Kaggle Competitions ①

1. Titanic : Machine Learning from Disaster

◯ Data

- train.head() → (891, 11)
- test.head() → (418, 10)
- 컬럼 설명

① Survival : 생존 여부. 0이면 사망, 1이면 생존
② Pclass : 티켓 등급. 1등석(1), 2등석(2), 3등석(3)
③ Sex : 성별. 남자(male)와 여자(female)
④ Age : 나이
⑤ SibSp : 해당 승객과 같이 탑승한 형제/자매 (siblings)와 배우자(spouses)의 총 인원 수
⑥ Parch : 해당 승객과 같이 탑승한 부모(parents)와 자식(children)의 총 인원 수
⑦ Ticket : 티켓 번호
⑧ Fare : 운임 요금
⑨ Cabin : 객실 번호
⑩ Embarked : 선착장. C는 셰르부르(Cherbourg) 프랑스, Q는 퀸스타운 (Queenstown) 영국, S는 사우스햄튼(Southampton) 영국 지역

◯ Environment

- Python 3.73 with `jupyter==1.0.0`
- pandas, numpy, matplotlib, seaborn
- scikit-learn

➢ Perpose

어떤 승객이 생존하며, 또한 어떤 승객이 사망하는지를
예측하는 예측 모델을 만드는 것

➢ Methodology

- 알고리즘 : DecisionTree, RandomForest(Coarse & Fine Search)
- 모델평가 : Accuracy

➢ Solution

- Exploratory data analysis
  - 남자 승객의 생존률은, 18.9% 여성 승객의 생존률은 74.2%
  - Pclass 2등급인 경우 생존률이 1/2(50%), 3등급인 경우 1/4(25%)
  - Cherbourg(C)에서 탑승할수록 생존할 확률이 높으며, Southampton(S)에서 탑승할수록 사망할 확률이 높음
  - 나이가 15세 이하인 승객은 생존 확률이 높음
  - 핵가족(Nuclear)의 생존률이 높고, Single(독신), Big의 생존률이 낮음
- Feature : 등급(Pclass), 성별(Sex_encode), 운임요금(Fare_fillin), 선착장(Embarked), Age(Child 여부), 가족 수(Sibsp+Parch), 이름(Name 내 Master 포함여부)
- Label: 생존 여부(Survived)
- 스코어 : DecisionTree 0.78476, RandomForest 0.81339 예측(상위6% Rank)