Research on Stock Return Forecasting Methods based on Time Series Models

Xiyuan Jiang*

International Business School, Dalian Minzu University, Dalian, China

*Corresponding author: sxq042@student.bham.ac.uk

Abstract. Accurately predicting the trend of stock return rate is a hot research issue. With the development of artificial intelligence, machine learning, big data and other technologies, it brings new potential to the prediction of the stock market. In order to accurately predict the trend of stock return, this paper mainly constructs the time series ARMA model and random forest model, uses the stacking method to fuse the models, and predicts the daily return of Yangtze River Electric Power stock. The final fusion model has an MSE of 1.757 on the training set and 1.274 on the test set. The overall prediction error of the model is within an acceptable range. At the same time, the fused model can weaken the problem of underfitting of a single model, which provides a valuable reference for model optimization research.

Keywords: Stock prediction; ARMA model; Random Forest; stacking.

1. Introduction

In financial markets, whether securities markets or other, all investors strive to precisely estimate the price trend of financial items. A large number of investors and scholars have tried to use various prediction models to predict the future of assets [1]. Among them, stock data is a typical time series data, and the theory of time series model is relatively solid and mature, so the example of using time series model to predict stock trend is more extensive. Many time series models, such as ARMA and ARIMA, can be used to predict the return rate of stocks, and the future stock price can be deduced according to the predicted return rate, so as to achieve the purpose of predicting the stock price [2-5]. So the core task of this paper is to build a model to predict stock returns.

Among the many forecasting methods and models for time series, Autoregressive (AR), moving average (MA), ARMA (combining autoregressive and autoregressive), and ARIMA are the primary models. However, time series models have the obvious disadvantage of being more accurate in short-term forecasting, but perform poorly in long-term forecasting, which is related to the principle of the model itself. Time series models rely only on the past characteristics of the series and do not involve external characteristics. The stock market movements, on the other hand, are usually related to external characteristics, such as the recent plunge in stocks due to interest rate hikes by central banks [6]. The level of interest rates and the movement of interest rates are external features, which cannot be captured by time series models. Therefore, to improve the accuracy of forecasting and to construct a more accurate model, this paper uses two sub-forecasting models: the ARMA model as well as the random forest model to forecast stock returns, combining both internal historical trends and external characteristics. The internal historical trends can be fitted using a time-series model, while the external characteristics can be analyzed by a machine learning mathematical model. And the two models above are model fused using the stack method to obtain a more robust forecasting model.

This paper is divided into four main sections. The first chapter is the introduction of this chapter, which mainly introduces the research background and significance. The second chapter introduces the model theory used in this paper. The third part is an example test, taking the daily transaction data of the Yangtze River Electric Power as the example data, using the above model to predict. In section IV, the prediction effect is analyzed and the experimental conclusions are drawn.

2. Methodology

2.1 ARMA

Before introducing the ARMA model, we introduce two basic time series models: the autoregressive model (AR) and the moving average (MA).

(1) AR models

AR models are also known as autoregressive models. However, unlike the general linear regression model, which uses the independent variable X to predict the dependent variable Y, the autoregressive regression uses its own data from the past to regress with the present data. For an autoregressive model of order p, the mathematical equation is as follows:

$$X_t = c + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + \varepsilon_t \quad (1)$$

Where is a constant term, (i = 1,2, ..., p) is the undetermined coefficient (autoregressive coefficient) of the model, is a random disturbance term with zero mean and non-zero variance, and is a time series.

If the back-shift operator is defined as B,that is:

$$BXt = Xt—1\ BkX_t = X_t—k1 \quad (2)$$

Then Equation (1) can be formulated as follows:

$$X_t = (\varphi 1B + \varphi 2B2 + \ldots\ldots + \varphi PBp)X_t + \varepsilon_t \quad (3)$$

AR models are widely used to predict economic and natural aspects, and the scope of use is quite extensive, even can be said to have no limit. However, in general, the time series is required to have high autocorrelation, otherwise the prediction error of the model will be large and the model will lose its significance.

(2) MA Model

The MA model is also known as moving average model. The MA model considers that the fluctuation of the series is mainly caused by white noise, which means that the series is stationary. For an autoregressive model of order q, the mathematical equation is as follows:

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \quad (4)$$

Which is assumed to be a normally distributed white noise sequence with mean 0, while is a stationary sequence, which means is the mean of the stationary sequence. ARMA model can be guessed from the structure of the name that the model is composed of AR model and MA model. There are two parameters of the ARMA model, which correspond to the orders p, q of the AR model and the MA model,

respectively. For the ARMA(p,q) model, the model equation is:

$$X_t = c + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \quad (5)$$

ARMA has the properties of both AR and MA models, capturing both the time series' own trend and white noise fluctuations. Given this property, ARMA is commonly used in economic and business forecasting trends.

The modeling of the ARMA model generally follows the process shown in Figure 1:

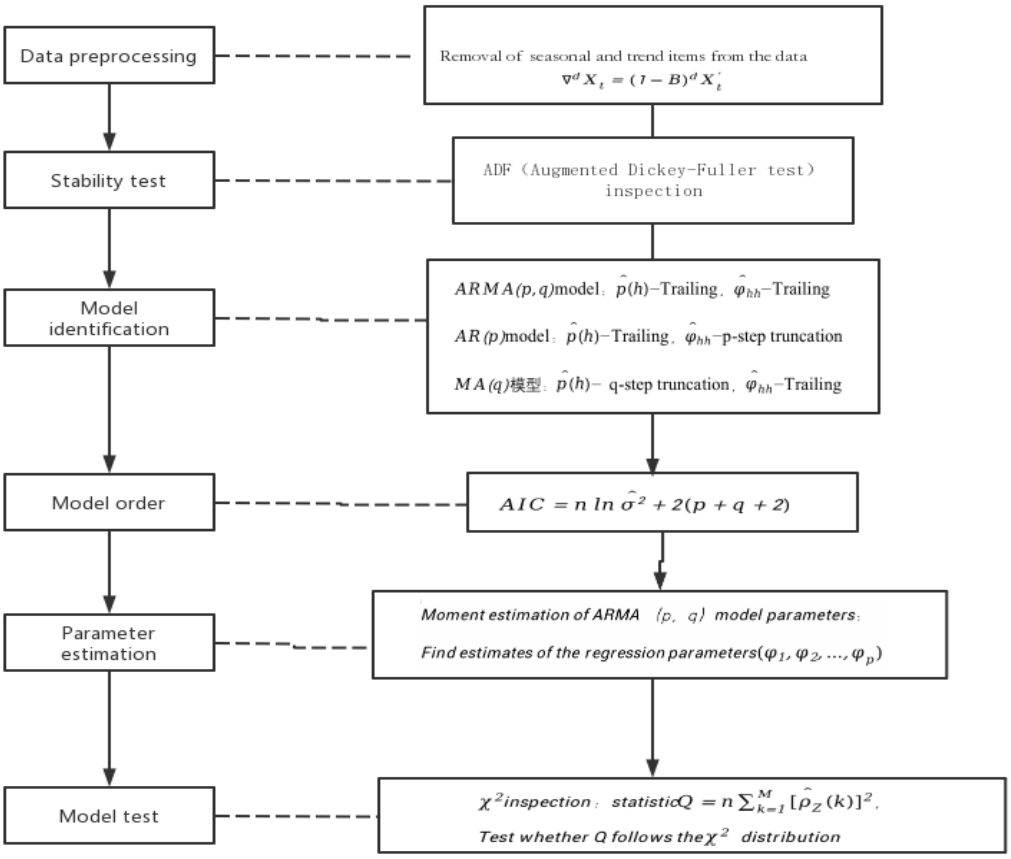

Figure 1. ARMA modeling flow chart

In practice, individual steps of the above process can be omitted depending on the source of the data and the model objectives. For example, in the data forecasting step, the first step can be skipped when the time series is non-seasonal.

2.2 Random Forest

Before understanding random forests, we first need to understand its constituent base model-decision trees. Decision trees are the basis of several supervised learning models, such as random forests, GBDT, etc [7-9]. They are commonly used for business decisions because of their strong explanatory power and the ability to determine the importance of features. A decision tree is a tree where each branch in the tree represents a different alternative and each leaf node represents a decision option.

To determine whether a feature node is the best division node, multiple indicators are usually used. For example, the feature selection criterion for ID3 decision tree is information gain, the feature selection criterion for C4.5 is information gain ratio, and CART decision tree uses Gini coefficient as the feature selection criterion. In general, decision trees that have not been pruned are prone to overfitting. We usually use early stopping (pre-pruning) and post-pruning to solve it. According to the general logic, a random forest includes at least many trees, the decision trees are constructed randomly, and the decision trees inside the forest are not correlated with each other [10-12].

As for the randomness of the construction of the decision tree, it can be reflected in the construction process of the random forest according to the following:

Step 1: For each of the N examples in the training set, we don't use those N examples directly as training; Instead, we choose N replacement samples at random (one at a time, and then go back to continue the selection). The decision tree is trained using these N samples.

Step 2: For a training set of M features, at each node of the decision tree, We choose M attributes at random from these m attributes, and then apply some approach (e.g., information gain) to choose one characteristic as the optimal split feature for that node.

Step 3: In step 2, each node is split until it can no longer be split to form the decision tree.

Step 4: Steps 1-3 should be followed to create a large number of decision trees and a random forest.

For the random forest prediction, we use a vote, where each decision tree has a consistent vote. For classification tasks, the label with the highest occurrence rate is used as the prediction result. For regression, the average of all the tree predictions is used as the final prediction of the random forest.

An ensemble model is a random forest. In comparison to a single decision tree, random forest has the characteristics of decision tree at the same time, and usually has more robust performance and generalization ability.

2.3 Model fusion

For the same prediction tasks, we can build a number of different models, the base learning can use a variety of methods continue to integrate a strong learning, one of the methods for stacking, the idea is to use learning of the prediction results as new features, and USES the model to fit these characteristics, get better learning. The thought flow is shown in Figure2:

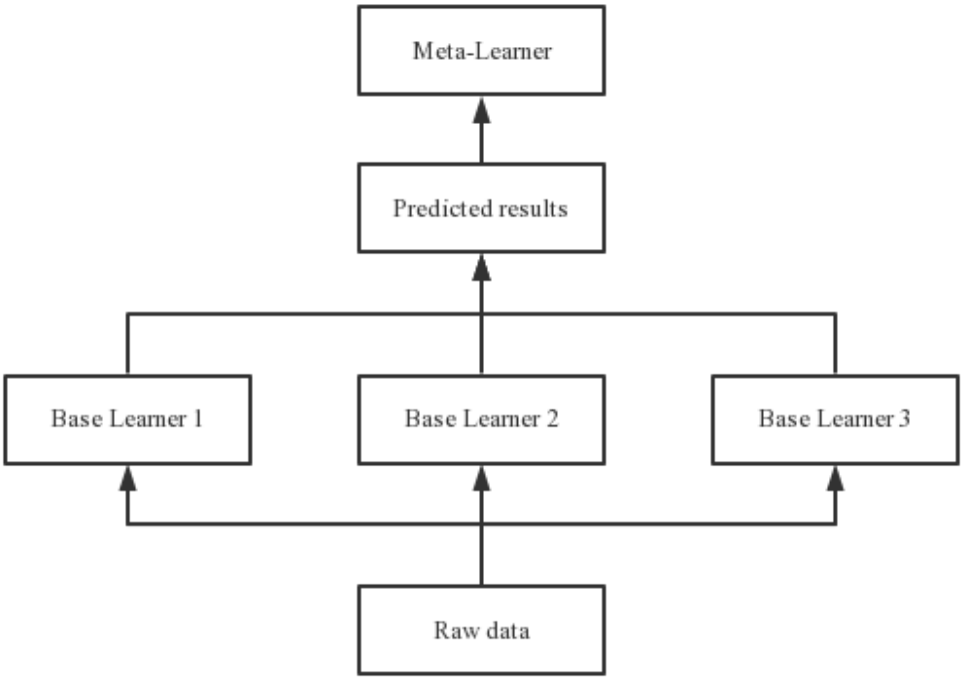

Figure 2. Model fusion flow chart

The fusion of ARMA model and random forest using stacking can improve the prediction accuracy of the model.

3. Experiment

Experiments were carried out mostly using real-world data to validate the efficacy of the suggested strategy in this work.

3.1 Experimental data introduction:

The data used in this experiment are the daily trading data of Yangtze River Power (stock code: 600900.SH), observed from the beginning of 2020 (the first trading day) to the present (late October

2022), and the raw data include daily opening price, closing price, highest level, lowest price, turnover amount (in million yuan), share price increase or decrease compared with the previous trading day, and increase or decrease rate (in percent). Where stock prices are all pre-compounded data. The data were obtained online using python through the API of tushare financial data platform. The first five rows of data were obtained as in Table 1:

Table 1. Data information

trade_date	ts_code	open	high	low	close	pre_close	change	pct_chg	vol	amount
2020-01-02	600900.SH	16.6218	16.7478	16.5587	16.6308	16.5497	0.0811	0.4900	191284.92	353770.322
2020-01-03	600900.SH	16.6578	16.7929	16.6128	16.7208	16.6308	0.0900	0.5412	154383.77	286562.117
2020-01-06	600900.SH	16.6938	16.7568	16.2526	16.3877	16.7208	-0.3331	-1.9921	415505.82	757921.841
2020-01-07	600900.SH	16.4057	16.4417	16.2976	16.3787	16.3877	-0.0090	-0.0549	232106.68	421457.432
2020-01-08	600900.SH	16.2796	16.3607	16.1986	16.2526	16.3787	-0.1261	-0.7699	264409.91	476879.600

Exploratory Analysis: Since the data are normalized by the platform, the data quality is intact, there are no missing values, etc., and can be used for direct analysis and modeling.

3.2 Experimental design

In this paper, we designed the following experiments:

(1) First, correlation analysis of each numerical variable was conducted to examine the relationship between the variables and the rate of return.

(2) Use ARMA and RF models for stock return forecasting and compare the forecasting performance of different models.

In order to properly assess the model's performance, the MSE and RMSE are mainly used as indicators in this experiment and are calculated as follows.

MSE indicator calculation formula:

$$\mathrm{MSE} = \frac{1}{\mathrm{m}}\sum_{\mathrm{i}=1}^{\mathrm{m}}(\mathrm{y_i} - \hat{y_i})^2 \tag{6}$$

Where m is the sample count, and $\hat{y_i}$ denotes the ith sample pair of predicted values.

RMSE indicator calculation formula:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{m}}\sum_{\mathrm{i}=1}^{\mathrm{m}}(\mathrm{y_i} - \hat{y_i})^2} \tag{7}$$

3.3 Experimental results

3.3.1 Correlation analysis

As Figure 3 shows the correlation between the variables, it can be seen that the four prices of the stock are almost linearly correlated with each other, while the target variable of our modeling, return, is almost uncorrelated with all other variables.

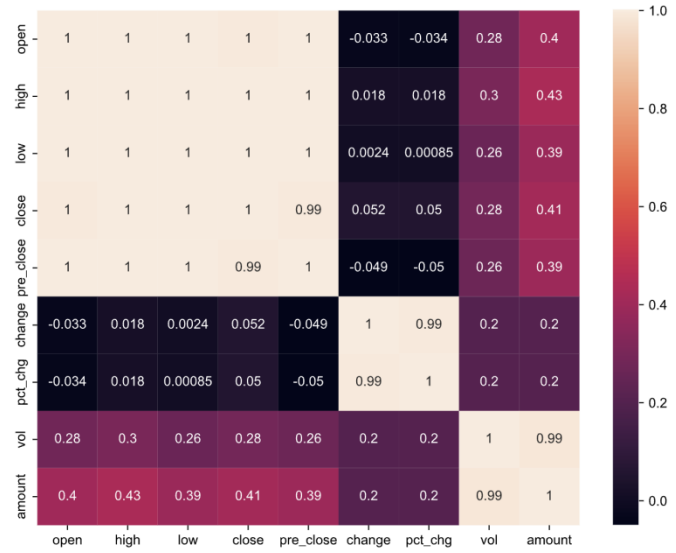

Figure 3. Related Matrix Heat Map

The histogram of the distribution is plotted for the return data in Figure 4. From the shape of the distribution, the returns obey a normal distribution, and in order to confirm this analytical conclusion more rigorously. We use the K-S test to check whether the test returns obey a normal distribution. According to the findings, the test statistic is 0.0452, with a p-value of 0.1265. There is insufficient evidence at the 0.05 significance level to reject the initial hypothesis that the distribution from which the sample data originated is not substantially different from the normal distribution. In other words, the daily stock returns follow a normal distribution.

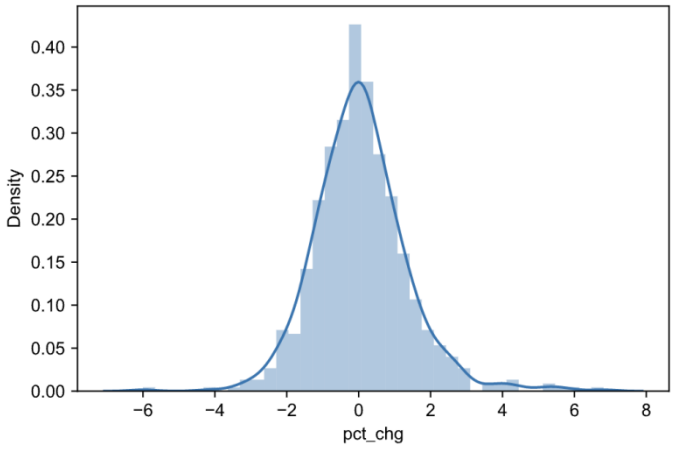

Figure 4. Histogram of daily return distribution

Subsequently, the model of time series ARMA requires stationary series data, and we utilize the ADF test to determine whether the daily return data are stationary. The test's initial premise is that the time series has a unit root and is non-stationary. The alternative theory is that there is no unit root and that the series is smooth, indicating that it lacks a time-dependent structure. The test statistic is -27.25, leading to a p-value of 0.0, according to the computed findings. If the time series is smooth, the original hypothesis can be rejected at a significance level of 0.05. As a result, it may be represented directly without the need for transformation processing such as differencing, etc.

And according to the autocorrelation plot of the series data and the first-order truncated tails of the partial autocorrelation plot (Figures 5), there is no autocorrelation in the returns. Combining the above analysis, we believe that the information of the return series is mainly concentrated on the stochastic disturbance term.

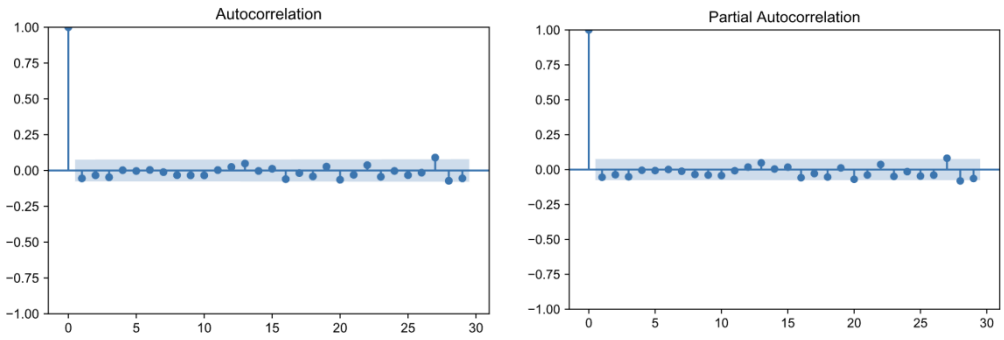

Figure 5. Autocorrelation and partial autocorrelation plots

3.3.2 Return forecast

(1) ARMA

We divide the data into a training set and a test set to make evaluating the model performance easier, where the beginning of 2020 to the end of June 2022 is used as the training set (a total of 593 trading days) and the data from the beginning of July 2022 to the present is used as the test set (a total of 75 trading days).

A grid search technique is used to find the best ARMA(p,q) parameters-model order p and q, and the model with the minimum AIC is chosen as the best model using AIC as the model selection index. The best is ARMA(1,1), and the model summary output is shown in Table 2 below:

Table 2. Model summary output results

	coef	**std err**	z	P>|z|	**[0.025**	**0.975]**
const	0.0706	0.006	11.004	0.000	0.058	0.083
ar.L1	0.9512	0.017	54.503	0.000	0.917	0.985
ma.L1	-0.9999	0.461	-2.167	0.030	-1.904	-0.096
sigma2	1.7395	0.800	2.175	0.030	0.172	3.307

After forecasting using the ARMA model, the residual series is calculated by subtracting the real value from the anticipated value, and the LB test is used to determine whether the residual series is white noise, and if it is white noise, the model performs better, and vice versa. And in our model, for any order of LB test, the p-value is greater than 0.05, implying that the best model is significant for whether the residual series is white noise or not.

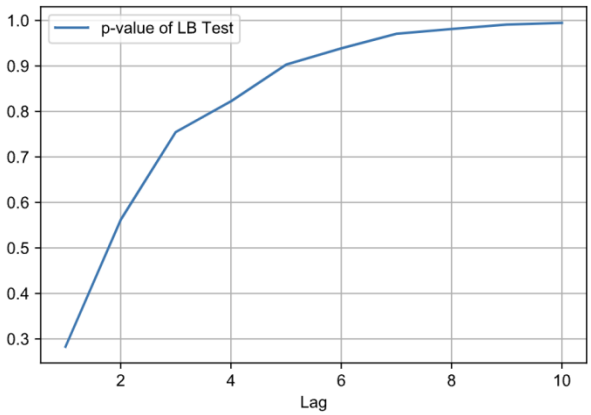

Figure 6. LB test p-values for different lag orders

However, in terms of the prediction results, the model has a large prediction variance, with a 95% confidence interval of almost -3% to 3% for the predicted values, and in practice, CK Power's returns

rarely exceed this volatility. Under the best model, the training MSE is 1.746, while the test MSE is 1.270.

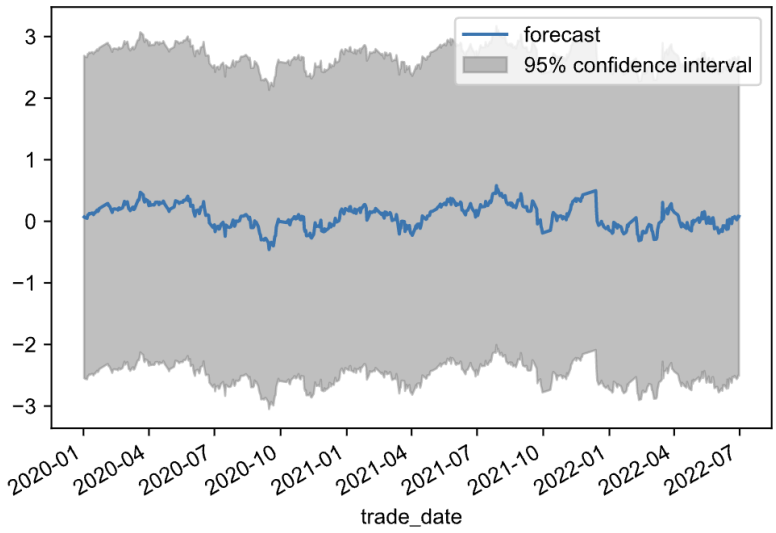

Figure 7. Predicted values and their 95% confidence intervals

(2) Random Forest

The ARMA model is mainly modeled for the sequence's own historical data, and it is difficult to capture non-self signals, so we use the random forest model to capture these non-self signals. First, we need to calculate a large number of trading indicators based on stock prices, such as the moving average of prices, MACD, Boll, immature indicator RSV, KDJ indicator, CCI indicator, RSI relative strength indicator, volume variance, etc., and construct multiple signal dummy variables based on the above indicators, such as whether the moving average is a golden cross or a dead cross, and whether the KDJ indicator is overbought or The KDJ indicator is overbought or oversold. Many stockholders are chartists and trade based on these stock indicators, which can affect stock price fluctuations. Therefore, Random Forest uses the above construction of multiple indicator signal characteristics as input.

Again, the data up to July 2022 is used as the training set, and the data after that as the test set. A crossover 5-fold validation is used, combined with a grid search algorithm to search for the best model parameters. With the optimal parameters, the training MSE is 1.771, while the test MSE is 1.198.

Finally, the projected values from the previous two models were utilized as input characteristics in the fusion model, which used linear regression to generate the final model predictions. The fusion model's MSE on the training set is 1.757, whereas the test MSE is 1.274. Overall the performance of all three models is basically the same, and the test error is significantly lower than the training error, which is underfitting, and the fusion model can attenuate the phenomenon of model underfitting, but the effect is not obvious.

4. Conclusion

In this paper, the daily trading data of Yangtze River Power in early 2020 to present are used as the research sample to construct ARMA and random forest models to predict the stock's return in the next trading day. A grid search algorithm is used to search for the best parameters of each model. For the ARMA model, the best time series model ARMA (1,1) is derived using AIC as the evaluation index of the ARMA model; for the random forest model, the best parameters are searched by combining cross-validation and grid search. After the two best sub-models are derived, the model fusion is performed using the stacking method. The training MSE of the fusion model is 1.759 and the test MSE is 1.274, which has the phenomenon of underfitting, but the fusion model can improve

the phenomenon of underfitting compared with the single sub-model. Therefore, the fusion model is considered more effective than the single model. Meanwhile, the shortcoming of the model is that it can only predict the rise and fall of the next trading day, but lacks long-term prediction. In actual trading, the advantage of the model for short term trading will be relatively obvious. However, the actual effect still needs to be left to the market to verify.

References

[1] Manish Agrawal et al. Stock Prediction Based on Technical Indicators Using Deep Learning Model[J]. Computers, Materials & Continua, 2022, 70(1) : 287-304.

[2] Zou Cunzhu,Luo Jiping,Bai Shengyuan,Wang Yuanze,Zhong Changfa,Cai Yi. Stock Time Series Prediction Based on Deep Learning[C]//.Proceedings of 2019 2nd International Conference on Mechanical,Electronic and Engineering Technology(MEET 2019).Clausius Scientific Press,2019:26-30.

[3] Wennian Yu and Il Yong Kim and Chris Mechefske. Analysis of different RNN autoencoder variants for time series classification and machine prognostics[J]. Mechanical Systems and Signal Processing, 2021, 149

[4] W.T. Ho and F.W. Yu. Predicting chiller system performance using ARIMA-regression models[J]. Journal of Building Engineering, 2021, 33: 101871-.

[5] Fang Zheng et al. Minimum Message Length in Hybrid ARMA and LSTM Model Forecasting[J]. Entropy, 2021, 23(12) : 1601-1601.

[6] Bo Zhang and Joshua C.C. Chan and Jamie L. Cross. Stochastic volatility models with ARMA innovations: An application to G7 inflation forecasts[J]. International Journal of Forecasting, 2020, 36(4) : 1318-1328.

[7] GUO Haofei,DU Jiaxin. Research on the Prediction of Intercity Passenger Waiting Time Based on Random Forest of Survival Analysis[C]//.Proceedings of the World Transport Conference 2022 (WTC2022) (Traffic Engineering and Air Transport Chapter).,2022:24-30.DOI:10.26914/c.cnkihy.2022.019761.

[8] 1VALET and Building Stack Enter Partnership[J]. Wireless News, 2021,

[9] Xu Xin Hong et al. Research on the Comprehensive Evaluation of the Higher Education System Based on FCE and ARMA Models[J]. Complexity, 2022, 2022

[10] Wu Jiujiang et al. Evaluating the accuracy of ARMA and multi-index for predicting winter wheat maturity date.[J]. Journal of the science of food and agriculture, 2021, 102(6) : 2484-2493.

[11] Alaa Kafafi. Grow a decision tree to support decision-making, machine learning[J]. ISE ; Industrial and Systems Engineering at Work, 2019, 51(8) : 40-45.

[12] Chai Weiwei et al. Short-term Load Prediction Based on the Combination of K-means and Random Forest[J]. Journal of Physics: Conference Series, 2022, 2166(1)