**Artificial Intelligence and Machine Learning Techniques for Stock Market Prediction**

Subrat Chetia

Department of Computer Science

Pandit Deendayal Upadhyaya Adarsha Mahavidyalaya

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

**ABSTRACT**

Predicting anything is very hard where the relationship between inputs and outputs are non-linear in nature. The prediction of stock market values is one of a challenging task for financial analyst due to intrinsic noisy environments and large volatility with respect to the market trends. The goal of this article is to demonstrate the use data mining and artificial intelligence techniques to address the issue of stock market prediction. The stock market prediction can be modelled based on two principal analyses called technical analysis and fundamental analysis. In the technical analysis approach, the regression machine learning (ML) algorithms are employed to predict the stock price trend at the end of a business day based on the historical price data. In contrast, in the fundamental analysis, the classification ML algorithms are applied to classify the public sentiment based on news and social media.

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

**I. INTRODUCTION**

Stock markets are always an appealing investment option 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 in the stock markets grows year after year, many people try to find a way to predict a stock market's future trend. Early studies attempted to predict stock market movements using AI strategies since the 1990s. Many research studies on the performance of AI approaches in stock market prediction have been published. The enormous daily volume of traded money in the stock markets drives researchers' interest in studying the stock market prediction problem.

Typically, analysing the stock markets is based on two key strategies:

* Technical analysis and
* Fundamental analysis.

In the technical analysis, stockholders attempt to evaluate the stock markets using historical price data and investigating the generated indicators based on this data, such as the RSI and the MACD. A machine learning model can do the same thing. It can be trained to detect a logical relationship between financial indicators and the closing price of a stock. 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 attempt to determine an actual stock value based on its owner company's financial reports, such as the market cap or dividends. If the estimated price is greater than the stock price, stockholders receive a selling signal; if the estimated price is less than the stock price, stockholders receive a holding/buying signal.

**II. THEORETICAL FRAMEWORK**

The prediction of the stock market using Machine Learning tools consists of four major steps.:

* 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. This dataset contains some features that are used to train the ML model. The training procedure can be carried out with or without the use of a set of labelled data known as target values. 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.

If the training is based on a set of labelled data, the training procedure is called supervised learning; whereas, unsupervised learning does not need any target values and attempts to find the hidden patterns in the training dataset.

**B. Data Engineering**

The data obtained from the proposed datasets requires to be pre-processed before being exploited in model training. Technical analysis has several indicators applied in the model training step and the most significant ones are simple moving average (SMA), Exponential Moving Average (EMA), MACD, RSI and On-balance Volume (OBV) to build the input features of a machine learning training model.

* **Simple Moving Average (SMA)**

A simple moving average (SMA) computes the average of a selected range of prices, usually closing prices, by the number of periods in that range.

The mathematical formula for SMA is

where

is the closing price of the stock at period *i*.

*N* is the total number of period

* **Exponential Moving Average (EMA)**

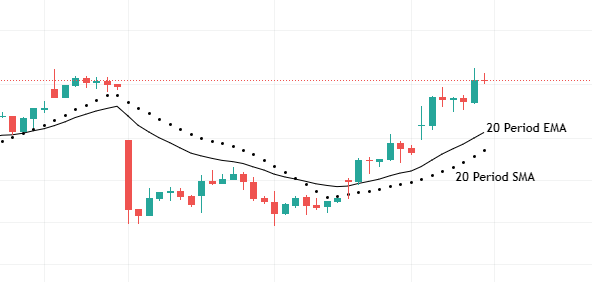
An exponential moving average (EMA) is a type of moving average (MA) that gives the most recent data points more weight and significance. The formula for EMA is given below

where

C*i* is the closing price of the stock at period *i*

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

*N* is the total number of periods



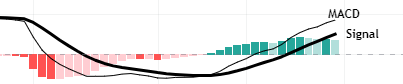
**Figure 2:** SMA and EMA

* **Moving average convergence divergence (MACD)**

Moving average convergence divergence (MACD) is a momentum indicator that demonstrates the relationship between two moving averages of a security's price. The MACD is calculated by subtracting the exponential moving average (EMA) of 26 periods from the EMA of 12 periods.

*MACD=12-Period EMA − 26-Period EMA*

The signal line ( a 9-day EMA of the MACD) is then plotted on top of the MACD line, which can act as a trigger for buy and sell signals. Traders can buy the stock when the MACD crosses above its signal line and sell it when it crosses below the signal line.



**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 indicate securities that are primed for a trend reversal or price correction. It can tell you when to buy and sell. Historically, an RSI reading of 70 or higher indicates an overbought condition. A reading of 30 or less indicates that the market is oversold. It can be calculated as

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

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

where

OBV = Current on-balance volume level

OBVprev = Previous on-balance volume level

volume = Latest trading volume amount

​

* **Fundamental Analysis**

Because fundamental indicators are unstructured, extracting data for fundamental analysis is difficult. 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 immediate impact 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 investigate news and social media through the use of the Internet to predict the impact of fundamental indicators on stock prices. This strategy is known as stock market sentiment analysis where 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 financial reports positive or negative impact on a specific stock.

**C. Machine Learning Model Training**

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

* Classification models, which attempt to assist investors in the decision-making process of buying, selling, or holding stock
* Regression models, which attempt to predict stock price movements such as the closing price of a stock.

According to research, over 90% of the algorithms used in stock market prediction are classification models. Few studies, however, attempted to predict exact stock prices using regression models. The Decision Tree (DT), Support Vector Machine (SVM), and Artificial Neural Networks (ANN) are the most popular Machine Learning algorithms used to forecast stock markets.

Classification strategies include Logistic Regression (LR), Gaussian Naive Bayes (GNB), Bernoulli Naive Bayes (BNB), Random Forest (RF), k-Nearest Neighbour (KNN), and XGBoost (XGB).

Linear regression and long short-term memory (LSTM) are used in regression problems.

* **The Decision Tree (DT)**

A decision tree algorithm performs a series of recursive actions before arriving at the end result, and when these actions are plotted on a screen, the visual resembles a large tree, hence the name 'Decision Tree.'

* **Support Vector Machine (SVM)**

The SVM algorithm's goal is to find the best line or decision boundary for classifying n-dimensional space  which we can easily place new data points in the correct category in the future.

* **Artificial Neural Networks (ANN)**

Artificial Neural Network (ANN) is a popular and relatively new method for making financial market predictions that also incorporates technical analysis. A set of threshold functions is included with ANN. These functions are trained on historical data and used to predict the future by connecting them with adaptive weights.

* **Logistic Regression (LR)**

By applying variable to logistic curves, logistic regression can be used to categorise a collection of independent variables into two or more mutually exclusive classes and can be used to predict the likelihood of good performing stocks.

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

The supervised learning algorithms Gaussian Naive Bayes and Bernoulli Naive Bayes are simple but very functional. The prior and posterior probabilities of the dataset classes are included in Gaussian Naive Bayes, whereas Bernoulli Naive Bayes only applies to data with binary-valued variables.

* **Random Forest (RF)**

The random forest algorithm consists of a series of decision trees whose goal is to produce an uncorrelated group of trees whose prediction is more accurate than any single tree in the group.

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

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

* **XGBoost**

XGBoost is a supervised learning algorithm for the accurate prediction of an aimed variable based on its simpler and weaker models estimation. It is a popular and open-source version of the gradient boosted trees algorithm.

**D. Model Evaluation Metrics**

To investigate their accuracy in the prediction procedure, all prediction models require some evaluation metrics.

For classification models, Confusion Matrix and Receiver Operator Characteristic (ROC) curve are available. Similarly, for regression models, R-squared, explanation variation, Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) are available as evaluation metrics.

* **Confusion Matrix**

The confusion matrix, also known as an error matrix, is a popular measure for solving classification problems. It can be used for both binary classification and multiclass classification problems. The template for any binary confusion matrix incorporates the four types of results positives, false negatives, false positives, and true negatives as well as the positive and negative classifications. The four outcomes can be expressed as follows in a 2x2 confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Actual value obtained by 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) : correctly identifies the presence of a condition or feature

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

False Positive (FP) : incorrectly indicates the presence 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 ROC curve is generated by plotting the true positive rate (TPR) versus the false positive rate (FPR) at various threshold levels. The true-positive rate is also known as sensitivity, recall, or detection probability. The false-positive rate is also referred to as the probability of false alarm.

* **R-squared (R2)**

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

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

The mean absolute percentage error (MAPE), also known as the mean absolute percentage deviation (MAPD), is a measure of a forecast system's accuracy. It calculates the average absolute percent error for each time period minus actual values divided by actual values to calculate this accuracy as a percentage. The formula for MAPE is

Where *n* is the number of fitted points

*At* is the actual value

*Ft* is the forcast value

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

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

* **Mean Absolute Error (MAE)**

The MAE is calculated as the sum of the absolute differences between the target and predicted variables. As a result, it evaluates 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 is

Where n is the number of errors and

|xi-x| is the absolute error

**E. Stock Market Prediction**

To predict the stock market, programming languages such as Python can be utilized to train ML models and forecast unpredicted data.  In this concern, the market prediction based on technical analysis is evaluated initially, and then the fundamental analysis is examined.

The dataset for developing a predictor model based on technical analysis is easily available on the "Yahoo Finance" website. It does, in fact, contain historical data for all well-known stocks. The dataset includes open, high, as well as low prices, as well as the moving average, MACD, and RSI. The close price, or the final price of a stock at the end of a business day, is the target. The most correlated attributes to the target are then chosen, and the repetitive features with a high correlation are merged. 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 learn how to predict the target value accessible to the algorithm during the training process. The model then evaluates the prediction's performance in relation to the validation data. Finally, 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.

Prediction of public sentiment using machine learning algorithms does not yield promising results. The accuracy of various machine learning algorithms ranges between 60% to 75%.

**III. SUMMERY**

Using ML algorithms, this chapter attempts to address the problem of stock market prediction. Many techniques like ANN, SVM, K-means etc., are generally used to predict the stock market. Until now, the accuracy of ML 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 model that combines technical and fundamental analysis in a single ML model to compensate for the shortcomings of the individual algorithms. This could improve prediction accuracy, implying an intriguing topic for future research. Based on this research, it appears that AI is not capable of accurately forecasting the stock market. Perhaps in the future, with AI advancement and, in particular, computation power, a more precise 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.