심리 성향에 따른
미국 선거 투표여부 예측

6조

최정석 이윤정 윤혜성 양지우

B.a.f

목차

01 데이터 소개

Train Set

45,532명의 관측치

78개의 변수

Index + Phycological + Non Phycological + Target(voted)

Test Set

11,383명의 관측치

77개의 변수

Index + Phycological + Non Phycological

Submission

11,383명의 관측치

2개의 변수

Index + Target (voted)

01 데이터 소개

Phycological Value
(심리 변수)

- 응답자의 심리성향을 파악하기 위한 질문
- Q_A / Q_E / TP__

Non Phycological Value
(비 심리 변수)

- 응답자의 신상정보 파악을 위한 질문
- Gender, Familysize, Race 등…

심리 변수 – Q_A

변수 명	성향	내용
QaA	T+	Secret
QbA	V−	The biggest difference between most criminals and other people is that the criminals are stupid enough to get caught.
QcA	V+	Anyone who completely trusts anyone else is asking for trouble.
QdA	T−	Secret
QeA	V+	P.T. Barnum was wrong when he said that there's a sucker born every minute.
QfA	T+	There is no excuse for lying to someone else.
QgA	V+	Secret
QhA	M+	Most people forget more easily the death of their parents than the loss of their property.
QiA	V−	Secret
QjA	T+	It is safest to assume that all people have a vicious streak and it will come out when they are given a chance.
QkA	V+	All in all, it is better to be humble and honest than to be important and dishonest.
QlA	T−	Secret
QmA	V−	It is hard to get ahead without cutting corners here and there.
QnA	V−	Secret
QoA	T−	The best way to handle people is to tell them what they want to hear.
QpA	T−	Secret
QqA	V+	Most people are basically good and kind.
QrA	T+	One should take action only when sure it is morally right.
QsA	T−	It is wise to flatter important people.
QtA	M−	Secret

- 마키아벨리즘 수치 측정 질문
- QaA ~ QtA 까지 총 20개로 구성
- 범주형 변수 (순서형)

1 : 동의 안함 / 2: 조금 동의안함 / 3: 보통 / 4: 조금 동의 / 5: 동의

심리 변수 – Q_E

- Q_A 질문에 응답하는데 걸린 상대적 시간
  - QaE ~ QtE 까지 총 20개로 구성
    - 수치형 변수

심리 변수 – TP_

변수 명	내용
TP01	Extraverted, enthusiastic
TP02	Critical, quarrelsome
TP03	Dependable, self-disciplined
TP04	Anxious, easily upset
TP05	Open to new experiences, complex
TP06	Reserved, quiet
TP07	Sympathetic, warm
TP08	Disorganized, careless
TP09	Calm, emotionally stable
TP10	Conventional, uncreative

- TIPI 성격 유형 설문에 대한 답변
- TP01~TP10 까지 총 10개로 구성
- 범주형 변수 (순서형)

0 ~ 6 (+ 7 : 무응답)으로 답변

(0 에 가까울수록 해당 항목이 높다는 평가)

비심리 변수

Age_group (연령대)

10s 50s
20s 60s
30s 70s+
40s

Education (교육수준)

- 1 : Less than high school
- 2 : High school
- 3 : University degree
- 4 : Graduate degree
- 0 : 무응답

Engnat (모국어)

- 1 : 영어
- 2 : 비영어
- 0 : 무응답

Familysize (형제자매 수)

수치형 변수

Gender (성별)

- Male
- Female

비심리 변수

Hand (필기하는 손)

- 1 : Right
- 2 : Left
- 3 : Both
- 0 : 무응답

Married (혼인여부)

- 1 : Never married
- 2 : Currently married
- 3 : Previously married
- 0 : Other

Wr__ (실존 단어의 정의를 앎)

Wr01 ~ Wr13

- 1 : 안다
- 0 : 모른다

Wf__ (허구 단어의 정의를 앎)

Wf01 ~ Wf03

- 1 : 안다
- 0 : 모른다

Voted (지난해 선거 참여 여부)

- 1 : 했다
- 0 : 안했다

비심리 변수

Race (인종)

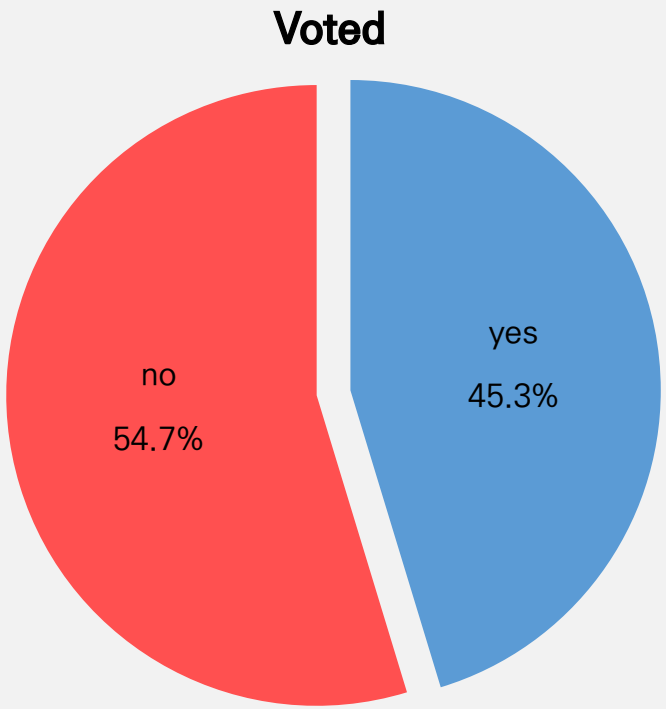
- Asian
- Arab
- Black
- White
- Indigenous Australian
- Native American
- Other

Religion (종교)

- Agnostic
- Atheist
- Buddhist
- Christian_Catholic
- Christian_Mormon
- Christian_Protestant
- Christian_Other
- Hindu
- Jewish
- Muslim
- Sikh
- Other

02 변수 분석 및 전처리

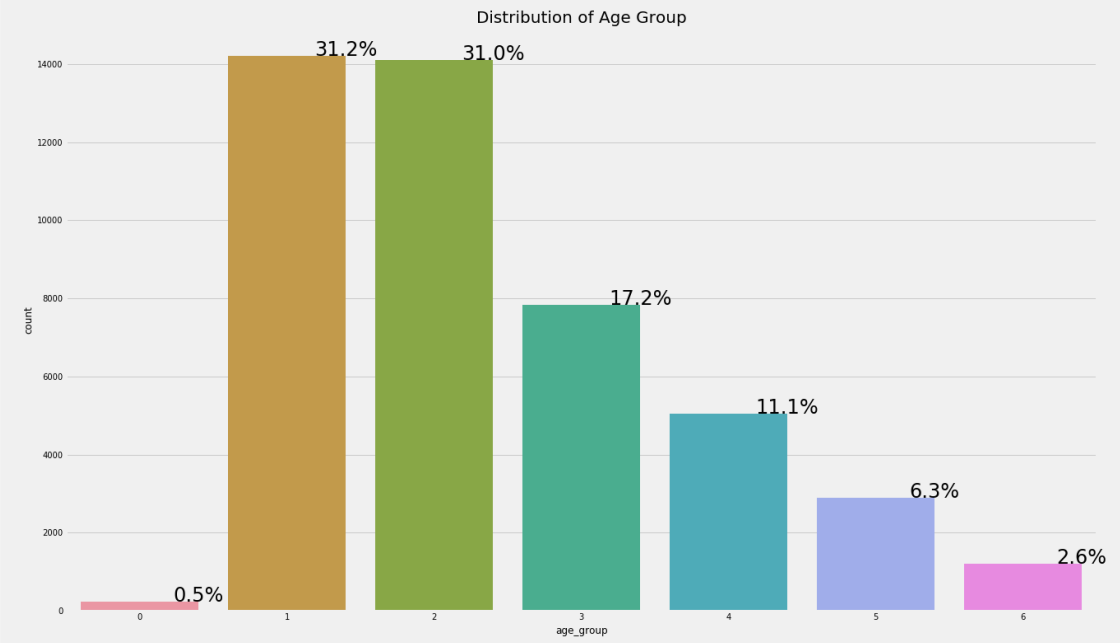
Voted

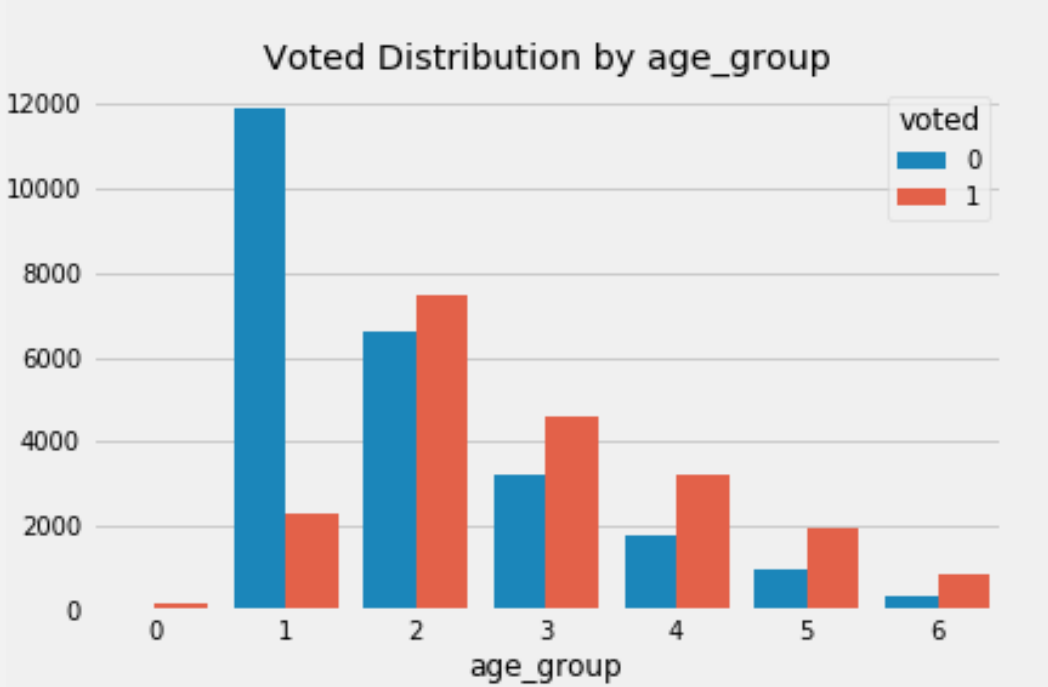


선거에 참가한 인원과 그렇지 못한 인원의 비율이 대체로 유사한 편

≫ 균등한 데이터!

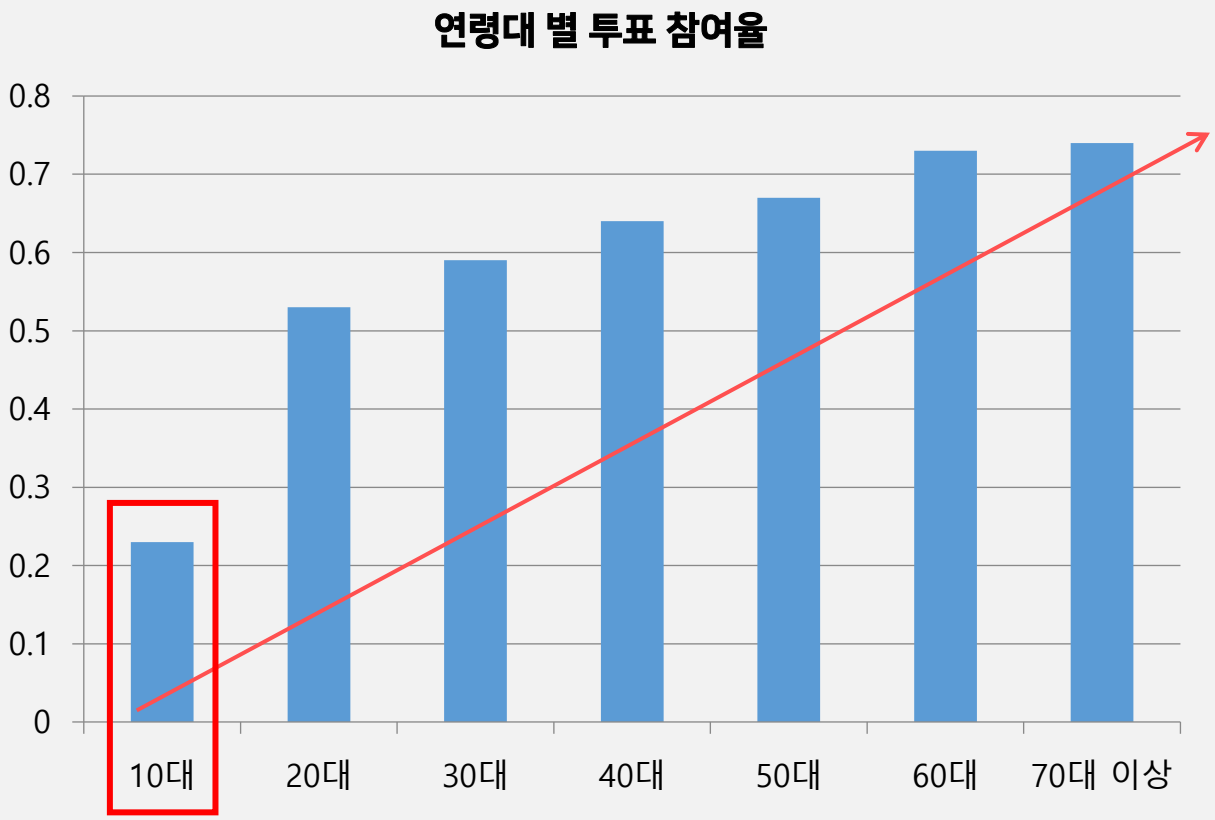
연령대에 따른 투표 참여율





데이터의 31.2%를 차지하는 10대의 투표 참여율 저조로 인한 결과

연령대에 따른 투표 참여율



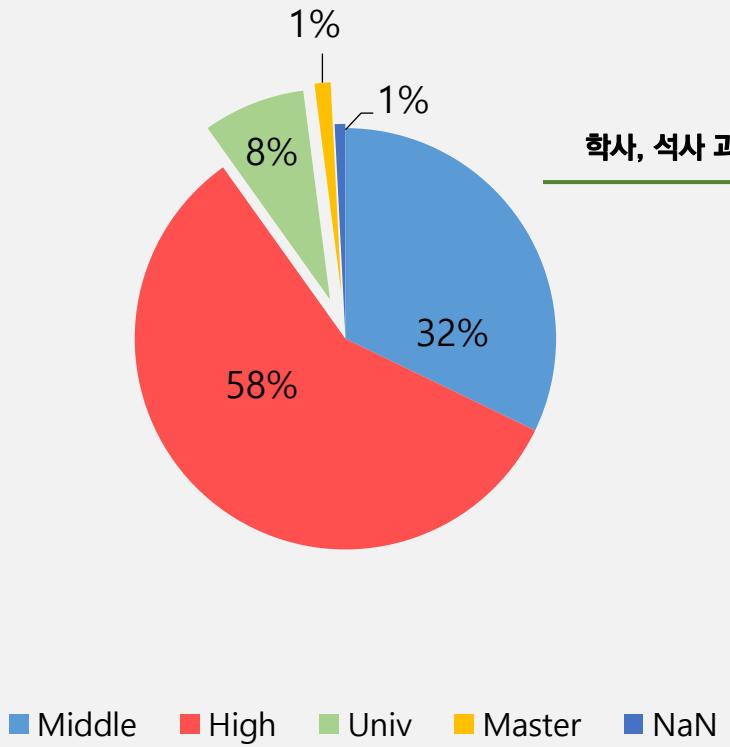
10대의 경우 선거권 미취득 인원 제외를 위해 고등학교 졸업 이상으로 제한

연령대의 증가에 따른 투표 참여율 증가

02 변수 분석 및 전처리

10대의 교육수준 (학력)

10대의 학력 분포



학사, 석사 과정을 마친 인원이 10대의 9%를 차지

10대의 인종 별 학력

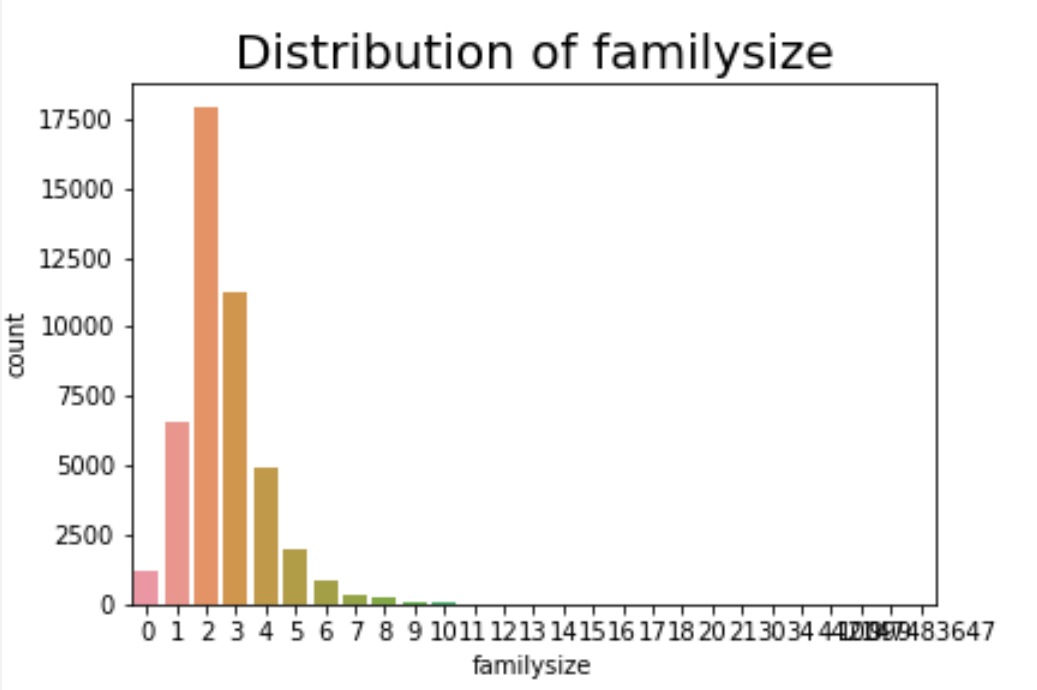
-	Middle	High	Univ	Master
Arab	0.01	0.01	0.02	0.01
Asian	0.15	0.18	0.26	0.43
Black	0.06	0.06	0.06	0.05
Indigenous Austrailian	0.00	0.00	0.00	0.01
Native American	0.02	0.01	0.01	0.01
Other	0.11	0.12	0.11	0.09
White	0.65	0.62	0.54	0.41

인도와 중국은 세계 경쟁력과 자원을 두고서 벌써 치열한 경쟁을 하고 있습니다. 여기에 더해 이 두 나라는 미국에서 공부 하고 있는 자국 학생 수에서 다시 한 번 경쟁하고 있습니다. 2008년부터 2012년 사이 인도는 168,034명의 학생을 미국으로 보냈습니다. 이는 미국에서 공부하고 있는 외국 학생의 15%에 해당합니다. 이는 284,173명에 달하는 중국 학생 다음으로 많습니다. 하지만 인도 학생들과 중국 학생들이 미국으로 건너오는 이유는 다릅니다. 대부분 인도 학생들은 석사 학위를 위해서 미국에 옵니다. 미국에서 공부하는 인도 학생 중 10%만이 학부나 박사 학위를 위해서 미국에 옵니다. 반면, 중국 학생 중 44%는 학부 교육을 위해서 미국으로 건너옵니다.

https://newspeppermint.com/2014/09/02/differencebetweenchineseandindianstudentsintheus/

고학력일수록 아시아 인의 비율이 높다는 점과 위의 기사를 토대로 10대이면서 고학력자에 대한 전처리 X

이상치 제거 – familysize



상식적으로 불가능 한 수치들에 대한 제거

(2147483647, 999, 100)

```
print(sorted(data['familysize'].unique(), reverse=True))
```

```
[2147483647, 999, 100, 44, 34, 30, 21, 20, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9,
8, 7, 6, 5, 4, 3, 2, 1, 0]
```

02 변수 분석 및 전처리

무응답 대체

관측치 제거

장점

- 전처리 과정이 복잡하지 않음
- 불성실한 답변에 대한 배제 가능

단점

- 데이터 정보량 손실
- Test set에도 무응답 존재하므로, test set의 관측치도 제거해야 됨

평균 및 최빈값 대체

장점

- 전처리 과정이 복잡하지 않음

단점

- 각 관측대상의 특성 무시

채택

Mice

장점

- 정보량의 손실 및 왜곡 최소화

단점

- 제거 및 평균 대체에 비해 소요시간이 큼

Education, Engnat, Hand, Married, Urban 변수에 대해 Mice를 통한 무응답 대체

LabelEncoding

Character 변수에 대한 Factor화 필요

변수의 개수가 늘어나지 않는 방법을 채택

One-Hot Encoding (Dummy화)

장점

- 각 응답항목이 갖는 특성 파악 용이

단점

- 변수의 개수가 많아짐

LabelEncoding

장점

- 변수의 개수가 늘어나지 않음

단점

- Encoding된 응답항목에 대한 match 필요

LabelEncoding

Character 변수에 대한 Factor화 필요

변수의 개수가 늘어나지 않는 방법을 채택

One-Hot Encoding (Dummy화)

```
In [3]: pd.get_dummies(train.race).head()
Out[3]:
   Arab  Asian  Black  Indigenous Australian  Native American  Other  White
0     0      0      0                      0                 0      0      1
1     0      1      0                      0                 0      0      0
2     0      0      0                      0                 0      0      1
3     0      1      0                      0                 0      0      0
4     0      0      0                      0                 0      0      1
```

LabelEncoding

```
Out[9]:
0    6
1    1
2    6
3    1
4    6
```

After	Before
0	Arab
1	Asian
2	Black
3	Indigenous Australian
4	Native American
5	Other
6	White

변수제거 – Q_E

- 상대적 답변에 대한 해석 기준 모호
- Target 변수와의 관계 無 판단

	Q_E_mean	voted
Q_E_mean	1.000000	-0.003894
voted	-0.003894	1.000000

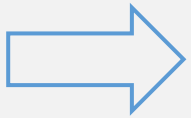
- Model Accuracy 확인 결과 감소 확인

해당 변수 제거

Scaling

- FamilySize 이상치 제거
- Q_E 변수 제거

Scale이 큰 변수 존재 X

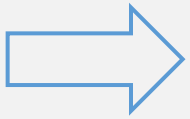


Scaling 미실시

변수개수 감축

Q_A, TP__, Wr_, Wf_ 변수가 전체 78개의 변수 중 46개를 차지

각 변수에 대한 Score 생성



8개의 변수로 변환

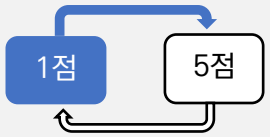
- Q_A : Mach_Score
- TP__ : TP_ex, TP_ag, TP_co, TP_em, TP_op
- Wr_ : Wr_more, Wr_less
- Wf_ : Wf_know

02 변수 분석 및 전처리

Mach_score

Q_A을 통한 마키아벨리즘 Score

각각 10개의 Positive, Negative 질문으로 이루어져 있어 음의 질문에 대한 역순화 필요



1. 주어진 5개의 Negative 질문에 대한 역순화 진행 (QeA, QfA, QkA, QqA, QrA)

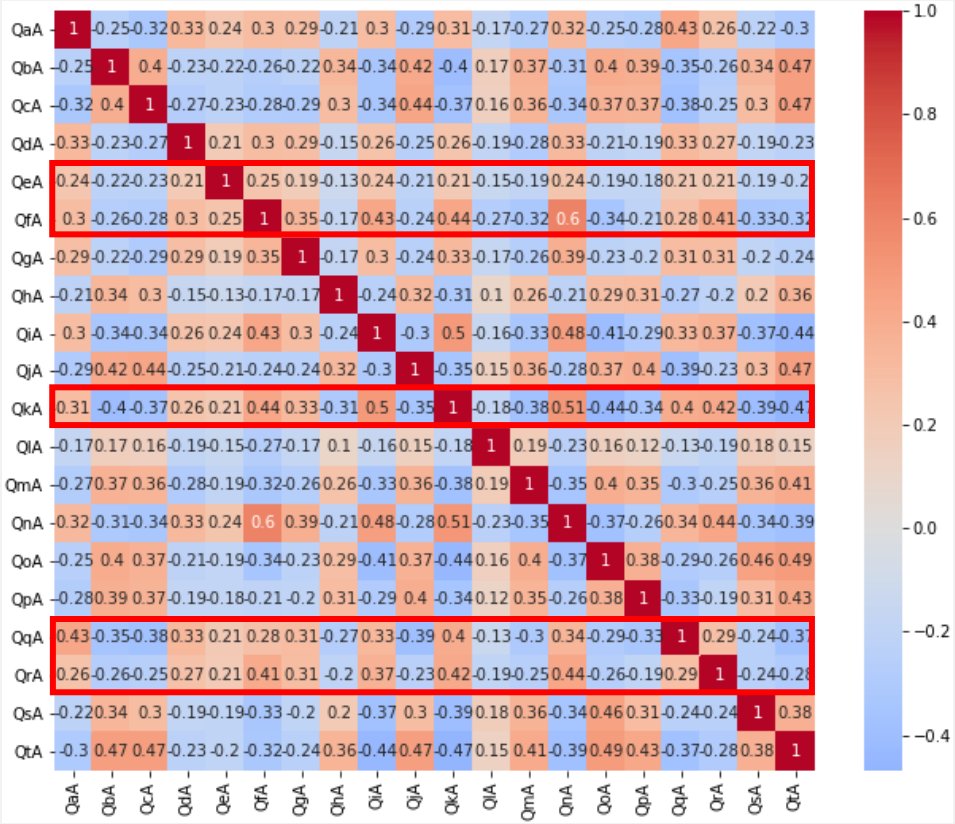
↓

2. Secret 처리된 8개의 질문 중 Negative 질문에 대한 역순화 진행

↓

3. Mean(Q_A) = Mach_score

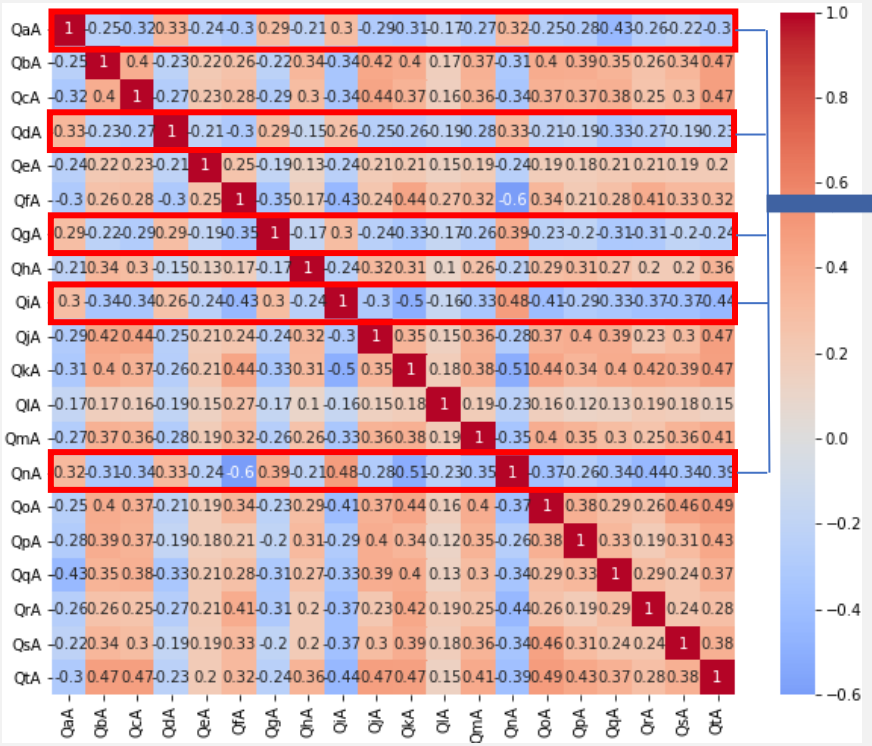
1. Original Correlation Heatmap



02 변수 분석 및 전처리

Mach_score

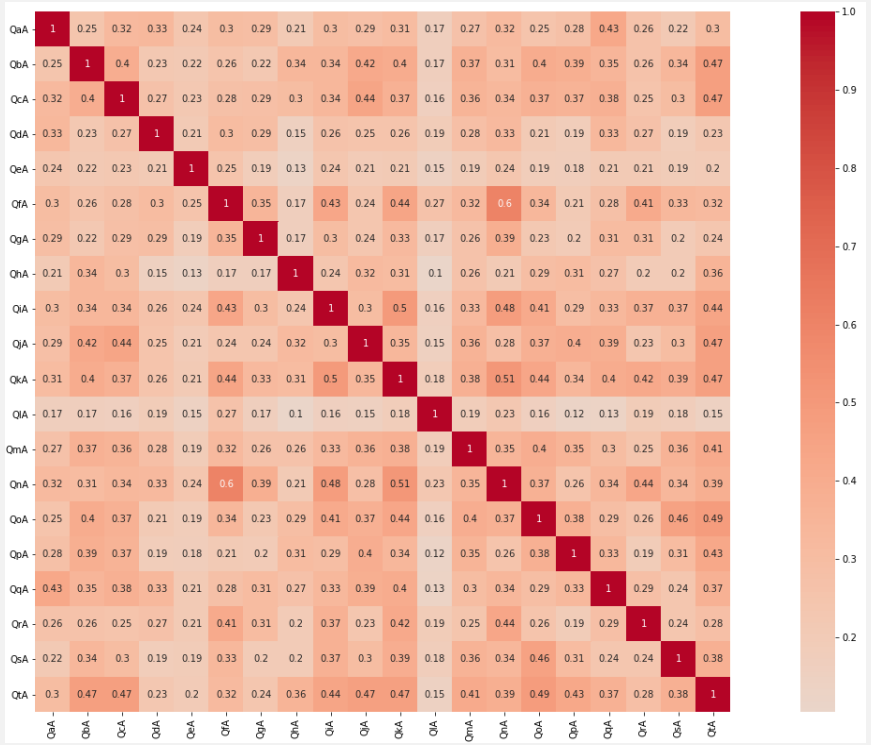
2. After cleaning known negative answers



Secret 처리된 음의 질문 식별
(QaA, QdA, QgA, QiA, QnA)

식별한 질문에 대한 역순화 적용

3. After cleaning all negative answers



02 변수 분석 및 전처리

TP_score

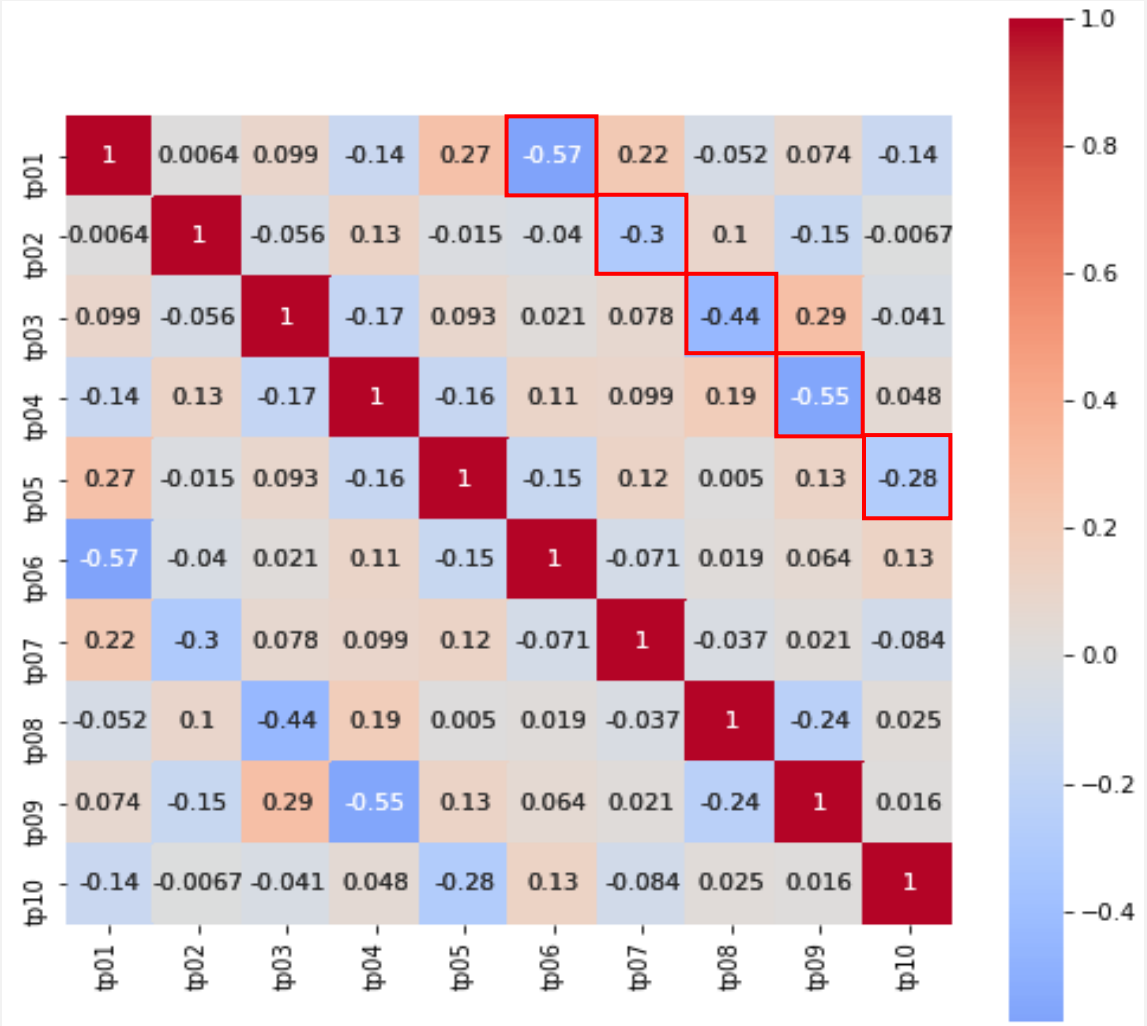
TP__ 을 통한 5개의 성향 변수

1. 무응답에 대한 평균 대체

2. 0에 가까울수록 높은 성향을 지닌 것에 대한 역순화 진행

3. 음의 상관관계를 갖는 질문 식별 및 성향 변수 생성

- TP_ex (외향성) = $\{TP01 + (7 - TP06)\}/2$
- TP_ag (친화성) = $\{TP07 + (7 - TP02)\}/2$
- TP_co (성실성) = $\{TP03 + (7 - TP08)\}/2$
- TP_em (정서 안정성) = $\{TP09 + (7 - TP04)\}/2$
- TP_op (경험 개방성) = $\{TP05 + (7 - TP10)\}/2$



02 변수 분석 및 전처리

Wr_more, Wr_less, Wf_know

Wr_more

- 대부분의 사람이 설명 가능한 실존하는 단어 중 해당 인원이 정의할 수 있는 단어의 수
- wr_01, wr_02, wr_04, wr_05, wr_07, wr_08, wr_10, wr_12, wr_13

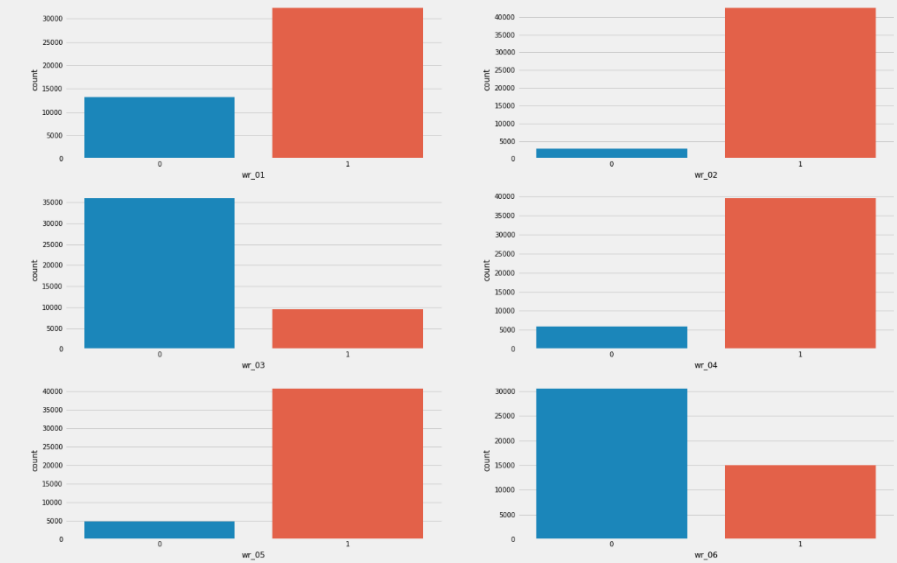
Wr_less

- 대부분의 사람이 설명 불가능한 실존하는 단어 중 해당 인원이 정의할 수 있는 단어의 수
- wr_03, wr_06, wr_09, wr_11

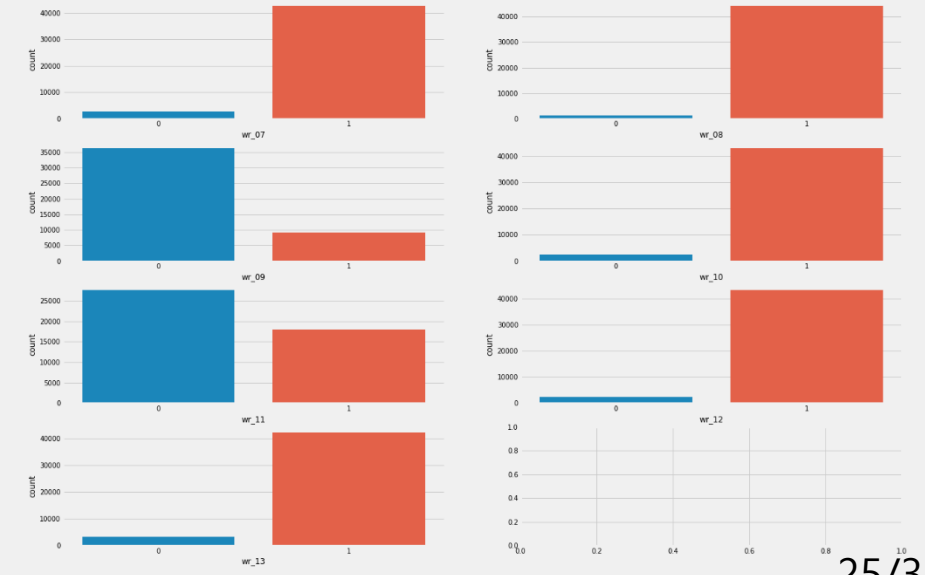
Wf_know

- 허구의 단어 중 해당 인원이 정의할 수 있는 단어의 수
- wf_01, wf_02, wf_03

Wr_01 ~ wr_06 분포



Wr_07 ~ wr_13 분포



최종 변수

독립 변수

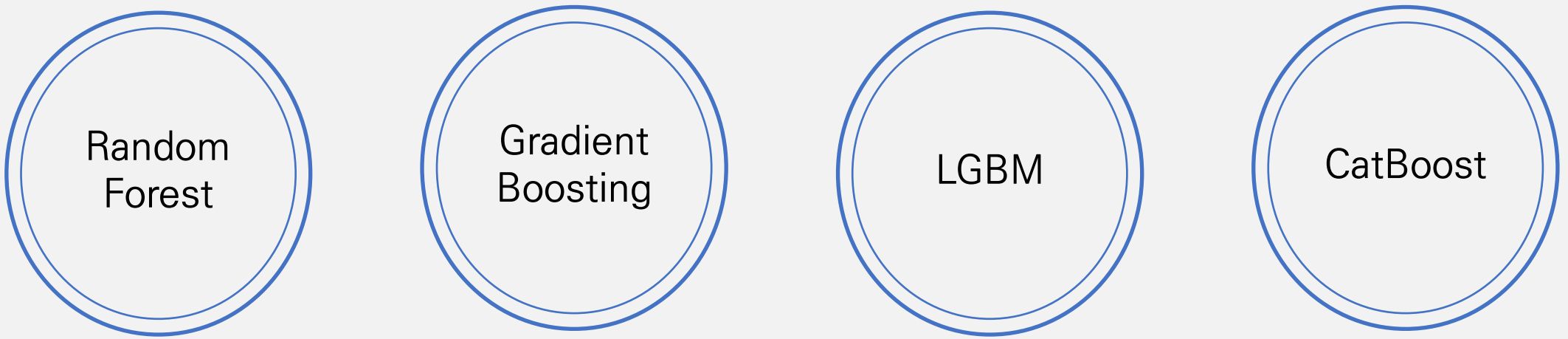
age_group, education, engnat,
familysize, gender, hand,
married, race, religion, urban,
tp_ex, tp_ag, tp_co, tp_em, tp_op,
wr_many, wr_less, wf_know,
Mach_score

종속 변수

voted

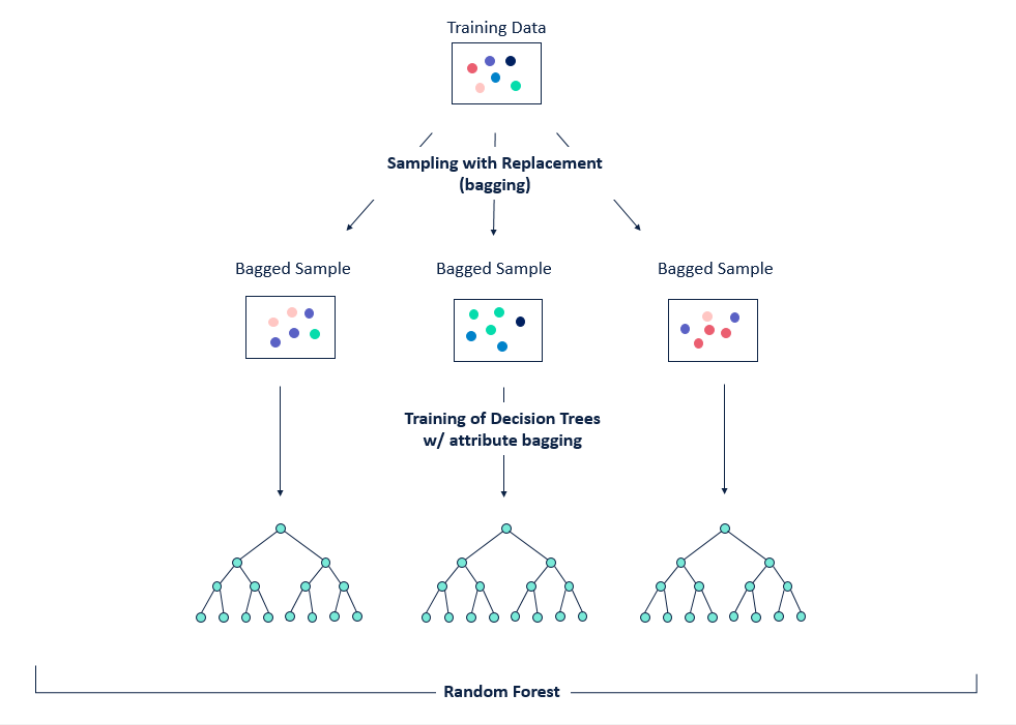
Taget feature = Categorical Variable

: 분류 모델 사용!



Random Forest

: 여러 개의 결정 트리(Decision Tree)를 활용한 배깅 방식의 알고리즘

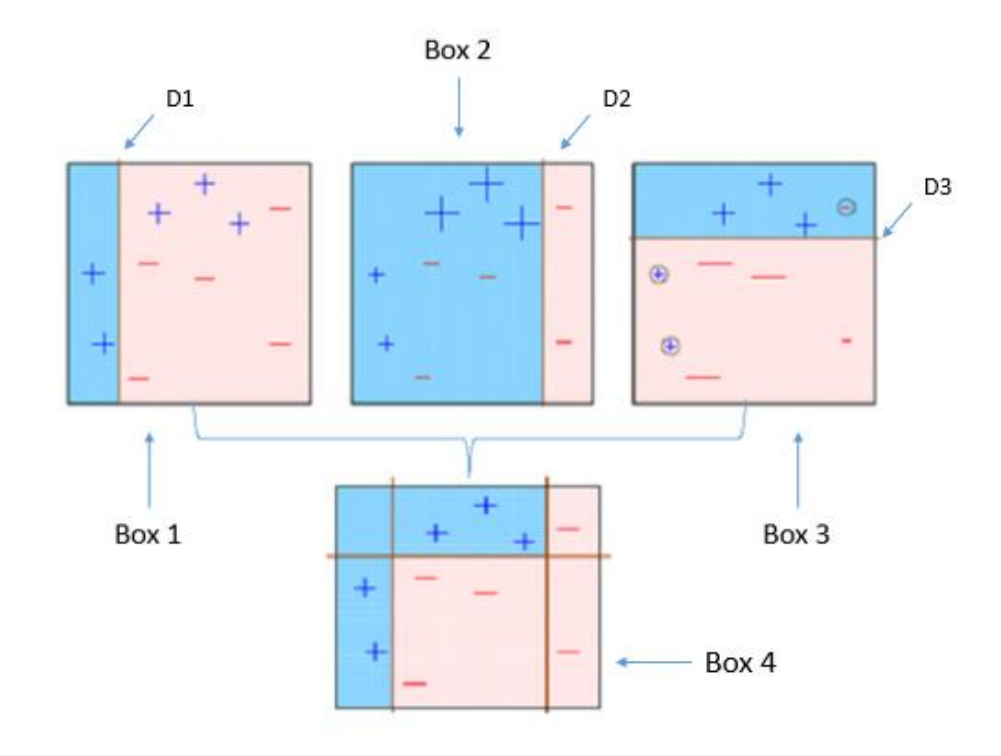


장점

- 결정 트리의 쉽고 직관적인 장점을 그대로 지님
- 앙상블 알고리즘 중 비교적 빠른 수행 속도를 지님
- 과적합 방지 가능

Gradient Boosting

: 손실 함수의 기울기를 바탕으로 여러 개의 약한 예측 모형을 단계적으로 생성한 뒤,
앙상블 방법으로 결합하는 알고리즘



장점

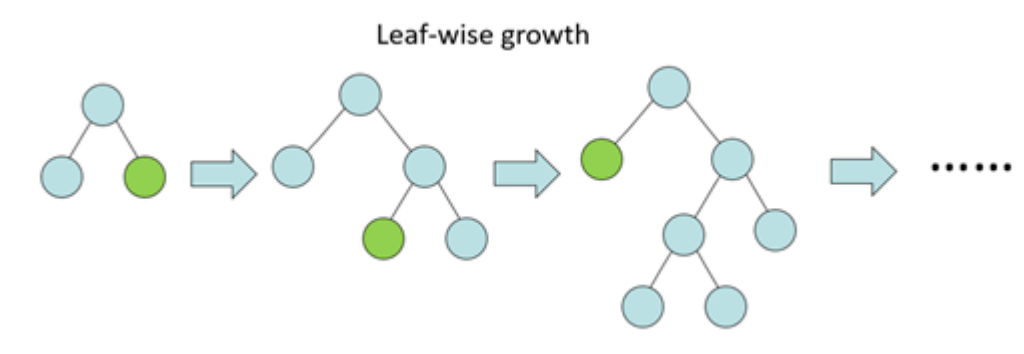
- 약한 예측력을 지닌 나무모형을 결합하는 데 적용되어 예측 성능이 높음

단점

- Greedy Algorithm을 사용하므로 과적합 우려
- 수행 시간이 오래 걸림

LGBM

: Gradient Boosting 프레임 워크로 Tree 기반의 학습 알고리즘



장점

- Leaf – wise 트리 성장을 통해 손실을 최소화
- 메모리를 적게 차지하고, 속도가 빠름

단점

- Leaf – wise 트리 성장을 통해 과적합에 취약

Cat Boost

: 잔차 추정의 분산을 최소로 하여 bias를 피하는 부스팅 기법

장점

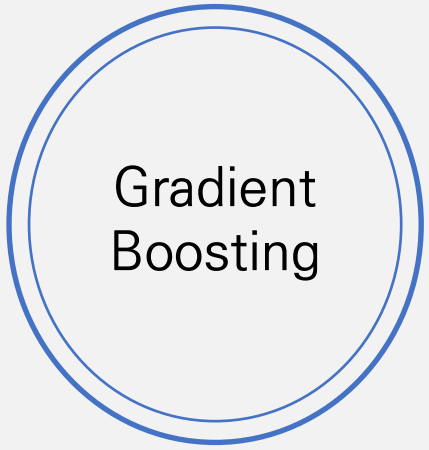
- Categorical 변수를 처리하는 데 효과적
- Ordered Boosting과 Random Permutation을 통해 과적합 방지
- Ordered target Encoding을 통해 Data Leakage 방지

단점

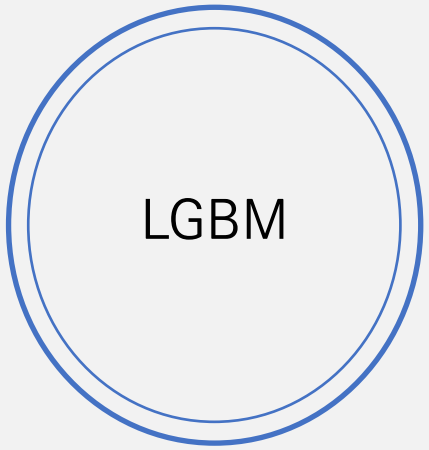
- 데이터 대부분이 Numeric 변수인 경우 LGBM보다 학습 속도가 느림

Modeling Accuracy

0.758629



0.768318

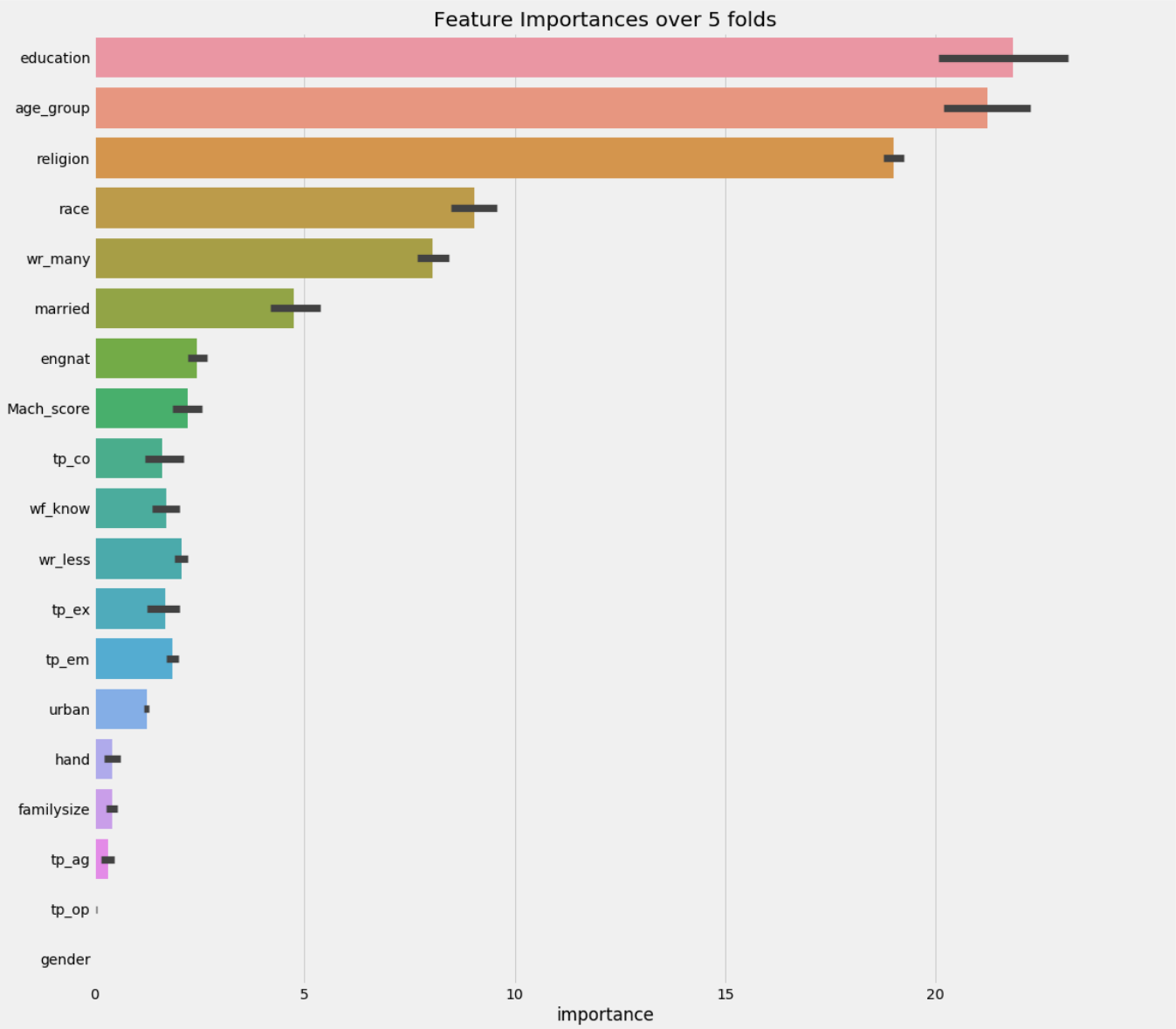


0.769074

0.771326

Best

Best Model's Feature Importance



한계점

1. 이상치 제거 기준이 명확하지 못한 점
2. 미국 연령 기준 18세부터 선거가 가능하다. 즉, 18세 미만의 청소년은 선거에 참여가 불가능하다. 대회에서 주어진 데이터는 10대와 20대의 비중이 가장 크기 때문에, 10대 중 법적으로 선거에 참여가 가능한 10대와 그렇지 못한 10대를 구분하는 것이 중요하다고 볼 수 있으나 별도의 처리를 하지 못했다.
3. 분석을 진행하면서 변수 별로 독특한 데이터를 확인할 수 있었다. 처음에 이를 이상치로 판단했으나, 관련된 domain 지식을 찾아본 결과 유의미한 데이터로 확인되었다. 이처럼 분석 외적으로 domain 지식이 부족하였다.
4. 시간 상의 문제로 각 모델에 대한 최적의 하이퍼 파라매터를 찾지 못하였다.

감사합니다