Technological University Dublin
ARROW@TU Dublin

Dissertations School of Computing

2018

Supervised Learning Models to Predict Stock Direction Within Different Sectors in a Bull and Bear Market

Tiffany Razy
arrow.admin@tudublin.ie

Follow this and additional works at: https://arrow.tudublin.ie/scschcomdis

Part of the Computer Engineering Commons

Recommended Citation

Razy, Tiffany (2018). *Stock direction within different sectors in a bull and bear market.* Masters dissertation, DIT, 2018.

This Dissertation is brought to you for free and open access by the School of Computing at ARROW@TU Dublin. It has been accepted for inclusion in Dissertations by an authorized administrator of ARROW@TU Dublin. For more information, please contact yvonne.desmond@tudublin.ie, arrow.admin@tudublin.ie, brian.widdis@tudublin.ie.

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License

Supervised Learning Models to Predict Stock Direction Within Different Sectors in a Bull and Bear Market

Tiffany Razy

A dissertation submitted in partial fulfilment of the requirements of
Dublin Institute of Technology for the degree of
M.Sc. in Computing (Data Analytics)

January 2018

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Knowledge Management), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the test of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: Tiffany Sara Razy

Date: 03 January 2018

ABSTRACT

Forecasting stock market price movement is a well researched and an alluring topic within the machine learning and financial realm. Supervised machine learning algorithms such as Random Forest (RF) and Support Vector Machines (SVM) have been used independently to gain insight on the market. With such volatility in the market the scope of this study will utilized the RF and SVM in a very volatility market to determine if these models will perform at a high level or outperform each other in both markets. This relative study is performed on 16 stocks in 4 different sectors over the bear market ”housing crash” of 2008 . The model utilized technical indicators as the respective parameters to assist in predicting the stock price movement when determining the performance of each model. Despite the No Free Lunch Theorem stating one model can not out perform another model, the study displayed higher accuracy for the RF model. Each model was evaluated using the confusion metrics to calculate the precision, recall, and F1 score.

Key words: *Support Vector Machines, Random Forest, Bull and Bear Market, Stock Price Forecasting, Technical Indicators*

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my supervisor Andre Rios for his assistance, patience, guidance, and constant support through the course of this dissertation.

I'd like to thank Sarah Jane Delany, Luca Longo, and Matthew Morey for their unconditional support, advice, and patience.

I also appreciate all my DIT lectures for assisting me in gaining the knowledge to be able to complete this dissertation.

I would like to thank my parents, siblings, and boyfriend for encouraging me to continue working hard, providing unconditional love, and for supporting me.

I would like to immensely thank Angie Hayes, and Roxanne Chenette for proofreading my proposal, dissertation, and for their enormous amount of kindness.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

List of Acronyms

ADX – Average Directional Index

BB- Bollinger Band

CCI – Commodity Channel Index

CMF – Chaikin Money Flow

EMA – Exponential Smoothing Average

MA – Moving Average

MF – Money Flow

RF – Random Forest

RSI – Relative Strenght Index

SMA – Smoothing Moving Average

S&P – Standard & Poor

SVM – Support Vector Machine

1 INTRODUCTION

1.1 Background

Forecasting the stock market has proven to be a highly researched topic and incredibly challenging problem. Predicting the stock market price, direction or next strategy has proven to be increasingly profitable therefore, numerous traders, economists, financial industries and academics form and create their own breakthrough strategy for this forecast. However, they do not gain the profit range that keeps their approach ahead of the market. Given its complexity, the attraction of minimizing risk and increasing odds has brought much attention towards finding the best strategy. The reason for this is that financial markets are highly influenced by noisy, non-linear, and non-stationary data (Kim and Han, 2000; Handa, Hota, Tandan, 2015; Khaidem, Saha, Dey, 2016).

The world has witnessed major economic crashes in the market during the following two periods; the great depression and the housing market crash. These major occurrences in history caused the stock market to fall at its lowest triggering the trend of all stock prices to fall, thus causing bear markets to erupt in the stock market. [1]The bear and bull market names emerged when the markets followed the way in which the animals attack their target. The bull market follows that of the attacking bull striking upward with its horns, while the bear market follows the downward swipe of the attacking bear. Though the market collapsing caused stock prices to crumble, certain sectors were in desperate need of assistance. This resulted in a desire for sector trends to be analysed and predicted to gain insight for traders to buy or sell prior to another market crash.

1.2 Research Project/problem

The stock market has been an ideal place to build and maximize financial security for the common businessman and investment companies. Strategies have been incorporated into different analytical techniques in order to further the ability to forecast and be ahead of the stock market. Though some strategists such as Stan Weinstein have profited off the market following tactics that focus on a more technical strategy by using technical

[1] http://www.businessinsider.com/why-we-call-it-bull-bear-market-2014-7?IR=T

trading strategies to understand when to buy and sell in a bull and bear market, many other traders and strategists focus on fundamental market strategies (Weinstein, 1988). In order to profit from the stock market, individuals analyse relationships regarding economic occurrences and company information. As mergers evolve, financial crises occur and currency rates change researchers have tried to find relationships within the data available to build analytical models to strategically predict the stock price index (Kim and Han, 2000).

As the market keeps evolving, research continues to expand on executing machine learning models such as support vector machines and neural networks to maximize the forecasting of the stock direction and price (Ou and Wang, 2009). However, due to the diverse trading techniques and ever evolving economy, the focus has shifted towards determining minimal error as well (Ou and Wang, 2009; Kim, 2003;).

The overall research regarding the use of machine learning algorithms to predict the stock price and direction with a higher accuracy and lower error based rate continues to be ongoing as changes continue to take place within the stock market. Though many strategists have focused on incorporating fundamental factors such as price per earnings, and technical features such as Bollinger Band into their models, few have taken the initiative of integrating the "Bull and Bear" market period when incorporating a variety of machine learning models (Cheng, Chen, and Wei, 2010; Khedkar and Argiddi, 2013). By studying these trends one would be able to gain a better understanding and gain an upper hand when investing. As expressed in greater depth in the next chapter, the general research focus in this area has been into feature exploration such as technical indicators and fundamental analysis and the efficient-market hypothesis (Fama,1970). An underexploited area of research can be seen where selecting specific technical indicators used for predicating stock trends in a bear and bull market is concerned (Kannan, Seka, Sathik, Arumugam, 2010). A vast amount of research can be presented on the stock market using machine learning algorithms. However, minimal research can be found on specific ensemble models, or more specifically which model performs better; an ensemble of decision trees or support vector machines?

This study will predict the stock price direction on a set of stock in various sectors to determine the best performing sector when utilizing a set of technical indicators. To obtain this goal, the following research question is presented:

Can a Support Vector Machine predict more accurately than a Random Forest in a bear and bull market when forecasting stock price direction in specific sectors utilizing a set of five technical indicators?

1.3 Research Objectives

Primarily, the aim of this research is to determine the forecasting capabilities of a support vector machine and whether it will perform better than a random forest learning model in forecasting the stock price direction in an upward market (bull) and a downward market (bear). Though to keep in mind with the relevant No Free Lunch Theorem stating no one model can be "better" than another due to the various parameters each model provides (Wolpert and Macready, 1997), this goal will analyse the performance of these models against that of the stock data as research has shown independently that both have performed well in relation to other machine learning models. This study consists of a variety of goals; utilizing five technical indicators to assist in forecasting the stock price direction and also determining the performance of each sector in the bull and bear market by only using historic price data.

The use of the technical indicators will be to assist in forecasting the stock direction movement and for presenting chart based visualizations. Prior to implementing the feature variables, a calculation of the return value will be formulated using a common technical indicator; the exponential moving average. As explored by Khaidem, Saha, and Dey (2016) the process of using this the exponential moving average is to smooth out the noise in the stock data which will be further discussed in the next chapter. Once modelled, the support vector machine and random forest models will be evaluated using the confusion matrix to identify accuracy, precision, recall and F1. This will provide the results of the null hypothesis. The null hypothesis in this study is: *the random forest model will perform with a higher accuracy when utilizing specific technical indicators in both a bull market against a support vector machine.*

To gain insight into the “better” performing model when used against the stock data the following tasks will be implemented:

- Study existing literature on stock market trends, stock market trading behavior, market trends such as those of bull and bear, technical indicators, and machine learning models to gain an in depth analysis of the research and tools used by academics and traders alike.
- Analyzing the data to split into train and test samples.
- Smoothing the data using the EMA technical indicator.
- Calculate the return price in order to determine the difference in direction from the initial close price.
- Perform the feature selection and analysis of the overall data to clean and prepare it for modeling.
- Build the models to implement the data into the support vector machine and random forest models.
- Evaluate the model performance by utilizing the confusion matrix and compare the models using accuracy, precision, recall and F1.

1.4 Research Methodologies

This research is a collective set of data that is measurable by mathematical expressions and quantitative methods. The mathematical models will consist of multiple machine learning algorithms which will test the best accuracy and evaluation when forecasting the stock price and direction. The data comprises of stocks acquired from Yahoo Finance which is a commonly used source of financial knowledge. This analysis comprises of empirical research as the different models will be testing the feasibility towards gaining a direct resolution to the testable prediction. Due to the nature of this study, a substantial portion of the research will be focused on observation and literature review. This study will form an inductive approach of reasoning building on both analytical observation and literature review.

1.5 Scope and Limitations

The scope of this study will cover two different periods in history to incorporate a "bear" market and a "bull" market. The time periods will cover that of the housing crisis that happened from 2007 to 2009. To gain an understanding of the performance of technical indicators, stocks, and machine learning models, this time period was selected as it provides a major dip in the stock market as the market crashed resulting in that of the bear market and was in recession stage when coming into the bull market. This will provide insight into the best performing sector to assist in learning if the models could detect any trends when only utilizing technical indicators. This will also provide insight into whether traders can accurately use only historic data in order to see if all sectors perform badly when the stock market crashes or if one sector stays levelled throughout the market crash. Thus, comparing the bear and bull market predictions will provide insight into the effects of the market trend on specific sectors as well as the effects on the machine learning algorithms. These insights could be useful in future where a trend is starting to show the market falling into a bear market as the investor or traders sell and purchase based on the performance of sectors with little effect.

The bear market will be covered from January 01, 2007 to March 31, 2009. Although initially the historic crash considered the bear or housing crisis collapsed the stock market from October 9,2007 to March 9, 2009 [2] (Lim, 2014), due to the short window of time, an additional few months were added to avoid a lack of non-applicable results when incorporating certain feature selections. The second period of time will be from April 1, 2009 until June 30, 2011 as the stock market changes into a bull trend. Within these time periods, the bear market will be split between test data and training data and the bull market will be split into test data and training data.

In addition to the time period, the experiment will cover four different sectors from the S&P 500. These sectors will have four different stocks within each sector. These stocks will be traded on the New York Stock Exchange and Nasdaq. The sectors are broken down into Technology, Financial, Healthcare, and Industrial. A more in depth break

[2] http://time.com/money/3482841/bear-market-anniversary/

down of the stocks and sectors will be incorporated in chapter three; “Design and Methodologies”.

1.6 Document Outline

Within this study a discussion of the following will occur:

- Chapter two – “Literature Review” will go deeper into the discussion of past research compiled of different algorithms and approaches towards the predication of stock market price and direction. This chapter will outline the various approaches commonly used by traders and researchers in order to ensure a high level of accuracy when forecasting. This chapter will also include an analysis of the reasoning behind the historic factors in today’s methodology regarding stock market prediction while also integrating the contradictive approach towards forecasting prediction price.

- Chapter three – “Design and Methodology” will outline the method breakdown of the experiment. Within this section of the study, the design portion will be broken down into sections similar to the CRISP-DM to provide an in depth understanding of the structure of the business ideology behind the study and will also include the collection and preparation of the data. Within this section, a special focus will be held on the feature selection that will assist in transporting the data into qualified inputs and clarifying its strengths and limitations. The remaining sections within this chapter will refine each machine learning model used.

- Chapter four – “Implementation and Results” provides a breakdown of each models implementation and the results of the four stages of the experiment provided. The four stages will be running four different models evaluating the accuracy and limitations of each.

- Chapter five – “Analysis, and Evaluation” will be comprised of the four stages of the experiment’s results along with the relevant research and analysis that has been provided in the Literature Review. This chapter will encompass the analytical aspect of the results while justifying the outcome of the hypothesis based on the undertaken experiment.

- Chapter six – "Conclusion" will be an overview of the completed study incorporating the workflow process of the entire experiment focusing on the limitations and scope of the stock market and forecasting research. This summary will reiterate the initial research question discussed in chapter one and provide areas for future work and recommendations.

2 LITERATURE REVIEW

With such a vast amount of literature provided this chapter specifically addresses two main parts of the overall study. These parts are focused on the financial background of the stock market and the machine learning algorithms that will be implemented. This chapter will be split into multiple sections and sub-sections in order to guide through the assortment of approaches towards forecasting stock market prices. As seen in Figure 2.1 the sections will consist of; Trading Strategies, Market Types, Machine Learning Algorithms, and Evaluation. While reviewing these sections the goal is to analyse relevant work that has been researched within various financial markets utilizing magnitudes of machine learning algorithms to gain a solid understanding of the complexity that follows predicting stock market prices. Due to the high complexity and influential factors that incorporate the stock market a considerable amount of research has concluded an array of hypotheses and theories. The research has shown a multitude of factors which encompass the basis of the financial market and the workings of various concepts behind the market. While many researchers have formulated intricate models utilizing technical indicators to determine a pattern to forecast the stock price, others have focused on researching and testing the Efficient Market Hypothesis (Arratia 2014; Fama 1970). The Efficient Market Hypothesis is built on the basis that stock prices are built on the relevant information present at the time of making an investment (Fama 1970; Nikfarjam, Muthaiyah, Emadzadeh 2010; Khaidem, Saha, Dey 2016).

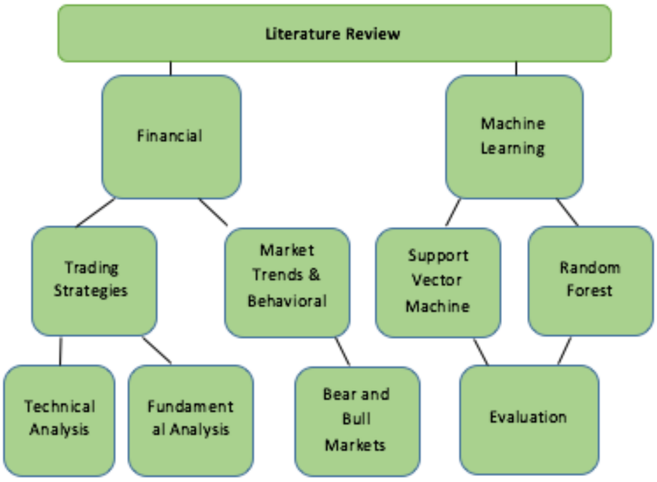


Figure 2.1 illustrates the layout of the following literature review

2.1 Trading Strategies

Trading strategies are set by individual traders, financial institutions, and academics. These strategies are complex and personalized based on the person or institution that is trading (Ou and Wang, 2009). With a broad variety of trading approaches used in the investment industry, the two best desired methods of investing are Technical and Fundamental analysis (Arratia, 2014; Atsalakis and Valavanis, 2009). Although technical indicators have proven profitable in research, a notable debate continues among researchers as to whether technical indicators are enough to forecast trends in the market. Technical strategists believe there is a pattern available in past prices (Austin, Bates, Dempster, Leemans, Williams, 2004).

2.1.1 Technical Analysis

Though technical analysis is one of the major investing strategies it is also a common approach used toward forecasting stock prices and direction. The technical strategy is primarily based on mathematical factors that follow a stock (Cheng, Che, Wei, 2010). The mathematical analysis is initially based on the historical price of the stock. The historical price stock would consist of open, close, high, low, volume, and adjust.

Although technical analysis is based on historical price, researchers such as Cheng et al. (2010) express that investor psychology formulates the historical price based on investor's response and actions toward price movements in the market, and so determine that the high fluctuation of the market follows human expectations rather than following a linear formation Cheng et al. (2010). To determine the human behaviour which influences trading stock and the trend of the stock, a commonly found charting type used by technical strategists was created by Japanese Munehisa Homma in the 1700s. This is called the candlestick chart and is used to analyse trends in the market (Prado, Ferneda, Morais, Luiz, Mastura, 2013). This charting process evolved into a charting technique that is used to combine trading rules to that of the market trend. The use of this charting technique in this study is to provide a visual understanding of the stock components of both markets to distinguish a better understanding of the initial data.

Historically, the use of technical analysis was said to be the most inefficient and highly representative approach to predicating the stock market trend due to the psychological factors behind the investors emotions and rationality towards various stocks (Neely, Weller, Dittmar, 1997). This method goes against the Efficient Market Hypothesis as this approach is said to follow economic factors and company related factors that are represented in the current price (Fama 1970; Atsalakis and Valavanis, 2009) as opposed to the use of historical values to build logically articulated formulas to determine conditions in the price trends. With the vast variety of technical trading techniques provided in the field, researchers and traders use a combination of them for the purposes of forecasting price movement and trend. Kanna et al. (2010) studied the use of five technical analysis trading techniques to find patterns in past trends. This study consisted of the typical price, Chaikin Money Flow indicator, Stochastic Momentum index, Relative Strength index and Bollinger Band. While their research displayed profitable insight on the Bollinger Band and Stochastic Momentum, other academics found striking results using the Relative Strength index, Stochastic Oscillator, Williams %, Moving Average, Convergence, Divergence, Price Rate of Change, and on Balance Volume.

Taking into account the hundreds of technical indicators that academics and traders have used, this study will cover a range of different technical indicators that have provided profitability in various studies. These indicators are; Commodity Channel Index, Relative Strength Index, Bollinger Band, Chaikin Money Flow, Average Directional Index (Handa, Hota, Tandan, 2015; Khaidem, Saha, Dey, 2016; Kanna, Sekar, Sathik, Arumugan, 2010, Kim, 2003). These technical indicators have been seen to positively aid in analysing and forecasting the stock price movement. Using a combination of these indicators has been unexplored when looking into the period of a major economic downfall such as that of the housing crisis. The below table will provide a visualization of the combined technical indicators that will be explored in this study.

Technical Indicators		
Name	Definition	Formula
Chaikin Money Flow	Illustrates the money flow within the stock to see the volume of purchases and sales; It assists in indicating a bullish/ bearish market; Utilizing a basis of 20 days for calculations	$CMF\ (t) = (t\left(\left((Close - Low) - \frac{High - Close}{High}\right)\right) * Volume)/Volume))$ $t\ =\ number\ of\ days$
Average Directional Index	Measurement of the trend strength; Utilizing a basis of 14 days for calculation Strong strength = higher value Weak Strength = lower value	$ADX = SMA[\left(\frac{+DI - (-DI)}{+DI + (-DI)}\right)]$ $DI += High - High(-t)$ $DI -= Low(-t) - Low$ $-t\ =\ day\ before$
Bollinger Band	Measures the unpredictable trend for price movement P(t) = Typical Price (Another Technical indicator) ;Utilizing a basis of 20 days for calculation	$Middle\ Band\ =$ $20\ day\ of\ the\ Simple\ Moving\ Average$ $Lower\ Band = SMA + (\partial \times 2)$ $Upper\ Band\ =\ SMA - (\partial x2)$ $(\partial)\ =\ the\ 20\ day\ of\ the\ close\ price$
Commodity Channel Index	Adaptable indicator for trends in the market or major crashes P(t) = Typical Price (Another Technical indicator)	$P(t)\ =\ High\ +\ Low\ +\ Close/3$ $CCI\ =\ (1/0.015)x((P(t) - \frac{SMA(P(t))}{(\partial(P(t)))}$
Relative Strength Index	Momentum indicator to determine if a stock is oversold or overbought; strength indicator of trend; Utilizing a basis of 14 days for calculation	$RS = \frac{Smoothing\ Moving\ Average(U,n)}{Smoothing\ Moving\ Average\ (D,n)}$ $RSI = 100 - (\frac{100}{(1 + RS)})$

Table 2. 1 [3]displays the technical indicators implemented in the study. The table provides the mathematical formulas behind each of the technical indicators.

[3] http://stockcharts.com/school/doku.php?id=chart_school:market_indicators

2.1.2 Fundamental Analysis

Although this study uses only the technical approach to understanding the stock movement in a bull and bear market, it is important to review fundamental as it has been argued that it can provide insight into market trends which will be a major aspect of this experiment.

As technical analysis formulates on historical stock prices and mathematical formation, fundamental analysis follows a vastly different concept. Macroeconomic data is the basis for fundamental analysis (Atsalakis and Valavanis, 2009). Referenced in the previous sections of this chapter, the concept of the EMH was said to follow a fundamental analysis methodology toward the relevant change in price.

This approach states that any economical change will be fully reflected in the price of a stock. The ideology behind fundamental analysis is said to reflect macro-economical changes within a sector in the price. The past prices of any stock would refrain from holding any pattern as the historical price would not display current changes in the market or sector. The current price would be the only imitative of the change in the market (Enke and Thawornwong, 2005).

2.2 Market Trends and Behavioural Investment

With the complexity of the market and non-linearity, research was expanded into the area of psychology and looked into examining the combination of market trends and investor behaviour. The psychological behaviour of following the market trend is seen as “herding” or “crowd effects” (Cont and Bouchaud, 2000; Bikhchandani and Sharma, 2001; Prechter. Jr , 2001). Herding behaviour, a psychological behaviour that consists of following another person or a group of peoples actions (Clement and Tse 2005, Cont and Bouchaud, 2000; Bikhchandani and Sharma, 2001; Prechter. Jr , 2001). Herding has been seen to have a volatile effect on the market trend when investors routinely follow each other’s decisions in the hopes of increasing revenue or to avoid any loss of revenue. Herding is considered to follow hand in hand with fundamental analysis as an investor can make a quick decision based on an economical change causing a sudden change in the purchase or selling of a stock (Bikhchandani and Sharma, 2001).

Although herding has been found in the changes of the stock price, a reflection of herding in the peaks of market stage such as that of a bull or bear market are still being researched (Christie and Huang,1995). Shiller (1987) studied the correlation between investors' behaviour and the stock market crash which found supporting evidence that the influx and fear of an investor is present when analysing market trends such as the 1987 stock market crash. Analysing the market trends with an investigation into behavioural investing could be a further area of research. As investors continue to buy and sell in a "herd" like behaviour, the market sees a reflection of sudden movement. These movements are seen a bear and bull market. The following sections will elaborate on the meaning of a bull and bear market trend.

2.2.1 Bull and Bear Market

A bull market is when a set of stock prices are increased and continue to increase on the market. This trend is when investors believe that the market will continue to rise so they keep purchasing which increases the price of stock thus providing a trend in the upward direction of the stock market. The market displays a "bullish" type market when returns of the stock are at a high and little volatility is displayed (Maheu and McCurdy, 2000). A bear market is the opposite to that of a bull market. A bear market is the decreased and continual downfall of the stock price. This is when an investor starts increasing the selling of stock and volume of stock to avoid the loss of value. A bear market can be seen when there is a sense of irrationality around stock selling due to fears from investors resulting in a drastic decrease in stock price (Day and Huang,1990).

2.3 Machine Learning

Machine learning is a type of strategy used to distinguish patterns in data. There are multiple forms of machine learning such as supervised, unsupervised and semi-supervised algorithms. Researchers have approached the stock market price direction utilizing an assortment of supervised machine learning algorithms. Supervised algorithms are learning models built based on a relationship between dependent and independent data (Kelleher, Mac Namee, D'Arcy, 2015; Carbonell, Michalski, Mitchell, 1983). Recently machine learning has become a widely used component in many sectors in industry across the world. A considerable amount of studies have been completed to forecast stocks using a variety of machine learning and data mining techniques to gain insight into the complexity of the non-linear trends (Francis and Tay, 2003; Hsieh, Yang, Wu, 2006; Huagn, Nakamori, Wang, 2005; Khaidem, Saha, Dey, 2016; Ou and Wang, 2009; Kim, 2003; Enke, Thawornwong, 2005). The use of machine learning has provided the ability to build models using separate areas of focus such as those of technical analysis which will be further discussed in this subsection to other forms of models such as textual analysis following a fundamental approach. Researchers found a variety of different supervised machine learning algorithms such as Artificial Neural Networks, Support Vector Machines, Decision Trees, Logistic Regression, Naïve Bayesian, and Random Forest (Francis and Tay, 2003; Hsieh, Yang, Wu, 2006; Huagn, Nakamori, Wang, 2005; Khaidem, Saha, Dey, 2016; Ou and Wang, 2009; Kim, 2003; Enke, Thawornwong, 2005) which could aid in forecasting stock price and price direction.

Though many academics have performed the desired task of predicating stock price return, stock price movement and stock trends using a variety of models, the No Free Lunch theorem states that though one researcher found one model to work the "best," this specified model may not perform better than another particular model when analysing a similar but different question (Wolpert, Macready, 1997). This theory asserts that each model has their own strengths and limitations with a variety of data. Thus, it is important to ensure that the models are built using parameters that are appropriate to their predicting abilities when applied to dependent and independent variables.

2.3.1 Support Vector Machines

The support vector machine is an error-based learning model used in a wide collection of literature when forecasting stock price and price movement (Ou, Wang, 2009; Huan, Nakamori, Wang, 2005; Kim, 2003; Hsu, Hsieh, Chih, Hsu, 2009; Cao, Tay, 2003; Pai, Lin, 2005). Huang et al. (2005) explored the support vector machines capabilities of predicting the price movement against other machine learning algorithms and found that the SVM forecasted a higher accuracy due to the minimal "structural risk" rather than the minimal "empirical risk". This leaves the SVM with less capabilities of over-fitting in comparison to the Elman Backpropagation Neural Network or Linear Discriminant Analysis. With a vast set of parameters available for SVM model training, a process used by Kim (2003) displays the use of polynomial and radial basis kernels. If improperly used, the SVM parameters such as the kernel and C can cause the predictive model to over-fit or under-fit. A focus across literature for classification SVM models has been the consistent use of the radial kernel based learning strategy. This is a parameter for the kernel function for the training stage of the algorithm (Huang, Yoshiteru, Nakamori, Wang, 2005; Ou, Wang, 2009; Pai, Lin, 2005). Another focus in the literature was set on the C and gamma parameters of the SVM model. These parameters ranged in value size based on the dataset acquired for the overall goal of forecasting the stock price or the stock price direction.

2.3.2 Random Forest

The random forest model is an ensemble model made up of multiple decision trees. The random forest model is an information based learning model built based on regression data or classification. A combination of academics have used either decision trees, or a random forest ensemble model for the purposes of classification. Ou and Wang (2009) utilized a type of decision tree called tree based classification as this form of decision tree is optimized by using various objects in the dependent variable from measuring on multiple independent variables. Though this type of model optimized well in Ou and Wang's research, other academics have found that utilizing the ensemble random forest provided a higher set of performance results when forecasting stock movement (Khaidem, Saha, Dey, 2016). The random forest model optimizes when building the decision trees and splitting at the mode of each output for each single tree. Khaidem et al. (2016) build their random forest using the splitting option of the Gini impurity.

2.4 Evaluation

The evaluation stage of the model will provide an understanding of the overall accuracy levels of the model. The evaluation stage will provide a level of efficiency and specify a range of results that are beneficial for future works on the model against the tested data. Similar to that of the machine learning models, the range of evaluation methods used to gauge the performance of each model execution has been seen as diverse in the scope of the literature gathered. Though few researchers provide their initial evaluation process and proceed to provide results, the following popular classification evaluation metric; the confusion matrix, will be utilized.

2.4.1 Confusion Matrix

A commonly used evaluation method in supervised machine learning for classification models is the confusion matrix. This matrix is used on a predictive model against a set of data taken from the dataset specifically for testing purposes. This evaluation process provides a large range of performance methods to emphasize the performance of the model. The following table displays the range of measures used to calculate the performance.

Confusion Matrix Table			
Target	Prediction		
		Positive	*Negative*
	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Table 2. 2 illustrates the confusion matrix table providing a visualization on how the measures work.

Each measure; true positive, false negative, false positive, and true negative are measurements that calculate the frequency of each of the outcomes available for the prediction of the target feature. This matrix measurement provides two different ways the predictive model can analyse the correct and incorrect predictions from the test set. The below table explains the measure meanings for each of the outcomes possible.

Measure Meanings	
Measure Name	Meaning
True Positive (TP)	Reflects the target feature in the test set that was positive and forecasted as positive
True Negative (TN)	Reflects the target feature in the test set that was negative and forecasted as negative
False Positive (FP)	Reflects the target feature that was negative and forecasted as positive (Type One Error)

False Negative (FN)	Reflects the target feature that was positive and forecasted as negative (Type Two Error)

Table 2. 3 elaborates on the measure meanings for each of the correlation prediction measurements.

From the confusion matrix a set of three performance measurements are commonly used in literature (Khaidem et al. 2016). These measures are; Precision, Recall and F1 score. The following table will provide a brief description and display of their formulas:

Performance Measures		
Measure Name	Meaning	Formula
Precision	Measures the forecasted positive target levels over all of the actually predicted positive level	$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$
Recall	Is a measurement of the sensitivity of the predictive model; measures the calculated positive features forecasted and the number of positive features forecasted over all	$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$
F1	Measures the misclassification of the predicating model and combines the performance measures of Precision and Recall	$F1 = 2\ x\ \frac{(\ Precision\ x\ Recall)}{(Precision\ +\ Recall)}$

Table 2. 4 demonstrates the formulas and definitions of the performance measures utilized by the confusion matrix.

This process was taken from the performance matrix by Khaidem et al. (2016) to evaluate their ensemble model’s performance when forecasting the stock market price direction for a time series.

2.5 Overview

With a variety of variables behind stock prices, investors and academics continue to find what is considered the "best" strategy towards getting ahead of the market and forecasting patterns utilizing past stock data. This desire has provided a scope of research overlooking the varied areas in finance and in machine learning. Researchers have implemented machine learning models such as support vector machines to analyse the non-linearity of the stock price and forecast the future price movement. This scope covers a range of complexity behind what influences the stock trend, what are trading strategies and what models perform well with time series type data. Academics have found a range of algorithms that produce high level of accuracies but with a varied set of parameters and featured variables. Thus following the No Free Lunch theorem that not one model can be considered as the best model for this dataset, a study will be conducted against two models found in the literature review that proved to be notable when forecasting stock price direction, but have not been tested against a combination of technical indicators in two types of market trends. The drive behind this experiment provides another insight into two high performing predictive models in a bull and bear market using a set of technical indicators to determine if one of these models predicts more accurately than the other in forecasting the stock price direction in specific sectors. The following research question will be assessed in the following chapters of this study.

Can a Support Vector Machine predict more accurately than a Random Forest in a bear and bull market when forecasting stock price direction in specific sectors utilizing a set of five technical indicators?

The following chapter "Design and Methodology" will provide a step by step walk through of the processes of the experiment. The chapter will provide the scope of the business objective, the data understanding, data preparation, modelling, evaluation, deployment, strengths and limitations.

3 DESIGN AND METHODOLOGY

3.1 Introduction

This stage of the study will cover the design and methodology approach of the experiment. This chapter will explore the Cross Industry Standard Process for Data Mining (CRISP-DM) strategy towards handling the data associated with the research and experiment (Kelleher, Mac Namee, D'Arcy, 2015). Many researchers have incorporated this formation into their studies in order to keep a clean process in the steps and overall workflow of the experiment. It provides the ability to understand the step to step sequence as well as providing better capabilities in terms of the future reproduction of the study.

The following chapter will be broken down as seen below:

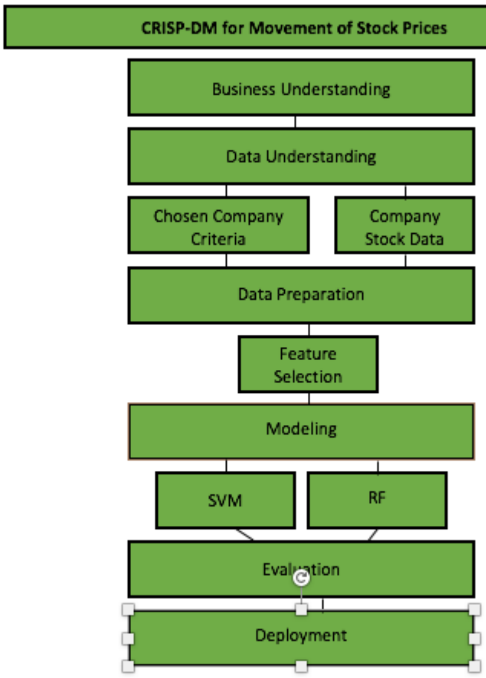


Figure 3. 1 provides an overview of the Design and Methodology chapter reviewing main aspects of the data, feature extraction, modelling, and evaluation stages. The processes are key toward forecasting the direction of the stock price in a bull and bear market.

Business Understanding: The business understanding will provide the business goal. This is what the study is initially trying to accomplish. To accomplish a successful experiment, it is crucial to have a theoretical understanding of the business problem in order to build an analytical strategy towards solving the research problem.

Data Understanding: The data understanding stage will transform the business objective into an analytical objective. This will involve gaining a thorough understanding of the stock market data that the study will focus on. Along with understanding the data in a conceptual manner, this section will cover the company selection process of the data, where the data is acquired from and the different variations of the sectors that the stock data can come from.

Data Preparation: The data preparation stage will transform the acquired data into the correct parameters for the machine learning algorithms that will be utilized in this study. This preparation stage is important as it will include the technical indicators that were previously mentioned in the literature review.

Modelling: The modelling stage integrates the data and the machine learning models. This section will discuss the chosen models and their development stage.

Evaluation: The evaluation stage is the most important stage of the CRISP-DM process and the experiment process as it will provide the accuracy of the models and the overall findings of the study.

Deployment: The deployment stage is the final stage of the CRISP-DM as it will analyse the scope and limitations of the study and also assess the necessary adjustments for the experiment that need to be implemented into real time stock trading.

Strength and Limitations: The strength and limitations section will discuss the strengths and the weaknesses of the design solution.

3.2 Business Understanding

Stock market investing, as seen in the Literature Review, has been and continues to be a highly desired way to increase gains and receive financial security. Given the vast set of investment strategies and parameters presented, the primary aim of this study is to determine stock price movement using technical indicators with specific parameters to determine if stocks in a specific sector perform better or worse in the bear or bull market. As previously mentioned, a crash in the market such as that expressed in a bear market can cause major financial indiscretions to a company and its investors. In the hope of gaining insight on stock price movement and predicting the transition of a downward trend, this study transforms into an analytical perspective that targets the price trend using machine learning models.

3.2.1 Business Goal

The business goal of this study is to assist in detecting the trends of the stock market movement in specified sectors using a set of stocks based on a return algorithm which utilizes a technical indicator for calculating the return. The additional parameters will operate a variety of appropriate technical indicators that were briefly covered in the Literature Review and will be further studied in the Data Preparation stage of this chapter. This study will determine the stock market price behaviour in each sector in the specified market type bull or bear. With this goal the study will implement the analytical aspect to operate supervised machine learning models to determine accuracy in forecasting the direction of the stock price. Providing the aim of the study the hypothesis is constructed: *The support vector machine model, trained using the stock dataset, will perform at a higher accuracy than that of the random forest in each sector in both a bull and bear market.*

To limit the volatility of the outcome of the various sectors in each market time, the parameters will be limited to any adjustments per sector. The models will fit each stock in their respective sector and will be based on one set of parameter type when running the technical indicators and the models.

3.3 Data Understanding

This section of the chapter will analyse the company selection process and company stock data. The section will analyse the processes for picking the stocks and what the stock data actually contains.

3.3.1 Company Selection

With an immeasurable amount of company stocks and trading markets to choose from, this study focuses on an index in order to pick equally equipped companies. The index chosen is known as the Standard & Poor U.S indices which contains a subdivision commonly known as S&P 500 index. [4]This index consists of the largest 500 companies capitalizing within an equal range for the market cap unlike those of the Dow Jones Industrial Average. The S&P 500 focuses on the largest companies based within their respective sector. Within these sectors the S&P 500 contains a range of the top 50 largest capitalized traded companies that are included within their 500 "constituent" companies, meaning companies that are merged or consolidated and can be a subsidiary of a larger company. Companies held within the S&P 500 index are publicly traded stocks and are focused on price index versus total return index.

3.3.2 Company Stock Criteria

For this study, the companies that were chosen from the S&P 500 index sectors must have maintained their standing on the S&P 500 list for the duration of the experiment. This ensures that they were within the criteria of market capitalization size, liquidity, domicile, eligible securities, and sector classification.

The market capitalization size is valued and maintained at a United States market price of 6.1 billion dollars or more as per referenced from the S&P U.S Indices Methodology provided from the S&P Dow Jones Indices: Index Methodology (2017). The liquidity criteria follows that the company stock shall be traded at a reflected reasonable price reflecting that of the adjusted close with the historical volume quantity. In relation to this key criteria requirement, it is essential to note that liquidity follows a specific set of technical strategy tools as it specifically reflects price with historical volume size following that of technical indicators. Domicile requires that the U.S companies follow a set of three specific rules such as filing 10-k annual reports, fixed assets, revenue plurality and their listing must be on a U.S exchange. Domicile is crucial to the study as it maintains that the companies chosen are U.S based companies and are traded within U.S based exchanges such as the New York Stock Exchange and NASDAQ. In addition,

[4]https://fred.stlouisfed.org/series/SP500

the third rule follows that of the criteria requirement - eligible securities. The eligible securities criteria ensures that the company stock is traded within a U.S exchange market. This stipulates that the company stocks are traded at the same currency value. For this study each company chosen within their respective sector are traded on U.S exchanges and are traded at the U.S dollar value. This constraint is significant in the fact that any effects to the economy are equally reflected in the value of the stock. As previously mentioned in chapter two, economical effects are triggers for trends such as the bear or the bull market. Sector classification is a contribution towards the balance of the sector weight. This contribution maintains a balance of the market capitalization range for the companies within their respected sectors. As this studies data is separated by sector to gain knowledge on the best performing sector when forecasting stock price trends, this criteria emphasizes the equality in company value within their sector. This removes smaller market capitalized companies within specified sectors which avoids an unequal lack of trade and trade value.

3.3.3 Company Stock Data

From the S& P 500 list a set of sectors were analysed to make a selection of the specified sectors that this study will explore. The analysis will be conducted over a set of 4 different sectors containing stocks for 4 different companies per sector. The sectors that were chosen for this study are the following: Finance, Technology, Healthcare and Industrial. These sectors provide a wide range of focus to conduct the hypothesis. A conclusive break down is displayed below in table 3.1.

Finance Sector				
Symbol	Name	Industry Type	Exchange Market	Market Capitalization : in Billions (Approx.)
BAC	Bank of America	Bank	NYSE	311.6
WFC	Wells Fargo and Company	Bank	NYSE	303.0
ALL	All State Corporation	Insurance	NYSE	37.4
MS	Morgan Stanley	Asset Management	NYSE	95.3
Technology Sector				
EBAY	eBay Inc.	Internet (Online Auction)	Nasdaq	39.4
IBM	International Business Machines Corporation	Technology Company	NYSE	141.18
CSCO	Cisco Systems Inc.	Networking products	Nasdaq	190.58
INTC	Intel	Semiconductor Companies	Nasdaq	218.56
Healthcare Sector				
CI	Cigna Corporation	Health Insurance	NYSE	50.15
UNH	UnitedHealth Group Incorporated	Health Insurance	NYSE	213.20
CNC	Centene Corporation	Health Insurance	NYSE	17.63
HUM	Humana Inc.	Health Insurance	NYSE	35.00
Industrial Sector				
MMM	3M Company	Diversified Industrial Equipment, Products	New York Stock Exchange	139.81
GE	General Electric Company	Diversified Industrial Equipment, Products	Nasdaq	151.76
LUV	Southwest Airlines Company	Airlines	NYSE	39.15
GD	General Dynamics Corporation	Aerospace and Defence Companies	New York Stock Exchange	60.51

Table 3.1 provides an overview of the four sectors used in this research. The sectors provide the four stocks per sector with their tickers, exchange markets, and market capitalization. The above details have been acquired by Yahoo Finance and represent their 2017 values.

To conduct this research, the data was obtained using daily historical values per stock. These daily values were gathered over a collective period of 55 months. This period was selected due to the research of this project running over a bear and bull market period. The period starts from January of 2007 to July of 2011. The bear market period stated in this project is set to run from January 1, 2007 to March 31, 2009 following the economic crash that hit the U.S economy in October 2007. The bull market period will take place from April 1, 2009 to July 1, 2011. This period was chosen to follow the recession and market trend after the economic crash. Although the bull market period did start in 2009 it has stayed on a record high until the present day, but due to the bear market's short window of time, the decision was made to make the markets equal in period length to test the parameters within the same restrictions. Each stock individually consists of 564 daily observations in the bear and bull market totalling 18,048 observations in total between both markets over the 55 months.

The historical data collected from Yahoo Finance for each stock consist of 6 daily features, Open, Close, High, Low, Adjusted Close and Volume. Each of these features will be applied to the various technical indicators increasing the number of selected features to 16 in total. These will be discussed in the Data Preparation stage of this chapter.

3.4 Data Preparation

Within this stage of the chapter a discussion and analysis of the major factors of the experiment will be conducted. The data preparation and feature selection stage is critical to the analysis as these are the parameters that provide insight into the data prior to forecasting using the models. The data preparation stage provides the ability to examine the data formation and data consistency preceding the implementation of the feature selection. As mentioned throughout chapter one and chapter two, stock market data has a non-linear noisy characteristic causing instability in forecasting its future behaviour. To rid this data of noise, the exponential smoothing average will assist in calming the noise and smoothing the data to fit the model parameters. Technical indicators are commonly used by traders but since there are a wide range of technical indicators available, a focus will be placed on indicators that were seen in the research and are commonly used by traders.

3.4.1 Exponential Smoothing Average

Though the Exponential Smoothing Average is considered a technical indicator it was introduced to this study as an aid to smoothing the noise in the data when calculating the return close value. This indicator will be used as a data transformer for the raw data to provide smoothing. The exponential smoothing average is a type of moving average such as the simple moving average. Traders such as Stan Weinstein have discussed the importance of the moving average in assisting with short and long-term predictions in the trend (1988). The exponential smoothing average is a formula which uses the data of recent daily traders for a specific time. The exponential smoothing average will be incorporated as a lag based calculation to assist with smoothing the data using the recent close values for said stock to provide a balanced set of close prices. As referred to by Handa, Hota, and Tanda, (2015) the exponential moving average minimizes the lag time by providing more weight towards the recent stock value. The amount of lag days in this study was tested over a period of 15, 30, and 50 days as the study is focused on long term trends. It has been found that analysis with long term lag time produced higher values (Khaidem, Saha, Roy Dey, 2016) the short-term trend in the bear and bull market were under studied. The formula for the exponential moving average is formulated taking the sum of the total number of lag day close price.

3.4.2 Average Directional Index

The Average Directional Index is a technical analysis tool used to calculate the strength of the trend rather than the direction. Though this study focuses on the prediction of the stock price movement this technical analysis is used on the basis of understanding if the trend in the stock price will continue staying strong. This technical indicator is used by traders to measure the strength or weakness of the trend for a specified stock in order to determine entering or exiting (Handa, Hota, Tandan, 2015; Kim, 2003). In addition to the strength of the trend, the Average Directional Index provides the detection of where the stock can lie on the trend itself. If the return value of the stock is at a lower value, the stock would be then at a point where an investor would sell causing the market the volume of stocks traded to decrease. The average directional index uses 3 values from the historical data. The formula for the Average Index takes the difference between the directional movement index (D+ and D-). The directional movement index calculates as the D- as high of minus one previous high value and high from the current day while D+ is low of the previous low value minus the current low. Then the calculation will be multiplied by that of the smooth moving average. The average directional index is interpreted as follows; when the calculated number is at higher value the trend is strong, if the value is low the trend is frail.

3.4.3 Bollinger Band

The Bollinger Band although unusual in academic analysis of stock market research, is a highly utilized technical strategy used by investors. Many investors incorporate this type of technical indicator on a candle stick as the Bollinger band shows a set of three lines; one upper band, one lower band displaying the volatility and the third band displaying the moving average. The wider the upper and lower bands are from the close price and moving average, the more unpredictability is displayed. Figure 3.2 displays a visualization on the eBay stock price when the price of the Bollinger Band has been implemented. Keep in mind the visualization below is used on a training set that has been smoothed by the EMA. The below figure displays a small range between the upper and lower bands around the moving average displaying little volatility in the change of price. This technique is unique in the sense that within the formula the use of moving average is incorporated based on a 20-day period. This tool utilizes the volatility of the stock price which focuses on the business objective of this study. Using this technical indicator will assist in visualizing and calculating a drop or spike in the market based on the closing price. This technical indicator uses the calculation of a moving average to then subtract or add the standard deviation of the stock price (Arratia, 2014). The formula for this technical indicator is split into three focuses. [5]The upper band uses the SMA (smoothing moving average calculation of the closing price) plus the standard deviation of the price over a 20-day period multiplied by 2. The lower band uses the SMA minus the standard deviation of the price over a 20-day period multiplied by 2. The middle band is the calculation of the SMA over the span of a 20-day period.

[5] http://stockcharts.com/school/doku.php?id=chart_school:market_indicators

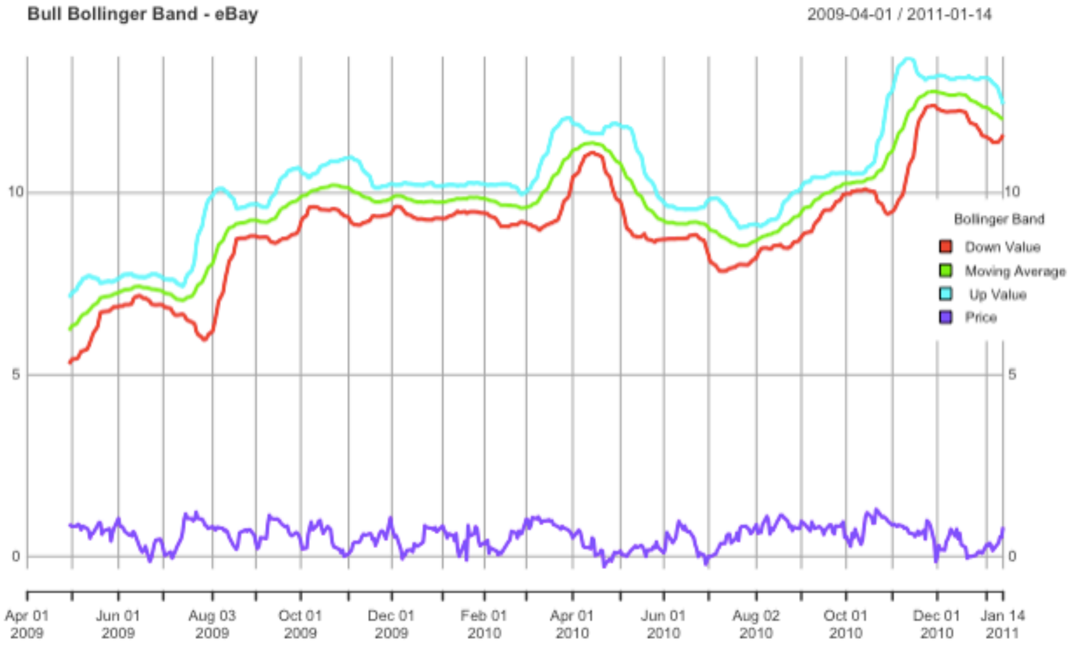


Figure 3. 2 illustrates the visual of the closing price of the eBay stock in the Bull Market over the training date range when used in the Bollinger Band technical indicator.

3.4.4 Chaikin Money Flow

The Chaikin Money Flow is another uncommon technique within the academic research field that is highly utilized within the trading community. The CMF uses the volume and the close price to determine the trading range. This tool is used to assist when signaling that a price is in a bull or bear type market with a specified indicator of .25. This technique assumes that a high close price and increased volume in a day-to-day trade expresses that of a bull market (Kannan, Sekar, Sathik, Arumugam, 2010; Handa, Hota, Tandan, 2015). The CMF is based on the initial measures of the MF (money flow) volume over a certain time. The calculations for the CMF are based on Money Flow which calculates the absolute means of the close value minus the low value and the high value minus the close value which is then divided by the high minus the low value. This calculation result is then multiplied by the volume of the specific period indicated and then divided by the initial sum of the volume over a set period. This period is at a basis of 20.

3.4.5 Commodity Channel Index

The Commodity Channel Index is a trading tool used by traders to determine the behaviour of a new trend or to predict those of long lasting conditions. This trading tool is commonly used to assist traders in observing the end of a current trend. This technical analysis tool indicates when a cycle change will occur (Handa, Hota, Tandan, 2015; Kim, 2003). The CCI provides an initial price indicator stating that when the price is above the normal average or influxed at a higher rate then the value of the CCI will be high and goes for the opposite if the price is exceptional below the average then the CCI will indicate the value at a lower rate. The CCI is calculated using the typical price value which is the high value plus the low value plus the close value. The sum of these three values then is divided by three in return taking the average of the three values for a specified date. The TP value is then subtracted by the day period (typically 20) by the SMA. This value will then be divided by the constant x - the mean deviation of the TP of the SMA day period.

3.4.6 Relative Strength Index

The Relative Strength Index is a tool used to determine the strength of the price of a stock by determining the average movement of the close price. This feature is considered as a momentum indicator establishing the strength of the close price. The standard time frame that this formula uses is a set number of 14 days. This technical indicator is a commonly used tool within the trading and academic realm (Khaidem, Saha, Dey, 2016; Handa, Hota, Tanda, 2015; Kim, 2003). The use of this indicator is to detect when a stock is being overbought such as that in an upward trend or oversold such as that in a downward trend (Khedkar, Argiddi, 2013; Kannan, Sekar, Sathik, Arumugam, 2010). This momentum will be a highly profitable trigger in detecting the better performing sector in the bull and bear market. It is notable to state that the RSI provides the measurement of speed change in the change in price. The calculation for the RSI uses the basis of RS which takes the average gain of stock divided by that of the average loss. This gain and loss is typically used on the basis of a 14-day period. The gains and losses are built averaging the price of gains and loss each over the past 14 days. Once the RS is then calculated the calculations of the RS are added to one and divided by 100. Once this value is formulated the total value is then subtracted from 100. The resulting values of the RSI are displayed as a zero value then the RSI represents a price decrease over the 14-day timeframe. If RSI then ranges towards 100 the price over the 14-day time frame is increased. An initial value of 70 in the range of 0 to 100 specifies that the RSI detects overbuying, while if the RSI value ranges below 30 the stock is then oversold.

3.5 Modelling

This section of the chapter will incorporate the prepared data from section 3.4 with supervised machine learning models. The goal of this research is to determine the accuracy of a support vector model against a random forest model to forecast the stock price direction in a bull and bear market. Two supervised machine learning classification models will be implemented. As referenced in the literature review, many academics and researchers have applied a variety of classification and regression machine learning algorithms to predict stock price and price movement.

A support vector machine (SVM) is an error-based supervised machine learning model, this model type has been a highly exploited algorithm within the stock market forecasting realm. As the SVM model is a classification algorithm, the model is trained differently than a regression model but has the same outcome or goal. The model builds a linear hyperplane separating the two classification values (0,1). With these variables split by the hyper-plane, the values neighbouring the linear hyperplane are those of a support vector (this may not be the case where the hyperplane is able to split the data by variables). In order to find the best optimal boundary for the hyperplane, it is necessary to train the SVM using the train dataset to forecast the target levels. The SVM is built using a set of parameters with a focus on the kernel type, the cost range, and gamma. The overall aim of the SVM model is to reduce the upper bound of the "generalization error".

Random forests (RF) are an ensemble type of decision tree algorithm built by aggregating trees. The RF algorithm is an information-based learning algorithm. Similar to the SVM, the RF algorithm can be used as both a classification and a regression model. The random forest model works by building an assortment of decision trees when training the model. The training of the model will utilize that of the random collection of feature variables to build the tree. If the data has any slight noise the model will grow in a different formation than if the noise was smoothed out. To avoid this problem, the random forest model then splits numerous decision trees using the feature variables, increasing the bias of the overall model.

With this type of model containing numerous split trees, the initial tree would be the only tree to actually see and utilize the entire training data set as each tree is split from the decision/ feature of the previous tree. The algorithm provides two optimal ways of splitting each partition in the tree, the Shannon Entropy or the Gini impurity. The Shannon Entropy is built on the basis that the trees are split on the root of the volatility of the data fed to the model. The Gini impurity splits the trees on the root of the level of misclassification when predicting the target levels of the dataset. This algorithm is built by avoiding overfitting unlike that of the SVM. The RF model incorporates and manages a large number of features. This model is specific towards the overall objective of this study as it triggers the best performing features used in the model. This can then show whether specific features perform better for a sector in the bear market rather than in the bull market. The way that the RF works is by using a boostrap sample of ntrees. Ntree is a type of parameter set for the RF. This means that this parameter then builds n number of decision trees based on the number of ntrees chosen. In addition to the ntree parameter the RF builds its parameters by randomly sampling its variables as the tree splits.

3.6 Evaluation

This section will provide the ability to see which model performs better to satisfy the hypothesis that the support vector machine model will outperform the random forest model. To determine the performance of these models, a confusion matrix will used. From the confusion matrix the precision, recall, and F1 will be evaluated to provide an in-depth scope of the classifiers. To assist in evaluating the models, a validation will be conducted for each model using the train and test cross validation.

3.6.1 Confusion Matrix

A confusion matrix provides a large range of performance measures and highlights different aspects of the performance of the models. The performance measurements are based on a basic set of four algorithms. These algorithms are measured as the following: true positive rate (TPR), true negative rate (TNR), false negative rate (FNR), and false positive rate (FPR). Using these individual measures, an algorithm is then computed to calculate the accuracy level. These measures will be based on the classification accuracy and misclassification of the predicated price direction of the stock. In addition to the four measures with the exception of the accuracy algorithm, precision, recall and F1 will measured from the confusion matrix. These three measures are optimal measures for classifying binary performance (Khaidem, Saha, Dey, 2016). As mentioned in the literature review, the precision, recall and F1 score measure the positive detected target variable, the negative positive target variable and the misclassification rate. These variables assist in evaluation the performance of the overall predictive model.

3.7 Deployment

In order to deploy this study in real time, a few adjustments must be made to make the model less artificial. As using certain features such as that of the EMA, the model transitions towards a normalized structure reducing real time noise and producing a less realistic outcome. The research will be reviewed and re-approached in order to analyse the data in a less artificial manner. This study will not be adopted for trading purposes as of yet, but will be applied in this instance for the purposes of this research.

3.8 Strengths and Limitations

This stage of the chapter will evaluate the strengths and weaknesses of the design model and methodology. A major strength in this model is the formation of the data into both the bear and the bull market while accurately being split by specified date to ensure formality within the experiment. Using three different models to predict the accuracy of the performance provides a diverse set of parameters per model which takes advantage of the various strengths of each model. The second primary strength is that of the technical indicators providing specified parameters that follow respective learning rules in regards to their feature variables. Having a set of 4 various sector types with 4 different stocks per sector gives a diversity to the data that the models will learn from. This proves or disproves the outcome of the hypothesis that specific sectors may work better in said market using specific technical indicators.

One limitation to the study is that using this type of data required EMA smoothing in order to rid it of noise. This caused the data to form a bias and become artificial. To ensure equality in the experiment, for performance accuracy, each model kept the same parameters and was treated in the same manner. This meant that although some features were better to incorporate for various stocks, each feature was set to the same formation of the previous stock.

3.9 Overview

The Design and Methodology chapter provides each stage of the procedure for how the experiment was organized and conducted. From a business perspective the models will assist in forecasting the “best” performing model in the bear and bull market while using a set of technical indicators. This chapter provided insight on the data and what the data contains with an analysis on the technical indicators used as feature variables. The dataset preparation stage prepares the data and gives the explanation of two predictive models to analyse data in two market types to predict the stock price direction (though a limitation was set on utilizing the parameters in a single formation to evaluate the sectors performance.) Another potential issue with the design and methodology is the use of EMA. Although EMA provides noise cancelation to the data, it also creates a less realistic formation of the data thus providing a set of artificial data. The following chapter will discuss the implementation of the design and methodology proving the results of the models and expressing the changes in respect to each market.

.

4 IMPLEMENTATION AND RESULTS

4.1 Introduction

This chapter will review the implementation of the data into the models, and their results. The following sections will be broken down into; data preparation and exploration, modeling, and evaluation. Although the formation deviates from chapter three's CRISP-DM layout, any changes made to the design process when implementing will be noted.

To reproduce this experiment, the software being used to conduct the research is R and Excel. While R contains a large number of available packages and libraries, a series of specific packages have been installed for implementing the time series technical indicators and supervised models. The packages used for this were; TTR, timeSeries, quantmod, pdfetch, xts, and ggplot. The packages and libraries used for the supervised machine learning algorithms were; RandomForest, e1071 and caret. Excel was used to confirm calculations and build a few graphs.

4.2 Data Preparation and Exploration

4.2.1 Data Preparation

This experiment is set to focus on two different time periods in history but with identical data and identical parameters to gain insight on the performance of the SVM model against the RF. Not only will the performance of the models be analysed, but the performance of the stocks in each sector will be reviewed to determine the "best" performing sector using technical indicators in the bear and bull market.

To accomplish this goal, the data in both markets are kept and acquired in the exact same way using the R library package PDFetch. This process keeps all of the data in the same format and containing the same variables for both market periods. As previously mentioned in the design and methodology chapter, the variables retrieved using the PDFetch library are: open, high, low, close, adjusted close, and volume. The acquired data from each stock was not combined with any of the other data in order to understand the features' effect on the individual stock. In the model, each stock went through the same process but at different stages.

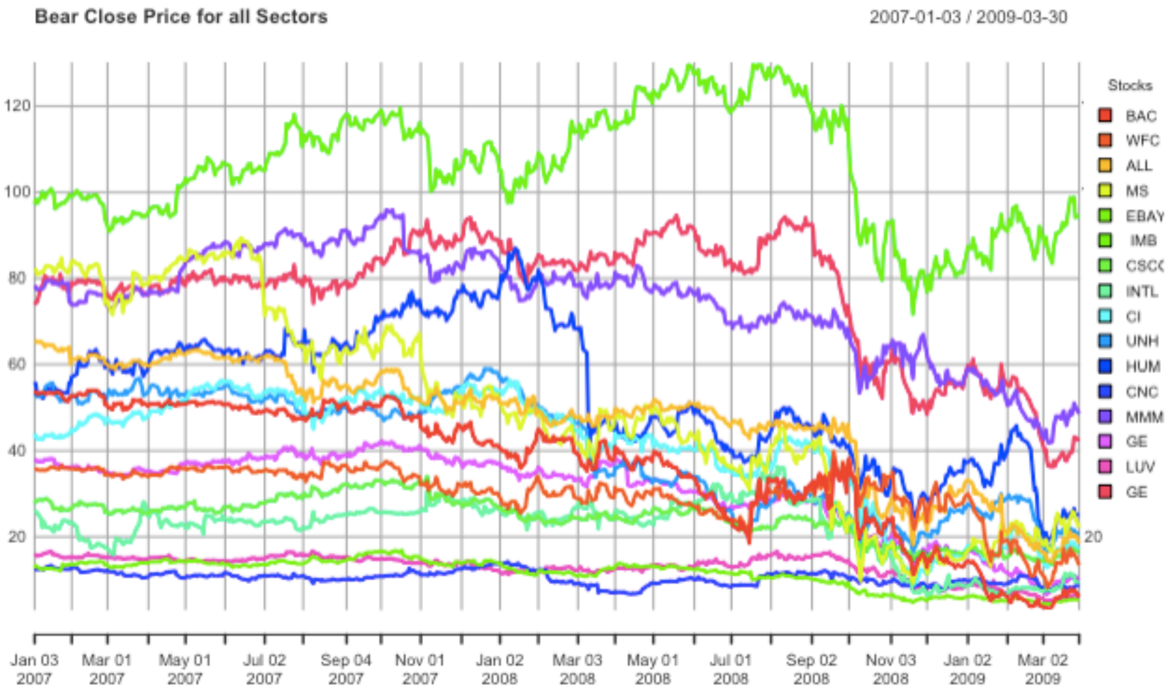


Figure 4. 1 displays an overview of the close price for each of the 16 stocks in the bear market. These prices are displayed in their raw format without any noise smoothing. Stocks such as eBay show a distinct visual of the close price visibly going downward. Other stocks as such Southwest Airlines maintain a steady closing price.

Following the data overview, the next step taken was to split the data into training and testing sets (hold out). This process followed an 80/20 percent ratio. To do this split, an R function in the zoo library was needed. A function called “windows” provided the ability to manually split the data by specified date. As the data in this study was a daily time series it was crucial that the data was continuous by date. The training data for the bear market was split from January 1, 2007 to October 15, 2008 while the test data period was from October 16, 2008 to March 31, 2009. The bull market training data was split from April 1, 2009 to January 14, 2011 and the test from January 18, 2011 to June 30, 2011. Due to the data being made up of daily values the actual days it consisted of are known as “trade only days” which means Monday to Friday as they are the only days considered in each month for the daily trade values. As referenced in the date split, a jump of two days in the bull market test set is seen as these were non-trade days.

A major phase in the preparation stage prior to the feature exploration was the use of converting the close price into a predictable value when creating the return value for forecasting the target variable; direction. As mentioned in the Design and Methodology chapter, an exponential moving average was first applied to the close value to smooth out the data prior to building the return. Using this technical indicator to smooth the data applies more weight towards newer close values which reduces the weight of past historical values. The purpose of this phase is to eliminate any noise provided from historical close values. This process has been used in recent research such as Khaidem, Saha, and Dey (2016). Using the EMA value a return calculation was formulated. The formula was devised using the log function and lag functions in R. The value was calculated and tested on a number of different lag days to determine a more accurate return. The lag days are the number of closing days prior to said closing day for example; yesterday or two days prior to yesterday. The return is then calculated based on the difference in close values. If the return value is positive then the direction will be indicated as a 1, if the value of the return is negative then the direction will be indicated as 0 (Khaidem, Saha, Dey, 2016; Kim, 2003; Enke, Thawornwong, 2005).

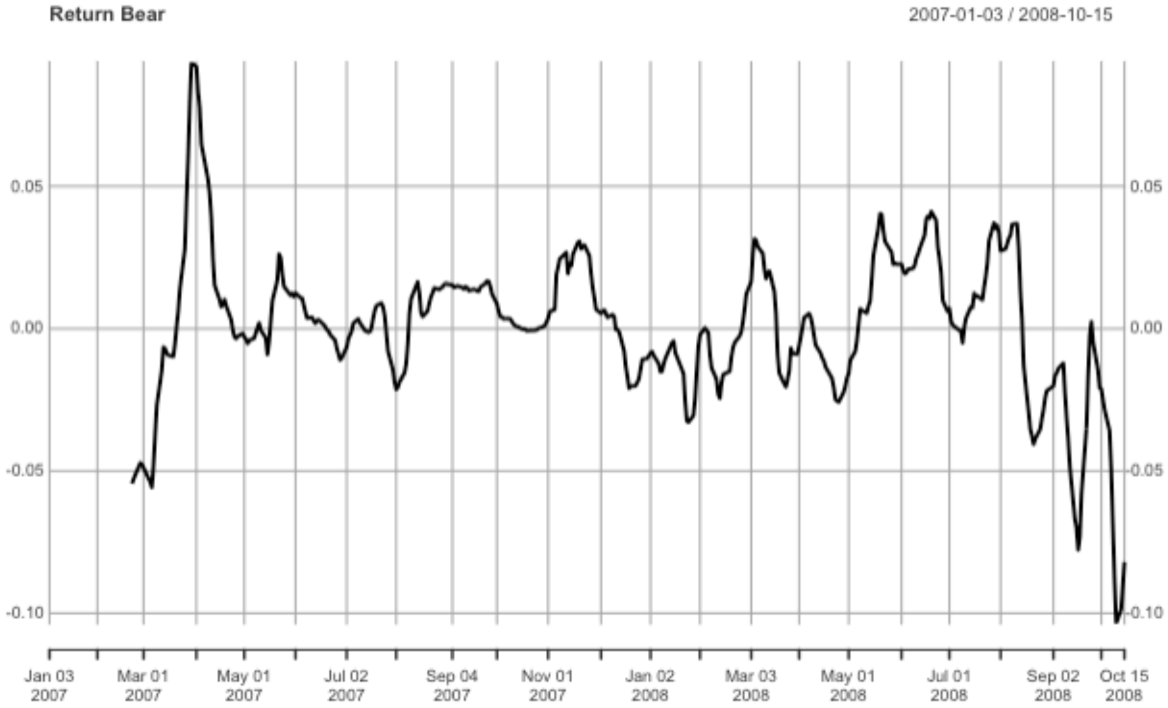


Figure 4. 2 demonstrates the return calculation for the Intel Company in the Bear Market. The return was based on a EMA of 30 Days with a return lag time of 5 days. Above displays an upward peak in in April of 2007 for the close price change following a downward price difference.

4.2.2 Feature Exploration

This section will review the feature exploration used in this study. Each technical indicator will be analysed. Though the EMA was used in the data preparation stage of the return model, many of the technical indicators rely on a moving average type, typically Weighted Moving Average, for their respective formulas. Figure 4.2.3 and 4.2.4 are correlated heat maps representing only the technical indicators used within both the bear and bull markets. Each of these technical indicators are duplicated with the same parameters for both markets in order to gain an understanding of the performance of each sector in each market.

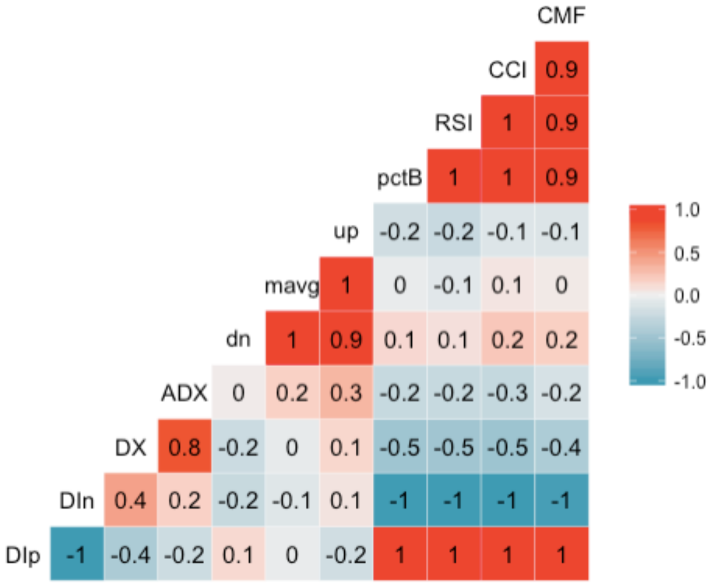


Figure 4. 3 represents the correlation of all the technical indicators presented for the Three M Company only in the Bull Market. There is a high positive correlation between respective technical indicators such as RSI, Dip, and CMF, and CCI.

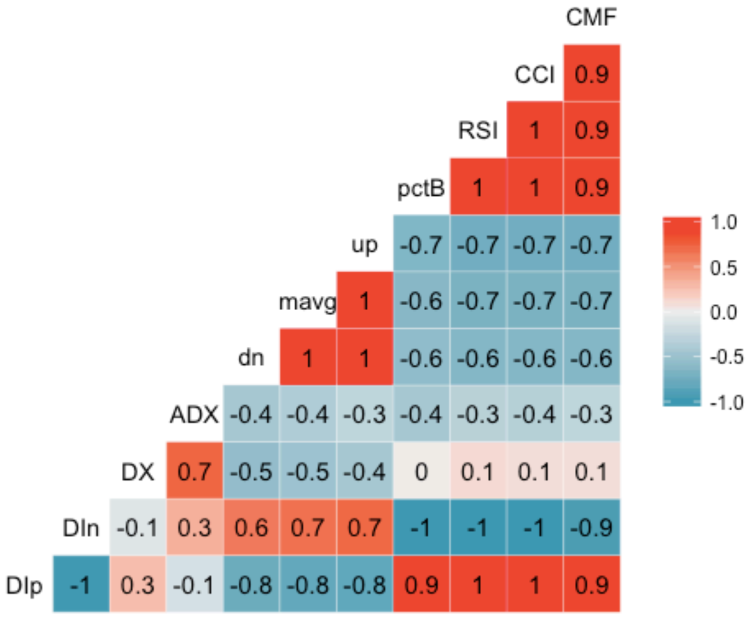


Figure 4. 4 signifies the correlation of all the technical indicators presented for the Three M Company only in the Bear Market. Some of the indicators such as the RSI and Dip are highly correlated as seen in the Bull market. Technical indicators such as the ADX and dn that were at a 0 correlation in the Bull market are even weaker in the bear market.

4.3 Modelling

Following the data preparation and feature extraction, models have been created and made based on the criteria needed to evaluate each model. As seen in the previous section the data has initially been split into 80/20. 80% of the data is training data that will be used to build and tune the models. 20% of the data will be used as the test data for the performance and prediction stage of the models. This split was completed to assist and avoid the random sampling of the entire dataset for both markets. Although this process is varied from k-fold cross validation, researchers such as Kim (2003), used this process when using time series data. In addition, utilizing a test set assists in preventing “peeking” when tuning the models (Kelleher, Mac Namee, D’Arcy, 2015). “Peeking” is when the model has been previously exposed to the trained and testing data. This would prevent a false high level of performance as the models will be tested using a set of data that has not been exposed to the tuning or build process.

To implement the best model with the data provided, the use of the tune and best model functions were presented with the SVM and RF algorithms. To help improve the SVM performance the SVM model used the tune, predict, and SVM functions from the library e1071. Within the SVM there are three hyper parameters that are fundamental to building the model. Hyper parameter tuning is used to implement a grid type search over a range of two parameters; the cost and gamma. The gamma and cost are initial in the tuning stage using the training data to test on. As the data is non-linear and the target variable is a binary of 0 or 1, a radial basis function (RBF) kernel and c-classification were set and tested. When tested using the radial kernel the results produce zeros for about twenty of the stocks in both markets, while other stocks showed poor overall results. This then led to the testing of the data against other kernels. The kernel that produced a high performance level was the linear kernel.

To determine the best range of parameters for the cost and gamma, a combination of ranges were tested. The ranges were vast as the C were from 0.001 to 100, while the gamma ranged from 0.5 to 5. For the purposes of this experiment the pre-established parameters were retained and used against every stock without adjustments for both the bear and bull market. The reasoning for this was to find an equal effect of the outcome for each model to perform in an equally designed experiment. It is important to note that when using the linear kernel, the use of the gamma is a coefficient in the SVM model and is a necessary parameter as mentioned in previous research. The performance of the model built using the "best" tuned values was then tested against the test set of data.

The RF, similar to that of the SVM utilizes the tune function to reduce the OOB (out of bag error) rate to determine parameters for the model. To apply the tuneRF function for the RF model, the randomForest package is used. To model this tune function, the data was used against the train data. The parameters that are looked at are; mtry and ntreeTry. Mtry is the number of features randomly sampled using the cross validation. This will determine using the OOB error rate as seen in Figure 4.5.1. The ntreeTry is the number of trees used at the tuning stage. The ntree value will be an important part of the RF model as it will determine the number of times a tree should be split. To determine the lowest OOB error rate, a combination of ntree's were executed resulting in a total of 400 ntrees and 6 mtry with the lowest OOB error rate.

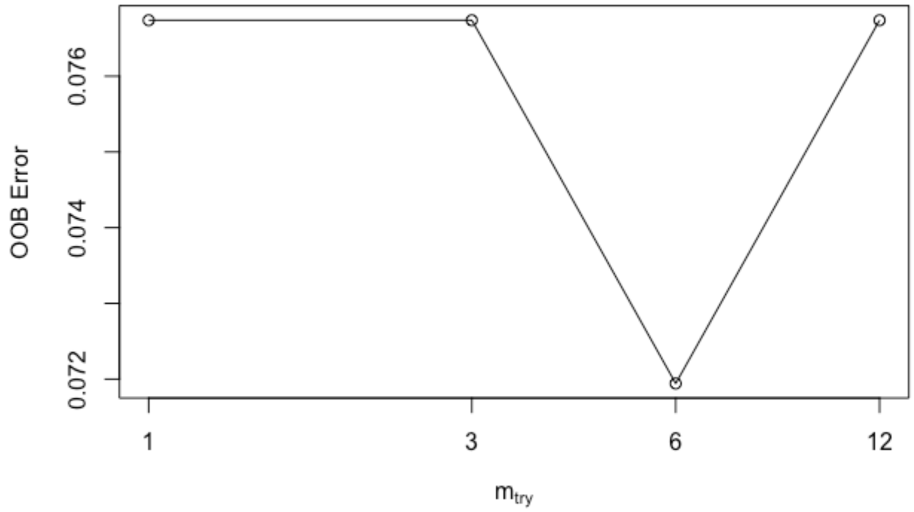


Figure 4. 5 displays the best number of mtry based on the lowest out of bag error rate.

4.4 Model Evaluation

This section of the chapter will cover the performance of the SVM and RF in order to gain an understanding for the overall objective; a SVM predicting with a higher level of accuracy than a RF in either a bull or bear market when forecasting the stock price direction. In addition to the overall objective, this section will determine the outcome of the two models used to determine which of the 4 sectors perform better in a bull or a bear marketing using a selected assortment of related technical indicators analyzed by traders.

The following section will be broken down into the 4 sectors, Financial, Technology, Healthcare, and Industrial. This is to provide an overall perspective of how the effects of each model and their respective parameters performed on each stock and how the overall sector performed in both markets. A further analysis of the evaluations will be presented in Chapter 5.

To be able to evaluate each of the models, the use of the caret package provided by R is necessary on the models that have been tuned using the dataset and predicted using the hold-out set. The predictive results will be evaluated using the "confusionMatrix" function. While each of the components of the confusion matrix were analysed, the tables presented for each stock in the 4 sectors focuses on the Accuracy, Precision, Recall, and F1. Each of these values were derived for the SVM and RF in both the bull and bear market.

As a brief reiteration, these calculations are a common use for performance evaluation. Accuracy provides insight on the measures of correctly classifying the predicted target. Accuracy will not be the important factor when interpreting the overall model results as it fails to provide accuracy if the testing dataset is biased toward one target variable. For example, if in the bear market, a stock's target variable has 99% of occurrences where the value is 0 then the accuracy of the model will be 0.99, but fail to correctly classify the 5% of 1 target variables. This results in an unbalanced test set causing a biased classifier and misleading the classifiers overall performance. For this reason, this study analyses additional metrics when using the confusion matrix to evaluate the performance of each model. Precision, Recall, and F1 provide a deeper scope of the models when determining the strength of the overall performance. The Precision indicates the confidence level when forecasting positive target levels. Positive levels predicted by a metric can be analysed as True Positive and True Negatives. Recall signifies that the confidence level of the model correctly classifies the true positive rate. The F1 score is a harmonizing tool that collectively analyses the performance of the model as it displays its shortcomings.

4.4.1 Financial Sector

<table>
<tr><th colspan="5">Bear Morgan Stanley</th></tr>
<tr><td></td><td>Accuracy</td><td>Precision</td><td>Recall</td><td>FI</td></tr>
<tr><td>SVM</td><td>0.56</td><td>0.95</td><td>0.36</td><td>0.52</td></tr>
<tr><td>RF</td><td>0.82</td><td>0.70</td><td>0.64</td><td>0.67</td></tr>
<tr><th colspan="5">Bull Morgan Stanley</th></tr>
<tr><td></td><td>Accuracy</td><td>Precision</td><td>Recall</td><td>FI</td></tr>
<tr><td>SVM</td><td>1</td><td>1</td><td>1</td><td>1</td></tr>
<tr><td>RF</td><td>0.96</td><td>0.96</td><td>1</td><td>0.98</td></tr>
<tr><th colspan="5">Bear Bank of America</th></tr>
<tr><td></td><td>Accuracy</td><td>Precision</td><td>Recall</td><td>FI</td></tr>
<tr><td>SVM</td><td>0.97</td><td>1</td><td>0.97</td><td>0.99</td></tr>
<tr><td>RF</td><td>0.95</td><td>1</td><td>0.94</td><td>0.97</td></tr>
<tr><th colspan="5">Bull Bank of America</th></tr>
<tr><td></td><td>Accuracy</td><td>Precision</td><td>Recall</td><td>FI</td></tr>
<tr><td>SVM</td><td>0.95</td><td>0.99</td><td>0.96</td><td>0.97</td></tr>
<tr><td>RF</td><td>0.95</td><td>0.99</td><td>0.96</td><td>0.97</td></tr>
<tr><th colspan="5">Bear Wells Fargo and Company</th></tr>
<tr><td></td><td>Accuracy</td><td>Precision</td><td>Recall</td><td>FI</td></tr>
<tr><td>SVM</td><td>0.73</td><td>.84</td><td>0.83</td><td>0.83</td></tr>
<tr><td>RF</td><td>0.76</td><td>0.87</td><td>0.83</td><td>0.85</td></tr>
<tr><th colspan="5">Bull Wells Fargo and Company</th></tr>
<tr><td></td><td>Accuracy</td><td>Precision</td><td>Recall</td><td>FI</td></tr>
<tr><td>SVM</td><td>0.94</td><td>.94</td><td>1</td><td>0.97</td></tr>
<tr><td>RF</td><td>0.86</td><td>0.86</td><td>1</td><td>0.93</td></tr>
<tr><th colspan="5">Bear Allstate Corporation</th></tr>
<tr><td></td><td>Accuracy</td><td>Precision</td><td>Recall</td><td>FI</td></tr>
<tr><td>SVM</td><td>0.89</td><td>0.96</td><td>0.87</td><td>0.91</td></tr>
<tr><td>RF</td><td>0.85</td><td>0.98</td><td>0.81</td><td>0.89</td></tr>
<tr><th colspan="5">Bull Allstate Corporation</th></tr>
<tr><td></td><td>Accuracy</td><td>Precision</td><td>Recall</td><td>FI</td></tr>
<tr><td>SVM</td><td>0.75</td><td>0.98</td><td>0.70</td><td>0.82</td></tr>
<tr><td>RF</td><td>0.85</td><td>0.82</td><td>0.90</td><td>0.86</td></tr>
</table>

Table 4. 1 illustrates the SVM and RF prediction results for the 4 stocks chosen for the financial sector.

4.4.2 Technology Sector

Bear eBay Inc.				
	Accuracy	**Precision**	**Recall**	**FI**
SVM	**0.60**	**1.00**	**0.58**	**0.74**
RF	**0.57**	**1**	**0.57**	**0.73**
Bull eBay Inc.				
	Accuracy	**Precision**	**Recall**	**FI**
SVM	**0.89**	**0.89**	**0.94**	**0.91**
RF	**0.83**	**0.74**	**1**	**0.85**
Bear IBM				
	Accuracy	**Precision**	**Recall**	**FI**
SVM	**0.63**	**0.90**	**0.76**	**0.83**
RF	**0.80**	**0.97**	**0.65**	**0.78**
Bull IBM				
	Accuracy	**Precision**	**Recall**	**FI**
SVM	**0.80**	**0.45**	**1**	**0.62**
RF	**0.75**	**0.31**	**1**	**0.47**
Bear Cisco Systems				
	Accuracy	**Precision**	**Recall**	**FI**
SVM	**0.61**	**1**	**0.61**	**0.76**
RF	**0.66**	**0.98**	**0.64**	**0.78**
Bull Cisco Systems				
	Accuracy	**Precision**	**Recall**	**FI**
SVM	**0.94**	**0.95**	**0.99**	**0.97**
RF	**0.91**	**0.92**	**0.99**	**0.95**
Bear Intel				
	Accuracy	**Precision**	**Recall**	**FI**
SVM	**0.81**	**0.95**	**0.81**	**0.87**
RF	**0.73**	**0.91**	**0.76**	**0.83**
Bull Intel				
	Accuracy	**Precision**	**Recall**	**FI**
SVM	**0.72**	**0.49**	**0.95**	**0.65**
RF	**0.67**	**0.40**	**0.94**	**0.56**

Table 4. 2 displays the SVM and RF prediction results for the 4 stocks in the technology sector.

4.4.3 Healthcare Sector

Bear Cigna				
	Accuracy	**Precision**	**Recall**	**FI**
SVM	**0.46**	**1**	**0.43**	**0.61**
RF	**0.89**	**0.76**	**0.96**	**0.85**
Bull Cigna				
	Accuracy	**Precision**	**Recall**	**FI**
SVM	**0.94**	**0.67**	**0.75**	**0.71**
RF	**0.91**	**0.56**	**0.63**	**0.59**
Bear United Health Group Incorporated				
	Accuracy	**Precision**	**Recall**	**FI**
SVM	**0.96**	**0.97**	**0.94**	**0.96**
RF	**0.99**	**1**	**0.97**	**0.99**
Bull United Health Group Incorporated				
	Accuracy	**Precision**	**Recall**	**FI**
SVM	**0.91**	**0**	**0**	**0**
RF	**0.86**	**0.14**	**0.17**	**0.15**
Bear Centene Corporation				
	Accuracy	**Precision**	**Recall**	**FI**
SVM	**0.90**	**0.88**	**0.95**	**0.91**
RF	**0.94**	**0.90**	**1**	**0.95**
Bull Centene Corporation				
	Accuracy	**Precision**	**Recall**	**FI**
SVM	**0.94**	**0.69**	**0.90**	**0.78**
RF	**0.85**	**0.08**	**1**	**0.14**
Bear Humana Inc.				
	Accuracy	**Precision**	**Recall**	**FI**
SVM	**0.96**	**1**	**0.93**	**0.96**
RF	**0.95**	**1**	**0.90**	**0.95**
Bull Humana Inc.				
	Accuracy	**Precision**	**Recall**	**FI**
SVM	**0.98**	**0**	**0**	**0**
RF	**1**	**0**	**0**	**0**

Table 4. 3 shows the SVM and RF prediction results for the 4 stocks in the healthcare sector.

4.4.4 Industrial Sector

Bear 3M Company				
	Accuracy	Precision	Recall	FI
SVM	0.82	1	0.82	0.90
RF	0.82	1	0.82	0.90
Bull 3M Company				
	Accuracy	Precision	Recall	FI
SVM	0.86	0.91	0.79	0.85
RF	0.80	0.55	0.90	0.68
Bear General Electric Company				
	Accuracy	Precision	Recall	FI
SVM	0.82	1	0.82	0.90
RF	0.92	0.98	0.93	0.96
Bull General Electric Company				
	Accuracy	Precision	Recall	FI
SVM	0.89	0.90	0.95	0.92
RF	0.83	0.77	1	0.87
Bear Southwest Airlines Company				
	Accuracy	Precision	Recall	FI
SVM	0.92	1	0.92	0.96
RF	0.89	0.99	0.90	0.94
Bull Southwest Airlines Company				
	Accuracy	Precision	Recall	FI
SVM	0.88	0.80	1	0.89
RF	0.89	0.86	0.96	0.91
Bear General Dynamics Corporation				
	Accuracy	Precision	Recall	FI
SVM	0.80	1	0.77	0.87
RF	0.80	0.98	0.78	0.87
Bull General Dynamics Corporation				
	Accuracy	Precision	Recall	FI
SVM	0.80	0.73	0.92	0.82
RF	0.74	0.63	0.91	0.75

Table 4. 4 shows the SVM and RF prediction results for the 4 stocks in the industrial sector.

4.5 Overview

This chapter provides the practical deployment of the design methodology of the models and the data. This chapter provides the actual results of each model's prediction for each stock in their respective sector with the aim of gaining insight into the performance of the SVM against the RF.

The data preparation stage reviews the practical stages of the retrieval of the data while explaining the pre-processing stage prior to the feature extraction. The key points to the preparation were the importance of the 20% hold out test set. This test set provided unseen data to the models when predicting in order to avoid any issues such as "peeking".

The feature exploration stage reviews the implementation of the EMA to remove noise in the data while providing a smoother set of close prices to calculate the return value to predict the stock price direction. Once the direction was calculated, the implementation of the 5 technical indicators were discussed and reviewed in a heat correlation matrix. This provided an overview of the effect in a bull and bear market on the technical indicators while illustrating a view of the highly correlated technical indicators. Though each stock provided different correlation to similar technical indicators, much of the variance of correlation shifts not only in the bull and bear market, but per stock in each sector. This reflects back to the initial 6 features, high, low, open, close, adjusted close, and volume acquired at the data retrieval stage. Following the feature exploration, the modelling and evaluation stage provide insight into the measures taken to tune and build both the SVM and RF model. To evaluate the performance of these models and stocks in each sector, the confusion matrix was implemented which provided the results seen in the tables 4.4.1 – 4.4.4. A complete analysis of these tables and results are explained in the following chapter in order to accept/reject the hypothesis.

5 EVALUATION

This chapter will review and analyse the findings of the experiment performed in chapter 4. The aim of this evaluation is to examine the performance of the SVM and RF model to accept or reject the hypothesis and analyse the findings. To reiterate; the hypothesis states that the use of a SVM model will outperform a RF model based on various stocks in specific sectors utilizing only technical indicators in both a bull and bear market. Although the No Free Lunch Theorem as mentioned in the literature review chapter of this study does imply that not one best algorithm can outperform another algorithm on the basis that they are built using different indicators (Wolpert and Macready, 1997), this study will then question the ability of the SVM model versus that of the RF model against predicating the stock price direction in a bull and bear market. This chapter will provide an in-depth analysis of each models findings based on the model, the sector, and the market.

This chapter will primarily review the performances of each model, then transition into interpretations of the results, and lastly cover the strengths and weakness of the results.

5.1 Result Evaluation

The evaluation will be split into four divisions. Each division will contain the bear and bull market for the sector to examine the F1 score for the SVM and RF. To analyse these models accurately, it is important to reiterate that the F1 score ranges from 0 to 1 and the closer the value is to one the better the model has performed. The focus on this metric was determined as it represents the weight of both the precision and recall providing an in-depth understanding. The confusion matrix for the models in both markets classified the positive value as 0 and the negative value as 1.

5.1.1 Financial Sector

As seen in Figure 5.1.1 and 5.1.2 the SVM performs 4 time at a higher F1 score rate than the RF out of the 8 stock implementations. The RF only contains a higher F1 score in 3 out of the 8 performances. This left one stock, Bank of America, with an equal level of measure at 0.97 in the bull market. With all the ranges above a 0.5 scale level, the performance of both models was equally high providing evidence that the models were well tuned. When analysing the F1 score of each stock in the bull and bear market, the Morgan Stately stock displays a poorer level of performance in the bear market. When analysing the recall and precision separately, the precision for the SVM displays a .95 which indicates a higher performance but a .36 recall value. Whereas the RF performs drastically higher for recall with a .64, but a .70 for precision. This implies that the RF was able to predict .70 positive instances, while the SVM predicted a larger number of negative instances with a .95 precision. Thus implying that the SVM misclassified many of the negative instances.

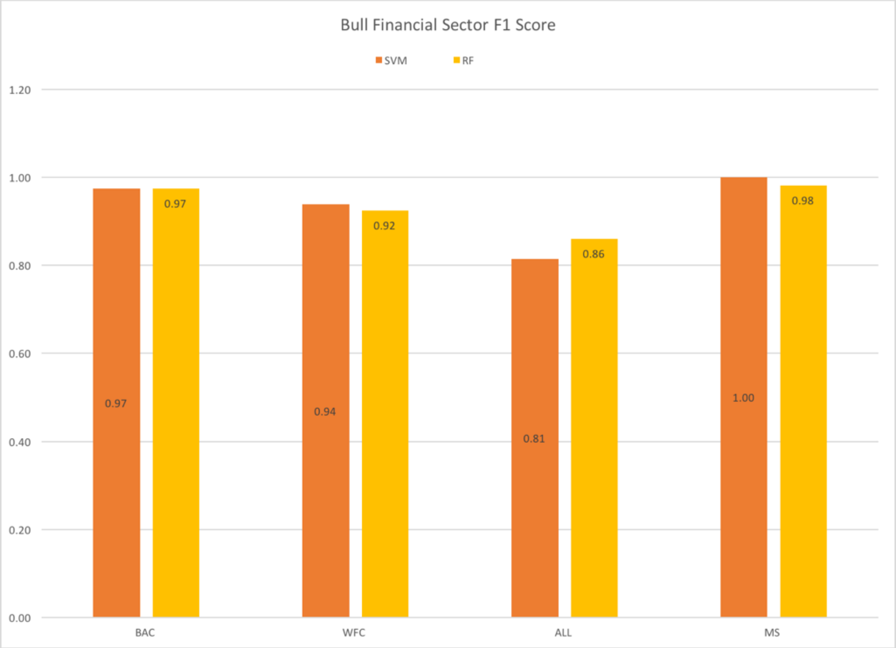


Figure 5. 1 illustrates the F1 score of the SVM and RF model for the Financial sector in the bull market.

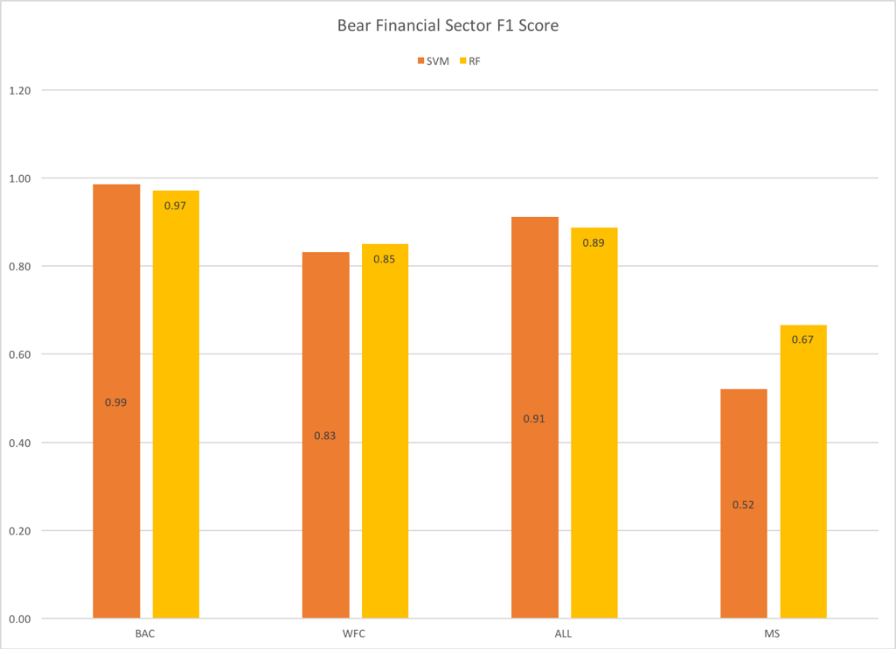


Figure 5. 2 illustrates the F1 score of the SVM and RF model for the Financial sector in the bear market.

5.1.2 Technology Sector

Unlike the Financial sector, the SVM model has a higher F1 score for 7 out of the 8 stocks in both markets. IBM has a (0.62) figure and Intel (0.65) in the bull market and contain a much higher F1 score. The RF provides lower scores for IBM (0.47) and Intel (0.56). When analysing the precision and recall rate for these stocks in the bull market, both of their values are high for recall and lower for precision, IBM recall (SVM 1, RF 1) and Intel (SVM 0.95, RF 0.94). Their precision scores though are varied, IBM (SVM 0.45, RF, 0.31) and Intel (SVM 0.95, RF 0.94). In both instances for these stocks, the recall provides a higher value, thus implying that the SVM model was better able to accurately predict the positive values. The RF model however, provided a higher F1 score overall for a stock in the bull market and a stock in the bear market. Analysing Cisco's F1 score for the SVM and RF display shows a small difference in performance level for the bear market with the SVM providing 0.76 score, and the RF at 0.78. The SVM's precision and recall score give the insight that the recall is lower a performing model at a 0.61 (lower than 1) score thus implying the model misclassified the false positives, while the RF model implies the same with a recall of 0.64, and a precision score of 0.98.

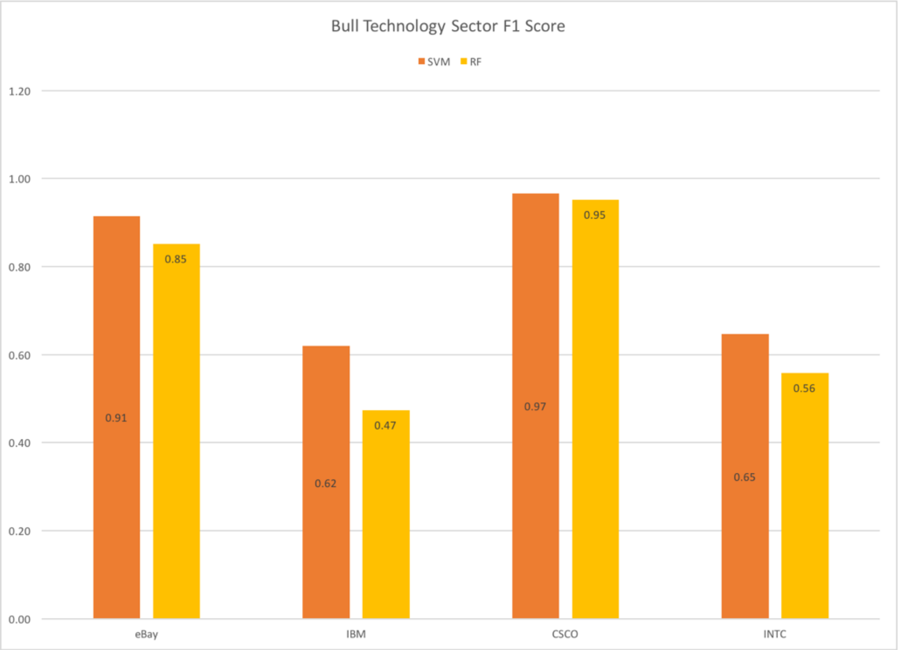


Figure 5. 3 illustrates the F1 score of the SVM and RF model for the Technology sector in the bull market.

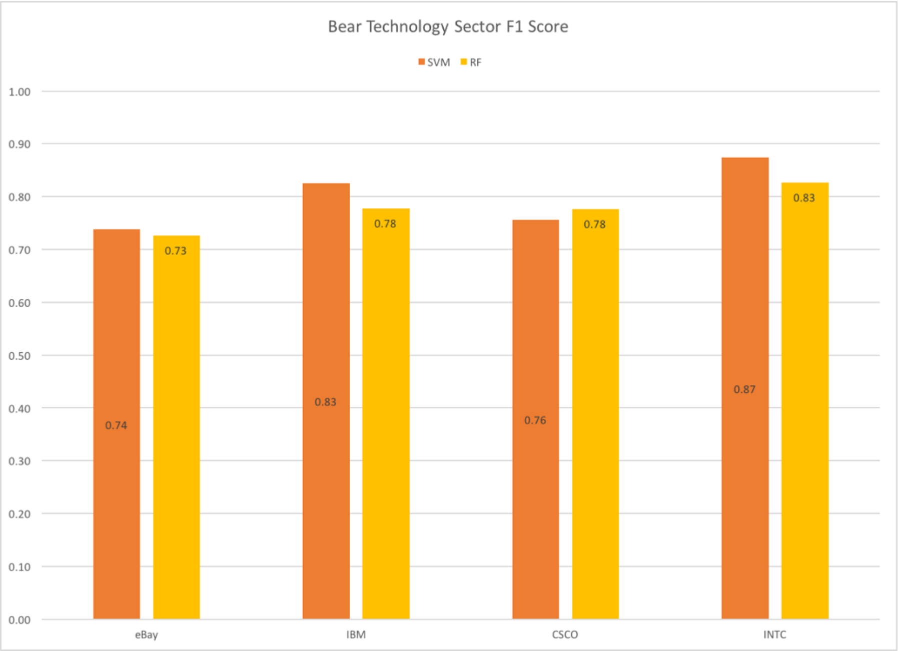


Figure 5. 4 illustrates the F1 score of the SVM and RF model for the Technology sector in the bear market.

5.1.3 Healthcare Sector

The Healthcare sector resulted with the most diverse model performances as two stocks, United Healthcare and Humana Inc., displayed 0 F1 scores for the bull market. United Healthcare, as seen in figure 5.1.3.1 has a 0 F1 score only for the SVM model, while the RF displays a 0.15 result. The precision and recall for the SVM in the bull market displays a 0 value for both, but a 0.91 value for accuracy. The RF model has a precision value of 0.14, but a recall of 0.17, thus leaving the F1 with a 0.15 value. The RF model was in turn able to provide a small amount of detection when analysing the overall selected positive values, however, both models still provided a poor performance level. The Humana Inc. stock in the bull market performed poorly overall for both SVM and RF with precision and recall resulting in 0 values for both models. This was alarming as this is the only 0 resulting stock for both models, a look back at the confusion matrix show that the models both resulted in forecasting the true negatives with the value of 1. This value in the bull market is what we are looking for and ideal as that means that the price is going up. The remaining stocks in the bull market were analysed and found to have a better performing F1 score for example in the SVM, Centene Corporation gave a result of 0.78 and both precision and recall were high values (0.69, 0.90), while the RF model contained a F1 score of 0.14, and precision and recall were unbalanced (0.08, and 1). While in the bear market the SVM and RF seem to have been predicating at a very similar rate, with the exception of Cigna. RF provides a higher F1 (0.85), and recall (0.96), but a lower precision (0.76), while the SVM presents a 0.61 F1 value with a higher precision (1), but a lower recall (0.43).

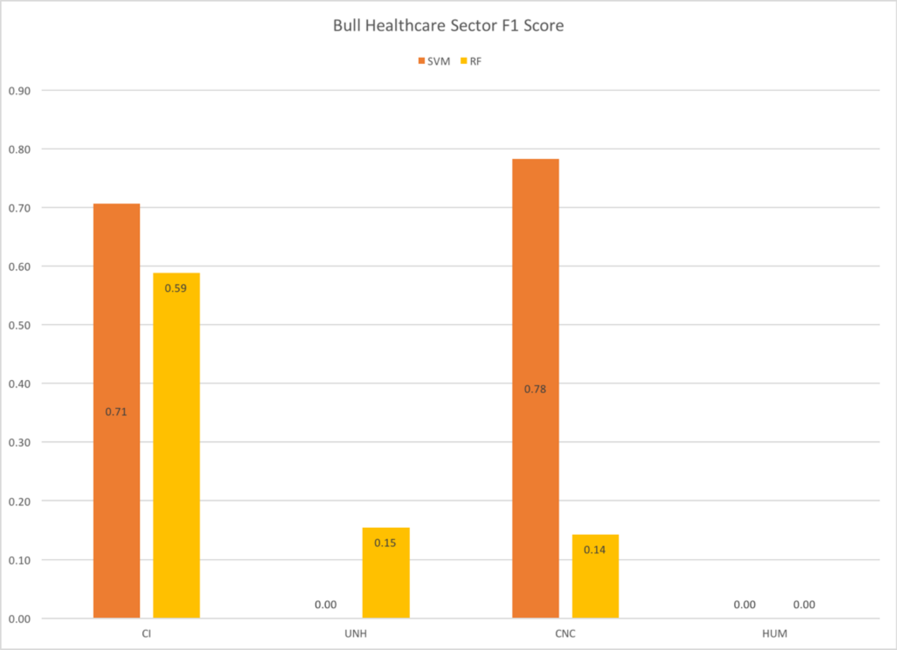


Figure 5. 5 illustrates the F1 score of the SVM and RF model for the Healthcare sector in the bull market.

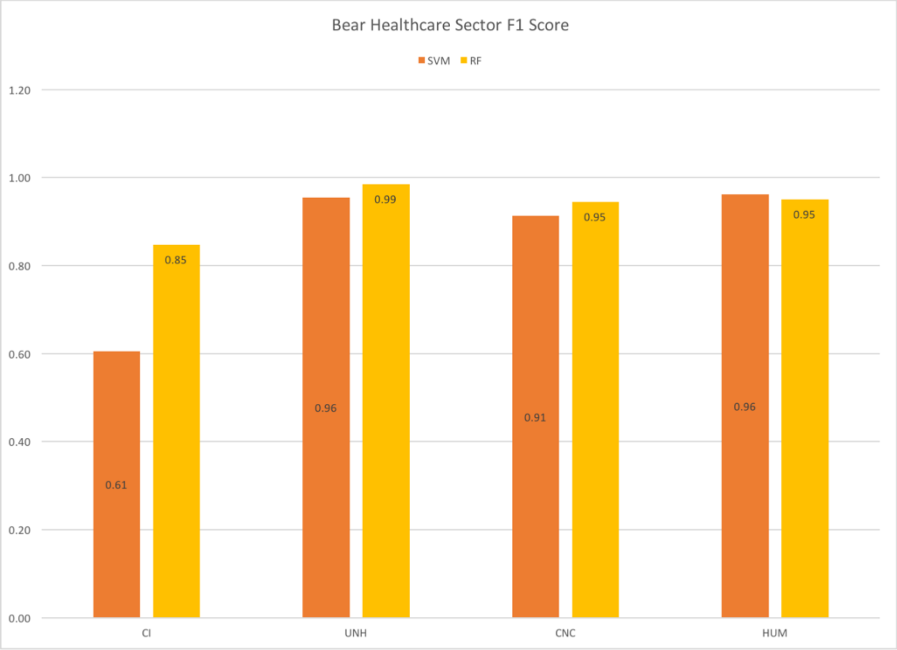


Figure 5. 6 illustrates the F1 score of the SVM and RF model for the Healthcare sector in the bear market.

5.1.4 Industrial Sector

Vastly different to the previous sector, the Industrial sector provides much higher results for both the SVM and RF model throughout the bull and bear market. The lowest recall within both sectors is seen in the General Dynamic stock in the bear market. The F1 scores are both rather low for each model with a SVM value of 0.87 and RF value of 0.87. When analysing their respective recall and precision values, the SVM contains a recall score of 0.92 and precision score of 0.73. The RF model provides a 0.91 recall score, and 0.63 precision score. Another stock that is particularly interesting is the 3M Company because the SVM and RF performed with identical results across both models.

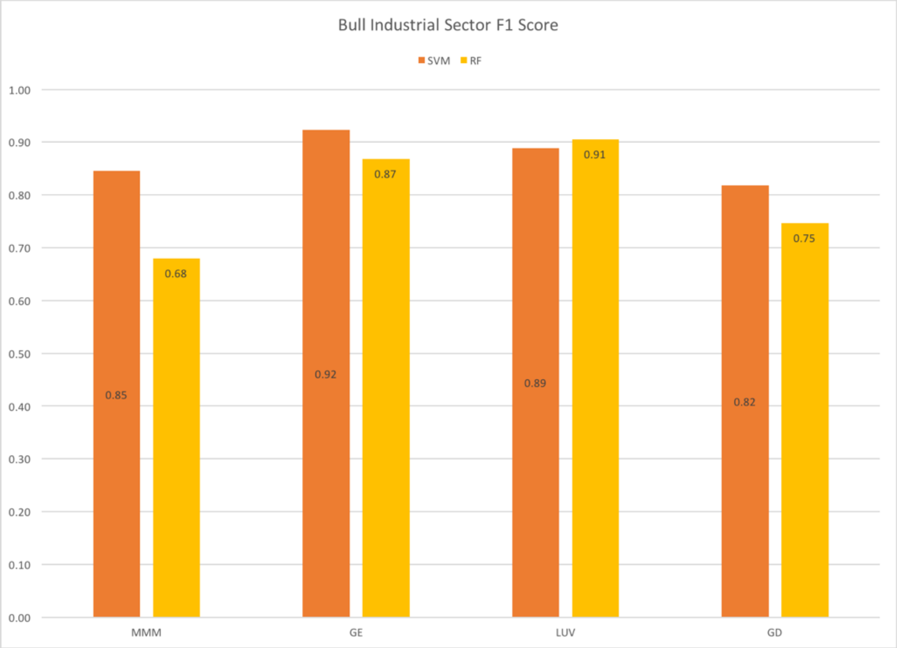


Figure 5. 7 illustrates the F1 score of the SVM and RF model for the Industrial sector in the bull market.

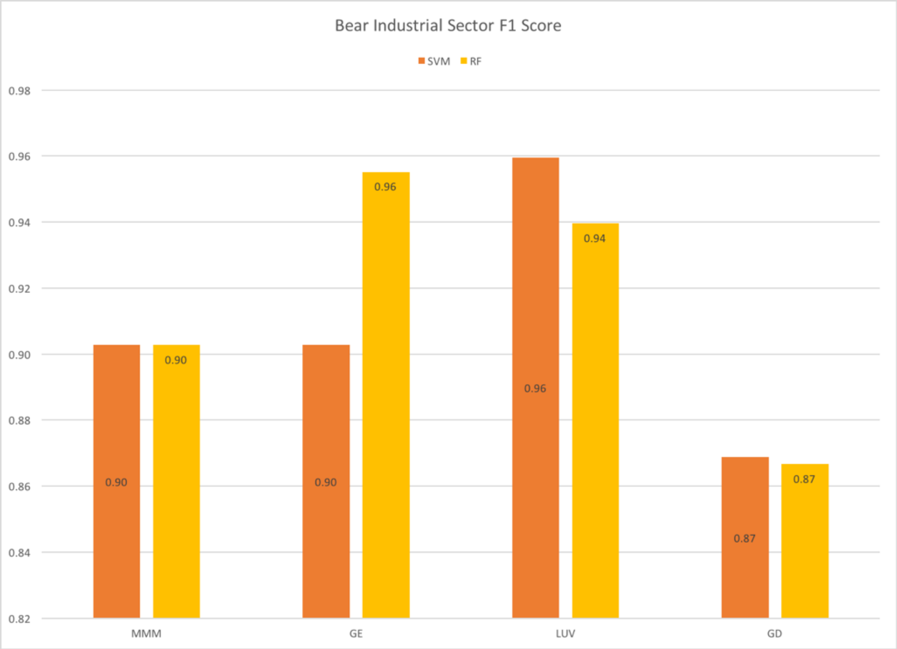
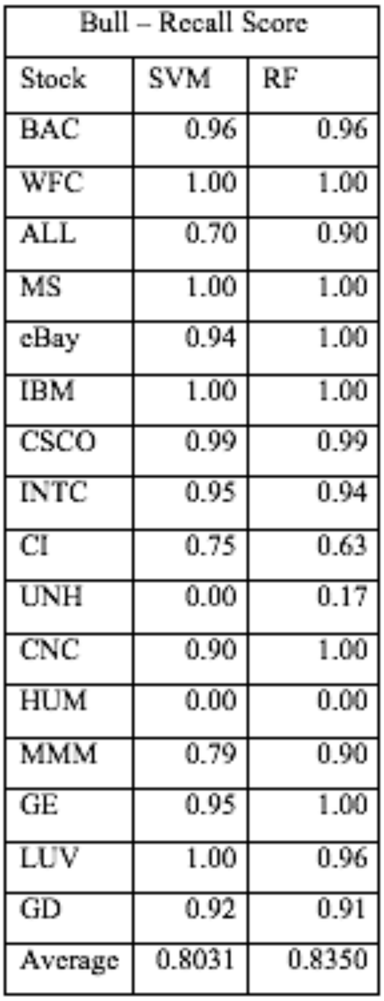


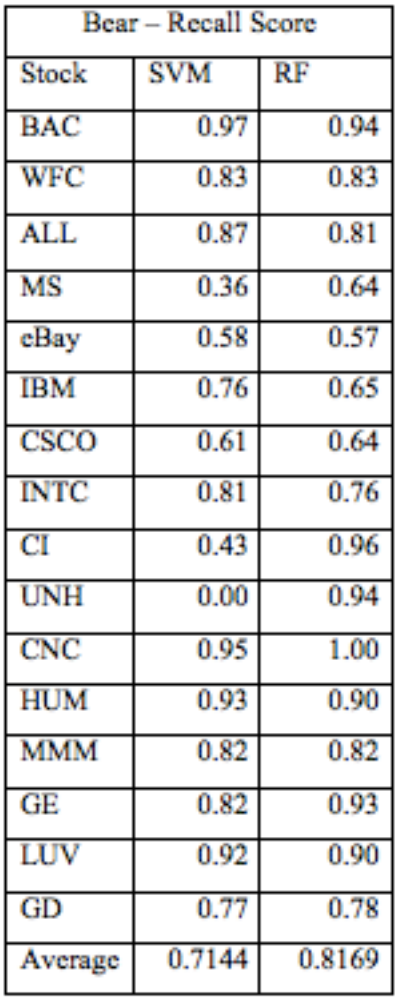
Figure 5. 8 illustrates the F1 score of the SVM and RF model for the Industrial sector in the bear market.

5.2 Analysis of Results

The results per sector displayed a varied range of F1 score performance between models, markets, stocks, and sectors. As the F1 score provides an equally weighted harmonic average of the recall and precision, an analysis of each stock in each market for each sector provided deep insight as to how the models performed. During the analysis stage, it was necessary to look at the precision and recall, as it was difficult for certain stocks to provide insight on the better performing stock with just the F1 score. Notably for this study, a trade-off between the precision and recall was required. As this study aims to provide realistic insight for future traders, the trade-off between recall and precision was necessary. The models were then analysed by determining the higher recall to specify accuracy. Initially when analysing tables 5.2.1 and 5.2.2 of the Recall score, the overall average for the SVM in the Bear market performed a at a lower level than the RF. In the Bull market, the overall average of the recall level for the RF was still higher than the SVM. The overall average of the F1 scores in each market showed that in the Bear market the SVM performed at average of 84% while the RF performed at an average of 87%.



Bull – Recall Score		
Stock	SVM	RF
BAC	0.96	0.96
WFC	1.00	1.00
ALL	0.70	0.90
MS	1.00	1.00
eBay	0.94	1.00
IBM	1.00	1.00
CSCO	0.99	0.99
INTC	0.95	0.94
CI	0.75	0.63
UNH	0.00	0.17
CNC	0.90	1.00
HUM	0.00	0.00
MMM	0.79	0.90
GE	0.95	1.00
LUV	1.00	0.96
GD	0.92	0.91
Average	0.8031	0.8350



Bear – Recall Score		
Stock	SVM	RF
BAC	0.97	0.94
WFC	0.83	0.83
ALL	0.87	0.81
MS	0.36	0.64
eBay	0.58	0.57
IBM	0.76	0.65
CSCO	0.61	0.64
INTC	0.81	0.76
CI	0.43	0.96
UNH	0.00	0.94
CNC	0.95	1.00
HUM	0.93	0.90
MMM	0.82	0.82
GE	0.82	0.93
LUV	0.92	0.90
GD	0.77	0.78
Average	0.7144	0.8169

Table 5. 1 displays the overall recall score for both markets for the SVM and RF model.

These results provide the results of the goal of this study; rejecting or accepting the hypothesis:

The support vector machine model, trained using the stock dataset, will perform at a higher accuracy than that of the random forest in each sector in both a bull and bear market.

After reviewing the individual F1 scores for each model along with the recall scores of each model, the SVM performed at an overall low in the bear market, while in the bull market the SVM performed at a higher accuracy. Therefore, this rejected the hypothesis that the SVM model would perform at a higher accuracy than the RF in both a bull and bear market for each sector. Although each model performed relatively well providing that both models did efficiently provide well performing predictive results.

5.3 Strengths and Limitations

The initial strength of this study provided insight into the high performance measures seen in the RF and SVM when applied to forecasting the direction of the stock price direction. The EMA of a 30-day lag day range, provided better results overall. Although a limitation of the EMA results in a less realistic and more artificial dataset when applied.

A second strength of this study specifies the ability to finely tune and train the SVM and RF model over a range of parameters per model. Each tuning stage was quick to produce and resulted in higher performing values providing the ability for replication when needed. Though one must be careful of the biased of data as with the SVM the model is prone to over fitting and under fitting.

5.4 Overview

This chapter evaluated the mechanisms of the model implementation stage and their respective results. Each of the sector's results for the F1 and recall score were analysed in depth to provide initial insight on the overall performance of the SVM and RF model in this experiment. The findings of the experiment were then discussed resulting in a rejection of the hypothesis as the RF model produced higher performing results. With these interpretations, a contribution of the strengths and limitations were then provided. The following chapter will provide a review of the entire study.

6 CHAPTER 6: CONCLUSION

This chapter will provide an overview of the study. It will go through the research aim and all the processes to gain insight on answering the question. This chapter will present the research overview/problem, design/implementation, evaluation and results, contributions and future work and recommendations.

6.1 Research Overview and Gap

As the stock market continues to fluctuate, investors and researchers continue to experiment and analyse the market to gain insight into the next trend. The overall goal for these individuals is to forecast the stock price, or stock price direction to obtain better knowledge on how the overall market will appear in days, weeks, month or years.

With such a vast set of literature presented in academia on stock forecasting, an under researched area was presented in the bull and bear market. The market trends provide a range of reasonings for cause of shift and volatility in the market. Researchers and investors specify a variety of trading approaches towards forecasting the stock trend. Once understanding the variety of strategies used in the forecasting the market trend, the area of focus was then set on implementing two machine learning models to predict stock price direction in 4 sectors in a bull and bear market. The literature review presented a range of machine learning models used to forecast both stock price movement, and stock price. A wide range of literature provided the insight that a SVM was a high performing machine learning algorithm when used to forecast stock movement, but an assortment of other studies conducted indicated that using decision tree models such as a random forest would be better. These models were analysed using a vast set of parameters such as fundamental indicators, and technical indicators. Though the range of time rarely focused on the stock market trend types such as the bear and bull.

Based on the literature review findings, the overall goal of the study was to *examine the accuracy level of an SVM versus an RF model in both a bull and bear market to forecast the stock price direction in 4 sectors*. By examining 16 stocks in 4 sectors, a diverse range of stock types were provided to analyse the trend in a bear and bull market. The use of 5 commonly utilized technical indicators in academia and by traders were the parameters that assisted in forecasting the trends. These parameters ranged from strength indicators to momentum indicators. Thus, providing a better insight in the market movement and assisting with the overall goal of forecasting using the SVM and RF to evaluate which model forecasts with a higher level of accuracy.

6.2 Design, Implementation, and Evaluation

The design and implementation of the overall research goal is presented in the following phases:

- Primarily reviewed and analyzed literature to grasp the ideologies behind the market trends, techniques towards forecasting the market trend, and machine learning integration.

- The dataset was initially built based on the intent of diversity to see the changes in the market from different sectors.
  The dataset was configured using a set of strict criteria to provide a level of balance in stock size.

- Due to the volatility and non-linearity of the stock prices, a smoothing tool (EMA) was applied as seen and referenced in a recent study.

- Implementation then took place to utilize the technical indicators, and formulate the stock return/direction.
- The models were then tuned, and executed.

- The confusion matrix was the metric used to analyze the overall performance of the models and strengths of the findings were then discussed.

Once the models were implemented a few measurements were ranged as the radial kernel performance of the SVM provided poor results while the linear model improved drastically thus providing more insight in as the literature primarily utilized the radial kernel. The models were then analysed using the confusion matrix performance measures. When comparing modes a range of stocks in each sector displayed close results for each model, but once precision and recall were reanalysed a further understanding was made. Since some of the results for the models provided the same results, a second form of measurement was to analyse the recall as a trade off was made between precision and recall. The overall performance though provided that RF and SVM both performed well, but SVM did not outperform the RF in both sectors.

6.3 Contributions

The experiment provided a varied set of contributes. The primary contribution was the higher levels of performance of both the SVM and RF rather than just one model. Thus implying that both models were able to handle the data and provide high accuracy efficiently.

When applying the correlation matrix, the analysis of correlation of technical indicators changed for each market and each stock. Therefore, instigating that the high correlation could of resulted in a more frequent change in price and volume or that the indicators would have been better used if mixed with some fundamental analysis.

The model did provide insight on the use of technical indicators when forecasting the trend. The models did provide a accurate forecast when analysing the trend change.

6.4 Future Work and Recommendations

The level of future work in this study is wide. There are many areas still that have been researched but have changed over time.

- The time frame can be analyzed to determine the effect on the market when analyzing technical analysis against fundamentals when there is a major dip in the market.
- An analysis can be done against the sectors using fundamental analysis rather then technical analysis to determine if a stock performs better in each market.

- The use of an ensemble model built using a few models such as the ANN, SVM and RF would be highly beneficial to see if the accuracy levels can be improved for some stocks.
- Analyzing the effect of more technical indicators to determine which technical indicators are actually valid as the ranges are varied in terms of which are utilized by investors and which are utilized by researchers.

7 BIBLIOGRAPHY

Abarbanell, J. S., & Bushee, B. J. (1997). Fundamental Analysis, Future Earnings, and Stock Prices. *Journal of Accounting Research*, *35*(1), 1. https://doi.org/10.2307/2491464

Adhikari, R., & Agrawal, R. K. (2011). A Homogenous Ensemble of Artifical Neural Networks for Time Series Forecasting. *International Journal of Computer Applications*, *32*(7), 1–7.

Anbalagan, T., & Maheswari, S. U. (2015). Classification and Prediction of Stock Market Index Based on Fuzzy Metagraph. *Procedia Computer Science*, *47*, 214–221. https://doi.org/10.1016/j.procs.2015.03.200

Ao, S. I., & International Association of Engineers (Eds.). (2010). *The 2010 IAENG International Conference on Artificial Intelligence and Applications, the 2010 IAENG International Conference on Bioinformatics, the 2010 IAENG International Conference on Computer Science, the 2010 IAENG International Conference on Data Mining and Applications, the 2010 IAENG International Conference on Internet Computing and Web Services, the 2010 IAENG International Conference on Software Engineering*. Hong Kong: IAENG.

Arik, S., Eryilmaz, S. B., & Goldberg, A. (2014). Supervised classification-based stock prediction and portfolio optimization. *arXiv Preprint arXiv:1406.0824*.

Arratia, A. (2014). *Computational Finance*. 29, avenue Laumiére 75019 Paris, France: Atlantis Press.

Atsalakis, G. S., & Valavanis, K. P. (2009a). Forecasting stock market short-term trends using a neuro-fuzzy based methodology. *Expert Systems with Applications*, *36*(7), 10696–10707. https://doi.org/10.1016/j.eswa.2009.02.043

Atsalakis, G. S., & Valavanis, K. P. (2009b). Surveying stock market forecasting techniques – Part II: Soft computing methods. *Expert Systems with Applications*, *36*(3), 5932–5941. https://doi.org/10.1016/j.eswa.2008.07.006

Austin, M. P., Bates, G., Dempster, M. A. H., Leemans, V., & Williams, S. N. (2004). Adaptive systems for foreign exchange trading. *Quantitative Finance*, *4*, C37–C45. https://doi.org/10.1080/14697680400008593

Baba, N., & Kozaki, M. (1992). An Intelligent Forecasting System of Stock Price Using Neural Networks. *IEEE*, I-371-I377.

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *The Journal of Economic Perspectives*, *21*(2), 129–151.

Bernard, V. L., & Thomas, J. K. (1990). Evidence that stock prices do not fully reflect the implications of current earnings for future earnings. *Journal of Accounting and Economics*, *13*(4), 305–340.

Bikhchandani, S., & Sharma, S. (2000). Herd behavior in financial markets. *IMF Staff Papers*, 279–310.

Bouchaud, J.-P., & Cont, R. (1998). A Langevin Approach to Stock Market Fluctuations and Crashes. *The European Physical Journal B*, *6*(4), 543–550. https://doi.org/10.1007/s100510050582

Cao, L. J., & Tay, E. H. (2003). Support Vector Machine With Adaptive Parameters in Financial Time Series Forecasting. *IEE Transactions on Neural Networks*, *14*(6), 1506–1518. https://doi.org/10.1109/TNN.2003.82.0556

Capocci, D., Corhay, A., & Hübner, G. (2005). Hedge fund performance and persistence in bull and bear markets. *The European Journal of Finance*, *11*(5), 361–392.

Carbonell, J. G., Michalski, R. S., & Mitchell, T. M. (1983). *Machine Learning An Artificial Intelligence Approach*. Palo Alto, California: TIOGA Publishing Co.

Chai, J., Du, J., Lai, K. K., & Lee, Y. P. (2015). A Hybrid Least Square Support Vector Machine Model with Parameters Optimization for Stock Forecasting. *Mathematical Problems in Engineering*, *2015*, 1–7. https://doi.org/10.1155/2015/231394

Chan, L. K. C., & Lakonishok, J. (1995). The Behavior of Stock Prices Around Institutional Trades. *The Journal of Finance*, *50*(4), 1147–1174.

Cheng, C.-H., Chen, T.-L., & Wei, L.-Y. (2010). A hybrid model based on rough sets theory and genetic algorithms for stock price forecasting. *Information Sciences*, *180*(9), 1610–1629. https://doi.org/10.1016/j.ins.2010.01.014

Christie, W. G., & Huang, R. D. (1995). Following the Pied Piper: Do Individual Returns Herd around the Market? *Financial Analysts Journal*, *51*(4), 31–37.

Clement, M. B., & Tse, S. Y. (2005). Financial analyst characteristics and herding behavior in forecasting. *The Journal of Finance*, *60*(1), 307–341.

Cont, R., & Bouchaud, J.-P. (2000). Herd Behavior and Aggregate Fluctuations in Financial Markets. *Macroeconomic Dynamics*, *4*, 170–196.

Day, R. H., & Huang, W. (1990). Bulls, bears and market sheep. *Journal of Economic Behavior & Organization*, *14*(3), 299–329.

De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Positive Feedback Investment Strategies and Destabilizing Rational Speculation. *The Journal of Finance*, *45*(2), 379–395. https://doi.org/10.1111/j.1540-6261.1990.tb03695.x

Enke, D., & Thawornwong, S. (2005). The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with Applications*, *29*(4), 927–940. https://doi.org/10.1016/j.eswa.2005.06.024

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, *25*(2), 383. https://doi.org/10.2307/2325486

Gavrilov, M., Anguelov, D., Indyk, P., & Motwani, R. (2000). Mining the stock market (extended abstract): which measure is best? In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 487–496). ACM.

gompers and metrick.pdf. (n.d.).

Gompers, P. A., & Metrick, A. (1998). *Institutional Investors and Equity Prices* (NBER Working Paper Series No. 6723). Cambridge, MA 02132: National Bureau of Economic Research. Retrieved from http://wwww.nber.org/papers/w6723

Hamao, Y., Masulis, R. W., & Ng, V. (1990). Correlations in Price Changes and Volatility across International Stock Markets. *The Review of Financial Studies Oxford University Press*, *3*(2), 281–307.

Handa, R., Hota, H. S., & Tandan, S. R. (2015). Stock Market Prediction with Various Technical Indicators Using Neural Network Techniques. *International Journal for Research in Applied Science and Engineering Technology*, *3*(VI), 603–608.

Hou, K., & Robinson, D. T. (2006). Industry concentration and average stock returns. *The Journal of Finance*, *61*(4), 1927–1956.

Hsu, S.-H., Hsieh, J. P.-A., Chih, T.-C., & Hsu, K.-C. (2009). A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression. *Expert Systems with Applications*, *36*(4), 7947–7951. https://doi.org/10.1016/j.eswa.2008.10.065

Huang, W., Nakamori, Y., & Wang, S.-Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, *32*(10), 2513–2522. https://doi.org/10.1016/j.cor.2004.03.016

Imandoust, S., & Bolandraftar, M. (2014). Forecasting the direction of stock market index movement using three data mining techniques: the case of Tehran Stock Exchange. *International Journal of Engineering Research and Applications*, *4*, 106–117.

Kaastra, I., & Boyd, M. (1996). Designing a neural network for forecasting financial and economic time series. *Elsevier Neurocomputing*, *10*, 215–236.

Kannan, S. K., Sekar, S. P., Sathik, M. M., & Arumugam, P. (2010). Financial Stock Market Forecast using Data Mining Techniques. *Proceedings of The International MultiConference of Engineers and Computer Scientist*, *1*(IMECS), 17–19.

Kara, Y., Acar Boyacioglu, M., & Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert Systems with Applications*, *38*(5), 5311–5319. https://doi.org/10.1016/j.eswa.2010.10.027

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics Algorithms, Worked Examples, and Case Studies*. Cambridge, Massachusetts, London, England: MIT Press.

Khaidem, L., Saha, S., & Dey, S. R. (2016). Predicting the direction of stock market prices using random forest. *arXiv Preprint arXiv:1605.00003*.

Khedkar, A., & Argiddi, R. V. (2013). To Study and Analyze to Foresee Market using Data Mining Technique. *International Journal of Engineering Trends and Technology (IJETT)*, *4*(9), 3718–3720.

Kim, K. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, *55*(1–2), 307–319. https://doi.org/10.1016/S0925-2312(03)00372-2

Kim, K., & Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Systems with Applications*, *19*(2), 125–132.

Lariviere, B., & Vandenpoel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, *29*(2), 472–484. https://doi.org/10.1016/j.eswa.2005.04.043

Lu, C.-J., Lee, T.-S., & Chiu, C.-C. (2009a). Financial time series forecasting using independent component analysis and support vector regression. *Decision Support Systems*, *47*(2), 115–125.

Lu, C.-J., Lee, T.-S., & Chiu, C.-C. (2009b). Financial time series forecasting using independent component analysis and support vector regression. *Decision Support Systems*, *47*(2), 115–125.

Lunde, A., & Timmermann, A. (2004). Duration Dependence in Stock Prices: An Analysis of Bull and Bear Markets. *Journal of Business & Economic Statistics*, *22*(3), 253–273. https://doi.org/10.1198/073500104000000136

Maheu, J. M., & McCurdy, T. H. (2000). Identifying Bull and Bear Markets in Stock Returns. *Journal of Business & Economic Statistics*, *18*(1), 100–112.

McNamara, S. (2016). Identifying Market Indicators and Content Quality from a Financial Micro-Blog Platform.

Mittermayer, M.-A. (2004). Forecasting intraday stock price trends with text mining techniques. In *system sciences, 2004. proceedings of the 37th annual hawaii international conference on* (p. 10–pp). IEEE.

Mukherjee, P., Bose, S., & Coondoo, D. (2002). Foreign institutional investment in the Indian equity market: An analysis of daily flows during January 1999-May 2002.

Nann, S., Krauss, J., & Schoder, D. (2013). Predictive analytics on public data-the case of stock markets.

Neely, C., Weller, P., & Dittmar, R. (1997). Is technical analysis in the foreign exchange market profitable? A genetic programming approach. *Journal of Financial and Quantitative Analysis*, *32*(4), 405–426.

Nikfarjam, A., Emadzadeh, E., & Muthaiyah, S. (2010). Text mining approaches for stock market prediction (pp. 256–260). IEEE. https://doi.org/10.1109/ICCAE.2010.5451705

Ou, P., & Wang, H. (2009). Prediction of stock market index movement by ten data mining techniques. *Modern Applied Science*, *3*(12), 28.

Pagan, A. R., & Sossounov, K. A. (2003). A simple framework for analysing bull and bear markets. *Journal of Applied Econometrics*, *18*(1), 23–46. https://doi.org/10.1002/jae.664

Pai, P.-F., & Lin, C.-S. (2005). A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega*, *33*(6), 497–505. https://doi.org/10.1016/j.omega.2004.07.024

Prado, H. A. do, Ferneda, E., Morais, L. C. R., Luiz, A. J. B., & Matsura, E. (2013). On the Effectiveness of Candlestick Chart Analysis for the Brazilian Stock Market. *Procedia Computer Science*, *22*, 1136–1145. https://doi.org/10.1016/j.procs.2013.09.200

Prechter, Jr., R. R. (2001). Unconscious Herding Behavior as the Psychological Basis of Financial Market Trends and Patterns. *The Journal of Psychology and Financial Markets*, *2*(3), 120–125.

Roberts, H. V. (1959). Stock-Market "Patterns" and Financial Analysis: Methodological Suggestions. *The Journal of Finance*, *14*(1), 1–10.

S1877050913009939.html. (n.d.).

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems*, *27*(2), 1–19. https://doi.org/10.1145/1462198.1462204

Shiller, R. J. (1987). *Investor behavior in the October 1987 stock market crash: Survey evidence*. National Bureau of Economic Research Cambridge, Mass., USA.

S&P 500. (2018, January 3). Retrieved January 3, 2018, from https://fred.stlouisfed.org/series/SP500

S&P Dow Jones. (2017). *S&P U.S Indices Methodology*. Retrieved from http://eu.spindices.com/indices/equity/sp-500

Tsaih, R., Hsu, Y., & Lai, C. C. (1998). Forecasting S&P 5oo stock index futures with a hybrid AI system. *Elsevier Decision Support System*, (23), 161–174.

Vijh, A. (1994). S&P 500 Trading Strategies and Stock Betas. *The Review of Financial Studies*, *7*(1), 215–251.

Weinstein, S. (1988). *Secrets for Profiting in Bull and Bear Markets*. Homewood, Illinois 60430: Dow Jones-Irwin.

Wolpert, D. H., & Macready, W. G. (1997). No Free Lunch Theorems for Optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82. https://doi.org/10.1109/4235.585893

Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., & Zhang, J. (1998). Daily stock market forecast from textual web data. In *Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on* (Vol. 3, pp. 2720–2725). IEEE.

Yang, D.-L., Hsieh, Y. L., & Wu, J. (2006). Using Data Mining to Study Upstream and Downstream Causal Relationship in Stock Market. Atlantis Press. https://doi.org/10.2991/jcis.2006.191

Zhang, P. G. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Elsevier Neurocomputing*, *50*, 159–175.

(n.d.).

8 APPENDIX A

~~~
---
title: "Bear Market"
author: "Tiffany Razy"
date: "January 3, 2018"
output: word_document
---
## Installing Packages and Libraries
```{r}
#suppressMessages(install.packages("rmarkdown"))
#suppressMessages(install.packages("quantmod"))
#suppressMessages(install.packages('TTR'))
#suppressMessages(install.packages('ggplot2'))
#suppressMessages(install.packages('reshape2'))
#suppressMessages(install.packages('zoo'))
#suppressMessages(install.packages('caret'))
#suppressMessages(install.packages('xts') )
#suppressMessages(install.packages('forecast'))
#suppressMessages(install.packages('stats'))
#suppressMessages(install.packages('e1071'))
#suppressMessages(install.packages('timeSeries'))
#suppressMessages(install.packages('tidyverse'))
#suppressMessages(install.packages('GGally'))
suppressMessages(library('quantmod'))
suppressMessages(library('timeSeries'))
suppressMessages(library('pdfetch'))
suppressMessages(library('xts'))
suppressMessages(library('TTR'))
suppressMessages(library('ggplot2'))
suppressMessages(library('caret'))
suppressMessages(library('forecast'))
suppressMessages(library('randomForest'))
~~~

~~~
suppressMessages(library('zoo'))
suppressMessages(library('markdown'))
suppressMessages(library('ROCR'))
suppressMessages(library('corrplot'))
suppressMessages(library('e1071'))
suppressMessages(library('GGally'))
suppressMessages(library('reshape'))
```
~~~

## Retreiving Stock Data from Yahoo (Bear Market Only 26 Months in total)

~~~
```{r}
#Financial Sector (BEAR)
bankofamericabear <- pdfetch_YAHOO("BAC", fields = c("open", "high", "low", "close", "adjclose", "volume"), from = as.Date("2007-01-01"), to = as.Date("2009-03-31"), interval = "1d")
bankofamericabear <-data.frame(bankofamericabear)
bankofamericabear <- cbind(date = rownames(bankofamericabear), bankofamericabear)
bankofamericabear$date<-as.Date(bankofamericabear$date)
colnames(bankofamericabear)<- c("date", "bacbear.open","bacbear.high","bacbear.low", "bacbear.close", "bacbear.adjclose", "bacbear.volume")

wellsfargoandcompanybear <- pdfetch_YAHOO("WFC", fields = c("open", "high", "low", "close", "adjclose", "volume"), from = as.Date("2007-01-01"), to = as.Date("2009-03-31"), interval = "1d")
wellsfargoandcompanybear <-data.frame(wellsfargoandcompanybear)
wellsfargoandcompanybear <- cbind(date = rownames(wellsfargoandcompanybear), wellsfargoandcompanybear)
wellsfargoandcompanybear$date<-as.Date(wellsfargoandcompanybear$date)
~~~

```
colnames(wellsfargoandcompanybear)<- c("date",
"wfcbear.open","wfcbear.high","wfcbear.low", "wfcbear.close", "wfcbear.adjclose",
"wfcbear.volume")

allstatecorporationbear <- pdfetch_YAHOO("ALL", fields = c("open", "high", "low",
"close", "adjclose", "volume"), from = as.Date("2007-01-01"), to = as.Date("2009-03-
31"), interval = "1d")
allstatecorporationbear <-data.frame(allstatecorporationbear)
allstatecorporationbear <- cbind(date = rownames(allstatecorporationbear),
allstatecorporationbear)
allstatecorporationbear$date<-as.Date(allstatecorporationbear$date)
colnames(allstatecorporationbear)<- c("date",
"allbear.open","allbear.high","allbear.low", "allbear.close", "allbear.adjclose",
"allbear.volume")

morganstanleybear  <- pdfetch_YAHOO("MS", fields = c("open", "high", "low",
"close", "adjclose", "volume"), from = as.Date("2007-01-01"), to = as.Date("2009-03-
31"), interval = "1d")
morganstanleybear <-data.frame(morganstanleybear)
mogranstanleybull <- cbind(date = rownames(morganstanleybear), morganstanleybear)
morganstanleybear <- cbind(date = rownames(morganstanleybear), morganstanleybear)
morganstanleybear$date<-as.Date(morganstanleybear$date)
colnames(morganstanleybear)<-  c("date","msbear.open","msbear.high","msbear.low",
"msbear.close", "msbear.adjclose", "msbear.volume")

#Technology Sector (BEAR)
ebayincbear  <- pdfetch_YAHOO("EBAY", fields = c("open", "high", "low", "close",
"adjclose", "volume"), from = as.Date("2007-01-01"), to = as.Date("2009-03-31"),
interval = "1d")
ebayincbear <-data.frame(ebayincbear)
ebayincbear <- cbind(date = rownames(ebayincbear), ebayincbear)
ebayincbear$date<-as.Date(ebayincbear$date)
colnames(ebayincbear)<-  c("date",  "ebaybear.open","ebaybear.high","ebaybear.low",
"ebaybear.close", "ebaybear.adjclose", "ebaybear.volume")
```

```
ibmbear <- pdfetch_YAHOO("IBM", fields = c("open", "high", "low", "close",
"adjclose", "volume"), from = as.Date("2007-01-01"), to = as.Date("2009-03-31"),
interval = "1d")
ibmbear <-data.frame(ibmbear)
ibmbear <- cbind(date = rownames(ibmbear), ibmbear)
ibmbear$date<-as.Date(ibmbear$date)
colnames(ibmbear)<- c("date", "ibmbear.open","ibmbear.high","ibmbear.low",
"ibmbear.close", "ibmbear.adjclose", "ibmbear.volume")

ciscosystemsincbear <- pdfetch_YAHOO("CSCO", fields = c("open", "high", "low",
"close", "adjclose", "volume"), from = as.Date("2007-01-01"), to = as.Date("2009-03-
31"), interval = "1d")
ciscosystemsincbear <-data.frame(ciscosystemsincbear)
ciscosystemsincbear <- cbind(date = rownames(ciscosystemsincbear),
ciscosystemsincbear)
ciscosystemsincbear$date<-as.Date(ciscosystemsincbear$date)
colnames(ciscosystemsincbear)<- c("date",
"cscobear.open","cscobear.high","cscobear.low", "cscobear.close",
"cscobear.adjclose", "cscobear.volume")

intelbear <- pdfetch_YAHOO("INTL", fields = c("open", "high", "low", "close",
"adjclose", "volume"), from = as.Date("2007-01-01"), to = as.Date("2009-03-31"),
interval = "1d")
intelbear <-data.frame(intelbear)
intelbear <- cbind(date = rownames(intelbear), intelbear)
intelbear$date<-as.Date(intelbear$date)
colnames(intelbear)<- c("date", "intlbear.open","intlbear.high","intlbear.low",
"intlbear.close", "intlbear.adjclose", "intlbear.volume")

#Healthcare Sector (BEAR)
cignacorporationbear <- pdfetch_YAHOO("CI", fields = c("open", "high", "low",
"close", "adjclose", "volume"), from = as.Date("2007-01-01"), to = as.Date("2009-03-
31"), interval = "1d")
```

```
cignacorporationbear <-data.frame(cignacorporationbear)
cignacorporationbear <- cbind(date = rownames(cignacorporationbear),
cignacorporationbear)
cignacorporationbear$date<-as.Date(cignacorporationbear$date)
colnames(cignacorporationbear)<- c("date", "cibear.open","cibear.high","cibear.low",
"cibear.close", "cibear.adjclose", "cibear.volume")
```

```
unitedhealthgroupincorporatedbear <- pdfetch_YAHOO("UNH", fields = c("open",
"high", "low", "close", "adjclose", "volume"), from = as.Date("2007-01-01"), to =
as.Date("2009-03-31"), interval = "1d")
unitedhealthgroupincorporatedbear <-data.frame(unitedhealthgroupincorporatedbear)
unitedhealthgroupincorporatedbear <- cbind(date =
rownames(unitedhealthgroupincorporatedbear), unitedhealthgroupincorporatedbear)
unitedhealthgroupincorporatedbear$date<-
as.Date(unitedhealthgroupincorporatedbear$date)
colnames(unitedhealthgroupincorporatedbear)<- c("date",
"unhbear.open","unhbear.high","unhbear.low","unhbear.close", "unhbear.adjclose",
"unhbear.volume")
```

```
centenecorporationincbear <- pdfetch_YAHOO("CNC", fields = c("open", "high",
"low", "close", "adjclose", "volume"), from = as.Date("2007-01-01"), to =
as.Date("2009-03-31"), interval = "1d")
centenecorporationincbear <-data.frame(centenecorporationincbear)
centenecorporationincbear <- cbind(date = rownames(centenecorporationincbear),
centenecorporationincbear)
centenecorporationincbear$date<-as.Date(centenecorporationincbear$date)
colnames(centenecorporationincbear)<- c("date",
"cncbear.open","cncbear.high","cncbear.low","cncbear.close", "cncbear.adjclose",
"cncbear.volume")
```

```
humanaincbear <- pdfetch_YAHOO("HUM", fields = c("open", "high", "low", "close",
"adjclose", "volume"), from = as.Date("2007-01-01"), to = as.Date("2009-03-31"),
interval = "1d")
```

```
humanaincbear <-data.frame(humanaincbear)
humanaincbear <- cbind(date = rownames(humanaincbear), humanaincbear)
humanaincbear$date<-as.Date(humanaincbear$date)
colnames(humanaincbear)<- c("date",
"humbear.open","humbear.high","humbear.low","humbear.close", "humbear.adjclose",
"humbear.volume")


#Industrial Sector (BEAR)
threemcompanybear <- pdfetch_YAHOO("MMM", fields = c("open", "high", "low",
"close", "adjclose", "volume"), from = as.Date("2007-01-01"), to = as.Date("2009-03-
31"), interval = "1d")
threemcompanybear <-data.frame(threemcompanybear)
threemcompanybear <- cbind(date = rownames(threemcompanybear),
threemcompanybear)
threemcompanybear$date<-as.Date(threemcompanybear$date)
colnames(threemcompanybear)<- c("date",
"mmmbear.open","mmmbear.high","mmmbear.low","mmmbear.close",
"mmmbear.adjclose", "mmmbear.volume")


generalelectriccompanybear <- pdfetch_YAHOO("GE", fields = c("open", "high",
"low", "close", "adjclose", "volume"), from = as.Date("2007-01-01"), to =
as.Date("2009-03-31"), interval = "1d")
generalelectriccompanybear <-data.frame(generalelectriccompanybear)
generalelectriccompanybear <- cbind(date = rownames(generalelectriccompanybear),
generalelectriccompanybear)
generalelectriccompanybear$date<-as.Date(generalelectriccompanybear$date)
colnames(generalelectriccompanybear)<- c("date",
"gebear.open","gebear.high","gebear.low","gebear.close", "gebear.adjclose",
"gebear.volume")
```

~~~~
southwestairlinescompanybear <- pdfetch_YAHOO("LUV", fields = c("open", "high",
"low", "close", "adjclose", "volume"), from = as.Date("2007-01-01"), to =
as.Date("2009-03-31"), interval = "1d")
southwestairlinescompanybear <-data.frame(southwestairlinescompanybear)
southwestairlinescompanybear <- cbind(date =
rownames(southwestairlinescompanybear), southwestairlinescompanybear)
southwestairlinescompanybear$date<-as.Date(southwestairlinescompanybear$date)
colnames(southwestairlinescompanybear)<- c("date",
"luvbear.open","luvbear.high","luvbear.low","luvbear.close", "luvbear.adjclose",
"luvbear.volume")

generaldynamicscorporationbear <- pdfetch_YAHOO("GD", fields = c("open", "high",
"low", "close", "adjclose", "volume"), from = as.Date("2007-01-01"), to =
as.Date("2009-03-31"), interval = "1d")
generaldynamicscorporationbear <-data.frame(generaldynamicscorporationbear)
generaldynamicscorporationbear <- cbind(date =
rownames(generaldynamicscorporationbear), generaldynamicscorporationbear)
generaldynamicscorporationbear$date<-
as.Date(generaldynamicscorporationbear$date)
colnames(generaldynamicscorporationbear)<- c("date",
"gdbear.open","gdbear.high","gdbear.low","gdbear.close", "gdbear.adjclose",
"gdbear.volume")

```

##Visulisation of Close for all of the Bear Stocks

```{r}
##Visual
stocks <- data.frame(bankofamericabear = bankofamericabear[, "bacbear.close"],
wellsfargoandcompanybear = wellsfargoandcompanybear[, "wfcbear.close"],
            allstatecorporationbear = allstatecorporationbear[, "allbear.close"],
morganstanleybear= morganstanleybear[,"msbear.close"], ebayincbear
=ebayincbear[,"ebaybear.close"],ibmbear =
~~~~

~~~
ibmbear[,"ibmbear.close"],ciscosystemsincbear
=ciscosystemsincbear[,"cscobear.close"], intelbear = intelbear[,"intlbear.close"],
cignacorporationbear = cignacorporationbear[,"cibear.close"],
unitedhealthgroupincorporatedbear =
unitedhealthgroupincorporatedbear[,"unhbear.close"], humanaincbear =
humanaincbear[,"humbear.close"], centenecorporationincbear=
centenecorporationincbear[,"cncbear.close"] , threemcompanybear=
threemcompanybear[,"mmmbear.close"], generalelectriccompanybear =
generalelectriccompanybear[,"gebear.close"],southwestairlinescompanybear =
southwestairlinescompanybear[,"luvbear.close"], generaldynamicscorporationbear =
generaldynamicscorporationbear[,"gdbear.close"])

stocks$date<-bankofamericabear[,"date"]
stocks1<-xts(stocks[,-17], order.by=stocks$date)



#rain<-rainbow(ncol(stocks1))
#plot(x=stocks1, col=rain, main="Bear Close Price for all Sectors")
#legend("right",inset=-.05, title="Stocks", c("BAC", "WFC", "ALL", "MS", "EBAY", "
IMB", "CSCO", "INTL", "CI", "UNH", "HUM","CNC", "MMM", "GE", "LUV",
"GD"), fill=rain, xpd = TRUE,cex=.5, box.lty=0)
#par(xpd = T, mar = par()$mar + c(0,0,0,7))



```



## Converting files to zoo formate from dataframe and zoo formate from
```{r}

#------------------Finance
~~~

~~~
#bear_xts<-xts(bankofamericabear[,-1],order.by = bankofamericabear$date)
#bear_xts<-xts(wellsfargoandcompanybear[,-1],order.by =
wellsfargoandcompanybear$date)
#bear_xts<-xts(allstatecorporationbear[,-1],order.by = allstatecorporationbear$date)
#bear_xts<-xts(morganstanleybear[,-1],order.by = morganstanleybear$date)
#------------------Technology
#bear_xts<-xts(ebayincbear[,-1],order.by = ebayincbear$date)
#bear_xts<-xts(intelbear[,-1],order.by = intelbear$date)
#bear_xts<-xts(ciscosystemsincbear[,-1],order.by = ciscosystemsincbear$date)
#bear_xts<-xts(ibmbear[,-1],order.by = ibmbear$date)
#------------------Healthcare
#bear_xts<-xts(cignacorporationbear[,-1],order.by = cignacorporationbear$date)
#bear_xts<-xts(unitedhealthgroupincorporatedbear[,-1],order.by =
unitedhealthgroupincorporatedbear$date)
#bear_xts<-xts(humanaincbear[,-1],order.by = humanaincbear$date)
#bear_xts<-xts(centenecorporationincbear[,-1],order.by =
centenecorporationincbear$date)
#------------------Industrail
bear_xts<-xts(threemcompanybear[,-1],order.by = threemcompanybear$date)
#bear_xts<-xts(generalelectriccompanybear[,-1],order.by =
generalelectriccompanybear$date)
#bear_xts<-xts(southwestairlinescompanybear[,-1],order.by =
southwestairlinescompanybear$date)
#bear_xts<-xts(generaldynamicscorporationbear[,-1],order.by =
generaldynamicscorporationbear$date)

```


##Candlestick and graphs
```{r}


#chartSeries(bear_xts[,4],name='Bear Close', type=("candlestick"), subset=NULL,
show.grid=TRUE,time.scale = NULL, bar.type="ohlc",
TA=c('addVo()','addBBands()'))
~~~

~~~~
#chartSeries(bear_xts,name='Bear',type='candle',subset='2007-03::2007-06',up.col = "black", dn.col = "red",theme="white", bar.type="ohlc")


#plot(bear_xts[,4], main = "Close Price Bear")

```


## Split into Training and Test for Bear Market

```{r}

train_bear<-window(bear_xts, start=as.Date("2007-01-01"), end=as.Date("2008-10-15"))
#head(train_bear,n=30)
#str(train_bear)


test_bear<-window(bear_xts, start=as.Date("2008-10-16"), end=as.Date("2009-03-31"))
#head(test_bear)
#str(test_bear)

```


####------------------------------------------------------TRAIN Section -----------------------------------

##(TRAIN) Creating the Return and Direction columns for Train Bear Data
~~~~

~~~
```{r}
#Expotential Moving Average for Calculating the Return and Direction
######check to see if you can modify the ratio parameter
################NOTE:removing noise
ema_train_bear_close<-EMA(train_bear[,4],n=30)



#Return
return_train_bear<-diff(log(ema_train_bear_close),lag=5,na.pad=TRUE)
#plot(return_train_bear, main="Return Bear")



#Calculating the Direction
direction<-ifelse(return_train_bear>0,1,0)


```


## (TRAIN) Technical Inidcators Train Data - Relative Strength Index
```{r}
#Relative Strength Index (BEAR Train)
rsi<-RSI(train_bear[,4], n=14, maType='WMA')

```


## (TRAIN) Technical Inidcators Train Data - Average Directional Index
```{r}
#Average Directional Index (BEAR Train )
~~~

~~~
dmi.adx<-ADX(train_bear[,c(2,3,4)])

```

## (TRAIN) Technical Inidcators Train Data - Bollinger Band
```{r}
#Bollinger Bands (BEAR Train)

bb20 <- BBands(train_bear[,4], sd=2)

```

#### (TRAIN) Technical Inidcators Train Data - Chaikin Money Flow
```{r}
#Chaikin Money Flow (BEAR Train) - Known for signaling for a bull or bearish market

cmf <- CMF(train_bear[,c(2,3,4)], train_bear[,6], n=20)


```



## (TRAIN) Technical Indicators Train Data - Commodity Channel Index (Bear Train)
```{r}
#Commodity Channel Index (BEAR Train)

cci <- CCI(train_bear[,c(2,3,4)], n=30 )

```

## (TRAIN) Combining all the technical indicators
```{r}
#combining the technical indicators & return & direction
~~~

~~~~
train_bear_ti<-data.frame(train_bear,dmi.adx,bb20)
train_bear_ti$RSI<-rsi
train_bear_ti$CCI<-cci
train_bear_ti$CMF<-cmf
train_bear_ti$Return<-return_train_bear #adding the calculated return column
train_bear_ti$direction<-direction #adding the direction 0 if price has gone down and 1 if price has gone up


```


##Creating Table and Correlation Matrix
```{r}


bear_model_train<-data.frame(train_bear_ti[,-c(1,2,3,4,5,6,18)])
bear_model_train<-na.omit(bear_model_train)
bear_model_train$direction<-as.factor(bear_model_train$direction)
#table(bear_model_train$direction)

#prop.table(table(bear_model_train$direction))
#barplot(prop.table(table(bear_model_train$direction)), col = rainbow(2), ylim= c(0, 0.7), main="Direction Distribution Bear")

#Correlation Matrix
#corrMatrix <- cor(bear_model_train[,-c(12)])
#print(corrMatrix)

#ggcorr(corrMatrix, label = TRUE)

#corrplot(corrMatrix, method="number")
#corrplot(corrMatrix, type = "lower", order = "hclust", tl.col = "black", tl.srt = 90)

```
~~~~

~~~
##--------------------------------------------------------TEST Section -------------------------------------


##(TEST) Creating the Return and Direction columns for Test Bear Data
```{r}

#TEST Expotential Moving Average for Calculating the Return and Direction

ema_test_bear_close<-EMA(test_bear[,4],n=30)



#Return
return_test_bear<-diff(log(ema_test_bear_close),lag=5,na.pad=TRUE)




#Calculating the Direction
direction_test<-ifelse(return_test_bear>0,1,0)


~~~

##(TEST) Technical Inidcators Test Data - Relative Strength Index
```{r}
#Relative Strength Index (BEAR Test)

rsi_test<-RSI(test_bear[,4], n=14, maType='WMA')

```

##(TEST) Technical Inidcators Test Data - Average Directional Index
```{r}
```

~~~
#Average Directional Index (BEAR Test )

dmi.adx_test<- ADX(test_bear[,c(2,3,4)], n=15)

```

##(TEST) Technical Inidcators Test Data - Bollinger Bands
```{r}
#Bollinger Bands (BEAR Test)

bb20_test <- BBands(test_bear[,4], sd=2)
#head(bb20_test,n=50)
#plot(bb20_test)

```

##(TEST) Technical Inidcators Test Data - Chailkin Money Flow
```{r}
#Chaikin Money Flow (BEAR Test) - Known for signaling for a bull or bearish market

cmf_test <- CMF(test_bear[,c(2,3,4)], test_bear[,6], n=20)

```

##(TEST) Technical Inidcators Test Data - Commodity
```{r}
#Commodity Channel Index (BEAR Test)

cci_test <- CCI(test_bear[,c(2,3,4)], n=30)

```


##Creating the Return and Direction columns for Test Bear Data
~~~

~~~
```{r}


test_bear_ti<-data.frame(test_bear,dmi.adx_test,bb20_test)
test_bear_ti$RSI<-rsi_test
test_bear_ti$CCI<-cci_test
test_bear_ti$CMF<-cmf_test
test_bear_ti$cal_return<-return_test_bear #adding the calculated return column
test_bear_ti$direction<-direction_test #adding the direction 0 if price has gone down and 1 if price has gone up


```

##Creating Table (BEAR TEST)
```{r}
#trying to understand Andre's code for the correlation matrix on his example ...

bear_model_test<-data.frame(test_bear_ti[,-c(1,2,3,4,5,6,18)])
bear_model_test<-na.omit(bear_model_test)
bear_model_test$direction<-as.factor(bear_model_test$direction)
#plot(bear_model_test)


```


##-------------------------------------------Machine Learning Algorithms ------------------------------------------

##Support Vector Machine
```{r}
#NOTE:Finding the best cost for the svm ; tune is a function that is used for cross validation and svm
~~~

```
set.seed(451)

#Support Vector Machine Tune
bear_svm_tune<-tune(svm,direction ~., data=bear_model_train,ranges=
list(cost=c(.001,0.01,0.1,1,5,10,100), gamma=c(0.5,1,2,3,4,5)))
summary(bear_svm_tune)

#Determining the best model
bestmodel <-bear_svm_tune$best.model
summary(bestmodel)

#Support Vector Model
bear_svm<- svm(direction ~., data=bear_model_train, method="ROC",
kernel='linear',cost=1, gamma=0.5, scale=FALSE)
summary(bear_svm)


#Prediction Model using Test
predict_bear_test<-predict(bear_svm,bear_model_test)
summary(predict_bear_test)

#confusion matrix
svm_tab<-table(predict_bear_test,bear_model_test$direction)
svm_tab
confusionMatrix(predict_bear_test, bear_model_test$direction)

#Misclassfication rate
#accurate_classfication <-sum(diag(svm_tab))/sum(svm_tab)
#head(accurate_classfication)

#missclassfication_rate<-1-sum(diag(svm_tab))/sum(svm_tab)
#head(missclassfication_rate)
```

~~~
#ROC
#NOTE:FPR - False Positive Rate $ TPR - True Positive Rate - ROC Curve

pred_svm<-predict(bear_svm, bear_model_test, type="prop")
bear_direct1 <- as.factor(bear_model_test$direction)
pred_svm_123<-prediction(as.numeric(pred_svm), as.numeric(bear_model_test$direction))

roc_svm<-performance(pred_svm_123,"tpr", "fpr")
plot(roc_svm, colorize=T, main="ROC Curve for SVM")#,# xlab= "Sensitivity", xlab="1-Specificity")
evaluation_svm<-performance(pred_svm_123,"acc")
plot(evaluation_svm)
abline(h=0.71, v=0.45)


SVM.perf <- performance(pred_svm_123, "prec", "rec")

summary(SVM.perf)


#Area Under the Curve (AUC)
Area<-performance(pred_svm_123, "auc")
Area<-unlist(slot(Area,"y.values"))
Area
```


## Retreiving Stock Data from Yahoo (bull Market Only 26 Months in total)

```{r}
#Financial Sector (bull)
~~~

```
bankofamericabull <- pdfetch_YAHOO("BAC", fields = c("open", "high", "low", "close", "adjclose", "volume"), from = as.Date("2009-04-01"), to = as.Date("2011-07-01"), interval = "1d")
bankofamericabull <-data.frame(bankofamericabull)
bankofamericabull <- cbind(date = rownames(bankofamericabull), bankofamericabull)
bankofamericabull$date<-as.Date(bankofamericabull$date)
colnames(bankofamericabull)<- c("date", "bacbull.open","bacbull.high","bacbull.low", "bacbull.close", "bacbull.adjclose", "bacbull.volume")

wellsfargoandcompanybull <- pdfetch_YAHOO("WFC", fields = c("open", "high", "low", "close", "adjclose", "volume"), from = as.Date("2009-04-01"), to = as.Date("2011-07-01"), interval = "1d")
wellsfargoandcompanybull <-data.frame(wellsfargoandcompanybull)
wellsfargoandcompanybull <- cbind(date = rownames(wellsfargoandcompanybull), wellsfargoandcompanybull)
wellsfargoandcompanybull$date<-as.Date(wellsfargoandcompanybull$date)
colnames(wellsfargoandcompanybull)<- c("date", "wfcbull.open","wfcbull.high","wfcbull.low", "wfcbull.close", "wfcbull.adjclose", "wfcbull.volume")

allstatecorporationbull <- pdfetch_YAHOO("ALL", fields = c("open", "high", "low", "close", "adjclose", "volume"), from = as.Date("2009-04-01"), to = as.Date("2011-07-01"), interval = "1d")
allstatecorporationbull <-data.frame(allstatecorporationbull)
allstatecorporationbull <- cbind(date = rownames(allstatecorporationbull), allstatecorporationbull)
allstatecorporationbull$date<-as.Date(allstatecorporationbull$date)
colnames(allstatecorporationbull)<- c("date", "allbull.open","allbull.high","allbull.low", "allbull.close", "allbull.adjclose", "allbull.volume")

morganstanleybull <- pdfetch_YAHOO("MS", fields = c("open", "high", "low", "close", "adjclose", "volume"), from = as.Date("2009-04-01"), to = as.Date("2011-07-01"), interval = "1d")
morganstanleybull <-data.frame(morganstanleybull)
mogranstanleybull <- cbind(date = rownames(morganstanleybull), morganstanleybull)
morganstanleybull <- cbind(date = rownames(morganstanleybull), morganstanleybull)
morganstanleybull$date<-as.Date(morganstanleybull$date)
colnames(morganstanleybull)<- c("date","msbull.open","msbull.high","msbull.low",
"msbull.close", "msbull.adjclose", "msbull.volume")

#Technology Sector (bull)
ebayincbull <- pdfetch_YAHOO("EBAY", fields = c("open", "high", "low", "close",
"adjclose", "volume"), from = as.Date("2009-04-01"), to = as.Date("2011-07-01"),
interval = "1d")
ebayincbull <-data.frame(ebayincbull)
ebayincbull <- cbind(date = rownames(ebayincbull), ebayincbull)
ebayincbull$date<-as.Date(ebayincbull$date)
colnames(ebayincbull)<- c("date", "ebaybull.open","ebaybull.high","ebaybull.low",
"ebaybull.close", "ebaybull.adjclose", "ebaybull.volume")

ibmbull <- pdfetch_YAHOO("IBM", fields = c("open", "high", "low", "close",
"adjclose", "volume"), from = as.Date("2009-04-01"), to = as.Date("2011-07-01"),
interval = "1d")
ibmbull <-data.frame(ibmbull)
ibmbull <- cbind(date = rownames(ibmbull), ibmbull)
ibmbull$date<-as.Date(ibmbull$date)
colnames(ibmbull)<- c("date", "ibmbull.open","ibmbull.high","ibmbull.low",
"ibmbull.close", "ibmbull.adjclose", "ibmbull.volume")

ciscosystemsincbull <- pdfetch_YAHOO("CSCO", fields = c("open", "high", "low",
"close", "adjclose", "volume"), from = as.Date("2009-04-01"), to = as.Date("2011-07-
01"), interval = "1d")
ciscosystemsincbull <-data.frame(ciscosystemsincbull)
ciscosystemsincbull <- cbind(date = rownames(ciscosystemsincbull),
ciscosystemsincbull)
ciscosystemsincbull$date<-as.Date(ciscosystemsincbull$date)
```

```
colnames(ciscosystemsincbull)<- c("date",
"cscobull.open","cscobull.high","cscobull.low", "cscobull.close", "cscobull.adjclose",
"cscobull.volume")

intelbull <- pdfetch_YAHOO("INTL", fields = c("open", "high", "low", "close",
"adjclose", "volume"), from = as.Date("2009-04-01"), to = as.Date("2011-07-01"),
interval = "1d")
intelbull <-data.frame(intelbull)
intelbull <- cbind(date = rownames(intelbull), intelbull)
intelbull$date<-as.Date(intelbull$date)
colnames(intelbull)<- c("date", "intlbull.open","intlbull.high","intlbull.low",
"intlbull.close", "intlbull.adjclose", "intlbull.volume")

#Healthcare Sector (bull)
cignacorporationbull <- pdfetch_YAHOO("CI", fields = c("open", "high", "low",
"close", "adjclose", "volume"), from = as.Date("2009-04-01"), to = as.Date("2011-07-
01"), interval = "1d")
cignacorporationbull <-data.frame(cignacorporationbull)
cignacorporationbull <- cbind(date = rownames(cignacorporationbull),
cignacorporationbull)
cignacorporationbull$date<-as.Date(cignacorporationbull$date)
colnames(cignacorporationbull)<- c("date", "cibull.open","cibull.high","cibull.low",
"cibull.close", "cibull.adjclose", "cibull.volume")

unitedhealthgroupincorporatedbull <- pdfetch_YAHOO("UNH", fields = c("open",
"high", "low", "close", "adjclose", "volume"), from = as.Date("2009-04-01"), to =
as.Date("2011-07-01"), interval = "1d")
unitedhealthgroupincorporatedbull <-data.frame(unitedhealthgroupincorporatedbull)
unitedhealthgroupincorporatedbull <- cbind(date =
rownames(unitedhealthgroupincorporatedbull), unitedhealthgroupincorporatedbull)
unitedhealthgroupincorporatedbull$date<-
as.Date(unitedhealthgroupincorporatedbull$date)
```

```
colnames(unitedhealthgroupincorporatedbull)<- c("date",
"unhbull.open","unhbull.high","unhbull.low","unhbull.close", "unhbull.adjclose",
"unhbull.volume")

centenecorporationincbull <- pdfetch_YAHOO("CNC", fields = c("open", "high",
"low", "close", "adjclose", "volume"), from = as.Date("2009-04-01"), to =
as.Date("2011-07-01"), interval = "1d")
centenecorporationincbull <-data.frame(centenecorporationincbull)
centenecorporationincbull <- cbind(date = rownames(centenecorporationincbull),
centenecorporationincbull)
centenecorporationincbull$date<-as.Date(centenecorporationincbull$date)
colnames(centenecorporationincbull)<- c("date",
"cncbull.open","cncbull.high","cncbull.low","cncbull.close", "cncbull.adjclose",
"cncbull.volume")


humanaincbull <- pdfetch_YAHOO("HUM", fields = c("open", "high", "low", "close",
"adjclose", "volume"), from = as.Date("2009-04-01"), to = as.Date("2011-07-01"),
interval = "1d")
humanaincbull <-data.frame(humanaincbull)
humanaincbull <- cbind(date = rownames(humanaincbull), humanaincbull)
humanaincbull$date<-as.Date(humanaincbull$date)
colnames(humanaincbull)<- c("date",
"humbull.open","humbull.high","humbull.low","humbull.close", "humbull.adjclose",
"humbull.volume")


#Industrial Sector (bull)
threemcompanybull <- pdfetch_YAHOO("MMM", fields = c("open", "high", "low",
"close", "adjclose", "volume"), from = as.Date("2009-04-01"), to = as.Date("2011-07-
01"), interval = "1d")
threemcompanybull <-data.frame(threemcompanybull)
threemcompanybull <- cbind(date = rownames(threemcompanybull),
threemcompanybull)
```

```
threemcompanybull$date<-as.Date(threemcompanybull$date)
colnames(threemcompanybull)<- c("date",
"mmmbull.open","mmmbull.high","mmmbull.low","mmmbull.close",
"mmmbull.adjclose", "mmmbull.volume")


generalelectriccompanybull <- pdfetch_YAHOO("GE", fields = c("open", "high",
"low", "close", "adjclose", "volume"), from = as.Date("2009-04-01"), to =
as.Date("2011-07-01"), interval = "1d")
generalelectriccompanybull <-data.frame(generalelectriccompanybull)
generalelectriccompanybull <- cbind(date = rownames(generalelectriccompanybull),
generalelectriccompanybull)
generalelectriccompanybull$date<-as.Date(generalelectriccompanybull$date)
colnames(generalelectriccompanybull)<- c("date",
"gebull.open","gebull.high","gebull.low","gebull.close", "gebull.adjclose",
"gebull.volume")

southwestairlinescompanybull <- pdfetch_YAHOO("LUV", fields = c("open", "high",
"low", "close", "adjclose", "volume"), from = as.Date("2009-04-01"), to =
as.Date("2011-07-01"), interval = "1d")
southwestairlinescompanybull <-data.frame(southwestairlinescompanybull)
southwestairlinescompanybull <- cbind(date =
rownames(southwestairlinescompanybull), southwestairlinescompanybull)
southwestairlinescompanybull$date<-as.Date(southwestairlinescompanybull$date)
colnames(southwestairlinescompanybull)<- c("date",
"luvbull.open","luvbull.high","luvbull.low","luvbull.close", "luvbull.adjclose",
"luvbull.volume")

generaldynamicscorporationbull <- pdfetch_YAHOO("GD", fields = c("open", "high",
"low", "close", "adjclose", "volume"), from = as.Date("2009-04-01"), to =
as.Date("2011-07-01"), interval = "1d")
generaldynamicscorporationbull <-data.frame(generaldynamicscorporationbull)
generaldynamicscorporationbull <- cbind(date =
rownames(generaldynamicscorporationbull), generaldynamicscorporationbull)
```

```
generaldynamicscorporationbull$date<-as.Date(generaldynamicscorporationbull$date)
colnames(generaldynamicscorporationbull)<- c("date",
"gdbull.open","gdbull.high","gdbull.low","gdbull.close", "gdbull.adjclose",
"gdbull.volume")

```

##Visulisation of Close for all of the bull Stocks

```{r}
##Visual
stocks <- data.frame(bankofamericabull = bankofamericabull[, "bacbull.close"],
wellsfargoandcompanybull = wellsfargoandcompanybull[, "wfcbull.close"],
            allstatecorporationbull = allstatecorporationbull[, "allbull.close"],
morganstanleybull= morganstanleybull[,"msbull.close"], ebayincbull
=ebayincbull[,"ebaybull.close"],ibmbull =
ibmbull[,"ibmbull.close"],ciscosystemsincbull =ciscosystemsincbull[,"cscobull.close"],
intelbull = intelbull[,"intlbull.close"], cignacorporationbull =
cignacorporationbull[,"cibull.close"], unitedhealthgroupincorporatedbull =
unitedhealthgroupincorporatedbull[,"unhbull.close"], humanaincbull =
humanaincbull[,"humbull.close"], centenecorporationincbull=
centenecorporationincbull[,"cncbull.close"] , threemcompanybull=
threemcompanybull[,"mmmbull.close"], generalelectriccompanybull =
generalelectriccompanybull[,"gebull.close"],southwestairlinescompanybull =
southwestairlinescompanybull[,"luvbull.close"], generaldynamicscorporationbull =
generaldynamicscorporationbull[,"gdbull.close"])

stocks$date<-bankofamericabull[,"date"]
#stocks1<-xts(stocks[,-17], order.by=stocks$date)
#head(stocks1)


#rain<-rainbow(ncol(stocks1))
#plot(x=stocks1, col=rain, main="Bull Close Price for all Sectors")
```

~~~
#legend("right",inset=-.05, title="Stocks", c("BAC", "WFC", "ALL", "MS", "EBAY", "IMB", "CSCO", "INTL", "CI", "UNH", "HUM","CNC", "MMM", "GE", "LUV", "GD"), fill=rain, xpd = TRUE,cex=.5, box.lty=0)
#par(xpd = T, mar = par()$mar + c(0,0,0,7))



```



## Converting files to zoo formate from dataframe and zoo formate from
```{r}

#------------------Finance
#bull_xts<-xts(bankofamericabull[,-1],order.by = bankofamericabull$date)
#bull_xts<-xts(wellsfargoandcompanybull[,-1],order.by = wellsfargoandcompanybull$date)
#bull_xts<-xts(allstatecorporationbull[,-1],order.by = allstatecorporationbull$date)
#bull_xts<-xts(morganstanleybull[,-1],order.by = morganstanleybull$date)
#------------------Technology
#bull_xts<-xts(ebayincbull[,-1],order.by = ebayincbull$date)
#bull_xts<-xts(intelbull[,-1],order.by = intelbull$date)
#bull_xts<-xts(ciscosystemsincbull[,-1],order.by = ciscosystemsincbull$date)
#bull_xts<-xts(ibmbull[,-1],order.by = ibmbull$date)
#------------------Healthcare
#bull_xts<-xts(cignacorporationbull[,-1],order.by = cignacorporationbull$date)
#bull_xts<-xts(unitedhealthgroupincorporatedbull[,-1],order.by = unitedhealthgroupincorporatedbull$date)
#bull_xts<-xts(humanaincbull[,-1],order.by = humanaincbull$date)
#bull_xts<-xts(centenecorporationincbull[,-1],order.by = centenecorporationincbull$date)
#------------------Industrail
bull_xts<-xts(threemcompanybull[,-1],order.by = threemcompanybull$date)
~~~

~~~~
#bull_xts<-xts(generalelectriccompanybull[,-1],order.by =
generalelectriccompanybull$date)
#bull_xts<-xts(southwestairlinescompanybull[,-1],order.by =
southwestairlinescompanybull$date)
#bull_xts<-xts(generaldynamicscorporationbull[,-1],order.by =
generaldynamicscorporationbull$date)

```

##Candlestick and graphs
```{r}

#chartSeries(bull_xts[,4],name='Bull Close - Morgan Stanely', type=("candlestick"),
subset=NULL, show.grid=TRUE,time.scale = NULL, bar.type="ohlc",
TA=c('addVo()','addBBand()'))



#chartSeries(bull_xts,name='bull',type='candle',subset='2009-04::2009-10',up.col =
"black", dn.col = "red",theme="white", bar.type="ohlc")


#plot(bull_xts[,4], main = "Close Price Bull")

```


## Split into Training and Test for bull Market

```{r}

train_bull<-window(bull_xts, start=as.Date("2009-04-01"), end=as.Date("2011-01-
14"))
#head(train_bull,n=30)
~~~~

~~~
#str(train_bull)


test_bull<-window(bull_xts, start=as.Date("2011-01-18"), end=as.Date("2011-07-01"))
#head(test_bull)
#str(test_bull)

```


####------------------------------------------------------TRAIN Section -----------------------------------

##(TRAIN) Creating the Return and Direction columns for Train bull Data
```{r}
#Expotential Moving Average for Calculating the Return and Direction
######check to see if you can modify the ratio parameter
################NOTE:removing noise
ema_train_bull_close<-EMA(train_bull[,4],n=30)
#summary(ema_train_bull_close)



#Return
return_train_bull<-diff(log(ema_train_bull_close),lag=5,na.pad=TRUE)
#plot(return_train_bull, main="Return bull")
#summary(return_train_bull)
#head(return_train_bull, n=60)


#Calculating the Direction
direction<-ifelse(return_train_bull>0,1,0)
~~~

~~~
```


## (TRAIN) Technical Inidcators Train Data - Relative Strength Index
```{r}
#Relative Strength Index (bull Train)
rsi<-RSI(train_bull[,4], n=14, maType='WMA')
#str(rsi)
#head(rsi, n=20)

```


## (TRAIN) Technical Inidcators Train Data - Average Directional Index
```{r}
#Average Directional Index (bull Train )

dmi.adx<-ADX(train_bull[,c(2,3,4)])
#head(dmi.adx,n=50)



```


## (TRAIN) Technical Inidcators Train Data - Bollinger Band
```{r}
#Bollinger Bands (bull Train)

bb20 <- BBands(train_bull[,4], sd=2)
#head(bb20,n=50)
#rain1<-rainbow(ncol(bb20))
#plot(bb20, col=rain1,  main="Bull Bollinger Band - eBay")
#legend("right",inset=-0.03, title="Bollinger Band", c("Down Value", "Moving Average"," Up Value", "Price"), fill=rain1, xpd = TRUE,cex=.5, box.lty=0)
~~~

~~~~
```


#### (TRAIN) Technical Inidcators Train Data - Chaikin Money Flow
```{r}
#Chaikin Money Flow (bull Train) - Known for signaling for a bull or bullish market

cmf <- CMF(train_bull[,c(2,3,4)], train_bull[,6], n=20)



```



## (TRAIN) Technical Indicators Train Data - Commodity Channel Index (bull Train)
```{r}
#Commodity Channel Index (bull Train)

cci <- CCI(train_bull[,c(2,3,4)], n=30 )

```


## (TRAIN) Combining all the technical indicators
```{r}
#combining the technical indicators & return & direction

train_bull_ti<-data.frame(train_bull,dmi.adx,bb20)
train_bull_ti$RSI<-rsi
train_bull_ti$CCI<-cci
train_bull_ti$CMF<-cmf
train_bull_ti$Return<-return_train_bull #adding the calculated return column
train_bull_ti$direction<-direction #adding the direction 0 if price has gone down and 1 if price has gone up
~~~~

~~~~
```

##Creating Table and Correlation Matrix
```{r}

bull_model_train<-data.frame(train_bull_ti[,-c(1,2,3,4,5,6,18)])
bull_model_train<-na.omit(bull_model_train)
bull_model_train$direction<-as.factor(bull_model_train$direction)
#head(bull_model_train)
#table(bull_model_train$direction)



#prop.table(table(bull_model_train$direction))
#barplot(prop.table(table(bull_model_train$direction)), col = rainbow(2), ylim= c(0, 0.7), main="Direction Distribution bull")

#Correlation Matrix
#corrMatrix <- cor(bull_model_train[,-c(12)])
#print(corrMatrix)

#ggcorr(corrMatrix, label = TRUE)

#corrplot(corrMatrix, method="number")
#corrplot(corrMatrix, type = "lower", order = "hclust", tl.col = "black", tl.srt = 90)

```

##-----------------------------------------------------TEST Section -----------------------------------

##(TEST) Creating the Return and Direction columns for Test bull Data
```{r}
~~~~

~~~
#TEST Expotential Moving Average for Calculating the Return and Direction

ema_test_bull_close<-EMA(test_bull[,4],n=30)


#Return
return_test_bull<-diff(log(ema_test_bull_close),lag=5,na.pad=TRUE)
#head(return_test_bull, n=30)


#Calculating the Direction
direction_test<-ifelse(return_test_bull>0,1,0)


```


##(TEST) Technical Inidcators Test Data - Relative Strength Index
```{r}
#Relative Strength Index (bull Test)

rsi_test<-RSI(test_bull[,4], n=14, maType='WMA')
#str(rsi)
#head(rsi, n=20)


```


##(TEST) Technical Inidcators Test Data - Average Directional Index
```{r}
#Average Directional Index (bull Test )

dmi.adx_test<- ADX(test_bull[,c(2,3,4)], n=15)
~~~

~~~
#head(dmi.adx_test,n=50)


```

##(TEST) Technical Inidcators Test Data - Bollinger Bands
```{r}
#Bollinger Bands (bull Test)

bb20_test <- BBands(test_bull[,4], sd=2)
#head(bb20_test,n=50)
#plot(bb20_test)

```

##(TEST) Technical Inidcators Test Data - Chailkin Money Flow
```{r}
#Chaikin Money Flow (bull Test) - Known for signaling for a bull or bullish market

cmf_test <- CMF(test_bull[,c(2,3,4)], test_bull[,6], n=20)


```

##(TEST) Technical Inidcators Test Data - Commodity
```{r}
#Commodity Channel Index (bull Test)

cci_test <- CCI(test_bull[,c(2,3,4)], n=30)



```
~~~

~~~~
##Creating the Return and Direction columns for Test bull Data
```{r}


test_bull_ti<-data.frame(test_bull,dmi.adx_test,bb20_test)
test_bull_ti$RSI<-rsi_test
test_bull_ti$CCI<-cci_test
test_bull_ti$CMF<-cmf_test
test_bull_ti$cal_return<-return_test_bull #adding the calculated return column
test_bull_ti$direction<-direction_test #adding the direction 0 if price has gone down and 1 if price has gone up


```

##Creating Table (bull TEST)
```{r}
#trying to understand Andre's code for the correlation matrix on his example ...

bull_model_test<-data.frame(test_bull_ti[,-c(1,2,3,4,5,6,18)])
bull_model_test<-na.omit(bull_model_test)
bull_model_test$direction<-as.factor(bull_model_test$direction)
#table(bull_model_test$direction)
#plot(bull_model_test)

#summary(return_test_bull)
#rain<-rainbow(ncol(bull_model_test))
#plot(return_test_bull, main="Bull Close Price for all Sectors")



```
~~~~

~~~
##-------------------------------------------Machine Learning Algorithms -------------------
-------------------------

##Support Vector Machine
```{r}
#NOTE:Finding the best cost for the svm ; tune is a function that is used for cross
validation and svm

set.seed(451)

#Support Vector Machine Tune
bull_svm_tune<-tune(svm,direction ~., data=bull_model_train,ranges=
list(cost=c(.001,0.01,0.1,1,5,10,100), gamma=c(0.5,1,2,3,4)))
summary(bull_svm_tune)

#Determining the best model
bestmodel <-bull_svm_tune$best.model
summary(bestmodel)

#Support Vector Model
bull_svm<- svm(direction ~., data=bull_model_train, method="ROC",
kernel='linear',cost=5, gamma=0.5, scale=FALSE)
summary(bull_svm)

#Prediction Model using Test
predict_bull_test<-predict(bull_svm,bull_model_test)
summary(predict_bull_test)

#confusion matrix
svm_tab<-table(predict_bull_test,bull_model_test$direction)
head(svm_tab)
svm_tab
~~~

```
confusionMatrix(predict_bull_test, bull_model_test$direction)


#Misclassfication rate
accurate_classfication <-sum(diag(svm_tab))/sum(svm_tab)
summary(accurate_classfication)

missclassfication_rate<-1-sum(diag(svm_tab))/sum(svm_tab)
summary(missclassfication_rate)

#ROC
#NOTE:FPR - False Positive Rate $ TPR - True Positive Rate - ROC Curve

pred_svm<-predict(bull_svm, bull_model_test, type="prop")
bull_direct1 <- as.factor(bull_model_test$direction)
pred_svm_123<-prediction(as.numeric(pred_svm),
as.numeric(bull_model_test$direction))

roc_svm<-performance(pred_svm_123,"tpr", "fpr")
plot(roc_svm)
#plot(na.omit(roc_svm, colorize=T, main="ROC Curve for SVM")#,# xlab=
"Sensitivity", xlab="1-Specificity")
evaluation_svm<-performance(pred_svm_123,"acc")
plot(evaluation_svm)
abline(h=0.71, v=0.45)


SVM.perf <- performance(pred_svm_123, "prec", "rec")

plot(SVM.perf)


#Area Under the Curve (AUC)
Area<-performance(pred_svm_123, "auc")
```

~~~~
Area<-unlist(slot(Area,"y.values"))
Area
```


##Random Forrest
```{r}

set.seed(451)

#Tuning the Random Forest Model
rf_tune<-tuneRF(bear_model_train[,-12],bear_model_train[,12], stepFactor = 0.5, plot=TRUE, ntreeTry = 400, trace=TRUE, improve=0.05)
plot(rf_tune)

#Random Forest Model
bear_rf<-randomForest(direction ~., data=bear_model_train, ntree=400, cv.fold=10, do.trace=50, importance=TRUE, mytry=1, proximity=TRUE)

print(bear_rf)
#attributes(bear_rf)
#plot(bear_rf)

#Prediction & Confusion Matrix - Train Data - To just view the difference
RF_train<-predict(bear_rf, type="response")
confusionMatrix(RF_train,bear_model_train$direction)
table(RF_train,bear_model_train$direction)



#Prediction & Confusion Matrix - Test Data
pred_rf<-predict(bear_rf, newdata=bear_model_test[,-c(12)])
rf_tab<-table(pred_rf,bear_model_test$direction)
head(rf_tab)
~~~~

~~~
confusionMatrix(pred_rf, bear_model_test$direction)


#Number of nodes for the trees
hist(treesize(bear_rf), main="Number of Nodes for the Trees", col = "blue")

#Variable Importance
varImpPlot(bear_rf, sort = T, n.var = 10, main= "Top - Variable Importance")
importance(bear_rf)
varUsed(bear_rf)
print(bear_rf)

#Extract Single Tree
getTree(bear_rf, labelVar=TRUE)

#Multi-dimensional Scaling Plot
MDSplot(bear_rf, bear_model_test$direction)


#ROC Curve
prediction_rf<-prediction(as.numeric(pred_rf),as.numeric(bear_model_test$direction))
roc_rf<-performance(prediction_rf, "tpr", "fpr")
plot(roc_rf, colorize=T, main="ROC Curve for RF")

    #, xlab= "Sensitivity", xlab="1-Specificity")
evaluation_rf<-performance(prediction_rf,"acc")
plot(evaluation_rf)

#Area Under the Curve (AUC)
Area<-performance(prediction_rf, "auc")
Area<-unlist(slot(Area,"y.values"))
Area
```
~~~