



KISTI Kaggle Competiton

김은별

대회 개요

◆ Introduction

- 한국과학기술정보연구원(KISTI)의 “2020년 하반기 과학기술 빅데이터 분석가 과정”에서 진행하는 캐글 대회.
- 순위가 중요한 게 아니라 데이터를 분석하고, 모델을 생성하고, 그 모델에 데이터를 돌려보는 과정에 익숙해지는 것이 목적.

◆ Competition background

- 역사상 가장 악명 높은 타이타닉 참사.
- 첫 항해 도중 빙산과 충돌 후 침몰.
- 어떤 특징을 가진 사람들이 생존할 가능성이 높았는지 분석.
- 이를 기반으로 머신러닝 모델을 만든 뒤 생존유무를 예측.

Feature Engineering

◆Pclass

- Pclass(승선등급)이 높을수록 생존율이 높다

◆Sex

- 남성에 비해 여성의 생존율이 높다
- 모든 Pclass에서 동일

◆Age


- Age가 어린 아동의 경우, 생존율이 높다
- Age Null값 채우기 - Initial로 성별&나이 추정 후 평균값으로 채운다
- Age를 5개의 구간으로 나누어 Age Band로 분류.
- Age Band를 사용하고 Age는 Drop.

Feature Engineering

◆ SibSp, Parch

- SibSp(형제자매)와 Parch(부모자식)이 있을 때의 생존율
- Family Size로 둘을 합쳐서 하나의 Feature로 만들어준다
- Family Size를 사용하고 SibSp와 Parch는 각각 Drop

◆ Fare

- Fare(요금) Range를 4개의 범위로 나눈다
- 복잡한 숫자인 Fare Range를 간단히 Fare Catagory화한다
-  Fare Range를 사용하고 Fare는 Drop

◆ Drop할 Column 정하기

- Cabin, Name, PassengerId, Ticket을 삭제한다

사용 모델 Catboost

◆모델 생성

- !pip install catboost
- from catboost import CatBoostClassifier, cv, Pool
- M_catboost = CatBoostClassifier(custom_metric=['AUC'],
random_seed=20, logging_level='Silent')

◆모델 학습

- M_catboost.fit(train_X, train_Y, eval_set=(test_X, test_Y), plot=True)

◆모델 적용

- model=M_catboost
- prediction=model.predict(df_test)

사용 모델 앙상블

◆모델 앙상블

- RandomForest, XGBoost, LGBM model, CatBoost, MLP
- 각 모델의 스코어를 봤더니 XGBoost가 0.87이고 나머지는 다 0.85
- XGBoost의 평균에만 가중치를 더 줘서 0.28로, 나머지는 0.18의 가중으로 진행

◆모델 적용

- `pred_wavg = np.round(np.average([pred_rf, pred_xgb, pred_lgbm, pred_cboost, pred_mlp], weights=[0.18, 0.28, 0.18, 0.18, 0.18], axis=0)).astype(int)`

점수 (Score)

EDA_submission (3).csv

2 days ago by KEB

M_catboost 사용

0.76076

0.77511

titanic_wavg (2).csv

20 hours ago by KEB

앙상블 모델 가중치 조정 (하나만 0.28로)

0.75598

0.73684

노하우 및 소감

- ◆ 모델을 만드는 것보다, 모델을 돌리기 전에 프로그램이 인식할 수 있게 데이터를 손보는 데이터 전처리 과정이 훨씬 더 복잡하고 까다로웠다
- ◆ 앙상블 모델 중에서 파라미터를 손본 버전이 Public 점수는 높았는데 Private 점수는 낮았다. 파라미터를 다시 수정해봐야 할 듯.
- ◆ 올려진 코드를 그대로 적용만 하는데도 왜.. 어렵지??
- ◆ 하나하나 뜯어보면서 적용하는 과정에서 모르는 구문이 나올 때마다 머리를 쥐어뜯었는데 그래도 재미있었다