Journal of Economic Theory and Econometrics, Vol. 36, No. 2, Jun. 2025, 59–82

Bankruptcy Prediction for Listed Companies in Korea Using Machine Learning and Oversampling Methods *

Si Hyun Noh † Heejoon Han ‡

Abstract This study compares the performance of statistical models and various machine learning methods in predicting corporate bankruptcy using financial and macroeconomic data from publicly listed companies. It also analyzes the effectiveness of oversampling methods in addressing the class imbalance problem. The empirical analysis employs logistic regression, random forest, XGBoost (extreme gradient boosting), and deep neural networks, along with oversampling techniques such as SMOTE (synthetic minority over-sampling technique) and ADASYN (adaptive synthetic sampling). The results show that XGBoost delivers the most accurate and balanced predictive performance across both the original dataset and the oversampled datasets. In contrast, logistic regression exhibits high recall but limited practical applicability due to its low precision. These findings suggest that combining oversampling techniques with machine learning models such as XGBoost provides a more effective and practical approach to bankruptcy prediction in the context of imbalanced data.

Keywords Corporate bankruptcy, bankruptcy prediction, machine learning, imbalanced data, SMOTE, ADASYN.

JEL Classification C01, C83, G33.

*This paper is a revised and extended version of the first author's master's thesis submitted to the Department of Quantitative Applied Economics at Sungkyunkwan University in 2024.

†Valuation Policy Team, NICE P&I, 11 Gukhoe-daero 68-gil, Yeongdeungpo-gu, Seoul, Republic of Korea 07237. E-mail: vs6609@naver.com.

‡Corresponding Author. Department of Economics and Department of Quantitative Applied Economics, Sungkyunkwan University, 25-2 Sungkyunkwan-ro, Jongno-gu, Republic of Korea 03063. E-mail: heejoonhan@skku.edu.

Received June 4, 2025, Revised June 15, 2025, Accepted June 16, 2025

Journal of Economic Theory and Econometrics, Vol. 36, No. 2, Jun. 2025, 59–82

머신러닝과 오버샘플링(oversampling)을 이용한 상장기업 부도예측 연구 *

노시현 †　　한희준‡

Abstract　본 논문은 상장기업의 재무 및 거시경제 데이터를 활용하여 기업부도를 예측하는 통계적 모형과 다양한 머신러닝 기법의 성능을 비교하고, 불균형 데이터 문제를 완화하기 위한 오버샘플링 기법의 효과를 분석하였다. 실증 분석에는 로지스틱 회귀, 랜덤 포레스트, XGBoost(extreme gradient boosting), 심층신경망 모형을 적용하였으며, 오버샘플링 기법인 SMOTE(synthetic minority over-sampling technique) 및 ADASYN(adapti-ve synthetic sampling)을 사용하였다. 분석 결과, XGBoost는 원자료뿐 아니라 오버샘플링을 적용한 경우 모두에서 가장 우수하고 균형 있는 예측 성능을 보였다. 반면, 로지스틱 회귀는 높은 재현율을 나타냈으나, 낮은 정밀도로 인해 실무적 활용에는 한계가 있었다. 이러한 결과는 불균형 데이터 환경에서 오버샘플링 기법과 XGBoost와 같은 머신러닝 모형을 결합하여 사용하는 것이 기업부도 예측에 있어 보다 효과적이고 실용적인 접근법이 될 수 있음을 시사한다.

Keywords　기업부도, 부도예측, 머신러닝, 불균형 데이터, SMOTE, ADASYN.

JEL Classification　C01, C83, G33.

---

*이 논문은 노시현의 성균관대학교 2024년 퀀트응용경제학과 석사학위논문을 수정보완한 것임.

†NICE피앤아이, 대체투자평가정책팀 연구원, 서울특별시 영등포구 국회대로 68길 11(여의도동) 07237. E-mail: vs6609@naver.com.

‡교신저자. 성균관대학교, 경제학과/퀀트응용경제학과 교수, 서울특별시 종로구 성균관로 25-2 03063. E-mail: heejoonhan@skku.edu.

Received June 4, 2025, Revised June 15, 2025, Accepted June 16, 2025

1. 서론

기업의 건전성과 안정성은 전체 경제의 지속 가능성과 밀접하게 연결되어 있으며, 특히 상호 연계성이 높은 현대 경제 구조에서는 한 기업의 부도가 연쇄적인 위기로 이어질 수 있다. 따라서 기업부도 예측은 투자자, 금융기관, 정책 입안자 모두에게 핵심적인 과제로, 투자 리스크를 줄이고 금융시장의 안정성을 유지하는 데 중요한 역할을 한다.

본 연구는 재무 데이터 및 거시경제 데이터를 기반으로 통계적 방법과 머신러닝 기법을 활용해 기업 부도를 예측하고자 하며, 이를 통해 부도 위험이 높은 기업을 조기에 식별하여 조기 경보 시스템으로서의 기능을 수행하고, 합리적인 투자 및 신용 판단을 돕는 것을 목표로 한다. 또한 정밀한 예측은 금융기관의 리스크 관리 향상, 시장의 투명성 제고, 나아가 경제 전반의 안정성 확보에 기여할 수 있을 것으로 기대된다.

최근에는 전통적인 통계 분석 방법에 더해, 다양한 머신러닝 기법들이 기업 부도 예측에 도입되고 있다. 로지스틱 회귀와 같은 통계적 방법은 해석력과 이론적 기반이 강점이지만, 비선형성과 변수 간 상호작용을 충분히 반영하지 못하는 한계가 있다. 반면 머신러닝 기법은 고차원 데이터, 복잡한 변수 구조, 비선형적인 관계 등을 보다 유연하게 학습할 수 있는 능력을 지닌다. 본 연구는 이러한 방법론적 차이를 비교하고, 실제 데이터를 활용하여 양자의 예측 성능을 실증적으로 분석하고자 한다.

한편, 기업부도 예측에서 중요한 도전 과제 중 하나는 데이터의 불균형성이다. 실제 현실에서 부도를 경험한 기업은 전체 기업 중 극히 일부에 불과하며, 이에 따라 데이터는 정상 기업에 편향되어 있다. 이러한 불균형 데이터(unbalanced data)는 예측 모형이 주로 다수 클래스(정상기업)의 특성만을 학습하고, 소수 클래스(부도기업)에 대한 예측 성능이 떨어지는 문제를 야기한다. 이를 해결하기 위해 본 연구는 SMOTE, ADASYN 등과 같은 오버샘플링 기법을 도입하여 모델의 성능을 개선하고자 한다.

선행연구 중 송현준 외 (2021)는 우리나라 부도기업 예측에서 불균형 데이터에 대응한 오버샘플링을 고려한 유일한 연구인데, 이는 대부분 비상장회사를 고려하고 설명변수로 5개의 재무비율만을 사용하고 있다. 이에 반해 본 연구는 상장기업을 대상으로 하고 재무비율뿐만 아니라, 거시경제 변수까지 포함한 총 43개의 설명변수를 이용하고 있는 것이 특징이다. 상장기업과 비상장기업은 적용되는 회계기준에서 차이를 보인다. 상장기업은 K-IFRS를 적용하여 재무제표를 작성하는 반면, 비상장기업은 일반기업 회계기준을 따르거나 선택적으로 K-IFRS를 적용할 수 있다. 이처럼 상장기업과 비상장기업은

회계기준이 달라 예측에 설명변수로 사용할 수 있는 재무 관련 데이터에 차이가 있다.

2000년에서 2019년까지의 상장기업 데이터를 대상으로 수행한 본 연구의 주요 결과는 다음과 같다. 첫째, SMOTE 및 ADASYN과 같은 오버샘플링 기법은 불균형 데이터 문제를 완화하며, 특히 재현율(recall) 향상에 효과적이었는데 이는 실제 부도기업 예측 사례를 더 많이 찾아내는 능력이 좋아졌다는 뜻이다. 둘째, SMOTE를 적용한 XGBoost 모형이 가장 우수한 예측 성능을 보였고, ADASYN 적용 시에도 유사한 결과를 나타냈다. 셋째, 로지스틱 회귀는 오버샘플링 기법과 결합할 때 높은 Recall을 보였지만, 정밀도(precision) 저하(부도라고 예측한 기업 중 실제로는 정상기업인 경우가 많다는 뜻)로 인해 정상기업의 오분류가 빈번하게 발생하여 실무적 한계가 있는 것으로 나타났다.

본 논문의 구성은 다음과 같다. 2장에서는 선행연구 및 모형을 설명한다. 3에서는 본 연구에서 고려한 데이터, 불균형 데이터 처리법, 예측 성능평가 지표를 설명한다. 4장에서는 예측 및 분석 결과 등 논문의 주요 결과를 설명하고, 5장에서 연구의 결론을 제시한다. 부록1은 설명변수 목록을 보여준다.

2. 선행연구 및 모형

2.1. 선행연구

국내외 학계와 실무에서는 Beaver (1966)와 Altman (1968)의 연구를 시작으로 전통적 재무비율 분석에서부터 통계적 모형, 머신러닝 기법에 이르기까지 폭넓은 방법론을 활용한 연구들이 활발히 이루어지고 있다. Beaver (1966)는 단변량 분석을 통해 부도 기업과 건전 기업의 재무 비율을 비교하여 부도 예측의 가능성을 탐구하였다. 이 연구는 재무 비율이 기업의 재무 건전성을 평가하는 데 중요한 도구임을 강조하였다. Altman (1968)은 다변량 판별분석(multiple discriminant analysis)을 활용하여 여러 재무 비율을 결합한 Z-Score 모형을 개발하였다. 이 모형은 운전자본 대비 총자산, 이익잉여금 대비 총자산, 세전이익 대비 총자산, 자기자본의 시장가치 대비 총부채, 매출액 대비 총자산의 5가지 재무 비율을 사용하여 기업의 부도 가능성을 수치화하였다. 이후에도 기업 부도 예측에 대한 다양한 연구가 진행되었습니다. Ohlson (1980)은 로지스틱 회귀분석을 활용한 O-Score 모형을 제안하였으며, Zmijewski (1984)는 프로빗 분석을 통한 부도 예측 모형을 개발하였다.

국내에서도 기업의 부도 예측을 위한 많은 연구가 이루어졌다. 이인로 · 김동철 (2015)은 국내 기업을 대상으로 부도 예측 모형의 예측력을 평가하며, 회계정보와 시장정보를 활용한 다양한 부도 예측 모형을 비교하였다. 이 연구

는 기존의 부도 예측 모형을 회계 모형, 시장모형, 해저드 모형으로 정의하고 각 모형별 예측력을 살펴보았다. 해당 연구에서 회계 모형(Z-점수와 O-점수), 시장모형(부도거리모형), 그리고 회계와 시장정보를 통합한 헤저드모형을 중심으로 분석을 진행하였다. 특히, Campbell *et al.* (2008)의 헤저드모형을 국내 기업에 적용한 결과와 이를 수정하여 새롭게 제안된 헤저드모형의 예측력을 비교하였다. 이인로 · 김동철 (2015)의 연구 결과, 국내 기업에 적합하도록 수정된 헤저드모형이 가장 높은 부도 예측력을 보였으며, 이는 기존의 미국 모형을 단순히 적용하기보다는 국내 실정에 맞게 수정할 필요성을 시사한다. 본 연구는 또한 2008년 글로벌 금융위기 기간을 포함하여 분석을 진행함으로써 부도 예측모형의 실증적 신뢰성을 제고하였다.

박종원 · 안성만 (2014)의 연구는 한국 외부감사 대상 기업을 대상으로 재무비율을 활용한 부도 예측 모형을 개발하고, 그 모형의 변별력과 정확도를 검증하였다. 본 연구는 기존의 선행연구와 더불어 회계적으로 기업의 건정성에 영향을 주는 주요 재무비율을 선정하였으며, 이들 변수의 결합을 통해 부도기업과 정상기업을 효과적으로 구별할 수 있는 예측 모형을 구축하였다. 기존의 기업 부도 예측 연구는 전통적으로 재무비율을 기반으로 한 통계적 모형을 활용한 방법론이 우세하여, 로지스틱 회귀, 판별분석, 다중회귀분석과 같은 통계적 기법을 기반으로 기업의 부도 가능성을 추정하는 연구가 주류를 이루었다. 그러나 최근에는 데이터 활용 폭증과 알고리즘의 발전을 바탕으로 머신러닝 기법의 연구의 필요성도 대두된다.

송현준 외 (2021)는 불균형 데이터를 고려하면서 머신러닝 기법을 활용하였는데, 2012년부터 2016년까지 5개년의 표본기간의 외감기업과 비외감기업의 데이터를 사용하였다. 이 연구는 대부분 비상장회사를 고려하고 설명변수로 5개의 재무비율만을 사용하고 있다.

2.2. 모형

2.2.1. 통계적 모형

로지스틱 회귀는 이진 분류에 적용하는 전통적인 통계적 기법으로, 설명변수들의 선형 결합(linear combination)을 사용하여 종속변수가 이산형(주로 0과 1)인 경우 특정 범주에 속할 확률을 예측하는 것이다. 설명변수의 벡터 x_i가 주어졌을 때 종속 변수 y_i가 일$y_i = 1$ 확률을

$$p(x_i) = P(y_i = 1|x_i)$$

라 하면, 로짓 변환(logit transformation)은 $p(x_i)$와 x_i 간의 비선형 관계를 선형 함수 형태로 변환하여 선형 회귀 모형의 방법론을 적용할 수 있게 한다. 이는 다음과 같이 정의된다.

$$log\Big(\frac{p(x_i)}{1-p(x_i)}\Big) = x_i^T\beta.$$

이로부터 로지스틱 회귀 모형은 다음과 같이 도출된다.

$$p(x_i) = \frac{1}{1+e^{(-x_i^T\beta)}}.$$

임곗값을 0.5로 할 경우, $p(x_i)$가 0.5보다 크게 예측되면 이를 $y_i = 1$로 분류한다.[1]

다만 본 연구에서는 총 43개의 설명변수를 사용하는데, 이를 모두 사용하여 로지스틱 회귀 모형을 추정하면 예측력이 매우 낮은 것으로 나타났다. 따라서 4.2절에서 설명하는 별도의 사전 테스트(pre-test)를 거쳐 선별된 설명변수만을 로지스틱 회귀 모형에 사용하였다. 이는 아래 설명하는 머신러닝 모형들을 추정할 때와 구별되는 차이점인데, 머신러닝 모형을 추정하고 예측치를 만들 때에는 별도의 사전 테스트를 거치지 않고 모든 설명변수를 사용하였다.

2.2.2. 머신러닝 모형

랜덤 포레스트(random forest)는 의사결정나무(decision tree)를 기반으로 하는 앙상블 학습 기법의 하나이다. 랜덤 포레스트는 배깅을 변형한 것이다. 배깅은 원래의 훈련 데이터에서 총 개의 부트스트랩 샘플을 만들고 각 부트스트랩 샘플에 의사결정나무 모형을 학습하는데, 부트스트랩 샘플들이 유사한 데이터를 포함할 수밖에 없으므로 각 부트스트랩 샘플에서 학습된 의사결정나무 모형은 서로 양의 상관관계(positively correlated)를 가지게 된다. 즉, 배깅만으로는 분산 감소 효과가 충분하지 않을 수 있다.

Breiman (2001)이 제안한 랜덤 포레스트는 상관성이 낮은(de-correlated) 의사결정나무들을 생성하여 전체적인 분산을 더욱 줄이는 방법을 사용한다. 랜덤 포레스트는 각 부트스트랩 샘플에서 분할을 수행할 때마다 사용할 설명

[1]부도 여부를 나타내는 종속 변수 y_i의 시점과 설명변수의 벡터 x_i의 시점은 동일하다. 만약 부도가 발생한 연도의 해당 기업 재무제표를 구할 수 없는 경우는 전년도 혹은 전전년도 재무제표를 활용하였다. 부도기업은 부도 이전부터 징후가 발생하는 경우가 많으므로, 부도가 발생하기 이전 연도의 정상기업 데이터에는 포함하지 않았다. 이러한 데이터 처리는 선행연구의 방식을 따른 것이다.

변수를 무작위로 선택한다. 각 노드에서 분할을 수행할 때, 전체 m개의 예측변수 중에서 m^*개만 무작위로 선택하여 후보로 사용한다. 선택된 m^*개의 변수 중 하나만을 이용하여 해당 노드에서 최적의 분할을 수행한다. 새로운 노드를 만들 때마다 다시 m^*개의 새로운 설명변수를 무작위로 선택한다(고정된 설명변수가 아님). m^*값은 일반적으로 회귀 문제에서는 $m^* = [m/3]$로 선택하고 분류 문제에서는 $m^* = [\sqrt{m}]$로 선택한다.

XGBoost는 그래디언트 부스팅(gradient boosting)을 개선한 알고리즘으로, 성능과 효율성을 극대화하기 위해 다양한 최적화 기법과 정교한 기능을 통합한 것이 특징이다. 이 기법은 Chen and Guestrin (2016)이 소개하였으며, 분류와 회귀 모두에 사용가능하며 예측 성능이 좋은 것으로 알려져있다. 부스팅은 앙상블 학습 방법 중 하나로, 여러 약한 학습기(weak learners)를 결합하여 강력한 모형을 만드는 것이다. 주로 의사결정나무를 약한 학습기로 사용하며, 이전 트리의 오류를 보완하는 방식으로 새로운 트리를 추가하는 과정이 반복된다. XGBoost는 정규화(regularization) 기능을 추가하여 기존 그래디언트 부스팅의 약점인 과대적합을 방지한다. 특히 XGBoost는 병렬 CPU 환경에서 병렬 학습이 가능하기 때문에 기존 순차적으로 학습하는 그래디언트 부스팅보다 빠르게 학습할 수 있다.

심층신경망(deep neural network, DNN)은 인공신경망(artificial neural network, ANN)의 확장된 형태로, 여러 개의 은닉층(hidden layer)을 포함하여 데이터를 점진적으로 높은 수준의 추상화로 표현할 수 있도록 설계된 모형이다. 기본적으로 DNN은 입력층, 은닉층, 그리고 출력층으로 이루어지며, 은닉층의 수가 많을수록 모형의 깊이는 더욱 깊어진다. 입력 데이터에 가중치(weight)와 편향(bias)을 적용하고, 활성화 함수(activation function)를 통해 비선형성을 부여하여 복잡한 패턴을 학습한다. 이러한 과정을 여러 층에 걸쳐 반복한다. 본 논문에서는 세 개의 은닉층으로 각각 64, 32, 16개의 노드를 지닌 것으로 설정하였으며 ReLu함수를 활성화 함수를 사용하였다.

3. 데이터, 불균형 데이터 처리법, 성능평가 지표

3.1. 부도기업 데이터 정의 및 표본 설정

기업 부도 예측 연구는 표본 기업 중 부도 기업의 정의와 재무제표 상 부도의 인식 범위를 어디까지 포함하는가, 부도 관측 기간을 어떻게 설정하느냐에 따라 연구의 방향성과 연구결과가 달라진다. 기업의 부도를 예측하는 모형 연구에서 정확한 모형을 도출하기 위해서는 기업의 부도에 대한 명확한 의미를

정하는 것이 중요하다. 증권거래소[2], 금융투자업규정[3], 금융감독원, 금융결제원 등 다양한 기관에서 기업의 부도에 대해 정의하고 있지만, 실제로 기업의 부도를 판단하는 기준은 연구 목적에 따라 달라질 수 있다. 또한 재무제표 수치상 실제로 기업이 이미 심각한 자본잠식과 같은 부실 징후가 관찰된 뒤에도, 상당한 시간이 흐른 후에야 비로소 공식적인 부도로 공시되는 사례가 빈번하게 발생한다. 따라서 기업 부도 예측 연구에서 더욱 유의미한 예측 모형의 결과를 얻으려면, 기업의 부도에 대한 명확한 정의를 수립하는 것이 필수적이다.

다수의 부도 관련 선행 연구에서도 다양한 기준으로 부도 사건을 정의하고 있다. 박종원 · 안성만 (2014)은 금융결제원에서 제공하는 당좌수표 정지 또는 약속어음 부도로 인해 당좌거래가 정지된 기업을 부도로 정의하였으며, 오세경 외 (2017)과 이인로 · 김동철 (2015)은 상장폐지가 결정된 기업 중 부도에 관련된 공시가 발생한 기업들을 부도 발생 기업으로 정의하였다. 부도를 인식하는 기준을 넓히면, 부실 징후가 미미한 상태에 있는 기업까지 조기에 탐지할 수 있지만 상대적으로 건전한 기업까지 부도 위험 기업으로 잘못 분류함으로써 모형의 신뢰도를 하락시킬 수도 있다. 본 연구는 이인로 · 김동철 (2015), 오세경 외 (2017) 등의 선행연구와 같이 유가증권시장 및 코스닥 시장에서 '상장폐지'가 결정된 기업 중 부도에 관련된 공시 [4] 가 발생한 기업들을 부도 발생 기업으로 정의하였다.

본 연구에서는 표본은 상장공시 사이트 KIND(https://kind.krx.co.kr)에서 관찰 연도별 마지막 영업일 기준 유가증권시장 및 코스닥 시장에 상장된 기업 [5] 을 표본으로 설정하였다. 재무비율 및 정보는 모두 FnGuide의 데이터를 사용하였다. 본 연구는 2000년부터 2019년까지의 공시 결산 재무정보와 회계정보를 사용하였다. 이때 전체 표본에 대해서 KSIC 11차 대분류 기준 "K. 금융 및 보험업(64 66)"에 해당하는 금융권 기업은 영업의 성격과 회계보고 특성이 비 금융업 기업과 차별적이므로 연구의 동질성을 위해서 표본에서 제외하였다. 표본기간 동안 상장폐지 이후 재상장하였다 할지라도 실제 회복 여부에 대한 확인이 어려움과 동시에 차주를 정상적인 차주로 보기 어렵다는 판단하에 정상기업의 표본에서 제외하였다. 본 연구에 사용한 정상기업과 부도기업의

[2] "유가증권상장규정 제37조(주권의 상장폐지기준)"에 따라 ①은행거래정지 ②감사의견부적정 ③감사의견거절 ④거래실적부진 ⑤자본전액잠식 ⑥최종부도 기업

[3] 제8-19조의9 ③의2에 따르면, 부도는 "원리금의 적기상환이 이루어지지 않거나 기업회생절차 또는 파산절차의 개시가 있는 경우"로 정의

[4] 상장시장의 이전 상장, 피흡수 합병, 지주회사(최대주주등)의 완전자회사화 등과 같은 상장폐지 공시의 경우 부도 사건과 상관없는 공시로 부도기업의 대상에서 제외하였다.

[5] 유가증권시장에서 "외국주권, 투자회사, 부동산투자회사, DR, 신주인스권증권, ELW, ETN, ETF, 수익증권"과 코스닥 시장의 "기업인수목적회사, 외국주권, DR, 신주인수권증권, 신주인수권증서"의 경우는 표본에서 제외하였다.

연도	정상기업	상장폐지	비율	연도	정상기업	상장폐지	비율
2000	1120	29	2.6%	2010	1608	76	4.7%
2001	1283	13	1.0%	2011	1625	47	2.9%
2002	1400	38	2.7%	2012	1615	43	2.7%
2003	1441	29	2.0%	2013	1628	26	1.6%
2004	1448	48	3.3%	2014	1653	16	1.0%
2005	1487	47	3.2%	2015	1719	20	1.2%
2006	1551	6	0.4%	2016	1764	11	0.6%
2007	1607	11	0.7%	2017	1856	11	0.6%
2008	1648	19	1.2%	2018	1912	15	0.8%
2009	1619	68	4.2%	2019	2012	7	0.3%

표 1: 연도별 정상기업과 부도기업 (상장폐지) 현황. 이 표는 정상기업과 부도기업의 수, 그리고 부도기업의 비율을 제시한다.
Table 1: ANNUAL STATUS OF SOLVENT AND BANKRUPT (DELISTED) LISTED COMPANIES IN KOREA. This table presents the number of solvent and bankrupt firms, as well as the proportion of bankrupt firms.

표본 구성은 다음 <표1>에서 제시한다.

3.2. 설명변수

국내외 기존 연구에서 부도 예측에 유의하게 영향을 미치는 것으로 알려진 36개의 회계 비율과 3개의 거시경제 지표를 <부록 A>과 같이 후보 변수로 선정하였다. 회계 비율은 수익성, 유동성, 규모, 건정성, 현금흐름, 활동성 등 6개 범주로 구분하였다. 본 연구에서 설명변수로 선택한 회계 관련 데이터는 재무 건전성, 수익성, 유동성 등 여러 측면에서 기업의 상태를 평가하며, 부실 가능성을 사전에 파악하는 데 중요한 역할을 한다.

먼저, TLTA(total liabilities to total assets)는 총부채를 총자산으로 나눈 비율로, 기업의 자산 대비 부채 수준을 나타낸다. 이 비율이 높을수록 기업이 많은 부채를 보유하고 있음을 의미하며, 재무적 위험이 증가할 가능성이 있다. 한편, EBITTA(earnings before interest and taxes to total assets)는 자산 대비 영업이익 창출 능력을 보여주는 지표로, 기업이 자산을 효율적으로 활용하고 있는지를 평가한다. EBITTA가 낮을 경우 수익성 부족으로 인해 부실 가능성이 높아질 수 있다. 수익성과 관련된 또 다른 지표인 NITA(net income to total assets)는 순이익을 총자산으로 나눈 비율로, 기업의 자산을 통해 창출된 순

이익 수준을 나타낸다. 이와 유사한 RETA(retained earnings to total assets)는 이익잉여금을 총자산으로 나눈 비율로, 과거에 벌어들인 이익이 얼마나 기업 내에 보유되고 있는지를 평가한다. 이 비율이 높을수록 재무 안정성이 높은 것으로 간주된다.

또한, WCTA[6](working capital to total assets)는 순운전자본을 총자산으로 나눈 비율로, 기업의 단기 유동성을 나타내는 중요한 지표이다. WCTA가 높을수록 기업이 단기 채무를 이행할 능력이 크다는 것을 의미한다. 현금흐름과 관련된 지표로는 CASHTA(cash to total assets)와 FFOEQ(funds from operations to equity)가 있다. CASHTA는 총자산 중 현금성 자산의 비중을 나타내며, 기업의 즉시 사용 가능한 자금 수준을 보여준다. FFOEQ는 영업활동으로부터 창출된 현금흐름을 자본으로 나눈 비율로, 자본 대비 현금 창출 능력을 평가하는 데 사용된다. 이와 더불어, FFOTA(funds from operations to total assets)는 영업활동으로부터 창출된 현금흐름을 총자산으로 나눈 비율로, 자산 대비 현금흐름의 효율성을 측정한다.

이처럼 다양한 재무 지표들은 기업의 부실 가능성을 예측하는 데 유용한 정보를 제공한다. 특히, 부채비율(TLTA)과 같은 건전성 지표는 기업의 재무적 위험을 직접적으로 나타내며, 수익성 지표(EBITTA, NITA, RETA)는 기업의 이익 창출 능력을 평가한다. 또한, 유동성 지표(WCTA, CASHTA)는 단기 채무 이행 능력을 평가하는 데 중요한 역할을 한다. 이러한 지표들의 종합적인 분석은 기업의 재무 상태를 정확히 평가하고, 부실 위험을 효과적으로 예측하는 데 기여한다.

거시경제 상황을 반영하는 변수로는 시장금리(CP 91일물), 선행종합지수, GDP 성장률을 고려하였다. 기업 부도를 예측하는 선행연구들이 대부분 회계 관련 데이터만을 사용하는데, 기업의 부도에 경기상황이 영향을 미칠 수 있기 때문에 거시경제 상황을 반영하는 변수들을 설명변수를 추가하였다. 후속 연구는 보다 광범위한 거시경제지표를 고려해 볼 수 있을 것이다.

3.3. 불균형 데이터 처리 및 데이터 분할

3.3.1. 불균형 데이터 처리법

기업의 부도는 일반적으로 전체 기업 중 극히 일부에서 발생하며, 정상 기업에 비해 부도 기업의 수는 현저히 적다. 이처럼 한쪽 범주에 속한 데이터의 빈도가 다른 범주에 비해 지나치게 적은 경우를 불균형 데이터라고 한다. 불균형 데이터의 주요 문제는 예측 모형이 데이터 분포에서 우세한 다수 클래스

[6]순운전자본은 유동자산에서 유동부채를 차감하여 구한다.

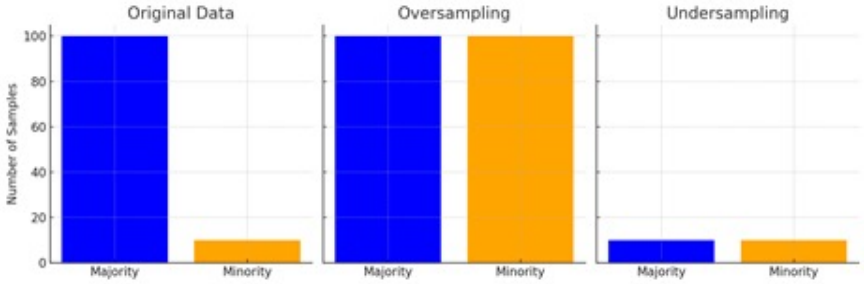

그림 1: 과대 표집과 과소 표집. 이 그림은 오버샘플링과 언더샘플링의 예시를 보여준다.
FIGURE 1: OVERSAMPLING AND UNDERSAMPLING. This figure illustrates examples of oversampling and undersampling.

(majority class)에 편향되기 쉽다는 점이다. 이에 따라 소수 클래스(minority class)에 대한 학습이 충분히 이루어지지 않아, 결과적으로 소수 클래스에 대한 예측 성능이 크게 저하될 수 있다. 이는 실제 분석과 예측 과정에서 중요한 오류를 유발할 수 있다.

이러한 문제를 해결하기 위한 다양한 방법들이 지속적으로 연구되어 왔다. 대표적인 방법은 샘플링(sampling) 기법으로, 이는 데이터의 분포를 조정하여 학습 모형의 균형을 맞추고자 한다. 과소 표집(undersampling) 방법은 다수 클래스의 데이터를 줄여 소수 클래스의 수에 맞추는 방식이다. 하지만, 이 경우 다수 클래스에 존재하는 유용한 정보가 손실될 수 있다. 과대 표집(oversampling) 방법은 소수 클래스의 데이터를 인위적으로 늘려 다수 클래스와 균형을 맞추는 방식으로, 이에 따라 과적합 우려와 더불어 높은 계산 비용이 발생할 수 있다. <그림1>은 과소 표집과 과대 표집의 예시를 나타낸다. 이처럼 불균형 데이터 문제는 예측 모형의 정확도와 신뢰성에 중대한 영향을 미치므로, 적절한 샘플링 기법의 선택과 적용이 필수적이다.

과소 표집은 다수 클래스의 데이터를 소수 클래스에 맞추어 일부를 제거하는 방식으로, 이 과정에서 정보 손실이 발생할 수 있다. 또한, 표본 추출 시 초기 설정에 따라 선택되는 관측치가 달라질 수 있어, 모형의 예측력이 변동될 수 있다는 한계가 있다. 이러한 제약으로 인해, 대부분의 불균형 데이터 연구에서는 과소 표집보다 과대 표집 기법이 더 널리 활용된다. 본 연구에서는 불균형 데이터 문제를 해결하기 위해 박근우 · 정인경 (2019)의 선행 연구에서 가장 효과적인 방법으로 제시된 SMOTE와 ADASYN 기법을 적용하여 데이터 불균형을 보정하였다.

SMOTE는 Chawla *et al.* (2002)에 의해 제안된 기법으로, 소수 클래스의 데이터를 단순히 복제하는 대신 기존 데이터 포인트들 사이에서 새로운 데이터를 합성하여 생성한다. 이에 따라 단순 복제 방식에 비해 과적합의 가능성을 낮출 수 있으며, 특히 분류 문제에서 소수 클래스의 학습 성능을 향상하는 데 효과적이다.

SMOTE의 작동 원리는 다음과 같다. 먼저, 소수 클래스의 각 데이터 포인트 x_i에 대해 k개의 최근접 이웃(k-nearest neighbors, KNN)을 찾는다. 이후, 선택된 k개의 이웃 중 무작위로 하나의 이웃 x_{zi}를 선택한다. 선택된 이웃과 원본 데이터 포인트 사이의 선형 보간(linear interpolation)을 통해 새로운 합성 데이터(new synthetic sample)를 생성한다. 이 과정은 소수 클래스의 데이터 분포를 보다 고르게 확장하여 학습에 도움을 준다. 해당 과정을 각 소수 클래스의 데이터에 대해 여러 번 반복함으로써, 충분한 양의 새로운 합성 데이터를 생성하고, 이를 원래의 소수 클래스 데이터와 결합하여 모형 학습에 사용한다.

ADASYN은 He *et al.* (2008)이 처음 제안한 기법으로, 기존의 SMOTE를 발전시켜 데이터 분포의 밀도에 따라 새로운 샘플의 생성 비율을 조절하는 특징을 지닌다. ADASYN의 핵심 아이디어는 소수 클래스 중 학습하기 어려운 데이터를 우선적으로 보강하는 데 있다. 이를 위해 먼저 소수 클래스와 다수 클래스 간의 불균형 비율을 계산하여 전체적인 불균형의 정도를 평가한다. 이후 k-최근접 이웃(KNN) 방법을 활용하여 각 소수 클래스 샘플 주변에 존재하는 다수 클래스 샘플의 비율을 측정하고, 이를 기반으로 학습 난이도를 계산한다. 학습 난이도가 높은 샘플일수록 더 많은 가중치를 부여하여 새로운 데이터를 생성한다. 이러한 데이터는 기존 샘플과 이웃 샘플 간의 선형 보간을 통해 생성되며, 결과적으로 모형이 복잡한 분포를 더 잘 학습하도록 돕는다.

ADASYN은 기존의 SMOTE와 몇 가지 중요한 차이점을 가진다. SMOTE는 모든 소수 클래스 샘플에 대해 균일한 수의 데이터를 생성하는 반면, ADASYN은 샘플별 학습 난이도에 따라 생성 수를 조절함으로써 보다 적응적(adaptive)인 방식으로 접근한다. 이를 통해 모형 학습이 어려운 샘플에 중점을 두어 결정 경계를 조정함으로써 예측 성능을 향상시킬 수 있다.

SMOTE와 ADASYN의 하이퍼 파라미터는 기본 설정값을 따라 정하였다. KNN에서 이웃의 개수 $k = 5$ 로, 소수 클래스 비율 beta=1로(소수 클래스의 샘플 수를 다수 클래스의 샘플 수와 동일하게 맞춤), 거리 측정 방식은 distance=Euclidean으로 설정하였다.

3.3.2. 데이터 분할

본 연구에서는 기업 부도 예측 모형의 학습 및 검증을 위해 전체 데이터를 학습용(train)과 평가용(test) 데이터로 7:3의 비율로 분할하였다. 이후 학습용 데이터의 70%는 실제 모형 학습에 사용되는 훈련 데이터로, 나머지 30%는 모형의 일반화 성능을 확인하기 위한 검증(validation) 데이터로 재분할하였다. 부도기업에 비해 정상기업의 수가 상대적으로 많은 불균형 구조를 가지고 있어, 학습 데이터에 한하여 오버샘플링 기법을 적용하였다.

3.4. 모형의 성능평가 지표

혼동 행렬(confusion matrix)은 분류 문제에서 모형의 성능을 평가하기 위한 가장 기본적인 도구이다. 분류 모형이 양성과 음성을 얼마나 잘 분류하는지를 평가하는 측정값을 기반으로 평가한다.

Confusion Matrix		Actual	
		Positive	Negative
Predict	Positive	TP	FP
	Negative	FN	TN

표 2: 혼동 행렬. 이 표는 혼동 행렬을 보여준다.
Table 2: CONFUSION MATRIX. This table presents the confusion matrix.

이진 분류의 혼동 행렬은 <표2>와 같이 2×2 형태의 매트릭스로 표현되며, True Positive(TP), False Positive(FP), True Negative(TN), False Negative(FN) 네 가지 값을 포함한다. TP는 실제로 양성(Positive, 본 논문에서는 부도기업)인 데이터를 모형이 올바르게 양성으로 예측한 경우이며, TN은 실제로 음성(Negative, 본 논문에서는 정상기업)인 데이터를 정확히 음성으로 예측한 경우이다. 반면, FP는 실제로 음성(정상기업)인 데이터를 잘못 양성(부도기업)으로 예측한 경우로, 제1종 오류(Type 1 Error)에 해당한다. FN은 실제로 양성(부도기업)인 데이터를 잘못 음성(정상기업)으로 예측한 경우로, 제2종 오류(Type 2 Error)를 의미한다. 혼동 행렬을 통해서 산출되는 주요 평가지표 중 본 논문에서 사용한 지표는 다음과 같다.

- 정확도(accuracy)
  정확도는 전체 샘플 중에서 모형이 정답을 맞춘 샘플의 비율을 의미한다. 이는 양성과 음성 여부에 관계없이 예측이 실제값과 일치한 비율을

나타내며, 분류 모형의 전반적인 성능을 평가하는 가장 기본적인 지표이다. 본 연구에서는 전체 기업 중에서 부도 여부를 모형이 정확히 예측한 비율을 의미한다.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

- 정밀도(precision)
  정밀도는 모형이 양성이라고 예측한 것들 중 실제로 양성인 비율을 나타내는 지표이다. 양성이라고 판단한 것 중에서 얼마나 맞췄는가를 측정한다. 본 연구에서는 부도기업이라고 예측한 기업 중 실제로 부도인 기업의 비율을 의미하며, 부도예측의 신뢰도를 판단하는 데 활용된다.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

- 재현율(recall)

  민감도(Sensitivity)라고도 불리는 재현율은 실제 양성 샘플 중 모형이 양성으로 예측한 샘플의 비율을 나타낸다. 양성 샘플을 얼마나 잘 찾아내는지를 평가한다. 본 연구에서는 실제 부도기업 중에서 부도기업이라고 예측한 비율을 의미한다.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

- 특이도(specificity)

  특이도는 실제 음성 샘플 중 모형이 음성으로 예측한 샘플의 비율을 나타낸다. 음성 샘플을 얼마나 잘 맞추는지를 평가한다. 본 연구에서는 실제 정상기업 중에서 정상기업이라고 예측한 비율을 의미한다.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}.$$

- F1-Score

  F1-Score는 정밀도와 재현율의 조화 평균으로, 두 지표 간의 균형을 평가하는 데 사용되는 지표이다. 정밀도와 재현율 사이의 상충 관계를 고려해야 할 때 특히 유용하며, 데이터의 클래스 불균형 상황에서도 성능을

효과적으로 평가할 수 있다. F1-Score는 한쪽 지표가 매우 높고 다른 쪽이 낮은 경우, 이러한 불균형을 보완해 모형의 실제 성능을 더 정확히 반영한다. F1-Score의 값이 1에 가까울수록 정밀도와 재현율 간의 균형이 잘 맞춰졌음을 의미하며,

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

- ROC 곡선(receiver operating characteristic curve)

  ROC 커브와 AUC(area under the curve)는 특히 이진 분류 문제에서 유용하게 사용되며, 클래스 불균형이 있는 데이터에서도 비교적 강건한 평가 도구로 알려져 있다. ROC 커브는 모형이 다양한 임곗값(threshold)에 따라 얼마나 잘 양성과 음성을 구분하는지를 시각적으로 나타낸 곡선이다. ROC 커브의 X축은 False Positive Rate(FPR), Y축은 True Positive Rate(TPR)로 구성된다. FPR은 실제 음성인 데이터 중에서 잘못 양성으로 예측한 비율이며, TPR은 실제 양성인 데이터 중에서 정확히 양성으로 예측한 비율이다.

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \ \text{TPR}(\text{Recall}) = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

  ROC 커브는 모형의 분류 기준(임곗값)을 연속적으로 변화시키며 측정한 FPR과 TPR의 조합을 연결한 곡선으로, 시작점은 (0,0), 끝점은 (1,1)을 기준으로 한다. ROC 커브가 왼쪽 위 모서리에 가까울수록, 즉 FPR은 낮고 TPR은 높은 영역에 위치할수록 모형의 분류 성능이 우수함을 의미한다.

- 곡선하면적(area under the curve, AUC)

  AUC는 ROC 커브 아래의 면적을 의미하며, 모형의 전반적인 성능을 단일 값으로 요약한다. AUC의 값은 0과 1 사이의 값을 가지며, 값이 클수록 모형의 분류 성능이 우수함을 나타낸다. AUC가 1에 가까울수록 모형이 모든 임곗값에서 높은 TPR과 낮은 FPR을 유지할 수 있음을 의미한다. 반대로 AUC가 0.5에 가까우면 모형이 무작위로 분류하는 것과 동일한 성능을 나타내며, 0.5 미만의 AUC는 성능이 오히려 무작위보다 낮음을 의미한다. 일반적으로 통용되는 기준은 AUC 값이 0.7~0.8인 경우 평균적인 성능의 모형으로, 0.8 이상인 경우 좋은 성능의 모형으로 판단한다.

4. 예측 및 분석 결과

4.1. 정상기업과 부도기업의 기초통계량

본격적인 부도 예측에 앞서 설명변수들의 기초통계량을 통해 변수별 정상기업과 부도기업 간의 특성을 살펴보았는데, 이는 <표3>에 제시되어 있다.

변수명	부도여부	평균	표준편차	최소값	최대값
BWTA	1(부도)	0.480	1.107	0.000	15.421
	0(정상)	0.153	0.170	-0.052	8.320
CACL	1	1.196	1.895	0.000	21.104
	0	2.548	4.744	0.005	426.877
CAR	1	-0.363	2.069	-25.559	0.995
	0	0.554	0.302	-25.477	0.991
CASHTA	1	0.042	0.080	-0.036	0.644
	0	0.086	0.088	-0.001	0.889
CATA	1	0.446	0.232	0.001	0.994
	0	0.501	0.180	0.002	0.995
CFOTA	1	-1.186	3.331	-54.820	2.427
	0	0.040	0.262	-12.582	28.537
CLCA	1	8.722	120.868	0.047	2937.145
	0	0.786	1.524	0.002	218.849
CLTA	1	0.921	1.250	0.003	15.761
	0	0.324	0.191	0.002	8.780
CLTL	1	0.743	0.267	0.009	1.004
	0	0.739	0.187	0.010	1.000
EBITTA	1	-0.430	1.845	-27.338	0.567
	0	0.036	0.098	-4.526	0.586
EBTA	1	-1.241	3.350	-54.884	2.407
	0	0.013	0.267	-13.093	28.531
FATA	1	0.846	1.822	-0.536	25.871
	0	0.225	0.233	-0.040	15.768
FATO	1	22144.944	526482.555	0.000	12817910.5
	0	14.808	133.993	0.002	9519.155
FFOEQ	1	-0.342	5.739	-71.685	79.633
	0	0.079	1.440	-185.911	103.859

변수명	부도여부	평균	표준편차	최소값	최대값
FFOTA	1	-0.251	1.367	-18.207	15.979
	0	0.044	0.110	-6.027	1.215
FUIL	1	0.566	0.266	-0.306	1.470
	0	0.446	0.277	-0.128	1.306
INSL	1	0.177	1.056	0.000	20.985
	0	0.014	0.109	-0.001	13.940
lnCE	1	5.381	2.863	-6.924	12.255
	0	8.888	2.074	-1.370	17.285
lnES	1	-8.997	5.647	-16.630	12.418
	0	7.126	7.945	-16.264	19.355
lnOP	1	-5.973	6.244	-12.975	11.826
	0	5.550	7.284	-15.002	17.891
lnTA	1	10.448	1.293	6.449	15.434
	0	12.078	1.530	6.196	19.681
lnTL	1	10.328	1.499	6.680	16.630
	0	11.114	1.793	5.253	18.916
lnTO	1	9.538	1.946	0.000	14.914
	0	11.777	1.643	0.681	19.312
lnTOI	1	1.756	9.415	-16.576	13.136
	0	11.289	2.195	-15.044	19.387
NCTA	1	0.554	0.232	0.006	0.999
	0	0.497	0.180	0.005	0.998
NISL	1	-13.710	125.999	-2401.094	13.438
	0	-0.198	21.279	-3532.055	83.322
NITA	1	-1.243	3.350	-54.884	2.407
	0	0.006	0.206	-13.093	9.688
OPM	1	-1.755	7.942	-113.074	0.879
	0	-0.135	20.974	-3500.519	0.998
OPTA	1	-0.430	1.845	-27.338	0.567
	0	0.036	0.099	-4.526	0.586
OPTL	1	-0.937	1.470	-14.585	4.939
	0	0.178	1.850	-260.699	33.590
RETA	1	-2.987	6.517	-95.609	0.826
	0	0.173	0.554	-28.112	6.209

변수명	부도여부	평균	표준편차	최소값	최대값
ROE	1	-0.940	1.470	-14.585	4.939
	0	0.007	0.266	-13.093	28.531
SLEQ	1	2.082	19.807	-186.947	233.555
	0	2.094	8.496	-453.053	679.346
SLFA	1	2.450	7.586	0.000	87.380
	0	2.463	4.205	0.001	224.493
SLLL	1	85.873	336.914	-275.540	3801.423
	0	19.523	422.906	0.010	53395.065
SLTA	1	0.968	4.080	0.000	71.797
	0	0.939	0.595	0.000	23.350
TLEQ	1	1.450	23.510	-173.698	299.340
	0	1.312	9.113	-665.957	845.102
TLTA	1	1.352	2.017	0.005	26.559
	0	0.446	0.302	0.009	26.477
WCTA	1	-0.476	1.257	-14.832	0.745
	0	0.178	0.260	-7.990	0.959
WCTO	1	-4.222	90.317	-2140.238	242.497
	0	8.358	681.267	-18651.436	91943.986

표 3: 설명변수의 기초통계량. 이 표는 설명변수의 평균, 표준편차, 최소값, 최대값을 부도 상태별로 구분하여 제시한다 (1 = 부도, 0 = 정상).

TABLE 3: DESCRIPTIVE STATISTICS OF THE EXPLANATORY VARIABLES. This table presents the mean, standard deviation, minimum, and maximum of the explanatory variables, reported separately by bankruptcy status (1 = bankrupt, 0 = solvent).

운전자본 대 총자산 비율(WCTA)을 살펴보면, 부도 기업(bankrupt firms)은 평균 –0.476으로 음수를 기록하여 운전자본이 부족한 상태임을 보였다. 반면 정상 기업은 0.178로 양(+)의 값을 유지하였다. 부도 집단의 표준편차가 1.257로 정상 집단(0.260)보다 약 다섯 배 커, 기업 간 유동성 격차가 현저히 크다는 점이 확인되었다.

유동부채 대 유동자산 비율(CLCA) 역시 부도 기업이 8.722로 정상 기업 0.786보다 11배 이상 높았고, 최대값이 2 937.145에 달해 일부 기업이 극심한 단기 지급불능에 처해 있음을 파악할 수 있었다. 유동부채 대 총자산(CLTA)도 부도 기업 0.921, 정상 기업 0.324로 큰 격차를 보였으며, 현금 및 현금성자산 비율(CASHTA)은 부도 기업 0.042로 정상 기업 0.086의 절반 수준에 머물렀

다. 종합하면 부도 기업은 전반적으로 심각한 유동성 부족 상태에 노출되어 있다.

부채 구조를 나타내는 총부채 대 총자산 비율(TLTA)은 부도 기업이 평균 1.352로 자산보다 부채가 많은 상황이었고, 최대 26.559까지 치솟았다. 정상 기업은 0.446로 비교적 안정적인 범위에 있었다. 자본적정성지표(CAR)는 부도 기업에서 –0.363으로 음수를 기록해 부도기업의 대부분이 자본잠식 상태를 나타냈으며, 최소값 –25.559는 극단적인 자본 손실을 보여준다. 총부채 대 총자본 비율(TLEQ) 역시 부도 기업의 변동폭이 훨씬 컸다. 이처럼 부도 기업은 높은 레버리지와 자본 훼손으로 인해 지급능력 위험이 현저히 크다.

수익성 지표에서도 차이가 뚜렷하다. 영업이익 대 총자산 비율(EBITTA)은 부도 기업이 –0.430으로 음수인 반면 정상 기업은 0.036으로 소폭 양수였다. 순이익 대 총자산(NITA)은 –1.243 대 0.006, 순이익 대 매출(NISL)은 –13.710 대 –0.198로 격차가 극심하였다. 자기자본이익률(ROE) 역시 –0.940 대 0.007로 부도 집단이 순손실에 머물렀다. 특히 NISL의 표준편차가 125.999에 달해 부도 기업 간 손익 변동성이 훨씬 크다는 점이 확인된다.

총자산 로그(lnTA)는 부도 기업이 10.448, 정상 기업이 12.078로 약 1.6배 차이가 있었고, 이는 자산 기준으로 약 5 배의 규모 격차에 해당한다. 총자본 로그(lnTO) 역시 부도 기업 9.538, 정상 기업 11.777로 유사한 폭의 차이를 보였다. 즉 부도 기업은 자산 · 자본 모두에서 상대적으로 소규모이며, 자본 기반이 충분히 축적되지 못한 상태에 머무르고 있음을 의미한다.

4.2. 로지스틱 회귀를 위한 변수 선택

머신러닝 모형의 경우 총 43개의 설명변수를 모두 사용하지만, 모든 설명변수를 사용하여 로지스틱 회귀 모형을 추정하면 예측력이 매우 낮은 것으로 나타났다. 따라서 사전 테스트(pre-test)를 거쳐 설명변수를 선별하고 그 설명변수들의 최적 조합을 찾는 별도의 과정을 거쳐 로지스틱 회귀 모형을 추정하였다.

우선 설명변수를 하나씩만 사용하여 로지스틱 회귀 모형을 추정하고, t 검정을 통해 5% 유의수준에서 통계적으로 유의하지 않은 설명변수를 제외하였다. 이를 통해 SLTA, CLCA, TLEQ, CLTL, FFOEQ, SLEQ, SLFA, FATO 등 8개의 변수가 제외되었다. 다음으로 t 값이 높은 순서대로 변수를 정렬한 후, 각 변수에 대해 분산 팽창 계수(variance inflation factor)가 10 이상인 강한 다중공선성을 지닌 변수들을 제외하였다. 이를 통해 WCTA, EBITTA, NITA, NISL, lnTA, lnTOI, lnTL, TLTA, CAR, CLTA, CATA, NCTA, EBTA, lnTO, FATA, ROE, OPM, OPTA, CFOTA 등 19개 변수가 제외되었다.

이러한 두 단계를 거쳐 SLLL, FFOTA, CASHTA, lnOP, CACL, RETA, BWTA, lnCE, lnES, OPTL, INSL, FUIL, WCTO, CP, CLI, GDP 등 16개 변수가 남게 되었는데, 단계적 변수 선정(stepwise variable selection)을 이용하여 이 변수들 중 최적 조합을 선정하였다. 이는 변수를 추가하고 제거하면서 최적 변수 조합을 찾는 것인데, AIC 기준을 적용하였다.

4.3. 예측 결과

본 연구에서는 상장기업을 대상으로 부도예측 모형의 성능을 비교 · 분석하기 위해 원자료(raw data)와 오버샘플링 기법인 SMOTE 및 ADASYN을 적용한 세 가지 데이터 세트를 구성하였다. 각 데이터 세트에 대해 로지스틱 모형, 랜덤 포레스트, XGBoost, 심층신경망 기법을 적용하여 총 12개의 데이터 세트와 예측 모형의 조합을 구축하였다. <표4>는 각 데이터 세트 및 예측 모형의 예측치를 이용하여 Accuracy(정확도), Precision(정밀도), Recall(재현율), Specificity(특이도), F1 Score(F1 점수), AUC(곡선하면적) 등의 예측력 평가지표를 계산한 결과를 제시한다.

모형 간 예측 결과는 AUC와 F1 Score를 중심으로 비교하였다. 3.4절에서 설명한 것처럼 AUC는 전체적인 예측 성능을 나타내는 대표적인 지표로, 양성 클래스와 음성 클래스 간의 분류 능력을 종합적으로 반영하며, 값이 1에 가까울수록 뛰어난 예측력을 의미한다. F1 Score는 Precision과 Recall의 조화평균으로, 특히 데이터가 불균형한 경우 모형의 실질적인 분류 성능을 보다 정교하게 반영한다. Precision 또는 Recall 중 하나에 과도하게 편중된 모형은 F1 Score가 낮게 나타난다.

원자료를 기반으로 분석한 결과, 랜덤 포레스트는 AUC가 0.980 그리고 F1 Score가 0.620를 나타내 가장 우수한 예측력을 보였고, Precision이 0.803 그리고 Specificity가 0.997인 것으로 나타나 높은 정분류 능력을 확인할 수 있었다. XGBoost 또한 AUC 0.979 그리고 F1 Score 0.626으로 비슷하게 예측력이 우수한 것으로 나타났다. 그러나, 대부분의 모형에서 Recall이 상대적으로 낮게 나타났는데, 이는 원자료에서 부도기업과 정상기업 간의 표본 불균형이 심하기 때문이다. 특히 로지스틱 회귀 모형의 경우 AUC가 0.660, F1 Score가 0.435로 다른 모형에 비해 전반적으로 낮은 성능을 보였다. Precision이 0.667로 비교적 양호하였으나, Recall은 0.323으로 가장 낮았으며, 이는 실제 부도기업을 제대로 찾아내지 못하는 한계를 나타낸다. Specificity는 0.996으로 높아 정상기업에 대한 오분류는 거의 없었지만, 부도 기업을 제대로 분류하지 못하는 문제가 컸다. 이는 소수 클래스인 부도기업을 거의 탐지하지 못하였다는 점에서 실무적으로 활용하기에는 제한이 있다는 것을 의미한다.

	Accuracy	Precision	Recall	Sepcificity	F1 Score	AUC
Raw data						
LG	0.982	0.667	0.323	0.996	0.435	0.660
RF	0.987	0.803	0.505	0.997	0.620	0.980
XGB	0.984	0.651	0.602	0.993	0.626	0.979
DNN	0.984	0.753	0.376	0.997	0.502	0.961
SMOTE						
LG	0.920	0.195	0.860	0.921	0.318	0.891
RF	0.981	0.547	0.591	0.989	0.568	0.976
XGB	0.985	0.661	0.608	0.993	0.633	0.974
DNN	0.954	0.277	0.773	0.958	0.408	0.961
ADASYN						
LG	0.876	0.139	0.903	0.876	0.241	0.890
RF	0.980	0.537	0.624	0.988	0.577	0.975
XGB	0.984	0.637	0.575	0.993	0.605	0.972
DNN	0.950	0.277	0.812	0.953	0.413	0.959

표 4: 예측 결과. LG는 로지스틱 회귀 모형, RF는 랜덤 포레스트, XGB는 XGBoost, DNN은 심층신경망 모형을 나타낸다.
Table 4: FORECAST RESULTS.LG, RF, XGB, and DNN refer to logistic regression, random forest, XGBoost, and deep neural network models, respectively.

SMOTE 기법을 적용한 데이터 세트의 경우, 전반적으로 Recall이 큰 폭으로 향상되어 부도기업 예측 성능이 개선되는 양상을 보였다. 로지스틱 회귀 모형의 경우 Recall이 0.860로 상승하며 높은 비율로 부도기업을 예측하였으나, Precision은 0.195로 크게 하락하였다. 이로 인해 F1 Score는 0.318로 낮게 나타났는데, 이는 정상기업을 과도하게 부도기업으로 오분류하는 경향이 있음을 의미한다. 반면 XGBoost는 Precision이 0.661 그리고 Recall이 0.608로 균형 있는 예측 성능을 보였으며, AUC이 0.974 그리고 F1 Score가 0.633로 모형 중 가장 안정적으로 우수한 예측력을 나타냈다. 랜덤 포레스트 또한 이와 유사하게 예측력이 우수한 것으로 나타났다. DNN 역시 Recall이 0.773로 상승하였으나, Precision이 0.277로 매우 낮아 F1 Score는 0.408에 불과하였다.

ADASYN 기법을 적용한 데이터 세트의 경우에도 유사한 결과가 나타났다. 전반적으로 Recall이 높아지고, 일부 모형에서는 Precision 감소 폭이 상대적으로 완화되었다. 특히 로지스틱 회귀분석은 Recall이 0.903로 가장 높게 나

타났으며, 이는 대부분의 부도기업을 정확하게 예측하였음을 의미한다. 그러나 Precision은 0.139로 매우 낮아 F1 Score는 0.241에 그쳤다. 또한 Specificity 역시 0.876로 상대적으로 낮게 나타났는데, 이러한 결과들은 부도기업을 부도기업으로 잘 예측했지만 정상기업을 부도기업으로 잘못 예측하는 오분류가 다수 발생했음을 의미한다.

XGBoost는 ADASYN을 적용한 경우에도 AUC가 0.972 그리고 F1 Score가 0.605로 우수한 예측 성능을 보였으며, Precision과 Recall 모두 안정적인 수준을 나타냈다. DNN도 Recall이 0.812로 향상되며 F1 Score가 0.413으로 개선되었고, AUC는 0.959로 높은 수준을 보였다. ADASYN 기법은 특히 DNN과 XGBoost와의 결합에서 예측 성능을 개선하는 데 기여하였으며, 불균형 문제 해결에 효과적으로 작용한 것으로 판단된다.

상장기업을 대상으로 수행한 본 연구의 예측 결과를 정리하면 다음과 같다. 첫째, 부도예측 모형의 성능은 사용된 불균형 데이터 처리와 모형의 조합에 따라 유의미한 차이를 보였다. 특히, 소수 클래스의 정보 손실을 완화하기 위해 적용된 SMOTE 및 ADASYN과 같은 오버샘플링 기법은 전통적인 원자료에 비해 모든 모형에서 Recall 향상에 기여하였다.

둘째, 모형별 예측력을 비교한 결과, SMOTE를 적용한 XGBoost 모형이 전체 모형 중 가장 뛰어난 예측력을 보였으며, Precision과 Recall 간의 균형도 우수하였다. 그리고 ADASYN을 적용한 경우에도 XGBoost 모형은 비슷하게 우수한 예측력을 보였다. 또한 랜덤 포레스트도 상대적으로 우수한 예측력을 보였는데, 원자료를 적용한 경우에 비해 SMOTE 또는 ADASYN을 적용할 경우 Recall이 다소 향상되지만 Precision이 상대적으로 크게 감소하여 결과적으로 F1 Score가 오히려 줄어드는 것으로 나타났다.

셋째, 로지스틱 회귀 모형은 일부 오버샘플링 기법과 결합 시 높은 Recall을 보이기도 하였으나, Precision이 크게 감소하여 정상기업에 대한 오분류가 빈번하게 발생하는 한계가 있었다. 특히, ADASYN을 적용한 로지스틱 회귀의 경우 부도기업을 거의 완벽히 예측하면서도 정상기업을 부도기업으로 잘못 분류하는 비율이 매우 높아 실무적으로 활용하기에는 제한적임을 보였다. 머신러닝 모형이 설명변수를 모두 사용한 것과 달리 4.2절에서 설명한 것처럼 로지스틱 회귀 모형을 위해 설명변수들의 최적 조합을 찾는 별도의 과정을 거쳤음에도 불구하고 로지스틱 회귀 모형의 예측력은 낮은 것으로 나타난 점에 유의해야 한다.

5. 결론

본 연구는 상장기업의 재무 및 거시경제 데이터를 활용하여 통계적 모형 및 다양한 머신러닝 기법의 예측 성능을 비교하였다. 특히 부도기업이 소수에 해당하는 불균형 데이터를 처리하기 위해 SMOTE와 ADASYN과 같은 오버샘플링 기법을 적용하여 모형 성능의 개선 가능성을 검토하였다. 분석 결과, 원자료를 기반으로 할 경우 전통적인 로지스틱 회귀와 심층신경망은 소수 클래스(부도기업) 예측에 한계를 보인 반면, XGBoost와 랜덤 포레스트는 높은 AUC와 F1 Score를 기록하며 상대적으로 우수한 예측력을 보였다. 특히 SMOTE 및 ADASYN을 적용한 경우 전반적으로 Recall이 향상되어 부도기업에 대한 탐지 성능이 개선되었으며, XGBoost는 가장 안정적이고 균형 잡힌 성능을 나타냈다. 반면, 로지스틱 회귀는 오버샘플링 기법과 결합할 경우 높은 Recall을 달성하기도 했지만, Precision이 과도하게 낮아져 실무적 활용에는 한계가 있었다.

결론적으로, 기업부도 예측에서는 불균형 데이터 문제를 적절히 보정하고, 복잡한 패턴을 효과적으로 학습할 수 있는 XGBoost와 같은 머신러닝 기반 모형이 실무적으로도 높은 활용 가능성을 지니는 것으로 판단된다. 향후 연구에서는 비재무 정보나 시계열 특성을 반영한 확장된 설명변수 구성을 통해 예측력을 한층 더 향상시키는 방안을 모색할 수 있을 것이다. 특히 후속 연구에서는 업종별 산업의 특성을 고려한 설명변수 선택 및 모형 설정으로 업종별로 부도기업 예측 정확도를 개선하는 방안을 검토해 볼 필요가 있을 것이다.

참고문헌

송현준 · 박도준 · 이준기 (2021). “머신러닝을 이용한 외감기업 및 비외감기업의 부도예측에 관한 연구,” *한국IT정책경영학회 논문지* 13, 2521-2527.

박종원 · 안성만 (2014). “재무비율을 이용한 부도예측에 대한 연구: 한국의 외부감사 대상기업을 대상으로,” *경영학연구* 43, 639-666.

박근우 · 정인경 (2019). “이분형 자료의 분류문제에서 불균형을 다루기 위한 표본재추출 방법 비교,” *응용통계연구* 32, 349-374.

이인로 · 김동철 (2015). “회계정보와 시장정보를 이용한 부도예측모형의 평가연구,” *재무연구* 28, 626-666.

오세경 · 최정원 · 장재원 (2017). “빅데이터를 이용한 딥러닝 기반의 기업 부도예측 연구,” KIF Working Paper 8, 1-113.

Altman, E. I. (1968). “Financial ratios, discriminant analysis and the prediction of corporate bankruptcy,” *Journal of Finance* 23, 589–609.

Breiman, L. (2001). “Random forests,” *Machine Learning* 45, 5–32.

Beaver, W. H. (1966). “Financial ratios as predictors of failure,” *Journal of Accounting Research* 4, 71–111.

Campbell, J. Y., Hilscher, J., and Szilagyi, J. (2008). “In search of distress risk,” in *Journal of Finance* 63, 2899–2939.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). “SMOTE: Synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research* 16, 321–357.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 1322–1328.

Ohlson, J. A. (2010). “Financial ratios and the probabilistic prediction of bankruptcy,” *Journal of Accounting Research* 18, 109–131.

Chen, T., and Guestrin, C. (2016). “XGBoost: A scalable tree boosting system,” In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

Zmijewski, M. E. (1984). “Methodological issues related to the estimation of financial distress prediction models,” *Journal of Accounting Research* 22, 59–82.

A. 부도예측 설명변수

본 연구의 부도예측에 사용한 설명변수의 목록은 아래 표에 제시되어 있다.

설명	변수명	설명	변수명
수익성 (8개)		건전성 (8개)	
영업이익/총자산	EBITTA	총부채/총자산	TLTA
영업이익/매출액	OPM	총부채/총자본	TLEQ
영업이익/총자산	OPTA	총자본/총자산	CAR
당기순이익/총자산	NITA	차입금/총자산	BWTA
당기순이익/총자본	ROE	금융부채/총부채	FUIL
당기순이익/매출액	NISL	금융부채/총자산	FATA
세전순이익/총자산	EBTA	영업활동이익/총부채	OPTL
이익잉여금/총자산	RETA	이자비용/매출액	INSL
유동성 (8개)		현금흐름 (4개)	
(유동자산-유동부채)/총자산	WCTA	영업활동현금흐름/총자산	FFOTA
유동자산/총자산	CATA	영업활동현금흐름/총자본	FFOEQ
유동자산/유동부채	CACL	현금및현금성자산/총자산	CASHTA
유동부채/총자산	CLTA	현금흐름/총자산	CFOTA
유동부채/총부채	CLTL		
유동부채/비유동부채	SLLL	활동성 (5개)	
유동부채/유동자산	CLCA	총자산/매출액	SLTA
비유동자산/총자산	NCTA	매출액/총자본	SLEQ
		매출액/고정자산	SLFA
규모 (7개)		매출액/(유동자산-유동부채)	WCTO
log(자산총계)	lnTA	매출액/유형자산	FATO
log(자본총계)	lnTOI		
log(영업이익)	lnOP	거시경제지표 (3개)	
log(부채총계)	lnTL	시장금리(CP 91일물)	CP
log(현금및현금성자산)	lnCE	선행종합지수	CLI
log(이익잉여금)	lnES	GDP 성장률	GDP
log(매출액)	lnTO		

표 A: 설명변수 목록. 이 표는 부도예측에 사용한 설명변수의 목록이다.

TABLE A: LIST OF EXPLANATORY VARIABLES. This table presents the list of explanatory variables used for bankruptcy prediction.