Stock Price Prediction using Machine Learning and Deep Learning Models

Project Group 14
Members: Mannan Rangoonia (mr07882) , Mian Abdul Wasay (mw08482)

1 Problem Statement

Stock market investors and analysts rely heavily on accurate forecasts of stock prices to make informed decisions. However, predicting the next-day closing price of a stock is highly challenging due to the volatile and non-linear nature of financial data. Traditional statistical models often fail to capture these complex temporal dependencies and patterns present in historical stock prices.

The problem this project addresses is how to develop a robust and comparative forecasting framework capable of predicting next-day stock closing prices using three different modeling strategies:

- LSTM (Long Short-Term Memory): a deep learning model that captures sequential dependencies.
- XGBoost (Extreme Gradient Boosting): a tree-based ensemble machine learning model known for its performance and interpretability.
- N-BEATS (Neural Basis Expansion Analysis for Time Series): a deep learning model specifically designed for univariate time series forecasting.

The ultimate goal is to compare their performances using the same dataset and preprocessing pipeline to determine which model achieves the highest predictive accuracy and generalization capability.

2 Potential Solution

The proposed solution involves building a five-stage machine learning/deep learning pipeline that automates data extraction, preprocessing, training, evaluation, and prediction.

2.1 Pipeline 1

The first stage of the solution involves collecting daily stock closing price data using freemium APIs (such as alpha vantage). The data is then cleaned and saved in a csv file.

2.2 Pipeline 2

The second stage deals with preparing the data for machine learning and deep learning models. This includes normalizing the data so that it lies between 1 and +1, creating n-day input sequences to predict the next day's closing price, and splitting the dataset into training and testing sets. The prepared data is then converted into a PyTorch compatible format for deep learning models and NumPy arrays for traditional models.

2.3 Pipeline 3

The third stage focuses on model training. The LSTM model is designed to capture long-term dependencies and sequential trends in the data, while the XGBoost model uses gradient boosting to identify nonlinear interactions between variables efficiently. The N-BEATS model, a deep neural architecture tailored for time series forecasting, learns complex temporal patterns by stacking several fully connected layers in a residual block structure. Each model will be trained with optimized parameters, and their training progress will be monitored to ensure convergence.

2.4 Pipeline 4

Once training is complete, the models are evaluated using unseen testing data. The evaluation metric used is Mean Absolute Percentage Error (MAPE), which measures the average deviation between predicted and actual values as a percentage. Accuracy is calculated as 100 minus MAPE, giving a straightforward indicator of performance.

2.5 Pipeline 5

Finally, each model predicts the next day's closing price using the most recent n days of available data.

3 Resources To Be Used

3.1 Libraries And Tools

- Programming Language: Python 3.11+
- Data Extraction: Alpha Vantage Api
- Data Manipulation: numpy, pandas
- Machine Learning Models: xgboost
- Deep Learning Models: torch
- Evaluation Metrics: sklearn.metrics
- Visualization: matplotlib

3.2 Datasets

- Source: Alpha Vantage Api
- Data Type: Daily closing prices of selected stocks
- Data Volume: From the time the data exists in the source database till present day (approx $\geq$ 10 years)