Bankruptcy and Liquidation Prediction with Weighted Likelihood and Bayesian Approach

Junwoo Yang* Jaeseon Lee†

December 8, 2020

Abstract

In this study, we examine three methods of logistic regression to predict bankruptcy and liquidation of North America companies: MLE, weighted likelihood, and Bayesian inference (Markov chain Monte Carlo). 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 into the response variable could alleviate the imbalanced data problem. Since there are more than 400 variables, we take preselection steps by t-test and VIF before stepwise selection by AIC and BIC. We forecast companies that will go bankrupt or liquidate within three years based on the certain fiscal year.

Keywords: bankruptcy and liquidation prediction, imbalanced data problem, logistic regression, weighted likelihood, Bayesian inference, Markov chain Monte Carlo, forecasting

1 Introduction

The main problems in constructing bankruptcy prediction models are the choice of the independent variables and the functional form between these variables. There are some major bankruptcy theories: the gambler's ruin model, perfect-access model (Scott, 1981). However theoretical approaches to variable selection are rare and too simplified. In most previous bankruptcy prediction studies, the independent variables are empirically chosen.

Most bankruptcy prediction research use linear discriminant analysis or logistic regression. Kuruppu, Laswad, and Oyelere (2003) used multiple discriminant analysis to develop the liquidation prediction model on New Zealand sample. Of the 63 ratios, 12 were selected through stepwise methodology. Linear model, however, requires an assumption that independent variables are normally distributed. Laitinen and Laitinen (2000) tested whether Taylor's series expansion can be used to solve the problems associated with the functional form of bankruptcy prediction models, and showed improvement of classification accuracy in some of their logistic regression models. However, they used three financial ratio variables: cash to total assets, cash flow to total assets, and shareholder's equity to total assets, relying on simple theoretical analysis of the isolation concept for variable selection. Hauser and Booth (2011) use robust logistic regression with the Bianco and Yohai estimator which improved both the classification

This paper is a project report of Math Capstone PBL (Data Analysis) in fall 2020.

*Department of Finance, Hanyang University

†Department of Economics and Finance, Hanyang University

of bankrupt firms in the training set and the prediction of bankrupt firms in the testing set. In an out of sample test, the BY robust logistic regression correctly predicted bankruptcy for Lehman Brothers while maximum likelihood logistic regression never. However, this study also designated five financial ratios as variables without empirically selecting.[1]

In this study, 414 variables containing both financial and non-financial are included in the candidates of variables in the final model. We empirically select variables describing bankruptcy predictions through several steps. To avoid the normality assumption of independent variables, this study also uses logistic regression. First, we evaluate the basic maximum likelihood logistic regression. However, this method is subject to the problems caused by imbalanced data. As a solution, we use adjusted likelihood function, called the weighted likelihood. A total of four models are derived by considering both AIC and BIC as statistics scoring each model. For these four model specifications, we re-estimate the coefficients with the Bayesian approach. Therefore, a total of eight models are constructed. We measure the performance of these models and use the simplest model to out of sample test.

The remainder of this paper is organized as follows. Section 2 presents how to handle missing values and preselect continuous variables. Section 3 derives mathematical notation for three logistic regression models and fits training data to eight final models. In the process, we apply two variable selection methods: AIC and BIC. Section 4 measures performance and forecasts future bankruptcies. Section 5 concludes by suggesting some directions for future research.

2 EDA

2.1 Data

The annual fundamental data used in this paper are from the Compustat Capital IQ database of WRDS. The data include a total of 981 variables and we merge these with US counties data in SimpleMaps[2] database by columns of state and city for Figure 1c. All variables are classified into following 7 categories: identifying information, company descriptor, balance sheet items, income statement items, cash flow items, miscellaneous items, supplemental data items. We remove 26 variables with the same value for all observation and 552 variables with missing values greater than 80%. There are left 414 variables including 349 continuous variables. These are listed in Table 1.

The observations are companies traded on the NYSE, AMEX, NASDAQ, TSX, and NYSE Arca from January 2000 through November 2020. In 2010, 7,978 of the 11,036 companies were headquartered in the United States and 1,887 in Canada. China, the United Kingdom, Israel, Bermuda and Hong Kong followed. However, none of companies headquartered in these

[1]Hauser and Booth (2011) used WCTA, RETA, EBITTA, MEDDEBT, and SALETA as independent variables where WCTA: working capital/total assets as a measure of the net liquid assets of the firm to total capitalization; RETA: retained earnings/total assets as a measure of cumulative profitability; EBITTA: earnings before interest and taxes/total assets as a measure of the true productivity of the firm's assets; MEDDEBT: market value of equity/book value of total debt as a measure of how much the firm's assets can decline in value before the liabilities exceed the assets and the firm becomes insolvent; SLAETA: sales/total assets as a measure of the sales generating ability of the firm's assets (for prediction capabilities of these ratios, see Altman, 1968; Boritz and Kennedy, 1995; Odom and Sharda, 1990; Zhang et al, 1999; Lee et al, 2005).

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

contries went bankrupt, and only five Singapore-headquartered companies were liquidated between 2011 and 2013. Thus, we consider only 9,865 companies headquartered in the U.S. and Canada, removing the rest from the population. Now the response variable is defined as

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

Note that it is not set for two consecutive years, but within three years based on a certain fiscal year. Table 4a shows the number of deleted companies due to bankruptcy and liquidation, respectively. Although bankruptcy and liquidation have been combined to define a response variable, the event is still rare. This is why 1,171 companies were deleted earlier.

2.2 Imputation

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 is assigned a 6-digit NAICS code. A company's NAICS code classifies the company on a production and/or process-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

See Table 2 for a detailed structure of NAICS. Meanwhile, there are other industry classification systems such as Standard Industry Classification (SIC) and Global Industry Classification Standard (GICS). However, SIC codes were not able to keep up with current industries, and as a result of the development of NAICS, more than 350 new industries were recognized.

We replaced each missing values of continuous variables with the average value of companies with the same 6-digit NAICS code.[3] If all companies with the same 6-digit have missing data, it expands to those with the same 5-digit code. Repeating this process to 2-digit code replaces most missing values. Oddly enough, the NAICS variable also had missing values for 560 samples, whose SIC codes were either 6722 or 6726. Only these missing values were replaced by the average of companies with the same 4-digit SIC code. We replaced all missing values existing in 349 continuous variables with appropriate averages.

2.3 Preselection by t-test and VIF

Considering all 414 variables for stepwise selection takes too much time. Therefore, we remove continuous variables in advance that do not show significant differences in mean between the two groups classified by the response variable. First, for each continuous variable, an F test is

[3] White et al (2015) used the average 6-digit NAICS for imputation of inputs.

performed to determine whether the variance ratio of the two groups is 1. If we accept the null hypothesis that the variance ratio is 1, we use the student's t-test, and if we reject it, we use the Welch's t-test. Both t-test determine if the means of two groups are significantly different from each other. Next, to reduce the multicollinearity between variables, we remove the variables with the highest VIF value one by one until the maximum VIF value does not exceed 10.

Table 3 shows the results of the preselection steps. As a result of t-testing 349 continuous variables, 105 variables did not show significant differences in means between two groups classified by the response variable. The remaining 244 variables were removed one by one until the highest VIF value is less than 10. As a result, 90 variables survived and along with the rest of the categorical variables became candidates for the variables used in the final model.

2.4 Summary statistics and correlation analysis

In Section 3.4, the MLLR-AIC model contains the largest number of variables which are described in Table 4b, leaving out explanations for other variables. Table 4c reports summary statistics of continuous variables of those. Because continuous variables were standardized to compare the size of the estimated coefficients, the mean is equal to 0 and the variance equal to 1. Meanwhile, Figure 1 shows various maps of United States. By state, Figure 1a represents the proportion of event (bankruptcy and liquidation), and Figure 1b represents the number of samples. Figure 1d refers to the cities in which a company's headquarter is located. The candidates for variables of final models include state, county, city, and census region of U.S.

The correlation analysis is also performed within variables contained in MLLR-AIC. There are three categorical variables including the response variable. Table 5 shows contingency tables and results of Pearson's chi-squared test and Fisher's exact test among them. While stalt and idbflag are associated with response variables respectively, 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 plots in Figure 2. 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 correlation coefficients between continuous variables are also mostly low in Figure 2a.

3 Logistic Regressions

3.1 Subsampling imbalanced data

As of 2010, there are 76 companies that would go bankrupt or be 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 insolvent companies (bankrupcy or liquidation) and 600 solvent companies at random. It gives more weight to insolvent 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. Here we care more about diagnosing a company that are going bankrupt will not go bankrupt than vice versa. 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 MLE

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 $\beta = (\beta_0, \beta_1, \cdots, \beta_k)^\top$. 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

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

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 likelihood

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

$$L_W(\beta|X, Y) = \prod_{i=1}^{n} (p_i)^{w_1 y_i}(1 - p_i)^{w_0(1-y_i)},$$

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 $w_i = w_1 y_i + w_0(1-y_i)$.

3.4 Stepwise 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 each model. Then the statistics of AIC and BIC are as follows:

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

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.

We chose the final model of maximum likelihood logistic regression (MLLR) and weighted logistic regression (WLR) for each AIC and BIC: MLLR-AIC, MLLR-BIC, WLR-AIC, and WLR-BIC. Table 6 reports fittig results. For example, the final MLLR-AIC model is

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

which contains the largest number of variables. The likelihood ratio test for this 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 response variable.

3.5 Bayesian approach

Bayesian inference estimates parameters using probability models for the observed data and the parameters of interest. First, we determine the prior distribution of parameters, and estimate the parameters using a posterior distribution combining the sampling distribution and prior distribution. Let $\beta \in \mathbb{R}^{k+1}$ be a parameters vector, and then the likelihood function for the observed data Y is

$$L(Y|\beta) = \prod_{i=1}^{n} (p_i)^{y_i}(1-p_i)^{1-y_i}.$$

The parameters are assumed to follow non-informative prior distributions to avoid bias and put more weight on sample mean rather than prior mean. That is, β follows the standard multivariate normal distribution. Now, with the assumption that the parameters are mutually independent, the prior and posterior distributions for β are defined as

$$p(\beta) = \prod_{i=1}^{n} p_i(\beta_i), \quad p(\beta|Y) = L(Y|\beta)p(\beta)$$

respectively.

We use the Markov chain Monte Carlo (MCMC) method and WinBUGS[4] software to estimate the parameters. The simulation for each parameter has two chains with different initial values and 15,000 iterations. After burning first 1,000 samples, 2,800 samples for each chain, in a total of 5,600, were used for posterior samples by extracting every fifth value (see Figure 3c for thinning setting). Table 7 reports the posterior distributions of all parameters of the four model specifications obtained in Section 3.4. Figure 3a shows the density plots of parameters in MLLR-AIC model specification with 95% intervals (colored). If the 95% interval contains zero, the parameter is considered insignificant. Meanwhile, we also checked the convergence by the Gelman-Rubin statistic and monitoring trace plot in Figure 3b. All parameters of the four Bayesian LR models converge.

4 Performance and forecasting

The performance reported in Table 8 is measured based on test data consists of 16 companies with $y_i = 1$ and 160 companies with $y_i = 0$. There are AUC, sensitivity, specificity, and Youden's J statistic.[5] We give the same weight to sensitivity and specificity. That is, the optimal cut-off is selected to maximize

$$J = \text{sensitivity} + \text{specificity} - 1$$

which is called Youden's J statistic. This is where the ROC curve in Figure 4 meets a 45 degree line at one point, which is X-marked. Table 8b shows confusion matrix of frequentist MLLR-AIC model with optimal cut-off 0.105.

We construct 8 models according to variable selection and coefficient estimation methods. The more variables used in the model, the more difficult it is to predict due to missing values. Thus, the simplest model, Bayesian WLR-BIC, was used to forecast future bankruptcy or liquidation. Table 9 reports forecasting results. Only 7 of the top 50 predicted companies are headquartered in the United States and the rest in Canada. Considering that the optimal cut-off value we used in Bayesian WLR-BIC model is around 0.1, it is hard to say that there is a company with a high probability of bankruptcy or liquidation.

5 Conclusion

We used three methods of logistic regression and two statistics for scoring and selecting a model. Considering that there are so many variables, we relied entirely on statistical tests, not

[4] http://www.mrc-bsu.cam.ac.uk/software/bugs

[5] AUC: area under ROC curve; sensitivity: True Positive/Actual Positive; specificity: True Negative/Actual Negative

theoretical approaches for variable selection. There are four model specification depending on whether MLLR or WLR, and whether AIC or BIC. Adding to the Bayesian approach, we constructed a total of eight models. The ROC curves were compared in each case as frequentist versus Bayesian.

Most of the models' performance was similar. The frequentist MLLR-AIC model had the highest Youden's J statistic with 78% sensitivity and 76% specificity. However, considering its complexity of 11 variables, the Bayesian WLR-BIC model is attractive in that it shows 65% sensitivity and 85% specificity with only two variables.

References

James Scott. The probability of bankruptcy: A comparison of empirical predictions and theoretical models. *Journal of banking & finance*, 5(3), 317–344, 1981.

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.

Altman, E. L. Financial ratios, discriminate analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23, 589–609, 1968.

Boritz, J. E. and Kennedy, D. B. Effectiveness of neural network types for prediction of business failure. *Expert Systems with Applications* 9, 503–512, 1995.

Odom, M. and Sharda, R. A neural network model for bankruptcy prediction. *Proceedings of the IEEE International Conference on Neural Networks* 2, 163–168, 1990.

Zhang, G., Hu, M. Y., Patuwo, B. E. and Indro, D. C. Artificial neural networks in bankruptcy prediction: general framework and cross-validation analysis. *European Journal of Operational Research* 116, 16–32, 1999.

Lee, K., Booth, D. and Alam, P. A comparison of supervised and unsupervised neural networks in predicting bankruptcy in Korean firms. *Expert Systems with Applications* 29, 1–16, 2005.

White, Reiter, and Petrin. Imputation in US manufacturing data and implications for within-industry productivity dispersion. *Proceedings of the Federal Committee on Statistical Methodology Research Conference*, 2015.

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.

D. Kang and J. Choi. Bayesian zero-inflated spatio-temporal modeling of scrub typhus data in Korea, 2010–2014. *Geospatial health*, 13(2), 2018.

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.

Table 1: Variable names

This table shows the names of the 414 variables in Compustat Capital IQ database of WRDS. Variables with missing values greater than 80% or all equal were removed. We derive columns of first 2- to 5-digit code from the 6-digit NAICS code, which represents a higher classification. A total of 414 variables became candidates for the variables used in the model. The variables in the colored cells are those included in the final model selected by the maximum likelihood logistic regression and AIC. Variables selected for other models do not deviate from the range of these variables. Because there are so many variables, variable descriptions and correlation analysis were performed within these variables.

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: Structure of 2017 NAICS

This table shows the structure of NAICS. The first 2-digit of NAICS code represents a sector. (b) shows the NAICS code of companies belonging to the manufacturing sector. Monster Beverage has the same 6-digit code with Coca Cola. NAICS Industry and National sometimes have the same description like Kellogg and Nike. We replaced each missing value with the average value of companies with the same 6-digit NAICS code. If all companies in the same National are missing a value for a variable, they move on to a 5-digit step. This process is carried out to the sector average, which is a 2-digit code.

(a) Sectors of NAICS

Sector	Obs	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	unclassifiable

(b) Examples of 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 3: Preselection by t-test and VIF

This table shows the results of the preselection steps of variables. As a result of t-testing 349 continuous variables, 105 variables did not show significant differences in means. The remaining 244 variables were removed one by one until the largest VIF value was less than 10. 90 surviving continuous variables and all categorical variables are candidates for the variables to be used in the final model.

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 4: Variable discription and summary statistics

(a) motivated us to include liquidation in the response. Considered that population size of 7380, even if liquidation is included, the response variable is still a rare event. There were mergers and acquisitions for other reasons for the deletion, but it was not considered as a response variable because it did not necessarily mean a dangerous financial state. (b) shows the description of variables which was selected by MLLR and AIC. 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 in (c).

(a) Number of deleted companies

year	All deletion	Bankruptcy	Liquidation	BL
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

(b) Variable description

aco	Current Assets Other Total
aqpl1	Assets Level1 (Quoted Prices)
caps	Capital Surplus/Share Premium Reserve
csho	Common Shares Outstanding
cstk	Common/Ordinary Stock (Capital)
glcea	Gain/Loss on Sale (Core Earnings Adjusted) After-tax
idbflag	International, Domestic, Both Indicator
optfvgr	Fair Value of Options Granted
spced	S&P Core Earnings EPS Diluted
stalt	Status Alert
stkcpa	After-tax stock compensation

(c) Summary statistics of continuous variables

	Obs	Mean	SD	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 5: Correlation analysis among categorical variables

This table shows the results of correlation analysis among three categorical variables: BL (response variable), stalt, and idbflag. While stalt and idbflag are associated with response variables respectively, stalt and idbflag are independent of each other. In terms of multicollinearity, variable selection is considered appropriate.

(a) Contingency tables

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

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

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

(b) The significance of the difference between the two proportions

	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 6: Results of logistic regressions

This table reports the results of logistic regression under the frequentist framework. There are four model specification depending on whether MLLR or WLR, and whether AIC or BIC. Non-significant variables were also included in three of four models because variables were selected based on AIC and BIC rather than parameter significance. While MLLR-AIC contains the largest number of variables, WLR-BIC contains only two variables and all parameters are shown to be significant.

	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)	
intercept	-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 7: Posterior distributions

This table reports posterior distributions of parameters in four model specifications obtained in Section 3.4. The simulation for each parameter has two chains with different initial values and 15,000 iterations. After burning first 1,000 samples, 2,800 samples for each chain, in a total of 5,600, were used for posterior samples by extracting every fifth value.

	Mean	SD	2.5%	25%	50%	75%	97.5%	Rhat	N.eff
Panel A: MLLR-AIC									
aco	−1.066	0.814	−2.713	−1.603	−1.040	−0.504	0.470	1.001	5,600
aqpl1	0.010	0.518	−1.228	−0.274	0.097	0.383	0.784	1.001	5,600
caps	−0.008	0.598	−1.283	−0.386	0.033	0.420	1.050	1.001	4,400
csho	−1.391	0.798	−3.048	−1.926	−1.360	−0.849	0.130	1.001	5,600
cstk	−0.988	0.625	−2.381	−1.381	−0.935	−0.537	0.073	1.001	5,600
glcea	1.046	0.418	0.184	0.780	1.055	1.329	1.838	1.002	1,900
idbflagD	−0.947	0.760	−2.511	−1.454	−0.900	−0.417	0.491	1.001	5,600
optfvgr	−1.146	0.923	−3.008	−1.743	−1.139	−0.514	0.648	1.003	820
spced	0.568	0.781	−1.015	0.039	0.577	1.096	2.073	1.001	2,600
stalt1	1.122	0.802	−0.475	0.591	1.128	1.666	2.677	1.001	4,000
stkcpa	−1.717	0.658	−3.090	−2.146	−1.697	−1.257	−0.493	1.001	5,600
intercept	−2.890	0.212	−3.329	−3.031	−2.884	−2.745	−2.493	1.001	2,800
deviance	359.627	5.755	349.697	355.600	359.200	363.200	372	1.002	1,700
Panel B: MLLR-BIC									
aco	−1.038	0.792	−2.659	−1.561	−0.995	−0.486	0.403	1.002	2,300
aqpl1	0.120	0.538	−1.173	−0.185	0.220	0.508	0.896	1.001	5,600
csho	−1.401	0.813	−3.029	−1.952	−1.386	−0.845	0.150	1.001	5,600
idbflagD	−1.003	0.746	−2.547	−1.500	−0.976	−0.484	0.339	1.001	2,900
optfvgr	−1.327	0.894	−3.125	−1.930	−1.304	−0.721	0.396	1.001	5,100
stkcpa	−1.821	0.656	−3.156	−2.252	−1.796	−1.360	−0.606	1.001	5,600
intercept	−2.756	0.199	−3.150	−2.889	−2.752	−2.620	−2.386	1.002	2,000
deviance	371.046	5.244	361.797	367.300	370.700	374.325	382.200	1.001	5,600
Panel C: WLR-AIC									
aco	−0.985	0.791	−2.629	−1.502	−0.952	−0.424	0.432	1.001	5,600
aqpl1	0.120	0.537	−1.137	−0.185	0.225	0.512	0.904	1.005	1,600
csho	−1.339	0.800	−2.973	−1.868	−1.310	−0.786	0.155	1.001	5,600
cstk	−0.829	0.674	−2.302	−1.257	−0.756	−0.337	0.296	1.001	5,600
idbflagD	−0.998	0.755	−2.565	−1.488	−0.961	−0.472	0.401	1.001	5,600
optfvgr	−1.269	0.898	−3.060	−1.870	−1.257	−0.657	0.460	1.001	5,600
stkcpa	−1.811	0.657	−3.128	−2.242	−1.792	−1.355	−0.560	1.001	5,600
intercept	−2.826	0.207	−3.246	−2.960	−2.820	−2.682	−2.438	1.001	5,600
deviance	369.900	5.192	360.900	366.200	369.600	373.100	381.100	1.001	5,600
Panel D: WLR-BIC									
aco	−1.611	0.750	−3.193	−2.103	−1.581	−1.079	−0.258	1.001	5,600
optfvgr	−1.510	0.894	−3.296	−2.102	−1.496	−0.903	0.214	1.001	5,600
intercept	−2.467	0.155	−2.777	−2.571	−2.463	−2.362	−2.166	1.001	5,600
deviance	389.530	4.149	382.100	386.700	389.300	392.200	398.302	1.001	4,400

Table 8: Performance

(a) reports the performance of eight models. AUC is area under the ROC curve in Figure 4. The optimal cut-off is selected to maximize Youden's J statistic. Most of the models' performance was similar. The frequentist MLLR-AIC model had the highest Youden's J with 78% sensitivity and 76% specificity. However, considering its complexity of 11 variables, the Bayesian WLR-BIC model is attractive in that it shows 65% sensitivity and 85% specificity with only two variables. (b) shows confusion matrix of frequentist MLLR-AIC with optimal cut-off 0.105.

(a) Performance

	Model	AUC	Sensitivity	Specificity	Youden's J
Frequentist	MLLR-AIC	0.8455278	0.7833333	0.76	0.6288611
	MLLR-BIC	0.8306944	0.8166667	0.69	0.5066667
	WLR-AIC	0.8309722	0.7833333	0.725	0.5083333
	WLR-BIC	0.7808056	0.6	0.8933333	0.4933333
Bayesian	MLLR-AIC	0.7969444	0.7	0.7883333	0.4883333
	MLLR-BIC	0.7873333	0.7333333	0.73	0.4633333
	WLR-AIC	0.7898889	0.7666667	0.7083333	0.475
	WLR-BIC	0.7783611	0.65	0.8533333	0.5033333

(b) Confusion matrix

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

Table 9: Forecasting

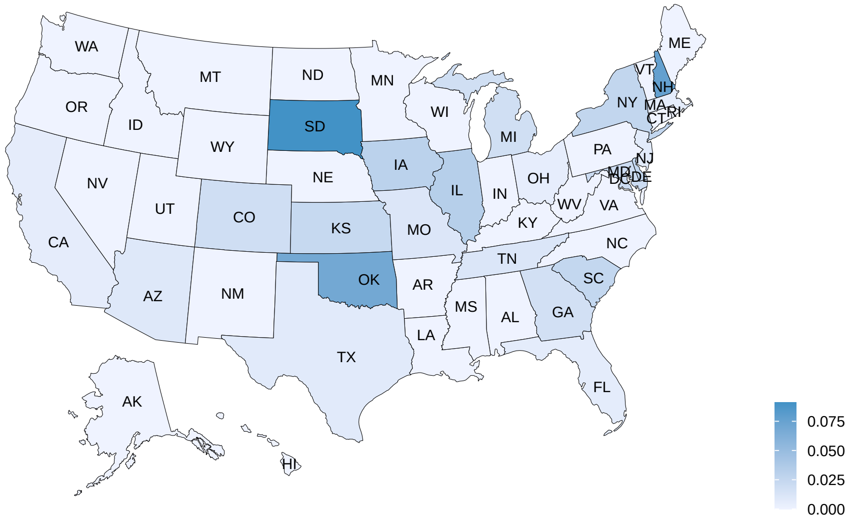
The table reports the top 50 companies with a high probability of bankruptcy and liquidation based on 2020 data. The results are predicted using the least complex Bayesian logistic regression model which simply has aco and optfvgr as explanatory variables. The gvkey is a global company key and loc is a current ISO country code where its headquarter locates.

rank	gvkey	company legal name	loc	aco	optfvgr	probability
1	20612	Nuinsco Resources Ltd	CAN	0.000	0.010	0.07712260
2	187545	Lupaka Gold Corp	CAN	0.002	0.010	0.07689359
3	116614	Uravan Minerals Inc	CAN	0.000	0.020	0.07605470
4	164631	Sphinx Resources Ltd	CAN	0.006	0.015	0.07590614
5	175071	Enertopia Corp	CAN	0.014	0.010	0.07553256
6	174368	Solar Alliance Energy Inc	CAN	0.005	0.020	0.07549061
7	26761	Advanced Proteome Therapeutics Corp	CAN	0.010	0.020	0.07493035
8	190576	Angkor Gold Corp	CAN	0.013	0.020	0.07459604
9	175527	Petrolympic Ltd	CAN	0.006	0.030	0.07433256
10	21874	Prosper Marketplace Inc	USA	0.000	0.040	0.07395952
11	170692	Wellness Center USA Inc	USA	0.001	0.040	0.07384926
12	16611	Soperior Fertilizer Corp	CAN	0.013	0.030	0.07356033
13	135145	Candente Copper Corp	CAN	0.013	0.030	0.07356033
14	170573	Pershimex Resources Corp	CAN	0.017	0.030	0.07312238
15	184946	Focus Graphite Inc	CAN	0.029	0.020	0.07283608
16	65688	Sienna Resources Inc	CAN	0.011	0.040	0.07275495
17	30749	QYOU Media Inc	CAN	0.030	0.020	0.07272736
18	140669	Rusoro Mining Ltd	CAN	0.030	0.020	0.07272736
19	170814	Asante Gold Corp	CAN	0.044	0.010	0.07222627
20	142501	Maxtech Ventures Inc	CAN	0.013	0.044	0.07213258
21	106789	CardioComm Solutions Inc	CAN	0.008	0.050	0.07206535
22	186296	Transatlantic Mining Corp	CAN	0.020	0.040	0.07178286
23	141309	Fuse Cobalt Inc	CAN	0.011	0.050	0.07174282
24	106561	Engold Mines Ltd	CAN	0.049	0.010	0.07168837
25	174026	Mobi724 Global Solutions Inc	CAN	0.040	0.020	0.07164838
26	65703	Inhibitor Therapeutics Inc	USA	0.032	0.030	0.07150139
27	20629	Moneta Porcupine Mines Inc	CAN	0.023	0.040	0.07146150
28	107693	Euromax Resources Ltd	CAN	0.035	0.030	0.07118120
29	25106	NewOrigin Gold Corp	CAN	0.008	0.060	0.07106208
30	25502	Lincoln Ventures Ltd	CAN	0.000	0.070	0.07091620
31	106517	Kiplin Metals Inc	CAN	0.000	0.070	0.07091620
32	15481	Urbanimmersive Inc	CAN	0.019	0.050	0.07088926
33	108841	Canadian Spirit Resources Corp	CAN	0.031	0.040	0.07061103
34	122273	Timberline Resources Corp	USA	0.014	0.060	0.07042665
35	107217	Playfair Mining Ltd	CAN	0.025	0.050	0.07025525
36	130402	Slam Exploration Ltd	CAN	0.003	0.074	0.07020315
37	184045	Belgravia Hartford Capital Inc	CAN	0.046	0.030	0.07001845
38	170401	Independence Gold Corp	CAN	0.000	0.080	0.06992773
39	125996	VoIP-PAL.com	USA	0.066	0.010	0.06988703
40	13825	Red Moon Resources Inc	CAN	0.007	0.075	0.06968572
41	178094	Xtierra Inc	CAN	0.055	0.030	0.06908020
42	176180	iCo Therapeutics Inc	CAN	0.029	0.060	0.06886097
43	26203	Forum Energy Metals Corp	CAN	0.011	0.080	0.06878394
44	18342	IOU Financial Inc	CAN	0.000	0.100	0.06798892
45	160541	Hercules Capital Inc	USA	0.000	0.100	0.06798892
46	22727	PureBase Corp	USA	0.005	0.100	0.06748027
47	184811	Bravada Gold Corp	CAN	0.024	0.080	0.06745454
48	17384	Banyan Gold Corp	CAN	0.054	0.050	0.06726419
49	179948	Rockhaven Resources Ltd	CAN	0.040	0.065	0.06725817
50	107048	ZincX Resources Corp	CAN	0.036	0.070	0.06718882

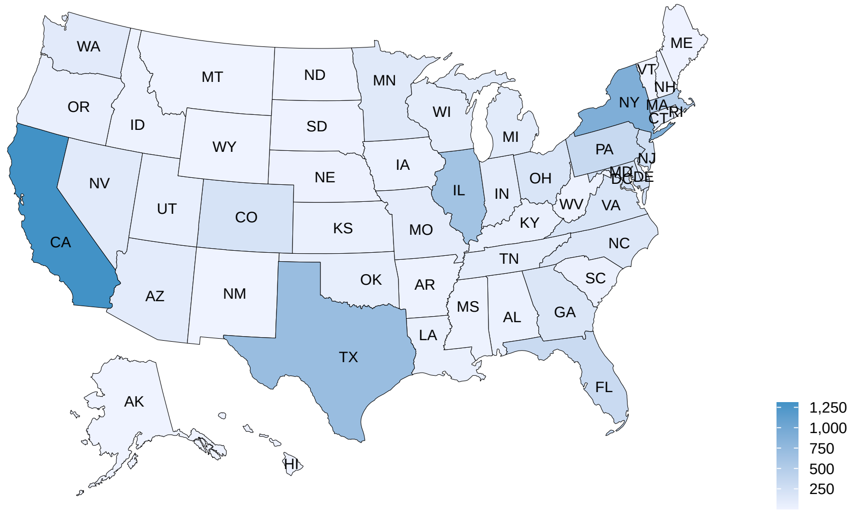
Figure 1: United States maps

These figures are the maps of states, counties, and cities in United States. (a) represents the proportion of event (bankruptcy or liquidation) by state. (b) and (c) show distribution of companies by states and counties. (d) refers to cities in which a company's headquarter is located. Not only is the sample geographically identified through the map, but it is also included in the variable candidates.

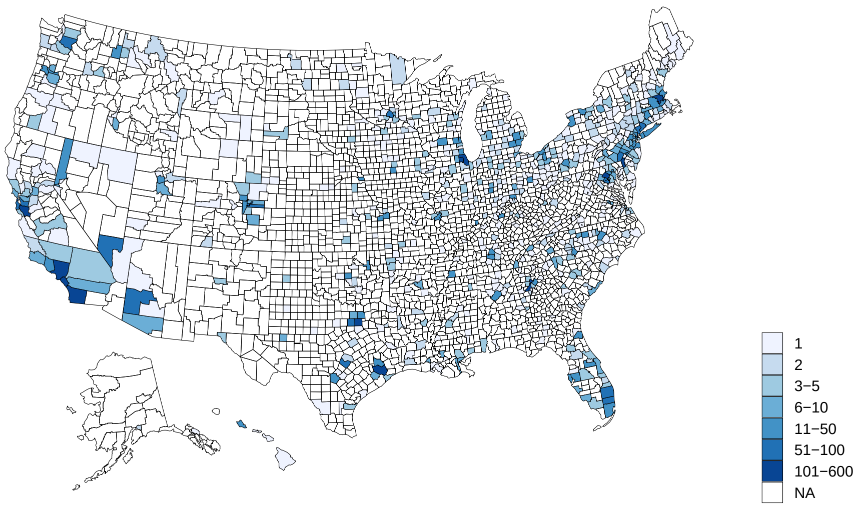
(a) Proportion of bankruptcy and liquidation by states



(b) Number of companies by states



(c) Number of companies by counties



(d) Plotting cities where a company is located

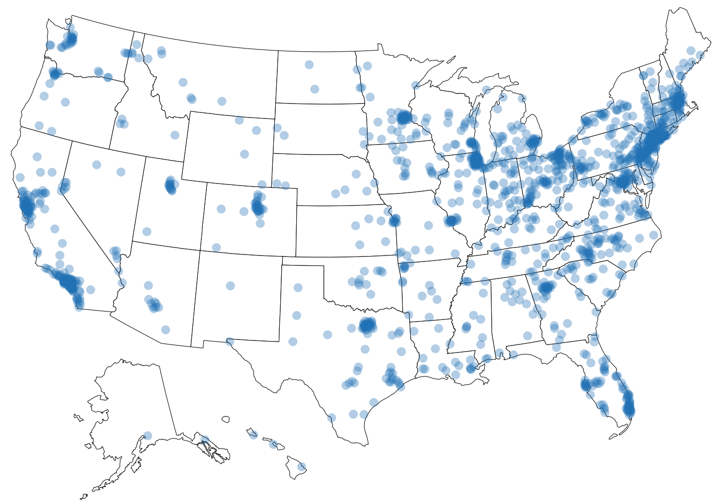
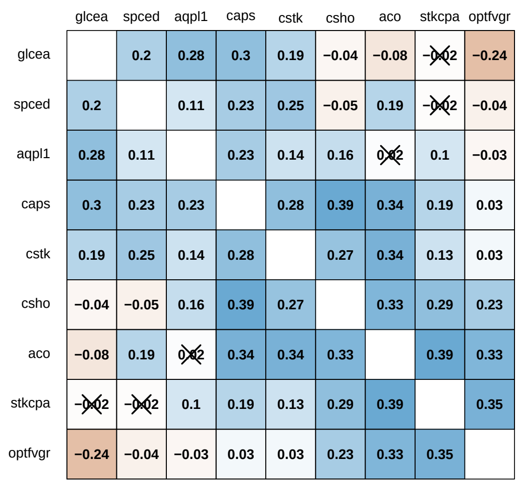


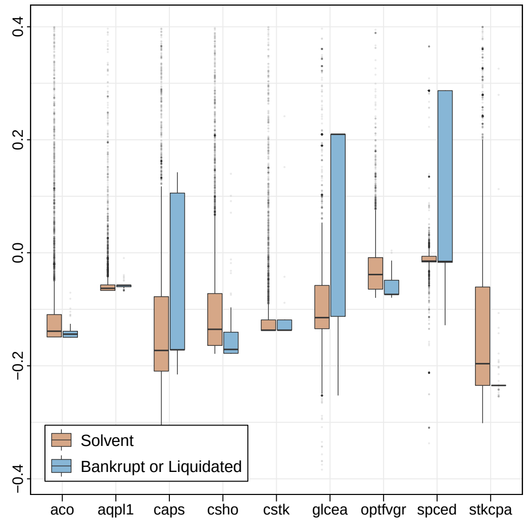
Figure 2: Correlation analysis with continuous variables

These figures show correlation analysis with continuous variables. Since all continuous variables are t-tested by BL, they differ visibly with BL. But idbflag and stalt differ relatively less significant with respect to continuous variables. The multicollinearity between the selected explanatory variables and their association with the response variable are reasonable.

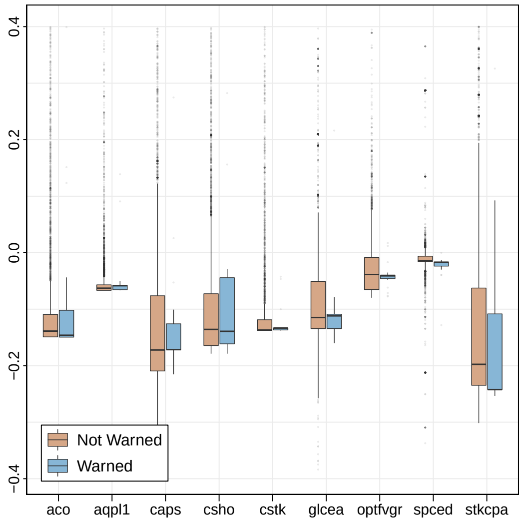
(a) Spearman correlogram with significance test



(b) Box plots by BL (truncated)



(c) Box plots by stalt (truncated)



(d) Box plots by idbflag (truncated)

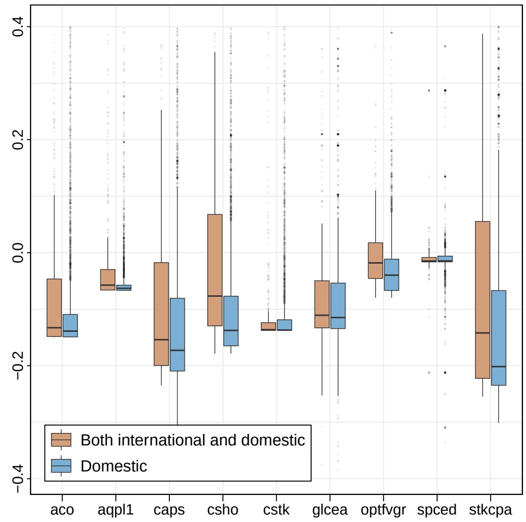
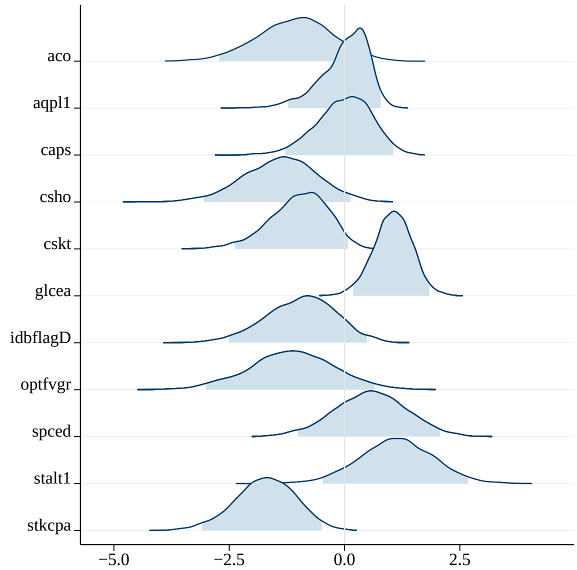


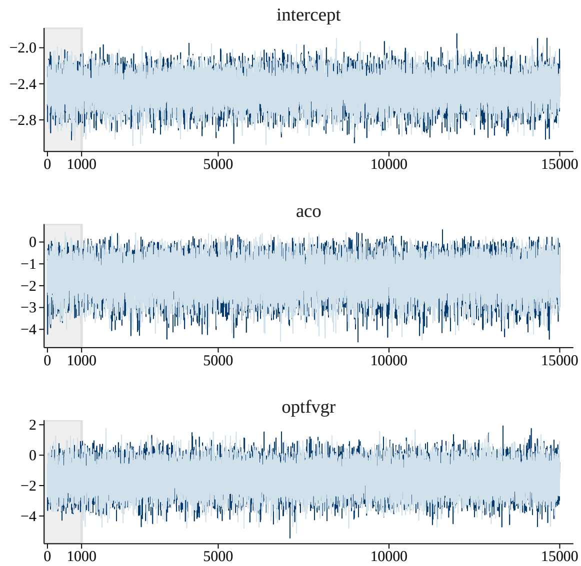
Figure 3: MCMC settings

(a) shows posterior density plot of parameters in MLLR-AIC model. The colored parts is the regions over 95% confidence intervals. If it contain zero, the parameter is not significant. (b) illustrates two chains of parameters in WLR-BIC model with 15,000 iterations and 1,000 burn-in. We monitor convergence visually. (c) refers to autocorrelation function of parameters in (b). In settings with high autocorrelation, i.e., there are no large jumps in the chain but sampled values are always close to the previous value, it may take many iterations before a sample is created that sufficiently represents the whole range of the posterior distribution. Thinning was set to 5 and appropriate.

(a) Density plot with 95% interval



(b) Trace plot with 1,000 warm-up



(c) Autocorrelation

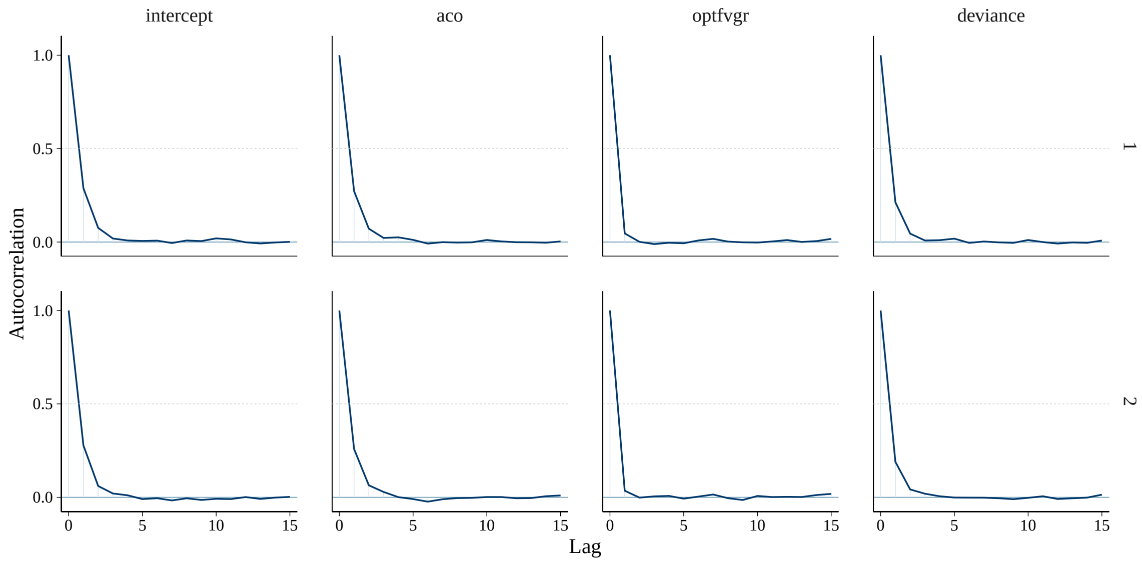


Figure 4: ROC curves

These figures show the ROC curves of final models. Subfigures are separated by variable selection methods and represent two path of frequentist and Bayesian. In other words, (a) includes aco, aqpl1, caps, csho, cstk, glcea, idbflagD, optfvgr, spced, stalt1, stkcpa, and (d) has aco, optfvgr as explanatory variables. The X-marked point is the position of the optimal cut-off given the same importance of sensitivity and specificity.