2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1) | 978-1-6654-5627-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAIA57370.2023.10169745

Survey of Stock Market Price Prediction Trends using Machine Learning Techniques

Paul Akash Gunturu
Dept of Computer Science Engineering,
IIIT Naya Raipur
Raipur, India
paul20100@iiitnr.edu.in

Rony Joseph
Dept of Computer Science Engineering,
IIIT Naya Raipur
Raipur, India
rony20100@iiitnr.edu.in

Emany Sri Revant
Department of Computer Science Engineering,
IIIT Naya Raipur
Raipur, India
amitsharma812010@gmail.com

Shailesh Khapre
Dept of Computer Science Engineering,
IIIT Naya Raipur
Raipur, India
shailesh@iiitnr.edu.in

***Abstract*— Investing in the stock market is an essential aspect of the financial sector. However, the task of identifying lucrative stocks is a challenging one that requires careful analysis. This study aims to address this challenge by comparing various Machine Learning and Deep Learning techniques for predicting stock trends. The research evaluates and compares different models, including Long Short-Term Memory (LSTM), Prophet (Automated Forecasting Procedure), Random Decision Forest, Auto-ARIMA, k-Nearest Neighbors (KNN), Linear Regression, and Moving Average techniques like SMA and EMA. Furthermore, a new hybrid model is proposed, which outperforms existing models in terms of accuracy. The models are trained and tested on a historical dataset of stocks from different industrial sectors and evaluated based on various performance metrics. The study provides insights into the accuracy of different prediction models and can help investors, traders, and financial analysts make informed investment decisions. Additionally, the findings of this research work can serve as a benchmark for future research on stock market prediction.**

Keywords*— *Stock Market, Machine Learning, Stock Price Predictions, Neural Networks

I. INTRODUCTION

In the fast-paced modern era of continuously evolving economies, it is crucial to have an in-depth understanding of financial investments, particularly in the realm of stock markets. The complexity of predicting the stock market has always posed a significant challenge for financial experts and statisticians. Stock traders are often interested in buying and selling the right stocks to maximize their profits. As a result, the trend of stock market prediction has become an ever-growing topic of research, with more and more investors, including young people, attempting to use analytics and technical analysis to gauge the market. Fundamental analysis and technical analysis can be broadly divided into two categories for stock market forecasting. Fundamental analysis considers several variables, including the company's overall financial health and the market's perception of a given sector or industry.

Financial analysts have long relied on technical analysis to predict the future values of stocks, using historical charts of stock patterns and knowledge of a company's financials. Today, technical analysis has advanced with the help of Data scientists and computer scientists are constantly creating and improving new statistical, machine-learning, deep learning, artificial intelligence, and AI approaches. These tools aid in stock price forecasting with little departure from current prices. In response to this, many organizations have invested in researching and developing stock-market predictors to offer customers stocks with a higher likelihood of gaining momentum.

The task of predicting the stock market remains an unresolved problem that presents significant challenges, despite the continuous efforts of various scientists and organizations. One of the main obstacles in developing a reliable stock market prediction model is the difficulty in accurately predicting the investment psychology of investors and market behavior influenced by sentimental investments. Moreover, the volatility of the stock market, caused by various factors such as inflation and government policies, adds to the challenge of creating an effective prediction model. Thus, the stock market prediction model offers a vast area for research and development.

The main aim of this research is to compare various prediction models using historical stock prices from different sectors. The study focuses on key parameters such as Day's High, Day's Low, Opening Stock Price, and Closing Stock Price, which have proven to be effective in predicting stock prices. The analysis focuses on deep learning techniques such as Long Short-Term Memory (LSTM), an artificial Recurrent Neural Network (RNN) that is well-suited for stock time series forecasting. Additionally, the study examines the highly complex forecasting model called Prophet, which was created by Facebook. Furthermore, the Random Decision Forest algorithm, also known as the Random Forest Algorithm, is a supervised learning algorithm that is primarily used for forecasting data using data categorization and regression techniques.

In this study, various techniques were employed to predict stock prices. One such technique is Auto ARIMA, It forecasts anticipated values using time series data. The k-Nearest Neighbors (kNN) algorithm, linear regression, and moving average methods including the simple moving average (SMA) and exponential moving average (EMA) were also utilised as supervised machine learning approaches. These methods allow for a more accurate prediction of stock prices based on historical data.

978-1-6654-5627-2/23/$31.00 ©2023 IEEE

The investigation carried out in this study delves into multiple Boosting Methods, a category of tree-based algorithms that employ time-series data to deliver projected stock price values. Adaboost and XGBoost are two prominent machine learning techniques used in this research.

The research focuses on comparing different Artificial Neural Networks and Neural Networks techniques to enhance the accuracy and precision of stock market forecasting. It is crucial to keep in mind that predicting the stock market is a complex process that involves considering various parameters for more precise outcomes.

II. Literature Survey

Reference [1] Min and Lee conducted a study to evaluate the effectiveness of different approaches in predicting insolvency. The methods they explored included multiple discriminant analysis, logistic regression analysis, Support Vector Machine, and three-layer fully connected back-propagation neural networks. According to Lee's findings, Support Vector Machines proved to be the most effective approach for assessing a company's credit rating. The study considered various financial indicators, such as interest coverage ratio, ordinary income to total assets, net income to stakeholders' equity, and current liabilities ratio, among others. The accuracy rate achieved was 60 percent. In predicting business credit ratings for the US and Taiwan markets using neural networks, the accuracy rate varied from 75 percent.

Reference [2] In a study conducted in 2015, Patel, Shah, Thakkar, and Kotecha explored the effectiveness of different input methodologies for predicting the stock prices of four companies listed on the Indian Stock Exchange (NSE). The study employed various prediction models such as SVM, random forest, naive bayes, and ANN. The first input methodology utilized trend deterministic data to represent technical features, while the second method used trading data to calculate ten technical measures like opening and closing prices. The accuracy of each prediction model was evaluated for both input methods.

Reference [3] In their study, Dai, Wu, and Lu (2012) proposed a time series prediction model for Asian stock markets that incorporates neural networks with a distinct feature extraction method called nonlinear independent component analysis (NLICA). This method is utilized to distinguish discrete sources from nonlinear mixed data in the absence of significant data mixing mechanisms. By transforming the existing time series data in the input space using NLICA, the authors aimed to create a feature space of independent variable components that captures the essence of the original data. The proposed model is designed to effectively capture complex nonlinear patterns in stock market data, potentially enhancing the accuracy of stock market predictions.

Reference [4] According to Guresen et al. (2011), artificial neural networks (ANN) are considered a promising approach for modeling the stock market because of their adaptability to market volatility and the absence of fixed formulas. ANN has the capability to learn from past data and interpolate and extrapolate based on it. Prior to problem-solving, the network must go through a learning phase during which it extracts patterns and creates a customized representation of the problem. In 1988, White used an ANN-based model to test Fama's efficient market hypothesis by analyzing daily returns on IBM stock. The results of this study were not particularly accurate, but it demonstrated the possibility of using ANN for such an analysis. Since then, numerous researchers have been working to develop more reliable forecasting models for the stock market.

III. Methodology

A. Linear Regression

This research utilizes supervised machine learning methods for statistical predictive analysis. Our approach involves employing a regression model that considers five parameters, of which four are independent variables, including Open Price, High Price, Low Price, and Average Price. The dependent variable we aim to predict is Last Price. To assess the model's performance, we split the dataset into training and testing subsets. The model is trained using the training data and then used to predict the Last Price of the testing data.

B. Auto Arima

In this study, a statistical technique called regression is utilized to predict future trends based on past time-series data. The ARIMA model is found to be suitable for time-series forecasting. The stationarity of the data is assessed using Augmented Dickey-Fuller, where a p-value greater than 0.05 indicates non-stationarity. The Auto-Arima model is used to determine the order of p, q, and d for the data. The model is then trained with the training data and tested against the testing data. The mean squared error is calculated to evaluate the error between the predicted and actual data. The model could generate a forecast if the mean squared error differs from the mean of the testing data. Closing price predictions are then made, and the model's accuracy is verified using MAPE, where an accuracy of less than 3% indicates high accuracy at
98.46%.

C. K-Nearest Neighbour

The proposed approach for making predictions is based on the nearest neighbours concept, which requires defining the number of neighbours to consider. This number is denoted as "k" and can be adjusted to meet the user's specific requirements. Depending on whether the algorithm is used for classification or regression, its behaviour differs after identifying the neighbours. In classification tasks, the algorithm uses a simple majority voting strategy to assign a label to the new data point. Conversely, for regression tasks, the algorithm computes the mean value of the neighbouring points to make the prediction.

D. Random Forest

The random forest technique utilizes a collection of decision trees, which are trained on different subsets of data, to make predictions. Each tree is designed to split nodes based on a finite set of attributes, and during training, random feature subsets are used to increase diversity. There are two key components to the random forest implementation: the random selection of data points for training each tree and the random selection of feature subsets for node separation. To generate its final predictions,

the algorithm takes the average of the projections made by each tree.

E. FB-Prophet

The Prophet method is a time series forecasting technique that utilizes an additive model to capture non-linear trends with seasonality, which include effects based on holidays, annual, weekly, and daily patterns. For best results, the Prophet method should be trained on historical data that spans multiple seasons and exhibits periodicity. The model is formulated as $y(t) = g(t) + s(t) + h(t) + e(t)$, where g(t) represents the trend estimated using a piecewise linear model. The seasonal changes are captured by s(t) occurring on a weekly, monthly, and yearly basis, while h(t) reflects the impact of holidays on the time series. Additionally, the model is computationally efficient and can handle outliers and missing data with ease. There is no need for data pre-processing, even when working with large datasets. The error term is represented by e(t).

F. LSTM

The long short term memory (LSTM) model is a variation of recurrent neural networks (RNNs) that enhances their memory capacity. RNNs have a short-term memory that can use previously processed data, making them useful for sequence modelling problems. However, they can suffer from the vanishing gradient problem when the same parameters are used repeatedly in RNN blocks. To overcome this issue, we introduce distinct parameters at each time step while maintaining a constant number of learnable parameters. We employ gated RNN cells such as the LSTM and GRU, which store internal variables called gates. The value of each gate depends on the information from each time step, including previous states, and is multiplied by appropriate factors to make a difference. We collect data in a time-series format to track changes over time. In this research, we utilize LSTM on a stock dataset to create a dependable automated forecasting framework for short-term stock price movements. A three-step procedure is used by LSTMs.

- In the first stage of LSTM, a sigmoid function is used to decide whether the cell's data should be included in the time step.
- The sigmoid function takes into account the previous state (ht-1) and the current input text to make this decision.
- The second layer of LSTM consists of two functions: sigmoid and tanh. The sigmoid function sets the acceptable values to 0 or 1, while the tanh function assigns weights to input values between -1 and 1 based on their importance.
- The final decision-making process involves creating a sigmoid layer to determine which components of the cell state should be output.
- The sigmoid gate output is then multiplied by the cell state after being sent through the tanh function to adjust the values between -1 and 1.

The primary goal of our automated prediction system is to offer a dependable framework for short-term forecasting of the stock market's price movement. Our system collects real-time data using web scraping techniques.

G. Hydrid Model

Our proposed approach for predicting stock prices involves combining Random Forest and Linear Regression, which has been statistically analyzed and found to be accurate. The model architecture is illustrated in the flow chart, depicting our ensemble empirical mode decomposition and linear regressionbased RF-LR hybrid approach. The process begins with the division of stock index sequences into decision trees using Random Forest. Linear Regression then utilizes the predicted outcomes of each sub-sequence as input to make predictions. Finally, the RF-LR approach is used to obtain the forecast result for the original stock index. The RF-LR prediction method consists of three modules: the fusion module, the module for predicting outcomes using linear regression, and the module for predicting outcomes using an ensemble of random forests. Three stages make up our suggested hybrid RF-LR prediction approach: input data, model prediction, and statistical model.

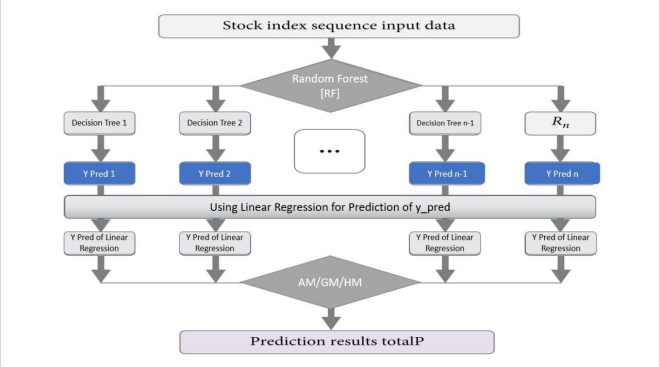

Fig. 1. A conceptual overview of the proposed solution.

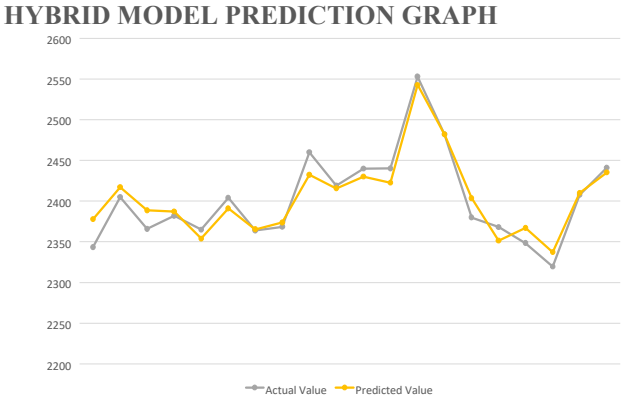

Fig. 2. Actual stock price's resemblance to the hybrid model's anticipated stock price. (NTPC Stock).

The hybrid RF-LR prediction method involves the following steps:

- The first step is to collect real-time stock market timeseries data from the NSE website and preprocess it to meet the requirements of ensemble decomposition, resulting in the creation of input data X for the hybrid prediction method.
- TThe second step involves using the Random Forest approach to divide the input data X into smaller

ensembles of decision trees. The average of all decision tree predictions is calculated to obtain RF y prediction.

- For each subsequence, multiple independent Linear Regression models (LRk) are constructed and trained, and n prediction values (LR predictionk) for the stock index time series are obtained using the results of the random forest as y actual for validation.
- In the third stage, the output of the random forest and multivariate linear regressions is used to obtain y pred, which is then processed using statistical means such as AM/GM/HM to select the best statistical standard.
- Finally, the predictions of random forest and linear regression are combined to generate a more accurate prediction.

IV. Results and Experimental Analysis

The Smart-Stock Prediction system is a machine learning-based tool developed using nsepy and Django, both Python tools. The nsepy automates machine learning models, while the latter deploys high-computational and machine-intensive algorithms with minimal load time. The system uses five major machine learning models, namely Linear Regression, AutoArima, to forecast the stock price's most recent price, K-Nearest Neighbor, Random Forest, Prophet, and Long Short Term Memory (LSTM) were used in addition to a native Support Vector Machine (SVM). The system includes a web application that meets the needs of the user. needs by providing detailed information on each model, accuracy charts, and predicted prices. It also offers user customization features and customer-related content to enhance the user experience.

Except for SVM, which exhibited dismal results with accuracy or R-Squared Value dipping as low as 0.52 in most equities, machine learning models Random Forest and Linear Regression have shown favourable outcomes. A new machine learning algorithm with an accuracy of 0.95 to 0.97 was developed by the system designers to address this. It mixes Linear Regression and Random Forest. The real value of the best-fit model is within ten times the final projected price.

The system aims to improve stock price prediction by incorporating cross-domain machine learning models. The study will explore time-series forecasting machine learning models and combination models that utilize arithmetic or geometric mean operations to represent outcomes. This approach may produce promising results. The research team also aims to develop a user-friendly web application that provides comprehensive information on the machine learning models used and their accuracy.

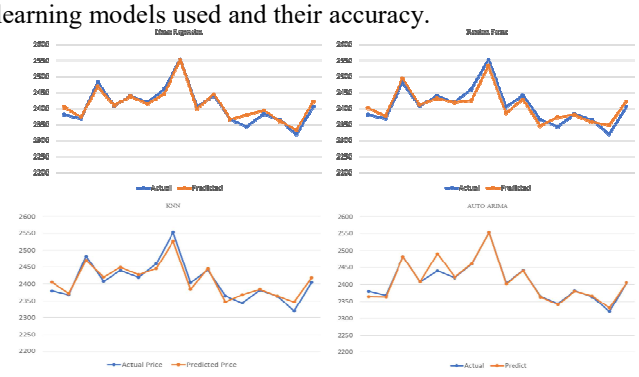

Fig. 3. Closeness of the actual Stock Price with the predicted Stock Price of Different Machine learning Models (NTPC Stock) (a) Linear Regression, (b) Random Forest, (c) K-Nearest Neighbour, (d)Auto Arima

Fig. 4. Stock overview.

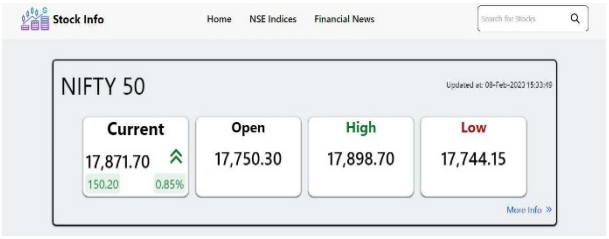

Fig. 5. Dashboard landing.

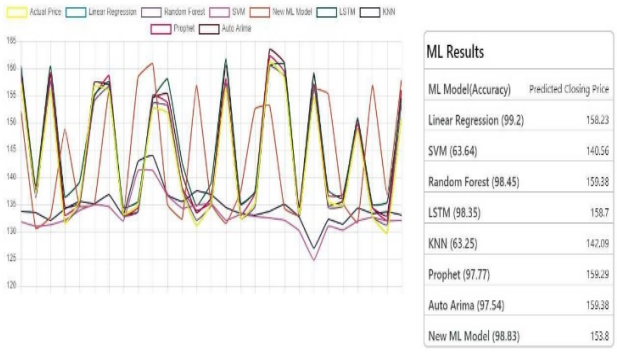

Fig. 6. Automated ML prediction results.

V. Conclusion

This web application will feature user customization and customer-related content to enhance the user experience. The research seeks to develop a robust and accurate stock prediction system that offers valuable insights to investors and traders. With the increasing demand for stock prediction systems, the study aims to develop an effective tool that predicts stock prices accurately while providing a user-friendly interface. Overall, the Smart-Stock Prediction system is a reliable and innovative tool that leverages machine learning models to provide valuable insights into the stock market.

REFERENCES

[1] N.M. Masoud, “The Impact of Stock Market Performance Upon Economic Growth”, *International Journal of Economics and Financial Issues*, vol 3, no. 4, pp.788-798, 2013.

[2] A. Murkute and T. Sarode, “Forecasting Market Price of Stock Using Artificial Neural Network”, International Journal of Computer Applications, vol. 124, no. 12, pp.11-15, 2015.

[3] Jung Hur, Manoj Raj, and Y. E. Riyanto, “Finance and Trade: A Cross Country Empirical Analysis on The Impact of Financial Development and Asset Tangibility on International Trade,” *World Development*, vol.34, no. 10, pp.1728-1741, 2006.

[4] L. Li, Y. Wu, Y. Ou, Q. Li, Y. Zhou and D. Chen, "Research on Machine Learning Algorithms and Feature Extraction for Time Series," in *Proc IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC),* Montreal, QC, Canada, 2017, pp. 1-5

[5] G.A. Seber and A.J. Lee, Linear Regression Analysis. John Wiley & Sons, 2003.

[6] T. T. L. Chong, and W. K. Ng, “Technical Analysis and the London Stock Exchange: Testing the MACD and RSI Rules using the FT30,” *Applied Economics Letters*, vol. 15, no. 14, pp.1111-1114, 2008.

[7] M. Obthong, N. Tantisantiwong, W. Jeamwatthanachai, and Gary Wills, “A Survey on Machine Learning for Stock Price Prediction: Algorithms and Techniques”, in *Proc 2nd International Conference on Finance, Economics, Management and IT Business,* Vienna House Diplomat Prague, Prague, Czech Republic, May 2020, pp. 63-71.