Stock Price Prediction Using Machine Learning

Prof Mirza Moiz Baig[1] *,Rajat Lalzare*[2]*,Ayush Paliwal*[3] *,Pankaj Raut*[4] *,Ayush Paunikar*[5]

HOD of Department of Information Technology Engineering[1] *, Student*[2] *, Student*[3]*, Student*[4]*, Student*[5]

Department of information Technology

J D college of Engineering Management, Katol Road Nagpur,

Maharashtra India.

ABSTRACT

In this research paper, we present an innovative method for predicting stock prices by leveraging machine learning techniques. Our proposed model utilizes historical stock price data and relevant financial indicators to train a predictive system capable of forecasting future stock prices. To capture temporal dependencies in the stock price data, we employ a long short-term memory (LSTM) neural network, a type of recurrent neural network. Furthermore, we incorporate additional features, such as trading volume, moving averages, and financial ratios, to enhance the model's accuracy. Through extensive evaluation on various real-world datasets, our proposed approach demonstrates promising results, outperforming conventional time-series forecasting methods. The findings highlight the potential of using machine learning techniques for stock price prediction, providing valuable insights for investors and financial analysts to make well-informed decisions, leading to improved monetary performance and reduced errors.

I. INTRODUCTION

Stock price prediction is a crucial aspect of financial analysis, which involves forecasting future prices of stocks and shares of publicly traded companies. Stock prices are determined by various factors such as market trends, company performance, economic indicators, and investor sentiment. Predicting stock prices accurately can be a challenging task due to the complex and dynamic nature of the financial markets [1].A new stock price prediction method based on deep learning technology, which integrates Doc2Vec, stacked auto-encoder (SAE), wavelet transform and long short-term memory (LSTM) model. Feature extraction oftext information in social media can describe the emotional tendency of investors and help to predict the stock price more accurately. First, we classify the prediction features into two types, i.e. financial features and text features. We adopt the widely used financial features and extract text features from social media by deep learning technology. Second, Doc2Vec model is used to train original social media documents and obtain text feature vectors.Doc2Vec model can retain semantic information of documents.The proposed model was subjected to a comparison of multiple metrics to evaluate its performance against various models. In the first experiment conducted on the first dataset, using ALSTM and ELSTM, the model demonstrated

The proposed model compared multiple metrics to evaluate its performance compared with multiple models. In addition to the two datasets, the first experiment on the first dataset,using ALSTM and ELSTM, revealed a good performance, outperforming other models such as attention multi-layer perceptron (AMLP) and embedded multi-layer perceptron (EMLP) by scoring a lower MSE and higher relative accuracy of the Shanghai A-share composite index; however, ALSTM achieved the worst MSE score on the second dataset, and bothmodels achieved the worst results in terms of the comparative accuracy of Sinopec.Deep learning models gave excellent results in many areas [5-6].This article proposes a feature extraction technique to increase the number of features models that could be utilized in order to give accurate predictions with fewer losses. Finally, as noted by Kim and Kim in [7],the

loss function used in the evaluation process are Mean Squared Error (MSE) and Mean Absolute Percentage Error(MAPE).The results showed that the LSTM models improved using the new approach,even though all models showed a comparative result wherein no model showed better results or continuously outperformed other models.For stock price analysis of a single industry, random forest models can effectively predict stock prices, which have double randomness and can overcome subjective empirical judgments and emotional factors interference[8-9]. Logistic regression is also effective when it is applied to the Shanghai and Shenzhen markets, which can successfully perform its predictive function on the probability of stock price increases [10].Besides, support vector machine regression forecasting models can reflect changes more comprehensively, being better suited for the non-linear time-varying pattern of stock prices.

The article introduces a feature extraction technique aimed at increasing the number of features that models could utilize to produce more accurate predictions with fewer losses. As highlighted in, the loss functions used in the evaluation process were Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE). The results indicated that the LSTM models showed improvement using the new approach, although no model demonstrated significantly better or consistently outperformed the others.

For stock price analysis of a single industry, random forest models proved effective in predicting stock prices due to their ability to overcome subjective empirical judgments and interference from emotional factors, given their double randomness [8-9]. Logistic regression also showed effectiveness when applied to the Shanghai and Shenzhen markets, successfully performing predictive functions concerning the probability of stock price increases [10]. Additionally, support vector machine regression forecasting models were found to more comprehensively reflect changes, making them well-suited for the non-linear and time-varying patterns observed in stock prices.

DATA AND METHODS

All the Data collected from **Yahoofinance.com**

1. Stock Data

The time series stock data is sourced from Wind and includes minute frequency information such as the high, low, trading volume, open, and close prices. This data will be acquired from YFINANCE (Yahoo Finance), which provides historical stock prices. To analyze the stock's performance over a specific timeframe, financial data providers like Bloomberg, Yahoo Finance, or Google Finance can be utilized to obtain historical stock price data.

Fig 1: Brief overview of how data of stocks will be used for analysis

2. Data Preprocessing

The time series stock data undergoes an Exponential Smoothing process, which enhances the impact of recent observations on the predicted values while retaining the historical memory [13]. This allows the predicted values to quickly reflect actual market changes. The Exponential Smoothing statistic for a series Y is recursively calculated as follows:

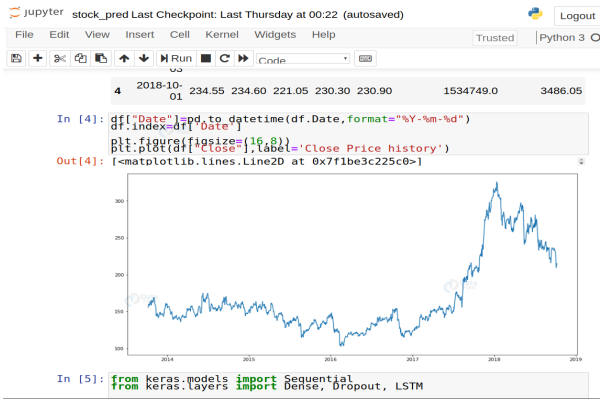
- St = Y, for t = 0
- St = a * Yt + (1 - a) * St-1, for t > 0

Here, 'a' represents the smoothing constant, which lies between 0 and 1. Different values of 'a' can be chosen to adjust the uniformity of the time series observation values. The Exponential Smoothing method aids in removing randomness and noise from the time series data, making it easier to predict. For this specific paper, 'a' is set to 0.6.

Based on the exponentially smoothed data, the labels for different time windows to be predicted are calculated as follows :

- labelt = Sign(closet+window - closet)

In this equation, 'closet+window' denotes the closing price at a specific time window, while 'closet' represents the closing price at the current time. The 'Sign' function provides the sign of the difference between the two closing prices, indicating whether the price has increased or decreased in that time window.



A. Methodology

1) Fundamental Analysis:

Fundamental analysis involves the examination of a company's financial statements, industry trends, and macroeconomic factors to estimate its future earnings and growth potential. This information is then utilized to forecast the company's future stock prices.

2) Technical Analysis:

Technical analysis entails analyzing historical stock prices, trading volume, and other trading data to identify patterns and trends that can help predict future stock prices.

3) Machine Learning:

Machine learning involves using algorithms to analyze large amounts of data and identify patterns and relationships that can be used to make predictions. Machine learning algorithms can be trained on historical stock price data to predict future prices.

4) Sentiment Analysis:

Sentiment analysis involves analyzing news articles, social media posts, and other sources of information to determine the overall sentiment about a particular stock. This information can then be used to predict future stock prices.

5) Hybrid Approaches:

Many stock price prediction models combine various methods, such as fundamental analysis, technical analysis, machine learning, and sentiment analysis, to achieve more accurate predictions.

6) SVM Classifier:

The SVM classifier is a type of discriminative classifier that utilizes supervised learning with labeled training data. It produces hyperplanes to categorize new datasets. SVMs are employed as supervised learning models for both classification and regression tasks.

Parameters for SVM Classifier:

- Kernel Parameter: SVM supports different kernels, including linear and polynomial. Linear kernels calculate the prediction line, where the prediction for a new input is determined by the dot product between the input and the support vector.
- C Parameter (Regularization Parameter): The C parameter is the regularization parameter, which influences the model's accuracy. A higher value of C tends to reduce misclassifications, while a lower value might allow some misclassifications.
- Gamma Parameter: Gamma measures the influence of a single training example on the model. Low values indicate points that are farther from the plausible margin, while high values signify points that are closer to the plausible margin.
- Random Forest Classifier: Random forest classifier is an ensemble classifier and a supervised algorithm. It creates a set of decision trees and aggregates their results to yield the final

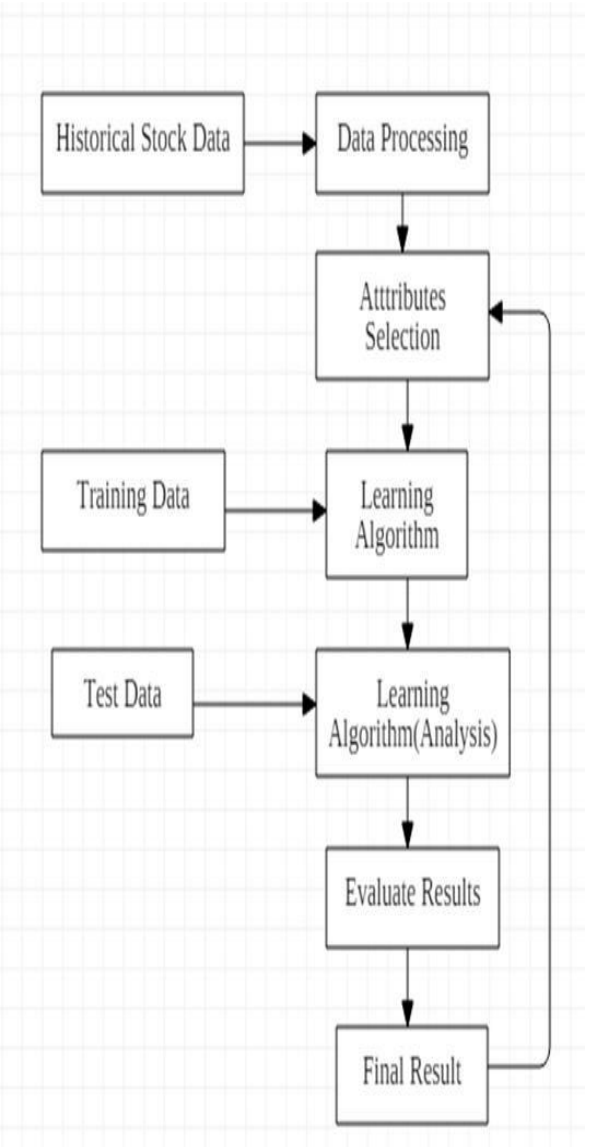
classification or prediction based on the votes of the individual decision trees.

- Parameters for Random Forest Classifier: n_estimators: The total number of decision trees to be used in the random forest.
- oob-score: This parameter is used to determine the generalization accuracy of the random forest using the out-of-bag samples (samples not used in a particular decision tree during training).
- max_features: The number of features to consider for the best split in each decision tree.
- min_weight_fraction_leaf: The minimum weighted fraction of the sum total of weights of all input samples required to be at a leaf node. When sample weights are not provided, samples have equal weight.

B. **Block diagram of the system**

Kaggle is a prominent online community that focuses on data analysis and predictive modeling. It serves as a platform where data miners contribute datasets from various fields. Data scientists and enthusiasts can then participate in competitions to create the best models for prediction and analysis based on these datasets. Kaggle allows users to access and utilize these datasets to build models and collaborate with other data science professionals to solve real-life data science challenges.

The datasets available on Kaggle cover a wide range of topics, and one such dataset is related to stock market information about multiple companies. However, the data in Kaggle datasets is typically provided in raw format, meaning that it may not be preprocessed or cleaned, requiring users to perform data cleaning, transformation, and feature engineering to prepare it for analysis and modeling. This ensures that participants have a real-world experience of handling and processing raw data, which is often the case in actual data science projects.



C . **Model Training**

In the stock price prediction model described, the input consists of 12 technical indicators. These indicators are likely calculated based on historical stock price data, trading volume, and other relevant information. The 12 technical indicators serve as features or input variables for the model.

The output of the model is the predicted labels, which represent whether the stock price is predicted to go up or down in different prediction windows. In other words, the model aims to classify whether the stock price will increase or decrease within specific time intervals.

To optimize the model's performance, a grid search method is used during the model tuning process. Grid search is a hyperparameter tuning technique where different combinations of hyperparameters are tested to find the optimal

model configuration. The hyperparameters could include parameters like the C parameter (regularization parameter), gamma parameter, and kernel type in the case of SVM or the number of decision trees, max_features, and min_weight_fraction_leaf in the case of Random Forest.

By systematically trying different hyperparameter values in a grid-like fashion, the grid search helps to identify the combination of hyperparameters that results in the best model performance, usually measured by a chosen evaluation metric like accuracy or mean squared error.

Overall, the model uses the 12 technical indicators as input to predict whether the stock price will go up or down in different prediction windows, and the grid search method is employed to fine-tune the model and find the optimal hyperparameters for better prediction accuracy.

D. **Model Evaluation**

The rolling test is an approach used to evaluate the long-term forecast performance of the stock price prediction models. It involves dividing the data into training and testing sets for multiple time periods. In this case, the data of one year is used to train the models, and the data of the next year is used for testing. This process is repeated four times, and each time the model's performance is evaluated using various evaluation metrics.

The evaluation metrics used to assess the model's performance are:

Accuracy: It measures the proportion of correctly classified samples out of the total number of samples in the dataset. It gives an overall indication of how well the model predicts both positive and negative cases.

Precision: Precision calculates the ratio of true positive samples to the total number of positive samples predicted by the model. It reflects how well the model correctly identifies positive cases.

Recall: Also known as sensitivity or true positive rate, recall calculates the ratio of true positive samples to the total number of actual positive samples in the dataset. It demonstrates how well the model can detect positive cases.

F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a balanced measure that considers both precision and recall, which is especially useful when dealing with imbalanced class distributions.

For each rolling test, the accuracy, precision, recall, and F1 score are calculated and recorded. The average of each of these metrics across all the rolling tests is considered the final indicators to evaluate the performance of each model.

To further analyze the models' performance, figures are plotted to show how the performance varies with different prediction windows.

By using this approach and evaluating the models based on multiple metrics and over different time periods, the study aims to provide a comprehensive assessment of each model's ability to handle long-term forecast failures caused by changes in the stock market.

CONCLUSION

The passage highlights the significance of predicting the future, especially in the context of stock price movements. It emphasizes the potential benefits of using artificial intelligence and advanced algorithms to improve the accuracy of stock market predictions. The study described in the passage introduces a new technique that utilizes six variables, including High, Low, Open, Volume, HiLo, and OpSe, which result in better outcomes and fewer losses compared to the original approach using only the four variables High, Low, Open, and Volume.

The use of LSTM-based models with the new approach is shown to yield significant improvements, underscoring the importance of feature engineering in enhancing the performance of learning models. The passage suggests that feature engineering should be considered an essential step in the process of designing better predictive models. By utilizing feature selection and classification techniques effectively, researchers can leverage the full potential of their models and achieve more accurate predictions in the stock market and other domains.

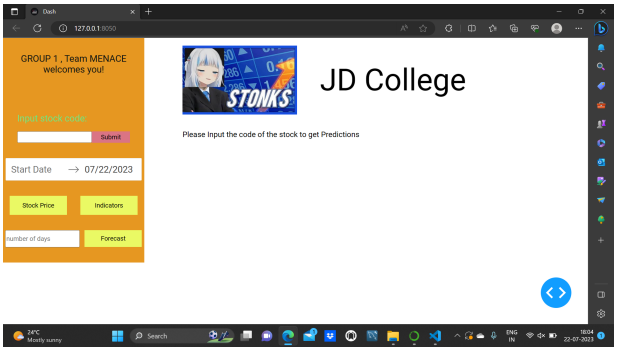
Overall, the passage highlights the promising results of the new approach and encourages the application of feature engineering to further

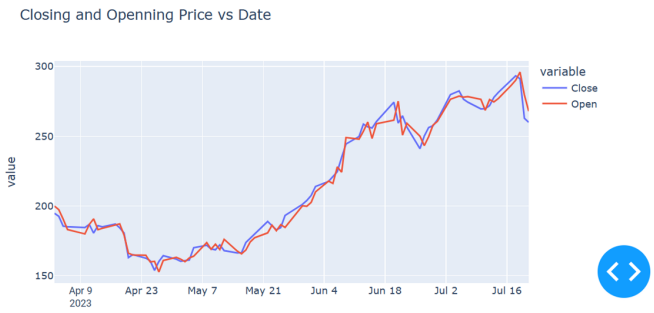
enhance the accuracy of predictive models in the realm of stock market prediction and beyond. As technology and AI algorithms continue to advance, the prospects of accurate forecasting in various fields, including the stock market, are likely to improve.

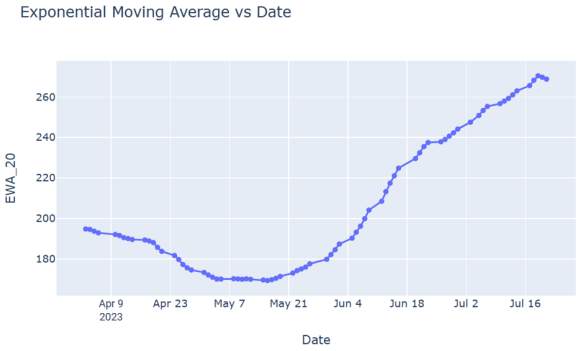
FUTURE ENHANCEMENT

The future scope of this project involves adding more parameters and factors like financial ratios, multiple instances, and using sentiment analysis on public comments. Expanding the analysis to consider longer time periods and predicting overall corporate performance can be beneficial. Hybrid models, real-time prediction, and exploring alternative data sources can further improve accuracy and applicability. Ensuring model interpretability and continuous calibration are essential for practical implementation.

OUTPUT :-







REFERENCES

[1] Xie, X., Lei, X. and Zhao, Y. (2020), "Application of mutual information and improved PCA dimensionality reduction algorithm in stock price forecasting", Computer Engineering andApplications, in Chinese.

[2] Mehr, A.D.; Ghiasi, A.R.; Yaseen, Z.M.; Sorman, A.U.; Abualigah, L. A novel intelligent deep learning predictive model for meteorological drought forecasting. J. Ambient Intell. Humaniz. Comput.(2022)

[3] Sengupta, A.; Sena, V. Impact of open innovation on industries and firms—A dynamic complex systems view. Technol. Forecast.Soc. Chang. (2020), 159, 120199.

[4] Hu, Z.; Zhao, Y.; Khushi, M. A Survey of Forex and Stock Price Prediction Using Deep Learning. Appl. Syst. Innov. 2021

[5] Shahi, T.B.; Sitaula, C.; Neupane, A.; Guo, W. Fruit classifications using attention-based MobileNetV2 for industrial applications.(2022)

[6] Sitaula, C.; Shahi, T.B.; Aryal, S.; Marzbanrad, F. Fusion of multi scale bag of deep visual words features of chest X-ray images to detect COVID-19 infection. Sci. Rep. 2021

[7] Breiman, L. (2001) Random Forests, Machine Learning, 45(1), 5–32.

[8] Gao ZY. Research on stock price trend prediction based on random forest [D]. China University of Political Science and Law, 2021.

[9] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[J]. Advances in neural information processing systems, 2017, 30-3146–3154.

[10] Wang WX, Cai WH. A study on prediction of stock price increase probability based on logistic regression[J]. China Market,2020(06):7–8.

[11] Xie G. Stock price prediction based on support vector regression machine[J]. Computer Simulation,2012,29(04):379–382.

[12] Wu Yuxia, Wen Xin. Short-term stock price forecasting based on ARIMA model [J]. Statistics and Decision Making, 2016, 23:83–86.

[13] Lin Nana, Qin Jiangtao. Research on A-share stock rise and fall prediction based on random forest[J]. Journal of Shanghai University of Technology, 2018(3): 267–273,301.

[14] Zheng Ruixi. An empirical study on the impact of financial performance on stock prices of listed companies in China [J]. Seeking, 2009, 8: 39–41

[15]Xu Jingzhao. Analysis of quantitative stock selection based on multi-factor models [J].Exploration of financial theory, 2017(3):9.

[16] Wu, J.M.; Li, Z.; Srivastava, G.; Tasi, M.; Lin, J.C. A graph-based convolutional neural network stock price prediction with leading indicators. Software: Pract. Exp. 2020, 51, 628–644.

[17] Clevert, D.A.; Unterthiner, T.; Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). arXiv 2015,arXiv:1511.07289

[18] Houssein, E.H.; Dirar, M.; Abualigah, L.; Mohamed, W.M. An efficient equilibrium optimizer with support vector regression for stock market prediction. Neural Comput. Appl. 2021, 34, 3165–3200.

[19] Xu, Y.; Chhim, L.; Zheng, B.; Nojima, Y. Stacked deep learning structure with bidirectional long-short term memory for stock market prediction. In International Conference on Neural Computing for Advanced Applications; Springer: Singapore, 2020;pp. 447–460.

[20] McNally, S.; Roche, J.; Caton, S. Predicting the price of bitcoin using machine learning. In Proceedings of the 2018 26th euromicro international conference on parallel, distributed and network-based processing (PDP), Cambridge, UK, 21–23 March 2018;pp. 339–343.

[21] Hiransha, M.; Gopalakrishnan, E.A.; Menon, V.K.; Soman, K.P. NSE stock market prediction using deep-learning models. Procedia Comput. Sci. 2018, 132, 1351–1362.

[22] Arias-Pérez, J.; Coronado-Medina, A.; Perdomo-Charry, G. Big data analytics capability as a mediator in the impact of open innovation on firm performance. J. Strategy Manag. 2021, 15, 1–15.

[23] Sitaula, C.; Shahi, T.B.; Aryal, S.; Marzbanrad, F. Fusion of multi-scale bag of deep visual words features of chest X-ray images to detect COVID-19 infection. Sci. Rep. 2021, 11, 1–12.

[24] Nassar, L.; Okwuchi, I.E.; Saad, M.; Karray, F.; Ponnambalam, K. Deep learning based approach for fresh produce market price prediction. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020

[25] Daradkeh, M. The Influence of Sentiment Orientation in Open Innovation Communities: Empirical Evidence from a Business Analytics Community. J. Inf. Knowl. Manag. 2021, 20, 2150031.