**ABSTRACT**

The study presented below is an experimental work in order to comprehend how machine learning algorithms can predict stock prices based on different input features and parameters with different accuracies. We have adopted two different paths, one predicting stock close price using historical data, and second predicting the direction of stock price high/low (1 or 0) using daily news headline data. Broadly speaking, we are performing historical data analysis along with sentimental analysis and matching them to prove our theory that the price of a stock is highly correlated to its historical performance and its general sentiment in the news. For historical data analysis, we are focusing on Linear Regression Algorithm and LSTM. For Sentiment analysis and classification, we are focusing on Logistic Regression and Random Forest Algorithms. The study presented below includes the survey of related work done by researchers in the past, problem definition, scope of work, data and dataset description, ML models used, implementation of our models, and result comparisons.

**INTRODUCTION**

The stock market refers to public markets that exist for issuing, buying, and selling stocks that trade on a stock exchange or over-the-counter. Stocks, also known as equities, represent fractional ownership in a company, and the stock market is a place where investors can buy and sell ownership of such investible assets. An efficiently functioning stock market is considered critical to economic development, as it gives companies the ability to quickly access capital from the public.

**RELATED RESEARCH**

Several researches have gone into this field. This section explains the analysis of the stock market prediction based on publication year, adapted methods, datasets used, evaluation parameters, software tools, and accuracy.

1. **Different stock market prediction techniques**

The stock market prediction techniques are broadly categorized into two types

**Prediction Techniques**

* Artificial Neural Network (ANN)
* Convolutional Neural Network (CNN)
* Decision Support System (DSS)
* Naive Bayes (NB)
* Neural Network (NN)
* Recurrent Neural Network (RNN)
* Support Vector Machine (SVM)
* Hidden Markov Model (HMM)

**Clustering Techniques**

* Filtering
* Fuzzy Based Method
* k-means
* Optimization

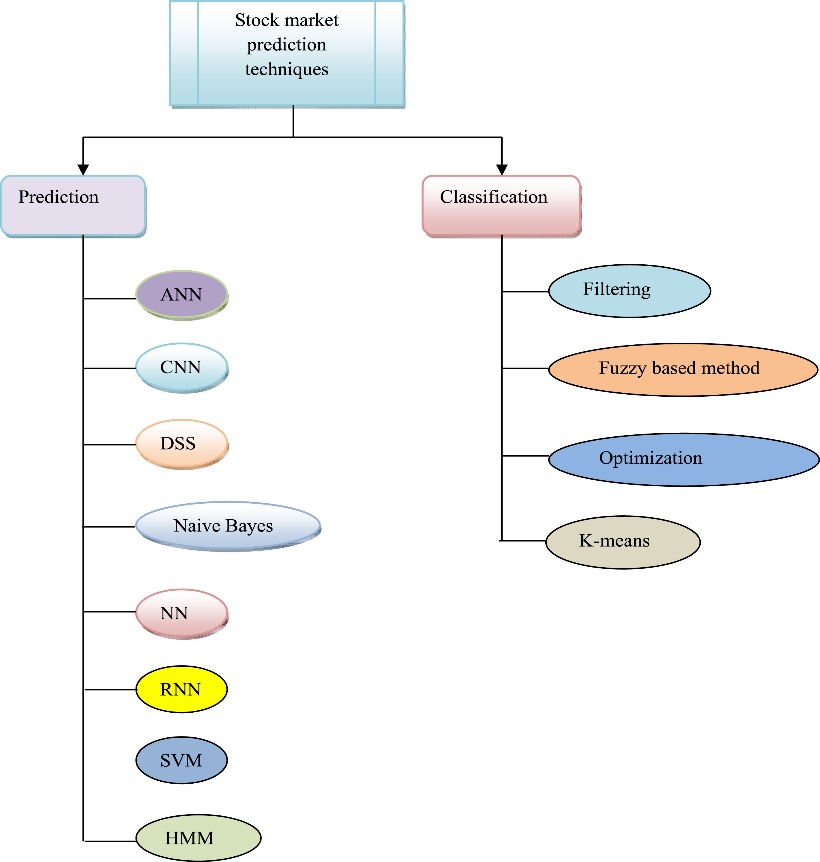
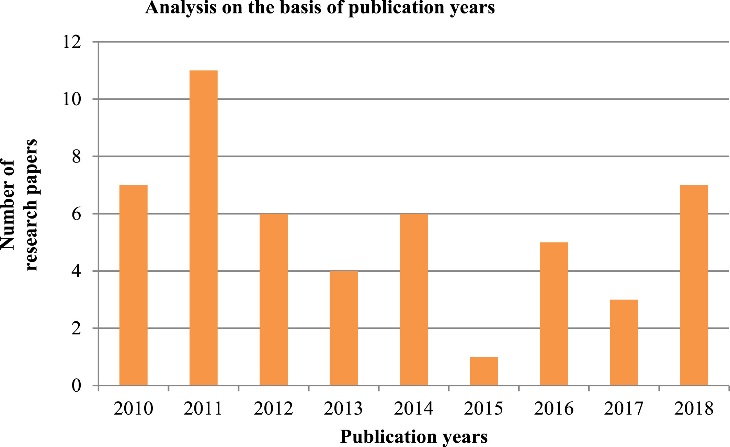


Fig 1 - Stock Market Prediction Techniques

1. **Related Work Analysis on the basic of publication Years**

* Figure 2 shows the number of research papers published in the years from 2010 to 2018.
* For this we took 50 papers from survey and then draw graph based on the publication year.
* For example 11 research papers were published in the year 2011. In 2018, seven research works were developed for stock market prediction.



# Fig 2 - Work Analysis on the basic of publication Years

# Analysis of related work based on prediction techniques used

* From Figure 3 we can see that 29% time ANN used for stock prediction 11% of the research papers used SVM, and 9% of the works are based on the SVR.
* The RNN is employed in 9% of the researches, decision support system is employed in 3% of the works and so on.
* So we can say that, NN and ANN are mostly employed techniques for stock market prediction.

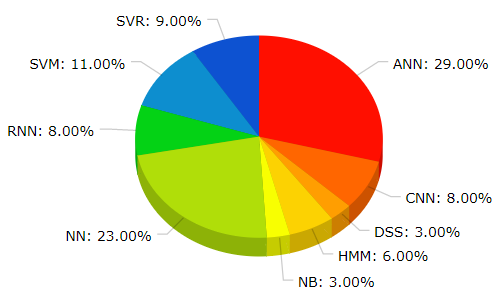


Fig 3 - Analysis of prediction techniques used.

1. **Analysis based on clustering techniques**

* Fig 4 shows the analysis of Clustering techniques used in previous papers.
* Here, 56% of the research papers are based on Fuzzy based system, 19% of papers used Filtering, 19% of papers utilized optimization techniques, and 6% of papers adapted the K-means clustering.
* So we can say that, the fuzzy-based technique was utilized in more research papers for stock market prediction.

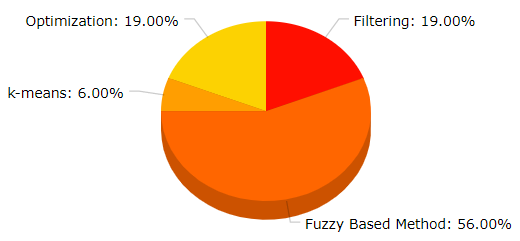


Fig 4 - Analysis of clustering techniques used.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Authors | Objective | Approach | Methodology | Findings / Limitation |
| Bing in [1] | Analysis of public sentiment for predicting stock movement. | Twitter data for sentiment analysis and then predict the stock movement of specific listed companies | SMeDA-SA | Using current public sentiment to predict real stock market price with a period of 3 days (i.e., T+3 days) has the best prediction accuracy. / Timeline gap in our SMeDA-SA |
| Hamid Beigy [3] | Develop a model using mood information in social media that could be used to predict stock market fluctuations (up or down) in the next day. | Use extracted features in one or two consecutive previous days and train a model in a supervised manner to predict the next day’s prices. | LDA-POS method | The LDA-POS method outperforms the presented Neural Network model by 4.62%./ limitations come from the regulators of the market. |
| Saman Haratizadeh[4] | capture the possible correlations among different variables for extracting combined features from a diverse set of input data from five major U.S. stock | Aggregating several variables in a CNN-based framework for feature extraction and market prediction. | **C**NN Method | The predictions of CNNpred as a base for the trading strategy of a trading system leads to good results in terms of Sharpe ratio and CEQ return measures./ designing the structures of CNNs for the stock prediction |
| M. M. Manohara Pai[6] | Prediction of stock market trend for that Two models are built one for daily prediction and the other one is for monthly prediction. | Sentiment of the company has been computed by using twitter data and news of the company. Outcome of sentiment analysis is considered along with open price, close price of stock with extracted statistical parameters to build model. | Supervised machine learning algorithms | Up to 70% of accuracy is observed using supervised machine learning algorithms on daily prediction model. |

1. **Research Papers Referred**

**PROBLEM DEFINITION**

Forecasting and predicting the trends of market is the most important applications of stock market. It also uncovers the future market behavior which always helps the investors to understand when and what stocks can be purchased for the growth of their investment.

All Stock prediction is a challenge in the field of finance as well as engineering and mathematics. Due to its financial gain, it has attracted much attention both from academic side and business side. Stock price prediction has always been a subject of interest for most investors and financial analysts.

Nevertheless, predicting how the stock market will perform is one of the most difficult things to do. Finding the best time to buy or sell has remained a tough nut for investors because there are numerous factors that may influence stock price such as news, social media data, fundamentals and production of the company, government bonds, historical price and country’s economics.

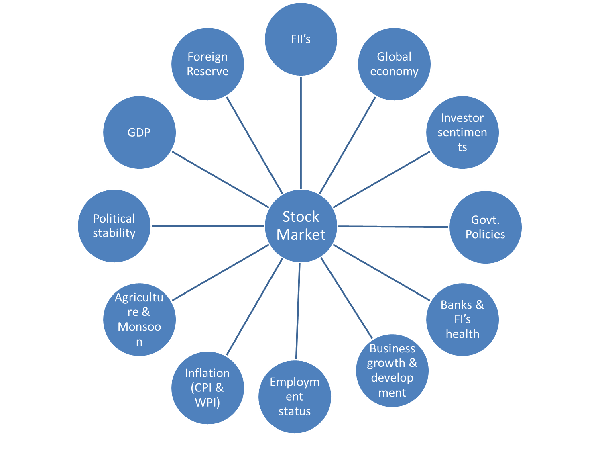
****

Fig 5 - Factors Affecting Stock Market

All these aspects combine to make share prices volatile and very difficult to predict with a high degree of accuracy.

**NEED AND SCOPE OF WORK**

Broadly, stock market analysis is divided into two parts – Fundamental Analysis and Technical Analysis.

* Fundamental Analysis involves analyzing the company’s future profitability on the basis of its current business environment and financial performance.
* Technical Analysis, on the other hand, includes reading the charts and using statistical figures to identify the trends in the stock market.

As you might have guessed, our focus will be on the technical analysis part. The study is primarily based on two components: sentiment analysis and historical data analysis.

* **Historical Data Analysis Component**

It would predict a function based on daily stock prices trained on previous years data.

Several attempts have been made in the past to enhance accuracy in the prediction of stock prices as well as overall performance of stock exchanges.

Most of this prediction models aimed to predict future performance of stocks, on from the analysis of historic data. None of these models succeeded in their prediction accuracy of at least 70–75%.

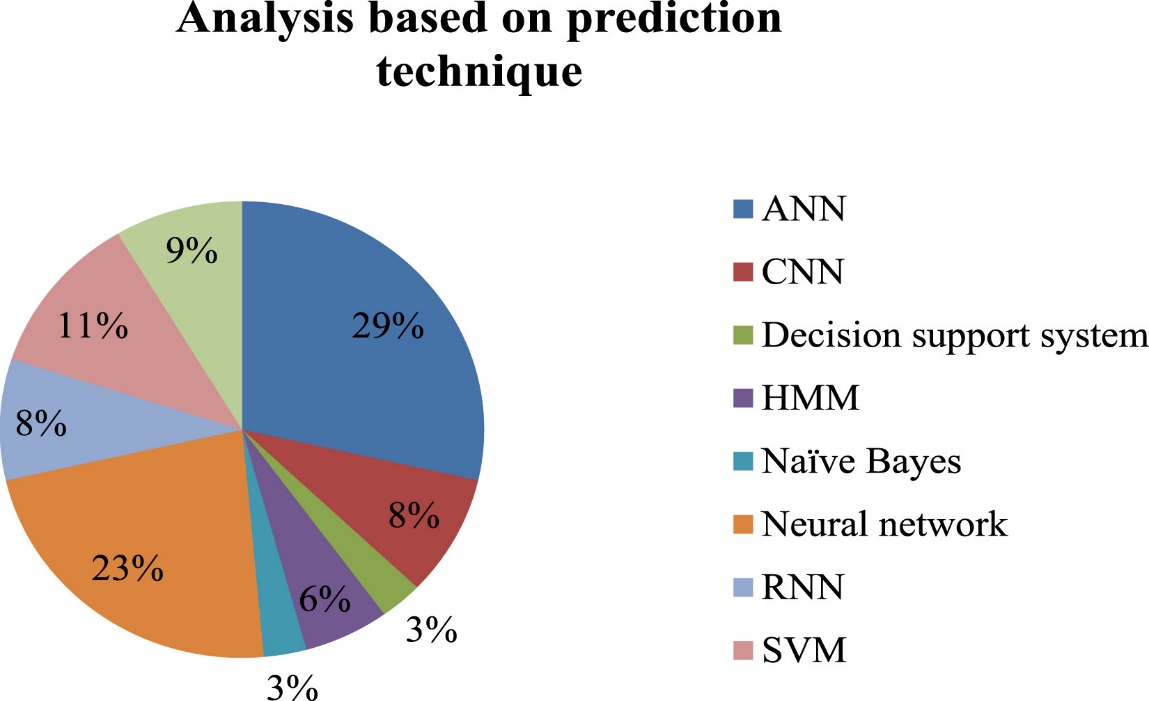


Fig 6 - Percentage of how many times an algorithm is used for prediction

To minimize the errors of prediction, researchers resorted to methods of machine learning, and sentiment analysis of stock reports on the social media. Such measures scarcely achieved their objectives due to uncontrollable trolls of unauthenticated reports on social media and polarized viewpoints of bloggers.

* **Sentiment Analysis Component**

People use social networking sites, like Facebook, Twitter, etc. to express their comments and opinions about a particular topic such as news, movie, event and remarks related to product. Sentiment classification could be done in word/phrase level, sentence level and document level. Sentiment analysis is used to extract such opinion and remarks of users by classifying them as positive, negative and natural sentiment.

Here we shall be using this approach to predict whether the stock price will increase or decrease (1 or 0) based on the top news headlines collected form one such social media platform.

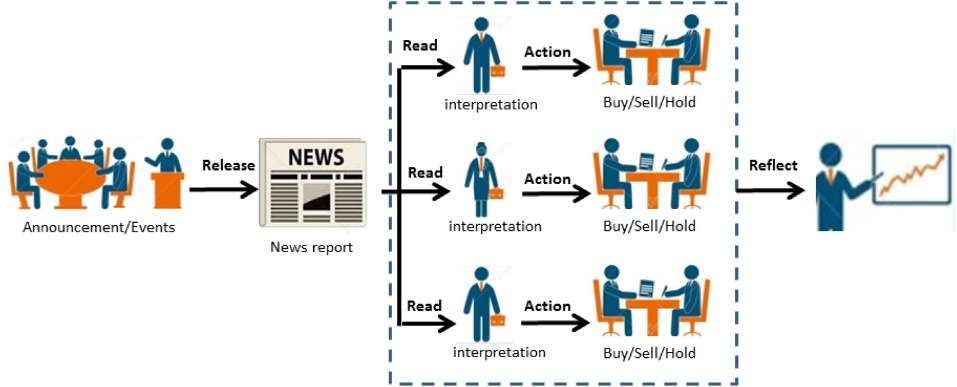


Fig 7 - Understanding the Daily News affect on Market

The figure above depicts how the stock market gets effected by daily news and how the prices of the stock are affected by constant buy and sell based on small events like appointment of new CEO, or Release of new product, etc.

**THE GOAL OF THE STUDY**

Our Goal here is to construct an effective and efficient model that can predict the future trends in the stock market as well as a stock sentiment based on top news headlines for a particular day.

We then compare the results of sentiment analysis as well as historical data analysis in order to see how closely they match, with the minimum error and maximum possible accuracy for the prediction.

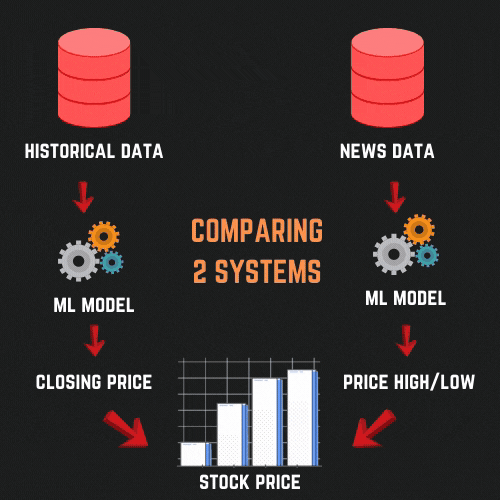
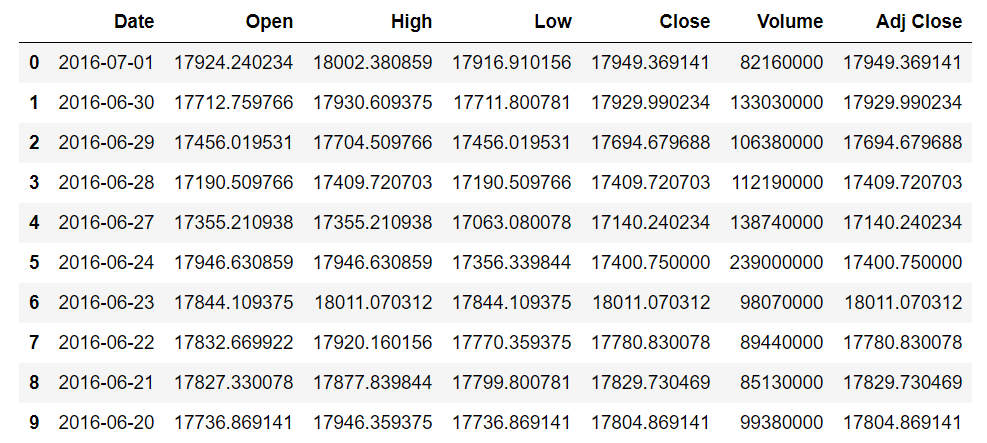


Fig 8 - Goal of our study

**DATASET AND DATA DESCRIPTION**

We are working with the DJI Stock Data. Da-Jiang Innovations, is a Chinese technology company acronymic as DJI. We have taken into the following two datasets for the two components of our study.

1. **Historical Data**



***Dataset Name****: DJI Stock Data*

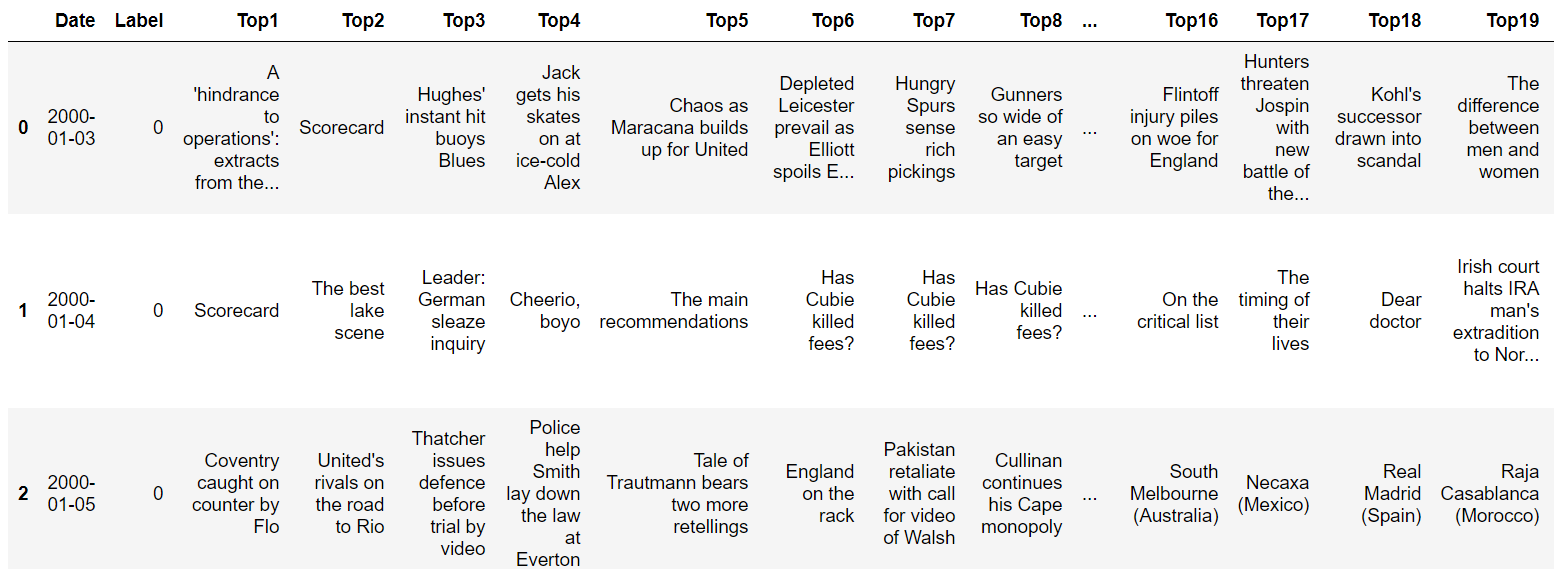
***Dataset Source****: Yahoo Finance*

***Data Description****: The dataset columns mean the following*

* + *Date – Date of the data generation*
  + *Open – The price of the stock when the market opened on that date*
  + *High – The highest price touched by the stock on that date*
  + *Low – The lowest price touched by the stock on that date*
  + *Close – The price of the stock when the market closed on that date*
  + *Volume - Number of shares of a stock traded between its daily open and close*
  + *Adj Close – Adjusted Close Value*

***Task Description****: We are aiming to predict the close price of the given stock DJI, based on the other parameters like Open, High, Low and Volume (for next day close price prediction)*

1. **Sentiment Data**



***Dataset Name****: DJI News Dataset*

***Dataset Source****: Reddit WorldNews Channel | Kaggle*

***Data Description****: The dataset columns mean the following*

* + *Date – Date of the data generation*
  + *Label – Represents whether the Dow Jones Industrial Average (DJI) rose or stayed as the same (1) or decreased (0) on that date*
  + *Rest of the columns are the top 25 news headlines as ranked on Reddit channel*

***Task Description****: We are aiming to predict the Label value (0 or 1) from the sentiment of news headlines.*

**MODELS USED**

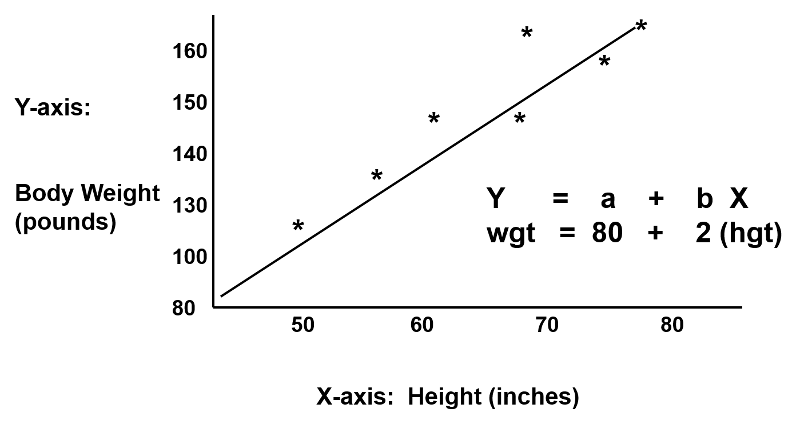
We have tried to implement the following model to achieve our goal.

1. **Historical Data Analysis**

Historical data analysis aims to analyze the daily trade data and predict the close price for a given stock. The proposed model can be used in live prediction on minute-to-minute basis and can also be efficient in implementing stop-loss or target features of famous trading apps like PayTM Money.

***Model-1: Linear Regression***

Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).



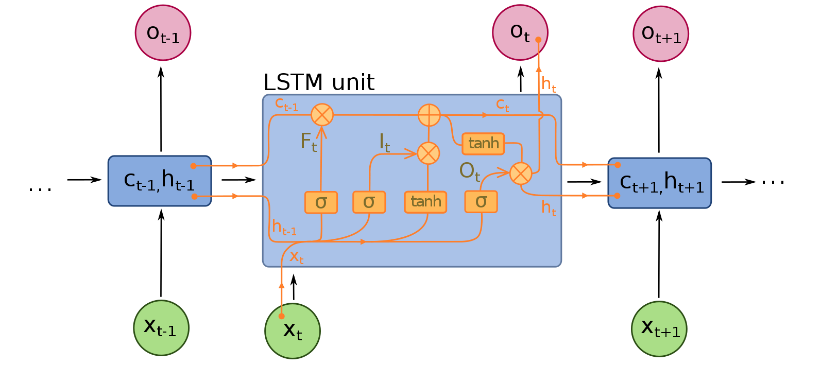
The equation of Linear Regression is as follows:

**y = B0 + B1\*x**

When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression.

We are observing the r2 value, and higher the value, better is the model fitted to the data.

***Model-2: LSTM (Experimental)***



As we can see in figure above, we are using LSTM for analyse Historical Data because LSTMs are widely used for sequence prediction problems and have proven to be extremely effective. The reason they work so well is because LSTM is able to store past information that is important, and forget the information that is not. LSTM has three gates:

* **The input gate:** The input gate adds information to the cell state
* **The forget gate:** It removes the information that is no longer required by the model
* **The output gate:**Output Gate at LSTM selects the information to be shown as output

1. **Sentiment Analysis**

Sentiment analysis component explains the analysis of stock-related news. The primary objective is to classify the reports to be either positive or negative. Several data preprocessing routines are performed on the stock reports. This action is followed by the classification of the news data using the lexical approach.

***Model-1: Logistic Regression***

The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It’s an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

**1 / (1 + e-value)**

Where e is the base of the natural logarithms and value is the actual numerical value that you want to transform. Below is a plot of the numbers between -5 and 5 transformed into the range 0 and 1 using the logistic function.



Logistic regression uses an equation as the representation, very much like linear regression. Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modeled is a binary values (0 or 1) rather than a numeric value.

Below is an example logistic regression equation:

**y = e^(b0 + b1\*x) / (1 + e^(b0 + b1\*x))**

**or**

**p(X) = e^(b0 + b1\*X) / (1 + e^(b0 + b1\*X))**

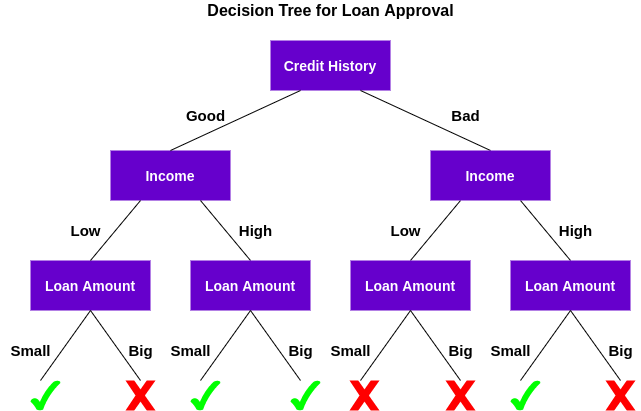
Where y or p(x) is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.

Logistic regression models the probability of the default class.

***Model-2: Random Forest***

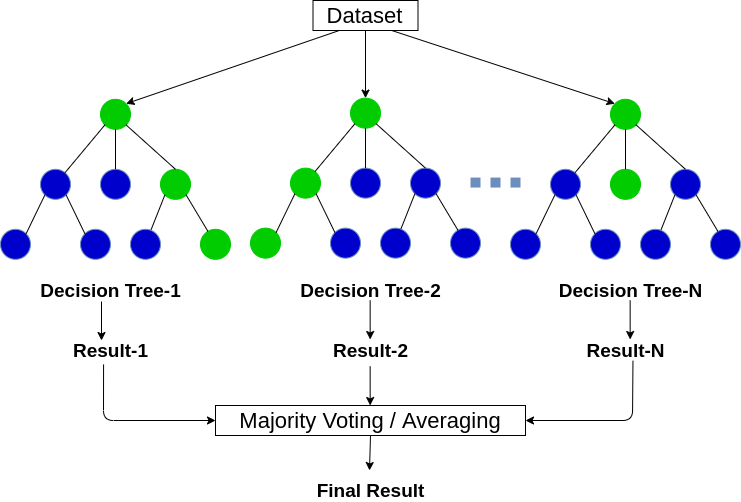
Random Forest Model is an extension or advanced version of the Decision Tree Classifier.

A decision tree is a supervised machine learning algorithm that can be used for both classification and regression problems. A decision tree is simply a series of sequential decisions made to reach a specific result. Here’s an illustration of a decision tree in action:



So, a decision tree makes a series of decisions based on a set of features/attributes present in the data, which in this case were credit history, income, and loan amount. Based on the outcomes from checking these three features, the decision tree decides if the customer’s loan should be approved or not.

The decision tree algorithm is quite easy to understand and interpret. But often, a single tree is not sufficient for producing effective results. This is where the Random Forest algorithm comes into the picture.



This process of combining the output of multiple individual models (also known as weak learners) is called Ensemble Learning. Random Forest is suitable for situations when we have a large dataset,

**RESULTS OBTAINED**

The implementation of algorithms is presented in the attached code files. Following is the comparison of results obtained for each algorithm.

* **Historical Data Analysis**

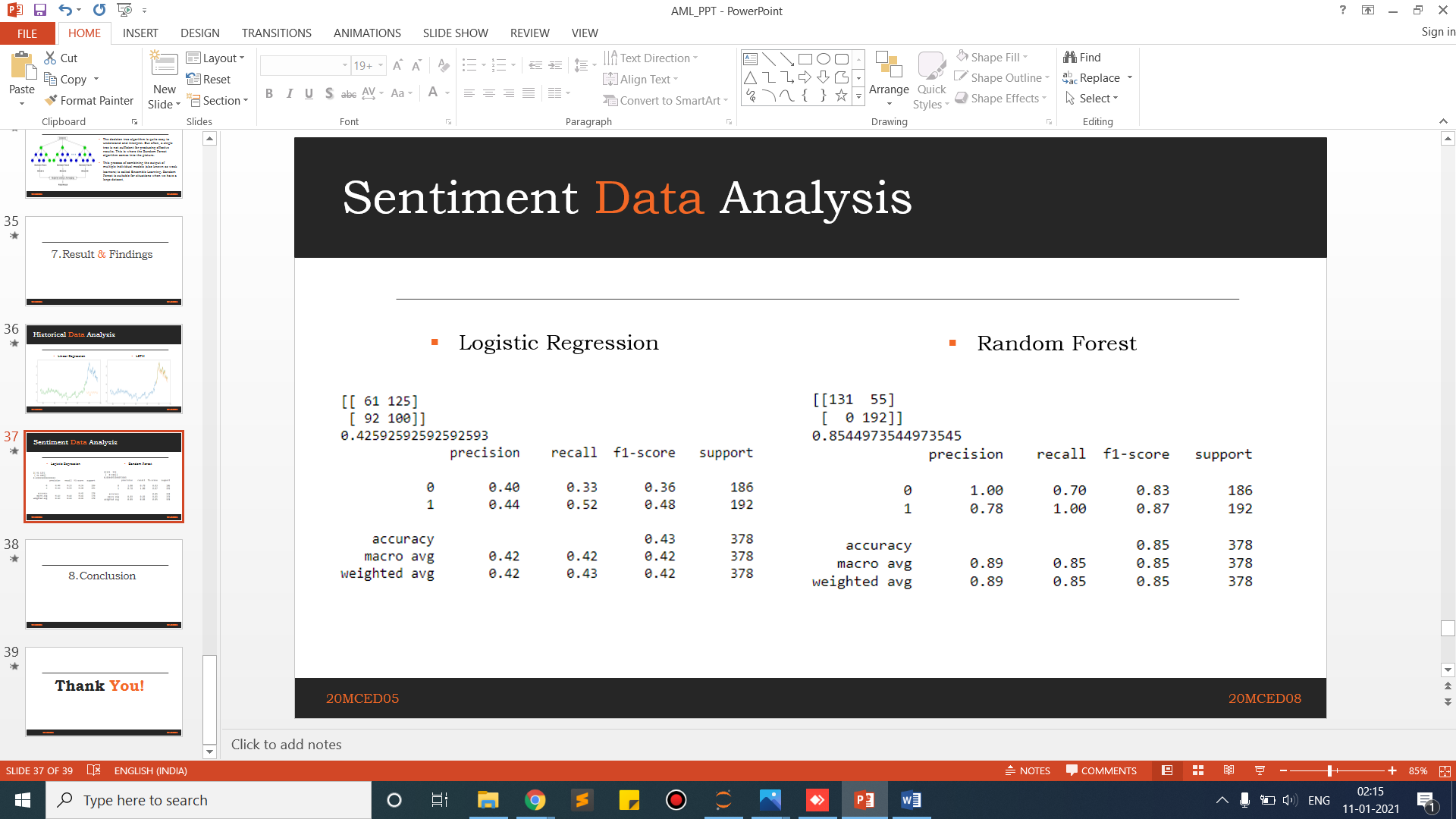
Using Linear Regression, we obtain an r2 = 0.9799

RMS = 135.072

Target Column = Close

Feature Column = Open

* **Sentiment Data Analysis**



Logistic regression became too simple of a model to classify and predict the price as to be going high or low using sentiment data like news headlines. While random forest predicted our stock price direction with higher accuracy based on the news headlines sentiment.

**CONCLUSION**

Since we know that market is volatile and keeps changing every minute, all these parameters are generated again and again. Thus, this prediction models can be used in building features like stop-loss or target threshold similar to those in famous trading apps like PayTM Money.

We also observed that our models are a little naïve and we need to implement more feature rich models to analyze stock market data like Support Vector Machine or Artificial Neural Network for more accurate predictions and sentiment classification.