Corporate Financial Distress and Bankruptcy Prediction in the North American Construction Industry

William (Gang) Li

August 4, 2014

Professor Connel Fullenkamp, Faculty Advisor

Duke University
Durham, North Carolina
2014

William Li graduated with Distinction in Economics in May 2014. He will be starting full-time at J.P. Morgan Securities in New York City after graduation. Please direct all related questions to billyli200465@gmail.com

Acknowledgement

I would like to thank Professor Connel Fullenkamp for his time and advice, without which this project would not be possible. I would also like to thank Professor Michelle Connolly for her support and guidance. Finally I would like to thank Professor Jun Yang and my friend Dayvid Le for their help in producing the machine learning algorithm.

Abstract

This paper seeks to explore the application of Altman's bankruptcy prediction model in the construction industry by measuring its percentage accuracy on a dataset consisting of 108 bankrupt and non-bankrupt firms selected across the timeline of 1985-2013. The main goal of this paper is to explore the predictive power of an expanded variable set tailored to the construction industry compared to the original Altman model. Specifically, this measuring process is done using machine learning algorithm based on scikit-learn library that transforms a financial statement dataset of a company into clean vectorized feature matrix. The algorithm provides various classifiers to cross-validate the training set. Naive Bayes, Logit Regression, Support Vector Machine, K N_Neighbors, Tree, and Grid Search classifiers are used in this paper. The result shows no single dominant classifier that manages to predict bankruptcy more accurately than others, but non-linear classifiers tend to outperform their linear counterparts. Additionally, there is no clear preference in terms of the original 5-variable set versus the newly expanded construction specific 14-variable set, meaning that the Altman model stands both valid and effective in the context of bankruptcy prediction in the North American Construction Industry.

JEL Classification: C5; C38; G33; G34

Keywords: Corporate, Bankruptcy, Distress, Discriminant Analysis, Machine Learning

I Introduction

The role of high leverage in corporate restructuring and the popularity of junk bonds have been important aspects of the corporate finance scene in the 1980s. The interest in the prediction of corporate bankruptcy is increasing due to the implication associated with this phenomenon for investors, creditors, competitors and government. As seen in the chain of events following the 2008 financial crisis, the disruptive effects of corporate bankruptcy can create severe volatility spillovers to the broader economy. Debt instruments such as corporate bonds, bridge the gap between capital and economy, creditors and debtors; and because of the legal implications embedded in such instruments, failure to deliver scheduled payments result in default which might in turn trigger bankruptcy procedure.

The definition of a distressed situation is when the entity in question lacks the ability to fulfill the terms of obligation dictated by its existing contracts. The financing contracts of a firm can be loosely categorized into hard and soft contracts. An example of a hard contract is a coupon debt contract that specifies periodic payments by the firm to the bondholders. If these payments are not made on time, the firm is considered to be in violation of the contract and the claimholders have specified and unspecified legal recourse to enforce the contract. Common stock and preferred stock are examples of soft contracts. Here, even though its claimholders have expectations of receiving current payouts from the firm in addition to their ownership rights, the level and frequency of these payouts are often policy decision made by the firm, which can be suspended or postponed based on the availability of resources remaining in the firm. The assets of a firm also have a natural categorization based on liquidity. Cash or cash-like securities are liquid assets whereas long-term investments (plant and machinery) are hard assets.

The above categorizations of the financing contracts of a firm and its assets give rise to a natural definition of financial distress – a firm is considered in financial distress at a given point in time when the liquid assets of the firm are not sufficient to meet the current requirements of its hard contracts. (John, 1993)

To rectify this mismatch, the firm has to either increase the liquidity of the assets through asset sales or decrease the "hardness" of the debt contracts through debt renegotiation. One way to deal with financial distress is by informal reorganization of corporate financial structures through private workouts where a existing debt contract is replaced by a new contract with (i) a reduction in the required interest or principal payments, (ii) extension of maturity, or (iii) placement of equity securities with creditors.

An alternative mechanism for dealing with financial distress is the formal, court-supervised bankruptcy process governed by Chapter 11 of the U.S. Bankruptcy Code which grants protection against creditor harassment, allows the firm to issue form to issue new debt that is senior to all "prepetition" debt (debt-in-possession) and exempt the entity from any interest accrued on unsecured debt while the firm is in bankruptcy.

Altman (1968) explores how fundamental financial data and equity market values can be combined to effectively predict whether firms would go bankrupt in the U.S. In this paper, I wish to extend that model and apply it to the construction industry so that this distress-prediction model can be an important indicator of the future success of firms not just limited to U.S. public manufacturing companies. In literature review (Part II) I will discuss the work that has already been done previously in order to place my work in the context of current academic research. In theoretical framework (Part III) I will explain in detail the relevant theories that have been put forth and form

an empirically testable view myself. In empirical specification (Part IV) I begin to introduce the model that I plan to use for this paper, including the dependent/independent variables, estimates of regression coefficients and why these initial interpretation makes sense. In the data section I wish to manipulate and process the data obtained thoroughly to reach convincing evidence that supports my previous claims. Lastly, the conclusion (Part V) summarizes the project findings and discusses some potential policy implications of my study.

II Literature Review

As more and more firms have defaulted on their debt and filed for bankruptcy in the recent recession, investors have become increasingly interested in understanding how firms deal with financial distress – and a very active academic literature category has been developed on the topic. Although extensive research has been done on different ways of managing financial distress and their effects on the company through either public or private restructuring, the literature on bankruptcy prediction is still limited to certain sectors.

The earliest work done in the area of ratio analysis and bankruptcy classification was performed by Beaver (1967) who found that a number of indicators could discriminate between matched samples of failed and sound firms for as long as five year prior to failure. This study implied a definite potential of ratios as predictors of bankruptcy. Beaver defined failure as a business defaulting on interest payments on its debt, overdrawing its bank account, or declaring bankruptcy. Using univariate discriminant analysis, he studied large asset-size firms that failed during 1954-1964 and a stratified sample of successful firms. Beaver chose to test debt/total assets, earnings after taxes/-total assets and cash flow/total debt for his paper and concluded that the cash flow to debt ratio

was the best single ratio predictor. In general, ratios measuring profitability, liquidity and solvency prevailed as the most significant indicators, however the weight on each was not clear since most papers focused on univariate attempts instead of multivariate analysis.

Altman (1968) first attempted to incorporate multiple variables into one predictive function by using a linear discriminant analysis method to study credit risk measurement and developed a famous five-variable Z-score model. This model was generated from analysis of 22 financial ratios through a statistical filter that was based on a sample that consisted of 33 distressed and 33 non-distressed manufacturing companies. The firms selected have total assets ranging from 0.7 million to 25.9 million and the distressed group all declared Chapter X during the period 1946 to 1965. Out of the 22 variables initially selected, Altman found five ratios to be incorporated into the discriminant function: working capital/total assets, retained earnings before interest and taxes/total assets, market value of equity/book value of total debt, and sales/total assets. F-ratios for each variable and the entire equation were calculated and found significant for all but the sales/total assets ratio. Altman validated the function using the 66-firm holdout sample and achieved 79 percent accuracy one year before failure. However, none of the 66 firms used by Altman in his paper were construction companies.

The original Altman model has since been developed to fit several other industries with modification made to the coefficients including small firms (Edmister, 1972), banks, insurance companies, railroads, savings & loan associations telecommunication companies (Simonoff, 2013) and construction companies (Punsalan, 1989). Edmister (1972) develops and tests a number of methods of analyzing financial ratios to predict small business failure. Using data provided by the Small Business Administration and Robert Morris Associates, Edmister selected 42 companies with three

consecutive annual statements that are available prior to when the loan was granted. The seven-variable function estimated on the tri-annual sample prove to be quite different from the Altman model – the variables are all defined as dummy variables that takes on the value of "1" when the underlying financial ratio exceed a certain threshold. The ratios include annual finds flow/current liabilities ratio, equity/sales, current liabilities/equity, inventory/sales and quick ratio. This function correctly discriminates in 39 out of 42 cases (93 percent) when the decision rule is to predict failure if $z < 0.520$ and non-failure if $z > 0.520$. At the end of the paper, Edmister quality his function with the provision that at least three consecutive financial statements be available for analysis of a small business whereas both Altman and Beaver showed that one financial statement is sufficient for a highly discriminant function for large businesses.

Simonoff (2013) extends the Altman model to telecommunications industry by examining a sample size of 50 telecom companies, half of which declared bankruptcy between May 2000 and January 2002. The paper then discusses the effect of the five original Altman independent variables on bankruptcy prediction of this specific industry.

Punsalan (1989) tackled the same issue but the attempt was thwarted by the privacy of many construction firms at that time. The smaller firms do not have strong financial accounting systems and access to financial data was "practically nil". However, he compiled a list of ratios that he tested to be "significantly" different between contraction companies and Altman's underlying manufacturing companies.

Mason and Harris (1979) developed a six-variable Z-score model based on a sample of 20 failed and 20 non-failed companies in the civil engineering sector of the U.K. Applying the MDA, the discriminant function was developed with 6 main variables. A positive Z-score indicates a

long-term solvency, while a company with a negative value was classified as a potential failure.

In this paper I wish to reference the methods incorporated in these past papers and build on Punsalan's initial model using current data on construction sector.

III Theoretical Framework

The most important theoretical framework and the foundation of modern bankruptcy prediction is the Altman 1968 Z-Score model. The emergence of this model changed the univariate approach of measuring corporate bankruptcy put forth first by Beaver (1967) and the general distrust of traditional ratio analysis in the academia space. Altman derived financial statements from dated one annual reporting period prior to bankruptcy from the Moody's Industrial Manuals for 33 manufacturing firms. These firms filed a bankruptcy petition under Chapter X of the National Bankruptcy Act from 1946 through 1965. The second group of manufacturing firms is paired samples chosen on a stratified random basis with matching asset sizes. (Altman, 1968)

Z-score bankruptcy model for public manufacturing companies:

$Z = 1.2 * X_1 + 1.4 * X_2 + 3.3 * X_3 + 0.6 * X_4 + .999 * X_5$

X_1 = *Working Capital / Total Assets*

X_2 = *Retained Earnings / Total Assets*

X_3 = *Earnings Before Interest and Taxes / Total Assets*

X_4 = *Market Value of Equity / Total Liabilities*

X_5 = *Sales/ Total Assets*

Zones of Discrimination:

$Z > 2.99$ -"Safe" Zones

1.81 < Z < 2.99 -"Grey" Zones

Z < 1.81 -"Distress" Zones

The five variables incorporated in the function were selected from a list of 22 potentially helpful variables and do the best overall job together in the prediction of corporate bankruptcy.

X1 – Working Capital/Total Assets (WC/TA)

This ratio is a measure of the net liquid assets of the firm relative to the total capitalization. Working capital is defined as the difference between current assets and current liabilities. Ordinarily, a firm experiencing consistent operating losses will have shrinking current assets in relation to total assets. This is found to be the more helpful than current ratio and quick ratio. (Altman, 1968)

X2 – Retained Earnings/Total Assets (RE/TA)

Retained earning is the account that reports the total amount of reinvested earnings and/or losses of a firm over its entire life. This was considered a new ratio at the time of publication, and the age of a firm is implicitly considered in this ratio. A relatively young firm will have lower RE/TA ratio because it does not have enough time to build up its cumulative profits. Therefore, young firms are somewhat discriminated against in this analysis and its chance of being classified as bankrupt is relatively higher than that of another older firm, which is true in the real world. In addition, firms with high RE/TA have financed their assets through retention of profits rather than taking on more debt, which a healthy sign of growth most of the time. (Altman, 1968)

X3 – Earnings Before Interest and Taxes to Total Assets (EBIT/TA)

This ratio is a measure of the rue productivity of the firm's assets, independent of any tax or leverage factors. Since a firm's ultimate existence is based on the earning power of its assets, this ratio appears to be particularly appropriate for studies dealing with corporate failure. This ratio

continually outperforms other profitability measures, including cash flow. (Altman, 1968)

X4 – Market Value of Equity to Total Liabilities (MVE/TL)

Equity is measured by the combined market value of all shares of stock, preferred and common, while liabilities include both current and long-term. This ratio shows how much the firms' assets can decline in value before the liabilities exceed the assets and the firm becomes insolvent. For example, a company with a market value of its equity of $1,000 and debt of $500 could experience a two-thirds drop in asset value before insolvency. However, the same firm with $250 equity will be insolvent if assets drop only one-third in value. This ratio adds a market value dimension, which most other failure studies did not consider. (Altman, 1968)

X5 – Sales to Total Assets (S/TA)

The capital turnover ratio is a standard financial ratio illustrating the sales generating ability of the firm's assets. It is one measure of management's capacity in dealing with competitive conditions. This final ratio is quite important because it is the least significant ratio on an individual basis. However, because of its unique relationship to other variables in the model, it contributes greatly to the discriminating ability of the model. (Altman, 1968)

With a cutoff point at 2.67, the Type I accuracy of the model ranged from 82%-94% based on data from one financial statement prior to bankruptcy or default on outstanding bonds. Altman used the most recent data from 1997 to 1999 and reached an impressive 84% accuracy. (Altman, 2000) We can therefore conclude that the Z-score model has retained its reported high accuracy despite its development over 30 years ago. The above Z-Score model has become an established tool for assessing the creditworthiness of manufacturing firms, and is used in credit analysis in a variety of ways.

Since the original Z-score model, Altman expanded his theoretical framework to encompass all kinds of corporations both inside and outside of the United States. The first adaptation was for the private companies because credit analyst, private placement dealers, accounting auditors, and firms themselves are concerned that the original model is only applicable to publicly traded entities. Altman (2000) advocates a complete reestimation of the model rather than a simple substitution of book value for the market value in X4.

Z-score bankruptcy model for private companies:

$Z = 0.7 * X_1 + 0.847 * X_2 + 3.107 * X_3 + 0.420 * X_4 + 0.998 * X_5$

X_1 = *Working Capital / Total Assets*

X_2 = *Retained Earnings / Total Assets*

X_3 = *Earnings Before Interest and Taxes / Total Assets*

X_4 = *Book Value of Equity / Total Liabilities*

X_5 = *Sales/ Total Assets*

Zones of Discrimination:

$Z > 2.90$ -"Safe" Zones

$1.23 < Z < 2.90$ -"Grey" Zones

$Z < 1.21$ -"Distress" Zones

The X4 now has less of an impact on the Z-Score with the modification and therefore the gray area is wider. All of this indicates that the revised model is probably somewhat less reliable than the original but only slightly less. The next modification of the Z-Score is targeted at public non-manufacturing companies as well as emerging market companies. This model analyzes the characteristics and accuracy of a model without X5 – Sales/Total Assets, which minimize the po-

tential industry effect that is more likely to take place when such an industry-sensitive variable as asset turnover is included. In addition, Altman, Hatzell and Peck (1995) have applied this enhanced Z-Score model to emerging markets corporates, specifically Mexican firms that had issued Eurobonds denominated in U.S. dollars. The book value of equity was used for X4 in this case.

Z-score bankruptcy model for non-manufacturing companies:

$Z = 6.56 * X_1 + 3.26 * X_2 + 6.72 * X_3 + 1.05 * X_4$

X_1 = *Working Capital / Total Assets*

X_2 = *Retained Earnings / Total Assets*

X_3 = *Earnings Before Interest and Taxes / Total Assets*

X_4 = *Book Value of Equity / Total Liabilities*

All of the coefficients are changed, as are the group means and cutoff scores.

Z-score bankruptcy model for emerging market companies:

$Z = 3.25 + 6.56 * X_1 + 3.26 * X_2 + 6.72 * X_3 + 1.05 * X_4$

In the EM model, Altman added a constant term of +3.25 so as to standardize the scores with a score of zero equated to a default rated bond to make up for the inherent foreign exchange risk as well as the industry risk. For relative value analysis, the corresponding U.S. corporates' credit spread is added to the sovereign bond's option-adjusted spread. He also pointed that models for specific industries are an even better method for assessing distress potential of like-industry firms. (Altman, 2002), which is exactly what this paper attempts to do for the construction industry.

In 1977, Altman, Haldeman and Narayanan (1977) constructed a second-generation model with several enhancements to the original Z-Score approach. The purpose of the study was to construct, analyze and test a new bankruptcy classification model that considers explicitly recent

developments with respect to business failures. The new study also incorporated refinements in the utilization of discriminant statistical techniques. The new model, which is called ZETA, is effective in classifying bankrupt companies up to five years prior to failure on a sample of corporations consisting of manufacturers and retailers. Since the ZETA model is a proprietary effort, Altman didn't fully disclose the parameters of the market.

IV Empirical Specification

Two kinds of models are generally adopted for bankruptcy prediction: accounting ratios abased models and market based models (Santos et al, 2006). In the former, classical statistical techniques such as discriminant analysis or logistic regression models have been used, while in the latter the Moody's KMV model was adopted. The KMV model is essentially based on the market value of equity and its volatility and the equity market value serves as a proxy or the firm's asset value. In this paper I wish to follow the first approach but at the same time integrating some of the indicators because the construction industry is defined to be a cyclical industry.

The goal at first is to investigate whether bankrupt firms in manufacturing are distinguishable from bankrupt firms in the construction industry. A first step is to identify major significant differences in between their financial reporting. Once a particular ratio or set of ratios is determined to be distinguishable between manufacturing and construction, the problem remains on how this can affect an existing bankruptcy prediction model. To accomplish this, the analysis of variance, one-way classification fixed effect model can be used to determine significant difference in financial ratios between the two sectors. Punsalan (1989) categorized the construction and manufacturing industry

into six sub-sectors respectively and obtained average financial ratio of these groups from Almanac of business and Industrial Financial Ratios, Dun & Bradstreet Industry Norms and Key Business Ratios and Robert Morris Associates Annual Statement Studies. Using each ratio from the twelve industry types (six construction and six manufacturing types), an F-test was constructed to verify the degree of significance of difference between these two sectors. The resulting F > $F_?$ led to the conclusion that there is significant difference between the value of financial data for construction and manufacturing.

X1 – Total Liabilities to Net Worth (TL/NW)

Total Liabilities (debt) are all current liabilities and all long-term liabilities. This ratio measures the "creditor's equity" in assets of the business exceeds owners equity. The higher the ratio, the more risk being assumed by the creditors. From the standard ratios by industry Almanac, the construction industry's TL/NW ratio is double in value to that for manufacturing industry for reported corporation with and without net income. (Punsalan, 1989)

X2 – Total Debt to Total Assets (TD/TA)

This leverage ratio shows the extent in which the firms are financed by debt and indicates the firms' financial risk. It is somewhat similar to total liabilities/net worth. The higher the ratio, the more risk for creditors. It is not surprising here that the construction industry has a higher debt ratio compared to manufacturing. (Punsalan, 1989)

X3 – Return on Equity (Profit/Net Worth)

This ratio measures the rate of return on the investment in the business. The tendency in the industry is to look at this ratio as a final criterion of profitability. A high ratio is generally indicative of positive performance. However an unusually high ratio could indicate a company with too little

investment. A low ratio may indicate poor performance, conservative management or a mature company that has accumulated a significant amount of wealth relative to its established volume level. This ratio was more than double for the construction industry than that of manufacturing. (Punsalan, 1989)

X4 – Retained Earnings/Net Income

This ratio is the percentage of earnings in the business. For corporations with net income only, construction had approximately a third more earnings than manufacturing relative to net income. (Punsalan, 1989)

X5 – Quick Ratio (Cash + Accounts Receivables/Current Liabilities)

This ratio reveals the protection of short-term creditors through the firms' cash and near cash assets. The higher the ratio, the greater the liquidity. But if too high, the firm may have too much capital that is idle. Construction is usually more liquid than manufacturing with larger percentage of the construction industries assets tied into cash and receivables than in the manufacturing industry. (Punsalan, 1989)

X6 – Current Liabilities/Inventory and Net Sales/Inventory

These ratios are a measurement of how management controls inventory. For both ratios, the construction industry was significantly higher when compared with manufacturing. This says that construction has a smaller amount of inventory relative to sales and total liability as compared to manufacturing. This fact is true since construction contractors use subcontractors and do not normally hold materials in storage for long period of time. A low sales/inventory ratio usually indicates excessively high inventory. By the very nature of the manufacturing industry, these ratios are significantly more important to them than in construction. (Punsalan, 1989)

X7 – Current Ratio (Current Assets/Current Liabilities)

The current ratio compares the amount of current assets with which payments can be made to the amount of current liabilities requiring payment. The higher the current ratio, the more capable the company is of meeting its current obligations. Manufacturing has higher current ratio than construction because the construction industry in general incurs higher debt. At the same time, construction has less material inventory tied up from capital than does the manufacturing industry. (Punsalan, 1989)

X8 – Revenue/Working Capital

This ratio measures how working capital is used in the business. Too high a ratio may indicate that the company is doing too much work for the available working capital and a high sensitivity to a cash flow interruption. Too low a ratio may indicate an inefficient use of working capital, possibly due to poor market conditions or a poor marketing program. On average construction had a higher revenue/working capital ratio than manufacturing, which means a higher profit and earnings can be realized. (Punsalan, 1989)

X9 – Percent Profit Before Tax/Total Assets

This ratio reflects the pre-tax return on total assets and measures the effectiveness of the firm in utilizing the available resources. The higher the ratio, the more effective and efficient is the performance of management. Manufacturing has a significantly higher ratio than construction, which says that construction is less efficient than manufacturing and this is probably due to higher overhead costs and numerous unrealized contracts from loss bidding. (Punsalan, 1989)

X10 – Retained Earnings/Total Assets

This ratio measures the cumulative earnings over time. This is considered an very important

ratio because the age of the firm is implicitly considered in this ratio. The younger the firm, and lower the ratio usually. Manufacturing companies tend to have higher retained earnings/total assets ratio, but this might be more attributed to construction firms having a larger total assets and equipment than manufacturing firms being older. (Punsalan, 1989)

X11 – Working Capital/Total Assets

This liquidity ratio measures the net liquid assets relative to the firms' total capitalization. Because a firm experiencing consistent operating losses will have shrinking current assets in relation to the total assets (Altman, 1968), a higher ratio indicates more liquidity and better health. Manufacturing has higher ratio than construction, which is understandable since the construction industry on average borrows more of its capital relative to manufacturing. (Punsalan, 1989)

X12 – Earnings Before Interest and Taxes/Total Assets

This ratio measures the true productivity of the firms' assets. It is similar to the percent profit before tax/total assets ratio and thus yields the same results. Manufacturing has a higher EBIT/TA ratio than the construction industry. Insolvency in a bankruptcy sense occurs when the total liabilities exceed a fair valuation of the firm's assets with value determined by the earning power of the assets, that's why this ratio is worth is looking at. (Punsalan, 1989)

X13 – Sales/Net Worth

The sale to net worth ratio compares revenues to equity. This ratio is often referred to as the "turnover of equity", because it measures how the company's investment is applied in the business, aka how effective the company is using its investment. Construction has a higher sales/net worth ratio, which coincides with the ratio of return on equity "profit/net worth". (Punsalan, 1989)

X14 – Current Debt/Net Worth

This ratio recognizes that as net worth increases in relation to creditors equity, the risk assumed by the current creditors decreases. The higher the ratio, the more risk is being assumed by the creditors. Conversely, a lower ratio indicates a company with more borrowing capacity and greater long-term financial stability. Construction had a higher average current debt to net worth ratio than manufacturing. This ratio is similar and coincides with the results of total liabilities/net worth because construction borrows more for financing projects. (Punsalan, 1989)

V Data Analysis

From *Bankruptcy.com* a set of 96 construction & supply companies that declared bankruptcy between 1985 and 2011 was extracted. Given each company's TIC tag, the WRDS database successfully matched the information to its proprietary GVKey and compiled a list of 53 companies that are tracked by the system. This will be our bankruptcy set. In addition, a set of current active construction companies with similar capital sizes is created from publicly available data on Bloomberg and NYSE. This will be our non-bankruptcy set. Using this list of companies, I was able to pull data from Compustat North America for all the available financial statements that contains critical measurements mentioned above. All of the data was dumped into Excel and subsequently cleaned. Each row represents a company with five factors (wc_ta, re_ta, edit_ta, mve_tl, s_ta) as well as the corresponding calculation of Z-score.

Table 1 shows the initial prediction accuracy on recent bankruptcies within in the construction industry the year of bankruptcy filing, two years before and three years before respectively. Type I accuracy here indicates the robustness of prediction on firms who went bankrupt; Type II on

Table 1: Z-Score Results

		Type I Accuracy			Type II Accuracy		
		Success	Failure	Total	Success	Failure	Total
One-year Period	Observations	45	7	52	18	38	56
	Percentage	86.54%	13.46%	100%	32.14%	67.86%	100%
Two-year Period	Observations	38	14	52	17	39	56
	Percentage	73.08%	26.92%	100%	30.36%	69.64%	100%
Three-year Period	Observations	32	17	49	18	34	52
	Percentage	65.31%	34.69%	100%	34.62%	65.38%	100%

the other hand involves correctly predicting heathy companies who have not filed for bankruptcy. The original Z-Score have an 86.54% Type I prediction accuracy using the 2.675 cut-off, which falls between Altman's 84%-94% usual success rate with period one data. However, the standard deviation for the data is relatively large and there are dramatic swings in the predictive power of the original Z-Score on a case-by-case basis. Because I am using financial data only one year before the bankruptcy, 86.54% is not expected to be a stable performance given the proximity of the financial statement date and bankruptcy filing date. This concern is validated by period-two and period-three data, which shows a steady decline in the accuracy of Type I prediction - from 86.54% to 73.08% then 65.31%. This result is within expectation because as the timeline moves away from the actual bankruptcy event, predictive power of the same set of metric is expected to decrease.

However, the same rule does not apply for Type II accuracy. The success rate of assessing the healthiness of non-bankrupt companies has remained steady over the three periods. The cor-

rectness percentage floated around 32% but most of it can be attributed to the change in the base statistics of the total number of companies. The total number of companies changes between different years because of the availability of data. There are several companies with only two most recent years of valid data and therefore could not be incorporated into the third set. In addition, if there is a data gap between the second most recent year and the third most recent year, that third year cannot be counted towards the dataset either because we are looking for third year immediately precede the bankruptcy filing, any data that dates further than that loses credibility.

To make matters more complicated, there are companies that reemerges from previous bankruptcy that I had to adjust for. WCI Communities Inc. filed for Chapter 11 bankruptcy back in 2008 but reemerged after debt restructuring. The most recent data is of 2012 but the period-one measure must take 2008 as the more accurate measure of company health. 2012 data on the other hand cannot be used in the non-bankruptcy group because debt-laden firms that just come out of bankruptcy have subpar financial data compared to longtime healthy companies and would therefore induce ambiguity to the compare group. Similar examples include U.S. Concrete Inc. which filed for Chapter 11 back in 2010 but has data as recent as 2012; its 2012 data is also excluded from the compare group. William Lyon Homes filed for bankruptcy in 2011, its 2012 data is also omitted from the non-bankruptcy group.

There are still exceptions to the rule above. American Homestar Group filed for bankruptcy back in 2001 but successfully emerged after. The most recent data is in 2008 and I chose to include 2006-2008 data to the non-bankruptcy group because 5 years is usually enough time for a company to get back on track - most of the restructured short-term corporate debt would be paid down and reforms would be in effect.

USG is a more complex case - it filed for bankruptcy twice, once in 1993 and another in 2001. The company managed to survive two rounds of restructurings and is still active today. To effectively use USG's data, I decided to treat it as three separate entries - two in bankruptcy group with 1993 and 2001 as most recent data for 1yr, one in non-bankruptcy group with 2012 being the period-one data.

VI Machine Learning

Most of previous work has been done through either Multiple Discriminant Analysis or Logit Analysis ever since Altman first introduced the multiple-factor linear model. Although the MDA is called a "continuous scoring" system, a discriminant score is simply an ordinal measure that allows the ranking of firms (Altman, 1968) and it has a lot of disadvantages in view of tough premises. For instance, a sample should consist of multivariate normally-distributed observations with equal variance-covariance matrices. Neglecting these problems in most cases leads to test biases (Balcaen & Ooghe, 2006). Under these conditions, the popularity of the MDA method declined after the 1980s, along with the emergence of methods based on the logit methods (Elena Makeeva & Ekaterina Neretina, 2013). Typically, the logic models assume a logistic distribution (Maddala, 1977) while the probit models assume a cumulative normal distribution (Theil, 1971).

In essence, the methods listed above are all ways of classifying existing data into groups, in our case bankruptcy and non-bankruptcy. At its heart, the multiple discriminant analysis is a classification method which projects high-dimensional data onto a line and performs classification in this one-dimensional space (Fisher, 1936). The projection maximizes the distance between the means of the two classes while minimizing the variance within each class (Klecka, 1980). In a classi-

fication context, the essence of the logic regression model is that i assigns firms to the failing or the non-failing group based on their logit score and a certain cut-off score for the model. Both models are based on the resemblance principle whereby firms are assigned to the group they most closely resemble.

Table 2: Sample Training Data

Company Name	WC/TA	RE/TA	EBIT/TA	MVE/TL	S/TA	Label
CASTLE INDUSTRIES INC	-0.17	-0.77	-0.37	-0.11	0.42	0
CHAMPION ENTERPRISES INC	-0.058	-0.29	-0.019	0.15	1.602	0
HECHINGER CO -CL A	0.13	-0.11	-0.0005	0.17	2.18	0
MUELLER WATER PRODUCTS INC	0.30	-0.98	0.077	0.34	0.87	1
ARMSTRONG WORLD INDUSTRIES	0.22	-0.12	0.07	0.33	0.92	1
QUANEX BUILDING PRODUCTS	0.20	0.31	0.03	2.67	1.66	1

Classification is a form of machine learning and can be implemented using existing libraries and python code. Each company in this case would be an object and all the relevant financial ratios would be considered "features" of that object. If we have a pool of objects with different features - or in this case a set of companies with their own financial statistics, we can classify them into group using predefined "classifiers". MDA is a classifier, so is logit/probit. I produced a python code (appendix) that takes in raw .csv files with data and extract the ratios in order to join them with the corresponding company. So now instead of a raw .csv file, we have a dictionary of companies, which coupled with all the information we need to distinguish them from each other.

Table 2 shows an example of how the data is aligned in .csv file. The ultimate goal of the

classification lies in the last column "label". This column is the end result of prediction with the bankrupt firms generating a zero and non-bankrupt firms generating a one. Using the classifier of our choice, we then plug the dictionary of company, which comes from feature extraction, into ten fold cross-validation. Cross-validation is a step in which the whole data set gets divided into two parts, one for "training" and one for "testing". The amount of training data versus testing data is dictated by the number of "folds" - ten fold split the space into 10 pieces, each time using 9 to train and 1 to test, and the algorithm does this for 10 times. The way to "train" the data is specified by the corresponding classifier, different classifier fit the features to labels differently. The end output is an average of all the test accuracy and training accuracy taken from all 10 sessions.

Sample output:

- *fold 1: train accuracy: 100.00% test accuracy: 60.00%*
- *fold 2: train accuracy: 100.00% test accuracy: 70.00%*
- *fold 3: train accuracy: 100.00% test accuracy: 60.00%*
- *fold 4: train accuracy: 100.00% test accuracy: 60.00%*
- *fold 5: train accuracy: 100.00% test accuracy: 60.00%*
- *fold 6: train accuracy: 100.00% test accuracy: 55.56%*
- *fold 7: train accuracy: 100.00% test accuracy: 66.67%*
- *fold 8: train accuracy: 100.00% test accuracy: 66.67%*
- *fold 9: train accuracy: 100.00% test accuracy: 55.56%*
- *fold 10: train accuracy: 100.00% test accuracy: 55.56%*

training accuracy: 100.00%

test accuracy: 61.00%

Above is a example of test output from the training program. Although this particular test uses logit as classifier, I incorporate a wide range of classifiers to play around with:

- **Multinomial Naive Bayes**: MultinomialNB (alpha=0.2) where alpha is the smoothing parameter: 0 for no smoothing, and larger alpha means more smoothing – i.e. we lean more towards our prior belief and less towards training data.
- **Logistic Regression**: Classic logit model incorporated into the training library, one of the main tools for predicting bankruptcy
- **K N_Neighbors**: KNeighborsClassifier (n_neighbors=1) where n_neighbors specifies how many nearest neighbors we look at when classifying an input.
- **Support Vector Machine**: SVC (kernel='rbf', gamma=$1e-3$, C=1). For kernel='linear', parameter C trades off misclassification of training examples against simplicity of the decision surface. A low C makes the decision surface smooth, while a high C aims at classifying all training examples correctly. For kernel='rbf', in addition to C as explained above, there is another parameter gamma, which defines how far the influence of a single training example reaches, with low values meaning "far" and high values meaning "close".
- **Decision Tree**: tree.DecisionTreeClassifier ()
- **GridSearchCV**: this is basically SVM with automatic parameter tuning

After running all combinations of training on the data we obtained predicting accuracy, namely "test accuracy" for both the Altman model and my model respectively. The results are also divided by period of lags range from one to three, similar to how the data was organized.

Test Accuracy numbers are similar for 5-Variable and 14-Variable with 5-Variable performing better in most of the classifiers except for SVM and Grid Search. The accuracy of both sets hover around the 40% to 65% range with 1yr lag having no significant advantage over 2yr or 3yrs. Training accuracy, although not shown on table, is mostly 100%.

VII Conclusion

The 14-variable set works better in SVM and GridSearch; the 5-variable set works better in the others. A significant factor here is the classifier itself. SVM and GridSearch are both non-linear classifiers. Linear classifiers achieve classification by making a classification decision based on the value of a linear combination of features. SVM (support vector machine) is a non-probabilistic binary linear classifier, but it can also efficiently perform a non-linear classification by shifting its kernel to "rbf". As the number of incorporated features increase, more sophisticated methods yield better result compared to the naive ones. This could be the reason for the discrepancy between test accuracy numbers under different classifiers. For now, it is hard to say which set is better at predicting bankruptcy.

Additionally, the generally low accuracy (40%-65%) might be due to overfitting. As I mentioned above, training accuracy is 100% across the board meaning that the classifier follows the training religiously. Although this means we have less average loss over the training data, it also means the classifier would not have enough "flexibility" to deal with anything that is slightly out of pattern during the test session. Therefore, we have too few entries for too many features; expanding the dataset to incorporate more companies or slightly decrease the set of variables may help.

Lastly, the low numbers in accuracy may be attributable to the fact that machine learning

incorporates both Type I and Type II tests. In the initial Z-score test, Altman's MDA equation has impressive predictive power regarding Type I but has poor predictive power regarding Type II. These two effects may end up canceling out in the machine learning process and therefore result in a number within the middle range.

Looking forward, there is still work to be done on this subject matter. For the machine learning process, features can be more carefully selected. One example would be to rank all features by some utility measure, and use only the top k. A popular utility measure would be χ^2 where a high χ^2 means it is unlikely that the feature value and the class label are independent. Also the classifiers did not include neural network because the scikit-learn neural network library does not have "score" as one of the classifier's attributes, therefore it is hard to assess the performance of NN as a classifier in this case. Data envelopment analysis and option models can also be used to monitor the process and to provide new insight into the current model.

Table 3: Machine Training Results

		Test Accuracy	
		Altman 5-Variable	Expanded 14-Variable
Naive Bayes	1yr lag	48.18%	42.22%
	2yr lag	48.18%	42.22%
	3yr lag	48.55%	41.53%
Logit Regression	1yr lag	62.82%	61.00%
	2yr lag	65.55%	56.78%
	3yr lag	60.27%	57.36%
SVM (rbf)	1yr lag	51.82%	57.78%
	2yr lag	53.64%	57.78%
	3yr lag	51.45%	58.47%
K N_Neighbors	1yr lag	48.18%	46.56%
	2yr lag	49.09%	44.22%
	3yr lag	48.55%	42.64%
Tree	1yr lag	62.82%	55.56%
	2yr lag	64.64%	56.67%
	3yr lag	63.27%	57.36%
Grid Search	1yr lag	51.82%	57.78%
	2yr lag	60.09%	57.78%
	3yr lag	63.27%	58.47%

References

[1] Altman, Edward I. (1968). 'Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy'. *The Journal of Finance.* Vol. 23, No. 4, 589-609

[2] Altman, Edward I. (1971). 'Corporate Bankruptcy in America', *Health Lexington Books*

[3] Altman, Edward I., Haldeman, Robert G. & Narayanan, P. (1977). 'Zeta Analysis, A New Model to Identify Bankruptcy Risk of Corporations', *Journal of Banking and Finance*, 109-131

[4] Altman, Edward I. (1993). *Corporate Financial Distress and Bankruptcy*, 1st and 2nd editions.

[5] Altman, Edward I., J. Hartzell, and M. Peck. (1995). 'Emerging Markets Corporate Bonds: A Scoring System', *Future of Emerging Market Flows*

[6] Altman, Edward I. (2000). 'Predicting Financial Distress of Companies: Revisiting the Z-score and ZETA Models', *Unpublished Manuscript*, NYU

[7] Altman, Edward I. (2002). 'Corporate Distress Prediction Models in a Turbulent Economic and Basel II Environment', *Unpublished Manuscript*, NYU

[8] Altman, Edward I. (2005). 'An emerging market credit scoring system for corporate bonds', *Emerging Markets Review*, No. 6, 311-323

[9] Altman, Edward I. (2011). 'Transparent and Unique Sovereign Default Risk Assessment', *The Journal of Applied Corporate Finance*, Vol. 23, No. 3, Winter 2011

[10] Balcaen, S., & Ooghe, H. (2006). '35 years of studies on business failure: An overview of the classic statistical methodologies and their related problems'. *The British Accounting Review*, 38 (1), 63-93.

[11] Beaver, W. (1967). 'Financial Ratios as Predictors of Failures', *Journal of Accounting Research*

[12] Beaver, William H., (1966). 'Financial Ratios as Predictors of Failure', *Empirical Research in Accounting: Selected Studies*, Supplement of Accounting Research, 71-111

[13] Edmister, Robert O. (1972). 'An Empirical Test of Financial Ratio Analysis for Small Business Failure Prediction', *Journal of financial and Quantitative Analysis*, 1477-1493

[14] Elena, M. & Ekaterina N. (2013). 'The Prediction of Bankruptcy in a Contruction Industry of Russian Federation', *Journal of Modern Accounting and Auditing*, Vol. 9, No.2, 256-271

[15] Fisher, R. A. (1936) 'The use of multiple measurements in taxonomic problems', *Annals of Eugenics*, 7, 179-188.

[16] John K. (1993). 'Managing Financial Distress and Valuing Distressed Securities: A Survey and a Research Agenda'. *Financial Management*. Vol. 22, No. 3, 60-78

[17] John, K. & Vasudevan, G.K. (1992). 'Bankruptcy and Reorganization: A Theory of the Choice Between Workouts and Chapter 11', *Unpublished Manuscript*, NYU

[18] Klecka, W. R. (1980) 'Discriminant Analysis'. *Sage Publications*, Beverly Hills, CA.

[19] Lee S. & Choi W. (2012). 'A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis', *Elsevier*, Vol. 40, 2941-2946

[20] Maddala, G.S. (1977) *Econometrics*. McGraw-Hill, New York.

[21] Mason, R.J. and Harris, F.C. (1979) Predicting company failure in the construction industry', *Proceedings Institution of Civil engineers*, 66, 301-307.

[22] Punsalan R.N. (1989). 'Bankruptcy prediction in the construction industry: financial ratio analysis', *Unpublished Manuscript*, Georgia Institute of Technology School of Civil Engineering

[23] Santos M.F., Cortez P., Pereira J. & Quintela H. (2006). 'Corporate bankruptcy prediction using data mining techniques', *WIT Transactions on Information and Communication Technologies*, Vol. 37, 349-357

[24] Simonoff J.S. & Lui J. (2013). 'Predicting bankruptcy in the telecommunications industry', *Unpublished Manuscript*, NYU

[25] Theil, H. (1971) *Principles of Econometrics*. J. Wiley and Sons, New York.

[26] Zhang, L., Altman, E.I. & Yen, J. (2010). 'Corporate financial distress diagnosis model and application in credit rating for listing firms in China', *Front computer science China*, 4 (2), 220-236

Appendix I

```
@author: LG

from sklearn import cross_validation

from sklearn import feature_extraction

from sklearn.naive_bayes import MultinomialNB

from sklearn.neighbors import KNeighborsClassifier

from sklearn.svm import SVC

from sklearn.grid_search import GridSearchCV

from sklearn import tree

from sklearn import linear_model

from sklearn.neural_network import BernoulliRBM

import csv

import numpy


def parseFile (filename):

    companies = list ()

    with open (filename) as csvfile:

        reader = csv.reader (csvfile, delimiter=',', quotechar='"')

        for index, line in enumerate (reader):

            #print index, line

            if  (index > 0 and index < 110):

                company_name, tl_nw, td_ta, p_nw, re_ni, quick, cl_inv, current, rev_wc,

                    gp_ta, re_ta, wc_ta, ebit_ta, s_nw, cl_nw, nonbankrupt = line

                #print company_name
```

```
            company = {\
                'company_name' : company_name,\
                'tl_nw' : tl_nw,\
                'td_ta' : td_ta,\
                'p_nw' : p_nw,\
                're_ni' : re_ni,\
                'quick' : quick,\
                'cl_inv' : cl_inv,\
                'current' : current,\
                'rev_wc' : rev_wc,\
                'gp_ta' : gp_ta,\
                're_ta' : re_ta,\
                'wc_ta' : wc_ta,\
                'ebit_ta' : ebit_ta,\
                's_nw' : s_nw,\
                'cl_nw' : cl_nw,\
                'nonbankrupt': int (nonbankrupt),\
            }
            companies.append (company)
    return companies


def extract_tl_nw_features (companies):
    tl_nw_list = list ()
    for company in companies:
        tl_nw_list.append (company['tl_nw'] * 10)
```

```
    tweet_vectorizer = feature_extraction.text.CountVectorizer ()
    # alternatives:
    # tweet_vectorizer = feature_extraction.text.TfidfVectorizer ()
    X = tweet_vectorizer.fit_transform (tl_nw_list).toarray ()
    return X


def extract_td_ta_features (companies):
    td_ta_list = list ()
    for company in companies:
        td_ta_list.append (company['re_ta'] * 10)
    tweet_vectorizer = feature_extraction.text.CountVectorizer ()
    # alternatives:
    # tweet_vectorizer = feature_extraction.text.TfidfVectorizer ()
    X = tweet_vectorizer.fit_transform (td_ta_list).toarray ()
    return X


def extract_p_nw_features (companies):
    p_nw_list = list ()
    for company in companies:
        p_nw_list.append (company['p_nw'] * 10)
    tweet_vectorizer = feature_extraction.text.CountVectorizer ()
    # alternatives:
    # tweet_vectorizer = feature_extraction.text.TfidfVectorizer ()
    X = tweet_vectorizer.fit_transform (p_nw_list).toarray ()
    return X
```

```
def extract_re_ni_features (companies):
    re_ni_list = list ()
    for company in companies:
        re_ni_list.append (company['re_ni'] * 10)
    tweet_vectorizer = feature_extraction.text.CountVectorizer ()
    # alternatives:
    # tweet_vectorizer = feature_extraction.text.TfidfVectorizer ()
    X = tweet_vectorizer.fit_transform (re_ni_list).toarray ()
    return X


def extract_quick_features (companies):
    quick_list = list ()
    for company in companies:
        quick_list.append (company['quick'] * 10)
    tweet_vectorizer = feature_extraction.text.CountVectorizer ()
    # alternatives:
    # tweet_vectorizer = feature_extraction.text.TfidfVectorizer ()
    X = tweet_vectorizer.fit_transform (quick_list).toarray ()
    return X


def extract_cl_inv_features (companies):
    cl_inv_list = list ()
    for company in companies:
        cl_inv_list.append (company['cl_inv'] * 10)
```

```
    tweet_vectorizer = feature_extraction.text.CountVectorizer ()
    # alternatives:
    # tweet_vectorizer = feature_extraction.text.TfidfVectorizer ()
    X = tweet_vectorizer.fit_transform (cl_inv_list).toarray ()
    return X


def extract_current_features (companies):
    current_list = list ()
    for company in companies:
        current_list.append (company['current'] * 10)
    tweet_vectorizer = feature_extraction.text.CountVectorizer ()
    # alternatives:
    # tweet_vectorizer = feature_extraction.text.TfidfVectorizer ()
    X = tweet_vectorizer.fit_transform (current_list).toarray ()
    return X


def extract_rev_wc_features (companies):
    rev_wc_list = list ()
    for company in companies:
        rev_wc_list.append (company['rev_wc'] * 10)
    tweet_vectorizer = feature_extraction.text.CountVectorizer ()
    # alternatives:
    # tweet_vectorizer = feature_extraction.text.TfidfVectorizer ()
    X = tweet_vectorizer.fit_transform (rev_wc_list).toarray ()
    return X
```

```
def extract_gp_ta_features (companies):
    gp_ta_list = list ()
    for company in companies:
        gp_ta_list.append (company['gp_ta'] * 10)
    tweet_vectorizer = feature_extraction.text.CountVectorizer ()
    # alternatives:
    # tweet_vectorizer = feature_extraction.text.TfidfVectorizer ()
    X = tweet_vectorizer.fit_transform (gp_ta_list).toarray ()
    return X


def extract_re_ta_features (companies):
    re_ta_list = list ()
    for company in companies:
        re_ta_list.append (company['re_ta'] * 10)
    tweet_vectorizer = feature_extraction.text.CountVectorizer ()
    # alternatives:
    # tweet_vectorizer = feature_extraction.text.TfidfVectorizer ()
    X = tweet_vectorizer.fit_transform (re_ta_list).toarray ()
    return X


def extract_wc_ta_features (companies):
    wc_ta_list = list ()
    for company in companies:
        wc_ta_list.append (company['wc_ta'] * 10)
```

```
    tweet_vectorizer = feature_extraction.text.CountVectorizer ()
    # alternatives:
    # tweet_vectorizer = feature_extraction.text.TfidfVectorizer ()
    X = tweet_vectorizer.fit_transform (wc_ta_list).toarray ()
    return X


def extract_ebit_ta_features (companies):
    ebit_ta_list = list ()
    for company in companies:
        ebit_ta_list.append (company['ebit_ta'] * 10)
    tweet_vectorizer = feature_extraction.text.CountVectorizer ()
    # alternatives:
    # tweet_vectorizer = feature_extraction.text.TfidfVectorizer ()
    X = tweet_vectorizer.fit_transform (ebit_ta_list).toarray ()
    return X


def extract_s_nw_features (companies):
    s_nw_list = list ()
    for company in companies:
        s_nw_list.append (company['s_nw'] * 10)
    tweet_vectorizer = feature_extraction.text.CountVectorizer ()
    # alternatives:
    # tweet_vectorizer = feature_extraction.text.TfidfVectorizer ()
    X = tweet_vectorizer.fit_transform (s_nw_list).toarray ()
    return X
```

```
def extract_cl_nw_features (companies):
    cl_nw_list = list ()
    for company in companies:
        cl_nw_list.append (company['cl_nw'] * 10)
    tweet_vectorizer = feature_extraction.text.CountVectorizer ()
    # alternatives:
    # tweet_vectorizer = feature_extraction.text.TfidfVectorizer ()
    X = tweet_vectorizer.fit_transform (cl_nw_list).toarray ()
    return X


def extract_all_features (companies):
    return numpy.concatenate ( (extract_tl_nw_features (companies), \
                            extract_td_ta_features (companies), \
                            extract_p_nw_features (companies), \
                            extract_re_ni_features (companies), \
                            extract_cl_inv_features (companies), \
                            extract_current_features (companies), \
                            extract_rev_wc_features (companies), \
                            extract_gp_ta_features (companies), \
                            extract_re_ta_features (companies), \
                            extract_wc_ta_features (companies), \
                            extract_ebit_ta_features (companies), \
                            extract_s_nw_features (companies), \
                            extract_cl_nw_features (companies), \
```

```
                        extract_quick_features (companies)), \
                        axis=1)


def generate_target (companies):
    # Non-Bankruptcies are coded 1, Bankruptcies 0:
    y = [company['nonbankrupt'] for company in companies]
    return numpy.array (y)


def get_classifier ():
    #
    ######################################################################
    # Multinomial Naive Bayes
    #return MultinomialNB (alpha=0.2)
    ######################################################################
    #Logit
    #return linear_model.LogisticRegression ()
    ######################################################################
    # KNN
    #return KNeighborsClassifier (n_neighbors=1)
    ######################################################################
    # SVM
    #return SVC (kernel='linear', gamma=1e-3, C=1)
    ######################################################################
    # Decision Tree Classifier
    #return tree.DecisionTreeClassifier ()
```

```
    ########################################################################
    # SVM, with automatic parameter tuning
    # Warning: SVM training can be really slow!
    classifier = SVC ()
    tuned_parameters = [{'kernel': ['rbf'], 'C': [0.1, 1, 10, 100]},
                        {'kernel': ['linear'], 'C': [0.1, 1, 10, 100]}]
    return GridSearchCV (classifier, tuned_parameters, scoring='f1', cv=5, verbose=2)
    ########################################################################
    #Neural Network
    #return BernoulliRBM (n_components=2)
    ########################################################################



def cross_validate (data, target, classifier, num_folds):
    cv = cross_validation.StratifiedKFold (target, n_folds=num_folds)
    train_accuracies = list ()
    test_accuracies = list ()
    print '{} folds:'.format (num_folds)
    k = 0
    coefs=[]
    for train, test in cv:
        k += 1
        print 'fold {}: '.format (k), # tracking the fold iteration
        X_train, y_train = data[train], target[train]
        X_test, y_test = data[test], target[test]
```

```
        classifier.fit (X_train, y_train)
        train_accuracy = classifier.score (X_train, y_train)
        train_accuracies.append (train_accuracy)
        coefs.append (classifier.best_estimator_)
        print 'train accuracy: {:.2%}'.format (train_accuracy),
        test_accuracy = classifier.score (X_test, y_test)
        test_accuracies.append (test_accuracy)
        print 'test accuracy: {:.2%}'.format (test_accuracy)
    avg_train_acc = sum (train_accuracies) / float (len (train_accuracies))
    avg_test_acc = sum (test_accuracies) / float (len (test_accuracies))
    print coefs[0]
    return avg_train_acc, avg_test_acc


companies = parseFile ("14-1yr-training-set.csv")
X = extract_all_features (companies)
y = generate_target (companies)


classifier = get_classifier ()
train_acc, test_acc = cross_validate (X, y, classifier, 10)
print 'training accuracy: {:.2%}'.format (train_acc)
print 'test accuracy: {:.2%}'.format (test_acc)
```

Appendix II

Chapter 7 Individual Bankruptcy

Chapter 7 bankruptcy, the most severe form of bankruptcy, governs the process of liquidation. This form of bankruptcy is available to both individuals and companies. When an individual files this form of bankruptcy, most unsecured debts are legally dismissed. But liens (mortgages or vehicles) are not discharged. Furthermore, Chapter 7 bankruptcy cannot dissolve debt such as child support and student loans.

Chapter 7 Business and Corporation Bankruptcy

Chapter 7 business and corporation bankruptcy typically causes a business to cease all operations, depending on the amount of and type of debt and the advisement of the bankruptcy trustee. For example, a company could be closed or it could be sold to another investor if approved by the bankruptcy trustee. Furthermore, in a Chapter 7 filing, corporations and partnerships do not receive a discharge on existing debts; instead, the corporation or partnership is dissolved. As in the case of an individual filing Chapter 7, however, most debts, depending on their nature, are discharged.

Chapter 11 Personal Bankruptcy & Business and Corporation Bankruptcy

Chapter 11 bankruptcy is available to both individuals and all types of corporations, businesses and partnerships. Instead of discharged debt like the Chapter 7 bankruptcy plan, Chapter 11 is a restructuring of existing debts. Moreover, a corporation, business or partnership remains in complete control of its assets, whereas in Chapter 7 business divisions are either closed or sold.

Chapter 13

Unlike Chapter 7 and 11, Chapter 13 bankruptcy exists mainly for individuals and is the rarest form of bankruptcy filed. In fact, it is nearly the opposite of chapter 7 bankruptcy. For example, in a Chapter 13, instead of an individual's non-secured debts being discharged, the individual presents a plan to repay the debt within a period of three to five years.