**Stock Market Prediction using Hybrid Machine Learning System**

**A PROJECT REPORT**

*Submitted by*

**SHRUSHTI GOVARDHAN [Reg No: RA1611008010099]**

**SHREYAS KULKARNI [Reg No: RA1611008010186]**

**PRATIK DHARAMDASANI [Reg No: RA1611008010298]**

**ASTHA MAHESH KUMAR [Reg No: RA1611008010752]**

*Under the Guidance of*

**Mr. S Deepan**

(Assistant Professor, Department of Information Technology)

*In partial fulfillment of the Requirements for the Degree*

*of*

**BACHELOR OF TECHNOLOGY**



**DEPARTMENT OF INFORMATION TECHNOLOGY**

**FACULTY OF ENGINEERING AND TECHNOLOGY**

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY**

**KATTANKULATHUR – 603203**

**MAY 2020**

**CHAPTER 2**

**LITERATURE STUDY**

**2.1 Detailed Literature Review**

It has been observed from the survey of the that since the rise of application of machine learning techniques in various fields, various attempts of predicting the stock market have been carried out. It is interesting to note that although the prediction accuracy obtained throughout the works have been fairly moderate, the results from machine learning have mostly been proved to be much more accurate than the contemporary prediction techniques.

In [2], the researchers created a dataset of Apple Inc. stock prices using various attributes of the market like Date, Open, High, Low, Close and Volume of the stock at the time of observation. The aim of the study was to compare the working of Random Forest classifier and Multilayer Perceptron. From the results of [2], random forest performed better than the neural network method, contrary to the commonly held belief. The comparison of performance was done based on correlation coefficient, RMS and Relative Absolute Error. It could be inferred from [2] that neural network is not always the best method for stock market prediction.

A comparison is made between the performance of several machine learning algorithms on Bombay Stock Exchange data in [3]. The value of BSE is predicted based on historical market features such as commodity prices and forex. Out of the four algorithms, AdaBoost outperformed all other methods while Random Forest reported the least accuracy. The dataset used in this study was also collected over a period of time.

It has been mentioned in [4] that time has a major role to play in the stock prices. This hypothesis is further tested using Yahoo Finance Dataset by selecting Opening and Closing price of the stock on the particular day and taking the Highest and Lowest price into consideration. The total volume of stock transaction is also taken as a feature. This study compares the performance of deep learning Long Short Term Memory (LSTM) architecture with a machine learning Regression model by taking R-Square value as the judging parameter. We infer two things from the results obtained. Firstly, the stock data is indeed dependent on time since LSTM model considers time-series values and outperforms regression model. Secondly, linear regression model is of less use because the stock data is non-linear in nature, as stated in [1].

In [5], a comparison is made between single layer perceptron, multi layer perceptron, radial basis function and SVM model. The dataset used for the study was collected over a period of time including relevant features like oil, gold, silver prices, forex, interest rates from banks and news articles regarding the stock prices. The news data was labeled as positive or negative depending on how it portrayed the future movement of the market. The aim of the study was predict the movement of the market as positive (increase in value) or negative (decrease in value). While all the models used for comparison gave similar results, multilayer perceptron reported the best accuracy while single layer perceptron and SVM gave similar accuracy. The performance of Radial basis function was in between these two. It can be inferred from this study that due to the complexity of the problem domain, simpler models such as SVM and single layer perceptron are bound to give limited results and hence models of higher complexity should be used for obtaining better results.

The performance of deep neural networks is compared in [1] on SPDR S&P 500 ETF dataset. Architectures with different number of hidden layers is trained and accuracy of each network is recorded on the training set. It is observed that with an increase in number of hidden layers from 12 to 35, there was a significant increase in the accuracy of the model. However, the accuracy of the model became stopped increasing and became almost constant as number of hidden layers became larger than 30. From this study we infer that there is a limit to the accuracy that can be obtained from deep neural networks and that use of highly complex models do not guarantee better results.

From the findings in the literature review, we conclude that in order to get accurate predictions of stock market, a complex function should be used. In addition to that, the function should be non linear in nature as the problem belongs to a non linear domain, as proved by [1].

**2.2 Project Modules and Time line**

The proposed system makes use of machine learning for predicting the price of a stock. The system consists of the following modules:

1. Data Preprocessing

* Removal or replacement of missing values
* Co-relation calculation and Removal of correlated data
* Visualization of correlation heatmap
* Split dataset into training and testing sets

2. Normalization of data

Normalizing refers to the transformation of data in a specific range. All the values in the dataset is normalized in the range of 0 to 1. This step helps in increasing computation speed as all the data is converted to same scale. Additionally, normalization is necessary before calculation of Principal Component Analysis, which is done in the next step.

3. PCA calculation

Principal Component Analysis is a method of calculating Eigen vectors and Eigen values from the data. This calculation helps in understanding the fundamental properties that the data follows. From the existing research it has been observed that algorithms give significantly higher accuracy on PCA transformed dataset as compared to the non-transformed data. It also helps in dimensionality reduction.

4. Training the regression model

Stock market data is highly non-linear in nature hence fitting linear models on this data would prove to be of less use. Therefore, Lasso and Ridge Regression are chosen for this task. These models would be trained on the same transformed dataset. The dataset is split for training and testing in order to obtain correct accuracy scores.

5. Testing model performance

The trained model is tested on test dataset and it is evaluated on the basis of accuracy achieved. The performance is reported and model is tuned for getting better results.

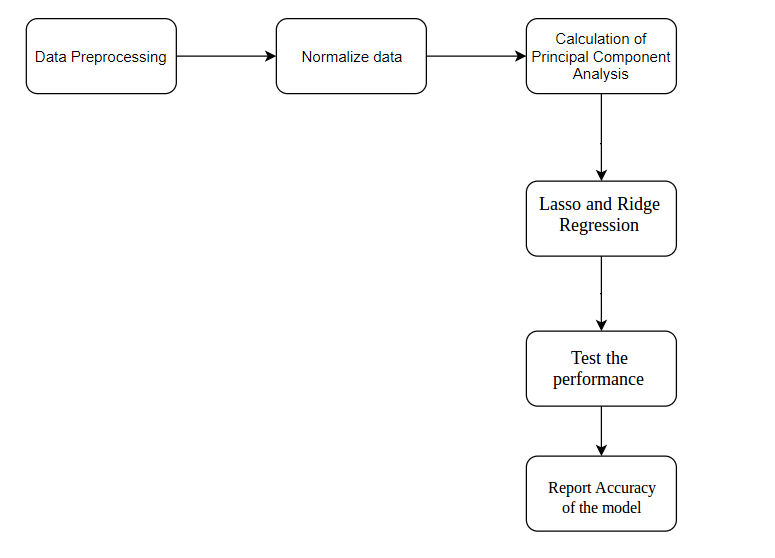
**2.3 Detailed Methodology of Implementation**

**System Architecture**

We propose the a method for stock market prediction which makes use of non-linear machine learning algorithms and a deep neural network architecture achieving accurate results. The results obtained would be compared on the terms of accuracy of price prediction and the best model would be chosen.

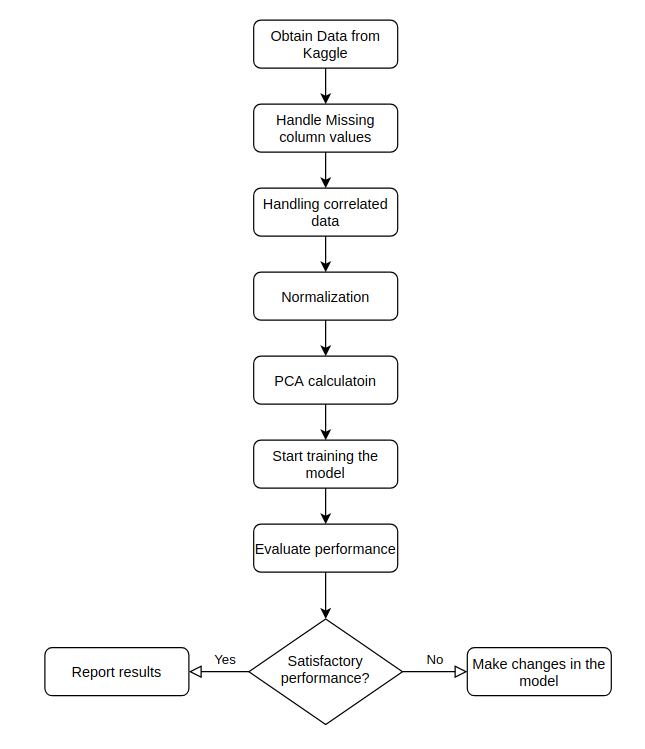
The method makes use of preprocessing concepts for handling the missing values and organizing the data. The data is normalized and PCA is calculated for it. It is done for extracting the essential information that is responsible for defining the data pattern. Three non-linear models are trained and tested on different parts of dataset for obtaining reliable results. The accuracy is compared and best model is reported and selected.

The diagram below represents the flow of the proposed system.

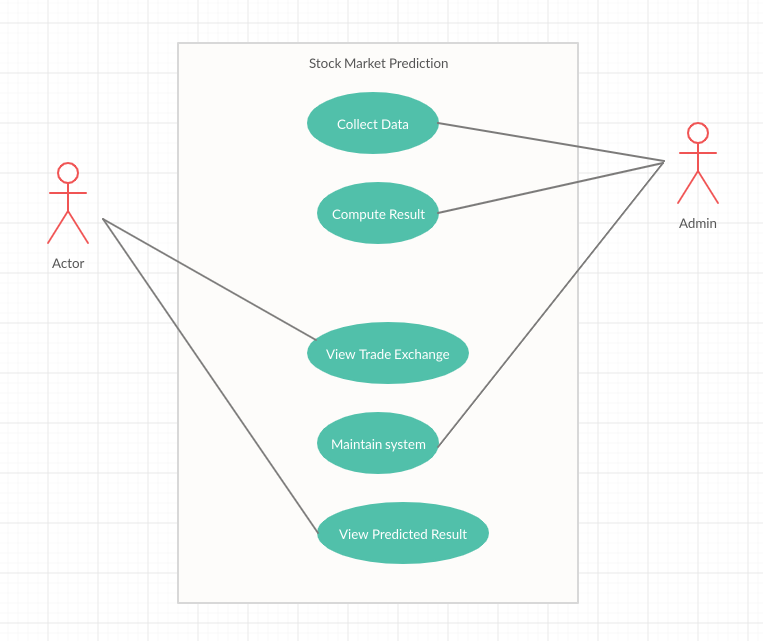


**Description of System Architecture**

**Flowchart**

****

**Use Case**

****

**2.4 Work Distribution**

|  |  |
| --- | --- |
| SHRUSHTI GOVARDHAN  [Reg No: RA1611008010099] |  |
| SHREYAS KULKARNI  [Reg No: RA1611008010186] |  |
| PRATIK DHARAMDASANI  [Reg No: RA1611008010298] |  |
| ASTHA MAHESH KUMAR  [Reg No: RA1611008010752] |  |

**Work list:**

**References**

[1] A. Nayak, M. Pai, R. Pai, Prediction models for Indian stock market. Proced. Comput. Sci. (2016)

[2] Z. Hu, J. Zhu, K. Tse, Stocks market prediction using support vector machine, in *2013 6th International Conference on InformationManagement, InnovationManagement and Industrial Engineering* (2013)

[3] WeiHuang,YoshiteruNakamori, Shou-YangWang, Forecasting stock market movement direction with support vector machine. Comput. Oper. Res. (2005)

[4] C. Cortes, V. Vapnik, Support vector networks. Mach. Learn. (1995)

[5] J. Ali, R. Khan, N. Ahmad, I. Maqsood, Random forests and decision trees. Int. J. Comput. Sci. Issues (2012)

[6] S. Yutong, H. Zhao, Stock selection model based on advanced AdaBoost algorithm, in *2015 7th International Conference on Modelling, Identification and Control (ICMIC)* (2015)

[7] P. Wu, H. Zhao, Some analysis and research of the AdaBoost algorithm, in *Communications in Computer and Information Science* (2011)