Analysis of SME Stock Price Forecasting System Based on Machine Learning

Ziju Hou

Abstract

In financial market development, various theories and hypotheses have been studied and analyzed through different methods to summarize stock prices, including random walk theory, efficient market hypothesis, and behavioral finance. Therefore, it is of great significance to combine the various algorithms of machine learning with the relevant theories of financial markets in quantitative finance. In the market economy, small and medium-sized enterprises(SMEs) absorb many workers in society and play a huge role in production, innovation, and entrepreneurship, which determines the importance of Chinese small enterprises in the financial market and stock market. Machine learning forecasts stock prices as a reference for SMEs and their investors. In other words, enterprises can adjust the direction and proportion of business promptly, and investors can also choose whether to invest according to the forecast results. In addition, this work shows that the forecasting effect of machine learning can meet the needs of investors and SMEs by comparing the stock price forecasting using the RM machine learning algorithm and comparing the forecasting results.

The machine learning algorithms commonly used in quantitative finance are briefly introduced, and the random forest algorithm's application principle in forecasting the stock price direction is described. Specifically, the stock price forecasting system is built on the platform of Python, the logic of the system is explained, and the feasibility of the system is explained through experiments and analysis, which reflects the advantage of machine learning in forecasting the stock price direction, and provides a new path for SMEs to forecast the stock price.

Keywords: Machine Learning; Random Forest Algorithm (RM); Python; Small And Medium-Sized Enterprises(SMEs); Stock Price Forecast

Chapter 1 Introduction

1.1 Research Background and Significance

The scope of application of machine learning algorithms in the financial field is gradually expanding, of which the most typical is the application in the field of financial quantification, including many practical application cases, such as credit score, fraud detection, risk management, portfolio optimization, trading strategy, etc., in which the application of machine learning in the stock market has attracted much attention. For many small and medium-sized innovative enterprises in the financial market, forecasting the stock price through machine learning can clarify the enterprise value and provide investment reference for investors, which can not reduce the risks faced by investors but also prevent the risks brought to the stock market due to information asymmetry to a certain extent.

SMEs absorb many workers in society and play a grand role in production, innovation, and entrepreneurship[1]. Shanghai Stock Exchange and Shenzhen Stock Exchange are the two channels for SME listing in China. Among them, Shanghai Stock Exchange founded the Science and Technology Innovation Board, which is similar to NASDAQ in the United States and focuses on the listing financing of start-up and growth enterprises under the registration system. Establishing the STAR Market broadens the financing channels of SMEs and enriches the financing forms. Still, the great stock price fluctuation of the STAR Market brings greater investment risks to investors and hinders the financing process of SMEs. This suggests that forecasting the stock price of SMEs through machine learning can lubricate the financing process of enterprises, which is of great significance to the financial market and stock market.

In the development of financial markets, various theories, and hypotheses have been studied and analyzed through different methods, including random walk theory, efficient market hypothesis, and behavioral finance. Different theories correspond to factors that affect stock prices with different interpretations. The random walk theory holds that it is completely impossible to forecast the next change in the stock price according to the previous situation in the stock price trend chart; the ideal random walk model holds that the change of the current stock price is completely independent of that of the past stock price; and the efficient market hypothesis acknowledges that the stock price is affected by many factors, such

as market supply and demand, news information, and the intrinsic value of the stock. In the stock market that meets the conditions, the relevant stock information will be reflected in the changing stock price trend in time. At the same time, behavioral finance believes that the stock price reflects the intrinsic value of the enterprise, which is affected by the intrinsic value and to a large extent by investors and other actors [2].

The stock analysis by machine learning has experienced the stage from simple linear regression to nonlinear fitting, in which all kinds of algorithms are constantly optimized[2][3]. With the accumulation of data sets and the breakthrough of hardware facilities, deep learning has also been a strong development. The various open-source frameworks written by researchers bring more people into the research and application of deep learning, which provides a hotbed for SMEs to use machine learning to forecast the company's stock price.

1.2 Research and Application Status at Home And Abroad

Machine learning is widely used in investment portfolios, credit loan scores, treasury bond futures, stock price forecasting, and asset allocation. Below, we will analyze the performance of machine learning in the stock market from two aspects: investment portfolio and stock price forecasting.

1.2.1 Selection of Portfolio by Machine Learning Algorithm

The investment portfolio is a product portfolio of all kinds of financial products (including stocks, bonds, etc.) held by investors or financial institutions. With the development of the financial industry, financial derivatives have become a part of the investment portfolio. The risk of the investment portfolio is lower than that of a single financial product, and a higher rate of return can be achieved by selecting appropriate investment projects, which indirectly implies that an investment portfolio accounts for a large proportion of financial market transactions.

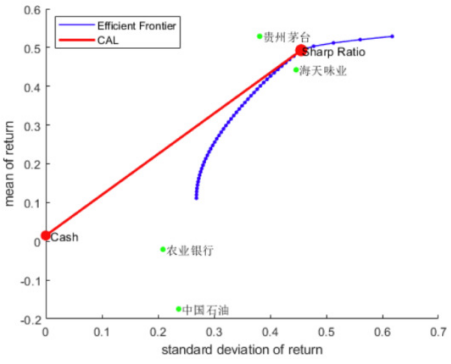


Fig.1 Portfolio Capital Allocation line [4]

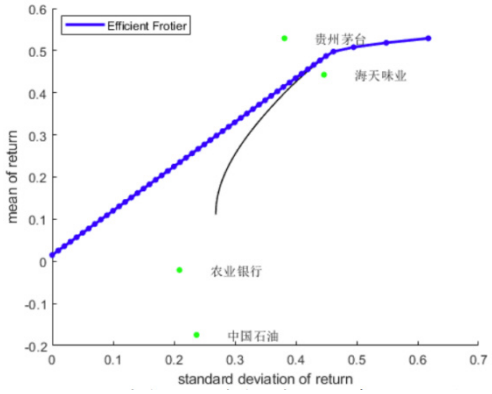


Fig. 2 Efficient Boundary Curve of Portfolio[4]

In the specific investment process, portfolio management means that investors "do not put eggs in one basket" but instead diversify their investment funds into different stocks and constantly adjust the weight of funds in pursuit of maximum returns and minimum risks[5]. There are usually two common ways to analyze financial markets: fundamental analysis and technical analysis. Among them, the fundamental analysis pays attention to the macroeconomic form, the development prospect of the industry, the operation status of the enterprise, etc., and takes this as the analysis object to judge the company's long-term investment value and safety boundary and insists on "value analysis." The technical analysis attaches importance to the data in the securities market, analyzes the changes in the data, and further makes the investment judgment, thinking that the stock price can be forecasted and, to a certain extent, holds the view that "history will repeat itself" [2]. In the stock market, investors often adopt the analysis method of the combination of fundamentals and technical aspects and make investment decisions based on the fundamental indicators disclosed by companies or enterprises and some technical indicators.

Some examples of machine learning applications in the portfolio direction are as follows:

1. GLASSO-MV portfolio strategy [6]**:** compared with the traditional approach, this portfolio approach can make full use of the pricing difference information between factor portfolios, estimate the weight of high-dimensional portfolios more effectively, and finally achieve better investment performance.

2. Ensemble portfolio strategy [7]**:** this portfolio strategy is based on the mean-variance (MV) portfolio strategy proposed by Markowitz [8]. To reduce the error caused by linear regression forecasting, this portfolio method combines mature ensemble learning technology with investment portfolios to improve the return on investment. In addition, the ensemble portfolio strategy also studies the properties of the portfolio loss function.

3. Multi-stage mean-semi-absolute deviation portfolio [9]**:** this portfolio can meet the diversified investment requirements, reduce investment risk, and include transaction costs and loan restrictions in the machine learning model. It can achieve better investment results [10].

1.2.2 Application of Machine Learning to Stock Price Forecasting

The stock market is an important channel for enterprises to raise production capital and can also reflect the real operation of enterprises, known as the "barometer" of the market economy, and plays an important role in the entire financial market. The uncertainty of stocks and the opacity of information have led to various problems in the stock market, which increases the risk for investors. As a result, "whether the stock market can be forecasted" and "how the stock market can be forecasted" have been considered topics of great interest to the financial community and investors.

Traditional stock forecasting adopts linear regression, such as Autoregressive Conditional Heteroscedasticity Model (ARCH) and Autoregressive Integrated Moving Average Model (ARIMA).

1. Autoregressive Integrated Moving Average Model (ARIMA): the model stabilizes the non-stationary time series [11] and then identifies the established model by the sample correlation coefficient [11]. In addition, they also proposed a set of modeling, estimation, testing, and control methods [11].

2. Autoregressive Conditional Heteroscedasticity Model (ARCH): the model holds that the sequence has no correlation, and its square or absolute value is sequence dependent[12].

Modeling and forecasting stock prices are vulnerable to interference by linear regression. When there is more data or noise, it is often unable to achieve the desired results, so this method is often used to forecast the short-term stock price [3]. In the actual stock market, the collected data often contain a lot of noise and uncertain factors. In the case of a long forecasting period, the forecasting deviation of the linear model increases due to its limitations. In this context, researchers use nonlinear models to develop different machine learning methods based on the principles of neural networks and decision trees and successfully apply these methods to stock forecasting [2].

1.3 SMEs Stock Price Forecast Machine Learning Method

Single and combined machine learning methods are two ways to use machine learning algorithms to study the financial market. Random forest (RF), artificial neural network (ANN), adaptive lifting (AdaBoost), and gradient lifting decision tree (GBDT) are all single machine learning methods with extensive and high representation. The combinatorial machine learning method is often used to forecast and analyze specific problems. Hassan et al. [13] argued that the use of combination algorithms could better forecast financial behavior. However, considering the problem of reducing the listing cost of small and medium-sized enterprises, the system chooses a single machine learning method to forecast the stock price of enterprises.

(1) Random forest (RF). The random forest algorithm is an ensemble algorithm developed based on a decision tree algorithm. The steps are as follows: first, random sampling of all the training samples is carried out to form multiple training sets, including the original samples, and the decision tree is drawn based on the data of the training set, and n features are randomly selected for the drawing of each decision tree. Subsequently, the feature selection is done to select the most feature-generating node so that each tree can grow fully without pruning. Finally, a simple vote is carried out to output the classification with the highest number of votes [14].

(2) Artificial Neural Network (ANN). Compared with the linear regression model, ARIMA, and other traditional methods [15], the neural network machine learning algorithm performs better regarding stock price forecasting. Based on the neural network algorithm, the rate of return of the Japanese stock market is studied, and it can be concluded that the forecasting ability of the artificial neural network is better than the traditional BP algorithm [16].

(3) Adaptive Boosting (AdaBoost). In forecasting the risk of derivative financial instruments, the AdaBoost model has a good performance [17]. Kim et al. (2014) [18] forecast the stock price by integrating emotional factors into the research framework and using the adaptive promotion model. The results show that the AdaBoost model with emotional factors has a better forecasting ability.

(4) Gradient Boosting Decision Tree (GBDT). It has higher accuracy than the traditional linear regression model following time series [19]. In addition, higher forecasting accuracy and classification authenticity are also outstanding advantages of the model [20].

In 2007, random forest technology was used to determine the enterprise credit evaluation index system [21]. Some researchers have constructed a more mature model to evaluate enterprise credit, enabling the enterprise credit evaluation index system to be established more quickly [22]. The fund heavy stock forecasting model based on random forest machine learning was established in 2008[23]. In 2018, a scholar took the random forest as the algorithm principle, selected the experimental subjects for the 500 constituent stocks of the China Securities Exchange

from 2010 to 2017, and constructed a quantitative stock selection model with 22 technical indicators as influencing factors. Subsequently, the model is used to forecast the key information of the stock market, such as stock rises, and several stocks with good forecasting results are invested as portfolios. Higher-than-average returns are obtained[24]. In 2019, a scholar combined random forest with principal component analysis to construct a multi-factor stock selection strategy model, which uses principal component analysis to reduce the dimension of features. Subsequently, the model's parameters are optimized, and the stocks with the highest performance are selected as the combination; the results reported that these combinations perform better than traditional stock selection [25]. In addition, a multi-factor stock selection model based on the random forest machine learning method was constructed in 2020. The experimental subjects of this model are stocks of the CSI 300 index from 2012 to 2018. After stock selection and investment according to the forecasted results, a higher rate of return is obtained [26]. Based on predecessors, this work is still based on a random forest machine learning method to study and forecast the stock price trend of SMEs.

1.4 Research Objectives

Much attention has been paid to the quantitative application of machine learning in financial quantification, especially in the stock market. For many small and medium-sized innovative enterprises in the financial market, machine learning can reduce enterprises' financing cost, clarify the enterprises' value, and provide investment references for investors. More importantly, it can reduce the risks faced by investors and prevent the risks brought to the stock market due to information asymmetry to a certain extent. Thus, the research carried out through combining a random forest algorithm, a highly representative algorithm model in machine learning, and stock price forecasting is extremely necessary and has obvious practical significance.

1.5 Research Methods and Ideas

There are two kinds of research methods.

Literature research method: it makes a deep study of the research object by searching, screening, analyzing, and comparing the relevant literature. In the research process, we actively consult the literature in various ways and systematically study the relevant theories by combining online and offline data and information.

Quantitative analysis: compared with qualitative analysis, it highlights the dependence on data, a technology that uses mathematical and statistical models to measure and analyze the research object, and researchers can draw research conclusions through determined values. The random forest's machine learning algorithm belongs to the quantitative analysis category.

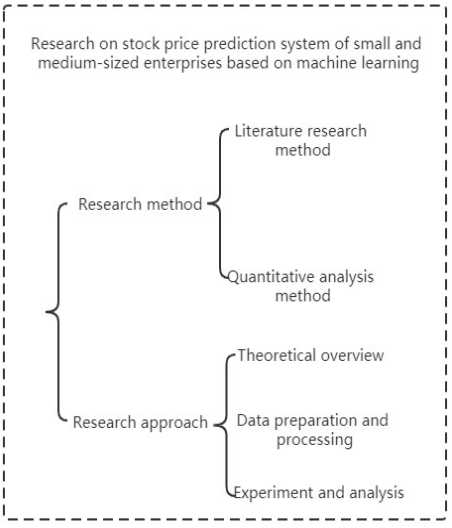


Fig. 3 Overall Flowchart

Overall research idea: A brief introduction of the principle of the decision tree, a detailed description of the random forest theory, the analysis and processing of the experimental data used in the system, and the visual analysis of the results. The specific research ideas are as follows:

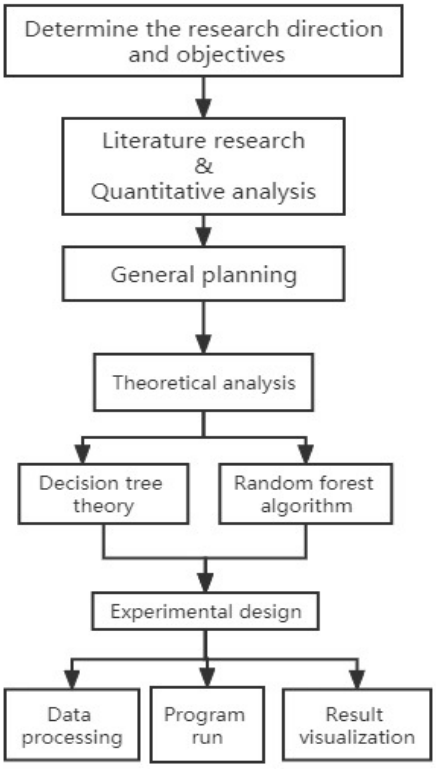


Fig. 4 Research Ideas

Chapter 2 Overview of Random Forest Machine Learning

2.1 Decision Tree Algorithm

A decision tree is essentially a flow chart named because its shape is tree-shaped, showing a series of decision results. Decision trees can be used for classification, decision-making, research, and analysis. Each branch of it represents a possible decision, result, or response. The furthest branch on the tree represents the final result. The main advantage of the decision tree is that it is easy to follow and understand.

The decision tree consists of three main parts: roots, leaves, and branches, in which roots and leaves are nodes. The starting point of the decision tree is the root node (shown in the following figure). The branches show the direction of different categories, which are displayed as arrows of connected nodes in the graph.

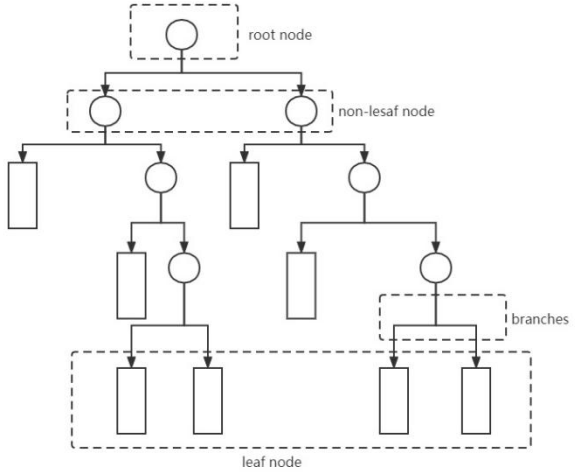


Fig. 5 Decision Tree

One of the important advantages of the decision tree, a machine learning algorithm, is that it is highly interpretable. The decision tree algorithm does not need a lot of data processing, standardization processing, and scaling processing in advance, which greatly reduces the workload of building the model. In addition, in the processing of variables, the decision tree can deal with both discrete and continuous variables. Any missing values in the raw data collected will not affect the decision tree [27].

It is worth noting that the advantage of decision trees in regression is not obvious. If the data samples used change significantly, the noise will affect the calculation of the decision tree. This suggests that to use the decision tree to deal with the classification problem, it is necessary to ensure the number and proportion of training sets, but even if the amount of data of the training set is guaranteed, the decision tree is still prone to problems when dealing with too many classification problems. If the data are not discretized correctly, the results obtained by the decision tree algorithm are likely to have large errors. In contrast, the random forest algorithm has a good effect on avoiding over-fitting phenomena and better performs classification problems[27].

2.2 Random Forest Algorithm

The random forest (RF) algorithm was proposed by Breiman[28] in 2001. He combined the Bagging method with the CART decision tree to form a new ensemble classification algorithm.

After the decision tree sample set and node-set are extracted, several decision tree models form a random decision tree forest. If a sample is entered at this time, each decision tree in the random forest will make a classification judgment and finally get the unique result according to the principle of the minority obeying the majority[29].

The performance of the random forest model needs to be evaluated after the successful construction of the random forest model, and the generalization error is often used to measure the performance of the random forest model. The performance of the model is inversely proportional to the generalization error. In other words, the lower the generalization error is, the higher the model performance is and the higher the classification accuracy is, and vice versa[29].

The performance index of generalization error is mainly affected by three factors: the number of trees, the correlation of trees, and the accuracy of tree classification. 1. When the number of decision trees in random forest increases, the generalization error of the random forest tends to decrease and finally approaches a constant value. 2. When the correlation between trees is large, the generalization error will be very high. 3. When the classification accuracy of trees is higher, the generalization error of random forest is lower.

Chapter 3 Data Analysis and Processing

3.1 Indicator Selection

In the process of forecasting the stock price of small and medium-sized enterprises, we focus on the indicators of Market indicators, Valuation indicators, and Profitability indicators. The details are as follows:

Table 1 Types and names of indicators

Category	Name
Market Indicators	Annual Turnover Rate
	Annual Amplitude
Valuation indicators	Price-Earnings Ratio
	Price to Book Ratio
	Market Sales Ratio
	Price to Cash Flow Ratio
Profitability indicators	Return on Net Assets
	Rate of Return on Total Assets
	Sales Profit Margin

Turnover rate: reflects the frequency of changing hands in stock trading. The higher the index, the more active the stock trading.

Stock amplitude: reflects the activity of the stock to a certain extent.

Price-earnings ratio: The ratio of a company's market capitalization to earnings, one of the most commonly used indicators of stock prices.

Price to book ratio, Market sales ratio, and Price to Cash Flow Ratio: Net assets, sales, and cash flow are all important financial indicators of listed companies. The ratio of the company's market value to it can be used to measure the stock price level of the company. The high ratio of this kind of company indicates that its share price level is relatively high, reflecting investors' strong confidence in the company or the existence of a certain bubble in the stock price.

Return on net assets: reflects the return on investment of shareholders.

Rate of return on total assets: the total liabilities are also included in the denominator to measure the level of return on the company's total assets, reflecting the overall operating capacity of the company.

Sales profit margin: It measures the level of profitability per unit of sales. If the index is low, it means that the profit of the company's products or services is low, and whether the pricing is too low or the cost control is poor should be considered.

3.2 Data Collection and Processing

Part of the research data comes from the Wind financial terminal, and part is provided by the author's cooperative company, in which the Wind financial terminal is the main data source of this study. In terms of time, the stock market data of listed companies in Shanghai and Shenzhen from 2015 to 2022 are selected, ensuring the validity and accuracy of the collected stock market data.

The relevant indicators are selected from the collected data, including opening price, closing price, highest and lowest price, turnover rate, amplitude, etc.. Still, these original sample data are missing in some years in order not to affect the operation of the random algorithm, delete the row where the null value is located. Finally, after eliminating invalid data, the remaining data sets are classified according to the training set and the testing machine and classified according to the proportion of the training set and the testing machine at 9:1.

Chapter 4 Experiment and Analysis of Stock Price Forecasting for SMEs Based on Random Forest Algorithm

4.1 Random Forest Model Construction

Random forest (RF) is composed of multiple decision trees. The sample data set is extracted from the original data set, and the decision tree algorithm processes the sample data set to output the results. Finally, the only final result is obtained by voting.

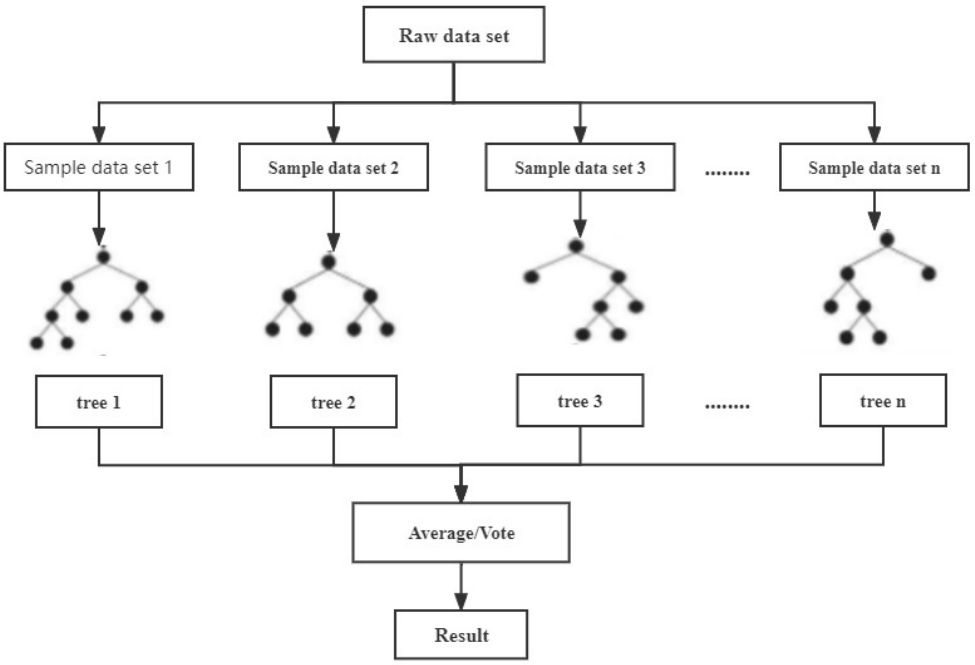


Fig. 6 Random Forest

To reduce the generalization error as much as possible, we construct as many decision trees as possible in the random forest model to make the generalization error tend to a constant value. In addition, the correlation and classification accuracy of decision trees should be improved, and each decision tree should be constructed according to the principles of "data random" and "feature random" to avoid over-fitting and improve the generalization ability.

4.2 Model Building

Python is used to build a stock price forecasting model in the development environment.

4.2.1 Generation of Stock Derivative Variables

First, the database that needs to be built is introduced, including the basic data of the stock (opening price, highest price, percentage of stock price change, etc.). The specific data is shown in the following figure:

date	open	close	high	low	volume	code	close-open	high-low	pre_close	price_change	p_change
2015-01-05	8.168	8.688	9.068	7.998	6560835.0	000002	0.063663	0.133783	NaN	NaN	NaN
2015-01-06	8.378	8.138	8.766	7.828	3346346.0	000002	-0.028646	0.120082	8.688	-0.55	-6.330571
2015-01-07	8.038	8.008	8.278	7.778	2642051.0	000002	-0.003732	0.064284	8.138	-0.13	-1.597444
2015-01-08	8.098	7.368	8.148	7.238	2639394.0	000002	-0.090146	0.125725	8.008	-0.64	-7.992008
2015-01-09	7.318	7.228	7.998	7.068	3294584.0	000002	-0.012298	0.131579	7.368	-0.14	-1.900109

Fig. 7 Basic Stock Data

Then, a simple derived variable is generated with the following code:

```
df[' close-open'] = (df[' close' ]—df[' open']) / df[' open' ]
df[' high-1ow'] = (df['high'] —df[' 1ow']) / df[' low' ]
df[' pre_ _close' ] = df[' close' ]. shift(1)
df[' price_ change' ]二df[' close' ]—df[' pre_ close' ]
df['p_ change'] = (df[' close']—df[' pre_ close' ]) / df[' pre_ _close'] * 100
df. head()
```

(Generate simple derived variables)

Furthermore, the rolling function generates the 5-day moving average MA5 and the 10-day moving average MA10 of the stock price.

```
df['MA5'] = df[' close']. rolling (5). mean()
df[' MA10'] = df[' close' ]. rolling (10). mean()
df. head(10)
```

(Use the rolling function)

After that, the TA-Lib library is used to generate the relative strength index SSI value.

```
import talib
df[' RSI'] = talib. RSI(df[' close'], timeperiod=12)
df. tai1(5)
```

(Generate relative strength indicator RSI value)

RSI value can reflect the contrast of stock price rise and fall in the short term, which has important reference significance for judging the rise and fall of stock price. The higher the RSI value, the stronger the rising trend relative to the falling trend, and the stronger the upward trend of the stock price; on the contrary, the weaker the rising trend relative to the falling trend, the downward trend of the stock price. The calculation formula is as follows:

$$RSI = \frac{NDailyaveragerisingprice}{sDailyaveragerisingprice + NDailyaveragefallingprice} \times 100\%$$

Then, the TA_Lib library generates the mom value, which can reflect the rising and falling speed of the stock price over time. The formula is as follows: moving average EMA.

$$MOM = C - C_n$$

C represents the closing price, and C_n represents the closing price n days ago.
The TA_ lib library is used to generate exponential moving averages EMA.

```
df[' EMA12'] = talib. EMA(df[' close' ], timeperiod=12)
df[' EMA26' ] = talib. EMA(df[' close'], timeperiod=26)
df. tai1()
```

(Generate EMA)

The EMA calculation formula is:

$$EMA_{today} = \alpha Price_{today} + £¨1 - \alpha £©EMA_{yesterday}$$

The TA_ Lib library generates similarities and differences moving average MACD values. MACD is a common indicator in the stock market, and its change represents the change in the market trend. The MACD of different K-line levels represents the buying and selling trend in the current level cycle.

4.2.2 Model Building

The following code introduces the libraries that must be built into the Python development environment.

```
import tushare as ts
import numpy as np
import pandas as pd
import talib
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
```

(Import Library)

Access to data, including stock basic data acquisition, simple derivative variable data construction, average related data construction, etc.

```
import tushare as ts
df = ts.get_k_data('000002',start='2015-01-01',end='2019-12-31')
df = df.set_index('date')


df['close-open'] = (df['close'] - df['open']) / df['open']
df['high-low'] = (df['high'] - df['low']) / df['low']
df['pre_close'] = df['close'].shift(1)
df['price_change'] = df['close'] - df['pre_close']
df['p_change'] = (df['close'] - df['pre_close']) / df['pre_close'] * 100


df['MA5'] = df['close'].rolling(5).mean()
df['MA10'] = df['close'].rolling(10).mean()
df.dropna(inplace=True)


df['RSI'] = talib.RSI(df['close'],timeperiod=12)
df['MOM'] = talib.MOM(df['close'],timeperiod=5)
df['EMA12'] = talib.EMA(df['close'],timeperiod=12)
df['EMA26'] = talib.EMA(df['close'],timeperiod=26)
df['MACD'],df['MACDsignal'],df['MACDhist'] = talib.MACD(df['close'],fastperiod=6,slowperiod
df.dropna(inplace=True)
```

(Collect data)

Then use the following code to extract feature variables and target variables.

```
X = df[['close','volume','close-open','MA5','MA10','high-low','RSI','MOM','EMA12','MACD','M
y = np.where(df['price_change'].shift(-1) > 0,1,-1)
```

(Extract variables)

Divide the training set and the test set according to the time series, and take the first 90% of the data as the training set and the last 10% as the test set. The code is as follows:

```
X_ length = X. shape[0]
split = int(X_ length
*
0.9)
X_ train,X_ test = X[ :split],X[split: ]
y_ train,y_ test = y[ :sp1it],y[split: ]
```

(Divide the Training Set)

Set model parameters. Set the maximum depth max_ _ depth of decision trees to 3, i.e., each decision tree has a maximum of 3 layers; The number of weak learners (i.e., decision tree models) n_ estimators set to 10, that is, there are a total of 10 decision trees in the random forest; The minimum number of samples for leaf nodes min_ samples_ leaf is set to 10, i.e., if the number of samples for leaf nodes is less than 10, the division stops; The purpose of the random state parameter random_ state is to keep the results consistent from run to run.

4.3 Model Testing

4.3.1 Forecast of The Rise or Fall of The Stock Price on The Next Day

The model was used to test the rise and fall of stock prices the next day.

```
y_ pred = model. predict(X_ test)
a = pd.DataFrame ()
a['预测值'] = list(y_ _pred)
a['实际值'] = list(y_ test)
a. head ()
```

(Test Code)

The test results are shown in the table below.

	Forecasted Value	Actual Value
0	-1	-1
1	-1	-1
2	-1	-1
3	-1	-1
4	-1	-1

4.3.2 Model Accuracy Detection

The accuracy of the overall forecasting of the model can be determined in the following code.

```
accuracy = accuracy_ score(y_ _pred, y_ test)
accuracy

model. score(X_ test, y_ test)
```

(Detection Accuracy)

The printout's score is the model's forecasting accuracy for the whole test set. The closer the score is to 1, the higher the forecasting accuracy.

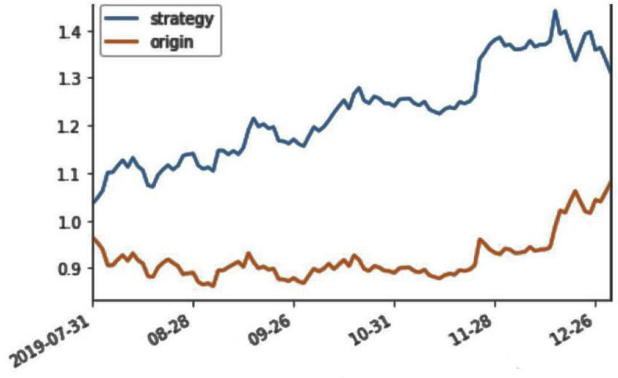
4.3.3 Yield Curve Visualization Results

The yield (net worth curve) in quantitative finance is widely concerned. Compared with the curve drawn by the actual data and the curve drawn by the model forecasting data, the results can be compared before and after using the model.

```
X_test['prediction'] = model.predict(X_test)
X_test['p_change'] = (X_test['close'] - X_test['close'].shift(1)) / X_test['close'].shift(1)
X_test['origin'] = (X_test['p_change'] + 1).cumprod()
X_test['strategy'] = (X_test['prediction'].shift(1) * X_test['p_change'] + 1).cumprod()
X_test[['strategy','origin']].dropna().plot()
```

(Comparison of Results)

The visualization results are shown in the following figure. The upper curve is the yield curve obtained according to the model, and the lower curve is the yield curve of the stock itself; it is obvious that the return obtained by the model is optimal.



(Comparison of visual results)

Chapter 5 Summary and Prospects

Taking the demand of SMEs as the cut-in point, this work analyzes the method of machine learning to forecast stock prices and the current situation at home and abroad and chooses the random forest (RM) machine learning method, a representative ensemble model, to forecast stock prices. The random forest model integrates many advantages of the decision tree model and avoids the shortcomings of the decision tree model, which is easy to over-fit, so it is extensively applied in the field of quantitative finance.

In the process of research, we formulate research objectives, select research methods, conduct more in-depth research and study the principle of random forest algorithm, select the indicators that determine the stock price combined with the information of the stock market, finalize the data set, build the model and training model based on Python as the development platform, compare and analyze the model effect, and draw a conclusion. But the stock market changes a lot, and there are many restrictions on stock trading. Accordingly, the author will constantly improve the model according to the changes in the stock market and adjust the model in time according to the market demand.

References

[1] WANG Lei, XIE Mingzhu. Stock Price Forecasting Of Listed Innovative Small And Medium-Sized Enterprises - Machine Learning Algorithm Based on Bayesian Optimization[J]. Journal of Jilin University of Business, 2023,39(02):79-88. DOI:10.19520/j.cnki.issn 1674-3288.2023.02.010.

[2] LIN Sheng, QI Ke, WEI Kaicong, et al. Review of Machine Learning In Stock Price Forecasting[J]. Economist, 2019,No.361(03):71-73+78.

[3] QIAN Qimiao, ZHANG Duo, WANG Yingjian, et al. Application of Machine Learning In Stock Price Forecasting[J]. China Market, 2022, No.1120(21):7-10. DOI:10.13939/j.cnki.zags. 2022.21.007.

[4] Xia Xue. An Empirical Study on Markowitz Portfolio Theory based on Matlab[J]. Bohai Rim Economic Outlook, 2022, No.339(12):145-147. DOI:10.16457/j.cnki.hbhjjlw.2022.12.054.

[5] Gurisi. Application of Machine Learning in Investment Portfolio [J]. Industrial Innovation Research, 2023, No.106(05):127-129.

[6] NI Xuanming, QIU Yuning, ZHAO Huimin. High-Dimensional Sparse Portfolio Optimization Based on Factor Characteristics[J]. Systems Science and Mathematical Sciences, 2021, 41(10): 2716-2729.

[7] QIAN Long, WEI Jiang, ZHAO Huimin, et al. Portfolio optimization based on AdaBoost[J]. Systems Science and Mathematical Sciences, 2022,42(02):271-286.

[8] Markowitz H. Portfolio selection. The Journal of Finance,1952,7(1):77-91.

[9] ZENG Yongquan, ZHANG Peng. Multi-stage Mean-Semi-Absolute Deviation Portfolio Optimization With Entropy constraint[J]. Chinese Journal of Management Science, 2021, 29(09):36-43. DOI:10.16381/j.cnki.issn1003-207x.2019.0897.

[10] ZHANG Peng. Discrete Approximation Iterative Method for Multi-Stage Mean-Mean Absolute Deviation Portfolio[J]. Journal of Systems Management,v2010,19(03):266-271.

[11] Wu Yuxia, Wen Xin. Short-Term Stock Price Forecasting Based on ARIMA Model[J]. Statistics and Decision, 2016, No.467(23):83-86.DOI:10.13546/j.cnki.tjyjc.2016.23.051.)

[12] Yang Qi, Cao Xianbing. Stock price analysis and forecasting based on the ARMA-GARCH model[J].Mathematics in Practice and Theory,2016,46(06):80-86.

[13]Hassan,M.R.,Nath ,B.,Kirley,M.A fusion model of HMM, ANN, and GA for Stock Market Forecasting[J]. Expert Systems with Applications, 2007,33(1):171-180.

[14] Ma Mengchen. Research on Credit Risk Assessment of Small And Medium-Sized Enterprises Based On Random Forest Method [J]. Special Economic Zone Economy, 2023, No.408(01):141-144.

[15] Mostafa, M.M.Forecasting Stock Exchange Movements Using Neural Networks: Tems With Applications, 2010,37(9):6302-6309

[16] Qiu, M.Song, Y., Akagi, F. Application of Artificial Neural Network for The Forecasting of Stock Market Returns: The Case of The Japanese Stock Market[J]. Chaos, Solitons & Fractals, 2016,85:1-7.

[17] ZHANG Jie, SUN Yuyao. Risk Forecasting Of Derivative Financial Instruments Based On Adaboost Combination Algorithm[J]. Statistics and Decision, 2012, No.355(07):41-44. DOI:10.13546/j.cnki.tjyjc.2012.07.006.

[18] Hassan, M.R., Nath, B., Kirley, M.A fusion model of HMM, ANN, and GA for Stock Market Forecacasting[J]. Expert Systems with Applications,2007,33(1):171-180.

[19] ZHANG Xiao, WEI Zengxin, YANG Tianshan. Application of GBDT Combinatorial Model in Stock Forecasting[J]. Journal of Hainan Normal University(Natural Science Edition), 2018, 31(01):73-80.

[20] LIU Pingshan, ZENG Ziming. Financial Credit Risk Evaluation of Pharmaceutical Supply Chain Based on GBDT[J]. Friends of Accounting,2021, No.664(16):24-31.

[21] Ross, S. A. Return, risk, and arbitrage. In: Friend, I., and Bicksler, J., Eds., Risk and Return in Finance, Ballinger, Cambridge, 1976.

[22] Eugene F. Fama, Kenneth R. French. The Cross Section of Expected Stock Returns[J]. The Journal of Finance, 1992, 47(2):427-465.

[23] Mark Carhart. On Persistence in Mutual Fund Performance[J]. The Journal of Finance, Vol.52, No.1. (Mar. 1997), pp. 57-82.

[24] Quinlan, J. R. Discovering Rules by Induction From Large Collections of Examples. In D. Michie (Ed.), Expert systems In The Micro Electronic Age. Edinburgh University Press, 1979.

[25] Quinlan, J. R. C4.5: Programs for Machine Learning[M]. San Francisco: Morgan Kaufmann, 1993.

[26] Markowitz, H. M. Portfolio Selection[J]. The Journal of Finance, 1952, 7(1):77-91.

[27] Wang Jingxiang. Computer Programming Skills and Maintenance, 2022, No.446(08): 54-56+72. DOI:10.16184/j.cnki.comprg.2022.08.043.

[28] Markowitz, H. M. Portfolio Selection[J]. The Journal of Finance, 1952, 7(1):77-91.

[29] GAO Ziyi. Research on Stock Price Trend Forecasting Based on Random Forest[D]. China University of Political Science and Law, 2021. DOI:10.27656/d.cnki.gzgzu.2021.000061.

Appendix

```
import tushare as ts
import numpy as np
import pandas as pd
import talib
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
```

```
import tushare as ts
df = ts.get_k_data('000002',start='2015-01-01',end='2019-12-31')
df = df.set_index('date')


df['close-open'] = (df['close'] - df['open']) / df['open']
df['high-low'] = (df['high'] - df['low']) / df['low']
df['pre_close'] = df['close'].shift(1)
df['price_change'] = df['close'] - df['pre_close']
df['p_change'] = (df['close'] - df['pre_close']) / df['pre_close'] * 100


df['MA5'] = df['close'].rolling(5).mean()
df['MA10'] = df['close'].rolling(10).mean()
df.dropna(inplace=True)


df['RSI'] = talib.RSI(df['close'],timeperiod=12)
df['MOM'] = talib.MOM(df['close'],timeperiod=5)
df['EMA12'] = talib.EMA(df['close'],timeperiod=12)
df['EMA26'] = talib.EMA(df['close'],timeperiod=26)
df['MACD'],df['MACDsignal'],df['MACDhist'] = talib.MACD(df['close'],fastperiod=6,slowperiod
df.dropna(inplace=True)
```

```
X = df[['close','volume','close-open','MA5','MA10','high-low','RSI','MOM','EMA12','MACD','M
y = np.where(df['price_change'].shift(-1) > 0,1,-1)
```