

Introduction to Explainable Boosting Machine (EBM)

2023-08-11

Korea University

Data Mining & Quality Analytics Lab.

심 세 진

발표자 소개



❖ 심세진 (Sejin Sim)

- 고려대학교 산업경영공학과 대학원 재학
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- 석박통합 과정 (2022. 03 ~ Present)

❖ 관심 연구 분야

- Deep Learning for Multichannel Signal Analysis
- Semi-supervised Regression

❖ 연락처

- ssj259@korea.ac.kr

Contents

- ❖ Introduction
- ❖ Explainable Boosting Machine (EBM)
- ❖ Conclusion
- ❖ References

Introduction

Introduction

❖ 관련 세미나 소개

부스팅 모델 (머신러닝) 관련

종료

Introduction to boosting

2021. 07. 02
Data Mining & Quality Analytics Lab.
발표자: 고은지
ekorh21@kma.ac.kr

Introduction to boosting

발표자:  **고은지**

 2021년 7월 2일
 오후 1시 ~
 온라인 비디오 시청 (YouTube)

[세미나 정보 보기 →](#)

XAI 기법 (SHAP, LIME) 관련

종료

DMQA open seminar
Applications of eXplainable Artificial Intelligence(XAI)

2022. 4. 8
Data Mining & Quality Analytics Lab.
발표자: 황석철
hs941125@kma.ac.kr

Applications of eXplainable Artificial Inte

발표자:  **황석철**

 2022년 4월 8일
 오후 12시 ~
 온라인 비디오 시청 (YouTube)

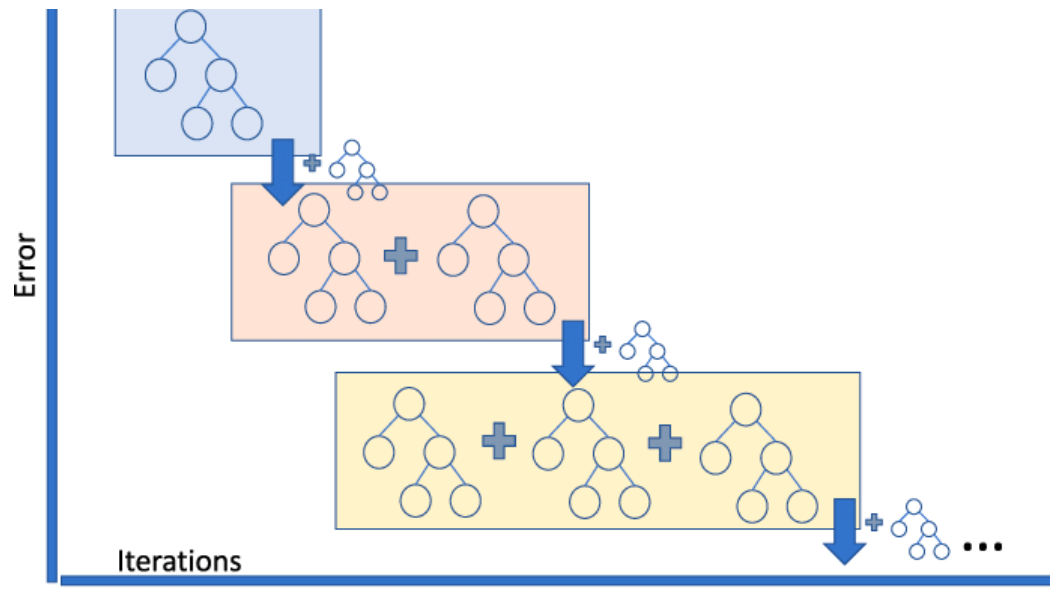
[세미나 정보 보기 →](#)

종료**Introduction to XAI(eXplainable AI)****Introduction to XAI(eXplainable AI)**

Introduction

❖ 머신러닝의 활용성 및 기존 한계점

- 머신러닝은 표 형식(tabular) 데이터를 다루는 경우, 많은 대회(Kaggle 등)에서 우수한 성과를 보임
- 우수한 성과를 보이는 머신러닝 모델들(ex. 부스팅 모델들: LGBM, XGB 등)은 복잡한 구조로 모델 의사결정 과정을 설명하기 어려움



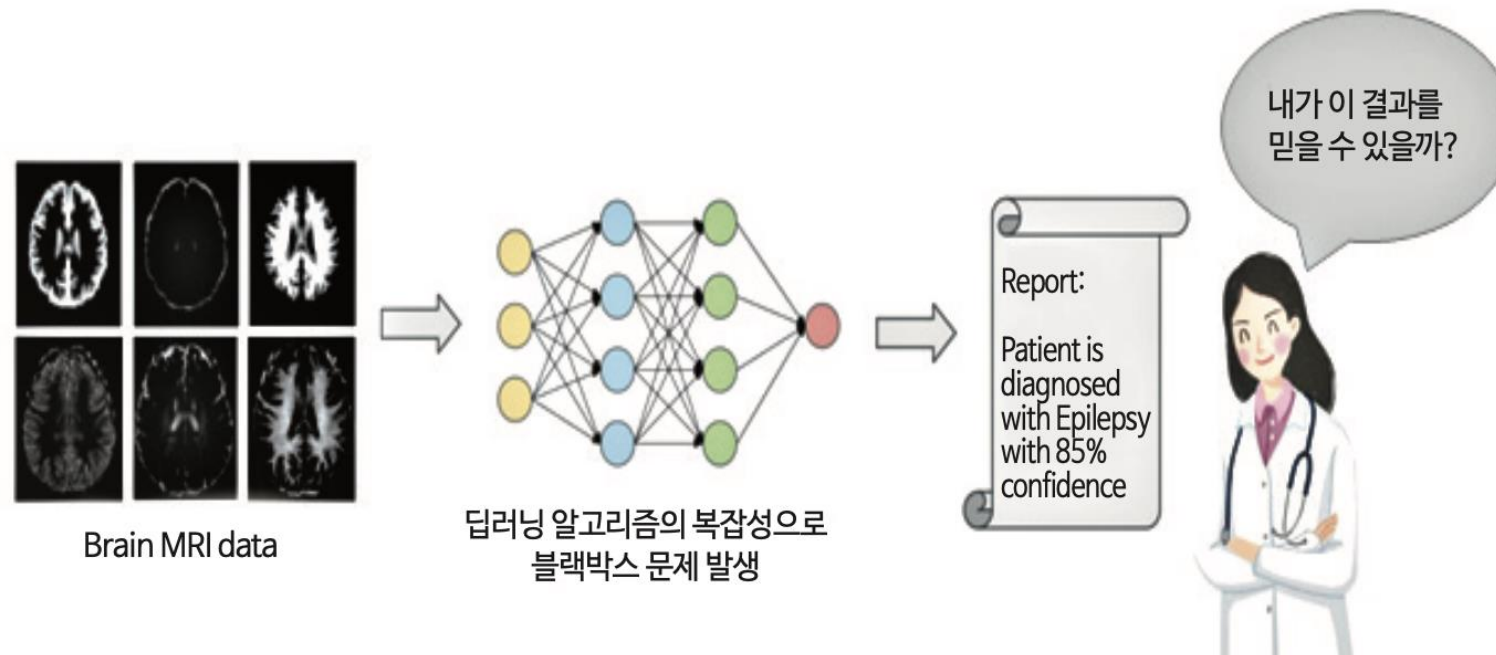
〈 Gradient Boosting 예시〉



Introduction

❖ 머신러닝의 활용성 및 기존 한계점

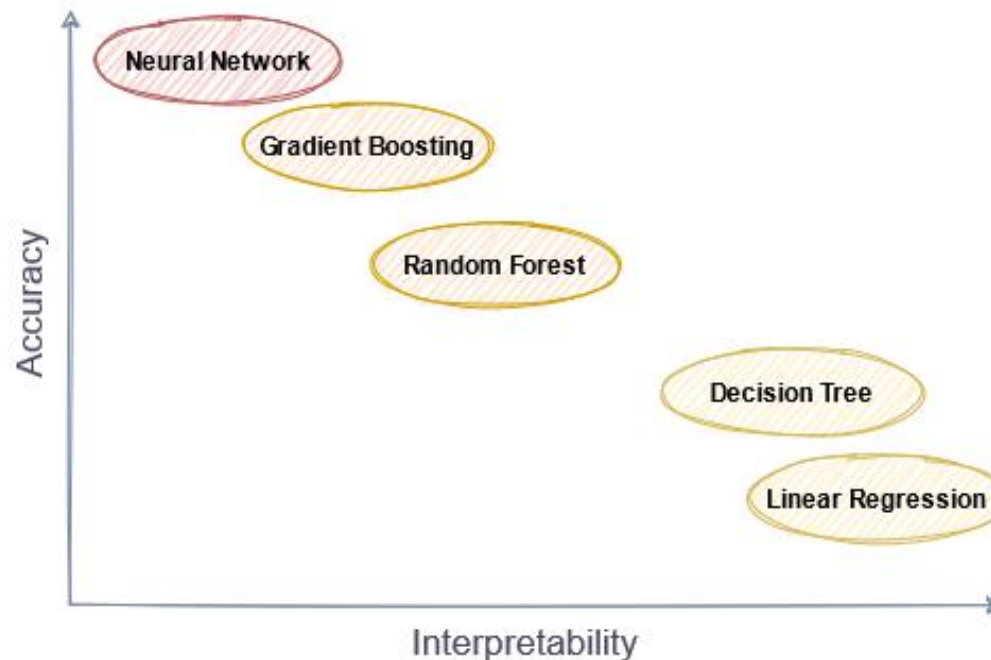
- 현업에서는 인공지능 예측 결과에 대한 근거를 필요로 하며 특히, 의료와 금융과 같은 분야는 성능보다는 결과에 대한 설명도가 중요한 요소로 자리잡음
- 예시) 어떠한 특징을 기반으로 질환을 진단하였는지에 대한 설명이 필요



Introduction

❖ 머신러닝의 활용성 및 기존 한계점

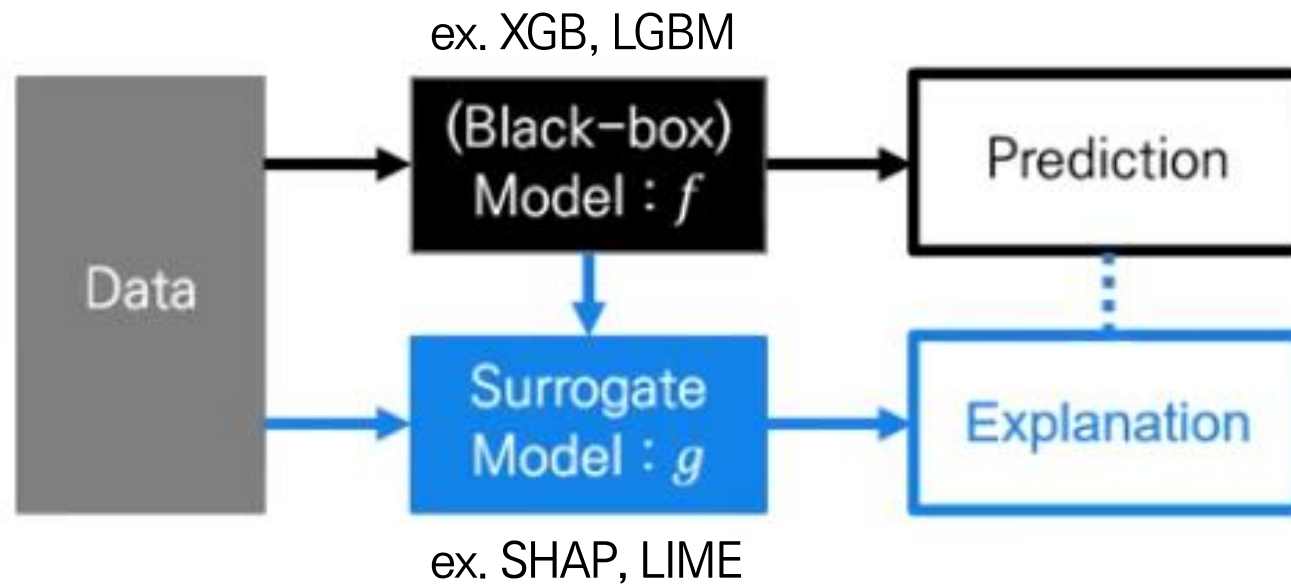
- 모델의 성능과 설명 가능성 사이에는 역의 상관관계가 존재함 (높은 성능 vs 높은 해석력)
- 해석이 쉬운 모델 : Linear Regression, Decision tree 등 해석이 쉽지만, 정확도가 낮아 활용도가 낮음
- 해석이 어려운 모델 : Ensemble 및 Boosting, Neural Network 등 성능이 높지만, 모델 자체 해석력이 낮음



Introduction

❖ 머신러닝의 활용성 및 기존 한계점

- 해석이 어려운 모델의 예측 결과를 해석하기 위해 설명 가능한 XAI 기법들(ex. SHAP, LIME)이 제안 됨
- 예측 모델과 해석 모델을 분리하는 Model-Agnostic 방법으로 모델의 의사 결정을 간접적으로 추론하는 한계를 가짐



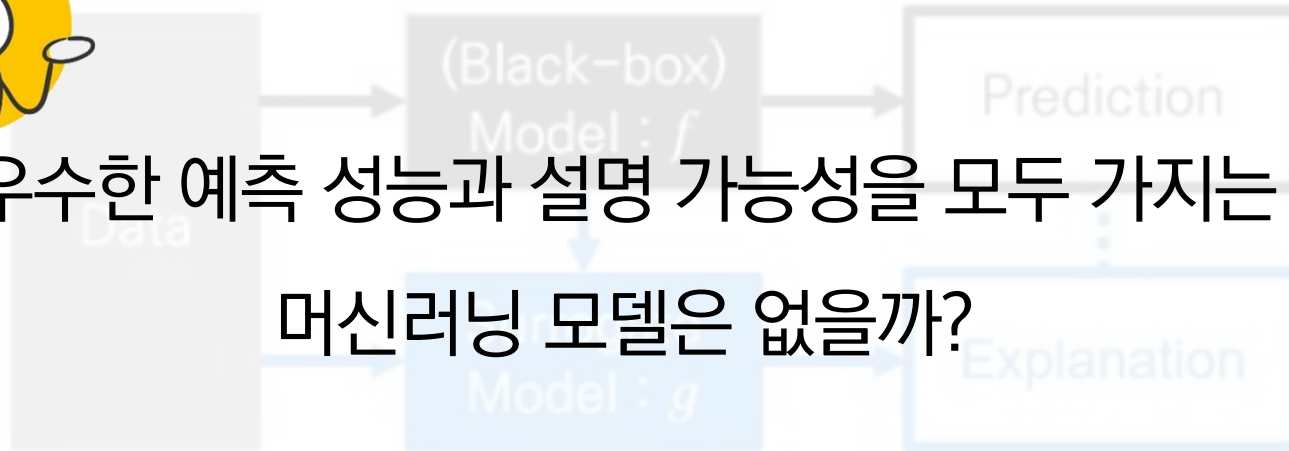
Introduction

❖ 머신러닝의 활용성 및 기존 한계점

- 모델의 예측 결과를 해석하기 위해 설명 가능한 XAI 기법들(ex. SHAP, LIME)이 제안 됨
- 예측 모델과 해석 모델을 분리하는 Model-Agnostic 방법으로 모델의 의사 결정을 간접적으로 추론하는 한계를 가짐



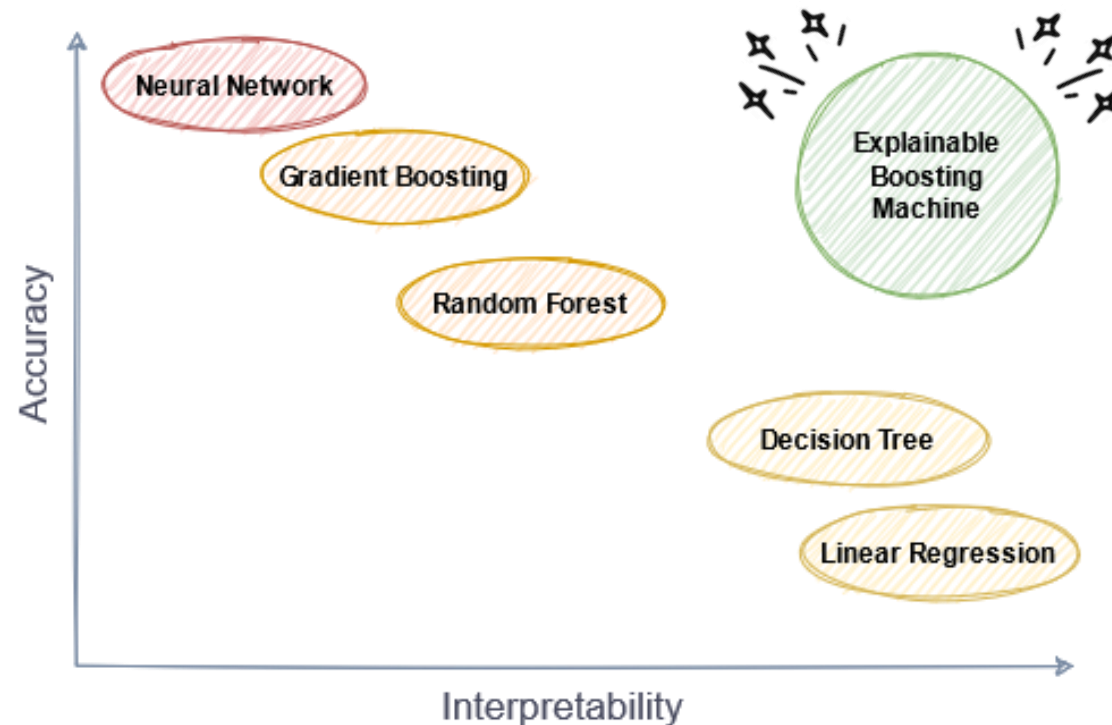
우수한 예측 성능과 설명 가능성을 모두 가지는
머신러닝 모델은 없을까?



Introduction

❖ 머신러닝의 활용성 및 기존 한계점

- 기존의 한계를 극복하는 Explainable Boosting Machine (EBM)이 등장함



Explainable Boosting Machine (EBM)

Explainable Boosting Machine (EBM)

❖ Interpretml : A unified framework for machine learning interpretability

- 2019년 마이크로 소프트에서 머신러닝 해석을 위한 python package인 InterpretML을 소개하며 EBM이 소개 됨 (2023년 8월 8일 기준 362회 인용)
- EBM은 기존에 높은 성능인 부스팅 모델에 준하는 성능을 가지며 의사결정 과정을 설명 가능한 머신러닝 모델임

InterpretML: A Unified Framework for Machine Learning Interpretability

Harsha Nori
Samuel Jenkins
Paul Koch
Rich Caruana
Microsoft Corporation
1 Microsoft Way
Redmond, WA 98052, USA

HANORI@MICROSOFT.COM
SAJENKIN@MICROSOFT.COM
PAULKOCH@MICROSOFT.COM
RCARUANA@MICROSOFT.COM

Editor:

Abstract

InterpretML is an open-source Python package which exposes machine learning interpretability algorithms to practitioners and researchers. InterpretML exposes two types of interpretability – **glassbox**, which are machine learning models designed for interpretability (ex: linear models, rule lists, generalized additive models), and **blackbox** explainability techniques for explaining existing systems (ex: Partial Dependence, LIME). The package enables practitioners to easily compare interpretability algorithms by exposing multiple methods under a unified API, and by having a built-in, extensible visualization platform. InterpretML also includes the first implementation of the Explainable Boosting Machine, a powerful, interpretable, glassbox model that can be as accurate as many blackbox models. The MIT licensed source code can be downloaded from github.com/microsoft/interpret.

Keywords: Interpretability, Explainable Boosting Machine, Glassbox, Blackbox

Explainable Boosting Machine (EBM)

❖ Before EBM

- Generalized Additive Model (GAM) – 일반화 가법 모형
 - ✓ 통계 모델링 기법 중 하나로, 변수 간 가산성(Additivity)을 유지하면서 비선형 관계를 모델링할 수 있음

Explainable Boosting Machine (EBM)

❖ Before EBM

- Generalized Additive Model (GAM) – 일반화 가법 모형
 - ✓ 통계 모델링 기법 중 하나로, 변수 간 가산성(Additivity)을 유지하면서 비선형 관계를 모델링할 수 있음
 - 연산 값의 합을 의사결정에 활용하기 때문에 각 변수가 어떤 의사결정에 영향을 미치는지 파악하기 쉬움

✓ GAM 수식

$$g(E(y)) = \beta_0 + \sum f_j(x_j) = \beta_0 + f_1(x_1) + \cdots + f_j(x_j)$$

- $g(\cdot)$ 는 Link Function(= regression or classification)
- $f(\cdot)$ 는 Shape Function(= ex. spline, tree, gradient boosting etc.)

Explainable Boosting Machine (EBM)

❖ Before EBM

- Generalized Additive Model (GAM) – 일반화 가법 모형

- ✓ 통계 모델링 기법 중 하나로, 변수 간 **가산성(Additivity)**을 유지하면서 비선형 관계를 모델링할 수 있음

- 연산 값의 합을 의사결정에 활용하기 때문에 각 변수가 어떤 의사결정에 영향을 미치는지 파악하기 쉬움

- 부스팅 모델은 $E(y) = f(x_1 + \dots + x_j)$ 형태로 예측하기에 각 변수가 의사 결정에 어떻게 기여하는지 쉽게 파악하기 어려움

- ✓ GAM 수식

$$g(E(y)) = \beta_0 + \sum f_j(x_j) = \beta_0 + f_1(x_1) + \dots + f_j(x_j)$$

- $g(\cdot)$ 는 Link Function(= regression or classification)

- $f(\cdot)$ 는 Shape Function(= ex. spline, tree, gradient boosting etc.)

- ✓ Linear Regression (참고)

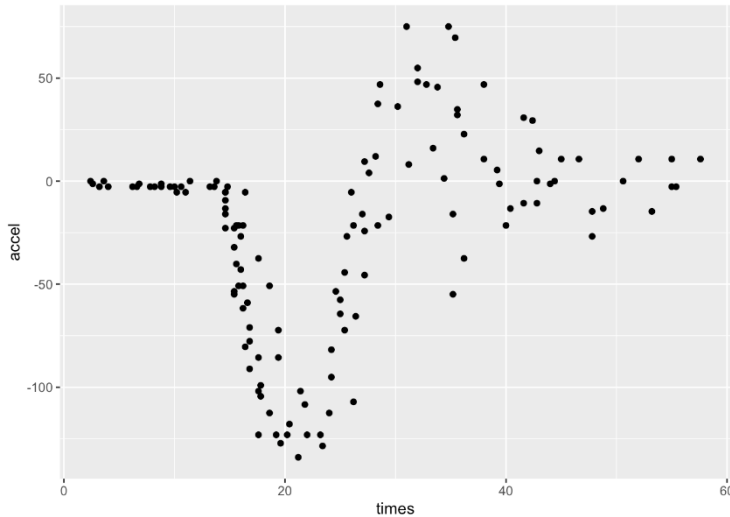
$$E(y) = \beta_0 + \sum \beta_j x_j = \beta_0 + \beta_1 x_1 + \dots + \beta_j x_j$$

Explainable Boosting Machine (EBM)

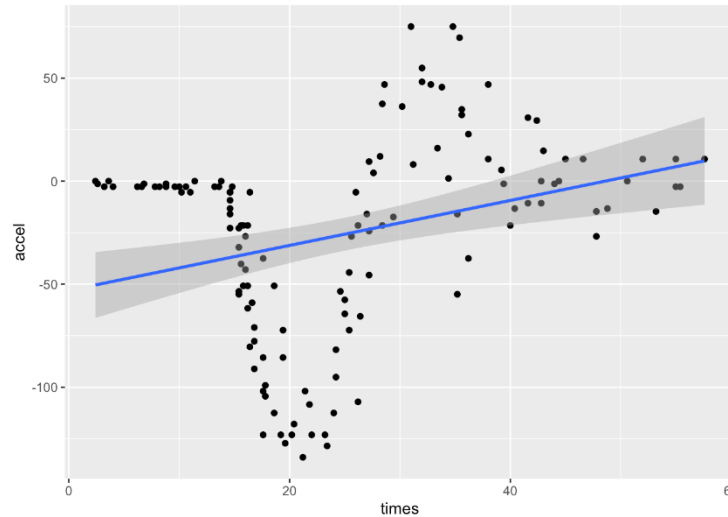
❖ Before EBM

- Generalized Additive Model (GAM) – 일반화 가법 모형
 - ✓ 통계 모델링 기법 중 하나로, 변수 간 가산성(Additivity)을 유지하면서 비선형 관계를 모델링할 수 있음
 - 입력 변수(Feature)와 출력 변수(Target)간의 비선형적 관계를 잘 표현할 수 있어 예측 성능이 선형 회귀 대비 높음

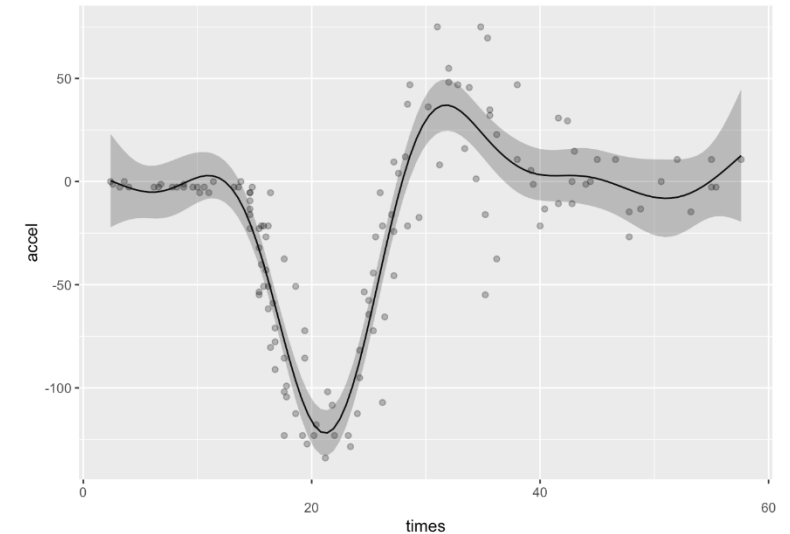
[예시 데이터]



[Linear Regression Model 적용]



[GAM 적용]



Explainable Boosting Machine (EBM)

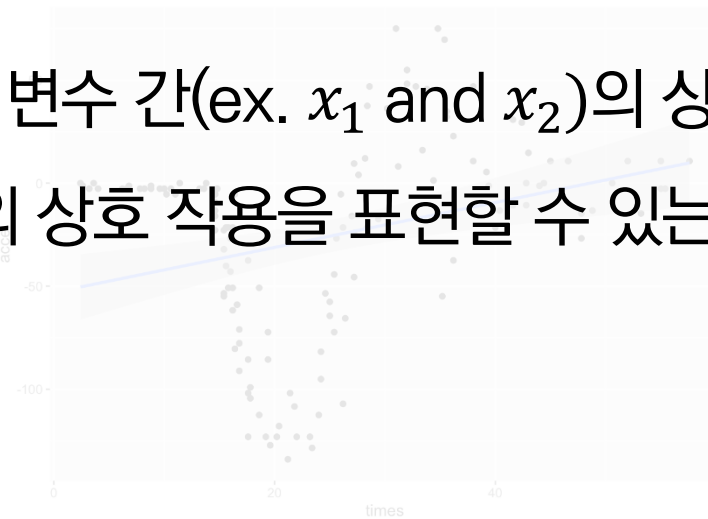
❖ Before EBM

- Generalized Additive Model (GAM) – 일반화 가법 모형
 - ✓ 통계 모델링 기법 중 하나로, 변수 간 가산성(Additivity)을 유지하면서 비선형 관계를 모델링할 수 있음
 - 입력 변수(Feature)와 출력 변수(Target)간의 비선형적 관계를 잘 표현할 수 있어 예측 성능이 선형 회귀 대비 높음

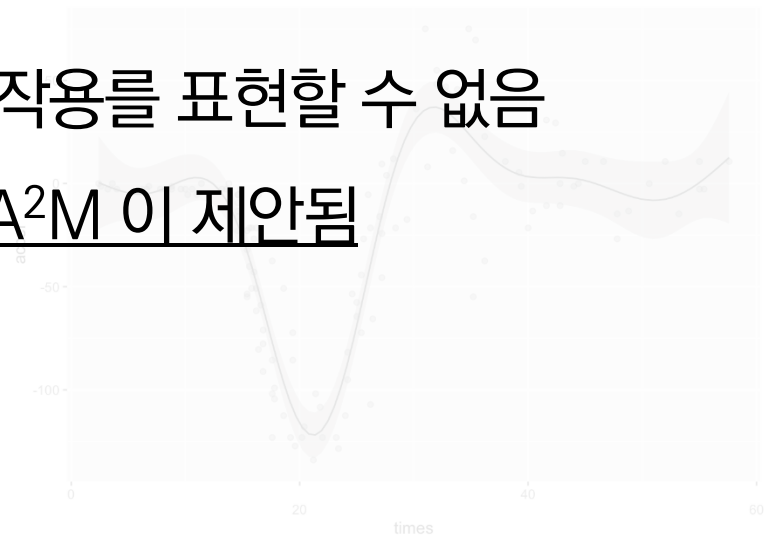
[예시 데이터]



[Linear Regression Model 적용]



[GAM 적용]



한계점 : GAM은 입력 변수 간(ex. x_1 and x_2)의 상호 작용을 표현할 수 없음
→ 입력 변수간의 상호 작용을 표현할 수 있는 GA²M 이 제안됨

Explainable Boosting Machine (EBM)

❖ Before EBM

- Generalized Additive Model Plus Interactions (GA²M)
 - ✓ 입력 변수간 상호 작용을 표현해 주는 pairwise interaction term 이 추가 된 모형

- ✓ GAM 수식

$$g(E(y)) = \beta_0 + \sum f_j(x_j)$$

- ✓ GA²M 수식

$$g(E(y)) = \beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j)$$

Explainable Boosting Machine (EBM)

❖ Before EBM

- Generalized Additive Model Plus Interactions (GA²M)
 - ✓ 입력 변수간 상호 작용을 표현해 주는 pairwise interaction term 이 추가 된 모형
 - 두 개 입력변수간 상호 작용을 시각화하기 위해 heatmap 형태로 추후 보여짐

✓ GAM 수식

$$g(E(y)) = \beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j)$$

한계점 : 입력 변수가 많아지면 모델 개수 $f_j(\cdot)$ 증가에 따라 학습 시간이 늘어남

→ 학습속도가 개선된 Explainable Boosting Machine (EBM)이 제안됨

✓ GA²M 수식

$$g(E(y)) = \beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j)$$

Explainable Boosting Machine (EBM)

❖ Explainable Boosting Machine (EBM)

- EBM 수식 (=GA²M 수식)

$$g(E(y)) = \beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j)$$

- $g(\cdot)$ 는 Link Function(= regression or classification)
- $f(\cdot)$ 는 Feature Function(= gradient boosting)

- GA²M과의 차이점

- ✓ 병렬 연산을 위한 라이브러리(joblib)를 사용하여 멀티 코어 및 멀티 머신 병렬화를 제공
- ✓ gradient boosting 모델을 Feature Function으로 고정하여 사용함

Explainable Boosting Machine (EBM)

❖ Explainable Boosting Machine (EBM)

- EBM 학습 과정 - 1단계 : 개별 입력 변수만을 고려한 학습

✓ 잔차(=실제 출력 변수와 예측 출력 변수의 차)를 기반 모델 학습을 업데이트

$$g(E(y)) = \beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j)$$

Iteration	x_1	x_2	...	x_j
1				

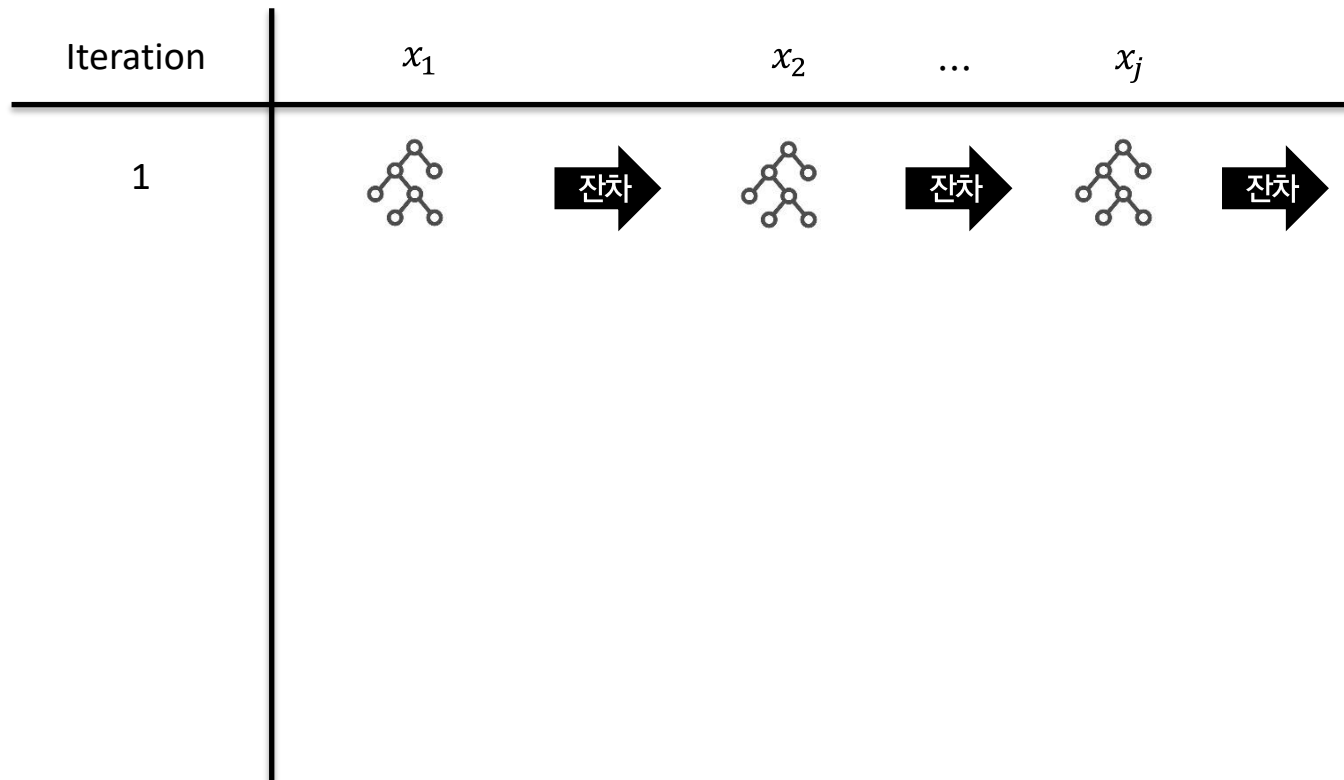
Explainable Boosting Machine (EBM)

❖ Explainable Boosting Machine (EBM)

- EBM 학습 과정 - 1단계 : 개별 입력 변수만을 고려한 학습

✓ 잔차(=실제 출력 변수와 예측 출력 변수의 차)를 기반 모델 학습을 업데이트

$$g(E(y)) = \beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j)$$



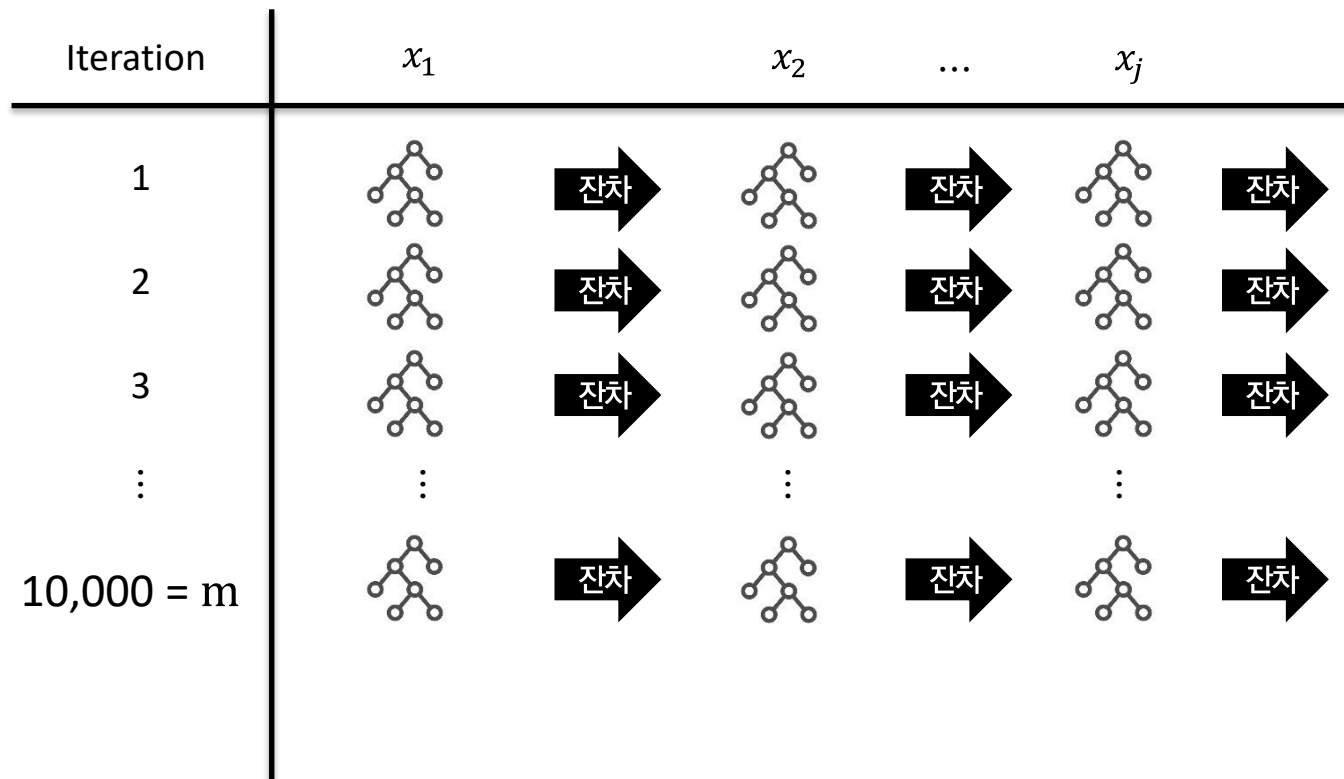
Explainable Boosting Machine (EBM)

❖ Explainable Boosting Machine (EBM)

- EBM 학습 과정 - 1단계 : 개별 입력 변수만을 고려한 학습

✓ 잔차(=실제 출력 변수와 예측 출력 변수의 차)를 기반 모델 학습을 업데이트

$$g(E(y)) = \beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j)$$

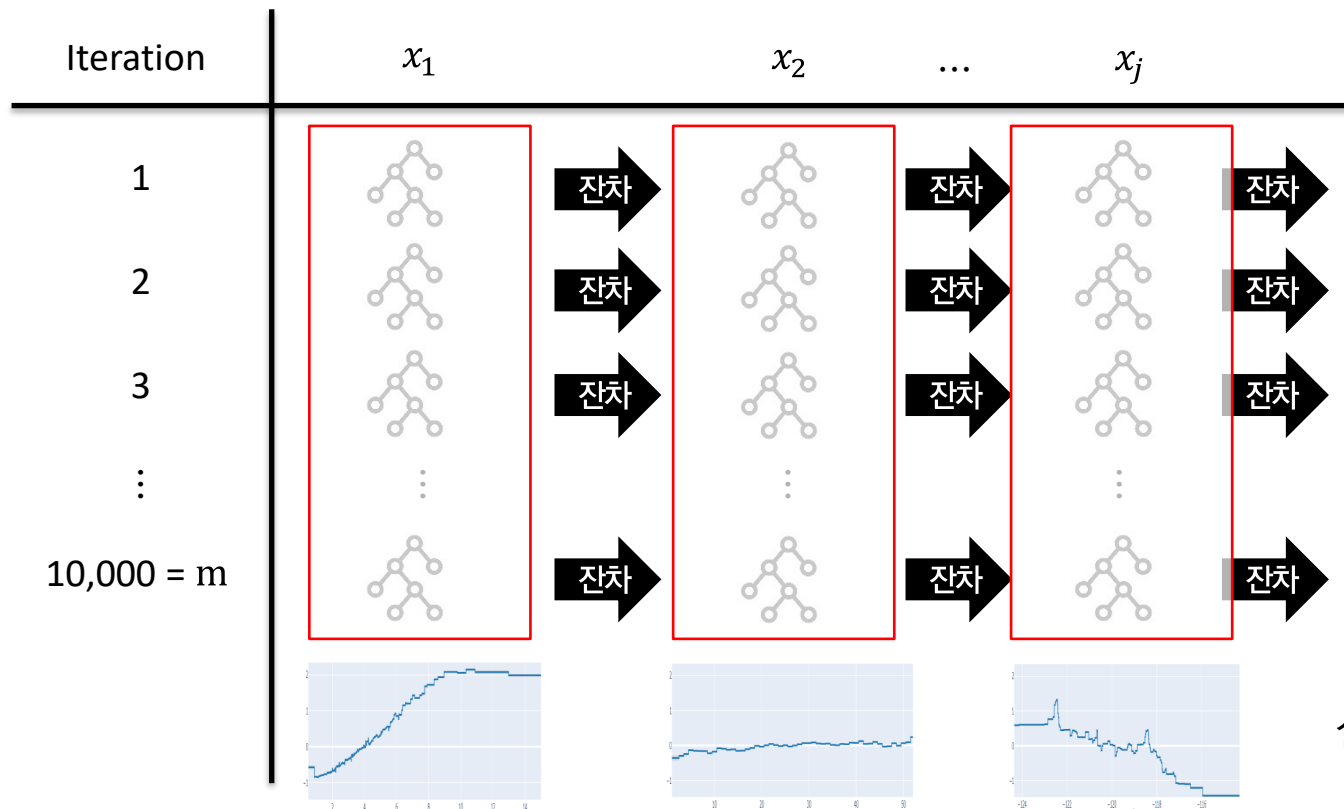


Explainable Boosting Machine (EBM)

❖ Explainable Boosting Machine (EBM)

- EBM 학습 과정 - 1단계 : 개별 입력 변수만을 고려한 학습
 - ✓ m개의 모델을 종합하여 학습 결과를 변수별로 표로 정리

$$g(E(y)) = \beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j)$$



실제로는 표 안에 입력 변수 별 값들이 존재

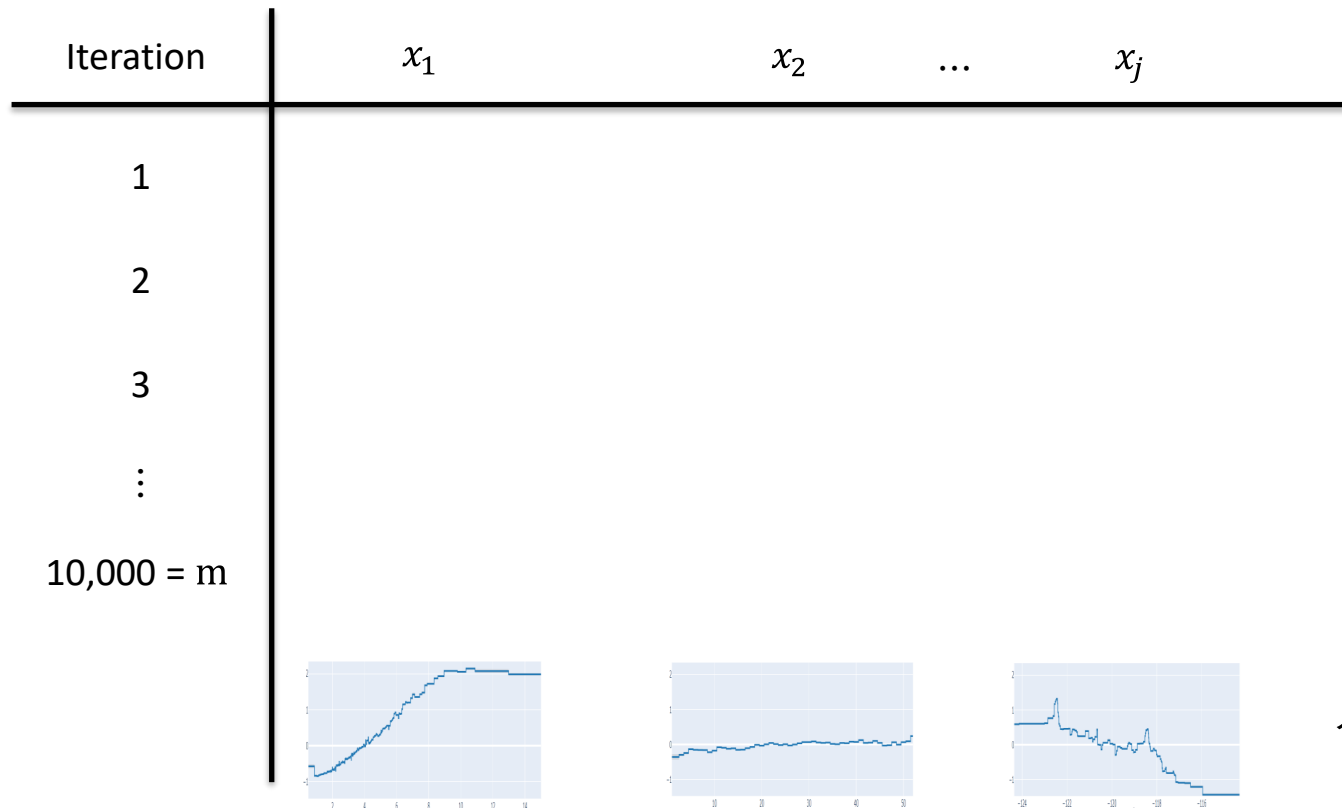
Explainable Boosting Machine (EBM)

❖ Explainable Boosting Machine (EBM)

- EBM 학습 과정 - 1단계 : 개별 입력 변수만을 고려한 학습

✓ 기존 개별모델들은 추론 과정에서 사용하지 않기 때문에 메모리를 위해 삭제함

$$g(E(y)) = \beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j)$$



실제로는 표 안에 입력 변수 별 값들이 존재

Explainable Boosting Machine (EBM)

❖ Explainable Boosting Machine (EBM)

- EBM 학습 과정 - 1단계 : 개별 입력 변수만을 고려한 학습

✓ 기존 개별모델들은 추론 과정에서 사용하지 않기 때문에 메모리를 위해 삭제함

$$g(E(y)) = \beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j)$$



안에 입력 변수 별 값들이 존재

Explainable Boosting Machine (EBM)

❖ Explainable Boosting Machine (EBM)

- EBM 학습 과정 - 1단계 : 개별 입력 변수만을 고려한 학습

$$g(E(y)) = \beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j)$$

- ✓ 기존 개별모델들은 추론 과정에서 사용하지 않기 때문에 메모리를 위해 삭제함



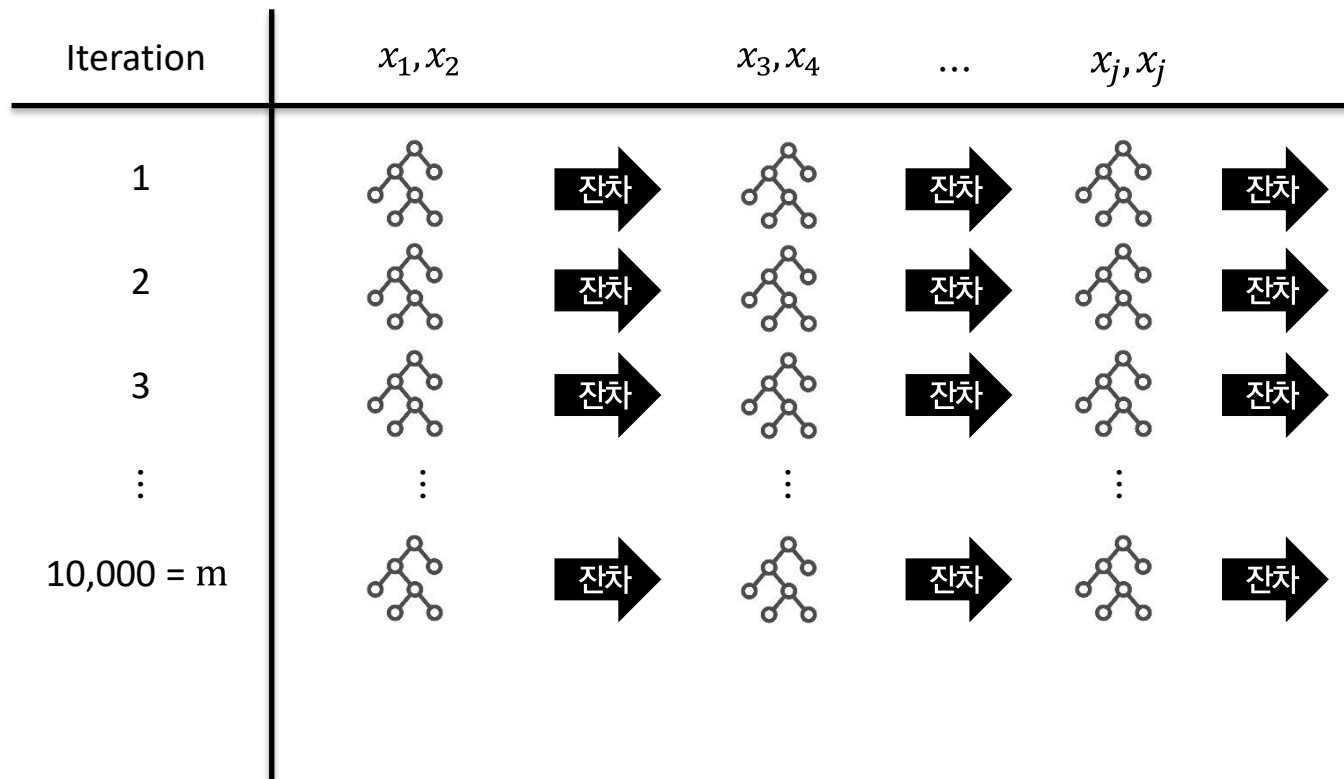
Explainable Boosting Machine (EBM)

❖ Explainable Boosting Machine (EBM)

- EBM 학습 과정 - 2단계 : 입력 변수간 상호 작용을 고려한 학습

$$g(E(y)) = \beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j)$$

- ✓ 잔차(=실제 출력 변수와 예측 출력 변수의 차)를 기반 모델 학습을 업데이트



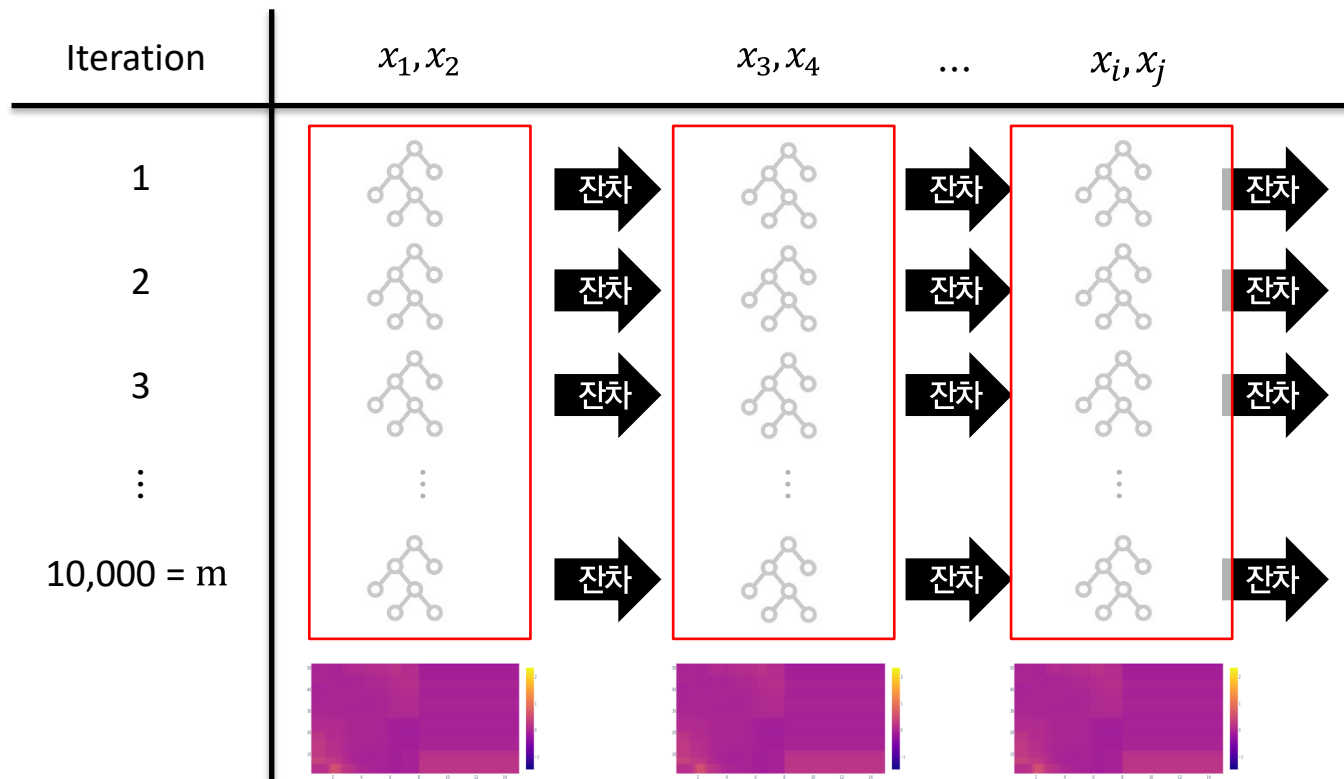
Explainable Boosting Machine (EBM)

❖ Explainable Boosting Machine (EBM)

- EBM 학습 과정 - 2단계 : 입력 변수간 상호 작용을 고려한 학습

$$g(E(y)) = \beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j)$$

- ✓ 잔차(=실제 출력 변수와 예측 출력 변수의 차)를 기반 모델 학습을 업데이트



Explainable Boosting Machine (EBM)

❖ Explainable Boosting Machine (EBM)

- EBM 학습 과정 - 2단계 : 입력 변수간 상호 작용을 고려한 학습 $g(E(y)) = \beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j)$
 - ✓ 잔차(=실제 출력 변수와 예측 출력 변수의 차)를 기반 모델 학습을 업데이트

Iteration	x_1, x_2	x_3, x_4	...	x_i, x_j
1				
2				
3				
⋮				
10,000 = m				

Explainable Boosting Machine (EBM)

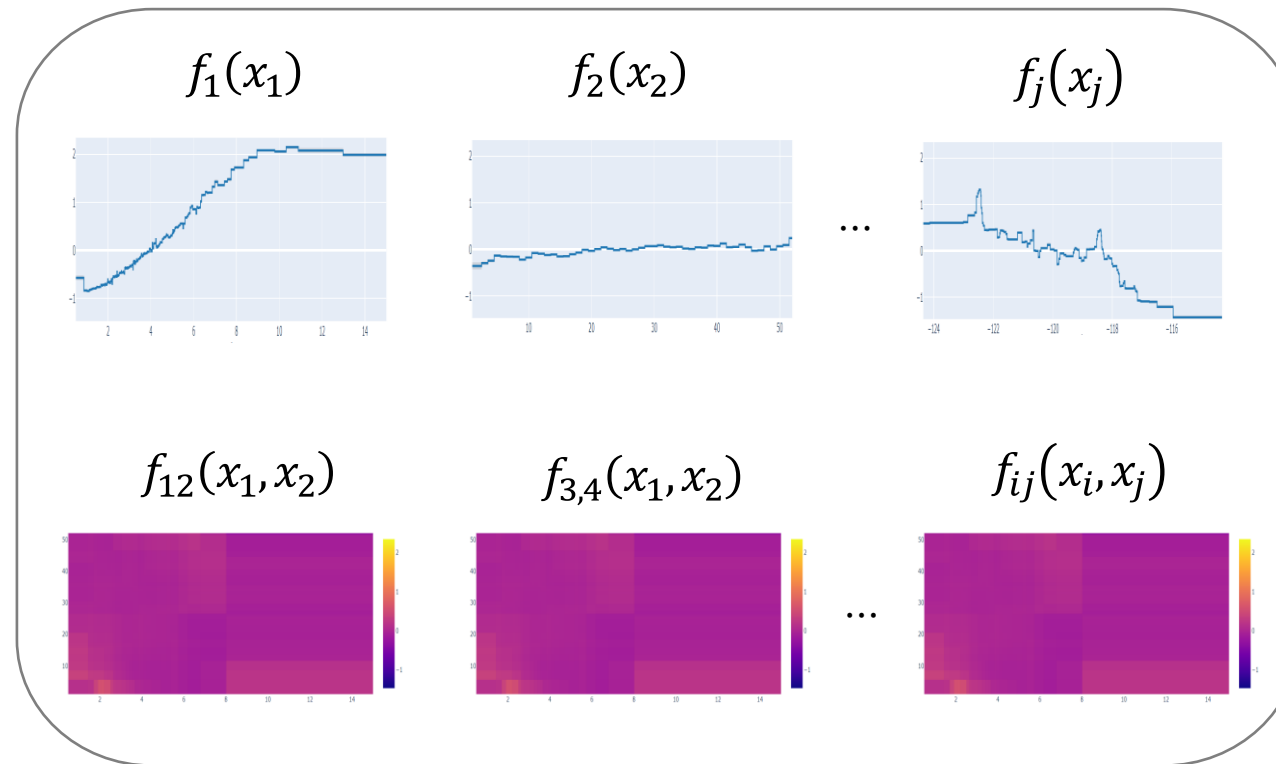
❖ Explainable Boosting Machine (EBM)

- EBM 학습 과정 - 최종 단계 : 한 개 표에 모든 값들을 저장

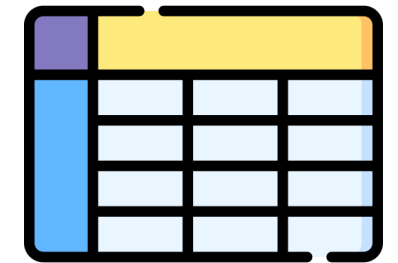
$$g(E(y)) = \beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j)$$

- ✓ 추론 과정(= Test data 예측)은 Lookup table 값을 기반으로 출력 변수를 예측 함

학습 과정 1단계



학습 과정 2단계



Lookup table

Explainable Boosting Machine (EBM)

❖ Explainable Boosting Machine (EBM)

- 코드 시현
 - ✓ 기존 머신러닝 패키지(ex. 사이킷런)들과 동일하게 간단하게 학습 및 평가가 가능함 (예시 : 집 값 예측-회귀)

```
from interpret.glassbox import ExplainableBoostingRegressor

ebm = ExplainableBoostingRegressor()
ebm.fit(X_train, y_train)

print(f'R2={r2_score(y_test, ebm.predict(X_test)):.2f}')
```

✓ 9.6s

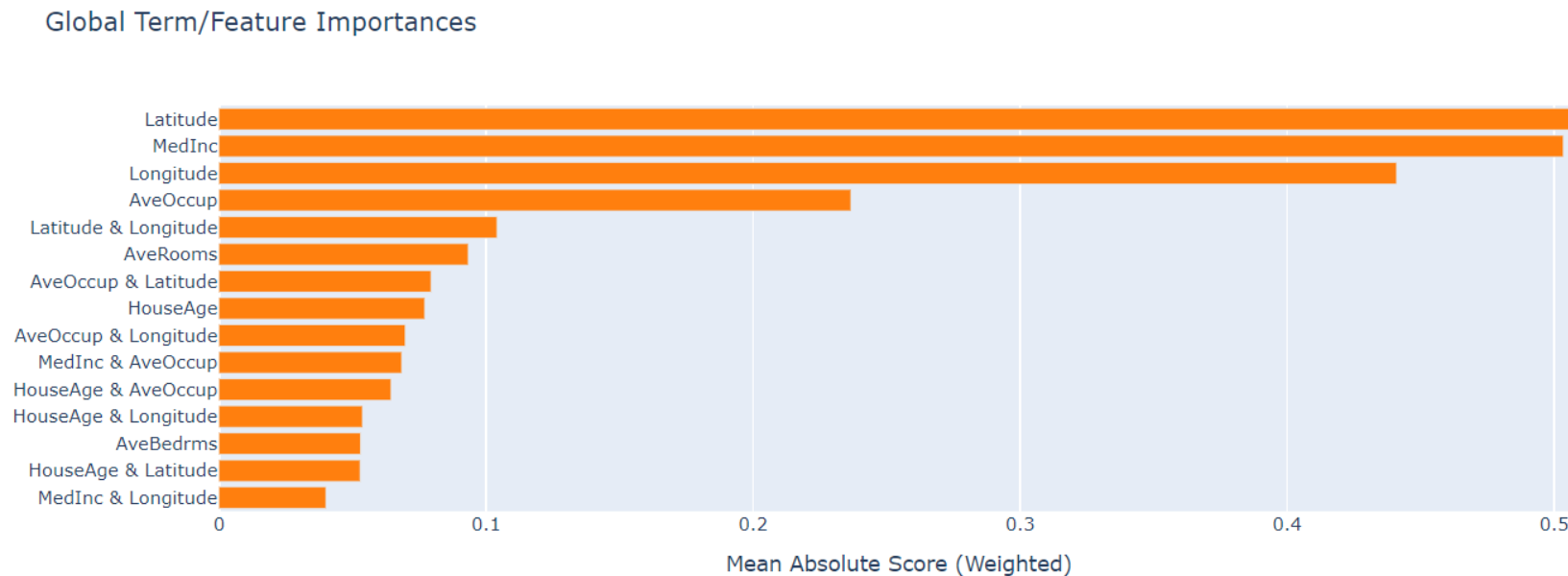
R2=0.81

〈EBM 학습 및 추론 코드〉

Explainable Boosting Machine (EBM)

❖ Explainable Boosting Machine (EBM)

- 시각화 결과
 - ✓ 학습한 모델에 대한 변수에 따른 기여도를 시각화 가능하게 제공



〈 입력 변수 별 기여점수 〉

Explainable Boosting Machine (EBM)

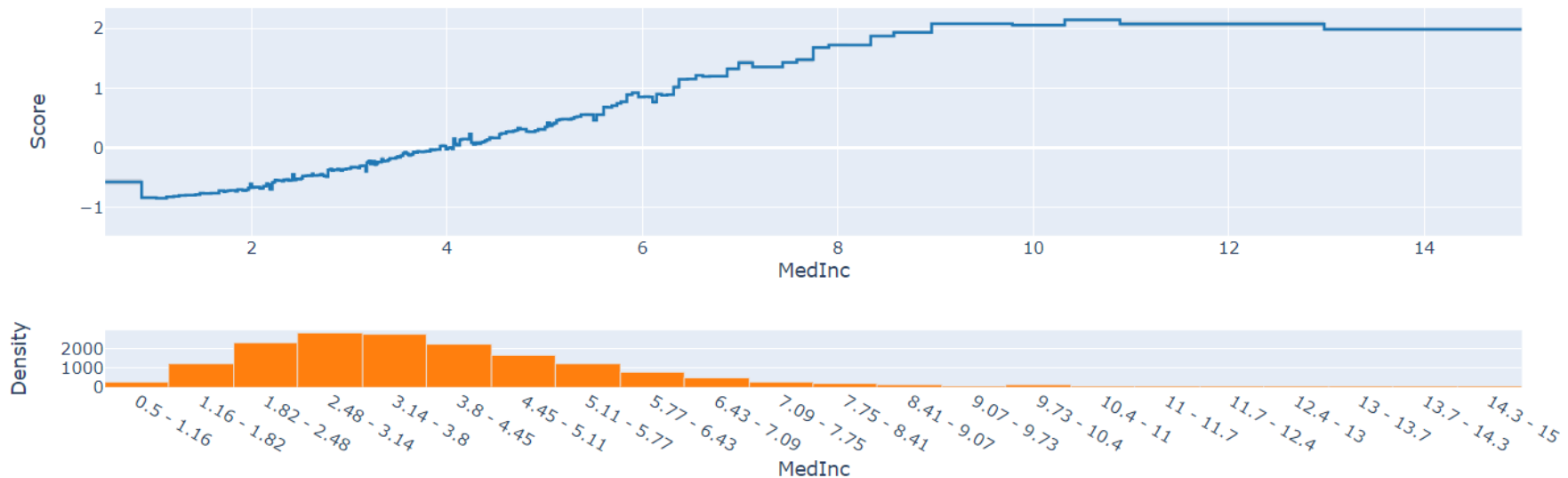
❖ Explainable Boosting Machine (EBM)

- 시각화 결과
 - ✓ 학습한 모델에 대한 변수에 따른 기여도를 시각화 가능하게 제공

$$g(E(y)) = \beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j)$$

$\searrow$
 $f_{MedInc}(x_{MedInc})$

Term: MedInc (continuous)



〈 MedInc 변수(소득) 변화에 따른 출력변수 기여도 〉

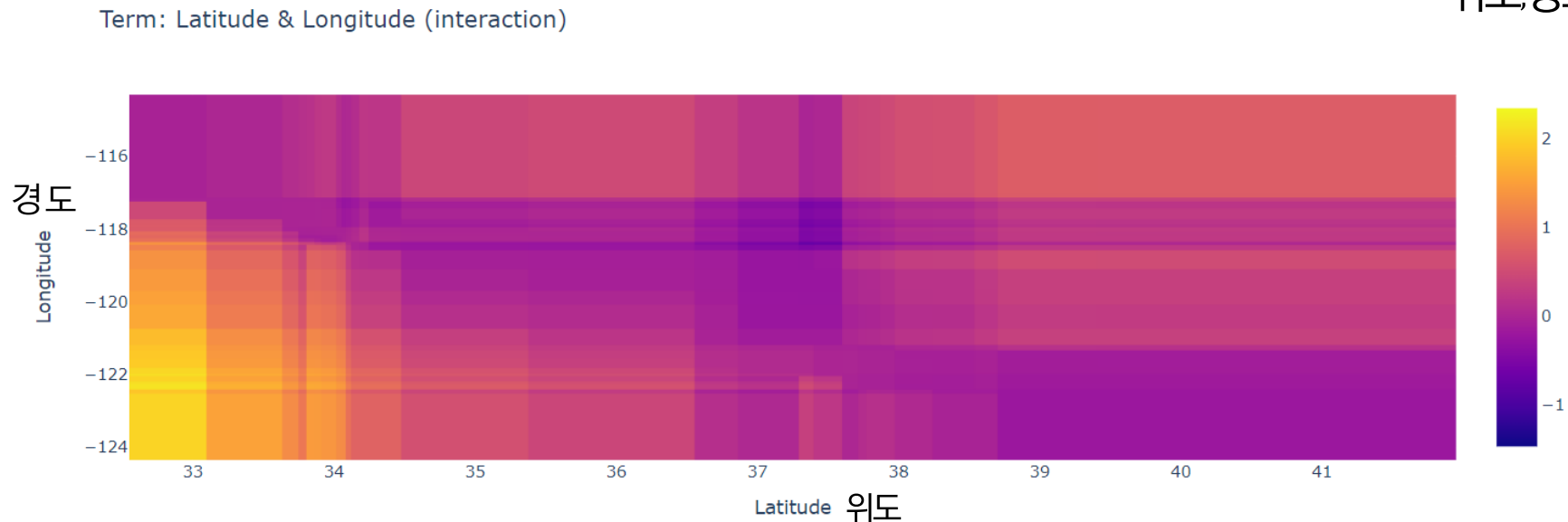
Explainable Boosting Machine (EBM)

❖ Explainable Boosting Machine (EBM)

- 시각화 결과
 - ✓ 학습한 모델에 대한 변수에 따른 기여도를 시각화 가능하게 제공

$$g(E(y)) = \beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j)$$

$f_{\text{위도,경도}}(x_{\text{위도}}, x_{\text{경도}})$



〈 두 개 입력변수(위도, 경도)간 상호작용을 통한 출력변수의 기여도〉

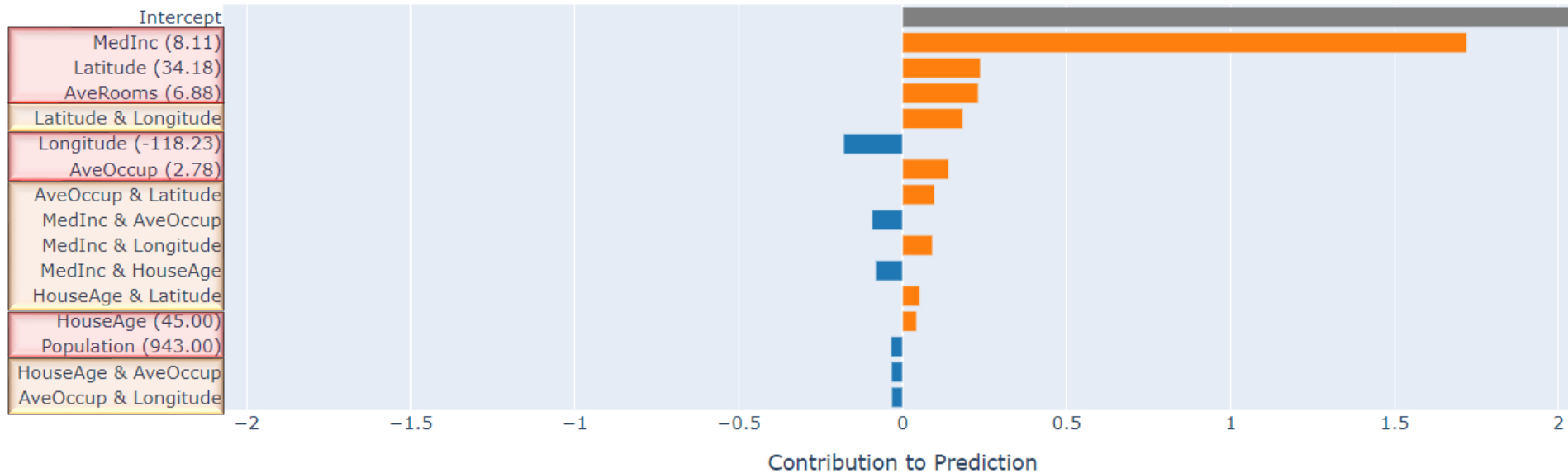
Explainable Boosting Machine (EBM)

❖ Explainable Boosting Machine (EBM)

- 시각화 결과
 - ✓ Test sample 예측 결과에 대한 변수 별 기여도를 시각화 가능하게 제공

$$g(E(y)) = \underset{\text{Intercept}}{\beta_0} + \sum f_j(x_j) + \sum f_{ij}(x_i x_j)$$

Local Explanation (Actual: 4.47 | Predicted: 4.4)



〈 예측 값 4.4에 대한 입력변수 별 기여도 〉

Conclusion

Conclusion

❖ Conclusion

- 기존 머신러닝의 문제
 - ✓ 기존 높은 성능을 내는 머신러닝은 복잡한 구조로 모델의 의사결정 과정을 설명하기 어려운 문제가 존재
 - ✓ 이를 해결하기 위한 XAI 기법(ex. LIME, SHAP) 등이 존재하지만, 예측 모델과 해석 모델을 분리하는 형식으로 간접적인 추론 한계가 남아있었음
- EBM
 - ✓ 부스팅 모델과 동등한 성능을 가지며, 모델 의사결정 과정을 설명 가능한 EBM을 소개
 - ✓ EBM은 가산성을 유지하기 때문에 개별 변수 및 변수 간 상호 작용 기여도를 확인 가능함
 - ✓ 간단한 파이썬 패키지로 제공되고, 예측에 대한 직관적 설명이 가능하기에 모델 의사결정 과정이 중요한 분야에서 많이 사용 될 것으로 예상 됨

References

1. Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223.*
2. Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013, August). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 623–631).
3. Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy, 23(1), 18.*
4. <https://www.youtube.com/watch?v=y6ajhgaT9MA&t=2314s>

Thank you