2021 방학세미나

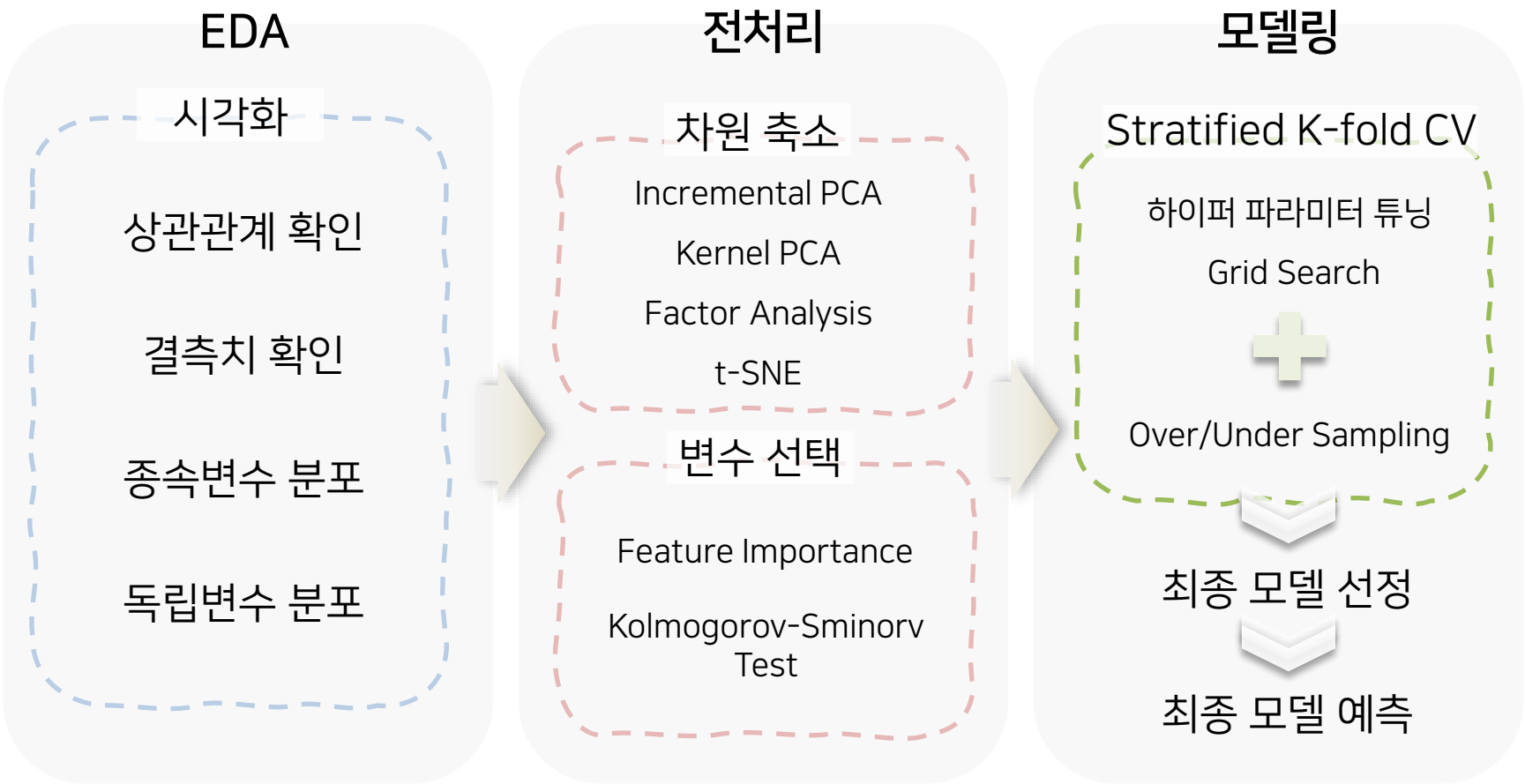
3팀

이승우
고경현
김지현
박세령
임주은

INDEX

0 시작하기 전에

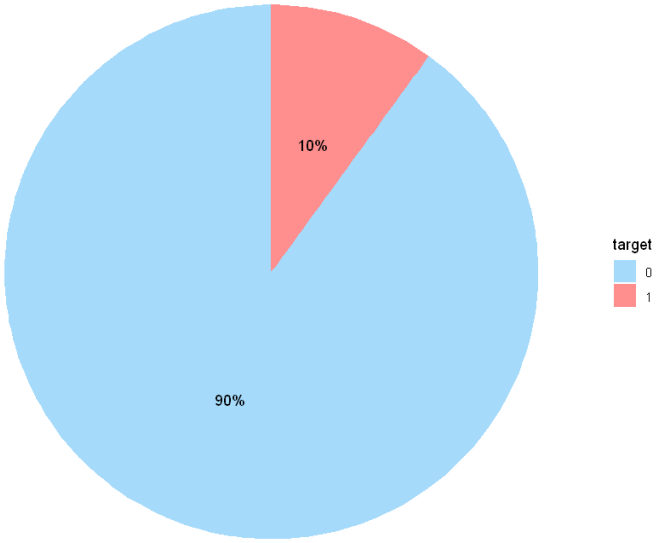
분석 흐름도



1

EDA

클래스 불균형



변수는 9:1
불균형의 분포를
이루고 있음

NA (결측치) 확인

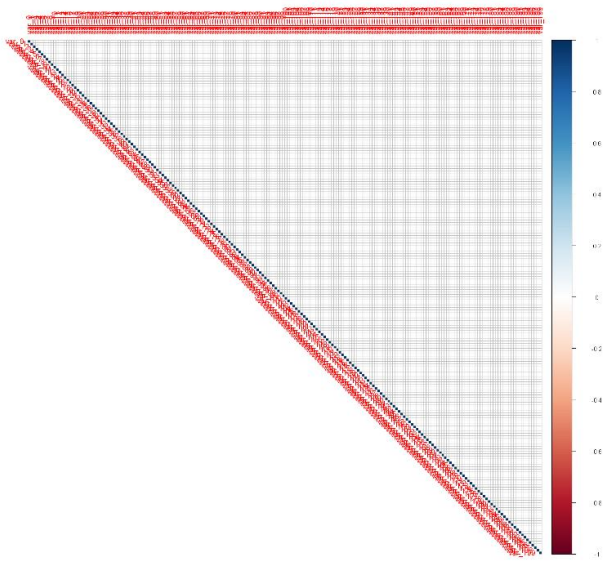
target	var_0	var_1	...
0	5.0702	-0.5447	
1	16.3699	1.5934	

⋮

28001 rows X 201 cols

“ NA가 존재하지 않아 결측치 처리를 해주지 않아도 됨 ”

상관관계 존재 여부



Plot을 그려본 후,

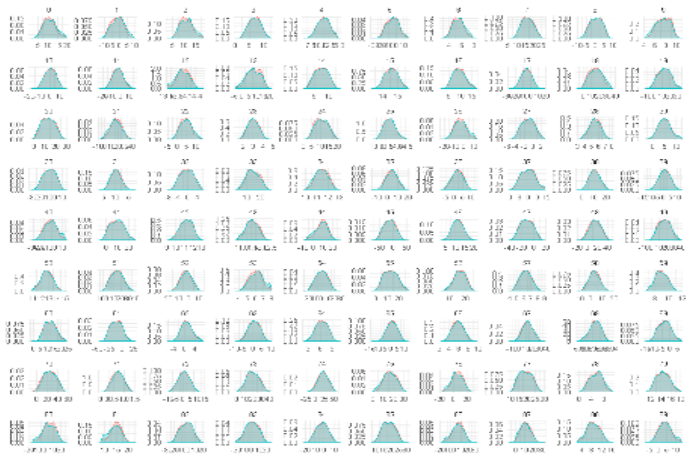
상관관계가 존재하지 않음

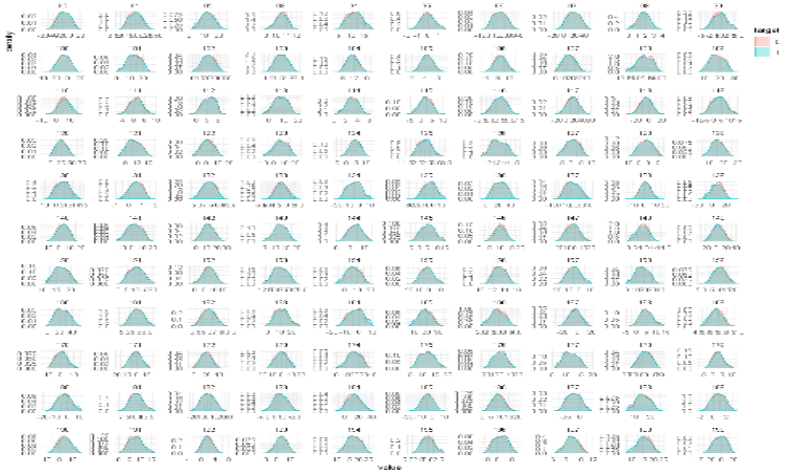
확인

특이사항 발견

1) Target 별 변수 분포 확인에서 특이사항 발견

타겟 클래스에 따라 독립변수의 분포를 시각화





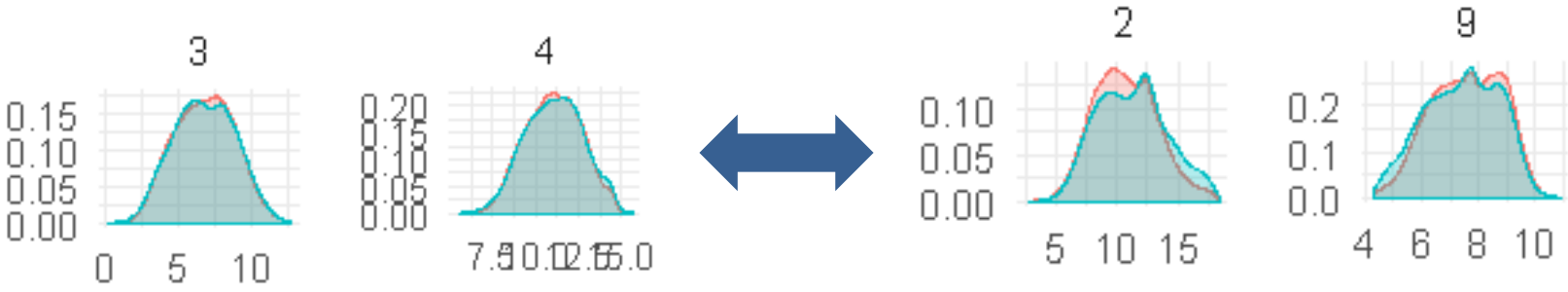
target 값과 상관없이 유사한 분포를 갖는 변수가 존재한다!

유사한 분포를 갖는 변수는 분류에 영향을 미치지 않을 것이라고 생각

특이사항 발견

1) Target 별 변수 분포 확인에서 특이사항 발견

타겟 클래스에 따라 독립변수의 분포를 시각화



target 값과 상관없이 유사한 분포를 갖는 변수가 존재한다!

유사한 분포를 갖는 변수는 분류에 영향을 미치지 않을 것이라고 생각

특이사항 발견

```
res = {}
for i in train:
    res[i] = train[i].value_counts().max()
```

```
dict(sorted(res.items(), key=lambda item: item[1], reverse=True))
```

	count
var_68	176
var_108	60
var_126	47
var_12	27
var_91	16
var_117	3
var_120	3
var_136	3
var_149	3
var_187	3

빈도 수가 높은 변수들을 발견

	var_108_count
14 1999	60
14 2003	43
14 2000	40
14 2001	39
14 2002	36
14 2005	35
14 1994	33
14 2007	31
14 1997	30
14 1996	30
14 1998	30
14 2004	28
14 1992	25
14 1990	23
14 2008	22
14 1995	19
14 1993	19
14 1987	17
14 2010	17
14 1991	16

이 변수들.. 연도가 아닌가..?
하는 의심을 하게 되었습니다 ..

2

전처리

Incremental PCA (IPCA)

학습 데이터넷을 미니배치로 나눈 뒤, IPCA 알고리즘에 주입하는 방식

Kernel PCA (KPCA)

다변량 자료를 저차원의 비선형적 공간에 시각화하는 방식

Factor Analysis

변수들 간의 상관관계를 고려해 저변에 내재된 요인들을 추출해내는 방식

t-SNE

높은 차원의 복잡한 데이터를 2차원에 차원 축소하는 방식

Incremental PCA (IPCA)

학습 데이터넷을 미니배치로 나눈 뒤, IPCA 알고리즘에 주입하는 방식

Kernel PCA (KPCA)

다변량 자료를 저차원의 비선형적 공간에 사상화

Factor Analysis

변수들 간의 상관관계를 고려해 저변에 내재된 요

t-SNE

높은 차원의 복잡한 데이터를 2차원에 차원 축소

변수 간 상관관계가 존재하지 않아서인지

4가지 차원축소 방법

모두 실패 ㅜㅜ

Feature Selection

유의미한 변수를 알아보고자 Lgbm. feature_importances를 사용

```
lgbm = lightgbm.LGBMClassifier(is_unbalance=True)
lgbm.fit(X_train, y_train)
```

```
imp = sorted(zip(lgbm.feature_importances_,X_train.columns))
feature_imp = pd.DataFrame(imp, columns=['Value','Feature'])
```

```
feature_imp.sort_values('Value')
```

Feature	Value
var_187	0
var_14	1
var_61	1
...	...
var_94	40
var_53	43

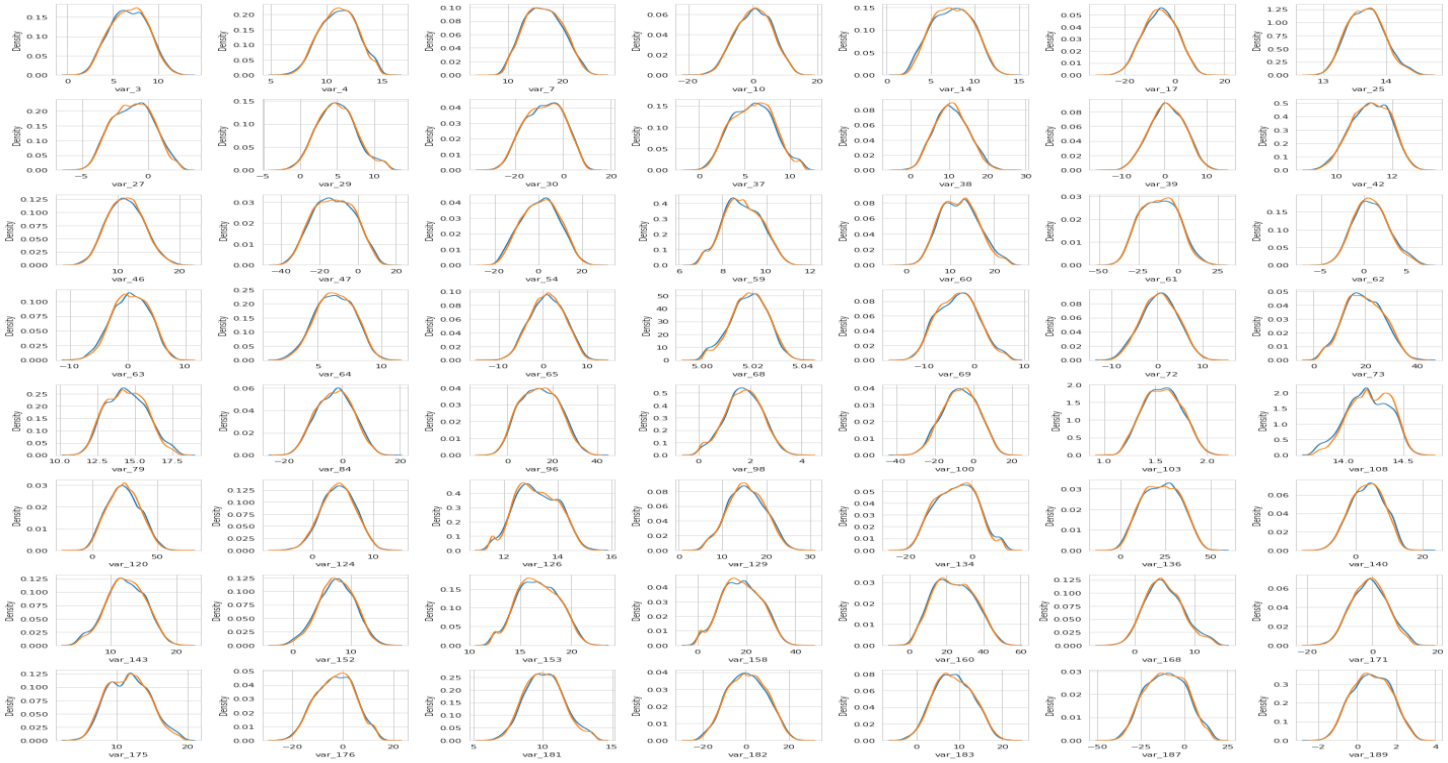
Target 값에 따른 변수 분포도를 확인하여
Feature를 줄이려는 계획!

Feature Selection

target 값이 0 일 때와 1 일 때 분포가 다른 변수.

분포가 다른 변수는 대체로 feature_importance 결과에서 높은 값을 기록

Feature Selection



target 값이 0 일 때와 1 일 때 분포가 같은 변수.

분포가 같은 변수는 대체로 feature_importance 결과에서 낮은 값을 기록

Feature Selection

KS-test로 분포 동질성 검사를 한 결과

동일한 분포를 갖는 54개의 변수는 제외하기로 함!

target	var_0	var_1	var_2	var_5	…	var_199
0	5.0702	-0.5447	9.5900	-18.8687	…	-7.6652
1	16.3699	1.5934	16.7395	5.9004	…	-10.8529
0	5.0615	0.2689	15.1325	-6.5477	…	11.1524

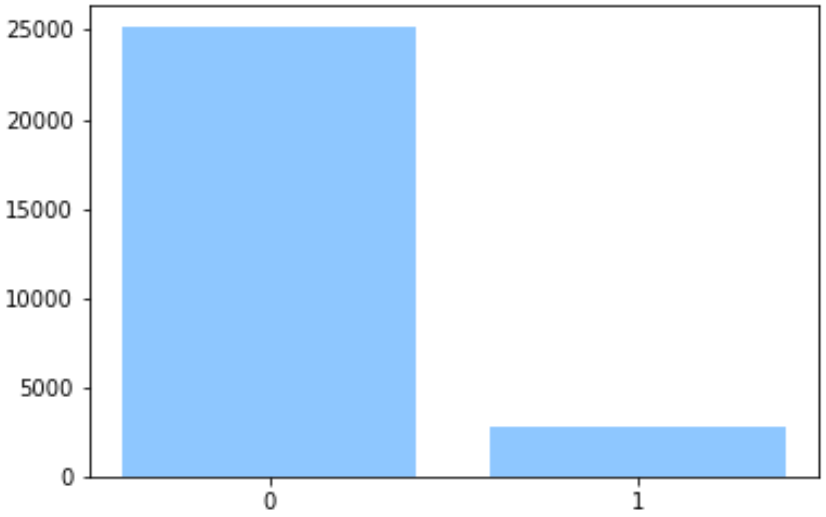
28000 X 146

Data set 확보

3

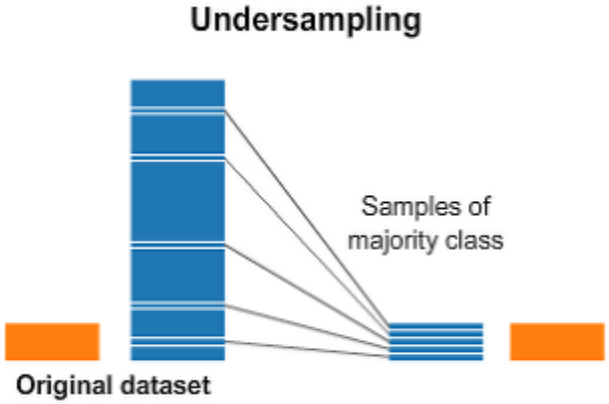
모델링

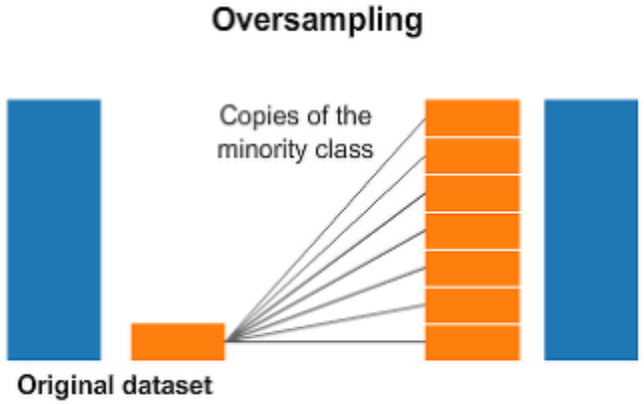
클래스 불균형 처리



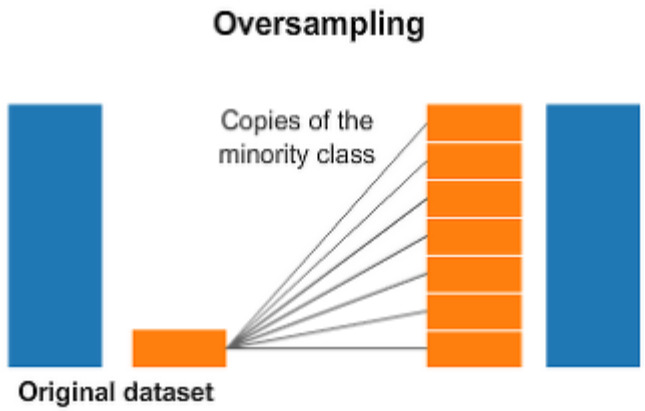
기존 train data는 targe의 비율이 약 9대1의 imbalanced data

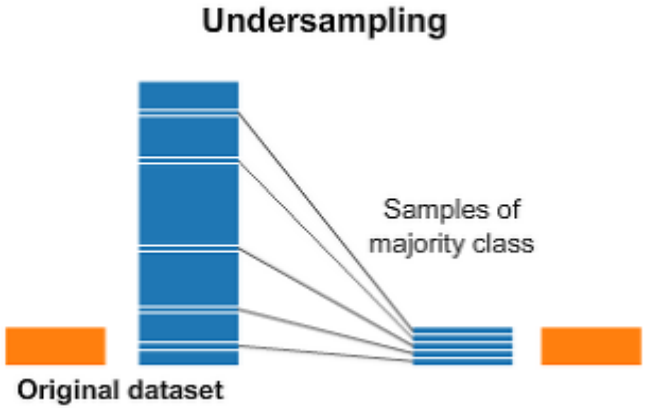
불균형 상태의 모델링으로 인한 편향된 학습을 피하고자 Sampling 방법 도입





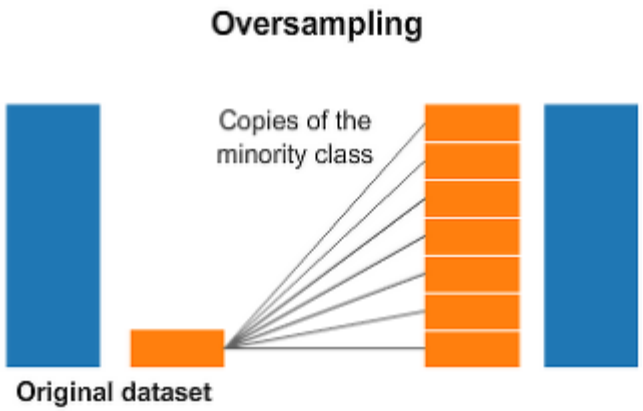
클래스 불균형 처리

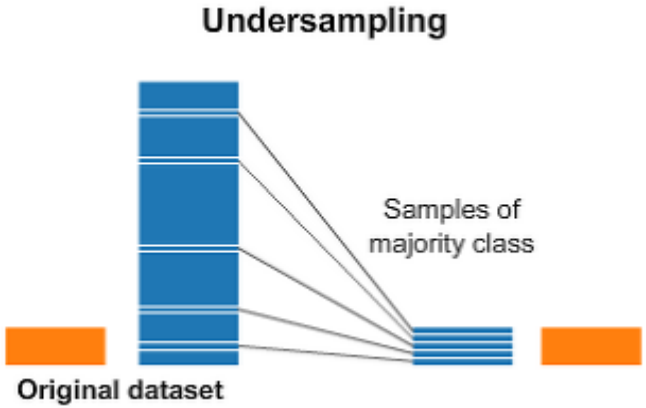




- SMOTE
- SVMSMOTE
- BorderlineSMOTE
- ADASYN
- KmeansSMOTE
- Random undersampling
- NearMiss
- Edited Nearest Neighbors
- One-sided selection
- Neighborhood cleaning rule

클래스 불균형 처리





- SMOTE
- SVMSMOTE
- BorderlineSMOTE
- ADASYN
- KmeansSMOTE
- Random undersampling
- NearMiss
- Edited Nearest Neighbors
- One-sided selection
- Neighborhood cleaning rule

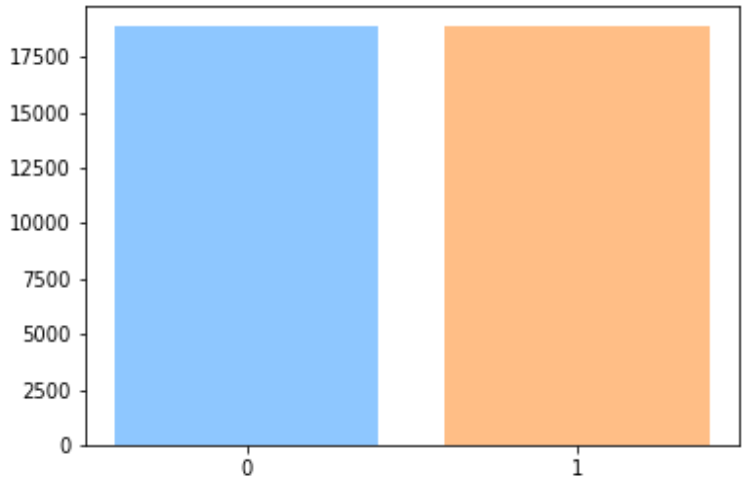
클래스 불균형 처리

SVM SMOTE

1.소수 클래스 포인트를 노이즈 포인트와 경계 포인트로 분류

2. 노이즈 포인트는 무시하고, 경계 포인트 데이터 하나를 선택

3. 선택된 데이터의 **k- 최근 접 이웃** 알고리즘을 사용하여 합성 데이터를 생성



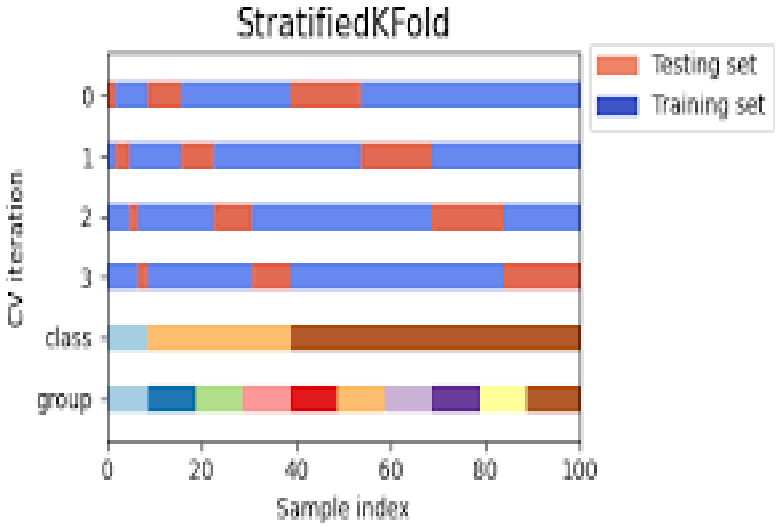
기존 train을 train_test_split으로 나눈 뒤 SVMSOTE를 적용하여 불균형을 해결

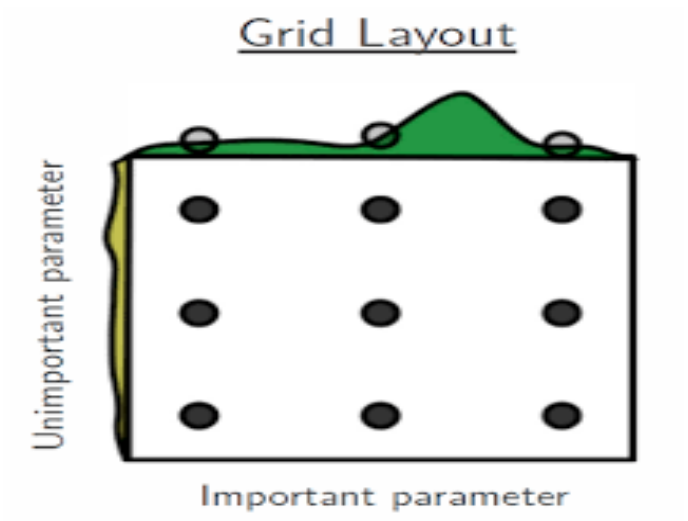
총 37,786 행의 train data 생성

파라미터 튜닝 (Stratified Kfold CV + Grid Search)

Stratified Kfold CV(계층별 교차 검증)

: 폴드 안의 클래스 비율이 전체 데이터 셋의 클래스 비율과 같도록 해준 뒤 교차 검증을 해주어 불균형 데이터셋에 유용한 교차검증





모델링

LDA

LinearDiscriminantAnalysis

: LDA는 PCA와 유사하게
데이터 셋의 차원을 축소하는 기법이지만,
PCA와 다르게
지도학습의 분류(Classification)에서 사용됨

parameter	value
n_components	0
priors	[0.7, 0.3]
Shirinkage	'auto'
Solver	'eigen'
Store_covariance	True
'tol'	1e-05

모델링

LGBM

LightGradientBoosting Model

: leaf-wise 트리분할을사용하는 Gradient boosting 모델로 속도가 빠른 장점이 있음

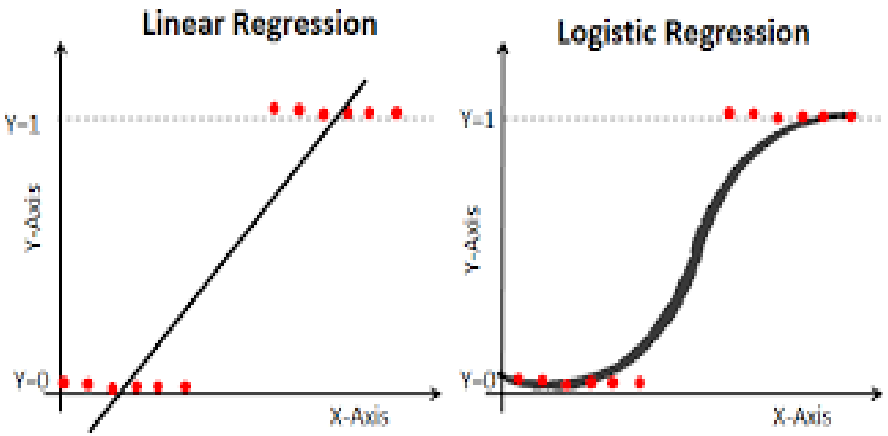
parameter	value
Is_unbalanced	True
learning rate	0.1
boosting	‘gbdt’
objective	‘binary’
num_leaves	31
max_depth	-1

모델링

LR

LogisticRegression

: 종속변수가 이진분류일 때 수행하는
회귀분석 모델



parameter	value
warm_start	False
tol	0.00045
penalty	L1
max_iter	300
intercept_scaling	1.5
fit_intercept	True
dual	False
class_weight	Balanced
C	0.5

모델링

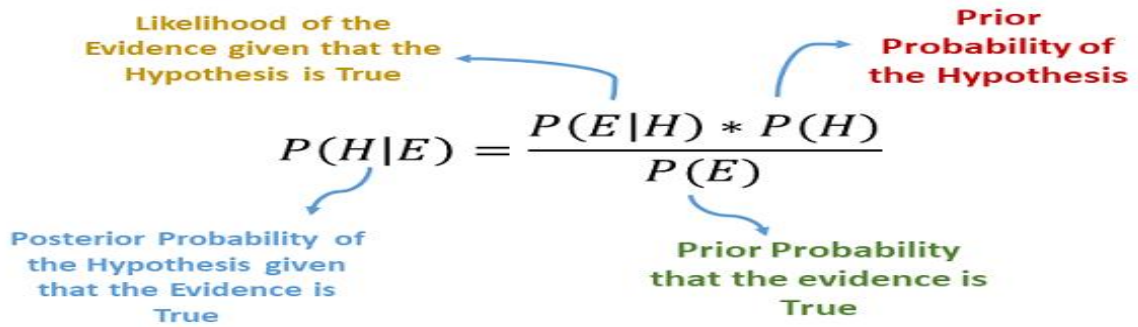
GNB

GaussianNaiveBayes

: Gaussian 정규분포를 따르고
독립변수가 연속 데이터일 때
사용하는 Naïve Bayes의 변형

parameter	value
priors	[0.8,0.2]
var_smoothing	1e-09

표본평균과 표본 분산을 가진
정규분포 하에서 베이즈 정리를 사용한 것!



4개의 모델 비교

<LightGBM>

F1 SCORE = 0.42654

<LDA>

F1 SCORE = 0.43404

<SVMsmote + Logistic Regression>

F1 SCORE = 0.35967

<Gaussian Naïve Bayes>

F1 SCORE = 0.49462

4

결론 및 의의와 한계

<최종결론>

Feature importance및 Kstest로 유사한 분포인 변수 삭제 → SVMsmote (Logistic Regression만) → Modeling

<Kaggle 결과>

F1 score = 0.49462

< 의의 >

1. FANCY 한 모델이라고 성능을 보장하는 것이 아님을 알게 되었다..
2. PCA 뿐 아니라, Feature importance와 Kstest를 통해 변수를 제거 하는 방법을 시도해 보았다.
3. 모델마다 좋은 성능을 발휘하는 oversampling의 기법이 다른 것을 알게 되었다.
4. 불균형 데이터를 가진 경우 교차검증과 그리드 서치를 진행하는 방법을 배울 수 있었다.

< 한계 >

1. 마의 0.5를 넘지 못하였다.. ㅠ^ㅠ
2. Unique값을 보았을 때, var108변수에서 유독 연도처럼 보이는 데이터들이 있었으나, 변수에 대한 정보가 없어 그 의미를 끝내 알 수는 없었다…

THANK YOU