2023 하계 URP 발표

심현구

최용원

강세현

INDEX

1

연구 소개

1 연구 소개

연구 주제

재무데이터와 비재무데이터를 활용한 중소기업 휴폐업 예측

비재무데이터

근로자 평균 연봉, 입사율, 퇴사율, 관련 뉴스 기사 등

재무관련 지표 외의 다양한 자료

1 연구 설계

연구 목적

휴폐업 예측 모형을 구축할 때 비재무적 요인의 중요성에도 불구하고 표준화와 주관성 등의 문제로 반영이 어려운 점을 극복하고 모형을 해석하고자 함

비재무적 요인 중 수집이 용이한 근로자 관련 데이터와 뉴스 텍스트데이터를 적극 활용하고자 함

재무지표만 사용한 모델보다 개선된 성능의 휴폐업 예측 모델 개발 및 개별 비재무 요인이 모델에 미치는 영향 해석

2

재무데이터 수집 및 전처리

2 재무데이터 수집 및 전처리

데이터 소개

데이터 종류	출처
중소기업 데이터	NICE DNB
근로자 평균 연봉	KreditJob
입사율 (연단위)	KreditJob
퇴사율 (연단위)	KreditJob
관련 뉴스 기사	NAVER news API

크롤링을
통해 수집

2 데이터 수집 및 전처리

변수 선택 | 단변량 분석

t-검정

재무 변수에 대해 t-검정을 통해

생존기업과 부도기업 간 평균의 차이가 있는지를 검토

P-value < 0.05 만족하는 변수만 선택

총 88개의 재무 변수에 대해

통계적으로 유의한 47개의 재무 변수 선택

※ 김용환(2022), 비정형 데이터를 활용한 머신러닝 기반 기업신용평가모형에 관한 연구

2 재무데이터 수집 및 전처리

재무데이터 소개

* 재무데이터는 보유 자료 중
가장 최신 자료를 기준으로 작성

* 변수명	* 변수명	* 변수명	* 변수명
유동자산	자본총계	매출액총이익률	유동부채비율
매출채권	매출액	매출액영업이익률	비유동부채비율
비유동자산	판매비와관리비	매출액순이익률	부채총계대매출액
유형자산	영업이익손실	경상수지비율	총자본회전율
자산총계	법인세비용차감전순손익	금융비용대매출액비율	재고자산회전율
유동부채	법인세비용	금융비용대부채비율	매출채권회전율
비유동부채	당기순이익손실	금융비용대총비용비율	매입채무회전율
부채총계	기업순이익률	부채비율	미수금
자본금	유보액총자산	차입금의존도	매출원가
이익잉여금결손금	유보액납입자본	자기자본비율	재고자산
			순운전자본비율

2 재무데이터 수집 및 전처리

파생변수 소개

** 변화율은 보유 자료 내에서
기업별 결산연도의 기점과 종점을 기준으로 산정

변수명	설명
재무결측치개수	해당 기업의 재무관련변수 결측치 개수
** 매출액총이익률 변화율	예) $\frac{\text{2020년도 매출액총이익률} - \text{2018년도 매출액총이익률}}{\text{2018년도 매출액총이익률}}$
** 금융비용대매출액비율 변화율	예) $\frac{\text{2021년도금융비용대매출액비율} - \text{2019년도금융비용대매출액비율}}{\text{2019년도금융비용대매출액비율}}$
** 금융비용대총비용비율 변화율	예) $\frac{\text{2019년도금융비용대총비용비율} - \text{2018년도금융비용대총비용비율}}{\text{2018년도금융비용대총비용비율}}$
변화율결측	변화율이 결측치인 재무관련변수의 개수
변화율이상치	변화율이 이상치인 재무관련변수의 개수
Target	휴폐업여부(binary) 0: 생존기업 1: 휴폐업기업

2 재무데이터 수집 및 전처리

재무 관련 데이터 결측치 전처리

결측치 포함 기업 단순 제거 시
너무 많은 데이터가 소실

평균 등 대표값으로 대체 시
업종 별 차이 반영이 어려움

재무 관련 데이터의 결측치는 0으로 대체

2 재무데이터 수집 및 전처리

재무데이터 최종 데이터셋

유동자산	매출채권	비유동 자산	…	미수금	매출원가	…	변화율 결측	변화율 이상치
615,988	185,615	42,385	…	0	2,503,923	…	3	5
30,077,990	6,034,264	15,251,977	…	968,621	22,532,564	…	0	0
28,431,311	15,498,763	21,334,735	…	0	34,484,342	…	1	0
5,409,774	943,751	4,856,884	…	40,558	4,294,008	…	3	0
2,196,004	990,945	174,426	…	269,350	10,056,444	…	1	0
12,027,994	2,281,118	10,155,381	…	0	0	…	8	0
…	…	…	…	…	…	…	…	…

3

재무데이터 모델링

3 재무데이터 모델링

재무데이터 모델 성능 비교

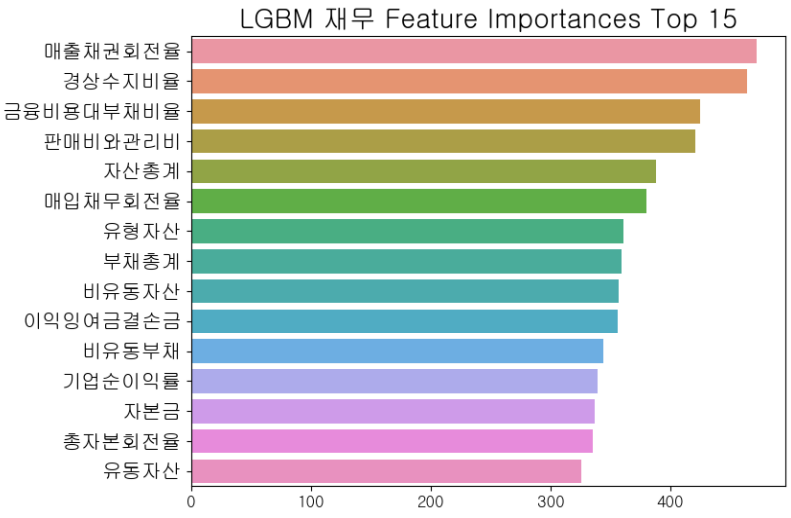
모델	AUC-ROC	F1 score	KS score
Decision Tree	0.820	0.769	0.640
Logistic Regression	0.845	0.630	0.491
SVM	0.611	0.364	0.220
Random Forest	0.962	0.810	0.694
XGBoost	0.969	0.810	0.694
LGBM	0.972	0.852	0.787
Tabnet	0.863	0.605	0.443

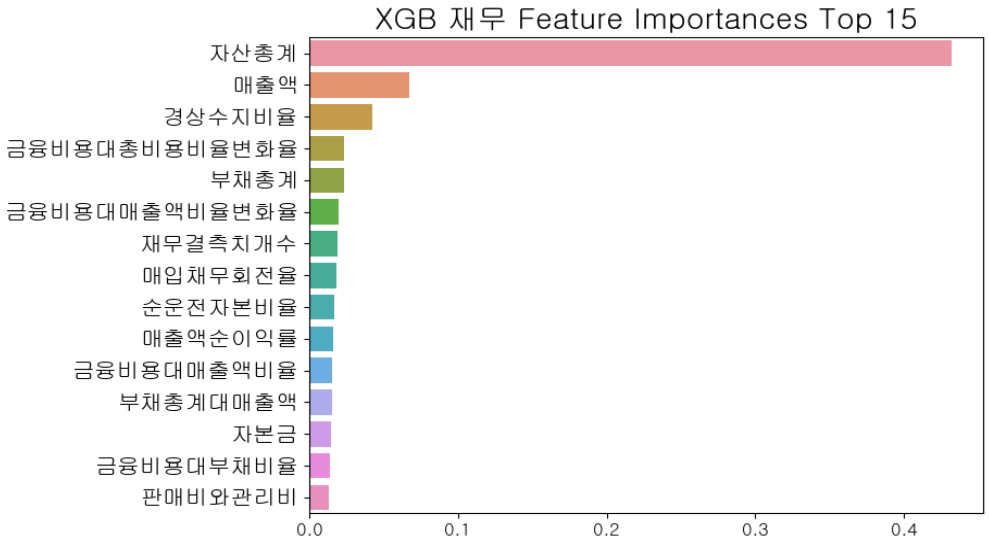
LGBM, XGBoost, Random Forest 등

트리 기반 머신러닝 모델의 성능이 우수

3 재무데이터 모델링

우수 재무 모델 Feature Importances





LGBM 중요변수

매출채권회전율, 경상수지비율

금융비용대부채비율

XGB 중요변수

자산총계, 매출액, 경상수지비율

4

비재무데이터 수집 및 전처리

4 비재무데이터 수집 및 전처리

비재무데이터 수집

네이버 뉴스

'뉴스의 감성분석이 부도예측 연구에서 예측 성능을 향상시킨다'는 선행연구

→ 기업명 기준 뉴스 기사 크롤링

크레딧잡

'인적자원관리가 기업성과에 영향을 준다'는 선행연구

→ 입사율, 퇴사율, 연봉을 크롤링

4 비재무데이터 수집 및 전처리

비재무데이터 수집

네이버 뉴스 API

변수명	설명
title	뉴스 제목
description	뉴스 요약문
기업명	기업명
pubDate	뉴스 발행 날짜
link	뉴스 URL
contents	뉴스 내용 전문

- 네이버 뉴스 API 이용해 대상 기업명과 정확히 일치하는 뉴스의 제목과 본문을 수집
- 분석 정확성을 위해 크롤링 대상을 네이버 뉴스로 한정

크레딧잡 (wanted insight)

변수명	설명
BIZ_NO	사업자번호
기업명	기업명
연봉	국민연금 출처 예상평균연봉
입사율	국민연금 출처 입사율
퇴사율	국민연금 출처 퇴사율

- 사업자 번호에 대응되는 기업명을 크레딧잡에 검색해 해당 기업의 연봉, 입사율, 퇴사율 수집
- 파이썬 Selenium을 이용해 크레딧잡의 공개 데이터를 크롤링

비재무데이터 수집

NICE DNB

변수명	설명
HDOF_BR_GB	본점지점구분 1.본점 2.지점
FR_IVST_CORP_YN	국외투자법인여부
VENT_YN	벤처기업여부
MDSCO_PRTC_YN	중견기업보호여부
ESTB_GB	설립구분
EMP_CNT	직원수
Industry	한국표준산업분류

- NICE DNB 제공
- EMP_CNT(직원수) 제외 모두 범주형 변수
- EMP_CNT(직원수) 결측치
  : 동일 산업군 내 평균으로 대체
- ESTB_GB(설립구분)
  : 01 = 주식, 02 = 합자, 04 = 유한,
  05 = 조합, 06 = 정부투자기관, 07 = 개인,
  10 = 단체협회, 99 = 기타, 결측치 = 미분류
- Industry : 통계청 제공 산업 분류 코드 사용.
  결측치는 '미분류' 처리

4 비재무데이터 수집 및 전처리

비재무데이터 전처리 | ① 뉴스 데이터

1. 회사명 中 식별력 낮은 기업 제거

(21세기, 엑소메드, 좌우지간, 코뿔소…)

2. 기사 내용에 회사 이름을 포함하지 않는 뉴스 데이터 제거
3. 구두점(.) 기준으로 분석 대상 문장 분리
4. 기타 불필요한 문장 제거 (언론사 소개 등)

최종적으로 **6,462개 기업**에 대해

약 1,400,000개의 유효한 문장 데이터 수집

4 비재무데이터 수집 및 전처리

텍스트데이터 자연어처리(NLP)

KLUE-BERT 모델에
금융감성 label 사전학습

사전학습된 모델에
수집한 텍스트데이터를
문장 단위로 입력

반환된 문장별 label의 비율과
기사 수를 데이터셋에 추가

문장 예시	감성 label
20% - 40% 범위의 장기적인 순매출 성장을 목표로 하고 있으며, 영업이익률은 순매출액의 10% - 20%를 목표로 하고 있습니다.	1 (긍정)

4 비재무데이터 수집 및 전처리

텍스트데이터 자연어처리(NLP)

KLUE-BERT 모델

- 벤치마크 데이터인 KLUE에서 베이스라인으로 사용되었던 모델
- 모두의 말뭉치, CC-100-Kor, 나무위키, 뉴스, 청원 등 문서에서 추출한 63GB의 데이터로 학습
- KoBERT에 비해 더 최신화되고 다양한 출처의 데이터로 학습되어 성능이 우수
- vocab size는 32,000(KoBERT의 약 4배), 모델의 크기는 111M

4 비재무데이터 수집 및 전처리

텍스트데이터 자연어처리(NLP)

파생변수 생성

전체 기사 수, 긍정·부정·중립 기사 비율

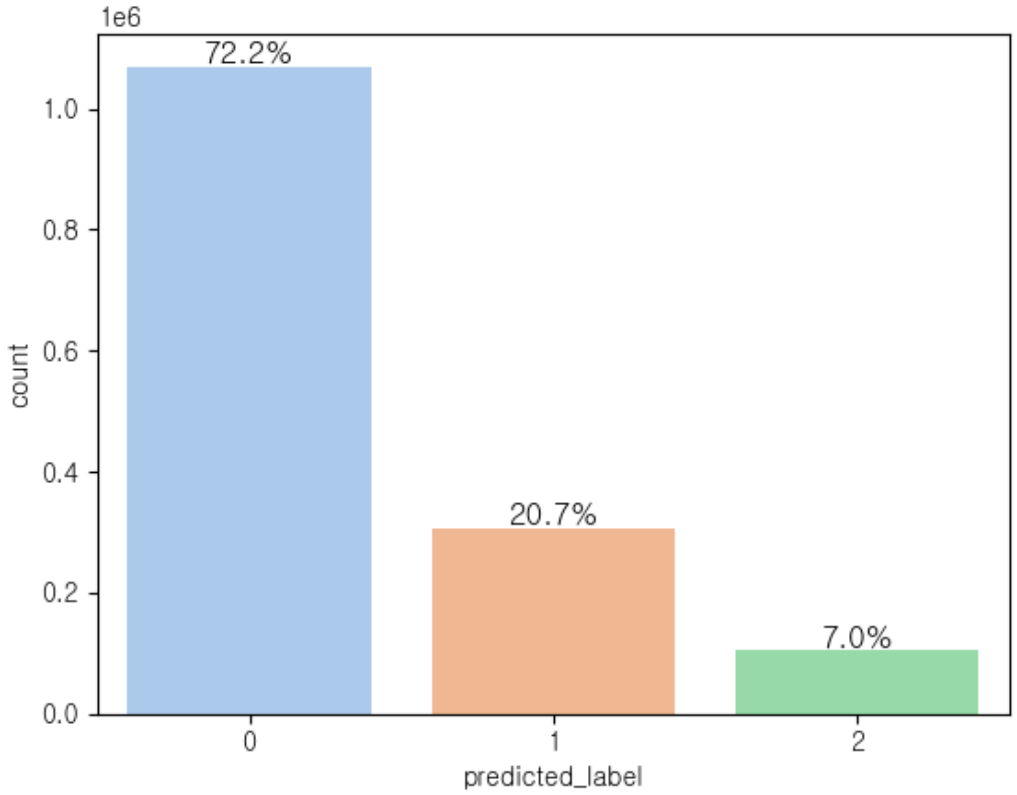
※ 결측치는 0으로 대체

기업명	기사 내 문장 예시
율성건설	23개사 모임으로 시작한 차세대 클럽은 10여년만인 올해 회원사가 72개로 늘었고...
율성건설	제 월급은 대기업 다닐 때보다 반으로 줄었고요

기업명	기사 수	긍정	부정	중립
율성건설	40	0.82	0.10	0.07

4 비재무데이터 수집 및 전처리

텍스트데이터 자연어처리(NLP)



금융 감성분석 결과

분류된 문장 수

중립(0) > 긍정(1) > 부정(2)

비재무데이터 전처리 | ② 직원 관련

결측치 처리

전체 35,767개의 기업에 대해

연봉 71%, 입사율 50.0%, 퇴사율 51.2%의 유효 데이터 획득

MICE 사용해 다중 대체 (Method = ‘rf’)

입퇴사율의 경우 결측 비율이 높아 제외 전·후 결과를 비교함

비재무데이터 전처리 | ② 직원 관련

모델	AUC-ROC	KS score
입퇴사율 제외 LGBM	0.983	0.818
입퇴사율 포함 LGBM	0.984	0.826

입퇴사율을 제외한 모델에 비해
입퇴사율을 포함한 모델 성능이 개선

입퇴사율을 모델링 변수에
포함하기로 결정

4 비재무데이터 수집 및 전처리

비재무데이터 추가 최종 데이터셋

유동자산	… (재무변수)	기사수	중립	긍정	부정	연봉(만원)	입사율	퇴사율
24,965,602	…	0	0	0	0	3,845	6	28
615,988	…	963	0.753	0.149	0.096	3,436	66	51
1,596,577	…	0	0	0	0	3,993	60	135
9,903,117	…	149	0.771	0.154	0.073	3,005	60	145
1,289,911	…	0	0	0	0	3,867	41	57

분석에 필요한 최종 데이터셋 구성 완료

설명변수 62(재무 47 + 비재무 15) & 종속 변수 1(휴폐업 여부)

5

재무+비재무데이터 모델링

5 재무+비재무데이터 모델링

재무+비재무 모델 성능비교

모델	AUC-ROC	F1 score	KS score
Logistic Regression	0.903	0.715	0.616
Random Forest	0.976	0.824	0.712
XGBoost	0.983	0.885	0.826
LGBM	0.984	0.880	0.826

LGBM, XGBoost , Random Forest 등

트리 기반 머신러닝 모델의 성능이 우수

5 재무+비재무데이터 모델링

재무 모델 vs 재무+비재무 모델

모델	AUC-ROC	F1 score	KS score
Logistic Regression	0.845	0.630	0.491
Random Forest	0.962	0.810	0.694
XGBoost	0.969	0.810	0.694
LGBM	0.972	0.852	0.787

[재무 모델]

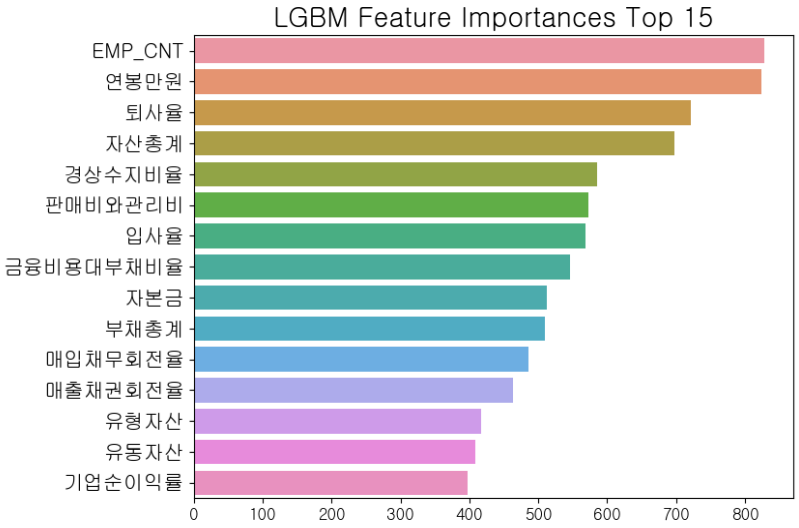
모델	AUC-ROC	F1 score	KS score
Logistic Regression	0.903	0.715	0.616
Random Forest	0.976	0.824	0.712
XGBoost	0.983	0.885	0.826
LGBM	0.984	0.880	0.826

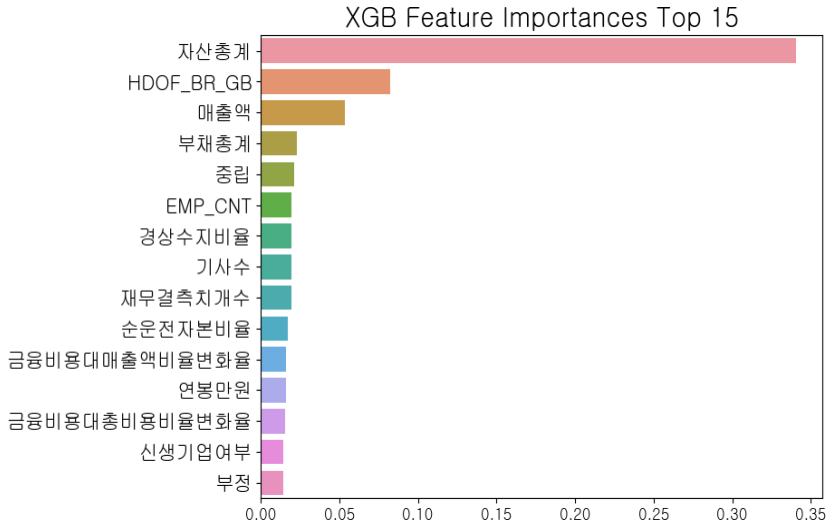
[재무+비재무 모델]

비재무 변수를 추가했을 때 전반적으로 **성능이 개선**

5 재무+비재무데이터 모델링

재무+비재무 모델 Feature Importances





LGBM 중요변수

총 직원수, 연봉, 퇴사율

XGB 중요변수

자산총계, 본점지점구분, 매출액

폐업 예측에 대한 SHAP

Feature importances의 한계

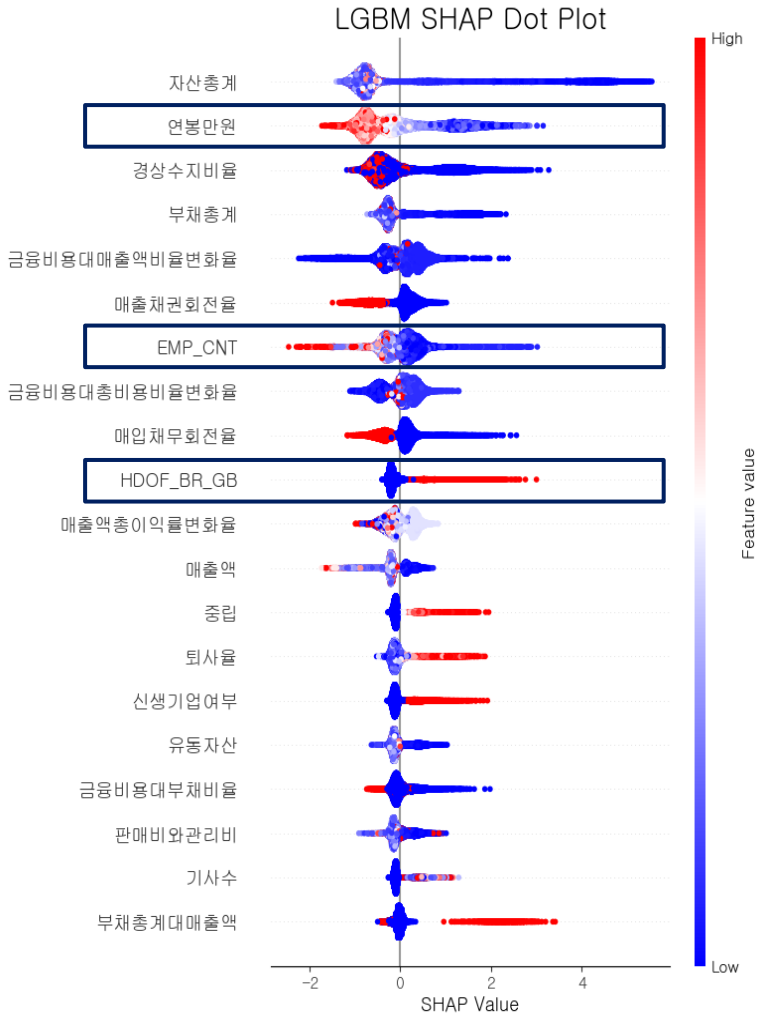
- 변수가 어떤 방향으로 영향을 미치는지 확인할 수 없음
- 중요도에 따라 모델 수정 시 해당 변수의 중요도가 inconsistent 함

SHAP(SHapley Additive exPlanation) 도입

변수중요도 뿐만 아니라 개별 예측값에 대한 각 변수들의 영향력을 모형 클래스에 상관없이 누적으로 배분하는 방식으로, 인간의 생각과 유사한 해석을 제공하며, consistent하고 변수중요도의 방향성을 알 수 있음

5 재무+비재무데이터 모델링

폐업 예측에 대한 SHAP

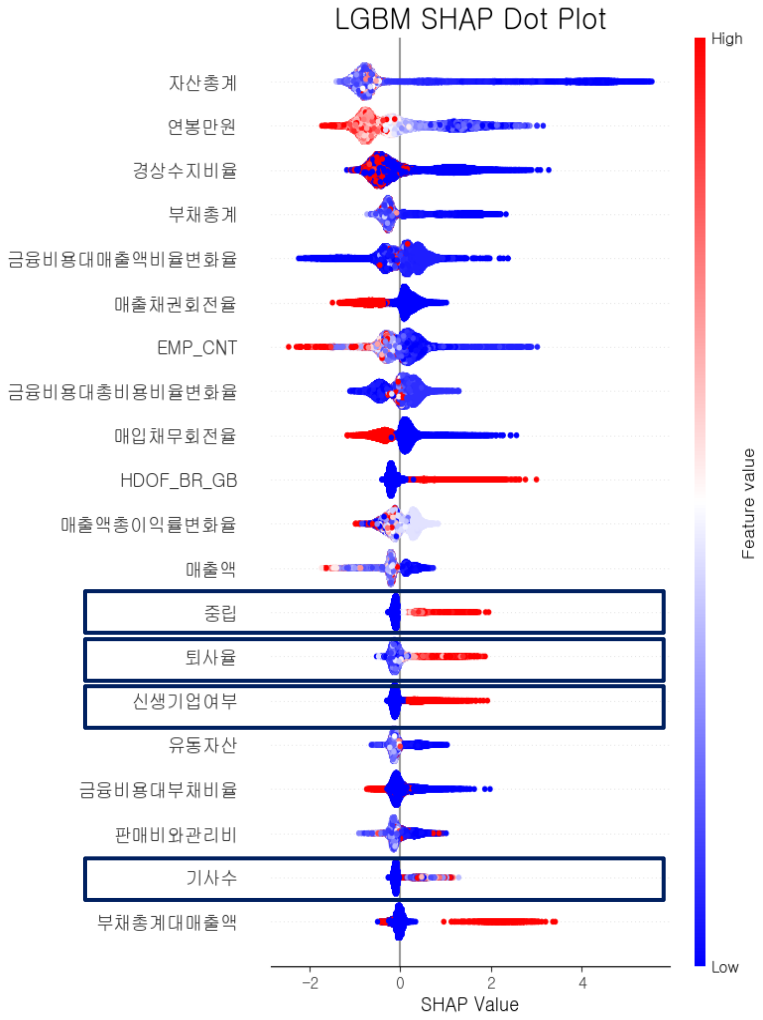


SHAP 기반 중요도 해석

- 연봉: 폐업률과 음의 관계
- EMP_CNT(직원 수): 폐업률과 음의 관계
- HDOF_BR_GB(본점기업코드): 지점 여부와 폐업률이 양의 관계

5 재무+비재무데이터 모델링

폐업 예측에 대한 SHAP



SHAP 기반 중요도 해석

- 중립(중립기사비율): 폐업률과 양의 관계
- 퇴사율: 폐업률과 양의 관계
- 신생기업여부: 폐업률과 양의 관계
- 기사 수: 영향력 파악이 어려움

6

결론

6 결론

기대 효과 | ① 연구적 측면

모델 성능 향상 확인을 통해
중소기업 휴폐업 예측에서
비재무데이터의 중요성 파악

비재무데이터의 중요성 제고로
양질의 비재무데이터
생산 및 수집 증가

6 결론

기대 효과 | ② 사회적 측면

기업

- 재무적 요소와 비재무적 요소의 균형 있는 성장 추구 → ESG 경영
- 재무 데이터가 충분하지 않은 신생 기업 평가 변수의 다양화

정부

- 초기창업기업 지원금 제공 시 기업의 잠재력을 고려한 선별적 지원 가능
- 중기금융 기관 대출 시 비재무지표를 고려한 심사로 부실채권의 비율 완화

개인

- 인적자원관리 능력을 중소기업의 전망과 결부시켜 기업 평가 가능
- 기업 투자 시 고려해볼 비재무적 요인에 대한 정보 제공

6 결론

제언

모델링 대상 중소기업 관련 기사 수가 부족하다는 한계

→ 대상 기업의 기사수가 충분하다면 '뉴스 기사'를 비재무변수로 활용 가능

휴폐업 예측에 유의미한 변수로 도출되었던 연봉, 퇴사율을

중소기업 신용평가 시 비재무변수로서 적극 사용할 것을 제안

경영자 개인의 역량 및 재무 상태를 파악할 수 있는 지표를 획득해 활용한다면

모델 성능 개선에 도움이 될 것으로 예상

별첨

별첨

통계청 제공 비재무변수

변수명	설명
상용근로자남	남자 상용근로자 수
상용근로자여	여자 상용근로자 수
임시근로자남	남자 임시근로자 수
임시근로자여	여자 임시근로자 수
조직형태코드	법인여부(binary) 회사법인: 0 회사이외법인: 1

통계청 제공 데이터 중 사업자등록번호 기준
비재무 지표를 추가해 모델링

별첨

통계청 제공 비재무변수 활용 모델링

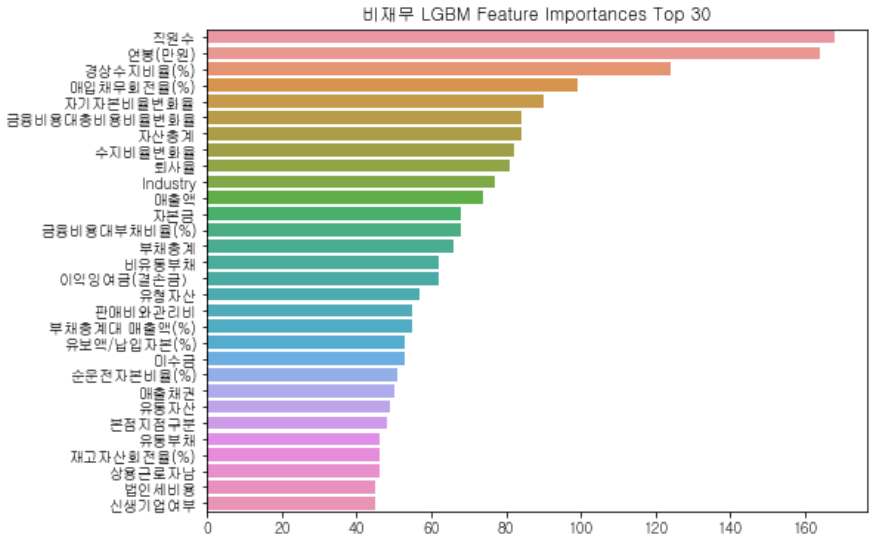
모델	재무 모델링			재무+비재무 모델링		
	AUC-ROC	F1 score	KS score	AUC-ROC	F1 score	KS score
Logistic Regression	0.838	0.64	0.577	0.893	0.70	0.652
SVM	0.802	0.58	0.604	0.851	0.69	0.702
Random Forest	0.962	0.79	0.672	0.973	0.79	0.667
XGBoost	0.966	0.81	0.719	0.975	0.84	0.761
LGBM	0.966	0.81	0.726	0.980	0.85	0.778

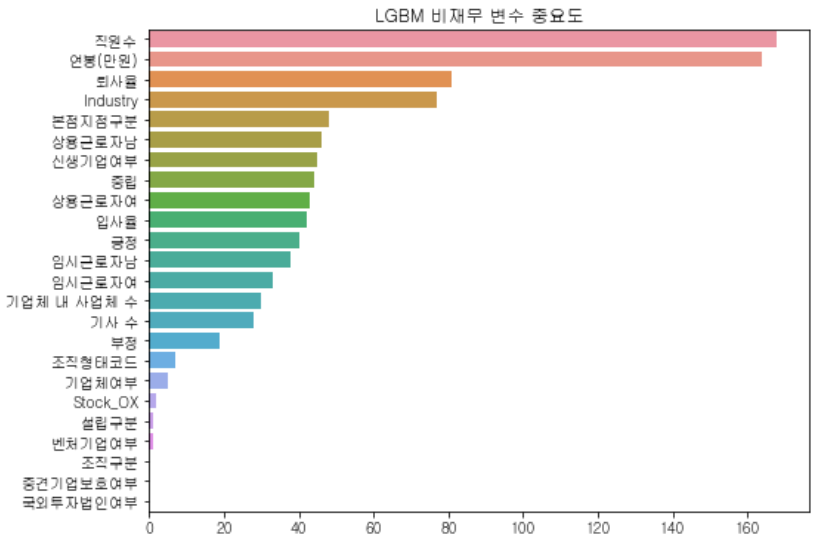
성능 증가폭이 가장 큰 모델: Logistic Regression

전반적인 성능이 가장 우수한 모델: LightGBM

별첨

통계청 제공 비재무변수 활용 모델링



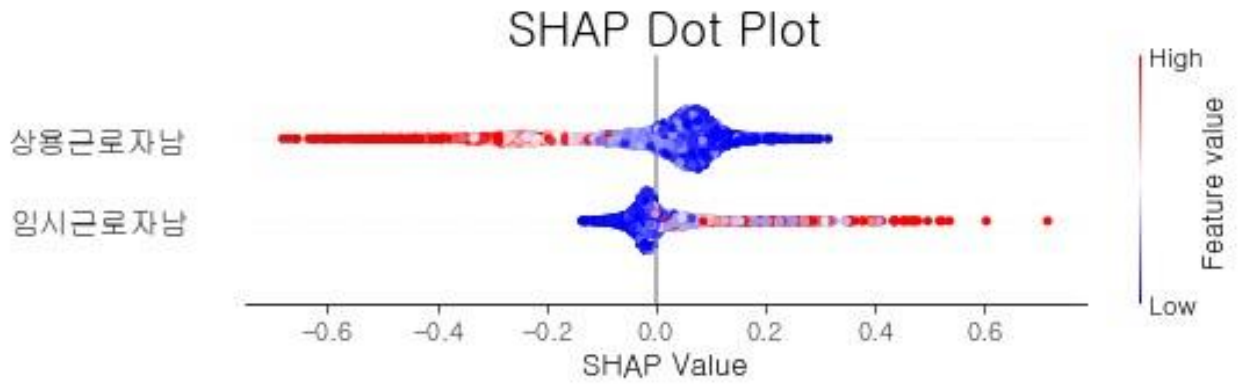


LightGBM 모델 기준 직원수, 연봉, 퇴사율 등의

비재무변수가 높은 중요도를 보임

별첨

통계청 제공 비재무변수 활용 모델링



상용근로자남 : 폐업률과 음의 관계

임시근로자남 : 폐업률과 양의 관계

참고문헌

- 현지원, 이준일, 조현권(2022). KoBERT를 이용한 기업관련 신문기사 감성 분류 연구. 회계학연구. 47권(4호)
- 정승호(2022). TabNet을 활용한 딥러닝 성능 비교와 설명가능한 AI 활용성에 대한 연구-기업 신용평가 모형을 중심으로. 서강대학교 대학원.
- 남윤미(2017). 국내 자영업의 폐업률 결정요인 분석. BOK 경제연구. 5
- 장제훈(2021). 머신러닝 기법을 활용한 자영업자 폐업 예측 모형 연구 서울시 25개 자치구를 중심으로. 인문사회. 12권(1호)
- 박준식, 최진, 이성호(2023). 머신러닝을 이용한 소상공인 창업기업의 폐업 예측 모형 개발-도소매산업을 중심으로. 한국창업학회지. 18권(1호)
- 유원규, 이철규(2015). 비재무적 요인이 중소벤처기업의 신용평가에 미치는 영향. 대한경영학회지. 28권(12호)
- 김용환(2022). 비정형 데이터를 활용한 머신러닝 기반 기업 신용평가 모형에 관한 연구. 숭실대학교 대학원.
- 방준아, 손광민, 이소정, 이현근, 조수빈(2018). 서울 치킨집 폐업 예측 모형 개발 연구. 한국빅데이터학회지. 3권(2호)
- 김용환, 김도형, 허재혁, 김광용(2022). 오차 앙상블모형을 활용한 기업 신용평가모형의 비교 연구. The Journal of Korean Institute of Communications and Information Sciences. 47권(1호)

감사합니다