기업 부도 예측분석 보고서

: 생존분석과 머신러닝 다양한 분석방법론 비교를 중심으로

2023021401 김혜연

1. 분석 개요

a. 주제 소개

본 분석에서는 다양한 분석방법론을 통해 기업 부도 예측 모형을 연구하고자 한다. 부도 예측을 위한 기업별 데이터를 구축하고 SMOTE, 생존분석 등 다양한 기법을 적용 후 비교 분석하여 최적의 모델을 선정한다. 또한 부도 예측에 유의미한 변수를 파악하여 해석에 기여하고자 한다.

b. 분석 배경 및 목적

법원통계월보에 따르면 올해 10월까지 전국 법원에 접수된 법인 파산신청 건수는 총 1,363건으로 집계되며 지난해 연간 건수(1,004)를 넘어섰다. 작년부터 이어진 고금리·고환율·고물가 등 '3고(高) 현상' 영향으로 기업 파산이 증가한 것으로 보인다. 기업의 부실화는 기업뿐만 아니라 국민경제에 막대한 손실을 초래한다. 따라서 기업 부도 예측 모형을 구축하여 기업들이 선제 대응을 하고 정부가 사전에 지원 정책을 수립할 수 있도록 한다.

본 분석의 목적은 높은 정확도와 해석력 둘을 모두 갖춘 예측 모델을 구축하는 것이다. 정확한 부도예측을 위해 전통적인 모델과 ml 기법을 비교 분석하며 실험을 수행하였고, 현재 기업 부도에 거시경제 요인의 영향이 큰 것을 고려하여 재무 데이터뿐만 아니라 거시경제요인과 뉴스심리지수 등 다양한 변수를 활용하였다. 이를 통해 이를 통해 부도에 유의한 변수를 파악하고자 한다.

c. 분석 배경 및 목적

본 분석에서는 부도를 상장폐지로 정의하였다. 상장폐지 사건은 부도와 반드시 연결된다고 볼 수는 없으나 거래 정지 및 주가 하락이 발생하면 투자자와 채권자가 큰

손실이 발생하므로, 상장폐지를 부도로 인식하는 것은 더욱 보수적인 기준에서 부도를 적절하게 평가하는 방법이라고 할 수 있다.

이때 '파산선고', '자본 잠식' 등 실적 부진과 관련 있는 사유의 상장폐지 기업을 부도 기업으로, '신규/이전상장', '피흡수합병' 등 관련 없는 사유의 상장폐지 기업을 정상기업이라고 정의하였다. 본 보고서에서는 이하 부도기업과 정상기업으로 칭한다.

2. 데이터 수집 및 전처리

a. 기업 표본

코스피 시장에서의 상장폐지보다 코스닥 시장에서의 상장폐지가 더 많이 발생한다는 점을 고려하여 코스닥 시장의 기업을 대상으로 분석을 진행하였다. 2022년의 데이터를 주로 활용하기 위하여 폐지일 기준 2023년까지의 상장폐지 기업과 상장일 기준 2021년까지의 상장기업을 수집한 후 조건에 맞게 전처리하였다. 그 결과 정상기업의 표본 수가 부도기업의 약 10배였고 데이터 불균형 문제를 확인하였다

b. 데이터 수집

데이터 수집 시점은 당시 기업과 경제 상황을 설명할 수 있는 데이터들을 타임스탬프가 맞도록 설정하였다. 부도 기업에 대해서는 부도 기준 수집 가능한 최신 1년의 기업 재무 데이터와 부도 전년도 1년의 거시경제 데이터를, 정상기업에 대해서는 2022년 1년 간의 데이터를 수집하였다.

기업 정보는 KIND에서 '기업명', '종목코드', '폐지일', '상장일', '지속기간', '수집기준연도'를, 재무 데이터는 OpenDartReader API를 활용하여 '자산총계', '부채총계', '자본총계', '매출액', '영업이익', '당기순이익', '유동부채', '유동자산', '비유동자산', '비유동부채'를 수집하였다. 거시경제 데이터는 kosis와 ecos를 통해 'CD91일', '콜금리', '국고채3년', 'GDP성장률', '원/달러 환율', '실업률', '코스피 종가', '코스닥 종가', '전산업생산지수', '경상수지', '경제심리지수', '뉴스심리지수', '소비자물가지수증감', '생산자물가지수증감', '주택가격지수'를 수집하였다.

기업 재무 데이터 수집 시 기업마다 공시하는 방식이 상이하여, 이후 재무비율 계산을 위해 결측치가 많지 않은 변수만 수집하였다. 또한 전자공시의 재무데이터는 2015년 이후의 데이터만을 제공하여 폐지일 기준 2016~2023년까지의 87개의 기업에 대해서만 분석을 진행하였다. 수집한 변수 중 결측치가 일부 존재하는 경우에는 KNN Imputer로 결측치를 보간하였다.

c. 데이터 전처리

기업 재무 데이터를 활용하여 건전성, 수익성, 성장성, 유동성, 활동성, 규모를 설명하는 19개의 재무비율을 산출했다. 두번째로 거시경제 데이터는 연도별 평균값을 기업별 데이터 기준연도에 맞게 조인하였다. 이때 기준연도는 정상기업은 2022년, 부도기업은 부도 전년도이다.

총 41개의 컬럼을 수집하였고, 경제 도메인인 만큼 컬럼 간의 상관성이 매우 높아 상관계수 0.95를 임계치로 하여 변수 제거를 우선적으로 진행하였고 그 결과 30개의 컬럼을 활용하였다.

3. 생존분석

a. 생존분석 개요

생존분석을 기업 부도 예측에 적용할 때, 생존기간은 기업의 지속기간인 수명, 사망은 기업의 부도가 된다. 생존분석은 다른 공변량을 활용하는지와 생존시간 분포를 가정하는지에 따라 크게 세 가지 방법으로 나뉘어지는데, 본 분석에서는 준모수적 방법과 모수적 방법을 활용하여 생존시간과 유의한 공변량을 파악하였다.

b. 준모수적 방법

콕스 비례위험 모형의 주요 관심사는 어떤 변수들이 생존시간에 어떤 영향을 미치는가를 규명하는 것이다. 즉 유의한 공변량을 파악하는 데에 쓰이는데, 해당 모형을 적합해보았을 때 총자산규모를 포함한 7개의 변수를 선택한 모델의 C-index가 가장 높았다.

그러나 콕스 비례위험 가정을 검정해보았을 때 만족하지 못하여 다음 모수적 방법으로 진행해 보았다.

c. 모수적 방법

AFT모델은 생존함수가 모수적 연속 분포를 따른다고 가정하며 다른 공변량을 활용할 수 있다. 모형 적합 결과 로그로지스틱으로 가정했을 때의 AFT가 가장 높은 성능을 보였다. 해당 모델에서 변수 선택 후 성능은 더 높아졌고 경제심리지수, 뉴스심리지수, 당기순이익증가율, 매출액증가율, 총자산규모, 총자산영업이익율 총 6개의 변수가 유의하다는 결과가 도출되었다.

지금까지 진행한 생존분석 결과는 데이터가 매우 불균형한 상태였기 때문에 높은 정확도로 편향되었을 수 있다고 판단하였고 다른 머신러닝 방법들을 시도하였다.

4. 기업 부도 예측 모델링

a. 모델링 전 전처리

부도 여부를 예측하는 모델에 쓰이는 공변량은 총 24개이다. 모델링 전 training set과 test set을 분리하였고, 학습 시 training set에서 교차 검증을 수행하였다. 모델에 쓰이는 공변량들은 모두 연속형이므로 Standard Scaling을 진행하였다.

다음 데이터 불균형 문제를 해결하기 위해 오버샘플링을 수행하였다. 불균형 데이터의 경우 다수 클래스만을 잘 예측하는 편향된 모델이 구축될 수 있으므로 SMOTE를 활용하여 training set의 클래스 비율을 같게 하였다. 이때 활용한 SMOTE는 새로운 샘플을 생성 시 소수 클래스의 데이터들을 기준으로 최근접 이웃 k개를 합성하는 알고리즘이다.

b. 모델링 전 전처리

전처리 후, Decision Tree, Random Forest, LightGBM, XGBoost, MLP, SVM, Naïve Bayes 총 7개의 모델로 학습을 진행하였다. Grid Search CV를 진행하여 최적의 하이퍼파라미터를 선정하였다. 그 결과 Random Forest가 정확도 0.9967, F1-score 0.975로 가장 성능이 좋았고 최적 모델로 선정하였다.

c. 변수 중요도 해석

SHAP는 머신러닝 모델에서 각 변수를 하나씩 빼고 더하면서 해당 변수가 y 예측의 정확도에 얼마나 기여하는지 계산하여 각 변수의 중요도를 수치화하는 방법이다. 최적 모델에서는 전산업생산지수, 원달러환율, 뉴스심리지수 등이 변수 중요도가 높았다. Summary Plot에서 변수 중요도를 해석해 본 결과 전산업생산지수가 낮으면 부도 경향이 있었다.

5. 결론

a. 분석 결과

기업 부도 예측 분석 결과 로그로지스틱 AFT 모델과 랜덤포레스트의 각 최적 모델에서 경제심리지수, 뉴스심리지수, 매출액증가율, 총자산규모, 총자산영업이익율 5개가 공통적으로 유의한 변수로 꼽혔다. 기업 부도 예측에 있어서 통계적 모델과 머신러닝 모델의 중요한 변수가 유사함을 확인할 수 있었다.

b. 주체별 기대효과

기업 부도 예측 모델 구축을 통해 기업은 기업 부도에 유의한 변수들에 대해 기업 재무 상황을 관리하고, 거시경제 상황을 고려하며 선제적으로 실적 부진에 대비하고, 정부는 거시경제 상황에 대비하여 사전에 정책을 수립하고, 금융기관은 기업 신용 및 실적 평가 시 데이터 기반의 기준을 수립 후 효율적인 평가를 할 수 있다.

c. 의의

본 분석의 의의는 다양한 변수들과 기법들을 활용하여 정확한 기업 부도 예측을 하고자 했다는 점이다. 생존분석을 통해 변수 유의도를, 머신러닝 알고리즘을 통해 변수 중요도를 확인하며 기업 부도 예측에 유의한 변수를 파악하고 해석력에 기여하였다. 추후에 뉴스심리지수를 활용하는 대신 직접 수집한 종목별 뉴스 텍스트 데이터로 감성분석을 수행해 보고, 다양한 오버·언더샘플링과 변수 선택을 실험해 보며 연구를 발전시키고자 한다.