ISSN: 2395-7852

International Journal of Advanced Research in Arts, Science, Engineering & Management

Volume 10, Issue 3, May 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 6.551

+91 9940572462 +91 9940572462 ijarasem@gmail.com www.ijarasem.com

International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 6.551 |Bimonthly, Peer Reviewed & Referred Journal|

| Volume 10, Issue 3, May 2023 |

Time Series Analysis-Based Amazon Stock Price Prediction Using Maching Learning Models

Dr.S.Soundararajan, M.E., PhD[1], Vakati Harshitha[2] , Parvathareddy Jhansi[3], Pokkireddy Varshini[4] , Chennareddy Preethi[5]

[1]Professor, Department of Computer Science and Engineering, Velammal Institute of Technology, Panchetti, Chennai, India

[2] UG Scholar, Department of Computer Science and Engineering, Velammal Institute of Technology, Panchetti, Chennai, India

[3] UG Scholar, Department of Computer Science and Engineering, Velammal Institute of Technology, Panchetti, Chennai, India

[4] UG Scholar, Department of Computer Science and Engineering, Velammal Institute of Technology, Panchetti, Chennai, India

[5] UG Scholar, Department of Computer Science and Engineering, Velammal Institute of Technology, Panchetti, Chennai, India

ABSTRACT: Stock prices are first determined by a company's initial public offering (IPO) when it first puts its shares into the market. Investment firms use a variety of metrics, along with the total number of shares being offered, to determine what the stock's price should be. Afterward, the several reasons mentioned above will cause the share price to rise and fall, driven largely by the earnings that can be expected from the company. Traders use financial metrics constantly to determine the value of the company, including its history of earnings, changes in the market, and the profit that it can reasonably be expected to bring in. Hence, stock price prediction has become an important research area. The aim is to predict machine learning based techniques for stock price prediction. The analysis of dataset by supervised machine learning technique (SMLT) using uni-variate analysis, bi-variate and multi-variate analysis. To propose a machine learning-based method to accurately predict the stock price. Proposed machine learning algorithm technique can be compared with best accuracy with precision, Recall and F1 Score.

KEYWORDS: Initial Public Offering (IPO), Supervised Machine Learning Technique (SMLT), F1 Score, machine learning-based method.

I. INTRODUCTION

As one of the world's most predominant organizations, Amazon (AMZN) possesses an unmistakable position. There has been some cooperation with an Amazon administration, whether a buyer item or business administration. On account of Amazon, a comparative venture would have created a multi-bagger return. The question remains, however, how long will Amazon's dominance last?

Amazon neednot bother with a concise presentation, however here is a fast outline of the organization's vital organizations and achievements. The organization offers electronic gadgets, clothing, furniture, food, toys, and numerous different items from different shippers. As well as giving video and music web-based, it additionally offers live real-time features.

Many clients utilize Amazon's cloud administration stage to have online applications. At the point when Amazon previously sent off, it was an internet-based bookshop. By and by, Jeff Bezos, Amazon's pioneer and previous President maintained that the organization should be something other than an internet-based store.

Data Science:

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains.

Artificial Intelligence:
Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving.

Machine Learning:
Machine learning is to predict the future from past data. Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning.

II. RELATED WORKS

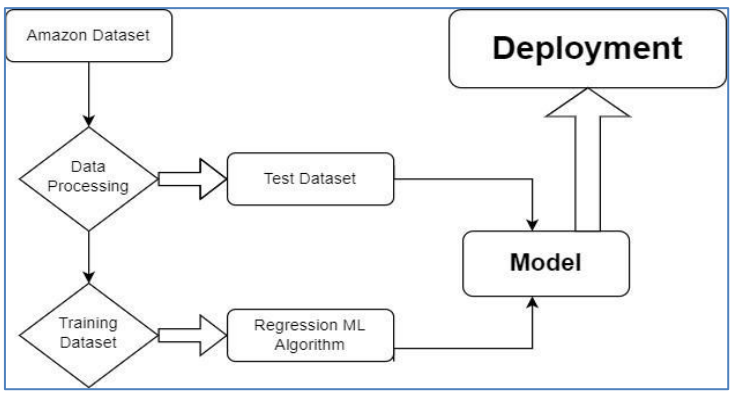
Haocheng Du In their work they researched on Amazon's stock price forecasting based on arbitrage pricing model based on big data, 2022.The generation of big data is based on the network data generated when people use Internet information systems to interact. Big data can reflect the general laws of specific fields and industries, provide more accurate references for decision makers and managers, and provide people with better Data services. Arbitrage pricing model has long been widely quoted by scholars as an alternative theory to capital asset pricing model, which is used to make a regression analysis on Amazon's stock price in this study. In our study, we aim to construct an arbitrage pricing model to make a regression analysis on Amazon's stock price, which is demonstrated to have a higher prediction accuracy and better fitting degree compared with the self-coding network.

Kevin Thomas In their work they proposed Time Series Prediction for Stock Price and Opioid Incident Location, 2019.Time series forecasting is the prediction of future data after analysing the past data for temporal trends. This work investigates two fields of time series forecasting in the form of Stock Data Prediction and the Opioid Incident Prediction. In this thesis, the Stock Data Prediction Problem investigates methods which could predict the trends in the NYSE and NASDAQ stock markets for ten different companies, nine of which are part of the Dow Jones Industrial Average (DJIA). A novel deep learning model which uses a Generative Adversarial Network (GAN) is used to predict future data and the results are compared with the existing regression techniques like Linear, Huber, and Ridge regression and neural network models such as Long-Short Term Memory (LSTMs) models.

Rashid, Abdul In their work they proposed Stock Prices and Trading Volume: An assessment for linear and nonlinear Granger causality.JournalofAsianEconomics, 18, pp. 595 – 612. 2019.Analyzing the association between returns and trading volume using the Granger causality test to the Karachi Stock Exchange data. For this study, the author finds that the causality from volume to stock price depends on the direction of the stock price movement, and he recognizes that the volume has a significant nonlinear explanatory power for stock price movement. For that reason, taking into consideration the attempt at the stock market, in this study, we are going to discuss about the relationship between stock price and the volume of e-commerce sales using a state space model.

III. PROPOSED METHOD

The proposed methodconsistof data's collected from various sources. The data will be checked for the cleanliness. The cleaned data will be generated as training and testing. The model is created by using machine learning method. Different algorithms are applied for finding the best algorithm. The best one is formed as a model. The data model is used for the prediction of the new stock price.



Architecture of Proposed Methodology

3.1 Data collection

The data set collected for predicting given data is split into Training set and Test set. Generally, 7:3 ratios are applied to split the Training set and Test set. The Data Model which was created using Machine Learning Regression Techniques are applied on the Training set and based on the test result accuracy, Test set prediction is done.

3.2 Data Pre-Processing:

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters.

Some of these sources are just simple random mistakes. Other times, there can be a deeper reason why data is missing. Before, joint into code, it's important to understand the sources of missing data. Here are some typical reasons why data is missing:

- User forgot to fill in a field.
- Data was lost while transferring manually from a legacy database.
- There was a programming error.
- Users chose not to fill out a field tied to their beliefs about how the results would be used or interpreted.

Variable identification with Uni-variate, Bi-variate and Multi-variate analysis:

- import libraries for access and functional purpose and read the given dataset
- General Properties of Analyzing the given dataset
- Display the given dataset in the form of data frame
- show columns
- shape of the data frame
- To describe the data frame
- Checking data type and information about dataset
- Checking for duplicate data
- Checking Missing values of data frame
- Checking unique values of data frame
- Checking count values of data frame
- Rename and drop the given data frame
- To specify the type of values
- To create extra columns

3.3 Data Validation/ Cleaning/Preparing Process

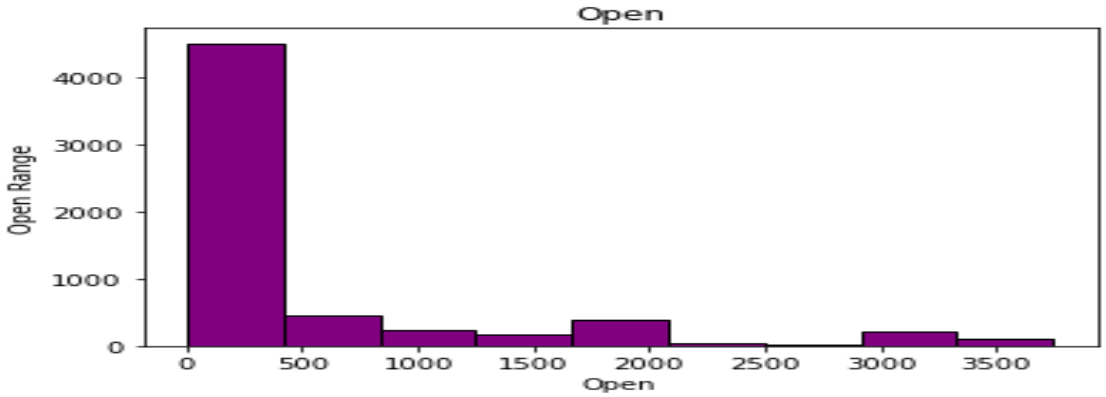
Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning models and procedures that you can use to make the best use of validation and test datasets when evaluating your models. and decision making.

	Date	Open	High	Low	Close	Adj Close	Volume
0	1997-05-15	2.437500	2.500000	1.927083	1.958333	1.958333	72156000
1	1997-05-16	1.968750	1.979167	1.708333	1.729167	1.729167	14700000
2	1997-05-19	1.760417	1.770833	1.625000	1.708333	1.708333	6106800
3	1997-05-20	1.729167	1.750000	1.635417	1.635417	1.635417	5467200
4	1997-05-21	1.635417	1.645833	1.375000	1.427083	1.427083	18853200

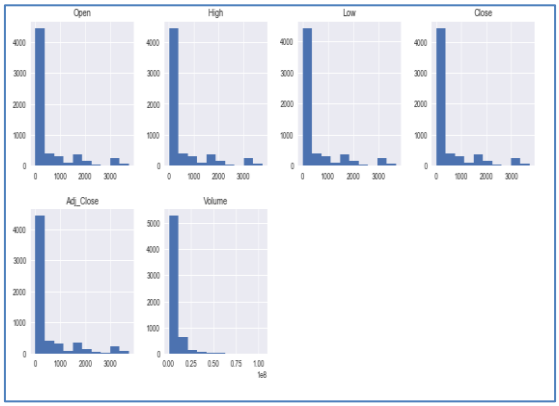
	Open	High	Low	Close	Adj Close	Volume
count	6155.000000	6155.000000	6155.000000	6155.000000	6155.000000	6.155000e+03
mean	520.556302	526.216132	514.277282	520.429832	520.429832	7.329010e+06
std	857.161696	865.821041	847.270905	856.668492	856.668492	7.149521e+06
min	1.406250	1.447917	1.312500	1.395833	1.395833	4.872000e+05
25%	38.750000	39.514999	38.104999	38.821251	38.821251	3.579350e+06
50%	92.669998	94.190002	90.750000	92.639999	92.639999	5.470000e+06
75%	528.949982	535.304993	521.950012	529.450012	529.450012	8.294950e+06
max	3744.000000	3773.080078	3696.790039	3731.409912	3731.409912	1.043292e+08

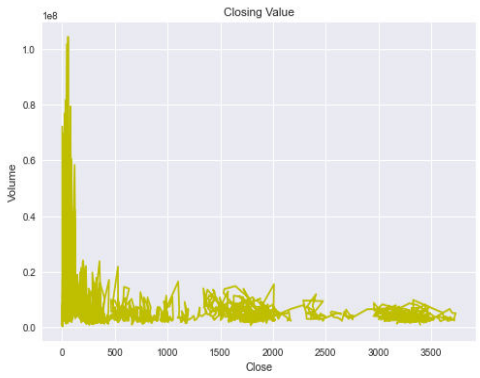
3.4 Exploration Data Analysis of Visualization

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend aa deeper dive into some the books mentioned at the end.



- How to chart time series data with line plots and categorical quantities with bar charts.
- How to summarize data distributions with histograms and box plots.





3.5 Linear Regression:

Linear Regression is a machine learning algorithm based on supervised learning. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output).

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 6.551 |Bimonthly, Peer Reviewed & Referred Journal|

| Volume 10, Issue 3, May 2023 |

```
MEAN ABSOLUTE ERROR VALUE IS: 2.757359381236852

MEAN SQUARED ERROR VALUE IS: 46.0574345755943

MEDIAN_ABSOLUTE_ERROR VALUE IS: 0.6966165052120346

ACCURACY RESULT OF LR IS: 99.99371054136492

R2_SCORE VALUE IS: 0.9999370740424633
```

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

3.6 K Nearest Neighbor

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) with some mathematics we might have learned in our childhood—calculating the distance between points on a graph. There are other ways of calculating distance, and one way might be preferable depending on the problem we are solving. However, the straight-line distance (also called the Euclidean distance) is a popular and familiar choice.

3.7 Support Vector Regression:

Support Vector Machines (SVM) are popularly and widely used for classification and regression problems in machine learning.

SVMs solve binary classification problems by formulating them as convex optimization problems. The optimization problem entails finding the maximum margin separating the hyperplane, while correctly classifying as many training points as possible. SVMs represent this optimal hyperplane with support vectors. The sparse solution and good generalization of the SVM lend themselves to adaptation to regression problems. SVM generalization to SVR is accomplished by introducing an ε-insensitive region around the function, called the ε-tube. This tube reformulates the optimization problem to find the tube that best approximates the continuous-valued function, while balancing model complexity and prediction error. More specifically, SVR is formulated as an optimization problem by first defining a convex ε-insensitive loss function to be minimized and finding the flattest tube that contains most of the training instances.

3.8 Elastic Net Regression:

Linear regression refers to a model that assumes a linear relationship between input variables and the target variable.

With a single input variable, this relationship is a line, and with higher dimensions, this relationship can be thought of as a hyperplane that connects the input variables to the target variable. The coefficients of the model are found via an optimization process that seeks to minimize the sum squared error between the predictions (yhat) and the expected target values(y).

- loss = sum i=0 to n (y_i – yhat_i)^2

One popular penalty is to penalize a model based on the sum of the squared coefficient values. This is called an L2 penalty. An L2 penalty minimizes the size of all coefficients, although it prevents any coefficients from being removed from the model.

- l2_penalty = sum j=0 to p beta_j^2

Another popular penalty is to penalize a model based on the sum of the absolute coefficient values. This is called the L1 penalty. An L1 penalty minimizes the size of all coefficients and allows some coefficients to be minimized to the value zero, which removes the predictor from the model.

- l1_penalty = sum j=0 to p abs(beta_j)

Elastic net is a penalized linear regression model that includes both the L1 and L2 penalties during training.

- elastic_net_penalty = (alpha * l1_penalty) + ((1 – alpha) * l2_penalty)

IV. CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The Best accuracy on public test set is higher accuracy score is will be find out. This

application can help out to find the Amazon Stock Price. Amazon Stock Price prediction to connect the AI Model. To automate this process by show the prediction result in web application or desktop application. To optimize the work to implement in Artificial Intelligence Environment.

REFERENCES

[1] Da Silveira Coelho, Lidiane, Rafaela Carvalho Oliveira, and Tatiana Martins Alméri. "O crescimento do e-commerce e osproblemas que oacompanham: a identificação da oportunidade demelhoriaemuma rede de comercioeletrôniconavisão do cliente." Revista de Administração do UNISAL 3.3 2013.
[2] Chen, Le, Alan Mislove, and Christo Wilson. "An empirical analysis of algorithmic pricing on amazon marketplace." Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee. 2016.
[3]Becker, K. "The impact of e-commerce and information technology in developing countries, In Krishnamurthy and Isaias" Proceedings of the IADIS International Conference eCommerce, December 7-9, Algarve, Portugal, pp. 230-235. 2007.
[4] Alves, Carla., Chaves, Renata., Penteado, Isis, and SimoneDa Costa. A Importância da Logística para o e-Commerce: O Exemplo da Amazon.com. <http://www.tecspace.com.br/paginas/aula/faccamp/TI/Texto07.pdf > 2015.
[5] Bughin, Jacques, and John Hagel III. "The operational performance of virtual communities-Towards a successful business model?."ElectronicMarkets, 10 (4), pp. 237 – 243 2000.
[6] Stone, Brad. "A loja de tudo: Jeff Bezosea era da Amazon." Rio de Janeiro, RJ: Intrínseca. 2014.
[7] Sadiku, Oladayo. "Internet Pioneers: Amazon from Start-up to Multi-Billion Dollar Business"
[8] Grego, Maurício. "Como Jeff Bezos fez fortuna con a Amazon", 2018.
[9] Berg, Nathalie. and Knights, Miya. Amazon: How the World's Most Relentless Retailer will Continue to Revolutionize Commerce. New York: Kogan Page Ltd. 2019.
[10] Donici, Andreea Nicoleta, AndreeaMaha, Ion Ignat, and Liviu-George Maha. "E-Commerce across United States of America: Amazon. com." Economy Transdisciplinarity Cognition 15, no. 1 2012.
[11] Rajgopal, Shivaram, Venkatachalam, Mohan and Suresh Kotha. The Value Relevance of Network Advantages: The Case of E-Commerce Firms. Journal of Accounting Research. Vol. 41, No. 1. 2003.
[12] Lastres, Helena MM, and Sarita Albagli. "Informação e globalizaçãona era do conhecimento."Rio de Janeiro: Campus 163. 1999.
[13] Lee, Hau L. "The triple-A supply chain." Harvard business review 82.10, pp. 102-113. 2004.
[14] Dukes, Anthony. and Liu, Lin. Online Shopping Intermediaries: The Strategic Design of Search Environments. Management Science. Articles in Advance, pp. 1 – 14. 2015.
[15] Mendes, Laura. "E-commerce: Origem, Desenvolvimento e Perspectivas." Porto Alegre: UFRGS. Available in: https://lume.ufrgs.br/handle/10183/78391.2013.
[16] Wigand, Rolf. "Electronic Commerce: Definition, Theory, and Context. The Information Society." An International Journal, 13:1, pp. 1 – 16. 1997.
[17]Vissotto, E."ComércioEletronico."Federico Westphalen: Universidade Federal deSanta Maria, ColegioAgricola de Federico Westphalen. 2013.
[18] Ramcharran, Harri. "E-commerce Growth and the Changing Structure of the Retail Sales Industry". International Journal of E-Business Research, 9(2), pp. 46 – 60. 2013.

International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| Mobile No: +91-9940572462 | Whatsapp: +91-9940572462 | ijarasem@gmail.com |

www.ijarasem.com