패럿 6기 ML 프로젝트

1. Dataset

Korea Income and Welfare

- id
- year : study conducted
- wave : from wave 1st in 2005 to wave 14th in 2018
- region: 1) Seoul 2) Kyeong-gi 3) Kyoung-nam 4) Kyoung-buk 5) Chung-nam 6) Gang-won &. Chung-buk 7) Jeolla & Jeju
- income: yearly income in M KRW(Million Korean Won. 1100 KRW = 1 USD)
- family_member: no. of family members
- gender: 1) male 2) female
- year_born
- education_level: 1) no education(under 7 yrs-old) 2) no education(7 & over 7 yrs-old) 3) elementary 4) middle school 5) high school 6) college 7) university degree 8) MA 9) doctoral degree
- marriage: marital status. 1) not applicable (under 18) 2) married 3) separated by death 4) separated 5) not married yet 6) others
- religion: 1) have religion 2) do not have
- occupation: this will be provided in separated code book
- company_size
- reason*none*worker

  \1) no capable 2) in military service 3) studying in school 4) prepare for school 5) preprare to apply job 6) house worker 7) caring kids at home 8) nursing 9) giving-up economic activities 10) no intention to work 11) others

특이사항

- 이 데이터셋은 동일한 사람을 일정 주기로 임금을 분석한 데이터입니다. 따라서 id가 동일한 데이터가 존재합니다.
  - 하지만 id가 같다고 해도 조건이 다르기 때문에(나이, 직업 등) 다른 사람으로 취급하고 분석을 진행하시면 됩니다.
- Training dataset은 약 7만개의 데이터이며, 시간이 오래걸릴 수 있습니다.
- Training dataset에는 income이 포함되어 있고, 여러분이 예측해야 할 Test dataset에는 income이 포함되어 있지 않습니다. 모델을 통해 income을 예측하시고, txt나 csv파일로 뽑아낸 뒤, 각 멘토에게 보내주시면, 적절한 평가지표, 혹은 시각적인 그래프를 통해 결과를 확인해드립니다.

2. 발표준비

중간발표 5/20

- 주피터 노트북을 이용해 코드 위주로 발표를 하셔도 좋고, ppt를 만드셔도 좋습니다.
- 다만 중간 발표는 힘을 빼고, 이런 저런 시도를 해보았다 정도만 해도 충분 합니다.

최종발표 5/27

- 최종발표도 중간발표와 마찬가지로 발표 방법은 자유입니다.
- 다만, 이 때는 모델에 대한 상세한 설명을 부탁드립니다.
    - 특정 데이터의 전처리의 이유, 특정 모델의 사용 이유 등
    - 학습을 진행하면서 세운 가설과 그 결과

  등 다방면으로 데이터를 뜯어보시면 좋을 것 같습니다.