Journal of Korean Society of Industrial and Systems Engineering
https://doi.org/10.11627/jksie.2022.45.2.065

Vol. 45, No. 2 : 65-76, June 2022
ISSN : 2005-0461 / EISSN : 2287-7975

A Methodology for Bankruptcy Prediction in Imbalanced Datasets using eXplainable AI

Sun-Woo Heo* · Dong Hyun Baek**†

*Department of Management Consulting, Graduate School of Hanyang University
**Division of Business Administration, Hanyang University

데이터 불균형을 고려한 설명 가능한 인공지능 기반 기업부도예측 방법론 연구

허선우* · 백동현**†

*한양대학교 일반대학원 경영컨설팅학과
**한양대학교 경상대학 경영학부

Recently, not only traditional statistical techniques but also machine learning algorithms have been used to make more accurate bankruptcy predictions. But the insolvency rate of companies dealing with financial institutions is very low, resulting in a data imbalance problem. In particular, since data imbalance negatively affects the performance of artificial intelligence models, it is necessary to first perform the data imbalance process. In additional, as artificial intelligence algorithms are advanced for precise decision-making, regulatory pressure related to securing transparency of Artificial Intelligence models is gradually increasing, such as mandating the installation of explanation functions for Artificial Intelligence models. Therefore, this study aims to present guidelines for eXplainable Artificial Intelligence-based corporate bankruptcy prediction methodology applying SMOTE techniques and LIME algorithms to solve a data imbalance problem and model transparency problem in predicting corporate bankruptcy.

The implications of this study are as follows. First, it was confirmed that SMOTE can effectively solve the data imbalance issue, a problem that can be easily overlooked in predicting corporate bankruptcy. Second, through the LIME algorithm, the basis for predicting bankruptcy of the machine learning model was visualized, and derive improvement priorities of financial variables that increase the possibility of bankruptcy of companies. Third, the scope of application of the algorithm in future research was expanded by confirming the possibility of using SMOTE and LIME through case application.

Keywords : Bankruptcy Prediction, Data Imbalance, eXplainable AI, SMOTE, LIME, Machine Learning

Received 29 May 2022; Finally Revised 23 June 2022;
Accepted 25 June 2022
† Corresponding Author : estarbaek@hanyang.ac.kr

1. 서 론

금융기관에게 있어 기업의 신용위험을 평가하는 것은 매우 중요한 문제이다. 이에 따라 금융기관들은 기업의 신용위험을 평가하기 위한 방법으로 다양한 기업부도예측 기법을 활용하여 기업의 신용위험을 평가하고 있다. 그러나, 금융기관과 거래하는 기업들의 부도율은 매우 낮아서, 부도 사례보다 정상 사례의 빈도가 월등히 높은, 데이터 불균형 문제가 발생하고 있다. 특히, 데이터 불균형은 인공지능 모델의 성능에 부정적 영향을 미치므로 우선적으로 데이터 불균형 처리 과정을 수행할 필요가 있는 상황이다 [15, 18, 35].

한편, 정밀한 의사결정을 위해 인공지능 알고리즘이 고도화 되면서 기능은 알지만 작동원리를 이해할 수 없는, 소위 블랙박스 모델이 등장하였다. 이에 EU에서는 2021년 생체 인식, 신용 평가 등 '고위험 AI'에 대한 규제를 포함한 인공지능법(Artificial Intelligence Act) 초안을 발표하며, 고위험 AI에 대한 설명 책임을 명시하였다. 또한 美연방거래위원회(FTC)에서는 지난 2020년 AI 및 알고리즘 활용(Using artificial intelligence and algorithm)이라는 보고서를 발표하며, AI 설계 5대 지침에 설명 가능성을 포함하는 등 AI 모델의 투명성 확보와 관련된 규제압력과 관심이 점차 증가하고 있는 상황이다.

4차 산업혁명 시대를 맞이해 머신러닝을 비롯한 인공지능 기술들이 전 산업으로 확대되는 가운데, 기업부도예측은 데이터 불균형이라는 내부적 문제점 극복과 모델 투명성 확보라는 외부적 요구에 응할 필요가 존재한다. 그러나, 국내외를 막론하고 데이터 불균형 문제와 모델의 투명성 확보 문제를 모두 고려한 방법론 및 사례적용 연구는 찾아보기 어려운 상황이다. Dastile et al.[9]에 따르면 2010년부터 2018년까지 진행된 74개의 머신러닝 기반 기업부도예측 및 기업신용평가 관련 연구 중 데이터 불균형을 처리하지 않은 연구는 전체의 82%, 모델의 설명 기능을 탑재하지 않은 연구는 전체의 92%인 것으로 나타났다.

따라서 본 연구는 기 연구된 향상된 데이터 분석기법들을 조합해 데이터 불균형과 모델 투명성 문제를 모두 고려한 기업부도예측 5단계 방법론을 제안하고, 그 방법론을 사례에 적용하여 타당성을 검증하고자 한다.

2. 이론적 배경

2.1 기업부도예측 선행연구

초기 부도예측 모델은 1960년대부터 본격적으로 연구가 진행되었다. 1966년, Beaver는 단변량 판별 분석을 활용해 기업별 재무 비율의 평균 차이를 중심으로 재무 변수의 기업부도예측 능력을 평가하였다[4]. 그러나 기업의 부도는 어떤 한 가지 요인에 의해 결정되는 것이 아니므로 다양한 변수를 하나의 모델에 통합하는 다변량 판별 분석의 필요성이 대두되었고, 1968년 Altman의 연구에서 부도 데이터를 다변량 판별 분석에 적용하면서 기업부도예측에 관한 연구가 본격적으로 진행되기 시작했다[2]. Altman은 1946년부터 1965년까지의 기업 중 부도 기업과 건전 기업을 각각 33개씩 추출하여 다변량 판별 분석 모델을 통해 예측모델을 구성하였으며, 기존 연구를 통해 영향력이 확인 된 22개의 재무변수 중에서 5개의 재무 변수를 예측모델에 사용하였다[2].

Altman의 다변량 판별 분석의 경우, 독립 변수의 분포가 다변량 정규분포를 따라야하며, 집단 간 분산과 공분산 행렬이 동일하다는 가정을 만족시켜야만 하는 통계적 한계점이 존재했다[17]. 이후, 이러한 한계점을 보완하기 위하여 후속 연구에서는 이항 반응 모델을 기업부도예측에 적용하는 연구가 실시되었다. 기업부도예측관련 이항 반응 모델 연구로는 1984년 Zmijewski[39]의 프로빗 분석과 1980년 Ohlson[24]의 로지스틱 회귀 분석이 대표적이다. 프로빗 분석과 로지스틱 회귀분석은 다변량 판별 분석과 달리 복잡한 통계적 가정을 만족시킬 필요 없이 사용할 수 있으며, 기업부도 원인의 확률적 해석이 가능하다는 장점이 존재한다. 그러나 예측 성능은 그다지 높지 않다는 분명한 한계점을 가지고 있었다.

<Table 1> Preliminary Studies on Bankruptcy Predictions Using Statistical Techniques

Research	Model
Beaver[4]	Univariable discriminant analysis
Altman[2]	Multi-variable discriminant analysis
Edmister[11]	Multiple regression analysis
Pinches et al.[27]	Principal components analysis
Ohlson[24]	Logistic regression(LR)
Zmijewski[39]	Probit regression analysis

1980년대 후반부터는 모수적 방법을 사용한 연구에서 발전해 인공신경망(ANN, Artificial Neural Network), 의사결정 나무(Decision Tree), SVM(Support Vector Machine)과 같은 비모수적 방법을 사용한 연구가 진행되었다.

특히, 가장 높은 예측력을 보였던 인공신경망이 기업부도예측 연구에서 가장 많이 활용되었다. 그 중에서 최초로 부도예측연구에 인공신경망을 적용한 Odom and Sharda[23]의 연구가 가장 대표적이다. 이 연구에서는 판별 분석과 인공신경망의 정확도를 비교하였고, 인공신경망이 기존의 판별 분석보다 더 높은 예측 성능을 낸다는 것을 보여주었다.

이후의 연구에서도 기업부도예측을 위한 다양한 연구가 진행 되었으나, 결과적으로 가장 높은 예측력을 보였던 인공신경망과 전통적인 통계 기법의 예측 정확도를 비교하는 연구로 발전되었다. Tam and Kiang[34]은 은행의 부도여부를 인공신경망과 의사결정나무, 판별분석, 로지스틱 회귀분석, k-NN(k-Nearest Neighbor)을 이용해 비교분석한 결과, 인공신경망이 다른 기법과 비교하여 우수하다는 결론을 내렸다. Wilson and Sharda[36] 또한 인공신경망의 성능이 사례기반 추론과 판별 분석보다 우수함을 확인하였다. Jo and Han[13]의 연구에서도 선형판별분석, 이차판별분석, k-NN에 비해 인공신경망의 성능이 더 뛰어남을 보였다. 이처럼, 인공신경망의 우수한 예측 정확도는 많은 실증연구를 통해 증명되었지만, 인공신경망은 과적합 문제가 발생 할 수 있음이 Altman et al.[3]의 연구에서 확인되었다. 또한, 연구자가 임의로 설계해주어야 하는 하이퍼파라미터가 많기 때문에 모델을 설계하는 것에 비교적 많은 시간이 소요된다는 문제가 있었다.

이러한 인공신경망의 문제점이 제기된 이후, SVM을 활용해 기업의 부도를 예측하고자 하는 연구가 이어졌다. SVM은 인공신경망과 달리 적은 수의 샘플로도 예측이 가능하고 연구자가 통제해야 하는 하이퍼파라미터의 수가 비교적 적다는 장점이 존재해 기업부도예측 연구에 활발히 쓰이게 되었다. Shin et al.[32]은 제조업 분야 기업의 부도 예측에 SVM과 인공신경망을 적용해 비교한 결과, SVM이 인공신경망에 비해 성능이 더 우수하다는 결론을 내렸다. Wu et al.[37]은 SVM의 하이퍼파라미터인 C와 σ^2를 유전자 알고리즘을 통해 최적화하고 판별분석, 로지스틱 회귀분석, 프로빗 분석, 인공신경망과 비교분석을 진행하였다. 비교결과, 정확도 면에서 SVM이 가장 우수한 모델임을 확인하였다.

<Table 2> Preliminary Studies on Bankruptcy Predictions Using Machine Learning Techniques

Research	Model
Odom and Sharda[23]	ANN, Discriminant Analysis
Tam and Kiang[34]	ANN, Decision Tree, Discriminant Analysis, LR and k-NN
Jo and Han[13]	ANN, Linear Discriminant Analysis, Quadratic Discriminant Analysis, and k-NN
Shin et al.[32]	SVM, ANN
Wu et al.[35]	SVM, LR, Probit Analysis and ANN

2.2 랜덤포레스트(Random Forest)

랜덤포레스트는 2001년 Breiman[5] 소개한 방법으로, 다수 의사결정나무의 결과를 결합해 하나의 결과를 도출하는 앙상블(Ensemble) 기법이다. 다른 앙상블 모델과의 큰 차이점은 부트스트랩된 표본을 다수 생성함으로서 임의성이 도입된다는 측면에서 의사결정나무 간의 상관관계를 낮춰 예측오차를 낮춰준다는 장점을 갖고 있다[38]. 또한 의사결정나무의 수가 증가할수록 예측오차가 줄어들며, 의사결정나무의 수가 많더라도 과적합 하지 않는다는 장점이 있다[5].

2.3 XGBoost(eXtreme Gradient Boosting)

XGBoost는 여러 개의 분류 및 회귀나무(CART)를 결합해 error 값을 낮추는 부스팅 기법을 활용한 의사결정나무 계열의 알고리즘이다[7]. XGBoost는 학습 손실을 최소화하면서 과적합을 방지하기 위해 나무의 복잡도를 통제하는 방식으로 최적화된 모델을 생성하며, XGBoost의 목적함수는 식 (1)과 같다. k는 나무의 수이고, Ω는 의사결정나무의 복잡도에 영향을 줄 수 있는 모든 상황을 포함한다. XGBoost는 빠른 처리 속도와 뛰어난 성능으로 다수의 데이터 분석 경진대회 우승 알고리즘으로 활용되어온 것으로 유명하다[9]

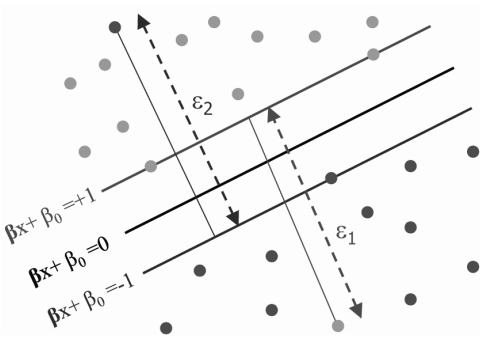
$$L^{(t)} = \sum_{k=1}^{n} l(y_k, \hat{y_k}^{(t-1)} + \phi(x_k)) + \Omega(\phi_t) \quad (1)$$

2.4 서포트벡터머신(Support Vector Machine)

SVM은 V. Vapnik[8]에 의해 제안된 모델로, 고차원 벡터 공간 내에서 클래스를 분리하는 최적의 초평면(Hyperplane)을 찾는 이진 분류모델이다. 선형 SVM의 경우 음극 초평면과 양극 초평면 사이의 마진은 식 (2)로 표현되며, 마진을 최대화하는 데 중점을 두어 최적의 초평면을 결정한다. 마지막으로 식 (3)에 따라 클래스를 분리하게 된다. 비선형 경우 커널 트릭[31]을 사용하여 고차원 벡터 공간에서의 최적의 초평면을 결정한다. <Figure 1>은 2차원 공간에서 최적의 초평면을 찾는 과정을 표현한 것이다.

$$\| a \| = \sqrt{\sum_{i=1}^{m} a^{2_i}} \quad (2)$$

$$y = \begin{cases} +1, & \text{if } \beta + a^T x \geq +1 \\ -1, & \text{if } \beta + a^T x \leq -1 \end{cases} \quad (3)$$

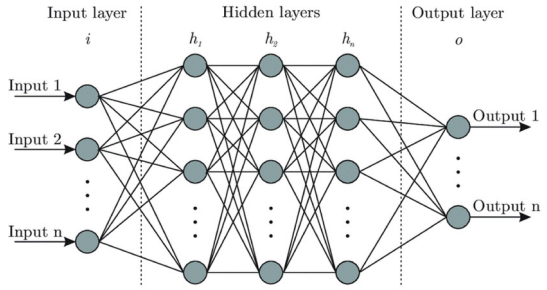


<Figure 1> Support Vector Machine

2.5 인공신경망(Artificial Neural Network)

인공신경망은 생물학의 신경망에서 영감을 얻은 학습 알고리즘이며, 일반적으로 입력층, 은닉층, 출력층으로 구성되어 있다. 인공신경망은 인공 뉴런(노드)의 결합을 통해 네트워크를 형성한 모델이 학습과정을 거쳐 문제 해결을 위한 최적의 가중치를 찾는 모델 전반을 가리킨다.

최초의 인공신경망 모델은 1943년 McCulloch and Pitts [20]에 의해 제안되었고, 이후 1958년 Rosenblatt[29]에 의해 지도 학습이 가능한 퍼셉트론 모델이 개발되었다. 1969년에는 Marvin and Seymour[19]에 의해 단층퍼셉트론이 XOR 문제를 해결하지 못한다는 것이 밝혀져 인공신경망에 대한 연구는 암흑기에 접어들었고, 1986년 오차 역전파 알고리즘[30]이 개발되면서 다시 발전하기 시작하였다. <Figure 2>는 인공신경망의 구조를 도식화 한 것이다.



<Figure 2> Artificial Neural Network

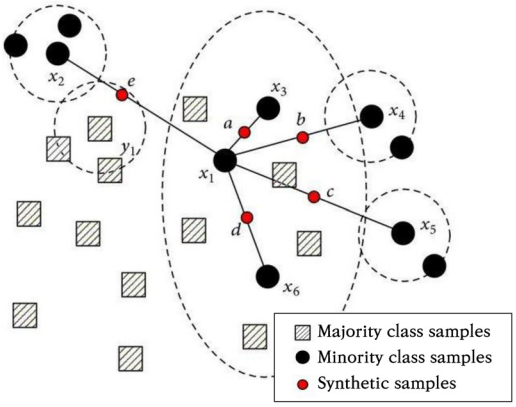
2.6 SMOTE(Synthetic Minority Over sampling Technique)

데이터 불균형이란 하나의 클래스(Class)에 해당하는 데이터의 수가 다른 클래스에 해당하는 데이터의 수에 비하여 현저히 많거나 적은 경우를 말한다[26]. 이때 더 많은 수의 데이터를 보유하고 있는 클래스를 다수 클래스(Majority class)라고 하며, 더 적은 데이터를 보유하고 있는 클래스를 소수 클래스(Minority class)라고 한다[33]. 만일 데이터 불균형 상태가 해결되지 않은 데이터셋에 머신러닝을 적용할 경우, 예측 모형은 다수 클래스에 편향된 학습을 하게 된다[22]. 이를 데이터 불균형 문제라고 하며 데이터 불균형은 결과적으로 모형의 성능에 부정적 영향을 미치게 된다[15, 18, 35]. 이러한 데이터 불균형을 해소하기 위한 대표적인 방법으로는 SMOTE가 있다.

SMOTE 기법은 부트스트래핑(Bootstrapping)과 k-NN을 통해 데이터를 생성하는 기법이다. 이 기법은 소수 클래스에 속하는 특정 데이터에 대해 소수 클래스 데이터로 구성된 K개의 최근접 이웃을 찾아 그 이웃과 선형의 연결 구조를 만들고 그 연결 구조 사이에 새로운 데이터를 생성한다[6]. 따라서, 재표본 추출을 통한 오버샘플링보다 과적합 가능성이 적으며, 언더샘플링처럼 데이터 손실의 우려도 없다는 장점이 존재한다.

이때, SMOTE를 수식으로 나타내면 수식 4와 같다. 이때 i와 j는 소수 클래스 데이터의 인덱스이며 λ는 0과 1 사이의 무작위 숫자이다. <Figure 3>은 SMOTE를 통해 데이터를 생성하는 과정을 표현한 것이다.

$$X_{create} = X_i + \lambda(X_j - X_i), \lambda \in [0,1] \quad (4)$$



<Figure 3> Data Generation Process of SMOTE

2.7 LIME

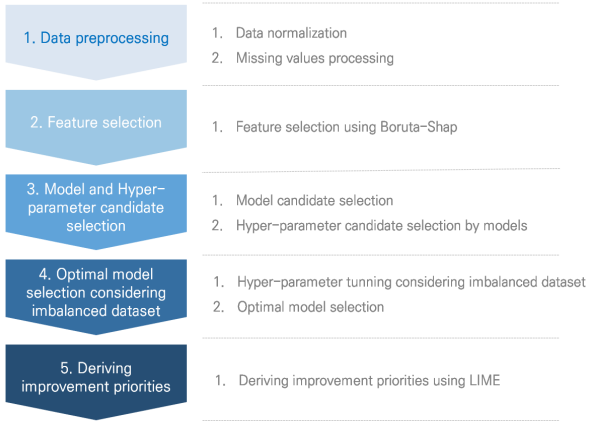
LIME(Local Interpretable Model-agnostic Explanation)은 Ribeiro et al.[28]이 제안한 설명가능한 인공지능(XAI, eXplainable AI) 기법으로서 설명할 수 없는 인공지능 모델이 현재 데이터의 어떤 영역에 집중하여 분석했고 어떤 영역을 분류 근거로 사용했는지 알려주는 기법이다[1]. 모델 학습 방법과 관계없이 적용할 수 있다는 장점이 있어 이미 사용하는 학습모델이 존재한다면 LIME을 적용하여 인공지능 모델을 설명 가능하게 변환 할 수 있다. LIME 알고리즘의 작동방식은 입력값 x에 대해 일정 범위 π_x내에서 블랙박스 모델 f과 유사한 결과를 보이는 설명 모델 g를 찾는 것이라고 요약 할 수 있다. 이를 수식으로 나타내면 식 (5)와 같다.

$$Exp(x) = \underset{g \in G}{argmin} L(f, g, \pi_x) + \Omega(g) \quad (5)$$

3. 기업부도예측 방법론

본 방법론의 목적은 SMOTE기법을 활용하여 데이터 불

균형을 해결하고 XAI 기법인 LIME을 통해 분류결과를 해석하는 것이다. 본 연구에서 제안하는 방법론의 수행단계는 <Figure 4>와 같이 5단계로 구성되어 있으며, 데이터 전처리, 특성 선택, 모델 및 하이퍼파라미터 후보군 선정, 데이터 불균형을 고려한 최적 분류 모델 결정, 개선 우선순위 도출 순으로 진행된다.



<Figure 4> XAI Methodology for Bankruptcy Prediction in Imbalanced Datasets

3.1 데이터 전처리

첫 번째 단계인 데이터 전처리 단계에서는 데이터 정규화와 결측치 처리 단계를 수행한다. 전처리 단계에서는 데이터 특성에 대해 고려할 필요가 있다. 본 연구에서 사용되는 데이터는 대부분 중소기업 데이터로 중소기업 재무데이터의 경우 기록이 불완전해 이상치, 결측치가 존재한다는 특성이 있다[40]. 따라서, 데이터 정규화 방식에는 이상치의 영향을 덜 받는 Z-score 정규화 방식을 사용하며, 데이터 결측치 처리에는 데이터 불균형 상황임을 고려하여 소수클래스 데이터의 소실을 방지할 수 있는 k-NN 대체법을 사용한다. k-NN 대체법은 결측이 나타나는 표본의 미결측 변수를 중심으로 유클리드 거리가 가장 가까운 k개의 다른 표본 데이터를 탐색한 뒤, 탐색 된 k개의 표본의 평균(연속형) 또는 최빈값(범주형)으로 각 변수를 대체하는 기법이다.

3.2 특성 선택

두 번째 단계인 특성 선택 단계는 고차원의 데이터 셋을 분석하기 위해서 반드시 필요한 과정이다[25]. 일반적으로 많은 머신러닝 알고리즘들은 너무 많은 수의 특성(변수)을 사용하게 되면 예측 정확도가 감소하며[14], 최고의 결과를 나타내는 최소의 특성 집합을 고르는 것이 실용적인 관점에서 적합하다고 알려져 있다[21]. 따라서 머신러닝 모형을 설계할 때 최적의 특성 집합을 선택하는 것은 모형의 정확도에 큰 영향을 주게 된다. 이에 따라 본 단계에서는 기업부도와 관련이 적은 특성을 제거하여 머신러닝의 효율성을 높이기 위해 특성 선택 알고리즘인 Boruta-Shap(Shapley Additive exPlanations)[16]을 활용하여 특성 선택을 수행한다. <Table 3>은 Boruta-Shap 알고리즘의 프로세스를 표현한 것이다.

<Table 3> Boruta-Shap Algorithm

Boruta-Shap[16]	
Step 1	Start by creating new copies of all the features in the data set and name them shadow + feature_name, shuffle these newly added features to remove their correlations with the response variable.
Step 2	Run a random forest classifier on the extended data with the random shadow features included. Then rank the features using a feature importance metric the original algorithm used permutation importance as it's metric of choice.
Step 3	Create a threshold using the maximum importance score from the shadow features. Then assign a hit to any feature that had exceeded this threshold.
Step 4	For every unassigned feature preform a two sided T-test of equality
Step 5	Attributes which have an importance significantly lower than the threshold are deemed 'unimportant' and are removed them from process. Deem the attributes which have importance significantly higher than than the threshold as 'important'.
Step 6	Remove all shadow attributes and repeat the procedure until an importance has been assigned for each feature, or the algorithm has reached the previously set limit of the random forest runs.

3.3 모델 및 하이퍼파라미터 후보군 선정

세 번째 단계인 모델 및 하이퍼파라미터 후보군 선정 단계는 모델 최적화에 사용할 하이퍼파라미터를 선정하는 단계이다. 하이퍼파라미터는 모델의 성능에 직접적인 영향을 주기 때문에 최종 모델 결정에 있어서 매우 중요하다. 각 하이퍼파라미터는 최적화에 필요한 계산 비용을 고려하여 연구자 임의로 모델 성능에 영향이 높은 몇 가지만을 선택해 다음 4단계 데이터 불균형을 고려한 분류모델 결정 단계에서 최적화를 수행한다.

3.4 데이터 불균형을 고려한 분류모델 결정

네 번째 단계인 데이터 불균형을 고려한 분류모델 결정 단계는 데이터의 불균형 상황을 고려하여 머신러닝 모델별 최적의 하이퍼파라미터를 결정하는 단계이다. 본 단계의 수행은 3단계로 요약된다.

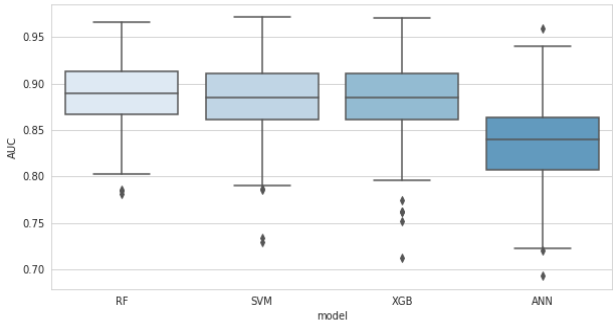
첫 번째, 각 모델의 하이퍼파라미터 튜닝을 위해 5-fold CV 방식으로 데이터를 분할하고, 각 Training Fold에 SMOTE를 적용하여 데이터 밸런싱을 수행한다.

두 번째, 직전 단계에서 선정된 하이퍼파라미터들에 베이지안 최적화, 그리드 서치 기법 등을 적용하여 모델별 최적 하이퍼파라미터 값을 탐색한다.

세 번째, 최종 탐색된 하이퍼파라미터 값을 각 모델에 적용하여 균형화된 데이터셋을 연구자가 정한 횟수만큼 학습, 예측한다. 학습 및 예측된 결과는 <Table 4>와 <Figure 5>같이 표와 상자그림으로 기록하고, 기록된 결과를 토대로 최적의 부도예측 모델을 결정한다.

<Table 4> Example of Model Evaluation

	RF	SVM	XGB	ANN
Max	0.961	0.962	0.962	0.952
Median	0.944	0.943	0.943	0.924
Mean	0.945	0.944	0.943	0.923
Min	0.922	0.924	0.918	0.801



<Figure 5> Example of Model Evaluation(Boxplot)

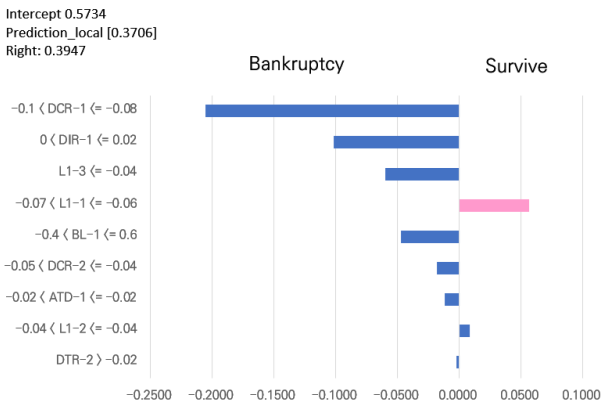
3.5 개선 우선순위 도출

다섯 번째 단계인 개선 우선순위 도출 단계는 부도예측 결과에 XAI의 일종인 LIME 알고리즘을 적용하여 재무변수의 개선 우선순위를 도출하는 단계이다.

본 단계의 프로세스는 크게 두 가지로 구성된다. 첫 번째, 직전 단계에서 결정된 부도예측 모델을 이용해 다시 한 번 기업부도예측을 수행하고 부도예측 결과에 LIME 알고리즘을 적용한다. 두 번째, LIME 알고리즘을 적용해 얻은 변수 민감도를 통해 개별 기업의 재무변수의 개선 우선순위를 도출한다.

LIME을 적용하여 얻을 수 있는 변수 민감도를 나타내는 산출물로는 <Table 5>와 <Figure 6>이 있다. 이중 <Table 5>는 세부결과 예시이며. <Figure 6>은 이를 시각화한 것이다.

세부결과 예시를 통해 확인할 수 있는 변수 민감도는 변수별 가중치와 변수 값을 곱한 값으로서, 변수 민감도의 절대값이 클수록 예측결과에 큰 영향을 주었다고 해석할 수 있다. 한편. 변수 민감도의 총합은 <Figure 6>의 y절편(intercept) 값과 더해져 LIME으로 기존 인공지능 모델을 모사한 생존율(Prediction_local)값이 된다.



<Figure 6> Example of Variable Sensitivity

<Table 5> Example of Variable Sensitivity

Feature(π_x)	Sensitivity (Contribution)	weight	Z-score
DCR-1 <= -0.10	-0.2106	2.1611	-0.0974
0 < DIR-1 <= 0.02	-0.1022	-4.9916	0.0205
-0.07 < L1-1 <= -0.06	0.0541	-0.7759	-0.0697
-0.4 < BL-1 <= 0.6	-0.0387	-0.3507	0.1105
-0.05 < DCR-2 <= -0.04	-0.0365	0.7886	-0.0462
-0.04 < L1-2 <= -0.04	0.0141	-0.3486	-0.0404
-0.02 < ATD-1 <= -0.02	0.0112	-0.5723	-0.0196
-0.02 < DTR-2 <= -0.02	-0.0053	0.3214	-0.0164
-0.04 < L1-3 <= -0.04	-0.0027	0.0658	-0.0408

4. 사례적용

본 연구에서는 XAI 기반 기업부도예측 방법론의 사례 적용을 위해 Zoricak et al.[40]의 연구에서 사용하였던 슬로바키아의 건설분야 중소기업 데이터를 사용하였다[11]. 제공 받은 데이터에서는 법적인 파산 상태와 기업개선 상태에 있는 기업을 부도 기업으로 정의하였다. 부도 기업에 대한 표본은 2014년부터 2016년까지 총 3년간 법적인 파산 상태 또는 개선 상태에 있는 89개를 기업을 활용하였고, 정상기업에 대한 표본은 부도기업과 동일한 기간 동안 법적인 파산 상태 혹은 개선 상태에 놓이지 않은 6,546개 기업을 정상기업으로 활용하였다.

본 연구에서 사용된 재무 변수는 총 63개로, 부도 여부 결정 직전 3개년도(해당년도 연말 기준)에 대한 21개 재무 데이터로 구성되어 있다. 21개 재무 데이터는 크게 활동성, 단기유동성, 수익성, 장기유동성의 4가지 재무 비율 지표로 구성되며, 활동성 지표에는 총자산회전율, 자산회전일, 총수취채권미결일, 재고자산회전일이, 단기유동성 지표에는 현금비율, 당좌비율, 유동비율이, 수익성 지표에는 자산이익률, 자본이익률, 매출이익률, 투자이익률, 인건비 비율, 노동분배율, 노동생산성이, 장기유동성비율에는 부채비율, 부채대자본비율, 레버리지 비율, 총부채상환비율, 채무상환능력계수비율, 자산보상비율, 은행부채비율이 포함되어있다. <Table 6>은 연구에 사용된 변수들과 변수의 약어를 정리한 것이다.

4.1 데이터 전처리

본 단계에서는 정확한 특성 중요도 산정과 변수의 균형 있는 학습을 위해 Z-Score 정규화를 실시하였고, <Table 7>과 같이 변수별로 결측비율을 정리하여 10% 이상의 결측비율이 발생한 13개 변수를 삭제, 10% 미만의 결측이 발생한 나머지 50개 변수에 대해 k-NN대체법을 적용하여 데이터 결측을 처리하였다.

<Table 6> Financial Variables

Financial Indicators	Financial Variables	Abbr.
Activity	Total Asset Turnover	TAT
	Asset Turnover Days	ATD
	Days Total Receivables Outstanding	DTR
	Inventory Turnover Days	ITD
Liquidity	Cash Ratio	L1
	Quick Ratio	L2
	Current Ratio	L3
Profitablity	Return on Assets	ROA
	Return on Equity	ROE
	Return on Sales	ROS
	Return on Investment	ROI
	Labor-to-Revenue Ratio	LRR
	Wages to Added Value Ratio	WAR
	Labor Productivity	LP
Solvency	Debt-to-Assets Ratio	DA
	Debt-to-Equity Ratio	DE
	Financial Leverage	FL
	Debt to Income Ratio	DIR
	Debt Service Coverage Ratio	DCR
	Asset Coverage Ratio	ACR
	Bank Liabilities to Debt Ratio	BL

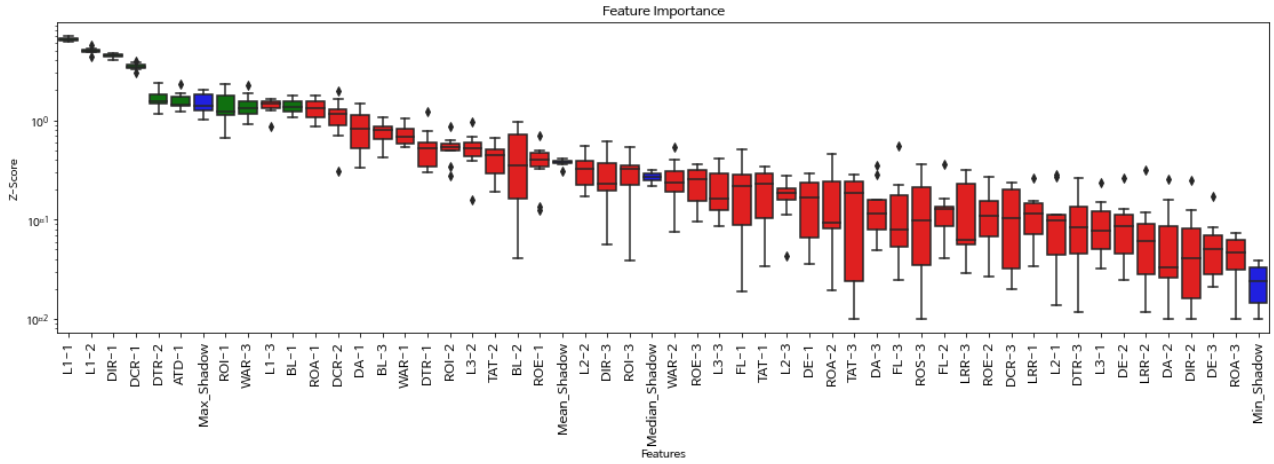
<Table 7> Missing Ratio in Financial Variables

Financial variables	Missing ratio(n = 6,635)			
	-1Y	-2Y	-3Y	MEAN
TAT	3.30%	4.64%	7.91%	5.29%
ATD	9.51%	11.15%	15.30%	11.99%
DTR	3.95%	5.50%	9.31%	6.25%
ITD	44.10%	42.65%	42.65%	43.13%
L1	2.17%	3.20%	6.41%	3.92%
L2	1.99%	3.07%	6.24%	3.77%
L3	2.06%	3.15%	6.35%	3.85%
ROA	0.69%	0.78%	0.80%	0.76%
ROE	0.71%	0.86%	0.84%	0.80%
ROS	23.48%	15.64%	8.82%	15.98%
ROI	0.69%	0.78%	0.80%	0.76%
LRR	3.29%	4.55%	7.73%	5.19%
WAR	1.91%	2.70%	3.87%	2.83%
LP	70.66%	64.40%	59.47%	64.84%
DA	1.67%	2.52%	4.58%	2.92%
DE	1.67%	2.56%	4.58%	2.94%
FL	0.71%	0.86%	0.84%	0.80%
DIR	0.75%	0.83%	0.81%	0.80%
DCR	1.81%	2.91%	5.83%	3.52%
ACR	24.17%	26.72%	30.60%	27.16%
BL	0.75%	0.84%	1.06%	0.88%

4.2 특성 선택

본 단계에서는 기업부도와 관련이 적은 특성을 제거하여 머신러닝의 효율성을 높이기 위해 특성 선택 알고리즘인 Boruta-Shap(Shapley Additive exPlanations)을 활용하였다[16]. Boruta-Shap의 특성 중요도 값을 산정하기 위한 트리 기반 머신러닝 알고리즘으로는 XGBoost가 사용되었다. 특성 선택을 위해 Boruta-Shap에 투입된 변수(특성)는 원본데이터의 63개 변수 중 완전제거법으로 삭제된 변수 13개를 제외한 50개 변수이며, Boruta-Shap 알고리즘을 500회 시행한 결과, 최종적으로 은행부채비율-1, 총부채상환비율-1, 자산회전일-1, 현금비율-1, 현금비율-2, 현금비율-3, 채무상환능력계수비율-1, 채무상환능력계수비율-2, 총수취채권미결일-2 등 9개 변수가 선택되었다. 이에 따라 <Figure 7>은 Boruta-Shap 알고리즘을 시행한 특성 중요도 산정 결과이며, <Table 8>은 선택된 변수를 정리한 것이다.

선택된 변수를 살펴보면 9개 변수 중 직전년도에 해당하는 변수가 절반 이상인 5개로 나타났고, 2년 전 변수가 3개, 3년 전 변수가 1개로 부도징후가 주로 1년 전에 나타난다는 것을 짐작해볼 수 있었다. 또한, 변수의 구분을 확인한 결과 활동성 지표가 2개, 단기 유동성 지표가 3개, 장기유동성 지표가 4개로 나타나 생산성 지표가 부도 여부에 큰 영향을 미치지 못한다는 것을 확인하였다. 이는



<Figure 7> Feature Impotance

총공사원가의 크기에 따라 공사의 진행률 및 매출액이 변동하는 건설분야의 특성으로 보여진다. 총공사원가는 공사의 진행과정에서 변동될 여지가 있으며, 이에 따른 매출규모도 달라질 여지가 존재한다. 따라서, 공사 미수금과 분양 미수금에 대한 최종적인 수익 인식을 의미하는 수취채권미결일 변수가 수익성지표보다 더 중요하게 작용한 것으로 해석된다.

<Table 8> Selected Variables

	Financial variables
Activity	Asset Turnover Days-1
	Days Total Receivables Outstanding-2
Liquidity	Cash Ratio-1
	Cash Ratio-2
	Cash Ratio-3
Solvency	Debt Service Coverage Ratio-1
	Debt Service Coverage Ratio-2
	Bank Liabilities to Debt Ratio-1
	Debt to Income Ratio-1

4.3 분류 모델 후보군 및 하이퍼파라미터 선정

본 단계는 문헌연구를 통해 부도예측에 효과적인 블랙박스 머신러닝 모델을 탐색하고 연구에 사용할 분류 모델의 후보군을 선정하는 단계이다. 본 단계에서는 Dastile et al.의 연구를 참고하여 2010년부터 2018년까지 신용평가 및 부도예측을 주제로 머신러닝 모델을 활용한 74개 연구에 사용된 16개 분류 모델을 <Table 9>와 같이 정리하였다[10]. 이를 토대로 SVM, 인공신경망, 랜덤포레스트, XGBoost를 기업부도예측 분류 모델 후보군으로 선정하고, 각 모델별로 선정된 하이퍼파라미터와 하이퍼파라미터의 탐색 범위를 결정하였다.

<Table 9> Machine Learning Models in Bankruptcy Prediction

Abbr.	Model	Frequency
LR	Logistic Regression	38
NB	Naïve Bayes	7
LDA	Linear Discriminant Analysis	5
XGB	XGBoost	4
DT	Decision Tree	23
ELM	Extreme Learning Machine	2
k-NN	k-Nearest Neighbor	10
SVM	Support Vector Machine	43
ANN	Artificial Neural Network	31
BA	Bagging	13
BO	Boosting	16
RF	Random Forest	13
RBM	Restricted Boltzmann Machine	4
DBN	Deep Belief Network	6
DMLP	Deep Multi-Layer Perceptron	4
CNN	Convolutional Neural Network	3

<Table 10> The Range of hyper-parameter by model

Model	Hyper-parameter	Range
SVM	kernel function	rbf
	C	0.0001 ~ 1000
RT	n_estimators	50 ~ 200
	min_sample_leaf	50 ~ 200
	min_samples_split	50 ~ 200
XGB	n_estimators	50 ~ 200
	min_sample_leaf	50 ~ 200
	min_samples_split	50 ~ 200
ANN	activation function	relu
	optimizer	adam
	number of neurons	16, 32
	number of hidden layers	3
	batch_size	64, 128, 256
	epochs	100, 200, 300

선정된 하이퍼파라미터와 하이퍼파라미터의 탐색 범위는 <Table 10>과 같다.

4.4 데이터 불균형을 고려한 분류모델 결정

본 단계는 데이터의 불균형을 고려하여 머신러닝 모델별 최적의 하이퍼파라미터를 결정하는 단계이다. 인공신경망을 제외한 모든 모델은 베이지안 최적화 기법을 활용하여 하이퍼파라미터를 결정하였고, 인공신경망의 경우 소요 시간 문제로 인해 그리드 서치 기법을 이용하여 각 하이퍼파라미터의 값을 결정하였다. 본 단계를 통해 최종 결정된 하이퍼파라미터의 값은 <Table 11>과 같다.

<Table 11> Selected Hyper-Parameter Values by Model

Model	Hyper-parameter	Selected value
SVM	kernel function	rbf
	C	989
RT	n_estimators	186
	min_sample_leaf	176
	min_samples_split	158
XGB	n_estimators	62
	min_sample_leaf	77
	min_samples_split	108
ANN	activation function	relu
	optimizer	adam
	number of neurons	32
	number of hidden layers	3
	batch_size	256
	epochs	100

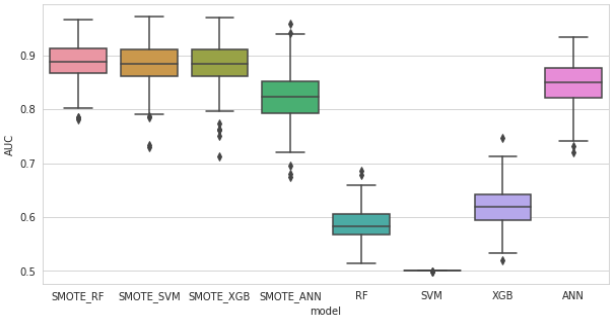
모델별 하이퍼파라미터 선정 후에는 최적의 부도예측 모델 결정을 위해 <Table 11>에서 결정된 하이퍼파라미터 값을 각 모델에 적용하여 300회의 성능 평가를 시행하였다. 추가적으로, SMOTE 및 하이퍼파라미터 조정의 효과성도 함께 파악하기 위하여 SMOTE와 하이퍼파라미터 조정 모두 적용하지 않은 모델도 동일하게 300회의 성능평가를 시행하였다. 단, SMOTE와 하이퍼파라미터 조정을 실시하지 않은 인공신경망의 경우 설계 자체에서 하이퍼파라미터 지정을 필요로 하기 때문에, 각 은닉층의 뉴런 수를 10, 에포크를 50으로 조정하여 비교분석을 진행하였다. 성능 지표는 데이터 불균형 연구에서 주로 활용되는 ROC-AUC를 사용하였으며[12], 비교 분석결과는 <Table 12>, <Figure 8>로 표현하였다.

비교분석 결과, 하이퍼파라미터 튜닝과 SMOTE를 적용한 모델은 인공신경망을 제외하면 SMOTE 미적용 모델과 비교하여 성능이 크게 개선됨을 확인하였다. 특히 랜덤포레스트와 SVM의 경우 중위 값을 기준으로 처음에는 0.5대의 ROC-AUC를 보이다가 하이퍼파라미터 튜닝과 SMOTE 적용 이후 0.8대의 ROC-AUC를 보여 성능이 크게 향상됨을 확인 할 수 있었다.

한편, ROC-AUC 최대값은 SMOTE_SVM이 가장 높았으나 SMOTE_RF가 중앙값, 평균, 최소값에서 가장 우수한 모델인 것을 확인할 수 있었다. 최종적으로 본 연구에서는 안정적으로 높은 성능을 보여주는 SMOTE_RF 알고리즘을 최적의 부도예측 모델로 결정하였다.

<Table 12> Model Performance(ROC-AUC)

	MAX	MEDIAN	MEAN	MIN
SMOTE_RF	0.9654	0.8888	0.8881	0.7814
SMOTE_SVM	0.9712	0.8849	0.8845	0.7294
SMOTE_XGB	0.97	0.8852	0.8855	0.7124
SMOTE_ANN	0.9589	0.8392	0.836	0.6926
RF	0.6852	0.583	0.5863	0.5137
SVM	0.5	0.5	0.4999	0.4997
XGB	0.7477	0.6194	0.6195	0.5207
ANN	0.9339	0.8495	0.8469	0.7198



<Figure 8> Model Performance(ROC-AUC)

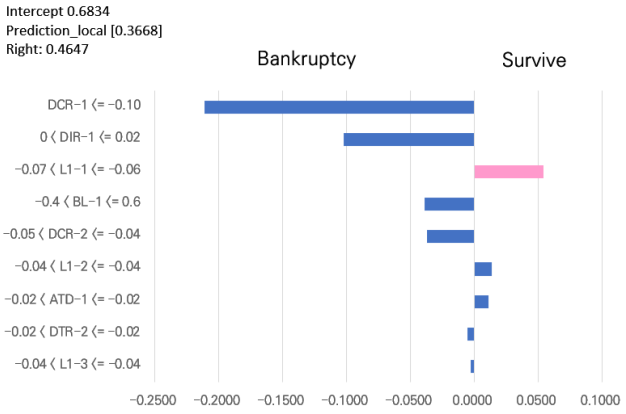
4.5 개선 우선순위 도출

본 단계는 부도예측결과에 XAI의 일종인 LIME 알고리즘을 적용하여 재무변수의 개선 우선순위를 도출하는 단계이다. 먼저, 최종 결정된 모델의 부도예측 결과를 얻기 위해 우선적으로 6,546개 기업의 부도 데이터를 학습데이터와 검증데이터로 분할하였다. 분할비는 7:3이 되도록 하였다. 그 후 최종 결정된 부도예측 모델인 SMOTE_RF를 통해 부도를 예측하고 모델의 성능을 검증하였다.

예측 모델링을 실행한 결과, 모델 정확도(Accuracy)는 94.52%를 보여주었으나, 블랙박스 모델의 특성상 모델의 정확도가 뛰어나다 하더라도 의료나 금융 등 신뢰를 기반으로 하는 시스템에 적용하는 것에는 한계가 존재한다. 이에 따라 본 단계에서는 금융소비자에게 보다 객관적인 정보를 제공하기 위해 블랙박스 모델인 SMOTE_RF에

LIME 알고리즘을 적용하여 어떠한 이유로 예측 결과가 도출되었는지 파악하고 LIME 알고리즘의 해석 값을 통해 재무변수의 개선 우선순위를 도출하였다.

<Figure 9>는 검증데이터에 속한 기업 중 하나를 임의로 선정하여 LIME 알고리즘을 적용한 결과이다. 적용결과, SMOTE_RF 알고리즘이 예측한 실제 1년 뒤 생존율(Right)은 46.47%, LIME으로 SMOTE_RF를 모사한 생존율(Prediction_Local)은 36.68%로 나타났다.



<Figure 9> Variable Sensitivity

<Table 13>은 부도예측 결과에 대한 민감도를 세부적으로 나타낸 것이다. <Table 13>을 통해 확인 한 재무 변수의 개선 우선순위는 채무상환능력계수비율-1, 총부채상환비율-1, 현금비율-1, 은행부채비율-1, 채무상환능력계수비율-2, 현금비율-2, 자산회전일-1, 총수취채권미결일-2, 현금비율-3 순으로 나타났다.

<Table 13> LIME 알고리즘 적용 결과

Feature(π_x)	Sensitivity (Contribution)	weight	Z-score
DCR-1 <= -0.10	-0.2106	2.1611	-0.0974
0 < DIR-1 <= 0.02	-0.1022	-4.9916	0.0205
-0.07 < L1-1 <= -0.06	0.0541	-0.7759	-0.0697
-0.4 < BL-1 <= 0.6	-0.0387	-0.3507	0.1105
-0.05 < DCR-2 <= -0.04	-0.0365	0.7886	-0.0462
-0.04 < L1-2 <= -0.04	0.0141	-0.3486	-0.0404
-0.02 < ATD-1 <= -0.02	0.0112	-0.5723	-0.0196
-0.02 < DTR-2 <= -0.02	-0.0053	0.3214	-0.0164
-0.04 < L1-3 <= -0.04	-0.0027	0.0658	-0.0408

채무상환능력계수비율은 순영업이익을 채무상환부담액으로 나눈 값으로 낮을수록 기업에게 불리한 유동성 지표이며, 총부채상환비율은 연간부채원리금상환액을 영업이익으로 나눈 값으로 높을수록 기업에게 불리한 유동성지표이다.

각 변수별 민감도를 살펴본 결과, 직전년도 채무상환능력계수비율과 총부채상환비율, 현금비율에 대한 민감도의 절대값 합이 0.36인 것으로 나타나 이들 재무 변수가 부도확률 산정에 큰 영향을 준 것으로 파악된다.

5. 결 론

본 연구는 Dastile et al.[9]의 연구에 따라 기업부도예측 연구에서 쉽게 간과되고 있는 데이터 불균형 문제와과 모델 투명성 문제를 모두를 고려한 기업부도예측 방법론을 제안하고 사례적용을 통해 방법론의 유용성과 타당성을 검증하는 것에 목적이 있다.

이에 따라 본 연구에서는 부도예측 모델을 생성하기 위해 SVM, 랜덤포레스트, XGBoost, 인공신경망 등의 머신러닝 알고리즘을 사용하였고. 데이터 불균형 문제와 모델 투명성 문제를 해결하고자 SMOTE와 LIME 알고리즘을 본 모델들에 적용하여 '데이터 불균형을 고려한 설명 가능한 인공지능 기반 기업부도예측 방법론'을 제안하였다. 본 연구의 시사점은 다음의 다섯 가지로 요약된다.

첫째, SMOTE를 사용하여 기업부도예측 문제에 있어서 쉽게 간과될 수 있는 데이터 불균형 문제를 효과적으로 해결 할 수 있음을 확인하였다.

둘째, LIME 알고리즘을 통해 블랙박스 모델의 기업부도예측 근거를 시각화하고, 개별기업의 부도 가능성을 높이는 요인을 찾아 개선 우선순위를 도출하였다.

셋째, 데이터 불균형 처리와 하이퍼파라미터 튜닝을 실시한 결과 랜덤포레스트의 성능이 가장 우수한 것으로 나타나 향후 후속 연구자가 인공지능 기반의 기업부도예측 모델을 설계하는데 참고할 수 있다는 점에서 의의가 있다.

넷째, 기업부도예측 분야에서의 사례적용을 통해 SMOTE와 LIME 알고리즘의 적용 범위를 확장하였다.

다섯째, 연구가 미진하였던 기업부도예측 분야에서의 데이터 불균형 문제와 모델 투명성 문제를 함께 해결 할 수 있는 가이드라인을 제시하였다.

한편, 본 연구의 한계점으로는 다음의 세 가지가 존재한다. 첫째, SMOTE의 경우 소수 클래스 데이터가 밀집 되어 있을 때 비슷한 특성을 가진 데이터를 과도하게 많이 생성할 가능성이 존재한다. 이는 머신러닝 기법의 과적합을 야기하므로 향후 연구에서는 데이터의 분포를 고려할 수 있는 딥러닝 기반 데이터 보강 기법인 적대적 생성 신경망(GAN, Generative Adversarial Networks)을 활용하여 연구를 진행할 필요가 있다.

둘째, LIME 알고리즘의 모델 해석 결과는 선형회귀식으로 표현 될 수 있으나, π_x의 범위가 너무 작아 선형회귀식 도출의 의미가 없는 경우가 존재한다. 따라서 향후 연

구에서는 대리모델의 π_x 범위를 고려하지 않아도 되는 XAI모델인 SHAP을 활용하여 서로 비교연구를 수행할 필요가 있다.

셋째, 사례 연구가 단일성으로 이루어져 다른 부도 데이터에서도 동일하게 효용성이 있는지는 확인 할 수 없다. 향후 연구에서는 국내외를 아우르는 다양한 데이터를 활용하여 다방면으로 검증해볼 필요가 있다.

위와 같은 한계점들이 보완된다면, 금융 소비자에게 보다 정확한 정보를 제공하여 기업의 의사결정에 큰 도움을 줄 수 있을 것으로 기대한다.

Acknowledgement

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea(NRF-2019S1A5C2A04083153).

References

[1] Ahn, J. H., *XAI, Dissects Artificial Intelligence*, Wiki Books, 2020.

[2] Altman, E. I., Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy, *The Journal of Finance,* 1968, Vol. 23, No. 4, pp. 589-609.

[3] Altman, E.I., Marco, G., and Varetto, F., Corporate Distress Diagnosis: Comparisons Using Linear Discriminant Analysis and Neural Networks (the Italian experience), *Journal of Banking & Finance*, 1994, Vol. 18, No. 3. pp. 505-529.

[4] Beaver, W.H., Financial Ratios as Predictors of Failure, *Journal of Accounting Research*, 1966, Vol. 4, pp. 71-111

[5] Breiman, L., Random forests, *Machine Learning*, 2001, Vol. 45, No. 1, pp. 5-32.

[6] Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P., SMOTE: Synthetic Minority oversampling Technique, *Journal of Artificial Intelligence Research*, 2002, Vol. 16, pp. 321-357.

[7] Chen, T. and Guestrin, C., Xgboost: A Scalable Tree Boosting System, In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, August 2016, pp. 785-794.

[8] Cortes, C. and Vapnik, V., Support-vector Networks, *Machine Learning*, 1995, Vol. 20, No. 3, pp. 273-297.

[9] Dastile, X., Celik, T., and Potsane, M., Statistical and Machine Learning Models in Credit Scoring: A Systematic Literature Survey, *Applied Soft Computing,* 2020, Vol. 91, pp. 1-21.

[10] Drotar, P., Gnip, P., Zoričak, M., and Gazda, V., Small- and Medium-Enterprises Bankruptcy Dataset, *Data in brief*, 2019. Vol. 25, pp. 1-6.

[11] Edmister, R. O., An Empirical Test of Financial Ratio Analysis for Small Business Failure Prediction, *Journal of Financial and Quantitative Analysis*, 1972, Vol. 7, No. 2, pp. 1477-1493.

[12] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G., Learning from Class- Imbalanced Data: Review of Methods and Applications, *Expert Systems with Applications*, 2017, Vol. 73, pp. 220-239.

[13] Jo, H. and Han, I., Integration of Case-Based Forecasting, Neural Network, and Discriminant Analysis for Bankruptcy Prediction, *Expert Systems with applications*, 1996. Vol. 11, No. 4, pp 415-422.

[14] John, G.H., Kohavi, R., and Pfleger, K., Irrelevant Features and the Subset Selection Problem, In *Machine Learning Proceedings*, 1994, pp. 121-129.

[15] Keany, E., Boruta-Shap: A Tree Based Feature Selection Tool which Combines Both the Boruta Feature Selection Algorithm with Shapley Values, 2019. [Website] (2021, Nov .27). https://github.com/Ekeany/Boruta-Shap.

[16] Kim, H., GAN-based Oversampling Technique for Imbalanced Bankruptcy Data Processing. Master's Thesis, Ewha Womans University, 2020.

[17] Kim, S.J. and Ahn, H.C., Application of Random Forests to Corporate Credit Rating Prediction, *The Journal of Business and Economics*, 2016, Vol. 32, No. 1, pp. 187-211.

[18] Kotsiantis, S., Tzelepis, D., Koumanakos, E., and Tampakas, V., Selective Costing Voting for Bankruptcy Prediction, *International Journal of Knowledge-based and Intelligent Engineering Systems,* 2007, Vol. 11, No. 2, pp. 115-127.

[19] Marvin, M. and Seymour, A.P., *Perceptrons,* MIT Press, 1969.

[20] McCulloch, W.S. and Pitts, W., A Logical Calculus of the Ideas Immanent in Nervous Activity, *The bulletin of Mathematical Biophysics*, 1943, Vol. 5, No. 4, pp. 115-133.

[21] Nilsson, R., Pena, J. M., Bjorkegren, J., and Tegnor, J., Consistent Feature Selection for Pattern Recognition in Polynomial Time, *The Journal of Machine Learning*

Research, 2007, Vol. 8, pp. 589-612.

[22] O'Brien, R. and Ishwaran, H., A Random Forests Quantile Classifier for Class Imbalanced Data, *Pattern Recognition*, 2019, Vol. 90, pp. 232-249.

[23] Odom, M.D. and Sharda, R., A Neural Network Model for Bankruptcy Prediction, In *1990 IJCNN International Joint Conference on Neural Networks,* June 1990.

[24] Ohlson, J.A., Financial Ratios and the Probabilistic Prediction of Bankruptcy, *Journal of Accounting Research*, 1980, Vol. 18, No. 1, pp. 109-131

[25] Ohn, S.Y., Chi, S.D., and Han, M.Y., Feature Selection for Classification of Mass Spectrometric Proteomic Data Using Random Forest, *Journal of the Korea Society for Simulation*, 2013, Vol. 22, No. 4, pp. 139-147.

[26] Park, J.R., A Study on Improving Turnover Intention Forecasting Power through Solving Imbalanced Data Problems: Focusing on SMOTE and Generative Adversarial Networks, Doctorial Dissertation, Chungbuk National University, 2021.

[27] Pinches, G.E., Mingo, K.A., and Caruthers, J.K., The Stability of Financial Patterns in Industrial Organizations, *The Journal of Finance,* 1973, Vol. 28, No. 2, pp. 389-396.

[28] Ribeiro, B. and Lopes, N., Deep Belief Networks for Financial Prediction, In *International Conference on Neural Information Processing*, 2011. pp. 766-773

[29] Rosenblatt, F., The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, *Psychological Review*, 1958, Vol. 65, No. 6, pp. 386-408.

[30] Rumelhart, D.E., Hinton, G.E., and Williams, R.J., Learning Representations by Back-Propagating Errors, *Nature*, 1986, Vol. 323, No. 6088, pp. 533-536.

[31] Scholkopf, B., The Kernel Trick for Distances, *Advances in neural information processing systems*, 2000, Vol.13.

[32] Shin, K.S., Lee, T.S., and Kim, H.J., An Application of Support Vector Machines in Bankruptcy Prediction Model, *Expert Systems with Applications*, 2005, Vol. 28, No. 1, pp.127-135.

[33] Sun, Y., Kamel, M.S., Wong, A.K., and Wang, Y., Cost-sensitive Boosting for Classification of Imbalanced Data, *Pattern Recognition*, 2007, Vol. 40, No. 12, pp. 3358-3378.

[34] Tam, K.Y. and Kiang, M.Y., "Managerial Applications of Neural Networks: The Case of Bank Failure Predictions, *Management Science*, 1992, Vol. 38, No. 7, pp. 926-947.

[35] Wang, B. X. and Japkowicz, N., Boosting Support Vector Machines for Imbalanced Data Sets. *Knowledge and Information Systems*, Vol. 25, No. 1, pp. 1-20.

[36] Wilson, R.L. and Sharda, R., Bankruptcy prediction using neural networks, *Decision Support Systems*, 1994, Vol. 11, No. 5, pp 545-557.

[37] Wu, C.H., Tzeng, G.H., Goo, Y.J., and Fang, W.C., *A Real-valued Genetic Algorithm to Optimize the Parameters of Support Vector Machine for Predicting Bankruptcy*, Expert systems with application, 2007. Vol. 32, No. 2, pp. 397-408

[38] Yoo, J.E., Random Forests, an Alternative Data Mining Technique to Decision Tree, *Journal of Educational Evaluation*, 2015, Vol. 28, No. 2, pp. 427-448.

[39] Zmijewski, M.E., Methodological Issues Related to the Estimation of Financial Distress Prediction Models, *Journal of Accounting Research*, 1984, Vol. 22, pp. 59-82

[40] Zoričak, M., Gnip, P., Drotar, P., and Gazda, V., Bankruptcy Prediction for Small-and Medium-sized Companies Using Severely Imbalanced Datasets, *Economic Modelling*, 2020, Vol. 84, pp. 165-176

ORCID

Sun-Woo Heo | https://orcid.org/0000-0002-3216-6422

Dong Hyun Baek | http://orcid.org/0000-0002-3107-9511