Stock Market Prediction with High Accuracy using Machine Learning Techniques

Mr. Rahul Ranjan, Assistant Professor, Department of Computer Science & Engineering,
Faiz Ahmad Khan, B.Tech, Computer Science & Engineering (DSAI),
Altamash Ahmad, B.Tech, Computer Science & Engineering (CSE),
Abu Baker Aftab Khan, B.Tech, Computer Science & Engineering (CSE),
Aman Siddiqui, B.Tech, Computer Science & Engineering (DSAI),
Mohd Faiz, B.Tech, Computer Science & Engineering (DSAI),
Department of Computer Science and Engineering, Integral University Lucknow-226026 India

Abstract

Stock market trading is a major and predominant activity when one talks about the financial markets. With the inevitable uncertainty and volatility in the prices of the stocks, an investor keeps looking for ways to predict the future trends in order to dodge the losses and make the maximum possible profits. However, it cannot be denied that as of yet there is no such technique to predict the upcoming trends in the markets with complete accuracy, while multiple methods are being explored to improve the predictive performance of models to an extent as large as possible. With the advancement in Machine Learning (ML) and Deep Learning (DL) over the past few years, many algorithms are being deployed for stock price prediction. This paper researches 5 algorithms namely K-Nearest Neighbors, Linear Regression, Support Vector Regression, Decision Tree Regression, and Long Short-Term Memory for predicting stock prices of 12 leading companies of the Indian stock market. After exhaustive research of the various aspects related to the application of ML in stock market, a data extensive implementation has been carried out as a part of this research work wherein the stock price dataset of 12 companies over the last 7 years was collected and used. The paper also highlights some more efficient and robust techniques that are used to forecast trends in the stock market. In detail, the methodology followed, to acquire the results, has been talked about step-wise. Furthermore, a detailed comparative analysis of the performances of the aforementioned algorithms for stock price prediction has been carried out with the results displayed in a legible tabulated and graphical form to analyze them better. The conclusions from this novel, data comprehensive research work have been presented and it has been inferred that the DL algorithm outperforms all the other algorithms for stock price or time series prediction and provides results with extensive accuracy.

Keywords: Stock market, Machine learning, K-Nearest Neighbour (K-NN), Linear Regression (LR), Support Vector Regression (SVR), Decision Tree Regression (DTR), and Long Short-Term Memory (LSTM)

1.Introduction

Due to extremely unpredictable trends and high market volatility, majorly all stock market enthusiasts wish to get their hands on something which can help them score higher profits by predicting the stock market trends reliably.

Basically, the stock market by definition is a market where there are buyers and sellers interested in buying stocks of a certain company, and the prices of these stocks vary widely as time passes. Regardless, it won't be wise to ignore the reasons for these drastic oscillations in the stock market, which could possibly be, politics, brand reputation, the current global scenario, for instance, the pharmaceutical companies experienced a good hike due to the pandemic phase. The factors listed above can greatly affect the opinions, and beliefs of the potential investors leading to swinging stock market trends. Therefore, even though it is essential to understand these possible factors causing changes, it is not sufficient to develop a method for accurate prediction of the trends due to everlasting global transformations and uncertainties. However, constant efforts are being made towards developing a model or algorithm which can help investors to predict the changes more accurately than before. One of the most widely known and adopted ways to form predictive models is through the application of machine learning (ML) algorithms. ML is a concept wherein the computers learn or predict things with the help of past knowledge and training, without any external program being involved. There are several ML algorithms which are quite handy for performing predictions in interdisciplinary domains, be it stock market, electricity demand, healthcare domain, etc. However, it is critically important to decide the best algorithm depending on the type of the dataset and the purpose that is to be served. Dataset or the problem could either be time-dependent, for example, the height and weight of a toddler, which would be changing as time passes, or time-independent, for example, the name of a person, which would remain the same even after a decade. Now, if the stock market is considered, it's a highly time-dependent area, where the prices fluctuate each given minute, therefore, time series analysis is an extensively adopted approach. One of the highly prevalent techniques to carry out time-series modeling is the ARIMA (Auto Regressive Integrated Moving Average) model as proposed by, however, since ARIMA model is that of linear type, it is unsuitable for stock market prediction as it is unable to look after the fluctuations in the data set due to high market volatility. Nonetheless, ML and data science have seen a lot of refinement in the past few years, leading to the development of some particular algorithms which are quite efficient for predictive analytics, be it in any field. A couple of techniques and algorithms related to ML have been studied and discussed in. In the past, there have been many research works wherein some common ML algorithms have been worked upon to perform predictive analytics. However, this paper aims at developing ML models using 5 different types of algorithms and further applying them to the stock market area for predicting the stock market trends. Thereafter, the five models that are implemented, will be compared with each other based on various performance metrics like Symmetric Mean Absolute Percentage Error (SMAPE), R^2 value (R-squared), Root Mean Square Error (RMSE) so as to highlight the best model. The ML algorithms are used to form models and then carry out the rest of the purpose of prediction and data analysis. The primary or first and foremost step is to train the models using good datasets, in order to make the models learn the inputs being given and later use these past experiences and knowledge to classify and predict things. This particular portion of the dataset that is meant for training models is known as the training dataset. Later, this knowledge gained by these models and the past encounters help them to predict more accurately. Therefore, ML is gradually gaining fame for use in the field of stock market and people, rather investors are beginning to rely on these ML models for investing in stock markets. To be specific, this paper includes 5 ML algorithms, which are Linear Regression (LR) algorithm, Lasso and Ridge regression algorithm, and Support Vector machine (SVM) algorithm.

2. About Stock Market

2.1. Stock Exchange

The stock exchange is a place where shares of listed companies can be traded. It is considered to be a secondary form of market. In order for a company to go public and list its shares to investors out in the market, it must make a place for itself on any of the recognized stock exchanges, and subsequently, a promoter must sell a substantial amount of its shares to the public retail investors, which once successfully done, then further trading can be carried on in the secondary market or stock exchange. In India, there are prevalently two major stock exchanges namely, the Bombay stock exchange (BSE) having approximately 5000 listed companies, and the National stock exchange (NSE) with around 1600 listed companies. Both the NSE and the BSE have similar functionality and trading mechanisms. Trading in the stock market mainly takes place through Demat and trading accounts. Stock exchanges help the public at large to pool their savings and channel their funds while the companies enjoy an inflow of investments into their ventures. The stock market has brought about a revolution in the arena of Indian investments. Faced with increasing inflation, declining bank interest rates, and hopes of better returns the middle-class investors are now shifting towards the equity market. This in totality sums up the ever-growing need and importance of stock exchanges.

2.2. Open-High-Low-Close Charts

Open-high-low-close charts (OHLC) are a kind of bar chart that displays the open, high, low, and closing price values of shares for a frequent period. OHLC charts comprise a vertical line and two short horizontal lines. The height of the vertical line signifies the intra-day range for a certain period, while the highest point is that period's highest value and the lowest point is that period's lowest value. Talking about the horizontal line, the points situated on the extreme left or the left extension signify the opening price while the points on the extreme right or the right extension signify the closing price for a certain time period. The collective structure is called the price bar. Rising prices are denoted by the right line being above the left one while falling prices are characterized by the right line falling below the left one. OHLC charts are versatile and may be used for understanding and depicting any sort of time period i.e., minutes, days, or any other as per the requirement. Even though OHLC charts are more informative when compared with line charts, they render the same amount and content of the information as that of candlestick charts, the only difference being in the presentation of the data i.e., OHLC charts use left-facing and right-facing horizontal lines to display the opening and closing values, while the candlestick charts display the same using real body.

2.3. Interpreting OHLC charts

There are several ways and methods of interpreting or analyzing OHLC charts. Fig.1 and Fig.2 shown
below display the markings and denotations to understand OHLC values.

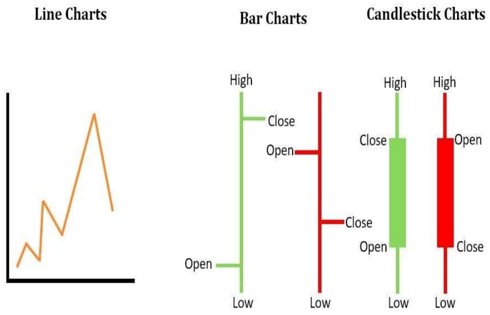


Fig 1:

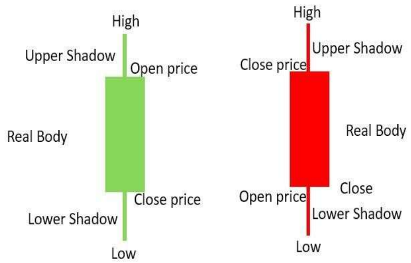


Fig. 2: Candlestick interpretation

1.Vertical height: The vertical height indicates the volatility of the stock market during a given period. The more the height of the vertical line, the more the volatility in the market.

2.Horizontal line: The leftmost and the rightmost points of the line imply the highest and lowest opening and closing values respectively. The similarity in the opening and closing values imply a condition of indecisiveness in the market.

3.Bar color: Black-colored bars imply an upward trend whereas red-colored bars imply a downtrend. Such information is handy when trying to analyze the trend strength and the direction.

4.Patterns: The major patterns are the inside bar, the outside bar, and a key reversal. Although key reversals do not appear very often, they are significant when they occur, giving reliable information to the traders regarding trend reversals of the signal whether upward accelerating or downward accelerating.

3. STOCK PREDICTION TECHNIQUES TAXONOMY

These techniques have gained popularity and have shown promising results in the field of stock analysis in the recent past.

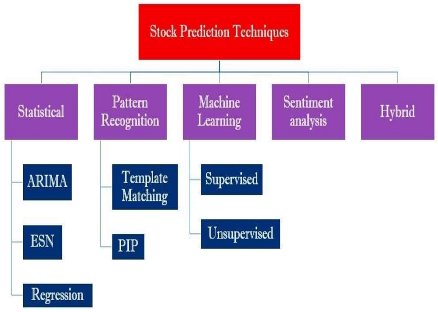


Fig. 3: Stock Prediction Techniques

If the recent developments related to stock prediction are to be discussed, there are four major sorts, namely, Statistical methods, ML, Pattern recognition, and Sentimental analysis. There are also ways to predict stock using multiple methods together giving rise to hybrid technologies. The flowchart depicted in the figure above shows the taxonomy of prevalent stock prediction strategies. Prior to the dawn of efficient ML algorithms, one of the wellknown ways to predict and study stocks were the statistical methods that presumed stationarity, normality, and linearity to help in stock prediction. A commonly discussed term in stock market examination called 'Time Series', is nothing but a rigorous cluster or compilation of data related to the day-to-day sales, stock prices, number of investors, etc. One of the corollaries of statistical approaches applied in the field of stock market prediction being a part of the univariate computation because they make use of the time series data as inputs are the AutoRegressive Integrated Moving Average (ARIMA) model, Auto Regressive Moving Average

(ARMA) Model, the Smooth Transition Autoregressive (STAR) model, and the Generalized Autoregressive Conditional Heteroskedastic volatility model (GARCH). In the case of the ARMA model, the Auto-regressive (AR) model and the Moving Average model are combined, out of which the former is responsible for justifying the impetus or momentum and the average regression impacts that are noticed during trading, whereas the latter attempts to grasp the sudden or shock effects that are marked in time series data. However, it is known that time-series data, more specifically, financial time-series data, is highly volatile hence ARMA model is not the best choice for this purpose as it does not take into account this drastic variation. Taking about the ARIMA model, it is an integration of the ARMA model, capable of converting a non-stationary data series into a stationary data series. The novel data points can be forecasted by fitting the ARIMA model to the time series data available. A separate branch of statistical techniques making use of numerous variables as inputs, namely Quadratic Discriminant Analysis (QDA), Linear Discriminant Analysis (LDA), and regressive algorithms have been further explained in. Talking about the second method i.e., Pattern Recognition, is pretty much similar to ML, however both the techniques are applied in a very different manner when it comes to use for stock market prediction. In patter recognition the main focus is on the trends and tendencies in the available data. The periodic trends or orders noticed in the Open-HighLowClose candlestick charts are the patterns which are being talked about in pattern recognition technique. These trends and candlestick charts are extensively used by avid traders and investors while participating in stock market purchases. The technical aspect of chart analysis is to a lot of extent dependent on the stock dataset. Primarily, a good level of visual study and analysis is required for the charts which are formulated using time series data exhibiting fluctuations of various attributes like volume, price, and indirectly deduced indicators like price momentum. Charting is considered to be a very robust technique to perform technical analysis when it comes to comparison of market value and past history of volumes with the chart trends to predict the upcoming tendencies and trends of prices based on the level of match. Some of the commonly observed patterns in charts are spikes, pennants, flags, gaps, triangles, saucers, wedges, tops and bottoms, head-and-shoulders. This technique is capable of apprising the trader and investor of the upcoming expected trends in a particular stock. To broadly classify, pattern recognition is of two types, namely, Perceptually Important Points (PIP), which diminish the dimensions of the time series data by the preservation of outstanding points, and the other type being template matching, wherein a stock trend or pattern is matched with an image meant for the purpose of object identification. With the increasing awareness and advancement in the field of ML, it is being used as a highly prevalent and potent method of predicting trends and tendencies in financial markets and specifically in stock markets. On a broader note, ML can be classified into two types, namely, supervised learning and unsupervised learning. In the former type, labelled dataset is used as input for training the model and the subsequently deduced data is available, whereas in the latter, unlabeled dataset is used as input. Supervised learning aims at training a model in a way that when new data is input in the model, it would automatically associate and match it to the given output data. However, the unsupervised form of learning aims at training a model, to find some similarity or correlation in the available data. Occasionally, this form learning may also act as an antecedent for tasks involving supervised learning.

4. ML Algorithms Applied

4.1. Linear Regression Algorithm

The linear regression algorithm falls under the category of supervised learning in ML. This algorithm makes predictions of values that are well within the range rather than predicting categories. It establishes a linear relationship between the dependent and the independent variable and does not work very well with the non-linear type of data sets due to the presence of outliers. Researchers used this algorithm for stock market predictions and came to the conclusion that this algorithm when used for predicting daily stock values, offered serious challenges that must be taken care of. Using this algorithm's prediction, investors cannot reliably invest money.

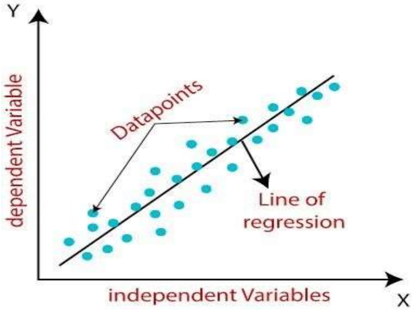


Fig. 4: Linear Regression [33]

4.2. K-Nearest Neighbours Algorithm

K-nearest-neighbor (K-NN) being one of the most essential and effective algorithms for data segregation is capable of becoming the primary choice for implementation especially when the given data is quite ambiguous. This algorithm was invented back in 1951 by Evelyn Fix and Joseph Hodges. The K-NN algorithm is positioned under the supervised type learning technique and although it is suited for classifying as well as regressing both, it is predominantly utilized for classifying objects. K-NN can also be referred to as the lazy learner algorithm, as the data set is only stored initially, but the learning process of the training data set does not take place until there is a demand for classification or prediction of the new data set. It is also non-parametric in nature, i.e., in K-NN there does not exist any predetermined method between the input and output.

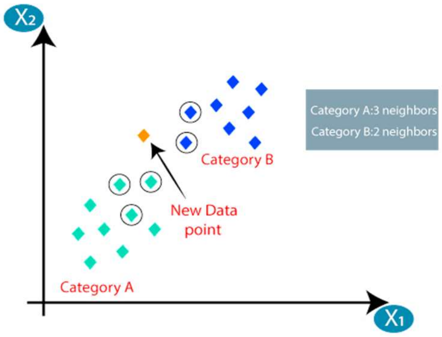


Fig.5: K-Nearest Neighbour Algorithm [36]

4.3. Support Vector Regression

Among one of the highly prevalent supervised type learning algorithms, the Support Vector Regression algorithm is intended for regression and classification problems. For support vector machine regression or SVR, a hyperplane is identified with maximum margin such that the maximum number of data points are within those margins. In SVR, the best fit line is the hyperplane that has the maximum number of points. In SVR algorithm, extreme vector points called Support Vectors are chosen which help in creating an appropriate hyperplane. It is similar to the Support Vector Machine (SVM) algorithm when it comes to the working principle. It is used for working with time series data.

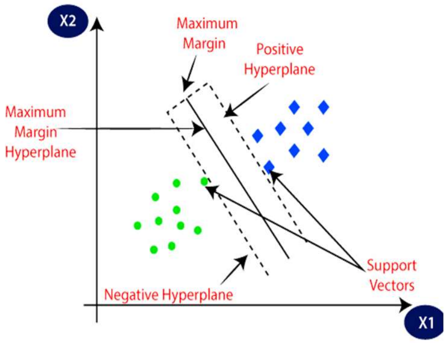


Fig.6: Support Vector Regression Algorithm [39]

4.4. Decision Tree Regression Algorithm

Decision Tree Regression algorithm, belonging to the supervised learning class of algorithms is mostly preferred for solving classification problems but either way, it may be used in classifying as well as in regressing cases. It consists of inner nodes representing the structures of the branches, dataset, representing the verdict given by the algorithm, and each leaf node representing an outcome. There are two nodes, first is the decision node, that is used to make a decision and has various branches; and second is the leaf node, which is the output of decision nodes and has no further branches. The root node is a starting point that further expands to various branches making it a tree-like structure. Decision tree simply forks the tree into subtrees on the basis of answer to question i.e., whether a Yes or a No.

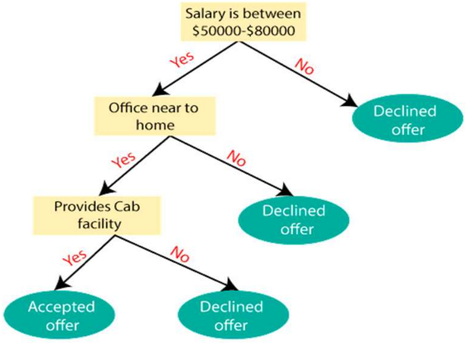


Fig. 7: Decision Tree Regression Algorithm [41]

4.5. Long Short-Term Memory Algorithm

Due to backpropagation with real-time recurrent learning or time, the error-incorporated signals running rearward in time are likely to disappear or blow up; the temporal shifts of the error incorporated signal to a great extent relies on the weight sizes. In case of blowing up, the weights are quite likely to start oscillating and in case of disappearance, either the time consumed to learn bridging longer time lags is out of bounds, or in the worst case it does not work. As a remedy, the Long Short-Term Memory (LSTM) algorithm, a novel type of recurrent neural network came into existence in 1991, developed by Sepp Hochreiter and Jurgen Schmidhuber to outperform the existing systems and overcome the error backpropagation issues discussed above. The primary version of this Long short-term memory algorithm only consisted of cells, input, and output gates. This algorithm is capable of bridging time breaks in excess of steps even when the sequences being used for input are incompressible or noisy in nature while preventing losses of short time break abilities.

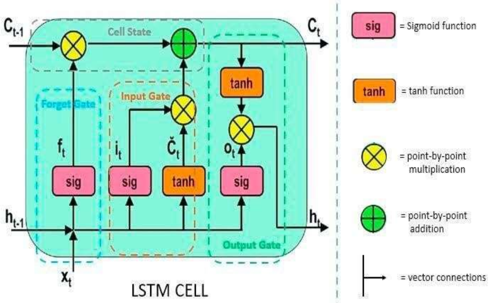


Fig.8: Long Short-Term Memory Algorithm

5. METHODOLOGY

The methodology for any research project is the core aspect of obtaining authentic and accurate results. Therefore, the methodology adopted for this research work has been carefully curated and scientifically designed so as to render reliable and wholesome conclusions from the implementation. The methodology used in this research project is majorly described through the following steps. The first step involves the collection of raw data from various open sources. The second step revolves around data preprocessing which includes data scaling, data standardization, data cleaning, and other technical ways to formulate the dataset. The third step consists of splitting the dataset into training and testing data. The next step paves way for the training of the five models which are built using five respective algorithms by feeding, which is carried out using the training dataset, following this, the models have been tested using the testing dataset in order to be able to notice the deviation from actual values in different models. Lastly, the five different models for each of the twelve companies' datasets were tested and evaluated using efficient performance metrics namely, Symmetric Mean Absolute Percentage Error (SMAPE), R2-Value (R-squared), and Root Mean Square Error (RMSE), in order to rate and grade the performance, draw some comparisons and reliable conclusions related to the algorithms, i.e., K-Nearest Neighbour algorithm, Linear Regression algorithm, Support Vector Regression algorithm, Decision Tree Regression algorithm, and Long Short-Term Memory Algorithm. The steps taken to execute this research implementation are shown in the flow diagram below written in a generalized form, while the details have been mentioned in the subsequent sub-sections.

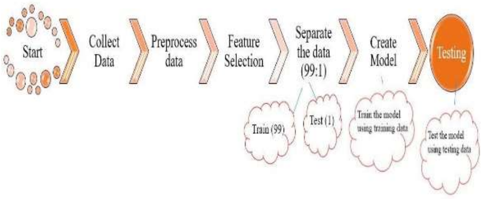
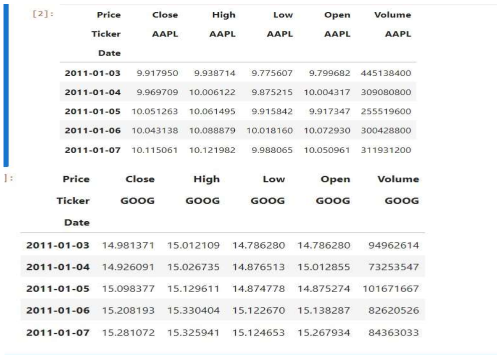


Fig.9:

5.1. Data Description

The first and foremost step in most ML predictive analytics projects is picking or acquiring a suitable dataset which involves. As part of this implementation research work, stock price datasets were collected from the Quandl Website, Bombay Stock Exchange, starting from the year January 2015 up till and including April 2021 for twelve well-known associations or companies that capture a huge chunk of market share and economy, especially in the Indian market. The following attributes of the stock price datasets were available namely, previous closing price, opening price, an all-time high, an all-time low, last price, and closing price of the stocks. While the twelve companies whose datasets have been used are listed as follows:-Apple, Google, Axis Bank, Housing Development Finance Corporation Limited (HDFC) Bank, Industrial Credit and Investment Corporation of India (ICICI) Bank, Kotak Bank, Hindustan Unilever Limited, Maruti, National Thermal Power plant Corporation.



[2]:

Price	Close	High	Low	Open	Volume
Ticker	AAPL	AAPL	AAPL	AAPL	AAPL
Date					
2011-01-03	9.917950	9.938714	9.775607	9.799682	445138400
2011-01-04	9.969709	10.006122	9.875215	10.004317	309080800
2011-01-05	10.051263	10.061495	9.915842	9.917347	255519600
2011-01-06	10.043138	10.088879	10.018160	10.072930	300428800
2011-01-07	10.115061	10.121982	9.988065	10.050961	311931200

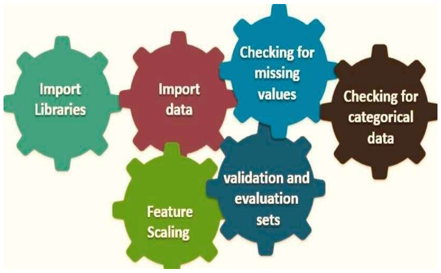
]:

Price	Close	High	Low	Open	Volume
Ticker	GOOG	GOOG	GOOG	GOOG	GOOG
Date					
2011-01-03	14.981371	15.012109	14.786280	14.786280	94962614
2011-01-04	14.926091	15.026735	14.876513	15.012855	73253547
2011-01-05	15.098377	15.129611	14.874778	14.875274	101671667
2011-01-06	15.208193	15.330404	15.122670	15.138287	82620526
2011-01-07	15.281072	15.325941	15.124653	15.267934	84363033

Fig.10 – Samples of stock datasets collected for twelve companies along with all the attributes

5.2. Data Pre-processing

Data preprocessing especially for data extensive projects is a critically important step as it involves the transformation of random and raw data in such a way that it enhances its quality by removing or cleaning the unwanted points in addition to standardizing it for normal use and making it capable of delivering useful insights. It is not the huge quantity of data that gives great outputs but rather data quality that creates an impact. It involves data cleaning, data segregation or organization, data scaling, data standardization, etc., i.e., data normalization and standardization as well as encoding categorical data. During the data preprocessing step of the project, Min-Max scalers were used to scale the data in order to scale and standardize it and the null values, missing values, and unknown values were cleaned and disparities if any were taken care of. The two main Python libraries that were used for the purpose of preprocessing the dataset were NumPy and Pandas, and for data visualization, Matplotlib was used. NumPy performed the functions of a scientific calculator whenever required during the manipulation of the datasets, while the Pandas library was appropriate for data analysis and manipulation. Matplotlib was used to visualize the data in the form of charts.


Import Libraries
Import data
Checking for missing values
Checking for categorical data
Feature Scaling
validation and evaluation sets

5.3. Splitting of data into train and test dataset

As discussed earlier, in order to move ahead of the data preprocessing step, it is required that the dataset is split into the training and testing datasets. As part of this project, the dataset was split into training and testing sets in the ratio of 99:1. This ratio was chosen as this is a predictive analytics project with loads of volatile data hence this ratio would suffice for such requirements. Therefore, out of 2312 values or trading days in the dataset, the last 8 days were devoted to the testing of models while the rest were kept for training. Since the research involves time series forecasting, therefore it is necessary to train the model on a majority of historic data owing to data interdependencies. A portion of the data was also set aside for cross-validation.

Furthermore, random rows were allotted to training data and testing data to ensure random sampling as it enhances the model's performance during testing. It is the Scikit learn library that handles the splitting of the dataset into training and testing data. The ratio in which training and testing data have been split is an essential factor since it impacts the performance of models. Having similar training data and test data can lead to the overfitting of models while having huge differences within dataset values can lead to underfitting. Therefore, it is necessary to have an appropriate ratio of train to test dataset so as to get a real picture of the performance of various models which in turn would render reliable results.

5.4. Training of models

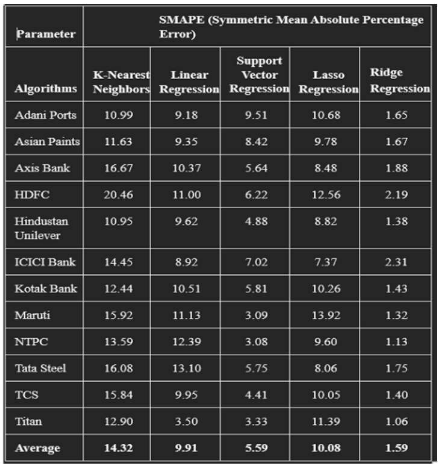
Training is the process that helps ML algorithms to extract useful information to train the models so that they are capable of producing accurate predictions and hence desired outcomes. The chosen algorithms for this project were the K-Nearest Neighbour algorithm, Support Vector Regression algorithm, Linear Regression algorithm, Decision Tree Regression algorithm, and Long ShortTerm Memory algorithm. These five models for each of the twelve companies' were trained using the training data keeping in mind the issues of overfitting or underfitting. Consistent learning by the models was chosen as the key to improving the predictive performance of these models. The algorithms used in this project are that of the supervised learning type and the kind of learning that has been adopted for this project implementation is also of the supervised type. The target attribute i.e., the desired value must be a part of the dataset as it forms an important basis for prediction. The highly important features like the time taken to train each model as well as the time lag error (i.e., number of steps required to shift data reverse in time) were also taken into consideration. This stage comprises an iterative process of learning called 'model fitting'. The weights of the model were initialized randomly as it offers more adjustability to algorithms.

5.5. Testing the models

This is the final stage of the process, which as the name suggests evaluates the performance of selected fully trained models based on some efficient and useful performance parameters or metrics. All of the five models i.e. K Nearest Neighbour, Support Vector Regression, Decision Tree Regression, Linear Regression, and Long Short-Term Memory were tested using the 8-value testing dataset which had all possible combinations to make testing as close to real-time as possible. The performance evaluation parameters or metrics considered were Symmetric Mean Absolute Percentage Error (SMAPE), Root Mean Square Error or Deviation (RMSE/RMSD) and R Squared value (R2). SMAPE is a well-known and efficient testing parameter based on percentage errors or relative errors. The lower the value of SMAPE, the higher the predictive accuracy of a given model. RMSE is defined as the square root of the mean of squared differences between actual and predicted values. The lower the RMSE value, the better the model. R Squared (R2), possesses an ideal value of '1', and the values could be negative or positive, the closer the value to the positive '1', the better the model. Subsequently, detailed analytical and comparative results were drawn after testing which are displayed in the results section.

6. RESULTS

In this section, the results after carrying out the successful implementation of the project have been presented in the form of tables and graphs. As part of this project algorithms i.e., K-Nearest Neighbour algorithm, Support Vector Regression algorithm, Linear Regression algorithm, Lasso and Ridge regression algorithms, and Long Short Term Memory algorithm were chosen for the prediction of stock prices of twelve different companies. The dataset was huge, starting from 2015 up to and including 2021. The models were tested for 8 trading days i.e., since the train to test ratio was 99:1, and the data for 2304 days was used for training whereas the remaining 8 days were allocated for testing the created models. The models were tested on three essential performance metrics, namely, Symmetric Mean Absolute Percentage Error (SMAPE), R-squared value (R2), and Root Mean Square Error (RMSE). These are well-known prevalent evaluation parameters that help researchers to draw conclusions about the different models that were being studied.



Parameter	SMAPE (Symmetric Mean Absolute Percentage Error)				
Algorithms	K-Nearest Neighbors	Linear Regression	Support Vector Regression	Lasso Regression	Ridge Regression
Adani Ports	10.99	9.18	9.51	10.68	1.65
Asian Paints	11.63	9.35	8.42	9.78	1.67
Axis Bank	16.67	10.37	5.64	8.48	1.88
HDFC	20.46	11.00	6.22	12.56	2.19
Hindustan Unilever	10.95	9.62	4.88	8.82	1.38
ICICI Bank	14.45	8.92	7.02	7.37	2.31
Kotak Bank	12.44	10.51	5.81	10.26	1.43
Maruti	15.92	11.13	3.09	13.92	1.32
NTPC	13.59	12.39	3.08	9.60	1.13
Tata Steel	16.08	13.10	5.75	8.06	1.75
TCS	15.84	9.95	4.41	10.05	1.40
Titan	12.90	3.50	3.33	11.39	1.06
Average	**14.32**	**9.91**	**5.59**	**10.08**	**1.59**

Table I displays the tabulated results for the SMAPE acquired when a particular model was tested for that particular company's dataset. With the ideal value for SMAPE being close to zero, from this table, it is observed that out of all the five different algorithmic models, the DL algorithm i.e., Lasso algorithm has rendered the best predictive performance, as it has the leastvalue of error (1.59), followed by Support Vector Regression, with a SMAPE of 5.59

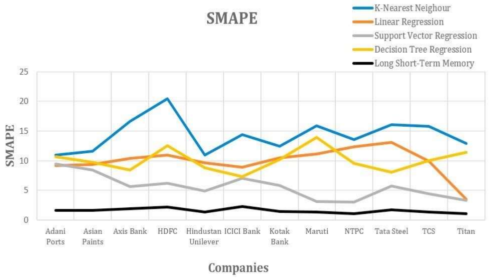


Fig. 12: The SMAPE for all algorithms plotted against the companies

The plot shown in Fig.12, depicts the predictive performance of all the five algorithms in terms of SMAPE plotted against the twelve companies. From this plot, it is quite noticeable that the LSTM algorithm has rendered the best results, as the low-lying black plot (LSTM) is much below compared to the rest of the algorithms. It is also possible to conclude that the SVR algorithm has provided the second-best performance, however, the remaining models do not seem to be a very good choice for predictive analytics, as they are extremely error-prone.

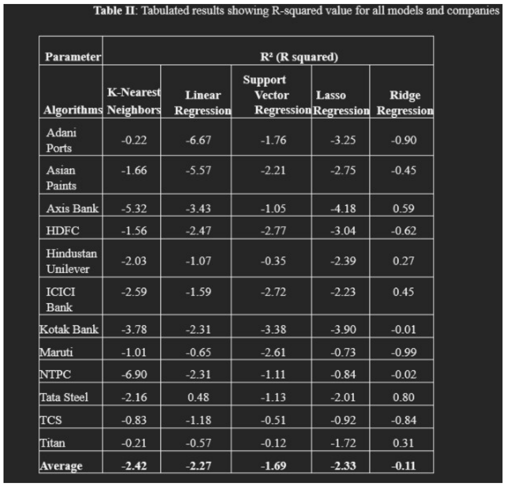


Table II: Tabulated results showing R-squared value for all models and companies

Parameter	R² (R squared)				
Algorithms	K-Nearest Neighbors	Linear Regression	Support Vector Regression	Lasso Regression	Ridge Regression
Adani Ports	-0.22	-6.67	-1.76	-3.25	-0.90
Asian Paints	-1.66	-5.57	-2.21	-2.75	-0.45
Axis Bank	-5.32	-3.43	-1.05	-4.18	0.59
HDFC	-1.56	-2.47	-2.77	-3.04	-0.62
Hindustan Unilever	-2.03	-1.07	-0.35	-2.39	0.27
ICICI Bank	-2.59	-1.59	-2.72	-2.23	0.45
Kotak Bank	-3.78	-2.31	-3.38	-3.90	-0.01
Maruti	-1.01	-0.65	-2.61	-0.73	-0.99
NTPC	-6.90	-2.31	-1.11	-0.84	-0.02
Tata Steel	-2.16	0.48	-1.13	-2.01	0.80
TCS	-0.83	-1.18	-0.51	-0.92	-0.84
Titan	-0.21	-0.57	-0.12	-1.72	0.31
Average	**-2.42**	**-2.27**	**-1.69**	**-2.33**	**-0.11**

Table II: Tabulated results showing R-squared value for all models and companies

Table II displays the tabulated results for the R-Squared value (R2) acquired when a particular model was tested for that particular company's dataset. With the ideal value for R-squared being close to a non-negative '1', from this table, it is observed that out of all the five different algorithmic models, the DL algorithm i.e., the Long Short-Term Memory algorithm has provided the best results, as it has R-squared quite close 1, (i.e., -0.11), followed by Support Vector Regression, with an R-squared value of -1.69, so on and so forth.

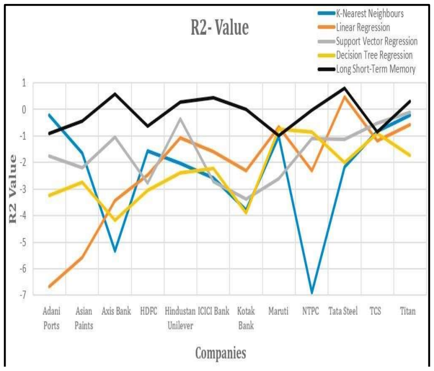


Fig. 13: The R-squared value for all algorithms plotted against the companies

The plot shown in Fig.13, depicts the predictive performance of all the five algorithms in terms of R-squared plotted against the twelve companies. Upon careful reading of this chart, it can be seen that the LSTM algorithm has rendered the best results, as the upper most situated black plot line (LSTM) is the one that is closest to '1', compared to the rest of the algorithms. It is also possible to conclude that the SVR algorithm has provided the second-best performance, with the remaining models not showing a very great and credible performance for predictions

Table III: Tabulated results showing RMSE value for all models and companies

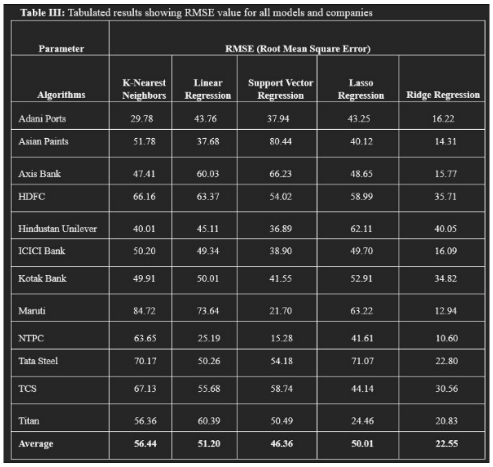


Table III: Tabulated results showing RMSE value for all models and companies

Parameter	RMSE (Root Mean Square Error)				
Algorithms	K-Nearest Neighbors	Linear Regression	Support Vector Regression	Lasso Regression	Ridge Regression
Adani Ports	29.78	43.76	37.94	43.25	16.22
Asian Paints	51.78	37.68	80.44	40.12	14.31
Axis Bank	47.41	60.03	66.23	48.65	15.77
HDFC	66.16	63.37	54.02	58.99	35.71
Hindustan Unilever	40.01	45.11	36.89	62.11	40.05
ICICI Bank	50.20	49.34	38.90	49.70	16.09
Kotak Bank	49.91	50.01	41.55	52.91	34.82
Maruti	84.72	73.64	21.70	63.22	12.94
NTPC	63.65	25.19	15.28	41.61	10.60
Tata Steel	70.17	50.26	54.18	71.07	22.80
TCS	67.13	55.68	58.74	44.14	30.56
Titan	56.36	60.39	50.49	24.46	20.83
Average	**56.44**	**51.20**	**46.36**	**50.01**	**22.55**

Table III displays the tabulated results for the Root Mean Square Error (RMSE) acquired when a particular model was tested for that particular company's dataset. With the ideal value for RMSE being zero, from this table, it is observed that out of all the five different algorithmic models, the DL algorithm i.e., the Long Short-Term Memory algorithm has shown the best performance, as it has an RMSE of 22.55, followed by Support Vector Regression, with an RMSE 46.36, so on and so forth.

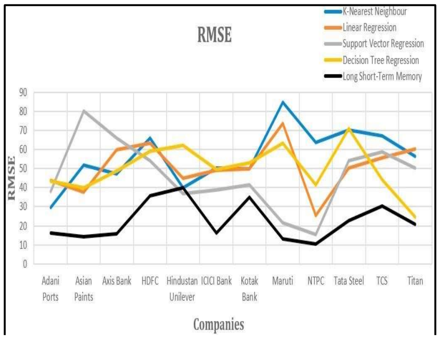


Fig.14: The RMSE for all algorithms plotted against the companies

The plot in Fig.14, depicts the predictive performance of all the five algorithms in terms of RMSE plotted against the twelve companies. Upon careful analysis of the chart, it can be seen that the LSTM algorithm has provided the best result, as the lower most situated black plot line (LSTM) is the one that is the closest to '0', compared to the rest of the algorithms. It is also possible to conclude that the SVR algorithm has more or less provided the secondbest performance, with the remaining models following these two in terms of predictive accuracy and performance. Therefore, after plotting and tabulating all the results, it can be seen that DL algorithm LSTM is the best choice for predictive analytics among the chosen algorithms, followed by the support vector Regression algorithm which has also provided decent results. Linear Regression and Decision Tree Regression have almost provided similar results, however K-NN has provided the worst results as it is predominantly a classification algorithm and not a prediction friendly algorithm.

7. Future Scope

With the increasing demand for ML in almost every possible place and situation, be it industries or business models, or healthcare domains, it is of utmost importance to make better models which can make more accurate and precise predictions from huge sets of data available. However, the analysis from the results above and various other literature reviews suggests that ML yields less authentic results when it comes to the prediction of time series data. Nevertheless, there is a potent solution to this which lies in DL and neural networks that offers great results during time series prediction when compared to ML. The results found after the implementation of this project conform to the theoretical facts related to the performance of the algorithms. As far as this project is considered, there may be more algorithms implemented in the future and different datasets may be used in order to offer a wider spectrum for comparison of the algorithms and on a wider note comparison of ML and DL. DL techniques comprise several layers which drive them a step closer to the way human brains function whereas, on the contrary, ML still requires a lot more human assistance. DL techniques are based on the concept of neural networks, similar to the neuron cells in the human brain. Further DL techniques equip the machine to itself identify features that make identification easier or in other words hierarchical arrangement of the features. A major advantage that comes with DL is that they continue to become more robust and efficient with increasing data, which is a quite favorable feature for data extensive projects and historic time-series predictions. The future lies in the development and betterment of more and more ML and DL based models requiring minimum human intervention, lesser prediction time, more accurate predictions, capability to handle huge data with the help of multiple layers, reduced complexity, and more affordability, etc. The upcoming researchers possess DL and ML possess high potential for the upcoming researchers to study, scientifically curate and develop much more efficient and robust models which could predict faster and more precise results based on real-time situations and would help theworld save and progress in multiple domains to give rise to holistic development in interdisciplinary areas with minimum to no human intervention.

8. Conclusions

This research work aimed at developing ML models which would be capable of predicting stock prices with an increased accuracy, so that interested traders and investors could make use of such methods to experience increased profits by investing on the right day at the right place. Successful implementation as part of this project was carried out wherein five algorithms i.e., K-Nearest Neighbours, Linear Regression, Support Vector Regression, Decision Tree Regression, and Long Short-Term Memory algorithms were deployed to create precise predictive models for application in stock price prediction of twelve prevalent Indian Companies namely, Adani Ports, Asian Paints, Axis Bank, Housing Development Finance Corporation Limited (HDFC) Bank, Industrial Credit and Investment

Corporation of India (ICICI) Bank, Kotak Bank, Hindustan Unilever Limited, Maruti, National Thermal Power plant Corporation (NTPC), Tata Steel, Tata Consultancy Services (TCS) and Titan, after which an elaborate comparative analysis of the performances of the algorithms during stock price prediction has been carried out. The stock prices collected were from 2015 to 2021, and after this exhaustive research, it can be concluded that DL algorithms have a substantial edge over simple ML algorithms when it comes to the prediction of time series data. out of the five chosen algorithms, the Long ShortTerm Memory algorithm was a DL algorithm that has provided the best results during stock price prediction. The results section of this research paper clearly displays the values acquired during the testing of the models in the form of tables and graphs for three evaluation metrics i.e., Symmetric Mean Absolute Percentage Error (SMAPE), R-Squared Value (R2), and Root Mean Square Error (RMSE). While carefully studying and analysing the result, it is concluded that the LSTM algorithm is the best choice among the given algorithms for time series prediction, because it has the least value or errors with SMAPE (1.59), R2 (-0.11), and RMSE (22.55). The second-best algorithm for this task was Support Vector Regression with an SMAPE value (5.59), R2 value (-1.69) and RMSE (46.36). While the algorithms Linear Regression and Decision Tree Regression have almost matched performances, KNN has shown the most inferior quality of prediction as it is predominantly a classification algorithm. Therefore, the implementation and the associated results have conformed to the theoretical analysis.

References

1. **Pei-Yuan Zhou, Keith C.C. Chan, Member, IEEE, and Carol XiaojuanOu, "Corporate Communication Network and Stock Price Movements: Insights From Data Mining", IEEE 2021**

2. **Atsalakis GS, Valavanis KP. Forecasting stock market short-term trends using a neuro-fuzzy based methodology. Expert Syst Appl. 2009;36(7):10696–707.**

3. **Ayo CK. Stock price prediction using the ARIMA model. In: 2022UKSim-AMSS 16th international conference on computer modelling and simulation. 2014.**

4. **Brownlee J. Deep learning for time series forecasting: predict the future with MLPs, CNNs and LSTMs in Python. Machine Learning Mastery. 2021. https://machinelearningmastery.com/timeseries-prediction-lstm-recurrent-neural networkspython-keras.**

5. **Shih D. A study of early warning system in volume burst risk assessment of stock with Big Data platform. In: 2019 IEEE 4th interna-**

tional conference on cloud computing and big data analysis (ICCCBDA). 2019. pp. 244–8.

6. Sirignano J, Cont R. Universal features of price formation in fnancial markets: perspectives from deep learning. Ssrn. 2018. https://doi.org/10.2139/ssrn.3141294.

7. Thakur M, Kumar D. A hybrid fnancial trading support system using multi-category classifers and random forest. Appl Soft Comput J. 2018;67:337–49. https://doi.org/10.1016/j.asoc.2018.03.006.

8. Tsai CF, Hsiao YC. Combining multiple feature selection methods for stock prediction: union, intersection, and multiintersection approaches. Decis Support Syst. 2020;50(1):258–69.
https://doi.org/10.1016/j.dss.2020.08.028.

9. Tushare API. 2018. https://github.com/waditu/tushare. Accessed 1 July 2020.

10. Wang X, Lin W. Stock market prediction using neural networks: does trading volume help in short-term prediction?. n.d.

11. Weng B, Lu L, Wang X, Megahed FM, Martinez W. Predicting short-term stock prices using ensemble methods and online data sources. Expert Syst Appl. 2020;112:258–73. https://doi.org/10.1016/j.eswa.2018.06.016.

12. Zhang S. Architectural complexity measures of recurrent neural networks, (NIPS). 2019. pp. 1–9.

13. Ashish Sharma, Dinesh Bhuriya, Upendra Singh. "Survey of Stock Market Prediction Using Machine Learning Approach", ICECA 2022.

14. Loke.K.S. “Impact Of Financial Ratios And Technical Analysis On Stock Price Prediction Using Random Forests”,

15. Xi Zhang1, Siyu Qu1, Jieyun Huang1, Binxing Fang1, Philip Yu2, “Stock Market Prediction via Multi-Source Multiple Instance Learning.” IEEE 2021.

16. VivekKanade, BhausahebDevikar, SayaliPhadatare, PranaliMunde, ShubhangiSonone.
“Stock Market Prediction: Using Historical Data Analysis”, IJARCSSE 2022.

17. SachinSampatPatil, Prof. Kailash Patidar, Asst. Prof. Megha Jain, “A Survey on Stock Market Prediction Using SVM”, IJCTET 2021.

18. https://www.cs.princeton.edu/sites/default/files/uploads/Saahil_magde. Pdf

19. Hakob GRIGORYAN, “A Stock Market Prediction Method Based on Support Vector Machines (SVM) and Independent Component Analysis (ICA)”, DSJ 2021.

20. RautSushrut Deepak, ShindeIshaUday, Dr. D. Malathi, “Machine Learning Approach In Stock Market 9. Prediction”, IJPAM 2022.