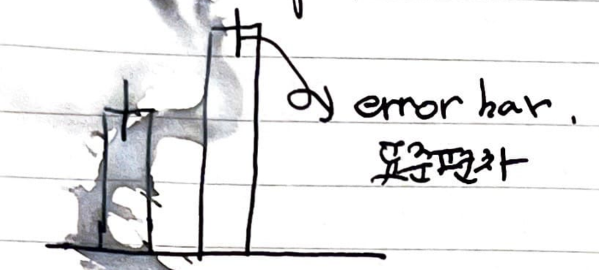
DieTanic EDA

2023. DATE 06. 29. NO.

data.pclass, data.survived 이 방식보다 data['Sex'], data['Pclass'] 추천합니다. 가시성이 훨씬 좋아요.

>> crosstab → sns.factorplot 으로 가시화
└ ex) groupby랑 비슷.

이거 안써 → catplot



Pclass feature는 중요하구나 유의미해!
target과의 관련성.

>>
Age값은 연속형 값은 히스토그램을 많이 그린다.
왜냐하면 함수형태인 kdeplot을 이용하자. 가우시안 커널.

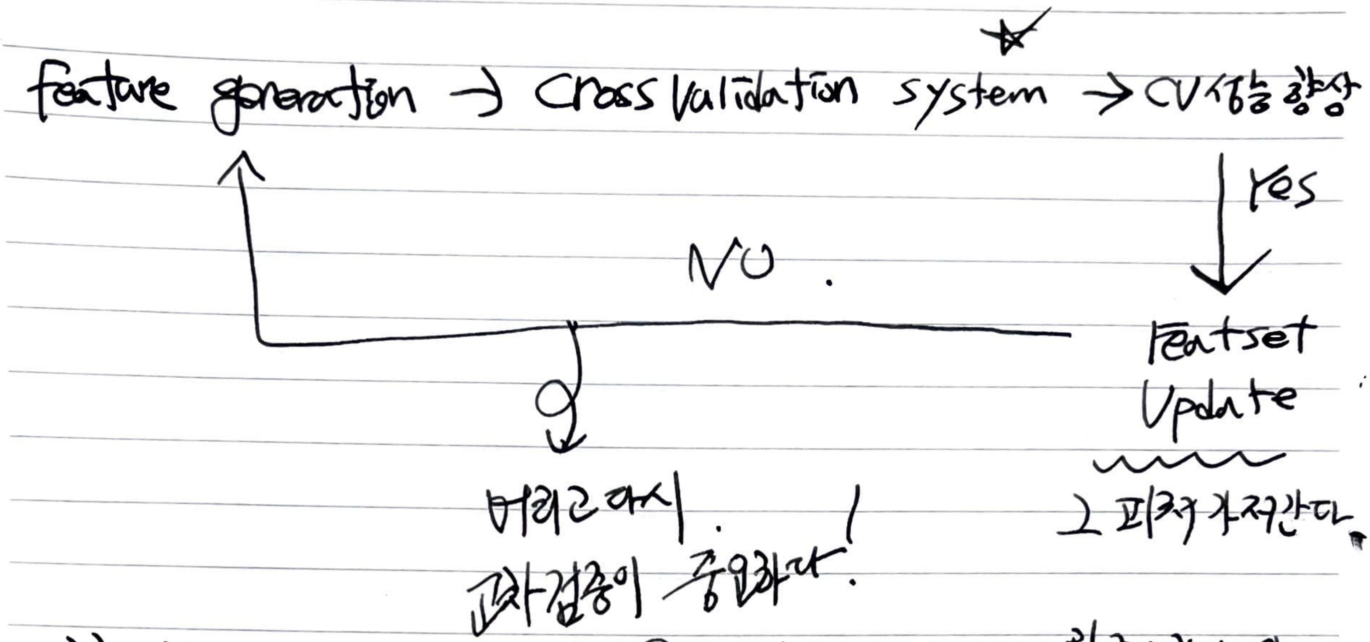
Null값을 해결할때 group화를 해서 평균을 구해 넣어주면 좀 낫다.

feature engineering은 데이터를 분석해서 insight를 얻거나 우리가 가진 상식을 가지고 파생화 시킨다음에 그것을 모델에 넣어주고 학습을 해서 더 좋은 성능을 얻는것.

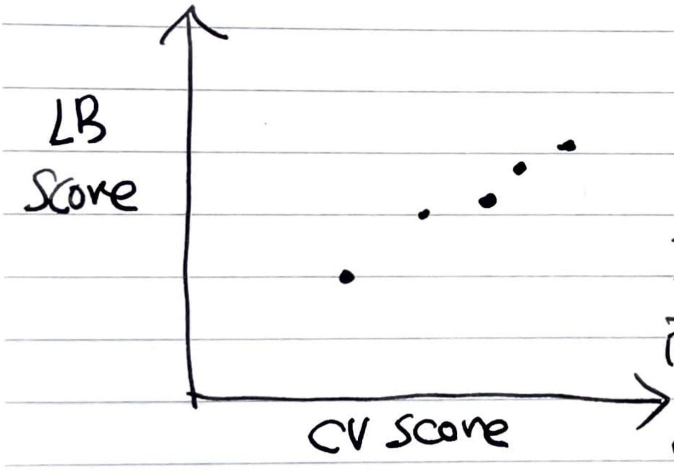
>>
feature를 변경해주었을 때 주의할 점은 `inplace = True`를 지정해서 원래 값에 덮는 방법이 있고, 다른 방법으로는 변수 안에 넣는 것이다.

>>
data.loc[(data.Age.isnull()), :]
→ location 안에 mask (조건)을 넣으면 참 편리하다.

1. data .isnull().sum()
2. EDA
3. 교차검증 System.
 (cross validation)



>> Feature generation과 Cross-validation은 함께 해야 한다.



	CV	● LB
Featset1	0.6	0.7
Featset1 + new feat(featset 2)	0.7	0.75
→ Featset2 + new feat2 (featset3)	0.8	0.8
Featset3 + Newfeat	0.75	0.775

→ CV와 LB가 같이 움직인다.
좋은 교차검증 끝에는 좋은 feat이 있다.

CV와 LB가 미묘하게 잘 움직이는
교차검증 system 만들자.

교차검증 fold?

K-fold
Loo
Lpo
.
.