저축은행 부실예측 모형의 머신러닝 기법으로의 전환과 예측력 개선의 주요 요인

이상헌(KIS채권평가 신사업개발실 실장)

https://doi.org/10.26588/kdic.2021.22.2.003

저축은행 부실예측 모형의 머신러닝 기법으로의 전환과 예측력 개선의 주요 요인*

이상헌**

〈요 약〉

본 연구는 기존의 금융기관 부실예측 모형을 머신러닝 기법으로 전환할 때 필요한 적용 방법론을 체계적으로 정리하고 저축은행을 대상으로 한 실증분석을 통해 대표적인 머신러닝 기법의 예측력 개선 효과를 분석한다. 이 과정에서 과대적합 및 편중된 분류 문제를 완화하기 위해 전진 교차검증과 SMOTE 기법을 적용하였다. 또한 부실예측 변수의 경제학적 해석이 가능하도록 부호 제약 Lasso 기법을 이용하여 변수를 축약하였다. 2008년부터 2019년까지의 저축은행 자료를 이용하여 실증 분석한 결과 머신러닝 모형은 기존의 로지스틱 회귀분석 모형에 비해 부실예측력을 높일 수 있음을 확인하였다. 주요 조건의 변화에 따른 민감도 분석 결과 머신러닝 기법에 따른 부실예측력 개선의 주요 요인은 이항 자료 불균형의 조정으로 나타났다. 따라서 기존의 통계적인 부실예측 모형을 머신러닝 기법으로 전환할 때 부실개체 샘플링 기법의 고도화가 중요함을 시사한다.

핵심 단어: 머신러닝, 저축은행, 부실예측, 이항 불균형 자료, SMOTE
JEL 분류기호: G12, G13, G24

접수일(2021년 2월 15일), 수정일(2021년 11월 19일), 게재확정일(2021년 11월 29일)

* 본 논문은 2020년도 예금보험공사 외부연구지원사업의 지원을 받아 작성되었습니다. 본 연구의 결론은 예금보험공사의 공식견해와는 관련이 없는 저자의 개인적인 결론임을 밝혀둡니다.
** KIS채권평가 신사업개발실 주소: 07328, 서울시 영등포구 국제금융로6길 38;
E-mail: shlee725@gmail.com; 전화: 02-3215-2924; Fax: 02-3215-1445.

Ⅰ. 서론

머신러닝 기법(머신러닝 모형과 자료 조정 방법)은 다양한 분야의 분류 및 예측 문제에서 우수한 성능을 나타내는 것으로 알려져 있다. 금융 분야에서도 이미 고객행동모형, 신용평가, 부도예측 등의 분류 또는 예측 문제에 머신러닝 기법이 활발히 적용되고 있다. 그런데 머신러닝 기법을 적용하는 과정에서 두 가지 사항을 주의 깊게 고려할 필요가 있다. 첫 번째는 과대적합(over-fitting)의 방지이다. 일반적으로 머신러닝 모형은 비선형 최적화 모형으로 볼 수 있으므로 내표본 적합도는 매우 높지만 그에 비해 외표본 예측성과가 낮게 나타나는 과대적합(또는 오차 최적화)의 문제에서 자유롭지 않으므로 머신러닝 기법의 적용 방법과 순서를 정교하게 구성해야 한다(Raschka; 2018). 두 번째는 개체 불균형(class imbalance) 문제의 조정이다. 일반적으로 부도 또는 부실 개체의 개수는 정상 개체보다 매우 작은 것이 일반적이다. 이 불균형을 조정하지 않으면 기존의 통계 모형과 마찬가지로 머신러닝 모형도 정상 개체의 예측에 더 높은 가중치를 두게 되어 보다 의미 있는 부도 또는 부실 개체의 예측력은 저하된다.

물론 과대적합 방지와 개체 불균형의 조정은 머신러닝 모형의 구축 과정에서 독립적으로 적용되는 것이 아니라 전체 프로세스에서 유기적으로 연결되어야 한다. 즉 머신러닝 기법의 적용 프로세스는 이 두 가지 블록을 포함하면서 변수 전처리, 변수 선택, 교차 검증 등이 하나의 프로세스로 정의되어야 한다. 이와 같이 완결된 하나의 프로세스를 통해 과대적합이 방지되고 특정 분류에 편중되지 않으면서 예측력이 제고된 소위 일반화(generalization) 성능을 가진 금융기관의 부실예측에 관한 머신러닝 모형을 구축하는 것이 중요하다. 이와 같은 접근법은 기존의 통계적인 부실예측 모형을 머신러닝 기법으로 전환할 경우에도 동일하게 적용된다.

금융안정이라는 목적 하에서 부실예측 모형의 고도화는 기존 통계모형의 개선뿐만 아니라 최신의 머신러닝 기법으로의 전환 및 확장까지 포함한다. 이와 같은 전환 과정에서 본 연구는 금융기관 부실예측에 관한 머신러닝 기법의 체계적인 적용 순서를 제시하고 이를 저축은행 부실예측에 적용하였다. 특히 본 연구는 방법론의 적용 순서에 초점을 두었는데 그 이유는 머신러닝 기법의 적용 순서가 기존 분석

절차에 비해 다소 복잡하므로 이 과정을 명확히 정리할 필요가 있기 때문이다. 체계적인 적용 과정의 구축을 통해 과대적합 및 정보누수(information leak)를 방지하는 것은 머신러닝 기법의 일반화 예측성능을 극대화하기 위한 전제조건이다. 실제로 변수 선택, 변수 스케일 조정, 편중된 분류 조정 등의 절차는 개별적으로는 명확하지만 머신러닝 모형의 학습과 결합될 때 그 적용 순서에 다소 중첩적이면서 모호한 부분이 존재한다. 이와 같은 머신러닝 기법의 적용 순서가 명확히 설정되면 그 이후에는 개별적인 모형이나 기법을 변경해가며 일반화 예측능력을 높이는 결과를 탐색할 수 있다.

머신러닝을 이용한 금융기관의 부실예측과 관련된 연구는 기본적으로 금융기관의 부실예측 확률을 높이기 위한 다양한 모형과 기법의 적용으로 귀결된다. 이때 설명변수는 부실변수와의 연관성이 높고 경제학적 해석이 가능한 변수들이며 SCOR (Statistical CAMELS Off-site Rating system) 모형 등에서 사용되는 자본적정성, 자산건전성, 수익성, 유동성 부문의 재무비율이 주로 사용된다. 또한 거시금융환경의 변화가 금융기관의 부실에 미치는 영향을 고려하기 위하여 거시경제 및 금융시장 변수가 포함되기도 한다(Suss and Treitel; 2019).

머신러닝을 이용한 금융기관 부실예측에 관한 연구는 매우 다양하지만 최근의 대표적인 연구로 Petropoulos(2017)은 은행의 부실예측을 위해 다양한 머신러닝 모형을 적용한 결과 Random Forest 모형이 Support Vector Machine, Artificial Neural Network 등에 비해 예측력이 우수함을 보여주었다. Carmona *et al.*(2018)은 Extreme Gradient Boosting 기법을 이용하여 이익잉여금, ROA, 위험가중 자본비율 등을 은행의 부실예측에 중요한 변수로 식별하였고 자산수익률이 지나치게 높을 경우 은행의 부실 위험을 높일 수 있다고 하였다. Suss and Treitel(2019)는 머신러닝 기법을 이용하여 은행 부실에 관한 조기경보 모형을 구축하였고 Random Forest 모형이 다른 머신러닝보다 예측력이 높다는 실증분석 결과를 제시하였다.

저축은행 부실예측에 관한 국내 주요 연구는 초기에는 로지스틱 회귀모형 등을 이용한 모형 구축 및 주요 변수 선정에 관한 내용이 주를 이루었고 최근에는 기존 지표의 설명력을 다시 검증하고 예측력이 높은 새로운 지표를 제시하는 과정을

통해 부실예측의 정밀도를 높이는 방향으로 진행되고 있다. 대표적인 국내 연구로써 남주하·진태홍(1998)은 로짓 부실예측 모형을 추정하여 예측성과가 93%~99%임을 보여주었다 또한 고정자산비율, 총자산영업이익률, 그리고 고정이하분류 여신비율 등이 상호신용금고의 주요 부실 변수라고 하였다. 장영광·김영기(2004)는 경영실태평가 지표를 이용한 도산예측모형을 구축 및 추정하고 그 타당성을 검증하였다. 김영기·정신동(2005)은 부실징후 상호저축은행을 조기에 판별하기 위한 조기경보모형을 개발하여 BIS기본자본비율 및 연체대출비율 등 4개 지표를 식별하고 자본적정성 및 자산건전성의 악화가 저축은행 부실화의 주요 원인이라고 해석하였다. 특히 이 연구는 이후의 저축은행 부실예측 연구의 벤치마크 역할을 하는 것으로 평가된다. 또한 강선민·황인태(2013)는 부채비율이 BIS비율보다 저축은행의 부실을 예측하는 데 있어 유용한 변수임을 보여주었다. 김남현·김민혁(2020)은 저축은행의 대안적인 부실예측지표로써 레버리지 지표인 예수금부채비율을 제안하였고 그 예측력이 높음을 보여주었다. 국내의 머신러닝을 이용한 금융기관의 부실예측 연구로써 김형준·류두진·조훈 (2019)은 기존의 기업부도예측에 관한 연구를 살펴보고, 통계적 모형과 기계학습 알고리즘의 대표적 방법론을 소개함으로써 관련 분야에 대한 이해를 돕고 있다. 이와 같은 국내외의 머신러닝을 이용한 금융기관 부실예측에 관한 연구는 새로운 변수의 중요성, 부실예측력의 제고 등에 대한 새로운 시사점을 제공해주고 있다.

이에 반해 본 연구는 부실예측력을 제고를 위한 모형 개발이라는 기존 연구와 목적은 같지만 보다 구체적으로 기존의 금융기관 부실예측 통계모형이 머신러닝 기반의 모형으로 전환되는 과정에서 어떤 단계가 중요한가를 실증분석을 통해 제시한다는 점에서 기존 연구와 차별성을 가진다. 이를 위해 본 연구는 다음과 같이 세 가지 측면에서 연구를 진행한다. 첫째, 대표적인 머신러닝 기법을 사용하되 일관된 프로세스의 적용에 초점을 두었다. 둘째, 변수 선택 기법으로써 부호제약 Lasso (least absolute shrinkage and selection operator) 모형을 이용하여 변수 축약과 경제학적 해석이 모두 가능하도록 하였다. 이는 금융감독원이나 예금보험공사 등의 기관에서 변수 선정 시 경제학적 해석을 매우 중요시하는 실무적인 요구를 충족시키기 위한 부분이다. 셋째, 기존 로지스틱 회귀모형을

벤치마크 모형으로 설정한 후 머신러닝 모형 및 조정 기법의 유무에 따른 결과를 다양한 각도에서 비교하였다. 이를 통해 부실예측력 개선의 요인이 머신러닝의 어떤 측면에서 도출된 것인지 파악할 수 있도록 하였다.

저축은행 자료를 이용한 실증분석 결과 머신러닝 기법은 기존의 로지스틱 회귀분석에 비해 부실예측력을 개선함을 확인하였다. 그러나 부실예측력 향상에 따른 오분류율(제1종 오류)도 증가할 가능성이 존재하였다. 이 결과는 부실예측력 향상과 제 1종 오류 간 상충관계가 존재한다는 것이므로 이 두 가지 사이의 최적화 즉 모형 튜닝이 중요함을 시사한다. 특히 다양한 조건을 변화시켜가며 검토한 민감도 분석 결과 머신러닝 기법의 부실 개체 예측력 개선에서 이항 불균형 조정이 매우 중요하였다.

이후 본 연구의 구성은 다음과 같다. II장은 본 연구에서 사용할 머신러닝 모형, 자료 처리 방법, 불균형 이항 자료 조정 기법 등을 다루고 머신러닝 방법론의 적용 순서 및 단계를 다룬다. III장은 실증분석이며 IV장은 실증분석으로부터 도출한 시사점을 다룬다. V장은 결론이다.

II. 방법론

1. 주요 머신러닝 모형

본 연구에서 다루는 머신러닝 모형에 대한 구체적인 내용은 이미 많은 연구에서 다루어졌으므로 간략한 특징과 내용을 다루되 각각의 머신러닝 모형의 학습을 위한 R의 대표적 패키지 라이브러리와 함수 그리고 초모수(hyper parameter)를 설명한다.

(1) Logistic Regression (LR)

Logistic 회귀모형(이하 LR 모형)은 X_{ij}와 p_i가 각각 설명변수(covariates)와

i번째 관측치의 분류확률일 때 0 또는 1을 가지는 종속변수의 분류 확률을 다음 식과 같이 추정하는 회귀 모형이다.

$$p_i = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots)}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots)} \quad (1)$$

이때 종속 변수가 이항 변수이므로 다음과 같은 우도 함수가 설정되며 이를 극대화하는 파라미터를 추정한다.

$$l(B; y, x) = \sum_{i=1}^{N} y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \quad (2)$$

LR 모형의 학습은 R의 glmnet 패키지의 glmnet() 함수가 이용되며 별도의 초모수는 없다.

(2) Decision Tree (DT)

Breiman *et al.*(1984), Quinlan(1986), Quinlan(1993) 등에서 제시된 DT 모형은 여러 나뭇가지가 하나의 뿌리에서 계층적으로 뻗어가는 것과 유사하게 중요도가 높은 설명변수가 가지는 범위를 분할해가며 순차적으로 분류하는 모형이다. DT 모형의 분류 기준으로 다음 식과 같은 Gini index와 Information gain (entropy)이 주로 사용된다.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2, \quad Entropy = \sum_{i=1}^{C} - p_i \log_2 (p_i) \quad (3)$$

여기에서 C(=2)는 부실/정상과 같은 분류 집단(class)의 인덱스, p_i는 i 클래스로 분류된 비율을 의미한다.

DT 모형의 학습은 R의 rpart 패키지의 rpart() 함수가 이용되며 초모수는 트리의 분할(splits)이 지나치게 많은 경우에 부여되는 벌칙을 의미하는 cp(complexity parameter), 트리의 높이(depth)를 의미하는 maxdepth, 부모 노드가 포함할 최소 관측치의 개수이며 분할의 종료 기준인 minsplit 등이 있다.

(3) Support Vector Machine (SVM)

Cortes and Vapnick (1995)의 SVM 모형은 마진 극대화 기법을 이용하여 분류 오차를 최소화하는 분리 초평면(hyper plane)을 찾는 머신러닝 기법이다. SVM 모형은 커널 함수(kernel function)를 이용하여 입력 벡터를 고차원의 특성 공간(feature space)로 변환함으로써 비선형 분류 경계를 생성할 수 있다(Kumar and Ravi; 2007). 또한 다양한 커널 함수를 이용하여 확장되며 보통 커널 함수의 종류도 초모수의 하나로 간주된다. 그러나 커널 함수에 따라 초모수에 차이가 있으므로 본 연구는 다음 식과 같은 선형 커널을 이용한 SVM-Linear Kernel (SVML)모형과 RBF 커널을 이용한 SVM-Radial Basis Kernel (SVMR)모형을 구분한다.

$$\begin{aligned} Linear\ \ kernel &= K(x_i, x_j) = x_i' x_j \\ Radial\ \ Basis\ \ kernel &= K(x_i, x_j) = \exp(\gamma x_i' x_j) \end{aligned} \tag{4}$$

SVML과 SVMR 모형의 학습은 모두 R의 e1071 패키지의 svm() 함수가 이용되지만 초모수에 차이가 있다. SVML 모형의 초모수는 마진 제약에 위배될 경우 적용되는 벌칙을 의미하는 cost이다. SVMR 모형의 초모수는 cost외에도 radial basis 커널함수의 모습을 결정하는 gamma가 있다.

(4) Random Forest (RF)

Breiman (2001)의 RF 모형은 앙상블 머신러닝 기법으로써 bagging (bootstrap aggregation)과 무작위 변수 선택을 결합하여 트리 모형의 집단 즉 숲을 구축함으로써 일반화 예측력을 높인 모형이다. 여기에서 bagging은 주어진 자료로부터 생성한 여러 개의 붓스트랩 샘플(random sampling)을 각각 모델링한 후 결합하여 최종 예측 모형을 산출하는 기법이다. RF 모형은 무작위로 특성변수를 선택함으로써 다수의 트리를 생성하기 때문에 과대적합을 피할 수 있다는 장점이 있으며 최종 분류는 다수의 트리가 각각 반환하는 분류 값의 다수결로 결정된다. RF 모형의 학습은 R의 randomForest 패키지의 randomForest() 함수가 이용되며 초모수는 샘플링할 특성변수의 개수를 의미하는 mtry와 트리의 개수를 의미하는 ntree 등이다.

(5) Artificial Neural Network (ANN)

ANN 모형은 인간의 뇌와 흡사하게 입력과 출력 사이의 구조를 여러 층의 인공 뉴런(artificial neurons) 또는 노드(nodes)로 연결하여 구성하여 분류 학습 문제를 해결하는 기법이다. 이와 같은 다층 구조는 은닉 층과 은닉 노드를 통해 이루어지므로 비선형성과 상호 작용을 효과적으로 다룰 수 있다는 장점이 있다. 또한 다른 모형과 달리 ANN 모형의 입력 변수는 0과 1사이의 값으로 조정되는 것이 학습 속도의 개선과 국지적인 해(local minima)의 문제를 완화하는데 도움이 되는 것으로 알려져 있다. 그러나 블랙박스 모형의 특성상 결과 해석이 쉽지 않으며 은닉 층과 은닉노드 수를 결정하는 것이 어렵다는 단점이 있다. 무엇보다 초모수 값에 따라 국지적인 해로 수렴될 수 있으며, 학습 과정에 많은 시간이 많이 소요되는 경향이 있다. ANN 모형의 학습은 R의 neuralnet패키지의 neuralnet() 함수가 이용되며 주요 초모수는 은닉 층의 구조를 나타내는 layer1, layer2와 학습 속도를 나타내는 learningrate 등이다.

(6) Gradient Boosting (GBoost)

Freund(1995), Freund and Schapire(1997), Friedman(2001) 등에 의해 제시된 GBoost 모형은 Gradient Tree Boosting 모형으로서 트리 모형과 경사하강(gradient descent)법에 따른 추가적인 학습을 결합한 것이다. 부스팅(Boosting)이란 약한 분류기를 결합하여 강한 분류기를 만든다는 의미이다. 구체적으로 GBoost 모형은 트리 모형을 순차적으로 추가하되 종속변수가 이전 트리 모형이 설명하지 못한 잔차항이라는 특징이 있다. 그런데 파라미터에 대한 목적함수의 최소화는 결국 잔차항의 최소화를 의미하므로 경사 벡터를 계산한 후 음(-)의 부호를 붙이면 그 방향이 결국 잔차항을 최소화하는 방향이 된다. 따라서 모수에 대한 학습은 음(-)의 경사 방향으로 어떤 크기(또는 비율)만큼 모수(학습률)를 변화시킴으로써 이루어진다. GBoost 모형의 학습은 R의 gbm 패키지의 gbm() 함수가 이용되며 주요 초모수는 부스팅 횟수를 나타내는 n.trees, 트리의 높이를 의미하는 나타내는 interaction.depth, 학습률을 의미하는 shrinkage 등이다.

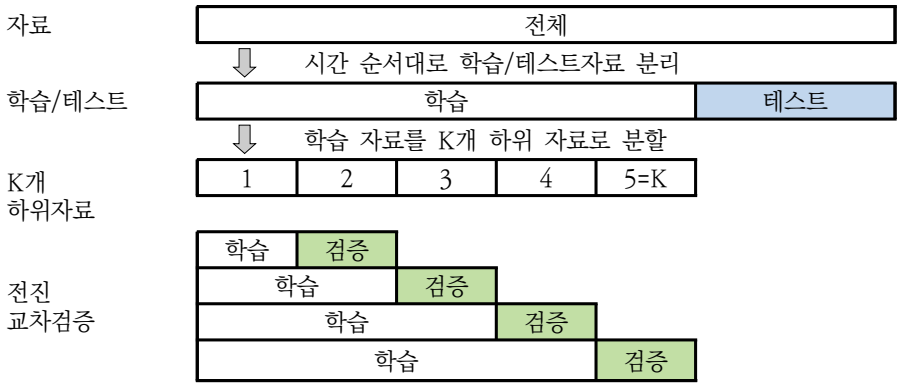
(7) Extreme Gradient Boosting (XGBoost)

Chen and Guestrin(2016)의 XGBoost 모형은 GBoost 모형을 모형의 간소화와 실행 속도 면에서 크게 개선한 모형이다. 모형의 간소화는 Lasso 모형을 이용한 모형 축소 및 변수 축약의 형태로 구현되었으며 과대적합을 방지하는데 효과적이다. 또한 트리 모형에서 부모 노드를 자식 노드로 분리하는 기준을 탐색하는 과정에 수치 근사법을 적용함으로써 실행 속도의 개선이 이루어졌고 기존 트리모형이 가지는 다차원 탐색 문제를 완화하였다. XGBoost 모형은 다양한 특성을 조합한 상당히 복잡한 머신러닝 모형으로서 초모수의 종류가 매우 많다. 본 연구에서는 세부적인 초모수는 기본 값을 사용하되 대표적인 초모수를 선택하여 최적화한다. XGBoost 모형의 학습은 R의 xbgoost 패키지의 xgboost() 함수가 이용되며 주요 초모수는 부스팅 횟수를 나타내는 nrounds, 트리의 높이를 의미하는 나타내는 eta, 학습률을 의미하는 max_depth 등이다.

2. 자료의 분할

머신러닝 학습과 예측을 위한 첫 번째 단계는 학습 자료(train sample)와 테스트자료(test sample)의 구분이다. 학습 자료는 모수 추정 등 학습을 위한 표본이고 테스트자료는 예측성과 비교를 위한 표본이다. 두 번째 단계는 학습 자료 내에서 학습 하위 자료(train fold)와 테스트 하위 자료(test fold)를 구분하는 것이다. 일반적으로 초모수를 결정하기 위해 학습 자료를 k개의 하위 자료로 구분한 후 각각의 하위 자료를 한 번씩 외표본으로 가정한 후(물론 이때 나머지 하위 자료들은 내표본으로 가정) 학습과 검증 과정을 k번 반복한다. 이를 k차-교차검증(k-fold cross validation)이라고 한다.

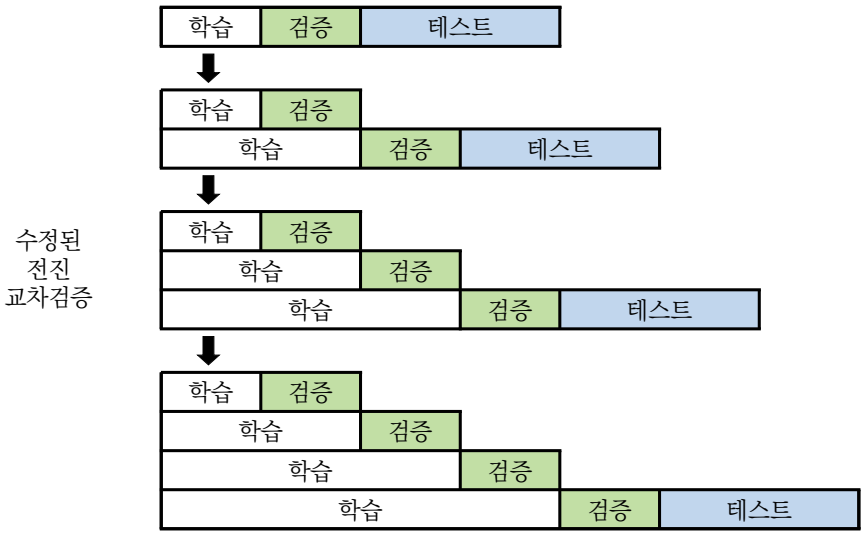
〈그림 1〉 자료의 분할과 명칭 및 전진 교차검증



그러나 재무/금융 분야에서 다루는 시계열 또는 패널 자료는 시간의 흐름에 따른 정보의 변화 또는 누적이 내재되어 있으므로 시간의 전후가 바뀌는 일반적인 교차검증은 적합하지 않다. 따라서 〈그림 1〉과 같은 전진 교차검증(forward chaining) 기법으로 시간의 순서를 고려하여 학습 자료를 순차적인 하위 자료로 구분한다[1].

1) 〈그림 1〉은 학습 하위 자료(train fold)를 확장구간(expanding window)으로 나타냈으나 이동구간(moving window)의 형태도 가능하다.

〈그림 2〉 수정된 전진 교차검증



본 연구는 〈그림 2〉와 같이 학습과 검증 단계 후 테스트 단계를 추가함으로써 더욱 일반화된 예측성과가 가능한 전진 교차검증 기법을 적용한 경우의 결과도 분석하였다.[2] 이 경우 테스트 기간은 고정되지 않고 각각의 전진 교차검증이 실시될 때 마다 다르다. 또한 최종 테스트 결과는 내부적으로 포함되는 형태의 교차검증(nested cross-validation) 후에 산출된 각 단계별 테스트 결과의 평균이다.

2) 전진 교차검증 방식은 테스트자료가 고정되어 있으므로 예측 모형의 일반화된 성능을 비교하기에는 한계가 있다. 특히, 본 연구와 같이 자료의 후반부에 부실 저축은행이 발생하지 않는 경우에 대해서는 예측 성능 평가가 어려울 수 있다. 이에 본 연구는 익명의 심사위원께서 제안하신 수정된 형태의 전진 교차검증(forward chaining nested cross-validation) 기법도 적용하였다. 이 방식은 시계열 데이터의 특성을 보존하면서 여러 번의 테스트를 수행할 수 있다는 장점으로 인해 더 일반화된 예측 성능 비교가 가능하다.

3. 불균형 이항 자료 샘플링 기법

일반적으로 부도 개체는 정상 개체에 비해 그 비율이 매우 낮기 때문에 불균형 이항 자료(class imbalance) 문제가 발생한다. 이 경우 모든 예측치를 정상으로 예측하더라도 정확도(accuracy)가 높아지는 경향이 있으므로 분석 목적인 부도 개체의 식별력(sensitivity 또는 recall)이 저하된다. 이 문제를 완화하기 위한 머신러닝 기법은 다양하지만 대표적으로 다음 표와 같은 샘플 조정 기법이 사용된다.

〈표 1〉 이항 분류 불균형 조정 기법

구분	내용
오버샘플링	소수 클래스(minority) 데이터를 복제함으로써 불균형 문제를 해결
언더샘플링	다수 클래스(majority) 데이터를 랜덤하게 제거함으로써 불균형 문제를 해결
SMOTE	오버샘플링과 언더샘플링을 합성한 방법으로 KNN을 이용한 랜덤 샘플링을 이용

언더샘플링은 자료의 개수를 현저히 줄임으로써 정보 손실의 문제가, 오버샘플링은 동일한 자료가 복제됨으로써 과대 적합의 문제가 있는 것으로 알려져 있다. 따라서 본 연구는 이와 같은 상충 관계를 고려한 Chawla *et al.*(2011)의 SMOTE(synthetic minority over-sampling technique) 기법을 이용한다.[3] SMOTE은 대표적인 불균형 이항 자료 조정 기법으로 소수 범주에 속하는 두 샘플 사이의 볼록 결합(convex combination)으로부터 가상의 샘플을 생성하는 방식이다.[4] 주의할 점은 Hulse *et al.*(2007), 김한용·이우주(2017)에서

3) 김한용·이우주(2017)에 따르면 불균형 이항 자료 조정 기법 중 SMOTE 기법이 항상 우위에 있는 것은 아니고 머신러닝 모형과의 조합에 따라 다르다는 것을 참고할 필요가 있다. 또한 불균형 이항자료 조정기법은 SMOTE 뿐만 아니라 매우 다양하므로 이를 고려하는 것이 바람직하지만 그 범위가 너무 넓어 본 연구의 범위를 벗어나므로 이 부분은 제외하였다.

4) 익명의 심사위원의 의견과 같이 금융데이터 관점에서 이러한 방식으로 두 부실 저축은행을 적절히 선형 결합한 새로운 부실 저축은행 데이터를 생성하는 것이 합리적인 가정인지 살펴볼 필요가 있다. SMOTE 기법은 기본적으로 KNN (K-Nearest Neighbors) 방법을 사용하므로 자료의 유사성이 높게 유지된다. 본 연구는 KNN의 최근접 이웃 자료의 개수를 3으로 지정하였으므로 시계열적으로 거리가 먼 자료의 경우 그 특성이 유사하지 않으면 선택될 확률이 낮기 때문에 큰 문제를 발생시키지 않을 것으로 판단된다.

강조되었듯이 편중된 개체 문제의 조정은 교차검증 과정에서 실행되어야 한다. 본 연구는 SMOTE 기법을 이용하여 1:1의 클래스 비율이 유지되도록 정상개체와 거의 같은 개수의 부실개체가 생성되도록 하였다[5].

4. 변수 스케일 조정

일부 머신러닝 모형의 성능은 설명변수의 스케일(값의 크기와 범위)에 영향을 받는 것으로 알려져 있다. 설명변수의 크기와 범위가 다를 경우 머신러닝 모형의 학습이 특정 변수의 영향을 과도하게 받는 경향이 있으므로 특정 변수의 영향이 지배적이지 않도록 조정할 필요가 있다. 모형에 따라 스케일 변환의 영향이 다르므로 선택적으로 적용할 수 있지만 모형 추정 전에 스케일 변환을 적용하는 것이 일반적이다.

스케일 변환 방법은 크게 두 가지로써 1) 변수의 범위를 0과 1 또는 -1과 1사이로 변환하는 정규화(normalization), 2) 변수의 평균과 표준편차를 각각 0과 1로 만드는 표준화(standardization)이다. 본 연구에서 변수 선택에 사용한 Lasso 모형은 계수의 크기에 벌칙을 주므로 변수의 스케일이 중요하며 보통 표준화 기법이 적용된다. ANN 모형의 경우 학습 속도 개선과 최적 결과 도출의 가능성을 높이기 위한 정규화가 필수적이다.

5. 특성 변수 축약(변수 선택)

변수 선택 즉 특성변수 축약은 머신러닝의 성능에 큰 영향을 주는 중요한 단계이다. 이 과정을 통해 1) 학습 속도 개선, 2) 모형의 복잡도 감소 및 이해도 제고, 3) 모형의 예측력 향상, 4) 과대적합 회피 등의 효과를 기대할 수 있기

5) 본 연구에서 사용한 R의 DMwR 패키지의 SMOTE()는 언더샘플링 비율(perc.under)과 오버샘플링 비율(perc.over)을 이용하여 두 클래스의 비율을 조정한다. 본 연구는 perc.under = 100, perc.over = (다수클래스 개수/소수클래스 개수)*100과 같이 지정하였다. 이때 기존 클래스 개수에 이 비율을 적용하면 최종 클래스 개수는 정수로 표현되므로 정확히 50:50이 아닌 예를 들어 49.9:50.1 정도로 나타난다. 이와 같은 이유로 클래스 비율이 1:1과 거의 같은 수준으로 유지된다고 표현한 것이다.

때문이다(Stańczyk; 2015). 그런데 부실 예측의 경우 결과의 경제학적 해석이 가능하기 위해서는 추정 계수의 부호가 이론 또는 기대 부호에 부합될 필요가 있다.[6] 변수 선택을 위해 일반적으로 많이 사용되는 단계적(stepwise) 회귀분석은 이론 부호 제약을 부여하기 어려운 한계를 가진다. 이에 본 연구는 변수 선택 기법으로써 Lasso 모형을 사용하되 식(5)와 같이 부호 제약 Lasso 모형을 적용하였다.[7] 따라서 본 연구에서 모형 추정 전 선택된 변수들은 모두 기대 부호에 부합된다.

$$\beta(Lasso)=\sum_{t=1}^{n}\left(y_t-\sum_{j=1}^{p}\beta_j X_{jt}\right)^2+\lambda\sum_{j=1}^{p}w_j|\beta_j|$$
$$w_j=\left[\tanh\left(-\beta_j s_j\right)\right]_+ +1=\begin{cases}\infty, & \beta_j s_j<0\\ 1, & \beta_j s_j\geq 0\end{cases} \quad (5)$$

여기에서 w_j는 계수별로 다른 추가적인 가중치이고 s_j는 사용자가 지정한 기대부호(-1, 0, 1 중 하나)이다.

6. 예측성과 비교 방법

일반적으로 이항분류 모형의 예측성과 지표는 두 가지로 구분된다. 우선 분류행렬(confusion matrix)로부터 계산한 다양한 성과 측정치들이 이용된다. 또한 분류의 기준인 임계값을 변화시켜가며 생성한 분류행렬들로부터 계산된 TPR (true positive rate)과 FPR (false positive rate)의 조합에 기반한 ROC (receiver operating characteristic) 곡선 또는 ROC곡선 아래의 면적인 AUROC (area under ROC)도 이용된다.

---

6) 머신러닝 모형은 일반적으로 블랙박스 모형이고 인공신경망 모형의 경우 비선형으로 인해 지역적으로 그 부호가 다를 수도 있으므로 분석 과정에서 부호 설정이 필수 과정은 아니다. 그럼에도 불구하고 본 연구에서 이론 부호 제약을 넣은 이유는 부실이나 위기 등의 예측과 관련된 연구에서 예측성과뿐만 아니라 그 해석 가능성도 매우 중요하기 때문이다. 그러나 머신러닝 모형의 학습 단계에서 이론 부호를 부여하는 것은 어렵기 때문에 변수 선택 단계에서만 이론 부호 제약을 부여하였다.

7) 일반적인 선형 제약을 도입한 Lasso 모형에 관해서는 James *et al.*(2013)을 참고하면 된다.

(1) 분류행렬

〈표 2〉는 실제값과 예측값의 비교 결과를 2×2 행렬로 나타낸 분류행렬이다. 이때 사용한 임계값은 0.5이다. 분류행렬의 행과 열은 각각 실제값과 예측값이고 두 값은 모두 0과 1 즉 음성과 양성으로 구분된다. 이때 실제값이 음성인 경우는 *Y*가 음성이라는 귀무가설(H0)이 참이라는 것을, 실제값이 양성인 경우는 귀무가설이 거짓이고 *Y*가 양성이라는 대립가설(H1)이 참이라는 것을 의미한다. 또한 예측값이 음성이면 귀무가설이 기각되지 못함을, 예측값이 양성이면 귀무가설이 기각됨을 의미한다.

〈표 2〉 분류행렬(confusion matrix)

		모형의 Y 예측값	
		음성 (H0 기각 못함)	양성 (H0 기각)
실제 Y값	음성 (H_0는 참)	TN (true negative)	FP (false positive) 제1종 오류
	양성 (H_0는 거짓)	FN (false negative) 제2종 오류	TP(true positive)

분류행렬로부터 다음 식과 같은 다양한 성과 측정치를 계산할 수 있다.

$$
\begin{aligned}
&TPR = TP/(TP+FN) = TPR = sensitivity = recall \\
&TNR = TN/(FP+TN) = TNR = specificity \\
&FPR = FP/(FP+TN) = FPR = 1 - specificity \\
&FNR = FN/(TP+FN) = FNR = 1 - recall \\
&accuracy = (TP+TN)/(TP+FN+FP+TN) \\
&precision = TP/(TP+FP) \\
&F1\,score = 2(precision \times recall)/(precision + recall) \\
&Cohen's\ kappa = \frac{p_0 - p_e}{1 - p_e}, \quad \begin{cases} p_0 = observed\ \ accuracy \\ p_e = expected\ \ accuracy \end{cases} \\
&balanced\ \ accuracy\ \ = (TPR+TNR)/2 \\
&G-mean = sensitivity \times specificity \\
&F2\ score = (1+\beta^2) \times \frac{precision \times recall}{(\beta^2 \times precision) + recall},\ \beta = 2 \\
&MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}
\end{aligned}
\tag{6}
$$

분류행렬의 기본 원소는 TP (true positive), TN (true negative), FP (false positive), FN(false negative)이며 순서대로 양성 예측이 맞은 경우, 음성 예측이 맞은 경우, 양성 예측이 틀린 경우, 음성 예측이 틀린 경우를 의미한다. 이 원소들의 비율인 주요 성과 측정 지표 중 정확도(accuracy)는 모형의 판정과 실제 양성/음성 여부가 동일한 경우의 비율을 나타낸다. 정밀도(precision)는 모형이 양성으로 판정했는데 실제 양성인 경우의 비율이다. 민감도(sensitivity)는 실제 양성일 때 양성으로 판정한 비율이고 특이도(specificity)는 실제 음성일 때 음성으로 판정한 비율이다.

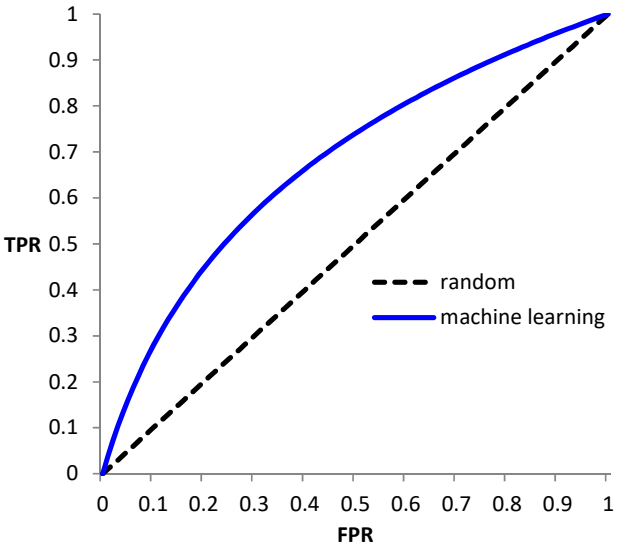
이진 불균형 자료의 특성을 고려한 대표적인 예측성과 지표에는 MCC, balanced accuracy, F2, G-mean, Cohen's kappa 등이 있다. Matthews correlation coefficient (MCC)는 이진 분류 결과의 균형척도(balanced measure)로써 -1(완전히 잘못된 모형)에서 1(완전히 예측된 모형) 사이의 값을 가지며 부실과

정상 개체를 동시에 고려하므로 이진 불균형 자료에 대한 성과지표로 유용한 것으로 알려져 있다 (Chicco; 2017). 나머지 측도들의 수식도 결국 희소한 부실개체의 비중이 강조되는 형식을 가진다.

(2) ROC와 AUROC

〈그림 3〉은 분류 모형의 임계값을 0에서 1까지 작은 간격으로 변화시켜가며 계산한 TruePositive와 FalsePositive를 동시에 나타낸 ROC곡선이다. ROC 곡선에서 45도 선은 분류의 성능이 랜덤하게 결정되는 모형을 의미한다. 만일 실제 양성인 개체에 대하여 양성으로 예측되는 확률이 높다면 ROC 곡선은 이 45선 위에 그려진다. 따라서 어떤 머신러닝 모형의 ROC 곡선이 45도 선에서 멀어질수록 모형의 양성 분류 성능이 높은 것으로 해석한다. ROC 곡선이 45도선에서 좌상향으로 멀어질수록 ROC곡선 아래의 면적(AUROC)은 확대된다. 따라서 AUROC는 머신러닝 모형의 분류 성능을 하나의 숫자로 나타낸 것이다.

〈그림 3〉 ROC 곡선의 예



7. 머신러닝 기법을 이용한 분석 과정

분류 또는 예측 문제에 머신러닝 기법을 적용할 경우 과대적합을 방지하고 일반화 예측(generalization prediction) 능력을 향상시키기 위하여 정보 누수가 발생하지 않도록 올바른 적용 프로세스가 구축되어야 한다. 〈표 3〉은 일반화 예측능력을 제고하기 위한 머신러닝 기법의 적용 순서를 나타내며 크게 세 단계로 구분된다.

〈표 3〉 머신러닝 기법의 적용 순서

1. 자료	
1) 원자료	분류대상(Y) 결정, 이론 및 경험적 근거로부터 설명변수 (X) 수집
2) 자료 탐색	① X의 이상치(outlier), 결측치(NA), 분산이 0이거나 거의 0인 자료(near zero variance) 확인 후 제거 또는 보정 ② 범주형 변수 변환(one hot encoding, entire label encoding)
3) 자료 분할	① data의 순서를 무작위로 변경 ② 70:30 정도로 학습 자료와 테스트자료로 구분 ③ 학습 자료를 K개의 fold로 구분, 시계열 또는 패널 자료의 경우 ① 과정 제외(시간 순서 유지)

⬇

2. 교차검증 : X변수, 모형, 초모수 선택	
1) train /test fold 구분	i번째 fold를 test fold (validation set)로, 이를 제외한 자료를 train fold 로 구분 (시계열 자료의 경우 forward chain 이용)
2) 이항 자료 불균형조정	① Y가 불균형 이항 자료인 경우 불균형을 조정하는 샘플링 기법들을 이용하여 train fold 만 조정
3) train fold의 특성변수 축약(feature selection)	① train fold 에 대해서만 특성변수 축약 방법 적용 - Filter: X와 Y, X그룹에 대한 상관관계 등 - Wrapper: 단계적 selection, RFE 등 - Shrinkage: lasso, ridge, elastic-net 등 ② 변수 축약 후 분석용 train fold의X 확정
4) train fold의 X 변환 (transformation)	① train fold 의 X만 이용하여 정규화 또는 표준화 등 변환 함수의 파라미터 추정 ② 추정된 파라미터로 설정된 변환 함수를 이용하여 train fold 의 X를 transformed train fold의 X로, test fold의 X를 transformed test fold 의 X로 변환 ③ Y는 transformed train/test fold의 Y로 이름만 변경
5) i번째 검증	① transformed train fold에 후보 모형들을 학습시킴

(validation)	② 학습된 모형들을 transformed test fold의 X에 적용하여 계산한 예측치를 transformed test fold의 Y와 비교 ③ 모형/샘플링 기법별 i번째 예측성과 측정치를 기록
교차검증 (cross validation)	① 1)~4) 단계를 i=1부터 i=K까지 test fold를 바꿔 가며 계산한 예측성과 측정치(Balanced Accuracy등)의 평균 등으로 모형 순위를 기록
모형 튜닝(tuning)	① 초모수(hyper parameter)를 변경하며 6) 단계 즉 교차검증을 반복 실행(grid search 등): 기록된 순위정보로부터 최적 모형과 초모수 선택

⬇

3. 최종 예측 : 변수 선택, 스케일 변환, 선택된 모형(초모수 고정) 추정 후 예측	
1) 특성 변수 축약	① train set을 대상으로 2.3)에서 선택한 변수 선택 및 축약 기법 적용하여 X 변수 확정
2) test set의 X 변환	① 학습 자료의 X로부터 2.4)에서 선택된 변환함수의 모수 추정 ② 추정된 변환 함수를 이용하여 test set의X를 transformed test set의 X로 변환 ③ Y는 변환 없이 transformed test set의 Y로 이름 변경
3) 예측	① 〈2. 교차검증〉에서 선택된 최적 모형(초모수 고정)을 transformed test set의 X에 적용하여 예측치 계산 ② 예측치 Y와 transformed test set의 Y를 비교하여 예측성과 측정

주) train fold : 하위 학습 자료, test fold : 하위 테스트자료, train set : 학습 자료, test set : 테스트자료

첫 번째 단계인 자료의 수집과 탐색 및 자료의 분할 단계는 부실개체에 대한 정의(부도 여부 또는 자기자본비율 등의 크기에 따른 분류), 기존 연구에서 중요하게 언급되거나 새롭게 도입할 설명변수(주요 재무비율) 수집, 내표본 적합도와 외표본 예측성과를 판단하기 위한 표본의 분리를 다룬다.

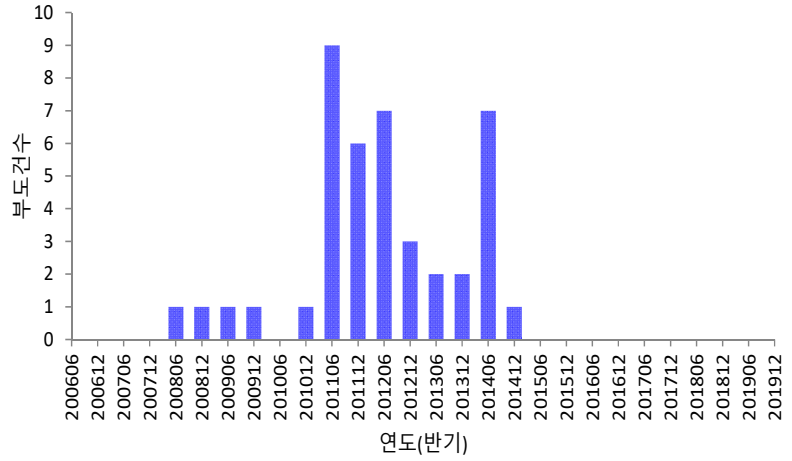
두 번째 단계인 교차검증 단계는 변수 변환, 변수 선택, 편중된 분류 문제 조정 등으로 구성된다. 이 과정은 초모수 선택과 모수 추정을 동시에 포함하며 가장 예측력이 좋은 모형의 구조를 선택한다.

세 번째 단계인 모수 추정과 예측 단계는 교차검증 단계에서 선택된 모형을 이용하여 내표본(학습 자료 전체)을 대상으로 변수 선택, 변수 변환의 과정을 다시 거쳐 모수를 추정하고 예측성과를 측정한다. 주의할 점은 이 단계에서는 편중된 분류 문제를 조정하면 안 된다는 것이다. 왜냐하면 이 단계는 외표본(테스트자료)에 최종 선택된 모형을 적용하여 예측력을 검증하는 것이기 때문이다.

Ⅲ. 실증 분석

1. 표본 구성

〈그림 4〉 저축은행 부도건수



주 : 부도건수는 반기 재무제표가 2019년 12월 현재 공시되지 않는 폐업 저축은행의 건수를 의미하며 부도시점은 마지막 재무제표 공시일을 의미한다.
출처 : 금융감독원 금융통계정보

본 연구의 실증분석은 머신러닝 기법을 이용하여 저축은행의 부실 여부를 예측하는 것이다. 분석 기간은 반기(6월, 12월) 기준으로 2008년 6월부터 2019년 12월까지이다. 부실여부는 재무제표 기준으로 판단한다. 즉 2019년 12월 이전에 재무제표가 마지막으로 보고된 시점이 있을 경우 해당 저축은행을 부실 개체로 간주하고 그 시점을 부실 시점으로 간주한다.[8] 이때 부실예측이 목적이므로 설명변수와 종속변수의 시차는 6개월이다.

〈그림 4〉는 반기별 저축은행 부도건수[9]를 나타낸다. 2014년 12월의 골든브릿지

8) 저축은행의 구조조정으로 인해 부실이 상당히 진행된 저축은행이 영업정지되거나 합병으로 인해 새로운 저축은행이 된 경우 이 개체들을 표본에서 제거할 수도 있으나 부실에 대한 정보를 제공한다는 점과 부실개체수의 부족을 보완한다는 점을 고려하여 본 연구에서는 별도로 제거하지 않았다.
9) 본 연구는 부도 이벤트의 시점을 마지막 재무제표가 공시된 반기의 다음 반기가 아니라 마지막 공시가 발생한 반기로 선택하였다. 그 이유는 마지막 공시 시점의 재무비율 변수(BIS자기자본비율

저축은행의 부도를 끝으로 더 이상 저축은행의 부도가 없는 실정이다. 저축은행의 부도 예측을 위한 내표본/외표본 분리는 생각보다 다소 어려운 부분이다. 왜냐하면 저축은행의 마지막 부도 관측 시점은 2014년 12월이므로 2015년부터 2019년까지 5년 동안 정상 개체만 관측되기 때문이다. 그래서 내표본을 2008년 6월부터 2013년 12월까지로 설정하고 분석을 위한 최소한의 부실 개체수가 포함되도록 외표본을 2014년 6월부터 2019년 12월까지로 설정하였다.

2. 자료

저축은행 부실예측을 위한 후보 설명변수는 국내외 문헌을 기초로 〈표 4〉와 같이 구성하였다. 특히 사전에 기대부호를 부여하여 변수 선택과정에서 제약조건으로 사용하였다. 후보 변수들의 기초통계량은 〈부록〉에 수록하였다.

〈표 4〉 후보 변수 목록

분류	변수명	변수 산식	기대부호
자본적정성	BIS기준 자기자본비율	BIS 기준 자기자본/위험가중자산	-
	BIS기준 기본자기자본비율	BIS 기준 기본자본/위험가중자산	-
	단순자기자본비율	단순자기자본(B/S)/총자산	-
	예수금부채비율	예수금/자기자본	+
	레버리지비율	총부채/자기자본	+
	예대율	대출금/예수금	0
자산건전성	순고정이하여신비율	(고정이하여신-대손충당금)/총여신	+
	연체율	연체금액/총여신	+
	고정이하여신비율	고정이하여신/총여신	+
	대손충당금적립비율(총여신대비)	대손충당금잔액/총여신	-

등)는 합리적인 범위를 벗어나는 경우가 대부분이므로 그 시점의 변수를 이용하여 다음 반기를 예측하는 것은 이미 알려진 정보를 이용하여 예측하는 것과 크게 다를 것이 없다고 판단했기 때문이다. 그래서 본 연구는 마지막 재무제표 공시가 발생한 반기 시점을 부도 이벤트로 설정하고 설명변수는 직전 반기의 재무비율 변수를 이용하였다. 이와 같은 설정은 예측을 더욱 어렵게 하는 요인이 되지만 예측의 목적에는 더욱 부합하는 설정으로 판단된다.

분류	변수명	변수 산식	기대부호
	대손충당금적립비율(요적립액대비)	대손충당금잔액/총요적립액	-
	대손충당금적립비율(고정이하여신대비)	대손충당금잔액/고정이하여신	-
	총 대출비중	대출금/총자산	0
	담보대출비중	담보대출/총대출	0
	보증대출비중	보증대출/총대출	0
	신용대출비중	신용대출/총대출	0
	비업무용자산비율	비업무용자산 / 총자산	-
수익성	총자산순이익률(ROA)	당기순이익/총자산	-
	자기자본순이익률(ROE)	당기순이익/자기자본	-
	수지비율	영업비용/영업수익	+
	총자산경비율	판매비와관리비/총자산	+
	영업이익경비율	판관비/(이자수익-이자비용+수수료수익-수수료비용-예금보험료)	+
	자기자본 대비 이익잉여금 증감	(금반기이익잉여금-전반기이익잉여금)/자기자본	-
	자기자본 대비 자본잉여금 증감	(금반기자본잉여금-전반기자본잉여금)/자기자본	+
유동성	유동성비율	유동성자산/유동성비율	-
	유형자산비율	유형자산/자기자본	-
	실가용자금비율	(현금 및 예치금+단기매매증권)/총예수금	-
	수신증가율	(금반기예수금-전반기예수금)/전분기예수금	0
	역고정 자산비율	자기자본 / (투자자산 + 유형자산 +무형자산 +비업무용자산)	-
거시경제	GDP 성장률		-
	주택가격상승률		-
	기간 스프레드		-
	신용 스프레드		+
	Log(총자산)	저축은행 규모 통제변수	0

주) 기대부호에서 +는 저축은행 부실과 양의 관계, -는 음의 관계를 나타내고 0은 부실과의 관계가 다소 불확실하여 모형이 결정하도록 부호제약을 유보한 것이다.

예수금부채비율은 김남현·김민혁(2020)에 기반하며 채무부담 한도를 대리하는 지표로서 자본 대비 예수부채로 계산된다. 저축은행의 고금리 수신에 대한 조치로써 2020년부터 시행 중인 저축은행의 예대율 변수를 추가하였다.[10] 그 외의 변수들은 대부분 연구에서 사용되는 변수들인데 그 중 대출 비중 관련 변수들의 이론 부호는 다소 불확실한 면이 있다. 그 이유는 대출 비중 관련 변수는 대출의 구성과 질에 따라 그 기대 부호가 다르게 나타나기 때문이다. 그래서 본 연구는 대출 관련 비중을 포함한 몇 개의 변수에 대한 기대 부호를 음(-) 또는 양(+) 대신 불확실한 것으로 가정하였다. 즉 이들 변수의 기대부호는 모형이 결정해주는 것으로 유보하였다. 이와 함께 거시경제의 변화는 실물, 금리, 부동산 측면을 나타내는 거시변수를 사용하였다. 또한 기존 문헌에서 기업 규모의 통제변수로 관행적으로 사용되는 log(총자산)을 추가하였다.

3. 전처리 과정의 적용 순서

SMOTE 같은 이항 불균형 조정 기법이 적용될 경우 자료의 전처리 과정은 이항불균형조정-변수선택-변수스케일변환 (sample-select-scale)의 순서를 따르는 것이 자료의 특성을 올바로 반영하는 방법이다. 〈표 5〉는 학습(train) 자료와 테스트(test) 자료에 적용한 변수 전처리 과정의 순서에 따른 X변수(BIS기준자기자본비율)의 표본 평균을 정리한 것이다. 이때 scale은 평균과 표준편차를 각각 0과 1로 표준화한 것이고, select는 부호제약 lasso 로지스틱 모형을 이용한 변수 선택, sample는 이항 불균형 조정기법인 SMOTE 기법의 적용을 의미한다.

10) 예대율의 정확한 계산을 위해 필요한 대출금리별 대출금액에 대한 정보가 없는 관계로 대출금/예수금으로 계산하였다.

〈표 5〉 변수 스케일 조정(scale), 변수 선택(select), 불균형 이항 자료 조정(sample) 적용순서에 따른 표본 평균 비교(BIS기준자기자본비율의 표본평균)

번호	적용 순서	x.train (all)	x.test (all)	x.train (y=0)	x.test (y=0)	x.train (y=1)	x.test (y=1)
1	scale-select-sample	-0.530	0.000	0.018	0.032	-1.063	-1.092
2	scale-sample-select	-0.535	0.000	0.018	0.032	-1.072	-1.092
3	select-scale-sample	-0.530	0.000	0.018	0.032	-1.063	-1.092
4	select-sample-scale	0.000	0.513	0.530	0.543	-0.515	-0.543
5	sample-scale-select	0.000	0.517	0.535	0.548	-0.519	-0.539
6	sample-select-scale	0.000	0.517	0.535	0.548	-0.519	-0.539

〈표 5〉를 보면 scale 적용 후 sample을 적용하는 1, 2, 3번의 경우 평균이 0이 아니므로 scale 결과가 변형되어 머신러닝 모형의 입력 변수로 스케일이 조정된 변수를 넣는 목적에 부합되지 않는다. 4, 5, 6번 중 4번의 경우 sample 적용 전에 select가 이루어짐으로써 부호제약 Lasso 로지스틱 모형은 학습 자료의 대부분을 차지하는 정상 개체를 잘 설명하는 변수를 선택할 것이므로 부실 개체에 대한 고려가 부족한 변수 선택이 이루어질 것으로 예상된다. 따라서 5와 6번이 변수 전처리 과정의 순서로 적절할 것이다. 본 연구는 변수 선택 후 스케일 변환이 이루어지는 일반적인 경우를 고려하여 sample-select-scale를 적용한다.

4. 머신러닝 모형의 초모수

〈표 6〉 머신러닝 모형의 초모수와 R 패키지 및 함수

모형	초모수	R패키지	R함수
LR	없음	glmnet	
DT	cp = (0.1,0.4,0.9), minsplit = (3,5,7) maxdepth = (3,5,7), loss = gini	rpart	
SVML	cost = (0.1,1,5)	e1071	svm
SVMR	cost = (0.1,1,5), gamma = (0.01 0.1,1),	e1071	svm
RF	mtry = (3,5,7), ntree = (100,250,500)	randomForest	
ANN	layer1 = (5,7), layer2 = (3,5) learningrate = (0.1,0.3,0.5)	neuralnet	

모형	초모수	R패키지	R함수
GBoost	n.trees = (100,250,500), interaction.depth = (3,5,7) shrinkage = (0.1,0.3,0.5)	gbm	
XGBoost	nrounds = (100,250,500), eta = (0.1,0.3,0.5) max_depth = (3,5,7)	xgboost	

초모수의 범위는 후보변수 선정과 마찬가지로 국내외 문헌을 참고하여 정하였다. 〈표 6〉은 본 연구에서 설정한 머신러닝 모형의 변수변환, 초모수, 학습에 사용한 R 패키지와 함수를 정리한 것이다.

5. 벤치마크 모형의 변수 선택과 추정 결과

머신러닝 모형과 비교할 벤치마크 모형(LR_base)은 부실예측에 주로 사용되는 로지스틱 회귀모형이며 이항조정 불균형 조정과 교차검증 과정이 제외된 모형이다. 따라서 벤치마크 모형의 변수 선택은 머신러닝 기법인 이항조정 불균형을 제외한 후 학습 자료 전체를 대상으로 부호제약 Lasso 기법을 적용한 결과이다.

〈표 7〉 저축은행 부실예측을 위한 기존 로지스틱 회귀모형 추정 결과

변수	Estimate	Std.error	z-stats	Prob
상수항	3.0394	2.8809	1.06	0.2914
BIS기준자기자본비율	-0.0405*	0.0171	-2.37	0.0180
연체율	0.0431**	0.0143	3.01	0.0026
대손충당금적립비율요적립액대비	-0.0493*	0.0216	-2.28	0.0227
총대출비중	-0.0361*	0.0143	-2.53	0.0114
보증대출비율	0.1529*	0.0711	2.15	0.0314
총자산이익률ROA	-0.1102**	0.0376	-2.93	0.0034
이익잉여금증감_자기자본대비	-0.0161	0.0130	-1.24	0.2164
실가용자금비율	-0.0544**	0.0182	-2.99	0.0028
수신증가율	-0.0625**	0.0228	-2.74	0.0061
신용스프레드	0.1670	0.2036	0.82	0.4121

〈표 7〉은 부호 제약 Lasso 모형으로 선택한 설명변수에 대한 로지스틱 회귀모형의 결과이다. 이익잉여금 증감과 신용스프레드를 제외하면 대부분의 설명변수가 통계적으로 유의한 것으로 나타났다. 예를 들어 연체율이 높으면 부실 확률이 높아지고 BIS기준 자기자본비율이나 총자산 이익률이 높아지면 부실 확률이 낮아지는 것으로 나타났다. 통계적으로 유의하지 않았지만 신용스프레드가 확대되면 부실 확률이 높아진다는 결과는 신용경색과 같은 거시경제 환경의 변화가 부실 확률에 영향을 줄 가능성이 있음을 시사한다. 머신러닝 모형의 변수 선택과 마찬가지로 학습 자료가 변하면 이 변수 선택 결과도 바뀌게 된다.

6. 머신러닝 기법의 변수 선택 결과

벤치마크 로지스틱 회귀 모형(LR_base)의 변수선택에 대응된 머신러닝 모형의 변수 선택 결과는 〈표 8〉과 같다. 이 변수들은 학습 자료 전체를 대상으로 SMOTE을 이용하여 이항조정 불균형을 조정한 후 부호제약 Lasso 모형으로부터 선택된 것이다. 이때 초모수인 벌칙 파라미터 λ는 자본적정성 등 각각의 분류 기준에서 최소 한 개의 변수가 선택되도록 제약을 주도록 설정되었다. 물론 학습 자료가 변하면 이 변수 선택도 바뀌게 된다. 또한 머신러닝 모형마다 변수 선택을 다르게 할 수 있으나 동일한 조건 하에서 예측성과를 측정한다는 가정 하에 동일한 설명변수가 사용되도록 변수 선택 또는 축약 과정을 모형 추정과 분리하였다.

〈표 8〉 부호 제약 Lasso 모형을 이용한 머신러닝 모형의 변수 선택 결과

분류	변수명
자본적정성	BIS기준 자기자본비율
자산건전성	연체율
	대손충당금적립비율(요적립액대비)
수익성	총자산순이익률(ROA)
유동성	수신증가율
거시경제	신용 스프레드

7. 예측성과 비교

예측성과의 측정은 내표본 전체를 이용하여 최종모형을 추정한 후 외표본의 실제 부실 여부와 비교하는 것이다. 이때 부실 판단을 위한 임계값은 50%로 설정하였고 예측 부실 확률이 50% 이상이면 부실(positive)로 예측된다. 예측성과를 판단하기 위한 외표본은 2014년 6월부터 2019년 12월까지이다.

〈표 9〉는 벤치마크 모형과 머신러닝 모형의 외표본(테스트자료)에 대한 예측성과를 나타낸 것이다. 우선 벤치마크 로지스틱 회귀모형(LR_base)의 경우 특이도는 0.98로 매우 높지만 민감도는 0으로 나타나 외표본의 부실 개체를 식별하지는 못했다. 이에 반해 로지스틱 회귀모형(LR)을 포함한 머신러닝 모형은 민감도가 0.125~0.75를 나타내어 8개의 부실 개체 중 일부를 식별하는 것으로 나타났다. 이항분류를 고려한 다양한 성과 측정 지표(F1, BA, Gmean, F2, MCC)를 보면 SVML 모형의 예측능력이 높은 것으로 나타났다. 그러나 LR 모형의 예측력도 SVML과 거의 유사하였다. 따라서 기존의 부실예측 통계 모형 즉 로지스틱 회귀모형(LR_base)에 머신러닝의 자료처리 기법이 적용될 경우 예측력이 개선되는 것으로 나타났다. 따라서 기존의 벤치마크 로지스틱 회귀모형도 머신러닝 기법이 적용되면 머신러닝 기법인 LR로 분류되고 다른 머신러닝 모형과 유사한 예측력을 나타냈다.

〈표 9〉 예측성과 비교 : scale(st), smote(y), metric(BA)

	LR base	LR	DT	RF	SVM L	SVM R	ANN	GB	XGB
TN	927	907	916	924	908	901	927	920	922
FP	22	42	33	25	41	48	22	29	27
FN	8	2	4	7	2	2	6	6	6
TP	0	6	4	1	6	6	2	2	2
Accuracy	**0.969**	0.954	0.961	0.967	0.955	0.948	**0.971**	0.963	0.966
Kappa	-0.012	**0.203**	0.166	0.047	**0.207**	0.182	0.114	0.090	0.096
Sensitivity	0	**0.750**	0.500	0.125	**0.750**	**0.750**	0.250	0.250	0.250
Specificity	**0.977**	0.956	0.965	0.974	0.957	0.949	**0.977**	0.969	0.972

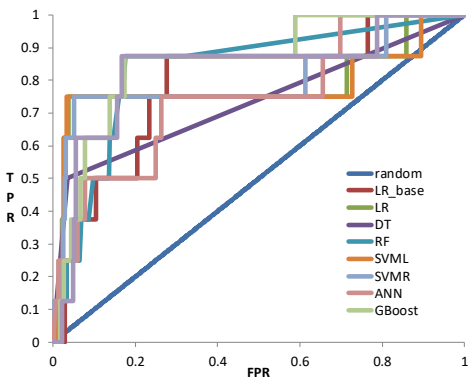
	LR base	LR	DT	RF	SVM L	SVM R	ANN	GB	XGB
Precision	0	**0.125**	0.108	0.038	**0.128**	0.111	0.083	0.065	0.069
F1		**0.214**	0.178	0.059	**0.218**	0.194	0.125	0.103	0.108
BA	0.488	**0.853**	0.733	0.549	**0.853**	0.850	0.613	0.610	0.611
Gmean	0	**0.306**	0.232	0.069	**0.309**	0.289	0.144	0.127	0.131
F2	0	**0.096**	0.059	0.006	**0.098**	0.086	0.024	0.019	0.020
MCC	-0.014	**0.294**	0.220	0.055	**0.298**	0.276	0.132	0.113	0.118
AUROC	0.786	0.785	0.733	0.834	0.779	0.803	0.748	**0.858**	0.800

주: scale(st) : 표준화 변수 스케일 변환, smote(y) : SMOTE 기법 적용, metric(BA) : 전진 교차검증에서 초모수 선택 기준으로 Balanced Accuracy를 선택. 굵게 표시한 숫자는 특정 평가 지표가 가장 높은 모형을 나타내며 두 번째로 높은 모형과 수치 차이가 작으면 두 번째로 높은 모형도 표시함

LR_base 모형과 LR 모형을 비교하면 SMOTE 기법이 적용됨으로써 민감도로 나타나는 부도 식별력(TPR)이 0%에서 75%로 개선되었지만 실제 부도가 아님에도 불구하고 부도로 예측된 경우가 약 2배 정도(22→42) 증가하였다. 이는 부도 예측의 식별력을 높이는 과정에서 일부 잘못 예측되는 빈도 또한 높아질 수 있음을 의미한다.

머신러닝 모형 간 예측성과를 비교하면 일종의 상충 관계가 나타나는데 그것은 부실을 정확히 예측하는 성능(TPR)이 높아짐에 따라 부실이 아닌 개체를 부실로 잘못 분류(FPR)하는 경우도 다소 증가한다는 점이다. 즉 부실예측력을 높이면 높일수록 제1종 오류 즉 부실이라고 예측했지만 실제로는 부실이 아닌 경우의 빈도가 높아질 가능성이 존재한다.

〈그림 5〉 ROC 곡선



임계값에 의존하지 않는 AUROC 값을 보면 GBoost 모형이 가장 높은 수치를 나타내고 있다. 이는 임계값이 고정되지 않은 일반화된 예측력 측면에서 다른 모형보다 낫다는 것을 의미한다. 이 결과는 〈그림 5〉의 모형별 ROC 곡선에서도 알 수 있다. ROC 곡선이 계단 모양으로 나타나는 이유는 외표본의 부실개체수가 8개로 매우 적기 때문이다.

〈표 10〉 학습자료와 테스트자료에 대한 분류행렬 비교 : scale(st), smote(y), metric(BA)

	학습자료				테스트자료			
	TNR	FPR	FNR	TPR	TNR	FPR	FNR	TPR
LR	0.96	0.04	0.25	0.75	0.90	0.10	0.57	0.43
DT	0.97	0.03	0.50	0.50	0.89	0.11	0.40	0.60
RF	0.97	0.03	0.88	0.13	0.96	0.04	0.70	0.30
SVML	0.96	0.04	0.25	0.75	0.91	0.09	0.50	0.50
SVMR	0.95	0.05	0.25	0.75	0.90	0.10	0.53	0.47
ANN	0.98	0.02	0.75	0.25	0.92	0.08	0.57	0.43
GBoost	0.97	0.03	0.75	0.25	0.96	0.04	0.63	0.37
XGBoost	0.97	0.03	0.88	0.13	0.96	0.04	0.67	0.33

주: scale(st) : 표준화 변수 스케일 변환, smote(y) : SMOTE 기법 적용, metric(BA) : 전진 교차검증에서 초모수 선택 기준으로 Balanced Accuracy를 선택

머신러닝 모형을 적용할 때 교차검증을 도입하는 이유는 학습 자료에서 높은 적합도를 보이더라도 테스트자료에서 큰 폭의 예측력 저하를 나타내는 과대적합 문제를 방지하기 위해서이다. 따라서 〈표 10〉과 같이 학습자료와 테스트자료를 대상으로 계산한 모형별 분류행렬의 4가지 비율(TNR, FPR, FNR, TPR)을 비교함으로써 과대적합 여부를 파악할 수 있다. 〈표 10〉을 보면 학습자료와 테스트자료 간 4개의 비율이 전반적으로 유사한 모습을 보이고 있으므로 과대적합의 문제는 없는 것으로 해석된다.

Ⅳ. 민감도 분석

본 연구에서 사용한 학습 대상은 기본적으로 sample-select-scale 순서로 이항 불균형 조정(SMOTE), 변수 선택(부호제약 Lasso 모형), 그리고 변수 스케일 변환(평균과 분산을 각각 0과 1로 표준화)을 거친 자료이다. 설명변수는 부도 시점과 6개월의 시차를 가진 과거변수이고 전진 교차검증에서 초모수 선택의 기준은 BA(balanced accuracy)이다. 이 조건을 기본조건이라고 하자. 이와 같은 기본조건하에서 도출된 결론은 자료 전처리, 변수 선택 및 머신러닝 모형 자체의 효과가 혼합된 것이므로 이들 조건을 개별적으로 변경시켜가며 분석 결과의 강건성(robustness)을 확인할 필요가 있다. 또한 민감도 분석 과정에서 어떤 조건 변화로 인한 결과의 차이가 유의미하다면 해당 조건 변경이 중요하다는 시사점을 얻을 수 있을 것이다.

1. 변수 스케일 변환 유무 및 기법의 변경

머신러닝 모형에 따라 변수 스케일 변화가 미치는 영향은 다르다. 특히 ANN 모형의 경우 변수 스케일 변환이 중요한데 스케일 변환 방법의 적용 유무 및 기법의 변경이 분석 결과의 변화에 미치는 영향을 파악할 필요가 있다. 〈표 11〉은 기본조건 중 스케일 변환이 없는 경우의 예측성과를 나타낸다.

〈표 11〉 예측성과 비교 : scale(n), smote(y), metric(BA)

	LR base	LR	DT	RF	SVM L	SVM R	ANN	GB	XGB
TN	927	907	916	924	908	929		920	922
FP	22	42	33	25	41	20		29	27
FN	8	2	4	7	2	4		6	6
TP	0	6	4	1	6	4		2	2
Accuracy	0.969	0.954	0.961	0.967	0.955	**0.975**		0.963	0.966
Kappa	-0.012	0.203	0.166	0.047	0.207	**0.240**		0.090	0.096
Sensitivity	0	**0.75**	0.5	0.125	**0.75**	0.5		0.25	0.25

	LR base	LR	DT	RF	SVM L	SVM R	ANN	GB	XGB
Specificity	0.977	0.956	0.965	0.974	0.957	**0.979**		0.969	0.972
Precision	0	0.125	0.108	0.038	**0.128**	0.167		0.065	0.069
F1		0.214	0.178	0.059	**0.218**	0.250		0.103	0.108
BA	0.488	**0.853**	0.733	0.549	**0.853**	0.739		0.610	0.611
Gmean	0	0.306	0.232	0.069	**0.309**	0.289		0.127	0.131
F2	0	0.096	0.059	0.006	**0.098**	0.089		0.019	0.020
MCC	-0.014	**0.294**	0.220	0.055	0.298	0.279		0.113	0.118
AUROC	0.786	0.785	0.733	0.834	0.784	**0.898**		0.858	0.800

주: scale(n) : 변수 스케일 변환 없음, smote(y) : SMOTE 기법 적용, metric(BA) : 전진 교차검증에서 초모수 선택 기준으로 Balanced Accuracy를 선택. 굵게 표시한 숫자는 특정 평가 지표가 가장 높은 모형을 나타내며 두 번째로 높은 모형과 수치 차이가 작으면 두 번째로 높은 모형도 표시함

〈표 12〉 예측성과 비교 : scale(01), smote(y), metric(BA)

	LR base	LR	DT	RF	SVM L	SVM R	ANN	GB	XGB
TN	927	907	916	924	907	902	906	920	922
FP	22	42	33	25	42	47	43	29	27
FN	8	2	4	7	2	3	5	6	6
TP	0	6	4	1	6	5	3	2	2
Accuracy	**0.969**	0.954	0.961	0.967	0.954	0.948	0.950	0.963	0.966
Kappa	-0.012	**0.203**	0.166	0.047	**0.203**	0.154	0.098	0.090	0.096
Sensitivity	0	**0.75**	0.5	0.125	**0.75**	0.625	0.375	0.25	0.25
Specificity	**0.977**	0.956	0.965	0.974	0.956	0.950	0.955	0.969	0.972
Precision	0	**0.125**	0.108	0.038	**0.125**	0.096	0.065	0.065	0.069
F1		**0.214**	0.178	0.059	**0.214**	0.167	0.111	0.103	0.108
BA	0.488	**0.853**	0.733	0.549	**0.853**	0.788	0.665	0.610	0.611
Gmean	0	**0.306**	0.232	0.069	**0.306**	0.245	0.156	0.127	0.131
F2	0	**0.096**	0.059	0.006	**0.096**	0.064	0.028	0.019	0.020
MCC	-0.014	**0.294**	0.220	0.055	**0.294**	0.231	0.140	0.113	0.118
AUROC	0.786	0.785	0.733	0.834	0.793	0.804	0.698	**0.858**	0.800

주: scale(01) : 0과 1 범위로 변수 스케일 변환, smote(y) : SMOTE 기법 적용, metric(BA) : 전진 교차검증에서 초모수 선택 기준으로 Balanced Accuracy를 선택. 굵게 표시한 숫자는 특정 평가 지표가 가장 높은 모형을 나타내며 두 번째로 높은 모형과 수치 차이가 작으면 두 번째로 높은 모형도 표시함

스케일 변환을 하지 않은 경우 가장 두드러진 특징은 ANN 모형의 학습에 문제가 발생하여 의미있는 결과가 산출되지 않았다는 점이다. 〈표 12〉는 기존 조건에서

스케일 변환을 0과 1사이로 정규화로 변경한 것이다. 〈표 11〉과 달리 ANN 모형의 결과가 문제없이 산출되고 있다. 그 외의 전반적인 결과는 기존 조건과 크게 다르지 않은 것으로 파악된다.

2. 이항자료 불균형 문제의 조정 여부

머신러닝 모형도 불균형 이항자료를 그대로 학습할 경우 정상개체를 정상으로 분류하는데 치중될 수 있다는 문제점은 이미 널리 알려져 있다. 따라서 SMOTE 기법의 적용 유무에 따른 분석 결과를 기존 결과와 비교함으로써 이항 자료 불균형 조정의 중요성을 파악할 수 있다.

〈표 13〉 예측성과 비교 : scale(st), smote(n), metric(BA)

	LR base	LR	DT	RF	SVM L	SVM R	ANN	GB	XGB
TN	927	927	949	949	949	949	944	942	948
FP	22	22	0	0	0	0	5	7	1
FN	8	8	8	8	8	8	8	8	8
TP	0	0	0	0	0	0	0	0	0
Accuracy	0.969	0.969	**0.992**	**0.992**	**0.992**	**0.992**	0.986	0.984	0.991
Kappa	-0.012	-0.012	0	0	0	0	-0.006	-0.008	-0.002
Sensitivity	0	0	0	0	0	0	0	0	0
Specificity	0.977	0.977	**1**	**1**	**1**	**1**	0.995	0.993	0.999
Precision	0	0					0	0	0
F1									
BA	0.488	0.488	**0.5**	**0.5**	**0.5**	**0.5**	0.497	0.496	0.499
Gmean	0	0					0	0	0
F2	0	0					0	0	0
MCC	-0.014	-0.014					-0.007	-0.008	-0.003
AUROC	0.786	0.786	0.500	**0.876**	0.651	0.809	0.723	0.718	0.860

주: scale(st) : 표준화 변수 스케일 변환, smote(n) : SMOTE 기법 미적용, metric(BA) : 전진 교차검증에서 초모수 선택 기준으로 Balanced Accuracy를 선택. 굵게 표시한 숫자는 특정 평가 지표가 가장 높은 모형을 나타내며 두 번째로 높은 모형과 수치 차이가 작으면 두 번째로 높은 모형도 표시함

〈표 13〉은 기존 조건에서 SMOTE 기법의 적용을 제외한 경우의 예측성과를 나타낸다. 이 경우 모든 모형에서 부실개체를 제대로 식별하지 못하고 있다. 이는 머신러닝 모형도 기존 모형과 같이 부실 개체수가 희소할 경우 정상개체 식별에 높은 비중을 두기 때문이다. 따라서 머신러닝 모형이 의미있는 부실 식별력을 가지기 위해서는 개체 불균형 조정 절차가 필요함을 시사한다.

〈표 14〉 예측성과 비교 : scale(01), smote(n), metric(BA)

	LR base	LR	DT	RF	SVML	SVMR	ANN	GB	XGB
TN	927	927	949	949	949	949	935	942	948
FP	22	22	0	0	0	0	14	7	1
FN	8	8	8	8	8	8	7	8	8
TP	0	0	0	0	0	0	1	0	0
Accuracy	0.969	0.969	**0.992**	**0.992**	**0.992**	**0.992**	0.978	0.984	0.991
Kappa	-0.012	-0.012	0	0	0	0	0.077	-0.008	-0.002
Sensitivity	0	0	0	0	0	0	0.125	0	0
Specificity	0.977	0.977	**1**	**1**	**1**	**1**	0.985	0.993	0.999
Precision	0	0					0.067	0	0
F1							0.087		
BA	0.488	0.488	0.5	0.5	0.5	0.5	**0.555**	0.496	0.499
Gmean	0	0					**0.091**	0	0
F2	0	0					**0.010**	0	0
MCC	-0.014	-0.014					**0.081**	-0.008	-0.003
AUROC	0.786	0.786	0.500	**0.876**	0.411	0.366	0.391	0.720	0.860

주: scale(01) : 0과 1 범위로 변수 스케일 변환, smote(n) : SMOTE 기법 미적용, metric (BA) : 전진 교차검증에서 초모수 선택 기준으로 Balanced Accuracy를 선택. 굵게 표시한 숫자는 특정 평가 지표가 가장 높은 모형을 나타내며 두 번째로 높은 모형과 수치 차이가 작으면 두 번째로 높은 모형도 표시함

〈표 14〉는 기존 조건에서 SMOTE 기법을 제외하되 변수 스케일 조정 기법으로 0과 1사이의 정규화를 선택한 경우의 예측성과를 나타낸다. 이 경우 대부분 모형에서 부실개체를 제대로 식별하지 못하는 것은 〈표 13〉의 결과와 유사하지만 ANN 모형의 부실 개체 1개를 식별했다는 점에서 차이가 있다. 물론 성과지표 상으로는 ANN 모형의 성과가 개선된 것으로 나타났으나 AUROC는 상대적으로 낮고 식별 개체수가 너무 적기 때문에 SMOTE 기법이 적용되지 않은 상태에서 정규화를

통해 ANN 모형의 예측력이 대폭 개선된 것으로 해석하는 것은 다소 어렵다고 판단된다.

3. 전진 교차검증 기법의 변경

기존의 전진 교차검증 기법에 비해 Ⅱ. 방법론의 2. 자료의 분할에서 다루었던 수정된 형태의 전진 교차검증 기법은 테스트자료를 더 확보함으로써 더욱 일반화된 예측성능을 비교할 수 있다는 장점이 있다. 따라서 이 두 가지 기법의 차이에 따른 예측성과를 확인함으로써 수정된 전진 교차검증의 효과를 파악할 수 있다. 〈표 15〉는 수정된 전진 교차검증기법을 사용한 경우의 테스트자료의 예측 결과이다.

〈표 15〉 예측성과 비교 : scale(st), smote(y), metric(BA), fc(nested)

	LR	DT	RF	SVML	SVMR	ANN	GB	XGB
TN	3501	3619	3808	3550	3682	3628	3805	3809
FP	417	299	110	368	236	290	113	109
FN	40	34	66	38	48	54	64	66
TP	39	45	13	41	31	25	15	13
Accuracy	0.886	0.917	**0.956**	0.898	0.929	0.914	**0.956**	**0.956**
Kappa	0.116	**0.187**	0.107	0.140	0.153	0.098	0.124	0.108
Sensitivity	0.494	**0.570**	0.165	0.519	0.392	0.316	0.190	0.165
Specificity	0.894	0.924	**0.972**	0.906	0.940	0.926	0.971	**0.972**
Precision	0.086	**0.131**	0.106	0.100	0.116	0.079	0.117	0.107
F1	0.146	**0.213**	0.129	0.168	0.179	0.127	0.145	0.129
BA	0.694	**0.747**	0.568	0.713	0.666	0.621	0.581	0.568
Gmean	0.205	**0.273**	0.132	0.228	0.213	0.158	0.149	0.132
F2	0.046	**0.079**	0.020	0.056	0.051	0.029	0.026	0.021
MCC	0.170	**0.245**	0.110	0.195	0.185	0.125	0.127	0.111
AUROC	0.574	0.683	**0.794**	0.658	0.571	0.789	0.653	0.776

주: scale(st) : 표준화 변수 스케일 변환, smote(y) : SMOTE 기법 적용, metric(BA) : 전진 교차검증에서 초모수 선택 기준으로 Balanced Accuracy를 선택, fc(nested) : 수정된 전진 교차검증 적용. 굵게 표시한 숫자는 특정 평가 지표가 가장 높은 모형을 나타내며 두 번째로 높은 모형과 수치 차이가 작으면 두 번째로 높은 모형도 표시함

수정된 전진 교차검증을 이용한 예측성과를 보면 F1, BA, Gmean, F2, MCC에서 DT 모형의 예측력이 높은 것으로 나타나고 있다. 전진 교차검증을 이용할 경우 주로 SVML 모형의 예측력이 지배적이었음을 고려하면 테스트자료의 범위를 설정하는 방법이 예측성과에 미치는 영향이 일부 있음을 알 수 있다. 물론 이 경우에도 SVML 모형의 예측력은 DT 모형 다음 순으로 나타나고 있다.

4. 변수 축약 기법의 미적용

머신러닝 기법은 고차원 변수 분석에 적합하므로 사전 변수 선정은 오히려 정보를 제한할 가능성도 존재하므로 전체 변수를 사용함으로써 변수 축약기법의 필요성을 검토할 필요가 있다. 〈표 16〉은 Lasso 변수 축약 기법을 사용하지 않은 테스트자료의 예측 결과이다.

〈표 16〉 예측성과 비교 : scale(st), smote(y), metric(BA), lasso(no)

	LR	DT	RF	SVML	SVMR	ANN	GB	XGB
TN	915	834	904	944	830	929	941	945
FP	34	115	45	5	119	20	8	4
FN	8	7	4	7	7	5	8	6
TP	0	1	4	1	1	3	0	2
Accuracy	0.956	0.873	0.949	**0.987**	0.868	0.974	0.983	**0.990**
Kappa	-0.014	0.000	0.128	0.137	0.000	0.183	-0.008	**0.281**
Sensitivity	0.000	0.125	**0.500**	0.125	0.125	0.375	0.000	0.250
Specificity	0.964	0.879	0.953	**0.995**	0.875	0.979	0.992	**0.996**
Precision	0.000	0.009	0.082	0.167	0.008	0.130	0.000	**0.333**
F1		0.016	0.140	0.143	0.016	0.194		**0.286**
BA	0.482	0.502	**0.726**	0.560	0.500	0.677	0.496	0.623
Gmean	0.000	0.033	0.202	0.144	0.032	0.221	0.000	**0.289**
F2	0.000	0.001	0.045	0.024	0.001	0.054	0.000	**0.091**
MCC	-0.018	0.001	0.187	0.138	0.000	0.210	-0.008	**0.284**
AUROC	0.460	0.470	0.726	**0.811**	0.553	**0.832**	0.635	0.623

주: scale(st) : 표준화 변수 스케일 변환, smote(y) : SMOTE 기법 적용, metric(BA) : 전진 교차검증에서 초모수 선택 기준으로 Balanced Accuracy를 선택, lasso(no) : 변수축약 미적용. 굵게 표시한 숫자는 특정 평가 지표가 가장 높은 모형을 나타내며 두 번째로 높은 모형과 수치 차이가 작으면 두 번째로 높은 모형도 표시함

변수 축약 기법이 적용되지 않은 경우의 예측성과(표 16)는 그렇지 않은 경우(표 9)에 비해 전반적으로 TP가 낮아짐에 따라 MCC와 AUROC가 낮아지는 것으로 나타났다. 즉 전체 변수를 모두 사용한 경우 부실개체의 예측력이 상대적으로 낮아지므로 변수 축약 기법은 예측력 제고를 위해 필요한 과정으로 판단된다.

5. 부호 제약이 없는 Lasso 모형을 이용한 변수 선택

재무금융 관점에서 부실이나 부도 또는 위기 변수에 대한 설명변수는 이론적 또는 실증적으로 검증된 인과 관계를 통해서 영향을 미치므로 이에 대한 이론 또는 기대부호가 존재하는 것이 일반적이다. 그러나 머신러닝 관점에서 볼 때 부호제약이 반드시 필요한 것은 아니다. 그 이유는 부호제약이 추가된 Lasso를 통해 변수를 선택할 경우 상관계수가 높은 자료서의 자연스러운 변수 선택을 방해할 수 있기 때문이다. 또한 구조적인 변화나 국면에 따라 기대부호가 변화할 가능성도 존재한다. 특히 부호제약을 하지 않고 기존의 예상과 반대의 결과가 나왔을 때, 그 원인을 찾아보는 것도 경제학적 해석 못지않은 의미를 가질 수 있다. 따라서 부호제약이 도입되지 않은 일반적인 Lasso 모형으로 변수 선택을 수행한 경우의 예측성과를 확인함으로써 사전적인 부호제약의 효과를 파악할 수 있다. 〈표 17〉은 부호 제약이 없는 Lasso 변수 축약 기법을 사용한 테스트자료의 예측 결과이다.

〈표 17〉 예측성과 비교 : scale(st), smote(y), metric(BA), lasso(w/o sign)

	LR	DT	RF	SVML	SVMR	ANN	GB	XGB
TN	948	844	914	944	871	921	927	931
FP	1	105	35	5	78	28	22	18
FN	7	4	4	7	5	5	7	5
TP	1	4	4	1	3	3	1	3
Accuracy	**0.992**	0.886	0.959	0.987	0.913	0.966	0.970	0.976
Kappa	**0.197**	0.054	0.159	0.137	0.053	0.142	0.053	**0.197**
Sensitivity	0.125	**0.500**	**0.500**	0.125	0.375	0.375	0.125	0.375
Specificity	**0.999**	0.889	0.963	0.995	0.918	0.970	0.977	0.981
Precision	**0.500**	0.037	0.103	0.167	0.037	0.097	0.043	0.143
F1	**0.200**	0.068	0.170	0.143	0.067	0.154	0.065	**0.207**
BA	0.562	0.695	**0.732**	0.560	0.646	0.673	0.551	0.678

	LR	DT	RF	SVML	SVMR	ANN	GB	XGB
Gmean	**0.250**	0.135	0.226	0.144	0.118	0.191	0.074	0.231
F2	**0.068**	0.020	0.056	0.024	0.016	0.041	0.007	0.059
MCC	**0.247**	0.112	0.213	0.138	0.096	0.178	0.061	0.221
AUROC	0.720	0.687	0.732	0.820	0.715	**0.844**	0.658	0.729

주: scale(st) : 표준화 변수 스케일 변환, smote(y) : SMOTE 기법 적용, metric(BA) : 전진 교차검증에서 초모수 선택 기준으로 Balanced Accuracy를 선택, lasso(w/o sign) : 부호 미제약 변수축약 적용. 굵게 표시한 숫자는 특정 평가 지표가 가장 높은 모형을 나타내며 두 번째로 높은 모형과 수치 차이가 작으면 두 번째로 높은 모형도 표시함

부호제약이 없는 경우의 예측성과는 부호 제약이 있는 경우(표 9)와 비교할 때 특정 성과지표 또는 모형에 따라 다소 우열의 변동이 있지만 TP가 소폭 낮아졌고 MCC나 AUROC 또한 다소 낮아졌다. 따라서 기대부호 제약은 예측성과를 개선하는 방향으로 작용하는 것으로 판단된다.

6. AUROC와 MCC를 이용한 최종 결과 비교

예측성과의 우수성을 평가할 때 즉 어떠한 방법론이 더 우수하다고 평가할 때에는 근거가 되는 지표가 필요하다. 특히 부실기업예측과 같이 극단적인 불균형 자료의 경우 AUROC가 판단 기준으로서 적합한 것으로 알려져 있다. 다른 지표들의 경우 임계값을 어떻게 설정하는가에 따라서, 즉 부실기업을 얼마나 엄격하게 정의할 것인가에 따라서 크게 달라질 수 있기 때문이다. AUROC와 함께 이진 불균형 자료에 대한 성과지표로 유용한 것으로 알려져 있는 MCC를 추가적인 판단기준으로 설정하였다. (Chicco; 2017). 〈표 18〉와 〈표 19〉은 각각 Ⅲ 장의 최종 결과 및 Ⅳ 장의 다양한 분석 결과를 AUROC와 MCC를 기준으로 정리한 것이다.

〈표 18〉 AUROC 예측성과 비교

(기준 : scale(st), smote(y), metric(BA), lasso(sign))

	LR	DT	RF	SVML	SVMR	ANN	GB	XGB
최종 결과	0.785	0.733	0.834	0.779	0.803	0.748	0.858	0.800
스케일 미조정	0.785	0.733	0.834	0.784	**0.898**		0.858	0.800
0-1	0.785	0.733	0.834	0.793	0.804	0.698	0.858	0.800
smote 미적용, 표준화	0.786	0.500	**0.876**	0.651	0.809	0.723	0.718	0.860

	LR	DT	RF	SVML	SVMR	ANN	GB	XGB
smote 미적용, 0-1	0.786	0.500	**0.876**	0.411	0.366	0.391	0.720	0.860
수정교차검증	0.574	0.683	0.794	0.658	0.571	0.789	0.653	0.776
전체변수 사용	0.460	0.470	0.726	0.811	0.553	0.832	0.635	0.623
부호미제약 변수선택	0.720	0.687	0.732	0.820	0.715	0.844	0.658	0.729

주: scale(st) : 표준화 변수 스케일 변환, smote(y) : SMOTE 기법 적용, metric(BA) : 전진 교차검증에서 초모수 선택 기준으로 Balanced Accuracy를 선택, lasso(sign) : 부호 제약 변수축약 적용. 굵게 표시한 숫자는 AUROC가 가장 높은 모형을 나타내며 두 번째로 높은 모형과 수치 차이가 작으면 두 번째로 높은 모형도 표시함

〈표 19〉 MCC 예측성과 비교 (기준 : scale(st), smote(y), metric(BA), lasso(sign))

	LR	DT	RF	SVML	SVMR	ANN	GB	XGB
최종 결과	**0.294**	0.220	0.055	**0.298**	0.276	0.132	0.113	0.118
스케일 미조정	**0.294**	0.220	0.055	**0.298**	0.279		0.113	0.118
0-1	**0.294**	0.220	0.055	**0.294**	0.231	0.140	0.113	0.118
smote 미적용, 표준화	-0.014					-0.007	-0.008	-0.003
smote 미적용, 0-1	-0.014					0.081	-0.008	-0.003
수정교차검증	0.170	0.245	0.110	0.195	0.185	0.125	0.127	0.111
전체변수 사용	-0.018	0.001	0.187	0.138	0.000	0.210	-0.008	0.284
부호미제약 변수선택	0.247	0.112	0.213	0.138	0.096	0.178	0.061	0.221

주: scale(st) : 표준화 변수 스케일 변환, smote(y) : SMOTE 기법 적용, metric(BA) : 전진 교차검증에서 초모수 선택 기준으로 Balanced Accuracy를 선택, lasso(sign) : 부호 제약 변수축약 적용. 굵게 표시한 숫자는 MCC가 가장 높은 모형을 나타내며 두 번째로 높은 모형과 수치 차이가 작으면 두 번째로 높은 모형도 표시함

AUROC를 기준으로 보면 변수의 스케일을 조정하지 않은 경우의 SVMR 모형의 예측성과가 가장 높게 나타났다. 특히 SMOTE을 이용한 이항 불균형 조정이 적용되지 않더라도 AUROC 기준의 예측성과가 급격히 낮아지는 것은 아니지만 MCC 기준의 이항불균형 설명력은 크게 낮아졌다. 따라서 이항불균형 조정기법은 예측력 개선을 위해 매우 중요함을 알 수 있다. 이와 유사하게 전체변수를 사용하거나 부호제약을 도입하지 않는 경우에도 AUROC 기준의 예측성과는 소폭 낮아지지만 더욱 문제가 되는 것은 MCC 기준의 이항불균형 설명력이 크게 낮아진다는 것이다. 따라서 변수 축약과 기대부호 제약은 부실개체 예측력을 개선할 수 있는 접근법으로 해석된다.

Ⅴ. 시사점

1. 이항자료 불균형 조정의 중요성

민감도 분석으로부터 알 수 있듯이 머신러닝 모형도 불균형도가 심한 이항자료를 대상으로 학습할 경우 정상 개체 예측에 초점을 둠으로써 부실개체 예측력이 현저히 저하되었다. 따라서 SMOTE 기법과 같은 이항자료의 불균형 조정이 부실예측력 개선의 중요한 요인임을 알 수 있었다. 본 연구는 연구 범위의 제약으로 인해 보다 다양한 이항 불균형 조정 기법을 적용하지 않았으나 실무 적용에서는 더욱 다양한 기법을 적용할 필요가 있을 것이다.

2. 변수 축약과 기대부호 도입의 중요성

본 연구는 최근 변수 선택에서 자주 사용되는 머신러닝 모형 중 하나인 Lasso 모형으로 변수 선택을 하되 경제학적 해석이 가능하도록 이론 부호 제약을 주었다. 특히 Ⅳ 장의 4와 5 절에서 변수 축약이 필요하며 변수 축약 과정에 기대부호 제약조건을 추가할 경우 예측력의 소폭 개선이 나타났다. Lasso 기법을 이용한 변수 선택은 특히 후보 변수가 매우 많을 때 유효하며 다양한 형태로 발전되고 있다. 실제 금융기관의 부실에 영향을 주는 요소는 매우 많고 후보 변수의 과거 값 즉 래그를 고려하면 그 개수는 매우 크게 늘어난다. 이 경우 변수 그룹 선택이 가능한 Group Lasso(Meier *et al.*; 2008)는 특정 설명변수와 그 과거 래그 값을 항상 선택하거나 제외할 수 있으므로 특히 유용할 수 있다.

3. 부실예측력과 제1종 오류 사이의 상충관계

본 연구의 실증분석 결과로부터 부실예측력과 제1종 오류 사이의 상충관계가 SMOTE 기법 적용 여부와 머신러닝 모형 간 예측성과 양쪽에 모두 나타났다. 이와 같은 상충 관계는 머신러닝 모형을 개발하거나 적용할 때 충분히 고려해야할

부분으로 판단된다. 불균형 자료 조정 기법인 SMOTE 기법은 부실 개체의 예측력 즉 검정력($1-\beta$)을 높이지만 이와 동시에 부실이 아닌 개체를 부실로 예측하는 제 1종 오류(α)도 일부 높이는 경향이 있었다.

그럼에도 불구하고 SMOTE 기법을 사용함으로써 동일한 조건하에서 벤치마크 모형이 전혀 식별하지 못한 부실 개체를 최대 50%까지 식별한 실증 분석 결과는 불균형 자료 조정 단계가 매우 중요하다는 것을 시사한다. 결국 머신러닝 모형의 성능을 제고하기 위한 기법을 추가적으로 고려하되 제1종 오류의 증가폭을 낮추기 위한 다양한 모형 탐색 및 튜닝이 필요할 것으로 판단된다.

Ⅵ. 결론

금융기관의 부실예측 고도화를 기존 통계모형에서 머신러닝으로의 전환으로 보는 관점에서 본 연구는 머신러닝 기법의 적용 단계를 상세히 논의하고 저축은행을 예로 들어 사례분석을 하였다. 본 연구가 강조하는 것은 머신러닝 기법이 올바른 방법론적 절차에 따라 적용될 때 비로소 일반화 예측 성능의 개선을 가져올 수 있다는 것이다. 저축은행의 부실예측에 관한 사례를 분석한 결과 머신러닝 기법은 기본적으로 기존의 로지스틱 회귀분석 모형에 비해 부실예측력을 높일 수 있었다. 이와 같은 결과는 일차적으로 불균형 이항자료 조정 기법이 효과적이기 때문이고 그 다음으로 머신러닝 모형 자체의 미세 튜닝에 기인한다.

그러나 머신러닝 기법의 적용이 항상 긍정적인 효과만 가져오는 것은 아니다. 즉 검정력을 높이기 위해 부실예측의 빈도가 다소 높아짐에 따라 오분류율 즉 제1종 오류도 다소 높아졌기 때문이다. 따라서 머신러닝 방법론을 이용함으로써 오분류율을 적은 폭으로 증가시키되 부실예측력을 큰 폭으로 높이기 위해서는 다양한 경우의 미세 튜닝이 필요할 것으로 판단된다. 이는 머신러닝 모형의 고도화뿐만 아니라 후보 변수의 최적화된 선택 그리고 이들 요소를 정보누수가 발생하지 않는 올바른 절차적 프로세스에 적용이 중요함을 의미함을 시사한다.

물론 본 연구도 몇 가지 한계점을 가지고 있다. 첫째, 부실 여부를 나타내는

종속변수를 세분화 할 필요가 있다. 종속변수와 관련하여 본 연구는 6개월 후 저축은행의 부실 여부를 예측하는 것이므로 예측 범위에 있어서 다소 한정적일 수 있다. 예를 들어 종속변수를 1년 또는 2년 안에 부실이 발생할 가능성이나 주요 건전성 지표가 특정 값 이하인 경우를 고려할 수 있을 것이다. 둘째, 후보 설명변수의 범위를 확장할 필요가 있다. 예를 들어 내부 자료를 이용하여 기존 연구에서 그 중요성이 높은 것으로 알려진 대출 비중을 만기, 업종, 익스포저 크기 등으로 세분화하는 것 등을 고려할 수 있을 것이다. 셋째, 다소 계산시간이 많이 소요되는 과정이지만 머신러닝 모형의 미세 튜닝을 위하여 초모수의 탐색 범위를 더욱 늘릴 필요가 있다고 판단된다.

참고문헌

강선민·황인태, “BIS비율과 부채비율: 상호저축은행 부실예측모형,”『경영학연구』, 제42권, 제1호, 2013, pp. 1~27.
(UCI: http://uci.or.kr/I410-ECN-0102-2013-320-002047341)

김남현·김민혁, “저축은행의 레버리지와 부실예측,”『기업경영연구』, 제27권, 제3호, 2020, pp. 145~172.
(UCI: http://uci.or.kr/I410-ECN-0102-2021-300-001114924)

김영기·정신동, “SCOR모형을 활용한 상호저축은행 조기경보시스템 연구,”『금융연구』제19권, 제1호, 2005, pp. 33~67.
(DOI: http://dx.doi.org/10.2139/ssrn.3025819)

김한용·이우주, “불균형적인 이항 자료 분석을 위한 샘플링 알고리즘들: 성능 비교 및 주의점,”『응용통계연구』, 제30권, 제5호, 2017, pp. 681~690.
(UCI: http://uci.or.kr/I410-ECN-0102-2018-300-000646804)

김형준·류두진·조훈, “기업부도예측과 기계학습,”『금융공학연구』제18권, 제3호, 2019, pp. 131~152.
(UCI: http://uci.or.kr/I410-ECN-0102-2021-300-000262401)

남주하·진태홍, “금융기관의 부실화 예측 모형 분석,”『국제경제연구』제4권, 제1호, 1998, pp. 33~57.
(UCI: http://uci.or.kr/I410-ECN-0102-2008-320-001683479)

장영광·김영기, “상호저축은행 경영실태평가지표 타당성 분석 및 도산 예측,”『금융동향』제9권, 제1호, 2004, pp. 1~39.
(UCI: http://uci.or.kr/I410-ECN-0102-2009-320-002781010)

Breiman, L., “Random Forests,” *Machine Learning,* Vol. 45, No. 1, 2001, pp. 5-32.
(DOI: https://doi.org/10.1023/A:1010933404324)

Carmona, P., F. Climent, and A. Momparler, “Predicting Failure in

the US Banking Sector: An Extreme Gradient Boosting Approach," *International Review of Economics and Finance*, Vol. 61, No. 1, 2018, pp. 304-324.
(DOI: https://doi.org/10.1016/j.iref.2018.03.008)

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research*, Vol. 16, No. 1, 2011, pp. 321-357.
(DOI: https://doi.org/10.1613/jair.953)

Chen, T., and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
(DOI: https://doi.org/10.1145/2939672.2939785)

Chicco, D., "Ten Quick Tips for Machine Learning in Computational Biology," *BioData Mining*, Vol. 10, No. 35, 2017, pp. 1-17.
(DOI: https://doi.org/10.1186/s13040-017-0155-3)

Cortes, C. and V. Vapnik, "Support-Vector Networks," *Machine Learning*, Vol. 20, No. 3, 1995, pp. 273-297.
(DOI: https://doi.org/10.1007/BF00994018)

Freund, Y., "Boosting a Weak Learning Algorithm by Majority," *Information and Computation*, Vol. 121, No. 2, 1995, pp. 256–285.
(DOI: https://doi.org/10.1006/inco.1995.1136)

Freund, Y. and R. E. Schapire, "A Decision-theoretic Generalization of On-line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, Vol. 55, No. 1,

1997, pp. 119–139.
(DOI: https://doi.org/10.1006/jcss.1997.1504)

Friedman, J., "Greedy Boosting Approximation: A Gradient Boosting Machine," *Annals of Statistics*, Vol. 29, No. 5, 2001, pp.1189–1232.
(DOI: https://doi.org/10.1214/aos/1013203451)

Hastie, T., R. Tibshirani, and JH. Friedman, *The Elements of Statistical Learning*, corrected edition, New York: Springer, 2003.
(URL: https://link.springer.com/book/10.1007/978-0-387-84858-7)

Hulse, J. V., T. M. Khoshgoftaar, and A. Napolitano, "Experimental Perspectives on Learning from Imbalanced Data," In Proceedings of the 24th International Conference on Machine Learning, 2007, pp. 935-942.
(DOI: https://doi.org/10.1145/1273496.1273614)

James, G.M., C. Paulson, and P. Rusmevichientong, "The Constrained Lasso," Technical report, University of Southern California, 2013.
(URL: http://www-bcf.usc.edu/~rusmevic/psfiles/CLasso.pdf)

Petropoulos, A., V. Siakoulis, E. Stavroulakis, and N. E. Vlachogiannakis, "Predicting Bank Insolvencies using Machine Learning Techniques," *International Journal of Forecasting*, Vol. 36, No. 3, 2017, pp. 1092-1113.
(DOI: https://doi.org/10.1016/j.ijforecast.2019.11.005)

Raschka, S., "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning," 2018, working paper.
(URL: https://arxiv.org/abs/1811.12808)

Suss, J. and H. Treitel, "Predicting Bank Distress in the UK with Machine Learning," Staff Working Paper No. 831, 2019, Bank of England.
(URL: https://www.bankofengland.co.uk/working-paper/2019/predicting-bank-distress-in-the-uk-with-machine-learning)

Meier, L., S. van de Geer, and P. Bühlmann, "The Group Lasso for Logistic Regression," *Journal of the Royal Statistical Society*, Vol. 70, No. 1, 2008, pp. 53-71.
(DOI: https://doi.org/10.1111/j.1467-9868.2007.00627.x)

〈부록〉

〈표〉 설명변수의 전체 및 개체 구분별 기초통계량

변수	전체표본		정상개체		부실개체	
	평균	표준편차	평균	표준편차	평균	표준편차
BIS기준자기자본비율	13.8	15.8	14.1	15.7	-0.7	14.9
기본자본비율	10.9	20.2	11.1	20.2	-1.8	14.3
단순자기자본비율	9.4	13.1	9.6	13.1	-1.7	11.3
예수금부채비율	14.2	104.7	13.3	91.9	58.3	376.3
레버리지비율	14.9	110.9	14.0	98.1	59.2	390.2
예대율	124.1	81.5	123.9	81.8	132.3	60.5
순고정이하여신비율	7.5	7.2	7.3	7.0	18.2	11.9
연체율	14.3	11.3	14.0	11.0	31.9	13.2
고정이하여신비율	12.6	10.9	12.2	10.5	30.2	16.3
대손충당금적립비율총여신대비	7.3	6.2	7.1	6.0	16.4	9.4
대손충당금적립비율요적립액대비	115.8	46.5	116.1	46.9	100.9	1.7
대손충당금적립비율고정이하여신대비	85.0	334.2	85.6	337.4	57.4	16.8
총대출비중	75.4	16.0	75.4	15.9	78.4	21.4
담보대출비율	76.7	19.0	76.9	19.0	70.5	17.7
보증대출비율	2.3	6.2	2.3	6.2	1.5	4.0
신용대출비중	19.4	16.2	19.2	16.2	26.4	17.1
비업무용자산비율	2.7	7.0	2.7	7.0	3.4	3.7
총자산이익률ROA	-0.1	2.9	0.0	2.3	-7.0	10.1
자기자본순이익률ROE	-5.3	397.2	-1.8	319.5	-184.9	1753.1
수지비율	103.9	74.1	102.2	71.4	193.8	135.2
총자산경비율	2.2	24.2	2.3	24.4	1.1	0.8
영업이익경비율	51.1	317.5	50.7	320.0	73.6	123.3
자본잉여금증감_자기자본대비	0.0	1.2	0.0	1.2	0.0	0.2
이익잉여금증감_자기자본대비	-0.1	4.0	0.0	3.2	-2.0	17.5
유동성비율	157.3	130.7	157.6	131.6	144.1	64.6
유형자산비율	81.7	598.2	79.8	582.2	176.7	1163.5
실가용자금비율	21.0	17.7	21.0	17.7	21.0	16.1
수신증가율	2.2	10.8	2.4	10.7	-5.0	14.0
역고정자산비율	61.7	624.5	58.9	627.7	204.6	403.7
GDP성장률yoy	3.5	2.1	3.5	2.1	3.9	1.9
관리재정수지GDP비율	-1.3	1.0	-1.3	1.0	-1.1	0.6
주택매매가격지수증감률	1.5	1.8	1.5	1.8	1.8	1.9
기간스프레드	0.5	0.3	0.5	0.3	0.6	0.3
신용스프레드	5.9	1.4	5.9	1.4	6.1	1.2
ln총자산	12.8	1.1	12.8	1.1	13.4	1.1

A Transition to Financial Distress Prediction Machine Learning Model and Determinants of Forecast Accuracy

Sang-Heon Lee*

〈Abstract〉

The purpose of this study is to present a machine learning process for predicting the insolvency of financial institutions, and to analyze the effect of machine learning techniques and models through empirical analysis. In this process, forward chaining cross-validation and SMOTE technique are applied to mitigate over-fitting and imbalanced classification problems respectively. From the empirical analysis using Korean savings bank data from 2008 to 2019, machine learning models can increase bankruptcy prediction power compared to existing logistic regression model. In particular, to account for expected sign constraints, a sign-restricted LASSO model is used. From the sensitivity analysis, I find that the most significant factor in improving the prediction power is the treatment of the imbalanced binary data problem. Therefore, in order to improve the financial distress prediction power using machine learning, it is necessary to put more emphasis on imbalanced data sampling techniques.

Keywords: Machine learning, Savings bank, Financial Distress Prediction, Imbalanced classification data, SMOTE

JEL Classification: G12, G13, G24

* KIS Pricing, New Business Development Group. Address: 38, Gukjegeumyung-ro 6-gil, Yeongdeungpo-gu, Seoul, 07328, Republic of Korea; E-mail: shlee725@gmail.com.