경영정보학연구
제26권 제2호, 2024년 5월 http://dx.doi.org/10.14329/isr.2024.26.2.169

부도 예측 모형 연구: 부도 데이터의 불균형 문제를 중심으로

A Study on Default Prediction Model: Focusing on The Imbalance Problem of Default Data

박 진 수 (Jinsoo Park) 동아대학교 경영정보학과 석사과정
이 강 배 (Kangbae Lee) 동아대학교 경영정보학과 교수
조 용 복 (Yongbok Cho) 동아대학교 경영정보학과 조교수, 교신저자

요 약

본 연구는 부도 예측 모형을 구축할 때 반드시 고려해야 하는 관측된 부도 데이터의 불균형 문제에 대한 개선 방안을 정리하고, 데이터 리샘플링 기법과 부도 임계치 조정에 따른 모형의 성능 개선 효과를 비교 분석한다. 실증분석 결과 데이터의 불균형 해소 수준이 높을수록, 그리고 모형의 부도 임계치가 낮을수록 모형의 민감도가 개선되는 것을 발견하였으며, 데이터의 불균형 해소 수준이 낮을수록, 그리고 모형의 부도 임계치가 높을수록 모형의 정밀도가 개선되는 것을 발견하였다. 또한 민감도 또는 정밀도 중 한 가지 지표만을 중심으로 불균형 문제를 개선할 경우, 상충 관계로 인해 나머지 성능 평가 지표가 지나치게 낮아지는 현상을 확인하였다. 본 연구는 기존 선행 연구와는 달리 부도 데이터의 불균형 문제 개선 방안과 부도 예측 모형의 성능 개선 효과의 관계에 초점을 두고 있다는 점에서 시사점을 찾을 수 있다. 또한 부도 예측 모형의 실무적 활용도 제고를 위해 모형의 활용 목적에 따라 불균형 문제 개선 방안을 달리 적용하는 것이 바람직하며, 모형의 주된 성능 평가 지표로는 F_β Score를 활용해야 할 필요가 있음을 확인하였다.

키워드 : 부도 모형, 데이터 불균형, Resampling, Decision Threshold, F_β Score

Ⅰ. 서 론

2020년 3월 세계보건기구의 코로나19 팬데믹 선언 이후, 전 세계는 경제 위기 극복과 금융 시스템 안정을 위해 막대한 유동성을 공급하였다. 하지만 이는 높은 수준의 인플레이션 문제로 이어졌고, 엔데믹 시대의 도래와 함께 각국의 금융 당국은 인플레이션 문제 해결을 위해 본격적인 기준금리 인상 및 유동성 회수 조치를 시작하였다.

한국은행 또한 2021년 8월을 기점으로 지속적인 기준금리 인상을 단행하였으며, 2023년 2월 이후 기준금리를 동결하고 있지만 2023년 11월 통화

† 이 논문은 동아대학교 교내연구비 지원에 의하여 연구되었음. 본고는 박진수 2022학년도 석사학위 논문 「부도 데이터의 불균형 문제 개선 방안 연구」의 일부를 발췌하여 재작성한 것임.

정책방향 관련 기자간담회를 통해 '인플레이션 둔화 흐름 등을 면밀히 점검하는 가운데 추가 인상 필요성을 판단한 것임'을 언급하는 등, 상당 기간 고금리 기조를 지속할 것을 시사하고 있다.

지속적인 기준금리 인상은 대출금리의 상승으로 이어졌고, 대표적인 여신 건전성 지표인 연체율 또한 크게 상승하였다. 이에 금융감독원은 2023년 8월 금융상황 점검회의를 통해 금융회사의 안정적 경영과 건전성 관리의 중요성을 강조하는 등, 여신 건전성 관리 이슈에 크게 주목하고 있다.

실무적인 관점에서 금융기관의 여신 건전성 관리 방안은 우량 차주 중심의 선별적인 여신을 취급하는 사전적 관리 방안과, 여신을 취급한 기존 차주 중 상대적으로 불량한 차주를 중점적으로 관리하는 사후적 관리 방안으로 구분할 수 있다. 하지만 2023년 11월 개최된 금융위원장 · 금감원장 · 은행장 간담회를 통해 확인할 수 있듯이, 금융기관의 자금 중개 기능과 사회적 역할 강화가 강조되는 분위기 속에서 사전적 건전성 관리 방안의 활용은 제한적일 수밖에 없기에 사후적 건전성 관리 방안의 중요성이 큰 상황이다.

사후적 건전성 관리의 가장 대표적인 방안으로 활용할 수 있는 부도 예측 모형의 경우, 모형 구축 시 데이터의 불균형 문제를 반드시 고려해야 한다. 부도 예측 모형의 목적 변수인 부도는 일반적으로 신용카드 이상 거래와 같이 매우 낮은 빈도로 발생하는 대표적인 불균형 데이터이며(Haixiang *et al.*, 2017), 이러한 불균형 데이터를 분류 모형에 그대로 적용할 경우 모형의 성능이 심각하게 저해될 수 있기 때문이다(He and Garcia, 2009). 부도 예측 모형을 다루고 있는 연구들은 대부분 부도 데이터의 불균형 문제를 고려하고 있으나, 부도 예측 모형의 성능 개선에만 초점을 두고 있는 경우가 많다.

본 연구는 부도 예측 모형의 성능 자체보다는 부도 데이터의 불균형 문제 개선 방안에 따른 성능 개선 효과를 비교 분석하는 것을 중심으로 하고 있다. 그리고 불균형 데이터를 다루는 분류 모형에서 널리 활용되고 있는 성능 평가 지표인 민감도 및 정밀도의 실무적 한계점을 지적하고, 모형의 활용 목적에 가장 적합한 불균형 문제 개선 방안을 선택할 수 있도록 F_β Score를 성능 평가 지표로 활용할 것을 제안한다. 이를 통해 부도 데이터의 불균형 문제 개선 방안에 관한 종합적인 관점을 제시하고, 부도 예측 모형의 실무적 활용도를 제고할 수 있을 것으로 기대한다.

이후 논문의 구성은 다음과 같다. 먼저 제Ⅱ장에서 부도 데이터의 불균형 문제와 관련된 이론적 배경을 서술한다. 이어서 제Ⅲ장과 제Ⅳ장에서는 연구 방법과 연구 결과에 대해 소개하고, 마지막으로 제Ⅴ장에서 본 연구의 결론과 시사점을 정리하였다.

Ⅱ. 이론적 배경

2.1 부도 예측 모형

부도가 예상되는 차주를 선별하는 부도 예측 모형은 Altman(1968)이 다변량 판별분석을 이용한 모형을 제안한 이래 Ohlson(1980)의 로짓 분석을 활용한 모형, Zmijewski(1984)의 프로빗 모형 등으로 발전하였다.

전통적인 통계 모형에 기반한 연구 이후로는 머신러닝 알고리즘을 활용한 부도 예측 모형 연구가 이루어졌다. Messier and Hansen(1998)은 Decision Tree를 활용한 연구를 진행하였고, Shin *et al.* (2005)은 Support Vector Machine을 활용한 부도 예측 모형을 제안하였다. 하지만 이와 같은 초기 선행 연구는 부도 예측 모형에 적용한 기법에 따른 모형의 성능 비교에만 초점을 맞추고 있다는 한계가 있다(김명종, 윤우섭, 2022). 즉, 부도 예측 모형의 성능 저하 요인 중 하나인 부도 데이터의 불균형 문제는 중요한 주제로 다루어지지 않았으며, 2000년대 초반 이후에 접어서야 불균형 문제 개선을 다루는 연구가 본격적으로 주목받기 시작하였다(김명종, 윤우섭, 2022).

Zhou(2013)는 Oversampling 기법 및 Under sampling 기법을 통한 부도 예측 모형의 성능을 비교하는 연구를 진행하였고, Kim *et al.*(2015)은 기하평균 정확도를 중심으로 하는 GMBoost를 제안하며 불균형 데이터로 이루어진 기업 부도 예측 문제를 연구하였다. Zhou *et al.*(2019)은 GBDT, XGBoost, Light-GBM을 활용하여 P2P 대출 부도 데이터의 불균형 문제를 연구하였고, 조성임, 김명종(2022)은 기업 부도 데이터의 불균형 문제 해결을 위해 비대칭 마진 Support Vector Machine의 최적화를 제안하기도 하였다.

2.2 데이터 불균형 문제 개선

현실 세계에 존재하는 데이터의 경우, 목적 변수의 범주 간 관측 빈도에는 불균형이 존재할 수밖에 없다. 따라서 엄밀히 말하자면 모든 데이터는 불균형 데이터로 볼 수 있지만(He and Garcia, 2009), 일반적으로 소수 범주에 해당하는 데이터의 수가 다수 범주에 해당하는 데이터 수의 10% 이하인 경우를 불균형 데이터로 본다(He and Garcia, 2009; 조용복 등, 2022).

서론에서 언급한 것처럼, 데이터 불균형 문제를 개선하지 않고 분류 모형을 구축할 경우 성능 저하 문제가 발생할 수 있다. 이는 대부분의 분류 모형 알고리즘이 데이터가 균형 잡힌 분포를 가지고 있거나, 오분류로 인해 발생할 수 있는 비용이 동일하다고 가정하기 때문이다(He and Garcia, 2009). 하지만 실제 데이터는 분포가 불균형한 경우가 많고, 오분류 비용이 동일하다는 가정 또한 현실적이지 않다. 가령 금융 사기 거래 탐지 문제의 경우, 사기 거래를 정상 거래로 식별하는 경우가 그 반대의 경우보다 더 큰 손실이 발생할 가능성이 높을 것이다(Haixiang *et al.*, 2017).

데이터 불균형 문제는 금융 분야 외에도 질병 진단(의료), 기계 내부 결함 진단(산업) 등 다양한 분야에서 발생하고 있으며, 학계와 산업계는 불균형 데이터 문제 연구를 통해 분류 모형의 성능을 크게 개선하고 있다(Haixiang *et al.*, 2017).

데이터 불균형 문제 개선 방안은 크게 데이터 전처리 방안(Preprocessing techniques)과 비용 민감 학습(Cost-sensitive learning) 구축으로 구분할 수 있으며(Haixiang *et al.*, 2017), 구체적인 내용은 다음과 같다.

2.2.1 데이터 전처리 방안

데이터 전처리 방안은 모형 구축 전 데이터의 불균형을 개선하는 것으로, Resampling 기법과 Feature Selection and Extraction 기법으로 구분할 수 있다(Haixiang *et al.*, 2017). 데이터 전처리 방안 구분 및 각각의 대표적인 예시는 <표 1>을 통해 확인할 수 있다.

Resampling 기법은 다수 범주 데이터의 제거 또는 소수 범주 데이터의 생성을 통해 데이터 분포를 조정하는 기법을 말한다. Resampling 기법은 분류 모형 알고리즘의 종류에 따른 영향을 받지 않는다는 장점이 있으며(López *et al.*, 2013), Oversampling, Undersampling, Hybrid sampling 기법으

〈표 1〉 데이터 전처리 방안 구분 및 예시

구분		예시
Resampling	Oversampling	ROS(Random Oversampling), SMOTE(Synthetic Minority Over-sampling Technique)
	Undersampling	RUS(Random Undersampling), Near-Miss
	Hybrid sampling	SMOTE-ENN, SMOTE-Tomek Links
Feature Selection and Extraction	Feature Selection	Filter, Wrapper, Embedded
	Feature Extraction	PCA(Principal Component Analysis)

로 나눌 수 있다.

Oversampling 기법은 소수 범주 데이터를 생성하여 불균형 문제를 개선하는 기법으로, 소수 범주 데이터를 임의로 생성하는 ROS(Random Oversampling) 기법과 Chawla *et al.*(2002)이 제안한 SMOTE (Synthetic Minority Over-sampling Technique) 기법 등이 널리 쓰이고 있다(Haixiang *et al.*, 2017).

Oversampling 기법은 소수 범주 데이터 생성 과정에서 과적합(Overfitting) 문제가 발생할 수 있다는 단점이 알려져 있다(López *et al.*, 2013). 이를 해결하기 위한 기법으로 Adaptive Synthetic Sampling (He *et al.*, 2008) 기법 및 Borderline-SMOTE(Han *et al.*, 2013) 기법 등이 제안되었으며(López *et al.*, 2013), 노정담, 최병구(2022)는 순환 생산적 적대 신경망을(Cycle GAN)을 SMOTE 기법과 결합하는 방안을 제시하기도 하였다.

Undersampling 기법은 다수 범주 데이터 제거를 통해 불균형 문제를 개선하는 기법으로, 다수 범주의 데이터를 임의로 제거하는 RUS(Random Undersampling) 기법이 대표적이다.

Undersampling 기법은 데이터 제거 과정에서 정보 손실이 발생할 수 있다는 단점이 있으며(López *et al.*, 2013), 정보 손실 문제를 최소화하면서 데이터의 불균형을 완화할 수 있는 Near-Miss 기법 등이 제안되었다(Mani and Zhang, 2003).

Hybrid Sampling 기법은 소수 범주 데이터 생성과 다수 범주 데이터 제거가 함께 이루어지는 기법을 말하며, Cateni *et al.*(2014)과 Dubey *et al.*(2014) 등은 Oversampling 및 Undersampling 기법을 결합하는 연구를 진행하였다.

한편, Feature Selection 기법은 데이터의 전체 특성 변수 중 일부 특성 변수를 선택하는 것으로 일반적으로 필터(Filter), 래퍼(Wrapper), 임베디드(Embedded) 기법으로 구분할 수 있으며, 대부분 분류 모형의 성능 개선을 위해 활용된다. 데이터의 불균형 문제가 심각한 경우 소수 범주 데이터는 잡음(Noise)으로 여겨질 수 있으나, Feature Selection을 통해 소수 범주와 관련 없는 불필요한 특성 변수를 제거할 경우 이 문제를 개선할 수 있다고 알려져 있다(Li *et al.*, 2016).

Feature Extraction 기법은 데이터 특성 변수의 차원을 축소하는 기법으로, 전체 특성 변수 중 일부 특성 변수를 선택하는 Feature Selection 기법과는 달리 새로운 특성 변수를 생성한다는 차이점이 있다. 가장 대표적인 예로는 PCA(Principal Component Analysis)가 있으며, 이미지나 텍스트 등과 같은 비정형 데이터를 다룰 때 자주 사용된다(Haixiang *et al.*, 2017).

2.2.2 비용 민감 학습

비용 민감 학습은 머신러닝 알고리즘 단계에서 데이터 불균형 문제를 개선하는 것으로, Haixiang *et al.*(2017)에 따르면 비용 민감 학습은 <표 2>와 같이 크게 세 가지로 구분할 수 있다.

〈표 2〉 비용 민감 학습 구분 및 내용

구분	내용
Methods based on training data modification	Modifying the decision thresholds or assigning weights to instance when resampling the training dataset according to the cost decision matrix
Changing the learning process or learning objective to build a cost-sensitive classifier	Modifying the objective function of SVM/ELM using a weighting strategy
	Tree-building strategies that could minimize misclassification costs
	Integrating a cost factor into the fuzzy rule-based classification system
	Cost sensitive error function on neural network
	Cost-sensitive boosting methods
Methods based on Bayes decision theory	Incorporating cost matrix into Bayes based decision boundary

먼저, Decision Threshold는 분류 문제 해결을 위한 양성 판정 기준을 의미한다. 머신러닝 알고리즘은 목적 변수(Y)가 양성(Y=1)일 확률($p(Y=1|X=x_1, x_2, x_3, ..., x_k)$)을 계산한 뒤, 그 확률을 Decision Threshold와 비교하여 양성 여부를 판정한다. 분류 모형에 별도의 분류 기준을 적용하지 않을 경우 양성 판정의 기준이 되는 Decision Threshold는 일반적으로 50%, 즉 목적변수가 양성일 확률(P)이 0.5로 설정된다. 하지만 데이터의 분포가 불균형한 경우 50%로 설정된 Decision Threshold는 최적값이 아닐 수 있으며(Esposito *et al.*, 2021), Decision Threshold의 조정을 통해 불균형 데이터 문제를 개선할 수 있다. 이와 같은 조정은 분류 모형 구축 시 대부분의 머신러닝 알고리즘에 쉽게 적용할 수 있으며, 불균형 데이터 문제 개선에 큰 성과를 보이는 것으로 알려져 있다(Sheng and Ling, 2006).

Cost-sensitive Classifier 구축은 소수 범주 데이터에 대한 오분류 비용을 상대적으로 크게 조정하여 분류 모형의 비용 민감도를 높이는 것을 의미한다. 이는 앞서 언급한 바와 같이, 소수 범주 데이터에 대한 오분류 비용은 다수 범주 데이터에 대한 오분류 비용보다 큰 경우가 많기 때문이다(Li *et al.*, 2016). Cost-sensitive Classifier 구축은 크게 앙상블 기반 분류 모형을 구축하거나, 머신러닝 알고리즘의 목적 함수(Objective function) 또는 비용 함수(Cost function)를 수정하는 것으로 분류할 수 있다(Haixiang *et al.*, 2017). 앙상블 기반 분류 모형 구축을 위한 기법으로는 배깅(Bagging), 부스팅(Boosting) 등이 있으며, 단일 분류 모형에 비해 모형의 성능이 더욱 뛰어난 것으로 알려져 있다(Barboza *et al.*, 2017). 배깅은 동일한 알고리즘을 통해 복수의 분류 모형을 구축한 뒤 그 결과를 결합하여 활용하는 기법을 말한다. Random Forest 알고리즘은 배깅 기법을 활용한 앙상블 기반 분류 모형 중 가장 대표적인 예시로, 모형의 과적합을 줄이는 데 효과적이다(Mellor *et al.*, 2015). 부스팅은 분류 모형의 순차적인 학습을 통해 성능을 개선하는 기법을 말하며, XGBoost 알고리즘이 가장 대표적인 예시이다. 부스팅 기법은 오분류된 관측치에 높은 가중치를 부여하여 데이터 불균형 문제를 효과적으로 개선하는 것으로 알려져 있다(Mellor *et al.*, 2015).

모형의 매개변수에 대한 사전 확률을 가정하는 베이즈 결정 이론(Bayes Decision Theory)을 통한 불균형 데이터 문제를 개선 또한 비용 민감 학습으로 분류할 수 있다. 이와 관련된 예시로는 Datta and Das(2015)가 제안한 NBSVM(Near-Bayesian Support Vector machine) 등이 있다.

이러한 비용 민감 학습은 데이터 전처리 방안, 특히 Resampling 기법에 비해 효율적인 계산이 가능하므로 많은 양의 데이터를 다루기에 적합하지만, 목적 함수 또는 비용 함수를 수정하거나 학습 알고리즘을 조정하는 것이 상대적으로 어렵다는 단점이 있다(Haixiang *et al.*, 2017).

2.3 성능 평가 지표

일반적으로 이진 분류 모형은 <표 3>과 같은 혼동 행렬을 통해 TN(True Negative), TP(True Positive), FN(False Negative), FP(False Positive) 값을 계산한 뒤, 정확도(Accuracy), 민감도(Recall), 정밀도(Precision) 등의 지표로 성능을 평가한다.

〈표 3〉 혼동 행렬

구분		예측	
		양성 (Positive)	음성 (Negative)
실제	양성 (Positive)	TP (True Positive)	FN (False Negative)
	음성 (Negative)	FP (False Positive)	TN (True Negative)

Weiss(2004)에 의하면 정확도는 다수 범주 데이터에 더 많은 가중치를 부여하므로, 부도 예측 모형과 같이 소수 범주 데이터에 집중하는 분류 모형에는 적합하지 않을 수 있다. 따라서 본 연구에

서는 민감도와 정밀도를 기준으로 모형의 성능을 평가하고자 한다.

한편, 민감도와 정밀도는 상충 관계에 있으므로(Buckland and Gey, 1994) 불균형 데이터를 다루는 분류 모형의 정확한 성능 평가를 위해서는 두 지표를 함께 고려할 필요가 있다. F_β Score는 민감도와 정밀도의 상대적인 가중치 조절을 통해 두 지표를 종합적으로 고려하면서 자유로운 변형이 가능하다는 장점이 있으며, F_β Score 최적화와 관련된 연구(Dembczynski *et al.*, 2013; Musicant *et al.*, 2003; Nan *et al.*, 2012) 또한 중요한 연구 영역으로 여겨지고 있다.

따라서 본 연구에서는 민감도와 정밀도의 조화평균 값인 F_1 Score를 주된 성능 평가 지표로 활용할 것이다. 민감도, 정밀도, 및 F_β Score의 산식은 아래 식 (1)~식 (3)을 통해 확인할 수 있다.

$$Recall = \frac{TP}{TP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$F_\beta\ Score = \frac{(1+\beta^2)\times Recall \times Precison}{Recall + (Precision \times \beta^2)} \quad (3)$$

Ⅲ. 연구 방법

3.1 연구 데이터

실증 분석을 위한 데이터는 캐글(Kaggle)에서 제공하는 Loan Default Prediction Dataset을 활용하였다. 이 데이터는 부도 예측 모형 구축을 위한 학습용 데이터로 코세라(Coursera)의 Data Science Coding Challenge를 통해 소개되었으며, 결측치 및 이상치가 존재하지 않는다는 특징이 있다. 따라서 결측치와 이상치 처리 방식에 따른 모형 성능 차이 문제를 피할 수 있다는 장점이 있으며, 총 255,347개의 데이터 중 부도 데이터는 29,653개로 약 13.14% 수준의 불균형 비율을 보이는 전형적인 불균형 데이터라는 점을 감안할 때 본 연구에 적합하다고 판단된다.

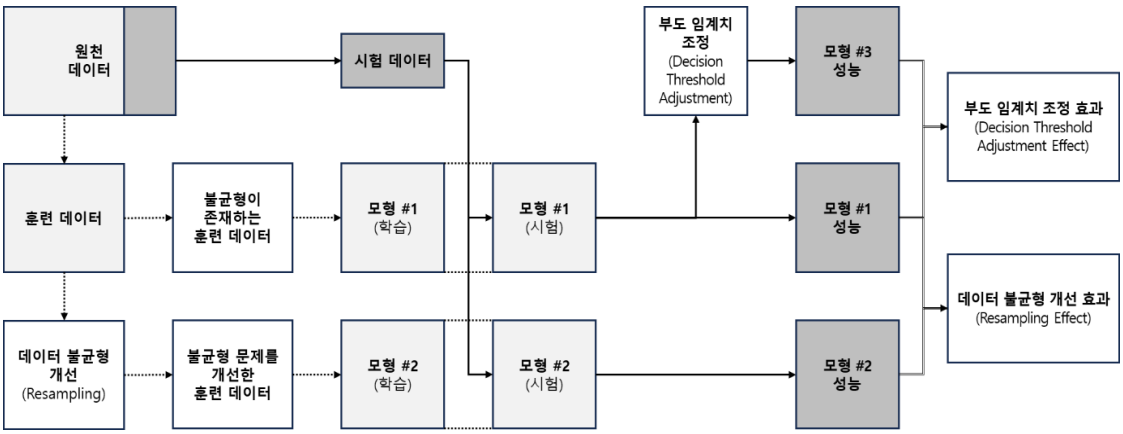
목적 변수인 부도 여부를 제외한 연구 데이터의 특성은 <표 4>를 통해 확인할 수 있다.

〈표 4〉 연구 데이터 특성

특성	내용
LoanID	고유 식별자
Age	연령
Income	연소득
LoanAmount	대출 금액
CreditScore	신용평점
MonthsEmployed	고용 기간
NumCreditLines	대출계좌 수
InterestRate	대출 금리
LoanTerm	대출 기간
DTIRatio	총부채상환비율
Education	교육 수준
EmploymentType	고용 상태
MaritalStatus	혼인 여부
HasMortgage	담보대출 유무
HasDependents	자녀 유무
LoanPurpose	자금용도
HasCoSigner	보증인 유무

3.2 연구 절차

본 연구의 절차는 다음과 같다. 먼저 훈련 데이터와 시험 데이터를 분리한 뒤, 훈련 데이터를 불균형이 존재하는 훈련 데이터와 Resampling 기법(이하 데이터 리샘플링 기법)을 통해 불균형 문제를 개선한 훈련 데이터로 구분하여 머신러닝 알고리즘을 학습한다. 이후 시험 데이터를 불균형이 존재하는 훈련 데이터를 학습한 모형(모형 #1)과 불균형 문제를 개선한 훈련 데이터를 학습한 모형(모형 #2)에 적용한 뒤, 각 모형의 부도 예측 성능(모형 #1 성능, 모형 #2 성능)을 평가한다. 이어서 불균형이 존재하는 훈련 데이터를 학습한 모형



〈그림 1〉 연구 절차

(모형 #1)의 Decision Threshold(이하 부도 임계치)를 조정하여 모형의 성능(모형 #3 성능)을 평가한다. 이를 통해 부도 예측 모형의 데이터 불균형 개선 효과(Resampling Effect)와 부도 임계치 조정 효과(Decision Threshold Adjustment Effect)를 비교 분석할 것이며, <그림 1>을 통해 전반적인 연구 절차를 파악할 수 있다.

분석에 활용된 데이터 리샘플링 기법은 총 6가지로 <표 5>를 통해 확인할 수 있으며, 적용 비율은 5% 단위(15%~100%)로 설정하였다. 부도 임계치 또한 5% 단위(5%~95%)로 조정하며 분석을 진행하였다. 모형에 적용한 머신러닝 알고리즘은 분류 문제에 널리 쓰이고 있는 Logistic Regression, Random Forest, XGBoost를 활용하였으며, 초모수 값은 별도로 조정하지 않았다.

〈표 5〉 데이터 리샘플링 기법

구분	상세
Oversampling	ROS
	SMOTE
Undersampling	RUS
	Near-Miss
Hybrid sampling	SMOTE-ENN
	SMOTE-Tomek

Ⅳ. 연구 결과

4.1 데이터 리샘플링 기법 적용

<표 6>은 불균형이 존재하는 훈련 데이터와 데이터 리샘플링 기법을 통해 불균형 문제를 개선한 훈련 데이터를 학습한 모형의 성능을 요약한 자료이다.

데이터 리샘플링 기법 적용 비율이 높아질수록 범주 간 불균형 해소 수준이 높아지는 것을 확인할 수 있으며, 모형의 민감도는 상승하는 반면 정밀도는 하락하는 경향을 발견할 수 있다. 예를 들어 ROS 기법과 Logistic Regression 알고리즘을 활용한 모형의 성능 평가 지표를 살펴보면, ROS 기법을 15% 비율로 적용할 경우 민감도는 2.05%, 정밀도는 57.05% 수준이나 ROS 기법을 100% 비율로 적용할 경우 민감도는 65.21%로 상승하는 반면, 정밀도는 20.36%로 하락하는 결과를 확인할 수 있다. 이는 데이터 리샘플링 기법 적용 비율이 높아질수록 모형을 통해 실제 부도 차주를 정상 차주로 잘못 예측할 가능성이 낮아지는 반면, 부도로 예측한 차주가 실제 부도 차주일 가능성 또한 낮아진다는 것으로 해석할 수 있다.

한편, 데이터 리샘플링 기법 적용 비율이 낮아질수록 범주 간 불균형 해소 수준 또한 낮아지며,

〈표 6〉 데이터 리샘플링 기법 적용에 따른 모형 성능 요약 (단위: 데이터 개수, %)

데이터 리샘플링 기법	적용 비율	부도 데이터 (훈련 데이터)		성능 평가 지표(시험 데이터)					
				민감도			정밀도		
		N	Y	LR	RF	XGB	LR	RF	XGB
None	None	157,985	20,757	0.71	5.00	7.77	70.79	57.64	52.79
ROS	100	157,985	157,985	65.21	9.48	60.51	20.36	50.15	23.08
	50		78,992	34.67	9.38	37.69	29.55	50.39	31.49
	15		23,697	2.05	5.76	10.03	57.05	60.52	50.91
SMOTE	100		157,985	65.09	23.54	19.45	20.71	24.52	32.00
	50		78,992	36.39	17.29	16.34	28.64	28.92	38.50
	15		23,697	0.37	6.44	8.82	63.46	56.12	52.06
RUS	100	20,757	20,757	61.21	68.69	66.41	20.08	21.22	21.11
	50	41,514		18.91	40.45	42.34	29.82	30.57	29.18
	15	138,380		2.18	6.86	10.49	58.08	56.17	50.35
Near-Miss	100	20,757		59.77	67.85	66.00	16.18	17.80	17.42
	50	41,514		23.21	36.38	38.22	24.18	26.45	24.84
	15	138,380		0.07	5.51	9.11	75.00	56.78	50.98
SMOTE-ENN	100	76,731	103,353	71.43	37.79	32.69	17.59	24.52	28.78
	50	89,126	32,282	26.10	19.06	17.42	30.89	33.51	37.76
	15	109,785	574	0.00	1.38	2.47		70.69	58.82
SMOTE-Tomek	100	145,978	145,978	64.73	24.99	21.10	20.10	25.20	31.58
	50	146,632	67,639	26.14	17.15	15.25	29.36	31.32	38.53
	15	147,547	13,259	0.03	3.88	5.82	75.00	63.89	58.60

모형의 정밀도는 상승하는 반면 민감도는 하락하는 모습을 보인다. 이는 데이터 불균형 개선 수준이 낮아질수록 모형을 통해 부도로 예측한 차주가 실제 부도로 이어질 가능성이 높아지는 반면, 실제 부도 차주를 정상 차주로 잘못 예측하는 비율 또한 높아진다는 것을 의미한다.

<표 6>을 통해 구체적인 데이터 리샘플링 기법과 머신러닝 알고리즘에 따라 각 모형의 성능 개선 효과는 차이가 있으나, 데이터 리샘플링 기법의 적용 비율과 모형의 성능 개선 효과는 세 가지 적용 모형(LR: Logistic Regression, RF: Random Forest, XGB: XGBoost)에서 모두 동일한 경향을 보인다는 것을 발견할 수 있다. 한편, Logistic Regression 알고리즘에 Hybrid sampling 기법 중 SMOTE-ENN 기법을 15% 비율로 적용할 경우 모형의 민감도가 0으로 낮아지면서 모형의 정밀도가 산출되지 않는 현상이 나타난다. 이는 혼동 행렬의 TP 및 FP 값이 0이 되었기 때문이다.

4.2 부도 임계치 조정

<표 7>은 불균형이 존재하는 훈련 데이터를 학습한 모형의 부도 임계치를 조정하였을 경우의 성능을 요약한 것으로, 데이터 리샘플링 기법 적용을 통해 데이터 불균형 문제를 개선하는 경우와 유사한 성능 변화 양상을 보인다.

모형의 부도 임계치를 낮은 수준으로 조정할 경우, 데이터 리샘플링 기법 적용 비율이 높아지

〈표 7〉 부도 임계치 조정에 따른 모형 성능 요약 (단위: 데이터 개수, %)

부도 임계치	부도 데이터 (훈련 데이터)		성능 평가 지표(시험 데이터)					
			민감도			정밀도		
	N	Y	LR	RF	XGB	LR	RF	XGB
5	157,985	20,757	94.24	93.39	88.77	13.69	13.67	15.61
20			39.16	49.13	42.00	27.91	26.78	29.75
35			8.59	18.19	19.28	45.75	41.84	42.60
50			0.71	5.00	7.77	70.79	57.64	52.79
65			0.00	0.92	2.83		68.33	60.58
80			0.00	0.08	0.80		100.00	87.65
95			0.00	0.00	0.00			

는 경우와 마찬가지로 모형의 민감도는 상승하는 반면 정밀도는 하락한다. 예를 들어, Logistic Regression 알고리즘을 활용한 모형의 부도 임계치를 조정하지 않은 경우(부도 임계치 50%) 민감도는 0.71%, 정밀도는 70.79% 수준이나 부도 임계치를 5%로 조정할 경우 민감도는 94.24%로 상승하고 정밀도는 13.69%로 하락한다.

그리고 모형의 부도 임계치가 높아질수록 데이터 리샘플링 기법 적용 비율이 낮아지는 경우와 같이 모형의 민감도는 하락하는 반면 정밀도는 상승한다.

부도 임계치 또한 일정 수준 이상으로 높아질 경우 모형의 민감도가 0으로 낮아지고 정밀도가 산출되지 않는 현상이 나타나며, 그 사유는 데이터 리샘플링 기법의 경우와 동일하다.

Ⅴ. 결 론

5.1 연구 결과 요약

<표 6> 및 <표 7>을 통해 데이터 리샘플링 기법과 모형의 부도 임계치 조정 모두 부도 모형의 예측(일반화) 성능 개선 효과가 있음을 확인하였다. 또한 모형에 활용한 데이터 불균형 문제 개선 기법 및 머신러닝 알고리즘에 따라 모형의 성능 개선 효과의 차이는 존재하나, 변화 양상은 동일한 것을 확인하였으며 연구 결과를 요약하면 다음과 같다.

먼저, 모형의 민감도를 중심으로 성능을 개선하고자 할 경우 데이터 리샘플링 기법을 높은 비율로 적용하여 데이터의 불균형 해소 수준을 높이거나, 모형의 부도 임계치를 낮은 수준으로 조정해야 한다. 하지만 이 경우 모형의 정밀도는 하락한다.

만약 모형의 정밀도 개선이 주 목적인 경우, 데이터 리샘플링 기법을 낮은 비율로 적용하여 비교적 낮은 수준으로 데이터의 불균형을 해소할 필요가 있으며, 모형의 부도 임계치를 높은 수준으로 조정하는 것이 바람직하다. 하지만 이 경우, 상충관계로 인해 민감도가 하락하게 된다.

5.2 연구 결과 토의

부도 예측 모형의 민감도가 높을 경우, 실제 부도 차주를 정상 차주로 잘못 예측할 가능성이 낮다는 장점이 있다. 하지만 민감도가 높아질수록 정밀도는 낮아질 수밖에 없으며, 이는 모형을 통해 부도로 예측한 차주가 실제 부도 차주일 가능성이 낮아진다는 문제점으로 이어진다.

실무적인 관점에서 볼 때, 민감도가 높은 부도 예측 모형은 실제 부도 차주를 예측하는 능력은 뛰어나지만 수많은 정상 차주를 부도로 예측한다

〈표 8〉 F_1 Score를 중심으로 한 데이터 불균형 문제 개선 방안별 모형 성능 요약 (단위: %)

알고리즘	성능 평가 지표	데이터 리샘플링 기법						부도 임계치 조정
		ROS	SMOTE	RUS	Near-Miss	SMOTE-ENN	SMOTE-Tomek	
Logistic Regression	민감도	45.99	52.51	43.41	52.30	43.48	58.24	39.16
	정밀도	25.84	24.33	27.27	20.87	26.57	22.47	27.91
	F_1 Score	33.09	33.25	33.50	29.84	32.99	32.43	32.59
Random Forest	민감도	9.73	23.31	51.82	40.89	37.79	24.75	49.13
	정밀도	49.68	25.60	26.66	25.19	24.52	25.69	26.78
	F_1 Score	16.28	24.40	35.21	31.18	29.74	25.21	34.67
XGBoost	민감도	44.01	19.83	49.96	42.67	32.69	20.85	42.00
	정밀도	29.22	32.49	27.01	23.83	28.78	32.64	29.75
	F_1 Score	35.12	24.63	35.06	30.58	30.61	25.45	34.83

는 점에서 지나치게 보수적인 모형이라고 할 수 있다. 이 경우 부도로 예측한 차주를 관리하기 위한 인적 · 물적 관리 비용이 과다하게 발생할 수 있으며, 정상 차주에 대한 제한적인 여신 공급 또는 여신 회수로 인한 수익성 하락이 발생할 수 있다.

반면, 정밀도를 중심으로 부도 예측 모형을 구축할 경우 모형을 통해 부도로 예측한 차주가 실제 부도로 이어질 가능성이 높다는 장점이 있다. 그러나 이 경우, 낮은 민감도로 인해 실제 부도 차주를 정상 차주로 잘못 예측하는 비율 또한 높다는 문제가 있다. 이러한 부도 예측 모형은 효율적인 부도 예상 차주 관리가 가능하지만 수많은 실제 부도 차주를 정상 차주로 예측한다는 한계점이 있다. 즉 부도 예측 모형의 운용 비용은 상대적으로 낮을 수 있지만, 실제 부도 차주를 정상 차주로 예측하면서 발생하는 부실 여신으로 인한 여신 건전성 악화 문제가 발생할 수 있다.

민감도와 정밀도 중 특정 지표만을 극대화할 경우, 앞서 언급한 바와 같은 큰 비용이 발생할 수 있다. 따라서 부도 예측 모형의 실무적인 활용도 제고를 위해서는 민감도와 정밀도의 균형을 반드시 고려해야 한다. 그러므로 부도 예측 모형의 성능 평가 지표는 민감도와 정밀도의 조화평균 값인 F_1 Score를 중심으로 한 F_β Score를 활용할 필요가 있다. <표 8>은 불균형 문제 개선 방안별로 F_1 Score가 가장 높은 경우의 민감도와 정밀도 값을 정리한 것이다. 이를 통해 F_1 Score 값을 극대화할 경우에도 민감도와 정밀도가 극단적으로 높거나 낮아지지 않는다는 것을 확인할 수 있다.

일반적인 상황에는 두 지표의 가중치가 동일한 F_1 Score를 극대화한 부도 예측 모형을 운용하되, 경기 악화 등으로 인해 보수적인 건전성 관리가 필요한 경우 민감도의 가중치가 높은 F_2 Score를 활용하는 것이 합리적일 것이다. 반면, 경기 확장 국면에 접어들거나 수익성 제고가 중요한 상황이라면 정밀도의 가중치가 높은 $F_{0.5}$ Score를 중심으로 모형의 성능을 개선할 필요가 있을 것이다.

5.3 학술적 시사점

본 연구는 부도 데이터의 불균형 문제 개선 방안을 체계적으로 정리하고, 여러 방안 중 데이터 리샘플링 기법의 적용과 부도 임계치의 조정에 따른 부도 예측 모형의 성능 개선 효과를 비교 분석하였다. 부도 예측 모형을 다루는 대부분의 선행연구는 모형의 성능 개선에 중심을 두고 있는 반면, 본 연구는 부도 데이터의 불균형 문제 개선 방안과 부도 예측 모형의 성능 개선 효과의 관계를 주로 다루고 있다는 점에서 기존의 선행 연구와 차이점이 있다.

5.4 실무적 시사점

본 연구에서는 부도 예측 모형의 활용 목적에 따라 부도 데이터의 불균형 문제 개선 방안을 달리 적용할 것을 제안하였다. 또한 민감도와 정밀도의 상충 관계로 인해 발생할 수 있는 비용 문제를 지적하였으며, 부도 예측 모형의 활용도 제고를 위해 F_1 Score를 중심으로 한 F_β Score를 성능 평가 지표로 활용할 것을 제안하였다는 점에서 실무적인 시사점을 찾을 수 있다.

5.5 연구 한계점 및 후속 연구 방향

본 연구는 부도 데이터의 불균형 문제를 개선할 수 있는 다양한 방안 중, 데이터 리샘플링 기법과 부도 임계치 조정에 한하여 분석하였다는 한계가 있다. 따라서 후속 연구에서는 Feature Selection 또는 Feature Extraction, Cost-sensitive Classifier 구축 등 본 연구에서 다루지 못한 다양한 불균형 문제 개선 방안을 활용하여 부도 예측 모형의 성능 개선 효과를 검토할 필요가 있다.

또한, 연구에 쓰인 실증 분석 데이터와 머신러닝 알고리즘이 제한적이라는 한계점이 있다. 본 연구는 캐글에서 제공하는 한 가지 종류의 가상 데이터만을 활용하고 있으며, 머신러닝 알고리즘 또한 세 가지 종류만 적용하였다. 따라서 향후 금융기관의 실제 부도 데이터를 통해 강건성을 검증할 필요가 있으며, 다양한 머신러닝 알고리즘을 통해 성능 개선 효과의 일반화 가능성을 보다 엄격하게 검증할 필요가 있다.

참 고 문 헌

[1] 금융감독원, “금감원, 금융상황 점검회의 개최”, 2023.08.24, Available at https://www.fss.or.kr/fss/bbs/B0000188/view.do?nttId=129829&menuNo=200218&cl1Cd=&sdate=&edate=&searchCnd=1&searchWrd=%EC%A0%90%EA%B2%80&pageIndex=3.

[2] 김명종, 윤우섭, “기업부도 예측 앙상블 모형의 최적화”, *경영정보학연구*, 제24권, 제1호, 2022, pp. 39-57.

[3] 노정담, 최병구, “불균형 정형 데이터를 위한 SMOTE와 변형 CycleGAN 기반 하이브리드 오버샘플링 기법”, *경영정보학연구*, 제24권, 제4호, 2022, pp. 97-118.

[4] 조성임, 김명종, “비대칭 마진 SVM 최적화 모델을 이용한 기업부실 예측모형의 범주 불균형 문제 해결”, *경영정보학연구*, 제24권, 제4호, 2022, pp. 23-40.

[5] 조용복, 조동우, 최보승, “불균형 시계열 자료를 위한 분류 알고리즘 적용방안: 기업 부도모형을 중심으로”, *Journal of The Korean Data Analysis Society(JKDAS)*, 제24권, 제2호, 2022, pp. 639-651.

[6] 한국은행, “통화정책방향 관련 총재 기자간담회(2023.11)”, 2023.11.30, Available at https://www.bok.or.kr/portal/bbs/B0000169/view.do?nttId=10080889&menuNo=200059&pageIndex=1.

[7] 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.

[8] Barboza, F., H. Kimura, and E. Altman, “Machine learning models and bankruptcy prediction”, *Expert Systems with Applications*, Vol.83, 2017, pp. 405-417.

[9] Buckland, M. and F. Gey, “The relationship between recall and precision”, *Journal of The American Society for Information Science*, Vol.45, No.1, 1994, pp. 12-19.

[10] Cateni, S., V. Colla, and M. Vannucci, “A method for resampling imbalanced datasets in binary classification tasks for real-world problems”, *Neurocomputing*, Vol.135, 2014, pp. 32-41.

[11] 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, 2002, pp. 321-357.

[12] Datta, S. and S. Das, “Near-Bayesian support vector machines for imbalanced data classification with equal or unequal misclassification costs”, *Neural Networks*, Vol.70, 2015, pp. 39-52.

[13] Dembczynski, K., A. Jachnik, W. Kotlowski, W. Waegeman, and E. Hüllermeier, “Optimizing the F-measure in multi-label classification: Plug-in rule approach versus structured loss minimization”, *In International Conference on Machine Learning*, 2013, pp. 1130-1138.

[14] Dubey, R., J. Zhou, Y. Wang, P. M. Thompson, J. Ye, and Alzheimer’s Disease Neuroimaging Initiative, “Analysis of sampling techniques for imbalanced data: An n = 648 ADNI study”, *NeuroImage*, Vol.87, 2014, pp. 220-241.

[15] Esposito, C., G. A. Landrum, N. Schneider, N. Stiefl, and S. Riniker, “GHOST: Adjusting the decision threshold to handle imbalanced data in machine learning”, *Journal of Chemical Information and Modeling*, Vol.61, No.6, 2021, pp. 2623-2640.

[16] Guyon, I. and A. Elisseeff, “An introduction to variable and feature selection”, *Journal of Machine Learning Research*, Vol.3, 2003, pp. 1157-1182.

[17] Haixiang, G., L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, “Learning from class-imbalanced data: Review of methods and applications”, *Expert Systems with Applications*, Vol.73, 2017, pp. 220-239.

[18] Han, H., W. Y. Wang, and B. H. Mao, “Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning”, In *International Conference on Intelligent Computing*, 2005, pp. 878-887.

[19] He, H., Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning”, *IEEE International Joint Conference on Neural Networks*, 2008, pp. 1322-1328.

[20] He, H. and E. A. Garcia, “Learning from imbalanced data”, *IEEE Transactions on Knowledge and Data Engineering*, Vol.21, No.9, 2009, pp. 1263-1284.

[21] Kaggle, “Loan Default Prediction Dataset”, NIK HIL, 2023, Available at https://www.kaggle.com/datasets/nikhil1e9/loan-default.

[22] Kim, M. J., D. K. Kang, and H. B. Kim, “Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction”, *Expert Systems with Applications*, Vol.42, No.3, 2015, pp. 1074-1082.

[23] López, V., A. Fernández, S. García, V. Palade, and F. Herrera, “An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics”, *Information Sciences*, Vol.250, 2013, pp. 113-141.

[24] Mani, I. and I. Zhang, “kNN approach to unbalanced data distributions: A case study involving information extraction”, In *Proceedings of Workshop on Learning From Imbalanced Datasets*, Vol.126, No.1, 2003, pp. 1-7.

[25] Mellor, A., S. Boukir, A. Haywood, and S. Jones, “Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin”, *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol.105, 2015, pp. 155-168.

[26] Messier, W. F. Jr. and J. V. Hansen, “Inducing rules for expert system development: An example using default and bankruptcy data”, *Management Science*, Vol.34, No.4, 1998, pp. 1403-1415.

[27] Musicant, D. R., V. Kumar, and A. Ozgur, “Optimizing F-Measure with Support Vector Machines”, *FLAIRS*, 2003, pp. 356-360.

[28] Nan, Y., K. M. Chai, W. S. Lee, and H. L. Chieu, “Optimizing F-measure: A tale of two approaches”, arXiv preprint arXiv:1206.4625, 2012.

[29] Ohlson, J. A., “Financial ratios and the probabilistic prediction of bankruptcy”, *Journal of Accounting Research*, 1980, pp. 109-131.

[30] Sheng, V. S. and C. X. Ling, “Thresholding for making classifiers cost-sensitive”, *Aaai*, Vol.6, 2006, pp. 476-481.

[31] Shin, K. S., T. S. Lee, and H. J. Kim, “An application of support vector machines in bankruptcy prediction”, *Expert Systems with Applications*, Vol.28, No.1, 2005, pp. 127-135.

[32] Weiss, G. M., “Mining with rarity: A unifying framework”, *ACM Sigkdd Explorations Newsletter*, Vol.6, No.1, 2004, pp. 7-19.

[33] Yijing, L., G. Haixiang, L. Xiao, L. Yanan, and L. Jinling, “Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data”, *Knowledge-Based Systems*, Vol.94, 2016, pp. 88-104.

[34] Zhou, J., W. Li, J. Wang, S. Ding, and C. Xia, “Default prediction in P2P lending from high-dimensional data based on machine learning”, *Physica A: Statistical Mechanics and Its Applications*, Vol.534, 2019.

[35] 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.

[36] Zmijewski, M. E., “Methodological issues related to the estimation of financial distress prediction models”, *Journal of Accounting Research*, Vol.22, 1984, pp. 59-82.

Information Systems Review
Volume 26 Number 2
May 2024

A Study on Default Prediction Model: Focusing on The Imbalance Problem of Default Data*

Jinsoo Park** · Kangbae Lee*** · Yongbok Cho****

Abstract

This study summarizes improvement strategies for addressing the imbalance problem in observed default data that must be considered when constructing a default model and compares and analyzes the performance improvement effects using data resampling techniques and default threshold adjustments. Empirical analysis results indicate that as the level of imbalance resolution in the data increases, and as the default threshold of the model decreases, the recall of the model improves. Conversely, it was found that as the level of imbalance resolution in the data decreases, and as the default threshold of the model increases, the precision of the model improves. Additionally, focusing solely on either recall or precision when addressing the imbalance problem results in a phenomenon where the other performance evaluation metrics decrease significantly due to the trade-off relationship. This study differs from most previous research by focusing on the relationship between improvement strategies for the imbalance problem of default data and the enhancement of default model performance. Moreover, it is confirmed that to enhance the practical usability of the default model, different improvement strategies for the imbalance problem should be applied depending on the main purpose of the model, and there is a need to utilize the F_β Score as a performance evaluation metric.

Keywords: Default Model, Imbalanced Data, Resampling, Decision Threshold, F_β Score

* This work was supported by the Dong-A University research fund. This work is a rephrased excerpt from Jinsoo Park's 2022 master's thesis titled "A Study on Solutions to Address the Imbalance Problem of Default Data".

** Graduate Student, Department of Management Information Systems, Dong-A University
*** Professor, Department of Management Information Systems, Dong-A University
**** Corresponding Author, Assistant Professor, Department of Management Information Systems, Dong-A University

◐ 저 자 소 개 ◑

박 진 수 (2373458@donga.ac.kr)

부산대학교 경영학과를 졸업하였으며, 동아대학교 디지털금융학과에서 석사학위를 취득하였다. 현재 동 대학원 경영정보학과 석사 과정 재학 중이며, 부산은행 리스크관리부에서 근무하고 있다. 주요 관심분야는 금융 데이터사이언스, 디지털 금융, 인공지능, 데이터 분석 등이다.

이 강 배 (kanglee@dau.ac.kr)

고려대학교에서 산업공학을 수학하였으며, 한국과학기술원에서 산업공학 석사 및 박사학위를 취득하였다. LG CNS 컨설팅 부문에서 컨설턴트로 근무하였고 부산가톨릭대학교 e-business학과에서 근무하였으며, 현재 동아대학교 경영정보학과 교수로 재직 중이다. 관심연구 분야는 제조 및 서비스 분야의 인공지능 활용을 위한 거대언어모형 응용과 이미지 및 시계열 데이터 등 다중모드 데이터 분석 모델 개발이다.

조 용 복 (ybcho@dau.ac.kr)

인하대학교 및 고려대학교에서 경영학과 통계학을 수학하였으며, 고려대학교 금융공학협동과정에서 석사 및 박사학위를 취득하였다. 금융기관 리스크관리 및 자산배분 모형 컨설팅으로 산업계 경력을 시작하고, 이후 증권사 채권영업 및 FICC(채권 및 외환)운용 업무를 다년간 수행하였다. 현재는 동아대학교 경영정보학과에서 금융 및 인공지능 강의를 담당하고 있으며 다양한 데이터 사이언스 방법론을 활용한 금융관리기법의 혁신에 관한 연구를 진행 중이다.

논문접수일 : 2024년 03월 06일 **게재확정일 : 2024년 05월 03일**
1차 수정일 : 2024년 04월 19일 **2차 수정일 : 2024년 04월 29일**