Proceedings of the 10th International Conference on Economic Management and Green Development
DOI: 10.54254/2754-1169/2026.36121

Machine Learning-Based Forecasting of Capital Market Sentiment in China's New Energy Vehicle Industry

Zipeng Deng

College of Art & Sciences, American University, Washington, USA
zd2513a@american.edu

Abstract. This paper takes listed companies in China's new energy vehicle industry chain as the research subject and constructs an experimental model for predicting industry capital market prosperity. Compared to existing research mainly focusing on predicting the direction of individual stock prices, identifying fluctuations in clean energy assets, and optimizing deep learning models, research on the overall capital market scenario of the new energy vehicle industry chain is still relatively limited. Research is conducted on representative companies in the three sub industries of vehicle, battery, and lithium resources to establish an equal weight index for the new energy vehicle industry. The industry index returns over the next 20 trading days are divided into three categories: Growth, Neutral, and Decline. The experiment used yfinance to obtain daily trading data from January 2018 to April 2026. In terms of models, this article compares the predictive performance of Majority Baseline, Logistic Regression, Random Forest, and LSTM. The results showed that logistic regression performed the best under the current feature system, indicating that there is a certain degree of predictability in changes in industry prosperity, but complex models were not significantly better than simple interpretable models.

Keywords: New energy vehicles, industry prosperity, capital market, random forest, LSTM

1. Introduction

The new energy vehicle industry has the characteristics of a long industrial chain, fast technological iteration, strong policy influence, and high attention from the capital market. If only predicting the short-term rise and fall of a certain stock, the research conclusions are easily affected by individual stock events, market sentiment, and short-term trading noise, and the academic significance is relatively limited. Therefore, this article will adjust the research object from "short-term trend prediction of individual stocks" to "industry capital market scenario sentiment prediction".

This article mainly answers three questions: firstly, whether the new energy vehicle industry as a whole is showing signs of growth, volatility, or decline in the capital market; Secondly, whether there is structural differentiation among the three sub industries of vehicles, batteries, and lithium resources; Thirdly, what are the differences in performance between traditional machine learning models and deep learning models in predicting industry prosperity.

The existing research on capital market forecasting mainly focuses on predicting individual stock returns, price directions, and industry sector volatility. Specifically, Liu et al. predicted changes in

DOI: 10.54254/2754-1169/2026.36121

stock prices of new energy vehicle companies based on multi-source data and mixed attention structures [1]. Sadorsky studied the relationship between clean energy stock prices, market volatility, and economic policy uncertainty; Ghosh et al. directly used LSTM and random forest to predict the direction of stock price changes [2]. For new energy vehicle companies, research has begun to combine multi-source data and mixed attention structures to predict changes in company stock prices, indicating that new energy vehicle related targets have a certain degree of predictability in the capital market [1]. In the broader study of clean energy assets, machine learning models have also been used to identify the relationship between market volatility, policy uncertainty, and stock prices. Research has shown that external risk variables have a significant impact on the pricing of the clean energy sector.

At the methodological level, existing research widely uses LSTM, Random Forest, and their combination models to predict the direction of stock prices and changes in returns. Specifically, previous studies have used LSTM and random forest for stock price direction prediction, integrated LSTM and random forest frameworks, and directly used LSTM to predict changes in stock returns [3-5]. There are also studies that improve stock prediction performance through symbolic genetic programming, multi feature screening, and improved recurrent neural network structures [6, 7]. In addition, some studies further introduce textual emotions, financial semantic information, and multi-stage machine learning frameworks to enhance the model's ability to recognize changes in market expectations [8, 9]. From the perspective of literature review, research on stock price prediction has shown a continuous evolution from traditional statistical models to machine learning and deep learning [10]. However, research on the overall capital market atmosphere of the new energy vehicle industry chain is still relatively insufficient [1-3].

This article obtains daily trading data of sample companies and market indices through the yfinance library in Python. The sample time range is from January 2018 to April 2026, and the data includes opening price, highest price, lowest price, closing price, and trading volume.

The sample company is divided into three sub industries based on the position of the new energy vehicle industry chain. Vehicle companies include BYD, Xiaomi, Great Wall Motors, and SAIC Group; Battery companies include CATL and EVE Energy; Lithium resource enterprises include Ganfeng Lithium and Tianqi Lithium. This classification can avoid treating the new energy vehicle industry as a completely synchronized whole, which helps analyze whether there is structural differentiation within the industry chain. It should be noted that Xiaomi was included in the sample because it entered the field of intelligent electric vehicles, which can represent the capital market expectation of technology companies crossing the new energy vehicle industry chain; However, Xiaomi's historical stock price does not fully represent the performance of its automotive business, so caution should be exercised when interpreting the results. This article also includes CSI300, Tesla, and Nasdaq as external market variables. Among them, CSI300 represents the overall stock market environment in China, Tesla represents the market performance of global new energy vehicle leaders, and Nasdaq represents the global technology growth stock market environment, which is consistent with the existing research that emphasizes the combined effects of multi-source market characteristics and external risk variables [1, 2].

2. Research design

2.1. Experimental steps

This article presents a five-step experimental design for predicting the prosperity of the new energy vehicle industry. Firstly, conduct a comparison of the performance of individual stock capital

DOI: 10.54254/2754-1169/2026.36121

markets, calculate the cumulative return rate, return rate in June and December, annualized volatility, and maximum drawdown index of each enterprise, and analyze the phenomenon of strong and weak differentiation of listed companies in the industrial chain. Secondly, establish an industry equity index. Standardize the initial benchmark of individual stocks and calculate the average values of the vehicle, battery, and lithium resource sub industries to obtain three sub industry indices. Then, construct a new energy vehicle industry index based on the average values of all samples to observe the overall capital market trend of the sector. Thirdly, establish a three-category business forecast label that does not target the next day's rise or fall of a single stock. Instead, use the industry index's future 20 day return rate as the forecast target: a return rate above 3 is considered a growth range, a return rate below 3 is considered a decline range, and a return rate between the two is considered a stable range, in order to characterize short-term industry business expectations. Fourth, construct predictive features from multiple dimensions and extract variables from four major dimensions: industry trends, market risks, sub industry structure, and external markets. Trend categories include industry returns on 5/20/60 and deviation from index relative moving averages; The risk indicator uses a 20 day volatility to measure market divergence; Comparing the relative strengths and weaknesses of the three major sub industries with the overall industry using structural variables reflects the internal differentiation of the industrial chain; External features are selected from the Shanghai and Shenzhen 300, Tesla, and Nasdaq 20 day returns to characterize the external shocks of the A-share market, overseas leaders, and global growth sectors, respectively. Fifth, time series partitioning of the dataset, with data before 2024 as the training set and 2024 and beyond as the testing set, abandoning random sampling, avoiding future data leakage issues, and fitting in with real market prediction scenarios.

2.2. Model setting and evaluation indicators

This article compares four types of models: Majority Baseline, Logistic Regression, Random Forest, and LSTM. The Majority Baseline serves as the simplest control model to uniformly predict the largest number of business categories in the training sample and is used to determine whether the complex model has incremental prediction ability. Logistic Regression, as a linear benchmark model, has strong interpretability; Random Forest is used to capture non-linear relationships and interactions between variables, and output feature importance; LSTM is used to process continuous time series features and test whether changes in industry characteristics over a period of time can help predict future business conditions. Previous studies have extensively applied LSTM and Random Forest to tasks such as predicting stock price direction and returns [3-5].

The reasons for choosing these four types of models are as follows. The Majority Baseline is used to provide the simplest benchmark for comparison. If a machine learning model cannot exceed this baseline, it indicates that the model has no actual predictive value. Logistic Regression has a simple structure and strong interpretability, making it suitable as a linear benchmark model for predicting industry prosperity. Random Forest can capture non-linear relationships and output feature importance, making it suitable for analyzing which variables are more critical for predicting industry prosperity. LSTM is a time series deep learning model that can test whether continuous feature changes over a period of time providing more predictive information than traditional machine learning models. Previous studies have also shown that symbolic genetic programming, multi feature screening, text emotion, and multi-stage machine learning frameworks can help improve the performance of stock prediction models [6-8].

The evaluation indicators include Accuracy, Macro Precision, Macro Recall, and Macro F1. The distribution of the number of three types of prosperity labels in the sample is uneven, and the

DOI: 10.54254/2754-1169/2026.36121

weighted F1 is affected by the interference of the majority class. Therefore, Macro F1 is used to estimate the accuracy of each class prediction. Macro F1 can more fairly measure the predictive ability of the model for different categories, avoiding the model being biased towards the category with the largest sample size.

3. Experimental results and analysis

From the data construction and descriptive analysis results, the experiment successfully obtained daily transaction data of 8 sample companies from 2018 to April 2026. There is a significant difference in the 12-month interval returns of each target, which can prove the stable market differentiation characteristics of the new energy upstream, midstream and downstream industry chain after excluding individual stock events.

This can be supported by Figure 1. At the same time, the overall index of the industry is not completely consistent with the three sub industry indices of vehicle, battery, and lithium resources. Figure 2 further illustrates that the new energy vehicle industry is not a completely synchronized whole, but there are obvious differences in the industrial chain structure.

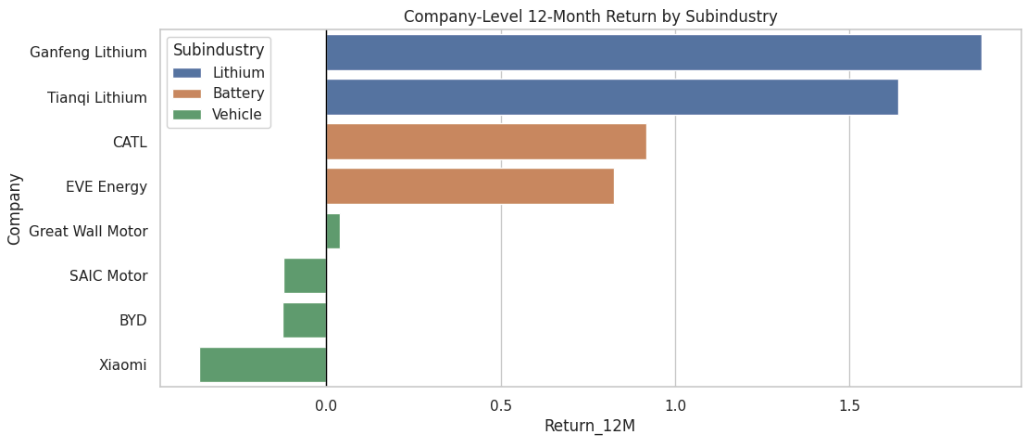

Figure 1. Comparison of the company's returns in the past 12 months

Picture credit: Original

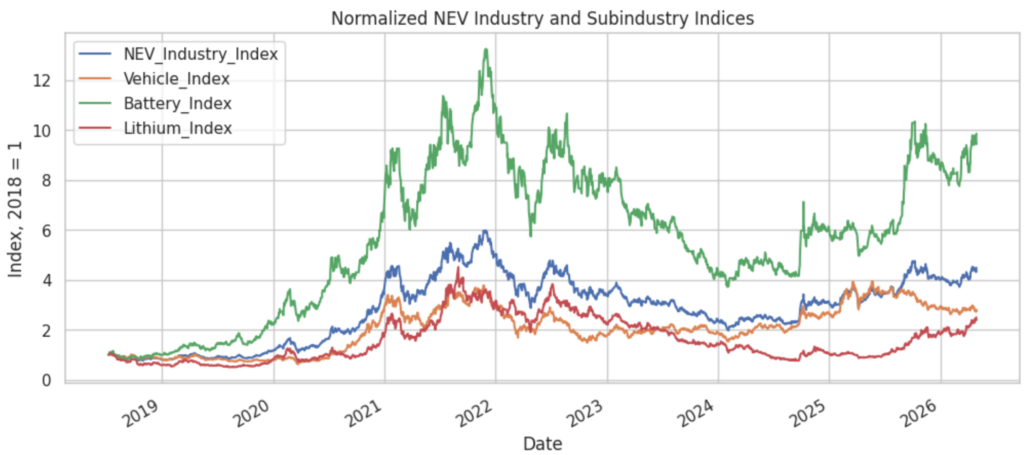

Figure 2. Trend of index for new energy vehicle industry and sub industry

Picture credit: Original

DOI: 10.54254/2754-1169/2026.36121

From the model prediction results, logistic regression performs the best. The experimental results show that the logistic regression comprehensive index is the best among the four models, but the macro F1 is only 0.3121. The overall ability to distinguish three categories of predictions is weak, and there is a large amount of unpredictable noise in the short-term prosperity of the industry. Table 1 summarizes the predictive indicators of four types of models, while Figure 3 visually illustrates the differences between each model on Macro F1. The performance of Random Forest and LSTM is slightly lower than that of Logistic Regression; From the Macro F1 perspective, both are superior to the Majority Baseline, but the advantage is not significant from the Accuracy perspective. This indicates that machine learning models can capture signals of changes in industry prosperity, but their overall predictive ability is still limited.

Table 1. Summary of model prediction results

Model	Accuracy	Macro Precision	Macro Recall	Macro F1
Majority Baseline	0.2810	0.0937	0.3333	0.1462
Logistic Regression	0.4112	0.4035	0.4019	0.3121
Random Forest	0.3306	0.2320	0.3552	0.2734
LSTM	0.3368	0.2919	0.3543	0.2779

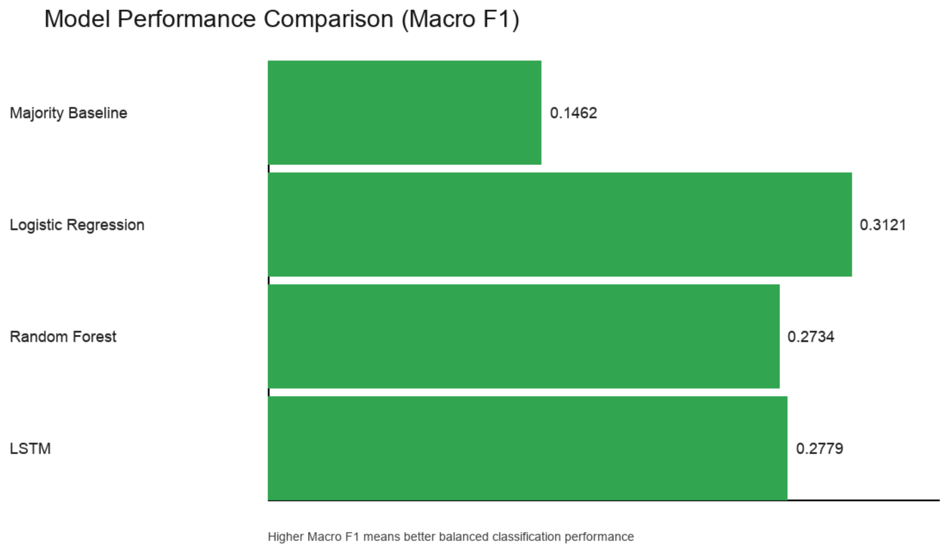

Figure 3. Comparison of macro F1 models

Picture credit: Original

From Macro F1, both Random Forest and LSTM outperform the Majority Baseline, but the improvement is limited; From the perspective of accuracy, the advantages of the two over the baseline are not significant. Therefore, this article cannot simply assume that complex models significantly improve predictive ability but should be understood as models that only capture a portion of industry prosperity signals. This result indicates that under the current sample and feature settings, industry prosperity prediction relies more on stable trends, volatility, and relative strength information, and complex models have not gained significant advantages due to their more complex

DOI: 10.54254/2754-1169/2026.36121

structures. Especially since LSTM did not significantly surpass traditional models, it indicates that relying solely on historical price sequences is difficult to fully explain changes in industry prosperity. From the perspective of feature importance results, the random forest feature importance shows that the lithium sector has the highest relative strength weight, and the industry logic is that the upstream lithium carbonate price cycle directly transmits the cost and market profit expectations of automotive companies. Figure 4 visualizes this result, and the industry prosperity forecast is driven by three types of characteristics: internal sector structure, self-volatility trend, and overseas markets. The overall weight of the differentiation index of upstream, midstream, and downstream sectors is higher than that of Nasdaq and Tesla Overseas Index, which is consistent with the conclusion of existing research emphasizing the importance of multi-source features and external risk variables [1, 9, 10].

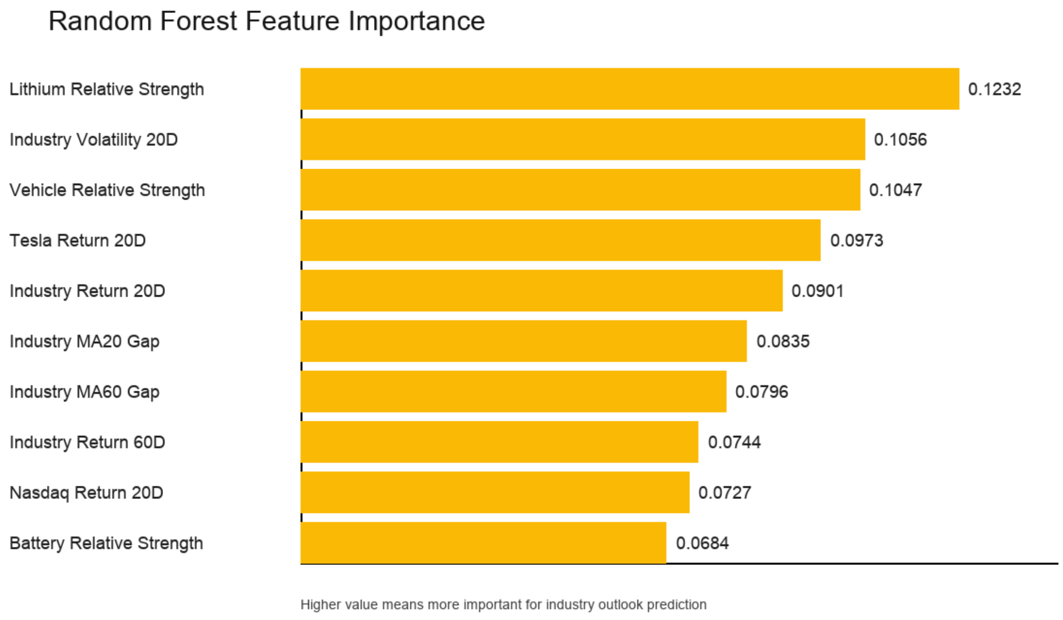

Figure 4. Random forest feature importance

Picture credit: Original

4. Conclusion

The market of new energy vehicles in the upstream, midstream, and downstream sectors has long been structurally differentiated, with only the trend of various targets tending to synchronize during the period of market volatility and sideways movement. There may be differences in the performance of the three sub industries of vehicles, batteries, and lithium resources in the capital market, among which the relative strength of the lithium resources sector has a high impact on the industry's prosperity forecast. This indicates that the capital market prospects of the new energy vehicle industry are not only determined by vehicle sales expectations but also influenced by upstream raw material cycles and the global new energy market environment.

The model results show that Logistic Regression performs the best in prediction. In the experimental setting where only single day cross-sectional features are used and time series samples are limited, simple linear logistic regression outperforms random forest and LSTM deep learning

DOI: 10.54254/2754-1169/2026.36121

models in prediction performance. Although Random Forest and LSTM can also capture some signals, they do not significantly surpass Logistic Regression; From the experimental results, it can be seen that the improvement of complex models compared to simple baselines is mainly reflected in Macro F1, and the degree of improvement is limited. This indicates that there is a certain degree of predictability in the capital market prosperity of the new energy vehicle industry, but the difficulty of prediction is high, influenced by factors such as market sentiment, policy changes, macro environment, and external technology stock trends.

Overall, this study has established a relatively complete framework for analyzing the prosperity of the capital market in the new energy vehicle industry through company level performance analysis, sub industry index construction, industry prosperity label design, and machine learning model prediction. It should be noted that this study predicts the atmosphere of the capital market scenario, rather than the actual sales, production, or profits of the new energy vehicle industry. Therefore, the model results should be understood as market expectation signals, rather than deterministic judgments on the true operating conditions of the industry. Firstly, this article uses stock prices and market index data, so the object of the prediction is the capital market scenario atmosphere, rather than the actual sales, production or profits of the new energy vehicle industry. Capital market prices reflect expectations in advance and are also influenced by investor sentiment and macro liquidity. Secondly, this article uses individual stock arithmetic weighting to construct sector indices, which have the advantage of simple calculation, but may underestimate the market driving effect of billion-dollar leading targets such as BYD and CATL. In the future, market value weighted indices can be used to optimize feature inputs. Subsequent research can try using market capitalization weighted index or trading volume weighted index. Thirdly, the LSTM model is highly sensitive to parameters, LSTM is highly sensitive to time series length, and hyperparameter combinations. In this paper, the time series sample period is relatively short; In the future, it is possible to expand the market for many years, build multiple days of continuous time series input, and optimize network parameters with grid search. Therefore, deep learning models may not be able to fully leverage their advantages. In the future, longer time series, richer macro variables, policy text indicators, or new energy vehicle sales data can be added to enhance the explanatory power of the model. Fourthly, this article sets three types of labels: Growth, Neutral, and Decline, with thresholds of plus or minus 3% of the next 20 trading days' returns. This threshold has a certain degree of subjectivity, and sensitivity analysis can be used to compare the performance of models under different thresholds in the future.

References

[1] Liu, X., Wu, Y., Luo, M., & Chen, Z. (2024). Stock price prediction for new energy vehicle companies based on multi-source data and hybrid attention structure. *Expert Systems with Applications, 255*, 124787.

[2] Sadorsky, P. (2022). Using machine learning to predict clean energy stock prices: How important are market volatility and economic policy uncertainty? *Journal of Climate Finance, 1*, 100002.

[3] Ghosh, P., Neufeld, A., & Sahoo, J. K. (2022). Forecasting directional movements of stock prices for intraday trading using LSTM and random forests. *Finance Research Letters, 46*, 102280.

[4] Park, H. J., Kim, Y., & Kim, H. Y. (2022). Stock market forecasting using a multi-task approach integrating long short-term memory and the random forest framework. *Applied Soft Computing, 114*, 108106.

[5] Qiao, R., Chen, W., & Qiao, Y. (2022). Prediction of stock return by LSTM neural network. *Applied Artificial Intelligence, 36*(1), 2151159.

[6] Li, Q., Kamaruddin, N., Yuhaniz, S. S., & Al-Jaifi, H. A. A. (2024). Forecasting stock prices changes using long-short term memory neural network with symbolic genetic programming. *Scientific Reports, 14*, 422.

[7] Chen, X., Cao, L., Cao, Z., & Zhang, H. (2024). A multi-feature stock price prediction model based on multi-feature calculation, LASSO feature selection, and Ca-LSTM network. *Connection Science, 36*(1), 2286188.

DOI: 10.54254/2754-1169/2026.36121

[8] Kim, J., Kim, H.-S., & Choi, S.-Y. (2023). Forecasting the S&P 500 Index using mathematical-based sentiment analysis and deep learning models: A FinBERT transformer model and LSTM. *Axioms, 12*(9), 835.

[9] Abolmakarem, S., Abdi, F., Khalili-Damghani, K., & Didehkhani, H. (2024). A multi-stage machine learning approach for stock price prediction: Engineered and derivative indices. *Intelligent Systems with Applications, 22*, 200449.

[10] Vuong, P. H., Phu, L. H., Nguyen, T. H. V., Duy, L. N., Bao, P. T., & Trinh, T. D. (2024). A bibliometric literature review of stock price forecastin: From statistical model to deep learning approach. *Science Progress, 107*(1).