
FMA4200 Project Report

Stock Price Forecasting and Efficient Portfolio Construction

Using Machine Learning and Time Series Models in the Chinese A-share Market

Group Members: Celine Williem (120040005), Yohanes James (120040006), Edward Jayadi Halim (120040019), Jefferson Joseph Tedjojuwono (120040023)

1. Abstract

This project evaluates the effectiveness of various forecasting models for estimating mean returns within the mean-variance optimization framework. Our stocks universe for the purpose of this project are stocks within the SSE50 that have over 120 monthly returns from 2012 to 2022, which amounts to 38 stocks. The data from 2012 to 2021 were used as training data and data in 2022 were used as the testing data. We contrasted the performance of several predictive models against two baselines: the simple average of past one-year monthly returns and the market return calculated as the mean of all selected stock returns. The models tested include both time series models such as ARIMA, SARIMAX, and Exponential Smoothing, as well as machine learning models that includes Gradient Boosting Model (GBM), Long Short-Term Memory (LSTM) networks, Support Vector Machines (SVM), Random Forest, and Linear Regression. Our initial findings indicate that traditional time series models like ARIMA and SARIMAX were outperformed by the market return. Notably, the Random Forest model achieved the highest cumulative returns, followed by GBM and Linear Regression. These results suggest that advanced machine learning models can potentially offer more accurate mean return estimates for portfolio optimization than simple historical averages, which may lead to enhanced decision-making in investment strategies.

2. Introduction

Investment strategies and financial portfolio management crucially depend on accurate forecasting of stock returns and effective estimation of risk. A fundamental challenge in portfolio management is optimizing the Sharpe ratio within the mean-variance framework, which hinges on precise estimates of expected returns and covariance matrices. Traditional approaches often rely on historical averages to estimate these parameters, which can lead to significant prediction errors and suboptimal investment decisions.

The objective of this project is to explore and compare the effectiveness of various forecasting models in accurately predicting stock returns, essential for optimizing the Sharpe ratio. By improving the estimation of expected returns and employing these forecasts to

determine optimal portfolio weights, investors can potentially achieve higher adjusted returns for a given level of risk.

Our study focuses on stocks included in the SSE50 index, specifically selecting those with a substantial history of monthly returns (over 120 months). This dataset provides a robust basis for testing and comparing different predictive models. We examine traditional time series forecasting methods such as ARIMA, SARIMAX, and Exponential Smoothing, as well as advanced machine learning techniques including Linear Regression, Gradient Boosting Models (GBM), Random Forest, Support Vector Machines (SVM), and Long Short-Term Memory (LSTM) networks. These models are evaluated based on their ability to forecast returns used in optimizing the Sharpe ratio.

This report will outline the implementation of these forecasting models, demonstrate how the forecasted returns and historical covariance data are used to maximize the Sharpe ratio, evaluate model performance against traditional benchmarks, and discuss the implications of our findings for portfolio optimization. Through this analysis, we aim to identify which models provide the most reliable and accurate forecasts of future stock returns, thereby informing better investment strategies and contributing to the field of financial analytics.

In subsequent sections, we will review the relevant literature, describe our methodology and data sources in detail, present our findings, and discuss the broader implications of our results for both academic research and practical investment management.

3. Literature Review

The article "Forecasting Stock Market Prices Using Machine Learning and Deep Learning Models: A Systematic Review, Performance Analysis and Discussion of Implications," by Sonkavde et al. (2023) provides a comprehensive review of the current state of machine learning (ML) and deep learning (DL) applications in financial forecasting. This study highlights various ML and DL techniques that have been utilized for predictive accuracy in stock price movements, encompassing a range of methodologies from simple regression models to complex neural networks.

The research emphasizes the effectiveness of ensemble methods and advanced neural networks in enhancing prediction accuracy. Notably, it discusses the integration of models like Random Forest, XG-Boost, and LSTM, illustrating their capability to outperform traditional statistical methods. The systematic review also touches on the importance of incorporating multiple data sources, including economic indicators and market sentiment, to improve the robustness of predictive models.

Inspired by the findings from this research, our project aims to apply advanced ML and DL techniques to the Chinese stock market, specifically focusing on stocks within the SSE50 index that have over 120 monthly returns from 2012 to 2022. This subset of 38 stocks provides a rich dataset for analysis, allowing for a detailed exploration of ML and DL models in a market known for its volatility and unique regulatory environment. By applying these

sophisticated forecasting tools, we seek to uncover deeper insights into market dynamics and improve investment strategies through more accurate predictions of stock movements.

This literature review sets the stage for a detailed examination of how modern computational techniques can be leveraged to enhance forecasting accuracy and financial decision-making.

4. Data Collection and Preprocessing

Data Description

This study utilizes monthly return data for stocks listed in the SSE50 index, specifically focusing on those with over ten years of monthly returns to ensure robust historical data for analysis. A complete list of all 50 stocks included in the SSE50 index can be found in the “SSE 50 Company Codes.txt” file and the filtered 38 stocks used is also provided under the “SSE Over 10 Years.txt” file. These stocks represent a diverse array of sectors, offering a comprehensive view of the market dynamics within the Shanghai Stock Exchange.

Data Sources

The monthly returns data, along with additional financial metrics such as the monthly risk-free rate, Price-to-Earnings (P/E) ratio, and Price-to-Book (P/B) ratio, were sourced from [CSMAR](#) (China Stock Market & Accounting Research Database).

Data Preprocessing

The preprocessing of the data was conducted to ensure the dataset was well-structured and suitable for analysis. This involved several key steps, which are detailed in the Python script “Data Preprocessing.py”. The primary steps included:

1. Stock Selection: We began by filtering the dataset to include only those stocks from the SSE50 index that had more than 120 months of returns. This criterion was set to ensure that each stock had sufficient historical data for robust analysis and model training.\
2. Data Consolidation: Since the monthly returns data was dispersed across multiple Excel files, a significant preprocessing step involved combining these files into a single dataset.
3. Data Structuring: The data was then pivoted to align all monthly returns by stock and date, facilitating easier access and manipulation in subsequent modeling stages.
4. Handling Missing Values: Any missing data points were either filled using appropriate imputation methods or dropped, depending on their nature and the potential impact on the integrity of the dataset.

Rationale for Data Selection

The selection of stocks from the SSE50 index was chosen to balance between the diversity and manageability of the dataset. Using a small amount of stocks would lead to a less diverse stock universe and too large amounts of stocks would lead to bad performance of the models due to the additional computational complexity. Using the stocks in the SSE50 index seems

optimal since the SSE50 represents a broad spectrum of the top 50 stocks by market capitalization on the Shanghai Stock Exchange, providing a representative sample without overwhelming computational resources. We further refined this to 38 stocks based on the availability of over 120 monthly returns from 2012 to 2022, ensuring each stock had a robust historical record for reliable analysis. This size strikes an optimal balance, preventing the complexities and increased computational demands associated with larger datasets, while still capturing a comprehensive view of the market trends.

5. Methodology

This section outlines our systematic approach to forecasting stock returns, optimizing portfolios based on these forecasts, and evaluating their performance. Central to our methodology is a function designed to maximize the Sharpe ratio by determining optimal portfolio weights.

Portfolio Optimization Function

Our function “maximize_sharpe_ratio” (which can be found in the code.ipynb file), calculates optimal portfolio weights to maximize the Sharpe ratio, which measures risk-adjusted returns. It takes as input the mean return vector, covariance matrix, and risk free rate and then output the optimal weights for those input data.

Components of the Function:

- Portfolio Variance: Calculates the variance of the portfolio's returns as a measure of risk. The variance of the portfolio's returns is computed using the formula:

$$\text{Protfolio Variance} = \mathbf{W}^T \mathbf{Cov} \mathbf{W}$$

where \mathbf{W}^T is the transpose of the weight vector, and \mathbf{Cov} is the covariance matrix of the stock returns.

- Portfolio Return (portfolio_return): Computes the expected return of the portfolio. The expected return of the portfolio is calculated by the dot product of the weights vector with the vector of expected returns:

$$\text{Portofolio Return} = \mathbf{W}^T \boldsymbol{\mu}$$

where $\boldsymbol{\mu}$ represents the mean returns vector for the stocks in the portfolio.

- Sharpe Ratio (sharpe_ratio): Defined negatively to facilitate minimization, integrating both portfolio variance and return. The Sharpe ratio is calculated by dividing the excess return of the portfolio over the risk-free rate by the standard deviation (square root of the variance) of the portfolio's returns:

$$\text{Sharpe Ratio} = \frac{\mathbf{W}^T \boldsymbol{\mu} - r_f}{\sqrt{\mathbf{W}^T \mathbf{Cov} \mathbf{W}}}$$

where r_f is the risk-free rate.

For our optimization we set the constraint that the weights must be non-negative (meaning no short selling) but we allow it to be as large as possible, the sum of the weights does not necessarily have to equal to 1. After obtaining the weights from the function we can simply divide them by the sum of the weights to normalize them and have them sum to 1, for fairness in model comparison.

Baseline Models

To assess the effectiveness of our predictive models, we establish two baseline models:

- Simple Mean Return Model: Calculates expected returns using the mean of past one-year monthly returns and then uses these simple mean returns as the return vector to be passed onto the optimization function.
- Market Return Model: Get the return of the market as the average returns of all stocks in that month without having to use the optimization function.

Forecasting Models:

- **Time Series Models**

- 1) ARIMA (AutoRegressive Integrated Moving Average): Models the dependencies in time series data.
- 2) SARIMAX (Seasonal ARIMA with eXogenous variables): Extends ARIMA for seasonal effects and external influences.
- 3) Exponential Smoothing: Forecasts future data by assigning exponentially decreasing weights over time.

- **Machine Learning Models**

- 1) LSTM (Long Short-Term Memory): A type of recurrent neural network capable of learning order dependence in sequence prediction problems.
- 2) SVM (Support Vector Machine): Used for regression challenges to model non-linear relationships.
- 3) GBM (Gradient Boosting Machine): An ensemble technique that builds models sequentially to correct the errors of prior models.
- 4) Random Forest: An ensemble of decision trees, typically used for regression and classification.
- 5) Linear Regression: Predicts outcomes based on linear relationships between input variables.

Results Visualization

The Results section will include plots comparing market returns to returns from each model portfolio, as well as comparing cumulative returns of the portfolio returns for each model with the market returns. These visuals are crucial for assessing the empirical performance of our models using real-world data.

Performance Evaluation

We evaluate each model based on the actual returns of optimized portfolios compared to our baseline models, determining the most effective forecasting method for enhancing portfolio performance.

To aid in the analysis:

- The monthly returns of all 38 stocks during the testing period were plotted to visualize any trends or anomalies. These plots are included in Appendix A.
- For each forecasting model, the predicted returns were plotted against the actual returns to assess model accuracy. These comparative plots are found in Appendix B.

6. Results and Conclusion

Baseline Models

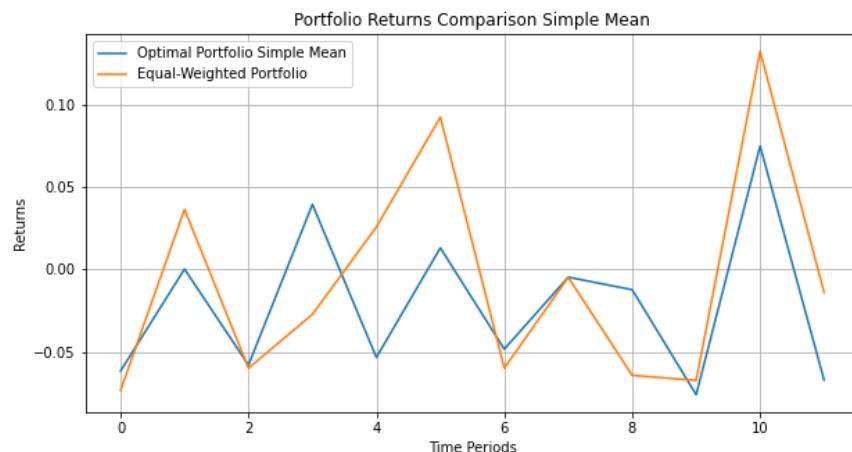


Figure 1 Baseline Model Returns (Simple Mean)

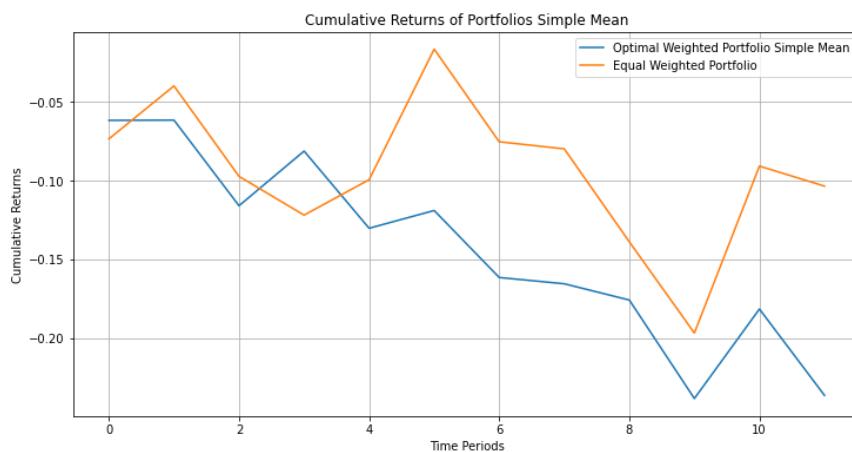


Figure 2 Baseline Model Cumulative Returns (Simple Mean)

ARIMA (AutoRegressive Integrated Moving Average)



Figure 3 ARIMA Returns

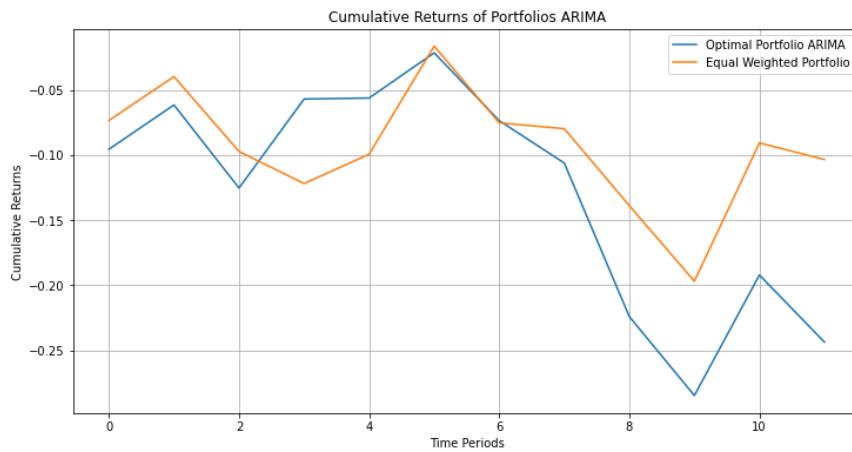


Figure 4 ARIMA Cumulative Returns

SARIMAX (Seasonal ARIMA with eXogenous variables)

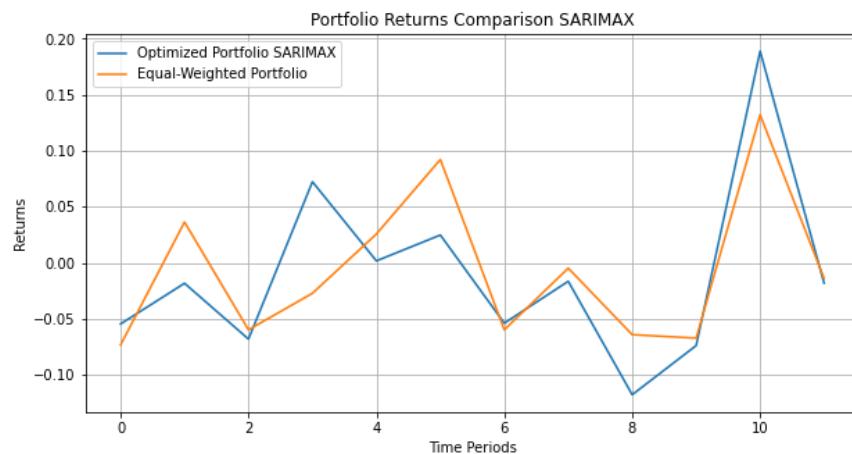


Figure 5 SARIMAX Returns

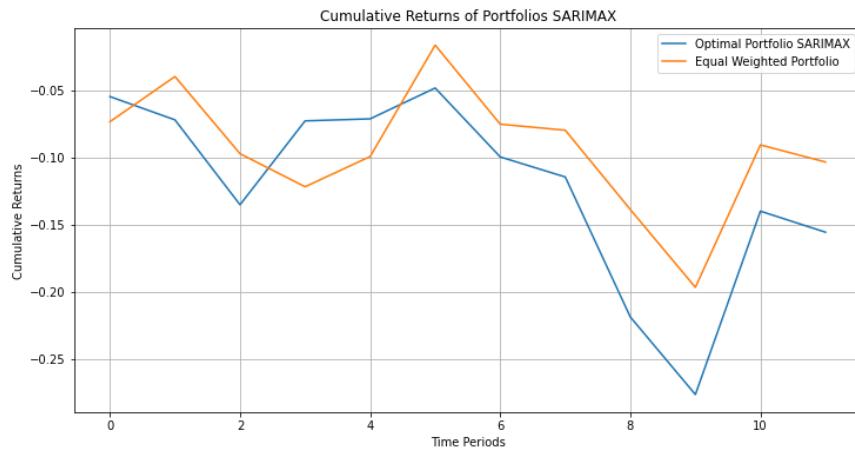


Figure 6 SARIMAX Cumulative Returns

Exponential Smoothing

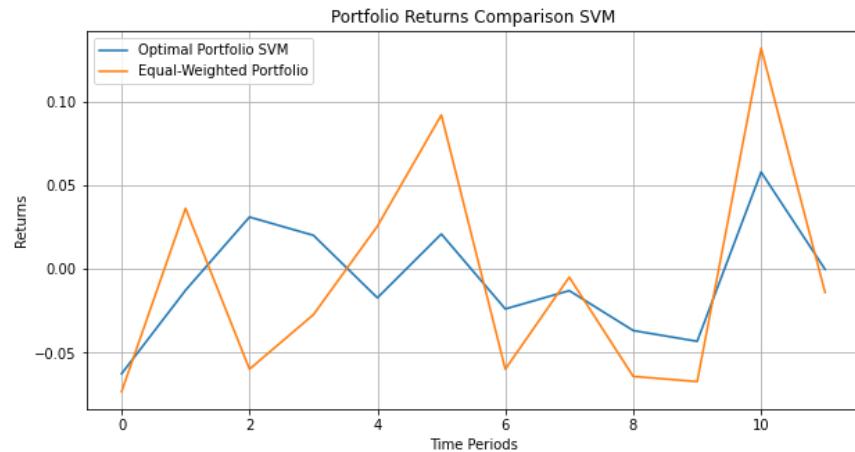


Figure 7 Exponential Smoothing Returns

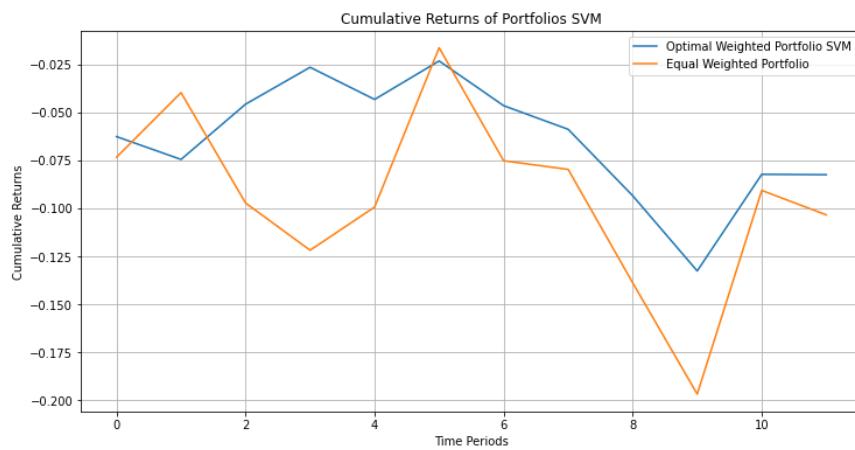


Figure 8 Exponential Smoothing Cumulative Returns

LSTM (Long Short-Term Memory)

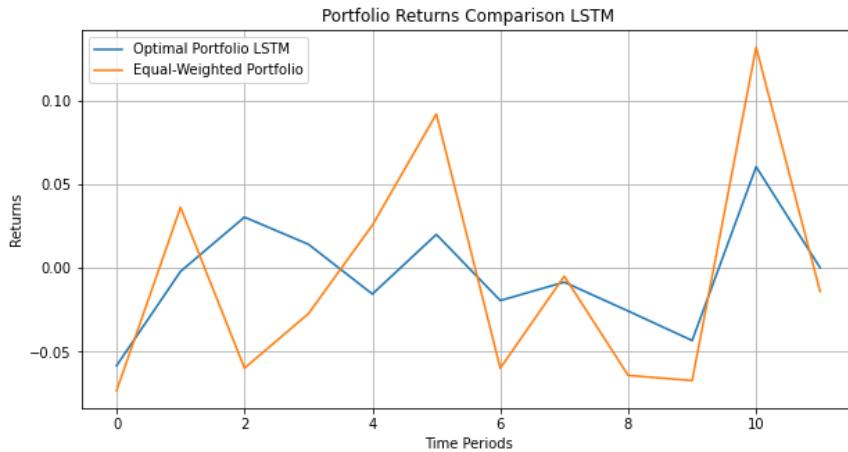


Figure 9 LSTM Returns



Figure 10 LSTM Cumulative Returns

SVM (Support Vector Machine)

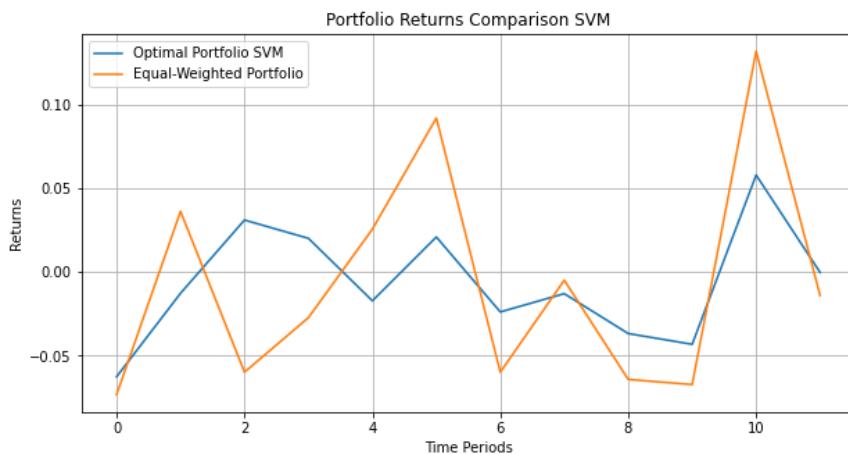


Figure 11 SVM Returns



Figure 12 SVM Cumulative Returns

GBM (Gradient Boosting Machine)

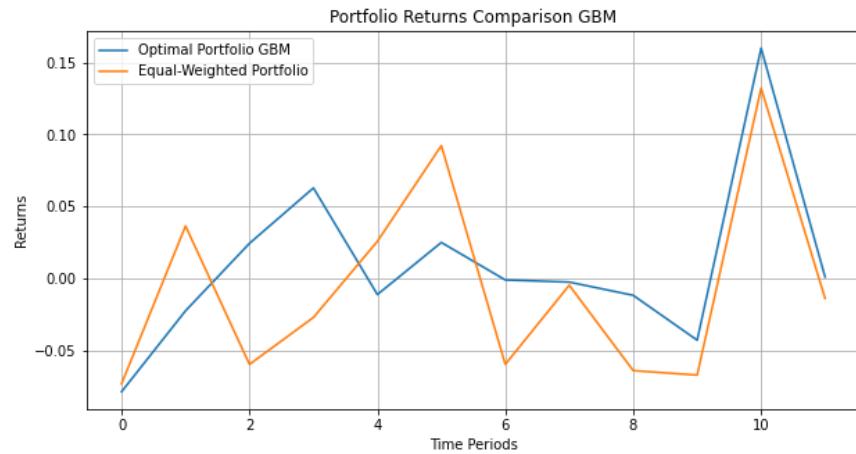


Figure 13 GBM Returns

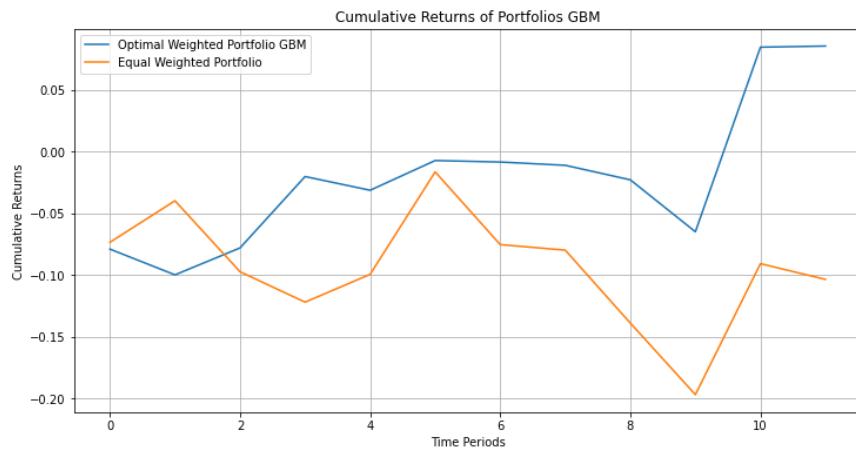


Figure 14 GBM Cumulative Returns

Random Forest

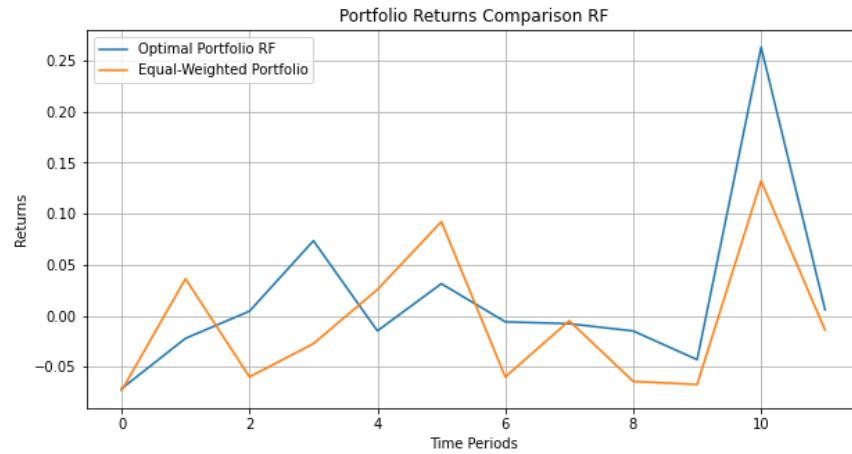


Figure 15 Random Forest returns

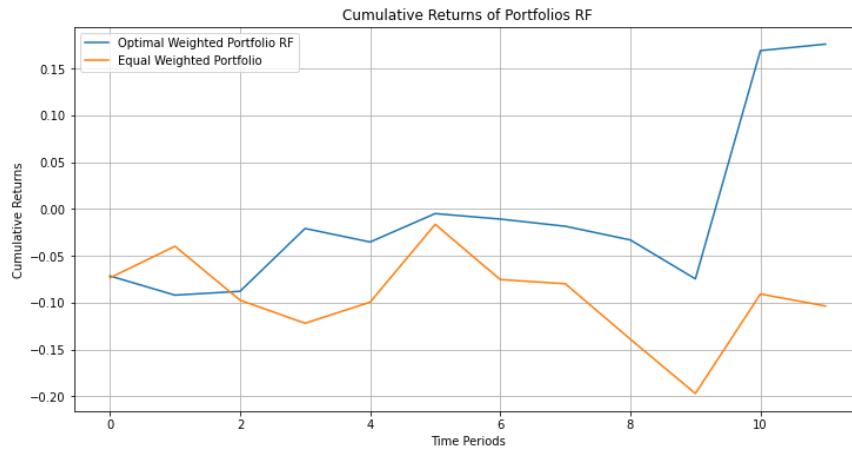


Figure 16 Random Forest Cumulative Returns

Linear Regression

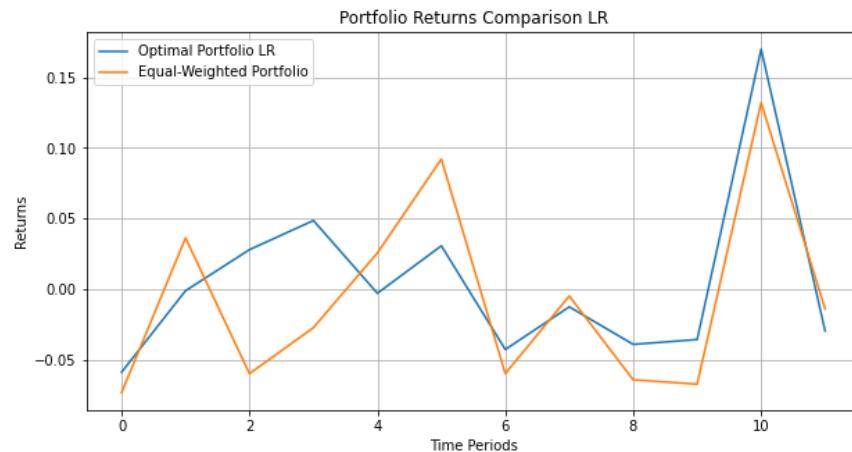


Figure 17 Linear Regression Returns

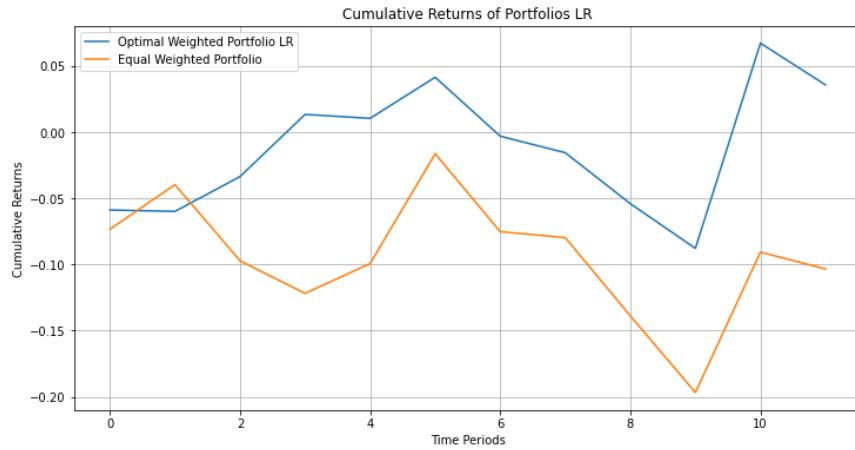


Figure 18 Linear Regression Cumulative Returns

All Models Comparison

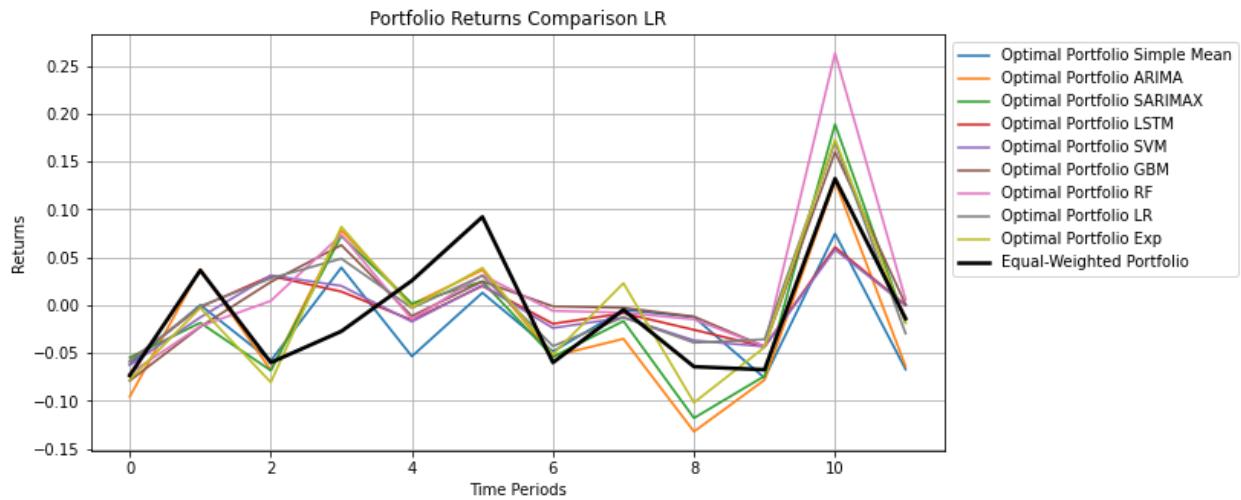


Figure 19 All Models Returns

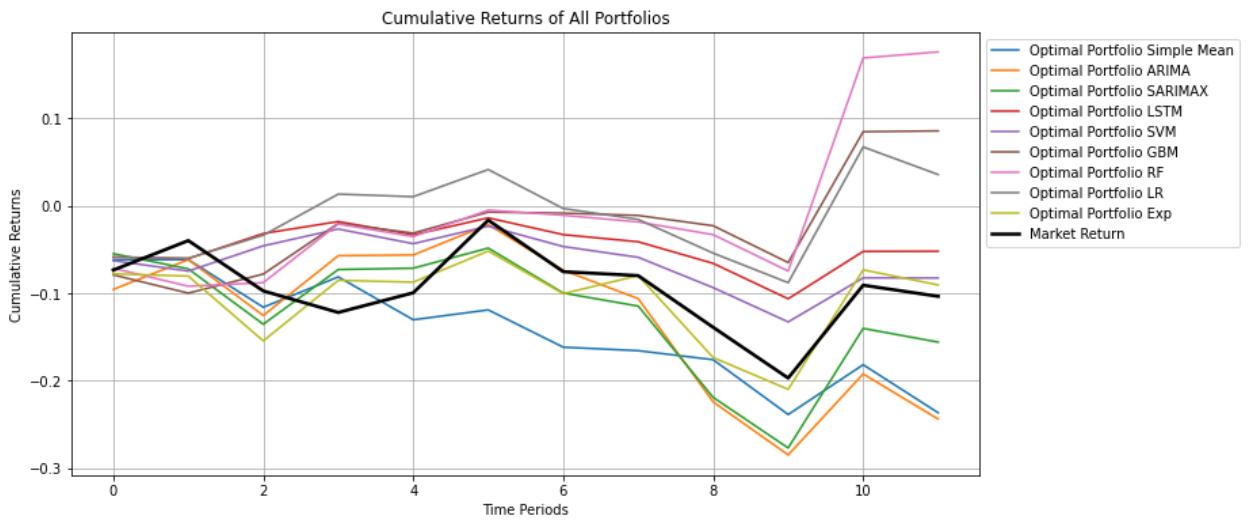


Figure 20 All Models Cumulative Returns

Discussion of Results

In the discussion of the results, it's evident that the performance of various models varies significantly, with distinct patterns emerging among the different types of models employed. Notably, the baseline Simple Mean model, along with the ARIMA and SARIMAX time series models, underperform relative to the market return in terms of overall cumulative returns. This underperformance could suggest limitations in the ability of traditional time series models to capture and leverage patterns in stock price movements effectively under the conditions tested.

In contrast, the Exponential Smoothing model shows a marginal improvement over the market return by the end of the testing period. This slight edge indicates that while simple time series methods might sometimes capture general market trends, they do not consistently provide superior returns, especially in volatile or complex market conditions. This result reinforces the notion that relying solely on traditional time series models may not be sufficient for optimizing investment strategies.

On the other hand, machine learning models demonstrate a more robust performance, consistently surpassing the market return. Notably, the Random Forest model achieves the highest final cumulative return, followed by the Gradient Boosting Machine (GBM) and Linear Regression models. The superior performance of these machine learning models suggests their greater efficacy in discerning and exploiting complex patterns in the data, likely due to their ability to model non-linear relationships and interactions that are not readily apparent or accessible to simpler statistical or time series models.

This variance in performance underscores the importance of choosing the right model based on the specific objectives and constraints of the investment strategy, with machine learning models providing a compelling option for scenarios where capturing complex market dynamics is critical.

The final cumulative returns of all the models on the last month of the testing period is listed in the table below ordered ascendingly:

Model	Final Cumulative Returns
ARIMA	-0.2436942325821847
Simple Mean	-0.23650128981019125
SARIMAX	-0.15575594113950597
Market Return	-0.10344144224650875
Exponential Smoothing	-0.0903985372848991
SVM	-0.08249145555763393
LSTM	-0.05189191530285808
Linear Regression	0.0357402929015731
GBM	0.08566223766447356
Random Forest	0.17592660511426317

Table 1 Final Cumulative Returns Comparison

7. Conclusion

This study evaluated various forecasting models' effectiveness in predicting stock returns and optimizing portfolio performance using a mean-variance framework. Machine learning models, notably Random Forest, GBM, and Linear Regression, demonstrated superior performance over traditional time series models and market returns during the one-year test period. These models excelled in capturing complex market dynamics, offering significant improvements over simpler approaches like ARIMA, SARIMAX, and the baseline Simple Mean model.

However, the analysis is constrained by its duration and scope; it only spans one year and is limited to 38 stocks from the SSE50 index. These limitations suggest the results may not fully represent different market conditions or generalize across broader portfolios. Future research should extend the timeframe and diversify the stock universe to enhance the robustness and applicability of the findings. Exploring more sophisticated machine learning techniques and integrating alternative data sources could also provide further insights into effective investment strategies.

8. References

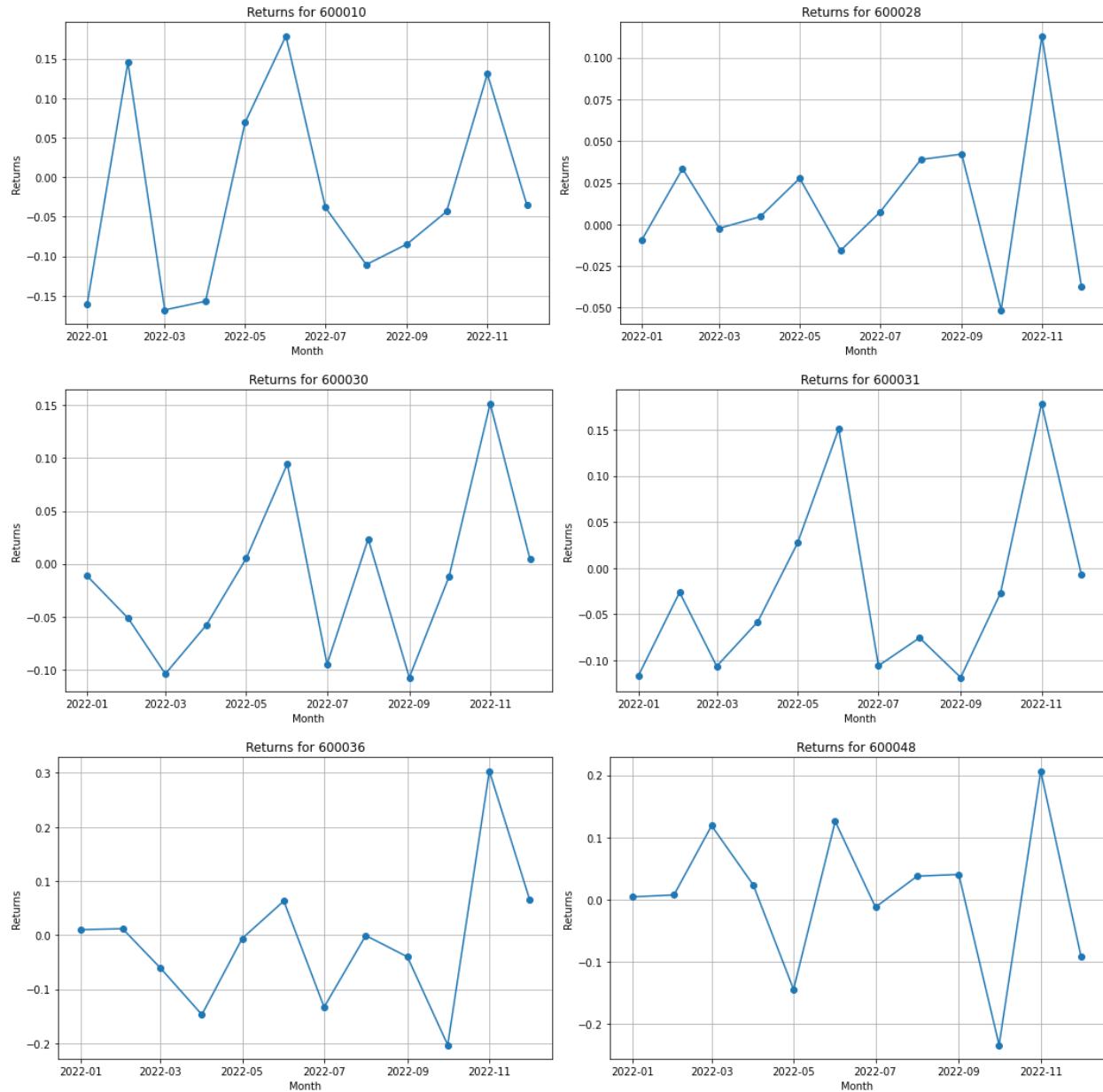
- Sonkavde, G., Dharrao, D. S., Bongale, A. M., Deokate, S. T., Doreswamy, D., & Bhat, S. K. (2023). Forecasting stock market prices using machine learning and Deep Learning Models: A systematic review, performance analysis and discussion of implications. International Journal of Financial Studies, 11(3), 94.
<https://doi.org/10.3390/ijfs11030094>

Member contribution:

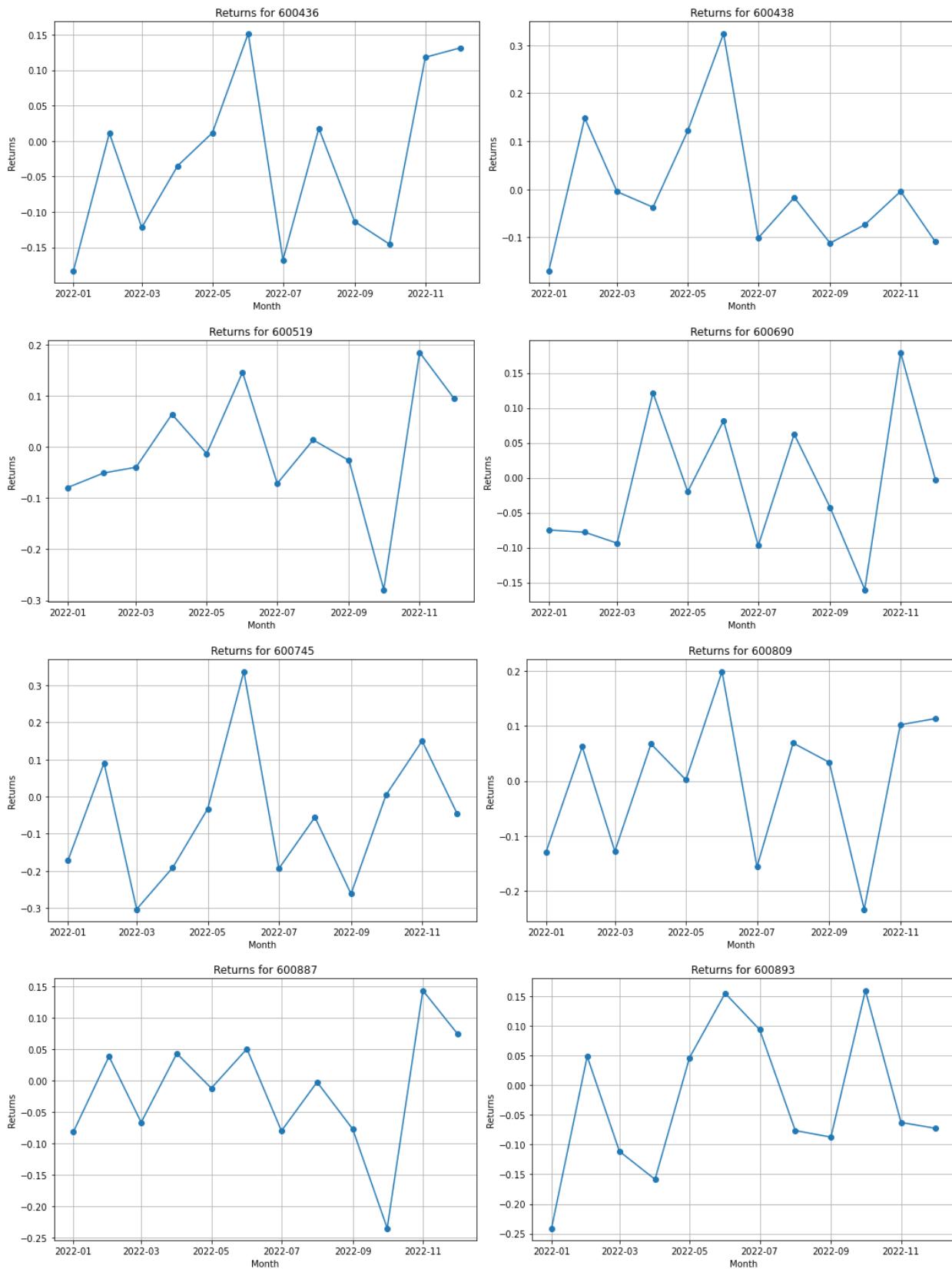
- 1) Celine Williem (120040005): Data collection and preprocessing, integration, and report
- 2) Yohanes James (120040006): Machine Learning Models
- 3) Edward Jayadi Halim (120040019): Time Series Models
- 4) Jefferson Joseph Tedjojuwono (120040023): Baseline Model, visualizations, and testing

Appendix

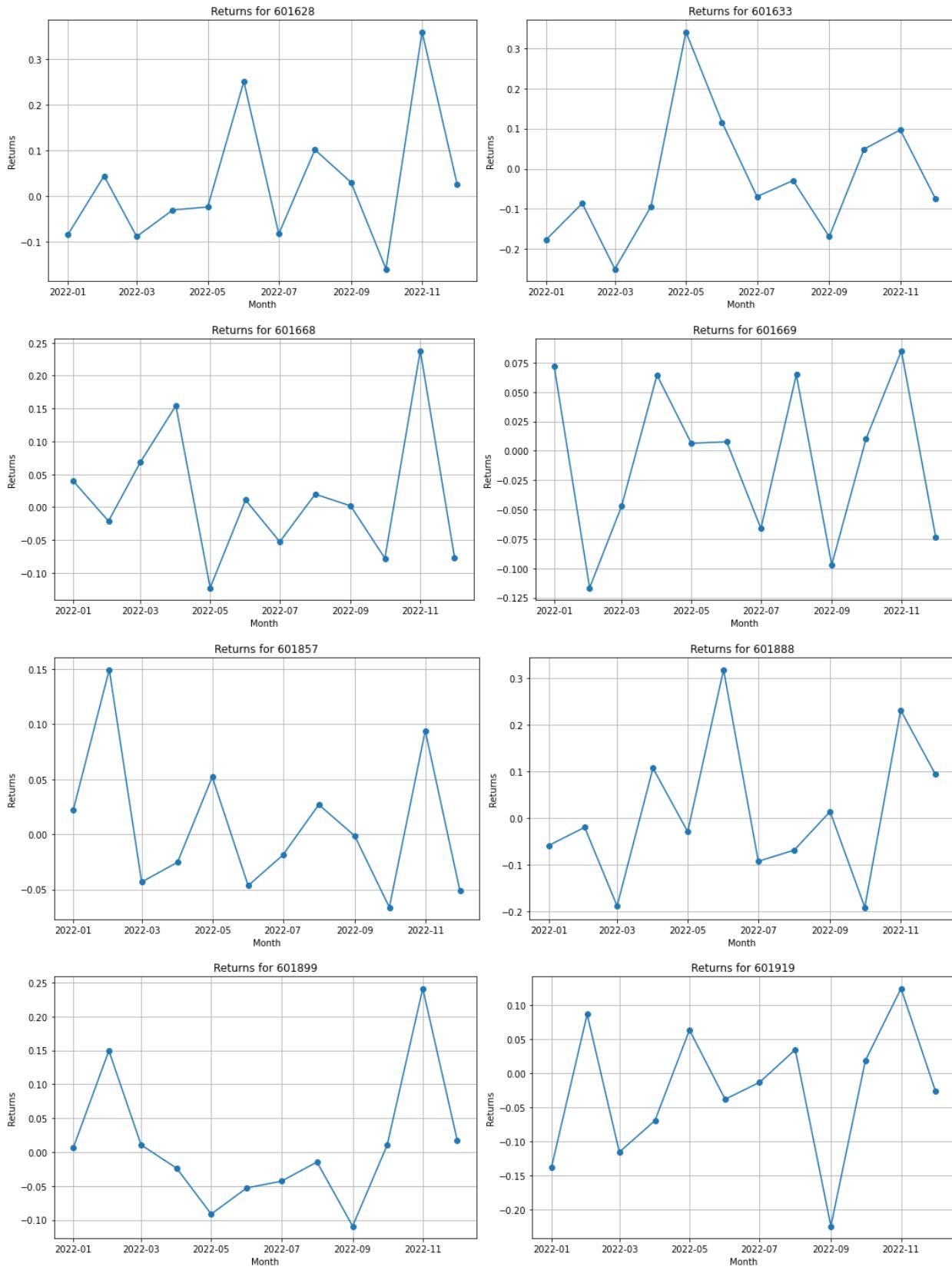
Appendix A: Visualization of individual stock returns











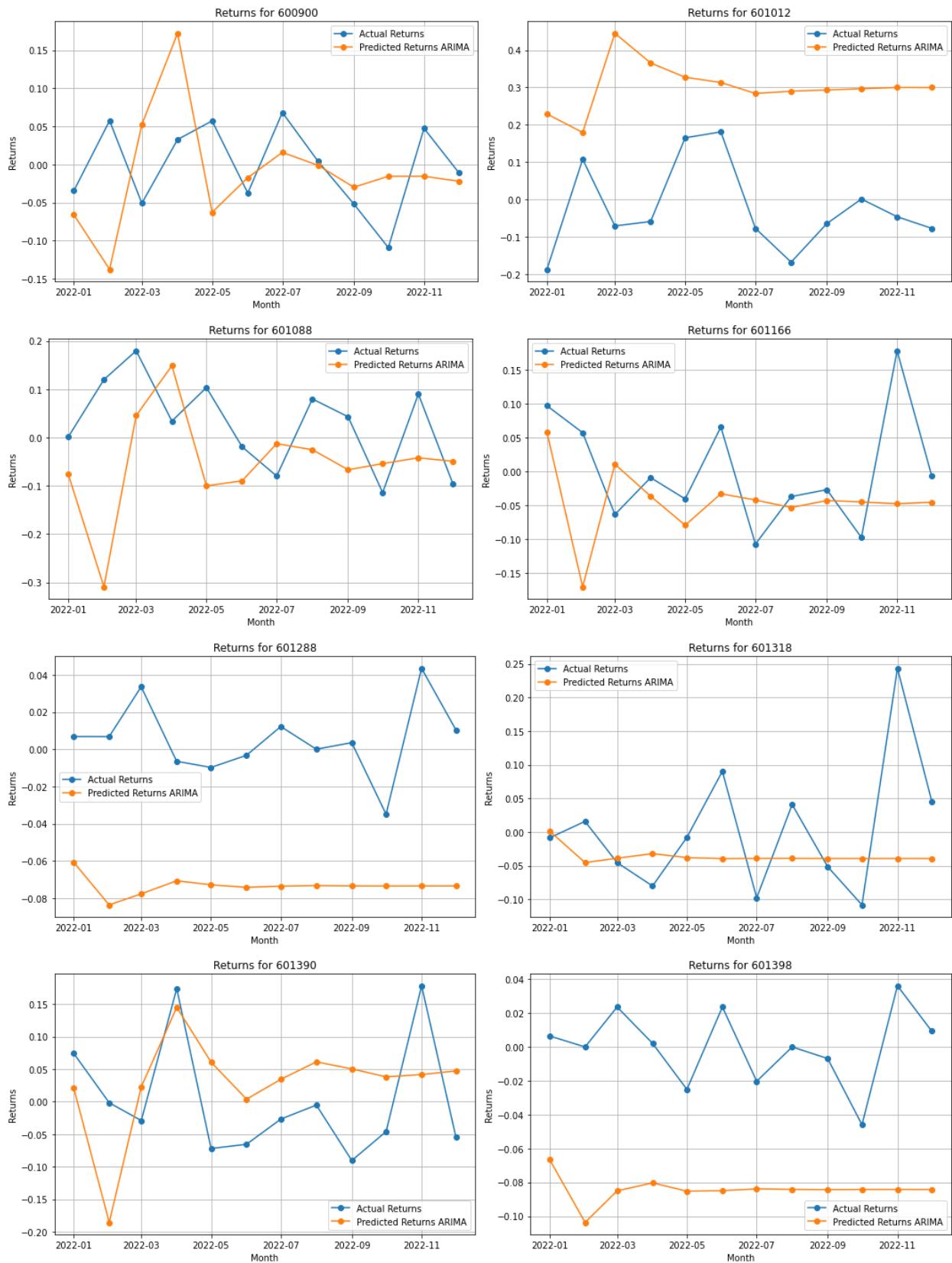
Appendix B: Visualization of predicted returns compared to individual stock returns

ARIMA





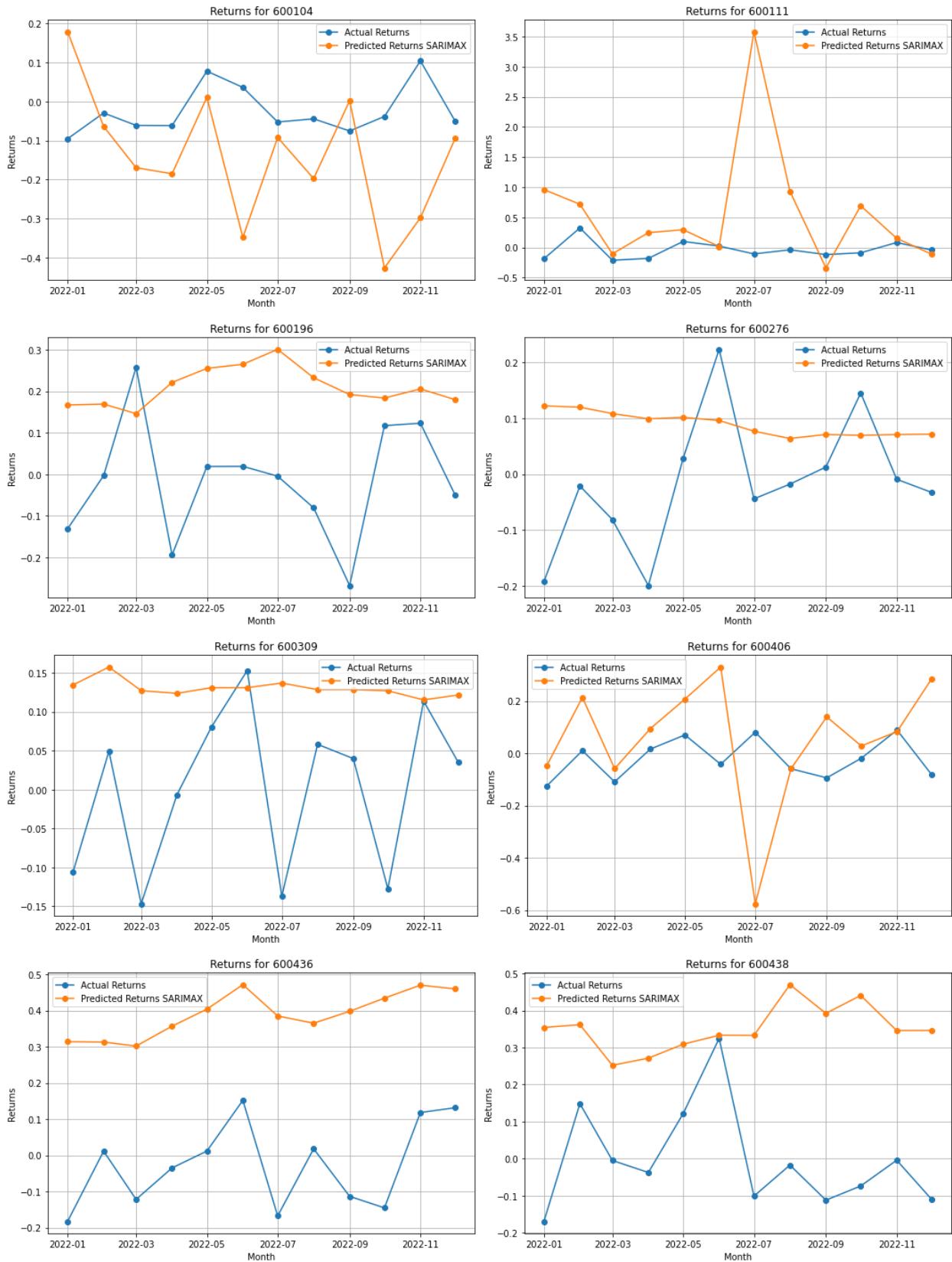




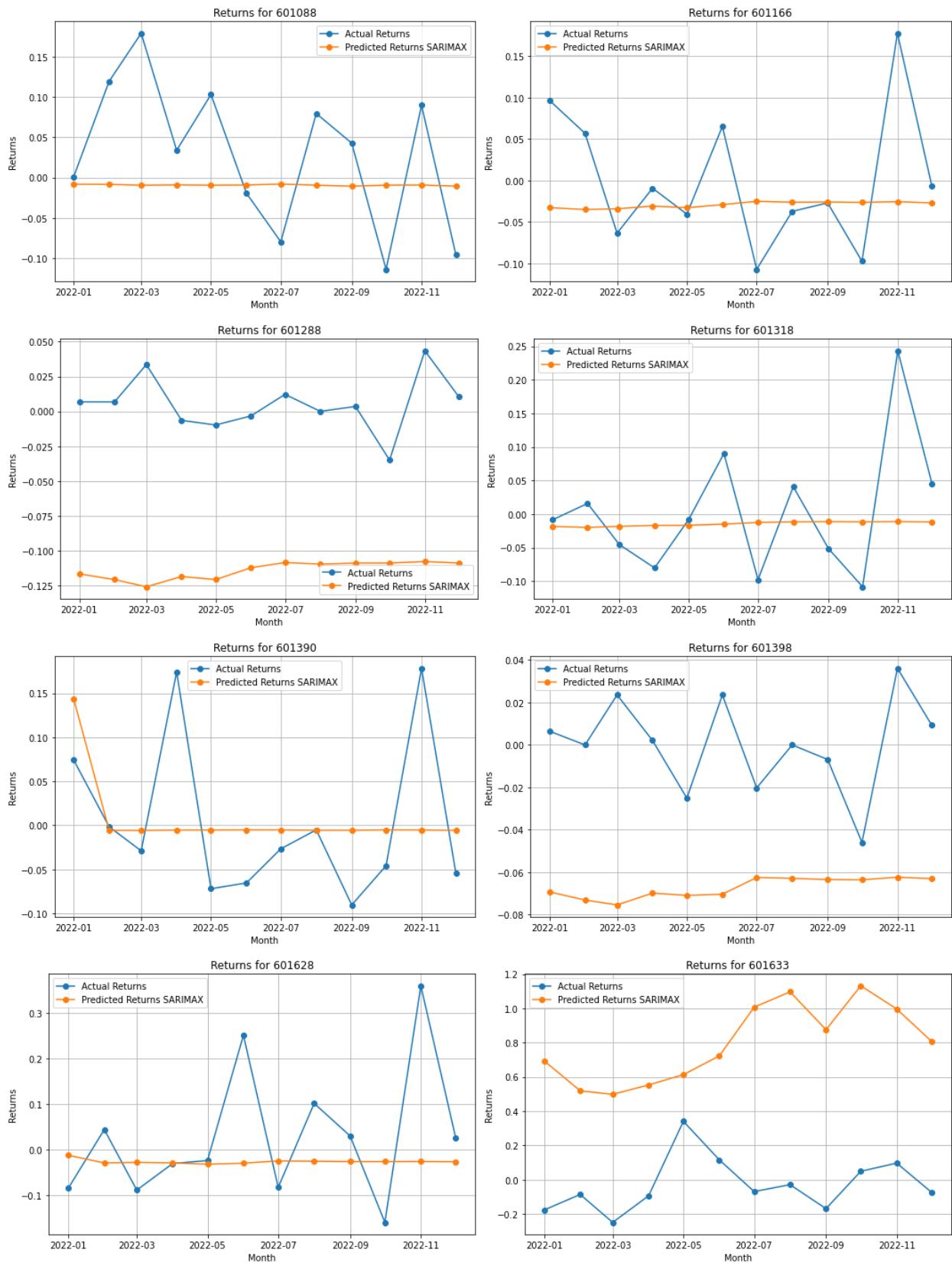


SARIMAX











Exponential Smoothing





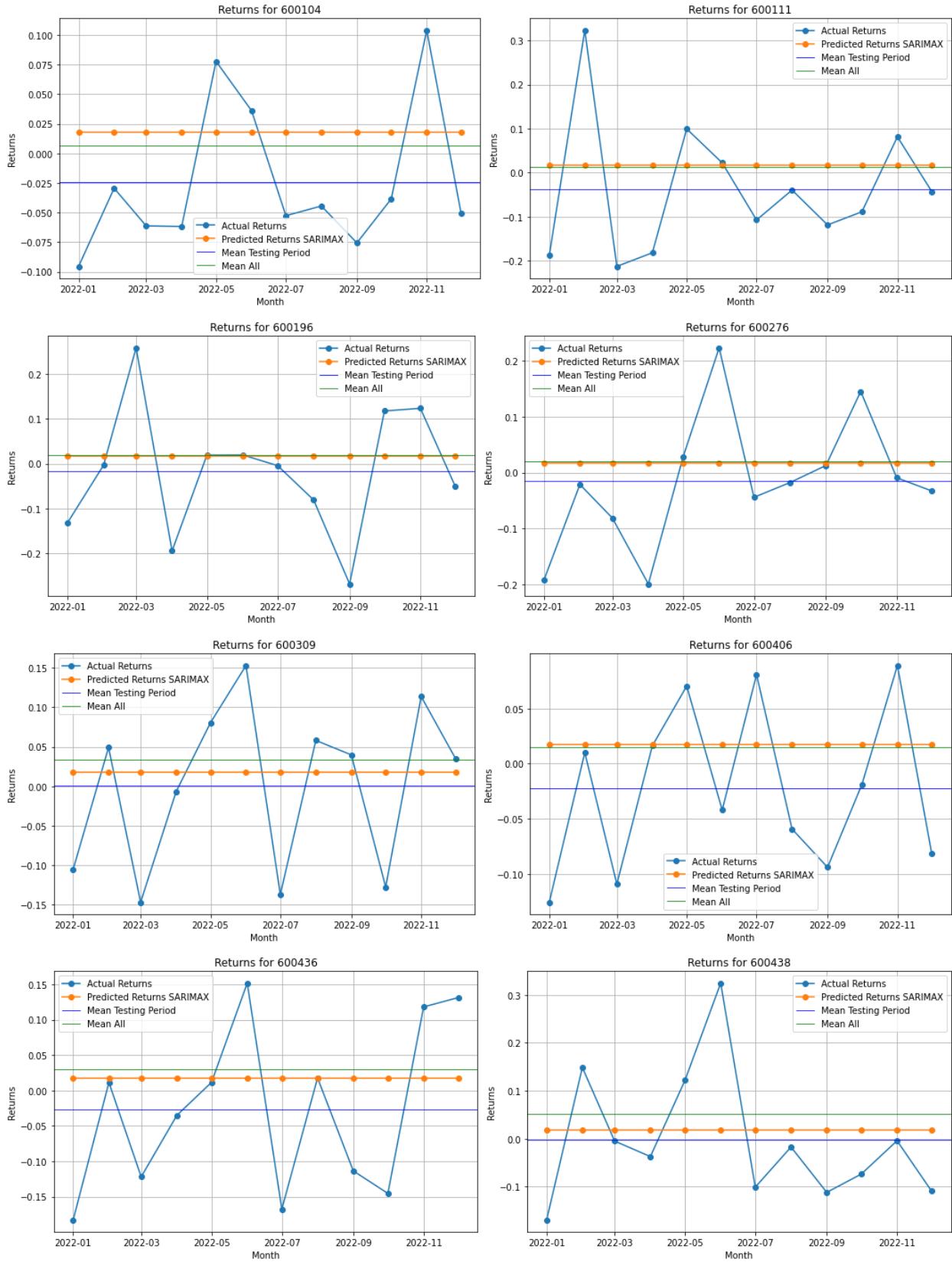






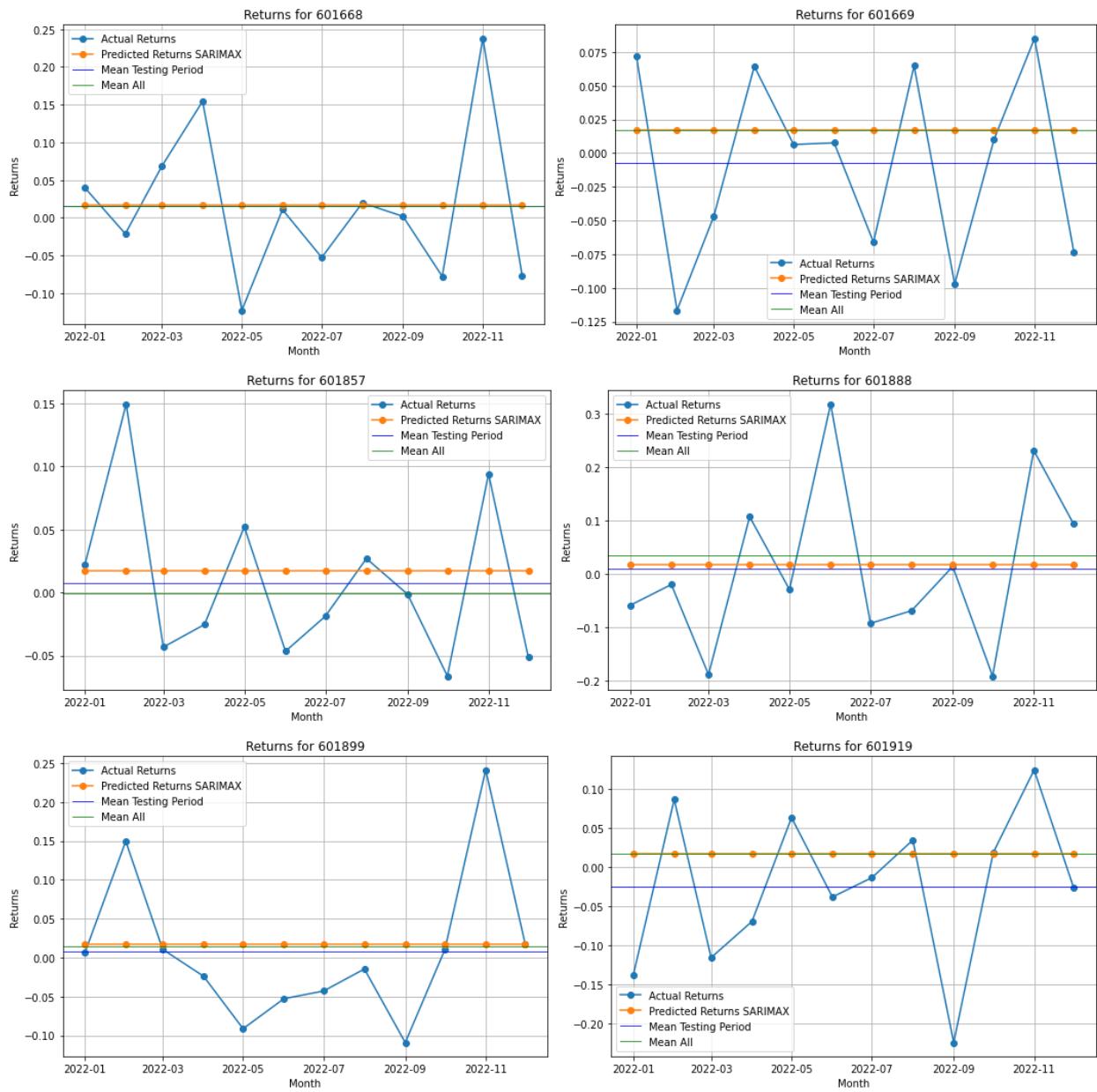
LSTM









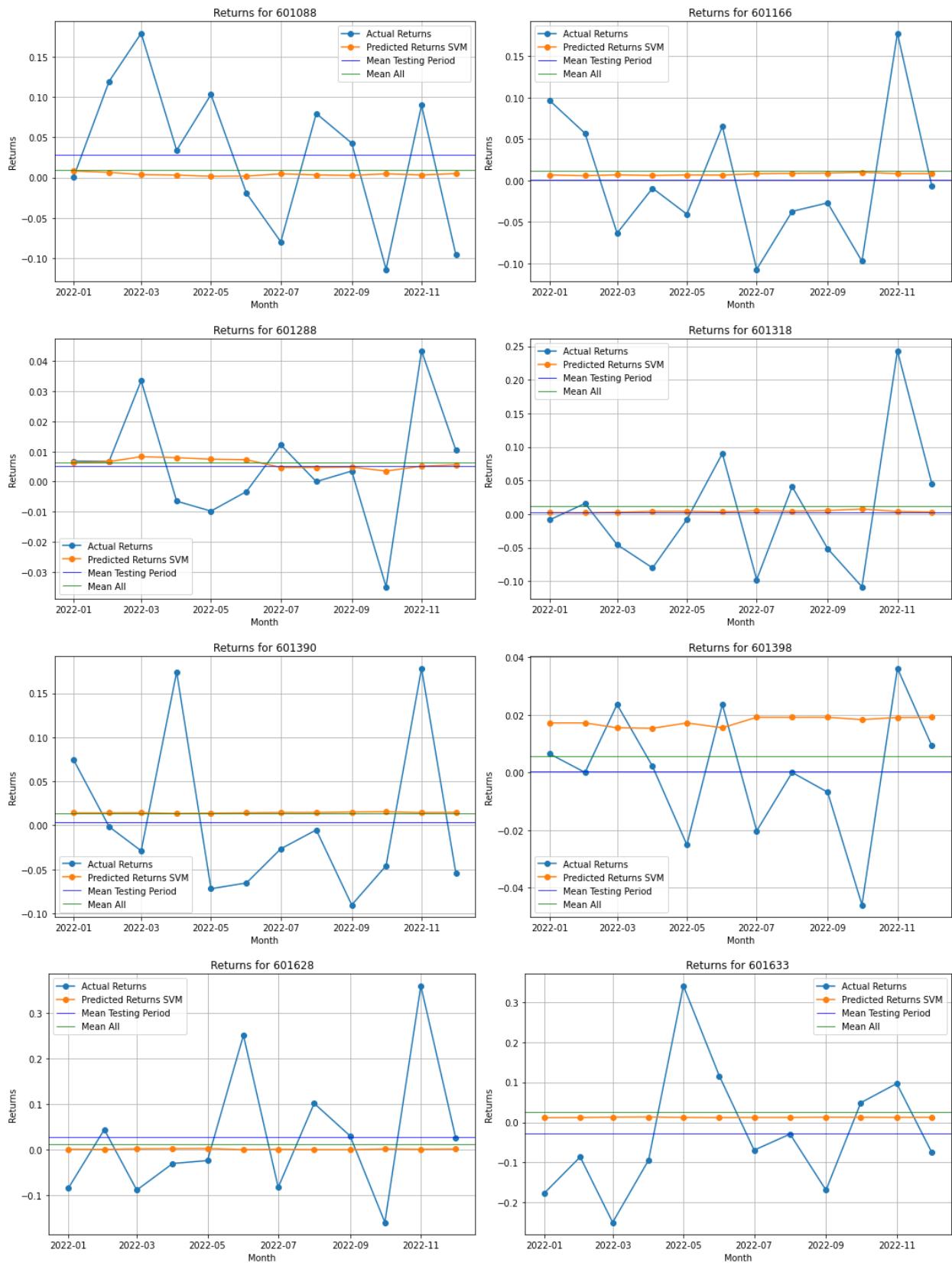


SVM







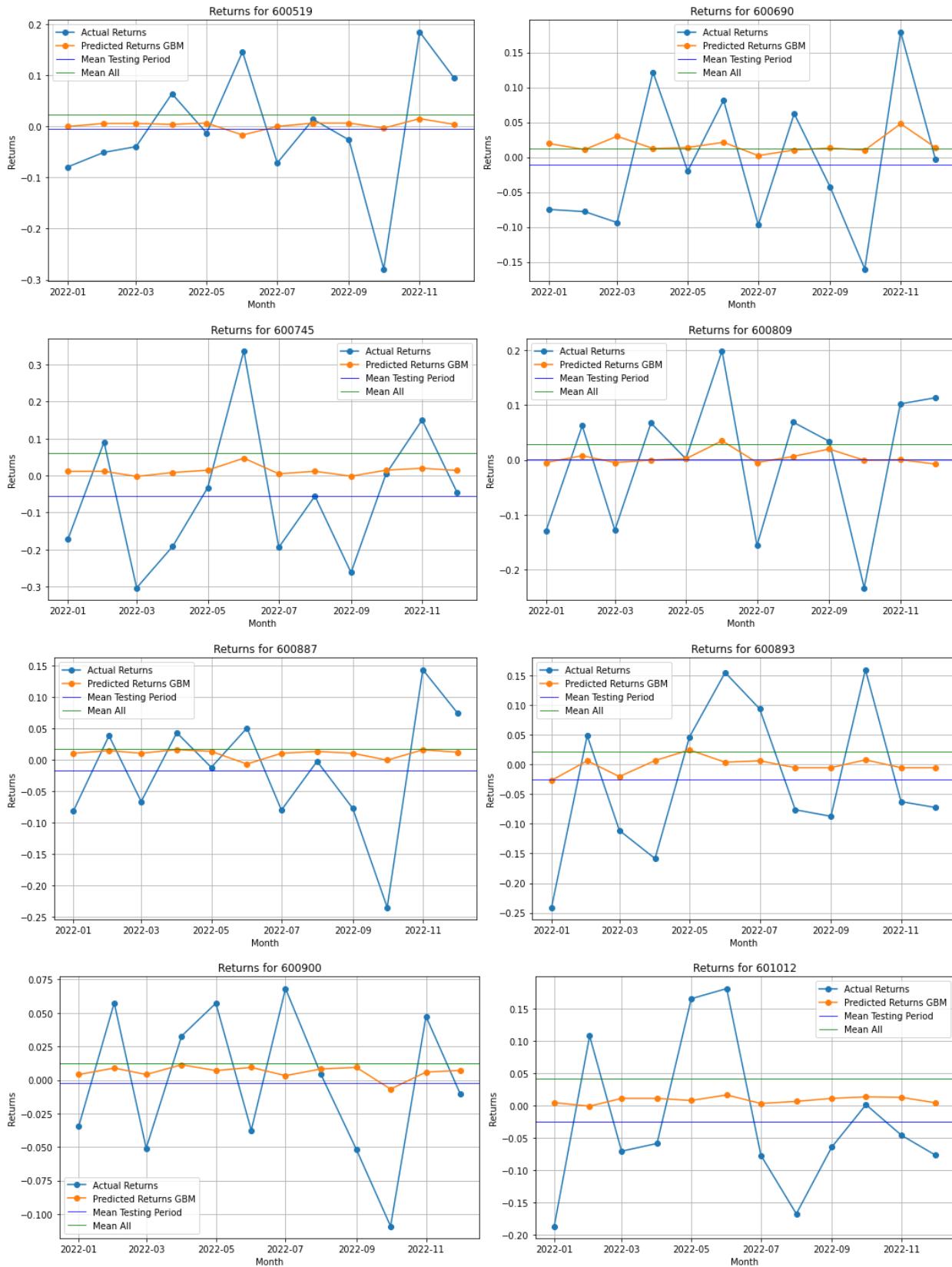




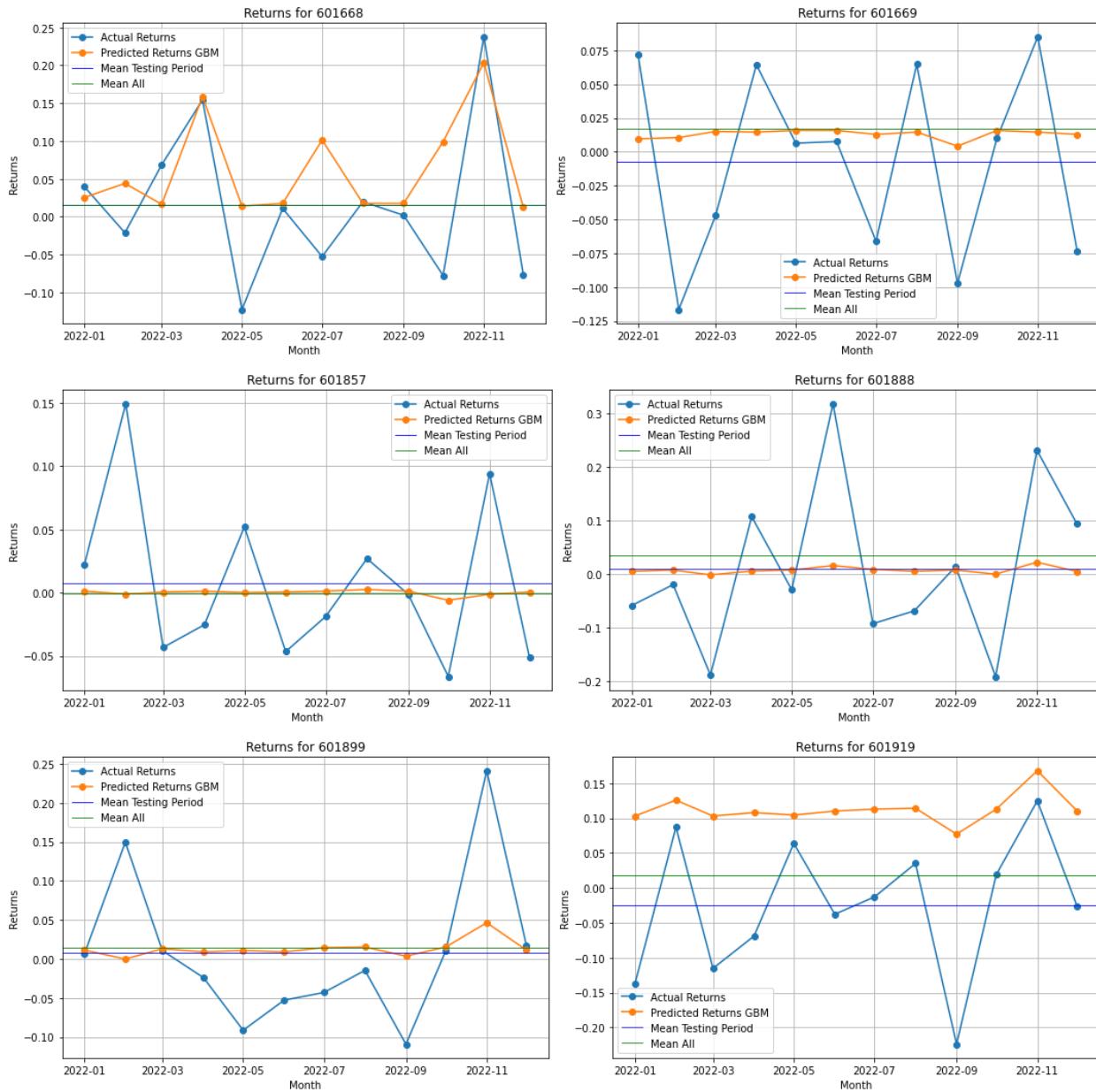
GBM









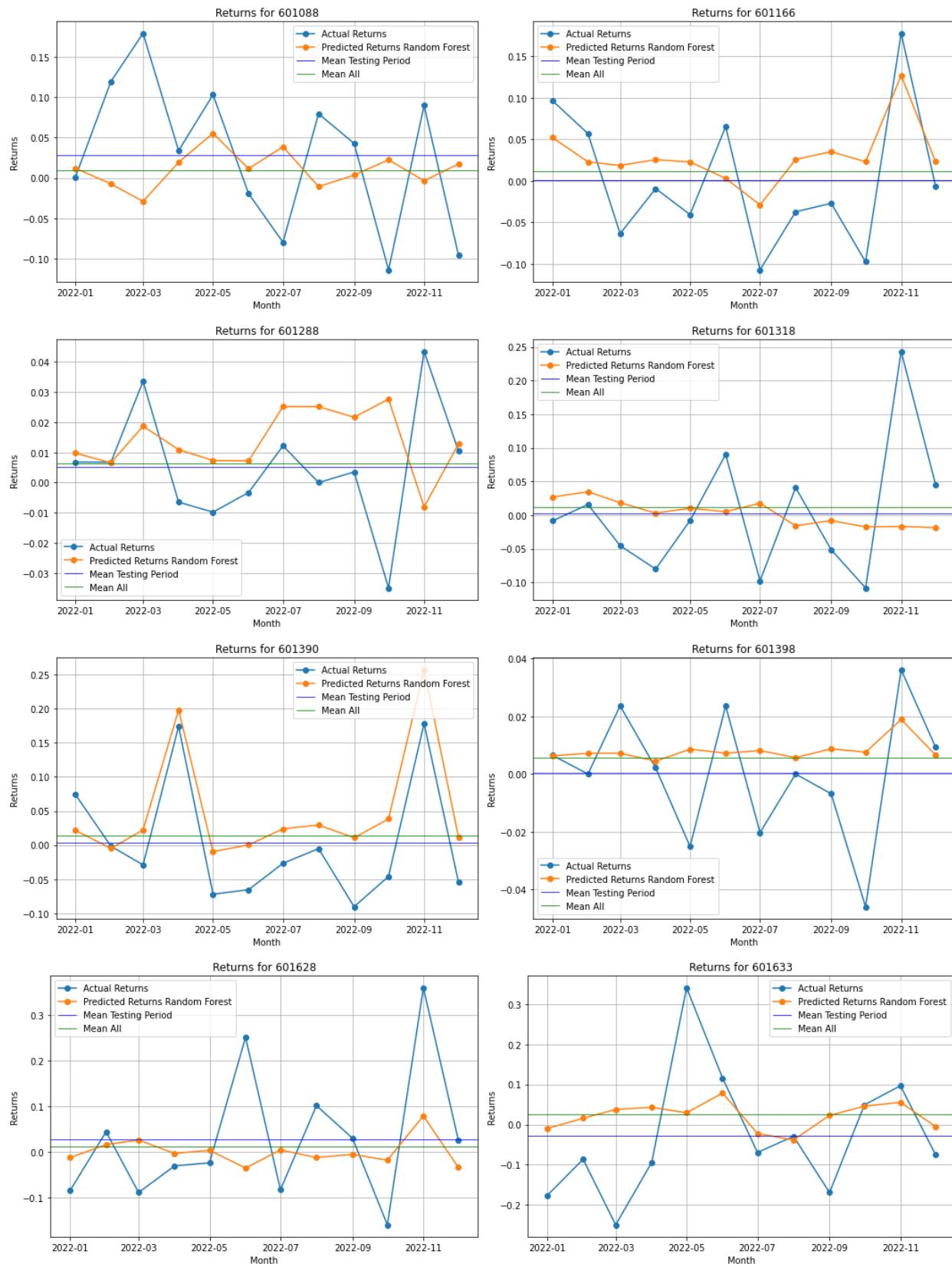


Random Forest











Linear Regression









