

심리성향예측

심리학 테스트 분석
알고리즘 개발

7조 안수빈 배정민 최지원 안도현



01

데이터 소개
및 분석 목적

02

데이터
전처리

03

모델링

04

결과 및
한계점



데이터 소개

train.csv : 관측치 45532개 78개의 변수 test.csv : 관측치 11382개 77개의 변수(voted외)



여러 인적 사항을 나타내는 변수

age_group

연령
(7 factors)

religion

종교
(12 factors)

gender

성별
(2 factors)

race

인종
(7 factors)

familysize

형제 자매 수

hand

손 쓰임
(3+1 factors)

education

교육 수준
(4+1 factors)

urban

유년기 거주 구역
(3+1 factors)

engnat

모국어 영어 여부
(2+1 factors)

married

혼인 상태
(3+1 factors)

※ +1 : 무응답



voted : 지난 해 국가 선거 투표 여부(2 factors, target)

데이터 소개



“마키아벨리즘”, “TIPI”에 대한 정보를 나타내는 변수

Q_A (a~t)

마키아벨리즘 테스트
문항 별 점수



Agree

Disagree

Q_E (a~t)

마키아벨리즘 테스트
문항 별 응답 시간
(상대적 시간 변수)

tp__ (01~10)

5가지 성격 요소 파악을 위한
10 항목 검사 문항



Agree

무응답

Disagree

wr_00 (01~13)

실존 단어의 뜻을
아는지에 대한 답변
(2 factors)

VCL01 : boat
VCL02 : incoherent
...
VCL16 : funny

wf_00 (01~03)

허구 단어의 뜻을
아는지에 대한 답변
(2 factors)

VCL06 : cuivocal
VCL09 : florted
VCL12 : verdid

데이터 소개

데이터의 기록된 정보들은 크게 2가지 심리 테스트에 대한 답변으로 기록

마키아벨리즘

- ▶ 국가의 발전을 위해 어떠한 수단이나 방법도 허용된다는 국가 지상적인 정치이념
- ▶ 여기에서 파생된 마키아벨리즘 테스트(Mach-IV)는 개인적인 욕구의 충족을 위해 남을 속이거나 조종하려는 욕구의 정도를 평가하는 테스트

성향이 높음



타인과의 소통에서 계산적이고 신중하게 접근하는 경향

성향이 낮음



타인과의 소통에서 개인적이며 감정이 이입된 접근하는 경향

데이터 소개

Mach-IV 질문들은 “요령이 없다면 앞으로 나아가기 힘들다”
“다른 이를 완벽히 믿는 어떤 사람은 문제를 제기한다”
...
“도덕적으로 올바른 때만 행동 해야한다”

등 총 20개가 존재

이때 질문 중 reverse  되는 항목이 존재

V+ : Qb, Qf, Qj, Qm,
V- : Qe, Qq

T+ : Qc, Qo, Qs
T- : Qf, Qr

M+
M- : Qk

데이터 소개

5가지 성격 특성 요소

- ▶ 심리학에서 경험적인 조사와 연구를 통해 정립한 성격 특성의 다섯 가지 주요한 요소

개방성(Openness to experience)
성실성(Conscientiousness)
외향성(Extraversion)
우호성(Agreeableness)
신경성(Neuroticism, 정서적 불안정성)

10항목 성격 검사(TIPI)

- ▶ 위 5가지 성격 특성 요소에 근거해 10가지 문항으로 짧은 시간에 성격을 파악할 수 있는 테스트
- ▶ 개방성, 성실성, 외향성, 우호성과 정서적 안정성에 대한 성격 파악

“활발하고 적극적이다”
“비판적이며 논쟁을 좋아한다”

외향성 +
친화성 -



Mach-IV처럼 reverse되는 항목이 있으며
각 특성 당 2가지 문항으로 구성

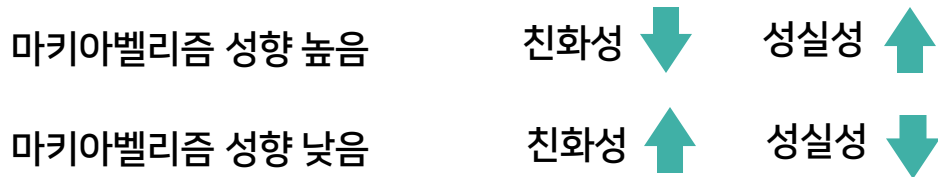
데이터 소개

연구에 따르면 마키아벨리즘이 높은 사람은 자신의 이익을 위해 상대를 배반하는 모습을 보임

즉 마키아벨리즘이 높은 사람은 이익을 위해 타인을 쉽게 기만하고 배신한다고 볼 수 있음

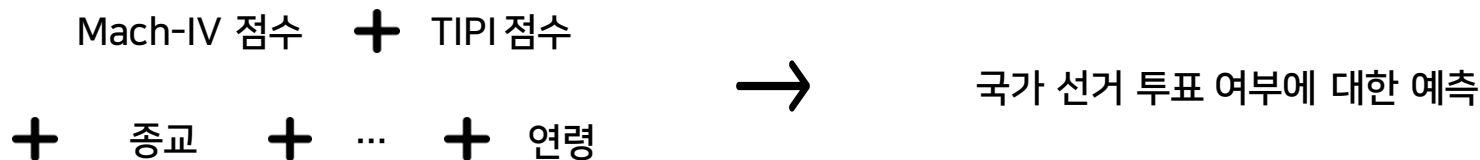
반대로 마키아벨리즘이 낮은 사람은 친화적이며 타인의 의견에 잘 순응함

또한 마키아벨리즘이 높은 사람은 계산적이기 때문에 자신의 이익을 위해 오히려 성실하게 행동한다고 볼 수 있음



데이터 소개

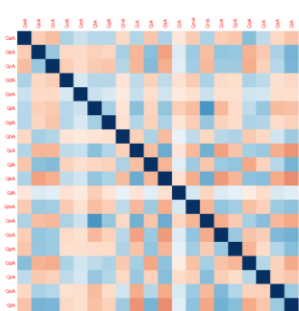
심리 테스트에 담겨있는 개인의 특성에 따라 국가 선거 투표 여부가 영향을 받을 것
따라서 마키아벨리즘 심리 테스트를 통해 투표여부에 대한 예측을 수행함



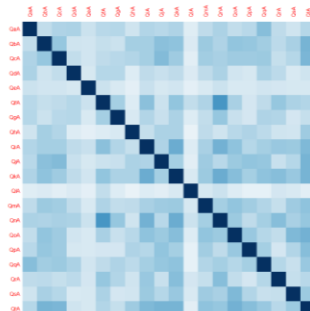
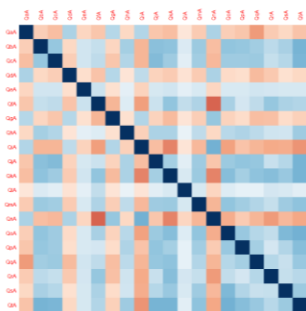
마키아벨리즘 성향이 높고, 친화성 수치가 낮으며 성실성 수치가 높은
20대 석사 학위의 미혼 남성이 국가 선거 투표를 할까?

데이터 전처리 파생변수생성

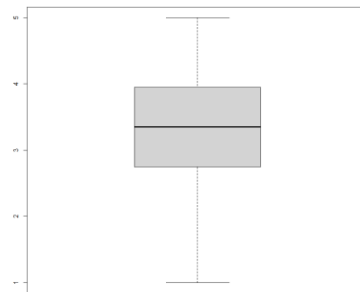
- (1) 20개의 마키아벨리즘 테스트 답변(Q_A)으로부터 마키아벨리즘 정도를 측정할 수 있는 변수 Mach_score 생성
- 각 문항마다 점수가 그대로 들어가는지, 역으로 들어가는지 설정돼있고 파악할 수 있는 변수는
- 점수를 역으로 계산, Secret 질문들은 점수들의 상관성을 파악해 reverse함



reverse 처리



Secret질문
reverse 처리



데이터 전처리 파생변수생성

- Q_A 문항
- Qb : 범죄자들과 다른 일반 사람들의 가장 큰 차이점은, 범죄자들이 잡힐 만큼 멍청하다는 것이다.
- Qc : 다른 이를 완벽히 믿는 어떤 사람은 문제를 제기한다.
- Qe : P.T. Barnum이 매 분마다 선천적으로 잘 속는 사람이 태어난다라고 말한 것은 틀렸다.
- Qf : 다른 이에게 거짓말하는 것은 변명할 여지가 없다.
- Qh : 대부분의 사람들은 그들의 재산을 잃은 것보다 부모의 죽음을 더 빨리 잃는다.
- Qj : 모든 사람은 악한 경향이 있고, 기회만 주어진다면 그 성향이 표현된다고 가정하는 것이 가장 안전하다.
- Qk : 대체적으로 중요하고 정직하지 못한 것보다 겸손하고 정직한 것이 더 낫다.
- Qm : 요령이 없다면 앞으로 나아가기 힘들다.
- Qo : 사람을 다루는 가장 좋은 방법은 그들이 듣고 싶은 말을 해주는 것이다.
- Qq : 대부분의 사람들은 기본적으로 선하고 친절하다.
- Qr : 도덕적으로 올바른 때만 행동해야 한다.
- Qs : 중요 인물(상관, 정부의 요인 등)에게 아첨하는 것은 현명하다

Qa, Qd, Qg, Qi, Ql Qn, Qp, Qt 질문은 secret 처리

데이터 전처리 파생변수생성

(2) Tipi 항목검사를 통해 5가지 성격 특성요소에 관한 변수 5개 생성

번호	영어	한글	특성	+/-
1	Extraverted, enthusiastic.	활발하고 적극적이다.	외향성	+
2	Critical, quarrelsome.	비판적이며 논쟁을 좋아한다.	친화성	-
3	Dependable, self-disciplined.	믿음직스럽고 자기 관리를 잘한다.	성실성	+
4	Anxious, easily upset.	걱정이 많고 쉽게 흥분한다.	정서적 안정성	-
5	Open to new experiences, complex.	새로운 경험에 개방적이며 복잡다단하다.	경험 개방성	+
6	Reserved, quiet.	내향적이며 조용하다.	외향성	-
7	Sympathetic, warm.	동정심이 많고 다정하다.	친화성	+
8	Disorganized, careless.	계획적이지 않으며 덤벙댄다.	성실성	-
9	Calm, emotionally stable.	차분하며 감정 기복이 적다.	정서적 안정성	+
10	Conventional, uncreative.	변화를 싫어하며 창의적이지 못하다.	경험 개방성	-

데이터 전처리 파생변수생성

(2) Tipi 항목검사를 통해 5가지 성격 특성요소에 관한 변수 5개 생성



O (개방성)

- 점수 : $(5\text{번 점수} + (8 - 10\text{번 점수})) / 2$

C (성실성)

- 점수 : $(3\text{번 점수} + (8 - 8\text{번 점수})) / 2$

E (외향성)

- 점수 : $(1\text{번 점수} + (8 - 6\text{번 점수})) / 2$

A (우호성)

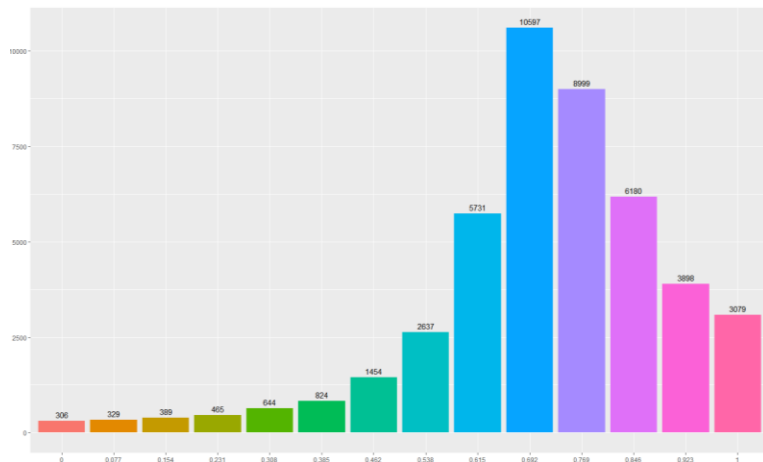
- 점수 : $(7\text{번 점수} + (8 - 2\text{번 점수})) / 2$

N (정서적 안정성)

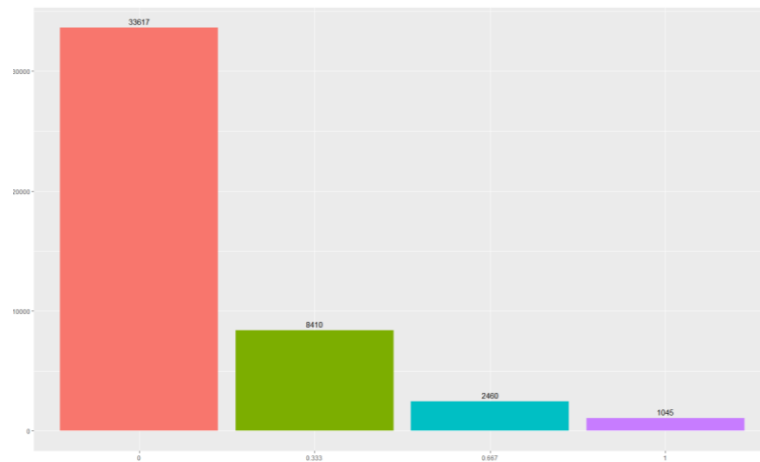
- 점수 : $(9\text{번 점수} + (8 - 4\text{번 점수})) / 2$

데이터 전처리 파생변수생성

(3) 실존/허구 단어에 대한 응답 평균 점수 변수 wr_mean, wf_mean 생성



wr_mean



wf_mean

데이터 전처리 파생변수생성

(4) Q_A 변수의 질문 특성에 따라 변수를 생성함

Tactic : 인간관계 전술과 관련된 변수 → QcA QfA QoA QrA QsA 평균

ex) 사람을 다루는 가장 좋은 방법은 그들이 듣고 싶은 말을 해주는 것이다.

Views : 가치관과 관련된 변수 → QbA QeA QhA QjA QmA QqA 평균

ex) 대부분의 사람들은 기본적으로 선하고 친절하다.

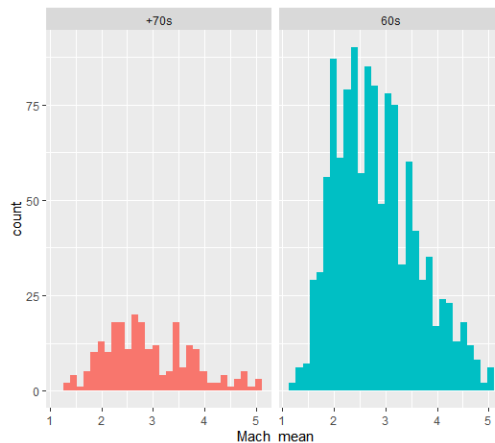
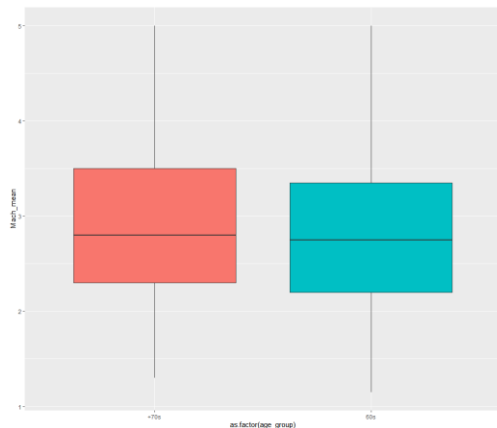
Morality : 도덕과 관련된 변수 → QkA

ex) 대체적으로 중요하고 정직하지 못한 것 보다 겸손하고 정직한 것이 더 낫다.

(5) QaE~QtE까지의 시간 변수를 응답자별로 평균내어 QE_Mean 변수로 생성함

범주형 변수의 재범주화

(1) age_group 재범주화



- ▶ 70대 이상의 빈도수가 적음
- ▶ 60대와 70대의 Mach_score 등 수치형 변수들과의 관계에서 큰 차이가 없어 두 범주를 병합

범주형 변수의 재범주화

(2) religion 재범주화



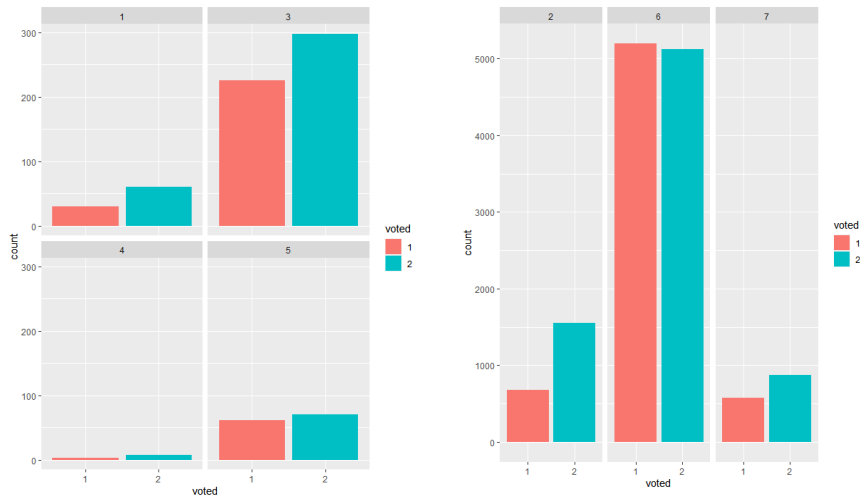
▶ 종교 범주가 많다고 판단하여 재범주화 진행

▶ Mach_score 등 수치형 변수들과의 관계에서 큰 차이가 없어,

Christian의 4가지 종교 병합 여부, 개체수가 적은 종교와 other 병합 여부 총 두가지 경우를 모두 관찰

범주형 변수의 재범주화

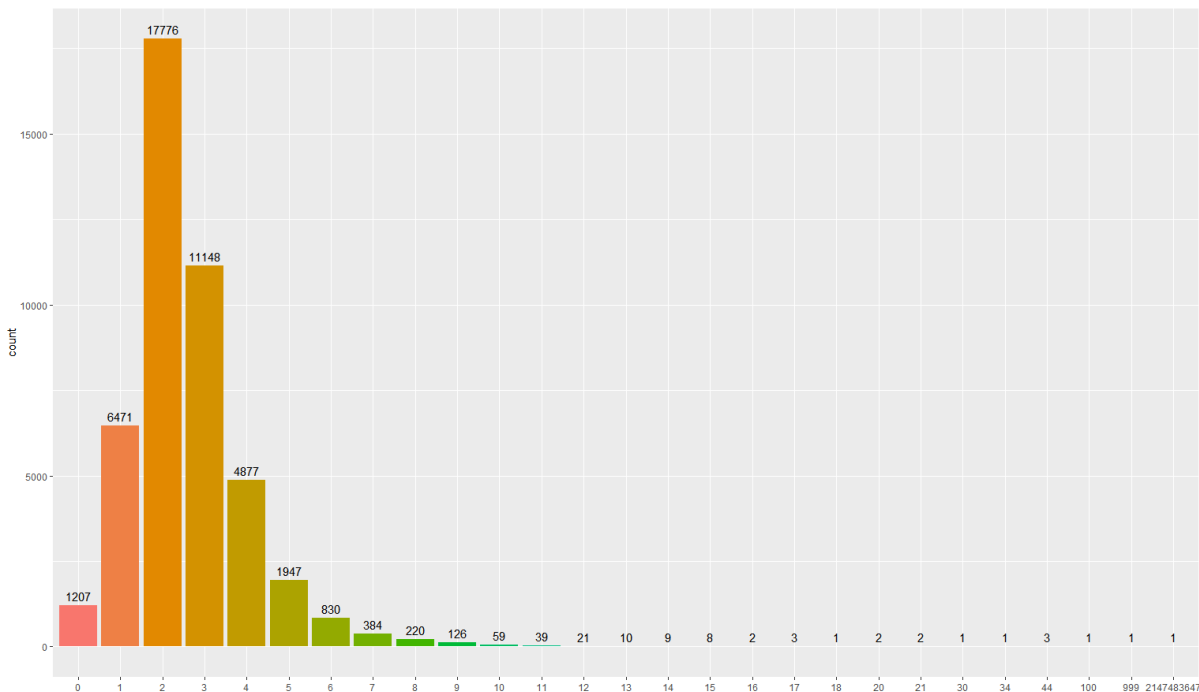
(3) race 재범주화



- ▶ Arab, Indigenous Australian, Native American 빈도수가 적음
- ▶ Mach_score 등 수치형 변수들과의 관계에서 큰 차이가 없어, Black의 Other로의 병합 여부에 따른 성능 비교 관찰

이상치 및 결측치 처리

(1) familysize 이상치 제거 ▶ 16명 이상인 행 제거 (percentile 0.995까지 사용 0.05제거)



이상치 및 결측치 처리

(2) engnat, hand, married, urban의 무응답은 최빈값으로 대체함

▶ engnat	0 (무응답)	→	1(yes)
▶ hand	0 (무응답)	→	1(right)
▶ married	0 (무응답)	→	1(never mind)
▶ urban	0 (무응답)	→	2(suburban)

(3) tp01~ tp10 변수의 무응답은 중앙 값인 3=Neither agree nor disagree로 대체

이상치 및 결측치 처리

(4) age_group & education

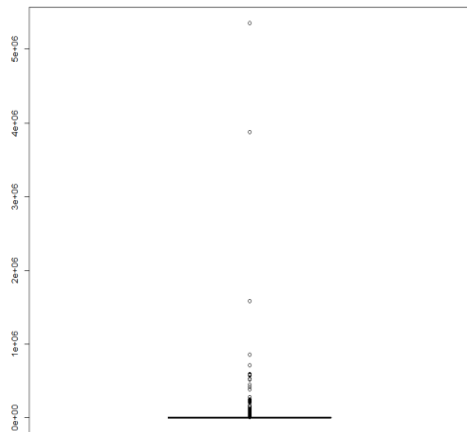
	무응답	중졸	고졸	학사	석사
10대	114	4,534	8,144	1,109	172
20대	133	204	4,131	6,922	2,602
30대	111	155	1,907	3,357	2,221
40대	77	80	1,273	2,106	1,483
50대	59	48	800	1,131	836
60대 이상	31	22	419	470	480



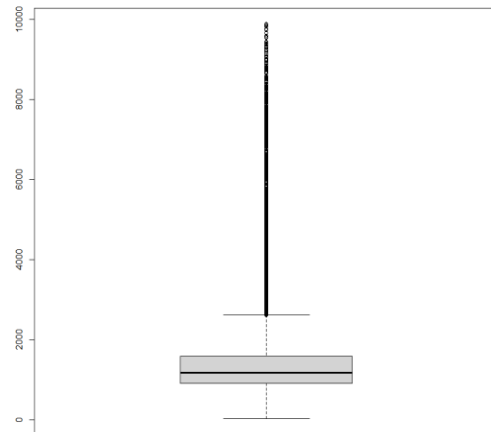
	중졸	고졸	학사	석사
10대	4,995	9,078	0	0
20대	204	4,131	7,055	2,602
30대	155	1,907	3,468	2,221
40대	80	1,273	2,183	1,483
50대	48	800	1,191	836
60대 이상	22	419	470	511

- ▶ 10대 무응답 인원 및 학/석사 인원 중졸:고졸 비율로 수정
- ▶ 나머지 그룹 무응답 인원 최빈값으로 대체

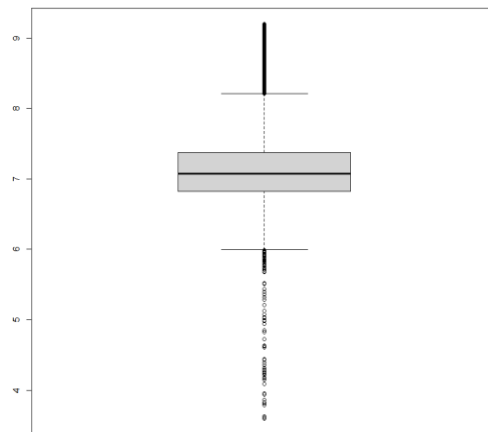
(5) 시간변수 이상치 처리



전처리 전 QE_Mean



이상치를 **중위수**으로
대체한 후의 QE_Mean



Log scale 처리 후의
QE_Mean

최종 변수 정리

Train

age_group	race	urban
education	religion	wr_mean
engnat	Op	wf_mean
familysize	Co	Mach_score
gender	Ex	Tactic
hand	Ag	Views
married	Ne	Morality

Target : **voted**

Test

age_group	race	urban
education	religion	wr_mean
engnat	Op	wf_mean
familysize	Co	Mach_score
gender	Ex	Tactic
hand	Ag	Views
married	Ne	Morality

모델링 소개

(1) LightGBM

- ▶ Tree 기반의 Gradient Boosting 프레임워크
- ▶ Leaf-wise 알고리즘을 활용하여 속도가 빠름
- ▶ GPU 지원

(2) *XGBoost*

- ▶ Tree 기반의 Gradient Boosting 프레임워크
- ▶ LGBM의 과적합 문제를 보완해줌
- ▶ 자체 분할 가능



0.5 : 0.5의 비율로 사용하여 **성능 개선**

TRAIN SET

▶ 재범주화 Case에 따라 성능 비교 후 유의미한 재범주화만 선택

- (1) race " Asian + White + **Black** + Other vs Asian + White + other "
 - ▶ Black(흑인)을 단순히 Other에 병합하기엔 Black의 빈도수가 애매하며 대표 인종으로 볼 수 있는 Black의 구분 또한 필요하기에 두 가지 경우를 체크
- (2) religion " 크리스찬 **병합** vs 크리스찬 **병합 X** "
 - " other **병합** vs other **병합 X** "
 - ▶ EDA상에서(plot확인) 빈도수는 작지만 다른 수치형 변수들 사이의 관계가 어느정도 있을 수 있다고 판단하여 이를 확인
- (3) age_group " 10대인데 투표한 인원 **제거** vs **제거 x** "
 - ▶ 일반적인 투표 연령 제한은 만 18세,
10대 인원 14,079명 중 2,285명이 투표를 했고 그 빈도가 높다고 판단해 이를 체크

성능 차이 **X**

성능 차이 **O**

모델링

TRAIN SET

Train01

Train02


Train03

•
•
•



Train15

Train16

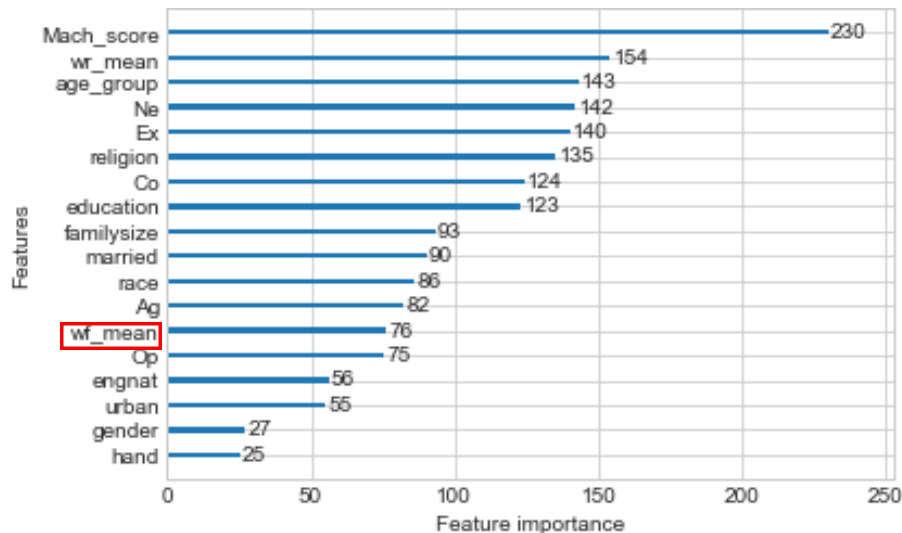
	 LightGBM	<i>XGBoost</i>
race	Asian+White+other	
age_group	10대인데 투표한 인원 제거	
religion - 크리스찬	병합	
religion - other	병합	

→ " Train12 "

TRAIN SET 전처리

(1) wf_mean, hand 제거

Religion 재범주화 후, 변수중요도에 따라 wf_mean을 제거하였음



TRAIN SET 전처리

(2) E(외향성), A(우호성) 제거

O(개방성), C(성실성), N(정서적 안정성), 성별과 투표율의 관계를 발견하여, 투표율과 관련없다고 판단된 E, A 변수를 제거함
또한 전처리 결과, 남성 O의 voted가 살짝 높았고 나머지 여자의 voted가 높았음

(3) O, C, N 재범주화

투표율에 영향을 끼칠 수 있다고 판단된 O, C, N 변수를
대상으로 Positive, Negative, Neutrality 세가지
지표를 활용해 재범주화 진행

$0 < x < 2.5 = \text{"Negative"}$

$2.5 < x < 5.5 = \text{"Neutrality"}$

$5.5 < x = \text{"Positive"}$

O	C	N
6.0	2.7	1.0
3.4	1.3	5.6
...		



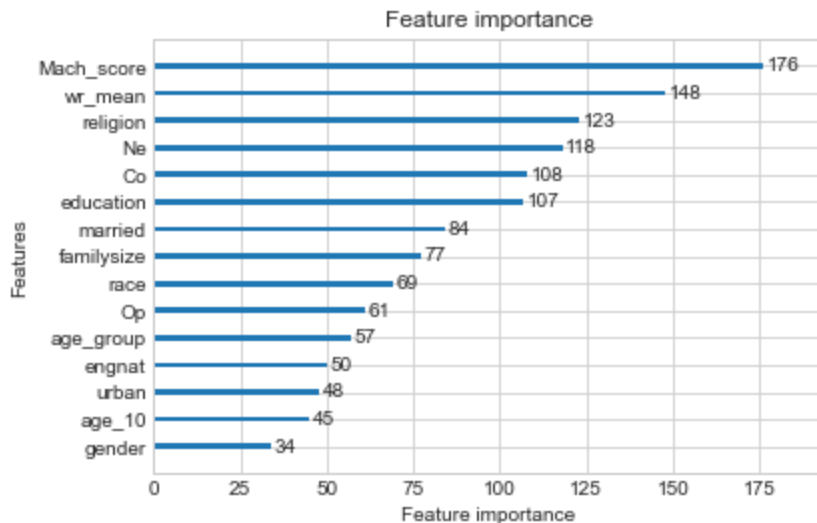
O	C	N
Positive	Neutrality	Positive
Neutrality	Negative	Positive
...		

TRAIN SET 전처리

(4) age_10 변수 생성

age 범주 중, 10대인지 아닌지를 나타내는 이분변수를 추가하였음

변수중요도를 확인해봤을 때, 중요도가 높았음



정확도(accuracy) : 0.7407

Train set score: 0.7554

Test set score: 0.7407

정밀도(precision) : 0.6317

재현율(recall) : 0.8377

F1 점수 : 0.7203

Roc_Auc score: 0.7570

결과

- ▶ AUC (Area under an Roc curve) : 임의의 curve에 대해 그 아래 면적을 계산한 것
AUC 정확도는 높을 수록 좋지만 너무 높으면 과적합이 될 수 있음

LGBM	XGBoost	Test Set
정확도: 0.7407	정확도: 0.7347	public: 0.500174756
정밀도: 0.6317	정밀도: 0.8475	private: 0.4985671192
재현율 : 0.8377	재현율: 0.6789	

→ Valid Set과 Test Set의 AUC 값 차이가 큼

한계점

- ▶ SVM, KNN, 나이브 베이즈 등 여러가지 모델링을 시도하였으나 영향력 높은 변수를 쉽게 알 수 없었고 분석 시간에 너무 많은 시간이 소요돼서 사용하지 못했음
- ▶ 이로 인해 다양한 모델을 혼합하지 못하고 부스팅 모델만을 이용해 새로운 모델을 구축하였음
- ▶ 응답자 대답 회피 문제로 인해 MACH 성향 파악에 한계가 있었음
- ▶ 설문조사의 특성 상 응답을 신뢰하기 어려워 한 사람이 모순된 값들을 가지게 되는 경우가 있었음
- ▶ 마키아벨리즘 테스트 답변이 모두 공개되지 않아 Views, Tactic, Morality의 변수중요도가 낮게 나옴
- ▶ TIPI의 경우 10점이 넘어가면 신뢰성이 떨어진다고 간주하여 제외하였는데 한계치를 넘어가는 문항만 제거를 하였으면 더 높은 결과를 기대할 수 있었을 거 같음
- ▶ Valid set으로 확인한 AUC보다 Test set으로 확인한 AUC가 현저히 낮았음

감사합니다.

7조 안수빈 배정민 최지원 안도현

