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 these parameters 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

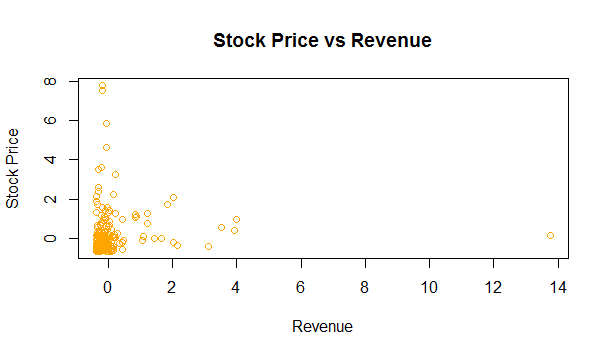
## Reading Financial Data from the PDF

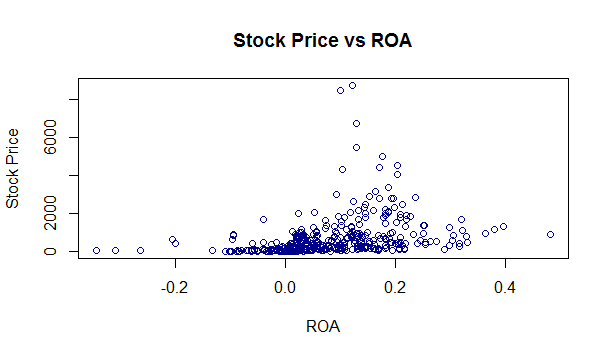
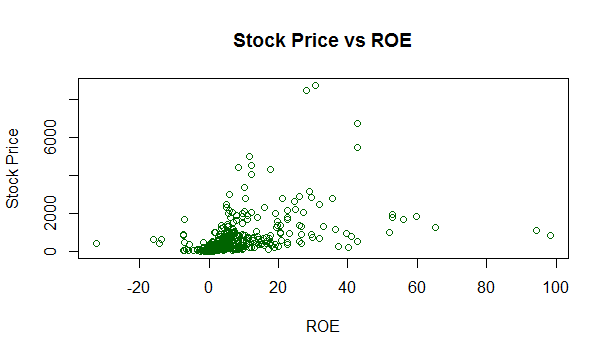
The initial process of obtaining desired information from the annual reports of various industries, involved manipulation of PDF documents. As these documents were meant to be read-only and had no underlying structure, extraction of a very highly structured form of data, the table, proved to be a predicament 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.

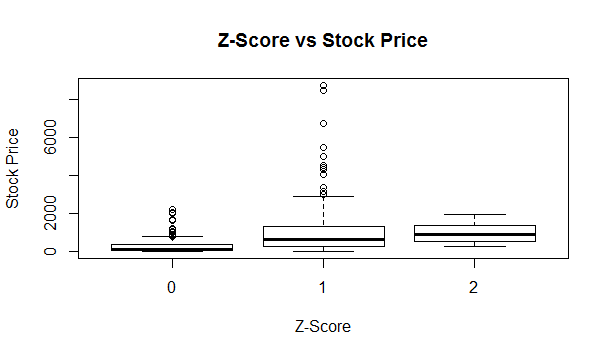
## Building a model to predict Stock Prices

Once we acquired the required data, we used IBM’s Watson to perform sentiment analysis and acquire sentiment. Using Watson’s Discovery API, we were able to obtain sentiment scores for companies. However, due to certain limitations, we were unable to calculate sentiment scores for all the years given. We then found various indicative ratios used in Z-Score and DuPont Analysis. We also had previously acquired sector data. We calculated rolling mean of closing stock prices for a 30 day window and took the mean of that for yearly stock price. We fed all the data into a Polynomial Regression classifier in R and used it to predict Stock Prices.

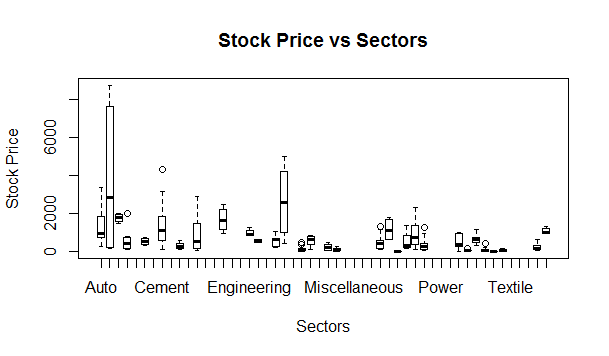
While determining the variables that should be part of our predictor, we found that many quantities obtained from the financial sections had little to no bearing on the final regressor. For Example, when revenue was plotted versus the Stock Price, there was almost no correlation between the two (shown 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. These were therefore added to the regressor as orthogonal polynomials of order 3.



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



The best predictors we found were the sectors and the stock prices over the past two years. The Sectors are clearly separable and serve as a good indicator of stock price.

The stock prices are also heavily correlated and serve as the basis of the linear regression.

# Prev Stock Price.png

# Detecting Forward Looking Sentences

As none of the data required was present in a tabular format, we were easily able to use “PDFMiner” and “Slate” to scrape data from the PDF. 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. We also utilised the dictionaries provided by Professor. Bill McDonald of the University of Notre Dam to get a basic sentiment analysis. Forward looking sentences were identified by keywords such as 'future','looking ahead',etc. We then stored and printed those sentences.

### 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.

**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] Slate 0.5.2 documentation -<https://pypi.python.org/pypi/slate>

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

[9] Linear regression - <http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html>