KNOW기반 직업 추천 알고리즘 경진대회

2021.12.06~2022.01.28
데이콘 주최
Url https://dacon.io/competitions/official/235863/overview/description

목차

1. 대회&데이터 설명

1. 대회&데이터 설명

KNOW(한국직업정보) 재직자 조사 -

한국 고용원이 청소년과 성인의 진로 및 경력설계, 진로상담, 구인, 구직 등에 도움을 주기 위해 2001년부터 개발, 운영하고 있는 조사

KNOW(한국직업정보) 데이터를 기반으로 직업 추천 모델을 만들고 직업과 연관성 높은 직무능력 탐색 발굴

목적

- KNOW 설문 데이터셋을 활용한 직업 추천 알고리즘 개발
- 직업과 연관이 높은 문항 분석 및 영향 변수 발굴

1. 대회&데이터 설명

갖고 있는 데이터

- 2017년 train&test dataset
- 2018년 train&test dataset
- 2019년 train&test dataset
- 2020년 train&test dataset

Test dataset은 직업 분류 코드(label)인 knowcode 열만 존재하지 않는다.

갖고 있는 참조 자료
- 각 년도 설문지
- 각 년도 변수값
- 각 년도 변수정보

2. EDA

2. EDA

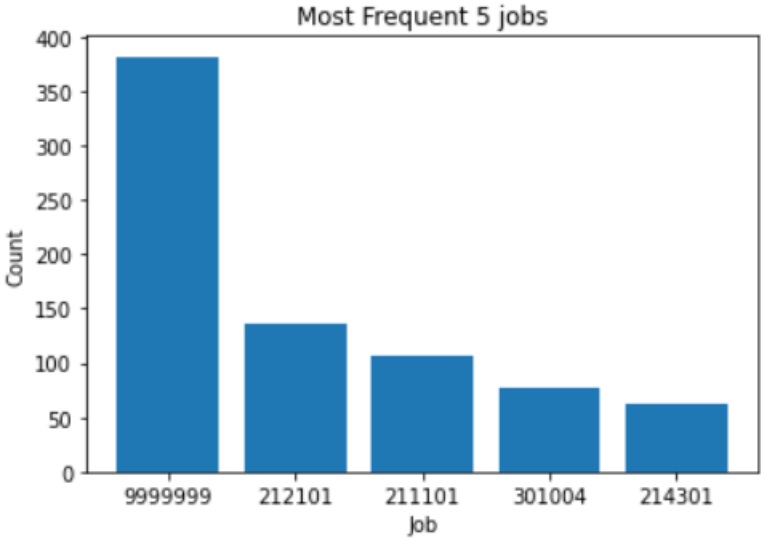
2017년에 대해서만 보이겠다.

데이터 타입 분포

```
학습용 데이터 type 분포
int64      94
float64    50
object     11
dtype: int64
```

2. EDA

가장 많이 등장한 직업 유형



9999999 : 해당년도 코드값이 부여되지 않은 직업이다.

212101 : 중, 고등학교 교사

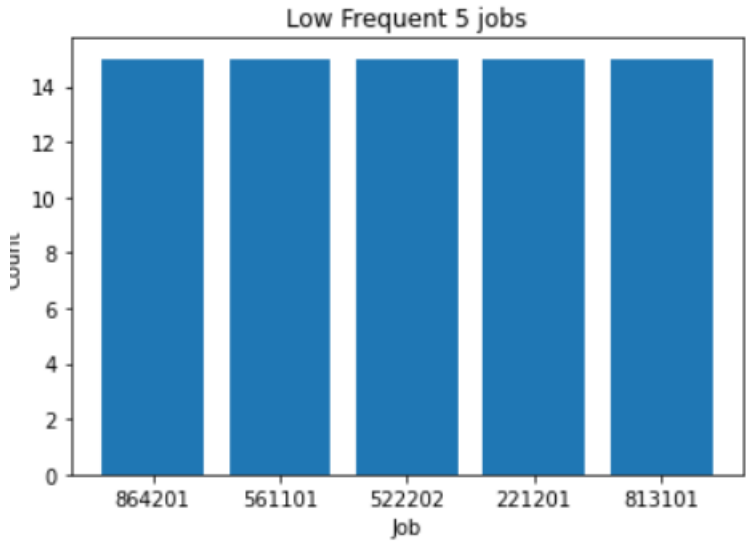
211101 : 판사

301004 : 간호사

314301 : 기술, 기능계 강사

2. EDA

가장 적게 등장한 직업 유형



864201 : 신발 제조기계 조작원

561101 : 청소원

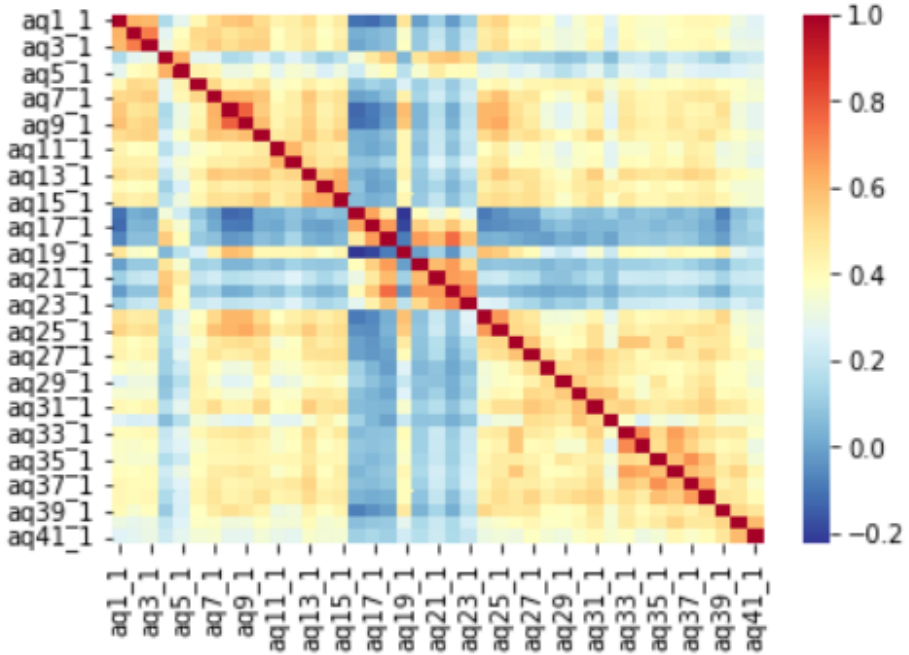
522202 : 열차 객실 승무원

221201 : 대학 시간 강사

813101 : 금형원

2. EDA

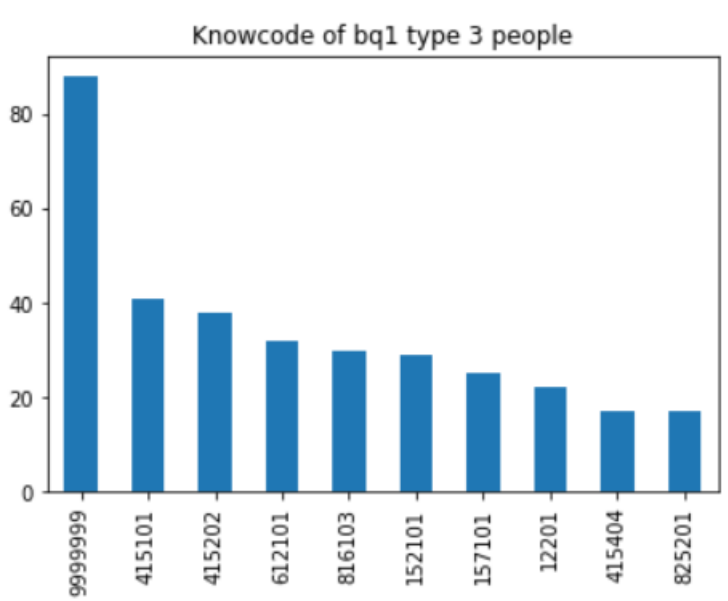
Aq 설문지 문항 응답간 상관관계



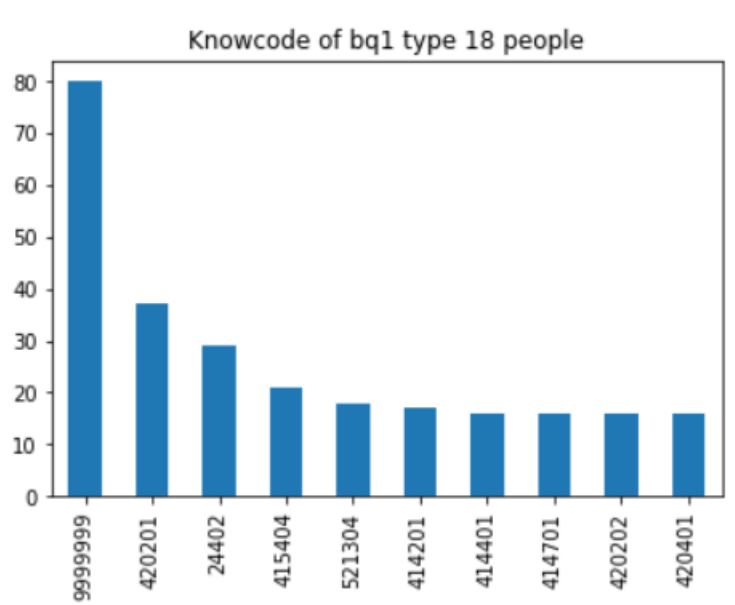
2. EDA

Bq1 선택에 따른 직업코드 분포 파악

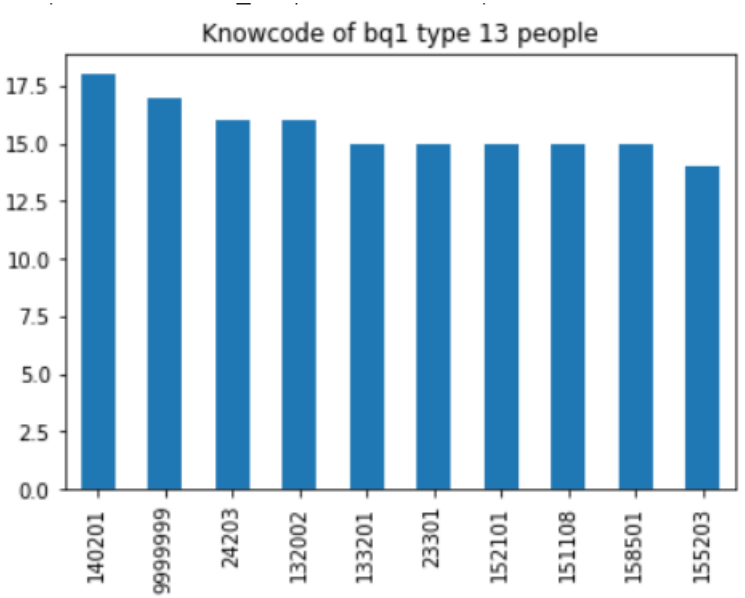
-bq1을 3을 선택하였을 때



-bq1을 18을 선택하였을 때



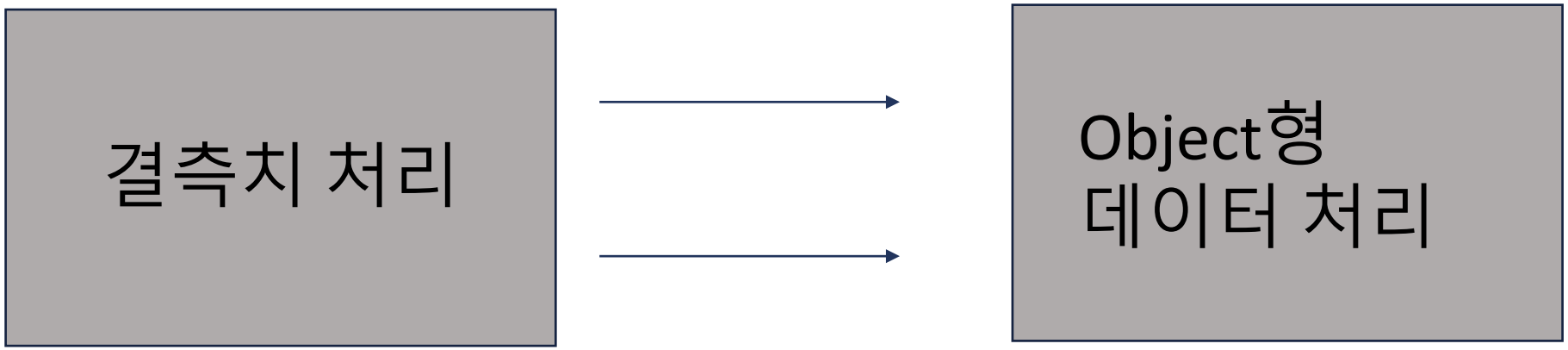
-bq1을 13을 선택하였을 때



Bq1 응답에 따른 직업코드의 분포가 확실히 다른 것을 알 수 있었다.
따른 bq 문항에 대해서도 위와 같은 EDA를 실행하였다.

3. EDA

3. 전처리

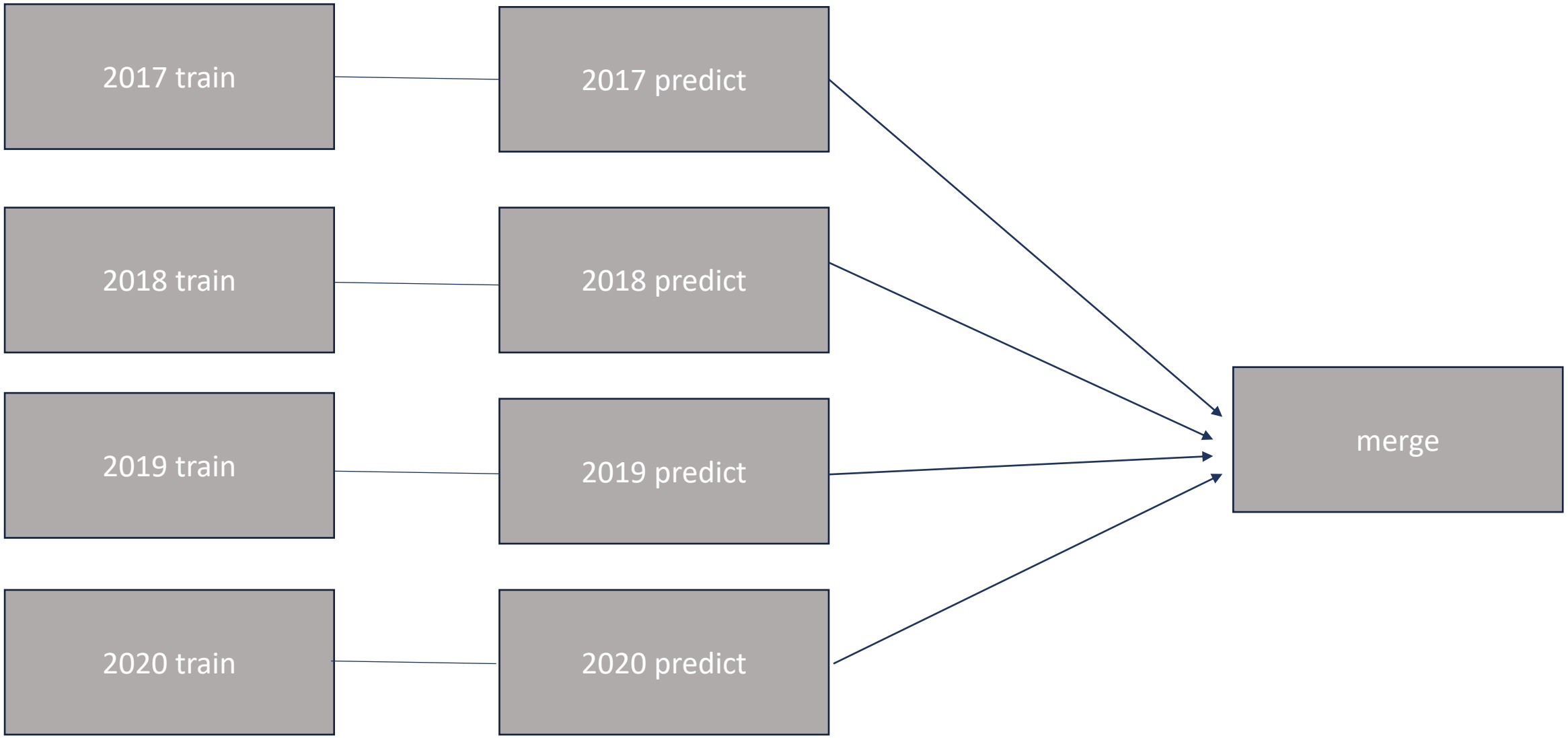


각 문항의 성격에 따른 적절한
전처리를 수행

Label Encoding

4. 모델링

4. 모델링



4. 모델링

Categorical Feature 예측을 잘하는 CatBoostClassifier 활용

Parameter

- interations=500, 1500, 1500, 1500
- Random state=123
- 나머지 - default

아쉬운 점

- 텍스트 문항들을 NLP로 처리하여 보았지만 그냥 Label Encoder로 한 것만 못했다.
- 더 나은 NLP로 하였으면 더 좋은 결과가 나왔을 것
- 모델링 시에 RF와 Catboost만 활용해보고 딥러닝 모델을 시도 못해보았다.