깔삼하게 대출받자!

P - SAT 김종혁 이성희
YBIGTA 김나현 허승민

목차

1 깔삼하게 대출받자!

1 주차 리뷰

지난 주차 진행 방향

데이터 셋

모델을 돌리기 위한 DATA SET을 만드는 작업까지 진행

- **최종 데이터 set**
  - FEATURE : 423
  - OBS : 30만개
- **NA 를 regression 으로 채워 넣음**
- **범주형 → One hot encoding**

지난 주차 진행 방향

FEEDBACK

IMPORTANCE PLOT 를 기준으로 하는 FEATURE SELECTION

Tree base 모델에서 feature importance 는
변수의 설명력에 대한 지표
Feature selection으로 적합하지 않다고 하지만,

모델링에서 설명력이 매우 미미한 변수들이 존재한다면

그 변수들은 제거 해도 된다고 판단했다.

2 깔삼하게 대출받자!

진행 방향 소개

데이터 전처리 방향 재설정

One-hot encoding 에서 Factorization으로 방향을 재설정

One-hot encoding

Color
Yellow
Yellow
Red
Green

→

Yellow	Red	Green
1	0	0
1	0	0
0	1	0
0	0	1

방법	범주개수 k라면, k-1개의 더미변수 생성
장점	범주에 관한 영향력 확인
단점	차원이 높아짐

Factorization

Color
Yellow
Yellow
Red
Green

→

Color
1
1
2
0

방법	– 범주를 0,1,2,3 과 같은 수치로 변환 – 범주형 데이터를 연속형 데이터로 간주
장점	예측력이 높다
단점	범주 해석이 불가능함

데이터 전처리 방향 재설정

NA를 하나의 범주로 간주하고 별도의 처리과정을 거치지 않는 방향으로 재설정

Regression을 이용한 NA Imputation

	count_NA	ratio_NA
EXT_SOURCE_1	157655	56.264
EXT_SOURCE_2	599	0.214
EXT_SOURCE_3	55268	19.724

? NA imputation이 변수 자체의 의미를 희석시키는 영향에 대한 우려

XG Boost
Light Gradient Boost
CAT Boost

모두 NA를 처리하는 방법이 내장되어 있다.

분류 예측력에 NA의 영향을 반영시키기로 결정

Unbalanced data를 위한

평가 지표

Accuracy 는 적절하지 않은 평가지표이다.

		Predicted		
		Positive	Negative	
Observed	Positive	TP	FN	P
	Negative	FP	TN	N

	0	1	Accuracy	F1
True	950	50	0.95	0
Predict	1000	0		

- 가장 대표적으로 사용되는 지표
- 실제 집단과 일치하게 분류한 비율
- 모델이 얼마나 정확하게 분류하는지를 의미

$$\frac{\text{제대로 분류된 DATA}}{\text{전체 DATA}} = \frac{TP+TN}{(TP+FP+FN+TN)=P+N}$$

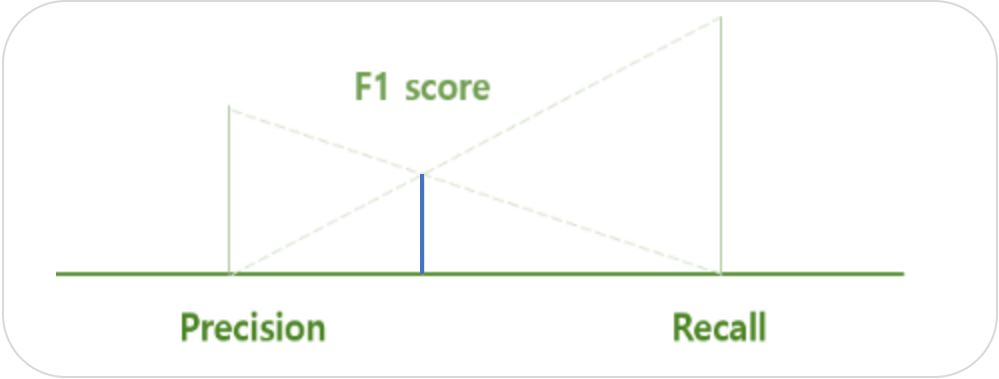
Accuracy는 데이터가 한쪽으로 치우쳐 있으면 정확한 예측이 어렵다

Unbalanced data를 위한

평가 지표

F1-Score

- Precision과 Recall 의 조화 평균



$$\text{F1 Score} = 2 \times \frac{\text{Precision X Recall}}{\text{Precision + Recall}}$$

Precision(정밀성)

- Positive로 예측한 내용 중에서 실제 Positive의 비율을 의미

- $\frac{TP}{TP+FP}$

Recall(재현율)

- 모델이 얼마나 정확하게 Positive 값을 찾는가를 의미하며 Sensitivity(민감도)라고도 한다.

- $\frac{TP}{TP+FN}$

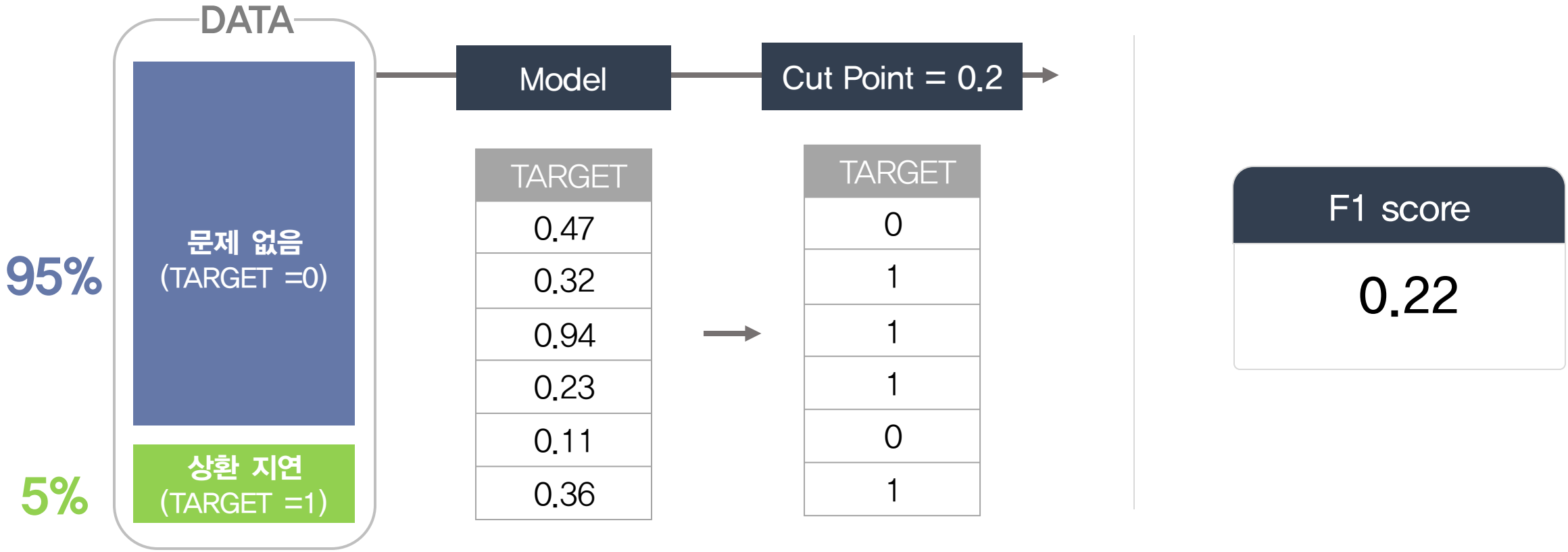
예측하고자 하는 타겟 값을 기준으로 평가하기 때문에 Unbalanced data에 적합

3 깔삼하게 대출받자!

높은 예측력을 위한 기법

예측력을 높이기 위한 **방법** (1)

cut point

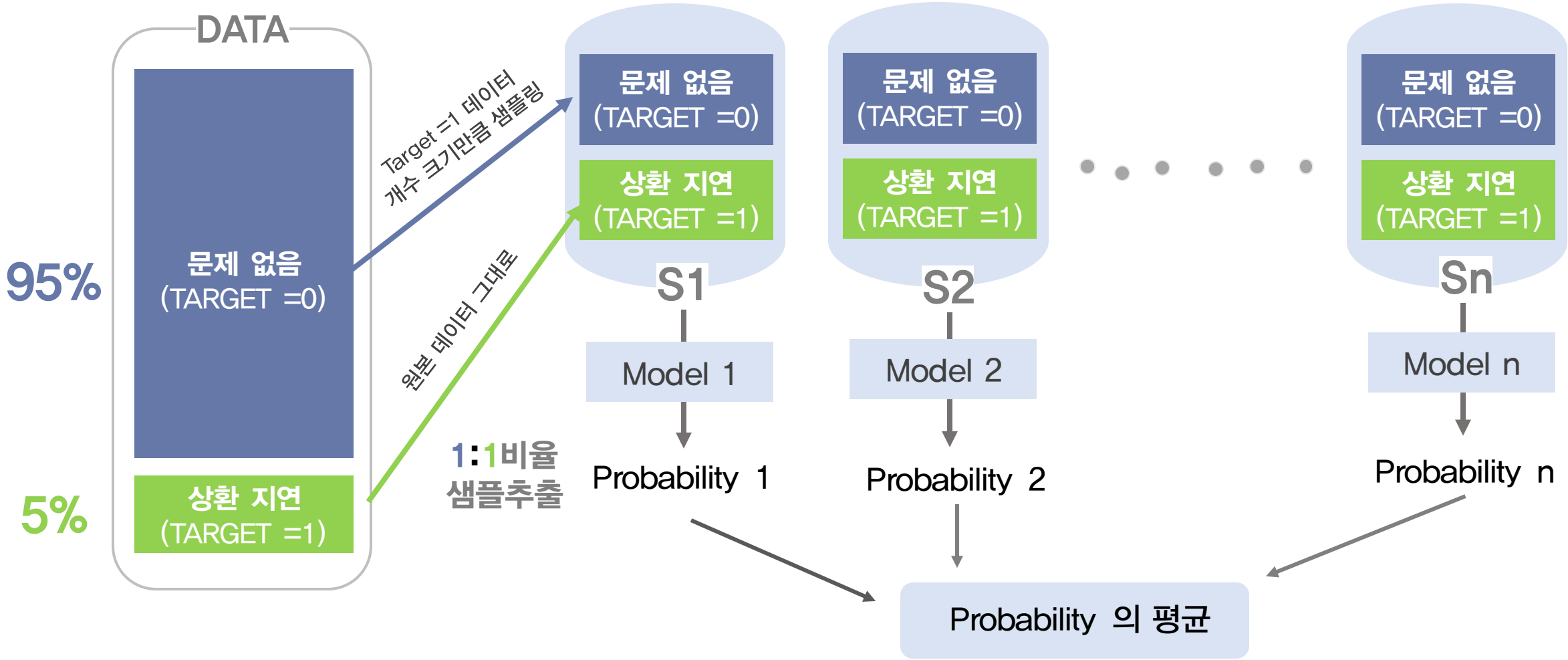


Cut point를 0.2로 지정하였을 때 낮은 예측력을 보였기에 디폴트인 0.5로 설정

예측력을 높이기 위한 **방법 (2)**

샘 플 링

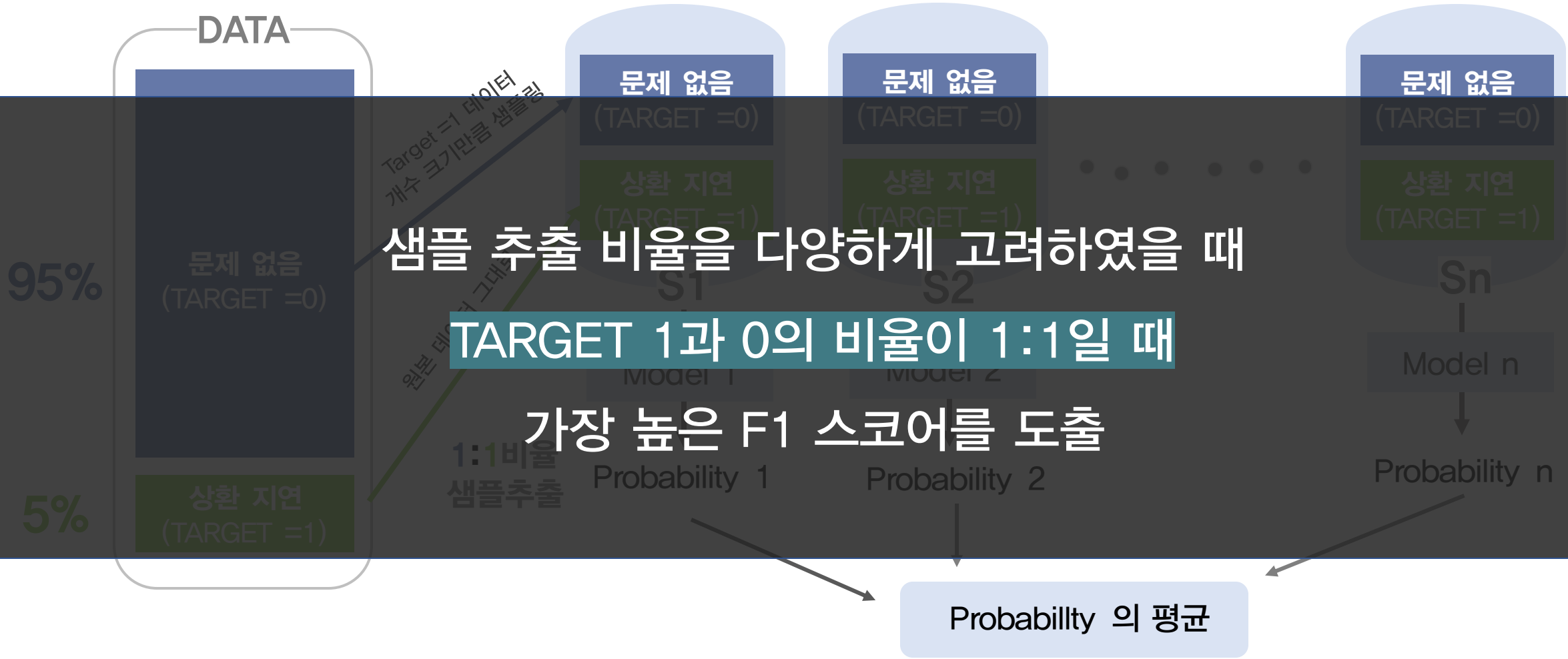
언더 샘플링 앙상블



예측력을 높이기 위한 **방법 (2)**

샘 플 링

언더 샘플링 앙상블



샘플 추출 비율을 다양하게 고려하였을 때

TARGET 1과 0의 비율이 1:1일 때

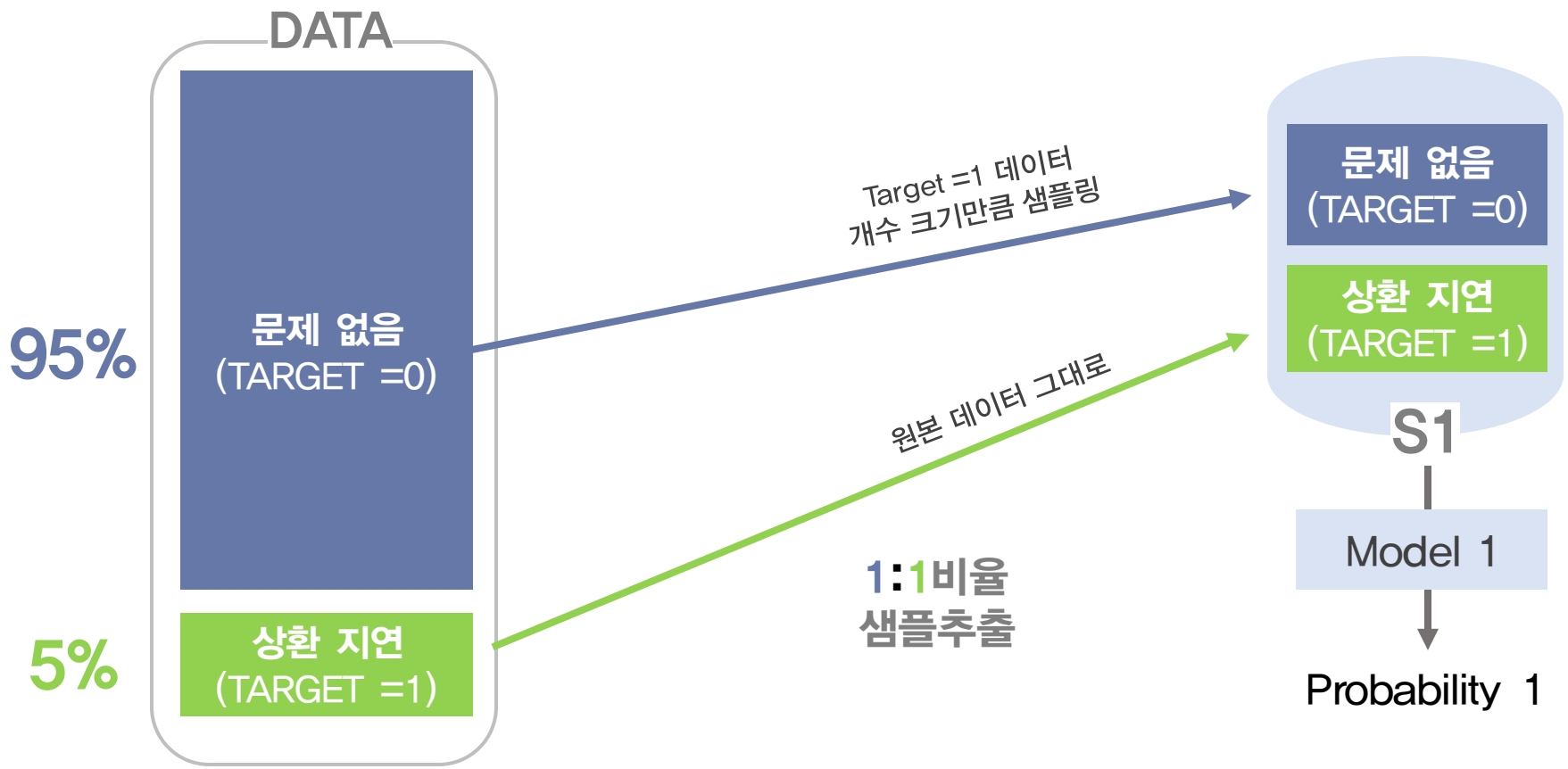
가장 높은 F1 스코어를 도출

예측력을 높이기 위한 **방법 (2)**

샘 플 링

언더 샘플링

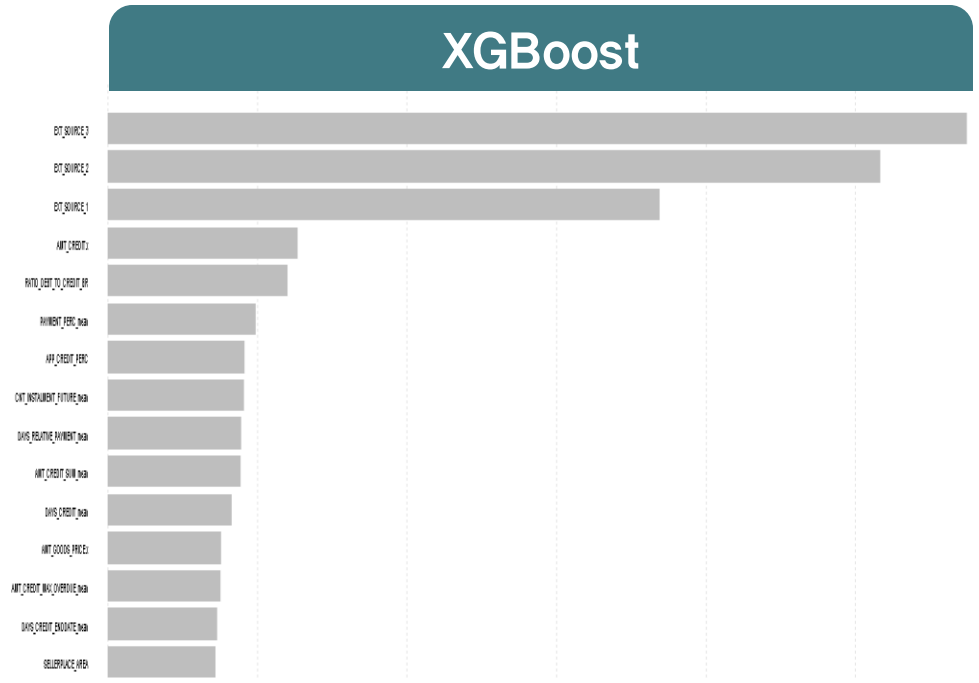
현실적으로, 컴퓨팅 파워를 고려하여 언더 샘플링을 채택



예측력을 높이기 위한 **방법 (3)**

파생 변수

external source를 활용한 변수 생성



SUM, MEAN, MODE, MEDIAN, MAX, MIN…

EXT_SOURCES_PROD — external source의 곱

' EXT_SOURCE1 ' * ' EXT_SOURCE2 ' * ' EXT_SOURCE3 '

EXT_SOURCES_WEIGHTED — external source와 가중치의 곱

' EXT_SOURCE2 ' *1+ ' EXT_SOURCE1 ' *2 + ' EXT_SOURCE3 '*3

EXT_SOURCES_MIN — external source의 최소값

EXT_SOURCE1, EXT_SOURCE2, EXT_SOURCE3 중 최소값

EXT_SOURCES_MAX — external source의 최대값

EXT_SOURCE1, EXT_SOURCE2, EXT_SOURCE3 중 최대값

예측력을 높이기 위한 **방법 (3)**

파생 변수

APPLICATION_TEST

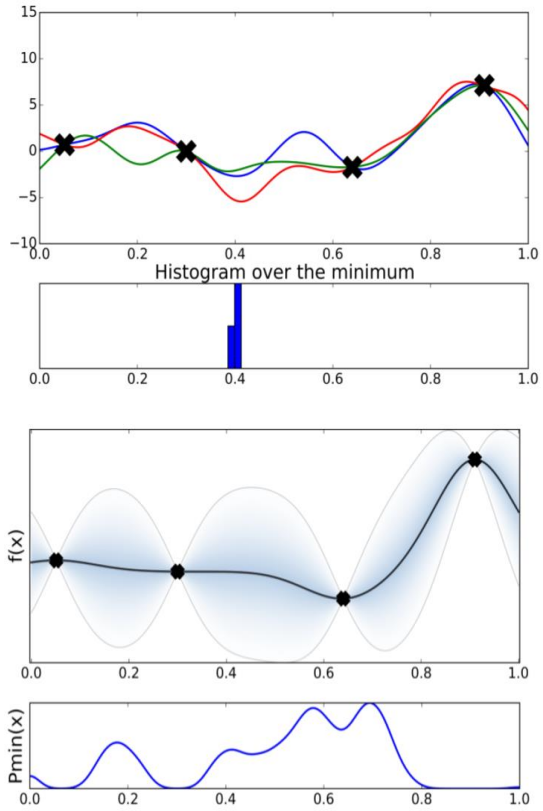
어플리케이션 데이터를

위주로 FEATURE 새로 생성

CAR_TO_BIRTH_RATIO	차를 소유했던 기간 비율
OWN_CAR_AGE / DAYS_BIRTH	
CAR_TO_EMPLOY_RATIO	직장을 가진 시기에서 차를 소유했던 기간 비율
OWN_CAR_AGE / DAYS_EMPLOYED	
PHONE_TO_BIRTH_RATIO	폰을 소유했던 기간 비율
DAYS_LAST_PHONE_CHANGE / DAYS_BIRTH	
CAR_TO_EMPLOY_RATIO	직장을 가진 시기에서 폰을 소유했던 기간 비율
DAYS_LAST_PHONE_CHANGE / DAYS_EMPLOYED	

예측력을 높이기 위한 **방법 (4)**

파라미터 튜닝



Bayesian optimization

- 최소한의 함숫값만으로 최적화 문제를 풀도록 하는 **데이터를 선택하는 알고리즘**
- 그리드 서치보다는 **시간이 단축**되고 랜덤 서치보다는 **정확성이 높아**진다

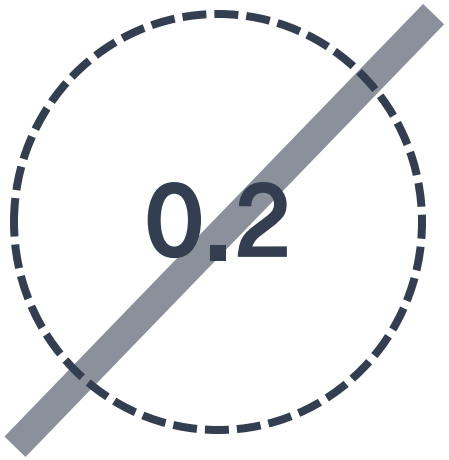
Grid Search

하이퍼 파라미터 공간에서 임의로 지정한 하위 집합을 단순하게 모든 조합을 다 탐색

예측력을 높이기 위한 **방법 (4)**

기법 최종 정리

cut point | 샘플링 | 파생변수 | 파라미터 튜닝



Target =1 와
Target =0 의
비율을 1:1로 하는

언더 샘플링

APPLICATION_TEST

EXT_SOURCE

OWN_CAR_AGE

DAYS_LAST_PHON
E_CHANGE

Bayesian optimization

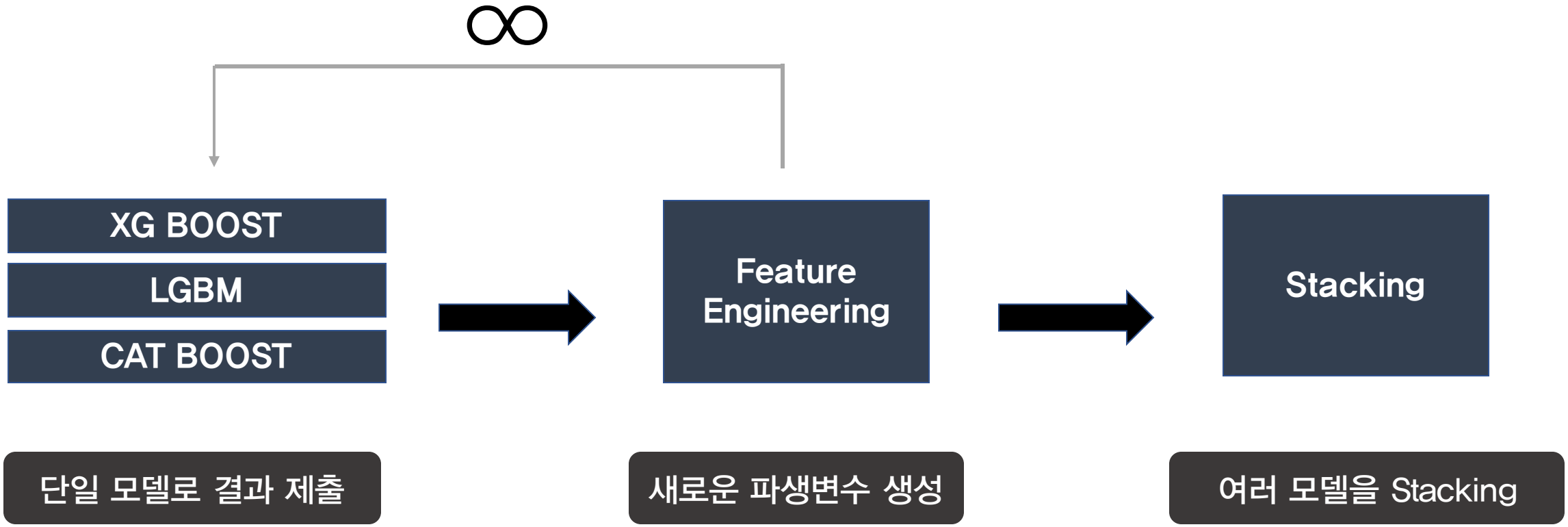
Grid Search

4 깔삼하게 대출받자!

모 델 링

모델링 기법 선정

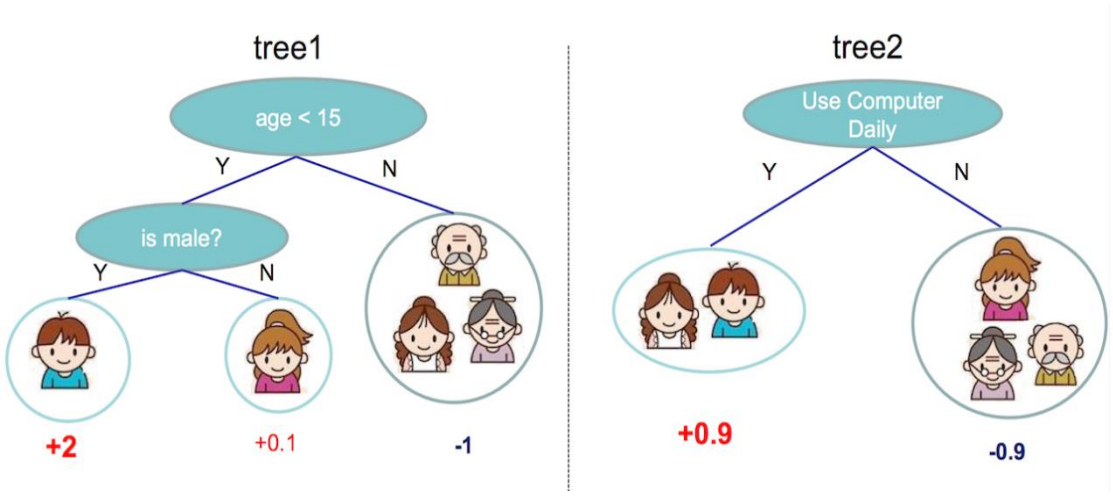
모 델 링


∞
XG BOOST
LGBM
CAT BOOST
Feature Engineering
Stacking
단일 모델로 결과 제출
새로운 파생변수 생성
여러 모델을 Stacking

모델링 기법 선정 (1)

XG Boost

CART(classification and regression tree)



decision tree에 비해 분류의 정확성 뿐만 아니라
같은 분류 결과를 가진 모델끼리도 스코어를 통해
모델의 우위를 비교할 수 있다.

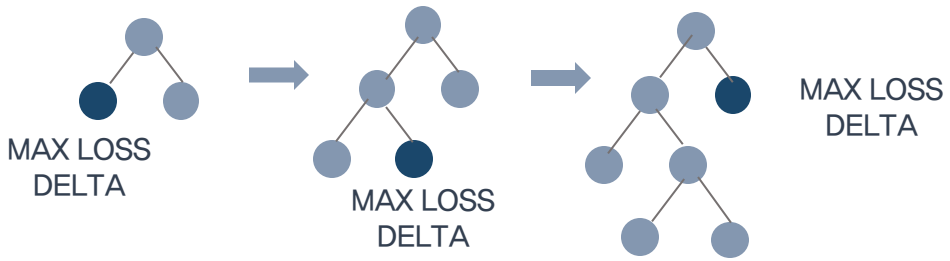
- Regularization 으로 과적합 방지
- Cross Validation 이 내장
- 병렬처리로 빠른 속도
- 평가지표의 customized metric로 유연한 해석이 가능

모델링 기법 선정 (2)

LGBM

- 피쳐들의 히스토그램을 만들어 근사치의 분할이 진행
- 대용량의 데이터를 빠르게 학습가능 (10000 이상)
- Over-fitting에 예민하기 때문에 적은 데이터에는 부적함

1 Gradient-based One-Side Sampling (GOSS)



Untrained 개체가 정보 획득에 더 기여하기 때문에

max loss delta 로 트리를 확장하며 loss를 줄인다

2 Exclusive Feature Bundling (EFB)

feature1	feature2	feature_bundle
0	2	6
0	1	5
0	2	6
1	0	1
2	0	2
3	0	3
4	0	4

변수 개수를 줄이기 위한 거의 손실 없는 방법

변수들은 0이 아닌 값을 동시에 갖는 일이 거의 없는 배타적인 변수들을 묶는다

모델링 기법 선정 (3)

CAT Boost

- 범주형 데이터가 많은 데이터 셋에서 높은 예측력
- 모델 튜닝 없이 default값으로만 좋은 성능을 보여준다.

1 Ordered TBS (Target-Based Statistics)

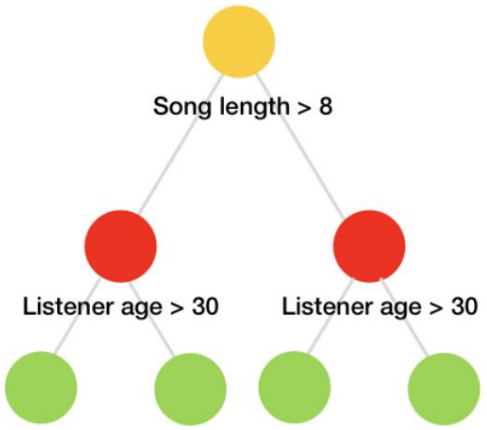
범주형 데이터를 연속형 데이터로 변환

Calculating ctr for the i-th bucket ($i \in [0; k-1]$):

$$ctr_i = \frac{countInClass + prior}{totalCount + 1}, \text{ where}$$

범주형 데이터를 분포만을 가지고 변환할 경우
Train 데이터와 test 데이터가 다른 분포를 가지면
예측력이 떨어지기 때문에 이를 보완!

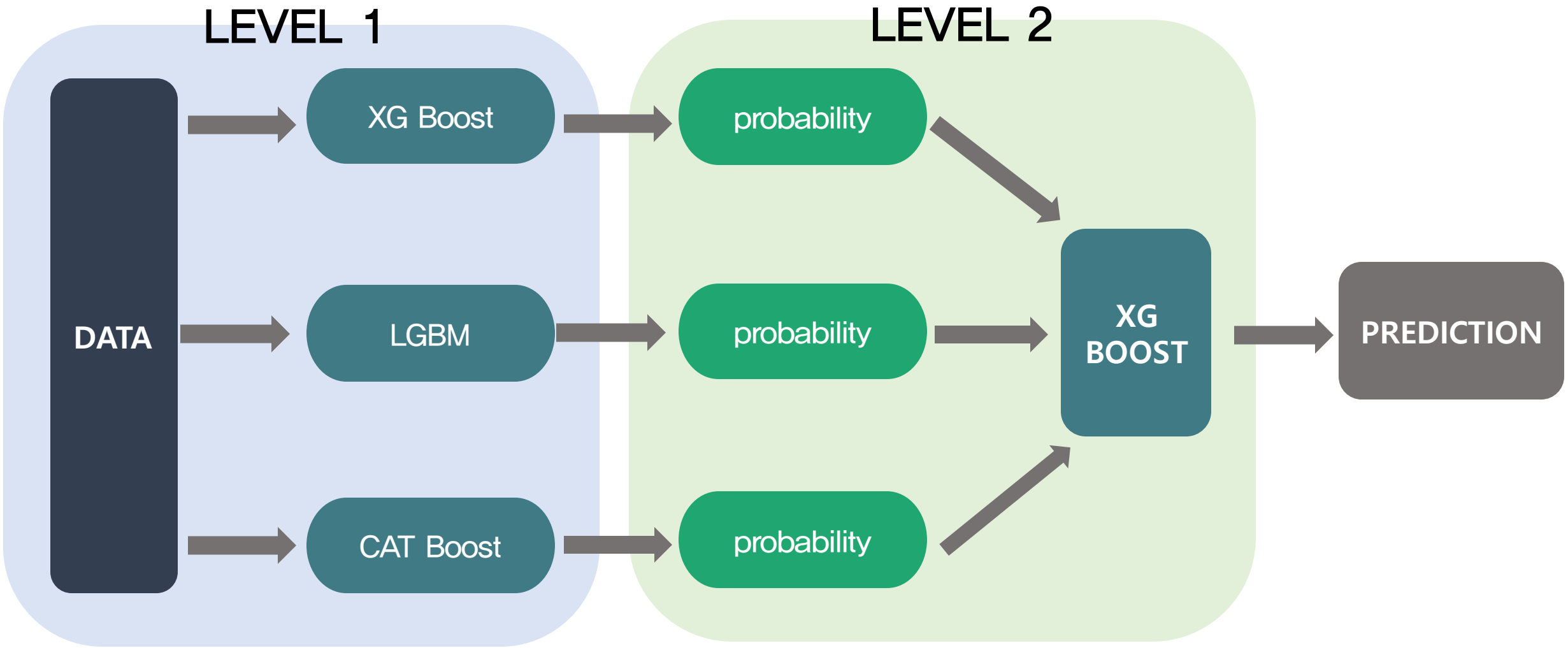
2 level wise tree



대칭형 트리로
Overfitting 을 방지

모델링 기법 선정 (4)

STACKING



모델링 기법 선정

손실 함수 vs 평가 지표

손실 함수

- 어떠한 모델 내부의 파라미터 (가중치) 최적인지 찾기 위한 함수
- XG Boost, LGBM, CAT Boost 모두 디폴트 손실 함수인 Log loss 활용

평가 지표

- Cross – Validation 에서 hyper parameter 학습 시 사용
- 모델 자체의 성능 평가 지표로도 사용

F1 스코어

손실 함수에는 미분이 불가능한 f-score를 활용할 수 없어 log loss를 활용

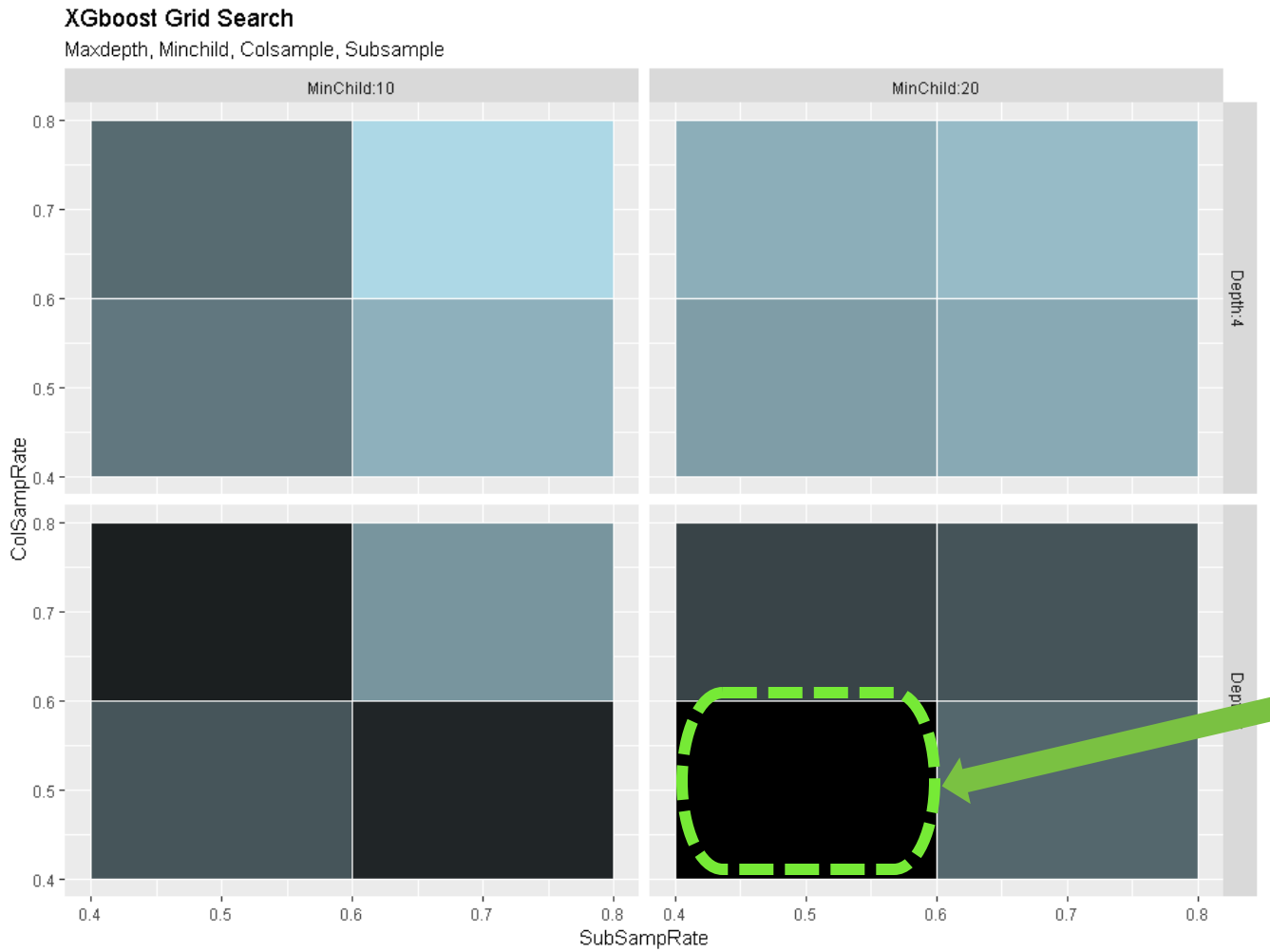
5 깔삼하게 대출받자!

모델링 결과

모델링 평가

XG Boost

Parameter tuning



Grid Search

가장 높은 F1값을 가지는
max_depth, minchild
Colsample, subsample 조합

모델링 평가

XG Boost

Parameter tuning

N round 및 parameter 결과값

Max_depth	트리의 최대 깊이	6
Min_child_weight	노드 분할 시 필요한 최소 instance weight	20
Subsample	트리에서 obs 샘플링 비율	0.5
Colsample_bytree	트리에서 feature 샘플링 비율	0.5

N round	1000

F1 score

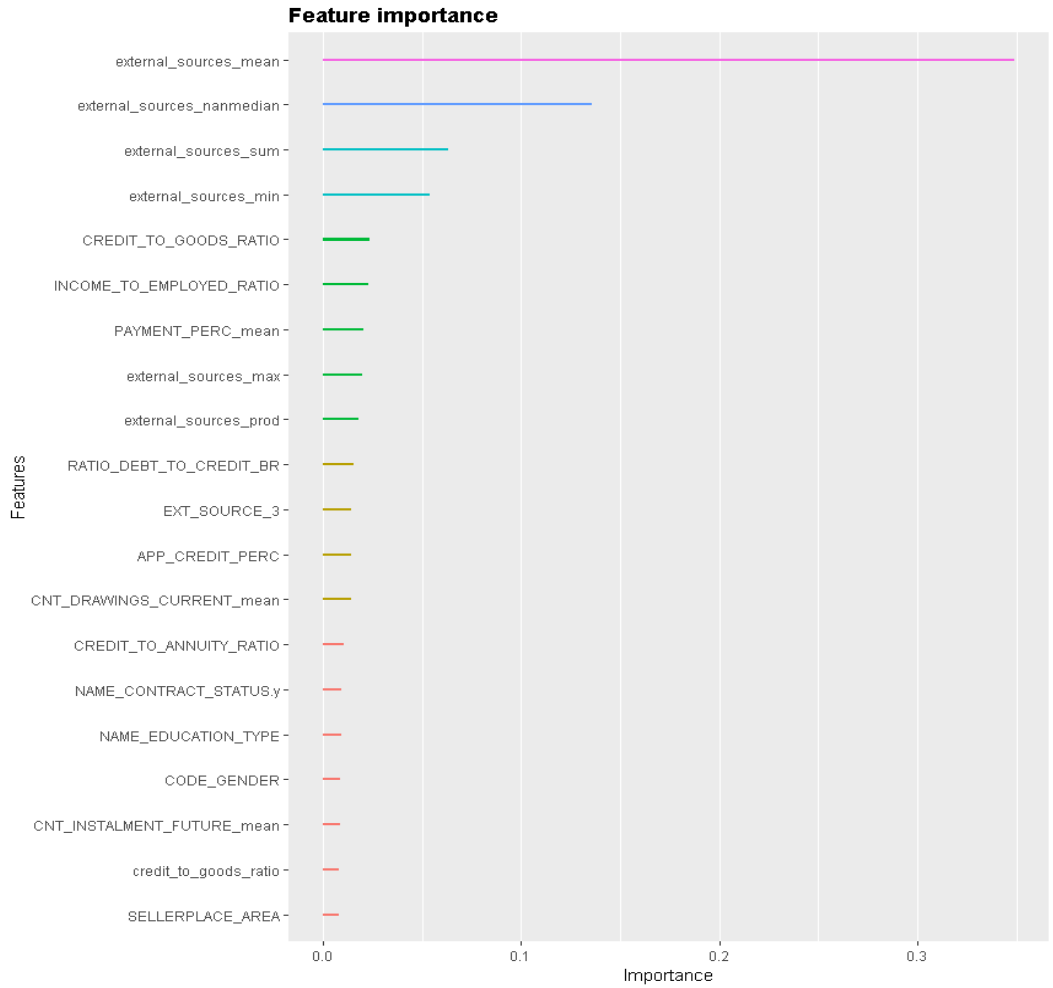
CV에 대한 F1 스코어

0.714

모델링 평가

XG Boost

Feature Importance



External_sources_mean

external source의 평균

External_sources_nammedian

external source의 중앙값

External_sources_sum

external source의 합

모델링 평가

LGBM

Parameter tuning

Bayesian optimization

Max_depth	트리의 최대 깊이	22
Num_leaves	노드 분할 시 필요한 instance	38
Min_child_sample	노드 분할 시 필요한 최소 instance	9
Min_child_weight	노드 분할 시 최소 instance weight	5
Subsample	트리에서 obs 샘플링 비율	0.39

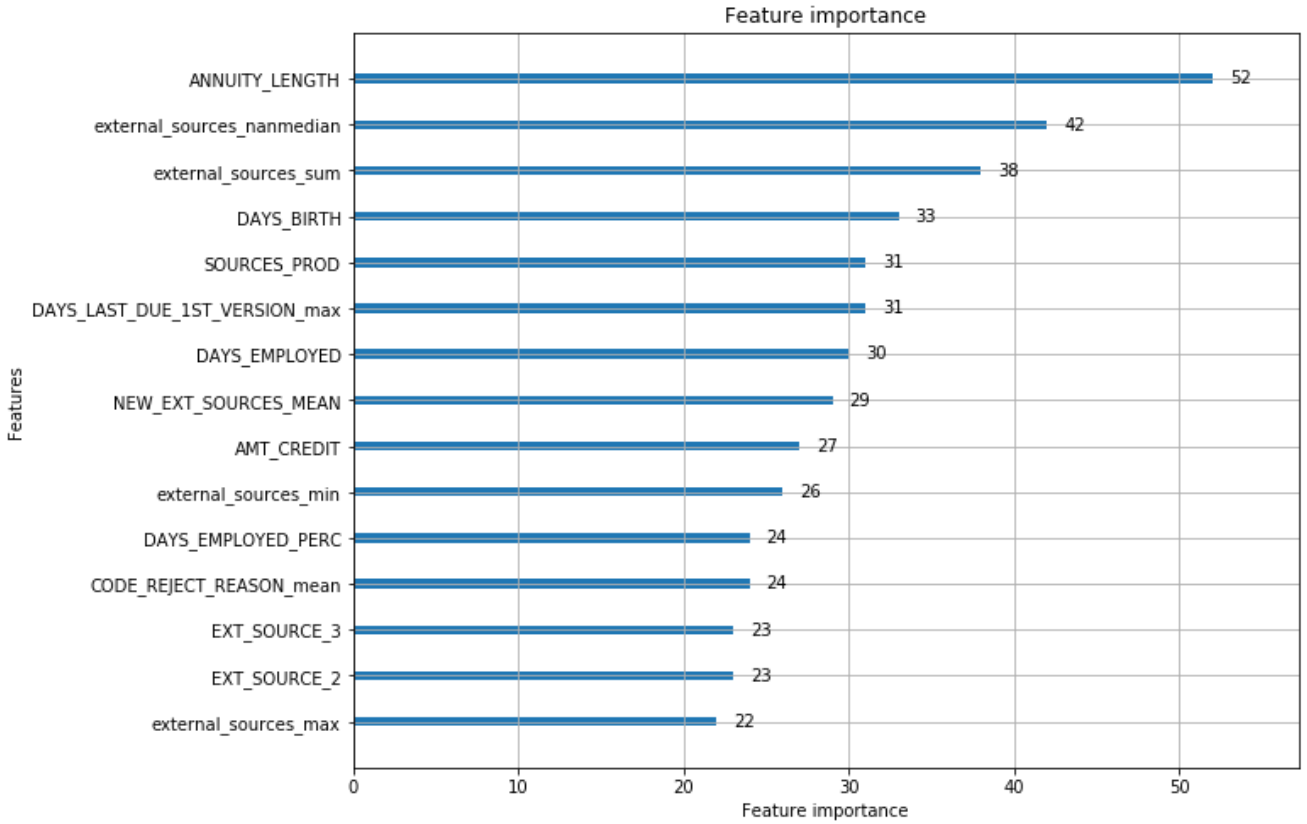
F1 score

CV에 대한 F1 스코어

0.721

모델링 평가

LGBM Feature Importance



annuity_length

연금을 받는 기간

external_sources_nammedian

external source의 중앙값

external_sources_sum

external source의 합

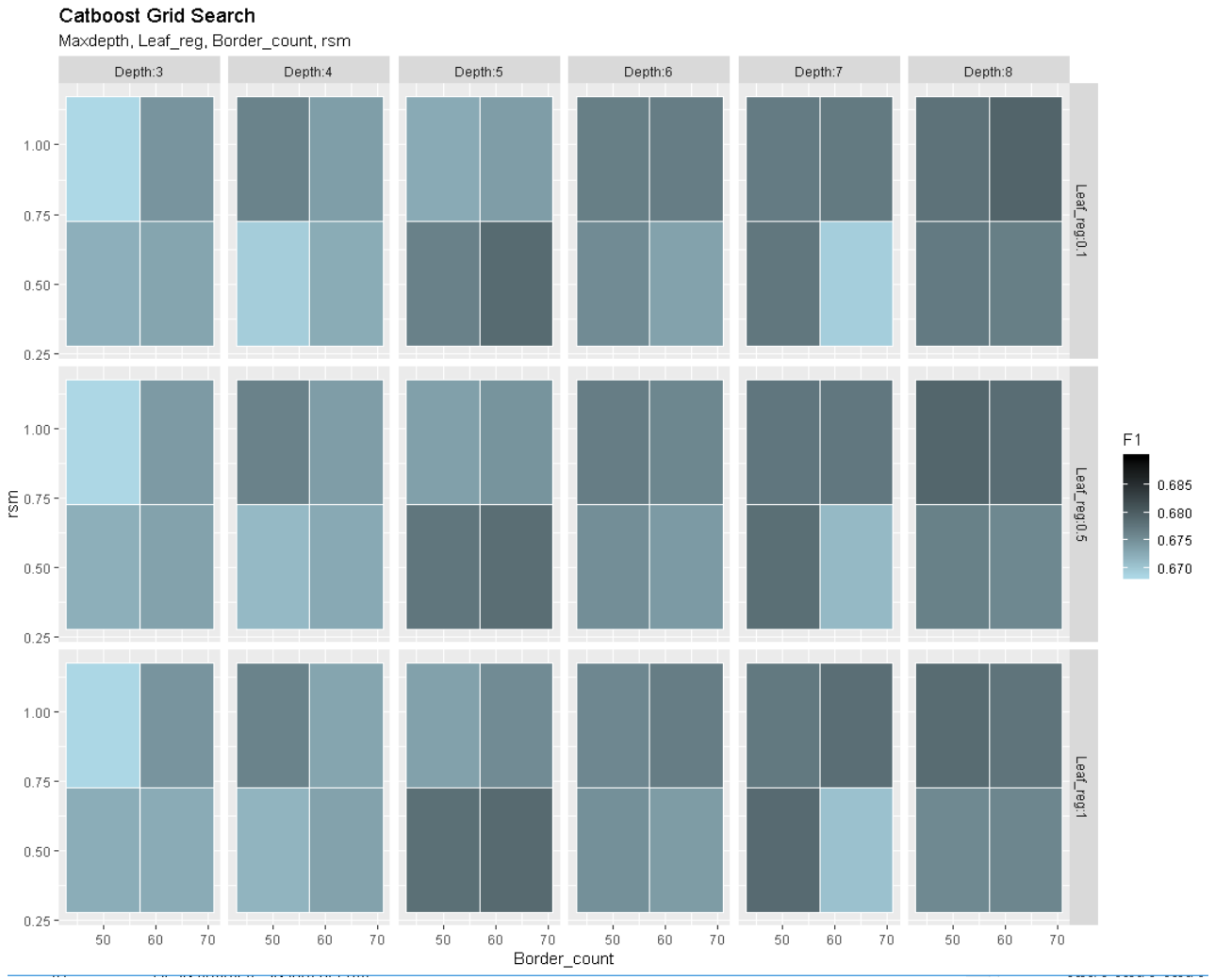
days_birth

나이

모델링 평가

CAT Boost

Parameter tuning



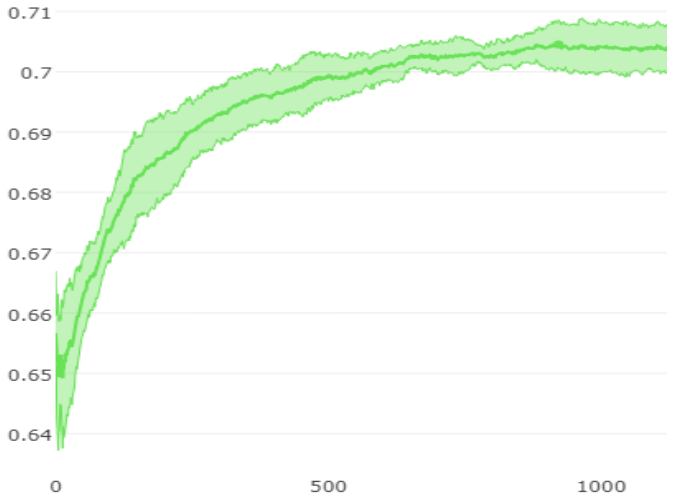
Grid Search

가장 높은 F1값을 가지는
max_depth, leaf_reg
Border_count, rsm 조합

모델링 평가

CAT Boost

depth	트리의 최대 깊이	8
L2_leaf_leg	L2 정규화에 대한 계수	1
rsm	트리에서 feature 샘플링 비율	0.5
border_count	feature combination 분리 개수	5



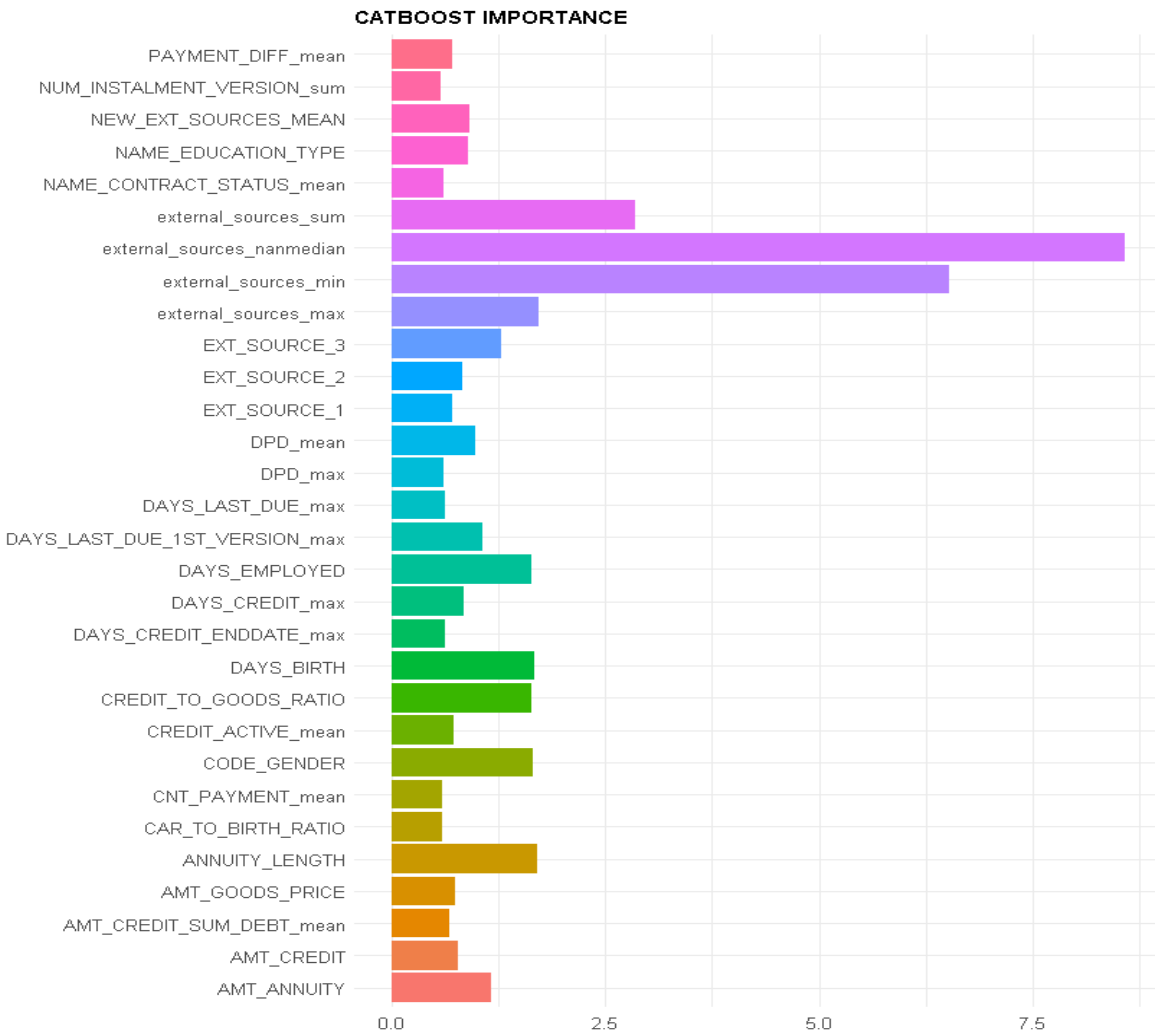
F1 score

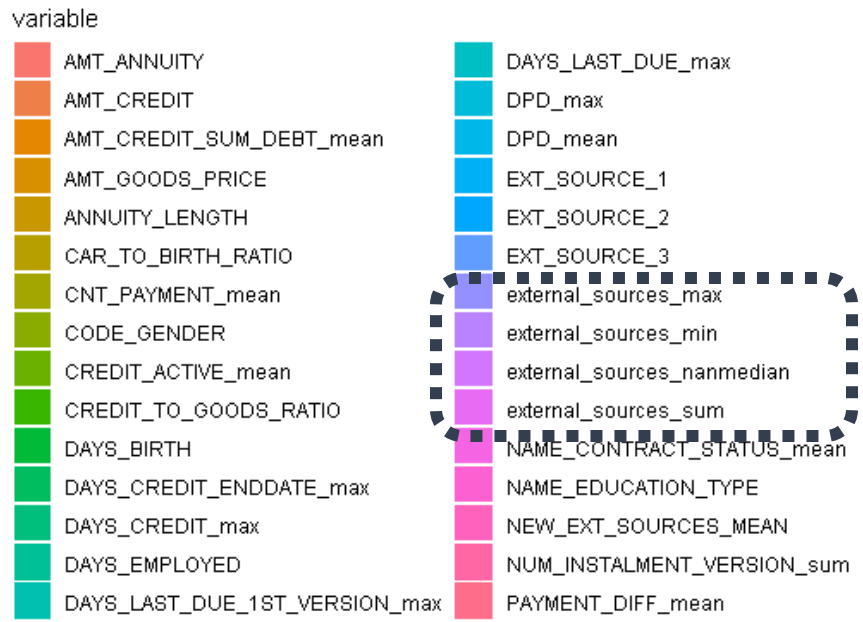
CV에 대한 F1 스코어

0.705

모델링 평가

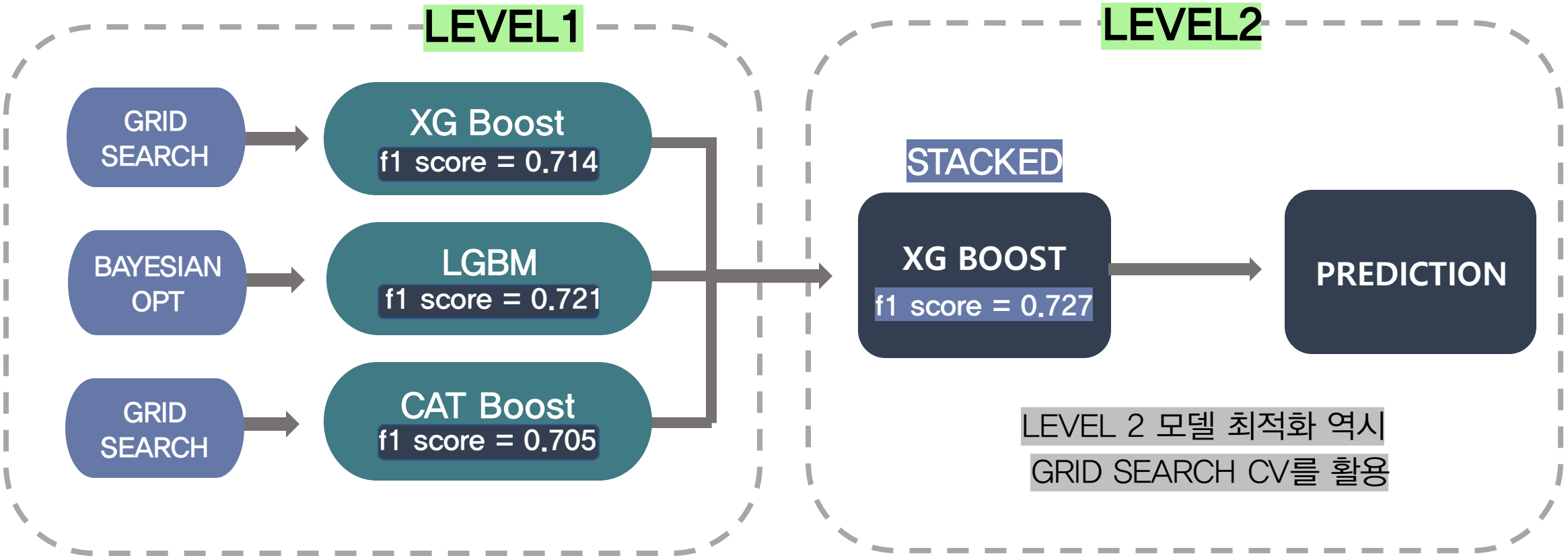
CAT Boost





모델링 평가

STACKING



모델링 평가

최종 모델

Model	단일모델			스태킹
	XG Boost	LGBM	CAT boost	XG Boost
Parameters used	max_depth :6 min_child_weight : 20 Subsample :0.5 Colsamplebytree:0.5 Nround=1000	max_depth : 22 num_leaves : 38 min_child_sample :9 feature_fraction : 5 bagging_fraction : 0.39	depth : 8 l2-leaf-reg :1 rsm : 0.5 Border_count :5	Vecstack package max_depth :5 min_child_weight : 18 Subsample :0.7 Colsamplebytree:0.5 Nround=462
F1 score (cross validation)	0.714	0.721	0.705	0.727

최종적으로 F1 스코어가 가장 높게 나온, ‘스태킹 모델’을 선정

모델링 평가

최종 모델 의의 및 한계점

의 의

1. unbalanced 데이터 예측을 샘플링을 통해서 해결
2. 다양한 파라미터 튜닝 시행(그리드서치, 베이지안)
3. 스태킹을 통해 모델들간의 앙상블 시도

한 계

예측 위주의 전처리로 인해 해석의 한계가 존재

컴퓨팅 성능의 한계로 인해 보다 로지컬한 샘플링을 하지 못함

Q&A