Research on Optimization of Random Forest Algorithm Based on Feature Engineering

Lei Yang, Yong Fan and Yungui Chen
Guangdong University of Science and Technology, Dongguan, China

Keywords: Random Forest, Feature Engineering, Machine Learning.

Abstract: Random forest is an ensemble learning method that builds a strong classifier by combining multiple weak classifiers. Feature engineering refers to the process of improving model performance by selecting and manipulating features. This paper selects 3 A-share stocks of China's Shanghai Stock Exchange as the research object, and proposes a prediction model based on the random forest optimization algorithm of feature engineering, and uses this model to predict the closing price of the stock. Our research results show that the optimized model performance of predictions can be improved.

1 INTRODUCTION

In machine learning, feature engineering refers to the process of improving model performance by selecting and manipulating features. Traditional feature engineering methods include feature selection, feature extraction, and feature transformation 0. Feature selection refers to the selection of features that have a significant impact on model performance through statistical methods or machine learning algorithms. Feature extraction refers to the extraction of valuable information of features through clustering, embedding, transformation and other methods 0. Feature conversion refers to converting features into a form that is more suitable for model learning.

This paper uses Spearman correlation to analyze the characteristics of the data, and selects the most useful features to construct a new data set by performing feature selection on the data set0. This paper uses the filtering method for feature selection, analyzes the characteristics of the data with the help of Spearman correlation, extracts the data with higher correlation and separates it into another data set, and uses the same random forest algorithm parameters 0. Two datasets for learning and prediction. Spearman correlation is a nonparametric method for measuring the correlation between two variables. It measures the monotonic relationship between variables, i.e. whether they follow the same trend. Unlike Pearson correlation, Spearman correlation does not require the relationship between variables to be linear, so Spearman correlation is more suitable when there is a nonlinear relationship between variables 0.

The value range of Spearman correlation is between -1 and 1, where 0 means that there is no monotonic relationship between two variables, and -1 means that there is a completely opposite monotonic relationship between two variables, that is, when one variable increases, the other variable will decrease, and 1 means that there is exactly the same monotonic relationship between the two variables, that is, when one variable increases, the other variable also increases 0.

Random forest is an ensemble learning method that builds a strong classifier by combining multiple weak classifiers. It is a probabilistic prediction model that builds a model through a large number of training samples and random feature selection. Its basic principle is to build a model through a large number of training samples and random feature selection methods, each weak classifier classifies the training samples, and finally summarizes the prediction results of all weak classifiers to obtain the final prediction result 0. Random forest has been widely used in many fields because of its good generalization ability, robustness and interpretability.

This paper proposes a random forest algorithm optimization method based on feature engineering. This method improves on the random forest algorithm by combining techniques such as feature selection, feature extraction, and feature transformation. Specifically, this method first selects features that have a significant impact on model performance through feature selection methods, then extracts valuable information about features through feature extraction methods, and finally converts features into

Yang, L., Fan, Y. and Chen, Y.
Research on Optimization of Random Forest Algorithm Based on Feature Engineering.
DOI: 10.5220/0012273400003807
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (ANIT 2023)*, pages 34-38
ISBN: 978-989-758-677-4
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

a form that is more suitable for model learning through feature conversion methods 0.

Experimental results show that this method can effectively improve the performance of random forest algorithm. Compared with the traditional random forest algorithm, the optimized model has achieved better accuracy and faster calculation speed on the dataset. In addition, the method is also scalable and portable, and can be applied to different fields and datasets.

2 DATASET AND METHOD

The data used in this research comes from the data of China's Shanghai Stock Exchange Index in the first half of 2022. The historical transaction data of three Chinese Shanghai Stock Exchange A-shares were selected as the research objects, namely Gree Real Estate Co., Ltd., Nanjing Gaoke Company Limited and China Railway Hi-Tech Industry Corporation Limited. The stock codes are 600185, 600064 and 600528 respectively.

This study analyzes and processes the data of the opening date, opening price, highest price, lowest price, closing price and trading volume of these three stocks, and uses random forest machine learning algorithm to predict the closing price of the stock. The original data set of each stock has 117 samples, each sample contains 8 features, the first 102 samples are used as the training set, and the remaining 15 samples are used as the test set.

Algorithm 1. Random forest creation.

INPUT: Training set *A*； Test set *B*； Trees number *M*;
OUTPUT: Predicting outcomes
Procedure: Function Random forest(*A*, *M*)
for *m*=*M* **do**
  Randomly create *A*';
  Create fully growing decision tree;
  Predict *B*;
end for
Vote on *M* forecasts and adopt the principle of majority;
Save forecast results;

We set the number of trees in the random forest to 12. When the number of subsets is less than 2, we will not split them 0. After the predicted results are output, we will visualize the trees in the random forest. Meanwhile, we also performed performance analysis of our predictive model using different evaluation metrics 0. The random forest algorithm is shown in the algorithm 1.

In order to optimize the prediction model of the random forest algorithm and better predict the stock price, we applied feature engineering to the original data set. We used a filter-style feature selection algorithm to select features by performing statistical tests or mathematical transformations on them. We analyze the characteristics of the data with the help of Spearman correlation, and perform correlation analysis on the 8 features in the data set. After Spearman correlation analysis, we finally selected the strongest 3 features plus the opening date, a total of 4 features, and constructed a new data set. Finally, we use the same random forest algorithm parameters to perform machine learning on the two data sets before and after. The results of the study show that the optimized model has indeed improved predictive performance.

3 EXPERIMENT AND ANALYSIS

Random forest is an ensemble learning method that consists of multiple decision trees, each of which is a weak classifier. Ensemble learning is a technique that combines multiple classifiers into a single strong classifier 0. The random forest has a strong generalization ability, and it can build a strong classifier by combining multiple weak classifiers, thereby improving the performance of the model. Random feature selection and random sample selection in random forest can reduce the model's dependence on training data and improve the robustness of the model. Feature importance can be used for feature selection and model interpretation, thereby improving model performance and interpretability0.

In random forests, a weak classifier is one that can only classify the data in the training samples. Weak classifiers have limited classification ability, so it is necessary to combine multiple weak classifiers to build a strong classifier. A strong classifier refers to a classifier that can classify both training samples and unknown samples, and it can predict unknown samples, so it has a higher classification ability 0. Each weak classifier in random forest is classified based on a single feature, which can reduce the model's dependence on training samples, thereby improving the robustness of the model.

Each decision tree in a random forest is constructed by means of random feature selection and random sample selection. Random feature selection refers to randomly selecting a part of features from all features for classification when the nodes of the decision tree are split 0. Random sample selection refers to randomly selecting a part of data from the training data for training when building a decision tree. This method can make each decision tree have a certain degree of difference, thereby avoiding

overfitting and improving the generalization ability of the model.

Random forest can improve the interpretability of the model by analyzing the impact of each feature on the model's prediction results to understand how the model works. Feature importance refers to the degree of influence of features on the prediction results of the model, which can be obtained by calculating the importance of each feature 0. Feature importance can be used for feature selection to improve model performance and interpretability. As shown in Figure 1, it is the forest model for stock price prediction of Gree Real Estate. The random forest models for the other 2 stocks are similar.

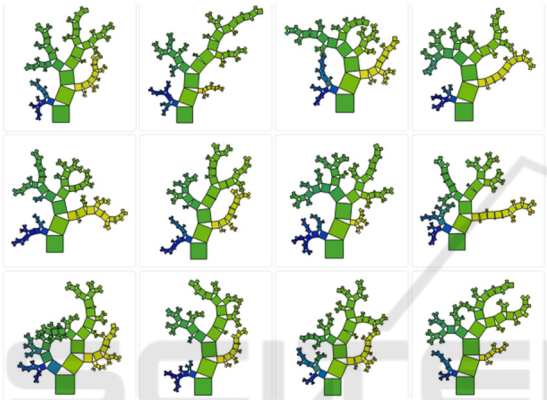

Figure 1. The random forest models with Gree Real Estate.

This paper analyzes the historical transaction data of three stocks and divides them into training set and test set. A total of 351 pieces of data are included, and the amount of data for each stock is 117 pieces. In order to better compare the effect of feature selection on the prediction model, a fixed sampling method is used to select the training set and the test set. The training set contains 102 samples, while the test set contains 15 samples. The prediction results are shown in Table 1.

Table 1. The predicted results and real values of the 3 stocks.

SN	Gree real	Gree pred	Gree devi	NJ GK real	NJ GK pred	NJ GK devi	CRHIC real	CRHIC pred	CRHIC devi
1	6.8	6.79	0.15%	10.85	10.84	0.07%	8.29	8.22	0.81%
2	6.54	6.53	0.18%	10.62	10.74	1.10%	8.07	8.13	0.72%
3	6.94	6.89	0.73%	10.82	10.80	0.22%	8.1	8.07	0.35%
4	6.77	6.84	0.97%	11.01	11.02	0.11%	8.25	8.27	0.21%
5	6.78	6.76	0.28%	10.86	10.97	0.98%	8.18	8.21	0.38%
6	7.23	7.24	0.08%	10.86	10.85	0.12%	8.12	8.11	0.08%
7	7.23	7.18	0.76%	10.94	10.94	0.05%	8.11	8.12	0.06%
8	7.3	7.20	1.39%	10.9	10.94	0.33%	8.07	8.12	0.62%
9	6.93	6.94	0.08%	10.75	10.80	0.48%	7.94	7.99	0.60%
10	6.89	6.81	1.10%	10.91	10.82	0.84%	8.03	8.03	0.06%
11	6.51	6.64	1.97%	10.9	10.89	0.09%	8.03	8.03	0.03%
12	6.28	6.29	0.21%	10.92	10.93	0.12%	8	8.00	0.01%
13	6.32	6.32	0.04%	11.08	11.06	0.17%	8.05	8.04	0.10%
14	6.34	6.36	0.32%	11.02	11.03	0.12%	8.05	8.06	0.14%
15	6.31	6.33	0.24%	11	11.02	0.21%	8.05	8.02	0.35%

It can be seen from the above table that most of the deviations between the predicted values of the 3 stocks and the real values are less than 1%. The curves of the predicted values and the real values of the 3 stocks are shown in Figure 2.

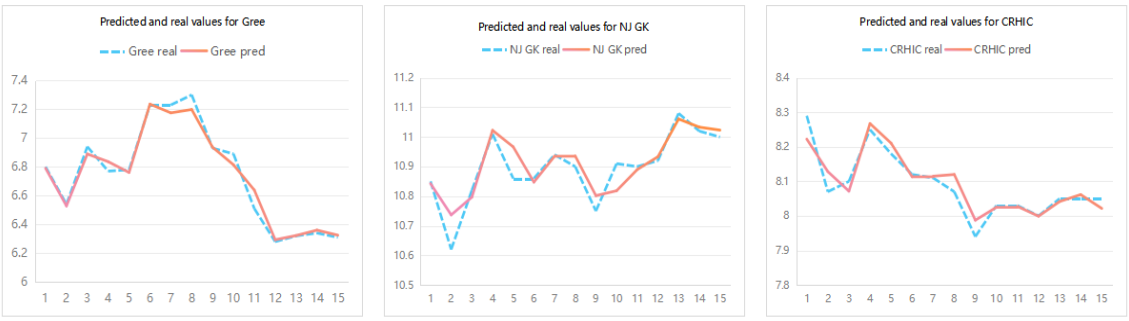

Figure 2. The Curves of the Predicted Value and the Real Value of the 3 Stocks.

In order to optimize our prediction model, for the original data set, we use Spearman correlation to analyze the characteristics of the data and perform correlation analysis on the 8 features in the data set. The correlation analysis of the three stocks is shown in Table 2.

Table 2. The correlation analysis of the 3 stocks' features.

SN	Gree			NJ GK			CRHIC		
1	0.974	close	high	0.972	close	low	0.971	close	low
2	0.967	close	low	0.969	close	high	0.964	close	high
3	0.938	close	open	0.932	close	open	0.921	close	open
4	0.432	amount	close	0.833	close	date	0.586	amount	close
5	0.32	close	volume	0.526	amount	close	-0.56	close	date
6	-0.013	close	date	0.421	close	volume	0.514	close	volume

As can be seen from Table 2, the three features of the three stocks, the opening price, the highest price of the day, and the lowest price of the day, have the strongest correlation with the closing price. Therefore, we choose these three features, plus the opening time, to form a the new data set uses the same random forest algorithm and parameters to perform machine learning on the new data set again, and the prediction results are shown in Table 3.

Table 3. The predicted results and real values of the new dataset.

SN	Gree real	Gree pred	Gree devi	NJ GK real	NJ GK pred	NJ GK devi	CRHIC real	CRHIC pred	CRHIC devi
1	6.8	6.77	0.49%	10.85	10.84	0.09%	8.29	8.24	0.60%
2	6.54	6.50	0.60%	10.62	10.67	0.49%	8.07	8.07	0.04%
3	6.94	6.89	0.67%	10.82	10.79	0.31%	8.1	8.07	0.36%
4	6.77	6.82	0.66%	11.01	11.00	0.08%	8.25	8.24	0.13%
5	6.78	6.73	0.75%	10.86	10.89	0.27%	8.18	8.22	0.54%
6	7.23	7.21	0.27%	10.86	10.84	0.21%	8.12	8.12	0.02%
7	7.23	7.16	1.03%	10.94	10.94	0.05%	8.11	8.12	0.16%
8	7.3	7.28	0.32%	10.9	10.92	0.20%	8.07	8.08	0.15%
9	6.93	6.99	0.87%	10.75	10.83	0.75%	7.94	7.96	0.19%
10	6.89	6.86	0.44%	10.91	10.87	0.36%	8.03	8.02	0.07%
11	6.51	6.61	1.51%	10.9	10.87	0.24%	8.03	8.04	0.10%
12	6.28	6.30	0.35%	10.92	10.94	0.15%	8	8.01	0.06%
13	6.32	6.31	0.20%	11.08	11.05	0.28%	8.05	8.05	0.06%
14	6.34	6.35	0.08%	11.02	11.02	0.03%	8.05	8.07	0.28%
15	6.31	6.30	0.11%	11	10.99	0.12%	8.05	8.04	0.12%

MSA (Mean Squared Absolute Error), RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and R2 (R Squared) are commonly used regression error indicators or scoring indicators. The value of R2 is between 0 and 1, the closer to 1, the better the fitting effect of the model. Table 4 shows the performance of random forest's predictive scoring indicators on the two data sets before and after.

Table 4. The performance of the random forest model.

Stock	MSE before	MSE after	RMSE before	RMSE after	MAE before	MAE after	R2 before	R2 after	R2 impr	AVG devi before	AVG devi after	AVG devi impr
Gree	0.003	0.002	0.054	0.045	0.039	0.038	0.974	0.982	0.82%	0.57%	0.56%	2.14%
NJ GK	0.003	0.001	0.051	0.033	0.036	0.026	0.784	0.912	16.33%	0.33%	0.24%	27.41%
CRHIC	0.001	0.001	0.033	0.021	0.025	0.016	0.866	0.943	8.89%	0.30%	0.19%	35.94%

As shown in Table 4, the performance of the prediction model after the feature engineering process has been improved in both R2 and average deviation, especially the average deviation index, which has increased by more than 25%. Our experimental results show that after feature engineering processing, the performance of machine learning model prediction can indeed be improved.

4 CONCLUSION

Based on feature engineering, this study improves the prediction model of random forest algorithm, and uses this model to predict the closing price of stocks. The research shows that the performance of the prediction model has been improved. Research on machine learning algorithm optimization based on feature engineering is an important research direction in the field of machine learning. By selecting the most representative and important features, the performance and efficiency of machine learning algorithms can be improved, and the model's generalization ability and interpretability. In addition, the randomness of the random forest will also have a greater impact on the performance of the model, which is also our future research direction.

ACKNOWLEDGMENTS

This research was funded by the following programs: Research Capability Enhancement Project of Guangdong University of Science and Technology: Application Research of Artificial Intelligence Technology Based on Kunpeng Computing Platform; Natural Science Project of Guangdong University of Science and Technology(GKY-2022KYZDK-12, GKY-2022KYZDK-9); Innovation and School Strengthening Project of Guangdong University of Science and Technology(GKY-2022CQTD-2, GKY-

2022CQTD-4, CQ2020062); 2022 Guangdong Province Undergraduate University Quality Engineering Construction Project - Exploration of Integrated Teaching Reform of "Curriculum Chain, Practice Chain, and Competition Chain"; Education Research and Reform Project of Online Open Course Alliance of Universities in Guangdong–Hong Kong–Macao Greater Bay Area(WGKM2023166); Quality Engineering Project of Guangdong University of Science and Technology(GKZLGC2022018, GKZLGC2022271).

REFERENCES

Illa P K, Parvathala B, Sharma A K. Stock price prediction methodology using random forest algorithm and support vector machine [J]. *Materials Today: Proceedings*, 2022, 56: 1776-1782.

Verma S, Sahu S P, Sahu T P. Discrete wavelet transform-based feature engineering for stock market prediction [J]. *International Journal of Information Technology*, 2023, 15(2): 1179-1188.

Yin L, Li B, Li P, et al. Research on stock trend prediction method based on optimized random forest [J]. *CAAI Transactions on Intelligence Technology*, 2023, 8(1): 274-284.

Wang Z, Xia L, Yuan H, et al. Principles, research status, and prospects of feature engineering for data-driven building energy prediction: A comprehensive review [J]. *Journal of Building Engineering*, 2022: 105028.

Abraham R, Samad M E, Bakhach A M, et al. Forecasting a stock trend using genetic algorithm and random forest [J]. *Journal of Risk and Financial Management*, 2022, 15(5): 188.

Htun H H, Biehl M, Petkov N. Survey of feature selection and extraction techniques for stock market prediction[J]. *Financial Innovation*, 2023, 9(1): 26.

Singh G. Machine learning models in stock market prediction [J]. *arXiv preprint arXiv:*2202.09359, 2022.

Argade S, Chothe P, Gawande A, et al. Machine learning in stock market prediction: A Review[J]. *Available at SSRN 4128716, 2022*.

Li P, Gu H, Yin L, et al. Research on trend prediction of component stock in fuzzy time series based on deep forest [J]. *CAAI Transactions on Intelligence Technology*, 2022, 7(4): 617-626.

Mathanprasad L, Gunasekaran M. Analysing the Trend of Stock Marketand Evaluate the performance of Market Prediction using Machine Learning Approach[C]//*2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*. IEEE, 2022: 1-9.

Koukaras P, Nousi C, Tjortjis C. Stock market prediction using microblogging sentiment analysis and machine learning[C]//*Telecom*. MDPI, 2022, 3(2): 358-378.

Li H, Li X, Liu X, et al. Prediction of blast furnace parameters using feature engineering and Stacking algorithm [J]. *Ironmaking & Steelmaking*, 2022, 49(3): 283-296.

Mndawe S T, Paul B S, Doorsamy W. Development of a stock price prediction framework for intelligent media and technical analysis [J]. *Applied Sciences*, 2022, 12(2): 719.

Mahfooz S Z, Ali I, Khan M N. Improving stock trend prediction using LSTM neural network trained on a complex trading strategy [J]. *International Journal for Research in Applied Science and Engineering Technology*, 2022, 10(7): 4361-4371.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS