Survey on Method of Stock Market Forecasting

Aayush Shah[1] Harmit Sampat[2] Shail Vira[3] Nikita Raut[4]
[1,2,3,4]Department of Computer Engineering
[1,2,3,4]K. J. Somaiya College of Engineering, Mumbai, India

Abstract— Stock market prediction has been always a centre of attraction from people working in financial domain, over the years there has been a lot of forecasting made through regression, Machine learning algorithms, sentimental analysis etc. We have studied how these methods have been efficient and accurate in forecasting the stock prices, all the methods that have been studied have different methodology and feature selection however it is mutually agreed upon that the stock prices depend on a lot of features and all of them can never be truly accounted for however the models have been quite successful in forecasting the stock market upto a great extent which is helping the people who are working in the financial domain.

Key words: Stock Market, Stock Price Movements, Current Market Value, Return on Equity, Neural Networks, Machine Learning, Accuracy

I. INTRODUCTION

Data mining is defined as a process used to extract usable data from a larger set of any raw data. It implies analysing data patterns in large batches of data using one or more software. Data mining has applications in multiple fields, like science and research. As an application of data mining, businesses can learn more about their customers and develop more effective strategies related to various business functions and in turn leverage resources in a more optimal and insightful manner. This helps businesses be closer to their objective and make better decisions. A forecast is a prediction, which means talking about the future. It is an assumption, sometimes based on realities or proofs, but not always. Forecasting is a foreseen activity based on a link between a thing you already know and what you want to predict. Forecasting is mainly concerned with accuracy. Prediction can be done through a variety of algorithms like Regression, K-means Clustering, Support Vector Machines, Self Organizing map, Artificial Neural Networks, Bayesian Network, Lexical Method, and Ensemble Learning. Prediction can be used for different purposes like predicting rain fall, stock market analysis, breast cancer detection, skin cancer detection, housing prediction, crime prediction. By using the above mentioned techniques, we can reduce the risks and minimize the error by improving accuracy in the prediction value.

A stock market, equity market or share market is the aggregation of buyers and sellers of stocks. It represents ownership claims on businesses which may include securities listed on a public stock exchange as well as those only traded privately. Stock exchanges list shares of common equity as well as other security types, e.g. corporate bonds and convertible bonds. Companies go public which means they make their shares available to common people. Such companies list their shares on stock exchange. Some of well-known stock exchanges throughout the world are NSE (National Stock Exchange, India), BSE (Bombay Stock Exchange), New York Stock Exchange.

The factors which influence the stock market are of dynamic nature. For the development of a good simulator, these factors would play a vital role. These factors would influence the methods to customize simulator, the technologies which would be used in the simulator and how effectively the simulator can be built.

In upcoming sections, we discuss some recent approaches, which perform accurate stock market prediction with the help of various algorithms.

II. LITERATURE REVIEW

A. Neural Networks

Stock market prediction are carried out on the basis of Efficient Market Hypothesis and Random walk theory. Here, we have made two assumptions, one is that Investors are rational and they have all the available information about the stock market but in real life scenario this is not possible. We are going to predict the stock market using Recurrent Neural Network using Grated Recurrent Units to predict the stock market using news from verified accounts on Social Media. The Stock prices have been predicted through sentiment analysis and news analytics, so online social media is also an important factor that has been used for prediction of stock market. This prediction has been done on Chinese Stock Market and the social media platform chosen is Weibo, 100 verified accounts were taken and their posts were crawled. They were filtered of posts like good morning and other irrelevant things that are of no use. The news that has been posted a day before plays a lot of relevance in prediction. News of natural disaster will have an adverse effect and declaration of an economic zone will have a positive effect.

$$V = \frac{1}{N}\sum_{j=1}^{n} v$$

Where,
n is the news items on day i.
v is the influence of each item.

B. Machine Learning

Machine learning is an application of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. Thus, machine learning is an excellent approach for prediction of stock market. The prediction methods used is regression. In statistics, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. Regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent

variables are held fixed. The various techniques for regression are as follows:

– Polynomial Regression
It is a form of linear regression in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial. We can model the expected value of y as an nth degree polynomial, yielding the general polynomial regression model.
$y = a_0 + a_1x + a_2x_2 + \ldots\ldots + a_nx_n$
Where, x = Input, y = Goal, a0, a1, a2,..an = Unknown Parameters

– RBF Regression
A radial basis function (RBF) is a real-valued function whose value depends only on the distance from the origin. Any function that satisfies the property is a radial basis function:
$\varphi(x) = \varphi(||x||)$

– Sigmoid Regression
A sigmoid function is a mathematical function having an "S" shape (sigmoid curve). Often, sigmoid function refers to the special case of the logistic function shown in the first figure and defined by the formula:

$$S(t) = \frac{1}{1 + e^{-t}}$$

– Linear Regression
Linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory (or independent variables) denoted X. If the goal is prediction, linear regression can be used to fit a predictive model to an observed data set of y and X values. After developing such a model, if an additional value of X is then given without its accompanying value of y, the fitted model can be used to make a prediction of the value of y.

C. Deep Learning

The stock market is a very volatile entity which changes very dynamically, while predicting stock price of a company we assume EMH (Efficient Market Hypothesis) but in reality we know that it is impossible to know beforehand all the information related to that stock. We would use Neural Network because they possess striking capacity to extract information from the estimated information. Secondly, Neural networks exhibit nonlinear relation, which is quite frequent in this field and the same network can be retrained for a different dataset under same conditions. We will now be looking at how we have implemented the Multilayer Perceptron Network. We have a very randomised set of data when we take a company's stock for a week in minute by minute data. So, we normalise it.

$$Xstd = \frac{(X - Xmin)}{Xmax - Xmin}$$

$$Xscaled = Xstd * (max - min) + min$$

MLP normalises this values which go in as the input. Another important aspect in the model is the selection of features, the financial indicators used can be categorized into three parts.

– Trend type which indicates whether the stock prices are bullish or bearish and they include moving average (short), moving average (long), exponentially moving weighted average (EWMA), and Moving Average Deviation Rate (MAD).
– Oscillator type which indicates the possibility of reversal of a trend which has like moving average convergence/divergence (MACD), rank correlation index (RCI) and relative strength index (RSI).
– Momentum which is rate of change of stock prices.

Two models have been implemented to forecast the stock prices, one is LSTM (Long Short Term Memory) and MLP (Multilayer Perceptron).

– LSTM

It has one input layer taking inputs on the basis on 3 aforementioned features and has 4 hidden layer fully connected to input and output layer, the output layer has one cell that will be predicting the stock price of the second minute of the data, this model is beneficial because LSTM are known to remember past values which is very critical in stock market forecasting, we have used a sliding window of past 20 prices which helps to treat the problem as time-series problem.

– MLP

They are similar to the LSTM and contain 4 hidden layers and similar number of neurons and functionalities as LSTM but they do not have the tendency to remember past values, so the predicted price is dependent on the previous minute value unlike the sliding window in LSTM.

The result have been obtained from test data which is 30% of dataset from the past year data. LSTM predicts the downward or upward scale perfectly, however the variation in actual price and predicted price is not a small number on the other hand MLP is able to predict the future increase or decrease in trend with high accuracy and accurate prices.

This research throws light on how neural networks can be used to predict stock prices based on features which are derived from the heuristic method of analysis which are beneficial for frequently traded stocks.

D. Semantic Association Rule Mining

Data mining aims at finding interesting patterns in data. Pattern discovery is an active aria mining task which is searching for any regularity in data. With the rise of the Semantic Web, an interest has grown in employing knowledge hidden in the Semantic Web. Traditional algorithms focused on either semantic data or the time frame, no discovery was made in respect to both of those things combined. The main criteria of getting to know the frequent patterns are confidence, and support. This paper focuses on combining the use of semantic data along with using a suitable time frame so that they can predict the stock prices with a better accuracy. The main contributions of this paper are: a) They find frequent sequences by applying sequential pattern mining techniques on triple data. b) To show how ontology based association inferences and decreases the at the start and at the end of the process. Such a strategy allows for a more versatile approach because it allows to reuse the data mining algorithms and techniques. The work is divided into numerous stages which are shown as follows:

– Pre-Inference Phase

This phase is used to decrease the number of triples and allows for a more efficient pruning of sequences.

– Middle Inference

We perform item relatedness transaction filter on large databases and try to predict the transactions of the data. The data which is not useful is pruned and not considered in the next phase.

– Post Processing

Minimum improvement constraint filter (MICF) which helped to get comprehensive rules instead of detailed rules is applied to prune a lot of rules at the end of the cycle. Thus, by applying MICF and then replacing it with triples will form a sequential form of data.

The results of this experiment showed that larger minimum support generates more transactions and rules, while smaller minimum support generates lesser transactions and rules. The results of this experiment show that data mining resulted in valuable rules which could be used for stock market prediction.

E. Modified Back Propogation Network

Back Propagation neural network is the most universal and successful prototype for composite multi-layer networks. A typical back propagation network consists of three layers: input layer, output layer and at least one hidden layer. The network depends on the no. of neurons at each layer and the no. of hidden layer to achieve the correct result. The experiment done with this approach will compare the actual value to the predict value and calculate the error minimization. The proposed algorithm is to pre-process the data and apply modified back propagation neural network with the current trends and events taken into consideration. The steps involved are:

– Pre-processing

The data that is gathered from various locations is cleaned, reduced, and data transformation and data discretization is performed on the data to produce only meaningful data.

– Current Trend & Event Analysis

The current performance of the trend is evaluated and the events like financial budget, quarterly performance, new project in pipeline, election, etc are considered for analysis.

– Feature Extraction

A procedure which contains different features like no. of volume, average no. of bid, no. of positive and no. of negative circuit during that time period is performed.

– Finding Threshold

After the completion of the above mentioned steps, if the predicted value is lesser than the current value, then the share should be considered bad. If the predicted value is higher than the current value, then the share should be considered as good and it should be considered to be bought. The past threshold values are required to be calculated for the same share which will allow to find the error and hence minimize the error by adding or removing the threshold value.

The experiment done can prove that the errors can be reduced and optimized in a better way by using the threshold value and considering the past events.

F. K-Means Clustering with Genetic Algorithm

The K-Means algorithms is one of the simplest and most popular unsupervised learning algorithms, and is widely used for performing clustering analysis in data mining operations. In K-Means, the first step is to randomly select K data points from the given dataset, to act as centroid. After the centroids are decided, all the remaining data points are assigned to a cluster corresponding to the centroid for which the value of the euclidean distance between the data point and the centroid is the least. A new centroid is then calculated for the clusters, the value of the new centroid will be the mean of all the data points in the cluster. The two steps are then repeated till the the stopping criteria i.e, centroids stop moving. K-means often fails to distribute data points equitably due to poor selection of initial centroids. To solve this problem, this approach was used in combination with Genetic Algorithm to optimise the results. The genetic algorithm iteratively modifies the population of possible solutions to in an attempt to search for the most optimum clustering solution. The implementation was carried in the following manner:

– Providing the Algorithm with the Input

Population size, Number of clusters, Data set, Maximum number of generations, Targeted sum of square distances.

– Sort all the vectors

In ascending or descending order – which will cause the scrambled vectors to rearrange. Then, it can be assumed that the vectors with close characteristics in terms of word weight will be grouped together.

– Generate k initial random centroids using the initial centroid selection optimization method.
– Concatenate the updated centroids with their solutions to form the final structure of the chromosome. Pass these chromosomes to the GA and start operating on them in an attempt to search for the most optimum clustering solution.

The GA will calculate the fitness of all chromosomes regarding their centroids by calculating the sum of square distances between cluster elements and their centroid is:

$$f(C1, C2, \ldots CN) = \sum_{i=1}^{K} \sum_{xj \,\epsilon\, Ci} || xj - zi ||$$

– Finally, a clustering solution will be acquired to resolve the problem of calculating the average accuracy and to compare with previous results.

Accuracy was found to be about 89.31% which is a significant improvement over pure k-means (83.3%) and pure GA (82.5%).

G. Causal Relationship Mining

Here, the proposed framework attempts to find upstream and downstream causal relationships in a comprehensive manner by exploring the inter-transaction relationships that are latent in stock datasets. It tackles a common problem across most mining association rules, such rules consider any item set which does not satisfy the minimum support as uninteresting. However, in stock market dataset item sets that are infrequent also can provide information needed towards converging decisions. Two items a and b have no direct relationship but they may have strong relationship through another item set Y. In this case the a, b pair is said to have indirection relationship or association. Mining such rules is the focus of this paper. Mining indirect association rules is done using the following algorithm:

- The first step focuses on candidate generation, it makes use of an algorithm like Apriori to extract frequent item sets. This knowhow is used in the algorithm for finding indirect associations.
- The second step is for pruning, item sets that do not have mediator dependence are pruned and final associations that have relationships through mediator are extracted and rules are generated based on the support and confidence.

The empirical results revealed that the proposed algorithm is useful in analysing causal relationships, especially for two stock tickers having no direct relationship but are related through a third ticker (mediator). This kind of relationship can provide valuable business intelligence.

H. Hidden Markov Model

Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with hidden states. The models are made in accordance to the insufficient accuracy and precision present in forecasting the stocks. In this model, normal conditions of the market were assumed. The following step are performed in the implementation of this model:

- Data Collection

The dataset was collected for a time interval of three years for three different industries. 70% of the data was used for training and 30% for testing.

- Data Pre-Processing

Four variables are set which show fluctuations in the market which are trading volume, main index, industry index, and closing price. Each row of the data is compared to its previous row value and the type of change is termed as increase or decrease. The steps are data selection -> data filtration -> data interpretation -> data normalization -> data storage.

- Model Design Phase

Firstly, all types of transitions from one mode to another are specified and since there are 3 variables in two increasing and decreasing modes, there are a total of 36 modes. The probability of each mode is hence calculated and the primary developed model is trained with prepared data by Baum Welch algorithm. To forecast data about the current day, the information of previous days is trained.

- Implementation Phase

The market behaviour is forecasted using progressive algorithms. By receiving the parameters related to three variables of trading volume, main index, and industry index in the current day and specifying its increase or decrease mode in comparison with the previous day. If the calculated probability is less than 0.55, the market is forecasted as decreasing, otherwise it is increasing.

Criteria for Evaluation of precision, Recall, F1 measure, Specificity and Accuracy are among the parameters for evaluating forecasting algorithms. The precision criteria shows what percentage of increasing behavior is accurately forecasted based on FP parameters. The recall criteria is used for evaluating the number of true positive factors based on FN parameter. The F1 measure is the combination of recall and precision criteria. Specificity criteria shows what percentage of behaviors forecasted as decreasing are really decreasing. Accuracy criteria evaluates the quality of the algorithm. The specificity criteria proved to be the best among all the criterias.

I. Sentiment Transfer Learning

Sentimental transfer learning is an approach wherein the characteristics of source domain are transferred into the target domain. This method has been deployed in stock market prediction by taking a company in a particular domain and gathering news article related to the company and the market and analysing how the news piece will affect the company's stock price for which accuracy. Loughran-McDonald financial sentiment dictionary (LMD) is has been used which has predefined 3911 words and 6 sentiment dimensions. Now there might be companies which are in the same particular domain but those companies are news-poor, here we use our knowledge of sentiment transfer and the model that we have trained using the data set of news-rich company is deployed on the news-poor company and the results are achieved.

$$y = \begin{cases} +1, & \text{if } r \geq \theta \\ 0, & \text{if } -\theta \leq r \leq \theta \\ -1, & \text{if } r \leq -\theta \end{cases}$$

Where, θ is 0.5% and r is simple return which is calculated based on open and closed prices.

$$r = \frac{\text{Close} - \text{Open}}{\text{Open}}$$

J. Recurrent Convolutional Neural Networks

Recurrent Neural network is an algorithm for sequential data which remembers its input due to internal memory, which makes it perfectly suited for machine learning. Convolutional neural network is a type of neural network which has vectors as input and has the ability for feature extraction. Word vectors and historical information are taken as the inputs and trained overtime to reduce the price error to predict the future prices. Local feature extraction is done with the help of convolution neural networks, and the temporal and long dependencies of news and price is handled by long short term memory (LSTM), a type of recurrent neural network which contains memory cells named gates which are implemented as activation functions like sigmoid function in order to decide the amount of information that should be passed through it.

The evaluation was done by the following methods:

- Data

The data was collected from a variety of sources for a period of 10 years from June 21, 2007 to Feb 13, 2017.

- Evaluation Baselines Method

The baseline model is divided into two methods: Traditional Financial approach where the attempt is to compare the profits after a period of 240 days; Second, machine learning algorithms, which predicts future stock prices with the help of historical sequence of data.

- Evaluation in Terms of Profit or RMSE

The return on investment of each model is calculated. The difference between the test data and the prediction is calculated by root mean square error method.

This paper shows the result that the price of RCN is better than LSTM with lower RMSE.

III. PROPOSED METHODOLOGY & FUTURE WORK

Stock price prediction is a challenging task primarily due to two main reasons: (i) the volatile nature of the stock market and (ii) the very large number of factors that can have a potential impact on stock prices. The first problem could be dealt with by using Machine Learning. We have seen various regression techniques give fairly accurate predictions but they still left much to be desired. The Deep Learning and Artificial Neural Network models that followed show a substantial improvement in the accuracy of stock price prediction. By analysing the empirical results, we can see that LSTMs are able to predict the downward or upward scale perfectly, however fall short in being able predict the actual price where on the other hand, MLPs are better at identifying future trends with a high degree of accuracy. Meanwhile, we have also seen that RCN, out of all the models discussed so far, has the lowest rate of error. A result having close to perfect accuracy is yet to be seen due to not being able to account for external factors, which is the second problem. Here, we have the seen the use of various data mining techniques to extract knowledge that is latent in inter-transaction relations as well as outside sources.

Our study shows the use of Neural Networks, especially RNNs will be able to provide promising results if the results can be supplemented with additional business intelligence data. Hence, for future work in stock market forecasting, we recommend further research on using predictive models in combination with intelligent data mining algorithms for comprehensive results.

IV. CONCLUSION

The aim of our research study is to help the stock brokers and investors for investing money in the stock market. The prediction plays a very important role in stock market business which is very complicated and challenging process dueto dynamic nature of the stock market. This paper describes about various algorithms, methods and models involved determining the stock prices. It illustrates many techniques for classifying and analyzing data. This survey focus on different models like K-means clustering, regression, Deep Learning, Hidden Markov Model, and many more which is mainly involved in stock price forecasting.

REFERENCES

[1] Li, Xiaodong, et al. "Market impact analysis via sentimental transfer learning." Big Data and Smart Computing (BigComp), 2017 IEEE International Conference on. IEEE, 2017
[2] Chen, Weiling, et al. "Stock market prediction using neural network through news on online social networks." Smart Cities Conference (ISC2), 2017 International. IEEE, 2017.
[3] Sharma, Ashish, Dinesh Bhuriya, and Upendra Singh. "Survey of stock market prediction using machine learning approach." Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of. Vol. 2. IEEE, 2017.
[4] Siddiqui, Sehba Shahabuddin, and Vandana A. Patil. "Proposed system for estimating intrinsic value of stock using Monte Carlo simulation." Intelligent Computing and Control Systems (ICICCS), 2017 International Conference on. IEEE, 2017.
[5] Khare, Kaustubh, et al. "Short term stock price prediction using deep learning." Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2017 2nd IEEE International Conference on. IEEE, 2017.
[6] Bharne, Pankaj K., and Sameer S. Prabhune. "Survey on combined swarm intelligence and ANN for optimized daily stock market price." Soft Computing and its Engineering Applications (icSoftComp), 2017 International Conference on. IEEE, 2017.
[7] Asadifar, Somayyeh, and Mohsen Kahani. "Semantic association rule mining: a new approach for stock market prediction." Swarm Intelligence and Evolutionary Computation (CSIEC), 2017 2nd Conference on. IEEE, 2017.
[8] Mithani, Fesal, Sahista Machchhar, and Fernaz Jasdanwala. "A modified bpn approach for stock market prediction." Computational Intelligence and Computing Research (ICCIC), 2016 IEEE International Conference on. IEEE, 2016.
[9] Coyne, Scott, Praveen Madiraju, and Joseph Coelho. "Forecasting Stock Prices Using Social Media Analysis." Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence & Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), 2017 IEEE 15th Intl. IEEE, 2017.
[10] Shyamala, G., and N. Pooranam. "A survey on online Stock forum using subspace clustering." Computer Communication and Informatics (ICCCI), 2016 International Conference on. IEEE, 2016.
[11] Spector, A. Z. 1989. Achieving application requirements. In Distributed Systems, S. Mullender
[12] Lee, Che-Yu, and Von-Wun Soo. "Predict Stock Price with Financial News Based on Recurrent Convolutional Neural Networks." 2017 Conference on Technologies and Applications of Artificial Intelligence (TAAI). IEEE, 2017.
[13] Desokey, Eslam Nader, Amr Badr, and Abdel Fatah Hegazy. "Enhancing stock prediction clustering using K-means with genetic algorithm." 2017 13th International Computer Engineering Conference (ICENCO). IEEE, 2017.
[14] Li, Zhixi, and Vincent Tam. "Combining the real-time wavelet denoising and long-short-term-memory neural network for predicting stock indexes." Computational Intelligence (SSCI), 2017 IEEE Symposium Series on. IEEE, 2017.
[15] Bhoopathi, Harchana, and B. Rama. "A novel framework for stock trading analysis using casual relationship mining." Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2017 Third International Conference on. IEEE, 2017.
[16] Liang, Qiubin, et al. "Restricted boltzmann machine based stock market trend prediction." Neural Networks

(IJCNN), 2017 International Joint Conference on. IEEE, 2017.
[17] Asad, Muhammad. "Optimized Stock market prediction using ensemble learning." Application of Information and Communication Technologies (AICT), 2015 9th International Conference on. IEEE, 2015.
[18] Gao, Guangliang, et al. "A survival analysis method for stock market prediction." Behavioral, Economic and Socio-cultural Computing (BESC), 2015 International Conference on. IEEE, 2015.
[19] Chang, Pei-Chann, and Jheng-Long Wu. "The weighted Support Vector Machines for the stock turning point prediction." Intelligent Systems Design and Applications (ISDA), 2014 14th International Conference on. IEEE, 2014.
[20] Farshchian, Maryam, and Majid Vafaei Jahan. "Stock market prediction with Hidden Markov Model." Technology, Communication and Knowledge (ICTCK), 2015 International Congress on. IEEE, 2015.
[21] Im, Tan Li, et al. "Analysing market sentiment in financial news using lexical approach." Open Systems (ICOS), 2013 IEEE Conference on. IEEE, 2013.

