Predicting Firm Bankruptcy

Adam Vogel

Department of Economics

University of California, Davis

June 6, 2021

## Abstract

The problem of assessing the solvency and health of firms presents challenges for both government regulators and businesspeople alike. The goal of this report is to predict whether or not a firm has gone bankrupt based on various observed financial metrics. If firms can be predicted as bankrupt or insolvent while they still exist, then this information might be useful to regulators, other government officials, or even private investors in aiding their ability to ensure solvent financial markets and navigate investment risk. Through the implementation of several machine-learning models (including logistic regression, LASSO regression, regression tree, and random forest models) we are able to categorize firms as being “bankrupt” or “not bankrupt” to a satisfactory degree of accuracy. Although misclassifications are common among all models, the LASSO regression model performs especially well at accomplishing this task, with relatively low rates of false positives compared to the other models.

## Introduction

The primary question of this report is whether we can predict a firm's bankruptcy status through the observation of various metrics from the firm's accounting statements. Being able to accurately and precisely predict the solvency of a firm provides huge benefits to all stakeholders involved, including investors, board members, and market regulators. The ability to predict the probability of bankruptcy for individual firms is of great importance in assessing the financial soundness of individual companies, as well as the overall economy. If policymakers are able to predict firm bankruptcy to a sufficient level of accuracy, it is possible that this may even have implications for economists' understanding of the business cycle.

Variables which may be of interest when assessing the likelihood of firm bankruptcy include ratios like debt ratio percent (Liabilities/Assets) and cash flow rate (Cash flow from Operations/Current Liabilities). There are a total of 94 variables which may be assessed within the dataset and included in our model. In order to maximize out of sample performance while simultaneously reducing the likelihood of overfitting the predicted outcome to our observed data, we prefer simpler models over complex ones when their explanatory power (MSE) is comparable. Later on, we will accomplish this task of finding the simplest yet sufficiently explanatory model by using LASSO and other methods.

As always, there are confounding factors that must be taken into account when attempting to predict probability of firm bankruptcy based on financial data. Therefore, we specify that the goal of this report is not to determine any possible causal link between the explanatory variables and firm bankruptcy, but merely to assess whether it is feasible to attempt to predict firm bankruptcy based on similar observed data. I plan to implement logistic regression, LASSO logistic regression, regression trees, and random forests in order to predict firm bankruptcy. Because we are interested more in prediction rather than any sort of association between variables, I will also apply Principal Component Analysis to the observed data in order to determine if doing so has any effect on the performance of the above models. Thus, we will have models which are input only the raw data, and other models which are input only the principal components (in turn sacrificing some interpretability).

Several other studies have previously attempted to predict the probability of firm bankruptcy based on financial data. In a study by Gepp and Kumar (2015), it was demonstrated that decision trees have good bankruptcy prediction accuracy when trained against cost ratios. This study is relevant to my own question because it shows how similar financial ratios to the ones contained in my dataset can be used in a regression trees to obtain a high level of prediction accuracy. Another study by Geng et al. (2015) of 107 firms listed on the Shanghai and Shenzhen Stock Exchanges identified several predictive signs of firm bankruptcy, including the net profit margin of assets, cash flow per share, and return on total assets (among others). These variables have been identified as suitable for forecasting the financial distress of companies, and therefore they should be a focus of my own analysis later on in the paper. A boxplot of return on assets against bankruptcy status is available in the appendix to demonstrate the possible explanatory power of this particular predictor. Models which contain all possible predictors available to the researcher also perform well when the model is a logistic regression. According to a paper by Ben Jabeur (2017), logistic regression allows integrating a large number of ratios in the model, providing the researcher with the opportunity to consider as many indicators as feasible when predicting bankruptcy. The same paper found that logistic regression produced satisfactory results.

One issue of note is that even a well-developed model in one country may not perform satisfactory in another country, possibly due to the differences in structural economic conditions between countries. According to Svabova et al. (2018), this issue should be mitigated by only running models on data within the same country. My paper follows this guidance, in that it seeks to predict firm bankruptcies for Taiwanese firms only, whose regulation and oversight is managed by the Taiwanese government.

## Data

The dataset to be analyzed contains firm data collected by the Taiwan Economic Journal between 1999 and 2009. The dataset was assembled by Professors Deron Liang and Chih-Fong Tsai, who are researchers at Taiwan's National Central University. This is a cross-sectional dataset, where the cross-sectional units are comprised of individual firms listed on the Taiwan Stock Exchange. There are a total of 6819 firms in the dataset, leaving us with $n = 6819$ observations to work with. This sample size will be adequate for testing the models we wish to employ. The response variable is binary, taking a value 0 for "not bankrupt" and 1 for "bankrupt". Here, bankruptcy is defined in accordance with the business regulations of the Taiwan Stock Exchange and takes the value of 1 if the observed firm filed for bankruptcy within the time frame 1999 to 2009. From summing over values in the "Bankrupt" column, we find that the dataset contains 220 firms which went bankrupt, and 6599 firms which did not go bankrupt. Consequently, the dataset is quite heavily weighted towards observations for which the firm is not bankrupt.

One column of data, "Net Income Flag", due to its zero variance causing problems for some of the models. We are left with a total of 94 predictor variables in the dataset. Most of these variables are ratios of financial metrics, such as "total debt/total net worth" and "cash/total assets". For the sake of out of sample prediction, and because not all 94 of these same variables may be known in other contexts, we attempt to find the simplest model possible which contains only a subset of the predictors, rather than the complete set of 94 variables.

## Model

### Logistic Regression (with PCA)

Because the outcome variable in question is binary (0 for 'not bankrupt', 1 otherwise), we will begin by implementing logistic regression models. Logistic regression models can be run on the variables themselves, a subset of the variables, or on principal component vectors formed from linear combinations of the variables. The latter method will be used initially as well as later on in the paper and will be compared to regression performed on the variables themselves.

To begin my analysis, I perform Principal Component Analysis on the complete set of 94 predictor variables. Principal Component Analysis works by extracting relevant information from complex datasets with many variables. In essence, PCA finds a sequence of basis vectors which are linear combinations of the original basis vector and orthogonal with respect to each other, such that the sequence of vectors summarizes as much information stored within the original variables as possible (Shlens 2014). As such, PCA is applied well in situations when the predictor variables themselves exhibit a high degree of collinearity, as the orthonormal basis vectors it forms are constructed to be as uncorrelated as possible. As our data is composed of financial ratios which themselves are composed of many accounting identities, we observe high levels of multicollinearity between certain variables (see the scatter plot of return on assets before interest and depreciation before tax versus return on assets before interest and depreciation after tax for an example of the multicollinearity caused by accounting identities).

PCA is a method that can remedy this multicollinearity (although it is not a huge issue either way).

After performing PCA on the 94 variables, we obtain their principal components. From a plot of the first two principal components, we can observe how the algorithm performed at separating our data. Figure one depicts the first two principal components formed from the data. Points which represent bankruptcy are red. Due to the low number of bankrupt firms captured in the dataset, it may be difficult to observe that most points which represent bankruptcy are clustered in the shape of a circle around a concentration of points which represent non-bankruptcy. Nonetheless, it would appear for now that PCA succeeded in its goal of summarizing the vast dataset into fewer vectors. We also observe in figure two the cumulative proportion of variance explained by the different principal components. The results indicate that PCA explains around 95% of the variance present in the dataset when between 50 to 60 principal components are included. The practical explanatory power of this PCA, however, should only be ascertained by evaluating its predictive performance in a logistic regression model as compared to the predictive capabilities of the variables themselves.

I begin by following the guidance of Jabeur et al. (2017) and estimate a logistic regression which contains all 94 predictor variables. Here, model (1) takes the form $P_i = \frac{e^{a+\beta_1 x_{1,i}+\cdots+\beta_{94}x_{94,i}}}{1+e^{a+\beta_1 x_{1,i}+\cdots+\beta_{94}x_{94,i}}}$, where $P$ is the probability of bankruptcy for the i-th firm, and $x$ represents the observed values for all 94 variables for the i-th firm. I also cross validate the regression by partitioning the data into $k = 5$ separate folds, in order to obtain an out of sample MSE which can be used to compare the predictive capabilities of models. The results of this initial regression are discussed in the “results” section.

Next, I formulate a similar logistic regression model, but on the complete set of 94 principal components formed by the above PCA process, rather than on the 94 variables themselves. Model (2) takes the form $P_i = \frac{e^{a+\beta_1 v_{1,i}+\cdots+\beta_{94} v_{94,i}}}{1+e^{a+\beta_1 v_{1,i}+\cdots+\beta_{94} v_{94,i}}}$, where $v$ now represents the principal component for the given variable. This model is also cross validated into 5 folds in order to obtain an out of sample MSE.

Ideally, we would like to formulate a model which does not depend on the presence of all 94 predictor variables. Such a model would be more practical for broader use in industry and research, as less data collection would be required for its implementation. In order to discover the optimal model which contains fewer predictors while still exhibiting sufficient explanatory power, we proceed in the next section with LASSO regression.

**LASSO Logistic Regression (With PCA)**

Lasso regression is a method which utilizes a tuning parameter, $\lambda$, which controls the number of coefficients that take a value of zero. As such, LASSO along with cross validation is commonly used to find a model which has sufficient explanatory capabilities, while not being overly-complex. The LASSO logistic regression model works by penalizing the log-likelihood of beta such that some components of beta become zero. Its equation when performed alongside k-fold cross validation is specified as:

$$\hat{\beta}_T^K = argmin\{\frac{1}{\frac{n(K-1)}{K}} dev_{-k}(\beta) + \lambda_t \sum_k |\beta_j|\}$$

Where $\hat{\beta}_T$ is a p-dimensional vector of the beta coefficients estimated for every fold besides the k-th fold, and deviance here is calculated as binomial deviance. It is apparent from the equation that higher values of lambda will penalize complexity in the model, while lower values of lambda will allow for more complexity. Our goal is to determine the optimal value of lambda for the bankruptcy prediction model through k-fold cross validation. We will perform LASSO logistic regression on both the principal components as well as the variables themselves, in order to determine which approach is superior in terms of out of sample error rate. We will select the optimal value of lambda by determining which value of lambda minimizes the out of sample error through cross validation.

**Classification Tree**

So far the methods we have discussed are parametric in nature. There are other non-parametric methods for prediction problems which have been applied to bankruptcy prediction models. Gepp and Kumar (2015) demonstrated that decision trees have good bankruptcy prediction accuracy when trained against cost ratios. Classification trees works by iterating over the data so as to minimize deviance. Due to the binary classification problem at hand, deviance here is again defined as binomial classification deviance. The tree begins with the dataset and iteratively breaks down the data into smaller and smaller subsets such that the child sets are split in a way which minimize deviance from expected values. For example, at the split $x_{i,j}$, we have the child sets defined as left: $\{x_k y_k : x_{kj} \leq x_{ij}\}$ and right: $\{x_k y_k : x_{kj} > x_{ij}\}$. The algorithm continues the process of splitting the dataset until the size of leaf nodes reaches a minimum threshold.

I implement two classification trees on the data. The first is unpruned, meaning that the number of leaf nodes is left unspecified. Because classification trees can be prone to overfitting when left unpruned, I also run a second classification tree which is pruned. Pruning reduces the size of the classification tree by turning some branch nodes into leaf nodes, thus resulting in less nodes and consequently less prone to overfitting. The optimal tree size for the pruned tree will be determined by the tree which yields the lowest cross-validated out of sample error. After determining the lowest cross-validated out of sample error, we find that the optimal tree size is 4.

**Random Forest**

Random forests are another non-parametric method for classification which may be useful in predicting firm bankruptcy. It is conceptually similar to the classification tree approach. It operates on the assumption that a large number of relatively uncorrelated trees will on aggregate outperform the individual models they are comprised of. By decorrelating the set of trees, errors in individual trees are less likely to have an outsized effect on the model. We first build a set of decision trees from bootstrapped training samples. For each split in each tree, only a random set of $m$ predictors may be used in the algorithm's decision making process. The fact that only a subset of predictors are available at each split allows the entire set of trees to be decorrelated from each other. A common value that we will use for $m$ is $\sqrt{p}$, where $p$ is the number of parameters in our model. Once again, the data will be divided into a test set and a training set, so that the number of trees constructed, $B$, can be determined based on out of sample error. We want B to be sufficiently large enough for the error rate to converge towards its minimum value.

## Results

In order to determine the optimal model for the purpose of predicting firm bankruptcy in Taiwan, we can compare each model's out of sample error using both MSE and misclassification error rate. For all models, a predicted firm bankruptcy was defined as the event that the given firm had a predicted probability of bankruptcy greater than 0.5. Model (1) was a logistic regression which contained all 94 variables. To estimate the out of sample error, perform cross validation with k=10 folds. The out of sample MSE for the model which contains all 94 predictor variables is 0.0438. This is to be expected, since models with more variables will tend to result in a lower MSE. Since this is essentially a binary classification problem, we can also evaluate model performance through the implementation of confusion matrices. We can run the fitted model against 1364 remaining values which were not present in the original model or the cross-validation stage. To be clear, we have already cross validated the model within the first 5456 observations using K=10 folds. Now, we use the remaining 1364 observations to obtain out of sample predictions so that we can form a confusion matrix. The confusion matrix for model (1) is in the appendix. Using the confusion matrix, we can calculate an out of sample misclassification error rate for all models. It is defined as the number of false positives plus the number of false negatives divided by the total number of positives plus the total number of negatives. The out of sample misclassification error rate for this model was 0.02568, which is relatively close to zero and thus satisfactory.

Next, we analyze the same model, but this time ran against the 94 principal components rather than the predictor variables. Call it model (2). The model ran against the set of all 94 principal components has an out of sample mean squared error of 0.07869011, which is higher than model (1). The model was once again validated on K=10 folds. We can also obtain a new confusion matrix, from which we estimate the out of sample classification error rate for this model to be 0.1371974. Based on the OOS MSE as well as the OOS misclassification error rate, it is apparent that model (1) which contains all the predictor variables rather than the principal components exhibits superior predictive capabilities to model (2).

As mentioned before, for the sake of practicality we would like to find a model which does not require the set of all predictor variables as input. We use LASSO to find such a model, on first the principal components, and then the variables themselves. We find that for model (3), which contains only the principal components subject to the LASSO procedure, our cross validated out of sample binomial deviance is minimized for $\lambda = 0.003806$. The minimum value of OOS binomial deviance we obtain for the corresponding value of lambda is 0.1986328, and its OOS MSE is 0.0322. We can once again compare all the models based on their OOS misclassification error rate. For model (3), the misclassification error rate is 0.03592375. In fact, if model (3) predicts that a given firm will go bankrupt, there is a 66% chance that the prediction will be correct. These results are comparable to model (1), while achieving significantly less complexity with only 36 principal components included rather than the complete set of 94.

For the LASSO model ran on the predictor variables themselves (model (4)), performance improves further. Our new out of sample binomial deviance calculated from K=10 cross validated folds is now 0.1863 for the minimum value of $\lambda = 0.0029$. The new binomial deviance is slight but non-negligible improvement from model (3). Our out of sample misclassification error rate is now 0.03519062, and our OOS MSE is greatly reduced to 0.0281. Additionally, if the algorithm predicts that a firm is bankrupt, there is a 63% chance that they actually are bankrupt. If a firm is not predicted to be bankrupt then there is only a 0.03% chance that they actually are bankrupt. These results are comparable, if not slightly improved, from model (3). The misclassification error rate is significantly lower than model (2), and comparable to model (1). Because model (4) obtains one of the lowest OOS MSE as well as one of the lowest misclassification error rates while being significantly less complex than all other models analyzed thus far (it only has 25 non-zero beta terms), it appears to be the best model in terms of balancing predictive accuracy with interpretability and practicality. A table summarizing the results of logistic regression model (4) is available in the appendix. A plot representing the cross validated OOS binomial deviance for different values of lambda is also included, and it shows that the fit of the model is not very sensitive with respect to nearby values of lambda, but becomes more sensitive as the log of lambda moves towards zero.

It was mentioned that regression trees can also be useful for predicting firm bankruptcy. I first estimate an unpruned tree (model (5)), and then proceed by pruning that tree through cross-validation in order to avoid overfitting the data (model (6)). We can once again compare this model with the others based on its out of sample misclassification error rate calculated from a confusion matrix. We find that model (5) produces a misclassification error rate of approximately 0.03228 and a OOS MSE of 0.04750733. This may seem comparable to the other models, but we realize that model (5) is somewhat less accurate than model (4) due to the fact that when model (5) predicts a firm to be bankrupt, there is only a 59% chance that they actually are bankrupt. This reduced accuracy may be due to the tree overfitting the training model. We can attempt to prune the tree to see if it improves the out of sample classification. We can determine the optimal number of terminal nodes, once again through cross-validation. When we do so, we find that out of sample deviance is minimized when we have four terminal nodes. For model (6), the pruned classification tree, we achieve a misclassification error rate of 0.03521643 and a slightly reduced OOS MSE of 0.04633. Additionally, if the tree predicts a firm to be bankrupt, there is now only a 50% chance that the firm is actually bankrupt. It does not appear as if the pruned tree performs any better than the unpruned tree. The pruned tree is included in the appendix. From the tree, it is apparent that “Net Value Growth Rate” was the most predictive variable, followed borrowing dependency, and finally continuous interest rate after tax. From this we may infer that firms with a net value growth rate below a certain threshold (approximately 0.00034) are more likely to be bankrupt.

We will now move forward with the random forest model. Following convention, we set the number of parameters considered at each node split to be equal to $\sqrt{p}$, where p is the number of parameters in the data (94). A measure of deviance for this binomial classification problem is out of bag error rate, and it converges towards 3.02% as the number of trees increases past 100. Unfortunately, because random forests act as a black box ensemble classifier, it is meaningless to plot any of the single trees. We should not directly compare the out of bag error rate to the MSE or binomial deviance estimated for the other models. However, we can once again obtain misclassification error rates in order to better understand how model (7) compares to the other models. Our test set misclassification error rate turns out to be 2.79%. However, it would be misleading to state that this model performs any better than the other ones. In fact, if model (7) predicts a firm to be bankrupt, there is only a 0.08% chance that the firm is actually bankrupt. This is significantly poorer performance than the other models we have analyzed. In total, the random forest model is too conservative in its decision making; it only classifies a total of 5 firms in the test data as being bankrupt.

Overall, the random forest is seen to perform the poorest job at predicting firm bankruptcy in terms of both OOS MSE and misclassification error rate. It may be the case that random forest performs poorly because of the class imbalance in the dataset. We have very few values of ‘1’ to represent bankruptcy, and it is known in the literature that random forests and decision trees in general are sensitive to class imbalance. According to a paper by Cieslak & Chawla, a training set consisting of different numbers of representatives from either class may result in a classifier that is biased towards the majority class. With this being known, the relative underperformance of the random forest and decision trees compared to the other models seems expected.

The logistic regression models exhibit the highest degree of predictive accuracy when evaluated by OOS MSE as well as misclassification error. And of the logistic models we have estimated, the LASSO procedure ran on the predictor variables themselves rather than their principal components balances predictive capabilities with simplicity. LASSO performs variable selection, and can yield a reduction in variance at the expense of increased bias (Taddy). LASSO performed well when applied to this data because it turns out that most of the predictors are irrelevant in predicting firm bankruptcy. Even when we use only 25 of the original 94 variables, LASSO achieves results comparable, if not superior to logistic regressions ran on all 94 variables. Obviously, when faced with two models with similar explanatory capabilities, we should always opt for the simpler one. Hence we conclude that the LASSO logistic regression model performs the best in predicting firm bankruptcy of all the models assessed.

Having identified model (4) as the model which best balances practicality with explanatory capability, we proceed by predicting one value for a firm's bankruptcy status given the set of 25 predictors. For firm #2100 in the test dataset, we have various values related to return on assets which are lower than most values observed in the dataset. Geng et al. (2015) identified return on assets as one of the most powerful predictors of firm bankruptcy status. Therefore we should expect firm #2100 to be bankrupt, since the values that multiple of its observed variables related to return on assets are well below average for the rest of the data in the test sample. It turns out that firm #2100 is in fact a bankrupt firm. Indeed, when we run observed values for firm #2100 through logistic regression model (4) with 25 non-zero beta coefficients, we obtain a predicted value of '1', indicating that the model correctly predicted the bankruptcy status of the firm.

## Conclusion

In this paper, we have demonstrated the predictive capabilities of various machine learning algorithms in predicting firm bankruptcy among a dataset of Taiwanese firms. Logistic regression with LASSO on the set of real variables (model (4)) proved to be the method which best balanced explanatory capability with practicality, in that it only used 25 of the 94 possible predictor variables. In both LASSO and the initial full models with 94 variables, the Principal Components yielded higher out of sample error rates as measured by both out of sample binomial deviance and misclassification error rate. The regression tree and random forest performed the worst of all the models in terms of their out of sample misclassification error rate.

More research is needed on the use of random forests and decision trees in predicting firm bankruptcy. Because Gepp and Kumar (2015) already demonstrated that regression trees can be useful in predicting firm bankruptcy, researchers should investigate whether there exists a method of re-weighting classes so as to avoid the bias these models experience when trained on unbalanced classes.

It is important to note that the predictive capabilities of models discussed here only apply to Taiwanese firms. Therefore, structural and regulatory differences which exist between different countries may reduce the accuracy of these models in when applied in other contexts.

## References


Gepp, Adrian, and Kuldeep Kumar. 2015. Predicting Financial Distress: A Comparison of Survival Analysis and Decision Tree Techniques. *Procedia Computer Science* 54: 396–404.

Geng, Ruibin, Indranil Bose, and Xi Chen. 2015. Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research* 241: 236–47.

Ben Jabeur, Sami. 2017. Bankruptcy prediction using Partial Least Squares Logistic Regression. *Journal of Retailing and Consumer Services* 36: 197–202.

Svabova, Lucia, Marek Durica, and Ivana Podhorska. 2018. Prediction of Default of Small Companies in the Slovak Republic. *Economics and Culture* 15: 88–95.

Shlens, A Tutorial on Principal Component Analysis, http://arxiv.org/abs/1404.1100,2014.

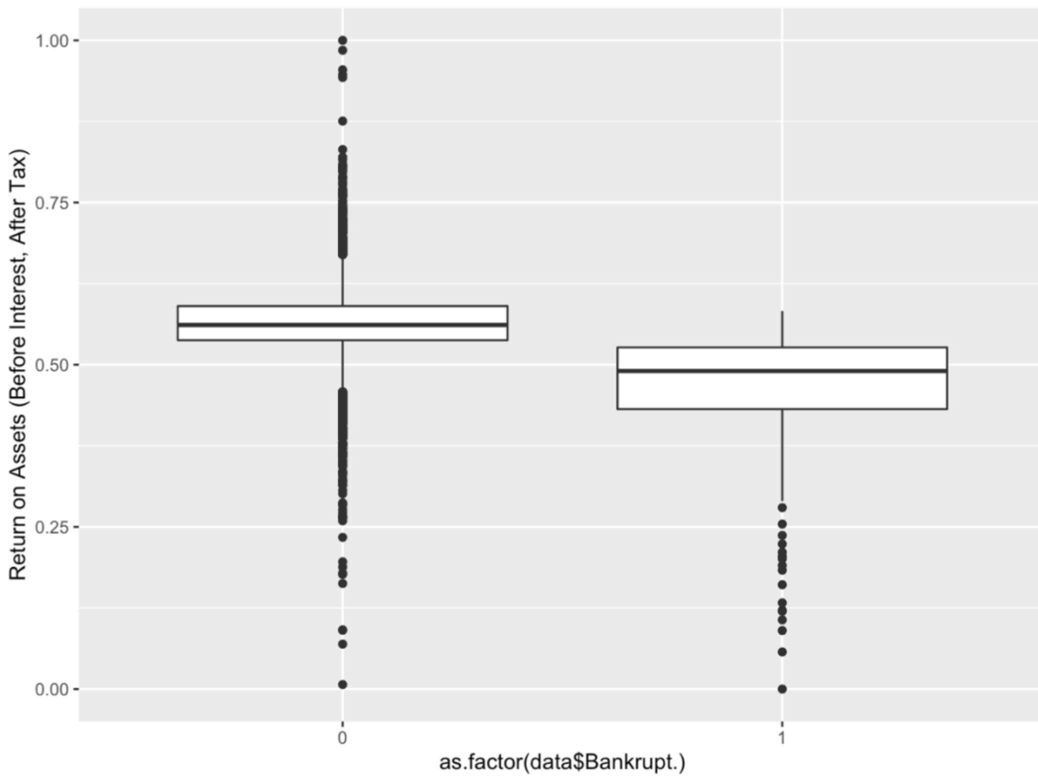
Chawla, N.V., Cieslak, D.A., Hall, L.O. et al. Automatically countering imbalance and its empirical relationship to cost. Data Min Knowl Disc 17, 225–252 (2008). https://doi.org/10.1007/s10618-008-0087-0

Taddy, M. (2019). *Business data science: combining machine learning and economics to optimize, automate, and accelerate business decisions*. McGraw-Hill Education.

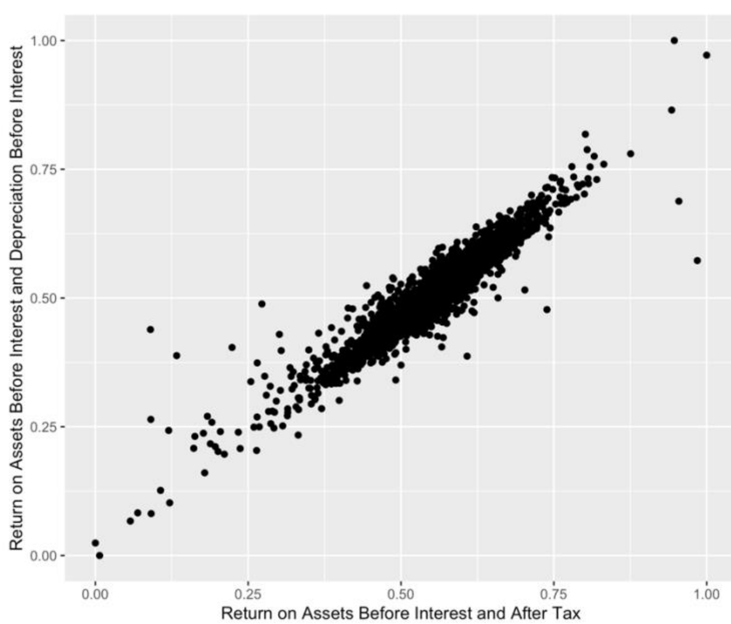
# Appendix

## A.1: Tables & Figures

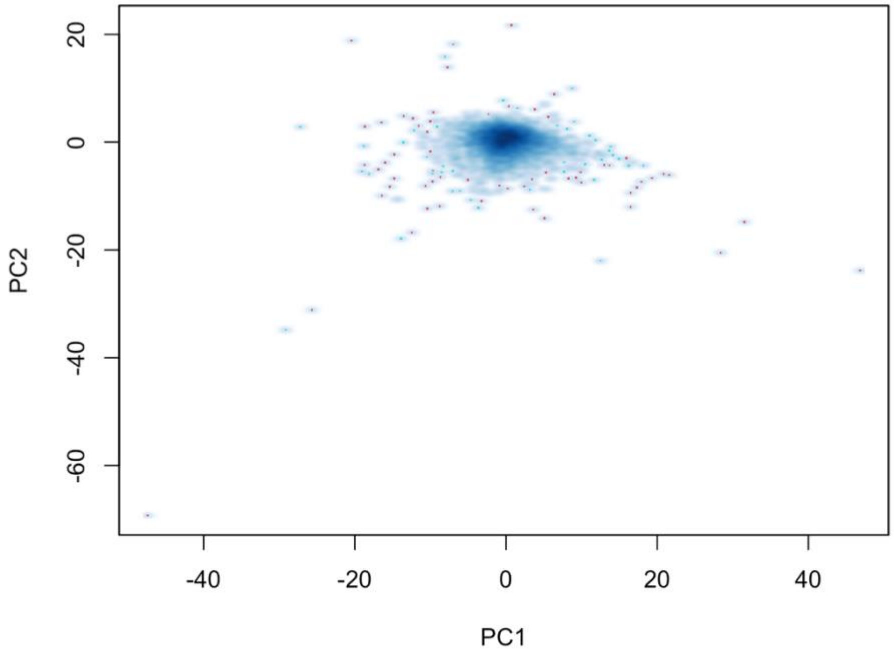
### Boxplot of Return on Assets by Bankruptcy Status (1=bankrupt)



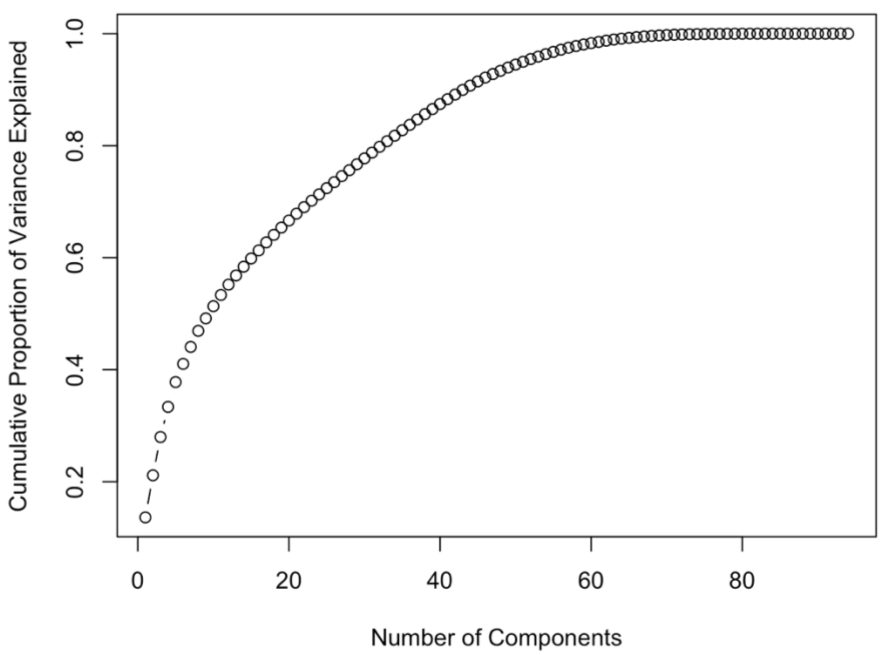

### Depiction of Multicollinearity Among Variables with Similar Accounting Identities



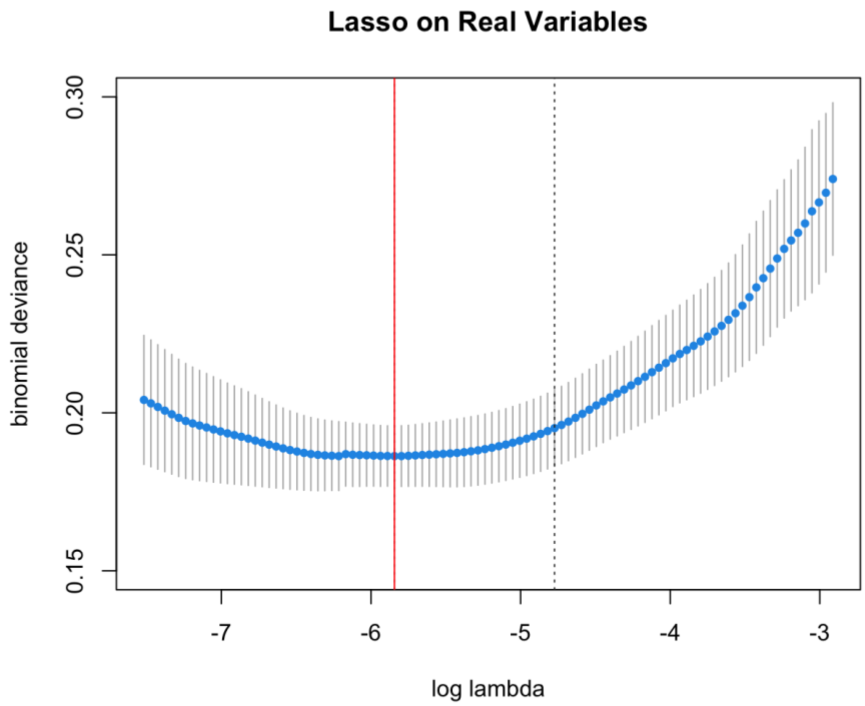
**Plot of First Two Principal Components (Red=Bankrupt)**



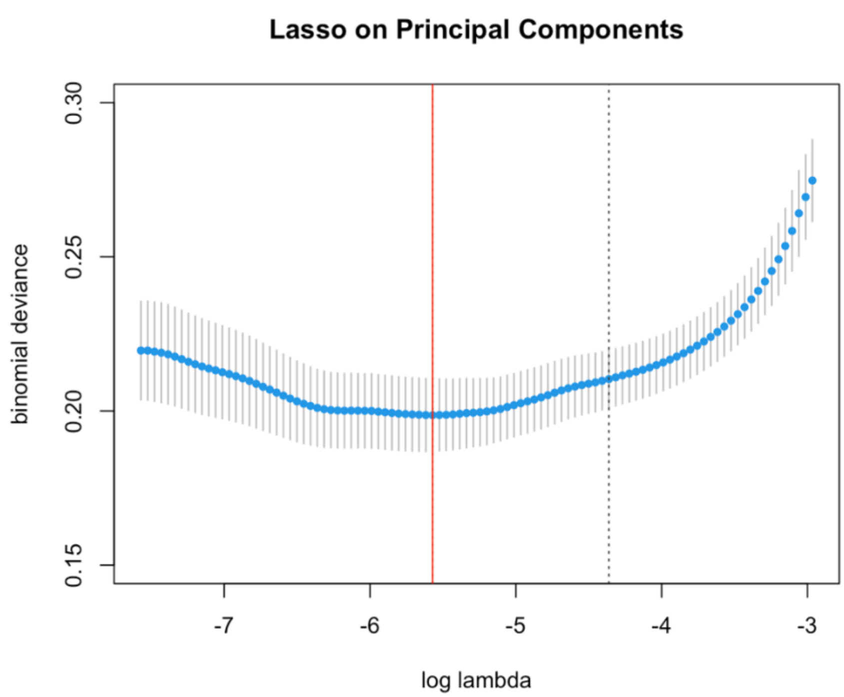

**Proportion of Cumulative Variance Explained by the Principal Components**



**Sensitivity of LASSO (Real Variables Only)**




**Sensitivity of LASSO (Principal Components Only)**



**Summarized Results for Best Model (4) (LASSO Logistic Regression on Real Variables)**
Note that certain coefficient values which have less than three significant figures are reported as 0

Table 1: Cross Validated LASSO Logistic Regression (Minimal Error)

| | Coefficient | Lambda Location | Value |
|---|---|---|---|
| 1 | intercept | seg64 | 10.176 |
| 2 | ROA.B..before.interest.and.depreciation.after.tax | seg64 | -3.425 |
| 3 | Operating.Profit.Rate | seg64 | 0.171 |
| 4 | Tax.rate..A. | seg64 | -0.178 |
| 5 | Net.Value.Per.Share..B. | seg64 | -4.357 |
| 6 | Persistent.EPS.in.the.Last.Four.Seasons | seg64 | -23.974 |
| 7 | Quick.Ratio | seg64 | 0 |
| 8 | Total.debt.Total.net.worth | seg64 | 0 |
| 9 | Debt.ratio.. | seg64 | 13.169 |
| 10 | Net.worth.Assets | seg64 | -2.560 |
| 11 | Borrowing.dependency | seg64 | 2.899 |
| 12 | Contingent.liabilities.Net.worth | seg64 | -26.904 |
| 13 | Total.Asset.Turnover | seg64 | -4.137 |
| 14 | Fixed.Assets.Turnover.Frequency | seg64 | 0 |
| 15 | Revenue.per.person | seg64 | 0 |
| 16 | Operating.profit.per.person | seg64 | -0.105 |
| 17 | Quick.Assets.Total.Assets | seg64 | -0.161 |
| 18 | Cash.Total.Assets | seg64 | -1.456 |
| 19 | Cash.Current.Liability | seg64 | 0 |
| 20 | Working.Capital.Equity | seg64 | -3.273 |
| 21 | Retained.Earnings.to.Total.Assets | seg64 | 3.365 |
| 22 | Cash.Turnover.Rate | seg64 | -0 |
| 23 | Cash.Flow.to.Liability | seg64 | -3.833 |
| 24 | Cash.Flow.to.Equity | seg64 | -4.014 |
| 25 | Liability.Assets.Flag | seg64 | -0.817 |
| 26 | Net.Income.to.Total.Assets | seg64 | -4.749 |
| 27 | OOS CV Binomial Deviance | | 0.1863 |
| 28 | OOS misclassification error rate | . | 0.03519062 |

**Model (1) Confusion Matrix**

| TestPredict | Not Bankrupt | Bankrupt |
|---|---|---|
| Predicted Not Bankrupt | 1287 | 22 |
| Predicted Bankrupt | 37 | 17 |

**Model (2) Confusion Matrix**

| TestPredict | Not Bankrupt | Bankrupt |
|---|---|---|
| Predicted Not Bankrupt | 1173 | 45 |
| Predicted Bankrupt | 142 | 3 |

**Model (3) Confusion Matrix**

| TestPredictPCR | Not Bankrupt | Bankrupt |
|---|---|---|
| Predicted Not Bankrupt | 1311 | 47 |
| Predicted Bankrupt | 2 | 4 |

**Model (4) Confusion Matrix**

| TestPredict | Not Bankrupt | Bankrupt |
|---|---|---|
| Predicted Not Bankrupt | 1309 | 44 |
| Predicted Bankrupt | 4 | 7 |

**Model (5) Confusion Matrix**

| TestPredict | Not Bankrupt | Bankrupt |
|---|---|---|
| Predicted Not Bankrupt | 1306 | 35 |
| Predicted Bankrupt | 9 | 13 |

**Model (6) Confusion Matrix**

| TestPredict | Not Bankrupt | Bankrupt |
|---|---|---|
| Predicted Not Bankrupt | 1306 | 39 |
| Predicted Bankrupt | 9 | 9 |

**Model (7) Confusion Matrix**

| | 0 | 1 | class.error |
|---|---|---|---|
| 0 | 1311 | 4 | 0.003041825 |
| 1 | 39 | 9 | 0.812500000 |

## Pruned Regression Tree with Optimal Size