AutoML 과 XAI 의 결합 : 기계학습 모델의 자동화와 해석력 향상을 위하여

손민혁 [1*], 김남훈 [2*], 이현지 [3*], 김도연 [4*]
*공동 주저자

[1] 서울과학기술대학교 글로벌융합산업공학과 학부생
[2] 한국공학대학교 IT 경영학과 학부생
[3] 고려대학교(세종) 빅데이터사이언스학부 학부생
[4] 명지대학교 데이터테크놀로지학과 학부생
shawn225@naver.com, kimnh097@naver.com, 0105lhl@korea.ac.kr, ehdusrla812@naver.com

Combining AutoML and XAI: Automating machine learning models and improving interpretability

Min Hyeok Son[1*], Nam Hun Kim[2*], Hyeon Ji Lee[3*], Do Yeon Kim[4*]
[1]Dept. of Industrial engineering, SeoulTech
[2]Dept. of IT management, Tech University of Korea
[3]Dept. of Big Data Science, Korea University Sejong Campus
[4]Dept. of Data Technology, Myungji University

요 약

본 연구는 최근 기계학습 모델의 복잡성 증가와 '블랙 박스'로 인식된 머신러닝 모델의 해석 문제에 주목하였다. 이를 해결하기 위해, AutoML 기술을 사용하여 효율적으로 최적의 모델을 탐색하고, XAI 기법을 도입하여 모델의 예측 과정에 대한 투명성을 확보하려 하였다. XAI 기법을 도입한 방식은 전통적인 방법에 비해 뛰어난 해석력을 제공하며, 사용자가 머신러닝 모델의 예측 근거와 그 타당성을 명확히 이해할 수 있음을 확인하였다.

1. 서론

현대의 기계학습 모델은 복잡도가 급증하며 다양한 알고리즘이 등장하고 있다. 이런 배경 속에서 AutoML 라이브러리가 모든 알고리즘을 포괄하는 것은 어려워지고 있다. 예를 들면, pycaret 은 M/L 에, Auto-Keras 는 딥러닝에 특화되어 있다. 모델의 예측력은 물론, 변수 중요도와 같은 해석력도 중요하다. 특히 딥러닝은 '블랙 박스'로 인식되기 쉽다.

이 문제를 해결하기 위해, 확장 가능한 AutoML 과 XAI 가 결합될 필요가 있다. 본 연구는 AutoML 에 SHAP, LIME 과 같은 XAI 해석 기법을 통합, 사용자가 모델의 효율과 투명성을 동시에 향상시키는 방안을 제시한다. 이 통합 솔루션은 다양한 알고리즘에 적용 가능하며, 신뢰성 있는 의사결정을 지원할 것으로 예상된다.

2. 본론

2.1 관련 연구

AutoML: 'Automated Machine Learning'을 지칭하며, 머신러닝 과정을 자동화하는 기술이다. 데이터 처리부터 모델 최적화까지를 포괄한다. 최근에는 주요 연구 분야로 각광받고 있으며, 많은 기업들이 투자하고 있다.

XAI: 'Explainable Artificial Intelligence'로, 머신러닝 모델의 결정 과정을 해석 가능하게 하는 기술이다. 특히 딥러닝 같은 복잡한 모델의 'Black Box' 성격을 해석하는 데 중요하다. 특히 SHAP[1]은 설명가능한 인공지능(XAI) 방법 중 하나로, 게임이론 기반으로 모델 예측을 설명하는 방식이다. 플레이어의 기여도 분석 개념을 변수의 기여도 평가에 활용한다.

2.2 실험 및 평가

본 연구는 AutoML 기술을 적용하여 최적의 예측 모델을 자동 탐색하였다. AutoML 을 통해 최적의 모델을 선별한 후, XAI 기법을 활용하여 모델의 예측 기여도를 분석하고 시각화하였다. 이를 통해 중요한 피처와 그들의 영향력을 정밀하게 파악하였다.

(a) AutoML 을 통한 최적의 모델 선정

본 연구에서는 AutoML 을 이용해 최적의 예측 모델을 찾았다. 실험 과정에서 오픈소스 라이브러리인 pycaret 이라는 AutoML 툴을 사용하였으며, 사용한 데이터는 'Boston House Price

Dataset'[2]으로 보스턴의 집값을 여러 피처들을 통해 예측하는 데이터이다.

(b) 기본적인 변수와 데이터 관련 시각화 툴 대비 XAI 기법을 활용한 설명력 향상

모델을 선택한 후, 초기 단계에서는 전통적인 Feature Importance 방법을 이용하여 변수들의 중요도를 확인했다. 그러나, 이 방법만으로는 복잡한 모델의 예측 과정을 완전히 이해하는 데에 한계가 있었다.

따라서 본 연구는 더 상세한 특징 중요도 파악을 위해 SHAP 를 활용하였다. [그림 1]은 SHAP 의 summary_plot 을 나타낸다. 이 플롯의 각 점은 피처의 SHAP 값을 표현하며, 세로축은 피처들을 나열한다. 점의 색상은 피처 값의 크기를, 위치는 SHAP 값의 크기와 방향을 나타낸다. 이 플롯을 통해 어떤 피처가 예측에 큰 영향을 미치는 지와 그 영향의 방향을 쉽게 파악할 수 있다.

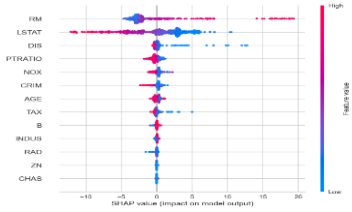

그림 1

그리고 변수가 예측에 미치는 영향을 구체적으로 파악하기 위해 PDP, SHAP 를 이용했다. [그림 2]는 중요한 피처의 기여도를 상세히 보여준다. dependence_plot 은 피처의 값과 그 SHAP 값을 관계화하여 시각화한다. x 축은 피처 값, y 축은 SHAP 값으로, 피처 값의 변화에 따른 SHAP 값의 변화를 나타낸다. 이로써 해당 피처 값이 모델 예측에 미치는 영향을 쉽게 파악할 수 있다.

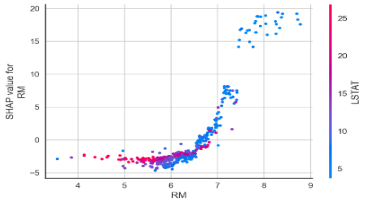

그림 2

[그림 3]에서 보여주는 PDP 는 피처의 특정 값들에 대한 예측결과와 평균적인 변화를 나타낸다. 이로써 각 변수가 모델에 어떻게 영향을 미치는지 알 수 있다.

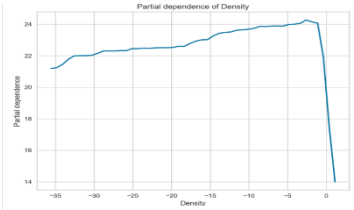

그림 3

그 후에는 지역적 변수의 분석을 ICE 와 LIME 기법을 통해 진행한다. 먼저, ICE 기법을 사용하여 모델의 작동 방식을 깊게 탐색한다. 개별 데이터 포인트에 대한 예측 반응을 시각적으로 표현하여, 각 데이터 포인트가 모델의 예측에 어떻게 기여하는지를 보여준다. 이로써, 동일한 피처 값의 다른 데이터 포인트들이 모델 내에서 어떻게 다르게 작용하는지도 파악할 수 있다.

그리고 [그림 4]에서 LIME 기법을 활용하여 지역적 변수 분석을 진행한다. 이 기법은 각 데이터 포인트 주변의 로컬 영역에서 모델을 근사화하며, 이를 통해 해당 포인트에서의 예측이 어떻게 이루어졌는지를 설명한다. 따라서 LIME 을 통해 모델의 복잡한 예측도 사용자가 이해할 수 있는 방식으로 해석될 수 있게 된다.

그림 4

3. 결론

본 연구에서는 AutoML 의 효율성을 활용하여 최적의 기계학습 모델을 자동 탐색하며, SHAP, LIME 과 같은 설명 가능한 인공지능(XAI) 기술을 도입하여 모델의 예측 결과에 대한 투명성과 해석력을 강화하였다. 이를 통해, Pycaret 라이브러리를 활용하여 사용자의 개입을 최소화하면서도 다양한 머신러닝 결과를 효율적으로 얻을 수 있게 되었다. 또한, XAI 알고리즘의 도입으로 모델의 예측 근거와 그 타당성을 사용자에게 명확하게 제시할 수 있게 되었다. 이러한 접근 방식은 머신러닝 모델의 효율성과 투명성을 동시에 실현하는 데 크게 기여할 것으로 기대된다.

사사표기

본 논문은 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT 멘토링 프로젝트 결과물입니다.

참고문헌

[1] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems* 30 (2017).

[2] 'Boston House Price Dataset'. "Kaggle", https://www.kaggle.com/datasets/fedesoriano/the-boston-houseprice-data