Bankruptcy and Liquidation Prediction Model

Math Capstone PBL (Data Analysis) – Project 2

Jaeseon Lee[1] Junwoo Yang[2]

[1]Department of Economics & Finance
Hanyang University

[2]Department of Finance
Hanyang University

December 8, 2020

Table of contents

Current section

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?

Raw data

WRDS Compustat – Capital IQ [1]

- 226,866 observations × 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 × 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

Variable groups

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

Variable names

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

Variable description 1

Response variable

The response variable BL is defined as binary as follows:

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

year	All deletion	Bankruptcy	Liquidation	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: Number of deleted companies

Variable description 2

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

Variable description 3

Explanatory variables: fundamentals of 2010

- 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)

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

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

Current section

Handling missing values with NAICS

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

Structure of 2017 NAICS

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

Example

Monster Beverage Corp		
	31	Manufacturing
	312	Beverage and Tobacco Product Manufacturing
	3121	Beverage Manufacturing
	31211	Soft Drink and Ice Manufacturing
	312111	Soft Drink Manufacturing

Nike Inc		
	31	Manufacturing
	316	Leather and Allied Product Manufacturing
	3162	Footwear Manufacturing
	31621	Footwear Manufacturing
	316210	Footwear Manufacturing

Table: Replacing order: ↑

Test of equality of two variances

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 $j = 1, 2$.

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

Statistical test[4]

$H_{F,0}$: Two normal populations have the same variance.

$H_{F,1}$: True ratio of variances is not equal to 1.

[4] $H_{F,0}$: Homogeneity of variance, $H_{F,1}$: Heteroscedasticity of variance

Tests of equality of two means

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

$H_{t,0}$: The two population means are equal.

$H_{t,1}$: True difference in means is not equal to 0.

Selected and Removed variables by t-test (245/350)

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	

Variance Inflation Factor (VIF)

VIF

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

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

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

Selected and Removed variables by VIF (90/245)

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	

Data scaling

Standardization

$$X_{\text{changed}} = \frac{X - \mu}{\sigma}$$

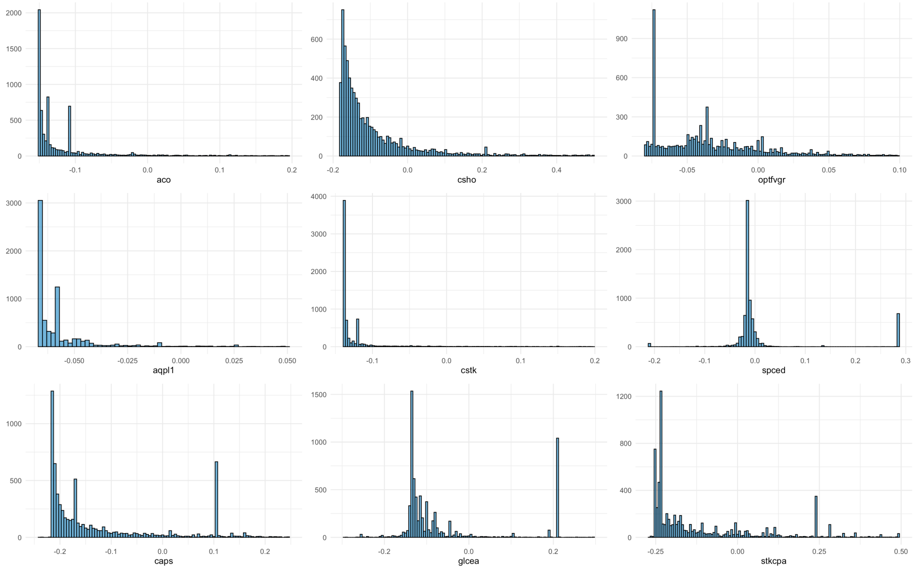
Standardized regression coefficients can be used to directly compare the effects of independent variables because standardized variables have the effect of eliminating the measurement unit or variation of the original variable.

Summary statistics of continuous variables

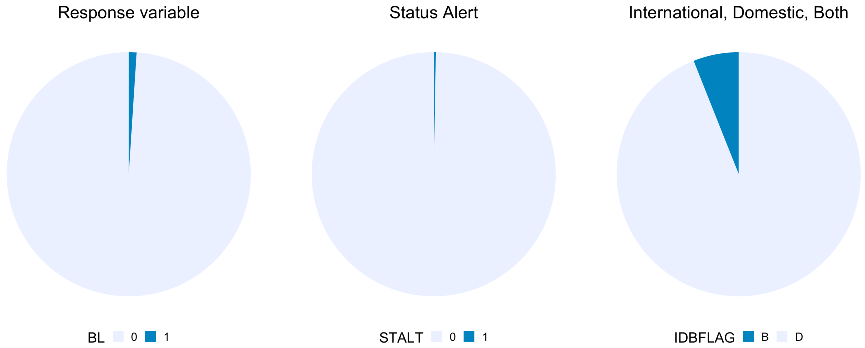
Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
aco	7,380	0.000	1.000	−0.150	−0.149	−0.107	42.624
aqpl1	7,380	0.000	1.000	−0.067	−0.067	−0.056	55.435
caps	7,380	0.000	1.000	−0.389	−0.209	−0.029	34.987
csho	7,380	0.000	1.000	−0.179	−0.163	−0.053	44.172
cstk	7,380	0.000	1.000	−0.138	−0.138	−0.119	35.533
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: Summary statistics

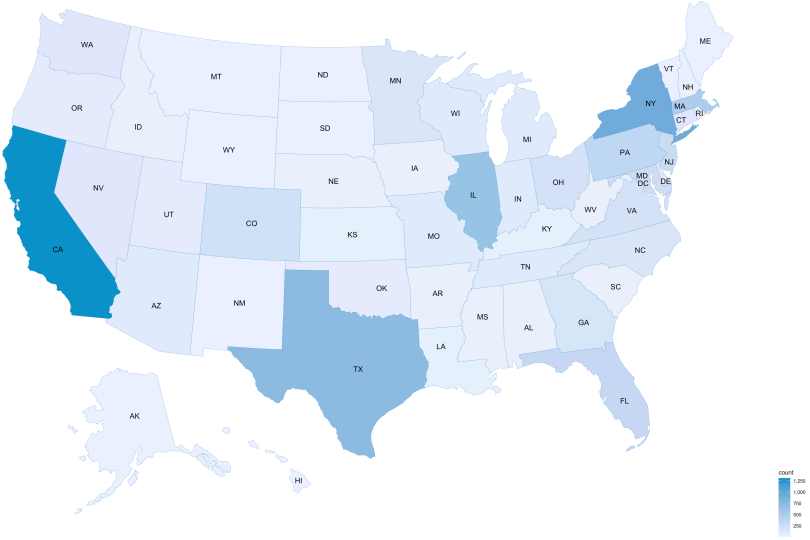
Histogram of continuous variables (truncated)



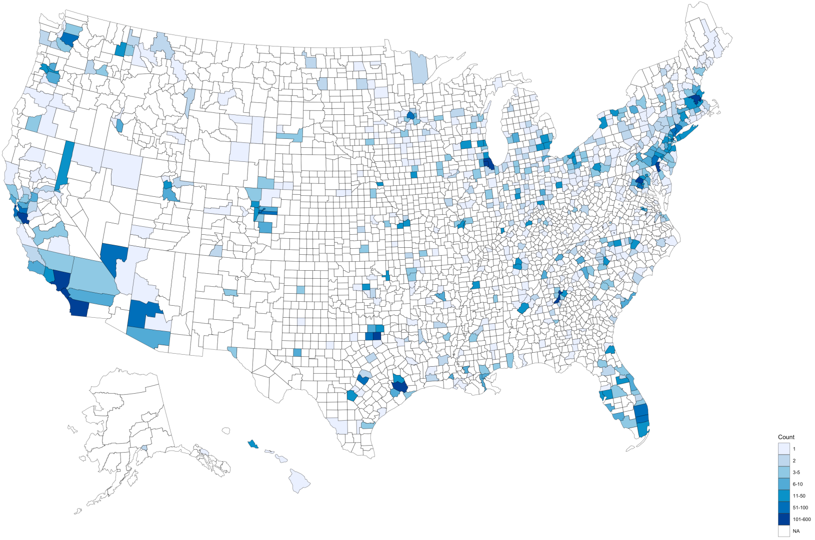
Pie chart of categorical variables



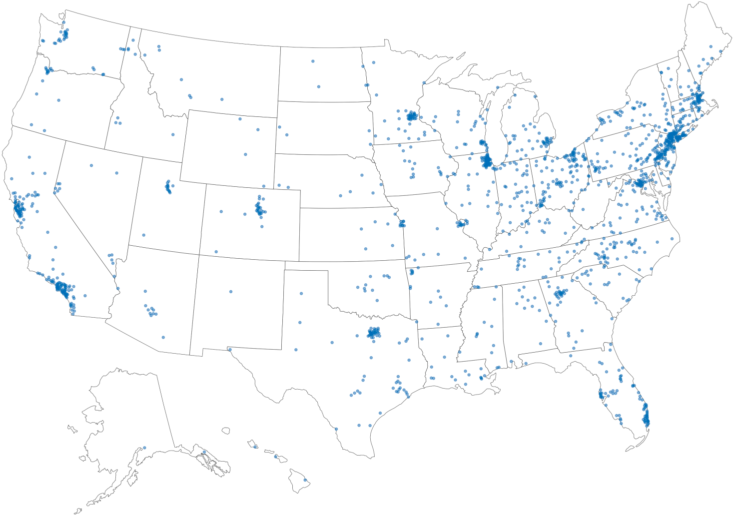
Number of companies by state



Number of companies by county



Plotting cities where the company is located

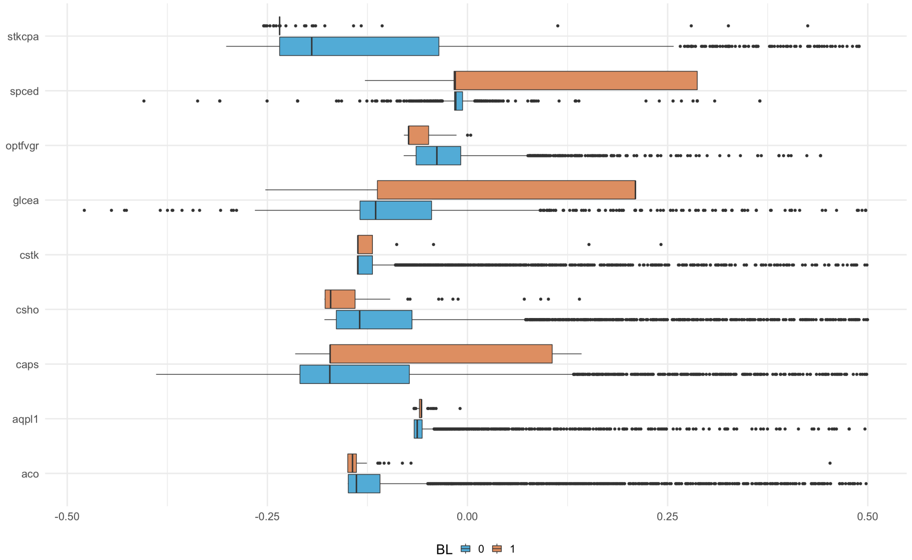


Correlation analysis

- BL $\sim$ continuous variables
- BL $\sim$ categorical variables (stalt, idbflag)
- between continuous variables
- continuous variables $\sim$ categorical variables (stalt, idbflag)
- between categorical variables (stalt $\sim$ idbflag)

Box plots (truncated)

BL ~ continuous variables



Contingency tables

BL ~ categorical variables

stalt

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

idbflag

		B	D	Total
BL	1	0	76	76
	0	445	6859	7304
	Total	445	6935	7380

Tests of independence of two categorical variables

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: $p_{ij} = p_{i.} p_{.j}$ for all i, j.

H_1: $p_{ij} \neq p_{i.} p_{.j}$ for some i, j.

Fisher's exact test

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

H_0: true odds ratio is equal to 1

H_1: true odds ratio is not equal to 1

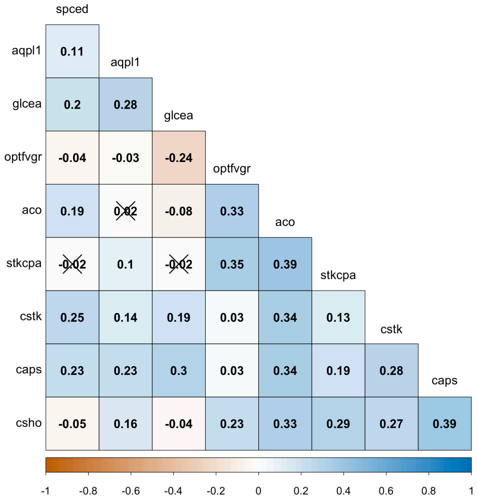
Result of tests

	Pearson's chi-squared test			Fisher's exact test			
	χ^2	df	p-value	odds ratio	95% confidence interval		p-value
stalt	53.376	1	2.755×10^{-13}	25.23833	5.993632	80.874026	4.441×10^{-5}
idbflag	3.9109	1	0.04797	∞	1.298749	∞	0.01416

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

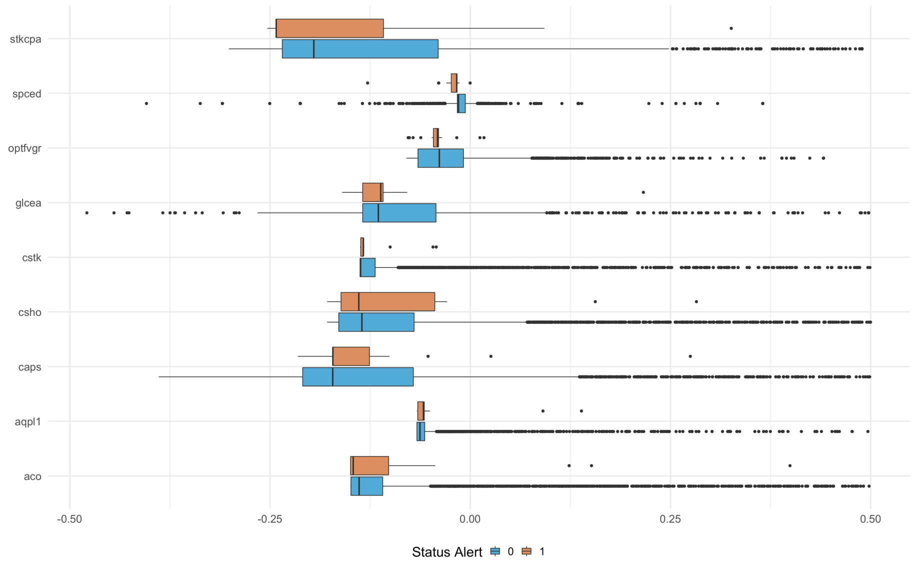
Spearman correlogram with significance test

between continuous variables



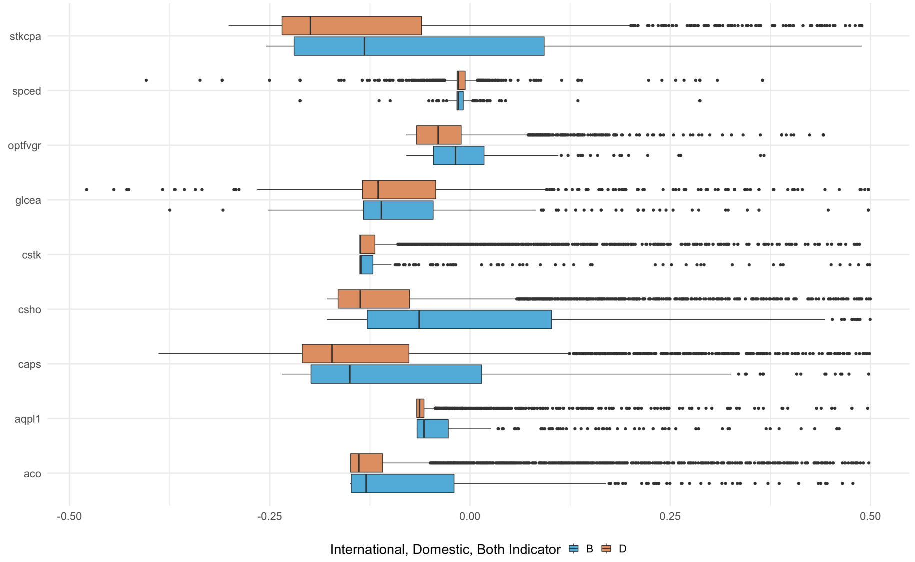
Box plots (truncated)

stalt ~ continuous variables



Box plots (truncated)

idbflag ~ continuous variables



Result of tests

between categorical variables

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

Pearson's chi-squared test			Fisher's exact test			
χ^2	df	p-value	odds ratio	95% confidence interval		p-value
2.6475×10^{-29}	1	1	1.219758	0.1930234	50.7951608	1

stalt and idbflag are independent!

Current section

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

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

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.

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

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: 44 variables before stepwise selection

Model fitting

	Estimate	Std. Error	z value	Pr(>|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: Coefficients

Final model

Final model

Finally, our logistic regression model is that

$$\begin{aligned}\log\frac{p_i}{1-p_i} = & -203.6 - 7.19x_{\mathrm{aco},i} + 8.9x_{\mathrm{aqppl1},i} + 2.12x_{\mathrm{caps},i} \\ & - 7.58x_{\mathrm{csho},i} - 13.51x_{\mathrm{cstk},i} + 1.56x_{\mathrm{glcea},i} + 197.01x_{\mathrm{idbflag_D},i} \\ & - 19.96x_{\mathrm{optfvgr},i} - 2.42x_{\mathrm{spced},i} + 2.66x_{\mathrm{stalt_1},i} - 3.31x_{\mathrm{stkcpa},i}\end{aligned}$$

Solving for p_i, this gives

$$p_i = \frac{1}{1 + e^{203.6 + 7.19x_{\mathrm{aco},i} + \cdots + 3.31x_{\mathrm{stkcpa},i}}}.$$

Likelihood ratio test

Likelihood ratio test

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

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

$$q = 659 - 648 = 11$$

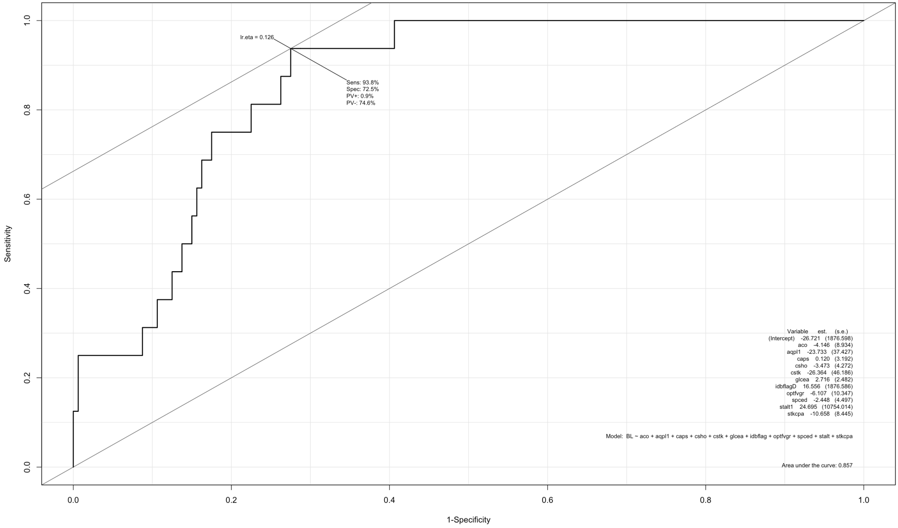
$$\text{p-value} = 1.210039 \times 10^{-14}$$

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 can be considered significant in predicting the response variable.

ROC curve



lr.eta = 0.126
Sens: 93.8%
Spec: 72.5%
PV+: 0.9%
PV-: 74.6%
Variable est. (s.e.)
(Intercept) -26.721 (1876.598)
aco -4.146 (8.934)
aqpl1 -23.733 (37.427)
caps 0.120 (3.192)
csho -3.473 (4.272)
cstk -26.364 (46.186)
glcea 2.716 (2.482)
idbflagD 16.556 (1876.586)
optfvgr -6.107 (10.347)
spced -2.448 (4.497)
stalt1 24.695 (10754.014)
stkcpa -10.658 (8.445)
Model: BL ~ aco + aqpl1 + caps + csho + cstk + glcea + idbflag + optfvgr + spced + stalt + stkcpa
Area under the curve: 0.857
Sensitivity
1-Specificity
0.0
0.2
0.4
0.6
0.8
1.0

Confusion matrix with cut-off 0.1267

		Predicted		
		Positive	Negative	Total
Actual	Positive	14	2	16
	Negative	43	117	160
	Total	57	119	176

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

Current section

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.

References I

S. and Hadi, A.S. (2012)
Regression Analysis by Example. 5th Edition.
Wiley, New York.

Stock J, Watson M. (2015)
Introduction to Econometrics. 3rd edition.
Pearson, Boston.

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

김기영 외. (2009)
예제로 배우는 *SAS* 데이터 분석 입문.
자유아카데미.

네이트 실버. (2014)
신호와 소음.
더 퀘스트, 223–268.

References II

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.