머신러닝을 이용한 외감기업 및 비외감기업의 부도예측에 관한 연구

송현준*, 박도준**, 이준기***

An Empirical Comparison of Bankruptcy Prediction of External Auditing and Non-External Auditing Companies Using Machine Learning Methods

Hyunjun Song*, Dojoon Park**, Zoonky Lee***

요 약

본 연구는 머신러닝 기법을 사용하여, 우리나라 외감기업과 비외감기업의 부도예측 모델들을 실증분석하였다. 자료의 불균형성을 균형화하기 위해 다양한 표집방법을 사용하고, 기존의 방법론과 머신러닝 방법론으로 모델을 설계하였다. 분석결과에서 외감기업에 대한 예측력이 비외감기업 대비 높게 나타났으며, 기존의 방법론 대비 머신러닝 방법론의 예측력이 일관되게 향상되었다. 또한, 과대표집 방법으로 일부 머신러닝 방법론의 예측력을 향상시킬 수 있었다.

ABSTRACT

The paper presents an empirical analysis of bankruptcy prediction models using financial data of external auditing and non-external auditing companies. To improve performance occurring from the imbalanced data, we apply four resampling techniques and compare the performance of conventional models and machine learning models. The result indicates that the machine learning models show consistently higher performance compared to conventional methods. In addition, the predictability of the model is higher for external auditing companies compared to non-external auditing companies.

키워드 : 부도율 예측, 머신러닝, 비외감기업, 로지스틱 회귀, SMOTE

Key Words : Bankruptcy prediction, Machine learning, Non-external auditing, Logistic regression, SMOTE

Ⅰ. 서 론

기업은 사업의 운영 또는 신규 투자 등에 필요한 자금을 회사채를 발행하거나, 은행에서 대출을 받는 방법으로 금융시장에서 조달한다. 기업의 부채는 원금 또는 이자를 상환하지 못하는 부도위험 및 신용위험이 있고, 따라서 투자자 또는 은행은 회사채에 대한 투자 또는 대출의 의사결정 및 부채의 이자율 결정에 기업의 부도 확률, 부도가 발생할 경우의 회수가능성 및 기타 관련 비용 등을 고려한다. 기업의 부도위험은 개별 기업의 사업위험, 재무위험 등의 내적 요인뿐만 아니라, 경제성장률, 인플레이션, 금리 변화 등의 경기 순환적 요인과 밀접한 관계가 있다. 기업의 부도를 예측하고, 신용위험을 잘 관리하는 것은 금융기관의 중요한 목표 중 하나이다. 하지만, 예측모델의 오류, 환경의 변화 등 다양한 이유로 예측과 달

* 연세대학교 정보대학원 비즈니스빅데이터분석학과(hsong1667@gmail.com), ** 연세대학교 경영대학(dojoon@yonsei.ac.kr)
***(교신저자) 연세대학교 정보대학원(zlee@yonsei.ac.kr)
§ 논문번호 : 211610, 접수일자 : 2021년 01월 20일, 수정일자 : 2021년 02월 25일, 심사완료일자 : 2021년 03월 28일
§ 이 논문은 2021년도 연세대학교 미래융합연구원(ICON)과 '4단계 두뇌한국21 사업(4단계 BK21 사업)'에 의하여 지원되었음.

리 기업의 부실이 발생하는 경우, 해당 기업의 관계자, 유관 기업, 투자자 및 금융기관에게 연쇄적인 피해가 발생할 수 있다. 이에 따라 기업의 부도예측은 신용위험 관리의 주체인 금융기관과 더불어 금융기관의 건전성을 관리 감독하는 감독기관에게도 매우 중요하다.

최근의 인공지능과 머신러닝의 발전은 금융 산업 분야에 있어서 새로운 시각을 전하여주고 있으며, 부도예측 분야에서도 머신러닝 기법을 적용해 보는 많은 시도가 일어나고 있다. 대부분의 연구는 Altman이 부도예측 모델에 사용한 5개 변수에 새로운 재무지표를 찾아서 DT(Decision Tree), NN(Neural Network), SVM(Support Vector Machine) 등의 머신러닝 기법을 적용하여 회귀분석이나 판별분석 등의 기존 방식과 비교하고 있다[1-5]. 연구들은 머신러닝 기법이 기존의 방식에 비해 약간의 정확성을 높이는 것으로 결과를 보고하고 있다. 하지만 재무제표의 지표들을 이용하는 부도예측 방식은 외부감사인의 업력, 비용 및 능력 등에 예측 정합도가 영향을 받는다고 알려져 있는데[6], 이것은 재무제표의 신뢰성, 적시성 등에 기인한 것으로 보인다. 또한 우리나라 기업 중 공식적인 외부감사를 받은 기업과 그렇지 않은 기업 간 회계자료에는 그 투명성에도 큰 차이가 있는 것으로 보고되어, 비외감기업의 부도예측은 상대적으로 어려움이 큰 것으로 판단된다[7]. 대부분의 부도가 상장되지 않은 중소기업에서 일어나고 있는 현실을 감안하면 비외감기업에 대한 연구가 더 요구되고 있는 현실에도 불구하고 데이터의 한계 등으로 인하여 중소기업 등의 비외감기업에 대한 부도예측 연구는 아주 제한적이다.

본 연구는 외감기업과 비외감기업에 대하여 기존 방법론, 머신러닝 방법론, 불균형데이터 균형방법(과소표집, 과대표집) 등을 각각 적용하여 머신러닝 방법론이 비외감기업에 대하여 정확성을 어느 정도 향상 시킬 수 있는가를 알아보는 연구이다. 좀 더 구체적으로, 추가 변수를 넣지 않고 기존의 Altman의 모델을 적용함으로써 회귀분석 방식들의 변수의 선형적 함수에 대한 가정의 완화가 어느 정도 정확도를 상승으로 이어지는 가를 연구하려 한다. 이 연구의 결과는 향후 추가 재무변수의 효과와 머신러닝의 효과를 분리하여 볼 수 있게 해주어, 선형모형의 가정의 완화가 머신러닝에 의하여 어떻게 정확도를 높일 수 있는가에 대한 이해도를 높여 줄 것이며, 실무적으로도 비외감기업에 대한 부도예측 정확도를 높이는 방식을 제안함으로써 중소기업의 부도예측에 도움이 될 것으로 기대한다.

II. 관련연구

부도예측 모델은 부채가 있는 기업을 부도위험이 높은 기업과 재무상태가 건전한 정상기업 두 개의 그룹으로 분류하는 것을 목적으로 한다. 통상적으로 전체 기업 중 부도가 발생한 기업의 개수가 상대적으로 매우 작기 때문에, 부도예측 연구에 사용되는 원데이터는 불균형데이터라는 특징이 있다. 부도예측 모델 구축의 일반적인 절차는 이러한 불균형데이터를 정상기업을 과소표집(undersampling)하거나 부도기업을 과대표집(oversampling)하는 방법을 균형화하고, 부도예측에 사용될 기업의 특성 변수를 선정한 후, 회귀분석, 판별분석 및 머신러닝 방법론 등을 활용하여 모델을 설계한다.

기업의 부도예측에 사용되는 부도관련 데이터는 정형화되어 있지 않기 때문에, 개별 연구자 또는 금융기관의 모델 설계에 어려움이 있다. 기업의 대출금 출자전환 요청 또는 협조적인 추가 융자 요청 등의 부분적인 채무불이행 위험을 포함하기 위해서는 데이터를 수기로 수집해야 하기 때문에, 대안적인 방법으로 상장폐지 데이터 등을 사용하는 경우도 있다. Altman은 저서에서, '부도예측의 연구에서는 데이터가 왕(king)이다.'라고 표현할 만큼 데이터의 중요성을 강조하였다[8]. 연구에 사용된 부도기업의 개수를 살펴보면, 초기 연구인 1960년대 Altman의 Z-score에는 65개의 부도기업이 사용되었으며[1], 국내 기업의 특수성을 고려해 Z-score를 수정한 K-score에는 34개의 부도기업이 사용되었다[9]. 최근 외국 연구사례를 살펴보면, 1981년부터 2009년까지 미국의 부도기업은 918개(전체의 1.10%)이며, 같은 기간 일본의 부도기업은 58개(전체의 0.16%)이다[10]. 우리나라의 부도예측 연구에서 사용된 상장폐지 기업의 수는 2001년부터 2013년까지 643개(전체의 3.16%)이다[11]. 데이터의 한계로 중소기업을 대상으로 한 연구 사례는 매우 드물다. 이탈리아의 중소기업(small and medium-sized enterprises)에 대한 연구사례에서는 2004년부터 2013년까지 부도기업 520개(전체의 3.58%)가 사용되었으며[12], 우리나라의 비상장 중소기업 자료를 사용한 연구는 2002년부터 2007년까지 부도기업 1,350개(전체의 6.00%)가 사용되었다[13]. 데이터의 한계로 인해 국가 간 부도예측력을 비교한 사례는 있으나, 한 국가에서 대기업과 중소기업으로 그룹을 구분하여 비교한 사례는 찾기 힘들다. 본 연구는 외감기업과 비외감기업을 동시에 실증분석하여 두 그룹 간 예측력의

정합도를 직접적으로 비교할 수 있었으며, 사용된 부도기업의 개수는 약 23,466개(전체의 3.32%)로 상대적으로 광범위하게 많은 부도 데이터를 사용하였다.

부도예측 모형은 사용된 기업의 특성 변수에 따라 회계모형, 시장모형 그리고 헤저드모형으로 구분이 가능하다. 초기의 부도모형은 주로 재무제표 정보를 활용한 회계모형으로, 판별분석, 로짓분석 및 프로빗분석 등의 방법론을 사용하였다[1,14,15,16]. Altman의 Z-score의 경우 5개의 재무비율을 사용하였는데, 이후 500여개의 다양한 재무정보가 부도예측 연구에 사용되었다[17]. 기업의 주식이 상장되어 있는 경우 옵션가격모형으로 주식의 가격정보에서 부도거리(distance to default)를 측정하거나[18], 예상부도확률(expected default frequency)을 계산할 수 있다. 이러한 부도예측 모형은 기업의 미래 수익정보를 포함하는 시장정보를 부도예측에 활용하였다는 특징이 있다. 헤저드모형은 회계정보와 시장정보를 모두 활용한다[19]. 본 연구에 포함된 대부분의 기업들은 비상장기업이기 때문에 예측모델에 시장정보를 사용할 수 없는 한계가 있어 재무제표 정보만을 활용하였다.

불균형데이터 균형방법과 부도 예측력의 관계에 대한 연구도 지속적으로 진행되어 왔다. 일반적으로 과소표집 방법은 유용한 정보를 모두 사용하지 못하는 반면, 과대표집 방법은 과적합의 가능성이 있고 모형 추정에 상대적으로 많은 계산시간이 필요하다[20,21]. 표집방법의 개선을 통해 모형의 예측력을 높이려는 노력이 지속되었는데, 과소표집을 하면서도 자료의 특성을 최대한 사용하여 부도예측을 향상시키는 방법론이 제시되었다[13], 한편, 다양한 설명변수를 포함시키고 고도화된 머신러닝 방법론의 활용을 위해 저빈도 데이터를 합성하는 방법으로 과대표집하는 여러 방법론들이 제시되었으며[22], 실증분석에서 과대표집을 하는 경우의 예측력이 향상되었다는 보고가 있었다[10,23]. 최근 컴퓨팅 기술의 발전으로 과대표집의 계산시간 문제는 대부분 완화되어, 본 연구는 과소표집 방법과 전체 자료의 특징을 활용하는 다양한 과대표집 방법을 사용하였다.

Ⅲ. 분석방법과 데이터

3.1 분석방법

본 연구는 전체 데이터를 7:3 비율로 학습 데이터(training set)와 검증 데이터(test set)로 구분하여 학습 데이터를 이용해 모형을 추정하고 검증 데이터로 추정된 모형을 평가하였다. 학습 데이터 및 검증 데이터는 모두 정상기업보다 부도기업의 수가 작은 불균형데이터인데, 학습 데이터에 대해서만 균형방법을 적용한 후 부도모델을 추정하고, 검증 데이터에는 실제 부도기업의 분포를 고려하여 균형방법을 적용하지 않고 모델을 검증하였다. 초기의 부도예측 연구는 모델의 검증에도 균형데이터를 사용했는데, 실제와 같은 분포의 데이터로 모델을 검증할 필요성 때문에 최근 연구 사례는 본 연구와 같이 불균형데이터를 사용하고 있다[10].

과소표집 방법으로는 빈도수가 낮은 부도기업의 개수만큼 정상기업 데이터에서 무작위로 데이터를 선정하였으며, 생성된 균형데이터로 모델 추정을 1000회 반복한 후 평균을 보고하였다. 과대표집 방법으로는 저빈도 데이터를 합성하는 기법인 SMOTE (Synthetic Minority Oversampling Technique), Borderline-SMOTE, 그리고 ADASYN (Adaptive Synthetic Sampling)을 사용한다. SMOTE는 저빈도인 부도기업 데이터 주변에서 다른 부도기업 데이터를 KNN(K-Nearest Neighbor) 기법을 이용해 추출하고, 무작위의 선형조합을 통해 새로운 부도기업 데이터를 생성하는 방식이다. Borderline-SMOTE는 저빈도 데이터 주변의 데이터 중 KNN으로 경계를 판단하여 추출하는 방식으로 새로운 데이터 조합에 경계 데이터를 사용하는 것이 특징이다. ADASYN은 SMOTE와 유사하나 저빈도 데이터를 모두 동일하게 사용하지 않고 일부 경계 자료에 중요도를 두기 때문에 생성된 데이터가 실제 데이터와 더 유사하다고 알려져 있다. 과대표집을 실시하여 부도기업의 수를 정상기업의 숫자 만큼 생성하였다.

학습 데이터로 추정하는 부도예측 모델은 기존 연구에 사용된 판별분석 (LDA:linear discriminant analysis)과 로지스틱 회귀분석에 더해 다음의 5개 머신러닝 기법이 추가되었다: DT(Decision Tree), RF(Random Forest), AB(Adaptive Boosting), GB(Gradient Boosting), NN(Neural Network). 결론적으로 외감기업과 비외감기업에 대해 각각 위에서 소개된 과소/과대 표집 4개의 표집방법을 적용한 후, 7개의 기법으로 모델을 추정하여 총 28개의 모델이 만들어졌다.

3.2 모델의 검증

그룹별로 추정된 28개 모델에서 검증 데이터의 부도 여부를 예측하였고, 실제 부도 여부와 비교하였다. 모델 간 비교에는 민감도, 특이도, F1 score

에 더하여 불균형 데이터에서 최근 많이 사용되는 G-Mean 과 AUC(Area Under curve)가 사용되어[24,25,26], 다음의 5가지 지표를 사용하였다.

- 민감도(sensitivity): 실제 부도가 발생한 기업 중 모델이 부도기업으로 예측한 비율(정확히 부도기업을 예측한 개수/전체 부도기업 개수)
- 특이도(specificity): 실제 정상인 기업을 모델이 정상기업으로 예측한 비율(정확히 정상기업을 예측한 개수/전체 부도기업 개수)
- F1 score: 모델이 부도기업으로 예측한 기업 중 실제 부도기업의 비율(정확히 부도기업을 예측한 개수/부도로 예측한 개수)인 정밀도(precision)를 사용하여, 민감도와 정밀도의 곱을 두 값의 평균으로 나는 값. F1 score가 높으면 모형의 부도 예측력이 높다고 해석할 수 있다.
- G-Mean: 민감도와 특이도의 기하평균 ($\sqrt{\text{민감도} \times \text{특이도}}$)
- AUC(Area Under Curve): 민감도와 특이도의 산술평균

3.3 표본기업 데이터와 기업의 특성변수

본 연구의 표본기간은 2012년부터 2016년까지 5년이다. 자료의 개수는 기업 및 연도별 총 706,280개이며, 이 중 부도가 발생한 기업의 수는 23,466개이다. 매년 각 기업의 재무비율을 계산하고, 1년간 부도 발생 여부를 관찰한다. 부도는 파산신청, 당좌거래 정지, 법정관리, 부도유예 협약, 화의신청, 워크아웃, 채무조정, 은행연합회 신용정보 관리 규약 중 신용 불량, 대손상각 등 특정 차주에 대한 신용손실 발생 및 90일 이상 연체 기업을 포함한다. 연구에 사용된 외감기업과 비외감기업의 개수와 정상기업 및 부도기업의 개수는 표 1과 같다.

표 1. 부도기업 데이터

	정상기업	부도기업	전체
외감기업	87,557	2,175	89,732
비외감기업	595,257	21,291	616,548
합계	682,814	23,466	706,280

외감기업은 전체 중 약 2.42%가 부도기업이고, 비외감기업은 약 3.45%가 부도기업이다. 비외감기업의 개수는 외감기업 대비 약 7배이고, 비외감기업 중 부도기업의 개수는 외감기업의 부도기업 대비 약 10배이다. 표본기간 동안 전체 부도의 약 90.7%가 비외감기업에서 발생한 것으로 나타났다.

부도예측 모델에 사용한 기업의 특성변수는 Altman의 Z-score에 사용된 운전자본비율(운전자본/자산총계), 누적수익성비율(이익잉여금/자산총계), 총자산영업이익률(영업이익/자산총계), 자본부채비율(자본의 시장가치/총부채의 장부가치), 총자산회전율(매출액/자산총계) 등 5개의 재무비율이다. 데이터의 대부분이 비상장회사이므로 시가총액을 사용하는 자본부채비율 대신 부채비율(부채총계/자산총계)을 사용하였다.

IV. 분석 결과

외감기업과 비외감기업의 데이터를 7:3 비율로 학습 데이터와 검증 데이터로 구분하였다. 학습 데이터는 외감기업의 경우 부도기업 1,495개와 정상기업 61,317개로 구성되며, 비외감기업의 경우 부도기업 14,865개와 정상기업 416,718개로 구성된다. 검증 데이터는 외감기업의 경우 부도기업 680개와 정상기업 26,240개로 구성되며, 비외감기업의 경우 부도기업 6,426개와 정상기업 178,539개로 구성된다. 4개의 표집방법과 7개 방법론을 사용하여 부도예측 모델을 추정 및 검증하는데, 모두 동일한 학습 데이터와 검증 데이터가 사용되었다.

외감기업과 비외감기업의 학습 데이터를 과소표집 및 과대표집 방법으로 데이터를 균형화하고, 모델을 추정한 후 검증 데이터를 사용해 검증한 결과가 표 2이다. 과소표집 방법의 경우 학습 데이터의 개수는 부도기업의 두 배가 되므로, 외감기업의 경우 2,990개이고, 비외감기업의 경우는 29,730개이다. 분석결과 예상대로, 외감기업의 예측력이 비외감기업의 예측력에 비하여 높게 나옴을 알 수 있었다. 표 2에서 보듯이 외감기업 예측모델의 AUC와 G-mean 값이 비외감기업의 예측모델 대비 약 17~18% 높게 나타났다. 표 2에는 포함되어 있지 않지만, 상장회사를 대상으로 한 기존 연구와의 비교를 위해 자산규모 600억 이상의 기업을 대상으로 동일한 분석을 진행하였다. 자료의 개수는 약 2만개(외감기업의 24%)이며, 이 중 1.76%가 부도기업이다. 분석결과는 LDA와 LR 기법의 AUC와 G-mean 값이 약 0.74~0.76이며, 머신러닝 기법의 경우 약 0.74~0.79 수준으로 분석되었다. 즉 상대적으로 신뢰성과 투명성이 보장된 대기업을 대상으로 한 부도 예측의 경우 비외감기업의 예측력에 비하여 약 25% 정도 높게 나타나 기존의 연구 결과와 비슷한 수준을 보였지만 비외감기업의 경우 예측력이 현저히 떨어짐을 볼 수 있었다.

예측력을 높이기 위한 머신러닝 적용의 결과는 외

감기업, 비외감기업 모두 기존의 모델 추정방법인 LDA와 LR 기법에 비해 머신러닝 기법의 예측력이 높게 나타났다. 외감기업의 경우, 기존의 LDA 기법의 예측력은 AUC와 G-mean 값이 약 0.62 이며, LR 기법은 약 0.70 이다. 머신러닝 기법의 AUC와 G-mean 값은 상대적으로 예측력이 좋았던 LR 기법 대비 5% 정도 향상되었으며, LDA 기법 대비는 약 20% 향상되었다. 비외감기업의 경우도, LDA와 LR 기법에 비해 대체로 머신러닝 기법의 예측력이 높게 나타났다. DT 기법을 제외한 나머지 기법들의 AUC와 G-mean 값은 기존의 LDA와 LR 기법 대비 4~5% 향상되었다.

과대표집(SMOTE, Borderline-SMOTE, ADASYN) 방법의 경우, 부도기업을 정상기업의 숫자만큼 합성하기 때문에 학습 데이터의 개수는 정상기업의 두 배가 된다. 외감기업의 경우 122,634개이고, 비외감기업의 경우는 833,436개이다. 과소표집과 과대표집 방법을 비교하면, 과대표집 방법의 예측력이 과소표집 대비 높지는 않지만, 일부 머신러닝 기법들은 과대표집 방

표 2. 검증 결과

	외감기업					비외감기업				
	민감도	특이도	F1 score	AUC	G-mean	민감도	특이도	F1 score	AUC	G-mean
	Undersampling									
LDA	0.5544	0.6892	0.7921	0.6218	0.6181	0.6083	0.5418	0.6752	0.5751	0.5741
LR	0.7029	0.7051	0.7040	0.7040	0.7040	0.6086	0.5943	0.7166	0.6014	0.6014
DT	0.7838	0.6940	0.7989	0.7389	0.7375	0.8075	0.4257	0.5768	0.6166	0.5863
RF	0.8019	0.6893	0.7960	0.7456	0.7434	0.7275	0.5148	0.6552	0.6212	0.6120
AB	0.8103	0.6457	0.7652	0.7280	0.7233	0.6947	0.5614	0.6927	0.6280	0.6245
GB	0.8147	0.6848	0.7931	0.7498	0.7469	0.7009	0.5709	0.7004	0.6359	0.6326
NN	0.8172	0.6837	0.7923	0.7504	0.7474	0.6780	0.5912	0.7153	0.6346	0.6325
	SMOTE									
LDA	0.6265	0.6424	0.7601	0.6344	0.6345	0.7200	0.3630	0.5130	0.5415	0.5113
LR	0.7015	0.7024	0.8034	0.7019	0.7019	0.6962	0.4079	0.5577	0.5521	0.5329
DT	0.8015	0.6684	0.7814	0.7349	0.7319	0.7828	0.4258	0.5764	0.6043	0.5773
RF	0.8029	0.6899	0.7964	0.7464	0.7443	0.6993	0.5425	0.6776	0.6209	0.6160
AB	0.7882	0.6497	0.7678	0.7190	0.7156	0.6807	0.5709	0.7000	0.6258	0.6234
GB	0.7750	0.7119	0.8110	0.7435	0.7428	0.6772	0.5844	0.7104	0.6308	0.6291
NN	0.7618	0.7120	0.8108	0.7369	0.7365	0.6660	0.6096	0.7294	0.6378	0.6372
	Borderline-SMOTE									
LDA	0.6250	0.6771	0.7848	0.6510	0.6505	0.6981	0.3916	0.5417	0.5449	0.5229
LR	0.6294	0.7456	0.8309	0.6875	0.6850	0.6685	0.4419	0.5895	0.5552	0.5436
DT	0.7485	0.7048	0.8058	0.7267	0.7263	0.5955	0.6306	0.7435	0.6131	0.6128
RF	0.7324	0.7418	0.8301	0.7371	0.7371	0.6089	0.6407	0.7511	0.6248	0.6246
AB	0.7000	0.7307	0.8223	0.7153	0.7152	0.5996	0.6511	0.7584	0.6254	0.6248
GB	0.6691	0.7751	0.8504	0.7221	0.7202	0.5984	0.6646	0.7680	0.6315	0.6306
NN	0.6721	0.7879	0.8585	0.7300	0.7277	0.5627	0.6823	0.7795	0.6225	0.6196
	ADASYN									
LDA	0.6412	0.6275	0.7494	0.6343	0.6343	0.5258	0.6007	0.7194	0.5632	0.5620
LR	0.7059	0.6976	0.8002	0.7017	0.7017	0.6802	0.4276	0.5763	0.5539	0.5393
DT	0.7824	0.6818	0.7905	0.9321	0.7304	0.7166	0.5098	0.6508	0.6132	0.6044
RF	0.8147	0.6829	0.7917	0.7488	0.7459	0.6925	0.5464	0.6807	0.6195	0.6151
AB	0.7853	0.6445	0.7640	0.7149	0.7114	0.6864	0.5647	0.6952	0.6256	0.6226
GB	0.7765	0.7029	0.8049	0.7397	0.7387	0.6757	0.5817	0.7083	0.6287	0.6269
NN	0.7426	0.7372	0.8273	0.7399	0.7399	0.6702	0.6002	0.7224	0.6352	0.6342

법에 의해 예측력을 향상된 것을 알 수 있다. SMOTE 방법과 ADASYN 방법을 사용한 경우 외감기업의 분석결과에서는 LDA와 RF 기법의 AUC와 G-mean 값이 향상되었으며, 비외감기업의 분석결과에서는 NN 기법의 AUC와 G-mean 값이 증가하였다. 각각의 과대표집 방법에서 기존의 LDA와 LR 기법의 예측력에 비해 머신러닝 기법의 예측력은 약 9~15% 향상되었으며, 과소표집 방법의 LDA와 LR 기법에 비해서는 과대표집 방법을 적용한 머신러닝 기법의 모델의 예측력은 AUC 값은 약 10~13%, G-mean 값은 약 5% 향상되었다. 과대표집을 하는 경우에서도 외감기업의 예측력이 비외감기업의 예측력에 비해 높게 나타났다.

분석결과를 종합해보면, 외감기업의 예측력이 비외감기업 대비 높게 나타났으며, 기존의 모델 추정 방법인 LDA와 LR 기법에 비해 머신러닝 기법을 사용하는 경우는 표본을 과소 또는 과대표집하는 경우에 상관없이 일관되게 예측력이 향상되는 점을 발견할 수 있었다. 과소표집과 과대표집 방법의 비교에서, 과대표집 방법이 일관적으로 예측력을 높이는 것은 아니지만, 일부 머신러닝 기법들은 과대표집 방법에 의해 예측력을 향상시킬 수 있었다.

V. 결 론

최근 인공지능에 대한 높은 관심과 투자로 머신러닝 방법론이 급격히 발전하고 있다. 금융 산업 분야에서도 금융과 인공지능이 융합된 핀테크(FinTech)가 활성화 되고 있으며, 은행, 보험, 투자, 재무, 위험관리 등 광범위한 영역에서 머신러닝을 사용하는 서비스가 확대되고 있다. 본 연구는 머신러닝 기법을 사용하여, 우리나라 기업의 부도예측 모델을 다양한 방법으로 설계하고, 이를 비교 검증하였다. 부도관련 자료의 특징인 불균형성을 균형화하기 위해 다양한 표집방법을 사용하였고, 기존의 방법론과 머신러닝 방법론으로 모델을 추정하고 그 예측력을 비교하였다. 분석결과 머신러닝을 사용한 모델의 예측력은 기존의 방법론을 사용한 모델에 비해 향상되는 것으로 나타났다. 또한, 일부 머신러닝 기법들은 과대표집 방법에 의해 예측력을 향상시킬 수 있었다.

대부분의 부도관련 연구 대상이 상장기업인 반면, 본 연구는 상장기업을 포함하는 외감기업과 비외감기업, 두 그룹을 대상으로 연구를 진행하였다는 차별점이 있다. 표본기간 중 전체 부도 기업의 90% 이상이 비외감기업에서 발생했다는 것을 감안하면 비외감기업에 대한 부도예측 연구는 학문적으로 뿐만 아니라 실무적으로도 중요한 의미를 갖는다. 회계자료 투명성에 차이가 있는 비외감기업의 부도예측은 외감기업에 비해 상대적으로 어려움이 많지만, 본 연구의 실증분석은 비외감기업의 부도예측에도 머신러닝 기법을 적용하여 부도의 예측력을 향상시킬 수 있음을 시사한다.

본 연구는 재무제표의 자료 중 5개의 재무비율을 부도예측에 사용하였고, 부도에 영향을 미치는 경기 순환적 요인은 고려하지 않았다. 추가적으로 재무비율과 경기 순환적 요인을 반영하는 변수를 모델에 포함시켜 예측력의 향상이 가능하며, 상장회사의 경우 주식가격에서 얻을 수 있는 시장정보를 모델에 포함시킬 수 있다. 향후 부도예측과 관련된 추가적인 변수를 선정해 설명변수를 다양화하고, 표집방법의 개선 및 머신러닝 기법의 고도화를 통해 부도모델의 예측력을 향상시킬 수 있는 개선의 여지가 남아있어, 머신러닝을 이용한 부도예측 분야 연구에 다양한 분석이 필요할 것으로 판단된다.

참고문헌

[1] Altman, E.I., "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy", The journal of finance, Vol. 23. No. 4, 1968, pp.589-609.

[2] Shaw, M.J. & Gentry, J.A., "Inductive learning for risk classification", IEEE Intelligent Systems, Vol. 5, No. 1, 1990, pp.47-53.

[3] Tam, K.Y. & Kiang, M.Y., "Managerial applications of neural networks: the case of bank failure predictions", Management science, Vol. 38, No. 7, 1992, pp.926-947.

[4] Barboza, F., Kimura, H. & Altman, E., "Machine learning models and bankruptcy prediction", Expert Systems with Applications, Vol. 83, 2017, pp.405-417.

[5] Lu, H., Li, Y., Chen, M., Kim, H. & Serikawa, S., "Brain intelligence: go beyond artificial intelligence", Mobile Networks and Applications, Vol. 23, No. 2, 2018, pp.368-375.

[6] Cenciarelli, V.G., Greco, G. & Allegrini, M., "External audit and bankruptcy prediction", Journal of Management and Governance, Vol. 22, No. 4, 2018, pp.863-890.

[7] 김성규, "중소기업 회계투명성 제고 방안", 중소기업금융연구, 제38권, 제3호, pp 3-45, 2018.

[8] Altman, E.I., Hotchkiss, E. & Wang, W., "Corporate financial distress, restructuring, and bankruptcy: analyze leveraged finance, distressed debt, and bankruptcy", John Wiley & Sons, 2019.

[9] Altman, E.I., Eom, Y.H. & Kim, D.W., "Failure prediction: evidence from Korea", Journal of International Financial Management & Accounting, Vol. 6, No. 3, 1995, pp.230-249.
[10] Zhou, L., "Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods", Knowledge-Based Systems, Vol. 41, 2013, pp.16-25.
[11] 이인로, 김동철, "회계정보와 시장정보를 이용한 부도예측모형의 평가 연구", 재무연구, 제28권, 제4호, pp.625-665, 2015.
[12] Altman, E.I., Esentato, M. & Sabato, G., "Assessing the credit worthiness of Italian SMEs and mini-bond issuers", Global Finance Journal, Vol. 43, 2020, p.100450.
[13] Kim, H.J., Jo, N.O. & Shin, K.S., "Optimization of cluster-based evolutionary undersampling for the artificial neural networks in corporate bankruptcy prediction", Expert systems with applications, Vol. 59, 2016, pp.226-234.
[14] Beaver, W.H., "Financial ratios as predictors of failure", Journal of accounting research, Vol. 4, 1966, pp.71-111.
[15] Ohlson, J.A., "Financial ratios and the probabilistic prediction of bankruptcy", Journal of accounting research, Vol. 18, No. 1, 1980, pp.109-131.
[16] Zmijewski, M.E., "Methodological issues related to the estimation of financial distress prediction models", Journal of Accounting research, Vol. 22, 1984, pp.59-82.
[17] Kumar, P.R. & Ravi, V., "Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review", European journal of operational research, Vol. 180, No. 1, 2007, pp.1-28.
[18] Merton, R.C., "On the pricing of corporate debt: The risk structure of interest rates", The Journal of finance, Vol. 29, No. 2, 1974, pp.449-470.
[19] Shumway, T., "Forecasting bankruptcy more accurately: A simple hazard model", The journal of business, Vol. 74, No. 1, 2001, pp.101-124.
[20] Vuttipittayamongkol, P. & Elyan, E., "Neighbourhood-based undersampling approach for handling imbalanced and overlapped data", Information Sciences, Vol. 509, 2020, pp.47-70.
[21] Susan, S. & Kumar, A., "SSOMaj-SMOTE-SSOMin: Three-step intelligent pruning of majority and minority samples for learning from imbalanced datasets", Applied Soft Computing, Vol. 78, 2019, pp.141-149.
[22] Chawla, N.V., Bowyer, K.W., Hall, L.O. & Kegelmeyer, W.P., "SMOTE: synthetic minority over-sampling technique", Journal of artificial intelligence research, Vol. 16, 2002, pp.321-357.
[23] Nekooeimehr, I. & Lai-Yuen, S.K., "Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets", Expert Systems with Applications, No. 46, 2016, pp.405-416.
[24] Gong, J. & Kim, H., "RHSBoost: Improving classification performance in imbalance data", Computational Statistics & Data Analysis, Vol. 111, 2017, pp.1-13.
[25] Veganzones, D. & Séverin, E., "An investigation of bankruptcy prediction in imbalanced datasets", Decision Support Systems, Vol. 112, 2018, pp.111-124.
[26] Zoričák, M., Gnip, P., Drotár, P. & Gazda, V., "Bankruptcy prediction for small-and medium-sized companies using severely imbalanced datasets", Economic Modelling, Vol. 84, 2020, pp.165-176.

저자소개

송 현 준(Hyunjun Song)
2020년 서울사이버대학교 경영학과 졸업(학사), 2020년~현재 연세대학교 정보대학원 비즈니스 빅데이터 분석 석사과정 재학, 주요 관심분야는 빅데이터 분석, 머신러닝, 자연어처리, 계산금융 등이다.

박 도 준(Dojoon Park)
1997년 중앙대학교 경제학 졸업(학사), 2002년 한국과학기술원 금융공학 졸업(석사), 2018년 연세대학교 경영학 박사를 취득하였다. 2020년~현재 연세대학교 경영대학 연구교수로 재직, 주요 관심분야는 자산가격이론, 금융경제학, 머신러닝 등이다.

이 준 기(Zoonky Lee)
1985년 서울대학교 계산통계학 졸업(학사), 1991년 카네기멜론대학 사회심리학 졸업(석사), 1999년 남가주대학교 경영정보학 박사를 취득하였다. 2004년~현재 연세대학교 정보대학원 교수로 재직, 주요 관심분야는 빅데이터 분석, 디지털 트랜스포메이션, 개방형 협업 등이다.