric methods, regression trees, and random forests, with hyperparameters optimized using cross-validation-based randomized and grid search. The selected model achieved strong test performance ($R^2 = 0.9753$, SMAPE = 9.26%), substantially outperforming the raw-feature linear baseline ($R^2 \approx 0.0484$). Traffic volume at time t was predicted using contemporaneously available environmental variables and lagged traffic observations. The main contribution is a lag-aware, multidimensional feature-engineering and ensemble-learning framework for data-driven intelligent traffic management.

Keywords: Traffic volume prediction, Intelligent Transportation Systems (ITS), Feature engineering

Mathematics Subject Classification (2010): 62M10, 68T05, 90B20.

^{*}Corresponding Author: p.fallah@atu.ac.ir

1. Introduction

Rapid urbanization and growing travel demand have posed serious challenges to the efficient management of urban transportation networks. Increased congestion on highways leads to travel-time losses, higher fuel consumption, and aggravated environmental pollution. Consequently, the development of Intelligent Transportation Systems (ITS) capable of forecasting traffic conditions and supporting proactive decision-making has become increasingly important (Saha, Haque and Sidebottom , 2025; Ermagun and Levinson , 2018). In this context, accurate short-term traffic-volume forecasting can substantially improve active traffic management, traffic-flow control, and infrastructure planning (Saha, Haque and Sidebottom , 2025).

Over the past decades, a wide range of statistical and data-driven methods have been developed for short-term traffic prediction. While these approaches have significantly improved predictive capability, forecasting performance remains constrained by the ability of predictive models to represent temporal dependencies, nonlinear relationships, and contextual factors encountered in real-world transportation systems (Pavlyuk , 2017, 2019).

At the same time, the increasing availability of traffic, environmental, and contextual data has shifted greater attention toward data-driven prediction methods. Beyond the selection of prediction algorithms, recent research has increasingly emphasized the importance of effectively representing available information through appropriate feature engineering and feature selection, since these factors directly influence both predictive performance and model interpretability (Pavlyuk , 2019; Ou *et al.* , 2017).

Recent advances in artificial intelligence have further expanded the range of approaches available for traffic forecasting. In particular, deep-learning and spatiotemporal models have demonstrated strong capability for learning complex traffic patterns and nonlinear relationships. However, these approaches often require large training datasets, extensive parameter tuning, and substantial computational resources, while their limited interpretability can restrict their practical deployment in operational traffic-management systems (Saha, Haque and Sidebottom , 2025; Wang *et al.* , 2025).

Consequently, recent research has increasingly shifted attention from improving predictive algorithms alone toward developing prediction frameworks that achieve an effective balance between predictive accuracy, computational efficiency, and model interpretability. Within this context, growing evidence suggests that forecasting performance depends not only on model selection but also on the systematic representation of temporal, environmental, and historical traffic information through appropriate feature engineering. Nevertheless, practical and transferable strategies for designing such feature sets remain limited, particularly for applications requiring

both reliable prediction accuracy and computational efficiency (Pavlyuk , 2019; Ou *et al.* , 2017; Yang and Qian , 2019; Ermagun and Levinson , 2018).

To address this research gap, the present study develops a systematic feature-engineering framework for short-term highway traffic-volume prediction. The framework is evaluated using real-world data collected from the I-94 highway in the Minneapolis–Saint Paul metropolitan area, consisting of hourly traffic-volume records and environmental observations from 2012 to 2018. Temporal features (e.g., hour of day, day of week, and periodic patterns), weather-related environmental variables, and lagged traffic-volume variables (representing temporal dependence) are extracted and integrated into the predictive framework to capture short-term traffic dynamics while maintaining moderate model complexity (Pavlyuk , 2019; Ou *et al.* , 2017). In the present study, traffic volume observed at time t is modeled using contemporaneously available environmental variables together with lagged traffic observations. This formulation enables the analysis of short-term traffic behavior while preserving a feature framework that can be naturally extended to future-horizon forecasting tasks.

To evaluate the effectiveness of the proposed feature framework, a diverse set of statistical and machine-learning models—including linear regression, Poisson regression, negative binomial regression, nonparametric regression methods (KNN and Nadaraya–Watson), principal component regression, decision tree, and random forest—is assessed under a unified experimental framework using standard performance metrics. Rather than proposing a new prediction algorithm, the primary objective of this study is to quantify the contribution of systematically engineered temporal, environmental, and lagged features to short-term traffic-volume prediction across models with different levels of complexity.

The main contribution of this study is the development of a data-driven framework for the systematic design and management of temporal and environmental features and the empirical evaluation of their effectiveness using moderately complex statistical and machine-learning models. By emphasizing feature engineering rather than increasing algorithmic complexity, the proposed framework seeks to achieve a practical balance among predictive accuracy, computational efficiency, and model interpretability, thereby improving its suitability for real-time traffic-management systems. In addition, feature-importance analysis provides further insight into the relative influence of temporal and environmental factors on traffic-volume variations (Ermagun and Levinson , 2018; Yang and Qian , 2019).

2. Literature Review

2.1 Classical Statistical Models for Short-Term Traffic Forecasting

A substantial portion of early traffic-forecasting research is grounded in multivariate statistical modeling. Pavlyuk (2017) reviews a family of models—including ARIMA, VARMA, VECM, STARMA, and state-space approaches—for short-term prediction of traffic flow and speed. The shared rationale is to capture temporal (and, in some variants, spatial) dependence through autoregressive and moving-average structures, typically under predominantly linear assumptions (Pavlyuk, 2017).

Advantages: These models are theoretically mature, offer parameter interpretability, can often be estimated with relatively limited data, and typically incur low computational cost (Pavlyuk, 2017).

Limitations: Assumptions such as stationarity and linearity, together with simplified or partial representations of spatial and environmental effects, may reduce effectiveness in real-world networks—especially under strong nonlinearity, recurrent congestion, and complex spatiotemporal interactions (Pavlyuk, 2017). In addition, Ermagun and Levinson (2018) show that even within spatiotemporal approaches, a persistent gap remains between true link-level spatial dependence in real networks and the simplified spatial structures adopted in many studies.

2.2 Machine-Learning and Data-Mining Methods in Traffic Forecasting

With the expansion of traffic, environmental, and contextual data, machine-learning methods have been widely adopted for transportation forecasting and analysis, largely because they can learn nonlinear relationships and complex patterns without relying on restrictive distributional assumptions (Ermagun and Levinson, 2018).

In the accident-severity domain, Obasi and Benson (2023) compare Naive Bayes, Random Forest, logistic regression, and neural networks, reporting that Random Forest and logistic regression achieve approximately 87% overall accuracy. Their feature-importance analysis further highlights variables that contribute most strongly to predictive performance (Obasi and Benson, 2023). Although their target is accident severity rather than traffic flow, the study is methodologically relevant because it illustrates how models such as Random Forest can provide both strong predictive performance and interpretable feature-importance outputs—capabilities that also translate to modeling determinants of traffic flow, speed, or travel time (Obasi and Benson, 2023; Yang and Qian, 2019).

More directly for traffic-flow forecasting, Ou *et al.* (2017) combine a data-

driven feature-selection strategy with a bias-corrected Random Forest. They report that permutation-importance-driven feature selection, together with the modified Random Forest, reduces forecasting error while improving training efficiency (Ou *et al.* , 2017). In the domestic literature, Heidari and Hanani (2021) also emphasize the potential of data-mining techniques (e.g., decision trees, regression, clustering, and neural networks) for extracting patterns from traffic-related big data, though some contributions are primarily review- or framework-oriented.

Advantages: Decision trees, Random Forests, and nonparametric/nonlinear regression techniques can capture complex interactions and, in many cases, support post-hoc interpretation via feature-importance measures (Obasi and Benson , 2023; Ou *et al.* , 2017).

Limitations: Pavlyuk’s systematic review indicates that feature selection and feature construction remain ad hoc in a substantial share of the literature; the lack of comprehensive and transferable feature-engineering guidance can limit reproducibility and generalizability across corridors, cities, and datasets (Pavlyuk , 2019). This challenge is especially pronounced in data-driven studies that rely on heterogeneous inputs (lagged variables, spatial features, weather factors, and event indicators) (Ou *et al.* , 2017; Pavlyuk , 2019).

2.3 Deep-Learning and Graph-Based Spatiotemporal Models

Deep learning has become a dominant paradigm in time-series forecasting, including traffic forecasting, due to its ability to learn latent representations and complex nonlinear dynamics (Ermagun and Levinson , 2018). In a related domain, Saha, Haque and Sidebottom (2025) show that combining LSTM-based models with transfer learning and wavelet-based data augmentation can improve multi-step-ahead forecasting accuracy for Internet traffic. They also report increasing error and reduced stability as the forecasting horizon grows, highlighting a practical trade-off among model complexity, data scarcity, and long-horizon uncertainty (Saha, Haque and Sidebottom , 2025). Although the application domain differs, the methodological implications are relevant for road-traffic forecasting, where longer horizons and limited labeled data can similarly degrade performance.

For explicitly spatiotemporal modeling, Cui, Wang and Yin (2026) propose DG-STMTL, a multitask framework that constructs graph adjacency via a hybrid mechanism (domain prior + data-driven dynamic graph) regulated by task-specific gating. They report improved accuracy relative to baseline methods (Cui, Wang and Yin , 2026). From an interpretability perspective, Wang *et al.* (2025) introduce counterfactual explanations for deep GCN-based traffic forecasting to clarify the influence of contextual features and support practical deployment (Wang *et al.* ,

2025). In the domestic literature, [Gheisari and Bagherzadeh Khalkhali \(2018\)](#) compare MLP and NARX models and report that recurrent structures can be more suitable for longer horizons, while added complexity does not necessarily yield consistent advantages for short-term forecasting.

Advantages: Deep and graph-based models can learn complex spatiotemporal dependencies and multiscale patterns, and many studies report high predictive accuracy ([Cui, Wang and Yin, 2026](#); [Wang *et al.*, 2025](#)).

Limitations: High data and computational requirements, substantial tuning complexity, and interpretability challenges remain major barriers. Consequently, for online forecasting systems and resource-constrained deployments, moderately complex models with stronger transparency and lower runtime cost remain attractive ([Saha, Haque and Sidebottom, 2025](#); [Wang *et al.*, 2025](#); [Gheisari and Bagherzadeh Khalkhali, 2018](#)).

2.4 Feature Selection and Feature Extraction in Spatiotemporal Traffic Forecasting

[Pavlyuk \(2019\)](#), reviewing 211 studies, show that identifying spatiotemporal structure and applying feature selection/extraction are often decisive for forecasting performance; however, comprehensive, scenario-specific guidance for choosing feature-engineering methods remains limited. At the applied level, [Ou *et al.* \(2017\)](#) demonstrate that data-driven feature selection paired with Random Forest modeling can reduce error and improve training efficiency. [Yang and Qian \(2019\)](#) further report that integrating spatiotemporal traffic-flow features with weather and incident information—and applying methods such as PCA and LASSO—improves predictive performance compared with using traffic data alone. Nonetheless, feature design and combination are frequently implemented in a stepwise, case-specific manner rather than formalized into transferable frameworks ([Pavlyuk, 2019](#); [Yang and Qian, 2019](#)).

2.5 Critical Synthesis and Identification of the Research Gap

Overall, the literature suggests that:

- Classical statistical models are transparent and computationally efficient but may be limited under strong nonlinearity, environmental influences, and complex spatiotemporal dependence ([Ermagun and Levinson, 2018](#); [Pavlyuk, 2017](#)).

- Machine-learning methods often improve accuracy and support feature-importance analysis; however, feature engineering remains insufficiently systematic in much of the literature (Obasi and Benson , 2023; Pavlyuk , 2019; Ou *et al.* , 2017; Yang and Qian , 2019).
- Deep and graph-based approaches can be highly accurate, yet computational burden, tuning complexity, and interpretability constraints can limit practical adoption—particularly for short-horizon operational settings (Cui, Wang and Yin , 2026; Saha, Haque and Sidebottom , 2025; Wang *et al.* , 2025).
- Evidence consistently indicates that feature selection/extraction and explicit incorporation of temporal-spatial-environmental components can substantially improve forecasting accuracy, but an operational and transferable guideline for doing so remains a gap (Ou *et al.* , 2017; Pavlyuk , 2019; Yang and Qian , 2019).

3. Methodology

3.1 Overall Research Design

This study is empirical-analytical in nature and, using real-world traffic and weather observations, develops and evaluates a data-driven framework for short-term highway traffic-volume prediction. The central problem is to predict traffic volume at time t using contemporaneously available environmental information (e.g., weather and calendar effects) together with lagged traffic observations, which is a key capability for Intelligent Transportation Systems (ITS) and proactive traffic management (Ermagun and Levinson , 2018; Pavlyuk , 2019).

While a large body of work has focused on developing increasingly sophisticated predictive models, evidence from systematic reviews and applied studies suggests that careful feature selection and feature construction can be as influential as model choice for improving forecasting performance and robustness (Pavlyuk , 2019; Ou *et al.* , 2017; Yang and Qian , 2019). Accordingly, rather than emphasizing model complexity alone, the present study focuses on the systematic management of temporal, environmental, and lagged features and evaluates a set of moderately complex models within a unified framework.

The overall research workflow comprises six main stages:

- **Collection of raw data:** At this stage, traffic and environmental data are collected from a public database containing records captured by highway sensor systems.

- **Data preprocessing:** The raw data are examined for quality, and cleaning procedures, detection of implausible/unrealistic observations, and variable transformations are performed. In this step, inconsistent records are removed, and categorical variables are encoded into numeric representations to enable their use across different models.
- **Feature extraction and feature engineering:** A set of derived features is generated from the original variables. These features include temporal variables, environmental variables, and lagged traffic-volume variables, aiming to reveal latent structure in the data and enhance model predictive capability.
- **Modeling:** A suite of statistical and moderately complex machine-learning models is trained to predict traffic volume. The models include linear regression, Poisson regression, negative binomial regression, nonparametric methods, PCA-based regression, decision tree, and random forest.
- **Model performance evaluation:** Model performance is compared using standard model-evaluation metrics. The purpose of this stage is to quantify predictive accuracy and to examine the effect of the extracted features on model performance.
- **Selection of the optimal model:** Finally, based on evaluation criteria and performance analysis, the model that provides the best balance among predictive accuracy, robustness/stability, and computational cost is selected as the proposed model.

This staged structure allows the contribution of each component of the workflow—particularly feature engineering—to be assessed transparently in relation to final model performance. Thus, the present study not only seeks the most accurate predictive model but also analyzes how feature design affects the behavior of different machine-learning models.

3.2 Data and Information Sources

This research uses the Metro Interstate Traffic Volume dataset. The dataset contains traffic-volume information for Interstate 94 (I-94) in the Minneapolis–Saint Paul metropolitan area in Minnesota, USA. Traffic measurements were collected from a fixed traffic-monitoring location on the westbound segment of Interstate 94 between Minneapolis and Saint Paul and are publicly available to researchers. The study uses the Metro Interstate Traffic Volume dataset, which is publicly available through the UCI Machine Learning Repository ([Hogue , 2019](#)) and is also accessible via

Kaggle. The dataset was originally compiled from traffic monitoring data collected on Interstate 94 in the Minneapolis–Saint Paul metropolitan area, Minnesota, USA.

The dataset represents observations from a single highway-monitoring location and does not contain measurements from multiple sensors, roadway segments, or spatially distributed traffic stations.

The final dataset contains 48,204 hourly observations collected between October 2012 and September 2018. Such data enable the analysis of both short-term and longer-term temporal traffic patterns.

The dataset includes a set of traffic, temporal, and environmental variables that may affect traffic volume. Weather-related variables are provided as part of the original dataset and include temperature, precipitation, snowfall, cloud coverage, and weather condition descriptors associated with each hourly observation. The key variables are as follows.

- **Target variable (`traffic_volume`):** Number of vehicles passing the highway in each hour. This variable serves as the response variable in the predictive models developed in this study.
- Environmental and temporal variables (Table 1).

Table 1: Environmental and temporal variables in the Metro Interstate Traffic Volume dataset.

Variable group	Variable name	Data type	Description
Meteorological	<code>temp</code>	Continuous	Air temperature
Meteorological	<code>rain_1h</code>	Continuous	Accumulated rainfall
Meteorological	<code>snow_1h</code>	Continuous	Accumulated snowfall
Meteorological	<code>clouds_all</code>	Continuous	Cloud cover percentage
Meteorological	<code>weather_main</code>	Categorical	General weather
Meteorological	<code>weather_description</code>	Categorical	More detailed weather description
Calendar	<code>holiday</code>	Categorical	Holiday indicator and holiday type
Calendar	<code>date_time</code>	Datetime	Observation timestamp

The concurrent availability of traffic, temporal, and environmental variables in this dataset enables the investigation of interactions among different factors influencing changes in traffic volume.

Preliminary statistical analysis of the data indicates that the response variable `traffic_volume` has a mean of approximately 3,259 vehicles per hour and a standard deviation of approximately 1,987. The maximum value of this variable

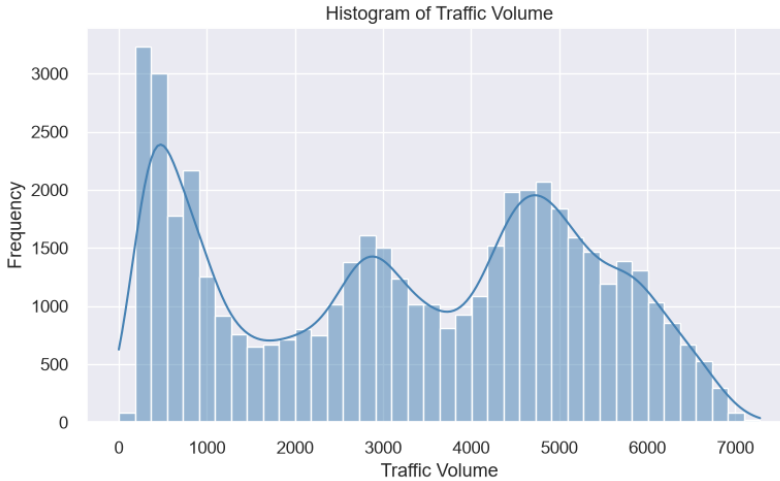


Figure 1: Histogram of traffic volume, showing the density of the target variable.

is reported as 7,280 vehicles per hour, which indicates peak traffic periods on the highway (Figure 1).

Among the environmental variables, `temp` has a mean of approximately 281 K, indicating diverse weather conditions over the observation period. In addition, `clouds_all` exhibits a relatively high dispersion and covers a wide range of atmospheric conditions. The distribution of general weather conditions in the dataset is illustrated in Figure 2.

3.3 Data Preprocessing

Before model development, the dataset was examined in terms of quality and consistency, and several preprocessing steps were applied.

Missing-data assessment.

In the initial assessment, no missing values were observed in the main variables of the dataset. During the temporal inspection, a small number of gaps between consecutive timestamps were identified, indicating the absence of some hourly records rather than missing values within existing records. Because these gaps were very limited in number and represented only a negligible fraction of the dataset, no imputation or interpolation procedure was applied, and the data were retained in their original form.

Detection and removal of unrealistic records.

The data inspection indicated that, in some records, the value of the variable `temp` had been recorded as 0 K. Since such a temperature is not physically possible under environmental conditions, these values were treated as unrealistic outliers

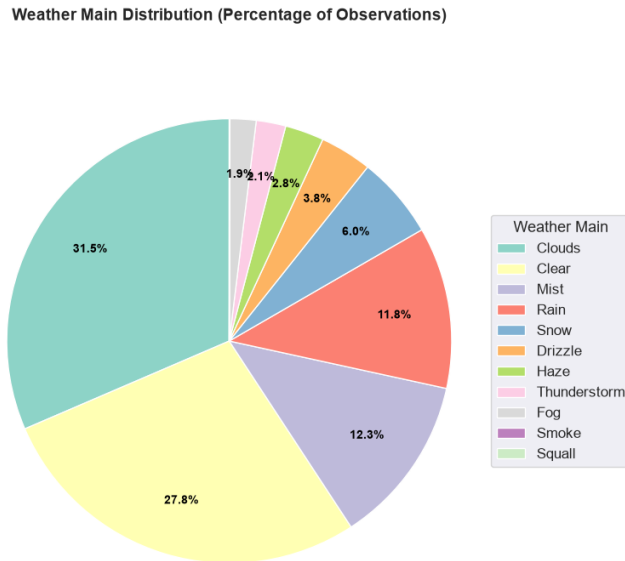


Figure 2: Pie chart of `weather_main` distribution.

and were removed from the dataset.

For the response variable `traffic_volume`, some very small or very large values were also observed. Given that, in real-world traffic data, traffic volume may be very low or very high during certain hours, these values were considered valid and were retained in the dataset.

Because the number of holidays was small, the values of the variable `holiday` were merged (rather than kept separately) to avoid having high-cardinality categories with low frequencies. The resulting new variable is binary with two classes: holiday and non-holiday.

Transformation of categorical variables.

To enable the use of categorical variables in different machine-learning models, variables such as `weather_main` and `weather_description` were transformed into numerical variables using dummy coding (one-hot encoding).

No records were removed due to missing values; only physically unrealistic observations were excluded during the preprocessing stage.

3.4 Variables and Indicators

Model inputs were defined as a combination of raw variables and engineered features. Feature selection was guided by evidence from prior traffic-forecasting studies emphasizing the importance of temporal structure, contextual information, and feature engineering, preliminary correlation analysis, and the time-series nature

of the prediction task (Pavlyuk , 2019; Ou *et al.* , 2017; Pavlyuk , 2017; Ermagun and Levinson , 2018).

Temperature variable.

The primary variable `temp` in the dataset records air temperature in Kelvin. For ease of interpretation and consistency with common practice in traffic engineering, temperature was converted to the Celsius scale:

$$\text{temp}_{\text{Celsius}} = \text{temp}_{\text{Kelvin}} - 273.15. \quad (3.1)$$

Temperature was retained as a continuous predictor. Prior work indicates that incorporating weather-related variables can improve short-term traffic prediction performance relative to using traffic history alone, reflecting the role of environmental conditions in travel behavior and network performance (Ou *et al.* , 2017; Yang and Qian , 2019; Ermagun and Levinson , 2018).

Holiday variable.

In the original dataset, the variable `holiday` is recorded as a categorical field (holiday name or None). For use in modeling, this variable was converted into a binary variable:

$$\text{holiday_binary} = \begin{cases} 1 & \text{if holiday,} \\ 0 & \text{if not holiday.} \end{cases} \quad (3.2)$$

This feature captures potential calendar-related shifts in travel demand, since traffic patterns on public holidays often differ from typical weekdays (Yang and Qian , 2019).

Time-of-day categorization.

Figure 3 presents the distribution of traffic volume by hour of day (panel a) and across four time-of-day categories (panel b). The hour component was extracted from the `date_time` variable and grouped into the following intervals:

- Night (00:00–05:59)
- Morning (06:00–11:59)
- Afternoon (12:00–18:59)
- Evening (19:00–23:59)

The resulting variable (`hour_cat`) was used as a categorical feature in the forecasting models. The purpose of this transformation was to represent recurrent daily traffic patterns in a more interpretable form while reducing the complexity associated with treating hour as a 24-level variable. This feature allows the models

to distinguish between broad traffic-demand regimes, including off-peak and peak-activity periods, without substantially increasing model dimensionality.

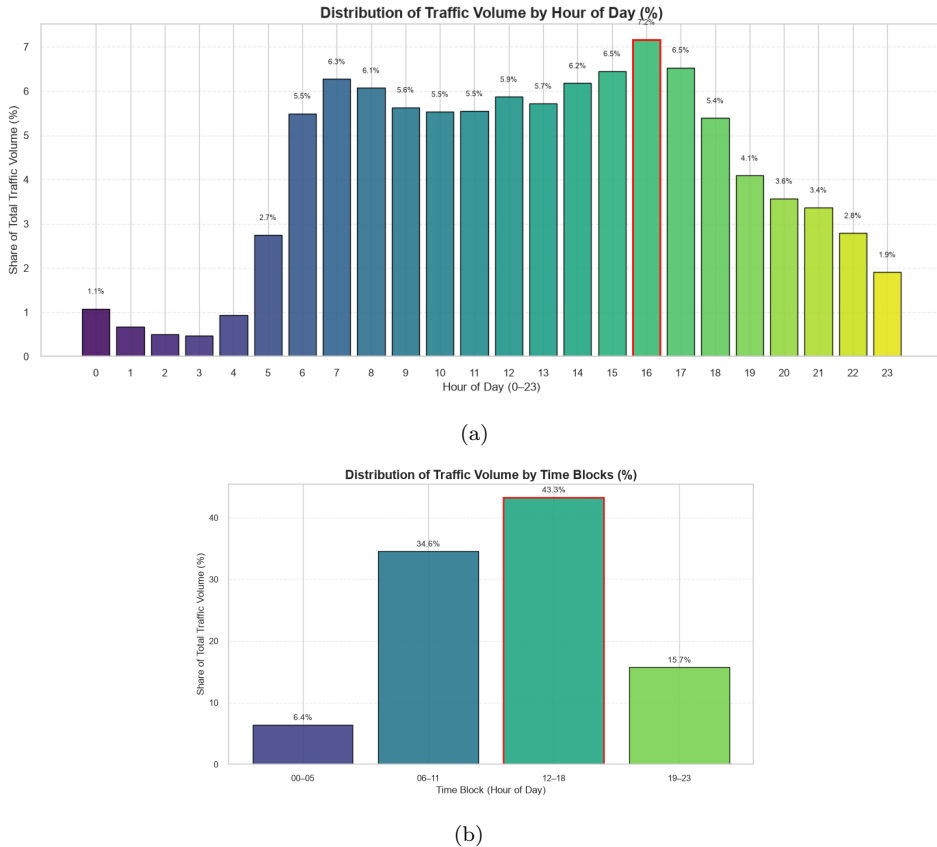


Figure 3: Distribution of traffic volume by hour and **hour_cat**: (a) share of total traffic volume by hour of day; (b) share by time-of-day block.

Season.

A seasonal indicator was derived from the observation date and categorized into four meteorological seasons: Spring (March–May), Summer (June–August), Autumn (September–November), and Winter (December–February). The resulting variable (**season**) was included as a categorical predictor to capture recurring annual patterns associated with weather conditions, travel demand, and highway utilization. Incorporating seasonal information enables the models to account for longer-term temporal variation beyond daily and weekly traffic cycles (Figure 4).

Weekday variable.

The variable **weekday** was extracted from the date and indicates whether the record corresponds to weekdays (working days) or the weekend. This variable was

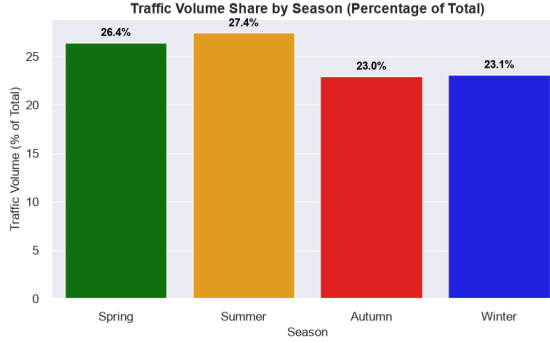


Figure 4: Traffic volume distribution by season.

defined in binary form:

$$\text{weekday} = \begin{cases} 1 & \text{if weekday,} \\ 0 & \text{if weekend.} \end{cases} \quad (3.3)$$

This feature captures systematic weekly seasonality, as traffic patterns on weekdays and weekends commonly differ, particularly around commuting peak periods (Yang and Qian , 2019; Ermagun and Levinson , 2018).

Lagged traffic-volume features.

Given the time-series nature of the data, traffic volume at a given hour is expected to depend on its recent historical values. To capture this temporal autocorrelation, lagged traffic-volume variables were generated from the target variable according to:

$$\text{LAG}_k = \text{traffic_volume}_{t-k}, \quad k = 1, 2, 3, \dots, 18. \quad (3.4)$$

An initial candidate set consisting of lag orders up to 40 hours was examined. The candidate lag range was evaluated using the autocorrelation function (ACF) and partial autocorrelation function (PACF) to identify the temporal dependence structure of the series and to determine a practically relevant lag horizon.

Visual inspection of the ACF/PACF patterns indicated that the strongest and most persistent autocorrelation signals were concentrated within the first 18 lag orders. Beyond this range, the autocorrelation values weakened substantially and contributed limited additional temporal information. Therefore, lag variables LAG_1 – LAG_{18} were retained and included as predictors in Scenario S_3 for all evaluated models (Figure 5).

The inclusion of lagged traffic-volume variables is widely recognized as an effective strategy for short-term traffic forecasting because it enables predictive

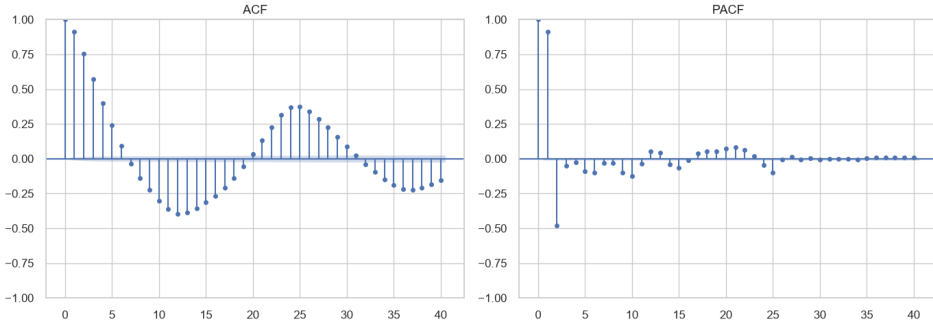


Figure 5: ACF and PACF of the traffic-volume series. The lag cutoff was selected at lag 18, corresponding to the last lag exhibiting a clear and practically meaningful autocorrelation signal before the dependence structure became substantially weaker.

models to capture temporal persistence, local temporal dependencies, and recurring traffic patterns (Pavlyuk , 2017; Ou *et al.* , 2017; Pavlyuk , 2019).

In the present study, traffic volume observed at time t is modeled using contemporaneously available environmental and calendar-related variables together with lagged traffic-volume observations recorded prior to time t . Specifically, each lagged predictor LAG_k represents the observed traffic volume at time $t - k$ and is therefore available at the prediction timestamp.

For observations belonging to the test period, lagged features were constructed exclusively from traffic-volume measurements that occurred before the prediction time, including observations immediately preceding the test interval and previously observed timestamps within the test period. No future observations were used during feature construction, model training, hyperparameter tuning, or model evaluation. This procedure preserves the chronological ordering of the data and prevents information leakage between training and test periods.

The response variable itself is not shifted to a future prediction horizon. Rather, temporal dependence is represented through engineered lag features that allow the models to learn short-term traffic dynamics while maintaining an interpretable supervised-learning framework. Because the methodology is based on explicit lagged predictors, it can be naturally extended to one-step-ahead or multi-step forecasting settings by redefining the response horizon in future research.

The selection of the above variables is grounded in three principles:

- **Coverage of calendar- and behavior-related factors:** The variables `hour_cat`, `season`, `weekday`, and `holiday_binary` were included to represent systematic calendar-driven and behavioral variations in travel demand. Prior traffic-forecasting studies consistently report strong periodic patterns across

the day, week, and special dates, and show that incorporating such temporal context can improve predictive accuracy and stability (Yang and Qian , 2019; Ermagun and Levinson , 2018).

- **Consideration of environmental conditions:** Weather-related variables (e.g., temperature, precipitation, cloud cover, and categorical weather conditions) were incorporated to capture environmental circumstances that may influence drivers' decisions and the effective operating conditions of the highway. The inclusion of weather variables is widely used in short-term traffic prediction and is frequently associated with improved performance relative to models that rely on traffic history alone (Ou *et al.* , 2017; Yang and Qian , 2019; Ermagun and Levinson , 2018).
- **Explicit modeling of temporal autocorrelation:** To account for temporal dependence in traffic volume, the study incorporates lagged traffic-volume features together with calendar-related and environmental variables. The resulting feature set enables the predictive models to simultaneously exploit temporal persistence, recurring demand patterns, and exogenous contextual information. This integrated representation provides a transparent and practically implementable framework for short-term traffic-volume prediction (Pavlyuk , 2019; Ermagun and Levinson , 2018).

In this study, candidate lag orders were initially examined through ACF and PACF diagnostics to identify the temporal dependence structure of the traffic-volume series. Based on these diagnostics, lag orders 1–18 were retained and used consistently across all models evaluated under Scenario S_3 . Feature-importance analyses were conducted subsequently as part of the model-interpretation stage and were used to assess the relative contribution of the retained lag variables within the final predictive models, particularly for tree-based methods such as Decision Tree and Random Forest (Pavlyuk , 2017; Ou *et al.* , 2017; Pavlyuk , 2019).

3.5 Methods, Models, and Algorithms

In this study, a set of statistical models and machine-learning algorithms with moderate complexity was employed to enable a comparable assessment of the effects of feature engineering and input-structure design within a consistent experimental framework. The models used are briefly introduced below.

3.5.1 Linear Regression

Linear regression provides a transparent baseline for modeling the relationship between a response variable and a set of explanatory variables and is widely used

as a benchmark in predictive studies (James *et al.* , 2021). Let y_t denote traffic volume at time t , and let $\mathbf{x}_t = (x_{t1}, \dots, x_{tp})^\top$ denote the corresponding vector of p predictors. The linear regression model is specified as:

$$\hat{y}_t = \beta_0 + \sum_{j=1}^p \beta_j x_{tj}, \quad (3.5)$$

where $\hat{y}_t$ is the predicted traffic volume, β_0 is the intercept, and β_j are regression coefficients. In this study, linear regression is used as a linear baseline for comparison with more flexible models.

3.5.2 Principal Component Regression (PCR; PCA Regression)

To reduce dimensionality and mitigate multicollinearity among predictors, we also considered principal component regression (PCR), which combines principal component analysis (PCA) with linear regression (Hastie, Tibshirani and Friedman , 2017). In PCR, PCA is first applied to transform the original predictors into orthogonal components (linear combinations of the input variables). Linear regression is then fit using the first m principal components, where m is chosen to retain most of the variability in the predictors while controlling model complexity. This approach is particularly useful when the feature set includes many derived variables (e.g., multiple lag terms and one-hot-encoded indicators), which can induce strong correlation structures and unstable coefficient estimates in ordinary least squares.

3.5.3 Poisson Regression

Because hourly traffic volume is a nonnegative count, Poisson regression was included as a classical count-data model (Cameron and Trivedi , 2013). Under this specification, the conditional distribution of y_t is assumed to be Poisson with mean parameter λ_t , and the relationship between λ_t and the covariates is modeled via a log link:

$$\log \lambda_t = \beta_0 + \sum_{j=1}^p \beta_j x_{tj}. \quad (3.6)$$

This model serves two purposes in our study: providing a principled baseline under a standard count-data assumption, and assessing whether the Poisson mean–variance relationship is adequate for the observed traffic counts.

3.5.4 Negative Binomial Regression

Traffic count data frequently exhibit overdispersion relative to the Poisson model, i.e., $\text{Var}(y_t | \mathbf{x}_t) > \mathbb{E}(y_t | \mathbf{x}_t)$. To accommodate this, we also estimated a negative

binomial regression model, which generalizes Poisson regression by allowing a more flexible variance structure through an additional dispersion parameter (Cameron and Trivedi, 2013). The systematic component remains log-linear in predictors, while the conditional variance can exceed the conditional mean. The negative binomial model was therefore used as a robustness check against Poisson misspecification driven by overdispersion.

3.5.5 Nonparametric Regression

To allow for nonlinear relationships between predictors and traffic volume without imposing a parametric functional form, two nonparametric regression approaches were considered: K-nearest neighbors (KNN) regression and kernel regression using the Nadaraya–Watson estimator (Hastie, Tibshirani and Friedman, 2017).

K-Nearest Neighbors (KNN) Regression.

In KNN regression, the prediction at a new point $\mathbf{x}$ is computed as an average of responses among the K closest training observations in the feature space, according to a chosen distance metric (Hastie, Tibshirani and Friedman, 2017). Because KNN is a local averaging method, it can capture nonlinearities and localized patterns but may be sensitive to feature scaling and the choice of K , particularly in higher dimensions.

Nadaraya–Watson Regression.

Kernel regression estimates the conditional mean function by computing a locally weighted average of observed responses, with weights determined by a kernel function and bandwidth h (Li and Racine, 2007). The Nadaraya–Watson estimator can be written as:

$$\hat{y}(\mathbf{x}) = \frac{\sum_{i=1}^n K_h(\mathbf{x} - \mathbf{x}_i) y_i}{\sum_{i=1}^n K_h(\mathbf{x} - \mathbf{x}_i)}, \quad (3.7)$$

where $K_h(\cdot)$ is a kernel function with bandwidth h , $\mathbf{x}_i$ are feature vectors, and y_i are observed traffic volumes. In practice, the bandwidth selection plays a central role in balancing bias and variance.

3.5.6 Decision Tree Regression

Decision tree regression is a nonparametric modeling approach based on recursive partitioning of the predictor space into regions with relatively homogeneous response values (Breiman *et al.*, 1984). By learning a sequence of split rules, decision trees can capture nonlinear effects and high-order interactions without explicitly

specifying them. In this study, decision trees were used both as predictive models and as diagnostic tools to provide model-based indications of feature relevance.

3.5.7 Random Forest Regression

Random forest is an ensemble learning method that improves predictive accuracy and stability by aggregating predictions from many decision trees built on bootstrap resamples of the training data, with randomized feature selection at each split (Breiman , 2001). This combination reduces variance relative to a single tree and typically yields strong out-of-sample performance in tabular prediction tasks. In addition to serving as a predictive model, random forest provides feature-importance signals that can support data-driven screening of candidate predictors.

3.6 Evaluation Procedure and Metrics

To evaluate predictive performance in short-term traffic-volume prediction, we report five standard accuracy metrics that are widely used in the time-series and highway-traffic prediction literature: the coefficient of determination (R^2), mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and symmetric mean absolute percentage error (sMAPE) (Hyndman and Athanasopoulos , 2021; Willmott and Matsuura , 2005). Lower values of MAE, RMSE, MAPE, and sMAPE indicate better predictive accuracy, while higher R^2 indicates better fit.

3.6.1 Coefficient of Determination (R^2)

The coefficient of determination measures the proportion of variance in the observed response explained by the model relative to a mean-only baseline. It is defined as (Hyndman and Athanasopoulos , 2021):

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad (3.8)$$

In general, with values closer to 1 indicating that the model explains a larger fraction of the variability in the test data. In this study, R^2 is reported as a global indicator of goodness-of-fit on the held-out test set.

3.6.2 Mean Absolute Error (MAE)

MAE quantifies the average magnitude of the absolute prediction error:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|. \quad (3.9)$$

MAE is straightforward to interpret and is comparatively less sensitive to extreme errors than squared-error measures (Hyndman and Athanasopoulos, 2021). Here, MAE is reported in vehicles per hour and can be read as the average hourly deviation between predicted and observed traffic counts.

3.6.3 Root Mean Squared Error (RMSE)

RMSE is computed as the square root of the average squared prediction error:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}. \quad (3.10)$$

Because squaring amplifies large residuals, RMSE penalizes occasional high-error periods more strongly than MAE. Reporting RMSE alongside MAE helps distinguish typical performance from performance under larger deviations, such as peak-demand hours (Hyndman and Athanasopoulos, 2021).

3.6.4 Mean Absolute Percentage Error (MAPE)

To express prediction error in relative, scale-free terms, we report MAPE:

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|. \quad (3.11)$$

MAPE is reported as a percentage and represents the average absolute error as a fraction of the observed value.

3.6.5 Symmetric Mean Absolute Percentage Error (sMAPE)

The Mean Absolute Percentage Error (MAPE) can become unstable when observed values approach zero, because the denominator contains the actual value. This issue is particularly relevant in traffic-volume prediction, where very low traffic counts may occasionally occur during off-peak periods. To provide a more robust scale-independent measure of prediction accuracy, we additionally report the Symmetric Mean Absolute Percentage Error (sMAPE) (Hyndman and Koehler, 2006):

$$\text{sMAPE} = \frac{100}{n} \sum_{i=1}^n \frac{2|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|}, \quad (3.12)$$

where y_i denotes the observed traffic volume and $\hat{y}_i$ denotes the corresponding prediction. Unlike MAPE, sMAPE normalizes the prediction error using both

observed and predicted values, reducing sensitivity to near-zero observations and providing a more balanced percentage-based assessment of predictive performance. Lower sMAPE values indicate higher predictive accuracy.

3.6.6 Evaluation Metrics

Model performance was assessed on the held-out test set using standard regression metrics:

- R^2 : proportion of variance explained;
- MAE: average absolute error in vehicle-count units;
- RMSE: penalizes larger errors more heavily than MAE;
- MAPE and sMAPE: relative error measures to compare accuracy across different traffic-volume levels.

3.7 Experimental Settings and Implementation

3.7.1 Data Splitting and Reproducibility

To evaluate model generalization under identical experimental conditions, the dataset was partitioned into two subsets:

- Training set (Train): 70% (33,735 observations)
- Test set (Test): 30% (14,459 observations)

Because the dataset consists of chronologically ordered traffic observations, a time-aware hold-out validation strategy was adopted rather than a random train-test split. This approach preserves the natural temporal structure of the data and prevents look-ahead bias and information leakage that may arise when future observations are used during model development.

Specifically, the observations were ordered according to their timestamps, and the earliest 70% of records were assigned to the training set, while the remaining 30% were reserved for out-of-sample evaluation. No random shuffling was performed at any stage of the data-splitting procedure.

Under this protocol, the training set contains 33,735 observations spanning from October 2, 2012 (09:00:00) to May 17, 2017 (21:00:00). The test set contains the subsequent 14,459 observations and covers the period from the first timestamp immediately following the training interval through September 30, 2018 (23:00:00). Consequently, the training and test sets are strictly non-overlapping and maintain the original chronological ordering of the series.

This chronological hold-out design provides a realistic assessment of predictive performance because each model is trained exclusively on historical observations and evaluated on future unseen data, closely reflecting the conditions encountered in practical traffic-forecasting applications.

Leakage control: All feature-engineering and preprocessing procedures (including scaling, encoding, and feature construction) were fitted using the training data only and subsequently applied to the test data, ensuring that the test set remained completely unseen during model development (Kuhn and Johnson, 2013). For observations in the test period, lagged features were generated using only traffic counts that would have been available at the prediction time: specifically, LAG_k variables for an observation at time t were calculated from observed traffic-volume values recorded at times $t - k$. Consequently, lagged predictors could include observations from the end of the training period and previously observed timestamps within the test period, but never any information from future timestamps. The test set was evaluated chronologically, preserving the natural temporal ordering of the data. No future observations were used during feature construction, model training, hyperparameter tuning, or model evaluation. This procedure preserves the temporal ordering of the data and prevents information leakage between training and test periods.

3.7.2 Hyperparameter Tuning

Hyperparameters were optimized using training data only through a two-stage search strategy. First, RandomizedSearchCV was employed to efficiently explore a broad hyperparameter space (Bergstra and Bengio, 2012). Subsequently, GridSearchCV was applied to refine the most promising region identified during the randomized search stage. The final hyperparameter configuration for each model was selected according to five-fold cross-validation performance on the training set. To improve reproducibility and transparency, the optimized hyperparameter values of the principal machine-learning models are reported in Table 2.

Table 2 summarizes the final hyperparameter settings obtained through the tuning procedure. These configurations were subsequently used for model training and evaluation on the held-out test set.

Table 2: Optimized hyperparameter configurations of the principal machine-learning models.

Model	Hyperparameter	Final Value
Decision Tree	<code>max_depth</code>	15
Decision Tree	<code>max_features</code>	None
Decision Tree	<code>min_samples_leaf</code>	5
Decision Tree	<code>min_samples_split</code>	100
Random Forest	<code>n_estimators</code>	200
Random Forest	<code>max_depth</code>	25
Random Forest	<code>min_samples_split</code>	5
Random Forest	<code>min_samples_leaf</code>	3
Random Forest	<code>max_features</code>	0.5
Random Forest	<code>bootstrap</code>	False
KNN	<code>n_neighbors</code>	5
KNN	<code>weights</code>	distance
KNN	<code>metric</code>	minkowski

4. Results and Research Findings

This section reports the out-of-sample results on the held-out test set under the experimental protocol described in Section 3.7, and compares models using the defined evaluation metrics to analyze predictive performance across traffic conditions.

4.1 Model-Performance Comparison and Effectiveness Assessment of the Feature-Engineering Framework

This section provides a quantitative and qualitative assessment of how the extracted features affect traffic-forecasting accuracy. To measure the contribution of each feature category to reducing prediction error, an ablation study was designed under three cumulative scenarios. The evaluations were conducted using two models with different functional forms—Linear Regression as a baseline model and Random Forest as an advanced nonlinear model—so that the generalizability of the proposed framework could be examined independently of the learning algorithm.

4.1.1 Experimental Scenario Configuration

The experimental scenarios were defined cumulatively to isolate the effect of each set of variables (Table 3):

Table 3: Experimental scenarios and their corresponding feature sets.

Scenario	Name	Feature Set Definition	Objective
S_1	Baseline	Basic problem-specific features	Establish a performance baseline and assess problem complexity.
S_2	Env + Time	Addition of temporal and environmental features	Evaluate the significance of contextual information.
S_3	Proposed	Addition of lagged features	Capture temporal dependencies through lagged traffic observations.

- **Scenario S_1 :** using the dataset’s original/raw features (no feature engineering).
- **Scenario S_2 :** adding temporal (calendar-cycle) and environmental (weather-condition) variables to S_1 .
- **Scenario S_3 :** the proposed framework; adding selected lag features derived from traffic volume to S_2 .

4.1.2 Results Analysis

The test-set performance results are reported in Table 4. It is noteworthy that, for the Linear Regression model in scenarios without lag features, MAPE was not reported due to the presence of zero values in the target variable and the resulting division-by-zero issue. MAPE values are reported when finite and interpretable, while sMAPE is reported for all models. While MAPE remains widely used in forecasting studies, sMAPE is included because it is less sensitive to near-zero observations and therefore provides a more stable relative-error measure.

4.1.3 Findings Analysis and Substantiation of the Claimed Contribution

The results confirm three key findings:

- **Insufficiency of raw-feature models (S_1):** The weak performance of models in Scenario S_1 ($R^2 < 0.1$) highlights the high complexity of the traffic-

Table 4: Predictive performance of the evaluated models across the three feature-engineering scenarios.

Scenario	Model	R^2	MAE	RMSE	MAPE (%)	sMAPE (%)
S_1	LR	0.0484	1675.07	1935.21	∞	62.13
S_1	RF	0.0967	1624.98	1892.44	249.11	60.77
S_2	LR	0.7128	781.06	1063.12	∞	33.78
S_2	RF	0.7415	746.81	1012.30	110.58	32.45
S_3	LR	0.9176	404.12	566.85	23.50	20.12
S_3	RF	0.9753	208.54	310.32	10.05	9.26

volume prediction task. Relying on raw features—without incorporating time-dependent and environmental variables—prevents the model from learning the dominant underlying patterns.

- **Determinant role of contextual information (S_2):** Incorporating temporal and environmental features yielded a substantial improvement in predictive accuracy, with the Random Forest model reaching $R^2 = 0.7415$. This finding indicates the intrinsic importance of calendar cycles and weather conditions in explaining traffic fluctuations.
- **Importance of temporal memory in the proposed framework (S_3):** The best performance was achieved by incorporating lag features. With these features, the Random Forest model attained the highest accuracy and the lowest error. This result supports the study’s claim that traffic flow exhibits temporal memory, and shows that the proposed framework—by modeling these dependencies—effectively learns complex interactions between lagged variables and external conditions.

4.1.4 Qualitative Residual Analysis

The residual diagnostics provide additional evidence regarding the adequacy and robustness of the proposed feature-engineering framework and the final Random Forest model. While the evaluation metrics quantify predictive accuracy, residual analysis enables an assessment of whether meaningful information remains unexplained after model fitting.

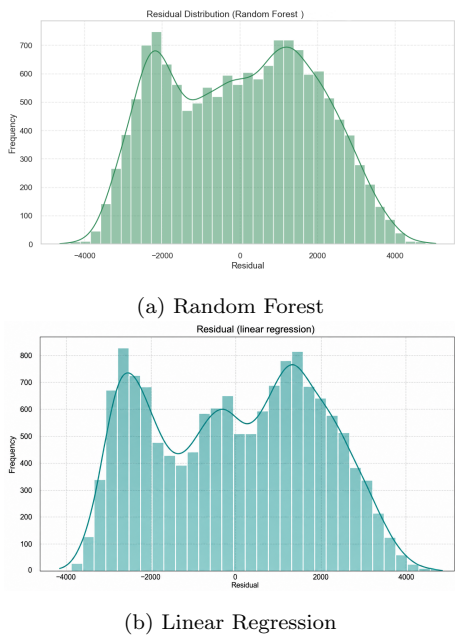


Figure 6: Residual histograms for Scenario S_1 models: (a) Random Forest; (b) Linear Regression.

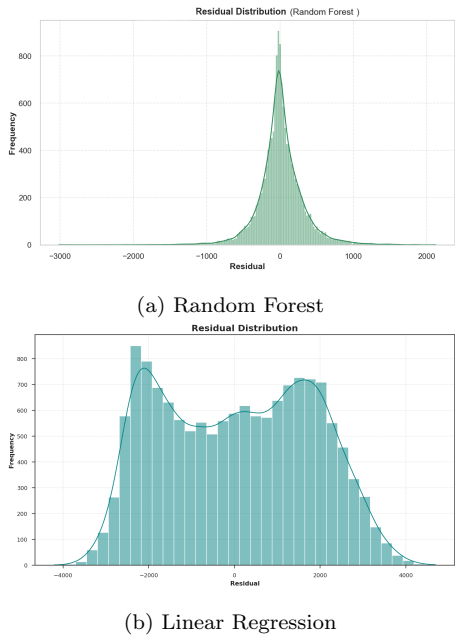


Figure 7: Residual histograms for Scenario S_3 models: (a) Random Forest; (b) Linear Regression.

Figures 6 and 7 compare the residual distributions obtained under the baseline

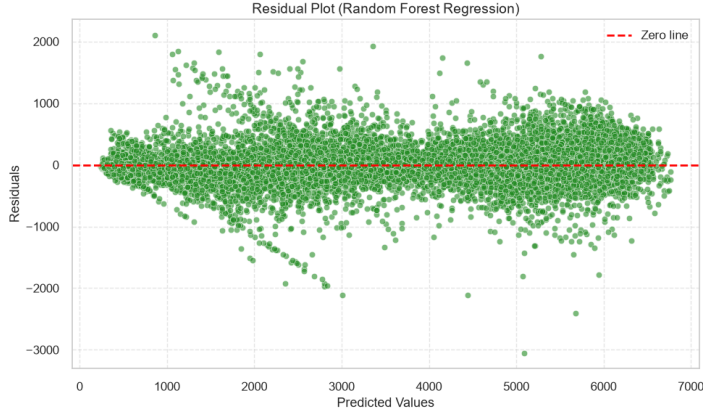


Figure 8: Residuals versus fitted values for the final Random Forest model under Scenario S_3 .

feature set (S_1) and the proposed feature-engineering framework (S_3). The baseline scenario exhibits a wider and more asymmetric residual distribution, indicating larger prediction errors and a greater frequency of extreme deviations. In contrast, the residuals produced by the final Random Forest model under Scenario S_3 are substantially more concentrated around zero, demonstrating a considerable reduction in forecasting errors. This observation is consistent with the improvements observed in R^2 , MAE, and RMSE and confirms the effectiveness of the proposed feature-engineering strategy.

Figure 8 presents the residual-versus-fitted plot. The residuals are distributed around the zero line throughout the prediction range without exhibiting a clear systematic trend or curvature. This behavior suggests that the model successfully captures the dominant nonlinear relationships between predictors and traffic volume and does not suffer from substantial systematic bias. Although a small number of outlying observations remain, no strong evidence of model misspecification is observed.

Figure 9 shows the autocorrelation function (ACF) of the residuals. The autocorrelation coefficients remain very close to zero across all examined lags, indicating the absence of meaningful temporal dependence in the residual sequence. This result is particularly important because traffic-volume forecasting inherently involves strong temporal correlations. The near-white-noise behavior of the residuals demonstrates that the proposed lag-based feature-engineering framework effectively captures the temporal structure of the traffic series, leaving little exploitable information in the prediction errors.

Figure 10 presents the normal Q–Q plot of the residuals. The central portion of the residual distribution follows the theoretical normal line reasonably well,

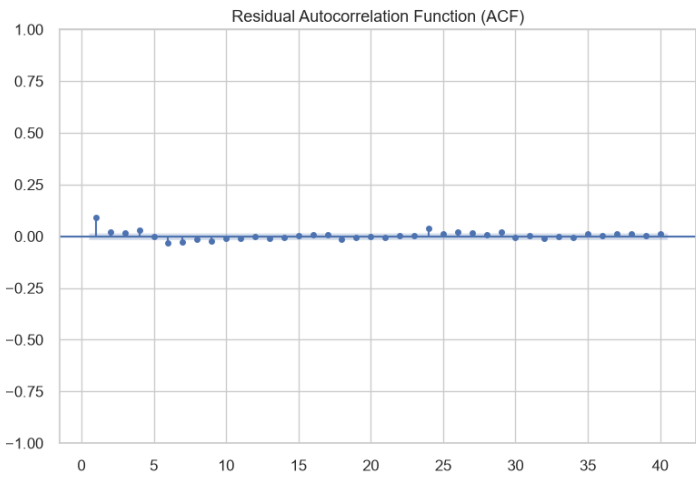


Figure 9: Autocorrelation function (ACF) of residuals for the final Random Forest model under Scenario S_3 .

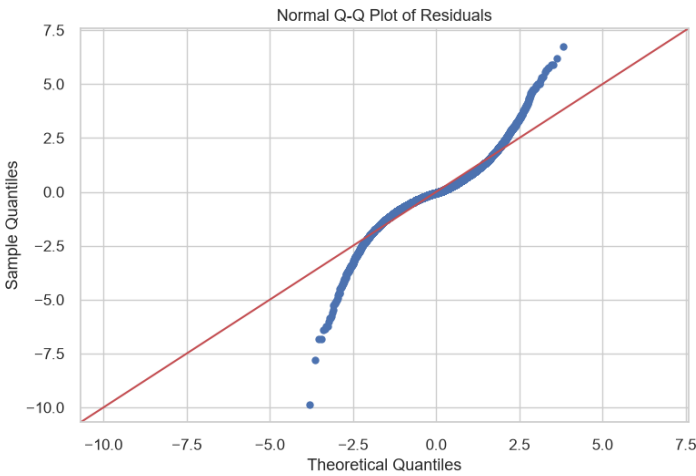


Figure 10: Normal Q-Q plot of residuals for the final Random Forest model under Scenario S_3 .

Table 5: Test-set model benchmarking results (Scenario S_3).

Model	R^2	MAE	RMSE	MAPE (%)	sMAPE (%)
Linear Regression	0.9176	404.12	566.85	23.50	20.12
PCA Regression	0.6661	889.87	1141.21	58.57	37.43
Poisson Regression	0.8900	468.52	655.10	29.87	25.44
NB Regression	0.8338	562.38	805.14	27.78	25.68
KNN Regression	0.8946	429.40	641.29	22.73	20.67
Nadaraya–Watson	0.9010	399.04	621.52	20.82	23.88
Regression Tree	0.9631	254.01	379.56	12.35	11.37
Random Forest	0.9752	209.13	310.99	10.05	9.26

whereas deviations are observed in the upper and lower tails. These departures indicate the presence of a limited number of large forecasting errors associated with unusual traffic conditions and extreme observations. Such behavior is common in traffic forecasting applications and is not unexpected for a non-parametric ensemble model such as Random Forest, which does not rely on the normality assumption for predictive validity.

Taken together, the residual diagnostics indicate that the final model produces approximately unbiased predictions, exhibits negligible residual autocorrelation, and leaves primarily random prediction errors. These findings provide additional support for the effectiveness of the proposed feature-engineering framework and confirm that the observed performance improvements arise from genuine extraction of temporal and contextual information rather than overfitting or spurious correlations.

4.2 Model Benchmarking

This section aims to support a principled selection of the final model and to strengthen the empirical evidence by benchmarking several model families—linear, dimensionality-reduction, count-based, neighbor-based, kernel-based, and tree-based approaches—under an identical evaluation protocol on the test set. The primary reported metrics are R^2 , MAE, RMSE, MAPE, and sMAPE (Table 5).

The Random Forest model achieved the highest R^2 and the lowest MAE/RMSE on the test set; therefore, in terms of numerical accuracy, it is the superior option (Figure 11).

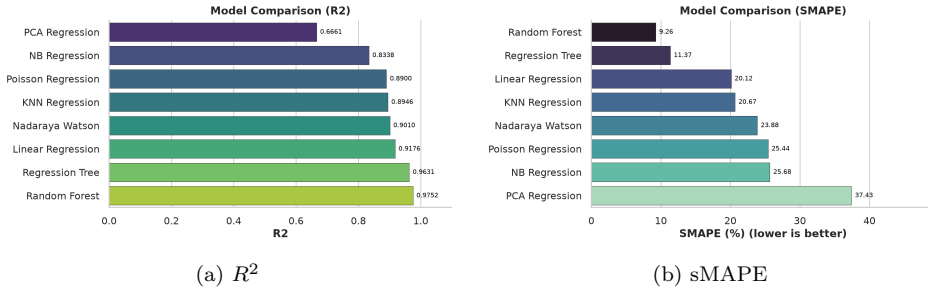


Figure 11: Bar charts of model-evaluation results across benchmarked models: (a) R^2 ; (b) SMAPE (lower is better).

4.3 Interpretability, Feature Contributions, and Stability Checks

4.3.1 Feature Contributions in the Final Model

Figure 12 shows a clear pattern in feature contributions within the final model. The top-ranked features are lagged traffic-volume variables, with LAG_1 exhibiting the highest importance. A set of additional lag features also appears among the most influential variables. This pattern is consistent with the time-series nature of traffic, as traffic volume at the current hour is directly affected by behavior in preceding hours.

Alongside lag variables, several temporal features also appear among the higher-ranked predictors. The placement of hour_{cat} features in subsequent ranks indicates that the model has captured daily cycles and differences between peak and off-peak hours. This explains why adding temporal features in earlier stages led to a meaningful improvement, and why nonlinear models, when combined with these features, exhibited better performance.

Moreover, the presence of at least one environmental variable such as temp among the top features suggests that weather-related variables play a complementary role. The effect of this feature class is typically condition-dependent—i.e., it contributes more to error reduction during certain intervals (e.g., temperature fluctuations or specific weather regimes)—but, compared with the traffic “temporal memory” captured by lag variables, it is not the primary driver.

4.3.2 Practicality and Cost-Effectiveness of the Feature Framework

The feature-importance results also convey an applied message: the proposed framework is practical because it relies on features that are inexpensive, readily available, and easy to update:

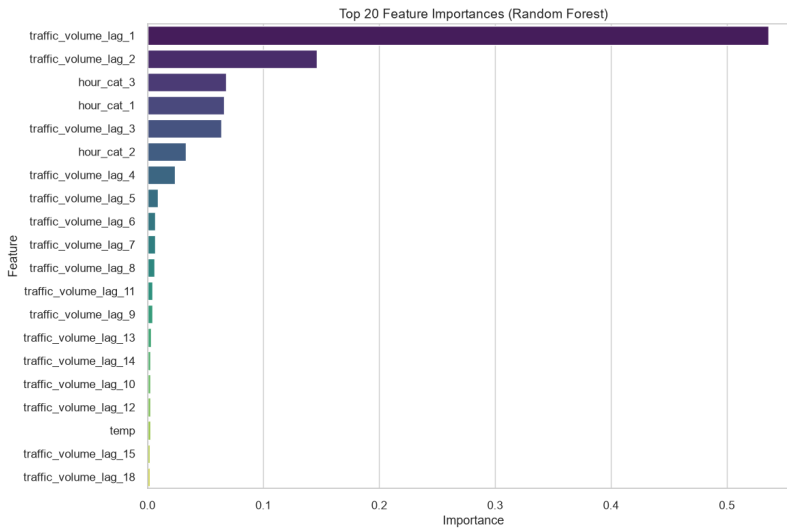


Figure 12: Top-20 feature importances in the Random Forest model.

- Lag features require only a short history of the target variable itself; such data are typically recorded continuously in traffic-monitoring systems and incur no additional acquisition cost.
- Temporal features (hour / hour category) are extracted from the timestamp and thus have negligible computational cost.
- Environmental features such as temperature are commonly obtainable via APIs or meteorological stations; given the resulting accuracy gains, the data-collection cost is justifiable.

From an efficiency perspective, the feature-importance profile indicates that a substantial portion of the model's predictive power is driven by a limited set of key features. Accordingly, in operational deployments, focusing on selected lag variables and a small number of core temporal features can achieve an appropriate balance between accuracy and complexity—an important consideration for real-world (near-real-time) traffic prediction.

4.3.3 Qualitative Assessment of Practical Performance

To qualitatively assess model behavior, Figure 13 shows the model's predictions over a randomly selected window of $n = 250$ samples. In this figure, the observed series is plotted alongside the raw predictions (Predicted Raw) as well as a smoothed version. Visualizing these three components simultaneously helps highlight two points:

- The model follows the overall traffic trend and reflects slow and cyclic variations well.
- In regions with sharper fluctuations, the discrepancy between observations and predictions increases; however, the smoothed trend curve indicates that the model still preserves the medium-term structure of the data.

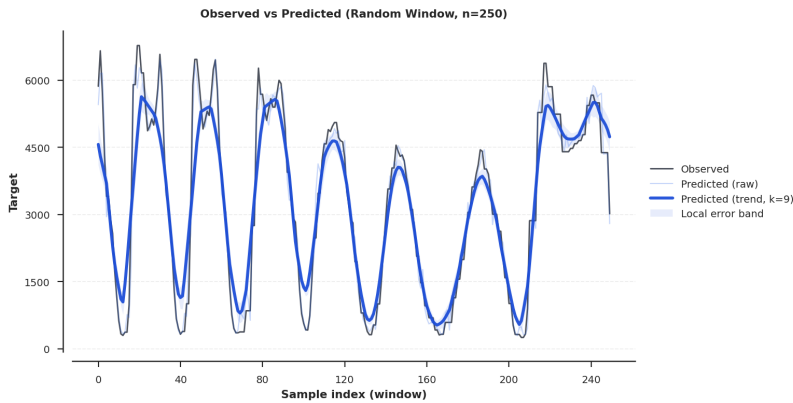


Figure 13: Observed vs. predicted traffic volume over a random window of 250 samples.

5. Conclusion

This study addressed traffic-volume prediction with an emphasis on systematic feature engineering and the evaluation of data-driven models. The objective was to clarify the role of temporal features, environmental features, and, in particular, lag features in improving predictive accuracy; accordingly, an ablation study was designed to quantify the contribution of each feature category separately.

Test-set results showed that model choice alone does not guarantee high accuracy; rather, enriching the feature space is the determining factor in enabling the learning of traffic patterns. Within the proposed framework, the Random Forest model delivered the best performance. These findings indicate that incorporating selected lag variables and nonlinear interactions substantially enhances the model's ability to reconstruct traffic fluctuations.

From an applied perspective, the proposed framework can serve as a decision-support tool in ITS applications for congestion anticipation and resource-allocation optimization. Nevertheless, extending these results to other cities or different time periods requires additional validation and careful consideration of local dynamics.

6. Recommendations and Future Work

In future work, it is recommended to assess the model's temporal transferability and geographic generalizability by testing across different seasons and on different routes/stations, and to investigate the effect of unusual events (e.g., holidays, crashes, road closures) via labeling, filtering, or explicit modeling, since such events are typically a major source of error at traffic peaks.

From a methodological standpoint, comparison with deep time-series models (e.g., TCN, LSTM/GRU, and Transformer architectures) is recommended as a benchmark. In addition, to improve the reliability of interpretability analyses, it is necessary to adopt more stable explainability approaches such as Permutation Importance and SHAP, and to examine the stability of feature rankings across different temporal slices. Finally, systematic optimization of lag selection (range and window length), along with the inclusion of rolling statistics, may yield a better balance between accuracy and complexity.

Operationally, the current study predicts traffic volume at time t using contemporaneously available contextual variables and lagged observations. Future work may extend this framework to one-step-ahead and multi-step forecasting by redefining the response variable at future horizons ($t + h$).

References

- Bergstra J. and Bengio Y. (2012), Random search for hyper-parameter optimization, *Journal of Machine Learning Research*, **13**, 281–305.
- Breiman L. (2001), Random forests, *Machine Learning*, **45**(1), 5–32.
- Breiman L., Friedman J.H., Olshen R.A. and Stone C.J. (1984), *Classification and Regression Trees*, Belmont, CA, Wadsworth.
- Cameron A.C. and Trivedi P.K. (2013), *Regression Analysis of Count Data*, 2nd ed., Cambridge, Cambridge University Press.
- Cui W., Wang P. and Yin F. (2026), DG-STMTL: A novel graph convolutional network for multi-task spatio-temporal traffic forecasting, *Neurocomputing*, Article 132421.
- Ermagun A. and Levinson D. (2018), Spatiotemporal traffic forecasting: Review and proposed directions, *Transport Reviews*, **38**(6), 786–814.
- Gheisari Y. and Bagherzadeh Khalkhali A. (2018), Short- and long-term traffic flow forecasting using neural-network methods, *Proceedings of the 5th National Conference on Applied Research in Civil Engineering, Architecture and Urban Management*, Tehran, Iran, Khajeh Nasir Toosi University of Technology. (in Persian)

- Hastie T., Tibshirani R. and Friedman J. (2017), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., New York, Springer.
- Heidari M. and Hanani A. (2021), A model for traffic prediction using big-data mining, *Proceedings of the 8th International Conference on Novel Solutions in Engineering, Information Science and Technology*, Dubai, United Arab Emirates. (in Persian)
- Hogue J. (2019), Metro Interstate Traffic Volume, *UCI Machine Learning Repository*, doi:10.24432/C5X60B.
- Hyndman R.J. and Athanasopoulos G. (2021), *Forecasting: Principles and Practice*, 3rd ed., Melbourne, OTexts.
- Hyndman R.J. and Koehler A.B. (2006), Another look at measures of forecast accuracy, *International Journal of Forecasting*, **22**(4), 679–688.
- James G., Witten D., Hastie T. and Tibshirani R. (2021), *An Introduction to Statistical Learning: With Applications in R*, 2nd ed., New York, Springer.
- Kuhn M. and Johnson K. (2013), *Applied Predictive Modeling*, New York, Springer.
- Li Q. and Racine J.S. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton, NJ, Princeton University Press.
- Obasi I.C. and Benson C. (2023), Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic accidents, *Heliyon*, **9**(8), e18812.
- Ou J., Xia J., Wu Y.J. and Rao W. (2017), Short-term traffic flow forecasting for urban roads using data-driven feature selection strategy and bias-corrected random forests, *Transportation Research Record*, **2645**(1), 157–167.
- Pavlyuk D. (2017), Short-term traffic forecasting using multivariate autoregressive models, *Procedia Engineering*, **178**, 57–66.
- Pavlyuk D. (2019), Feature selection and extraction in spatiotemporal traffic forecasting: A systematic literature review, *European Transport Research Review*, **11**, 6.
- Saha S., Haque A. and Sidebottom G. (2025), Overcoming data limitations in internet traffic forecasting: LSTM models with transfer learning and wavelet augmentation, *Computer Communications*, **242**, 108280.
- Wang R., Xin Y., Zhang Y., Perez-Cruz F. and Raubal M. (2025), Counterfactual explanations for deep learning-based traffic forecasting, *Communications in Transportation Research*, **5**(2), 100176.
- Willmott C.J. and Matsuura K. (2005), Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, *Climate Research*, **30**(1), 79–82.
- Yang S. and Qian S. (2019), Understanding and predicting travel time with spatio-temporal features of network traffic flow, weather and incidents, *IEEE Intelligent Transportation Systems Magazine*, **11**(3), 12–28.