Stock market forecasting: an extension of the Engle-Granger approach to machine learning and ensemble learning techniques.

Alberto Sartini

Abstract

This thesis investigates the use of econometric and machine learning methods for forecasting short-term stock market returns, with a focus on the S&P 500 index.

The objective is to assess whether the linear dynamics underlying the Engle-Granger approach can effectively forecast one-period-ahead S&P 500 returns, or if predictive accuracy can be improved using nonlinear learning models.

The analysis builds on the work of W.T. Ziemba et al., who identified a long-run equilibrium relationship between the earnings yield and interest rates.

Using this relationship as a foundation, a rolling backtesting cycle is implemented. At each iteration, data up to time *t* are loaded; the Engle-Granger two-step procedure is applied, with the Error Correction Model estimated in the second step. In parallel, several machine learning algorithms and ensemble methods are trained. Their forecasts are stored and evaluated.

Results show that among the machine learning models, Random Forest achieves the lowest RMSE (3.946), slightly outperforming the Error Correction Model. Even simple ensemble models outperform three out of four individual models. The best performance is obtained by the "Combined Forecast (BIC)" ensemble, with an RMSE of 3.912 compared to 3.994 for the Error Correction Model.

In conclusion, integrating the Engle-Granger approach with nonlinear models may enhance short-term stock market forecasting. Further improvements could be achieved using higher-frequency data or more refined hyperparameter tuning.

Keywords: stock market forecasting, Engle-Granger approach, error correction model, machine learning, ensemble learning