

Kaggle Project

발표자 : 김유진

1. 대회 개요

- 대회명 : KISTI KAGGLE COMPETITION(4TH)
- 2020년 하반기 KISTI 과학기술 빅데이터 분석가 과정 캐글대회
- 기간 : 2020년 11월 23일 ~ 2020년 12월 03일 오전 10시
- 내용 : [Titanic: Machine Learning From Disaster] 데이터 활용
생존 여부 분류 문제 해결 (Binary Classification)

2. Feature Engineering

Before

```
PassengerId Survived Pclass Name Sex Age SibSp Parch Ticket Fare Cabin Embarked
0 1 0 3 Braund, Mr. Owen Harris male 22.0 1 0 A/5 21171 7.2500 NaN S
1 2 1 1 Cumings, Mrs. John Bradley (Florence Briggs Th... female 38.0 1 0 PC 17599 71.2833 C85 C
2 3 1 3 Heikkinen, Miss. Laina female 26.0 0 0 STON/O2. 3101282 7.9250 NaN S
3 4 1 1 Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35.0 1 0 113803 53.1000 C123 S
4 5 0 3 Allen, Mr. William Henry male 35.0 0 0 373450 8.0500 NaN S
time: 22.7 ms

! train.shape
(891, 12)time: 2.17 ms
```

After

```
FamilySurvived FamilyDied FamilySize IsAlone Age_band rich_woman men_3 Pclass_1 Pclass_2 Pclass_3 Sex_female Embarked_C Embarked_Q Embarked_S Cabin_A Cabin_B Cabin_C Cabin_D Cabin_E Cabin_F Cabin_G Cabin_X
0 0.0 0 0 2 0 1 0 1 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 1
1 1.0 0 0 2 0 2 1 0 1 0 0 1 1 0 0 0 0 1 0 0 0 0 0 0
2 1.0 0 0 1 1 1 0 0 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0 1
3 1.0 0 1 2 0 2 1 0 1 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0
4 0.0 0 0 1 1 2 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 1
time: 46 ms

! train.shape, test.shape
((891, 23), (418, 22))time: 3.01 ms
```

Columns (Survived 제외) : 11개 → 22개

2. Feature Engineering

- 원본 데이터에 대한 가공 내용 (추가/삭제/변경 등)
 - ✓ 추가 : LastName, FamilySurvived, FamilyDied, FamilySize, IsAlone, Initial, Age_band, rich_woman, men_3

LastName	Name에서 첫 번째 단어 추출
FamilySurvived, FamilyDied	LastName과 Ticket이 같으면 가족으로 간주. Survived 합계
FamilySize	SibSp + Parch + 1(본인)
IsAlone	FamilySize가 1인 경우
Initial	Mr, Miss 등 Name에서 호칭 단어 추출
Age_band	나이대를 5개 그룹으로 분류 (16, 32, 48, 64 기준)
rich_woman, men_3	Pclass와 Sex를 이용하여 컬럼 생성 <ul style="list-style-type: none">- rich_woman : Pclass 1 + female- men_3 : Pclass 3 + male

2. Feature Engineering

- 원본 데이터에 대한 가공 내용 (추가/삭제/변경 등)

- ✓ 삭제 : PassengerId, Age, Ticket, LastName, SibSp, Parch, Sex_male, Name, Initial, Fare, Cabin_T

Ticket, Name, LastName, Initial	str 데이터인 경우
SibSp, Parch, Sex_male, Age	데이터 성질이 중복될 경우
Cabin_T	train 데이터셋에는 있지만 test 데이터셋엔 없는 경우
Fare	모델 학습 시 Feature Importance가 현저히 높은 경우

- ✓ 변경 : Embarked, Cabin, Pclass, Sex

Cabin	구역을 나타내는 알파벳으로 변경. 빈값은 X
Embarked, Cabin, Pclass, Sex	One-hot encoding


3. 사용 모델

- Random Forest
- XG Boost
- Light GBM
- Cat Boost
- 평균, 가중치를 통한 4개 모델 앙상블


✓ 실험을 통해 성능이 가장 높은 모델 선택 : **XG Boost**

4. 점수

- Public Leaderboard : **0.79904**

Yujin Kim		0.79904	22
-----------	---	---------	----

- Private Leaderboard : **0.77033**

Yujin Kim		0.77033	22
-----------	--	---------	----

5. 노하우 및 소감

- Feature를 잘 변경 하는 것이 모델 성능에 영향을 준다