모델링, 너 뭐 돼?

구산팔해(九山八海)가 모여 벨 수 없는 것은 없으니!

- 롤로노아 조로 -

ESC 26기 고정민

목차 Contents

목차라고는 하지만 딱히 할 말은 없는데 또 안 적으면 이상하니까 아무 말이나 하다가 말아야지
이걸 읽고 있는 당신! 그대는 방금 전 10초를 날리셨습니다! 축하합니다! 쿠쿠루삥빵뽕 봉구스 밥버거

Boosting

빠르고 유연하다!
거기다 연계 활용성까지!

Deep Learning

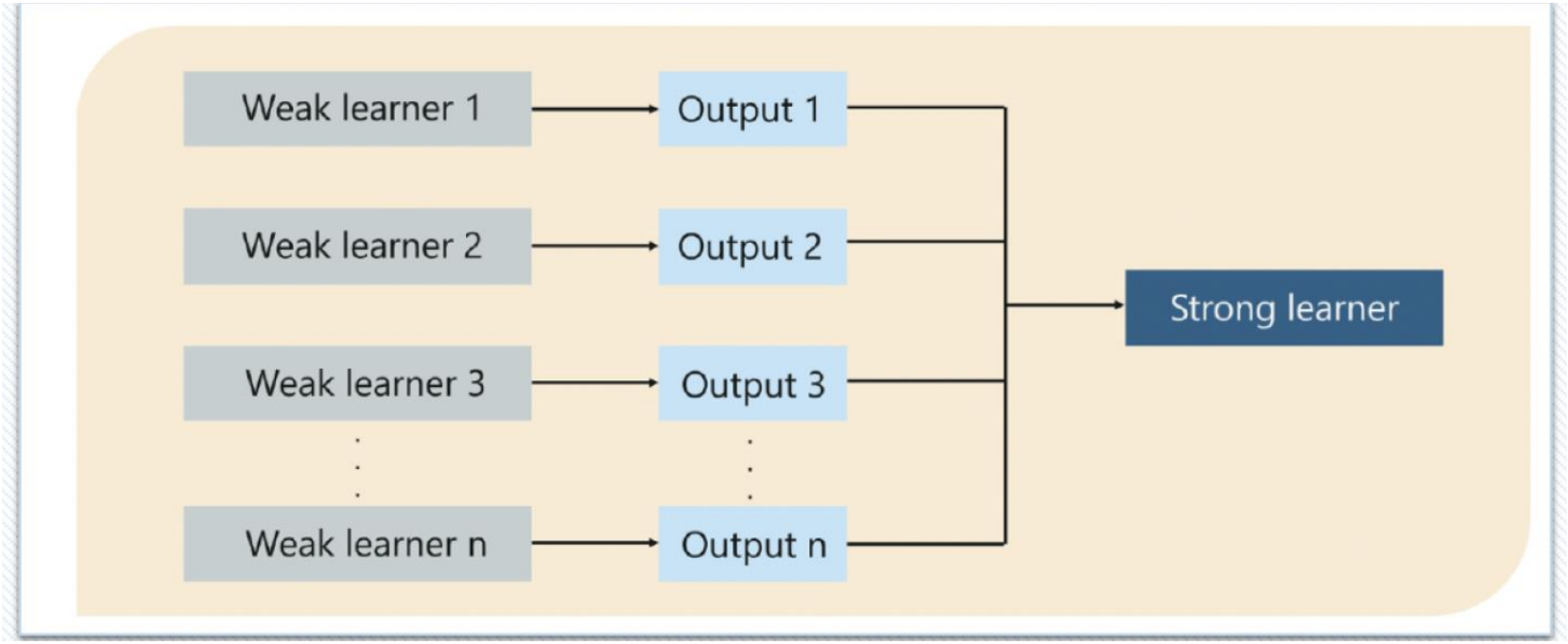
나 : 아 노력 없이 성공하고 싶따..
? ? ? : 이 모델은 공짜로 해줍니다

Explainability

그래서 네 말이 맞았다는 걸
어떻게 증명할건데?

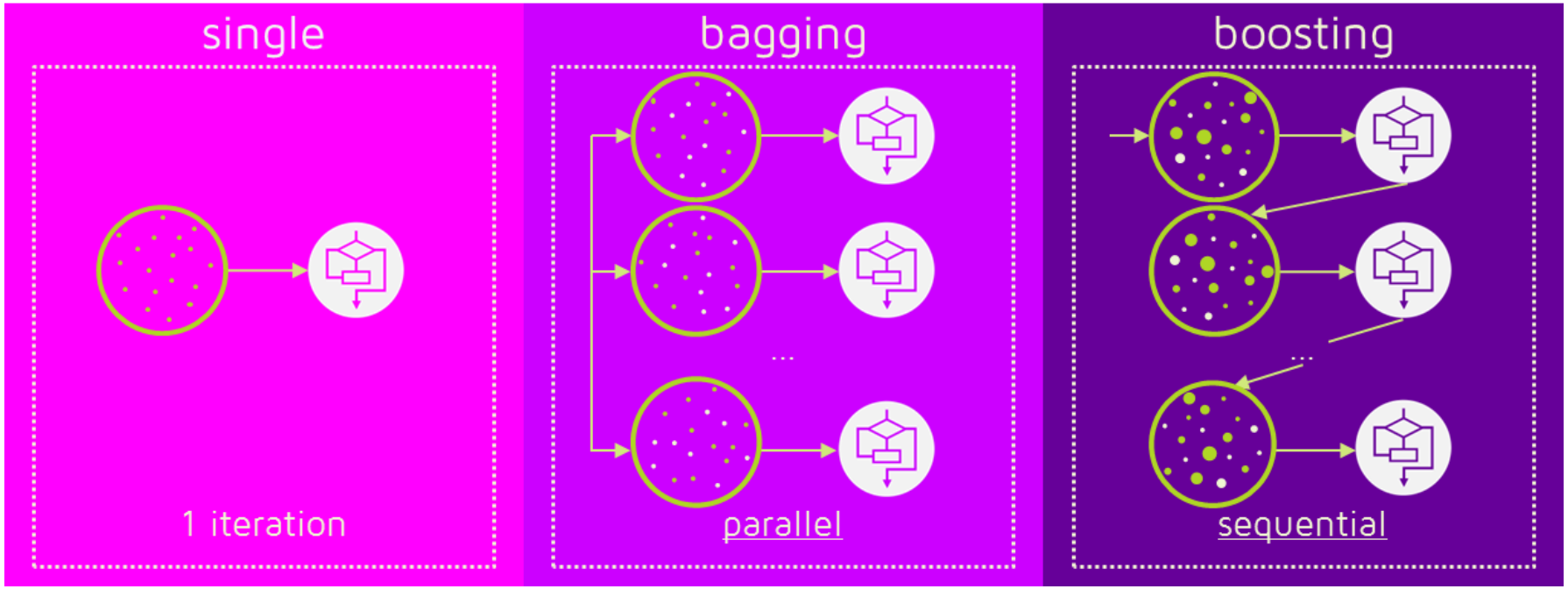
앙상블 학습 (Ensemble Learning)

백지장도 맞들면 낫다! 집단지성! 아니 이게 되네..?
오늘 우리가 다룰 앙상블 기법은 트리 기반 부스팅!



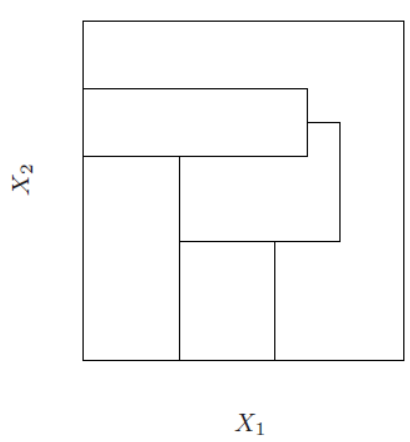
앙상블 학습 (Ensemble Learning)

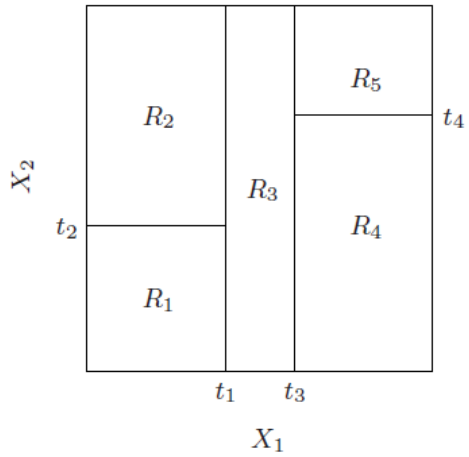
백지장도 맞들면 낫다! 집단지성! 아니 이게 되네..?
오늘 우리가 다룰 앙상블 기법은 트리 기반 부스팅!

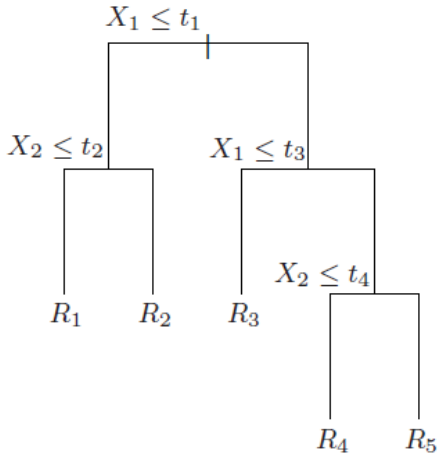


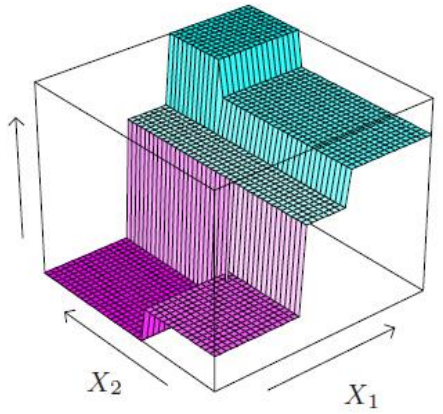
앙상블 학습 (Ensemble Learning)

백지장도 맞들면 낫다! 집단지성! 아니 이게 되네..?
오늘 우리가 다룰 앙상블 기법은 트리 기반 부스팅!





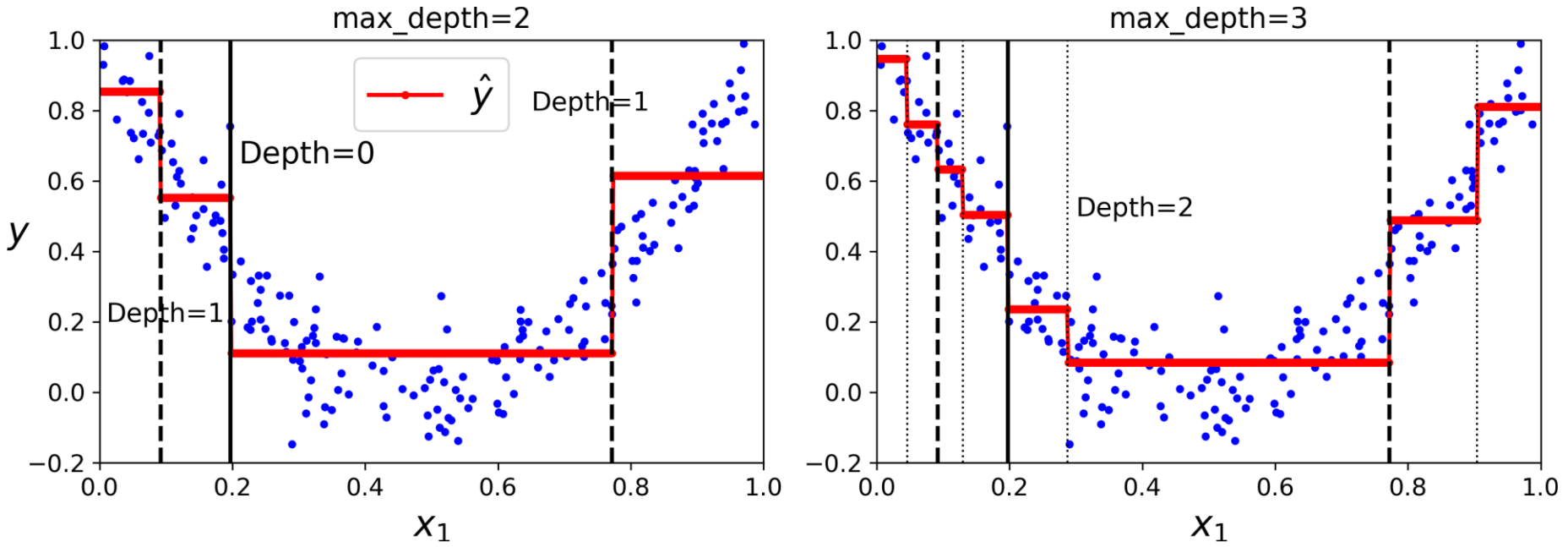




ESC-2021FALL - Tree & Boosting : https://youtu.be/2A5mxMo5jlo

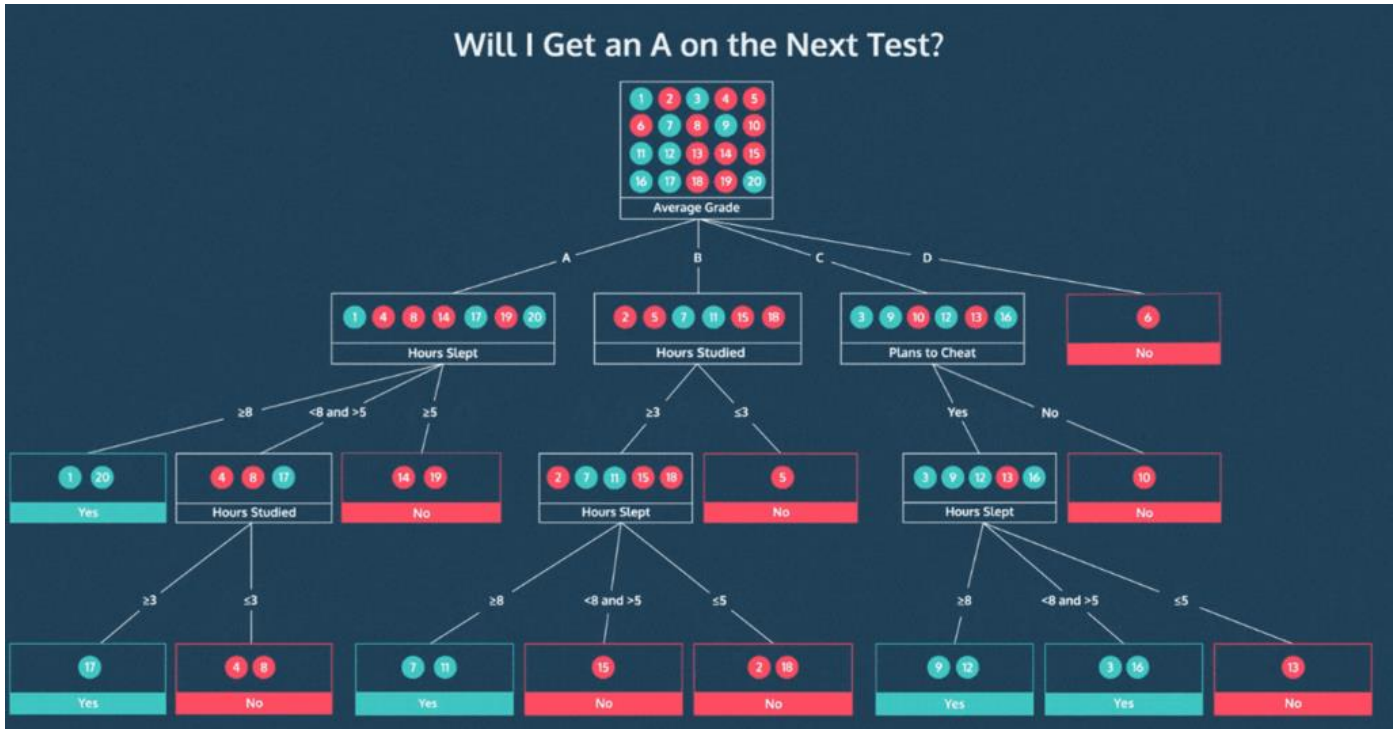
앙상블 학습 (Ensemble Learning)

백지장도 맞들면 낫다! 집단지성! 아니 이게 되네..?
오늘 우리가 다룰 앙상블 기법은 트리 기반 부스팅!



앙상블 학습 (Ensemble Learning)

백지장도 맞들면 낫다! 집단지성! 아니 이게 되네..?
오늘 우리가 다룰 앙상블 기법은 트리 기반 부스팅!



ESC-2021FALL - Bagging & RandomForest : https://youtu.be/8-rEGknAURs

부스팅 Boosting

가중치를 업데이트하면서 찐따 모델에서 헬창 모델로 강화시키자! 메이플 보보보는 왜 안 뜰까...

1

AdaBoost

강자만이 살아남는다!
가중치 폭격기!

2

XGBoost

적을 알고 나를 알면
지피지기 백전백승!

3

LightGBM

더 가볍고 빠르다!
(애플 보고 있냐?)

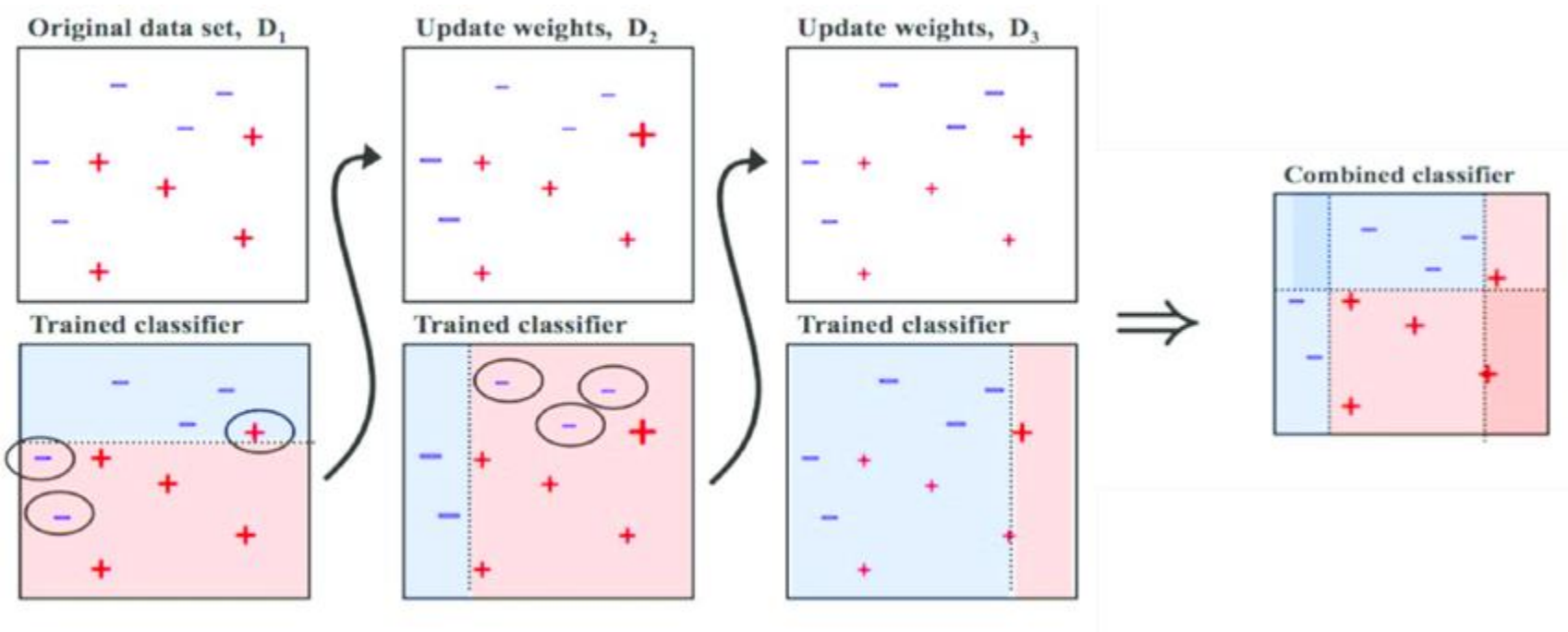
4

CatBoost

고양이 아니라고!
고양이 아니라고!

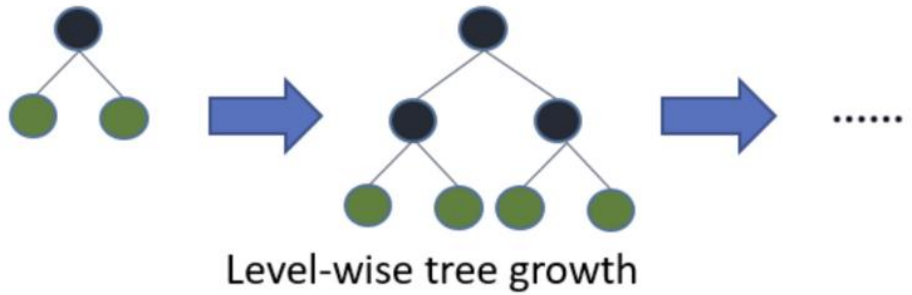
에이다부스트 AdaBoost

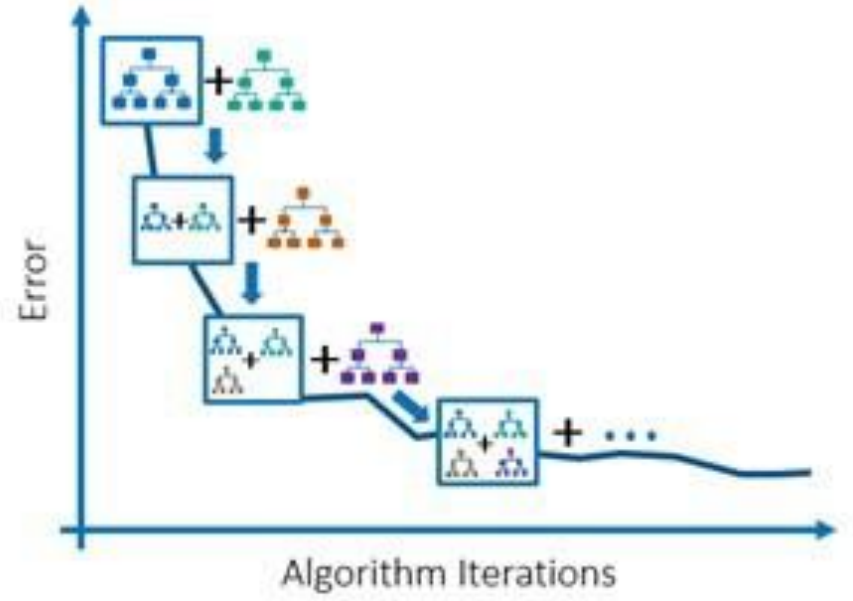
이전 classifier가 잘못 분류한 부분을 adaptive하게 바꾸어 가자! 어떻게? 가중치를 부여해서!
그리고 모든 classifier을 합쳐서 최종 결과로 출력!



익스트림 그레디언트 부스팅 XGBoost

경사하강법을 사용해서 트리 모형의 손실함수를 최소화 시키는 방향으로 병렬 부스팅하자!
XGBoost와 LightGBM의 차이는 Tree Growth 방식에 있다!

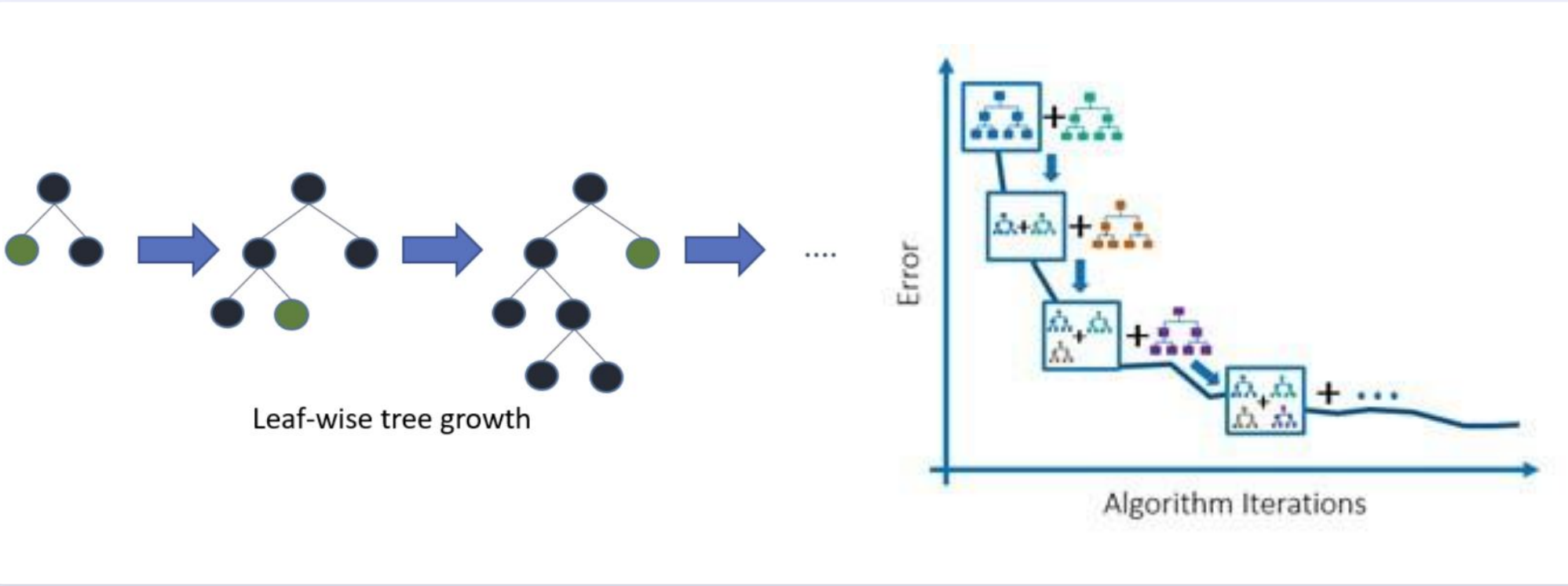




XGBoost: A Scalable Tree Boosting System : https://arxiv.org/abs/1603.02754

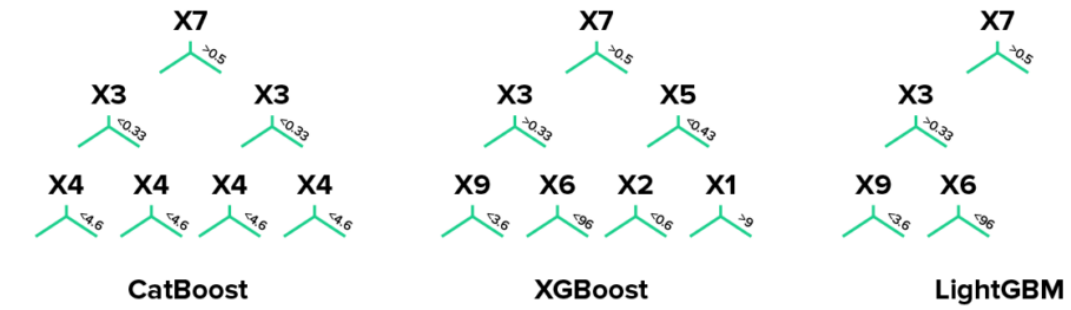
라이트 그레디언트 부스팅 LightGBM

왕위를 계승하는 중입니다.. 두둥..! XGBoost보다 더 빠르면서 성능도 준수한 기특한 녀석!
하지만 둘 중 누가 더 낫다는 섣부른 판단은 금물! 모델은 연습장일 뿐, 답지가 아님!



카테고리컬 부스팅 CatBoost 야옹

범주형 feature를 처리하는 데 전 집중 호흡! 범주형 feature 인코딩, 중복 feature 처리 등등!
But, 수치형 데이터가 대부분이거나, 결측치가 많이 있는 자료에는 적용하기 힘들다!



country	hair color	class_label
India	black	1
India	black	1
India	black	1
india	black	1
russia	white	0
russia	white	0
russia	white	0
russia	white	0

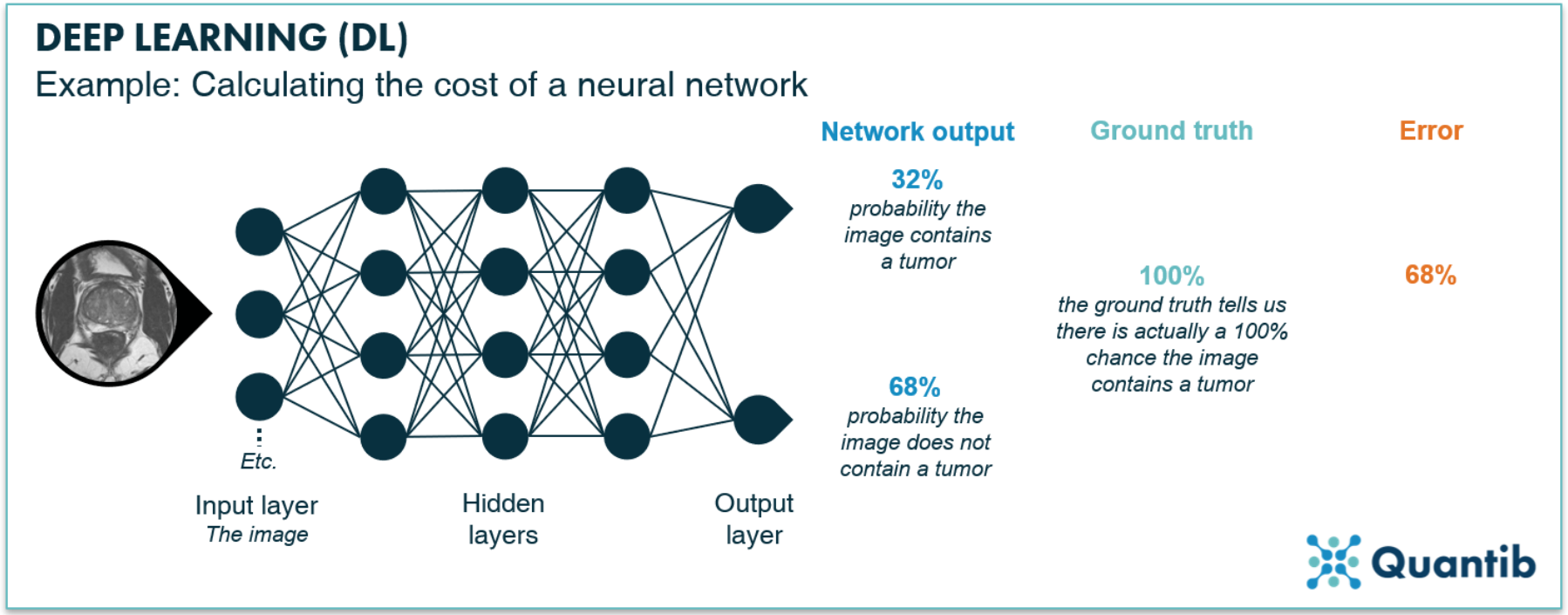
https://towardsdatascience.com/how-do-you-use-categorical-features-directly-with-catboost-947b211c2923

캐글 산탄데르 고객 만족 예측

https://www.kaggle.com/c/santander-customer-satisfaction

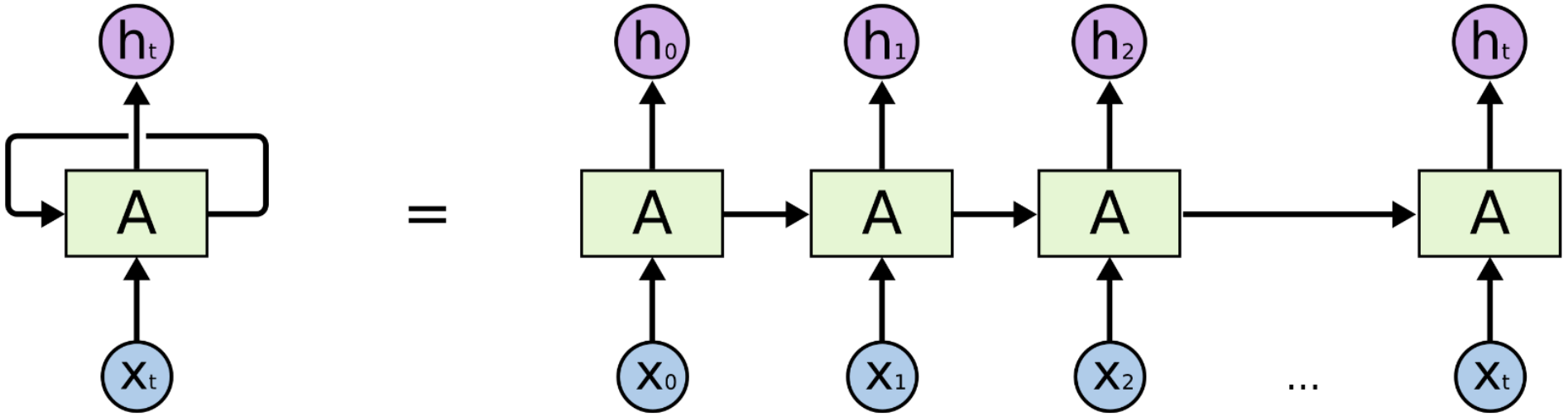
딥러닝 Deep Learning

Input, output은 있는데, 괜찮은 예측 모델을 못 찾겠을 때의 강력한 대안! 회귀분석의 상위 호환!
But! 과유불급이라 하거늘! input의 정보가 output까지 잘 전달되는 것이 제일 중요!



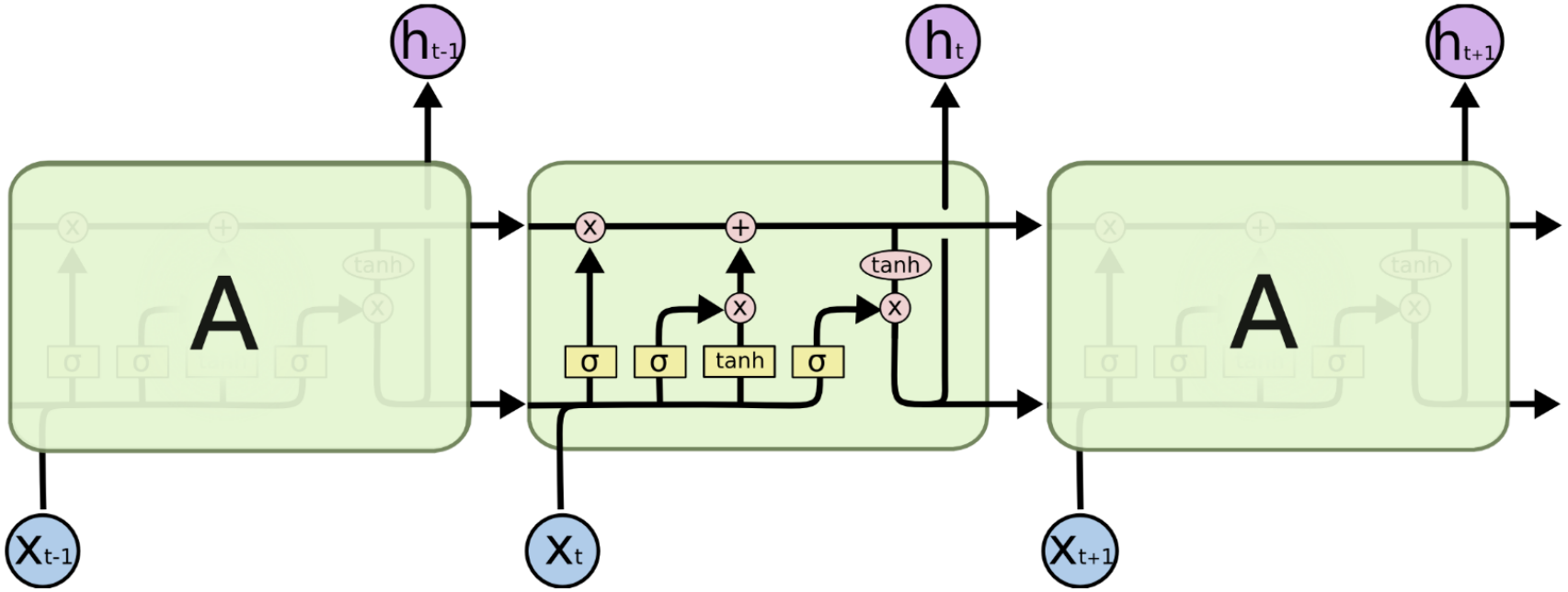
순환신경망 Recurrent Neural Network

그... 내가... 뭐라... 했더라...? 음성인식, 언어 모델링, 번역 등등 안 되는 게 없다!
문제는 장기 의존성! 순서상 데이터 사이의 갭이 커질수록 두 데이터의 정보를 연결하기 힘들어진다!



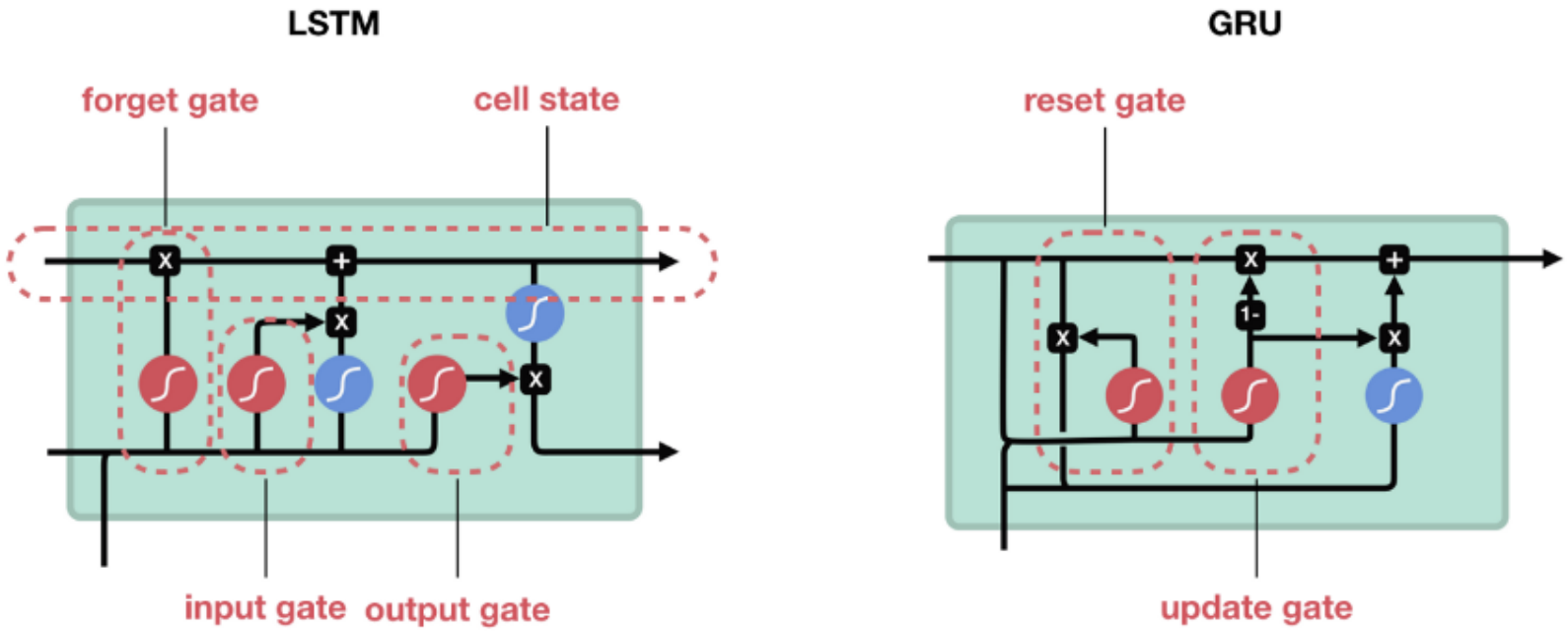
LSTM (Long Short-Term Memory Model)

잊어버릴 건 잊고, 새로 추가할 건 추가하고, 남길 건 남겨 놓자! 그래서 Cell State에 중요한 정보들이 남아있도록!



GRU (Gated Recurrent Units)

4달라는 너무 많소. 1달라로 합시다. 1달라... 1달라..! 오케이 땡큐 1달라!
LSTM보다 간단한 구조! 그래서 학습할 파라미터 수도 훨씬 적다! 하지만 항상 둘 다 확인해보자!

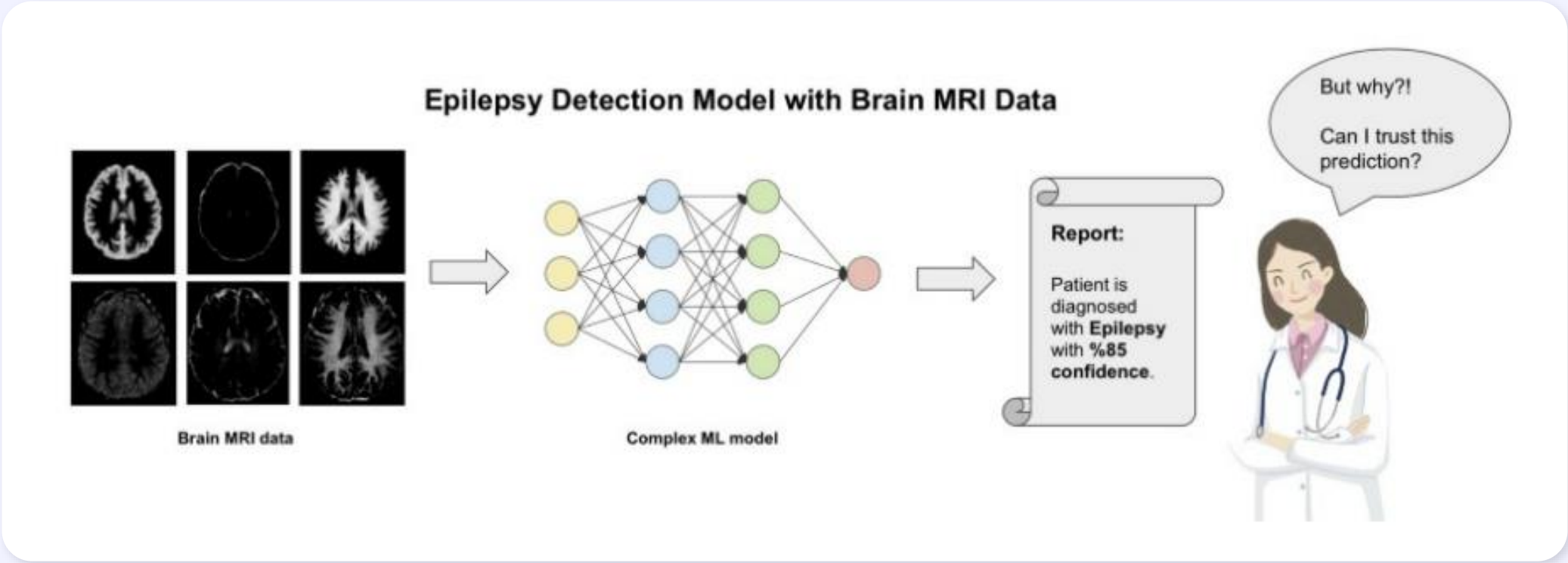


삼성전자 주식 예측

https://finance.yahoo.com/quote/005930.KS?p=005930.KS

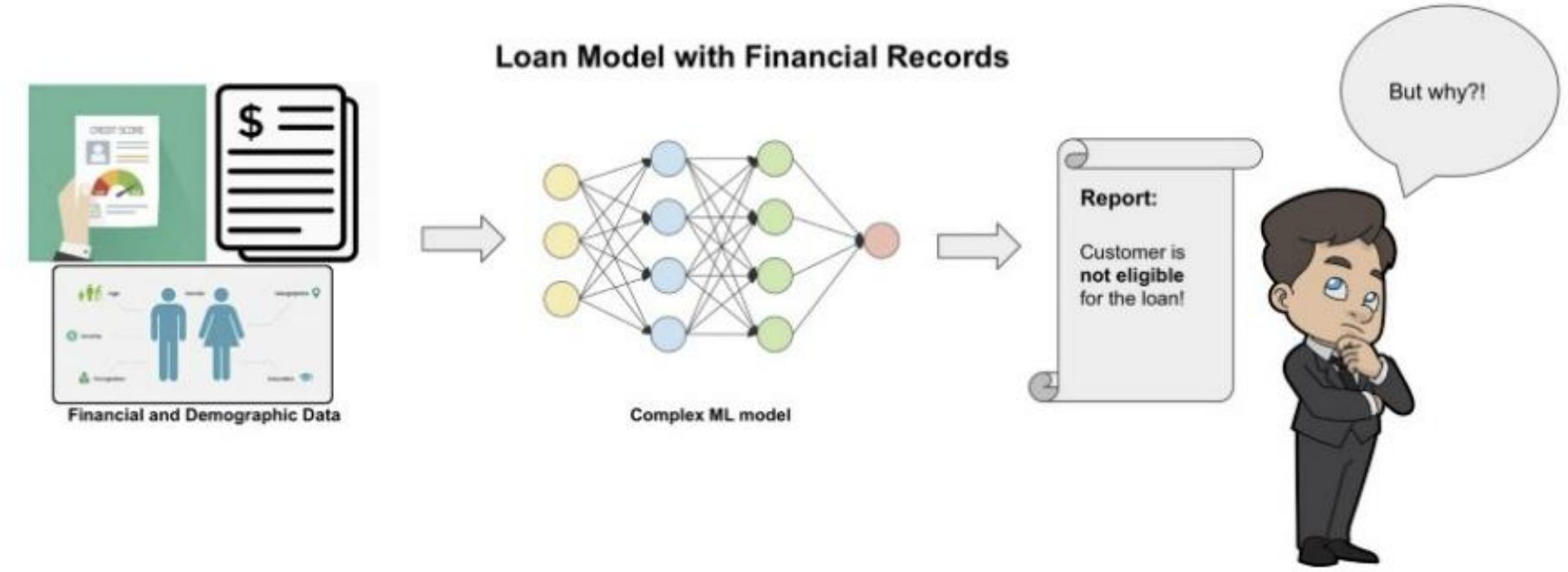
설명 가능성 Explainability

‘과거를 이해하고 미래를 디자인한다’ - 그래픽 디자이너 레오나르도 소놀리 -
모델의 학습 결과만 사용하는 데 그치지 말고, 모델이 어떠한 체제로 동작하고 동작하지 않는지, 시스템이 왜 실패하고 성공하는지를 파악하자!



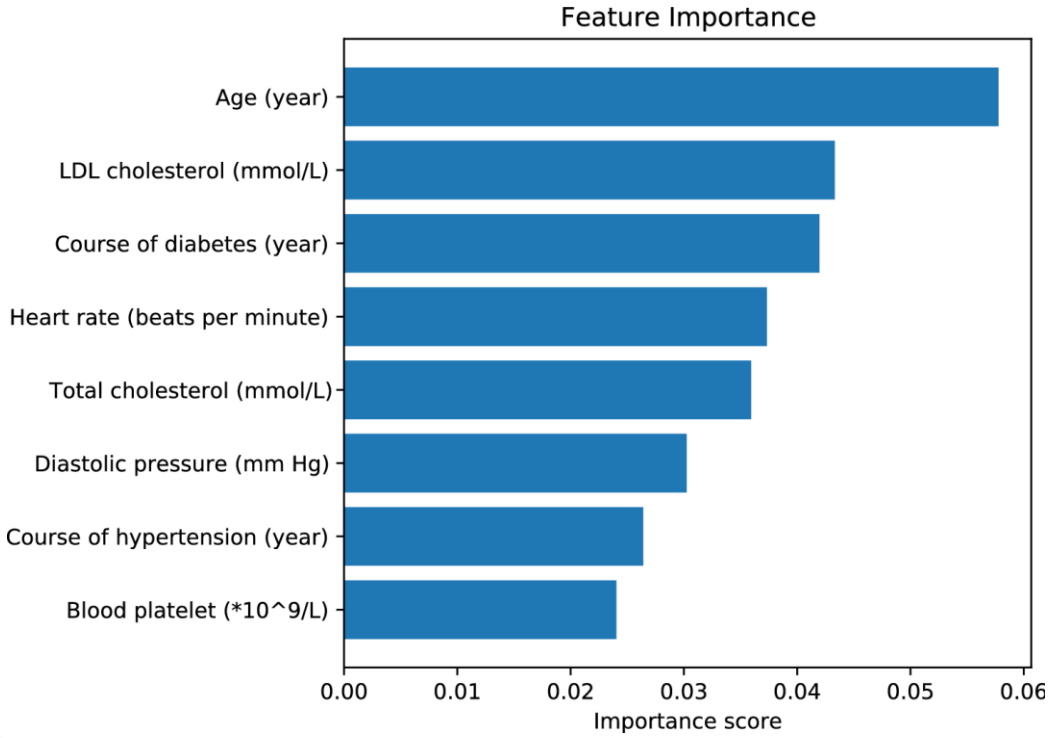
설명 가능성 Explainability

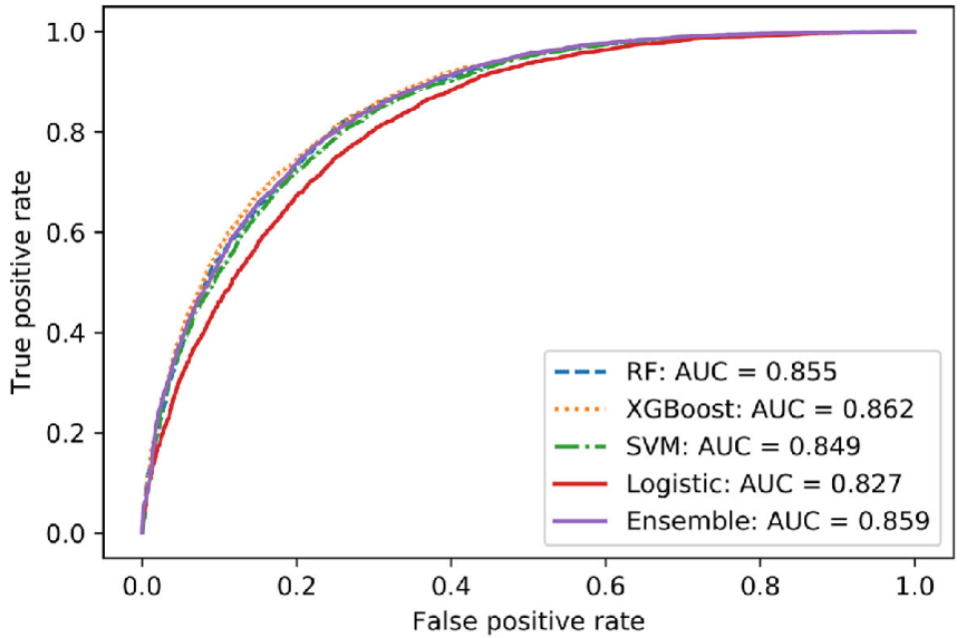
‘과거를 이해하고 미래를 디자인한다’ - 그래픽 디자이너 레오나르도 소놀리 -
모델의 학습 결과만 사용하는 데 그치지 말고, 모델이 어떠한 체제로 동작하고 동작하지 않는지, 시스템이 왜 실패하고 성공하는지를 파악하자!



피처 중요도 Feature Importance

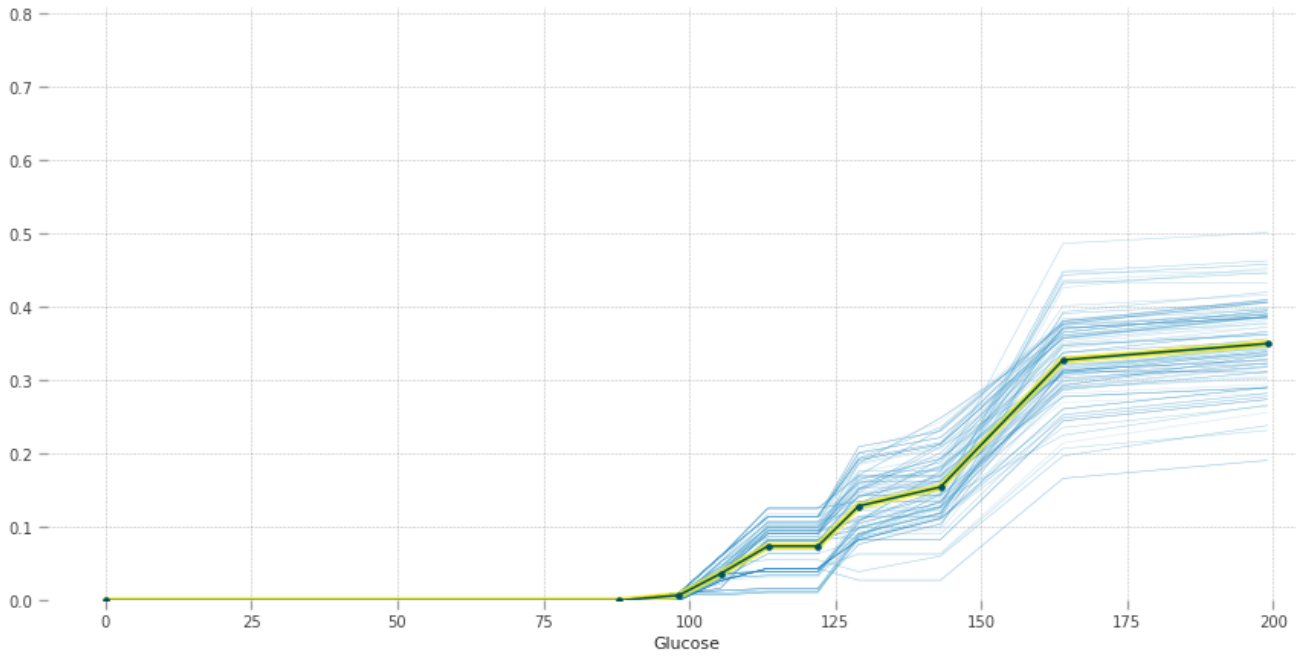
데이터의 피처가 알고리즘의 정확한 분류에 얼마나 큰 영향을 미치는지 분석해보자!





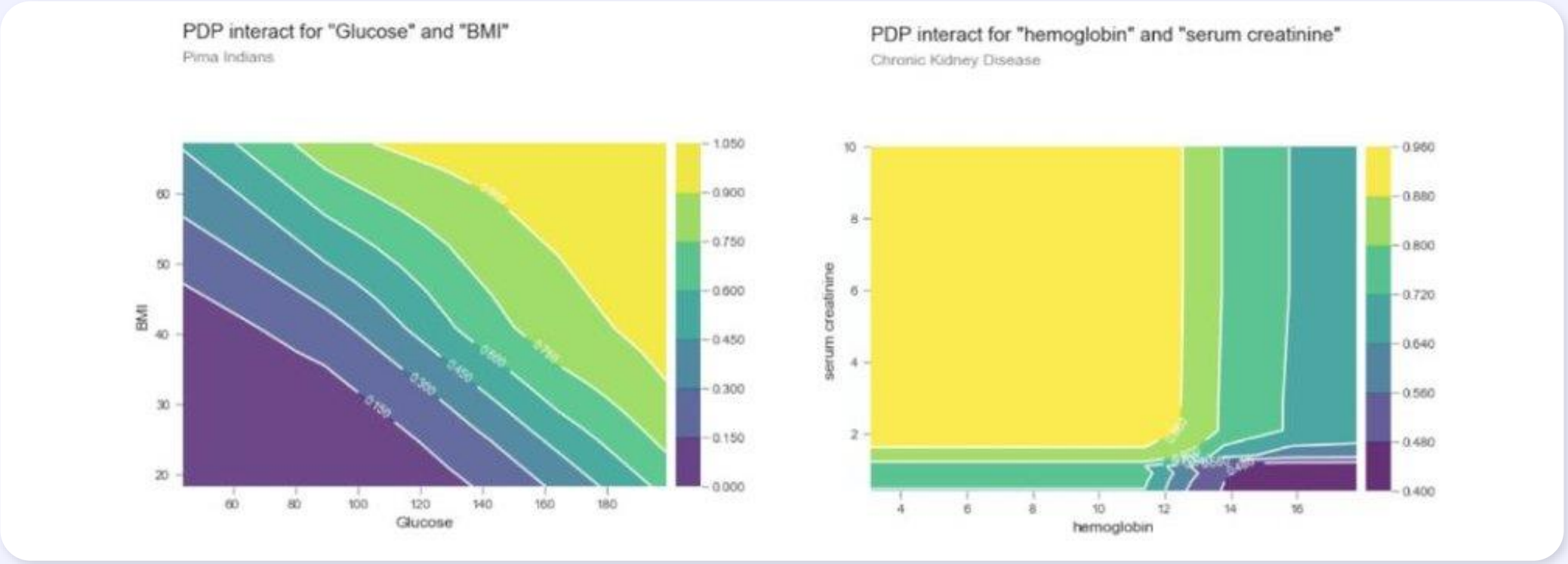
부분 의존성 플롯 Partial Dependence Plots

피처의 수치를 선형적으로 변형하면서 알고리즘 해석 능력이 얼마나 증가하고 감소하는지 관찰하는 방식!
피처의 값이 변할 때 모델에 미치는 영향을 시각적으로 이해할 수 있다!



부분 의존성 플롯 Partial Dependence Plots

피처의 수치를 선형적으로 변형하면서 알고리즘 해석 능력이 얼마나 증가하고 감소하는지 관찰하는 방식!
피처의 값이 변할 때 모델에 미치는 영향을 시각적으로 이해할 수 있다!

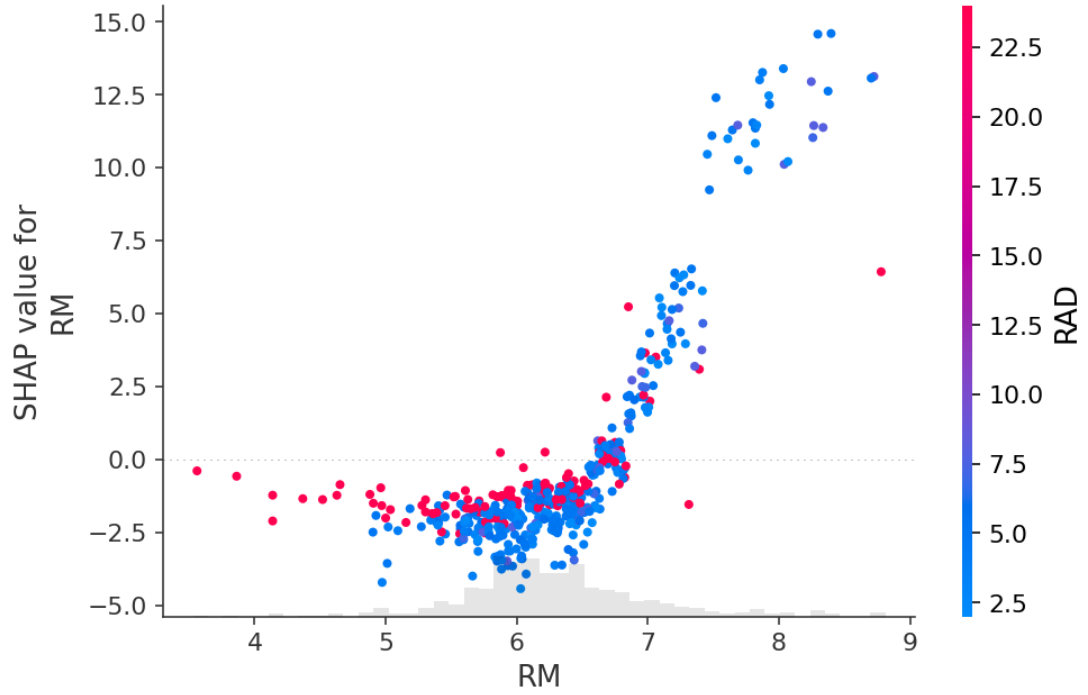


SHAP(Shapley Additive exPlanations)

섀플리 값이란? 전체 성과를 창출하는 데 각 참여자가 얼마나 공헌했는지를 수치로 표현한 값!
force_plot 메서드는 특정 데이터에 대한 섀플리 값을 상세하게 분해하고 시각화해준다! 빨간색은 긍정적 영향, 파란색은 부정적 영향!

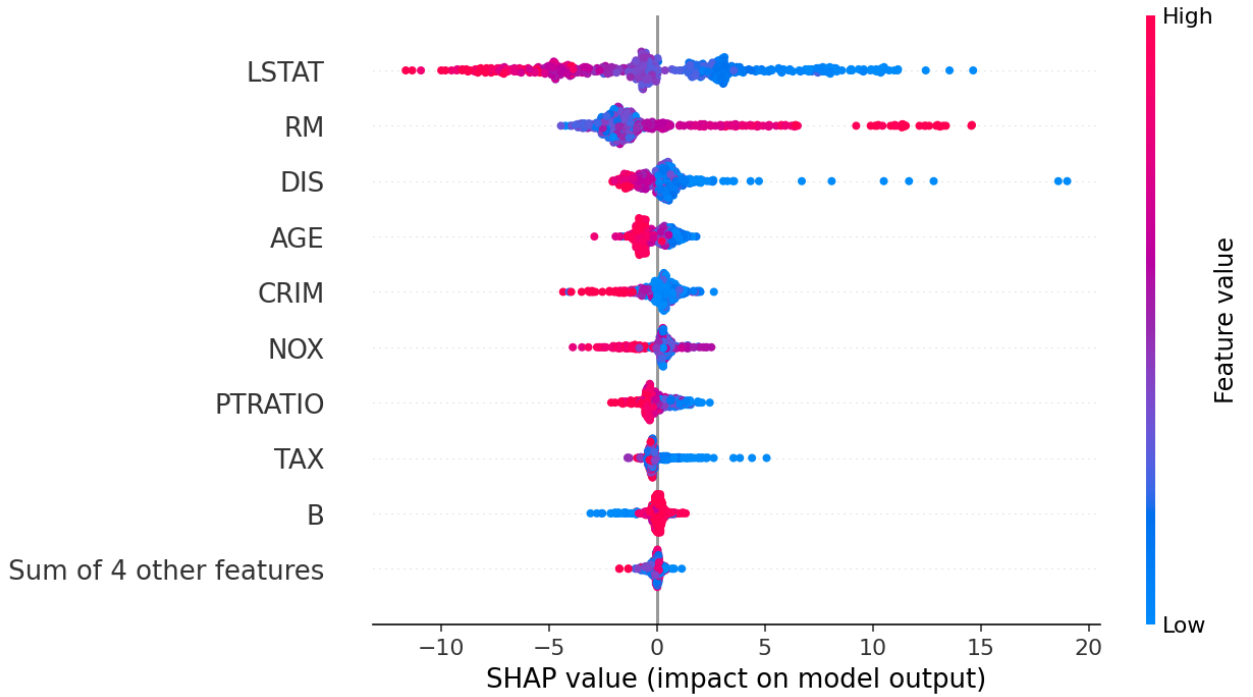
SHAP(Shapley Additive exPlanations)

dependence_plot 메서드는 하나의 피처가 전체 예측에 미치는 영향력을 계산하고 시각화해준다!
빨간색은 다른 피처들보다 이 피처의 영향을 많이 받는 데이터! 파란색은 반대!



SHAP(Shapley Additive exPlanations)

summary_plot 메서드는 전체 피처들이 샤플리 값 결정에 어떻게 관여하는지 시각화해준다!
빨간색은 그 지점에 해당하는 행 피처의 영향이 컸음을 의미! 파란색은 반대!



보스턴 주택 가격 결정 요소 구하기

https://github.com/slundberg/shap

감사합니다

띠용띠용 워익워익
띠용띠용 워익워익