비재무 데이터를 활용한
중소기업 휴·폐업 예측 챌린지

팀장 - newdoin

팀원 - ollon

팀원 - 김바바나

목 차

요 약

재무데이터와 비재무데이터를 활용하여 중소기업의 휴폐업 여부를 예측하고자 한다. 제공받은 재무/비재무데이터에 더해 재무비율, 파생재무비율, 지방지표, 산업특성, 재무등급, 신용등급으로 구분되는 변수를 새롭게 생성했다. 그 후 통계검정을 통해 휴폐업과 관련성을 보이는 변수 52개를 선정해 예측모델을 구축하는데 사용했다.

다양한 모델에 수많은 변수조합을 사용하여 성능을 비교한 결과 XGBoost가 가장 좋은 성능을 보였다. 이후 하이퍼파라미터를 최적화하면서 일반화가 되는 변수 조합을 분석했다. 결과적으로 기존의 재무데이터와 비재무데이터를 사용했을 때보다 인구특성, 지역특성, 산업특성이 결합된 데이터에서 더 나은 성능을 보였다.

I. 서론

기존에 기업의 건실성을 파악할 때는 재무데이터를 많이 활용했다. 재무데이터는 기업의 자금흐름이나 영업활동의 결과를 명확하게 보여줄 수 있기 때문이다.

하지만 모든 것이 데이터로 만들어지고, 활용할 수 있는 데이터가 많아지는 현시점에서 재무데이터에 더해 휴폐업과 관련 있는 비재무데이터를 찾아 함께 분석하는 것이 더 현명할 수 있다.

따라서 인구특성, 지역특성, 산업특성과 관련된 비재무데이터를 결합하여 중소기업의 휴폐업을 예측하는 모델을 구축하고자 한다.

II. 분석방법

2.1. 제공받은 데이터

(1) 재무데이터

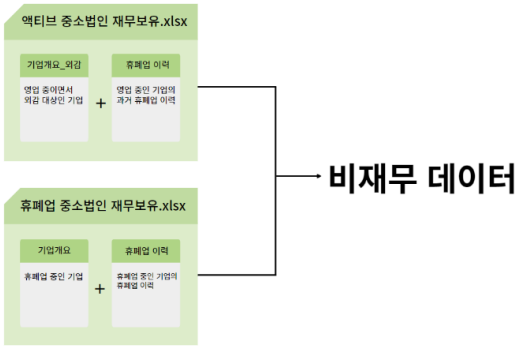
재무데이터는 기업들의 재무결산내역을 담고 있다. 결산시점의 기업재무상태를 나타내는 재무상태표 정보와 영업기간의 활동을 나타내는 포괄손익계산서 정보, 이를 이용하여 계산한 재무비율 정보가 함께 주어졌다.

재무데이터를 살펴본 결과, 'NaN'과 같은 결측치와 '8.888890e+11, 1.000000e+12' 같은 이상치가 상당수 존재했고 자본금이 음수(-)가 되는 경우나 자산총계가 0인 기업도 존재한다는 점을 발견했다.

(2) 비재무데이터

비재무데이터는 '액티브 중소법인'과 '휴폐업 중소법인'데이터가 주어졌다. 각각의 데이터는 현재 영업중인 기업과 휴폐업 중인 기업으로 구분된다. 이 데이터는 기업명, 산업코드 등과 같은 기업개요와 휴폐업 이력에 관한 정보를 담고 있다.

본 챌린지에서는 외감법인의 데이터를 평가대상으로 삼고, 나머지 데이터는 참고사항으로 간주하고 있기 때문에 비재무데이터의 기업개요 데이터 중 '외감기업'을 대상으로 한 데이터만 사용하였다.



<그림 1> 비재무데이터 구조

2.2. 파생변수

모델의 예측력을 높이고 본 챌린지의 목적을 달성하기 위해 다양한 파생변수를 생성하는 것이 매우 중요하다. 따라서, 제공받은 데이터를 가공하거나 새로운 데이터를 수집하여 파생변수를 생성했다. 이는 <표 1>과 같이 '재무비율', '파생재무비율', '지방지표', '산업특성', '재무등급', '신용점수'과 같이 크게 6가지로 정리했다.

<표 1> 파생변수 목록

구분1	구분2	변수	개수
재무비율	성장성	총자본증가율, 영업이익증가율, 당기순이익증가율, 자기자본증가율, 매출액증가율	5
	수익성	매출총이익률, 매출액영업이익률, 매출액경상이익률, 매출액순이익률, 총자산영업이익률, 자기자본영업이익률, 자기자본순이익률, 금융비용부담률, 수지비율, 사내유보 대 자기자본비율, 총자산순이익률	11
	활동성	총자본회전율, 자기자본회전율, 타인자본회전율, 유동자산회전율, 재고자산회전율, 당좌자산회전율, 순운전자본회전율, 운전자본회전율	8
	생산성	유보율	1
	안정성	자기자본비율, 유동비율, 당좌비율, 재고자산 대 순운전자본비율, 매출채권 대 매입채무비율	5
파생재무비율	재무상태변동성	재무상태변동성(차이), 재무상태변동성(비율)	60
	산업수준상태변동성 (상위10개 기업)	산업수준상태변동성(top10차이), 산업수준상태변동성(top10비율)	60
	산업수준상태변동성 (전체 기업)	산업수준상태변동성(all차이), 산업수준상태변동성(all비율)	60
지방지표	인구	인구수, 1인가구비율, 노인비율, 외국인비율,	4
	소득과 소비	가구별소득	1
	고용과 노동	고용률, 실업률, 취업자증감, 경제활동참가율	4
	주거와 교통	도시면적	1
	성장과 안정	재정자립도, 특허, 광공업생산지수	3
	환경	전기사용량	1
산업특성	산업	업종 중분류, 업종 대분류	2
	지역 특화산업	지역 특화산업 해당여부	1
	산업재해	산업재해 해당여부	1
	에너지	전기요금	1
재무등급		재무비율 상위 40% 이상인 컬럼 개수에 따른 SSS ~ F 등급	1
신용점수		재무점수, 비재무점수	2

(1) 재무비율

재무비율은 재무정보를 이용해서 상대적인 비율을 구하는 것으로 기업의 현재와 과거상태를 분석하여 재무적인 건실성을 파악하는데 사용된다. 기본재무데이터와 더불어 기업의 규모 차이를 고려한 재무비율을 분석에 사용하고자 새로운 재무비율을 생성했다.

따라서, 기업부실화 예측 관련 논문[1]을 참고하여 성장성, 수익성, 활동성, 생산성, 안정성이라는 5가지 측면에서 총 50개의 재무비율 항목을 선정했고 그 중 30개의 재무비율을 생성했다.

(2) 파생재무비율[1]

파생재무비율은 시간의 흐름에 따른 변화를 고려한 '재무상태변동성'과 산업별 경쟁정도를 고려한 '산업수준상태변동성'으로 구분된다.

'재무상태변동성'의 경우, 경영성과나 재무상태의 변화를 반영하기 위해 생성했다. 앞서 산출한 재무비율에서 당기와 전기의 차(-) 또는 비(/)로 산출했다.

'산업수준상태변동성'의 경우, 휴폐업 위기를 겪는 기업은 산업내 우량기업 또는 산업평균의 재무상태보다 불량할 것이라고 판단하여 생성했다. 상위 10개 기업의 총자산수익률 평균과 산업별 총자산수익률 평균을 구하여 차(-) 또는 비(/)로 산출했다.

(3) 지방지표

기업이 위치하고 있는 지역의 특성을 반영하여 휴폐업을 분석하기 위해 기업 소재지를 크롤링했다. 사업자등록번호와 기업명을 통해 나이스비즈인포[2]에서 수집했다. 기본 크롤링은 7일이 소요될 것으로 예상되었지만 병렬크롤러를 제작하여 소요시간을 1/15로 줄였다. 또한 검색결과가 나오지 않는 12개의 기업은 결측치로 간주하여 제거했다.

지방지표와 관련 데이터는 통계청의 e-지방지표[3]에서 14가지 항목을 수집했다. 각 지표를 수집할 때 세웠던 가설은 <표 2>와 같다. e-지방지표 14가지 항목을 각각의 비재무변수로 설정하여 분석에 사용했다.

<표 2> 지방지표

지표	가설
취업자증감 고용률 경제활동참가율 재정자립도 가구별소득 실업률	해당 지표와 지역경제의 관련성으로 휴폐업에 영향을 미칠 것이다.
인구수 1인가구비율 노인비율 외국인비율 도시면적	해당 지표에 따라 지역의 발달정도가 달라 휴폐업에 영향을 미칠 것이다.
특허 광공업생산지수 전기사용량	해당 지표가 높을수록 지역산업의 생산성이 높고 휴폐업은 줄어들 것이다.

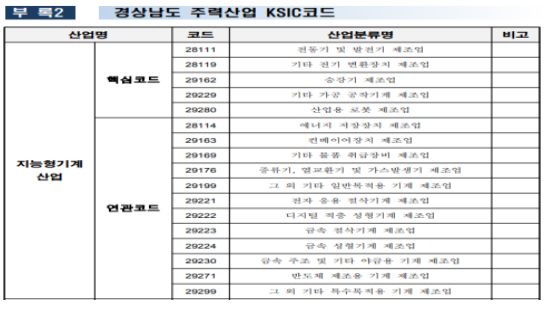
(4) 산업특성

산업분류의 경우, 비재무데이터 중 '업종'이라는 변수가 있었지만 기업의 상세 업종을 잘 나타내지 못하고 있었다. 따라서 산업코드1을 기준으로 산업코드 10차수[4]로 업종중분류, 업종대분류 정보를 생성했다.

기업별 소재지와 산업정보를 이용해 '지역특화산업 해당여부, '산업재해 해당여부', '전기요금'이라는 파생변수를 만들었다.

먼저 '지역특화산업 해당여부'의 경우, 어떤 기업의 산업이 특화산업에 해당한다면 지자체로부터 혜택을 받을 수 있기 때문에 휴폐업확률이 낮아진다고 가정했다. 따라서 한국산업기술진흥원(KIAT)에서 연도별 지역특화산업 추진 보고서[5]를 통해 핵심산업을 찾았다.

다만, 수도권의 경우 특화산업이 없기 때문에 결측치로 간주하고 보간했으며, 수도권에 가장 많이 분포된 업종 n개를 특화산업으로 지정했다. 업종의 개수는 특화산업의 개수가 가장 적은 지역을 기준으로 설정했다.



부 록2 경상남도 주력산업 KSIC코드

산업명		코드	산업분류명	비고
지능형기계산업	핵심코드	28111	전동기 및 발전기 제조업	
		28119	기타 전기 변환장치 제조업	
		29162	승강기 제조업	
		29229	기타 가공 공작기계 제조업	
		29280	산업용 로봇 제조업	
	업관코드	28114	에너지 저장장치 제조업	
		29163	컨베이어장치 제조업	
		29169	기타 물품 취급장비 제조업	
		29176	공유기, 열교환기 및 가스발생기 제조업	
		29199	그 외 기타 일반목적용 기계 제조업	
		29221	전자 응용 절삭기계 제조업	
		29222	디지털 제어 성형기계 제조업	
		29223	금속 절삭기계 제조업	
		29224	금속 성형기계 제조업	
		29230	금속 주조 및 기타 야금용 기계 제조업	
		29271	반도체 제조용 기계 제조업	
		29299	그 외 기타 특수목적용 기계 제조업	

<그림 2> 지역별 특화산업

다음으로 '산업재해 해당여부'의 경우, 산업재해가 발생한 사업장에는 벌금 또는 징역과 같은 처벌로 인해 경영상 어려움을 겪을 것이라고 판단했다. 따라서 고용노동부에서 공표한 연도별 산업재해 발생 현황자료[6]를 이용하여 우리가 가진 기업명단과 비교했다.

우선, PDF파일로 만들어진 산업재해 발생현황자료를 데이터프레임으로 불러올 수 있도록 XLSX로 변환하고 기업명을 정제했다. 사업자등록번호로 비교할 수 없는 상황이었기 때문에 기업명이 일치하는지 여부로 산업재해 해당여부를 판단했다.

1. 중대재해 발생 사업장으로 산업재해율이 규모별 같은 업종의 평균 재해율 이상인 사업장

○ '18년 중대재해 발생 사업장(503개소)

지역	업종명	규모	사업장명 (현장명)	소재지	중대재해자수 (명)	근로자수 (명)	재해자수 (명)	재해율 (%)	규모별 동종업종평균 재해율 (%)
서울	건설업	100~299인	㈜진성씨티건설(원청) ㈜진성산업건설(하청) (마곡지구 업무용지 C11-2블럭 복합시설 신축공사)	서울 강서구 내발산동 200-3번지 마곡지구 상업용지 C11-2블럭	1	107	2	1.87	0.47
서울	건설업	50인~99인	(주)아주디자인그룹 (청일학원 증축 및 대수선공사)	서울 용산구 용산동2가 1-206	1	69	1	1.45	0.72
서울	건설업	50인미만	㈜디케이건설(원청) 주식회사대원이엔씨(하청) (방이동50-10 복합시설 신축공사)	서울 송파구 오금로11길 63-13 (방이동)	2	14	2	14.29	1.8
서울	건설업	50인미만	㈜로이넥스건설(원청) 동원산전주식회사(하청) (길동 413-37외2필지 오피스텔 공사)	서울 강동구 진황도로43길 13-18 (길동)	1	22	2	9.09	1.8
서울	건설업	50인미만	(주)신성글로벌네트워크(원청) 일성네트웍(하청) (수도권_관악2,금천구,광명시(선로) '18년 NW유선공사 계약 (강	서울 금천구 가마산로 70 (가산동) 일대	1	3	2	66.67	1.8
서울	건설업	50인미만	(주)지투원(원청) 주식회사원중공업(하청) (길동 413-45번지 판매시설 증, 개축공사)	서울 강동구 양재대로 1449 (길동413-45)	1	33	2	6.06	1.8

<그림 3> 산업재해 발생현황

'전기요금'의 경우, 산업별로 전기요금이 다르며 전기요금이 비싼 산업일수록 비용부담이 될 것이라는 가정하에 선정했다. 한국전력공사의 산업분류별 전력사용량 데이터[7]로 연도, 지역, 산업별 평균 전기요금을 구했다.

전기요금과 휴폐업간의 관계를 더 정확하게 살펴보기 위해 정규화(min-max scaler)를 했다. 연도와 지역을 그룹으로 묶어 산업별로 정규화한 뒤, 0.1, 0.2, 0.3, 0.4 범위로 나누어 휴폐업비율을 계산한 결과는 <표 3>과 같다.

<표 3> 정규화 범위

스케일 범위	정규화 범위	휴폐업 비율
0.1	0.0 ~ 0.1	0%
	0.1 ~ 0.2	0%
	0.2 ~ 0.3	0%
	0.3 ~ 0.4	4.6%
	0.4 ~ 0.5	5.0%
	0.5 ~ 0.6	4.5%
	0.6 ~ 0.7	4.2%
	0.7 ~ 0.8	6.2%
	0.8 ~ 0.9	4.7%
	0.9 ~ 1.0	7.9%
스케일 범위	**정규화 범위**	**휴폐업 비율**
0.2	0.0 ~ 0.2	0%
	0.2 ~ 0.4	4.5%
	0.4 ~ 0.6	4.7%
	0.6 ~ 0.8	5.2%
	0.8 ~ 1.0	5.7%
스케일 범위	**정규화 범위**	**휴폐업 비율**
0.3	0.0 ~ 0.3	0%
	0.3 ~ 0.6	4.7%
	0.6 ~ 0.9	5.0%
	0.9 ~ 1.0	7.9%
스케일 범위	**정규화 범위**	**휴폐업 비율**
0.4	0.0 ~ 0.4	4.5%
	0.4 ~ 0.8	5.0%
	0.8 ~ 1.0	5.7%

휴폐업비율의 선형성이 가장 잘 보이는 정규화 범위가 0.2 단위였기 때문에 정규화 범위를 0.2로 구분한 전기요금을 변수로 사용했다.

(5) 재무등급

파생변수 중 재무비율 데이터를 이용하여 재무등급이라는 새로운 지표를 만들었다. 재무등급 변수는 다양한 관점으로 재무데이터를 분석함과 동시에 사용하는 재무 변수를 줄이고자 만들었다.

지표에 사용된 항목은 재무비율 데이터에서 성장성, 수익성, 활동성, 생산성, 안정성 5가지 항목을 기준으로 26개를 선택했다. 각 항목마다 상위 n%에 해당하는 것에 아래와 같이 등급을 나누었다.

<표 4> 등급 기준

선정 기준	등급
상위 n% 이상인 컬럼 24개 이상	SSS
상위 n% 이상인 컬럼 22개 이상	SS
상위 n% 이상인 컬럼 20개 이상	S
상위 n% 이상인 컬럼 18개 이상	AAA
상위 n% 이상인 컬럼 16개 이상	AA
상위 n% 이상인 컬럼 14개 이상	A
상위 n% 이상인 컬럼 11개 이상	B
상위 n% 이상인 컬럼 9개 이상	C
상위 n% 이상인 컬럼 7개 이상	D
상위 n% 이상인 컬럼 5개 이상	E
상위 n% 이상인 컬럼 5개 미만	F

상위 n%의 기준을 정하기 위해 30%, 40%, 50%로 수행 후 비교하여 가장 나은 n값을 선정했다. 아래 <표 5>는 수행한 결과이다.

<표 5> 등급별 기업분포

등급	상위 30%	상위 40%	상위 50%
SSS	124	1190	5623
SS	397	2501	5843
S	1109	4147	6946
AAA	2497	5434	7204
AA	3912	6505	7115
A	5584	7250	7215
B	11340	7899	7803
C	4875	9292	4215
D	11658	10897	9147
E	14115	10961	8102
F	21739	11274	8137

수행 결과 상위 40%에서 등급별 기업 분포의 선형성이 가장 잘 나타났다. 전체 데이터 중 휴폐업률과 등급안에서의 휴폐업률은 <표 6>과 같이 등급이 높아질수록 낮아졌다.

<표 6> 등급별 휴폐업률

등급	전체 중 휴폐업(%)	등급 내 휴폐업(%)
SSS	0.06%	3.87%
SS	0.14%	4.28%
S	0.26%	4.75%
AAA	0.34%	4.86%
AA	0.36%	4.26%
A	0.40%	4.26%
B	0.48%	4.74%
C	0.67%	5.58%
D	0.86%	6.09%
E	0.77%	5.47%
F	0.80%	5.52%

(6) 재무점수 및 비재무점수[8)]

실제 신용평가모델에 사용되는 기법으로 재무데이터와 비재무데이터를 점수화(Scoring)하기 위해서 스코어카드에 근거하여 재무점수, 비재무점수를 생성했다.

재무점수의 경우, 재무비율 30개를 변수별로 구간화한 뒤 구간별 가중치(WoE; Weight of Evidence)를 산출했다. 산출한 가중치를 로지스틱 회귀 모형에 학습시켜 산출된 회귀계수를 통해 구간별 점수를 생성했다. 이렇게 만들어진 스코어카드를 바탕으로 재무비율을 대입했을 때 나오는 총점이 재무점수이다.

<표 7> 재무 스코어카드

변수명	구간	점수
총자본증가율	(-inf, -0.14)	31.377
총자본증가율	[-0.14, 0.0)	18.461
...	...	...
매출채권대매입채무비율	[4.5, 8.5)	16.380
매출채권대매입채무비율	[8.5, inf)	23.470

비재무점수의 경우, 통계데이터 14종과 산업재해 해당여부, 전기요금 변수에 대해서 재무점수에서 산출한 방식과 동일하게 스코어카드를 만들었다. 이를 바탕으로 총점을 산출했다.

<표 8> 비재무 스코어카드

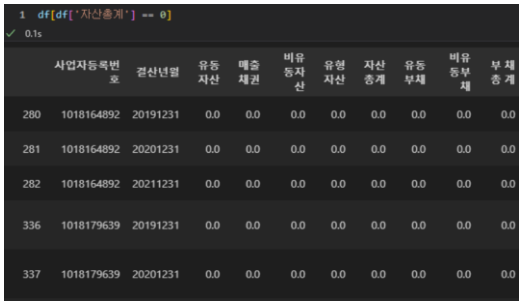
변수명	구간	점수
1인가구	(-inf, 27.10)	48.612
1인가구	[27.10, 33.75)	37.468
...	...	...
전기요금	[0.73, 0.89)	37.715
전기요금	[0.89, inf)	40.875

2.3. 데이터 전처리

(1) 재무데이터

재무데이터에 존재하는 결측치와 이상치는 0으로 변경하여 분석에 사용했다. 결측치는 재무 기록시 0의 값을 가지는 데이터는 표기하지 않는 점을 고려했고 이상치는 비율계산과정에서 생긴 예외처리에 따른 것이라는 점을 고려했다.

다음으로 자본금이 음수(-)가 되는 경우는 조세특례제한법 제120조 5항[9)]을 참고하여 0원으로 보는 것이 합리적이라고 판단했다. 자산총계가 0인 행의 경우, 모든 재무데이터가 0이었기 때문에 분석에 불필요하다고 판단하여 제거했다.



```
df[df['자산총계'] == 0]
```

0.1s

	사업자등록번호	결산년월	유동자산	매출채권	비유동자산	유형자산	자산총계	유동부채	비유동부채	부채총계
280	1018164892	20191231	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
281	1018164892	20201231	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
282	1018164892	20211231	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
336	1018179639	20191231	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
337	1018179639	20201231	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

<그림4> 자산총계가 0인 데이터

마지막으로, 재무데이터 특성상 너무 높거나 낮은 극단치가 존재할 수 있으며 이는 분석에 악영향을 미칠 수 있다고 판단했다. 따라서, 상위 1%이상, 하위 99% 이하 값은 각각 1%, 99% 경계값으로 변환하는 윈저라이징(winsorizing)을 진행했다.

(2) 비재무데이터

휴폐업 이력은 날짜가 일치하지 않는 등 데이터의 정합성이 맞지 않는 경우가 있었다. 모델링에 악영향을 미칠 수 있기 때문에 과거데이터를 삭제하고 신뢰성 있는 최신데이터만 사용하기 위해서 종료일자가 '99991231'인 데이터만 사용했다.

다음으로 비재무데이터의 기업개요와 휴폐업 이력을 병합할 때, 기업개요에 존재하지 않는 사업자등록번호는 신뢰할 수 없는 데이터로 간주하여 사용하지 않았다. 사업자등록번호가 여러 개인 기업은 가장 결산을 많이 한 사업자등록번호 하나만 사용했다.[10)]

비재무데이터의 변수는 총 31개로 이루어져 있다. 이 중 고유 값이 1개인 변수는 분석과 예측력 향상에 도움이 되지 않으므로 제거했다. 또한, 고유 값이 많은 변수들 중 홈페이지URL, 주요사업내용, 대표자명과 같은 변수도 예측에 도움이 되지 않는다고 판단하여 제거했다. 따라서, 분석에 1차적으로 사용한 비재무데이터 변수는 아래표의 마크가 되어 있는 9개이다.

<표 9> 비재무데이터 목록

피처명	자료형	결측값	고윳값
기업접두명	object	38302	4
기업명	object	0	35529
기업접미명	object	73655	4
기업영문명	object	93	34578
업종	object	0	8
기업규모	object	0	1
공기업구분	object	0	2
개인법인구분	object	0	1
본점지점구분	object	0	2
국외투자법인여부	object	0	2
벤처기업여부	object	0	2
상장코드	object	108713	134
산업코드차수	float64	19	1
산업코드1	float64	184	1411
산업코드2	float64	104154	432
산업코드3	float64	108067	138
공공기관유형	object	0	3
중견기업보호여부	object	0	2
본점기업코드	float64	103539	2468
설립일자	float64	1975	9899
설립구분	object	0	8
상장일자	float64	108425	190
주요사업내용	object	403	24118
국가명	object	102570	64
홈페이지URL	object	40109	17619
대표자명	object	2	22987
직원수	float64	1649	759
종료일자	float64	97718	1
시작일자	float64	97718	710
휴폐업구분	object	97718	4
상태발생일자	float64	98892	932
피처 총 개수	31개		

2.4. 베이스라인 모델

베이스 라인 모델은 LightGBM으로 선정했다. 초기 LogisticRegression으로 학습 및 예측을 시도했으나 성능이 좋지 않아 베이스라인으로 부적합하다고 판단했다. 따라서 성능이 보장된 GBM 모델 중 가볍고 과적합을 방지할 수 있는 LightGBM으로 선정했다.

학습에 사용한 데이터는 제공받은 재무데이터, 비재무데이터가 합쳐진 (101962, 56) 크기의 데이터셋이다. 기본적인 성능을 알아보기 위한 작업이었으므로 추가 생성한 파생변수는 사용하지 않았다.

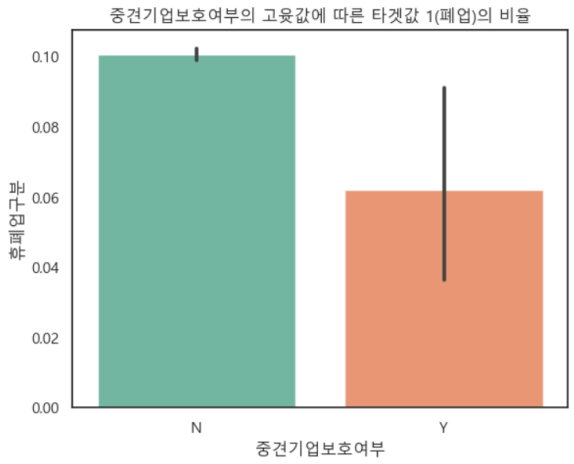
사용	재무 데이터 비재무 데이터 (공기업구분, 본지점구분, 국외투자법인여부, 벤처기업여부, 공공기관유형, 업종)
제거	기업접두명, 기업접미명, 기업영문명, 산업코드1, 산업코드2, 산업코드3, 본점기업코드, 설립일자, 설립구분, 상장일자, 주요사업내용, 국가명, 홈페이지URL, 대표자명, 종료일자, 시작일자, 상태발생일자, 중소기업보호여부, 직원수

(1) 타겟값 정의 및 EDA

모델을 학습하기 전 타겟값을 정의하고 탐색적 데이터 분석(EDA)을 실시했다. 본 챌린지는 휴업 및 폐업을 예측하는 것으로 휴업의 경우 장래 영업활동 재개 의사를 가지고 있지만, 신용평가기관입장에서는 주의가 필요한 상태라고 판단했다. 따라서 휴업과 폐업을 하나로 묶어 타겟값으로 정의했다.

EDA시 고려한 사항은 신뢰구간이 너무 넓어 통계적 유효성이 떨어지거나 변수의 고유값 간 휴폐업비율이 차이가 나지 않는 두 가지 경우이다. 이는 예측력 향상에 도움이 되지 않으므로 변수를 제거했다.

아래 그림인 중견기업보호여부는 신뢰구간이 매우 넓어 통계적 유효성이 떨어지기 때문에 제거했다.



<그림 5> 중견기업 보호여부

(2) 학습결과

모델의 평가지표는 F1-Score이다. 이는 재현율을 중시하되, 정밀도도 고려하는 평가지표로써 모델의 성능을 가장 잘 나타낼 수 있다고 판단했다. 학습결과, F1-Score는 0.7058의 성능을 기록했다.

```
오차 행렬
[[19700   249]
 [  878  1352]]
정확도: 0.9492, 정밀도: 0.8445, 재현율: 0.6063,    F1: 0.7058, AUC:0.9365
```

<그림 6> 학습결과

III. 모델링 결과

3.1. 변수선정

베이스라인 모델에 사용한 재무/비재무 데이터 변수 56개와 모델 성능을 향상시키기 위해 생성한 파생변수 232개로 총 288개의 변수가 있다. 이 변수들 중에는 같은 의미를 담고 있는 변수가 존재하여 다중공선성 문제를 일으킬 수 있다.

따라서 휴폐업과 유의미한 관계를 가지고 있는 변수를 선정하여 분석에 사용하기 위해 재무/비재무데이터, 재무비율/파생재무비율, 재무등급, 지방지표에 통계검정을 실시했다. 연속형 데이터의 경우 Two Sample T-Test, 범주형 데이터의 경우 Two Sample Chi-Squared-Test를 실시하여 P-value가 0.05보다 작은 변수를 선정했다.

파생변수 중 '지역특화 산업여부'와 '산업재해 해당여부'는 P-value가 0.05보다 컸기 때문에 통계적으로 유의하지 않다고 판단하여 제거했다. 선정된 변수는 <표 10>과 같다.

<표 10> 선정된 변수

재무 데이터	
유동자산, 매출채권, 비유동자산, 유형자산, 자산총계, 유동부채, 비유동부채, 부채총계, 이익잉여금, 자본총계, 매출액, 판매비와관리비, 영업이익, 법인세비용차감전순손익, 법인세비용, 당기순이익, 기업순이익률, 유보액대총자산비율, 금융비용대총비용비율, 차입금의존도, 자기자본비율, 총자본회전율, 매출원가, 재고자산	24 개
비재무 데이터	
본점지점구분, 국외투자법인여부, 업종중분류, 업종대분류	4 개
재무비율 / 파생재무비율	
총자산영업이익률, 총자산순이익률, 총자본회전률, 자기자본비율, 유동비율, 당좌비율	6 개
총자산영업이익률(top10_차이), 유동비율(top10_차이), 총자산순이익률(top10_차이), 총자본회전률(top10_차이), 자기자본비율(top10_차이), 당좌비율(top10_차이)	6 개
지방지표	
1인 가구, 고용률, 경제활동참가율, 광공업생산지수, 재정자립도, 가구별소득, 노인비율, 외국인비율, 도시면적	9 개
기 타	
재무등급, 재무점수, 비재무점수	3 개

3.2. 최종모델 선정과정

휴폐업 예측에 사용된 머신러닝 모델은 앙상블모델인 LightGBM, XGBoost, RandomForest, 회귀모델인 LogisticRegression, 통계모델인 NaiveBayes로 5개이다. 이 모델들의 기본 파라미터를 사용하여 학습 및 예측을 시도했다.

위 **3.1.변수선정** 단계에서 선정한 변수들의 모든 경우를 조합하여 학습데이터로 사용했다. LogisticRegression과 NaiveBayes의 경우 베이스라인 데이터부터 성능이 좋지 않았기 때문에 추가분석은 시도하지 않았다. 각 알고리즘의 성능이 높은 상위 2개와 최하위 결과를 요약하여 <표 11>에 나타냈다.

분석 결과, 모든 알고리즘에서 파생변수가 적게 포함될수록 높은 성능을 내는 것을 알 수 있었다. 학습할 데이터가 점점 증가할 때 파라미터값은 변하지 않아 이런 현상이 발생했다고 판단했다.

또한 알고리즘이 훈련데이터셋과 테스트데이터셋에서만 높은 성능을 발휘하지 않고 새로운 데이터에도 동일한 성능을 발휘할 수 있는 파라미터를 찾고자 했다. 따라서, 가장 준수한 성능을 보인 XGBoost의 하이퍼파라미터를 최적화시키면서 일반화가 되는 변수 조합을 분석했다. 분석 결과는 <표 12>에 나타냈다.

Random Search, Bayesian Search를 통해 하이퍼파라미터를 최적화한 결과, 많은 정보가 포함된 변수 조합일수록 준수한 성능을 보였다. 가장 좋은 성능을 보인 변수조합은 통계검정으로 선별한 재무데이터/비재무데이터/재무비율과 재무등급, 지방지표, 재무점수, 비재무점수 조합이었다.

<표 11> 알고리즘별 성능

α : T-Test 재무데이터 + Chi-Squared-Test 비재무데이터
β : T-Test 재무비율, γ : T-Test 파생재무비율
δ : 재무등급, ε : 지방지표, ζ : 재무점수, η : 비재무점수

머신러닝 알고리즘	구분	사용 변수 조합	예측 성능	
			Recall	F1 Score
LightGBM	베이스라인	기본 재무/비재무데이터	0.6331	0.7264
	1st	$\alpha+\beta+\delta$	0.6344	0.7280
	2nd	$\alpha+\beta+\gamma$	0.6291	0.7263
	Worst	$\alpha+\beta+\gamma+\zeta$	0.6088	0.6960
XGBoost	베이스라인	기본 재무/비재무데이터	0.6411	0.7322
	1st	$\alpha+\beta+\gamma$	0.6464	0.7394
	2nd	$\alpha+\beta$	0.6397	0.7352
	Worst	$\alpha+\beta+\varepsilon+\eta$	0.6142	0.7042
Ramdom Forest	베이스라인	기본 재무/비재무데이터	0.5987	0.7152
	1st	$\alpha+\beta$	0.6000	0.7156
	2nd	α	0.5947	0.7127
	Worst	$\alpha+\beta+\gamma+\zeta+\eta$	0.5676	0.6864
Logistic Regression	베이스라인	기본 재무/비재무데이터	0.0040	0.0078
Naive Bayes	베이스라인	기본 재무/비재무데이터	0.9642	0.1003

<표 12> 하이퍼파라미터 최적화 및 사용 변수

머신러닝 알고리즘	구분	사용 변수 조합	예측 성능	
			Recall	F1 Score
XGBoost	베이스라인	기본 재무/비재무데이터	0.6456	0.7269
	1st	$\alpha+\beta+\gamma+\delta+\epsilon+\zeta$	0.6777	0.7597
	2nd	$\alpha+\beta+\delta+\epsilon+\zeta+\eta$	0.6744	0.7589
	3rd	$\alpha+\beta+\varepsilon$	0.6423	0.7335

IV. 결론

이번 챌린지는 재무데이터와 비재무데이터를 결합하여 중소기업의 휴폐업을 예측하는 모델 구축하는 것이 목표였다. 인구특성과 지역특성을 나타내는 지방지표 데이터와 지역 특화산업 및 산업재해 해당여부, 산업별 전기요금을 나타내는 산업특성과 같은 비재무데이터를 찾아 분석에 사용했다.

학습을 위해 수집한 데이터를 선별하기 위해 EDA와 통계 검정을 실시한 결과, 288개의 중 52개의 변수를 선정했다. 5가지 알고리즘에 학습 및 예측을 진행했고 F1-Score 성능이 가장 나은 모델의 하이퍼파라미터 값을 조정하며 일반화했다.

그 결과, 제공받은 재무데이터와 비재무데이터만 사용했을 때 0.7269의 성능을 보였지만 인구특성, 지역특성, 산업특성이 결합된 데이터에서 0.7597의 성능을 보이며 5% 가량 성능을 향상 시킬 수 있었다. 또한 특정 비재무데이터가 성능에 큰 비중을 차지하지 않았고 여러 특성들이 적절히 조화를 이룰 때 휴폐업 예측에 도움이 되는 것을 확인했다.

특히 통계청, 공공기관, 공공데이터포털 등에서 쉽게 수집할 수 있는 비재무데이터만으로도 예측모델의 성능을 끌어올렸다는 점에서 비재무데이터를 결합한 모델구축의 가능성을 볼 수 있었다.

한계점으로는 시/군/구별 지방지표가 2021년부터 제공되었기 때문에 시/도별 지방지표만 사용했다는 점이다. 이는 사업체가 속한 지역의 대분류이기 때문에 지역특성을 정확하게 대변하지 못했다. 따라서 이후 데이터에 대하여 예측 모델을 구축 시, 더 세분화된 분류인 시/군/구별 지방지표를 사용하여 분석할 필요가 있다.

Reference

[1] 김량형, 유동희 and 김건우. (2016). 데이터마이닝 기법을 이용한 기업부실화 예측 모델 개발과 예측 성능 향상에 관한 연구. Information Systems Review, 18(2), 173-198.

[2] "나이스비즈인포", 나이스평가정보(주), URL:https://url.kr/4k1sr6

[3] "KOSIS e-지방지표", 국가통계포털, URL:https://url.kr/9oy83t

[4] "산업코드 10 차수", 국가통계포털, URL:https://url.kr/96ygxe

[5] "지역별 지역특화산업", 한국산업기술진흥원, URL:https://url.kr/acnp56

[6] "고용노동부 사전정보 공표목록", 고용노동부, URL:https://url.kr/813hgt

[7] "한국전력공사 산업분류별 전력사용량", 공공데이터포털, URL:https://url.kr/1za2yi

[8] "Credit Scoring with Machine Learning", Medium, URL:https://url.kr/4l1z8g

[9] "개인기업에서 법인으로 전환하는 사업장의 순자산가액이 음수(-)인 경우", 세무법인 넥스트, URL:https://url.kr/cgqkso

[10] "무료 세무상담 Q&A", 찾아줘세무사, URL:https://url.kr/wuj4td

[11] “기업신용평가체계”, 신용보증기금 URL: https://url.kr/x5kvqo

[12] "신용평가방법론", 나이스신용평가, URL:https://url.kr/3obqgr

[13] 장제훈. (2021). 머신러닝 기법을 활용한 자영업자 폐업 예측 모형 연구: 서울시 25 개 자치구를 중심으로. 인문사회 21, 12(1), 1081-1096.

[14] 오희장. (2005). 도산예측에서 신용등급정보의 유용성. 경제연구, 23(2), 173-208.

[15] 이금숙 and 박소현. (2019). 업종별 창업 및 폐업의 지리적 특성 분석. 한국경제지리학회지, 22(2), 178-195.

[16] 방준아, 손광민, 이소정, 이현근 and 조수빈. (2018). 서울 치킨집 폐업 예측 모형 개발 연구. 한국빅데이터학회 학회지, 3(2), 35-49.