WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Stock Market Prediction Using Machine Learning

[1]Prof. Krishna M Aldar, [2]Niranjan Kabade, [3]Rajat Gatade, Megha Rathod, [4]Pradnya Kambale.

Sanjay Ghodawat Institutions

ABSTRACT –

In the capitalist world stock swapping is one of the most significant exercises. Stock exchange cast is a demonstration of trying to decide the unborn estimation of a stock. This paper clarifies the cast of a stock exercising Machine literacy. The technical and major or the time arrangement examination is employed by the maturity of the stockbrokers while making the stock prognostications. The programming language is employed to prevision the fiscal exchange exercising AI in Python. Right now we propose a Machine Learning (ML) approach that will be trained from the accessible stocks information and afterward utilizes the procured information for a precise prediction. Right now study utilizes an AI procedure called Linear Regression Algorithm and Random Forest Algorithm to foresee stock costs.

Key Words: Stock Market, Machine Learning, Predictions, Linear Regression, Random Forest

1. INTRODUCTION

These days, as the associations between overall economies are fixed by globalization, outer aggravations to the money related markets are never again residential. With developing capital markets, an ever increasing number of information is being made day by day.

The inherent estimation of an organization's stock is the worth controlled by assessing the normal future inflows of a stock and limiting them to the present, which is known as the book value. This is different from the market estimation of the stock, that is controlled by the organization's stock cost.

Investing in the stock market is among the most common ways investors try to grow their moneybags, but it's also among the unsafe investment options available. Understanding the introductory conception of the stock market is the first step in getting an informed investor. While the stock market is an extremely complex system, its introductory traits are much more simple.

In this approach, with the use of supervised learning classifier to predict stock price movement based on financial index report, and evaluate their potency is proposed. Statistical analytic methods in the financial market have become stock modeling. Here in this paper we used two algorithms viz. Linear Regression algorithm and Random Forest Algorithm. We tried to use different types of features and kernels in order to achieve the results. We took the Data Set from Quandl- an online data set provider.

We took the data set of Ford company which works in the automobile sector. By splitting the data in various ways we tried to minimize the loss as possible. We also added our features to increase the accuracy of the training model.

2. METHODOLOGY

Linear Regression -

Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Linear regression analysis is **used to predict the value of a variable based on the value of another variable**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

Direct relapse plays out the errand to foresee a needy variable worth (y) in light of a given autonomous variable (x). Along these lines, this relapse procedure discovers a straight connection between x (input) and y(output). Subsequently, the name is Linear Regression.

In the figure above, X (input) is the work understanding and Y (yield) is the pay of an individual. The relapse line is the best fit line for our model. Hypothesis function for Linear Regression :

$$y = \theta_1 + \theta_2.x$$

While preparing the model we are given :
x: input preparing information (univariate – one info variable(parameter))
y: labels to information (directed learning)
When preparing the model – it fits the best line to foresee the estimation of y for a given estimation of x. The model gets the best relapse fit line by finding the best θ1 and θ2 values.
θ1: intercept
θ2: coefficient of x
When we locate the best θ1 and θ2 values, we get the best fit line. So when we are at long last utilizing our model for expectation, it will anticipate the estimation of y for the info estimation of x.

Step by step instructions to refresh θ1 and θ2 values to get the best fit line ?
Cost Function (J):

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

Cost function(J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (pred) and true y value (y)

Advantages:

1. Linear Regression is simple to implement and easier to interpret the output coefficients.
2. When you know the relationship between the independent and dependent variable have a linear relationship, this algorithm is the best to use because of it's less complexity to compared to other algorithms.
3. Linear Regression is susceptible to over-fitting but it can be avoided using some dimensionality reduction techniques, regularization (L1 and L2) techniques and cross-validation.

Disadvantages:

1. On the other hand in linear regression technique outliers can have huge effects on the regression and boundaries are linear in this technique.
2. Diversely, linear regression assumes a linear relationship between dependent and independent variables. That means it assumes that there is a straight-line relationship between them. It assumes independence between attributes.
3. But then linear regression also looks at a relationship between the mean of the dependent variables and the independent variables. Just as the mean is not a complete description of a single variable, linear regression is not a complete description of relationships among variables.

Random Forest -

Random Forest is a Supervised Learning algorithm which utilizes outfit learning strategy for classification and relapse. Random Forest is a bagging technique and not a boosting technique. The trees in random forests are run in equal. There is no association between these trees while building the trees. It works by building a huge number of choice trees at preparing time and yielding the class that is the mode of the classes (classification) or mean expectation (regression) of the individual trees. An arbitrary backwoods is a meta-estimator (for example it consolidates the aftereffect of numerous forecasts) which aggregates numerous choice trees, with some supportive adjustments:

1. The number of highlights that can be part on at every hub is constrained to some level of the aggregate (which is known as the hyperparameter). This guarantees the troupe model does not depend too vigorously on any individual component, and makes fair utilization of all conceivably prescient highlights.
2. Each tree draws an arbitrary example from the first informational collection while producing its parts, including a further component of arbitrariness that prevents overfitting. The above adjustments help keep the trees from being excessively exceptionally associated. For Example, See these nine choice tree classifiers underneath :

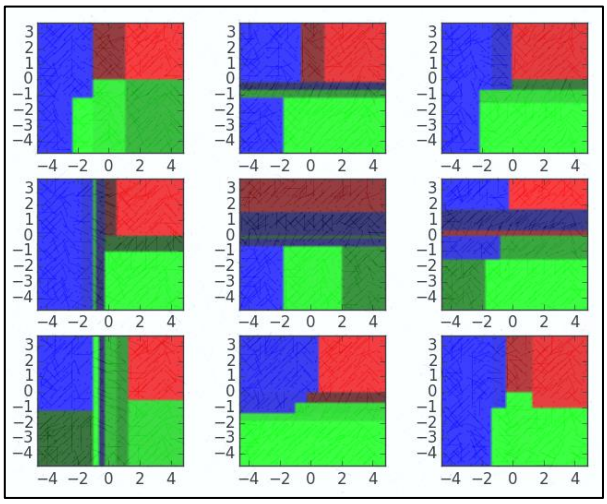


Fig. Nine Different Decision Tree Classifiers

These choice tree classifiers can be collected into an arbitrary woodland group which combines their information. Think about the even and vertical tomahawks of the above choice tree yields as highlights x1 and x2. At specific estimations of each element, the choice tree yields a characterization of "blue", "green", "red", and so forth. These above results are accumulated, through model votes or averaging, into a solitary gathering model that winds up beating any individual choice tree's yield. The totaled outcome for the nine choice tree classifiers is appeared underneath :

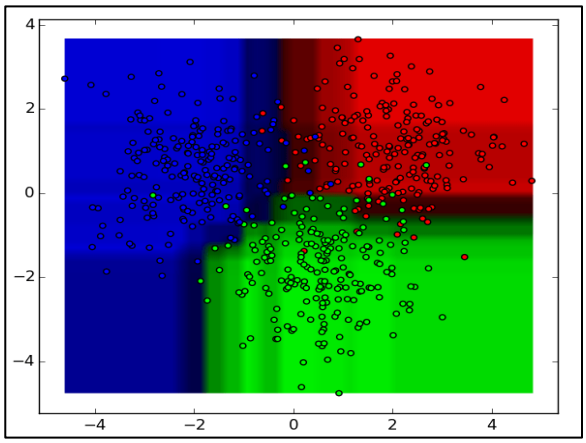


Fig. Random Forest ensemble for the above Decision Tree classifiers

Feature and Advantages of Random Forest :

1. It's one of the most accurate literacy algorithms available. For numerous data sets, it produces a largely correct classifier.
2. It runs efficiently on large databases.
3. It can handle thousands of input variables without variable omission. It gives estimates of what variables that are important in the bracket.
4. It generates an internal unprejudiced estimate of the conception error as the forestland structure progresses.
5. It has an effective system for estimating missing data and maintains delicacy when a large proportion of the data are missing.

Disadvantages of Random Forest :

1. Random forests have been adhered to overfit for some datasets with noisy category/ regression assignments.
2. For data carrying categorical variables with different number of situations, random forests are turned in favor of those attributes with further situations. thus, the variable significance scores from random forest aren't dependable for this type of data.

Model Creation and Evaluation Methods

i. Preprocessing and Cleaning :

Interpolating or recovering the missing data and removing the redundant data. This step also involves creating any useful feature from the existing ones.

ii. Feature Extraction :

This progression includes looking in the space of conceivable element subsets. We at that point pick the subset that is ideal or close ideal concerning some goal work. This is done as such as to stay away from issues of overfitting/underfitting the dataset.

iii. Data Normalization :

Information is should have been standardized for better precision by guaranteeing that all highlights are not given inordinate/low weightage.

a. Classification Methods

This phase would involve supervised classification methods like Support Vector Machines, Neural Networks, Naive Bayes, Ensemble classifiers (like Adaboost, Random Forest Classifiers), etc.

Fig. Level 0 DFD

b. Regression Methods:

These models would be used to get the expected numerical value of the interested stock. This phase would involve supervised regressions methods like Linear Regressions, Support Vector Regressions, Usage of Kernel Methods, etc.

i. Social Media Sentiment Analysis

a. Analyzing the current market situation from the latest news headlines and social media platform such as Twitter to gain insights into the future of stock prices.

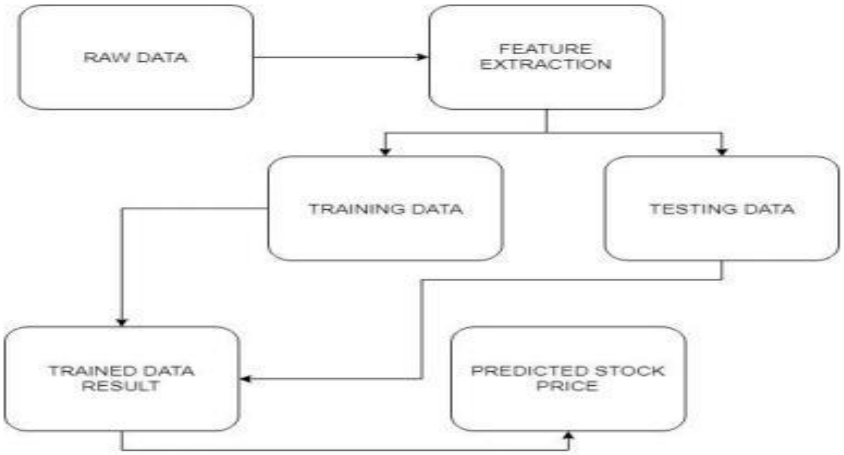
ii. Credit Assignment Problem

b. This progression includes the doling out of fitting weightage to various ways utilized for information assortment.

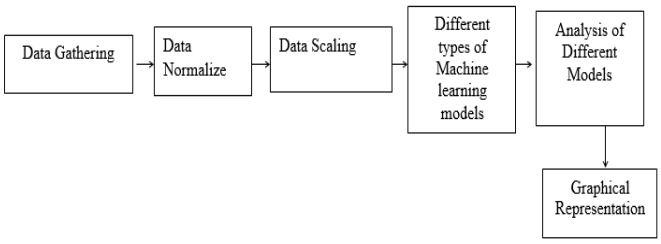
iii. Analysis of Different Models

Examination between the different techniques and models actualized over the datasets. anticipated test component by anticipating it into face space and contrasting with preparing.

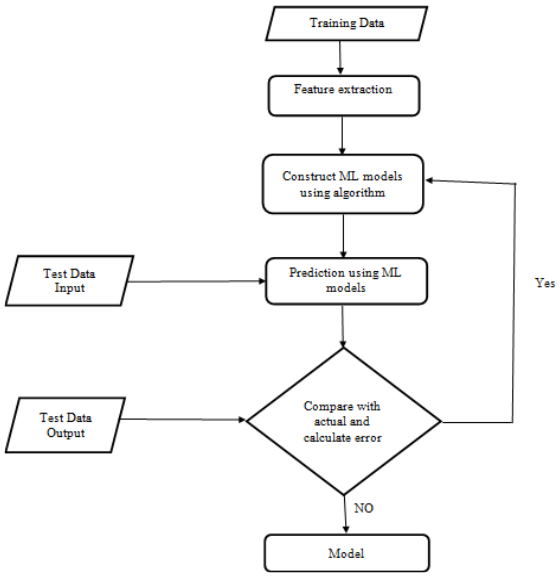
System Architecture :



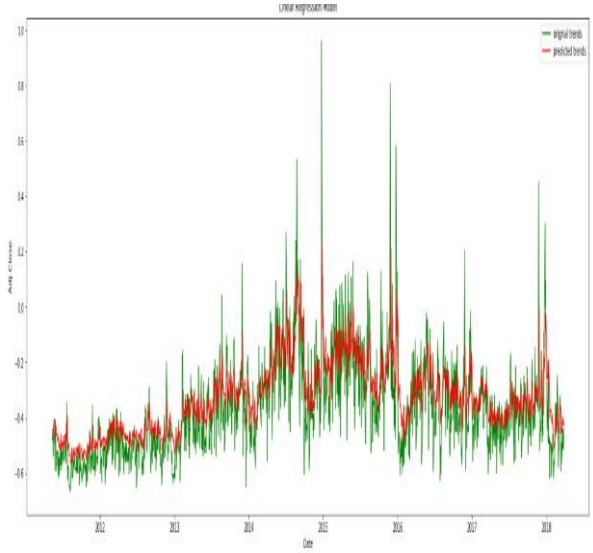
System Processing :



Flow Chart :

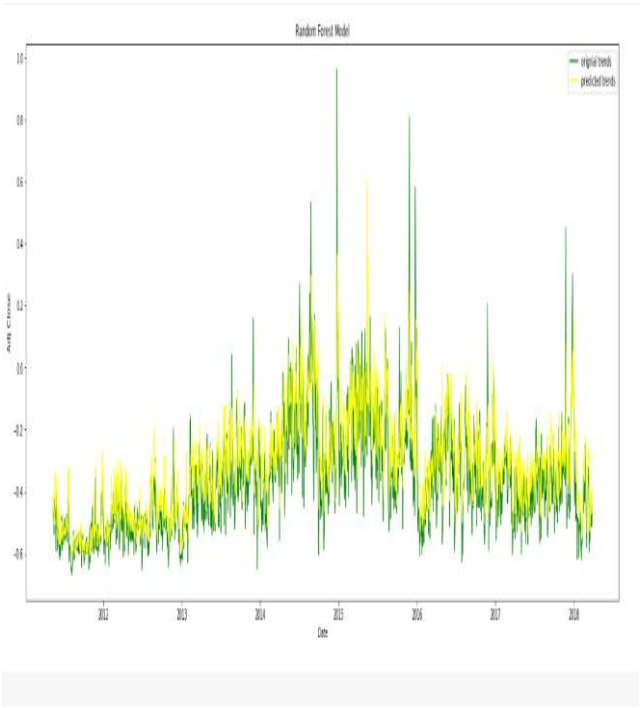


3. RESULTS



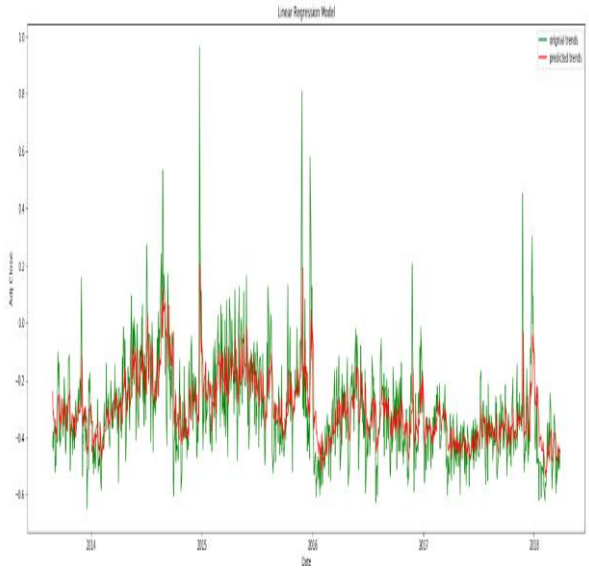
Training =70% , Testing = 15%, Forecast = 15%

Fig.3.1 Linear Regression Model



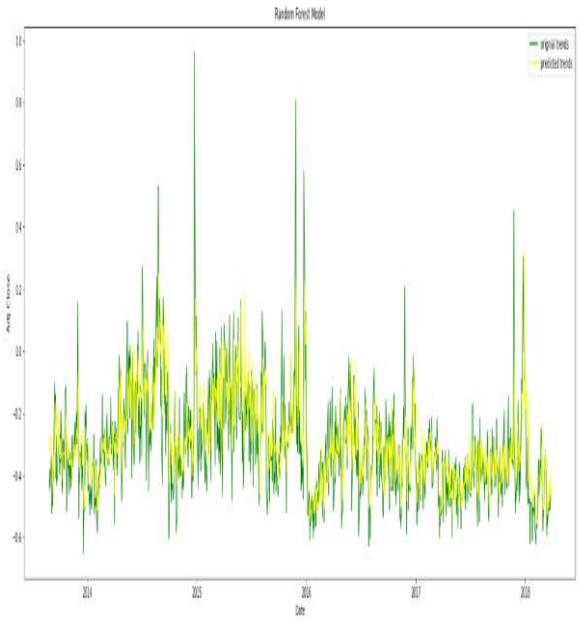
Training =70%, Testing = 15%, Forecast = 15%

Fig.3.2 Random Forest Model



Training =80% , Testing = 10%, Forecast = 10%

Fig.3.3 Linear Regression Model



Training =80% , Testing = 10%, Forecast = 10%

Fig.3.4 Random Forest Model

We implemented Linear Regression and Random Forest algorithms on dataset in Two ways. We divided dataset into training, testing and forecasting data.

1.Training = 70% Testing = 15% and Forecasting = 15%

2. Training = 80% Testing = 10% and Forecasting = 10%

The Linear Regression calculation works with exactness 0.80 for 70-15-15 methodology and 0.64 for 80-10-10.

The Random Forest calculation works with exactness 0.80 for 70-15-15 methodology and 0.63 for 80-10-10.

Direct Regression and Random Forest produce comparable outcomes in both the cases. The two calculations have practically same exactness.

In 80-10-10 methodology the model overfits the preparation information. We need increasingly summed up model so as to deliver exact outcomes.

4. CONCLUSION

In the task, we proposed the utilization of the information gathered from various worldwide budgetary markets with AI calculation so as to anticipate the stock file developments.

The further studies may include adding sentiment analysis which has pretty good effect on the stock prices. This will marginally increase the performance of the model and give us real world results.

The developed application does its predication of stock market prices with minimum amount of error. It was also observed from the experimental results that optimal prediction can be achieved by using Linear Regression Algorithm. The initialization scheme may be improved by estimating weights between input nodes and hidden nodes, instead of random initialization. Enrichment of more relevant inputs such as fundamental data and technical data from derivative markets may improve the predictability of the network. Applying Linear Regression Algorithm and Random Forest Algorithm in training data for stock prediction has been shown in this paper to be an efficient tool in developing applications for stock prediction.

References

1. https://www.researchgate.net/publication/328930285_Stock_Market_Prediction_Using_ Machine_Learning
2. https://www.researchgate.net/publication/259240183_A_Machine_Learning_Model_for_Stock_Market_Prediction
3. https://ieeexplore.ieee.org/document/8212715
4. http://www.ijaerd.com/papers/special_papers/RTDE014
5. One Up On Wall Street By Peter Lynch
6. Decision Trees and Random Forests: A Visual Introduction for Beginners By Chris Sm