Arun Kumar 21079279

Stock Price Prediction of UK Banks Using Multiple Machine Learning Models and Technical Indicators

The stock market is a complex system affected by economy, geopolitical events, investor sentiments and a lot of other hidden factors Traditional statistical methods do not capture the complex nonlinear relationships among these factors that's where the power of AI comes in use that is able to figure out complex relationships in these patterns and gives more accurate predictions especially when combined with technical indicators. This project aims to predict the stock prices of major UK banks by using multiple machine learning models. These include Long Short-Term Memory (LSTM), Random Forest, XGBoost, Support Vector Machine (SVM)and Timeseries Transformer, combined with various technical indicators such as SMA, EMA, MACD, RSI, and Bollinger Bands. The project focuses on key UK banks such as Barclays, Lloyds, HSBC, Standard Chartered PLC – London and NatWest. These models will be compared based on their prediction accuracy.

The 5-10 years of historical stock data for Barclays, Lloyds, HSBC, Standard Chartered PLC – London and NatWest were available publicly at Yahoo Finance UK. Data will be collected using the yfinance Python library. This library is ideal for downloading large amounts of financial data efficiently and research purposes. Data will be stored in CSV files. Each file will contain columns for Date, Open, High, Low, Close, Adjusted Close, Volume, and Technical Indicators etc. Each bank's data is expected to be between 5-10 MB, with the total dataset size estimated at 25-50 MB.

The data will be subjected to proper cleaning and preprocessing techniques so that redundant or missing data can be tackled before subjected to feature engineering which uses advance techniques like SMA, EMA etc. The preprocessing and feature engineering is vital aspects for more precise and accurate predictions as the Machine Learning models are more likely to perform better on clean data rather than raw disoriented text.

All code and data will be version-controlled via GitHub, with weekly commits to track changes. This ensures that progress is documented and accessible to others. ReadMe file will include an overview of the project, instructions on how to run the code and a description of the dataset and how to use it. Weekly backups will be made to GitHub and OneDrive to ensure data safety. The GitHub repository will be shared with the supervisor and university staff. Access will be managed to ensure data security.

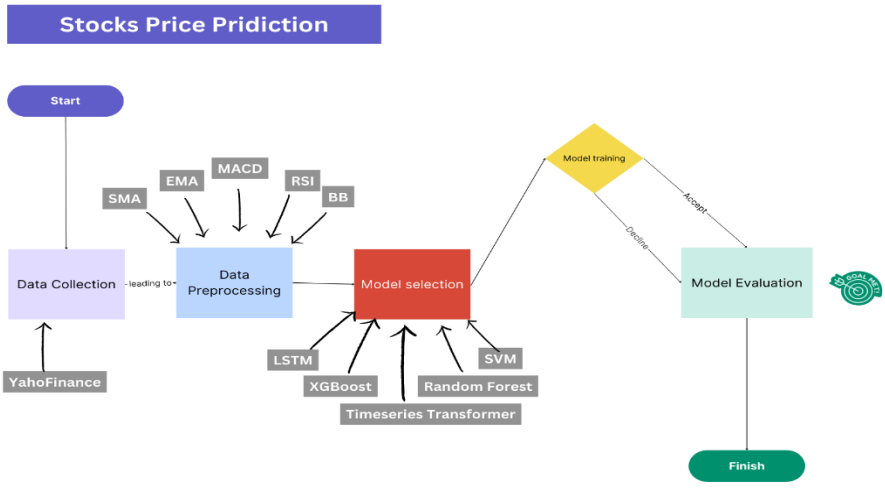

Figure 1 Workflow Diagram

Ethical Considerations

To ensure ethical conduct and protect data privacy, the project will strictly adhere to the following regulations and policies such as GDPR Compliance, University of Hertfordshire Ethical Policies and Permission for Data Use

Technical Indicators and Usage

Technical indicators help detect patterns, trends, and momentum in stock price data. Machine learning models can use these indicators as additional features, which enhances their ability to predict future price movements. Some methods that will be used are Simple Moving Average (SMA), Exponential Moving Average (EMA), Moving Average Convergence Divergence (MACD), Relative Strength Index (RSI) and Bollinger Bands (BB).

We will be using these four basic models with proper experimentations LSTM, Random Forest, XGBoost and SVM. If there is an ample amount of time the fifth model Timeseries Transformer will be used or the model will be replaced with any other model. Below are some ML models and what are our expectations towards their performance.

Long Short-Term Memory (LSTM)

LSTM networks are designed to handle sequential data, such as time series, which is exactly what stock prices represent. LSTMs can capture both short-term and long-term dependencies in data, which makes them ideal for predicting stock price movements over time. LSTM is best model for capturing patterns in stock price trends over a long period (such as weeks, months, or years) because it solves the vanishing gradient problem that arises in traditional neural networks when processing time series data.

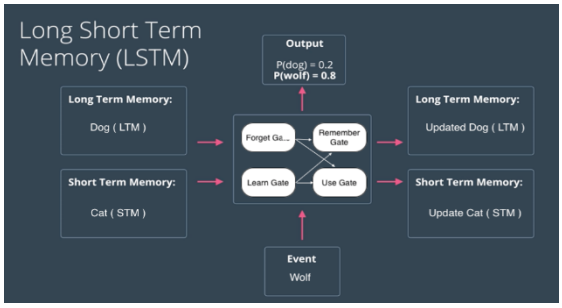

Figure 2 LSTM Architecture

Random Forest

Random Forest is a tree-based ensemble model that excels in regression tasks, such as predicting stock prices. It aggregates multiple decision trees to reduce overfitting and improve accuracy. This model is great for capturing non-linear relationships in stock data and for handling high-dimensional data such as technical indicators combined with stock prices.

XGBoost

XGBoost is a highly efficient and scalable machine learning algorithm, known for its ability to handle large datasets quickly. It is built on the concept of boosting, where each subsequent model corrects the errors of the previous ones. XGBoost is especially effective in capturing complex patterns and interactions in stock price data and is often used in competitions for predictive tasks because of its high accuracy and performance.

Support Vector Machine (SVM)

SVM is a robust model that performs well in both classification and regression tasks. It is particularly effective in cases where the dataset is high-dimensional and contains non-linear patterns. SVM is great for stock prediction where

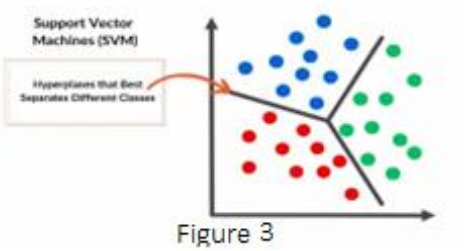

Figure 3

we want to classify price movements (up/down) or forecast trends. It is known for its generalization performance in a variety of tasks.

Timeseries Transformer

Transformers have the self-attention mechanism that makes it unique from other models while the Timeseries Transformers are specifically design for such type of data that we have. If there is an ample time this model can be applied or any other model be replaced with this model to get the stock predictions.

For better performance each model will be subjected to hyperparameter tuning such as changing the learning rate, test train split ratio, number of epochs etc. These experimentations will help to achieve best possible performance. For model evaluation the confusion matrix will be used other methods like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) or Loss Function are also a good option to assess the model performance.

Time line

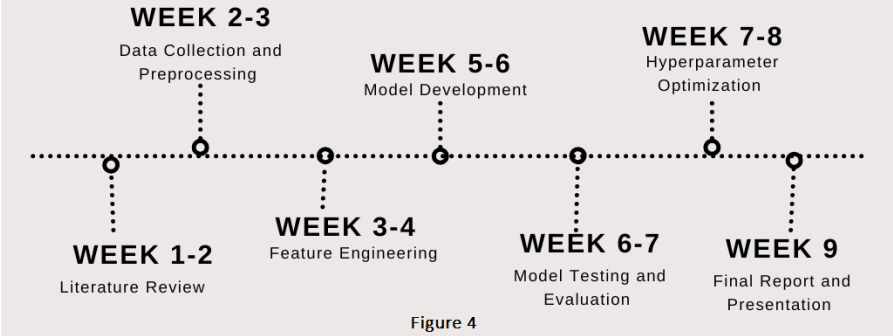

Figure 4

Conclusion

This Project and Data Management Plan details how we will use machine learning models (LSTM, Random Forest, XGBoost, and SVM) in combination with technical indicators to predict stock prices of major UK banks. By using a mix of technical and advanced machine learning models, we aim to find the best approach for accurate stock price forecasting. For the future works Timeseries Transformer can be applied to get the results which is a most recent Deep Learning model also the real-time updates of stocks and news from authentic sources can be integrated with highly performing pretrained model using techniques like transfer learning. Subjected to time and resources influence of social media on stocks can be measured and included in the model in the future.

References

i. Hochreiter, S. and Schmidhuber, J. (1997) demonstrated the power of LSTMs in solving time-series problems like stock price prediction.
ii. Breiman, L. (2001) introduced Random Forest as a versatile model that performs well in complex regression tasks.
iii. Chen, T. and Guestrin, C. (2016) showed how XGBoost can outperform other models in tasks that involve tabular datasets like stock prices.
iv. Figure 1 Analytics Vidhya. (2021, January 1). Understanding the Architecture of LSTM. [Figure 1] Retrieved from https://www.analyticsvidhya.com/blog/2021/01/understanding-architecture-of-lstm/
v. Cortes, C. and Vapnik, V. (1995) introduced SVM as a powerful technique for classification and regression, which makes it ideal for financial market predictions.