[4주차 예습과제]
09. 분류실습 - 캐글 산탄데르 고객 만족 예측
https://www.kaggle.com/c/santander-customer-satisfaction/data

산탄데르 고객 만족 예측 분석은 370개의 피처로 주어진 데이터 세트 기반에서 고객 만족 여부를 예측하는 것임.
클래스 레이블 명은 TARGET이며, 이 값이 1이면 불만을 가진 고객, 0이면 만족한 고객임.

cf) describe() 메서드를 이용해 각 피처의 값 분포를 살펴 보았을 때,
*var3 칼럼의 경우 min 값이 -999999임. NAN이나 특정 예외 값을 변환시킨 것일 거임. -> 편차가 너무 심하므로 가장 값이 많은 2로 변환함.

cf) ROC-AUC의 값이 하이퍼 파라미터 튜닝 이후 개선됨. 시간을 투자한 것 만큼 개선된 건 아니지만, 캐글과 같이 순위 경쟁이 필요한 경우에는 이 정도의 수치 개선은 도움이 될 거임!!
-> XGBoost가 GBM 보다는 빠르지만, 아무래도 GBM을 기반으로 하고 있기 때문에 수행 시간이 상당히 더 많이 요구됨.
앙상블 계열 알고리즘에서 하이퍼 파라미터 튜닝으로 성능 수치 개선이 급격하게 되는 경우는 많지 않음. (앙상블 계열 알고리즘은 과적합이나 잡음에 기본적으로 뛰어난 알고리즘이기에 그러함!)

LightGBM
*LightGBM을 수행해보면, XGBppst보다 학습에 걸리는 시간이 좀 더 단축됐음을 느낄 수 있음! 소스 코드는 XGBoost와 크게 다르지 않으며, 단지 LGBMClassifier 객체를 생성하는 부분만 달라짐.

*fmin()을 호출하여 최적 하이퍼 파라미터를 도출할 수 있음.

10. 분류 실습 - 캐글 신용카드 사기 검출
https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

해당 데이터 세트의 레이블인 Class 속성은 매우 불균형한 분포를 가지고 있음.
Class는 0과 1로 분류되는데 0이 사기가 아닌 정상적인 신용카드 트랜잭션 데이터, 1은 신용카드 사기 트랜잭션을 의미함. 전체 데이터의 약 0.172%만이 레이블 값이 1, 즉 사기 트랜잭션임.
-> 일반적으로 사기 검출이나 검출과 같은 데이터 세트는 레이블 값이 극도로 불균형한 분포를 가지기 쉬움. 그 이유는 사기와 같은 이상 현상은 전체 데이터에서 차지하는 비중이 매우 적을 수밖에 없기 때문임.

레이블이 불균형한 분포르 ㄹ가진 데이터 세트를 학습시킬 때 예측 성능의 문제가 발생할 수 있는데, 이는 이상 레이블을 가지는 데이터 건수가 정상 레이블을 가진 데이터 건수에 비해

너무 적기 때문에 발생함. 즉, 이상 레이블을 가지는 데이터 건수는 매우 적기 때문에 제대로 다양한 유형을 학습하지 못하는 반면에 정상 레이블을 가지는 데이터 건수는 매우 많기 때문에 일방적으로 정상 레이블로 치우친 학습을 수행해 제대로 된 이상 데이터 검출이 어려워지기 쉬움.

*지도 학습에서 극도로 불균형한 레이블 값 분포로 인한 문제점을 해겨하기 위해서는 적절한 학습 데이터를 확보하는 방안이 필요한데, 대표적으로 오버 샘플링과 언더 샘플링 방법이 있음.

언더 샘플링 : 많은 데이터 세트를 적은 데이터 세트 수준으로 감소시키는 방식임. 이렇게 정상 레이블 데이터를 이상 레이블 데이터 수준으로 줄여 버린 상태에서 핚븡르 수행하면 과도하게 정상 레이블로 학습/예측하는 부작용을 개선할 수 있지만, 너무 많은 레이블 데이터를 감소시켜서 정상 레이블의 경우 제대로 된 학습을 수행할 수 없는 문제가 발생할 수도 있음.

오버 샘플링 : 이상 데이터와 같이 적은 데이터 세트를 증식하여 학습을 위한 충분한 데이터를 확보하는 방법임. 동일한 데이터를 단순히 증식하는 방법은 과적합이 되기 때문에 의미가 없으므로 원본 데이터의 피처 값들을 아주 약간만 변경하여 증식함. 대표적으로 SMOTE 방법이 있음.

SMOTE 적은 데이터 세트에 있는 개별 데이터들의 K 최근접 이웃을 찾아서 이 데이터와 K개 이웃들의 차이를 일정 값으로 만들어서 기존 데이터와 약간 차이가 나는 새로운 데이터들을 생성하는 방식임.
-> imbalanced - learn

로지스틱 회귀는 선형 모델임. 대부분의 선형 모델은 중요 피처들의 값이 정규 분포 형태를 유지하는 것을 선호함.

*로그 변환은 데이터 분포도가 심하게 왜곡되어 있을 경우 적용하는 중요 기법 중에 하나임. 원래 값을 log 값으로 변환해 원래 큰 값을 상대적으로 작은값으로 변환하기 때문에 데이터 분포도의 왜곡을 상당 수준 개선해 줌.
-> 로그 변환은 넘파이의 log1p() 함수를 이용해 간단히 변환이 가능함.

*레이블이 극도로 불균일한 데이터 세트에서 로지스틱 회귀는 데이터 변환시 약간은 불안정한 성능 결과를 보여주고 있음.

outlier(이상치 데이터) : 전체 데이터의 패턴에서 벗어난 이상 값을 가진 데이터이며, 아웃라이어라고도 불림.
이상치로 인해 머신러닝 모델의 성능에 영향을 받는 경우가 발생하기 쉬움.
이상치를 찾는 방법에는 여러 가지가 존재. ex. IQR 방식(사분위 값의 편차를 이용하는 기법으로 흔히 박스 플롯 방식으로 시각화 할 수 있음)

*매우 많은 피처가 있을 경우 이들 중 결정값과 가장 상관성이 높은 피처들을 위주로 이상치를 검출하는 것이 좋음. 모든 피처들의 이상치를 검출하는 것은 시간이 많이 소모되며, 결정값과 상관성이 높지 않은 피처들의 경우는 이상치를 제거하더라도 크게 성능 향상에 기여하지 않음.

*SMOTE를 적용하면 재현율은 높아지나, 정밀도는 낮아지는 것이 일반적임. 때문에 정밀도 지표보다는 재현율 지표를 높이는 것이 머신러닝 모델의 주요한 목표인 경우 SMOTE를적용하면좋음. 좋은 SMOTE 패키지일수록 재현율 증가율은 높이고 정밀도 감소율은 낮출 수 있도록 효과적으로 데이터를 증식함.