Capstone Project Plan

Ri-on Kim, Manuelito Bagasina, Youngjun Jhee

February 22, 2025

1 Brief Background

Forecasting stock prices in the past has been particularly challenging with the vast array of financial, economic, and psychological forces that need consideration. The Efficient Market Hypothesis (EMH) and Random Walk Theory are among the conventional financial theories suggesting stock market movement is random in nature, and prediction is made difficult. Recent innovations in Machine Learning (ML) approaches and increased access to vast databases have enabled researchers to design more advanced models combining stock prices over a period with macroeconomic (GDP and interest rate), microeconomic (P/E and EPS), and media and Twitter-derived sentiment analyses. These newer models are trying to identify finer forces in the stock market not captured in conventional approaches.

Despite the progress made, a number of problems still remain, such as problems with information quality, the need for near-instantaneous signals in economies, uncertainties in the process of sentiment analysis, and complexity in combining multiple information streams. Subsequent research efforts will need research on alternate and better ways of obtaining and pre-processing information, better feature selection methods, and models with increased ability and fault tolerance in order to produce better predictive results. In addition, the ongoing effort is aimed at developing a completely automated pipeline suitable for production. The primary intent is to deploy the model via a simple web interface, and in doing so, enable the availability and usefulness of predictions in a production setup.

2 Motivation

Current models often ignore real-time sentiment and industry-specific dynamics, limiting their accuracy. An integrated approach that consolidates multiple sources of data has the potential to improve predictive powers for investors. In addition, automating the entire workflow through an end-to-end data pipeline will enable real-time predictions, thus making the model more relevant in real-world applications. The project also aims to create a simple web platform that enables users to interact with the model and see live predictions, thus demonstrating its real-world application and deployment potential.

3 Research Problem

How can stock prices be effectively predicted using historical stock data, economic indicators, and sentiment analysis?

4 Sub-Questions

4.1 How do different types of data contribute to stock price prediction accuracy?

Knowledge of the impact of different sources of data, such as historical share prices, macroeconomic data, and sentiment analysis, can be used to tune predictive models. Our hypothesis is that macroeconomic variables and sentiment analysis will be significant contributors, and historical share data by itself will have little predictive capability.

4.2 Which tools and approaches in sentiment and natural language processing (NLP) are the most accurate in representing market mood in forecasting stock prices?

This study explores the comparative performance of the diverse methodologies and approaches utilized in sentiment analysis and natural language processing (NLP), such as lexicon models, machine learning methods (e.g., Naïve Bayes and Support Vector Machines), and deep models (e.g., BERT and FinBERT). The major hypothesis is that transformer models, particularly FinBERT, coupled with advanced NLP methods such as word embedding and attention, are going to outperform conventional approaches in sentiment analysis because they are better positioned to identify complex linguistic features and situational nuances in financial text.

4.3 Which features (stock data, macroeconomic, macroeconomic, sentiment) contributed the most to predicting stock price movement?

The features will be those that help predict the movements of the stock price and their use will help refine the accuracy of the models. Historical stock data, including price and volume, provides the foundation for time-series analysis, with technical indicators derived from this data often being key predictors. Sentiment scores from social media and financial news articles are significant predictors, reflecting investor sentiment, which can drive market movements. Identifying the most influential features allows researchers to focus on collecting and refining the most relevant data, thereby improving the accuracy of stock market predictions.

The hypothesis is this: Sentiment scores (from NLP analysis of news/social networks) and technical indicators (e.g. moving averages, RSI) will dominate feature importance. Microeconomic factors (EPS, P/E ratio) will lag behind due to quarterly reporting delays.

4.4 How do different machine learning models compare in terms of accuracy and computational complexity in forecasting stock prices?

The present research explores the trade-offs involved with varying machine learning models, e.g., Random Forest, XGBoost, LSTM, and Transformer, with reference to predictive performance and associated computational costs. The hypothesis is that while complex deep models such as LSTM and Transformer could offer better predictive performance, basic models such as Random Forest and XGBoost could exhibit better computational efficiency and are better suited for faster prediction.

5 Datasets

5.1 Stock Price Indicators

We will fetch past stock prices from financial market APIs like Alpha Vantage or Yahoo Finance. The data will consist of daily open, high, low, and close (OHLC) prices, trading volume, and computed technical indicators like moving averages and relative strength index (RSI).

5.2 Macroeconomic Indicators

Macroeconomic data will be obtained from the Federal Reserve Economic Data (FRED) and other publicly accessible financial databases. The dataset will cover major indicators such as GDP growth rates, unemployment rates, interest rates, inflation, and consumer confidence indices. These indicators set the wider background for the stock market move.

5.3 Microeconomic Indicators

Microeconomic data will be sourced from company financial reports and filings, as well as through the `yfinance` API, which provides access to a wide range of financial data. The dataset will include:

- Earnings Per Share (EPS): Measures profitability distributed per share.
- Price-to-Earnings (P/E) Ratio: Indicates if a stock is over- or under-valued.
- Debt-to-Equity Ratio: Reflects financial leverage and associated risk.
- Return on Assets (ROA): Evaluates asset efficiency in profit generation.
- Return on Equity (ROE): Measures profit generated from shareholders' equity.

In addition to the above-mentioned indicators, over 25 other financial features are going to be obtained with the `yfinance` API. The Recursive Feature Elimination (RFE), SHAP values, and Principal Component Analysis (PCA) methods are going to be utilized in order to discover the features that are having a substantial impact on increasing the predictive power of the models. The method ensures the model puts a greater weight on the most impactful variables, thus increasing the accuracy of the predictions.

5.4 Sentiment Analysis Data

5.4.1 Reddit

Sentiment analysis data will be gathered from various Reddit communities including r/worldnews. The posts and comments will be collected using Reddit's API, focusing on:

- Post and comment sentiment (positive, negative, neutral) using NLP models.
- Volume of discussion related to specific stocks.
- Time-series analysis of sentiment trends leading up to stock price changes.

The objective here is capturing investor sentiment in real-time, something that is capable of influencing short-run stock price volatility.

5.4.2 Stock News

The Stock News API served as the primary data source for web-scraping financial news articles. Data collection targeted US-based technology companies over a five-year historical window, with extracted information organized into datasets per company. The news articles varied in source, some coming from Reuters, Investopedia, Business Insider, and many more. Each data set comprises five key elements: publication timestamp, article headline, full textual content, source URL, and a pre-classified sentiment label (positive, negative, neutral). This approach allows for systematic analysis of temporal news patterns and sentiment trends in the technology sector.

This comprehensive EDA will not only reveal essential insights about the dataset but also assist in identifying which features should be prioritized during model development, contributing to a more accurate and robust prediction framework.

6 Methodology

The methodological process involves a systematic and integrative pipeline with efficient data preprocessing, model development, and deployment, all with the purpose of real-time accurate prediction of stock prices in real-time. The systematic process incorporates data engineering, machine learning, sentiment analysis, and deployment, leading to a working product.

- **Data Collection:** Gather stock price data, macroeconomic indicators, microeconomic indicators, and sentiment data from multiple sources:

  - **Financial Data:** Extract historical stock prices and financial ratios using the `yfinance` API.
  - **Macroeconomic Data:** Collect macroeconomic indicators from publicly available databases, such as the Federal Reserve Economic Data (FRED).
  - **Sentiment Data:** Scrape data from Reddit and news articles using APIs and web scraping tools.

- **Data Preprocessing:** Clean and transform the dataset to prepare it for analysis:
  - Handling missing values using imputation techniques.
  - Removing or adjusting outliers using Z-score or IQR methods.
  - Normalizing or standardizing data to ensure consistent feature scaling.
  - Synchronizing timeframes across different data sources for time-series alignment.
  - **Sentiment Analysis Preprocessing:** Perform natural language preprocessing on text data:
    - Tokenization, stop-word removal, and stemming/lemmatization.
    - Sentiment score extraction using NLP models such as VADER, FinBERT, or custom-trained models.
    - Cleaning noise from text data, including removing irrelevant links, emojis, and special characters.
    - Encoding Sentiment-based features.

- **Exploratory Data Analysis (EDA):** Conduct a comprehensive analysis to better understand the data:
  - **Time-series visualization:** Visualize trends in stock prices alongside macroeconomic shifts to understand how economic indicators influence stock market movements.
  - **Correlation analysis:** Use heatmaps to identify relationships between sentiment scores, technical indicators such as the relative strength index (RSI), and stock price movements. This will help to understand how different factors correlate with each other and with stock prices.
  - **Sentiment Trend Analysis:** Analyze differences in sentiment based on Reddit comments and financial media reports, and correlate these differences with stock price fluctuations via the usage of time-series decomposition.
  - **Cluster Analysis:** Perform unsupervised clustering (e.g., K-means) on sentiment scores and stock price patterns to identify distinct market behavior regimes.
  - **Outlier Detection:** Utilize methods like the Interquartile Range (IQR) and Z-score method in order to identify and handle potential outliers whose distortion could compromise predictive models.
  - **Principal Component Analysis (PCA):** Perform PCA in order to get dimensionality reduction and portray how multiple features affect the total variance within the dataset.

  This comprehensive EDA will not only reveal essential insights about the dataset but also assist in identifying which features should be prioritized during model development, contributing to a more accurate and robust prediction framework.

- **Feature Engineering:** Develop new features that enhance the model's predictive performance:
  - Generate technical indicators (e.g., moving averages, RSI, MACD).
  - Create lagged features for time-series forecasting.

- **Feature Selection:** Identify the most impactful features through:
  - **Recursive Feature Elimination (RFE):** Iteratively eliminate less important features.
  - **SHAP Values:** Interpret each feature's contribution to model outputs.
  - **Principal Component Analysis (PCA):** Reduce dimensionality while preserving variance.
- **Model Development:** Perform multiple predictive model training and evaluation:
  - **Baseline Models:** ARIMA for time-series forecasting and Linear Regression as a benchmark.
  - **Machine Learning Models:** Random Forest and XGBoost for structured data.
  - **Deep Learning Models:** LSTM networks and Transformer-based models (e.g., FinBERT) for integrating time-series and sentiment data.
  - **Hybrid Models:** Combine LSTM for sequential data patterns and XGBoost for handling structured numerical features.
- **Model Evaluation:** Assess model performance using:
  - The Mean Absolute Percentage Error (MAPE), and the Root Mean Squared Error (RMSE) would be precision indicators.
  - Directional Accuracy to evaluate the ability to predict market trends.
  - Cross-validation techniques to ensure model robustness.
- **Automated Data Pipeline:** Create a fully automated pipeline where integration is seamless
  - Automated data ingestion from APIs and web scrapers.
  - Real-time data preprocessing and feature updating.
  - Periodic model retraining based on new data.
- **Deployment and Production:** Deploy the model into a practical environment:
  - Create a platform with on-demand services with instant access for users.
  - Develop a dashboard for visualizing predictions, feature importance, and sentiment trends.
  - Build a dashboard for the visualization of predictions, feature importance, and sentiment trends.
- **Documentation and Reporting:** Provide clear and comprehensive documentation:
  - Justify model selection, highlight limitations, and explain evaluation results.
  - Present visualizations for insights derived from EDA and predictions.
  - Suggest future research pathways and possibilities for model enrichment.