Carbon Neutral Concept Stock Price Forecast and Public Opinion Analysis System based on Deep Learning

Hao Yang[1, a], Longxiang Xiao[2, b], Zichao Zhang[2, c], Chunzhao Chen[2, d]

[1]Anhui University of Finance and Economics, School of Finance, Internet Finance, Bengbu, China

[2]Anhui University of Finance and Economics, Bengbu, China

[a]1725737893@qq.com, [b]727650500@qq.com, [c]1196597924@qq.com, [d]1114606598@qq.com

Abstract

The rapid development of financial technology has expanded the way of stock market forecasting in the securities market. The random forest model and BP neural network model based on deep learning can explore the deep features of text data and find special clues from the scattered information on the network. Quantitative analysis of public opinion in the market provides a solution. Based on the national "carbon neutrality" background, this article takes text comments of stockholders of carbon neutral concept stocks as samples, and uses real stock price fluctuations as labels. The sample data is preprocessed by clearing and text vectorization, etc. Random Forest's stock price prediction model, after training and evaluation, can judge the stock price rise and fall by up to 62%, and the stockholder comment sentiment analysis model based on the BP neural network has an accurate score of 0.98, which is highly accurate.

Keywords

Financial Technology; Random Forest; Carbon Neutral Concept Stock; Neural Network.

1. Introduction

In the stock market, the rise and fall of stock prices are affected by various factors such as price indicators, liquidity indicators, and activity levels. In the stock market, people hope to effectively predict the trend of stock prices, so as to avoid unnecessary losses, and analyze important factors that affect stock price fluctuations. But in the stock market, the fluctuation of stock price itself is a non-linear, dynamic and unstable process. The fluctuation process itself contains large or small noise, which has a significant impact on the trend of stock prices. Therefore, how to more accurately predict the stock price trend and the degree of volatility under the characteristics of multi-dimensional data has become a concern of many scholars at home and abroad. On September 22, 2020, during the general debate of the 75th UN General Assembly, "carbon neutrality" was highlighted. President Xi Jinping said at the meeting that China will increase its nationally determined contributions, adopt more powerful policies and measures, strive to reach its peak carbon dioxide emissions by 2030, and strive to achieve carbon neutrality by 2060. Because Shenzhen Energy holds an important share in the Shenzhen Carbon Emissions Exchange, as a direct beneficiary of the national "carbon neutral" strategy, we select Shenzhen Energy as the stock for quantitative analysis in this article.

2. Literature Review

Machine learning models have been more and more widely used in the field of financial investment because of their ability to deal with the collinearity and nonlinear relationships between variables. Zhou Liang [1] used 28 common factors to establish a random forest model to predict the return rate of the CSI 50 index constituent stocks and construct a portfolio, and found that the random forest model can better fit and predict the relative return rate of individual stocks. The annualized rate of return and Sharpe ratio of the portfolio are as high as 27.31% and 1.59, respectively. He Pingren[2] studied the predictability of 41 characteristic variables of listed companies on the out-of-sample of my country's stock returns. The research results show that the combined LASSO-Logistic algorithm can effectively identify the complex relationship between characteristic variables and expected returns. Its investment The strategy of portfolio asset allocation can obtain higher excess returns than the traditional multi-logistic algorithm, support vector machine (SVM) algorithm and random forest algorithm. Yan Zhengxu et al. [3] proposed a new combination model method of random forest based on Pearson coefficient based on random forest, which can realize short-term prediction regression of stock prices and reduce the influence of noise on stock price prediction. Yu Cilong et al.[4] natural language processing technology based on deep learning can explore the deep characteristics of text data, find special clues from scattered information on the Internet, and provide a solution for the quantitative analysis of public opinion in the financial market. Zhao Tingting et al.[5] will take stock time series data forecasting as an example to introduce the time series data forecasting methods in detail, focusing on the analysis of nonlinear forecasting methods, and discuss their future development trends. Wei Jian et al.[6] The proportion of the combination forecasting method that correctly predicts the closing price rise and fall is much higher than that of the single forecast model, and the proportion of correct prediction reaches 94.33%. In other error standards, the combination model also has certain advantages.

3. Research and Design

3.1. Random Forest Model

Random forest is an ensemble learning algorithm that combines multiple weak classifiers into a strong classifier. Random Forest uses bootstrap to randomly sample m samples with replacement on the training set, and selects random features for each decision tree on the basis of bagging. Build m decision tree models from these m samples. Finally, the results are obtained by voting through these m decision tree models. The specific algorithm steps of random forest are as follows:

(1) Enter the training set D.

(2) Using bootstrap sampling to form k training subsets $D_{k.}$

(3) Randomly extract m features from the original features.

(4) Perform training on the training subset D_k, and optimally segment the randomly selected m features to obtain k decision tree prediction results.

(5) Voting based on k prediction results to get the prediction result with the highest number of votes.

Random forest is a machine learning algorithm that is easily affected by its own parameters and characteristic variables. In order to improve the prediction effect of random forest, this article first builds a Multi-factor model, extracts feature variables and target variables, divides the training set and test set, and obtains the prediction results through the random forest composed of k decision trees. The algorithm process is shown in Figure 1.

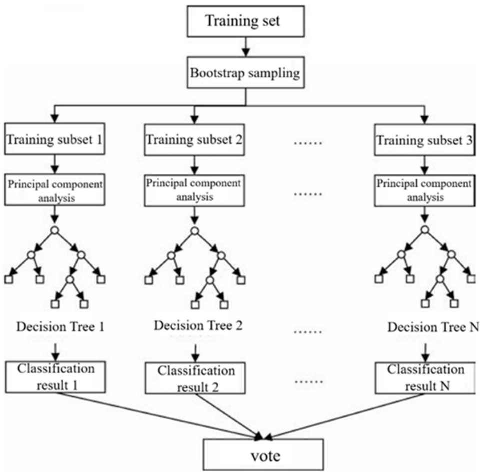

Figure 1. Random forest model

3.2. BP Neural Network Model

The type of neural network involved in the model established in this article is BP neural network, which is a multi-layer feedforward neural network. BP neural network was proposed by a scientific group headed by Rumelhart and McCelland in 1986. It is a multi-layer feedforward network trained by error reversal propagation algorithm and is currently one of the most widely used neural network models [7]. The BP neural network can learn and store a large number of input/output pattern mapping relationships without revealing the mathematical equations describing this mapping relationship in advance. Its learning rule is to use the steepest descent method to continuously adjust the weights and thresholds of the network through backpropagation to minimize the sum of squared errors of the network. The topological structure of BP neural network includes input layer, hidden layer and output layer.

The general model of BP neural network is shown in Figure 2. BP neural network is similar to other neurons, the difference is that the transfer function of BP neuron is a non-linear function, the most commonly used are logsig and tansig functions, and some also use linear functions. The output is a=log sig (W_b+b).

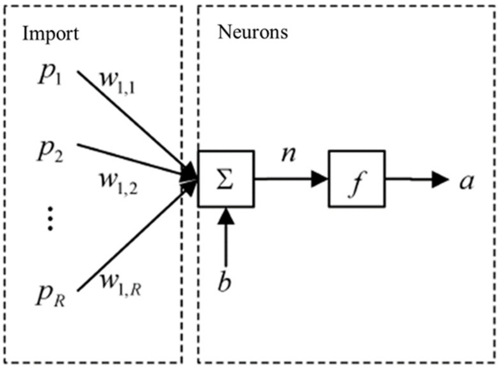

Figure 2. General model of BP neural network

The BP neural network is generally a multilayer neural network, and the information of the BP neural network flows from the input layer to the output layer. If the output of the multilayer BP neural network uses an S-type transfer function (such as logsig), as shown in Figure 3, its output value will be limited to a smaller range such as (0,1); while a linear transfer function is used Can take any value.

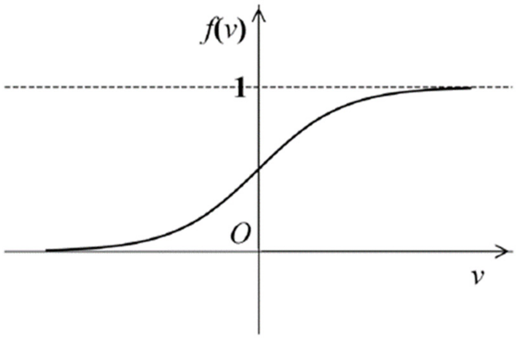

Figure 3. Sigmoid activation function

The BP neural network model includes its input/output model, action function model, error calculation model and self-learning model. The hidden node output model of BP neural network is:

$$\mathrm{O}_j = f(\sum W_{ij}X_{ij} - q_j) \tag{1}$$

The output model of the output node is:

$$\mathrm{Y}_k = f(\sum T_{jk}\mathrm{O}_j - q_k) \tag{2}$$

Among them, f is the non-linear action function, and q is the neuron threshold.
The action function is a function that reflects the stimulation pulse intensity of the lower layer input to the upper layer node, also known as the stimulation function, generally takes the continuous value sigmoid function within (0,1): $f(x) = \frac{1}{1+e}$; error calculation model It is a function reflecting the size of the error between the expected output of the neural network and the calculated output: $\mathrm{E}_q = \frac{1}{2}\sum(t_{pi} - \mathrm{O}_{pi})$, where tpi is the expected output value of the node, and O_{pi} is the calculated output value of the node; The learning process is the setting and error correction process of the weight matrix W_{ij} connecting the lower node and the upper node.

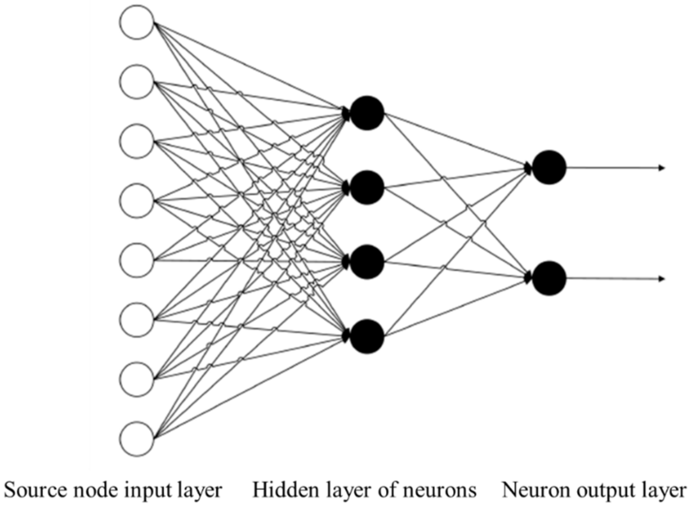

Figure 4. Single-layer forward network

4. Empirical Analysis

4.1. The Stock Price Prediction Model based on Random Forest

4.1.1. Multi-factor Model Construction

(1)Data acquisition

First, use Python’s Tushare library to try to obtain historical stock data. Here, take the carbon neutral stock Shenzhen Energy as an example, the stock code is 000027, and the daily K-line data obtained from January 1, 2020 to June 30, 2021[8].

(2)Extraction of feature variables and target variables

First, select some characteristic variables, namely factors, to assign to the variable x through the panda’s library to select data, as shown in Table 1, and then construct the target variable y.

Table 1. Constructing characteristic variable index system

Characteristic variable	instruction
close	Closing price of the stock
volume	Stock volume
close-open	Opening price—closing price, which refers to how much the stock price changes in a day
MA5	The average value of the price in the last 5 days
MA10	The average value of the price in the last 10 days
high-low	The biggest change in stock price in one day
RSI	Relative Strength Index
MOM	Momentum indicator
EMA12	12-day moving average
MACD	Moving Average Convergence and Divergence
MACDsignal	9-day exponential moving average
MACDhist	MACD value is half

Because the stock price data of the day is used to predict the stock price fluctuations of the next day, the target variable y should be the stock price fluctuations of the next day. This is because many data in the characteristic variables can only be determined after the end of the day's trading, so the stock price rise and fall during the day's trading is unpredictable, and when the market closes, although the required data is complete, the stock price rises and falls on the day It is a foregone conclusion, and there is no need to predict, so the stock price data of the day is used to predict the stock price rise and fall of the next day.

```
y=np.where(df['price_change'].shift(-1)>0,1,-1)
```

The code uses the where () function in the NumPy library, and the meaning of the three parameters passed in are the judgment condition, the assignment that meets the condition, and the assignment that does not meet the condition. Among them, the shift() function is used to move all the data in the column price change (stock price change) up by one row, so that each row corresponds to the next day's stock price change. The prediction result has only two classifications of 1 or -1, so the judgment condition here is whether the stock price change of the next day is greater than 0, if it is greater than 0, it means that the stock price will rise next day, then y is assigned a value of 1; if it is not greater than 0, it means Next day if the stock price remains unchanged or falls, y is assigned a value of -1.

(3)Data division of training set and test set

Next, we need to divide the original data set. The reason is that the trend of stock prices is time-sensitive. If we divide randomly, the temporal characteristics will be destroyed. Therefore, the training set and the test set are divided according to time series. Because we use the data of the day to predict the stock price rise and fall of the next day, not the stock data of any day to predict the stock price rise and fall of the next day. Therefore, we use the first 90% of the data as the training set, and the last 10% of the data as the test set.

(4)Model building

Set the model parameters: the maximum depth of the decision tree is set to 3, that is, each decision tree has at most 3 layers; the weak learner, that is, the number of decision tree models n_estimators is set to 10, that is, there are 10 decisions in the random forest Tree; the minimum number of samples of leaf nodes min samples_leaf is set to 10, that is, if the number of samples of leaf nodes is less than 10, the splitting stops; the function of the random state parameter random state is to keep the result of each run consistent, the number set here is 1.

4.1.2. Model Use and Evaluation

(1)Forecast the rise and fall of the next day

After the model training is completed, the model can be used to make predictions and evaluate the prediction effect of the model. 1 means that the stock price will rise on the next day, and -1 means that the stock price will remain unchanged or fall on the next day.

As shown in Table 2, the prediction accuracy of the first 5 rows of data is 80%.

Table 2. Comparison of predicted value and actual value

Forecast value	Actual value
1	1
-1	-1
-1	1
-1	-1
-1	-1

At the same time predict the probabilities belonging to each category, the first column is category-1, that is, the probability of the stock price unchanged or falling next day, and the second column is category 1, that is, the probability of the stock price rising next day. Part of the content is shown in Table 3.

Table 3. Classification probability distribution table

Probability of classification as -1	Probability of classification as 1
0.48	0.52
0.50	0.50
0.51	0.49
0.54	0.46
0.53	0.47

(2) Model accuracy evaluation

Finally, the overall prediction accuracy is 0.62, indicating that the model predicts about 62% of the data in the entire test set correctly. The accuracy of this forecast is not particularly high, and it does meet the ever-changing characteristics of the stock market.

4.1.3. Analyze the Importance of Data Characteristics

By summarizing the names of feature variables and their feature importance, and sorting them in descending order according to the "feature importance" field, the results are shown in Table X, the momentum indicator MOM value, the closing price of the day close, high-low, MACD related indicators and other feature variables It has a greater impact on the accuracy of the prediction of the stock price rise and fall results for the next day.

Table 4. The importance of feature vector ranking table

Sort	feature	feature importance
7	MOM	0.167096
2	close-open	0.155537
9	MACD	0.142302
10	MACDsigna	0.124385
11	MACDhist	0.084143
0	close	0.075084
3	MA5	0.074306
5	high-low	0.055287
8	EMA12	0.046613
6	RSI	0.035524
4	MA10	0.034885
1	volume	0.004837

4.1.4. Drawing the Return Test Curve

The prediction accuracy of the model has been evaluated before, but in financial practice, we are more concerned about its return-testing curve, also known as the net worth curve, which means to see whether the results obtained based on the built model are better than those obtained without using the model. it is good.

First calculate the price change of a stock after the forecast, the price change of the original data and the daily stock price change rate, then the cumulative return rate can be calculated, and the return test curve can be drawn based on the return rate. Secondly, calculate the rate of return predicted by the model, and predict the stock price rise and fall of the next day based on the stock price data of the day. If the forecast is 1, then buy on the next day; if the forecast is -1, then sell on the next day. The visualization results are shown in Figure 5. The upper curve in the figure is the yield curve obtained from the model, and the lower curve is the yield curve of the stock itself. It can be seen that the benefits obtained by using the model are still good.

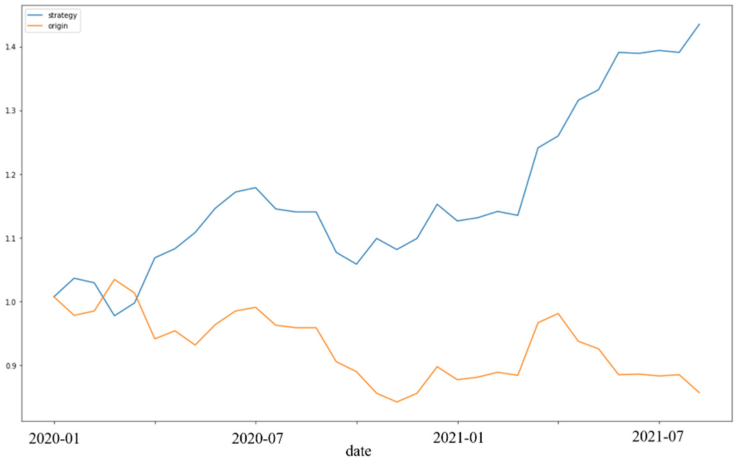

Figure 5. The return test curve

4.2. The Sentiment Analysis Model of Stockholders' Comments based on BP Neural Network

4.2.1. Comment Crawling

Through from selenium import webdriver, automated testing is carried out on the Shenzhen Energy Stock webpage of Oriental Fortune Bar, crawling from May to July of 2021, 1,000 shareholder comments, and data sorting and cleaning through regular expressions.

4.2.2. Predictive Effect

Before building a neural network, we need to read the comment data, and then process the text data into numerical data. Because text data cannot be directly used for training, it is necessary to use the jieba library for text segmentation, construct a word frequency matrix, and then use it to fit the model.

Set the test set data to account for 10%, and the training set data to account for 90%. By predicting the test set data and comparing it with the actual value, as shown in Table 5, observing the first five comparison data shows that the prediction accuracy is 100%.

Table 5. Numerical comparison between predicted and actual values

Forecast value	Actual value
1	1
1	1
1	1
1	1
1	1

The obtained model has an accurate score of 0.981, which means that the model's prediction accuracy has reached 98.1%.

4.2.3. Model Comparison

In order to illustrate the prediction effect of the neural network model, the following is to establish a Gaussian Naive Bayes model for comparison.

X is a data sample of a given unknown category, and the naive Bayes classifier will predict that X belongs to the class with the largest posterior probability (under condition X). That is, to classify X into class C$_i$ if and only if:

$$P(C_i|X) > P\left(C_j\middle|X\right) 1 \leq j \leq m, j \neq i \tag{3}$$

Among them, the class C$_i$ with the largestis called the largest $P(C_i|X)$ posterior hypothesis.

According to Bayes' theorem:

$$P(C_i|X) = \frac{P(X|C_i)}{P(X)} \tag{4}$$

Assuming that P(X) is the same for any category, if $P(C_i|X)$ is the largest, only $P(C_i|X)\ P(C_i)$ is the largest. Among them, the values of $P(C_i)$ and $P(C_i|X)$ can be determined according to the parameter estimation of the naive Bayes classifier.

In order to classify the unknown sample X, the corresponding $P(C_i|X)\ P(C_i)$ of each category C$_i$ is estimated, and the sample X is assigned to the category C$_i$, if and only if

$$P(X|C_i)P(C_i) > P\left(X\middle|C_j\right)P\left(C_j\right)\ 1 \le j \le m, j \ne i \tag{5}$$

The final accuracy score of the model obtained is 0.92, which means that the prediction accuracy of the model reaches 92%, indicating that the prediction effect of the Gaussian Naive Bayes model is slightly inferior to the BP neural network model.

5. Conclusion and Discussion

Market information is the vane of market trends, and improving the utilization rate of market information plays an important role in improving market efficiency and promoting the rational allocation of economic resources. This paper analyzes the emotional expression of individual stocks in the Internet public opinion in the financial market, and establishes a stock price prediction model based on random forest and a sentiment analysis based on BP neural network. The experimental results show that after training and evaluation, the stocks can be judged up and down by 62%, and the stockholder comment sentiment analysis model based on BP neural network has an accurate score of 0.98, which is higher than the prediction effect of the naive Bayes model and has a higher accuracy.

The system can also combine technical indicators such as Bollinger Bands, price change rate, average amplitude, and macro indicators such as exchange rates and interest rates for comprehensive analysis to further improve the accuracy and comprehensiveness of forecasting. Finally, the size of the data set in this article is limited by the hardware equipment and cannot be increased. If the data can be further expanded, and the data sources and types can be enriched, it will be helpful to give full play to the advantages of deep learning, and at the same time solve the over-fitting problem of this model. Better results.

References

[1] Zhou Liang. Research on stock Multi-factor investment based on random forest model[J]. Financial Theory and Practice, 2021(07): 97-103.

[2] He Ping, Lan Wei, Ding Yue. Can my country’s stock market be predicted?--Based on the perspective of combined LASSO-logistic method [J/OL]. Statistical research: 1-15 [2021-07-27].

[3] Yan Zhengxu, Qin Chao, Song Gang. Random forest model stock price prediction based on Pearson feature selection [J/OL]. Computer Engineering and Applications: 1-12 [2021-07-27].

[4] Yu Cilong, Shi Zhenyu, Xie Yunhao, etc. Public opinion analysis and stock price forecast system based on natural language processing [J/OL]. System Engineering: 1-12 [2021-07-27].

[5] Zhao Tingting, Han Yajie, Yang Mengnan, etc. Research review of time series data prediction methods based on machine learning [J/OL]. Journal of Tianjin University of Science and Technology: 1-9 [2021-07-27].

[6] Wei Jian, Zhao Hongtao, Liu Dunnan. Short-term forecasting method of stock closing price based on combination model [J]. Economic Research Guide, 2021(09): 75-79.

[7] Zhu Fan, Wang Yinqi. Research on user information clustering and prediction based on k-means and neural network machine learning algorithm[J]. Information Science, 2021, 39(07): 83-90.

[8] Hu Zhen, Gong Xue, Liu Hua. Research on the prediction of household consumption carbon emissions in western cities based on the BP model--Taking Xi'an as an example [J]. Arid Land Resources and Environment, 2020, 34(07): 82- 89.