Hackathon Submission

**ICADABAI 2017**

Predicting Stock Prices and Detecting Forward Looking Sentences

horizontal line

# machine_learning-1024x724.jpg

# Introduction

This report covers the work we did on the submission for the Hackathon, the resources and techniques we employed and how we arrived at the results. The first part of the report describes how we built a model that predicts stock prices and the second part covers the process by which we detect forward looking sentences.

# Problem Statement

The objective of this analytics competition is to develop a statistical model that is able to predict various financial parameters while exhibiting comprehensive knowledge of the subject. As is the case with running any industry/business, there is always a huge amount of stress caused by several financial conundrums. Being able to predict such a factor is beneficial to people planning on investing in one of those companies. This leads to the first issue at hand – assessing financial stress of a company using its stock market data. But a high value of stress need not necessarily mean that investing in those companies is not a good idea. If that were the case, we’d all be rich. One of the parameters this depends upon is the sector of the company at hand. Even after taking all these into account, there still lies a tenuous amount of parameters. A company that might be on the downside may become better depending on the actions that it is willing to take in the future for the purpose of development. Thus giving the final problem statement – to identify forward looking sentences from the company’s annual report.

# Methodology

# Predicting Stock Prices

## Reading Financial Data from the PDF

The initial process of obtaining desired information from the annual reports of various industries, involved manipulation of PDF documents. As these documents were meant to be read-only and had no underlying structure, extraction of a very highly structured form of data, the table, proved to be a predicament we were not able to resolve. We used several techniques such as conversion of the PDF documents into HTML before data extraction, converting the documents into ppm images and applying optical character recognition in order to obtain the required data, using various third party libraries, with slim to no documentation, but failed to achieve consistent results.

As usage of provided data (annual reports) was our main priority at the time, we did not opt for any external information and ended up uncomfortably close to the proposed deadline. Thus, we made use of financial data provided by the website – [www.moneycontrol.com](http://www.moneycontrol.com). This website provides data for various financial parameters over the course of five years for all the 186 industries. We made use of the “requests” library in order to mine and obtain required information from the given URL. This HTML data was manipulated using the “BeautifulSoup” library and stored in the csv format. We then use this tabular data in order to conduct data analysis and predict the best approximation of the result.

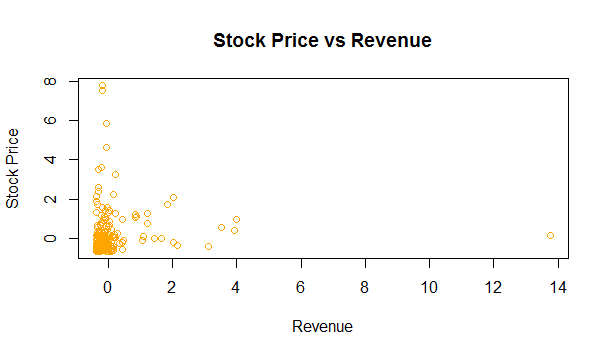
## Detecting Forward Looking Sentences

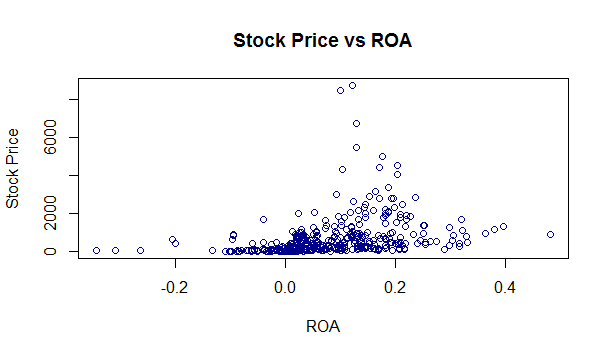
As none of the data required was present in a tabular format, we were easily able to use commonly available python packages to scrape data from the PDFs. The library “PDFMiner” seemed to be most popular; however, ultimately we settled on PyPDF2 for its speed of reading (PyPDF2 took less than a fifth of the time!). We then cleaned the raw data acquired, tokenized the text into sentences and scanned the various sentences for keywords/phrases indicating it was a forward looking sentence. This was done with the help of NLTK library in Python. Some examples of the keywords used are “looking forward”, “in the future”, etc.

We then performed sentiment analysis on the identified sentences to obtain the general sentiment amongst the company directors. To perform sentiment analysis, we used two different approaches - the inbuilt sentiment analyser from NLTK (Vader) as well as a naive dictionary based approach utilising the dictionaries provided by Professor Bill McDonald of the University of Notre Dam. These dictionaries were used as they were specifically compiled for financial statements.

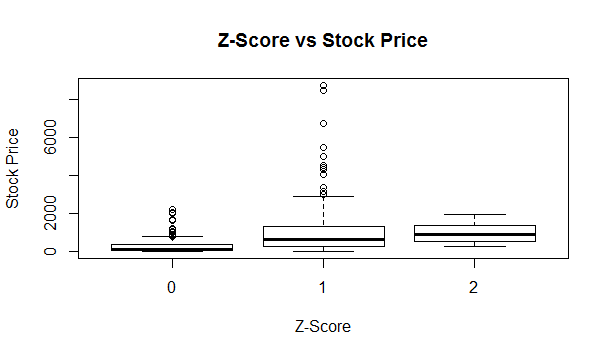
## Building a model to predict Stock Prices

Apart from what we found above, we also found various indicative ratios used in Z-Score and DuPont Analysis. These were done on the assumption that these ratios would play a vital role in predicting stock prices. Additionally, we determined the sector each company belonged to - the hypothesis being that the sector as a whole could help predict the general trend of the company. Finally, the rolling mean of closing stock prices for a 30 day window was determined and its mean for each year was taken as the value to predict. We fed all the data into a Random Forest Regression classifier in R and used it to predict Stock Prices.

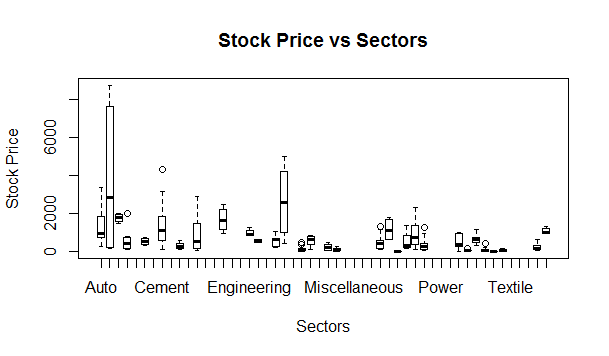
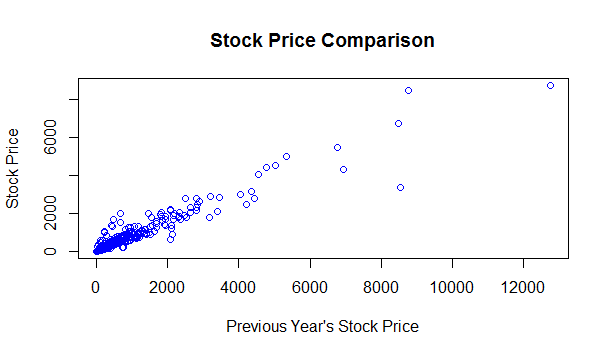
While determining the variables that should be part of our predictor, we found that many quantities obtained from the financial sections had little to no bearing on the final regressor. For Example, when revenue was plotted versus the Stock Price, there was almost no correlation between the two (shown left).

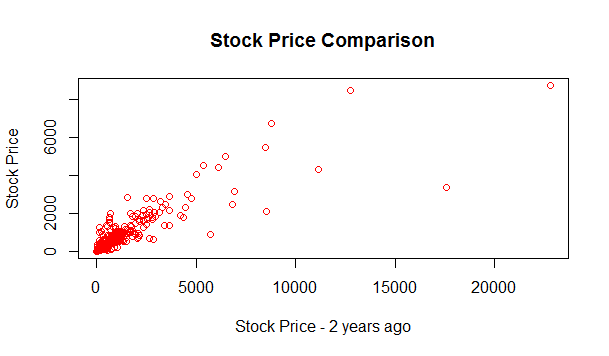


However, we found better relations between stock prices and diagnostic rations such as ROE (Return On Equity) and ROA (Return On Assets). Looking at the graphs, it can be inferred that both ROE and ROA seem to have cubic functions.



On the other hand, well known metrics such as the Altman Z-Score showed little impact on the stock price.

Our hypothesis of the sector playing a part in stock price seemed to hold good however as can be seen from the box plot below. Each sector seems to be distinctly separable which is very helpful during prediction. More than anything else, the best predictors we found were the stock prices over the past two years.



The stock prices are also heavily correlated and serve as the basis of our regression model. Seeing as the stock prices were so important, we also computed the overall trend of stock prices for a year. This was done by directly fitting each year’s stock prices into a linear regressor and obtaining the coefficients - A negative coefficient corresponding to a downward trend in prices and vice versa. Therefore, the final variables we used for prediction were - Stock price for the past two years, the sector each company belongs to, stock price trends for the past two years, ROE and ROA for the past year and the sentiment of the past year obtained from the general report.

# Result

We built our model on the years 2010 to 2014 and tested the model on the year 2015. Although our results were not very close to the actual stocks, we feel that they provide a good approximation of the trend followed by the stock. The model used was a Random Forest Regressor with 15,000 trees. The predicted and actual stock prices for the year 2015 can be viewed at Output.csv. The Mean Squared Error we obtained was 199,100.3

# References

[1] Beautiful Soup 4.4.0 documentation - <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

[2] Requests 2.13.0 documentation -<http://docs.python-requests.org/en/master/>

[3] Data acquisition -<http://www.moneycontrol.com/>

[4] Financial knowledge acquisition -<https://www.youtube.com/user/sentdex>

[5]<https://en.wikipedia.org/wiki/>

[6] CSV documentation -<https://docs.python.org/2/library/csv.html>

[7] PyPDF2 1.26.0 documentation - <https://pythonhosted.org/PyPDF2/>

[8] Sentiment analysis on financial data -<http://www3.nd.edu/~mcdonald/Word_Lists.html>

[9] Random Forest R package - <https://cran.r-project.org/web/packages/randomForest/randomForest.pdf>

[10] Random Forest Regression - <http://www.bios.unc.edu/~dzeng/BIOS740/randomforest.pdf>