「2023년 통계데이터 활용대회」 데이터분석 보고서

제 목	비재무데이터를 활용한 중소기업 휴폐업예측 모델 제안 및 결정요인 분석

비재무데이터를 활용한 중소기업 휴페업 예측 모델 제안 및 결정요인 분석

1. 배경

□ 주제 선정

○ **주제 선정 배경**

고금리, 고물가, 고환율에 수출과 내수 침체가 겹친 복합 위기가 길어지면서 많은 중소기업이 재정위기에 놓였다. '노란우산공제'에 따르면, 2022년 지급된 폐업 공제금이 역대 최대 규모로 집계되는 등 중소기업의 경영난이 심화되고 있다. 이러한 상황에서 대기업 및 중견기업에 비해 투자유치에 한계가 있는 중소기업의 경우, 자금 조달의 수단으로 대출을 택할 수밖에 없는 현실이다(정승호, 2022). 이에 중소기업 대상 여신 리스크 관리의 필요성이 대두되고 있다. 따라서 대출 시행 전 건전성이 높은 기업을 선별하기 위해 중소기업 휴폐업 요인을 파악하는 것은 중요한 과제이다.

이에 본 연구에서는 중소기업의 휴폐업 예측을 위해 사용되는 전통적 지표인 재무변수와 더불어 비재무변수를 활용한 모델링을 통해 부실기업 선별에 대한 새로운 지표를 마련하고자 한다.

□ 분석 필요성(문제점) 및 전략

○ **재무적 요인만을 활용한 중소기업 휴폐업 예측의 한계**

기존의 휴폐업 예측 모델은 재무데이터 위주로 접근했다. 하지만 기존의 재무데이터는 수치적인 정보만을 포함하기에 기업의 전망 및 영업 환경과 같은 질적 정보는 포함하지 못하고, 이는 결국 중소기업의 잠재성 및 성장성을 온전히 반영하지 못하는 한계로 이어졌다(유원종 외, 2015).

또한, 외부감사 대상이 아닌 중소기업의 경우 대기업이나 중견기업에 비해 상대적으로 재무적 신뢰성이 열악하다. 따라서 재무적 요인으로만 휴폐업을 예측하는 것은 바람직하지 못한 예측 결과를 초래할 수 있다는 문제점이 있다(김소정 외, 2022).

○ **비재무 요인의 제한적 활용**

일부 연구에서 중소기업의 경영 능력에 대한 비재무변수의 중요성을 파악했음에도 회계 관련 비재무적 데이터만을 사용하거나(유원종 외, 2015), 뉴스 데이터의 경우 그 키워드만을 분석에 사용하는 등(김용환, 2022) 신용평가에 대한 비재무 데이터의 활용은 다소 제한적으로 이뤄졌다.

이에 본 연구는 다양한 비재무변수를 적극 활용해 기본적인 통계모델인

로지스틱 회귀 모형과 머신러닝 알고리즘으로 중소기업의 휴·폐업 예측을 진행하고자 한다.

2. 데이터 분석

□ 데이터 선정(사용한 데이터 및 이유 등)

○ **재무데이터 소개 및 선정 이유**

- 중소기업 재무데이터(사업자등록번호 기준)
  - 통계데이터센터에서 제공하는 기업별 비재무데이터를 추가하기 위해 사업자등록번호 또는 기업명이 명시된 재무제표 데이터를 활용했다.
- 파생변수
  - 기업의 재무 상태 변화를 반영하기 위해 '변화율'을 파생변수로 추가했다.
  - 기업의 다수의 재무상태가 비공개된 점도 휴폐업 여부에 영향을 미칠 것이라 예상하여 '재무결측치 개수'를 파생변수로 추가했다.
  - 데이터의 무결성 여부를 반영하기 위해 '변화율 결측치 개수'와 '변화율 이상치 개수'를 파생변수로 추가했다.

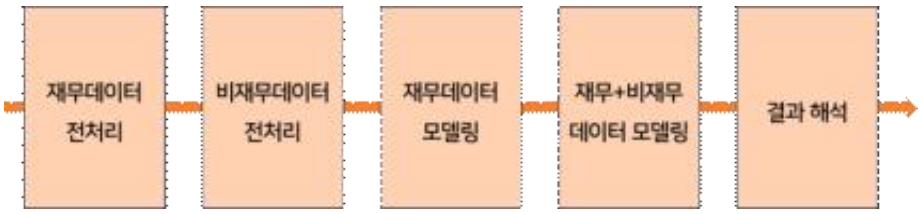
○ **비재무데이터 소개 및 선정 이유**

- 인적자원관리 관련 데이터(기업 통계 등록부 中 상용·임시 근로자 관련 변수, 총 직원수, 입사율·퇴사율 변수, 연봉 변수)
  - '인적자원관리 능력이 기업성과에 영향을 준다'는 선행연구를 토대로, 관련한 변수를 선정했다(김광현 외, 2019). 특히 '입·퇴사율'의 경우 재무성과에 양의 영향력을 준다는 연구결과가 있었고, 상용·임시 근로자 관련 변수의 경우 고용형태가 휴폐업에 미치는 영향을 볼 수 있다는 점에서 모델링에 사용할 변수로 채택했다. '총 직원 수'의 경우 인적자원의 양적지표로서 가장 적절하다고 판단해 모델링에 사용했다.
- 기업 관련 텍스트 데이터(기업별 긍정·부정·중립 기사 비율, 전체 기사 수)
  - '뉴스 감성분석이 부도 예측 연구에서 예측 성능을 향상시킨다'는 선행연구 결과를 토대로, 기업 이름을 기준으로 네이버 뉴스 API 활용 크롤링을 진행했다(김찬송 외, 2019). 그리고 해당 뉴스 데이터에 대한 감성분석을 실시해 뉴스 문장별 긍정·부정·중립 라벨을 얻고 관련 파생변수를 생성했다. 뉴스의 양에 따른 라벨 수의 차이를 완화하고자 라벨의 '비율'을 모델링의 변수로 선정했으며, 동시에 뉴스 기사량의 차이

는 반영하기 위해 ‘전체 기사수’ 도 변수로 선정했다.

□ 데이터 분석(분석 프로세스, 분석방법, 접근방법 등)

○ **분석 프로세스**



○ **데이터 전처리**

- 재무데이터 전처리
  - 변수선택: t-검정을 통해 변수마다 생존기업과 부도기업 간 평균의 차이가 있는지 확인 후 p-value가 0.05 미만인 47개의 변수만을 선택했다.
  - 결측치 및 이상치: 결측치의 경우 분석에 영향을 미치지 않는 값인 0으로 대체했다(김용환, 2022). 머신러닝 모형에서는 반드시 이상치를 제거할 필요성이 없기에 이상치 제거를 하지 않고 진행했다.
- 비재무 데이터 전처리(결측치 대체)
  - ‘총 직원수’의 경우에는 동일 산업군 직원수 평균으로 결측치를 대체했다.
  - ‘뉴스 기사’의 경우에는 기업과 무관한 기사를 수작업으로 삭제하는 과정을 거쳤고, 감성 분석 모델링을 위해 구두점(.)을 기준으로 데이터를 문장 단위로 분할했다. 그리고 문장 단위로 분류된 기사에 대해 감정 라벨링에 필요하지 않은 언론사와 기자 소개 등이 포함된 문장을 삭제했다.
  - ‘연봉’, ‘입퇴사율’ 데이터의 경우에는 결측치의 양이 타 데이터에 비해 많아 MICE[1]를 활용해 다중대체했다.

○ **금융 감성분석 모델링**

- 한국어 대용량 사전학습 언어모델인 KLUE-BERT 모델을 사용해 금융 감성분석 모델을 구축했다. 해당 모델은 자연어처리 분야에서 많이 사용되어

1) 하나의 값으로 결측치를 채워넣는 일반적인 결측치 대체 방법과 달리, 결측치를 여러 변수를 고려해 채워넣는 방법.

온 KoBERT 모델에 비해 더 최신화되고 다양한 출처의 데이터로 학습이 되어 성능이 우수하다.

- 금융 뉴스 데이터에서 추출된 4,840개의 문장으로 사전학습된 모델에 수집한 중소기업 관련 뉴스 텍스트 데이터를 입력해 긍부정 라벨링을 시행했다. 해당 라벨링 결과 중립 기사의 비율이 가장 높았다.

○ **재무데이터 모델링**

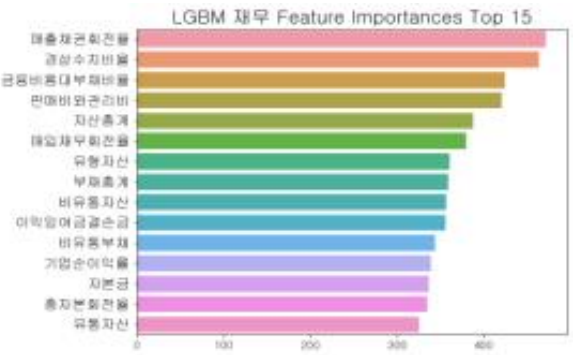
- 재무데이터만을 사용한 모델 중 최고성능의 모델을 찾고자 5가지 모델을 사용하여 성능을 비교했다.

모델	AUC-ROC	F1 score	KS score
Logistic Regression	0.838	0.64	0.577
SVM	0.802	0.58	0.604
Random Forest	0.962	0.79	0.672
XGBoost	0.966	0.81	0.719
LGBM	0.966	0.81	0.726

표1 - 재무데이터 모델링 성능 비교

- 휴폐업 예측 모델의 성능평가지표로 가장 많이 사용되는 AUC-ROC와 KS score를 기준으로 트리기반모델인 XGBoost(XGB)와 LightGBM(LGBM)의 성능이 가장 우수하여 해당 모델의 변수중요도를 확인해보았다.

그래프 1 - XGB 재무변수 중요도



그래프 2 - LGBM 재무변수 중요도

- LGBM에서는 매출채권회전율, 경상수지비율, 금융비용대부채비율 등이 중요변수로 작용했고, XGB에서는 자산총계, 매출액, 경상수지비율 등이 중요변수로 작용했다. 성능평가지표를 개선하고 다양한 데이터를 활용하고자 수집한 비재무변수를 추가하여 모델을 구축했다.

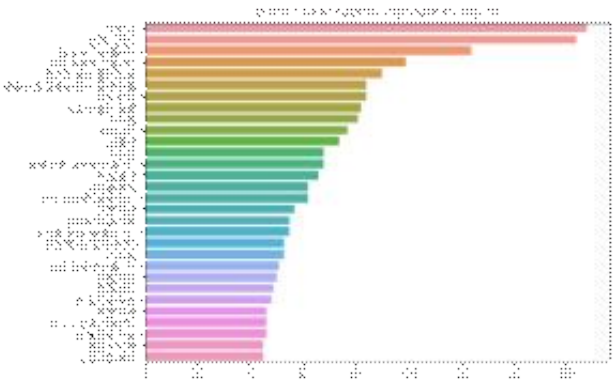
○ **재무·비재무 데이터 통합 모델링**

- 재무 및 비재무 데이터를 통합한 모델링은 재무데이터를 활용한 모델링과 마찬가지로 기본적인 통계모형인 로지스틱 회귀 모형과 머신러닝 모형인 SVM, Random Forest, XGB, LGBM 등을 통해 분석이 진행되었으며, 그 전반적인 성능은 아래와 같다.

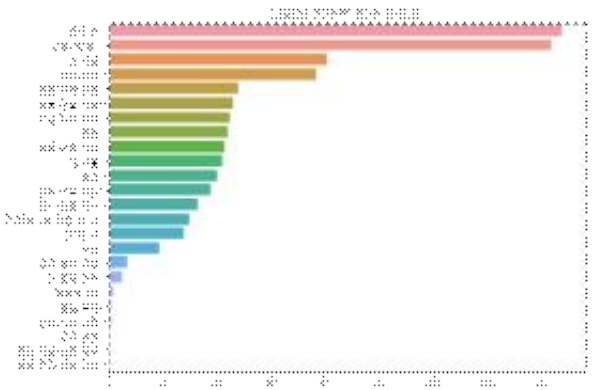
모델	AUC-ROC	F1 score	KS score
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
XGBoost	0.975	0.84	0.761
LGBM	0.980	0.85	0.778

표 2- 재무・비재무 모델링 성능 비교

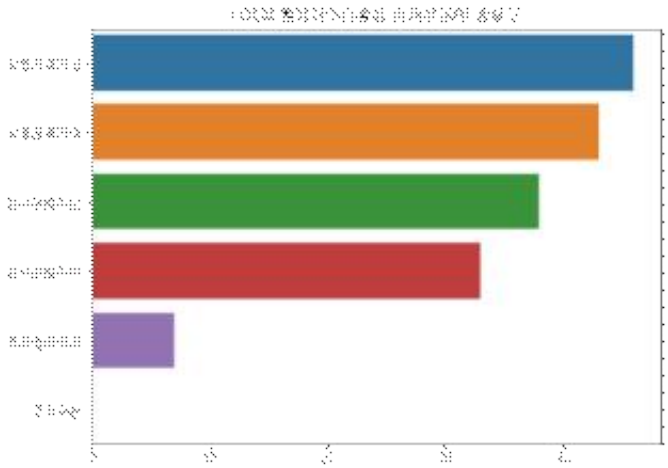
- 재무와 비재무 데이터를 동시에 고려한 모델링 역시 XGB와 LGBM 등 트리 기반 모델이 전반적으로 우수한 성능을 드러냈다. 비재무 데이터의 추가로 인한 성능의 증가폭은 로지스틱 회귀 모델이 가장 컸으며, 전반적으로는 LGBM의 성능이 가장 우수했기에 이후의 분석 및 해석은 LGBM을 기준으로 함을 밝힌다.



그래프 3 - LGBM 재무・비재무 변수 중요도



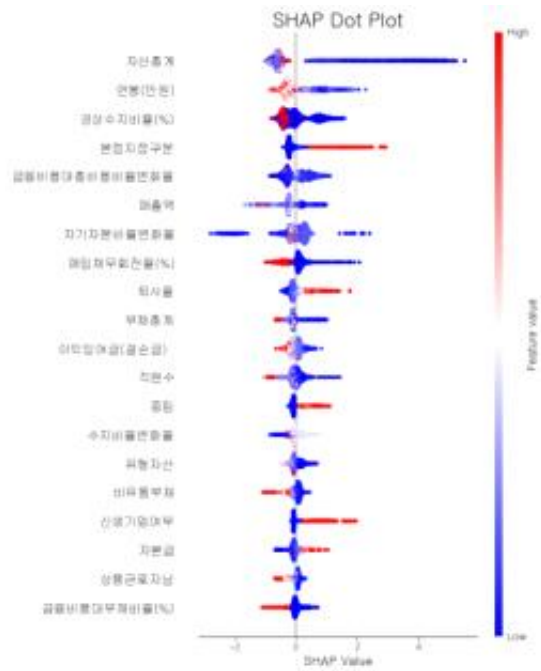
그래프 4 - LGBM 비재무 변수 중요도



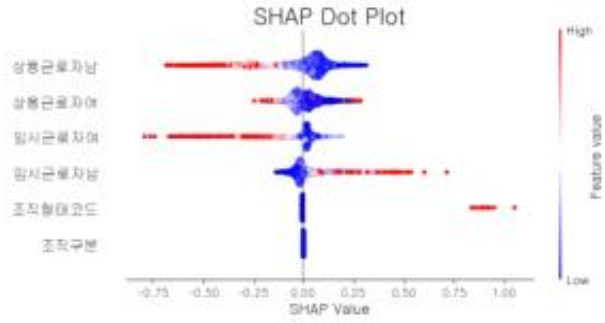
그래프 5 - LGBM 인적 자원 관련 비재무 변수 중요도

- 변수중요도를 확인한 결과 직원수, 평균 연봉, 경상수지비율(%) 등이 예측에 있어 중요변수로 작용한 것을 확인할 수 있었다. 비재무변수의 경우, 상기한 직원수와 평균 연봉 이외에 퇴사율 및 산업 구분, 본점·지점 구분 여부 등이 중요변수로 작용함을 파악할 수 있었다.

- 보다 정확한 변수의 영향력을 파악하기 위해 SHAP[2]을 통해 해석한 결과는 다음과 같았다.



그래프 6 - LGBM 재무·비재무 변수 SHAP



그래프 7 - 통계청 비재무 변수 SHAP

- 비재무변수의 경우, 연봉과 직원수는 그 값이 적은 기업의 폐업률이 높은 경향을 보이는 음의 관계가 나타났으며, 퇴사율, 중립뉴스비율은 그 값이 큰 기업이 폐업률이 높은 경향을 보이는 양의 관계가 드러났다. 또한 본점보다는 지점이, 기존 기업보다는 신생기업이 폐업률이 더 높은 경향을 보였다.
- 사용한 통계청 비재무 데이터 중에서는 상용 남자 근로자와 임시 여자 근로자가 그 값이 높은 기업의 폐업률이 낮은 경향을 보이는 음의 관계가 나타났다. 반면, 임시 남자 근로자는 그 값이 높은 기업의 폐업률이 높은 경향을 보이는 양의 관계가 드러났다.

3. 분석 활용 전략

□ 기대효과 및 향후과제

○ **중소기업**

- 기업 평가 시 비재무적 요소가 함께 고려된다는 점을 인지하여 중소기업들이 재무적 요소와 비재무적 요소의 균형성장을 추구할 것이다.
- 비재무데이터를 활용한 신용평가를 통해 재무적 데이터가 부족해도 우수한 비재무적 자원을 보유한 신생기업의 신용도 제고를 기대할 수 있다.

2) 머신러닝 모델의 예측 근거를 제시하고, 모델에서 각 피처(Feature)의 중요성을 결정하는 도구

○ **공공기관 및 공기업**

- 현재 정부는 창업기업을 대상으로 금전적 지원을 하는 초기창업패키지 등의 사업을 운영하고 있다. 하지만 정부 지원금만 받고 제대로 된 매출을 발생하지 않는 '체리피커형 창업'에 대한 문제가 제기됐다(성상훈, 2021). 본 연구를 통해 제시하는 비재무 변수를 사용한다면 성장잠재력이 높은 기업을 선별해 '유의미한 지원'으로 연결하는 데에 큰 도움이 될 수 있다.
- 정책금융기관 대출 시에 비재무 지표를 고려한 심사로 부실채권의 비율을 완화할 수 있다. 정책금융기관 특성상 중소기업 지원이 활발해 신용위험이 높은 기업 여신이 많다(강수인, 2019). 정책금융기관이 대출을 시행할 때 부실기업을 사전에 선별할 수 있도록 비재무 데이터를 활용한 휴폐업 예측 모델을 사용하는 것은 큰 도움이 될 것이다. 특히 본 연구를 통해 도출한 휴폐업 예측에 유의미한 비재무변수인 근로 형태, 연봉, 입·퇴사율 등을 적극 활용하면, 부실기업에 대한 변별력을 높일 수 있을 것이다.

○ **사회**

- 중소기업의 가치에 대해 보다 정확한 판단이 가능해짐에 따라, 개인 및 기업의 적극적 투자가 가능해진다. 이는 곧 유망 중소기업에 대해 성장동력으로 작용할 것이다.
- 입·퇴사율 등의 비재무 변수를 활용한 휴폐업 예측으로 인적자원관리 능력의 중요성이 증대된다. 이는 경영자의 업무환경 개선을 위한 노력으로 이어져 직원복지 수준이 상승할 것이다. 중소기업의 전반적 업무환경 개선은 중소기업 취업 활성화 및 우수 인재 유입이라는 긍정적 결과를 불러올 것이다.

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)

김광현, & 동학림. (2019). 사람중심 기업가정신이 중소기업 핵심역량과 기업성과에 미치는 영향. 한국콘텐츠학회논문지, 19(5).

김찬송, & 신민수. (2019). 부도예측 모형에서 뉴스 분류를 통한 효과적인 감성분석에 관한 연구. 한국 IT 서비스학회지, 18(1), 187-200.

성상훈.(2021). 청년창업기업 3곳 중 2곳 '매출 0'. 한국경제신문, https://www.hankyung.com/politics/article/2021101507481.

강수인. (2023). 기업은행, '부실채권 압도적' 건전성 적신호…"어쩔 수 없다". NSP통신. https://www.nspna.com/news/?mode=view&newsid=634495.