**Airbnb EDA and Xgboost**

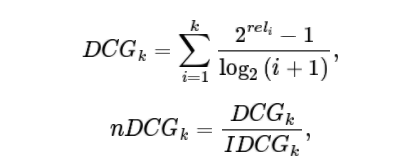
한현수

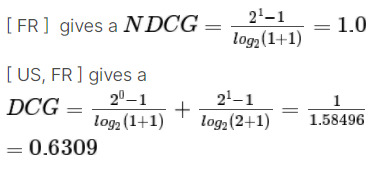
**1. Description**

Airbnb new users booking 이 competition은 Airbnb 새로운 사용자의 첫번째 예약 장소를 추천하는 것이다.

**2. Evaluation**

평가지표는 NDCG(Normalized discounted cumulative gain)를 사용한다.

는 i의 relevance of the result 이다.



예를 들어 추천이 FR 하나인 경우 NDCG는 1이 되고 , US,FR의 경우는 0.6309로 계산된다.

**3. Data**

  \*   id: user id

  \*   date\_account\_created: 계정 생성일

  \*   timestamp\_first\_active: 첫 번재 활동 시간

  \*   date\_first\_booking: 첫 번째 예약 날짜

  \*   gender: 성별

  \*   age: 나이

  \*   signup\_method: 회원 가입 방법, 여부

  \*   signup\_flow: 등록하러 온 경로

  \*   language: 언어

  \*   affiliate\_channel: 마케팅 종류

  \*   affiliate\_provider: 마케팅 제공사

  \*   first\_affiliate\_tracked: 처음으로 사용한 마케팅

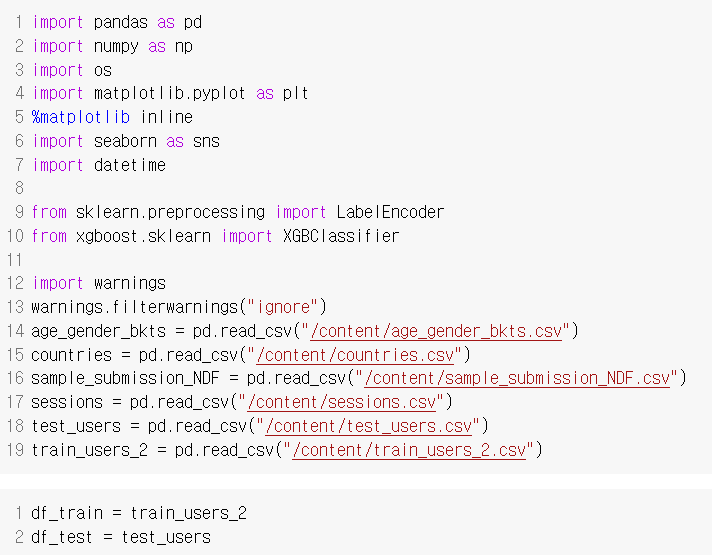
  \*   signup\_app: 앱 가입 여부

  \*   first\_device\_type:

  \*   first\_browser:

\*   country\_destination: target variable로 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU', 'NDF' (no destination found), and 'other' 가 있음. 'NDF'는 예약 안함을 의미한다.

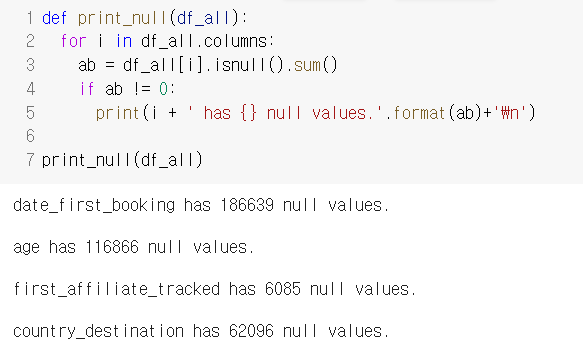
**4. import**

****

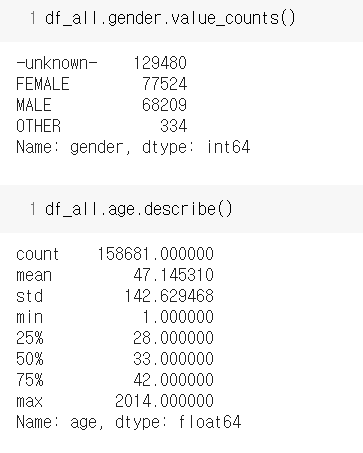
**Categorical data info**

****

**Print Null**

****

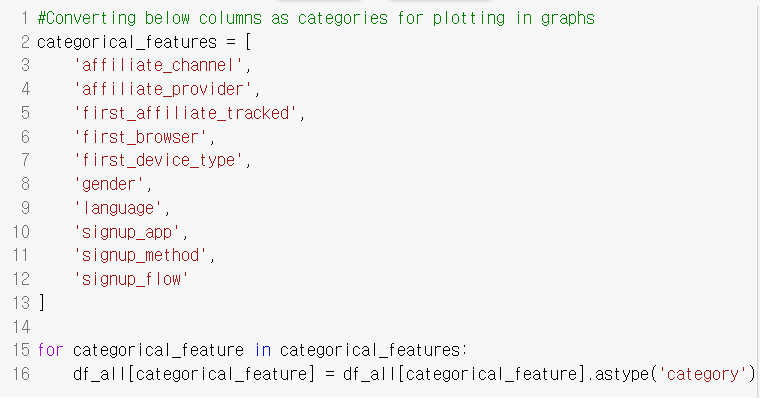
**Gender, Age**

****

이를 보아 성별을 밝히지 않은 사용자들이 약 50퍼정도임을 알 수 있었고, 남성보다 여성 사용자의 비율이 많다는 것을 확인할 수 있었다.

나이의 경우, min이 1, max가 2014임을 보아 이상치가 존재함을 확인할 수 있다.

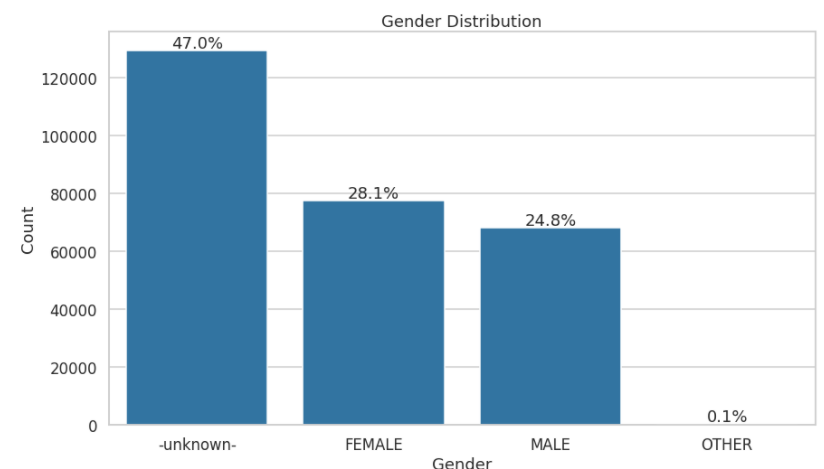
그래서 나이가 15보다 적거나, 100 이상인 것들을 nan으로 처리해준다. 15보다 작은 경우는 airbnb의 사용 연령제한이 15세이기 때문이며, 100 이상은 사람의 수명은 100세를 넘기기 힘들기 때문이다

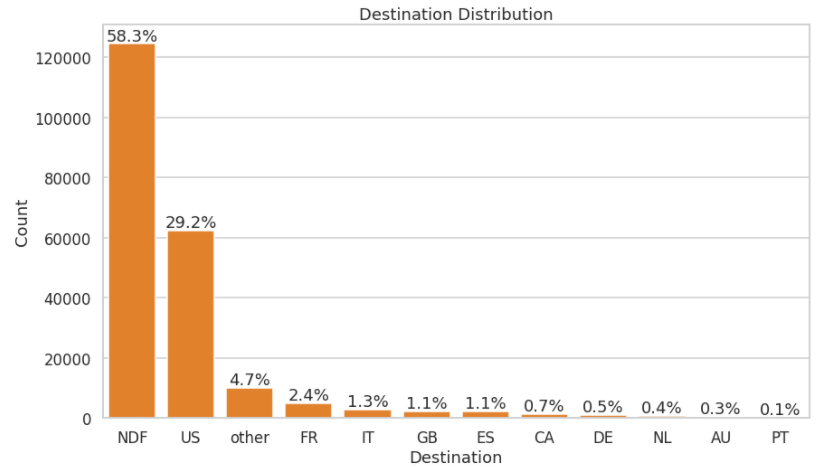
.

위의 feature들을 category type으로 변환한다. Categorical data는 반복된 데이터를 코드화하여 데이터 사이즈를 줄여서 메모리 사용량이 줄고, 처리속도가 빨라진다고 한다.

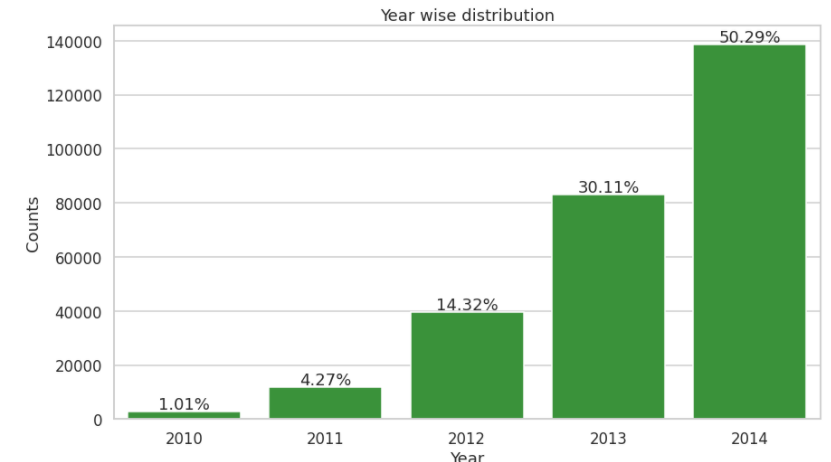


날짜 데이터들을 datetime type으로 변환한다.

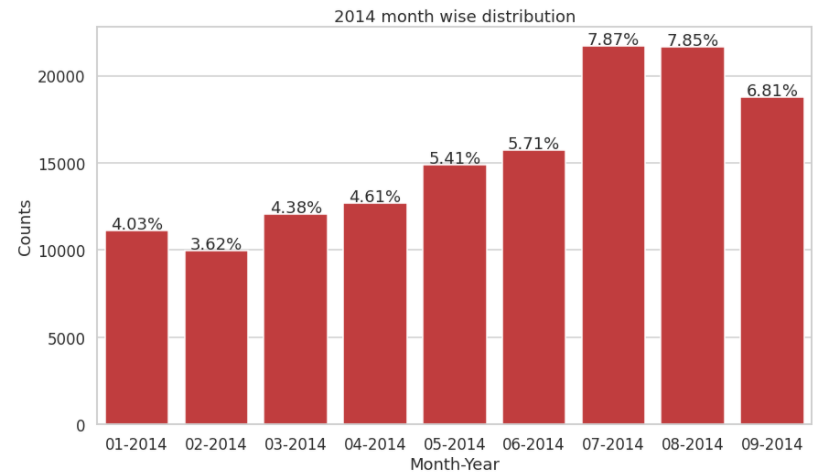
**Visualizations**성별분포를 확인한 결과, 성별을 알리지 않은 경우가 가장 많았고, 여성의 비율이 남성 비율보다 많다는 것을 확인할 수 있다.



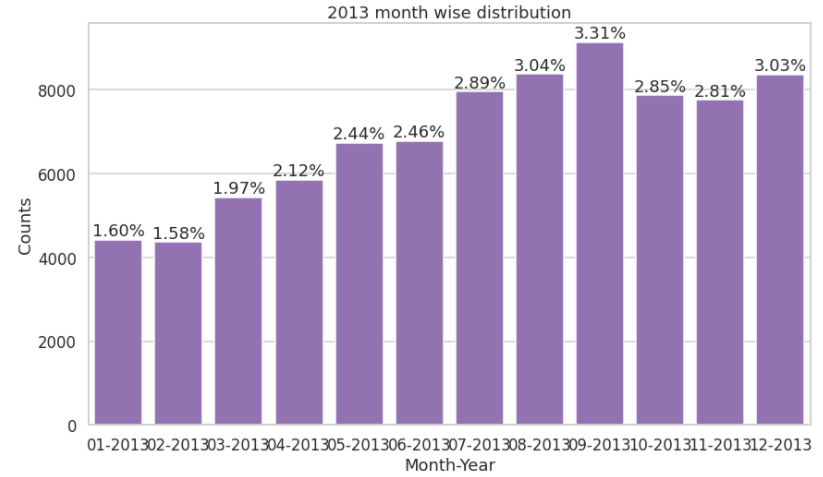
Train data의 destination 분포이다. 대부분은 NDF였는데 이는 예약을 하지 않음을 의미한다. Other은 예약은 했지만, destination list에 포함되지 않은 경우이다. 그래서 이 분포를 확인해본 결과, NDF, US가 각각 58.3%, 29.2%로 가장 많아서 클래스 불균형 문제가 있을 것이다.



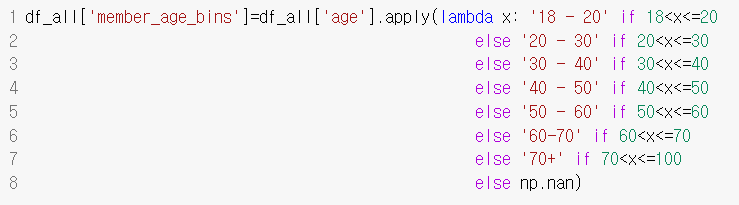
연도별 데이터 개수이다. 시간이 지날수록 데이터 개수들이 많아진다는 것을 확인할 수 있었는데, 점점 user가 증가한다고 볼 수 있지만, 2012년 이전에는 데이터수가 현저히 적음을 확인할 수 있다.

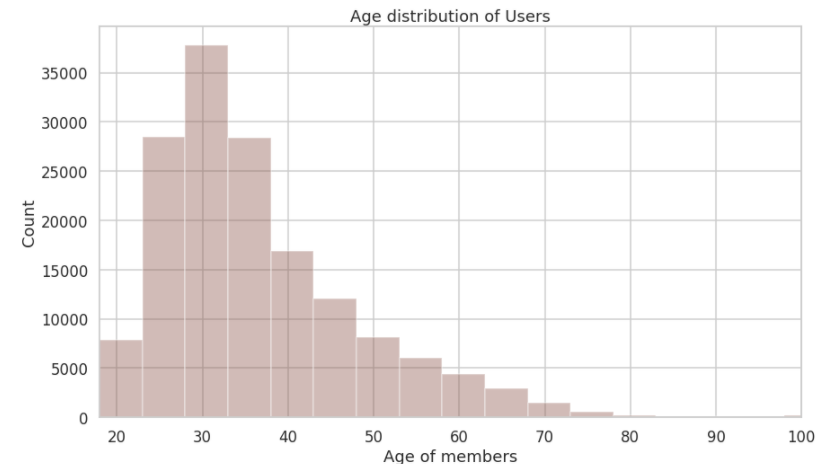


2014년의 월별 분포이다. 위의 년도 별 분포와 같이 시간이 지남에 따라 데이터 개수가 증가함을 확인할 수 있다. 그리고 2014년 9월까지만 있는 것으로 보아 이 데이터셋은 2014년 9월까지의 데이터셋임을 알 수 있다.

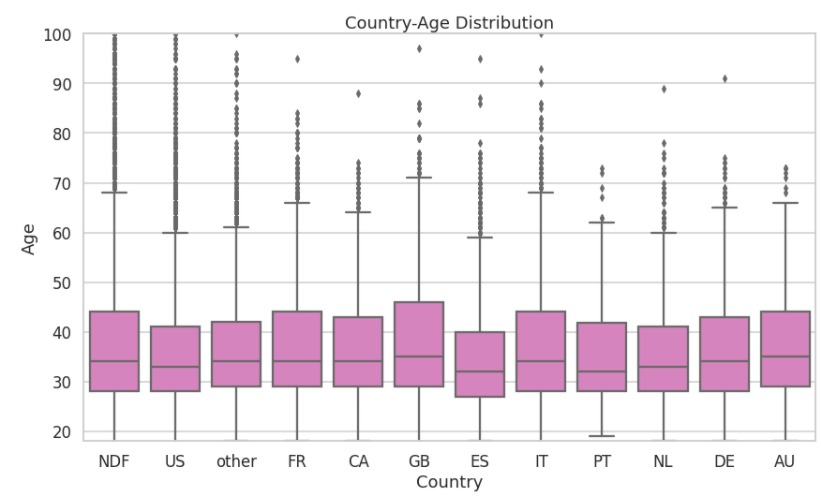


2013년의 월별 분포이다. 이도 시간이 지남에 따라 데이터 개수가 증가하는데, 2014년과 같이 7,8,9월에 가장 많은 것을 보아 7,8,9월에 사용자들이 많은 것인지, 아니면 단순히 시간이 지남에 따라서 사용자들이 늘어난 것인지 확인해 볼 필요가 있어 보인다.

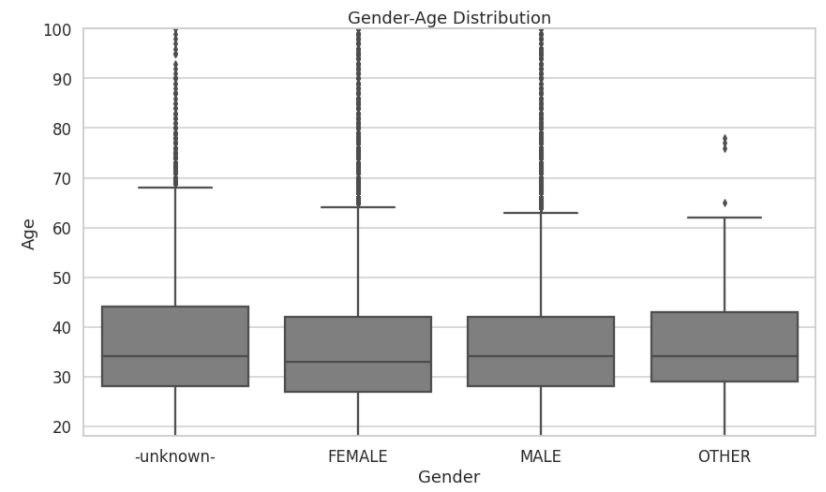


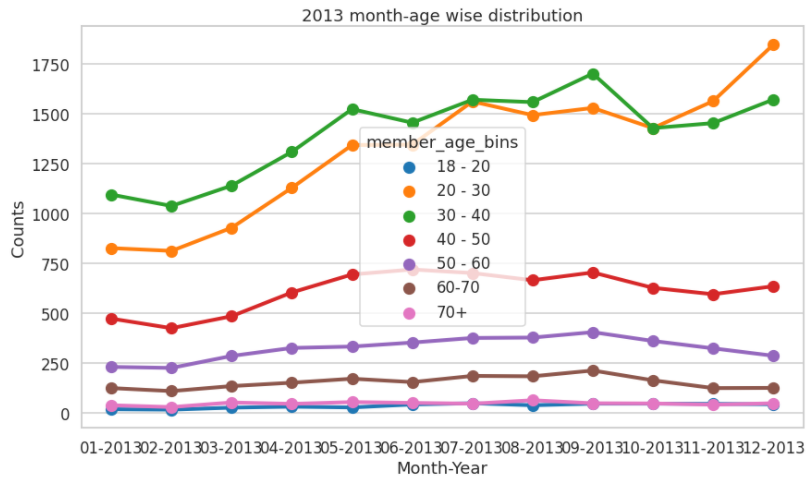
이는 나이대를 10대, 20대, 30대 이렇게 categorical 하게 바꾸며, 18세 이하와, 100세 이상은 nan으로 처리한다.

나이 분포를 보아 사용자들의 나이는 대부분 30대임을 확인 할 수 있고, 그 다음으로 20대 40대가 두번째로 같음을 볼 수 있었다.

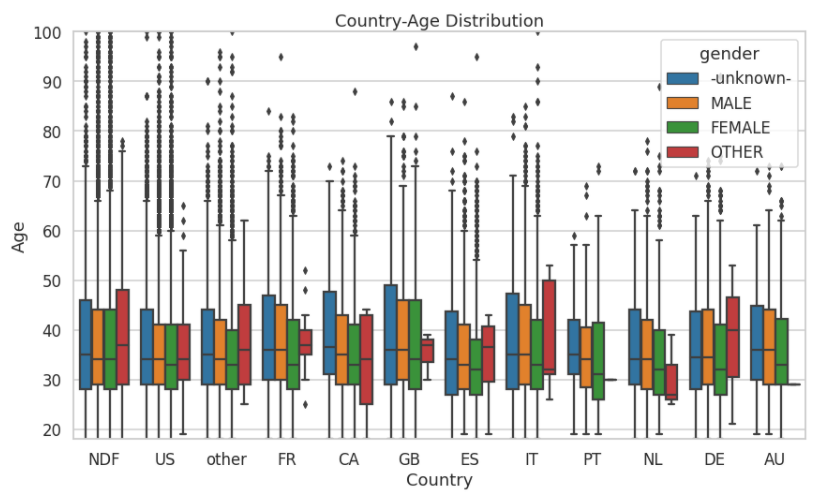


이는 도시와 나이 관계를 boxplot으로 시각화 한 것이다. NDF에서 이상치가 아주 촘촘히 있는 것을 확인할 수 있는데, 이는 예약을 하지 않았기 때문에 나이 정보를 제대로 입력하지 않은 것으로 예측된다.



이는 성별과 나이를 boxplot으로 시각화 한 것이다.

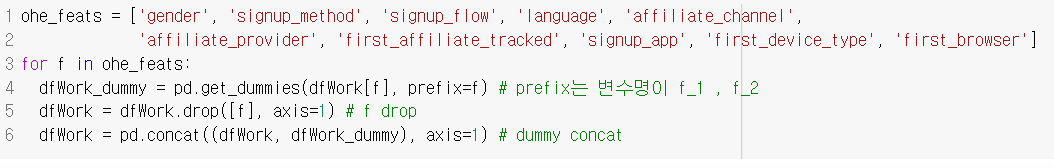
원래 사용자들은 30대가 가장 많았지만, 2013년 10월 이후 20대가 30대를 추월함을 확인할 수 있다. 이는 20대와 30대가 대부분의 사용자가 될 것으로도 예측할 수 있다.

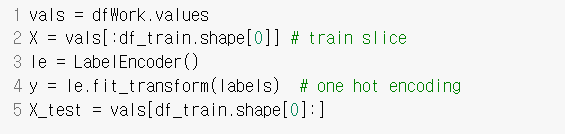


이는 도시와 나이, 성별을 boxplot으로 시각화 한 것이다. 지금까지 시각화된 자료들을 보니, NDF와 age이상치들, 성별을 알리지 않은 사용자들 간의 관계 있어 보인다. 예를 들어 ‘난 예약 안 할 거니까, 개인정보를 정확히 입력하지 않을 거야’ 와 같은 경우도 있을 것이고, 개인 정보에 예민한 사람들도 있을 것으로 예측된다.

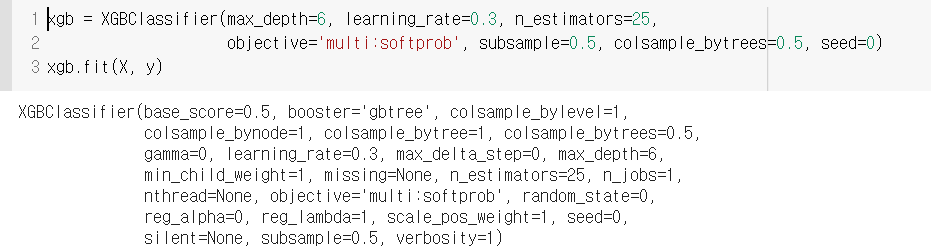
**Computation for the booking destination**

****

이는 train data를 전처리 하는 것으로, nan data를 -1로 채우기, 날짜 데이터를 int자료형으로 바꾸기, 나이 이상치(age<14, age>100)들을 -1로 변환한다.이는 categorical data를 one-hot encoding 하는 것으로 get\_dummies에서 prefix = f 는 one-hot encoding의 column명을 f\_1, f\_2와 같은 형식으로 encoding한다.



이는 train, test를 split하고, y를 one-hot encoding한다.



이는 XGBoost classifier를 사용하는 것으로 사용한 파라미터는

max\_depth: 트리의 깊이 수

learning\_rate: 학습률

n\_estimaors: 생성할 weak learner 수

objective: 목적함수 multi:softprob은 다중분류, 확률 반환

subsample: data중 sampling

colsample\_bytrees: column 중 sampling

이렇게 사용되었다.

**XGBoost(eXtra Gradient Boost)**

|  |  |
| --- | --- |
| 항목 | 설명 |
| 뛰어난 예측 성능 | 분류와 회귀에서 예측 성능이 뛰어남 |
| GBM 대비 빠른 수행시간 | GBM에 비해 시간이 빠름(Parallel Computing) |
| 과적합 규제(Regularization) | XGBoost 자체에 Regularization 있어 과적합에 강함 |
| Tree pruning | GBM은 tree pruning이 negative loss가 발생하면 멈춘다. 하지만 XGBoost는 Max\_depth까지 진행한 후 lossFunction에서 개선이 threshold에 못미치면 역방향 pruning진행함 |
| 자체 내장 교차검증 | CV function이 내장되어 있음 |
| 결측치 자체 처리 | 결측치를 내부적으로 처리함 |

참고: <https://medium.com/@peteryun/ml-kaggle%EC%97%90-%EC%A0%81%EC%9A%A9%ED%95%B4%EB%B3%B4%EB%8A%94-xgboost-f1650342ba93>, 파이썬 머신러닝 완벽가이드