# 설문조사를 통한 지지정당 예측

행복 범주형자료분석팀

김찬영 이혜인 김서윤 심은주 진수정

# CONTENTS


01. 1주차 피드백

02. DATA 정리

03. 모델링

04. 결과 해석

05. 한계와 의의

# 피드백 반영

01.
1주차 피드백

02.
DATA 정리

03.
모델링

04.
결과 해석

05.
한계와 의의

## *Solution 2* Imputation 변수 선택 알고리즘 ver.2.0 재정비

### STEP1

Imputation을 위한 모델링에서 다중공선성을 최소화 하기위해 10개의 1:1 조합만 선택

| | v1 | v2 |
|---|---|---|
| Comb1 | Life_Q_livealone | ps_Q_LeftHanded |
| Comb2 | Life_Q_watchTV | env_Q_p_spank |
| Comb3 | Life_Q_gun | ps_Q_PowerOfPositive |
| Comb4 | mo_Q_fulltimejob | mo_Q_minwage_job |
| | | ... |
| Comb9 | Life_Q_drink | re_Q_newromance |
| Comb10 | ps_Q_Creative | re_Q_havesibling |

이 때의 척도 역시 Gktau measure!

BUT, 비대칭성을 고려하여
선후관계가 바뀌어도
연관 있는 조합 중 1:1 조합을 선택

# 피드백 반영

01.
1주차 피드백

02.
DATA 정리

03.
모델링

04.
결과 해석

05.
한계와 의의

## *Solution 2* Imputation 변수 선택 알고리즘 ver.2.0 재정비

### STEP2

각 조합에 포함된 변수와 Gktau measure가 낮은 질문 변수들을 선택하여
각 조합의 변수 15개로 완성

| | v1 | v2 | V14(추가) | V15(추가) |
|---|---|---|---|---|
| Comb1 | Life_Q_livealone | ps_Q_LeftHanded | re_Q_extendedfamilly | re_Q_meetoffline |
| Comb2 | Life_Q_watchTV | env_Q_p_spank | Life_Q_ownTool | ps_Q_Socializing |
| Comb3 | Life_Q_gun | ps_Q_PowerOfPositive | ps_Q_BetterAfter5y | mo_Q_has_debt |
| Comb4 | mo_Q_fulltimejob | mo_Q_minwage_job | ex_Q_cry60D | ps_Q_SupportCharity |
| | | ... | | |
| Comb9 | Life_Q_drink | re_Q_newromance | Life_Q_tabwater | ps_Q_BuyHappiness |
| Comb10 | ps_Q_Creative | re_Q_havesibling | mo_Q_carpayment | Life_Q_glasses |

# 피드백 반영

01.
1주차 피드백

02.
DATA 정리

03.
모델링

04.
결과 해석

05.
한계와 의의

*Solution 3* ***인적변수들은 모두 MNAR로 가정, NA imputation 진행 X,***

***'무응답 ' 으로 처리***

| Age | Education Level | Income |
|---|---|---|
| 20s | 1 | NA |
| 30s | 2 | 1 |
| 30s | NA | 0.74 |
| NA | 2 | 0.74 |
| 40s | 1 | NA |
| NA | 1 | 1 |

| Age | Education Level | Income |
|---|---|---|
| 20s | 1 | 0 |
| 30s | 2 | 1 |
| 30s | 0 | 0.74 |
| non_answer | 2 | 0.74 |
| 40s | 1 | 0 |
| non_answer | 1 | 1 |

*무응답은 Ordinal Encoding 시에 0으로 처리*

01.
1주차 피드백

02.
DATA 정리

03.
모델링

04.
결과 해석

05.
한계와 의의

- **파생변수 유의성 판단**

새롭게 만든 파생변수 (조나단 파생변수, Life sum, Edu sum) 가 유의할까?

⟶ 각각의 파생변수와 Y변수 'Party' (지지정당)간의 독립성 검정 시행

```
> chisq.test(train_mca_jo_sum$Life_sum, train_mca_jo_sum$Party)

        Pearson's Chi-squared test

data:  train_mca_jo_sum$Life_sum and train_mca_jo_sum$Party
X-squared = 2.3767, df = 4, p-value = 0.6669
```

*Y와 독립이라면?*

해당 파생변수는 유의하지 않으므로 제거하자!

01.
1주차 피드백

02.
DATA 정리

03.
모델링

04.
결과 해석

05.
한계와 의의

- **질문 변수 선별**

Gktau를 기준으로 서로 연관 있는 질문들의 조합을 탐색한 후에
그 중 Y와 연관성이 없는 질문 삭제

제거된 질문변수

| 데이터셋 | Train_nonanswer | Train_rf, Train_mca, Train_mean, Train_NA |
|---|---|---|
| 제거된 질문 변수 | mo_Q_minwage_job, mo_Q_fulltimejob, Life_Q_collectHobby, mo_Q_has_enoughcash_now, re_Q_newromance, ps_Q_PowerOfPositive, edu_Q_parents_college, env_Q_single_parent ,ps_Q_LikeFirstName, re_Q_likepeople, Life_Q_watchTV, ps_Q_LeftHanded,re_Q_havesibling | Nothing |

01.
1주차 피드백

02.
DATA 정리

03.
모델링

04.
결과 해석

05.
한계와 의의

# *"최종 데이터 분포 비교"*

- Mac or PC 변수 동질성 검정 결과

| 데이터 종류 | P-value |
|---|---|
| NA 데이터 | 0.4598 |
| Non_answer 데이터 | 0.8854 |
| RF 데이터 | 0.7396 |
| MCA 데이터 | 0.3453 |
| Mean 데이터 | 0.4489 |

**p-value > 0.05**

**: 모두 귀무가설 채택 !**

데이터 간의 분포가 동일

01.
1주차 피드백

02.
DATA 정리

03.
모델링

04.
결과 해석

05.
한계와 의의

# *"찐 최종 데이터"*

**NA → 무응답**
으로 처리한
데이터

가장 최적 값이 나온
**nonanswer 데이터**를
최종 데이터로 결정!

- NA로 남겨놓은 기존 데이터
- NA →MCA 로 채운 데이터
- NA → Mean 으로 채운 데이터
- NA → RF 로 채운 데이터

*넘모 슬퍼,,우리의 1주차 고생 다 어디에,,*

01.
1주차 피드백

02.
DATA 정리

03.
모델링

04.
결과 해석

05.
한계와 의의

# GLM

## *Logistic with Lasso penalty*

**최적의 파라미터 조합**

- Log lambda : -4.677
- Cutoff : 0.526

Test Accuracy

0.64224

01.
1주차 피드백

02.
DATA 정리

03.
모델링

04.
결과 해석

05.
한계와 의의

RandomForest

# Random Forest Classifier

- 모델에 사용할 변수 선택

  → RFE를 적용해서 선택!

**RFE(Recursive Feature Elimination)란?**

모든 변수를 다 포함시킨 후 반복해서 학습을 진행하면서
중요도가 낮은 변수를 하나씩 제거하는 방식

→ 일종의 Backward Selection 방법

**선택된 변수들**

careness
fairness
ps_Q_Feminist
Life_Q_medipray
Life_Q_gun
Life_Q_MacPC
Ps_Q_LifePurpose
⋮

01.
1주차 피드백

02.
DATA 정리

03.
모델링

04.
결과 해석

05.
한계와 의의

RandomForest

## Random Forest Classifier

**변수 선택**

RFE를 통해
200개의 변수
선택

**최적의 파라미터 조합**

- n_estimators: 1200
- max_features: 'sqrt'
- max_depth: 30
- min_samples_split: 16
- min_samples_leaf: 9

Test Accuracy

0.64511

# 모델링



01.
1주차 피드백

02.
DATA 정리

03.
모델링

04.
결과 해석

05.
한계와 의의

## XGBoost

### eXtreme Gradient Boosting

**최적의 파라미터 조합**

- max_depth : 4
- min_child_weight: 1
- learning_rate : 0.1
- n_estimators: 100
- gamma : 0
- reg_lambda : 20

Test Accuracy

0.65948

# 모델링



01.
1주차 피드백

02.
DATA 정리

03.
모델링

04.
결과 해석

05.
한계와 의의

## Light GBM

### *Light Gradient Boosting Model*

**최적의 파라미터 조합**

- learning_rate:0.003
- n_estimators : 900
- num_leaves: 30
- reg_alph:0
- reg_lambda: 40
- colsample_bytree: 0
- bagging_fraction:0.0002
- feature_fraction: 0.7
- min_child_weight: 40
- max_bin: 800

Test Accuracy

0.66091

# 모델링



01.
1주차 피드백

02.
DATA 정리

03.
모델링

04.
결과 해석

05.
한계와 의의

## CatBoost

### Categorical Boosting

**최적의 파라미터 조합**

- iterations : 400
- depth : 7
- random_strength : 0.099234
- learning_rate : 0.018487
- l2_leaf_reg : 8

Test Accuracy

0.64367

# 모델링



01.
1주차 피드백

02.
DATA 정리

03.
모델링

04.
결과 해석

05.
한계와 의의

Ensemble

## Stacking Ensemble

**사용한 모델**

- Random Forest
- XGBoost
- Lasso Regression
- Logistic Regression
- CatBoost
- Extra Tree Classifier
- LGBM
- SVM

meta model
: XGBoost

Test Accuracy

0.65517

*찢어따,,, 앙상블,,,*

# 모델링 결과 해석

01.
1주차 피드백

02.
DATA 정리

03.
모델링

04.
결과 해석

05.
한계와 의의

SHAP value

## Shapley value

특정 결과에 각 특성이 얼마나 영향을 미쳤는지 나타내는 수치

→ 게임이론에서 파생된 개념으로 확고한 이론적인 기반을 갖는다!

$$\varphi_i(v) = \frac{1}{number\ of\ players} \sum_{coalitions\ excluding\ i} \frac{marginal\ contribution\ of\ i\ to\ coalition}{number\ of\ coalitions\ excluding\ i\ of\ this\ size}$$

*(갓키피디아에서 shap value의 예시를 수식으로 나타낸 것을 첨부했다!)*

→ 직관적으로 보자면, [A라는 플레이어의 행위를 포함한 결과]에서
[marginal 하게 A 플레이어의 행위를 제외한 결과]를 뺀 값

# 모델링 결과 해석

01.
1주차 피드백

02.
DATA 정리

03.
모델링

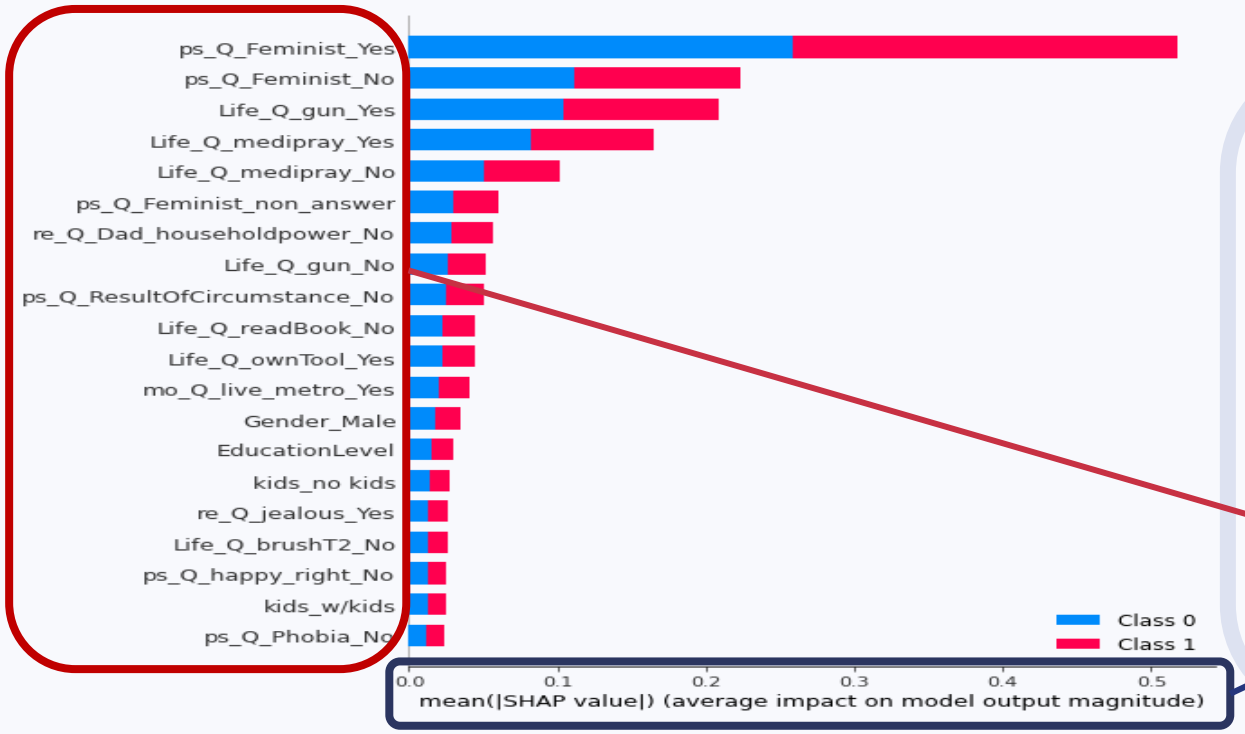
04.
결과 해석

05.
한계와 의의

SHAP value

## Shapley value

What is shap.summary plot?




shap. summary plot

타겟 변수에 대한 X변수들의
평균적인 영향력을 나타낸 Plot

Class 0 → Democrat
Class 1 → Republicans

Y축 : 영향력이 높은 순서대로 변수 나열

X축 : mean|*Shap value*|

# 모델링 결과 해석

01.
1주차 피드백

02.
DATA 정리

03.
모델링

04.
결과 해석

05.
한계와 의의

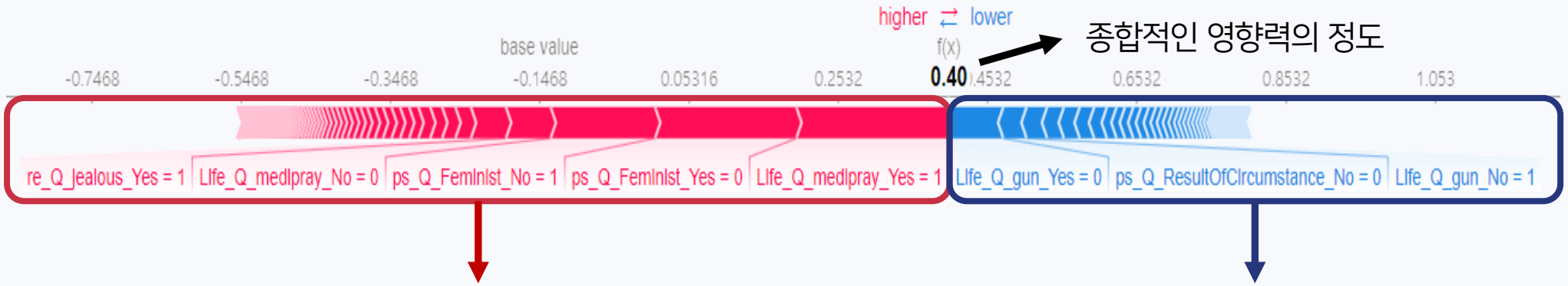
SHAP value

## Shapley value

What is shap.force plot?

→ 각 설문 참여자의 응답결과를 shap value를 활용하여 시각화한 plot

→ **각 설문 참여자들의 예측 결과에 대한 해석이 가능**해진다!




빨간색은 타겟 변수에 대한
양(+)의 영향력을 나타낸다!

파란색은 타겟 변수에 대한
음(-) 영향력을 나타낸다!

# 모델링 결과 해석

01.
1주차 피드백

02.
DATA 정리

03.
모델링

04.
결과 해석

05.
한계와 의의

SHAP value

## Shapley value

shap.force plot 해석 : 민주당 지지자


기존 범주 : Republican

Republican으로 예측함에 있어
양의 영향력을 가지는 질문과 대답
Q. ps_Q_Phobia → A: No

Republican으로 예측함에 있어
음의 영향력을 가지는 질문과 대답
Q. Life_Q_gun → A: No
Q. ps_Q_Feminist → A: Yes
Q. Life_Q_medipray → A: No

*(기존 범주가 Republican이므로 음의 영향력을 가지는 질문과 대답이 민주당의 특성을 반영한다!)*

# 모델링 결과 해석

01.
1주차 피드백

02.
DATA 정리

03.
모델링

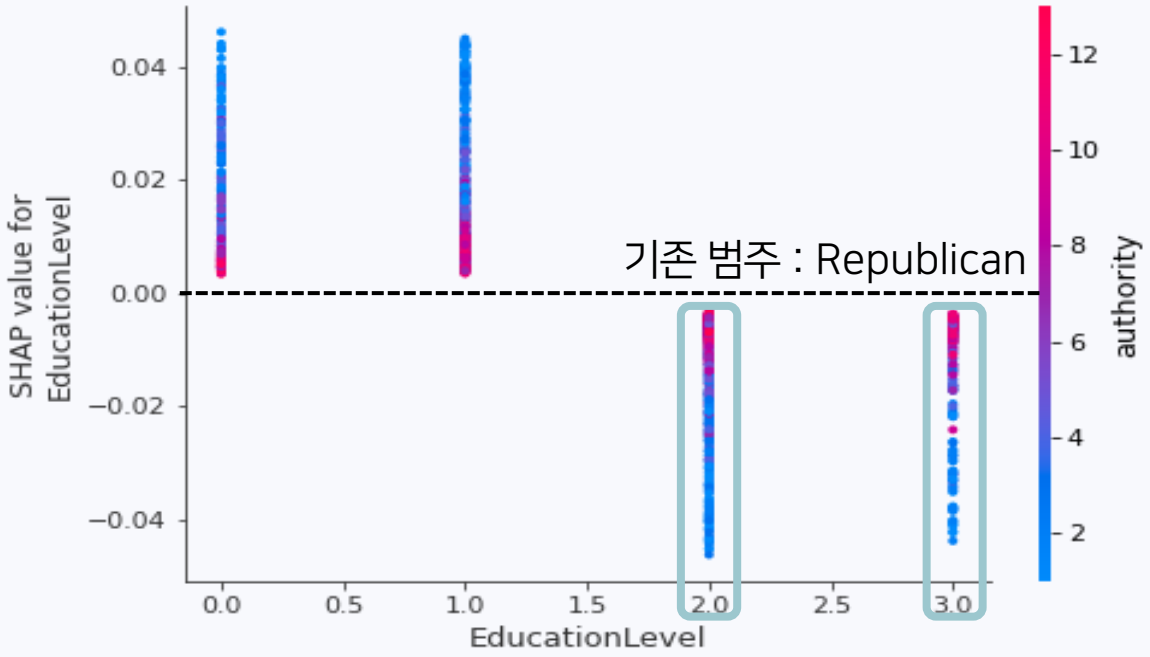
04.
결과 해석

05.
한계와 의의

SHAP value

## Shapley value

What is shap.dependence plot? & 해석




shap.dependence plot

X변수들 간의 연관성을
shap value를 통해 나타낸 Plot

교육수준에 따른
권위에 대한 복종심은 큰 차이가 없다!

높은 교육수준은 Republican에 대해
음의 영향력을 가진다!

*(즉, 민주당의 특성을 반영한다!)*

# 우리의 한계와 의의는?

01.
1주차 피드백

02.
DATA 정리

03.
모델링

04.
결과 해석

05.
한계와 의의

- 의의

범주가 범-주 했다
(100개 이상의 범주형 변수 ^_^)

팀원 모두 한 개 이상의 모델을 이용한 모델링 진행!

R과 Python 모두 마-스터

그 외 오조 오억 개

워라밸 최고 행복 범-주♡