Stock Market Prediction Using ML Algorithms

Shivam Singhal
SRM Institute of Science and Technology

Manthan Solanki
SRM Institute of Science and Technology

S. Sharanya
SRM Institute of Science and Technology

Abstract – **The stock market has been a topic of great deliberation due to its diverse and convoluted nature. Today's financial investors are plagued by sudden and notable fluctuations in the market. They cannot easily comprehend as to which stocks they should buy or sell in order to get profitable outcomes. However, with rapid advancements in machine learning, stock market prediction has become plausible. This paper proposes a stock price prediction system that utilizes an ensemble model coupled with a separate LSTM model to make predictions. The ensemble model makes use of Random Forest (RF), K-Nearest Neighbors (KNN), and Gradient Boosting (GB) classifiers to determine whether an investor should buy or sell stocks on a particular day. A separate LSTM model analyzes the historical stock data to predict the closing stock prices in the future. The combined model assists the investors to make the buy/sell call on a particular day with an approximation of the closing prices for better and safer investments.**

Keywords: - stock market prediction, Random Forest, K-Nearest Neighbors, Gradient Boosting, ensemble model, Long Short Term Memory.

I. INTRODUCTION

1.1. Background

Investing in the stock market has been a lucrative temptation for both novice and expert investors alike for the past few decades. However, its dynamic and complex nature makes it intricately perplexing for investors to make the right choice for remunerative trading. Such a predicament divides the market experts on the possibility of making calculated predictions for the right investments at the right times. Some believe that as per the efficient-market hypothesis theory, the stock market reacts by assimilating newly available information. Therefore, it is not possible to make accurate predictions without possessing any prior future information of the stocks. However, other analysts argue that even though movements might seem random, they actually are correlated, and several statistical indicators can help establish a pattern. Based on historical stock market data, some trends can be discerned about the behaviour of stocks. This can be used to make close to precise predictions. Accurate predictions using technology provide investors with an opportunity to make steady financial gains. It also assists researchers in determining how different statistical indicators together can be used to improve accuracy. Machine Learning algorithms find its application in almost all the fields right from failure prediction in machines till forecasting economic growth [8]. This work deploys a hybrid model that integrates the prowess of Long Short Term Memory (LSTM) and ensemble model to forecast the trends in stock market.

1.2 Definitions

Relative Strength Index (RSI): It is a momentum oscillator that measures the magnitude of recent price movements. It oscillates between 0 and 100 and examines the overvalued (above 70) or undervalued (below 30) conditions of the stock prices.

Moving Average Convergence Divergence (MACD): MACD [5] is a momentum oscillator that evaluates a trend in stock prices. It determines the relation between two trend-following indicators, the moving averages (MA), by subtracting the higher MA from the lower MA.

Stochastic Oscillator (STOCH): It [9] is a momentum oscillator that depicts the relative location of the closing price of a stock to its range of prices over a specified period. It is used to identify overbought and oversold trading signals.

Accumulation/Distribution Line (ADL): It is a cumulative volume-based indicator that assesses the money flow into and out of a stock. It determines whether the market trend is inclined towards accumulation or distribution and measures the strength of a trend.

Average True Range (ATR): It is an indicator used for measuring the price volatility of commodities. It also accounts for any gaps in the price movement.

Market Momentum (MOM): It is a market indicator that reflects the comparison between the current market price and the price 'n' periods ago.

Money Flow Index (MFI): It is a market indicator equivalent to a volume-weighted RSI. It examines overbought and oversold trade signals on the basis of both magnitude and volume of prices.

Rate of Change (ROC): It [5] is a simple momentum oscillator that computes the percentage change in price from

the current price to a price 'n' periods ago. The oscillator forms a graph that oscillates above (positive change) and below (negative change) the zero-line. Overbought and oversold zones can be adjusted as per the market conditions.

On Balance Flow (OBV): It is a momentum indicator that predicts stock price changes based on the flow of volume. It measures the buying and selling pressure by summing volume on up days and subtracting it on down days.

Commodity Channel Index (CCI): It is a momentum oscillator that evaluates price trends and overvalued/ undervalued conditions. It computes the current price level relative to that of the historical average price level.

Ease of Movement (EMV): It [10] is an indicator that quantifies the price-volume relationship to determine the ease at which prices move upwards or downwards.

Vortex Indicator (VI): It is an indicator comprising of two oscillator lines – an uptrend (VI+) line to capture positive trends and a downtrend (VI-) line to capture negative trends. It is used to examine continuations and changes in trends.

Random Forest (RF): It [6] is an ensemble machine learning technique based on the bagging method. RF combines multiple decision trees to provide the final output. The aggregated result of multiple uncorrelated decision trees delivers more accurate results than the individual constituent trees.

K-Nearest Neighbours (KNN): It is a supervised ML algorithm that classifies new data into different categories based on the similarity of available data. Thus, whenever new data arrives, it is placed into a category that is similar to itself.

Gradient Boosting (GBM): It is an ML boosting algorithm [7] that derives the result by ensembling multiple weak learners to form a strong learner. With regression trees as the weak learners, each subsequent tree in the series is built on the residual errors of the predecessor trees thereby, minimizing the loss function.

Voting classifier: It [3] [8] is a classification technique that utilizes an ensemble of multiple classifiers. It makes predictions based on their highest probability of the chosen class as the output. The first type of voting is hard voting, where the output is simply the mode of individual predictions of the constituent classifiers of an ensemble. The other type of voting is soft voting, where the class with the greatest sum of weighted probabilities is delivered as the output.

Long Short-Term Memory (LSTM): It [6] is a Recurrent Neural Network (RNN) that showcases the ability to demarcate between recent and relatively older examples. LSTM assigns the former with higher weights and the latter with lower weights while forgetting data that seem irrelevant for the next set of outputs. Hence, it can handle long sequences of data better than other RNNs, which can store only a short series of data in memory. Thus, it is a much more suited neural network for the prediction of time-series data as compared to others.

II. LITERATURE SURVEY

Title	Merits	Demerits
Deep learning with long short-term memory networks for financial market predictions, 2018 [1]	Provides efficient predictions for large-scale financial markets.	Requires more subtle patterns of LSTM neural networks.
Global stock market investment strategies based on financial network indicators using machine learning techniques, 2019 [2]	Network indicators provide better results for global markets.	Insufficient for measuring latent factors in complex financial markets.
Stock market index prediction using deep neural network ensemble, 2017 [3]	For index predictions, the relative errors of high and low are less than a percent.	The relative errors of high, low, and close predictions are higher when the market index fluctuates fiercely.
Predicting and Beating the Stock Market with Machine Learning and Technical Analysis, 2018 [4]	At 85% confidence level, ML outperformed technical analysis during up-market.	At 90% confidence level, technical analysis outperformed ML during down-market.

III. PROPOSED WORK

The system proposed in this paper emphasises on using an ensemble model to make accurate predictions. While an individual algorithm-based model might have higher accuracy, however, an ensemble model boosts the overall confidence and reliability of the system. The proposed system has the following characteristics:

A. Buy/sell decision for stocks

A combination of Random Forest classifier, K-Nearest Neighbours classifier, and Gradient Boosting classifier forms an ensemble model [3] which predicts whether an investor should buy/sell stocks on a particular day.

B. Closing price prediction

An LSTM model [12] predicts the closing prices of the stock using historical datasets. The general stock direction trends can also be extracted from the dataset to predict the future behaviour of the stocks.

IV. IMPLEMENTATION

4.1 Methodology

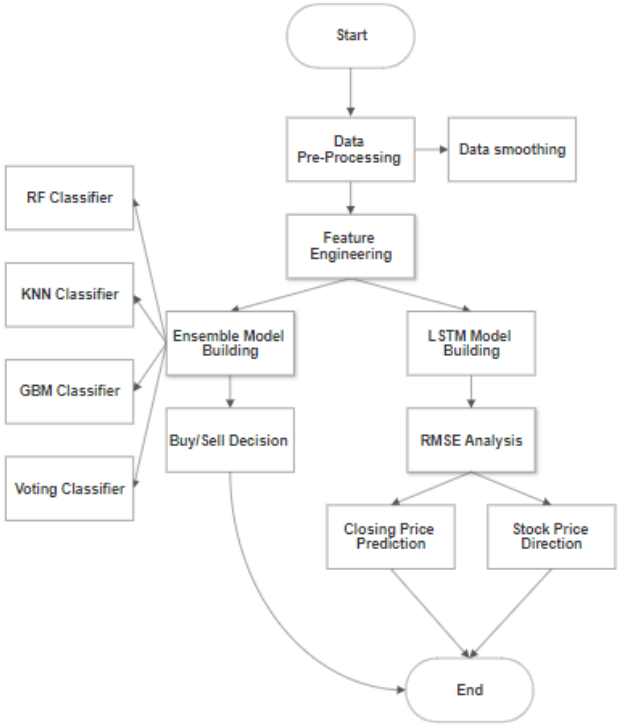

Fig 1. Flow of the proposed system

The complete flow of processes used to build and deploy the system is depicted in the flowchart (Fig. 1). The utilization of three different algorithms to build the ensemble model amalgamated with a separate LSTM model raises the overall accuracy and confidence of the system.

The proposed system is deployed under a series of modules, namely – data pre-processing, feature engineering, ensemble model building to predict the final call, and finally, the LSTM model building to predict the closing prices of stocks. The ensemble model outputs the final buy/sell call to be made for a stock by the investor. A sample testing dataset's results are represented in a confusion matrix that delineates the model's accuracy in terms of the buy/sell calls made. The final output of the LSTM model is a graphical representation of the predicted, and actual closing prices with the root mean squared error (RMSE) [12] analysis that depicts the accuracy of the model. The model also delivers an approximation of the future stock behaviour to further aid the investors in making the right decisions.

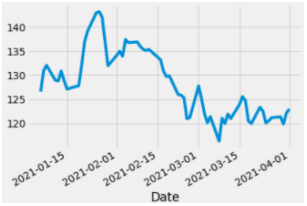

Fig 2. Data from 'close' column before smoothing

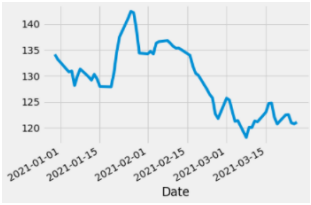

Fig 3. Data from 'close' column after smoothing

4.1.1. Data Processing and Feature Engineering

In the preliminary step, the dataset obtained is plotted (Fig. 2), and a list of technical indicators derived from the FINTA library is determined. Since stock prices are found to be affected by a myriad of different factors, the proposed model takes into account multiple factors to yield better predictions. The various technical indicators taken as input include RSI, MACD, STOCH, ADL, ATR, MOM, MFI, ROC, OBV, CCI, EMV, and VI.

Before deriving features from the indicators, the dataset is exponentially smoothed (Fig. 3) to remove the noise, which can become problematic for the model while predicting trends. The processed stock data is utilized for computing the various technical indicators. These technical indicators ensure that different aspects such as price changes, volume variations, price volatility, stock trends, and gaps in price movements, to name a few, are incorporated as essential parameters of the model. In parallel, the exponential moving averages (EMA) [11] at different average lengths along with a normalized volume value are calculated. The EMA is calculated using the following formula:

$$EMA_{today} = \alpha Price_{today} + (1 - \alpha) EMA_{today-1}$$

$$\alpha = \frac{2}{(n+1)}$$

Where,

- EMA_{today} = EMA of today
- $Price_{today}$ = Price of today
- $EMA_{today-1}$ = EMA of yesterday
- α = Smoothing factor
- n = Number of days

For example, a 21-day smoothing factor will be computed equal to $0.0909 \approx 9.09\%$.

The next step is to generate the truth values by observing the prices window rows ahead, examining whether the prices increased or decreased. An increase in prices yields the truth value as buy (1), whereas a decrease in prices yields sell (0) as the truth value. This step completes the data pre-processing and feature engineering modules.

4.1.2. Ensemble Model Building

The Random Forest, K-Nearest Neighbours, and Gradient Boosting algorithms are combined to form an ensemble model. A voting classifier is created which utilizes soft voting, i.e., it predicts the buy/sell decisions based on the average of the predicted results of all the constituent classifiers used in the ensemble model. In order to avoid a look-ahead bias, it is essential to perform cross-validation. By iterating over the data with multiple evenly-sized chunks, the data is partitioned. The partitioned data is bifurcated into testing and training data. The look-ahead bias can be easily avoided by not shuffling or randomizing the data in the train and test split function. For the last step, the models are incorporated for the cross-validation, and the final results are produced.

4.1.3. LSTM Model Building

A simple buy/sell call must not be the only basis of judgement available for the investors. Any investor must have a rough estimate of the closing prices so as to be prompted to make the right call. Accordingly, the system utilizes an LSTM model further to predict the closing prices and the general stock direction.

Utilizing the same technical indicators, the date and close columns of the dataset are first filtered and subsequently normalized using a min-max scaler. The LSTM model is built and trained on the dataset. The Adam optimization algorithm [3] is used along with the mean squared error as the loss function while compiling the model. Once the model is built, a testing dataset is taken to test the model for predicting closing prices. Lastly, an RMSE analysis is done to check for the accuracy of the predicted prices, and the results are plotted graphically along with the actual prices.

V. RESULTS AND DISCUSSION

5.1 Ensemble Model

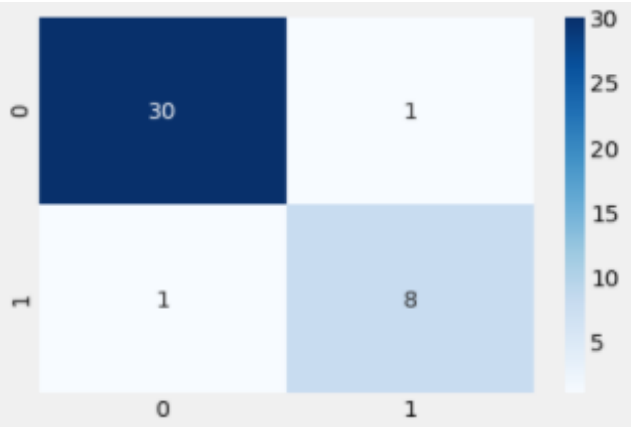

Fig 4. Confusion matrix

```
RF Accuracy = 71.54%
KNN Accuracy = 67.33%
GBM Accuracy = 70.26%
ENSEMBLE Accuracy = 71.59%
```

A sample testing data of 40-days is utilized to measure the accuracy of the ensemble model. The results are shown in the confusion matrix (Fig. 4). The accuracy obtained for the sample testing dataset is 95%. For two days, the predicted results were contradictory to the actual results, thereby rendering the misclassification rate as 5%.

5.2 LSTM Model

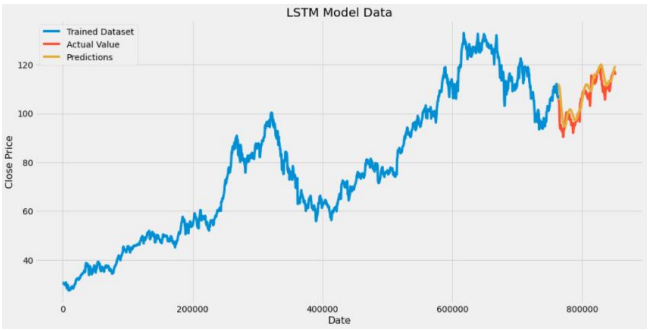

Fig 5. Closing Price Prediction using LSTM

The graph (Fig. 5) depicts the actual closing prices versus the predicted closing prices. It is observed that the predicted and actual values overlap at various points and intervals of time. Although there are intervals where the decline in actual prices is not precisely depicted by the predicted prices, the stock direction clearly indicates the eventual decline. To evaluate the difference, RMSE analysis is done, which yields a value of 4.12. The low value obtained depicts that the model is a good fit for making predictions.

The separation of the ensemble model and the LSTM model ensures that any inaccuracies in the predicted closing prices do not affect the buy/sell decision predicted by the ensemble model. Another major advantage of the system lies in the reinforced stock price trend prediction. Both the ensemble and LSTM models predict the price trend of stocks, thereby lessening the risks associated with sudden fluctuations in the market. By observing the results of both the models, the investor can easily discern the future behaviour of the stocks and make safer investments.

VI. CONCLUSION

The ensemble model predicts the final buy/sell call with an accuracy of 71.59%. Coupled with an LSTM model that predicts the closing prices of the stock, the entire system ensures reliability, confidence, and robustness. As a future prospect, sentimental analysis of investors can be taken into account to bolster the model's accuracy further.

REFERENCES

[1] Fischer, Thomas & Krauss, Christopher. (2017). Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research. 270. 10.1016/j.ejor.2017.11.054.

[2] Lee, Tae & Cho, Joon & Kwon, Deuk & Sohn, So. (2018). Global stock market investment strategies based on financial network indicators using machine learning techniques. Expert Systems with Applications. 117. 10.1016/j.eswa.2018.09.005.

[3] Yang, Bing & Gong, Zi-Jia & Yang, Wenqi. (2017). Stock market index prediction using deep neural network ensemble. 3882-3887. 10.23919/ChiCC.2017.8027964.

[4] Macchiarulo, A. (2018). Predicting and Beating the Stock Market with Machine Learning and Technical Analysis. The Journal of Internet Banking and Commerce, 23, 1-22.

[5] Aguirre, Alberto & Medina, Ricardo & Méndez, Néstor. (2020). Machine learning applied in the stock market through the Moving Average Convergence Divergence (MACD) indicator. Investment Management and Financial Innovations. 17. 44-60. 10.21511/imfi.17(4).2020.05.

[6] Pawar, Kriti & Jalem, Raj & Tiwari, Vivek. (2019). Stock Market Price Prediction Using LSTM RNN: Proceedings of ICETEAS 2018. 10.1007/978-981-13-2285-3_58.

[7] Momin, Faisal & Patel, Sunny & Shinde, Kuldeep & Sahane, Prof & Syed, Habeebullah Hussaini. (2020). Stock Market Prediction System Using Machine Learning Approach. SSRN Electronic Journal. 7. 190-194.

[8] S Sharanya, Revathi Venkataraman, G Murali (2020). Analysis Of Machine Learning Based Fault Diagnosis Approaches In Mechanical And Electrical Components. International Journal of Advanced Research in Engineering and Technology (IJARET). Volume 11, Issue 10, October 2020, pp.80-94,

[9] Abraham, Cerene & Elayidom, M.Sudheep & Santhanakrishnan, T.. (2019). Analysis and Design of an Efficient Temporal Data Mining Model for the Indian Stock Market: Proceedings of IEMIS 2018, Volume 2. 10.1007/978-981-13-1498-8_54.

[10] Hu, Hongping & Tang, Li & Zhang, Shuhua & Wang, Haiyan. (2018). Predicting the Direction of Stock Markets Using Optimized Neural Networks with Google Trends. Neurocomputing. 285. 10.1016/j.neucom.2018.01.038.

[11] Shastri, Malav & Roy, Sudipta & Mittal, Mamta. (2018). Stock Price Prediction using Artificial Neural Model: An Application of Big Data. ICST Transactions on Scalable Information Systems. 6. 156085. 10.4108/eai.19-12-2018.156085.

[12] Yu, Pengfei & Yan, Xuesong. (2020). Stock price prediction based on deep neural networks. Neural Computing and Applications. 32. 10.1007/s00521-019-04212-x.