NICE DNB 데이터를 활용한 중소기업 휴폐업 예측 챌린지

1 서론

코로나 19를 기점으로 산업구조·국제환경 등이 빠르게 변화하면서, 기업의 산업 환경이 크게 변화하였다. 먼저 코로나 19로 인한 직접접인 산업구조 변화는 다음과 같다. 1) 비대면 서비스가 각광을 받으며 정보통신산업, IT산업, 운수 및 창고산업이 부상했고, 2) 백신 생산과 관련된 바이오 산업 역시 수혜를 받았으며, 3) 질병 이슈를 계기로 환경에 대한 관심이 강화되어 전기차 및 2차전지 산업도 성장하였다. 4) 한편 오프라인 중심의 숙박·음식점, 예술·스포츠·여가 업종은 쇠퇴했다.

국제환경의 변화 역시 이러한 산업 환경 변화에 일조했는데, '기업에 대한 글로벌 패러다임 변화'가 그 중 하나이다. 질병이슈를 계기로 환경문제·사회불평등 등 각종 사회문제에 대한 개개인의 관심이 높아짐으로써, 성장뿐만이 아니라 사회문제도 해결 가능한 '지속가능한 발전'이 요구되기 시작한다. 이에 따라 사회적 책임을 이행할 수 있는 기업·산업, 즉 성장잠재력과 지속가능성이 모두 높은 기업·산업이 사회적으로도 정책적으로도 각광을 받는다. 대표적으로 정보통신·전기차·바이오 산업의 기업체가 그 가능성을 인정받았으며, 해당 산업·산업내 기업체는 각종 뉴딜 정책을 통해 전폭적인 지원을 받고 성장중이다.

이처럼 코로나 19를 계기로 산업 환경은 확실히 변화했다고 봐야하며, 산업이 변화한 만큼 기업의 생존을 결정짓는 요인도 코로나 19 이전과는 확연히 다를 수 있다. 따라서 코로나19 전후 중소법인의 폐업 여부에 대해 분석할 시 새롭게 개편된 산업 구조에서의 생존 요인을 추론할 수 있을 것으로 기대된다. 이 점에서 2020년 결산 재무제표를 기준으로 분석 기간인 2021년에서 2022년 6월까지의 폐업을 예측하여 코로나 19 유행 전후 기업의 생존 요인을 파악하고자 한다. 이때 중소법인의 신용평가에서 비재무 정보를 결합하는 것이 중요한 점에서 본 대회의 목적 상 여러 비재무 정보를 결합하여 중소법인의 폐업 여부를 예측하고자 한다. 궁극적으로, 이를 통해 코로나 19 이후 금융기관의 여신심사 정확도를 높이는데 기여할 수 있을 것으로 판단된다.

2 선행연구 분석

1) 선행 연구 1 : 금융기관 여신심사 시의 재무정보 활용

일반적으로 금융기관에서 기업을 평가할 시 재무비율을 활용하여 기업의 신용 및 상환가능성 등을 검토한다. 이때 대표적으로 활용하는 재무지표로 **활동성·안정성·수익성·성장성** 지표가 존재한다. (한국은행, 2007) 활동성은 기업이 조달한 자본 혹은 투하한 자산을 얼마나 효율적으로 운용하고 있는지를 나타내는 지표로, 재고자산회전율·총자본회전율 등이 있다. 다음으로, 안정성은 자기자본과 타인자본으로 구성된 자본이 기업의 자산에 얼마나 적절히

배분되고 있는지 측정하는 것으로, 부채비율・차입금의존도 등이 있다. 수익성은 이익창출능력이나 경영성과 등을 보는 것으로, 매출액순이익률・기업순이익률 등이 있다. 마지막으로 성장성은 일정 기간 동안 기업 활동 성과나 기업의 경영규모가 얼마나 변했는지를 측정하는 것으로, 순이익증가율・매출액 증가율 등이 있다. 이러한 재무분석을 통해 기업 재무구조의 안정성 및 가변성 여부를 확인 가능하다. 하지만 재무 분석만으로는 기업의 모든 것을 평가하기는 어려우며, 보유기술・경영진의 역할 등 비재무적 요소 역시 고려되어야 한다. (홍태호・신택수, 2009) 특히 정기적으로 외부감사를 받는 일정 규모 이상 기업의 재무자료는 어느 정도 신뢰할 수 있으나, 중소기업의 재무자료는 상대적으로 신뢰하기가 어려울 수 있다. 그러므로 중소기업은 재무 분석만으로 도산예측을 한다면 부정확한 결과가 초래될 수 있으며, 비재무요인도 함께 적절히 고려해야 할 것이다. (김용성・유왕진・이철규・이동명. 2012)

2) 선행 연구 2 : 강소기업 여부와 기업의 재무상태 간의 관계

강소기업의 정의를 참고한다면, **강소기업에 분류된 기업은 재무구조가 건전하다는 것을 알 수 있다.** 오한석・최경현 (2021)은 강소기업이란 제품 경쟁력이 있어 지속적으로 이익을 내는 성장을 하고, 시장 점유율이 높고, **재무구조가 견실한 중소기업**이라고 보았다. **이러한 강소기업은 우수한 재무성과를 낼 수도 있다.** 강소기업의 유형은 6개로 일자리친화・글로벌역량・기술력우수・지역선도기업・재무건전성・사회적가치로 나뉘는데, 기술력우수에 속하는 '이노비즈' 기업과 재무건전성에 속하는 '메인비즈' 기업은 재무성과가 우수하다는 것이 입증된 바 있다. (신상혁・김문겸, 2013) 이는 강소기업에 선정되어 금융기관 및 정부로부터 다양한 지원을 받았기에 가능했던 것으로 사료된다.

3) 선행 연구 3 : 업종과 부실기업간의 관계

최경연(2021)에 의하면, 상장기업은 기업 규모와 무관하게 **한계기업의 비중이 높은 업종에 포함될수록 한계기업이 될 확률이 높다.** 한편 한계기업은 정상적인 영업활동으로 이자상환조차 어려운 상태이므로, 정상기업보다 폐업・도산 등 부실의 위험이 높기도 하다. 이를 종합해서 본다면, **기업이 속한 업종이 한계기업의 비중이 높은 업종일수록 폐업의 위험이 크다고 볼 수 있다.** 박찬우(2022)에 의하면 '21년 기준 한계기업 수는 제조업(1,294개), 부동산 및 임대업(1,182개), 도소매업(367개) 순으로 많다. 제조업 분야가 한계기업 수가 가장 많은 것은 사실이나, 전체 기업수 대비 한계기업의 비중은 12.0%로 중간 수준이다. 도소매업도 전체 기업 수 대비 한계기업의 비중이 11.4%로 상대적으로 양호하다. 전체 기업 수 대비 한계기업 비중이 높은 산업은 숙박 및 음식점업(46.0%)이나 부동산 및 임대업(35.3%)이었다. **다시 말해 제조업・도소매업보다는 숙박 및 음식점업・부동산 및 임대업에 속한 기업일수록 폐업의 위험이 크다고 할 수 있을 것이다.**

4) 선행 연구 4 : 기업 입지와 기업 성장간의 관계

기업 입지는 기업의 성장 및 생존과 유의미한 연관이 있다는 선행연구가 존재한다. (Falck, 2007)는 신생기업이 외곽지역보다는 도심지역에 입지하였을 때 생존확률이 높아진다고 하였다. 한편 (Renski, 2011)는 국지화경제(입지상계수) 및 도시화경제(지역 규모 및 산업의 다양성) 등의 공간적 외부효과도 기업의 생존가능성을 높일 수 있다고 보았다. Wennberg and

Lindqvist(2010)도 스웨덴의 금융서비스, 가전제품, 전기통신, 의료장비, 정보기술, 제약 부문 등 5개 산업을 대상으로 분석한 결과, 집중적으로 산업 클러스터가 형성된 지역이 그렇지 않은 지역보다 기업의 생존률을 높이는데 도움이 된다는 것을 발견했다. 이때 클러스터의 절대적인 수준을 나타내는 상권의 집중도가, 상대적인 수준을 나타내는 입지상계수보다 영향력이 크다는 것 또한 파악하였다. 신혜원・김의준(2014)은 다양한 규모의 기업 중에서도 특히 중소기업의 생존확률이 이러한 접근성과 공간적 외부효과에 민감하게 영향을 받는다고 하였다. 그러므로 중소기업이 많은 지역에 1) 접근성을 개선하는 교통시설, 2) 정보교환을 돕는 네트워크 등이 존재한다면, 중소기업의 생존율이 높아질 것이라고 주장했다. 네트워크 요인으로서 지역별 액셀러레이터 수와 대표자 창업 네트워크를 변수로 사용하였다. 이외에도 기업의 경영 상태나 역량을 나타낼 수 있는 상장법인 여부, 국외법인 여부, 공동창업자 존재 여부, 사업화기반구축점수, 직원 수, 홈페이지 존재 여부를 비재무 정보로 활용하였다.

3 활용데이터

1) 활용 데이터

□ 나이스 제공 데이터 - 중소법인 휴폐업 이력 및 기업 정보

1. 기간 : 2018.01 ~ 2022.06
2. 데이터 목록 : 휴폐업_중소법인_외감.csv, 휴폐업_중소법인_휴폐업이력.csv, 액티브_중소법인_종합.csv, 액티브_중소법인_휴폐업이력.csv

□ 나이스 제공 데이터 - 재무정보

1. 기간 : 2018년 ~ 2021년 결산 재무제표
2. 데이터 목록 : 재무데이터.txt수

□ 외부 데이터 : 통계청 - 한국표준산업분류(10차)

1. 수집 목적 : 나이스 제공 데이터 내 IND_CD1 대분류 등 매칭
2. 수집 방법 : 홈페이지 내 다운로드 후 excel로 전처리
3. 데이터 목록 : 업종코드-표준산업분류 연계표.csv

□ 외부 데이터 : 고용노동부 - 강소기업 선정결과

1. 수집 목적 : 분석 기업 중 강소기업 선별 및 기업 주소 수집
2. 수집 방법 : 홈페이지 내 다운로드 후 excel로 전처리
3. 데이터 목록 : 2020_강소기업.csv, 2019_강소기업.csv, 2018_강소기업.csv, 2017_강소기업.csv

□ 외부 데이터 : 금융감독원 - 기업개황

1. 수집 목적 : 입지 분석을 위한 기업 주소 수집
2. 수집 방법 : 금융감독원 API를 통한 수집
3. 데이터 목록 : Dart_기업개황.csv

4. 비고 : Dart로 수집하지 못하는 폐업 기업의 경우 잡코리아, 원티드 등을 통해 수기로 수집

□ 외부 데이터 : 통계청 - 시군구 GRDP(지역내총생산)
1. 수집 목적 : 입지 분석을 위한 입지상계수 계산
2. 수집 방법 : KOSIS 내 시군구 경제활동별 지역내총생산 다운로드 및 결합
3. 데이터 목록 : GRDP.csv
4. 기간 : 2015년 기준 2019년 GRDP

□ 외부 데이터 : 소상공인시장진흥공단 - 상가(상권)정보
1. 수집 목적 : 입지 분석을 위한 상가 포화도 계산
2. 수집 방법 : 공공데이터포털 내 소상공인시장진흥공단_상가(상권)정보
3. 데이터 목록 : 상가(상권)정보 폴더 내 소상공인시장진흥공단_상가(상권)정보_광주_202012
4. 기간 : 2020년 12월 기준

□ 외부 데이터 : 창업진흥원 - 액셀러레이터 등록 현황
1. 수집 목적 : 입지 분석을 위한 창업 생태계 분석
2. 수집 방법 : 홈페이지 내 다운로드 후 python으로 전처리
3. 데이터 목록 : 액셀러레이터_전처리후.csv

□ 외부 데이터 : 알리콘 - 창업자 네트워크, 벡스인텔리전스 - 사업화보유역량
1. 수집 목적 : 창업자 역량 지표 수집
2. 수집 방법 : 디지털 산업혁신플랫폼 내 다운로드 후 python으로 전처리
3. 데이터 목록 : 기업네트워크.csv, 사업화기반구축수준등급.csv

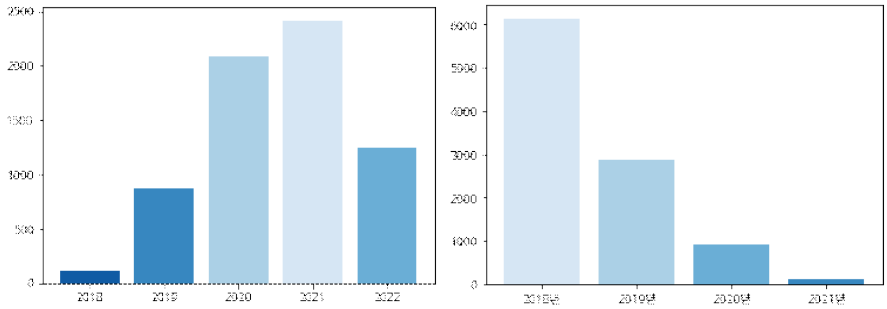
4 데이터 분석

1) 분석 목적 및 개념 정의

본 분석의 경우 코로나19 유행 전후에 따른 폐업 기업을 예측하고자 한다. <그림 1>에 나타난 연도별 폐업 기업 추이를 보면 2020~2021년에 폐업 기업 수가 가파르게 상승하는 점에서 코로나19에 따른 영향을 간접적으로 확인할 수 있다. 이때 코로나19 유행이 2020년 상반기에 시작된 점에서 2019년 결산 재무제표를 통해 기업의 폐업을 예측하는 것을 고려할 수 있다. 하지만 특정 한 시점의 경영 성과를 통해 예측하는 것은 기업의 가치를 과소·과대평가할 수 있는 점에서 여러 시점의 경영 성과를 통한 예측이 모델의 성능을 높이는 것이 필요하다. 따라서 <그림 2>에서 볼 수 있듯이 2020년 결산 재무제표의 수가 2019년보다 적음에도 여러 해의 걸친 경영 성과를 반영하기 위해 2020년을 기준으로 향후 폐업 여부를 예측하고자 한다.

이 점에서 **폐업 기업의 정의는 '2020년말 이후 폐업한 기업'** 이며 코로나19 전후 변화한 산업 환경에 적응하지 못한 낮은 경쟁력의 기업이라고 할 수 있다. 중도로 휴업하거나 폐업 취소된 기업은 폐업 기업으로 보지 않는다.

(그림1) 연도별 폐업 기업 수 추이 (그림2) 연도별 폐업 기업 재무제표 존재 수



2) 활용 모듈(Python) 및 구현 환경

- 데이터 전처리 및 시각화 : Pandas, Numpy, Matplotlib, Seaborn 등
- Factor analysis : FactorAnalyzer
- Modeling : sklearn, lightgbm, xgboost, imblearn, optuna
- Modeling 결과 분석 : shap
- 구현환경 : Google Colab Pro

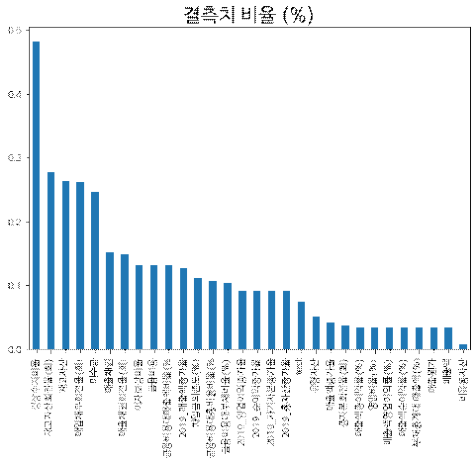
3) 데이터 전처리 - 결측치 및 이상치 제거

나이스 제공 데이터를 보면 액티브 기업의 경우 외감 기업만을 사용하며 폐업 기업은 외감 여부가 구분되지 않는다. 20년을 기준으로 존재하는 외감 대상 중소법인과 폐업 기업을 학습 데이터로 활용하여 분석을 진행한다. 이들 기업의 2019~2020년 재무정보를 분석하며, 금융업을 제외한 제조업 기업만을 분석에 활용하였다. 금융업 기업을 제외한 이유는 금융업이 영업활동 및 재무제표 계정의 특성이 타 업종과 상이함을 고려기 위함이다(강순심, 김정재, 2014, 14) 금융업은 'IND_CD1' 컬럼이 아닌 'BZ_TYP' 컬럼을 통해 판별하였다. ('M' 인 경우) 또한, 산업코드나 설립년월이 존재하지 않은 기업은 제외하였으며 본점인 기업만을 분석에 활용하였다. 이는 제공된 재무제표를 볼 경우 본점과 지점의 재무제표가 동일하며 대표자명이 동일한 등 각 지점 고유의 신용 상태를 나타내는 것으로 볼 수 없었기 때문이다.

동시에 중소기업 부실 예측 등에서 총자산이 지나치게 낮은 기업은 이상치로 간주하고 제거하였다. 이는 **중소기업의 재무자료의 신뢰성 문제를 감안하여 일정 규모의 기업만을 분석의 대상으로 해야 하기 때문이다.** (김상문, 2011) 또한, 회계기준에 기초하여 매출액이 존재하지 않거나 일치하지 않는 등 발생가능성이 없는 재무제표 역시 신뢰성이 부족한 점에서 제거했다. **이상치 제거 기준은 총자산이 1억원 미만 기업, 대차가 10만원 이상 차이나는 기**

업(단수 차이 고려) 매출액순이익률이 존재하지 않은 경우이다.

기업 재무제표 데이터에 대한 결측치는 우선적으로 다른 재무정보를 통해 유추할 수 있는 경우 해당 값으로 대체한다. 남은 결측치는 제거 시 생존한 데이터에 대해 편향 문제가 발생할 수 있으므로 중앙값(Median) 대체를 통하여 처리하였다. 또한 결측치 비중이 높은 변수의 경우 대체된 값에 따라 편향된 결과가 나타날 수 있으므로 해당 변수(경상수지비율, 재고자산 등)는 제외하고 선행연구 등으로 중요한 재무변수만을 선정하였다. 마지막으로 분석과정에 영향을 미칠 수 있는 이상치(Outlier)를 처리하기 위해 윈저라이징(Winsorizing)을 적용하여 이상치를 변수별로 분포상 누적확률 0.01, 0.99에 해당하는 값들로 변환하였다. 이외에 비재무 지표에 대한 결측치는 모두 최빈값으로 대체하였다.

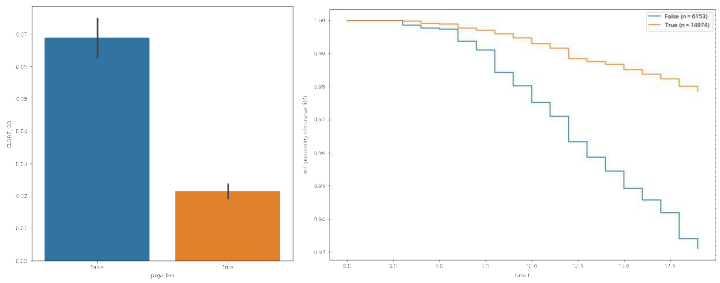


4) EDA

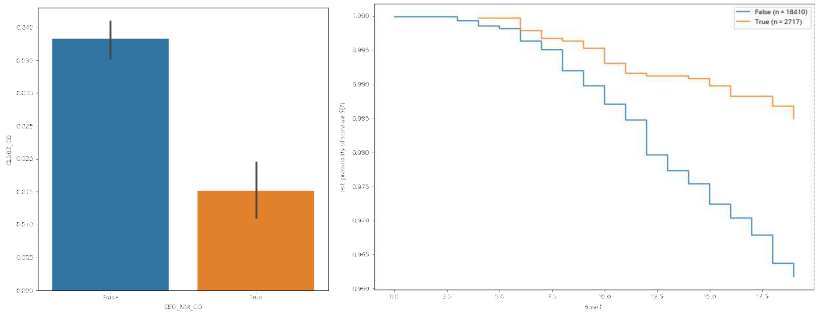
□ 비재무 지표별 시각화 및 생존분석

선행연구를 통해 기업의 재무제표를 통해 구할 수 있는 기업의 활동성, 안정성, 수익성, 성장성 지표는 기업의 부실예측 및 신용등급 평가 등에 효과적인 것으로 나타난 반면, 비재무 지표에 대한 선행연구는 부족하다. 따라서 비재무 지표에 따른 폐업 기업 비중을 시각화하고 생존분석을 통해 해당 비재무 지표의 타당성을 확인하고 모델링에 활용하였다.

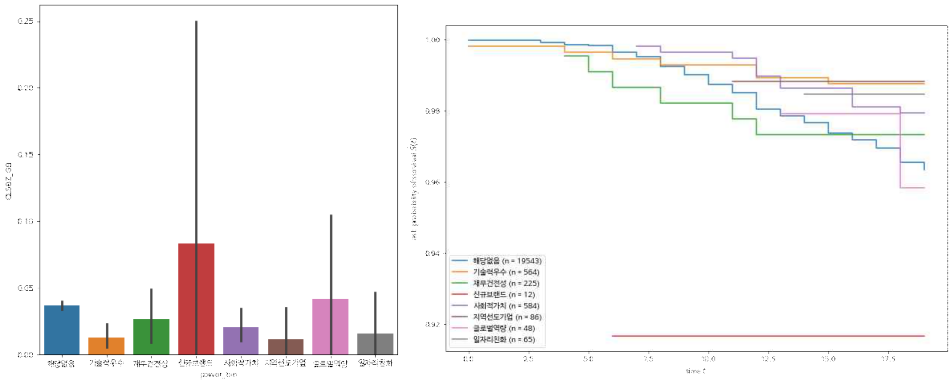
<예시 - 웹페이지 존재 여부에 따른 기업의 생존확률 추정>



<예시2 - 공동 창업자 존재 여부에 따른 기업의 생존확률 추정>



<예시3 - 강소기업 분류에 따른 기업의 생존확률 추정>



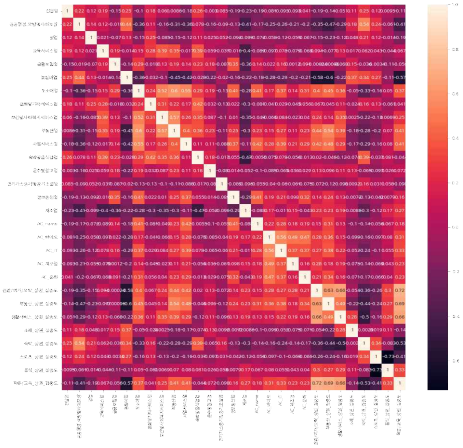
□ **요인 분석을 통한 요인 생성**

기업의 경영환경 관련 변수는 기업이 위치한 '시도' 변수뿐만 아니라 각 업종별 입지상계수와 액셀러레이터 수와 주요 산업별 액셀러레이터 비중, 업종별 상권 집중도를 수집하였다. 각 변수 간에는 상관관계가 다수 존재하는 점에서 요인 분석을 통해 변수 간의 상호관계가 함축된 변수를 추출하여 분석에 활용하고자 한다. 이를 통해 지역별 창업 입지 특성 해석을 용이하게 하고자 하며 변수 수를 줄여 모델의 성능을 높일 것으로 기대된다.

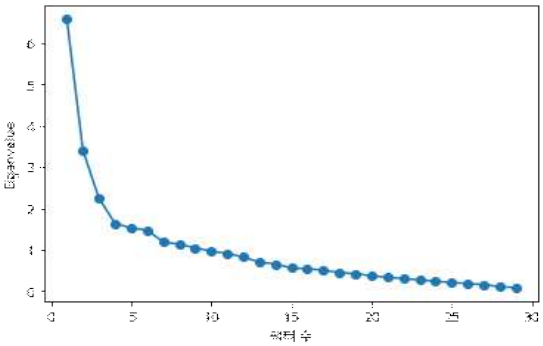
구형성 검정은 Bartlett 'test와 Kaiser-Meyer-Olkin(KMO) 검정을 통해 이루어졌으며 Bartlett 'test 결과, p-value가 귀무가설을 기각할 수준으로 나타났으며 KMO 측도는 0.72로 비교적 우수한 편으로 요인 분석을 하기에 적합하다고 판단된다. 스크리 도표를 통해 요인 수를 결정했으며 1 이상이며 꺾이는 지점인 5~6개 중 6개로 결정하였다. 6개로 결정한 이유는 Tableau 등을 통한 시각화 시 레이더 차트를 통해 지역의 특성을 한눈에 확인할 수 있다

는 실용적인 측면에서 결정하였다. 일반적으로 활용되는 varimax 방식을 통해 6개의 요인을 추출하고 각 변수에 대한 ±0.5인 높은 요인적재량을 통해 요인을 정의하였다.

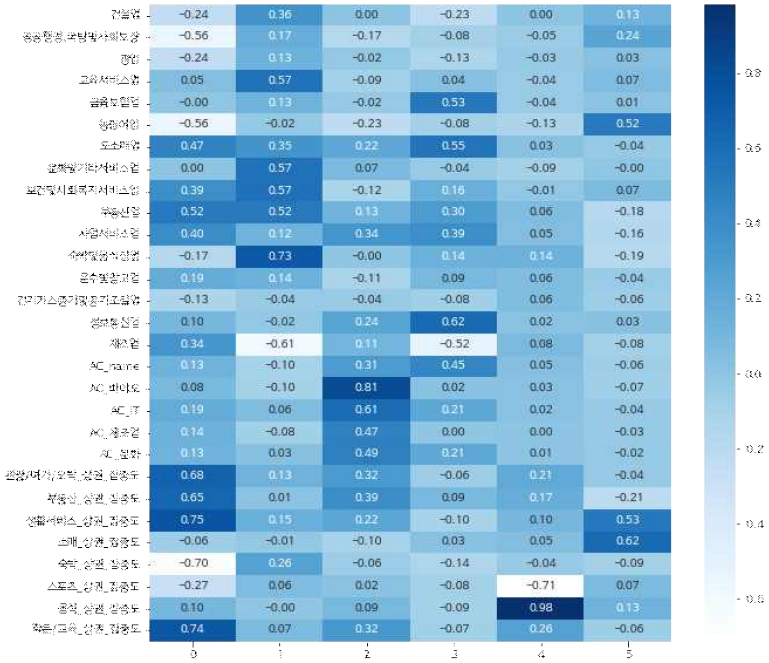
< 경영환경 변수 간 상관관계 분석 >



< 요인 수 결정 - 스크리 도표 >



< Factor loading matrix >



< 표. 요인 정의 >

요인	요인 정의	정의 근거
Factor_0	학교 근접 수준	관광·여가·오락, 생활서비스, 교육업의 상권 집중도 계수는 높으나 숙박업의 상권 집중도 계수는 (-)임. 이는 학교 근처 상권의 모습과 유사하다는 점에서 학교 근접 수준를 나타내는 변수로 정의함.
Factor_1	도심 근접 수준	숙박 및 음식점업, 교육서비스업의 계수가 높음. 수도권과 같은 도심 지역일수록 숙박 및 교육 서비스업이 발달했다는 점에서 도심 근접 수준을 나타내는 변수로 정의함.
Factor_2	AC 밀집 수준	바이오와 IT 산업의 AC 계수가 높다는 점에서 AC 밀집 수준을 나타내는 변수로 정의함
Factor_3	IT산업 발달 수준	정보통신업 계수가 높은 점에서 IT산업 발달 수준을 나타내는 변수로 정의함
Factor_4	서울 근접 수준	음식점 상권 집중도 계수가 높음. 음식점 상권 집중도가 높은 지역이 서울 마포구·종로구라는 점에서 해당 변수는 서울이거나 서울에 근접한 수준을 나타내는 변수로 정의함
Factor_5	농촌 근접 수준	농림어업 계수가 높고 소매, 생활서비스업의 상권 집중도 계수가 높다는 점에서 농촌이거나 농촌에 근접한 수준을 나타내는 변수로 정의함

5) 모델링

모델링에 활용된 변수는 데이터별 분포를 고려하여 선정하였다. 지나치게 데이터 분포가 편향된 변수는 모델에 활용될 경우 과적합이 나타날 수 있거나 의미 없을 우려가 존재한다. 따라서 PSN_CORP_GB(개인법인구분), CMP_SCL(기업 규모) 등을 제외하였다. 또한, 일부 변수에서 범주별 기업의 수가 적은 경우에도 과적합이 나타날 수 있는 점에서 산업코드가 '보건업 및 사회복지 서비스업'과 '공공 행정, 국방 및 사회보장 행정'인 경우 제외하였다. 범주형 변수의 경우 Shap value를 구하기 위해 더미변수로 변환하였다.

데이터 전처리 과정을 통해 최종적으로 분석에 활용되는 기업은 액티브 기업 20,382개와 폐업기업 745개이며 활용되는 변수는 다음과 같다.

< 표. 선정된 변수 >

분류	column	feature 정의	type
비재무 변수	NATN_NM_Bin	국외법인 여부	bool
	LIST_BIN	상장법인 여부	bool
	page_bin	홈페이지 존재 여부	bool
	power_bin	강소기업 분류	bool
	CEO_NM_CO	공동 창업자 존재 여부	bool
	대분류	산업코드 - 대분류	category → dummy
	ESTB_time	기업연령(월)	float64
	EMP_CNT	직원 수	float64
	network_count	창업자 네트워크 수	float64
	사업화기반구축수준점수	사업화기반구축수준점수	float64
경영환경 변수	adres	시도	category → dummy

분류	column	feature 정의	type
	Factor_0	학교 근접 수준	float64
	Factor_1	도심 근접 수준	float64
	Factor_2	AC 밀집 수준	float64
	Factor_3	IT산업 발달 수준	float64
	Factor_4	서울 근접 수준	float64
	Factor_5	농촌 근접 수준	float64
재무 변수	자산총계	기업 규모	float64
	재고자산회전율(회)	활동성 지표	float64
	매출채권회전율(회)	활동성 지표	float64
	매입채무회전율(회)	활동성 지표	float64
	총자본회전율(회)	활동성 지표	float64
	유동비율	안정성 지표	float64
	부채비율(%)	안정성 지표	float64
	차입금의존도(%)	안정성 지표	float64
	이자보상비율	안정성 지표	float64
	기업순이익률(%)	수익성 지표	float64
	매출액순이익률(%)	수익성 지표	float64
	매출액총이익률(%)	수익성 지표	float64
	총자산증가율	성장성 지표	float64
	영업이익증가율	성장성 지표	float64
	순이익증가율	성장성 지표	float64
	자기자본증가율	성장성 지표	float64
	무형자산_비율	무형자산 비율	float64

선정된 변수를 바탕으로 모형을 학습한 후 예측 결과에 대한 평가 척도는 F1-Score를 활용하고자 한다. 종속변수가 폐업 여부인 점에서 이진 분류 문제에 해당하며 폐업 기업의 경우 전체 데이터의 약 3.54% 밖에 되지 않는 점에서 데이터 불균형 문제가 있는 상황이다. 따라서 단순히 예측 값의 정답률을 나타내는 Accuracy는 적절한 평가 척도로 활용될 수 없다고 본다. 이 점에서 재현율과 정밀도의 조화평균인 F1-Score를 중점으로 모델의 성과를 판단하고자 한다. 다만 폐업 예측을 통해 기업 신용평가 등 금융권에 활용할 수 있는 점을 고려할 때 실제 폐업 기업을 폐업 기업으로 분류한 비율인 재현율(Recall)을 고려할 필요가 있다. 부실기업을 정상 기업으로 잘못 예측한다면 부실기업에 대한 대출 승인으로 디폴트 시 막대한 손실이 발생할 수 있기 때문이다. 따라서 중소기업의 부실화를 사전에 예측하여 폐업이 예상되는 기업에 보수적인 여신심사를 통해 성공적으로 리스크를 관리하고자 한다.

예측 모형은 F1-score를 최대화하며, 재현율을 동시에 고려하여 선택하였으며, 전통적인 회귀 방법인 로지스틱 회귀분석 등 과거의 예측 모형보다 높은 예측률을 보이는 랜덤포레스트, XGBoost, LGBM(Light Gredient Boosting Machine)의 결과를 비교하고 최종적으로 모델을 선택한다.

이때 폐업 기업의 수가 적은 데이터 불균형 문제를 해결하기 위하여 오버샘플링 방법을 사용하고자 하며, Chawla et al.(2002)이 제시한 오버샘플링 기법인 SMOTE를 사용하였다. 또한, 모델의 성능을 높이기 위해서 최적의 하이퍼파라미터(Hyper parameter)를 찾아내는 것이 중요하다. (김소정・이군희, 2021) 이점에서 모델의 하이퍼파라미터 튜닝을 위해 5-fold 교차검증(5-fold Cross Vaildation) 방식으로 데이터를 분할하고, 각 Training Fold에 SMOTE를 적용하여 데이터 밸런싱을 수행하고 Optuna를 통해 하이퍼파라미터 튜닝을 진행하였다. Optuna는 베이지안 최적화 기반의 하이퍼파라미터 최적화 프레임워크로 사용자가 하이퍼파라미터를 일정 범위로 정하고 이 중 최적화를 위해 설정된 측정 지표에 따라 하이퍼파라미터를 지속해서 수정해나가는 방식이다. 하이퍼파라미터 조정 전 각 모델의 성능을 비교한 후 높은 성능을 보이는 모델에 대해 하이퍼파라미터 조정 후 최종 모델을 선택하였다

6) 모델링 결과

< 모델링 결과 요약 >

SMOTE 이전					
	Accuracy	Precision	Recall	F1-score	rou-auc
Random Forest	97.9%	95.1%	43.3%	0.59	0.93
LGBM	98.1%	87.4%	55.8%	0.68	0.92
XGBoost	98.0%	89.%	49.6%	0.64	0.93
SMOTE 이후					
Random Forest	97.8%	70.9%	64.3%	0.67	0.93
LGBM	97.6%	66.4%	66.1%	0.66	0.90
XGBoost	95.2%	39.8%	71.9%	0.51	0.89
하이퍼파라미터 조정 후					
LGBM	98.4%	92.4%	60.3%	0.73	0.91
XGBoost	98.3%	86.4%	59.4%	0.70	0.92

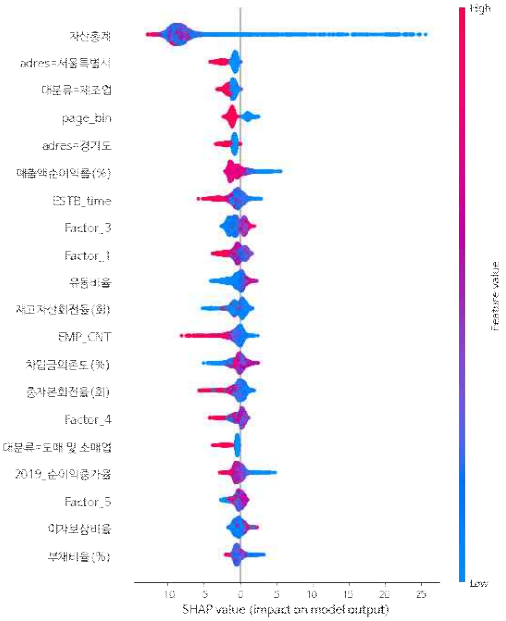
SMOTE 전후를 비교했을 때 모든 모델에서 SMOTE를 통한 오버샘플링 이후 재현율이 전부 상승하였지만 정밀도는 하락한 것으로 나타났다. 특히 모델에서 재현율이 가장 좋은 것은 XGBoost 모형으로 폐업 기업을 다른 모형 대비 정확히 예측하는 것으로 나타났다. 하지만 XGBoost의 F1-score가 SMOTE 이후 낮아졌으며 Precision이 지나치게 낮다는 문제가 존재한다. 특히 재현율은 다소 낮을지라도 F1-score는 XGBoost보다 LGBM이 더 높은 성능보였으며, SMOTE 이전 재현율은 LGBM이 더 높았다. 따라서 하이퍼파라미터 조정 이후 두 모델의 성능을 비교하였다. 하이퍼파라미터 조정 이후 재현율은 다소 낮아졌을지라도 Precision이 개선된 한편, F1-score가 상승하면서 높은 성능을 보이고 있다. 낮아진 재현율의 경우 5-fold 교차검증을 통한 평가로 과적합 문제가 완화된 것으로 생각되며 하이퍼파라미터 조정 이전보다 강건한 모델이라고 판단된다. 이에 하이퍼파라마터가 조정된 모델링을 기준으로 LGBM과

XGBoost 중 F1-score와 재현율이 모두 높은 LGBM을 최종 모델로 선택하였다.

7) 모델 해석

한편, AI 기반 모델링이 복잡하고 다양한 변수가 모델에 활용되면서 모델의 평가 기준을 블랙박스와 같이 알 수 없는 문제가 존재한다. 특히 AI 관련 윤리 이슈가 강해지면서 발표한 금융위원회의 「금융분야 인공지능(AI) 가이드라인」에 따르면 신용평가 및 언더라이팅 등 금융거래 및 계약체결 여부 결정 시 고객의 설명요구 및 정정요구권이 있음을 고지하고 AI 평가결과 및 주요 평가기준 등을 고객에게 설명할 수 있어야한다. 따라서 모델의 설명력을 확보하는 동시에 변수의 영향도를 파악하고 중요한 변수를 선별하고자 한다. 동시에 중소기업의 재무 및 비재무 정보 수집에 따른 비용을 생각할 때 신용평가에 유의미한 변수만을 수집하는 것이 신용평가사에 있어 비용 효율성을 높일 것으로 기대된다. 선행연구 등을 참고하여 XAI를 위한 방법론으로 샤플리 값(Shapley values)을 이용한 Adadi, Berrada(2018)의 모형인 SHAP을 사용하였다. (김소정・이군희, 2022) 최종적으로 선택한 모델의 shap value은 다음과 같이 시각화할 수 있다.

< Summary Plot >



먼저 변수별 중요도를 보면 재무정보의 경우 자산총계(기업규모), 매출액순이익률, 유동비율, 재고자산회전율, 차입금의존도, 총자본회전율, 2019년 순이익증가율, 이자보상비율, 부채비율 순으로 높은 영향을 미친 변수로 파악된다. 기업 규모는 중요도 수준이 가장 높음에도 shap value가 일관성이 다소 부족한 것을 볼 수 있다. 다만 우측 꼬리가 긴 것을 통해 기업규모가 클수록 폐업률이 낮을 수 있을 것으로 추정된다. 수익성 지표인 매출액순이익률과 성장성 지표인 순이익증가율은 높을수록 기업의 생존 여부에 (+)의 영향을 주는 것으로 확인되었다.

한편, 안정성 비율인 유동비율과 부채비율을 볼 경우, 유동비율은 낮을수록 폐업에 (+)의 영향을 주며, 부채비율은 높을수록 폐업에 (-)의 영향을 주고 있다. 높은 유동비율과 낮은 부채

비율이 기업의 안정성을 높여 생존에 도움이 된다는 통념과 반대되는 상황이다. 이는 오히려 차입을 통한 투자가 적거나 유동자산을 투자에 활용하지 않으면서 기업의 성장성을 제고하는데 실패했기 때문으로 추정된다. 활동성지표인 총자본회전율은 일반적인 통념과 같이 낮아질수록 폐업률이 높아진 것으로 나타났다. 한편, 통념과 반대로 재고자산회전율의 경우 낮을수록 폐업률이 낮아지는 것으로 나타났는데 이는 재고자산회전율 결측치가 대부분 폐업 기업에 몰려있어 중간값으로 대체하면서 나타난 문제 때문에 노이즈로 작용했을 가능성도 존재한다고 본다.

한편, 비재무 정보를 보면 수도권 여부나 업종이 영향을 높은 영향을 준 것으로 보이는 한편, 6개의 Factor 중 4개가 높은 영향을 주고 있는 것으로 보인다. 수도권이 아닐수록 폐업률이 높아진 것으로 나타났으며 도소매업·제조업이 아닌 경우 폐업률이 높은 것으로 나타났다. Factor를 보면 Factor 3이 낮을수록 폐업률이 낮아지는 한편, 올라갈수록 폐업률이 높아진다. 정보통신업 입지상계수가 높은 지역, 즉 IT 산업이 발달한 지역에 위치한 기업은 폐업할 확률이 높을 수 있다는 뜻이다. 이는 코로나 19 이후 기술기업의 창·폐업이 활발했던 것에 기인한 듯 하다. 실제로 '20년 기준 30대 청년층의 IT 등 기술창업은 3.8% 증가하여 역대 최고치를 보였으나(중소벤처기업부, 2021), 신생기업 중 절반 정도는 1년도 넘기지 못한 채 폐업하였다.(통계청, 2020)

Factor 1이 낮을수록 폐업률이 올라가는 것을 볼 수 있는데, 이는 숙박/음식점업 및 교육서비스업 입지상계수가 낮을수록 폐업률이 올라가는 것을 의미한다. 숙박 및 교육 서비스업이 발달한 지역이 수도권 등 도시의 특성을 보이는 점을 고려할 때 도시 지역에서 창업하는 것이 생존률을 높일 수 있을 것으로 보인다. 동시에 Factor 4를 보면 음식점 상권 집중도가 낮을수록 폐업률이 올라가고 있다. 음식점 상권 집중도가 높은 지역이 서울 마포구·종로구임을 상기할 때, 해당 요인 역시 도심에서의 창업이 기업의 생존에 유리함을 나타낸다고 볼 수 있다. 즉 Factor 1과 Factor 4를 종합적으로 고려한다면, 도심 지역에 위치한 기업이 그렇지 않은 기업보다 생존률이 높다는 것을 알 수 있다.

5 참고문헌

양재룡 외 6인. (2007). 기업경영분석해설. 한국은행

홍태호, 신택수. (2007). 부도확률맵과 AHP를 이용한 기업 신용등급 산출모형의 개발. 정보시스템연구, 16(3), 1-20.

김용성, 유왕진, 이철규, 이동명. (2012). 기업사례를 통한 중소벤처기업의 도산예측을 위한 비재무적 요인에 관한 연구. 한국경영공학회지, 17(1), 245-258.

오한석, 최경현. (2020). 기술혁신 강소기업 지원사업의 경제적 성과 분석: World Class 300 프로젝트를 중심으로. 벤처창업연구, 15(4), 121-133.

신상혁, 김문겸. (2013). 혁신형중소기업인증이 재무성과에 미치는 영향: 이노비즈와 메인비즈를 중심으로. 경영컨설팅연구, 13(3), 193-217.

최현경. (2021). 한계기업 현황과 지원기준. 월간 KIET 산업경제, 276(), 31-41.

박찬우. (2022). 한계기업 현황과 시사점. KDB산업은행. KDB미래전략연구소

Falck, O. 2007. Survival chances of new businesses: Do regional conditions matter?. Applied Economics 39, issue 16: 2039-2048.

Renski, H. 2011. External economies of localization, urbanization and industrial diversity and new firm survival. Papers in Regional Science 90, no.3:473-502

Wennberg and Lindqvist, G. 2010. The effect of clusters on the survival and performance of new firms. Small Business Economics 34, issue 4:221-241.

신혜원, 김의준. (2014). 기업 입지유형 및 규모가 신생기업의 생존에 미치는 영향. 국토연구, (), 17-30.

강순심, 김정재. (2014). IFRS도입이후 분·반기 보고기준에 따른 연결 및 별도재무제표의 가치관련성 분석－유가증권시장과 코스닥시장의 비교－. 경영연구, 29(1), 223-261.

김상문. (2011). 생존분석을 이용한 중소기업 부실예측과 생존시간 추정. 신용보증기금

김소정, 이군희. (2022). 심층신경망의 설명가능성과 하이퍼파라미터 특성에 관한 연구 -중소기업 신용평가를 중심으로-. 중소기업금융연구, 42(1), 3-37.

A. Adadi and M. Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI),“ in IEEE Access, vol. 6, pp. 52138-52160, 2018, doi: 10.1109/ACCESS.2018.2870052.