STOCK PRICE PREDICTION USING DATA SCIENCE TECHNIQUES

Koti Muni Yogeshswar Reddy

Department of Computing
Technologies
SRM Institute of Science and
Technology
Chennai, India
ky7113@srmist.edu.in

Medida Shivananda Reddy

Department of Computing
Technologies
SRM Institute of Science and
Technology
Chennai, India
mr8397@srmist.edu.in

Dr. D.Vinod

Department of Computing
Technologies
SRM Institute of Science and
Technology
Chennai, India
vinodd@srmist.edu.in

***Abstract*— Predicting stock price during a crucial procedure is the aim of the work.**

Predicting the stock market's behaviour is typically one of the most difficult tasks. Prediction may be viewed as one of the most crucial procedures.

This is a difficult endeavour with many unknowns. To avoid this issue in employing machine learning, one of the most fascinating (or possibly most lucrative) time series data. As a result, stock price forecasting has grown in significance as a field of study. The objective is to use machine learning-based approaches to forecast stock price projections with the maximum level of accuracy.

—

The supervised machine learning technique (SMLT) is used to analyse a dataset in order to gather a variety of data, including parameter identification, single analysis, bivariate analysis, multivariate and missing data. The whole given dataset will be subjected to value treatments, data analysis, data cleaning/preparation, and data visualisation.

***Keywords*—SMLT, Data Science**

I. Introduction

In the multidisciplinary subject of data science, information and insights are derived from both organised and unstructured data using scientific methods, procedures, algorithms, and systems. Then, across a wide range of application domains, these insights and expertise are put to use. In order to give computer science a new name, Peter Naur created the phrase "data science" in 1974. The first meeting with a distinct data science focus took place in 1996 at the International Federation of Classification Societies. On the phrase, however, there was still debate. The phrase "data science" was first used by Patil and Hammerbacher, the creative directors of LinkedIn's and Facebook's data and analytics departments. for the first time in 2008. After 10 years, it has become one of the most well-liked and in-demand professions. The academic field known as "data science" combines subject-matter experience, programming aptitude, and competency in mathematics and statistics in order to extract relevant insights from data. Data science is an interdisciplinary discipline that brings together machine learning, business savvy, technology, and mathematics to assist in uncovering hidden patterns or insights from the raw data that may be very useful in the formulation of important business decisions. Data scientists consider which problems must be addressed as well as how to find the necessary data. They have analytical and business acumen and can gather, clean, and display data. Businesses utilising data scientist source, manage, and analyse vast volumes of unstructured data. Programming skills in Python, SQL, Scala, R, Java, and MATLAB are prerequisites for a data scientist.

- Clustering, classification, and natural language processing are all examples of machine learning.

- Data visualisation libraries for Tableau, SAS, D3.js, Python, Java, and R.

Oracle, Microsoft Azure, MongoDB, and Cloudera are some big data platforms.

Artificial intelligence is the process of creating intelligent beings that mimic human behaviour and thought processes (AI). In 1956, the study of artificial intelligence became a recognised academic discipline. Since then, it has seen a number of waves of optimism, disappointment, and funding cuts (known as "AI winter"), followed by new tactics, successes, and further investment. As AI research has progressed, a variety of approaches have been tested, including formal logic, emulating animal behaviour, using vast knowledge bases, and human problem-solving. The area has been dominated by statistical machine learning with a significant mathematical component throughout the first two decades of the twenty-first century. Both industry and academia have adopted this approach to address several challenging issues. The several AI research fields are concentrated on certain goals.

II. STATE OF THE ART (LITERATURE SURVEY)

[1] Title: Predicting Fresh Produce Market Prices Using Deep Learning

Authors include Muhammad Saad, Ifeanyi Emmanuel Okwuchi, and Lobna Nassar. Due to the biggest mean absolute percentage error, the ARIMA model performs the poorest when compared to the conventional ML models (MAPE). Gradient Boosting (GB), which has the lowest MAPE error among traditional approaches, is also the best. Last but not least, for two FP, the LSTM basic DL model outperforms all examined standard ML models (Watermelon and Bok Choy). This is due to the fact that there are less markets for these two FP, giving us data with a structure more akin to a time series. Additionally, it is found that the ATTCNN-LSTM compound DL model outperforms the ML and basic DL models in terms of price prediction accuracy, especially for small sample sizes when it includes attention.

[2] The title: For the prediction of stock price, an advanced fuzzy neural deep hybrid Hammerstein-Wiener Network is used.

Authors include Xie Chen, Chai Quek, and Deepu Rajan

In this work, a deep hybrid fuzzy neural Hammerstein-Wiener model has been suggested (FNHW). The fuzzy rule base developed throughout training serves as the foundation for neuro-implications fuzzy's and inferences. The training data must be able to faithfully depict the complete system's behaviours. The distribution shift of the time series domain, however, may alter the test data. Furthermore, although the training data may be acquired from major data changes in certain situations, such as a financial crisis, the test data may be derived from continuously changing constant-state data. On steady-state data, the neuro-fuzzy system achieves sound rule base inference, and on dynamically changing data, it inherits the strong asymptotic tracking and good approximation accuracy benefits of the Hamerstein-Wiener model. Utilizing information from two financial stock price prediction datasets, the recommended method's efficacy is assessed. Decision tree is easily over-fitted, creating an excess of branches, and it may represent abnormalities caused by noise or outliers. In trials using two different datasets for financial stock price prediction, we discovered that our model significantly outperformed state-of-the-art neuro-fuzzy systems.

[3] Title: Stock Price Prediction Using Sentiment Analysis

Authors: Min Chen and Rubi Gupta

Analysis of StockTwits data with a focus on understanding how sentiments affect changes in stock prices. The following work will be further improved, according to their plans. First, we use bullish (positive) and bearish (negative) attitudes in this work (negative). By introducing neutral sentiment, the work's accuracy might be improved while noise is reduced. Second, we can only analyse five businesses in our investigation. An expansion to a larger group of businesses or to the entire StockTwits data set may produce greater insights into the data and more useful applications in stock price prediction. The optional sentiment labels provided by StockTwits users are used as the foundation for the model training. The prediction of stock price movement is improved by using sentiment data in addition to past stock time series data. It is demonstrated the usefulness of the proposed work in forecasting stock prices by doing tests on five different companies. People are increasingly adopting StockTwits, a relatively new microblogging platform, to express comments and ideas regarding stocks and the financial markets. Solid proof that the use of attitudinal data increases the forecasting accuracy of stock price changes.

[4] Title: Using an attention mechanism, predict financial big data stock trends

Authors: Feifei Kou, Du Junping , and Jiannan Chen

Historically, the main focus of study in the field of financial big data has been stock trend prediction. Because stock prices fluctuate, sophisticated nonlinear data on stocks is dynamic and evolves over time. This study offers a big data financial analysis. Using stock data characteristics, an algorithm for predicting stock trends with attention is shown (STPA). For the purpose of capturing the long-term time dependency of data, we use attention mechanisms and bidirectional gated recurrent units (BGRU). Three tiers make up the three-layer, attention mechanism-based STPA reduction technique that is advised. The shifting patterns in financial big data stocks can be predicted using STPA. The Bidirectional Gated Recurrent Unit design is used by STPA, which also introduces attention mechanism technology. In the financial stock price data set, the STPA technique surpasses the widely used algorithms in forecasting stock movements, proving the viability of the proposed strategy. The experimental results show that the suggested technique performs better at forecasting stock movements in the financial stock price data set than the routinely employed mainstream algorithms.

[5] Title: Using Principal Component Analysis to Identify Key Drivers of Stock Price in the Consumer Goods Sector

Authors: Rahma Firsty Fitriyana, Brady Rikumahu, and Andry Alamsyah are the authors.

Stock price is a crucial element in reaching By connecting a stock's price to its influencing elements, one can typically estimate the return on investing in stocks. The problem is that there are several factors that may be utilised to forecast stock values, which makes it difficult for a potential investor to choose which variables to use. Without losing information, this study examined data from five different firms to apply the principal component analysis as a dimension reduction approach to discover the important variables that influence stock prices. By adding other variables, such as macroeconomic conditions and additional financial ratios, the analysis method can be utilised to determine the primary drivers of stock prices. There are a number of factors influencing stock prices, including macroeconomic issues that were left out of this study. These variables may be examined in future research to determine how they affect stock prices.

III. PRPOSED WORK

A. Stock Forecast Exploratory Data Analysis:

After combining a large number of datasets from multiple sources to produce a generalised dataset,the most accurate results might be drawn by using a range of machine learning algorithms to spot trends.

B. Data Manipulation:

The information will be input into this section of the report, verified as accurate, and then ready for analysis by being trimmed. Ensure that you properly track your cleaning decisions and provide justifications.

Data gathering:

To forecast the supplied data, a training set and a test set of the obtained data are created. For uniformity, the training set and test set are frequently divided 7:3 apart. The training set is used to apply the machine learning-created data model, and the forecast for the test set is determined by the precision of the test findings.

Construction of the categorization model:

- The justifications provided below clarify why the prediction model for the stock problem using the decision tree method is successful: It gives better solutions to the classification problem.
- It causes out of bag estimate error, which has been shown in several studies to be neutral and is fairly simple to fix.

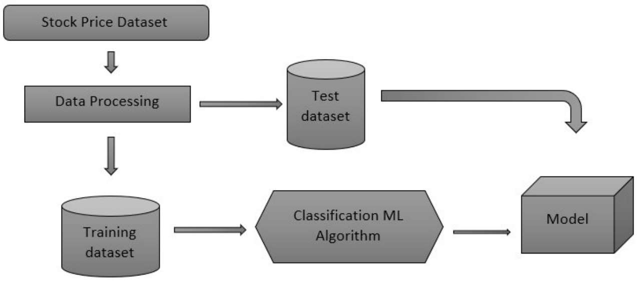


Fig: Architecture of the proposed model

The exploration of the application of machine learning algorithms for stock price prediction under operational situations is the focus of these reports, which also emphasise some remarks on the concerns, challenges, and demands of future research.

IV. Implementation

A supervised learning method known as classification employs data input from the user to teach the computer programme how to classify incoming observations. This data set may be multi-class or it may just be bi-class (for example, indicating whether the individual is male or female or if the message is spam or not). Recognition of speech, handwriting, biometric identity, categorization of documents, etc. are a few instances of classification issues. Algorithms are taught by supervised learning using labelled data. After gaining a comprehension of the data, the algorithm decides which label fresh data should receive based on patterns and associations with the unlabeled new data.

USED PYTHON PACKAGES:

sklearn:

I. A few machine learning algorithms are included in the Python machine learning package called Sklearn.

II. In this section, we make use of some of its modules, including accuracy score, train test split, and decision tree classifier or logistic regression.

NumPy:

III. NumPy is a numerical Python package that offers quick math operations for calculations.

IV. It can read data from Numpy arrays and be used to manipulate data.

Pandas:

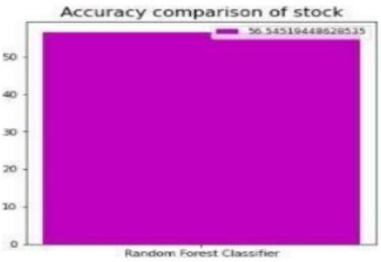
V. Applied to read and write various file types.

VI. Data manipulation is made easier by data frames.

Logistic Regression:

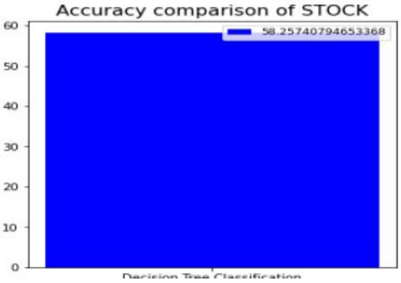
It is a statistical approach used to examine a collection of data when the outcome is influenced by one or many independent factors. When measuring the output, a dichotomous variable is utilised, and there are (two possibilities are feasible). Logistic regression looks for the model that best fits the data in order to effectively depict the link between a set of independent (predictor or explanatory) components and a relevant dichotomous feature (dependent variable = response or outcome variable). When a rule is learned, the tuples that it applies to are eliminated. On the practise set, this procedure is continued until a termination requirement is satisfied. In logistic regression, the target variable is a binary variable with data coding as 1 (true, success, etc.) or 0 (false) (false, failure, etc.)

RANDOM FOREST: For classification, regression, and other tasks, the ensemble learning technique known as random forests—also known as random choice forests—is utilised. In order to get the class that reflects the mean of the predictions produced by each individual tree (in the case of regression) or they produce a big number of decision trees throughout the training phase, which is the mode of the courses (in the case of classification). Random choice forests combat decision trees' tendency to overfit their training set. Random forest is a supervised machine learning method based on ensemble learning. The approach creates a forest of trees by using several techniques of the same kind or different decision trees, thus the name "Random Forest". The random forest method has applications in both classification and regression.It uses a formula called "Random Forest" that mixes equations from several decision trees of the same kind. As a result, a forest of trees is created. Applications of the random forest approach include both classification and regression. If a regression issue arises for a brand-new record, each tree in the forest forecasts a value for Y. By summing together all of the forecasts made by each tree in the forest, the total number may be determined. Each tree in the forest can foretell which category the new data will fall into in the event that classification proves to be difficult. The most popular category is eventually awarded the new record.



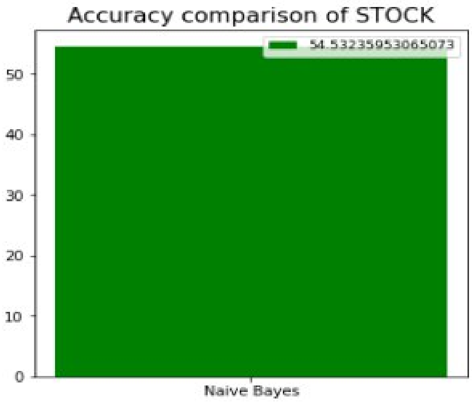
• DECISION TREE CLASSIFIER:

This approach is both one of the best and most popular. The decision-tree algorithm is a member of the supervised learning algorithms category. It functions with categorical and continuous output variables. Through the use of a tree structure, a decision tree develops classification or regression models. It segments a data collection into ever-smaller pieces while gradually building an associated decision tree. A leaf node, which contains two or more branches and represents a classification or judgement, is used to symbolise a decision node. Decision trees can be used to handle categorical and numerical data. It uses a set of exhaustive and mutually exclusive if-then rules to categorise data. The rules are successively learnt using one training batch of data at a time. Each time. A decision tree may reflect anomalies brought on by noise or outliers and is readily over-fitted, producing an excess number of branches.



NAÏVE BAYES ALGORITHM:

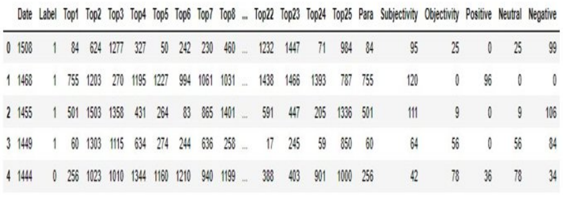
The Naive Bayes approach makes predictions by using the probabilities that each attribute belongs to each class. It is the supervised learning strategy that you would employ to probabilistically resolve a predictive modelling issue. The naive bayes technique assumes that the likelihood of every attribute belonging to a particular class value is independent of all other features, simplifying the computation of probabilities. Despite the fact that this is a significant assumption, the technique it produces is straightforward and efficient. The conditional probability is the likelihood that an attribute value will result in a class value. By summing the conditional probabilities for each characteristic for a given class value, we may determine the chance that a data instance belongs to a particular class. We can calculate the probability that an instance will belong to each class in order to produce a forecast. The class value with the highest likelihood can then be chosen. One of the simplest supervised learning methodologies is this one.The Naive Bayes algorithm is a dependable, quick, and accurate one.. Naive Bayes classifiers function rapidly and precisely on huge datasets. The Naive Bayes classifier works on the premise that the effects of various characteristics on a class are independent of one another.



V. RESULTS DISCUSSION

DATA PRE-PROCESSING/CLEANING:

loading the specified dataset while importing the library packages. To analyse the data type, shape, and size in relation to the duplicate values, missing values, and variable identification.A validation dataset is a sample of data that was not included in model training and is used to assess model competency while completing models. Methods for making the most of validation and test datasets when reviewing your models. Data cleaning and preparation procedures for analysing the uni-variate, bi-variate, and multi-variate processes include renaming the supplied dataset, eliminating columns, etc. The data cleaning process will vary depending on the dataset and the techniques and methods employed. In order to maximise the value of data in analytics and decision-making, data cleaning's primary goal is to identify and eradicate errors and anomalies.



	Date	Label	Top1	Top2	Top3	Top4	Top5	Top6	Top7	Top8	...	Top22	Top23	Top24	Top25	Para	Subjectivity	Objectivity	Positive	Neutral	Negative
0	1508	1	84	624	1277	327	50	242	230	460	...	1232	1447	71	984	84	95	25	0	25	99
1	1468	1	755	1203	270	1195	1227	994	1061	1031	...	1438	1466	1393	787	755	120	0	96	0	0
2	1455	1	501	1503	1358	431	264	83	865	1401	...	591	447	205	1336	501	111	9	0	9	106
3	1449	1	60	1303	1115	634	274	244	636	258	...	17	245	59	850	60	64	56	0	56	84
4	1444	0	256	1023	1010	1344	1160	1210	940	1199	...	388	403	901	1000	256	42	78	36	78	34

Table-1

DATA ANALYSIS OF VISUALIZATION:

In applied statistics and machine learning, data visualisation is a crucial ability. In fact, the main focus of statistics is on numerical estimates and descriptions of data. Data visualisation offers a crucial set of tools for gaining a qualitative insight. When examining and learning about a dataset, this may be helpful for spotting trends, corrupt data, outliers, and much more. With a little topic knowledge, data visualisations may be used to express and show important relationships in graphs & charts that are much more visceral and compelling for stakeholders than measures of association or relevance. It will suggest a closer look at some of the books suggested at the conclusion because data visualisation and exploratory data analysis are entire areas in themselves.

Data may not always make sense unless it is presented visually, such as through charts and graphs. Both applied statistics and applied machine learning value fast visualisation of data samples and other objects. It will show you how to utilise various plot types to analyse your own data and the many plot types you'll need to be familiar with when visualising data in Python.

- The use of bar charts to display categorical data and line plots to display time series data.
- The proper methods for summarising data distributions using histograms and box plots.

```
Text(0.5, 1.0, 'Subjectivity of distribution of News')
```

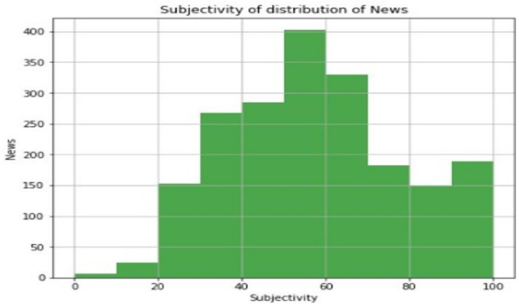
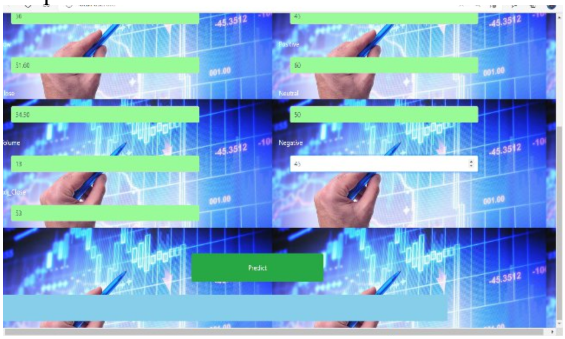


Fig :1.1

Test cases:

Test case Description	Result anticipated	Actual Results	Test Status(P/F)
Price Difference	Stock price increase/decrease	Stock price increase/decrease	P
Highest Value	Stock price increase/decrease	Stock price increase/decrease	P
Lowest Value	Stock price increase/decrease	Stock price increase/decrease	P
Share Volume	Stock price increase/decrease	Stock price increase/decrease	P
Closing price	Stock price increase/decrease	Stock price increase/decrease	P
General Index	Stock price increase/decrease	Stock price increase/decrease	P
Daily Share Volume	Stock price increase/decrease	Stock price increase/decrease	P
Monthly Share Volume	Stock price increase/decrease	Stock price increase/decrease	P
Opening Price	Stock price increase/decrease	Stock price increase/decrease	P

Output Screenshot:



36

ow

51.60

lose

34.50

olume

13

dj_Close

53

45

Positive

60

Neutral

50

Negative

45

VI. Future work and Conclusion

- Prediction of stock prices linked to the cloud.
- To streamline the work that needs to be done in an environment with artificial intelligence

The suggested system attempted to use analytical methods on this data beginning with knowledge purification and method, missing value, beta analysis. The best accuracy on the public examination set will be determined by a greater accuracy score. This programme will make it easier to find stock value predictions.

VII REFERENCES

[1] A Deep Learning Based Approach for Predicting the Price of Fresh Produce. Ifeanyi and Lobna Nassar, authors Muhammad Saad and Emmanuel Okwuchi 19-24 July 2020

[2] "Unsupervised real-time anomaly detection for streaming data," Neurocomputing, vol. 262, pp. 134–147, Nov. 2017, S. Ahmad, A. Lavin, S. Purdy, and Z. Agha.

[3] A Hammerstein-Wiener deep hybrid fuzzy neural network for predicting stock prices. Authors: Xie Chen, Chai Quek, and DeepuRajan International Conference on Artificial Intelligence in Information and Communication, 16 April 2020 (ICAIIC).

[4] A Deep Learning Based Approach for Predicting the Price of Fresh Produce. Ifeanyi and Lobna Nassar, authors Muhammad Saad and Emmanuel Okwuchi 19-24 July 2020 Unsupervised real-time anomaly detection for streaming by S. Ahmad, A. Lavin, S. Purdy, and Z. Agha

[5] Deep Learning Based Approach for Predicting the Price of Fresh Produce. Ifeanyi Emmanuel LobnaNassar is the author. Saad Muhammad Okwuchi 19-24 July 2020