인공지능 모형의 신뢰성 확보 방안에 관한 고찰

-설명 가능한 인공지능의 활용사례를 중심으로-

김윤명[1], 김영묵[2]
[1]건국대학교 산업공학과
[2]인하대학교 컴퓨터공학과

qwertytoki@konkuk.ac.kr[1], mukmookk@inha.edu[2]

A Study on the Strategies for Ensuring Trustworthiness of Artificial Intelligence Modeling

- Focusing on eXplainable AI's Use Cases-

Yoon-Myung Kim[1], Youngmuk Kim[2]
[1]Dept. of Industrial Engineering, Kon-Kuk University
[2]Dept. of Computer Engineering, INHA University

요 약

본 논문에서는 설명가능한 머신러닝 모델과 관련된 다양한 도구를 활용해보고, 최근 각광받는 주제인 신뢰성 에 대해서도 고찰해보았다. 근래의 인공지능 모델은 설명력을 덧붙여 정보 장벽을 낮추는 방향으로 진화하고 있다. 이에 따라 AI 모델이 제공하는 정보량이 늘고 사용자 친화적인 방식으로 바뀌면서 사용자층이 확대되고 있는 추세이다. 또한 데이터 분석 분야의 영향력이 높아지고 연구 주체들이 다양해지면서, 해당 모델이나 데이터에 관한 신뢰성을 확보해야한다는 요구가 많아지고 있다. 이에 많은 연구자들이 인공지능 모델의 신뢰성의 확보를 위해 노력하고 있다. 본 연구에서는 이러한 노력의 발자취를 따라가보면서 인공지능의 설명가능성에 관하여 소개하려고 한다. 그 과정에서 민감한 데이터를 다루어보면서 신뢰성 확보의 필요성에 대해서도 논의해보려고 한다.

I. 서론

2021년 11월 Gartner에 따르면 전 세계에 인공지능(Artificial Intelligence, 이하 AI) 소프트웨어 시장이 2022년에는 62만 달러(약 75억원) 크기로 성장할 것으로 예측되었다. 이러한 성장은 AI 기술이 산업계에 미치는 영향력이 점차 커지고 있다는 것과, 새로운 기회가 창출되는 정도 및 기술 성숙도(Maturity)가 높아지고 있음을 의미한다.

위에서 기술한 시장 변화에 힘입어 현재 데이터 분석과 기계 학습 분야에서 다양한 기법이 소개되고 있고, 이를 활용한 서비스 또한 활성화되고 있는 추세이다. 그 중 하나가 설명 가능한 인공지능(eXplainable AI, 이하 XAI)이다. XAI는 전문가들만의 영역이었던 데이터 분석

분야의 지평을 넓힘으로써 인공지능 모델의 사용자층을 확대하는 결과를 가져왔다.

물론 XAI가 제공하는 맞춤 정보들은 친절하고 자세하지만, 데이터 분석 결과로 결정적인 판단을 내려야 할 경우에 해당 정보들로 전문지식이 적은 고관여층까지 설득하기에는 어려움이 있다. 실제로 데이터 및 모델의 신뢰성 확보는 현업에 종사하는 많은 데이터 분석가들에게 가장 큰 난제 중 하나이고, 이 간극을 메우기 위해 AI 컨설턴트라는 새로운 직업까지 등장하기에 이르렀다.

이러한 전문가-비전문가 사이의 간극을 줄일 수 있다면 많은 정보 자원을 창출함과 동시에 사회적 비용을 대폭 줄일 수 있을 것으로 기대된다. 이러한 노력의 일환으로 개발자들은 기존 XAI에 신뢰성을 더할 여러 요소들을 고려하고 있다. 그 중에서도 특히 인구통계학적 데이터에 녹아있는 편향성을 완화할 수 있는 Fairness 측면이 주목받고 있다.

이미 EU에서는 이러한 요소들을 반영한 guideline을 공표하여 사용 중에 있다. 이러한 변화에 발맞추어, 본 논문에서는 Biased data를 활용하여 인공지능 모델을 구축하고 해당 모델에 오픈 소스 XAI 도구를 적용하여 현존하는 Black box 문제를 다양한 시각에서 접근한다.

II. 설명 가능한 인공지능 모델(XAI)

광범위한 산업 분야에서 더욱 복잡하고 발전된 AI가 사용됨에 따라, AI를 적용하기에 앞서 알고리즘의 동작 과정을 이해하고 설명해야하는 문제가 발생하였다. XAI는 이러한 요구에 발맞추어 사용자가 기계 학습 알고리즘의 결과 및 출력을 이해할수 있도록 하는 일련의 프로세스 및 전략으로 LIME· SHAP·LRP와 같은 도구를 제시하였다.

이 중 설명 가능한 부스팅 머신(Explainable Boosting Machine, 이하 EBM)은 자동 상호 작용 감지가 있는 트리 기반 순환 Gradient Boosting 일반화 부가 모델이다. EBM은 최신 블랙박스 모델만큼의 정확도를 가지면서도 Glass box 모델의 일종으로서 해석력이 좋으며, 여타 알고리즘에 비해 훈련 속도가 느린 대신 예측에서는 빠른 속도를 보장하는 것이 특징이다.

설명력을 고려하지 않은 블랙 박스 모델의 경우, 입력이 주어졌을 때 이에 대한 결과가 도출되기까지의 과정에 대해 사용자로서는 알 도리가 없다. 그렇다면 금융·의료·행정·사법과 같이 자칫 치명적일 수 있는 분야에 인공지능을 도입하는 것이 실무자 입장에서는 부담으로 작용할 수 밖에 없다.

본 절에서는 심근 경색 예측 데이터를 활용하여 XAI 모델(InterpretML)[1]을 구축하고 여기에 존재하는 한계점에 대해서 알아보고자 한다. 먼저 데이터는 BMI·알코올 섭취 주기·당뇨·흡연 등 심장 질환 발병 여부와 관련이 있다고 추정되는 22개의 독립변수와 종속변수인 심장 질환 여부로 구성되어 있다. 일부는 Dummy 변수 형태로 활용했으며, 대표적인 Glassbox 모델인 EBM을 통해 Local/Global 두 가지 관점에서 모델의 설명력을 확보하였다. 추가적으로 PII을 통한 비식별화 과정을 거쳐 대상 특정 가능성을 배제하였다.

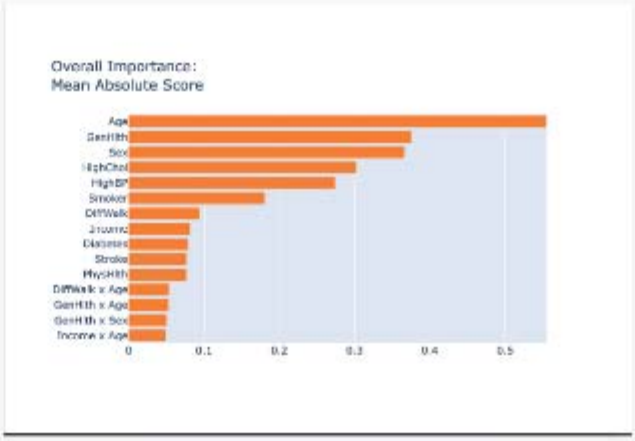

그림 1. 심근 경색 데이터를 활용한 EBM 모델의 Global explainability 시각화

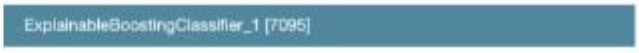

그림 2. 심근 경색 데이터를 활용한 EBM 모델의 Local Explainability 시각화

분석 결과 그림 1의 Global Explainability 측면에서는 Age(나이), GenHLTH(본인 건강에 대한 주관적 평가), Sex(성별), HighChol(높은 콜레스트롤) 순으로 특성의 중요도가 높은 것을 확인할 수 있었다. 그림 2에서는 Local Explainability 측면에서 시각화된 대시보드를 통해, GenHLTH(본인 건강에 대한 주관적 평가)·Stroke(뇌졸중 발생 여부)등이 양의 가중치, Age가 음의 가중치에 해당함을 확인하였고, 이를 계산에 활용하여 심근 경색 예측에 활용하였다.

이와 같이 XAI는 Feature에 대한 해석력을 보강해주는 역할을 중점으로 발전해왔다. 많은 경우 Blackbox 모델을 수학적 해석 기법을 통해 해석하거나 해석력이 포함된 Glassbox 모델을 활용한다. 이러한 방식은 인공지능 및 수학적 지식이 부족한 일반 사용자들에게 충분한 설득력을 갖지는 못하므로, 다양한 사용자층에게 폭넓은 신뢰를 얻을 수 있는 접근이 필요하다.

III. 신뢰 가능한 인공지능 모델(Trustworthy AI)

신뢰 가능한 인공지능 모델(Trustworthy AI, 이하 TAI)은 Responsible AI 라고도 불리며, 기존의 XAI 개념에 ethics 측면을 보강한 것이다. 현재 데이터 분석과 기계학습(Machine Learning, 이하 ML)분야는 빠른 속도로 성장하면서 기술의 성숙도가 높아지고 있는 추세이다.

그러나 이를 무분별하게 AI에 접목시키면 원치 않는 결과를 맞닥뜨릴 수 있다. 데이터 속에는 bias(편향)를 비롯한 사회의 어두운 면이 녹아있고, 인공지능 모델은 데이터를 기반으로 학습하기 때문이다.

따라서 사회에 대한 고찰을 통해 개념을 구체화하여 AI에게 알고리즘·데이터·관행 측면에서의 Ethics를 가르쳐야 한다. AI에게 이를 가르치지 않고 무분별하게 기술만 재생산하게 되면, 데이터에 존재하는 사회의 어두운 면으로 인해 치명적인 결과를 초래할 수 있다.

핵심은 초기 개발 단계에서 올바른 ethics 개념을 AI에게 항목화된 지표로써 전달하고, 판단 기준도 적절히 조정해야 한다는 것이다. 그러려면 이 분야에 지식이 없는 사람에게도 설득력을 지닐만큼 Plausible한 AI 모델을 구축해야 하고, 이 과정에서 Explainability에 더불어 Fairness와 Transparency 등을 확보하여야 한다.

4. 결론 및 제언

선행연구[2] 혹은 연구 기관마다 신뢰가능성을 정의하는 기준은 다르다. IBM AI Research[1]는 TAI를 인공지능 알고리즘과 그 결정을 이해할 수 있는 프레임워크라고 정의하는 한편, Fairness·Robustness, Explainability·accountability 등 사회의 특성을 담아낼 수 있는 요소들을 갖추어야 한다고 말한다. Caltech AI Research[2]에서는 여기에 더해 TAI를 불확실성(Uncertainty)에 대응하는 기술적 도전의 일환으로 본다.

데이터는 인간의 산물이므로 언제나 원치않는 편향과 불확실성이 내재되어 있다. 아마도 이러한 부정적 요소들을 걷어내고 중요한 판단을 수행하는 권한을 인공지능에게 넘기기까지는 수많은 장벽이 존재할 것이다. 그러나 인간은 언제나 장벽을 넘어왔고, 이미 많은 사회적 비용을 인공지능을 통해 해소해온 바 있다.

인공지능을 만능 열쇠처럼 여기자는 풍조에 일조하고자 하는 것은 아니다. 인류의 역사를 가만히 돌아보면 수많은 결정들이 편향으로 인해 그릇된 결과를 불러온 바 있다. 하지만 이미 우리는 궂은 일을 대신 해주는 보조 도구로써 인공지능을 성공적으로 활용하고 있다. 여기에 신뢰성을 확보하여 사회의 부정적 편향을 덜어낼 공정한 보조 심판관을 우리 곁에 둘 수 있다면 막대한 사회적 비용을 줄일 수 있을 것이라고 기대한다.

참조 문헌

[1] "InterpretML: A Unified Framework for Machine Learning Interpretability" (H. Nori, S. Jenkins, P. Koch, and R. Caruana 2019)

[2] “Why Should I Trust You?”: Explaining the Predictions of Any Classifier (Ribeiro et al., NAACL 2016)

※ 본 프로젝트는 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT 멘토링 프로젝트 결과물입니다.

[1] https://research.ibm.com/topics/trustworthy-ai
[2]https://scienceexchange.caltech.edu/topics/artificial-intelligence-research/trustworthy-ai