ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 16, No. 2, 2021, *pp*.108-117

Analysis of feature selection for stock price prediction with LSTM: A case study on China's new energy leading stocks

Wanbao Zhou

Nanjing University of Information Science and Technology, Nanjing, China

(Received June 21, 2021, accepted September 07, 2021)

Abstract: Stock price prediction has always been the focus of investors' attention in the stock market. In recent years, deep learning technology has been widely used in this field. In the era of big data, feature selection is a necessary part of data preprocessing. Feature selection is a data dimensionality reduction technology, and its main purpose is to select the relevant features that are most beneficial to the algorithm from the original data, reduce the dimensionality of the data and the difficulty of learning tasks, and improve the efficiency of the model. This paper has performed analysis of input feature selection with three feature selection methods: Multiple linear regression analysis, Correlation matrix heatmap, Feature importance. Plus the original features set, four different input features sets were provided for predicting stock price of ten China's new energy leading stocks with LSTM. From the conducted experiments, it is found that after using the feature selection method, the prediction results of all ten stocks are performed better than the prediction results under the original features.

Keywords: Stock Price Prediction, LSTM, Feature Selection, Multiple linear regression analysis, Correlation matrix heatmap, Feature importance

1. Introduction

With the rapid development of China's economy, finance has become the core force of the modern economy and an important core competitiveness of the country. The effective development of the financial market plays a vital role in promoting the development of China's overall economy. The stock market is an important part of the financial market. The stable development of the stock market has become an important prerequisite for the sustained and stable development of the Chinese economy.

Environmental protection is a basic national policy in China. With the deepening of economic reform and development, especially after joining the World Trade Organization, environmental protection has attracted more and more attention. Coal prices have risen sharply this year, and many provinces and cities have issued power rationing orders on companies. From the perspective of environmental protection, this aspect has been affected by the 30-year carbon peak and 60-year carbon neutrality of the "14th Five-Year Plan". On the one hand, it is also affected by energy scarcity. Environmental issues once again sounded the alarm. The transformation of the energy structure is the general trend, and the replacement of traditional energy by new energy is an irreversible process.

In recent years, the source and scale of stock market data have increased rapidly. Deep learning models have been introduced into many research scenarios in stock market forecast due to their excellent large-scale data processing capabilities [1]. Recurrent Neural Networks (RNNs) is a mainstream deep neural network used to sequence modeling, which solves the problem that traditional feedforward neural networks cannot handle variable-length sequences. It has received a great amount of attention due to their flexibility in capturing nonlinear relationships [2-4]. However, traditional RNNs suffer from the problem of vanishing gradients and thus have difficulty in capturing long-term dependencies [5]. Long Short-term memory units (LSTM) have overcome this limitation and achieved great success in many applications, such as sequence prediction [6].

For a particular learning algorithm, which feature is effective is unknown. Therefore, it is necessary to select relevant features that are beneficial to the learning algorithm from all the features. And in practical applications, the problem of dimensional disasters often occurs. If only some of the features are selected to construct the model, the running time of the learning algorithm can be greatly reduced, and the interpretability of the model can also be increased. First of all, the definition of feature selection was looking for the smallest feature subset for the model to be effective under ideal conditions [7]. Subsequently, the definition of feature selection evolved to try to find the smaller feature subset between the feature and the original data in the case of similar set distributions [8]. The ensemble learning idea was first proposed by Breiman from the statistics Bootstrap idea. This idea extended the classification advantages of the base learner and greatly promoted the research and development of classification models. After that, Breiman and Cutler first proposed the Bagging idea and random subspace combination method as an effective integrated learning classification prediction tool [9]. This method is continuously improved by combining the decision tree algorithm to form Random Forest, laying the foundation for the integrated learning algorithm. Random survival forest was proposed to effectively process high-dimensional data [10]. Extremely randomized trees were proposed to improve the overall analysis of variance and bias for Random Forest which does not include attribute selection [11].

This paper has performed an analysis of look-back period used with Long Short-term Memory (LSTM) for predicting stock prices. First, using the data of a stock to determine the parameters for model LSTM. Then, three feature selection methods as follows were used for obtaining new input features subset: Multiple linear regression analysis, Correlation matrix heatmap, Feature importance. Last, plus the original features set, four different input features sets were provided for predicting stock price. The main concern is the difference in prediction results under different input features. Ten China's new energy leading stocks are analyzed in this research work.

2. Long Short-term Memory Network

Long Short-term Memory Network (LSTM) is a type of Recurrent Neural Network (RNN). RNN has huge difficulties in dealing with long-term dependencies, because the calculation of connections between distant nodes will involve multiple multiplications of the Jacobian matrix, which will cause the vanishing gradient problem or exploding gradient problems. In order to solve this problem, researchers have proposed many solutions. Among them, the most successful and widely used is the gated RNN , and LSTM is the most famous kind of the gated RNN. Leaky units allow RNN to accumulate long-term connections between distant nodes by designing the weight coefficients between connections. And gated RNN generalizes this idea, allowing the coefficient to be changed at different times, and allowing the network to forget the current accumulation Information. LSTM is such a gated RNN, and its single node structure is shown in the Fig.1. The ingenuity of LSTM is that by increasing the input gate, forget gate and output gate, the weight of the self-loop is changed. In this way, when the model parameters are fixed, the integration scale at different moments can be dynamically changed, combining short-term memory with long-term memory which avoids the problem of the vanishing gradient problem or exploding gradient problems to a certain extent.

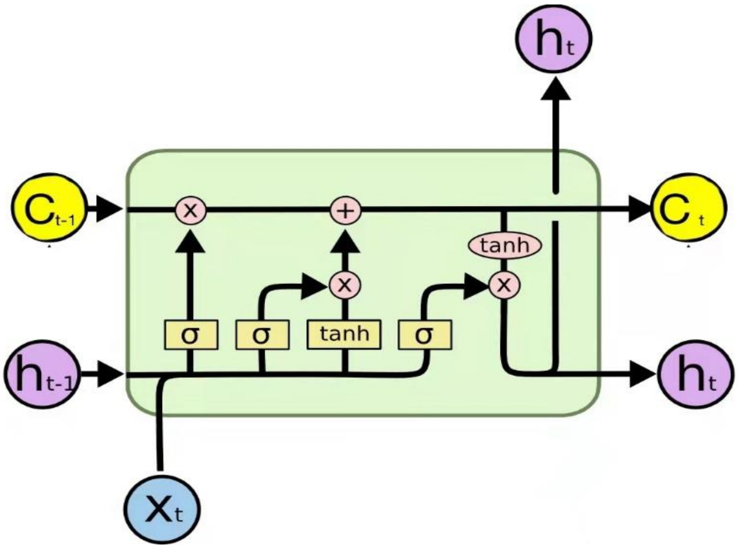


Fig.1. LSTM Architecture

DOI https://doi.org/2024-JICS-22368 | Generated on 2025-04-12 03:46:34

An LSTM is well-fitted to predict time-series data. Each LSTM unit has a memory cell with the state C_t at time t. Access to memory cell will be controlled by three sigmoid gates: forget gate f_t, input gate i_t and output gate o_t. Every LSTM cell computes new values of cell state and hidden state. The update of an LSTM unit can be summarized as follows:

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \tag{1}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{2}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{3}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{4}$$

$$h_t = o_t \odot \tanh(C_t) \tag{5}$$

where $[h_{t-1}, x_t]$ is a concatenation of the previous hidden state h_{t-1} and the current input x_t. W_f, W_i, W_o, W_C and b_f, b_i, b_o, b_C are parameters to learn. σ and $\odot$ are a logistic sigmoid function and an elementwise multiplication, respectively.

3. Feature Selection

3.1 Multiple linear regression analysis

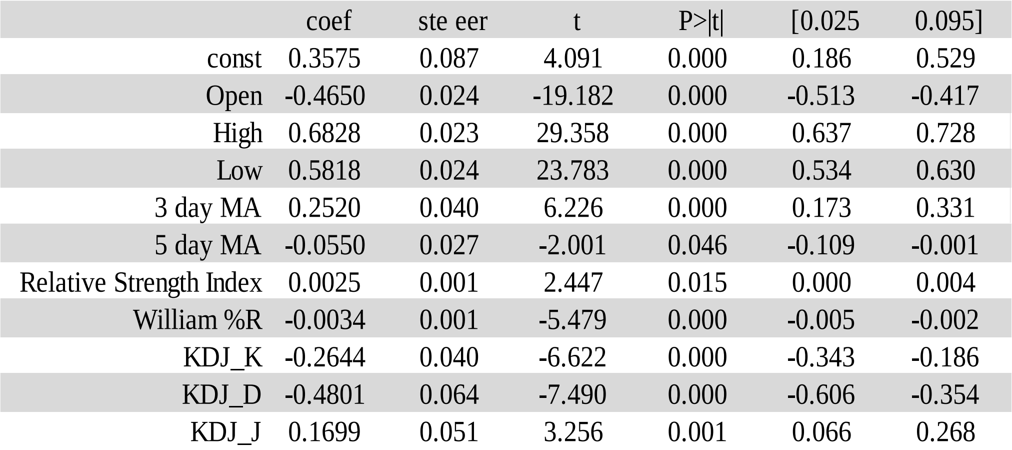
In statistics, regression analysis refers to a statistical analysis method that determines the quantitative relationship between two or more variables. In big data analysis, regression analysis is a predictive modeling technique that studies the relationship between the dependent variable (target) and the independent variable (predictor). This technique is commonly used in predictive analysis, time series models, and discovering causal relationships between variables.

Multiple linear regression is a statistical analysis method that studies the linear relationship between a dependent variable (quantitative) and multiple independent variables. First, performing a linear relationship test. The linear relationship test is to test whether the relationship between the dependent variable and multiple independent variables is significant, also known as the overall significance test. As long as there is a significant linear relationship between one independent variable and the dependent variable, the F test can pass, but this does not necessarily mean that the relationship between each independent variable and the dependent variable is significant. The regression coefficient test is a separate test for each regression coefficient. It is mainly used to test whether each independent variable has a significant influence on the dependent variable. If an independent variable fails the test, it means that the independent variable has no significant influence on the dependent variable, and it may not be necessary to put this independent variable into the regression model.

When building a regression model based on multiple independent variables, trying to introduce all the independent variables into the regression model may not be able to effectively explain the established model. If the collected independent variables can be filtered before the model was built, and those unnecessary independent variables can be removed. Not only will it be easier to build the model, but also it will make the model more maneuverable and easier to explain.

In this paper, using stepwise regression to filter out the subset of independent variables which is most relevant to the predictive variable closing price, and then plus the predictive variable closing price as new input features subset.

DOI https://doi.org/2024-JICS-22368 | Generated on 2025-04-12 03:46:34



	coef	ste eer	t	P>|t|	[0.025	0.095]
const	0.3575	0.087	4.091	0.000	0.186	0.529
Open	-0.4650	0.024	-19.182	0.000	-0.513	-0.417
High	0.6828	0.023	29.358	0.000	0.637	0.728
Low	0.5818	0.024	23.783	0.000	0.534	0.630
3 day MA	0.2520	0.040	6.226	0.000	0.173	0.331
5 day MA	-0.0550	0.027	-2.001	0.046	-0.109	-0.001
Relative Strength Index	0.0025	0.001	2.447	0.015	0.000	0.004
William %R	-0.0034	0.001	-5.479	0.000	-0.005	-0.002
KDJ_K	-0.2644	0.040	-6.622	0.000	-0.343	-0.186
KDJ_D	-0.4801	0.064	-7.490	0.000	-0.606	-0.354
KDJ_J	0.1699	0.051	3.256	0.001	0.066	0.268

Fig.2. The final result of Multiple linear regression analysis of Longji Shares

Taking Longji Shares for example, given significance level α=0.05, using t-statistics to test regression coefficients. We think the impact is significant if pass the t-statistic test. Then using stepwise regression to remove variables. Finally, we obtain variables which are significant to predictor variable closing price as shown in Fig.2. Plus the variable closing price, we obtain new input features subset which is selected by multiple linear regression analysis.

3.2 Correlation matrix heatmap

Correlation represents the relationship between the variable and the target variable. The correlation coefficient is a statistic calculated based on sample data to measure the strength of the linear relationship between two variables. In statistics, for a specific correlation coefficient, based on experience, the degree of correlation can be divided into the following situations: when the absolute value of the correlation coefficient is greater than or equal to 0.8, it can be regarded as highly correlated; when the absolute value of the correlation coefficient is greater than or equal to 0.5 and less than 0.8, it can be regarded as a moderate correlation; when the absolute value of the correlation coefficient is greater than or equal to 0.3 and less than 0.5, it can be regarded as a low correlation; when the absolute value of the correlation coefficient is less than 0.3, it indicates that the correlation degree is extremely weak between the two variables and can be regarded as irrelevant.

The heatmap makes it easy to identify which elements are most relevant to the target variable. The darker the color, the higher the correlation.

In this paper, selecting feature variables that are moderately correlated and highly correlated with the predictive variable closing price, and then plus the predictive variable closing price as new input features subset.

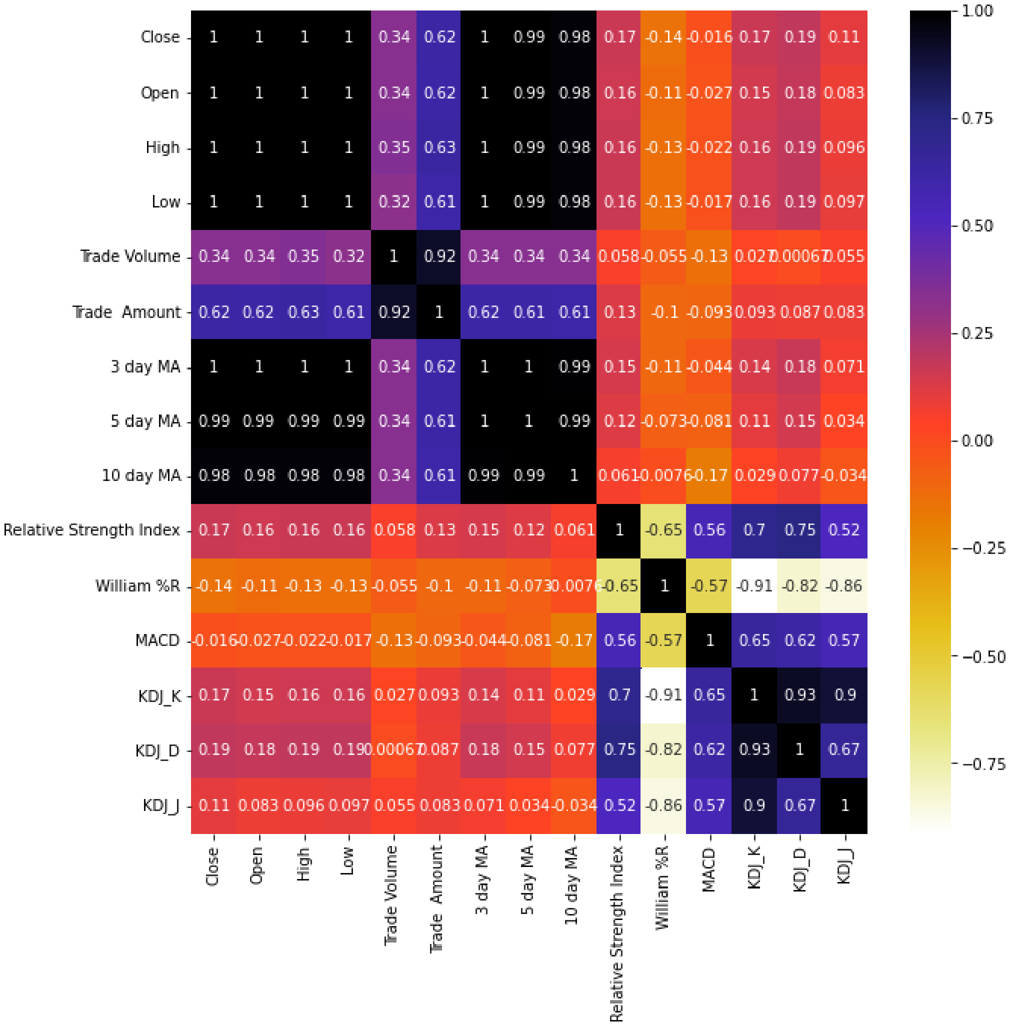


Fig.3. Correlation matrix heatmap of Longji Shares

Taking Longji Shares for example, we obtain new input features subset include Close, Open, High, Low, Trade Amount, 3 day MA, 5 day MA and 10 day MA as shown in Fig.3. using the feature selection method Correlation matrix heatmap.

3.3 Feature importance

Feature importance is to give a score to each feature in the data. The higher the score, the more important the feature is. By using the element importance attribute of the model, the element importance of each element of the dataset can be obtained.

Extremely Randomized Trees Classifier is an ensemble learning technology that aggregates the results of multiple decorrelated decision trees collected in the forest to output the classification results. Each decision tree of the extremely random tree is constructed from the original training samples. At each test node, each tree has a random sample, and there are multiple features in the sample. Each decision tree must select the best feature from these feature sets, and then use some mathematical indicators (generally the Gini index) to split the data. This random feature sample leads to the generation of multiple uncorrelated decision trees.

In the process of constructing the forest, for each feature, calculating the normalized total reduction of the mathematical index (such as the use of the Gini index) used to divide the feature decision. This value is called the importance of the Gini element. After the Gini importance is sorted in descending order, the top number of features can be selected as needed.

In this paper, selecting the nine most important features for variable close as new input features subset. Taking Longji Shares for example, we remove the variable Trade Volume, KDJ_D, Trade Amount, KDJ_K and Relative Strength Index which is represented by the lowest 5 values as shown in Fig.4. Other variables

DOI https://doi.org/2024-JICS-22368 | Generated on 2025-04-12 03:46:34

plus the predictive variable closing price can be used as the new input features subset.

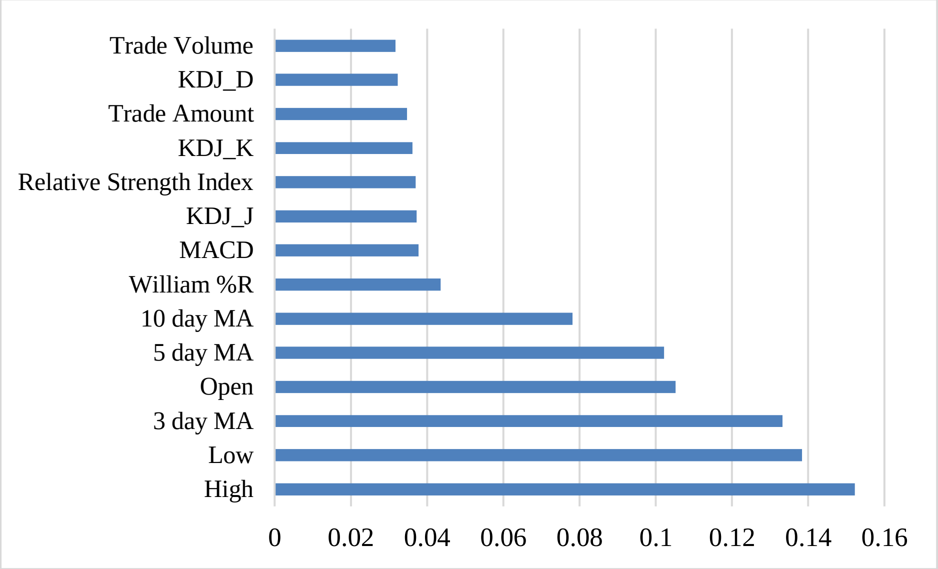


Fig.4. Feature importance of Longji Shares

4. Experiment

In this section, we first describe ten China's new energy leading stocks datasets for empirical studies. Then, we introduce the parameter settings of LSTM and the evaluation metrics. Finally, we compare prediction results of origin features against prediction results of features selected by three feature methods.

4.1 Datasets and Setup

Ten China's new energy leading stocks respectively are Longi Shares(601012) , Tianqi Lithium (002466), Desay Battery (000049) , Tinci Materials (002709) , Sungrow Power (300274) , Chint Electrics(601877)，BYD (002594), Energy Very Endure(300014), Inovance Technology(300124), Lead Intelligent (300450) , the value after the bracket is the code of this stock.

For each stock, historical data used in this research work is from Sep 1, 2016 to Sep 5, 2021. There are 15 variables in the original data set, which are “Close”, “Open”, “High”, “Low”, “Trade Volume”, “Trade Amount”, “3day MA”, “5day MA”, “10day MA”, “Relative Strength Index”, “William %R” , “MACD”, “KDJ_K”, “KDJ_D”, “KDJ_J”. The train/validation/test set split of the dataset is done at almost a 7:2:1 ratio. The data is pre-processed before using it for training and testing LSTM by using z-score normalization with the attributes of the training set. According to different feature selection methods, we select different feature subsets as input attribute and “Next Day’s Close Price” will be used as the prediction attribute. Anyway, “Close” is always used as the input variable. Finally, stock prices are predicted and the predicted stock prices are converted back to its original form by using inverse z-score transformation.

4.2 Parameter Settings and Evaluation Metrics

There are two parameters we need to confirm, i.e., the look-back period T and the size of hidden states m for the LSTM. To determine the look-back period T, we conducted a grid search over $T\in\{3,5,10,15,25\}$. Taking Longji Shares for example, the one (T=5) that achieves the best performance over validation set is used for test as shown in Table 1. For the size of hidden states m for LSTM, we conducted a grid search over $m\in\{16,32,64,128\}$. Also, taking Longji Shares for example, the one (m=64) that achieves the best performance over validation set is used for test as shown in Table 2.

DOI https://doi.org/2024-JICS-22368 | Generated on 2025-04-12 03:46:34

Table 1: The performance results of the look-back period of Longi Shares (601012).

T	MAPE (x100)	MAE	RMSE
3	4.645	4.17	5.38
5	4.582	4.097	5.352
10	4.688	4.208	5.483
15	5.027	4.513	5.796
20	4.644	4.175	5.449

Table 2: The performance results of the size of hidden states of Longi Shares (601012).

m	MAPE (x100)	MAE	RMSE
16	5.321	4.74	6.061
32	5.12	4.591	5.844
64	4.582	4.097	5.352
128	4.499	4.036	5.333

To measure the effectiveness of various input feature subsets for time series prediction, we consider three different evaluation metrics. Among them, root mean squared error (RMSE) and mean absolute error (MAE) are two scale-dependent measures, and mean absolute percentage error (MAPE) is a scale-dependent measure. Specifically, as summing y_t is the target at time t and $\hat{y}_t$ is the predicted value at time t, RMSE [12] and MAE is defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_t^i - \hat{y}_t^i\right)^2} \tag{6}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_t^i - \hat{y}_t^i\right| \tag{7}$$

When comparing the prediction performance across different datasets, mean absolute percentage error is popular because it measures the prediction deviation proportion in terms of the true values. It is defined as follows:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_t^i - \hat{y}_t^i}{y_t^i}\right| \times 100\% \tag{8}$$

For model, we use single layer of LSTM and 20% dropout to prevent overfitting. For each dataset, we train them 10 times and compare the average of three evaluation metrics of 10 times.

4.3 Results

Using three feature selection methods: Multiple linear regression analysis, Correlation matrix heatmap and Feature importance to obtain new input features. Plus the original input features, four different input feature sets of 10 stocks were established to test the prediction results of different input dataset. We provide feature numbers with different way of feature selection and three different evaluation metrics of prediction results of 10 stock as shown in Table3-Table12. We fixed the number of input features of Feature importance, noticing that the feature number of Correlation matrix heatmap is always least and its prediction result is better than the other two in 9 stocks. And Feature importance performs better than Multiple linear regression analysis in most cases. What's more, all of three input feature sets obtained by feature selection methods perform better

DOI https://doi.org/2024-JICS-22368 | Generated on 2025-04-12 03:46:34

than original feature set.

Table 3: The performance results of the four different input feature sets of Longi Shares (601012).

Input features	Longi Shares (601012)			
	feature numbers	MAPE (x 100)	MAE	RSE
Original	15	4.582	4.097	5.352
Multiple linear regression analysis	11	4.194	3.757	5.065
Correlation matrix heatmap	8	3.67	3.299	4.544
Feature importance	10	3.955	3.535	4.877

Table 4: The performance results of the four different input feature sets of Tianqi Lithium (002466).

Input features	Tianqi Lithium (002466)			
	feature numbers	MAPE (x 100)	MAE	RSE
Original	15	8.098	7.359	12.332
Multiple linear regression analysis	14	8.063	7.323	12.261
Correlation matrix heatmap	7	5.422	4.348	6.871
Feature importance	10	7.939	7.123	11.828

Table 5: The performance results of the four different input feature sets of Desay Battery (000049).

Input features	Desay Battery (000049)			
	feature numbers	MAPE (x 100)	MAE	RSE
Original	15	3.543	1.761	3.122
Multiple linear regression analysis	13	3.486	1.73	3.149
Correlation matrix heatmap	7	2.918	1.462	2.797
Feature importance	10	3.39	1.69	2.192

Table 6: The performance results of the four different input feature sets of Tinci Materials (002709).

Input features	Tinci Materials (002709)			
	feature numbers	MAPE (x 100)	MAE	RSE
Original	15	7.31	7.181	9.394
Multiple linear regression analysis	13	6.827	6.676	8.708
Correlation matrix heatmap	8	6.244	6.067	8.522
Feature importance	10	6.3	6.036	8.459

DOI https://doi.org/2024-JICS-22368 | Generated on 2025-04-12 03:46:34

Table 7: The performance results of the four different input feature sets of Sungrow Power (300274).

Input features	Sungrow Power (300274)			
	feature numbers	MAPE (x 100)	MAE	RSE
Original	15	11.551	13.749	18.968
Multiple linear regression analysis	13	11.07	13.285	18.499
Correlation matrix heatmap	7	7.273	8.665	12.207
Feature importance	10	10.669	12.844	17.857

Table 8: The performance results of the four different input feature sets of Chint Electrics(601877).

Input features	Chint Electrics(601877)			
	feature numbers	MAPE (x 100)	MAE	RSE
Original	15	6.09	2.842	5.165
Multiple linear regression analysis	13	5.814	2.711	4.927
Correlation matrix heatmap	9	4.541	2.086	3.747
Feature importance	10	5.745	2.673	4.908

Table 9: The performance results of the four different input feature sets of BYD (002594).

Input features	BYD (002594)			
	feature numbers	MAPE (x 100)	MAE	RSE
Original	15	5.913	13.359	17.226
Multiple linear regression analysis	12	5.671	12.869	16.661
Correlation matrix heatmap	8	4.397	9.69	12.425
Feature importance	10	4.66	10.679	14.073

Table 10: The performance results of the four different input feature sets of Energy Very Endure(300014).

Input features	Energy Very Endure(300014)			
	feature numbers	MAPE (x 100)	MAE	RSE
Original	15	4.519	4.494	5.689
Multiple linear regression analysis	10	4.286	4.321	5.554
Correlation matrix heatmap	8	3.742	3.744	5.04
Feature importance	10	4.045	4.076	5.38

5. Conclusion

After using three feature selection methods: Multiple linear regression analysis, Correlation matrix heatmap and Feature importance, the prediction results of all ten stocks are performed better than the prediction results under the original features. Through feature selection, the complexity of the problem can be reduced, and the prediction accuracy, robustness and interpretability of the learning algorithm can be improved. In addition, we find that among three feature selection methods, in most cases, Correlation matrix heatmap performs best, Feature importance performs better than Multiple linear regression analysis. Feature numbers

DOI https://doi.org/2024-JICS-22368 | Generated on 2025-04-12 03:46:34

also affect the prediction results. The smaller subset of features may be perform better.

Table 11: The performance results of the four different input feature sets of Inovance Technology(300124).

Input features	Inovance Technology(300124)			
	feature numbers	MAPE (x 100)	MAE	RSE
Original	15	5.581	4.211	5.769
Multiple linear regression analysis	13	5.168	3.919	5.389
Correlation matrix heatmap	7	4.359	3.332	4.972
Feature importance	10	4.662	3.529	5.144

Table 12: The performance results of the four different input feature sets of Lead Intelligent (300450).

Input features	Lead Intelligent (300450)			
	feature numbers	MAPE (x 100)	MAE	RSE
Original	15	3.957	2.766	4.621
Multiple linear regression analysis	9	3.497	2.45	4.379
Correlation matrix heatmap	7	3.603	2.522	4.424
Feature importance	10	3.529	2.476	4.395

6. References

[1] Chong E, Han C, Park F C. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. Expert Systems with Applications, 2017,83: 187-205.

[2] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by backpropagating errors. Nature, 323(9):533–536, 1986.

[3] Paul J Werbos. Backpropagation through time: what it does and how to do it. Proceedings of the IEEE, 78(10):1550–1560, 1990.

[4] Jeffrey L Elman. Distributed representations, simple recurrent networks, and grammatical structure. Machine learning, 7(2-3):195–225, 1991.

[5] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2):157–166, 1994.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural Computation, 9(8):1735-1780,1997.

[7] KIRA K. The feature selection problem: traditional methods and a new algorithm. Proc. AAAI-92, 1992.

[8] KOLLER D. Toward optimal feature selection// Proc.13th International Conference on Machine Learning. Morgan Kaufmann, 1996.

[9] Breiman L. Baggingpredictors. Machine learning, 1996, 24(2): 123-140.

[10] Lunetta K. L., Hayward L. B., Segal J., et al. Screening large-scale association study data: exploiting interactions using random forest. Bmc Genetics, 2004, 5(1): 1-13.

[11] Geurts P., Ernst D., Wehenkel L. Extremely randomized trees. Machine Learning, 2006, 63(1): 3-42.

[12] Mark Plutowski, Garrison Cottrell, and Halbert White. Experience with selecting exemplars from clean data. Neural Networks, 9(2):273–294, 1996.