Project Plan

Project Title:

Predictive Modeling and Analysis of Stock Price Movements Using NYSE Historical Data

Research Question:

How stock price movements be predicted and trading strategies be developed using previous stock price data and financial metrics?

Project Objectives:

1. Perform EDA to understand the composition of previous stock price data as well as basic financial measurements.
2. Preprocess the data by handling outliers, and missing values.
3. Features engineering includes moving averages, volatility metrics and fundamental ratios.
4. Use the forecasting AI model and compare their results.
5. Make and evaluate trading strategies such as mean-reversion and momentum strategies based on projections.

Summary of Project and Background:

Our objective is to forecast stock prices and develop trading techniques using the NYSE dataset. The data includes daily stock prices, prices for splits of stocks, firm descriptions, and financial information extracted from SEC 10K filings. In light of algorithmic trading, we aim to investigate ML algorithms that can be used to automate stock trading.

We will preprocess the data, design features, and use models on the dataset. The accuracy of their prediction and their capacity to profitable trading methods will be the criteria used to test their performance. The goal is to provide insight into the performance of different trading techniques and models within the framework of the New York Stock Exchange by the project's conclusion.

References:

1. Nabipour, M., Nayyeri, P., Jabani, H., Mosavi, A., Salwana, E. and S, S., 2020. Deep learning for stock market prediction. *Entropy*, *22*(8), p.840.
2. Pawar, K., Jalem, R.S. and Tiwari, V., 2019. Stock market price prediction using LSTM RNN. In Emerging Trends in Expert Applications and Security: Proceedings of ICETEAS 2018 (pp. 493-503). Springer Singapore.
3. Zheng, A. and Jin, J., 2017. Using ai to make predictions on stock market. *cs229. stanford. edu*.

Task List and Project Timeline

Task List

Literature Review (June 9 - June 21)

- Examine pertinent research on trading tactics, stock price forecasting, and predictive modeling in finance.
- Summarize key findings and identify gaps in existing research.

Data Collection and Exploration (June 9 - June 28)

- Gather datasets from the New York Stock Exchange (NYSE) and perform initial data exploration.

Data Preprocessing (June 28 - July 19)

- Clean and preprocess the data, handling missing values, and outliers, and merging datasets.

Feature Engineering (July 8 - July 19)

- Engineer relevant features from historical stock prices and fundamental financial metrics.

Model Development (June 19 – August 14)

- Implement predictive models including LSTM, RNN, and SVM.

Model Evaluation (August 1 - August 16)

- Evaluate model performance using RMSE, MAE, and accuracy metrics etc.

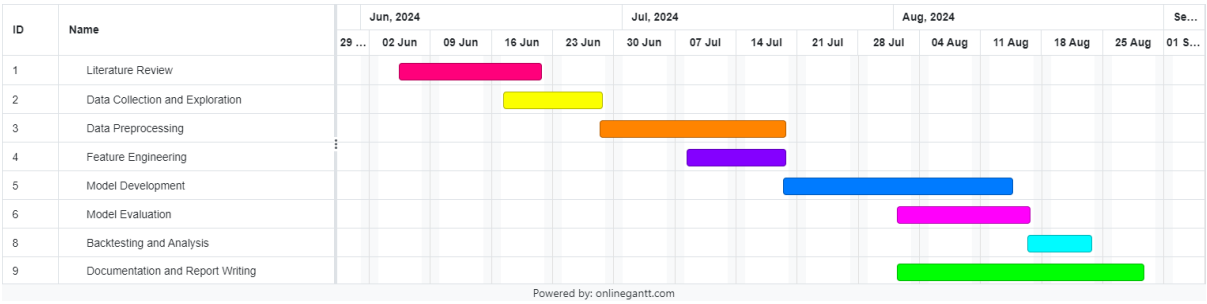
Backtesting and Analysis (July 21 - July 27)

- Backtest trading strategies and analyze their profitability and risk.

Documentation and Report Writing (July 28 - August 10)

- Prepare project documentation and write the final report.

Gantt Chart



Data Management Plan

Overview of the Dataset

The dataset consists of historical stock price data from the New York Stock Exchange (NYSE), including daily prices, adjusted prices for stock splits, company descriptions, and fundamental financial metrics extracted from SEC 10K filings. The data was sourced from Yahoo Finance for stock prices and Nasdaq Financials for fundamental metrics. The original purpose of collecting the data was for financial analysis and research in the domain of stock market forecasting and trading strategies.

Data Collection

The data will be collected from the Kaggle platform, specifically from the dataset repository for the New York Stock Exchange (NYSE). The dataset includes multiple CSV files containing raw and adjusted stock prices, company descriptions, and fundamental financial metrics.

Summary of Data

- The dataset includes four files: fundamentals.csv, prices-split-adjusted.csv, prices.csv, and securities.csv.
- **fundamentals.csv:** Contains financial metrics from SEC 10K filings.
- **prices-split-adjusted.csv:** Provides adjusted prices for stock splits.
- **prices.csv:** Contains raw daily stock prices.
- **securities.csv:** Describes each company with sector divisions.
- The total size of Version 3 is 105.84 MB.

Document Control

GitHub Repository: https://github.com/yourrepository

GitHub will be used for version control to track the development of code and record all commits in the project logbook. File naming conventions will follow a structured approach, with clear indications of file contents and versions.

Metadata

A User Document (ReadMe file) will be provided in the GitHub repository. This ReadMe file will include a brief overview of the project, instructions for accessing and using the dataset, descriptions of the data files, explanations of the code structure, and any dependencies required to run the code.

Security and Storage

Regular backups will be carried out and updates to the GitHub repository will be made following codebase additions or modifications. The data and code backups will be kept on cloud storage services like Dropbox and Google Drive in addition to local storage. Only the project team, supervisor, and student will have access to data assuring adherence to security and ethical guidelines.

Ethical Requirements

1. **GDPR Compliance:** The data does not contain personally identifiable information and is derived from publicly accessible financial sources.
2. **UH Ethical Policies:** The project adheres to the University of Hertfordshire's ethical standards, which ensure integrity, propriety, and confidentiality in study practices.
3. **Permission to utilize Data:** Since the dataset is openly available on Kaggle and intended to be used for research and teaching, we are permitted to utilize it.
4. **Ethical Data Collection:** Yahoo Finance and Nasdaq Financials, the original data providers, adhere to ethical standards when gathering data to guarantee the validity and integrity of the dataset.