S&P Index Returns Forecast Based on Ridge and Random Forest

Junyan Tong[1, *]

[1] Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai, China

Abstract. The S&P 500 is a crucial factor in numerous investment decisions. Accurate predictions of its yield data are of paramount importance. This paper employs Ridge and Random Forest models to predict the yield data of the S&P 500 index from September 1929 to August 2023. By comparing the models' Mean Squared Error (MSE) in the prediction set, there is evidence that the Random Forest model outperforms the Ridge model in forecasting. Additionally, a maximum depth setting of 5 for the Random Forest model appears more suitable than 10 or 20. Furthermore, it's worth noting that the number of estimators has a more significant impact on the training set compared to the prediction one. In the case of Ridge regression, there are predictable and unpredictable time intervals. In future prediction studies, it may be advantageous to use the Random Forest model more extensively for forecasting the daily returns of the S&P 500.

1 Introduction

In today's rapidly changing and dynamic financial markets, accurately predicting changes in asset prices is of paramount importance for both investors and financial professionals. Therefore, improving the accuracy of asset price forecasting has been a prominent topic in the field of finance.

Over the past few years, certain academics have endeavored to forecast S&P 500 index yield data by employing regime-switching models or jump-diffusion models [1,2]. Despite the extensive collection of literature regarding asset price prediction, this field still faces challenges. When dealing with datasets with numerous independent variables, the influence of outliers becomes more significant, particularly when using traditional linear regression models [3].

Recently, with the emergence of machine learning technologies, researchers have begun to explore the application of these emerging techniques in asset price prediction. For instance, the Random Forest model is considered to have relatively good predictive power for gold and silver prices [4,5]. In addition, OLS, Lasso, Ridge, and Elastic Net methods have also been used to predict indices from various industries and compare their predictive performance [6]. Some studies have indicated that random forests have predictive capabilities for the direction of individual stock returns [7]. However, current research in this area is relatively limited, especially concerning the comparison between Ridge and Random Forest models and the identification of optimal hyperparameter settings for the S&P 500.

This study aims to address these research gaps. By applying Ridge regression and Random Forest models, this paper seeks to enhance the ability to accurately predict changes in the yield data of the S&P 500 index.". In the experiments, we will focus on the stability of model performance, the impact of different hyperparameter settings on predictive outcomes, and a comparative analysis of the predictive capabilities of the two models in both training and test datasets. In addition to the comparison between these two models, what sets this study apart is its investigation into the dynamic variations of optimal hyperparameters for the models.

This paper follows this outline: In Section 2, we present the data and methods employed. Section 3 delves into the empirical results, while Section 4 serves as the conclusion.

2 Data and methods

This paper primarily employs Ridge and Random Forest regression models to forecast the performance of the S&P Index for the period spanning from September 1929 to August 2023. This chapter introduces the data preprocessing methods and each of the models, which predict returns during these distinct time intervals. The prediction performances are compared by calculating the MSE between the predicted yields and the real ones, followed by the visualization of the prediction results.

2.1 Data

Due to the representative nature of the S&P 500 Index in the stock market and the availability of a sufficient amount of data, this study collected a substantial dataset from Yahoo Finance using Python. Table 1 displays the statistical summary of the daily returns for the S&P 500 Index. Based on the descriptive statistics and the results of the Shapiro-Wilk test, it appears that the returns may not necessarily follow a normal distribution.

The daily return (Daily_Return) is calculated as the percentage change of the closing price between

* Corresponding author: junyan_t@shu.edu.cn

consecutive trading days. Daily returns are a fundamental feature in financial analysis, reflecting the daily performance of the market and serving as the primary dependent variable for return prediction models.

Lagged returns (Lagged Return 1 to Lagged Return 21) represent the past daily returns for a range of 1 to 21 days. Lagged returns are used to capture momentum and mean-reversion effects in financial markets. The inclusion of multiple lags allows the model to detect patterns over different time horizons.

Moving averages (MA 5, MA 10, MA 20, MA 50, MA 100) are calculated as the average closing prices over the specified periods. Moving averages smooth out short-term price fluctuations and highlight longer-term trends, making them a popular tool for trend analysis and trading signals.

Moving standard deviations (STD 5, STD 10, STD 20, STD 50, STD 100) represent the volatility of the closing prices over the specified periods. Volatility is a critical factor in risk assessment and trading strategies. Moving standard deviations provide insights into the market's stability and potential price fluctuations.

The Relative Strength Index (RSI 14, RSI 28, RSI 42, RSI 56, RSI 70) is a momentum oscillator that measures the speed and change of price movements over varying periods. RSI is widely used to identify overbought and oversold conditions in the market, offering potential reversal points that can enhance trading strategies.

Momentum indicators (Momentum 10, Momentum 20, Momentum 60) are calculated as the percentage change in the closing price over the specified periods. Momentum indicators are essential for identifying trends and gauging the strength of price movements, aiding in momentum-based trading strategies.

Volume-based features (Volume MA 5, Volume MA 10, Volume MA 20, Volume MA 50) represent the moving averages of trading volume over different periods. Trading volume is a key indicator of market sentiment and liquidity. Volume-based features help in understanding the underlying market dynamics, especially when combined with price data.

Price ratio features (Open Close Ratio, High Low Ratio) represent the ratios of opening to closing prices and the highest to lowest prices within a trading day. These ratios provide insights into intraday price dynamics and market behavior, often used to detect anomalies and potential reversal points.

Seasonal features (Month, Quarter, Day_of_Week) represent the time of year and day in the trading calendar. Financial markets exhibit seasonal patterns that can influence asset prices. Incorporating seasonal effects helps capture these recurring behaviors and enhances model predictions.

Advanced combinations of the above features include ratios and products of momentum indicators, moving averages, standard deviations, and lagged returns, as well as combinations with seasonal effects. These complex features aim to capture interactions between different market factors, potentially revealing more sophisticated patterns and relationships that single features might miss.

Table 1. Basic Information about S&P 500

	N	AVE	Std	Min	25%	50%	75%	Max	SW Test	SW P
R	1149	0.006	0.053	-0.29	-0.019	0.009	0.035	0.391	0.91660	1.18E-24

2.2 Ridge

Ridge regression is a form of regularization method that excels in dealing with multicollinearity [8].

According to the model observations, when the hyperparameter of Ridge regression approaches infinity, the coefficients of the independent variables tend to approach 0, leaving only the constant term. On the other hand, when the hyperparameter approaches 0, the coefficient estimates are approximately the same as those obtained by ordinary least squares (OLS) estimation.

2.3 Random Forest

Random forests are a machine learning technique that deals with tree models. Initially, they randomly select a dataset of the same size as the training set from it using bootstrap sampling with replacement. Then, within this new training set, they apply a random subset of variables, which is smaller than the total number of independent variables, to generate a full-sized tree [8]. This process is repeated, and the average of these trees' predictions is used as the final prediction. Like kernel regression, random forests give higher weight to points closer to the actual values in the averaging process, effectively increasing the weight of more accurate values [9]. Random Forest is an advancement of the Bagging technique, designed to mitigate the variance of a statistical model by simulating the inherent data variability [10].

3 Results

This section presents the empirical results of the Linear Regression, Ridge Regression, and Random Forest models, which were applied to predict the daily returns of the S&P 500 Index. The models' performance was evaluated based on the Mean Squared Error (MSE) on the test set, the significance of selected features, and the stability of the models' predictions across different hyperparameter settings.

To ensure the robustness of the research results, this study employs various methods. Firstly, it selects the daily returns data of the S&P 500 from 1929 to 2023. By increasing the volume of data, the performance of the models under different economic periods can be considered. In terms of the forecasting method, this

research assumes that the future index returns are related to the past 20 periods of return data.

Simultaneously, this study assumes that after approximately 300 periods, there will be significant changes in the economic environment. Both model hyperparameters and parameters are likely to change, and therefore, the optimal model should be re-estimated. As a result, the sliding window size is set to 50. Within each window, 240 data points are used as the training set to estimate the model's hyperparameters, while the remaining 60 data points serve as the testing set to assess the model's predictive performance.

After obtaining the optimal hyperparameters within the window period, the optimal model is used to estimate the prediction set within that same window period. The evaluation of the model's predictive performance is then conducted by comparing the Mean Squared Error (MSE) on the prediction set.

3.1 Linear Regression Model

The general linear regression model, which incorporates all features, resulted in a Cross-Validation Mean Squared Error (CV MSE) of 6.36E-06 and a Test Mean Squared Error (Test MSE) of 0.001757.

Table 2: Average MSE for Linear Regression

CV MSE	Test MSE
6.36E-06	0.001757

Another Linear Regression model was refined by iteratively excluding features that were statistically insignificant, as indicated by their p-values. This iterative process resulted in the exclusion of features such as 'RSI_28_STD_50' and 'High_Low_Ratio,' among others. The final model included a comprehensive set of features that significantly contributed to predicting daily returns. The Test MSE for this model was 0.001562, indicating robust predictive capability.

The final set of selected features and their corresponding coefficients are presented in Table 3. The model demonstrated high R-squared and Adjusted R-squared values, confirming its ability to effectively explain the variance in daily returns.

Table 3: Final Model Coefficients in the Linear Regression Mode

Feature	Coefficient
Intercept	0.000246
High	0.010921
Low	-0.00501
Volume	-0.00043
Lagged_Return_1	-0.00646
Lagged_Return_2	-0.01031
…	…

3.2 Ridge Regression Model

The Ridge Regression model was evaluated using a range of alpha values to determine the optimal level of regularization. The best performance was observed with an alpha of 1.0, resulting in a Test MSE of (0.001554). Although this MSE was slightly higher than that of the Random Forest model, the Ridge Regression model effectively handled multicollinearity among the predictors.

Table 4 shows the average MSE for different alpha values across the training and test sets. The model's coefficients indicated that Momentum_10 remained a significant predictor with a coefficient of 0.0439, further supporting its importance in forecasting daily returns.

Table 4: Average MSE for Different Alpha Values in Ridge Regression

Alpha	Train MSE	CV MSE	Test MSE
0.01	6.3E-06	9.13E-06	0.001745
0.1	6.19E-06	9.17E-06	0.001695
1	6.23E-06	9.89E-06	0.001554
10	6.74E-06	1.14E-05	0.001713
100	1.16E-05	1.72E-05	0.002039
1000	6.61E-05	8.95E-05	0.002388

By setting coefficients with absolute values less than 0.1 to zero and displaying the rest normally, as seen in the sliding window coefficient tables, it becomes apparent that the optimal hyperparameters can remain relatively stable across consecutive windows. Furthermore, the parameter estimates in these consecutive periods are similar, with consistent parameter signs and significant lag periods. This suggests that the model's effectiveness can be sustained over a period.

The optimal Ridge model and the feature-selected linear regression model achieved roughly the same predictive performance. This indicates that both the Ridge model and the feature-selected linear regression model have a similar effectiveness in preventing overfitting.

3.3 Random Forest Model

The Random Forest model was optimized using a maximum depth of 20, with a sqrt selection for the maximum features, a minimum of 1 sample per leaf, a minimum of 5 samples for a split, and 300 estimators. This configuration resulted in a Test MSE of 8.39e-05, highlighting the model's strong predictive ability.

Table 5: Average MSE for Different Combinations of Hyper-parameters in the Random Forest Model

depth	features	leaf	split	Test MSE
5	sqrt	1	2	0.000135
5	sqrt	1	5	0.000135
5	sqrt	1	10	0.000134
5	sqrt	2	2	0.000134

5	sqrt	2	5	0.000134
5	sqrt	2	10	0.000134
5	sqrt	4	2	0.000135
20	log2	2	5	0.000121
20	log2	2	10	0.000121
20	log2	4	2	0.000121
20	log2	4	5	0.000121
20	log2	4	10	0.000122

The feature importance analysis revealed that Momentum_10 was the most significant predictor, with an importance value of 0.065, followed by MA_5_over_STD_5 (0.052), and MA_5_over_MA_20 (0.050). The complete list of feature importances is provided in Table 6. Notably, features such as 'Stock Splits' and 'Dividends' were assigned zero importance, indicating their lack of contribution to the model's predictive performance.

Table 6: Feature Importances in the Random Forest Model

Feature	Importances
Momentum_10	0.065159
MA_5_over_STD_5	0.052206
MA_5_over_MA_20	0.050475
Momentum_10_MA_5	0.040676
RSI_14	0.030272
...	...

When comparing the Ridge model to the Random Forest model, it's clear that the Random Forest model consistently exhibits lower average MSE and relatively stable optimal hyperparameter settings across various periods. The predictive performance of the Ridge model appears to change over time, whereas the Random Forest model with its optimal hyperparameters seems to consistently perform well on the S&P 500 index.

4 Conclusions

In this study, three models—Linear Regression, Ridge Regression, and Random Forest—were employed to predict the daily returns of the S&P 500 Index. The empirical results indicated that the Random Forest model outperformed both the Linear Regression and Ridge Regression models in terms of predictive accuracy, as evidenced by the lowest Test MSE of 8.39e-05.

The feature importance analysis highlighted that Momentum_10 was a consistently significant predictor across all models, emphasizing its relevance in forecasting daily returns. The Linear Regression model, through an iterative feature selection process, achieved a Test MSE of 0.001562, demonstrating robust predictive performance. However, the Ridge Regression model, despite its effective handling of multicollinearity, had a slightly higher Test MSE.

The optimal hyperparameters and corresponding model parameters exhibit a degree of persistence, meaning they do not change their signs or values dramatically over a relatively long period.

Based on the second conclusion, this paper defines both unpredictable and predictable time windows. It confines the predictions of the Ridge regression model to periods where the optimal hyperparameter is less than 1, achieving better predictive performance.

Larger numbers of estimators and smaller maximum depths for hyperparameters are likely to yield better performance in the training set.

Depth has a significant impact on the prediction set, with larger depths leading to notably better predictive performance, while the other hyperparameters show minimal influence.In comparison to the Ridge regression model, the Random Forest model exhibits significantly smaller Mean Squared Error (MSE), lower volatility, and relatively stable optimal hyperparameters. Overall, the Random Forest model demonstrates superior predictive performance.

The article does have some limitations, as it evaluates the two models solely on the S&P 500 single-index and uses a limited set of hyperparameter settings. This may introduce some bias into the conclusions.

References

1. Haase, F., Neuenkirch, M. ERN: North America (Developed Markets), (2021).
2. Megaritis, A., Vlastakis, N., Triantafyllou, A. Journal of International Money and Finance **113**, (2021).
3. Qu, L. International Journal of Forecasting **37**, (2021).
4. Sadorsky, P. Journal of Risk and Financial Management **14**, (2021).
5. Basher, S. A., Sadorsky, P. Machine Learning with Applications **9**, (2022).
6. Wang, X., Wang, W., Zhang, S. In Proceedings of the 2022 International Conference on Bigdata Blockchain and Economy Management **5**, (2022).
7. Ghosh, P., Neufeld, A., Sahoo, J. K. Finance Research Letters **46**, (2022).

8. Nagel, S. Machine learning in asset pricing **8**, (2021).
9. Athey, S., Tibshirani, J., Wager, S. Generalized random forests, (2019).
10. Aria, M., Cuccurullo, C., Gnasso, A. Machine Learning with Applications **6**, (2021).