Dacon 15회 원자력발전소 상태 판단 모델링 경진대회

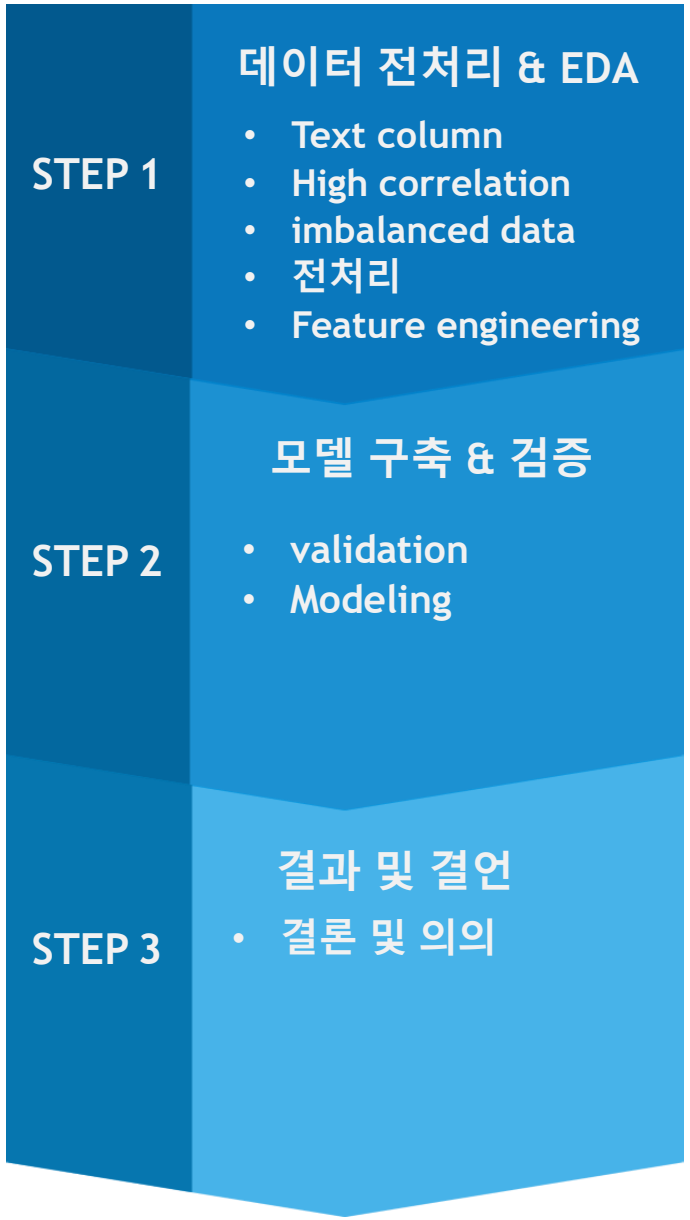
팀 : 생물학적 수처리

목차

1 EDA 및 전처리

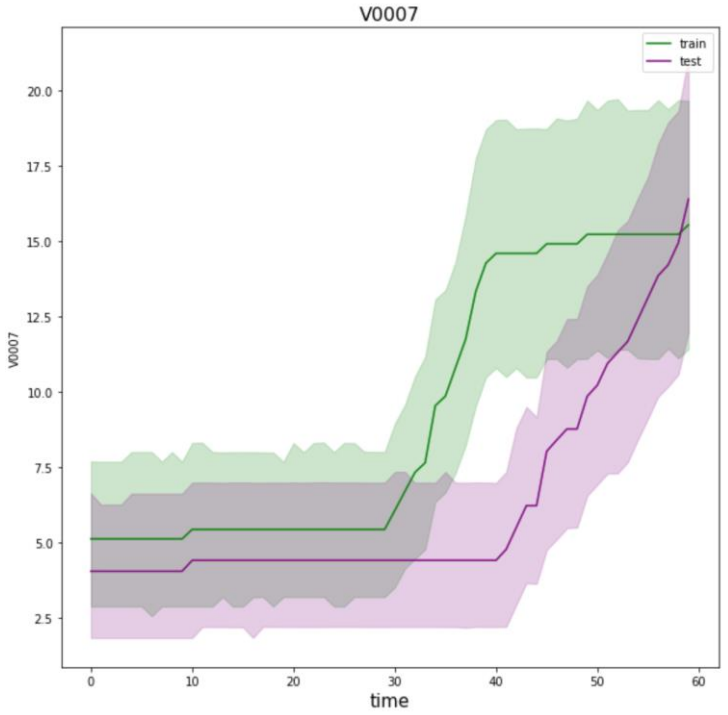
2 모델 구축 및 검증

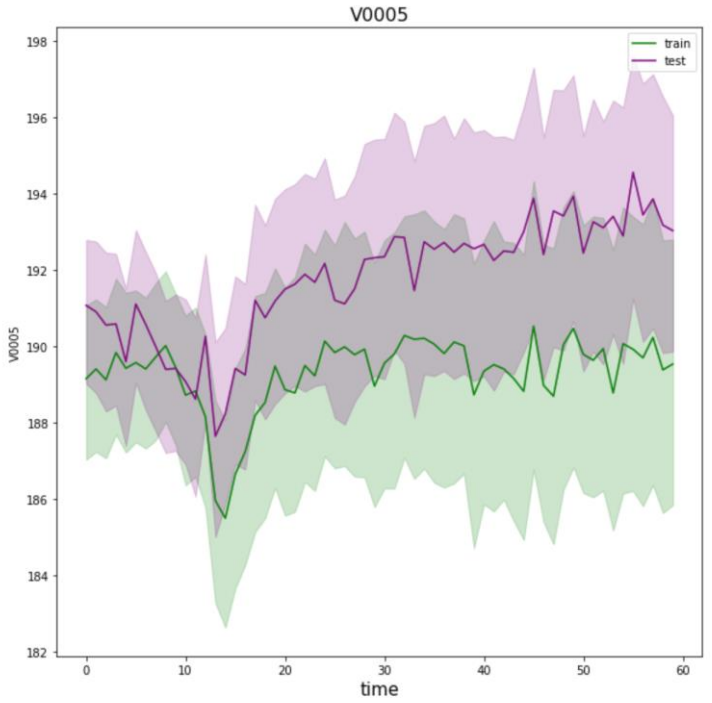
3 결과 및 결언



1. EDA & 데이터 전처리

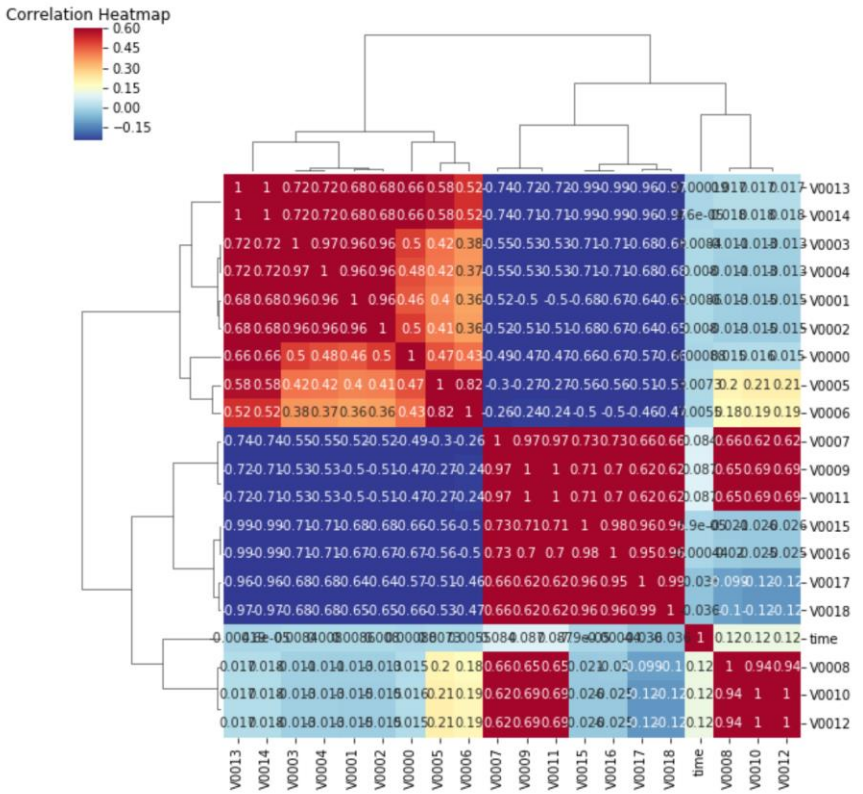
1) 각 column들은 시간이 지남에 따라 값이 계속 변한다.
따라서 Test 데이터가 60초까지 있으므로, Train에서도 60초까지만 사용한다.





1. EDA & 데이터 전처리

2) Column끼리 correlation이 높은 케이스가 많다.
또한 완전 동일한 값을 갖는 column 들도 있다.



1. EDA & 데이터 전처리

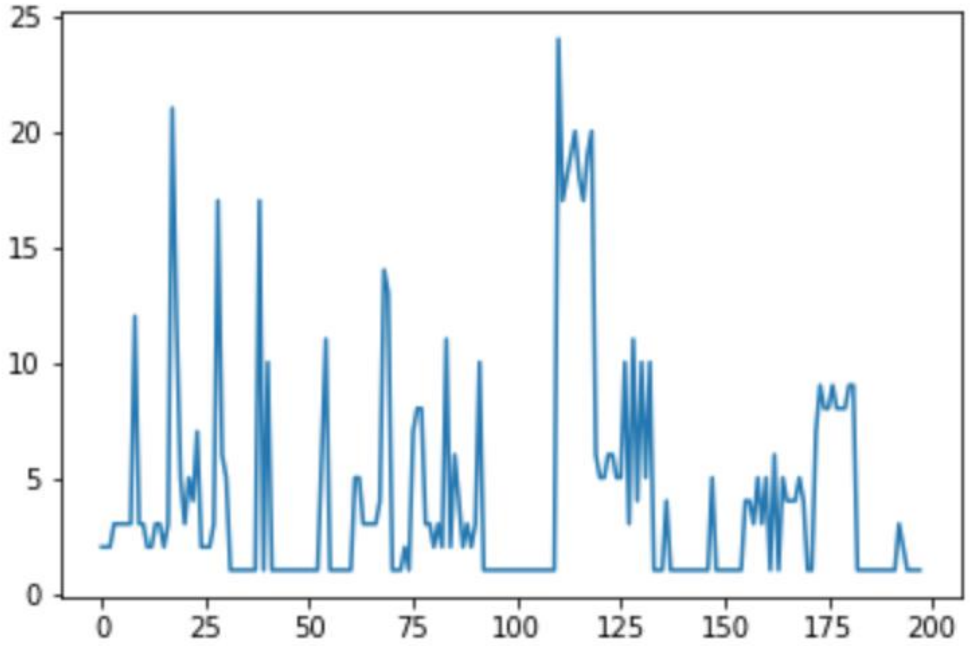
3) 실제 데이터는 일부 column에서 text 값을 갖는다.

	V3989	V3990	V3991	V3992	V3993	V3994	V3995	V3996	V3997	V3998	...	V4189	V4190	V4191	V4192	V4193	V4194	V4195	V4196	V4197	V4198
13440	OFF	ON	ON	ON	ON	ON	OFF	OFF	OFF	OFF	...	63.00769	63.084593	OFF	NaN	OFF	ON	ON	ON	ON	ON
13441	OFF	ON	ON	ON	ON	ON	OFF	OFF	OFF	OFF	...	63.00769	63.084593	OFF	NaN	OFF	ON	ON	ON	ON	ON
13442	OFF	ON	ON	ON	ON	ON	OFF	OFF	OFF	OFF	...	63.00769	63.084593	OFF	NaN	OFF	ON	ON	ON	ON	ON
13443	OFF	ON	ON	ON	ON	ON	OFF	OFF	OFF	OFF	...	63.00769	63.084593	OFF	NaN	OFF	ON	ON	ON	ON	ON
13444	OFF	ON	ON	ON	ON	ON	OFF	OFF	OFF	OFF	...	63.00769	63.084593	OFF	NaN	OFF	ON	ON	ON	ON	ON
13445	OFF	ON	ON	ON	ON	ON	OFF	OFF	OFF	OFF	...	63.00769	63.084593	OFF	NaN	OFF	ON	ON	ON	ON	ON
13446	OFF	ON	ON	ON	ON	ON	OFF	OFF	OFF	OFF	...	63.00769	63.084593	OFF	NaN	OFF	ON	ON	ON	ON	ON
13447	OFF	ON	ON	ON	ON	ON	OFF	OFF	OFF	OFF	...	63.00769	63.084593	OFF	NaN	OFF	ON	ON	ON	ON	ON
13448	OFF	ON	ON	ON	ON	ON	OFF	OFF	OFF	OFF	...	63.00769	63.084593	OFF	NaN	OFF	ON	ON	ON	ON	ON
13449	OFF	ON	ON	ON	ON	ON	OFF	OFF	OFF	OFF	...	63.00769	63.084593	OFF	NaN	OFF	ON	ON	ON	ON	ON
13450	OFF	ON	ON	ON	ON	ON	OFF	OFF	OFF	OFF	...	63.00769	63.084593	OFF	NaN	OFF	ON	ON	ON	ON	ON
13451	OFF	ON	ON	ON	ON	ON	OFF	OFF	OFF	OFF	...	63.00769	63.084593	OFF	NaN	OFF	ON	ON	ON	ON	ON
13452	OFF	ON	ON	ON	ON	ON	OFF	OFF	OFF	OFF	...	63.00769	63.084593	OFF	NaN	OFF	ON	ON	ON	ON	ON
13453	OFF	ON	ON	ON	ON	ON	OFF	OFF	OFF	OFF	...	63.00769	63.084593	OFF	NaN	OFF	ON	ON	ON	ON	ON
13454	OFF	ON	ON	ON	ON	ON	OFF	OFF	OFF	OFF	...	63.00769	63.084593	OFF	NaN	OFF	ON	ON	ON	ON	ON
13455	OFF	ON	ON	ON	ON	ON	OFF	OFF	OFF	OFF	...	63.00769	63.084593	OFF	NaN	OFF	ON	ON	ON	ON	ON
13456	OFF	ON	ON	ON	ON	ON	OFF	OFF	OFF	OFF	...	63.00769	63.084593	OFF	NaN	OFF	ON	ON	ON	ON	ON
13457	OFF	ON	ON	ON	ON	ON	OFF	OFF	OFF	OFF	...	63.00769	63.084593	OFF	NaN	OFF	ON	ON	ON	ON	ON
13458	OFF	ON	ON	ON	ON	ON	OFF	OFF	OFF	OFF	...	63.00769	63.084593	OFF	NaN	OFF	ON	ON	ON	ON	ON
13459	OFF	ON	ON	ON	ON	ON	OFF	OFF	OFF	OFF	...	63.00769	63.084593	OFF	NaN	OFF	ON	ON	ON	ON	ON

1. EDA & 데이터 전처리

4) Label이 unbalanced하다.
따라서 Train & Valid를 나눌 때 특별하게 split 해야 한다.

Nunique id



label

1 EDA & 데이터 전처리

	Hard 전처리	Soft 전처리
Text column	Text column이 존재하면 해당 column을 삭제	Text column이 존재하면 해당 row 값을 NA로 변경
Duplicated column	모든 row에서 같은 value를 가지면 동일한 column이라고 간주, 하나의 column만 남김	모든 row에서 같은 value를 가지면 동일한 column이라고 간주, 하나의 column만 남김
Constant column	모든 row에서 하나의 같은 값을 가지면 해당 column 삭제	모든 row에서 하나의 같은 값을 가지면 해당 column 삭제

- Train 데이터와 test 데이터를 합쳐서 전처리 진행

1. EDA & 데이터 전처리

FE(Feature Engineering) : hard 전처리

1. 변수들의 type 유추

- 모든 변수가 익명화 되어 있어, FE에 어려움을 겪음.

- 다양한 FE를 위해 변수들의 type을 최대한 유추해보았음
    1. Class label 수가 2개 : binary 변수
    2. Class label 수가 3개 이상 10개 이하 : categorical 변수
    3. 그 외 : numeric 변수

1. EDA & 데이터 전처리

FE(Feature Engineering) : hard 전처리

2. 변수들의 type에 기반한 FE

- categorical 변수를 frequency encoding
- Numeric 변수를 frequency encoding

Step1. Numeric 변수를 소수 둘째 자리에서 반올림

why? 유사한 값을 군집화 => 해당 군집은 비슷한 상태를 가질 것이라 생각.

Step2. Time 변수와 반올림 numeric 변수를 concat

why? 특정 시간대에 특정 군집이 나타나는 것이 상태와 관련이 있다고 생각

Step3. step2의 concat 변수를 frequency encoding

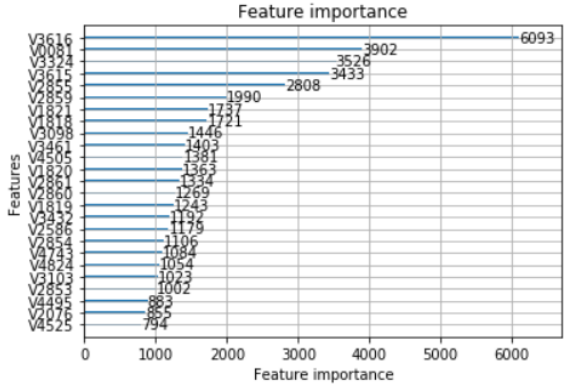
why? 특정 시간대에 특정 군집의 value count 정보를 모델에 제공

1. EDA & 데이터 전처리

FE(Feature Engineering) : hard 전처리

3. Interaction feature 생성

- FE 수행 이전 초기 모델링에서 Feature importance 추출
- Top 5 변수로 feature 생성
    1. 덧셈, 뺄셈, 곱셈, 나눗셈



4. 모든 변수에 대해 rolling mean 수행

- Id별로 rolling mean / window=5 수행

Why? EDA에서 데이터 값의 흔들림이 크다는 것을 인지하여, smoothing 하였음.

1. EDA & 데이터 전처리

FE(Feature Engineering) : soft 전처리

- Hard 전처리와 다르게, 별도의 FE를 시행하지 않음

2. 모델 구축 및 검증

Validation strategy : train / validation split

1. Train / test split 방식

- Train set과 test set의 id가 겹치지 않음

  → id based split 방식

2. Train label imbalance

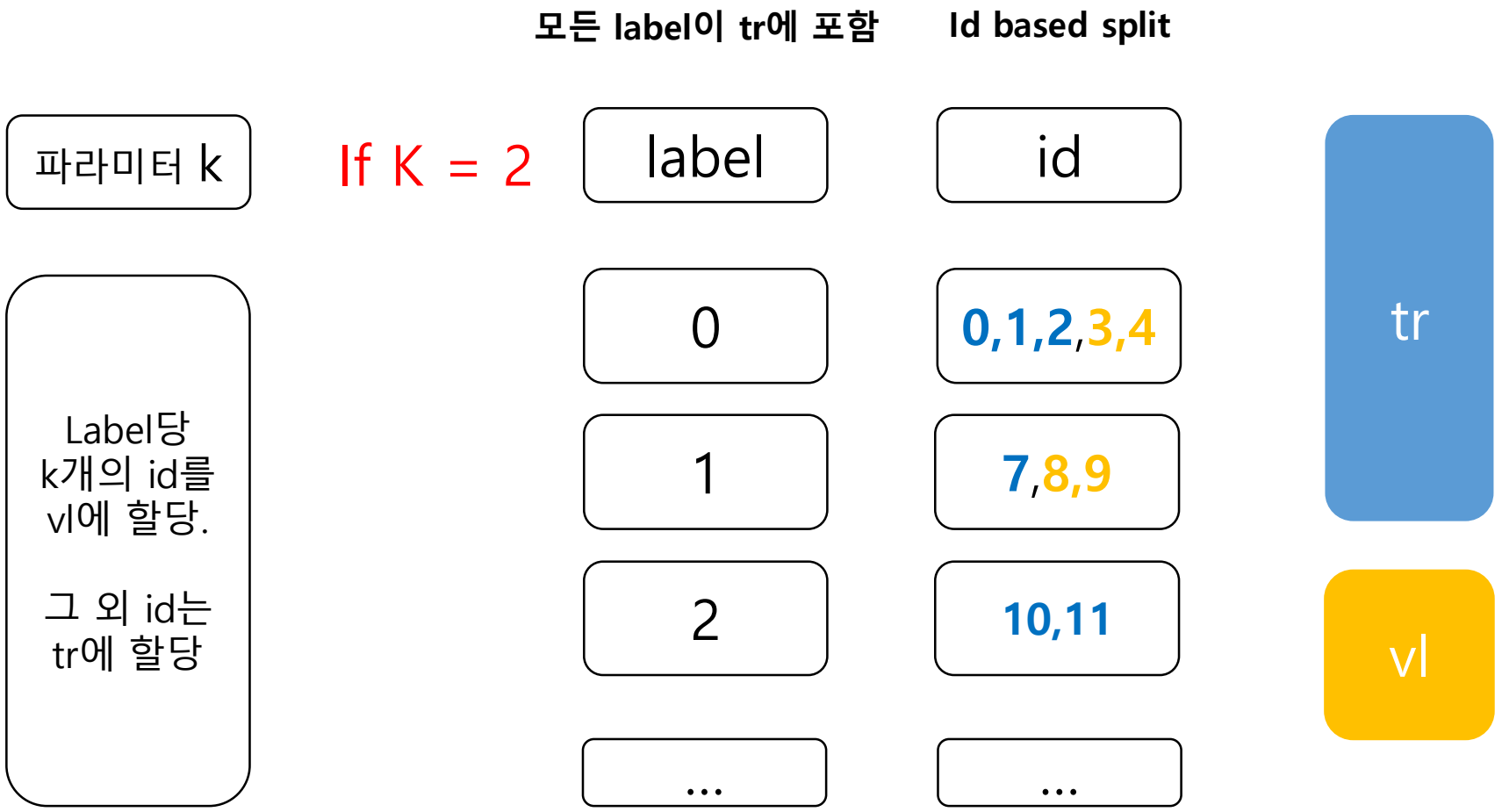
- Train label이 imbalanced하게 분포

  → 모든 label이 최소 1번은 train에 포함되도록

두 방식을 모두 고려한 validation strategy 고안

2. 모델 구축 및 검증

Validation strategy : train / validation split



2. 모델 구축 및 검증

Validation strategy : Half and Half method

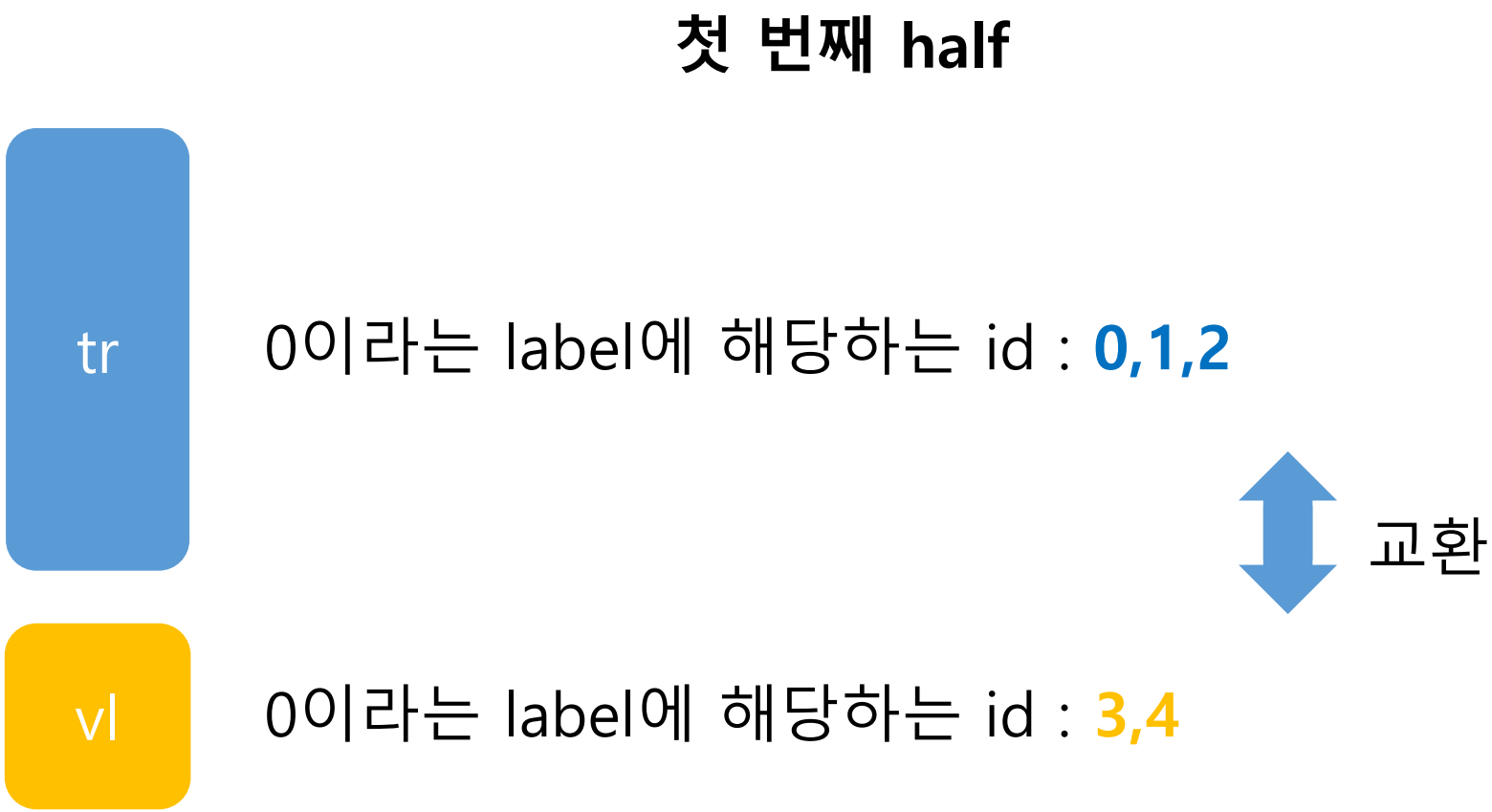
첫 번째 half

0이라는 label에 해당하는 id : **0,1,2**

0이라는 label에 해당하는 id : **3,4**

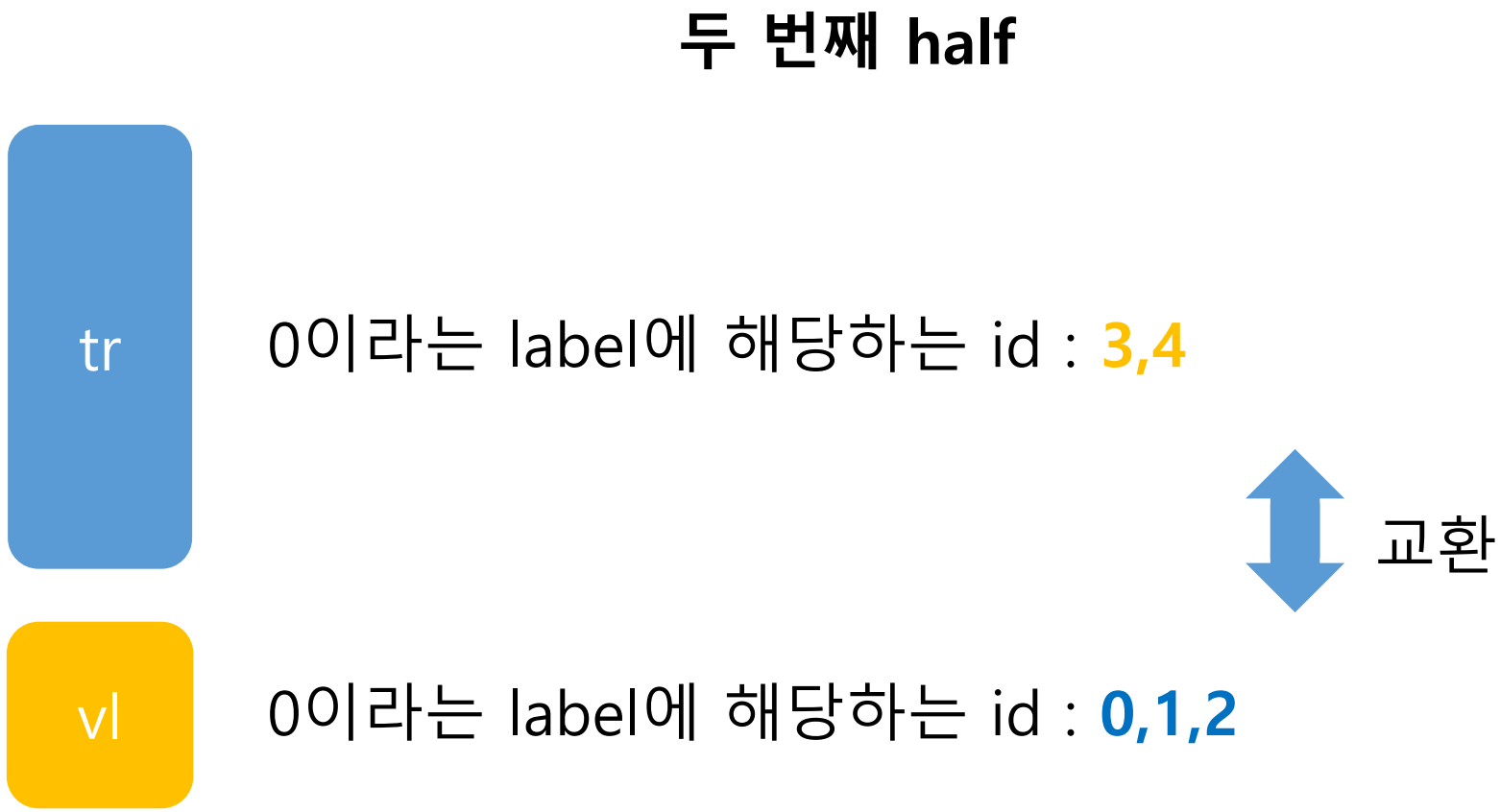
2. 모델 구축 및 검증

Validation strategy : Half and Half method



2. 모델 구축 및 검증

Validation strategy : Half and Half method



2. 모델 구축 및 검증

Validation strategy : Half and Half method

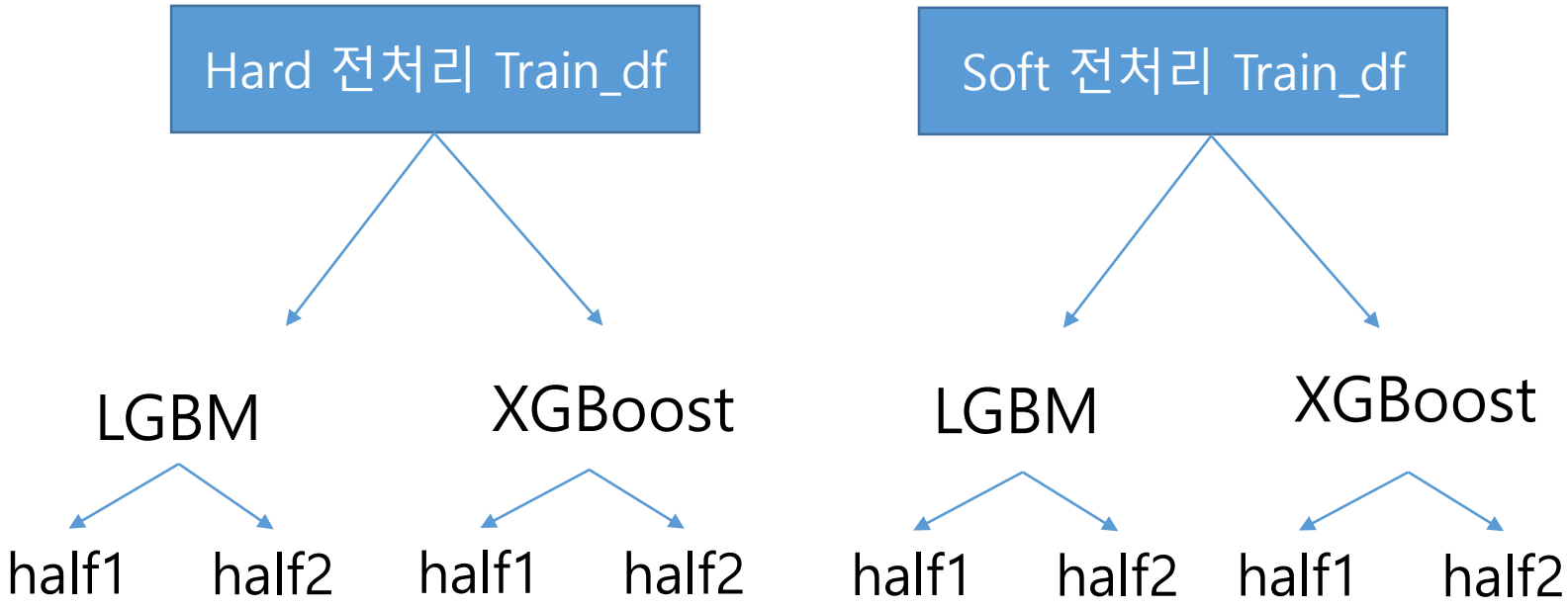
첫 번째 half prediction

두 번째 half prediction

Simple stacking
(same weight)

2. 모델 구축 및 검증

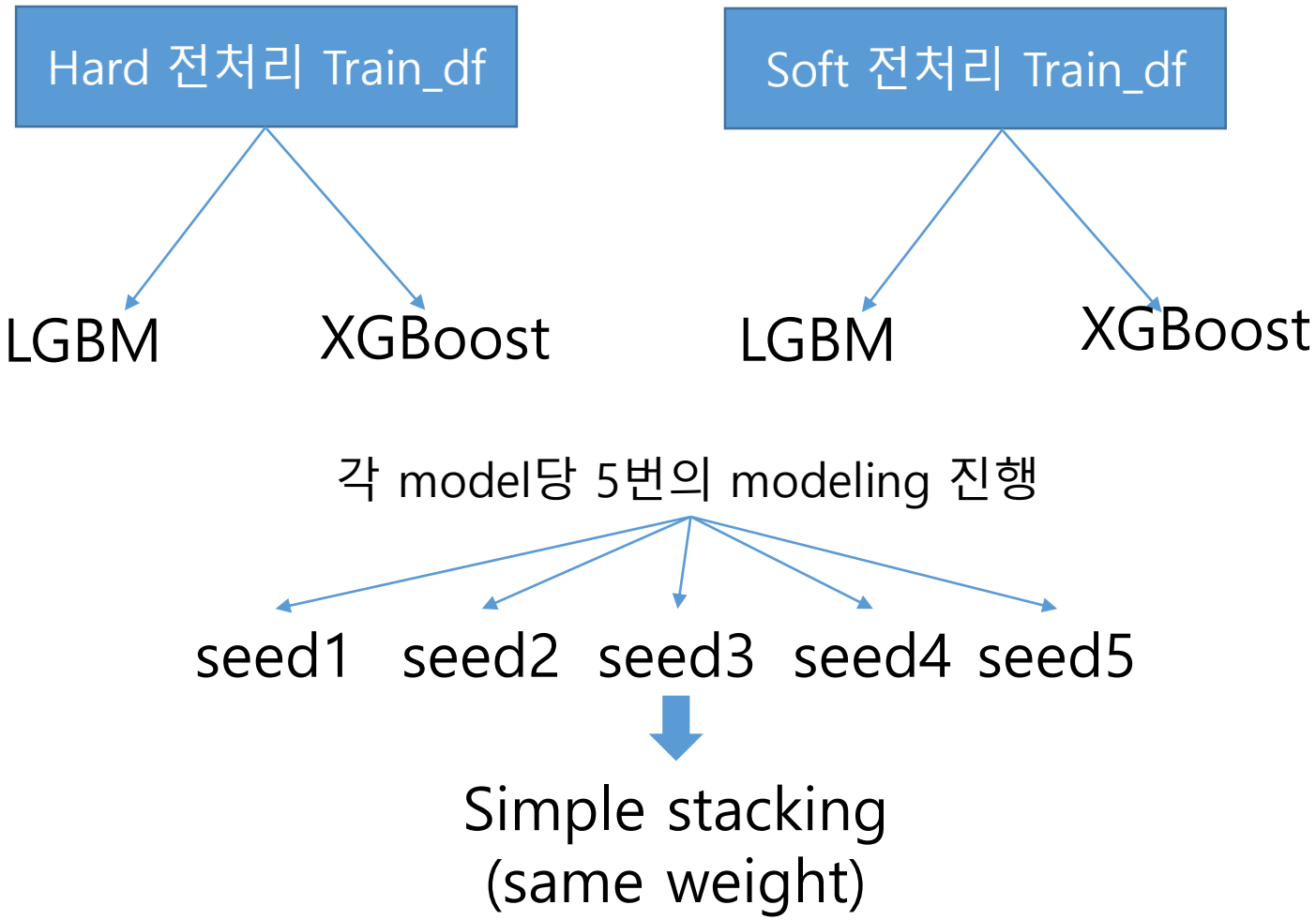
Modeling : Light GBM, XGBoost



Tree based model인 Light GBM과 XGBoost 두 개를 학습에 사용함

2. 모델 구축 및 검증

Modeling strategy : Robustness

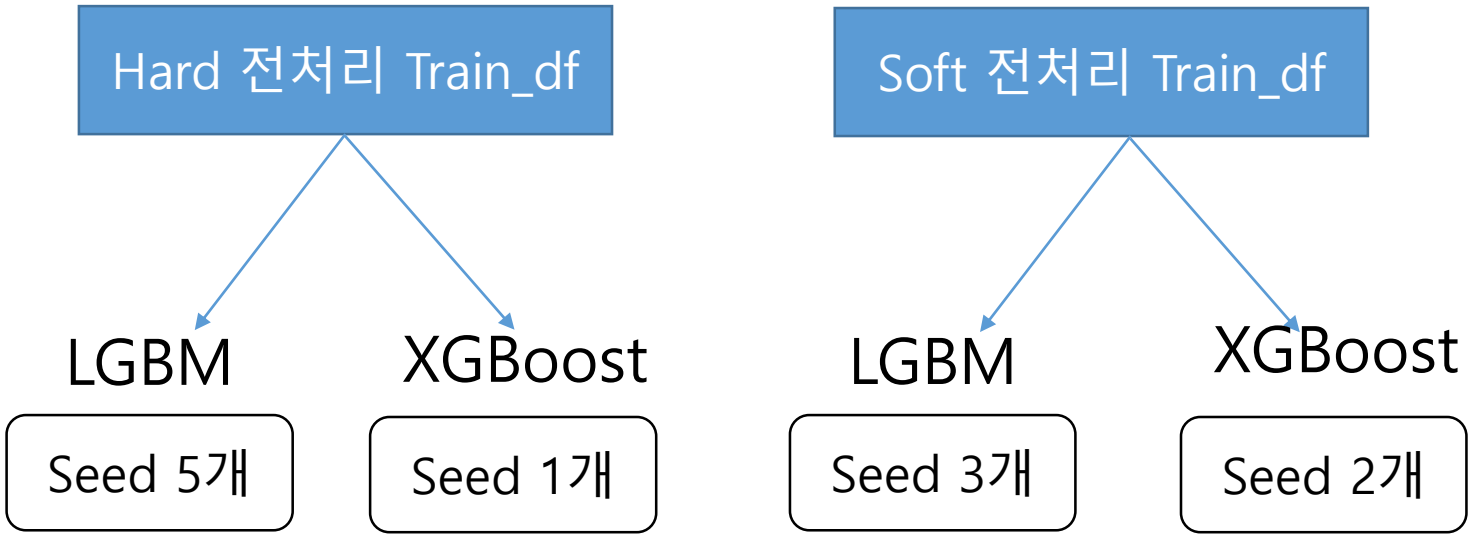


2. 모델 구축 및 검증

Modeling strategy : Robustness

아쉬운 점

- 기존 계획은 앞선 장표와 같았으나, 대회 막바지에 접어드니 시간이 부족하였습니다.
- 그래서 실제로는 이렇게 진행하였습니다.



3. 결과 및 결언

- 각기 다른 전처리 방법을 활용하여, 데이터에서 각기 다른 패턴을 찾아내고자 노력하였습니다.
- 모델의 robustness를 위해 여러 개 seed로 modeling을 진행하였습니다.
- LGBM XGBoost의 Ensemble을 통해 모델의 variance를 줄이고자 하였습니다.
- 결과적으로 public 점수와, private 점수의 차이가 아주 작았습니다.
  - Public score : 0.5476110949
  - Private score : 0.5418394542