STOCK MARKET TRADING USING MACHINE LEARNING

Vrutik Shah
Department of Information Technology
K.J. Somaiya College of Engineering
1814056
vrutik.shah@somaiya.edu

Jash Shah
Department of Information Technology
K.J. Somaiya College of Engineering
1814054
jash12@somaiya.edu

Jill Shah
Department of Information Technology
K.J. Somaiya College of Engineering
1814055
jill25@somaiya.edu

Prof. Nilkamal More
Department of Information Technology
K.J. Somaiya College of Engineering)
Machine Learning Professor
neelkamalsurve@somaiya.edu

***Abstract*—Trading is the most important thing in the finance world. Money management is arguably the most crucial and underappreciated aspect of a successful trading career. The stock market, often known as the stock exchange, is one of the most intricate and sophisticated ways to do business. Small businesses, banking industry and brokerage firms, all rely on this body to generate money and split risks; it's a sophisticated mechanism. It is critical to correctly anticipate the stock market in order to maximise profits. As these data exist in enormous numbers and are very complicated, there is always a need for a more efficient machine learning model for daily forecasts by utilizing open-source libraries and algorithms. Stock price predictions can considerably assist consumers in determining where and how to invest in order to reduce the danger of losing money. This survey paper will focus on various machine learning Regression algorithms like Linear Regression, Polynomial Regression, Support Vector Machine Regression, Decision Tree Regression, Random Forest Regression and Classification algorithms like K-Nearest Neighbors, Naive Bayes, Logistic Regression, Decision Tree Classification, Random Forest Classification and Support Vector Machine. It also discusses the hybrid model by appending two algorithms and gives the final comparison of all the algorithms based on the accuracy.**

***Keywords*— Linear Regression, Hybrid model, Machine learning, Random Forest Classification, Stock market, Stock price prediction, Stock forecasting, SVM (Support Vector Machine).**

I. Introduction

Stock market is one of the oldest means of trading stocks, earning money and making investments from businesses who sell a piece of themselves on this market. A share market or a stock market is a place where the sellers and customers of shares / stocks are gathered. Shares are usually the inventory alternative for most huge agencies. This results in more liquid inventory and hence it interests most of the investors. Over the country, trading is done for most of the unique stocks by the provider. When there is a downfall of these inventory markets, it is referred to as Bear markets. Whereas the rise in the same is known as the Bull market. Different terminologies can be used to determine the stock level or the trend, etc. like the trend indicators, stochastic oscillators, commodity channel index, relative energy index, etc.

Machine learning is used to train a model and analyze and predict the solutions for the problems faced. There are two types of learning algorithms. Supervised and unsupervised learning algorithms. The dataset is formed using the data and then preprocessing is done on the dataset. Then the dataset is split into training and testing data. Training the model is the first thing done which helps to analyze and provide solutions to the problem. Then the algorithm is applied on the testing dataset to do the finalization. There are many algorithms in machine learning. Each of them has its own purpose and the need of the problem is used for identifying the algorithm.

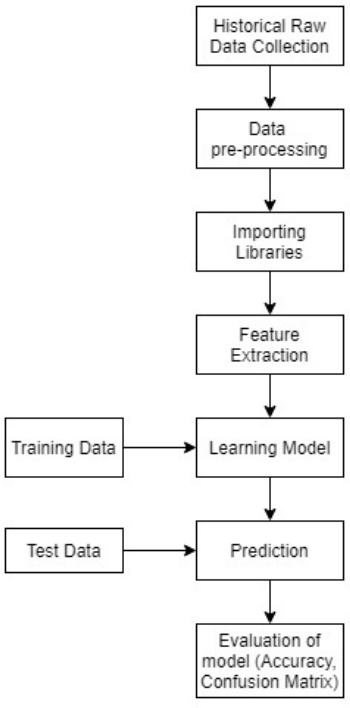

Fig. 1. Steps for developing a Machine Learning model

These techniques have the ability to uncover previously unseen patterns and insights, which may then be utilised to generate impeccably precise predictions. In machine learning, we utilize features to train the classifier, which then predicts the value of the label with a specified accuracy that can be validated throughout the classifier's training and testing. To make a classifier accurate, there is a need to choose the correct features and have enough data to train it. The quantity of data supplied to the classifier and the attributes chosen have a direct relationship with the accuracy of the classifier.

Now, stock market forecasting is a technique for predicting the direction of the stock market and estimating the value of a stock or other financial asset in the future. 90 percent of the world's data has been generated in the last few years as a result of the daily generation of 2.5 quintillion bytes of data. The financial market generates a significant amount of data. Recognizing a pattern and then devising an ideal method for making judgments is extremely tough for a trader. One of the most difficult things to accomplish is to predict how the stock market will perform. It is critical to develop a system that will function with optimum precision and take into account all key elements that may impact the outcome. Because there are numerous causes for a share's profit or loss, the stock market can do well even while the economy is in decline.

II. Methodology

A. Paper I - Stock Market Analysis using Supervised Machine Learning [1]

In this paper, the data for the program is collected from www.quandl.com, a leading dataset provider platform.
The following are the features of this dataset: Open – Stock Opening Price, High – Stock Highest Price, Low – Stock Lowest Price, Close – Stock Closing Price and Volume – Total Times Traded.
The attribute "Close" is the predicting variable depending on the above features. The features "Adj. Open, Adj. Volume Adj. High, Adj. Low and, Adj. Close are used because these numbers have already been processed and are free of frequent data gathering problems.
The set of features that will be used for the classifier are derived from the graphing parameters used in OHLCV graph.
They are:

- **Adj.Close** - the market's opening price is determined for the next day.
- **HL_PCT** – Defined as the difference between Adj.High and Adj.Low divided by Adj.Close and multiplied by 100. The High-Low feature is important since it aids in the formulation of the OHLCV graph's shape.
- **PCT_Change** - Defined as the difference between Adj.Close and Adj.Open divided by Adj.Open and multiplied by 100. Additionally, it assists us in reducing the quantity of redundant features. They're both important in our prediction model, and they help us decrease the amount of duplicate characteristics.
- **Adj.Volume** - This is a critical decision element since the volume traded has the greatest influence on future stock prices.

The characteristics retrieved are highly unique to the subject at hand, and they will undoubtedly differ from one subject to the next.

Now we'll move on to the stage of Training and Testing

To train our model, we'll use the Python libraries Scikit-learn, SciPy, and Matplotlib, and then train it using the labels and features we retrieved before testing it with the same data.

We'll use the most basic classifier, **Linear Regression**, which is described in the Sklearn package. The Scikit-learn package is an open-source learning framework. Linear regression is a widely used data analysis and forecasting tool. It simply employs key characteristics to anticipate relationships between variables based on their inter dependencies with other variables. It recalls the combination of attributes and the label that corresponds to them, in our instance the stock price a few days later. Then it continues on to figuring out what pattern the features are using to create their individual labels. This is how it works with supervised machine learning.

For Testing, in supervised machine learning, we input a set of features into the trained classifier and compare the classifier's output to the real data. label. This aids us in determining the precision of our data classifier.

Results:

Fig. 2. From 2005 through July 2018, the stock price of GOOGL was graphed. The red line represents the supplied data, while the blue line represents the stock's anticipated or predicted value.

We have achieved accuracy of 97.6%. Accuracy is a component that every machine learning developer is continuously striving to improve. Following the development of the model, an endless amount of effort is expended in order to improve the model's accuracy. The following are some standard approaches for calculating accuracy in machine learning: Confusion matrix for classification issues, R2 value of the model, RMSE value, Adjusted R2 value, and many more.

B. Paper II - Prediction of Stock Prices using Machine Learning (Regression,Classification) Algorithms [2]

In this paper, regression and classification are the two approaches used by the suggested system. The system predicts the closing price of a company's stock in regression, and it predicts whether the closing price of a company's stock will grow or drop the next day in classification.
The raw data or the dataset is taken from Yahoo Finance. Using in-built APIs, Yahoo Finance makes it simple to retrieve any historical stock values for a firm using the ticker-name. It has a function that allows you to receive pricing based on the starting and finish dates you enter.

The data includes the greatest, lowest values, starting price, ending price, and trading volume of stocks for a specific day and time.
The data must be changed or preprocessed before being fed into the machine learning model in order for the model to give the most accurate results feasible. The closing price is mostly utilized to input the model as an attribute.

We'll now go to the algorithmic method.

Following parameters are added into the dataset:

- **Momentum** - The current day's momentum is 1 if the current day's closing price is higher than the previous day's closing price, otherwise it is 0.
- **Volatility** - It is calculated as the difference between previous day stock closing price and current day stock closing price divided by previous day stock closing price.
- **Index Momentum** - It is the average of the index momentum over the preceding five days.
- **Index Volatility** - It's computed as the average of the index's volatility over the preceding five days.
- **Sector momentum** - It is determined as the average of the momentum values of all firms in a certain sector.
- **Stock momentum** - It represents a company's average momentum over the preceding 5 days.
- **Stock price volatility** - It is the average of a company's volatility over the preceding five days.

Regression: The system forecasts a company's stock closing price. The dataset will be applied on following methods: Simple Linear Regression, Polynomial Regression, Decision Tree Regression, Support Vector Regression and Random Forest Regression.

Classification: Classification models determine if a given input corresponds to one of the aforementioned classes. A binary classification is performed, with a Yes/No outcome. The outcome of this application will be determined by momentum. Based on all of the previous data, the model will predict whether the stock will go high or low the next day.

The following are the many classification algorithms used: Support Vector Machine (SVM), Logistic Regression, K – Nearest Neighbours (KNN), Random Forest Classification , Decision Tree Classification, Naïve Bayes.

Results:

TABLE I
RESULT ANALYSIS OF REGRESSION MODELS

Model	Accuracy	Time(in seconds)
Simple Linear Regression	81.52	0.77
Polynomial Regression	91.45	0.98
Support Vector Regression (SVR)	87.41	1.16
Decision Tree Regression	98.09	0.79
Random Forest Regression	99.57	1.06

C. Paper III - Study of Machine learning Algorithms for Stock Market Prediction [3]

The dataset has been downloaded from Kaggle. The dataset contains data of National Stock Exchange of India for the years 2016-2017. Then, the handling of missing data was the first step towards data pre-processing. The next step was one hot encoding of data. It transforms categorical data to a quantitative variable since data in the form of a string or an object is useless for data analysis. The last step was normalization of data. They have used various machine learning algorithms like:

TABLE II
RESULT ANALYSIS OF CLASSIFICATION MODELS

Model	Accuracy	Time(in seconds)
Support Vector Machine (linear)	68.41	158.48
Support Vector Machine (poly)	64.80	195.38
Support Vector Machine (rbf)	67.86	201.15
Support Vector Machine (sigmoid)	58.65	160.81
K – Nearest Neighbors	61.50	19.02
Logistic Regression	68.27	10.51
Naïve Bayes	67.10	10.14
Decision Tree Classification	57.99	198.57
Random Forest Classification	63.33	202.54

- **Random Forest classifier** - It's a form of supervised algorithm and ensemble learning software. It is created on the basis of decision trees in which multiple decision trees are created and merged for getting results. It can function good on large datasets and can work on classification as well as regression problems. It uses a huge number of trees so the speed can sometimes become an issue. **Algorithm** - In this algorithm we first randomly select m features then, find the best split and split the node based on it. Then we repeat these steps and build a forest by repeating all these steps.
- **Support Vector Machine (SVM)** - It is a supervised model which finds a separator after mapping data to a high-dimensional feature space. It locates an n-dimensional space in which data points are classified. The various tuning parameters are kernel parameter, gamma parameter and regularization parameter. New input is predicted by Linear Kernel by dot product between input and support vector. Kernelling is the process of data transformation to a higher-dimensional space. There are various types of kernel functions like linear, polynomial, RBF and Sigmoid. C is the regularization parameter with a value set to 10 by default. If there isn't enough regularisation, the categorization will be incorrect. A low gamma value indicates that the data region could not be found. The model may be improved by raising the significance of each data's categorization. A pro of this algorithm is it's an excellent approach for estimating in high-dimensional spaces, and it's memory-friendly. A con is that it is prone to over-fitting and that it performs admirably on small datasets
- **KNN (k-nearest neighbours)** - It only gives you results when you ask for them. Because there is no learning period, it is referred to as a lazy learner. KNN has the advantage of being one of the simplest algorithms because it just needs to compute the value of k and the Euclidean distance. KNN has the disadvantage of not going through the learning step, which means the method may not generalise effectively. It takes longer to sort a huge dataset because it must compute all the distances from the unknown item. **Algorithm** - First a value for k is chosen and Euclidean distance of all cases is calculated. Any k data points are selected around the unknown data. The unknown data will be found in most of the cases in a set of k neighbors.
- **Logistic Regression** - Used when response is binary. Multiclass and binary classification can be done using this. It produces the most precise results of all, but it necessitates the discovery of the greatest potential feature to suit.

Results
Random Forest Classifier and Logistic Regression showed the greatest accuracy where as KNN is the worst when it comes to processing time and accuracy both.

TABLE III
RESULTS OF THE FOUR ALGORITHMS

Algorithm	Accuracy	Recall	Precision	F-score
Random Forest	80.7	78.3	75.2	76.7
SVM	68.2	65.2	64.7	64.9
KNN	65.2	63.6	64.8	64.1
Logistic Regression	78.6	76.6	77.8	77.1

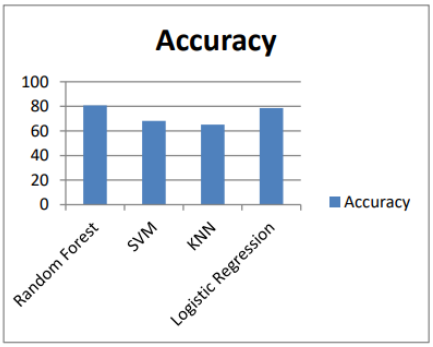

Fig. 3. Accuracy plot from [3]

D. Paper IV - Analysis of various machine learning algorithm and hybrid model for stock market prediction using python [4]

In this paper a comparison is made between normal machine learning algorithms and a hybrid approach that is made by combining two algorithms.

Proposed hybrid model approach - In the proposed hybrid model a combination of two algorithms is suggested. The Linear Regression model is used as base model as it showed least error when all these algorithms were individually tested. So, then each model was coupled with Linear Regression. Furthermore, the dataset is then imported using the pandas, resulting in a 70:30 split between training and testing. Then they predicted the values of the testing dataset using linear regression, which will be used as input for the next algorithm, and the output of the first algorithm is then divided into a 70:30 ratio for training and testing the dataset, yielding a result from combining two algorithms in a hybrid form model. Then they suggested that we can combine any two algorithms, but results from model testing show that linear regression provides much more accurate results than other algorithms, so it will be beneficial if linear regression is used first, and then any other algorithm is used in combination with it; as a result of this hybrid method, much more accurate results are expected.

Results :
It can thus be concluded that the hybrid models coupled with Linear Regression gave better results as compared to normal models.

TABLE IV
ERROR VALUES FOR HYBRID MODELS

	MAE	MSE	RMSE
LR + LR	3.0899217871739 e^{-16}	3.12975351019934 e^{-31}	1.75781733612272 e^{-8}
LR + KNN	0.0168824060	0.000717925	0.12993231
LR + SVM	0.0217266217	0.000875966	0.14739953
LR + DT	0.0268039095	0.0016172521	0.16371899
LR + RF	0.0147869564	0..0004964863	0.121601630

E. Paper V - Stock Market Prediction using Linear Regression and SVM [5]

The author has implemented the price prediction model using two algorithms - SVM and Linear Regression. Accuracy is calculated for both algorithms used and a result is made.

Dataset: The author has made dataset using the web scrapping method. It is done using BeautifulSoup libarary from the yahoo finance website. A plot is shown for the stocks of Amazon from 1 oct to 31 dec 2019.

Fig. 4. Amazon stock plot from [5]

SVM (STATE VECTOR MACHINE)

The author has explained the basics, architecture, working and the applications of this algorithm. SVM learns by giving labels to different objects. Then the objects are plotted and a hyperplne is made such that the features or characteristics are divided. The hyperplasia acts as a separator. This helps to predict the values accordingly by generating an equation or relation between two features. This can help to predict the stock prices based on the given data. Using the classes we can predict the stock price based on situation of the the market.

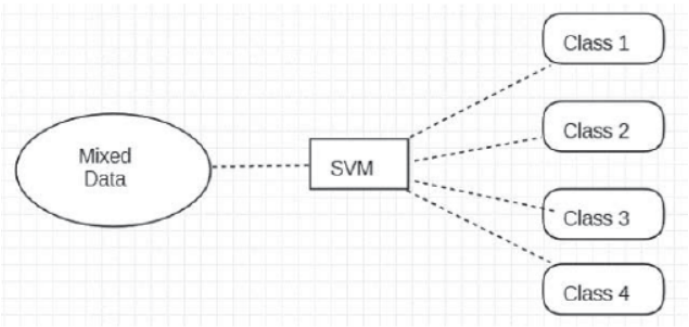

Fig. 5. SVM from [5]

Linear Regression
Linear Regression model shows the relation between the independent

and dependent data.It can be used for prediction of stock price based on the other values of stock market like the price of shares currently and the purchases done by customers. The author has used linear regression for prediction of stock prices of Amazon.

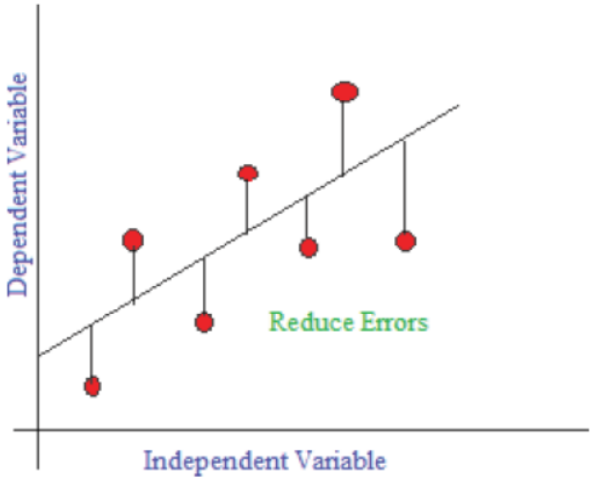

Fig. 6. Linear regression from [5]

Results

Using the SVR and Linear regression, the accuracy produced is 98.76% for the Linear regression and 94.32% for the state vector regression. This shows that Linear regression is better for stock price prediction than SVM in case of Amazon stocks.

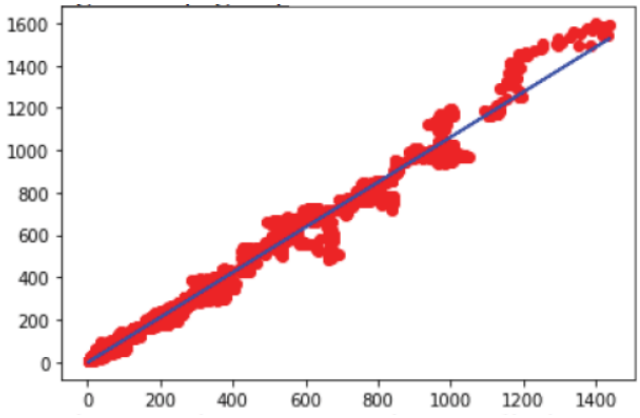

Fig. 7. Plot of linear regression from [5]

F. Paper VI - Recursive Stock Price Prediction With Machine Learning And Web Scrapping For Specified Time Period [6]

The algorithm used in the paper is Random forest regression. In this, while training the machines, the algorithm constructs many trees and trains them individually. Then the prediction of all the trees is combined to generate the final prediction. This method is called Ensemble method or technique.

The node probability is calculated to determine whether the data trained reaches the node in order to decrease the impurity. It is calculated by the number of samples that reach the node divided by total samples. Better the value, more important the feature is. The author has used Scikit learn library for the implementation. This is done for each tree and then the final feature importance value is calculated by sum of all tree values. Below are the steps for the process.

Step 1: Collection of data

This is done using web scrapping from NSE and BeautifulSoup library. Data collection is the most important part in any prediction. Here the data is real time as it is scrapped from the real stock market values.

Step 2: Data pre processing

This is done to discard the useless or unimportant values or features. Data preprocessing is a part of Data mining. By doing this, we can get a good accuracy of prediction for the data.

Step 3: Training the model

Using the above algorithm, the author has trained the machine. Moving average is calculated. Analysis for the company is done using the price per share divided by earnings per share. In this way, the model is trained to predict the future stock price.

Step 4: Recursion

The predicted stock price is updated in the dataset. One can continue the process to predict the price. But the accuracy of prediction decreases with increase of future prediction price value.

Results

The author has attached some snapshots of the experimented result. The snapshots are of dataset with predicted and actual values which are scrapped. It also has the comparison of price predicted and the change percentage of prices with respect to the actual price.

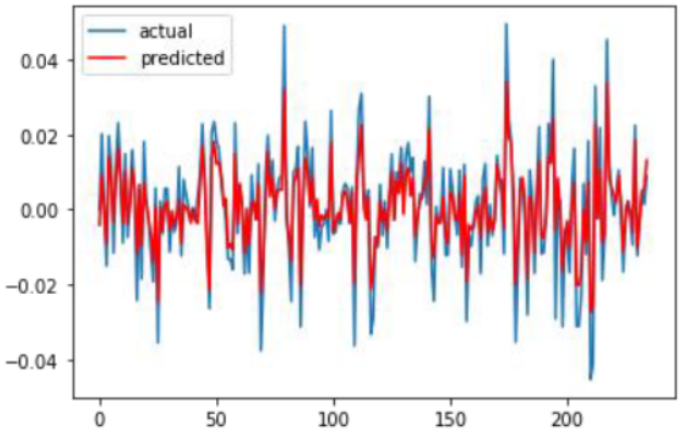

Fig. 8. Result from [6]

III. Final Result

The table shows algorithms which are most used in some papers and their accuracy.

	Paper I	Paper II	Paper III	Paper V	Average Accuracy
Simple Linear Regression	97.6	81.52	-	98.76	92.63
Polynomial Regression	-	91.45	-	-	91.45
Support Vector Regression (SVR)	-	87.41	-	-	87.41
Decision Tree Regression	-	98.09	-	-	98.09
Random Forest Regression	-	99.57	-	-	99.57
Support Vector Machine (linear)	-	68.41	68.2	94.32	76.97
Support Vector Machine (poly)	-	64.80	-	-	64.80
Support Vector Machine (rbf)	-	67.86	-	-	67.86
Support Vector Machine (sigmoid)	-	58.65	-	-	58.65
K – Nearest Neighbors	-	61.50	65.2	-	63.35
Logistic Regression	-	68.27	78.6	-	73.435
Naïve Bayes	-	67.10	-	-	67.10
Decision Tree Classification	-	57.99	-	-	57.99
Random Forest Classification	-	63.33	80.7	-	72.015

IV. Conclusion

Stock market trading being the most important factor in business and finance, it was necessary to develop a model that can predict the stock prices of future based on the factors which were currently in identified. The authors of the different papers have applied different machine learning algorithms to predict the future stock price. The papers have different datasets and some of them even have real time data as it is web scrapped from websites. The survey shows the algorithms used in papers and their accuracy for developing the model with respect to their dataset. One paper also had a hybrid model for predicting the stock price. A final result is shown depicting the accuracy's of algorithms that are used the most in all survey papers. Prediction of stock price in advance is a very good advantage for the companies as they can yield a very significant profit for their business. This ultimately benefits the customers.

References

[1] Kunal Pahwa, Neha Agarwal "Stock Market Analysis using Supervised Machine Learning", International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), 2019

[2] Srinath Ravikumar, Prasad Saraf "Prediction of Stock Prices using Machine Learning (Regression,Classification) Algorithms", International Conference for Emerging Technology (INCET),2020

[3] Ashwini Pathak, Sakshi Pathak "Study of Machine learning Algorithms for Stock Market Prediction",International Journal of Engineering Research Technology (IJERT), Vol. 9 Issue 06, June-2020

[4] Sahil Vazirani, Abhishek Sharma and Pavika Sharma " Analysis of various machine learning algorithm and hybrid model for stock market prediction using python", International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE 2020)

[5] B. Panwar, G. Dhuriya, P. Johri, S. Singh Yadav and N. Gaur, "Stock Market Prediction Using Linear Regression and SVM," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2021

[6] B. B. P. Maurya, A. Ray, A. Upadhyay, B. Gour and A. U. Khan, "Recursive Stock Price Prediction With Machine Learning And Web Scrapping For Specified Time Period," 2019 Sixteenth International Conference on Wireless and Optical Communication Networks (WOCN), 2019