Stock Prediction using Machine Learning

Sanju Abraham Binu
Master of Computer Applications
Amal Jyothi College of Engineering
Kottayam, Kerala, India
sanjuabraham321@gmail.com

Rony Tom
Master of Computer Applications
Amal Jyothi College of Engineering
Kottayam, Kerala, India
ronytom@amaljyothi.ac.in

Abstract: Due to the volatility and non-linear nature of economic inventory markets, as it should be predicting stock market returns is extraordinarily tough. because the creation of machine studying and stepped forward processing functionality, programmable prediction techniques have established to be more green in predicting stock values.

Keywords: Stock, Machine Learning, Linear Regression

I. INTRODUCTION

The stock marketplace is dynamic, unpredictable, and nonlinear in nature. Predicting inventory fees is a tough assignment because they're affected by a spread of factors, which include but not restrained to political conditions, the worldwide financial system, a enterprise's economic reviews and overall performance, and so on. for this reason, in an effort to maximize income and decrease losses, techniques for predicting inventory values earlier by way of reading the trend over the previous couple of years may be extremely useful for making stock market moves [1][2]. Traditionally, two techniques to predicting a company's inventory rate have been proposed. The technical analysis approach predicts the destiny charge of a stock by the usage of ancient facts such as the stock's final and establishing fee, quantity traded, adjoining close values, and so on. The second sort of analysis is qualitative, and it is carried out by way of financial analysts who use external elements which includes company profile, marketplace scenario, political and economic considerations, and textual statistics inside the shape of monetary new articles, social media, or even blogs. Superior smart techniques based on both technical and fundamental analysis at the moment are used to forecast stock prices. For inventory market evaluation, the records length is specially large and non-linear. To address one of these various set of facts, a model able to detecting hidden patterns and complex relationships in one of these massive statistics set is required. Device learning strategies were proven to boom efficiency in this area by means of 60-86 percent whilst in comparison to previous techniques [4].

One approach for becoming a straight line of the form y = mx + c to a graph in this sort of way that it passes through the best range of statistics points in the dataset. Match a straight line on a graph thru the factors of the dataset so that the square of the space between each point and the line is as small as viable. The speculation is a line that predicts the y value for any given x. The prediction method is referred to as Linear Regression, and the formula used is the Least Squares method. This approach is well-known among statisticians and is likewise a essential idea in machine gaining knowledge of.

Linear Regression's hypothesis function has general form:

$$y = h\square(x) = \square 0 + \square 1 x \quad (1)$$

Take note that this is similar to the equation for a straight line. The values $\square 0$ and $\square 1$ is passed to $h\square(x)$ to obtain the estimated output y. It should be noted that the effort is to obtain various values of $\square 0$ and $\square 1$ in order to find values that provide the best possible "fit" or most representative "straight line" through the data points mapped on the x-y plane. The accuracy of the hypothesis is assessed using a cost function that averages all of the results of the tests. The accuracy of the hypothesis is determined by a cost function that averages all of the hypothesis's results using inputs from x's compared to the actual output y's.

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

This function is also known as the "Squared error function" or the "Mean squared error." The mean is halved (1/2m) for ease of computation of the gradient descent, because the derivative term of the square function cancels out the1/2 term.

II. METHODOLOGY

a) Environment

The Jupyter notebook improvement surround built-in and a statistical language which integrated Python were used to create this survey. As a result, built-in the observations with Python functions makes it simple to get right of entry to them later. Because Linear Regression is a general method utilized in nearly all facts technological know-how built-in, the built-in Python features are getting used for this.

b) Time - Series Forecasting

A time collection is a collection of facts points which have been indexed (or listed or graphed) in chronological order. A time series is most usually defined as a sequence of images taken at same intervals across time. As a end result, it's miles a discrete-time information series. Time series analysis is the method of studying time collection facts on the way to extract

DOI:10.5281/zenodo.6365072
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

significant information and other facts traits.

Using a version to expect destiny values based on previously determined values is referred to as time series forecasting. Prediction is a part of statistical inference in data. Forecasting is the method of transferring data across time, frequently to particular factors in time [2chrome].

III. System Setup and Experimental Results

a) Data Collection

Kaggle's historical inventory fee repository was used. With a simple click, Kaggle affords a training and trying out dataset that may be used to retrieve a organization's stock rate for any time variety. This eliminates the want for manual records mining through different manner. We used stock data from The Tesla Corporation for this survey, which spans exactly nine years, from June 2010 to March 2019. Date, open, low, high, close, adj near, and volume are some of the 2193 values inside the dataset.

b) Analysis Method

To assess one of the machine gaining knowledge of, it's far enough to show that the predicted model suits the records as accurately as feasible. Instead of creating a prediction, the effectiveness of Linear Regression is verified by using the use of the schooling statistics as the check records set. The technique can be used to forecast real destiny cost by demonstrating how properly the model suits. A easy Linear Regression in one Variable, particularly the ultimate inventory charge or quit of Day rate, is used to forecast the price. The linear regression is an easy mathematical tool which can be used to provide correct consequences in prediction.

i. Jupyter Notebook

JupyterLab is an interactive pocket book, code, and records development surroundings this is handy thru the web. The usage of the platform's flexible interface, customers can configure and set up workflows in statistics technology, medical computing, computational journalism, and machine gaining knowledge of. A modular structure lets in for extensions that boom and improve capability.

ii. Python and Data Analysis

Python will undoubtedly be compared to the numerous other domain-particular open source and commercial programming languages and equipment in extensive use for statistics analysis and interactive, exploratory computing and statistics visualization, inclusive of R, MATLAB, SAS, Stata, and others. In recent years, Python's growing library guide (particularly pandas) has made it a great choice for records processing workloads. Python's energy in popular reason programming makes it an amazing preference as an unmarried language for developing information-centric apps.

iii. NumPy, Pandas, Matplotlib, Chart Studio

NumPy is a scientific computing environment based totally on Python. Pandas is a Python open-source library that provides efficient and person-pleasant facts shape and evaluation skills. Matplotlib is a Python toolkit that permits you to create graphs, charts, and different records visualization tools. you'll want the Chart Studio package deal to apply it, which includes utilities for running with Plotly's Chart Studio provider (both Chart Studio cloud and Chart Studio On-Prem).

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline

import chart_studio.plotly as py
import plotly.graph_objs as go
from plotly.offline import plot

#for offline plotting
from plotly.offline import download_plotlyjs, init_notebook_mode, plot, iplot
init_notebook_mode(connected=True)
```

iv. Statistical Values

Mean, count, standard deviation, minimum value, quartiles and maximum value are being depicted in the table.

	Open	High	Low	Close	Adj Close	Volume
count	2193.000000	2193.000000	2193.000000	2193.000000	2193.000000	2.193000e+03
mean	175.652882	178.710262	172.412075	175.648555	175.648555	5.077449e+06
std	115.580903	117.370092	113.654794	115.580771	115.580771	4.545398e+06
min	16.139999	16.629999	14.980000	15.800000	15.800000	1.185000e+05
25%	33.110001	33.910000	32.459999	33.160000	33.160000	1.577800e+06
50%	204.990005	208.160004	201.669998	204.990005	204.990005	4.171700e+06
75%	262.000000	265.329987	256.209991	261.739990	261.739990	6.885600e+06
max	386.690002	389.609985	379.350006	385.000000	385.000000	3.716390e+07

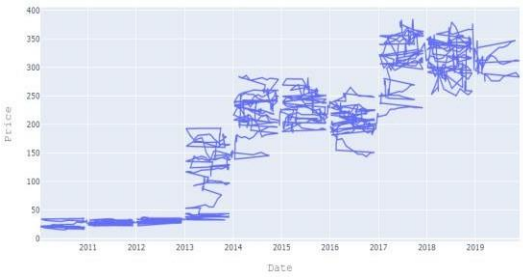
v. Stock Candlestick Pattern

Candlestick charts are used by traders to forecast price movements based on historical patterns. Candlesticks are useful in trading because they show four price points (open, close, Adj close, high, and low) over a given time period. Many algorithms make use of the same price information as candlestick charts. Emotion is frequently the driving force behind trading, as evidenced by candlestick charts. The Candlestick pattern is depicted in Figure 1.

DOI:10.5281/zenodo.6365072
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

vi. Plotting

The price value is plotted against date so that a plot chart is made for the full years.



vii. Linear Regression

Linear regression is the most basic and widely used type of predictive analysis. The purpose of regression analysis is to look at two things: (1) Can an outcome (dependent) variable be forecasted using a set of predictor variables (2) Which variables are significant predictors of the outcome variable, and how do they influence it (as indicated by the size and sign of the beta estimates)? These regression estimations show how one dependent variable interacts with one or more independent variables.

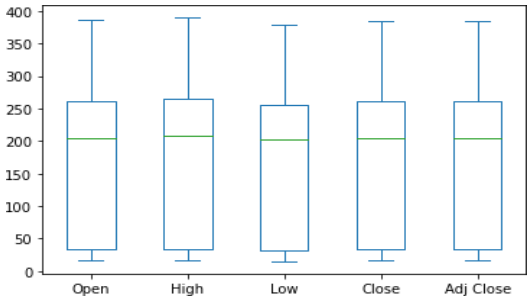
The most basic version of the regression equation with one dependent and one independent variable is y = c + b * x, where y is the estimated dependent variable score, c is the constant, b is the regression coefficient, and x is the independent variable score. There are several names for the dependent variable in a regression. The dependent variable in a regression is known by a variety of names. It is also referred to as a regressand, criterion variable, endogenous variable, or result variable. Exogenous variables, predictor variables, and regressors are all terms for independent variables.

Regression analysis has three main applications: assessing predictor strength, anticipating an impact, and forecasting trends.
The Equation for linear regression with values is depicted as:

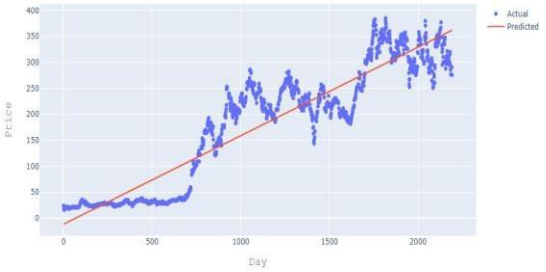
$$Y = a + bX$$

$$a = \frac{[(\sum y)(\sum x^2) - (\sum x)(\sum xy)]}{[n(\sum x^2) - (\sum x)^2]}$$

$$b = \frac{[n(\sum xy) - (\sum x)(\sum y)]}{[n(\sum x^2) - (\sum x^2)]}$$



The final plot earned by using functions in python is given in figure



c) Performance Evaluation Methods

The r2 score is defined as "(total variance explained by model) / total variance." If it is 100 percent, the two variables are completely linked and have no variance. A low number indicates a low level of correlation, implying that the regression model is not always valid.

The mean square error is the average of the squares of the errors (MSE). The greater the number, the greater the error. The difference in observed values y1, y2, y3,... as well as the projected values pred(y1), pred(y2), and pred(y3) (y3),... In this case, the term "error" applies. To avoid cancelling out negative and positive values, we square each difference (pred(yn) – yn) ** 2.

From the analysis the values are as follows:

Metric	Train	Test
r2_score	0.8658871776828707	0.8610649253244574
MSE	1821.3833862936174	1780.987539418845

Conclusion and Future work

Predicting stock market returns is a tough task because inventory values are constantly converting and are dependent on multiple parameters that shape complicated styles. The historical dataset available on the enterprise's website consists of only some functions inclusive of high, low, open, close, adjoining near price of inventory prices, quantity of shares traded, and so on, which can be insufficient. to improve the accuracy of the expected fee, new variables had been created via combining current variables.

IV. References

[1] Masoud, Najeb MH. (2017) "The impact of stock market performance upon economic growth." International Journal of Economics and Financial Issues 3 (4): 788–798.

[2] Murkute, Amod, and Tanuja Sarode. (2015) "Forecasting market price of stock using artificial neural network." International Journal of Computer Applications 124 (12): 11-15.

[3] Li, Lei, Yabin Wu, Yihang Ou, Qi Li, Yanquan Zhou, and Daoxin Chen. (2017) "Research on machine learning algorithms and feature extraction for time series." IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC): 1-5.

DOI:10.5281/zenodo.6365072
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam