STOCK MARKET ANALYSIS AND PREDICTION

Boopathy K[1], Kanagaraj P[2], Prasanth G[3], Thirumal P C[4]

[1-3] Student, Department of Information Technology, Kumaraguru College Of Technology, Coimbatore, India
[4]Associate Professor, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India

---***---

Abstract - *Analysis and prediction of NSE stock market is one of the attractive topic to researchers from different fields. In particular, numerous studies have been conducted to predict the movement of stock market using machine learning algorithms such as support vector machine (SVM), Long short term memory (LSTM), Auto regressive integrated moving average (ARIMA), Random forest and Linear regression. In this project, we compared these algorithms based on the accuracy calculated using RMSE (root mean square error) for all the models, we predicted prices of industries.*

Key Words: **Data mining techniques, Machine Learning Algorithms, Prediction, Forecasting stocks.**

1. INTRODUCTION

Forecasting of NSE stock market is a way to predict future prices of stocks to find best time to buy and sell. It is one of the attractive topic for researcher and investors. The NSE Stock prices are dynamic day by day, so it is not possible to decide what is the best time to buy and sell stocks. Machine Learning provides a wide range of algorithms, which has been very effective in predicting the future stock prices.

In this project, we explored different Machine Learning algorithms to forecast and analyse stock market prices for NSE stock market. Our aim is to compare various algorithms and evaluate models by comparing prediction accuracy. We performed a few models including Linear regression, ARIMA, LSTM, Random Forest and Support Vector Regression. Based on the accuracy derived using RMSE of all the models, we predicted prices and ranged algorithms which can be used for stock prediction of different industries. For forecasting, we used historical data of NSE stock market and applied a few pre-processing methods to make prediction more accurate and relevant.

2. TOOLS & TECHNOLOGIES

Language and libraries: Python, SciPy, NumPy, Pandas, Sci-Kit Learn, Keras. Keras is required to implement LSTM model. the other libraries are required to process data and implement machine learning algorithms. Pandas made data pre-processing relatively easy.

Tool: GOOGLE colab is convenient to use and is very fast.

3. ALGORITHMS

3.1 SVM (Support Vector Machine for Regression):

SVM is considered as one of the most important breakthroughs in machine learning field and can be applied in classification and regression. In this project, SVR is considered to solve a regression problem as it avoids difficulties of using linear functions.

3.2 LSTM (Long Short-Term Memory):

It is a recurrent neural network (RNN) architecture that learns about values using intervals. LSTM keeps track of the past values and use those changes to predict future values. In our project we have stock values for each day which can be treated as sequence of values. for its ability to act as memory unit, LSTM can be treated as one of the best algorithms for time-series analysis problems.

Y is present value and X is past value by one day. LSTM will link between X and Y to predict future value.

x	Y
22	35
35	48
48	52

3.3 ARIMA (Auto Regressive Integrated Moving Average):

ARIMA model works appropriately for modelling time series with trend characteristics, random walk processes, and seasonal and non-seasonal time series . It has simple structure that enables to model our time series dataset characteristics properly.

3.4 Random Forest Algorithm:

Random tree is an ensemble of multiple trees. It has its own way of predicting values. We don’t know whether the data is linear or not. In such cases, Random forest is effective.

3.5 Linear Regression:

It is one of the most used algorithm for regression analysis. This algorithm is implemented to check how it works compared to other algorithms.

4. SYSTEM DESIGN AND WORKFLOW

The input data is pre-processed by cleaning of data and splitting them into proper sets of training and test. This is in turn is fed to the learning algorithms for the main phase of analysis. Based on the output from the algorithm, the values are predicted and new data has been generated. Generated data are been used for the evaluation of the predicted values to find the accuracy of the algorithms efficiency.

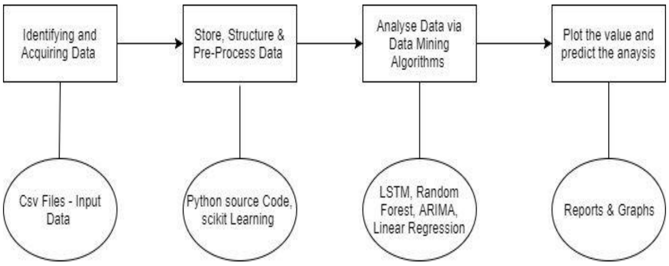

Identifying and Acquiring Data
Store, Structure & Pre-Process Data
Analyse Data via Data Mining Algorithms
Plot the value and predict the anaysis
Csv Files - Input Data
Python source Code, scikit Learning
LSTM, Random Forest, ARIMA, Linear Regression
Reports & Graphs

Fig-1: system design

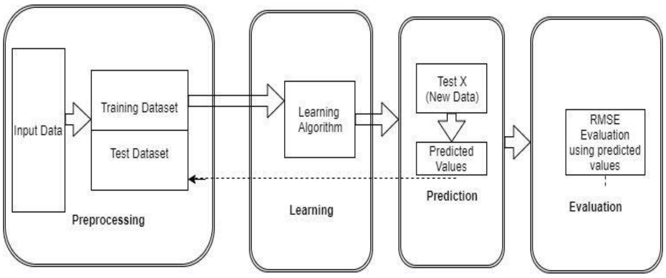


Fig-2: workflow

5. EXPERIMENTS / PROOF OF CONCEPT EVALUATION

5.1 DATASET

In the project, we chose the National Stock exchange collected from. This dataset includes India stocks and our index covers a diverse set of sectors featuring many Indian companies. Our aim was to focus on making general and unbiased model, which works on every type of scenario irrespective of company or financial sector. It helps to validate our predictive algorithm and provide more accurate stock prediction.

Our dataset includes eight features such as company Index, Date, Time, Open, Close, High, Low values and Volume of trading (prices are in INR). The dataset covers 440 companies every minute since 2015. We took this dataset as it's size is quite large (~2gb) and it can be used to evaluate several companies using our algorithm. With the primary dataset prepared, we applied pre-processing methods to carry out individual experiments.

Data pre-processing: It is a data mining technique that involves transforming raw data into an understandable format. Our dataset has some limitations such as it contains invalid values, null values and missing records etc. We applied following techniques to pre-process our data to make accurate prediction.

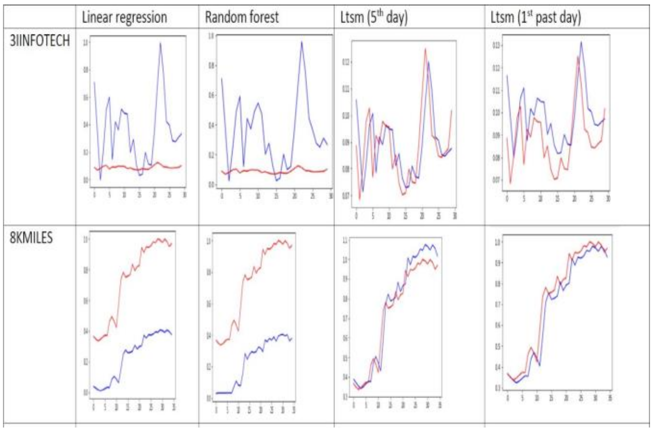
1. **Data cleaning:**
   The purpose of data cleaning is to fill in missing values and correct inconsistencies in the data. Index, Date, time closing prices of NSE dataset are used as input. There were some missing values due to public holidays. We removed null values and invalid indexes. There were few irrelevant columns in the dataset which were not used as input. So we eliminated those columns to reduce the complexity of our prediction model.
2. **Data Transformation:**
   As our dataset contains minute-wise stock prices and we needed daily basis prices to fit in our model, so we grouped the data on daily basis prices and took mean of all the rows. Also we applied min-max scaling for a few algorithms to get more accurate prediction.

5.2 Methodologies:

In this project, we have made a time-series analysis and it doesn't need n-fold cross validation methodology since it's sequential data. We split our dataset in train and test data. Top 80 percent of data will be Train data and the remaining will be test.

6. COMPARISON

Below are the few companies with graphs plotted for predicted values(Red) versus actual values (Blue) for different algorithms. We can see that LSTM and Arima performs better compared to random forest and linear regression.



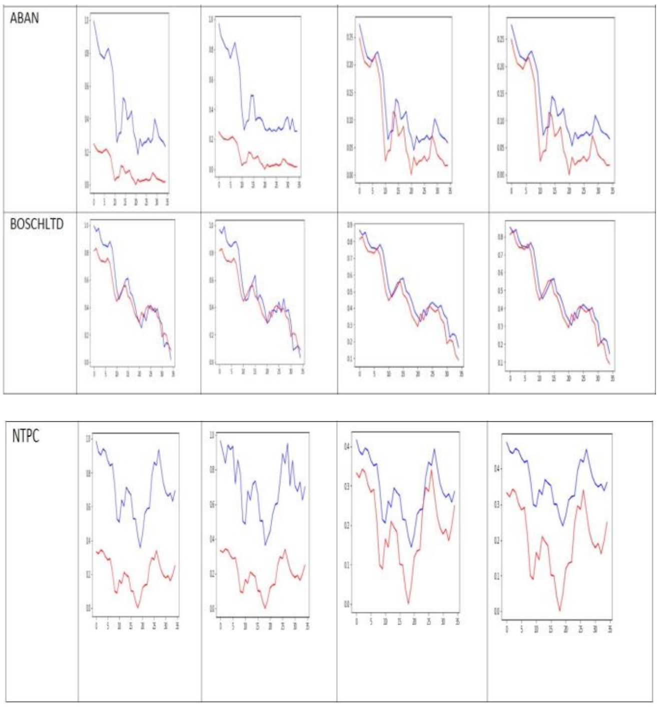


Fig-3: Actual closing price index and its predicted value from LR, RF, LSTM models

6.1 Arima model (monthly basis):

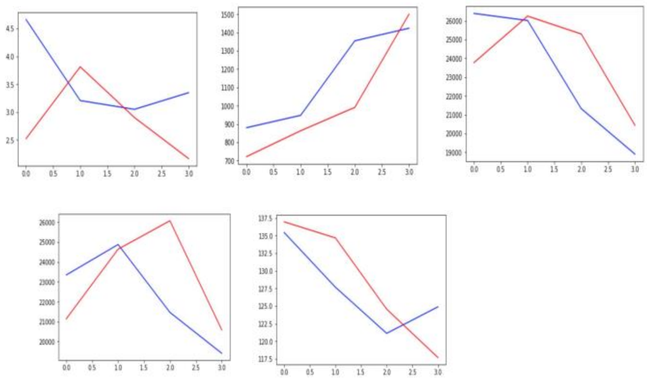


Fig -4: Graph Comparison for five companies (Left to right) Infotech, 8kmiles, Aban, Bosch Ltd, NTPC

7. EVALUATION

The accuracy of prediction is referred to as "goodness of fit". In this project, most popular and statistical accuracy measure RMSE is used for comparison of different algorithms on same dataset, which is defined as:

$$\mathrm{RMSE}_{fo} = [\sum_{i=1}^{N} (z_{f_i} - z_{o_i})^2/N]^{1/2}$$

Below is the table of evaluation of all used algorithms:

Table -1: RMSE comparison for three companies

Model	3IINFOTECH	8KMILES	ABAN
Linear Regression	0.334	0.502	0.415
Support Vector Regression	NA	NA	NA
LSTM (relating present and past fifth day)	0.011	0.064	0.043
LSTM (relating present and past day)	0.013	0.054	0.039
Arima	1.263	206.344	23.707
Random Forest	0.345	0.502	0.402

8. DECISION MADE

We decided to make analysis using close values of the stock on a particular day or month and predict the closing values for future. Decided to list these algorithms from top 1 to top 5 based on the performance of all the models.

9. CONCLUSION

In the project, we proposed the use of different algorithms to predict the future stock prices of almost twenty companies. Although comparison is shown for only five companies (randomly selected) in the report due to space constraint, the behaviour can be known for any company by using the same code. Long short term memory algorithm worked best in case of forecasting and also we ranged from first to last algorithms for forecasting stock market,

- LSTM
- ARIMA
- SVR
- RF & LR

In future, we will extend the project for other effective methods that might result a better performance. Our algorithms can be used to maximize profit of investors but it has to be improved for real time conditions.

10. REFERENCES

1) http://markdunne.github.io/public/mark-dunne-stock-market-prediction.pdf

2) http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.278.6139&rep=rep1&type=pdf

3) https://www.kaggle.com/ramamet4/nse-company-stocks

4) https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

5) https://ec.europa.eu/eurostat/sa-elearning/arima-model