

Bankruptcy and Liquidation Prediction Model

Math Capstone PBL (Data Analysis) – Project 2

Jaeseon Lee

Student ID: 2015024284
Department of Economics & Finance
Hanyang University

Junwoo Yang

Student ID: 2017004093
Department of Finance
Hanyang University



December 8, 2020

Contents

Contents	i
List of Tables	iii
List of Figures	iv
Abstract	v
1 Introduction	1
1.1 연구주제 및 방향 설정	1
1.2 자료 소개	2
1.2.1 자료 출처 및 소개	2
1.2.2 변수 설명	3
1.3 통계 방법론 소개	6
1.3.1 로지스틱 회귀모형	6
2 EDA	10
2.1 자료 구축	10
2.1.1 결측치 처리	10
2.1.2 1차 변수선택	12
2.1.3 Data scaling and summary statistics	15
2.2 종속변수와 설명변수 간 상관분석	17
2.2.1 종속변수와 연속형 설명변수	19
2.2.2 종속변수와 범주형 설명변수	19
2.3 설명변수 간 상관분석	20
2.3.1 연속형 설명변수	21
2.3.2 연속형 설명변수와 범주형 설명변수	21

2.3.3 범주형 설명변수	22
3 Modeling	25
3.1 Model fitting	25
3.1.1 Subsampling with imbalanced data	25
3.1.2 Stepwise selection by AIC	26
3.2 Measuring performance	29
3.2.1 ROC curve	30
3.2.2 Confusion matrix	31
4 Conclusion	33
4.1 분석 결과	33
4.2 향후 분석방향 제시	34
References	36

List of Tables

1.1	Variable names	4
1.2	Number of deleted companies	5
2.1	Structure of 2017 NAICS	11
2.2	Examples	12
2.3	Selected and Removed variables by t-test and VIF	14
2.4	Summary statistics of continuous variables	17
2.5	Contingency tables	20
2.6	Result of tests	20
2.7	Contingency table	23
2.8	Result of tests	24
3.1	44 variables before stepwise selection	27
3.2	Coefficients	28
3.3	Confusion matrix with cut-off 0.1267	32

List of Figures

1.1	Graph of logistic function	6
2.1	Histogram of continuous variables (truncated)	16
2.2	Pie chart of categorical variables	16
2.3	Number of companies by state	17
2.4	Number of companies by county	18
2.5	Plotting cities where the company is located	18
2.6	Box plots of BL (truncated)	19
2.7	Spearman correlogram with significance test	21
2.8	Box plots of stalt (truncated)	22
2.9	Box plots of idbflag (truncated)	23
3.1	ROC curve	30

Abstract

기업의 파산 및 청산 예측은 중요하다. 채권자와 기업 투자자는 수익성 있는 결정을 위해 기업의 재무 불이행 가능성을 올바르게 판단해야 한다. 은행으로선, 기업 파산 및 청산에 대한 정확한 예측은 기업을 대상으로 안전한 대출 업무를 가능하게 하며 리스크를 적절히 반영하는 이자율을 부과할 수 있도록 한다. 또한 회계법인이 피감사회사의 파산 가능성에 대해 부적절한 예측을 할 경우 소송에 휘말릴 수 있다. 한편 파산 및 청산 예측 모형을 구축할 때 설명변수의 선택과 이를 변수 간 함수 형식 선택이 핵심 문제로 대두된다. E. K. Laitinen & T. Laitinen(2000)은 기업 내 현금이 많을수록, 순 현금 흐름이 많을수록, 그리고 외부로부터의 자금 조달이 유연할수록 지급 불능 위험이 줄어드는 사실에 착안하여, 총자산 대비 현금, 현금흐름, 그리고 자기자본 비율을 설명변수로 분석을 진행하였다. E. K. Laitinen & T. Laitinen(2000)에 따르면 파산 1년 전 재무상태를 바탕으로 분석을 진행하였을 때, 총자산 대비 현금 비율과 자기자본 비율만이 통계적으로 유의한 것으로 나타났다. 한편 Kuruppu et al.(2003)은 뉴질랜드 기업의 청산 예측 모형을 고안하였는데, 이들은 재무제표상 63개의 변수를 모형에 포함하였다. 이 중 총자산 대비 자기자본 비율 등 12개의 변수만이 통계적으로 유의하다고 나타났다. 본 연구는 기업의 재무적 요인뿐만 아니라 지리적 요인 등 다양한 변수를 분석에 포함함으로써 기업의 파산 및 청산에 영향을 미칠 수 있는 요인을 종합적으로 검증하고자 하였다.

1

Introduction

1.1 연구주제 및 방향 설정

기업의 파산 및 청산 예측은 중요하다. 채권자와 기업 투자자는 수익성 있는 결정을 위해 기업의 채무 불이행 가능성을 올바르게 판단해야 한다. 은행으로선, 기업 파산 및 청산에 대한 정확한 예측은 기업을 대상으로 안전한 대출 업무를 가능하게 하며 리스크를 적절히 반영하는 이자율을 부과할 수 있도록 한다. 또한 회계법인이 피감사회사의 파산 가능성에 대해 부적절한 예측을 할 경우 소송에 휘말릴 수 있다. 한편 파산 및 청산 예측 모형을 구축할 때 설명변수의 선택과 이를 변수 간 함수 형식 선택이 핵심 문제로 대두된다. E. K. Laitinen & T. Laitinen(2000)은 기업 내 현금이 많을수록, 순 현금 흐름이 많을수록, 그리고 외부로부터의 자금 조달이 유연할수록 지급 불능 위험이 줄어드는 사실에 착안하여, 총자산 대비 현금, 현금흐름, 그리고 자기자본 비율을 설명변수로 분석을 진행하였다. E. K. Laitinen & T. Laitinen(2000)에 따르면 파산 1년 전 재무상태를 바탕으로 분석을 진행하였을 때, 총자산 대비 현금 비율과 자기자본 비율만이 통계적으로 유의한 것으로 나타났다. 한편 Kuruppu et al.(2003)은 뉴질랜드 기업의 청산 예측 모형을 고안하였는데, 이들은 재무제표상 63개의 변수를 모형에 포함하였다. 이 중 총자산 대비 자기자본 비율 등 12개의 변수만이 통계적으로 유의하다고 나타났다. 본 연구는 기업의 재무적 요인뿐만 아니라 지리적 요인 등 다양한 변수를 분석에 포함함으로써 기업의 파산 및 청산에 영향을 미칠 수 있는 요인을 종합적으로 검증하고자 하였다.

파산을 주제로 하는 연구는 일반적으로 변수 간 선형(선형 판별 분석법)이나 로지스틱(로지스틱 회귀분석) 함수 관계를 상정한다. 특히 로지스틱 회귀분석은 개인의 암 발병 여부나 기업의 파산 여부처럼 양분된 내용의 자료를 분석하는 데 효과적이다. 실제로 E. K. Laitinen & T. Laitinen(2000)은 로지스틱 회귀모형과

테일러 전개를 이용하여 기업의 파산 예측 모형을 고안하고자 노력하였다. 한편 Hauser & Booth(2011)은 로지스틱 회귀모형을 바탕으로 파산 예측 모형을 도출하였다. 이들은 계수의 추정 방향을 Bianco-Yohai 추정량과 최대우도 추정량으로 나눈 후 두 가지 모형을 분석하였다. 그 결과 2006년에서 2007년 자료를 통해 전자의 모형만 리먼 브라더스의 파산을 정확하게 예측하였다. 본 연구는 로지스틱 회귀모형을 이용하여 기업의 파산 및 청산 예측 모형을 고안하고자 한다. 자료를 바탕으로 적합을 진행하고, 적합한 모형의 성능을 평가함으로써 예측 모형으로서 유용한지 파악해보고자 한다.

본 연구는 중장기적 파산 및 청산 예측 모형을 만들고자 한다. 일반적으로 기업의 파산 및 청산은 단기간에 결정되는 것은 아니며, 이에 대한 신호는 오랜 기간에 걸쳐 전해질 가능성이 크다. 즉, 기업의 특정 해의 데이터가 해당 기업이 곧 파산 및 청산 상태에 접어들 것을 진단하더라도, 그 이듬해가 아닌 2년 혹은 3년이 지난 후 파산 및 청산 상태에 접어들 수 있다. 한편 기간을 지나치게 장기적으로 설정하면 모형의 설명력이 떨어지게 된다. 따라서 본 연구는 3년이라는 적절한 기간을 상정함으로써 대부분의 경우를 포괄하고자 하였다. 또한 파산 및 청산 직전에는 기업의 재무상태가 매우 나쁠 가능성이 크기 때문에, 합리적인 중기적 예측 모형을 고안할 수 있다면 이는 예측의 실용성 측면에서도 매우 우수할 것이다. 2011년부터 2013년까지 3년 동안 기업의 파산 및 청산 여부를 2010년 자료를 바탕으로 분석하여 설명하고자 노력하였다. 본 연구를 통해 먼저 어떤 설명변수가 파산 및 청산 여부와 유의미한 상관관계가 있는지 파악해보고, 최종 모형과 2020년 현재 자료를 바탕으로 앞으로 3년 이내에 파산 및 청산하게 될 기업을 예측해보고자 한다.

1.2 자료 소개

1.2.1 자료 출처 및 소개

본 연구는 Wharton Research Data Service(WRDS) Compustat – Capital IQ¹와 United States Cities Database²에서 자료를 얻었다. 우선 2000년부터 2020년 11월까지 기간 내 뉴욕증권거래소(NYSE), 아메리카증권거래소(AMEX), 나스닥(NASDAQ), 토론토증권거래소(TSX), 또는 NYSE ARCA에 상장한 연간 기업 데이터를 구축하였다. 이는 총 981개 변수, 226,866개의 관측치를 포함한다. 한편 기업 펀더멘털 데이터는 개별 기업이 존재하는 주(State)와 도시(City)에 대한 정보만 가지고 있었다. 기업의 지역별 분포를 State, County, City 별로 한눈에 파악하고, 더 나아가 지리적 요건에 따라 기업의 파산 및 청산 여부가 다르게 나타나는지 상세히 분석해보고자 하였다. 따라서 기업 데이터 외에 미국의 지리적 특징을 반영하는 자료

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>

를 추가로 구축하였다. 해당 자료는 City, County, State의 이름과 County FIPS, 도시의 경도와 위도 등을 포함한다. 편더멘털 데이터와 City, State를 기준으로 병합한 County 별 자료를 지도를 활용하여 시각적으로 파악할 수 있도록 했으며, 경도 및 위도를 활용하여 기업의 좌표를 표시함으로써 분포를 한 눈에 파악할 수 있도록 노력하였다.

한편 본사가 미국 이외의 다른 국가에 있는 기업을 분석에서 제외하고자 하였다. 이를 통해 이후 지리적 요건과 다른 변수와의 관계성을 살펴보는 과정에서 편의성을 확보할 수 있었다. 또한 자료 내 파산 및 청산 기업의 비율을 증가시킴으로써 자료가 지나치게 불균형성을 보이는 것을 조금이나마 낮출 수 있었다. 우선 본사가 미국 외 국가에 존재하는 3,058개의 기업을 분석에서 제외하였다. 또한 미국 내 존재하는 기업에 대해서도 주에 대한 정보가 없는 4개의 기업을 분석에서 제외하였다. 한편 기업이 모종의 이유로 2011년 이전에 사라진 19개의 기업과 푸에르토리코와 팜에 있는 7개의 기업을 분석에서 제외하였다.

1.2.2 변수 설명

자료 내 변수는 크게 8개의 항목으로 구분할 수 있다. 회사명, NAICS 등 기업의 식별을 위한 항목(Identifying Information), 기업이 따르는 회계 기준 등 개별 기업의 특성을 나타내는 항목(Company Descriptor), 대차대조표상 항목(Balance Sheet Items), 손익계산서상 항목(Income Statement Items), 현금흐름표상 항목(Cash Flow Items), 발행된 보통주의 수 등을 포함한 기타 항목(Miscellaneous Items), 기업의 주가와 관련된 보충 항목(Supplemental Data Items), 그리고 지도 항목(Maps Items)으로 분류할 수 있다.

본 연구에 활용한 기업 데이터는 981개의 변수를 아우른다. 따라서 개별 변수의 의미를 모두 설명하는 것에는 어려움이 있으며, 분석과 관련해서도 무의미하다고 생각하였다. Table 1.1은 981개의 변수 중 일부(414개)를 나열한 것이다. 이 중에서도 최종 모형에 포함된 변수에 대해서만 상세히 설명하고자 한다.

acctstd	auop	costat	dlc	ebit	gind	ivch	naics4	reajo	tfvce
acdo	auopic	county_fips	dlcch	ebitda	glcea	ivncf	naics5	rech	tfvl
aco	bkvpls	county_name	dldte	ein	glced	ivstch	naics6	recco	tic
aco	BL	cshfd	dlsrn	emp	glceeps	ivstch	naicsh	recd	tlcf
acomine	busdesc	cshi	dltis	epsfi	glcep	lat	ni	rect	tstk
acox	caps	csho	dlto	epsfx	gp	lco	niadj	recta	tstkc
act	capx	cshpri	dlfp	epspi	gsector	lcox	nopi	rectr	tstkn
add1	capxv	cshr	dltr	epspx	gsbind	lcoxdr	nopio	reuna	tstkp
addzip	census_region	cshter_c	dltt	esoptc	gvkey	lct	np	revt	txach
adjex_c	ceoso	cshter_f	dm	esopdt	ib	liffr	oancf	sale	txbco
adjex_f	ceq	cstik	dn	esopnr	ibadj	lifrp	oiaidp	txbcof	txc
ajex	ceql	cstkcv	do	esopt	ibc	lno	oibdp	txdb	txw
ajip	ceqt	cstke	donr	esub	ibcom	lo	opeps	txdbc	upd
aldo	cfoso	curcd	dp	esubc	ibmii	lol2	oprepx	txdbc	wcap
am	ch	curncd	dpacl	exchg	icapt	long	prcc_c	txdbc	weburl
ano	che	currtr	dpc	exre	idbflag	loxdr	prcc_f	txdbc	xacc
ao	chech	cusip	dpvieb	fath	idit	lppl1	optex	txdc	xad
aocidengl	ci	datadate	drc	fatc	incorp	lse	prch_c	txdfed	xi
aociother	cibegni	dc	drlt	fatc	intan	lt	optfvgr	txdfio	xido
aocipen	cicurr	dclo	ds	fatl	intano	lul3	optigr	txdi	xidoc
aociseegl	cidergl	dcomm	dt	fatn	intc	mib	optlife	txditc	xint
aodo	cik	dcpstk	dudd	fato	intpn	mibn	optosby	txds	xintopt
aol2	cimii	dcs	dv	fatp	invch	miit	optosey	txfed	xopr
aoloch	clother	devsra	dvc	fax	invfg	mii	optprcby	txfo	xpp
aox	cipen	devsub	dvp	fca	invo	mkvalt	optprcca	txndb	xpr
ap	cisecgl	dcvt	dvpva	fdate	invrn	mrc1	optprcex	txndha	xrd
apalch	citotal	dd	dvpssp_c	fiao	invt	mrc2	optprcey	txndbl	xrdp
appdedate	city	dd2	dvpssp_f	fic	invwip	mrc3	optprcgr	txndbr	xrent
aqc	cld2	dd3	dvpssx_c	fincf	ipodate	mrc4	optprewa	bstkrv	xsga
aqi	cld3	dd4	dvpssx_f	fopo	ismod	mrc5	optrfr	rdip	txp
aqpl1	cld4	dd5	dvt	itcb	itcb	mrct	optvol	rdipeps	txpd
aqs	cogs	dfs	dxd2	fyr	itci	mrcta	pdate	rdipd	txr
at	comm	diladj	dxd3	fyrc	ivaco	msa	pddur	rdipeps	txs
au	conn	dilavx	dxd4	gdwl	ivaeq	naics2	phone	re	txt
au3	connml	dxd5	dxd5	gggroup	ivao	naics3	pi	rea	txtubadjust

Table 1.1: Variable names

다음은 최종 모형에 포함된 설명변수이다. 먼저 aco는 대차대조표상 현금, 현금등가액, 미수금, 또는 채고 등에 포함되지 않은 자산을 의미한다. aqpl1은 공정가치로 계산된 자산을 나타낸다. caps는 자본잉여금에 해당하며, csho는 연말 시점 시장에 유통되고 있는 보통주의 수를 의미한다. cstk는 모든 일반 자본의 총 액면가를 나타낸다. glcea란 S&P의 계산에서는 포함되지 않은 매출액 증감의 세후 금액을 뜻한다. idbflag란 자료의 출처를 나타내는 범주형 변수로서, 미국 내(idbflag=D), 미국 외(idbflag=I), 그리고 둘 다(idbflag=B)의 세 가지 경우로 나뉜다. 자료의 출처가 미국 외(idbflag=I)인 경우는 없었으므로 사실상 두 가지 범주를 가지고 있다. 한편 자료 출처가 국내와 국외를 가리지 않는 기업(B)의 경우 다국적 기업으로 해석할 수 있다. optfvgr은 한 해 동안 허가된 옵션 공정가치의 가중 평균을 나타낸다. spced는 S&P 기준에서의 희석된 주당 순이익을 의미한다. 한편 stalt는 해당 기업이 파산 위험에 빠졌거나 기업 인수 진행 중에 있음(stalt=1)을 나타내는 범주형 변수이다. stkcpa는 손익계산서상 비용 처리된 주식기준보상액을 의미한다. 본 연구는 기업이 파산 및 청산하는 경우 1, 그렇지 않으면 0을 나타내는 BL 변수를 종속변수로 정의하고 분석을 진행하였다. WRDS로부터 구축한 데이터 내, 기업이 사라진 원인을 14가지로 구분한 DLRSN 변수에서 파산과 청산에 해당하는 관측치만 분류하여 BL 변수를 새롭게 정의할 수 있었다. Table 1.2는 원자료 내 DLRSN 변수에서 2011년부터 2020년까지의 기간 중 파산 및 청산한 기업의 수를 요약한 것이다.

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 1.2: Number of deleted companies

특히 본 연구는 기업의 파산 및 청산 과정이 단기간에 이루어지지 않는다는 점에 착안하여 파산 및 청산에 대한 중장기적 예측 모형을 고안하고자 한다. 따라서 종속변수의 기간으로서 개별연도를 정의하는 대신 3년이라는 충분한 기간을 설정하였다. 본 연구는 2011년부터 2013년까지의 3개 연도를 선별하였다. 즉 종속변수 BL은 다음과 같이 2011년부터 2013년 사이에 기업이 파산 및 청산을 한 경우 1, 그렇지 않은 경우 0으로 정의된다.

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

1.3 통계 방법론 소개

1.3.1 로지스틱 회귀모형

모든 기업은 상호 배타적인 두 집단으로 분류할 수 있다. 파산 및 청산한 기업 또는 그렇지 않은 기업. 또한 개별 기업은 파산 및 청산 예측에 사용 가능한 재무적, 비재무적 정보를 수집하여 공개하고 있다. 이러한 조건 하에 본 연구에서는 로지스틱 모형(logistic model)을 이용하고자 한다.

로지스틱 모형은 이항(binary)의 종속변수와 여러 설명변수 사이의 관계를 분석하기 위하여 고안된 비선형 회귀 모형으로서, 기업의 파산 및 청산 예측에 가장 일반적으로 사용되는 통계 방법론이다. 로지스틱 함수(logistic function)은 로지스틱 모형이 기저로 상정하는 수학적 형태를 나타낸다. 로지스틱 함수 $f(z)$ 는 다음과 같이 정의된다.

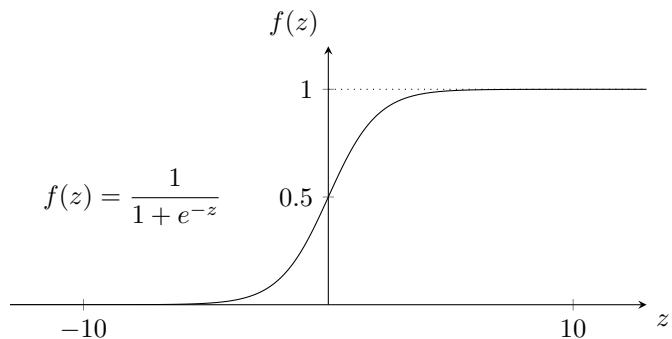


Figure 1.1: Graph of logistic function

즉 정의역은 실수 전체이며, z 가 $-\infty$ 에 가까워질 때 $f(z)$ 는 0으로 수렴하며, z 가 ∞ 에 가까워질 때 $f(z)$ 는 1로 수렴한다. 즉, $f(z)$ 의 값은 언제나 0과 1 사이에 분포한다. 그러므로 로지스틱 모형을 바탕으로 어떠한 위험 추정치를 구하더라도, 그 값은 0과 1 사이에서 나타나게 된다. 확률에 대한 추정이 이루어질 때 일반적으로 로지스틱 모형이 활용되는 근거가 여기에 있다. 한편 본 연구에서 관심 있는 확률은 개별 기업이 파산 및 청산할 수 있는 위험이다.

로지스틱 함수의 형태는 로지스틱 모형이 자주 쓰이는 두 번째 이유이다. 그림에서 알 수 있듯 로지스틱 함수 $f(z)$ 는 z 가 $-\infty$ 근처에서는 0에서 머물다가, z 가 커지면서 1을 향해 급격히 상승한 뒤, z 가 ∞ 에 가까워지면서 1 아래에서 변동이 없어진다. 그 결과 그림과 같은 S자 곡선 형태를 보인다. 변수 z 를 다양한 위험 요인이 결합된 지수, $f(z)$ 를 주어진 z 에서 기업의 파산 위험이라고 생각한다면, 함수의 S자 곡선 형태는 매우 합리적이다. 함수의 S자 형태는, 위험 요인이 낮은 수준에서는 개별 기업의 위험에 별다른 영향을 미치

지 못하지만, 특정 지점에 도달하면 위험은 빠른 속도로 증가하고, 위험 요인이 매우 높은 수준에 도달해서는 파산 위험도 엄청나게 높은 수준을 유지하게 된다고 해석할 수 있다. 이러한 비선형적 사고방식은 기업 파산 등 현실 세계를 설명할 때 매우 합리적이며, 다양한 분야에 적용할 수 있다.

로지스틱 모형은 로지스틱 함수로부터 고안할 수 있다. X_1, \dots, X_k 을 설명변수, $\alpha, \beta_1, \dots, \beta_k$ 을 미지의 모수(parameter)를 나타낸다고 하자. 이때 z 는 다양한 설명변수가 결합된 지수라고 생각할 수 있으며, 이는 선형식으로 표현 가능하다. z 를 나타내는 선형식을 로지스틱 함수 $f(z)$ 에 대입하면 다음과 같이 정리할 수 있다.

로지스틱 모형은 기업 파산과 관련하여 다음과 같이 활용할 수 있다: 연구 대상 기업에 관한 X_1, \dots, X_k 독립변수와, 파산 시 1, 파산하지 않았을 시 0을 나타내는 개별 기업의 파산 상태에 관한 변수가 존재한다. 이를 바탕으로 특정 기간 내 기업이 파산할 확률을 알아보고자 한다. 파산 확률은 조건부 확률의 형태, 즉 $P(B = 1 | X_1, \dots, X_k)$ 로 모형화 가능하다. 한편 개별 기업의 파산 확률은 로지스틱 함수 형태로 정리할 수 있는데, 해당 모형은 로지스틱 모형으로 정의된다. 모형 내 모수 추정은 기업의 데이터를 활용하여 이루어진다. 또한 모형 내 모수와 개별 기업에 관한 모든 정보를 아는 경우, 우리는 모형에 값을 대입하여 개별 기업이 파산하게 될 확률을 도출할 수 있다.

한편 로지스틱 모형의 모수 추정은 최대우도 추정법(ML: maximum likelihood)을 활용한다. 우도 함수(likelihood function)는 미지의 모수에 대한 함수이다. 최대우도 추정법은 우도 함수의 값을 최대화하는 모수에 대한 추정량을 도출한다. 특히 최대우도 추정량은 독립, 동일분포 하에서 점근적으로 최소오차 추정량과 일치하며, 정규분포를 따르고, 불편성(unbiasedness), 일치성(consistency), 효율성(efficiency)을 가진다. 한편 로지스틱 모형에서 추정된 개별 회귀계수에 대한 검정은 우도비 검정(likelihood ratio)이나 Wald 검정으로 진행된다. Wald 검정 시 귀무가설 하 검정 통계량이 근사적으로 표준정규분포를 따르므로, 검정 통계량의 제곱은 자유도가 1인 카이제곱 분포를 따르게 돼 두 검정 방법은 대표본에서 동일한 결과를 도출한다.

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

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

또한 모형 적합 시 승산비(OR: odds ratio)를 이용한다. 승산비는 추적 연구가 아닌 후향적 연구가 이루어지는 경우 활용되는 통계량으로, 성공 혹은 실패와 같이 이항으로 나뉜 자료에서 실패 확률에 대한 성

공 확률의 비율을 의미한다. 즉 $P(X)$ 가 연구의 주제가 되는 확률을 나타낼 때, 승산비는 $\frac{P(X)}{1-P(X)}$ 이다. 기업의 파산에 관한 연구의 경우 $P(X)$ 는 $P(B = 1|X_1, \dots, X_k)$ 이다. 한편 승산비에 자연로그를 취한 것을 로짓(logit)으로 정의하는데, 로짓 변환을 통해 로지스틱 모형 내 승산비를 포함할 수 있다.

로짓 함수에 로지스틱 모형을 대입하면 다음과 같은 대수적 과정을 거쳐 선형식으로 정리할 수 있다. 일 반적으로 로지스틱 모형은 아래와 같이 로짓을 포함한 형태를 의미하는데, $P(X)$ 를 포함한 형태와 본질적으로 동일하다.

로지스틱 모형에 대한 해석은 로짓 단위로 이루어진다. α 는 모든 설명변수가 0인 상황에서 로짓, 즉 승 산비에 자연로그를 취한 값으로 해석할 수 있다. 이는 로짓의 기초값(background, or baseline log odds)으 로도 해석된다. 한편 회귀계수 β_i 는, 다른 설명변수가 모두 고정된 상태에서, 설명변수 X_i 가 한 단위 변할 때 로짓의 변화량을 의미한다.

한편 로지스틱 모형의 적합도(goodness of fit)는 일반적으로 이탈도(deviance) 통계량을 바탕으로 측정 된다. 이탈도는 현재 모형과 포화 모형(saturated model)을 비교하는 우도비 통계량으로, 현재 모형의 우도 와 포화 모형의 우도가 유사할수록 좋은 모형이라고 판단할 수 있다. 즉 두 모형의 우도가 동일한 경우, 이 탈도는 0이다. 한편 적합 모형의 우도가 포화 모형의 우도보다 매우 작은 경우, 이탈도는 매우 커진다. 따라서 이탈도는 0 이상의 값을 나타내며, 우도비 검정 시 사용되는 카이제곱 통계량과 유사한 성질을 가진다. 그러므로 이탈도의 통계적 유의성에 대한 검정은 카이제곱 분포를 이용하게 된다. 또한 모형의 적합도에 대 한 검정은 이탈도와 카이제곱값을 비교하는 방식으로 진행된다.

한편 로지스틱 모형의 성능 평가는 혼동 행렬(confusion matrix)와 ROC 곡선(Receiver Operating Curve)을 바탕으로 진행된다. 혼동 행렬을 통해 민감도(sensitivity)와 특이도(specificity)를 파악할 수 있 다. 민감도는 양성 예측이 올바르게 이루어지는 비율(TPR: True Positive Rate)을 의미하며, 특이도는 음 성 예측이 올바르게 이루어지는 비율(TNR: True Negative Rate)을 의미한다. 특정 임계값(cut-off value)에 서 민감도와 특이도가 1에 가까울수록 모형의 분류 성능이 우수하다고 판단한다. 한편 임계값별 민감도 와 1-특이도의 양상을 나타낸 것을 ROC 곡선이라고 한다. 특히 1-특이도는 위양성 비율(False Positive Rate)을 의미한다. 이상적으로 민감도와 특이도 모두 1에 가까워야 하는데, ROC 곡선과 (1-특이도, 민감 도)가 (0, 1)인 좌표가 가장 가까운 곳에서 최적의 임계점이 결정된다. 한편 로지스틱 모형에서 도출된 ROC 곡선은 전반적인 모형의 성능을 평가하게 된다. ROC 곡선 아래 면적을 AUC(Area Under Curve)라 하는 데, AUC가 클수록 우수한 모형으로 판단한다. AUC가 0.5인 경우 무작위 추정과 다름없는 수준의 모형이 며, 1이면 완벽한 모형이다. 일반적으로 AUC가 0.8 이상이면 우수한 모형으로 판단한다.

결론적으로 본 연구에 활용된 로지스틱 모형은 다음과 같이 정리할 수 있다.

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

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

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

2

EDA

2.1 자료 구축

2.1.1 결측치 처리

자료 분석에 앞서 누락된 데이터를 올바르게 처리하는 것은 매우 중요하다. 사회 현상을 연구하는 경우 자료에 결측치가 포함되는 경우가 빈번한데, 결측치 처리가 부적절하게 이루어질 시 심각한 오류를 일으킬 수 있다. 우선, 결측치는 귀무가설이 거짓일 때 이를 올바르게 기각할 확률을 나타내는 통계적 검정력(Statistical Power)을 감소시킨다. 또한, 결측치는 모수 추정에 있어서 편향을 낳을 수 있다. 세 번째로 결측치는 표본의 대표성을 줄이며, 마지막으로 데이터 분석을 복잡하게 만든다.¹ 따라서 결측치 처리는 데이터 분석의 유효성과 맞닿아 있는 중요한 문제이며, 이를 적절한 방법에 따라 수행해야만 한다.

본격적인 결측치 처리에 앞서 모든 관측치에서 동일한 값을 가지는 26개의 변수를 제거하였다. 모든 기업이 동일한 값을 가지는 변수는 파산 및 청산 여부를 설명하는 데 아무런 도움이 되지 않기 때문이다. 또한 결측치가 80% 이상인 변수 552개도 분석에서 제외하였다. 이를 통해 연속형 변수 349개와 범주형 및 기타 정보 변수 65개, 총 414개의 변수와 7,461개 관측치로 자료를 정리하였다.

다음으로 연속형 변수 내 결측치를 산업 부문별 평균값으로 추정하여 분석을 진행하고자 하였다. 이는 기업 관련 데이터의 경우 기업 활동이 이루어지고 있는 산업 부문에 따라 특징이 상이하게 나타날 수 있기 때문이다. 결측치를 전체 평균값이 아닌 산업별 평균값으로 나타냄으로써 산업 부문에 따른 특징을 유지하고자 노력하였다. 본 연구는 NAICS 변수를 기업 산업 부문 분류의 기준으로 설정하였다. 일반적으로 사용

¹Kang H. (2013). The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64(5), 402–406.

되는 SIC 코드는 4자리로 산업을 분류하며, 1987년 개정 이후 바뀌지 않았다는 한계점이 있다. 이는 산업 흐름이 급변하는 현대 사회에서 다소 부적절한 분류 기준일 수 있다고 판단했다. 따라서 6자리 수로 SIC보다 더 상세하게 산업을 분류하며 비교적 최신인 2017년 개정된 NAICS 코드가 더 적절한 기준이라고 생각했다.

NAICS의 처음 두 자리는 기업 경제활동에 따른 일반적인 분류를 나타내며, 세 번째 자리는 그 하위 부문, 네 번째 자리는 산업 집단을 나타낸다. 다섯째 자리에서는 NAICS 구분에 따른 산업 집단, 그리고 마지막 여섯째 자리에서는 국가 산업을 나타낸다. Table 2.1은 NAICS의 구조를 나타낸 것이다. 1열에는 NAICS의 처음 두 자리를 의미하며, 2열은 자료 내 관측치 수를 나타낸다. NAICS 처음 두 자리 숫자에 대한 자세한 서술이 3열에 기술되어 있다. 예를 들어 NAICS 코드의 처음 두 자리가 31인 기업은 제조업으로 분류되며, 52인 경우 해당 기업은 금융업이나 보험업으로 분류된다.

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

Table 2.1: Structure of 2017 NAICS

에너지 드링크로 유명한 미국의 몬스터 음료 주식회사의 NAICS 코드가 312111이다. Table 2.2의 몬스터 음료 주식회사의 예시에서 31에서 312, 3121 등 자릿수가 증가할수록 분류가 더 세분화되는 것을 파악할 수 있다. 한편 나이키의 경우 제조업으로 분류되어 몬스터 음료 주식회사와 같은 31로 시작하지만, 세부 분류는 상이하다. 나이키의 경우 가죽 및 관련 제품 제조업을 나타내는 316으로 분류되어, 최종 316210으로 나타난다. 한편 NAICS 코드 상 3162 이하의 산업 부문은 신발 제조업으로 동일하다.

본 연구는 NAICS별 평균으로 결측치를 추정하였다. 6자리의 평균을 시작으로, 5자리의 평균, 4자리 평

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 2.2: Examples

균 등 하위 분류가 불가능할 시 상위 분류로 나아가는 방식으로 진행하였다. 한편 NAICS 변수도 560개의 결측치가 있었는데, 해당 관측치는 다른 변수와의 관계를 바탕으로 합리적으로 추정할 수 있었다. NAICS 값이 결측치인 관측치는 기업의 SIC 코드가 6722, 6726 둘 중 하나였다. 따라서 이들은 기업의 SIC 별 평균으로 대체하여 분석을 진행하였다. 위의 결측치 처리 과정을 통해 연속형 변수 350개에 존재하는 결측치를 모두 처리할 수 있었다.

2.1.2 1차 변수선택

연속형 변수 349개와 범주형 및 기타정보 변수 65개, 총 414개 변수에 대한 상관분석을 진행하기에 앞서 분석상 불필요한 변수를 일차적으로 제거하고자 하였다. 특히 종속변수를 설명하는 데 도움이 되지 않는 변수 와, 설명변수 간 상관성이 지나치게 높은 변수를 우선으로 선별하려고 노력하였다. 우선 파산 및 청산 여부 그룹별로 연속형 설명변수 사이에 평균의 차이가 존재하는지 T검정을 실시한 후, P값을 기준으로 종속변수 BL과 상관성이 없다고 판단되는 연속형 설명변수를 일차적으로 분석에서 제외하였다. 구체적으로 변수 내 집단의 분산이 같다는 귀무가설을 F검정을 이용하여 검정하였으며, 이 결과를 바탕으로 평균의 차이에 대해 T검정(귀무가설: 평균이 같다)을 실시하였다. F검정 결과 분산이 다르다고 판단한 변수에 대해서는 Welch Two Sample t-test를 진행하였으며, 분산이 같다고 판단한 변수는 Two Sample t-test를 진행하였다. T검정을 통하여 349가지 연속형 변수 중 105가지를 분석에서 제외하였다.

다음으로 설명변수 간 다중공선성(Multi-collinearity) 문제가 존재하는지 살펴보고, 이를 해결하였다. 설명변수 간의 완전한 또는 거의 완전한 선형종속 관계를 의미하는 다중공선성은 추정한 회귀계수의 분산을 매우 크게 만들어 신뢰성을 떨어뜨린다. 또한 회귀모형의 해석에도 문제를 일으킨다. 회귀계수에 대한 해석은 다른 설명변수를 모두 고정한 상태를 전제하는데, 다중공선성의 존재는 해당 가정에 모순을 일으킨

다. 또한 회귀계수의 추정치의 부호가 경험적 또는 이론적 기대와 상반된다. 본 연구는 분산확대인자(VIF: Variance Inflation Factor)를 이용하여 다중공선성 확인을 진행했다. 설명변수의 VIF를 조사하여 해당 수치가 10 이상의 값을 가지는 동시에 가장 큰 변수를 분석에서 순차적으로 하나씩 제외함으로써 다중공선성을 해결하고자 노력하였다. 해당 과정을 통하여 t-test를 통해 제외되고 남은 244개 변수 중 154가지를 분석에서 제외, 총 90개의 변수를 분석에 포함하였다. 위의 두 가지 과정을 거쳐 연속형 변수 90개와 범주 및 기타정보 변수 65개, 총 155개의 변수로 자료를 정리할 수 있었다. Table 2.3은 이를 정리한 것이다.

Removed variables by t-test (105/349)															
Removed variables by VIF (154/244)															
Selected variables (90)															
adjex_c	aoloch	currtr	do	epspi	glceeps	lno	oprepsx	preaeps	postkr	spi	txo	txtubadjust	txtubxitnis	txo	
adjex_f	apalch	dcom	dnor	epspx	invch	lol2	oprica	prcc_c	pstkrv	sstk	txtubxitnis	xi	xi	txtubxitnis	
ajex	aqi	dcpstk	dvp	esopdt	invt	long	optlife	prcc_f	rdip	tfr	txtubxitnis	xido	xido	txtubxitnis	
ajp	che	dcvsr	dvpsp_c	esopnr	ivaco	lqp11	optfr	prch_c	rdipa	tfvl	txtubxitnis	xintopt	xintopt	txtubxitnis	
ano	cicur	dcvt	dvpsp_f	extre	ivao	mib	pncwia	prch_f	rea	tstkp	txtubxitnis	xi	xi	txtubxitnis	
aocidergl	cidergl	diladj	dypsx_c	fat1	ivch	msa	pncwip	prcl_c	recco	txach	txtubxitnis	xintopt	xintopt	txtubxitnis	
aociother	ciother	dlc	dvpssx_f	fca	ivnclf	nopi	pnrsht	pvc1_f	seqo	txdfed	txdi	txtubxitnis	xi	xi	txtubxitnis
aocisegl	cshr	dltis	epsfi	fiao	ivst	np	pca	pstkc	siv	txndbr	txndbr	txtubxitnis	xi	xi	txtubxitnis
aol2	cstkev	dlttr	epsfx	gled	lat	opeps	prcad	pstkl	spce	txndbr	txndbr	txtubxitnis	xi	xi	txtubxitnis
acodo	ceq	cshfd	dlech	dxd3	ibc	loxdr	ni	pifo	reuna	txdba	txditc	txtubposinc	txtubposinc	txtubposinc	txtubposinc
acominc	ceql	cshi	dlto	dxd4	ibcom	lse	niadj	pncaebs	revt	txds	txditc	txtubxittrs	txtubxittrs	txtubxittrs	txtubxittrs
acox	ceqt	cshpri	dltt	dxd5	ibmii	lt	oancf	ppegt	sale	txfo	txndb	xacc	xacc	xacc	xacc
act	ch	cshtr_f	dn	ebit	icapt	hub3	oiadp	pment	seq	txfo	txndb	xint	xint	xint	xint
am	ci	estke	dp	ebitda	intan	mibn	oibdp	ppeveb	speeps	txndba	txndba	xopr	xopr	xopr	xopr
ao	cibegmi	dd	dpc	esub	intpn	mibt	optosby	pstkm	spiv	txndbl	txndbl	xpr	xpr	xpr	xpr
aodo	cimii	dd1	dpo	folo	lco	mkvalt	optosey	pstkm	stko	txpd	txpd	xrd	xrd	xrd	xrd
aox	ciseegl	dd2	dpvieb	fpopx	lcox	mrc1	optprchy	ridpd	teq	txrd	txrd	xrdp	xrdp	xrdp	xrdp
ap	citotal	dd3	ds	gdwl	lcoxdrl	mrc2	optprcca	re	tlcf	txt	txt	xrent	xrent	xrent	xrent
at	cld2	dd4	dt	glcep	lct	mrc3	optprcex	reajo	tstkl	txtubbegin	txtubbegin	xsga	xsga	xsga	xsga
aul3	cld4	dd5	dv	gp	lifr	mrc4	optprcey	recd	tstkc	txtubend	txtubend	xsga	xsga	xsga	xsga
capx	cld5	dfs	dvt	ib	lifrp	mrc5	optprcgr	rect	txc	txtubposinc	txtubposinc	xad	xad	xad	xad
capxv	cogs	dilawx	dxd2	ibadj	lo	mrcrt	pi	rectr	txdb	txtubposdec	txtubposdec	xad	xad	xad	xad

Table 2.3: Selected and Removed variables by t-test and VIF

마지막으로 원자료를 기준으로 결측치와 0 값이 6,000개를 초과하는 변수도 분석에서 제외하였다. 이는 1차 변수선택 과정에서 분석상 무의미한 변수를 최대한 선별하고자 하는 목적이었는데, 결측치가 80%가 넘지 않는 않더라도 0인 값이 지나치게 많으면 결측치 처리가 합리적으로 이루어질 수 없다고 판단했기 때문이다. 즉, 결측치가 60%이고 나머지 30%가 0 값을 가진 변수는 나머지 10%의 관측치를 통해 결측치를 추정하게 되므로 설득력이 떨어진다고 생각했다. 해당 과정을 통해 76개의 변수로 자료를 정리하였다. 한편 기업의 전화번호 등 무의미한 범주형 변수 및 기업에 관한 기타정보 변수를 분석에서 제외하였다. 결과적으로 1차 변수 선택과정을 통해 44개의 변수, 7,380개의 관측치를 선별하여 분석을 진행할 수 있었다.

2.1.3 DATA SCALING AND SUMMARY STATISTICS

전처리 과정에서 설명변수의 단위, 범위 등 특성을 조정해주는 작업은 중요하다. 변수마다 단위가 다른 경우, 더 큰 값을 가진 변수가 종속변수에 더 큰 영향을 미치는 결과가 나타날 수 있다. 또한 자료의 값이 너무 크거나 혹은 지나치게 작은 경우 모형이 0으로 수렴하거나 무한대로 발산하는 식으로 문제를 일으킬 수 있다. 자료의 특성을 조정해주는 작업을 특성 스케일링(Feature Scaling)이라고 하는데, 스케일링 방법은 매우 다양하다. 본 연구에서는 분류 모형 구축 시 유용한 표준화(standardization)를 통해 자료를 정리하고자 노력하였다. 표준화는 개별 변수의 평균을 0, 분산을 1로 만드는 스케일링 작업이다.

다음으로 개별 변수의 기초통계량을 파악하였다. 연속형 변수의 경우 히스토그램을 살펴보며 전반적인 분포를 파악하고자 하였다. 또한 평균, 최솟값, 최댓값 등을 파악하여 변수의 특징을 수치적으로도 살펴보았다. 한편 자료 내 변수가 상당히 많은 까닭에 분량상 제 1절부터 제 3절까지는 최종 모형에 포함된 변수만을 대상으로 표와 그림을 작성하였다. 다음의 Figure 2.1은 최종 모형 적합에 사용된 연속형 변수의 히스토그램을 나타낸 것이다. 한편 7,380개 관측치를 모두 포함하여 히스토그램을 그릴 시 범위가 지나치게 넓은 변수는 분포 파악이 어려웠다. 따라서 양 극단에 있는 소수의 관측치를 절단하고 히스토그램을 그렸다. 히스토그램 상 제외된 관측치는 최대 511개이다. 한편 Table 2.4는 최종 모형 내 연속형 변수의 기초통계량을 정리한 것이다. 평균과 표준편차의 경우 표준화 작업을 통해 각각 0과 1로 정리된 것을 확인할 수 있었다.

본 연구는 범주형 변수도 분석의 대상으로 포함한다. 따라서 이들에 대한 통계량을 살펴보는 것도 중요 한데, Figure 2.2은 최종 모형에 포함된 범주형 변수의 비율을 나타낸 파이 도표(Pie-chart)이다. 해당 도표를 통해 각 변수별 항목의 비율을 시각적으로 파악할 수 있다. BL과 stalt의 경우 기업의 파산 및 청산을 의미하는 1의 비율이 상대적으로 매우 작은 것을 알 수 있다. 또한 idbflag의 파이 도표를 통해서는 기업 자료의 출처가 대부분 미국 내에 국한된 것을 알 수 있다.

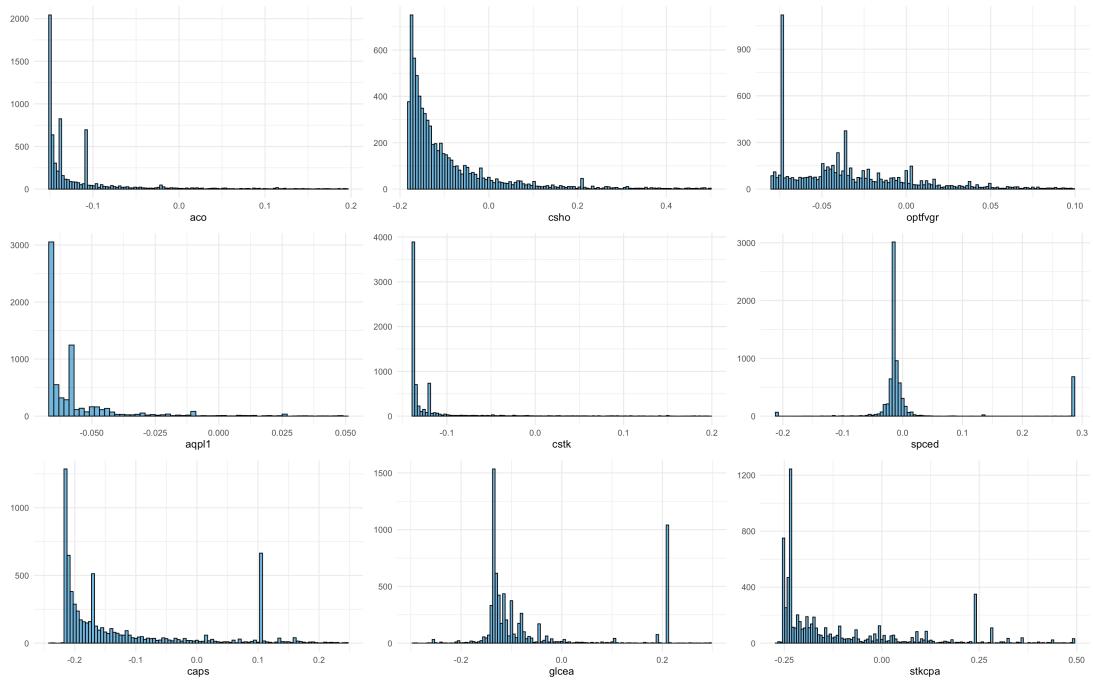


Figure 2.1: Histogram of continuous variables (truncated)

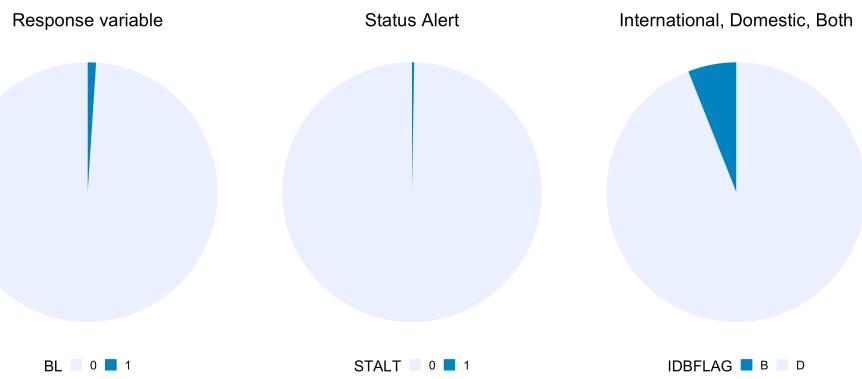


Figure 2.2: Pie chart of categorical variables

	Obs	Mean	Std.Dev	Min	25%	75%	Max
aco	7,380	0.000	1.000	-0.150	-0.149	-0.107	42.624
aql11	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 2.4: Summary statistics of continuous variables

한편 지리적 요건을 나타내는 범주형 변수를 분석에 추가하였다. 다음 Figure 2.3, 2.4, 2.5는 미국의 주, 카운티, 그리고 도시별 기업의 분포를 나타낸 지도이다. 실리콘밸리가 위치한 캘리포니아와 미국 북동부에 많은 기업이 분포하는 것으로 나타났다.

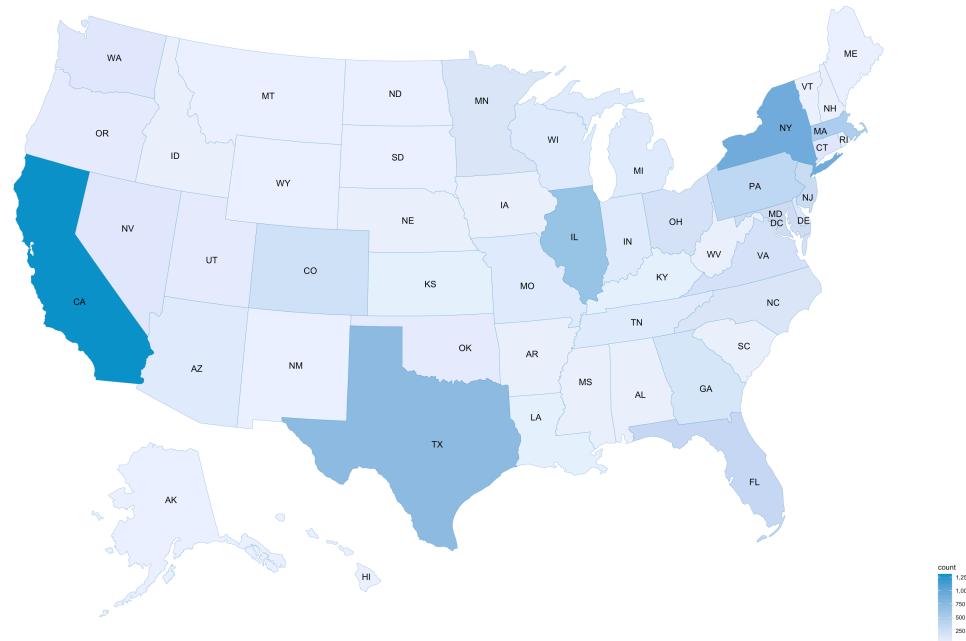


Figure 2.3: Number of companies by state

2.2 종속변수와 설명변수 간 상관분석

종속변수인 BL과 모든 설명변수간 상관분석을 진행하였다. BL은 0 또는 1의 값을 가지는 범주형 변수이므로 설명변수의 유형에 따라 상이한 방법으로 변수간 상관성을 분석하였다.

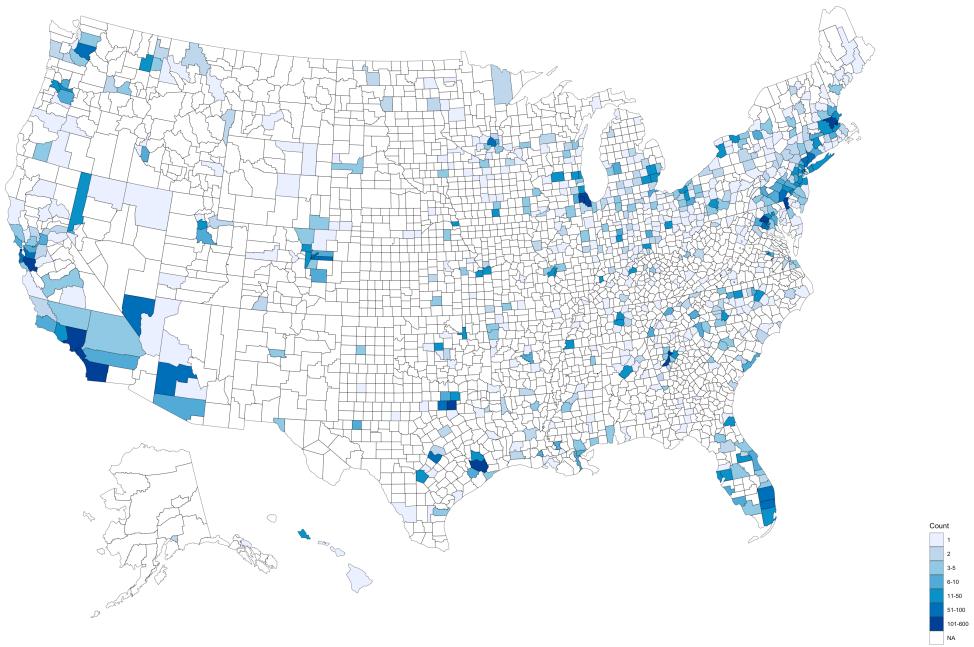


Figure 2.4: Number of companies by county

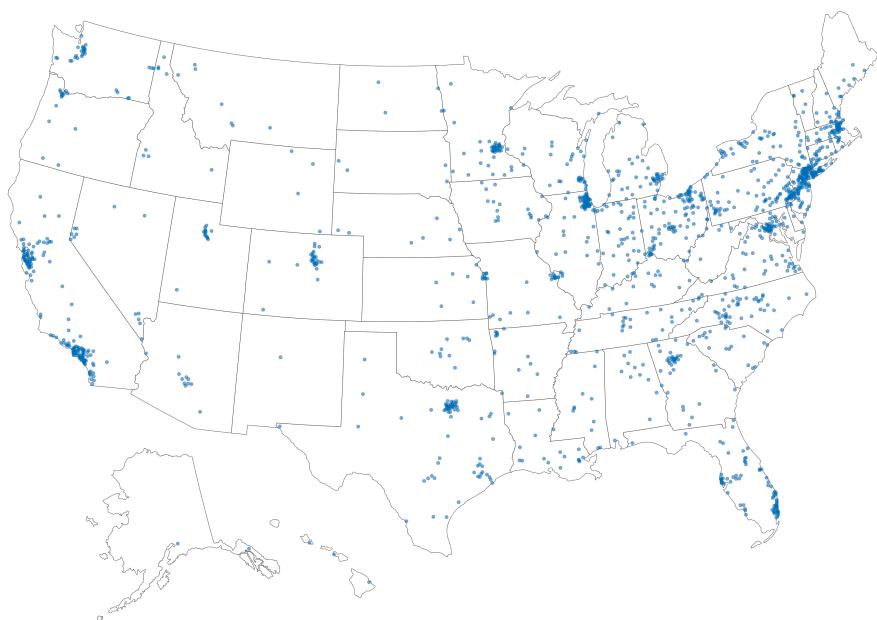


Figure 2.5: Plotting cities where the company is located

2.2.1 종속변수와 연속형 설명변수

종속변수 BL은 기업의 청산 및 파산 여부를 나타내는 범주형 변수이다. 설명변수가 연속형 변수인 경우 box plot을 통해 두 변수 사이의 상관성을 시각적으로 분석하였다. 이후 변수 내 집단의 분산이 동일하다는 귀무가설을 F검정을 이용하여 검정하였으며, 이 결과를 바탕으로 평균의 차이에 대해 T검정(귀무가설: 평균이 동일하다)을 실시하였다. Figure 2.6은 최종 모형에 포함된 연속형 변수 aco, aqpl1, caps, csho, cstk, glcea, optfvgr, spced, 그리고 stkcpa와 종속변수 BL 사이의 box plots을 나타낸 것이다. 그림을 통해 기업의 파산 및 청산 여부에 따른 연속형 변수 평균의 차이가 존재하는지를 시각적으로 살펴보고자 하였다. 모형 적합 전 1차 변수 선택과정을 거치며 종속변수와 유의한 상관성이 있는 연속형 변수만 선별하였기 때문에 box plots 상에서도 대부분 뚜렷한 차이를 파악할 수 있었다.

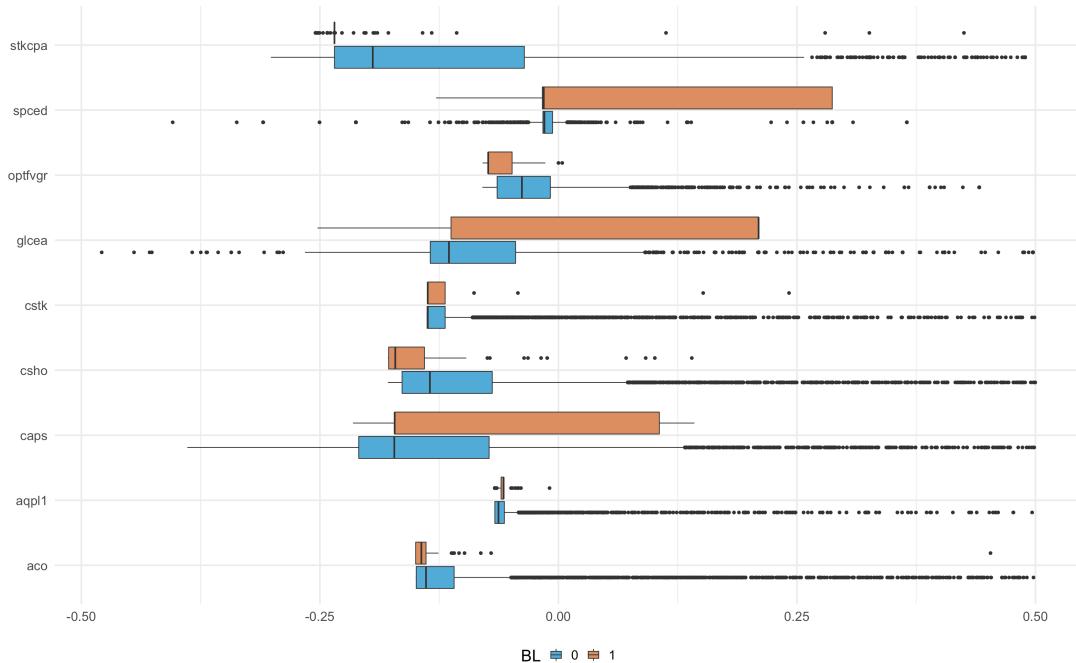


Figure 2.6: Box plots of BL (truncated)

2.2.2 종속변수와 범주형 설명변수

범주형 설명변수의 경우 교차분석을 진행했다. 종속변수와 범주형 설명변수에 대한 분할표를 만들고 독립성검정을 진행하였다. 즉 종속변수 BL과 범주형 설명변수가 확률적으로 독립이어서 아무런 연관 관계가 없는지를 파악하고자 하였다. Pearson 카이제곱 검정과 Fisher 정확 검정 두 가지 방향으로 분석을 진행하였다. 두 검정 모두 ‘두 변수가 독립이다’라는 귀무가설을 상정한다. 다음 Table 2.5는 종속변수 BL과 범주형

설명변수 stalt, idbflag 간 교차표를 나타낸다.

STALT			IDBFLAG				
BL	1	0	Total	Both	Domestic	Total	
	1	4	72	1	0	76	
	0	16	7288	0	445	6859	
Total	20	7360	7380	Total	445	6935	7380

Table 2.5: Contingency tables

한편 Pearson 카이제곱 검정과 Fisher 정확 검정의 검정통계량은 다르다. 전자의 경우 이름에서 알 수 있듯 카이제곱 분포를 따르는 통계량을 이용하는 반면, 후자는 초기하 분포를 따르는 통계량을 이용한다. 그러나 두 검정의 귀무가설의 생김새는 다르지만 의미하는 바는 동일하다. 즉 Pearson 카이제곱 검정과 Fisher 정확 검정의 귀무가설은 모두 ‘두 집단은 독립이다’를 뜻한다. 이와 관련된 자세한 내용은 아래에 설명하였다.

종속변수 BL과 범주형 설명변수 stalt의 검정 결과 귀무가설을 기각하여, 두 변수가 연관성이 있다고 판단하였다. 또한 범주형 설명변수 idbflag도 검정 결과 귀무가설을 기각하여, 종속변수 BL과 연관성이 있다고 판단할 수 있었다. Table 2.6은 두 변수의 검정 결과를 기술한 것이다.

	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 2.6: Result of tests

2.3 설명변수 간 상관분석

회귀분석 시 설명변수 간 독립성을 확보하는 것은 중요하다. 이 가정은 회귀분석을 통해 구해지는 최소 제곱해(Least Square Solution)의 유일성을 보장하기 위하여 필수적이며, 가정이 어긋나는 경우 다중공선성(multi-collinearity) 문제를 일으킨다.² 다중공선성이 존재하는 경우 회귀모형이 F검정을 통해서는 유의하게 나타나지만, 모든 변수는 T검정 상 유의성을 확보하지 못한다. 또한 회귀분석을 통해 추정한 계수의 부호나 값이 선행연구나 탐색적 자료분석에서 추정한 값과 다르게 나타난다. 따라서 자료분석을 하는 과정에서 독립변수 간 상관성을 분석하는 작업도 매우 중요하며, 본 절은 설명변수 사이의 상관분석을 진행하고

²Chatterjee, S. and Hadi, A.S. (2012) Regression Analysis by Example. 5th Edition, Wiley, New York, 98.

자 한다.

2.3.1 연속형 설명변수

연속형 설명변수 사이의 상관분석을 위해 상관계수를 파악하였다. 특히 각 변수는 정규분포를 따르지 않기 때문에 Spearman 상관계수를 이용하였다. 아래의 Figure 2.7은 연속형 설명변수 간 상관계수를 요약한 행렬이다. 행렬 내부 ×표시는 유의수준 0.05 하에서 유의하지 않다고 판단한 것이다. 한편 stkcpa와 aco, caps와 csho 상관계수가 0.39로 가장 높게 나왔으나, 이는 두 변수가 완전한 선형종속 관계라고 진단할 수준은 아니라고 판단할 수 있다. 이처럼 모든 연속형 설명변수의 조합에서 선형 상관성이 두드러지지 않게 나타난 까닭은 1차 변수선택 과정에서 VIF를 이용하여 다중공선성의 진단을 진행했기 때문이다.

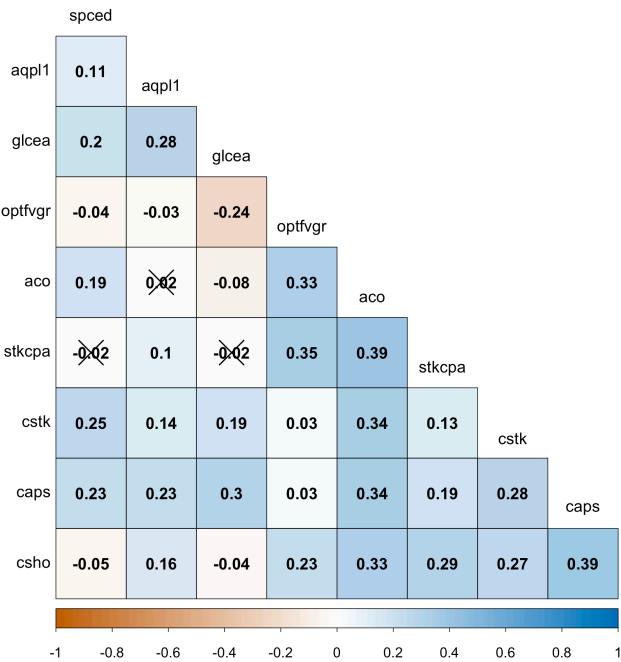


Figure 2.7: Spearman correlogram with significance test

2.3.2 연속형 설명변수와 범주형 설명변수

연속형 설명변수와 범주형 설명변수 사이의 상관분석은 box plot을 그려 두 변수 사이의 상관성을 시각적으로 파악하였다. 이후 변수 내 집단의 분산이 동일하다는 귀무가설을 F검정을 이용하여 검정하였으며, 이 결과를 바탕으로 평균의 차이에 대해 T검정(귀무가설: 평균이 동일하다)을 실시하였다. 한편 최종 모형 내

stalt와 idbflag, 두 가지 범주형 설명변수가 포함되었다. 이들과 9가지 연속형 설명변수 사이의 box plot은 Figure 2.8, 2.9에 제시하였다. 연속형 변수 stkcpa를 제외하고는 범주형 변수 stalt와 상관성이 의심되는 변수는 없었다. 한편 범주형 변수 idbflag의 경우 연속형 변수 stkcpa, optfvgr, 그리고 csho 정도가 상관성이 있을 것으로 추측할 수 있었다.

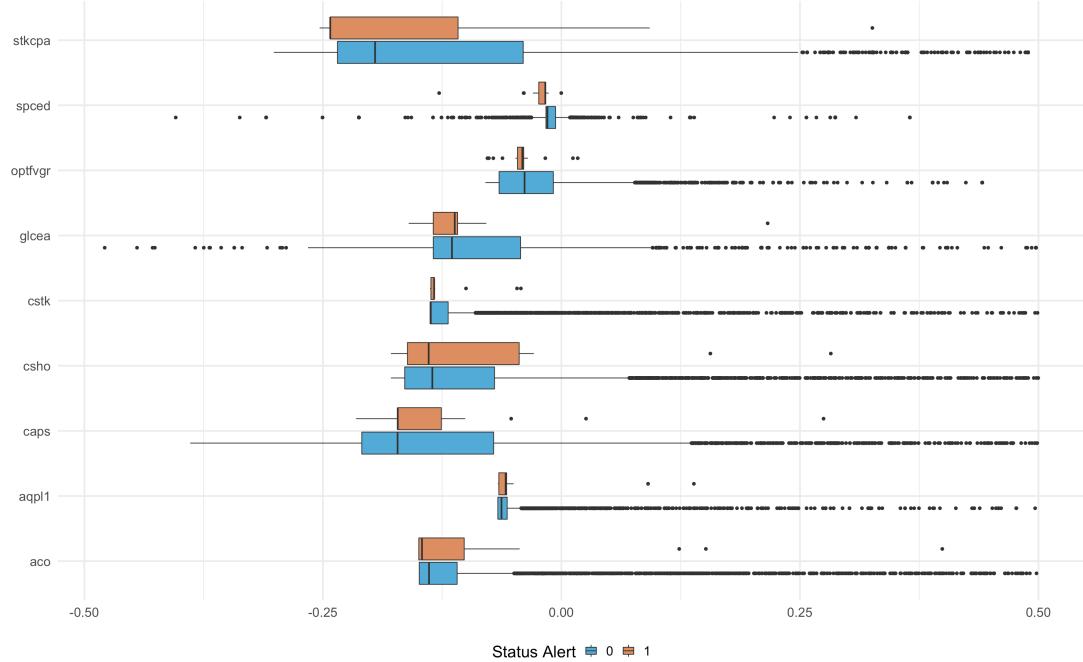


Figure 2.8: Box plots of stalt (truncated)

2.3.3 범주형 설명변수

범주형 설명변수 간 연관성을 교차분석을 통해 실시하였다. 두 범주형 설명변수의 분할표를 나타낸 후 독립성검정을 시행하였다. 즉 두 변수가 확률적으로 독립이어서 아무런 연관관계가 없는지 통계적으로 파악해 보고자 노력하였다. 종속변수의 상관분석과 마찬가지로 Pearson 카이제곱 검정과 Fisher 정확 검정을 실시하였다. 두 검정의 귀무가설은 ‘두 변수가 독립이다’로 동일하다.

Table 2.7, 2.8는 범주형 설명변수 stalt와 idbflag 간 교차표와 검정 결과를 요약한 것이다. Pearson 카이제곱 검정과 Fisher 정확 검정 결과 귀무가설을 기각하지 못하였다. 두 검정 모두 p-값이 1에 매우 가깝게 나타났으며, 두 변수가 독립임을 기각할 수 없었다. 즉 두 변수의 연관성을 통계적으로 주장할 수 없다고 판단하였다.

요약하면 연속형 설명변수 사이의 상관계수를 파악하여 다중공선성의 존재를 확인해보고자 하였는데,

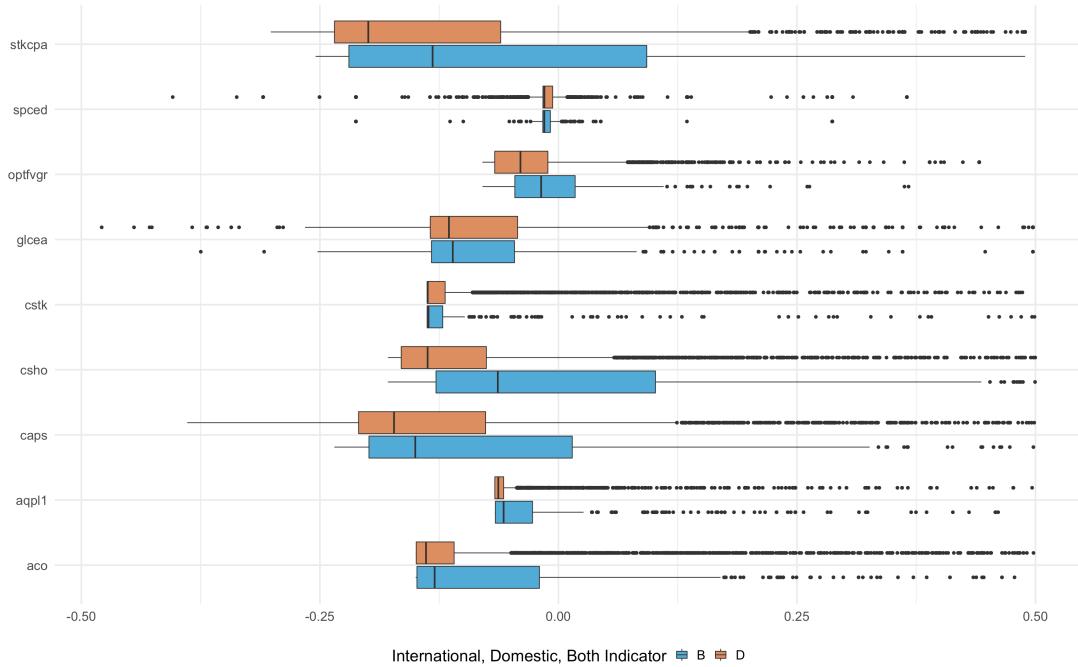


Figure 2.9: Box plots of idbflag (truncated)

IDBFLAG			
STALT			Total
	Both	Domestic	
	1	19	20
0	444	6916	7360
Total	445	6935	7380

Table 2.7: Contingency table

완전한 선형관계가 존재한다고 할 수 있을 만큼 강한 선형 상관성을 나타나지 않았다. 연속형 설명변수와 범주형 설명변수의 경우 box plot을 통해 상관성을 시각적으로 파악해보고자 노력하였다. 상관성이 있는 것으로 보이는 변수 조합이 있었지만 평균의 차이가 엄청나게 큰 편은 아니어서 문제가 될 것으로 판단하지는 않았다. 한편 범주형 설명변수 간 Pearson 카이제곱 검정과 Fisher 정확 검정을 시행하였다. 그 결과 두 범주형 설명변수가 독립이라고 나타나 연관성이 크지 않음을 확인할 수 있었다.

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 2.8: Result of tests

3

Modeling

3.1 MODEL FITTING

3.1.1 SUBSAMPLING WITH IMBALANCED DATA

과잉 적합(overfitting)은 더 나쁜 예측을 유도한다. 통계 모델은 과거의 관측 사실에 적합하는 방식으로 추정되는데, 이 적합성은 지나치게 느슨할 수도 있고, 지나치게 빽빽할 수도 있다. 전자를 과소적합(underfitting), 후자를 과잉적합이라고 한다. 과잉적합의 경우 자료에 내재한 본질적 구조를 발견하지 못하고 자료 속 소음에 적합하도록 모형을 맞춘다. 과잉적합 모델은 대부분의 통계 테스트에서 더 나은 점수를 기록하지만, 실제 현실을 설명하는 데는 적정적합 모델보다 훨씬 덜 정확하게 된다.¹ 과잉적합을 피하기 위해서는 모든 데이터를 모형 적합에 활용하는 대신 자료를 훈련용 데이터(Training set)와 테스트 데이터(Test set)로 나누어 분석을 진행해야 한다. 훈련용 데이터를 바탕으로 모형 적합을 진행하고, 테스트 데이터를 통해 모형의 성능을 평가하며 과잉적합을 경계해야 한다.

불균형 데이터(Imbalanced data)는 자료 분석에 어려움을 초래한다. 우선 대다수의 전통적 통계 모형은 자료 내 계층이 대칭적으로 분포한다는 전제를 두고 있다. 그러나 대부분의 현실 자료는 비대칭적 양상을 보이는데, 비대칭성이 극단적일 경우 rare events problem이나 imbalanced data problem으로 지칭하는 문제를 일으킨다. 또한 실제 현실에서는 극히 희귀한 사건이 분석의 주제가 되는 경우가 많다. 희귀한 사건이란 상당히 낮은 빈도수로 발생하는 사건으로, 기업의 파산이나 부정 신용카드 거래, 지진과 같은 예시가 있다. 따라서 불균형 데이터를 활용한 자료 분석 시 추가적인 주의가 필요하며, 자료를 올바르게 분석하기

¹ 네이트 실버. (2014). 신호와 소음. 더 케스트, 223–268.

위해서는 가중 로지스틱 모형(Weighted logistic model)을 활용하거나² 자료를 이단추출(subsampling)하는 방법을 고려할 수 있다.³ 한편 두 방법은 실제 양성을 양성이라고 올바르게 진단하는 비율과 실제 음성을 양성이라고 진단하는 비율 모두를 증가시키게 된다. 즉 2종 오류가 감소하는 반면, 1종 오류는 증가하게 된다. 기업의 파산 및 청산의 경우, 실제로 파산하게 될 기업을 파산하지 않는다고 예측하는 것이 반대의 경우보다 위험하다고 판단하였다. 따라서 1종 오류가 증가하는 한계점에도 앞선 방법을 통해 불균형 자료가 일으킬 수 있는 문제를 해결하고자 노력하였다.

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

where w represents proportion of events in the population.

기업의 파산 및 청산은 매우 드문 사건이어서 자료 내 비대칭성이 상당히 존재하였다. 본 연구는 E. K. Laitinen & T. Laitinen(2000)의 선행연구를 참고, 불균형 자료의 이단추출 방식으로 문제를 해결하고자 노력하였다. 모형 적합에 사용할 자료 내 파산 및 청산 기업(BL=1)과 파산 및 청산하지 않을 기업(BL=0)의 비율을 1:10으로 설정하였다. 즉 BL=1인 기업은 총 76개였는데, BL=0인 기업 760개를 임의로 추출하였다. 또한 선별한 자료를 적합에 사용할 훈련용 데이터와 성능 평가에 사용할 테스트 데이터로 무작위로 분류하였다. 훈련용 데이터는 BL=1인 기업 60개와 BL=0인 기업 600개로 구성하였고, 테스트 데이터는 BL=1인 기업 16개, BL=0인 기업 160개로 구성하였다.

3.1.2 STEPWISE SELECTION BY AIC

훈련용 데이터를 바탕으로 모형을 적합하였다. 한편 모형 선택은 AIC(Akaike information criteria) 통계량을 기준으로 진행하였다. AIC는 모형 선택 과정에서 적합도(fit)와 단순함 사이 균형을 유지할 수 있게끔 도와주는 척도이다. AIC가 작을수록 우수한 모형이라고 판단하는데, 식에서 알 수 있듯 변수의 수가 많을수록 AIC가 높아진다. 특히 AIC는 축소형 모형(nested model)이 아닌 것들 사이의 비교를 가능하게 하는 장점이 있다. 예를 들어 설명변수가 X_1, X_2, X_3 인 모형과 X_4, X_5 인 모형의 적합성 비교는 F검정으로는 불가능하지만, AIC를 기준으로는 가능하다. 적합성 관점에서 AIC가 2 이하로 차이가 나는 모형은 큰 차이가 없다고 판단하며, 그 이상의 차이에서는 AIC가 작은 모형이 더 우수한 모형이다.⁴

²Maalouf, M., and Siddiqi, M. (2014). Weighted logistic regression for large-scale imbalanced and rare events data. *Knowledge-Based Systems*, 59, 142–148.

³Laitinen, 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

⁴S. and Hadi, A.S. (2012). *Regression Analysis by Example*. 5th Edition. Wiley, New York.

모형의 변수 선택은 단계식 선택 방법(Stepwise Selection Method)를 활용하였다. 단계식 선택 방법은 전진 선택법(Forward Selection)에 후진 소거법(Backward Elimination)을 결합한 것으로서, 매 단계 선택과 제거를 반복하면서 중요한 변수를 찾아내는 방법이다. 이 방법은 중요한 변수를 하나씩 추가로 선택하면서 이미 선택된 변수들이 제거될 수 있는지를 단계마다 검토한다.⁵ 즉 초기 단계에서 모형에 포함된 변수는 이후의 단계에서 소거될 수도 있다.⁶ 본 연구는 단계식 선택 방법을 통해 가장 낮은 AIC를 나타내는 모형을 최종 모형으로 판단하였다. 즉 현재 단계에서 추가로 변수를 더하거나 제외하여도 AIC를 더는 낮출 수 없을 때 변수 선택이 완료된 것으로 생각하였다. 한편 AIC를 기준으로 변수 선택을 고려하였기 때문에, 개별 변수의 통계적 유의성은 고려하지 않았다. 모형 적합에는 앞의 모든 과정을 거친 44개의 변수, 660개의 관측치가 포함된 훈련용 데이터를 활용하였다. Table 3.1은 모형 적합에 사용된 변수를 정리한 것이다.

variable	type	variable	type	variable	type	variable	type
aco	numeric	dvc	numeric	naics2	factor	spced	numeric
aqpl1	numeric	emp	numeric	nopio	numeric	stalt	factor
bkvlp	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	optprcw	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 3.1: 44 variables before stepwise selection

적합 결과 9개의 연속형 설명변수와 2개의 범주형 설명변수, 총 11개의 설명변수가 최종 모형에 포함되었다. 유의수준 0.05 하 7개의 설명변수 aqpl1, caps, csho, glcea, optfvgr, stalt, 그리고 stkcpa가 통계적으로 유의하다고 나타났다. 최종 모형의 계수에 대한 정보는 Table 3.2에 간략히 기술하였다.

따라서 최종 모형은 다음과 같다.

$$\begin{aligned} \log \frac{p_i}{1 - p_i} = & -203.6 - 7.19x_{aco,i} + 8.9x_{aqpl1,i} + 2.12x_{caps,i} \\ & - 7.58x_{csho,i} - 13.51x_{cstk,i} + 1.56x_{glcea,i} + 197.01x_{idbflag_D,i} \\ & - 19.96x_{optfvgr,i} - 2.42x_{spced,i} + 2.66x_{stalt_1,i} - 3.31x_{stkcpa,i}. \end{aligned}$$

⁵김기영 외. (2009). 예제로 배우는 SAS 데이터 분석 입문. 자유아카데미.

⁶S. and Hadi, A.S. (2012). *Regression Analysis by Example*. 5th Edition. Wiley, New York.

	Estimate	Std. Error	<i>z</i> value	<i>P</i> (> <i>z</i>)
(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 3.2: Coefficients

한편 p_i 에 대해 정리하면 다음과 같다.

$$p_i = (1 + \exp(203.6 + 7.19x_{aco,i} - 8.9x_{aqpl1,i} + \dots + 3.31x_{stkcpa,i}))^{-1}.$$

모형에 대한 해석은 다음과 같다. 개별 설명변수에 대한 해석은 다른 설명변수가 모두 고정된 상황, *ceteris paribus*를 전제한다. 다른 변수가 고정된 상황에서 대차대조표상 현금, 현금등가액, 미수금, 그리고 재고 이외의 자산(aco)이 100만 달러 증가할수록 로그 승산비는 7.19 감소, 즉 파산 및 청산 예측 확률은 3.793×10^{-89} 감소한다. 자산이 늘어날수록 기업의 파산 및 청산 가능성성이 작아진다는 해석은 합리적이다. 연말 시점에 시장에 유통되고 있는 보통주의 수(csho)가 100만 달러 늘어날수록 로그 승산비는 7.575만큼 감소하며, 이는 파산 및 예측 확률은 3.795×10^{-89} 감소와 같다. 또한 모든 일반 자본의 총 액면가(cstk)가 100만 달러 증가할수록 로그 승산비는 13.510만큼 감소한다. 이는 기업의 파산 및 청산 확률 3.796×10^{-89} 감소와 같다. 증권시장 내 기업의 보통주 유통 증가와 자본의 액면가 증가는 기업의 자금 조달 흐름을 원활하게 하여 파산 및 청산 확률을 낮추게 된다. 기업 활동이 미국 내에서만 이루어지는 기업 (idbflag_D)은 미국 외에서도 활동하는 기업에 비해 로그 승산비가 197.005 높았다. 이는 활동이 미국 내에 한정된 기업이 그렇지 않은 기업에 비해 파산 및 청산 확률이 0.137% 높다고 해석할 수 있다. 한 해 동안 허가된 옵션 공정가치의 가중 평균(optfvgr)이 100만 달러 증가할수록 로그 승산비는 19.957 감소하는 것으로 나타났다. 이는 기업의 파산 및 청산 확률이 3.796×10^{-89} 감소한다는 것과 동일하다. 옵션의 가치가 높을수록 기업의 파산 및 청산 가능성이 줄어든다는 해석은 합리적이다. S&P 기준에서의 회석된 주당 순이익(spced)의 1달러 증가는 로그 승산비의 2.42 감소를 의미한다. 이는 기업의 파산 및 청산 확률의 3.459×10^{-89} 감소하는 것

과 동일하다. 주당 순이익의 증가는 기업의 경영 활동이 원활하게 이루어진다고 해석할 수 있으며, 경영이 건강하게 이루어질수록 기업의 파산 및 청산 확률은 낮아진다. 또한 순익계산서상 비용 처리된 주식기준보상액(stkcpa) 100만 달러의 증가는 로그 승산비를 3.308 감소시킨다. 이는 파산 및 청산 확률 3.658×10^{-89} 감소와 동일하다. 주식 보상금의 증가는 기업의 경영 상태가 개선되었다고 해석할 수 있어 파산 및 청산 가능성과 반대 방향으로 움직인다. 마지막으로 해당 기업이 파산 위험에 빠졌거나 기업 인수 진행 중인 경우(stalt1)는 그렇지 않은 기업에 비하여 파산 및 청산 가능성이 5.064×10^{-88} 높은 것으로 드러났다. 즉 기업의 상황이 부정적인 방향으로 급변하는 양상은 실제로도 파산 및 청산과 관련이 있는 것으로 나타났다.

한편 공정가치로 평가된 자산(aqpl1)이 100만 달러 증가할수록 로그 승산비는 8.901 증가한다. 즉 파산 및 예측 확률은 2.786×10^{-85} 상승한다고 해석할 수 있다. 자본잉여금(caps) 100만 달러의 증가는 로그 승산비를 2.117 증가시키며, 이는 기업의 파산 및 청산 확률 2.774×10^{-88} 증가와 같다. 다음으로 S&P의 계산에서는 포함되지 않은 매출액 증감의 세후 금액(glcea)이 100만 달러 증가할수록 로그 승산비는 1.56증가, 파산 및 청산 확률은 1.429×10^{-88} 증가한다. 즉 해당 세 변수의 파산 확률에 대한 효과는 매우 미미하지만, 기존 상식과 반대되는 해석을 내놓았다. 자산, 자본, 그리고 매출액의 증가는 기업의 재무 상태를 건강하게 하여 파산 및 청산 확률을 낮추는 방향이 올바르다. 다음과 같은 한계점은 모형 적합 과정에서 자료의 이단추출을 시행한 것과 AIC를 기준으로 변수 선택이 이루어진 점 등에 기인한다고 판단했다.

마지막으로 최종 모형을 대상으로 우도비 검정을 시행하였다. 이는 적합한 최종 모형과 설명변수가 모두 제거된 모형의 적합도를 비교하는 절차이다. 모든 회귀계수가 0이라는 귀무가설(모든 설명변수의 설명력이 없다)을 상정한 후, 각 모형의 우도 비교를 통해 가설을 통계적으로 검정하였다. 우도비 검정 결과 p-값이 1.21×10^{-14} 로 매우 작게 나타나 귀무가설을 기각할 수 있었다. 즉 적어도 하나의 설명변수는 종속변수에 설명력을 지니고 있다고 판단할 수 있었다.

3.2 MEASURING PERFORMANCE

모형의 성능 평가는 테스트 데이터를 바탕으로 이루어진다. 테스트 데이터는 BL=1인 기업 16개와 BL=0인 기업 160개로 이루어져 있다. ROC 곡선을 그려 최적 임계값을 파악하고, AUC를 통해 모형의 우수성을 판단하였다. 또한 혼동 행렬을 바탕으로 분류 정확도 등을 파악하는 방식으로 모형의 성능을 평가하였다.

3.2.1 ROC CURVE

ROC곡선은 임계값 변화에 따른 민감도와 1-특이도의 양상을 나타낸 것이다. 특히 1-특이도는 위양성 비율(False Positive Rate)을 나타낸다. 민감도와 특이도가 1에 가까울수록 우수한 모형이라고 판단하며, 따라서 ROC곡선과 (1-특이도, 민감도)값이 (0,1)인 좌표가 가장 가까운 곳에서 최적 임계점이 결정된다. 한편 로지스틱 모형으로부터 도출된 ROC 곡선은 전반적인 모형의 성능을 평가하게 된다. ROC 곡선 아래 면적을 AUC라 하는데, AUC가 클수록 우수한 모형으로 판단한다. AUC가 0.5인 경우 무작위 추정과 다름없는 수준의 모형이며, 1이면 완벽한 모형이다. 일반적으로 AUC가 0.8 이상이면 우수한 모형으로 판단한다.

Figure 3.1은 최종 모형의 ROC 곡선을 나타낸 것이다. ROC 곡선상 최적 임계점은 0.1267로 나타났다. 이는 예측된 파산 및 청산 확률이 0.1267을 초과하는 경우 해당 기업을 파산 및 청산으로 분류할 때 민감도와 적합도가 가장 우수하다는 것을 의미한다. 따라서 최적 임계점 기준에서, 추정된 파산 및 청산 확률이 0.1267를 초과하는 경우, 해당 기업은 파산 및 청산 기업으로 분류된다. 반대로 추정 확률이 0.1267 이하의 경우, 해당 기업은 파산 및 청산하지 않을 것으로 예측된다. 한편 AUC는 0.857로 나타났다. 이를 바탕으로 최종 모형이 비교적 우수하다고 판단할 수 있었다.

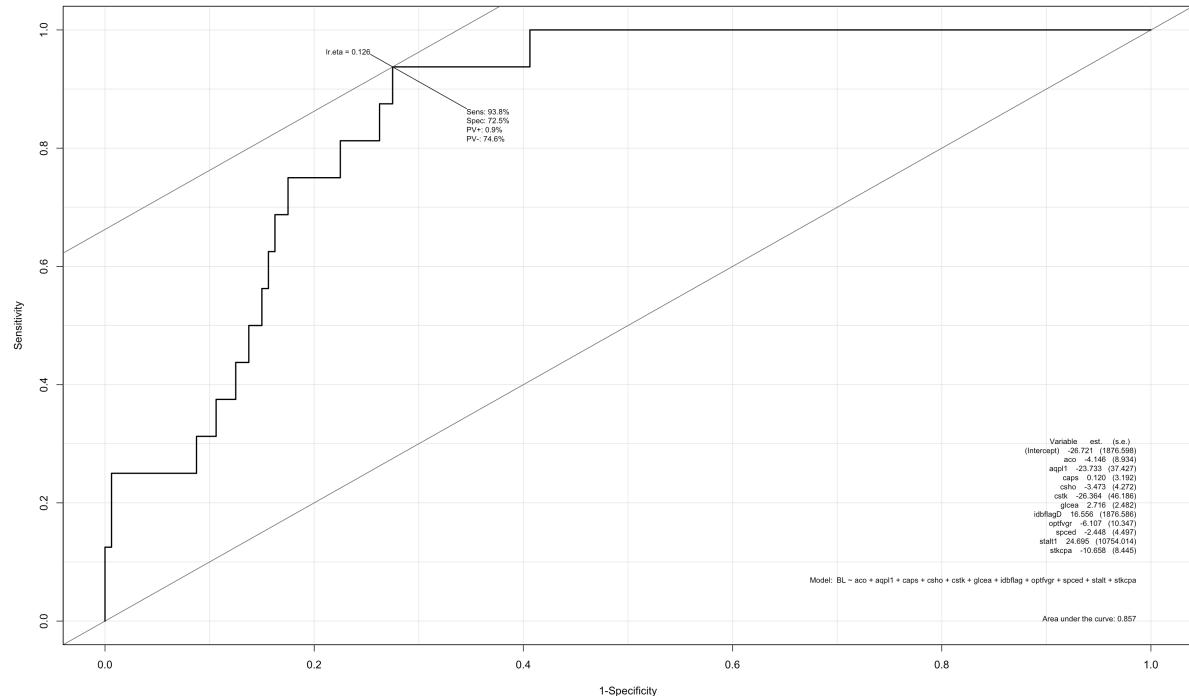


Figure 3.1: ROC curve

3.2.2 CONFUSION MATRIX

시작에 앞서 용어를 정리하고자 한다. 특히 혼동 행렬 내 Positive, Negative는 양성과 음성으로 해석할 수 있어 분석에 혼란을 일으킬 수 있다. 따라서 용어를 기업의 파산 및 청산의 관점으로 새롭게 해석함으로써 이후의 과정에서 발생할 수 있는 혼란을 최소화하고자 노력하였다. 예측 모형의 분류 결과는 다음의 네 가지로 정리할 수 있다. TP는 예측 모형에서 파산 및 청산으로 분류한 기업이 실제로 파산 및 청산한 경우를 의미하며, TN는 모형이 파산 및 청산하지 않을 것으로 분류한 기업이 실제로도 파산 및 청산하지 않은 경우를 나타낸다. 한편 FP는 모형이 특정 기업을 파산 및 청산할 것으로 예측했으나 실제로는 파산 및 청산하지 않은 경우이며, FN은 모형이 파산 및 청산하지 않을 것으로 예측하였으나 실제로는 파산 및 청산을 한 경우로서, 2종 오류에 해당한다. 특히 예측 시 파산 및 청산할 기업을 그렇지 않을 것으로 분류하는 사례는 매우 중대한 오류이다. 예를 들어 예측 모형을 바탕으로 투자 등의 사업 결정이 이루어진다면 해당 오류는 엄청난 손실을 초래할 수 있다. 앞서 언급한 것처럼 본 연구는 이를 중점에 두고 분석을 진행하였다.

혼동 행렬을 통해 민감도와 특이도를 파악할 수 있다. 민감도는 양성 예측이 올바르게 이루어지는 비율(TPR: True Positive Rate)을 의미하며, 특이도는 음성 예측이 올바르게 이루어지는 비율(TNR: True Negative Rate)을 의미한다. 특정 임계값에서 민감도와 특이도가 1에 가까울수록 모형의 분류 성능이 우수하다고 판단한다.

다음의 Table 3.3은 최적 임계점 0.1267을 기준으로 혼동 행렬을 나타낸 것이다. 최종 모형은 16개의 실제 파산 및 청산 기업 중 14개의 기업을 파산 및 청산 기업으로 분류하였다. 또한 160개의 파산 및 청산하지 않은 기업 중 117개를 올바르게 분류하였다. 즉 모형의 민감도는 87.5%이며, 특이도는 73.125%였다. 민감도와 특이도가 모두 1에 가까울수록 우수한 모형으로 판단하는데, 두 수치 모두 비교적 우수하게 나타났다. 한편 모형이 올바르게 분류한 비율을 나타내는 정확도(accuracy)는 74.43%였으며, 오분류율은 25.57%였다. 파산 및 청산으로 분류한 기업 중 실제 파산 및 청산한 기업의 비율을 의미하는 정밀도(precision)는 24.56%로 나타났다. 파산 및 청산하지 않은 기업을 파산 및 청산할 것으로 분류한 비율은 26.875%였다. 즉 최종 모형이 파산 및 청산으로 판단하는 사례가 실제의 사례보다 빈번했다. 하지만 앞서 언급했듯 해당 한 계점은 파산 및 청산할 기업을 그렇지 않다고 예측하는 것이 반대의 경우보다 더 위험하다고 판단했기 때문에 나타난 결과라고 해석할 수 있다.

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

Table 3.3: Confusion matrix with cut-off 0.1267

4

Conclusion

4.1 분석 결과

1개의 종속변수와 11개의 설명변수, 총 12개 변수가 최종 모형에 포함되었다. 종속변수 BL은 2011년부터 2013년까지 기업이 파산 및 청산한 경우 1, 그렇지 않으면 0을 나타내는 범주형 변수이다. 한편 설명변수는 9개의 연속형 변수와 2개의 범주형 변수로 구성된다. 대차대조표상 현금, 현금등가액, 미수금, 또는 재고 등에 포함되지 않는 자산(aco), 공정가치로 계산된 자산(aqpl1), 자본잉여금(caps), 연말 시점 시장에 유통되고 있는 보통주의 수(csho), 일반 자본의 총 액면가(cstk), S&P 계산에서 제외된 매출액 증감의 세후 금액(glcea), 한 해 동안 허가된 옵션 공정가치의 가중 평균(optfvgr), S&P 기준 희석된 주당 순이익(spced), 손익계산서상 비용 처리된 주식기준보상액(stkcpa)이 연속형 설명변수이며, 자료의 출처를 나타내는 변수(idbflag)와 기업이 파산 위험에 빠졌거나 인수 진행 중임을 나타내는 변수(stalt)가 범주형 설명변수이다. 이들 중 aco, csho, cstk, optfvgr, spced, 그리고 stkcpa 변수는 기업의 파산 및 청산 확률과 반대 방향으로 변동하는 반면, aqpl1, caps, 그리고 glcea는 기업의 파산 및 청산 확률과 같은 방향으로 변동하는 것으로 나타났다. 한편 다른 조건이 고정된 상태에서, 기업 활동이 미국 내에 국한된 경우(idbflag=D)는 미국 외 국가에서도 활동하는 기업에 비해 파산 및 청산 확률이 높은 것으로 드러났으며, 현재 기업이 파산 위험에 처했거나 인수가 진행 중인 경우(stalt=1)는 그렇지 않은 기업에 비해 파산 및 청산 확률이 높은 것으로 나타났다.

모형의 성능은 테스트 데이터를 활용하여 ROC 곡선과 혼동 행렬을 나타냄으로써 평가하였다. ROC 곡선을 통해 최적의 임계값과 AUC를 도출하였다. 최적 임계값은 0.1267이었는데, 이는 파산 및 청산으로 분

류하는 기준을 0.1267로 설정했을 때 민감도와 적합도가 가장 우수하다는 것을 의미한다. 한편 곡선 아래의 면적을 의미하는 AUC는 0.857이었는데, 이를 바탕으로 적합된 모형이 우수하다고 판단할 수 있다. 한편 최적 임계값을 바탕으로 혼동 행렬을 구성하였다. 민감도는 0.875, 적합도는 0.731, 그리고 정확도는 0.744였다. 한편 위양성 비율은 0.269로 비교적 높게 나타났다. 이는 파산 예측의 경우 파산할 기업을 파산하지 않을 것이라고 예측하는 경우가 반대의 경우, 즉 위양성 경우보다 심각하다고 판단하여 2종 오류를 낮추는 방향으로 분석을 진행했기 때문이다. 본 연구는 이단 추출을 바탕으로 2종 오류를 낮추었는데, 이는 어렵게도 1종 오류를 증가시키기도 했다.

4.2 향후 분석방향 제시

기존 상식과 부호가 반대되는 변수도 모형에 존재했다. 이는 회귀분석 시 다중공선성으로 인해 흔히 나타나는 현상이지만, 본 연구에서는 다중공선성의 문제를 진단하였다. 해당 한계는 모형 적합 과정에서 이단 추출을 시행한 것과 AIC를 기준으로 변수 선택이 이루어진 점 등에 기인한다고 판단했다. 따라서 이단 추출 대신 다른 방법으로 불균형 데이터의 한계를 보완하거나, 혹은 AIC가 아닌 다른 기준을 설정하여 변수 선택을 진행하는 방법으로 상식에 부합하는 결과를 도출할 수 있을 것으로 기대한다.

파산하게 될 기업을 파산하지 않는다고 진단하는 것이 반대의 경우보다 심각한 오류라고 판단했다. 따라서 본 연구는 훈련용 데이터의 수를 임의로 조정하는 방식으로 해당 오류를 줄일 수 있었다. 하지만 실제로 파산하지 않을 기업을 파산할 것으로 예측하는 오류, 즉 1종 오류가 증가하는 것을 감내해야 했다. 연구 목적에 따라 민감하게 반응하는 오류가 다를 것이며, 이에 따라 임계값 조정 역시 다르게 이루어질 것이다. 두 오류 사이의 적절한 균형에 대해서는 연구자의 신중한 판단이 요구된다.

본 연구는 현재의 자료와 적합한 모형을 바탕으로 앞으로 3년 내 파산 및 청산할 기업을 예측하고자 하였다. 그러나 2019년과 2020년 자료의 경우 최종 모형에 포함되는 변수의 관측치 모두가 결측치인 경우가 많았다. 따라서 적절한 방법으로 결측치 처리가 불가능했으며, 파산 및 청산 기업을 예측하는 단계까지 분석을 진행하지 못했다. 결측치 처리를 2000년부터 2020년에 걸친 원자료를 바탕으로 진행한다면 현재의 데이터로 파산 및 청산 기업 예측을 할 수 있으리라 기대한다.

기업의 파산 및 청산 직전 연도의 자료는 다른 어느 해보다 많은 신호를 담고 있다. 즉, 기업이 파산 및 청산 상태에 빠진다면 이에 대한 경고는 전년도 자료에 고스란히 남아있을 것이다. 또한 E. K. Laitinen & T. Laitinen(2000)의 연구에서는 파산 직전 연도의 자료를 바탕으로 분석을 진행하였을 때, 모형의 설명력이 가장 우수한 것으로 드러났다. 이외에도 많은 선행연구는 파산 전년도 자료를 바탕으로 예측 모형을 고

안하였다. 다만 본 연구에서는 기존 방향을 답습하는 것보다 새로운 방향을 시도하는 것에 의미를 두고 3년에 걸친 중장기적인 예측 모형을 도출하고자 노력하였다. 이미 증명된 바와 같이 예측력이 우수한 모형을 만들고자 할 때는 파산 직전 연도의 자료를 활용하는 것이 도움이 될 것이다.

References

- [1] S. and Hadi, A.S. (2012). *Regression Analysis by Example*. 5th Edition. Wiley, New York.
- [2] Stock J, Watson M. (2015). *Introduction to Econometrics*. 3rd edition. Pearson, Boston.
- [3] Kleinbaum, D. G. (2010). *Logistic regression: A self-learning text*. New York: Springer.
- [4] 김기영 외. (2009). 예제로 배우는 SAS 데이터 분석 입문. 자유아카데미.
- [5] 네이트 실버. (2014). 신호와 소음. 더 케스트, 223–268.
- [6] 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.
- [7] Kuruppu, N., Laswad, F., and Oyelere, P. (2003). The efficacy of liquidation and bankruptcy prediction models for assessing going concern. *Managerial auditing journal*.
- [8] White, M. J. (1989). The corporate bankruptcy decision. *Journal of Economic Perspectives*, 3(2), 129–151.
- [9] Maalouf, M., and Siddiqi, M. (2014). Weighted logistic regression for large-scale imbalanced and rare events data. *Knowledge-Based Systems*, 59, 142–148.
- [10] Kang H. (2013). The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64(5), 402–406.