

(주)PMI사 온라인 설문 분류 모델

- 설명: 패널의 정보와 설문 정보 데이터를 통해 어떤 패널이 온라인 설문에 응답할 가능성이 높은지 분류
- 기간: 2021.10 ~ 2021.12
- 사용 언어: python(pandas, matplotlib, seaborn, sklearn, tensorflow, optuna)
- 주관: 국민대 머신러닝 강의
- 비고: A+
- 주요 역할: 데이터 전처리, 모델링

프로젝트 내용

데이터

1. 패널 정보 데이터(약 17,000개)
 - 패널ID, 가입 시 설문조사 데이터(성별, 지역, 직업, 소득, 핸드폰 통신사, 요금제, 자녀 나이 등)
2. 설문 데이터(약 6,000개)
 - 설문ID, 설문 타이틀, 난이도(응답 가능성), 설문 시 걸릴 시간, 패널 응답 리워드 포인트(설문 참여 시 받는 포인트)
3. train(약 4,500,000개)
 - 패널ID, 설문ID, 설문 시기(20.06~21.05), 응답 여부(target)
4. test(약 1,400,000개)
 - 패널ID, 설문ID, 설문 시기(21.06~21.09)

데이터 전처리

1. 설문 조사 데이터 전처리
 - a. 패널의 가입 시 설문조사 데이터는 중복이 허용되는 객관식 문항으로 함수를 정의해 데이터 처리
2. 결측값 처리
 - a. 패널의 가입 시 설문조사 데이터는 개인정보가 다수 포함되어있어 결측값이 많음
 - b. 대부분의 경우 결측값이 많은 질문은 삭제했지만, 설문은 미응답도 의미가 있을 것이라 생각해 0이라는 새로운 값으로 결측값을 처리해 성능 향상에 기여

모델링



































1. XGB

a. optuna를 사용해 하이퍼파라미터 튜닝

DNN

a. 과적합 방지를 위해 1개의 은닉층만 사용

b. kerastuner를 사용해 하이퍼파라미터 튜닝

#	△pub	Team Name	Notebook	Team Members	Score 🏆	Entries	Last
1	—			 	0.89212	74	15m
2	—			  	0.88441	79	3h
3	—	김재민+석원희		 	0.87964	52	4h
4	—				0.87751	49	2m
5	—				0.87653	41	2h
6	—			 	0.87252	44	8h
7	▲ 1			 	0.87133	46	13h
8	▼ 1			  	0.87128	78	2h
9	—			 	0.86931	103	2m
10	▲ 1			 	0.86922	12	3h
11	▼ 1			 	0.86920	34	1h
12	—			 	0.86905	43	2h
13	—			  	0.86846	75	17m
14	—			 	0.86638	25	2h
15	▲ 1			  	0.86510	24	6m
16	▼ 1			 	0.86466	89	2h