DEPARTMENT OF COMPUTER ENGINEERING & APPLICATIONS
INSTITUTE OF ENGINEERING & TECHNOLOGY

B.Tech IV Year CSE

Project Report

On

"Comparative Analysis of various Machine Learning Algorithms in Stock Market"

Under the supervision of:

Mr. Maneet Singh

Submitted by:

1. Vaibhav Rai (B-67/151500593)
2. Vatsal Tewari (B-69/151500599)
3. Yash Agarwal (D-75/151500636)
4. Vivek Mishra (C-84/151500628)

Group No: G-10

Odd Semester: 2018-19

1. **Introduction**
   1.1 Overview & Motivation
   1.2 Objective
   1.3 Scope

2. **Software Requirement Analysis**
   2.1 Feasibility study: To discuss the feasibility of the system to be developed.
   2.2 Requirement analysis: Requirements of the software to be developed is to be established
   2.3 System modules: To determine system modules with the mapping of the requirements in each of the modules.
   2.4 Use case diagram
   2.4 SRS

3. **Design**
   3.1 DFD
   3.2 ER Diagram

4. **Additional Literature Survey**
   4.1 Summary of relevant papers
   4.2 Summary

Chapter 1
Introduction

1.1 Overview and Motivation

A stock market or share market is the aggregation of buyers and sellers (a loose network of economic transactions, not a physical facility or discrete entity) of stocks (also called shares), which represent ownership claims on businesses; these may include securities listed on a public stock exchange, as well as stock that is only traded privately. Examples of the latter include shares of private companies which are sold to investors through equity crowd funding platforms. Stock exchanges list shares of common equity as well as other security types, e.g. corporate bonds and convertible bonds.

Trade in stock markets means the transfer (in exchange for money) of a stock or security from a seller to a buyer. This requires these two parties to agree on a price. Equities (stocks or shares) confer an ownership interest in a particular company.

Participants in the stock market range from small individual stock investors to larger investors, who can be based anywhere in the world, and may include banks, insurance companies, pension and funds. Their buy or sell orders may be executed on their behalf by a stock exchange trader.

In order to trade in stock market there are several brokers which tend to invest the client's money to regain profit and for their return they take a percentage of fee from the profit.

Machine Learning is a vast field and has provided solutions to many. It even lays feet in the world of stock market as many brokers use Machine Learning algorithms to predict the value of the stock client wants to invest in. The project tends to show which of the algorithms or their hybrids are best for stock predictions.

The project covers algorithms such as Linear Regression, Decision Trees, Random Forest Regression, Artificial Neural Networks and Recurrent Neural Networks. There are various research papers for stock predictions using Artificial Neural Networks and Recurrent Neural Networks, but the classical algorithms are not used, so it will reflect the results of those algorithms as well.

1.2 Objective

The purpose of the project is to analyze the machine learning algorithms on the large dataset of different stock prices and their categories in training set part of dataset to predict the prices of stock, so that we can decrease human effort in deciding the prices of stock and come up with better guidance in order to help people to decide which stock to invest in so as to maximize the return and also to know which algorithm or hybrid of various algorithms gives a better prediction result.

1.3 Scope

This project consists of Analytics Module that provides predictions, and real statistical analysis. It can be used by people who want to invest in stock market without help of broker and their advice.

In the end it will try to uncover which algorithm is best for which part of stock predictions and its individual benefits.

Chapter 2
Software Requirement Analysis

2.1 Feasibility Study

Stock values cannot be accurately predicted. They are like complex problems, far too many variables to be predicted. The stock market is a place where buyers and sellers converge. When there are more buyers than sellers, the price increases and vice versa. It is very difficult to try to predict where markets is going, thus making it unpredictable.

The proposed will not always produce accurate results since it does not account for human behaviors as there are many algorithms and the project tends to produce results to show which gives more accurate result.

The main objective of the project is to show which algorithm gives better result than the other. It does not depend it will be for long term. There are too many variables to decide for the long term values.

2.2 Requirement Analysis

2.2.1 External Interface Requirements

2.2.1.1 User Interfaces

The interface will meet the following requirements to conform to the users' needs. It will be simple and easy to understand. The interface will include user inputs as well as results, outlined below.

Input: It will consist of dataset which shows certain stock values and its required columns which will help in the prediction.

Output: This will consist of the results represented in form of tables and graphs. The results will be displayed in a clear and meaningful manner that allows the user to easily interpret the result over all analysis sessions.

2.2.1.2 Hardware Interfaces

The solution makes extensive use of several hardware devices. These devices include:

- Python/R Tool Server's Data Warehouse.
- Windows and Linux users' computers.

2.2.1.3 Software Interfaces

The software requirements are to support windows operating system with support to Python frameworks.

2.2.1.4 Communications Interfaces

Python tool is required in order to make use of several functions and to be executed such as understanding, viewing and analyzing.

2.2.2 Functional Requirements

2.2.2.1 Retrieving Input

The software will receive input data (Stock details). Data will be entered by the user for analysis and prediction purpose.

2.2.2.2 Processing

The software will take input, process data, and display output. This will enforce that the snapshot provided by the simple gauge is a current view of the situation.

2.2.2.3 Prediction

This will be done on the processed data to find out the possible price of a certain stock so to remove human error.

2.2.2.4 Output

The results of analysis and predictions will be outputted in the form of simple graphs and tables. In addition, the software may output statistics pertaining to a topic. This output should be clear and easy to understand.

2.2.3 Non-functional Requirements

2.2.3.1 Reliability

The project will meet all of the functional requirements without any unexpected behavior. At no time should the gauge output display incorrect or outdated information without alerting the user to potential errors

2.2.3.2 Availability

The functionality of the project will depend on any external services such as availability of data and Python tool that are required. If those services are unavailable, the user should be alerted.

2.2.3.3 Security

The tool should never disclose any personal information of the buyers, and should collect no personal information from its own users.

2.2.3.4 Maintainability

The software should be written clearly and concisely. The code will be well documented. Particular care will be taken to design the software modularly to ensure that maintenance is easy.

2.2.3.5 Portability

This software will be designed to run on any windows and Linux system.

2.3 System Modules

2.3.1 Data Source

This will consist of data from various sources. Data will be either from yahoo finance which provides real time data but with payment or from QuandL which is a package in python.

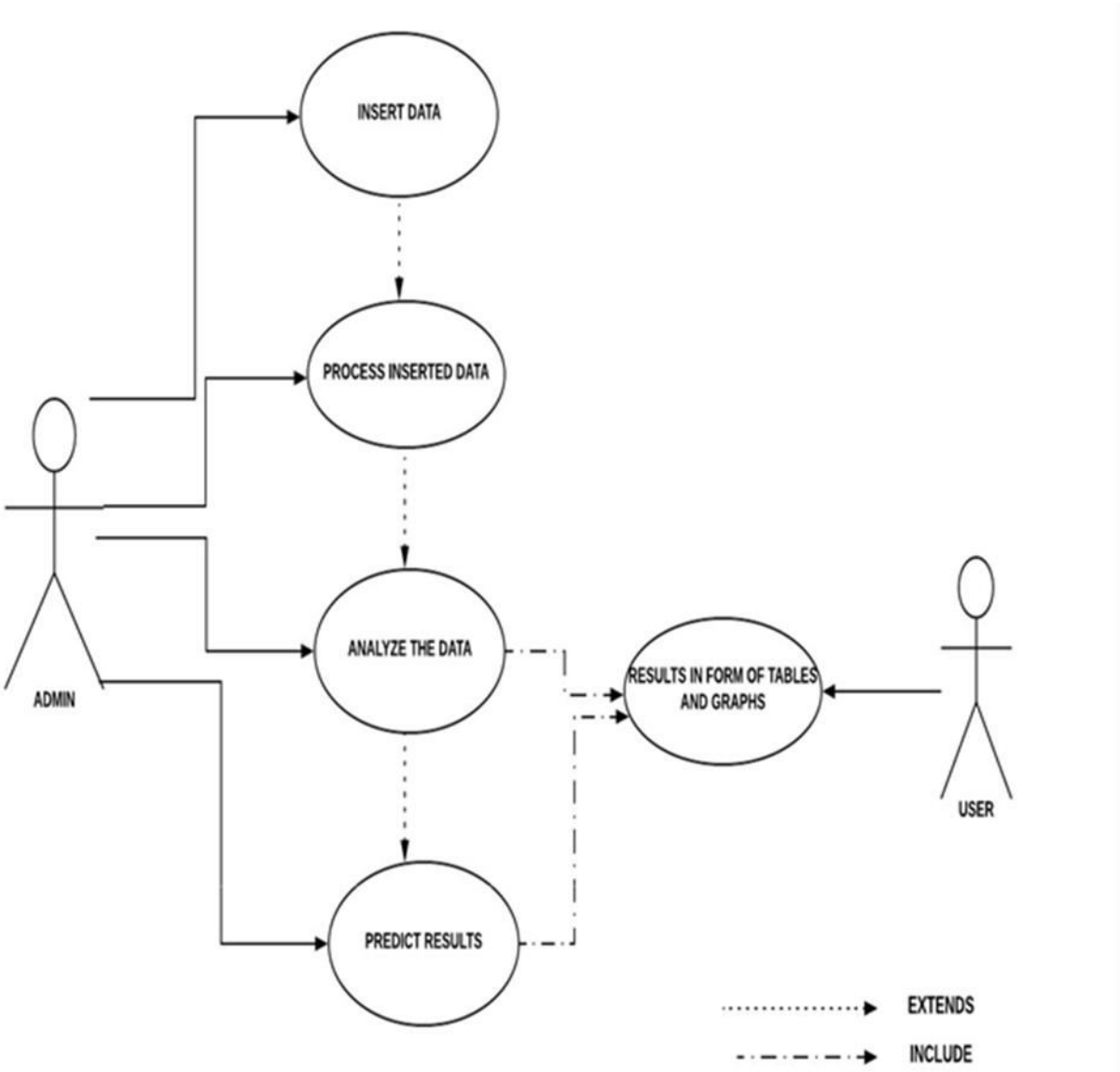
2.3.2 Stock Database

Data coming from above mentioned data sources will be stored in csv(Comma Separated Values) files from where it can be directly fetched and used.

2.3.3 Python/R Tool

Now this data will be fed into various machine learning algorithms which will be made by using either python or r, whichever fits the purpose, as each one has individual features.

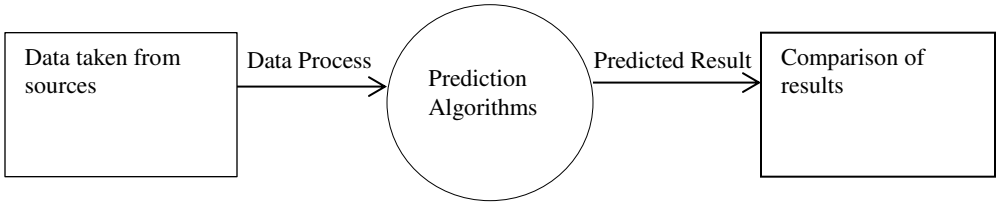
2.4 Use Case Diagram



Chapter 3
Design

3.1 Data Flow Diagram (DFD)

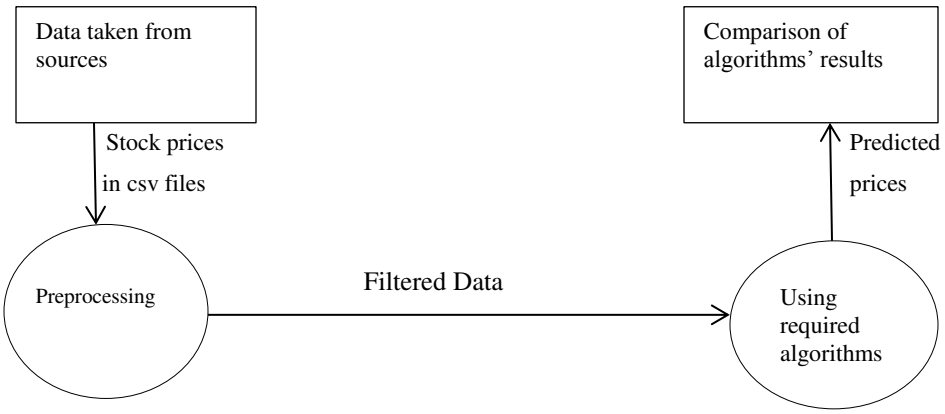
3.1.1 DFD Level – 0



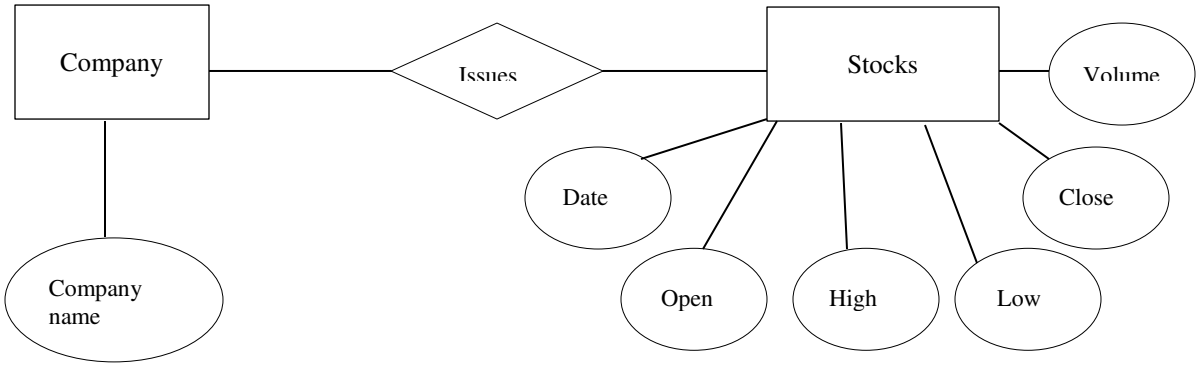
Data is collected from either Yahoo Finance or QuandL package. Further, the data is preprocessed for modeling. When the anomalies are removed, the required prediction algorithm will be modeled. The results will be compared to determine which algorithm is better for prediction.

3.2.2 DFD Level – 1

Data will be collected from the required sources in csv (Comma Separated Values) format. Data is sent for preprocessing to remove the anomalies. The filtered data will be sent for the prediction of prices. Results of various algorithms will be compared to determine which algorithm is better.



3.2 E-R Diagram



The Company's name will be reflected as the file the file name.

The attributes of the stock are as follows –

- Date – Date at which stock values are provided.
- Open – Opening value of the stock.
- High – Highest value of the stock throughout that day.
- Low – Lowest value of the stock throughout that day.
- Close – Ending value of the stock.
- Volume – Indicator provided by the stock market to give an idea about the next day's value.

Chapter 4
Additional Literature Survey

4.1 Summary of Relevant Papers

4.1.1 Paper – 1

Title – Stock market prediction using Linear Regression

Publish Date – 18 December 2017

Summary - It is a serious challenge for investors and corporate stockholders to forecast the daily behavior of stock market which helps them to invest with more confidence by taking risks and fluctuations into consideration. In this paper, by applying linear regression for forecasting behavior of TCS data set, we prove that our proposed method is best to compare the other regression technique method and the stockholders can invest confidentially based on that using Linear Regression.

URL - https://ieeexplore.ieee.org/abstract/document/8212716

4.1.2 Paper – 2

Title **-** Predicting customer retention and profitability by using random forests and regression forests techniques.

Publish Date – August 2005

Summary - In an era of strong customer relationship management (CRM) emphasis, firms strive to build valuable relationships with their existing customer base. In this study, we attempt to better understand three important measures of customer outcome: next buy, partial-defection and customers' profitability evolution. By means of random forests techniques we investigate a broad set of explanatory variables, including past customer behavior, observed customer heterogeneity and some typical variables related to intermediaries. We analyze a real-life sample of 100,000 customers taken from the data warehouse of a large European financial services company. Two types of random forests techniques are employed to analyze the data: random forests are used for binary classification, whereas regression forests are applied for the models with linear dependent variables. Our research findings demonstrate that both random forests techniques provide better fit for the estimation and validation sample compared to ordinary linear regression and logistic regression models. Furthermore, we find evidence that the same set of

variables have a different impact on buying versus defection versus profitability behavior. Our findings suggest that past customer behavior is more important to generate repeat purchasing and favorable profitability evolutions, while the intermediary's role has a greater impact on the customers' defection proneness. Finally, our results demonstrate the benefits of analyzing different customer outcome variables simultaneously, since an extended investigation of the next buy–partial-defection–customer profitability triad indicates that one cannot fully understand a particular outcome without understanding the other related behavioral outcome variables.
URL - https://www.sciencedirect.com/science/article/pii/S0957417405000965

4.1.3 Paper – 3

Title - Stock market prediction system with modular neural networks.
Publish Date – 21 June 1990
Summary - A discussion is presented of a buying- and selling-time prediction system for stocks on the Tokyo Stock Exchange and the analysis of internal representation. The system is based on modular neural networks. The authors developed a number of learning algorithms and prediction methods for the TOPIX (Tokyo Stock Exchange Prices Indexes) prediction system. The prediction system achieved accurate predictions, and the simulation on stocks trading showed an excellent profit.
URL - https://ieeexplore.ieee.org/abstract/document/5726498

4.1.4 Paper – 4

Title – Stock market prediction using different Neural Network Classification Architectures
Publish Date – January 1996
Summary - In recent years, many attempts have been made to predict the behavior of bonds, currencies, stocks, or stock markets. The Standard and Poors 500 Index is modeled using different neural network classification architectures. Most previous experiments used multilayer perceptron for stock market forecasting. A Multilayer perceptron architecture and a probabilistic neural network are used to predict the incline, decline, or steadiness of the index. The results of trading with the advice given by the network are then compared with the maximum possible performance and the performance of the index. Results show that both networks can be trained to

perform better than the index, with the probabilistic neural network performing slightly better than the multi-layer perceptron.
URL - http://scholarsmine.mst.edu/engman_syseng_facwork/225/

4.2 Summary

After various papers, it is difficult to decide which algorithm is better for stock prediction. There are algorithms such as RNN, ARIMA and HMM which are really good for stock prediction, but there are several algorithms untested such as decision trees, KNN or combination of several models which might provide good results. The parameters of the algorithms will be altered in order to get better results.