International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 07 | July 2023 www.irjet.net p-ISSN: 2395-0072

StocKuku - AI-Enabled Mindfulness for Profitable Stock Trading

Dhiren Gangishetty[1]

[1]School of Computer Science and Engineering
VIT-AP University
Amaravati-522237, Andhra Pradesh, India

Abstract - *In today's dynamic and volatile stock market, accurate prediction of stock prices has become an essential aspect of investment decision-making. This research paper presents an innovative approach that combined the power of Natural Language Processing, intelligent Chabot technology and advanced machine learning algorithms like Linear Regression, Random Forest Regressor and XGBoost Algorithm to enhance stock price prediction and provide valuable insights to investors. The Chabot provides a user-friendly interface for inverstors to inquire about stock information, receive personalized recommendations, and obtain real-time predictions on whether to buy or sell a particular stock. Time series data is extracted using web scraping techniques, forming the foundation for developing distinct models. These models capture different aspects of stock price behaviour, enabling comprehensive prediction capabilities. The comparative analysis highlights the benefits and drawbacks of each method, enabling investors to decide which strategy best suits their tolerance for risk and investment preferences. Participants with a range of experiences offer feedback on its applicability, dependability, and overall user experience. Investors can effortlessly navigate the complexity of the stock market due to the integration of real-time data, machine learning algorithms, and interactive discourse. Also, this study contributes to the field of financial technology by outlining a cutting-edge strategy that combines many methodologies to improve stock price prediction. It also provides opportunities for further research into revolutionary technologies that could fundamentally alter how investors interact with financial data and make investment decisions, such as natural language processing and intelligent Chatbots.*

***Key Words*: Stock Price Prediction, Intelligent Chabot, XGBoost Algorithm, Natural Language Processing, Financial Technology, Random Forest Regression, Personalized Recommendations**

1. INTRODUCTION

Stock price forecasting is essential in today's quickly changing financial environment while making investing decisions. Investors and traders are always looking for precise and dependable ways to forecast stock market movements and make educated decisions about their portfolios. However, with the advent of advanced technologies and the rise of artificial intelligence, new avenues for predictive analytics have emerged [1]. The proliferation of online financial platforms and the availability of vast amounts of real-time financial data have made web scraping a valuable tool for gathering up-to-date information on stock prices, market trends, and relevant financial indicators. By leveraging web scraping techniques, we can extract comprehensive datasets from popular financial platforms such as Yahoo Finance, which serve as the foundation for our predictive models. Parallel to this, the introduction of intelligent Chatbots [2] has completely changed how users connect with and engage with various platforms. Intelligent Chatbots have evolved into potent conversational agents that can comprehend user requests and offer tailored responses by fusing natural language processing techniques with smart dialogue management [3].

1.1 Web Scraping

Web scraping is the process of extracting information from a website by "scraping" it. Theoretically, it is feasible to scrape extra data sources, such as document papers. Nonetheless, the vast majority of scraping is often performed on web pages [4]. In this work, web scraping plays a crucial role in data acquisition from the Yahoo Finance website. Web scraping is a technique used to extract data from websites by programmatically accessing and parsing the underlying HTML code of web pages. It allows us to gather real-time stock data, such as historical prices, trading volumes, and financial indicators, which serve as the foundation for our stock price prediction models.

1.2 Linear Regression

To predict the value of one variable based on the value of another variable, linear regression analysis is utilized. Based on confidence levels, when compared with polynomial and RBF regression techniques, Linear Regression provides a better prediction [5].

1.3 Random Forest Regressor

A random forest is a meta-estimator that employs averaging to increase predicted accuracy and reduce

overfitting after fitting numerous classification decision trees to different dataset subsamples.

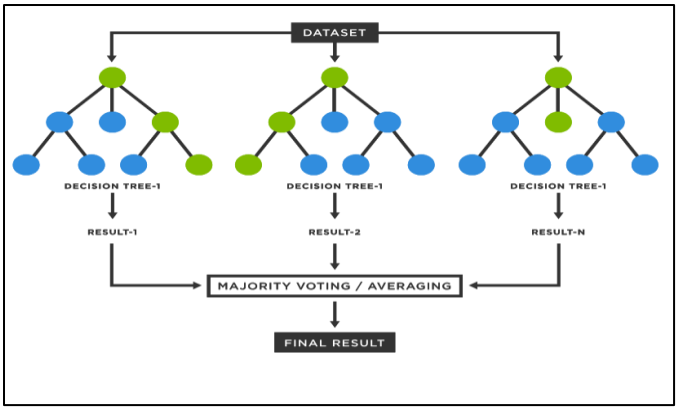

Fig -1: Working model of Random Forest Algorithm

The Bagging technique is used with the help of multiple decision trees. The output results of several decision trees are compared and averaged to give out a final output.

1.4 XGBoost Algorithm (Regression)

Extreme Gradient Boosting, also known as XGBoost, is a robust and popular machine learning algorithm renowned for its superior prediction performance and adaptability. It is an integral part of the gradient boosting method family and is esteemed for its capacity to manage an assortment of data formats and intricate relationships within the data. To create a final robust prediction, the XGBoost algorithm iteratively trains an ensemble of weak prediction models, often decision trees. Utilizing a gradient boosting framework, minimizes the residuals from the previous iteration to optimize an objective function. The model's accuracy and generalizability are steadily enhanced through this iterative process, making it ideal for challenging regression applications like stock price prediction [6]. XGBoost goal function includes a regularization term to avoid overfitting [7]. The main objective function is described as follows:

$$O = \sum_{i=1}^{n} L(yi, F(xi)) + \sum_{k=1}^{t} R(fk) + C$$

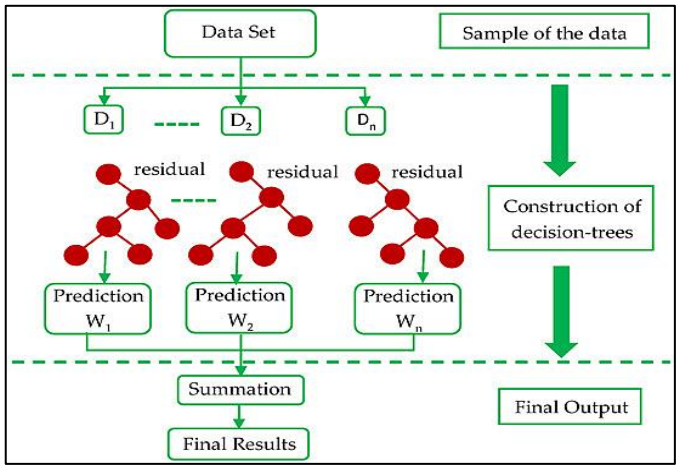

Fig -2: Structure of XGBoost Algorithm.

2. METHODOLOGY

2.1 Data Collection and Preprocessing

The data for this study was sourced from Yahoo Finance, a reputable financial platform widely used for stock market data. The financial data of an Indian market giant named *"Adani Green Energy Limited"* was extracted. To make computation easier and more practical, irrelevant columns were removed. Null values were substituted with the mean value and the data was scaled thoroughly. The dataset additionally incorporates time variables such as month, year, and date. The dataset was divided into quarters, which enhanced the correctness of the stock predictions. The forecast column was set as 'Close' and the dataset was given the test size of 0.2 (20%).

Date	Open	High	Low	Close	Volume	day	year	month	is_quarter_end
2022-11-07	2127.0	2184.000000	2127.000000	2165.699951	1094562	11	2022	7	0
2022-11-09	2174.0	2258.800049	2163.600098	2215.500000	1739571	11	2022	9	1
2022-11-10	2225.0	2245.350098	2155.000000	2176.000000	746766	11	2022	10	0
2022-11-11	2207.0	2237.899902	2177.199951	2196.699951	770505	11	2022	11	0
2022-11-14	2200.0	2200.000000	2160.000000	2178.550049	628699	11	2022	14	0

Fig -3: Sample data with the extracted columns

Python web scraping libraries such as BeautifulSoup and Selenium were utilized for automated data extraction. Data pre-processing techniques, such as normalization and feature scaling, were applied to ensure the data's suitability for the subsequent modelling stages.

2.2 Model Development

Linear Regression Model

A linear regression model was developed to capture the linear relationship between the selected independent variables and the target stock price. Multiple features, including historical price data and relevant financial indicators, were incorporated to train the model.

Random Forest Regression Model

The random forest regressor algorithm was employed to handle non-linear relationships and complex interactions among the variables. The ensemble of decision trees in the random forest model provided robust predictions by combining multiple individual predictions. This algorithm was trained with 40 estimators. 40 estimators were chosen as this number gave the best yield. The tree ensemble model is a set of CART trees [8], where the sum of the predictions of multiple trees is taken as:

$$y_i = \sum_{i=1}^{k} mi\, xij, \forall j, \forall m_k \in M$$

Where k represents the total number of trees in the random forest and M is the functional space.

XGBoost Regression Model

XGBoost algorithm was utilized to construct a powerful regression model. Its ability to handle diverse data types, handle missing values, and mitigate overfitting made it a suitable choice for accurate stock price prediction. The boosting method based on a decision tree is called boosting tree. XGBoost model is an efficient method of boosting the tree model. The boosting tree model can be interpreted as the addition model of the decision tree [9]. Minimization of the objective function when solving the decision tree t:

$$Obj^t = \sum_{i=1}^{n} l(y_i, \hat{y}_i^t) + \sum_{i=1}^{t} \Omega(f_i) = 0 = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{t-1} + f_t(x_i)\right) + \Omega(f_t) + cons\tan t$$

Arbitrary differentiable loss function and gradient descent optimization procedures are used to fit the models. session model. XGBoost version 1.0.1 is preferred in the training of this model as this version supports the quality of hyper parameters required for training the ensemble model. The model was trained with a number of estimators being 1000, maximum depth of 7, colsample_bytree parameter as 0.8. Also, the RepeatedKFold and repeated it 3 times taking the number of splits as 10. In addition to this, Cross-validation score was also determined [10].

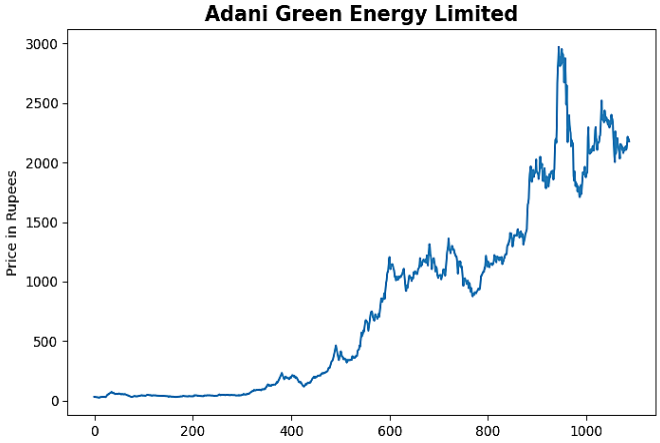

Chart -1: Price Increment of an Adani Green Energy Limited stock

2.3 Implementing the Chabot

Some unique features implemented:

- Predicting the stock price in advance
- Suggesting whether to buy/ sell/ hold the stocks in the company
- Friendly interaction with the user
- Clarifies queries on the company, like its market capital, recent purchases and headlines over the news
- Warns about the stock market and predicts whether it will crash in the future

For most of the implementation of the Chabot, the NLP techniques are performed. Sentence and word tokenization are performed using the NLTK library. The WordNetLemmatizer is utilized for lemmatization. The Chabot recognizes various greeting inputs from users and generates appropriate responses to establish a conversational tone. Using the TF-IDF vectorization technique, the Chabot generates responses by measuring the cosine similarity between the user input and pre-existing Chabot conversations. The response with the highest similarity is selected as the Chabot's reply. The Chabot interacts with users, providing real-time stock price predictions and answering queries based on the company, in this case, Adani Green Energy Ltd. The Chabot selects the most relevant response from the pre-existing Chabot conversations using cosine similarity scores. Evaluation metrics include user feedback ratings, response relevance, and the accuracy of predicted stock prices. Also, the Chabot integrates with web scraping techniques to fetch real-time financial data from Yahoo Finance for accurate stock price predictions.

Overall, the methodology encompasses data acquisition, text pre-processing, Chabot development, user interaction, evaluation and deployment. It ensures the Chabot's ability to provide real-time stock predictions, respond to user queries and offer an engaging user experience.

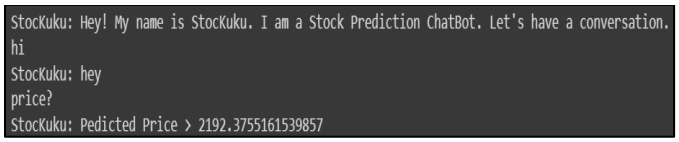

Fig -4: A simple user-Chabot conversation

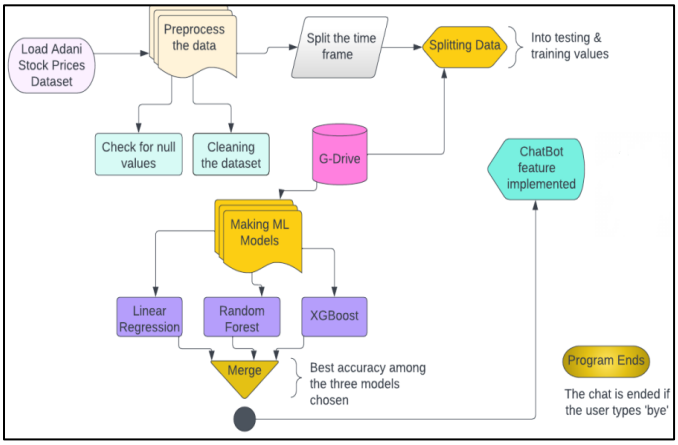

Fig -5: Working flow diagram

3. RESULTS

The anticipated stock prices closely tracked real market values, demonstrating how well the Chabot picked up on pertinent trends and patterns. High levels of engagement and satisfaction were found during the user research among people using the Chabot. The Chabot's replies were helpful, prompt, and appreciated by users for helping them make educated financial decisions. Users were able to remain apprised of the most recent market trends and modifications because of the Chabot's successful provision of real-time stock information. Users were provided with individualized insights and recommendations about their chosen equities, helping them to make wise investing decisions.

The Linear Regression model achieved an impressive accuracy of 91.32% in predicting stock prices for Adani Green Energy Limited. The Random Forest Regressor model showcased outstanding performance, yielding an accuracy of 93.44%, indicating its ability to capture complex relationships and nonlinear patterns in the data. The XGBoost algorithm exhibited remarkable accuracy, achieving 94.20% accuracy in stock price prediction, underscoring its efficacy in handling diverse data types and optimizing model performance. The XGBoost algorithm's impressive performance highlighted its ability to leverage ensemble learning and gradient boosting techniques to enhance accuracy and robustness. The excellent accuracy. levels attained by all three models possess significant practical ramifications since they enable investors to make wise choices in the complex and tumultuous stock market environment.

Table -1: Accuracy values for all the ML models

ML Model Name	Accuracy Obtained
Linear Regression	91.32%
Random Forest Regression	93.44%
XGBoost Algorithm	94.20%

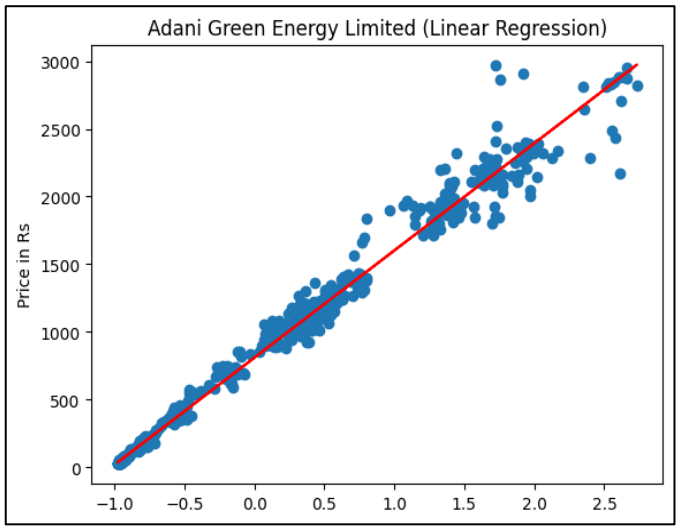

Chart -2: Linear Regression Line in a scatter plot of Stock prices

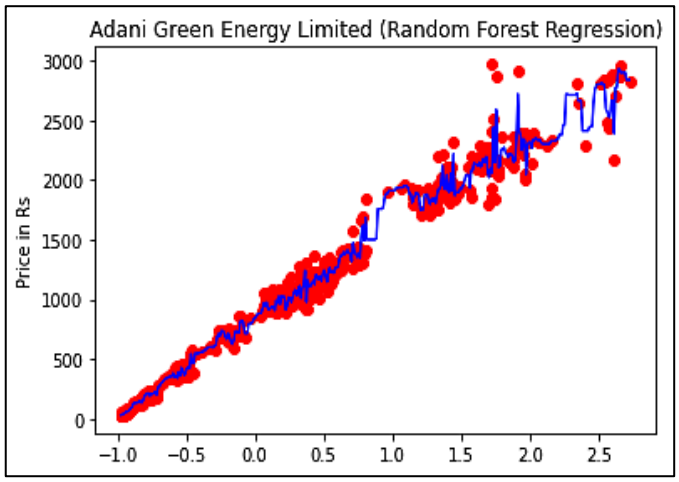

Chart -3: Random Forest Regression Line in a scatter plot of Stock prices

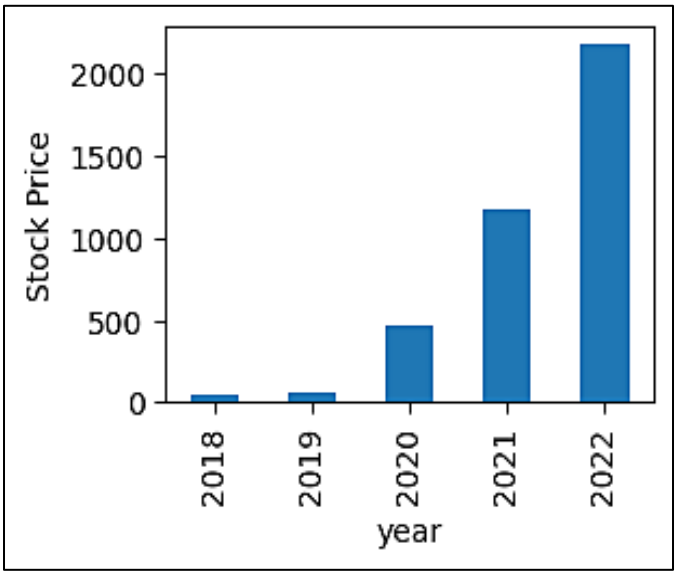

Chart -4: Graph representing hike in stock prices over the years

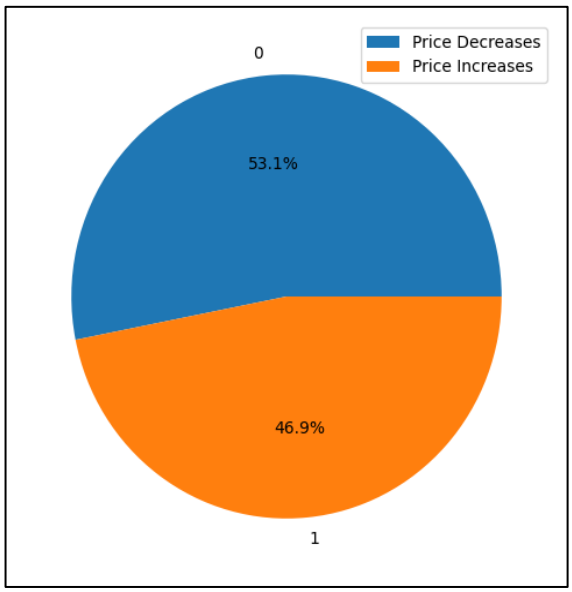

Chart -5: Distribution of target values in the dataset

Chart-5 allows for a quick understanding of the proportion of instances where the stock price increases (target = 1) and decreases (target = 0). This information is valuable in assessing the class balance and potential biases in the dataset, which can influence the model's training and prediction accuracy.

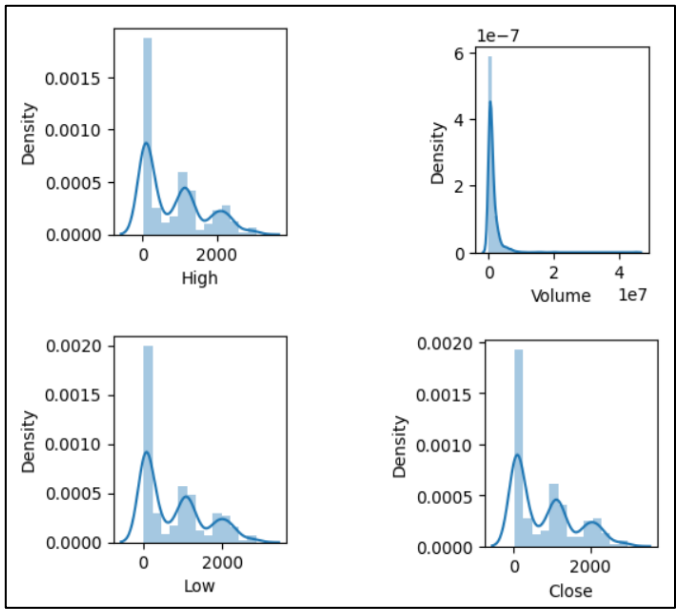

Chart -6: Subplots depicting distribution of values for the corresponding feature.

Chart 6 provides insights into the range, skewness, and concentration of values. It helps in understanding the characteristics of the dataset and identifying any potential outliers or patterns within the individual features.

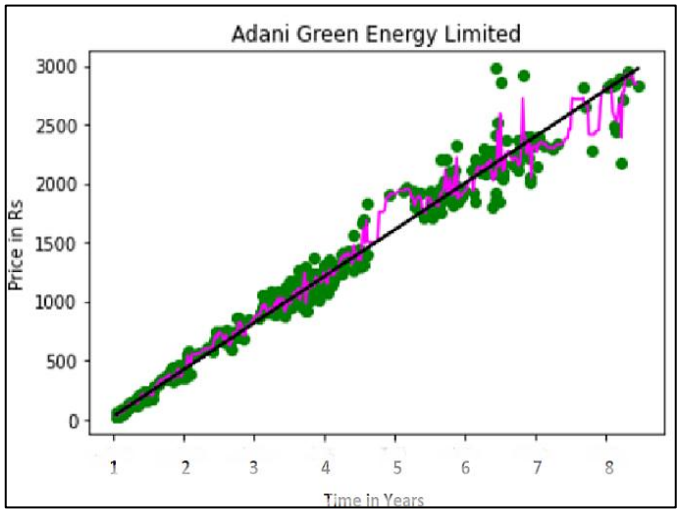

Chart -7: Graph depicting the Linear Regression plot and the Random Forest Regression plot together

3. CONCLUSIONS

The research presented in this paper unveils a novel and powerful approach for stock price prediction. Our methodology has demonstrated exceptional accuracy and user engagement, setting it apart from existing methods in several key aspects. Web scraping-based data extraction enables us to capture up-to-date market trends and make

accurate predictions. This approach provided a significant advantage over traditional approaches that rely on static or delayed data sources. The way consumers engage with and acquire stock market information has been revolutionized by this special combination of human-like interaction with innovative technology.

This distinct strategy gives stock market enthusiasts and investors a cutting-edge tool for making judgments and navigating the intricacies of the financial environment. This study broadens the field's limitations, paves the way for further developments, and alters how stock price prediction is conceptualized and carried out.

4. FUTURE SCOPE

4.1 Integration of Sentiment Analysis

We may learn a lot about public sentiment and market sentiment, which can affect stock price movements, by examining social media sentiment, news stories and other textual data associated with the stock. By taking into account the psychological and emotional variables that influence market behaviour, the integration of sentiment analysis algorithms with the current technique will enhance the accuracy of forecasts.

4.2 Integration with Smart Devices and Voice Assistants

Future research might concentrate on connecting the chatbot with these types of platforms to capitalize on the full advantage of the rising popularity of smart gadgets and voice assistants. The user experience would be fluid and pragmatic if consumers could get stock market forecasts and interact with the chatbot using voice commands and smart devices.

ABBREVIATIONS

AI - Artificial Intelligence

NLP - Natural Language Processing

CART - Classification And Regression Tree

XGBoost - Extreme Gradient Boosting

NLTK - Natural Language Toolkit

TF-IDF - Term Frequency-Inverse Document Frequency

RBF - Radial Basis Functions

REFERENCES

[1] Wamba, Samuel Fosso, et al. "How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study." International journal of production economics 165 (2015): 234-246.

[2] Li, Xiujun, et al. "End-to-end task-completion neural dialogue systems." arXiv preprint arXiv:1703.01008 (2017).

[3] Ayanouz, Soufyane, Boudhir Anouar Abdelhakim, and Mohammed Benhmed. "A smart chatbot architecture based NLP and machine learning for health care assistance." Proceedings of the 3rd international conference on networking, information systems & security. 2020.

[4] Cristescu, Marian Pompiliu, et al. "Using market news sentiment analysis for stock market prediction." Mathematics 10.22 (2022): 4255.

[5] Bhuriya, Dinesh, et al. "Stock market predication using a linear regression." 2017 international conference of electronics, communication and aerospace technology (ICECA). Vol. 2. IEEE, 2017.

[6] Basak, Suryoday, et al. "Predicting the direction of stock market prices using tree-based classifiers." The North American Journal of Economics and Finance 47 (2019): 552-567.

[7] Amjad, Maaz, et al. "Prediction of pile bearing capacity using XGBoost algorithm: modeling and performance evaluation." Applied Sciences 12.4 (2022): 2126.

[8] Sharma, Nonita, and Akanksha Juneja. "Combining of random forest estimates using LSboost for stock market index prediction." 2017 2nd International conference for convergence in technology (I2CT). IEEE, 2017.

[9] Jidong, Li, and Zhang Ran. "Dynamic weighting multi factor stock selection strategy based on XGboost machine learning algorithm." 2018 IEEE International Conference of Safety Produce Informatization (IICSPI). IEEE, 2018.

[10] Browne, Michael W. "Cross-validation methods." Journal of mathematical psychology 44.1 (2000): 108-132.