**조 설명**

- ㅅㅅㅎㅎ, 조원

**프로젝트 개요**

- 데이콘 심리 성향 예측 AI 경진대회

- 주소 : https://dacon.io/competitions/official/235647/overview/

**변수 설명**

- 변수 설명 파일

**1. 각 변수의분포 및 전처리 어떻게 했는지 왜 그렇게 했는지(그래프)**

- married, : 미혼, 이혼, 기혼 이외의 값이 추가되어 있으므로 (other : 93) 미혼으로 추가

- age\_Group, : 60대 이후의 인원이 현저히 떨어지므로 하나로 묶을 여부 생각

- race : 대표적인 인종을 제외한 나머지 인종을 그룹화

- education : 무응답(528)을 Less than high school으로 추가

**2. 새로 추가한 변수에 대한 설명(age 그룹, mach\_score, Q\_E 그룹)**

- Q\_E그룹 분류 이유

시간 변수는 상대적인 변수로 판단하시면 됩니다

절대적 수치에 신경 x , 일정한 기준을 잡아 오래걸렸다 아니다 등으로 판단

25%까지 : 짧다, 25-75 : 보통, 75 이상 : 길다

- age\_Group 분류 이유

10대 이하의 투표율이 현저히 떨어지므로 10대만 따로 분류/ 생애주기이론에 따른 분류(미성년/성년/중년/노년)

- wr\_new :

실존하는 단어들의 의미를 아는가에 대한 질문에 4가지의 질문만 현저하게 모른다 답변이 많아서 따로 분류해서 변수생성

- wf\_new

허구의 단어들의 의미를 하나라도 안다고 체크한 여부를 새로운 컬럼으로(신뢰도와 관련)

- tp : 각 tp 를 그룹화하여 대표적인 성향을 분류 -> 코드 설명

- mach\_score 설명(reverse)

- race\_white : 백인과 백인 아닌 사람들을 구분했을 때가 가장 상관계수가 높았다.

**3. 선별한 변수와 그 이유 -> 상관계수 표 (버린 거 버린 이유)**

선별한 변수

education, age\_group, married, wr\_, race(race\_white), tp\_score, mach\_score

**4. 선정한 머신러닝 종류(catboost, lgbm)와 머신러닝 선정 이유, 각각 어떤 변수를 선별했는가, 정확도는 얼마인가**

Catboost

- 공식사이트 : https://catboost.ai

- 범주형 피처를 지원하는 Gradient Boosting 오픈소스 라이브러리

- 오버피팅 감소, 범주형 피처 지원, 사용자 친화적인 API 인터페이스

- 기존 부스팅 모델 느린 학습속도와 오버피팅 문제 해결이 가능한 모델.

- 그리디언트 부스팅의 구현, 기본예측기로는 의사결정나무를 사용한다.

- 범주형 변수를 처리하는데 다른 알고리즘보다 효과적, 하지만 대부분의 수치형인 경우는 큰 효과를 못볼 가능성이 있다.

- XGBBoost, Light LGBM, Catboost -> catboost 사용

- 모형 사진 첨부

- 기존의 부팅방법은 모든데이터에서 잔차를 일괄계산 했다면, catboost는

1. 먼저 x1 의 잔차만 계산하고, 이를 기반으로 모델을 만든다. 그리고 x2 의 잔차를 이 모델로 예측한다.

2. x1, x2 의 잔차를 가지고 모델을 만든다. 이를 기반으로 x3, x4 의 잔차를 모델로 예측한다.

3. x1, x2, x3, x4 를 가지고 모델을 만든다. 이를 기반으로 x5, x6, z7, x8 의 잔차를 모델로 예측한다.

4. ... 반복

- catboost 성과

앙상블모델(pycaret)

- pycaret 공식 홈페이지 : <https://pycaret.org/>

- 여러 개의 모델들의 예측 결과들을 이용한다면 더 정황산 예측값을 구할 수 있다.

- 여러 개의 결정 트리를 결합하여 하나의 결정 트리보다 더 좋은 성능을 내는 머신러닝기법. 여러 개의 약 분류기를 결합하여 강 분류기를 만드는 것, 정확성이 향상된다.

- 한줄의 코드로 머신 러닝 모델을 교육하고 배포하기 위한 오픈소스 머신러닝 라이브러리

- 간단하고 사용하기 쉽다

**5. 느낀점 및 한계점**

- 데이터 자체의 수 부족, 마키아벨리즘 테스트 자체의 신빙성 부족, 기본지식의 한계(공부해야하지만 시간이 없다), 전처리에 따라 모델 적합도가 달라져서 점수가 달라진다! 모델에 맞는 전처리 필요

- 활용방안 : 범주형 데이터에서는 부스팅 방법을 사용하는 것이 좋다.

범주형 데이터 전처리의 다양한 방법이 있다. -> 기업의 고객조사를 통한 데이터 분석에 활용가능, 마케팅 수단으로 활용가능!

**6. 참고사항**

- 깃헙 주소 : <https://github.com/Chachaeul/TeamProject>

문성민씨에게 전화