**Artificial Intelligence and Machine Learning Methods for Predicting the Stock Market**

Subrat Chetia

Department of Computer Science

Pandit Deendayal Upadhyaya Adarsha Mahavidyalaya Dalgaon

Darrang-784116, Assam, India
subrat.chetia@gmail.com

**ABSTRACT**

It is highly challenging to anticipate anything when there is a non-linear relationship between the inputs and the outputs. One of the most difficult tasks for financial analysts is to predicting the stock market values because of the environments' inherent noise and their high volatility in relation to market movements. The goal of this article is to demonstrate the use artificial intelligence and machine learning techniques to address the issue of stock market prediction. The two basic analyses that can be applied to model the prediction of the stock market are technical analysis and fundamental analysis. Regression machine learning (ML) algorithms are used in the technical analysis approach to forecast the stock price trend at the end of a business day based on historical price data. Contrarily, in the fundamental analysis, the public attitude is classified using machine learning algorithms based on news and social media.

**Keywords:** SVM, KNN, ANN, XGB, SMA, RSI, MACD, OBV

**I. INTRODUCTION**

Stock markets are consistently a desirable investment choice for capital growth. In recent decades, as communication technology has advanced, the stock market has grown in popularity among individual investors. While the number of shareholders and companies on the stock markets increases year after year, many people look for a method to predict the direction of the stock market in the future. Since the 1990s, early studies have attempted to anticipate stock market movements using AI methodologies. Multiple research studies on the effectiveness of AI techniques in stock market prediction have been published. The massive daily volume of traded money in the stock markets drives researchers' interest in studying the issue of stock market forecasting.

There are two main approaches used in stock market analysis.:

* Technical Analysis and
* Fundamental Analysis.

In technical analysis, stockholders attempt to estimate the stock markets using past price data and looking into the indicators which are created based on this data, such as the EMA, RSI and MACD. A machine learning model can do the same thing. It can be trained to detect a logical relationship between a stock's closing price and financial indicators. This can lead to the development of a prediction model that forecasts the stock price at the end of a business day.

Fundamental analysis, on the other hand, makes an effort to estimate an actual stock value based on its owner company's financial records, such as the market cap, profit and loss statements, balance sheets and dividends paid. If the projected price is greater than the stock price, investors get a selling signal; if the estimated price is less than the stock price, investors get a holding or purchasing indication.

**II. THEORETICAL FRAMEWORK**

There are four main processes in the prediction of the stock market using machine learning algorithms.

* Dataset building
* Data engineering
* Model training, and
* Prediction.



**Figure 1:** ML Framework for Stock Market Prediction

**A. Dataset Generation**

Having access to a dataset is the first step in developing a Machine learning model. Several features of this dataset are used to train the machine learning model. The majority of the necessary data for predicting the stock market is available online, such as historical stock prices or public sentiment in the news.

The training procedure can be carried out with or without the use of a set of labelled data known as target values. When training procedure is conducted with a collection of labelled data, the process is known as supervised learning.; whereas, unsupervised learning does not require any target values and tries to find the unseen patterns in the training dataset.

**B. Data Engineering**

Before being used in model training, the data obtained from the proposed datasets must be pre-processed. In the model training stage, several indicators from technical analysis are used. The most significant ones are moving averages (simple and exponential), MACD, RSI and On-balance Volume (OBV) to construct the input characteristics of a machine learning training model.

* **Simple Moving Average (SMA)**

A simple moving average (SMA) computes the average of a given price range, usually closing prices, divided by the number of periods in that range. The mathematical formula for SMA is

$SMA=\frac{C\_{1}+C\_{2}+…+C\_{N}}{N}$ where

 $C\_{i}$ is the stock's closing price at period *i,*

*N* is the total number of period.

* **Exponential Moving Average (EMA)**

An exponential moving average (EMA) gives more weight and significance to recent data points. The formula for EMA is given below

$$EMA\_{i}=\left(C\_{i}×(\frac{SF}{1+N})\right)+EMA\_{(i-1)}×(1-(\frac{SF}{1+N}))$$

where

C*i* is the stock's closing price at period *i,*

*SF* is Smoothing Factor. The most common value is 2,

*N* is the total number of periods.



**Figure 2:** SMA and EMA

* **Moving Average Convergence Divergence (MACD)**

Moving average convergence divergence (MACD) is a momentum indicator that shows how two moving averages of a security's price relate to one another. The MACD is calculated by subtracting the exponential moving average of 26 periods from the exponential moving average of 12 periods.

*MACD=12 Period EMA – 26 Period EMA*

The signal line (9 period exponential moving average of the MACD) is then plotted on top of the MACD line, which can serve as a hint for buy and sell signals. When the MACD crosses above its signal line, traders can purchase the stock, and when it goes below, they can initiate a sell order.



**Figure 3:** MACD

* **Relative Strength Index (RSI)**

RSI evaluates overvalued or undervalued conditions in a security's price by measuring the speed and magnitude of recent price changes. It can also specify securities that are on the verge of a price correction or a trend reversal. It can indicate when to buy and sell. Historically, an RSI value of 70 or higher specifies an overbought condition. An RSI value of 30 or less specifies that the market is oversold. It can be calculated as

$$RSI=100-\left[\frac{100}{1+\frac{Average Gain}{Average Loss}}\right]$$

* **On-Balance Volume (OBV)**

On-balance volume (OBV) is a momentum indicator of momentum. OBV uses volume changes to forecast price movements. OBV reflects crowd sentiment and can forecast a bullish or bearish consequence. The formula for OBV is

$$OBV=OBV\_{prev}+\left\{\begin{array}{c}+Volume, if close>previous close\\0, if close=previous close\\-Volume, if close<previous close\end{array}\right.$$

where

*OBV* = Current OBV level

*OBVprev* = Previous OBV level

*Volume* = Latest trading volume amount

​

* **Fundamental Analysis**

Because fundamental indicators are unstructured, mining data for fundamental analysis is problematic. This data could be information from a company's financial report. It is obvious that changes in a company's financial report can have an instant influence on public viewpoint in the news and on social media. Thus, financial reports can be used to assess the impact of fundamental data on market trends. A machine learning model can use the internet to investigate news and social media in order to predict the impact of fundamental indicators on stock prices. This strategy is known as stock market sentiment analysis. In this analysis, the input data for training a model is largely unstructured and is imported into the model in text format. The goal of is to produce a binary value that indicates the positive or negative effect of financial reports on a particular stock.

**C. Machine Learning Model Training**

Many machine learning algorithms have been used in research studies to forecast stock markets. To address this problem, there are two main categories of models:

* Classification models, which attempt to support shareholders in the decision-making process of buying or selling a stock.
* Regression models, which try to forecast stock price activities such as the low, high or closing price of a stock.

According to study, more than 90% of the algorithms used in stock market prediction are based on classification models. The Decision Tree (DT), Support Vector Machine (SVM), Artificial Neural Networks (ANN), Logistic Regression (LR), Gaussian Naive Bayes (GNB), Bernoulli Naive Bayes (BNB), Random Forest (RF), k-Nearest Neighbour (KNN) and XGBoost (XGB) are the most common machine learning algorithms used to forecast stock markets.

Few studies, however, attempted to guess exact stock prices with the help of regression models. Linear regression and long short-term memory (LSTM) are used in regression problems.

* **The Decision Tree (DT)**

A decision tree algorithm executes a series of recursive actions before arriving at the end result, and when these actions are plotted on a screen, the pictorial representation look like a large tree, hence the name 'Decision Tree.'

* **Support Vector Machine (SVM)**

The SVM algorithm's aim is to find the decision boundary for classifying n-dimensional space so that new data points can be easily placed in the correct category in future.

* **Artificial Neural Networks (ANN)**

Artificial Neural Network (ANN) is a popular as well as relatively new technique for making financial market estimations that also integrates technical analysis. A set of threshold functions is comprised with ANN. These functions are trained on historical data and used to forecast the future by tying them together with adaptive weights.

* **Logistic Regression (LR)**

Logistic regression can be used to categorize a collection of independent variables into two or more mutually exclusive classes and can be used to predict the probability of good performing stocks by applying variable to logistic curves.

* **Gaussian Naive Bayes (GNB) and Bernoulli Naive Bayes (BNB)**

The Gaussian Naive Bayes and Bernoulli Naive Bayes are simple but very efficient supervised learning algorithms. Gaussian Naive Bayes comprises the prior and posterior probabilities of the dataset classes. Bernoulli Naive Bayes, on the other hand, is only applicable to data with binary-valued variables.

* **Random Forest (RF)**

The RF algorithm consists of a sequence of decision trees whose goal is to yield an uncorrelated cluster of trees whose prediction is more precise than any solitary tree in the cluster.

* **k-Nearest Neighbour (KNN)**

The KNN is a well-recognized classification algorithm that uses test data to determine how an unclassified point should be classified as. The Manhattan distance and Euclidean distance are the methods used in this algorithm to calculate the distance between the unclassified point and its similar points.

* **XGBoost**

XGBoost is a supervised learning algorithm that predicts an aimed variable correctly based on the approximation of simpler and weaker models. It is a widespread and open-source form of the gradient boosted trees algorithm.

* **Linear regression**

Linear regression is a useful measure for technical and quantitative analysis in financial markets that analyses two separate variables to define a single relationship. Stock prices plotted along a normal distribution (bell curve) can help traders determine when a stock is overbought or oversold.

* **Long Short Term Memory (LSTM)**

The Long Short Term Memory is a deep learning algorithm. The feedback connections in its architecture make it a recurrent network. It has an advantage over traditional neural networks because it can process the entire data sequence.

**D. Model Evaluation Metrics**

All prediction models need some evaluation metrics to examine their accuracy in the prediction procedure. For classification models, Confusion Matrix and Receiver Operator Characteristic (ROC) curve are available as evaluation metrices. Similarly, R-squared (R2), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE) are available as evaluation metrics for regression models.

* **Confusion Matrix**

The confusion matrix, also called as an error matrix, is a widespread measure for resolving classification problems. It is applicable to both binary and multiclass classification problems. The prototype for any binary confusion matrix combines the four types of results true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN) as well as the positive and negative classifications. The four results can be stated as follows in a 2x2 error matrix:

|  |  |
| --- | --- |
|  | **Actual value obtained by the experiment** |
| Positive | Negative |
| **Predicted Value** | Positive | True Positive (TP) | False Negative (FN) |
| Negative | False Positive (FP) | True Negative (TN) |

**Table 1:** Confusion Matrix

True Positive (TP) : properly identifies the existence of a condition or feature

True Negative (TN) : properly identifies the absence of a condition or feature

False Positive (FP) : incorrectly indicates the existence of a specific condition or attribute

False Negative (FN) : incorrectly indicates the absence of a specific condition or attribute

* **Receiver Operator Characteristic (ROC) curve**

The Receiver Operator Characteristic curve is constructed by comparing the true positive rate (TPR) to the false positive rate (FPR) at numerous threshold levels. The TPR is also termed as detection probability, recall or sensitivity. The FPR is referred as the probability of false alarm as well.

$$TPR= \frac{TP}{TP+FN} ;FPR= \frac{FP}{FP+TN} $$

* **R-squared (R2)**

The R Squared (R2) is a statistical measure which represents the proportion of the variance explained by an independent variable or variables in a regression model for a dependent variable. In investing, R-squared (R2) refers to the percentage of a fund's or security's movements that can be explained by movements in a benchmark index. An R-squared (R2) measure of 100% indicates that activities in the index (or the independent variable(s) of interest) fully explain all movements in the security (or other dependent variables). The following equation shows the formula for calculating the R Squared (R2) metric.

$$R^{2}=1-\frac{Unexplained Variation}{Total Variation}$$

* **Mean Absolute Percentage Error (MAPE)**

The mean absolute percentage error (MAPE), also known as the mean absolute percentage deviation (MAPD), is a metric used to assess the accuracy of a forecast system. It computes the average absolute percent error for each time period minus actual values divided by actual values to estimate this accuracy as a percentage. The formula for MAPE is

$$MAPE=\frac{1}{n}\sum\_{t=1}^{n}\left|\frac{A\_{t}-F\_{t}}{A\_{t}}\right|$$

Where *n* is the number of fitted points

*At* represents the actual value

*Ft* denotes the forcast value

* **Root Mean Square Error (RMSE)**

Root Mean Square Error (RMSE) is the standard deviation of the prediction errors. Prediction errors (or residuals) are a measure of how far the data points are from the regression line. The RMSE measures of how these residuals are evenly distributed. Root Mean Square Error is never negative, and a value of 0 (which almost never happens in practise) indicates a faultless fit to the data. A lower value is generally preferable to a higher one.

* **Mean Absolute Error (MAE)**

The Mean Absolute Error (MAE) is computed as the sum of the absolute differences among the predicted variables and target. As a result, it estimates the average magnitude of errors in a set of predictions without taking into account their directions. This metric's lower values indicate a better prediction model. The formula for calculating the Mean Absolute Error is

$$MAE=\frac{1}{n}\sum\_{i=1}^{n}\left|x\_{i}-x\right|$$

Where n represents the number of errors and

|xi-x| represents the absolute error

**E. Stock Market Prediction**

To predict the stock market, programming languages such as Python can be utilized to train Machine Learning models and forecast unpredicted data.  In this concern, the market estimation based on technical analysis is evaluated initially. Fundamental analysis is examined after the evaluation of technical analysis.

The dataset for developing a predictive model based on technical analysis is easily obtainable from the internet. It does, in fact, contain historical data for the all well known stocks. The dataset includes open, high, as well as low prices, as well as the moving average, MACD, and RSI. The closing price of a stock at the end of a trading day is the target. The most correlated attributes to the target are then chosen, and the repetitive features with a high correlation are combined. The training process consumes a large portion of the data, while the rest is used for validation and testing. The algorithm uses the training data to study how to forecast the target value during the training process. The ML model then estimates the prediction's performance in relation to the validation data. It can predict the unanticipated target values of the testing dataset to compare with the true target values. Finally, the evaluation metrics can be calculated using the predicted and actual closing price values.

The estimation of public sentiments using available machine learning algorithms does not yield encouraging results. The accuracy of various machine learning algorithms ranges between 60% to 75%.

**III. SUMMERY**

This article attempts to introduce various machine learning algorithms which can be used for predicting the stock market. Many techniques like ANN, SVM, K-means etc., are generally applied to guess the stock market. Until now, the accuracy of machine learning algorithms in predicting whether a stock should be bought, sold, or held is insufficient and no credible model has outperformed the stock market so far. Nonetheless, many research studies in this field use a hybrid prototype that combines fundamental and technical analysis in a single machine learning model to compensate for the deficiencies of specific algorithms. This may perhaps improve the prediction precision, suggesting an exciting topic for forthcoming research studies. Until now, it appears that artificial intelligence is incapable of accurately forecasting the stock market. Possibly in the near future, with the advancement of artificial intelligence and machine learning techniques and, in particular, computation power, a more accurate model of stock market prediction will be available.

**REFERENCE**

[1] Dattatray P. Gandhmal, K. Kumar, “Systematic analysis and review of stock market prediction techniques”, Computer Science Review 34 (2019) 100190, Elesvier

[2] Sohrab Mokhtari, Kang K Yen, Jin Liu, “Effectiveness of artificial intelligence in stock market prediction based on machine learning”, International Journal of Computer Applications (0975 - 8887) 2018

[3] G. S. Navale, Nishant Dudhwala, Kunal Jadhav, “Prediction of stock ,market using data mining and artificial intelligence”, International Journal of Computer Applications (0975 – 8887) Volume 134 – No.12, January 2016

[4] Prakash Ramani, Dr. P. D. Murarka, “Stock Market Prediction Using Artificial Neural Network” in International Journal of Advanced Research in Computer Science and Software Engineering. ISSN: 2277-128x, Volume 3, Issue 4, April 2013

[5] Zahid Iqbal, R. Ilyas, W. Shahzad, Z. Mahmood and J. Anjum, “ Efficient Machine Learning Techniques for Stock Market Prediction” in Int. Journal of Engineering Research and Applications, ISSN : 2248-9622, Vol. 3, Issue 6, Nov-Dec 2013, pp.855-867.

[6] Zhou, Z., Gao, M., Liu, Q., & Xiao, H. (2020), “Forecasting stock price movements with multiple data sources: Evidence from stock market in China”. Physica A: Statistical Mechanics and its Applications, 542, 123389.