Available Online: 29 April 2026 DOI: 10.54254/3029-0880/2026.33093

A comparative study of machine learning models for monthly new energy vehicle sales forecasting: evidence from China

Sunong Wu[1], Rui Wang[2*]

[1]School of Transportation Engineering, East China Jiaotong University, Nanchang, China
[2]College of Management, Shenzhen University, Shenzhen, China

*Corresponding Author. Email: wangrui22825@163.com

Abstract. Accurate sales forecasting of New Energy Vehicles (NEVs) is critical for automakers, policymakers, and energy infrastructure planners. To address the multifaceted drivers of NEV sales, this study proposes a machine learning-based forecasting framework using monthly data from China spanning January 2020 to December 2025. Two key innovations are introduced, namely a composite station factor that integrates charging station count, battery swap station count, and gasoline prices to capture infrastructure and fuel cost effects, and a time-varying categorical variable that encodes major policy shocks, including subsidy phase-outs, COVID-19 lockdowns, purchase tax adjustments, and trade-in incentives. A systematic comparison of multiple models shows that Extreme Gradient Boosting (XGBoost) equipped with the proposed features achieves the best performance, with a Mean Absolute Error (MAE) of 61,169, Root Mean Square Error (RMSE) of 71,972, Mean Absolute Percentage Error (MAPE) of 6.09%, and Symmetric Mean Absolute Percentage Error (SMAPE) of 6.00%. This represents improvements over the univariate baseline of 62.76% in MAE, 63.14% in RMSE, 63.62% in MAPE, and 63.19% in SMAPE. Finally, ablation experiments and SHAP analysis are performed to validate the effectiveness of the proposed features. The findings demonstrate that the proposed framework offers a practical and interpretable tool for NEV sales forecasting and policy simulation.

Keywords: new energy vehicles, sales forecasting, XGBoost, infrastructure factor, policy impact.

1. Introduction

The rapid growth of New Energy Vehicles (NEVs) has become a central pillar of China's strategy to reduce carbon emissions and transform the automotive industry [1]. According to the China Association of Automobile Manufacturers, NEV sales in China exceeded 16 million units in 2025, accounting for 68.3% of the global market. Accurate monthly sales forecasting is therefore critical for manufacturers to plan production, for grid operators to manage charging demand, and for policymakers to design effective incentive schemes [2].

However, NEV sales are influenced by a complex set of factors: traditional time-series momentum, seasonal patterns, the availability of charging and battery-swap infrastructure, volatile gasoline prices, and

frequent policy interventions [3]. Most existing studies rely solely on historical sales as the predictor, neglecting the multiple concurrent factors that influence NEV sales. Moreover, systematic comparisons of machine learning models for NEV sales forecasting are scarce, and the interpretability of black-box models remains a challenge.

This study addresses these gaps by making the following contributions:

• We construct a novel composite station factor that integrates charging station count, battery swap station count, and gasoline prices, capturing the joint effect of infrastructure and fuel cost.

• We manually code major policy events from 2020 to 2025 into a time-varying policy factor, reflecting negative, neutral, or positive expected impacts on NEV sales.

• We conduct a systematic comparison of multiple models, along with ablation experiments and SHAP interpretability analysis to validate the effectiveness of the proposed station factor and policy factor.

2. Literature review

Early research on vehicle sales forecasting relied on classical time series methods. For example, Qi [4] developed an ARIMA model to forecast China's NEV sales from 2025 to 2027, projecting a compound annual growth rate of 14.4% based on historical data from 2015 to 2024. However, such models assume linearity and cannot readily incorporate exogenous variables, leading to limited forecasting accuracy in complex real-world settings.

More recently, machine learning models have gained popularity for sales forecasting tasks. Qu et al. [5] applied support vector regression optimized by a grey wolf optimizer to automobile sales forecasting, achieving improved performance through metaheuristic tuning. Si et al. [6] combined grey forecasting with random forest to predict NEV sales, while Wang et al. [7] compared random forest and ARIMA models, finding that random forest better captured nonlinear relationships. Zhang et al. [8] proposed an ARIMA-XGBoost combined model for plug-in hybrid electric vehicle sales, which outperformed each individual model by leveraging the strengths of both linear and nonlinear approaches.

Deep learning models such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are theoretically attractive for sequence prediction and have been increasingly applied to NEV-related forecasting tasks. For instance, Li et al. [9] developed a Convolutional Neural Network (CNN) -LSTM model to analyze the demand and supply gap of China's charging infrastructure, demonstrating the ability of hybrid deep learning architectures to capture spatial-temporal dependencies. Similarly, Li [10] proposed a WOA-BiGRU model for monthly NEV sales forecasting in urban areas and achieved promising accuracy. However, these deep learning models require large training samples and careful hyperparameter tuning. In small-sample monthly settings, they often underperform tree-based models due to overfitting and unstable convergence.

Despite these advances, existing studies still lack a systematic integration of infrastructure and policy factors into a composite feature, as well as a comprehensive model comparison under a unified framework.

3. Research design

3.1. Data sources

We collect monthly data from January 2020 to December 2025, resulting in 72 observations. The dependent variable is the NEV sales volume of the current month, and the main independent variables include the NEV sales, infrastructure indicators (battery swap station count and charging station count), gasoline price, and a manually coded policy factor. The NEV sales data are obtained from the Yiche website (https://car.yiche.com).

The numbers of battery swap stations and charging stations are sourced from the Electric Vehicle Charging Infrastructure Promotion Alliance (EVCIPA, https://evcipa.com). The monthly average gasoline price is taken from the Eastmoney website (https://data.eastmoney.com). The policy factor is manually coded based on official government announcements (https://www.gov.cn) and major news reports, with detailed assignments listed in Section 3.2.3. All independent variables are lagged by one month to avoid look-ahead bias. We chronologically split the dataset into a training set (January 2020 to December 2024, 60 months) and a test set (January 2025 to December 2025, 12 months).

3.2. Variable construction

3.2.1. Time encoding

To capture the periodic nature of monthly sales (e.g., year-end spikes and post-holiday dips), we transform the month number into sine and cosine components:

$$month\_sin = \sin\left(2\pi\frac{month}{12}\right) \tag{1}$$

$$month\_cos = \cos\left(2\pi\frac{month}{12}\right) \tag{2}$$

where $month$ takes values from 1 to 12. This encoding preserves the cyclic proximity (e.g., December and January are close) and avoids the discontinuity of one-hot encoding.

3.2.2. Station factor

We construct a composite index that reflects the joint influence of charging and swapping infrastructure and gasoline prices on NEV adoption. Intuitively, more stations and higher fuel prices encourage consumers to switch to electric vehicles. Each component is min-max normalized to the range from zero to one. The station factor is then the arithmetic mean of the three normalized variables:

$$station\_factor = \frac{1}{3}\left(norm\left(C_{swap}\right) + norm\left(C_{charge}\right) + norm\left(P_{gas}\right)\right) \tag{3}$$

$$norm\left(C\right) = \frac{C - C_{min}}{C_{max} - C_{min}} \tag{4}$$

Where C_{swap} is the count of battery swap stations, C_{charge} is the count of charging stations, and P_{gas} is the gasoline price. The resulting station factor ranges between 0 and 1, with higher values indicating more favorable conditions for NEV sales.

3.2.3. Policy factor

We manually review major policy events and unexpected shocks that could affect NEV sales in China from 2020 to 2025. The timeline begins with the nationwide COVID-19 lockdowns from January to May 2020. Thereafter, early 2021 saw a 30% reduction in NEV subsidies. The situation further deteriorated with a severe lockdown in Shanghai (a major hub for NEV manufacturing and sales) from March to June 2022. The year 2023 marked the complete phase-out of national NEV subsidies on January 1, followed by a purchase tax

adjustment effective January 1, 2024, which capped the tax exemption for each NEV at 30,000 RMB. Finally, a trade-in incentive program for NEVs was launched in April 2024.

Based on these events, we construct a time-varying policy factor. Each event is assigned a value of –1 for negative impact, 0 for neutral, or +1 for positive impact. For interval-type events (the COVID-19 outbreak and the Shanghai lockdown), the assigned value directly applies to the entire period. For point-in-time policies (subsidy reduction, subsidy phase-out, tax adjustment, and trade-in program), we extend the impact by one additional month to account for the lagged transmission of policy signals to consumer decisions. The resulting policy factor coding is illustrated in Figure 1.

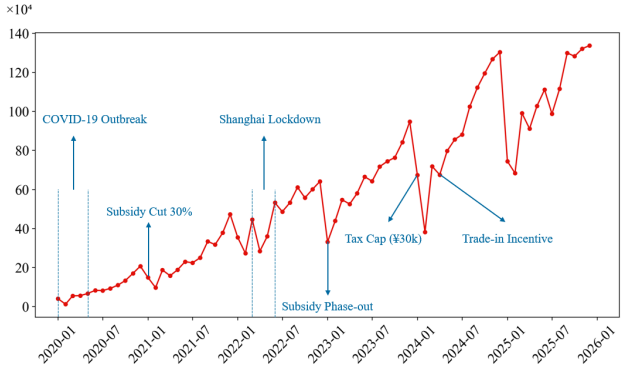


Figure 1. NEV monthly sales volume and policy factor coding timeline

3.3. Baseline and competing models

We compare twelve models spanning three categories: traditional time series (ARIMA [11]), classical machine learning (Linear Regression [12], Ridge Regression [13], Support Vector Regression [14], Random Forest [15], LightGBM [16], CatBoost [17], and XGBoost [18]), and deep learning (BPNN [19], RNN [20], GRU [21], LSTM [22]). For each model, we run two input configurations: ‘s’ uses only the sales, and ‘m’ uses all constructed features (sales, time encoding, station factor, and policy factor). For the deep learning models, given the limited sample size (60 training months), we adopt a simple architecture with a single hidden layer of 16 neurons, a dropout rate of 0.2 to mitigate overfitting, the Adam optimizer with a learning rate of 0.001, and train for 300 epochs. For all other models, we perform dedicated hyperparameter optimization using Optuna with 30 trials and TimeSeriesSplit with 5 folds, applying early stopping with 30 rounds to avoid overfitting.

3.4. Evaluation metrics

We evaluate forecasting accuracy using four standard metrics. Let y_i be the actual sales and $\hat{y}_i$ the predicted sales for month i in the test set.

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{5}$$

Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2} \tag{6}$$

Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y_i}}{y_i}\right| \times 100\% \tag{7}$$

Symmetric Mean Absolute Percentage Error (SMAPE)

$$SMAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i - \widehat{y_i}\right|}{\left(|y_i| + \left|\widehat{y_i}\right|\right)/2} \times 100\% \tag{8}$$

Lower values of all metrics indicate better forecasting performance. MAPE and SMAPE are reported as percentages.

4. Results

4.1. Overall model comparison

Table 1 presents the test set performance of all twelve models under the single (s) and multivariate (m) input settings, along with the percentage improvement. The results clearly demonstrate that incorporating the engineered features dramatically improves forecasting accuracy for every model.

Table 1. Main experimental comparison

Model	MAE	RMSE	MAPE	SMAPE
ARIMA (s)	139,309.22	183,341.15	14.86%	13.82%
ARIMA (m)	-	-	-	-
ARIMA (imp)	-	-	-	-
LR (s)	140,310.66	204,528.74	15.47%	14.12%
LR (m)	126,737.92	195,389.70	14.28%	12.67%
LR (imp)	9.67% ↑	4.47% ↑	7.69% ↑	10.27% ↑
Ridge (s)	141,628.36	202,607.80	15.47%	14.26%
Ridge (m)	124,821.73	191,963.61	14.10%	12.50%
Ridge (imp)	11.87% ↑	5.25% ↑	8.86% ↑	12.34% ↑
SVR (s)	142,640.00	211,313.30	15.97%	14.13%
SVR (m)	119,894.7775	188,495.4451	13.60%	12.13%
SVR (imp)	15.95% ↑	10.80% ↑	14.84% ↑	14.15% ↑
BPNN (s)	181,022.68	208,891.05	17.35%	17.79%
BPNN (m)	156,220.52	195,448.91	16.33%	14.76%

Table 1. Continued

BPNN (imp)	13.70% ↑	6.43% ↑	5.88% ↑	17.03% ↑
RNN (s)	172,581.55	237,411.81	19.44%	16.68%
RNN (m)	141,252.81	182,479.97	16.03%	14.49%
RNN (imp)	18.15% ↑	23.14% ↑	17.54% ↑	13.13% ↑
GRU (s)	169,773.67	239,572.11	19.32%	16.38%
GRU (m)	134,096.20	172,074.84	14.87%	13.90%
GRU (imp)	21.01% ↑	28.17% ↑	23.03% ↑	15.14% ↑
LSTM (s)	170,526.78	248,719.94	19.82%	16.41%
LSTM (m)	162,288.89	190,878.64	17.62%	16.01%
LSTM (imp)	4.83% ↑	23.26% ↑	11.10% ↑	2.44% ↑
RF (s)	155,752.92	193,787.83	16.49%	15.58%
RF (m)	140,680.65	189,696.66	15.90%	13.83%
RF (imp)	9.68% ↑	2.11% ↑	3.58% ↑	11.23% ↑
LightGBM (s)	162,507.45	190,534.88	16.45%	15.97%
LightGBM (m)	90,535.92	103,290.71	9.51%	9.10%
LightGBM (imp)	44.29% ↑	45.79% ↑	42.19% ↑	43.02% ↑
CatBoost (s)	160,280.56	188,778.83	16.41%	15.87%
CatBoost (m)	96,160.94	119,699.12	10.64%	9.89%
CatBoost (imp)	40.00% ↑	36.59% ↑	35.16% ↑	37.68% ↑
XGBoost (s)	164,247.68	195,266.24	16.74%	16.30%
XGBoost (m)	61,169.18	71,972.87	6.09%	6.00%
XGBoost (imp)	62.76% ↑	63.14% ↑	63.62% ↑	63.19% ↑

Note: ARIMA (m) results are not reported because the model cannot directly incorporate multiple features without extension. The improvement percentages are calculated as $\frac{V_s - V_m}{V_s} \times 100\%$.

XGBoost with the full feature set (m) achieves the lowest errors among all models: MAE = 61,169, RMSE = 71,973, MAPE = 6.09%, and SMAPE = 6.00%. Compared to its univariate version (s), the improvement is substantial, with reductions of 62.76% in MAE, 63.14% in RMSE, 63.62% in MAPE, and 63.19% in SMAPE. LightGBM and CatBoost also perform well, with MAPEs of 9.51% and 10.64% respectively, but still lag behind XGBoost. Deep learning models (RNN, GRU, LSTM) show moderate improvements but cannot match the tree-based ensembles, likely due to the limited sample size (60 training months). Traditional linear models (LR, Ridge) and SVR produce MAPEs around 13–14%, while ARIMA without exogenous variables gives 14.86%.

Figure 2 visualizes the four evaluation metrics across all models in a bar-chart format, highlighting the dominance of XGBoost. Figure 3 presents the time-series true-vs-predicted plots for each model. Taken together, these results demonstrate that incorporating the engineered features consistently improves forecasting accuracy across all models, with XGBoost achieving the lowest errors and the closest fit to actual sales. Specifically, XGBoost tracks both the upward trend and the monthly fluctuations most accurately, especially capturing the sharp drop in December 2024 and the subsequent recovery.

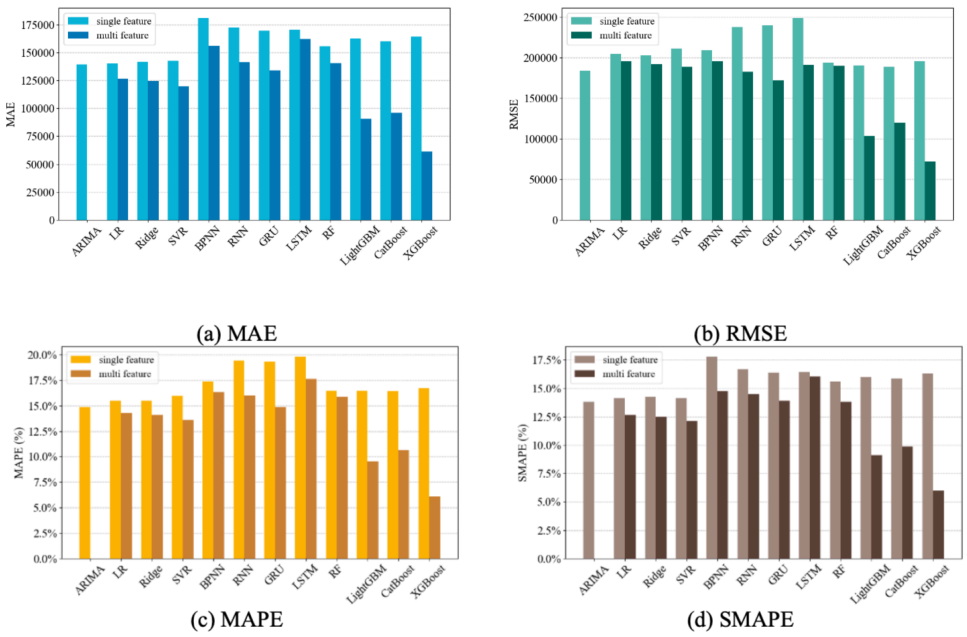
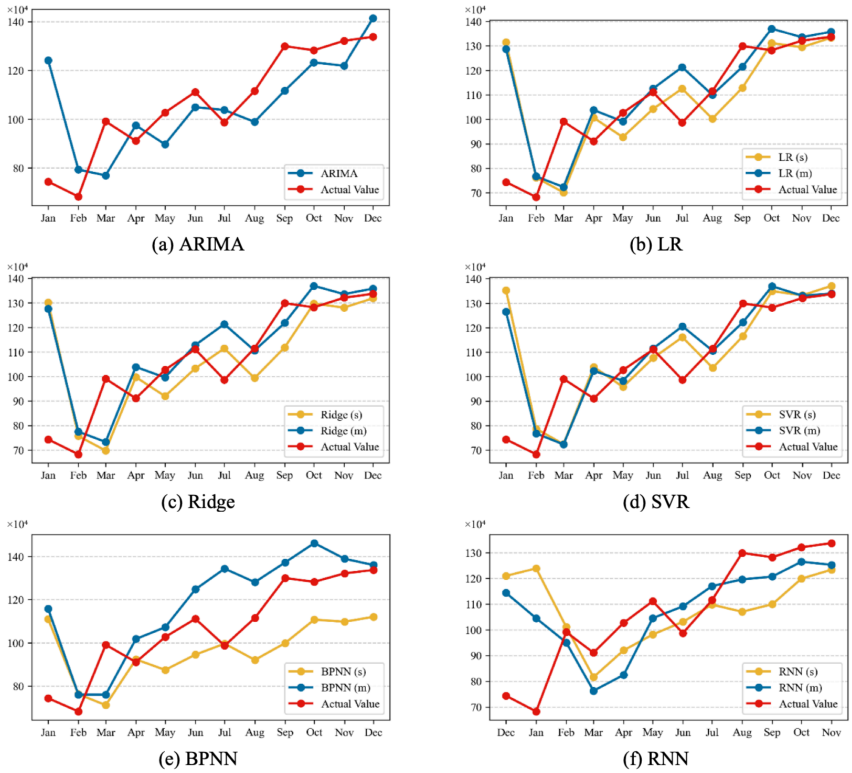


Figure 2. Comparison of MAE, RMSE, MAPE, and SMAPE across models



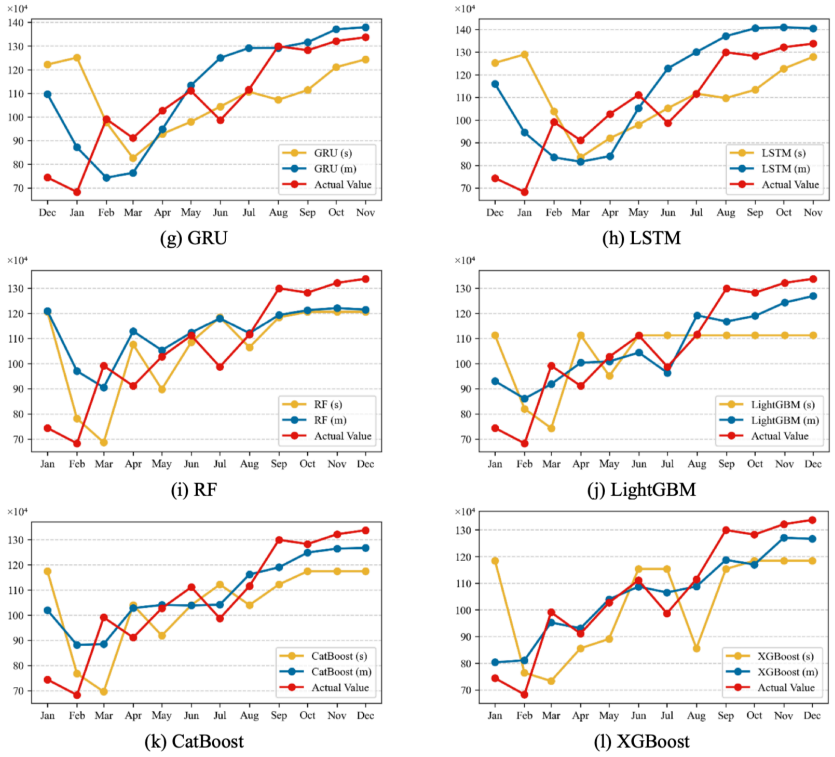


Figure 3. True versus predicted sales across models

4.2. Ablation study

To quantify the individual contributions of the station factor and the policy factor, we conduct an ablation experiment on the XGBoost model. We train three additional variants by removing different feature groups, resulting in four configurations in total: (1) the full model with all features (Ours), (2) the model without the station factor (w/o station), (3) the model without the policy factor (w/o policy), and (4) the model without both auxiliary features (w/o station & policy), which retains only the lagged sales and time encoding.

The evaluation results are shown in Table 2. Removing both auxiliary features increases MAPE from 6.00% to 9.13%, indicating that the combined effect of infrastructure and policy factors is substantial. Removing only the policy factor raises MAPE to 6.56%, while removing only the station factor raises MAPE to 6.93%. This suggests that both factors make significant contributions to forecasting accuracy, with the station factor having a slightly larger impact in terms of sales. Figure 4 visually confirms that the full model tracks the actual sales most closely across the test period, further underscoring the value of integrating both proposed features.

Table 2. Ablation study results

Model	MAE	RMSE	MAPE	SMAPE
w/o station	74,583.33	92,392.98	6.93%	7.26%
w/o policy	62,395.45	78,902.74	6.56%	6.30%
w/o station & policy	95,870.70	112,150.10	9.13%	9.60%
Ours	61,169.18	71,972.87	6.09%	6.00%

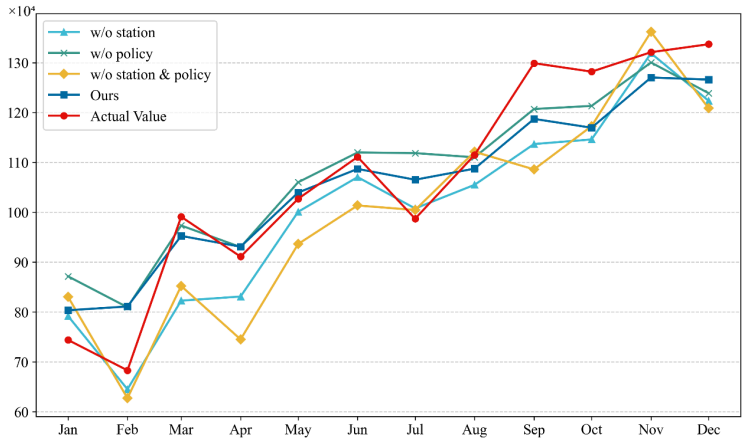


Figure 4. True and predicted sales for the four ablation configurations

4.3. Interpretability analysis

This section presents an interpretability analysis of the proposed NEV sales forecasting model using the SHAP method to quantitatively assess the importance of input features. SHAP is rooted in cooperative game theory and reveals the role of each feature in the prediction process by calculating its marginal contribution to the model output [23].

Figure 5 presents the SHAP summary plot based on the constructed features, which visualizes the contributions of different feature values to the model output. As shown in the figure, sales exerts the most significant effect on the prediction results, with its SHAP values predominantly concentrated in the positive range and higher feature values shifting further to the right. This indicates that higher sales in the previous month reliably lead to higher sales in the current month, reflecting a strong persistence effect. The station factor closely follows, displaying consistently positive SHAP values in the high-value region, suggesting that improved charging infrastructure contributes positively to sales. This finding is consistent with consumer behavior, where better charging accessibility and favorable cost conditions encourage NEV adoption, highlighting infrastructure as a key external driver. The time-related features exhibit more dispersed SHAP value distributions with both positive and negative contributions. This suggests that seasonality affects sales in a non-monotonic manner, with certain periods promoting sales while others suppress them. Together, these features effectively capture the cyclical pattern of NEV sales over time. The policy factor shows SHAP values concentrated on the negative side with a relatively small magnitude, indicating a weak but consistently negative contribution. This is consistent with the variable design, as the policy variable is dominated by restrictive measures in the sample, most of which exert a negative influence on NEV sales.

Overall, the SHAP analysis confirms that past sales and infrastructure factors are the dominant drivers of NEV sales, while the policy factor plays a secondary but directionally consistent role under the current specification. These findings are well aligned with domain knowledge and further support the validity of the feature construction in the proposed model.

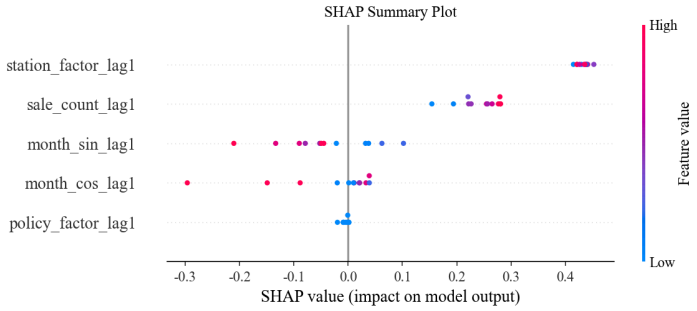


Figure 5. SHAP summary plot for the XGBoost model

5. Conclusion

5.1. Main findings

This study proposes two key features for NEV sales forecasting, namely a station factor that integrates charging stations, battery swap stations, and gasoline prices, and a time-varying policy factor that encodes major policy shocks. To validate the effectiveness of these features, a systematic comparison of twelve models is conducted under both univariate and multivariate settings. The results demonstrate that incorporating the proposed features consistently improves forecasting accuracy across all models. Among them, XGBoost with the full feature set achieves the best performance, with a MAE of 61,169, RMSE of 71,972, MAPE of 6.09%, and SMAPE of 6.00%, representing improvements of 62.76%, 63.14%, 63.62%, and 63.19% over the univariate baseline, respectively. Furthermore, ablation experiments confirm that both factors contribute independently, with the station factor having a slightly larger effect, and SHAP analysis further validates the interpretable roles of these features in driving NEV sales.

5.2. Limitations and future directions

Despite these achievements, the study has limitations. The sample size is modest (72 months), which may limit the performance of deep learning models. The policy factor uses a coarse three-level coding, future work could assign intensity weights or use natural language processing to quantify policy sentiment. Additionally, the framework currently focuses on national aggregate sales; extending it to provincial or city-level data would offer more granular insights. NEVertheless, the proposed model is ready for practical deployment and can serve as a benchmark for future research on NEV sales forecasting.

Funding project

This work was supported by the Science and Technology Research Project of Jiangxi Provincial Department of Education (Grant No. GJJ2200630).

References

[1] Yuan, X., Liu, X., & Zuo, J. (2015). The development of new energy vehicles for a sustainable future: A review. *Renewable and Sustainable Energy Reviews*, 42, 298-305.

[2] Han, G., & Li, Z. (2026). A structured review of electric vehicle sales research: Multi-level driving factors and forecasting pathways over the past decade. *World Electric Vehicle Journal*, *17*(3), 122.

[3] Wang, L., Fu, Z. L., Guo, W., Liang, R. Y., & Shao, H. Y. (2020). What influences sales market of new energy vehicles in China? Empirical study based on survey of consumers' purchase reasons. *Energy Policy*, *142*, 111484.

[4] Qi, Z. (2025). Analysis of influencing factors on China's new energy vehicle sales and forecasting with ARIMA model. In *SHS Web of Conferences* (pp. 03016). EDP Sciences.

[5] Qu, F., Wang, Y. T., Hou, W. H., Zhou, X. Y., Wang, X. K., Li, J. B., & Wang, J. Q. (2022). Forecasting of automobile sales based on support vector regression optimized by the grey wolf optimizer algorithm. *Mathematics*, *10*(13), 2234.

[6] Si, Q., Li, Y., Sun, J., Yu, X., Li, J., Wang, J., Dong, X., & Guo, K. (2024). New energy vehicle forecasting based on gray forecasting and random forest. In *International Conference on Smart Applications and Sustainability in the Artificial Intelligence of Things* (pp. 634-646). Cham: Springer Nature Switzerland.

[7] Wang, X., Shen, Z., & Hua, J. (2024). A study of new energy vehicles in China using random forest and ARIMA time series forecasting models. In *2024 IEEE 2nd International Conference on Image Processing and Computer Applications* (ICIPCA) (pp. 527-533).

[8] Zhang, B., Yang, S., Liu, C., & Xie, Z. (2024). Market sales forecast of PHEV based on ARIMA and XGBoost combined model. In *Proceedings of the 2024 5th International Conference on Big Data Economy and Information Management* (pp. 533-538).

[9] Li, B., Lv, X., & Chen, J. (2024). Demand and supply gap analysis of Chinese new energy vehicle charging infrastructure: Based on CNN-LSTM prediction model. *Renewable Energy*, *220*, 119618.

[10] Li, X. (2025). A study on monthly sales forecasting of new energy vehicles in urban areas using the WOA-BiGRU model. *PloS One*, *20*(4), e0320962.

[11] Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control.* John Wiley & Sons.

[12] Su, X., Yan, X., & Tsai, C. L. (2012). Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(3), 275-294.

[13] McDonald, G. C. (2009). Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(1), 93-100.

[14] Awad, M., & Khanna, R. (2015). Support vector regression. In *Efficient learning machines: Theories, concepts, and applications for engineers and system designers* (pp. 67-80). Berkeley, CA: Apress.

[15] Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

[16] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*.

[17] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, *31*.

[18] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

[19] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533-536.

[20] Medsker, L. R., & Jain, L. (2001). Recurrent neural networks. *Design and Applications*, *5*(64-67), 2.

[21] Cho, K., Van Merriënboer, B., Gulçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP) (pp. 1724-1734).

[22] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780.

[23] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4768–4777.