

# ♡ 31기 방학세미나 2팀 ♡

김보현

이상혁

장다연

천예원

최용원

# INDEX

---

1. INTRO

2. EDA

3. 데이터 전처리

4. 모델링

1

INTRO

## 분석 목표 및 기준

### 목적함수

FN(False Negative): 양성임에도 음성으로 예측한 경우

FP(False Positive): 음성임에도 양성으로 예측한 경우

$$\text{Cost Function} = 250 \times \text{FN} + 5 \times \text{FP}$$

↘ FN에 큰 가중치 부여

		예측 결과	
		TRUE	FALSE
실제 정답	TRUE	TP ( True Positive )	FN ( False Negative )
	FALSE	FP ( False Positive )	TN ( True Negative )

## 분석 목표 및 기준

### 목적함수

FN(False Negative): 양성임에도 음성으로 예측한 경우

FP(False Positive): 음성임에도 양성으로 예측한 경우

$$\text{Cost Function} = 250 \times \text{FN} + 5 \times \text{FP}$$

↘ FN에 큰 가중치 부여

		예측 결과	
		TRUE	FALSE
실제 정답	TRUE	TP ( True Positive )	FN ( False Negative )
	FALSE	FP ( False Positive )	TN ( True Negative )



목적 함수를 **최소화**하는 이진 분류 모델의 구성

## 분석 흐름



변수 선택

K-S Test



파생 변수 생성

데이터 분포 기반 파생 변수, 행 별 이상치 수,  
행 별 결측치 수, 행 별 0 수, 열 별 결측 여부



모델링

**LightGBM**, XGBoost, CatBoost,  
Random Forest, Logistic Regression, One class SVM

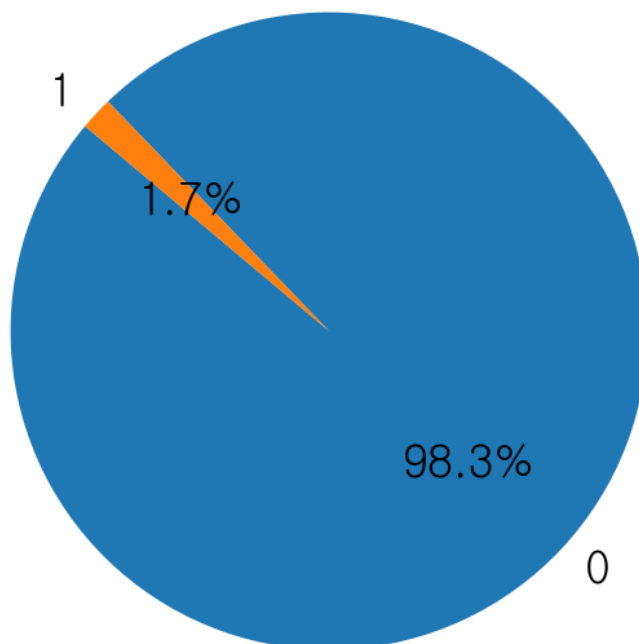
2

EDA

## 데이터 구조 확인

🚨 Class 0과 Class 1의 비율이 98:2로 불균형 존재 🚨

Class 변수 분포 확인





## 데이터 구조 확인



Class 0과 Class 1의 비율이 98:2로 불균형 존재



불균형 해소 위해  
다양한 전처리 시도!



랜덤 오버 샘플링

SMOTE

class weight 파라미터 설정...

*전처리에서 더 자세히!*

## 데이터 구조 확인



ID, class 제외 총 170개의 설명변수 존재

X1	X2	X3	X4	X5	X6	X7a	X7b	...	X105c	X105d	X105e	X105f	X105g	X105h	X105i	X105j	X106	X107
3490	0.0	8.000000e+01	212.0	0.0	0.0	0.0	0.0	...	12986.0	7612.0	17044.0	13682.0	19594.0	74564.0	5270.0	0.0	0.0	0.0
92	0.0	1.400000e+01	10.0	0.0	0.0	0.0	0.0	...	512.0	120.0	332.0	344.0	964.0	1414.0	0.0	0.0	0.0	0.0
10	0.0	1.800000e+01	2.0	4.0	6.0	0.0	0.0	...	132.0	22.0	32.0	24.0	26.0	54.0	0.0	0.0	0.0	0.0
390692	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
156758	NaN	2.130706e+09	408.0	0.0	0.0	0.0	0.0	...	1652370.0	852932.0	1494660.0	1026644.0	779338.0	480460.0	784792.0	33582.0	0.0	0.0

## 데이터 구조 확인



ID, class 제외 총 170개의 설명변수 존재

X1	X2	X3	X4	X5	X6	X7a	X7b	...	X105c	X105d	X105e	X105f	X105g	X105h	X105i	X105j	X106	X107
3490	0.0	8.000000e+01	212.0	0.0	0.0	0.0	0.0	...	12986.0	7612.0	17044.0	13682.0	19594.0	74564.0	5270.0	0.0	0.0	0.0
92	0.0	1.400000e+01	10.0	0.0	0.0	0.0	0.0	...	512.0	120.0	332.0	344.0	964.0	1414.0	0.0	0.0	0.0	0.0
10	0.0	1.800000e+01	2.0	4.0	6.0	0.0	0.0	...	132.0	22.0	32.0	24.0	26.0	54.0	0.0	0.0	0.0	0.0
390692	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
156758	NaN	2.130706e+09	408.0	0.0	0.0	0.0	0.0	...	1652370.0	852932.0	1494660.0	1026644.0	779338.0	480460.0	784792.0	33582.0	0.0	0.0



모든 변수는 수치형 변수로 구성

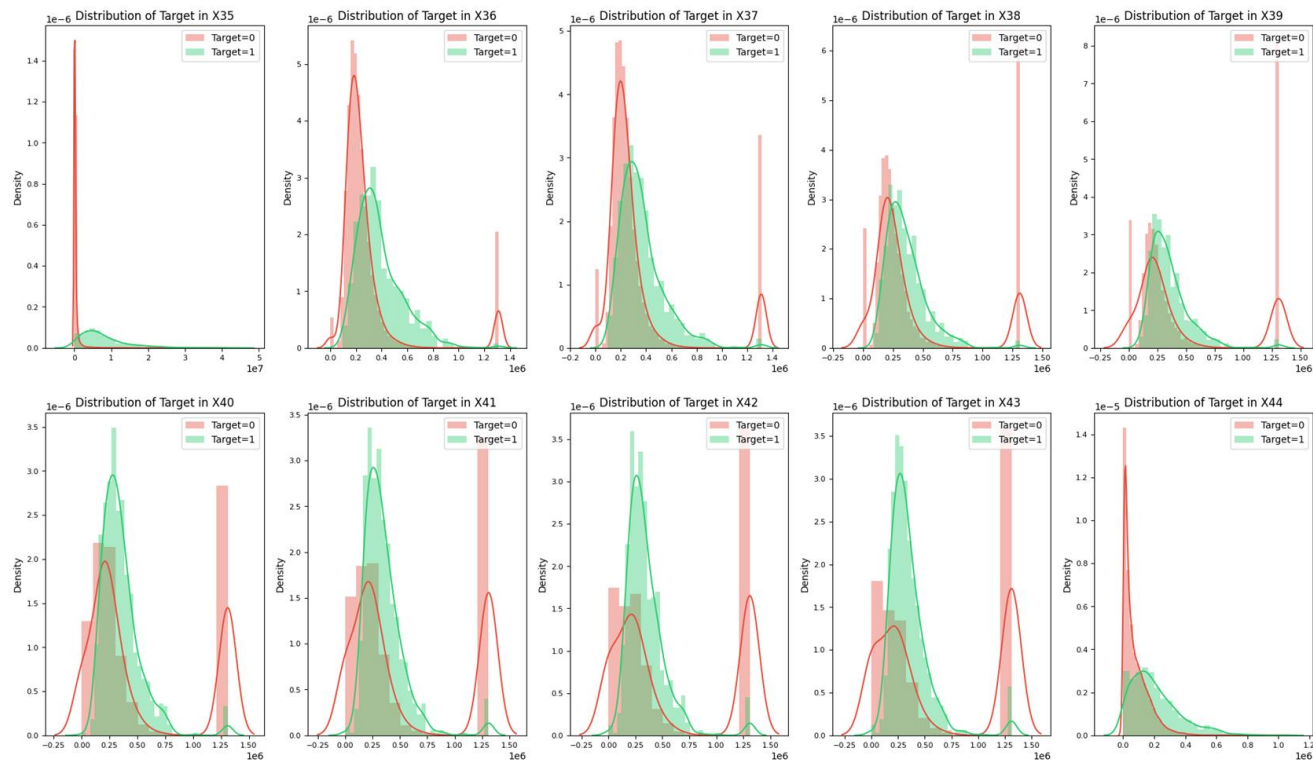
### 기술통계량

	count	mean	std	min	25%	50%	75%	max
X1	55000.0	5.912640e+04	1.445713e+05	0.0	818.0	30750.0	48532.0	2.434708e+06
X2	12585.0	6.944776e-01	3.305901e+00	0.0	0.0	0.0	0.0	2.040000e+02
X3	51939.0	3.554258e+08	7.943500e+08	0.0	16.0	150.0	962.0	2.130707e+09
X4	41404.0	2.077730e+05	4.218748e+07	0.0	24.0	124.0	430.0	8.584298e+09
X5	52723.0	6.714489e+00	1.643493e+02	0.0	0.0	0.0	0.0	2.105000e+04
...	...	...	...	...	...	...	...	...
X105h	54390.0	3.454091e+05	1.738760e+06	0.0	108.0	41000.0	167980.0	1.195801e+08
X105i	54390.0	1.380590e+05	4.431082e+05	0.0	0.0	3775.0	139213.5	1.456059e+07
X105j	54390.0	8.299358e+03	4.640349e+04	0.0	0.0	0.0	2011.5	3.810078e+06

### 자료형 확인

X1	int64
X2	float64
X3	float64
X4	float64
X5	float64
...	...
X105h	float64
X105i	float64
X105j	float64
X106	float64
X107	float64

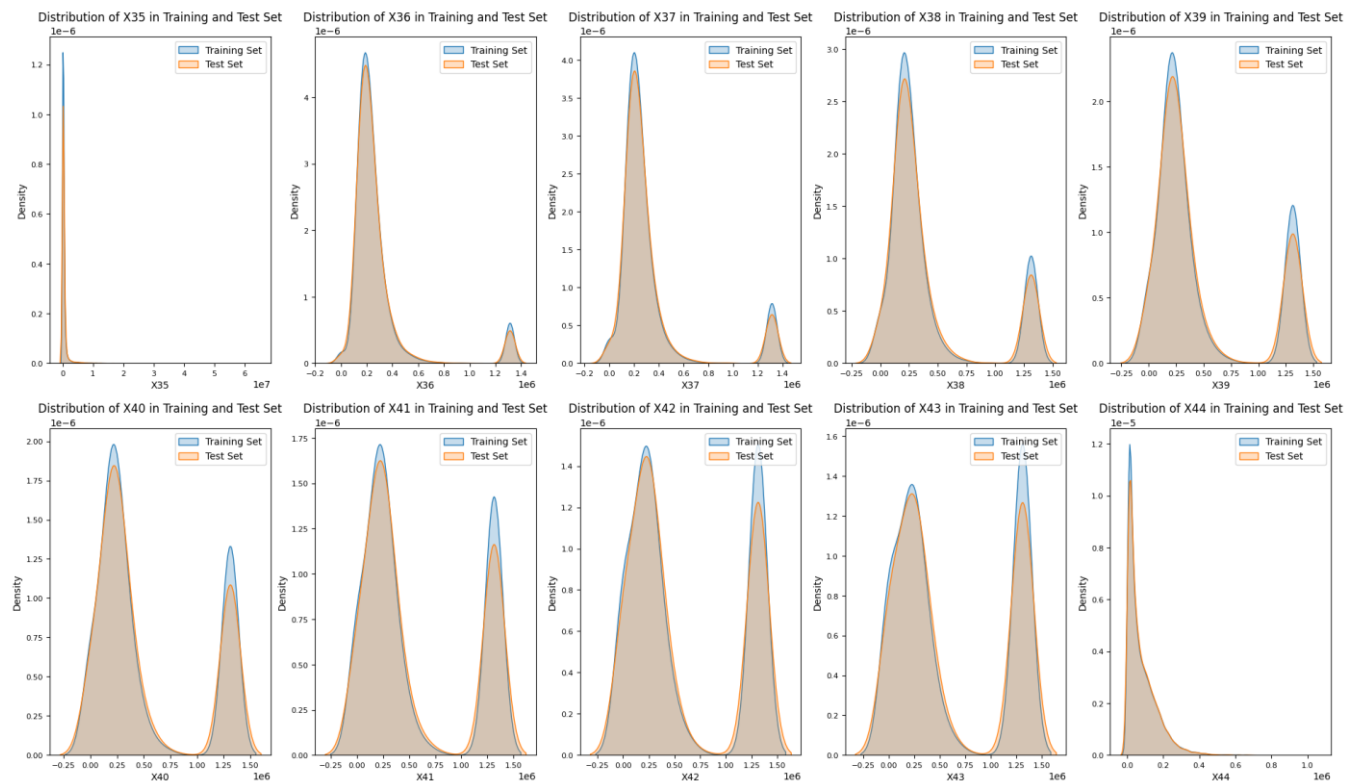
## 데이터 구조 확인



대부분의 설명변수가 왼쪽으로 치우쳐 있음을 확인

Class 별 유사한 분포를 보임

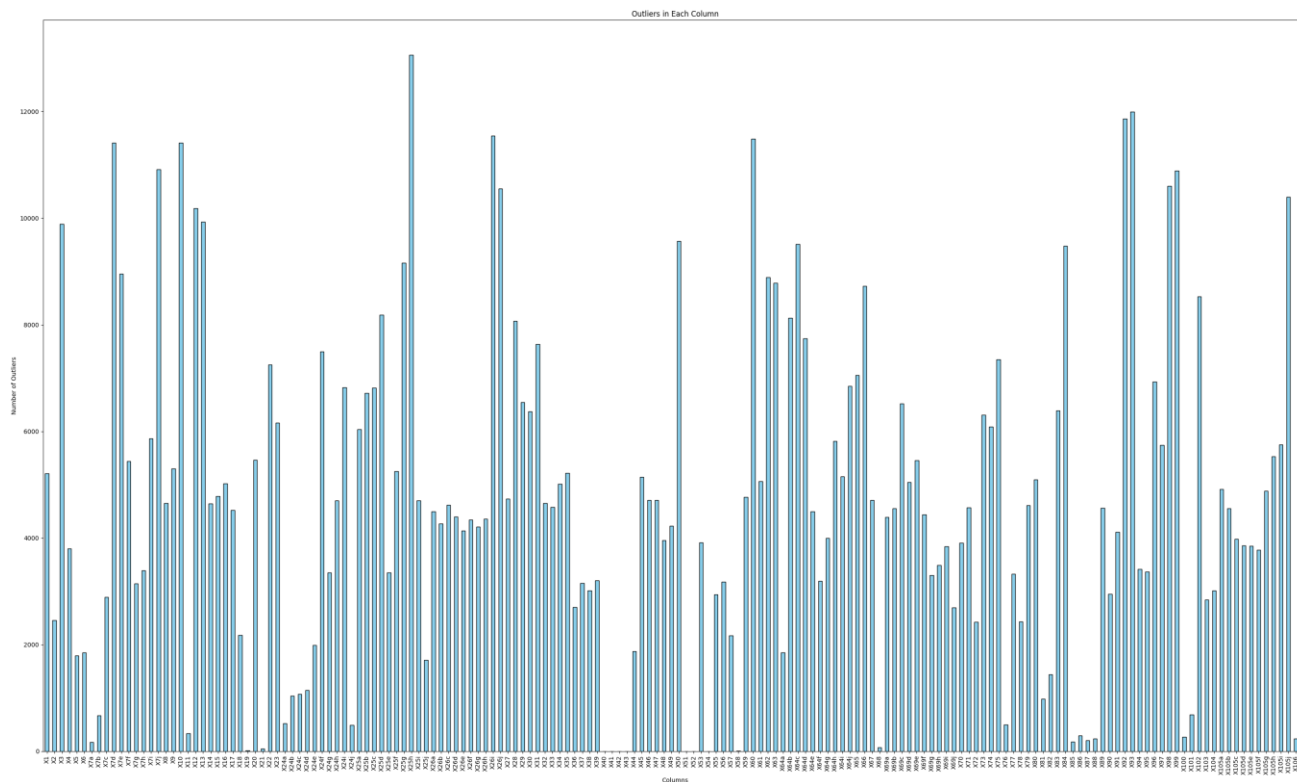
## 데이터 구조 확인



train과 test 데이터의 분포 확인 결과

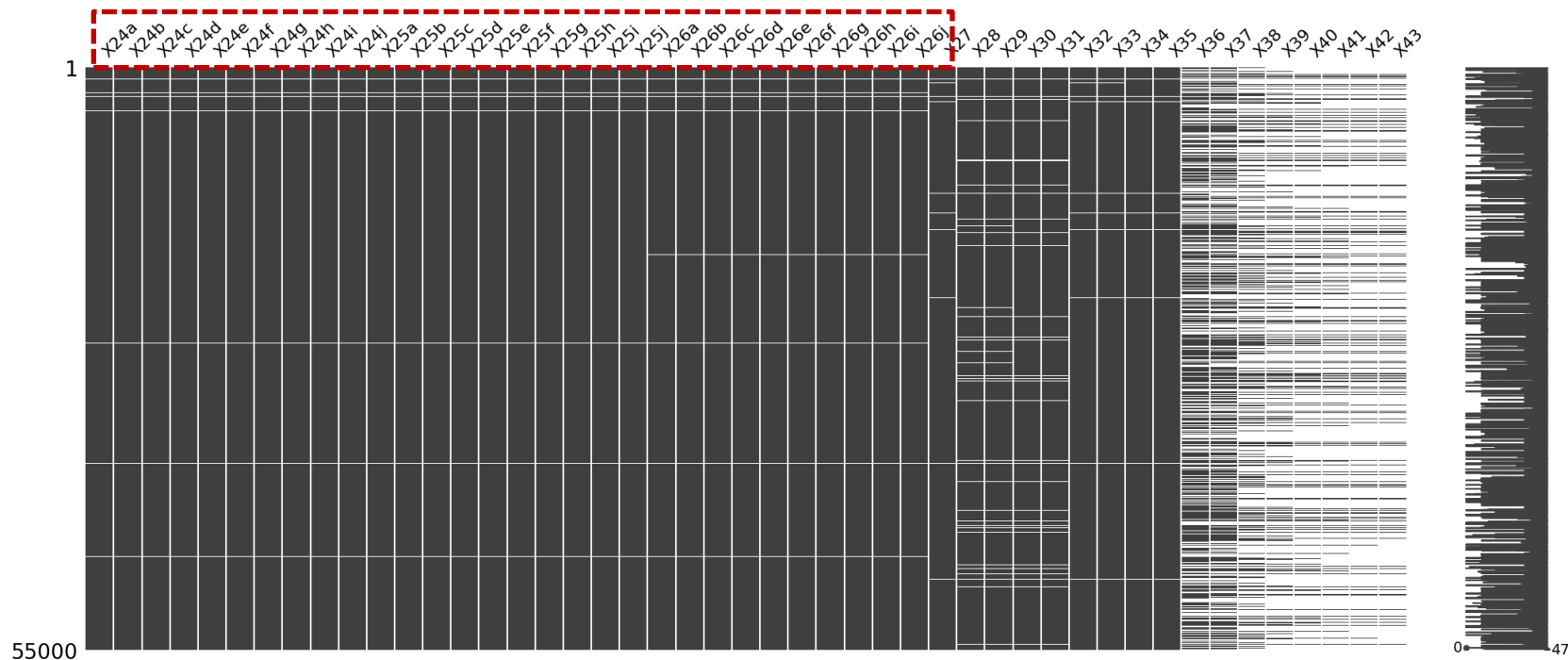
유사한 분포를 가지는 것을 확인

## 데이터 구조 확인



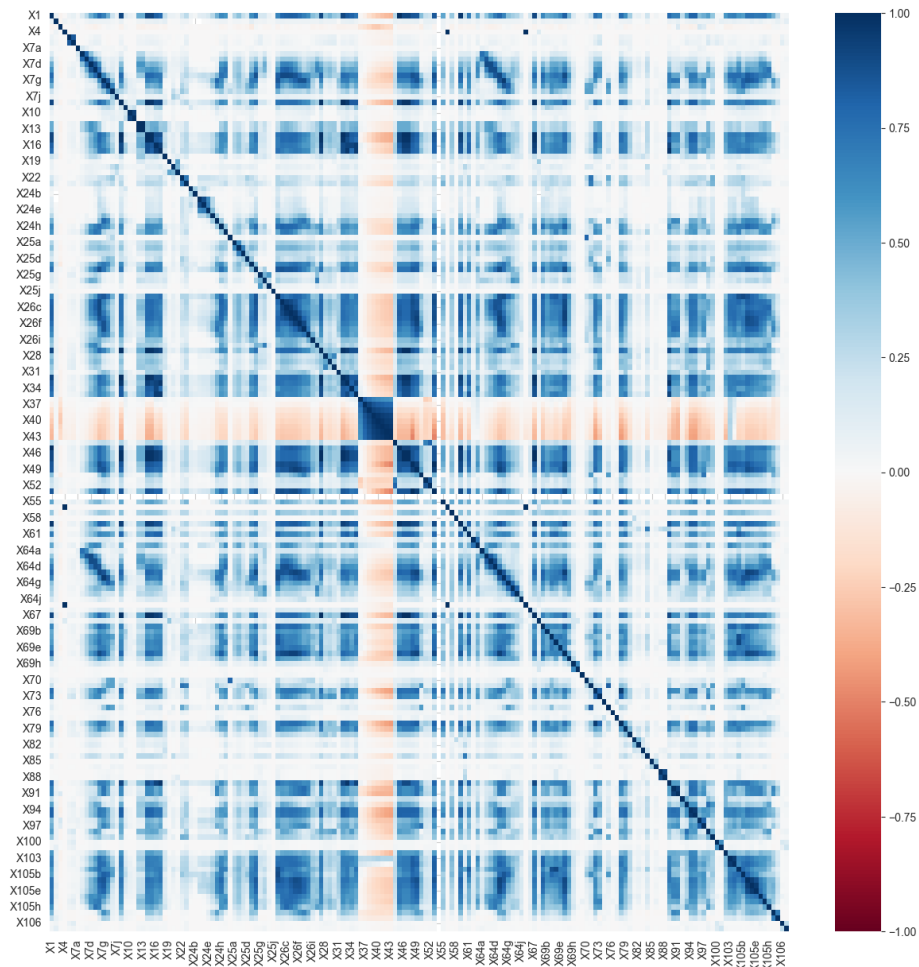
IQR 사용해 열 별 이상치 탐지 결과  
대부분의 변수에서 많은 이상치 발견

## 데이터 구조 확인



열 이름 'X+숫자' 조합이 같은 변수는 동일한 결측치를 가짐  
 유의미한 관련성 추측 가능

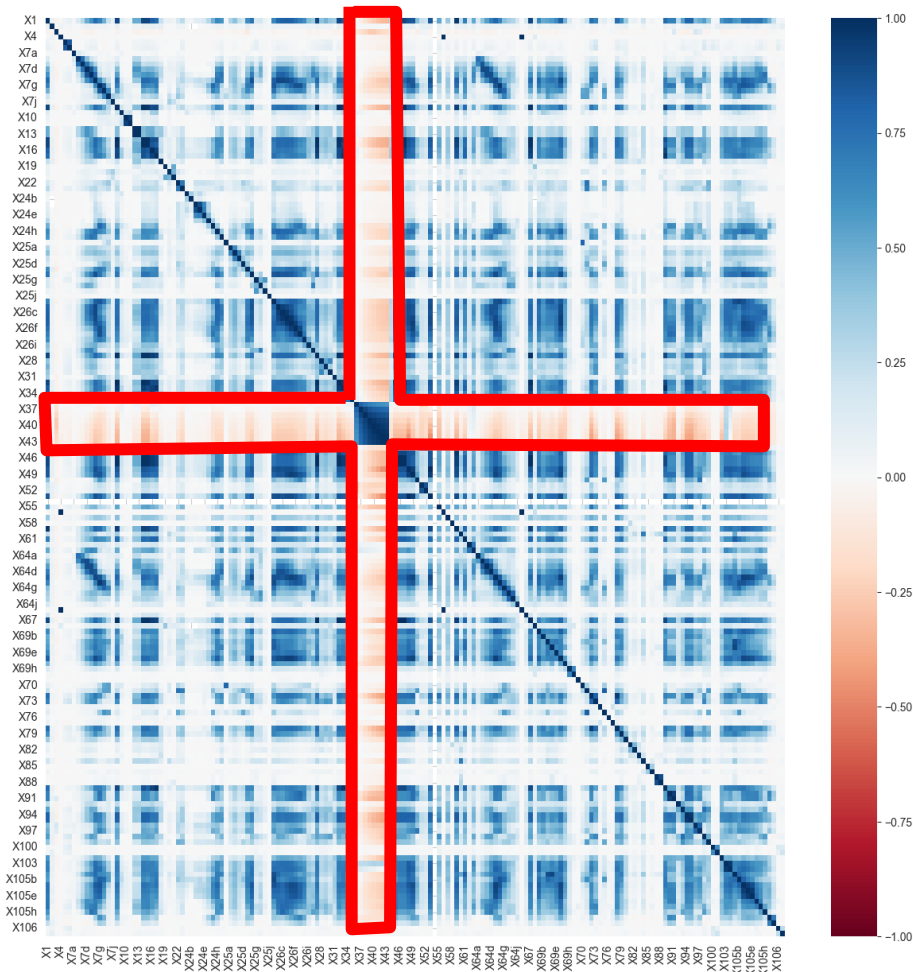
## 데이터 구조 확인



X 변수 간 상관관계 존재 확인



## 데이터 구조 확인



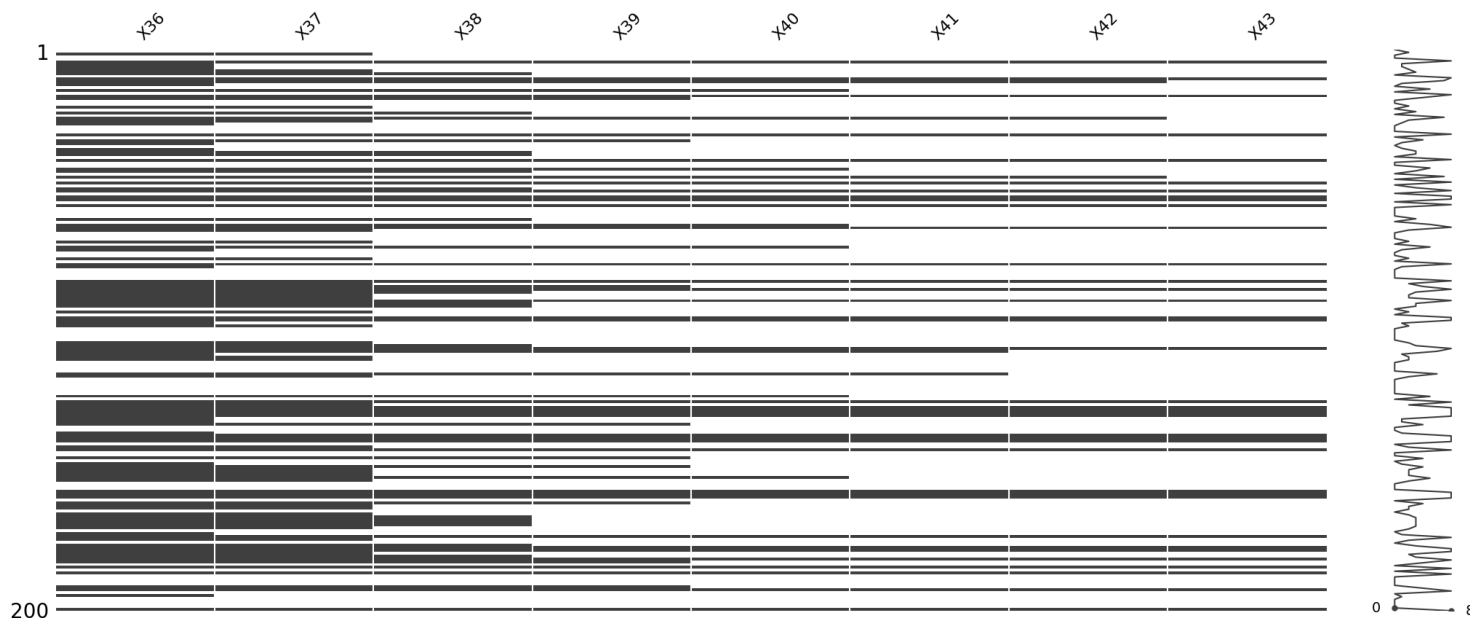
X 변수 간 상관관계 존재 확인



X36 ~ X43

다른 변수와 구별되는 패턴 !

## 데이터 구조 확인



상관관계와 결측치 분포를 통해

X36 ~ X43 변수가 연속적 특성을 지니고 있음을 추론

## 동질성 검정

### t 검정

정규분포 하에서 두 그룹 모집단 평균의 동일 여부 검정

Class에 따라 각 변수 별 동질성 확인

$H_0$  : 두 집단의 평균이 동일하다.

## 동질성 검정

### t 검정

정규분포 하에서 두 그룹 모집단 평균의 동일 여부 검정

Class에 따라 각 변수 별 동질성 확인

$H_0$  : 두 집단의 평균이 동일하다.



총 170개의 설명 변수에 대해

11개의 변수가 통계적으로 유의하지 않음을 확인

$X_4, X_{25j}, X_{54}, X_{56}, X_{65}, X_{69i}, X_{69j}, X_{76}, X_{85}, X_{87}, X_{88}$

## 동질성 검정

### Kolmogorov-Smirnov Test (K-S 검정)

회귀분석팀 2주차 클린업 참고!

하나의 모집단이 어떤 특정한 **분포함수**를 갖는지 알아보는 방법  
귀무가설 하에서 표본분포함수가 어떤 이론적 분포함수와 유사한지를 검정  
이론분포함수와 표본분포함수의 차가 크면  $H_0$ 을 기각

$H_0$  : 두 분포가 동일 분포이다.

## 동질성 검정

### Kolmogorov-Smirnov Test (K-S 검정)

회귀분석팀 2주차 클린업 참고!

하나의 모집단이 어떤 특정한 **분포함수**를 갖는지 알아보는 방법  
귀무가설 하에서 표본분포함수가 어떤 이론적 분포함수와 유사한지를 검정  
이론분포함수와 표본분포함수의 차가 크면  $H_0$ 을 기각

$H_0$  : 두 분포가 동일 분포이다.



총 170개의 설명 변수에 대해  
5개의 변수가 통계적으로 유의하지 않음을 확인

$X2, X19, X54, X68, X69j$

## 동질성 검정

### Kolmogorov-Smirnov Test (K-S)



하나의 모집단이 어떤 특정한 분포함수를 갖는지 알아보는 방법

귀무가설 하에 표본분포함수가 어떤 이론적 분포함수와 유사한지를 검정

분포 가정 등을 고려해, **K-S 검정**을 통한 변수 선택 실행

이론분포함수와 표본분포함수의 차가 크면  $H_0$ 을 기각

시계열팀 주제분석 변수선택법!

$H_0$  : 두 분포가 동일 분포이다.



총 170개의 설명 변수에 대해

5개의 변수가 통계적으로 유의하지 않음을 확인

$X_2, X_{19}, X_{54}, X_{68}, X_{69j}$

## 동질성 검정

## Kolmogorov-Smirnov Test (K-S)



하나의 모집단이 어떤 특정한 분포함수를 갖는지 알아보는 방법

귀무가설 하에  $H_0$  : 분포분포함수가 어떤 이론적 분포함수와 유사한지를 검정

분포 가정 등을 고려해, **K-S 검정**을 통한 변수 선택 실행

이론분포함수와 표본분포함수의 차가 크면  $H_0$ 을 기각

시계열팀 주제분석 변수선택법!

$H_0$  : 두 분포가 동일 분포이다.



5개의 변수 제거 후

총 170개의 설명 변수에 대해

**165개**의 변수 사용

5개의 변수가 통계적으로 유의하지 않음을 확인

X2, X19, X54, X68, X69j

X2, X19, X54, X68, X69j



# 3

## 데이터 전처리

## 파생변수 생성

데이터의 분포와 특성을 고려하여  
이를 반영할 수 있는 파생변수를 생성

X36 ~ X43  
범위 변수

X36 ~ X43  
변동계수 변수

이상치 개수 변수

결측치 개수 변수

0의 개수를  
이용한 변수

결측치 유무를  
이용한 변수

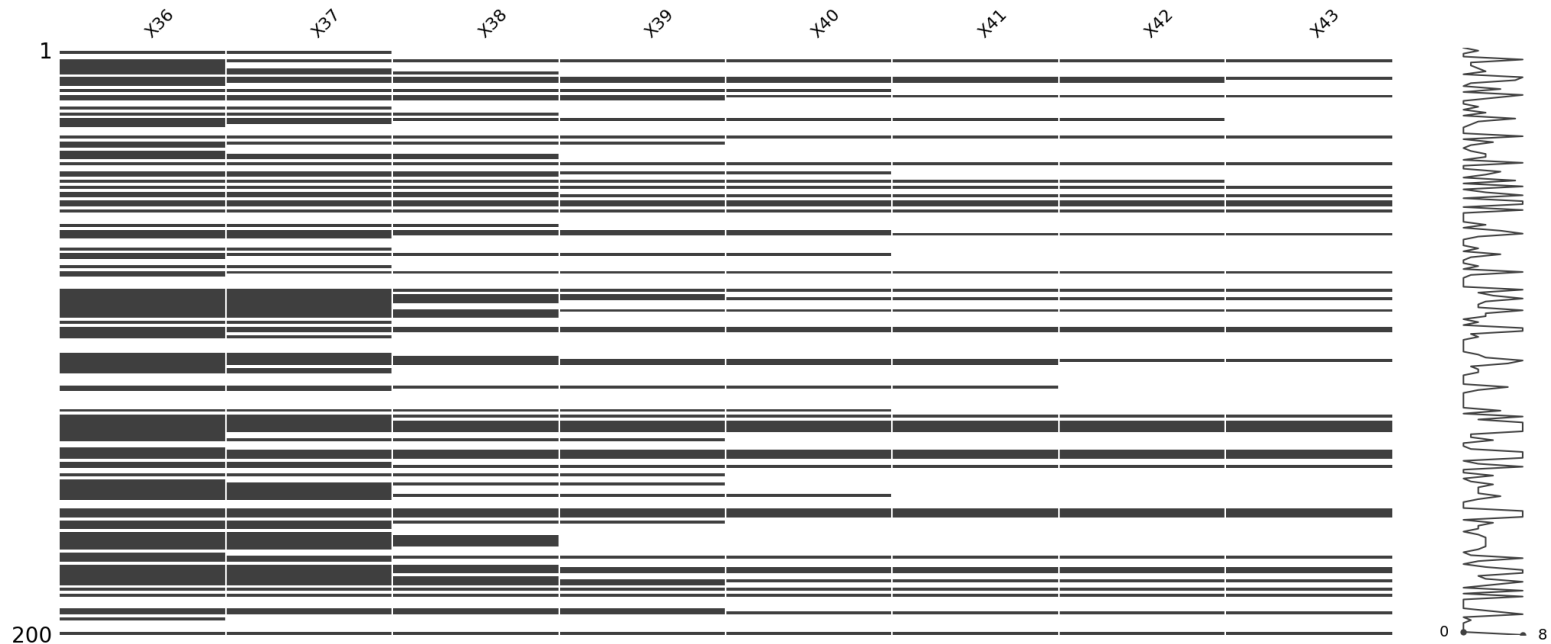
## 파생변수 생성

데이터의 **분포**와 **특성**을 고려하여  
이를 반영할 수 있는 **파생변수**를 생성

## 데이터의 특징

- ✓ X36부터 X43까지 패턴이 존재함
- ✓ 결측치와 이상치가 많음
- ✓ 샘플들이 0 주변에 주로 분포되어 있음

## 파생변수 생성



결측치의 개수가 X43으로 갈수록 줄어듦

시간의 흐름에 따라 기록된 **시계열 데이터**일 수 있다는

가정하에 파생변수 생성

## 파생변수 생성 - ①

range\_X36\_to\_X43(범위 변수)

데이터의 연속적인 특징을 살리기 위해  
각 샘플별로 X36~X43의 범위를 계산

X36부터 X43까지 패턴이 존재함

⋮

시계열 데이터의 특징을 보이기 때문에  
Min 값과 Max 값의 차이를 보여주는  
변수를 생성

## 파생변수 생성 - ②

coef\_variation\_X36\_to\_X43(변동계수 변수)

결측치가 많은 데이터 특징을 살리기 위해  
각 샘플별로 X36~X43의 **변동 계수**를 계산

X36부터 X43까지 패턴이 존재함

⋮

시계열 데이터의 특징을 보이기 때문에

**상대적인 산포도를 비교**하기 위해

변동 계수를 계산함

$$\text{변동 계수}(\text{coefficient of variation, CV}) = \frac{\sigma}{\bar{x}}$$

## 파생변수 생성 - ③

outliers\_count(이상치 개수 변수)

데이터의 분포를 통해 다수의 이상치를 확인해  
모델이 이를 고려할 수 있도록 **이상치 개수**를 변수로 생성

결측치와 이상치가 많음

⋮

class0이 가진 특징이 오히려 이상치로

판단될 수 있다는 점을 고려,

각 샘플이 가진 이상치 정보가 분류에

도움이 될 수 있다고 판단함

## 파생변수 생성 - ④

`count_NA(결측치 개수 변수)`

데이터에 결측치가 많기 때문에 이를 따로 보간하지 않고  
모델이 고려할 수 있도록 **결측치 개수**를 변수로 생성

결측치와 이상치가 많음

⋮

각 샘플별로 결측치의 개수를 파생변수로 생성하면  
클래스 분류에 유의미한 영향을 줄 수 있다고 판단함



## 파생변수 생성 - ⑤

X1\_missing, ..., X170\_missing(결측치 개수 변수)

샘플 뿐만 아니라 각 변수에도 많은 결측치가 있기 때문에  
샘플이 각 컬럼에 대해 결측치를 가지는지의 유무를 표현함

결측치와 이상치가 많음

⋮

165개의 변수에 대해서 샘플이  
각 컬럼에 대해 결측치가 있으면 1을,  
없으면 0을 갖는 범주형 파생변수를 생성함

## 파생변수 생성 - ⑥

zero\_counts\_by\_row(0 개수 변수)

0 주변에 많은 데이터가 몰려있기 때문에  
클래스 분류에 유의미한 차이를 확인하기 위해 **0 개수**를 변수로 생성

샘플들이 0 주변에 주로 분포되어 있음

⋮

각 샘플별로 0의 개수를 파생변수로 생성하면  
클래스 분류에 유의미한 영향을 줄 수 있다고 판단함

## 결측치 보간 - ①

## MICE (Multiple Imputation by Chained Equations)

여러 개의 독립적인 모델을 사용하여 반복적으로 결측치를 예측하고, 결측치의 불확실성과 분산을 반영해 최종 데이터셋을 구성하는 방법

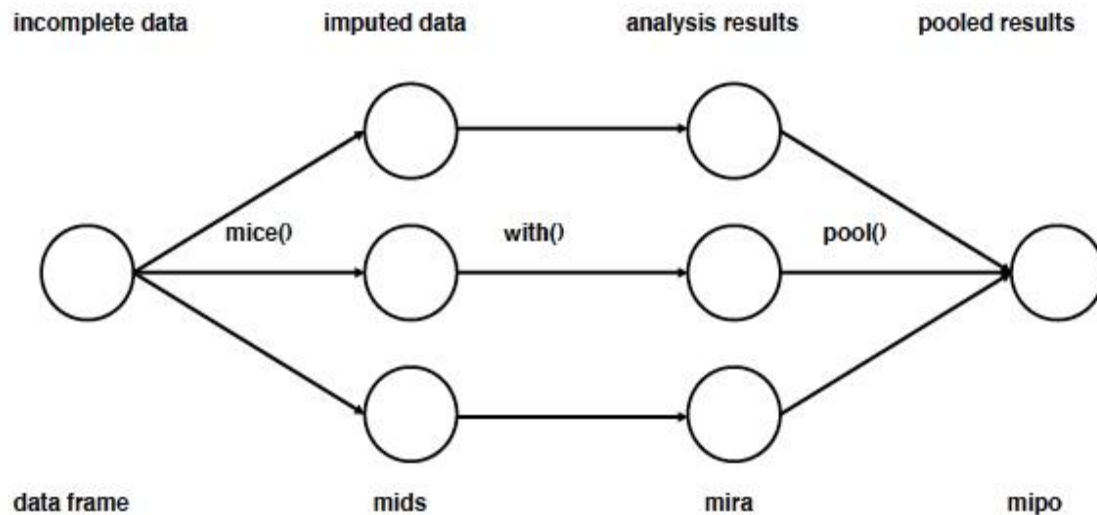


Figure 1: Main steps used in multiple imputation.

## 결측치 보간 - ①

MICE (Multiple Imputation by Chained Equations)

여러 개의 독립적인 모델을 사용하여 반복적으로 결측치를 예측하고,  
결측치의 불확실성과 분산을 반영해 최종 데이터셋을 구성하는 방법



결측치 대체 시간이 오래 걸리며, 과대적합 발생 가능성 존재



최종 모델에 사용하지 않음

## 결측치 보간 - ②

## Iterative Imputation

다른 변수들과의 상호작용을 고려하여 반복적으로 대체하는 방법  
대체된 결측치를 기반으로 다른 결측치를 채워나감



고차원 데이터이기 때문에 회귀 기반 보간법은 적절하지 않다고 판단



⋮

최종 모델에 사용하지 않음

### 3 데이터 전처리

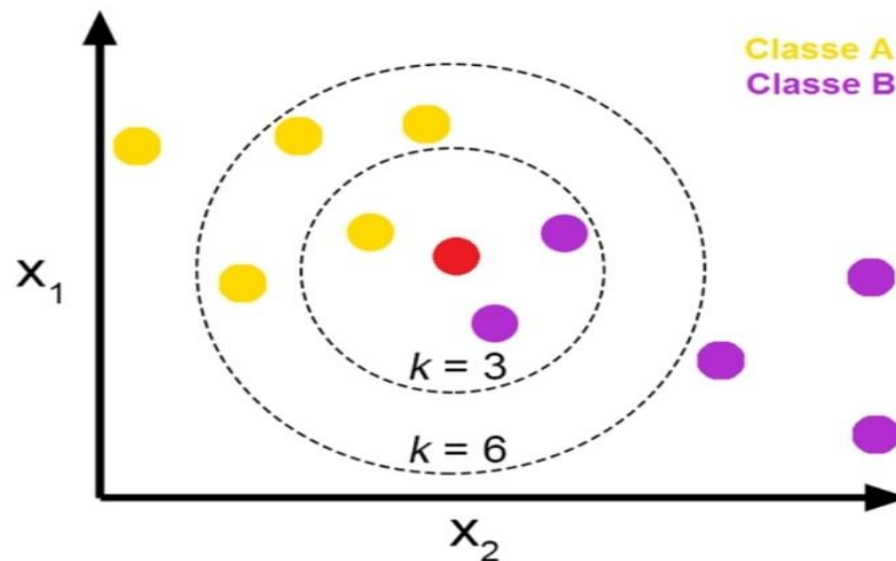
#### 결측치 보간 - ③

K - 최근접 이웃 대체 (K-Nearest Neighbors Imputation)

결측치와 가장 유사한 K개의 이웃을 찾아 대체하는 방법

주로 유클리드 거리나 맨해튼 거리를 사용해 계산됨

*데마팀 1주차 클린업 참고!*



## 결측치 보간 - ③

K - 최근접 이웃 대체 (K-Nearest Neighbors Imputation)

결측치와 가장 유사한 K개의 이웃을 찾아 대체하는 방법

주로 유클리드 거리나 맨해튼 거리를 사용해 계산됨

*데마팀 1주차 클린업 참고!*



차원이 증가함에 따라 거리 계산의 정확도가 떨어짐



최종 모델에 사용하지 않음

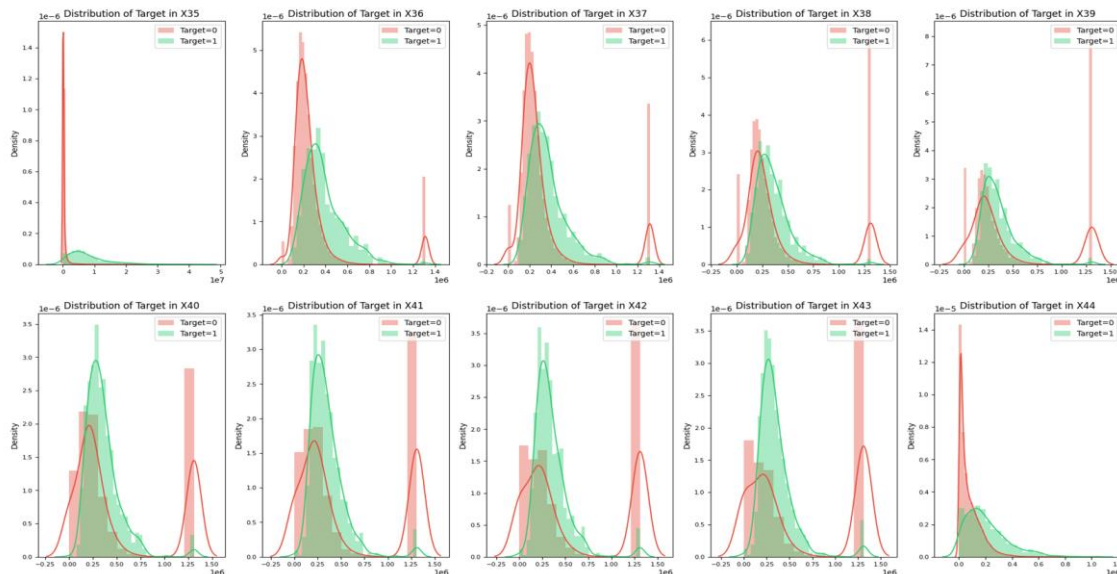
# 3 데이터 전처리

## 결측치 보간 - ④

### Single Imputation

① 평균 대체

② 중앙값 대체



분포 시각화를 통해 데이터를 대표하기에 적절하지 않다고 판단해 사용하지 않음







## 결측치 보간 - ④ Single Imputation

Single Imputation 최종적으로 결측치 처리 없이 모델링 진행!

사용한 모델인 XGB, LGBM은  
 ① 평균 대체 ② 중앙값 대체  
 자체적으로 결측치를 처리해 모델링을 진행하기 때문

분포 특성으로 인해 대표값이 될 수 없다는 판단~

⋮

선형회귀모델이 만족해야 하는 제약들이 곧 선형회귀의 기본 가정



## 결측치 보간 - ④ Single Imputation

Single Imputation 최종적으로 결측치 처리 없이 모델링 진행!

사용한 모델인 XGB, LGBM은  
 ① 평균 대체 ② 중앙값 대체  
 자체적으로 결측치를 처리해 모델링을 진행하기 때문



XGB

분포 특성으로 인해 대표값이 될 수 없다는 판단~

트리를 분기할 때 결측치만으로 하위 트리를 구성한 후,

가장 높은 점수를 낸 방향으로 모든 결측치 처리

선형회귀모델이 만족해야 하는 제약들이 곧 선형회귀의 기본 가정  
 자세한 내용은 2023 여름방학 논문 스터디 참고!



## 결측치 보간 - ④ Single Imputation

Single Imputation 최종적으로 결측치 처리 없이 모델링 진행!

사용한 모델인 XGB, LGBM은  
 ① 평균 대체 ② 중앙값 대체  
 자체적으로 결측치를 처리해 모델링을 진행하기 때문



LGBM

분포 특성으로 인해 대표값이 될 수 없다는 판단~

분할 시 자체적으로 결측치를 대체하거나,

결측치 유무를 기준으로 분할을 진행

선형회귀모델이 만족해야 하는 제약들이 곧 선형회귀의 기본 가정

## 기타 전처리

최종적으로 채택되지는 않았으나, 다양한 방식의 전처리 진행

로그 변환

Box-Cox  
Transformation

PCA

오버샘플링  
스케일링

## 기타 전처리

## 1. 로그 변환

`np.log1p()`로 왜도(skew)가 높은 변수들에 대해 로그 변환 처리

편차를 줄여 변수 분포의 정규성을 향상시키기 위해  
전체 로그 변환 / 왜도가 높은 상위 50% 로그 변환 시도



유의미한 성능 차이를 보이지 않았음

## 기타 전처리

## 2. Box-Cox Transformation

통계적인 검정에 따라 변수를 비선형 변환하는 방법

회귀팀 2주차 클린업 참고!

$$y(\lambda) \begin{cases} (y^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0 \end{cases}$$

$\lambda$ 를 변화시키면서,  $y$ 가 정규성 & 등분산성을 만족하도록 변환

⋮

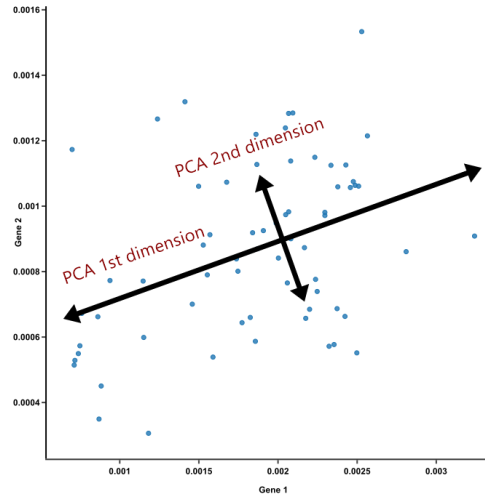
☹️ 유의미한 성능 차이를 보이지 않았음

## 기타 전처리

## 3. 주성분 분석 (Principal Component Analysis)

변수 간의 상관관계가 존재하는 다차원 공간의 데이터를  
저차원 공간의 데이터로 변환하는 차원축소 기법

선대팀 3주차 클린업 참고!



변수별 상관관계가 존재하므로 차원 축소를 위해 PCA 진행

## 기타 전처리



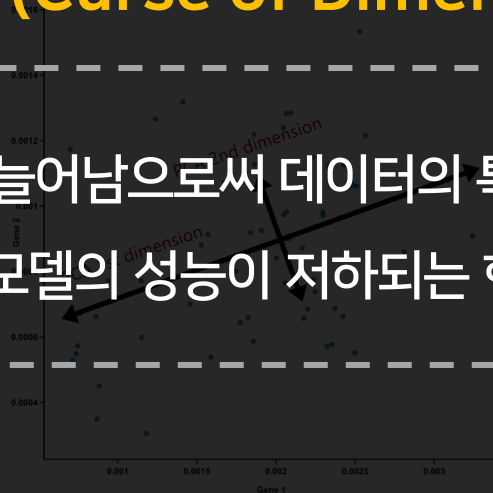
### 3. 주성분 분석 (Principal Component Analysis)

왜 차원 축소를 해야 할까?

변수 간의 상관관계가 존재하는 다차원 공간의 데이터를  
저차원 공간의 데이터로 변환하는 차원 축소 기법

## 차원의 저주 (Curse of Dimensionality)

차원의 수가 늘어남으로써 데이터의 특징이 너무 많아  
모델의 성능이 저하되는 현상



선대팀 3주차 클린업 참고!

데마팀 1주차 클린업 참고!

변수별 상관관계가 존재하므로 차원 축소를 위해 PCA 진행



## 기타 전처리: 주성분 분석 (PCA)



가장 큰 분산을 차지하는 주성분이 **분산 35%**, 그 외는 4-5%로 나타남

⋮

PCA는 비지도학습이므로 적절한 기준을 찾기 어렵고,  
변수의 개수가 유의미하게 줄어들지 않아 사용하지 않음

## 기타 전처리

## 4. 오버 샘플링 (Over Sampling)

소수의 클래스를 다수의 클래스에 맞춰 관측치를 증가시키는 기법

*범주형 3주차 클린업 참고!*

Random Over  
Sampling

랜덤으로 소수의 클래스를 **복제**

## SMOTE

KNN 알고리즘을 활용해  
**가상의 소수 클래스 데이터 생성**



언더샘플링은 관측치를 삭제해 정보의 손실이 발생할 수 있어 제외

## 기타 전처리: 오버 샘플링

클래스 불균형은 해결되었으나, 오버피팅 발생하여 채택하지 않음

⋮



Random Over  
Sampling

과적합에 취약

SMOTE

노이즈가 발생해  
고차원 데이터에는 효율적 X

## 기타 전처리

## 5. 스케일링

수치형 변수의 범위를 조절해주는 과정

## Standard Scaling

평균 0, 분산 1로 표준 정규 분포화

## Min-Max Scaling

최솟값이 0, 최댓값이 1이  
되도록 정규화

## 기타 전처리

## 5. 스케일링

수치형 변수의 범위를 조절해주는 과정

## Standard Scaling

평균 0, 분산 1로 표준 정규 분포화

## Min-Max Scaling

최솟값이 0, 최댓값이 1이  
되도록 정규화

☹️ 유의미한 성능 향상을 보이지 않아 채택 X

## 변수 선택 (Feature Selection)

데이터의 특성을 가장 잘 설명하는 변수를 추가 / 제거해서 모델을 적합시키는 방법

*데마팀 1주차 클린업 참고!*

### 변수 선택 종류

Forward Selection

Backward Elimination

Step-wise Selection

## 변수 선택 (Feature Selection)

데이터의 특성을 가장 잘 설명하는 변수를 추가 / 제거해서 모델을 적합시키는 방법

*데마팀 1주차 클린업 참고!*

### 변수 선택 종류

Forward Selection  
Backward Elimination  
Step-wise Selection

### 이번 방세에서는 ...

RFE

Feature Importance  
& Voting

## 변수 선택 (Feature Selection)

### 1. RFE (Recursive Feature Elimination)

전체 features를 활용해 모델을 생성하고, 원하는 개수의 피처가 남을 때까지 feature importance가 낮은 변수를 제거하는 **backward** 방식

- ✓ RFECV (Recursive Feature Elimination with Cross Validation)를 통해 feature 개수 지정 없이 성능 도출 가능!

```
# 최적의 feature 개수 : 111개  
selector.support_.sum()
```

```
111
```

▲ LGBM 모델의 RFECV 결과

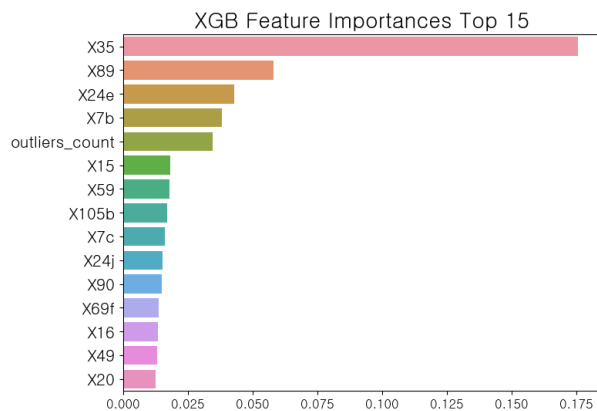


## 변수 선택 (Feature Selection)

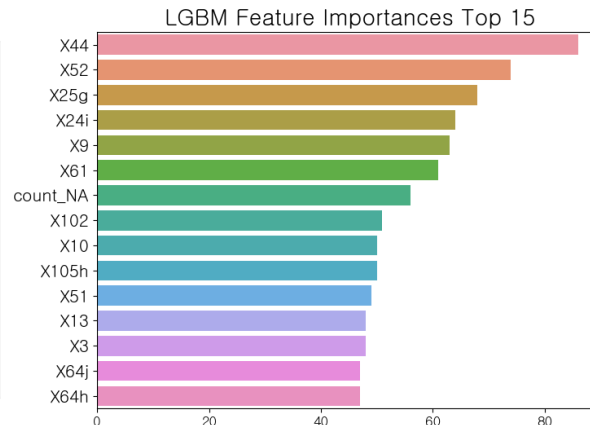
### 2. Feature Importance & Voting

LightGBM, XGBoost, RandomForest의 Feature Importance

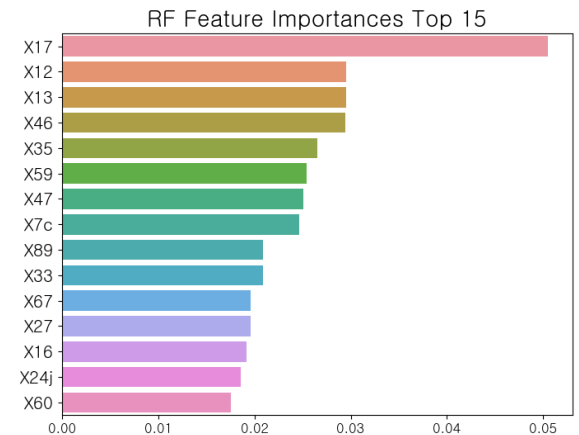
활용하여 **Voting** 진행



▲ XGB Feature Importance



▲ LGBM Feature Importance



▲ RF Feature Importance

## 변수 선택: Feature Importance & Voting

column	LGBM	XGB	RF	Voting
X1	0.002353	0.006914	0.007579	0.016846
X3	0.012272	0.002386	0.003007	0.017666
X7b	0.001971	0.001258	0.010822	0.014051
X7c	0.055422	0.005226	0.026379	0.087027
X7d	0.001503	0.003056	0.016704	0.021263
...	...	...	...	...
X105g	0.004321	0.002145	0.004265	0.010730
X105h	0.002807	0.003561	0.008648	0.015016
X105i	0.001601	0.002141	0.006743	0.010485
count_NA	0.001091	0.002198	0.002484	0.005773
outliers_count	0.012566	0.054539	0.005033	0.072139

모델 별 각 feature의 importance 값을 더하고,  
Voting 값을 기준으로 변수 선택을 시도

\* 각 모델별 importance 값의 크기가 달라 총합이 1이 되도록 조정



RFE와 Feature Importance & Voting 변수 선택법은  
유의미한 성능 향상에 기여하지 않아서 채택하지 않음...

## 변수 선택 (Feature Selection)

### 최종 변수 선택

KS test로 변수 5개 제외

Unique한 값이 1개 뿐인 파생변수 1개 제외

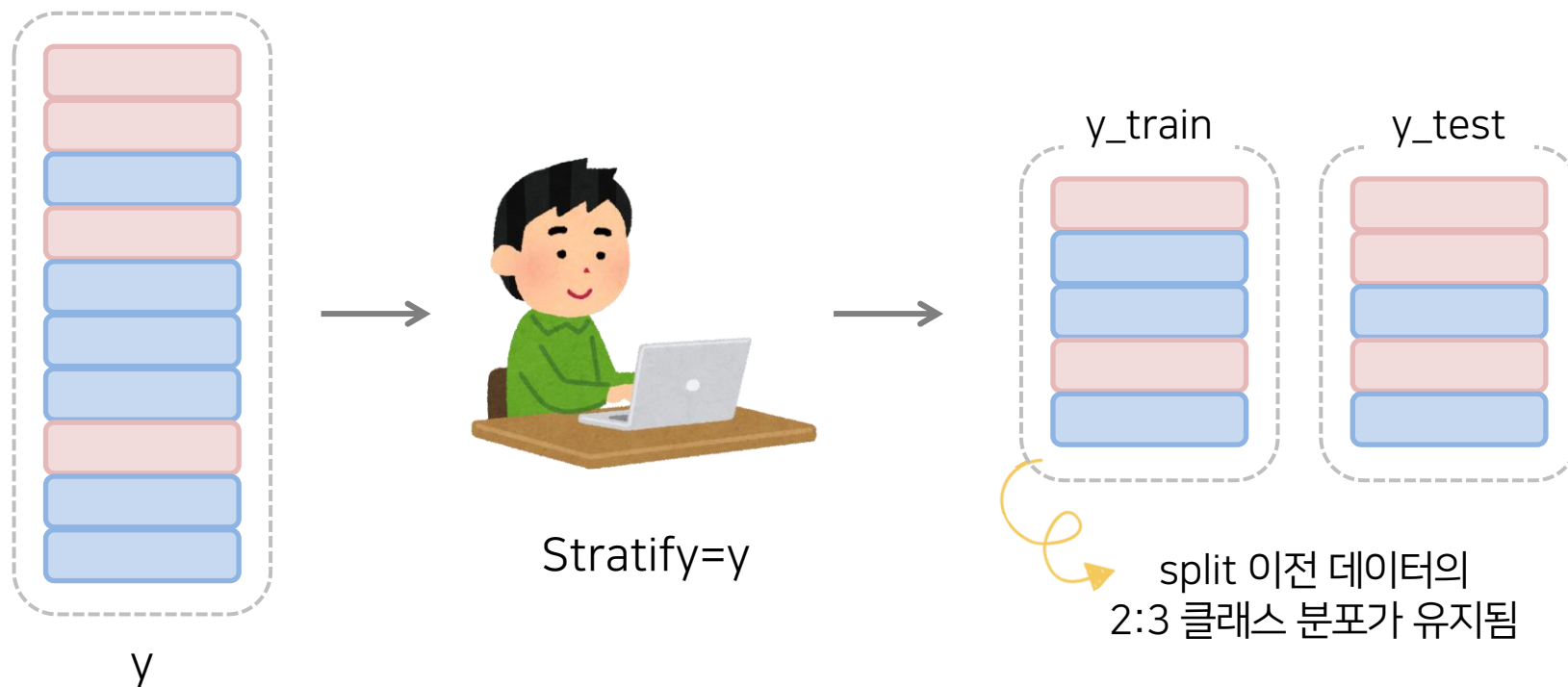
최종적으로 334 columns 활용해 모델링 ✨

# 4

모델링

## train\_test\_split *데마팀 1주차 클린업 참고!*

train\_test\_split() 함수에 stratify 파라미터를 적용하여  
train 데이터와 test 데이터의 **클래스 분포를 동일하게 분할**



## 학습에 사용한 분류 모델 *데마팀 2주차 클린업 참고!*

LogisticRegression

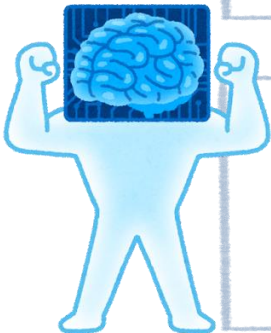
RandomForest

CatBoost

XGBoost

One-Class SVM  
(OCSVM)

LightGBM



## 학습에 사용한 분류 모델

### XGBoost

Decision Tree를 기본 학습기(Base Learner)로 가지며,  
여러 개의 학습기를 연결시켜 이전 모형의 약점을 보완하는 방향으로 학습  
기존의 Gradient Boosting 모델에 과적합 방지 파라미터를 추가한 모델



## 학습에 사용한 분류 모델

### XGBoost

Decision Tree를 기본 학습기(Base Learner)로 가지며,  
여러 개의 학습기를 연결시켜 이전 모형의 약점을 보완하는 방향으로 학습  
기존의 Gradient Boosting 모델에 과적합 방지 파라미터를 추가한 모델

### CatBoost

Level-wise로 트리를 적합하는 부스팅 계열의 모델

과적합 가능성이 적음

데이터 대부분이 범주형 변수인 경우 학습 속도가 빠름



범주형 파생변수를 많이 만들었기 때문에

분류 성능 향상에 도움이 될 수 있을 것 같아 시도해 봄!



## 학습에 사용한 분류 모델

### One-Class SVM (OCSVM)

하나의 클래스만 학습하여, 학습된 클래스와 학습되지 않은 나머지 클래스를 분류하는 최적의 support vector를 찾는 비지도학습 모델.

## 학습에 사용한 분류 모델

### One-Class SVM (OCSVM)

하나의 클래스만 학습하여, 학습된 클래스와 학습되지 않은 나머지 클래스를 분류하는 최적의 support vector를 찾는 비지도학습 모델.



Train data의 class0과 class1 불균형이 심하므로,  
OCSVM 모델에 class0 샘플들만을 학습시킨다면  
test data에서 class1의 특징을 이상치로서  
탐지(분류) 할 수 있을 것 같아 시도!

## 학습에 사용한 분류 모델

One-Class SVM (OCSVM)

하나의 클래스만 학습하여, 학습하지 않은 나머지 클래스를 분류하는 최적의 support vector를 찾는 비지도학습 모델.



Train data의 class0과 class1 불균형이 심하므로,

OCSVM 모델에 class0 샘플들만을 학습시킨다면

**하지만 XGBoost, CatBoost, OCSVM 모두**

**야심찬 마음과 다르게 높은 cost function 출력을 보여주었고 ...**

**돌고 돌아 가장 성능이 좋았던 LightGBM을 최종 모델로 선정!**

## 학습에 사용한 분류 모델 - LightGBM

### LightGBM

Decision Tree를 기반으로 한 Gradient Boosting 계열의 모델.  
Level-wise로 확장되는 XGBoost와 달리 **leaf-wise**로 모델이 확장됨.



자세한 내용은 여름학기 방학세미나 참고!

⋮

파라미터	파라미터 범위
Learning_rate	0.005, 0.01, 0.05, 0.1, 0.15, 0.25
max_depth	10, 11, 12, ..., 16, 17
reg_lambda	0.1, 0.3, 0.5, 0.7
min_data_in_leaf	10, 15, 20, 30, 40
n_estimators	100, 300, 500, 700

## LightGBM - 하이퍼파라미터 튜닝

가장 많이 이용되는 하이퍼파라미터 튜닝 방식을 모두 고려하였으나,

GridSearchCV

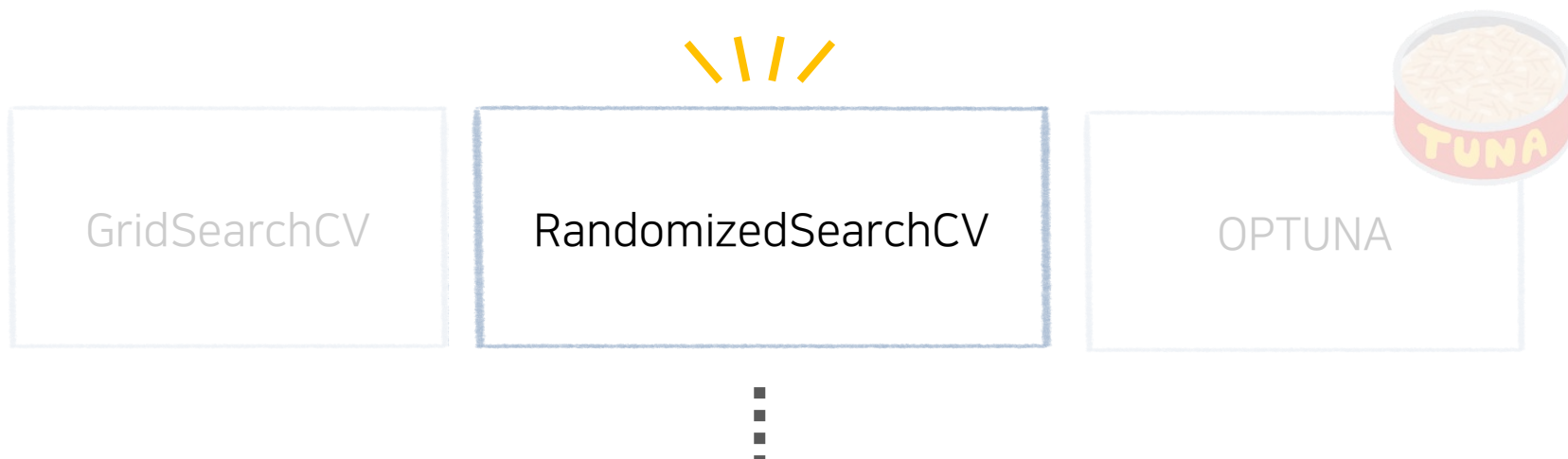
RandomizedSearchCV

OPTUNA



## LightGBM - 하이퍼파라미터 튜닝

가장 많이 이용되는 하이퍼파라미터 튜닝 방식을 모두 고려하였으나,



시간적인 효율을 고려하여 RandomizedSearchCV 방식을 채택!

## LightGBM - 하이퍼파라미터 튜닝

**Q. 랜덤하게 하이퍼파라미터를 탐색하면 모델의 성능이 떨어지지 않을까요?**

**A. 랜덤하게 탐색한 파라미터로 적합한 모델이 그리드 서치로 추정된 파라미터로 적합한 모델보다 성능이 떨어질 수는 있겠지만, 그 정도는 매우 작다!**  
**랜덤 서치에서 60개의 파라미터 조합이 있다면, 최적의 성능에서 5% 이내의 파라미터 조합을 찾을 가능성은 95%나 된다.**

시간적인 효율을 고려하여 RandomizedSearchCV 방식을 채택!



## LightGBM - 하이퍼파라미터 튜닝

RandomizedSearchCV의 scoring 파라미터로 Recall 지표를 사용.



$$\text{Recall (재현율)} = \frac{TP}{TP + FN}$$

## LightGBM - 하이퍼파라미터 튜닝

RandomizedSearchCV의 scoring 파라미터로 Recall 지표를 사용.



$$\text{Recall (재현율)} = \frac{TP}{TP + FN}$$

⋮

Cost function =  $250 \times FN + 5 \times FP$  인 만큼,

FN을 최소화하는 평가지표를 사용해야 한다고 판단하였기 때문!

## LightGBM - 하이퍼파라미터 튜닝

### learning\_rate

Gradient boosting 계열의 모델에서  
cost function 위를 이동하는 step의 크기를 조절하는 파라미터



Learning\_rate가 너무 크면 : cost가 커지는 방향으로 학습될 수 있음

learning\_rate가 너무 작으면 : 학습에 많은 시간이 소요됨

*여름방학 이론 스터디 1주차 참고!*

### max\_depth

Decision Tree학습기의 최대 깊이를 설정하는 파라미터.


max\_depth의 크기가 너무 크면 과적합될 위험이 있음

## LightGBM - 하이퍼파라미터 튜닝

*회귀분석팀 클린업 3주차 참고!*

reg\_lambda

모델의 L2-regularization을 설정하는 파라미터로, **모델의 과적합을 방지함**



L1-Regularization의 경우 reg\_alpha 파라미터를 이용하면 됨!

min\_data\_in\_leaf

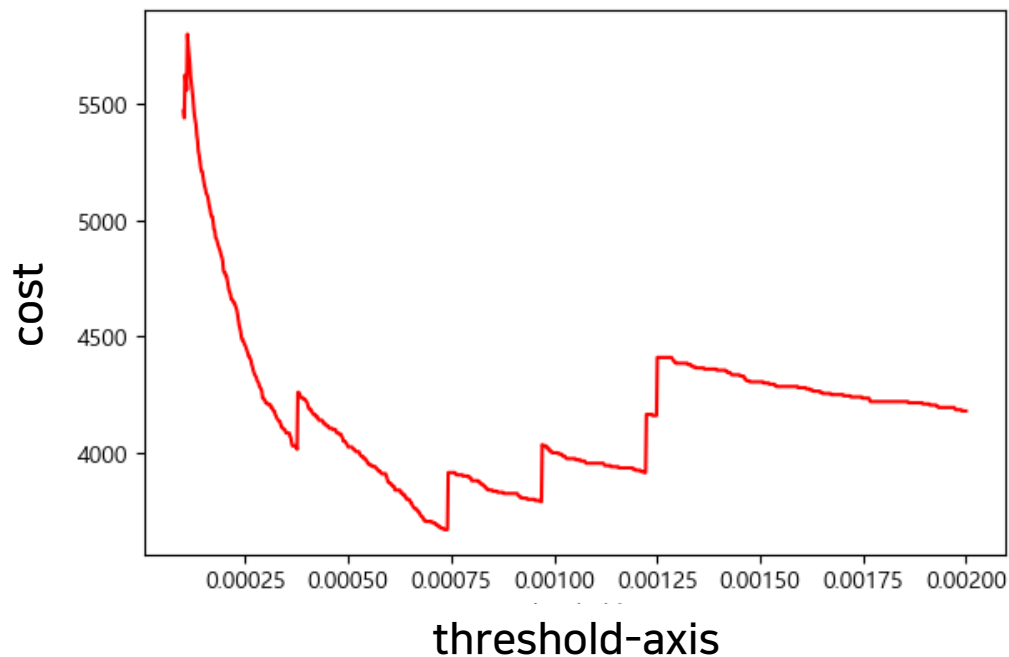
Decision Tree의 leaf가 분기를 중지할 수 있는 최소 샘플의 개수를 설정하는 파라미터로, **크기가 작을수록 모델이 과적합될 위험이 있음**

n\_estimators

기본 학습기의 개수를 설정하는 파라미터로,  
**크기가 클수록 모델이 과적합될 위험이 있음**

## LightGBM – threshold를 이용한 결과 보정

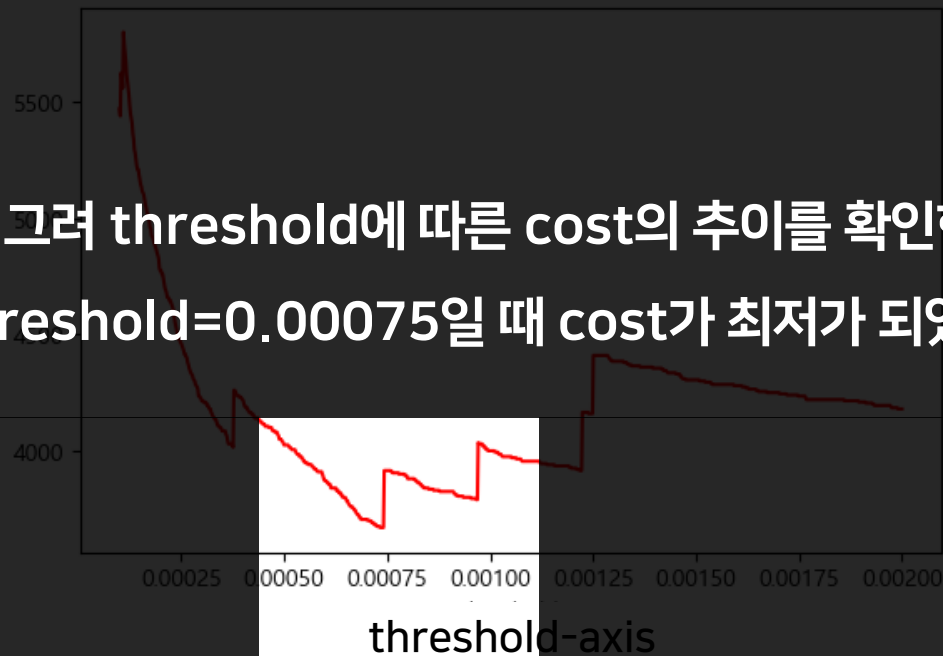
cost function 결과 최적화를 위하여 클래스 예측확률(predict\_proba())의 threshold를 조정해가며 cost function의 추이를 확인함



## LightGBM - threshold를 이용한 결과 보정

cost function 결과 최적화를 위하여 클래스 예측확률(predict\_proba())의 threshold를 조정해가며 cost function의 추이를 확인함

Plot을 그려 threshold에 따른 cost의 추이를 확인한 결과,  
Threshold=0.00075일 때 cost가 최저가 되었음






## LightGBM - 하이퍼파라미터 튜닝 결과


파라미터	파라미터 튜닝 결과
Learning_rate	0.05
max_depth	10
reg_lambda	0.1
min_data_in_leaf	15
n_estimators	300

최적의 파라미터와 threshold를 이용해 최종 predict를 진행해봅시다!

## 최종 캐글 결과

#	Team	Members	Score
1	2팀_공유 😊🌟		3150.00000

#	△	Team	Members	Score
1	—	2팀_공유 😊🌟		2535.00000

Public (3150) & Private (2535) 모두 1등 달성!







## 소감



방세하느라 정말 빠르게 흘렀던 5일... 성능이 잘 나오지 않아서 힘들었던 시간도 있었지만  
너무너무 똑똑했던 팀원들 덕분에 목표했던 성능 달성해서 너무 기뻐어요!!!  
끝까지 최선을 다해서 마무리까지 잘 한 것 같아서 아주 만족스럽습니다 ^\_^  
짧은 기간이었지만 정말 많이 배웠고, 스스로 성장한 기분(?)입니다.  
함께했던 2팀 팀원들 너무너무 수고 많았고 고마웠고 남은 피셋 1학기도 재밌게 보내자!!

다들 덩고 바쁜 와중에도 정말 고생 많았습니다!  
짧은 기간이지만 다 같이 괜찮은 결과를 낼 수 있었던 것 같아  
다행이라고 생각합니다ㅎㅎ 어려운 데이터였지만 팀원들이 잘 도와준  
덕분에 재밌게 분석하면서 많이 배울 수 있던 시간이었습니다. 다른  
팀들도 너무 고생많으셨고 2학기도 다 같이 힘내서 화이팅합시다!!





## 소감



선배들에게 말로만 듣던 무시무시한 방학세미나를 드디어 마무리했네요!  
짧은 시간이었지만 능력자 팀원들 덕분에 너무 든든하고 많이 배웠고  
많이 먹었습니다(?) 귀소본능 2팀 빨리 집어가자 !!! ☺

EDA부터 전처리, 모델링까지... 2023 가장 피곤한 일주일이었지만  
짧은 기간에 정말 많은 것을 배울 수 있었습니다. 처음 시작할 때는  
데이터 정보가 적어서 막막했는데 브레인 2팀 덕분에 좋은 결과를 낼 수 있었습니다!  
마음 잘 맞는 말랑한 팀원들 만나서 5일 동안 너무 즐거웠습니다 ㅎㅎ  
남은 한 학기도 화이팅합시다~~~~♡



주제분석을 5일 만에 끝낸 기분입니다. 처음에는 막막했지만 훌륭한 팀원들과 함께  
해서 잘 마무리하지 않았나 싶습니다. 2팀 덕분에 많이 배우고 성장한거 같습니다.  
모두들 피곤할텐데 최선을 다해주셔서 고맙고 덕분에 좋은 추억을 하나 만드느거  
같습니다. 2팀 화이팅~



지금까지 2팀이었습니다~!





THANK YOU

