[머신러닝] Kaggle 심장질환 예측

요약

일반적으로 불균형 데이터로 머신러닝을 진행하면 소수 계층에 대한 성능은 무시되고 저하되며 예측 모델은 부정확하게 편향될 수 있다. 이번 프로젝트를 통해 불균형 데이터 세트를 해결하기 위한 접근 방식으로 STMOE를 이용했고 Light GBM, LogisticRegression, KNN, DecisionTree, XGBoost, RandomForest 알고리즘을 통해 확인하였다. 그 결과, XGBoost가 90% 이상의 정확도와 정밀도, F1-score라는 가장 높은 점수를 보였다. 또한 LightGBM은 모든 점수에서 89%라는 수치를 보였으며, KNN은 모든 항목에 대해 80% 이상의 점수를 얻었다. 특히, 재현율은 94%라는 가장 높은 점수를 얻었다.

서론

 데이터가 불균형을 이루는 상태로 학습 데이터가 높은 성능을 보였다고 해서 예측 성능이 반드시 좋은 것은 아니다. 일반적으로 이러한 데이터 불균형 문제는 금융사고 또는 질병 식별 등 이상 감지가 중요한 데이터에 다양하게 나타난다. 본 프로젝트에 사용된 데이터는 Kaggle의 Heart Disease Dataset을 사용하였다.

불균형 데이터를 데이터 샘플링 기법을 통해 균형 있는 데이터 집합으로 만들 수 있다. 다수의 샘플을 제거하는 언더 샘플링과 소수의 샘플을 다수의 샘플에 맞춰 생성하는 방법이 있다. 언더 샘플링은 학습 가능한 전체 데이터 수를 감소시키고, 분류에 중요한 데이터를 학습 데이터에서 배제하여 성능 저하를 일으킬 수 있다. 때문에 이번 프로젝트에서는 데이터 정보의 손실을 최소화 하기 위해 오버샘플링 기법 중 하나인 SMOTE를 활용할 예정이다.

또한, 데이터는 자료형 데이터와 수치형 데이터가 혼합되어 있다. 때문에 자료형 데이터는 원 핫 인코딩을 통해 처리하고 수치형 데이터는 표준화를 통해 처리할 예정이다.

SMOTE는 데이터 세트의 사례 수를 균형 있게 늘릴 수 있는 통계적 기법이다. 다수의 클래스를 변경하지 않고 구현하며 소수의 클래스를 기반으로 새로운 데이터를 생성한다. 단순 무작위 생성하는 것에 비해 과적합 발생 가능성은 낮다.

본론

Light GBM은 100개 이상의 매개변수를 포함한다. 때문에 파라미터 튜닝이 필수적이다. Tree모델은 num_leaves의 값이 2^max_depth값 보다 같거나 적어야 한다. 이것보다 많은 값은 과적합을 유발할 수 있다.

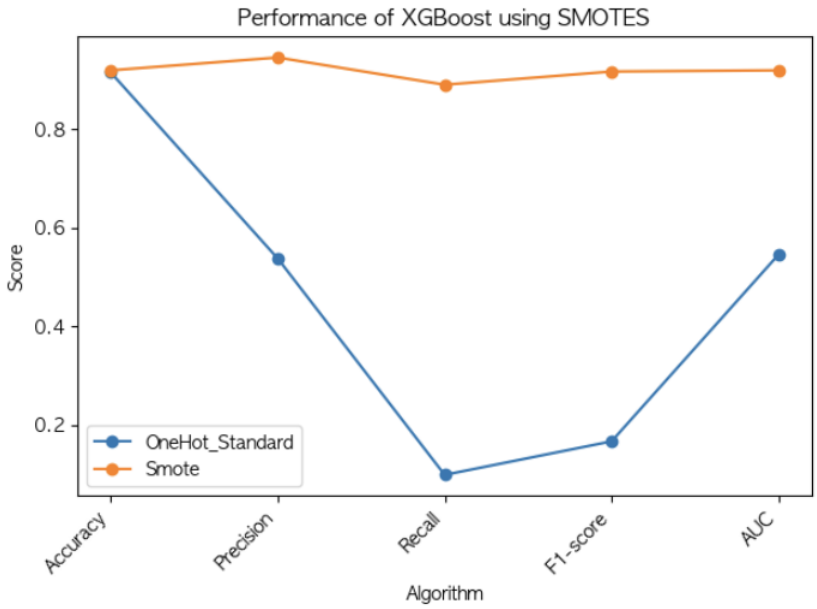
<그림 1> 은 원핫인코딩과 표준화만 적용한 데이터를 활용한 머신러닝 개별 성능 결과이고 <그림 2> 는 원핫 인코딩과 표준화를 적용한 데이터에서 SMOTE를 활용한 머신러닝 개별 성능 결과이다.

Algorithm	Accuracy	Precision	Recall	F1-score	AUC
LightGBM	0.914596	0.557978	0.081066	0.141564	0.537478
LogisticRegression	0.914233	0.533846	0.099928	0.168344	0.770710
KNNClassification	0.905140	0.361748	0.120374	0.180640	0.550085
DecisionTreeClassifier	0.913595	0.546599	0.031246	0.059112	0.514390
XGBoost	0.914320	0.537255	0.098632	0.166667	0.545275
RandomForestClassifier	0.914095	0.725146	0.017855	0.034851	0.508605

Algorithm	Accuracy	Precision	Recall	F1-score	AUC
LightGBM	0.892785	0.894503	0.890587	0.892540	0.892785
LogisticRegression	0.770708	0.757054	0.797207	0.776612	0.770710
KNNClassification	0.863034	0.810212	0.948139	0.873766	0.863041
DecisionTreeClassifier	0.749513	0.753154	0.742254	0.747664	0.749512
XGBoost	0.918905	0.944866	0.889711	0.916459	0.918902
RandomForestClassifier	0.773020	0.787451	0.747862	0.767146	0.773018

그림 1 그림 2

Light GBM, LogisticRegression, KNN, DecisionTree, RandomForest 모델 모두 정확도가 감소했으며 XGBoost는 아주 약간 증가했다. 또한, 정확도와 정밀도, F1-score은 모두 증가했다. AUC는 LogisicRegression은 유지되었고, 나머지 알고리즘에서는 20 ~ 30% 증가하였다. XGBoost가 전체적으로 우수한 성능을 보여주었다.



XGBoost 모델. SMOTE적용 전 후 점수 비교

결론

프로젝트를 통해 SMOTE를 이용하여 XGBoost 모델이 다른 알고리즘 보다 데이터 불균형 해결에 평균적으로 향상됨을 확인할 수 있었다.

참조

본론의 Light GBM 내용 :
https://scholarworks.bwise.kr/hanyang/bitstream/2021.sw.hanyang/182260/1/KCI_FI002906733.pdf

http://www.incodom.kr/SMOTE

데이터 : https://www.kaggle.com/datasets/abubakarsiddiquemahi/heart-disease-dataset