Gaussian Process Regression Based on Stochastic Segmentation of Data and Its Application to Stock Price Prediction

Zishun Liu *

School of Science, China University of Mining and Technology Beijing, Beijing, China, 100083

* Corresponding Author Email: zishun.liu.math@foxmail.com

Abstract. Gaussian Process Regression (GPR) is a powerful model for stock price prediction in both research and practice within financial markets. However, when applying GPR to stock price prediction, the model faces overfitting and underfitting problems. Moreover, when the amount of processed data is large, computational complexity and resource consumption become significant factors limiting its practical application. To address the above challenges, this study proposes a Gaussian process regression model based on stochastic data segmentation. It aims to optimize the computational efficiency of the model and improve the prediction performance. This method not only significantly reduces the computational complexity, but also improves the prediction performance through integrated learning of sub-models. It also improves the generalization ability of the model through the integrated learning of sub-models. In addition, the introduction of the model averaging strategy effectively mitigates the overfitting and underfitting problems by weighting the uncertainty measures of the sub-models. To verify the effectiveness of the proposed method, this study first analyzes a large number of simulation experiments. The performance of the model is systematically evaluated. Secondly, by selecting the stock data of listed companies in a number of different industries as the research object, the application value and robustness of this method in the real market environment are further confirmed.

Keywords: Gaussian Process Regression, Stock Price Forecast, Stochastic Segmentation.

1. Introduction

The continuous development and evolution of financial markets have brought unprecedented data volume and complexity, among which stock price forecasting occupies a central position in the field of financial quantitative research, which is related to the accuracy of investment decisions and market analysis [1]. The accuracy of forecasting directly affects the investor's rate of return and risk control; therefore, the development of effective stock price forecasting model has important theoretical value and practical significance. The volatility of stock prices is affected by a variety of factors such as macroeconomics, company performance, and market sentiment, etc. With the continuous evolution of the financial market, the volatility of stock prices increases, which makes the accurate prediction of stock prices a challenging task. Therefore, exploring new models that can accurately capture the characteristics of stock price volatility not only improves the prediction accuracy, but also provides market participants with more reliable decision support.

Traditional stock price forecasting methods such as linear regression and time series analysis have shown some effectiveness in dealing with simple linear relationships. However, when faced with the nonlinear nature of the market and high-dimensional data, these methods often show limitations. In recent years, with the development of machine learning technology, a series of prediction methods based on nonlinear models have gradually emerged, including neural networks, support vector machines, etc., which have demonstrated better performance in capturing the dynamic changes of stock prices [2]. However, the computational burden, overfitting, and underfitting problems of these models, when dealing with large-scale data, still need to be addressed. In addition, most of the models fail to provide an effective measure of the uncertainty of the predicted value, which is an important omission in practical applications because investors need to know not only the predicted value of the stock price, but more importantly, the credibility of the prediction, i.e., the uncertainty interval of the

predicted value. To this end, this study focuses on Gaussian Process Regression (GPR), which is a nonparametric statistical method. The advantage of GPR is that it allows the model to adaptively learn complex nonlinear relationships in the data, while being able to naturally give an estimate of the uncertainty in the predicted value. Although the GPR method has the advantages of easy implementation, adaptive acquisition of hyperparameters, and probabilistic significance of the predicted outputs, it still has some problems, mainly focusing on two aspects: one is the kernel matrix computation, and the other is model prediction robustness [3]. When applying GPR to stock price prediction, how to improve the robustness of prediction and the computational efficiency of the model becomes a key challenge.

To overcome these challenges, Zhang Xinyu, Zou Guohua's research [4] proposed a machine learning process that incorporates the GPR method into Bootstrap. In this study, an improved GPR model based on stochastic data segmentation is proposed from the other side. By stochastically partitioning the dataset into multiple smaller subsets, applying GPR independently on each subset, and then combining the prediction results of each sub-model using a model averaging strategy, this approach not only significantly improves the computational efficiency, but also enhances the generalization ability of the model by reducing the risk of overfitting and underfitting. This study systematically evaluates the performance of the proposed model through simulation experiments and analysis of actual stock price data. The results show that compared with the traditional GPR and other prediction models, the improved model has significant improvement in prediction accuracy and robustness, especially the advantage of providing a weighted uncertainty measure of the predicted values, which provides a new perspective and tool for stock price prediction. This study not only expands the application of GPR in the field of financial market forecasting, but also provides new perspectives and methods for dealing with complex financial data.

2. The Gaussian process regression model based on stochastic data segmentation

Let the data set $D = \{x_i, y_i\}_{i=1}^n$, where $X = \{x_i\}_{i=1}^n$ denotes the set of input matrices and $Y = \{y_i\}_{i=1}^n = \{f(x_i)\}_{i=1}^n$ denotes the response variables. Gaussian process regression is a nonparametric Bayesian regression method, which assumes that Y has a joint distribution of Gaussian processes in a finite set of given data D, and that all the statistical characteristics of the Gaussian processes consist of the mean function $m(x)$ and the covariance function $K(x, x')$, i.e., for any $x, x' \epsilon X$, satisfying,

$$\begin{cases} m(\mathbf{x}) = \mathbf{E}[f(\mathbf{x})], \\ K(\mathbf{x}, \mathbf{x}') = \mathbf{E}\big[\big(f(\mathbf{x}) - m(\mathbf{x})\big)\big(f(\mathbf{x}') - m(\mathbf{x}')\big)\big] \end{cases} \tag{1}$$

The basic form of Gaussian process regression can be modeled as follows:

$$Y = f(X) + \varepsilon \tag{2}$$

where $f(X)$ is an unknown function, ε is added as noise, usually assumed to be an independent identically distributed normal random variable $\varepsilon \sim N(0, \sigma_n^2)$.

Thus, the prior distribution of the observation y can be obtained as:

$$y \sim N(0, K(X, X) + \sigma_n^2 I_n) \tag{3}$$

And the Gaussian joint prior distribution of observations y and predictions f_*:

$$\begin{bmatrix} \mathbf{y} \\ f_* \end{bmatrix} \sim N\left(0, \begin{bmatrix} K(X, X) + \sigma_n^2 I_n & K(X, \mathbf{x}_*) \\ K(\mathbf{x}_*, X) & K(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix}\right) \tag{4}$$

where $K(X, X) = K_n = (k_{ij})$ is the $n \times n$ order symmetric positive definite covariance matrix (called **the kernel matrix**).

From this, the main GP regression equation can be solved, and the posterior distribution of the predicted value f_* can be calculated as

$$f_*|X,\mathbf{y},\mathbf{x}_* \sim N(\bar{f}_*, \text{cov}(f_*)) \tag{5}$$

were

$$\bar{f}_* = K(\mathbf{x}_*, X)[K(X,X) + \sigma_n^2 I_n]^{-1}\mathbf{y} \tag{6}$$

$$\text{cov}(f_*) = K(\mathbf{x}_*, \mathbf{x}_*) - K(\mathbf{x}_*, X)[K(X,X) + \sigma_n^2 I_n]^{-1}K(X, \mathbf{x}_*) \tag{7}$$

Then $\hat{\mu}_* \triangleq \bar{f}_*, \hat{\sigma}_{f_*}^2 \triangleq \text{cov}(f_*)$ is the mean and variance of the corresponding predicted value of f_* at test point x_*. Combined with the nature of normal distribution, that is, for each GPR, we can get the predicted value $x' = \hat{\mu}_*$ and the uncertainty measure $\sigma' = \hat{\sigma}_{f_*}^2$ of the test point x_*.

In the above Gaussian process, each gradient calculation needs to invert the covariance matrix $K_n + \delta_n^2 I_n$, so the computational complexity is $O(n^3 p)$, where p is the number of gradient calculations; on the other hand, the model fitted by highly redundant data is very prone to overfitting and underfitting problems. Among many methods to reduce the computational complexity, the simplest one is the data subset approximation method, i.e., only a small subset of dimension m from the original n-dimensional training set is selected as the new training set for GPR prediction. Although this method seems simple, it has no additional computational and memory overhead compared to other more complex approximation methods, and in the case of stock price prediction, where daily stocks generate highly redundant datasets, the additional data points provide very little information about the function, and it is not necessary to sacrifice computation to other complex approximation methods in order to obtain a negligible improvement in performance[5].

Based on this, the key to applying the data subset approximation method is how to select an appropriate data subset. In order to use a suitable selection method, we borrowed the resampling methods of Bootstrap and Random Forest, and fully considered the possible imbalance of weights in each random segmentation and used the weighted random segmentation selection method (SBGPR) for sampling. The specific steps are as follows:

(1) At the ith data partitioning, the original data set is $\{X, Y\}$.The original data set is re-sampled (i.e., sampled with put-back) into an m-dimensional (m<<n) training set $\{X_i, Y_i\}$, and record the number of columns in which the samples were taken $\{i_1, i_2 \ldots i_m\}(1 \le i_1 \le i_2 \ldots \le i_m \le n)$.

(2) For X_i ,using Gaussian process regression methods, its corresponding posterior distribution is calculated. In particular, the mean of the posterior distribution can be derived $\hat{\mu}_i = \bar{f}_{i*}$, and the variance$\sigma_i = \text{cov}(f_{i*})$.

(3) Using the uncertainty measure (variance) of this sampling σ_i, the weights for this sampling are calculated

$$\alpha_i = \sum_k \sigma_i\,(k) \tag{8}$$

(4) Repeat the above operation k times to obtain a dataset of Gaussian process regression based on data segmentation $\Delta = \{X_i, Y_i, \hat{\mu}_i, \alpha_i\}_{i=1}^k$

(5) Predictions for Gaussian process regression based on data segmentation at this point in time are calculated for the corresponding

$$Y^* = \sum_{p=1}^{k} \left(\frac{\alpha_p}{\sum_{n=1}^{k} \alpha_n} \hat{\mu}_i\right) \tag{9}$$

Through this Gaussian process regression method based on stochastic data segmentation, we not only improve the computational efficiency, but also enhance the generalization ability and predictive stability of the model through the integrated learning of sub-models. More importantly, the weighted averaging process effectively utilizes the uncertainty information of each submodel, optimizes the uncertainty measure [6,7] of the overall prediction, and provides a more accurate risk assessment for decision making. We call this new approach the stochastic division Gaussian process (SBGP).

3. Data analysis

In this section, the application of Gaussian process regression model in stock price prediction and its performance evaluation are introduced in detail, which is centered on the analysis of simulated data and real data

3.1. Analog data analysis

The aim of this study is to evaluate the performance of Gaussian Process Regression (GPR) models based on random segmentation of data in the case of complex data.

Let the input variable $X \sim N(\mu, \sigma^2)$. To test the performance of this model, a nonlinear function model that fits a Gaussian distribution containing Gaussian noise will be utilized, and the response variable will be generated by the following function [eq10] In this section, the application of Gaussian process regression model in stock price prediction and its performance evaluation are introduced in detail, which is centered on the analysis of simulated data and real data

$$y = \sin\frac{\pi}{2}x_{i1} + \cos\frac{2}{3}\pi(x_{i2} + x_{i7}) + 2(x_{i3} + x_{i4} + x_{i5} + x_{i6}) + \varepsilon \tag{10}$$

where $X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6}$ is the input variable, and ε are the random variables modeled as Gaussian noise, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, and σ^2 is the variance of the noise. In order to compare the performance of GPR models based on data partitioning, this study selects several classical machine learning prediction methods: decision tree regression (Dtree), XGBoost regression, random forest regression (RF), KNN regression, and MLP regression as comparisons to comprehensively evaluate the performance of GPR models based on data partitioning. The models are evaluated by the mean square error (MSE) between the predicted and true values. The total sample size of the benchmark model is 100, and the ratio of the training set to the test set is 2:8. The final mean and standard deviation of these experiments are used to represent the performance of the model. This not only measures the accuracy of the model's predictions, but also assesses its robustness. In order to ensure the reliability of the experimental results and to eliminate chance as much as possible.

With the above parameter settings, simulation experiments were conducted on the generated data in the standard form, by varying the number of parameters, parameter dimensions, the proportion of the training set, and the standard deviation of the Gaussian noise parameter, and the results are shown in the following tables, Table 1, Table 2, Table 3, Table 4 and Table 5, respectively.

Table. 1. Comparison results under the baseline model

Methods	MSE	std
DTree	0.902447	0.293875
XGBoost	0.544162	0.294718
RF	0.484042	0.163326
KNN	0.388188	0.121717
MLP	0.998065	0.718209
SBGP	0.013702	0.00988

Table. 2. Changing the model parameters (without Xi4) under the comparison results

Methods	MSE	std
DTree	0.964503	0.451806
XGBoost	0.590251	0.184033
RF	0.433210	0.149367
KNN	0.344876	0.113781
MLP	0.568095	0.716736
SBGP	0.043736	0.024473

Table. 3. Comparison results with n=50

Methods	MSE	std
DTree	1.156637	0.730146
XGBoost	0.694155	0.336267
RF	0.522900	0.251641
KNN	0.417127	0.220812
MLP	1.089756	0.483421
SBGP	0.014712	0.008236

Table. 4. Comparison results for the 0.85 training set share case

Methods	MSE	std
DTree	1.249027	0.696175
XGBoost	0.794229	0.370592
RF	0.702679	0.354389
KNN	0.744427	0.238452
MLP	1.530264	0.947163
SBGP	0.045742	0.050680

Table. 5. Comparison results at 0.1 standard deviation of noise parameters

Methods	MSE	std
DTree	1.685602	0.665414
XGBoost	0.953289	0.559650
RF	0.787941	0.208242
KNN	0.577604	0.191934
MLP	1.492221	0.555818
SBGP	0.023671	0.009232

The results of the above experimental data were analyzed:

•**Decision tree regression (DTree)** may not be effective enough with data containing Gaussian noise and complex nonlinear patterns, resulting in relatively weak performance and relatively high MSE across simulation experiments.

•**XGBoost regression (XGBoost)** outperforms decision tree regression due to its gradient boosting framework, which is able to continuously reduce the prediction error by adding more trees, and its strong ability to capture the distribution of features and nonlinear relationships in the data. However, although XGBoost performs relatively well in the simulation experiments, it still shows some limitations compared to the SBGP model, especially in dealing with high-dimensional data and complex data relationships.

•**Random Forest regression (RF)** still does not perform as well as the SBGP model, probably due to the fact that Random Forests are still subject to a certain degree of overfitting risk in the face of extremely complex and noisy data.

•**K Nearest Neighbor Regression (KNN)** performs moderately well in this experiment, suggesting that choosing the right k value and distance metric becomes critical when the data have complex nonlinear relationships. The performance of KNN is limited by its "dimensionality catastrophe" problem for high-dimensional data, and its inability to effectively learn complex patterns in the data.

•The performance of **the two-layer BP neural network** in the simulated data experiments is not outstanding, which may be due to the fact that the neural network requires a large amount of data for effective training and is very sensitive to the choice of hyperparameters. In addition, overfitting and underfitting are also key factors affecting the performance [8].

•**SBGP** performs optimally among all comparative models, with the lowest MSE. This is due to the nonparametric nature of Gaussian process regression itself, which makes it more flexible in capturing complex nonlinear relationships and patterns in the data. SBGP further improves the prediction accuracy and model robustness by applying GPR independently on multiple data subsets and integrating the predictions of these submodels. In addition, the strategy of randomized data partitioning reduces the computational complexity, enabling SBGP to effectively handle large-scale and high-dimensional datasets.

In summary, the Gaussian process regression model based on stochastic segmentation of data shows significant advantages in simulated data analysis, which is mainly attributed to its nonparametric properties and the superiority of the powerful ability of Gaussian process in capturing complex data relationships. Compared with traditional machine learning models, SBGP not only provides more accurate predictions, but also effectively evaluates the prediction uncertainty with good stability and robustness.

3.2. Real data analysis

In the real data analysis section, we focus on evaluating the application effect of Gaussian process regression model based on data random segmentation in actual stock price prediction. Therefore, we selected stock data of four listed companies in different industries as the research object. By accessing the daily interface [9] from the Tushare library in Python, we selected the daily K-line market trends of four different industries, namely ShenLiangKongGu(SZKG), ShenSaiGe(SSG), ShiHuaJiXie (SHJX), and ZhongGuoChangCheng(ZGCC), from January 3 to December 29, 2023. These include trading date, opening price, highest price, lowest price, closing price, yesterday's closing price (previously reinstated), price increase/decrease, price increase/decrease (not reinstated), trading volume (hands), and trading volume (in thousands of yuan). Table 6 shows the data values of some stocks obtained, and Figure 1 shows the stock price fluctuations of these four stocks during this period.

Table. 6. Selected daily market data for SZKG

trade_date	open	high	low	close	pre_close	change	pct_chg	vol	amount
20231229	7.45	7.55	7.45	7.54	7.51	0.03	0.3995	41720.9	31328.048
20231228	7.48	7.52	7.4	7.51	7.49	0.02	0.267	47721.43	35630.65
20231227	7.28	7.5	7.26	7.49	7.29	0.2	2.7435	49805.4	36867.038
20231226	7.3	7.37	7.26	7.29	7.31	-0.02	-0.2736	31152.01	22783.05
20231225	7.35	7.35	7.23	7.31	7.35	-0.04	-0.5442	30368.7	22133.364
20231222	7.4	7.45	7.31	7.35	7.4	-0.05	-0.6757	43375.1	31938.158
20231221	7.31	7.46	7.25	7.4	7.34	0.06	0.8174	53641.08	39380.618
...	...	...	...	...	...	...	...	...	...

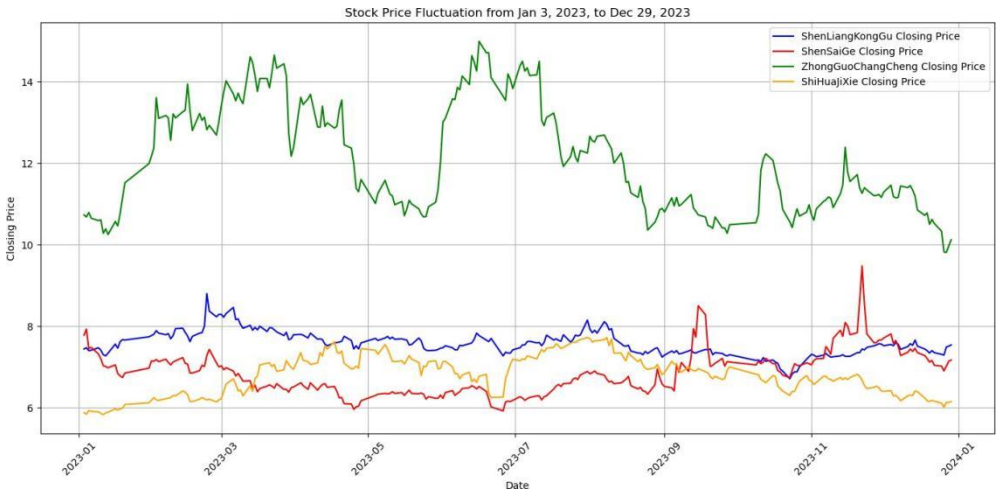

Figure 1. Chart of share price volatility for four stocks

In the real data analysis, based on the characteristics of the stock price data, we use rolling forecasts to evaluate the performance of the model. Specifically, we take 90 days as the training period, forecast the stock price for the next 30 days, and then update the data window on a rolling basis. This rolling forecasting approach not only matches the actual investment decision scenarios, but also tests the model's ability to adapt to market changes. In order to compare the prediction performance of Gaussian process regression model, we also choose decision tree regression, XGB regression, random forest regression, KNN regression and MLP regression as comparison models [10]. The performance of all models will be evaluated based on the mean square error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). Through the experiments, the MSE, MAE, and MAPE of each stock are averaged to obtain the results of the data analysis and evaluation of the different models for the four stocks, as shown in the following Table 7, Table 8 and Table 9, respectively.

Table. 7. MSE Evaluation of Four Stocks by Different Models

STOCK	DTree	XGBoost	RF	KNN	MLP	SBGP
SZKG	0.011475	0.008998	0.008297	0.008311	0.014603	0.006897
SSG	0.150875	0.039548	0.03913	0.038308	0.075373	0.047806
SSJX	0.173045	0.099469	0.115897	0.081274	0.137924	0.050828
ZGCC	0.01376	0.007902	0.011839	0.009589	0.009897	0.005932

Table. 8. MAE Evaluation of Four Stocks by Different Models

STOCK	DTree	XGBoost	RF	KNN	MLP	SBGP
SZKG	0.08525	0.069554	0.06759	0.075167	0.09688	0.068243
SSG	0.31205	0.163866	0.170465	0.145	0.220981	0.170238
SSJX	0.2859	0.231555	0.247315	0.217667	0.291342	0.171366
ZGCC	0.10995	0.069798	0.09216	0.086	0.076935	0.059903

Table. 9. MAPE Evaluation of Four Stocks by Different Models

STOCK	DTree	XGBoost	RF	KNN	MLP	SBGP
SZKG	1.487672	1.060304	0.995131	1.158275	1.752552	0.9951
SSG	3.965863	3.567115	2.916498	1.115329	6.973674	1.29627
SSJX	4.471303	3.182844	3.533321	2.37295	4.366706	1.019139
ZGCC	3.294254	1.49511	2.484124	2.02902	1.852126	1.069964

The experimental results show that the Gaussian process regression model based on stochastic partitioning of data basically exhibits the lowest mean square error, average absolute error and average absolute percentage error on all four stocks of different industries, and the best performance on all stocks of all industries, which shows its strong predictive ability and model generalization, which is mainly attributed to the nonparametric characteristics of Gaussian process regression and the ability of learning the intrinsic laws of data, especially its advantages in dealing with complex data and simulating uncertainty. This is mainly due to the nonparametric nature of Gaussian process regression and its ability to learn the intrinsic patterns of the data, especially its advantages in handling complex data and modeling uncertainty. Further, by analyzing the performance of the model over different forecasting periods, we find that the model has good temporal stability. Even in the long-term prediction, the performance of the model does not show significant degradation, which further validates its reliability and effectiveness in practical applications.

4. Conclusion and Prospect

In this study, we explore in detail a Gaussian process regression method based on random data segmentation, aiming to improve the accuracy and stability of stock price prediction. By randomly

partitioning the original data set into several small subsets, each of which is independently predicted by Gaussian process regression, and then combining these predictions, we not only improve the efficiency of the overall prediction, but also optimize the accuracy and robustness of the prediction through the weighted average method. The core advantage of this innovative method is that it effectively combines the nonparametric characteristics of Gaussian process regression and the idea of integrated learning, which provides a new perspective for the prediction of financial time series such as stock prices. We found that this method outperforms the traditional Gaussian process regression and other existing stock price prediction models in terms of prediction accuracy and model robustness. Especially when dealing with complex financial data, the proposed model can better capture the nonlinear characteristics and dynamic changes of the data, which shows a significant advantage.

Of course, there are further directions to improve this model. On the one hand, we will explore the possibility of further improving the performance of the model, especially considering the integration of more powerful learning algorithms to secondary process and analyze the posterior distributions of the sub-models, in order to further enhance the stability and accuracy of the predictions. On the other hand, considering the successful application of this model in stock price prediction, we plan to extend its application to other fields, such as economic indicator prediction, environmental change monitoring, etc., in order to verify its wide applicability and effectiveness. Most importantly, the results of this study will be applied to actual financial trading and investment decisions to further enhance the performance and usefulness of trading strategies by designing efficient trading strategies and combining them with the accuracy and uncertainty measures provided by the model predictions.

References

[1] Zeng Zhaoyou. Bootstrap Improved Algorithm for Cointegration Test of Spatial Panel Models and Power Analysis[J/OL]. Systems Engineering, 1-10[2024-03-13]

[2] Pandove, D., Goel, S., \& Rani, R. (2018). Systematic Review of Clustering High-Dimensional and Large Datasets. ACM Transactions on Knowledge Discovery from Data (TKDD), 12, 1 – 68

[3] He Zhikun, Liu Guangbin, Zhao Xijing, et al. Overview of Gaussian Process Regression Methods[J]. Control and Decision, 2013, 28(08): 1121-1129+1137. DOI: 10.13195/j.kzyjc.2013.08.018.

[4] Zhang Xinyu, Zou Guohua. Model Averaging Methods and Their Applications in Prediction[J]. Statistical Research, 2011, 28(06): 97-102. DOI: 10.19343/j.cnki.11-1302/c.2011.06.018.

[5] TALEIZADEH ATA ALLAH, TAFAKKORI KEIVAN, THAICHON PARK. Resilience toward supply disruptions:A stochastic inventory control model with partial backorstxdering under the base stock policy[J]. Journal of Retailing and Consumer Services, 2021, 58: 102291.

[6] Li Feng, Gao Xiaohai, Zheng Pengfei, et al. Typical Scenario Extraction Technique for Power System Planning Based on Gaussian Process Regression and Uncertainty Coupling Relationship[J]. Journal of Electric Power Science and Technology, 2022, 37(01): 64-73. DOI: 10.19781/j.issn.1673-9140.2022.01.008.

[7] Incoronata E T, Davide C, Elisa C. Nonlinear UGV Identification Methods via the Gaussian Process Regression Model for Control System Design[J]. Applied Sciences,2022,12(22):11769-11769.

[8] Xu Yiwei, Lu Wanrong. Research on the Prediction of Baijiu Stock Prices Based on XGB-LSTM Combined Model[J]. Electronic Components and Information Technology, 2022, 6(06): 64-68. DOI: 10.19772/j.cnki.2096-4455.2022.6.016.

[9] Xie Meifen. Personal Financial Data Acquisition and Analysis Based on Python[J]. Digital Technology and Application, 2023, 41(05): 118-122. DOI: 10.19695/j.cnki.cn12-1369.2023.05.37.

[10] Chen Yongtu. Interpretability Analysis of Stock Price Prediction Problems Based on Several Machine Learning Models[D]. East China Normal University, 2023. DOI: 10.27149/d.cnki.ghdsu.2023.002004.