Predicting Stock Price Direction Through Data Mining and Machine Learning Techniques

An Economics/Computer Science Interdepartmental Thesis

Conor Carey

June 10, 2015

Abstract

In this paper, we attempt various ways to test stock data by sector to determine whether sector influence can improve on stock price trend prediction accuracy. The stock market garners a lot of attention from its ability to make fortunes and take them away just as fast. Due to the power of the stock market, many people have built livelihoods depending on their ability to determine successful portfolios for others. In the past ten years, the stock market has seen one of its most dramatic falls with an even more dramatic recovery since its inception, causing many people to lose their jobs, which shows how important it is to make the right investment choices. This paper aims to strengthen an analyst's ability to choose a portfolio through using fundamental, technical and macroeconomic variables to improve stock market prediction accuracy. Existing models use historical stock price, volume and macroeconomic variables to predict the future stock price trend with an accuracy rate between 50 and 73%. This thesis attempts to develop a model for predicting stock price trend with at least 74% accuracy rate and to explore various ways in which a company's sector can improve prediction accuracy. It will implement an appropriate machine learning algorithm to analyze a sample of 50 high volume traded stocks. The algorithm used in this thesis is the Support Vector Machine (SVM). The SVM is chosen for its ability to handle large datasets and separate them into non-linear sections to more easily classify the data without overfitting the model. It is hypothesized that manipulating the data in various ways by sector will improve on the prediction accuracy. The results from the many tests do not support my hypothesis. The most significant difference found between the unchanged dataset and the changed dataset was a 5.6 percentage point increase. This increase was not statistically significant. This thesis shows the methods used have no significant effect on future stock price trend prediction.

Contents

List of Tables

1 Introduction

The stock market invites academic and professional study from around the world due to its promise of wealth. Current academic theories dictate that the stock market is unpredictable. The Efficient Market Hypothesis, one of the most prominent of these academic theories states, it is impossible to beat the market because the stock market efficiency causes existing share prices to always incorporate and reflect all relevant information. Despite the academic theories, researchers are still determined to predict the stock market due to the reward of unlimited wealth. This thesis explores different ways of manipulating stock data by sector to improve prediction accuracy.

With the promise of unlimited wealth, an abundance of research has attempted to predict the stock market. Although none of these studies have been entirely successful, the results are enough to show that the stock market is not as efficient as previously believed. Huang et al. (2005) successfully predicted the future weekly trend in the NIKKEI 225 index 73% of the time using machine learning. Huang et al. (2005) had the most successful model for stock market prediction even though they used the same machine learning method as Shah (2007) and Wang and Choi (2013). This is because machine learning algorithms are data dependent. Huangs dataset consisted of macroeconomic variables, including interest rates, consumer price index, industrial production, government consumption, private consumption, gross national product and gross domestic product. The purpose of Huangs study was to determine the best machine learning algorithm for stock market price trend prediction, however, the macroeconomic dataset in Huangs study proved to be more successful than the datasets in other studies that consisted only of historical stock prices and other index prices. This thesis examines the possibility of improving datasets by including sector information. Through changing the dataset created in this thesis by sector, we hoped to determine that sector information could influence future stock price trend prediction accuracy positively.

The stock market has not always been as efficient as it is now. Before 1960, the Dow Jones Industrial Average (DJIA) was strongly correlated with the movement of it's book-to-market ratio. [Pontiff and Schall, 1998] After 1960, the DJIA's book-to-market value's predictive ability dissipated. This suggests that the stock market previous to 1960 incorporated the book-to-market ratio with a lag. After 1960, the stock market's efficiency increased, causing the book-to-market ratio to be fully incorporated into the stock price with-

out lag, causing this variable to lose it's predictive ability. This thesis hypothesizes that by testing sectors within the stock market separately, patterns could be more easily identifiable. Such as, if the US interest rate increases from quarter 1 to quarter 2, the companies in the Financial sector may have lower stock prices in quarter 3 because that increase in the interest rate caused banks to borrow less money from the Federal Reserve, thus lowering the amount of money they loan to consumers, earning the companies in the Financial sector less money in the future. This same increase in interest rates could have no effect on the Basic Materials sector. Basic Materials include oil, steel, agricultural chemicals, etc. these goods are considered to be inelastic. Inelastic goods are products that are necessary to purchase; fluctuations in these good's prices does not effect the amount of that good that is purchased. With this in mind, this thesis attempts to find such patterns within sectors using machine learning algorithms.

Machine learning is a technique used for finding patterns in large datasets that would not usually be easily found. While machine learning has been around for decades, it has only recently started to be used as a technique for finding patterns in stock price trends to predict future movement. Previous techniques for finding trends involved either observing the performance of a company, or observing the previous trend in the stock price charts. These fundamental and technical analysis techniques have shown to be no more useful than choosing a stock portfolio at random [Cragg and Malkiel, 1982].

Stock market trend forecasting only observes whether the stock price will increase or decrease in the future. In machine learning terms, that makes predicting the price trend to be a classification problem. The machine learning algorithm will observe each line of data and see which factors contribute to a stock price increasing or decreasing in the future. Therefore, based on the observations of the algorithm, each data point will be classified as increasing or decreasing depending on the input factors for each variable.

As this is a classification study, a classification algorithm must be used to produce the desired output. Many previous studies have compared each machine learning technique to find the machine learning algorithm that yields the best accuracy with the lowest error rate. All studies, including Huang et al. (2005), have determined the best algorithm for stock market prediction is the Support Vector Machine (SVM).

The SVM algorithm is a classification algorithm that creates a hyper plane to separate data into classes. After the hyper plane is defined by the training data, the SVM evaluates the testing data on it. Due to the

SVMs ability to handle large datasets and avoid overfitting the data, the SVM will be the learning algorithm used in this study. The ratio of the correctly classified instances over the total number of classified instances is what gives the accuracy percentage of the model. A model with an accuracy percentage higher than 73% is the goal of this thesis.

For this thesis, the nominal variable of sector information will be added to the dataset to determine if it will positively affect the trend prediction accuracy of the model. A second purpose for this study is to create a model that improves upon the previous models by compiling the best available data from previous studies into one dataset. The improved model serves as the benchmark to test the effects of the addition of the nominal sector variable. To perform this task, it is important to use the variables that have been proven to predict stock price trends in the past. Many studies attempt to find variables that are highly correlated with the overall stock market prices. Arbarbanell and Bushee (1997) find that Inventory, Gross Margin, Effective Tax Rate, Earnings quality and Labor Force are correlated to future stock price movements. Hong et al. (2005) discover Inflation, Market Dividend Yield, Default Spread, Monthly Market Volatility, Industrial Production, and an index of economic activity affect future earnings, which affects the market price. Chen also discerned that inflation has predictive potential, but he also found that interest rate spreads have predictive potential for stock price trends. Following Arbarbanell and Bushee, Hong et al. and Chen, this study includes fundamental data, such as inventory, gross margin, effective tax rate, labor force and historical stock prices, as well as macroeconomic data, such as inflation, industrial production, market dividend yield and interest rate spreads, in developing a model for the fifty chosen stocks.

Due to the nature of the financial industry, the inventory variable is not available for this sector. This variable is still included in the model; it simply wont provide any information that could potentially benefit the prediction accuracy of the financial sector. SVM algorithms are also known for their ability to handle sparse data. Even though the data point may be missing a large portion of data, the SVM gives more weight to the other attributes contributing to that data point and creates support vectors for that specific data point differently.

Adding a nominal sector variable into the dataset can potentially improve the accuracy of the model by its ability to further group the data. As machine learning algorithms learn patterns on datasets, this extra

nominal variable will allow the machine learning algorithm to break the dataset down into smaller pieces to learn patterns within sectors. It is potentially true that some variables may affect certain sectors more strongly than others, such as raised interest rates can have a larger effect on the financial industry in the positive direction, where the same change in interest rates may have a small negative effect on the energy industry. The hypothesis of this thesis is that adding the nominal sector variable will increase the prediction accuracy of the model it is added to.

In chapter 2, this paper describes previous studies' methods of achieving significant variables as well as other studies examining the different machine learning techniques. Chapter 3 explains how the machine learning technique works and discusses hypotheses to be tested. Chapter 4 discusses the construction of the data set and presents the empirical results. Chapter 5 presents the conclusions.

2 Literature Review

2.1 Regression Analysis

The performance of a machine learning algorithm depends on the quality of the data provided by the researcher. Many past studies attempted to find economic variables that are significantly related to the stock market's future trend. For example, Abarbanell and Bushee (1997) found that the Inventory, which is the goods that a company holds that is intended for sale, Gross Margin, and Effective Tax Rate, which is the average rate at which a company's profits are taxed. They also demonstrated Earnings Quality, which is how reasonable the reported earnings are and Labor Force, which is the total work force in the United States. These variables are statistically significantly related at the 5% level to one-year-ahead earnings in the direction anticipated. They also found the accounts receivable, or money owed to the company, variable to have a positive correlation to future earnings at a 10% significance level. Through their research, they determined that investors have a reasonable justification for using a few of the studied fundamental variables to aid in future trend prediction. The variables found to be significant at the 5% level in Arbarbanell and Bushee's study are included in this thesis.

Hong et al. (2005) investigated industry portfolio returns and how they might affect future stock price.

Within their research, it has been found with statistical significance that twelve of the thirty-eight industry portfolios they tested were able to forecast the stock market by one month. In order to ensure their findings, they compared these portfolios with the classic market forecasting variables, such as Inflation, Market Dividend Yield, Default Spread, Monthly Market Volatility, Industrial Production, and an index of economic activity. They tested the Metal sector against these six variables and found that inflation was the best stock market predicting variable and the Metal sector variable had a comparable performance in comparison.

Chen observed macroeconomic variables in order to predict bear markets. His study looked at interest rate spreads, inflation rates, money stocks, aggregate output, unemployment rates, federal funds rate, federal government debt and nominal effective change rates. Chen found interest rate spreads and inflation rates are the best predictors for bear market forecasting. To test for robustness of his findings, Chen observed how a one dollar investment in a buy and hold strategy did compared to his model. His model only buys shares when the possibility of a bear market was below 30% and selling when the risk became higher. Over a forty year period, the one dollar in the buy and hold strategy earned 18.98 dollars, while his model using inflation as the bear market indicator earned 264.95 dollars. Using his model, interest rate spreads earned 109.47 dollars. Predicting the downturn of the market allows us to sell at the stock price peak. [Chen, 2009]

Though this thesis will not be including option prices due to the lack of free, public data, it is worth mentioning that Pan and Poteshman found a strong relationship between option prices option volume and future stock prices. They have found that the option markets have more people with insider knowledge trading. Hence, the options with high jumps in volume are positively related with the corresponding future stock trends. [Pan and Poteshman, 2006]

Although there are immeasurable amounts of possible datasets, these studies prove helpful in discovering some of the variables that relate to future trends in market prices, which provides this thesis with the variables needed for machine learning analysis.

2.2 Machine Learning

Machine learning provides methods for finding patterns in large amounts of data. These previous studies strongly suggest that the support vector machine (SVM) provides the best algorithm for stock trend prediction. The SVM is a classification algorithm. Classification algorithms observe datasets and, based on the differences observed between classes, decides which class the future data points belong in. SVMs are known for their ability to handle large amounts of cluttered data better than other machine learning algorithms. A large amount of previous research has been done to test different machine learning algorithms on a single dataset to determine the statistically significantly better method for solving this particular problem.

Shah (2007) observes the prediction ability of four different algorithms on historical prices of Google and Yahoo. This study examines seven daily variables over two years: Date, Open price, High price, Low prices, Close price, Trade Volume, and Adjusted Close price. The four algorithms Shah tested are: Decision Stump, Linear Regression, SVM, and the AdaBoostM1 algorithm applied to an SVM. The decision stump is an algorithm that determines the single best variable and creates a single rule based on that variable to make all future predictions. The boosting algorithm runs many variations of the algorithm it is applied to and learns how to weigh the given variables in the algorithm it is applied to for better accuracy and lower error rates. Shah's study proved the SVM with the applied boosting algorithm provided the best results in terms of both accuracy and error rates out of the four algorithms tested. It achieved a four-percentage point increase in accuracy as compared to the SVM without the boosting algorithm applied.

Ou and Wang tested their data using 10 different machine learning techniques: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), K-nearest Neighbor Classification (KNC), Naive Bayes based on kernel classification, Logit model, Tree Based Classification, Neural Network, Bayesian Classification with Gaussian process, SVM and Least Squares SVM (LS-SVM). Ou and Wang tested these 10 algorithms on future stock prediction and compared them to find which had the best overall results including percentage accuracy and root mean square error. Their data is based on technical data using five variables: Open price, High price, Low price, S&P 500 index price and the Exchange rate of the US dollar against the Hong Kong dollar. These values were observed for future forecasting of the Hang Seng index price in Hong Kong.

According to Ou and Wang, the SVM and LS-SVM generate superior predictive performances among the other models. The LS-SVM provides another SVM with a booster to improve results. However, the improvements in the results were not significant enough to consider this method superior to the non-boosted SVM.

Another study examines the SVM in comparison to linear discriminant analysis, quadratic discriminant analysis, random walk and neural networks [Huang et al., 2005]. Huang et al. observe nine macroeconomic variables: interest rates, short-term interest rates, long-term interest rates, consumer price index, industrial production, government consumption, private consumption, gross national product and gross domestic product. Huang's results show the SVM performs significantly better than the other proposed models. The neural network algorithm was historically used in stock market prediction before the current SVM was invented in 1995. This study shows the SVM outperforms its historical counterpart by 4 percentage points, earning itself a 73% hit ratio.

Many of these studies include variables that are used in this research, although none of them included the dataset that this thesis tests. The studies all show performances of over 50%, which suggests these algorithms and datasets provide more than a random guess, however, the focus of these studies was aimed at the best algorithm for prediction, rather than testing whether the best algorithm worked with what is considered to be the best data.

Given the benefits of the boosting algorithm and the significant success of the SVM compared to other algorithms, these are the algorithms that are used in this thesis.

2.3 Conclusion

None of the many previous studies on stock market price trend prediction have researched the many studies on which economic variables are highly correlated with future returns. Some of the variables, such as the interest rates spread were included in the studies because they are widely considered to be a top indicator in an economy. This study observes the previous research on relevant variables to use and which algorithm to use the chosen data with.

3 Methodology

This study consists of ten independent variables: inventory, gross margin, effective tax rate, labor force, historical stock prices, inflation, industrial production, market dividend yield, interest rate spreads, as well as the sector that each company belongs to, and one dependent variable, which is the class variable in machine learning to determine the increase and decrease in the future trend. The dependent variable is generated by a python script that observes the difference of the stock price in the current quarter and next quarter; if the price increases, a BUY signal is generated, if the price decreases, a DNB (Do Not Buy) signal is generated. The model will be treated in such a way that at the end of every quarter, all stocks owned will be sold. The current quarter's BUY/DNB signal predictions will dictate which stocks to buy for future profit. All independent variables, with the exception of the sector data, are intended to build a model with higher accuracy than that of previous models. Then, using this model as a baseline indicator, the dataset is then manipulated by sector to determine if this method improves the prediction accuracy of the model.

The data used in this experiment has been obtained from various sources. The fifty stocks with the highest trade volumes over the previous thirty days can be found at `www.NYSE.com`. All other data used in this experiment can be collected at `www.ycharts.com`. Ycharts is a monthly subscription service provided by Yahoo Finance.

The variables chosen in Ycharts for each company downloads into a Microsoft Excel spreadhseet. The data must be sorted by date for the python script to accurately determine whether the classification for each data point is BUY or DNB. The python script observes the stock price at the end of a quarter and the stock price at the end of the next quarter. If the price at the end of the next quarter is higher than the current quarter, the script will generate a BUY signal. If the price at the end of the next quarter is lower than the current quarter, the script will generate a DNB signal.

Once these classifications are added to the individual stock's dataset, the nominal sector variable is then added. If the dataset currenty being handled belongs to Bank of America, a further column will be added containing the word FINANCIAL next to each data point. This indicates that Bank of America is a company that belongs to the financial sector.

Once these variables are added to the individual stock's dataset, the individual datasets are then ag-

gregated into one large dataset containing all 4200 observations, 84 observations for each of the 50 stocks. Due to the quarterly availability of certain variables, the model built predicts if the stock price will increase or decrease by the end of the next quarter. Therefore, the data collected in this thesis starts with the first quarter being 12/31/1993 and the last quarter being 9/30/2014. In order to have an even 21 years of data, an extra quarter past the last quarter must be collected to correctly determine if the final quarter should be classified as a BUY or DNB. Once the python script determines the correct classification, the extra data point may be discarded.

This data is tested using a support vector machine, as implemented in WEKA. WEKA is a free machine learning software tool that allows a user to preprocess data and test the processed data using machine learning algorithms.

Machine learning algorithms are techniques used to find patterns in large datasets. They were developed to increase the output accuracy and decrease time consumption. There are many styles of machine learning techniques, such as: classification, regression, and clustering. For this experiment, the support vector machine classification algorithm is used to determine the class of the dependent variable. The classes that the dependent variable will belong to either BUY or DNB, respectively, which indicates the increase or decrease in the price of the stock.

The support vector machine that is used in this thesis builds a model based on the independent variables given, and learns the patterns within the independent variables that distinguish the outcome of BUY or DNB. Using the cross-validation technique, we have confidence that our outcome is not the result of overfitting.

Ten-fold-cross-validation splits the dataset into ten different sections. It then trains the machine learning algorithm on nine of the sections and tests its discovered pattern on the remaining tenth. This process repeats ten times; the section to be tested will be changed upon repetition to ensure that the results are not solely based on one test that may not be applicable to future data.

The hypothesis mentioned earlier is tested using these techniques. For this thesis, in order to test the performance of the SVM on the individual sectors compared to the performance of the SVM on each sector within the aggregated dataset, manual ten-fold-cross validation must be performed. When WEKA per-

forms cross-validation, it first randomizes the data to avoid learning heavily on a sorted section of the data that could throw off the prediction. This is a crucial tool for the cross-validation to work properly, however, WEKA outputs the individual predictions with no indication of which data point it is attempting to predict. Normally, this does not matter, in a machine learning experiment, but for our purposes, we need the individual predictions for each data point to compare the prediction outcomes of the aggregate dataset to the individual sector dataset.

To do this, the data must be randomized manually. In Microsoft Excel, it is possible to do this by generating a random number next to each datapoint and sorting them by their random number. The data set must then be manually divided ten times using the same process as cross-validation. Take a tenth of the dataset out, use the other nine-tenths to train the algorithm, and then test the trained algorithm on the tenth of the dataset it did not train on. By using a separately created test dataset, WEKA no longer automatically randomizes the data, allowing us to take the predicted outcomes and insert them next to each data point in the test dataset.

This process is only necessary with testing the aggregate dataset. Once all data points in the aggregate dataset have a predicted outcome associated with them, ten-fold-cross-validation as implemented in WEKA can be used for the individual datasets. The prediction accuracy of each sector is then compared to that sectors prediction accuracy when trained on the aggregate dataset to attempt to prove our hypothesis; that manipulating the dataset using sector information will increase stock price future trend prediction accuracy. The improvement on the prediction accuracy must be significant at the 5% level to reject the null hypothesis.

Two other methods were attempted to see their effect on prediction accuracy. These methods are not complex and their table of results can be found in the appendix.

4 Results

These tests investigate whether the SVM performs better while being tested on the dataset of a single sector, or if it performs better by learning a model based on a large dataset with many sectors and then tested on that individual dataset. Throughout the tests, the baseline percentage for the aggregate dataset is 58%. If the SVM predictsa BUY signal for every datapoint, the test would produce a 58% accuracy percentage. The

aggregate dataset includes all data points from every sector and includes all independent and dependent variables, including the nominal sector variable.

As can be seen in table 1, the Basic Materials sector performed only slightly better when tested spearately from the aggregate dataset. The SVM favored the BUY category very heavily in the total dataset, with only 12 of the 1176 data points predicted as a DNB signal. The SVM completely favored the BUY signal in the individual dataset. The baseline percentage for the test on the Basic Materials sector is 58.50%, that is, if the machine learning predicts every instance as a BUY, the prediction accuracy will be 58.50%.

Result	**Actual**	**Aggregate Dataset**	**Sector Dataset**
False Buy	0	480	488
True Buy	688	676	688
False DNB	0	12	0
True DNB	488	8	0
Percentage:	100%	58.16%	58.50%

Table 1: Basic Materials

As can be seen in table 2, the Consumer Goods sector performed only slightly better when tested spearately from the aggregate dataset. The SVM favored the BUY category very heavily in the total dataset, with only 13 of the 588 data points predicted as a DNB signal. The SVM completely favored the BUY signal in the individual dataset. The baseline percentage for the test on the Consumer Goods sector is 59.52%.

Result	**Actual**	**Aggregate Dataset**	**Sector Dataset**
False Buy	0	235	238
True Buy	350	340	350
False DNB	0	10	0
True DNB	238	3	0
Percentage:	100%	58.33%	59.52%

Table 2: Consumer Goods

As can be seen in table 3, the Basic Materials sector performed oconsiderably better when tested spearately from the aggregate dataset. The SVM favored the BUY category very heavily in the total dataset, with only 38 of the 776 data points predicted as a DNB signal. The test on the individual dataset shows an increase of 5.56 percentage points from the test on the aggregate dataset. The baseline percentage for the test on the

Financial sector is 54.25%.

Result	Actual	Aggregate Dataset	Sector Dataset
False Buy	0	319	278
True Buy	421	399	410
False DNB	0	12	11
True DNB	335	26	57
Percentage:	100%	56.21%	61.77%

Table 3: Financial

As can be seen in table 4, the Healthcare sector did not perform better when tested spearately from the aggregate dataset. The SVM favored the BUY category very heavily in the total dataset, with only 3 of the 336 data points predicted as a DNB signal. The SVM completely favored the BUY signal in the individual dataset. The baseline percentage for the test on the Healthcare sector is 58.93%.

Result	Actual	Aggregate Dataset	Sector Dataset
False Buy	0	134	138
True Buy	198	194	198
False DNB	0	4	0
True DNB	138	4	0
Percentage:	100%	58.92%	58.93%

Table 4: Healthcare

As can be seen in table 5, the Industrial Goods sector performed slightly better when tested spearately from the aggregate dataset. The SVM favored the BUY category very heavily in the total dataset, with only 3 of the 252 data points predicted as a DNB signal. The SVM completely favored the BUY signal in the individual dataset. The baseline percentage for the test on the Industrial Goods sector is 59.92%.

Result	Actual	Aggregate Dataset	Sector Dataset
False Buy	0	101	101
True Buy	151	148	151
False DNB	0	3	0
True DNB	101	0	0
Percentage:	100%	58.73%	59.92%

Table 5: Industrial Goods

As can be seen in table 6, the Metal sector performed slightly worse when tested spearately from the aggregate dataset. The SVM favored the BUY category very heavily in the total dataset, only 1 of the 84 data points predicted as a DNB signal. The SVM poorly predicted an equal amount of BUY and DNB signals. The baseline percentage for the test on the Metals sector is 50%.

Result	Actual	Aggregate Dataset	Sector Dataset
False Buy	0	42	18
True Buy	42	41	24
False DNB	0	1	26
True DNB	42	0	16
Percentage:	100%	48.81%	47.62%

Table 6: Metals

As can be seen in table 7, the Service sector performs equally as well as when tested spearately from the aggregate dataset. The SVM favored the BUY category very heavily in the total dataset, with only 18 of the 505 data points predicted as a DNB signal. The SVM also heavily favored the BUY signal in the individual dataset with only 4 of the 505 data points predicted as a DNB signal. The baseline percentage for the test on the Service sector is 58.42%.

Result	Actual	Aggregate Dataset	Sector Dataset
False Buy	0	200	207
True Buy	295	286	293
False DNB	0	9	2
True DNB	209	9	2
Percentage:	100%	58.53%	58.53%

Table 7: Service

As can be seen in table 8, the Technology sector performed slightly worse when tested spearately from the aggregate dataset. The SVM favored the BUY category very heavily in the total dataset, with only 15 of the 504 data points predicted as a DNB signal. The SVM completely favored the BUY signal in the individual dataset. The baseline percentage for the test on the Technology sector is 57.74%.

Result	Actual	Aggregate Dataset	Sector Dataset
False Buy	0	204	213
True Buy	291	285	291
False DNB	0	6	0
True DNB	213	9	0
Percentage:	100%	58.33%	57.74%

Table 8: Technology

5 Discussion

The results from this experiment are not able to reject the null hypothesis. The majority of test outcomes show little to no improvement from the baseline percentage. The SVM heavily favors the BUY signal for both the individual sectors and sectors tested from the model learned on the aggregate dataset with two exceptions.

The Financial and Metal sectors produces results with considerably more DNB predictions than the other sectors. A notable difference is that the model built from the Financial sector achieves a much higher accuracy rate than that of any of the other sectors tested, while the model built from the Metals sector achieves the lowest prediction accuracy than that of the other sectors tested. A possible explanation for these exceptions can be explained by the nature of the SVM.

The Financial sector data included in our dataset is the sparsest data; it is the only sector in the dataset that is missing every value for the Inventory and Gross Profit Margin variables. The SVM classifies data points by creating a hyper plane that maximizes the distance between the algorithms consideration of the most similar data points and the other class.

The SVM is unable to correctly classify the testing data if it has to create hyper planes for too many variables. The feature space, or n-dimensional space, which is where the SVM separates the data points with the hyper planes, can become cluttered. The Financial sector relieves the SVM of having to create two extra hyper planes that may have cluttered the feature space, which causes the SVM to more easily classify the data points in the Financial sector. This would also explain the 5.56 percentage point increase from the model trained on the aggregate dataset. All other sectors include these variables, which clutter the feature space, causing the prediction accuracy for the financial variables to be markedly worse.

The Metals sector's poor results can be attributed to its lack of data points. The Metals sector is the only sector in the dataset that only includes one stock. Therefore, while other sectors have 200 or more observations to train on, the Metals sector only has a total of 84 data points. This means, with ten-fold-cross-validation, it only has 76 observations to learn from and only 8 chances to make a correct prediction for each fold. The lack of data points can also explain the increase in DNB signals due to the lack of clutter in the feature space. Other than the Metals and Financial sectors, the models created by the SVM for the other sectors are fairly consistent with heavy BUY signal predictions.

The frequency of the BUY signals can also be attributed to clutter in the feature space. If too many variables with similar values are all taken into consideration concurrently with the attempt to place a test data point into a class, then the SVM will simply place the test data point in the class that has the greater number of data points in that class. This way, the SVM will achieve an above 50% accurate classification as opposed to choosing the class that will achieve a below average classification rate.

Although the clutter in the feature space explains everything in our data, including the outliers, there may be other causes. Huang et al. (2007) conducted a study with one fewer variable than this study and achieved the highest accuracy rate of the previous studies researched in the making of this project. Their study also used only 640 data points to train their model with only 36 data points for testing.

Even with Huang's successful prediction accuracy rate that contradicts the reasoning behind the feature space clutter being the issue with this study, the clutter is still the best explanation for this thesis' insignificant results.

6 Conclusion

The desired outcome of this thesis was to achieve improved prediction accuracy of the learned models created by the SVM. The results of this thesis proved to be insignificant at the 5% level, therefore, we can not reject the null hypothesis; manipulating stock data by sector does not improve the SVM model's prediction accuracy. Through testing multiple methods of manipulating the aggregate dataset by sector, we have discovered that the particular methods used in this thesis will not be making extra money anytime soon.

6.1 Future Works

Future work for this project would include, but is not limited to, testing these methods using more data points, test further attempts to partition the data using sector information that could potentially increase prediction accuracy, and to test these methods on a different set of stocks.

More data points could allows the SVM to build a better feature space to allow for better, more accurate classifications.

The way the datasets are partitioned in this thesis does not provide a statistically significant increase in prediction accuracy. Testing these dataset using sector information differently could produce a more accurate model.

The stocks used in this thesis are the current 50 most highly traded stocks on the New York Stock Exchange. The stocks' consistent high trade volume suggests the stocks' price fluctuates with little volitility. This lack of volitility could mean the effects of the independent variables used in this study have a diminished affect on the price fluctuations of these particular stocks. Collecting a dataset with volitile price fluctuations within the sectors could allow the SVM to have a better defined separation hyperplane.

Another problem to look into is why the SVM predicted the BUY marker heavily. Although the BUY classification makes up 58% of the classifications, the majority is not large enough to explain the large amount of tests with only BUY predictions.

7 Acknowledgements

I would like to thank my advisors, Professor Aaron Cass and Professor Suthathip Yaisawarng, for providing all the help I needed to complete this project and more. I would also like to extend my appreciation to David Fuller in Data Services for teaching me how to efficiently collect data. I would also like to thank My Hoang for the generous amount of her time she dedicated to helping proofread my paper multiple times. I would finally like to thank all of the professors in the Economics Department and the Computer Science department for giving the advice necessary for the completion of this project.

References

[Chen, 2009] Chen, S.-S. (2009). Predicting the bear stock market: Macroeconomic variables as leading indicators. *Journal of Banking & Finance*, 33(2):211–223.

[Cragg and Malkiel, 1982] Cragg, J. G. and Malkiel, B. G. (1982). *Expectations and the structure of share prices*. University of Chicago Press.

[Huang et al., 2005] Huang, W., Nakamori, Y., and Wang, S.-Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, 32(10):2513–2522.

[Pan and Poteshman, 2006] Pan, J. and Poteshman, A. M. (2006). The information in option volume for future stock prices. *Review of Financial Studies*, 19(3):871–908.

[Pontiff and Schall, 1998] Pontiff, J. and Schall, L. D. (1998). Book-to-market ratios as predictors of market returns. *Journal of Financial Economics*, 49(2):141–160.

8 Appendix

A 1

The results in Table 9 still suggest the SVM is favoring the BUY signal too heavily. The next test performed was to determine if the SVM simply heavily chooses the BUY signal because the dataset has 58% of the actual signals as BUY. This next test split the training dataset evenly with 1765 datapoints with the marker of BUY and 1764 datapoints with the marker of DNB. As Table 3 shows, this partition doubled the amount of DNB signals predicted, however BUY was still heavily favored. Only 20% of the datapoints were classified as DNB. Table 3 displays the output of the SVM on this test and shows that the sector variable does not make a significant difference in the output.

Result	Actual	Aggregate Dataset	Sector Dataset
False Buy	0	1419	1474
True Buy	1765	1419	1499
False DNB	0	346	266
True DNB	1764	345	290
Percentage:	100%	49.99%	50.69%

Table 9: Evenly Partitioned by Percentage Sector Composes Dataset

B 2

The SVM builds a feature space from the input variables and four of the variables: US Interest Rates, Labor Force, US Industrial Production and US Inflation, have the same value for every company at the same date. Therefore, the next test performed removed these specific variables to test whether the SVM weighted the BUY signal so heavily due to the large feature space filled with data that would not allow the SVM to differentiate between datapoints. Table 10 shows this test with and without the nominal sector variable. The results are not significant enough to support my hypothesis.

Result	Actual	With Sector Variable	Without Sector Variable
False Buy	0	1588	1606
True Buy	2436	2216	2243
False DNB	0	220	193
True DNB	1764	176	158
Percentage:	100%	56.95%	57.16%

Table 10: Reduced Feature Space Through Backwards Elimination of Features