DMQA Open Seminar

Explainable Machine Learning with Cyclic Boosting Machines (CBM)

2024-12-13

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

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

❖ 연락처

- ssj259@korea.ac.kr

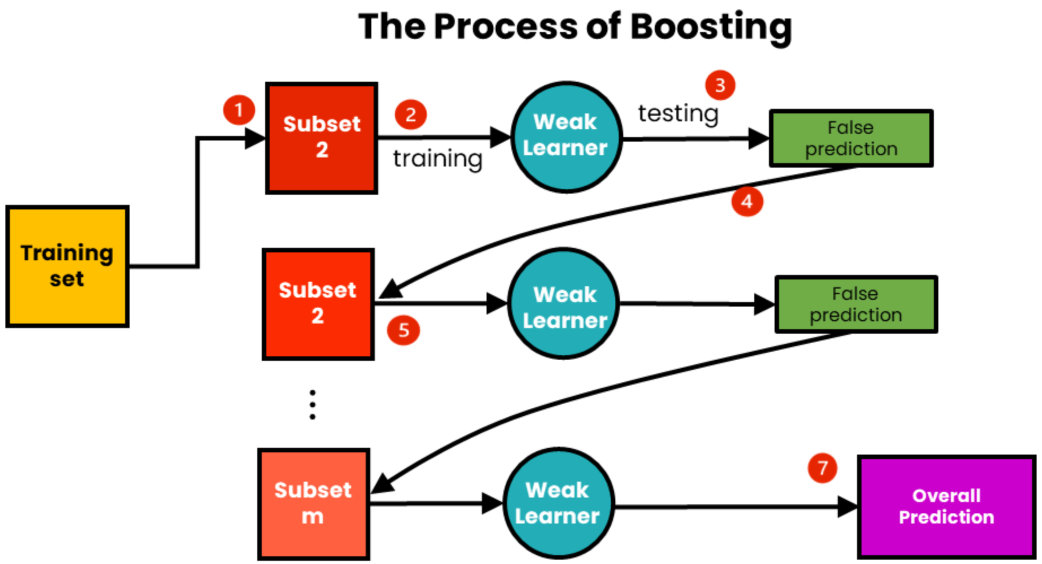
Contents

Introduction

Introduction

❖ 기존 연구의 한계점 – 머신러닝

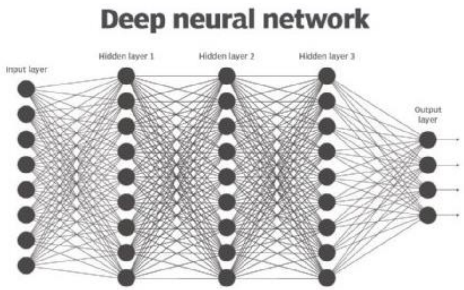
- 머신러닝은 특히 표 형식(tabular) 데이터를 다루는 경우 높은 성능을 보임
- 우수한 성능을 보이는 부스팅 계열의 머신러닝 모델들(LGBM, XGB 등)은 복잡한 구조로 <u>모델 의사결정 및 결과에 대해 명확히 이해하기 어려움</u>

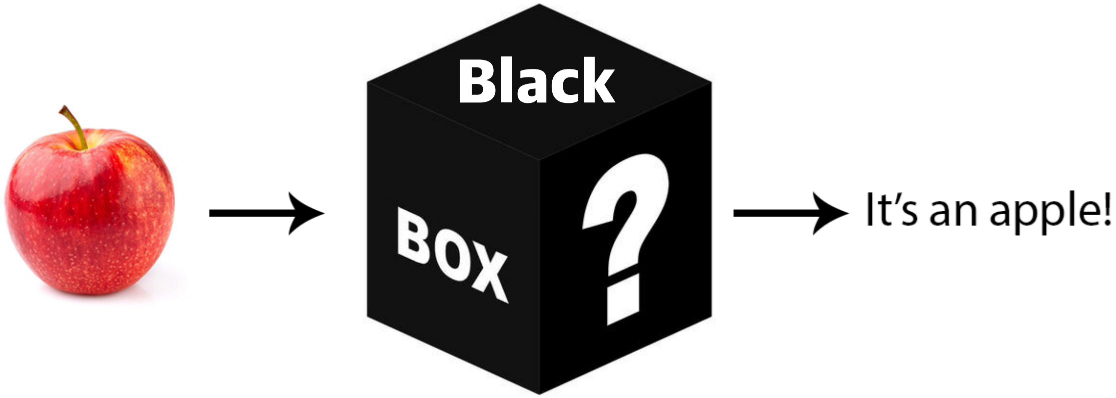


Introduction

❖ 기존 연구의 한계점 – 딥러닝

- 딥러닝은 이미지, 텍스트과 같은 데이터에서 머신러닝보다 높은 성능을 보이지만, 학습의 복잡성과 비선형성으로 <u>모델 자체 의사결정 과정 및 결과에 대한 해석이 매우 어려움 (Black box)</u>

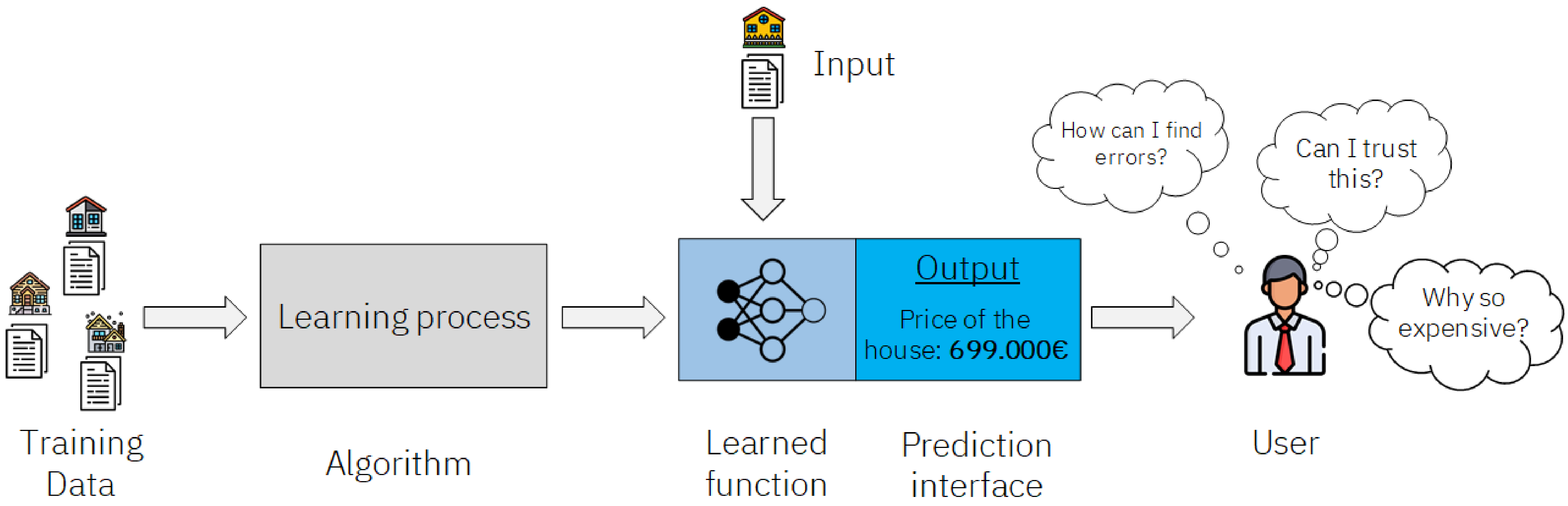




Introduction

❖ 설명 가능성의 중요성

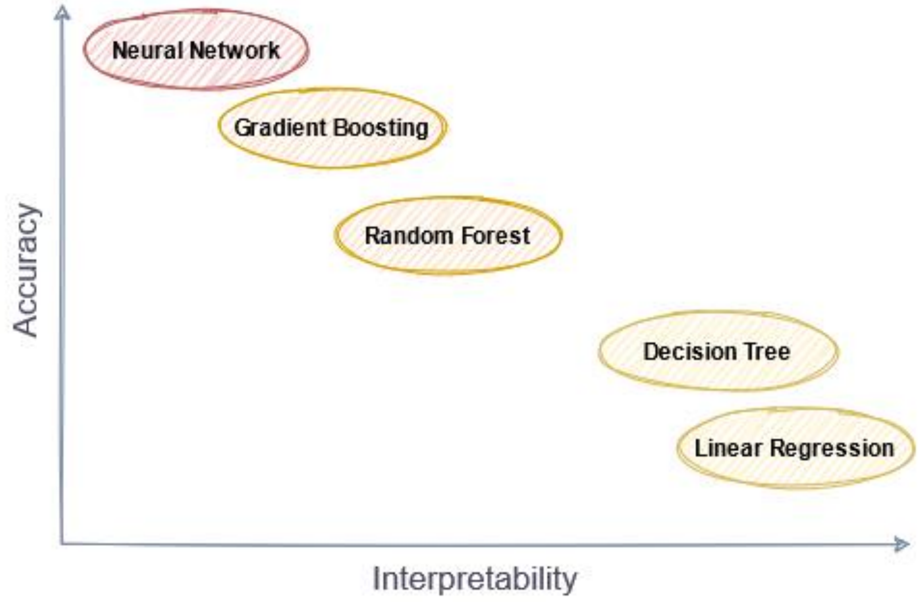
- 현업에서는 모델 의사결정 과정에 대한 근거를 필요로 하며, 의료와 금융 분야에서는 성능보다 설명 가능성을 더 중요하게 보는 경우가 많음(ex. 의사결정나무를 아직도 많이 사용)



Introduction

❖ 설명 가능성의 중요성

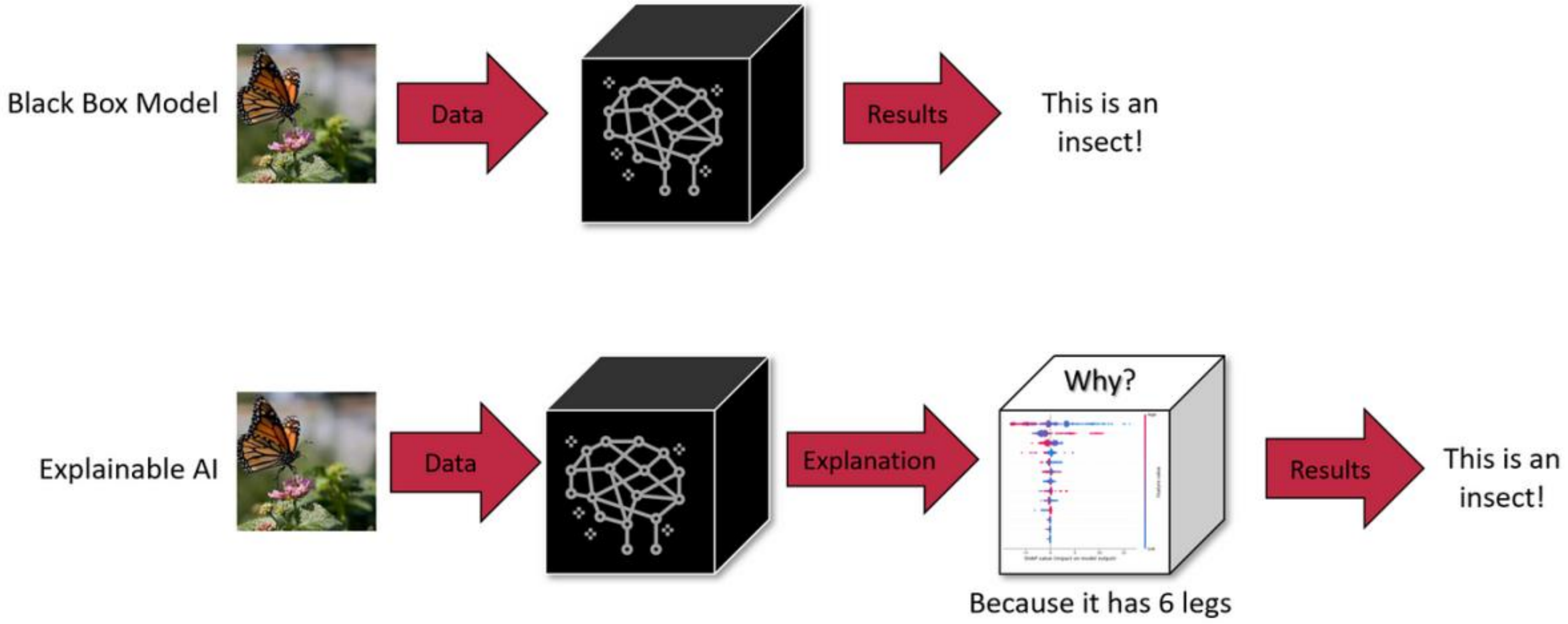
- 모델의 성능과 설명 가능성 사이에는 역의 상관관계가 존재함 (높은 성능 vs 높은 해석력)
- 해석이 **쉬운 모델** : Linear Regression, Decision tree 등 해석이 쉽지만, 정확도가 낮아 활용도가 낮음
- 해석이 어려운 모델 : Ensemble 및 Boosting, Neural Network 등 성능이 높지만, 모델 자체 해석력이 낮음



Introduction

❖ 설명 가능성의 중요성 – XAI 기법

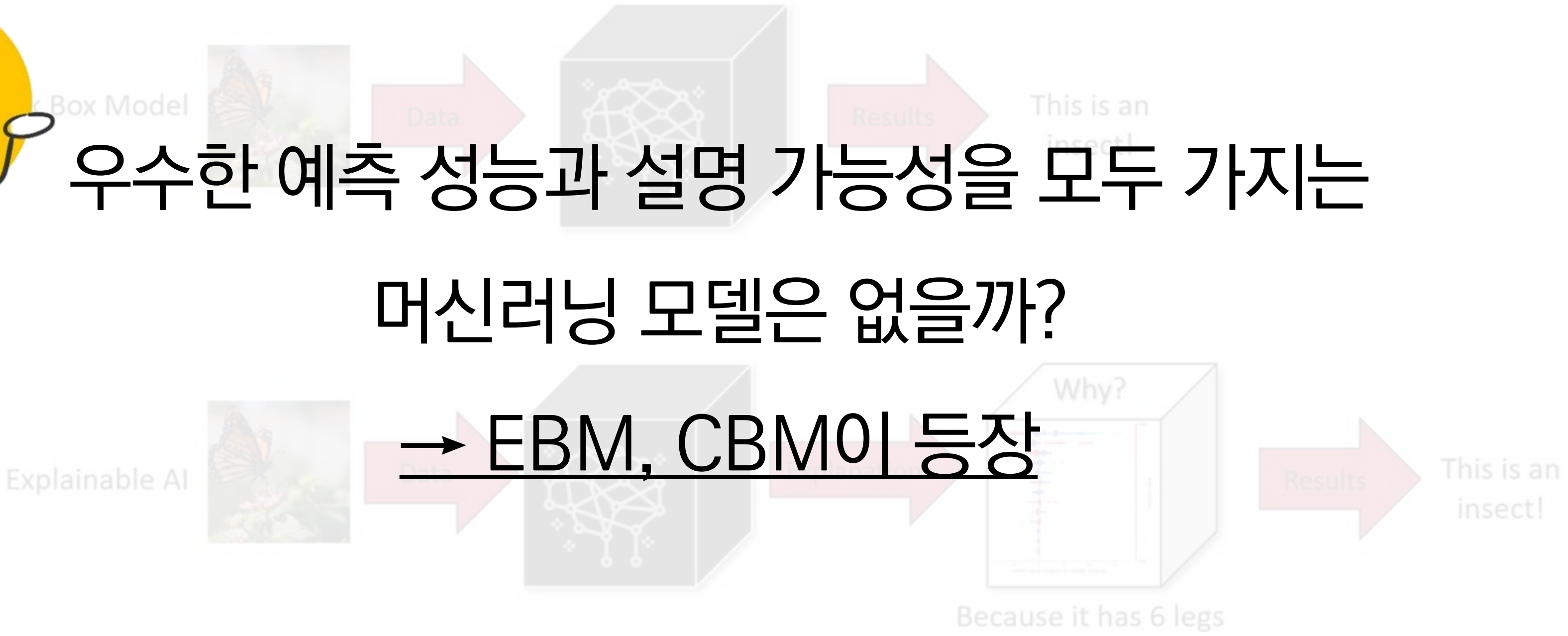
- 해석이 어려운 모델의 예측 결과를 해석하기 위해 설명 가능한 XAI 기법들(ex. SHAP, LIME)이 제안 됨
- 예측 모델과 해석 모델이 분리되어 있어, 모델의 본질적인 작동 방식을 이해하기 어려움 → 간접적으로 추론하는 한계



Introduction

❖ 설명 가능성의 중요성 – XAI 기법

- 해석이 어려운 모델의 예측 결과를 해석하기 위해 설명 가능한 XAI 기법들(ex. SHAP, LIME)이 제안 됨
- 예측 모델과 해석 모델이 분리되어 있어, 모델의 본질적인 작동 방식을 이해하기 어려움 → 간접적으로 추론하는 한계



우수한 예측 성능과 설명 가능성을 모두 가지는

머신러닝 모델은 없을까?

<u>→ EBM, CBM이 등장</u>

Introduction

❖ 관련 세미나 소개

XAI 기법 (SHAP, LIME) 관련

설명가능한 머신러닝 – EBM 관련

종료

DMQA open seminar

Applications of eXplainable Artificial Intelligence(XAI)

Data Mining & Quality Analytics Lab

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)

DMQA Open Seminar

2021.09.03

오혜령

Introduction to XAI(eXplainable AI)

발표자: **오혜령**

2021년 9월 3일

오후 1시 ~

온라인 비디오 시청 (YouTube)

세미나 정보 보기 ⟶

종료

Introduction to Explainable Boosting Machine (EBM)

2023-08-11

Korea University

Data Mining & Quality Analytics Lab.

심세진

Introduction to Explainable Boosting Mac

발표자: **심세진**

2023년 8월 11일

오전 12시 ~

온라인 비디오 시청 (YouTube)

세미나 정보 보기 ⟶

Cyclic Boosting Machines (CBM)

Cyclic Boosting Machines (CBM)

❖ Cyclic Boosting – an explainable supervised machine learning algorithm

- 2019년 마이크로 소프트에서 설명가능한 머신러닝인 CBM을 소개함 (2019 ICMLA, 현재기준 8회 인용)
- CBM은 판매 수요량을 예측하는 회귀 데이터셋에서 준수한 성능을 보이며, 해석을 제공하는 머신러닝 모델임

Cyclic Boosting - an explainable supervised machine learning algorithm

Felix Wick *1, Ulrich Kerzel †2, and Michael Feindt ‡1

[1] Blue Yonder GmbH (Karlsruhe, Germany)
[2] IUBH Internationale Hochschule (Erfurt, Germany)

Supervised machine learning algorithms have seen spectacular advances and surpassed human level performance in a wide range of specific applications. However, using complex ensemble or deep learning algorithms typically results in black box models, where the path leading to individual predictions cannot be followed in detail. In order to address this issue, we propose the novel "Cyclic Boosting" machine learning algorithm, which allows to efficiently perform accurate regression and classification tasks while at the same time allowing a detailed understanding of how each individual prediction was made.

Keywords: **machine learning, explainable algorithms, demand forecasting**

DOI: 10.1109/ICMLA.2019.00067

In order to address both the explainability and handling of rare effects, a novel machine learning algorithm called "Cyclic Boosting" is proposed, which is able to learn a model $\hat{p}(Y|\boldsymbol{X}, \vec{\theta})$ efficiently and accurately, while allowing to precisely follow the path how individual predictions were made.

As will become apparent below, Cyclic Boosting can be categorized as a generalized additive model [2], where the target Y belongs to the family of exponential distributions such as the Poisson, Gaussian, or Bernoulli distribution, together with a suitable link function. As such, the Cyclic Boosting algorithm can be used in the following three scenarios:

- Multiplicative regression mode: $Y \in [0, \infty)$

Cyclic Boosting Machines (CBM)

❖ Before CBM

- CBM은 일반화 가법 모델(Generalized Additive Model, GAM) 중 하나로 분류될 수 있음
- GAM?
  - ✓ **회귀 분석의 확장된 형태**로 데이터의 비선형 관계를 설명하기 위해 고안된 모델
  - ✓ 연산 값의 합을 의사결정에 활용하기 때문에 **각 변수가 어떤 의사결정에 영향을 미치는지 파악하기 쉬움**

  - ✓ GAM 수식 (CBM의 기초가 됨)

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

    - ➢ $g(\cdot)$= Link Function(= 회귀 or 분류)
    - ➢ $f(\cdot)$= 변수 x_j에 적용되는 **비선형 함수**(= ex. spline etc.)
    - ➢ β_0= **절편** (모델의 상수 값)

Cyclic Boosting Machines (CBM)

❖ CBM 상황별

- Y(= target 변수)의 종류에 따른 3가지 경우
  1. 곱셈 회귀 모드 – Multiplicative regression mode: $Y \in [0, \infty)$
     - ➢ $\hat{y} = \mu \cdot \prod f_j^k$
     - ➢ 판매량: 특정 상품의 하루 판매량(0개 또는 그 이상)
     - ➢ 이벤트 빈도: 특정 시간 내에 발생한 사건 수
  2. 덧셈 회귀 모드 – Additive regression mode: $Y \in (-\infty, \infty)$
     - ➢ $\hat{y} = \mu + \sum f_j^k$
     - ➢ 주식 변화율: 하루 주식 가격 변화(증가 또는 감소)
     - ➢ 온도 예측: 온도 변화가 음수(-10도)나 양수(30도)일 수 있음.
  3. 분류 모드 – Classification mode: $Y \in [0, 1]$
     - ➢ $\frac{\hat{p}}{1-\hat{p}} = \mu \cdot \prod f_j^k$
     - ➢ 스팸 이메일 분류: 이메일이 스팸인지 아닌지를 예측(0: 스팸 아님, 1: 스팸)
     - ➢ 병리학적 진단: 환자가 질병에 걸렸는지 여부(0: 없음, 1: 있음).

Cyclic Boosting Machines (CBM)

❖ CBM 상황별

- Y(= target 변수)의 종류에 따른 3가지 경우
  1. 곱셈 회귀 모드 – Multiplicative regression mode: $Y \in [0, \infty)$
     - ➢ $\hat{y} = \mu \cdot \prod f_j^k$
     - ➢ 판매량: 특정 상품의 하루 판매량(0개 또는 그 이상)
     - ➢ 이벤트 빈도: 특정 시간 내에 발생한 사건 수
  2. 덧셈 회귀 모드 – Additive regression mode: $Y \in (-\infty, \infty)$
     - ➢ $\hat{y} = \mu + \sum f_j^k$
     - ➢ 주식 변화율: 하루 주식 가격 변화(증가 또는 감소)
     - ➢ 온도 예측: 온도 변화가 음수(-10도)나 양수(30도)일 수 있음.
  3. 분류 모드 – Classification mode: $Y \in [0, 1]$
     - ➢ $\frac{\hat{p}}{1-\hat{p}} = \mu \cdot \prod f_j^k$
     - ➢ 스팸 이메일 분류: 이메일이 스팸인지 아닌지를 예측(0: 스팸 아님, 1: 스팸)
     - ➢ 병리학적 진단: 환자가 질병에 걸렸는지 여부(0: 없음, 1: 있음).

Cyclic Boosting Machines (CBM)

❖ 곱셈 회귀 모드 – Multiplicative regression mode: $Y \in [0,\infty)$

- y와 x간의 관계를 곱셈형태로 표현
- 타겟 변수 : 빈도, 카운트
- 수식 → 어떻게 도출 되는 걸까?

$$\hat{y}_i = \mu \cdot \prod_{j=1}^{p} f_j^k \quad \text{with } k = \{x_{j,i} \in b_j^k\}$$

  - ➢ μ = y의 평균
  - ➢ f_j^k= j번째 변수와 k번째 bin(=구간)에 해당되는 모델 파라미터

Cyclic Boosting Machines (CBM)

❖ CBM 구조 – 곱셈 회귀 모드 : GAM에서 링크 함수를 통해 CBM으로 연결 됨

- GAM 수식

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

링크 함수 g를 로그 ln로 대입

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

곱셈적 구조를 위해 양변에 지수 exp 를 적용

지수의 성질($\exp(a+b) = \exp(a) * \exp(b)$) 사용

$$E(y) = \exp\left(\beta_0 + \sum f_j(x_j)\right) = \exp(\beta_0) \cdot \prod \exp\left(f_j(x_j)\right)$$

- CBM 수식

$$E(y) = \mu \cdot \prod f_j^k$$

$\exp(\beta_0)$가 y의 평균 μ 으로

$\exp\left(f_j(x_j)\right)$가 특징 j에 구간 k 로 정의된 f_j^k로

Cyclic Boosting Machines (CBM)

❖ CBM 학습 과정

- **곱셈 회귀 모드 – Multiplicative regression mode: $Y \in [0, \infty)$**
  - ✓ 모델 학습 과정
    1. 모든 y의 평균인 μ 계산
    2. j번째 변수와 k번째 bin(=구간)에 해당되는 모델 파라미터 $f_j^k \leftarrow 1$ (초기화)
    3. 모든 변수 $j = 1 \sim p$를 순환하면서 각 구간 k에 대해 아래와 같이 계산

$$g_{j,t}^k = \frac{\sum_{x_{j,i} \in b_j^k} y_i}{\sum_{x_{j,i} \in b_j^k} \hat{y}_{i,\tau}} \quad \text{where} \quad f_{j,t}^k = \prod_{s=1}^{t} g_{j,s}^k$$

$$\hat{y}_{i,\tau} = \mu \cdot \prod_{j=1}^{p} f_{j,\tau}^k$$

    4. 모든 변수에 대해 순환한 뒤, 기준(ex. 반복 횟수 등)이 충족되면 종료

Cyclic Boosting Machines (CBM)

❖ CBM 학습 과정 – 1.

- 곱셈 회귀 모드 – Multiplicative regression mode: $Y \in [0, \infty)$

✓ 예시

Index(=i)	X_1(나이)	X_2(소득수준)	Y(월 지출액 $)
1	20	많음	500
2	40	중간	300
3	30	적음	400
4	50	많음	600
			μ = 450

✓ 모델 학습 과정

1. <u>모든 y의 평균인 μ 계산</u>
2. j번째 변수와 k번째 bin(=구간)에 해당되는 모델 파라미터 $f_k^j \leftarrow 1$ (초기화)
3. 모든 변수 $j = 1 \sim p$를 순환하면서 각 구간 k에 대해 아래와 같이 계산
4. 모든 변수에 대해 순환한 뒤, 기준(ex. 반복 횟수 등)이 충족되면 종료

Cyclic Boosting Machines (CBM)

❖ CBM 학습 과정 – 2.

- 곱셈 회귀 모드 – Multiplicative regression mode: $Y \in [0,\infty)$

✓ 예시

Index(= i)	X_1(나이)	X_2(소득수준)	Y(월 지출액 $)
1	20	많음	500
2	40	중간	300
3	30	적음	400
4	50	많음	600

구간 1 = 0~20
구간 2 = 21~40
구간 3 = 41~60

구간 1 = 많음
구간 2 = 중간
구간 3 = 적음

μ = 450

$$f_1^1 = 1 \quad f_1^2 = 1 \quad f_1^3 = 1$$

$$f_2^1 = 1 \quad f_2^2 = 1 \quad f_2^3 = 1$$

✓ 모델 학습 과정

1. 모든 y의 평균인 μ 계산
2. j번째 변수와 k번째 bin(=구간)에 해당되는 모델 파라미터 $f_k^j \leftarrow 1$ (초기화)
3. 모든 변수 $j = 1 \sim p$를 순환하면서 각 구간 k에 대해 아래와 같이 계산
4. 모든 변수에 대해 순환한 뒤, 기준(ex. 반복 횟수 등)이 충족되면 종료

Cyclic Boosting Machines (CBM)

❖ CBM 학습 과정 – 3.

- 곱셈 회귀 모드 – Multiplicative regression mode: $Y \in [0,\infty)$

✓ 예시

Index(= i)	X_1(나이)	X_2(소득수준)	Y(월 지출액 $)
1	20	많음	500
2	40	중간	300
3	30	적음	400
4	50	많음	600

구간 1 = 1~20
구간 2 = 21~40
구간 3 = 41~60

μ = 450

➢ X_1에 대해

$$f_1^1 = 1$$
$$f_1^2 = 1$$
$$f_1^3 = 1$$

구간 2

$$g_{1,t}^2 = \frac{300 + 400}{450 + 450} \approx 0.78$$

$$\hat{y}_2 = 450 \cdot 1 \cdot 1 = 450$$

$$\hat{y}_3 = 450 \cdot 1 \cdot 1 = 450$$

구간 1, 3

$$g_{1,t}^1 = \frac{500}{450} \approx 1.11$$

$$g_{1,t}^3 = \frac{600}{450} \approx 1.33$$

✓ 모델 학습 과정

1. 모든 y의 평균인 μ 계산
2. j번째 변수와 k번째 bin(=구간)에 해당되는 모델 파라미터 $f_k^j \leftarrow 1$ (초기화)
3. 모든 변수 $j = 1 \sim p$를 순환하면서 각 구간 k에 대해 아래와 같이 계산

$$g_{j,t}^k = \frac{\sum_{x_{j,i} \in b_j^k} y_i}{\sum_{x_{j,i} \in b_j^k} \hat{y}_{i,\tau}} \quad \text{where} \quad f_{j,t}^k = \prod_{s=1}^{t} g_{j,s}^k$$

$$\hat{y}_{i,\tau} = \mu \cdot \prod_{j=1}^{p} f_{j,\tau}^k$$

Cyclic Boosting Machines (CBM)

❖ CBM 학습 과정 – 3.

- 곱셈 회귀 모드 – Multiplicative regression mode: $Y \in [0, \infty)$

✓ 예시

Index(=i)	X_1(나이)	X_2(소득수준)	Y(월 지출액 $)
1	20	많음	500
2	40	중간	300
3	30	적음	400
4	50	많음	600

구간 1 = 1~20
구간 2 = 21~40
구간 3 = 41~60

μ = 450

➢ X_1에 대해

$$f_1^1 = 1$$
$$f_1^2 = 1$$
$$f_1^3 = 1$$

구간 2

$$f_{1,t}^2 = f_1^2 \cdot g_{1,t}^2 = 1 \cdot 0.78$$

구간 1, 3

$$f_{1,t}^1 = f_1^1 \cdot g_{1,t}^1 = 1 \cdot 1.11$$
$$f_{1,t}^3 = f_1^3 \cdot g_{1,t}^3 = 1 \cdot 1.33$$

✓ 모델 학습 과정

1. 모든 y의 평균인 μ 계산
2. j번째 변수와 k번째 bin(=구간)에 해당되는 모델 파라미터 $f_k^j \leftarrow 1$ (초기화)
3. 모든 변수 $j = 1 \sim p$를 순환하면서 각 구간 k에 대해 아래와 같이 계산

$$g_{j,t}^k = \frac{\sum_{x_{j,i} \in b_j^k} y_i}{\sum_{x_{j,i} \in b_j^k} \hat{y}_{i,\tau}} \quad \text{where} \quad f_{j,t}^k = \prod_{s=1}^{t} g_{j,s}^k$$

$$\hat{y}_{i,\tau} = \mu \cdot \prod_{j=1}^{p} f_{j,\tau}^k$$

Cyclic Boosting Machines (CBM)

❖ CBM 학습 과정 – 3.

- 곱셈 회귀 모드 – Multiplicative regression mode: $Y \in [0,\infty)$

✓ 예시

Index(= i)	X_1(나이)	X_2(소득수준)	Y(월 지출액 $)
1	20	많음	500
2	40	중간	300
3	30	적음	400
4	50	많음	600

구간 1 = 많음
구간 2 = 중간
구간 3 = 적음

μ = 450

➢ X_2에 대해

$$f_2^1 = 1 \quad f_2^2 = 1 \quad f_2^3 = 1$$

구간 1

$$g_{2,t}^1 = \frac{500+600}{450+450} \approx 1.22$$

$$\hat{y}_1 = 450 \cdot 1 \cdot 1 = 450$$

$$\hat{y}_4 = 450 \cdot 1 \cdot 1 = 450$$

구간 2, 3

$$g_{2,t}^2 = \frac{300}{450} \approx 0.67$$

$$g_{2,t}^3 = \frac{400}{450} \approx 0.89$$

✓ 모델 학습 과정

1. 모든 y의 평균인 μ 계산
2. j번째 변수와 k번째 bin(=구간)에 해당되는 모델 파라미터 $f_k^j \leftarrow 1$ (초기화)
3. 모든 변수 $j = 1 \sim p$를 순환하면서 각 구간 k에 대해 아래와 같이 계산

$$g_{j,t}^k = \frac{\sum_{x_{j,i} \in b_j^k} y_i}{\sum_{x_{j,i} \in b_j^k} \hat{y}_{i,\tau}} \quad \text{where} \quad f_{j,t}^k = \prod_{s=1}^{t} g_{j,s}^k$$

$$\hat{y}_{i,\tau} = \mu \cdot \prod_{j=1}^{p} f_{j,\tau}^k$$

Cyclic Boosting Machines (CBM)

❖ CBM 학습 과정 – 3.

- 곱셈 회귀 모드 – Multiplicative regression mode: $Y \in [0,\infty)$

✓ 예시

Index(= i)	X_1(나이)	X_2(소득수준)	Y(월 지출액 \$)
1	20	많음	500
2	40	중간	300
3	30	적음	400
4	50	많음	600

구간 1 = 많음
구간 2 = 중간
구간 3 = 적음

μ = 450

➢ X_2에 대해

$$f_2^1 = 1$$
$$f_2^2 = 1$$
$$f_2^3 = 1$$

구간 1

$$g_{2,t}^1 = \frac{500+600}{450+450} \approx 1.22$$

$$f_{2,t}^1 = f_2^1 \cdot g_{2,t}^1 = 1 \cdot 1.22$$

$$\hat{y}_3 = 450 \cdot 1 \cdot 1 = 450$$

구간 2, 3

$$g_{2,t}^2 = \frac{300}{450} \approx 0.67$$

$$f_{2,t}^2 = f_2^2 \cdot g_{2,t}^2 = 1 \cdot 0.67$$
$$f_{2,t}^3 = f_2^3 \cdot g_{2,t}^3 = 1 \cdot 0.89$$

$$g_{2,t}^3 = \frac{\quad}{450} \approx 0.89$$

✓ 모델 학습 과정

1. 모든 y의 평균인 μ 계산
2. j번째 변수와 k번째 bin(=구간)에 해당되는 모델 파라미터 $f_k^j \leftarrow 1$ (초기화)
3. <u>모든 변수 $j = 1 \sim p$를 순환하면서 각 구간 k에 대해 아래와 같이 계산</u>

$$g_{j,t}^k = \frac{\sum_{x_{j,i} \in b_j^k} y_i}{\sum_{x_{j,i} \in b_j^k} \hat{y}_{i,\tau}} \quad \text{where} \quad f_{j,t}^k = \prod_{s=1}^{t} g_{j,s}^k$$

$$\hat{y}_{i,\tau} = \mu \cdot \prod_{j=1}^{p} f_{j,\tau}^k$$

Cyclic Boosting Machines (CBM)

❖ CBM 학습 과정 – 3.

- 곱셈 회귀 모드 – Multiplicative regression mode: $Y \in [0,\infty)$

✓ 예시

Index(= i)	X_1(나이)	X_2(소득수준)	Y(월 지출액 $)
1	20	많음	500
2	40	중간	300
3	30	적음	400
4	50	많음	600

구간 1 = 0~20
구간 2 = 21~40
구간 3 = 41~60

구간 1 = 많음
구간 2 = 중간
구간 3 = 적음

μ = 450

$f_1^1 = 1.11$
$f_1^2 = 0.78$
$f_1^3 = 1.33$

$f_2^1 = 1.22$
$f_2^2 = 0.67$
$f_2^3 = 0.89$

➢ $\hat{y}_{i,\tau}$ 계산

$$\hat{y}_{1,\tau} = 450 \cdot 1.11 \cdot 1.22 = 609.39$$

➢ τ = 현재 iteration

➢ 파라미터 f_k^j가 iteration이 늘어날 수록 따라 $\sum y_i$ 실제 / $\sum \hat{y}_i$ 예측 만큼 업데이트 됨

✓ 모델 학습 과정

1. 모든 y의 평균인 μ 계산
2. j번째 변수와 k번째 bin(=구간)에 해당되는 모델 파라미터 $f_k^j \leftarrow 1$ (초기화)
3. 모든 변수 $j = 1 \sim p$를 순환하면서 각 구간 k에 대해 아래와 같이 계산

$$g_{j,t}^k = \frac{\sum_{x_{j,i} \in b_j^k} y_i}{\sum_{x_{j,i} \in b_j^k} \hat{y}_{i,\tau}} \quad \text{where} \quad f_{j,t}^k = \prod_{s=1}^{t} g_{j,s}^k$$

$$\hat{y}_{i,\tau} = \mu \cdot \prod_{j=1}^{p} f_{j,\tau}^k$$

Cyclic Boosting Machines (CBM)

❖ CBM 학습 과정 – 4.

- 곱셈 회귀 모드 – Multiplicative regression mode: $Y \in [0, \infty)$

✓ 예시

1. 최대 반복 횟수 도달
   - ➢ 예: 10회 반복 후 종료
2. 평가 함수 개선 없음
   - ➢ MSE or MAD 개선 X 종료
   - ➢ MAD : 예측 값과 실제 값 사이의 절대 차이의 평균.

✓ **모델 학습 과정**

1. 모든 y의 평균인 μ 계산
2. j번째 변수와 k번째 bin(=구간)에 해당되는 모델 파라미터 $f_k^j \leftarrow 1$ (초기화)
3. 모든 변수 $j = 1 \sim p$를 순환하면서 각 구간 k에 대해 아래와 같이 계산
4. <u>모든 변수에 대해 순환한 뒤, 기준(ex. 반복 횟수 등)이 충족되면 종료</u>

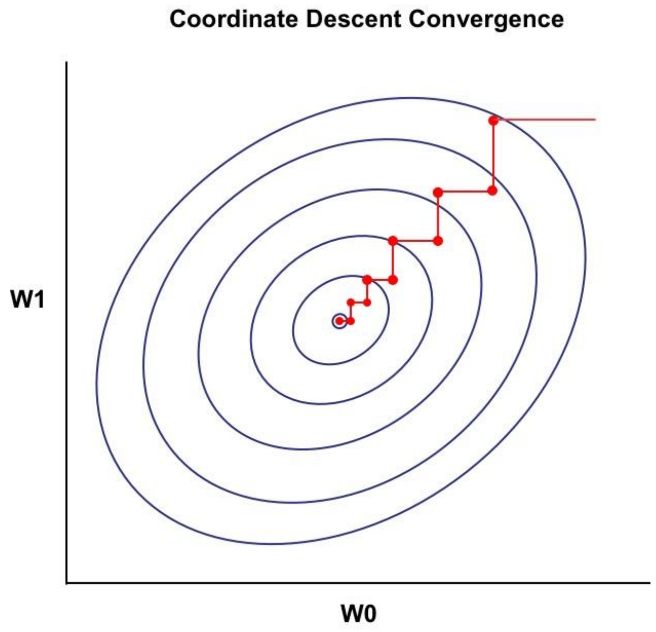
Cyclic Boosting Machines (CBM)

❖ CBM 학습 과정 – 좌표 하강 알고리즘

- 파라미터 f_k^j가 iteration이 늘어날 수록 따라 $\sum y_i$ 실제 / $\sum \hat{y}_i$ 예측 만큼 업데이트 됨

  = 좌표 하강 알고리즘(Coordinate descent)

  ➢ 최적화 알고리즘으로, 각 변수(좌표)별로 하나씩 최적화하여 목표 함수를 점진적으로 최소화 함



Cyclic Boosting Machines (CBM)

❖ CBM – 정규화(Regularization)

- CBM은 데이터 노이즈를 완화하고 안정성을 강화하기 위해 **정규화**와 스무딩을 적용함
  - ✓ 어떤 bin(구간)의 값이 지나치게 크거나 작으면 전체 모델이 왜곡될 수 있음. 이를 방지하기 위해 정규화 적용

$$g_j^k = \frac{\alpha_j^k}{\beta_j^k}$$

with $\alpha_j^k = \alpha_{\text{prior}} + \sum y_i$

and $\beta_j^k = \beta_{\text{prior}} + \sum \hat{y}_i$

✓ 예시

Index(=i)	X_1(나이)	X_2(소득수준)	Y(월 지출액 \$)
1	25	많음	500
2	40	중간	300
3	30	적음	400
4	50	많음	600

➢ X_2에 대해

구간 1

$$g_{2,t}^1 = \frac{500+600}{450+450} \approx 1.22$$

$$\hat{y}_2 = 450 \cdot 1 \cdot 1 = 450$$

$$\hat{y}_3 = 450 \cdot 1 \cdot 1 = 450$$

➢ 학습 과정 3.를 보면 파라미터 즉, 기여도 f_k^j는 **감마 분포** $g = \frac{\alpha}{\beta}$로 **표현 됨**

➢ 최초 감마 파라미터는 median이 1이 되도록 $\alpha_{\text{prior}} = 2$, $\beta_{\text{prior}} = 1.67834$로 설정 $\left(\frac{\alpha - 1/3}{\beta} = \frac{2-1/3}{1.67834} = 1\right)$ → 이상치 덜 민감

✓ 결론 : 초기 파라미터와 실제와 예측 간의 비율을 조정하여 bin(구간) 기여도가 모델에 과도하게 영향을 미치지 않도록 함

Cyclic Boosting Machines (CBM)

❖ CBM – 스무딩(Smoothing)

- CBM은 데이터 노이즈를 완화하고 안정성을 강화하기 위해 정규화와 **스무딩**을 적용함

✓ 데이터의 변동성(노이즈)을 완화하여 예측이 지나치게 급격하게 변화하지 않도록 파라미터 즉, 기여도 f_k^j에 스무딩이 적용 됨

① 로그 변환 : f_k^j 값이 너무 크거나 작을 경우, 이를 부드럽게 처리하기 위해 로그 변환 $(0, \infty) \rightarrow (-\infty, \infty)$

$$f_j'^k = \ln(f_j^k)$$

② 감마 분포와 로그-정규 분포 평균 및 분산 매칭 : 로그 변환된 $f_j'^k$ 의 불확실성 $\sigma^2_{f_j'^k}$ 추정을 위해 분산 매칭

A. 평균 매칭: $\boxed{\frac{\alpha}{\beta}} = \boxed{e^{\mu+\frac{\sigma^2}{2}}} \rightarrow \mu = \ln\left(\frac{\alpha}{\beta}\right) - \frac{\sigma^2}{2}$ B 분산 매칭에 A.에서 간소화 한 μ를 대입

B. 분산 매칭 : $\boxed{\frac{\alpha}{\beta^2}} = \boxed{\left(e^{\sigma^2} - 1\right) \cdot e^{2(\mu+\frac{\sigma^2}{2})}} \rightarrow (e^{\sigma^2} - 1) \cdot \left(\frac{\alpha}{\beta}\right)^2 \cdot e^{\sigma^2} = \frac{\alpha}{\beta^2} \rightarrow \sigma^2_{f_j'^k} = \log(1 + \alpha_j^k) - \log(\alpha_j^k)$

③ 스무딩 적용 : $\sigma^2_{f_j'^k}$을 스무딩 스플라인 강도로 사용

✓ 결론 : 감마 분포와 로그-정규 분포 매칭을 통해 계산된 불확실성을 활용하여 데이터 노이즈 완화(=스무딩)

Cyclic Boosting Machines (CBM)

❖ 실험 성능

- Kaggle 수요 예측 데이터
  - ✓ 타겟 변수(y) : 일일 판매량 (0 이상),
  - ✓ 성능 결과

방법론	SMAPE (↓)	훈련, 테스트 기간
Kaggle 1등 방법론 (세부 정보 X)	12.58%	2013~2017년, 2018년 1~3월
CBM	13.20%	2013~2016년, 2017년 1월~3월

※ 훈련 및 테스트 기간이 다른 이유 : Kaggle에서 2018년 1~3월 실제 y값을 공개하지 않음

  - ✓ 주장
    ① CBM이 기존 보다 훈련 기간이 1년 짧아 적은 데이터를 가지기에, 1년이 추가되면 유사할 것이라 예상
    ② 성능은 기존과 크게 차이 있지 않지만 설명 가능성이 있다는 장점을 가짐

→ 개인적으로 논문에서 비교방법론 표가 없는 게 처음이라 매우 생소하고 다른 데이터셋을 사용 했어야 한다고 생각함

Cyclic Boosting Machines (CBM)

❖ 실험 성능

- 변수별 기여도 시각화 예시
  - ✓ 개별 변수(요인)가 예측 값 $\hat{y}$에 에 미치는 영향을 시각화

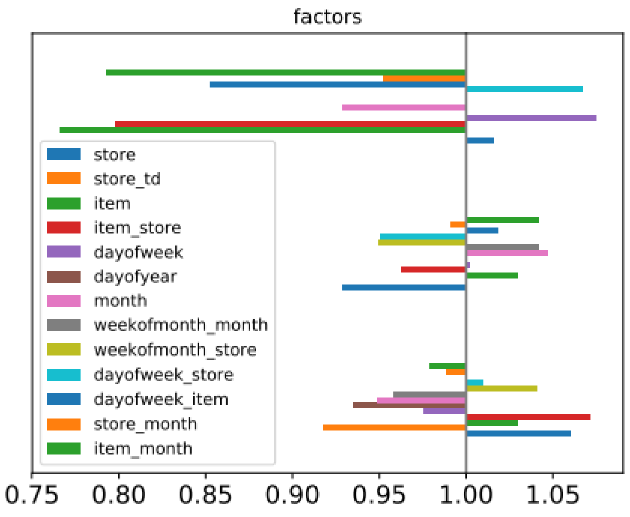


Figure 3: Illustration of the individual factors f_j^k from which the prediction $\hat{y}_i$ is calculated for three individual observations (displayed on top of each other).

❖ 코드 관련

- https://github.com/microsoft/CBM
- https://cyclic-boosting.readthedocs.io/en/latest/index.html

Conclusion

Conclusion

❖ Conclusion

- 기존 머신러닝 및 딥러닝 모델은 성능은 높으나, 설명 가능성이 낮아 의사결정 분야(예: 의료, 금융)에서 활용에 제약이 따르고 있음
- CBM은 준수한 예측 성능과 GAM의 설명 가능성을 결합해 복잡한 데이터에서도 투명성과 성능을 동시에 추구함
- CBM 핵심
  - ✓ 각 변수를 구간화(bin)하여 구간별 요인(기여도)을 계산하고 이를 바탕으로 예측 값 생성
  - ✓ 정규화 및 스무딩을 통해 모델을 강건화

❖ 개인 의견

- Related works에서 EBM이 등장한 만큼 설명가능성이 높은 머신러닝 모델들과 성능 비교하는 정량적 평가가 이뤄지기 바람

Thank you