NUMBLE X NICEDNB

최종보고서

이용우 김예지 박지영

목차

01. EDA 및 전처리

[제공된 데이터셋 확인]

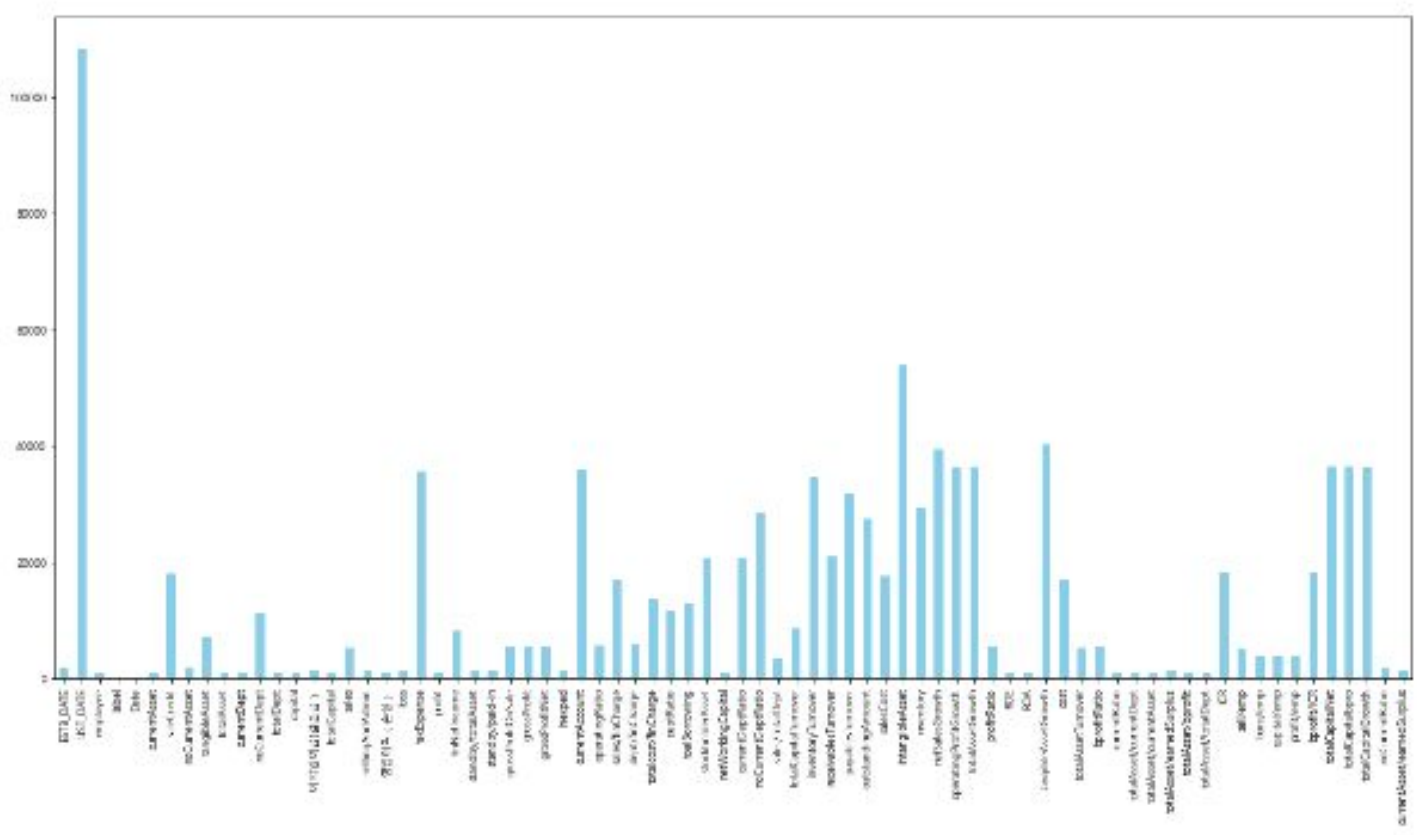
액티브 중소법인 재무보유: (28,982개 데이터, 28개 변수)

휴폐업 중소법인 재무보유: (6,739개 데이터, 28개 변수)

재무데이터: (109,142개 데이터, 76개 변수)

01. EDA 및 전처리

[NULL값 확인]



대부분 재무데이터에서 null값이 존재했으며, 가장 많은 결측치를 가진 변수는 약 11,000개를 가진 '상장일자' 였음.

01. EDA 및 전처리

[결측치 제거]

오류값(#VALUE, #N/A, #DIV/0!) 등 결측치 제거

[이상치 Cleansing Rule]

변수를 7가지 구간으로 Binning하여 이상치 처리

01. EDA 및 전처리

[파생변수 생성(23개)]

- 설립일자와 상장일자는 업력으로 변경(월 단위)
- 논문 참고하여 추가 생성한 변수

수익성	ROE(%)	**안정성**	이자비용	**성장성**	매출액증가율(%)
	ROA(%)		유동비율(%)		영업이익증가율(%)
	영업이익률(%)		유동부채대총자산(%)		총자산증가율(%)
	이익잉여금대총자산(%)		유동자산대총자산(%)		유형자산증가율(%)
	이자보상비율(%)		비유동비율(%)		총자본증가율(%)
	총자본순이익률(%)		이익잉여금대유동자산비율(%)		
	총자본영업이익률(%)		총부채대총자산비율(%)		
	영업이익대총자산비율(%)		영업이익대비이자보상비율(%)		
	당기순이익률(%)	**활동성**	총자산회전율(회)		

01. EDA 및 전처리

[데이터 병합]

- 데이터 라벨링(휴폐업: 0, 액티브: 1)
- 사업자번호(BIZ_NO)를 Key 값으로 재무데이터와 휴폐업 및 액티브 기업개요 데이터 병합

=> (109,142개 데이터, 78개 변수)

01. EDA 및 전처리

[데이터 Binning]

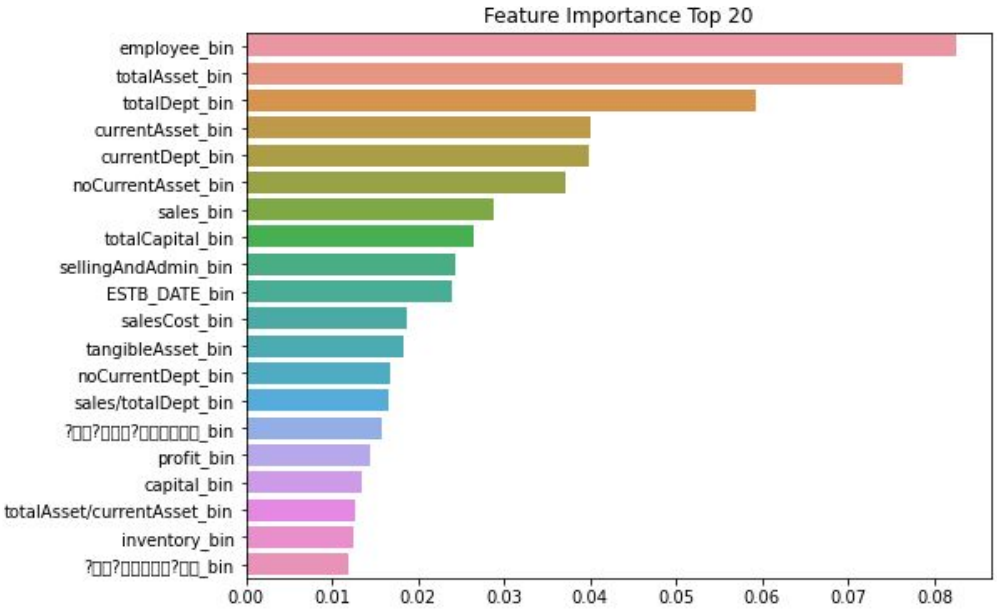
- 변수들을 7개로 구간화

	A	B	C	D	E	F	G	H	I	J	K
1	BIZ_NO	label	ESTB_DAT	LIST_DATE	employee_	currentAs	saleCredit_	noCurrent	tangibleAs	totalAsset_	currentDe
2	1268145633	0	5		5	2	2	3	4	2	2
3	1268145633	0	5		5	2	3	3	4	2	3
4	1268145633	0	5		5						
5	1268145633	0	5		5						
6	1388101749	0	7		3	2		2	2	1	5
7	1068179240	0	6		1	1	1	1	1	1	1
8	2248114507	0	6		2	1		1	1	1	1
9	1168132785	0	6		5	3	3	2	2	1	2
10	1168132785	0	6		5	2	2	2	2	1	2

02. 모델링1 : 랜덤포레스트

[변수 선택]

랜덤포레스트로 산출된 변수 중요도를 기준으로 상위 20개의 변수 선택



종업원수	자산총계	부채총계	유동부채
유동자산	비유동자산	매출액	자본총계
판매비와관리비	설립일자	매출원가	유형자산
비유동부채	부채총계대매출액(%)	영업이익(손실)	당기순이익(손실)
자본금	유동자산대총자산(%)	재고자산	이익잉여금(결손금)

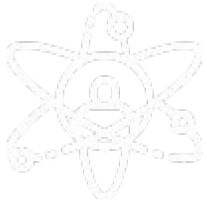
=> 총 78개의 변수 중 20개의 유의미한 변수 선정

02. 모델링1 : 랜덤포레스트

[랜덤포레스트 모델링]

```
#기본적인 randomforest모형

from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score # 정확도 함수

clf = RandomForestClassifier(n_estimators=20, max_depth=5,random_state=0)
clf.fit(train_x,train_y)

predict1 = clf.predict(test_x)
# AUC score

from sklearn.metrics import roc_auc_score

predict2 = clf.predict(test_x)
print('Random Forest AUC Score :', roc_auc_score(test_y,predict2))
```

```
Random Forest AUC Score : 0.8233455585904601
```



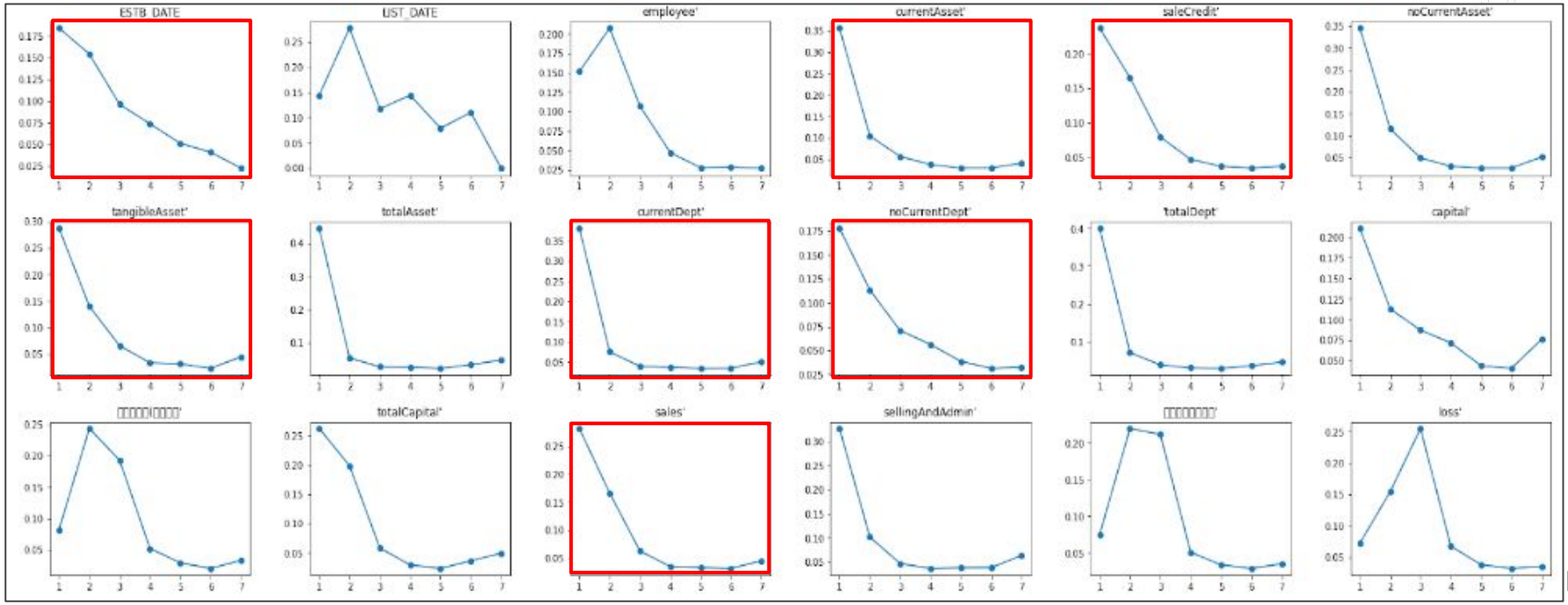
하이퍼파라미터 튜닝을 통해 얻은

가장 높은 **AUC = 0.823**

02. 모델링2 : 구간별 부도율 추세 + 단변량분석

[1차 변수 필터링]

구간별 부도율 시각화하여 그래프가 상승 또는 하강 추세일 때 모델링 변수로 선택



[18개의 Sample]

02. 모델링2 : 구간별 부도율 추세 + 단변량분석

[1차 변수 필터링]

부도율 그래프 일관성(상승 또는 하강) 기준으로 선택된 변수

설립일자	유형자산	매출액	수지비율(%)	미수금	유동자산대총자산(%)
당기순이익률(%)	자산총계	판매비와관리비	영업비율(%)	매출원가	
유동자산	유동부채	법인세비용	유동부채비율(%)	이자비용	
매출채권	비유동부채	매출액영업이익률(%)	부채총계대매출액(%)	총자산회전율(회)	
비유동자산	부채총계	매출액순이익률(%)	총자본회전율(회)	영업이익률(%)	

=> 총 78개의 변수 중 26개의 유의미한 변수 선정

02. 모델링2 : 구간별 부도율 추세 + 단변량분석

[2차 변수 필터링]

1차 필터링된 변수에 대해 단변량 분석 후 유의미한 P값을 가지는 변수

	Pr(>|t|)	
(Intercept)	2.09e-08	***
ESTB_DATE_bin	< 2e-16	***
inventory_bin	0.74139	
profitRatio_bin	0.95193	
currentAsset_bin	< 2e-16	***
saleCredit_bin	0.13323	
noCurrentAsset_bin	< 2e-16	***
tangibleAsset_bin	2.98e-05	***
totalAsset_bin	0.09052	.
currentDept_bin	0.84522	
noCurrentDept_bin	1.14e-07	***
totalDept_bin	5.85e-05	***
sales_bin	9.01e-08	***
sellingAndAdmin_bin	0.00368	**
texExpense_bin	0.00104	**
grossProfit_bin	0.77765	
grossProfitNet_bin	0.95212	
inexped_bin	0.31084	
operatingRatio_bin	0.59575	
currentDeptRatio_bin	< 2e-16	***
sales.totalDept_bin	0.03801	*
totalCapitalTurnover_bin	0.51269	
outstandingAmount_bin	0.00055	***
salesCost_bin	0.02661	*
cost_bin	2.30e-05	***
totalAssetTurnover_bin	0.02491	*
bprofitRatio_bin	0.31558	
totalAsset.currentAsset_bin	< 2e-16	***

설립일자	유동자산	비유동자산	유형자산
비유동부채	부채총계	매출액	판매비와관리비
법인세비용	유동부채비율(%)	부채총계대매출액(%)	미수금
매출원가	이자비용	총자산회전율(회)	유동자산대총자산(%)

[1차 필터링된 변수에 대해 단변량 분석 후 선택된 변수]

=> 총 26개의 변수 중 16개의 유의미한 변수 선정

02. 모델링2 : 구간별 부도율 추세 + 단변량분석

[모델링]

	Model	Accuracy	AUC
et	Extra Trees Classifier	0.9493	0.9137
rf	Random Forest Classifier	0.9488	0.9181
lightgbm	Light Gradient Boosting Machine	0.9476	0.9109

[AUC가 가장 높은 **et, rf, lightgbm** 모델을 Ensemble]

	Model	Accuracy	AUC
0	Voting Classifier	0.9507	0.9366

[최종 도출된 결과]

제공된 데이터 셋만을 사용하여 AutoML을 실행하였을 때 AUC가 **0.9366**인 결과를 도출

03. 외부 데이터

[산출 배경]

서울경제와 중앙일보의 기사 등을 통해 종업원 수가 기업 부도율 간의 상관관계가 존재한다는 Insight를 도출

수익성(영업이익, 당기순이익)과 안정성(자본금, 판매비와 관리비)의 주요지표와 종업원 수의 비율을 데이터로 활용

03. 외부 데이터

[산출한 변수]

판매비와관리비/종업원수	영업이익/종업원수	자본금/종업원수	당기순이익/종업원수

[기존 모델에 변수 추가 학습]

	Model	Accuracy	AUC
et	Extra Trees Classifier	0.9525	0.9316
rf	Random Forest Classifier	0.9512	0.9326
lightgbm	Light Gradient Boosting Machine	0.9492	0.9231

[AUC가 가장 높은 **et, rf, lightgbm** 모델을 Ensemble]

	Model	Accuracy	AUC
0	Voting Classifier	0.9531	0.9527

[최종 도출된 결과]

비재무 데이터를 결합하여 AutoML을 실행하였을 때 AUC가 **0.9366에서 0.9527로 향상**됨

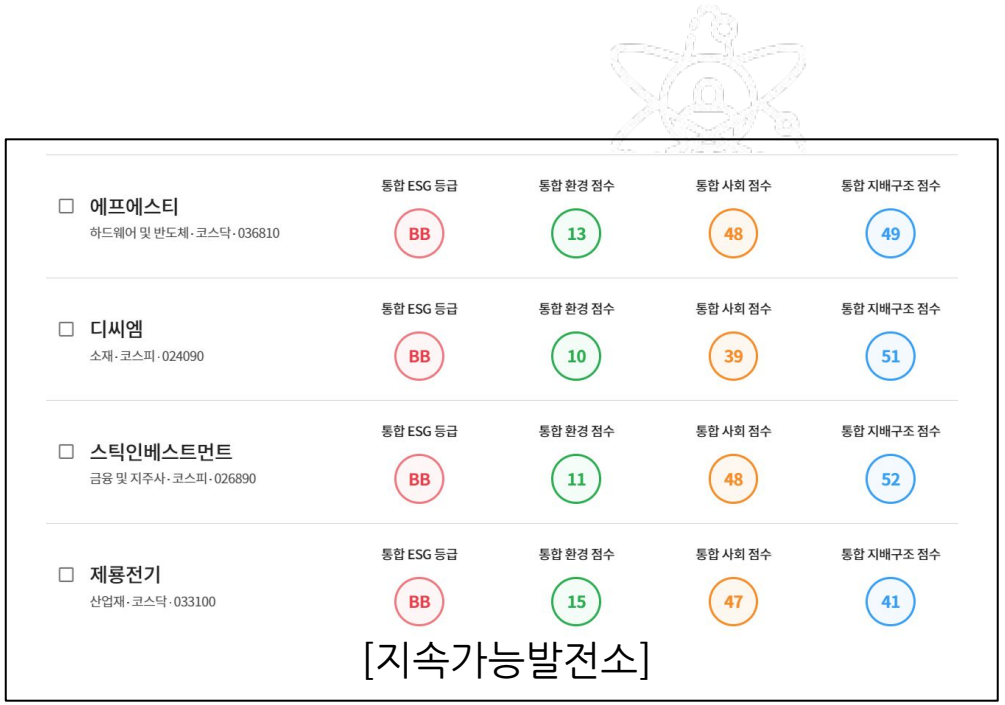
03. 외부 데이터

[추가 아이디어]

1. ESG 데이터

NO	기업명	기업코드	ESG등급	환경	사회	지배구조	평가년도
995	CJ ENM	035760	A	B	A+	A	2021
994	CJ대한통운	000120	A	A	A+	A	2021
993	CJ씨푸드	011150	B	C	B+	B+	2021
992	CJ제일제당	097950	A	A	A+	A	2021

[한국ESG기준원]



[지속가능발전소]

ESG 데이터를 크롤링하여 휴폐업 예측 모델링 변수로 선택(한국ESG기준원의 경우 기업명으로, 지속가능발전소의 경우 법인등록번호로 매핑) -> **ESG 지수가 높을수록 휴폐업 확률 낮음**

03. 외부 데이터

[추가 아이디어]

2. 산업코드별 매출액

- 산업코드별 매출액 데이터를 기준으로 휴폐업 평균 매출액 산출
- 산출된 매출액 기준으로 기업의 매출액이 해당 매출액보다 적다면 휴폐업 확률이 높다고 예측

04. 결과 및 인사이트

1. 그래프(상승 또는 하강)의 일관성과 단변량 분석을 통해 휴폐업 예측에 유의미한 재무 데이터를 확인 및 모델링

2. 경제 기사 및 논문을 통해 얻은 인사이트를 토대로 종업원 수를 활용한 비재무 데이터를 생성, 재무 데이터 모델에 결합

3. 비재무데이터 결합으로 인해 모델의 평가지표(AUC)가 **0.9366에서 0.9527로** 향상

4. 종업원 수로 도출해낸 비재무데이터가 기업의 휴폐업 예측에 유의미하다는 인사이트를 발굴

5. 추가 아이디어(ESG 데이터 및 산업코드별 매출액) 제공

05. 참고문헌

강문상, "정부R&D 참여 중소기업 폐업의 재무적 영향 요인에 관한 연구", 예산정책연구 제5권 제1호, 2016년 5월

김상봉, Philip Ji, 조경준, "부도예측모형을 이용한 기업부실화의 원인분석", 2011년

박종원, 안성만, "재무비율을 이용한 부도예측에 대한 연구: 한국의 외부감사대상기업을 대상으로", 2014년 3월 24일

배재권, "재무비율을 이용한 기업부도예측 모형의 예측력 비교 연구", 2006년 1월

서울경제, '대기업 부도, 지난해 전혀 없었다", 서울경제, 2006년 1월 18일

오우석, 김진화, "인공지능기법을 이용한기업부도예측", 2017년 6월 29일

윤우섭, 김명종, "AUROC기반의 부도예측 앙상블 모형", 2021년 9월 30일

장제훈, "머신러닝 기법을 활용한 자영업자 폐업 예측 모형 연구: 서울시 26개 자치구를 중심으로", 2020년 12월 19일

중앙일보, "부도 통계 꼼꼼하게 낸다-자본금.종업원수 별로 전산처리", The JoongAng, 1996년 8월 17일자

한영철, 이용회, "신BIS협약이 경기순응성에 미치는 영향 - 국내은행의 필요자기자본 규모 변동에 관한 실증분석을 중심으로", 2004년 10월

감사합니다
Thank you!

이용우 김예지 박지영