UNDERSTANDING PORTFOLIO THEORY WITH DATA SCIENCE AND MACHINE LEARNING

ARUN S BHARADWAJ, SPRINGBOARD DATA SCIENCE INTENSIVE CAPSTONE PROJECT

INTRODUCTION

The goal of the Springboard Data Science Intensive capstone project is to understand portfolio theory using data science, machine learning and Python programming. The intended audience for this project are hedge fund managers, academics and students of portfolio theory. The project uses fundamental company data collected from StockPup.com to develop a relationship between stock price and predictor variables. The StockPup data contains 739 csv files with quarterly financial statement information from 1993 to 2017 for 739 companies listed in the US stock market.

SIGNIFICANCE OF THE PROJECT

This project tries to answer some important questions in the realm of investment finance. Investors, hedge fund managers and academics have always been interested in understanding the factors driving stock price growth. The project tries to find if fundamental data can be used to predict stock prices and a stock`s annual rate of return.

WHY IS THIS PROBLEM STATEMENT IMPORTANT

Some academics claim that stock prices are a random function that have only a minor relationship with other economic variables. In contrast, many market participants and other academics claim that stock price movements can be predicted using company specific and macroeconomic variables. If stock prices are a random function, investors will be better off buying index funds since stock selection will be a futile exercise. If stock prices are a function of some underlying variables, stock picking is likely to be more profitable than buying an index fund. This project tries to answer some of these questions at a basic level.

ADDITIONAL DATA POINTS

In addition to the company fundamental data, the project also uses GDP, Inflation and S&P500 returns data. GDP and Inflation are some of the most important macro-economic variables driving stock prices. The S&P500 data is used as a benchmark to determine if a stock`s rate of return exceeds the benchmark return. The S&P500 data is useful in developing the dependent variable in our machine learning models.

DATA EXTRACTION AND REPLACING MISSING VALUES

Data extraction for the project required the import of 739 csv files into Jupyter notebook. Since the name of the company was only displayed in the csv filename and not anywhere inside the csv file, the company name had to be included in each row of the respective csv file. For replacing missing values, Kalman Filtering`s expectation maximization algorithm was used since the data has a time series format.

EXPLORATORY DATA ANALYSIS

MARKET CAPITALIZATION AND PRICE TO EARNINGS RATIO

The scale and complexity of the company fundamental data allows for a lot of exploratory analysis. Figure 1 shows the market capitalization of our dataset from 1994 to present.

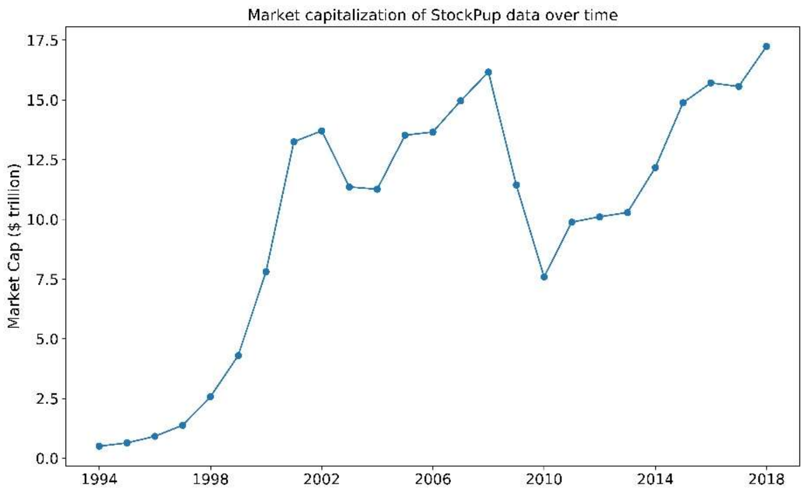
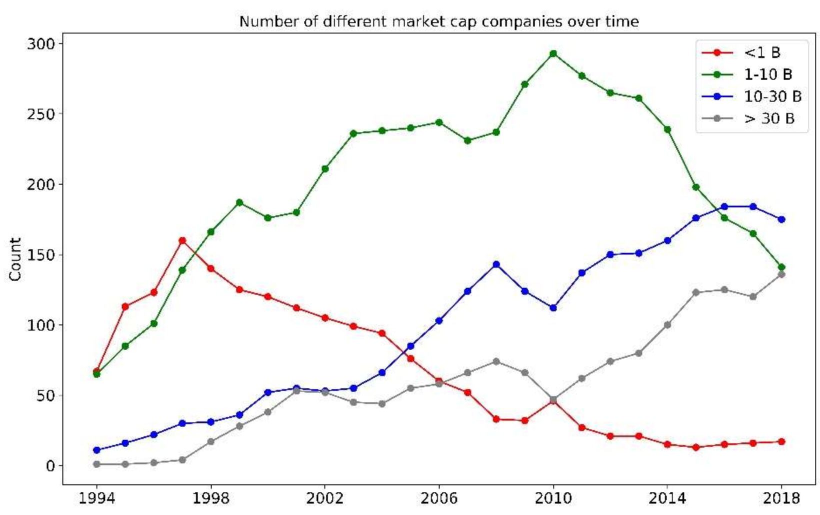


Figure 2 shows the number of different market capitalization companies over time. There are 4 different types of companies divided based on market capitalization: < $ 1 billion, $ 1-10 billion, $ 10-30 billion and > $ 30 billion.

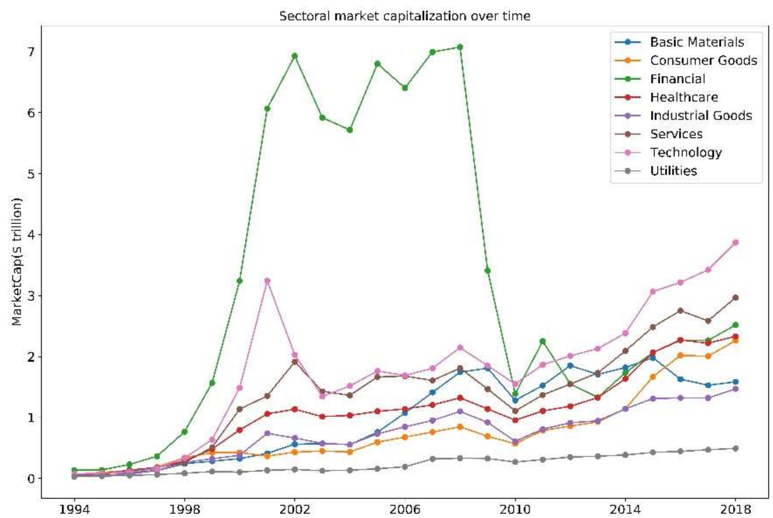
Figure 1



A company`s sub-type may change from year to year, depending on its market capitalization. From figure 2, it can be inferred that from 2010, the number of $ 1-10 billion market capitalization companies has halved while the number of $ 10-30 billion and > $ 30 billion companies have increased substantially.

Figure 2

Figure 3 breaks down market capitalization of our dataset into sub-types based on sectors. While six of these sectors have grown consistently over time, financials and technology have seen the largest increases in market capitalization at different points of time. From 1994 to 2008, financials saw the largest gains in market capitalization and was the largest sector based on market capitalization.



The technology sector was mostly the second largest sector by market capitalization from the year 2000 and overtook financials in the year 2012 to become the largest sector by market capitalization. The utilities sector has seen the slowest growth in market capitalization over time.

Figure 3

The P/E ratio stands for price to earnings ratio. Easiest way to interpret the P/E ratio is to think of it as the number of years needed to breakeven the investment in a company if the company expects to generate the same level of earnings into the future. So, a high P/E ratio implies a longer time to breakeven and hence a riskier investment.

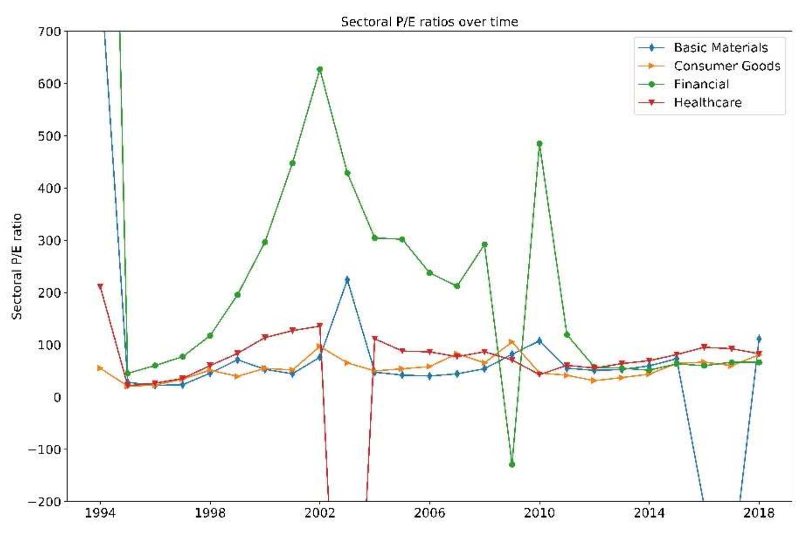


Figure 4

Figure 4 shows two large peak P/E ratios for the financials sector in the years 2002 and 2010. These were years that succeeded financial crises and these peaks were likely caused by large drops in earnings due to the crisis, which was not matched by a proportional drop in market capitalization. Figure 4 shows that the consumer goods sector has seen the lowest volatility in sectoral P/E ratios over time.

LONG TERM DEBT AND EARNINGS

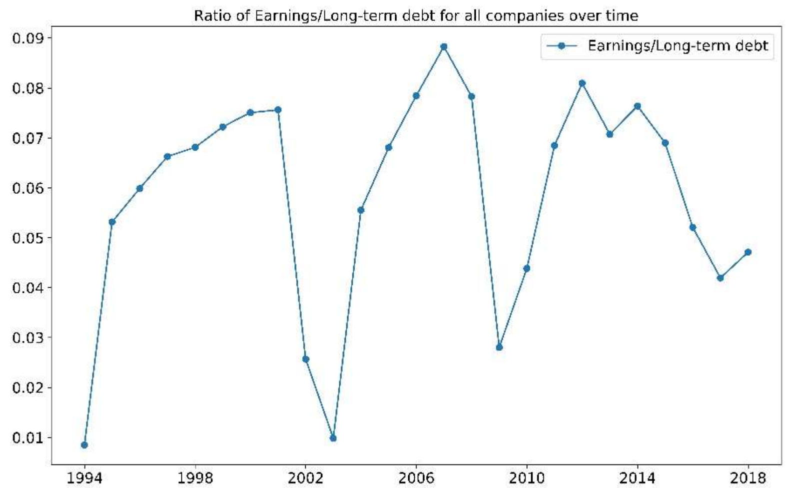
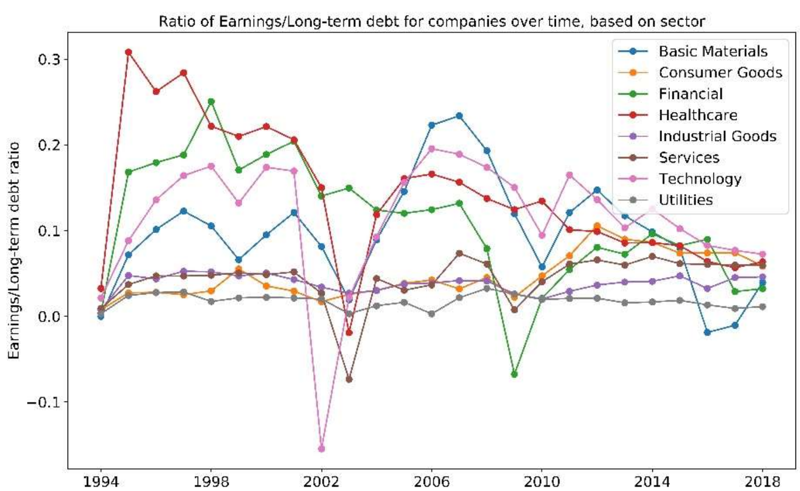


Figure 5 shows large dips in earnings to long term debt ratio after the 1999 tech bubble and after the 2008 financial crisis. This happened due to the sharp drops in earnings experienced by companies after these bubbles, while the amount of long term debt remained constant.

Figure 5



In figure 6, the ratio of earnings to long-term debt based on sectors shows a lot of volatility. Predictably, the ratio dips in the year 2002 for the technology sector and in 2009 for financials. In 2017, all sectors have earnings to long term debt ratio between 0 and 0.1.

Figure 6

STOCK PRICE ANNUAL RATE OF RETURNS

Figure 7 shows the annual rate of return of all companies in the dataset. The annual rate of return is calculated using market capitalization and the cumulative dividends paid out for the year. Figure 8 gives some descriptive statistics about the annual rate of return.

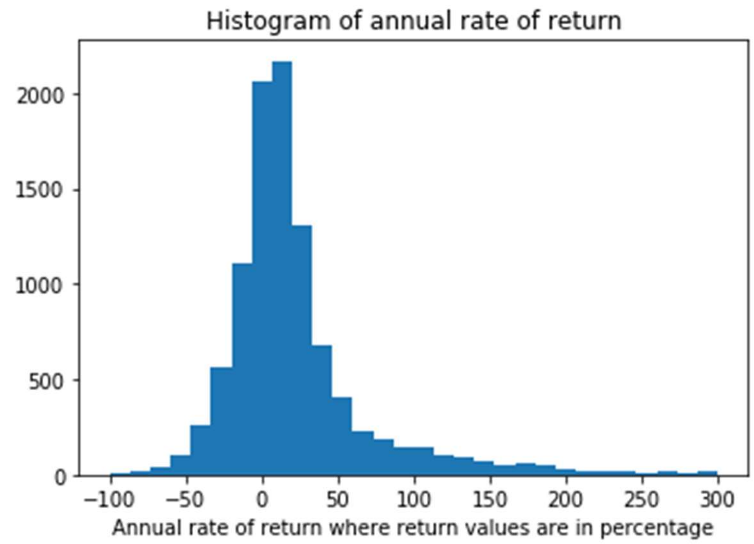


Figure 7

Mean	29.24%
Standard deviation	106%
Minimum	-99%
25%	-3.22%
Median	11.71%
75%	32.62%
Max	3926%

Figure 8

The mean value of 29% is abnormally high since an average investor, investing in a randomly selected sample of companies, will never be able to achieve an annual return of 29%. This mean value is due to the presence of large positive outliers. The median value of 11% is more reasonable since an average investor will be able to achieve this annual return when investing in a randomly selected group of companies.

Figure 9, figure 11 and figure 12 are annual rate of return histograms split based on time periods. The three time periods that were used are 1993-2001, 2002-2009 and 2010-2017. These time periods were chosen since they are of nearly equal length and because they signify three different eras in the evolution of US stock markets.

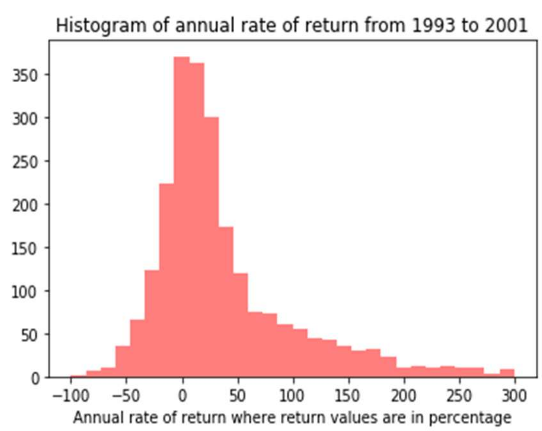


Figure 9

	1993-2001	**2002-2009**	**2010-2017**
Mean	60.05%	22.17%	14.89%
Standard deviation	184.04%	78.15%	38.13%
Minimum	-90.75%	-94.91%	-99.99%
25%	-0.87%	-9.69%	-0.39%
Median	20.29%	7.79%	9.93%
75%	64.29%	30.56%	22.63%
Maximum	3926%	2544%	858%

Figure 10

Figure 10 describes the findings of the three histograms. The mean of annual rate of return has been falling consistently over time. This also corresponds to a drop in standard deviation or market volatility. The median annual rate of return in the 90`s was substantially higher compared to the 2010-2017 time period. From the histogram, we see that the returns for the 1993-2001 time period has a short peak and the returns are heavily scattered around the mean of the distribution. For the 2002-2009 time period, the peak is taller and the scattering of returns are smaller compared to the earlier period. The 2010-2017 time period has the tallest peak and the lowest scattering of returns around the mean.

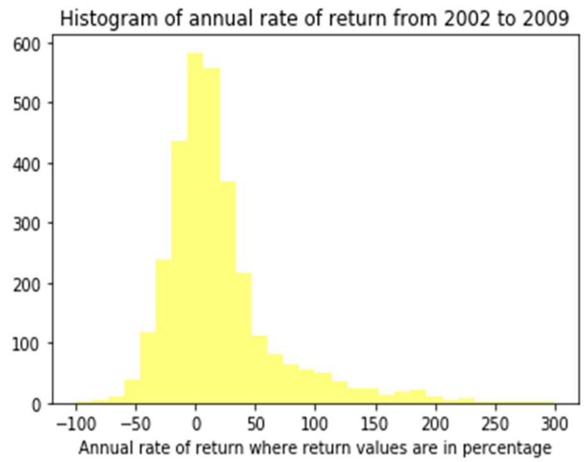


Figure 11

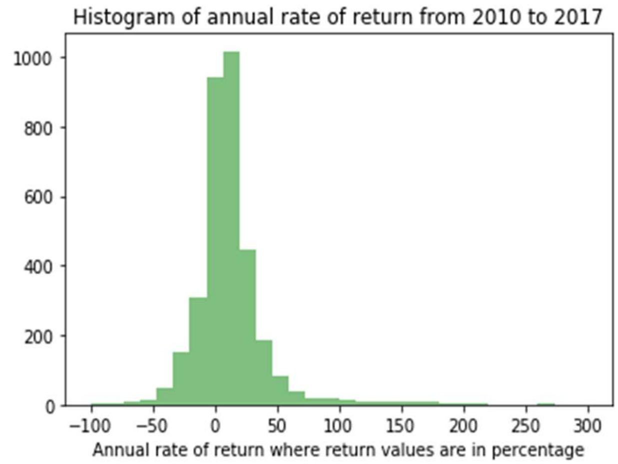


Figure 12

Based on figure 13, the < $ 1 billion market capitalization companies have the highest median and standard deviation of annual rate of return. While the highest median shows that small companies have the highest expected growth rates, the highest standard deviation or volatility shows that investors perceive investment in small companies as riskier than larger companies. The mean annual rate of return constantly decreases as the size of the company increases, except for > $ 30 billion market capitalization companies, which have a higher mean annual rate of return compared to $ 10-30 billion market capitalization companies. But the trend is maintained in the median annual rate of return metrics, which are inversely correlated with size of the companies.

	< $ 1 billion	**$ 1 – 10 billion**	**$ 10-30 billion**	**>$ 30 billion**
Mean	43.04%	18.98%	8.99%	11.66%
Standard deviation	108.71%	45.08%	23.16%	43.97%
Minimum	-90.03%	-76.36%	-55.7%	-82.57%
Median	17.69%	10.51%	7.08%	6.18%
Maximum	1988%	750%	352%	655%

Figure 13

Figure 14 shows the summary statistics of annual rates of return for companies in different sectors. The utilities sector had the lowest median annual rate of return, while financials, healthcare and industrial goods had the highest median annual rate of returns. The highest volatility in annual rate of return was seen in the technology sector, followed by the services sector.

	Basic materials	**Consumer goods**	**Financials**	**Healthcare**	**Industrial goods**	**Services**	**Technology**	**Utilities**
Mean	23.62%	25.75%	22.36%	35.4%	22.88%	27.94%	47.11%	18.31%
Standard deviation	58.84%	67.6%	50.81%	99.96%	46.39%	102.9%	194.1%	86.44%
Minimum	-94.6%	-59.97%	-94.91%	-70.24%	-71.78%	-87.6%	-99.9%	-83.5%

Median	11.04%	11.08%	13.14%	13.82%	13.62%	11.35%	11.83%	9.58%
Maximum	774%	1246%	533%	1988%	358%	2544%	3926%	1594%

Figure 14

ANNUAL RATE OF RETURNS VERSUS STANDARD DEVIATION OF RETURNS

Figure 15 shows that the mean annual rate of returns increases with increasing standard deviation of returns. This implies that stocks that have higher annual rate of returns also have higher standard deviation of returns. The curve is very similar to the Markowitz efficiency frontier. The Markowitz efficiency frontier is a right facing parabola drawn between standard deviation of stock returns and actual returns. According to Markowitz theorem, the points that fall below the vertex or the turning point of the parabola are inefficient. Also, points that are far off from the frontier are inefficient. Each point corresponds to a stock and the points below the vertex have higher risk for a lower return compared to the points above. This is true also for points away from the frontier. The points above the vertex are considered efficient compared to their peers below the vertex since they generate better returns for the same level of risk. The relationship between standard deviation of returns and rate of returns is quadratic and non-linear. Our plot confirms this.

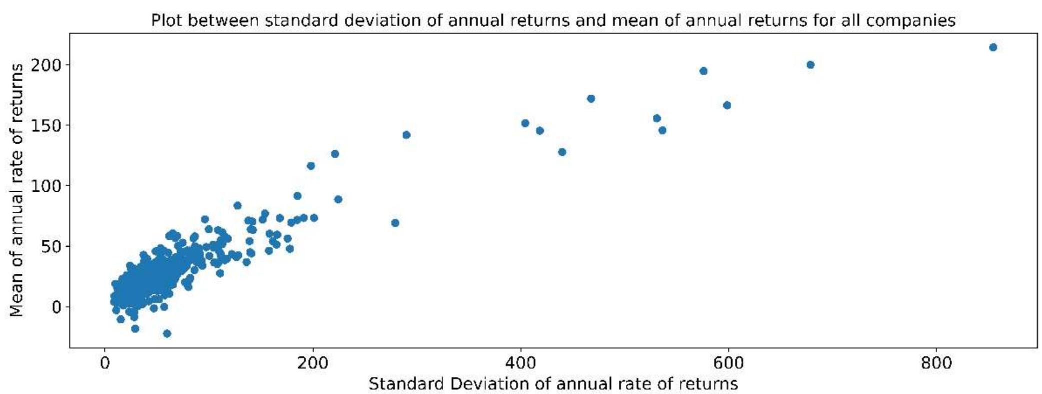


Figure 15

Finally, figure 15 does not have any data points below the vertex. This means that all the 739 companies in our dataset are efficient, according to the Markowitz theorem. The very likely reason for this phenomenon is survivorship bias. The companies data downloaded from StockPup.com is probably composed mostly of successful companies whose share prices do well. An unbiased dataset composed of both successful and unsuccessful companies will likely produce a more complete version of the Markowitz frontier.

STATISTICAL INFERENCE

Tests of statistical inference were performed on some of the results that were received from the exploratory data analysis part of the capstone project. Five hypothesis were developed

1) *The differences between the sample means of the sectoral P/E ratio of different sectors is not significant.*
2) *The differences between the sample means of the sectoral earnings to long term debt ratio of different sectors is not significant.*

3) *The differences between the mean of annual rate of returns divided based on the 3 timeperiods is not significant.*
4) *The differences between the mean of annual rate of returns divided based on company size is not significant.*
5) *The differences between the mean of annual rate of returns divided based on sector is not significant.*

Since we are working with a timeseries dataset, the hypothesis testing cannot be done directly on the raw data. Instead, bootstrap resampling with replacement is performed on the raw data, generating new samples. Hypothesis testing is done on these new samples of data using the tukey hsd multiple comparison of means approach in sklearn. Only the results where the null hypothesis is rejected are shown below

HYPOTHESIS TEST 1		
GROUP 1	**GROUP 2**	**REJECT NULL**
Basic Materials	Financials	True
Consumer Goods	Financials	True
Financials	Healthcare	True
Financials	Industrial Goods	True
Financials	Services	True
Financials	Technology	True
Financials	Utilities	True
HYPOTHESIS TEST 3		
GROUP 1	**GROUP 2**	**REJECT NULL**
1993-2000	2001-2009	True
2010-2017	1993-2000	True
2001-2009	2010-2017	True
HYPOTHESIS TEST 4		
GROUP 1	**GROUP 2**	**REJECT NULL**
1-10 B	10-30 B	True
1-10 B	< 1 B	True
1-10 B	> 30 B	True
10-30 B	< 1 B	True
< 1 B	> 30 B	True
HYPOTHESIS TEST 5		
GROUP 1	**GROUP 2**	**REJECT NULL**
Basic Materials	Technology	True
Consumer Goods	Technology	True
Financials	Healthcare	True
Financials	Technology	True
Healthcare	Technology	True
Healthcare	Utilities	True
Industrial Goods	Technology	True
Services	Technology	True
Technology	Utilities	True

MACHINE LEARNING

For the machine learning part of the project, the goal was to understand the factors that drive stock price growth. This is done by predicting a stock`s market capitalization and its annual rate of return. While predicting market capitalization can be done using a linear regression, the annual rate of return is subtracted by the S&P500 return that creates a new variable called AlphaReturn. If the annual rate of return is higher than S&P500 return, AlphaReturn equals 1 else it equals 0. AlphaReturn is modelled using logistic regression, random forest and support vector machines.

ALGORITHM	**DEPENDENT VARIABLE**	**INDEPENDENT VARIABLE**	**MODEL PARAMETERS**	**MODEL PERFORMANCE AND VALIDATION**
OLS Linear Regression	Market Capitalization	Earnings Free cash flow	R-squared = 0.382 Earnings [Co-Efficient, p-value] = [27.0, 0.0] Free cash flow [Co-Efficient, p-value] = [24.0, 0.0]	Scatter plot between actual and predicted values of market capitalization shows poor correlation Plot of residuals versus fitted values has a very unnatural shape Normal Q-Q plot shows that residuals are not normally distributed
Logistic Regression	AlphaReturn (equals 1 if Stock return > S&P500 return. Else 0)	GDP Inflation Price to book ratio	Model accuracy = 0.538 GDP [Co-Efficient, p-value] = [32.0, 0.0] Inflation [Co-Efficient, p-value] = [7.5, 0.047] Price to book ratio [Co-Efficient, p-value] = [0.017, 0.49]	K-fold cross validation gives model accuracy of 0.53 Grid search cross validation at C=100 gives model accuracy of 0.57 Positive values of co-efficients make intuitive sense and show that model can be used to predict alpha return
Random Forest	AlphaReturn	GDP Inflation Price to book ratio	Model accuracy = 0.6 Feature Importance (top 3) =	Useful for identifying features with highest importance. These features can in-turn

			[(GDP, 0.075), (Inflation, 0.064), (Price to book ratio, 0.06)] All other features have importance between 0.04 to 0.051	be used in logistic regression.
Support Vector Machine	AlphaReturn	GDP Inflation Price to book ratio	Model accuracy = 0.53	Model accuracy not better than logistic regression

CONCLUSIONS AND FUTURE WORK

1) Standalone fundamental accounting data is inadequate for modeling market capitalization. Fundamental data misses out on many other sources of information that are unstructured in nature. Quality of management, future earnings, intangible value are some other features that maybe useful in modeling market capitalization.
2) Fundamental accounting data suffers from multi-collinearity. Since accounting data is governed by the accounting equation, any change in assets is reflected on liabilities, owners equity and vice versa. Linear and logistic regressions are unable to model correlated features.
3) Random forest is best suited for modelling stock price returns and performs far better than logistic regression. This shows that more complex algorithms like random forest and neural networks are needed for this type of modelling. Simple, high bias algorithms like linear regression work poorly on fundamental data.
4) Future earnings and future free cash flows are some of the biggest drivers of market capitalization. Fundamental data does not include these projected figures. Models that incorporate these features will perform substantially better.
5) Finding a way to incorporate unstructured data and technical price movement data is important for building a model that can predict stock price returns. Features related to unstructured data are very complex, difficult to incorporate but are likely to predict stock price returns better than fundamental accounting data.

APPENDIX

https://anaconda.org/arunbharadwaj2009/capstone_dsintensive_final/notebook