ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2020 June: 26(2): 105~129 http://dx.doi.org/10.13088/jiis.2020.26.2.105

머신러닝 기반 기업부도위험 예측모델 검증 및 정책적 제언: 스태킹 앙상블 모델을 통한 개선을 중심으로

엄하늘
고려대학교 과학기술학 협동과정
(eomsky49@korea.ac.kr)

김재성
고려대학교 과학기술학 협동과정
(mdmstyle@msn.com)

최상옥
고려대학교 행정학과
(sangchoi@korea.ac.kr)

본 연구는 부도위험 예측을 위해 K-IFRS가 본격적으로 적용된 2012년부터 2018년까지의 기업데이터를 이용한다. 부도위험의 학습을 위해, 기존의 대부분 선행연구들이 부도발생 여부를 기준으로 사용했던 것과 다르게, 본 연구에서는 머튼 모형을 토대로 각 기업의 시가총액과 주가 변동성을 이용하여 부도위험을 산정했으며, 이를 통해 기존 방법론의 한계로 지적되어오던 부도사건 희소성에 따른 데이터 불균형 문제와 정상기업 내에서 존재하는 부도위험 차이 반영 문제를 해소할 수 있도록 하였다. 또한, 시장의 평가가 반영된 시가총액 및 주가 변동성을 기반으로 부도위험을 도출하되, 부도위험과 매칭될 입력데이터로는 비상장 기업에서 활용될 수 있는 기업 정보만을 활용하여 학습을 수행함으로써, 포스트 팬데믹 시대에서 주가 정보가 존재하지 않는 비상장 기업에게도 시장의 판단을 모사하여 부도위험을 적절하게 도출할 수 있도록 하였다. 기업의 부도위험 정보가 시장에서 매우 광범위하게 활용되고 있고, 부도위험 차이에 대한 민감도가 높다는 점에서 부도위험 산출 시 안정적이고 신뢰성 높은 평가방법론이 요구된다. 최근 머신러닝을 활용하여 기업의 부도위험을 예측하는 연구가 활발하게 이루어지고 있으나, 대부분 단일 모델을 기반으로 예측을 수행한다는 점에서 필연적인 모델 편향 문제가 존재하고, 이는 실무에서 활용하기 어려운 요인으로 작용하고 있다. 이에, 본 연구에서는 다양한 머신러닝 모델을 서브모델로 하는 스태킹 앙상블 기법을 활용하여 개별 모델이 갖는 편향을 경감시킬 수 있도록 하였다. 이를 통해 부도위험과 다양한 기업정보들 간의 복잡한 비선형적 관계들을 포착할 수 있으며, 산출에 소요되는 시간이 적다는 머신러닝 기반 부도위험 예측모델의 장점을 극대화할 수 있다. 본 연구가 기존 머신러닝 기반 모델의 한계를 극복 및 개선함으로써 실무에서의 활용도를 높일 수 있는 자료로 활용되기를 바라며, 머신러닝 기반 부도위험 예측 모형의 도입 기준 정립 및 정책적 활용에도 기여할 수 있기를 희망한다.

주제어 : 부도위험 예측, 스태킹 앙상블 모델, 머튼 모형, 랜덤 포레스트, 합성곱 신경망

논문접수일 : 2020년 4월 30일 논문수정일 : 2020년 6월 22일 게재확정일 : 2020년 6월 26일
원고유형 : 일반논문 교신저자 : 최상옥

1. 서론

기업의 부도위험을 올바르게 판단하는 것은 시장의 투명성을 확보하고 건전한 시장 경제를 달성하기 위해 매우 중요하다. 강건한 기업은 낮은 금리로 자금을 조달하여 성장동력을 확보할 수 있게 되고, 부실기업은 각고의 노력을 통해 경영상태를 개선하도록 만들 수 있기 때문이다. 여기서, 부도는 원리금의 적기상환이 이루어지지 않거나 기업회생절차, 파산절차의 개시가 있는 경우를 의미하며, 부도위험은 부도가 발생할 가능성을 의미한다고 정의할 수 있다. 현재 부도위험은 신용평가사의 신용등급 형태로 측정 및 제공되고 있지만, 신용등급은 갱신 주기가 길고

비상장 및 초기 단계 기업에 적용이 어려우며, 신용 사건이 발생되고 나서야 등급이 조정되기도 한다는 점에서 후행적이라는 비판도 존재하고 있다. 이에 따라, 신용평가사에서 제공하는 신용등급과는 별개로 부도위험에 대한 적시적이고 객관적인 평가 방법론이 요구되고 있고, 특히, 자본 조달이 절실하게 요구되지만 기존 신용평가 체계 상에서 제약이 존재하는 비상장 기업들에 대해 그 필요성이 높게 나타나고 있다.

부도위험에 대한 예측 방법은 1960년대부터 판별분석, 로짓 및 프로빗모형 등의 통계적 방법을 중심으로 발달해왔다. 하지만 이들 모형은 회계적 재무정보를 기반으로 부도위험을 산정한다는 점에서 시장에서 평가하고 있는 부도위험을 충분히 반영하지 못하고 있다는 한계가 존재했다. 이에 비해 주가 및 시가총액에 기반하여 내재부도율을 추정하는 Merton 모델은 시장가치를 반영하고 기업의 자본구조를 적용할 수 있다는 측면에서 합리적인 방법론으로 평가받고 있다. 특히, 코로나 19 팬데믹과 같은 외적 충격이 기업의 주가를 통해서 반영될 수 있다는 점에서 Merton 모델의 유용성을 확인할 수 있다. 이와 별도로 최근에는 부도사건 자체와 재무정보 등의 관계를 머신러닝을 통해 파악하고 이를 토대로 부도를 예측하는 연구가 활발하게 이루어지고 있다. 하지만, 부도사건이 매우 희소하다는 점에서 오버샘플링이나 언더샘플링이 필수적으로 요구되고 있고, 이는 정보 왜곡을 발생시키는 요인으로 작용할 가능성이 높다. 또한, 부도 여부만을 기준으로 부도위험을 산출한다는 점에서 시장에서 요구하고 있는 각 기업별 세밀한 부도위험 차이를 충분히 구분하지 못하고 있다는 한계가 존재했다. 그리고, 대부분의 머신러닝 기반 연구가 특정한 예측모델에 기반하고 있다는 점에서 모델 자체적으로 존재하고 있는 편향이 예측에 반영될 위험도 존재하고 있다. 본 연구에서는 기업의 모든 정보와 시장의 판단이 반영되어 있는 주가 및 시가총액에 기반하여 부도위험을 도출하고, 산출된 부도위험과 기업의 다양한 재무정보들과의 관계를 복수의 머신러닝 기반 예측모델들로 파악한 뒤, 각 모델들을 스태킹 앙상블 학습 모형으로 결합하여 모델이 가질 수 있는 편향 위험을 경감시켰다. 이에 따라, 시장정보가 반영된 주가 및 시가총액에 기반한 부도위험과 각 재무제표 및 재무비율 지표들간의 복잡한 비선형적 관계들을 포착할 수 있으며, 이를 토대로 주가 정보가 존재하지 않는 비상장 기업에게 시장의 판단을 모사하여 부도위험을 적절하게 도출할 수 있다. 본 연구는 머신러닝 기반 부도위험 예측모형의 안정적인 예측력 확보를 위한 방안을 제안하며, 본 연구가 머신러닝 기반 부도위험 예측 모형의 도입 기준과 정책 수립에도 활용될 수 있도록 한다.

2. 선행 연구

2.1 재무적 부도위험 예측 연구

부도위험 예측에 대한 연구는 1960년대 후반부터 본격적으로 시작되었다. Horrigan은 재무비율과 채무변제 우선순위를 독립변수로 두고 신용상태를 종속변수로 두어 다중회귀분석을 수행하였으며(Horrigan, 1966), Altman은 운전자본/총자산, 이익잉여금/총자산, 영업이익/총자산, 자기자본 시장가치/총부채의 장부가치, 매출액/총자산 5개 재무비율을 이용하여 부도위험 예측을 위한 판별함수를 도출하였다(Altman, 1968). Altman

의 판별분석 기반 Z-Score 모형을 개선하기 위해 Ohlson은 로짓분석(Ohlson, 1980)을 Zmijewski는 프로빗모형(Zmijewski, 1984)을 사용하였다. 국내 연구로 다중판별분석과 로지스틱 회귀분석을 사용하여 부도위험에 영향을 미치는 6개의 재무비율을 도출한 연구(Jeon, 1986)와 우량 제조업과 우량 기업으로 표본을 구분하여 Altman 모형을 개선한 kems1과 kems2 모형을 제시한 연구가 있으며(Jo, 1998), Jeong은 재무정보 기반의 부도예측모형을 적용하기 위해서는 재무정보에 대한 적시성 및 투명성 확보를 전제되어야 하고, 미래 도산 가능성 예측이 과거 추세가 미래에 지속될 것이라는 가정이 필요하다는 것을 제시했다(Jeong, 2002). 이러한 재무정보에 기반한 부도 예측모형의 한계를 극복하기 위한 대안으로 자본시장의 정보를 이용하는 부도위험 예측 모형이 활발히 연구 및 활용되고 있다. 특히, 부채비율이 높고 신용등급이 낮을수록 Merton 모형의 유용성이 높은 것으로 나타났다(Byun, 2004). Kang은 배리어 옵션을 사용하여 주어진 만기 이전에 대한 부도 발생을 포함한 예상 부도율을 산출하는 방법을 제시하였으며, 부도예측 모형의 모수추정 시 2단계 반복 갱신법을 적용할 때 정확도가 향상함을 보였다(Kang, 2011; Kang, 2014).

2.2 머신러닝 기반 부도 예측 연구

머신러닝 기반 부도 예측 시 MDA와 인공신경망 모형을 휴리스틱한 방법으로 결합한 부도예측모형이 높은 예측력을 가진 것으로 나타났으며(Lee, 1993), 중소기업의 재무정보 불충분성에 따라 비재무정보를 포함한 인공신경망 기반 부도예측 모형도 제시되었다(Lee, 1995). 기업부실 예측데이터의 불균형 정도에 관계없이 GM-Boost (Geometric Mean-based Boosting) 및 MGM-Boost(Multiclass GM-Boost) 알고리즘이 우수한 예측력을 보이며, 유전자 알고리즘을 이용한 SVM 앙상블 모델이 기존의 방법론보다 부도 예측력이 더 우수한 것으로 나타났다(Kim, 2009; 2010; 2012). 기업부도예측을 위해 통계적 방법과 인공지능 방법을 Voting 알고리즘으로 결합한 모형이 제안되었으며(Bae, 2010), 랜덤포레스트 모형이 다변량 판별분석, 인공신경망, MSVM 모형보다 부도예측력이 뛰어난 것이 확인되었다(Kim, 2014). Wang은 신용위험 평가 시 Lasso-logistic regression learning ensemble 모형이 효과적임을 보였으며, 불균형 데이터 상에서 신용위험 예측 모델로 Lasso-logistic regression learning ensemble이 CART, LLR, RF보다 우수함을 확인했다(Wang, 2015). 부도위험 예측에 있어. 사례선택을 활용한 배깅(Bagging) 모형이 기존 SVM보다 예측력이 뛰어난 것으로 나타났으며(Min, 2014), 두 개의 인공신경망 모델을 이용하는 하이브리드 모형을 토대로 부도 예측을 제안한 연구(Cho, 2015; 2016)와 딥러닝 시계열 알고리즘인 RNN과 LSTM 기반의 부도예측모형이 다른 알고리즘에 비해 성능이 우수함을 확인한 연구가 제시되었다(Cha, 2018).

3. 분석방법론

3.1 부도위험 기준 설정

본 연구에서 부도위험 기준으로 사용하는 머튼(Merton) 모형은 회사가치가 식 (1)과 같은 기하브라운운동(Geometric Brownian Motion)을 따른다고 가정한다. 여기서 W_t는 표준 브라운 운

동, σ_A는 자산의 변동성. μ는 자산의 수익률을 의미한다. 동 모형에 의하면 주식 가치는 다음 식(2)과 같이 표현될 수 있다. 동 모형은 부도가 기업의 자산가치와 부채수준에 의해 결정되며, 주식가치를 옵션가격결정모형으로 산정할 수 있다고 가정하며, 기업의 부도확률을 부채변제 만기 시점에서 자산가치가 부채가치보다 적을 확률로 계산한다. 동 계산과정에서 자산가치(V)와 자산가치의 변동성(σ_V)의 경우 연립방정식 (3)에 의거 반복적 과정을 거쳐 계산한다. 이의 계산을 위해서는 시장에 알려진 주식가치 E와 주식가치 변동성 σ_E로부터 자산가치 V와 자산가치변동성 σ_V를 알아내는 것이 필요하며, 이를 위해 Crosbie(2003), Vassalou(2004) 연구에서 소개된 반복적 알고리즘(iterative procedure)을 활용한다.

$$\frac{dV}{V} = \mu dt + \sigma_A dW_t \tag{1}$$

$$E = VN(d_1) - e^{-rT}F(d_1) \tag{2}$$

$$d_1 = \frac{\ln(V/F) + (r + 0.5 \times \sigma_V^2)}{\sigma_V\sqrt{T}}$$

$$d_2 = d_1 - \sigma_V\sqrt{T} \tag{3}$$

$$\sigma_E = \frac{V}{E}N(d_1)\sigma_V$$

V: 자산가치(주식가치 + 부채가치)
F: 부채가치 = (유동부채 + 고정부채*0.5)
r: 무위험이자율 (국고채 3년 수익률)
σ_V: 자산가치 변동성
σ_E: 주식가치변동성

자산가치 V와 자산가치변동성 σ_V를 추정해 내게 되면 Merton모형의 내재부도확률(implied probability of default) π은 다음 식(4)를 이용하여 구할 수 있다.

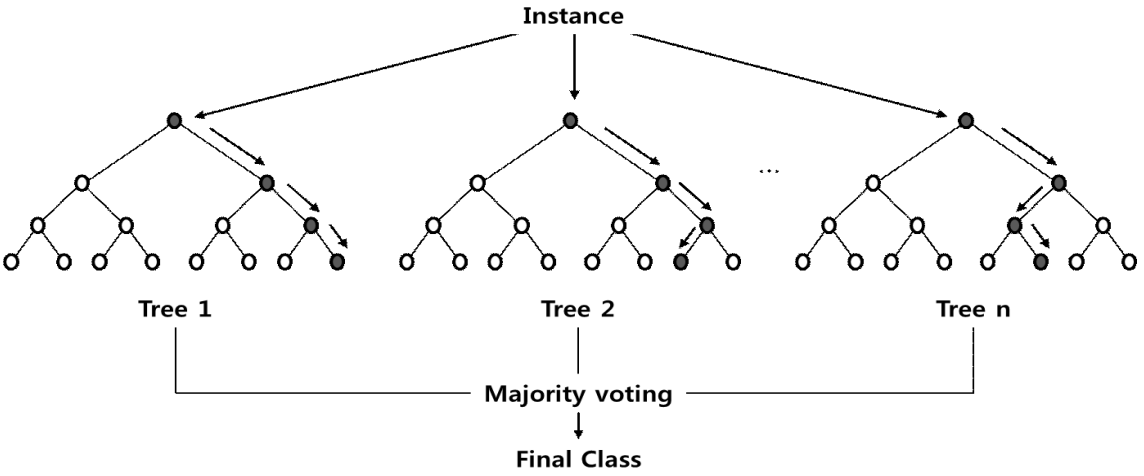
$$\pi = N\left(-\left(\frac{\ln(V/F) + (r - 0.5 \times \sigma_V^2)}{\sigma_V\sqrt{T}}\right)\right) \tag{4}$$

3.2 부도위험 서브(Sub) 예측모델 설정

3.2.1 랜덤 포레스트(Random Forest)

랜덤 포레스트는 다수의 결정 트리들을 학습하는 방법으로 랜덤 노드 최적화와 배깅을 결합한 방법과 같은 분류회귀트리(CART)를 사용해 상관관계가 없는 트리들로 포레스트를 구성한다. 여기서, CART는 설명변수 또는 예측 인자의 비선형성(nonlinearity)과 상호작용(interactions)을 최대한 활용하여 반응 변수(종속변수)에 대한 영향을 판단하는 기법으로, 설명변수를 중요도 기준에 따라 줄기(branch)를 만들어 나가며, 마지막 노드(node: 마디)에서 반응변수에 대해 판단을 내린다. CART는 반응변수가 이항변수, 다항변수, 그리고 연속형인 경우에도 사용할 수 있으며, 예측인자(설명변수) 역시 연속형과 범주형 변수의 구분 없이 선택 할 수 있다. 랜덤 포리스트 역시 CART와 마찬가지로 반응변수가 범주형과 연속형 모두에 사용할 수 있다.

트리를 만드는 방법에 있어서도 CART와 동일한 알고리즘을 사용한다. 하지만 CART의 경우 하나의 결정트리가 도출되는 반면에 랜덤 포리스트에서는 수많은 결정트리로 구성된 숲(forest)을 형성하는 과정을 거치는 것이 다른 점이다. 다수의 결정트리를 만들기 위해 예측인자와 관측치에 대한 무작위 샘플링을 반복하게 된다. 수많은 결정트리에서 예측범주를 얻은 후 다수결투표



〈Figure 1〉 Random Forest Method

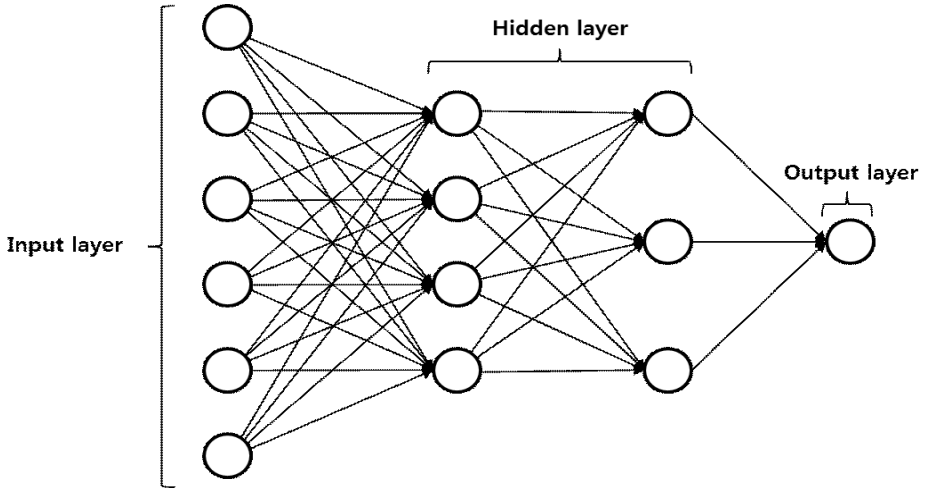
방식으로 최종 범주예측을 결정한다. 결정트리 형성에 무작위성을 부여함으로써 독립적인 결정트리를 반복 적으로 만들 수 있으며, 이 방식을 통해 예측오차를 줄일 수 있다. 예측인자와 관측치의 무작위 선택에는 붓스트래핑(bootstrapping) 기법이 사용된다. CART의 경우, 하위 노드가 많아질수록 예측오차의 편의(bias)는 줄어들지만 분산은 증가하는 문제가 있다. 반면, 랜덤 포리스트에서는 동일하게 분포 된 결정트리를 반복적으로 생성함으로써 예측오차의 분산을 줄일 수 있다.

3.2.2 다층 퍼셉트론(Multiple Layers Perceptron) 모델

다층 퍼셉트론은 투입층(input layer), 은닉층(hidden layer), 출력층(output layer)으로 구성되어 있는 신경망 모형이다. 인공신경망에서 은닉층을 다수 구성해서 분석하는 방법론은 다층 퍼셉트론(Multiple Layer perceptron; MLP)으로 부른다. 이러한 다층 퍼셉트론 모형은 패턴 분류, 인식, 예측에 있어서 널리 사용되는 분석 방법론으로 은닉층의 형태 및 활성함수에 따라서 보다 고도화된 인공신경망 분석으로 확장된다. 이는 기본적인 하나의 투입층과 은닉층으로 구성되어 있는 단층신경망 분석에서 점차 고도화된 알고리즘의 형태라고 볼 수 있다.

Figure 2을 기본적인 수식으로 나타내면 식 (5)와 같은 형태로 나타낼 수 있다. 여기서 v_i는 투입층 혹은 이전 은닉층 신호이고 b_j와 b_k는 은닉층과 출력층의 편향(bias)을 나타낸다. 그리고 w_{ij}와 w_{jk}는 각각 은닉층과 출력층의 계수값을 의미한다. f는 활성함수(activation function)를 나타내는데, Sigmoid, ReLU 함수 등이 보편적으로 사용된다. 식 (5)의 경로를 통해서 출력층의 결과값 (Y_k)을 얻을 수 있다.

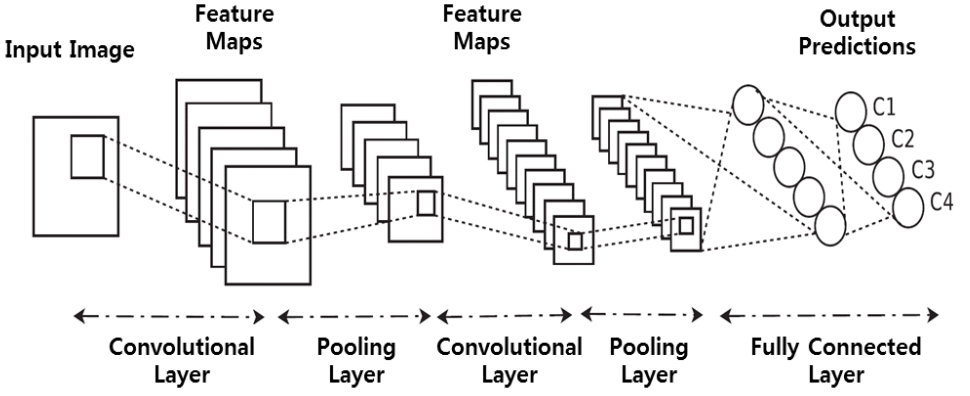
$$Y_k = \sum_{j=1}^{m} f\left(\sum_{i=1}^{n} v_i w_{ij} + b_j\right) w_{jk} + b_k \qquad (5)$$



〈Figure 2〉 Multi-layer Perceptron (MLP) Method

3.2.3 합성곱 신경망(Convolution Neural Network) 모델

합성곱 신경망(CNN)은 이미지 식별 분야에서 괄목할 만한 성능을 보여주고 있다. 이는 이미지의 일정부분만을 인식하는 값을 필터를 통해서 새로운 값을 만들어주고 이 값들을 지속적으로 연결시켜줘서 이미지의 특징을 보다 잘 이해할 수 있도록 하는 방법론이다. 이러한 CNN은 입력층과 필터의 합성곱(Convolution)을 이용하는 방법론으로 하나의 입력계층과 출력계층, 하나 이상의 합성곱 계층(Convolution layer)과 풀링 계층(Pooling layer)로 구성되어 있다. 입력층을 통해 입력하고자 하는 데이터를 입력하고 합성곱 계층을 통해 필터링되어 적절한 특징을 추출한다. 이 때 필터의 개수에 따라 특징맵의 수가 정해진다. 예시로 아래의 Figure 3와 같이 학습을 통해



〈Figure 3〉 Convolutional Neural Networks (CNN) Method

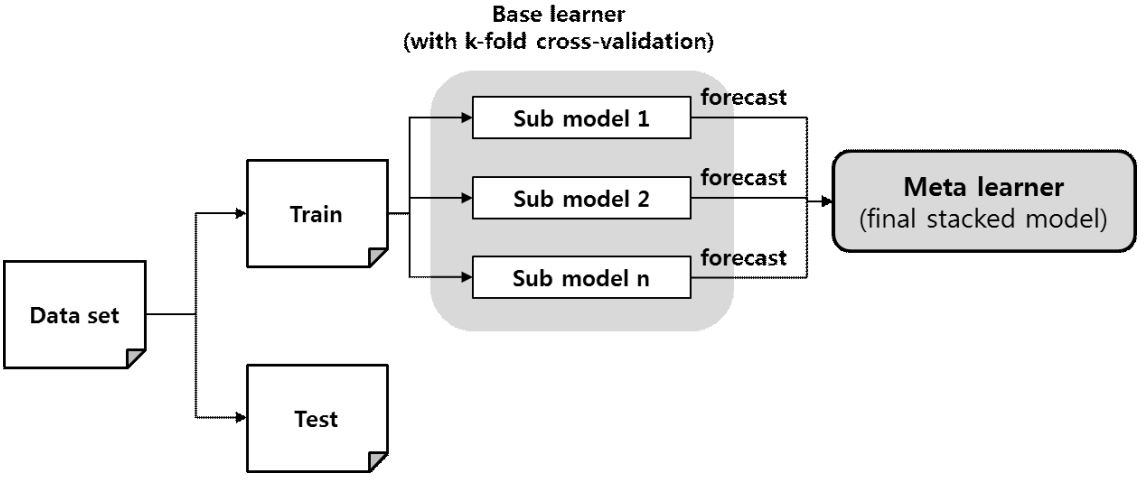
변화되는 가중치를 가지고 있는 합성곱 필터가 입력층에 대해 왼쪽에서 오른쪽으로 위에서 아래로 전체 영역을 훑고 지나가면서, 가중치를 곱하여 합한 결과들이 출력층의 출력값이 된다.

3.3 부도위험 서브 예측모델 통합: 모델 스태킹(Model Stacking)

모든 상황에서 우수한 성능을 보일 수 있는 단일한 모델은 존재하지 않는다. 본 연구에서 사용하는 스태킹 앙상블(Stacking Ensemble) 모델은 서로 다른 모델들을 조합해서 최고의 성능을 내는 모델을 생성한다. 스태킹 모델은 다양한 알고리즘을 조합하여 구성할 수 있으며, 이러한 조합을 통해서 각 알고리즘의 장점을 취하면서 약점을 보완할 수 있다. 이를 통해, 일반적으로 학습하는 단일 모델보다 성능을 향상시킬 수 있다. 본 연구에서는 랜덤 포레스트, 다층퍼셉트론, 합성곱 신경망 모델을 이용하여 스태킹 앙상블 모델을 구성한다.

3.4 모델 비교

Table 1은 기업 부도위험 예측에 사용된 랜덤 포레스트, MLP, CNN, 스태킹 앙상블 모델의 주요 로직과 데이터 구조를 나타낸다. Model 1 ~ Model 3은 스태킹 앙상블 모델의 예측력을 비교하기 위한 대조군이며, Model 4에 사용된 Base learner는 Model 1 ~ Model 3의 로직과 동일하되 트레이닝 데이터를 7개로 나눈 세트에서(7-fold) 훈련 및 테스트된 모델들이다. 여기서 Meta learner는 각 fold의 테스트 집합에서 산출된 Base learner들의 예측치를 토대로 훈련된다. 이때, Base learner들의 예측치를 결합시키는 알고리즘으로는 MLP을 사용했다.



〈Figure 4〉 Stacking Ensemble Model

〈Table 1〉 Model comparison

Division	Model 1	Model 2	Model 3	Model 4
	Random Forest (RF)	MLP	CNN	Stacking Ensemble
Main logic	Decision tree	Neural network	Neural network	Combining predictions using MLP
Input data for training	Training set	Training set	Training set	Base learner (RF, MLP, CNN) prediction results using segmented training set (7-fold)
Input data for testing	Test set	Test set	Test set	Base learner prediction results in test set
Training data dimension	(7382, 160)	(7382, 160)	(7382, 160)	(7382, 4)
Test data dimension	(3163, 160)	(3163, 160)	(3163, 160)	(3163, 4)

4. 실증 분석

4.1 데이터 설명

분석을 위해 2,194개 상장기업을 대상으로 K-IFRS가 정착된 2012년부터 2018년까지 연도별 기업 데이터를 토대로 분석을 진행했다. 종속변수인 부도위험 변수는 각 기업별/연도별 시가총액과 주가 변동성 정보를 기반으로 Merton의 모형을 사용하여 산출했다. 여기서, Merton 모형의 투입변수에 포함되는 자산가치와 자산변동성의 경우, 시가총액과 주식변동성을 토대로 수치최적화를 통한 추정이 요구되며, 이를 위해 파이썬의 Scipy 라이브러리를 이용하여 자산가치와 자산변동성을 산출하였다. 본 연구에서는 주요 선행연구에서 사용되고 있는 재무비율 데이터와 더불어 기업의 회계정보를 담고 있는 재무제표를 분석에 직접 포함시키도록 한다. 따라서, 설명변수로 기업의 재무적 정보를 파악할 수 있는 재무상태표와 포괄손익계산서, 현금흐름표의 주요 항목을 포함시키고, 기업의 안정성, 성장성, 수익성, 활동성을 나타내는 대표적인 재무비율 지표를 분석에 포함시켰다. 데이터는 시장정보 제공 업체인 에프앤가이드에서 제공받은 기업별 재무상태표 상 38개 항목, 포괄손익계산서 상 26개 항목, 현금흐름표 상 11개 항목과 더불어 안전성, 성장성, 수익성, 활동성, 상대가치 등을 판단할 수 있는 76개 재무비율 지표를 사용했으며, 연도별 Dummy 변수 7개, 타겟 변수인 부도위험 변수를 포함하여 총 160개의 변수를 사용했다. 여기서 연도별 Dummy는 외부환경 요인을 통제하기 위하여 사용되었다.

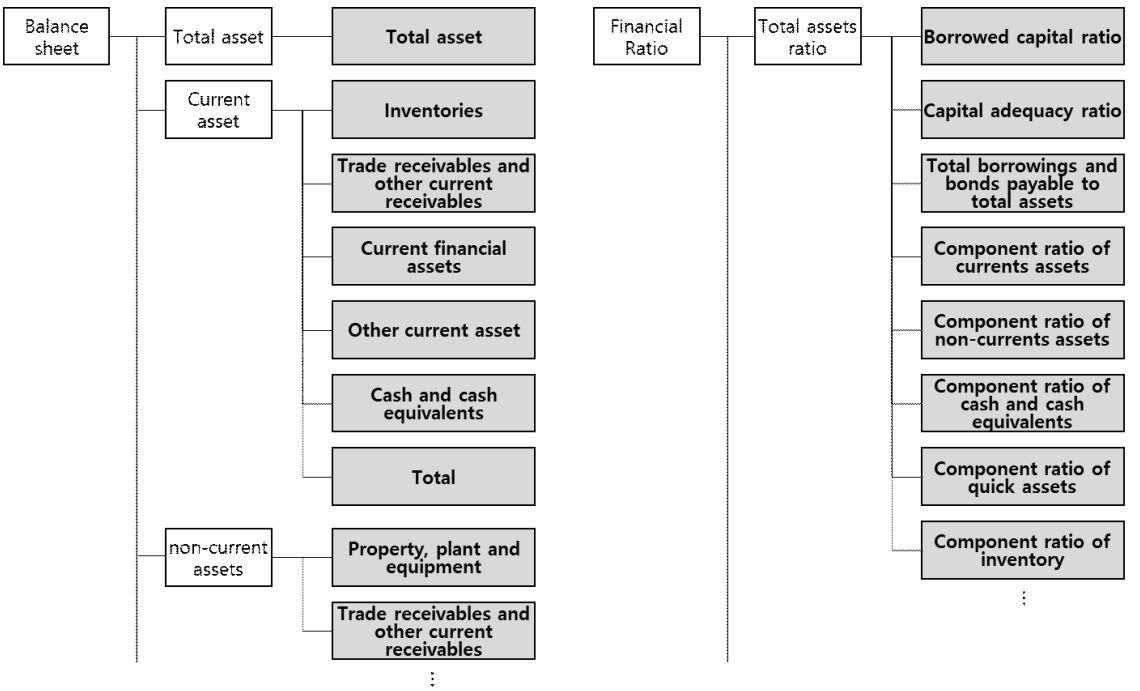
〈Table 2〉 Input-data descriptions

Division		Num of Variables	Division		Num of Variables
Balance sheet	Total asset	1	Financial Ratio	Total assets ratio	14
	Current asset	6		total capital ratio	5
	Non-current assets	8		asset liability ratio	5
	Total liabilities	1		Cashflow ratio	4
	Current liabilities	9		Profit-related ratio	3
	non-current liabilities	9		Balance item	3
	capital assets	1		Sales profit item	12
	capital surplus	1		Return on Capital	7
	Comprehensive Income	1		Turnover Related Items	9
	Retained Earnings	1		Period related items	2
State-ment of comprehensive income	Sales	1		EPS	1
	Cost of Goods sold	1		BPS	1
	Gross profit or loss	1		SPS	1
	S&A expenses	8		CFPS	1
	Operating Income/Loss	1		EBITDAPS	1
	Financial income	4		PBR	1
	finance costs	3		PSR	1
	Other Income	1		Beta	1
	Other costs	1		WACC	1
	ILBIT	1		Total Cashflow	1
	Income Taxes	1		CAPEX	1
	Net Income or loss	1		EBITDA	1
	Other income	1	Year dummy variables	2012	1
	Total income	1		2013	1
State-ment of cash flows	Operating activities	6		2014	1
	Investing activities	1		2015	1
	Financing activities	1		2016	1
	Cash & cash equivalents	1		2017	1
	Cash at beginning	1		2018	1
	Cash at end	1	Other	Number of employees	1
Target	Default risk				1
Total	Total number of variables				160

분석에 사용된 데이터는 행을 기준으로 총 10,545개(약 2,194개 기업을 대상으로 2012년부터 2018년까지의 연도별 데이터 사용, 누락은 제외)이며, 160개의 열로 구성된다.

Table 3은 부도확률, 시가총액, 현금흐름표, 종업원 수 외에 재무상태표의 자산/부채/자본 세부항목을 중분류 단위로 요약한 데이터에 대한 기술통계량을 나타낸다. Figure 6은 Table 3에 나타난 각 변수들에 대한 상관관계 분석 결과이다. 재무지표들 간에 높은 상관관계 경향성이 나타나는 것과 다르게, 부도확률(V1)은 대부분의 변수들과 강한 상관관계를 나타내지는 않는 것을 확인할 수 있다.

이는 부도확률이 재무지표가 아닌 주가, 변동성과 같은 시장의 판단을 토대로 도출되었고, 재무지표와 시장의 판단이 차이를 보이기 때문으로 해석할 수 있다. 부도확률과 비교적 양의 상관관계($\rho > 0.2$)를 보이는 변수로는 유동부채, 자본금, 기타비용이 있으며, 비유동부채, 비유동자산, 자본잉여금, 매출액 등은 이보다 약한 양의 상관관계($\rho > 0.1$)를 보이는 것으로 나타났다.

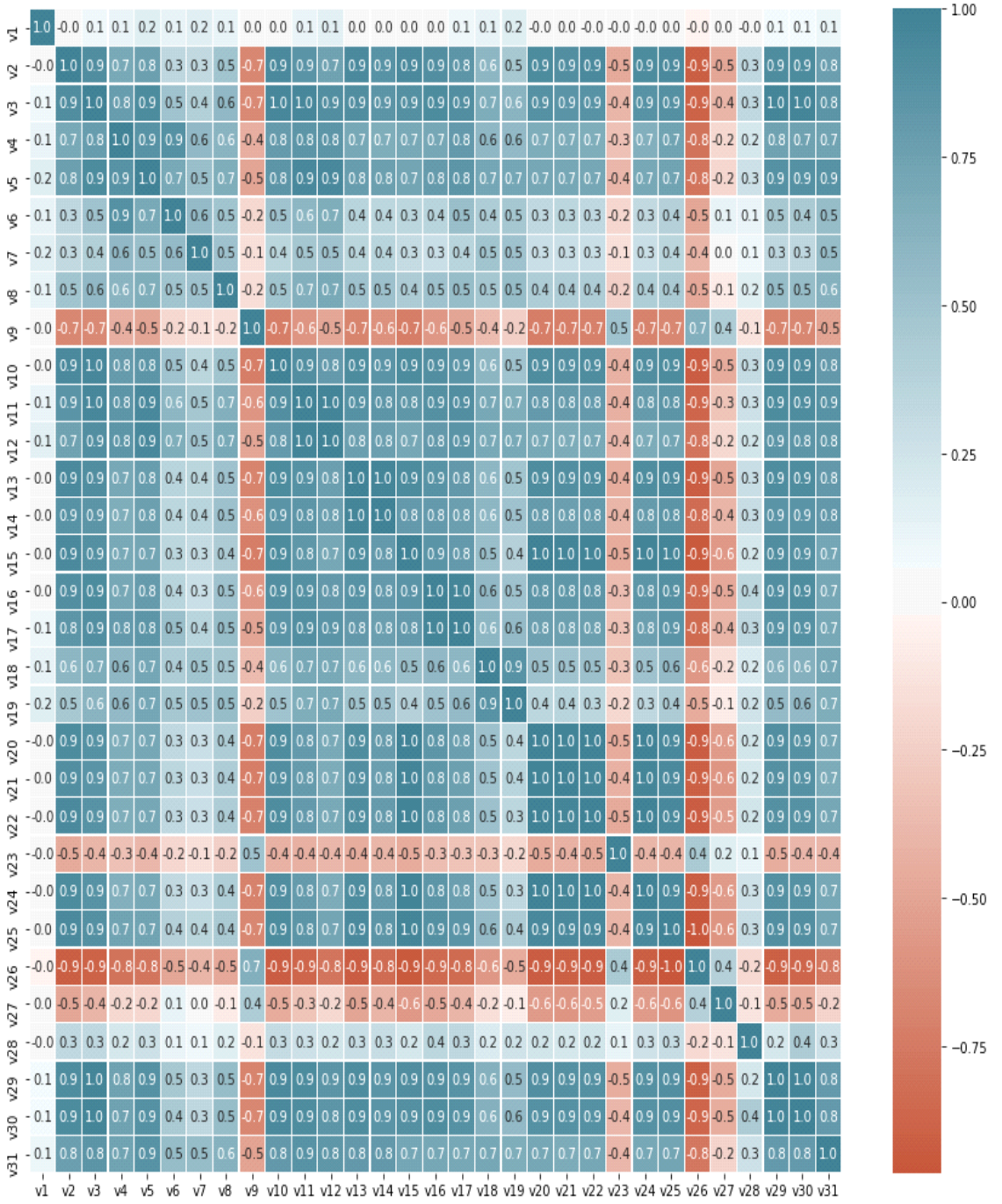


〈Figure 5〉 Input-data samples

〈Table 3〉 Descriptive statistics for representative variables

Division		Obs	Mean	Std. Dev.	Min	Max
V1	Default probability	10,545	0.34819	0.36766	0.0	1.0
V2	Market cap	10,545	761.0	6,550.0	3.3	329,000.0
V3	Current assets	10,545	598.0	4,070.0	0.9	175,000.0
V4	Non-current assets	10,545	926.0	6,240.0	0.1	166,000.0
V5	Current liability	10,545	458.0	2,440.0	0.1	69,100.0
V6	Non-current liability	10,545	348.0	2,940.0	0.0	92,300.0
V7	Capital stock	10,545	49.3	208.0	0.0	3,660.0
V8	Capital surplus	10,545	117.0	425.0	-574.0	7,060.0
V9	Accumulated Other Comprehensive Income	10,545	3.5	165.0	-7,990.0	1,940.0
V10	Retained earnings	10,545	530.0	5,450.0	-3,590.0	243,000.0
V11	Sales	10,545	1,280.0	7,720.0	0.0	244,000.0
V12	Cost of sales	10,545	1,000.0	5,610.0	0.0	138,000.0
V13	Gross profit	10,545	286.0	2,620.0	-1,580.0	111,000.0
V14	Selling, general and administrative expense	10,545	201.0	1,650.0	-273.0	56,600.0
V15	Operating income	10,545	84.9	1,090.0	-3,270.0	58,900.0
V16	Financial income	10,545	25.6	279.0	0.0	11,400.0
V17	Financial cost	10,545	34.5	289.0	0.0	10,700.0
V18	Other income	10,545	22.4	137.0	-0.6	4,220.0
V19	Other expense	10,545	28.1	157.0	0.0	4,420.0
V20	Continuing income and loss before income taxes	10,545	78.5	1,160.0	-4,060.0	61,200.0
V21	Tax expense	10,545	21.5	290.0	-985.0	16,800.0
V22	Net income	10,545	59.0	878.0	-3,080.0	44,300.0
V23	Accumulated other comprehensive income	10,545	-3.2	82.7	-5,500.0	1,990.0
V24	Total comprehensive income	10,545	55.8	842.0	-3,400.0	44,300.0
V25	Cash flow from operating activities	10,545	115.0	1,400.0	-3,460.0	67,000.0
V26	Cash flow from investing activities	10,545	-109.0	1,170.0	-52,200.0	4,330.0
V27	Cash flow from Financing activities	10,545	0.0	356.0	-15,100.0	8,380.0
V28	Cash and cash equivalents	10,545	5.4	168.0	-2,510.0	9,470.0
V29	Cash and cash equivalents at beginning of period	10,545	92.7	682.0	-5.4	32,100.0
V30	Cash and cash equivalents at end of period	10,545	98.1	738.0	-5.5	32,100.0
V31	The number of employees	10,545	876	3961	0	103011

[decimal point, one billion KRW, person]

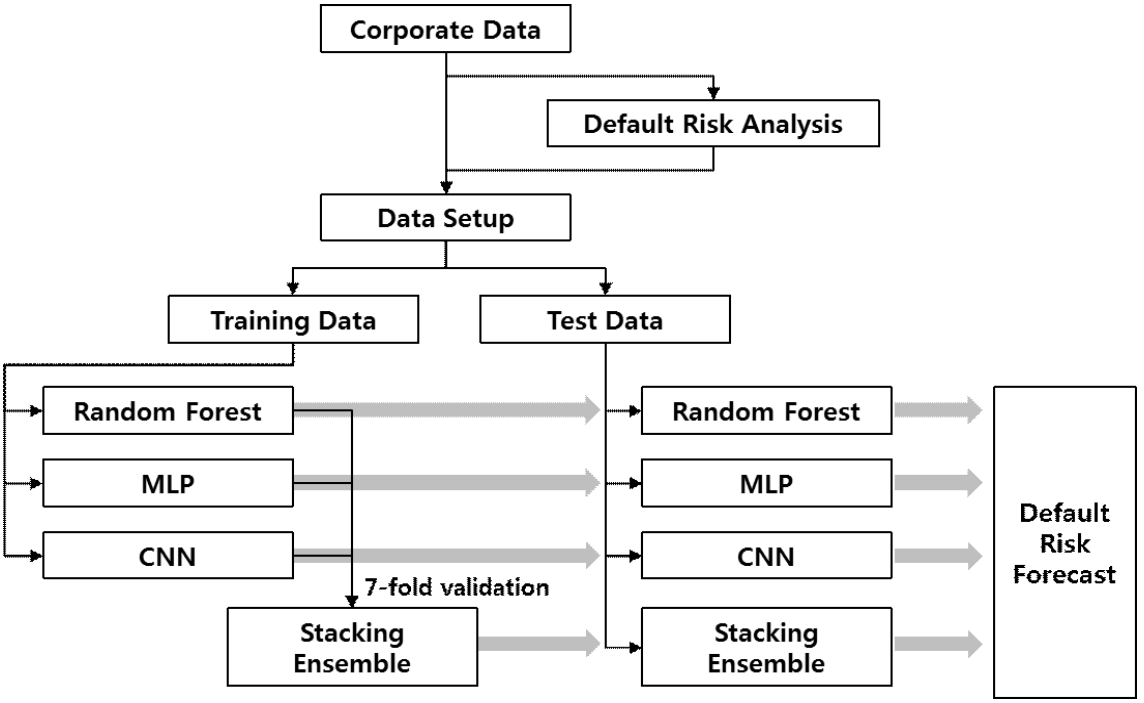


〈Figure 6〉 Correlation analysis results for representative variables

4.2 분석 모형

스태킹 앙상블 모델을 활용한 신용위험 예측 알고리즘은 Figure 7와 같이 구성된다. 분석대상 기업은 코스피 및 코스닥 상장기업이며, 기업별 부도위험을 산정하기 위해 각 기업별 시가총액과 주가 변동성 데이터를 에프앤가이드로부터 제공받았다. 머튼 모형으로 신용위험을 산출하기 위해서는 자산 가격과 자산 변동성 산출이 필요함에 따라, 시가총액과 주가 변동성 데이터를 토대로 관계식에 기반한 반복적 알고리즘을 통해 자산 가격과 자산 변동성을 산출했다. 부도거리 산정을 위해 KMV 방법론에 기반하여 부도거리를 산출하였으며, 자산 가격과 자산 변동성, 부도거리, 무위험이자율, 만기 정보를 블랙 숄즈 모형에 대입하여 부도위험을 산출했다.

모델 학습을 위한 데이터 중 타겟 값은 앞서 산출된 각 기업별 신용위험 산출값을 사용하며, 입력 데이터는 각 기업의 재무상태표, 포괄손익계산서, 현금흐름표외에 안전성, 성장성, 수익성, 활동성을 측정할 수 있는 재무비율 지표를 에프앤가이드로부터 제공받아 사용했다. 모델 트레이닝 및 검증을 위해, 전체 데이터를 무작위적으로 트레이닝 셋 70%, 테스트 셋 30%로 구분했으며, 트레이닝 셋에서는 스태킹 앙상블 모형 적용을 위해 별도의 7-fold 기반 트레이닝-검증 셋을 생성했다. 그리고, 7개의 트레이닝-검증 세트를 기반으로 랜덤 포레스트 서브 모델, MLP 서브 모델, CNN 서브 모델을 학습시켰다. 그리고 각 서브 모델의 예측치와 타깃값을 MLP 기반 스태킹 앙상블 모델에 학습시켰다.



〈Figure 7〉 Credit risk prediction algorithm

예측을 위한 서브모델에 대한 하이퍼 파라미터 설정값은 Table 4와 같다. 서브 모델로 랜덤 포레스트 모델은 생성할 트리의 개수로 1000개를 설정하였으며, 오차함수로는 MSE를 사용하였다. MLP 모델은 1개의 입력층과 2개의 은닉층 1개의 출력층으로 구성하였으며, 은닉층의 노드는 각각 100개, 옵티마이저로는 'Adam', 손실함수로는 MSE를, 출력함수는 확률값 출력을 위해 시그모이드 함수를 사용했다. CNN 모델은 1차원 CNN 모델을 사용했으며, 필터 사이즈는 2X2, Pool 사이즈는 2를 사용했으며, 출력함수는 확률값 출력을 위해 시그모이드 함수를 사용했다.

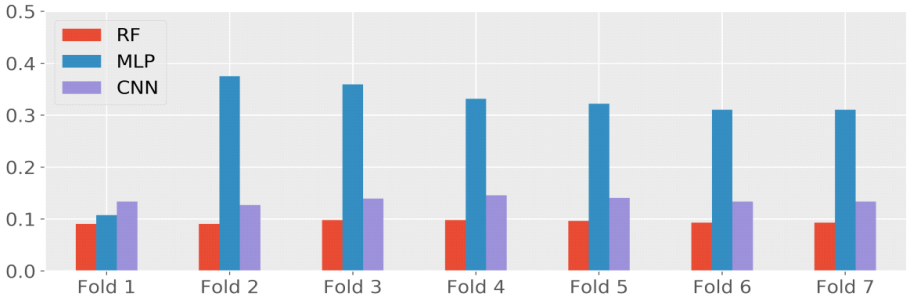
4.3 분석 결과

4.3.1 서브모델 구현 결과 및 모델 스태킹

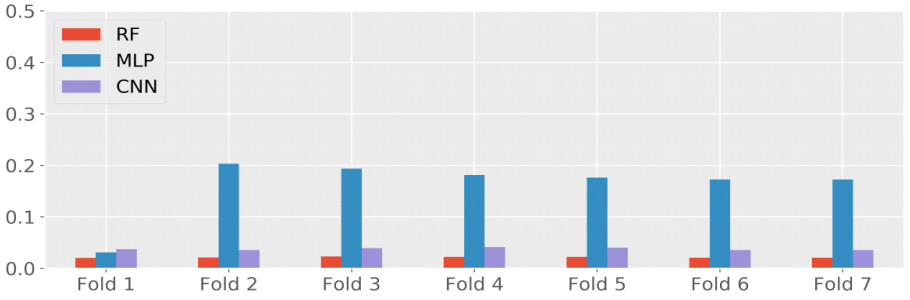
스태킹 앙상블 모형의 인풋데이터로 사용할 서브 모델 별 예측값 산출을 위해, 트레이닝 세트 상에서 7-fold로 훈련-검증데이터를 구성하고, 각 fold별로 서브 모델 훈련 및 예측값 산출을 수행했다. Figure 8 ~ Figure 10는 산출된 서브모델들의 각 fold 별 예측 오차결과이다. 트레이닝 세트 상에서는 랜덤 포레스트 모델이 가장 높은 정확도를 보였으며, 다음은 CNN, MLP 순으로 정확도가 높게 나타났다. 스태킹 앙상블 모델은 상

〈Table 4〉 Hyper-parameter setting results

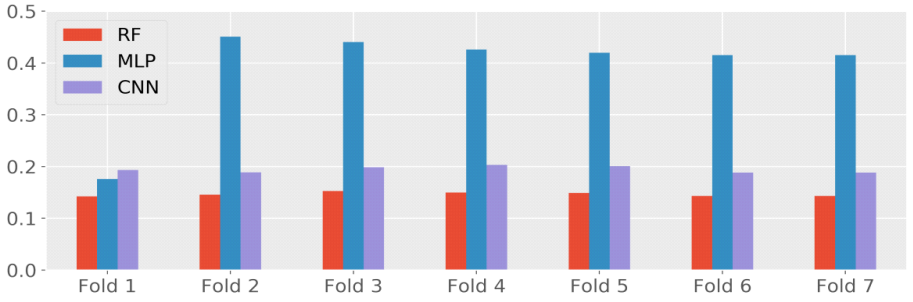
Division	Model	Hyper parameter	Value
Sub Model	Random forest	Number of trees	1000
		Loss function	MSE
		Random state	1
	MLP	Number of nodes	100
		Number of hidden layers	2
		Epoch	300
		Loss function	MSE
		Optimizer	Adam
	CNN	Dimension	1
		Filter size	2
		Kerner size	2
		Pool size	2
		Epoch	300
		Loss function	MSE
		Optimizer	Adam
Stacking Ensemble Model	MLP	Number of nodes	20
		Number of hidden layers	2
		Epoch	300
		Loss function	MSE
		Optimizer	Adam



〈Figure 8〉 Prediction errors of meta-learner models (MAE)



〈Figure 9〉 Prediction errors of meta-learner models (MSE)



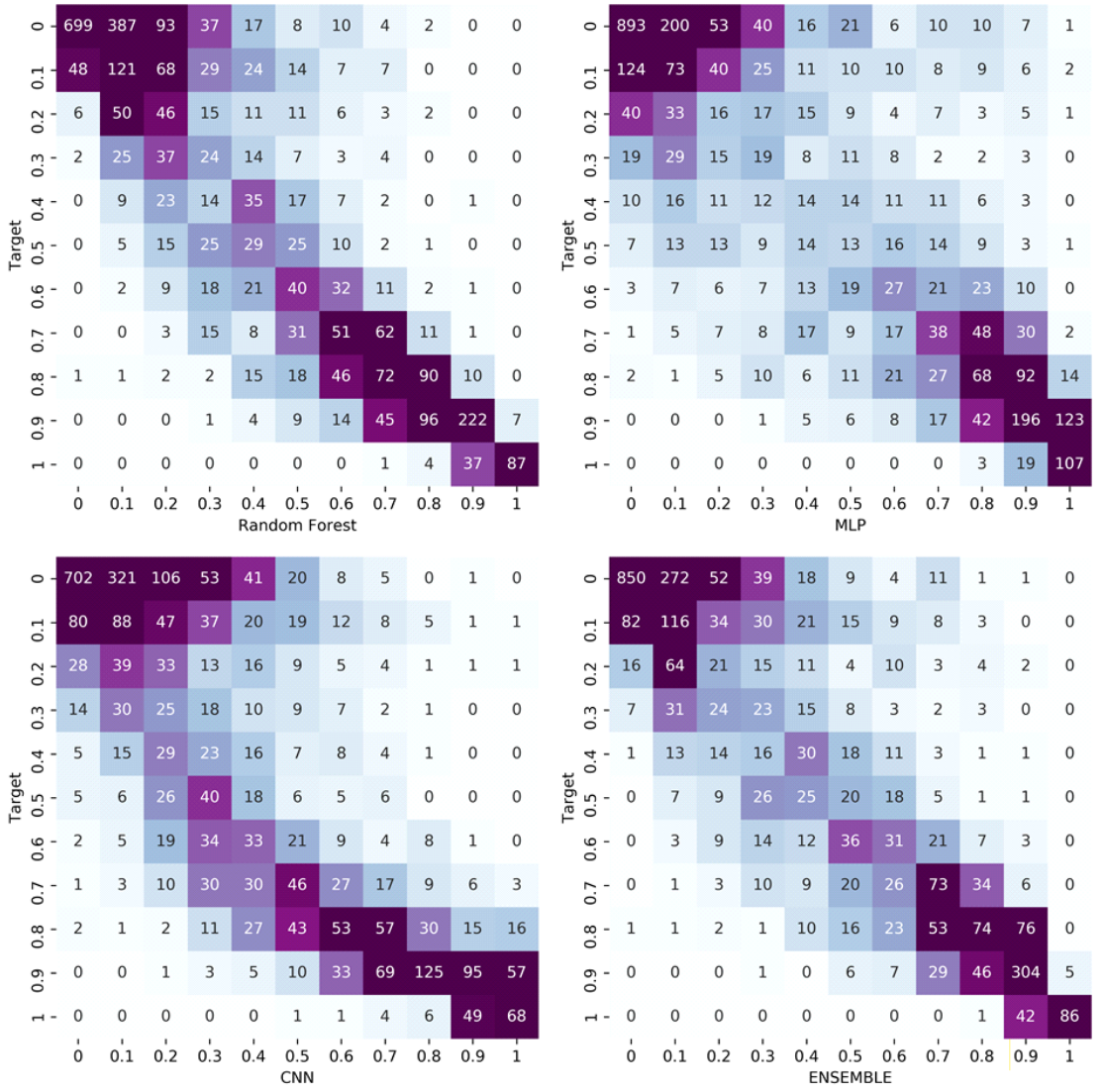
〈Figure 10〉 Prediction errors of meta-learner models (RMSE)

기 서브모델의 예측값과 실제 타깃값을 인풋데이터로 하며, 적은 노드로 세팅된 MLP를 기반으로 트레이닝을 실시하였다.

4.3.2 부도위험 예측결과

스태킹 앙상블 모델의 예측력을 비교하기 위해, 랜덤포레스트, MLP, CNN 모델을 전체 트레이닝 데이터로 훈련시킨 후, 테스트 셋 상에서 각 모델의 예측력을 검증했다. Figure 11는 각 모델이 신용위험을 얼마나 정확하게 예측했는지에 대한 빈도를 보여주고 있다. 여기서, 좌상단은 랜덤 포레스트 모델, 우상단은 MLP, 좌하단은

CNN, 우하단은 스태킹 앙상블 모형의 부도위험 예측 결과를 나타낸다. 각 그림에서 가로축은 모델이 산출한 예측값을 의미하며, 세로축은 실제 타깃값을 의미한다. 따라서, 우하향 대각선에 정중앙에 가까울수록 신용위험을 정확하게 예측한다고 해석할 수 있다. 스태킹 앙상블 모델과 랜덤 포레스트 모형에 비해서 MLP 모형과 CNN 모형은 상대적으로 예측력이 떨어지는 것을 확인할 수 있다.



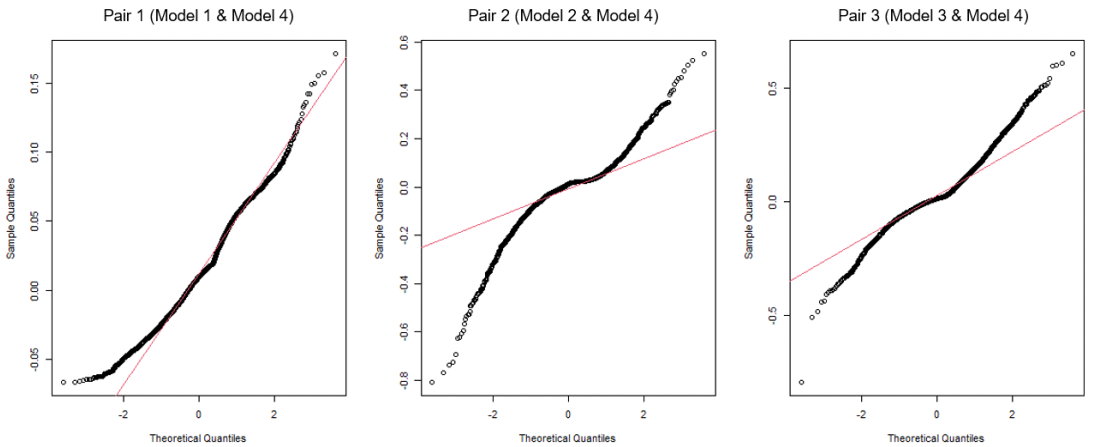
〈Figure 11〉 Prediction comparison (Horizontal axis: prediction value, Vertical axis: target value)

Table 5는 각 모델에 대한 오차 결과를 나타낸다. Mean Absolute Error(MAE), Mean Squared Error(MSE), Root Mean Squared Error(RMSE)를 기준으로 할 때, 스태킹 앙상블 모델이 가장 낮은 오차를 보이는 것을 확인할 수 있었다. 스태킹 앙상블 모형의 예측력은 MAE 기준 시, RF 모형의 1.044배, MLP 모형의 1.203배, CNN 모형의 1.486배 높은 것으로 나타났으며, MSE 기준으로는 RF 모형의 1.004배, MLP 모형의 1.626배, CNN 모형의 1.691배 높은 것으로 나타났다. RMSE 기준으로는 스태킹 앙상블 모형의 예측력이 RF 모형의 1.002배, MLP 모형의 1.275배, CNN 모형의 1.300배 우월한 것으로 나타났다.

다음으로 스태킹 앙상블 모델과 각 개별 모델의 예측값들이 통계적으로 유의미한 차이가 있는지 확인하기 위해, 스태킹 앙상블 모델과 각 개별모델 간 Pair를 구성하였다. Figure 12는 Pair를 구성하는 두 모델 예측값들의 차이가 정규성을 따르는지 확인하기 위한 Q-Q Plot이며, Table 6은 Shapiro-wilk normality test를 통한 정규성 검정 결과이다. Q-Q plot 상에서의 Pair 1, 2, 3 모두 붉은색 직선 바깥에 놓여져 있는 것을 확인할 수 있으며, Shapiro-wilk normality test의 결과에서도 모든 Pair가 정규성을 따르지 않는 것으로 나타났다.

〈Table 5〉 Results of prediction error

Division	Model 1	Model 2	Model 3	Model 4
	Random Forest	MLP	CNN	Stacking Ensemble
MAE	0.094751	0.111714	0.134892	0.09079
MSE	0.022263	0.036053	0.037492	0.022175
RMSE	0.149209	0.189877	0.193628	0.148912



〈Figure 12〉 Results of normality test using normal Q-Q plot

〈Table 6〉 Results of normality test (Shapiro-wilk normality test)

Division	Pair composition		Shapiro-wilk normality test		
	Set 1	Set 2	W	p-value	Results
Pair 1	Stacking Ensemble	Random Forest	0.98303	< 2.2e-16	Non-normal
Pair 2	Stacking Ensemble	MLP	0.89026	< 2.2e-16	Non-normal
Pair 3	Stacking Ensemble	CNN	0.95428	< 2.2e-16	Non-normal

〈Table 7〉 Results of wilcoxon rank sum test with continuity correction

Division	Pair composition		Wilcoxon rank sum test		
	Set 1	Set 2	W	p-value	Results
Pair 1	Stacking Ensemble	Random Forest	4920409	0.2596	Non-difference
Pair 2	Stacking Ensemble	MLP	4687668	1.479e-05	difference
Pair 3	Stacking Ensemble	CNN	4753811	0.0006236	difference

이에 따라, 비모수 방법인 Wilcoxon rank sum test를 통해 Pair를 구성하는 두 모델 예측값들이 통계적으로 유의미한 차이를 보이는지 확인했다. Table 7은 각 Pair의 Wilcoxon rank sum test 결과이다. 분석 결과, 스태킹 앙상블 모델 예측값과 랜덤 포레스트 모델 예측값(Pair 1)은 통계적으로 유의미한 차이를 보이지 않는 것으로 나타난 반면, 스태킹 앙상블 모델 예측값과 MLP 모델 예측값(Pair 2)는 통계적으로 유의미한 차이를 보인 것을 확인할 수 있다. 스태킹 앙상블 모델 예측값과 CNN 모델 예측값(Pair 3) 역시 통계적으로 유의미한 차이를 보였다.

5. 결론

본 연구에서는 부도위험을 머튼 모형을 기반으로 산출하고, 스태킹 앙상블 모형을 통해 부도위험을 예측했다. 이를 통해 기존 대부분의 연구들이 가지고 있던 희소한 부도사건에 따른 데이터 비대칭 문제와 이를 해소하기 위한 오버샘플링 및 언더샘플링이 데이터의 왜곡을 유발한다는 한계를 극복할 수 있었다. 본 연구에서는 부도여부가 아닌 각 기업의 시가총액과 주가변동성을 기반으로 시장에서 평가받고 있는 부도위험을 역으로 도출했다는 점에서 외적 위험이 기업의 생존에 직접적인 영향을 미칠 수 있는 포스트 팬데믹 시대에서 부도위험 산출 값에 대한 정합성과 활용도를 높였다. 또한 본 연구는 스태킹 앙상블 기법을 통해 각 서브 모델의 편향을 제어하고 안정적인 부도위험을 산출하였으며, 대조군과의 비교결과 스태킹 앙상블 모델은 단일 모델 상에서 가장 우수한 성능을 보였던 랜덤 포레스트 모델의 예측력을 상회했다. 스태킹 앙상블 모델의 예측값과 대조군인 개별 모델 예측값이 통계적으로 유의미한 차이를 보이는지

확인하기 위해 정규성 검정과 윌콕슨 순위합 검정을 수행했으며, 분석결과 스태킹 앙상블 모델의 예측값이 MLP 모델과 CNN 모델의 예측값과 통계적으로 유의미한 차이를 보인 것을 확인할 수 있었다.

본 연구의 학문적 기여는 메타-서브 구조 도입을 통한 과적합 문제 경감과 머튼 모형과의 연계를 통한 부도위험 세분화 및 머튼모형 자체의 한계점 개선으로 나타낼 수 있다. 먼저, 본 연구는 복수의 부도위험 예측 모델들을 서브모델로 도입하고 각 서브모델들이 산출하는 예측치를 메타러너가 학습해서 최종 부도위험을 산출하는 구조적 접근을 취했다는 점에서 단일 부도위험 예측 모델을 사용하고 있는 기존의 선행연구와 방법론적 차별성을 갖는다. 이는 개별 모델이 과거 기업 데이터에 과도하게 적합되어 일반화 능력이 떨어지는 과적합 문제를 해소하는데 기여될 수 있다. 또한, 본 연구는 예측을 위한 종속변수로 기존 선행연구에서 사용되어져 왔던 부도발생 여부 대신에 옵션가격결정모델에 기반한 머튼 모형과 Crosbie의 반복적 알고리즘을 통해 계산한 부도확률을 사용하였으며, 이를 통해 우량기업과 부실기업의 단순 구분에 초점을 맞춘 기존 선행연구들의 한계를 극복하는데 기여할 수 있다. 이와 더불어 주가를 토대로 부도확률을 계산하는 머튼 모형이 단독으로 사용될 경우 개별기업의 비정상적인 주가 변화에 따라 부도확률이 과대·과소 평가될 수 있는 반면, 본 연구는 대량의 회계 정보와 부도확률 간의 관계성을 만개 이상의 케이스를 통해 파악하고 이를 토대로 부도위험을 예측한다는 점에서, 머튼 모형의 한계를 극복하는데 기여할 수 있다.

본 연구의 실무적 기여는 주가 정보가 존재하지 않는 비상장 기업에게도 적합한 부도위험을 평가할 수 있도록 했다는 것과 전통적인 신용평가 모델을 예측모델에 반영할 수 있는 유연성을 제공했다는 것으로 나타낼 수 있다. 먼저, 본 연구는 시장의 평가를 반영하고 있는 부도확률 종속변수와 비상장 기업에서도 파악할 수 있는 회계적 변수들과의 관계를 파악하되, 예측 시에는 비상장 기업이 사용할 수 있는 회계적 정보를 이용하여 부도위험을 예측한다는 점에서, 주가 정보가 없는 비상장 기업에 대해서도 시장의 판단을 모사하여 적정한 부도위험을 평가할 수 있다. 이를 통해 중소기업이나 스타트업 등 전통적인 신용평가모델로는 적정한 부도위험 판단이 어려운 비상장 기업에 대해서도 부도위험평가 서비스를 제공할 수 있다. 또한, 본 연구는 머신러닝 기반 서브모델과 더불어 전통적인 신용평가모델도 서브모델로 반영하여 최종 부도확률을 산출할 수 있다는 점에서, 기존 신용평가업체의 전문성 및 노하우를 활용할 수 있도록 하였다.

오늘날 부도위험에 대해 AAA+부터 D까지 정형화된 구분이 존재하고 있고, 시장에서 해당 구분을 기준으로 의사결정이 이루어지고 있다는 것에서 볼 때, 본 연구에서 산출한 부도위험이 0~1을 범위로 갖는 확률값으로 산출된다는 점은 본 연구의 한계점으로 남는다. 이에 따라, 부도위험 예측치를 시장에서 사용되는 등급으로 Mapping하고 기존 신용등급 산출치와 비교하는 후속 연구가 필요할 것으로 판단된다. 또한, 본 연구에서는 분석의 용이성을 위해 서브모델을 머신러닝 분야에서 대표적인 세 가지 모델로 한정시켰으나, 스태킹 앙상블 기법은 다수의 서브모델이 존재할 때 더욱 안정적인 결과를 산출할 수 있다는 점에서 후속 연구에서는 더 많고 다양한 유형의 서브모델 활용이 필요할 것으로 판단된다.

기업 부도위험 평가는 해당 결과가 많은 산업에 영향을 미칠 수 있고 그 중요성이 높다는 점에서 산출값의 높은 신뢰성이 요구되며, 산출 방법에 대해서도 엄격한 기준을 요구하고 있다. 금융위원회가 금투업규정에서 정하고 있는 신용평가방법에서는 신용평가에 관한 과거의 통계자료 및 경험, 미래의 시장환경 변화 등을 고려하여 평가방법의 적정성 검증 등을 포함한 평가방법을 마련할 것을 요구하고 있다. 이러한 점에서, 머신러닝 기반 부도위험 예측모델이 도입되기 위해서는 높은 예측력을 갖출 뿐만 아니라, 금투업규정의 요구사항에 부합될 수 있어야만 하며, 본 연구에서 제안하는 스태킹 앙상블 기법은 서브모델을 금투업 규정에서 요구하는 적정성 기준에 부합되도록 세분화시킬 수 있다는 점에서 서비스 운영에 도움을 줄 수 있다. 본 연구가 머신러닝 기반 부도위험모델 도입에 제약요인으로 여겨졌던 문제들을 경감시켜 도입 활성화에 기여할 수 있기를 바란다.

참고문헌(References)

Altman, E., A. Resti, and A. Sironi, “Default Recovery Rates in Credit Risk Modelling: A Review of the Literature and Empirical Evidence,” Economic Notes, 33(2), (2004), 183-203.

Horrigan, J. O., “The Determination of Long-Term Credit Standing with Financial Ratios,” Journal of Accounting Research, 4, (1966), 44-62.

Ohlson, J. S., “Financial Ratios and Probabilistic Prediction of Bankruptcy,” Journal of Accounting Research, 18(1), (1980), 109-131.

Zmijewski, M. E., “Methodological Issues Related to the Estimation of Financial Distress Prediction Models,” Journal of Accounting Research, 22, (1984), 59-82.

Jeon, Yong., Ki., “Empirical study of the commercial credit rating and predictability of financial statements information”, Korea University PhD Thesis, 1986.

Jo, Ji., Ho., “A Study on Credit Rating System in Korea Bond Market”, Asian Review of Financial Research, V11, No.15, 1998, 149-190.

Jeong, Wan., Ho., “A Study on Forecasting Corporate Default: Based on Stock Price Information”, Asian Review of Financial Research, Vol 15, No.15(2002), 217-249.

Kang, Dae., IL., “An Analysis of Default Portfolios using the First Passage Time Stochastic Process”, Vol. 28, No.2,149~187

Kang, Dae., IL., “Stock Portfolio Composition and Trading Strategy Considering Bankruptcy Risks”, National Pension Service Research Report, 2014, 1~158

Yoon, J., M, “Effectiveness Analysis of Credit Card Default Risk with Deep Learning Neural Network”, Journal of Money & Finance, Vol. 33, No. 1(2019), 151~181.

Kwon, H., C, “A Study on the Work-time Estimation for Block Erections Using Stacking Ensemble Learning”, Journal of the Society of Naval Architects of Korea, Vol. 56, No. 6(2019), 488~496.

Choi, Y., S, “Expected Probability of Default Combining Financial Data and Stock Prices”, Korea Defivatives Association(2007)

Oh, S., K, “Structural Credit Risk Model

Incorporating the First Hitting Condition of Random Default Barrier", The Korean Journal of Financial Management, Vol 32, No. 4(2015), 23~51.

Cha, S., J, "Corporate Default Prediction Model Using Deep Learning Time Series Algorithm, RNN and LSTM", Journal of Intelligence and Information Systems, Vol 24, No. 4(2018), 1~32.

Byun, J., K, "The Effectiveness of Merton Model and Pricing of Credit Risk of Corporate Bonds.", journal of Korean Association of Applied Economics, Vol 6, No. 3(2004), 49~85

Lee, K. C., "Comparative Study on the Bankruptcy Prediction Power of Statistical Model and Al Models : MDA , Inductive Learning , Neural Network)", Journal of the Korean Operations Research and Management Science Society, Vol.18, No.2, (1993), 57~81.

Lee, J. S. and J. H. Han, "Test of Non-Financial Information in Bankruptcy Prediction using Artificial Neural Network - The Case of Small and Medium - Sized Firms -)", Journal of Intelligence and Information Systems, Vol.1, No.1, (1995), 123~134.

Kim, M. J., "Ensemble Learning for Solving Data Imbalance in Bankruptcy Prediction", Journal of Intelligence and Information Systems, Vol.15, No.3(2009), 1~15.

Kim, M. J., H. B. Kim and D. K. Kang, "Optimizing SVM Ensembles Using Genetic Algorithms in Bankruptcy Prediction", Journal of information and communication convergence engineering, Vol.8, No.4(2010), 370~376.

Kim, M. J., "Ensemble Learning with Support Vector Machines for Bond Rating", Journal of Intelligence and Information Systems, Vol.18, No.2(2012), 29~45.

Bae, J. K., "An Integrated Approach to Predict Corporate Bankruptcy with Voting Algorithms and Neural Networks", Korean Business Review, Vol.3, No.2(2010), 79~101.

Kim, S. J. and H. C. Ahn, "Corporate Bond Rating Model using Random Forest." , Journal of Intelligence and Information Systems, Spring Conference, (2014), 371~376.

Wang, H., Q. Xu and L. Zhou, "Large Unbalanced Credit Scoring Using Lasso-Logistic Regression Ensemble", PLoS One, San Francisco, Vol.10, No.2, (2015).

Yeh, S.-H., Wang, C.-J. and Tsai, M.-F., "Corporate Default Prediction via Deep Learning," in 24th Wireless and Optical Communication Conference (WOCC, 2014)

Min, S. H., "Bankruptcy prediction using an improved bagging ensemble", Journal of Intelligence and Information Systems, Vol.20, No.4, (2014), 121~139.

Min, S. H., "Simultaneous optimization of KNN ensemble model for bankruptcy prediction", Journal of Intelligence and Information Systems, Vol.22, No.1, (2016), 139~157.

Jo, N. O., H. J. Kim and K. S. Shin. "Bankrupcy Type Prediction Using A Hybrid Artificial Neural Networks Model." Journal of Intelligence and Information Systems, Vol.21, No.3, (2015), 79~99.

Jo, N. O. and K. S. Shin. "Bankrupcy Prediction Modeling Using Qualitative Information Based on Big Data Analytics", Journal of Intelligence and Information Systems, Vol.22, No.2, (2016), 33~56.

Addal, S., “Financial forecasting using machine learning”, African Institute for Mathematical Science, (2016), 1~32.

Yali Amit; Donald Geman (1997). “Shape quantization and recognition with randomized trees”. 《Neural Computation》 9 (7): 1545–1588. doi:10.1162/neco.1997.9.7.1545.

Abstract

Machine learning-based corporate default risk prediction model verification and policy recommendation: Focusing on improvement through stacking ensemble model

Eom, Haneul* · Kim, Jaeseong** · Choi, Sangok***

This study uses corporate data from 2012 to 2018 when K-IFRS was applied in earnest to predict default risks. The data used in the analysis totaled 10,545 rows, consisting of 160 columns including 38 in the statement of financial position, 26 in the statement of comprehensive income, 11 in the statement of cash flows, and 76 in the index of financial ratios. Unlike most previous prior studies used the default event as the basis for learning about default risk, this study calculated default risk using the market capitalization and stock price volatility of each company based on the Merton model. Through this, it was able to solve the problem of data imbalance due to the scarcity of default events, which had been pointed out as the limitation of the existing methodology, and the problem of reflecting the difference in default risk that exists within ordinary companies. Because learning was conducted only by using corporate information available to unlisted companies, default risks of unlisted companies without stock price information can be appropriately derived. Through this, it can provide stable default risk assessment services to unlisted companies that are difficult to determine proper default risk with traditional credit rating models such as small and medium-sized companies and startups. Although there has been an active study of predicting corporate default risks using machine learning recently, model bias issues exist because most studies are making predictions based on a single model. Stable and reliable valuation methodology is required for the calculation of default risk, given that the entity's default risk information is very widely utilized in the market and the sensitivity to the difference in default risk is high. Also, Strict standards are also required for methods of calculation. The credit rating method stipulated by the Financial Services

* Department of Science and Technology Studies, Korea University,
** Department of Science and Technology Studies, Korea University,
*** Corresponding author: Choi, Sangok
Department of public administration, Korea University
145, Anam-ro, Seongbuk-gu, Seoul, Republic of Korea
Tel: +82-2-3290-2284, E-mail: sangchoi@korea.ac.kr

Commission in the Financial Investment Regulations calls for the preparation of evaluation methods, including verification of the adequacy of evaluation methods, in consideration of past statistical data and experiences on credit ratings and changes in future market conditions. This study allowed the reduction of individual models' bias by utilizing stacking ensemble techniques that synthesize various machine learning models. This allows us to capture complex nonlinear relationships between default risk and various corporate information and maximize the advantages of machine learning-based default risk prediction models that take less time to calculate. To calculate forecasts by sub model to be used as input data for the Stacking Ensemble model, training data were divided into seven pieces, and sub-models were trained in a divided set to produce forecasts. To compare the predictive power of the Stacking Ensemble model, Random Forest, MLP, and CNN models were trained with full training data, then the predictive power of each model was verified on the test set. The analysis showed that the Stacking Ensemble model exceeded the predictive power of the Random Forest model, which had the best performance on a single model. Next, to check for statistically significant differences between the Stacking Ensemble model and the forecasts for each individual model, the Pair between the Stacking Ensemble model and each individual model was constructed. Because the results of the Shapiro-wilk normality test also showed that all Pair did not follow normality, Using the nonparametric method wilcoxon rank sum test, we checked whether the two model forecasts that make up the Pair showed statistically significant differences. The analysis showed that the forecasts of the Staging Ensemble model showed statistically significant differences from those of the MLP model and CNN model. In addition, this study can provide a methodology that allows existing credit rating agencies to apply machine learning-based bankruptcy risk prediction methodologies, given that traditional credit rating models can also be reflected as sub-models to calculate the final default probability. Also, the Stacking Ensemble techniques proposed in this study can help design to meet the requirements of the Financial Investment Business Regulations through the combination of various sub-models. We hope that this research will be used as a resource to increase practical use by overcoming and improving the limitations of existing machine learning-based models.

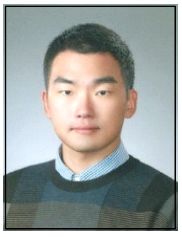
Key Words : Corporate default risk prediction, Merton model, Random forest, CNN

Received : April 30, 2020 Revised : June 22, 2020 Accepted : June 26, 2020
Publication Type : Regular Paper Corresponding Author : Choi, Sangok

저 자 소 개

엄하늘

숭실대학교에서 산업·정보시스템공학 학사, 성균관대학교에서 기술경영학 석사 학위를 취득하였으며, 고려대학교 과학기술학 협동과정 박사과정에 재학 중이다. 현재 에프앤자산평가에서 주식연계채권 평가 업무를 수행 중이다. 주요 관심분야는 기계학습, 알고리즘 트레이딩, 금융공학이다.



김재성

서강대학교 경영학 학사, 기술경영학 석사 학위를 취득하였고, 고려대학교 과학기술학 협동과정 박사과정에 재학 중이다. 현재 한국생산성본부에 R&D분야 직무교육 및 컨설팅 업무를 수행 중이다. 관심분야는 과학기술인력, 조직 관리, R&D 역량이다.

최상옥

Florida State University에서 행정학 박사학위를 취득하였으며, 현재 고려대학교 행정학과 교수로 재직 중이다. 연구분야는 인사, 조직/제도, 과학기술정책, 재난 및 위기관리, 네트워크이론이다.