An Overview of Model-Agnostic Interpretation Methods

Minjae Lee

2020. 09. 18

Contents

01 | Introduction

Seminar Topic

- Model-Agnostic Interpretation methods

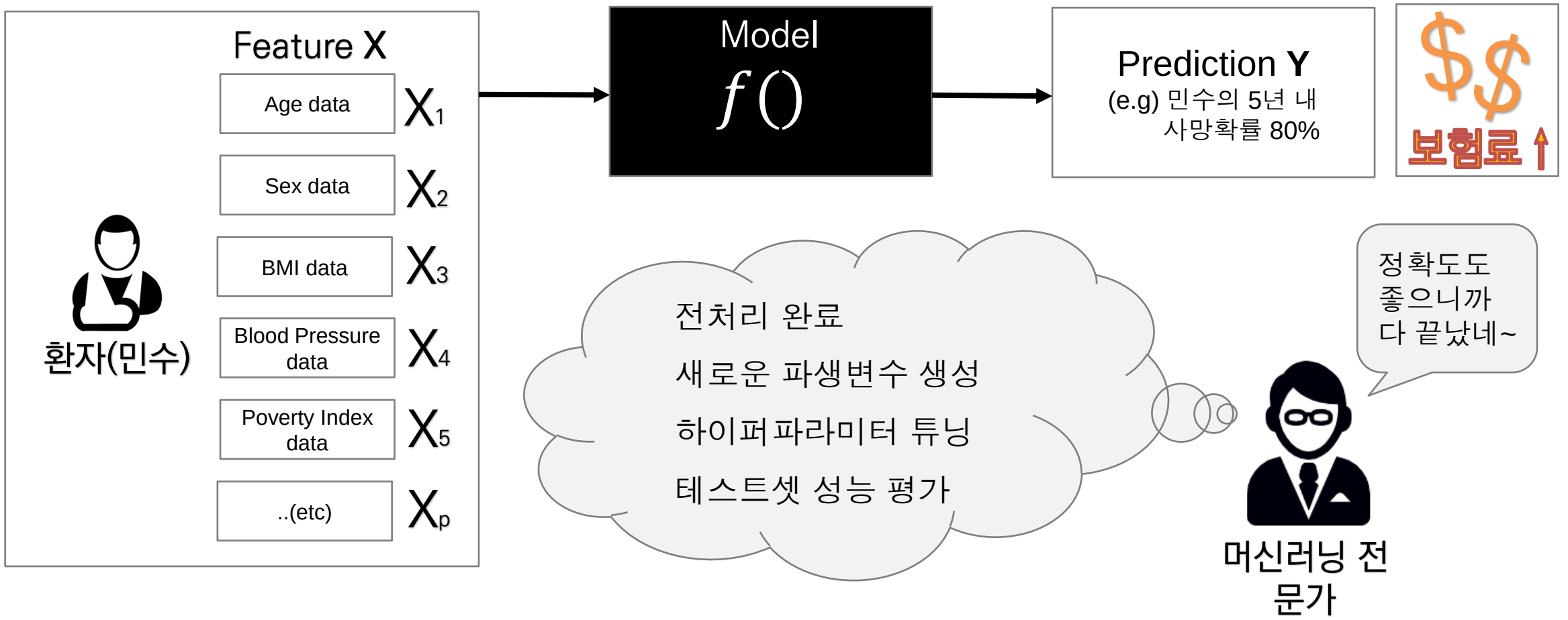
Model-Agnostic Interpretation method

원래 모델 종류에 구애 받지 않는 모델 해석 방법

01 | Introduction

Model Interpretability (모델 해석력)

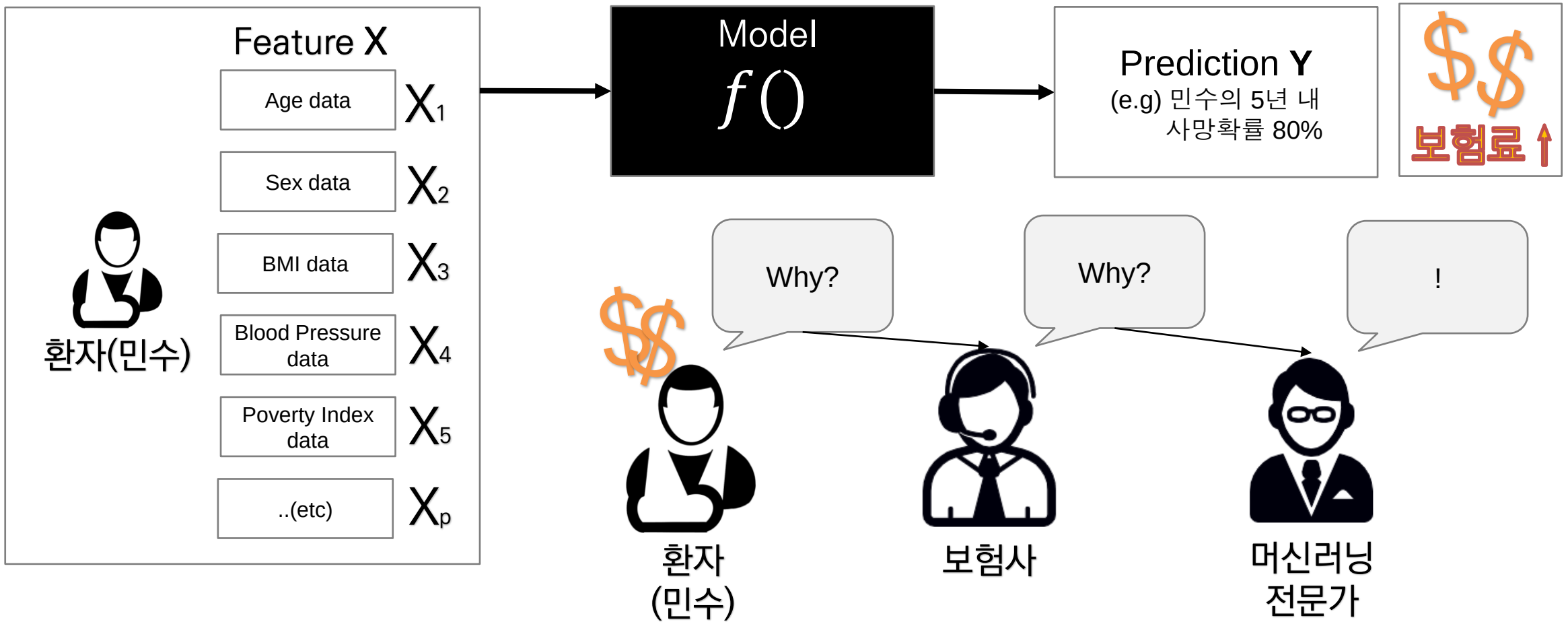
- 예시) 새로운 보험 가입자의 보험료 산출



01 | Introduction

Model Interpretability

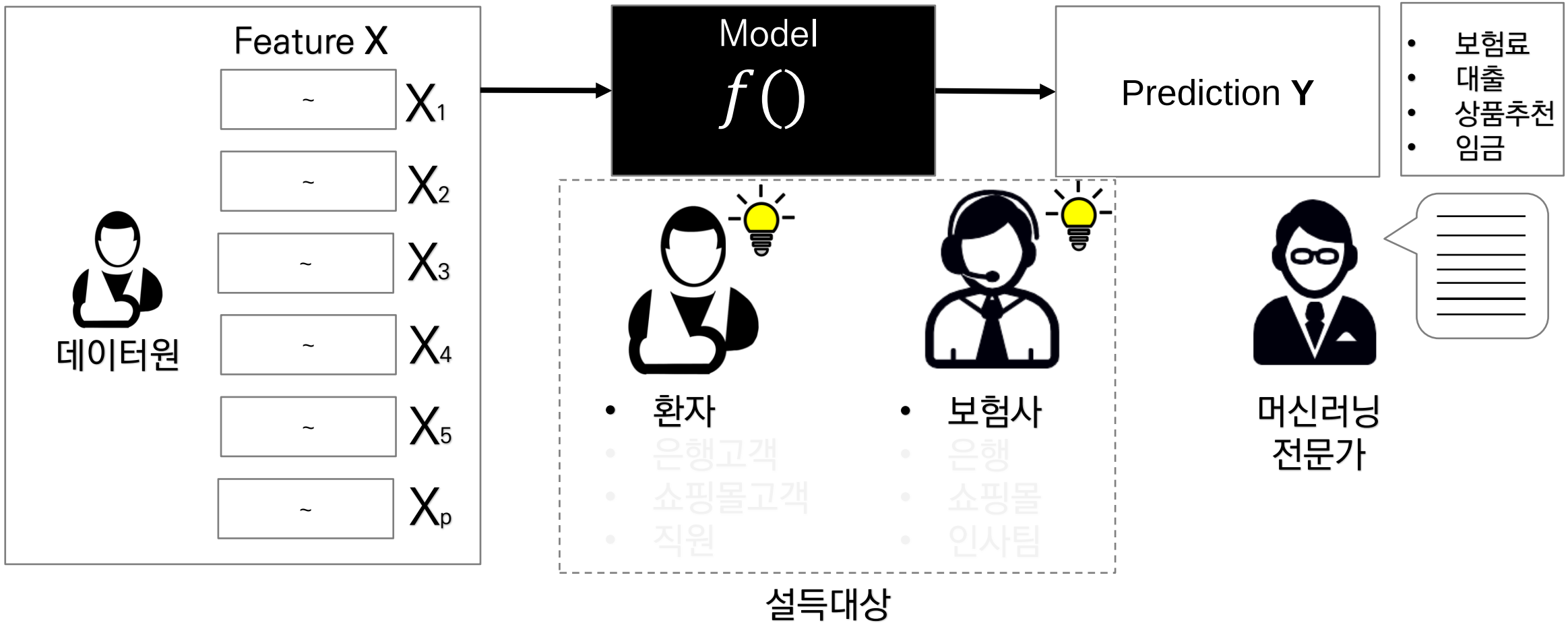
- 정확도가 좋은 모델이라도 고객은 '왜'를 궁금해함 → 이 '왜 ' 에 대한 답변을 제공하는 능력이 Interpretability



01 | Introduction

Model Interpretability

- 고객에게 모델 결과를 자연스럽게 '납득' 시켜주는 수단으로서 의의를 지님 (의의_1)



01 | Introduction

Model Interpretability

- 자연스럽게 모델 결과를 '고객'에게 납득시켜주는 수단으로서 의의를 지님 (의의_1)
- 나아가 모델 파이프라인을 개선시키기 위한 디버깅(Debugging) 수단으로서 의의 지님(의의_2)

  E.g) 자율 주행차의 사고 원인 분석

  E.g) 인종/성별/나이에 따른 모델 편향 분석

〈 Tesla 자율주행차 사물 오인식으로 인한 사고〉, YTN, 2016

〈범죄율 예측 모델이 특정 인종에 대한 모델 편향을 갖는 모습〉,
COMPAS Software Results', Julia Angwin et al. (2016)

01 | Introduction

Seminar Topic

- Model-Agnostic Interpretation methods

Model-Agnostic Interpretation method

원래 모델 — 종류에 구애 받지 않는 — 모델 해석 방법

대표적인 해석가능한 모델

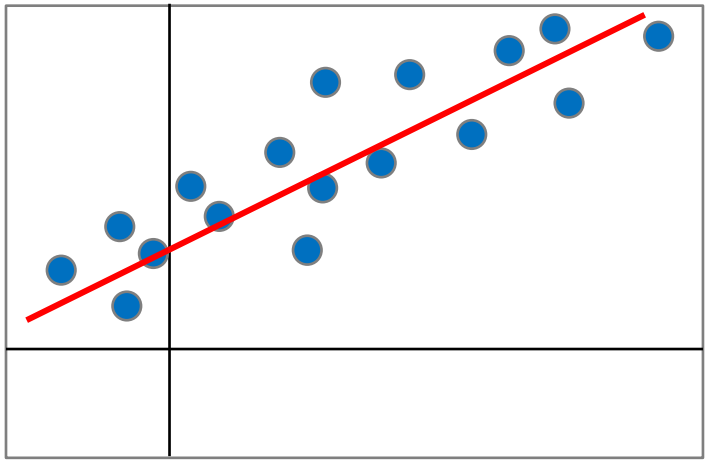
- Linear Regression (상대적인 해석 난이도 : 下)

  상관계수(β)를 '해석력'지표로서 간단히 참고가능

  나아가, 크기 및 (±) 효과를 파악할 수 있음

  t-statistic 등 고전통계기법으로 β의 유의성 검증 가능

  비선형관계 파악을 위해서는 x 에 대한 조작이 필요

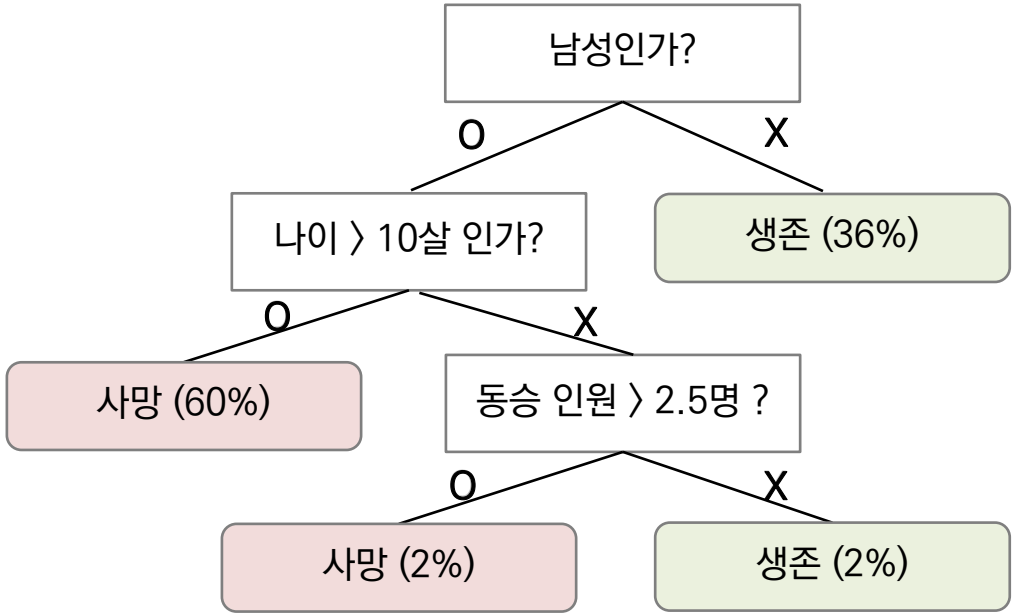


$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

- Decision Tree (상대적인 해석 난이도 : 下)

  트리가 적당히 얕다면, 트리 위부터 아래로 따라가면서

  모델의 의사결정 과정을 자연스럽게 파악 가능



타이타닉 데이터 생존자 요인 트리 분석

01 | Introduction

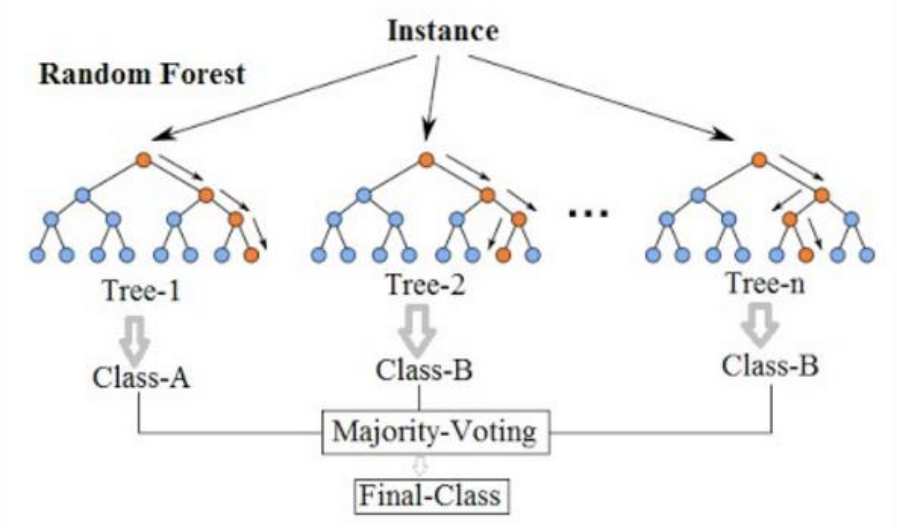
대표적인 해석가능한 모델

- Tree Ensemble Models (상대적인 해석 난이도 : 中)

  Random forest, Gradient Boosting tree

  앞선 두 방법보다 복잡한 데이터에 대해서 좋은 성능을 냄

  Tree 분기 시 해당 변수가 불순도를 평균적으로 얼마나 감소 시키는가 : 변수중요도로 사용가능

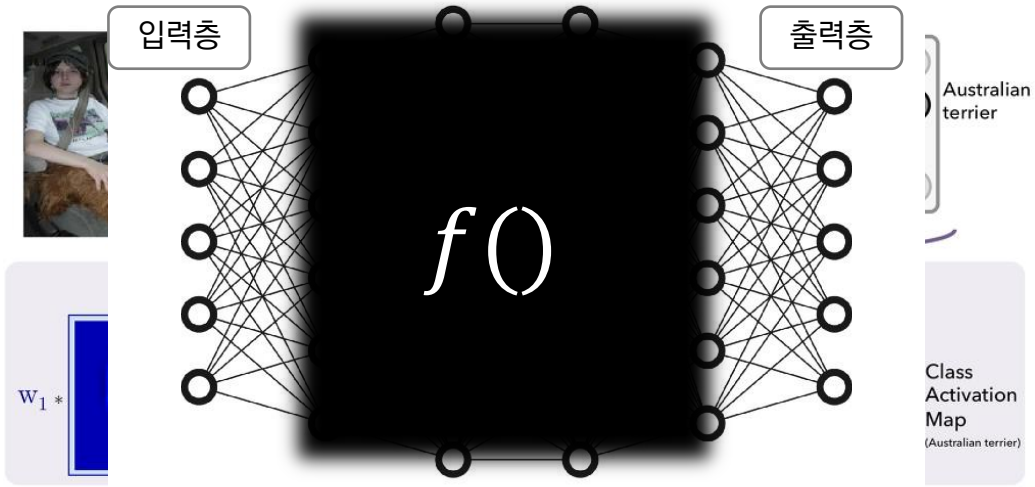


- Deep Neural Network (상대적인 해석 난이도 : 大)

  사실상 Black-box 모델로 해석하기 어려운 모델!

  Attention, Class-Activation-Map 등 추가적인 구조를 추가하여 사용한다면 부분적으로 해석가능하나,

  추가적인 구조 필요로 모델 개발의 부담이 있음

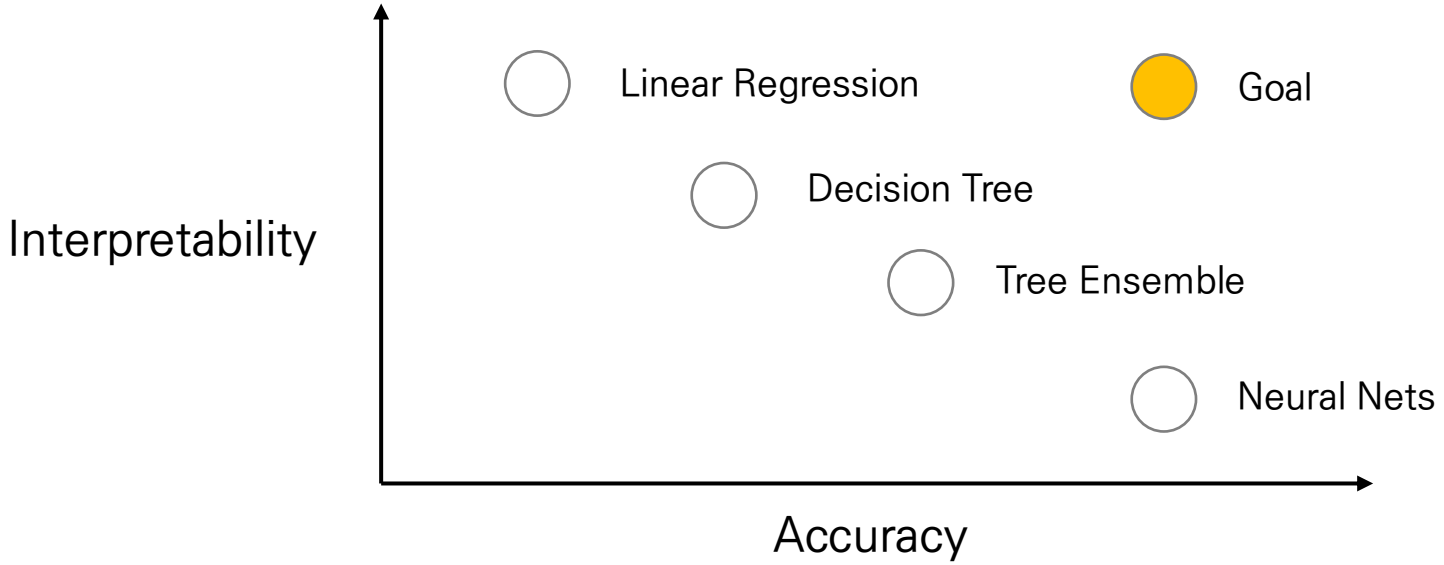


대표적인 해석가능한 모델들의 아쉬운 점

- 1. Accuracy 와 Interpretabilty 동시에 이루기 어려움
- 2 . 모델별 해석방법이 해당 모델에 종속되어 있기에, 다른 모델간 비교를 함부로 하기 어려움

대안 방법론 :Model-Agnostic* Interpretation method

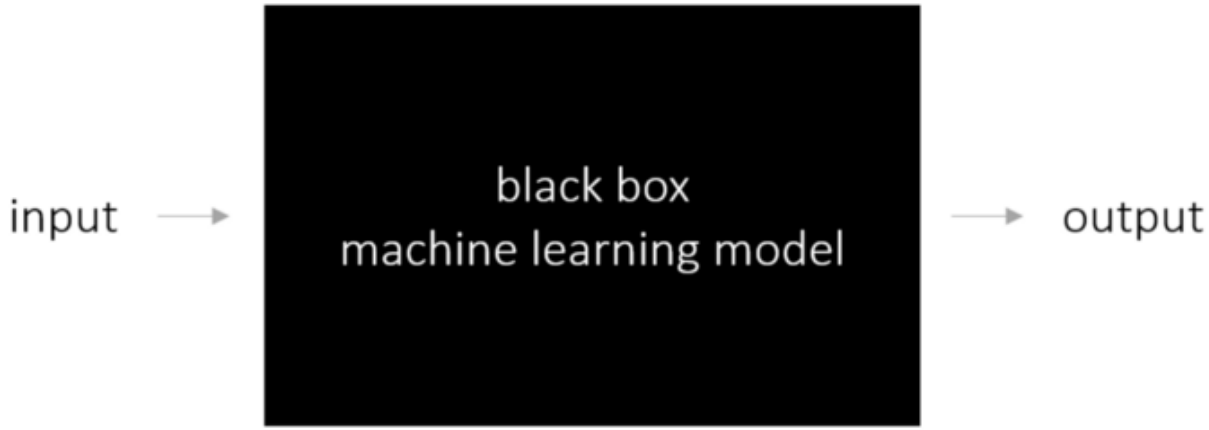
* 모든게 늘 그렇듯이 Trade-Off는 있다!



02 | Model-Agnostic Methods

Model-Agnostic Methods

- 원래 모델을 Black-box 모델로 본다
- 따라서, 모델 내부 변수에 접근하지 않고
- 모델의 input과 output을 조절하면서 그 변동에 대해서 해석!
- cf) Model-Specific model(e.g Linear Regression)은 내부변수(e.g. 상관관계 C)를 보고 처리



02 | Model-Agnostic Methods

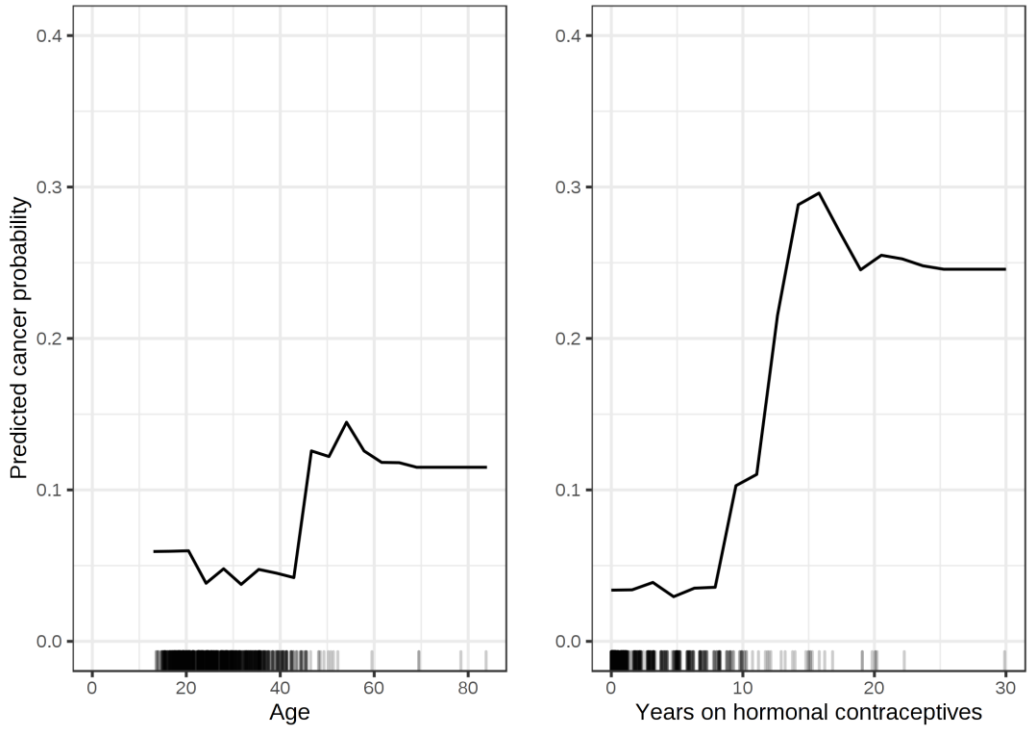
Model-Agnostic Methods

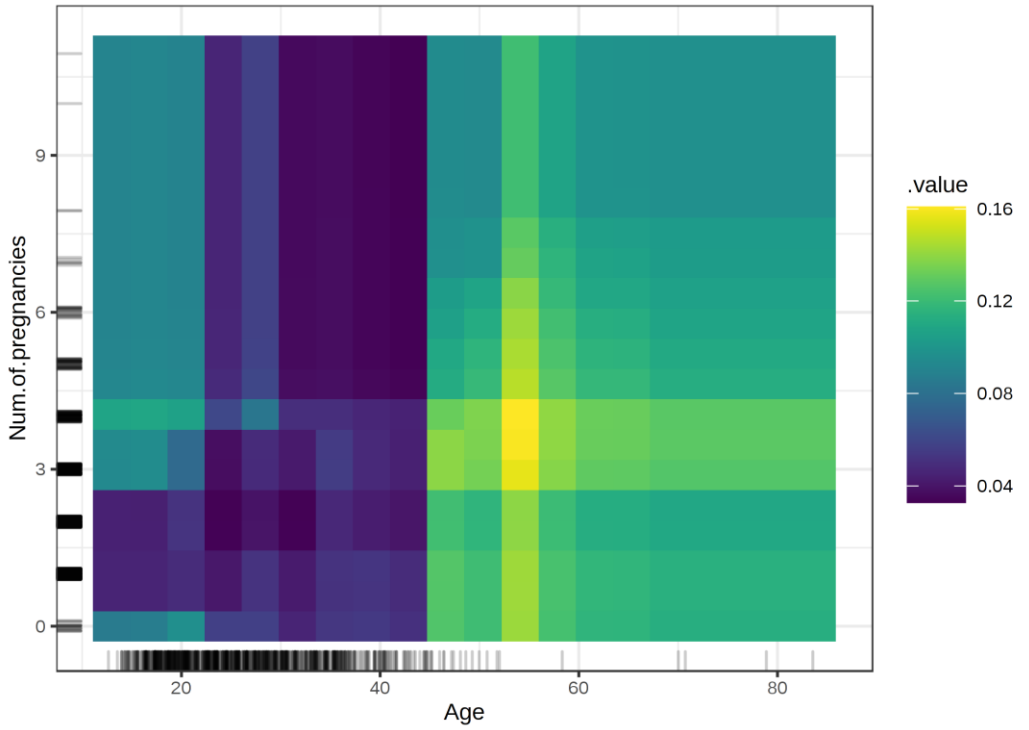
- Partial Dependence Plot (PDP)
- Individual Conditional Expectation (ICE)
- Permutation Feature Importance
- Local Surrogate (LIME)
- SHAP (Shapley Additive explanations)

02 | Model-Agnostic Methods

Partial Dependence Plot (PDP)

- Feature p개($X_1, X_2, \ldots, X_p$)로 사전학습된 모델 $\hat{f}(X)(= \hat{f}(X_s, X_c))$가 있을때
- Partial: 관심대상인 변수 1,2개의 feature 집합과 (X_s) e.g) X_s =(나이) or (피임약 복용 기간)
- Dependence Plot: Target과 간 관계를 그려 살펴보자 e.g) Target(y) = 자궁암 양성(1)/음성(0)





02 | Model-Agnostic Methods

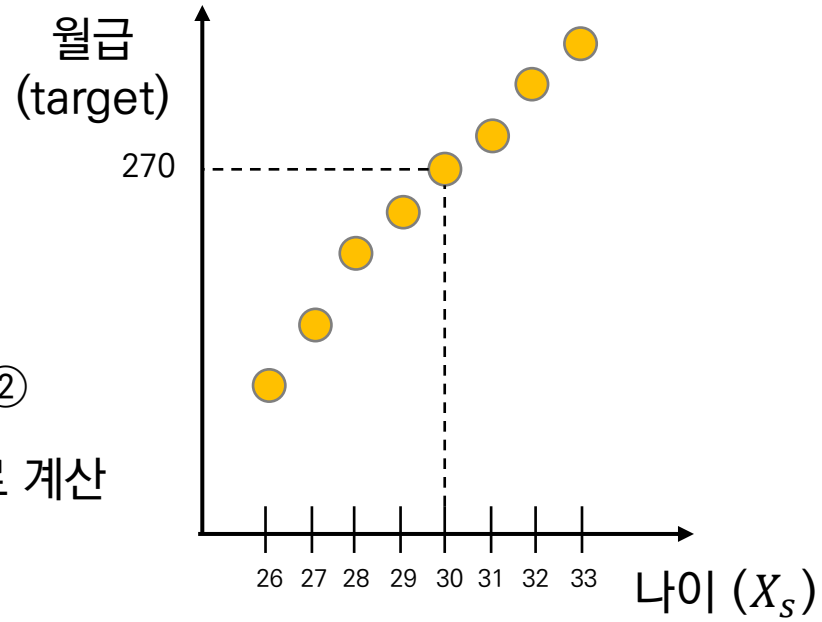
Partial Dependence Plot(PDP)

- $X = X_S \cup X_C$ [S: Selected(관심 대상)인 feature set), C: 그 이외(marginal) feature set]
- $\hat{f}_{X_S}(X_S) = E_{X_C}[\hat{f}(X_S, X_C)] = \int \hat{f}(X_S, X_C)\hat{f}_{X_C}(X_C)dX_C = \sum \hat{f}(X_S, X_C)\hat{f}_{X_C}(X_C) = \frac{1}{n}\sum_i^n \hat{f}(X_S, x_{ic})$ ...①

예시

- 1. $X = (키, 나이)$, $X_S = (나이)$, $x_S = 30$, $X_C = (키)$, n:데이터갯수

X	150	160	170	180
$p(X_C)$	1/10	3/10	5/10	1/10
$f(X_S = \widehat{30}, X_C)$: 예측 월급값	200	250	300	350



- 2. $\hat{f}(X_S = 30) = \frac{1}{20} * 200 + \frac{3}{10} * 250 + \frac{5}{10} * 300 + \frac{1}{10} * 350 = 270$ ...②
- 3. 하지만, $p(X_C)$ 는 주어지지 않기 때문에 ②는 ①처럼 몬테카를로 방법으로 계산
- 4. 다른 나이값 (…,28,29,31,32,…)에 대해서도 마찬가지로 계산

02 | Model-Agnostic Methods

Partial Dependence Plot (PDP)

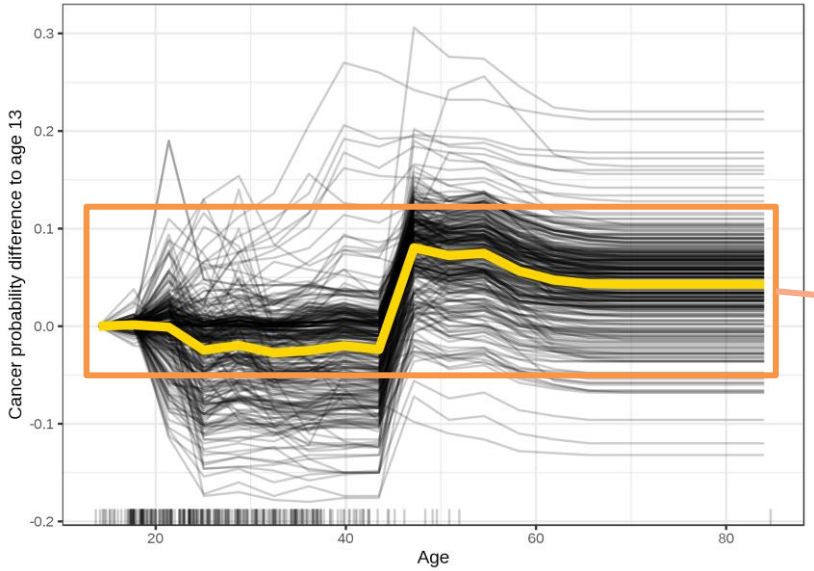
- 장점
  1. 해석이 직관적이고 명확함
  2. 구현하기 쉬움 (관심 feature에 대해 marginalize 하기만 하면 됨)
- 단점
  1. 한 Plot에 그려서 사람이 직관적으로 해석할 수 있는 feature 갯수는 2개
  2. 모든 X_S 의 feature space(e.g 나이)에 대해 X_C(e.g 키, 몸무게,…)를 다 돌아야 하므로 계산량이 많다.
  3. 각 특성(feature)가 독립이라고 가정 (사실 다중공선성(multicollinerity)는 본 모델만이 겪는 일반적인 문제)
  4. $\hat{f}_{X_S}(X_S) = E_{X_C}[\hat{f}(X_S, X_C)]$ 로 주어진 모든 n개의 X_c 에 대해 기댓값(평균값)을 취하기 때문에 임의의 고정된 X_S = (나이) 에 대해서 각 n개의 점들이 갖는 분포가 무시된다.

개선 ver. → Individual Conditional Expectation (ICE)

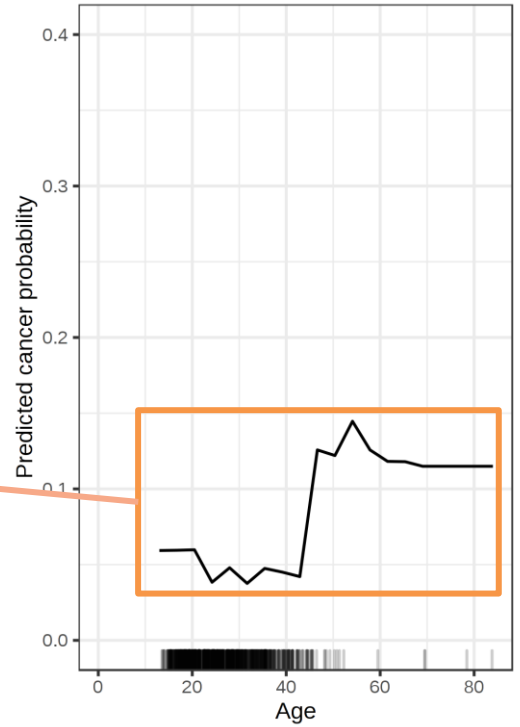
02 | Model-Agnostic Methods

Individual Conditional Expectation (ICE)

- Individual : 각 관측치에 Dependence를 개별적으로 시각화
- 관심변수 X_s 가 변화될때 어떻게 예측값(target)이 변하는지, 모든 train 데이터에 대해 보여준다.
- 따라서, PDP는 ICE의 각 line들의 평균선이다.
- ICE line : $\hat{f}(X_s)^i = \hat{f}(X_s, X_c^i), i = 1,2, \ldots, n$



〈 ICE lines, Average line of ICE Lines(yellow) 〉

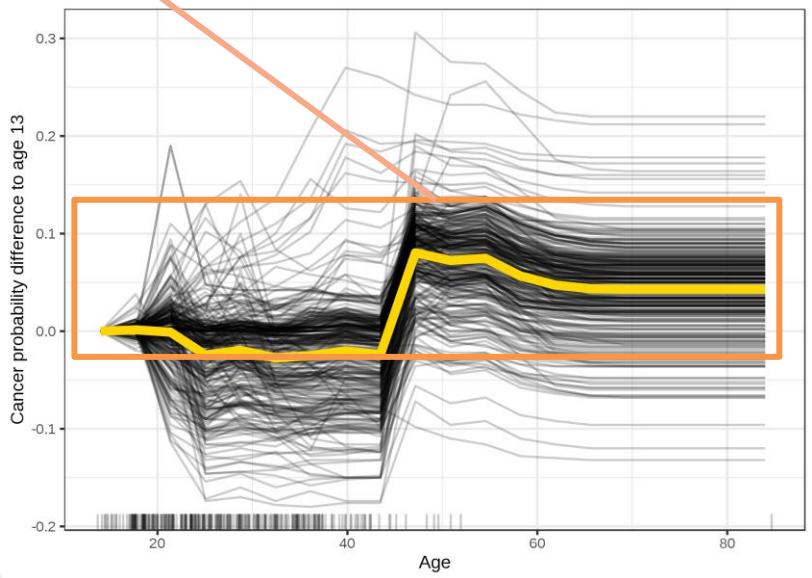


〈 PDP Line, equivalent to yellow line 〉

02 | Model-Agnostic Methods

Individual Conditional Expectation (ICE)

- 장점
  1. 해석이 직관적이고 명확함
  2. PDP처럼 기댓값을 취하지지 않기에, 각 관측치에 대응되는 선을 그릴 수 있음
- 단점
  1. 관측치 수가 많을 경우 너무 조밀하게 plot되어 제대로 파악하기 어려울 수 있음
  2. 그 외 단점은 PDP와 유사



Permutation Feature Importance

- 데이터 행렬 $X(=n\ x\ p)$로 사전학습된 모델 $\hat{f}$ 가 있을 때 (데이터 n개, 특성치 p개)
- Permutation: 확인하고자 하는 특성치(j열)을 순서만을 shuffle한 새 데이터 행렬 X^{perm} 을 만들어
- Base 성능과의 차이를 Feature importance(FI^j)로 사용. ($j = 1,2,..,p$ 에 대해 시행)
- e.g) $FI^j = e^{perm} - e^{base} = L\left(y, \hat{f}(X^{perm})\right) - L(y, \hat{f}(X))$
- e.g) 박지* 선수를 후반전에서 뺐더니(permute), 후반전에 5점 먹힘! → 박시*선수는 굉장히 팀에 중요한 사람이구나!

X_1:나이(yrs)	X_2:피임약 복용기간(yrs)	⋯	X_p:흡연(1/0)
25	2	⋯	1
30	4	⋯	0
⋯	⋯	⋯	⋯
49	10	⋯	0

X

Permute X_2

X_1:나이(yrs)	X_2:피임약 복용기간(yrs)	⋯	X_p:흡연(1/0)
25	10	⋯	1
30	2	⋯	0
⋯	⋯	⋯	⋯
49	4	⋯	0

X^{perm}

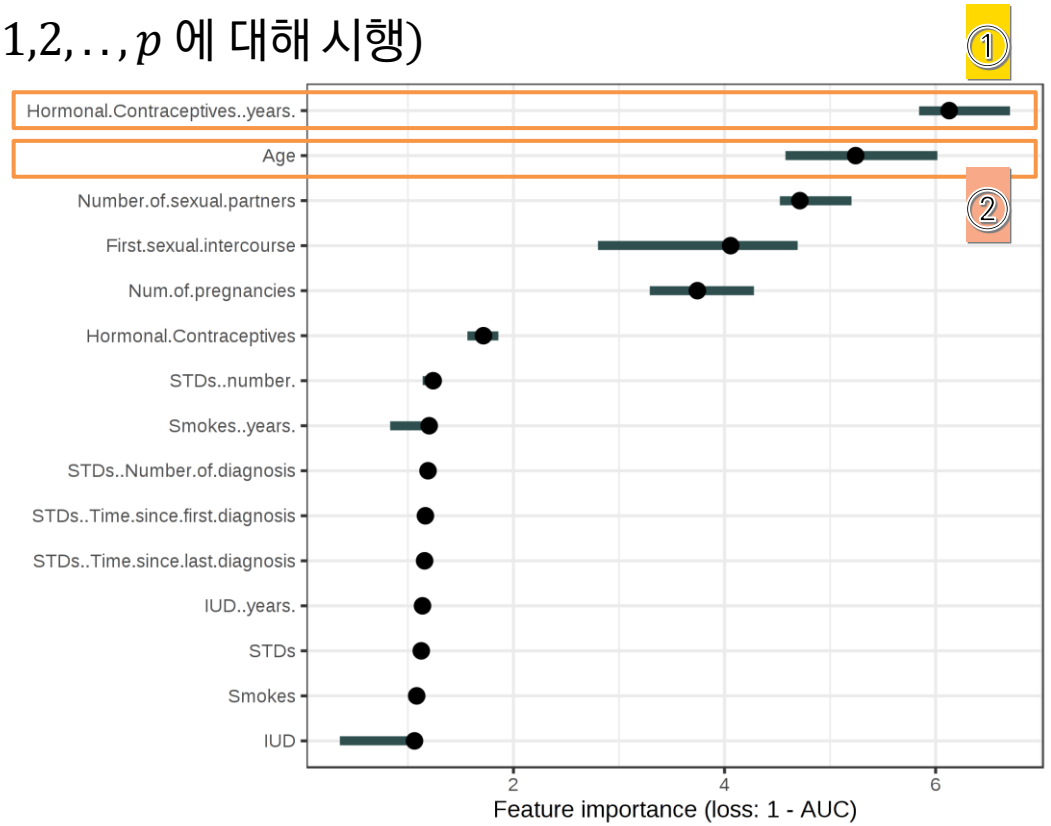
02 | Model-Agnostic Methods

Permutation Feature Importance

- 데이터 행렬 $X(=n\ x\ p)$로 사전학습된 모델 $\hat{f}$ 가 있을 때 (데이터 n개, 특성치 p개)
- Permutation: 확인하고자 하는 특성치(j열)을 순서만을 shuffle한 새 데이터 행렬 X^{perm} 을 만들어
- Base 성능과의 차이를 Feature importance(FI^j)로 사용. ($j = 1,2,..,p$ 에 대해 시행)
- e.g) $FI^j = e^{perm} - e^{base} = L\left(y, \hat{f}(X^{perm})\right) - L\left(y, \hat{f}(X)\right)$

  or $FI^j = e^{perm}/e^{base}$

- 우측 그림은 $FI^j = e^{perm}/e^{base}$ 을 x축으로 씀
- 실제 사용시에는 FI^j 는 여러 번 돌린 값을 쓰게 되어 우측처럼 바(bar) 형태로 plot 된다. (e.g 피임약 복용기간) ①
- 나이를 permute해서 돌릴 경우 baseline로부터 성능저하가 크므로, '나이'가 중요변수라고 유추할 수 있음 ②



02 | Model-Agnostic Methods

Permutation Feature Importance

- 장점
  1. $FI^j = e^{perm}/e^{base}$ 로 오류 비율로 정의할 경우 FI값이 정규화되어 서로 다른 문제끼리 비교 가능
  2. Baseline 모델로 많이 쓰이는 Tree 계열모델에 손쉽게 사용가능
- 단점
  1. 통상 레이블이 있는 supervised-learning 에서만 사용가능 (Loss 구할때 필요)
  2. Permute 시 비현실적인 데이터가 발생할 수 있음
  3. Permute는 무작위로 섞는 것이기 때문에, 아주 많이 시험하지 않는 이상 FI 간 순서가 바뀔 여지가 다분 → $SHAP$ 에서 해결
- (참고)

  Tree 계열의 Impurity기반 feature importance 산출 방식의 대안으로 사용됨

  (https://scikit-learn.org/stable/modules/ensemble.html#feature-importance-evaluation)

  Tree 계열모델에서 Permutation Feature Importance 사용시 주의점

  (https://explained.ai/rf-importance)

02 | Model-Agnostic Methods

LIME (Local Interpretable Model-agnostic Explanations)

- MT Ribeiro et al. (2016). ACM-KDD. Cited by 3718

"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

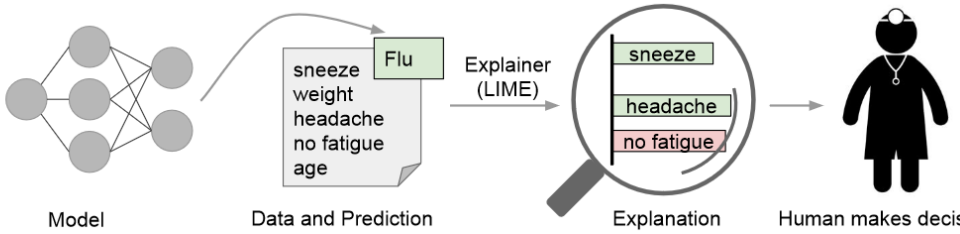


Figure 1: Explaining individual predictions. A model predicts that a patient has the flu, and LIME highlights the symptoms in the patient's history that led to the prediction. Sneeze and headache are portrayed as contributing to the "flu" prediction, while "no fatigue" is evidence against it. With these, a doctor can make an informed decision about whether to trust the model's prediction.

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 technique that explains the predictions of *any* classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. We also propose a method to explain models by presenting representative individual predictions and their explanations in a non-redundant way, framing the task as a submodular optimization problem. We demonstrate the flexibility of these methods by explaining different models for text (e.g. random forests) and image classification (e.g. neural networks). We show the utility of explanations via novel experiments, both simulated and with human subjects, on various scenarios that require trust: deciding if one should trust a prediction, choosing between models, improving an untrustworthy classifier, and identifying why a classifier should not be trusted.

02 | Model-Agnostic Methods

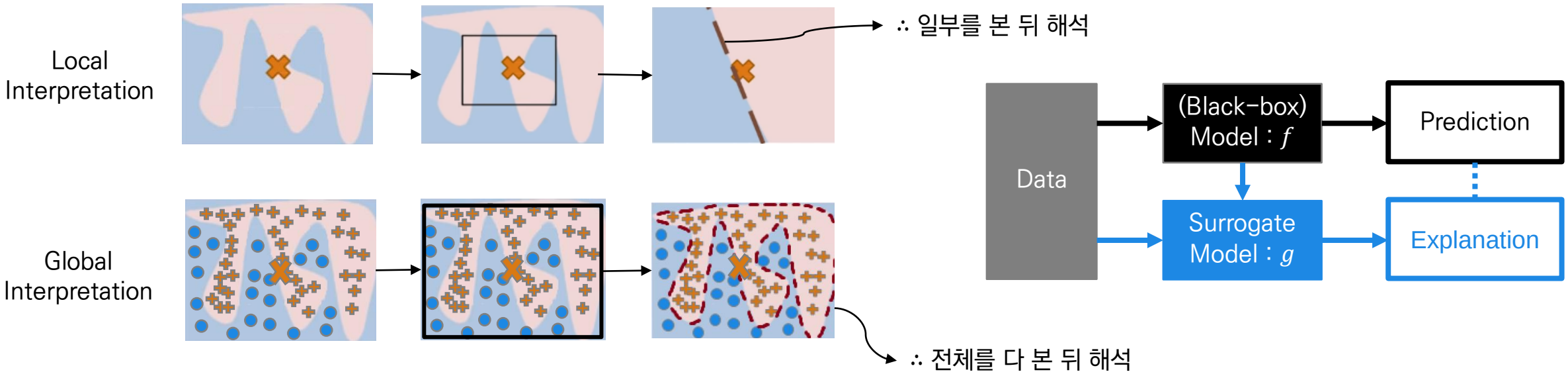
LIME (Local Interpretable Model-agnostic Explanations)

- Local: 단일 관측치(혹은 데이터셋 일부분)에 대한 모델 예측값 해석에 대해 초점을 둠

  e.g) "**왜** 어떤 한 관측치가 어떤 특정 클래스로 구분이 되었는가?" 에 대한 답

  Cf) Global: 모델의 각 부분(or 학습된 전체 모델)이 예측값을 어떻게 만드는지에 대해 초점을 둠

- Surrogate Model: 원래 모델 자체로 해석하기 어려울 때 외부에 구조가 간단한 대리(surrogate) 모델 두어 해석

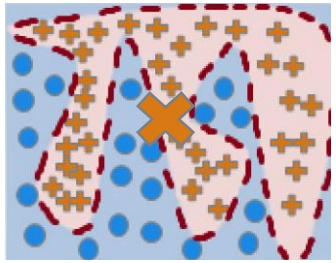


02 | Model-Agnostic Methods

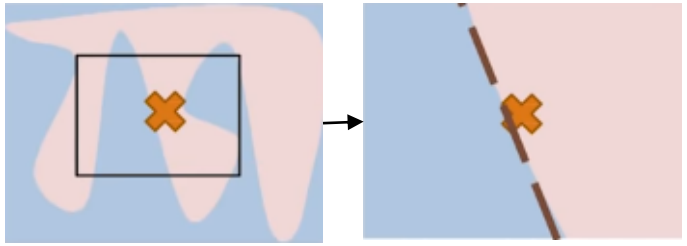
LIME (Local Interpretable Model-agnostic Explanations)

- 아이디어

1. 복잡한 데이터에 적합된 복잡한 모델의 전역적인 해석(Global Interpretation)은 어렵다.

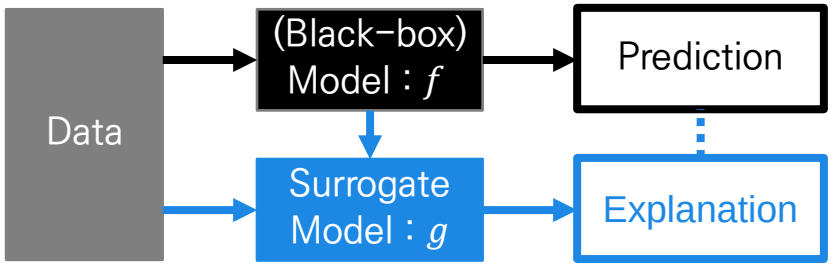


2. 국소적(local)으로는 비교적 해석이 간단한 모델(기능적 상 Surrogate Model)로
   근사시킬 수 있다고 가정하면, 국소적인 해석(Local Interpreation)으로 설명해보자!
   (해석이 간단한 모델의 예 : Linear Regression)



- 핵심 질문!
  - (Question) Linear Regression 모델처럼 간단한 모델에서 Raw 데이터(복잡한 데이터, e.g:이미지)를 그대로 집어넣게 되면 사람이 해석을 도대체 어떻게 하나요? 픽셀 단위로 보는건 사람에게 해석하기 어려운 일이잖아요!
  - (Answer) 네, 좋은 질문입니다. 결국에 사람이 '해석'할 수 있는 꼴로 바꿔주기 위해서 Surrogate Model에는 pixel 단위가 아닌, Super-Pixel을 input으로 사용합니다. 이를 위해 사전 Segmentation 전처리가 필요하구요.

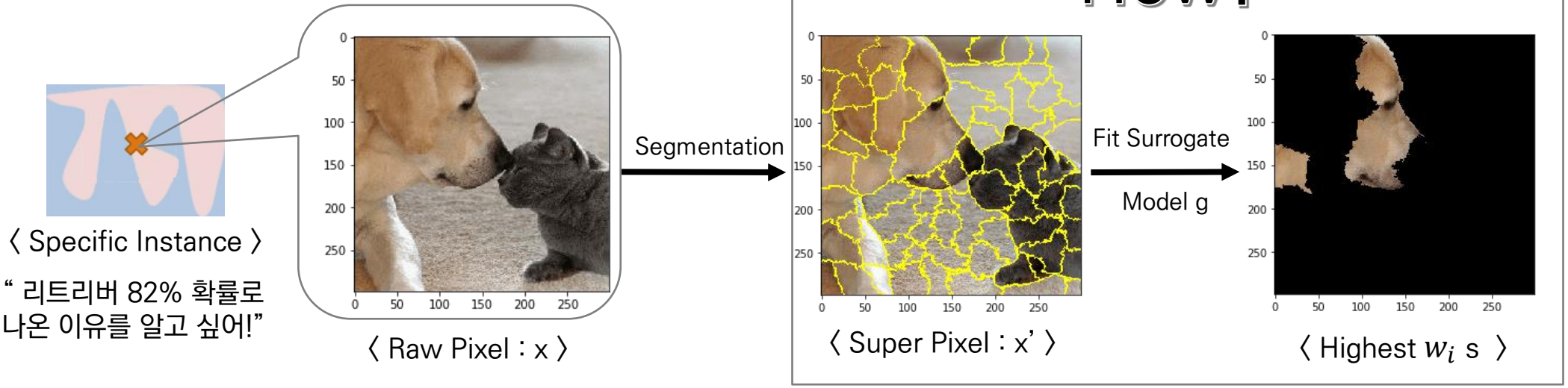
02 | Model-Agnostic Methods



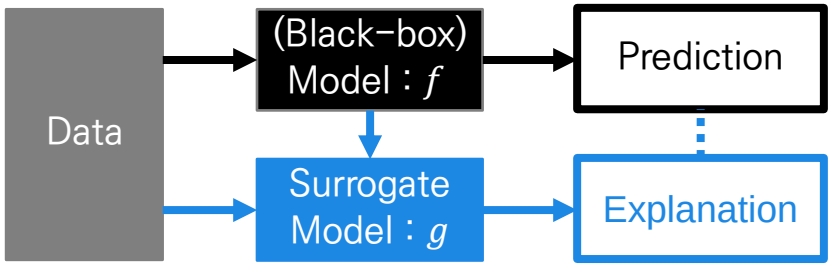
LIME (Local Interpretable Model-agnostic Explanations)

- Interpretable representation (x')
  - 사람이 쉽게 해석가능한 모델은 예를 들어 Linear model 이 있다 → Surrogate Model(=g)로 안성맞춤
  - Surrogate 함수가 linear model이고 data가 이미지인 경우를 예로 들어보자.
  - Interpretable: 사람도 해석가능해야 하지만, explanation model인 linear model에 알맞은 데이터 형태(SuperPixel) 필요
  - $e.g)\ g = g(x') = w_1 * x'_1 + \cdots + w_p * x'_p,\quad w_i: SuperPixel_i$의 가중치, $x_i: SuperPixel$을 넣느냐(1)/빼느냐(0)

How?



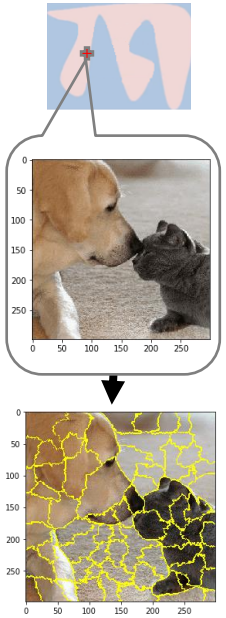
02 | Model-Agnostic Methods



LIME (Local Interpretable Model-agnostic Explanations)

- Steps

1. 해석하고자 하는 관측치(x)를 고르고 x'(interpretable representation: SuperPixel)을 구한다.
2. x' 값을 uniform하게 perturb(=셔플)하여 z' 점 n개를 만든다.
3. $h: z' \rightarrow z$ 로 다시 원래 이미지 차원으로 mapping하고 사전에 학습된 Black-box 모델 f에 z를 입력하여, Surrogate Model(:g) 입력으로 쓸 입력 $\{z', f(z)\}_i \ (i = 1,2,3 \ldots, n)$ 을 구축한다.

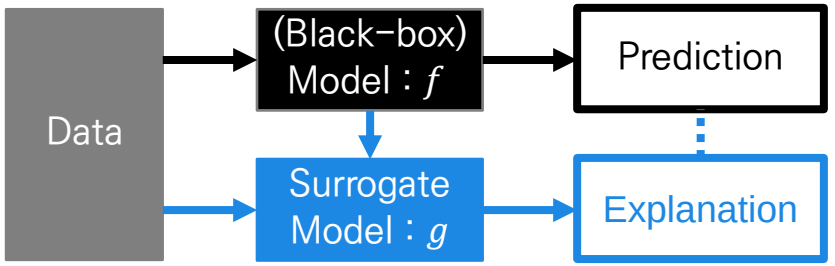


〈 1.〉

〈 2. x'를 perturbation하여 만든 z' 들〉

〈 3. 원래 이미지 hyper-space에 표시된 점 (z,f(z)) 들〉

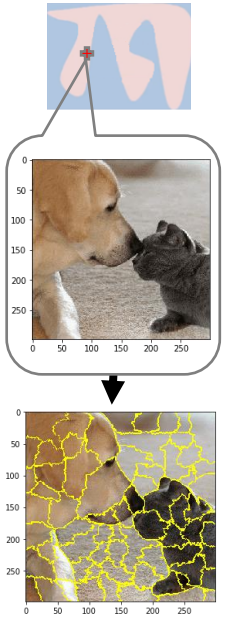
02 | Model-Agnostic Methods



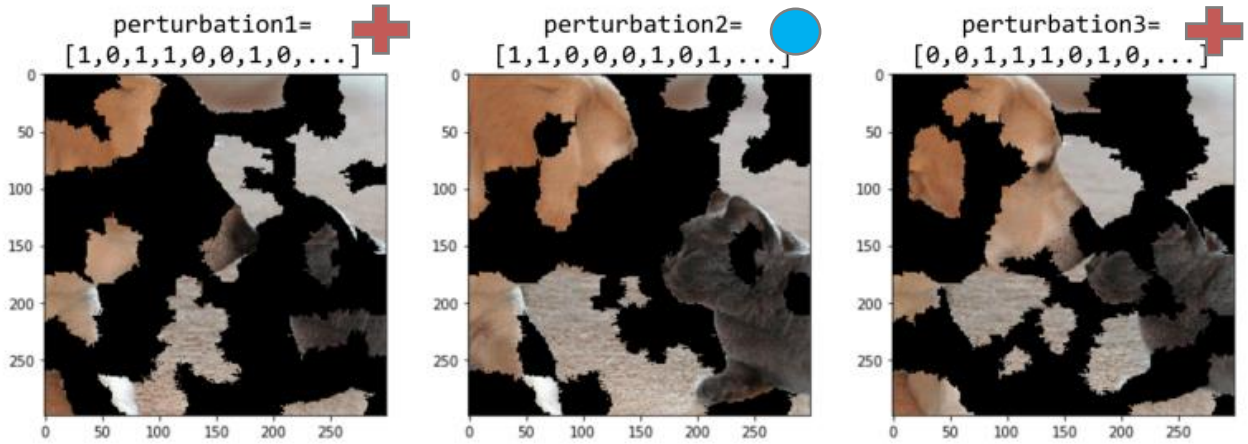
LIME (Local Interpretable Model-agnostic Explanations)

- Steps

1. 해석하고자 하는 관측치(x)를 고르고 x'(interpretable representation: SuperPixel)을 구한다.
2. x' 값을 uniform하게 perturb(=셔플)하여 z' 점 n개를 만든다.
3. $h: z' \rightarrow z$ 로 다시 원래 이미지 차원으로 mapping하고 사전에 학습된 Black-box 모델 f에 z를 입력하여, Surrogate Model(:g) 입력으로 쓸 입력 $\{z', f(z)\}_i \ (i = 1,2,3 \ldots, n)$ 을 구축한다.



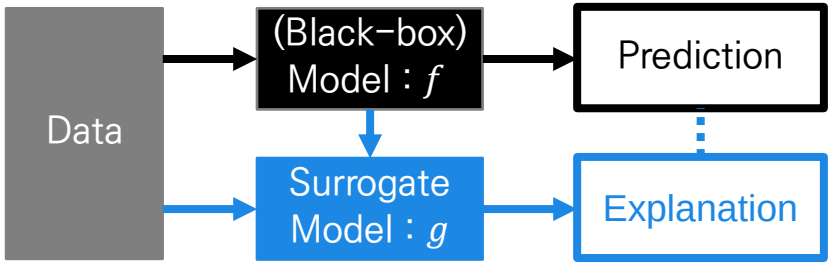
〈 1.〉



〈 2. x'를 perturbation하여 만든 z' 들〉

〈 3. 원래 이미지 hyper-space에 표시된 점 (z,f(z)) 들〉

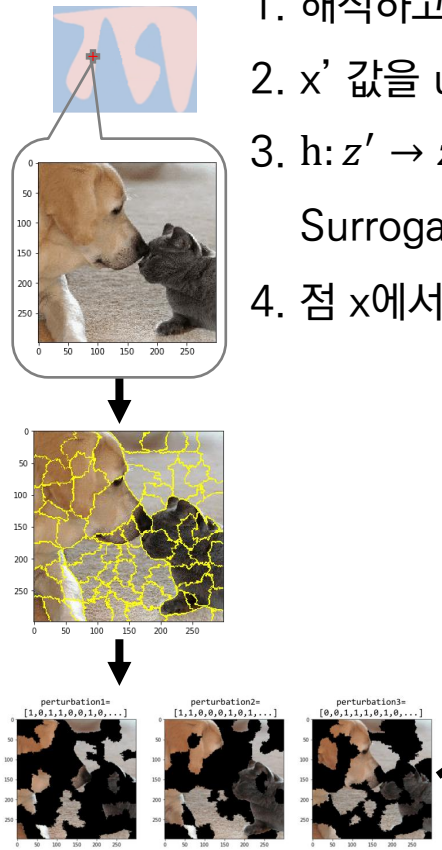
02 | Model-Agnostic Methods

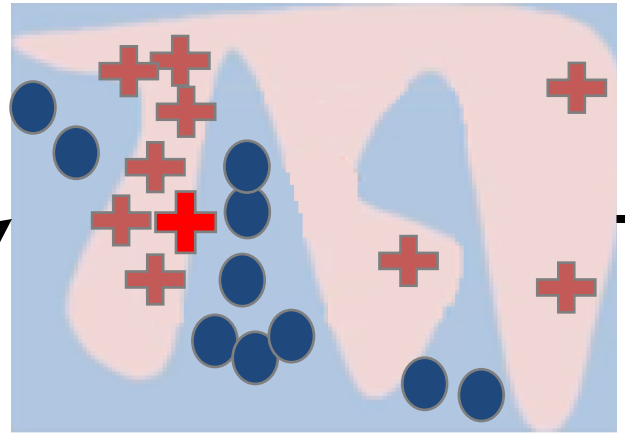


LIME (Local Interpretable Model-agnostic Explanations)

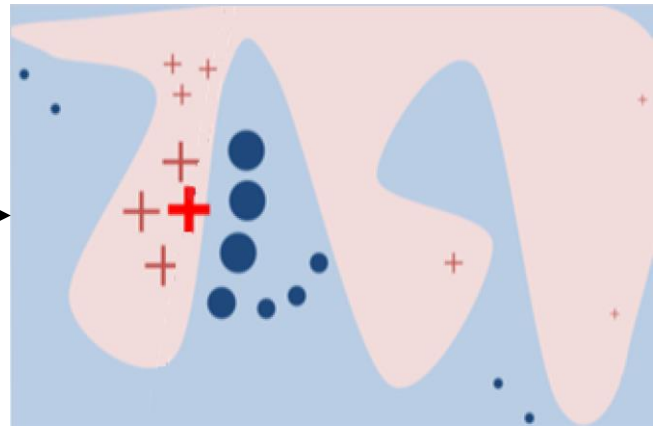
- Steps

1. 해석하고자 하는 관측치(x)를 고르고 x'(interpretable representation: SuperPixel)을 구한다.
2. x' 값을 uniform하게 perturb(=셔플)하여 z' 점 n개를 만든다.
3. h: $z' \rightarrow z$ 로 다시 원래 이미지 차원으로 mapping하고 사전에 학습된 Black-box 모델 f에 z를 입력하여, Surrogate Model(:g) 입력으로 쓸 입력 $\{z', f(z)\}_i \ (i = 1,2,3 \ldots, n)$ 을 구축한다.
4. 점 x에서 가까이 있을수록 가중치를 크게 주고, 멀리 있으면 작게 만들어, x 주변 Local 의 특성을 최대한 살린다.



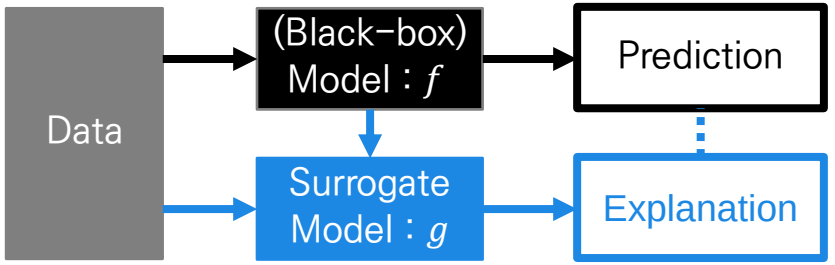


〈 3. (x,f(z)) 와 (z,f(z)) 〉



〈 4. 가중치 함수 적용 후 〉

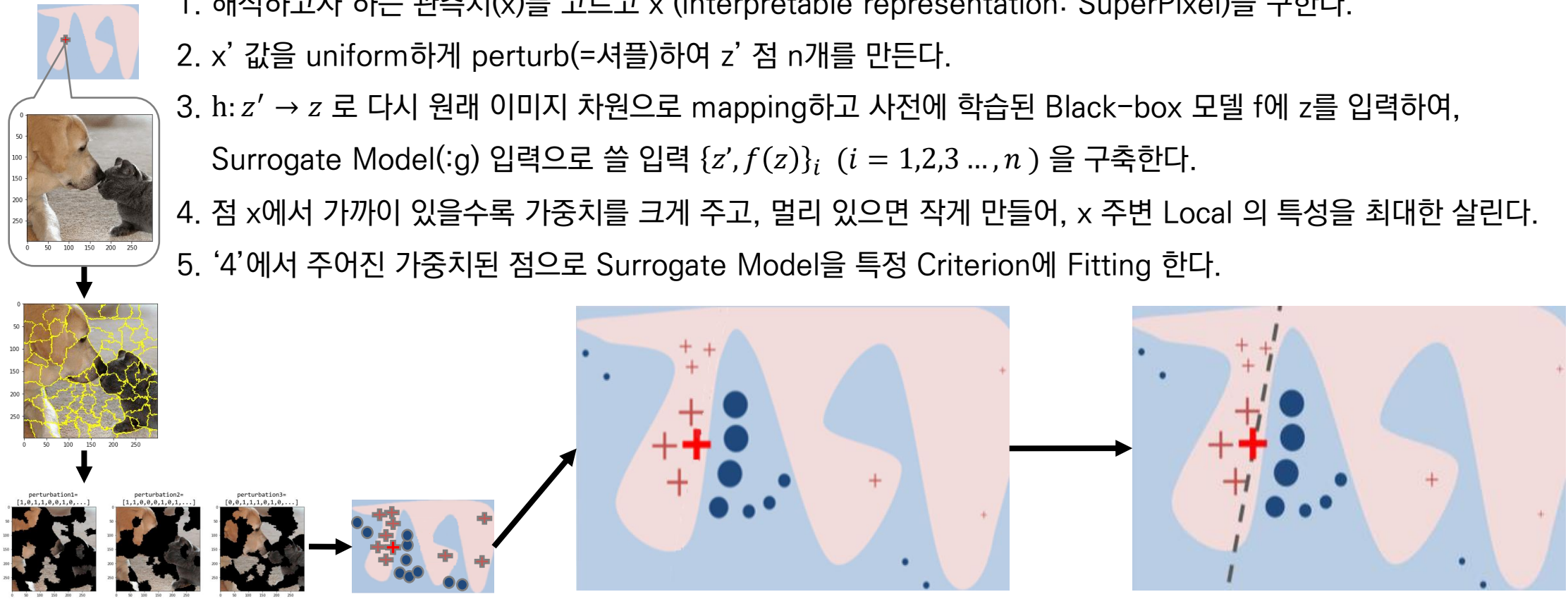
02 | Model-Agnostic Methods



LIME (Local Interpretable Model-agnostic Explanations)

- Steps

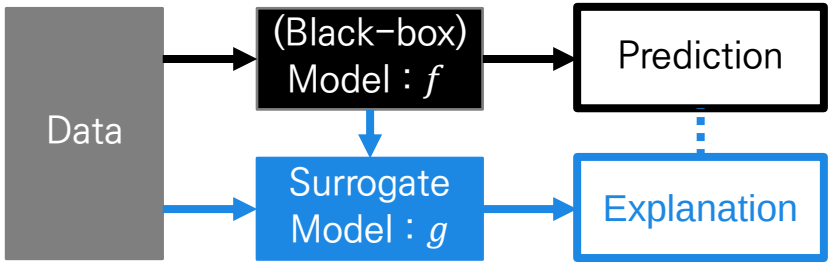
1. 해석하고자 하는 관측치(x)를 고르고 x'(interpretable representation: SuperPixel)을 구한다.
2. x' 값을 uniform하게 perturb(=셔플)하여 z' 점 n개를 만든다.
3. h: $z' \rightarrow z$ 로 다시 원래 이미지 차원으로 mapping하고 사전에 학습된 Black-box 모델 f에 z를 입력하여, Surrogate Model(:g) 입력으로 쓸 입력 $\{z', f(z)\}_i \ (i = 1,2,3 \ldots, n)$ 을 구축한다.
4. 점 x에서 가까이 있을수록 가중치를 크게 주고, 멀리 있으면 작게 만들어, x 주변 Local 의 특성을 최대한 살린다.
5. '4'에서 주어진 가중치된 점으로 Surrogate Model을 특정 Criterion에 Fitting 한다.



〈 4. 가중치 함수 적용 후 〉

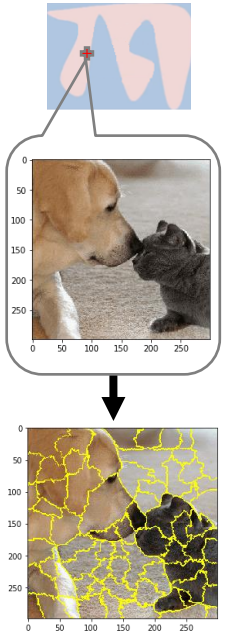
〈 5. Surrogate Model Fitting〉

02 | Model-Agnostic Methods

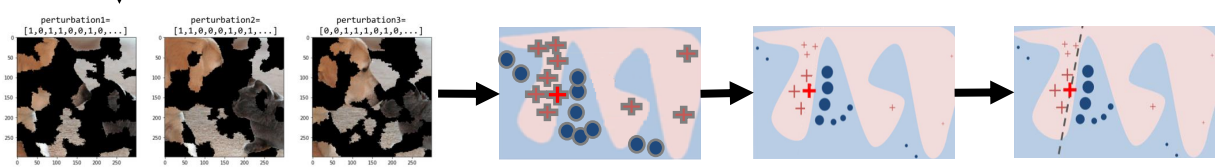


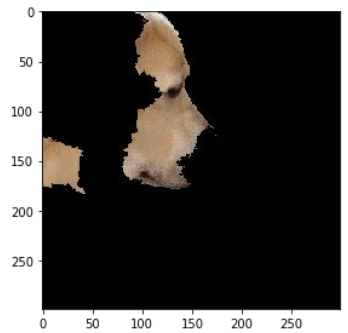
LIME (Local Interpretable Model-agnostic Explanations)

- Steps



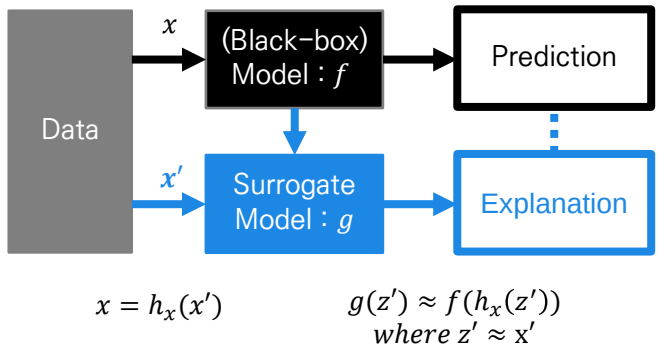
1. 해석하고자 하는 관측치(x)를 고르고 x'(interpretable representation: SuperPixel)을 구한다.
2. x' 값을 uniform하게 perturb(=셔플)하여 z' 점 n개를 만든다.
3. h: $z' \rightarrow z$ 로 다시 원래 이미지 차원으로 mapping하고 사전에 학습된 Black-box 모델 f에 z를 입력하여,
   Surrogate Model(:g) 입력으로 쓸 입력 $\{z', f(z)\}_i \ \ (i = 1,2,3 \ldots, n)$ 을 구축한다.
4. 점 x에서 가까이 있을수록 가중치를 크게 주고, 멀리 있으면 작게 만들어, x 주변 Local 의 특성을 최대한 살린다.
5. '4'에서 주어진 가중치된 점으로 Surrogate Model을 특정 Criterion에 Fitting 한다.
6. $g = g(x') = w_1 * x'_1 + \cdots + w_p * x'_p$, w_i: $SuperPixel_i$의 가중치, x_i: $SuperPixel$을 넣느냐(1)/빼느냐(0) 였음을 상기하고,
   w가 큰 순서대로 뽑아서 해당되는 SuperPixel를 다시 표시하면,
   모델f가 점x에 대해 "래브라도 82%" 라는 결과의 이유를 살펴볼 수 있다.





〈 6. SuperPixels with High weights〉

02 | Model-Agnostic Methods



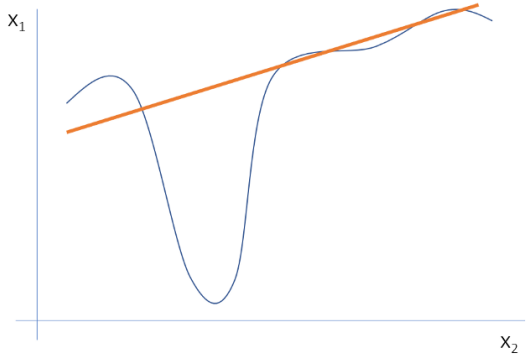
LIME (Local Interpretable Model-agnostic Explanations)

- Steps (Summary : 참고)

1. 해석하고자 하는 관측치(x)를 고른다.
2. x에 대응되는 interpretable representation x' 주위로
   interpretable space에서 Uniform 하게 점(z')들을 n개 Sampling 한다.
3. h:z'->z 로 다시 매핑한 뒤, 이에 대응되는 f(z) 값을 surrogate 함수 g 의 target값으로 쓴다.
4. 1~3에 의해 surrogate 함수g는 {z',f(z)}_i , i=1,2,3…,n 의 데이터셋에 의해 아래 식을 최적화한다.
5. $\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g)$ (1) $\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2$ (2)
6. Surrogate함수 g가 linear model이라고 하면, $g(z') = w_g^T \cdot z'$ 로 나타낼 수 있으며, 이때 x와
   멀리 떨어져있는 점은 낮은 가중치를 주는 커널함수 $\pi = \exp(-\frac{Distance(x,z)^2}{\sigma^2})$로 L를 optimize한다.
7. L은 함수g가 f를 얼마만큼 근사하지 못했느냐의 척도이며, Ω는 g의 복잡도 함수로 linear model일때는 상관계수의 개수다.
8. Optimize, 즉 surrogate함수 g를 적합시키면 그에 따른 w_g 들이 각 interpretation representation의 변수 중요도가 된다.

02 | Model-Agnostic Methods

LIME (Local Interpretable Model-agnostic Explanations)



- 장점/의의
  1. Global한 해석이 아닌 개별 데이터 인스턴스에 대한 local 해석력 제공
  2. Perturbation 의 방식을 다르게 하면 model-agnostic하게 해석할 수 있는 도구를 제공
  3. 추후에 나오는 SHAP보다 계산량이 적음
- 단점
  1. 데이터 분포가 국소(local)적으로도 매우 비선형적이면 local에서 선형성을 가정하는 LIME은 설명력에 한계를 갖게 됨
  2. $\pi = \exp(-\frac{Distance(x,z)^2}{\sigma^2})$ 의 하이퍼파라미터에 따라서 샘플링 성능이 들쭉날쭉하는 불안정성(Inconsistent)
  3. Data 종류(이미지, 텍스트,..)에 따라서, 그리고 어떤 surrogate모델을 고르느냐에 따라서 데이터 perturbation 방식이 다르게 해야하므로, model-agnostic 방법이 갖는 장점인 “유연성”을 다소 퇴색시킴

03 | SHAP (Shapley Additive exPlanations)

SHAP (Shapley Additive exPlanations)

- Scott Lundberg et al. (2017). NeurIPS. Cited by 1571

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

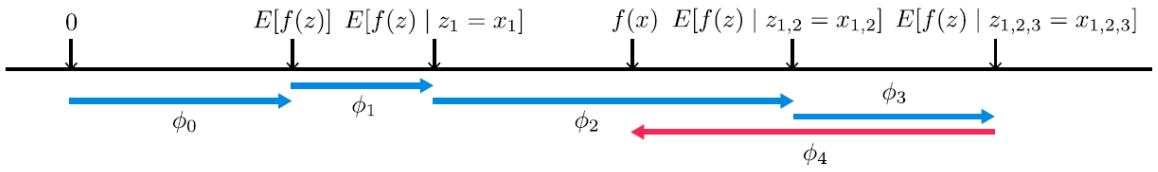


Figure 1: SHAP (SHapley Additive exPlanation) values attribute to each feature the change in the expected model prediction when conditioning on that feature. They explain how to get from the base value $E[f(z)]$ that would be predicted if we did not know any features to the current output $f(x)$. This diagram shows a single ordering. When the model is non-linear or the input features are not independent, however, the order in which features are added to the expectation matters, and the SHAP values arise from averaging the ϕ_i values across all possible orderings.

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 *accuracy* and *interpretability*. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and when one method is preferable over another. To address this problem, we present a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations). SHAP assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing there is a unique solution in this class with a set of desirable properties. The new class unifies six existing methods, notable because several recent methods in the class lack the proposed desirable properties. Based on insights from this unification, we present new methods that show improved computational performance and/or better consistency with human intuition than previous approaches.

SHAP (Shapley Additive exPlanations)

- Additive Feature Attribute methods

Shapley Additive exPlanations

03 | SHAP (Shapley Additive exPlanations)

SHAP (Shapley Additive exPlanations)

- Additive Feature Attribute methods

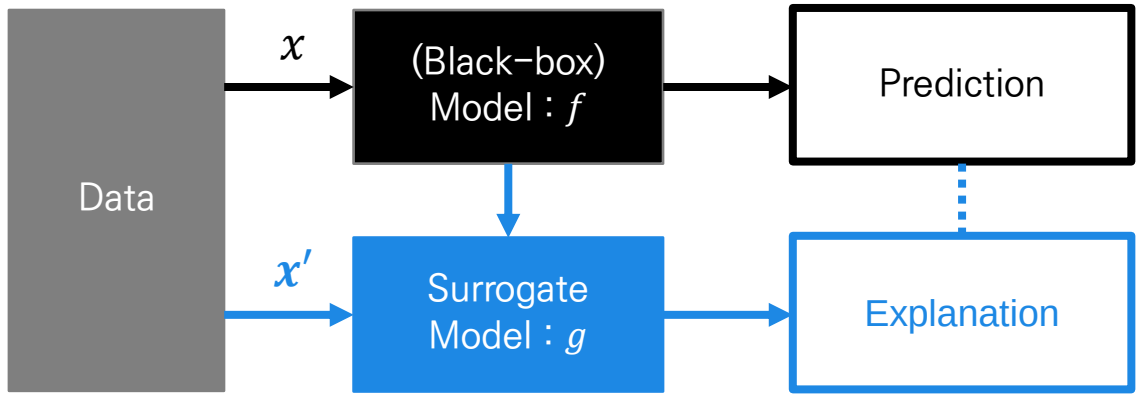
[Review : LIME 개요]

복잡한 모델 $f(x)$를 해석이 간단한 모델인 $g(x)$를 통해 해석을 하곤 한다. (e.g LIME)

x 의 해석용이성을 위해 간단화된 변수 x' 는 $x = h_x(x')$ 인 mapping 함수로 정의되며

$z' \approx x'$ 일 때 $g(z') \approx f(h_x(z'))$ 이도록 Surrogate 함수는 학습된다.

일반적으로 $g(x)$는 Linear model을 많이 쓴다.



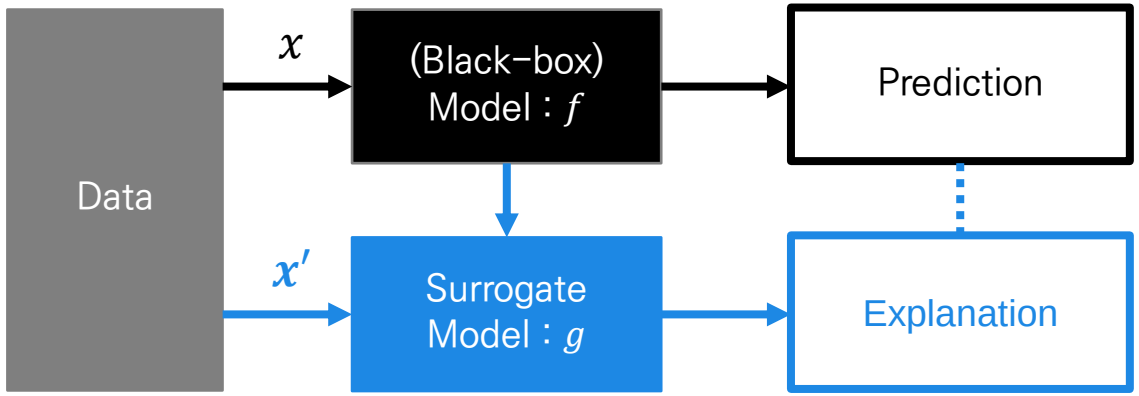
03 | SHAP (Shapley Additive exPlanations)

SHAP (Shapley Additive exPlanations)

- Additive Feature Attribute methods

  정의 : Binary(1/0) 변수의 선형결합으로 이루어진 Explanation 함수(g)

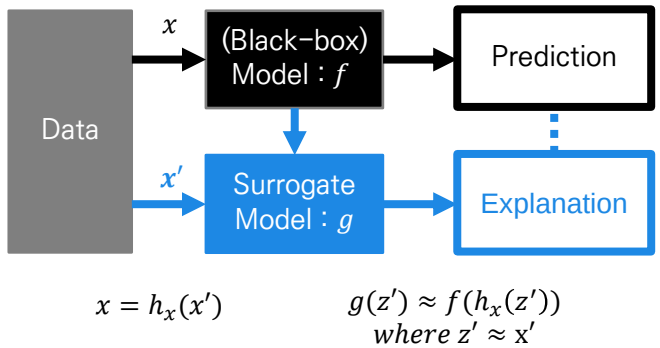
  $g(z') = \varphi_0 + \sum_{i=1}^{M} \varphi_i z_i' \text{, where } z' \in \{0,1\}^M, \varphi_i \in R$



$x = h_x(x')$

$g(z') \approx f(h_x(z'))$
$where\ z' \approx x'$

03 | SHAP (Shapley Additive exPlanations)



SHAP (Shapley Additive exPlanations)

- Additive Feature Attribute methods (협업 게임 이론 관점에서…)

  [Review : LIME 특징] : $g(z') = \varphi_0 + \sum_{i=1}^{M} \varphi_i z_i'$

  - 결과값(g): $team\_game\_score$, z_i': $player_i$, φ_i: $individual\_game\_score$

  1. Surrogate Model $g(z') = \varphi_0 + \sum_{i=1}^{M} \varphi_i z_i'$, where $z' \in \{0,1\}^M, \varphi_i \in R$

     → 각 팀원 점수를 합하면 전체 점수가 된다. (공평!)

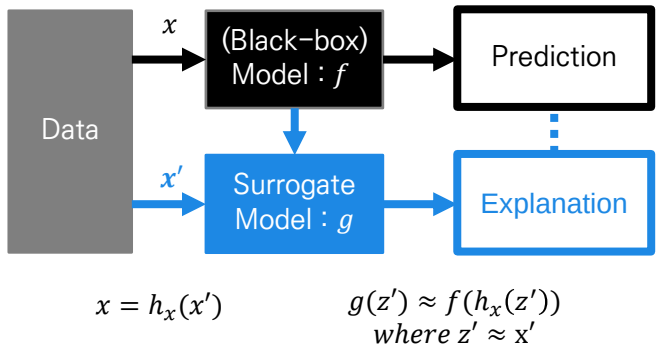
  2. Surrogate Model(: g)를 적합시키 위해 주위 perturbation은 uniform 분포에서 랜덤으로 발생하여 같은 점 x 를 뽑았더라도 랜덤성에 의해 최종 구해지는 g의 내부변수 φ 가 달라지게 됨

     → 매번 똑같은 방식으로 플레이했는데 개인 게임 점수 매겨지는게 다름. (불공평)

  3. [$x_i' = 0 \Rightarrow \varphi_i = 0$] 이 보장되지 않음. 즉, Null한 feature에 대해서 변수 효과가 0으로 취급되지 않음

     → 팀플레이에 참여하지 않았는데 개인 게임 점수가 0이 아님. (불공평)

03 | SHAP (Shapley Additive exPlanations)



SHAP (Shapley Additive exPlanations)

- Additive Feature Attribute methods (협업 게임 이론 관점에서…)

$$g(z') = \varphi_0 + \sum_{i=1}^{M} \varphi_i z_i'$$

1. 각 팀원 점수(z_i')를 합하면 전체 점수($g(z')$)가 된다. (공평!)
2. 매번 똑같은 방식으로 플레이했는데 개인 게임 점수(φ_i) 매겨지는게 다름. (불공평)
3. 팀플레이에 참여하지 않았는데 개인 게임 점수(φ_i) 0이 아님. (불공평)

바라는 특성

1. Addivity
2. Consistency
3. Missingness

불공평한 점을 종결시키는 **유.일.한** 방법!

Shapley Additive exPlanations

Shapley Value 사용

03 | SHAP (Shapley Additive exPlanations)

SHAP (Shapley Additive exPlanations)

- Shapley Value

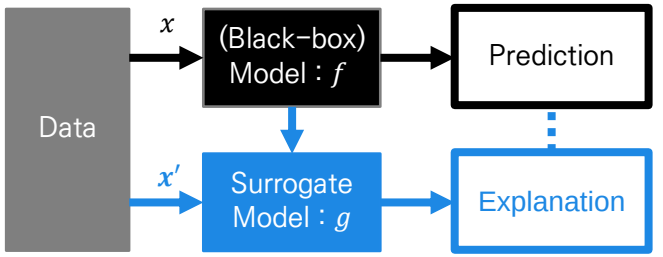
f : $game$
N: $all\ players$
S : $subset\ of\ players$
i : $specific\ player$

① $Importance\ of\ i = f(with\ i) - f(without\ i)$

② $Shapley\ Value\ for\ player_i\ in\ game\ f = Average\ over\ all\ players'\ subsets\ S \subseteq N/\{i\}$ $[Marginal\ contribution\ f(S \cup \{i\}) - f(S)]$

③ $\varphi_i = \sum_{S \subseteq N/\{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (f(S \cup \{i\}) - f(S))$

03 | SHAP (Shapley Additive exPlanations)



$x = h_x(x')$ $\quad g(z') \approx f(h_x(z'))$ $where\ z' \approx x'$

SHAP (Shapley Additive exPlanations)

1. .φ_i 는 협동 게임 이론에서 공평하다고 생각할만한 특성인 Addivitiy, Consistency, Missingness를 모두 갖춤
2. 따라서, 위 특징을 만족하지 못해서 LIME에서 발생하던 문제들 해소
3. 하지만, Shapley Value φ_i 직접적으로 계산하는 것은 모든 순열조합에 대해 체크하기 때문에 효과적인 계산 필요!
4. "모델f 특징에 따라서 계산법을 달리하여 빠른 계산속도 얻음 : KernelSHAP, TreeSHAP, DeepSHAP"

$$g(z') = \varphi_0 + \sum_{i=1}^{M} \varphi_i z_i' \text{, where } z' \in \{0,1\}^M, \varphi_i \in R$$

Shapley Additive exPlanations

$$\varphi_i = \sum_{S \subseteq N/\{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (f(S \cup \{i\}) - f(S))$$

φ_i

03 | SHAP (Shapley Additive exPlanations)

SHAP variations

- KernelSHAP : Truly Model-Agnostic / Relatively Slow / Approximate calculation
- TreeSHAP : For Tree Models / Fast / Exact calculation
- DeepSHAP, GradientSHAP : For Deep Learning Models

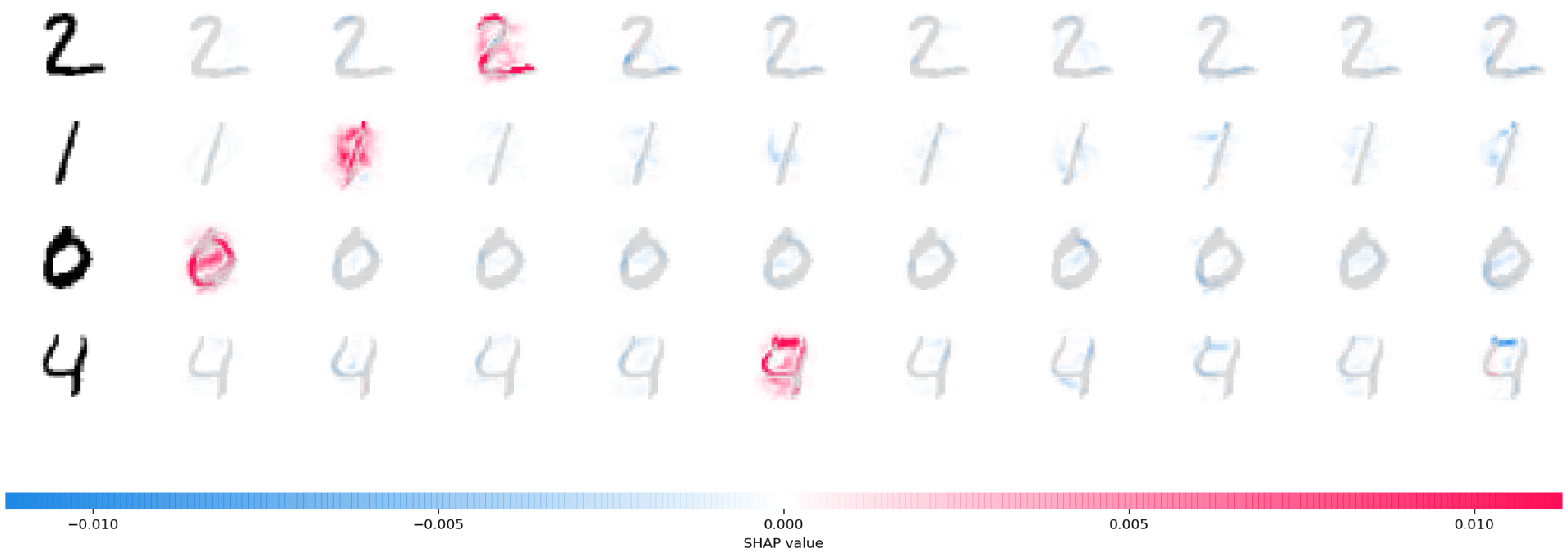
03 | SHAP (Shapley Additive exPlanations)

SHAP plot interpretation

- DeepSHAP : MNIST classification

  모든 Class 마다 SHAP 값을 갖는다는 것을 감안하자

  4 와 9를 구분하는데에 있어서 위 닫히는 변 부위가 핵심적인 파트임을 알 수 있다



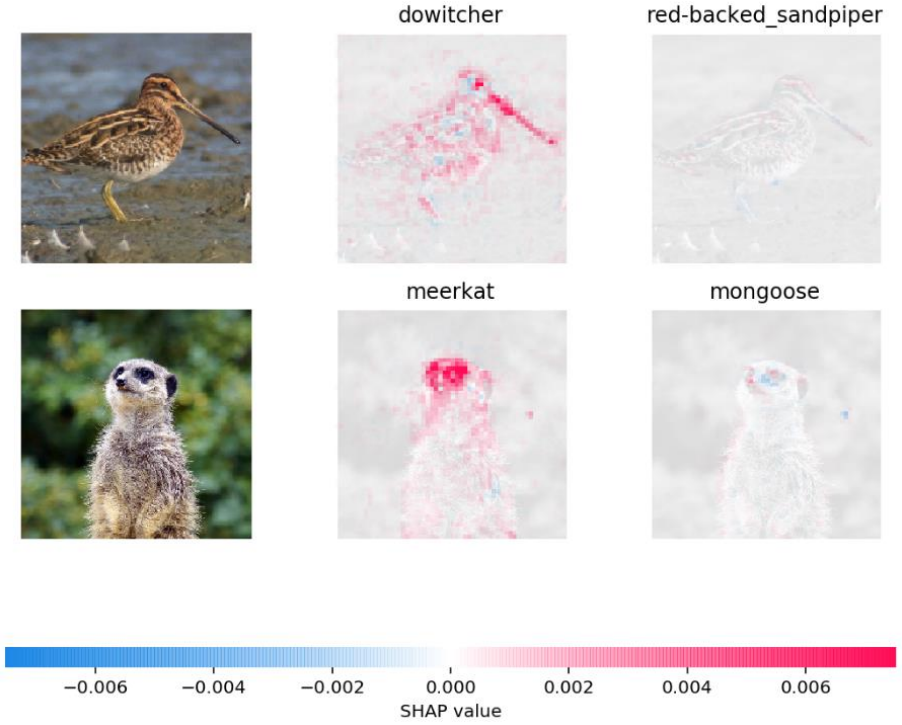
https://github.com/slundberg/shap#deep-learning-example-with-deepexplainer-tensorflowkeras-models

03 | SHAP (Shapley Additive exPlanations)

SHAP plot interpretation

- GradientSHAP : ImageNet classification

  각 Class를 예측해내는데 있어서 어느 부분이 큰 기여를 했는지(SHAP의 크고 작음) 볼 수 있다



https://github.com/slundberg/shap#deep-learning-example-with-gradientexplainer-tensorflowkeraspytorch-models

03 | SHAP (Shapley Additive exPlanations)

SHAP plot interpretation

- TreeSHAP: NHANES Survival Model (summary plot) : 사망률(mortality) 예측

  (우측): SHAP를 가로축으로 모든 관측치에 대해 Bee Pollen Plot을 그려 각 feature 내 크기에 따른 SHAP 분포 확인 가능

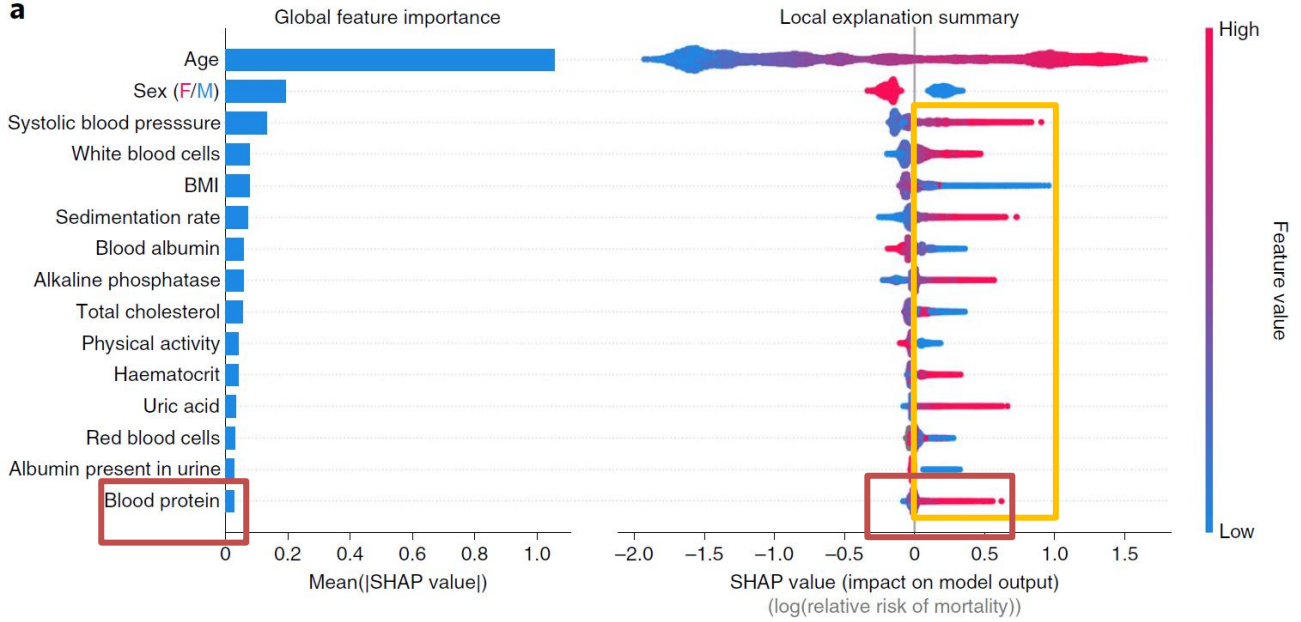
  당연한 결과이지만, 나이가 많을수록 SHAP value가 높고, 젊을수록 낮다 : 나이가 사망률 기여에 큰 기여를 함!

  남성이라는 것 자체가 사망률에 기여률이 여자일 때보다 크다

  안에 있는 부분에 포함되면 사망하기 딱 좋다 [a]

  (좌측) 우측의 관측치에 대응되는 SHAP값을 모두 합쳐서 global feature importance를 각 feature마다 나타낼 수도 있음

  는 전체 관측치에서 그리 큰 importanc를 갖지 않지만, 개별 관측치에서 큰 SHAP까지 분포가 long tail로 이어지므로 눈여겨볼 부분



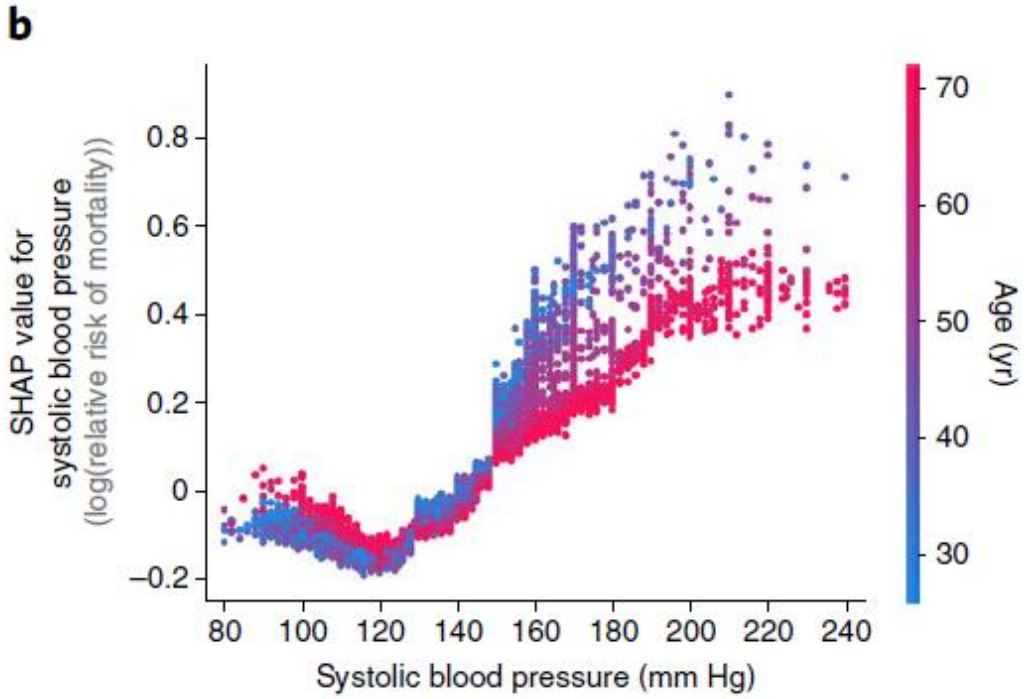
Scott Lundberg et al. (2020). “From local explanations to global understanding with explainable AI for trees”. Nature Machine Intelligence

03 | SHAP (Shapley Additive exPlanations)

SHAP plot interpretation

- TreeSHAP: NHANES Survival Model (dependency plot)

  혈압이 높은 사람 중 젊은 사람일수록 사망률에 큰 SHAP value를 갖는다



Scott Lundberg et al. (2020). “From local explanations to global understanding with explainable AI for trees”. Nature Machine Intelligence

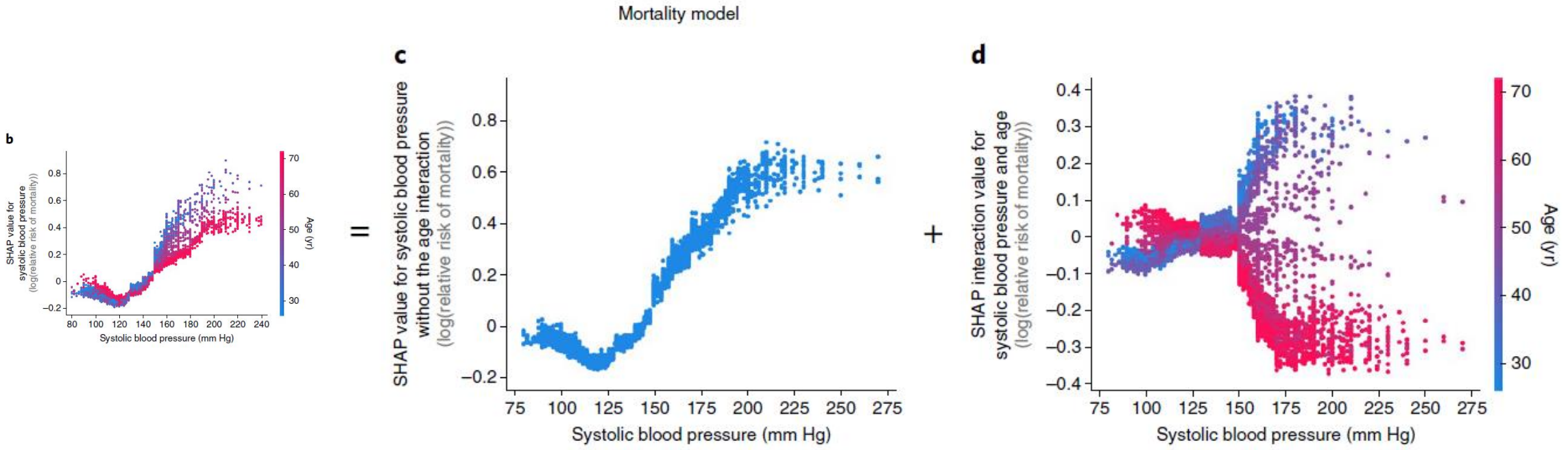
03 | SHAP (Shapley Additive exPlanations)

SHAP plot interpretation

- TreeSHAP: NHANES Survival Model (dependency plot + interaction value)

  SHAP interaction value를 사용하면 interaction(無) plot와 interaction value plot을 decompose 하여 해석할 수 있음

  Dependency plot에서 볼 수 없던 관계를 찾아낼 수도 있음



Scott Lundberg et al. (2020). "From local explanations to global understanding with explainable AI for trees". Nature Machine Intelligence

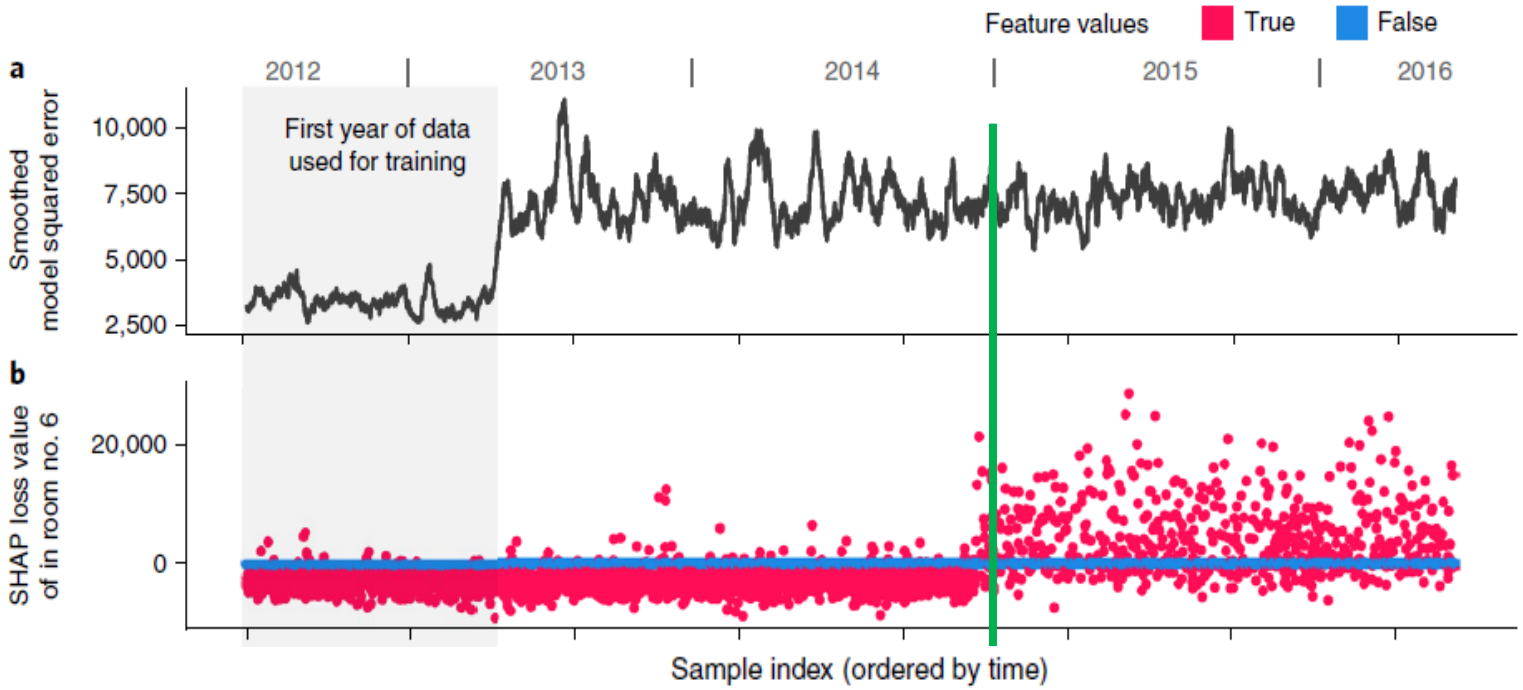
03 | SHAP (Shapley Additive exPlanations)

SHAP plot interpretation

- 수술 시간 예측 모델 (시뮬레이션) (monitor plot)

  인위적으로 두 방 번호를 방 번호(6,13)를 바꾸었을 때 (초록색 바)

  모델 일반 Loss상에서는 이상신호 감지 하지 못하지만, 6번 방의 SHAP loss 값으로는 확연한 SHAP value 차이를 볼 수 있음



Scott Lundberg et al. (2020). “From local explanations to global understanding with explainable AI for trees”. Nature Machine Intelligence

03 | SHAP (Shapley Additive exPlanations)

SHAP

- 장점
  1. Model-Agnostic 방법론 중에서 Explanation model이 가져야할 좋은 특성들이 이론적으로 잘 증명됨.
  2. 각 관측치에 대한 Local Explanation뿐만 아니라, 각 feature 별 SHAP mean으로 Global Explanation도 얻을 수 있다.
  3. 사견) 현재(2020.09) 쉽고 빠르게 쓸 수 있는 SOTA Model-Agnostic Interpretation Method
- 단점
  1. KernelSHAP의 경우 속도가 느리다.
  2. 자칫하면 SHAP value를 원인/결과로 해석할 여지가 있음

Reference

Papers

- MT Ribeiro et al. (2016). “LIME: “Why Should I Trust You? : Explaining the Predictions of Any Classifiers”. ACM-KDD
- Scott Lundberg et al. (2017). “SHAP: A Unified Approach to Interpreting Model Predictions”. NeurIPS.
- Scott Lundberg et al. (2018). “TreeSHAP: Consistent Individualized Feature Attribution for Tree Ensembles”. Arxiv.
- Scott Lundberg et al. (2020). “From local explanations to global understanding with explainable AI for trees”. Nature Machine Intelligence

Codes

- SHAP, https://github.com/slundberg/shap
- LIME, https://github.com/marcotcr/lime
- Scikit-Learn, Tree Feature Importance evaluation, https://scikit-learn.org/stable/modules/ensemble.html#feature-importance-evaluation
- Scikit-Learn, PDP & ICE, https://scikit-learn.org/dev/auto_examples/inspection/plot_partial_dependence.html

Tutorials

- Human Center AI 2019 seminar, https://human-centered.ai/seminar-explainable-ai-2019/

Books

- Christoph Molnar.『Interpretable Machine Learning』. lulu(2020)