Information Systems and Economics (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/infse.2024.050203
ISSN 2523-6407 Vol. 5 Num. 2

Research on Stock Price Prediction Model Based on Weighted Sufficient Dimension Reduction and Bagging Framework

Yi Shen[1], Chi Qin[2]

[1]China University of Petroleum (Beijing), Beijing, 100000, China
[2]Guangxi University of Finance, Nanning, 530007, China

Keywords: Stock Price Prediction, Model Averaging, Factor Mining

Abstract: Research on stock price prediction models based on multi-factor models has always been one of the hot directions in quantitative finance. The key lies in accurately mining factors that significantly impact stock prices and constructing prediction models that are both precise and robust. In light of this, we propose a stock price prediction method based on sufficient dimension reduction and the idea of model averaging. On the one hand, this method utilizes a weighted version of sliced inverse regression and mean-variance estimation as tools for factor mining. While reducing the curse of dimensionality, it can theoretically retain all the effective information of the original factors on stock prices completely. On the other hand, this method introduces the bagging model, which can effectively balance the variance and bias in the prediction model, thereby significantly enhancing the model's generalization ability. The results of actual data analysis show that compared to other methods, the proposed method has a smaller mean squared error and absolute error, and it possesses a certain degree of robustness. Moreover, when the sub-models use interpretable machine learning algorithms, the proposed method can not only perform accurate stock price predictions but also reveal the feature importance of each quantitative factor in stock price prediction.

1. Introduction

Stock price prediction has always been a core issue in financial market research. Accurate stock price forecasts are crucial for investors to formulate effective investment strategies, corporate management to plan market strategies, and policymakers to devise market regulation policies. With the increasing complexity of the global financial markets and the surge in data volume, traditional stock price prediction methods face new challenges and limitations. Methods based on macroeconomic indicators analysis or technical analysis often rely on limited data dimensions and subjective judgments, making it difficult to capture the market's diversity, complexity, and dynamic changes. Furthermore, these methods have shortcomings in large-scale data processing, real-time analysis, and model adaptability. In recent years, with the development of big data technology and advanced statistical methods, the application of multi-factor models in stock price prediction has been increasing. These models can integrate more dimensions of information, including

fundamental factors, market sentiment, macroeconomic indicators, etc., providing a more comprehensive and accurate perspective for stock price prediction.

Recent developments in research based on multi-factor prediction models have been rapid. Liu Xiao[1] and others used Long Short-Term Memory (LSTM) neural networks to predict stock prices, focusing on analyzing the characteristics of financial data, especially the high correlation found within the same type of factors. Wu Jiawei[2] and others, based on fundamental and technical factors of stock prices, constructed a factor scoring model and a KPCA-GA-CatBoost stock price trend prediction model. This research involves screening and scoring stocks in the Shanghai and Shenzhen[3] stock markets and using Genetic Algorithm (GA) for model tuning. Wang Xianhe[4] and others focused on using Recurrent Neural Network (RNN) algorithms for stock selection, combined with multi-factor analysis, providing new insights into the application of RNN algorithms in the financial field. Yuan Yifang[5] and others focused on a portfolio prediction stock selection model based on machine learning models, introducing an Alpha hedging strategy. Hong Changjiang and others built a stock price trend prediction model based on the LightGBM algorithm, paying particular attention to stocks held by northbound funds, and building strategies through the prediction of price trends. Additionally, Yang Xuenin and others used neural network technology to screen and learn the factors of the stock market, aiming to describe the trend of the stock market based entirely on objective data and find high-return stock portfolios.

In the complex environment of financial markets, the accuracy of stock price prediction has a decisive impact on investment decisions. Although traditional methods of stock price prediction are effective in certain contexts, they have limitations in handling big data and market dynamics. To address these issues, this study proposes an innovative multi-factor stock price prediction method that combines sufficient dimension reduction and model averaging. Specifically, in terms of factor mining, this study employs a weighted version of sliced inverse regression technology and mean-variance estimation. Compared to traditional methods of information extraction, this technique theoretically retains the effective information of quantitative factors on stock prices while reducing dimensionality. In model construction, the proposed method utilizes the bagging approach. By integrating the predictions of multiple sub-models, it significantly reduces the variance of the prediction model and enhances its generalization ability. Moreover, when sub-models use interpretable machine learning algorithms, the proposed method can also provide the importance levels of each quantitative factor in stock price prediction

2. Theory and Method

2.1 Weighted Sufficient Dimension Reduction

In statistics, Sufficient Dimension Reduction (SDR) is a paradigm for analyzing data that combines the ideas of dimension reduction and the concept of sufficiency. Sliced Inverse Regression (SIR) is a classic method of sufficient dimension reduction. SIR investigates the regression of a multivariate variable X on a univariate Y , rather than the regression of a univariate variable on a multivariate variable X .In the process of dimension reduction for the multivariate explanatory variable X , the information of Y is fully utilized. The general model is:

$$y = f(\beta_1^T x, ..., \beta_K^T x, \in) \tag{1}$$

$\in$ is a random variable independent of x and f is an unknown link function;the conditional distribution of y given x can be represented by K linear combinations of $\beta_1^T x, ..., \beta_K^T x$ without losing the original information contained ;This is equivalent to the response variable y being independent of the explanatory variable x given these K linear combinations of x . When K is much smaller

than the dimension of x, the goal of dimension reduction is achieved. Sliced Average Variance Estimation (SAVE) is another statistical method used for dimension reduction and exploratory data analysis. Firstly, based on the values of the response variable, the dataset is divided into several groups or 'slices'. For each slice, its covariance matrix is calculated $\sum_i$.The covariance matrix is a statistical tool used to measure the relationships between variables within a slice. The average of the covariance matrices of all slices is then calculated. By performing eigendecomposition on the average covariance matrix $\overline{\sum}$ the principal eigenvectors are extracted, which define a low-dimensional representation of the data. The calculation of the weighted covariance matrix can be completed by assigning different weights to each observation. Assume X is an $n \times p$ matrix, where n is the sample size and p is the number of variables. There is a weight vector w, of length n.The formula for the weighted covariance matrix C is:

$$C=\frac{1}{\sum w_i}\sum_{i=1}^{n} w_i(x_i-\bar{x}_w)(x_i-\bar{x}_w)^T \tag{2}$$

2.2 Weighted Bagging Algorithm

One way to reduce the variance of estimates is to average multiple estimates together. When averaging the obtained function, the contributions of the variance terms tend to cancel out, thereby improving the prediction. For example, it's possible to train M different trees on different subsets of the data using bootstrap sampling (resampling with replacement) and then perform ensemble calculations, such as taking the average.

$$\mathrm{ycom}(x)=\frac{1}{M}\sum_{m=1}^{M} y_m(x) \tag{3}$$

This process is known as bootstrap sampling or bagging (bootstrap aggregation). Since bagging is a method of random sampling with replacement, some data will not be selected, and the probability of not being selected is about 1/3. This unselected data is referred to as out of bag samples, which can be used to test the model. Below, we illustrate with regression why bagging is better than a single model. Assume the regression function to be predicted is $h(x)$ then the output of each model can be written as the true value plus an error:

$$y_m(x)=h(x)+\in_m(x) \tag{4}$$

The mean squared error for a single model is:

$$E_x\left[\{y_m(x)-h(x)\}^2\right]=E_x\left[\in_m(x)^2\right] \tag{5}$$

Then the mean squared error for all models is:

$$E_{\mathrm{AV}}=\frac{1}{M}\sum_{m=1}^{M} E_x\left[\in_m(x)^2\right] \tag{6}$$

The expected mean squared error is:

$$E_{\mathrm{COM}}=E_x(ycom(x)-h(x))^2 \tag{7}$$

The results indicate that the average error of a model can be reduced by a factor of M by simply averaging M versions of the model, that is, bagging reduces bias. This depends on a key assumption that the errors caused by each model are uncorrelated. In practice, errors are often

highly correlated, and the overall reduction in error is usually small. However, it can be shown that the expected error will not exceed the expected error of the constituent models, and therefore, $E_{\mathrm{COM}} \leq E_{\mathrm{AV}}$ for more significant improvements, more complex techniques are employed. Assuming that the errors have a mean of zero and are uncorrelated, we have:

$$E_{\mathrm{x}}(\in_m (x)) = 0, E_{\mathrm{x}}(\in_m (x) \in_l (x)) = 0, m \neq l \tag{8}$$

Thus, Bagging reduces the variance of the model, enhances its stability, and decreases the risk of overfitting. It effectively deals with high-variance base models, such as decision trees, by combining them into a more robust ensemble model. In Bagging, it is assumed that each sub-model is equally important, hence given the same weight, which is clearly unreasonable. Therefore, it needs to be weighted, and the simplest weighting method is by calculating the correlation coefficient between the predicted values and the actual values in the validation set. Calculating the correlation coefficient: For each sub-model i, calculate the correlation coefficient Y_i between its predicted values Y and the actual values r_i in the validation set. The formula for the correlation coefficient is

$$r_i = \frac{\sum_{j=1}^{n} (Y_{ij} - \overline{Y}_i)(Y_j - \overline{Y})}{\sqrt{\sum_{j=1}^{n} (Y_{ij} - \overline{Y}_i)^2 \sum_{j=1}^{n} (Y_j - \overline{Y})^2}} \tag{9}$$

where n is the number of samples in the validation set, Y_{ij} is the predicted value for the ith sample by sub-model j Y_i is the average predicted value by sub-model j Y_j is the actual value for the ith sample, and $\overline{Y}$ is the average of the actual values. Next, these correlation coefficients are used as weights to combine the predictions of all sub-models to obtain the final weighted predicted value. The formula for weighted prediction is

$$\overline{Y}_{final} = \frac{\sum_{i=1}^{m} r_i Y_i}{\sum_{i=1}^{m} |r_i|} \tag{10}$$

where m is the total number of sub-models, r_i is the correlation coefficient of sub-model i and Y_i is the predicted value by sub-model i.

2.3 Stock Price Prediction Model Based on Weighted Sufficient Dimension Reduction and Weighted Bagging Algorithm

This paper proposes a stock price prediction model based on weighted sufficient dimension reduction and weighted Bagging algorithm. This method combines dimension reduction techniques with ensemble learning, aiming to enhance the accuracy and stability of stock price prediction. The model building process involves several key steps, including data preprocessing, dimension reduction, model training, and evaluation.

Step 1: Preprocess the stock price data, which includes data cleaning and standardization. The cleaning process involves removing outliers and missing data. Divide the data into training, validation, and test sets.

Step 2: Input the training set data into the weighted sufficient dimension reduction to obtain the projection matrix, and perform dimension reduction on the training, validation, and test sets to obtain the dimensionally reduced data.

Step 3: Input the dimensionally reduced training set data matrix into each sub-model of the weighted Bagging algorithm, estimate the parameters to be estimated for each sub-model, and use the validation set data to estimate the weight parameters of each sub-model.

Step 4: Input the projected data matrix of the test set into the trained weighted Bagging algorithm to obtain the predicted stock price values

3. Data Analysis

3.1 Data Source and Experimental Setup

Firstly, the data is sourced from the Wind database. This paper extracts data of three stocks for the period from January to October 2023, which includes China Medicine (600056), China Shipbuilding (600150), and China Nuclear Power (601985), as the subjects of study. This encompasses 9 different quantitative indicators including volume ratio, turnover rate, etc. Subsequently, the data undergoes preprocessing, which includes handling missing values, detecting and replacing outliers, and data normalization. Then, various technical indicators are calculated, such as the Moving Average (MA), Moving Average Convergence Divergence (MACD), and On-Balance Volume (OBV). These technical indicators help to reveal market trends and momentum.

For the experimental setup, the dataset is first randomly divided into training and test sets, and 100 simulation tests are conducted to assess the performance of the proposed method. The comparative methods include LASSO, Ridge Regression, Random Forest, Decision Tree, Support Vector Machine, and Extreme Gradient Boosting Tree. Mean squared error and absolute error are used as evaluation metrics, with their specific calculation formulas as follows:

$$MSE = \frac{1}{n}\sum_{n=1}^{n}(Y_i - Y_i)^2 \tag{11}$$

$$MAE = \frac{1}{n}\sum_{n=1}^{n}\left|Y_i - Y_i\right| \tag{12}$$

Similarly, n represents the total number of samples, Y_i is the actual value of the i th sample, and, Y_i is the predicted value of the i h sample. Regarding the selection of hyperparameters in the model, we use the cross-validation method for optimization.

3.2 Comparative Results

Tables 1-3 present the performance of seven different models (LASSO, Ridge, Tree, RF (Random Forest), SVM (Support Vector Machine), GBDT (Gradient Boosting Decision Tree), and the stock price prediction model FSAVE-IMB based on weighted sufficient dimension reduction and weighted Bagging algorithm) for three different stocks across 100 simulation experiments. These results are measured by two metrics: the mean and standard deviation of SRE (which may be a specific type of error rate), and the mean and standard deviation of SME (which may be another type of error measurement)

The analysis indicates that FSAVE-IMB demonstrates significant stability in terms of standard deviation for both SRE and SME errors. This model has the smallest standard deviation in these two error metrics among all models tested. This demonstrates FSAVE-IMB's high consistency and reliability in different simulation experiments, showing minor performance fluctuations even under data variation or different experimental conditions. Although FSAVE-IMB may not be the

best-performing model in terms of average error, its performance is still at a good level, especially considering its high stability. This means that FSAVE-IMB can maintain a relatively low error rate while avoiding extreme or unstable results. For applications that value model performance stability, such as financial market forecasting, medical diagnosis, or any scenario requiring reliable predictions, FSAVE-IMB could be an excellent choice. Future improvements could focus on enhancing its average error performance, which could be achieved through parameter optimization, feature engineering, or ensemble learning methods, aiming to maintain or enhance this stability while trying to improve overall performance.

Table 1: Comparative Results on Chinese Pharmaceutical Stocks

Method	SRE_mean	SME_mean	SRE_std	SME_std
LASSO	0.668408	0.493610	0.066392	0.086378
Ridge	0.643304	0.454819	0.080972	0.096086
Tree	0.700408	0.562467	0.055438	0.075756
RF	0.678854	0.523100	0.057172	0.072098
SVM	0.694204	0.569605	0.045527	0.058939
GBDT	0.666389	0.499051	0.063411	0.075456
FSAVE-IMB	0.646210	0.473585	0.030813	0.040957

Table 2: Comparative Results on Chinese Nuclear Power Stocks

Method	SRE_mean	SME_mean	SRE_std	SME_std
LASSO	0.585535	0.377079	0.036072	0.046407
Ridge	0.504327	0.280687	0.069792	0.068425
Tree	0.559284	0.365304	0.058127	0.080915
RF	0.541563	0.331664	0.046905	0.055697
SVM	0.548267	0.341121	0.029599	0.032314
GBDT	0.526779	0.313220	0.054164	0.060284
FSAVE-IMB	0.480260	0.258397	0.023982	0.022352

Table 3: Comparative Results on Chinese Shipbuilding Stocks

Method	SRE_mean	SME_mean	SRE_std	SME_std
LASSO	0.815847	0.750694	0.082115	0.127542
Ridge	0.806112	0.730109	0.085365	0.128585
Tree	0.885637	0.900292	0.049147	0.088415
RF	0.845511	0.817981	0.055007	0.094295
SVM	0.877608	0.899039	0.049109	0.085039
GBDT	0.836209	0.798216	0.065275	0.105109
FSAVE-IMB	0.756740	0.676067	0.026495	0.047724

In summary, the main advantages of FSAVE-IMB lie in its stability and good error control capabilities, making it particularly suitable for applications sensitive to model volatility. At the same time, it also offers room and directions for improvement.

4. Conclusion and Outlook

In this study, the stock price prediction method based on sufficient dimension reduction and the concept of model averaging, which combines sliced inverse regression, a weighted version of mean-variance estimation, and the bagging model, has demonstrated significant predictive performance and robustness. Efficient Factor Mining: Utilizing sliced inverse regression and a

weighted version of mean-variance estimation, this method successfully retains factors that significantly impact stock prices while reducing data dimensionality, thereby enhancing prediction accuracy.

Improved Model Generalization: The introduction of the bagging strategy has significantly enhanced the model's adaptability and generalization capability across different datasets. By integrating multiple sub-models, this method has achieved notable success in reducing the model's variance and bias. In practical data analysis, this method has shown smaller mean squared error and absolute error compared to traditional methods, proving its superior robustness. When sub-models use interpretable machine learning algorithms, this method not only provides accurate stock price predictions but also identifies and explains the specific impact of each quantitative factor on stock price prediction. Looking at the model's generalization: More suitable for complex financial markets: Due to its effectiveness in handling high-dimensional and complex data, it can be extended to more types of financial markets and different financial product predictions.

References

[1] Liu Xiao. Research on Multi Factor Stock Forecasting Based on Quantitative Trading -Take the Shanghai and Shenzhen 300 Index an Example [D]. Soochow University, 2022.

[2] Wu Jiawei. Research on composite multi-factor quantitative stock selection scheme combining fundamental and technical factors [D]. Shanghai Normal University, 2022.

[3] Wang Xianhe. Research on a Dynamic Multi Factor Stock Selection Model Based on Recurrent Neural Network [D]. Northeast University of Finance and Economics, 2022.

[4] Banglong L, Jie L, Guanghui Y. Research on Stock Price Prediction Model based on GA Optimized SVM Parameters [J]. International Journal of Security & Its Applications, 2016, 10(7):269-280. DOI: 10. 14257/ijsia. 2016. 10.7.24.

[5] Du X, Chen K, Zhang T, et al. Multistep-Ahead Stock Price Forecasting Based on Secondary Decomposition Technique and Extreme Learning Machine Optimized by the Differential Evolution Algorithm [J]. Mathematical Problems in Engineering, 2020. DOI:10.1155/2020/2604915.