Weighted LR vs. Bayesian LR: Bankruptcy and Liquidation Prediction

Junwoo Yang* Jaeseon Lee†

July 1, 2021

Abstract

In this study, we examine three methods of Logistic Regression (LR) to predict bankruptcy and liquidation of US companies: Maximum Likelihood LR, Weighted LR, and Bayesian LR. While most previous prior studies focused on predicting bankruptcy, this study predicts liquidation as well as bankruptcy. Since bankruptcy is an extremely rare event, including liquidation in the response variable could alleviate the problem caused by imbalanced data. Since there are about 400 financial variables, we take preselection steps by t-test and VIF before selecting variables by AIC. We predict companies that will go bankrupt or liquidate within three years based on a certain financial year.

Keywords: fundamental, bankruptcy, liquidation, imbalanced data, variable selection, maximum likelihood LR, weighted LR, Bayesian LR

1 Introduction

Business bankruptcy and liquidation forecasts are important. Creditors and entity investors must correctly assess the probability of an entity's default on its obligations for profitable decisions. For banks, accurate forecasts of corporate bankruptcy and liquidation enable safe lending operations for businesses and impose interest rates that adequately reflect risks. In addition, if an accounting firm makes inappropriate predictions about the possibility of bankruptcy of the audited company, it may be caught in a lawsuit. On the other hand, the choice of explanatory variables and the choice of functional form between these variables are key issues when constructing bankruptcy and liquidation prediction models.

Previous studies on bankruptcy prediction employ economic implication of variables. For example, Laitinen and Laitinen (2000) analyzed the ratio of cash, cash flow and equity to total assets as explanatory variables, based on the fact that the more cash flows, the more net cash flows, and the more flexible financing from outside. According to Laitinen and Laitinen (2000), only the ratio of cash to total assets and equity capital were statistically significant when the analysis was conducted based on financial position a year before bankruptcy. Kuruppu, Laswad, and Oyelere (2003) devised a model for forecasting liquidation for New Zealand entities, which included 63 variables in their financial statements. Of these, only 12 variables, in-

This paper is a project report of Math Capstone PBL (Data Analysis) taught at Hanyang Univeriesty.

*Department of Finance

†Department of Economics and Finance

cluding the ratio of equity to total assets, were statistically significant. Theoretical approaches to variable selection are rare.[1] In this study, approximately 900 financial variables are included in the candidates of variables in the final model.

Bankruptcy-themed studies typically postulate linear (linear discriminant analysis) or logistic (logistic regression) function relationships between variables. In particular, logistic regression analysis is effective in analyzing bipartisan data, such as whether an individual has cancer or whether a company has gone bankrupt. Indeed, Laitinen and Laitinen (2000) tried to devise a corporate bankruptcy prediction model using the logistic regression model and Taylor deployment. Meanwhile, Hauser and Booth (2011) derives a bankruptcy prediction model based on the robust logistic regression model. They analyzed two models after dividing the estimated direction of the coefficients by the Bianco-Yohai estimate and the maximum likelihood estimate. As a result, data from 2006 to 2007 showed that only the former model accurately predicted Lehman Brothers' bankruptcy. This work seeks to devise a corporate bankruptcy and liquidation prediction model using a logistic regression model. Based on the data, we would like to proceed with the fit and evaluate the performance of the fit model to determine whether it is useful as a predictive model.

This study seeks to model mid- to long-term bankruptcy and liquidation forecasts. In general, an entity's bankruptcy or liquidation is not determined in a short period of time, and signals about it are likely to be transmitted over a long period of time. That is, even if the entity's data for a particular year diagnoses that the entity is about to enter bankruptcy or liquidation, it may enter bankruptcy or liquidation two or three years later than the following year. On the other hand, setting the period too long will reduce the explanatory power of the model. Therefore, this study sought to cover most of the cases by presupposing an appropriate period of three years. In addition, because the financial position of the entity is likely to be very bad just before bankruptcy and liquidation, it would be very good in terms of the practicality of the forecast if a reasonable medium-term forecast model could be devised. For three years from 2011 to 2013, the company tried to analyze and explain whether it was bankrupt or liquidated based on 2010 data. Through this study, we would like to first identify which explanatory variables are significantly related to bankruptcy and liquidation, and predict which companies will go bankrupt and liquidated within the next three years based on the final model and the current data in 2020.

The remainder of this paper is organized as follows. Section 2 presents the formal model. Section 3 characterizes the unique symmetric equilibrium of our auction model, and Section 4 studies the effects of (varying) information costs on the symmetric equilibrium outcome. Section 5 extends the analysis in order to account for the reserve price. Section 6 examines two simple asymmetric environments. Section 7 concludes by suggesting some directions for future research.

[1] Laitinen and Laitinen (2000)

2 EDA

2.1 Data

The annual fundamental data used in this paper are from the Compustat Capital IQ database of WRDS. The observations are firms traded on the NYSE, AMEX, NASDAQ, TSX, and NYSE Arca from January 2000 through November 2020. This data set includes a total of 981 variables and 226,866 observations.

The location data are from the United States cities database of SimpleMaps[2]. Meanwhile, corporate fundamentals data only contained information about the states and cities in which individual companies existed. We wanted to see the distribution of companies by state, county, and city at a glance, and further analyze in detail whether the company's bankruptcy and liquidation are different depending on geographical requirements. Therefore, in addition to the enterprise data, additional data reflecting the geographical characteristics of the United States were built. These materials include the names of City, County and State, County FIPS, and the longitude and latitude of the city. We used maps to visually understand the county-specific data merged with the fundamental data based on City and State, and used longitude and latitude to show the company's coordinates to understand the distribution at a glance.

In the meantime, the headquarters wanted to exclude companies outside of the United States from the analysis. This provided convenience in the process of examining the relationship between geographical requirements and other variables. Increasing the proportion of bankruptcy and liquidation companies in the data also allowed for a slight reduction in the data's excessive imbalance. First of all, the headquarters excluded 3,058 companies that exist outside the United States from the analysis. It also excluded four entities that do not have information about the state from the analysis for entities existing entities in the United States. On the other hand, the analysis excluded 19 entities that disappeared before 2011 and seven in Puerto Rico and Guam for some reason.

The variables in the data are divided into the following eight categories: Identifying Information, company descriptor, balance sheet items, income statement items, cash flow items, miscellaneous items, supplemental data items, maps items.

The enterprise data utilized in this study encompasses 981 variables. Therefore, we found it difficult to explain all the meanings of individual variables, and it was meaningless when it came to analysis. Table 2 lists 414 of the 981 variables. Among them, we only want to elaborate on the variables included in the final model.

First, aco means an asset that is not included in the balance sheet in cash, cash equivalents, uncollected, or inventories. aqpl1 represents an asset calculated at fair value. Caps are capital surplus and csho is the number of common shares in the market at the end of the year. The cstk represents the total face value of all general equity. Glcea is the after-tax amount of sales increases and decreases not included in the calculation of S&P. idbflag is a categorical variable that represents the source of the data, divided into three cases: within the United States (idbflag=D), outside the United States (idbflag=I), and both (idbflag=B). Since the source of the data has never been outside the United States (idbflag=I), it has virtually two categories.

[2] https://simplemaps.com/data/us-cities

On the other hand, companies (B), whose sources are domestic and foreign, can be interpreted as multinational companies. Optfvgr represents a weighted average of the fair value of the option granted over the year. Spced refers to diluted earnings per share on an S&P basis. On the other hand, stalt is a categorical variable indicating that the entity is at risk of bankruptcy or is in the process of acquiring the entity (stalt=1). stkcpa refers to the amount of share-based compensation that has been treated as an expense in the income statement.

This study defines BL variables that represent 1, if and when an entity goes bankrupt and liquidates, as dependent variables and proceeds with the analysis. Within the data built from WRDS, the BL variable could be redefined by classifying only observations corresponding to bankruptcy and liquidation in DLRSN variables that separated the entity from 14 causes of disappearance. Table 6 summarizes the number of companies that went bankrupt and liquidated during the 2011-2020 period in the DLRSN variable in the raw material.

In particular, this study seeks to devise a mid- to long-term forecast model for bankruptcy and liquidation, considering that the bankruptcy and liquidation process of an entity does not take place in a short period of time. Therefore, instead of defining an individual year as a period of the dependent variable, the entity postulates a sufficient period of three years. This study selected three years from 2011 to 2013. Thus, the response variable is defined as follows:

$$y_i = \begin{cases} 1 & \text{if firm } i \text{ was bankrupt or liquidated between 2011 and 2013,} \\ 0 & \text{otherwise. (solvent company)} \end{cases}$$

2.2 Handling missing value

Missing values reduce the statistical power, which indicates the probability of correctly rejecting the null hypothesis when it is false, and furthermore, can produce bias in parameter estimation. Thirdly, missing values reduce the representation of samples, and finally complicate data analysis.[3]

Prior to full-scale processing of missing values, 26 variables with the same value were removed from all observations. This is because variables that all entities have the same value do nothing to explain whether they are bankrupt or liquidated. We also exclude 552 variables with missing values above 80% from our analysis. This summarizes the data into 349 continuous variables, 65 categorical and other informational variables, a total of 414 variables and 7,461 observations.

Next, the analysis was conducted by estimating the missing values within the continuous variables as the industry sector-specific average. This is because for enterprise-related data, characteristics may vary depending on the industry sector in which the enterprise is engaged. By representing missing values as industry-specific average values rather than overall average values, we tried to maintain characteristics according to industry segments of the industry.

The North America Industry Classification System (NAICS) was developed jointly by the U.S., Canada, and Mexico to provide new comparability in statistics about business activity across North America. Each North America company in Xpressfeed is assigned a 6-digit NAICS code. A company's NAICS code classifies the company on a production and/or process-

[3] Kang (2013)

oriented basis at five different levels: Sector, Subsector, Industry Group, NAICS Industry, and National. For example, Microsoft Corporation's NAICS code **511210** breaks down as follows:

Sector	51	Information
Subsector	511	Publishing Industries (except Internet)
Industry Group	5112	Software Publishers
NAICS Industry	51121	Software Publishers
National	511210	Software Publishers

The commonly used SIC code categorizes industries into four digits, with a limitation that has not changed since its revision in 1987. It was judged that this may be a somewhat inappropriate classification standard in modern society, where industrial flows are rapidly changing. Therefore, the NAICS code revised in 2017 was considered a more appropriate criterion, classifying the industry in more detail than the SIC with six digits.

The first two digits of the NAICS represent a general classification of the entity's economic activities, the third being its sub-sector and the fourth being its industrial group. The fifth represents the NAICS classification of industrial groups, and the last sixth represents the national industry. Table 3 represents the structure of NAICS. Column 1 represents the first two digits of NAICS, and column 2 represents the number of observations in the data. A detailed description of the first two digits of NAICS is described in column 3. For example, an entity with the first two digits of the NAICS code 31 is classified as a manufacturing industry, and if 52 is classified as a financial or insurance industry.

NAICS code for Monster Beverage Inc., known for its energy drinks, is 312111. An example of Monster Beverage Inc. in Table 4 shows that the classification becomes more granular as the number of digits increases from 31 to 312, 3121. Nike, meanwhile, is classified as a manufacturing company, starting with the same 31 as Monster Beverage Inc., but the detailed classification is different. Nike was classified as 316 for leather and related product manufacturing, indicating the final 316210. In the NAICS code, the industrial sector below 3162 is the same as the shoe manufacturing sector.

This work estimates missing values as NAICS-specific averages. Starting with the 6-digit average, we proceeded with the upper classification when it was impossible to classify the lower classification, such as the 5-digit mean and 4-digit mean. On the other hand, there were 560 missing NAICS variables, which could reasonably be estimated based on their relationship to other variables. The observation with missing NAICS values was either the entity's SIC code: 6722 or 6726. Therefore, they conducted the analysis by replacing the SIC-specific averages of the entity. The above processing of missing values allowed us to handle all the missing values present in 350 continuous variables.

2.3 Preselection by t-test and VIF

Before proceeding with the correlation analysis of 349 continuous variables and 65 categorical and other informational variables, we wanted to eliminate the analytical unnecessary variables. In particular, we try to first screen variables that do not help explain dependent variables and variables that have an overly high correlation between explanatory variables. First, after conducting a t-test to see if there is a difference in the mean between the continuous explanatory variables for each group of bankruptcy and liquidation, a continuous explanatory variable that is considered irrelevant to the dependent variable BL was primarily excluded from the analysis. Specifically, the null hypothesis of equal variance in groups within variables was tested using the F-test, and based on this result, a t-test (the null hypothesis: equal mean) was conducted for the difference in means. As a result of the F-test, we conducted a Welch's two sample test for variables that determined that the variance was different, and a variable that determined that the variance was the same proceeded with a two sample test. The t-test excluded 105 of the 349 continuous variables from the analysis.

Next, we investigate the existence of a multi-collinearity problem between explanatory variables and solve it. Multicollinearity, which means a complete or almost complete linear dependent relationship between explanatory variables, makes the variance of the estimated regression coefficients very large, which reduces reliability. It also poses problems with the interpretation of regression models. The interpretation of regression coefficients presupposes the fixed state of all other explanatory variables, where the existence of multicollinearity contradicts the assumption. Furthermore, the sign of the estimate of the regression coefficient contradicts empirical or theoretical expectations. This study conducted multicollinearity verification using the Variance Inflation Factor (VIF). We investigate the VIFs of explanatory variables and try to solve the multicollinearity by excluding the largest variables sequentially one by one from the analysis at the same time that the corresponding figure has a value of 10 or higher. A total of 90 variables were included in the analysis, excluding 154 of the remaining 244 variables that were excluded through t-test throughout the process. Through the above two processes, we were able to organize the data into 90 continuous variables and 65 categories and other informational variables, a total of 155 variables. Table 5 is a summary of this.

Finally, the analysis also excludes missing values and variables with more than 6,000 zero values based on raw materials. This was aimed at screening as much meaningless variables as possible in the first-order variable selection process, as we judged that too many values with zero could not reasonably be processed, even if the missing values were not more than 80%. In other words, we found that variables with missing values of 60% and the remaining 30% have zero values are less convincing as they result in estimating missing values from observations of the remaining 10%. Through this process, the data was organized into 76 variables. In the meantime, the analysis excluded meaningless categorical variables, such as an entity's telephone number, and other information variables about the entity. As a result, the first-order variable selection process allowed us to select 44 variables and 7,380 observations to proceed with the analysis.

2.4 Summary statistics and correlation analysis

MLLR.AIC contained the largest number of variables in Section 4. Correlation analysis of the variables contained in MLLR.AIC was performed, leaving out explanations for all variables. Table 7 shows descriptive statistics of continuous variables in the MLLR model. Because the variables were standardized to compare the size of the estimated coefficients, the mean is equal to 0 and the variance equal to 1. Figure 2 shows a map of states, counties, and cities in the

United States. By State, (a) represents the bankruptcy ratio and (b) represents the number of companies. (d) refers to the city in which the corporate headquarters are located. Not only is the sample geographically identified through the map, but it is also included in the variable candidates.

There are three categorical variables in total, including dependent variables. Table 9 shows the results of cross-tabulation analysis between them. While stalt and idbflag are associated with dependent variables, stalt and idbflag are independent of each other. In terms of multicollinearity, variable selection is considered appropriate. It can also be visually identified by box plot in Figure 3. Since BL is t-tested, continuous variables differ visibly with BL, but idbflag and stalt differ relatively less significant with respect to continuous variables. The multicollinearity of the selected explanatory variables and their association with the response variables cannot be fully reflected, but they are at a satisfactory.

3 Methods of logistic regression

This section provides model and notation of Logistic Regression (LR). We begin by estimating parameters of Maximum Likelihood LR (MLLR) and Weighted LR (WLR), and then determine the distribution of parameters of Bayesian LR (BLR) by using R packages and WinBUGS.

3.1 Subsampling imbalanced data

In 2010, there were 7,380 U.S. companies, 76 of which went bankrupt or liquidated between 2011 and 2013. The proportion of events in the population is 1.03% which is rare and has been established in the literature that these variables are difficult to predict and explain. For example, the problem arises that the fitted model boasts high accuracy even if all X_i are classified into negative. Thus, we manipulate the proportion by constructing training set with 60 $Y_i = 1$ and 600 $Y_i = 0$ at random. It gives more weight to group of $Y_i = 1$.

However, this increases not only the rate at which the actual positive is correctly diagnosed as positive, but also the rate at which the actual negative is diagnosed as positive. That is, type 1 error increases instead of reducing type 2 error. In this case, it was considered more tremendous to predict that the company that would actually go bankrupt would not go bankrupt than to the contrary. Therefore, we try to solve the problem caused by imbalanced data through subsampling despite the increase in type I error. The test set consists of 16 bankrupt companies and 160 non-bankrupt companies.

3.2 Maximum Likelihood LR

Let $X \in \mathbb{R}^{n(k+1)}$ be a data matrix where n is the number of observation and k is the number of variables, and $Y \in \mathbb{R}^n$ be a binary outcomes vector. For all observation $X_i \in \mathbb{R}^{k+1}$ (a row vector in X), the outcome is either $y_i = 1$ (positive) or $y_i = 0$ (negative). The goal is to classify X_i as positive or negative. It can be treated as a Bernoulli trial with an expected value $E(y_i)$ or probability p_i. The logistic function commonly used to model each X_i with its expected outcome is given by the following formula:

$$E[y_i|X_i,\beta] = p_i = \frac{e^{X_i\beta}}{1+e^{X_i\beta}} = \frac{1}{1+e^{-X_i\beta}}$$

where $y_i \sim \text{Bernoulli}(p_i)$ and $\beta = (\beta_0, \beta_1, \cdots, \beta_m)^\top$. Let η be the logit link which is the logarithm of the odds ratio. It is defined as

$$\eta_i = \text{logit}(p_i) = \log\frac{p_i}{1-p_i} = X_i\beta,$$

and can be written as $\eta = X\beta$ in matrix form. Now, assuming that the observations are independent, the likelihood function is

$$L(\beta|X,Y) = \prod_{i=1}^{n}(p_i)^{y_i}(1-p_i)^{1-y_i} = \prod_{i=1}^{n}\left(\frac{e^{X_i\beta}}{1+e^{X_i\beta}}\right)^{y_i}\left(\frac{1}{1+e^{X_i\beta}}\right)^{1-y_i} = \prod_{i=1}^{n}\frac{e^{X_i\beta y_i}}{1+e^{X_i\beta}}.$$

and the regularized log likelihood function is defined as

$$\log L(\beta|X,Y) = \sum_{i=1}^{n}\log\frac{e^{y_iX_i\beta}}{1+e^{X_i\beta}} - \frac{\lambda}{2}\|\beta\|^2 \tag{1}$$

where the regularization (penalty) term was added to obtain better generalization. Since the log likelihood function is strictly concave, $\hat{\beta}_{\text{MLE}}$ which maximizes the log likelihood can be found.

3.3 Weighted LR

In addition to subsampling, we can use adjusted likelihood function, called weighted likelihood (King and Zeng , 2001; Maalouf and Siddiqi , 2014). Let $\overline{y}$ and τ be proportion of events in the sample and in the population, respectively. The weighted likelihood is defined as

$$L_W(\beta|X,Y) = \prod_{i=1}^{n}(p_i)^{w_1y_i}(1-p_i)^{w_0(1-y_i)},$$

where $w_1 = \tau/\overline{y}$, and $w_0 = (1-\tau)/(1-\overline{y})$. Now, instead of maximizing Eq.(1), we can maximize the weighted log likelihood

$$\log L_W(\beta|X,Y) = w_1\sum_{y_i=1}\log(p_i) + w_0\sum_{y_i=0}\log(1-p_i) = -\sum_{i=1}^{n}w_i\log(1+e^{(1-2y_i)X_i\beta})$$

where $w_i = w_1y_i + w_0(1-y_i)$.

3.4 Bayesian LR

3.5 Variable selection by AIC and BIC

Both Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are appropriate for models fit under the maximum likelihood estimation framework. Let $\hat{L}$ be the maximum value of the likelihood function for the model. Then the values of AIC and BIC are the following.

$$\text{AIC} = 2k - 2\log\hat{L}, \quad \text{BIC} = k\log n - 2\log\hat{L}$$

where k is the number of estimated parameters in the model and n is the number of observation. Compared to the BIC method, the AIC statistic penalizes complex models less, meaning that it may put more emphasis on model performance on the training set, and, in turn, select more complex models.

The selection of variables in the model utilizes the stepwise selection method. It is a compromise between Forward Selection and Backward Elimination, in which important variables are found while repeating each step selection and elimination. The method reviews step by step whether the already selected variables can be eliminated, while selecting additional important variables one by one. In other words, variables included in the model in the early stages can also be erased in later stages. This work determines the model representing the lowest AIC as the final model through a stepwise selection method. In other words, we thought that the selection of variables was complete when the AIC could no longer be lowered even with the addition or exclusion of additional variables in the current stage. On the other hand, the statistical significance of individual variables was not taken into account because variable selection based on AIC was considered. Model fitting utilizes training data containing 44 variables and 660 observations that have gone through all of the preceding processes. Table 10 shows the result of four regressions. The selected MLLR model by AIC method is

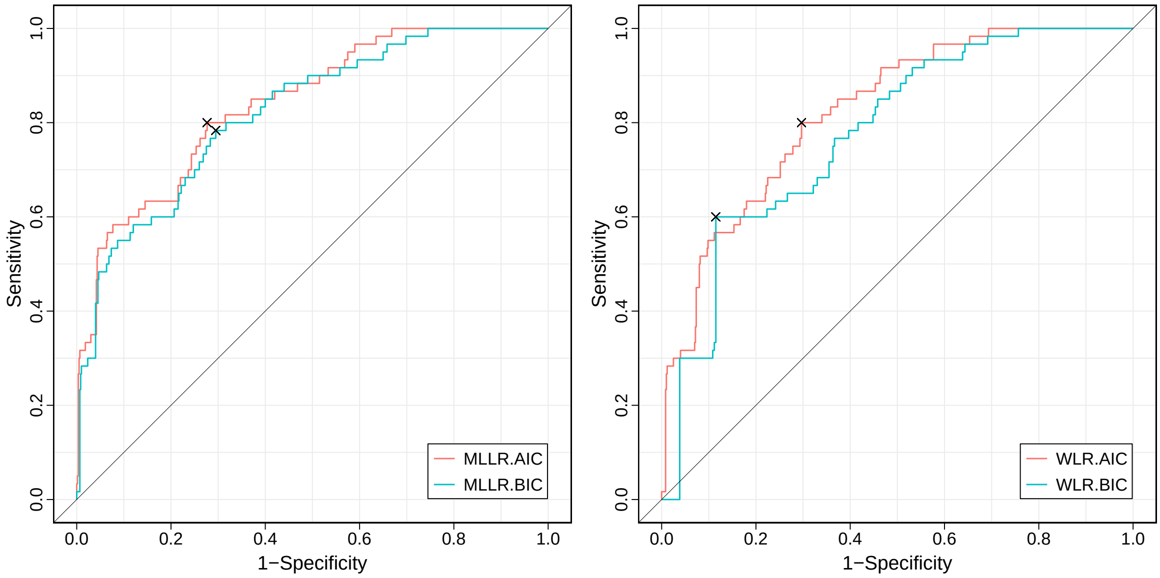
$$\begin{aligned}\eta_i = &-203.6 - 7.2x_{\text{aco},i} + 8.9x_{\text{aqppl1},i} + 2.1x_{\text{caps},i} - 7.6x_{\text{csho},i} - 13.5x_{\text{cstk},i} \\ &+ 1.6x_{\text{glcea},i} + 197.0x_{\text{idbflag},i} - 20.0x_{\text{optfvgr},i} - 2.4x_{\text{spced},i} + 2.7x_{\text{stalt},i} - 3.3x_{\text{stkcpa},i}.\end{aligned}$$

The likelihood ratio test for the final model resulted in a very small p-value of 1.21×10^{-14}, which allowed the null hypothesis to be rejected. In other words, at least one explanatory variable could be determined to have explanatory power over the dependent variable.

4 Performance

The performance evaluation of the model is based on test data. The test data consists of 16 companies with $BL = 1$ and 160 companies with $BL = 0$. We determine the optimal threshold by drawing ROC curves, and determine the superiority of the model through AUC. We also evaluate the performance of the model by understanding classification accuracy, etc., based on the confusion matrix.

The ROC curve represents an aspect of sensitivity and $1-$ specificity with threshold changes. Specifically, the $1-$ singularity represents a false positive rate. The closer sensitivity and specificity are to 1, the better the model is, and thus the optimal critical point is determined closest to the ROC curve and the coordinates with ($1-$ specificity, sensitivity) values (0,1). Meanwhile, the ROC curve derived from the logistic model evaluates the overall performance of the model. The area under the ROC curve is called AUC, and the larger the AUC, the better the model is. If the AUC is 0.5, it is a model of the same level as random estimation, and if it is 1, it is a



(a) Maximum Likelihood LR

(b) Weighted LR

Figure 1: ROC curves

perfect model. Generally, if the AUC is 0.8 or higher, we judge it as an excellent model.

Figure 1 shows the ROC curve of the final model. The optimal cut-off value on (a) is 0.105. This means that if the predicted probability of bankruptcy and liquidation exceeds 0.105, the sensitivity and fit of the entity is best when classifying it as bankruptcy and liquidation. Therefore, on an optimal cut-off basis, an entity is classified as a bankruptcy and liquidation entity if the estimated probability of bankruptcy and liquidation exceeds 0.105. Conversely, if the estimated probability is less than 0.105, the entity is not expected to go bankrupt and liquidate. Meanwhile, the AUC was 0.857. Based on this, the final model could be judged to be relatively good.

Before we begin, I want to summarize the terms. In particular, Positive, Negative within a confusion matrix can be interpreted positively and negatively, which can cause confusion in analysis. Consequently, an entity sought to minimize possible confusion in subsequent processes by reinterpreting the term in terms of bankruptcy and liquidation of an entity. The classification results of the prediction model can be summarized in four ways: TP refers to when an entity classified as bankruptcy and liquidation in the forecast model has actually gone bankrupt and liquidated, and TN refers to when an entity classified as insolvent and liquidated in the model has not actually gone bankrupt and liquidated. On the other hand, FP predicts that the model will bankrupt and liquidate a particular entity, but in reality it is bankrupt and not liquidated, meaning type 1 error. FN predicted that the model would not go bankrupt and liquidate, but in reality it would go bankrupt and liquidate, which is class II error. In particular, the case of classifying bankruptcy and liquidation entities as not likely to be the case when forecasting is a very serious error. For example, if business decisions, such as investment, are made based on forecasting models, the error can result in massive losses. As mentioned earlier, this study focused on this and proceeded with the analysis.

Confusion matrices allow us to determine sensitivity and specificity. Sensitivity refers to the rate at which positive predictions are made correctly (TPR), and specificity refers to the rate at which negative predictions are made correctly (TNR). We determine that the closer the sensitivity and specificity at a particular threshold to 1, the better the classification performance of the model.

Table 1 shows a confusion matrix based on the optimal cut-off 0.105. The final model classified 14 of the 16 actual bankrupt or liquidated companies (BL) as bankrupt or liquidated companies. In addition, 117 out of 160 companies that are neither bankrupt nor liquidated (NBL) were correctly classified. That is, the sensitivity of the model was 87.5%, and the specificity was 73.125%. The closer both sensitivity and specificity are to 1, the better the model is judged, both of which are relatively superior. On the other hand, the accuracy representing the correct classification ratio of the model was 74.43%, and the misclassification rate was 25.57%. Among the companies classified as bankruptcy and liquidation, precision, which means the ratio of actual bankruptcy and liquidation companies, was 24.56%. The ratio of bankruptcy and non- liquidation companies to bankruptcy and liquidation was 26.875%. In other words, there were more cases in which the final model was judged to be bankrupt and liquidated than in the case. However, as noted earlier, it can be interpreted that the threshold is the result of determining that it is more dangerous to predict otherwise the entity to go bankrupt and liquidate.

		Predicted: BL	Predicted: NBL	
Actual	BL	14	2	16
	NBL	43	117	160
		57	119	176

Table 1: Confusion matrix with cut-off 0.105

5 Conclusion

Nine continuous variables and two categorical variables were used in the final model. The net amount of cash, cash equivalent, or assets calculated at fair value (aqpl1), and the total amount of ordinary shares distributed at the end of the year (glcea). The explanatory variables are aco, csho, cstk, optfvgr, spced, and stkcpa variables, while aqpl1, caps, and glcea are shown to fluctuate in the same direction as the probability of bankruptcy and liquidation of an entity. On the other hand, when other conditions are fixed, business activities are found to be more likely to go bankrupt and liquidate than those operating outside the United States (idbflag=D), and those currently at risk of bankruptcy or undergoing acquisitions (stalt=1).

The performance of the model is evaluated by leveraging test data to represent the confusion matrix with the ROC curve. We derive optimal thresholds and AUCs through the ROC curve. The optimal threshold was 0.1267, which means that the sensitivity and goodness of fit are the best when the criteria for classification as bankruptcy and liquidation are set at 0.1267.

On the other hand, the AUC, which means the area under the curve, was 0.857, which makes it possible to judge that the fitted model is excellent. Meanwhile, we construct a confusion matrix based on optimal thresholds. The sensitivity was 0.875, fit was 0.731, and accuracy was 0.744. On the other hand, the false positive ratio was 0.269, which was relatively high. This is because the analysis was conducted in the direction of lowering class II errors, judging that the prediction of bankruptcy would be worse than the opposite case, i.e., false positive case, to predict that an entity would not go bankrupt. Based on heresy extraction, this study lowered type 2 errors, which unfortunately increased type 1 errors.

Variables with opposite common sense and signs also existed in the model. This is a common phenomenon due to multicollinearity in regression analysis, but in this work, we diagnose the problem of multicollinearity. The limitations were determined to be attributed to the implementation of heresy extraction in the process of fitting the model and the selection of variables based on AIC. Therefore, we expect to produce results consistent with common sense by supplementing the limits of unbalanced data in other ways instead of heresy extraction, or proceeding with variable selection by setting a criterion other than AIC.

It was judged that diagnosing that a company that would go bankrupt would not go bankrupt would be a more serious error than the opposite. Therefore, this work has been able to reduce the corresponding error by arbitrarily adjusting the number of training data. However, the error of predicting that a company that would not actually go bankrupt, or type 1 error, would go bankrupt, had to be tolerated. Depending on the purpose of the study, sensitive errors will vary, and threshold adjustments will also be made differently. The proper balance between the two errors requires careful judgement by the researcher.

Based on current data and appropriate models, this study sought to predict which companies will go bankrupt and liquidate in the next three years. However, for the 2019 and 2020 data, observations of both variables included in the final model were often missing. Therefore, it was not possible to deal with the missing values in an appropriate way, and the analysis was not carried out to the stage of predicting bankruptcy and liquidation entities. If we proceed with the processing of missing values based on raw materials from 2000 to 2020, we expect to be able to predict bankruptcy and liquidation companies with current data.

Data for the year immediately before the bankruptcy and liquidation of the entity contains more signals than in any other year. In other words, if a company goes bankrupt or liquidated, the warning will remain intact in the previous year's data. Furthermore, Laitinen and Laitinen (2000) found that the model had the best explanatory power when the analysis was conducted based on data from the year just before bankruptcy. In addition, many prior studies have devised predictive models based on data from the previous year of bankruptcy. However, in this work, we tried to derive a mid- to long-term prediction model over three years, meaning to try a new direction rather than to follow the existing direction. As has already been proven, it would be helpful to use data from the year just before bankruptcy when trying to create a predictive model.

References

Erkki K. Laitinen and Teija Laitinen. Bankruptcy prediction: Application of the Taylor's expansion in logistic regression. *International review of financial analysis*, 9(4), 327–349, 2000.

Nirosh Kuruppu, Fawzi Laswad, and Peter Oyelere. The efficacy of liquidation and bankruptcy prediction models for assessing going concern. *Managerial auditing journal*, 2003.

Richard P. Hauser and David Booth. Predicting bankruptcy with robust logistic regression. *Journal of Data Science*, 9(4), 565–584, 2011.

Michelle J. White. The corporate bankruptcy decision. *Journal of Economic Perspectives*, 3(2), 129–151, 1989.

Gary King and Langche Zeng. Logistic regression in rare events data. *Political analysis*, 9(2), 137–163, 2001.

Maher Maalouf and Mohammad Siddiqi. Weighted logistic regression for large-scale imbalanced and rare events data. *Knowledge-Based Systems*, 59, 142–148, 2014.

Hyun Kang. The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64(5), 402–406, 2013.

Samprit Chatterjee and Ali S. Hadi. *Regression analysis by example*. John Wiley & Sons, 2015.

James H. Stock and Mark W. Watson. *Introduction to Econometrics*. Pearson, 2015.

David G. Kleinbaum and Mitchel Klein. *Logistic regression: A self-learning text*. Springer, 2010.

acctstd	auop	costat	dlc	ebit	gind	ivch	naics4	pidom	reajo	tfvce	txtubbegin
acdo	auopic	county_fips	dlcch	ebitda	glcea	ivncf	naics5	pifo	recch	tfvl	txtubend
aco	bkvlps	county_name	dldte	ein	glced	ivst	naics6	pnca	recco	tic	txtubposdec
acodo	BL	cshfd	dlrsn	emp	glceeps	ivstch	naicsh	pncad	recd	tlcf	txtubposinc
acominc	busdesc	cshi	dltis	epsfi	glcep	lat	ni	pncaeps	rect	tstk	txtubpospdec
acox	caps	csho	dlto	epsfx	gp	lco	niadj	pncwia	recta	tstkc	txtubpospinc
act	capx	cshpri	dltp	epspi	gsector	lcox	nopi	pncwip	rectr	tstkn	txtubsettle
add1	capxv	cshr	dltr	epspx	gsubind	lcoxdr	nopio	pnrsho	reuna	tstkp	txtubsoflimit
addzip	census_region	cshtr_c	dltt	esopct	gvkey	lct	np	ppegt	revt	txach	txtubtxtr
adjex_c	ceoso	cshtr_f	dm	esopdlt	ib	lifr	oancf	ppent	sale	txbco	txtubxintbs
adjex_f	ceq	cstk	dn	esopnr	ibadj	lifrp	oiadp	ppeveb	scf	txbcof	txtubxintis
ajex	ceql	cstkcv	do	esopt	ibc	lno	oibdp	prca	seq	txc	txw
ajp	ceqt	cstke	donr	esub	ibcom	lo	opeps	prcad	seqo	txdb	upd
aldo	cfoso	curcd	dp	esubc	ibmii	lol2	oprepsx	prcaeps	sic	txdba	wcap
am	ch	curncd	dpact	exchg	icapt	long	optca	prcc_c	sich	txdbca	weburl
ano	che	currtr	dpc	exre	idbflag	loxdr	optdr	prcc_f	siv	txdbcl	xacc
ao	chech	cusip	dpvieb	fatb	idit	lqpl1	optex	prch_c	spce	txdc	xad
aocidergl	ci	datadate	drc	fatc	incorp	lse	optexd	prch_f	spced	txdfed	xi
aociother	cibegni	dc	drlt	fate	intan	lt	optfvgr	prcl_c	spceeps	txdfo	xido
aocipen	cicurr	dclo	ds	fatl	intano	lul3	optgr	prcl_f	spcindcd	txdi	xidoc
aocisecgl	cidergl	dcom	dt	fatn	intc	mib	optlife	priusa	spcseccd	txditc	xint
aodo	cik	dcpstk	dudd	fato	intpn	mibn	optosby	prsho	spcsrc	txds	xintopt
aol2	cimii	dcs	dv	fatp	invch	mibt	optosey	prstkc	spi	txfed	xopr
aoloch	ciother	dcvsr	dvc	fax	invfg	mii	optprcby	pstk	sppe	txfo	xpp
aox	cipen	dcvsub	dvp	fca	invo	mkvalt	optprcca	pstkc	sppiv	txndb	xpr
ap	cisecgl	dcvt	dvpa	fdate	invrm	mrc1	optprcex	pstkl	src	txndba	xrd
apalch	citotal	dd	dvpsp_c	fiao	invt	mrc2	optprcey	pstkn	sstk	txndbl	xrdp
apdedate	city	dd1	dvpsp_f	fic	invwip	mrc3	optprcgr	pstkr	stalt	txndbr	xrent
aqc	cld2	dd2	dvpsx_c	fincf	ipodate	mrc4	optprcwa	pstkrv	state	txo	xsga
aqi	cld3	dd3	dvpsx_f	fopo	ismod	mrc5	optrfr	rdip	state_name	txp	
aqpl1	cld4	dd4	dvt	fopox	itcb	mrct	optvol	rdipa	stkco	txpd	
aqs	cld5	dd5	dxd2	fyr	itci	mrcta	pdate	rdipd	stkcpa	txr	
at	cogs	dfs	dxd3	fyrc	ivaco	msa	pddur	rdipeps	stko	txs	
au	conm	diladj	dxd4	gdwl	ivaeq	naics2	phone	re	teq	txt	
aul3	conml	dilavx	dxd5	ggroup	ivao	naics3	pi	rea	tfva	txtubadjust	

Table 2: Variable names

Sector	N	Description
11	18	Agriculture, Forestry, Fishing and Hunting
21	426	Mining, Quarrying, and Oil and Gas Extraction
22	248	Utilities
23	78	Construction
31–33	2193	Manufacturing
42	169	Wholesale Trade
44–45	235	Retail Trade
48–49	148	Transportation and Warehousing
51	652	Information
52	2122	Finance and Insurance
53	341	Real Estate and Rental and Leasing
54	233	Professional, Scientific, and Technical Services
55	0	Management of Companies and Enterprises
56	111	Administrative and Support and Waste Management and Remediation Services
61	26	Educational Services
62	117	Health Care and Social Assistance
71	43	Arts, Entertainment, and Recreation
72	106	Accommodation and Food Services
81	17	Other Services (except Public Administration)
92	0	Public Administration
99	105	Nonclassifiable

Table 3: Structure of 2017 NAICS

Monster Beverage Corp		Kellogg Co	
31	Manufacturing	31	Manufacturing
312	Beverage and Tobacco Product Manufacturing	311	Food Manufacturing
3121	Beverage Manufacturing	3112	Grain and Oilseed Milling
31211	Soft Drink and Ice Manufacturing	31123	Breakfast Cereal Manufacturing
312111	Soft Drink Manufacturing	311230	Breakfast Cereal Manufacturing
Coca Cola Consolidated Inc		**Nike Inc**	
31	Manufacturing	31	Manufacturing
312	Beverage and Tobacco Product Manufacturing	316	Leather and Allied Product Manufacturing
3121	Beverage Manufacturing	3162	Footwear Manufacturing
31211	Soft Drink and Ice Manufacturing	31621	Footwear Manufacturing
312111	Soft Drink Manufacturing	316210	Footwear Manufacturing

Table 4: Examples of NAICS

Removed variables by t-test (105/349)											
adjex_c	aoloch	currtr	do	epspi	glceeps	lno	oprepsx	prcaeps	pstkr	spi	txo
adjex_f	apalch	dcom	donr	epspx	invch	lol2	optca	prcc_c	pstkrv	sstk	txtubadjust
ajex	aqi	dcpstk	dvp	esopdlt	invt	long	optlife	prcc_f	rdip	tfva	txtubxintis
ajp	che	dcvsr	dvpsp_c	esopnr	ivaco	lqpl1	optrfr	prch_c	rdipa	tfvl	xi
ano	cicurr	dcvt	dvpsp_f	exre	ivao	mib	pncwia	prch_f	rea	tstkp	xido
aocidergl	cidergl	diladj	dvpsx_c	fatl	ivch	msa	pncwip	prcl_c	recco	txach	xintopt
aociother	ciother	dlc	dvpsx_f	fca	ivncf	nopi	pnrsho	prcl_f	seqo	txdfed	
aocisecgl	cshr	dltis	epsfi	fiao	ivst	np	prca	pstkc	siv	txdi	
aol2	cstkcv	dltr	epsfx	glced	lat	opeps	prcad	pstkl	spce	txndbr	
Removed variables by VIF (154/244)											
acodo	ceq	cshfd	dlcch	dxd3	ibc	loxdr	ni	pifo	reuna	txdba	txtubpospinc
acominc	ceql	cshi	dlto	dxd4	ibcom	lse	niadj	pncaeps	revt	txditc	txtubtxtr
acox	ceqt	cshpri	dltt	dxd5	ibmii	lt	oancf	ppegt	sale	txds	txtubxintbs
act	ch	cshtr_f	dn	ebit	icapt	lul3	oiadp	ppent	seq	txfo	xacc
am	ci	cstke	dp	ebitda	intan	mibn	oibdp	ppeveb	spceeps	txndb	xint
ao	cibegni	dd	dpact	esub	intpn	mibt	optosby	pstk	sppiv	txndba	xopr
aodo	cimii	dd1	dpc	fopo	lco	mkvalt	optosey	pstkn	stkco	txndbl	xpr
aox	cisecgl	dd2	dpvieb	fopox	lcox	mrc1	optprcby	rdipd	teq	txpd	xrd
ap	citotal	dd3	ds	gdwl	lcoxdr	mrc2	optprcca	re	tlcf	txt	xrdp
at	cld2	dd4	dt	glcep	lct	mrc3	optprcex	reajo	tstk	txtubbegin	xrent
aul3	cld4	dd5	dv	gp	lifr	mrc4	optprcey	recd	tstkc	txtubend	xsga
capx	cld5	dfs	dvt	ib	lifrp	mrc5	optprcgr	rect	txc	txtubposinc	
capxv	cogs	dilavx	dxd2	ibadj	lo	mrct	pi	rectr	txdb	txtubpospdec	
Preselected variables (90)											
acdo	caps	dclo	dvc	fate	intc	ivstch	optgr	rdipeps	txbco	txr	xidoc
aco	chech	dcs	dvpa	fatn	invfg	mii	optprcwa	recch	txbcof	txs	xpp
aldo	cipen	dcvsub	emp	fato	invo	mrcta	optvol	recta	txdbca	txtubposdec	
aocipen	cld3	dltp	esopct	fatp	invrm	nopio	pidom	spced	txdbcl	txtubsettle	
aqc	csho	dm	esopt	fincf	invwip	optdr	pnca	sppe	txdc	txtubsoflimit	
aqpl1	cshtr_c	drc	esubc	glcea	itcb	optex	pncad	stkcpa	txdfo	txw	
aqs	cstk	drlt	fatb	idit	itci	optexd	prsho	tfvce	txfed	wcap	
bkvlps	dc	dudd	fatc	intano	ivaeq	optfvgr	prstkc	tstkn	txp	xad	

Table 5: Preselected variables by t-test and VIF

year	All deletion	Bankruptcy	Liquidation	B + L
2011	241	1	16	17
2012	363	6	29	35
2013	348	8	38	46
2014	369	3	47	50
2015	348	8	36	44
2016	356	10	31	41
2017	273	6	1	7
2018	243	8	1	9
2019	266	16	0	16
2020	99	4	0	4

Table 6: Number of deleted companies

	Obs	Mean	Std.Dev	Min	25%	75%	Max
aco	7,380	0.000	1.000	−0.150	−0.149	−0.107	42.624
aqpl1	7,380	0.000	1.000	−0.067	−0.067	−0.056	55.435
caps	7,380	0.000	1.000	−0.389	−0.209	−0.029	34.987
csho	7,380	0.000	1.000	−0.179	−0.163	−0.053	44.172
cstk	7,380	0.000	1.000	−0.138	−0.138	−0.119	35.533
glcea	7,380	0.000	1.000	−6.029	−0.134	−0.043	71.202
optfvgr	7,380	0.000	1.000	−0.080	−0.065	−0.008	66.953
spced	7,380	0.000	1.000	−83.980	−0.016	−0.006	5.143
stkcpa	7,380	0.000	1.000	−1.930	−0.235	−0.005	42.503

Table 7: Summary statistics of continuous variables

		stalt: 1	stalt: 0	
BL	1	4	72	76
BL	0	16	7288	7304
		20	7360	7380

		idbflag: B	idbflag: D	
BL	1	0	76	76
BL	0	445	6859	7304
		445	6935	7380

		idbflag: B	idbflag: D	
stalt	1	1	19	20
stalt	0	444	6916	7360
		445	6935	7380

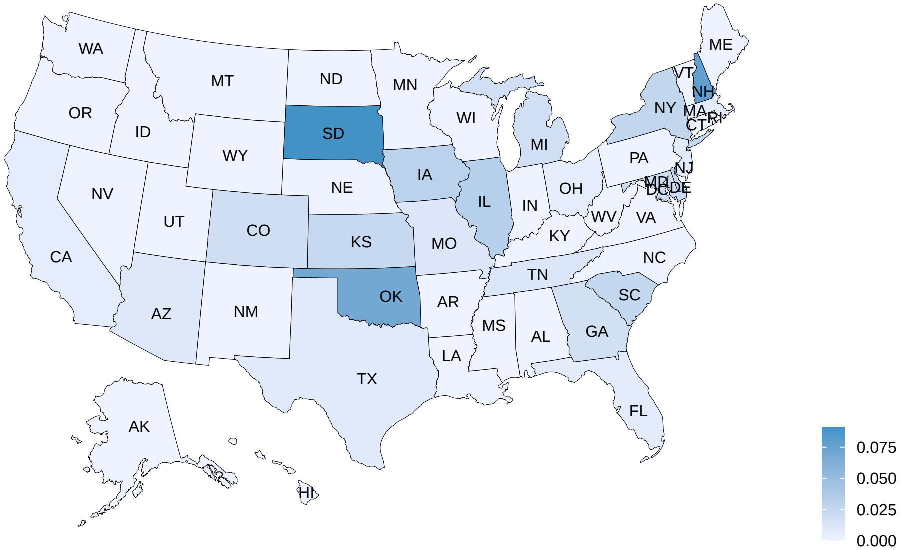
Table 8: Contingency tables

	Pearson's chi-squared test			Fisher's exact test			
	χ^2	df	p-value	odds ratio	95% CI		p-value
BL – stalt	53.376	1	2.755×10^{-13}	25.238	5.994	80.874	4.441×10^{-5}
BL – idbflag	3.911	1	0.048	∞	1.299	∞	0.014
stalt – idbflag	2.648×10^{-29}	1	1	1.219	0.193	50.795	1

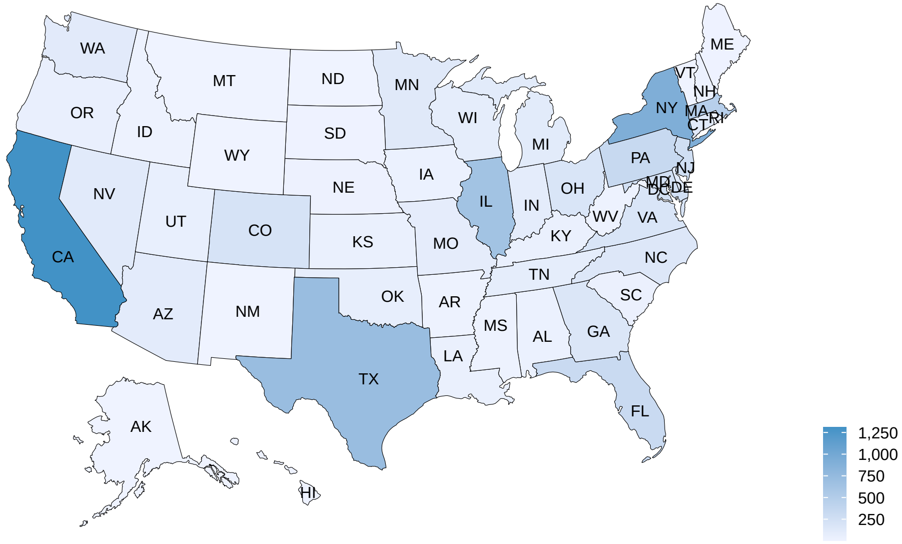
Table 9: Result of tests

	Maximum Likelihood LR		Weighted LR	
	AIC	BIC	AIC	BIC
aco	-7.190	-7.303	-5.595	-15.430^{*}
	(6.036)	(7.196)	(6.785)	(9.297)
aqpl1	8.901^{**}	6.869^{*}	8.451	
	(3.520)	(3.762)	(5.564)	
caps	2.117^{**}			
	(0.961)			
csho	-7.575^{**}	-7.104^{**}	-4.672	
	(3.088)	(3.109)	(2.876)	
cstk	-13.510		-15.164	
	(11.405)		(12.998)	
glcea	1.561^{**}			
	(0.675)			
idbflagD	197.005	137.557	185.161	
	(970.673)	(658.688)	(1,554.751)	
optfvgr	-19.957^{***}	-21.830^{***}	-22.422^{***}	-25.100^{***}
	(6.620)	(6.349)	(7.920)	(7.745)
spced	-2.420			
	(1.488)			
stalt1	2.663^{**}			
	(1.242)			
stkcpa	-3.308^{**}	-3.291^{**}	-2.900^{*}	
	(1.449)	(1.420)	(1.559)	
Constant	-203.596	-142.864	-189.460	-3.235^{**}
	(970.743)	(658.668)	(1,554.795)	(1.260)
Observations	660	660	660	660
Log Likelihood	-155.707	-163.898	-30.190	-33.437
Akaike Inf. Crit.	335.414	341.796	76.379	72.873
Note:			$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$	

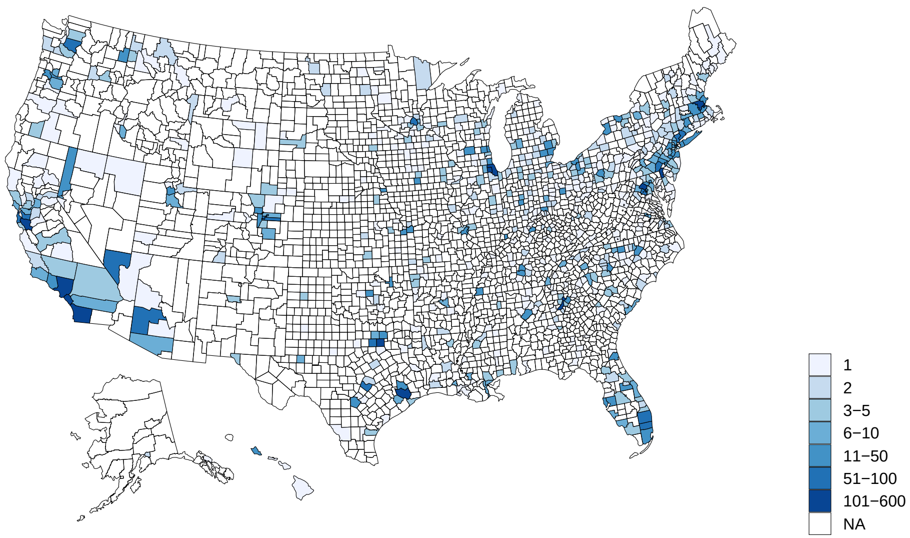
Table 10: Results of logistic regressions



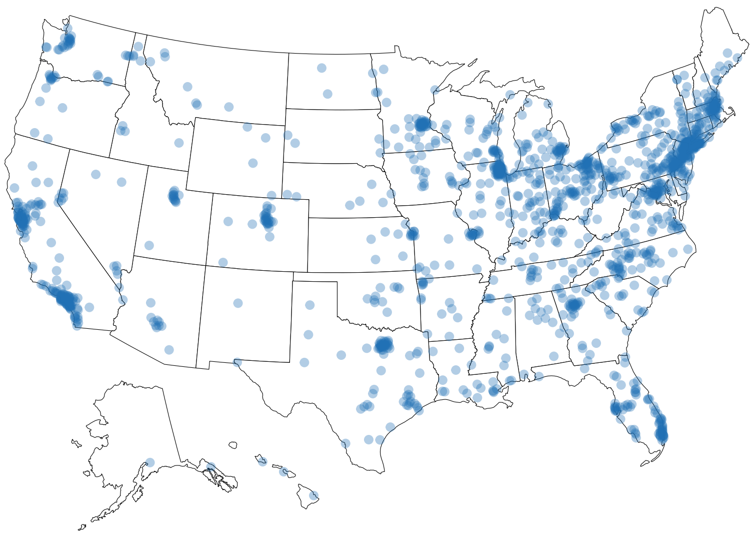
(a) Proportion of bankruptcy and liquidation by states



(b) Number of companies by states

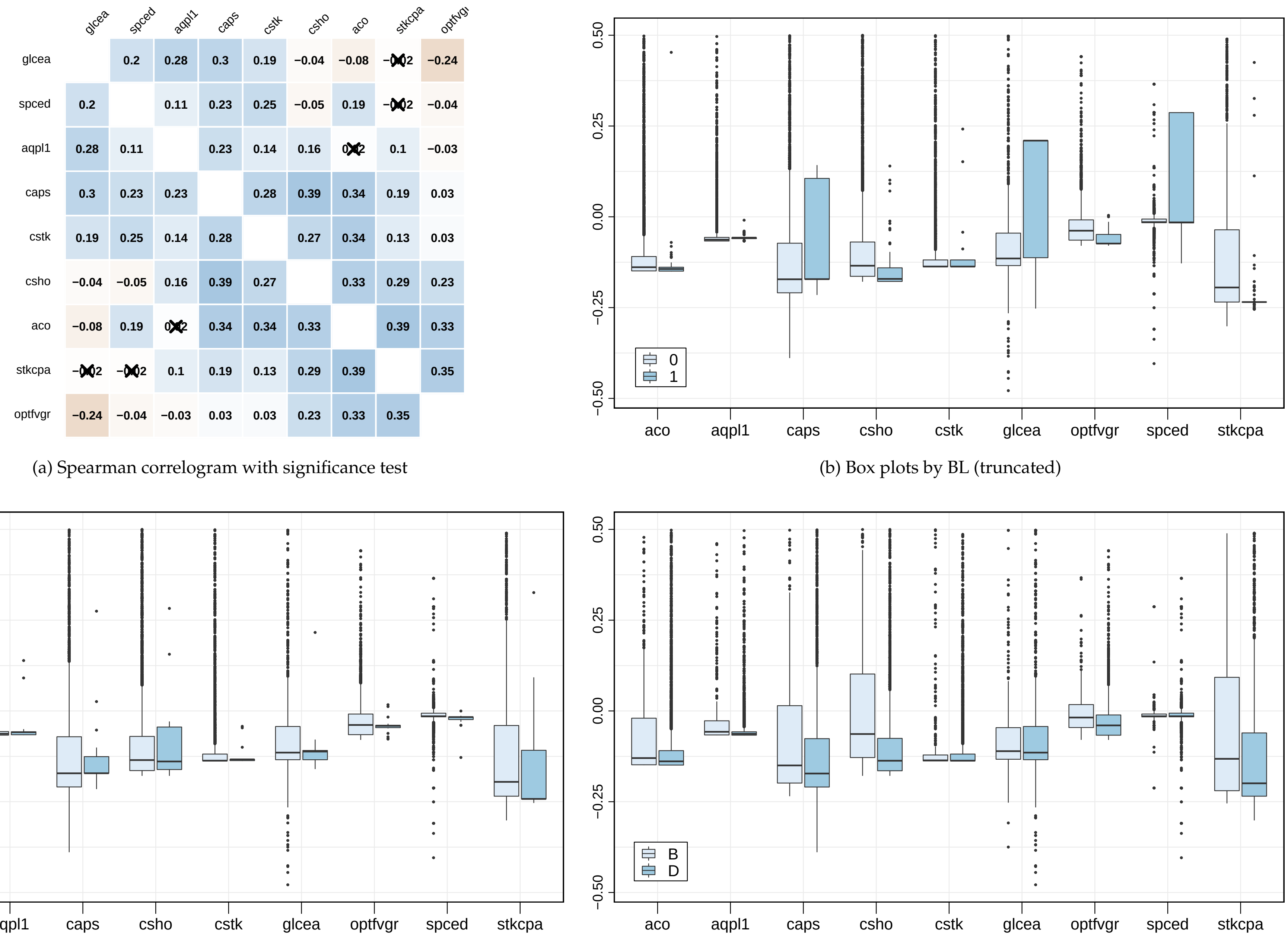


(c) Number of companies by counties



(d) Plotting cities where a company is located

Figure 2: US map by states, counties, and cities



(a) Spearman correlogram with significance test

(b) Box plots by BL (truncated)

(c) Box plots by stalt (truncated)

(d) Box plots by idbflag (truncated)

Figure 3: Correlogram and box plots