Forecasting Corporate Bankruptcy: Applying Feature Selection Techniques to the Pre-and Post-Global Financial Crisis Environments

Parker S. Levi

Professor Andrew J. Patton, Faculty Advisor
Professor Michelle P. Connolly, Faculty Advisor

Honors Thesis submitted in partial fulfillment of the requirements for Graduation with Distinction in Economics in Trinity College of Duke University.

Duke University
Durham, North Carolina
April 20, 2020

Acknowledgements

I would like to thank my two faculty advisors, Dr. Andrew Patton and Dr. Michelle Connolly, for their continued support and guidance throughout the past eight months. Dr. Patton's expertise helped me overcome various roadblocks related to the data collection and the analysis processes of my research, and I am very grateful for his many hours of assistance. Additionally, Dr. Connolly's feedback on both my honors thesis seminar presentations and my paper itself allowed me to improve the quality of my research topic, as well as the way in which I present my results. I also sincerely appreciate the consistent questions, comments and feedback that I received from my classmates in my honors thesis seminar. Lastly, I am grateful for the resources provided by Duke University, including access to Wharton Research Data Services, JSTOR and Google Scholar.

Abstract

I investigate the use of feature selection techniques to forecast corporate bankruptcy in the years before, during and after the global financial crisis. Feature selection is the process of selecting a subset of relevant features for use in model construction. While other empirical bankruptcy studies apply similar techniques, I focus specifically on the effect of the 2007-2009 global financial crisis. I conclude that the set of bankruptcy predictors shifts from accounting variables before the financial crisis to market variables during and after the financial crisis for one-year-ahead forecasts. These findings provide insight into the development of stricter lending standards in the financial markets that occurred as a result of the crisis. My analysis applies the Least Absolute Shrinkage and Selection Operator (LASSO) method as a variable selection technique and Principal Components Analysis (PCA) as a dimensionality reduction technique. In comparing each of these methods, I conclude that LASSO outperforms PCA in terms of prediction accuracy and offers more interpretable results.

JEL Classification: G01, G1, G33, C3, C53, C58

Keywords: corporate bankruptcy, forecasting, feature selection, global financial crisis

Introduction

This paper explores the topic of forecasting corporate bankruptcy for public companies in the United States. Corporate bankruptcy occurs when a business cannot repay its outstanding debt. It is a set of legal proceedings carried out to allow businesses freedom from their debts, while also providing creditors the opportunity for repayment. I ultimately hope to shed new light on corporate bankruptcy forecasting in two significant ways. First, I aim to identify sets of predictive variables correlated to bankruptcies before, during and after the 2007-2009 financial crisis to highlight any potential differences between these sets of predictors that may potentially occur from shifts in lending standards in the financial markets. Second, I seek to compare the prediction accuracy of two sophisticated bankruptcy forecasting techniques; in particular, I apply the Least Absolute Shrinkage and Selection Operator (LASSO) method and Principal Component Analysis (PCA).

The global financial crisis has been analyzed in many different ways, but there has been no published literature about the impact of the crisis on corporate bankruptcy forecasting. Of course, there have been plenty of published studies on bankruptcy related to the 2007-2009 financial crisis, as the aftermath of the crisis led to a large hike in the number of corporate bankruptcies in the United States. However, corporate bankruptcy *forecasting* is a much smaller subset of research within bankruptcy; corporate bankruptcy forecasting applies various statistical and econometric models with the goals of identifying relevant predictive variables and improving bankruptcy prediction accuracy. This is a more quantitative and niche approach compared to traditional bankruptcy research. This paper focuses on this smaller field of corporate bankruptcy research, as I shed light on how the global financial crisis may have impacted the set of relevant bankruptcy predictors relied upon by researchers. Additionally, while LASSO and

PCA are used as feature selection techniques in a wide variety of econometric research studies, both of these techniques are relatively new to the field of corporate bankruptcy forecasting. I aim to become the first study to directly compare the accuracy of LASSO to that of PCA in the context of forecasting corporate bankruptcy.

A variety of financial and market variables, such as stock price, stock volatility, book leverage and market leverage, have already been found statistically significant for forecasting bankruptcy in prior research. I aim to offer empirical insight related to an aspect of variable selection that has not yet been explored in any published bankruptcy forecasting studies by comparing pre-and post-global financial crisis predictive variable sets. In selecting these sets of variables, I utilize two feature selection techniques; while LASSO is a variable selection technique that simplifies my exhaustive list of variables to a sparse variable set solution, PCA identifies the underlying principal components of my variable set and variable subgroups, effectively creating new variables in doing so. I determine which of these methods performs better by measuring and comparing the prediction accuracy of each one.

The implications of bankruptcy forecasting are vast. Deriving variables that can predict bankruptcy may establish warning signs for companies that may not yet be under financial distress and could also highlight a potential market inefficiency for investors. Additionally, bankruptcy forecasting may help regulators design more efficient lending regulations in the credit markets. The other significant implication is the opportunity to test prior bankruptcy theories; prior studies have debated whether to include market variables in forecast models in addition to accounting variables, as well as whether financial ratios that incorporate the market values of equity and leverage are better predictors than ratios that incorporate the book values of equity and leverage. By applying feature selection techniques to the pre-and post-global financial

crisis time periods, I investigate whether distinct predictive variable sets exist for each of these environments.

In order to effectively conduct my empirical analysis, I construct a bankruptcy database by merging daily CRSP equity data with annual COMPUSTAT accounting data, spanning the 1989-2018 period. I use this data to compile a comprehensive candidate bankruptcy predictor set of 25 accounting-based and market-based variables that have been used in prior bankruptcy forecasting studies.

After compiling my dataset, I combine various approaches to forecasting corporate bankruptcy that are discussed in notable papers (Tian, Yu & Guo 2015 and Tsai 2009) in order to find sets of relevant predictive variables. I first execute the LASSO method on my data, using my extensive list of variables that I have compiled from previous studies to select those that are the most relevant bankruptcy predictors. LASSO provides a sparse variable set solution through a selection method in which it zeroes some coefficients. The technique uses a roughness penalty, which allows it to select the subset of variables that best forecasts bankruptcy.

The PCA method reduces the dimensionality of a dataset in which there are a large number of correlated variables. PCA achieves this by transforming the input data vector into a new variable set, which contains linearly uncorrelated and orthogonal principal components. These principal components are ordered so that the first several retain most of the variance present in the original dataset. I apply the PCA method to derive the principal components of both my entire variable set and my variable subgroups.

I use the techniques described above to select the most relevant predictive variables for subsamples before, during and after the financial crisis of 2007-2009 with the goal of highlighting any potential differences between bankruptcy predictors. All of the prior research

that I have read on my topic uses datasets that do not capture the periods during and after the financial crisis; these time periods exhibited both regulatory and macroeconomic changes in the financial services industry that may have potentially impacted the optimal set of bankruptcy predictors. In addition to selecting predictive variables corresponding to these different financial environments, I also analyze and compare the accuracy of my two feature selection techniques. In particular, I apply the Discrete Hazard Model, used in several relevant forecasting papers, to model bankruptcy risk for each of these techniques. The Discrete Hazard Model is simply a time-varying logistic regression; it takes into account the fact that firms change over time and can be used as a standardized method for measuring the prediction accuracy of each of my techniques.

Literature Review

Economists and statisticians have developed variable selection methods for forecasting corporate bankruptcy to identify relevant predictive variables and to improve prediction accuracy. Original research studies conducted by Beaver (1966), Altman (1968) and Ohlson (1980) focus exclusively on accounting variables for estimating default risk. Shumway (2001) and Campbell, Hilscher & Szilagyi (2008) attempt to improve the accuracy of these original accounting variable models by including market-based variables in their research. While their models exhibit improvements in accuracy over the original models, their studies do not apply any variable selection techniques to determine the relative importance of their newly incorporated variables compared with previously used bankruptcy predictive variables. Tian, Yu & Guo's (2015) study aims to fill this gap by applying the Least Absolute Shrinkage and Selection Operator (LASSO) variable selection technique to corporate bankruptcy forecasting.

LASSO provides a sparse variable set solution by applying a selection method that zeroes coefficients of variables that are not relevant predictors. Tian et al. (2015) utilizes the LASSO method with the goal of identifying a parsimonious set of the most relevant bankruptcy predictors. They choose to use LASSO because it has several advantages over other commonly used variable selection techniques. First, LASSO naturally overcomes multicollinearity, which otherwise poses a significant problem when considering an exhaustive set of correlated variables. Second, the technique is computationally efficient and can consider a large number of candidate predictive variables. Moreover, LASSO's process of selecting variables results in a variable selection path that can be used to interpret the relative importance of selected variables.

The bankruptcy predictors selected through Tian et al.'s (2015) LASSO technique offer interesting insight regarding the variables selected in previous bankruptcy models. Their study compiles an exhaustive list of 39 financial and market variables that are used as predictors in prior bankruptcy studies. The study first focuses on selecting variables for forecasting one year ahead of bankruptcy; LASSO selects seven predictive variables using data spanning 1980-2009. The seven selected variables are price, firm stock volatility, net income to market value of total assets, total debt to total book assets (book leverage), total liabilities to market value of total assets (market leverage), excess stock return and current liabilities to total book assets. These variable selection results reinforce the opinions of Shumway (2001) and Campbell et al. (2008), who both argue that both market variables and accounting variables should be incorporated in forecasting.

The LASSO variable selection analysis also sheds light on the difference of opinion between Shumway (2001) and Campbell et al. (2008) regarding the predictive power of ratios that incorporate the market value of equity versus the book value of equity. While Shumway

(2001) argues that the net income to total assets ratio and the total liabilities to total assets ratio constructed from accounting data are statistically significant predictors, Campbell et al. (2008) suggests that these ratios should be modified to use the market value of assets. LASSO selects both of these variables proposed by Campbell et al. (2008) but does not select either of Shumway's (2001) accounting-based ratios. While Tian et al.'s (2015) results imply that a firm's market value of equity is a more accurate gauge of bankruptcy than its book value of equity, they find that book leverage is a statistically significant bankruptcy predictor that is complement to the predictive power of market leverage. Campbell et al. (2008) stresses the importance of market leverage in predicting bankruptcy, and LASSO determines that in addition to market leverage, a firm's current liabilities to total assets ratio and total debts to total assets ratio constructed from accounting data also hold predictive power. This finding can be justified by the fact that book leverage can serve as a proxy for precautionary actions taken by a firm to reduce its bankruptcy risk; firms rarely alter their capital structure as a result of stock price fluctuations, but they may take action on the premise of limiting risks that appear on their balance sheets.

In addition to selecting variables to forecast one year ahead of bankruptcy, Tian et al.'s (2015) study explores how variable selection changes with forecast horizon. They find that LASSO selects 5 market variables and 2 accounting variables for forecast horizons within 2 years, while it selects 5 accounting variables and 2 market variables for 3-and 5-year forecast horizons. The researchers reason that firms are more susceptible to large idiosyncratic shocks as a result of stock price fluctuations in the short-term, while this idiosyncratic risk balances out in the long run, leading to a relative increase in predictive power for accounting variables.

Principal Component Analysis (PCA) is another approach to variable selection that has been applied to bankruptcy forecasting. The main goal of PCA is to reduce the dimensionality of

a dataset in which there are correlated variables, while retaining as much of the variance as possible from the original data. PCA accomplishes this by transforming the input data vector to a linear combination of uncorrelated components. Tsai (2009) finds in his study that PCA is more accurate in predicting bankruptcy for certain datasets than other variable selection methods. PCA is particularly useful when analyzing a large set of interrelated variables, as it solves the issue presented by methods that may select correlated variables and overlook their underlying principal components.

Various research studies conducted by Shumway (2001), Chava and Jarrow (2004), Campbell et al. (2008), Tian et al. (2015) and others have modeled bankruptcy risk using the Discrete Hazard Model. According the Shumway (2001), this model is widely accepted because it takes into account that firms change over time by using time-varying panel data, which poses significant advantages over static models used in older studies, such as Altman (1968) and Ohlson (1980). Tian et al. (2015) utilizes the Discrete Hazard Model as a consistent method for estimating bankruptcy risk across different sets of selected variables; their study measures the prediction accuracy of the variables selected by the LASSO technique compared to the variables used in Campbell et al.'s (2008) model by plugging each of these variable sets into the Discrete Hazard Model and comparing several evaluation metrics. Tian et al. (2015) ultimately concludes that LASSO-selected variables outperform those advocated by Campbell et al. (2008).

This prior research has provided me with a comprehensive set of candidate bankruptcy predictors, as well as an in-depth understanding of various forecast techniques. I use the methods discussed in these research papers, focusing on applying LASSO and PCA, to identify different sets of relevant bankruptcy predictors for the pre-and post-global financial crisis environments.

Empirical Framework

As discussed in the Literature Review section, economists have developed variable selection techniques for forecasting corporate bankruptcy with the objectives of (1) identifying relevant predictive variables and (2) improving prediction accuracy. With regard to this first objective, I aim to test whether the variables selected from my initial set of bankruptcy predictors change in the aftermath of the global financial crisis. The only published study to apply the LASSO technique to bankruptcy forecasting, Tian et al. (2015), concludes that variables selected by LASSO do not change across subsample periods; the same set of predictive variables is selected over the 1980-2000, 1980-2002, 1980-2005 and 1990-2009 subsample periods. Other variable selection studies using different techniques also find the same consistency across subsample periods. However, to the best of my knowledge, no published bankruptcy forecasting study analyzes data after 2009. Following the global financial crisis, corporate lending standards became more stringent as a result of regulatory changes. Through my empirical analysis, I identify sets of relevant bankruptcy predictors corresponding to bankruptcies in the years before, during and after the 2007-2009 financial crisis. This sheds light on any differences between these sets of bankruptcy forecasting factors that could possibly be due to these stricter lending standards.

Regarding my second objective of improving prediction accuracy, I analyze the prediction accuracy of two feature selection methods that have not yet been compared with one another in the same study: LASSO and PCA. Both LASSO and PCA have been applied to bankruptcy forecasting, but given the recent development of these techniques, they have not yet been compared with one another in the same study. While LASSO is a variable selection method that selects a set of the most relevant predictive variables, PCA is a dimensionality reduction

method that generates a linear combination of all predictive variables that are given as an input. In conducting my analysis, I apply the following empirical models and techniques:

Discrete Hazard Model:

The Discrete Hazard Model, as applied by Shumway (2001), Chava and Jarrow (2004), Campbell et al. (2008), Tian et al. (2015) and others, is used to model a firm's bankruptcy risk over a given future period of time. It is structured as a logistic regression in order to constrain the bounds of the predicted probability of bankruptcy between zero and one. The Discrete Hazard Model for bankruptcy risk prediction *j* months in advance is expressed in the following equation (Tian et al., 2015):

1) $$P\left(Y_{i,t+j} = 1 \middle| Y_{i,t+j-1} = 0, X_{i,t}\right) = \frac{e^{\beta_{0,j} + \beta'_j X_{i,t}}}{1 + e^{\beta_{0,j} + \beta'_j X_{i,t}}}$$

$X_{i,t}$ represents the vector of time-varying, firm-specific independent variables observable at time *t* for each firm *i* in my dataset. In the case of my LASSO analysis, my vector of independent variables is the sparse variable set selected by LASSO. In the case of my PCA analysis, my independent variable vector is a set of first principal components of PCA subgroups, or the first K principal components of my entire variable set, where I vary K from 1 to 5. My entire input vector of independent variables is listed in the Data section; these 25 variables consist of both market variables and financial variables that have been used in prior literature. $Y_{i,t}$ serves as a bankruptcy indicator dependent variable that equals one if firm *i* files for bankruptcy and zero otherwise. Additionally, β is the vector of parameters for each independent variable, and β_0 represents the intercept parameter. In my analysis, I will be applying the Discrete Hazard Model for bankruptcy risk prediction both 1 year (12 months) in advance and 2 years (24 months) in advance; this highlights any potential differences in selected

variables or statistically significant PCA groups as a result of factors such as idiosyncratic risk balancing out over a longer forecasting period.

The Discrete Hazard Model serves as a logistic regression. The popularity of this model stems from its ability to model the probability of a binary event: in this case, whether or not a company will declare bankruptcy. Additionally, one of the key features of the Discrete Hazard Model is that it is time-varying, as its model of bankruptcy risk is lagged *j* months prior to the potential bankruptcy event. As a result, the model takes into account that each firm may change over time.

LASSO:

LASSO is used to select the most relevant variables from a comprehensive variable set. By applying the Discrete Hazard Model in Equation 1 as my bankruptcy risk model, LASSO parameter estimates are obtained by maximizing the log-likelihood function and placing a roughness penalty–called the "l_1 penalty"–on the sum of the absolute value of the explanatory variables' parameters (Tian et al., 2015).[1] This is illustrated in the below equations as used by Tian et al. (2015):

2) $\sum_{i=1}^{n}\left(Y_{i,t+j}(\beta_0 + \beta' X_{i,t}\right) - log(1 + exp(\beta_0 + \beta' X_{i,t}))) \; subject \; to \; \sum_{k=1}^{p}|\beta_k| \leq s$

or equivalently:

3) $\sum_{i=1}^{n}\left(Y_{i,t+j}(\beta_0 + \beta' X_{i,t}\right) - log(1 + exp(\beta_0 + \beta' X_{i,t}))) - \; \lambda \sum_{k=1}^{p}|\beta_k|$

In these above equations, *n* represents the number of firms, and *p* represents the number of predictive variables. The independent and dependent variables are the same as those used in

[1] The MATLAB code that I use to obtain my LASSO parameter estimates is based on ordinary least squares (OLS) regressions as opposed to logistic regressions. Extending the LASSO code to use a logistic regression instead of OLS is left for future research.

the Discrete Hazard Model. λ is derived from s in the constraint $\sum_{k=1}^{p}|\beta_k| \leq s$ in Equation 2, and I am able to control the level of shrinkage applied through the *"l_1 penalty"* by changing the value of the parameter λ (or s); a larger λ (or equivalently, a smaller s) decreases the magnitude of the log-likelihood function, which entails a stricter constraint and consequently results in a more parsimonious set of selected variables. The inclusion of the *"l_1 penalty"* results in LASSO selecting variables by zeroing coefficients that have very little statistical significance, eliminating them from consideration for variable selection. Through the process of relaxing the value of λ, LASSO adds more variables into the predictive regression, and the parameter estimates increase in magnitude (Tian et al., 2015). It is important to note that if λ is set to zero, the objective function simplifies to the typical log-likelihood function, and no selection or shrinkage is done.

As discussed in the Literature Review section, these features of the LASSO technique pose several advantages over other commonly used variable selection techniques. LASSO overcomes multicollinearity, and its process of selecting variables results in a variable selection path that can be used to interpret the relative importance of selected variables. Thus, while an input set of 25 independent variables may seem far too large, LASSO is able to deal with large variable sets efficiently and generate a parsimonious output variable set. LASSO's selection method is a unique way of identifying the most relevant variables, and it is one of the techniques that I use to select my set of bankruptcy predictors.

PCA:

PCA offers a different approach to bankruptcy forecasting that reduces the dimensionality of a variable set. I conduct PCA because of the large number of correlated variables in my set of candidate bankruptcy predictors. As discussed in the Literature Review

section, PCA accomplishes dimensionality reduction by transforming the input data vector into a new vector of uncorrelated linear combinations, which are called principal components. Each principal component is a new variable that is a linear combination of the original variables, with the first principal component retaining as much as possible of the variance present in the dataset. The first principal component is the one that I focus on analyzing most closely given that it retains the closest variance to the original variable set.

In conducting my PCA analysis, I execute two different logistic regression methods. I first execute regressions on the first principal components of five variable subgroups, constructing these subgroups based on the categories of my independent variables: Assets & Liabilities Ratios Group, Earnings & Income Ratios Group, Cash Ratios Group, Sales Ratios Group and Market Variables Group. Each of these subgroups captures a different aspect of financial health or public market health for a given firm. I then derive a principal components matrix and extract the first principal component for each of these subgroups. This allows me to analyze linear combinations that represent each of these five types of financial metrics, while retaining the closest possible variance to that of the variables in the given subgroup. One typical flaw with the PCA method is its lack of interpretability; while LASSO's output is a set of bankruptcy predictors, PCA's output is often a vector of uninterpretable linear combinations. However, this method of analysis yields rather interpretable results; I know exactly which type of financial metric that each variable subgroup represents. Moreover, I am able to conduct t-tests on the principal components corresponding to each of these variable subgroups in order to determine statistical significance, and I can also compare goodness-of-fit metrics for each of my regressions of these subgroups.

While each of these variables subgroups captures the characteristics of their respective financial categories, this method of constructing subgroups may seem subjective. In order to address this issue, I apply an additional PCA method in which I conduct regression analysis on the first five principal components of my entire input variable set. I first derive a principal components coefficient matrix, which shows the weight assigned to each of my 25 variables for each of the five principal components. For a given principal component, I take note of which variables are assigned a weight with magnitude greater than 0.30; this implies that the given variable is weighted heavily for that principal component. By observing which variables are heavily weighted for a given principal component, I am able to interpret the types of variables that the principal component represents most. For example, if a principal component includes heavy weightings for stock volatility, price and excess return variables, I am able to conclude that the principal component places a heavier weight on market variables. However, it is important to note that these weights simply serve as an informal guide, as they are not part of my formal statistical analysis. After this preliminary analysis, I conduct logistic regressions on the first principal component and then subsequently add the next principal component to the regression; this results in five logistic regression models that illustrate the statistical significance of each principal component and the change in goodness-of-fit from its addition to the regression. This aspect of my PCA analysis helps provide additional insight to the findings of my variable subgroup method; while it may lack in interpretability compared to the subgroup method, it serves as an objective approach to PCA that analyzes all independent variables simultaneously.

Prediction Accuracy:

After selecting variables for each of my different financial environments, I compare the prediction accuracy of these feature selection techniques. In order to do so, I plug the set of variables selected by LASSO, as well as the first principal components from each PCA variable subgroup and the set of principal components from my entire variable into the Discrete Hazard Model. I use this time-varying logistic regression model to derive coefficient estimates, t-statistics and standard errors for each of these variables and determine whether they are statistically significant according to this model. Additionally, I calculate evaluation metrics on these sets of selected variables to measure which output is the most accurate. The main metric that I rely on across both LASSO and PCA analysis is Pseudo-R^2, which is a log-likelihood based information measure that is adjusted from the traditional R^2 metric to account for goodness-of-fit of a binary event. Pseudo-R^2 differs from the R^2 metric used in OLS in that it analyzes the goodness-of-fit of a model relative to a constant, which offers insight into the *relative* explanatory power of a given model.

Data

I construct my bankruptcy database with annual datapoints for each firm by merging annually updated COMPUSTAT accounting data with daily CRSP equity data, spanning the 1989-2018 period for all publicly traded companies in the United States. Over my 30-year period of observation period, my bankruptcy database has an entry for each firm-year with market and financial variables, as well as bankruptcy information. I construct a bankruptcy indicator variable for each firm, setting the indicator variable to one when a given firm is deleted from the database as a result of filing for either Chapter 7 (liquidation) or Chapter 11 (reorganization) bankruptcy.

Each firm is assigned a value of zero for all years that it remains in the database, and it maintains its indicator value of zero if it is deleted from the database for non-bankruptcy reasons, such as mergers and acquisitions. My dataset contains 200,489 firm-years and 875 bankruptcies.[2]

My set of candidate bankruptcy predictors consists of 25 of the same financial and market variables used by Tian et al. (2015).[3] As discussed in the empirical framework section, LASSO is able to simplify a large vector of input variables into a sparse variable set solution through its selection method. Because of the advantages of LASSO, such as its ability to overcome multicollinearity, I elect to make my set of candidate bankruptcy predictors as large as possible so as to avoid ruling out any potential variables from proving to be statistically significant. I choose to use the variables from the Tian et al. (2015) paper because their study assesses a comprehensive set of variables used in studies dating back to 1966, and there have been no significant additions to the list of accepted bankruptcy predictors in any published literature since 2015. This exhaustive list of predictive variables is drawn from previous bankruptcy forecasting studies, including Beaver (1966), Altman (1968), Ohlson (1980), Shumway (2001), Chava and Jarrow (2004), Bharath and Shumway (2008), Campbell et al. (2008) and others. The variables are listed below in Table 1:

[2] My dataset, which I obtained from Wharton Research Data Services (WRDS), is missing some firm-year entries over my sample period according to the WRDS support representative with whom I communicated. As a result, it does not include all of the bankruptcies of public companies that occurred from 1989 to 2018.

[3] Tian et al. (2015) assesses 39 candidate bankruptcy predictors in their study. While I initially aimed to analyze the same 39 predictive variables, I elected to remove variables from my dataset that had more than 20% of their values missing.

Table 1: List of Predictive Variables

Variable	Description
APSALE	Accounts payable/sales
CASHAT	Cash and short-term investment/total assets
CASHMTA	Cash and short-term investment/(market equity + total liabilities)
CHAT	Cash/total asssets
(EBIT+DP)/AT	(Earnings before interest and tax + amortization and depreciation)/total assets
EBITAT	Earnings before interest and tax/total assets
EBITSALE	Earnings before interest and tax/sales
EXCESS RETURN	Execess return over the S&P 500 Composite Index
FAT	Total debts/total assets
INVTSALE	Inventories/sales
LTAT	Total liabilities/total assets
LTMTA	Total liabilities/(market equity + total liabilities)
LOG(AT)	log(total assets)
MB	Market-to-book ratio
NIAT	Net income/total assets
NIMTA	Net income/(market equity + total liabilities)
NISALE	Net income/sales
OIADPAT	Operating income/total assets
OIADPSALE	Operating income/sales
PRICE	log(price)
REAT	Retained earnings/total assets
RSIZE	log(market capitalization)
SALEAT	Sales/total assets
SEQAT	Equity/total assets
SIGMA	Stock volatility

These market and financial variables serve as my independent variables in my LASSO regression. I construct these market and financial ratios by combining the relevant components of

the equity data from CRSP with the relevant components of the accounting data from COMPUSTAT for each firm-year. Additionally, I calculate an inflation adjustment factor (with 2018 as the index year) for each year in my dataset and create inflation-adjusted versions of each of my original variables to ensure that they are in real terms.

As discussed in the empirical analysis section, I divide these 25 candidate bankruptcy predictors into five variable subgroups for my PCA analysis. I specifically analyze the first principal component of each of these subgroups; each of these principal components represents a linear combination of the corresponding set of variables. Given that each of these variable sets relates to a given financial metric category, I am able to interpret the results of my PCA analysis. The variable subgroups are listed below in Table 2:

Table 2: PCA Variable Subgroups

Assets & Liabilities Ratios Group	
FAT	Total debts/total assets
LTAT	Total liabilities/total assets
LTMTA	Total liabilities/(market equity + total liabilities)
LOG(AT)	log(total assets)
SEQAT	Equity/total assets
Earnings & Income Ratios Group	
(EBIT+DP)/AT	(Earnings before interest and tax + amortization and depreciation)/total assets
EBITAT	Earnings before interest and tax/total assets
EBITSALE	Earnings before interest and tax/sales
NIAT	Net income/total assets
NIMTA	Net income/(market equity + total liabilities)
NISALE	Net income/sales
OIADPAT	Operating income/total assets
OIADPSALE	Operating income/sales
REAT	Retained earnings/total assets

Cash Ratios Group	
CASHAT	Cash and short-term investment/total assets
CASHMTA	Cash and short-term investment/(market equity + total liabilities)
CHAT	Cash/total asssets
Sales Ratios Group	
APSALE	Accounts payable/sales
SALEAT	Sales/total assets
INVTSALE	Inventories/sales
Market Variables Group	
EXCESS RETURN	Execess return over the S&P 500 Composite Index
MB	Market-to-book ratio
PRICE	log(price)
RSIZE	log(market capitalization)
SIGMA	Stock volatility

The method that I use to construct these groups is to first place any variable with a pure assets and liabilities ratio into the Assets & Liabilities Ratios Group. While the Earnings & Income Group includes variables such as EBITSALE and NIAT that have denominator components that may fit into other groups, these denominators normalize the earnings metrics by ensuring that the effect of differences in sizes of each company is not taken into account. The Cash Ratios Group, Sales Ratios Group and Market Variables Group are constructed based on which remaining variables fit into each respective category. While I am methodical and thoughtful in choosing each of these groups, I also conduct a second PCA analysis method in which I analyze the first five principal components of my entire independent variable set. This method is an objective form of analysis that also retains some interpretability by assessing the weights of the variables in each principal component.

Results

As previously discussed, I conduct my analysis with the objectives of (1) comparing the prediction accuracy of LASSO and PCA and (2) identifying sets of predictive variables correlated to bankruptcies before, during and after the global financial crisis to highlight any potential differences between these sets of predictors. I therefore first present findings from my entire 30-year sample and then discuss specific time subsamples to consider my question regarding the financial crisis.

LASSO vs. PCA

I first focus on the forecast of one-year-ahead bankruptcy to compare the predictive accuracy of LASSO regression analysis with that of PCA. A forecast horizon of one year is the most commonly used horizon in existing literature. I also conduct my analysis with a forecast of two-year-ahead bankruptcy in order to assess any changes in feature selection over forecast horizons and see if there are any changes in relative predictive accuracy of the two models. My results show convincingly that, in addition to being more interpretable, LASSO outperforms PCA in terms of prediction accuracy.

My LASSO results for the one-year forecast horizon over the span of my 30-year dataset are shown below in Table 3. Note that although the intercept row is not shown in this table nor any results table, I do include an intercept in all of my LASSO and PCA regression models.

Table 3: LASSO Regression Results (1989-2018)

Logit Regression	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6*	Model 7	Model 8
Net Income/Total Assets	-3.045	-2.639	-2.332	-1.973	-1.498	-1.524	-1.461	-2.443
[T-Statistic]	-[8.154]	-[6.497]	-[5.513]	-[4.423]	-[2.770]	-[2.809]	-[2.629]	-[3.431]
Excess Return		0.005	0.005	0.005	0.005	0.005	0.006	0.006
[T-Statistic]		[3.942]	[3.549]	[3.714]	[3.863]	[3.886]	[4.122]	[3.964]
Stock Volatility			0.028	0.028	0.029	0.029	0.029	0.028
[T-Statistic]			[2.916]	[2.937]	[3.092]	[3.077]	[3.019]	[2.886]
Total Liabilities/Total Assets				1.804	1.954	1.911	1.609	1.593
[T-Statistic]				[2.907]	[3.154]	[3.076]	[2.560]	[2.547]
Earnings Before Interest and Tax/Sales					-0.065	-0.050	-0.053	-0.081
[T-Statistic]					-[1.879]	-[1.325]	-[1.408]	-[2.004]
Inventories/Sales						0.834	0.804	0.874
[T-Statistic]						[1.434]	[1.385]	[1.503]
Market-to-Book Ratio							-0.102	-0.104
[T-Statistic]							-[1.608]	-[1.583]
Earnings Before Interest and Tax/Total Assets								1.819
[T-Statistic]								[1.798]
Pseudo-R^2	*0.0679*	*0.0887*	*0.1010*	*0.1146*	*0.1197*	*0.1227*	*0.1301*	*0.1351*

* Indicates that model corresponds to minimum cross-validated mean-squared error

This table illustrates the order in which LASSO selects its first 8 variables, as well as each variable's estimated coefficients and t-statistics across models. As discussed in the Empirical Framework section, LASSO's process of selecting variables results in a variable selection path that can be used to interpret the relative importance of selected variables; the first several variables that LASSO selects hold greater relative importance than the last several variables selected. Although I show model 1 through model 8 to portray the order in which the variables enter the model, I use LASSO's cross-validation function in MATLAB to confirm that only the first 6 variables correspond to the minimum cross-validated mean-squared error (MSE). Thus, model 6 (boxed in red in Table 3) is considered the optimal model.[4]

The variables selected by LASSO offer insight into some of the issues discussed in prior literature. For example, Shumway (2001) and Campbell et al. (2008) advocate for the inclusion of market variables in bankruptcy forecasts; they argue that stock price is forward looking in that

[4] Cross-validation is a technique used to protect against overfitting in a predictive model. The cross-validation function in MATLAB estimates MSE and finds the model that corresponds to the minimum MSE. In doing so, it creates a fixed number of folds (I use 10-fold cross-validation) of the data, runs analysis on each fold and then averages the overall error estimate.

it includes all available information, as well as the fact that a firm's market value of assets is a more accurate measure than its book value of assets. Table 3 shows that Excess Return and stock return volatility (SIGMA) are selected by LASSO as the second and third variables respectively. Additionally, market-to-book ratio (MB) is selected as the seventh variable and incorporated in model 7, although this is just outside of the optimal model. Tian et al. (2015) also find Excess Return and SIGMA are selected by LASSO in their model.

Additionally, Shumway (2001) argues that the book value of equity should be used in constructing the net income to total assets ratio (NIAT) and the total liabilities to total assets ratio (LTAT), while Campbell et al. (2008) advocates for the incorporation of the market value of equity in these predictive variables. My LASSO results side with Shumway (2001), as NIAT is selected as the first variable, and LTAT is selected fourth. This also goes against the LASSO results found by the Tian et al. (2015), as their model selects both market equity ratios—net income to market value of total assets (NIMTA) and total liabilities to market value of total assets (LTMTA). However, it is certainly possible that LASSO found the predictive power of NIAT and LTAT to only slightly exceed that of NIMTA and LTMTA; once LASSO incorporates a variable such as NIAT into the model, it will not select a highly correlated variable such as NIMTA because that would add very limited information.

As discussed in Empirical Framework section, my PCA analysis consists of two components: a variable subgroup regression and a multi-principal component regression on the entire variable set. My PCA variable subgroup results for the one-year forecast horizon over my 30-year sample are shown below in Table 4:[5]

[5] When looking at model 1 through model 5, it is important to note that positive or negative signs are indeterminate for univariate regressions in the PCA method.

Table 4: PCA Variable Subgroup Regression Results (1989-2018)

Logit Regression	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Assets & Liabilities Ratios PCA Group	-0.279					-0.230
[T-Statistic]	-[2.797]					-[2.287]
Earnings & Income Ratios PCA Group		-0.204				-0.161
[T-Statistic]		-[7.044]				-[3.696]
Cash Ratios PCA Group			0.177			-0.043
[T-Statistic]			[2.017]			-[0.419]
Sales Ratios PCA Group				0.261		0.029
[T-Statistic]				[2.336]		[0.245]
Market Variables PCA Group					0.498	0.311
[T-Statistic]					[5.020]	[2.741]
Pseudo-R ^{2}	*0.0124*	*0.0497*	*0.0057*	*0.0075*	*0.0372*	*0.0743*
AIC	*0.0042*	*0.0041*	*0.0043*	*0.0042*	*0.0041*	*0.0040*
BIC	*0.0042*	*0.0041*	*0.0043*	*0.0043*	*0.0041*	*0.0041*

As discussed in prior sections, I split my set of 25 variables into the five variable subgroups listed in Table 4. I then extract the first principal component of each of these subgroups and conduct a series of regressions. Table 4 shows that the Assets & Liabilities Ratios Group, Earnings & Income Group and Market Variables Group are statistically significant at the 95% confidence level for both their individual regressions and the model 6 regression, which incorporates all five subgroups. These results are consistent with my LASSO results, which show that a set of variables related to assets and liabilities ratios, earnings and income ratios and market ratios possesses the most predictive power. In terms of prediction accuracy, the Pseudo-R^2 of each LASSO model is consistently significantly higher than the Pseudo-R^2 of each of the PCA models. Most importantly, LASSO's model 6, which is its optimal model based on MSE cross-validation, has a Pseudo-R^2 value of 0.1227. This is 65% higher than the Pseudo-R^2 value of the PCA subgroup model 6, which boxed in red to signify that it is the optimal model in this table based on its AIC value.[6]

[6] Akaike Information Criterion (AIC) and Schwarz's Bayesian Information Criterion (BIC) are the two most widely used model selection criteria. While performance measures like Pseudo-R^2 will generally pick the largest model,

The second component of my PCA analysis assesses the first five principal components of the entire variable set. These results are shown below in Table 5:

Table 5: Entire Variable Set PCA Regression Results (1989-2018)

Logit Regression	Model 1	Model 2	Model 3	Model 4	Model 5
First PC	0.201	0.193	0.212	0.192	0.192
[T-Statistic]	[7.051]	[6.923]	[6.813]	[5.410]	[5.373]
Second PC		-0.183	-0.194	-0.190	-0.188
[T-Statistic]		-[2.650]	-[2.662]	-[2.548]	-[2.510]
Third PC			-0.118	-0.091	-0.103
[T-Statistic]			-[1.443]	-[1.078]	-[1.134]
Fourth PC				0.137	0.145
[T-Statistic]				[1.634]	[1.654]
Fifth PC					-0.039
[T-Statistic]					-[0.356]
Pseudo-R^2	*0.0517*	*0.0624*	*0.0659*	*0.0705*	*0.0707*
AIC	*0.0042*	*0.0041*	*0.0043*	*0.0043*	*0.0042*
BIC	*0.0042*	*0.0041*	*0.0043*	*0.0044*	*0.0043*

These regression results reinforce the results from both the LASSO regression analysis and the PCA subgroup analysis. The first principal component and second principal component of my variable set are the only two principal components (PCs) that are statistically significant across each model. The first PC contains five overweight variables—variables that have a coefficient with a weight greater than 0.30— that all fall into the Earnings & Income Group bucket. The second PC also contains five overweight variables from the Assets & Liabilities Group, Earnings & Income Group and Market Variables Group. Thus, these three variable subgroup categories illustrate predictive power across both PCA analysis components and the LASSO analysis. Additionally, similar to the PCA subgroup analysis, the Pseudo-R^2 values of each of the five models in Table 5 are consistently significantly lower than the values in each LASSO model. LASSO's model 6 Pseudo-R^2 value is 97% greater than that of model 2 in Table

these criteria incorporate a trade-off between goodness-of-fit and increased estimation error due to forecasting extra parameters. The model that *minimizes* these selection criteria is considered to be the optimal model.

5, which is boxed in red to signify that it is optimal model in this table based on both AIC and BIC model selection criteria.

From the results of my LASSO and PCA regression analysis, I am able to conclude that LASSO outperforms PCA in terms of prediction accuracy. As discussed, LASSO's optimal model (model 6 in Table 3) has a significantly greater Pseudo-R^2 value than both the PCA subgroup optimal model (model 6 in Table 4) and the PCA entire independent variable set optimal model (model 2 in Table 5). Moreover, by utilizing LASSO, I am able to see exactly which variables are included in a given model, and I understand how each of these variables are constructed. While I have a sense for the types of variables in each PCA subgroup and the heavy weighted variables in each principal component of the entire variable set, PCA's form of analysis is not nearly as interpretable.

These conclusions hold true when forecasting with a forecast horizon of two years, as the Pseudo-R^2 values for each of LASSO's models exceed the values in the PCA models. Additionally, the variable selection results in both LASSO and PCA do not change significantly when the forecast horizon is extended from one year to two years. While LASSO selects a combination of market variables, asset and liabilities variables and earnings and income variables, the PCA regressions reinforce the predictive power of each of these three main variable subgroups. These two-year-ahead forecast results are shown in the Appendix.

Variable Selection Before, During and After the Global Financial Crisis

The second significant component of my research is analyzing the sets of predictive variables for subsamples corresponding to before, during and after the global financial crisis. I construct these subsamples as follows: my first subsample is 1989-2006, which captures the time

period prior to the financial crisis. My second subsample is 2007-2012, which consists of both the time during the crisis (2007-2009) and the three years that followed. Given the fact that no literature on forecasting corporate bankruptcy during the financial crisis exists to provide guidance on the exact years to consider in each subsample, I made the decision on my own to extend the subsample to include the three years following the crisis. I include these three additional years in this subsample in order to capture any effects of the crisis on bankruptcies in the immediate aftermath. Specifically, as Tian et al. (2015) show in their paper, bankruptcy filings exhibit strong countercyclical patterns with peaks following business recessions. Thus, the effects of the crisis may have impacted bankruptcies in the years following the recession. Finally, my third subsample is 2013-2018, which captures the time period following the financial crisis and the subsequent three years. Similar to my analysis of the prediction accuracy of LASSO compared to PCA, I first focus on the forecast of one-year-ahead bankruptcy. Table 6, which is shown below, illustrates the breakdown of each of my subsamples for the one-year forecast horizon.

Table 6: One-year-ahead Bankruptcy Forecast Subsample Breakdown

Subsample	*Period*	*Number of Observations*	*Number of Bankruptcy Cases*
Subsample 1	1989-2006	91,362	16
Subsample 2	2007-2012	25,986	11
Subsample 3	2013-2018	23,718	5
Total	**1989-2018**	**141,066**	**32**

Table 7, Table 8 and Table 9 show condensed versions of my LASSO and PCA results for each of my subsamples. Specifically, each table presents the LASSO regression model that corresponds to the minimum cross-validated MSE (the optimal LASSO model), as well as the PCA model 6 that incorporates all five subgroup principal components into its regression. I choose to focus on the PCA variable subgroup technique for this part of my analysis instead of

the principal components of my entire variable because the variable subgroup method retains a higher level of interpretability. Moreover, the regressions on the principal components of my entire variable set do not provide much additional information with regard to variable selection in these three subsamples. For each set of subsample results, I highlight every market variable that is selected by LASSO with the color red in order to show this change in variable types across subsamples.

The LASSO and PCA results for subsample 1 (1989-2006) are shown below in Table 7:

Table 7: Pre-Crisis Period (1989-2006)

Logit Regression			
Net Income/Total Assets	-0.281	Assets & Liabilities Ratios PCA Group	-0.346
[T-Statistic]	-[0.282]	[T-Statistic]	-[2.385]
Total Liabilities/Total Assets	3.677	Earnings & Income Ratios PCA Group	-0.167
[T-Statistic]	[3.564]	[T-Statistic]	-[2.906]
Earnings Before Interest and Tax/Sales	-0.025	Cash Ratios PCA Group	-0.271
[T-Statistic]	-[0.308]	[T-Statistic]	-[1.215]
Log(Total Assets)	-0.478	Sales Ratios PCA Group	0.119
[T-Statistic]	-[2.257]	[T-Statistic]	[0.896]
Excess Return	0.004	Market Variables PCA Group	0.229
[T-Statistic]	[1.720]	[T-Statistic]	[1.325]
Total Debts/Total Assets	0.018	*Pseudo-R* ^{2}	*0.0958*
[T-Statistic]	[1.583]		
Inventories/Sales	0.762		
[T-Statistic]	[0.817]		
Accounts Payable/Sales	0.087		
[T-Statistic]	[0.959]		
Pseudo-R ^{2}	*0.1290*		

This table shows that only one market variable—Excess Return—is selected by LASSO for its minimum MSE optimized model, while seven accounting variables are selected. Additionally, Excess Return is the fifth variable selected by LASSO, meaning that the four variables that precede it hold a higher level of importance in predicting bankruptcy. Moreover, the t-statistic of Excess Return is less than 1.96 in absolute value, meaning that this variable is not statistically significant at the 95% confidence level. The PCA results show that the first principal components of the Assets & Liabilities Ratios Group and the Earnings & Income

Ratios Group are the only two statistically significant variables in this regression. Thus, market variables do not hold statistical significance in forecasting bankruptcy in this first pre-crisis subsample according to PCA.

Table 8 shows my LASSO and PCA results for subsample 2 (2007-2012), which captures the entirety of the global financial crisis, as well as the ensuing three years after the crisis.

Table 8: Crisis Period (2007-2012)

Logit Regression			
Net Income/Total Assets	-1.951	Assets & Liabilities Ratios PCA Group	-0.013
[T-Statistic]	-[2.266]	[T-Statistic]	-[0.067]
Excess Return	0.006	Earnings & Income Ratios PCA Group	-0.188
[T-Statistic]	[2.425]	[T-Statistic]	-[2.526]
Stock Volatility	0.040	Cash Ratios PCA Group	-0.207
[T-Statistic]	[2.517]	[T-Statistic]	-[1.043]
Operating Income/Sales	-0.035	Sales Ratios PCA Group	-0.086
[T-Statistic]	-[0.600]	[T-Statistic]	-[0.364]
Pseudo-R^2	*0.1548*	Market Variables PCA Group	0.430
		[T-Statistic]	[2.175]
		Pseudo-R^2	*0.0961*

The optimal model of the LASSO regression for the 2007-2012 crisis period contains only four variables—two market variables and two accounting variables. LASSO selects Excess Return and SIGMA as the second and third variables in the model, respectively. Thus, besides adding SIGMA to its variable set for this subsample, LASSO also places a higher value of importance on Excess Return than it does in the pre-crisis period, as Excess Return is selected as the second variable instead of the fifth variable. Furthermore, both are statistically significant at the 95% confidence level, which is not the case for Excess Return in the pre-crisis period. The log of market capitalization (RSIZE) is another market variable that is selected fifth and incorporated in model 5, although this is just outside of the optimal LASSO model. The PCA results in Table 8 reinforce this evident shift from accounting variables to market variables shown by LASSO. While the first principal component of the Assets & Liabilities Group loses statistical significance in subsample 2, the predictive power of the Market Variables Group's first

principal component increases, as this variable is statistically significant at the 95% confidence level. Moreover, while PCA model 6 is the optimal model in the subsample 1 based on AIC, PCA model 5 has the lowest AIC and BIC values in subsample 2. Thus, the model that uses only the Market Variables Group becomes the optimal PCA model in the crisis period. This further illustrates the greater predictive power of market variables in subsample 2 as compared to subsample 1.

My LASSO and PCA results for the 2013-2018 post-crisis period continue to show this emphasis on market variables as bankruptcy predictors. These results are shown below in Table 9:

Table 9: Post-Crisis Period (2013-2018)

Logit Regression			
Excess Return	0.009	Assets & Liabilities Ratios PCA Group	-0.089
[T-Statistic]	[2.937]	[T-Statistic]	-[0.299]
Net Income/Total Assets	-5.858	Earnings & Income Ratios PCA Group	0.029
[T-Statistic]	-[4.478]	[T-Statistic]	[0.185]
Stock Volatility	0.066	Cash Ratios PCA Group	0.287
[T-Statistic]	[2.865]	[T-Statistic]	[1.120]
Cash/Total Assets	8.233	Sales Ratios PCA Group	0.027
[T-Statistic]	[3.520]	[T-Statistic]	[0.065]
Earnings Before Interest and Tax/Total Assets	10.167	Market Variables PCA Group	0.493
[T-Statistic]	[3.923]	[T-Statistic]	[1.969]
Inventories/Sales	2.573	*Pseudo-R^2*	*0.1043*
[T-Statistic]	[2.063]		
Pseudo-R^2	*0.4764*		

In this final subsample, LASSO selects two market variables and four accounting variables for its optimal model, which contains six total variables. Excess Return is now selected first, indicating that it holds the greatest level of predictive power in this subsample.[7] This makes this LASSO subsample the only one that does not select NIAT as the first variable. Additionally, similar to the crisis period regression, SIGMA is the third variable selected. Both of these

[7] Note the high Pseudo-R^2 value for the LASSO regression; 0.4764 is a significantly higher value for this metric than any of the values from the other regressions. The main cause for this discrepancy is that the third subsample for the one-year forecast horizon contains a limited number of bankruptcy cases compared to the other two subsamples. Thus, this subsample regression is more susceptible to large swings in its Pseudo-R^2 metric.

variables maintain their statistical significance at the 95% confidence level according to each of their t-statistics. The PCA results continue to reinforce the LASSO regression results, as the Market Variables Group is the only variable subgroup that is statistically significant at the 95% confidence level for the post-crisis period. Additionally, PCA model 5, which regresses only the Market Variables Group, is also the optimal PCA model in the post-crisis period based on both AIC and BIC model selection criteria. This further emphasizes the importance of market variables as bankruptcy predictors relative to all types of accounting variables.

Discussion of Results

For one-year-ahead forecasts, the changes in variable sets across these three subsamples illustrates a shift in predictive power from accounting variables before the global financial crisis to market variables during and after the crisis. This trend can be explained by looking at the dynamic of public companies across each of the three subsamples. With the exception of the dotcom bubble in the early 2000s, the United States economy experienced a bull market for the entirety of the pre-crisis period (1989-2006). This means that for the majority of this period, companies were able to raise capital through the credit markets with relative ease, allowing them to utilize the credit markets to fund their operations and continue to grow their businesses. Prior to the financial crisis, banks lent directly to companies and held this risk on their own balance sheets. The inability of these banks to assess and evaluate risk on their balance sheets ultimately served as one of the causes of the global financial crisis.

The aftermath of the financial crisis prompted calls for significant financial reform to ensure that a similar crisis will never happen again. One of the most important components of this reform was an increase in the capital requirements of banks in order to improve the

resiliency of the financial markets in stressful economic environments; a capital requirement is the amount of liquid capital a bank or other financial institution is required to keep on hand to support the nature of their balance sheet. In early 2009, former Secretary of the Treasury Timothy Geithner initiated the stress test exercise to determine how much capital banks needed to remain viable financial institutions under varying economic environments. This eventually led to the passage of the Dodd-Frank Wall Street Reform and Consumer Protection Act of 2010, which mandates that US regulators increase capital requirements for large banks. Furthermore, Basel III (or the Third Basel Accord) is a global regulatory framework on bank capital adequacy and market liquidity risk that was agreed upon by the members of the Basel Committee on Banking Supervision in 2011. Basel III international capital standards require all banks to hold more capital based on risk-based capital metrics and leverage requirements including annual liquidity stress tests. While Basel III is voluntary framework, the U.S. Federal Reserve implemented the accord's capital rules in the US in July 2013.

These banking reforms related to capital requirements led banks to limit their overall capital deployment and shift to a more risk-averse lending appetite. As a result, lending standards for public companies became more stringent, and "bad" companies with lower quality credit found it harder to raise capital through the traditional credit markets. I believe that the lack of accessibility to the credit markets that these companies experienced may have materialized in the public equity markets prior to manifesting in the financial statements of these companies because investors took note of this fundamental shift in lending standards. The financial crisis redefined creditworthiness for corporations, and this led to "bad" companies with lower creditworthiness experiencing steep declines in their stock prices and heightened volatility relative to other companies. While this is one possible explanation for this shift from accounting variables to

market variables, there is no way to completely distinguish the reason for this change. The financial regulatory reform related to capital requirements did not occur in a vacuum, as significant macroeconomic changes, such as sharp declines in market sentiment and investor confidence, also took place in the immediate aftermath of the crisis. Thus, determining the exact cause of this shift in predictive variables would be difficult to accomplish.

The post-crisis period (2013-2018) captures another bull market, but by this time the financial services industry had lost its increased sense of caution regarding lending to corporations. US equities performed well over this period, and the only companies that did not perform well tended to have either unviable business models or significant headwinds. The companies that declared bankruptcy may have failed not as a direct result of an inability to raise capital but because they had failed to innovate and adapt to an evolving world. As a result, one possible reason that market variables, such as Excess Return and SIGMA, serve as two of the most statistically significant bankruptcy predictors for this subsample period is that companies that declared bankruptcy were far from the rest of the pack of growing US companies with regard to innovation. While the struggles of these companies may have reflected most prominently in the public equity markets, this theory is also difficult to definitively prove.

This shift from accounting variables to market variables is not as evident in two-year-ahead bankruptcy forecasts. Feature selection in these forecasts is much more consistent across subsamples, and there is no significant increase in the relative predictive power of market variables during and after the financial crisis for this forecast horizon. This may be because longer forecast horizons generally tend to weight accounting variables more heavily than market variables given that idiosyncratic shocks related to stock price balance out over time, which is one of the conclusions made by Tian et al. (2015).

Conclusion

I draw two main conclusions related to my interests of prediction accuracy and variable selection. In comparing my two forecasting methods, LASSO and PCA, I conclude that LASSO outperforms PCA in terms of prediction accuracy and offers more interpretable results. Additionally, with regard to variable selection for the pre-crisis, crisis and post-crisis time periods, I conclude that the set of bankruptcy predictors shifts from accounting variables before the financial crisis to market variables during and after the financial crisis for one-year-ahead forecasts. This shift in predictive variables offers insight into the effects of stricter lending standards that occurred in the aftermath of the financial crisis. One possible explanation for these findings is that financial crisis banking reforms related to capital requirements, such as Dodd-Frank and Basel III, led to stricter lending standards that limited capital access for companies with lower creditworthiness. The struggles of these companies that eventually filed for bankruptcy may have materialized first in the equity markets during the crisis and post-crisis periods, leading to a relative increase in predictive power for market variables during these periods. However, this is just one possible explanation, as this shift in predictive variables could be due to a variety of other factors such as macroeconomic changes in the aftermath of the crisis; it is not possible to distinguish the exact cause of these findings using my dataset.

As the financial markets have experienced heightened turmoil amid the coronavirus crisis, corporate bankruptcy forecasting has become increasingly relevant. Transportation, hospitality and retail companies are among many of the industries that face great uncertainty regarding the prospect of bankruptcy. Many of these companies have lost significant revenue as a result of the pandemic, and quite a few of them—namely airlines—will likely need the government to bail them out of bankruptcy. The findings of my paper may shed light on this issue, as the lack of viability

of these struggling companies has manifested in the equity markets, with heightened volatility and falling stock prices serving as a leading indicator for the companies that are most at-risk of bankruptcy. One significant question that remains is whether this shift in predictive power from accounting variables before the global financial crisis to market variables during and after the crisis has held true through the coronavirus crisis. The effects of the coronavirus crisis environment on bankruptcy predictors would be an interesting topic to investigate given the fundamental differences between this crisis and the global financial crisis.

Appendix

Appendix Figure 1: LASSO Regression Results (1989-2018), *h=2*

Logit Regression	Model 1	Model 2	Model 3*	Model 4	Model 5, 6, 7 [1]	Model 8
Net Income/Total Assets	-2.867	-2.440	-2.100	-2.121	-1.702	-1.768
[T-Statistic]	-[16.976]	-[13.220]	-[10.782]	-[10.664]	-[7.752]	-[7.798]
Stock Volatility		0.031	0.032	0.032	0.028	0.027
[T-Statistic]		[7.573]	[7.685]	[7.645]	[6.627]	[6.378]
Equity/Total Assets			-2.119	-2.398	-1.801	-0.894
[T-Statistic]			-[6.957]	-[3.887]	-[2.882]	-[1.306]
Total Liabilities/Total Assets				-0.292	0.330	0.721
[T-Statistic]				-[0.522]	[0.578]	[1.221]
Log(Market Capitalization)					-0.131	-0.166
[T-Statistic]					-[4.730]	-[5.622]
Cash/(Market Equity + Total Liabilities)						-1.526
[T-Statistic]						-[2.763]
Sales/Total Assets						0.115
[T-Statistic]						[0.995]
Excess Return						0.002
[T-Statistic]						[2.981]
Pseudo-R^2	*0.0630*	*0.0802*	*0.0982*	*0.0983*	*0.1063*	*0.1139*

(1) LASSO selects 5 variables for degrees of freedom values of 5, 6 and 7

* Indicates that model corresponds to minimum cross-validated mean-squared error

Appendix Figure 2: PCA Variable Subgroup Regression Results (1989-2018), *h=2*

Logit Regression	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Assets & Liabilities Ratios PCA Group	-0.276					-0.209
[T-Statistic]	-[6.643]					-[4.774]
Earnings & Income Ratios PCA Group		-0.181				-0.184
[T-Statistic]		-[13.973]				-[10.111]
Cash Ratios PCA Group			-0.031			-0.244
[T-Statistic]			-[0.614]			-[4.464]
Sales Ratios PCA Group				-0.045		-0.225
[T-Statistic]				-[0.697]		-[3.763]
Market Variables PCA Group					0.495	0.344
[T-Statistic]					[11.868]	[7.241]
Pseudo-R^2	*0.0152*	*0.0431*	*0.0001*	*0.0002*	*0.0453*	*0.0918*
AIC	*0.0217*	*0.0211*	*0.0220*	*0.0220*	*0.0210*	*0.0200*
BIC	*0.0217*	*0.0211*	*0.0220*	*0.0220*	*0.0210*	*0.0202*

Appendix Figure 3: Entire Variable Set PCA Regression Results (1989-2018), *h=2*

Logit Regression	Model 1	Model 2	Model 3	Model 4	Model 5
First PC	0.174	0.170	0.203	0.132	0.132
[T-Statistic]	[13.515]	[13.467]	[14.123]	[6.364]	[6.220]
Second PC		-0.169	-0.198	-0.182	-0.181
[T-Statistic]		-[5.816]	-[5.972]	-[4.979]	-[4.948]
Third PC			-0.288	-0.273	-0.273
[T-Statistic]			-[7.593]	-[6.504]	-[6.450]
Fourth PC				0.301	0.300
[T-Statistic]				[7.075]	[6.845]
Fifth PC					0.005
[T-Statistic]					[0.087]
Pseudo-R^2	*0.0420*	*0.0533*	*0.0751*	*0.0961*	*0.0961*
AIC	*0.0211*	*0.0208*	*0.0204*	*0.0199*	*0.0200*
BIC	*0.0211*	*0.0209*	*0.0205*	*0.0200*	*0.0201*

Appendix Figure 4: Two-year-ahead Bankruptcy Forecast Subsample Breakdown

Subsample	*Period*	*Number of Observations*	*Number of Bankruptcy Cases*
Subsample 1	1989-2006	78,950	116
Subsample 2	2007-2012	24,691	41
Subsample 3	2013-2018	22,362	27
Total	**1989-2018**	**126,003**	**184**

Appendix Figure 5: Pre-Crisis Period (1989-2006), *h=2*

Logit Regression			
Net Income/Total Assets	-1.544	Assets & Liabilities Ratios PCA Group	-0.128
[T-Statistic]	-[5.350]	[T-Statistic]	-[2.286]
Equity/Total Assets	-2.437	Earnings & Income Ratios PCA Group	-0.203
[T-Statistic]	-[5.692]	[T-Statistic]	-[9.808]
Stock Volatility	0.023	Cash Ratios PCA Group	-0.262
[T-Statistic]	[4.146]	[T-Statistic]	-[3.398]
Log(Total Assets)	-0.372	Sales Ratios PCA Group	-0.230
[T-Statistic]	-[4.881]	[T-Statistic]	-[3.145]
Excess Return	0.002	Market Variables PCA Group	0.272
[T-Statistic]	[2.867]	[T-Statistic]	[4.342]
Pseudo-R^2	*0.1095*	*Pseudo-R^2*	*0.0873*

Appendix Figure 6: Crisis Period (2007-2012), *h=2*

Logit Regression

Net Income/Total Assets	-2.125
[T-Statistic]	-[5.353]
Stock Volatility	0.035
[T-Statistic]	[4.207]
Total Liabilities/Total Assets	2.325
[T-Statistic]	[3.897]
Pseudo-R ^{2}	*0.1146*

Assets & Liabilities Ratios PCA Group	-0.314
[T-Statistic]	-[3.064]
Earnings & Income Ratios PCA Group	-0.120
[T-Statistic]	-[2.761]
Cash Ratios PCA Group	-0.136
[T-Statistic]	-[1.143]
Sales Ratios PCA Group	-0.115
[T-Statistic]	-[0.998]
Market Variables PCA Group	0.413
[T-Statistic]	[3.951]
Pseudo-R ^{2}	*0.0929*

Appendix Figure 7: Post-Crisis Period (2013-2018), *h=2*

Logit Regression

Stock Volatility	0.042
[T-Statistic]	[3.961]
Net Income/Total Assets	-1.120
[T-Statistic]	-[1.920]
Equity/Total Assets	-1.793
[T-Statistic]	-[2.582]
Sales/Total Assets	0.386
[T-Statistic]	[1.533]
Market-to-Book Ratio	-0.161
[T-Statistic]	-[1.551]
Log(Price)	-0.093
[T-Statistic]	-[0.983]
Accounts Payable/Sales	-0.220
[T-Statistic]	-[1.191]
Pseudo-R ^{2}	*0.4764*

Assets & Liabilities Ratios PCA Group	-0.336
[T-Statistic]	-[2.594]
Earnings & Income Ratios PCA Group	-0.173
[T-Statistic]	-[3.092]
Cash Ratios PCA Group	-0.232
[T-Statistic]	-[1.623]
Sales Ratios PCA Group	-0.447
[T-Statistic]	-[2.585]
Market Variables PCA Group	0.430
[T-Statistic]	[4.001]
Pseudo-R ^{2}	*0.1423*

Reference List

Altman, E.I., 1968,

Financial ratios, discriminant analysis and prediction of corporate bankruptcy, Journal of Finance, 23, pp. 589-610

Beaver, W.H., 1966,

Financial ratios as predictors of failure, Journal of Accounting Research, 4, pp. 71-111

Bharath, S., Shumway, T., 2008,

Forecasting default with the merton distance to default model, Review of Financial Studies, 21, pp. 1139-1369

Campbell, J., Hilscher, J., Szilagyi, J., 2008,

In search of distress risk, Journal of Finance, 63, pp. 2899-2939

Chava, S., Jarrow, R.A., 2004,

Bankruptcy prediction with industry effects, Review of Finance, 8, pp. 537-569

Ohlson, J.S., 1980,

Financial ratios and the probabilistic prediction of bankruptcy, Journal of Accounting Research, 19, pp. 109-131

Shumway, T., 2001,

Forecasting bankruptcy more accurately: a simple hazard model, Journal of Business, 74, pp. 101-124

Tian, S., Yu, Y., Guo, H., 2015,

Variable selection and corporate bankruptcy forecasts, Journal of Banking & Finance, 52, pp. 89-100

Tsai, C.-F., 2009,

Feature selection in bankruptcy prediction, Knowledge-Based Systems, 22 (2), pp. 120-127