2020년 하반기 4차인재 양성사업 과학기술 빅데이터 분석가 양성 과정
KISTI Kaggle Competition (4TH)

Titanic

2020.12.03.

이한별

www.kaggle.com/c/kisti-kaggle-competition-4th

대회 개요

- KISTI 빅데이터 분석가 과정 내부 대회
- 타이타닉 호의 생존자를 머신러닝을 통해 예측하기
- 주어진 feature들을 가공 및 활용하여 생존 여부 예측

Features

- Age

- Initial을 생성하여 평균으로 Null 값 채움

- Mr_e로 비교적 나이가 많은 남성 따로 추출

```
# initial 치환
df_train['Initial'].replace(['Mlle','Mme','Ms','Dr','Major','Lady','Countess',
                             'Jonkheer','Col','Rev','Capt','Sir','Don','Dona'],
                            ['Miss','Miss','Miss','Other','Mr_e','Mrs','Mrs',
                             'Other','Mr_e','Mr_e','Mr_e','Mr_e','Mr_e','Mrs'],inplace=True)
```

Initial	Age
Master	5.482642
Miss	21.834533
Mr	32.252151
Mr_e	47.176471
Mrs	37.046243
Other	42.875000

Features

- Age_cat - categorize

- 10살 간격이 5살 간격보다 상관관계가 높았음

- Age 삭제, Age_cat 활용

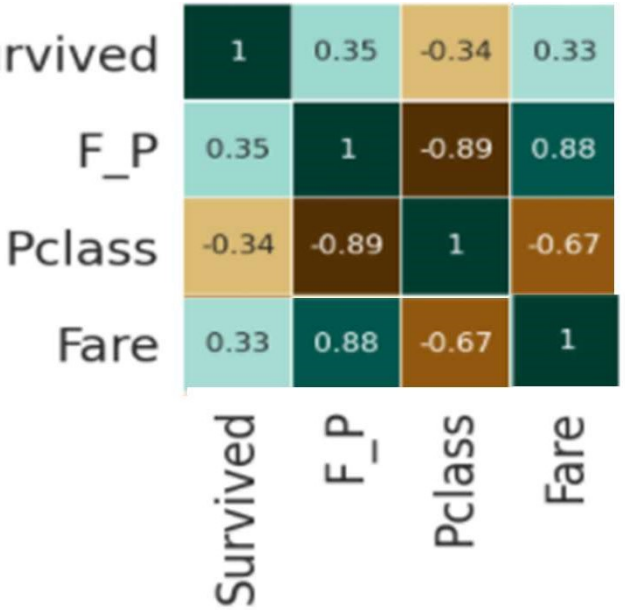
Features

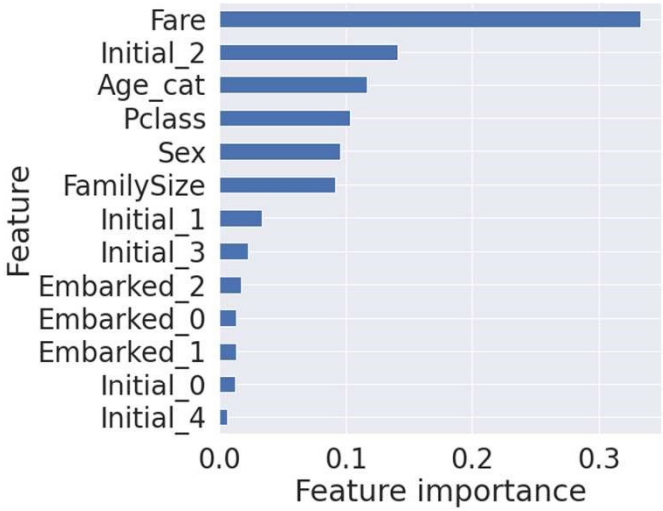
- Embarked
- 가장 많이 탑승한 S로 Null 값 채움
- One hot encoding
- Pclass
- One hot encoding
- Cabin
- Null 값이 많아 제거
- SibSp, Parch
- 가족 관련 feature들과 중복 -> 제거

New columns

- F_P
- Fare(정규화한 값) / Pclass
- Fare를 넣었을 때 영향력이 크기 때문에 가공해 봄

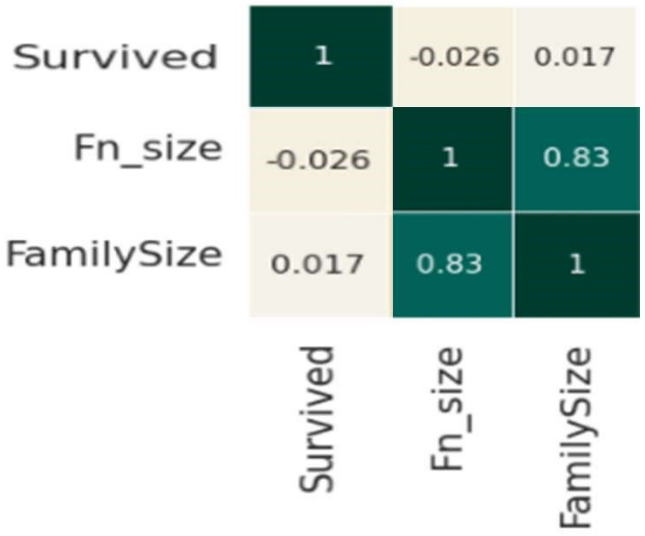
Pclass	Fare	F_P
3	2.085672	0.695224
1	4.690430	4.690430
3	1.981001	0.660334





New columns

- FN_size
- Name에서 성(Family name) 추출
- 같은 성을 가진 사람들을 count
- 부모자녀, 형제자매, 배우자 외 다른 가족들을 포함할 가능성
- 그냥 성이 같은 남일 가능성




```
'Foreman': 1,
'Fortune': 6,
'Fox': 2,
'Francatelli': 1,
'Franklin': 2,
'Fraunenthal': 3,
'Frolicher': 1,
'Frolicher-Stehli': 2,
'Frost': 1,
'Fry': 1,
'Funk': 1,
'Futrelle': 2,
'Fynney': 1,
'Gale': 2,
'Gallagher': 1,
'Garfirth': 1,
'Garside': 1,
'Gaskell': 1,
'Gavey': 1,
'Gee': 1,
'Geiger': 1,
'Gheorgheff': 1,
'Gibson': 2,
'Giglio': 1,
'Gilbert': 1,
'Giles': 3,
```

New columns

- TicketNo

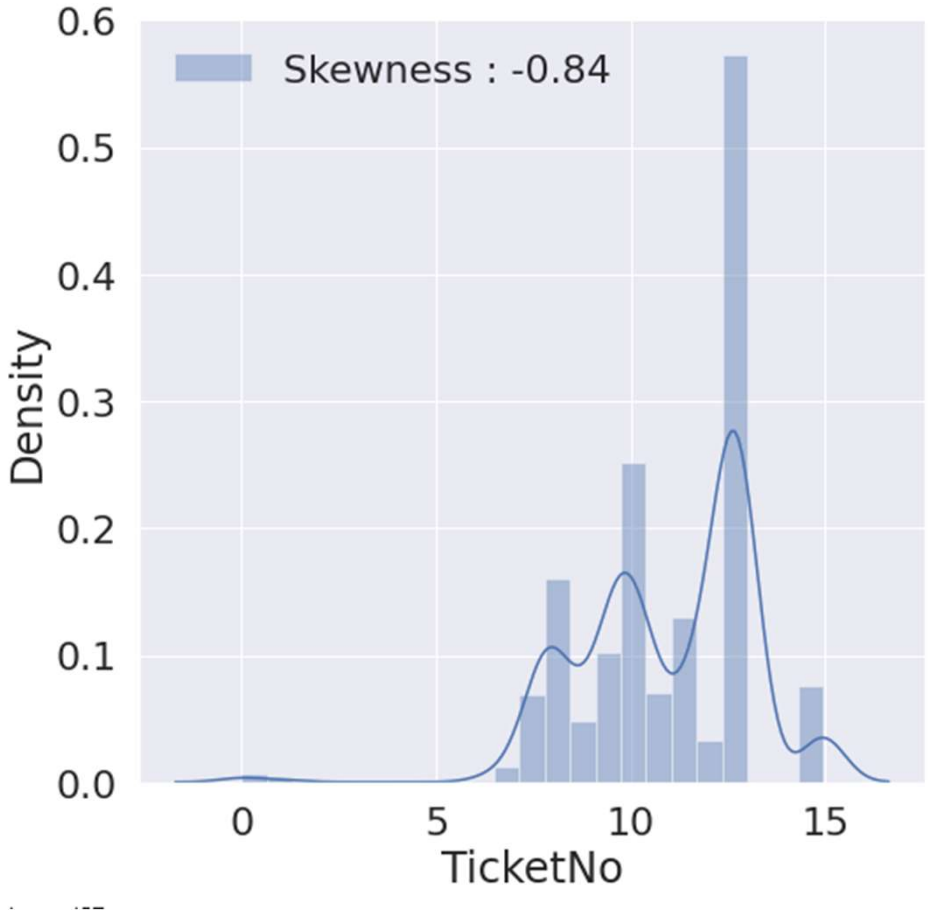
- Ticket에서 숫자만 추출(문자열만 있으면 0)

- Ticket번호가 유사하면 객실이 비슷하거나 일행이지 않을까?

Passenger	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1081	2	Veal, Mr. James	male	40	0	0	28221	13		S
1082	2	Angle, Mr. William A	male	34	1	0	226875	26		S
1083	1	Salomon, Mr. Abraham L	male		0	0	111163	26		S
1084	3	van Billiard, Master. Walter	male	11.5	1	1	A/5. 851	14.5		S
1085	2	Lingane, Mr. John	male	61	0	0	235509	12.35		Q
1086	2	Drew, Master. Marshall Bri	male	8	0	2	28220	32.5		S

New columns

- TicketNo
  - Ticket번호가 편차가 커서 정규화

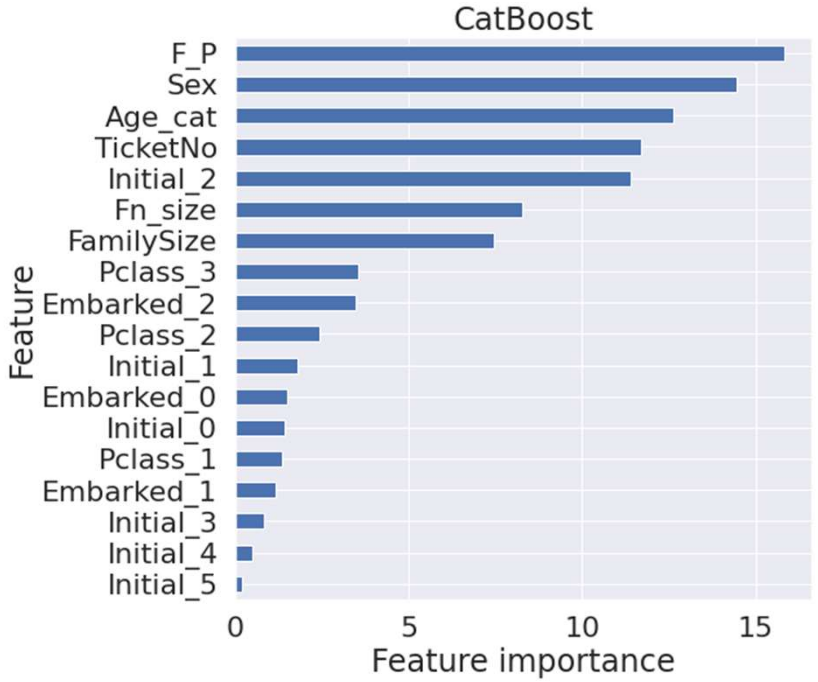


Features 선택

- Sex ~ Initial_0-5 의 importance가 높게 나타남
- 상관관계가 있는 다른 column들도 함께 사용해 봄
  - Pclass ~ F_P
  - FamilySize ~ Fn_size

- 상관관계가 있는 column 중 한쪽을 삭제하지 않는 게 score 높았음

-> 여러 차례의 검증 필요함!!



사용 모델

- 수업 시간 Ensemble 활용

-Randome Forest

-XGBoost

-LightGBM

-Catboost

- 계층적 샘플링

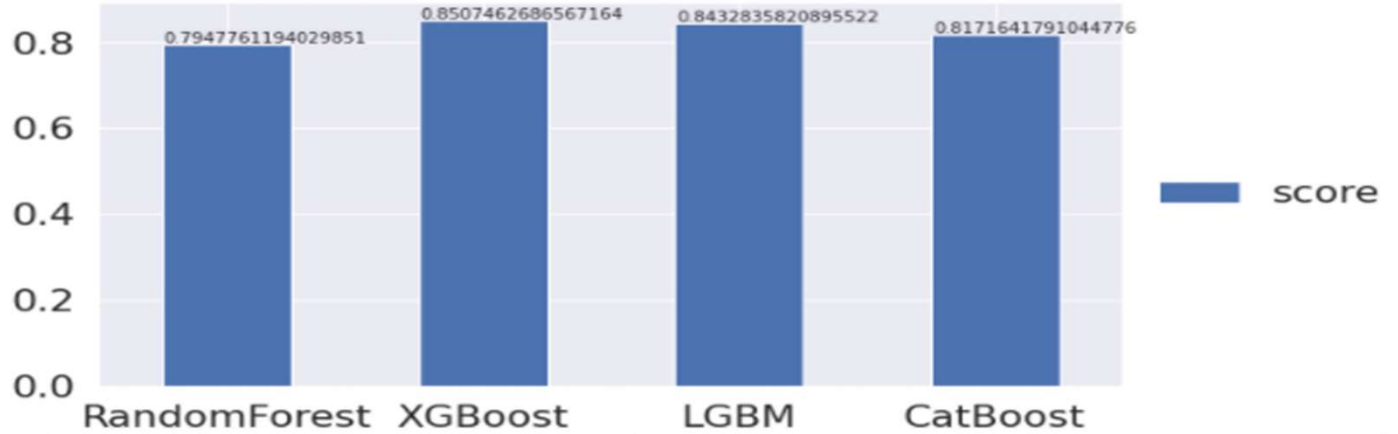
```
from sklearn.model_selection import StratifiedShuffleSplit

split = StratifiedShuffleSplit(n_splits=1, test_size=0.3, random_state=27)
for train_index, vld_index in split.split(df_train, df_train["Sex"]):
  strat_train_set = df_train.loc[train_index]
  strat_test_set = df_train.loc[vld_index]

df_train["Sex"].value_counts() / len(df_train)
```

```
1    0.647587
0    0.352413
Name: Sex, dtype: float64time: 80.1 ms
```

사용 모델

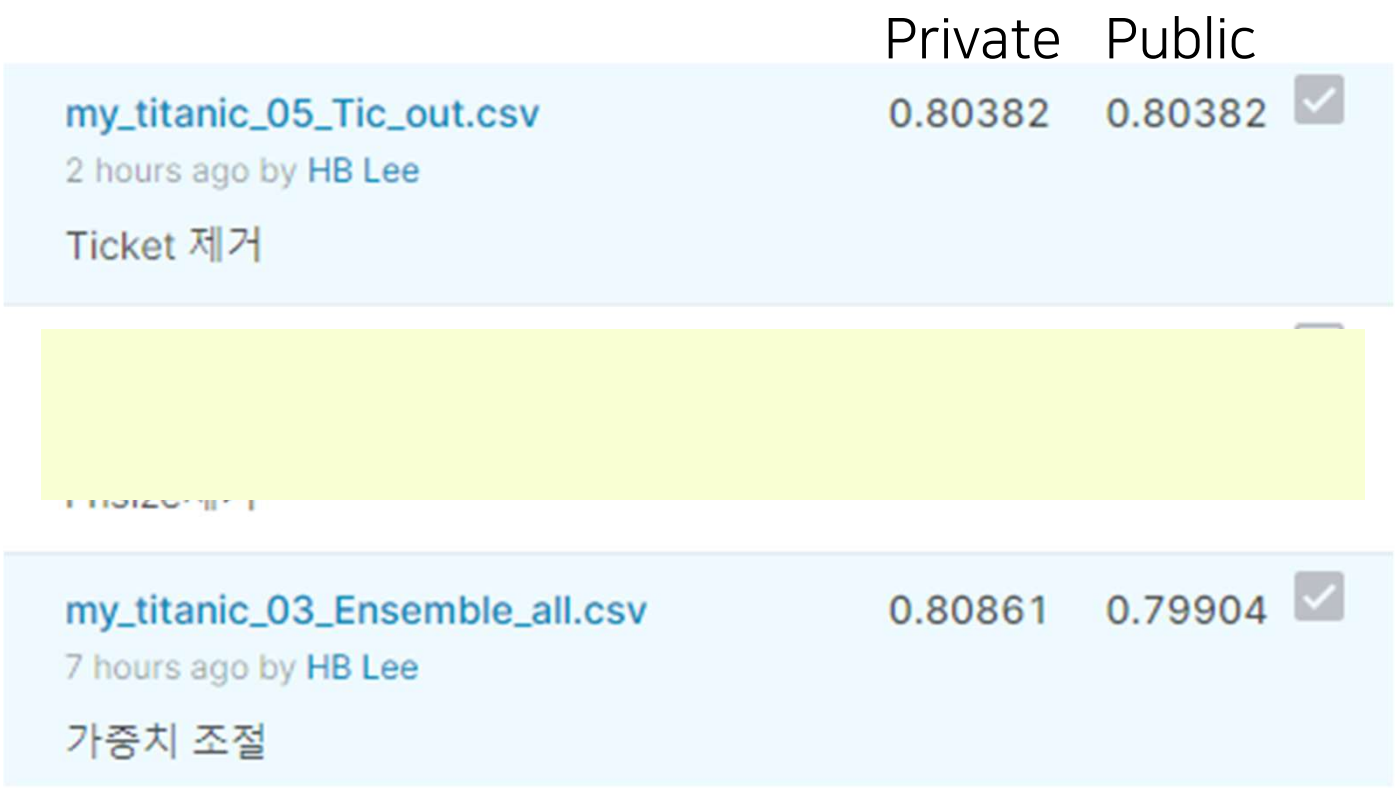



```
pred_wavg = np.round(np.average([pred_rf, pred_xgb, pred_lgbm, pred_cboost],
                                weights=[0.1, 0.5, 0.2, 0.2], axis=0)).astype(int)
score_wavg = metrics.accuracy_score(pred_wavg, y_vld)
score_wavg = round(score_wavg, ROUND_NUM)
print("Weighted Average: ", score_wavg)
```

```
Weighted Average:  0.8619402985074627
time: 19.5 ms
```

- 단독 모델 사용 때보다 score가 높은 가중치를 선택

점수 비교


Private
Public
my_titanic_05_Tic_out.csv
0.80382
0.80382
2 hours ago by HB Lee
Ticket 제거
my_titanic_03_Ensemble_all.csv
0.80861
0.79904
7 hours ago by HB Lee
가중치 조절

앞으로의 과제

- 중간중간 데이터 탐색 및 메모 잘 하기
  - 그래프를 자유자재로 그릴 수 있으면 유용할 듯함
  - 값을 어떻게 조정했는지, 결과는 어떤지 메모
- 다양한 모델을 활용해보고 튜닝, 검증도 해봐야 함