**Machine Learning for Long Term Stock Price Prediction**

“submitted towards partial fulfilment of the criteria for award of PGPBA by GLIM”

Submitted by: Sahil Makkar, Utkarsh Kulshrestha,

Ashwini Murthy, Deepti Ramani,

Pavan Kumar Thatha, Showbhik

Group No: 4

Batch: PGPBA-BI 2017, Section A

Project Mentor: Neelesh Singh



GREAT LAKES INSITUTE OF MANAGEMENT, BANGALORE

# **Acknowledgements**

We wish to place on record our deep appreciation for the guidance and help provided to us by our Mentor Mr. Neelesh Singh from Great Lakes Institute of Management for guiding us in completing this project on time.

We would also like to place on record our appreciation for the guidance provided by Dr. P K Vishwanathan for giving us valuable feedback and being a source of inspiration in helping us to work on this project.

Last but not the least we wish to thank Project Office, Bangalore for their support.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Date: 25 November 2017

Place: Bangalore

Sahil Makkar

UtkarshKulshrestha

Ashwini Murthy

Deepthi Ramani

Pavan Thatha

Showbhik

**GREAT LAKES INSTITUTE OF MANAGEMENT**

Postgraduate Program Business Analytics and Business Intelligence



***Certificate***

This is to certify that the PGPBA Capstone Project titled “***Machine Learning for Long Term Stock Price Prediction***” is a bonafide record of the Project work carried out by *Sahil Makkar, Utkarsh Kulshrestha, Ashwini Murthy, Deepthi Ramani, Pavan Thatha and Showbhik* in fulfilment of requirements for the award of PGPBA-BI in Great Lakes Institute of Management.

November 2017

**Neelesh Singh** **Dr. P K Vishwanathan**

Project Mentor Program Director

Date: November 2017

Place: Bangalore

# **Abstract**

The Stock market process is full of uncertainty and is affected by many factors. Hence the Stock market prediction is one of the important exertions in finance and business. There are two types of analysis possible for prediction, technical and fundamental. In this paper both technical and fundamental analysis are considered. Technical analysis is done using historical data of stock prices by applying machine learning and fundamental analysis is done using social media data by applying sentiment analysis. Social media data has high impact today than ever, it can aide in predicting the trend of the stock market. Assuming that news articles have impact on stock market, this is an attempt to study relationship between news and stock trend. The method involves collecting news and social media data and extracting sentiments expressed by individual. Then the correlation between the sentiments and the stock values is analysed. The learned model can then be used to make future predictions about stock values. It can be shown that this method can predict the sentiment and the stock performance and its recent news and social data are also closely correlated.

* Techniques: Predictive modelling and Machine Learning
* Tools: R, Python, Tableau

# **Table of Contents**

[Acknowledgements 2](#_Toc499651725)

[Abstract 4](#_Toc499651726)

[Table of Contents 5](#_Toc499651727)

[List of Tables 6](#_Toc499651728)

[List of Figures 7](#_Toc499651729)

[Chapter 1 8](#_Toc499651730)

[Introduction 8](#_Toc499651731)

[Importance of Stock Price Prediction 10](#_Toc499651732)

[Important Terminologies in Stock Market 11](#_Toc499651733)

[Chapter 2 12](#_Toc499651734)

[Financial Ratios 12](#_Toc499651735)

[Chapter 3 15](#_Toc499651736)

[Data under Study 15](#_Toc499651737)

[Analytical Approach 17](#_Toc499651738)

[Exploratory Data Analysis 19](#_Toc499651739)

[Chapter 4 24](#_Toc499651740)

[Sentiment Analysis 24](#_Toc499651741)

[Predictive Analytics 25](#_Toc499651742)

[Chapter 5 29](#_Toc499651743)

[Recommendations 29](#_Toc499651744)

[Conclusions 30](#_Toc499651745)

[Bibliography 31](#_Toc499651746)

[Annexure 32](#_Toc499651747)

# **List of Tables**

# **List of Figures**

# **Chapter 1**

## **Introduction**

The stock market is essentially a non-linear, nonparametric system that is extremely hard to model with any reasonable accuracy. Investors have been trying to find a way to predict stock prices and to find the right stocks and right timing to buy or sell. To achieve those objectives, and per some research used the techniques of fundamental analysis, where trading rules are developed based on the information associated with macroeconomics, industry, and company. The authors of some of the research said that fundamental analysis assumes that the price of a stock depends on its intrinsic value and expected return on investment. Analysing the company’s operations and the market in which the company is operating can do this. Consequently, the stock price can be predicted reasonably well. Most people believe that fundamental analysis is a good method only on a long-term basis. However, for short- and medium-term speculations, fundamental analysis is generally not suitable. Some other research used the techniques of technical analysis, in which trading rules were developed based on the historical data of stock trading price and volume. Technical analysis as illustrated in and refers to the various methods that aim to predict future price movements using past stock prices and volume information. It assumes that history repeats itself and that future market directions can be determined by examining historical price data. Thus, it is assumed that price trends and patterns exist that can be identified and utilized for profit. Most of the techniques used in technical analysis are highly subjective in nature and have been shown not to be statistically valid.

However, now data mining techniques and artificial intelligence techniques like decision trees, rough set approach, and artificial neural networks have been applied to this area. Artificial neural networks can be trained by using an appropriate learning algorithm. Following the assumption of technical analysis that patterns exist in price data, it is possible in principle to use data mining techniques to discover these patterns in an automated manner. Once these patterns have been discovered, future prices can be predicted.

In this project, we will Introduce a framework in which we integrate user predictions into the current machine learning algorithm using public historical data & text mining to improve our results. Following the assumption of technical analysis that patterns exist in price data, it is possible in principle to use data mining techniques to discover these patterns in an automated manner. Once these patterns have been discovered, future prices can be predicted.

**Objective:**

The main objective of our project is to find out how to predict the future market movement using sentiment data which has analysed news articles & popular news channels. Another objective is to use machine learning to stabilize the algorithm and see if it can lower the risk factors involvement.

**Fundamental Analysis:**

Fundamental Analysis is a method of evaluating a security in an attempt to measure its intrinsic value, by examining related economic, financial and other qualitative and quantitative factors. Fundamental analysts study anything that can affect the security’s value, including macroeconomic factors such as the overall economy and industry conditions and microeconomic factors such as financial conditions and company management. The end goal of fundamental analysis is to produce a quantitative value that an investor can compare with a security’s current price, thus indicating whether security is undervalued or overvalued.

Fundamental analysis involves the process of:

* Looking at the basics or fundamental quantitative and financial environment of a business.
* Examining key ratios to determine financial health
* Estimating the value of the stock

For stocks and equity instruments, this method uses revenues, earnings, future growth, ROE, profit margins and other data to determine a company's underlying value and potential for future growth. In terms of stocks, fundamental analysis focuses on the financial statements of the company being evaluated.



*Fig : Fundamental Approach*

The goal of analysing a company’s fundamentals is to find a stock’s intrinsic value, a term which gives what a stock in really worth – as opposed to a value at which it is being traded. If the intrinsic value is more than the current share price, our analysis is showing that the stock is worth more than its price and that it makes sense to buy this stock. In terms of stocks, fundamental analysis focuses on the financial statements of the company being evaluated.

## **Importance of Stock Price Prediction**

In the finance field, stock market and its trends are extremely volatile in nature. It attracts researchers to capture the volatility and predicting its next moves. Investors and market analysts study the market behaviour and plan their buy or sell strategies accordingly. As stock market produces large amount of data every day, it is very difficult for an individual to consider all the current and past information for predicting future trend of a stock. Mainly there are two methods for forecasting market trends. One is Technical analysis and other is Fundamental analysis. Technical analysis considers past price and volume to predict the future trend whereas Fundamental Analysis on the other hand, Fundamental analysis of a business involves analysing its financial data to get some insights. The efficacy of both technical and fundamental analysis is disputed by the efficient-market hypothesis which states that stock market prices are essentially unpredictable. This research follows the Fundamental analysis technique to discover future trend of a stock by considering news articles about a company as prime information and tries to classify news as good (positive) and bad (negative). If the news sentiment is positive, there are more chances that the stock price will go up and if the news sentiment is negative, then stock price may go down. This research is an attempt to build a model that predicts news polarity which may affect changes in stock trends. In other words, check the impact of news articles on stock prices. We are using supervised machine learning as classification and other text mining techniques to check news polarity. And, be able to classify unknown news, which is not used to build a classifier. We have taken historical, up to last 5 years’ data for analysis.

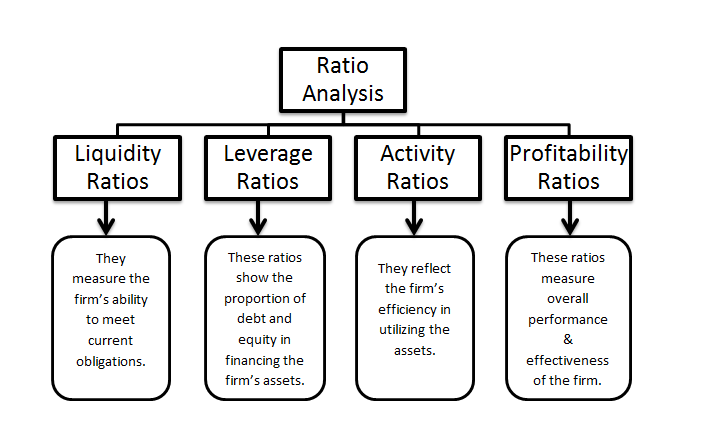
## **Important Terminologies in Stock Market**

# **Chapter 2**

## **Financial Ratios**

The main objective of a financial a report is to provide information on a company’s performance to the internal and external users to take decisions. From the perspective of information economics, accounting and financial reporting play a vital role in an efficient capital market. One of the most common ways of assessing the relative values of stocks among practitioners is to compare the numbers listed in financial statements by using financial ratios. The main advantage of using financial ratios instead of amounts from the income statement is that they are independent of the size of the company. The comparison of financial ratios is used to assess companies’ financial condition, operations and attractiveness as an investment.

Financial ratios defined as the numerical value created from two or more values taken from a company’s financial statements i.e. its balance sheet, income statement or statement of cash flow. Typically, financial ratios are presented as a quantified metric in the form of a percentage, multiple or a ratio which aims to evaluate the financial, operational performance and competitiveness of a company.



*Fig : Financial Ratio Analysis*

**Liquidity Ratios**:

The liquidity focuses on a firm’s ability (current assets and current liabilities) to meet its short-term debt obligations. In other words, it lets you know the resources available for a firm to use to pay its current obligation and expenses. If a company cannot maintain a short-term debt-paying ability, it will not be able to maintain a long-term debt-paying ability, nor will it be able to satisfy its stockholders. The main ratio in this group is the current ratio and acid test ratio.

The ***Current Ratio*** is a liquidity ratio that measures a company's ability to pay short-term and long-term obligations. To gauge this ability, the current ratio considers the current total assets of a company (both liquid and illiquid) relative to that company's current total liabilities.

Current Ratio is the most widely used of all analytical devices based on the balance sheet. It establishes the relationship between Current assets & Current liabilities.

Current Ratio = Current Assets / Current Liabilities

The current ratio is mainly used to give an idea of the company's ability to pay back its liabilities (debt and accounts payable) with its assets (cash, marketable securities, inventory, accounts receivable). As such, current ratio can be used to take a rough measurement of a company's financial health.

It compares a firm's current assets to its current liabilities, and is expressed as follows: The current ratio is an indication of a firm's liquidity. Acceptable current ratios vary from industry to industry.

Ideal Ratio: 2:1

High Ratio indicates under trading and over capitalization.

Low Ratio indicates over trading and under capitalization.

**Leverage Ratio**:

A leverage ratio is any one of several financial measurements that look at how much capital comes in the form of debt (loans), or assesses the ability of a company to meet its financial obligations. The leverage ratio is important given that companies rely on a mixture of equity and debt to finance their operations, and knowing the amount of debt held by a company is useful in evaluating whether it can pay its debts off as they come due.

Too much debt can be dangerous for a company and its investors. However, if a company's operations can generate a higher rate of return than the interest rate on its loans, then the debt is helping to fuel growth in profits. Nonetheless, uncontrolled debt levels can lead to credit downgrades or worse. On the other hand, too few debts can also raise questions. A reluctance or inability to borrow may be a sign that operating margins are simply too tight.

The most well-known financial leverage ratio is the ***debt-to-equity ratio***.

D/E Ratio = Total Debt / Total Equity

The D/E ratio indicates how much debt a company is using to finance its assets relative to the value of shareholders' equity.

A high debt/equity ratio generally indicates that a company has been aggressive in financing its growth with debt. This can result in volatile earnings because of the additional interest expense. If the company's interest expense grows too high, it may increase the company's chances of a default or bankruptcy. Typically, a D/E ratio greater than 2.0 indicates a risky scenario for an investor however this yardstick can vary by industry.

**Profitability Ratio:**

The profitability ratios group, also known as performance ratios, assesses the company ability to earn profits on sales, assets and equity, it measures the return earned on a company’s capital and the financial cushion relative to each dollar of sales, These are critical to determining the attractiveness of investing in company shares, and investors in using these ratios widely, much like the operational performance ratios, these ratios give users a good understanding of how well the company utilized its resources(assets) in generating profit and shareholder value.

Some examples of profitability ratios are profit margin, [return on assets](https://www.investopedia.com/terms/s/seasonality.asp) (ROA) and [return on equity](https://www.investopedia.com/terms/p/profitmargin.asp) (ROE). Profitability ratios are the most popular metrics used in [financial analysis](https://www.investopedia.com/terms/f/financial-analysis.asp).  
The long-term profitability of a company is vital for both the survivability of the company as well as the benefit received by shareholders.

# **Chapter 3**

## **Data under Study**

**Financial Data**

The volume and variety of data used by stock market investors have exploded in recent years. Gone are the days when stock prices and company fundamentals sufficed to build a robust investment strategy.

Today, analysts and traders use a far wider set of data to inform their investment decisions: market structure data, sentiment and psychology data, consensus and survey data and much more.

**Quandl** has the richest collection of publicly available equity data on the internet. Quandl offers a simple API for stock market data downloads. Our daily data feeds deliver end-of-day prices, historical stock fundamental data, harmonized fundamentals, financial ratios, indexes, options and volatility, earnings estimates, analyst ratings, investor sentiment and more.

Quandl’s data products come in many forms and contain various objects, including time-series and tables. Through APIs and various tools (R, Python, Excel, etc.), users can access/call the premium data to which they have subscribed.

We are studying the fundamental stock ratios available in Quandl using Python API.

Reliable stock quotes are the foundation of any trading strategy. Analysts require accurate current stock quotes, to judge prevailing market conditions and bias-free historical stock prices, for back testing and research.

**Stock Quotes**

There are a multitude of stock quote providers; however, not all of them are equivalent in quality. At a minimum, a good stock data publisher should have excellent, robust, transparent methodology for data collection, ideally directly from primary sources (exchanges). Historical stock prices need to be correctly adjusted for splits, dividends, spinoffs, mergers and other corporate actions.

Comprehensive documentation is a must: You should ask how the data is sourced, how it is organized, how it is adjusted if necessary, and how it can be accessed by the user. The data itself must be pristine: free of errors, gaps and outliers. Timeliness, reliability and consistency are essential for professional applications. Responsive customer support is another way that providers can add value to their data.

Surprisingly few data publishers satisfy all these criteria; nonetheless, the best stock market analysts insist on these. Quandl has aggregated several databases from a variety of publishers that score highly on all these components. These publishers cover both current and historical stock quotes, spanning different exchanges and stock classes. A list of different types of stock quotes follows:

The following five end-of-day stock price databases cover a single exchange each, with comprehensive quotes for that exchange.

* NYSE Stock Prices – End-of-day and historical quotes for all 11,500 stocks traded on the New York Stock Exchange.
* NASDAQ Stock Prices – End-of-day and historical prices for all 28,500 NASDAQ stocks.
* NYSE MKT (AMEX) Prices – Historical and end-of-day quotes for all 12,000 NYSE MKT stocks (formerly American Stock Exchange).
* OTC Bulletin Board Prices – Prices for 12,000 over-the-counter bulletin board stocks, with history going back to 2007.
* OTC Markets Group Prices – Current and historical quotes for over 48,000 OTC Markets Group stocks (formerly Pink Sheets).

The above five stock market databases all include daily quotes, historical prices, dividends, splits and other adjustments.

< financial ratios dataset > < python code to extract data from Quandl >

**Sentiment Data**

For performing sentimental analysis, we have collected data from Twitter related to the stocks that we have selected.

Tweets are classified as positive, negative and neutral based on the sentiment present. Approx. 1000 tweets out of the total tweets are examined manually and scaled as 1 for Extremely negative, 2 for Negative, 3 for Neutral, 4 for Positive and 5 for Extremely positive emotions. For classification of other tweets a machine learning model is trained whose features are extracted from the manually scaled tweets.

*Fig : Postive – Negative sentiments*

< code to extract twitter data> < sample tweets file>

## **Analytical Approach**

The stock market is tricky. It's one of the more difficult predictive challenges; there are inherent numerical patterns as well as a wide range of interactions with the outside world. As a market, it is continuously adjusting to the latest prediction theories in continuous coevolution.

With the advent of the digital computer, stock market prediction has since moved into the technological realm. The most prominent technique involves the use of artificial neural networks (ANNs) and Genetic Algorithms. The most common form of ANN in use for stock market prediction is the feed forward network utilizing the backward propagation of errors algorithm to update the network weights. These networks are commonly referred to as Backpropagation networks. Another form of ANN that is more appropriate for stock prediction is the time recurrent neural network (RNN) or time delay neural network (TDNN).

There are usually two approaches taken for forecasting different time horizons: independent and joint. The independent approach employs a single ANN for each time horizon, for example, 1-day, 2-day, or 5-day. The advantage of this approach is that network forecasting error for one horizon won't impact the error for another horizon—since each time horizon is typically a unique problem. The joint approach, however, incorporates multiple time horizons together so that they are determined simultaneously. In this approach, forecasting error for one time horizon may share its error with that of another horizon, which can decrease performance. There are also more parameters required for a joint model, which increases the risk of overfitting.

The use of Text Mining together with Machine Learning algorithms received more attention in the last years, with the use of textual content from Internet as input to predict price changes in Stocks and other financial markets. The collective mood of Twitter messages has been linked to stock market performance.

The model that we have right now consist of several stages:

* First, we perform adaptive filtering of the raw price data
* Then we feed filtered data into the NN model and train it
* Then we perform predictions and prepare the recommendations using those predictions.

Few highlights:

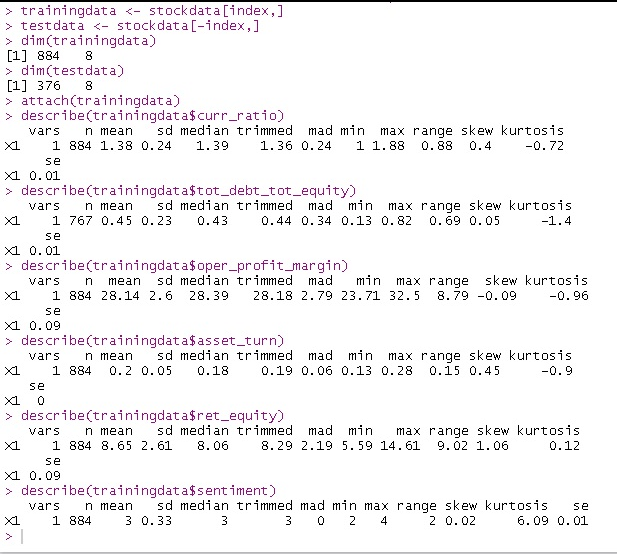
* A data analytic approach is proposed for stock market prediction.
* The proposed decision support system effectively uses fewer variables.
* Higher prediction accuracy is achieved compared to the literature.
* Sophisticated data analytics methods present capability to capture nonlinearity.
* This generic methodology is applicable to the other emerging markets as well.

## **Exploratory Data Analysis**

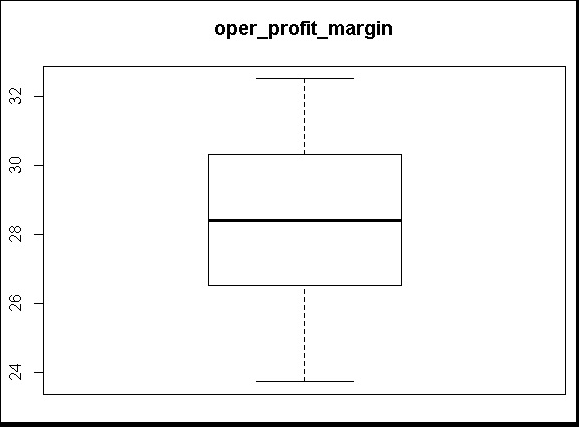
In case of Financial data, we did not face any discrepancies as the data was clean.

Tweets consists of more noisy words such as many numbers, more white spaces, tabs, punctuation characters, stop words, acronyms, emoticons and unnecessary data like pictures and URL’s. So, tweets are pre-processed to represent correct emotions of public. For pre-processing of tweets, we employed three stages of filtering: Tokenization, stop words removal and regex matching for removing special characters.

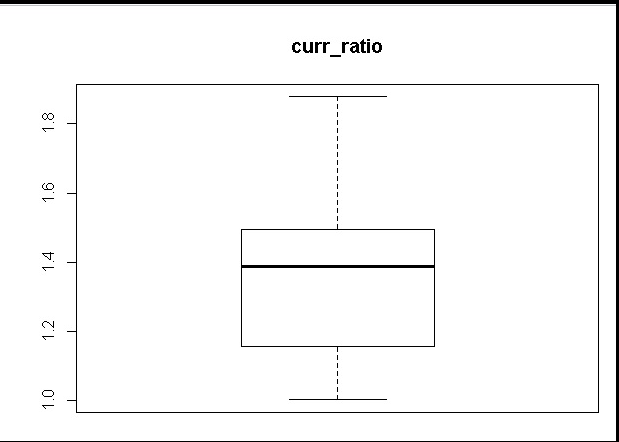
We have picked few sample stocks for EDA and here are the findings.



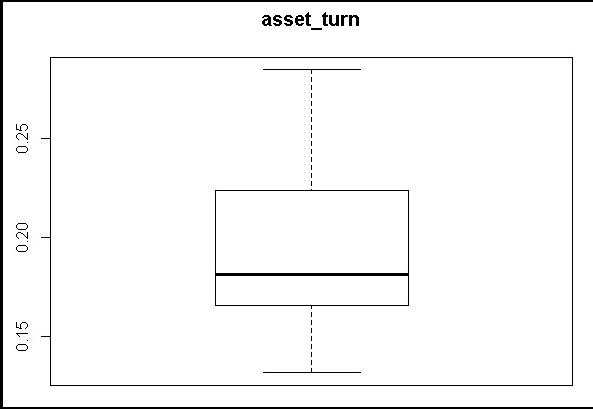
*Fig : Exploratory Data Analysis*



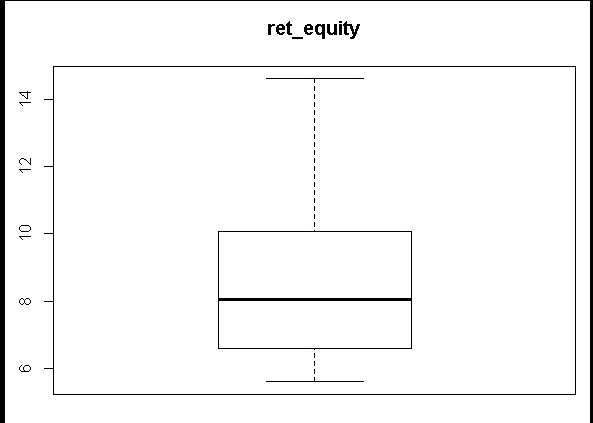
*Fig : Operational Profit Margin*



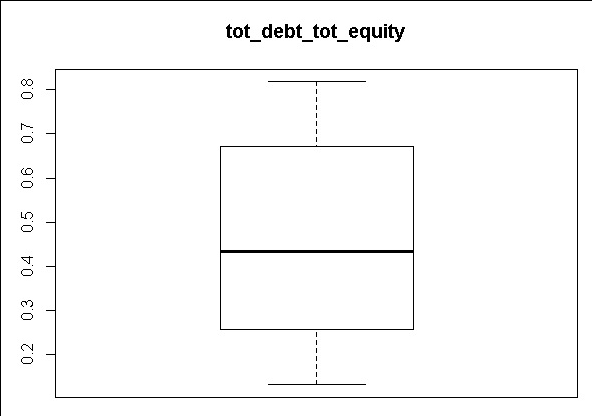
*Fig : Current Ratio*



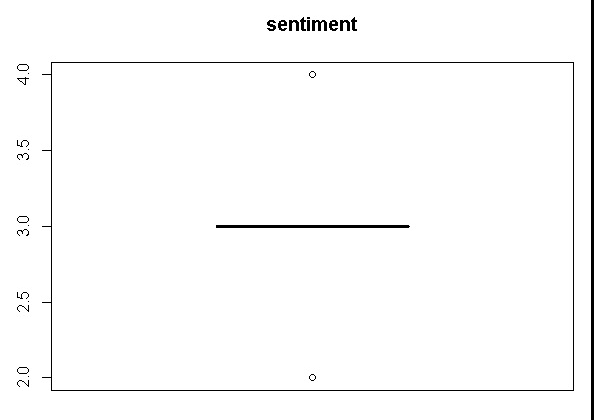
*Fig : Asset Turn*



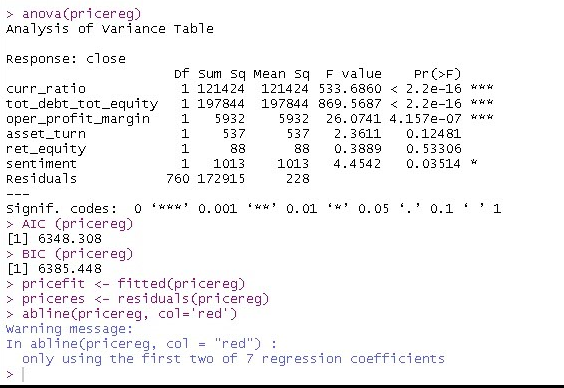
*Fig : Return on Equity*



*Fig : Total Debt to Total Equity*



*Fig : Sentiment*



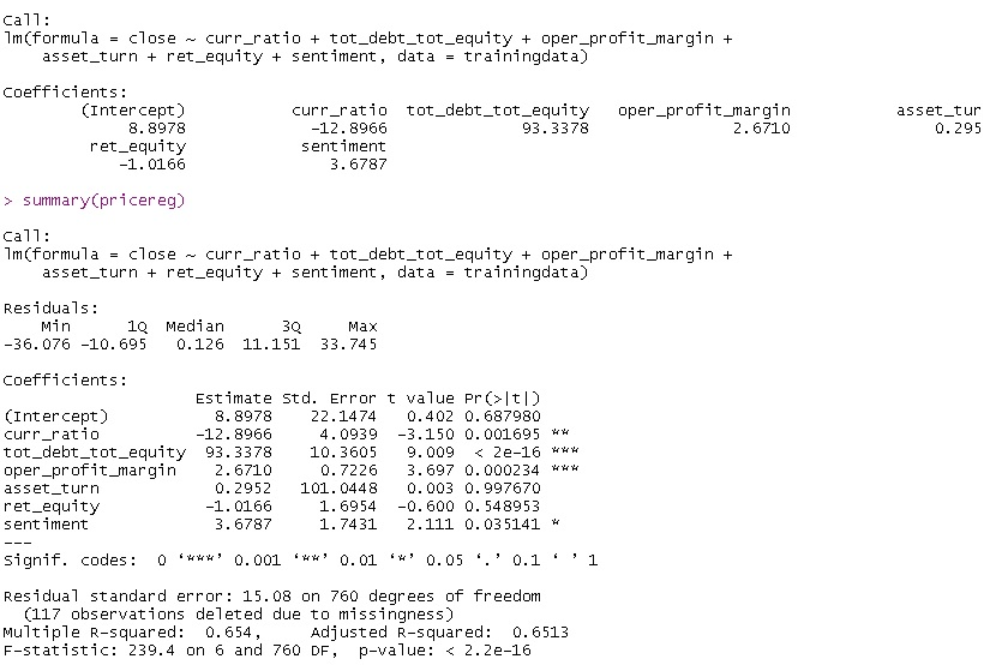
# **Chapter 4**

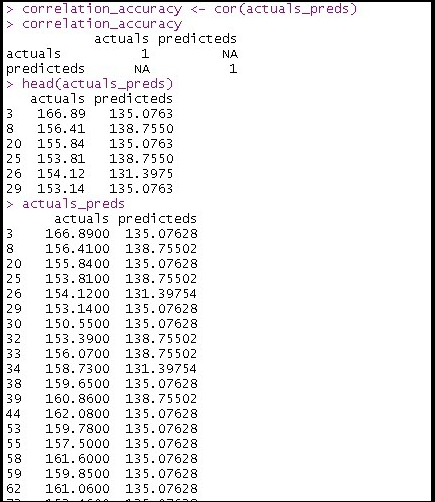
## **Sentiment Analysis**

## **Predictive Analytics**

In order to find a suitable technique for predicting the stock price accurately, we tried a pool of 14 prediction models (7 with regression and other 7 for Neural Networks)

Regression with Sentiment scores couldn’t result a reasonable prediction model, here are the few sample screenshots.





**Neural Network** model yielded a much better prediction model with the below sample outputs.

**Neural Model:**

We have multi-layer model having two LSTM layers, followed by one feed forward layer and one softmax based output layer.

Layer (type) Output Shape Param #

=================================================================

lstm\_3 (LSTM) (None, 5, 128) 69120

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

lstm\_4 (LSTM) (None, 128) 131584

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

dropout\_2 (Dropout) (None, 128) 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

dense\_3 (Dense) (None, 64) 8256

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

dense\_4 (Dense) (None, 3) 195

=================================================================

Total params: 209,155

Trainable params: 209,155

Non-trainable params: 0

Neural model classifies Stocks based on 5 days of sequence and predicting 5th day. It's multi feature sequence model.

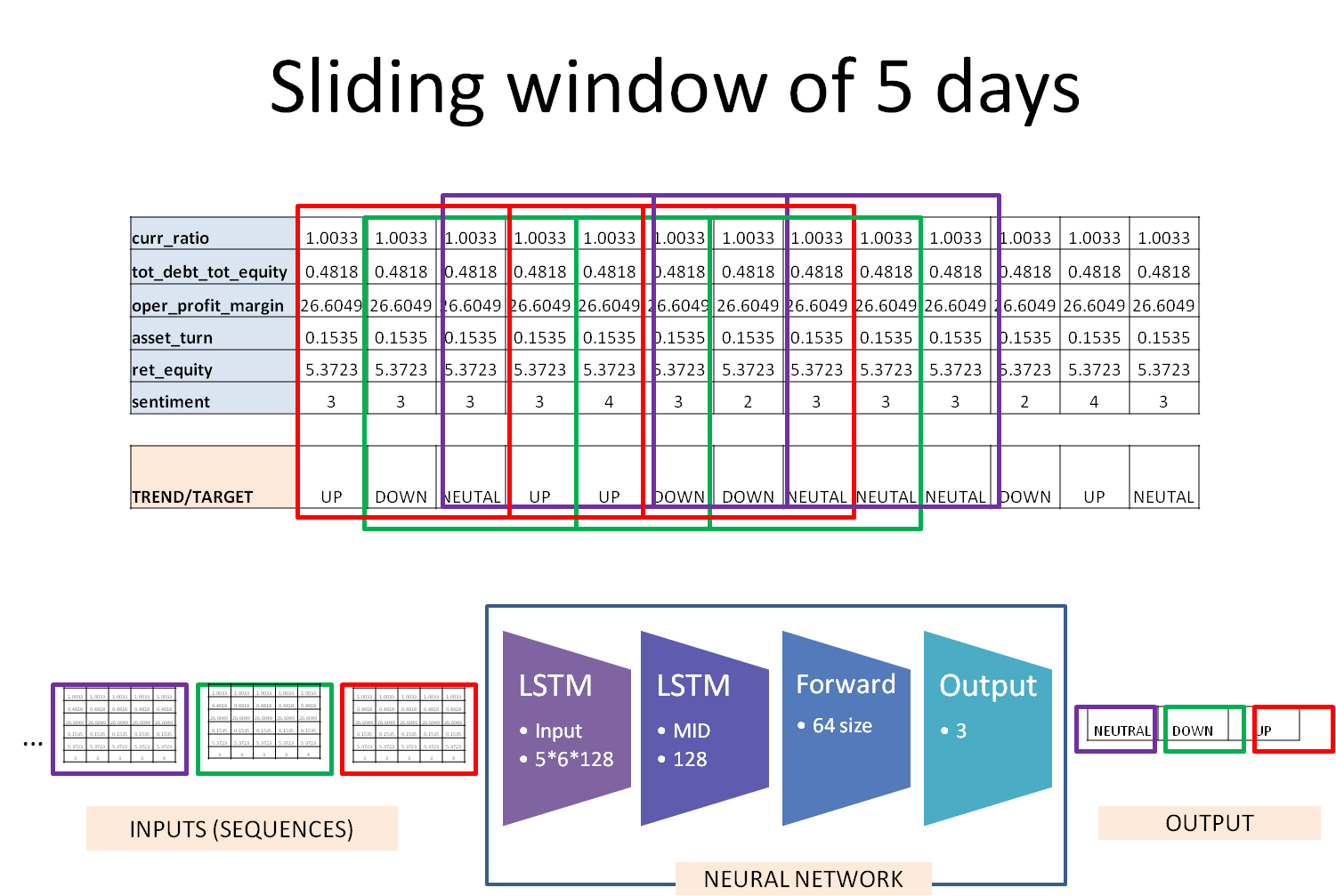
We have 3 classes :

1. Positive recommendation (where a day gain is more than 1%),
2. Negative recommendation (where a day loss is greater than 1%) and
3. Neutral recommendation for all other cases

**Approach:**

**Key features are:**

1. Inputs to ML algorithm (neural network) are formed with 5 days window to understand weekly impact.
2. 3 class neural network to predict: SELL/BUY/HOLD is done
3. Neural network has 3 layers, 2 LSTM to capture the sequence impact, one feed forward and then finally 3 class output classifiers.
4. Regression based model because of limited data is not considered, but same model can be converted to regression to predict prices itself.



Given model gives **96% of accuracy** in predictions. Out-sampling cases need to be handled better. Lag of data can help in better accuracy across classes.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

# **Chapter 5**

## **Recommendations**

## **Conclusions**

Every investor has their own strategy, such as particular fundamentals they tend to be fond of and level of risk they are willing to accept. These types of analysis alone are becoming outdated and more effective investing tools have become evident to lending a helping hand in increasing portfolio performance. Machine-based algorithms, which can analyze many stocks simultaneously and determine quantifiable founded objective predictions, are becoming increasingly more popular to investors as an improved strategy for optimizing returns and mitigating prospective risk. There still is no flawless way of picking stocks but utilizing advanced algorithms based on predictive analytics in conjunction with reasonable risk management and fundamental analysis can potentially help improve portfolio performance for investors.

We have shown that a strong correlation exists between rise/fall in stock prices of a company to the public opinions or emotions about that company expressed on twitter through tweets. We have developed a sentiment analysis tool that can judge the type of sentiment present in the tweet. The tweets are classified into three categories: positive, negative and neutral. At the beginning, we claimed that positive emotions or sentiment of public in twitter about a company would reflect in its stock price. Our speculation is well supported by the results we have achieved.

# **Bibliography**

# **Annexure**