23.03.20 월요일 줌회의

팀장(이도원) 이 제시해준 발표 방향성

모든 칼럼 설명 더 넣어야 됨.

* 이 칼럼은 분석에 쓸데없어 보인다. => 왜 이런 결론에 도출했는지
* 이 칼럼은 결측치를 채워야 한다. => 채울 결측치는 뭘 기준으로 넣었는지
* 인코딩을 했다. => 뭘 기준으로 인코딩을 해서 결론적으로 어떤 형태가 나오는 지.EDA 소결

위에서 설명한 EDA 내용 요약하는 동시에, 알고리즘 선택을 자연스럽게 유도하도록.

=> 로버스트 스케일러 + 그래디언트 부스트 분류기 조합사용해본 알고리즘 종류

LogisticRegression, RandomForest, GradientboostingClassifier, HistGradientBoostingClassifier, SVM모델링

* kfold 적용 => 적용 이유 : 과적합 방지
* 하이퍼파라미터가 어떻게 나왔는지 => GridSearchCV
* 검증 데이터 평균값이 테스트 데이터 값이랑 별로 차이가 안난다.
* 페이즈를 3개로 나눈 이유 : 강사님이 시켜서
* 전체 column에 대해 회귀분석을 했는데 Parch, Fare 요 두개 칼럼이 유효하지 않았음. [P-Value 0.05 초과]
* 나는 빼려고 했는데 강사님이 넣어도 보고, 빼도 보고
* Cabin 다중공선성이 존재했음.
* 빼려고 했는데, 강사님 조언 왈 : 요즘 추세가 하나라도 더 넣어서 정확도를 조금이라도 올리자.
* 실험 셀을 넣을거면, 강사님이 알려준 라이브러리랑 다른 라이브러리를 썼음.
* Name column은 일부로 넣지 않았다.

23.03.16 목

세미프로젝트 1이 시작되었습니다!복습에 가장 포커스를 맞춰주시고,

에러나 각종 문제가 발생하셨을 때에도 기존 자료를 우선적으로 검토해주세요!

**지난 학습내용 & 소스코드 검토**

우선 차근차근 **각 열들을 살펴보며 EDA & 시각화**를 진행해주세요!EDA & 시각화를 마친 후 팀 내에서의 논의를 바탕으로 차근차근 전처리를 시작해주시고,전처리를 진행할 때에는 **"어떤 열에" / "왜" / "어떻게"** 전처리를 적용했는지 설명을 준비해주시기 바랍니다.

23.03.20 오전

**Age 열**

의 경우는 ML 모델을 활용해 결측치를 예측하여 넣는 편이 좋은데,

그 흐름이 다소 어려운 편이에요!

아래 개괄적인 흐름을 알려드리니 한번 적용해보셔요

**x & y data split**

(최종적으로 예측해야할 대상인 survived 열은 age 예측에 활용되면 안됩니다)

1-2)

**train & test split**

(test 데이터의 행들은 age 열을 예측하는 ML 모델 훈련 시 활용되면 안됩니다.)2) train\_x 기준으로 기존 index 열을 일반 열로 만들어줍니다. 일반 열이 된 기존 index 열은 (y 데이터와의 행 기준 순서를 일치시키기 위해) 나중에 age열 결측치를 채운 후 다시 df을 합친 다음 재정렬할 때 활용합니다.

: df.reset\_index()3) 일반 열이 된 기존 index 열을 포함한 train\_x 데이터프레임을 age 열을 기준으로 결측치가 있는 행들과 결측치가 없는 행들로 나눠 2개의 데이터프레임으로 쪼갭니다.

ex)

x\_train['Age'] = x\_train['Age'].fillna(0)

x\_train\_age\_null = x\_train[ x\_data['Age'] == 0 ]

x\_train\_age\_notnull = x\_train[ x\_data['Age'] != 0 ]4-1) age 열에 결측치가 없는 행들의 df 을 기준으로 하여, age 열과 기존 index 열을 제외한 나머지 열들을 train\_x 로, age 열을 train\_y 로 쪼갭니다.

4-2) age 열에 결측치가 있는 행들의 df 을 기준으로 하여, age 열과 기존 index 열을 제외한 나머지 열들을 test\_x 로, age 열을 test\_y 로 쪼갭니다.5-1) 위에서 만든 train\_x 로 train\_y (age) 를 맞추는 ML 모델을 만들고 학습시킵니다.

5-2) 위에서 만든 ML 모델로 test\_x를 대상으로 age를 predict 하여 별도의 변수로 만듭니다 (ex. predicted\_age)6-1) age 열에 결측치가 있는 행들의 df 을 기준으로 하여, 위에서 만든 예측된 나이값(ex. predicted\_age)을 가져와 age 열을 덮어씌웁니다.

6-2) age 열에 결측치가 없는 행들의 df 와 age 열에 결측치가 있었던(이제는 채워진) 행들의 df 를 위아래로 합칩니다. (pd.concat 활용)7-1) survived 열과의 행 기준 순서를 일치시키기 위해, 일반 열로 만들어두었던 기존 index 열을 기준으로 재정렬합니다. (sort\_values)

7-2) 일반 열로 만들어두었던 기존 index 열을 제거합니다.8) 위 1-2) 에서 나눠두었던 test data를 기준으로 age 열의 값을 채워넣는 작업을 적용합니다.

* 기존 index 열을 일반 열로 변경
* age 열에 결측치가 있는 행들만 모아 df 로 구성
* age 열과 기존 index 열을 제외한 나머지 열들을 x 로 하여 age를 예측하고 age 열에 덮어씌움
* 결측치가 채워진 상태에서 결측치가 없던 행들의 df 와 위아래로 합침
* 기존 index 열 기준으로 재정렬 & 기존 index 열 삭제9) survived 열을 맞추기 위한 모델을 만듭니다.