Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Gaussian process regression model based on stacking algorithm and application to stock price prediction

Fan Wang [1, #], Tong Wang [2, #], Jiayu Fan [3, *, #]

[1] College of Science, Central University for Nationalities, Beijing, China, 100081

[2] Eurasia International, School of Henan University, Kai Feng, China, 475001

[3] School of Mathematics and Statistics, Hubei Normal University, Huangshi, China, 435000

* Corresponding author: ffan2003@163.com

[#]These authors contributed equally.

Abstract. Gaussian process regression, as a nonparametric statistical method capable of fitting nonlinear functions, holds an important place in the realm of quantitative finance. However, when applied to the prediction of noisy stock price time series, a single Gaussian process regression is prone to overfitting, and its computational efficiency diminishes with an increase in data volume. Considering this, we propose an ensemble learning model for Gaussian processes based on Bootstrap and stacking algorithms. The proposed approach balances the training and testing errors of individual predictive models through the concept of model averaging. Additionally, it enhances the computational efficiency of Gaussian process regression by employing subsamples instead of the full sample. Furthermore, by utilizing the stacking model, it can surpass the predictive limits of Gaussian processes, thus enhancing the predictive performance of the method. Simulated data analysis indicates that under the assumption of a Gaussian process model, the proposed method exhibits smaller mean squared error, absolute error, and relative error compared to some classical methods. Ultimately, when applied to the task of predicting stock price movements, it demonstrates higher accuracy and stability.

Keywords: Stock Price Forecasting; Model Averaging; Gaussian Process Regression; Computational Efficiency.

1. Introduction

Quantitative trading is a trading method based on mathematical models and algorithms, which relies on a large number of historical data to find the law of price changes. With the rapid development of computer technology and the advent of the era of big data, the amount of information in the stock market is growing explosively, and the traditional analysis methods are often difficult to cope with such a huge amount of data. As a powerful data analysis tool, machine learning is gradually integrated into the field of quantitative trading. How to design an efficient stock price trading strategy based on machine learning methods to promote the development of the stock market in the direction of intelligence and efficiency has always been an important research direction in the field of quantitative finance. In the research of quantitative trading based on machine learning methods, it is mainly divided into prediction models based on time series and multi-factor prediction models. For time series forecasting model, the commonly used methods are exponential smoothing, ARIMA, GARCH and long and short-term memory neural network. Specifically, Zhou Hongyong [1] et al. Used the seasonal exponential smoothing method to predict the demand trend of electric energy meters, and combined with the LSTM model to verify the results. Benvenuto D [2] et al. predicted the epidemiological trends of COVID-19 prevalence and incidence based on ARIMA model for epidemiological data; The prediction method based on LSTM neural network by Huo Jing [3] et al. carried out research on the surface wind pressure of coal shed structure, and proposed a multi-scale neural network prediction model. However, in the time series prediction model, only the time index and historical data information are involved, which is obviously insufficient for the complex and changeable task of stock price series. Therefore, this paper focuses on the use of multi-factor models

to predict the rise and fall of stock prices. In the multi-factor prediction model, and two important problems are encountered: the curse of dimensionality and the unknown link structure. Shengzheng W [4] realized the variable selection of feature variables based on LASSO regression algorithm; By analyzing the influence of various factors on electricity consumption, Hu Chunfeng [5] et al. proposed a monthly electricity consumption forecasting method based on the elastic network model, and compared the VAR model, BP neural network and Lasso. However, the relationship between the factor and the stock price is often nonlinear or unknown. In this case, some nonparametric methods may be suitable. For example, Youmeilu's [6] cotton fiber micronaire grade prediction model based on LightGBM algorithm makes full use of cotton fiber inspection data and simplifies the cotton fiber inspection process. Tzenios N [7] uses random forests to understand whether the deployment of educational technology in schools can improve students'academic performance, Cui K [8] established the prediction of geotechnical parameters on the basis of analyzing the characteristics of geotechnical materials, the distribution of geotechnical sediments and geotechnical parameters.

However, these nonparametric regression methods converge slowly when they encounter the curse of dimensionality, for example, when neural networks encounter high-dimensional problems, they need to use thousands of parameters to set the model, which will increase the computational cost and inefficiency, and it is easy to cause overfitting phenomenon. To sum up, this paper will focus on Gaussian process regression, which is a model that can consider both stock price volatility information and factor information. According to the actual needs, we can choose the appropriate kernel function to fit the copulas with different complexity, and the complexity of the model is related to the sample size, but not to the dimension, so there is no impact of the curse of dimensionality. For example, Zeng Xiongzhi [9] et al. Fitted the relationship between precipitation and environmental variables through Gaussian process regression. Liu Zhongyong[10] believed that the classical linear model of permanent magnet synchronous motor could not be applied to complex and changeable working conditions, and constructed a Gaussian process regression model for the flux-current relationship.

As Gaussian Process Regression is also a non-parametric model, single Gaussian Process Regression can easily fall into overfitting in high noise time series prediction, and the computational efficiency of Gaussian Process Regression decreases with the increase of data volume. In response to these issues, this paper proposes a Gaussian Process Ensemble Learning Model based on Bootstrap and stacking algorithms. On one hand, it balances the training error and testing error of a single prediction model through the idea of model averaging. On the other hand, this method improves the computational efficiency of Gaussian Process Regression by using subsamples instead of the full sample. Moreover, by applying the stacking model, it can break through the prediction limit of Gaussian Process, enhancing the predictive performance of the method. Simulated data analysis shows that under the Gaussian Process model assumption, the proposed method has smaller mean square error, absolute error, and relative error compared to some classical methods. Finally, when the proposed method was applied to stock price fluctuation prediction tasks, it demonstrated higher accuracy and stability.

2. Gaussian process regression based on the Boosting and stacking algorithms

Firstly, based on the bootstrap concept, data is randomly sampled from $P=\{(x,y)\,|\,i=1,2,...,n\}$ to form I sub-samples. Among them, $d=\{(x_{subi},y_{subi})\}_{i=1}^{I}$ is the training subset. Each sub-sample is then inputted into an independent Gaussian process, resulting in several estimated Gaussian processes denoted as $Z=\{GP_i\}_{i=1}^{I}$. The principle of a single Gaussian process regression model is as follows.A Gaussian Process (GP) is a probabilistic process in which the joint distribution of any finite number of random variables follows a multivariate Gaussian distribution.Specifically, for any finite number of time points $t_1,t_2,\ldots,t_n$, corresponding to random variables $X(t_1),X(t_2),\ldots,X(t_n)$, they follow a multivariate Gaussian distribution. Given a defined training data $G=\{(x_i,y_i)\,|\,i=1,2,\ldots,n\}=(X,y)$,

where: $x_i \in R^m$ is m-dimensional training vector. $X = [x_1, x_2, ..., x_n]$ training matrix of $m \times n$ dimensions. $y_i \in R$ is the output scalar of x_i the output vector is $Y = [y_1, y_2, ..., y_n]$ For any given set of input points $x_i (i = 1, 2, \ldots, n)$ .The corresponding output value $f(x_i)$ forms a random vector space $(f(x_1), f(x_2), \ldots, f(x_n))$. The joint probability distribution over the entire input space follows an n-dimensional Gaussian distribution:

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \sim N\left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{bmatrix}, \begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & \ddots & \vdots \\ K(x_n, x_1) & \cdots & K(x_n, x_n) \end{bmatrix} \right) \tag{1}$$

The statistical properties of Gaussian processes can be described by their mean function $m(X)$ and covariance function $K_f(X, X')$:

$$\begin{cases} m(X) = E[f(X)] \\ K_f(X, X') = E[(f(X) - m(X))(f(X') - m(X'))] = \mathrm{cov}[f(X), f(X')] \end{cases} \tag{2}$$

Therefore, the formula for the probability distribution of a Gaussian Process (GP) can also be expressed by its mean function and covariance function as:

$$f(X) \sim GP(m(X), K_f(X, X')) \tag{3}$$

To simplify the definition and derivation process of Gaussian processes, this paper performs data centering by setting the mean function to zero.

At this moment, the joint prior distribution of training inputs x and testing inputs x_*, as well as the training outputs y and testing outputs y_*, is as follows:

$$\begin{bmatrix} Y \\ y_* \end{bmatrix} \sim N\left(0, \begin{bmatrix} K(X, X) & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix} \right) \tag{4}$$

To obtain the posterior distribution of the function, we need to condition this joint prior distribution on the observed outcomes.

$$\begin{aligned} y_* \mid X_*, X, y \sim N(&K(X_*, X))K(X, X)^{-1} y, \\ &K(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*)) \end{aligned} \tag{5}$$

According to the given equation evaluates the mean and covariance matrix, we can draw samples of function values corresponding to the input from the joint posterior distribution. To measure the importance of each Gaussian process regression model, we use a generalized weighting method and propose that the SBGP model also considers that there are too many Gaussian process models, which is easy to cause overfitting. We use the idea of Stacking, and the output of each Gaussian process regression is regarded as a new input variable. Then using the elastic network as a new enhanced regression model to correct the accuracy of the model. Specifically, the obtained output variable Z will be used to create a new synthesized dataset Y .Then, the stacking model will be trained to fit the

$\{Z, Y\}$ model. When new samples are available, the corresponding columns are extracted using the bootstrap method and added to the i-th Gaussian process. The estimated values are then obtained by incorporating these processes into the model. The specific modeling steps are shown in Figure 1.

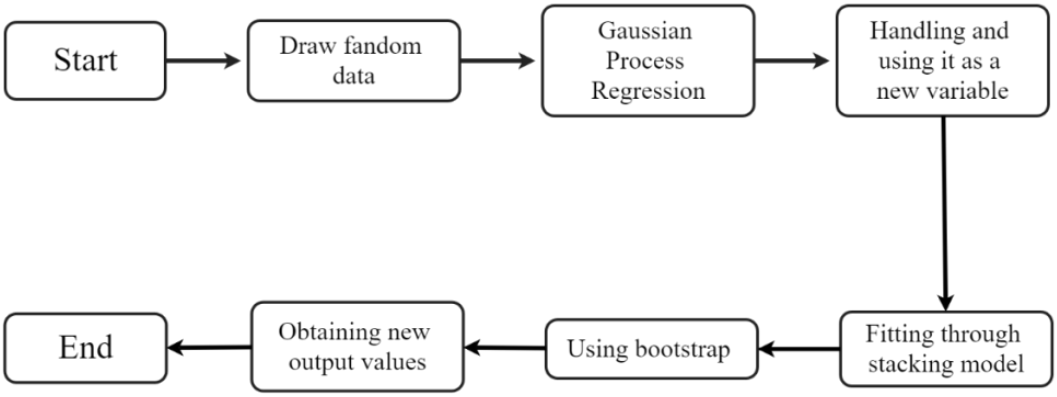

Figure 1. Flowchart of SBGP model establishment

3. Data Analysis

3.1. Analysis of Simulation Data

Comparing Equation (6) with Equation (7), it is obvious that the two equations are identical except for the variable x4 in Equation (7).By comparing the prediction results of the two sets of simulated data, it is possible to determine whether x4 has an impact on the model, and thus identify whether x4 is an independent variable.

$$Y_1=\sin(\frac{\pi}{2}*x1)+\cos(\frac{3}{2}\pi*(x2+x7))+2*(x3+x4+x5+x6)+\varepsilon \quad (6)$$

$$Y_2=\sin(\frac{\pi}{2}*x1)+\cos(\frac{3}{2}\pi*(x2+x7))+2*(x3+x5+x6)+\varepsilon \quad (7)$$

Where x1, x2, x3, x4, x5, x6, and x7 are independently and identically distributed in U(0, 1), ε obeys N(0, 0.01).

Set the total number of samples generated by each function n to be 100, and the total sample sets are sample1 and sample2, respectively. The proportion of the test set is 0.1, that is, 100*0.1=10 samples.In order to avoid the contingency of the experimental results, set the number of experiments D=10.In order to verify the superiority of the prediction effect of the Gaussian process regression model based on the Stacking algorithm (referred to as SBGP model), five nonlinear comparison models are set in this paper, respectively Decision Tree Regression, XGB regression, Random Forest regression, KNN regression and MLP regression.Calculate the mean and standard deviation of the mean square error of 10 experiments of each model, and use them as the standard to evaluate the model.The smaller the mean value is, the higher the accuracy is, and the smaller the standard deviation is, the more robust the model is. The following is the formula for calculating the mean square error:

$$\mathrm{MSE}(\hat{\theta})=\mathrm{E}(\hat{\theta}-\theta)^2 \quad (8)$$

Based on this, we use python programming to achieve the above operation, and the results are shown in Table 1 and Figure 2.

Table 1. Prediction effect of two groups of simulation data

sample1	mean	std
DTree	1.01	0.55
XGBoost	0.45	0.24
RF	0.35	0.23
KNN	0.29	0.15
MLP	0.18	0.38
SBGP	0.02	0.02

sample2	mean	std
DTree	0.76	0.27
XGBoost	0.51	0.25
RF	0.45	0.26
KNN	0.41	0.10
MLP	0.76	0.83
SBGP	0.04	0.03

The left table of Table 1 is the mean and standard deviation of the prediction results and true values of sample1 after 10 experiments using the above six models, and the right table of Table 1 is the mean and standard deviation of the prediction results and true values of sample2 after 10 experiments using the above six models. Obviously, as can be seen from the bar chart, the prediction effect and robustness of the two groups of simulation data based on the SBGP model are far better than those of the other five models.

As can be seen from the Table 1, the average mean square error of the prediction results of SBGP model for sample1 is only 0.02, among the other models, the best prediction effect is MLP regression, with an average mean square error of 0.18, and the prediction effect of SBGP model is better than that of MLP regression by nearly 87%. The average mean square error of the prediction results of SBGP model for sample2 is only 0.04, among the other models, the best prediction effect is KNN regression, with an average mean square error of 0.41, and the prediction effect of SBGP model is better than that of KNN regression by nearly 90%. In addition, the standard deviation of the two groups of simulation data based on the SBGP model is the smallest, which is not prone to overfitting, and the prediction results are relatively robust. SBGP model makes the model robust by using Bootstrap robustness, and uses Stacking model to break through the prediction limit of Gaussian process, which enhances the prediction performance of this method.

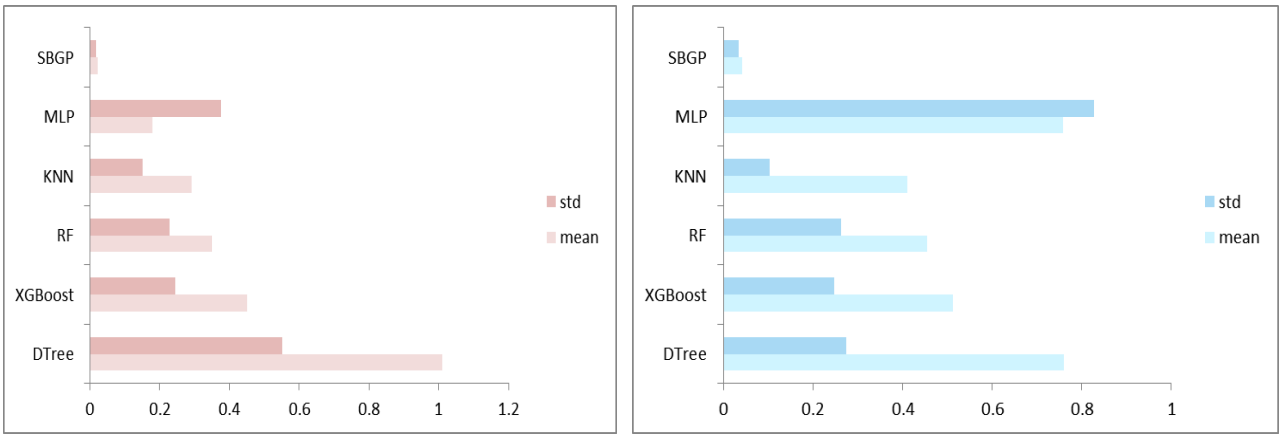

Figure 2. Visualization of prediction effect of two groups of simulation data

3.2. Real data analysis

This paper obtains the stock data of Ping An Bank from Tushare platform, making Y the rise and fall of the stock price, x1 the opening price, x2 the highest price, x3 the lowest price, x4 the closing price, x5 the yesterday's closing price, x6 the rise and fall amount, x7 the rise and fall amount, x8 the trading volume, and x9 the trading volume, a total of 9 factors. The rolling method is designed to divide the training set and test set for experiments, that is, the predicted stock price month is used as the test set, and the stock data of the first three months are used as the training set, and the experiment is carried out by rolling forward. A total of 8 experiments are carried out. Based on the stock price information of Ping An Bank from February to November in 2023, the stock price rise and fall of Ping An Bank from May to December in 2023 is predicted. In the empirical link, in addition to the mean square error consistent with the simulated data, two errors commonly used in the stock price are calculated, namely absolute error and relative error. The three errors are combined as the standard to evaluate the quality of the model. The following results are shown in Table 2.

According to the comprehensive analysis and comparison of Figure 3-5, the broken line representing MLP regression has the largest fluctuation, and MLP regression is easy to fall into overfitting in the prediction of stock price rise and fall. While the three errors calculated by the SBGP model represented by the red line are occasionally slightly larger than other models in each month, but the broken line fluctuation is not large, and the prediction results are the most robust. On the whole, the error is the smallest, and the prediction results are relatively accurate, that is, the SBGP model achieves excellent results in both accuracy and robustness in the quantitative trading of stocks.

Table 2. Mean square, absolute and relative error of the prediction results

		5	6	7	8	9	10	11	12
DTree	MSE	0.07	0.10	0.05	0.04	0.03	0.02	0.02	0.05
	AE	0.20	0.23	0.16	0.18	0.14	0.11	0.14	0.17
	RE	2.24	4.42	3.02	2.47	3.43	2.47	3.15	4.54
XGBoost	MSE	0.06	0.06	0.04	0.03	0.02	0.02	0.01	0.02
	AE	0.19	0.17	0.13	0.16	0.12	0.11	0.09	0.10
	RE	2.08	2.87	2.11	1.95	2.60	2.04	1.47	1.97
RF	MSE	0.05	0.04	0.03	0.03	0.01	0.03	0.01	0.02
	AE	0.18	0.15	0.12	0.16	0.11	0.12	0.08	0.12
	RE	1.84	2.22	1.58	2.02	2.29	2.66	1.24	2.81
KNN	MSE	0.04	0.03	0.03	0.02	0.01	0.03	0.01	0.02
	AE	0.16	0.13	0.12	0.10	0.10	0.14	0.08	0.11
	RE	1.30	1.65	1.39	0.80	1.91	3.15	1.27	2.43
MLP	MSE	0.07	0.05	0.10	0.02	0.01	0.25	0.02	0.02
	AE	0.21	0.17	0.29	0.14	0.11	0.45	0.09	0.11
	RE	2.50	2.57	6.86	1.65	2.36	14.08	1.35	2.40
SBGP	MSE	0.04	0.03	0.03	0.02	0.01	0.01	0.01	0.01
	AE	0.16	0.14	0.11	0.13	0.07	0.07	0.07	0.09
	RE	1.01	1.89	1.57	1.01	1.23	1.21	1.22	1.02

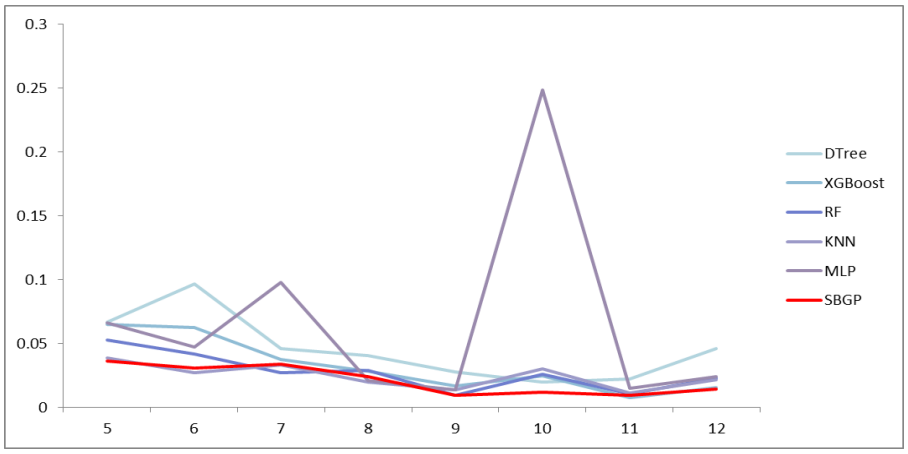

Figure 3. The mean square error of prediction results of six models

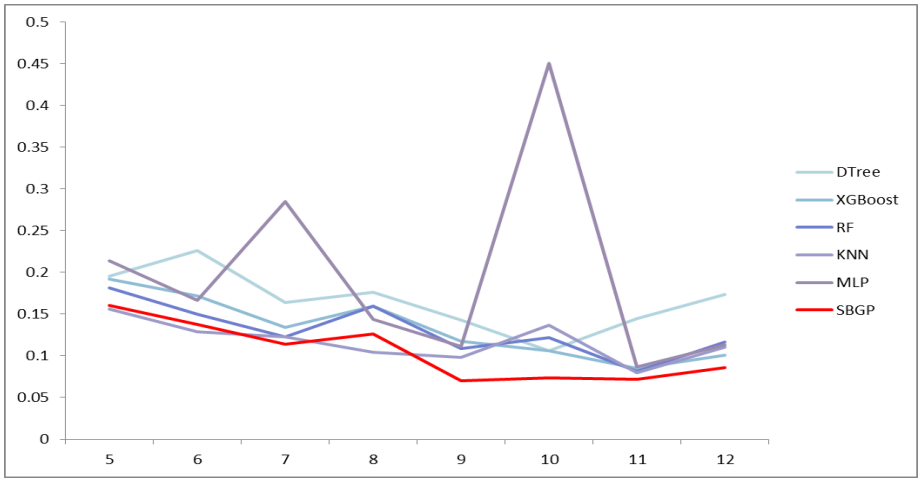

Figure 4. The absolute error of prediction results of six models

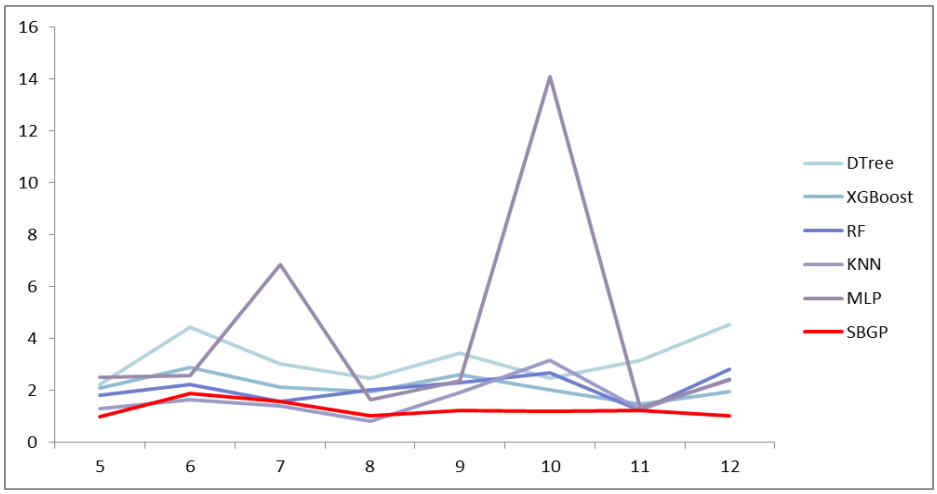

Figure 5. The relative error of the prediction results of six models

4. Conclusion

In the field of financial transactions, with the development of time, there has been a significant increase in information volume. The conventional Gaussian process regression may encounter the issue of overfitting when predicting time series of stock prices with high levels of noise. In order to

address this issue, this paper attempts to combine Gaussian processes with Bootstrap and stacking algorithms to obtain a novel Gaussian process ensemble learning model. The introduction of the stacking algorithm enables the new model to effectively harness the unique strengths of each individual base model. The traditional univariate and low-dimensional predictions of Gaussian process regression are thus transformed into multidimensional predictions, thereby further enhancing the predictive performance. On the other hand, by combining multiple Gaussian process regression models and using the idea of model average, the training error and test error of a single prediction model are weighed, which further improves the accuracy of the prediction. Furthermore, the proposed model also enhances the computational efficiency of conventional Gaussian process regression, thereby enabling traders to capture stock information and enhance the precision and effectiveness of trading decisions more accurately.

The SBGP model, in summary, presents a more precise and efficient approach for stock price prediction in the financial trading domain. This assists traders in effectively addressing the challenges posed by escalating information volume and high-noise environments. This model can not only be used for quantitative stock trading, but also for solving other economic problems. For example, it can be applied to handle non-Euclidean data in economic research, combining macro and micro studies, utilizing high-frequency data for nonlinear dynamic simulations, and applying soliton theory for live testing to further promote the application of the model. The selection of stacking sub-models in this study is based on the utilization of simple linear regression. The SBGP model exhibits greater stability and advanced predictive capabilities compared to traditional Gaussian processes. however, further experimentation and comparison are necessary to ascertain whether there exist alternative models that can enhance the representation of stacking concepts.

References

[1] Zhou Hongyong, Sun Yuting, Zhang Yanzhan. Analysis of energy meter demand forecast based on seasonal exponential smoothing method [J]. Electricity Demand Side Management, 2024, 26 (02): 95 - 99.

[2] Benvenuto D, Giovanetti M, Vassallo L, et al. Application of the ARIMA model on the COVID-2019 epidemic dataset [J]. Data in brief, 2020, 29: 105340.

[3] Huo J, Liu S, Zhang Z, et al. A multiscale long and short-term memory neural network prediction method for coal shed surface wind pressure [J/OL]. Engineering Mechanics, 2024: 1 - 9.

[4] Wang S, Ji B, Zhao J, et al. Predicting ship fuel consumption based on LASSO regression [J]. Transportation Research Part D: Transport and Environment, 2018, 65: 817 - 824.

[5] Chunfeng Hu, Shiming Tian, Hang Su. Monthly electricity consumption forecasting method based on elastic network model [J]. Power Engineering Technology, 2020, 39 (03): 166 - 172.

[6] You Meilu, Liang Huixiang, Abduzhixiti-Buaimaiti et al. Prediction of cotton fiber mclonal value grade based on data mining [J]. Modern Textile Technology, 2024: 1 - 8.

[7] Tzenios N. Examining the Impact of EdTech Integration on Academic Performance Using Random Forest Regression [J]. ResearchBerg Review of Science and Technology, 2020, 3 (1): 94 - 106.

[8] Cui K. Research on prediction model of geotechnical parameters based on BP neural network [J]. Neural Computing and Applications, 2019, 31: 8205 - 8215.

[9] Zeng Xiongzhi, Yu Chao. Study on spatial prediction of precipitation in Guangdong Province based on Gaussian process regression algorithm [J]. People's Pearl River, 2023, 44 (S2): 52 - 55.

[10] LIU Zhongyong, FAN Tao, HE Guolin et al. Nonlinear magnetic chain identification of permanent magnet synchronous motor based on Gaussian process regression [J]. Journal of Power Supply, 2024: 1 - 15.