International Journal of Information System Modeling and Design
Volume 13 • Issue 9

Analysis and Prediction of Healthcare Sector Stock Price Using Machine Learning Techniques:
Healthcare Stock Analysis

Daiyaan Ahmed, Vellore Institute of Technology, India
Ronhit Neema, Vellore Institute of Technology, India
Nishant Viswanadha, Vellore Institute of Technology, India
Ramani Selvanambi, Vellore Institute of Technology, India*

ABSTRACT

Healthcare sector stocks are a very good opportunity for investors to obtain gains faster most of the time in a year and mostly during this COVID pandemic. Purchasing a healthcare stock of a certain company indicates that you hold a part of the company shares. Specifically, various examinations have been led to anticipate the development of financial exchange utilizing AI calculations, such as SVM and reinforcement learning. A collection of machine learning algorithms are executed on Indian stock price data to precisely come up with the value of the stock in the future. Experiments are performed to find such healthcare sector stock markets that are difficult to predict and those that are more influenced by social media and financial news. The impact of sentiments on predicting stock prices is displayed and the accuracy of the final model is further increased by incorporating sentiment analysis.

KEYWORDS

Healthcare Stocks, Machine Learning, Neural Networks, Recurrent Neural Network, Sentiment Analysis, Stock Market Prediction

INTRODUCTION

The task of predicting future value of stocks accurately is Stock Market prediction. A group of a large section of investors and vendors is a stock market in simple terms. When a person or a group owns a stock, it denotes ownership rights of a company or organization to that particular group or individual. The challenging and demanding task of predicting future value of stocks accurately is Stock market prediction. Efficiency, robustness and accuracy are a must criterion for the prediction. The model must be built in such a way that it can adapt to a real world scenario and must perform well in it. The model is required to take each and every factor on which the stock market depends into consideration when the model is built.

DOI: 10.4018/IJISMD.303131 *Corresponding Author

The process of this prediction has entered in the world of technology by the exponential advancement of the digital age. Machine Learning is the process of making a machine to learn a task by making it observe a set of operations performing the same task with each iteration increasing its efficiency of performing the task. The inclusion of Machine learning in this process of predicting has been ground-breaking with algorithms using Neural Networks like ANN and RNN showing great promise in the field of prediction. Without being manually coded repeatedly, the core logic of AI is utilized by Machine learning that makes past experiences serve extraordinarily as the model gains a lot of important insights and experience from it (*Sirimevan N. et. al., 2019*).

The activity of predicting the worth of stocks on a small forecast horizon is an unsystematic and arbitrary process. When we consider the price of a stock on a long timeline, the graph formed by its prices over the years is fairly linear. It is the tendency of the general public to gain maximum profit, so they always are in line to purchase the stocks whose values are inclined to rise in the coming months or even years. In real life, the stock markets are largely erratic as they depend on a large number of factors. To predict the constantly changing and irregular stock prices in the stock market, the model includes a time series forecasting and analyses the data practically to model the system. Usually, the datasets used for predicting stocks contain data about various variables like the closing stock price and its opening price for each and every day of the forecast horizon that are absolutely essential for the stock value prediction process. The model is intended to learn various strategies and hidden patterns about the rise or fall of stock prices by applying the selected efficient machine learning algorithms on the historical dataset. (*V. Kranthi Sai Reddy, 2018*).

Prediction of Stock Market is an intriguing topic for investors and researchers since a number of years because of its inherent instability. Its nature of changing constantly and being unpredictable makes precise prediction a difficult and challenging task. Therefore predicting worth of stock market regularly is a tough and demanding challenge for buyers as well as shareholders. A large section of data scientists have been working on stock market prediction to get a more defined and accurate model since a few years. In the project we have implemented some of the best machine learning techniques based on the most recent research papers on the historical dataset to find out the algorithm which performs the best i.e., the algorithm which precisely predicts the stock price. We also know that the value of these stocks vary drastically depending upon the current economical situation. So we have decided to incorporate sentimental analysis to show the effect of news and current affairs on stock market prices. The accuracy of prediction of this efficient model is further increased by incorporating sentimental analysis

BACKGROUND

The efficiency of the stock market prediction model is constrained by the fact that the prices of stock prices are highly unpredictable, volatile and constantly changing. The accuracy of prediction is also affected by a set of other economic and social factors which cannot be digitalized as per today's technology. Precisely predicting the value of a stock also depends on various conditions like supply and demand, psychology and behaviour of investor etc. are some of the theories that cannot be incorporated into the model right now but there are chances of technologies emerging in the future which take advantage of such factors to generate a more accurate result. However, we have included twitter sentiment analysis which improves the accuracy of prediction by incorporating twitter data about current ongoing events. (*P. Ladyzynski et. al., 2013*).

Alternatives to this mode of stock market prediction include prediction methods of the past like Fundamental Analysis and Technical analysis. The former method involves analyzing the different factors that the stock market depends upon. Factors like balance sheets, cash flow and different statements from banks. Information about companies stakeholders, assets and issues are given by the balance sheet. Balance sheet provides data about the companies' assets and debts. Income statement lets us know if the company is making or losing money by displaying its revenues and

expenditures. The financial position of a company is decided by the fundamental analysis by making use of documents. Technical Analysis is used to foresee the sale of the stock depending upon price and amount of stock traded. Technical analysis takes various economic factors into account to unearth patterns of the stock prices rising or falling

Apart from the past methods we could also use various other algorithms for out model but due to their inefficiency of prediction or inability to fit to our problem are not used. The machine learning techniques used in our project have an edge over the earlier methods that require lots of time and resources to calculate predictions manually.

MAIN FOCUS OF THE ARTICLE

The project focuses on creating a model that predicts the future price of a stock using various using machine learning algorithms like Linear Regression, RNN with GRU, Random Forest Regressor and Support Vector Regressor. The model plots graphs for the stock market price by each of these algorithms. The model is fitted to the dataset and trained from the data in it from the past four years to finally make predictions of the values of stocks in the future. The data from NSE India website is used to train the model with the help of numerous libraries of machine learning. Widely used libraries like Numpy and Scikit are employed in the system. To clean the noisy data and scale it on a common basis converting it into a structure readily usable for training is done with the help of Numpy. The final predictions are made and machine learning algorithms are extracted and applied on the data by Scikit. Stock markets datasets extracted from the NSE India dataset is the dataset used in the project, the dataset is also split into different segments for training, validation and testing. The data from 2016, 2017 and 2018 stock prices for training, first 6 months of 2019 for cross validation and next 6 months of 2019 for testing.

Supervised learning is often used to find and recognise the hidden patterns and relations inside the dataset present in the training segment and exercise these new insights on the testing segment to find the accuracies of prediction. To combine the various datasets used in the project into one whole unified dataset, panda's library from python is operated. This final modified dataset prepares the data in way that makes it significantly easier to extract the required features from it. The features of the dataset are mainly dates of the timeline of the dataset used and closing prices of stocks of a certain day. These features are employed to make the system learn and train on all the algorithms to predict the price of stock, which is the price of it on a given day. We then compared the accuracies of all the prediction and chose the one with the highest accuracy. Using this algorithm we demonstrate the affect and impact of sentimental analysis on the accuracy of predicting stock prices by Web Scraping (*K. Hiba Sadia et. al., 2019*).

DESIGN METHODOLOGY

The initial step in our project is collecting the data. Once the data is collected it is pre-processed. The dataset employed in our project majorly contains the previous four years stock price observations from NSE India dataset that is analyzed and the and necessary pre-processing is done to remove noise and to scale the features. The data is then split into 3 sets: training, testing and validation sets. The training set contains the data which is used by the algorithms to understand various patterns and gain insights about the prediction. In simple terms it trains the algorithm to predict future stock prices. The cross validation compares the performances while tuning model's hyper parameters. The test set contains data which has been never seen by the algorithm. It estimates the performance of the model and accuracy of prediction.

We implement a set of efficient machine learning algorithms and compare their efficiencies and pick the one with the best performance for our machine learning model. The different algorithms chosen for comparison over our dataset are discussed here.

Linear Regression

Linear Regression is the most fundamental and basic machine learning techniques to predict linearly increasing data. This is included in the project to show the contrast of the drastic difference between basic and efficient techniques. The sole purpose of using this simple algorithm is to contrast the accuracy between one of the most basic algorithms and some of the best algorithms. Its major flaw is that it is highly susceptible to underfitting and sensitive to extreme data points or anomalies known as outliers which don't follow the regular pattern. The core idea behind linear regression technique is to find a relation linking dependent variable with the independent variables that results in a equation between them expressed as:

$$Y = \theta_1 x_1 + \theta_2 x_2 + ...\theta_n x_n \tag{1.1}$$

In the above equation, x1, x2 and so on till xn denote the independent variables whereas the quantities θ1, θ2 and so on till θn denote the weights. In our dataset, we have dates instead of a collection of independent variables; therefore, features regarding the date like day, month and year are extracted from the dataset to train the linear regression model.

RNN With GRU (Recurrent Neural Network With Gated Recurrent Unit)

RNN with GRUs are efficient algorithms which are known for performing exceptionally well on sequence prediction problems. RNN with GRUs have an advantage over conventional neural networks as they can remember patterns selectively for relatively long durations of time. GRU has a unique property that improves their efficiency and performance over others which is to remember information from the past that is considered important while choosing to forget the data that isn't. It is significantly better than LSTM for our dataset as it can be trained faster, is computationally efficient and it needs lesser data to unearth the hidden properties. It has two gates:

Figure 1. Linear regression

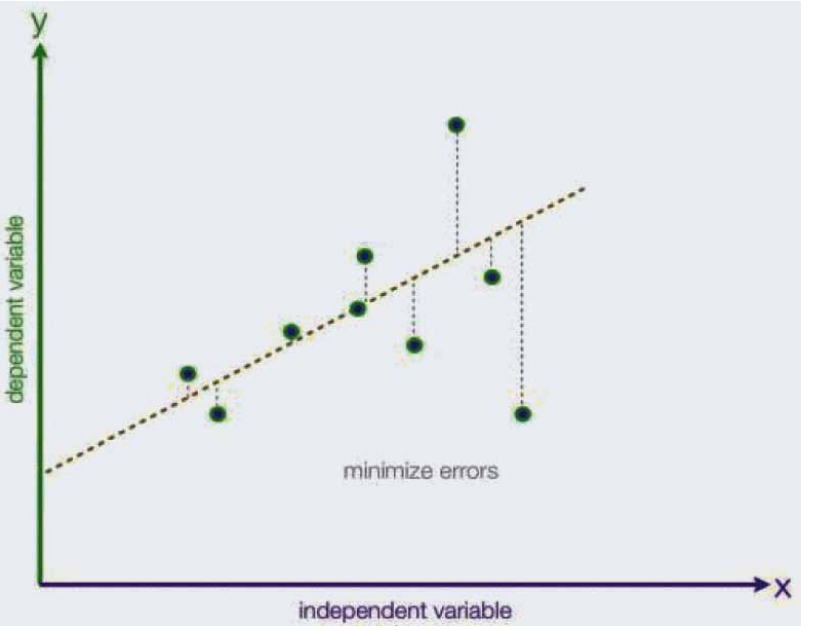
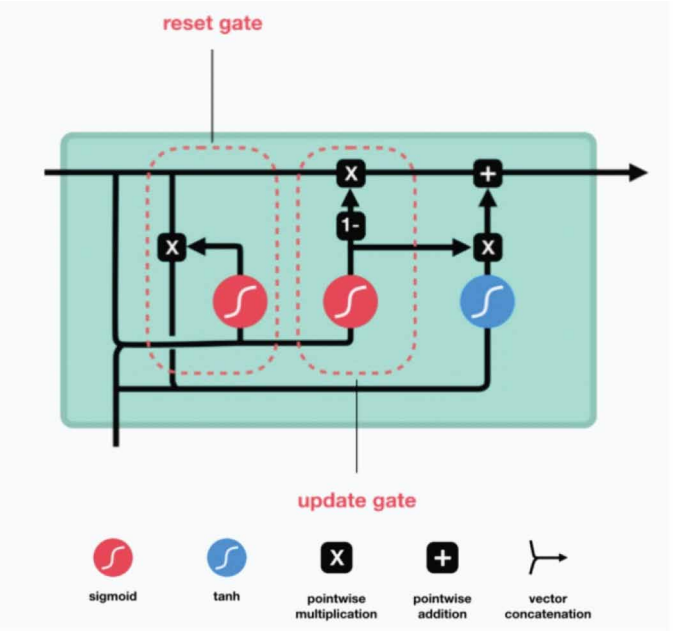


Figure 2. RNN with GRU (Source: Towards Data Science, 2018)



1. **The update gate:** It decides what information is relevant and what information is to be discarded.
2. **The reset gate:** Decided the amount of past information to be forgotten.

Random Forest Algorithm

One of the most efficient and flexible techniques utilized in the classification problems is the Random Forest algorithm. Because of the constant change and instability in stock market, predicting stock prices has always been a challenging task. Random forest regressor is used in prediction that has hyper-parameters similar to that of a decision tree. Revenue cost, Closing price and other economical factors are used to make the decision on the possible price of the stock by this algorithm. The random forest algorithm builds decision tress based on randomly selected features and observations and then finally it uses the aggregate of the results of this decision tree. The attributes and features of the problem decide the way in which the data is split. This algorithm is expected to perform really well as it can easily handle missing values and prevents overfitting. However it takes relatively longer to train than other algorithms. The stock price observations from previous four years extracted from the NSE India website is the dataset for creating Random Forest regression on the model.

Support Vector Machine Algorithm

Support Vector Machine algorithm works on the principle of finding out N-dimensional spaces that divides the observations into categories. The value of the variable N denotes the number of features. A set of possible hyperplanes exist among classes of observations and SVM aims to discover a plane that maximizes the margin between these classes. Increasing the distance between the observations of the classes is maximizing the margin. The advantage of increasing the margin is providing corroboration such that classification of observations can be easier later on. Hyperplanes are nothing but the decision boundaries that are required for better classification of observations between classes. The hyperplane divides the observations into various classes based on the location of these observations in the dataset.

Figure 3. Random Forrest algorithm

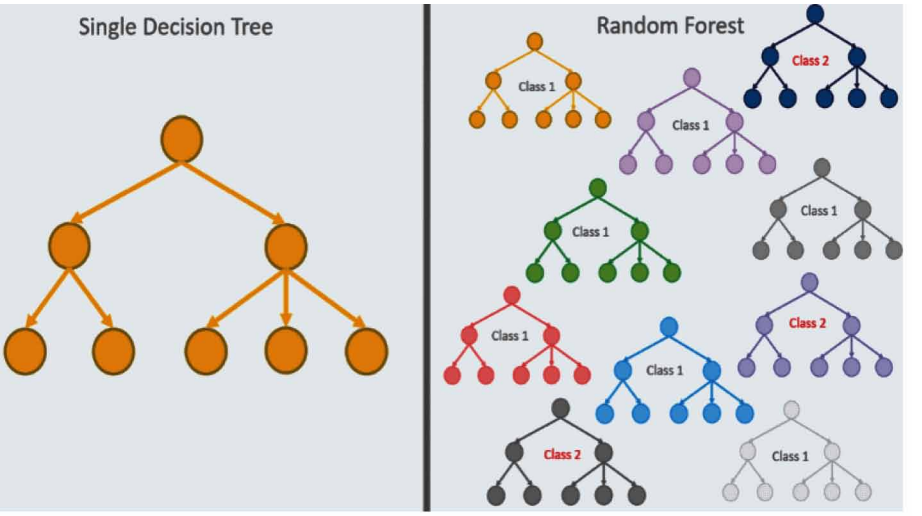
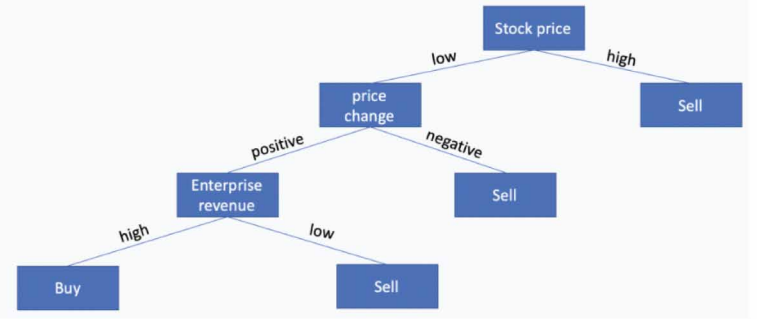


Figure 4. Each decision tree in the Random Forest



The number of features in a given challenge decides the dimension of the hyperplane, for example the hyperplane is a line if the features involved in the problem are just two or the hyperplane can be multi- dimensional if more than two features are involved. We picked SVM for predicting future stock values due to its ability to prevent overfitting by generalizing, its memory efficient and scales really well over our dataset. A few drawbacks of this model are longer training times, relatively harder to comprehend the model and to choose the optimum kernel. (*S. S. Patil et. al., 2016*).

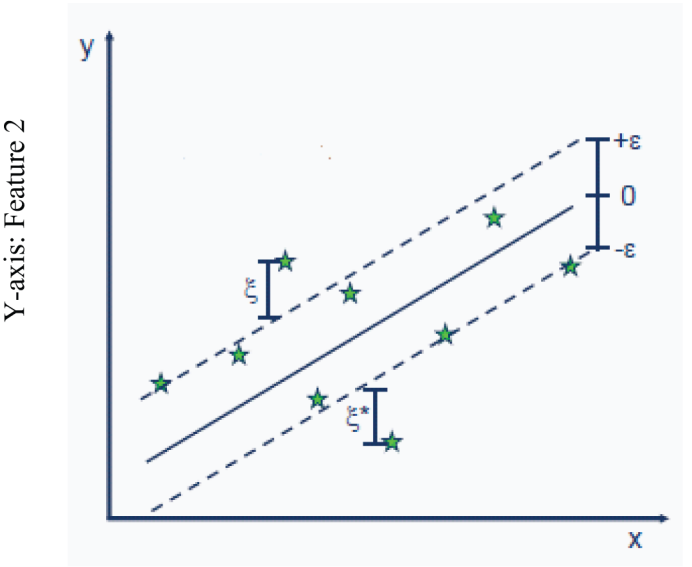
Minimize:

$$1 / 2w^2 + C\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right) \tag{1.2}$$

Constraints:

$$y_i - wx_i - b \leq \varepsilon + \xi_i \tag{1.3}$$

Figure 5. Support Vector Machine (Source: Towards Data Science, 2019)



$$wx_i + b - y_i \leq \varepsilon + \xi_i \tag{1.4}$$

$$\xi_i, \xi_i^* \geq 0 \tag{1.5}$$

where C is the constant in an epsilon sensitive tube

Sentiment Analysis Using Web Scraping

Sentiment Analysis is the process of mining text to extract information that helps to find important insights about emotions and sentiments of the general public. This technique is employed in many companies to excel as it involves gathering tons of data for research and with the development of technologies like Neural Networks, this technique has become very efficient in gathering vital insights which benefit the business. It is usually used in social media monitoring, brand monitoring and customer feedback.

Figure 6. Overview of Web Scraping

One of the methods to incorporate sentiment analysis is using Web Scraping. Web Scraping is a technique used to retrieve or scrape vast amounts of data that is stored on a local device or a database of a table. Since the Internet is an endless sea of data, this technique intelligently gathers relevant data automatically from the websites based on your requirement within seconds. Web Scraping can be done with the help of either a scraper or a crawler. The data that is extracted is used to make various vital decisions and processes.

PROPOSED MODEL

Data Collection

This is a fundamental part and one of the most essential and the primary stage of the project. It is about collecting the correct and exact information from a relevant dataset. The dataset used in this project is collected from NSE India that provides information on the stock prices of various Indian companies. This dataset is chosen as it contains minimal noise and the data is precise without unwanted inconsistencies or errors. Data collection is vital step as collecting and extracting irrelevant data can result in significant errors in the final system. Our data mostly contains healthcare stock price data of the years 2016, 2017, 2018 and 2019. The most important fields in the dataset are opening and closing prices of stocks of each day and features like price volatility and momentum which are used to estimate the direction of the trend. All of the models are trained by utilizing these values. The NSE India dataset is analysed in this stage and it is important that we gather the right data as this will further be utilised for performing the next stages of the project perfectly.

Pre-Processing

The process of transforming raw information to a meaningful format is data pre-processing. Pre-processing of data comes under Data mining. The problem with raw data is that it is generally incoherent or lacking containing various kinds of errors. Searching for values that are missing, checking the entire dataset to find out any categorical data, dividing the extracted data into various subsets are some of operations performed during this stage. Finally all the features are scaled into a certain range so that they can be evaluated on common ground. Our paper involves minimal pre-processing as the data used is already clean. The dataset is also split into different segments for training, validation and testing. The data from 2016, 2017 and 2018 stock prices for training, first 6 months of 2019 for cross validation and next 6 months of 2019 for testing (*Raut Sushrut Deepak et. al., 2017*).

Training the Model

This step is almost like providing the required information to each of the algorithms applied in our system to train the model for performing a specific task. The segments of the dataset put aside for training are utilised now to fit the dataset to the algorithms. Since the models must be judged based on data that is not seen by the algorithm, the section of dataset put aside for testing is never used in training the system. Various techniques are involved in training a model and an approximate performance of the system on training set is also obtained after training. Parameter tuning is applied in this stage where the hyper-parameters like the number of trees in a random forest are specifically tuned for our problem. The best hyper-parameters are then chosen by calculating a score for each of the set of these parameters. The training process of the system on a dataset begins with some estimated initial values that are altered and optimized to best fit the model. Until the required accuracy is obtained, the parameters are optimized repeatedly. The project we work on utilises data from 2016, 2017 and 2018 stock prices from NSE India website for training the model (*Vivek Kanade et. al., 2017, Osman Hegazy et. al., 2013*).

Testing the Data

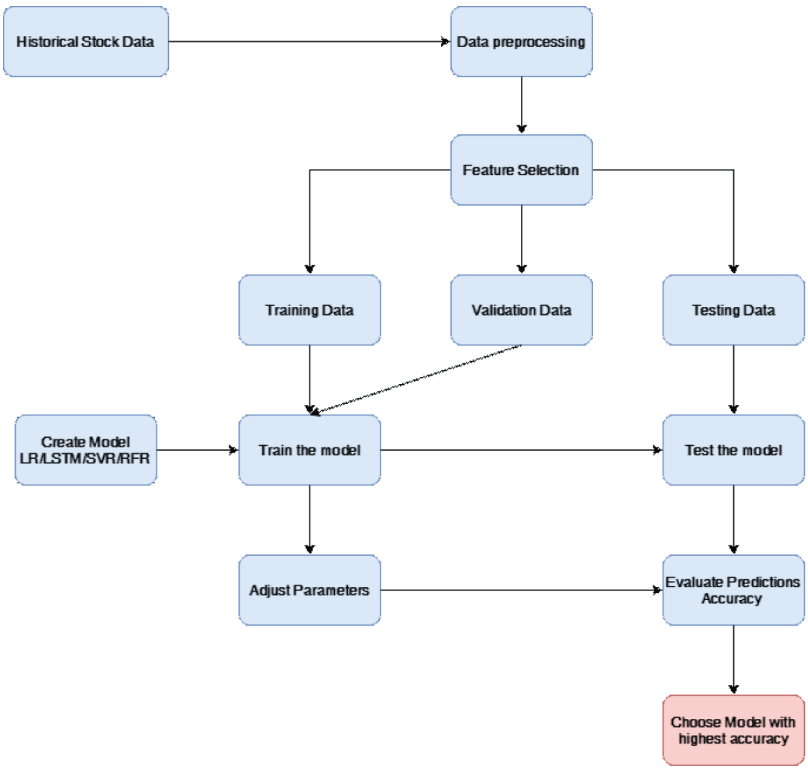
Test set is a collection of observations from the dataset which measure the performance of the algorithms based on a performance metric. The data in the test set must never contain observations from the training set. It will be a difficult task to determine if the model has learned to predict new values from the training dataset or has simply memorised it. Assessment of the algorithm if it has learned or just memorised will be difficult if the test set contains examples from training set. Precision of predicting new answers is provided by the test data after the model is trained. The 6 months of observations of 2019 which was stored for testing is now utilized to test the accuracy of each of the algorithm.

In predicting numerical data, Root Mean Square Error (RMSE) is a traditional way to compute the error of a model. We use this RMSE metric to determine the performance of each algorithm. The accuracies of all the algorithms are compared and the one with the lowest RMSE value is chosen for our model (*Prakash Ramani et. al., 2013*).

Sentiment Analysis Using Web Scraping

Sentiment Analysis is the process of mining text to extract information that helps to find important insights about emotions and sentiments of the general public. Web Scraping is a technique used to retrieve or scrape vast amounts of data that is stored on a local device or a database of a table. Since the Internet is an endless sea of data, this technique intelligently gathers relevant data automatically from the websites based on your requirement within seconds. In the project we use spider crawler for web scraping. We used web scraping to show the impact of sentiments on healthcare sector. The crawler gathers information about healthcare stocks from moneycontrol.com. The crawler searches

Figure 7. Stock Script Prediction Model



for the word "Pharma" in the entire website and provides relevant information about the emotions and current events regarding the company. It also assigns a score which determines the affect of news on the current stock market. This is incorporated to display the affect of sentiment analysis on stock price prediction and how it increases the accuracy (*Wasiat Khan et. al., 2019*).

RESULTS AND DISCUSSION

The normalization of data to a common scale of values ranging from 0 to 1 helps the performance of the model. The data is normalized by using MinMaxScaler function. Redundancies and inconsistencies are removed during this phase. Linear Regression, Support Vector Regressor, Random Forest Regressor and RNN with GRU models are fitted to the historical datasets and the future stock prices are calculated along with the accuracy of each of the models. The graphs are plotted with predicted stock values for each of the models.

Web Scraping and Sentimental Analysis

Scraping: Data is scraped from moneycontrol.com using spider web crawler.

This information is merged with existing historical data and the most efficient algorithm i.e., RNN with GRU is implemented on this new dataset to show the impact of sentiment on predicting stock values.

So, it is clear that RNN with GRU is the best algorithm for stock price prediction for the used historic data since it has lowest RMSE value (0.051). Hence we picked RNN with GRU for our model and performed sentiment analysis.

As we can see form figure 13 and 14, the accuracy of prediction has clearly increased after the incorporation of news sentiment in our RNN with GRU model. Hence we can say that implementing news sentiment in stock market prediction has a positive effect on the precision of the prediction.

Figure 8. Process Model

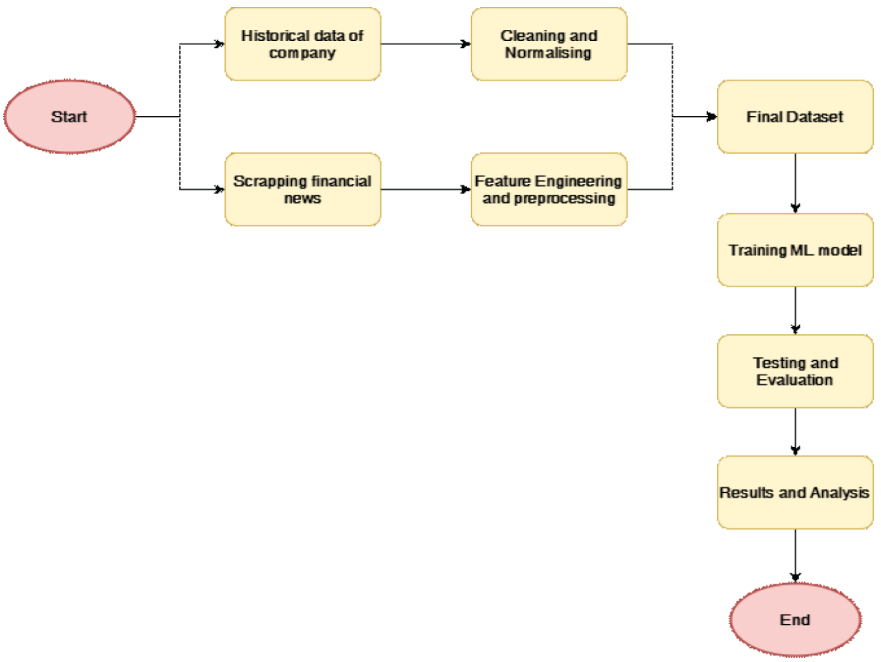


Figure 9. Linear Regression graph

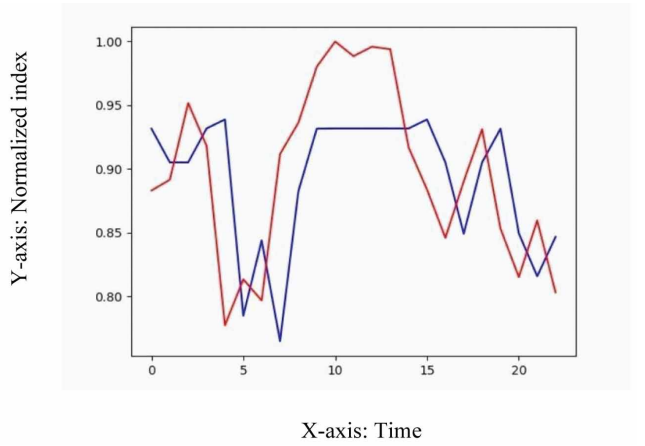


Figure 10. SVM graph

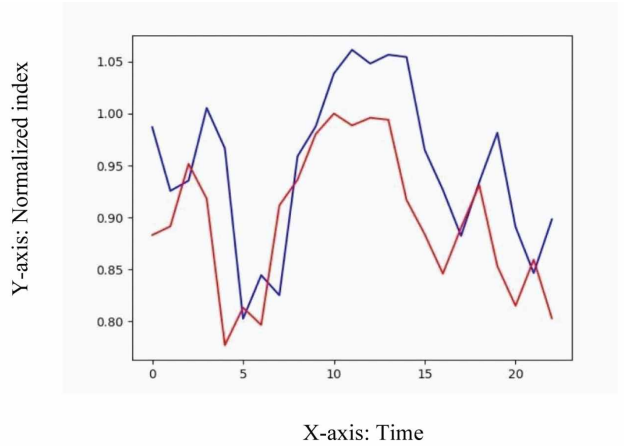


Figure 11. Random Forest graph

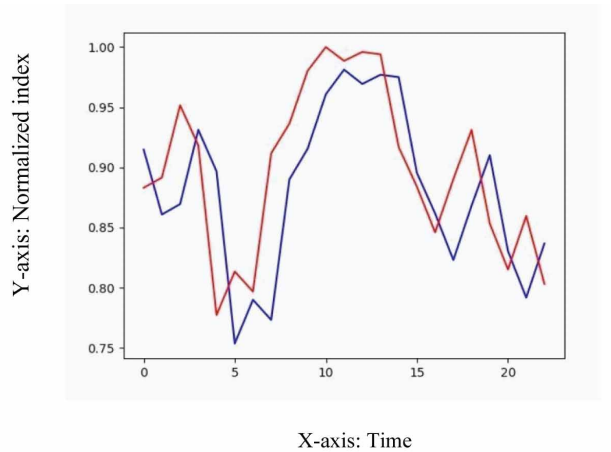


Figure 12. RNN with GRU graph

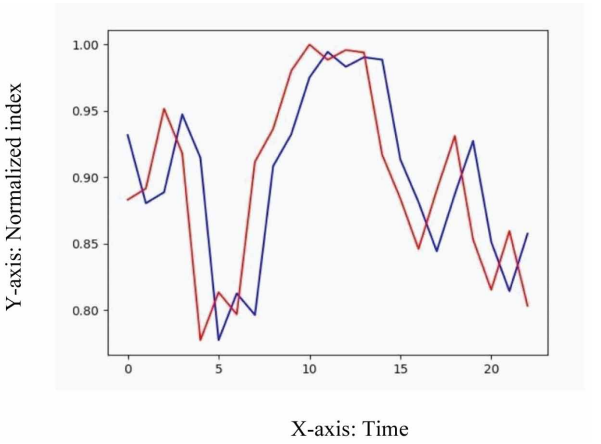


Figure 13. Analyzing the Scraped news

	A	B	C	D	E	F	G	H	I	J	K
1	Date	violate	slaps	fine	loss	poor rating	unstable	ied perforr	weakness	bearish	dued demo
2	18-12-201	0	0	1	2	0	0	0	0	0	0
3	05-12-201	0	0	0	1	0	0	0	0	0	0
4	28-11-201	0	0	1	0	0	0	0	0	0	0
5	28-11-201	0	0	1	1	0	0	0	0	0	0
6	22-11-201	1	0	1	2	0	0	0	0	0	0
7	21-11-201	0	0	1	1	0	0	0	0	0	0
8	20-11-201	0	0	1	1	0	0	0	0	0	0
9	18-11-201	1	0	0	1	0	0	0	0	0	0
10	11-11-201	0	0	1	0	0	0	0	0	0	0
11	08-11-201	0	0	0	1	0	0	0	0	0	0
12	07-11-201	1	0	1	0	0	0	0	1	0	0
13	06-11-201	0	0	0	1	0	0	0	0	0	0
14	06-11-201	0	0	0	2	0	0	0	0	0	0

CONCLUSION

After applying a set of Machine Learning algorithms to our dataset, we can say that RNN with GRU is the best for stock market prediction since it has the lowest RMSE value (0.051). Using this algorithm for our final model, we have implemented Sentiment analysis using web scraping. The results from Fig 16 clearly show that the error after applying sentiment analysis is lower than that of the model without it. Hence, we can say that RNN with GRU is the most efficient model for our dataset consisting of healthcare stock market price observations of four years (2016, 2017, 2018, and 2019) from NSE India and incorporating Sentiment analysis has a positive impact on stock price prediction and improves the accuracy of prediction. Utilizing this final model one can gain tremendous profit by investing and selling stocks in the health sector at the suggested time. Machine learning is an emerging technology with unlimited potential, it can be leveraged in health sector for improved drug discovery, better storage of EMR, predicting onset of a disease in an individual.

Figure 14. Calculating Sentiment of news

	A	B	C	D	E
1	Date	POS_SCOR	NEG_SCOR	Ratio	Net Score
2	28-11-2019	8.133333	3	2.711111	5.133333
3	22-11-2019	24.4	7	3.485714	17.4
4	21-11-2019	20.33333	8	2.541667	12.33333
5	20-11-2019	8.133333	4	2.033333	4.133333
6	18-11-2019	8.133333	8	1.016667	0.133333
7	11-11-2019	4.066667	5	0.813333	-0.93333
8	08-11-2019	12.2	2	6.1	10.2
9	07-11-2019	8.133333	3	2.711111	5.133333
10	06-11-2019	16.26667	10	1.626667	6.266667
11	05-11-2019	20.33333	8	2.541667	12.33333
12	04-11-2019	12.2	3	4.066667	9.2
13	02-11-2019	16.26667	8	2.033333	8.266667
14	31-10-2019	24.4	18	1.355556	6.4
15	25-10-2019	4.066667	4	1.016667	0.066667
16	23-10-2019	8.133333	3	2.711111	5.133333
17	22-10-2019	16.26667	6	2.711111	10.26667
18	21-10-2019	12.2	12	1.016667	0.2
19	16-10-2019	12.2	3	4.066667	9.2
20	14-10-2019	24.4	7	3.485714	17.4
21	11-10-2019	16.26667	14	1.161905	2.266667
22	09-10-2019	28.46667	11	2.587879	17.46667

Table 1. RMSE values of all the models

Algorithm	RMSE value
Linear Regression	0.080
RNN with GRU	0.051
SVM	0.079
Random Forest	0.065

Figure 15. Graph without News Sentiment Analysis

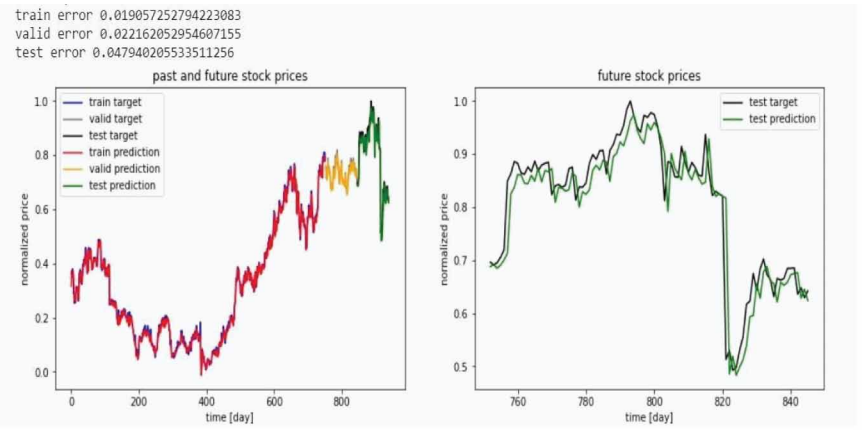
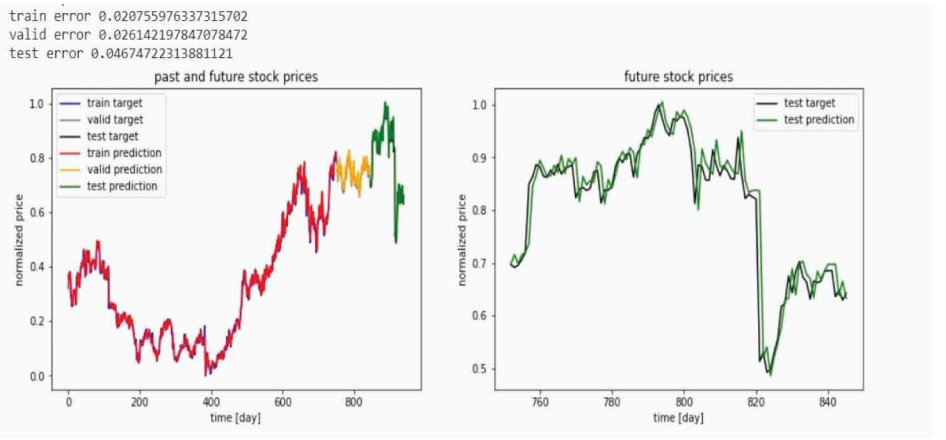


Figure 16. Graph with News Sentiment Analysis



ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

FUNDING AGENCY

Open Access Funding for this article has been covered by the authors of this manuscript.

REFERENCES

Deepak, Uday, & Malathi. (2017). Machine Learning Approach In Stock Market Prediction. *IJPAM, 115*(8), 71-77.

Hegazy, O., Omar, S. S., & Salam, M. A. (2013). A Machine Learning Model for Stock Market Prediction. IJCST, 4(12), 17-23.

Kanade, V., Devikar, B., Phadatare, S., Munde, P., & Sonone, S. (2017). Stock Market Prediction: Using Historical Data Analysis. IJARCSSE, 7(1), 267-270.

Khan, W., Malik, U., Ghazanfar, M. A., Azam, M. A., Khaled, H. A., & Ahmed S. A. (2019). Predicting stock market trends using machine learning algorithms via public sentiment and political situation analysis. Springer.

Kranthi Sai Reddy, V. (2018). Stock Market Prediction Using Machine Learning. In *International Research. Journal of Engineering Technology*, *5*(10).

Ladyzynski, P., Zbikowsk, K., & Grzegorzewski, P. (2013). Stock Trading With Random Forests Trend Detection Tests and Force Index Volume Indicators. In Artificial intelligence and soft computing (vol. 1, pp. 441-452). Academic Press.

Patil, Patidar, & Jain. (2016). A Survey on Stock Market Prediction Using SVM. IJCTET, 2(1), 1-7.

Ramani, & Murarka. (2013). Stock market Prediction Using Artificial Neural Network. *International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(4).

Sadia, Sharma, Paul, Padhi, & Sanyal. (2019). Stock Market Prediction Using Machine Learning Algorithms. IJEAT, 8(4).

Sirimevan, N., Mamalgaha, I. G. U. H., Jayasekara, C., Mayuran, Y. S., & Jayawardena, C. (2019). Stock Market Prediction Using Machine Learning Techniques. *International Conference on Advancements in Computing (ICAC)*. doi:10.1109/ICAC49085.2019.9103381