The Journal of the Convergence on Culture Technology (JCCT)
Vol. 10, No. 4, pp.751-756, July 31, 2024. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2024.10.4.751

JCCT 2024-7-88

신용평가에서 설명가능 인공지능의 활용에 관한 연구

Study on use of Explainable Artificial Intelligence in Credit Rating

윤영인*, 김성욱**, 정혜영***

Young-In Yoon*, Seong W. Kim, Hye-Young Jung*****

요 약 모델의 정확도와 결과에 대한 설명가능성은 동시에 고려되어야 할 중요한 요소이다. 최근에는 설명가능한 인공지능을 적용하는 응용 사례가 증가하였고 결과에 대한 해석이 특히 중요시되는 금융에서도 많이 적용되고 있다. 본 논문에서는 오픈 API의 신용평가 자료를 다양한 머신러닝 기법의 성능을 비교하고 모델로부터 설명가능한 인공지능 기법인 SHAP과 LIME을 통해 정확도와 결과에 대한 설명력을 보이고자 한다. 이에 따라 금융 시장에서 머신러닝의 적용가능성을 보일 것으로 기대된다.

주요어 : 설명가능성, 인공지능, 신용평가, SHAP, LIME

Abstract The accuracy of the model and the explanation of the results are important factors that should be considered simultaneously Recently, applications of explainable artificial intelligence are increasing, and it is especially widely applied in the financial field where interpretation of results is important. In this paper, we compare the performance of open API credit evaluation data using various machine learning techniques. In addition, existing financial logic is verified through explainable artificial intelligence technologies, SHAP and LIME. Accordingly, it is expected to demonstrate the applicability of machine learning in the financial market.

Key words : Explainable, AI, Credit evaluation, SHAP, LIME

Ⅰ. 서 론

인공지능의 발전으로 기계학습 기술은 기술적으로 큰 편의와 높은 성능을 보이게 되었고 신용평가와 같은 사회경제학적으로도 큰 영향을 미칠 수 있는 분야에도 활용하고자 시도되고 있다[1]. 그러나 높은 성능을 위해 모델이 복잡해졌고 모델에 대한 직관적인 이해가 어려워져 응용분야에 적용하기 위해서는 인공지능 기술이 올바른 판단을 내렸는지 신뢰할만한 모델인지에 대한 논의가 필요함이 대두되고 있다. 따라서 기계학습 모델이 판단하는 결과의 설명력을 위하여 설명가능 인공지능(XAI, eXplainable Artificial Intelligence)이 제안되었다. 이로써 복잡한 모델에 대해서도 결과를 신뢰할 수 있다.

금융에서도 바젤 Ⅱ 협약 이후 모델의 고도화에 대한 필요성이 증가하며 기존의 연구들은 기계학습 모델을 적용하려 시도되고 있음에 따라 인공신경망 모형 등의 인공지능 방법들이 활용되면서 우수한 성능을 인정받은 사례가 있다[5]. 그러나 설명력이 부족하여 많은 활용은 이루어지지 않고 있으며 실제 산업에서는 성능

*준회원, ㈜ 씨큐센 (제1저자)
**정회원, 한양대학교 수리데이터사이언스학과 (제2저자)
***정회원, 한양대학교 수리데이터사이언스학과 (교신저자)
접수일: 2024년 5월 15일, 수정완료일: 2024년 5월 30일
게재확정일: 2024년 6월 21일

Received: May 15, 2024 / Revised: May 30, 2024
Accepted: June 21, 2024
***Corresponding Author: hyjunglove@hantang.ac.kr
Dept. of Mathmatical Data Science, Hanyang Univ, ERICA

이 낮더라도 설명력이 있는 전통적인 통계 기법인 MDA, 회귀분석, 로지트(logit) 등의 기법들을 활용하고 있다. 본 논문은 기계학습 모델을 사용하여 성능이 좋으며 결과에 대한 설명 부분을 해석할 수 있도록 XAI를 활용하여 실제 산업에 적용하는데 기여하고자 한다. 이는 결과에 대한 해석 또한 고려되어 고객이 결과를 이해할 수 있도록 한다. 또한 전반적으로 출시될 인공지능 모델을 통한 금융 시장 활성화에 도움이 될 것이라 기대된다.

본 논문은 오픈된 다양한 금융 데이터에 선형 모델, 트리 모델, 비선형 모델, 은닉층 2개의 NN(Neural Network) 모델을 사용하여 전통적인 통계 모델과 성능 비교를 하고자 한다. 또한 모델 종류마다 성능이 좋은 모델이 갖는 결과의 해석이 통계 모델과 일치하는지 검증하고자 한다. 모델의 설명력은 XAI 기법 중 SHAP과 LIME을 통해 확인하고 전통적인 통계 모델은 회귀분석을 통해 진행한다.

2장에서는 설명가능 인공지능의 SHAP과 LIME에 대해 설명하고 3장에서는 연구에 대한 설계를 기술한다. 4장과 5장에서는 실제 금융 데이터를 적용한 실험을 통해 전통적인 통계 모델과 성능을 비교하고 모델의 해석력을 보인다. 6장에서는 결론 및 향후 연구 방향에 대해 제시한다.

II. 관련 연구

1. SHAP (SHapley Additive exPlanations)

SHAP은 게임이론인 Shapley value를 기반으로 한 XAI 기법이다. SHAP은 각 변수의 모든 조합을 고려하여 예측값 $\hat{f}(x)$에 대한 기여도를 계산한다. 기여도는 그 영향에 따라 양의 기여도, 음의 기여도로 나눌 수 있다. 모든 모델에서 적용 가능하며 모델의 종류에 따라 Linear SHAP, Tree SHAP, Kernel SHAP 등으로 변수의 중요도를 해석할 수 있다.

$$\begin{aligned} &E[f_{S\cup\{i\}}(x_{S\cup\{i\}})]-E[f_S(x_S)] \\ &=E[f(x_1,x_2,\cdots,x_i,\cdots,x_{n-1},x_n)] \\ &-E[f(x_1,x_2,\cdots,X_i,\cdots,x_{n-1},x_n)] \end{aligned} \quad (1)$$

식 (1)에서 S는 변수 전체 집합을, x는 각 변수값을, X는 해당 변수의 전체 값을 의미한다. 각 변수 대한 기여도의 합은 예측값 $\hat{f}(x)$로, 식 (2)와 같다.

$$\hat{f}(x)=\phi_0+\sum_{i=1}^{n}\phi_i \quad (2)$$

식 (2)에서 예측값 $\hat{f}(x)$는 각 변수의 기여도 ϕ_i와 해당 기여도들의 기준을 나타내는 base value ϕ_0의 합으로 표현된다[1].

2. LIME(Local Interpretable Model-agnostic Explanations)

LIME은 XAI 기법 중 단순화한 설명 모델을 통해 개별 모델의 예측을 해석하는 기법이다. 모든 모델에 적용 가능하며 개별적인 해석이 가능하다는 특징이 있다. LIME은 새로운 데이터를 생성할 때 기준인 데이터와 거리가 짧은 데이터를 가중치를 주어 회귀분석 모델을 학습하며 SHAP과 마찬가지로 그 영향에 따라 양의 기여도, 음의 기여도로 나눌 수 있다. 식 (3)은 LIME의 목적함수이다[1].

$$\xi=\arg\min(g\in G)L(f,g,\pi_x)+\Omega(g) \quad (3)$$

식 (3)에서 f는 해석하고자 하는 복잡한 모델이며 g는 설명 가능한 단순화한 설명 모델이다. π_x는 임의로 생성한 데이터와 기준인 데이터의 거리를 계산한 유사도이다. $L(f,g,\pi_x)$는 학습하는 모델을 적합하는 항이고, $\Omega(g)$는 정규화항이다.

III. 연구 설계

신용평가에 기계학습을 적용하기 위해서는 실제 금융 산업에서 적용되는 기준과 유사한지 확인해야한다. 신용등급은 신용평가 회사에서 기업의 재무 및 비재무 정보와 미래 예측 가능한 전망을 고려하여 채무불이행 발생 가능성을 평가하는 평가지표이다. 특히 공신력 있는 신용평가 회사가 제시하는 신용등급은 객관적이고 신뢰성이 있는 정보로 활용되며 국내에서는 한국신용평가, 한국신용정보평가, 그리고 한국기업평가가 존재한다. 본 논문은 공신력 있는 신용평가 회사 3개의 지표를 토대로 비교하고자 한다.

데이터의 변수 중 신용등급에 연관되어있는 변수는 상환 이력, 부채 수준, 신용거래 기간, 신용 형태, 및 상

환 여부이다.

본 논문은 여러 모델을 통해 유사한 성능을 보이며 일관된 해석을 보이는지 비교 분석하고자 한다. 이에 따라 상환 이력, 부채 수준은 신용평가 점수와 반대의 방향성을 가지고 신용거래 기간, 상환 이력, 및 상환여부는 신용평가 점수와 동일한 방향성을 가지는지 실험을 통해 확인하고자 한다.

Ⅳ. 실험 및 결과

기계학습 모델을 금융 데이터에 적용하기 위해서는 불량 신용등급에 대한 높은 정확도와 예측 결과에 대한 설명력이 함께 고려되어야 한다.

본 논문은 여러 실제 데이터를 사용하여 선형 모델, 트리 모델, 비선형 모델, NN 모델을 선정하여 전통적인 통계 모델과 성능을 비교하였다. 또한 모델의 종류마다 XAI의 LIME과 SHAP을 통해 모델의 설명력을 확인하여 전통적인 통계 모델의 결과인 회귀분석과 비교하고자 한다.

1. 데이터

UCI의 German 데이터, Australian 데이터, 그리고 Kaggle의 Taiwan 데이터를 사용하였고 자세한 설명은 표 1과 같다. 개인 채무불이행 가능성을 추정하기 위해 불량 신용등급을 1로 설정하여 분석하였다. UCI의 German 데이터와 Kaggle의 Taiwan 데이터의 경우 데이터 불균형이 존재했고 다수 샘플에 대한 편향된 결과가 나올 수 있기에 SMOTE를 이용하여 오버샘플링 후 분석을 진행했다.

표 1. 데이터 설명
Table 1. Datasets Summary

Data	총 개수	변수 개수	y=1 (불량)	y=0 (양호)
German	1000	21	300	700
Australian	690	15	307	383
Taiwan	30000	25	23364	6636

2. 성능 지표 및 성능 비교

예측 결과를 신뢰하기 위해 모델은 모두 최적의 성능일 때를 비교하였다. 정확도만 가지고 성능이 좋다고 판단하기 어렵기에 정확도 이외의 정밀도, F점수, G평균, 그리고 ROC curve의 AUC를 통하여 성능을 비교하였고 데이터에 따라 다음 표 2, 3, 그리고 표 4와 같다.

표 2. German 데이터의 모델 성능 비교
Table 2. Comparison of models on German dataset

Model	Accuracy	Recall	F-measure	G-means	AUC
Logistic Regression	65.5	83.3	61.5	68.7	70.0
Linear SVM	71.5	78.8	64.6	73.1	73.3
XGBoost	74.0	57.6	59.4	68.7	69.8
Random Forest	70.5	65.2	59.3	69.0	69.1
KNN	75.5	63.6	63.2	71.9	72.5
RBF SVM	78.5	51.5	61.3	69.8	71.7
NN	68.0	78.1	61.0	70.3	70.7

표 3. Australian 데이터의 모델 성능 비교
Table 2. Comparison of models on Australian dataset

Model	Accuracy	Recall	F-measure	G-means	AUC
Logistic Regression	84.8	78.7	84.9	65.0	85.3
Linear SVM	87.0	92.6	84.7	86.0	86.4
XGBoost	86.2	81.7	85.9	86.4	86.5
Random Forest	87.7	85.1	87.0	87.8	87.8
KNN	86.2	81.7	85.9	86.4	86.5
RBF SVM	87.0	92.6	84.7	86.0	86.4
NN	85.5	79.7	85.5	85.7	86.0

표 4. Taiwan 데이터의 모델 성능 비교
Table 2. Comparison of models on Taiwan dataset

Model	Accuracy	Recall	F-measure	G-means	AUC
Logistic Regression	58.7	69.8	43.3	62.2	62.6
Linear SVM	59.6	68.7	43.5	62.6	62.8
XGBoost	66.5	66.7	47.4	66.6	66.6
Random Forest	73.3	54.7	48.2	65.7	66.7
KNN	70.7	54.6	45.7	64.1	65.0
RBF SVM	64.0	69.5	46.7	65.9	66.0
NN	58.2	74.8	44.7	63.1	64.0

Ⅴ. 설명가능성 해석

설명가능성을 보이기 위해 성능이 보장된 데이터세 대한 실험 결과를 분석하고자 한다. 가장 성능이 좋은 Australian 데이터는 기밀성 보호하기 위해 무의미한 기호로 변경되어 이전 변수에 대해 알 수 없으므로 두 번째로 성능이 좋은 German 데이터에 대한 실험 결과

를 분석하고자 한다. 전통적인 통계모델에서는 전역적인 결과를 확인하나 각 기계학습 모델은 동일한 사람에 대한 데이터로 지역적인 해석과 전역적인 해석 모두 확인하였다.

1. 전통적인 통계모델의 해석

German 데이터의 변수 중 신용평가회사의 지표와 관련된 변수는 총 10가지로 표 5의 변수이다. 회귀분석을 통해 유의미한 변수를 확인하였고 결정계수는 0.536이며 DW검정은 1.991로 모델이 데이터를 잘 설명하고 있음을 알 수 있다.

표 5. German 데이터의 전통적인 통계 모델의 해석
Table 5. Interpretation of traditional statistical model of German dataset

변수	coef	std err	t	$P>\|t\|$
duration	0.0665	0.014	4.592	0.000
other debtors A103	-0.0455	0.011	-4.069	0.000
credit history A32	-0.0455	0.017	-2.637	0.008
credit history A34	-0.0802	0.017	-4.599	0.000
existing checking A12	-0.0300	0.013	-2.325	0.020
existing checking A13	-0.0436	0.011	-3.800	0.000
existing checking A14	-0.1374	0.014	-9.671	0.000
saving account A64	-0.0254	0.011	-2.329	0.020
saving account A65	-0.0441	0.012	-3.798	0.000
installment plans A143	-0.0305	0.012	-2.559	0.011

2. 선형 모델의 설명 가능성

그림 1은 선형 모델 중 성능이 높은 Linear SVM 모델의 LIME에 관한 결과이다. 이는 불량 신용등급을 받은 고객에 대한 데이터로 'existing checking A14' 변수와 'credit history A34' 변수는 불량 신호등급으로 예측하는데 중요한 변수임을 알 수 있다. 이는 통계 모델의 결과와 동일한 예측이며 통계 모델에서 중요하다고 인식된 'saving account A65'는 양호 신용등급으로 예측함을 보인다.

그림 2는 Linear SVM 모델에 대한 SHAP 결과이다. 값이 커질수록 중요도가 높아지거나 낮아지는 점진적인 결과를 보이며 LIME과 마찬가지로 'credit amount' 변수가 중요한 변수이며 넓게 분포되어있음을 알 수 있다. 이는 LIME에서 값의 변화가 크기에 불량 신용등급에 기여함을 알 수 있다. 높게 나왔던 'existing checking A14'와 'credit history A34' 변수는 특정 값으로 인해 불량 신용등급으로 나타남을 보인다. 이를 통해 선형 모델의 LIME과 SHAP을 통해 모델의 예측 결과 또한 신뢰할 수 있다고 해석된다.

그림 1. Linear SVM 모델의 LIME
Figure 1. LIME of Linear SVM model

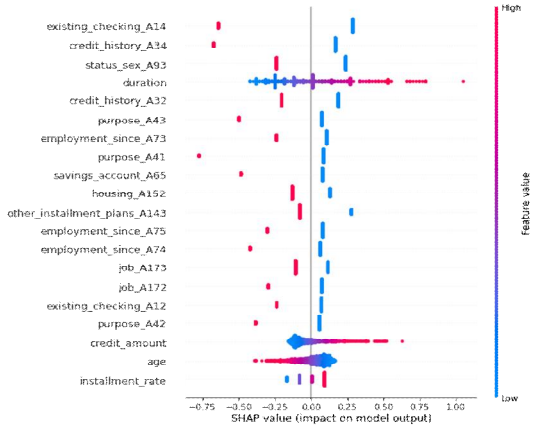

그림 2. Linear SVM 모델의 SHAP
Figure 2. SHAP of Linear SVM model

3. 트리 모델의 설명 가능성

그림 3은 트리 모델을 포함한 모든 모델 중에서 성능이 높은 XGBoost 모델의 LIME에 관한 결과이다. existing checking A14' 변수가 가장 높은 기여도를 가지며 'credit history A34' 변수와 함께 불량 신호 등급으로 예측되는 높은 기여도를 갖는다. 'saving account A65'는 양호 신용등급으로 예측함을 알 수 있다.

그림 4는 SHAP에 관한 결과로 선형 모델과 달리 점진적인 결과가 아닌 부분은 비선형적인 관계가 있음을 나타낸다. 이는 선형 모델에서는 비선형과 같은 고차원 데이터의 특성은 잡히지 않음을 보인다. 또한 절댓값 평균으로 본 결과 'existing checking A14' 변수와

'duration' 변수가 SHAP에서 중요한 변수로 인식됨을 보인다. 이를 통해 트리 모델은 통계 모델과 비교하였을 때 성능도 높고 해석도 일치하는 신뢰할 수 있는 모델임을 확인하였다.

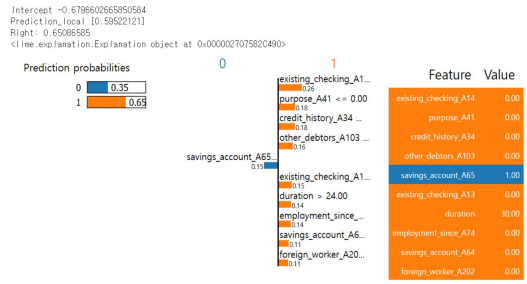

그림 3. XGBoost 모델의 LIME
Figure 3. LIME of XGBoost model

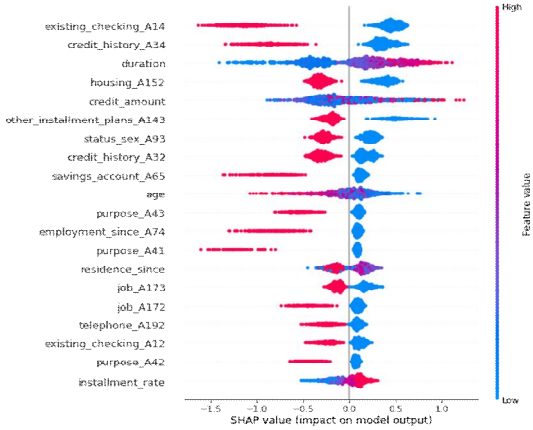

그림 4. XGBoost 모델의 Tree SHAP
Figure 4. Tree SHAP of XGBoost model

4. 비선형 모델의 설명 가능성

그림 5은 KNN 모델의 LIME에 대한 결과로 'duration' 변수가 두 번째로 중요한 변수임이 나타난다. 그림 6은 'duration' 변수에 대한 잔차 그래프로 선형성을 확인할 수 있다. 예측값에 따라 잔차가 크게 변하여 선형성이 없는 것으로 보아 트리 모델과 비선형 모델이 비선형 변수에 대해서도 파악하며 통계 모델과 유사한 양상임을 알 수 있다. 이는 금융 데이터에 기계학습 모델이 적용 가능함을 나타낸다.

그림 7은 KNN 모델의 SHAP에 대한 결과로 불량 신호등급에 대한 변수 중요도를 나타낸다. LIME의 결과와 마찬가지로 'existing checking A14' 변수가 가장 중요한 변수이며 'credit history A34'가 두 번째로 중요도에 기여한 변수임을 나타낸다. 통계 모델의 유의미한 변수 중에서 'existing checking A12' 변수와 'existing checking A13' 변수를 제외한 모든 변수가 중요한 변수로 포함되어있어 유사한 양상을 나타냄을 확인하였다.

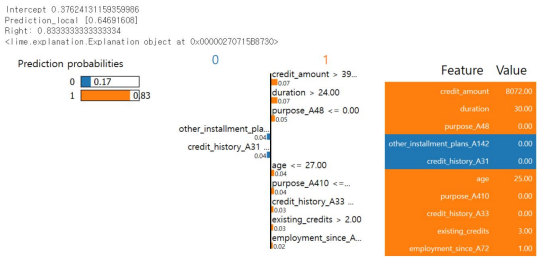

그림 5. KNN 모델의 LIME
Figure 5 LIME of KNN model

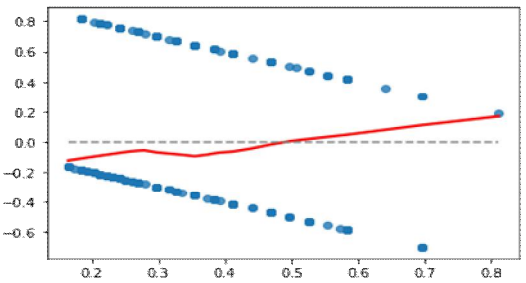

그림 6. 'duration' 변수의 선형성 확인
Figure 6. Check the linearity of the 'duration' variablel

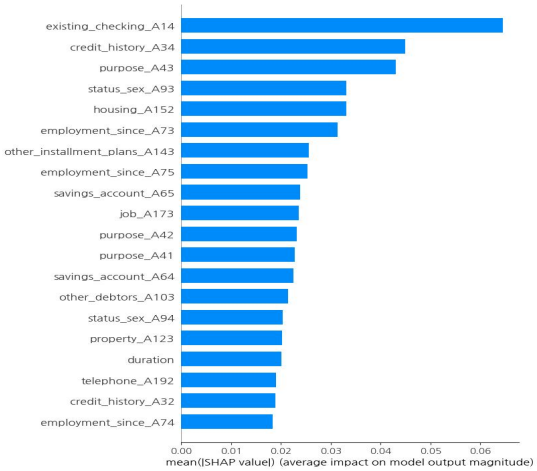

그림 7. KNN 모델의 SHAP
Figure 7. SHAP of KNN model

그림 8과 9는 NN 모델의 LIME과 SHAP에 대한 결과이다. LIME에서 가장 중요한 변수가 달라진 것은 LIME의 무작위성으로 인한 결과임으로 해석된다.

전역적인 해석력을 갖는 SHAP은 다른 기계학습 모델과 유사한 변수 중요도를 나타내며 통계 모델의 유의

미한 변수 중 ‘other debtors A103’, ‘existing checking A12’, ‘existing checking A13’, 그리고 ‘saving account A64’ 변수를 제외한 모든 변수가 중요한 변수로 나타난다.

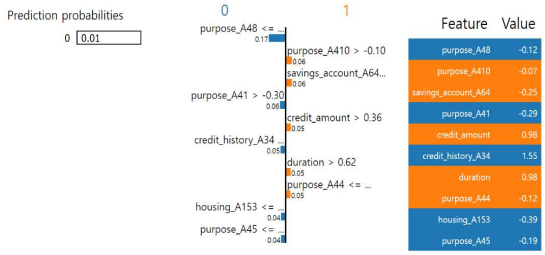

그림 8. NN 모델의 SHAP
Figure 8 SHAP of NN model

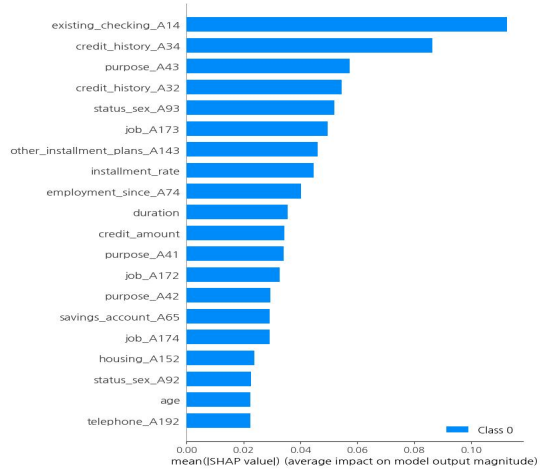

그림 9. NN 모델의 SHAP
Figure 9. SHAP of NN model

Ⅵ. 결론 및 향후 연구

본 연구는 금융 시장의 신용평가에서 기계학습 모델의 적용 가능성을 보이고자 하였다. 또한 전통적인 통계 모델과 비교하여 성능뿐만 아니라 모델이 갖는 설명력이 유사한지 비교 분석하였다.

실험 결과, 신용평가 데이터에 대해 대체적으로 트리 모델과 비선형 모델의 성능이 가장 좋으며 그 다음으로 선형 모델, 통계 모델 순으로 성능이 좋음을 확인하였다. 설명가능 인공지능의 LIME과 SHAP을 통한 기계학습 모델의 해석력은 전통적인 통계 모델과도 유사한 결과 해석을 보인다. 따라서 단순한 데이터에서 복잡한 데이터까지 기계학습 모델은 높은 성능과 해석력을 가짐을 확인하였고 기계학습 모델의 결과를 신뢰할 수 있음을 입증하였다.

또한 비선형 데이터에서도 잘 분류하는 트리 모델과 비선형 모델에서의 LIME과 SHAP의 해석은 선형 모델에서는 발견되지 않았던 비선형 관계가 있는 중요한 변수를 발견하였다. 이는 복잡한 기계학습 모델을 통해 전통적인 통계 모델에서는 발견되지 않았던 비선형 관계의 변수를 추가적으로 확인될 수 있음을 기대한다. 이에 따라 본 연구는 금융 시장에서 인공지능 모델의 도입에 큰 도움이 될 것을 기대한다.

향후 연구로는 NN 모델의 은닉층을 추가하여 더 복잡한 블랙박스 모델에서도 통계 모델의 결과와 유사한 성능과 해석력을 가지는지 확인하고자 한다. 또한, 개인의 민감 정보를 포함하는 Australian 데이터와 같은 동형암호를 도입했을 때에도 전통적인 통계 모델의 결과와 유사한 해석력을 가지는지 검증하여 더욱 금융 시장에서의 기계학습 모델의 적용 가능성을 보이고자 한다.

References

[1] Y. Ahn, S. Ryu, H. Lee, and M. Park, “Prediction of Food Franchise Success and Failure Based on Machine Learning,” The journal of the convergence on culture technology, vol. 8, no. 4, pp. 347 - 353, Jul. 2022.

[2] Lundberg, Scott M., and Su-In Lee. “A unified approach to interpreting model predictions.” Advances in neural information processing systems 30, 2017.

[3] Shen, Feng, et al. “A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation.” Physica A: Statistical Mechanics and its Applications 526, 2019.

[4] Misheva, B. H., Osterrieder, J., Hirsa, A., Kulkarni, O., & Lin, S. F. (2021). Explainable AI in credit risk management. arXiv preprint arXiv:2103.00949.

[5] 천예은, 김세빈, 이자윤, & 우지환. (2021). 설명가능한 AI 기술을 활용한 신용평가 모형에 대한 연구. 10.7465/jkdi.2021.32.2.283

※ 본 논문은 한국연구재단 연자지원사업(NRF-2022R1F1A1074939)의 지원을 받아 수행된 연구임.