A BEST ALGORITHM FOR SHORT-TERM AND LONG-TERM STOCK PRICE PREDECTION

Surat Banerjee

Abstract— **Stock price forecasting is a popular and important topic in financial and academic studies. Share Market is an untidy place for predicting since there are no significant rules to estimate or predict the price of share in the share market. Many methods like technical analysis, fundamental analysis, time series analysis and statistical analysis, etc. are all used to attempt to predict the price in the share market but none of these methods are proved as a consistently acceptable prediction tool.**

In this project we attempt to implement different algorithms (Moving Average, Linear Regression, KNN, ARIMA, LSTM) to predict stock market prices and will check which algorithm gives us best prediction. Artificial Neural networks are very effectively implemented in forecasting stock prices, returns, and stock modeling, and the most frequent methodology. We select a certain group of parameters with relatively significant impact on the share price of a company. With the help of statistical analysis, the relation between the selected factors and share price is formulated which can help in forecasting accurate results. Although, share market can never be predicted, due to its vague dons in, this project aims at applying different Time Series Analysis Algorithm and Non Time series Analysis Algorithm in forecasting the stock prices.

Keywords: Machine Learning, Data Preprocessing, Data Mining, Moving Average, Linear Regression, KNN, ARIMA, LSTM, RNN, Backpropagation

I. INTRODUCTION

The stock market is basically an aggregation of various buyers and sellers of stock. A stock (also known as shares more commonly) in general represents ownership claims on business by a particular individual or a group of people. The attempt to determine the future value of the stock price is known as a stock price prediction. The prediction is expected to be robust, accurate and efficient. The system must work according to the real-life scenarios and should be well suited to real-world settings. The system is also expected to take into account all the variables that might affect the stock's value and performance. There are various methods and ways of implementing the prediction system like Fundamental Analysis, Technical Analysis, Machine Learning, Market Mimicry, and Time series aspect structuring. With the advancement of the digital era, the prediction has moved up into the technological realm. The most prominent and promising technique involves the use of Artificial Neural Networks, Recurrent Neural Networks, that is basically the implementation of machine learning. Machine learning involves artificial intelligence which empowers the system to learn and improve from past experiences without being programmed time and again. Traditional methods of prediction in machine learning use algorithms like Backward Propagation, also known as Backpropagation errors. Many researchers are using more of ensemble learning techniques. It would use low price and time lags to predict future highs while another network would use lagged highs to predict future highs.

Stock price prediction for short time windows appears to be a random process. The stock price movement over a long period of time usually develops a linear curve. People tend to buy those stocks whose prices are expected to rise in the near future. The uncertainty in the stock market refrain people from investing in stocks. Thus, there is a need to accurately predict the stock market which can be used in a real-life scenario. The methods used to predict the stock market includes a time series forecasting along with technical analysis, machine learning modeling and predicting the variable stock market. The datasets of the stock market prediction model include details like the closing price opening price, the data and various other variables that are needed to predict the object

[1]Surat Banerjee, Department of I.T specialist A.I of MAKAUT

variable which is the price in a given day. The aim is to design a model that gains from the market information utilizing machine learning strategies and gauge the future patterns in stock value development.

Stock market is the important part of economy of the country and plays a vital role in the growth of the industry and commerce of the country that eventually affects the economy of the country. Both investors and industry are involved in stock market and wants to know whether some stock will rise or fall over certain period of time. The stock market is the primary source for any company to raise funds for business expansions. It is based on the concept of demand and supply. If the demand for a company's stock is higher, then the company share price increases and if the demand for company's stock is low then the company share price decrease.

Another motivation for research in this field is that it possesses many theoretical and experimental challenges. The most important of these is the Efficient Market Hypothesis (EMH), the hypothesis says that in an efficient market, stock market prices fully reflect available information about the market and its constituents and thus any opportunity of earning excess profit ceases to exist.

II. PROBLEM DEFINITION

Stock price prediction is basically defined as trying to determine the stock value and offer a robust idea for the people to know and predict the stock prices. The problem with estimating the stock price will remain a problem if a better stock price prediction algorithm is not proposed. Predicting how the stock market will perform is quite difficult. The movement in the stock price is usually determined by the sentiments of thousands of investors. Stock price prediction, calls for an ability to predict the effect of recent events on the investors. These events can be political events like a statement by a political leader, a piece of news on scam etc. It can also be an international event like sharp movements in currencies and commodity etc. All these events affect the corporate earnings, which in turn affects the sentiment of investors. It is beyond the scope of almost all investors to correctly and consistently predict these hyper-parameters. All these factors make stock price prediction very difficult. Once the right data is collected, it then can be used to train a machine and to generate a predictive result.

III. OBJECTIVE

The aims of this project are as follows:

1) To identify factors affecting share price.
2) Extract the pattern from large set which is effect on stock price changes.
3) To predict an approximate value of share price.
4) Implement different algorithms on the same dataset and predict the stock price for the same period (Short-Term and Long-Term).
5) Check with Actual stock price with the predicted stock price.
6) Conclude which algorithm is the best.

The project will be useful for investors to invest in stock market based on the various factors. The project target is to apply different algorithm that analyses previous stock price data for a company and implement these values in data mining and machine learning algorithm to determine the value that particular stock will have in near future with suitable accuracy.

The main feature of this project is to generate an accurate forecasting output using different algorithm and create a general idea of future values based on the previous data by generating a pattern. The scope of this project does not exceed more than a generalized suggestion tool and also suggest the best prediction algorithm.

IV. FINDING

1) Monday returns remained less than other days and Friday returns remained greater than other days.
2) Positive Return is noticed for the months of January, February, June, July, August, and December higher than other months. Maximum return in the month of February.
3) Negative returns are noticed for the months of March, April, May, September, October, and November.
4) Semi-Month Effect is noticed in BSE, where higher return is noticed during first

half months than second half month.

5) The above literature survey confirms that the January effect stock returns are noticed in India.
6) Indian Tax year ends in March the above literature survey confirmed tax-loss selling in Indian market to tax year effect.

V. SYSTEM ARCHITECTURE

Kaggle is an online community for data analysis and predictive modeling. It also contains dataset of different fields, which is contributed by data miners. Various data scientist competes to create the best models for predicting and depicting the information. It allows the users to use their datasets so that they can build models and work with various data science engineers to solve various real-life data science challenges. The dataset used in the proposed project has been downloaded from Kaggle. However, this data set is present in what we call raw format. The data set is a collection of stock market information about a few companies.

The first step is the conversion of this raw data into processed data. This is done using feature extraction, since in the raw data collected there are multiple attributes but only a few of those attributes are useful for the purpose of prediction. So, the first step is feature extraction, where the key attributes are extracted from the whole list of attributes available in the raw dataset. Feature extraction starts from an initial state of measured data and builds derived values or features. These features are intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps. Feature extraction is a dimensionality reduction process, where the initial set of raw variables is diminished to progressively reasonable features for ease of management, while still precisely and totally depicting the first informational collection.

The feature extraction process is followed by a classification process wherein the data that was obtained after feature extraction is split into two different and distinct segments. Classification is the issue of recognizing to which set of categories a new observation belongs. The training data set is used to train the model whereas the test data is used to predict the accuracy of the model. The splitting is done in a way that training data maintain a higher proportion than the test data.

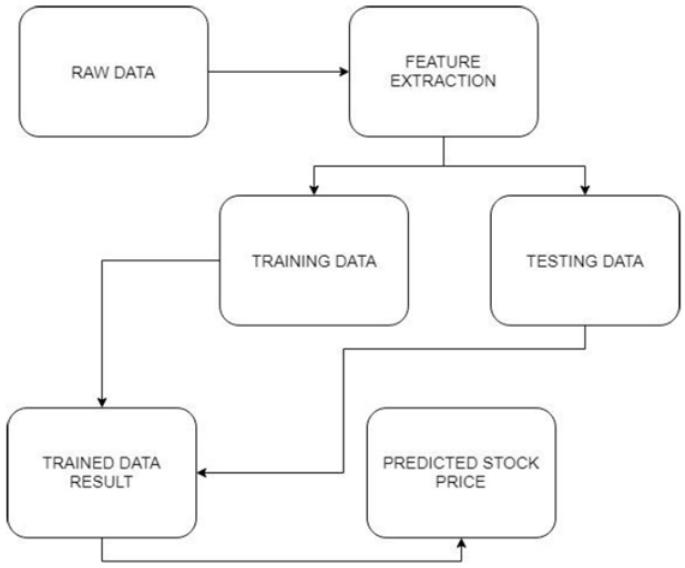

Fig. 1. System Flow Diagram

VI. MODULES AND FUNCTIONALITY

A. Data Collection

Data collection is a very basic module and the initial step towards the project. It generally deals with the collection of the right dataset. The dataset that is to be used in the market prediction has to be used to be filtered based on various aspects.

Data collection also complements to enhance the dataset by adding more data that are external. Our data mainly consists of the previous year's stock prices. Initially, we will be analyzing the NSE – TATA GLOBAL dataset from Kaggle and according to the accuracy, we will be using the model with the data to analyze the predictions accurately.

B. Pre-Processing

Data pre-processing is a part of data mining, which involves transforming raw data into a more coherent format. Raw data is usually, inconsistent or incomplete and usually contains many errors. The data pre-processing involves checking out for missing values, looking for categorical values, splitting the data-set into training and test set and finally do a feature scaling to limit the range of

variables so that they can be compared on common environs.

C. Training the Machine

Training the machine is similar to feeding the data to the algorithm to touch up the test data. The training sets are used to tune and fit the models. The test sets are untouched, as a model should not be judged based on unseen data. The training of the model includes cross-validation where we get a well-grounded approximate performance of the model using the training data. Tuning models are meant to specifically tune the hyperparameters like the number of trees in a random forest. We perform the entire cross-validation loop on each set of hyperparameter values.

Finally, we will calculate a score, for individual sets of hyperparameters. Then, we select the best hyperparameters. The idea behind the training of the model is that we some initial values with the dataset and then optimize the parameters which we want to in the model. This is kept on repetition until we get the optimal values. Thus, we take the predictions from the trained model on the inputs from the test dataset. Hence, it is divided in the ratio of 80:20 where 80 percent is for the training set and the rest 20 percent for a testing set of the data.

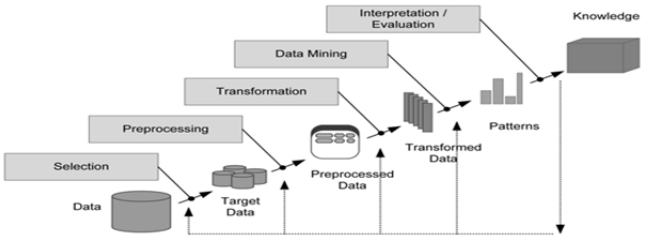

Fig. 2. Module Diagram

D. Data Scoring

The process of applying a predictive model to a set of data is referred to as scoring the data. Based on the learning models, we achieve interesting results. The last module thus describes how the result of the model can help to predict the probability of a stock to rise and sink based on certain parameters. It also shows the vulnerabilities of a particular stock or entity. The user authentication system control is implemented to make sure that only the authorized entities are accessing the results.

VII. METHODOLOGIES

WE WILL IMPLEMENT THE FOLLOWING ALGORITHMS FOR 15-DAYS, 45-DAYS AND 90-DAYS PREDICTION AND COMPARE RESULT TO FIND THE BEST ALGORITHM TO PREDICT THE SHARE PRICE

1) Moving Average
2) Linear Regression
3) K-Nearest Neighbors
4) ARIMA
5) Long Short-Term Memory (LSTM)

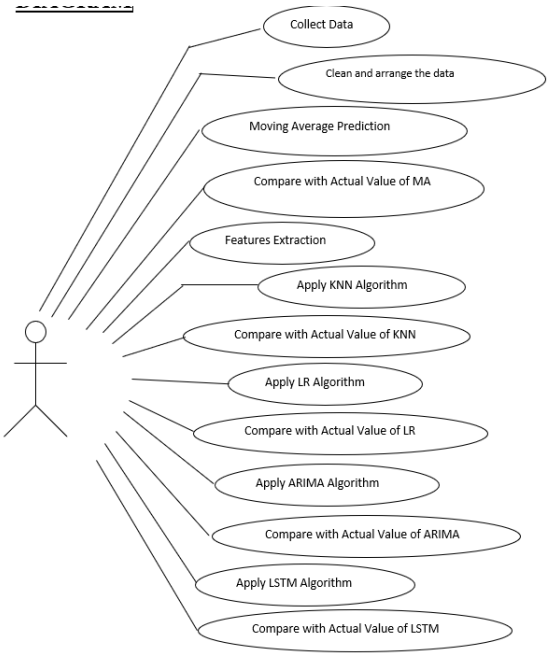

Fig. 3. Use Case Diagram

VIII. ALGORITHM COMPARISON WITH RESULT

LSTM is suitable for Time Series Analysis (Stock Prediction) for all time period to predict the stock price. Here RMSE value is nearly same for long-term and short-term prediction, so this algorithm can be used for any time period prediction.

ALGORITHM	TIME IN DAYS	RMSE	OBSERVATION
MOVING AVERAGE (Window Size 7)	**15**	**4.0762**	**SUITABLE FOR ALL TIME PERIOD PREDICTION**
	45	4.2988	
	90	5.7359	
MOVING AVERAGE (Window Size 90)	15	9.4363	NOT Suitable because RMSE Value is Large
	45	15.2000	
	90	19.4395	
LINEAR REGRESSION	**15**	**5.3079**	Suitable for Short- and Medium-Term Prediction
	45	6.4737	
	90	16.7328	
KNN	15	89.8367	NOT SUITABLE
	45	98.4237	
	90	105.8836	
ARIMA	**15**	**4.4175**	Suitable for Short Term Prediction
	45	14.7645	
	90	24.1911	
LSTM	15	6.2628	**SUITABLE FOR ALL TIME PERIOD PREDICTION**
	45	4.6568	
	90	6.5316	

Fig. 4. All model comparison Results

IX. Conclusion

Moving Average is suitable for Time Series Analysis (Stock Prediction) for all time period, if we take small window size (i.e. 6 previous value). Here RMSE value is nearly same for all time period prediction.

If we take long window size(90), it we will not suitable for prediction for any time period.

Linear Regression is suitable for Short Term and Mid Term Prediction. Not Suitable for Long term Prediction

KNN is not suitable for Stock Price Prediction because RMSE value is very high.

ARIMA is used for Time series Analysis but on this data set it is not performing well for Mid-Term and Long-Term Prediction but for Short Term it performs good.

LSTM is suitable for Time Series Analysis (Stock Prediction) for all time period to predict the stock price. Here RMSE value is nearly same for long-term and short-term prediction, so this algorithm can be used for any time period prediction.

References

[1] R. Wilson and R. Sharda, "Bankruptcy prediction using neural networks", Decision Support Systems, vol. 11, no. 5, pp. 545-557, 1994.

[2] Elliott wave theory and neuro-fuzzy systems, in stock market prediction: The WASP system. Expert Systems with Applications, 38, 9196–9206

[3] K. Tsai and J. Wang, "External technology sourcing and innovation performance in LMT sectors", Research Policy, vol. 38, no. 3, pp. 518-526, 2009.

[4] K. Han and J. Kim, "Genetic quantum algorithm and its application to combinatorial optimization problem", Evolutionary Computation, 2000, vol. 2, pp 1354-1360, 2000.

[5] "Market Capitalisation Definition from Financial Times Lexicon", Lexicon.ft.com,2016. [Online].

[6] "MarketTrak's Forecast Model Overview", Markettrak.com, 2016. [Online].