

## 3 팀

김다민	김보근
서유진	송승현
심현구	조성우

# INDEX

---

1. EDA 및 데이터 전처리

2. 모델링

3. 결론

# 1

## EDA 및 데이터 전처리

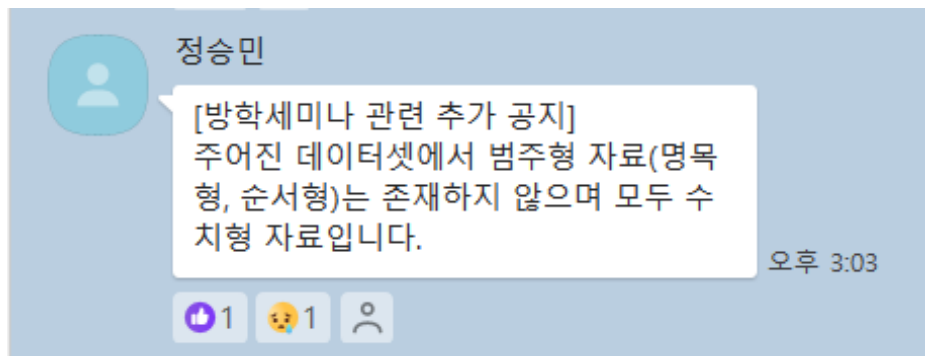
# 1

## EDA 및 데이터 전처리

### 데이터셋 구조

단호한 학회장팀...

변수명은 **확인 불가**하며, **자료형만 확인 가능합니다!**



training\_set

55000 \* 171

Int, float만으로 구성

test\_set

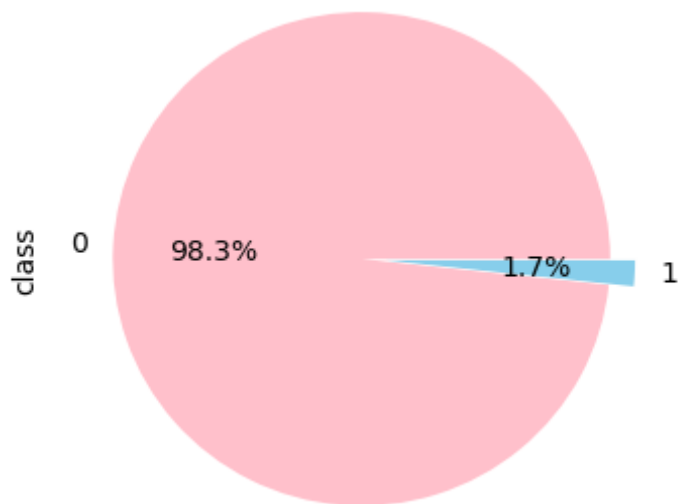
21000 \* 170

Int, float만으로 구성

# 1

## EDA 및 데이터 전처리

### 클래스 불균형



**클래스 불균형이 심각**

(0:1의 개수가 58:1의 비율로 존재)



학습 과정에서 소수 데이터에 대한 학습이  
잘 이루어지지 않을 가능성이 높음

데이터 샘플링(sampling)을 고려해볼 수 있음

## 데이터 샘플링 기법

### 오버샘플링(oversampling)

적은 클래스의 관측치를 늘리는 방법  
랜덤오버샘플링, SMOTE 등



심각한 클래스 불균형 하에서는  
**과적합 위험성** 증대

### 언더샘플링(undersampling)

많은 클래스의 관측치를 줄이는 방법  
Tomek Links 등



중요한 데이터가 탈락되어  
**정보의 소실** 발생 가능

## 클래스 불균형 해결방법

알고리즘을 통한 데이터 샘플링의 한계 확인



**비용민감학습(Cost-sensitive learning) 고려**

심혈관질환 위험 예측을 위한 비용민감 학습 모델(이유나 외, 2021)

**비용민감학습 (Cost-sensitive learning)**

모델링 학습 과정 중 **더 적은 클래스**를 잘못 예측했을 때 **큰 가중치**를 주는 방법

## 1

## EDA 및 데이터 전처리

## 샘플링과 비용민감학습 비교 (XGBoost)

SMOTE	예측값(N)	예측값(P)
실제값(N)	15268	217
실제값(P)	1	15453

비용민감학습	예측값(N)	예측값(P)
실제값(N)	15291	181
실제값(P)	13	164

Method	SMOTE	비용민감학습
validation cost	453	4220
실제 Cost	4565	3995

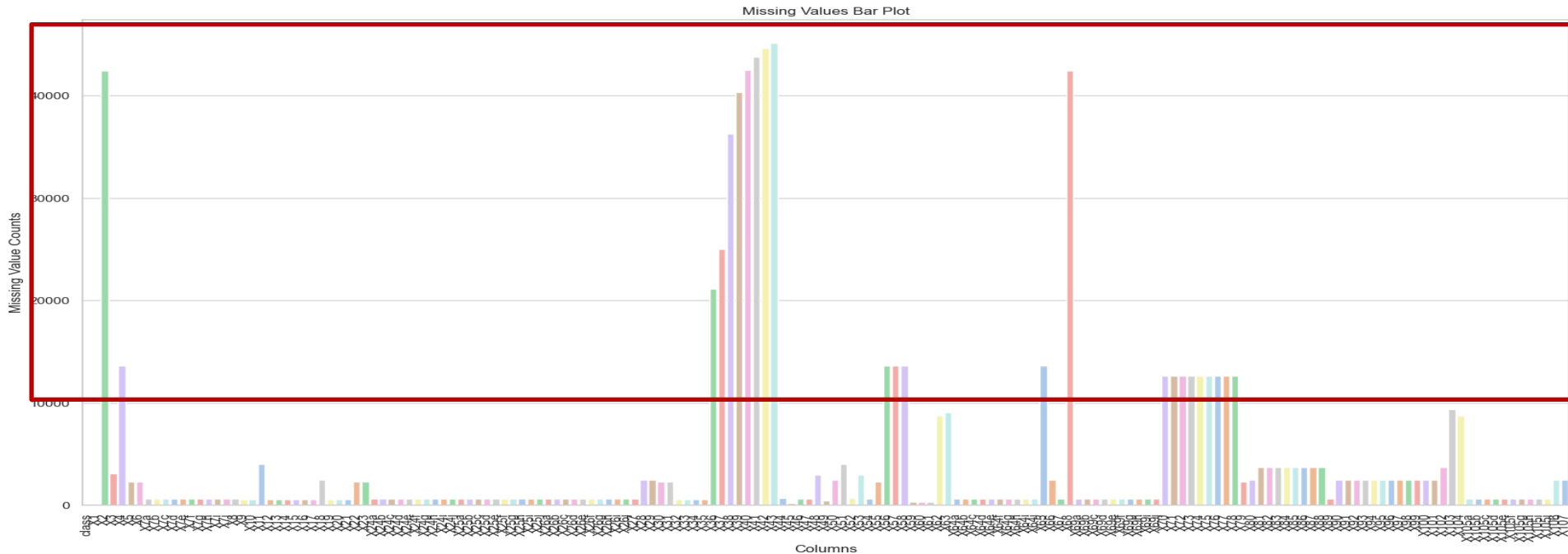
SMOTE에서는 과적합이 일어남 → 비용민감학습 채택



## 1

## EDA 및 데이터 전처리

## 열 결측치 확인



결측치 20%(11000개)이상 열 24개는 해당 **행 결측치 여부**로 대체

# 1

## EDA 및 데이터 전처리

### 열 결측치 확인



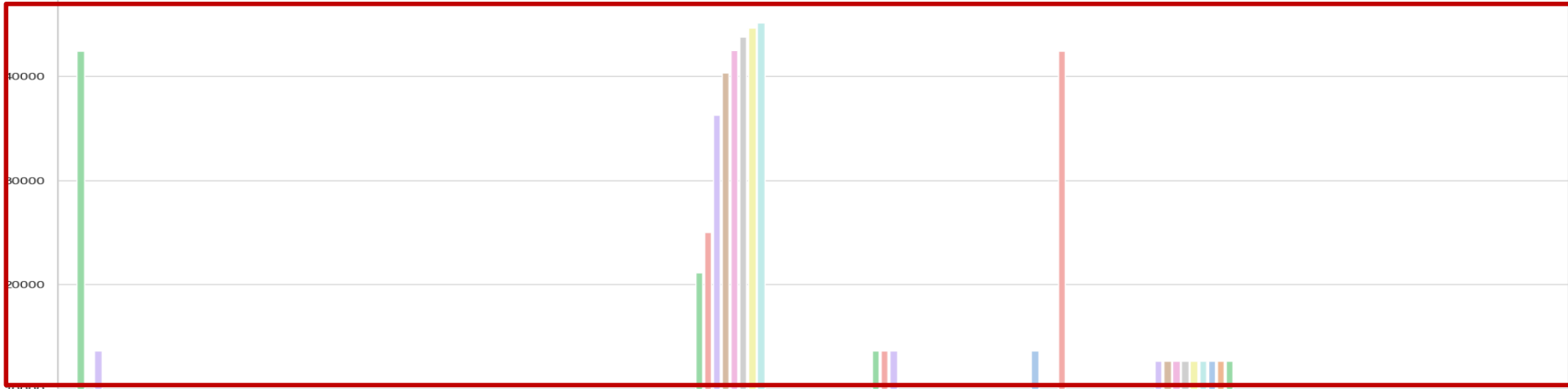
Missing Values Bar Plot

#### 열 제거 기준을 20%로 설정한 이유

실제 변수 모집단이 완전한 분포일 경우(ex. 정규분포),  
median 등으로 보간 시에 분포가 왜곡되어 많은 정보의 손실을 야기  
정규분포 가정이 있는 모델 사용에도 부정적인 영향을 미친다.

결측치 20%(11000개)이상 열 24개는 해당 행 결측치 여부로 대체

## EDA 및 데이터 전처리

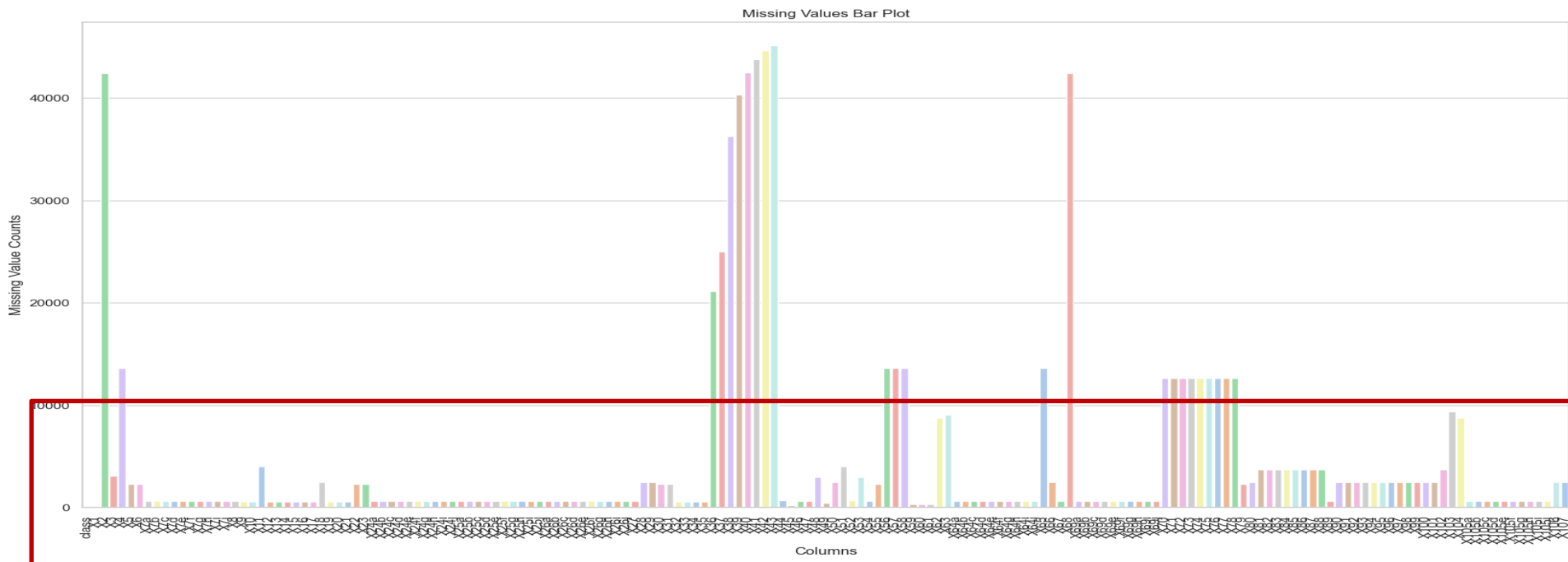


→ 'class'와 독립으로 검토된 X58 삭제

## 1

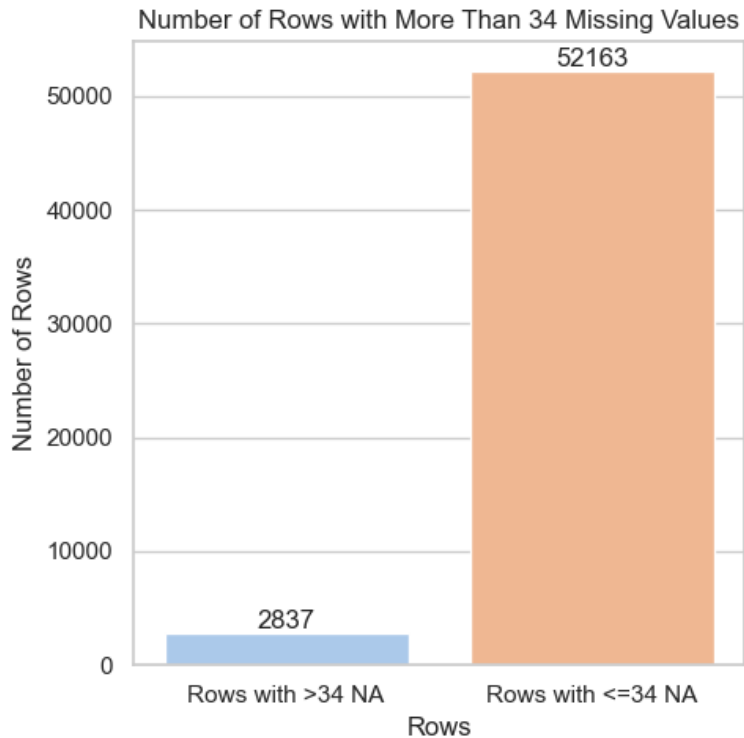
## EDA 및 데이터 전처리

## 열 결측치 확인



결측치 11000개 이하 열은 결측치를 보간하기로 결정

## 행 결측치 확인



본격적인 열 결측치 처리 전  
결측치 20%(34개) 이상인  
관측치 2837개는 정보량이  
부족하다고 판단하여 삭제

## 열 결측치 처리

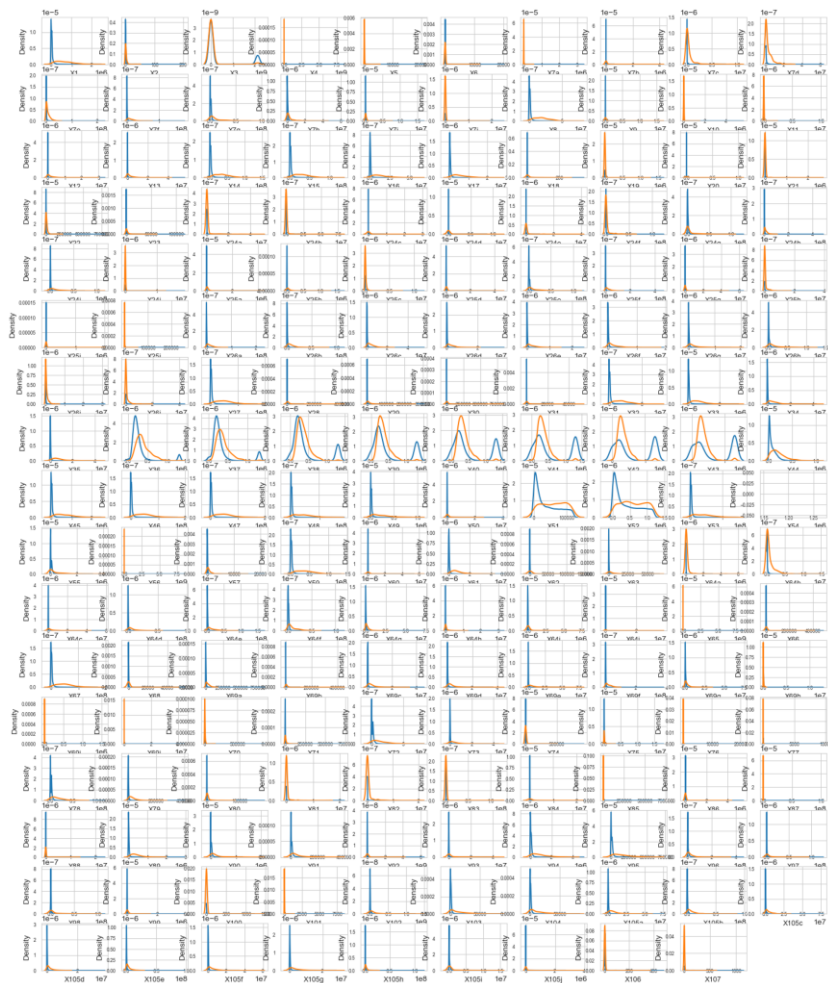
### 결측치 보간

변수별 밀도함수와 상관계수를 확인 후 비슷한 패턴을 보이는 변수들을 묶어서  
처리하거나 mode, median 등으로 대체

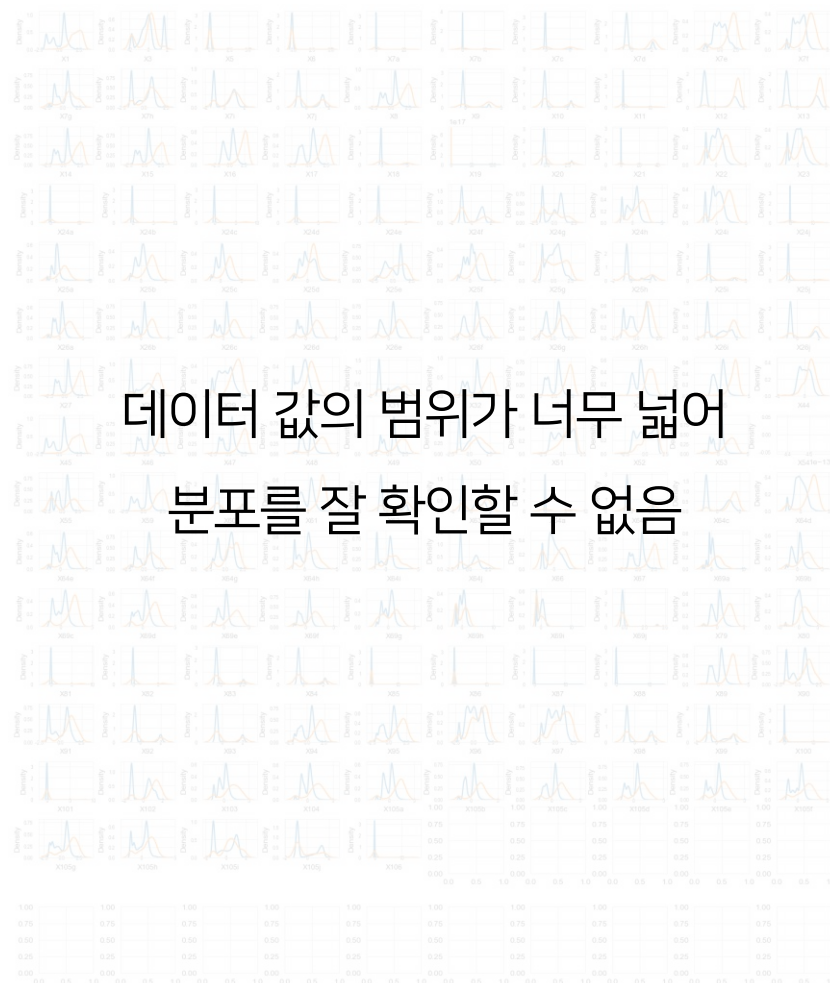
## 1

## 1

## 변수별 Distribution 확인



## Yeo-Johnson 변환

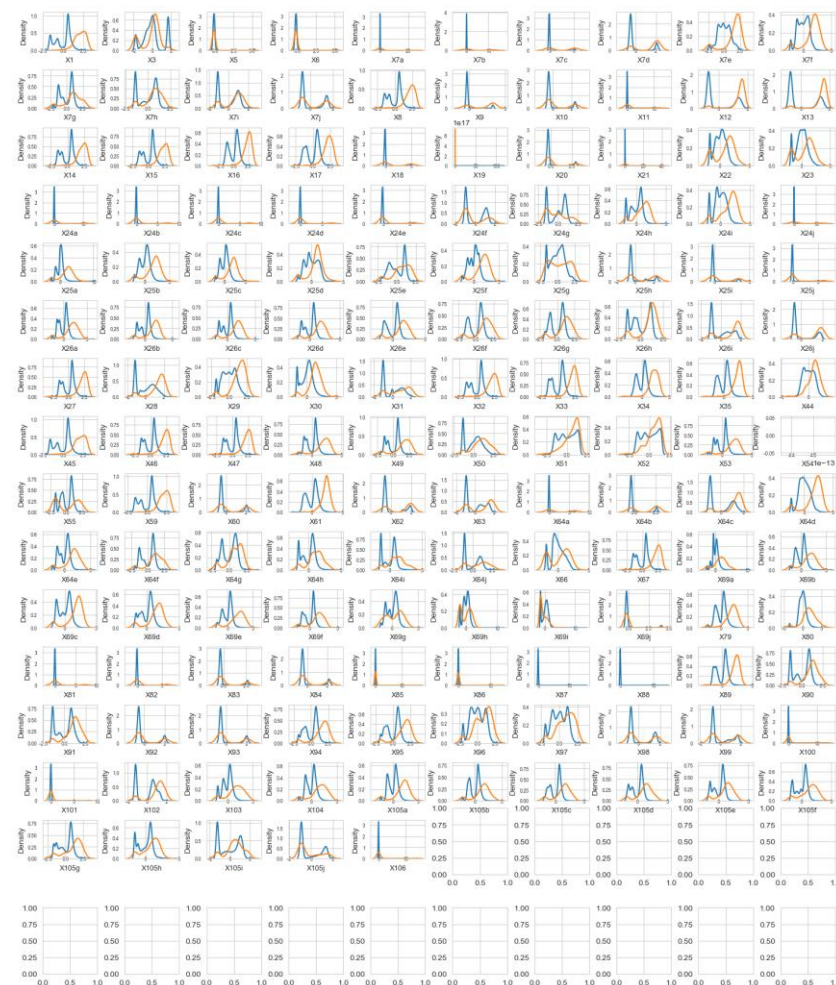


데이터 값의 범위가 너무 넓어  
분포를 잘 확인할 수 없음

## 1

## EDA 및 데이터 전처리

## 변수별 Distribution 확인





## 변수별 Distribution 확인



Yeo-Johnson 변환

Yeo-Johnson 변환

Yeo-Johnson 변환

- Box-cox 변환을 일반화한 변수변환법

Box-cox를 일반화한 변수변환법

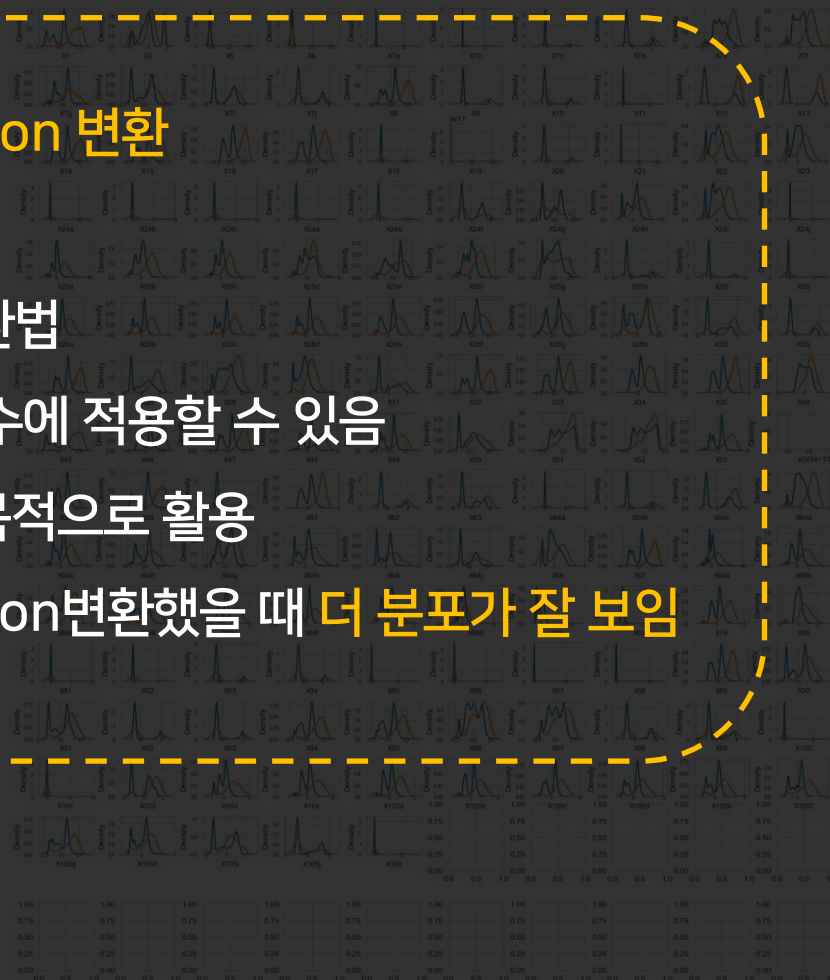
- 실수 전체 구간에서 정의된 확률변수에 적용할 수 있음

실수 전체 구간에서 정의된 확률변수에

- Yeo-Johnson 변환을 **Scaling** 목적으로 활용

적용할 수 있다.

- Log 변환 했을 때보다 Yeo-Johnson 변환했을 때 **더 분포가 잘 보임**



## 1

## EDA 및 데이터 전처리

## 변수별 Distribution 확인



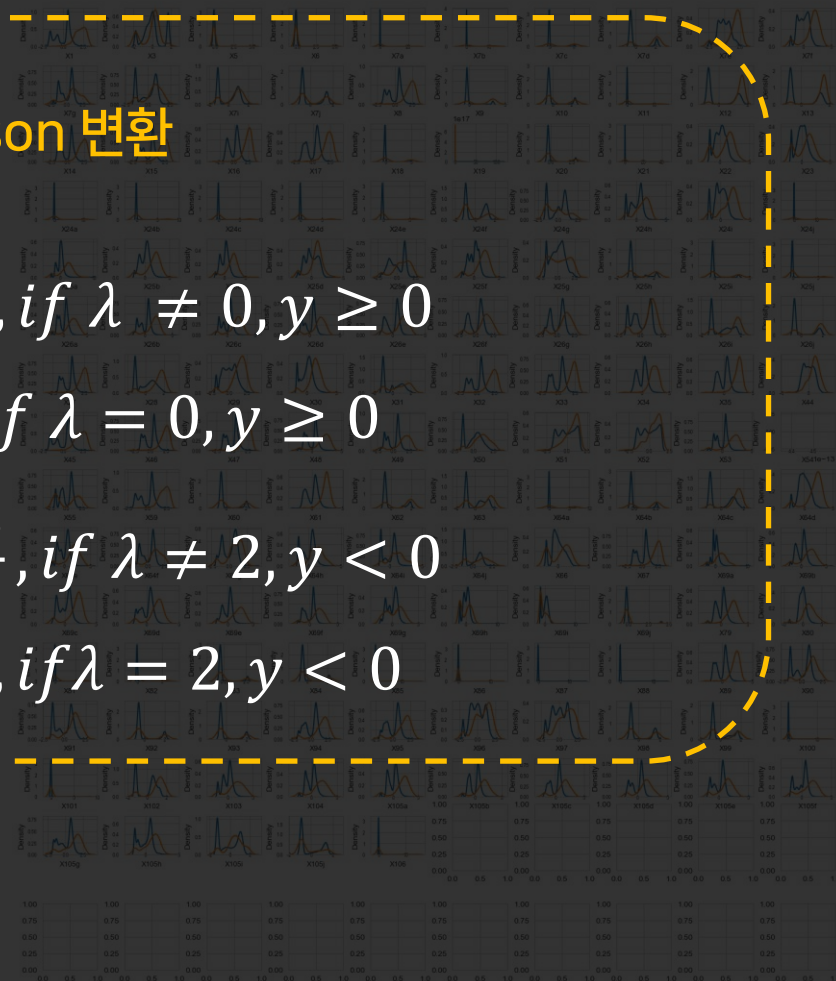
Yeo-Johnson 변환

Yeo-Johnson 변환

Yeo-Johnson 변환

Box-cox를 일반화한 변수변환법  
 실수 전체 구간에 적용 가능한 변환변수에  
 적용할 수 있다.

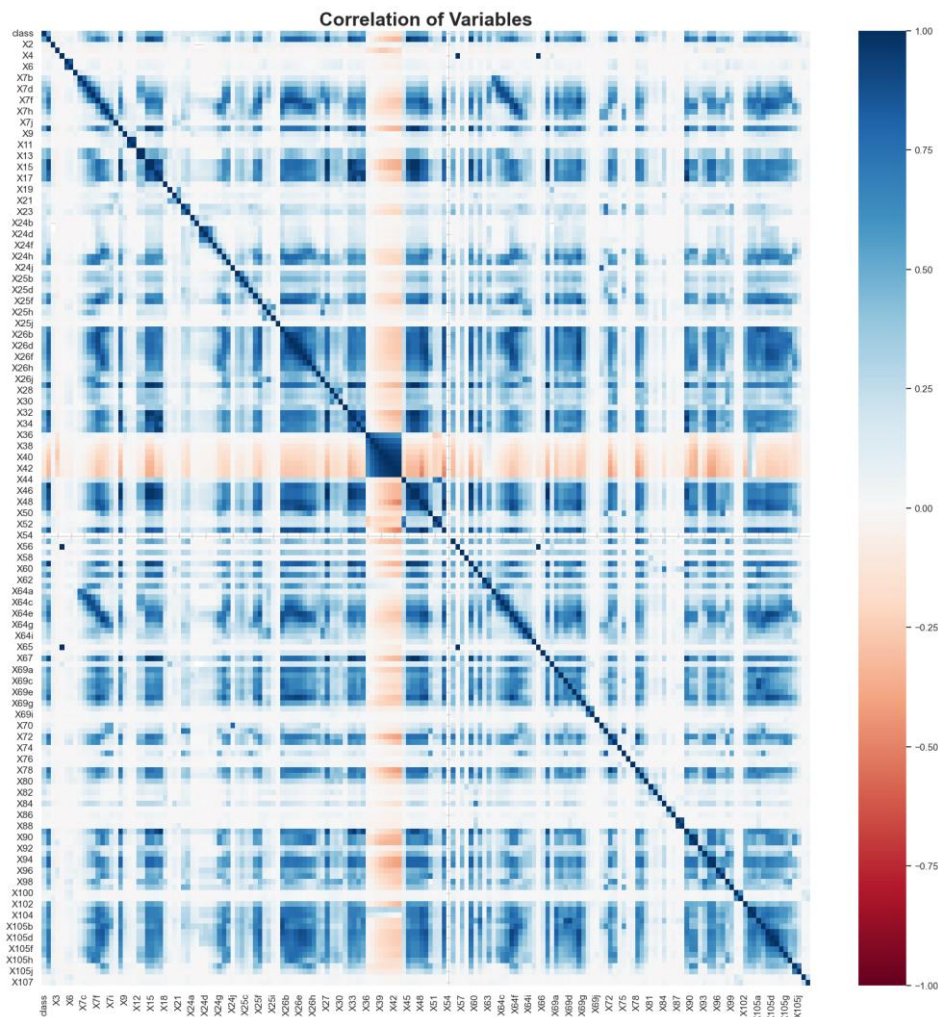
$$x(\lambda) = \begin{cases} \frac{(y+1)^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0, y \geq 0 \\ \ln(y+1), & \text{if } \lambda = 0, y \geq 0 \\ \frac{(1-y)^{2-\lambda} - 1}{\lambda - 2}, & \text{if } \lambda \neq 2, y < 0 \\ -\ln(1-y), & \text{if } \lambda = 2, y < 0 \end{cases}$$



## 1

## EDA 및 데이터 전처리

## 변수 간 상관계수 시각화



상관계수 높은 변수가 다수



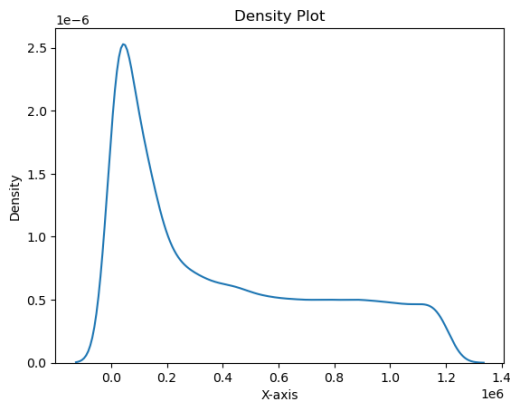
결측치 보간에 참고

## 1

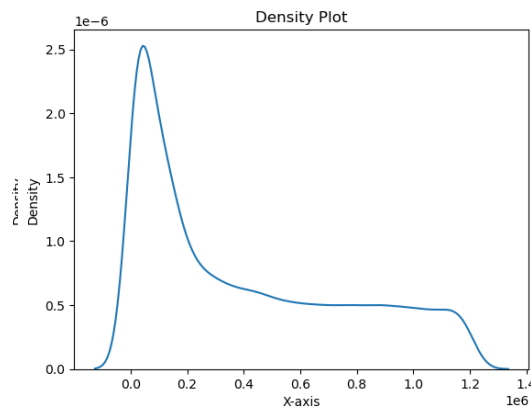
## EDA 및 데이터 전처리

## 열 결측치 처리

변수명	특징	처리
X14	X1과 상관계수 높음	X1을 기준으로 선형보간
X15		
X16		
X17		



&lt;보간 전&gt;



&lt;보간 후&gt;

결측치가 비슷한 패턴을  
보이는 변수는  
**상관계수가 높음**  
다른 변수를 기준으로 선형보간



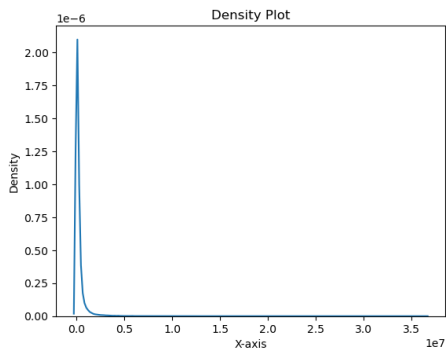
보간 전과 후의  
밀도함수가 거의 일치



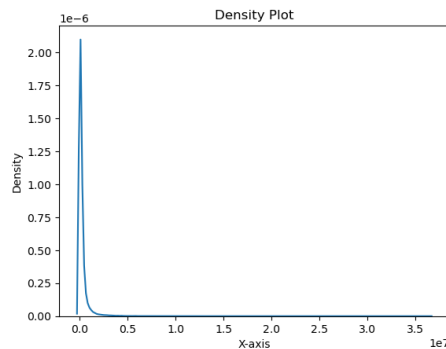
**보간이 잘 되었다고 판단**

## 열 결측치 처리

변수명	특징	처리
X28	결측 256, unique 작음(2681)	mode
X29	결측 257, unique 작음(3417)	
X30	결측 153, unique 작음(3732)	
X31	결측 151, unique 작음(1031)	



&lt;보간 전&gt;



&lt;보간 후&gt;

첨도가 높은 변수는  
**mode**로 선형보간



보간 전과 후의  
밀도함수가 거의 일치



**보간이 잘 되었다고 판단**

## 파생변수 추가

### Row\_Nas\_count

20% 이상 결측치를 보유한 obs는 삭제한 후 결측치의 개수가 의미있는 변수가 될 수 있을 것이라고 판단

### Cluster\_label

Feature importance top2 기준으로  
군집화 (Kmeans, n=2) 후  
군집별 클래스 비율 확인

군집1 0: 50934개, 1: 226개

군집2 0: 631개, 1: 372개

### PCA

기존 변수들로 PCA를 진행, 분산을  
80%만큼 설명하는 PC 3개를 선택하여  
파생변수로 사용

### Isol\_score

Isolation forest score

# 2

모델링

## 모델링 방향

다양한 모델로 성능 평가 후 최종 앙상블에 사용할 모델 top5 선정



앙상블을 통해 submission data 생성



## 비용민감학습(Cost-sensitive learning)

Cost Function

$$cost = 250 \times FN + 5 \times FP$$



가중치가 높은 **FN**(False Negative)을 줄이는 게 핵심

즉, 실제값이 1인데 예측값이 0인 경우를 줄이기 위해

Xgb, lgbm, catboost에서 **소수클래스에 weight를 부여**하며 학습을 진행

## 후보 모델

## XGBoost

트리 기반의 앙상블 모델  
뛰어난 예측 성능과 빠른 학습 시간  
과적합 방지  
결손 데이터 자체 처리

## LightGBM

트리 기반 Gradient Boosting 모형  
Leaf-wise 확장 방식으로 빠른 속도  
대용량 데이터에 적합

learning_rate	트리 학습 비율
max_depth	트리 최대 깊이
n_estimators	생성할 트리 개수
subsample	각 트리별 훈련 데이터 비율
colsample_bytree	각 트리별 피처 비율

n_estimators	결정나무의 개수
max_depth	결정나무의 최대 깊이
min_child_samples	과적합 방지 위한 파라미터
num_levels	개별 트리가 가질 수 있는 최대 리프 개수

## 후보 모델

### CatBoost

과적합 방지에 집중한 트리 기반 부스팅 모델

Level-wise 트리 확장 방식

데이터 일부만 사용하여 잔차 구하는 방식

범주형 변수 데이터에 효과적

max_depth	결정나무의 최대 깊이
l2_leaf_reg	L2 정규화의 정도
iterations	반복 횟수
border count	각 피처 공간에서 수행할 분할의 수
class_weights	클래스 불균형이 있는 경우 비율 지정

### Tabnet

트리와 신경망의 특성 결합한 딥러닝 모델

유연하고 재귀적인 특성 선택

병렬 처리를 통한 과적합 방지

n_steps	특성 선택 매커니즘 단계 설정
n_decision_layers	각 단계 선택할 특성 수
n_attention	재귀적인 특성 선택 반복 수
n_shared	공유 특성 사용 여부

## 후보 모델

## Support Vector Machine

결정경계 (decision boundary)를 정의  
 마진 최대화  
 커널 기법을 통한 비선형 문제 해결 가능

C	소프트 마진 규제
kernel	커널 함수 지정
gamma	결정 경계 곡률 조정

## Random Forest

부트스트랩 샘플링을 통한 학습 데이터 형성  
 랜덤 특성 선택  
 앙상블 방식을 통한 과적합 방지

n_estimators	결정나무의 개수
max_depth	결정나무의 최대 깊이
max_features	결정나무를 분지할 때 고려하는 특성 수
min_samples_split	노드 분할 위한 최소 샘플 데이터 수
min_samples_leafs	리프노드가 되기 위해 필요한 최소 샘플 데이터 수

## 후보 모델



클래스 불균형이 심하기 때문에 class = 0을 정상치로 학습하면  
class = 1인 데이터를 이상치로 탐지하지 않을까?

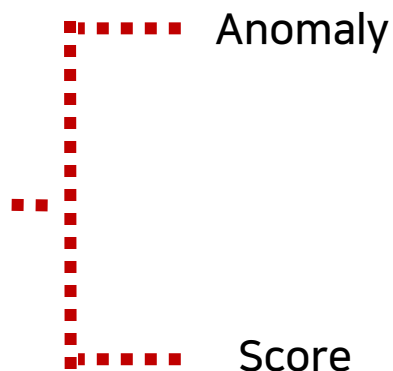


## Isolation Forest

전체 데이터셋을 decision tree을 통해서 각 노드에  
하나의 데이터포인트만 포함되도록 분리하는 모델로,  
이 때 분리 횟수가 적은 데이터들을 이상치로 탐지한다.

## 후보 모델

Isolation Forest  
output



각 데이터포인트가 이상치인지 정상치인지 판단하여  
정상치는 1, 이상치는 -1로 구성된 리스트

각 데이터포인트에 대한 점수로 1에 가까워질수록  
이상치로 판단 (range : (0,1))

n_estimators	생성할 Isolation tree 개수
max_samples	각 트리별 샘플 최대 비율
contamination	이상치의 비율

## 모델 설정

## XGB

$$\text{Weight} = \frac{\text{class 0 개수}}{\text{class 1 개수}}$$

CV 기법: Optuna

Threshold tuned

## LGBM

XGB와 유사하지만 Regression을  
통해서 예측을 진행하여  
threshold가 확률값이 아님

Threshold tuned

## Catboost

Weight = 'balanced'

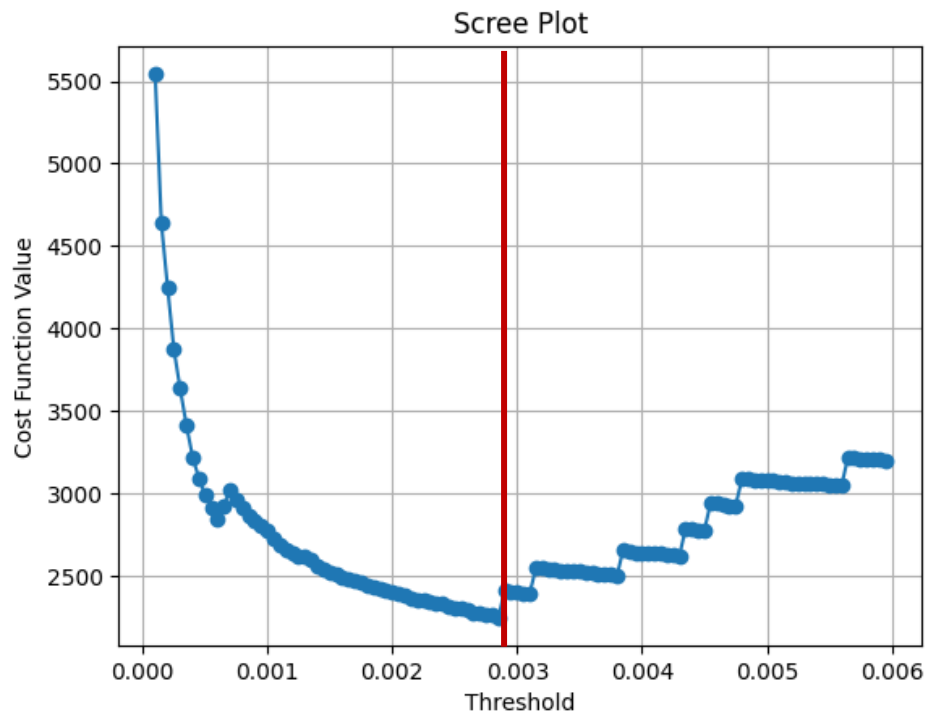
Threshold tuned

## Isolation Forest

Contamination = 0.08

Contamination 값 기준 CV 진행

## Threshold Tuning



1로 분류하는 기준확률(threshold)을 낮게 설정해 1로 많이 분류하기 위해  
0과 가까운 구간에서 cost를 최소화하는 threshold 탐색



## 모델링 성능 (validation 기준)

모델	cost	rank
XGBoost	2408	1
XGBoost_tuned	2670	3
LightGBM	3086	4
LightGBM_tuned	2593	2
Catboost	3576	5
Tabnet	6468	8
Isolation Forest	6380	7
Support Vector Machine	4864	6
Random Forest	8696	9

성능이 좋은 5개 모델로 앙상블 결정

## 양상블



## Hard Voting

Test set의 결과 중 N개 이상의 모델이 1로 예측한 경우만 최종 1로 판정  
이 때의 N은 voting에서 사용한 모델의 개수에 따라서 상이하게 설정함

XGB	XGB_ tuned	LGBM	LGBM _tuned	CatBoost	threshold	score
1	1				1	4635
1	1				2	4265
1	1	1			2	3555
1	1	1	1	1	3	4545
2	2	2	1	1	5	3950


## 양상블

XGB	XGB_ tuned	LGBM	LGBM _tuned	CatBoost	threshold	score
1	1				1	4635
1	1				2	4265
1	1	1			2	3555
1	1	1	1	1	3	4545
2	2	2	1	1	5	3950

동일한 가중치 하의 5개 모델 중 threshold인 3개 이상의 모델이 해당 data point를 1로 예측하였다면 1로, 그렇지 않다면 0으로 분류하는 양상블

## 양상블

보간을 완료한 데이터셋으로 학습시킨 모델



XGB	XGB_ 보간	LGBM	LGBM _보간	CatBoost	threshold	score
1	1		1		2	3745
1	1		1	1	2	4105
2	1	1	1		3	3865
2		1			2	4055
1	1		1		2	3910
2	3	1	3		6	3825
5	3	1				3975

⋮

위의 결과를 포함한 수많은 시도...

## 양상블

XGB	XGB_tuned	LGBM	LGBM_tuned	CatBoost	threshold	score
1	1				1	4635
1	1				2	4265
1	1	1			2	3555
1	1	1	1	1	3	4545
2	2	2	1	1	5	3950
...	...	...	...	...	...	...

최종 최고점 3555 with XGB, XGB\_tuned, LGBM

<<결측치 보간 안 한 모델이 1위>>

3개 모델 중 2개 이상이 1로 예측해야 최종 1

# 3

결론

## 인사이트

## XGB가 성능 면에서 좋을 수밖에 없었던 이유

- XGB가 과적합에 비교적 강건
- 결측치에 대해 sparsity-aware split finding 알고리즘을 통해 들어온 결측치를 한쪽 leaf에 몰아서 더 좋은 split point를 찾을 수 있도록 돕는다

Value  
Class

1.3		1.1	0.2		1.9	0.5		1.5	1.8
1	0	1	0	0	1	0	0	1	1

Value  
Class

0.2	0.5	0.8	1.1	1.3	1.5	1.9			
0	0	1	1	1	1	1	0	0	0

Value  
Class

			0.2	0.5	0.8	1.1	1.3	1.5	1.9
0	0	0	0	0	1	1	1	1	1

Best split, default direction = left

## 인사이트

## 최종적으로 앙상블에 모델을 3개만 사용한 이유

- 회귀 문제에서는 나쁜 모델(분산이 큰 모델)들을 많이 앙상블할 경우, 모델 전체의 분산이 줄어드는 효과가 있어 더 정확한 예측이 가능
- 그러나 **분류 문제**에서, 특히 Hard Voting의 경우 **전체 성능에 악영향**을 끼치는 투표가 이루어 질 수 있음



## 인사이트

## Isolation score &amp; anomaly를 앙상블에 사용하지 않은 이유

- Isolation Forest는 목적함수가 없기 때문에 분류 정확도를 높이는 방향으로 튜닝할 수 없음 → 분류에 적합한 모델 X
- isolation forest를 통해 이상치 탐지로 해당 문제를 접근해보았고, 그 결과 validation set에서 4800대의 cost가 나왔다. 분류 문제에 전혀 적합하지 않은 모델임에도 쓸만한 성능이 나왔기에, 그 output인 score와 anomaly를 파생변수로 사용함.

## 의의

## 데이터 전처리

- 클래스 불균형이 심각한 데이터에 대한 전처리 시 샘플링을 하는 방법과 모델 학습 과정에서 비용민감학습을 진행하는 방법의 성능을 비교해보고 어떤 상황에서 어떤 방법이 나은 지 알게 됨
- 차원이 매우 큰 데이터 분석 시 다양한 기준을 통한 변수 제어와 결측치 보간법을 적용해보고 성능을 평가해 봄

## 의의

## 모델링

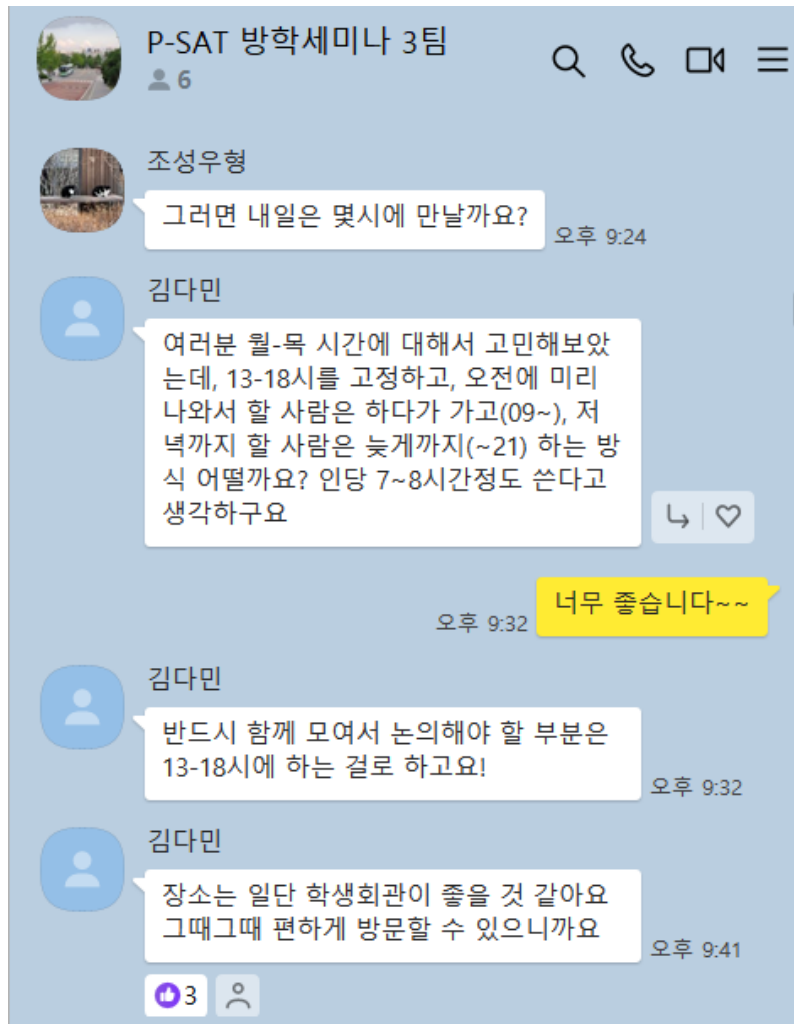
- 클래스 불균형이 심각한 데이터라는 점에서 착안하여 소수 class를 이상치로 탐지하도록 하는 방법인 isolation forest로 예측을 시도하는 창의적인 분석 시도
- 성능평가지표에서 FN(False Negative)의 영향력이 큰 상황에서 threshold 최적화를 통해 hyperparameter tuned model을 threshold tuning해서 모델을 fit하는 창의적인 분석방법 시도

## 의의

## 모델링

- **overfitting**이 발생하는 상황에서 preprocessing으로 돌아가 모델 개선에 영향을 미칠만한 방법을 찾는 과정에서의 많은 논의
- 보다 강건한 모델링을 위해 **각 모델별 가중치를 부여한 앙상블** 시도

## 3팀 개설




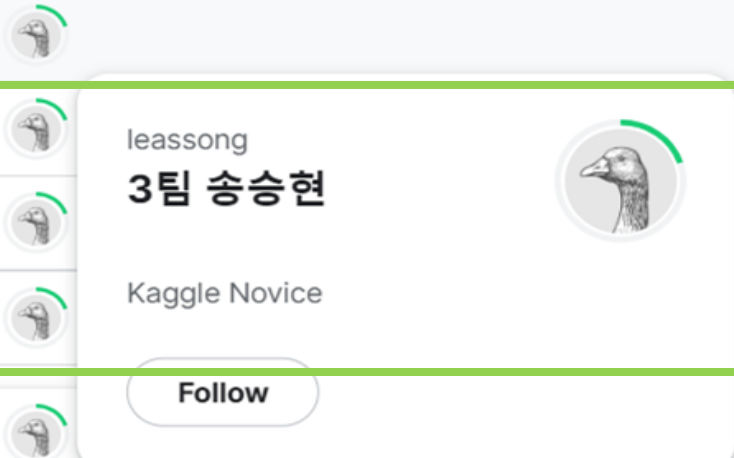
## 3팀 복지

- 자율출퇴근제
- 민트 캔디 / 간식 상비
- 웃음벨













선대회귀 많은 사랑 바랍니다

## 3팀 장기집권기

\*\*\* 5위 안에 들어야 3팀 타이틀 유지 가능








#	Team	Members	Score
1	3팀 심현구		4090.00000
2	3팀 김다민		4215.00000
3	3팀 조성우		4820.00000
4			5005.00000
5	3팀 서유진		5785.00000
6	3팀 송승현		5860.00000
7			8125.00000
8			9250.00000
9			10305.00000
10	그냥 김보근		13595.00000

## 3팀 난세

1			3150.00000	6	5h
2			3150.00000	11	12m
3			3230.00000	6	9h
4	3팀 송승현		3555.00000	6	9h
5			3590.00000	7	10h
6			3770.00000	9	12h
7	그냥 서유진		3950.00000	11	3h
8	그냥 조성우		3995.00000	8	3h
9	그냥 심현구		4090.00000	7	11h
 Your Best Entry! Your submission scored 4335.00000, which is not an improvement of your previous score. Keep trying!					
10	그냥 김다민		4135.00000	7	1d
11	그냥 김보근		4265.00000	9	4h

## 3팀 도움 탈환 시도

실패!

4	멧돼지 대장		3555.00000	6	1d
5	2팀 이상혁		3590.00000	7	1d
6	멧돼지1		3745.00000	10	10h
7	2팀 최용원		3770.00000	9	1d
8	멧돼지2		3825.00000	12	1h
9	멧돼지3		3865.00000	12	11h
10	멧돼지4		3910.00000	9	10h



Your Best Entry!  
Your most recent submission is a Great job!

11 멧돼지5







---

# 감사합니다

---

