Comparison of Future Price Prediction in Context of Machine Learning Approach

Lu Xu[1, *]

[1]Department of Economics, Sichuan University, Beijing, China

*Corresponding author: xulu3@stu.scu.edu.cn

Abstract. Contemporarily, futures markets have become prosper on account of its intrinsic risk hedging function as well as double-side trading advantages compared to Chinese stock market. In this case, finding the suitable and appropriate way to predict future price changes will be a crucial thing for investors and traders of commodity. In context of theoretical analysis in terms of summary of previous models, this article firstly selects features and market information of all trading days from 2016 to 2021 as input. The research sample is the CSI300 stock index futures, and 1-daylow-frequency data is selected. Three price prediction models were built in terms of the data, i.e., linear regression, random forest, and support vector machine algorithms. Accordingly, the random forest algorithm has the lowest error, high accuracy and stability in the prediction of stock index futures price. These results shed light on guiding further exploration of future price prediction based on specific machine learning approach.

Keywords: Stock index futures; Price prediction; Machine learning.

1. Introduction

Futures refers to a contract with a certain physical commodity or financial asset as its subject matter, which has the characteristics of standardization and tradability. In addition, the buyers and sellers of the contract will trade the subject matter of the contract according to the specified price at the time agreed in advance.

Contemporarily, futures trading has developed rapidly in many countries and regions. The target of stock index futures is various stock indexes, which refers to the cash delivery instead of physical delivery by both parties according to the contract price difference after the product expires.

Due to the instability and higher risk of stocks, stock index futures have an indispensable risk avoidance function, which is one of the important reasons for stock market participants to choose it. Although it was born later compared with other products, it became the most successful product since the last century owing to rapid expansion of trading scale and huge market influence [1].

China's futures market took shape and developed later than the international market, Under the condition that commodity futures have a certain foundation, China's first landmark stock index futures variety, i.e., CSI 300 stock index futures, were listed to participate in trading in April 2010. Afterwards, its scale has been expanding continuously. Two years later, with the gradual relaxation of trading restrictions, institutional investors can apply for long or short positions to hedge.

Stock index futures have been popular among all kinds of market participants since their listing. It can not only stabilize the fluctuation of stock prices in a reasonable range, but also better promote market growth through spot hedging of futures. With the development of the market in exploration, many researches related to stock index futures price prediction are useful [2]. Market price is affected by many factors and changes in related conditions and various models are proposed. However, the best and most suitable approaches have not been determined [3].

Up to now, many various machine learning scenarios are constructed to financial asset price forecasting, and achieved considerable research results. Drucker first proposed to predict the prices of various types of securities by combining SVM with regression problem [4]. Based on empirical research, Kim stated that SVM can replace the traditional stock forecasting methods [5]. Liu and Fan proposed that the composite model of two SVM can obtain better futures price prediction results than a single SVM [6]. The prediction object of Liu and Ma is oil futures price [7]. They use the model least square support vector machine to study and analyze oil futures price. The results show that the

prediction performance of this model has good accuracy and fitting ability, and its performance even surpasses that of neural network model. Zhou screened the features of Xinhua FTSE 50 and S&P 500 Index based on the prediction of price trend by support vector machine [8]. Wang used random forest to predict the price changes of CSI 300 Index, and the results were considerable [9]. Blair and Sun predict the rise and fall of different indexes based on several integrated algorithms of random forest, Bagging and Boosting [10]. The research shows that no one algorithm is suitable for all stock indexes, and different integrated algorithms are suitable for predicting different stock index types.

This paper intends to select multiple factors that have an impact on the stock index futures price as characteristics. The research purpose of this paper is to obtain a relatively accurate forecasting model through empirical research. The other parts will be separated as follows. The Sec. 2 will introduce the methodology and the results will be analyzed in the Sec. 3. Finally, the Sec. 4 will give a brief summary of the whole research.

2. Methodology

2.1 Data

The data used in this paper is collected from the Shanghai Futures Exchange. The research period is from January 2016 to June 2022. The subjects of this study are the main contracts of Shanghai and Shenzhen 300 (CSI300) stock index futures. The data frequency is 1 day and 1 day. There are 1557 rows of data in the sample.

This paper selects the basic quotations and technical indicators of 300 stock index futures in Shanghai, Shenzhen and Shanghai as input features: the basic quotation indicators include opening price (OPEN), closing price (CLOSE), highest price (HIGH), lowest price (LOW), volume of transaction (VOLUME), and holding volume (OI); Technical indicators include the obedience rate (BIAS), Bollinger band (BOLL), trend criterion (DMI), smooth moving average (EXPMA), random index (KDJ), moving average (MA), smooth similarity and difference average (MACD), and relative strength index (RSI).

Table 1. Description of scaled data.

	Inclose	Inopen	Inhigh	Inlow	PriceMA5	oi_dev	middle_10_tema	middle7_tema
count	1534	1534	1534	1534	1534	1534	1534	1534
mean	0.0002	0.0001	0.0002	0.0002	3972.4714	29.5248	3973.2865	3973.3518
std	0.0133	0.0127	0.0112	0.01314	670.1616	10997.1657	671.5963	672.3448
min	-0.1064	-0.1085	-0.0918	-0.0982	2874.2000	-44221.0000	2850.1473	2852.4515
25%	-0.0061	-0.0061	-0.0054	-0.0054	3413.5600	-3999.2500	3408.2045	3408.3533
50%	0.0003	-0.0001	-0.0003	0.0006	3856.2800	-591.5000	3853.1160	3857.8179
75%	0.0071	0.0069	0.0058	0.0064	4598.3700	1109.7500	4586.5441	4580.3370
max	0.0716	0.0701	0.0759	0.1364	5708.7600	121304.0000	5758.8531	5697.7897

Since the selected features are all financial time series data, through correlation analysis, independent variables without significant correlation are removed. For the stock index futures price, this paper will use the logarithmic yield to make an empirical analysis, i.e., the closing price is logarithmic poor, the expression is

$$r_t = \ln \frac{p_t}{p_{t-1}} \tag{1}$$

For other characteristic variables, the absolute amount is distorted as

$$y_t = \ln \frac{x_t}{x_{t-1}} \tag{2}$$

Secondly, considering the magnitude differences between the various indicators, if regression processing directly affects the significant situation of the indicators, it is necessary to standardize the sample data before regressing the model. The processed eigenvalues have eliminated autocorrelation and range in [-1,1]. The statistical descirption of the normalized data is given in Table. 1. As shown in Table.1, all the fugures are scaled between -1 and 0,and standard errors are different.

2.2 Models

Ordinary least square is the most common type method to fit linear relationship between variables and factors. The advantage of linear models is that they are simple in form and easy to model, which can be described as

$$f(x) = w_1x_1 + w_2x_2 + \cdots + w_mx_m + b \quad (3)$$

The Random Forest algorithm, containing several decision trees, allows random forests to perform classification and regression processing after training data samples. In context of the existing characteristic variable M, the final results can be derived in various ways. After calculation of the mean value, one derives the final results. The final result is voted for by each decision tree, with only one vote for any tree.

Support Vector Machines (SVM) is one of the latest algorithms in artificial intelligence, and it is widely used as a data mining tool. Classification surface is a spatial plane that can divide data samples into two classes, and the key goal of SVM is to define optimal boudaries. The problem can be separated into linear separable and linear inseparable problems according to whether a linear function can be used to separate the sample data. Defining an optimal classification surface also allows you to use linear equations to identify several samples that are closest to the optimal classification surface, which can be expressed as follows.

$$r_i(x_i'w + b) - 1 = 0 \quad (4)$$

Here, w represents the normal vector, the direction of the optimal classification plane; B represents the displacement, the distance between the origin and the optimal classification plane. By calculating the projection of the difference between two different support vectors in the normal vector direction, the classification interval expressed by $\frac{2}{\|W\|}$ can be obtained. Since the goal of support vector machine is to achieve the maximum classification interval and minimize $\|W\|$, the objective problem can be transformed into:

$$min\frac{1}{2}\|W\|^2, s.t.\, y_i(x_i'w + b) \geq 1 \quad (5)$$

The advantages of SVM include the ability to solve high-dimensional problems, i.e., being more friendly to large feature spaces. Ability to handle small sample situations in machine learning. It also solves the interaction of non-linear characteristics. Compared with the neural network model, SVM has no local minimum problem and stronger generalization ability. However, the disadvantage is that the computational efficiency is not very high when the sample size is large.

2.3 Metrics

For the sake of comparing the predictive performance of each model, three metrics are used, i.e., MAE, MSE and R-square. With regard to MAE, The smaller the MAE, the better the model.

$$MAE = \frac{1}{m}\sum_1^m|f(x_i) - y_i| \quad (6)$$

For MSE, the value range of MSE is from 0 to positive infinity, and the value range is also from 0 to positive infinity. Similar to the MAE, TThe smaller the MSE, the better the model.

$$MSE = \frac{1}{m}\sum_{i=1}^m(f(x_i) - y_i)^2 \quad (7)$$

R-square is the ratio of explained variance to total variance. A quantity that measures the fitting degree of a model. The larger the R square, the higher the fitting degree. When the model does not make any mistakes, R is 1.

$$R^2 = \frac{\sum_{i=1}^m(f(x_i)-y)^2}{\sum_{i=1}^m(y_i-y)^2} \quad (8)$$

3. Results & Discussion

3.1 Correlation analysis

In the selected relevant indicators, in addition to price related indicators, there are also position indicators and tema. Position refers to the total amount of futures or option contracts that have been concluded but have not been delivered or hedged. The position is the wind direction of the in-depth development of the futures market, which will affect the trading ability of a certain price or range. The triple index moving average is constructed o identify trends based on traditional moving average tema responds to price changes faster than traditional Ma or EMA. As shown in Fig.1, lnopen, lnopen, lnhigh, lnlow is highly related to the return.

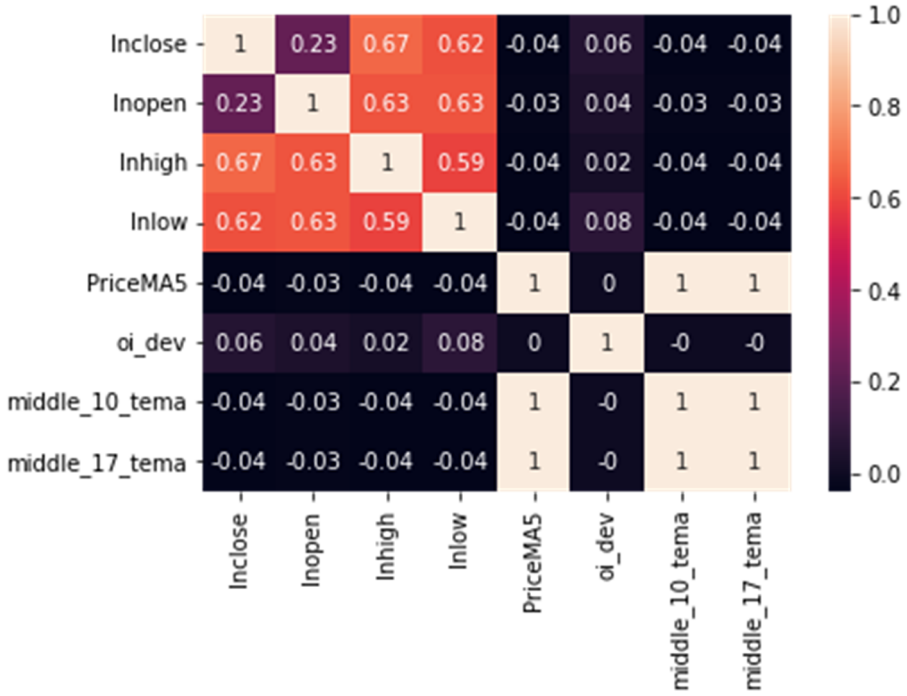

Fig. 1 Coefficient correlation

As listed in Table. 2, from both MAE and MSE, these indexes of the random forest model are slightly smaller than others, i.e., it is not difficult to find that there is little difference between the three models.

Table 2. Valuation indexes

	r-square	MSE	MAE	score
Linear	0.70128	4.6686513316695085e-05	0.00515	0.66008
svm	0.70459	4.616904648121196e-05	0.00500	0.65679
rf	0.72066	4.365860935076133e-05	0.00487	0.66826

3.2 Explanation

The linear model is used as the benchmark, and the above indicators are used as the dependent variables. The R-square is 0.70, which indicates that the model has certain prediction ability, while the R-square of random forest and SVM is not significantly improved on this basis. In the determination of SVM model, we try linear kernel function, Gaussian kernel function, sigmoid kernel function and polynomial kernel function respectively.

In addition to the linear kernel function, the other test scores are very low. Thus, this paper selects the linear kernel function for parameter adjustment. For SVM, C is mainly adjusted. C is the penalty coefficient, that is, the degree of punishment for errors in the process of classification. Increasing the value of C can make the classification more accurate.

However, the increase of C value will also increase the difficulty of classifying other data samples with the model and reduce the generalization ability of the model. Therefore, the focus of selecting C value is how to balance the generalization ability and classification accuracy of the model. Determining a suitable C value can make the model have better generalization ability and higher accuracy.

The prediction result of random forest is the best for the following reasons. After the training, the characteristics that are more important can be given by random forest. The influence between features can also be detected. Moreover, the data of some days for which the technical indicators cannot be calculated are deleted from this data set, while the accuracy of random forest can still be maintained. For unbalanced data sets, random forest can also balance errors.

3.3 Limitations

Nevertheless, the results given in this paper have some shortcomings and defects. Due to the limited performance of my computer and limited computing power and time, this study did not optimize each parameter of neural network and random forest, or even optimize each model in a shorter period, in order to improve the prediction ability. Although relevant studies have shown that the prediction effect of neural networks and other models is poor, and the operation is complex. Neural network, XGBoost and LSTM should be considered. The correlation of the original data set is particularly poor, especially the correlation coefficient between various technical indicators and yield is less than 0.01. The correlation between technical indicators and return is very low. In this case, more factors with better explanation ability ought to be added into the model. From 2016 to 2021, the stock 300 index was volatile.

4. Summary

In conclusion, this paper investigates future price forecasting based on various multiple factors approaches. To be specific, three regression prediction models, linear model, SVM and random forest, are used to predict China's stock index futures index, and the support vector machine and random forest of linear kernel function are optimized respectively. According to the empirical research, it is proved that the stochastic forest models are capable of achieving good results when used in the regression prediction of Chinese stock index futures. They are easy to be disturbed by accidental factors just launched, so it is more difficult to predict compared with other mature market indexes. Whereas it can still achieve good prediction results, so the model has good generalization ability. However, the current model does not involve deep learning. In the future, scholars ought to use the more state-of-art deep learning models and various strong factors or indicators for prediction. These results offer a guideline for future price prediction when selecting different machine learning approach.

References

[1] ZH Liang Cheng. Stock index futures price prediction based on wavelet kernel support vector machine regression. Master's thesis, Shanghai Normal University. 2018.
[2] Cortes C, Vapnik V. Support Vector Network. 1995, 20(3):273-297.
[3] Conqueret, G. Machine and Deep Learning Lecture Notes, London: Imperial College Business School, 2019:1-8.
[4] Drucker, H. Improving regressors using boosting techniques, in Machine Learning: Proceedings of the Fourteenth International Conference, 1997, pp. 107– 115, Morgan Kaufmann, Burlington, Mass.
[5] Kim K J. Financial time series forecasting using support vector mechines. Neurocomputing, 2003, 55(1-2): 307-319.
[6] Liu Fan. A Hybird Support Vector Machines and Discrete Wavelet Transform Model in Futures Price Forecasting. ISNN, LNCS 3973, 2006: 485-490.
[7] Liu Lixia, Ma Junhai. Least Squared Support Vector Machine for petroleum futures price prediction.Computer Engi-neering and Applications, 2008, 44(32): 230-231.

[8] Zhou Lei. Research on prediction model of stock index futures based on rough set and support vector machine Shandong science, 2010, 23 (5): 66-70
[9] Wang Wei. Research on the prediction of CSI300 index movement based on random forest method. Master's thesis, Tianjin University of Finance and economics, 2017.
[10] Wang Yue,SunDeshan. Stock Index Prediction. Journal of Quantitative Economics，2018，35(04): 28-30.