최재영

**[1] 데이터셋의 크기**

- TRAIN: X: (10000,10), y :(10000,1)

- TEST: X: (4000,10), y: (4000,1)

: 5차원\*2차원으로 펼쳐진 데이터로 이해됩니다.

5차원\*2차원을 펼쳐놓은 형태로 10개의 변수로서 활용할 수 있는 데이터입니다.

**[2] 데이터셋의 분포**

- x 변수의 분포

Calendar

Description automatically generated

: 변수에 따라 분포해 있는 y를 관찰할 수 있다. (파랑:0 / 주황: 1 ).

: 0,1이 명확하게 구분되는 변수가 있으나 대부분 중첩되었다.

- Train y의 분포 : 0으로 치우친 불균형한 데이터입니다.

Chart, bar chart

Description automatically generated

**[3] 분석 진행 – 불균형의 해결**

- 불균형한 데이터를 해결하기 위해 Algorithm Over Sampling기법을 사용합니다.

(1)SMOTE : Knn을 기반으로 적은 샘플을 oversampling 하는 기법입니다.

Chart, bar chart

Description automatically generated

(2) ADASYN: SMOTE의 개선된 방안으로 노이즈를 섞어 Oversampling을 진행하는 기법입니다.

Chart, bar chart

Description automatically generated

: 두 기법 모두 9000개에 인접한 균형된 데이터를 만들었습니다.

**[4] 모델링**

(1) 최적 데이터 선정 : train 데이터를 train/val로 나누어 진행했습니다. 모델링에 사용한 데이터는 original data, smote data, adasyn data, 총 3가지의 데이터입니다. Pipeline을 통해 scaling을 진행한 다음 이들의 classification score을 산출했습니다

Text

Description automatically generated. Text

Description automatically generatedText

Description automatically generated

Validation set의 classification score을 기준으로 smote 데이터가 recall, precision,f1 score이 우수하단 점을 이용해 **SMOTE 데이터를 선정해 분석**을 진행했습니다.

(2) 최적의 모델 선정 : SGDClassifier, svm, logisticRegression, Ridge Classifier 그리고 random forest classifier 중에서 그리드 서치를 이용해 가장 우수한 validation score을 산출했습니다. 이중 best estimator인 **random forest를 선정**했습니다.

(3) 파라미터 조정: Random Forest의 파라미터 탐색을 통해 최적의 모델/파라미터를 학습했습니다. 이때의 **validation score**은 아래로 산출되었습니다.

Text

Description automatically generated

(4) TEST score 산출: smote의 전체 데이터를 활용해 학습을 진행한 다음 test data를 예측했습니다. 아래의 score은 **test data 성능**입니다.

Text

Description automatically generated

(+) 추가 분석 : Feature importance를 통해 7번 피처가 중요하게 작용했음을 알 수 있습니다.

Chart, bar chart

Description automatically generated