Bankruptcy and Liquidation Prediction Model

Math Capstone PBL (Data Analysis) – Project #2

Jaeseon Lee[1] Junwoo Yang[2]

December 8, 2020

[1]Department of Economics & Finance
Hanyang University

[2]Department of Finance
Hanyang University

Table of contents

Introduction

Topic and objective

Bankruptcy and Liquidation Prediction by Logistic Regression

- What are the variables associated with bankruptcy and liquidation?
- Which company does model suggest will go bankrupt or liquidate in the next three years?

Data sources

WRDS Compustat – Capital IQ [1]

- 226,866 observations $\times$ 981 variables
- Fundamentals annual of companies that are actively trading on the NYSE, AMEX, NASDAQ, TSX, or NYSE/Arca exchanges from 2000 to 2020

United States Cities Database [2]

- 29,488 observations $\times$ 17 variables
- This data include city name, state abbr., state name, county fips, county name, longitude and latitude of city, etc.

[1] http://wrds-web.wharton.upenn.edu.ssl.access.hanyang.ac.kr/wrds/ds/compd/funda/index.cfm?navId=83

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

Data description

Variable groups

- Identifying Information
- Identifying Information, cont.
- Company Descriptor
- Balance Sheet Items
- Income Statement Items
- Cash Flow Items
- Miscellaneous Items
- Supplemental Data Items
- Map Items

Data description

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 1: Variable names

Data description

Response variable

The response variable BL is defined as binary as follows:

$$\mathtt{BL} = \begin{cases} 1 & \text{if it went bankrupt or liquidated in 2011–13} \\ 0 & \text{otherwise (solvent company).} \end{cases}$$

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

Table 2: Number of deleted companies

Data description

Explanatory variables: fundamentals of 2010

`aco`: current assets that are not included in cash, cash equivalents, receivables or inventory on the Balance Sheet.

`aqpl1`: assets measured at fair value using observable inputs based on unadjusted quoted prices for identical instruments in active markets.

`caps`: a group of capital accounts other than capital stock or retained earnings.

`csho`: net number of all common shares outstanding at year-end, excluding treasury shares and scrip.

`cstk`: total par, carrying, or stated value of all common/ordinary capital.

Data description

Explanatory variables: fundamentals of 2010

`glcea`: after-tax gain or loss on a sale that is excluded from the Standard & Poor's Core Earnings calculation.

`idbflag`: source of data for the company.

`optfvgr`: weighted average fair value of options granted during the year.

`spced`: Standard & Poor's Core Earnings EPS diluted value.

`stalt`: status alert as to whether the company is in bankruptcy or undergoing a leveraged buyout.

`stkcpa`: amount of stock-based compensation expensed on the Income Statement during the current period on an after-tax basis.

Methodology

Logistic regression model

$$y_i \, (= \mathrm{BL}_i) \sim \mathrm{Bernoulli}(p_i)$$

$$\mathrm{logit}(p_i) = \log \frac{p_i}{1-p_i} = \beta X_i, \quad p_i = \frac{1}{1+e^{-\beta X_i}}$$

$$\text{where } \beta = \begin{bmatrix} \beta_0 & \beta_1 & \cdots & \beta_m \end{bmatrix}, \; X_i = \begin{bmatrix} 1 & x_{1,i} & \cdots & x_{m,i} \end{bmatrix}^{\mathrm{T}}$$

Maximum Likelihood Estimation (MLE)

$$L(\beta | X_1, \cdots, X_n) = \prod_{i=1}^{n} (p_i)^{y_i} (1-p_i)^{1-y_i}$$

$$\log L(\beta | X_1, \cdots, X_n) = \sum_{i=1}^{n} y_i \log p_i + \sum_{i=1}^{n} (1-y_i) \log(1-p_i)$$

EDA

North American Industrial Classification System [3]

NAICS is a hierarchical structure and can consist of up to six digits/levels. It is a comprehensive system covering all economic activities. There are 20 sectors and 1,057 industries in 2017 NAICS United States.

NAICS vs. SIC

The NAICS was developed to eliminate the inconsistent logic utilized in the SIC system and to increase specificity from the 4 digit SIC system by creating a 6 digit NAICS code. The last revision of the SIC was in 1987.

[3] `http://www.census.gov/epcd/www/naics.html`

Handling missing values

Sector	#	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

Handling missing values

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: Replacing order is upward.

F-test

Let $X_{j,1}, \cdots, X_{j,n_j}$ be i.i.d. random variables with normal density and $\overline{X}_j$ be sample means for $j = 1, 2$. Then

$$F = \frac{s_1^2}{s_2^2} \sim F(n_1 - 1, n_2 - 1) \quad \text{where } s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (X_{j,i} - \overline{X}_j)^2.$$

$$H_{F,0} \colon \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad \text{vs.} \quad H_{F,1} \colon \frac{\sigma_1^2}{\sigma_2^2} \neq 1.$$

Variable selection

Student's t-test (when $H_{F,0}$ is accepted)

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

$$\text{where } s_p = \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}}$$

Welch's t-test (when $H_{F,1}$ is accepted)

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}} \sim t(\nu) \quad \text{where } \nu = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Statistical test

$$H_{t,0}\colon \mu_1 - \mu_2 = 0 \quad \text{vs.} \quad H_{t,1}\colon \mu_1 - \mu_2 \neq 0.$$

Variable selection

acdo	capxv	cshtr_c	dm	emp	ibcom	lifrp	oiadp	pncaeps	spced	txditc	txtubxintbs
aco	ceq	cshtr_f	dn	esopct	ibmii	lo	oibdp	ppegt	spceeps	txds	txw
acodo	ceql	cstk	dp	esopt	icapt	loxdr	optdr	ppent	sppe	txfed	wcap
acominc	ceqt	cstke	dpact	esub	idit	lse	optex	ppeveb	sppiv	txfo	xacc
acox	ch	dc	dpc	esubc	intan	lt	optexd	prsho	stkco	txndb	xad
act	chech	dclo	dpvieb	fatb	intano	lul3	optfvgr	prstkc	stkcpa	txndba	xidoc
aldo	ci	dcs	drc	fatc	intc	mibn	optgr	pstk	teq	txndbl	xint
am	cibegni	dcvsub	drlt	fate	intpn	mibt	optosby	pstkn	tfvce	txp	xopr
ao	cimii	dd	ds	fatn	invfg	mii	optosey	rdipd	tlcf	txpd	xpp
aocipen	cipen	dd1	dt	fato	invo	mkvalt	optprcby	rdipeps	tstk	txr	xpr
aodo	cisecgl	dd2	dudd	fatp	invrm	mrc1	optprcca	re	tstkc	txs	xrd
aox	citotal	dd3	dv	fincf	invwip	mrc2	optprcex	reajo	tstkn	txt	xrdp
ap	cld2	dd4	dvc	fopo	itcb	mrc3	optprcey	recch	txbco	txtubbegin	xrent
aqc	cld3	dd5	dvpa	fopox	itci	mrc4	optprcgr	recd	txbcof	txtubend	xsga
aqpl1	cld4	dfs	dvt	gdwl	ivaeq	mrc5	optprcwa	rect	txc	txtubposdec	
aqs	cld5	dilavx	dxd2	glcea	ivstch	mrct	optvol	recta	txdb	txtubposinc	
at	cogs	dlcch	dxd3	glcep	lco	mrcta	pi	rectr	txdba	txtubpospdec	
aul3	cshfd	dldte	dxd4	gp	lcox	ni	pidom	reuna	txdbca	txtubpospinc	
bkvlps	cshi	dlto	dxd5	ib	lcoxdr	niadj	pifo	revt	txdbcl	txtubsettle	
caps	csho	dltp	ebit	ibadj	lct	nopio	pnca	sale	txdc	txtubsoflimit	
capx	cshpri	dltt	ebitda	ibc	lifr	oancf	pncad	seq	txdfo	txtubtxtr	
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	

Table 5: Selected and removed variables by t-test (245/350)

How to solve multicollinearity?

Variance Inflation Factor (VIF)

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

where R_i^2 is the coefficient of determination of the regression equation

$$X_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_{i-1} X_{i-1} + \beta_{i+1} X_{i+1} + \cdots + \beta_n X_n + \varepsilon.$$

We eliminate X_j with the highest VIF. We recalculate the VIF with the rest of the variables except X_j, and eliminate a variable with the highest VIF again. Repeat this process until $\max \mathrm{VIF} < 10$.

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	
acodo	ceq	cshfd	dlcch	dxd2	ibadj	lo	mrct	pi	rectr	txdb	txtubpospdec
acominc	ceql	cshi	dldte	dxd3	ibc	loxdr	ni	pifo	reuna	txdba	txtubpospinc
acox	ceqt	cshpri	dlto	dxd4	ibcom	lse	niadj	pncaeps	revt	txditc	txtubtxtr
act	ch	cshtr_f	dltt	dxd5	ibmii	lt	oancf	ppegt	sale	txds	txtubxintbs
am	ci	cstke	dn	ebit	icapt	lul3	oiadp	ppent	seq	txfo	xacc
ao	cibegni	dd	dp	ebitda	intan	mibn	oibdp	ppeveb	spceeps	txndb	xint
aodo	cimii	dd1	dpact	esub	intpn	mibt	optosby	pstk	sppiv	txndba	xopr
aox	cisecgl	dd2	dpc	fopo	lco	mkvalt	optosey	pstkn	stkco	txndbl	xpr
ap	citotal	dd3	dpvieb	fopox	lcox	mrc1	optprcby	rdipd	teq	txpd	xrd
at	cld2	dd4	ds	gdwl	lcoxdr	mrc2	optprcca	re	tlcf	txt	xrdp
aul3	cld4	dd5	dt	glcep	lct	mrc3	optprcex	reajo	tstk	txtubbegin	xrent
capx	cld5	dfs	dv	gp	lifr	mrc4	optprcey	recd	tstkc	txtubend	xsga
capxv	cogs	dilavx	dvt	ib	lifrp	mrc5	optprcgr	rect	txc	txtubposinc	

Table 6: Selected and removed variables by VIF (90/245)

Data scaling

Standardization

$$X' = \frac{X - \mu}{\sigma}$$

Standardized regression coefficients can be used to directly compare the effects of independent variables because standardized variables have the effect of eliminating the measurement unit or variation of the original variable.

Summary statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(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

Visualization

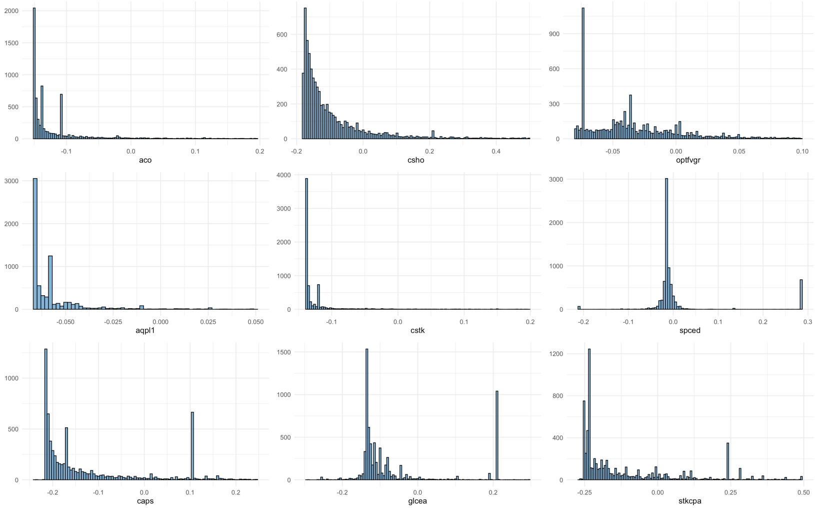

Figure 1: Histogram of continuous variables (truncated)

Visualization

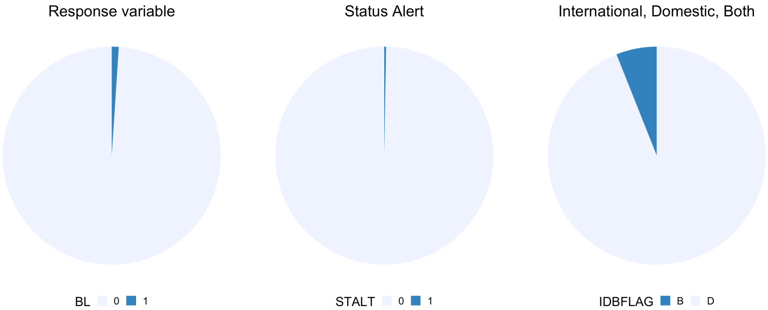

Figure 2: Pie chart of categorical variables

Visualization

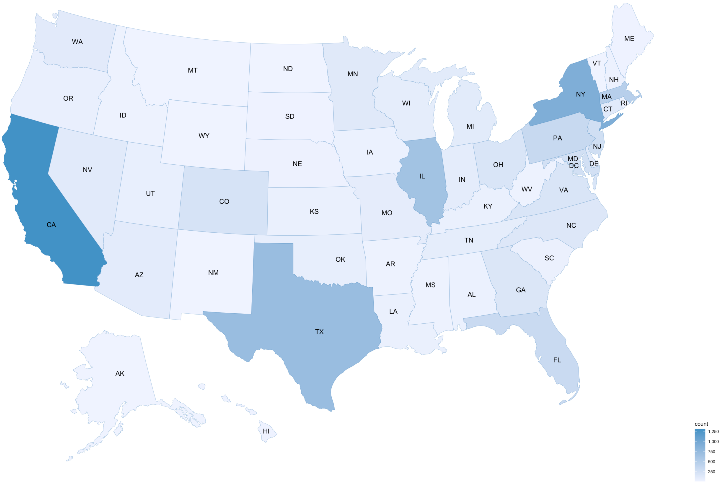

Figure 3: Number of companies by state

Visualization

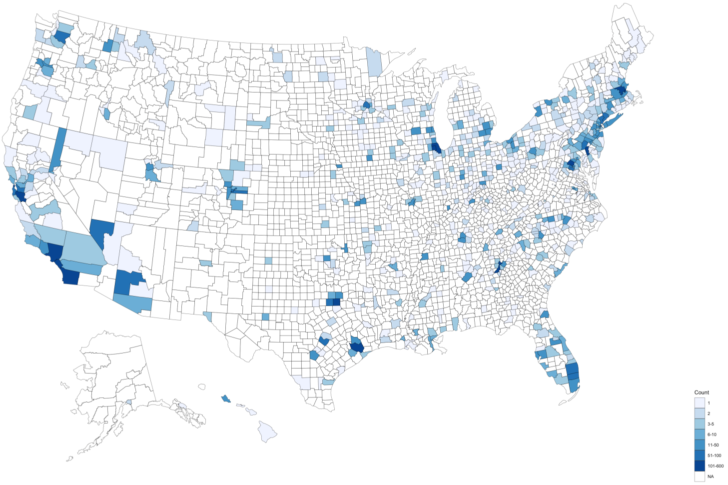

Figure 4: Number of companies by county

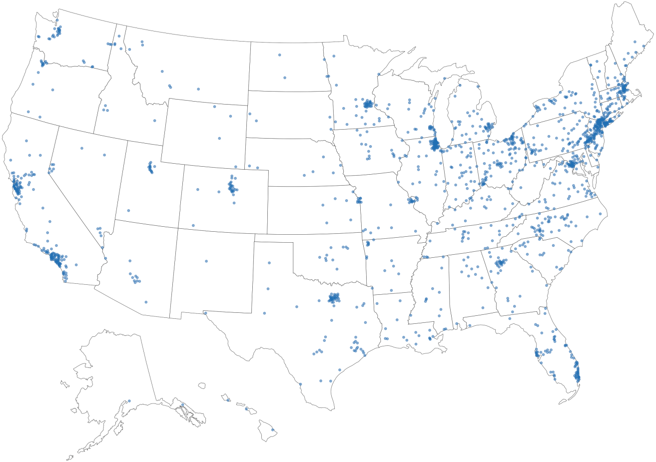

Figure 5: Cities that companies are located

Correlation Analysis

BL — continuous variable

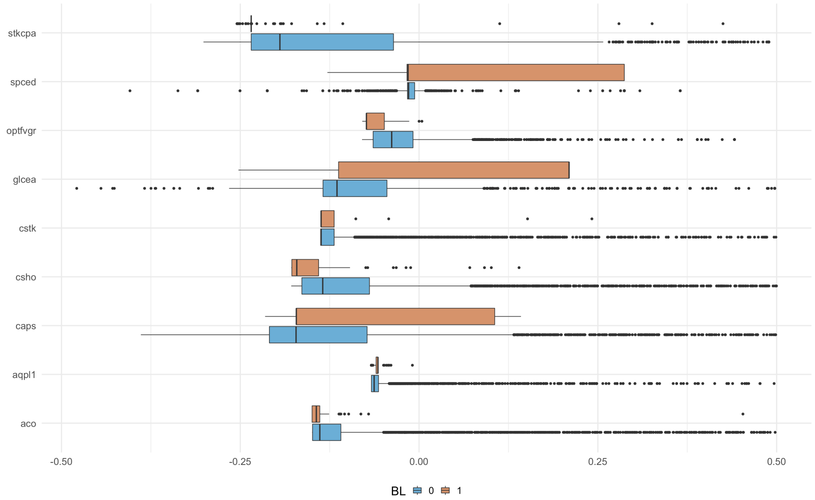

Figure 6: Box plots by BL (truncated)

BL — categorical variable

Pearson's chi-squared test

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} \sim \chi^2((r-1)(c-1)) \quad \text{where } e_{ij} = \frac{n_{i.}n_{.j}}{n}$$

$$H_0: p_{ij} = p_{i.}p_{.j}\ \forall i,j \quad \text{vs.} \quad H_1: \exists\, i,j \text{ s.t. } p_{ij} \neq p_{i.}p_{.j}.$$

Fisher's exact test

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} \sim \text{Hypergeometric}(n, a+b, a+c)$$

$$H_0: OR = 1 \quad \text{vs.} \quad H_1: OR \neq 1$$

where OR is true odds ratio.

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

Table 8: Result of tests

Neither `stalt` nor `idbfalg` are independent of `BL`. That is, the two categorical variables are associated with `BL`.

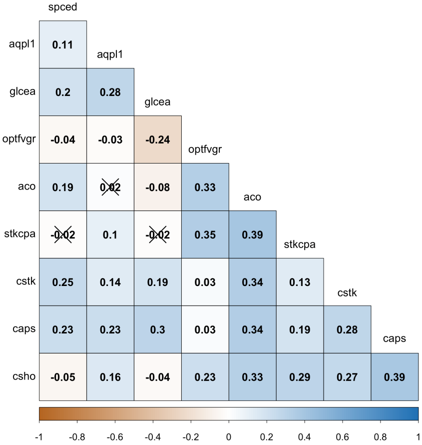

Figure 7: Spearman correlogram with significance test

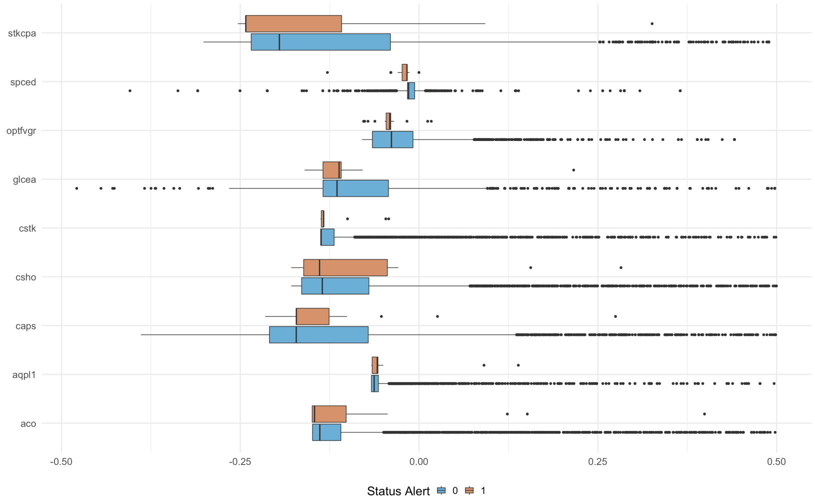

Figure 8: Box plots by `stalt` (truncated)

Between continuous and categorical variables

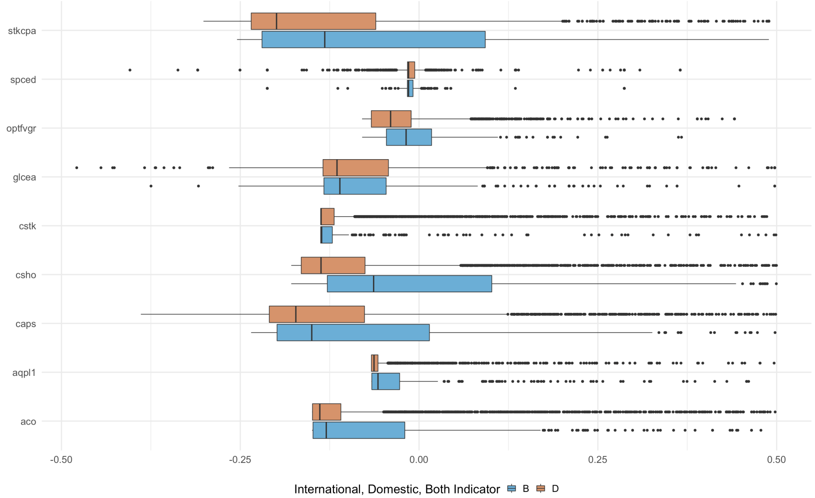

Figure 9: Box plots by `idbflag` (truncated)

Between categorical variables

Pearson's chi-squared test			Fisher's exact test			
χ^2	df	p-value	odds ratio	95% CI		p-value
2.648×10^{-29}	1	1	1.219	0.193	50.795	1

Table 9: Result of tests

Therefore, `stalt` and `idbflag` are independent!

Modeling

Imbalanced data

Method 1: subsampling training data

The first method is to subsample the negative set to reduce it to be the same size as the positive set, then fit the logistic regression model with the reduced data set.

Method 2: weighted logistic regression

For a data set containing 5% positives and 95% negatives, we can assign each positive observation a weight of 0.95, and each negative observation a weight of 0.05. The weighted likelihood can be written as

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

where w represents proportion of events in the population.

Type I error vs. Type II error

Pros and cons of both methods

Both of them predict a fair amount of true positives as positives and true negatives as positives. This means that Type II error decreases, but Type I error increases. However, it is more dangerous for a company that is actually going to go bankrupt to be predicted not to go bankrupt!

Subsampling – Training set vs. Test set

Training set: 60 bankrupt companies and 600 not bankrupt companies

Test set: 16 bankrupt companies and 160 not bankrupt companies

Stepwise selection by AIC

Akaike Information Criterion (AIC)

Let k be the number of estimated parameters in the model and $\hat{L}$ be the maximum value of the likelihood function for the model.

$$\mathrm{AIC} = 2k - 2\log\hat{L}$$

variable	type	variable	type	variable	type	variable	type
aco	numeric	dvc	numeric	naics2	factor	spced	numeric
aqpl1	numeric	emp	numeric	nopio	numeric	stalt	factor
bkvlps	numeric	exchg	factor	optex	numeric	state	factor
BL	factor	fate	numeric	optexd	numeric	stkcpa	numeric
caps	numeric	fic	factor	optfvgr	numeric	tstkn	numeric
census_region	factor	fincf	numeric	optgr	numeric	txdbca	numeric
chech	numeric	glcea	numeric	optprcwa	numeric	txdc	numeric
csho	numeric	idbflag	factor	optvol	numeric	txfed	numeric
cshtr_c	numeric	idit	numeric	prstkc	numeric	txs	numeric
cstk	numeric	intano	numeric	recch	numeric	wcap	numeric
dm	numeric	mrcta	numeric	recta	numeric	xad	numeric

Table 10: 44 variables before stepwise selection

Model fitting

	Estimate	Std. Error	z value	$P(>\|z\|)$
(Intercept)	−203.596	970.743	−0.210	0.834
aco	−7.190	6.036	−1.191	0.234
aqpl1	8.901	3.520	2.529	0.011
caps	2.117	0.961	2.204	0.028
csho	−7.575	3.088	−2.453	0.014
cstk	−13.510	11.405	−1.185	0.236
glcea	1.561	0.675	2.312	0.021
idbflag_D	197.005	970.673	0.203	0.839
optfvgr	−19.957	6.620	−3.015	0.003
spced	−2.420	1.488	−1.627	0.104
stalt_1	2.663	1.242	2.144	0.032
stkcpa	−3.308	1.449	−2.284	0.022

Table 11: Coefficients

Final model

Final model

Finally, our logistic regression model is

$$\begin{aligned}\log \frac{p_i}{1-p_i} = &-203.6 - 7.19x_{\text{aco},i} + 8.9x_{\text{aqppl1},i} + 2.12x_{\text{caps},i} \\ &- 7.58x_{\text{csho},i} - 13.51x_{\text{cstk},i} + 1.56x_{\text{glcea},i} + 197.01x_{\text{idbflag}_\text{D},i} \\ &\quad - 19.96x_{\text{optfvgr},i} - 2.42x_{\text{spced},i} + 2.66x_{\text{stalt}_1,i} - 3.31x_{\text{stkcpa},i}\end{aligned}$$

solving for p_i,

$$p_i = (1 + \exp(203.6 + 7.19x_{\text{aco},i} - 8.9x_{\text{aqppl1},i} + \cdots + 3.31x_{\text{stkcpa},i}))^{-1}.$$

Likelihood ratio test

Likelihood ratio test

$$LR = 2(ULF - RLF) \sim \chi^2_{df=q} \quad \text{where } q \text{ is \# of restrictions.}$$

$$ULF - RLF = 402.12 - 311.41 = 90.71$$

$$q = 659 - 648 = 11$$

$$\text{p-value} = 1.210039 \times 10^{-14}$$

H_0: $\beta_1 = \cdots = \beta_n = 0$

H_1: $\beta_i \neq 0$ for at least one i

Therefore, at least one explanatory variable is significant in predicting the response variable.

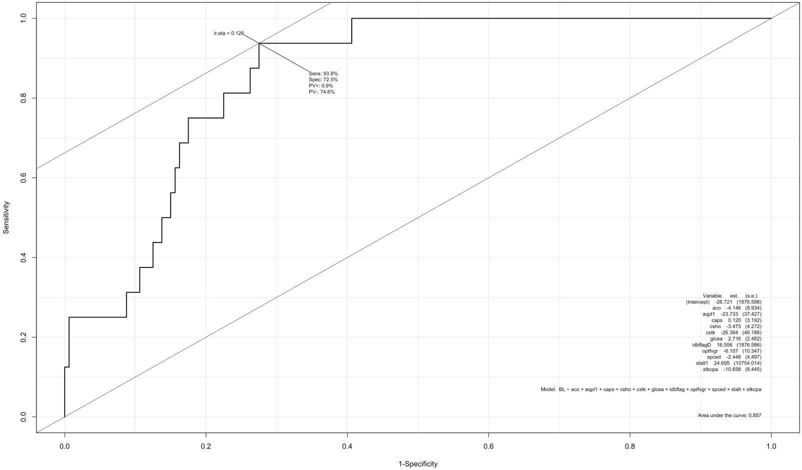

Figure 10: ROC curve

Confusion matrix with cut-off 0.1267

Terminology

- True Positives (TP)
  - These are cases in which we predicted positive (they go bankrupt or liquidate), and they actually went bankrupt or liquidated.
- True Negatives (TN)
  - We predicted negative, and they didn't actually go bankrupt or liquidate.
- False Positives (FP)
  - We predicted positive, but they didn't actually go bankrupt or liquidate.
  - Also known as a Type I error.
- False Negatives (FN)
  - We predicted negative, but they actually went bankrupt or liquidated. We cared more about this error.
  - Also known as a Type II error.

Measuring performance 1

- Accuracy
  - Overall, how often is the classifier correct?
  - $$\frac{\text{TP+TN}}{\text{Total}} = \frac{14 + 117}{176} = 74.43\%$$
- Misclassification Rate (Error Rate)
  - Overall, how often is it wrong?
  - equivalent to 1 minus Accuracy
  - $$\frac{\text{FP+FN}}{\text{Total}} = \frac{43 + 2}{176} = 25.57\%$$

Measuring performance 2

- True Positive Rate (Sensitivity, Recall)
  - When it's actually positive, how often does it predict positive?
  - $$\frac{\text{TP}}{\text{Actual Positive}} = \frac{14}{16} = 87.5\%$$
- False Positive Rate
  - When it's actually negative, how often does it predict positive?
  - $$\frac{\text{FP}}{\text{Actual Negative}} = \frac{43}{160} = 26.875\%$$
- True Negative Rate (Specificity)
  - When it's actually negative, how often does it predict negative?
  - equivalent to 1 minus False Positive Rate
  - $$\frac{\text{TN}}{\text{Actual Negative}} = \frac{117}{160} = 73.125\%$$

Measuring performance 3

- Precision
  - When it predicts positive, how often is it correct?
  - $$\frac{\text{TP}}{\text{Predicted Positive}} = \frac{14}{57} = 24.56\%$$
- Prevalence
  - How often does the positive condition actually occur in test set?
  - $$\frac{\text{Actual Positive}}{\text{Total}} = \frac{16}{176} = 9.09\%$$
- AUC (Area Under an ROC Curve) $= 0.857$

Conclusion

11 selected variables

aco, aqpl1, caps, csho, cstk, glcea, idbflag, optfvgr, spced, stalt, stkcpa

Performance

Accuracy: 74.43%, Sensitivity: 87.5%, Specificity: 73.13%

Forecasting

Due to many missing values in above variables on 2020, we failed in forecasting. If variables are chosen in consideration of the 2020 missing values, the prediction will be successful because of good performance.

S. and Hadi, A.S. (2012)
Regression Analysis by Example.
Wiley, New York.

Stock J, Watson M. (2015)
Introduction to Econometrics.
Pearson, Boston.

Kleinbaum, D. G. (2010)
Logistic regression: A self–learning text.
New York: Springer.

Laitinen, E. K., Laitinen, T. (2000)
Bankruptcy prediction: Application of the Taylor's expansion in logistic regression.
International review of financial analysis, 9(4), 327–349.

Kuruppu, N., Laswad, F., and Oyelere, P. (2003)
The efficacy of liquidation and bankruptcy prediction models for assessing going concern.
Managerial auditing journal.

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

Maalouf, M., and Siddiqi, M. (2014)
Weighted logistic regression for large–scale imbalanced and rare events data.
Knowledge–Based Systems, 59, 142–148.

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