DMQA Open Seminar

Introduction to XAI(eXplainable AI)

2021.09.03

Data Mining & Quality Analytics Lab.

오혜령

발표자 소개

❖ 오혜령

- 고려대학교 산업경영공학과
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- M.S. Student (2021.03 ~)

❖ Research Interest

- Machine Learning & Deep Learning
- Anomaly Detection

❖ Contact

- E-mail : hyeryeong5@korea.ac.kr

Contents

1. Introduction

Introduction

XAI란?

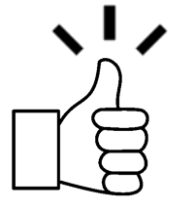
의사 A

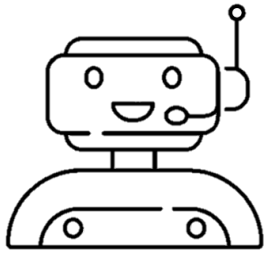
폐렴이 의심됩니다.

의사 B

좌측 상단에 뚜렷한 음영이 확인되네요.

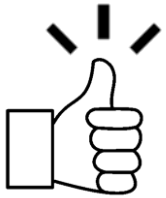
폐렴이 의심됩니다.





· 예측 결과 : 폐렴

· 해석 : 이미지 좌측 상단에 뚜렷한 음영 확인



Introduction

XAI란?

eXplainable AI (XAI)

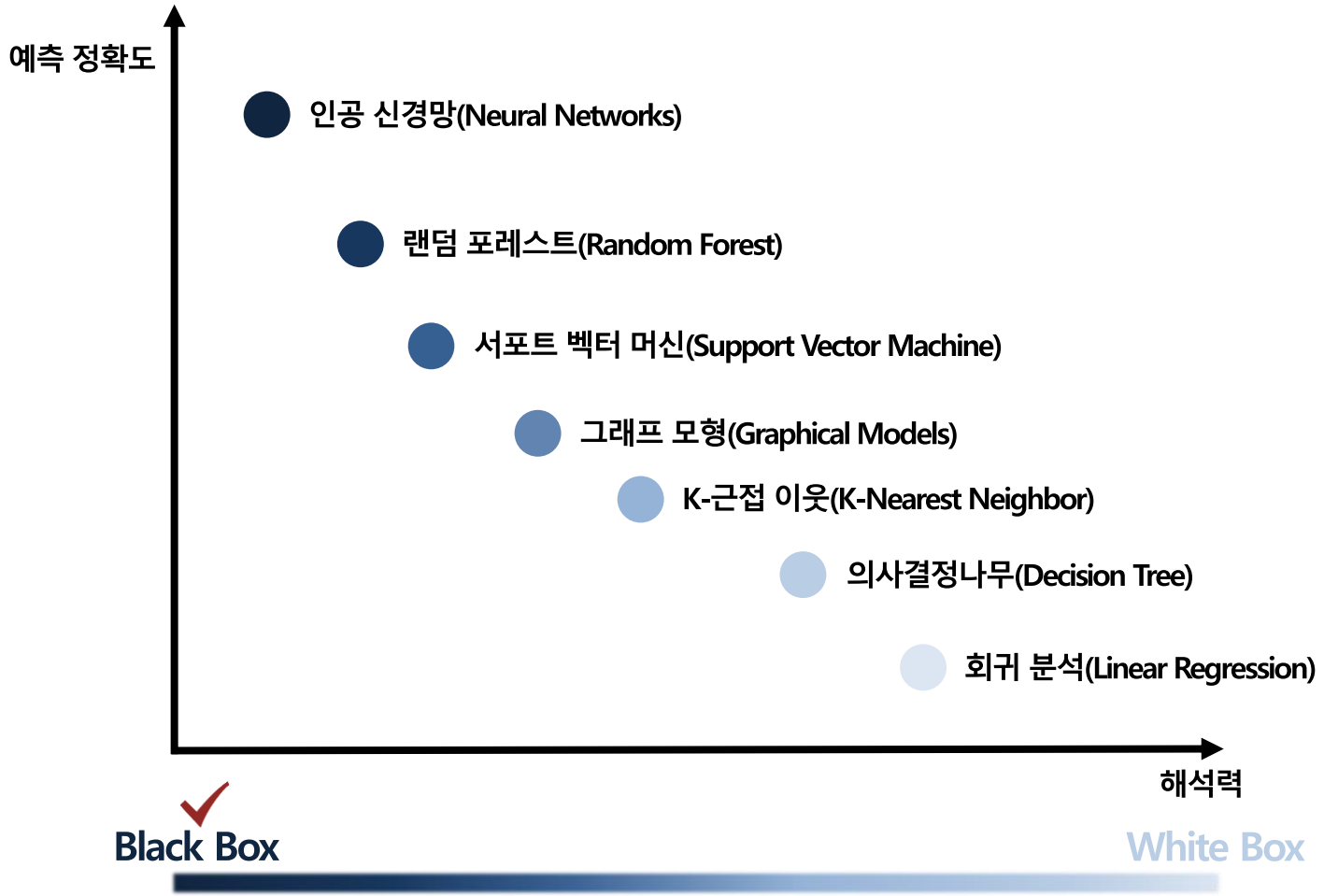
설명(해석)할 수 있는

머신 러닝 알고리즘이 예측한 결과를 사람이 이해할 수 있도록 해석을 제공하는 방법론

XAI는 해석력이 낮은 모델의 해석을 도와준다 !

Introduction

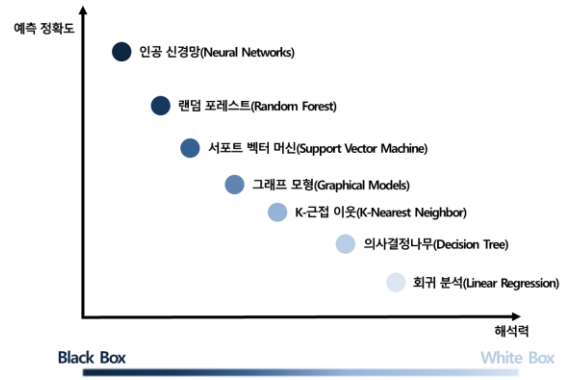
White Box / Black Box



Morocho-Cayamcela, Manuel Eugenio, Haeyoung Lee, and Wansu Lim. "Machine learning for 5G/B5G mobile and wireless communications: Potential, limitations, and future directions." IEEE Access 7 (2019): 137184-137206.

Introduction

White Box – 해석력이 높은 모델

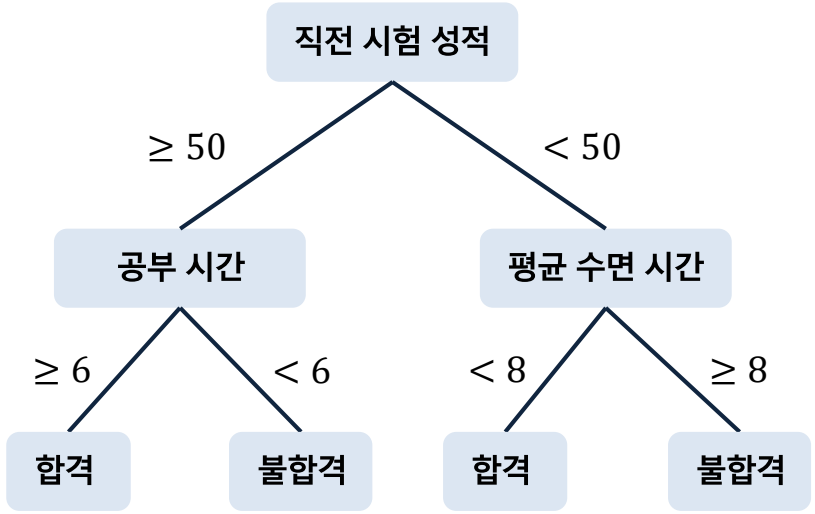


평균 수면 시간(x_1)	공부 시간(x_2)	직전 시험 성적(x_3)	합격/불합격
10	2	40	불합격
6	7	90	합격
…	…	…	…

Logistic Regression

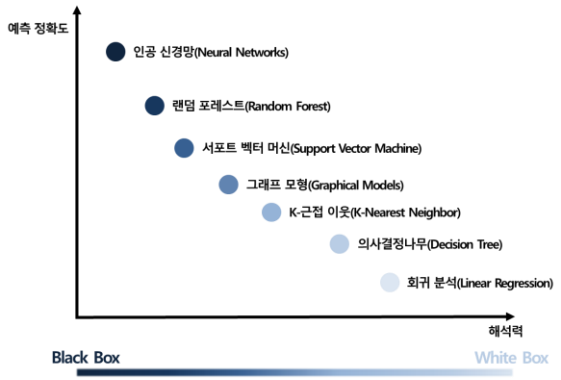
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Decision Tree

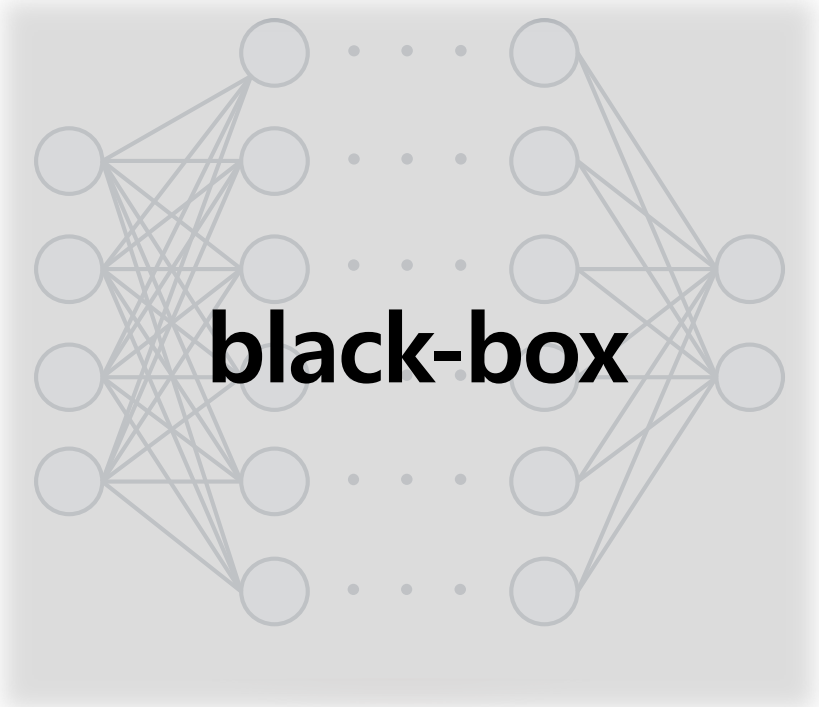


Introduction

Black Box – 해석력이 낮은 모델



Neural Networks



폐렴

Introduction

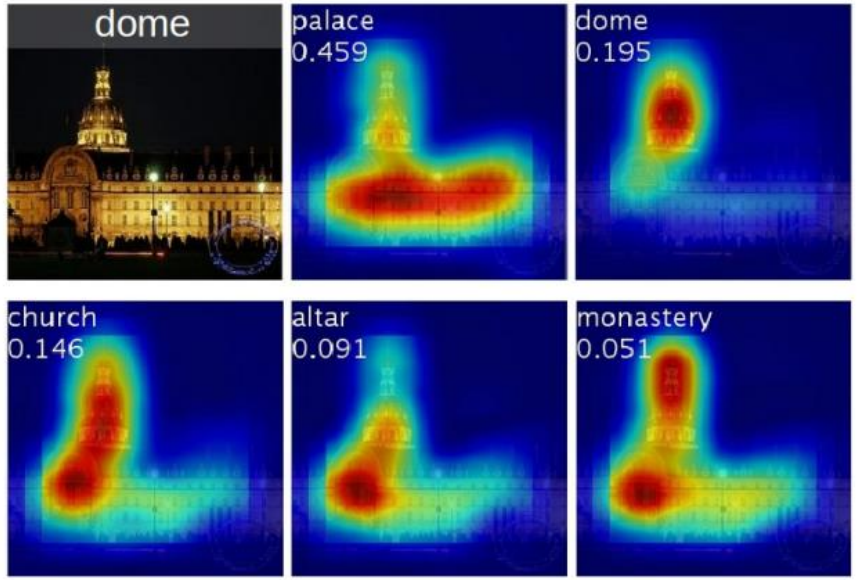
XAI Categorization

XAI

적용 가능 범위

Model-agnostic : 모델의 특성을 이용하지 않기 때문에 모델에 상관 없이 적용 가능

Model-specific : 특정 종류의 모델에만 적용 가능



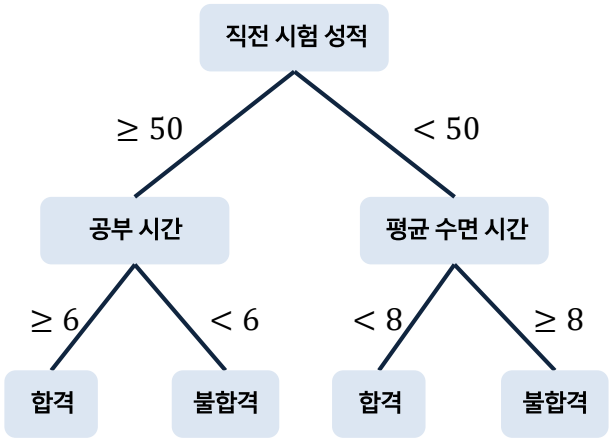
Class Activation Map(CAM)

이미지 출처 : Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2921-2929).

Introduction

XAI Categorization

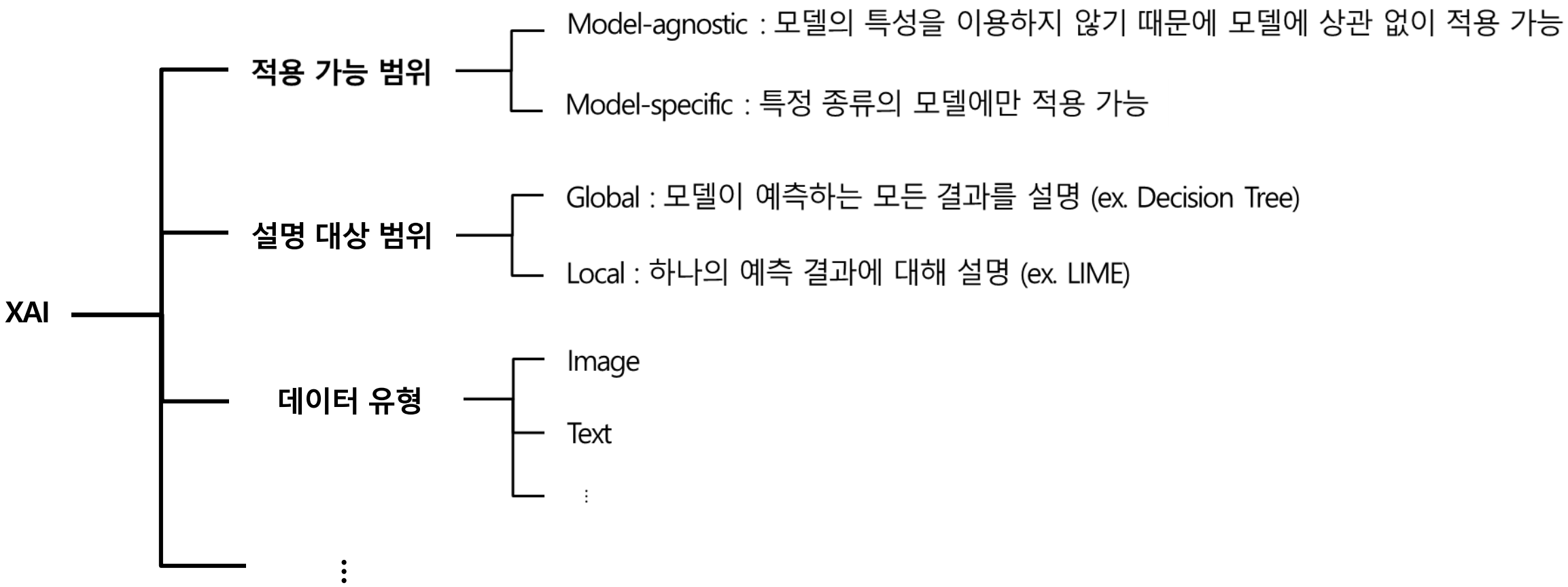
- XAI
  - 적용 가능 범위
    - Model-agnostic : 모델의 특성을 이용하지 않기 때문에 모델에 상관 없이 적용 가능
    - Model-specific : 특정 종류의 모델에만 적용 가능
  - 설명 대상 범위
    - Global : 모델이 예측하는 모든 결과를 설명 (ex. Decision Tree)
    - Local : 하나의 예측 결과에 대해 설명 (ex. LIME)
  - ⋮



Decision Tree

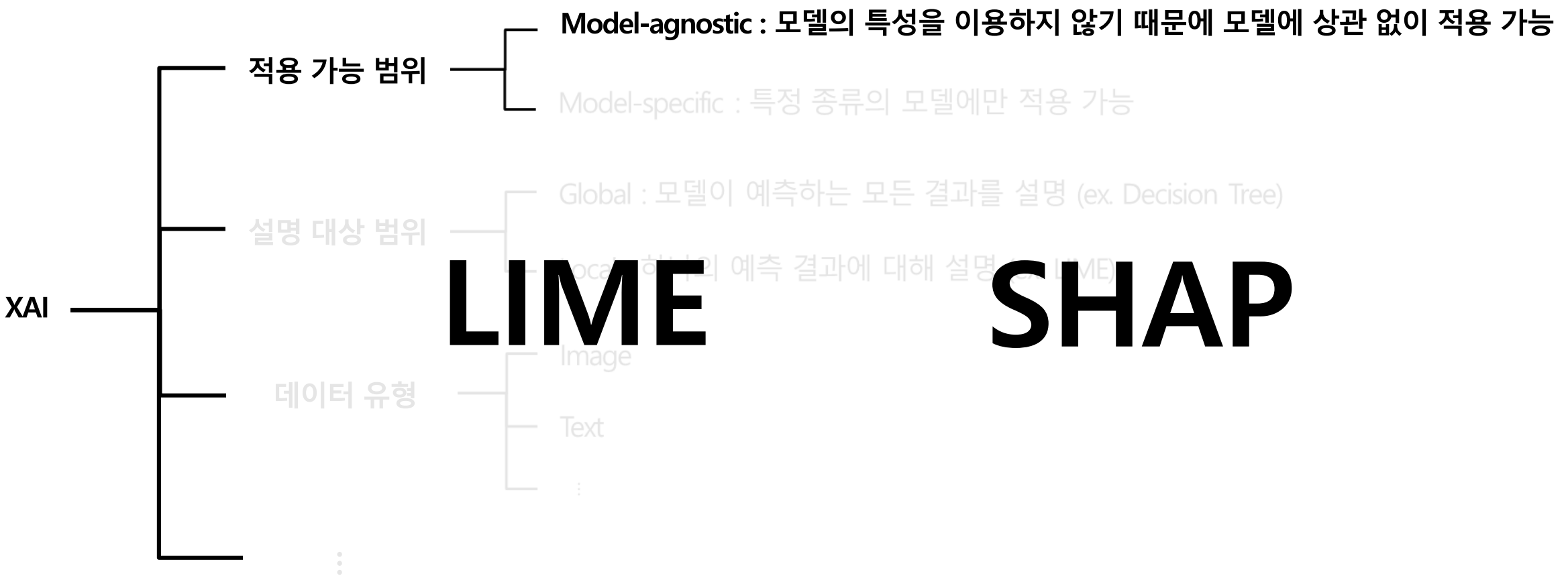
Introduction

XAI Categorization



Introduction

XAI Categorization



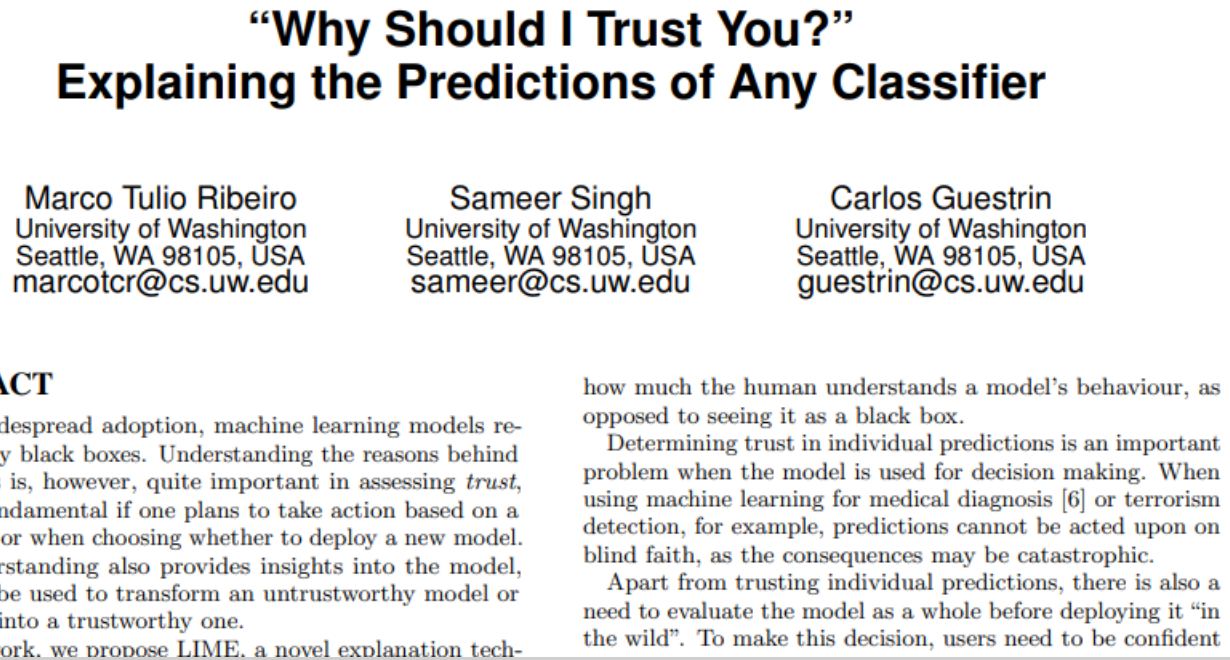
2. LIME

(Local Interpretable Model-agnostic Explanations)

LIME

❖ **"Why Should I Trust You?": Explaining the Predictions of Any Classifier**

- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).
- 2021년 8월 30일 기준 6,307회 인용
- Local approximation 생성을 통해 Black box 모델에 대한 해석 제공



"Why Should I Trust You?"
Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

ABSTRACT

Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing *trust*, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one.

In this work, we propose LIME, a novel explanation tech-

how much the human understands a model's behaviour, as opposed to seeing it as a black box.

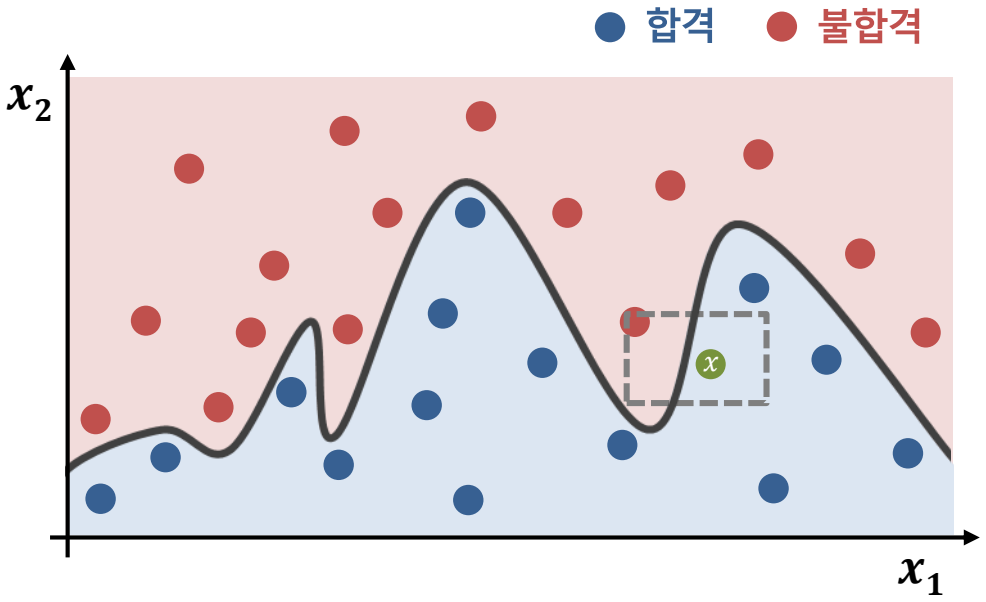
Determining trust in individual predictions is an important problem when the model is used for decision making. When using machine learning for medical diagnosis [6] or terrorism detection, for example, predictions cannot be acted upon on blind faith, as the consequences may be catastrophic.

Apart from trusting individual predictions, there is also a need to evaluate the model as a whole before deploying it "in the wild". To make this decision, users need to be confident

LIME

❖ **LIME(Local Interpretable Model-agnostic Explanations)**

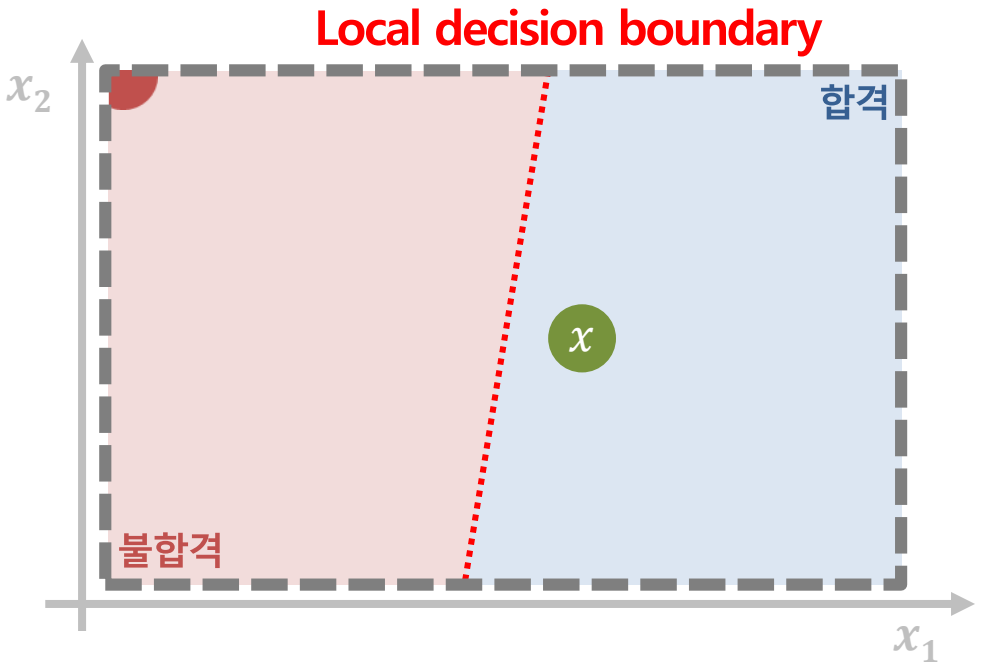
- Local Surrogate Model을 통해 관측치 하나의 예측에 대한 해석 제공
  - Local Surrogate Model : 전체 데이터셋이 아닌 특정 관측치에 대한 예측 설명 시 사용할 수 있는 해석 가능한 모델



LIME

❖ **LIME(Local Interpretable Model-agnostic Explanations)**

- Local Surrogate Model을 통해 관측치 하나의 예측에 대한 해석 제공
  - Local Surrogate Model : 전체 데이터셋이 아닌 특정 관측치에 대한 예측 설명 시 사용할 수 있는 해석 가능한 모델



관측치 x에 대한 좋은 설명 모델 g

- **해석할 수 있는 단순한 모델**
- **해석하고자 하는 모델의 예측과 유사하게 예측**

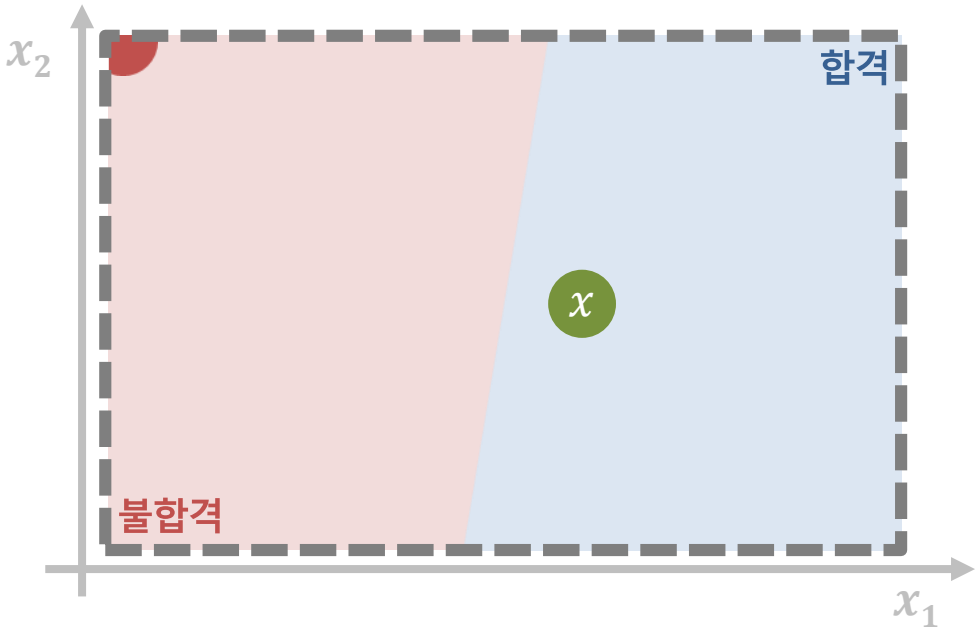
$$\xi(x) = \underset{g \in G}{argmin}\, \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

LIME

❖ LIME(Local Interpretable Model-agnostic Explanations)

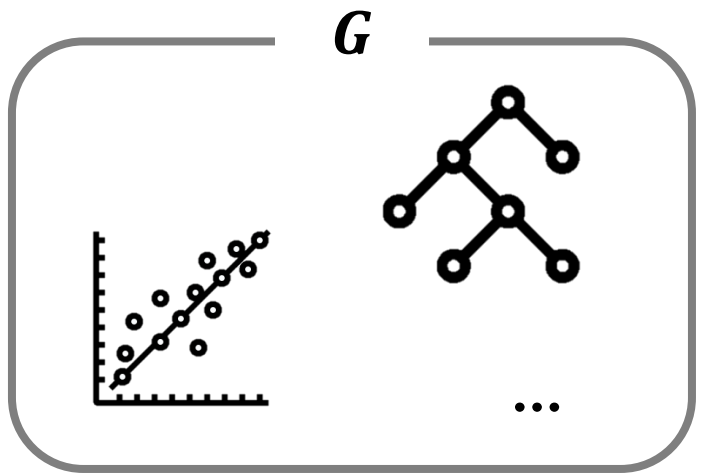
관측치 x에 대한 좋은 설명 모델 g

- 해석할 수 있는 단순한 모델
- 해석하고자 하는 모델의 예측과 유사하게 예측



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

해석력이 좋은 모델들의 집합



LIME

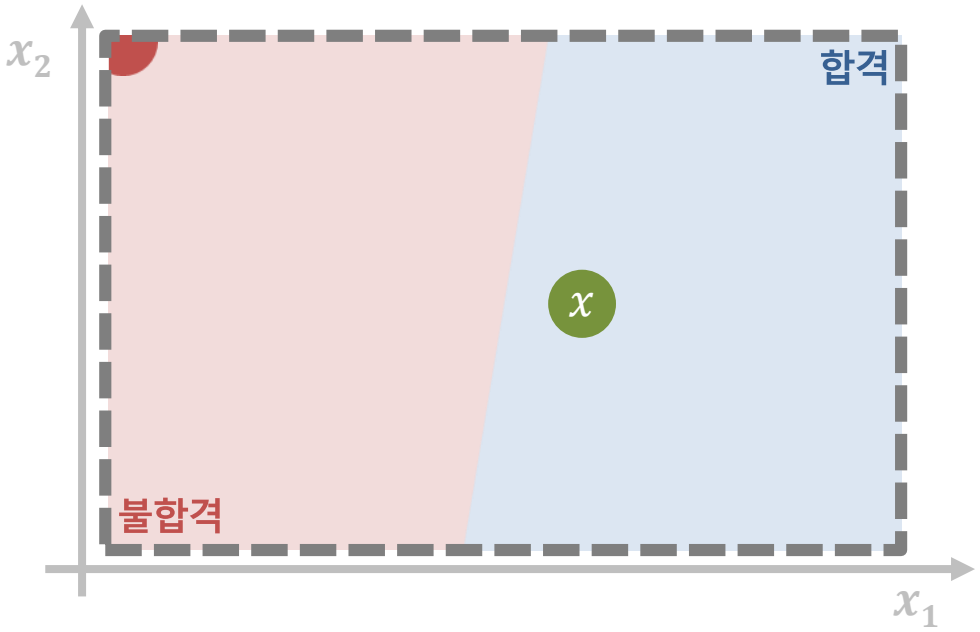
❖ LIME(Local Interpretable Model-agnostic Explanations)

관측치 x에 대한 좋은 설명 모델 g

- 해석할 수 있는 단순한 모델
- 해석하고자 하는 모델의 예측과 유사하게 예측

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

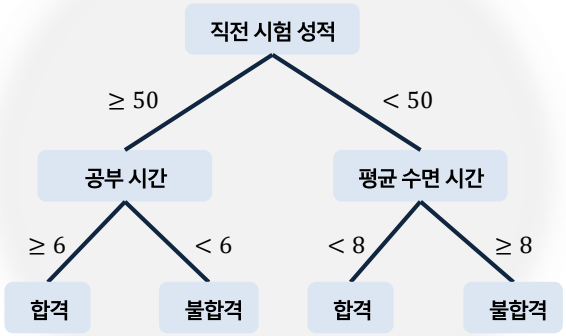
모델 g가 복잡한 정도



Logistic Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Decision Tree

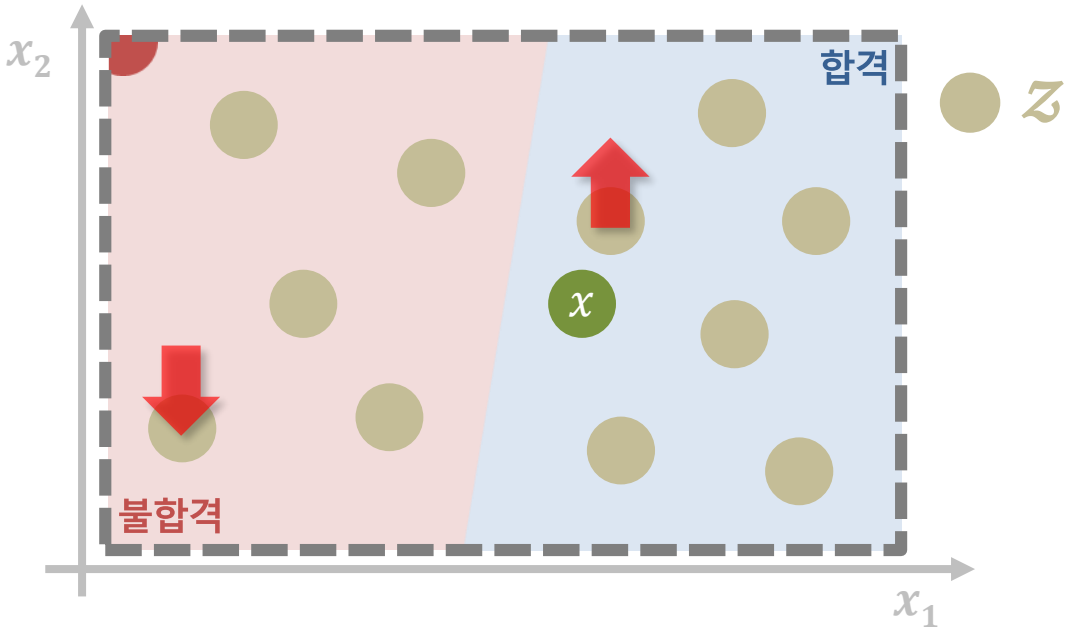


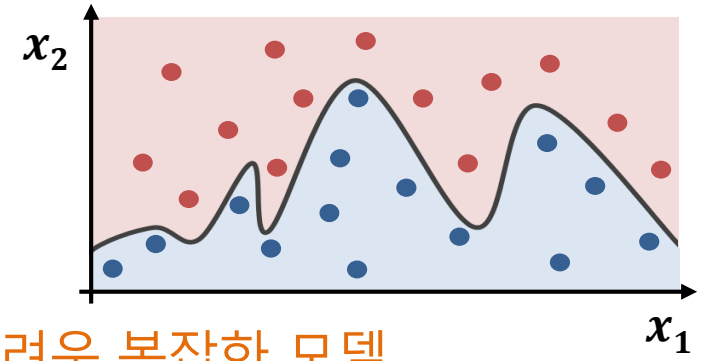
LIME

❖ LIME(Local Interpretable Model-agnostic Explanations)

관측치 x에 대한 좋은 설명 모델 g

- 해석할 수 있는 단순한 모델
- 해석하고자 하는 모델의 예측과 유사하게 예측





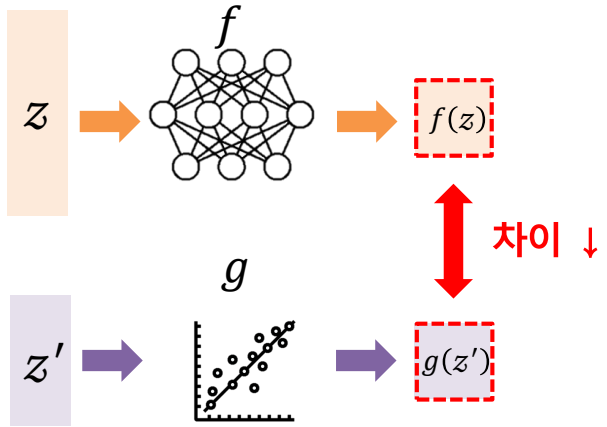
해석이 어려운 복잡한 모델

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

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in Z} \pi_x(z) \left(f(z) - g(z')\right)^2$$

각 z의 가중치

차원 축소된 z



LIME

❖ LIME(Local Interpretable Model-agnostic Explanations)

해석 하고자 하는 모델 예측과 설명 모델 예측의 차이 ↓

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

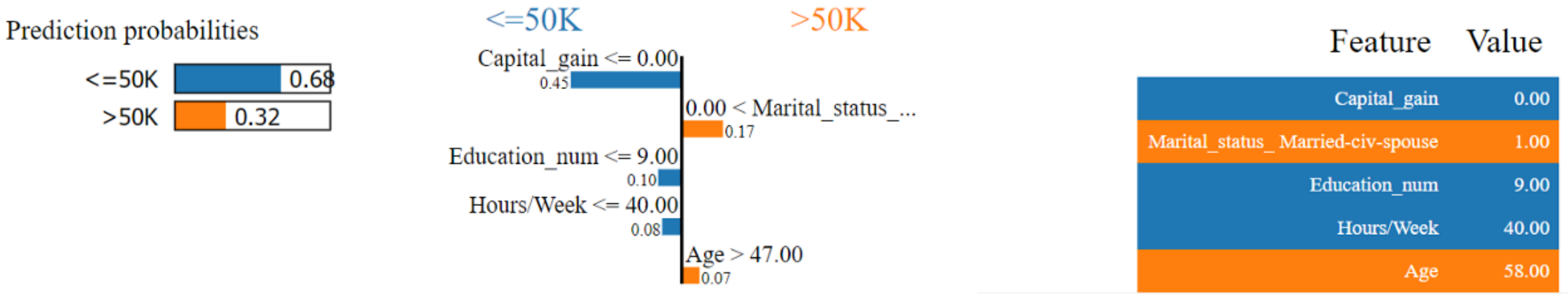
↓ ↓

모델이 복잡한 정도 ↓

LIME

❖ **LIME(Local Interpretable Model-agnostic Explanations)**

- 미국 성인 소득 데이터셋 이진 분류 문제 (5만 달러 이하 / 5만 달러 초과)
- LightGBM(Light Gradient Boosting Machine) 해석 결과

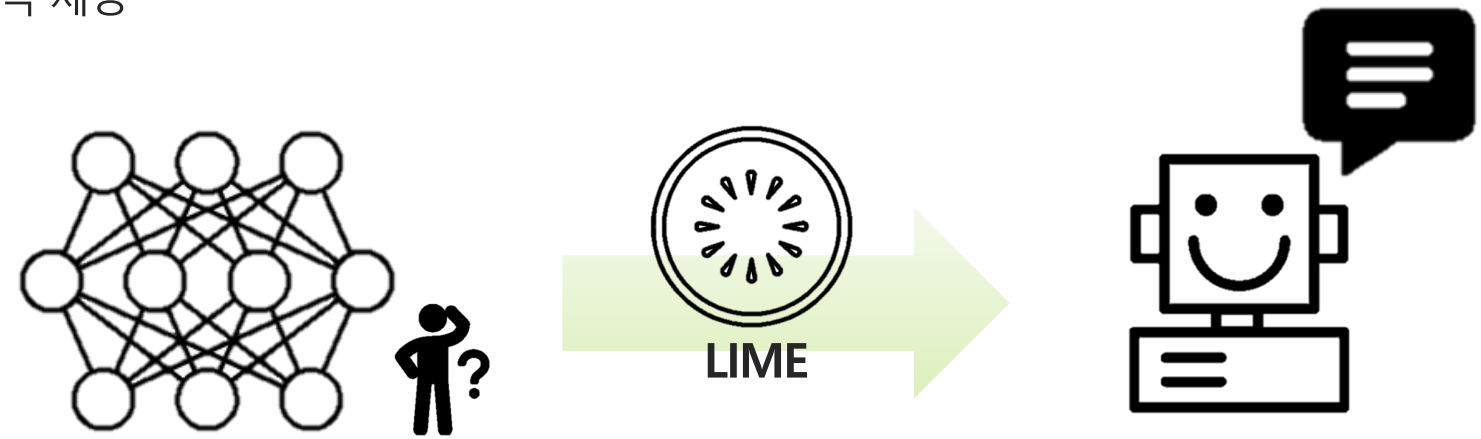


이미지 출처 : https://www.kdnuggets.com/2019/08/open-black-boxes-explainable-machine-learning.html

LIME

❖ **LIME(Local Interpretable Model-agnostic Explanations)**

- 인공지능 모델의 예측 근거를 이해하는 것은 모델에 대한 신뢰성을 확보할 수 있기 때문에 매우 중요하다는 배경으로 제안됨
- 뉴럴 네트워크, 랜덤 포레스트 등 모든 모델에 제한 없이 적용 가능하며 다양한 데이터 형식에도 유연하게 사용 가능
- 전체 데이터셋에 대한 decision boundary가 아닌 특정 관측치 하나의 decision boundary에 근사한 설명 모델을 통해 예측에 대한 해석 제공



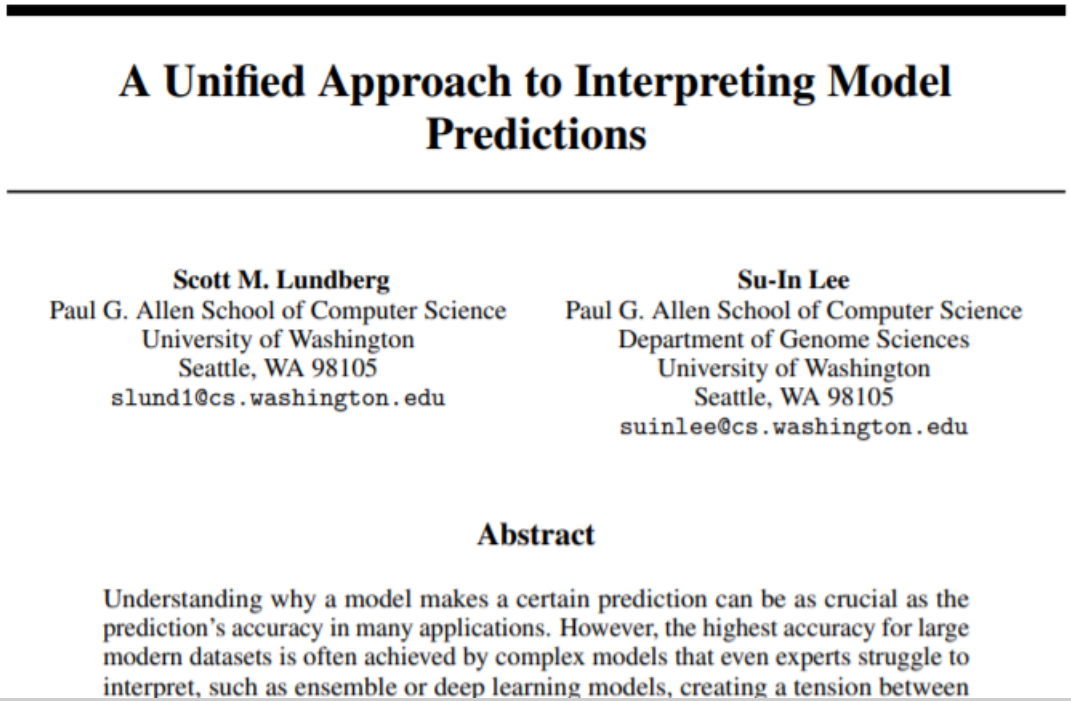
3. SHAP

(SHapley Additive exPlanations)

SHAP

❖ **A Unified Approach to Interpreting Model Predictions**

- Lundberg, S. M., & Lee, S. I. (2017). NeurIPS.
- 2021년 8월 30일 기준 4,704회 인용
- Shapley Values를 기반으로 예측 값에 대해 각 피처가 미치는 기여도를 측정하여 예측에 대한 해석 제공



A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

Abstract

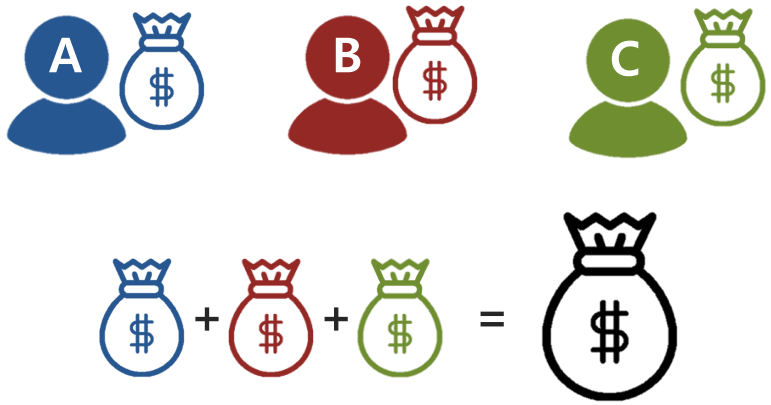
Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between

SHAP

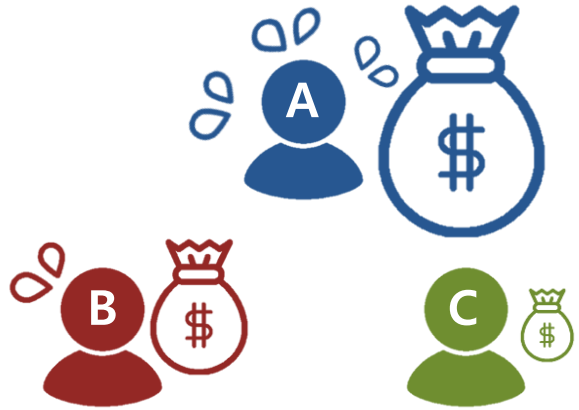
❖ **Shapley Values**

- 1951년 Lloyd Shapley가 제안하였으며 가장 합리적인 분배 이론 중 하나
- 게임 이론(Game theory)을 바탕으로 게임에서 각 플레이어의 기여도에 따라 상금을 공정하게 할당하기 위한 방법

1. 각 플레이어에게 할당되는 상금의 총합 = 게임의 상금



2. 플레이어의 기여도가 높을 수록 높은 상금 할당

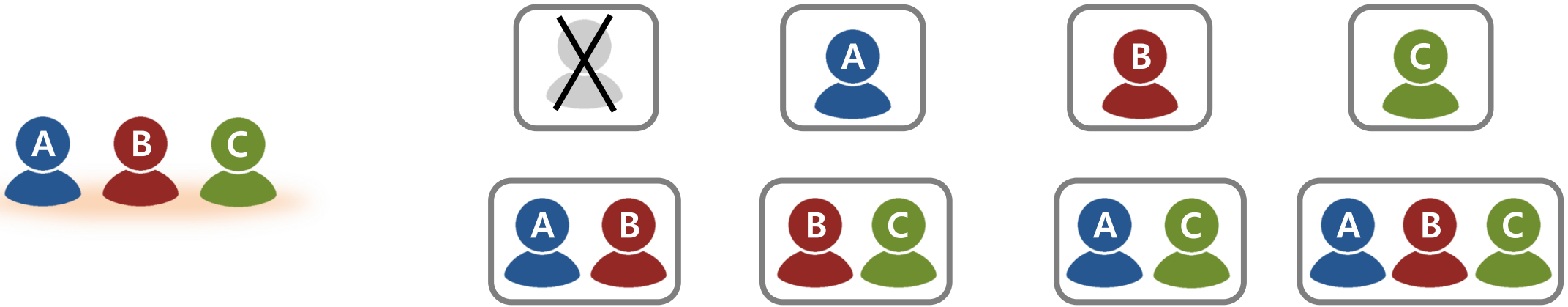


공정한 상금 할당을 위한 조건 2가지

SHAP

❖ **Shapley Values**

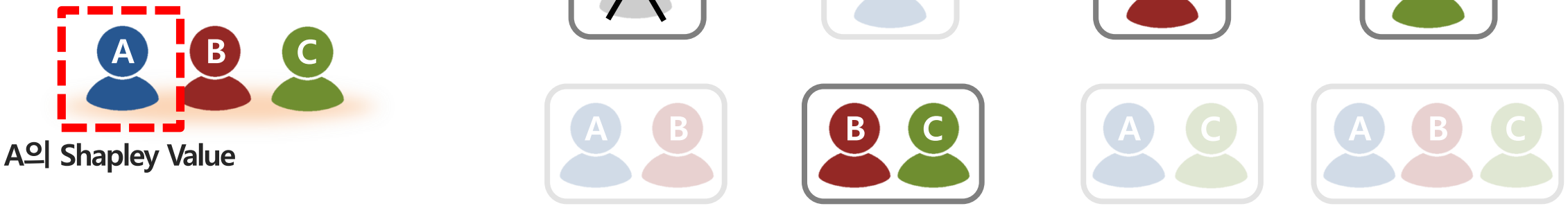
- 플레이어의 **Marginal contributions**를 계산하여 가중 평균한 값
- **Marginal contributions** : 플레이어 전체 집합에 대해 가능한 모든 부분 집합 마다 **특정 플레이어 존재 여부**에 따른 상금 변화량



SHAP

❖ **Shapley Values**

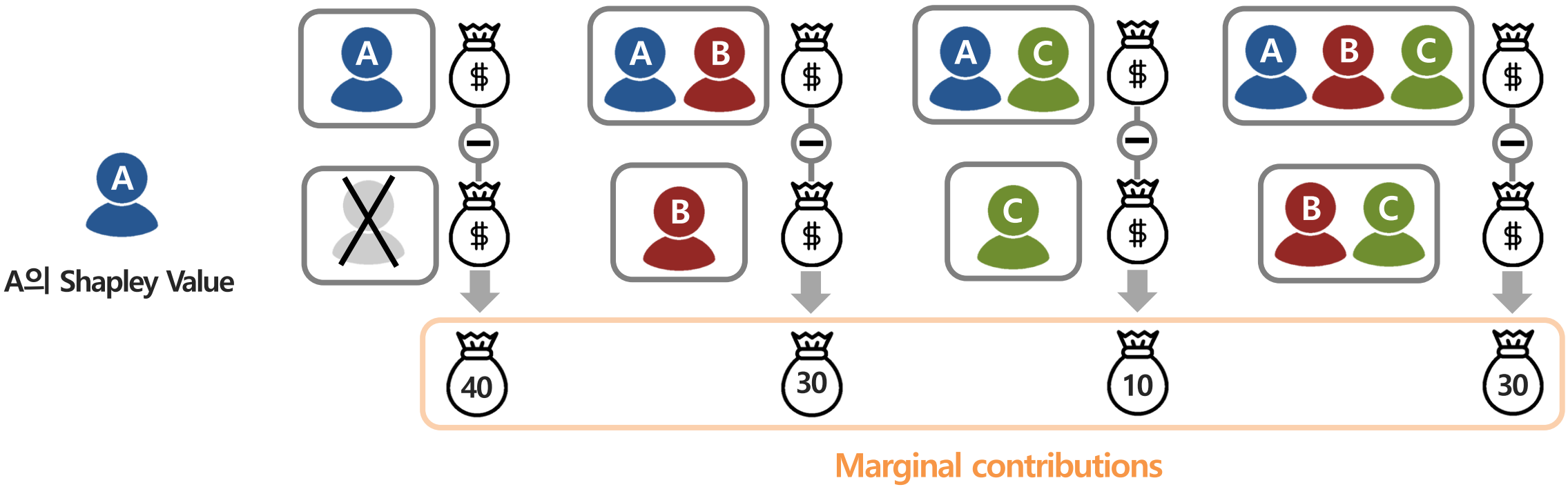
- 플레이어의 **Marginal contributions**를 계산하여 가중 평균한 값
- **Marginal contributions** : 플레이어 전체 집합에 대해 가능한 모든 부분 집합 마다 **특정 플레이어 존재 여부**에 따른 상금 변화량



SHAP

❖ **Shapley Values**

- 플레이어의 **Marginal contributions**를 계산하여 가중 평균한 값
- **Marginal contributions** : 플레이어 전체 집합에 대해 가능한 모든 부분 집합 마다 특정 플레이어 존재 여부에 따른 상금 변화량



SHAP

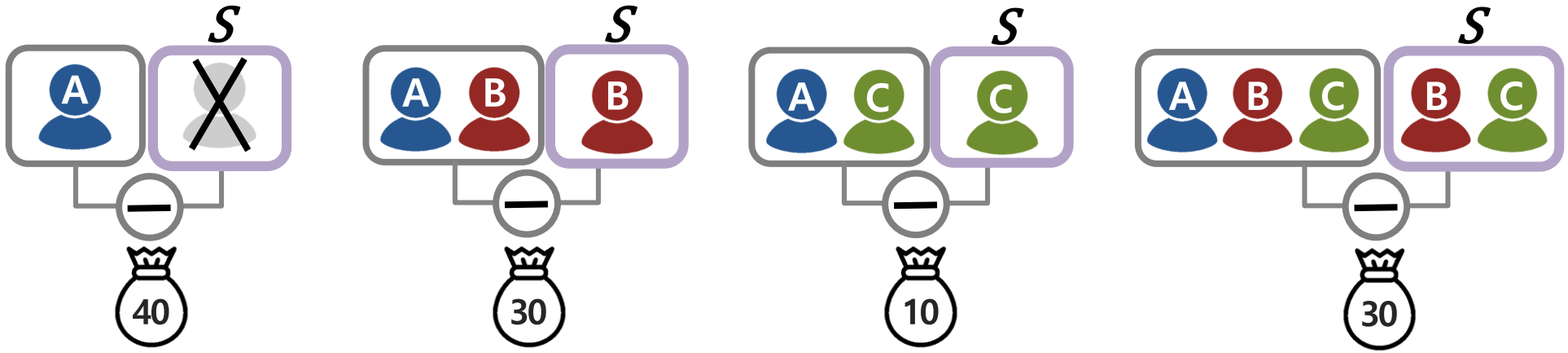
❖ Shapley Values

- 플레이어의 Marginal contributions를 계산하여 가중 평균값
- Marginal contributions : 플레이어 전체 집합에 대해 가능한 모든 서로 집합에서 특정 플레이어 존재 여부에 따른 상금 변화량

$$\frac{|S|!\,(|F|-|S|-1)!}{|F|!}$$

- F : 플레이어 전체 집합
- S: Shapley Value를 계산하는 플레이어가 제외된 각 집합

$F = \{$ A B C $\}$



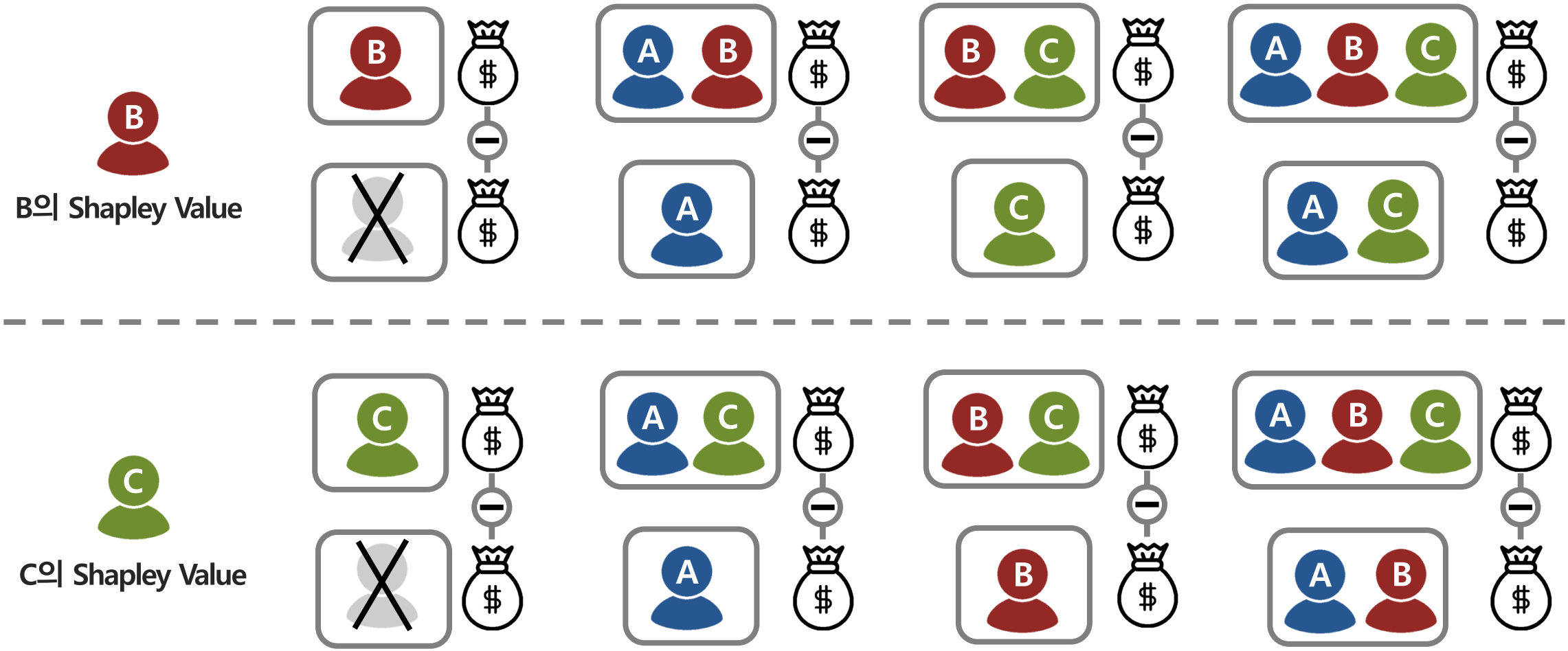
$\frac{0!(3-0-1)!}{3!} = \frac{1}{3}$ $\frac{1!(3-1-1)!}{3!} = \frac{1}{6}$ $\frac{1!(3-1-1)!}{3!} = \frac{1}{6}$ $\frac{2!(3-2-1)!}{3!} = \frac{1}{3}$

A의 Shapley Values

$40 \times \frac{1}{3} + 30 \times \frac{1}{6} + 10 \times \frac{1}{6} + 30 \times \frac{1}{3} = 30$

SHAP

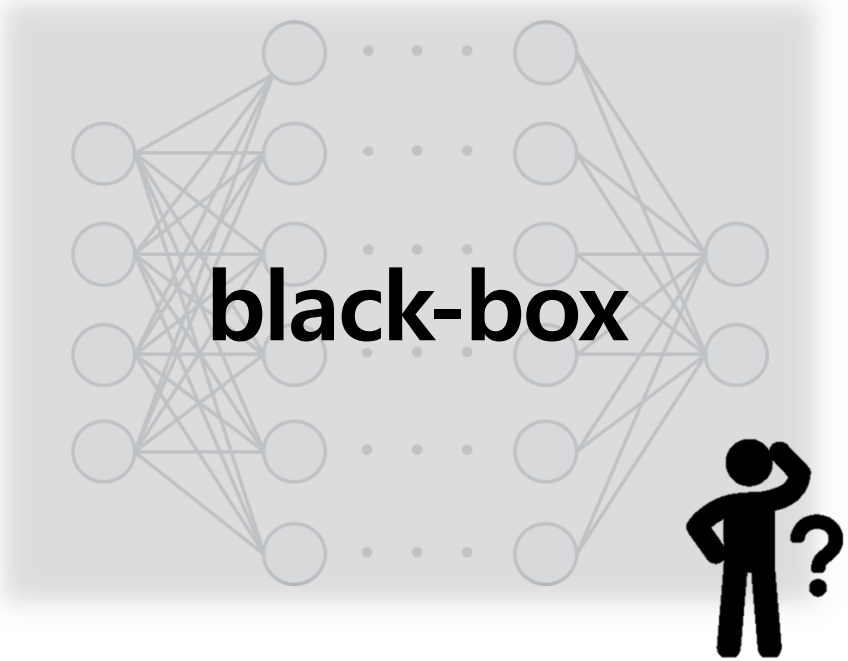
❖ Shapley Values



SHAP

❖ **SHAP(SHapley Additive exPlanations)**

- Shapley Values를 기반으로 예측 값에 대해 각 피처가 미치는 기여도를 측정하여 예측에 대한 해석 제공



플레이어

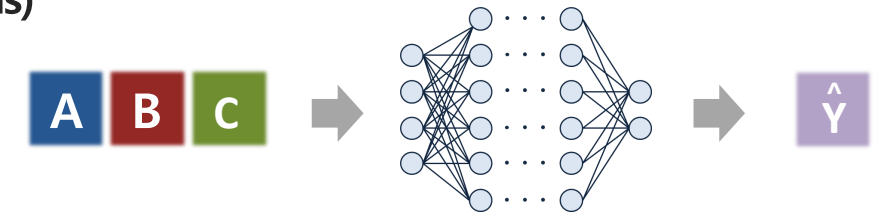
모델의 input features

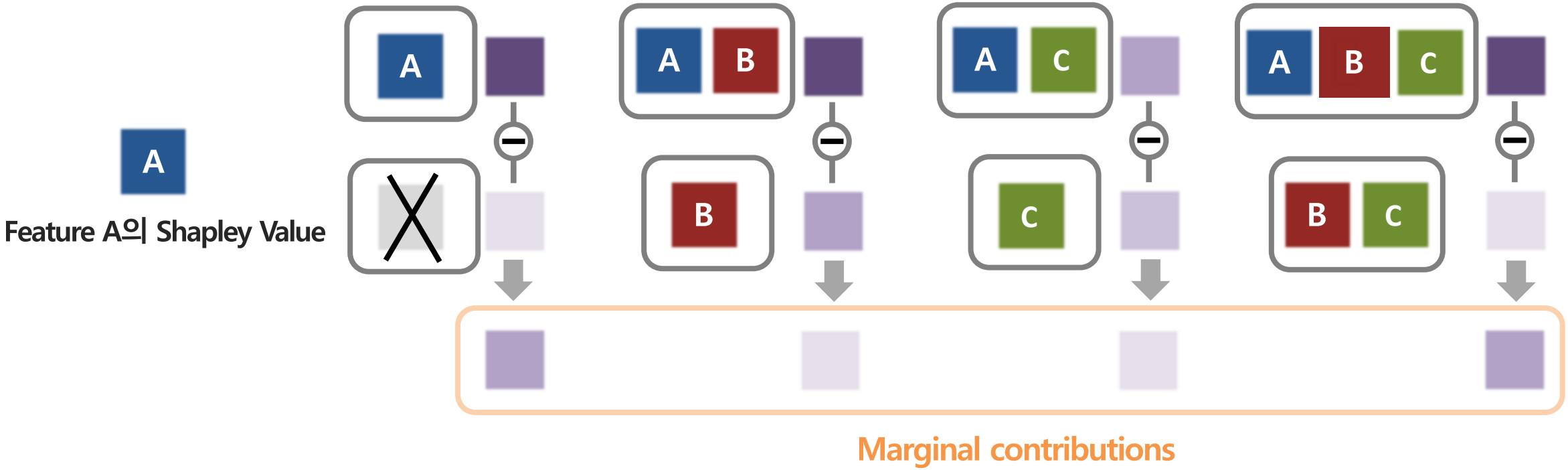
상금

모델의 예측

SHAP

❖ SHAP(SHapley Additive exPlanations)

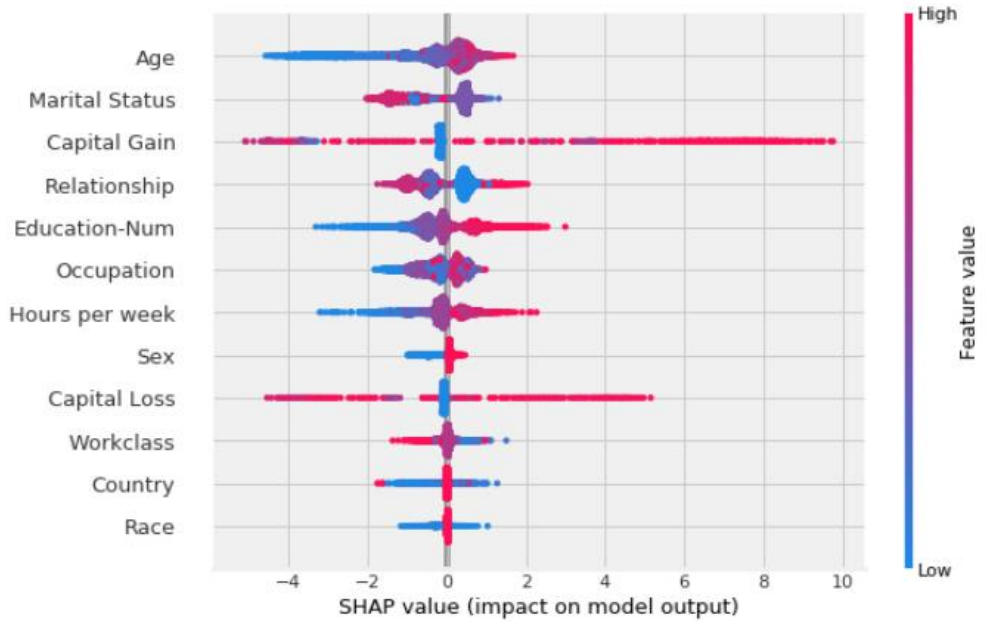




SHAP

❖ **SHAP(SHapley Additive exPlanations)**

- 미국 성인 소득 데이터셋 이진 분류 문제 (0 : 5만 달러 이하 / 1 : 5만 달러 초과)
- 나이가 많을수록 소득이 5만 달러를 초과할 확률에 큰 기여
- 교육 수준이 높을수록 소득이 5만 달러를 초과할 확률에 큰 기여

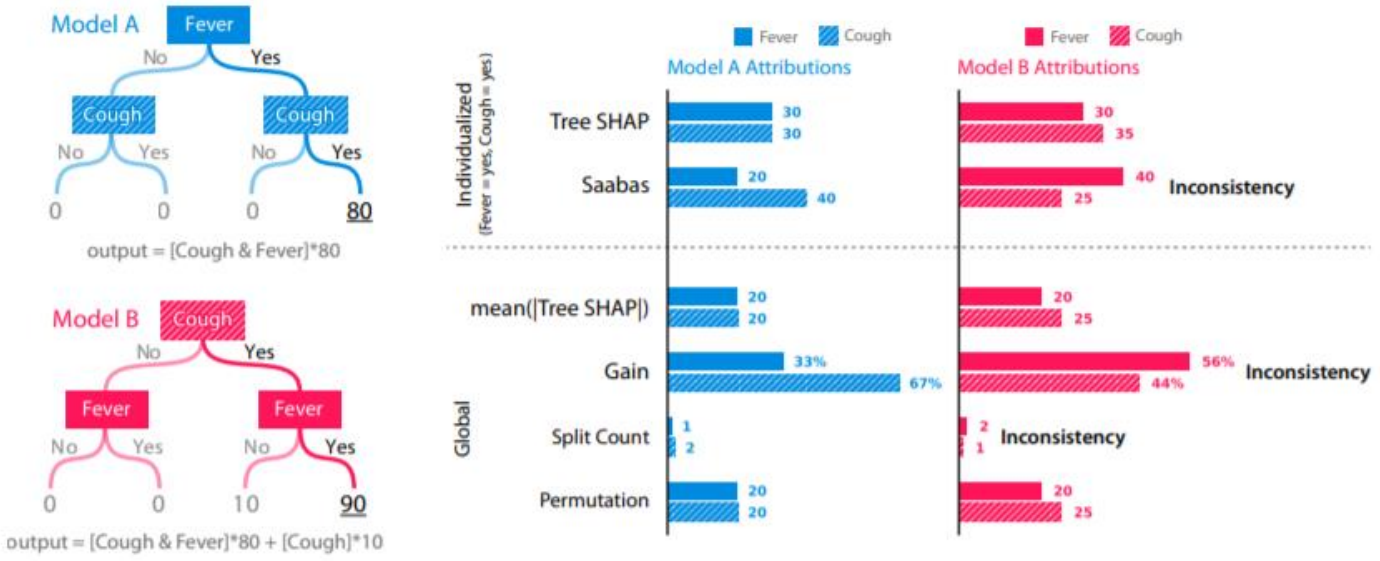


이미지 출처 : https://towardsdatascience.com/explainable-artificial-intelligence-part-3-hands-on-machine-learning-model-interpretation-e8ebe5afc608

SHAP

❖ **SHAP(SHapley Additive exPlanations)**

- 게임 이론(Game theory)을 바탕으로 하는 Shapley Values를 사용하여 예측에 대한 각 피처의 기여도 계산
- Shapley Values 계산 시 수많은 경우에 대해 연산이 이루어지므로 오랜 시간이 걸린다는 단점 존재
- Shapley Values 단점 보완을 위해 Kernel SHAP, Deep SHAP, Tree SHAP 등 다양한 구조의 방법 제안



Tree SHAP

이미지 출처 : Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2921-2929).

4. Conclusions

Conclusions

❖ XAI의 필요성

- 예측에 대한 해석이 어려운 Black box 모델의 경우 XAI 알고리즘을 통해 해석 가능

❖ Model-agnostic XAI

- 대표적인 방법론 : LIME, SHAP
- LIME : 관측치 하나에 대한 설명 모델을 추정하여 Black box 모델 해석 제공
- SHAP : Shapley Values를 기반으로 예측 값에 대한 각 피처의 기여도를 계산하여 Black box 모델 해석 제공
- SHAP은 계산해야 하는 피처 조합의 수가 많아지는 경우 연산 시간이 길어진다는 단점 존재

❖ 연구 동향

- LIME : KL-LIME, DLIME, Quadratic-LIME(QLIME) 등의 후속 연구
- SHAP : extended SHAP, DeepSHAP-LSTM 확장 등의 후속 연구

Thank you

References

1. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " 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 (pp. 1135-1144).
2. Lundberg, S. M., & Lee, S. I. (2017, December). A unified approach to interpreting model predictions. In Proceedings of the 31st international conference on neural information processing systems (pp. 4768-4777).
3. Das, Arun, and Paul Rad. "Opportunities and challenges in explainable artificial intelligence (xai): A survey." arXiv preprint arXiv:2006.11371 (2020).
4. https://www.youtube.com/watch?v=f2XqxOny3NA&t=711s
5. http://dmqm.korea.ac.kr/activity/seminar/297
6. https://www.youtube.com/watch?v=d6j6bofhj2M&t=608s