Amazon Stock Price Prediction Using Machine Learning Techniques

Navjot Kaur[1]
College of Computing
Michigan Technological University
Houghton, MI, USA
navjotk@mtu.edu

Prathamesh Jadhav[2]
College of Computing
Michigan Technological University
Houghton, MI, USA
pjadhav2@mtu.edu

Surya Ravula[3]
College of Computing
Michigan Technological University
Houghton, MI, USA
sravula2@mtu.edu

***Abstract*—The COVID-19 has changed the capital market sentiments for investors and companies around the world. The falling stock prices have made investors' situation more stressful before investing and pulling back the invested stocks. While the world fought with drastic fluctuations of capital markets, the eCommerce companies got a competitive advantage. Amazon's stock prices are outperforming its peers during pandemic hit. In this paper, we described the use of machine and deep learning techniques to anticipate Amazon's stock price. The aim of this experiment is to identify a suitable automated model to help Amazon investors to make their decision based on tomorrow's prediction. The model has been built using four popular regression machine learning algorithms, and recurring neural network's LSTM algorithms. The model performance matrices such as R-Squared, MSE, and MAE provide the comparison selecting the suitable model. Ten years of stock price data of Amazon.com has been scrapped from the S&P 500. The linear regression performed better among the resulted five models with reduced MSE and RMSE.**

Keywords—Amazon, Stock Price, Stock Price Prediction, Machine Learning, Deep Learning, Random Forest Regressor, Decision Tree Regressor, XGBoost, Long Short-Term Memory Network, RMSE, MSE, R-Squared

I. Introduction

The stock markets are the backbone of economic growth. They provide a common platform for people to invest in the different economies of the world. These markets explain the health of an organization or a country. The stock price direction affects the investors' behavior in terms of buying or selling the stocks. Various factors affect the stock price fluctuations such as economic, political, company's financial performance, and company's chair. Recently, there is a substantial change in the stock markets because of COVID-19. The market saw a tremendous economical lockdown where physical markets faced challenges to achieve sustainability. The current markets have given rise to more research study in stock markets and being a reason become a well-liked topic for researchers.

Historically, the stock price predictions used to be performed manually using statistical analysis on stock price data and qualitative analysis on the company's goodwill, location, product domain, and chairperson status. The continuous development in machine learning, robust cloud databases, and powerful computers simplified the task to develop efficient models to predict stock price. Automated systems not only just benefit the stock market brokers or organizations, but also to individuals to easily understand the trend of the high or low price to make their financial decisions. Everyone focuses on profit maximization from their investment and reduces the risk of losing money. The highly unpredictable nature of the stock prices makes it difficult for individuals or companies to keep track of the fluctuations and take an appropriate maneuver. The proposed system can settle down the problem and help to determine the next day's stock price in lesser time.

Amazon has been selected for the experiment to implement the system for the investors of e-commerce and cloud services giant. Amazon has always been one of the top retail companies in the stock market. During the pandemic, the sales of Amazon rapidly rocketed because the buyers transited to buy products online.

We aimed to get ten years of stock prices daily data about the stock prices use the machine and deep learning algorithms to build a simplified model. The problem has been considered a regression problem. The concentration is to determine the future value or amount of the stock and predict it.

II. Background and Motivation

Many research studies have been performed in the field of studying the capital markets and their behavior, while the rise of automated models gave a new platform to these studies. An intelligent trading system usually picks up the demand to meet the uncertainly changing trends. For the first time, Qiu J, Wang B, and Zhou C introduced the LSTM model of Recurring Neural Networks to study the stock market prices. While only regression or classification models cannot consider the non-linear and time-series data. They addressed this unpredictive and volatile nature of the stock markets using LSTM[6].

In one study of "Stock Closing Price Prediction using Machine Learning techniques", the authors, Vijh, Chandola, Tikkiwal, Kumar suggested Artificial Neural Network and Random Forest Approach predict the future stock price of five different companies and predict the price[1]. Penglei Gao, Rui Zhang, and Xi Yang discussed the Neural Network application to predict the stock index price in this [2] work. They described the benefits of using artificial neural networks to explain the non-linearity of the stock prices due to noise. Krollner et. al. studies the time-series nature of the financial markets using machine learning techniques. The explained the use of ANN using evolutionary optimization techniques in forecasting the stocks[3]. The short-term stock price trends are also a popular area of study. Shen, J., and Shafiq followed a comprehensive approach to study short-term trends[4]. These studies show the role of emerging artificial intelligence in changing the predictive models. In the year 2020, Daniel Štifanić et. al. conducted a study to check the influence of COVID-19 on stock price forecasting using bidirectional recurring neural networks[5].

Our study has identified the need for a machine learning model based on the current market fluctuations. Especially, the e-commerce boom during the pandemic. Amazon is performing excellent in this segment and we developed an approach to develop a predictive model for this company. The interesting research findings from previous works motivated us to build a comprehensive system that can generate accurate predictions. The study has included the Regression and Recurring Neural Networks along with detailed evaluation criteria to validate the model performance. The study has been considered Regression because of the continuous nature of the predictor variables. In the end, we also included the comparison of the five resulted models and presented our thought to select one go-to model for the implementation.

III. Methods

The proposed solution includes the various Machine and Deep Learning algorithms and techniques to fit the best performing model.

Linear Regression: It is a statistical way of check the linear or non-linear relationship between one or more dependent and independent variables. The linear regression work based on the linearity, independence, homoscedasticity, and normality of all the independent variables. It indicates the effect of the predictor (X) variable(s) on the response (Y) variable. The linear regression equation is.

$$y = \theta^{T} + \epsilon$$

Where y = response variable

Θ = coefficients of independent variables where linear combinations are:

$$\sum_{i=1}^{p} \theta_i \, x_i$$

E = (epsilon) error terms

Linear regression assumes that the error terms are normally distributed N(0, σ^2).

Process of conducting linear regression analysis

1. Analyze the Correlation of the data
2. Check the distribution of the data
3. Estimate the linear model by fitting the straight line
4. If the model indicates the high deviation of the error terms, then minimize the cost function.
5. Refit the model
6. Evaluate the model and perform model selection.

Decision Tree: This is a popular model to solve classification and regression models. It forms a tree structure form containing various nodes to solve the problems. For the classification problem, it takes the input and predicts in terms of 0 and 1. For the continuous input, it predicts the probability of the continuous response variable.

The algorithm is popularly known as a non-parametric algorithm that does not make any assumption of the data in the space. This algorithm answers a long list of questions just like a human brain check various conditions to conclude.

Process of Conducting Decision Tree Regressor

1. The data will be in the parent node that gets split into further branches to make a decision.
2. The next nodes are called decision nodes where the algorithm makes the decision in terms of True/False.
3. The decision nodes further get divide into two types of nodes, if the decision has been made then it will further the Terminal node which is also called a leaf node. Otherwise, the decision node will split into a further decision node.
4. The iteration goes on until one terminal node reached that is also called as child node.

To use the Decision Tree for regression problems, the model takes into consideration the standard deviation reduction rather than information gain using entropy or GINI index. The primary challenge with the Decision Tree algorithm is model overfitting. As the tree gets trained to its full depth the model overfits the data. Tree pruning techniques and Hyperparameter tuning can address this challenge.

Random Forest Regressor: This algorithm is also called as an ensemble algorithm and is widely used to solve classification and regression problems. It works well against the model overfitting by using bagging and boosting techniques.

The random forest creates multiple decision trees and based on the results of each decision tree, it takes the majority vote count (for classification) and means of the results (for regression).

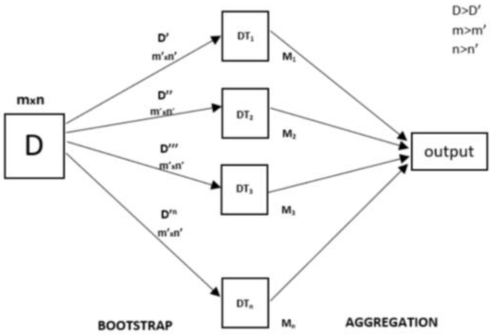

Figure 1: Process of Random forest Regressor

From figure 1, the detailed process of the random forest regressor has shown.

1. Input = D (data)
2. Repeat the steps K times
3. Randomly selects the subset trees before the split.
4. Fit each tree on the data
5. It does not perform tree pruning at an early stage.
6. Then computes the average results to give the new prediction.

XGBoost Regressor: Similar to the random forest algorithm, XGBoost is also an extensively used machine learning model for classification and regression problems. It works in the same fashion as the random forest does, but it performs the gradient algorithms to minimize the loss function.

This is also called an ensemble algorithm. Usually, decision tree train the model once and then make predictions. That likely to be an overfit model. The random forest overcomes the overfitting using ensemble technique but this algorithm does not perform the tree pruning at an early stage. While the XGBoost helps to overcome these all challenges by using ensemble technique along with gradient boosting at each level of creating a new decision tree. That means each subset decision tree first corrects its previous mistakes by minimizing the cost function. Then it finally gives the prediction based on the average of each subset tree.

Some benefits of XGBoost

1. Regularize the parameters
2. Control on data sparsity
3. Parallelize the process of sub-setting the trees

Recurring Neural Networks: This is also a deep learning neural network, but variant from the traditional feed-forward principle-based deep learning networks. These networks process the data in each layer, then the output of each layer gets treated as input of the next layer. These networks are popularly used in natural language processing models.

Long-Short Term Memory (LSTM) is a type of recurring neural network. It is a highly potential algorithm that can learn well in the long run by utilizing the learnings from the past. Long term learning utilization is also called "Long-term Dependencies". LSTM creates a long chain of networks that are input-output-input for each layer. Historically, RNN used to be built on one layer, but later the idea of using multiple layers in RNN has been introduced. Therefore, the long-short term memory network has used four layers, where each state-run horizontally [2]. Each layer has a cell state (C_t). The network can alter the state of each cell in the layer using the Gate system.

Notation of the gates is $\sigma(W_i X + b_i)$ and the sigmoid layer's output is between 0 and 1, where 0 means no pass allowed and 1 means the allowed passes. Then the three layers input, output, and forget layers regulate the information flow.

Evaluation Methods

R-Squared: It explains the proportion of variability in the response variable (Y) explained by the predictor variable (Xi). It is also sometimes addressed as the Coefficient of Determination. That means that if we draw a linear line from the axis, how many data points will be close or far from the line. In the regression analysis, the data should be on the best fit line, and considering the noise in the data, the data points should have a minimum deviation from the line. The successful model has a minimum cost function and explains higher variability. Generally, the R-Square measures between 0 – 1. Closer to one indicates a good fit of the model. The R-Square can be calculated as below.

$$R^2 = 1 - \frac{RSS}{TSS}$$

Where RSS = sum of squared residuals

TSS = total sum of squares

Mean Squared Error (MSE): It indicates the deviation of the regression line from data points that are unobserved. This is a type of distance measure therefore always resulted in a positive value. The distance between the error points calculates then squared and at last, their average gets calculated to check the errors. This value should be low possible because higher error indicated the model is not fitting correctly.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{Y_\iota}) \text{\^{}}2$$

Where, n = total number of data points

Yi = Observed values

Y (hat) = predicted value

Root Mean Squared (RMSE): It is the root of mean squared errors to check the standard deviation of the data points from the regression line.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{Y_\iota}) \text{\^{}}2}$$

Mean Absolute Error (MAE): This measure explains the absolute standard deviation of the predicted values and expected value along the best fit line.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

IV. Experimental Design

We designed a simple architecture for our experiment. The experiment includes the dataset information, data preprocessing, exploratory data analysis, application of methods and algorithms, and resulted in models and their comparison.

Data Description

We used historical stock price data of Amazon.com (AMZN) for ten years. The data have been scrapped from S&P 500 using an appropriate scrapping method. The imported data includes records of all registered 500 companies from 2010 – 2020. Out of which the Amazon records have been selected. The description of each variable in the data set that we are going to use is as below.

Variable Name	**Names in the Dataset**	**Variable Type**
Date	Date	Datetime
High	High	Continuous

Low	Low	Continuous
Open	Open	Continuous
Close	Close	Continuous
Volume	Volume	Continuous
Adjusted Close	Adj_Close	Continuous
Moving Averages	Moving_avg	Continuous
Increase in Volume	Increase_in_Vol	Continuous
Increase in Adjusted Close	Increase_in_adj_close	Continuous
Year	Year	Datetime

Table 1: Variables and their Types

For the current study, we added three new variables: (1) moving averages, (2) increase in volume, (3) increase in the adjusted close. These variables will help us to check the variability of the stock prices. We also added "Year" column by extracting year from the "Date" column.

Moving averages has been calculated by taking mean of the adjusted close. Increase in Volume has been calculated by subtracting the volume from 2020 to 2019. It will explain the yearly change in the volume. Similarly, we checked the increase in adjusted close by subtracting the adjusted close from year 2020 to 2019. These calculations will explain the actual fluctuations. The final dataset which we are going to use contains 2769 rows and 8 columns.

Overview Architecture

The below diagram shows the flow and overview of our work.

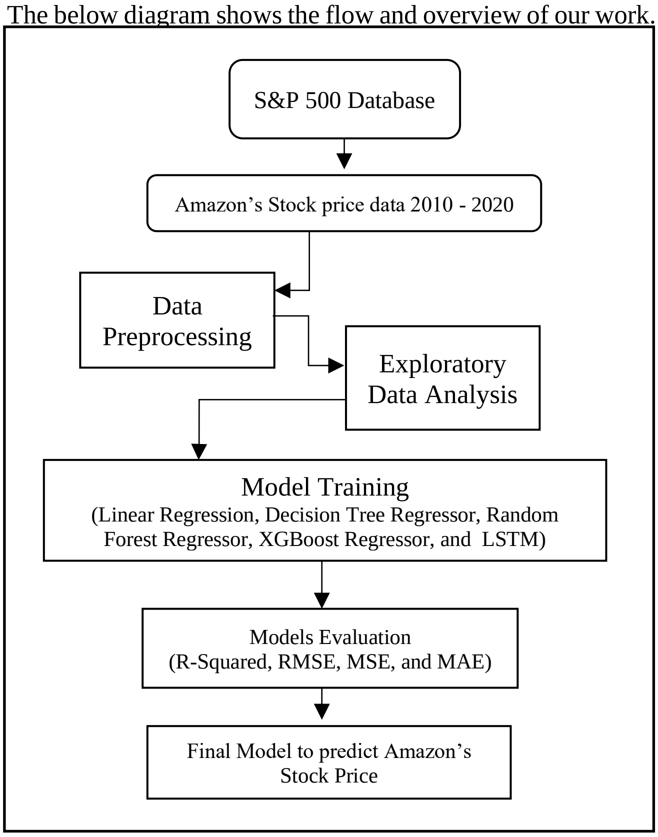
S&P 500 Database
Amazon's Stock price data 2010 - 2020
Data Preprocessing
Exploratory Data Analysis
Model Training
(Linear Regression, Decision Tree Regressor, Random Forest Regressor, XGBoost Regressor, and LSTM)
Models Evaluation
(R-Squared, RMSE, MSE, and MAE)
Final Model to predict Amazon's Stock Price

Figure 2: Overview Architecture

We followed this basic architecture to implement our experiment.

Data Preprocessing: We first scrapped the data from the data source (S&P 500) using HTML parsing query to connect to the S&P 500 database. Then, we filter the Amazon's stock price data from 2010- 2020. Then we checked the need of the data cleaning and feature engineering. To verify these, we applied some basic techniques on our data.

We checked the data for the any type of missing values (NA, NULL, NAN) but the data was cleaned. Then we checked the correlation matrix to identify the correlation among each variable. Then we added the four new variables to the data as they were required to identify the patterns during our experiment. Before splitting the data for training and validating our models, performed some exploratory data analysis.

Exploratory Data Analysis: The basic exploratory data analysis has been performed to explore the data we have used for the experiment. These analysis help to identify the pattern, distribution, skewness, and trends based on time.

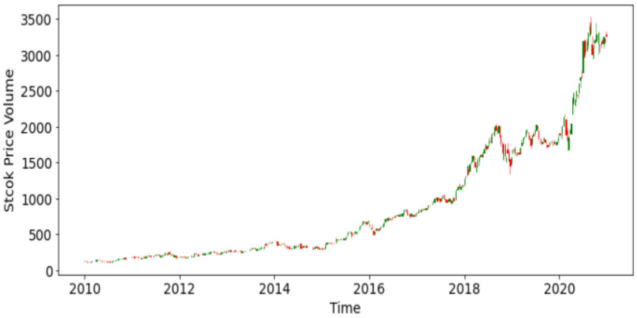

Figure 3: Amazon's Stock Price Volume Trend against Time

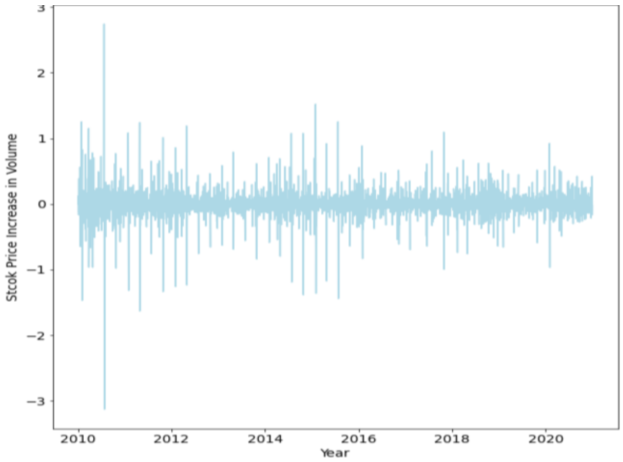

Figure 4: Amazon's yearly trend of Increase in Volume

The figures 3 and 4 are indicating the yearly trends in the volume of Amazon's stock price. Figure 3 indicates that the volume of the stocks highly increased from the year 2010 to 2020. It indicates that people have shown more interest in the Amazon's stocks. Whereas the Figure 4 indicates the yearly increase and decrease in the volume.

In the next process, we split our dataset into training (80%) and testing (20%). The shape of the training and testing set is 2104 and 526 respectively.

V. Results

We perform the experiment on five different regression models and build models. We have built Linear, Decision Tree, Random forest, XGBoost, and Long-Short Term Memory regression models. In this section, we have described the models along with their performance matrices.

Model 1: Linear Regression Model

Evaluation Matrices	Results
RMSE	5.54
R-Squared	0.99
MAE	2.92
MSE	30.77

Table 2: Performance Matrices of the Linear Regression Model

The model indicates that the predictor variables Open, High, Low, Volume, Increase in Volume, and Increase in Adjusted Close is explaining 99% variability in Closing price (Y). The MSE and RMSE are explaining the standard deviation of the error terms from the fitted regression line is 30.77 and 5.54, respectively. While MAE indicates the absolute value of predicted and expected values is 30.77.

Model 2: Decision Tree Regressor

We executed the Decision Tree Regressor twice with two different types of values of max depth. We set the max depth to control the model overfitting.

Evaluation Matrices	DT max depth 5	DT max depth 7
RMSE	26.88	14.17
R-Squared	0.99	0.99
MAE	18.83	8.01
MSE	722.82	200.78

Table 3: Performance Matrices of the Decision Tree Model

We found the Decision Tree Regressor is fitting the model and shows that 99% variability of response variable has been explained by the predictor variables. While the predicted values are showing high deviation from the regression line. When we tried the different values of max depth the errors values decreased significantly. The model is unable to perform well as compared to linear regression model.

Model 3: Random Forest Regressor

We executed the Decision Tree Regressor twice with two different types of values of max depth. We set the max depth to control the model overfitting.

Evaluation Matrices	Performance of model 3
RMSE	10.69
R-Squared	0.99
MAE	5.25
MSE	114.38

Table 4: Performance Matrices of the Random Forest Model

The random forest regressor performed better by reducing the error rate that we can the MSE and RMSE significantly reduced. The model tells that 99% variable in closing price can be explained by the independent variables.

Model 4: XGBoost Regressor

Evaluation Matrices	Results
RMSE	10.34
R-Squared	0.99
MAE	5.93
MSE	106.94

Table 5: Performance Matrices of the XGBoost regressor Model

XGBoost model reduced the mean squared error and root mean squared error significantly. The model also explains 99% variation in the data. The Absolute value is higher in the XGBoost model.

Model 5: Long-Short Term Memory

Evaluation Matrices	Results
RMSE	76.22
R-Squared	0.69
MAE	58.81
MSE	5810.53

Table 6: Performance Matrices of the LSTM Model

Comparison of the Model sand Model Selection

Based on the results of all the models, we can see the LSTM model is not performing well on the current data. The Linear regression model performed better than other four models as it is explaining the more variability than the other models; as well as it is giving less error rate against the regression line. The Decision tree regressor overfitted the model, but relatively improved when depth size was increased. With the current experiment we can choose linear regression as go-to model. And we can consider the optimization of the our other models in the future work.

Conclusion

Stock Markets are the driving factor of an economy that impacts the companies and their market value. In this paper, machine and deep learning have been addressed to build a framework for Amazon investors to predict future prices. Amazon investors can get benefitted by making appropriate decisions regarding their investments. The study utilized the ten years of historical stock price data to predict the future price. Five different models have been built and compared based on their performance criteria. R-Squared, RMSE, MSE, and MAE have been used to evaluate the model performance. We were optimistic about the performance of the LSTM model, but the model did not improve its performance. From our background work, we found that the LSTM can fit in this problem statement. This challenge is always being considered while studying the uncertain capital markets. We have selected the linear regression model as a base model and for future consideration, we have selected the

LSTM model. Future work can address these challenges and evolutionary optimization techniques to improve the predictions. We planned to include more companies of S&P 500 index to build a robust system.

Author Information and Contributions

1. Navjot Kaur: Student of Master of Science in Data Science at Michigan Technological University, Houghton MI, USA. Author Conrtributed in Experiment Design and Methods Specifications, wrote first draft of the document, implemented the XGBoost and Document review and editing.
2. Prathamesh Jadhav: Student of Master of Science in Data Science at Michigan Technological University, Houghton MI, USA. Author Conrtributed in investigating the regression methods and models, implemented the Decision Tree, Random Forest, and Linear Regression
3. Surya Ravula: Student of Master of Science in Data Science at Michigan Technological University, Houghton MI, USA. Author Conrtributed to introduce the topic, Implemented the data preprocessing, exploratory data analysis, and implementation of LSTM model and writing the respective section.

All the authors mutually reviewed and edited the paper as per requirements and declaring no any type of disagreements.

References

[1] Vijh, Mehar & Chandola, Deeksha & Tikkiwal, Vinay & Kumar, Arun. (2020). Stock Closing Price Prediction using Machine Learning Techniques. Procedia Computer Science. 167. 599-606. 10.1016/j.procs.2020.03.326.

[2] Penglei Gao, Rui Zhang, and Xi Yang, “The Application of Stock Index Price Prediction with Neural Networks”, 2020, mdpi

[3] Krollner, B., Vanstone, B., & Finnie, G. (2010). Financial time series forecasting with machine learning techniques: A survey. In *Proceedings of the 18th European Symposium on Artificial Neural Networks (ESANN 2010): Computational Intelligence and Machine Learning* (pp. 25-30)

[4] Shen, J., Shafiq, M.O. Short-term stock market price trend prediction using a comprehensive deep learning system. *J Big Data* **7,** 66 (2020). https://doi.org/10.1186/s40537-020-00333-6

[5] Daniel Štifanić, Jelena Musulin, Adrijana Miočević, Sandi Baressi Šegota, Roman Šubić, Zlatan Car, "Impact of COVID-19 on Forecasting Stock Prices: An Integration of Stationary Wavelet Transform and Bidirectional Long Short-Term Memory", *Complexity*, vol. 2020, Article ID 1846926, 12 pages, 2020. https://doi.org/10.1155/2020/1846926

[6] Qiu J, Wang B, Zhou C (2020) Forecasting stock prices with long-short term memory neural network based on attention mechanism. PLOS ONE 15(1): e0227222. https://doi.org/10.1371/journal.pone.0227222

[7] Budiharto, W. Data science approach to stock prices forecasting in Indonesia during Covid-19 using Long Short-Term Memory (LSTM). *J Big Data* **8,** 47 (2021). https://doi.org/10.1186/s40537-021-00430-0