타이타닉 데이터 실습

201600779 김영민

- Dataset

- Dacon에 있는 Titanic Data를 사용하였으며 train,val 비율은 8:2로 설정
- 평가지표는 AUC
- survived : 0은 사망, 1은 생존

- EDA 및 Preprocessing

- missing data 확인
  - missingno 모듈을 통해 확인
  - Age 및 Cabin에 결측치 다수 분포
- 성별에 따른 생존여부 countplot
  - 여자가 비율적으로 더 많이 생존
  - 긴급 상황 발생 시 여성 및 노약자 배려로 더 많이 생존한 것으로 추측
- pclass에 따른 생존여부
  - pclass 1일수록 생존할 가능성이 높다.
- 성별과 pclass를 고려한 생존 여부 그래프
  - pclass가 1이고 여자인 사람이 현저하게 생존여부가 높음
- Embarked의 결측치 처리
  - 최빈값으로 대체
- fare의 결측치 처리
  - pclass 와 fare의 상관관계가 -0.55로 음의 상관관계를 가짐
  - 따라서 각 pclass의 평균값으로 fare의 결측치 대체(groupby 사용)
- Ticket의 큰 의미가 없다 생각하고 삭제
- 연속형 변수인 Age를 범주형 범위로 변환
  - 20세 미만은 10대, 나머지는 20대,30대..로 정의하고 60세 이상은 old라고 정의함
- 이름에서 Mr, Mrs 등의 사회적 지위를 추출
- 형제 자매 및 자식의 여부의 컬럼을 합쳐서 family_size라는 컬럼을 생성
  - 이를 범주형으로 바꿈. 1은 isalone, 2는 small, 5이상은 big으로 정의
- 기존에 이름에서 뽑은 사회적 지위의 카테고리가 너무 많으므로 비슷한 카테고리를 기준으로 묶음
- cabin이란 값이 있으면 1, 없으면 0으로 처리
- 다른 변수와 종속되는 변수들을 삭제

- Modeling

- Automl 사용
  - Classification을 위한 여러 모델을 비교하기 위해 automl로 여러 가지 classification 모델을 비교
  - AUC score 기준으로 상위 5개 모델 선택
  - LGBM, Catboost, ExtraGradientBoost, GradientBoost, Logitstic Regression 선택
  - 선택된 모델을 Soft Voting을 통해서 하나의 모델을 도출 후 예측
  - 최종 validation 모델 auc score : 90.23
- Tunning 모델 추가
  - LGBM, Catboost를 RandomSearch로 튜닝(AUC score 1,2위)
  - LGBM 결과 기존의 모델보다 더 성능이 안좋음
  - CatBoost 결과 기존의 모델보다 성능이 좋음
  - 최종 예측 모델을 위에서 뽑았던 성능이 좋았던 5개의 모델, Voting한 모델, 튜닝한 Catboost 모델을 합쳐서 soft voting 2차 진행
  - validation auc score = 88.85

- Test set submit

- 첫 번째 automl만을 사용하여 voting한 모델(validation auc score : 90)의 test 결과는 82.6%
- 두 번째 튜닝한 모델을 추가한 모델(validation auc score : 88)의 test 결과는 82.7%