Stock Price Prediction

1st Ashwin Dhanasamy
Dept. of Computer Science (of Aff.)
Stevens Institute of Techonology (of Aff.)
Hoboken, United States
adhanasa@stevens.edu

2nd Daniel Salib
dept. of Computer Engineering (of Aff.)
Stevens Institute of Technology (of Aff.)
Hoboken, United States
dsalib@stevens.edu

3rd Aman Gupta
Dept. of Computer Science ,
Stevens institute of technoogy
Hoboken, United States
agupta47@setevens.edu

***Abstract*—Here is ONE short paragraph summarizing your project including the problem statement, ML algorithms you use to solve the problem, experimental results, and major contribution of this work (e.g., advantages over existing solutions if any).**

I. Introduction

The accuracy of stock prediction models has significant implications for investors and traders, as it can directly impact their financial decisions. Traditional methods of stock analysis, such as fundamental and technical analysis, have limitations, and the advent of Machine Learning has opened up new possibilities for predicting stock prices. The team's focus on MLP as the primary algorithm for stock prediction is based on its proven effectiveness in this area. However, the team will also explore other popular algorithms, such as LSTM and ANN Backpropagation, to compare their performance with MLP. This approach will allow for a more comprehensive understanding of the strengths and weaknesses of each algorithm and help in selecting the best algorithm for this specific task. The team's four-phase implementation plan aims to ensure a systematic approach to the development of the stock prediction model. The first phase involves conducting extensive research on MLP algorithms and acquiring relevant stock market datasets. The team will then form and implement the model, fine-tuning it to ensure that it produces accurate predictions. In the final phase, the team will individually research and compare the performance of MLP with alternate algorithms, such as LSTM and ANN Backpropagation. This phase will enable the team to identify the most effective algorithm for predicting stock prices. Through this project, the team hopes to contribute to the development of more accurate and reliable stock prediction models. By leveraging the power of Machine Learning algorithms and conducting thorough research, the team aims to achieve its objective of accurately predicting stock prices. Ultimately, this project will contribute to a better understanding of the field of stock prediction and facilitate more informed financial decision-making.

II. Related Work

In the field of stock market prediction, the research being conducted can be generalized into 4 topics:

1) Forecasting: In this topic, research carries out to estimate, predict, or predict stock price returns using forecasting or regression algorithms
2) Grouping: In this topic, the research implementation uses a classification algorithm to classify research into two or more classes such as "UP and Down," "Buy and Hold," or "Buy, Sell, Hold.
3) Clustering: In this topic, the research implementation uses a clustering algorithm in which stocks will be grouped based on investment decision-making.
4) Association: In this topic, research will find the relationship between stock price movements from one thing, for example, the relationship between the emergence of bullish and bearish signal indicators for stock price movements.

Several forms of analysis are carried out with different attributes in predicting stock prices. The three most core types of the analysis found in this study are technical analysis, sentiment analysis, and fundamental analysis. Technical analysis uses historical stock market data or technical data as attributes, which will then be used to predict the stock market. The attributes used are different; some researchers use historical data directly, such as open prices, closing prices, volumes, etc. [3], [8], [14]. In addition, some use attributes in the form of technical data such as Moving Average, Bollinger Bands, Weighted Moving Average (WMA) and share other technical data attributes [4], [6], [11]. Sentiment analysis is carried out by analyzing data or sentiment from an object in the form of news and social media to predict stock prices. In research [15] conducted sentiment analysis using Twitter data with tweet and date indicators. Furthermore, fundamental analysis is an analysis that uses a company's financial data to look at the financial health of the company, the attributes that exist also vary.

In addition, it is possible to use more than one form of analysis, in other words combining two or more forms of analysis above to predict stocks. Sentiment and technical analysis mean using technical and sentiment attributes to predict stocks; the attributes used are different. As done by [1], [5], [9], [10], [12] using a combination of technical attributes and analysis sentiment to be an attribute of predicting stock prices. Then also several studies using a combination of technical and fundamental analysis in which technically and fundamental attributes company's used to predict stock prices [2], [7], [13].

The use of technical attributes is the most commonly found in predicting the stock price by 56%. The combined use of technical attributes and other sentiments by 23%, the combined use of fundamental and technical attributes by 15%, the use of fundamental attributes by 3%, sentiment analysis by 3%, and there has been no combined use of fundamental attributes and sentiment analysis, and the use of all three at once.

III. Our Solution

This section elaborates your solution to the problem.

A. Description of Dataset

The primary dataset used in this project is the daily stock price data sourced from Yahoo Finance. This dataset provides historical daily stock price information for a variety of companies. In addition to the stock price data, the team will also incorporate additional datasets such as corporate actions, financial statements, and key financial ratios to enhance the predictive power of the model. The daily stock price dataset contains information on the opening price, closing price, highest and lowest prices, and volume traded for each day. The data covers a wide range of companies and indices, making it a valuable resource for financial analysis and modeling. The additional datasets, such as corporate actions, financial statements, and key financial ratios, provide valuable information that can be used to improve the accuracy of the predictive models, particularly when using ANN Backpropagation algorithm. Corporate actions, such as mergers and acquisitions, stock splits, and dividend payments, can impact the stock prices of the affected companies. By incorporating this information into the model, the team can capture the effects of these events and adjust their predictions accordingly. Financial statements, such as income statements, balance sheets, and cash flow statements, provide insights into a company's financial health and performance. These metrics can be used as input features in the ANN model to capture the impact of the financial performance on stock prices. Key financial ratios, such as price-to-earnings (P/E) ratio, price-to-book (P/B) ratio, and return-on-equity (ROE) ratio, provide insights into a company's valuation, profitability, and efficiency. These ratios can be used as additional input features in the ANN model to capture the relationship between these financial metrics and stock prices. Incorporating these datasets into the ANN Backpropagation model can help to improve the accuracy and reliability of the predictions, by providing more comprehensive and nuanced information about the companies and their performance. Additionally, by using ANN Backpropagation, the model can learn the complex and non-linear relationships between the input features and the output (stock prices) through a process of iterative learning, which can further enhance the predictive power of the model. Overall, the Yahoo Finance datasets provide a rich source of information for developing and testing predictive models. The dataset is sufficiently large enough and there are enough parameters that the team should not run into overfitting/underfitting our model. By incorporating additional datasets such as corporate actions, financial statements, and key financial ratios, the team can gain deeper insights into the factors that influence stock prices and develop more accurate prediction models.

B. Machine Learning Algorithms

Based on a survey of the most widely used algorithms for stock market prediction, we chose the following:

- LSTM
- MLP
- Random Forest

1) LSTM LSTM networks are a popular choice for stock price prediction due to their ability to capture long-term dependencies, memory of past information, non-linear modeling capability, flexibility in handling input data, robustness to sequence length, and scalability. These properties make LSTM networks well-suited for modeling the complex and dynamic nature of stock price data and predicting future stock prices with accuracy.

In our system, the number of units is set at *64*, the error function used is *Mean Square Error*, the activation function used is *ReLU* activation function.

2)MLP MLP is suitable for stock market prediction because it can capture non-linear relationships between inputs and outputs. It can also handle complex data sets with many features. There are many ways to use MLP for stock market prediction. For example, one way is to use MLP for forecasting up and down stock prices the next day. Another way is to use MLP as a pattern recognition model that uses machine learning methods to improve stock trading decisions.

3)Random Forest Random Forests are suitable for stock market prediction because they can take into account various financial indicators such as company earnings, dividends, and economic data. The algorithm can handle noisy data and reduce overfitting. The purpose of a random forest is to reduce the variance of the prediction of individual decision trees. The random forest technique can handle large data sets due to its capability to work with many variables running to thousands. The hyperparameters such as $\max_d epth, min_samples_leaf, min_samples_split, n_estimatorsthatwew$ $max_depth = 15, min_samples_leaf = 1, min_samples_split = 2, n_estimators = 500$.

C. Implementation Details

Data Loading and Preprocessing: The first step in implementing a stock price prediction model is to load and preprocess the data. This involves obtaining historical stock price data from reliable sources, here we have used the dataset from the Yahoo Finance. The data is typically represented as time-series data, with each data point consisting of the stock's historical prices, volume, and other relevant features. The data is then preprocessed to remove any missing values, outliers, or redundant features that may adversely affect the model's performance. Techniques such as data imputation, scaling, and normalization can also be applied to ensure that the data is in a suitable format for training the model.

LSTM Model Architecture: A prominent variety of recurrent neural network (RNN) that works well with time-series data is called Long Short-Term Memory (LSTM). LSTMs are excellent for predicting stock price because they can identify long-term dependencies and sequential patterns in the data. The architecture of an LSTM model typically comprises of numerous layers of LSTM cells, with hidden states and gates in each cell to control the flow of input. A list of past stock prices serves as the LSTM model's input, and its forecast stock price for the following time step serves as its output. To avoid overfitting, the architecture can be further altered by modifying hyperparameters like the number of LSTM layers, the quantity of units in each layer, the activation functions, and the dropout rate.

Model Compilation and Training: It is necessary to build the LSTM model architecture with the proper loss functions, optimizers, and evaluation metrics after it has been defined. For regression problems, it is usual practice to employ Mean Squared Error (MSE) or Root Mean Squared Error (RMSE), and to optimize the model weights using stochastic gradient descent (SGD) or the Adam optimizer. Following that, the model is trained using historical stock price data, with a subset of the data set set aside for validation to track the model's progress during training. The gradient of the loss function and the optimizer's learning rate are used to iteratively update the model weights during training. Early stopping strategies can be used to prevent overfitting and choose the top-performing model depending on the model's performance after being trained for numerous epochs.

Model Prediction and Evaluation: The model can be used to forecast stock prices based on unobserved data once it has been trained. The model makes forecasts for the following time step using historical stock prices as input. The performance of the model can be assessed by contrasting the anticipated stock prices with the actual stock prices. The accuracy of the model's predictions can be evaluated using metrics like MSE, RMSE, or MAE. The model's predictions can also be visually examined and contrasted with the actual stock prices using visualization techniques like line plots or candlestick charts. To make sure that the model achieves the desired accuracy, it is critical to assess its performance on a variety of assessment measures and compare it to benchmark models or baselines. Hyperparameter Tuning: Hyperparameters, such as learning rate, batch size, or number of epochs, are variables that affect how the model behaves. The performance of the model can be dramatically impacted by tuning these hyperparameters. Techniques for hyperparameter tuning include grid search, random search, and Bayesian optimization. In contrast to random search, which chooses hyperparameter values at random, grid search involves testing the model with various combinations of hyperparameter values in a predetermined grid. A more sophisticated method called Bayesian optimization uses probabilistic models to direct the search for the ideal hyperparameter values.

A different validation set is frequently used to evaluate the model's performance for various hyperparameter values, which is utilized to modify the hyperparameters. The best-performing hyperparameter values are chosen based on the validation findings after the model has been trained and validated using various hyperparameter values. Overfitting the hyperparameters to the validation set should be avoided as it could lead to too optimistic performance estimations. To lessen this risk, strategies like cross-validation or time-series cross-validation might be used.

Feature Engineering: Another crucial step in putting a stock price prediction model into practice is feature engineering. The effectiveness and accuracy of the model can be considerably impacted by the selection of pertinent features. Other pertinent information, such as technical indicators, sentiment analysis of news or social media data, economic indicators, or market sentiment, can be added as input features to the model in addition to historical stock prices and volume. The most pertinent features for the model can be found using feature selection approaches like correlation analysis, feature importance, or recursive feature removal. To avoid overfitting

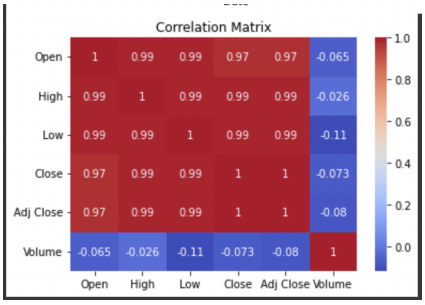

Fig. 1. Correltaion matrix

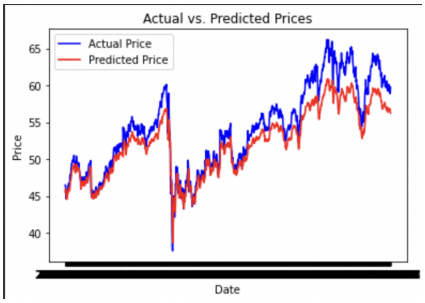

Fig. 2. Actual vs. predicted prices

or noise from irrelevant features, it's critical to find a balance between incorporating enough pertinent features.

Hyperparameter Tuning: Hyperparameters are parameters that control the behavior of the model, such as the learning rate, batch size, or number of epochs. Tuning these hyperparameters can significantly impact the performance of the model. Grid search, random search, or Bayesian optimization are commonly used techniques for hyperparameter tuning. Grid search involves testing the model with different combinations of hyperparameter values in a predefined grid, while random search randomly selects hyperparameter values for testing. Bayesian optimization is a more advanced technique that uses probabilistic models to guide the search for optimal hyperparameter values. To tune the hyperparameters, a separate validation set is often used to assess the model's performance for different hyperparameter values. The model is trained and evaluated multiple times with different hyperparameter values, and the best-performing hyperparameter values are selected based on the validation results. It is important to avoid overfitting the hyperparameters to the validation set, as it may result in over-optimistic performance estimates. Techniques such as cross-validation or time-series cross-validation can be employed to mitigate this risk.

Feature Engineering: Feature engineering is another important aspect of implementing a stock price prediction model. The choice of relevant features can significantly impact the model's accuracy and performance. Besides historical stock prices and volume, other relevant features such as technical indicators, sentiment analysis of news or social media data, economic indicators, or market sentiment can be included as input features to the model. Feature selection techniques such as correlation analysis, feature importance, or recursive feature elimination can be used to identify the most relevant features for the model. It is important to strike a balance between

including enough relevant features and avoiding overfitting or noise from irrelevant features.

Model Monitoring and Update: Once the stock price prediction model is deployed in a production environment, it is crucial to monitor its performance and update it periodically to ensure its accuracy and reliability. Monitoring techniques such as tracking prediction errors, monitoring model drift, or evaluating model performance against updated data can help identify and address any degradation in model performance. Model retraining or updating can be scheduled based on a predefined frequency or triggered by certain events or performance thresholds. It is important to continuously evaluate the model's performance and make necessary updates to maintain its accuracy and reliability in a dynamic market environment.

Performance Evaluation and Model Selection: The performance of the stock price prediction model is evaluated using various metrics such as MSE, RMSE, MAE, or accuracy. These metrics are used to compare the performance of different models or ensembles and select the best-performing model for deployment. Additionally, benchmark models or baselines can be used as a reference for performance comparison. It is important to evaluate the model's performance on multiple metrics and compare it with benchmark models or baselines to ensure its accuracy and reliability. The selected model or ensemble should meet the desired performance criteria and exhibit robustness across different evaluation metrics and test datasets.

Interpretability and Explainability: Interpretability and explainability of the stock price prediction model are important aspects, especially in regulated financial markets. Techniques such as model interpretability algorithms, feature importance analysis, or model-agnostic interpretability methods can be employed to explain the model's predictions and provide insights into the factors driving the predictions. Explainable models or interpretable ensembles can provide stakeholders with a better understanding of the model's predictions and build trust in the model's reliability.

IV. Comparison

The nature of the data, the intricacy of patterns in the data, the availability of the data, model assumptions, hyperparameter settings, and other factors can all affect how well a machine learning algorithm performs. Here are some potential explanations for why one algorithm might be superior to another:

Model Complexity: If the data contains complex patterns or relationships, more complicated algorithms like LSTM and ANN may perform better than simpler ones like linear regression or ARIMA because they are better at capturing nonlinear patterns and dependencies in the data.

Data accessibility: complicated models, such as LSTM and ANN, may be better able to extract more complicated patterns from the data than simpler models, such as linear regression or ARIMA, which may struggle when faced with sparse data.

Model Assumptions: Stationarity and linearity are two unique assumptions of the ARIMA and linear regression models, respectively. The performance of the stock price data could be affected if these hypotheses are incorrect.

V. Future Directions

Combining multiple models to improve the overall prediction performance. Different models, such as multiple LSTM models with different hyperparameter settings, or other types of models such as ARIMA, GARCH, or XGBoost, can be combined to form an ensemble. Ensemble methods such as stacking, bagging, or boosting can be employed to combine the predictions from multiple models. Ensemble methods can help mitigate the limitations or biases of individual models and result in more accurate and robust prediction

VI. Conclusion

This section summarizes this project, i.e., by the extensive experiments and analysis, do you think the problem is solved well? which algorithm(s) might be better suitable for this problem? Which technique(s) may help further improve the performance?

Last but not the least, don't forget to include references to any work you mentioned in the report.

VII. REFERENCES

1) G. Attanasio, L. Cagliero, P. Garza, and E. Baralis, "Combining news sentiment and technical analysis to predict stock trend reversal," in Combining news sentiment and technical analysis to predict stock trend reversal, Nov 2019, vol. 2019-Novem, pp. 514–521, doi: 10.1109/ICDMW.2019.00079.
2) A. Namdari and Z. S. Li, "Integrating Fundamental and Technical Analysis of Stock Market through Multi-layer Perceptron," Okt 2018, doi: 10.1109/TEMSCON.2018.8488440.
3) F. Zhou, Q. Zhang, D. Sornette, and L. Jiang, "Cascading logistic regression onto gradient boosted decision trees for forecasting and trading stock indices," Appl. Soft Comput. J., vol. 84, Nov 2019, doi: 10.1016/j.asoc.2019.105747.
4) S. Boonpeng and P. Jeatrakul, "Decision support system for investing in stock market by using OAA-Neural Network," in Proceedings of the 8th International Conference on Advanced Computational Intelligence, ICACI 2016, Apr 2016, pp. 1–6, doi: 10.1109/ICACI.2016.7449794.
5) K. M, K. J, E. R. T, and A. S, "Stock Market Prediction with Historical Time Series Data and Sentimental Analysis of Social Media Data," Proc. Int. Conf. Intell. Comput. Control Syst., 2020.
6) IEEE Computational Intelligence Society, Institute of Electrical and Electronics Engineers, and B. C. IEEE World Congress on Computational Intelligence (2016: Vancouver, "Equity Price Direction Prediction For Day Trading Ensemble Classification Using Technical Analysis Indicators With Interaction Effects," IEEE Comput. Intell. Soc. Inst. Electr. Electron. Eng. IEEE World Congr. Comput. Intell. (2016 Vancouver, B.C.), 2016.
7) L. S, "Impact of Financial Ratios and Technical Analysis on Stock Price Prediction Using Random Forests," Ethical Integr. Comput. Drone Technol.

Humanit. Sustain. 9th-11th Nov. 2017, Kuching, Sarawak, Malaysia, 2017.

8) O. B. Sezer, M. Ozbayoglu, and E. Dogdu, "A Deep Neural-Network Based Stock Trading System Based on Evolutionary Optimized Technical Analysis Parameters," in Procedia Computer Science, 2017, vol. 114, pp. 473–480, doi: 10.1016/j.procs.2017.09.031.
9) X. Zhang, J. Shi, D. Wang, and B. Fang, "Exploiting investors social network for stock prediction in China's market," J. Comput. Sci., vol. 28, pp. 294–303, Sep 2018, doi: 10.1016/j.jocs.2017.10.013.
10) W. Chen, C. K. Yeo, C. T. Lau, and B. S. Lee, "Leveraging social media news to predict stock index movement using RNN-boost," Data Knowl. Eng., vol. 118, no. December 2017, pp. 14–24, 2018, doi: 10.1016/j.datak.2018.08.003.
11) S. Lauguico, R. Concepcion, J. Alejandrino, D. Macasaet, R. R. Tobias, and E. Bandala, "A Fuzzy Logic-Based Stock Market Trading Algorithm Using Bollinger Bands," 2019 IEEE 11th Int. Conf. Humanoid, Nanotechnology, Inf. Technol. Commun. Control. Environ. Manag. (HNICEM)., 2019.
12) V. Sharma, R. Khemnar, R. Kumari, and D. B. R. Mohan, "Time Series with Sentiment Analysis for Stock Price Prediction," 2019 2nd Int. Conf. Intell. Commun. Comput. Tech. Manipal Univ. Jaipur, Sep. 28-29, 2019.
13) E. Beyaz, F. Tekiner, X. J. Zeng, and J. Keane, "Stock Price Forecasting Incorporating Market State," in Proceedings - 20th International Conference on High Performance Computing and Communications, 16th International Conference on Smart City and 4th International Conference on Data Science and Systems, HPCC/SmartCity/DSS 2018, Jan 2019, pp. 1614–1619, doi: 10.1109/HPCC/SmartCity/DSS.2018.00263
14) Y.-L. Cai, K. Kannan, Y.-H. Xie, and L. Zhao, "E-Commerce: Stock Market Analysis Blended With Mining and Ann," 2019 IEEE Int. Conf. Ind. Eng. Eng. Manag., 2019.
15) N. N. Reddy, N. E, and V. Kumar, "Predicting Stock Price Using Sentimental Analysis Through Twitter Data," Proc. IEEE Conecct 2020 6th Int. Conf. Electron. Comput. Commun. Technol. July 2-4, 2020, 2020.

REFERENCES