Laboratory of Artificial Intelligence and Data Science

Project 2 - Stock Portfolio Forecasting and Optimization on S&P500 Using Machine Learning and Search Methods

2024/2025

1 Introduction

This project will enable students to explore time series forecasting techniques and optimization methods to build an efficient investment strategy based on S&P500 stock data. Using classical machine learning algorithms and deep learning, students can predict stock behavior and apply search methods such as Monte Carlo or genetic algorithms to optimize portfolio selection, seeking to maximize return over a one-month horizon.

2 Context

The stock market is highly volatile and unpredictable, making stock price forecasting and portfolio optimization challenging tasks. Investors seek strategies that can provide risk-adjusted returns efficiently. This project focuses on utilizing machine learning algorithms to forecast future stock prices of the S&P500 index and applying optimization methods to select the best set of stocks for daily investment. It combines data-driven forecasts with optimization techniques to maximize return or minimize risk, creating a solution for real-world financial challenges.

2.1 Related Works

Several studies have applied machine learning techniques to stock market forecasting. Algorithms such as Random Forest, SVM, and Neural Networks have been widely used for stock price predictions. For example, Patel et al. (2015) [PSTK15] combined Random Forest and SVM to predict stock market movements with high accuracy using ensemble models and multi-step prediction approaches. Furthermore, Bao et al. (2017) [BYR17] applied LSTM networks to model financial time series, demonstrating the effectiveness of these networks in capturing long-term temporal dependencies in volatile financial data.

Portfolio optimization using Monte Carlo simulations and genetic algorithms has also been extensively explored. For example, Glasserman (2003) [Gla03] discusses the application of Monte Carlo methods to simulate future asset prices and optimize portfolios in financial

engineering. Lin and Gen (2007) [LG08] applied genetic algorithms to solve complex multi-stage combinatorial optimization problems, including portfolio optimization, demonstrating the advantages of evolutionary strategies in achieving superior risk-adjusted returns.

These methods provide a robust foundation for developing financial prediction and optimization solutions, combining historical data forecasts with optimization techniques to maximize risk-adjusted returns.

3 Dataset: S&P500

In this project, the students will work with historical stock data from the S&P500 index, which can be obtained from sources such as Yahoo Finance or Alpha Vantage. The dataset will include daily information such as the opening price, the closing price, the trading volume, and technical indicators like moving averages and volatility.

Students are required to use data from 2010 to 2023 to train their machine learning models and simulate investment operations for January 2024. This setup ensures that the models are trained in a sufficiently long historical period to capture various market conditions while evaluating their performance in a recent, unseen year.

A sample time series (Figure 1) of the asset GOOG (Alphabet) is shown below to provide a visual reference for the data.

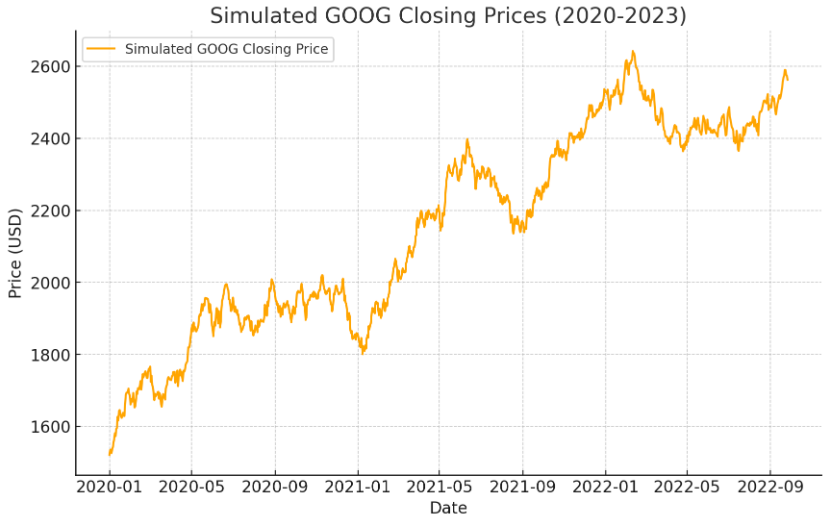

Figure 1: Simulated GOOG Time series

4 Work to develop

- Students will develop a machine learning-based solution to predict stock prices using classical algorithms (e.g. Random Forest, Decision Trees, SVM) or deep learning (e.g. LSTM, MLP).
- The solution must also include a portfolio optimization step using search methods such

as Monte Carlo simulations, MinMax, genetic algorithms, or other approaches to select the best stocks for daily investment.

- Results should be evaluated using financial metrics such as the cumulative return.

4.1 Submission of the solution

- Final code solution, as a notebook;
  - you should document your notebook, explaining your decisions and discussion about the results obtained;
- Link to a video summary. This is a team video, but each member should participate in it. This is a very short and to-the-point video (maximum of 5 minutes), summarizing the following:
  - the problem;
  - your solution;
  - the results and the impact you think this has.
- One-page document, including possible ethical and legal implications and the framework for current and future regulation issues.
- Auto-evaluation file provided by professors.

4.2 Guidelines for the solution

- Assess data quality and perform cleaning if necessary.
- Execute data pre-processing, including normalization and feature creation based on time windows.
- Perform Exploratory Data Analysis (EDA).
- Implement machine learning models for stock price prediction and evaluate their performance using metrics like RMSE, MAE, and MAPE.
- Apply optimization techniques to select the best set of stocks, using methods like Monte Carlo or genetic algorithms.
- Visualize and interpret the results, ensuring that the model is interpretable and justified.
- Explore financial metrics to assess portfolio performance.
- Consider including additional datasets to improve model generalization.

4.3 Evaluation Criteria

Your work will be evaluated on the following criteria:

- 15% Product: Does the proposed solution meet the needs of the prediction and portfolio optimization problem?
- 20% Business: Applicability and impact of the solution in the financial context.
- 40% Technical skills: Overall technical evaluation of the solution, including model quality and optimization techniques.
- 15% Soft-Skills: Communication skills, especially during the presentation of results.
- 10% Ethical and Legal Considerations: Understanding of the ethical and legal aspects involved in financial optimization projects.

4.4 Some Tips

Be creative in your solution! Think, for example, of how you can use certain approaches in an unusual way.

- Consider business constraints: understand the challenge well and identify any business constraints regarding this challenge;
- Mention the constraints you are considering for the solution in the notebook;
- Work as a team: The time is very short; our suggestion is that you distribute tasks well amongst the team;

References

[BYR17] Wei Bao, Jun Yue, and Yulei Rao. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PloS one*, 12(7):e0180944, 2017.

[Gla03] Paul Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer New York, 2003.

[LG08] Chi-Ming Lin and Mitsuo Gen. Multi-criteria human resource allocation for solving multistage combinatorial optimization problems using multiobjective hybrid genetic algorithm. *Expert Systems with Applications*, 34(4):2480–2490, 2008.

[PSTK15] Jigar Patel, Sahil Shah, Priyank Thakkar, and Ketan Kotecha. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert systems with applications*, 42(1):259–268, 2015.