International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

STOCK PRICE PREDICTION: A STUDY BETWEEN STATISTICAL APPROACH AND MACHINE LEARNING APPROACH

Kavin S
1st B.Tech Artificial Intelligence and Data science
Kongu Engineering College
India,Tamil Nadu,Erode
Kavins.22aid@kongu.edu

Madan Raj M A
1st B.Tech Artificial Intelligence and Data science
Kongu Engineering College
India,Tamil Nadu,Erode
madanrajma.22aid@kongu.edu

Arunachalam M
1st B.Tech Artificial Intelligence and Data science
Kongu Engineering College
India,Tamil Nadu,Erode
arunachalamm.22aid@kongu.edu

Bairavi E
1st B.Tech Artificial Intelligence and Data science
Kongu Engineering College
India,Tamil Nadu,Erode
bairavie.22aid@kongu.edu

Abstract: Stock price prediction uses historical data and financial indicators to estimate future stock prices. Machine learning algorithms are commonly used for this task, analyzing factors such as past stock prices, earnings, news and market trends. The goal is to assist investors in making informed decisions. However, the prediction of stock prices is challenging due to the many factors that influence them. Despite this, advancements in machine learning models and data availability have improved prediction accuracy.

I. INTRODUCTION

Stock price prediction refers to the estimation of future stock prices based on analysis of historical data and financial indicators. The aim is to provide valuable insights to investors in order to assist them in making informed decisions. This prediction is typically carried out using machine learning algorithms, which analyze data such as past stock prices, earnings reports, news articles and market trends. Despite the complex nature of the stock market, advancements in technology have enabled more accurate predictions.

II. OBJECTIVES OF STOCK PRICE PREDICTION

The primary objectives of stock price prediction are:

1. To help investors make informed investment decisions by providing insight into the potential future performance of a stock.
2. To identify market trends and patterns that can be used to make predictions about future stock prices.
3. To provide a quantitative assessment of the risks and potential rewards associated with a particular stock.

III. METHODOLOGIES OF STOCK PRICE PREDICTION

1. STATISTICAL METHODS:-

Simple Moving Average: Simple Moving Average (SMA) is a commonly used technical analysis indicator in stock price prediction. It is calculated by taking the average of the past "n" stock prices, where "n" is a user-defined parameter. An unweighted mean of a specific number of previous data is considered to be the predicted value for the next day**.**

The Formula used for SMA:-

$$F_t = \frac{A_{t-1} + A_{t-2} + A_{t-3} + \cdots + A_{t-n}}{n} \quad (1)$$

In (1),

F_i = Predicted closing price for i^{th} day

A_i = Actual closing price at i^{th} day

n = Number of days considered for prediction

Weighted Moving Average: Weighted Moving Average (WMA) is another technical analysis indicator used in stock price prediction. Similar to Simple Moving Average (SMA), it is calculated by taking the average of a set of past stock prices. However, unlike SMA, WMA assigns a weight to each stock price in the calculation, giving more importance to the most recent prices.

The difference between SMA and WMA is that a weight is used with the previous values to predict the future value

The formula used for WMA :-

$$F_t = \frac{A_{t-1}w_{t-1} + A_{t-2}w_{t-1} + \dots + A_{t-n}w_{t-1}}{100} \quad (2)$$

$$w_{t-i} = W * (n - i) \quad (3)$$

$$W = \frac{100}{1 + 2 + 3 + \dots + n} \quad (4)$$

Here, w_i = Weight used for i^{th} day

W = Unit weight

EXPONENTIAL SMOOTHING: Exponential Smoothing is a time series forecasting method used in stock price prediction. It is a method for calculating a weighted average of past stock prices, with the weights assigned to each price decaying exponentially as the prices get older. This approach gives more importance to the most recent prices, similar to Weighted Moving Average (WMA), but allows for more flexibility in adjusting the weight assigned to each price.

A smoothing constant, α is used for smoothing the prediction value from the previous prediction.

This smoothing constant maximizes prediction accuracy from the last prediction.

The formula used for exponential smoothing

$$F_t = A_{t-1} + \alpha * (A_{t-1} - F_{t-1}) \quad (5)$$

Here,

α = Smoothing constant

2. MACHINE LEARNING METHOD

***REGRESSION*:** Regression is a statistical method used in stock price prediction. It involves fitting a mathematical model to historical stock price data, in order to make predictions about future stock prices. The goal of regression is to identify relationships between independent variables, such as economic indicators, and the dependent variable, the stock price.

- Simple linear regression algorithm is one of the fundamental supervised machine learning algorithms used for regression.
- Ridge is one of the techniques which reduces model complexity and prevents over-fitting.

K-NEAREST NEIGHBOUR:

K-Nearest Neighbour (KNN) is a machine learning method used in stock price prediction. It is based on the idea that stock prices that are similar in the past are likely to be similar in the future. The method works by finding the "k" nearest historical stock prices to a given target stock price, and using those prices to make a prediction about the future price.

RANDOM FOREST:

Random Forest is a machine learning method used in stock price prediction. It is an ensemble learning technique that builds multiple decision trees and combines their predictions to make a final prediction. Each decision tree is constructed using a random subset of the training data and features, and the predictions of all the trees are combined through a process such as majority voting or weighting.

SUPPORT VECTOR MACHINES:

Support Vector Machines (SVMs) is a machine learning method used in stock price prediction. It is a type of supervised learning algorithm that is used for classification and regression tasks. In stock price prediction, SVMs can be used to predict whether a stock will rise or fall, or to predict the actual stock price.

PERFORMANCE MEASURES:

Mean Squared Error (MSE)

$$MSE = \frac{\sum_{t=1}^{n}(A_t - F_t)^2}{n} \quad (6)$$

Mean Absolute Percentage Error (MAPE)

$$MAPE = \left(\frac{100}{n}\right) * \left|\frac{A_t - F_t}{A_t}\right| \quad (7)$$

IV. PROPOSED SYSTEM

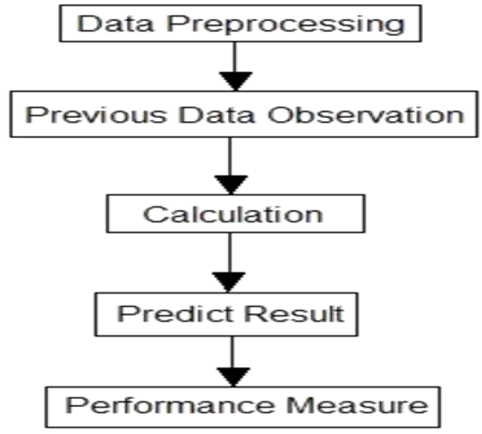


Fig 1. Flow Diagram of statistical methods

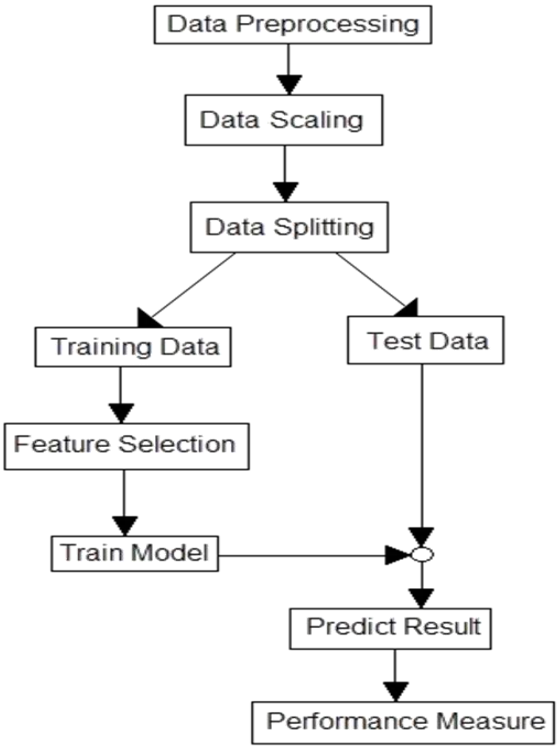


Fig.Flow Diagram of machine learning methods

1. DATASET

Two datasets have been used in the proposed system to predict the stock price:

i. The stock prices of **Tesla**

Start : 29-06-2010

End: 17-03-2017

Days: 1693

ii. The stock prices of **AppleStart:**

02-01-2014

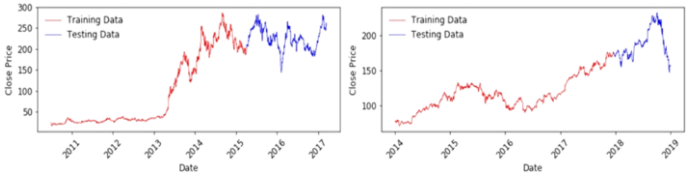
End: 31-12-2018

Days: 1259

Each row represents the information of a singleday.

There are six columns for each row.

- 1st column - the date
- 2nd column - the opening price of that day
- 3rd column - the highest price of that day
- 4th column - the lowest price of that day
- 5th column - the closing price of that day
- 6th column - the volume of shares traded on thatday.



2. DATA PREPROCESSING

Data preprocessing includes-

- Checking out for missing values and discardsthose data from the dataset
- Looking for categorical values
- Drop out unnecessary information in thedataset.

3. DATA SPLITTING

The dataset has been split into two parts as training dataand test data.

(a) For Tesla dataset,

Training Data (1200 days): 29-06-2010 to 06-04-2015

Testing Data (492 days): 07-04-2015 to 17-03-2017.

(b) For Apple dataset,

Training Data (1000 days): 02-01-2014 to 19-12-2017

Testing Data (258 days): 20-12-2017 to 31-12-2018

International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

4. FEATURE SELECTION

For time series prediction, selection of features is an important task. Because selection of worst features can direct the prediction in a wrong way. In this system, three features have been selected.

- The opening price
- The highest price
- The lowest price.

5. PREDICTION

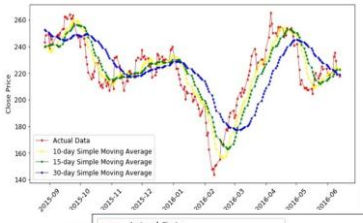
As statistical methods, predictions have been performed using 10-day, 15-day, 30-day Simple Moving Average and Weighted Moving Average, Exponential Smoothing with $\alpha = 0.3, 0.5, 0.75$ and naïve approach.

As Machine Learning methods, predictions have been performed using Simple Linear Regression, Lasso Regression and Ridge Regression, K-Nearest Neighbor, Random Forest with different number of estimators, Support Vector Machine and Neural Network Models like SLP, MLP and LSTM.

After predictions, MSE and MAPE values are calculated.

V. RESULT AND DISCUSSION

Table I. Performance Measures of Different Simple Moving Average Methods



SMA	Errors for Tesla Dataset	
	MSE_T	$MAPE_T$
30-Day SMA	322.4261	6.4752
15-Day SMA	141.0808	4.3776
10-Day SMA	87.1748	3.3761
SMA	**Errors for Apple Dataset**	
	MSE_A	$MAPE_A$
30-Day SMA	125.42905	4.6513
15-Day SMA	53.5439	3.1054
10-Day	32.6763	2.3691

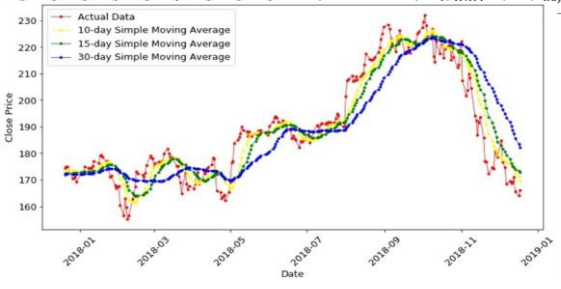


Table II. Performance Measures of Different Weighted Moving Average Methods

MA	Unit Weight W	Errors for Tesla Dataset	
		MSE_T	$MAPE_T$
-Day MA	0.22	184.6105	4.9482
-Day MA	0.83	81.1320	3.2736
-Day MA	1.82	50.4788	2.8051
MA	**Unit Weight W**	**Errors for Apple Dataset**	
		MSE_A	$MAPE_A$
-Day MA	0.22	69.8537	3.5392
-Day MA	0.83	30.3559	2.2954
-Day MA	1.82	18.3376	1.7408

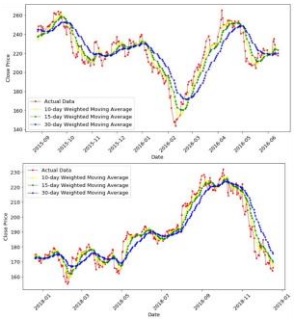


Table III. Performance Measures of Different Regression Methods

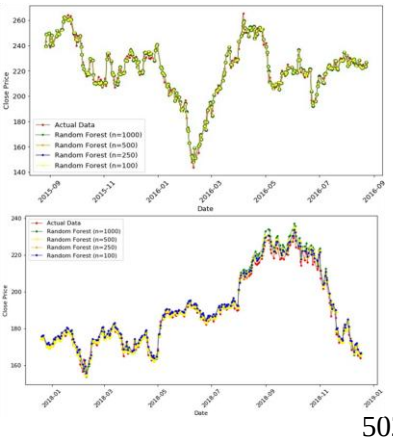
Regression Method	Errors for Tesla Dataset	
	MSE_T	$MAPE_T$
Simple Linear Regression	10.2627	1.1311
Lasso Regression	10.1814	1.1287
Ridge Regression	9.3313	1.1045
Regression Method	**Errors for Apple Dataset**	
	MSE_A	$MAPE_A$
Simple Linear Regression	7.4275	1.0987
Lasso Regression	7.4111	1.0978
Ridge Regression	7.2778	1.0512

Table IV. Performance Measures of KNN and SVM

Method	Errors for Tesla Dataset	
	MSE_T	$MAPE_T$
K-Nearest Neighbors	6.6241	0.8869
Support Vector Machine	8.3624	0.9947
Method	**Errors for Apple Dataset**	
	MSE_A	$MAPE_A$
K-Nearest Neighbors	6.6314	0.9423
Support Vector Machine	4.1864	0.8111

Table V. Performance Measures of Random Forest with Different Number of Estimators

Number of estimators	Errors for Tesla Dataset	
	MSE_T	$MAPE_T$
100	7.1052	0.9131
250	6.9909	0.9129
500	7.0017	0.9144
1000	6.9326	0.9119
Number of estimators	**Errors for Apple Dataset**	
	MSE_A	$MAPE_A$
100	6.5189	0.9785
250	6.5121	0.9778
500	6.0489	0.9433
1000	6.6759	0.9827



VIII. CONCLUSION

A comparative study between statistical and machine learning approaches has been done in terms of prediction performances. Machine learning methods, especially MLP and LSTM are found to be the most accurate to predict stock prices.

ACKNOWLEDGEMENT

We would like to express my sincere gratitude to all those who have contributed to this project. We are deeply grateful to our **HOD**, **Gandhimathi S**, for their invaluable guidance, support, and encouragement throughout the entire project.

We would also like to thank our colleagues and friends who provided us with valuable feedback and insights. We are especially thankful to **Dharshan P P**, **Lakshya S**, **Karthik Raja S K**, **Abishek S** for their time, patience, and expertise in helping us bring this project to fruition.

Finally, We would like to acknowledge the support of **Kongu Engineering College**, whose funding and resources made this project possible.

Thank you all for your dedication and support

REFERENCES

1. Saptashwa, “Ridge and Lasso Regression: A Complete Guide with Python Scikit-Learn,”https://towardsdatascience.com/ridge-and-lassoregressiona-complete-guide-with-python-scikit-learn-e20e34bcbf0b, Sep 26, 2018.
2. D. S. Sayad, “K Nearest Neighbors - Regression,” http://saedsayad.com/k nearest neighbors reg.htm.
3. J. Patel, S. Shah, P. Thakkar, and K. Kotecha, “Predicting stock market index using fusion of machine learning techniques,” Expert Syst. Appl., vol. 42, pp. 2162–2172, 2015.
4. D. S. Sayad, “Support Vector Machine - Regression (SVR),” https://www.saedsayad.com/support vector machine reg.htm.