XAI
(eXplainable Artificial Intelligence)

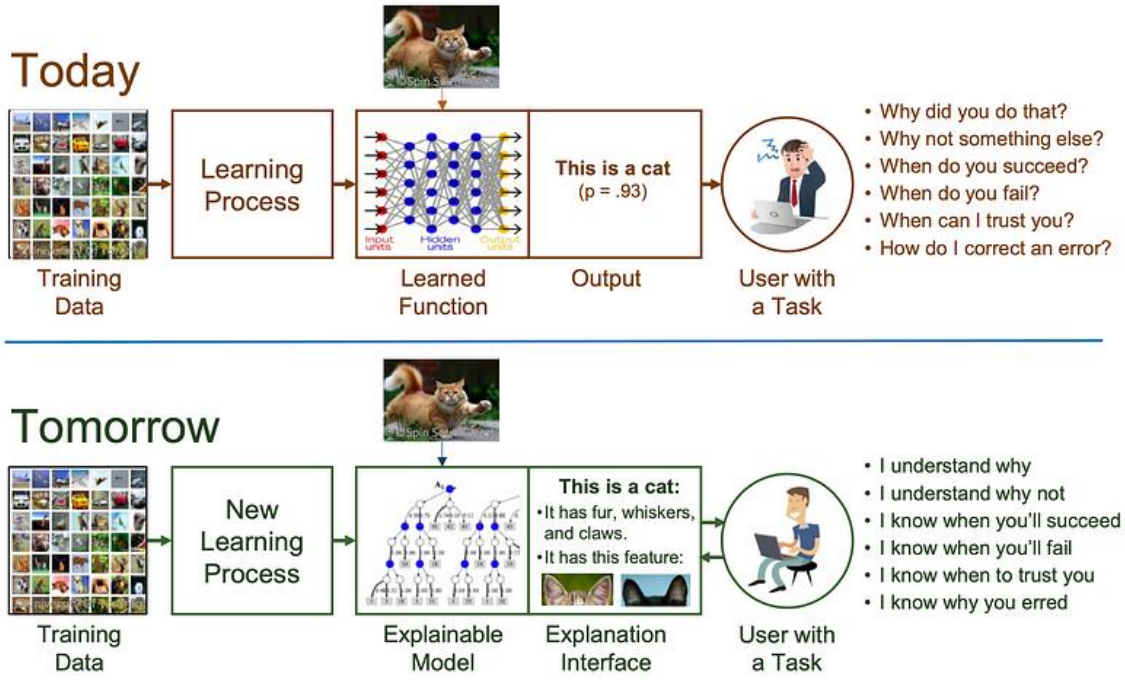
MILab Undergraduate student, Kim Taehyeon

2024. 1. 25

목차

1. XAI 란?

1. XAI 란 무엇인가?



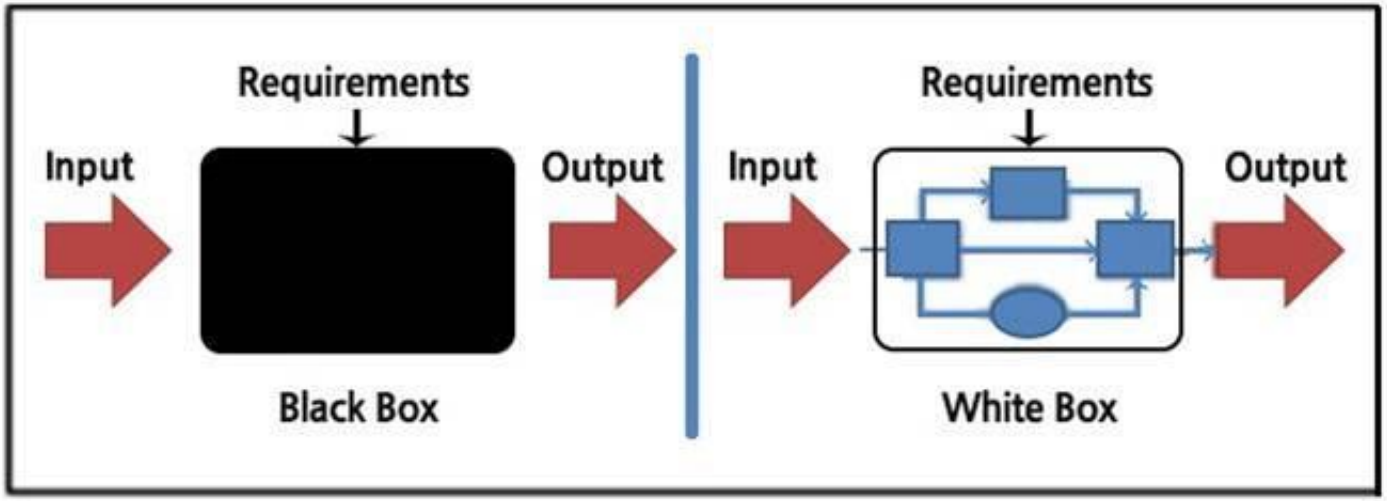
설명 가능한 AI(eXplainable AI, XAI)는

학습된 AI 모델을 바탕으로 사람이 이해할 수 있도록 설명하는 기술이다.

목차

2. XAI 기술 탄생 배경

1. 블랙박스 문제

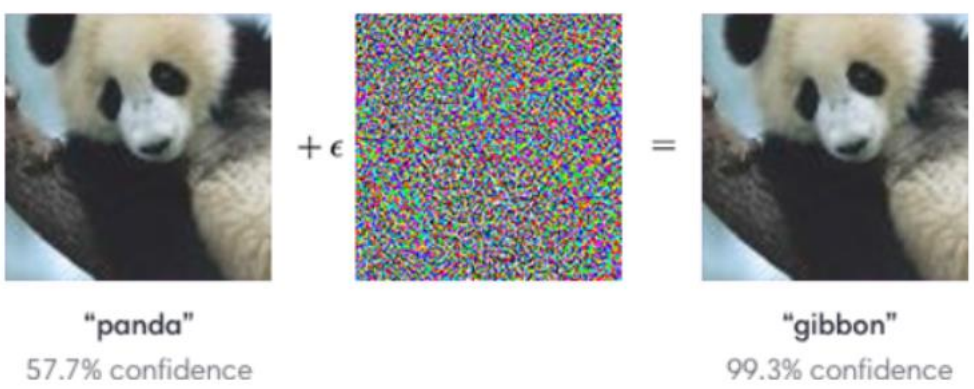


기존 DNN 등 신경망에 대한 블랙박스 모델에서의 추론 결과는

어떠한 이유로 만들어졌는지 파악하기가 어려움.

2. XAI 기술 탄생 배경

2. 적대적 공격 (Adversarial Attack) 문제



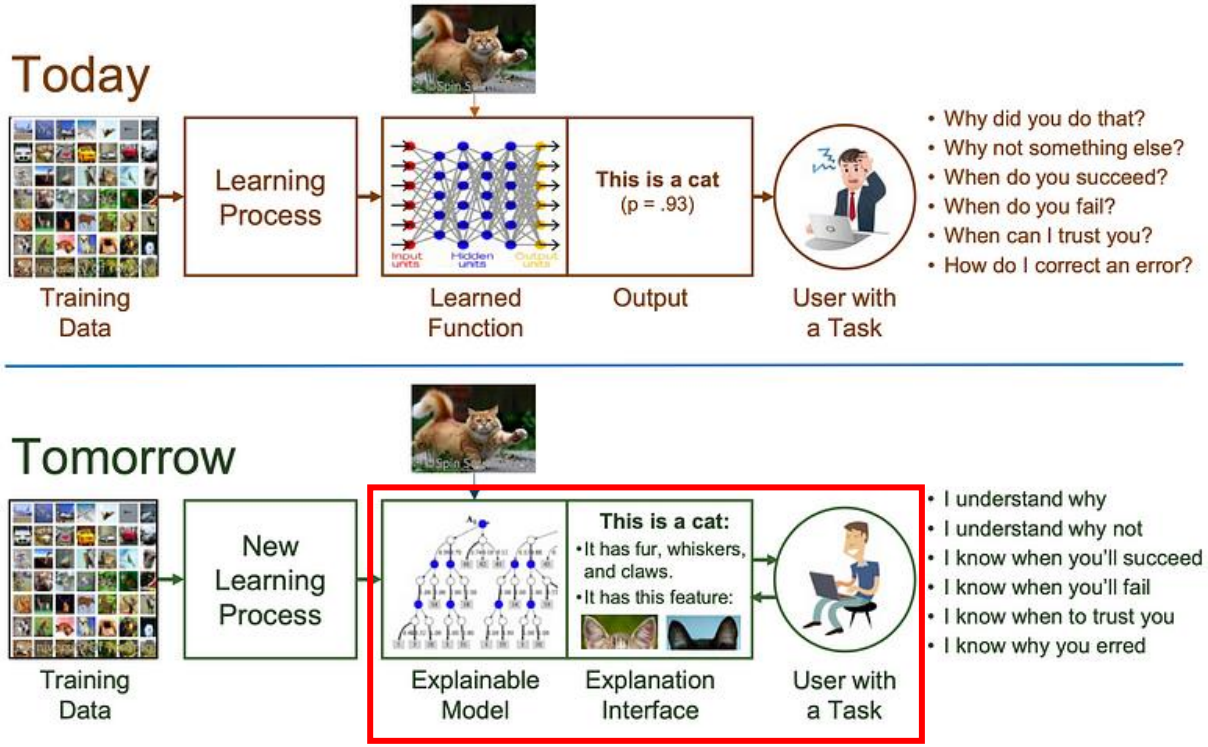
기본 CNN 등 블랙박스 모델의 경우, 모델의 출력 결과값에 대한 의문을 품을 수 있으며, 인간이 인지하지 못하는 노이즈를 추가하는 방식인 적대적 공격 (Adversarial Attack) 을 통해 다른 출력을 유도할 수 있음.

→ 이러한 상황에서 **AI 의 신뢰성을 높이는 것이 주 목적**

목차

3. XAI 기술 설명

1. XAI – 설명 모델 및 설명 인터페이스 적용

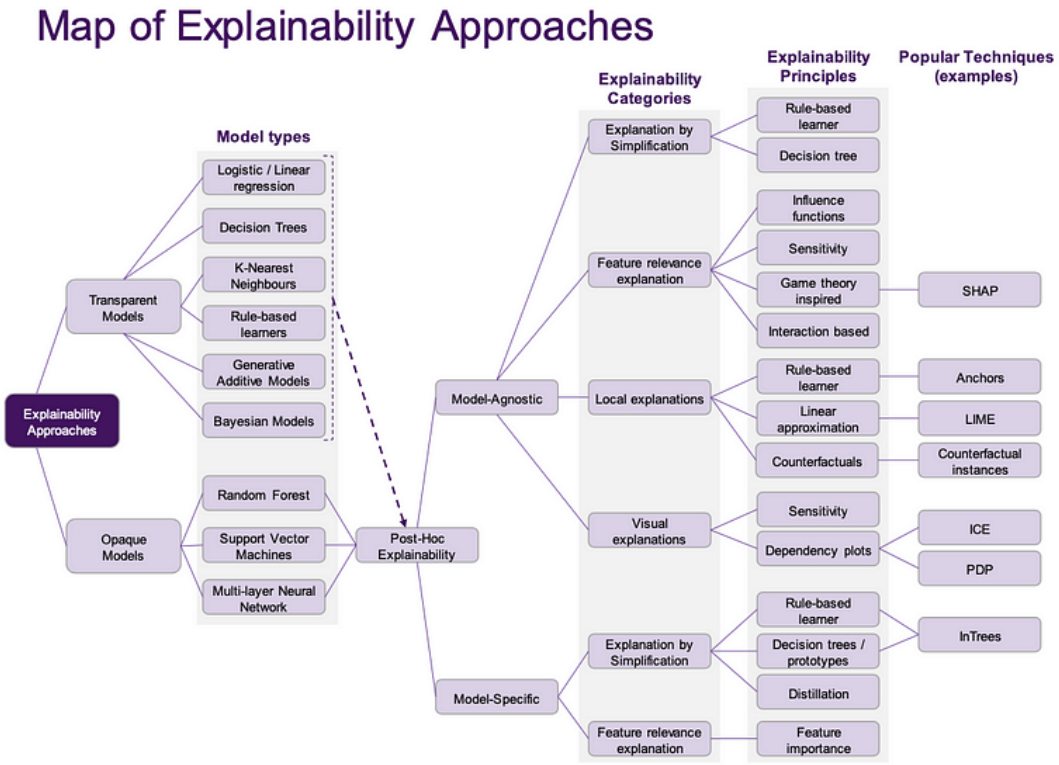


설명 가능한 직관적인 모델을 적용하여 **추론 및 검증 결과 뿐만 아닌**,

추론 결과를 도출하기까지의 **논리적인 구조를 설명 인터페이스로 제공**한다.

3. XAI 기술 설명

2. XAI Approaches – Taxonomic Framework

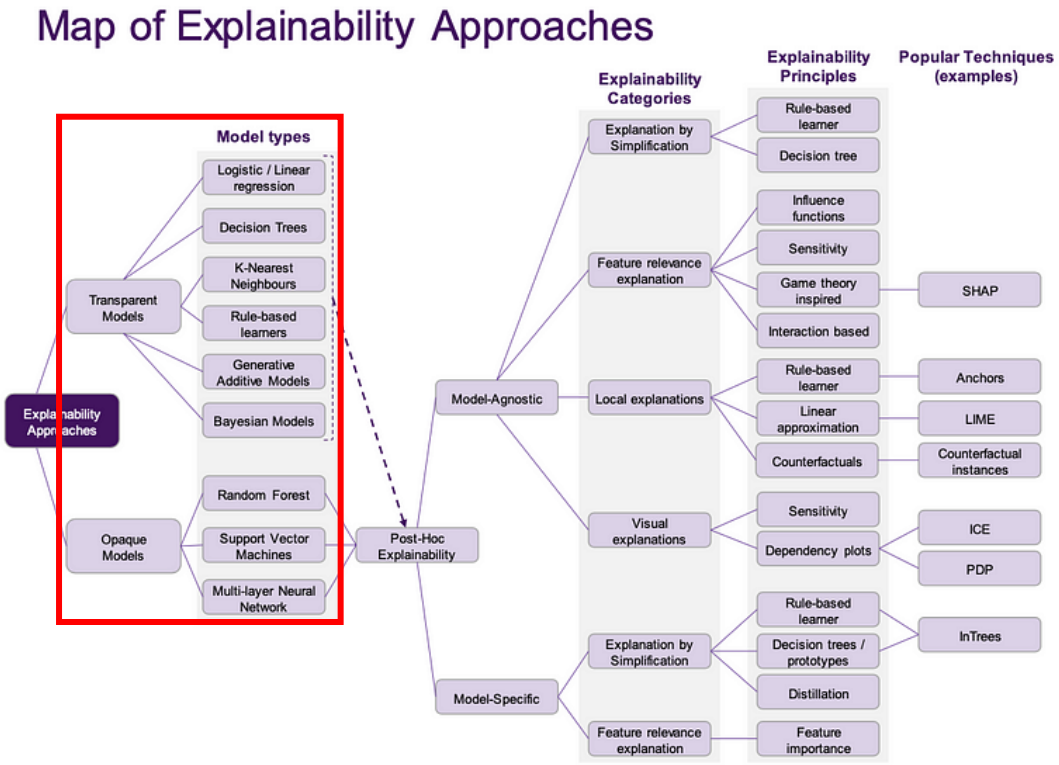


XAI 방법론 중에 하나인 **Taxonomic Framework[1]** 를 통해 전반적인

XAI 의 구성을 살펴보고자 한다.

3. XAI 기술 설명

3. XAI Approaches – Transparent vs Opaque

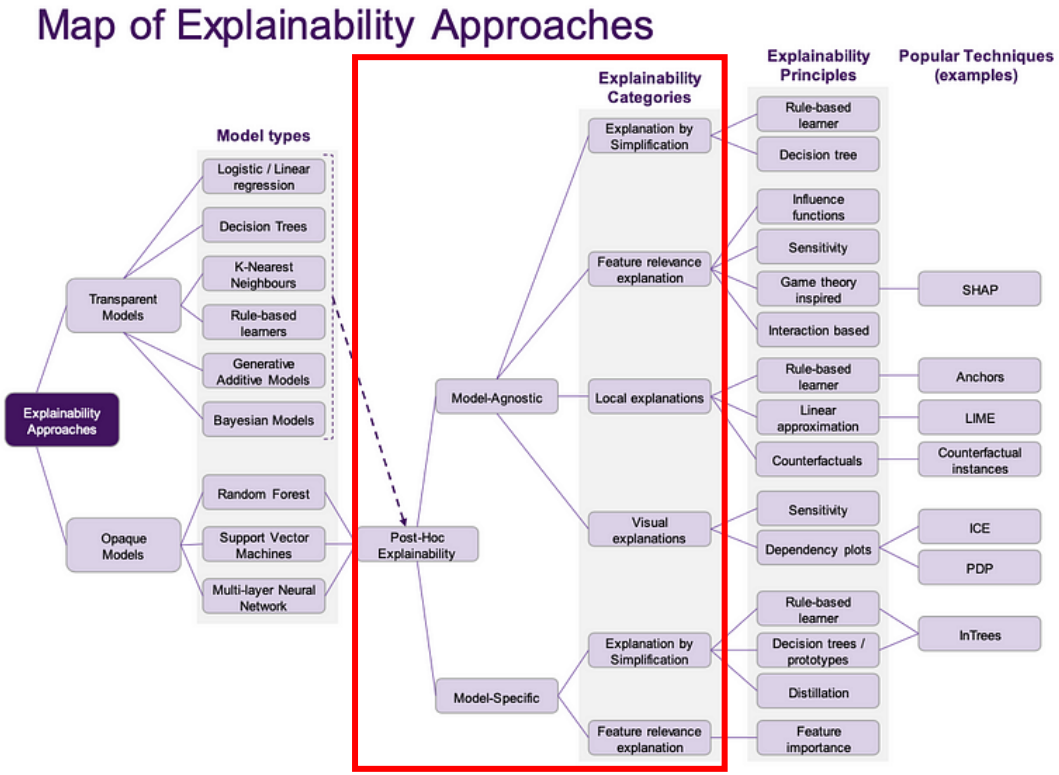


크게 Transparent Models (단순) 과 Opaque Models (복잡) 로 나눔.

모델의 복잡성과 설명력은 Trade-Off 관계이다. 적절한 균형이 필요함.

3. XAI 기술 설명

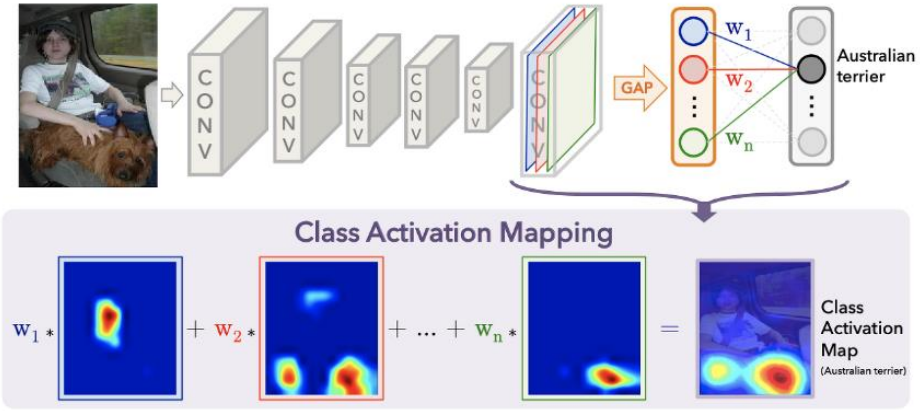
4. XAI Approaches – Model-Specific vs Model-Agnostic

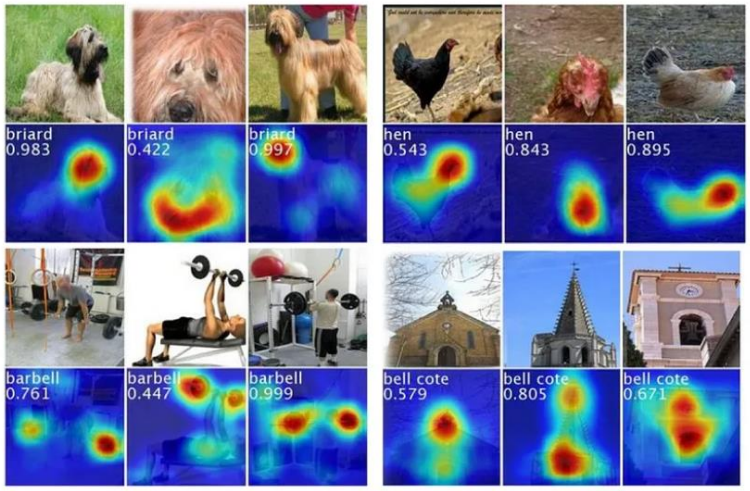


사후 설명력 기법(Post-hoc Explainability)에서 **모델의 본질적인 구조를 이용하여 설명력을 제공하는 Model-specific** 과 **범용적인 Model-agnostic 기법**이 있다.

3. XAI 기술 설명

5. XAI Approaches – Model-Specific

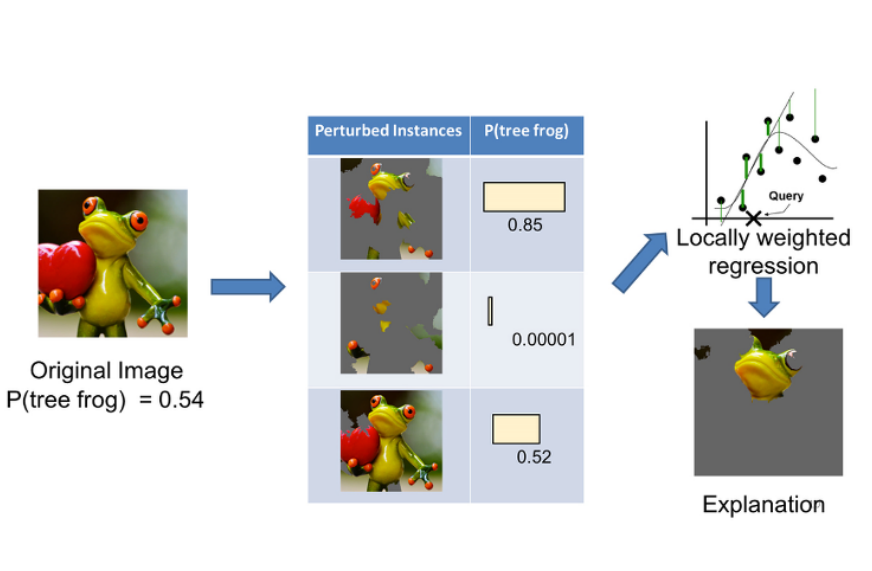


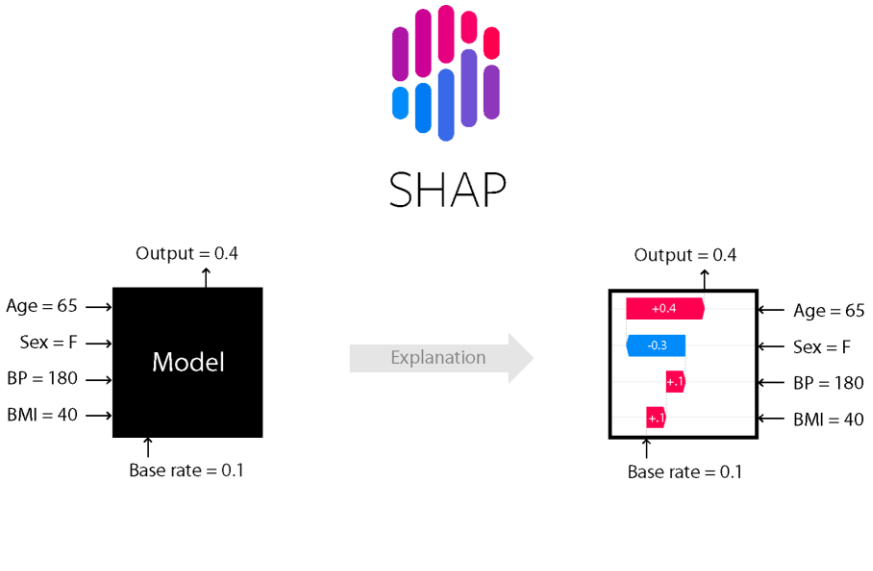


모델 구조를 바탕으로 설명 불가능한 CNN 에는 **CAM(Class Activation Mapping) 기법을 적용[2]**하여 설명력을 부여한다. **기존 Fully Connected Layer 방식이 아닌 GAP(Global Average Pooling) 을 통해 위치에 대한 새로운 Weight 를 추가**한다.

3. XAI 기술 설명

6. XAI Approaches – Model-agnostic

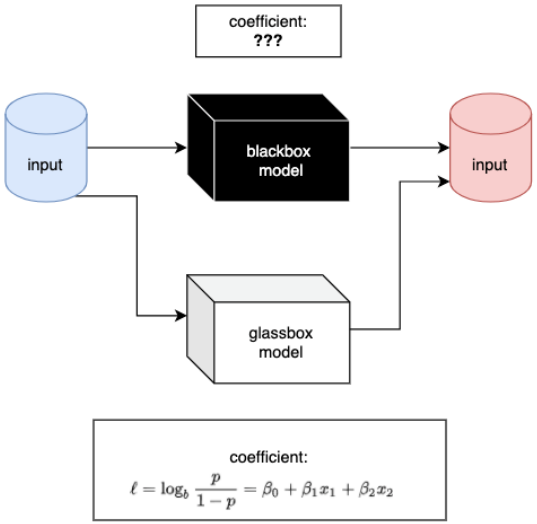




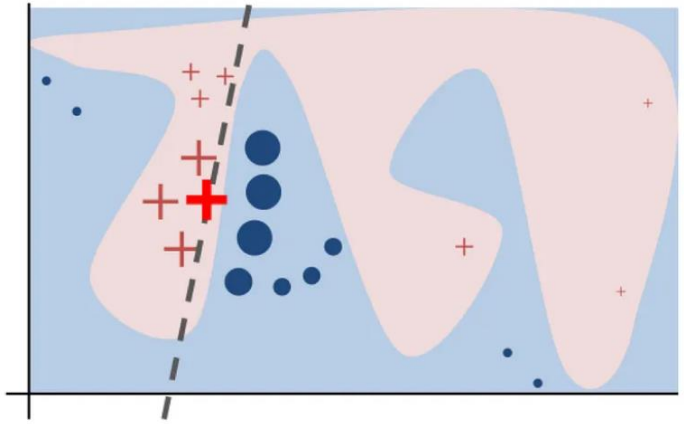
Model-agnostic 기법은 일반적으로 post-hoc (사후검증) 하며, 모든 알고리즘에 적용할 수 있는 장점이 있다. 대표적인 XAI 기법은 PDP, ICE, **LIME**[3], SHAP 이 있다.

3. XAI 기술 설명

7. LIME (Local Interpretable Model-agnostic Explanation)



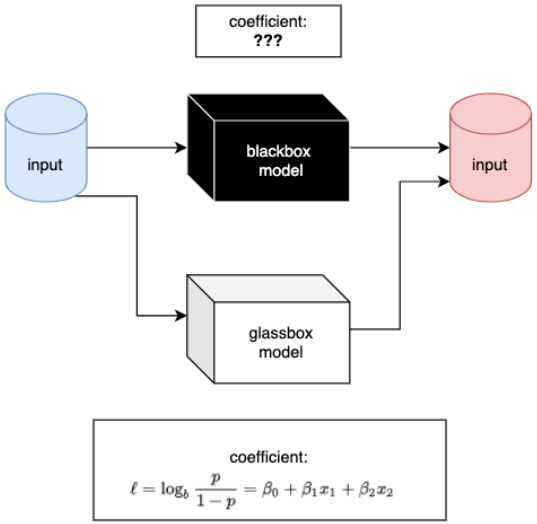
$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g)$$



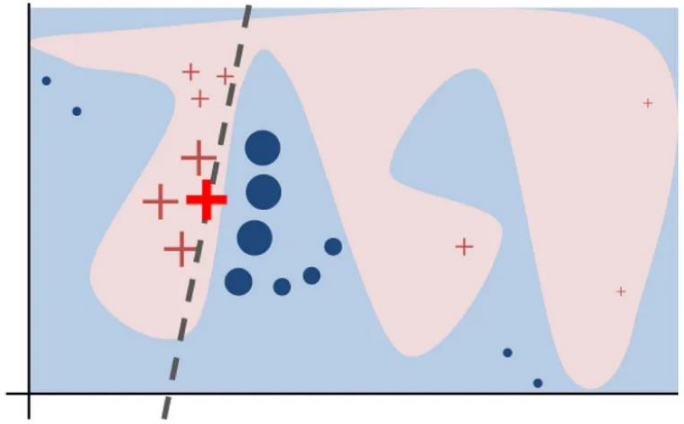
XAI 에서의 대리 분석(Surrogate Analysis)은 설명하고자 하는 **원래 모델이 지나치게 복잡해서 해석하기 어려울 때, 해석 가능한 대리 모델(Surrogate Model)을 사용하여 기존의 모델을 해석**하는 기법이다. SVM 모델 (Opaque) 에 대하여 간단하지만 설명 가능성이 높은 Logistic Regression 모델(Transparent)을 대리 모델로 사용하여 해당 모델의 계수를 기반으로 모델의 판단 메커니즘을 어림짐작한다.

3. XAI 기술 설명

7. LIME (Local Interpretable Model-agnostic Explanation)



$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g)$$



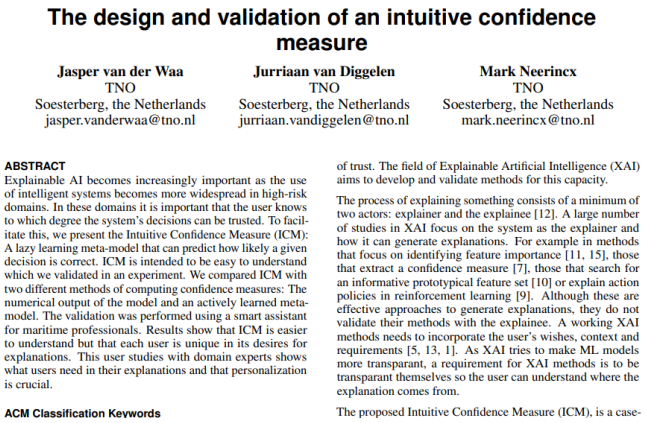
LIME[7]은 **특정 예측 값 근처에서의 지역적 해석력을 도출**하는 기법이다.

앞서 말한 대리분석에서의 Local Surrogate (로컬 대리분석)이라고도 한다.

(f : 설명하고자 하는 모델 / g, G : Interpretable model / Ω = complexity of model / L = f 에 대한 g 의 설명력 measure (클수록 bad))

3. XAI 기술 설명

8. XAI 기술의 한계



The design and validation of an intuitive confidence measure

Jasper van der Waa
TNO
Soesterberg, the Netherlands
jasper.vanderwaa@tno.nl

Jurriaan van Diggelen
TNO
Soesterberg, the Netherlands
jurriaan.vandiggelen@tno.nl

Mark Neerincx
TNO
Soesterberg, the Netherlands
mark.neerincx@tno.nl

ABSTRACT
Explainable AI becomes increasingly important as the use of intelligent systems becomes more widespread in high-risk domains. In these domains it is important that the user knows to which degree the system's decisions can be trusted. To facilitate this, we present the Intuitive Confidence Measure (ICM): A lazy learning meta-model that can predict how likely a given decision is correct. ICM is intended to be easy to understand which we validated in an experiment. We compared ICM with two different methods of computing confidence measures: The numerical output of the model and an actively learned meta-model. The validation was performed using a smart assistant for maritime professionals. Results show that ICM is easier to understand but that each user is unique in its desires for explanations. This user studies with domain experts shows what users need in their explanations and that personalization is crucial.

ACM Classification Keywords

of trust. The field of Explainable Artificial Intelligence (XAI) aims to develop and validate methods for this capacity.

The process of explaining something consists of a minimum of two actors: explainer and the explainee [12]. A large number of studies in XAI focus on the system as the explainer and how it can generate explanations. For example in methods that focus on identifying feature importance [11, 15], those that extract a confidence measure [7], those that search for an informative prototypical feature set [10] or explain action policies in reinforcement learning [9]. Although these are effective approaches to generate explanations, they do not validate their methods with the explainee. A working XAI methods needs to incorporate the user's wishes, context and requirements [5, 13, 1]. As XAI tries to make ML models more transparant, a requirement for XAI methods is to be transparant themselves so the user can understand where the explanation comes from.

The proposed Intuitive Confidence Measure (ICM), is a case-

Open Access Article

Evaluation Metrics Research for Explainable Artificial Intelligence Global Methods Using Synthetic Data

by Alexandr Oblizanov [1], Natalya Shevskaya [1], Anatoliy Kazak [2,*], Marina Rudenko [3] and Anna Dorofeeva [2]

[1] Faculty of Computer Science and Technology, Saint Petersburg Electrotechnical University Leti, 197376 Saint-Petersburg, Russia
[2] Humanitarian Pedagogical Academy, V.I. Vernadsky Crimean Federal University, 295007 Simferopol, Russia
[3] Institute of Physics and Technology, V.I., Vernadsky Crimean Federal University, 295007 Simferopol, Russia
* Author to whom correspondence should be addressed.

Appl. Syst. Innov. **2023**, *6*(1), 26; **https://doi.org/10.3390/asi6010026**

Submission received: 3 January 2023 / Revised: 16 January 2023 / Accepted: 5 February 2023 / Published: 9 February 2023

Download Browse Figures Versions Notes

설명 가능력을 높이기 위한 XAI 기법들 중 **절대적인 성능 지표가 없다**.

설명력의 질을 파악하는 것이 어렵다는 것이 그 이유인데, 이러한 상황에 대하여 XAI 자체의 신뢰도를 높이고 비교분석을 진행하는 연구가 진행되고 있다.

이러한 어려움에도 불구하고, AI 의 단점을 보완해줄 수 있는 기술이라고 생각한다.

목차

4. Related Works

1. Current Related Works


금융·핀테크

카카오뱅크, 카이스트와 금융 분야 설명가능 인공지능 (XAI) 공동 연구

설명가능 인공지능 기술 역량 내재화 선도

강진규 기자 입력 2023.11.15 09:33

댓글 0

가

kakaobank | KAIST

금융 분야 설명가능 인공지능 공동 연구

[HTML] Dermatologist-like explainable AI enhances trust and **confidence** in diagnosing melanoma

T Chanda, K Hauser, S Hobelsberger… - Nature …, 2024 - nature.com

… between **XAI** and dermatologist explanations. We also show that dermatologists' **confidence** in their diagnoses, and their trust in the support system significantly increase with **XAI** …

☆ 저장 인용 4회 인용 관련 학술자료 전체 2개의 버전

[HTML] Improving trust and **confidence** in medical skin lesion diagnosis through explainable deep learning

C Metta, A Beretta, R Guidotti, Y Yin, P Gallinari… - International Journal of …, 2023 - Springer

… However, often **XAI** approaches are only tested on … **confidence** of users towards automatic AI decision systems in the field of medical skin lesion diagnosis by customizing an existing **XAI** …

☆ 저장 인용 관련 학술자료

Improving deep neural network classification **confidence** using heatmap-based eXplainable AI

E Tjoa, HJ Khok, T Chouhan, G Cuntai - arXiv preprint arXiv:2201.00009, 2021 - arxiv.org

… We empirically show existing **XAI** methods have the potential to improve classification **confidence**. … for AX process on existing **XAI** methods, formal definition GAX process and the results. …

☆ 저장 인용 4회 인용 관련 학술자료 전체 2개의 버전

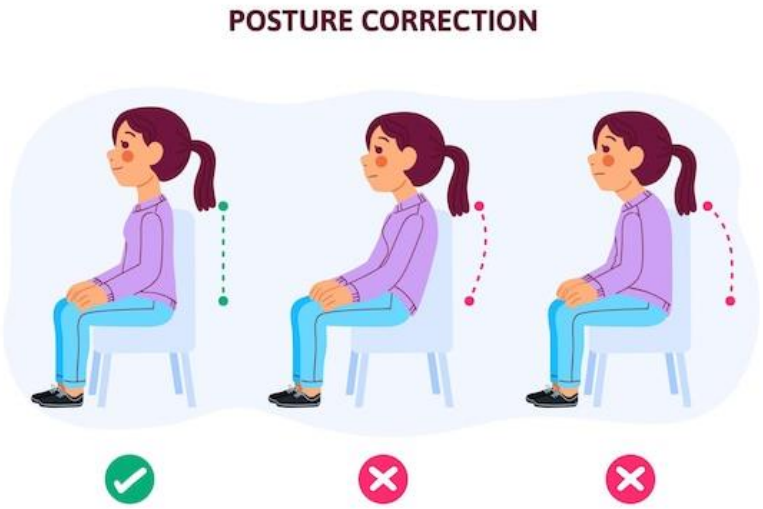
최근에 **금융권에서의 XAI 적용**에 대한 연구 시도를 계속해서 진행 중이다.

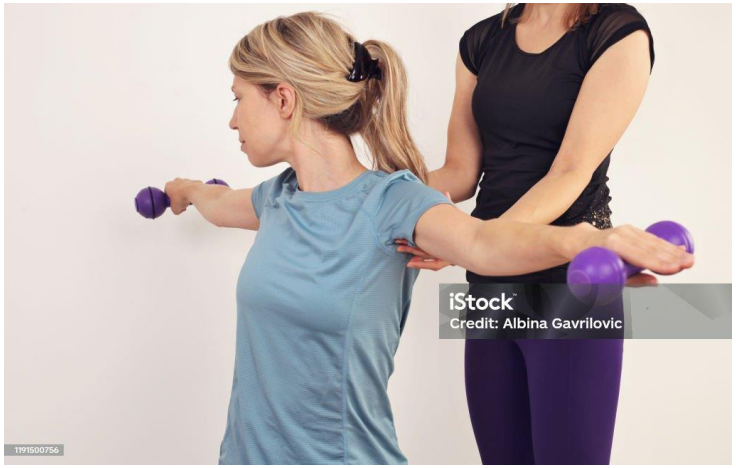
또한, **피부과의 XAI 적용** 및 **기존 AI 비전 인식 시스템에 제공할 수 있는 기능**으로 도입이 되고 있는 상황이다.

앞으로의 **XAI 적용이 점차 확장되고 있는 방향**으로 움직이고 있다.

4. Related Works

2. Expected Applications





보조해줄 수 있는 AI 기능 + 티칭 알고리즘 등

다양한 적용을 기대해볼 수 있다.

단순히 맞았다/틀렸다가 아닌, 더 나은 방향성으로 제시가 가능해짐.

Thank you!

MILab Undergraduate student, Kim Taehyeon

2024. 1. 25

Reference

<websites>

- https://www.incodom.kr/XAI%28explaniable_AI%29
- https://bitnine.tistory.com/408
- https://medium.com/daria-blog/%EC%84%A4%EB%AA%85%EA%B0%80%EB%8A%A5%ED%95%9C-%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5-explanable-ai-xai-%EC%9D%B4%EB%9E%80-4e51d9e7bb59
- https://wikibook.co.kr/xai/
- https://github.com/karthiksekaran/logistics-cost-prediction-XAI/blob/main/Logistics-Cost-XAI.pdf
- https://velog.io/@sjinu/%EB%85%BC%EB%AC%B8%EB%A6%AC%EB%B7%B0-Explainable-Artificial-Intelligence-XAI-An-Engineering-Perspective
- https://korea-sw-eng.blogspot.com/2017/04/1.html
- https://www.researchgate.net/profile/Jasper_Waa/publication/323772259_The_design_and_validation_of_an_intuitive_confidence_measure/links/5aaa2ba0aca272d39cd64796/The-design-and-validation-of-an-intuitive-confidence-measure.pdf
- https://www.mdpi.com/2571-5577/6/1/26

<papers>

- *[1] V. Belle and I. Papantonis. Principles and practice of explainable machine learning.* arXiv preprint arXiv:2009.11698, 2020.
- [2] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba, Learning Deep Features for Discriminative Localization, https://arxiv.org/abs/1512.04150
- *[3] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the predictions of any classifier. In* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.*
- *[4] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In* Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc.*