

# Dacon Analysis Study

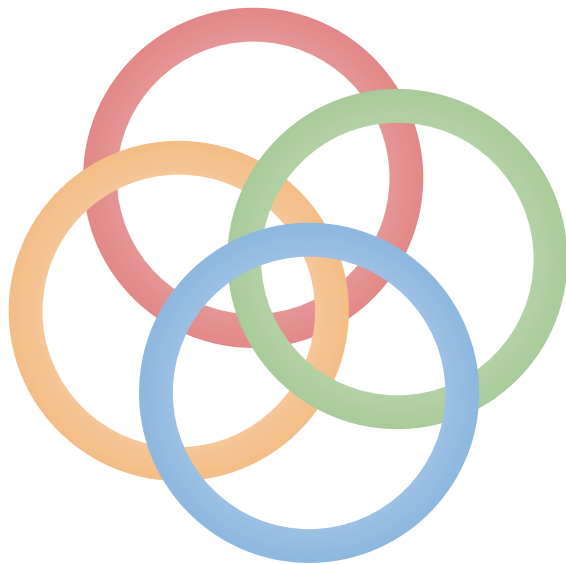
Chap 4) 심리 성향 예측 AI 모델

17 강신현


17 김건우

17 송원진


17 신도현



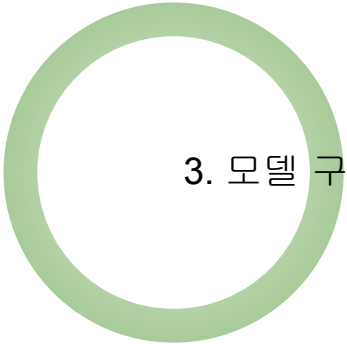
# 목차



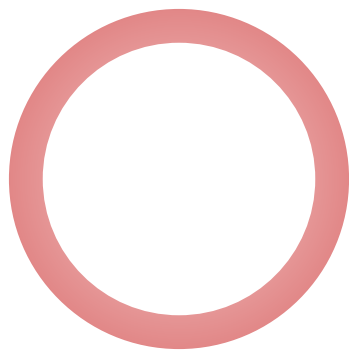
1. 자료 분석



2. 전처리, EDA (데이터 시각화)



3. 모델 구축



## 1. 자료 분석

## 주어진 데이터 살펴보기

- train.csv : 훈련용 데이터
  - shape : (45532, 78)

```
Index(['QaA', 'QaE', 'QbA', 'QbE', 'QcA', 'QcE', 'QdA', 'QdE', 'QeA', 'QeE',  
      'QfA', 'QfE', 'QgA', 'QgE', 'QhA', 'QhE', 'QiA', 'QiE', 'QjA', 'QjE',  
      'QkA', 'QkE', 'QlA', 'QlE', 'QmA', 'QmE', 'QnA', 'QnE', 'QoA', 'QoE',  
      'QpA', 'QpE', 'QqA', 'QqE', 'QrA', 'QrE', 'QsA', 'QsE', 'QtA', 'QtE',  
      'age_group', 'education', 'engnat', 'familysize', 'gender', 'hand',  
      'married', 'race', 'religion', 'tp01', 'tp02', 'tp03', 'tp04', 'tp05',  
      'tp06', 'tp07', 'tp08', 'tp09', 'tp10', 'urban', 'voted', 'wf_01',  
      'wf_02', 'wf_03', 'wr_01', 'wr_02', 'wr_03', 'wr_04', 'wr_05', 'wr_06',  
      'wr_07', 'wr_08', 'wr_09', 'wr_10', 'wr_11', 'wr_12', 'wr_13'],  
      dtype='object')
```

---

# Index 살펴보기 (Public)

(Q\_A: /Q\_E(a~t) : 질문을 답할 때까지의 시간)

**Qb:** 범죄자들과 일반 사람들 사이의 가장 큰 차이점은 범죄자들은 잡힐 만큼 어리석다는 것이다.

**Qc:** 다른 누군가를 전적으로 신뢰하는 사람은 문제를 요구하고 있습니다.

**Qe:** P. T. Barnum이 매 분마다 멍청이가 태어난다고 주장한것은 틀렸다.

**Qf:** 다른 사람에게 거짓말을 하는 것에는 변명의 여지가 없다.

**Qh:** 대부분의 사람들은 재산을 잃는 것보다 부모님의 죽음을 더 쉽게 잇는다.

**Qj:** 모든 사람들이 악랄한 경향을 가지고 있으며, 기회가 주어질 때만 나타난다고 생각하는 것이 가장 안전하다.

**Qk:** 영향력 있지만 정직하지 못한 것보다는, 겸손하고 정직한 것이 더 낫다.

**Qm:** 편법을 쓰지 않고는 출세(성공)하기 어렵다.


**Qo:** 사람들을 다루는 가장 좋은 방법은 그들이 듣고 싶어하는 말을 들려주는 것이다.

**Qq:** 사람들은 기본적으로 착하고 친절하다.

**Qr:** 오직 도덕적으로 옳은 것이 확실할 때만 행동을 취해야 한다.

**Qs:** 중요한 사람들에게 아침하는 것은 현명하다.

# Index 살펴보기 (Private)




Qa Qd Qg Qi

Ql Qn Qp Qt

> 비식별화를 위해 **secret** 문항 처리

# 마키아벨리즘



<https://dacon.io/competitions/official/235647/codeshare/1711?page=1&dtype=recent&ptype=pub>

# Index 살펴보기

1: 전혀 동의하지 않는다 / 2: 약간 동의하지 않는다 / 3: 보통이다 / 4: 약간 동의한다 / 5: 완전 동의한다

age\_group: 연령

education: 교육수준 (1: 고등교육 미만 / 2: 고등학교 졸업 / 3: 학사 / 4: 석사 / 0: 무응답)

engnat: 모국어가 영어 (1: 그렇다 / 2: 아니다 / 0: 무응답)

familysize: 형제자매 수

gender: 성별 (Male, Female)

hand : 필기하는 손 (1: 오른손잡이 / 2: 왼손잡이 / 3: 양손잡이 / 0: 무응답)

married: 혼인여부 (1: 미혼 / 2: 기혼 / 3: 이전에 결혼 / 0: 기타)

race: 인종 (Asian, Arab, Black, Indigenous Australian, Native American, White, Other)

religion: 종교 (Agnostic, Atheist, Buddhist, Christian\_Catholic, Christian\_Mormon,  
Christian\_Protestant, Christian\_Other, Hindu, Hewish, Muslim, Sikh, Other)

urban: 유년기의 거주 구역

1=Rural (country side), 2=Suburban, 3=Urban (town, city), 0=무응답



# Index 살펴보기 (tp\_\_(01~07))

Q. 나는 나 자신을 \_\_\_\_ 하다고 생각한다.

tp01: 외향적이고 열정적이다.

tp02: 비판적이고, 다투기 좋아한다.

tp03 : 의지할수 있고, 자기 훈련이 되어있다.

tp04 : 불안하고 쉽게 기분이 상한다

tp05 : 새로운 경험과 문제에 열린 마음을  
가졌다.

tp06: 내성적이고, 조용하다.

tp07 : 동정심있고 따뜻하다.

tp08 : 체계적이지 않고 부주의하다.

tp09: 차분하고 정서적으로 안정되어있다.

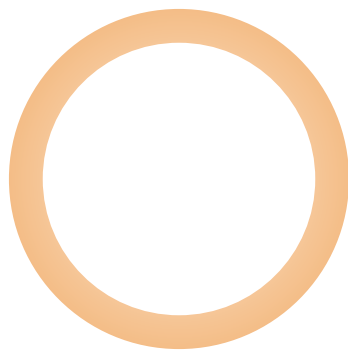
tp10: 보수적이고, 창조적이지 않다.

urban: 고향 (1: 시골 / 2: 도시외곽 / 3: 도시 / 0: 무응답)

wr(01~13): 실존하는 단어의 정의를 알고있다.

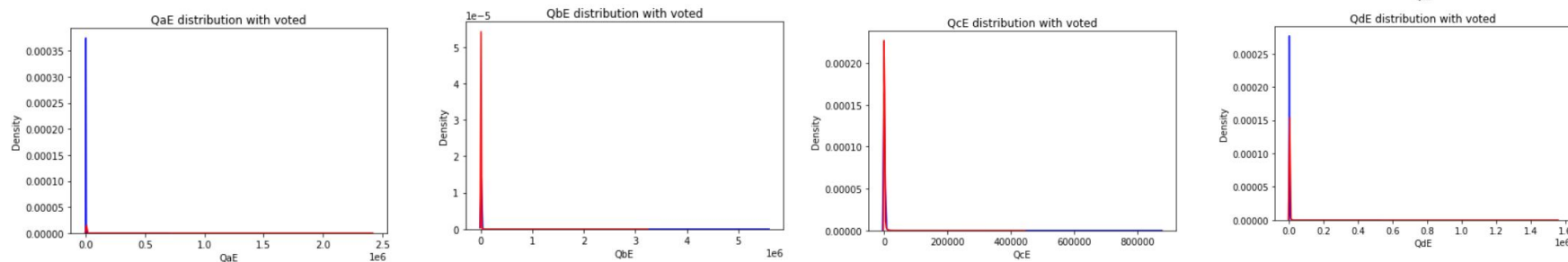
wf(01~03): 허구인 단어의 정의를 알고있다.

voted: 작년 국가선거 참여 여부 (1: 참여 / 2: 미참여)



## 2. 전처리 (pre-processing), EDA( 데이터 시각화)

# 데이터: Answer\_time 데이터의 치우침 문제



Answer\_time의 index별 데이터를 plot 한 이미지이다.

모든데이터(x축)에 비해서 값이 0인 데이터의 밀도가 상당히 크게,  
지나치게 치우쳐져 있음을 확인할 수 있다.

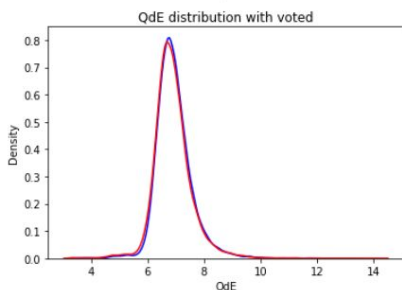
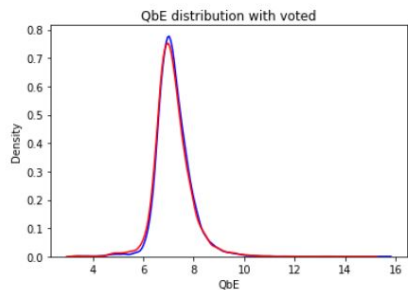
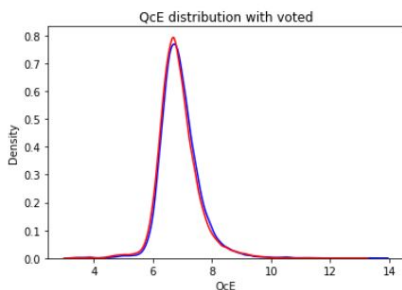
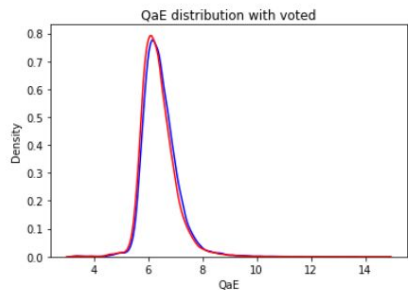
=> 데이터값에 로그를 취하여 문제를 해결하여본다

## 데이터: Answer\_time 데이터의 치우침 문제해결: Log

```
[ ] log_Answer_time = train[Answers_time_only].copy()
    log_Answer_time[Answers_time_only] = np.log1p(train[Answers_time_only])
    train[Answers_time_only] = log_Answer_time
```

```
▶ log_Answer_time_t = test[Answers_time_only].copy()
  log_Answer_time_t[Answers_time_only] = np.log1p(test[Answers_time_only])
  test[Answers_time_only] = log_Answer_time_t
```

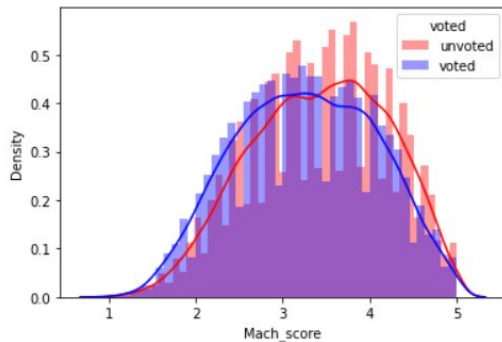
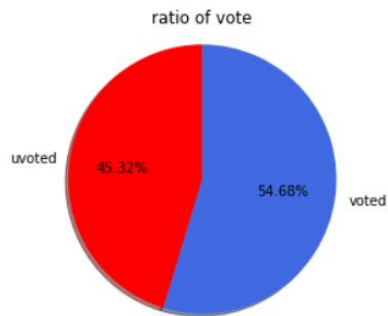
# 데이터: Answer\_time 데이터의 치우침 문제해결: Log

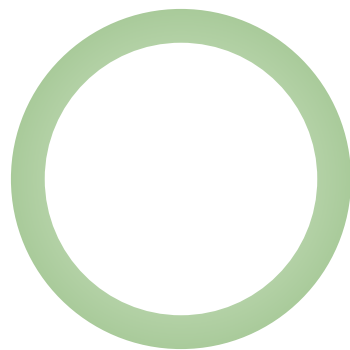


데이터에 로그를 취한뒤 plot하였을때, 치우침 문제가 개선되었음을 확인할 수 있었다.

# EDA: 데이터 시각화

각각의 데이터 **feature**들을 시각화하고, 분포를 관찰하여본다. (코랩으로!)





### 3. 모델 구축

# XGBoost 선정 이유 (장점)



1. 병렬 처리 사용 -> 빠른 학습, 분류 속도
2. 좋은 유연성 -> 평가 함수를 포함한 다양한 최적화 옵션 제공
3. greedy - algorithm을 사용한 자동 가지치기 -> 과적합 (Overfitting)  
이 잘 일어나지 않음
4. 다른 알고리즘과의 좋은 연계 활용성



# XGBoost 기본 원리

- 부스팅 기술 (Boosting or Additive Training)
- 약한 분류기를 세트로 묶어서 정확도 예측

ex)

1. 학습기 M에 대하여 Y를 예측할 확률 :  $Y = M(x) + \text{error1}$

2. 학습기 G에 대하여 Y를 예측할 확률 :  $Y = G(x) + \text{error2}$

3. 학습기 H에 대하여 Y를 예측할 확률 :  $Y = H(x) + \text{error3}$

-> 1 에 2,3 적용 :  $Y = M(x) + G(x) + H(x) + \text{error4}$

-> 학습기 M을 단독으로 사용했을 때보다 높은 정확도

# XGBoost 개선사항

M, G, H 각각 분류기의 성능이 다른데, 모두 같은 비중을 두고 있음

$$Y = 1 * H(x) + 1 * G(x) + 1 * H(x) + \text{error4}$$

-> 임의의 x에 대하여 서로 간섭하여 오류를 높이는 결과를 낼 수 있음

-> 각 모델 앞에 비중(weights)을 두어 해결

$$Y = a * H(x) + b * G(x) + c * H(x) + \text{error4}$$



## DACON Evaluation System



Thank you for your work.

수고하셨습니다.

제출 완료 되었습니다.

순위가 가장 높은 파일로 선택이 변경되었습니다.

Your file has been submitted successfully.

OK

## 심리 성향 예측 AI 경진대회

월간 데이콘 8 | 심리 테스트 분석 | AUC | 분류

🏆 상금 : 100만원+애플워치

🕒 2020.09.28 ~ 2020.11.16 17:59 [+ Google Calendar](#)

👥 1,485팀 📅 마감



참여중

[대회안내](#)[데이터](#)[코드 공유](#)[토론](#)[대회문의](#)[리더보드](#)[제출](#)[PUBLIC](#)[PRIVATE](#)[AWARDS](#)[RANKING CHART](#)[순위기준](#)

● WINNER ● 1% ● 4% ● 10%

#	팀	팀 멤버	점수	제출수	등록일
352	또또신		0.77184	3	몇 초 전
1	1996		0.78632	151	2달 전
2	YoungHoonShin		0.78471	62	4달 전
3	harryjo97		0.78412	133	6달 전

