

# Kaggle Project

KISTI KAGGLE COMPETITION(4TH)

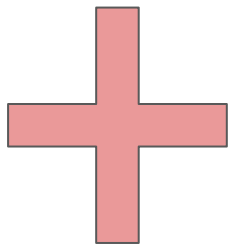
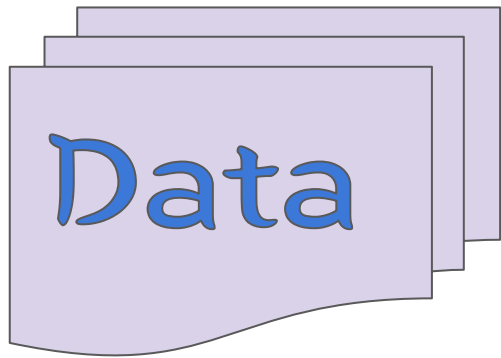
서 유정

2020/12/03

# Contents


1. 대회 개요
2. Feature Engineering
3. 사용 모델
4. 랭킹
5. 소감

## 1. 대회 개요



Survived or Not ?

## 2. Feature Engineering

Pclass	Name	Ticket 	Fare	Embarked
3	Bing, Mr. Lee	1601	56.4958	S
3	Ling, Mr. Lee	1601	56.4958	S
3	Lang, Mr. Fang	1601	56.4958	S
3	Foo, Mr. Choong	1601	56.4958	S
3	Lam, Mr. Ali	1601	56.4958	S
3	Lam, Mr. Len	1601	56.4958	S
3	Chip, Mr. Chang	1601	56.4958	S

- ticket 번호가 동일하고 ,embarked도 동일한 데이터 들이 존재
- 이러한 데이터들의 요금은 대체로 비싼 편
- 따라서 ticket의 갯수만큼 요금을 나눠 Fare를 재 처리

## 2. Feature Engineering

data

	Survived	Pclass	Sex	Age_band	Family_Size	Fare_cat	Embarked_C	Embarked_Q	Embarked_S
0	0	3	1	2	2	1	0	0	1
1	1	1	0	3	2	4	1	0	0
2	1	3	0	2	1	2	0	0	1
3	1	1	0	3	2	4	0	0	1
4	0	3	1	3	1	2	0	0	1

- Fare 분류 할 때, 0원도 따로 '0'이라고 분류
- initial은 Age null값 채우는 용도로 쓰였기에 그냥 drop
- 위 column 들이 전처리 후의 데이터 들

### 3. 사용모델 (Catboost)

```
[146] from catboost import CatBoostClassifier, cv, Pool
```

```
[9] M_catboost = CatBoostClassifier(custom_metric=['AUC'],  
                                   random_seed=42, logging_level='Silent')  
M_catboost.fit(X_tr,y_tr, eval_set=(X_vld, y_vld),  
               plot=True)
```

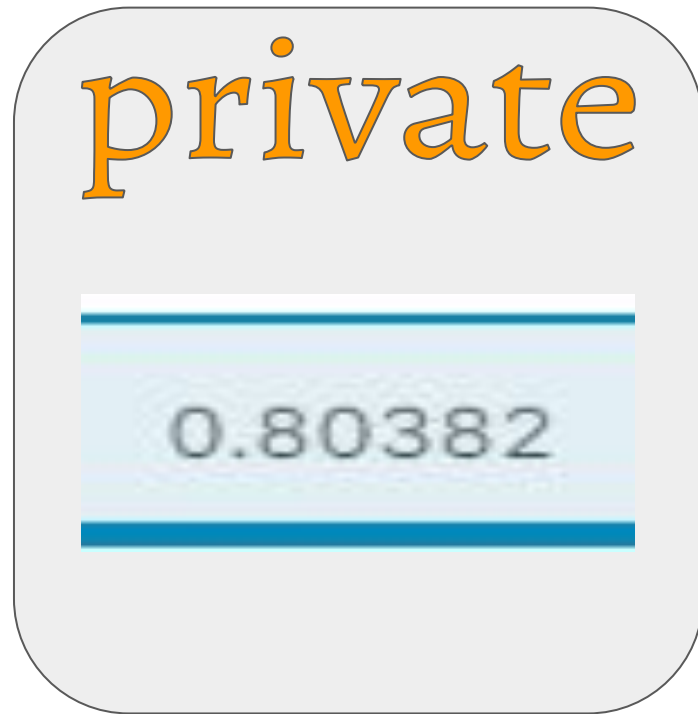
```
↳ <catboost.core.CatBoostClassifier at 0x7f9935521e10>
```

```
[148] prediction=M_catboost.predict(X_vld)  
print('The acc for Catboost is:', metrics.accuracy_score(prediction,y_vld))  
result = cross_val_score(M_catboost,X_train,Y_train,cv=10,scoring='accuracy')  
print('The cross val score for catboost is:', result.mean())
```

The acc for Catboost is: 0.8656716417910447

The cross val score for catboost is: 0.8171535580524344

#### 4. 랭킹 (?!)



## 4. 랭킹 (?!)

Submission and Description	Private Score	Public Score
<a href="#">baseline_submission02.csv</a> 7 days ago by Yujeong Seo 02	0.84126	0.71575
<a href="#">baseline_submission04.csv</a> 7 days ago by Yujeong Seo LogisticRegression	0.83333	0.73287
<a href="#">titanic_submission_20201126_01.csv</a> 7 days ago by Yujeong Seo	0.83333	0.71232



1. grid search가 모델마다 조금씩 다른건지, 아니면 내가 요령이 없는건지 모르겠지만 너무 오래걸리고 어려웠다.
2. 데이터 전처리 할 때, 여러 생각은 떠올랐지만 코딩 능력이 부족하여 실행할 수 없었다.
3. 머신러닝의 7할은 전처리와 데이터 분석이라 생각한다. 2할은 데이터 기반으로 모델선정  
1할은 심신의 안정,,,,
4. 가르쳐 주신 모든 모델을 전부 돌려봤다. 분명 다른 사람들도 나와 같은 모델들을 돌렸을 텐데 score가 다른 것을 보면서 전처리의 중요성을 다시 한번 느꼈다.
5. 훗날, 내 분야에서 특정 data를 가지고 작업을 할 때를 위해서라도 , 전공은 절대 손에서 놓지 말아야겠다.
6. 꾸준히 공부하는 습관을 가져야만 한다고 매우, 속 깊이 느끼는 중이다.



끝

down the curtain