Individual Plan

Mahad Ahmed
mahadah@kth.se

Ismail Mohammed
ismmoh@kth.se

April 2024

Examiner: Prof. Håkan Olsson
hakano@kth.se
Supervisor: Zhenyu Li
zhenyuli@kth.se

1 Background & Objective

This project lies at the intersection of financial technology and machine learning, focusing on the prediction of the Standard and Poor's 500(S&P 500) index [2]. The S&P is a stock market index that measures the stock performance of 500 large companies listed on stock exchanges in the United States. The index includes companies from all sectors of the economy, making it a significant indicator of the performance and condition of the United States economy.

The field of financial technology is increasingly applying machine learning (ML) models to predict stock market trends [5], moving beyond traditional forecasting methods. ML models excel at identifying complex patterns within large datasets and are more adaptable, automatically adjusting to new data that influences the financial markets in a better way. Unlike traditional methods that often require extensive manual intervention and are constrained by the analysts' expertise, ML models can dynamically learn and adapt, enhancing predictive performance.

The predictive analysis of the S&P 500 using machine learning offers tangible benefits, aiding investors and analysts in making data-driven decisions to optimize returns and reduce risk. This capability is especially valuable for financial strategists aiming to enhance portfolio performance and economic policymakers seeking insights into market behaviors [3].

1.1 Project Objective

The primary goal of our thesis, is focusing on the use of machine learning (ML) techniques for stock market prediction. Our aim is to critically assess and compare the efficacy of various ML models, such as Long Short-Term Memory(LSTM), decision trees [1] and polynomial regression [4], in how the accurately forecast the fluctuations within the S&P 500.

Anticipated outcomes include demonstrating the potential superiority of complex models like LSTM in achieving greater predictive accuracy and detailing the relative strengths and weaknesses of each model tested. This work endeavors to contribute meaningful knowledge to the fields of financial technology and machine learning, providing a solid basis for future innovations in stock market analysis.

1.2 Background Knowledge

To successfully carry out the thesis project on index prediction using machine learning models, a foundational knowledge in statistics, programming, and finance is essential for understanding the terminologies and concepts involved. Additionally, sufficient time will be allocated to mastering specific tools and technologies, including various Python libraries, TensorFlow, and Keras, which are pivotal for the project's execution.

Our academic endeavors at KTH have equipped us with a robust foundation and the necessary background knowledge to undertake this thesis within the given timeframe. This educational journey has not only honed our skills in the relevant disciplines but also prepared us to navigate and learn advanced tools and methodologies required for this research.

2 Research Question & Method

2.1 Research Question

We aim to compare the predictive performance of three machine learning models on short-term S&P 500 movements: A) Polynomial Regression, B) Decision Trees, and C) LSTM networks.

- **Polynomial Regression(PR)** is a form of regression analysis that models the relationship between dependent variable and one or more independent variables as (nth) degree polynomial. PR's ability to model non-linear patterns makes it a strong candidate for capturing the often non-linear and complex movements of financial markets. Compared to LSTMs, PR can be less computationally intensive and provide quicker insights, which is beneficial for analyses where speed is critical.

- **Decision Trees (DT)** is flowchart-like structure that use branching methods to illustrate every possible outcome of a decision. In the context of stock market data, they can help identify crucial decision points that affect stock prices. Compared to LSTMs, DT are typically faster to train and can handle effectively without the need for extensive pre processing.

- **LSTM** are designed to recognize patterns in sequence data, beneficial for financial markets where past trends can influence future ones. They have internal mechanisms to remember past information, which can be crucial for time-series data like stock prices.

In a comparative study, while LSTMs may excel in time-series data were context and order are significant, PR and DT may offer compelling advantages in scenarios where quick, clear and where the relationship between variables is highly non-linear and complex.

2.2 Hypothesis

We assume the financial market, particularly the S&P 500 index, exhibits a degree of stability and predictability over short-term intervals. This stability is predicated on the assumption that certain key factors influencing market movements, such as economic indicators and market sentiment, remain constant or change within predictable bounds over the time frame of our analysis.

2.3 Objectives

Evaluate the predictive accuracy of various machine learning models (LSTM, Decision Tree, Polynomial Regression) in forecasting short-term movements of the S&P 500 index. Identify the strengths and weaknesses of each model in the context of financial market predictions. Determine the most effective model for accurate and reliable short-term market movement predictions.

2.4 Tasks

- **Data Collection:**
  1. Identify and list potential data sources for the S&P 500 historical prices, including official stock exchange archives and financial data APIs.
  2. Download daily S&P 500 index data for the designated time period.
  3. Perform initial data cleaning to remove any missing or incorrect entries. Quantifiable metric: less than 0.5% missing or incorrect data after cleaning.
  4. Validate the quality and completeness of the data by comparing with multiple sources. Outcome: Consistency across at least 95% of the dataset.

- **Model Implementation:**
  1. For each considered model, identify key parameters and features that are hypothesized to impact prediction accuracy. These parameters might include the learning rate, the number of layers (for neural networks), the depth of trees (for decision trees) etc. Document the rationale for selection.
  2. Implement the initial version of each model using Python and relevant libraries (e.g., scikit-learn for Polynomial Regression and Decision Trees, TensorFlow/Keras for LSTM).
  3. Train each model on a subset of the collected data, applying cross-validation to optimize parameters. Measure: Improvement in validation accuracy with parameter tuning.

- **Model Evaluation:**
    1. Define evaluation criteria based on the project's objectives: Mean Absolute Error(MAE), Mean Squared Error(MSE) and Root Mean Squared Error(RMSE) for accuracy, and additional metrics if necessary for overfitting assessment.
    2. Analyze the models' sensitivity to market volatility by segmenting the test data based on volatility periods and reassessing performance.
    3. Identify the model with the best trade-off between accuracy and efficiency.

2.5 Method

A combination of statistical analysis and machine learning modeling will be employed, leveraging Python libraries such as TensorFlow and Keras for implementation. These methods are appropriate due to their proven effectiveness in handling time-series data and predicting financial market trends.

2.6 Limitations

The study is limited to predicting short-term movements of the S&P 500 index using publicly available data. It will not consider private data or detailed individual stock movements within the index.

2.7 Risks

Risks include potential data inaccuracies, model overfitting, and unforeseen changes in market dynamics. Mitigation strategies include rigorous data validation, regular model evaluation, and adaptive model updating to reflect new market conditions.

3 Evaluation

The prediction will be done for each considered ML model under different market volatility cases. The MAE, MSE, and RMSE will be checked for the purposes of evaluating the performance of the ML models. The computational efficiency such as training time and computational complexity will also be evaluated. Based on the results we obtained, the strengths and weaknesses of each ML model are expected to be drawn.

4 Pre-Study

The aim of the literature study is to get a solid foundation and understanding of ML application in the financial market prediction.

4.1 Focus Areas

The literature review will concentrate on ML in finance, temporal data analysis, and S&P 500 predictive modeling.

4.2 Obtaining Necessary Knowledge

To gather the required background and state-of-the-art knowledge, we will Utilize academic databases such as Google Scholar, JSTOR, and IEEE Xplore for peer-reviewed articles and journals. Consult financial databases and archives for real-world data and analysis on the S&P 500 index. Engage with online courses and recent publications to understand the latest developments in machine learning applications in finance. This could be YouTube or other resources.

4.3 Preliminarily Important References

1. Achmad Fauzan et al. *"Predicting Stock Market Movements Using Long Short-Term Memory (LSTM)"*[2].
2. Ibtissam Medarhri et al. *"Predicting Stock Market Price Movement using Machine Learning Techniques"* [5].

5 Project timeline

Every step of the way we are both contributing to everything. No fixed tasks.

- **Week 1: Project Planning:** Finalize thesis topic and define research question. Conduct initial research to understand the scope and relevance of the project. Set up meetings with supervisor to discuss the project plan and get feedback. Begin writing the individual plan
- **Week 2: Literature Study:** Conduct a comprehensive literature review to understand previous work in the field of financial market predictions using machine learning. Identify gaps in the literature and potential methodologies used in similar studies. Begin compiling a bibliography of sources to support the research and methodology. Finish writing the individual plan
- **Week 3: Data Collection and Preprocessing:** Identify and collect historical daily data on the S&P 500. Clean the data to handle missing values, anomalies, and normalize features. Begin writing the thesis intro, and background.
- **Week 4: Model Development (Initial Models):** Implement initial models such as Polynomial Regression and decision trees. Conduct initial training sessions and evaluate performance. Begin drafting sections of the thesis related to methodology.
- **Week 5-6: Model Development (Complex Models):** Implement and train more complex models, specifically LSTM. Adjust models based on validation performance. Continue drafting the thesis, focusing on initial model findings.
- **Week 7: Model Evaluation and Comparison:** Finalize model training and conduct comprehensive evaluations. Compare model performances to identify the most effective approach. Continue writing the thesis, focusing on the results of all models.
- **Week 8: Finalization and Submission Preparation:** Analyze results and draw conclusions. Complete the writing of the thesis, including discussions and conclusions. Proofread, format, and prepare for submission and presentation.

REFERENCES

[1] Hendrik Blockeel et al. "Decision trees: from efficient prediction to responsible AI". In: *Frontiers in Artificial Intelligence* 6 (July 26, 2023). Publisher: Frontiers. ISSN: 2624-8212. DOI: 10.3389/frai.2023.1124553. URL: https://www.frontiersin.org/articles/10.3389/frai.2023.1124553 (visited on 04/04/2024) (cit. on p. 1).

[2] Achmad Fauzan et al. "Predicting Stock Market Movements Using Long Short-Term Memory (LSTM)". In: *2023 4th International Conference on Artificial Intelligence and Data Sciences (AiDAS)*. 2023 4th International Conference on Artificial Intelligence and Data Sciences (AiDAS). Sept. 2023, pp. 144–147. DOI: 10.1109/AiDAS60501.2023.10284713. URL: https://ieeexplore.ieee.org/document/10284713 (visited on 03/26/2024) (cit. on pp. 1, 3).

[3] Shihao Gu, Bryan Kelly, and Dacheng Xiu. "Empirical Asset Pricing via Machine Learning". In: *The Review of Financial Studies* 33.5 (May 1, 2020), pp. 2223–2273. ISSN: 0893-9454. DOI: 10.1093/rfs/hhaa009. URL: https://doi.org/10.1093/rfs/hhaa009 (visited on 04/04/2024) (cit. on p. 1).

[4] Dastan Maulud and Adnan M. Abdulazeez. "A Review on Linear Regression Comprehensive in Machine Learning". In: *Journal of Applied Science and Technology Trends* 1.2 (Dec. 31, 2020). Number: 2, pp. 140–147. ISSN: 2708-0757. DOI: 10.38094/jastt1457. URL: https://jastt.org/index.php/jasttpath/article/view/57 (visited on 04/04/2024) (cit. on p. 1).

[5] Ibtissam Medarhri et al. "Predicting Stock Market Price Movement using Machine Learning Techniques". In: *2022 8th International Conference on Optimization and Applications (ICOA)*. 2022 8th International Conference on Optimization and Applications (ICOA). ISSN: 2768-6388. Oct. 2022, pp. 1–5. DOI: 10.1109/ICOA55659.2022.9934252. URL: https://ieeexplore.ieee.org/document/9934252 (visited on 03/26/2024) (cit. on pp. 1, 3).