Journal of the Korean Data & Information Science Society
2021, **32**(2), 283–295

http://dx.doi.org/10.7465/jkdi.2021.32.2.283
한국데이터정보과학회지

설명 가능한 AI 기술을 활용한 신용평가 모형에 대한 연구

천예은[1] · 김세빈[2] · 이자윤[3] · 우지환[4]

[1234]신한은행 AI 통합센터 · [4]고려대 기술경영전문대학원

접수 2020년 12월 21일, 수정 2021년 1월 21일, 게재확정 2021년 1월 27일

요 약

인공지능 기술이 발전함에 따라, 금융 산업에도 인공지능 기술을 적용하는 사례들이 증가하고 있다. 그러나 인공지능 기술의 경우 많은 부분 비선형성이 높기 때문에 결과를 도출하는 과정에 대한 이해가 직관적이지 않는 것이 문제이다. 이런 특성으로 인해서 인공지능으로 결과를 도출하는 과정을 블랙박스로 표현하기도 한다. 최근 EU에서 새로운 개인정보 보호 규정을 만들면서, 인공지능 알고리즘을 통해 도출된 결과에 대해서 고객이 서비스 제공자에게 설명을 요청할 수 있는 권리를 보장하였다. 즉, 금융 산업에서 인공지능 기술을 적용하기 위해서는, 높은 정밀도뿐만 아니라 설명할 수 있는 능력도 고려해야 한다는 것이다. 본 논문에서는 외부에 오픈된 다양한 신용 정보 데이터를 활용하여, 인공지능 기반의 신용평가 알고리즘을 제안하였다. 이와 함께, 인공지능이 도출한 결과에 대해서, 데이터의 다양한 특성들 중에서 어떤 특성이 결과 도출에 큰 영향을 끼쳤는지 도출하는 알고리즘을 제안하였다. 또한, 이를 확장해서 인공지능이 도출한 결과의 변동이 있었을 때, 변동 결과를 설명하는 방법을 금융 데이터에 적용하였다. 제안된 방법을 통해서, 금융 서비스에서 인공지능 기술을 도입할 때, 설명력을 제공할 수 있음을 확인하였다는 점에서 큰 의미가 있다.

주요용어: 디지털 금융, 설명 가능한 AI, 신용평가.

1. 서론

전통적으로 대출업을 영위하는 금융기관들은 리스크 관리 및 효과적인 규제 대응을 위하여 신용등급 및 대출심사를 위한 다양한 모형을 구축하여 운영하고 있다. 바젤II협약 (국제 결제은행에서 은행들의 자기자본 비율 (BIS비율) 설정의 국제기준을 제시한 것으로 은행의 재무건정성을 확보하기 위한 국제협약) 이후 이러한 모델의 고도화에 대한 필요성이 증가함에 따라, 금융기관들이 보유한 데이터를 기반으로 외부 신용평가사 정보를 결합하여 통계적 기법을 활용한 모델을 사용하고 있다 (Lee 등, 1996). 이러한 모형들은 주로 이분법 분류를 담당하는 로지스틱 선형 회귀 기법을 기반으로 설계되었다. 따라서 모델의 예측력은 다른 머신러닝 기법에 비해 상대적으로 낮지만 결과에 대한 설명이 직관적인 장점이 있다 (Kim, 2012).

4차 산업혁명 시대를 맞이하여 빅데이터와 인공지능 기술이 전 산업으로 확대되는 가운데, 금융 산업도 기존에 사용되던 신용평가 모형 등에 인공지능 기술을 도입하는 시도를 진행하고 있다 (Zhen, 2019). 특히 데이터 분석 기법과 머신러닝 알고리즘이 고도화됨에 따라 기존 통계 기반 모형의 예측력 향상을

[1] (04513) 서울특별시 중구 세종대로 55, 신한은행 AI Competency Center, 대리.
[2] (04513) 서울특별시 중구 세종대로 55, 신한은행 AI Competency Center, 과장.
[3] (04513) 서울특별시 중구 세종대로 55, 신한은행 AI Competency Center, 행원.
[4] 교신저자: (04513) 서울특별시 중구 세종대로 55, 신한은행 AI Competency Center, 부부장.
Email: jihwan_woo@korea.ac.kr

위한 다양한 시도가 진행되었다 (Zhen, 2019). 주요 내용으로는, 기존의 금융 데이터뿐만 아니라 대안 데이터를 활용하여 분석의 범위를 넓히는 방법과 부스팅 기반의 머신러닝 모델이나 딥러닝과 같은 복잡한 모형을 적용하는 방법들이 존재한다 (Zheng 등, 2018). 그러나 인공지능 기반 모델이 복잡해질수록 모델의 내부가 블랙박스와 같이 알 수 없다는 단점이 존재한다 (Qiu와 Choi, 2019; Jang 등, 2020). 따라서 기존에는 큰 문제가 되지 않았던 모델 결과에 대한 설명력을 확보하는 방안이 요구되고 있다. 결과에 대한 설명력 확보를 통해서, 투명하고 결과에 대한 해석 가능한 시스템을 만들 수 있기 때문이다. 특히 AI 윤리의 이슈와 GDPR (general data protection regulation의 약자로 기존 개인정보 보호 지침 (data protection directive)을 모든 회원국에 직접적인 법적 구속력을 갖는 규정 (regulation)으로 강화함) 등 설명력 확보에 대한 외부 규제가 점차 강화되고 있으며, 복잡한 모델의 결과 값에 대한 금융사 임직원의 이해도를 높이고 고객들에게 충분한 설명을 제공해야하는 내부적인 요구도 증가하고 있다 (Lee, 2020). 이러한 배경을 바탕으로, 설명 가능한 인공지능, XAI (eXplainable AI, 설명 가능한 인공지능의 약자)에 대한 금융권 모형 도입이 필수적이다 (Kim, 2020).

이에 대응하기 위해서, 각종 금융회사들은 대리 모델 (surrogate model)을 기반으로 하여 결과 값에 대한 변수간의 기여도를 측정하여 블랙박스인 머신러닝 모델에 대한 해석력을 확보하는 시도를 진행하고 있다. 이러한 방법들은 개별 결과에 대한 설명력을 제공한다는 장점이 있다. 그러나 금융회사의 고객들은 정기적으로 신용도에 대한 평가를 받기에 개별 결과보다는 과거 결과와의 비교를 통한 변화 정도에 대한 설명을 제공할 수 없다는 한계점이 존재한다. 또한 개별 결과에 대한 해석만으로는 금융회사의 규제모형에서 필요한 투명성과 신뢰성에 미치지 못하기 때문에 금융 산업에서 인공지능 기술의 도입에 큰 걸림돌이 되고 있다.

본 논문에서는 이러한 배경을 바탕으로 인공지능 모델이 도출한 신용평가 결과 변화에 대한 다양한 요인들의 영향도를 계산하여 기존의 대리 모델을 보완하는 방법론을 제안하고자 한다.

2. 문헌연구

2.1. 머신러닝 알고리즘

머신러닝은 AI의 한 분야로 데이터를 바탕으로 내재되어 있는 패턴을 알고리즘을 통해 학습하고자 한다. 이러한 학습을 통해 새로운 데이터에 대해서 예측 및 분류하는 일에 활용할 수 있다. 머신러닝은 학습되는 데이터의 형식과 목표에 따라 지도학습 (supervised learning), 비지도학습 (unsupervised learning), 강화학습 (reinforcement learning)으로 나눌 수 있다. 지도학습의 경우 정답 (ground truth)이 포함된 데이터를 바탕으로 분류 및 평가하는 학습 방법이며, 비지도학습은 데이터의 분포를 분석하여 클러스터를 찾고 분류하는 학습 방법이다. 마지막으로 강화학습은 수행 결과에 따른 보상을 통하여 주어진 환경에서 최적의 방법을 찾는 방법이다. 본 연구는 지도학습에서 사용되는 모델링 기법을 바탕으로 생성된 모델을 대상으로 신용평가 모형을 제시한다. 이를 위해서, 의사결정나무 (decision tree)와 부스팅 (boosting) 계열의 알고리즘을 사용하였다. 그리고 인공지능 모델의 도출한 결과에 대한 설명력 강화를 목표로 한다.

2.2. 의사결정나무

의사결정나무는 각 데이터들이 가진 속성들을 여러 단계의 패턴으로 구분하여 분류 및 예측하는 방법이다. 이때 모델의 의사결정 규칙은 최초의 과정부터 점차 하위로 발산하기에 마치 나무 가지처럼 갈라지는 모습을 보인다. 따라서 의사결정나무라는 이름을 사용한다. 갈라지는 모습을 이용하여 의사결정나무는 정답을 탐색하는 과정을 도식화 할 수 있다는 장점이 있다. 의사결정나무의 구조는 가장 상

위의 뿌리 노드에서부터 시작되며 일정한 임계값을 가지고 있는 분기 (가지분할)를 통해 부모 마디로부터 하위의 자식 마디들이 갈라져 나온다. 이와 같은 의사결정나무는 모든 학습 데이터를 활용한 풀 트리 (full tree) 방법을 사용할 수 있다. 그러나 이러한 방법은 과적합 (over fitting)이 발생할 가능성이 크다. 그러므로 특정 임계점 보다 낮은 수준의 정보들에 대해서 발생하는 분기를 제거하는 가지치기 (pruning)를 통하여 단순화한 모델을 만들 수 있다 (Barros 등, 2012).

2.3. 부스팅 알고리즘

부스팅 알고리즘은 학습 모델의 성능을 개선하기 위해 분류기를 수정하는 아이디어에서 시작한다. 부스팅 알고리즘은 샘플 데이터를 정제하여 여러 개의 분류 모델을 생성하는 방법이다. 여러 개의 단순한 분류기 조합을 통해, 복잡한 분류기의 결과보다 우수한 결과를 얻는 것이 목표이다. 전 단계에서 학습된 결과를 바탕으로 다음 단계의 분류 모델의 학습 샘플 데이터에 대한 가중치를 조정한다. 이와 같은 과정을 통해 이전 단계의 학습 결과가 다음 단계의 학습 결과에 영향을 주게 된다. 그러므로 학습이 진행될수록 분류 경계선 상의 데이터의 가중치가 증가하여 더욱 강력한 분별력을 가질 수 있게 된다. 이러한 특성을 통해서 2.2절에서 소개한 의사결정나무 모형의 과적합 문제를 해결 할 수 있다. 부스팅 알고리즘 기법의 성능을 개선하기 위해서 다음 절에서 소개하는 다양한 방법의 부스팅 알고리즘이 등장하였다 (Natekin와 Knoll, 2013).

2.4. 그레디언트 부스팅 알고리즘

그레디언트 부스팅 기법은 부스팅 알고리즘의 한 종류로 부스팅 기법을 통한 최적의 모델 탐색을 위하여 남은 잔차 (residual)를 개선하는 방법을 사용한다. 부스팅을 통하여 지속적으로 모델을 개선할 경우 학습 데이터를 잘 설명하는 예측 모형 구축이 가능하다. 그레디언트 부스팅은 이러한 잔차를 줄여주는 최적의 파라미터를 찾기 위하여 경사하강방법 (gradient descent)을 사용한다. 이러한 방법은 모델의 편향도를 줄여주지만 과적합이 일어날 수 있기에 샘플이나 정규화 등을 통하여 보완하며 사용하여야 한다 (Natekin와 Knoll, 2013).

2.5. XGBoost (extreme gradient boosting)

XGBoost 방법은 캐글 (Kaggle) 등 다양한 데이터 분석 대회에서 뛰어난 성적을 나타내는 알고리즘으로, 수치 데이터 예측 모델로 다양하게 활용되고 있다. XGBoost는 탐욕 알고리즘 (greedy algorithm)을 사용하여 내부에 생성된 다양한 모델들의 성능을 보완하는 가중치를 탐색한다. 그리고 이러한 가중치는 CART (classification and regression trees)라 불리는 앙상블 모델을 사용하며 모든 최종 노드들이 최종 스코어를 계산하는데 사용된다. 따라서 각 하위 모델들의 분류 성능을 측정할 수 있으며 앙상블에 대한 가중치를 설정할 수 있게 된다. 내부 하위 모델의 앙상블을 통해서 XGBoost는 과적합 문제를 해결할 수 있고 동시에 병렬 처리를 통해서 GBM 대비 학습시간을 단축 할 수 있다는 장점이 존재한다 (Chen과 Guestrin, 2016).

2.6. 설명 가능한 인공지능

설명 가능한 인공지능 (XAI)는 블랙박스와 같은 인공지능 모델을 분석하여 입력에 따른 결과의 변화를 사람들이 이해할 수 있게 하는 기술이다. 딥러닝 이전의 모델들은 인간의 설계에 의해 제작되었다. 따라서 복잡도가 상대적으로 낮기 때문에 모델의 결과 해석에 대한 필요성이 적었다. 그러나 딥러닝 모델의 보급 이후 고차원의 계산을 수행하는 모델이 등장함에 따라 결과에 대한 효율적인 설명을 요구하고

있다 (Kim, 2020). 이를 위해서 모델의 구조나 결과를 해석하는 기법들이 활발하게 연구되고 있다. 특히 GDPR 이후 인공지능 모델 사용에 있어서 설명 가능한 인공지능의 도입은 필수가 되었다.

2.7. 대리 모델 (surrogate model)

복잡한 인공지능 모델의 분석을 보다 유용하게 수행하기 위하여 유사한 기능을 하는 모델을 여러 개 만들어 본래 모델을 역으로 분석하는 기법이다. 대리 모델은 원 모델의 특정 기능을 모사하기 때문에 특정한 상황에서는 동일한 결과를 낼 수 있다. 그러나 대리 모델의 구조는 단순하여 해석이 쉬워야 한다는 제약조건이 있다. 원 모델을 블랙박스로 간주하여 사전 지식이 없어도 대리 모델을 구축 할 수 있게 만들어야 한다. 이러한 대리 모델 생성 방법에 따라 연구 방법이 나눠진다. 학습 데이터를 전부 혹은 일부를 사용하여 대리 모델을 생성하고 해석하는 방법을 글로벌 대리 분석 (global surrogate)이라 한다. 반면에 개별 데이터 하나를 해석하는 과정을 로컬 대리 분석 (local surrogate)이라고 한다. 본 연구에서는 로컬 대리 분석을 중심으로 LIME (local interpretable model-agnostic explanations)과 SHAP (Shapley additive explanations) 모델을 사용하였다 (Adadi와 Berrada, 2018).

2.8. LIME (local interpretable model-agnostic explanations)

LIME은 어떠한 블랙박스 모형에 대해서도 설명 가능한 로컬 대리 분석 기법 중 하나이다. LIME에서는 개별 예측 결과를 설명하기 위하여 입력 데이터를 변형하여 원 모델에 순차적으로 넣어서 나온 값을 해석한다. 이를 통해 입력 값 중에서 변화의 정도가 약하지만 예측 값을 크게 변형하는 변수를 탐색하고 그 정도를 수치화 하여 보여준다. 원 모델을 블랙박스 모형으로 보고 선형으로 근사하는 설명 모형을 만들기에 다양한 모델에 적용이 가능하다. 또한, 이미지나 텍스트 등을 포함한 다양한 데이터 형식에 적용할 수 있다. 하지만 데이터의 변수가 많아질수록 성능이 저하되는 단점이이 존재한다 (Ribeiro 등, 2016).

2.9. SHAP (Shapley additive explanations)

SHAP (Lundberg와 Lee, 2017)는 LIME과 같은 로컬 대리 분석 기법으로 결과 값에 기여하는 각 변수들의 상관관계가 어떤 의사결정이나 행동을 하는지 해석하기 위해서 게임 이론에 기반한 샤플리 값 (Shapley value)을 이용한다. 이러한 샤플리 값은 모델에 활용된 모든 특징 변수들을 이용해서 생성 가능한 모든 조합을 만들고 모든 조합에 대해서 특정 변수의 결과 변화에 따른 기여도 측정을 통해 계산된다. 따라서 모델 전체를 설명하는 것이 가능하고 게임 이론을 통한 이론적 설명이 가능하다는 장점이 있다. 하지만 대용량의 데이터를 모두 처리하기에 연산량이 크다는 단점이 존재한다. 또한 기여도 기반의 모델로 특정 변수의 변화에 대한 설명에는 어려움이 있다. 마지막으로 신규 데이터 입력 시 기존에 학습된 정보가 없다면 학습 데이터를 바탕으로 이와 비슷한 가상의 데이터를 만들어야 하는 한계점이 존재한다 (Lundberg와 Lee, 2017).

3. 제안 모델

본 연구에서는 개인의 신용평가등급 변화에 대한 이유를 제공하기 위해, 기존 공개된 개인 신용관련데이터로부터 가상의 신용평가등급 변화 데이터를 생성하고, 이를 활용하여 개인별 신용평가등급 변화의 이유를 설명하는 모델을 제안한다.

개인이 각기 다른 시점에서 상이한 신용평가등급을 받은 경우, 이와 같은 변화를 납득하기 위해서 신용평가등급 변화에 주요하게 영향을 준 신용정보 변수를 설명할 필요가 있다. 그러나 기존 공개된 신용평가 관련데이터는 어느 한 시점에서 신용평가를 실시한 결과로 동일인의 신용평가등급 변화를 학습하기에 적합하지 않다. 따라서 본 연구에서는 신용평가 변수 변화량에 따른 신용평가등급 변화를 설명하기 위해, 기존 신용평가 데이터를 활용하여 가상의 신용평가등급 변화 데이터를 생성하고, 이를 활용하여 각 신용평가 변화량을 학습 후 영향력이 큰 변수를 설명하는 모델을 제시한다. 전체적인 모델 구성은 Figure 3.1과 같다.

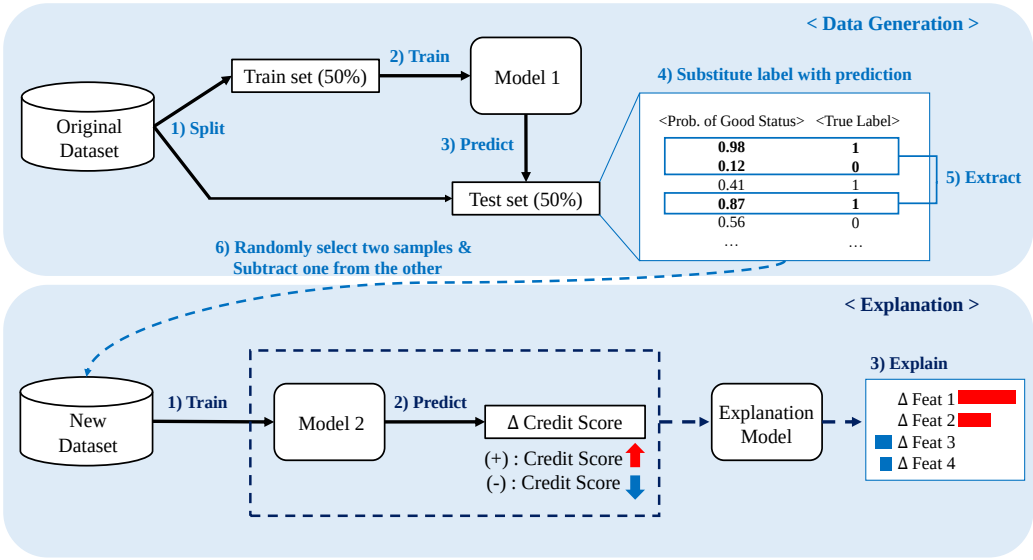

Figure 3.1 Explainable credit changes architecture

3.1. 데이터 생성

데이터 생성을 위해 기존 공개된 HELOC (home equity line of credit) Dataset이 사용되었다 (Yeh와 Lien, 2009). 해당 데이터 셋은 FICO Explainable Machine Learning Challenge에서 제공하는 개인 신용관련데이터로 총 10,459명의 데이터로 구성되어 있으며, 각 사람별로 23개의 신용정보 (채무이행 비율, 거래량 등)를 포함한다. 이 중에서 대출 가능과 불가한 데이터의 수는 각각 5,000개, 5,459개로 약 1:1 비율로 구성되어 있다. 원 데이터는 22개의 변수와 1개의 정답레이블로 구성되어 있으나, '가장 큰 채무불이행 (MaxDelqEver)' 항목의 경우 범주형 변수 (categorical variable)로 원-핫 인코딩 (one-hot encoding)을 시행하여 최종 28개의 변수와 1개의 정답레이블로 구성된 데이터로 변형시켰다 (Daly 등, 2016).

전체데이터를 1:1 비율로 학습데이터와 평가데이터로 분할 후, 학습데이터로 모델1을 학습한다. 학습된 모델1로 평가데이터에 대하여 대출 가능한 확률을 예측하고 해당 확률 값을 평가데이터의 레이블로 대체한다. 이러한 이유는 각 사람의 신용평가수준이 모두 동일하지 않으며, 대출 가능 범위에서도 다양하게 분포하고 있을 것이라고 가정하였다. 그리고 모델1의 불확실성이 평가데이터 레이블에 반영되는 것을 방지하기 위하여, 실제 레이블과 방향성이 일치 (0.5이하: 대출 불가능 / 0.5이상: 대출 가능)하는 평가데이터만을 추출하였다. 예를 들어, 실제 레이블은 '대출 가능'이며 예측 레이블이 0.5이하인 경우, 해당 데이터는 제거된다. 반면에 실제 레이블은 '대출 불가능'이며 예측 레이블이 0.5이하인 경우, 해당 데이터는 유지한다. 이와 같은 방법으로 추출된 데이터 셋에서 무작위로 샘플 두개씩을 선택 후 각 변수별 변화량을 계산하여 새로운 데이터 셋을 생성하였다.

3.2. 학습 및 설명 모델

3.1절에서 생성된 데이터 중 28개 신용정보 변수의 변화량은 모델2의 입력 값으로 사용되며, 출력 값은 대출 가능한 확률의 변화량으로 -1에서 1사이의 값을 가진다. 각 데이터 샘플에서 대출 가능 확률의 변화량에 영향을 준 변수를 분석하기 위해, 모델2를 학습 후 SHAP을 사용하여 예측결과를 설명하였다.

4. 실증분석

4절에서는 기존 신용평가 데이터를 학습하여 신용평가모델별 성능을 비교한 후 설명모델을 사용하여 신용평가 등급에 영향을 주는 변수분석을 진행하고 3절에서 제시한 모델을 여러 평가환경에서 분석하여 최적 데이터 생성량을 찾고 데이터 추출단계의 효과를 검증하였다. 그리고 대출 가능 확률의 변화량에 영향을 준 변수를 분석하여 시각화하였다.

4.1. 신용평가등급 모델 성능 비교

신용평가등급 모델의 성능 비교를 위해, 기존 공개된 세 가지 데이터 셋에 XGBoost, GBM (gradient boosting algorithm)과 랜덤 포레스트 (random forest) 분류기로 학습 및 평가를 시행하였다 (Chen과 Guestrin, 2016; Friedman, 2002; Liaw와 Matthew, 2002). 사용된 데이터는 HELOC, Lending club (미국 대출 회사 Lending club에서 개인 신용평가를 위해 2007년부터 2011년까지 수집된 데이터 셋 `http://www.lendingclub.com/info/download-data-action`)과 Default of credit card clients dataset (UCI) (UCI machine learning repository에서 공개한 개인의 파산 가능성 예측 데이터 셋; `https://community.fico.com/s/explainable-machine-learning-challenge?tabset-3158a=2`)로 모두 범주형 변수를 포함하고 있어 해당 변수들에 대해 원-핫 인코딩을 시행하여 데이터 전처리 후 모델을 학습하였다. 성능 평가는 AUPRC (area under the precision recall curve), AUROC (area under the receiver operating characteristic curve), 균형 정확도 (balanced accuracy) 및 정확도 (accuracy)를 사용하여 측정하였으며 해당 결과는 Table 4.1에 나타내었다.

Table 4.1 Test performance of classification models on various credit dataset (%)

Dataset	Metric	Classification model		
		XGBoost	GBM	Random forest
HELOC	AUPRC	77.5	75.8	74.3
	AUROC	78.4	77.3	75.4
	Balanced Accuracy	71.3	70.6	68.5
	Accuracy	71.5	70.7	68.8
Lending Club	AUPRC	56.8	56.1	53.0
	AUROC	90.3	90.3	88.5
	Balanced Accuracy	60.0	60.4	61.2
	Accuracy	87.1	87.1	86.6
UCI	AUPRC	52.3	47.6	45.3
	AUROC	75.4	72.1	69.3
	Balanced Accuracy	61.7	62.1	52.7
	Accuracy	81.1	80.2	78.6

Table 4.1의 결과와 같이 XGBoost 모델은 HELOC 데이터 셋에서 GBM 및 랜덤 포레스트와 비교했을 때, 가장 높은 성능을 보였다. Lending Club과 UCI 데이터 셋의 경우 균형 정확도는 GBM 분류기가 가장 높은 성능을 보였으나, 그 외의 평가에서는 XGBoost 모델이 높은 성능을 보였다. 각 데이터 셋에서 XGBoost의 AUPRC는 HELOC에서 가장 높았으며, Lending Club 및 UCI 데이터 셋에서 얻은 AUPRC와 큰 성능 차이를 보였다.

HELOC와 같은 경우, 대출 가능과 불가능의 비율이 약 1:1인 것과는 달리, Lending club 및 UCI 데이터는 비율이 3.5:1로 불균형하게 구성되어 있다. 본 연구의 목표는 다양한 구간의 신용평가 변수의 변화량을 학습하여 대출 가능 변화량에 주요하게 영향을 주는 변수를 분석하는 것이다. 따라서 불균형 데이터로 변수의 변화량 데이터를 생성하게 될 경우, 대출 불가능 구간 데이터 사이에서 생성된 데이터가 다수를 이루는 데이터 셋이 생성되며, 다양한 구간의 변수 변화량을 포함하지 못하는 문제점이 생긴다. 따라서 대출 가능과 불가능의 비율이 고르게 분포된 HELOC 데이터를 사용하는 것이 변수 변화량 데이터를 생성하는데 적합하다고 판단하였다.

Figure 4.1에서는 HELOC 데이터 셋을 GBM Classifier으로 학습한 결과를 SHAP을 사용하여 변수들의 중요도를 분석하였다. 평가데이터 100개 샘플에서 변수들의 SHAP value의 평균 절대 값을 구하여 순서대로 나열하였다. 평균 절대 값이 가장 높은 변수로는 '신용한도 대비 리볼빙 잔금 비율 (NetFractionRevolvingBurden)', '개인의 모든 대출, 상환 기록의 평균 개월 수 (AverageMinFile)', ' 지난 7일을 제외한 가장 최근 신용 거래 조회로부터 경과 개월 수 (MSinceMostRecentInqexcl7days)'가 있었다. 해당 변수들은 Figure 4.1에서 나타낸바와 같이 신용등급이 높거나 낮은 경우 모두에게 영향을 미침을 알 수 있다. 또한, SHAP 설명 모델은 개별 데이터별로 중요하게 영향을 준 변수를 분석할 수 있으며 Figure 4.2와 같이 나타낼 수 있다. 해당 샘플은 GBM 모델이 대출 가능 확률을 0.27로 예측하였으며, 이와 같은 결과에 '개인의 모든 대출, 상환 기록의 평균 개월 수 (AverageMinFile)'가 부정적으로 작용한 것을 확인할 수 있다.

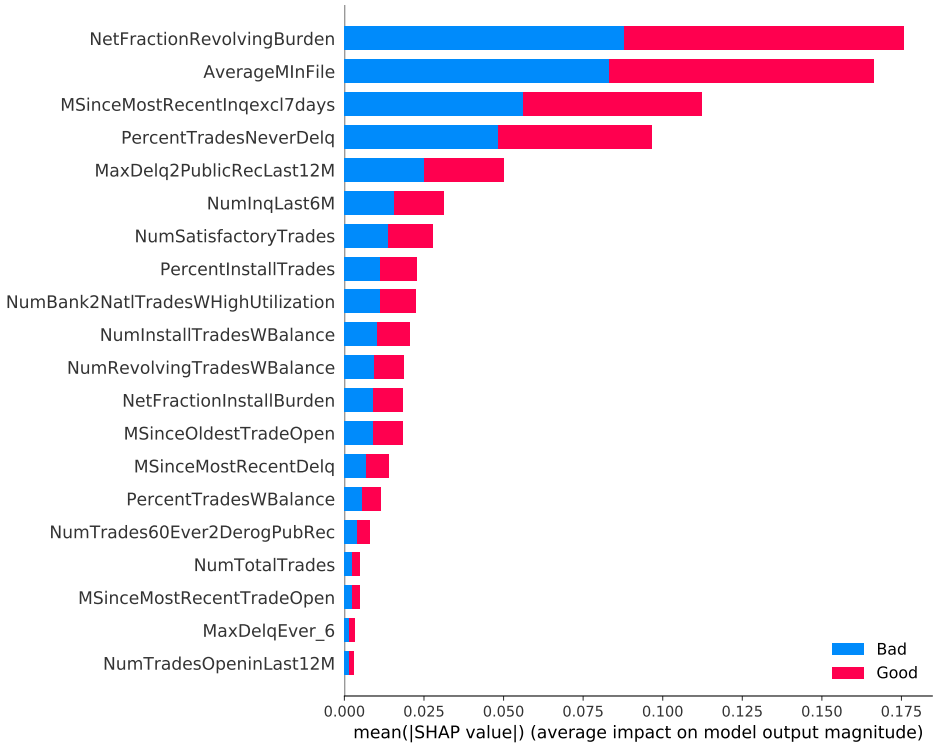

Figure 4.1 Summary plot of SHAP values

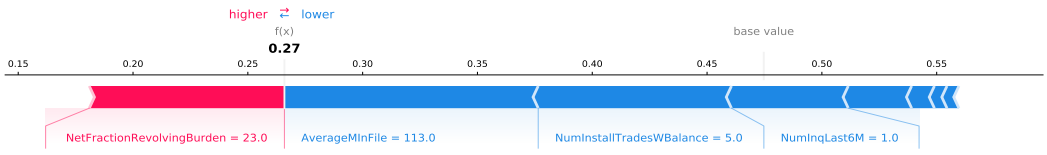

Figure 4.2 SHAP values of individual sample (credit score)

4.2. 데이터 수에 따른 성능 평가

3.1절에서 설명한 것과 같이 평가데이터의 실제 정답을 참고하여 추출된 데이터 셋에서 무작위로 샘플 두개씩을 선택 후 각 변수별 변화량을 계산하여 새로운 데이터 셋을 구성하게 된다. 이때 생성된 데이터 셋을 8:2로 나누어 각각 학습데이터와 평가데이터를 생성하도록 하였다. 그리고 무작위로 샘플을 선택하는 횟수에 따른 성능을 비교하고자 10,000부터 200,000까지 10,000 단위로 증가시키며 학습데이터를 생성하였고, 평가데이터는 20,000개로 고정하였다. 또한, 모델1을 XGBoost와 GBM Classifier를 각각 사용하여 서로 다른 학습데이터 및 평가데이터를 생성하고 모델2는 GBM Regressor를 사용하는 것으로 통일하여 성능을 비교하였다. 성능평가는 각 평가데이터의 예측 값과 정답의 평균 제곱근 오차 (root mean square error; RMSE)를 측정하였다. 검증평가 성능은 10-fold 교차검증 (cross validation)을 하여 평균 제곱근 오차의 평균을 구하였다. 추가로 정답 참고 추출과정의 효과를 검증하기 위해 추출단계 전 데이터 셋으로도 위의 단계를 반복하여 실험하였다. 이때, 평가는 추출단계를 거친 후 생성된 평가데이터로 통일하였다. 모델별 학습 성능 비교와 학습 데이터 수에 따른 성능비교는 Figure 4.3과 Table 4.2에 나타낸 바와 같다.

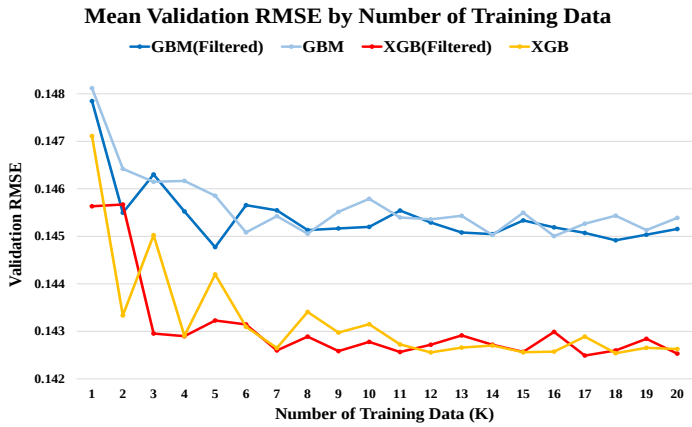

Figure 4.3 Mean validation RMSE by number of training data

Figure 4.3과 Table 4.2의 결과처럼 각 모델은 학습데이터가 증가함에 따라 평균 제곱근 오차가 감소하는 경향을 보였다. GBM (Filtered)은 GBM으로 학습 후 예측한 값과 실제 정답을 참고하여 추출단계를 거친 후 가공된 데이터를 학습한 모델로, 학습데이터가 50,000개일 때 검증평가에서 가장 낮은 평균 제곱근 오차, 14.48%를 보였으며 이때 평가 성능은 14.61%였다. 반면에, GBM은 추출단계를 거치지 않은 데이터를 학습한 모델로, 학습데이터가 140,000개와 160,000개일 때 가장 높은 성능을 보였다. 이때, 평가 성능은 14.63%와 14.66%로 모두 GBM (Filtered) 보다 낮은 성능을 보였다. 마찬가지로, XGB (Filtered)와 XGB를 비교하였을 때, XGB (Filtered)가 높은 성능을 보였다. 이는 데이터 생성에 있어 정답을 참고하는 추출단계가 효과적임을 뒷받침한다.

Table 4.2 Validation(V) and test(T) RMSE (%)

Training Data (K)	GBM(Filtered) V	GBM(Filtered) T	GBM V	GBM T	XGB(Filtered) V	XGB(Filtered) T	XGB V	XGB T
10	14.78	14.84	14.81	14.90	14.56	14.51	14.71	14.60
20	14.55	14.65	14.64	14.67	14.57	14.34	14.33	14.40
30	14.63	14.78	14.61	14.71	14.30	14.16	14.50	14.39
40	14.55	14.75	14.62	14.66	14.29	14.07	14.29	14.30
50	14.48	14.61	14.59	14.68	14.32	14.27	14.42	14.35
60	14.57	14.62	14.51	14.63	14.31	14.29	14.31	14.23
70	14.55	14.73	14.54	14.64	14.26	14.13	14.26	14.31
80	14.51	14.64	14.51	14.68	14.29	14.11	14.34	14.32
90	14.52	14.60	14.55	14.57	14.26	14.18	14.30	14.26
100	14.52	14.67	14.58	14.63	14.28	14.07	14.32	14.38
110	14.55	14.57	14.54	14.67	14.26	14.19	14.27	14.31
120	14.53	14.71	14.54	14.63	14.27	14.13	14.26	14.23
130	14.51	14.58	14.54	14.62	14.29	14.21	14.27	14.26
140	14.50	14.60	14.50	14.63	14.27	14.15	14.27	14.25
150	14.53	14.65	14.55	14.66	14.26	14.18	14.26	14.20
160	14.52	14.68	14.50	14.66	14.30	14.19	14.26	14.32
170	14.51	14.62	14.53	14.63	14.25	14.19	14.29	14.29
180	14.49	14.58	14.54	14.62	14.26	14.22	14.25	14.27
190	14.50	14.64	14.51	14.60	14.28	14.08	14.27	14.28
200	14.52	14.58	14.54	14.59	14.25	14.21	14.26	14.30

4.3. 대출 가능 확률의 변화량에 영향을 준 변수 분석

4.2절의 결과에서 나타낸 것과 같이 XGBoost로 예측한 값을 추출단계를 거쳐 GBM Regressor로 학습한 모델 중 학습데이터 170,000개를 사용한 모델이 가장 높은 성능을 보였으므로 이를 사용하여 SHAP 분석을 진행하였다. Figure 5은 대출 가능 확률의 변화량에 영향을 준 변수를 시각화 한 결과이다.

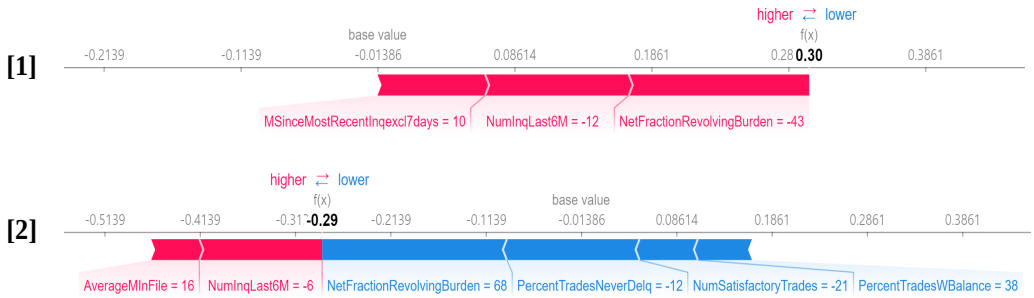

Figure 4.4 SHAP values of individual sample (Changes in credit score)

Figure 4.4에서 붉은색은 신용등급평가 향상에 영향을 준 변수이며, 푸른색은 신용등급평가 하락에 영향을 준 변수이다. [1]의 경우 '신용한도 대비 리볼빙 잔금 비율 (NetFractionRevolvingBurden)'이 43 줄어든 것이 신용등급 향상에 가장 큰 영향을 주었다. [2]의 경우 '최근 6달 동안 신용 거래 조회 수 (NumInqLast6M)'가 감소한 것은 신용등급 향상에 영향을 주었으나, '신용한도 대비 리볼빙 잔금 비율'이 68 증가한 것이 신용등급 하락에 더 큰 영향을 주었다.

5. 결론

본 연구에서는 HELOC, Lending club과 Default of credit card clients dataset (UCI)과 같은 실제 신용 평가 데이터를 활용하여 인공지능 기반의 신용평가 모형과 그 결과를 설명하는 알고리즘을 제안하였다. 신용평가 모형을 생성하기 위해서, XGBoost와 GBM classifier를 사용하였다. XGBoost와 GBM를 세 개의 다른 데이터 셋에서 실험한 결과, 정확도는 두 방법에 대해 비슷하게 하게 나와서, 향후 후속 연구자가 인공지능 기반의 신용평가 모형을 설계하는데, 참고할 수 있다는 점에서 의미가 있다. 머신러닝을 통해서 도출된 등급을 설명하기 위해서, 대리인 기반의 설명 가능한 인공지능 방법을 적용하였다. 그 결과, 인공지능이 개인의 신용 등급을 평가하는 데 필요한 중요한 특징 변수들이 무엇, 무엇, 무엇 인지 확인할 수 있었다. 마지막으로 제안된 알고리즘을 통해서, 신용 등급이 변동되었을 때, 변동 원인이 되는 이유를 설명 할 수 있다. 이 부분은 기존의 설명 가능한 인공지능 알고리즘이 단순히 어떤 결과를 도출했을 때, 중요한 원인이 되는 특징들을 선별하는 데 그치는 한계를 극복하였다. 단순한 결과 설명 분만 아니라, 다른 두 결과 사이의 변동을 위한 방법을 제시하였다는 점에의 의의가 있다.

지난 2018년 EU는 새로운 개인정보 보호규정인 GDPR을 시항하였다. 이 규정을 통해서, EU에 포함된 회원 국가들 사이의 개인정보 보호 수준을 표준화 하고, 자유로운 데이터 거래를 촉진시키고자 하고 있다. 따라서 EU 내에서 데이터와 인공지능을 이용한 경제와 산업의 디지털화가 예상된다. 이 보호 규정에서 중요한 사항 중 하나는, 알고리즘에 의한 자동화된 처리를 진행할 때, 소비자는 관련된 설명을 요구할 권리를 삽입한 것이다. 즉, 인공지능을 이용한 알고리즘을 가지고 제품과 서비스를 제공하는 기업들은 그 결과에 대해 설명할 수 있어야 한다. 금융 산업에도 인공지능을 도입할 때, 주의해야 할 사항이다. 인공지능 기반의 금융 서비스 결과에 대해서, 설명력을 제공하지 못할 경우에는, 서비스 품질의 우수함에도 불구하고 적용할 수 없기 때문이다. 따라서 금융 산업에서 설명 가능한 인공지능 기술에 대한 연구의 중요성은 더욱 더 높아질 것으로 예상된다. 본 연구는 이러한 시대의 흐름에 맞추어, 설명 가능한 인공지능 연구를 처음으로 금융 산업에 도입하였다는 점에서 큰 의미가 있다. 금융 서비스를 제공하는 금융 기관에서는 고객에게 정밀한 신용 등급 평가의 결과를 제공하는 것과 함께, 그 결과에 대한 설명도 제공할 수 있다. 또한, 동일한 개인에 대해서 신용 등급의 변화에 대한 설명도 제공할 수 있다.

따라서, 본 논문은 아래와 같은 시사점을 도출한다. 첫째, 은행에서 실제 신용평가 데이터를 활용하여 인공지능 기술을 적용하였다는 점에서 후속 엔지니어들이 실제로 비슷한 시스템을 구축할 때, 이론적, 기술적 근거가 될 수 있다. 둘째, 신용평가에서 인공지능 기술을 실제로 적용했을 때, 신용 등급 결과를 해석하는 시스템을 최초로 제안하였고, 금융 고객들이 자신의 신용 등급에 결과에 대한 이해도를 높일 수 있다. 마지막으로, 신용평가에 대한 산출 근거뿐 만 아니라, 신용 등급 변화 원인에 대한 이유를 설명하는 시스템을 최초로 제안하였다는 점에서 의미가 있다, 이를 통해서 금융 기관은 고객이 신용 등급 변화의 이유를 질의했을 때 정량적으로 답을 제시할 수 있다.

하지만, 제안된 알고리즘은 변화를 위한 최적의 방안을 제시하지는 못한다는 한계가 존재한다. 고객들이 필요한 것은 자신의 신용등급에 대한 설명뿐만 아니라, 다른 등급으로 이동하기 위한 방법이기 때문이다. 등급 변화를 위해 필요한 다양한 특징 변수들 중에서는 짧은 시간에 변화 시킬 수 없는 특징들도 존재한다. 따라서 이러한 특징 변수들의 특성에 대한 이해를 바탕으로, 최소한의 노력으로 최대의 등급 변동의 효과를 설명할 수 있는 방법에 대한 연구가 뒷받침 된다면, 인공지능을 이용한 신용평가 서비스 도입에 큰 도움이 될 것으로 기대한다.

References

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, **6**, 52138-52160.

Barros, R. C., Basgalupp, M. P., De Carvalho, A. C. P. L. F. and Freitas, A. A. (2012). A survey of evolutionary algorithms for decision-tree induction. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, **42**, 291-312.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Cho, S. (2011). Restoring the role of credit rating agencies as gatekeepers. *KDI Journal of Economic Policy*, **33**, 81-110.

Daly, A., Thijs, D. and Hess, S. (2016). Dummy coding vs effects coding for categorical variables: Clarifications and extensions. *Journal of Choice Modelling*, **21**, 36-41.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, **38**, 367-378.

Jang, D. H., Ha, I. D., Park, D. J., Park, I. H. and Lee, S. J. (2020). Statistical model development for high-dimensional big data analytics. *Journal of the Korean Data & Information Science Society*, **31**, 1009-1020.

Kim, J. (2020). AI filtering and explainable AI. *LAW & TECHNOLOGY*, **16**, 83-92.

Kim, M. G. (2012). *A study on credit assessment using WOE measure, Logistic regression and evolutionary programming*, Master's thesis, Department of Mathematics, Hanyang University.

Lee, J. (2020). Access to finance for artificial intelligence regulation in the financial services industry. *European Business Organization Law Review*, **21**, 731-757.

Lee, K. C., Han, I. G. and Kim, M. J. (1996). A study on the credit evaluation model integrating statistical model and artificial intelligence model. *Korean Management Review*, **21**, 81-100.

Liaw, A. and Matthew, W. (2002). Classification and regression by random forest. *R News, 2.3*, 18-22.

Lundberg, S. M. and Lee, S. (2017). A unified approach to interpreting model predictions. *In Advances in Neural Information Processing Systems*, 4765-4774, Curran Associates, Inc.

Natekin, A. and Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, **7**, 21.

Qiu, X. and Choi, P. (2019). A study on discrimination in mortgage lending in the United States: A revisit by random forest method. *Journal of the Korean Data & Information Science Society*, **30**, 261-270.

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.

Yang, N. (2020). AI assisted internet finance intelligent risk control system based on reptile data mining and fuzzy clustering. *In 2020 Fourth International Conference on I-SMAC*, 533-536, IEEE.

Yeh, I. C. and Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, **36**, 2473-2480.

Zhen, W. (2019). *Machine learning applications in finance: Some case studies*, Ph. D.'s thesis, Department of Computing, Imperial College London.

Zheng, X., Zhu, M., Li, Q., Chen, C. and Tan, Y. (2019). FinBrain: When finance meets AI 2.0. *Frontiers of Information Technology & Electronic Engineering*, **20**, 914-924.

Journal of the Korean Data & Information Science Society
2021, **32**(2), 283–295

http://dx.doi.org/10.7465/jkdi.2021.32.2.283
한국데이터정보과학회지

Study on credit rating model using explainable AI

Ye Eun Chun[1] · Se Bin Kim[2] · Ja Yun Lee[3] · Ji Hwan Woo[4]

[1234]AI Competency Center, Shinhan Bank
[4]Korea University School of Management of Technology

Received 21 December 2020, revised 21 January 2021, accepted 27 January 2021

Abstract

As artificial intelligence technology develops, cases of applying it to the financial industry are increasing. However, the biggest drawback is that understanding the process how the results are derived is not intuitive because most of its relationships are non-linear. Therefore, the process of deriving results using artificial intelligence is sometimes expressed as a black box. Recently, the EU created a new privacy regulation, guaranteeing the right of customers to request service providers for explanations about the results obtained by artificial intelligence algorithms. In other words, to apply artificial intelligence technology in the financial industry, not only high precision but also the ability to explain the results must be considered. In this paper, using various externally disclosed credit information data, an artificial intelligence-based credit rating algorithm was proposed. Also, for the results derived by artificial intelligence, we introduced an algorithm to calculate and distinguish which of the various characteristics of the data has the most significant effect. Finally, we further expanded this by applying the method of explaining the modified result to the financial data to explain when there is a change in the result derived by artificial intelligence. This research has great significance as it confirms that the proposed method can provide explanatory power when introducing artificial intelligence technology in financial services.

Keywords: Credit rating model, digital finance, explainable AI, XAI.

[1] AI Competency Center, Shinhan Bank, 55, Sejong-daero, Jung-gu, Seoul 04513, Korea.
[2] AI Competency Center, Shinhan Bank, 55, Sejong-daero, Jung-gu, Seoul 04513, Korea.
[3] AI Competency Center, Shinhan Bank, 55, Sejong-daero, Jung-gu, Seoul 04513, Korea.
[4] Corresponding author: AI Competency Center, Shinhan Bank, 55, Sejong-daero, Jung-gu, Seoul 04513, Korea. E-mail: jihwan_woo@korea.ac.kr