이론통계학 II

Project #8. 기업 부도 예측

발표일 2021. 11. 17

1조

202STG26	박지윤
202STG27	이수현
212STG04	김이현
212STG12	박윤정

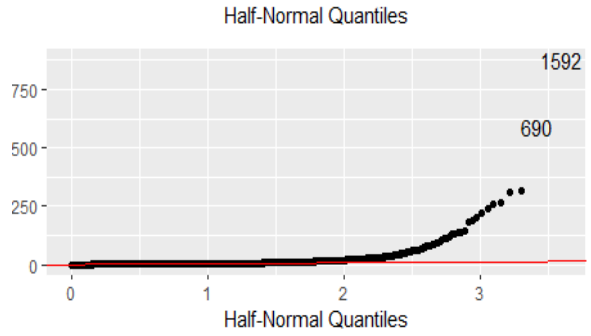
데이터 및 변수 설명

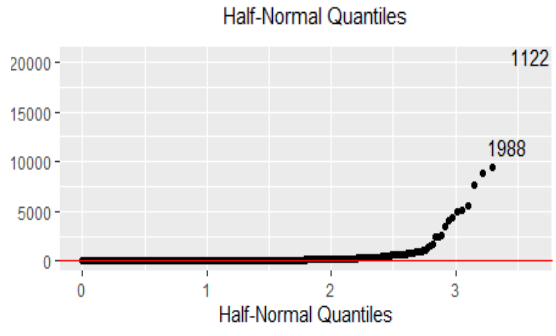
변수명	설명	변수명	설명	변수명	설명
x1	금융비용/총비용비율	x15	총차입금/(총차입금+자기자본)비율	x29	총자본순이익율
x2	고정부채비율	x16	총 CF/차입금비율	x30	총자본영업이익율
x3	고정비율	x17	CF/차입금비율	x31	총자산사업이익율
x4	부채비율	x18	순운전자본/총자산비율	x32	경영자본회전율
x5	부채총계/자산총계비율	x19	유동부채구성비율	x33	고정자산회전율
x6	순부채/총자산비율	x20	현금비율	x34	매입채무회전율
x7	유동부채비율	x21	총자산투자효율	x35	매출채권회전율
x8	유동비율	x22	매출채권증가율	x36	자기자본회전율
x9	유보액/총자산비율	x23	재고자산증가율	x37	자본금회전율
x10	자기자본비율	x24	경영자본순이익율	x38	재고자산회전율
x11	차입금의존도	x25	금융비용/총부채비율	x39	총자본회전율
x12	고정자산/차입금비율	x26	자기자본순이익율	x40	기업 나이
x13	차입금/자기자본비율	x27	자본금순이익율	x41	로그매출액
x14	고정재무비보상배율	x28	총자본경상이익율	x42	로그자산
x43	업종	x44	규모	y	부도시간(365=정상)

① Outlier(이상점) 고려

원자료의 값이 9999.99, -9999.99인 경우 결측치로 판단하여 NA로 대체하였다.

이후 ②에서 적절한 값으로 대체하고자 한다.





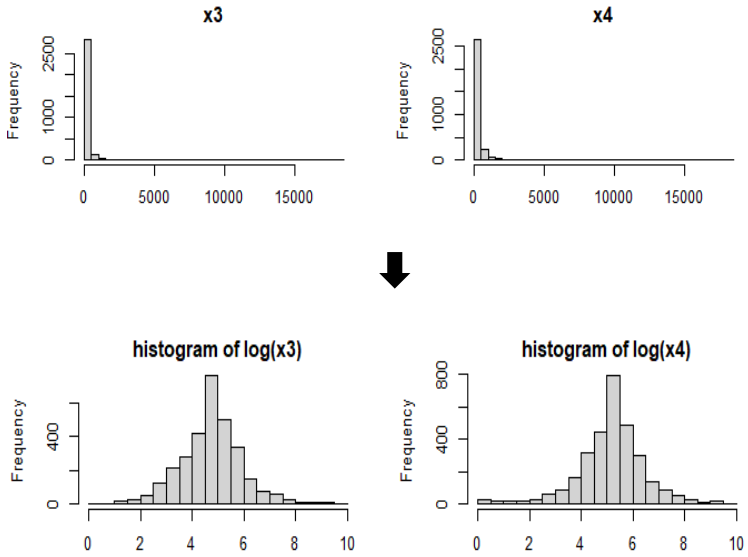
R의 halfnorm 함수를 사용하여 각 변수별 이상점을 찾아 총 34개의 행을 제거하였다.

② NA(결측치) 고려

결측치가 500개 이상인 변수의 경우, 데이터의 20%정도가 손실되므로 제거하였다. 이 때 총 3개의 변수가 삭제되었다.

이후, 각 변수별 결측치는 해당 변수의 median(중간값)으로 대체하였다.

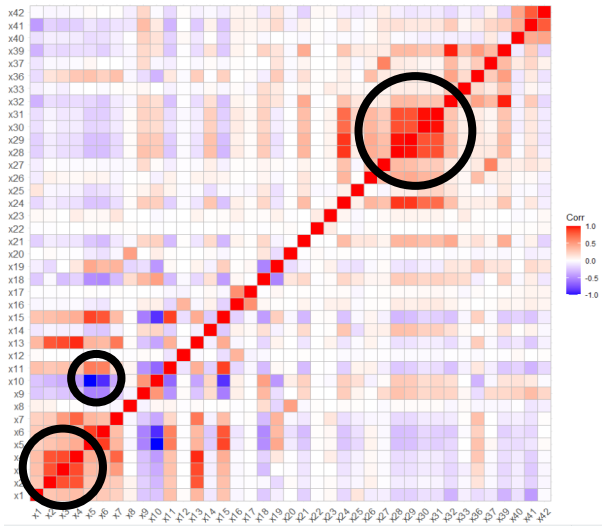
③ 변수 변환



히스토그램과 산점도를 통해 변수별 분포를 살펴본 후, 각 변수에 적합하게 log변환, 역수 변환, 제곱근 변환을 진행하였다.

log변환(11)	고정부채비율, 고정비율, …
역수변환(1)	기업 나이
제곱근변환(2)	차입금의존도, 유동부채구성비율

④ 상관관계가 높은 변수 제거



Correlation matrix를 통해 상관관계가 높아 다중공선성이 존재할 것이라고 판단되는 변수를 삭제하였다.

상관성이 높은 변수	**X2**	**X3**	**X4**	**X13**
	X5	**X6**	**X10**	
	X28	**X29**		
	X30	**X31**		
	X32	**X39**		

위와 같이 상관관계가 높은 변수들을 추출한 후 가장 중요하다고 판단되는 변수를 제외하고 삭제하였다. 최종적으로 8개의 변수가 제거되었다.

⑤ 범주형 변수 Factor화

업종(x43), 규모(x44) 변수는 범주형 변수이므로 factor로 지정해주었다.

Part 1 : Logistic GLM/GAM을 이용한 부도 예측

1. GLM을 이용한 분석

Link Function	변수 선택	선택된 변수 개수	AIC
Probit	x	44	913.25
	o	20	887.89
Logit	x	44	912.29
	o	19	887.07
Gompit	x	44	912.38
	o	20	888.37

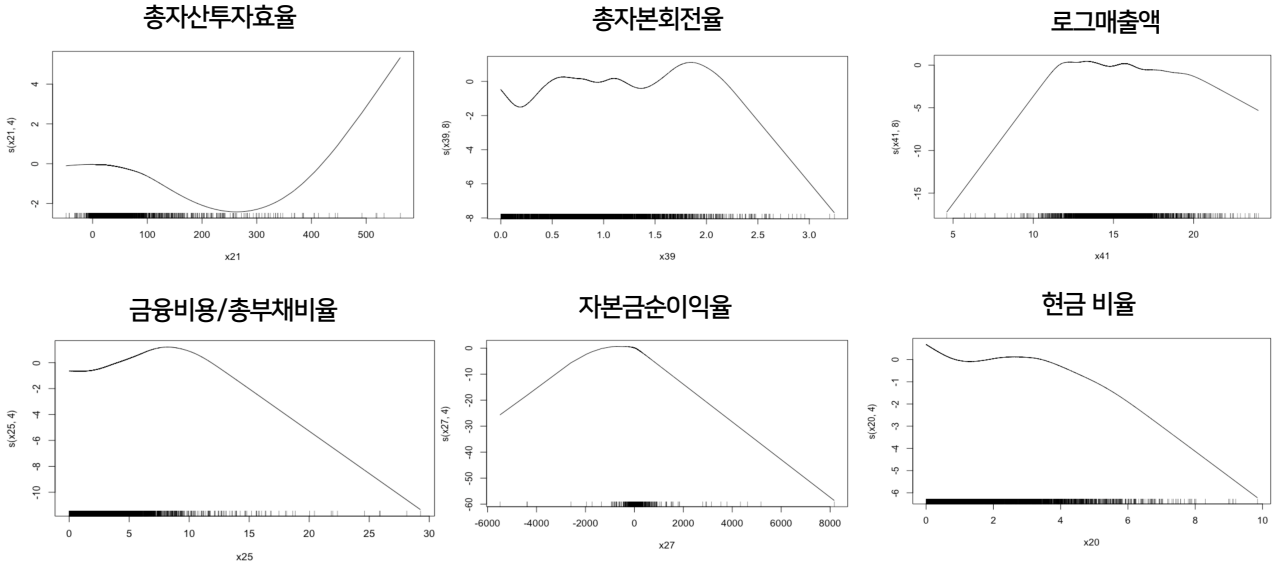
선택된 변수
금융비용/총비용비율, 유동부채비율, 고정자산/차입금비율, 순운전자본/총자산비율, 현금비율, 매출채권증가율, 경영자본순이익율, 자기자본순이익율, 총자본순이익율, 고정자산회전율, 기업 나이, 로그매출액 , 로그자산, 규모

AIC를 기준으로 Link function은 logit을 사용하고 변수 선택한 모형을 최종 예측 모형으로 선정한다. 이때의 AIC 값은 약 887이다.

2. GAM을 이용한 방법

선택된 변수	AIC
순부채/총자산비율, 유동부채비율, 유동비율, 고정자산/차입금비율, 고정재무비보상배율, 순운전자본/총자산비율, 유동부채구성비율, 현금비율, 총자산투자효율, 매출채권증가율, 경영자본순이익율, 금융비용/총부채비율, 자기자본순이익율, 자본금순이익율, 총자본순이익율, 고정자산회전율, 총자본회전율, 기업 나이, 로그매출액 , 규모	836

GAM 모형 적합 후 변수 선택을 진행한 결과, 위와 같은 변수가 선택되었고 이때의 AIC값은 836이다.



위는 GAM 모형 내에서 spline 함수를 씌운 변수의 효과를 나타낸 그래프이다.

총자산투자효율은 250 지점까지 곡선형태로 부도확률이 감소하다가 이후 지점에서는 증가하는 추세를 확인할 수 있다. 총자본회전율은 총자본이 몇 회의 매출액을 실현하였는가를 나타내는 지표이므로 높을수록 부도확률이 낮아지는 결과가 타당하다고 볼 수 있다. 또한, 로그매출액, 금융비용/총부채비율, 자본금순이익율, 현금비율 모두 값이 낮을 때 부도확률이 다소 증가하는 추세를 보이다가 어느 지점부터 지속적으로 감소하는 추세를 보여준다. 네 변수 모두 값이 클수록 부도를 막아낼 능력이 있음을 보여주는 지표이므로 그래프가 타당하다고 판단할 수 있다.

이 외 변수들은 부도확률에 선형적인 영향을 주는 것을 확인할 수 있었다.

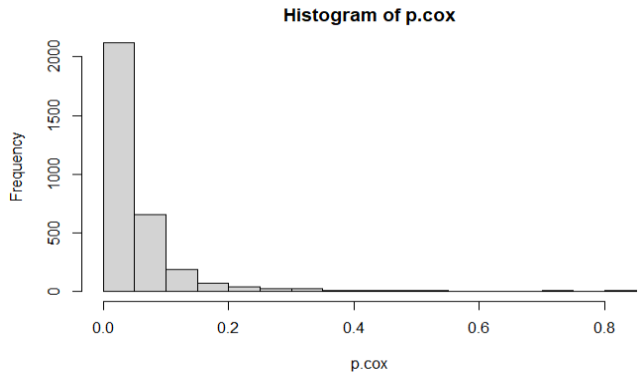
Part 2 : 생존분석 기법을 이용한 부도 예측 방법

부도 확률에 대한 최적 Cox PHM

선택된 변수	AIC
금융비용/총비용비율, 유동부채비율, 고정자산/차입금비율, 순운전자본/총자산비율, 현금비율, 매출채권증가율, 재고자산증가율, 경영자본순이익율, 금융비용/총부채비율, 자기자본순이익율, 총자본순이익율, 고정자산회전율, 기업 나이, 로그매출액 , 로그자산, 규모	1748

Cox PHM 모형을 적합한후 변수 선택을 진행한 결과, 위와 같은 변수들이 선택되었다. 이때의 AIC값은 1748로, 변수 선택하기 전보다 AIC가 작아짐을 확인하였다.

또한, 최적 Cox 모형에서 각 기업이 1년 이내 부도날 확률 score의 분포를 히스토그램에 나타낸 결과는 다음과 같다. 대부분이 0과 0.2 사이에 분포되어있다.



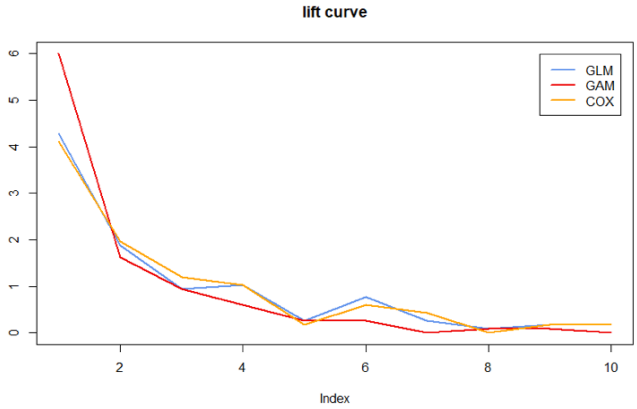
Part 3 : Lift Chart을 이용한 예측력 비교

1. **위에서 구한 최적 예측모형을 이용하여 모형별 부도확률을 구하시오. 또, 값의 크기순으로 10개 등급을 매긴 후 구간별 실제 부도율을 비교하는 표 및 Lift Chart를 겹쳐 그려보고 세 방법의 장단점을 비교 검토하시오.**

구간(상위)	GLM		GAM		CoxPHM	
	실제 부도율	Lift	실제 부도율	Lift	실제 부도율	Lift
10%	0.158	4.285	0.221	5.999	0.151	4.113
10% - 20%	0.069	1.885	0.060	1.628	0.073	1.971
20% - 30%	0.035	0.943	0.035	0.943	0.044	1.200
30% - 40%	0.038	1.028	0.022	0.600	0.038	1.028
40% - 50%	0.009	0.257	0.009	0.257	0.006	0.171
50% - 60%	0.028	0.771	0.009	0.257	0.022	0.600
60% - 70%	0.009	0.257	0.000	0.000	0.016	0.428

70% - 80%	0.003	0.086	0.003	0.086	0.000	0.000
80% - 90%	0.006	0.171	0.003	0.086	0.006	0.171
90% - 100%	0.007	0.200	0.000	0.000	0.007	0.200

3168개 기업을 부도 확률과 socre값 크기순으로 10개의 구간으로 나눈 후, 각 구간별로 실제 부도율과 Lift값을 나타낸 표이다. 세 모형 모두 상위권 구간에 해당할수록 Lift값이 높으므로 분석이 적절히 시행되었음을 판단할 수 있다.



위 표를 기반으로 Lift Chart를 그려본 결과는 좌측과 같다. 역시 세 모형 모두 상위권 구간에서 lift 값이 가장 높으며 하위 구간으로 내려갈수록 값이 감소함을 확인할 수 있다. 특히 GAM 모형의 경우, 상위 10% 구간에서 가장 높은 부도확률을 보이므로 가장 좋은 성능을 보인다고 판단할 수 있다. 또한, GLM과 COX모형은 매우 비슷한 형태를 보인다.

2. 확률 및 점수값이 상위 10%에 해당하는 기업들의 실제 부도확률을 구하시오.

모형	GLM	GAM	Cox PHM
상위 10% 구간의 부도확률(%)	15.773	22.082	15.142

3. 위에서 최종 선택한 GLM/ GAM/ Cox PHM 예측모형의 AIC값을 각각 제출하시오.

모형	GLM	GAM	Cox PHM
AIC	887	836	1748

4. 검정용 자료에 대하여 GLM/GAM 부도확률 및 Cox PHM 부도점수의 추정값과 이들 상위 10%에 해당하는 기업들의 지시변수를 구하시오.

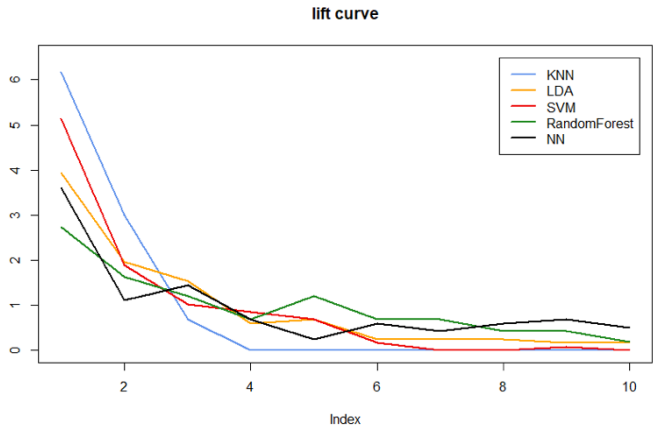
다음은 GLM, GAM CoxPHM 모형의 부도확률 및 score가 상위 10%에 해당하는 구간에는 1, 나머지는 0으로 지정하였을 때의 결과이다. 세 모형 모두 비슷하게 예측되었음을 확인할 수 있다.

ID	GLM	GAM	CoxPHM
3169	0	0	0
3170	1	1	1
3171	0	0	0
3172	0	0	0
3173	0	0	0
3174	0	0	0
3175	0	0	0

Part 4 : Data Mining기법을 이용한 분석

위에서 사용한 방법 외에 아래의 다양한 Data Mining 기법을 이용한 분석을 추가하여 Lift Chart를 서로 겹쳐서 그려보고 각 방법들의 장단점을 서로 비교 검토해 보시오.

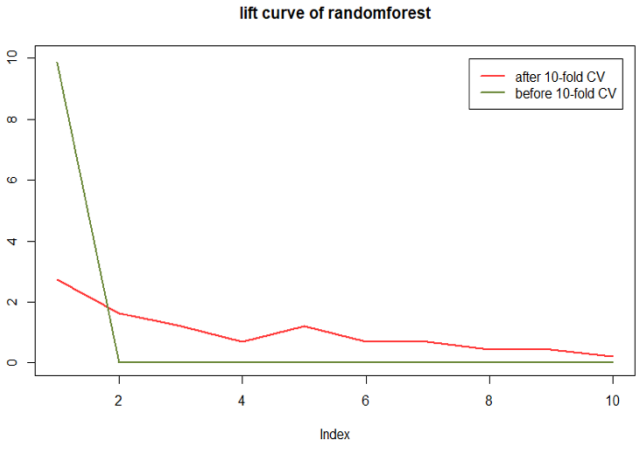
위에서 사용한 모형 외에 LDA, KNN, SVM, Random Forest, Neural Network 분석을 추가로 시행하여 기업의 부도 확률을 예측해보고자 한다.



분석을 시행한 후 모델 별 lift chart를 겹쳐 그려본 결과는 다음과 같다. 앞서 확인했던 그래프와 마찬가지로 모든 모델이 비교적 좋은 성능을 내고 있음을 확인할 수 있다.

특히, 예측력은 KNN과 SVM이 좋은 편이며 Random Forest와 Neural Network는 다소 낮을 수 있음을 추론해볼 수 있다.

머신러닝 분석 시 과적합을 방지하기 위해 Cross Validation을 시행한다. 본 과제에서도 이러한 이유로 모든 머신러닝 기법에서 Cross Validation을 진행하였고 그 유무에 따른 결과 차이를 비교해보고자 하였다. 그 중 Random Forest에서 10fold CV 전후의 결과를 나타낸 결과는 우측 그래프와 같다. CV 전에는 상위 10% 구간에 모든 부도 확률이 높은 기업들이 몰려있어 과적합이 이루어졌음을 확인할 수 있다. 반면, 10fold CV를 진행하고 나니 해당 문제가 해결되었다.



Part 5 : R- Shiny

기업의 재무정보를 입력하면 해당 기업이 1년 이내에 부도날 확률을 각 모형별로 계산해주는 어플리케이션이다.

(https://2hyeon.shinyapps.io/Default_Forecasting/)