**1번 파일**

1. Import

필요한 패키지 다운.

2. Read Data

필요한 데이터를 불러온다. / Panel 데이터에서 ‘REGION’ 컬럼은 무의미해서 삭제

3. Feature Create

Questions\_count = Panel 데이터 안에서 전체 질문에서 응답한 문항 비율을 구하는 피쳐

이후 train을 2021-02-01을 기준으로 before\_train, train으로 분리

Points\_earned = 유저ID별 응답한 설문의 CPI의 합을 구해본다(before\_train기준)

TIME데이터들을 통해 DAY, WEEKDAY, HOUR, MONTH 피쳐 추가

Points\_earned\_mean = 유저ID별 응답한 설문의 CPI의 평균을 구해본다(before\_train기준)

TITLE = TITLE 정규표현식 피쳐를 사용

TYPE\_mean = TYPE별 응답확률

g\_i = 성별 별 IR 평균

Response\_Probability = 응답한 유저의 응답률

ID\_(IR,LOI,CPI)\_mean = 유저ID별 응답한 IR, CPI, LOI 의 평균

STATUS\_ratio\_( IR,LOI,CPI,HOUR) = IR,LOI,CPI,HOUR별 응답률 평균

CATEGIRIES = CATEGORIES 열 라벨인코딩

Trend\_M = 달간격으로 추이 기울기를 만들고 각 userID당 응답수를 가중치로 두어 곱해준 피쳐

4. Feature Engineering

무의미한 'userID', 'surveyID', 'TIME' 삭제

수치형 변수의 이상치를 제거해 주고 StandardScaler 실시

범주형 변수는 원핫인코딩 실시

5. Feature Selection

Shap 패키지를 이용해, LGBM모델을 통한 Feature Importance를 탐색.

중요도가 0인 피쳐는 학습에 있어 어떤 영향도 주지 않기에 제외.

6. Modeling

lgbm모델을 통해 학습하여 lgbm.csv 서브미션 생성

RandomForestClassifier를 Out\_of\_fold기법을 활용하여 rf\_oof.csv 서브미션 생성

**2번 파일**

1. Import

필요한 패키지 다운.

2. Read Data

01.2.make\_feature\_4M\_shift\_data.pkl을 활용해 데이터를 불러온다.

3. 데이터 정리

예측을 위한 ID를 따로 저장 후 train, test 를 나눠준다.

4. Feature Engineering

결측치를 범주형과 출생년도(BIRTH)는 최빈값으로, 수치형은 0으로 채워준다.

이후 응답률 추이 파생변수를 추가

수치형 변수의 이상치를 제거해 주고 StandardScaler 실시

범주형 변수는 원핫인코딩 실시

5. Modeling

Lgbm을 통해 new\_lgbm.csv 서브미션을 생성

**3번 파일**

1. Import

필요한 패키지 다운.

2. Read Data

필요한 데이터를 불러온다. / Panel 데이터에서 ‘REGION’ 컬럼은 무의미해서 삭제

3. Feature create

패널들의 질문에 대한 성의 판단을 하기 위해 panel 데이터 프레임에 패널 질문 COUNT feature를 만든다.

train데이터를 '2021-01-01 00:00:00를 기준으로 before train과 train으로 나눠준다.

패널 ID별 누적 포인트를 before train기준으로 넣어준다

train, test데이터를 data로 놓고 TIME 컬럼의 타입을 datetime64로 바꾼 후 시간 관련 feature을 생성한다. 또, 설문제에 대한 정보인 IR, LOI, CPI의 범위별 feature 생성한다.

train데이터의 STATUS를 기준으로 획득 리워드 포인트 feature를 생성한다.

TITLE 정규표현식 피쳐를 활용한다.

TYPE별 STATUS의 평균을 낸 피쳐를 만들고 train, test에 붙여준다.

유저의 아이디 별 응답 확률을 알려주는 응답률 feature를 만들어준다.

CATEGORIES 열을 라벨인코딩해준다.

4. Feature Engineering

무의미한 'userID', 'surveyID', 'TIME' 삭제

수치형 변수의 이상치를 제거해 주고 StandardScaler 실시

범주형 변수는 원핫인코딩 실시

5. Feature Selection

Shap 패키지를 이용해, LGBM모델을 통한 Feature Importance를 탐색.

중요도가 0인 피쳐는 학습에 있어 어떤 영향도 주지 않기에 제외.

6. Modeling

Dnn을 활용해 dnn.csv 서브미션 생성

**앙상블 코드**

위에서 만든 submission들은 e\_submission에 저장돼있음.

e\_submission파일안의 csv를 모두 가져와 산술, 멱, 가중평균을 할 수 있는 코드.