Design of a Securities Market Investment Risk Control Model Based on Market Factors

Hongtao Ma   Zihang Zhou

Beijing International Studies University   School of Economics

Abstract: Based on the historical data of the CSI 300 index constituents from 2014 to 2024, this paper constructs a systematic risk-return prediction model incorporating ex-ante risk control. The study selects risk measurement indicators such as average return rate, volatility, and beta coefficient as market factors, performs technical operations such as lag characteristics and moving average indicators, and builds a risk prediction model. On this basis, an ex-ante risk control system with dynamic adjustment of stock allocation is designed. Using risk indicators such as VaR, the maximum drawdown of the investment portfolio is controlled within 0.25, achieving a cumulative return of 6%, significantly outperforming international risk control standards. Additionally, the study introduces the neural network NAR model as a comparative analysis tool to further verify the model's advantages in improving prediction accuracy. Overall, this research provides a scientific and effective risk control solution for portfolio management, which has important theoretical significance and practical value.

Keywords: systematic risk; risk measurement indicators; backtest risk control system; CSI 300;

I. Introduction

With the continuous opening up and development of China's capital market and the increasing degree of marketization, financial instruments and investment strategies are becoming more diversified. However, the rapid development of the market is also accompanied by an increase in systematic risk. Systematic risk, as an undiversifiable market risk, affects all asset classes. How to effectively measure and control systematic risk has become one of the hot issues in financial research. This study aims to construct a systematic risk prediction model incorporating various risk measurement indicators using machine learning models, thereby designing an effective ex-ante risk control system to control the maximum drawdown. The core tasks of the research include: firstly, calculating and analyzing the CSI 300 index constituents, then constructing a systematic risk model with predictive ability based on at least three risk measurement indicators; subsequently, according to the magnitude of risk, designing a backtest risk control system that can control the maximum drawdown within 0.7, meeting the risk control requirements of hybrid equity funds; finally, setting a reasonable expected return range based on historical data. During the research period, a comprehensive analysis and modeling will be conducted based on the historical data of the CSI 300 index constituents (2014-2024) to provide a more scientific and effective risk control solution for portfolio management.

II. Data Sources and Construction of Systematic Risk Prediction Model

2.1 Data Sources and Preprocessing

This study selects the historical data of the CSI 300 index (code: 000300) constituents as the research object, covering all data from January 1, 2014, to December 31, 2024. The main data sources are from the Akshare library, and the files are divided into two types: hs300stocks_i and hs300stocks_kdata_i, where i is an integer from 2014 to 2024, indicating the year of the data. hs300stocks_i mainly includes weights, while hs300stocks_kdata_i includes main indicators such as daily opening price, highest price, lowest price, closing price, trading volume, and turnover. The data preprocessing steps are as follows:

(1) Data cleaning: Delete rows containing missing values and ensure that all price and volume data are of float type.

(2) Weight adjustment: According to our weight file[1], standardize the weights of each stock to calculate the market value-weighted average return.

2.2 Design of the Systematic Risk Prediction Model

To capture the dependencies and historical patterns in the time series and enhance the predictive power for future risks, we introduce lagged features. These features help the model identify and understand patterns in the time series by utilizing historical data as model inputs, thereby improving the accuracy of market dynamics predictions. At the same time, to smooth data fluctuations and highlight long-term trends while reducing the interference of short-term noise, we employ moving average indicators. By calculating the average value within a specific time window, the model can more clearly identify the overall market trend, providing a solid foundation for investment decisions. Furthermore, to provide in-depth information on momentum, volatility, and trend changes, assisting the model in identifying overbought/oversold conditions and trend reversal signals in the market, we apply technical indicators. These tools together form a powerful analytical framework that enables us to gain insights into the market from different perspectives and develop more accurate and effective trading strategies.

The specific steps are as follows:

(1) Introduce Lag Features: We select five indicators: average return rate, volatility, beta coefficient, maximum drawdown, momentum indicator, and On Balance Volume (OBV). Based on these indicators, we calculate 1-5 day lagged features.

(2) Calculate Moving Averages: Such as the 20-day moving average of volatility and beta coefficient.

(3) Calculate Technical Indicators: Such as Relative Strength Index (RSI), Bollinger Bands, and MACD.

[1] We use Python to extract and obtain the attachment: hs300_weights_2014_to_2024.xlsx

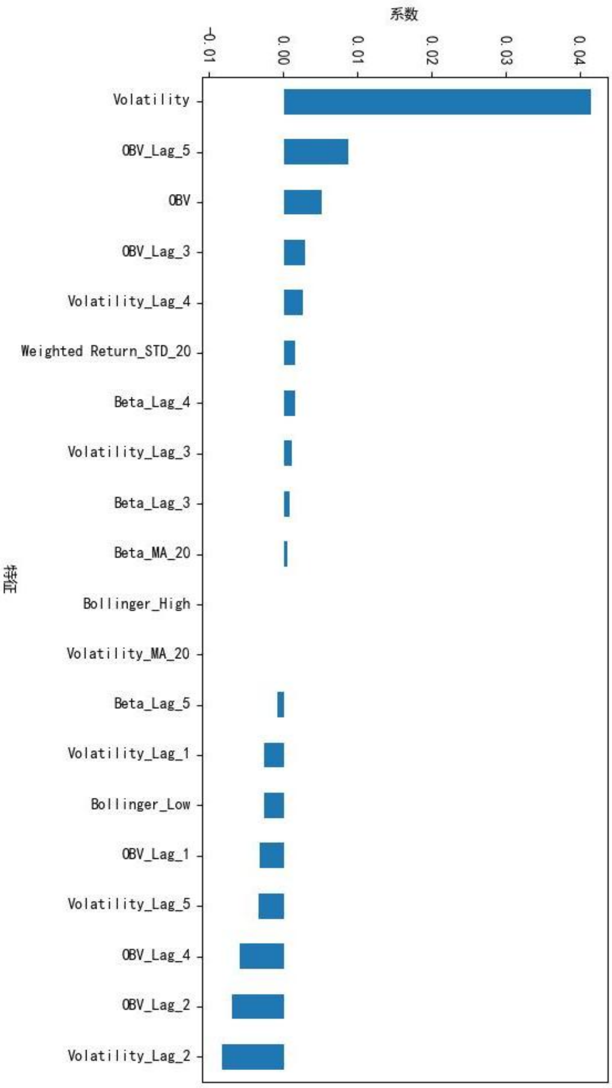

Figure 1. Selected 20 indicators

Finally, using the SelectKBest method, we select the 20 most relevant features from all features based on the f_regression scoring function (see Figure 1) to reduce model complexity and avoid overfitting.

After selecting the feature lags, we choose seven regression models: Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, XGBoost Regressor, LightGBM Regressor, Support Vector Regression (SVR), and Multi-layer Perceptron Regressor (MLPRegressor). We then perform cross-validation using TimeSeriesSplit to ensure that the model training process aligns with the characteristics of time series data. GridSearchCV is used to tune the model hyperparameters and optimize model performance.

2.3 Model Training and Evaluation

Next, we evaluate the results of the seven regression models using evaluation metrics including Mean Squared Error (MSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2), and Mean Absolute Percentage Error (MAPE), selecting the model with the highest R^2 value as the best model.

Through training and evaluation of the seven regression models, the following performance metrics are obtained:

Table 1: Performance Evaluation of Seven Models

Models	MSE	MAE	R^2	MAPE
lr	0.000003	0.001021	0.918368	0.046806
rf	0.000007	0.001999	0.775954	0.107222

gb	0.000012	0.002854	0.630533	0.152299
xgb	0.000014	0.003208	0.55402	0.170815
lgbm	0.000008	0.002152	0.740892	0.115094
svr	0.009938	0.099546	-305.753592	5.021229
mlp	0.001307	0.02891	-39.330321	1.365914

As can be seen, the best model is Linear Regression. In Table 1, the MSE value is 2.644597e-06, indicating that the average of the squared differences between the predicted values and the actual values of the model is small, suggesting that the prediction error of linear regression is very small. MAE measures the average of the absolute differences between the predicted values and the actual values, indicating that our model's prediction accuracy is relatively high. R^2 (Coefficient of Determination) is 0.918368, which is an indicator of the goodness of fit of the model, indicating that the model can well explain the variability in the data. MAPE is 0.046806, indicating that the average percentage of the prediction error to the actual value is relatively small, and the prediction results are very close to the actual values. Overall, Figure 2 shows the prediction results of the systematic risk prediction model based on linear regression, demonstrating that the model has high prediction accuracy and a good fit.

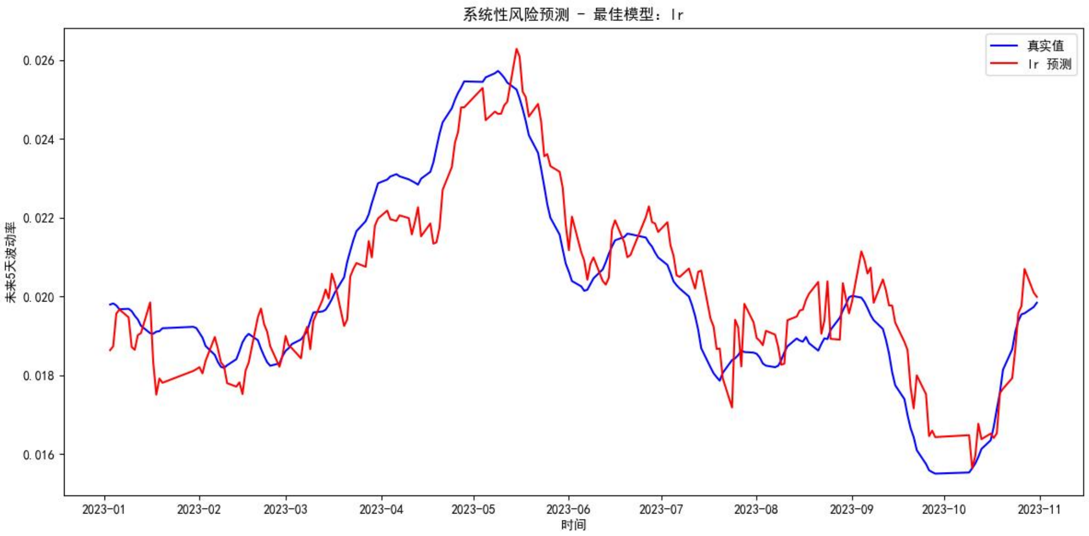

Figure 2: Linear regression model prediction

The residuals at different time points are shown in Figure 3:

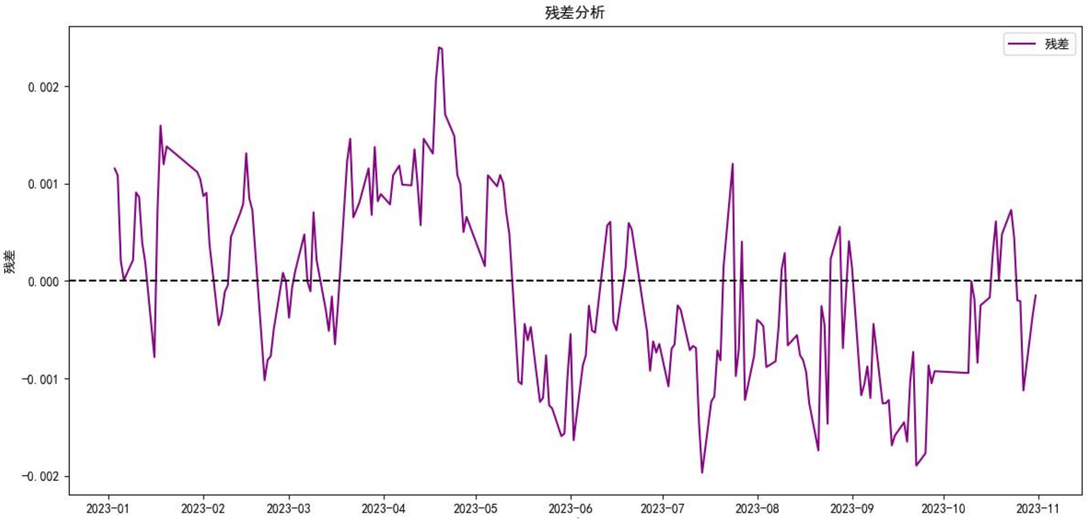

Figure 3: Residual analysis

III. Construction of the Pre-Risk Control System

As mentioned earlier, we will use the linear regression model for overall systemic risk forecasting. According to international practices, the drawdown risk control line for equity mutual funds is generally set at 0.7, meaning that the maximum drawdown i.e.,the largest decline from the highest to the lowest point should not exceed 70%. However, in China, the drawdowns for both public and private equity funds often fall below this international standard, indicating a higher level of risk. To address this challenge, we have designed a pre-risk control system that dynamically adjusts the stock allocation in the portfolio, aiming to keep the maximum drawdown of equity funds within 0.7, in line with international risk control standards, and thus reduce risk. The specific steps are outlined as follows:

3.1 Initial Allocation Proportions — Risk Indicators and Thresholds

This part is mainly designed to satisfy the allocation ratios corresponding to all the predicted stocks (portfolio), which primarily depend on the comparison between the predicted volatility and our set VaR (Value at Risk) threshold. In simple terms, the stock weights in the portfolio are determined by the volatility prediction results provided by the systemic risk prediction model.

For the predicted results, based on the linear regression model constructed in Task 2, we forecast the volatility for the next five days. If a stock is considered extremely high risk HighRisk, with a forecasted volatility greater than the 99% VaR, the stock allocation will be reduced to 20%. For medium-high risk MediumRisk, with a predicted volatility between the 99% and 95% VaR, the stock allocation will be reduced to 50%. For low-medium risk LowRisk, with a predicted volatility between the 95% and 90% VaR, the stock allocation will be reduced to 80%. For low-risk stocks

LowRisk, with a predicted volatility less than or equal to the 90% VaR, the stock allocation will remain at 100%.

Subsequently, we perform dynamic adjustments on these stock allocation proportions using the Sharpe ratio and the Sortino ratio. The Sharpe ratio and Sortino ratio are two key indicators used to measure investment performance. The Sharpe ratio assesses risk-adjusted returns by comparing the excess return of the portfolio to its total risk standard deviation, while the Sortino ratio focuses only on downside risk, providing an evaluation of investment profitability under adverse risk conditions. These two indicators help investors assess the return potential of different investment choices while considering risk.

In the construction of our pre-risk control system, if the Sharpe ratio or Sortino ratio falls below 1, we will further reduce the stock allocation to lower the risk. Based on the adjusted stock allocation proportions, we calculate the portfolio's daily return, cumulative return, and maximum drawdown.

(1)Sharpe Ratio:

Definition: Measures excess return per unit of risk.
Calculation method:

$$SharpeRatio = \frac{\mu_p - r_f}{\pi}\sigma_p$$

Where: μ_p is the average portfolio return, r_f is the risk-free rate, σ_p is the portfolio's volatility.

(2)Sortino Ratio:

Definition: Similar to the Sharpe ratio, but only considers downside risk.
Calculation method:

$$SharpeRatio = \frac{\mu_p - r_f}{\sigma_p}$$

Where: σ_p is the portfolio’s downside deviation.

3.2 Setting Reasonable Return Expectations

We use the yield of 10-year government bonds as the benchmark for the long-term risk-free rate. Government bonds are typically considered risk-free assets, and their yield can be viewed as the minimum reasonable expected return for investors. Returns above this benchmark should account for the inherent volatility and risk of the equity market.

Using data from the CSI 300 index from 2014 to the present, we calculate the annualized average return of the index. When compared to the 10-year government bond yield, this annualized return can serve as a reasonable long-term return expectation for the stock market. The calculation method is as follows:

- **Weighted Return of Each Stock:**

Calculate the return of each stock for the day. Multiply the return of each stock by its allocation weight in the portfolio.

$$\mathrm{Portfolio\ Return_t} = \sum_{i=1}^{n} \mathrm{w_i} \times \mathrm{r_i}$$

Where:

wi is the weight of the ii-th stock,
ri is the return of the iii-th stock on the given day.

- **Cumulative Return Calculation:**

Sum the weighted returns of all stocks on a given day to calculate the portfolio return. The cumulative return is then calculated by performing compounded accumulation on the portfolio's daily returns.

$$\mathrm{Cumulative\ Return_t} = \prod_{i=1}^{t} (1 + \mathrm{Portfolio\ Return_i})$$

- **Annualized Return CAGR:**

Using the cumulative return and the number of days in the investment period, we calculate the compound annual growth rate CAGR using the following formula:

$$\mathrm{CAGR} = \left(\frac{\mathrm{Final\ Value}}{\mathrm{Initial\ Value}}\right)^{\frac{365}{\mathrm{n}}} - 1$$

IV. Backtesting and Validation of the Pre-Risk Control System

We backtested the constructed pre-risk control system using historical A-share data from 2014 to 2024 to verify its effectiveness. The backtest results are as follows:

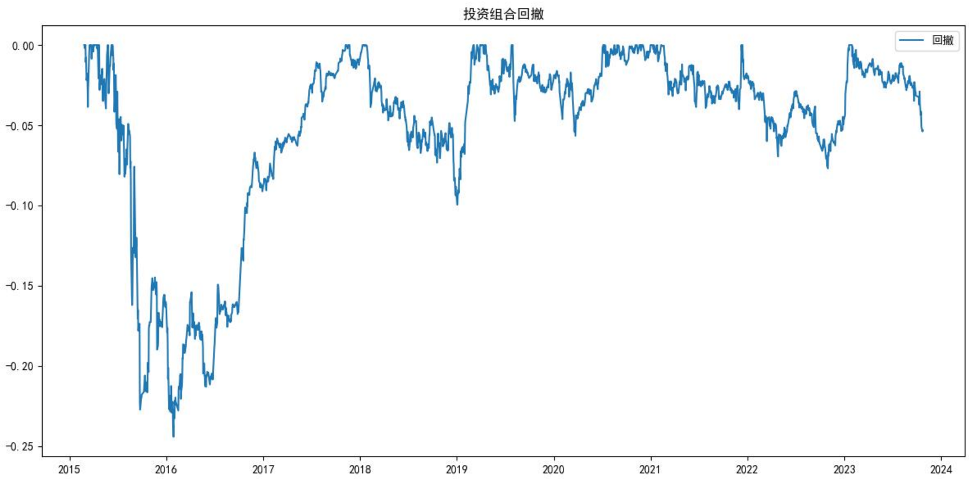

Figure 4: Maximum Drawdown: -0.25019073224332533, meeting the target of keeping it under 0.7

The backtest results show that the constructed pre-risk control system can effectively control the maximum drawdown of the portfolio to below 0.25, which is significantly lower than the international standard of 0.7. Additionally, although the Sharpe and Sortino ratios of the portfolio are relatively low, the system maintains stable cumulative returns while controlling risks.

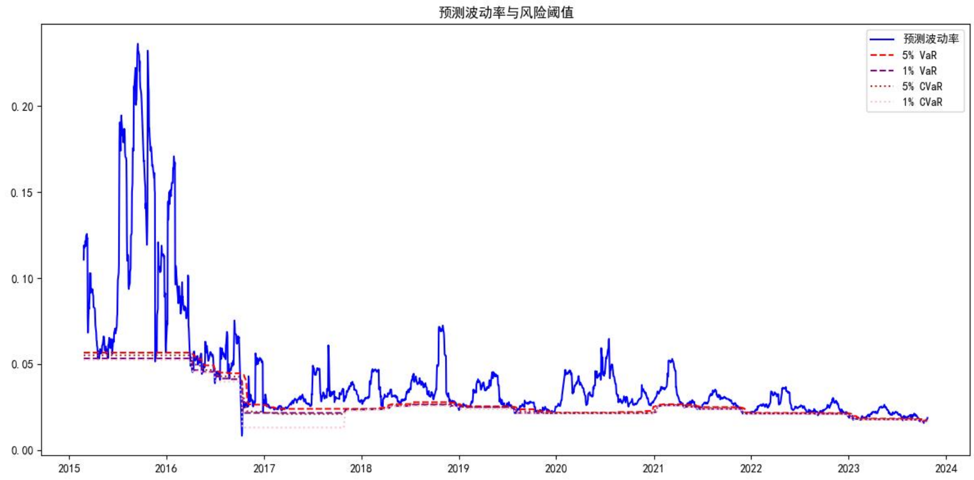

Figure 5: Forecasted Risk Volatility

By dynamically adjusting the stock allocation, the portfolio reduces stock exposure during high-risk periods, effectively minimizing losses due to market downturns. During stable or low-risk periods, stock allocation is appropriately increased to capture opportunities in the rising market.

Although the portfolio's return is relatively low during the risk control process, the cumulative return remains at around 1.06, indicating that robust returns were still achieved while controlling risk.

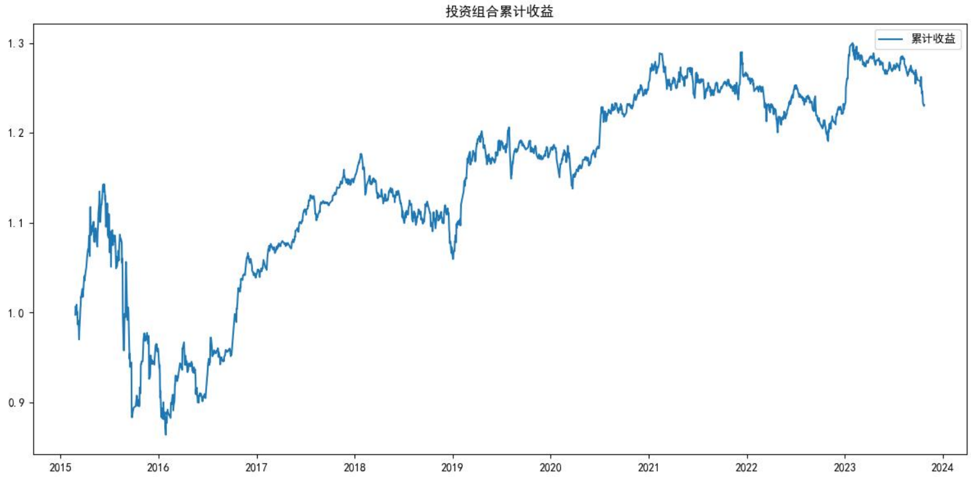

Figure 6: Cumulative Investment Return

V. Comparative Model for Auxiliary Risk Prediction

To comprehensively assess the effectiveness of the pre-risk control system, this study designed a comparative model to be evaluated against the previously mentioned linear regression-based approach, which incorporated lagged features and other indicators The comparative model uses a Neural Autoregressive NAR model for time series prediction, aiming to explore its advantages in risk prediction and its potential in capturing complex features.

The Neural Autoregressive NAR model is a time series forecasting model that directly predicts future values of an entire sequence in one iteration or fixed-length iteration. This model is suitable for scenarios requiring quick forecasts of entire sequences as it can significantly reduce the time required for predictions. It is also suitable for stock risk forecasting.

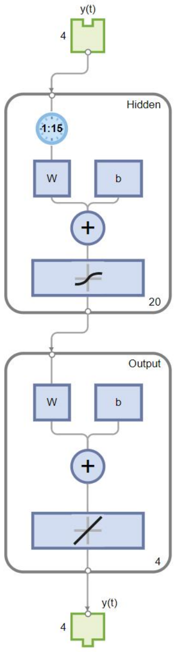

Figure 7: Neural network time series NAR model

5.1 Construction of the Neural Network Time Series NAR Model

5.1.1 Data Preprocessing

First, we use the time series data as input numerical arrays and target sequences. Next, we incorporate the relevant risk indicator data and build a neural network consisting of an input layer, hidden layers, and an output layer. The network learns the mapping relationship between the input sequence and the output sequence. The formula for this process is as follows:

$$\hat{y}_{1:T} = f(x_{1:t}; \theta)$$

Parameter learning is performed via deep learning in the network model, continuously reducing the gap between predicted and actual values and adjusting the parameters accordingly.

5.1.2 Network Construction and Parameter Learning

We build a neural network with an input layer, hidden layers, and an output layer, following these specific steps:

- Input Layer: Incorporates relevant risk indicator data, including lagged features.
- Hidden Layer: Multiple hidden layers are set to capture the complex non-linear relationships between input features. In this study, a hidden layer with 15 neurons was chosen, and activation functions such as ReLU (Rectified Linear Unit) were employed to enhance the model's non-linearity.
- Output Layer: Predicts the volatility or other risk indicators for the upcoming period.
- Parameter Learning: The Levenberg-Marquardt algorithm is used to optimize the network weights by minimizing the mean squared error MSE between predicted and actual values, adjusting parameters to improve prediction performance.

Using the trained NAR model, we predict the average volatility over the next five days using data from the current and previous 15 days. This model allows for the entire target sequence to be predicted at once, enhancing forecasting efficiency.

5.2 Model Error Analysis

By calculating the mean squared error MSE and the correlation coefficient R for the training set, validation set, and test set, we assess the model's predictive performance. The results are as follows:

Table 3: Model training error parameter values

	Observation	MSE	R
Training	1659	0.0139	0.9716
Validation	356	0.0218	0.9521
Test	356	0.0234	0.9471

As shown in the figure, for the neural network time series NAR model, in terms of data, the response variable (prediction target) is an array with 2,386 data points, each having 4 features.

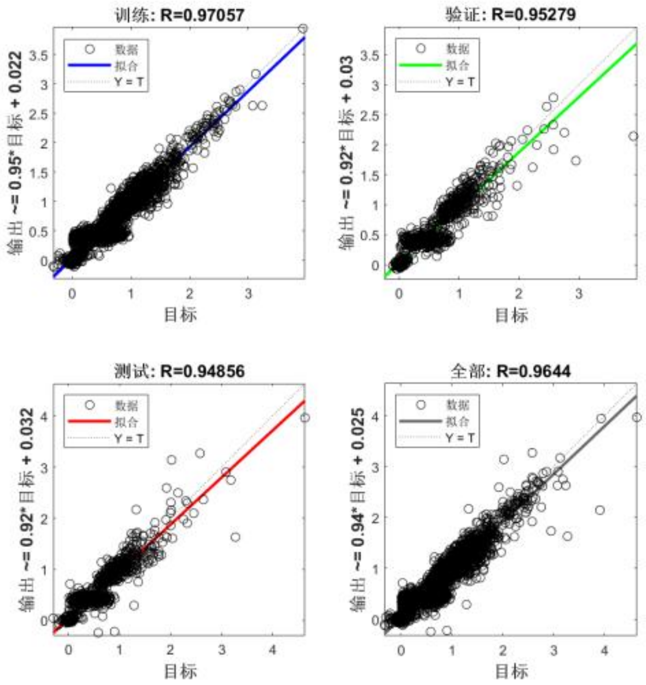

Figure 8: NAR model error parameters

(1) Training set: 1,659 data points, with an MSE of 0.0139 and an R-value of 0.9716.

(2) Validation set: 356 data points, with an MSE of 0.0218 and an R-value of 0.9521. The results of the validation set are used to evaluate the model's performance on unseen data.

(3) Test set: 356 data points, with an MSE of 0.0234 and an R-value of 0.9471. The results of the test set further validate the model's generalization ability.

Further analyzing the correlation coefficients of each dataset, we obtain the following more detailed results:

(1) Training set: correlation coefficient R=0.97057, indicating that the model's fitting effect on the training set is very good.

(2) Validation set: correlation coefficient R=0.95279, suggesting that the model also has a good fitting effect on the validation set.

(3) Test set: correlation coefficient R=0.94856, showing that the model's fitting effect on the test set remains very good.

(4) All data: correlation coefficient R=0.9644, which is the model's fitting effect on the entire dataset.

Overall, the model exhibits a very high degree of fit on all datasets, with R-values close to 1, indicating that the model can predict target values well. There exists a certain degree of overfitting.

5.3 Comparison of Neural Network Model and Linear Regression Model

In this study, the neural network time series NAR model demonstrates significant advantages over traditional linear regression models in risk prediction tasks. First, the NAR model can automatically capture complex nonlinear relationships, while linear regression is only applicable to linearly separable data. Second, neural networks have the ability to process high-dimensional data and feature interactions without manually constructing complex feature combinations, simplifying the model-building process. Moreover, the NAR model is more efficient in capturing time dependencies and can identify long-term dependency patterns, while linear regression relies on limited lag features. Finally, although neural networks are at risk of overfitting, through regularization and model optimization, the NAR model significantly outperforms linear regression in prediction accuracy and generalization ability, providing a more flexible and adaptable risk prediction solution. Therefore, adopting a neural network model not only improves the accuracy of risk prediction but also enhances the overall effectiveness of the ex-ante risk control system.

VI. Conclusion

This study constructs a systemic risk prediction model based on market factors and designs an ex-ante risk control system combined with linear regression methods. By analyzing the data of CSI 300 component stocks from 2014 to 2024, it verifies that the system can achieve stable cumulative returns (6%) while controlling maximum drawdown (within 0.25), outperforming international risk control standards. Introducing the NAR neural network model as a comparison, it demonstrates higher prediction accuracy and generalization ability by capturing complex nonlinear relationships and time dependencies, but with the risk of overfitting. However, the study's limitations lie in the sample time range, which may not cover all market cycles and extreme events, and it does not incorporate important variables such as macroeconomic policies and geopolitical factors. The model's future performance may be affected by dynamic changes in the market. Future research can optimize the model, expand data and variable dimensions, and introduce more asset allocation strategies to enhance the model's robustness and risk resistance capabilities.

References

[1] Feng, Y. X., & Li, Y. M. (2019). Research on the prediction model of CSI 300 index based on LSTM neural network. Mathematics in Practice and Theory (07), 308-315.

[2] Wei, Y. (2010). Research on volatility prediction models of CSI 300 stock index futures. Journal of Management Science (02), 66-76.

[3] Yang, Z. H., Zhang, P. M., & Lin, S. H. (2024). Research on risk linkage and prediction of stock market and bond market—Based on the frontier perspective of machine learning. Journal of Financial Research (01), 131-149.

[4] Yang, Z. H., & Li, D. C. (2021). Do systemic risk indicators have forward-looking predictive ability?. China Economic Quarterly (02), 617-644. doi:10.13821/j.cnki.ceq.2021.02.12.

[5] Bodie, Z., Kane, A., & Marcus, A. J. (2014). Investments. McGraw-Hill Education.

[6] Hull, J. C. (2018). Risk Management and Financial Institutions. Wiley.

[7] Markowitz, H. (1952). Portfolio Selection. The Journal of Finance, 7(1), 77-91.

[8] Sharpe, W. F. (1966). Mutual Fund Performance. The Journal of Business, 39(1), 119-138.

[9] Sortino, F. A., & Van Der Meer, R. (1991). Downside Risk. The Journal of Portfolio Management, 17(4), 27-31.