

# 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

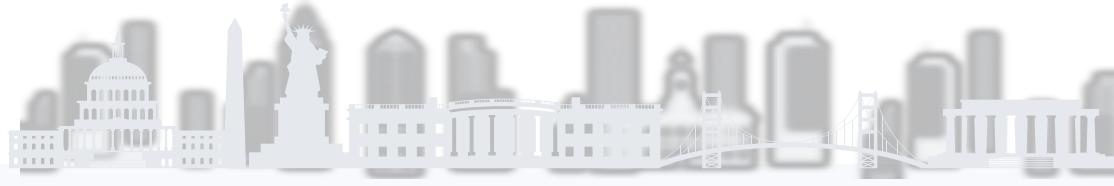


00

# 주제 소개



# 주제 소개



01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

# What

정치성향과 관련 없는 개인적인 질문들로 지지 정당을 파악할 수는 없을까?

Question	Answer
Do you have any siblings?	Yes/No
Does life have a purpose?	Yes/No
Do you have more than one pet?	Yes/No
Are you good good/effective liar?	Yes/No
Do you personally own gun?	Yes/No

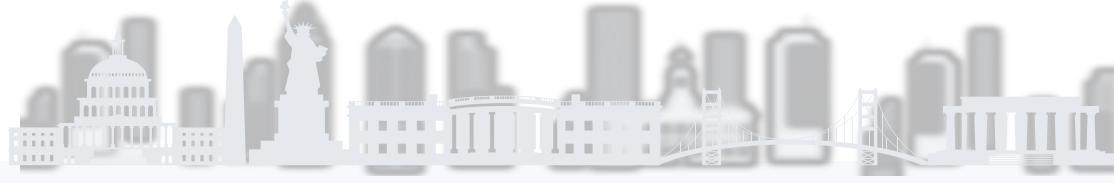
:

응답에 따르면...

공화당 VS 민주당

지지자겠구나!

# 주제 소개



- 설문조사 응답 데이터

101개의 질문에 대한 설문 응답 데이터

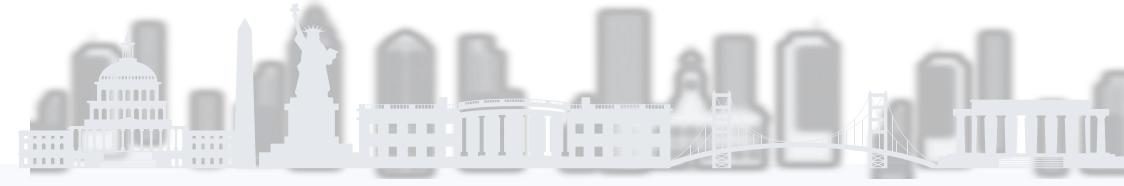
User_ID	Q1	Q2	Q3	Q4	Q5	Q6	...	Q100	Q101
1	No	NA	No	No	No	Yes		No	Yes
4		NA	Yes	No	No	Yes		Yes	No
5	NA	Yes	Yes	No	NA	Yes		No	No
8	No	Yes	No	Yes	No	No		No	Yes
9	No	Yes	No	No	No	Yes		No	Yes
10	NA	NA	NA	NA	No	Yes		NA	NA

:

질문에 대한 대답은 모두 Yes/No 이거나

질문에 따라 두 가지 선택지 중 하나를 선택하는 것으로 **이분화** 되어 있음

# 주제 소개



**모든 전처리/파생변수 생성 과정을 거쳐 만들어진 통합 설문 데이터**

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

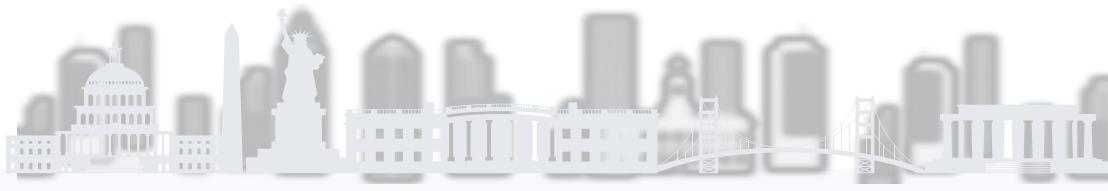
04.  
결과 해석

05.  
한계와 의의

USER_ID	Gender	Income	Level	Education					
				env_Q_p_fight	ifoU	mo_Q_minwage_job	mo_Q_minwage_job	care	sanctity
1	Male	2	1		1		1	1	0
4	Female	5	1		NA		1	1	0
5	Male	3	0		1		0	NA	1
8	Male	4	1		NA		NA	1	0
9	Female	2	0		1	0	0	0	0
10	Female	5	0		NA	1	1	1	0
11	Male	1	1		1		1	NA	1
12	Male	3	0		0		NA	0	0
13	Female	3	0		NA		1	1	0
14	Male	1	1		1		1	NA	1
15	Male	3	0		0		NA	0	1

:

# 주제 소개



## 다양한 데이터와 모델링

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

NA → 무응답  
으로 처리한  
기준 데이터

RF 이용한  
MICE로 채운  
데이터

NA → MICE  
로 채운  
데이터

NA → Mean  
→ MICE  
로 채운  
데이터

NA →  
MCA 이용한  
MICE로 채운  
데이터

5개의 데이터에 대해

*GLM · RandomForest  
XGBoost · CATBoost*

등의 모델을 사용해  
예측력 높이기!

+ *LGBM* 도 해볼게여!

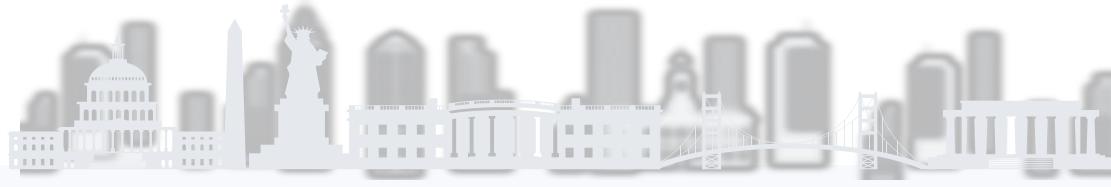


01

# 1주자 스파스 백



# 피드백 반영



## Feedback 1 Income 변수의 범위가 다르다는 문제점

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

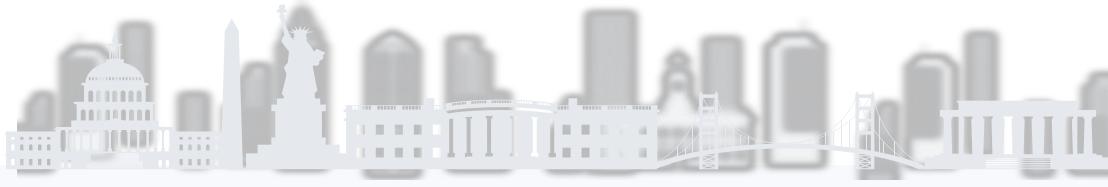
INCOME
under \$25,000
\$25,001
- \$50,000
\$50,001
-\$75,000
\$75,001
- \$100,000
\$100,001
-\$150,000
over \$150,000

INCOME
0
1
2
3
4
5

*Ordinal  
Encoding*

...우리도 몰랐던 문제점...  
감사합니다 딥팀장님^^

# 피드백 반영



01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

## Solution 1 Income 변수를 재범주화

INCOME	소득분위 누적비율
under \$25,000	18.01%
\$25,001 - \$50,000	42.69%
\$50,001 - \$75,000	60.79%
\$75,001 - \$100,000	73.45%
\$100,001 - \$150,000	87.87%
over \$150,000	100%

<2014년 기준 소득분위 누적비율>

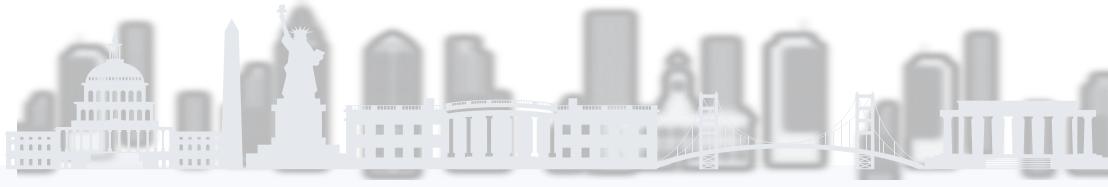
*Ordinal  
Encoding*

INCOME	
0.18	
0.43	
0.61	
0.74	
0.99	
1	

# 피드백 반영



PSAT의 루피 says...



*Feedback 2 MICE에서 변수 선택하는 법을 다시 고민해보심이..!*

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

## STEP1

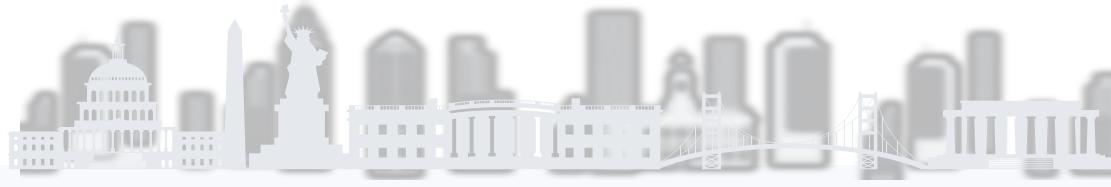
Imputation을 위한 모델링에서 다중공선성을 최소화 하기위해 10개의 1:1 조합만 선택

	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



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

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

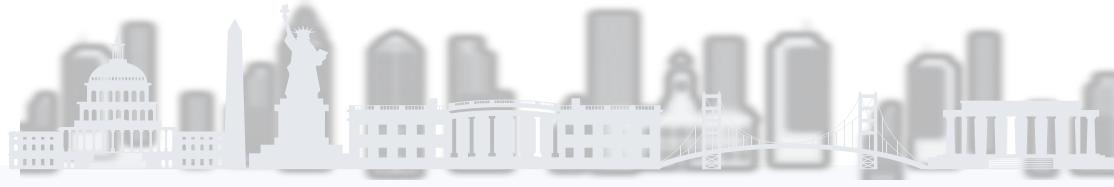
### STEP1

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



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

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

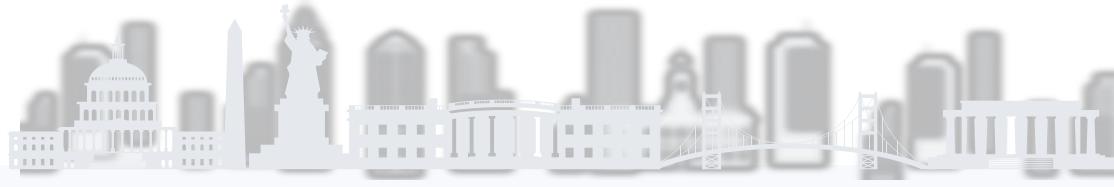
05.  
한계와 의의

### 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_extendedfamily	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



*Feedback 3* 인적변수의 NA는 MCAR이 아닙니다?

랜덤으로 발생하는 결측치

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

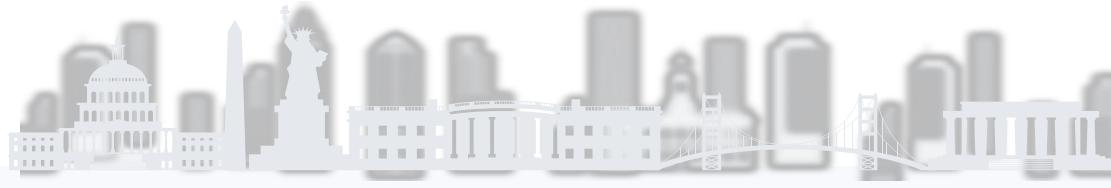
Age	Education Level	Income
20s	1	NA
30s	2	1
30s	NA	0.74
NA	2	0.74
40s	1	NA
NA	1	1

*NA Imputation  
with MICE*

Age	Education Level	Income
20s	1	0.43
30s	2	1
30s	3	0.74
20s	2	0.74
40s	1	0.61
10s	1	1

<기존의 인적 변수 NA 처리 방식>

# 피드백 반영



결측 여부가 해당 변수의 값에 의해 결정되는 NA

*Solution 3* 인적변수들은 모두 MNAR로 가정, NA imputation 진행 X,  
→ '무응답'으로 처리

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

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



02

# DATA 정리





- 주관적이지 않은 설문지 내부 파생변수 생성 시도

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

Life_Q_exercise3	일주일에 3번 이상 운동?
Life_Q_breakfast	매일 아침 먹는지?
Life_Q_earlalarm	알람을 의도적으로 몇분씩 빠르게 맞추어 놓는지?
Life_Q_standard	표준적인 시간에 주로 활동하는지 (9-5시)?

규칙적인 사람인지의 여부를 묻는

'Life\_sum' 질문 생성

<Life Category에 있는 질문>

4 개의 질문에 Yes로 대답한 개수에 따라 Scoring 진행



- 주관적이지 않은 설문지 내부 파생변수 생성 시도

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

edu_Q_math	수학 잘하니?
edu_Q_all A	고등학교나 대학교에서 올 A 받은적 있는지?
edu_Q_plandiploma	석사나 박사학위 있는지 or 계획 있는지?
edu_Q_prdiploma	부모님 두 분 다 대학 학위가 있는지?

교육수준을 판단할 수 있는  
'edu\_sum' 질문 생성

<edu Category에 있는 질문>

4 개의 질문에 Yes로 대답한 개수에 따라 *Scoring* 진행



- **파생변수 유의성 판단**

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

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!



- 파생변수 유의성 판단

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

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

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

### 제거된 파생변수

데이터셋	Train_nonanswer	Train_rf	Train_mca	Train_mean	Train_NA
제거된 파생변수	- Life_sum - Edu_sum	- Life_sum	- Life_sum - Edu_sum	- Life_sum	- Life_sum



- 질문 변수 선별

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

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

- 새로운 결측치 대체법 시도

## MICE with Random Forest

### STEP1

인적변수는 MNAR 가능성 고려,  
NA값에 대하여 “non\_answer”라는 무응답 범주 새롭게 제작하여 입력



### STEP2

설문지 구성 방식을 고려, 질문 변수의 NA값을 무응답으로 가정하고 진행



### STEP3

모든 질문 변수에 대하여 Random Forest 적용한 결과를 대치 값으로 선택!

# “최종 데이터 분포 비교”

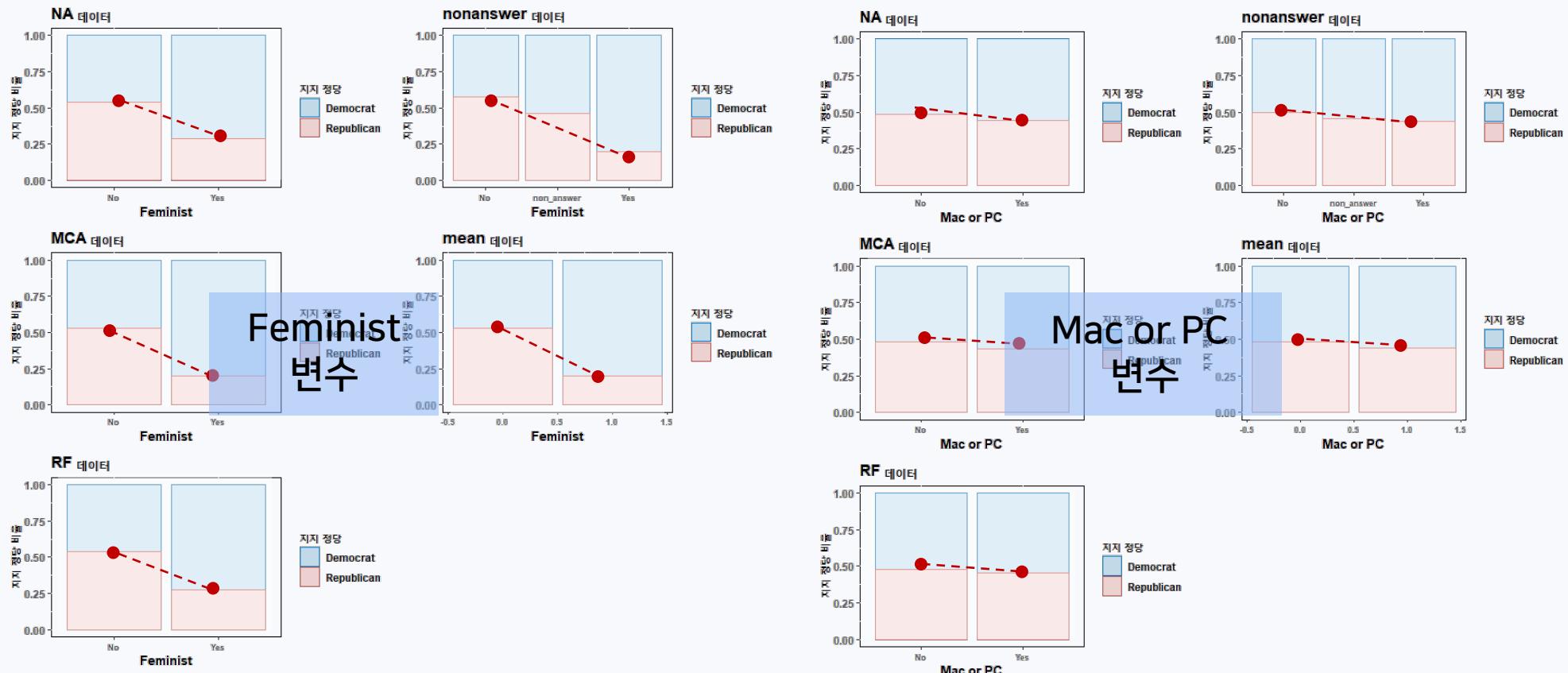
01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의





# 그래프 뿐 아니라 통계적으로도 검정해보자!

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

- 동질성 검정



귀무가설 ( $H_0$ )

두 데이터 간의 분포가 동일하다

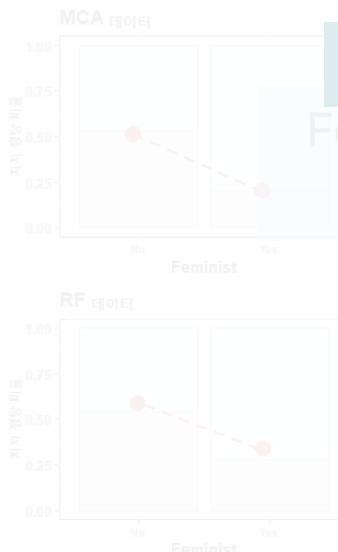
대립가설 ( $H_1$ )

두 데이터 간의 분포가 동일하지 않다



P-value 가 0.05 이상이면 귀무가설 채택!

데이터 간의 분포가 동일하다!





## “최종 데이터 분포 비교”

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

- Feminist 변수 동질성 검정 결과

p-value < 0.05 : 귀무가설 기각

p-value > 0.05 : 귀무가설 채택

데이터 종류	P-value
NA 데이터	0.0006099
Non_answer 데이터	2.2e-16
RF 데이터	0.001603

데이터 종류	P-value
MCA 데이터	0.4922
Mean 데이터	0.5081



## “최종 데이터 분포 비교”

01.  
1주차 피드백

- Feminist 변수 동질성 검정 결과

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

p-value < 0.05 : 귀무가설 기각

p-value > 0.05 : 귀무가설 채택

귀무가설 기각 O

NA 데이터

nonanswer 데이터

RF 데이터

귀무가설 기각 X

mean 데이터

MCA 데이터



## “최종 데이터 분포 비교”

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

- 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



## “최종 데이터 분포 비교”

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

- Mac or PC 변수 동질성 검정 결과

귀무가설 기각 X

데이터 종류	P-value	p-value > 0.05	보구 귀무가설 채택!
NA 데이터	0.8854	0.8854	데이터 간의 분포가 동일
Non_answer 데이터	0.7396	0.7396	
RF 데이터	0.3433	0.3433	
MCA 데이터	0.4489	0.4489	
Mean 데이터			



01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

동질성 검정 결과를 바탕으로  
mean 데이터와 MCA 데이터를 중점적으로 모델링 진행

NA로  
남겨놓은  
기존 데이터

NA → 무응답  
으로 처리한  
데이터

NA → Mean  
으로 채운  
데이터

NA → MCA  
로 채운  
데이터

NA → RF  
로 채운  
데이터

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의



그런데 말입니다...

NA → Mean  
으로 채운  
데이터

NA → MCA  
로 채운  
데이터

두 데이터에 대한 모델링의 성능 발전 X

5개의 데이터 모두 모델링해보자!

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  
기준 데이터

NA  
→ MCA  
로 채운  
데이터

NA →  
Mean  
으로 채운  
데이터

NA → RF  
로 채운  
데이터



넘모 슬퍼,, 우리의 1주차 고생 다 어디에,,



03

# 보통인



# GLM

## *Logistic with Lasso penalty*

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

타겟 변수 Y가 이항변수일 때 사용하는 대표적인 회귀 모델

→ 범주의 타겟은 Democrat VS Republican 적-절!

- logit을 link function으로 사용해 범위 문제 해결!
- 일반적인 회귀 모델에 비해 가정으로부터 자유롭다!

→ 독립성만 만족되면 진행 가능!

3 로지스틱 회귀모형 정의 특징 해석 모형 비교

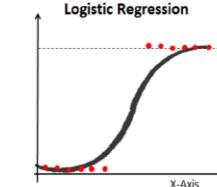
**Logistic Model** : logit을 link function으로 사용

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^T \beta \Leftrightarrow \pi_i = \frac{e^{x_i^T \beta}}{1+e^{x_i^T \beta}}$$

• 범위 문제 해결

$$0 \leq \pi_i \leq 1, \quad 0 \leq 1 - \pi_i \leq 1$$

$$0 \leq \frac{\pi_i}{1-\pi_i} < \infty$$

$$\Rightarrow -\infty < \log\left(\frac{\pi_i}{1-\pi_i}\right) < \infty$$


(로지스틱에 대해 더 궁금하다면, 이번 학기 범주팀 교안을 참고해보자!)

# GLM

## *Logistic with Lasso penalty*

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

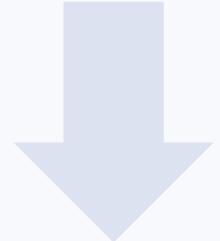
04.  
결과 해석

05.  
한계와 의의

모델에 사용할 변수 선택

→ 로지스틱을 사용하기에 질문 변수만 99개로 Too many..

→ 변수선택법 역시 과도한 computing power 요구!



Lasso penalty를 적용해보자!



(수식적인 배경은 이번학기 회귀팀의 랜쏘를 찾아가보자!)

# GLM

## *Logistic with Lasso penalty*

01.  
1주차 피드백

02.  
DATA 정리

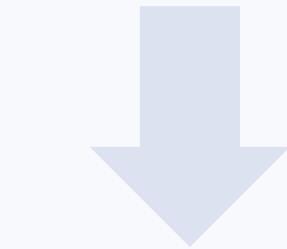
03.  
모델링

04.  
결과 해석

05.  
한계와 의의

모델에 사용할 변수 선택

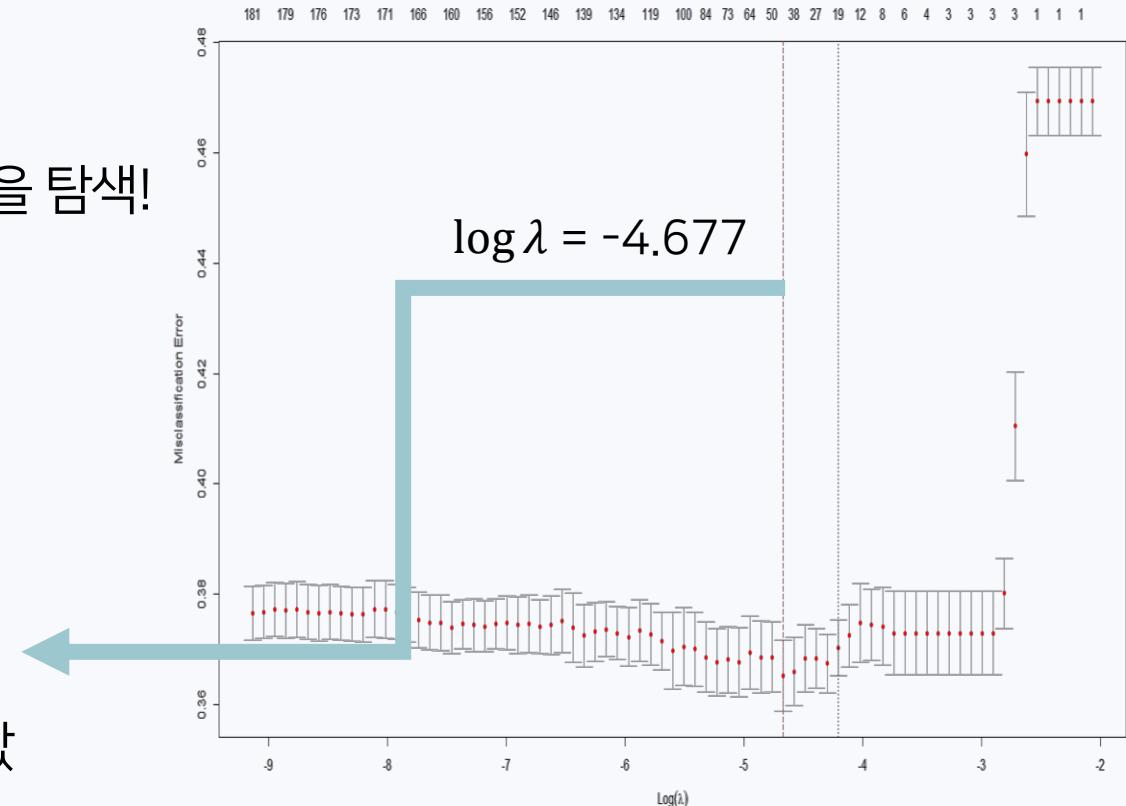
→ CV를 활용하여, 최적의 람다 값을 탐색!



오분류율 최소화



변수선택의 기준이 되는 값



# GLM

## Logistic with Lasso penalty

01.  
1주차 피드백

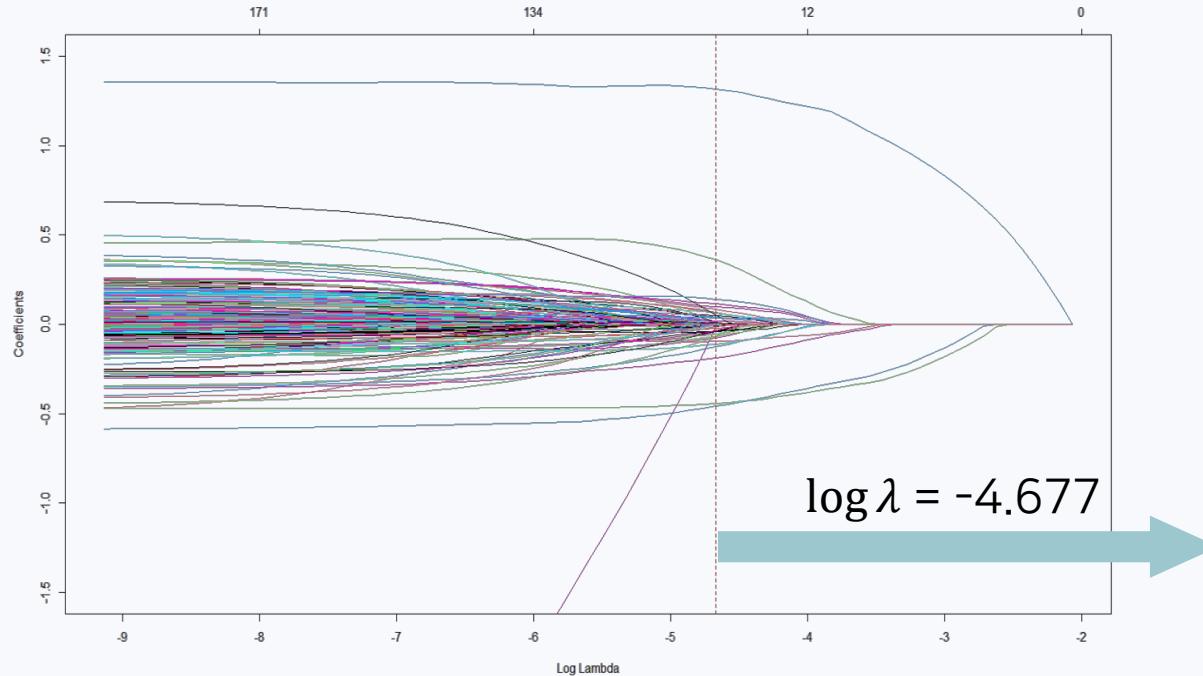
02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

### 모델에 사용할 변수 선택



총 100개의 변수 중 21개의 변수 선택

Gender  
HouseholdStatus  
ps\_Q\_Science\_Art  
Life\_Q\_gun  
Life\_Q\_medicare  
Life\_Q\_MacPC  
ps\_Q\_Feminist,  
ps\_Q\_LifePurpose  
Life\_Q\_drink  
careness  
⋮

(살아남았구나 조나단..!)

# GLM

## *Logistic with Lasso penalty*

01.  
1주차 피드백

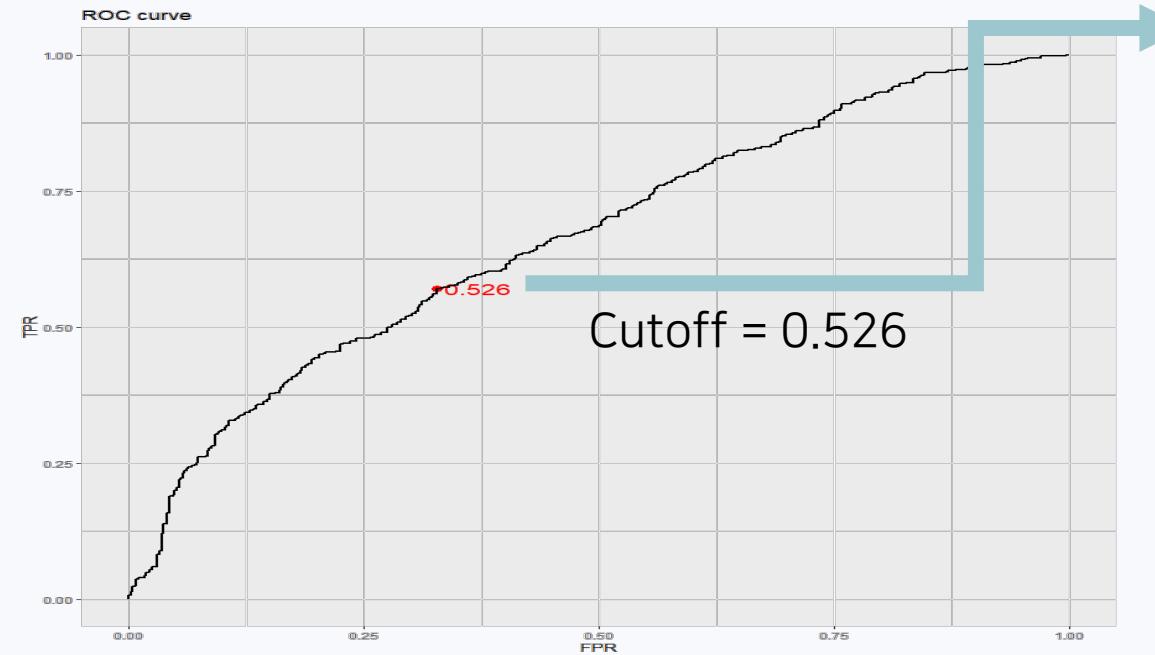
02.  
DATA 정리

03.  
모델링

04.  
결과 해석

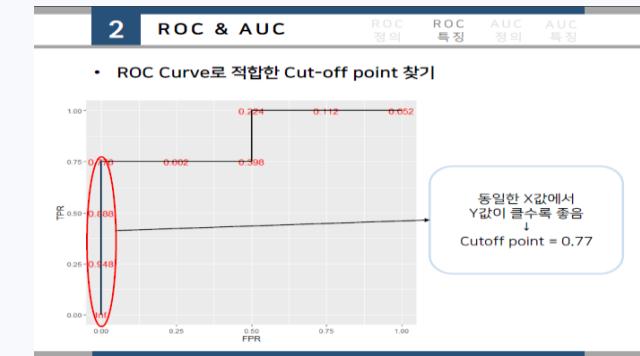
05.  
한계와 의의

ROC curve를 통해 최적의 cut-off값 탐색

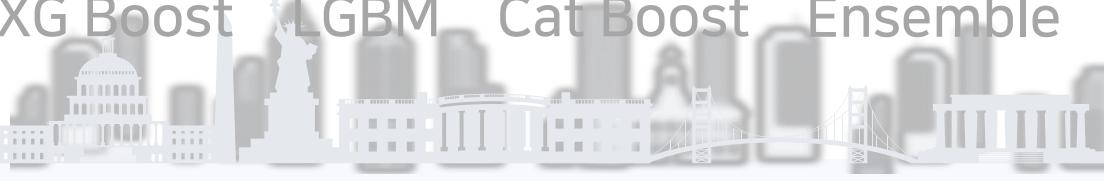


ROC curve는 모든 cutoff를 고려

(0,1)에 가장 가까운  
최적의 cutoff를 탐색!



(다시 한 번 떠올리자 이번학기 범주팀 클린업 3주차..)



# GLM

## *Logistic with Lasso penalty*

01.  
1주차 피드백

02.  
DATA 정리

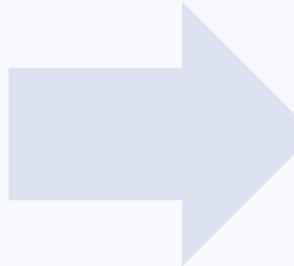
03.  
모델링

04.  
결과 해석

05.  
한계와 의의

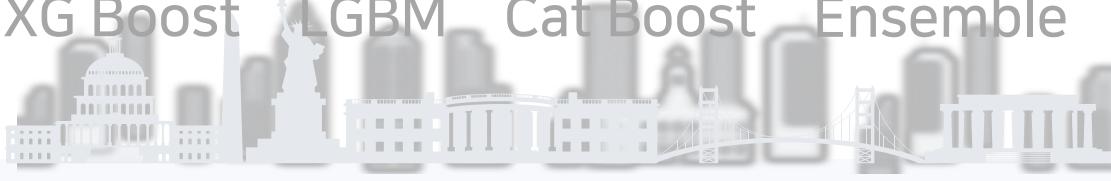
### 최적의 파라미터 조합

- Log lambda : -4.677
- Cutoff : 0.526



Test Accuracy

0.64224



# RandomForest

## Random Forest Classifier

01.  
1주차 피드백

02.  
DATA 정리

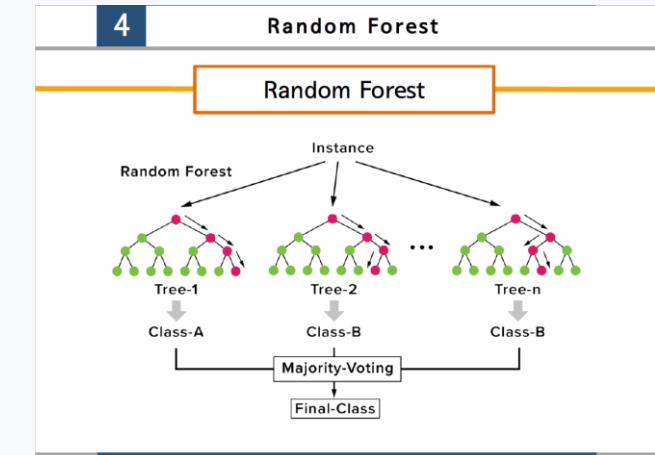
03.  
모델링

04.  
결과 해석

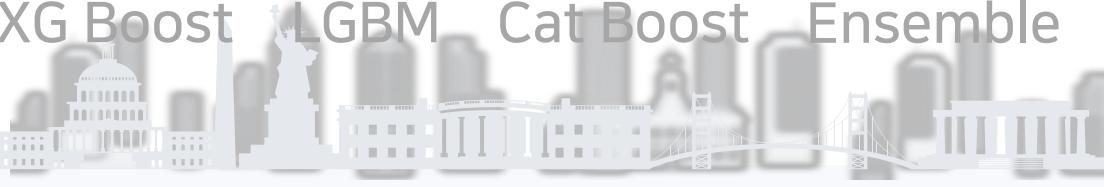
05.  
한계와 의의

개별적으로 학습시킨 여러 개의 트리를 통해 예측을 하는 알고리즘

- 양상을 알고리즘 중 수행 속도가 비교적 빠름
- 다양한 영역에서 높은 예측 성능



(자세한 내용은 지난학기 데마팀의 자료를 참고하자!)



# *RandomForest*

## *Random Forest Classifier*

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

- 모델에 사용할 변수 선택  
→ RFE를 적용해서 선택!

### 선택된 변수들

careness  
fairness  
ps\_Q\_Feminist  
Life\_Q\_medicpray  
Life\_Q\_gun  
Life\_Q\_MacPC  
Ps\_Q\_LifePurpose  
:

### RFE(Recursive Feature Elimination)란?

모든 변수를 다 포함시킨 후 반복해서 학습을 진행하면서  
중요도가 낮은 변수를 하나씩 제거하는 방식  
→ 일종의 Backward Selection 방법



01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

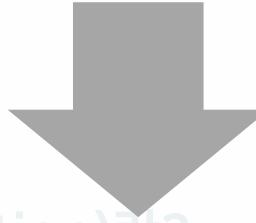
04.  
결과 해석

05.  
한계와 의의

Multicollinearity가 존재할 때 변수의 중요도를 이용하여  
변수를 삭제하는 것은 문제가 될 수도 있다!

- 모델에 사용할 변수 선택

→ RFE를 적용해서 선택!



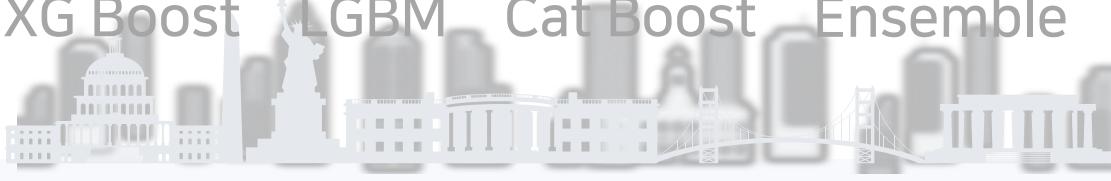
선택된 변수들

careness  
fairness  
ps\_Q\_Feminist  
Life\_Q\_medicpray  
Life\_Q\_gun  
Life\_Q\_MacPC  
LifePurpose

- RFE(Recursive Feature Elimination)란?
- Gktau값을 통해 연관성이 높은 변수는 사전에 제거함
  - Shapley value를 통해 나온 변수들과 비교하는 과정을 거침

모든 변수를 다 포함하면서 차례로 변수를 제거하면서  
중요도가 낮은 변수를 하나씩 제거하는 방식

→ 일종의 Backward Selection 방법



# *RandomForest*

## *Random Forest Classifier*

- 랜덤 포레스트 파라미터

파라미터	파라미터 설명
n_estimators	결정 트리의 개수
max_features	분할시 고려할 최대 feature 개수
max_depth	트리의 최대 깊이
min_samples_split	노드를 분할하기 위한 최소 샘플 수
min_samples_leaf	말단 노드의 최소 샘플 수

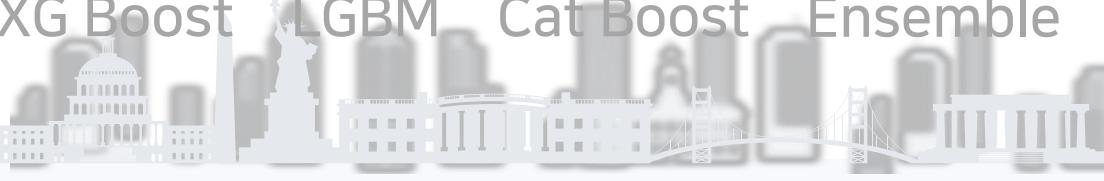
01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의



# RandomForest

## Random Forest Classifier

- 랜덤 포레스트 파라미터

파라미터	파라미터 설명
n_estimators	결정 트리의 개수
max_features	분할시 고려할 최대 feature 개수
max_depth	트리의 최대 깊이
min_samples_split	노드를 분할하기 위한 최소 샘플 수
min_samples_leaf	말단 노드의 최소 샘플 수

클수록 성능 좋을 수 있음  
깊이 깊어지면 과적합 가능성  
작게 설정하면 과적합 가능성

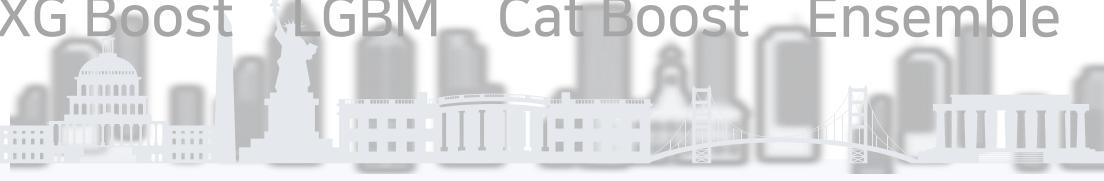
01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의



# RandomForest

## Random Forest Classifier

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

n\_estimators [200,2000]

max\_features ['auto', 'sqrt', 'log2']

max\_depth [10,110]

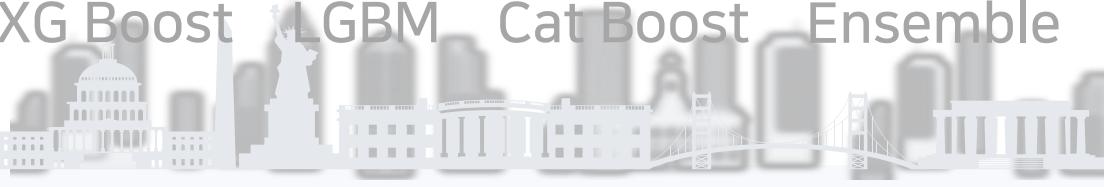
min\_samples\_split [2,20]

min\_samples\_leaf [1,10]

3-Fold CV  
&  
Random Search

RandomSearch  
최적의 파라미터 조합

- n\_estimators: 1200
- max\_features: 'sqrt'
- max\_depth: 20
- min\_samples\_split: 14
- min\_samples\_leaf: 9



# RandomForest

## Random Forest Classifier

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

n\_estimators [1000,1200,1400]

max\_features ['sqrt']

max\_depth [10,20,30]

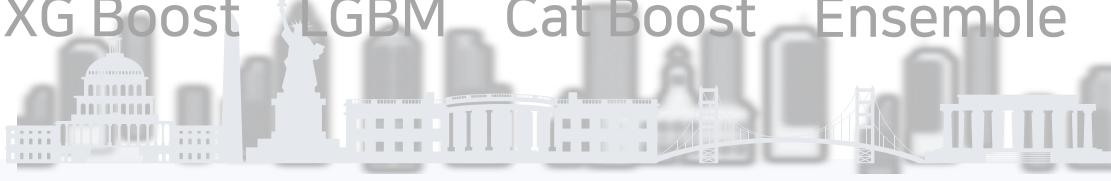
min\_samples\_split [12,14,16]

min\_samples\_leaf [8,9,10]

3-Fold CV  
&  
Grid Search

### 최적의 파라미터 조합

- n\_estimators: 1200
- max\_features: 'sqrt'
- max\_depth: 30
- min\_samples\_split: 16
- min\_samples\_leaf: 9



# RandomForest

## Random Forest Classifier

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

변수 선택

최적의 파라미터 조합

RFE를 통해  
200개의 변수  
선택

- n\_estimators: 1200
- max\_features: 'sqrt'
- max\_depth: 30
- min\_samples\_split: 16
- min\_samples\_leaf: 9

Test Accuracy  
0.64511

# XGBoost

*extreme Gradient Boosting*

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

Gradient Boosting 알고리즘 중 하나로,  
**병렬처리와 최적화**를 장점으로 내세우는 알고리즘

- CART(Classification And Regression Trees)를 기반으로 만들어짐
- 다양한 Hyper-parameter 존재
  - 하이퍼 파라미터를 얼마나 잘 조정하는지가 중요!



2014년 등장 이후 Kaggle 대회를 휩쓸며

Boosting 계열에서 많은 사랑을 받고 있는 모델!

01.  
1주차 피드백

02.  
DATA 정리

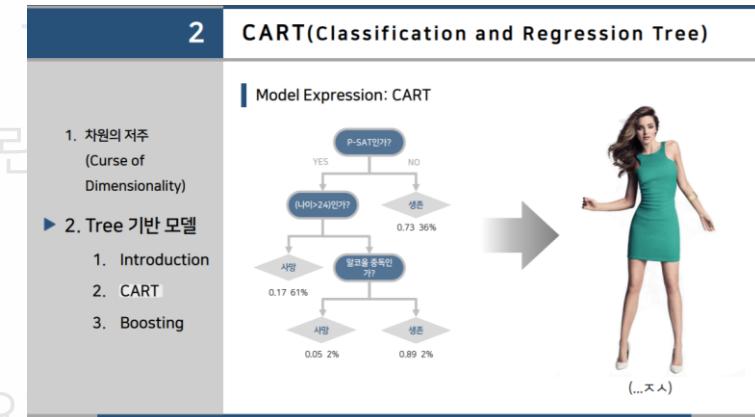
03.  
모델링

04.  
결과 해석

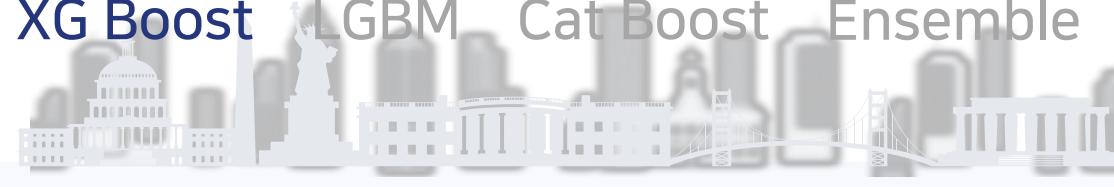
05.  
한계와 의의

## 장점

- 뛰어난 예측 성능
- 과적합 규제(Overfitting Regularization)
- Tree pruning(트리 가지치기)
- 자체 내장된 교차 검증
- 광범위한 하이퍼파라미터 튜닝
- 단점**
- 학습 시간이 오래 걸린다



(CART가 궁금하다면 20년도 2학기 데마 2주차 클린업 자료를 참고하자!)



# XGBoost

*extreme Gradient Boosting*

- XGBoost 파라미터 튜닝

파라미터	파라미터 설명
max_depth	트리의 최대 깊이
min_child_weight	자식노드에 필요한 최소 인스턴스 수
learning_rate	각 예측기마다의 학습의 가중치
n_estimators	결정 트리의 개수
gamma	트리의 가지치기 조정
reg_lambda	정규화 변수 람다를 조정

# XGBoost

*extreme Gradient Boosting*

- XGBoost 파라미터 튜닝

파라미터	파라미터 설명
max_depth	트리의 최대 깊이
min_child_weight	자식노드에 필요한 최소 인스턴스 수
learning_rate	각 예측기마다의 학습의 가중치
n_estimators	결정 트리의 개수
gamma	트리의 가지치기 조정
reg_lambda	정규화 변수 람다를 조정

값이 클수록  
보수적인 모델이  
된다

# XGBoost

*extreme Gradient Boosting*

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

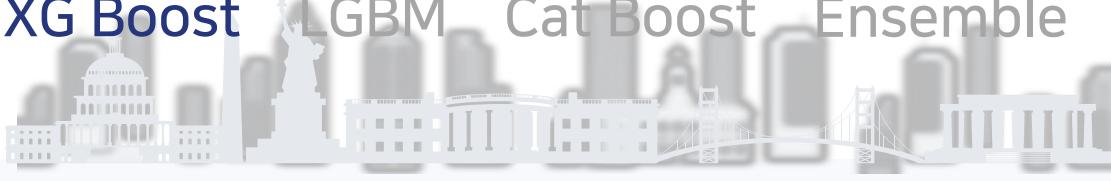
05.  
한계와 의의

max_depth	[3, 4, 5]
min_child_weight	[1, 2, 3, 5]
learning_rate	[1, 0.5, 0.1, 0.01, 0.05]
n_estimators	[50, 100, 200, 500]
gamma	[0, 0.25, 1.0]
reg_lambda	[0, 1, 10, 20, 100]

5-Fold CV  
&  
Grid Search

## 최적의 파라미터 조합

- max\_depth : 4
- min\_child\_weight: 1
- learning\_rate : 0.1
- n\_estimators: 100
- gamma : 0
- reg\_lambda : 20



# XGBoost

*extreme Gradient Boosting*

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

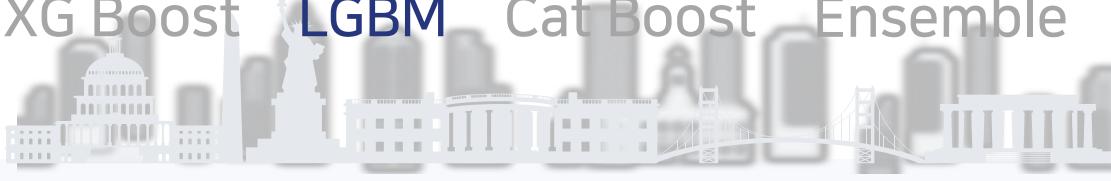
05.  
한계와 의의

## 최적의 파라미터 조합

- max\_depth : 4
- min\_child\_weight: 1
- learning\_rate : 0.1
- n\_estimators: 100
- gamma : 0
- reg\_lambda : 20

Predict

Test Accuracy  
0.65948

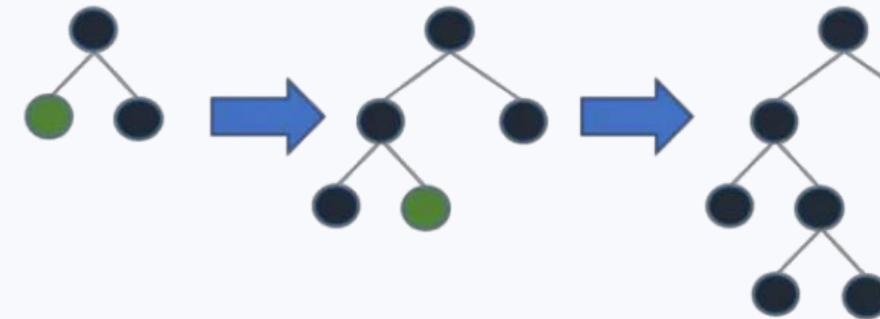


# **LGBM**

## *Light Gradient Boosting Model*

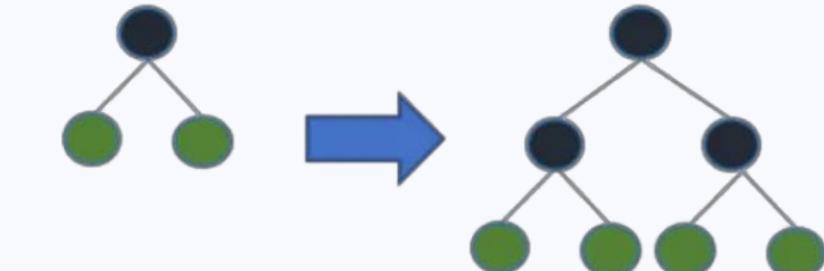
: XG Boost의 느린 학습속도를 극복하기 위해 개발된 알고리즘

- 빠른 학습과 예측 수행 시간, 이에 비해 작은 메모리 사용량
- 기존의 gradient boosting 알고리즘과 다르게 leaf-wise 트리분할을 사용



*<Leaf-wise tree growth>*

VS



*<Level-wise tree growth>*



# Light GPM

차이점을 제대로 알고 가보자!

01.  
1주차 피드백

02.  
DATA 정리

03.  
**모델링**

04.  
결과 해석

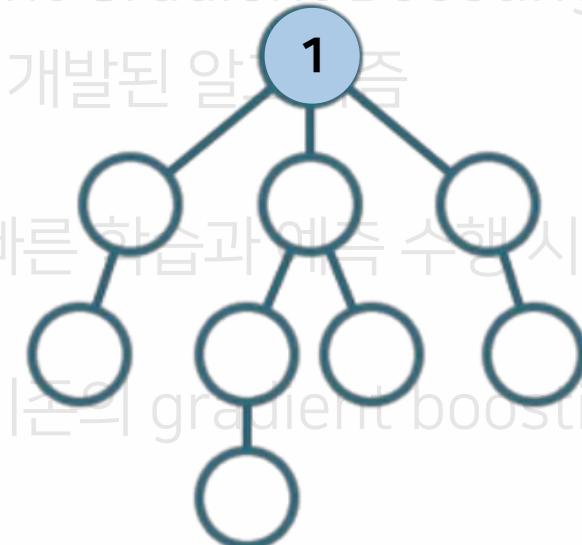
05.  
한계와 의의

## *<Leaf-wise tree growth>*

: Light Gradient Boosting Model, XG Boost의 느린 학습속도를 극복하기

위해 개발된 알고리즘

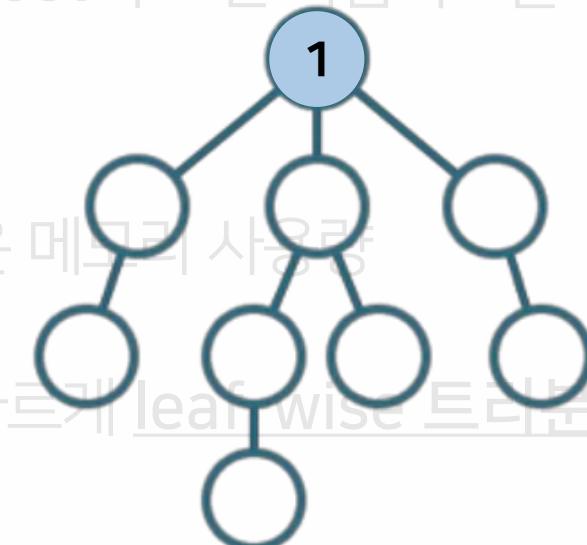
- 빠른 학습과 예측 수행 시간, 이에 비해 작은 메모리 사용량
- 기존의 gradient boosting 알고리즘과 다르게 leaf-wise 트리 분할을 사용



*LGBM*의 알고리즘

## *<Level-wise tree growth>*

: Light Gradient Boosting Model, XG Boost의 느린 학습속도를 극복하기



*Ef boosting* 모델의 알고리즘



# Light GPM

차이점을 제대로 알고 가보자!

01.  
1주차 피드백

02.  
DATA 정리

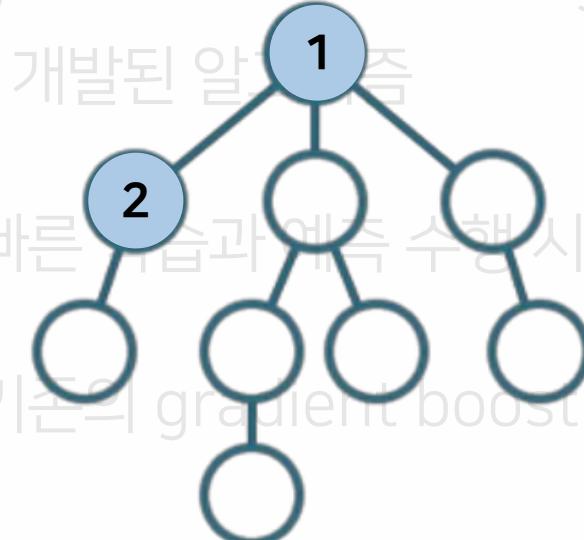
03.  
**모델링**

04.  
결과 해석

05.  
한계와 의의

## *<Leaf-wise tree growth>*

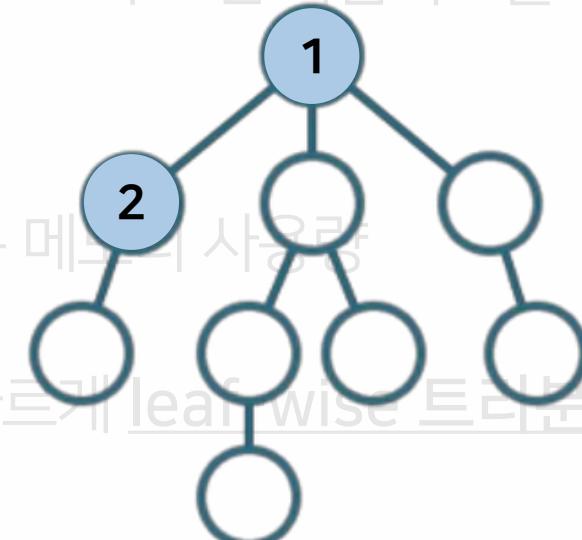
: Light Gradient Boosting Model, XG Boost의 느린 학습속도를 극복하기 위해 개발된 알고리즘



*LGBM*의 알고리즘

## *<Level-wise tree growth>*

: 기존의 gradient boosting 알고리즘과 다르게 leaf-wise 트리 분할을 사용



*Ef boosting* 모델의 알고리즘



# Light GPM

차이점을 제대로 알고 가보자!

01.  
1주차 피드백

02.  
DATA 정리

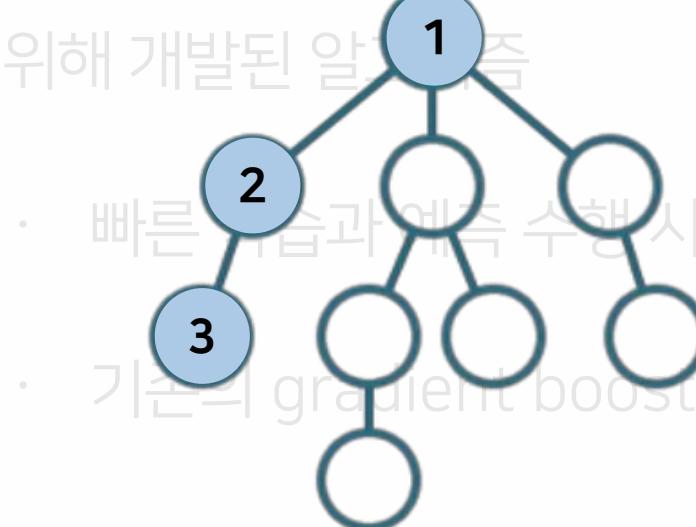
03.  
**모델링**

04.  
결과 해석

05.  
한계와 의의

## *<Leaf-wise tree growth>*

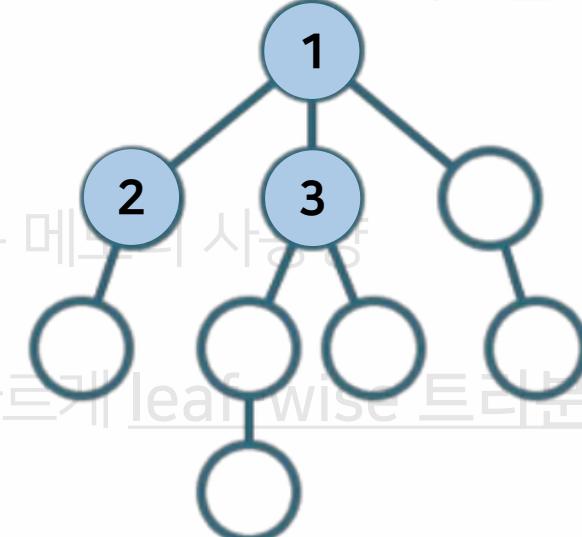
: Light Gradient Boosting Model, XG Boost의 느린 학습속도를 극복하기 위해 개발된 알고리즘



*LGBM*의 알고리즘

## *<Level-wise tree growth>*

: 기존의 gradient boosting 알고리즘과 다르게 level-wise 트리 분할을 사용



*Ef boosting* 모델의 알고리즘



## Light GPM

차이점을 제대로 알고 가보자!

01.  
1주차 피드백

02.  
DATA 정리

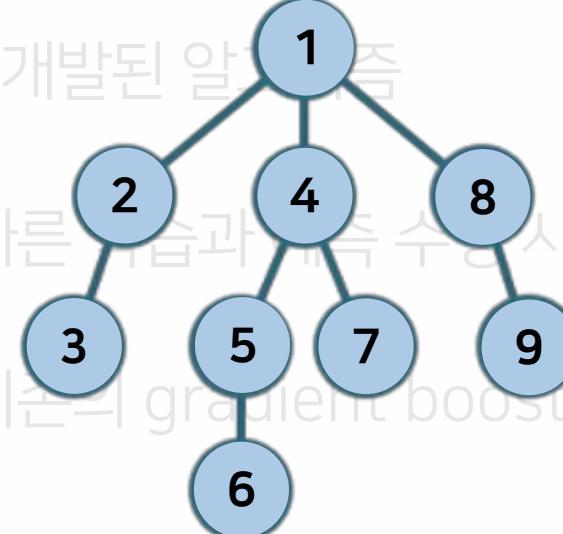
03.  
**모델링**

04.  
결과 해석

05.  
한계와 의의

*<Leaf-wise tree growth>*

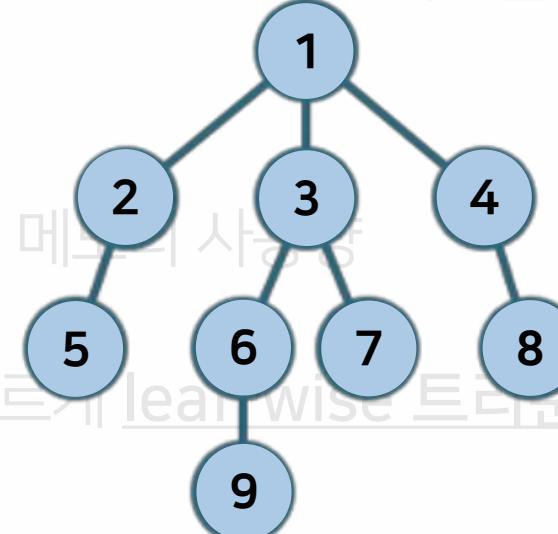
: Light Gradient Boosting Model, XG Boost의 느린 학습속도를 극복하기 위해 개발된 알고리즘



*LGBM*의 알고리즘

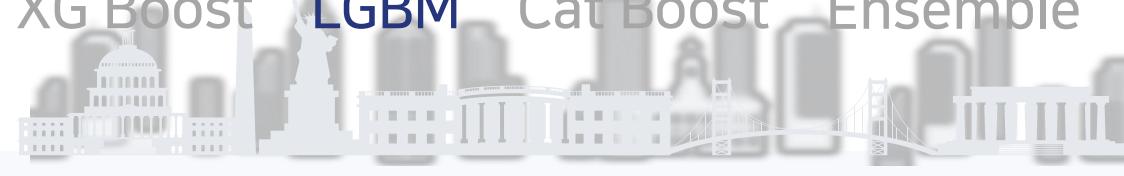
*<Level-wise tree growth>*

: 기존의 gradient boosting 알고리즘과 다르게 level-wise 트리 분할을 사용



*Cat boosting* 모델의 알고리즘

이런 방식이기 때문에 타 부스팅 모델에 비해 빠른편!



# *Light GBM*

## *Light Gradient Boosting Model*

- LGBM은 튜닝할 수 있는 파라미터가 굉장히 많다

파라미터	파라미터 설명
application	모델의 어플리케이션 설정 Ex) regression, binary, multiclass
boosting	실행하고자 하는 알고리즘 타입
learning_rate	각 예측기마다의 학습의 가중치
n_estimators	결정 트리의 개수
num_leaves	전체 트리의 leave 수
bagging_fraction	데이터 일부 사용하는 bagging 비율

# *Light GBM*

## **Light Gradient Boosting Model**

- LGBM 파라미터 튜닝 과정 (Bayesian Optimization)

num\_leaves

subsample

learning\_rate

reg\_alpha

n\_estimators

bagging\_  
fraction

이 과정에서 몇몇 파라미터 고정 후

후 핵심 파라미터들에 대해

Grid Search 진행

:

Bayesian Optimization을 통해  
다양한 파라미터의 범위 탐색

# 모델링

GLM Random Forest XG Boost LGBM Cat Boost Ensemble

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

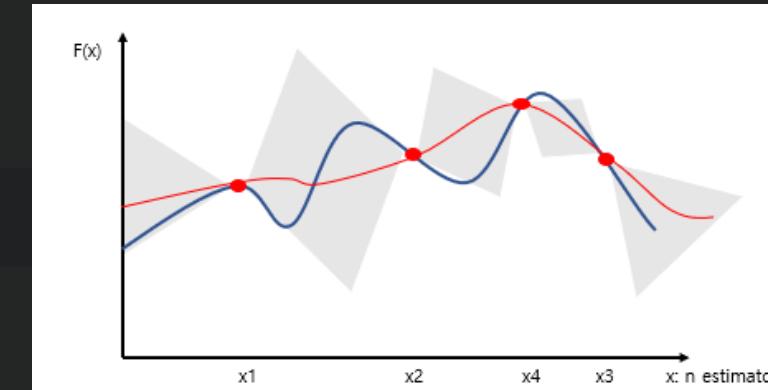
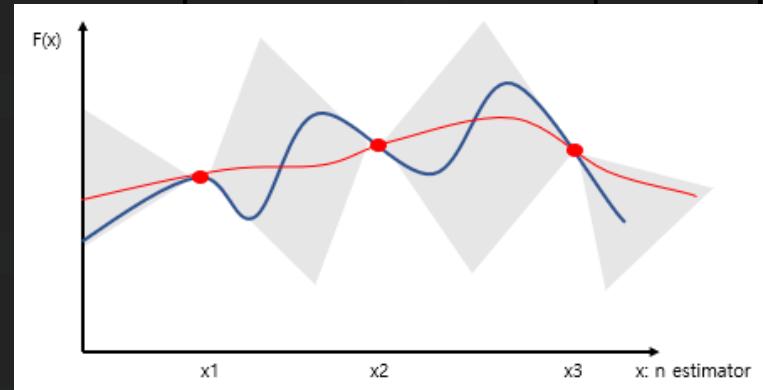
04.  
결과 해석

05.  
한계와 의의



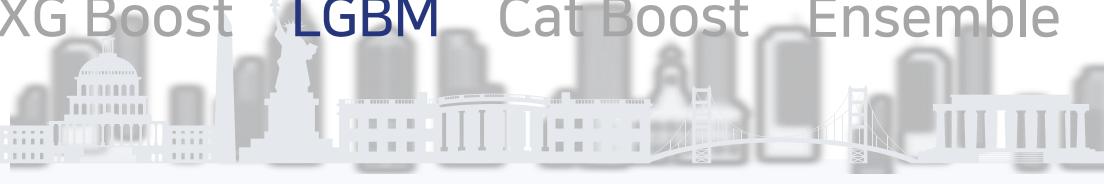
## Bayesian Optimization이란?

- Light Gradient Boosting Model : 이전의 탐색결과(prior knowledge)를 다음 탐색에 반영하면서 LGBM 파라미터 튜닝하는 방법



Surrogate model을 통해  
현재까지 조사된 데이터를 바탕으로 통해  
목적함수 추정  
다양한 파라미터의 범위 탐색

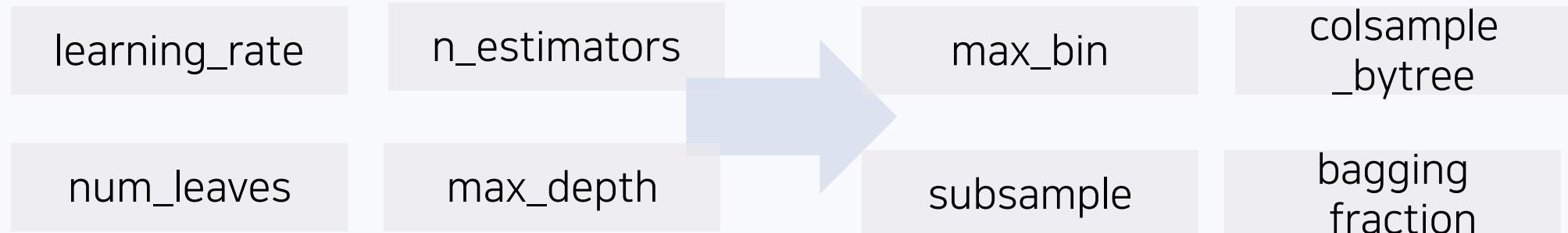
Acquisition Function을 통해  
다음 입력값을 추천받아  
그 관측치를 포함해 다시 목적함수 추정



# *Light GBM*

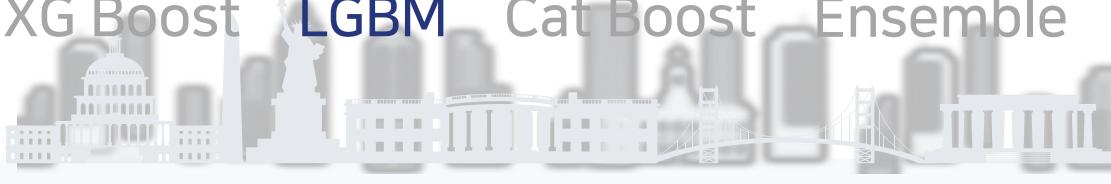
## **Light Gradient Boosting Model**

- LGBM 파라미터 튜닝 과정 (Grid Search)



가장 핵심적인 파라미터 먼저  
Grid Search로 탐색해 고정

이후 추가적인 파라미터 튜닝



# *Light GBM*

## *Light Gradient Boosting Model*

01.  
1주차 피드백

02.  
DATA 정리

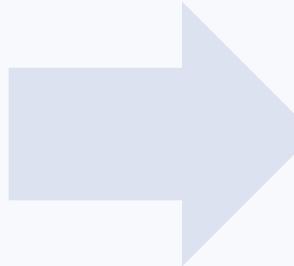
03.  
**모델링**

04.  
결과 해석

05.  
한계와 의의

### 최적의 파라미터 조합

- learning\_rate:0.003
- n\_estimators : 900
- num\_leaves: 30
- reg\_alpha: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**



# CatBoost

*Categorical Boosting*

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

: Gradient Boosting 알고리즘 중 하나로,  
**범주형 변수**가 많은 데이터를 처리하는데 유용한 알고리즘

- 범주형 변수에 대해 **자체적인 인코딩** 진행  
→ 빠른 속도 & 중요한 feature를 잃는 것을 막음
- **Ordered Boosting** 기법 사용  
→ overfitting 방지 & 효율적인 테스트



# CatBoost

*Categorical Boosting*

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

파라미터	파라미터 설명
iterations	생성할 수 있는 최대 tree 개수
depth	tree의 깊이
learning_rate	각 예측기마다의 학습의 가중치
l2_length_reg	비용 함수의 L2 regularization term의 계수
random_strength	각 split에 점수를 매길 때 사용할 randomness의 양 → 과적합 방지



# CatBoost

*Categorical Boosting*

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

depth

iterations

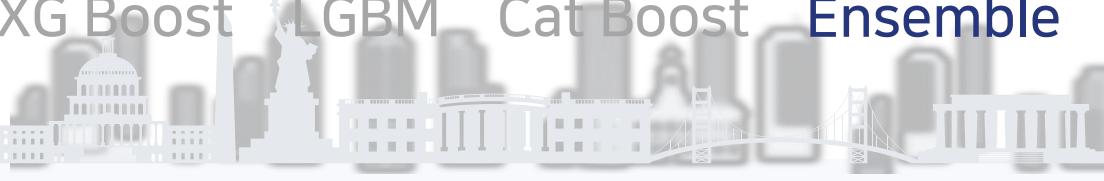
learning\_rate

L2\_leaf\_reg

random\_strength

Grid Search를 통해  
최적의 파라미터 고정

Bayesian Optimization을 통해  
최적의 파라미터 고정



# CatBoost

*Categorical Boosting*

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

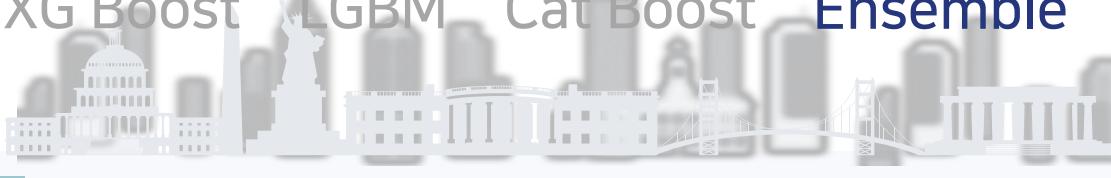
05.  
한계와 의의

## 최적의 파라미터 조합

- iterations : 400
- depth : 7
- random\_strength : 0.099234
- learning\_rate : 0.018487
- l2\_leaf\_reg : 8



Test Accuracy  
**0.64367**



# Ensemble

## Stacking Ensemble

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

: 여러 모델들을 활용해 각각의 예측 결과를 도출한 뒤,  
그 예측 결과를 결합해 최종 모델로 학습해 최종 예측 결과를 만들어내는 것

- 단일 모델보다 성능 향상  
(가끔은 함께보단 혼자가 좋을 때도..)
- Overfitting의 가능성 ↑  
→ CV 세트 기반 Stacking Ensemble 진행



01.  
1주차 피드백

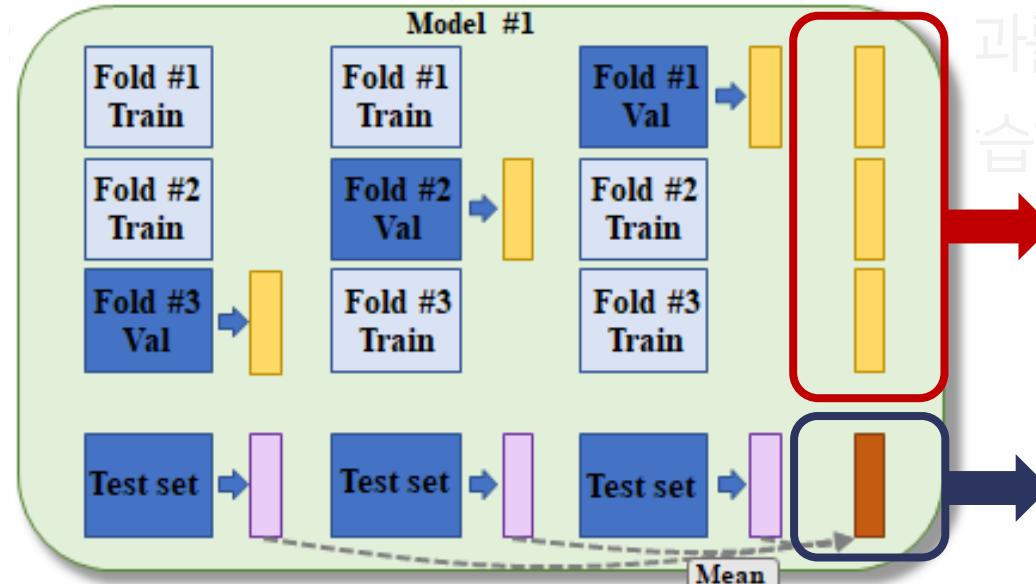
02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

## STEP 1



과를 도출한 뒤,  
습해 최종 예측 결과를 만들어내는 것

각 validation fold마다  
예측 값을 계산하여 결합

각 fold에서 얻은 모델을 통해  
test set에 대해 예측하고,  
그 값들의 평균을 냄

→ CV 세트 기반 Stacking Ensemble 진행

01.  
1주차 피드백

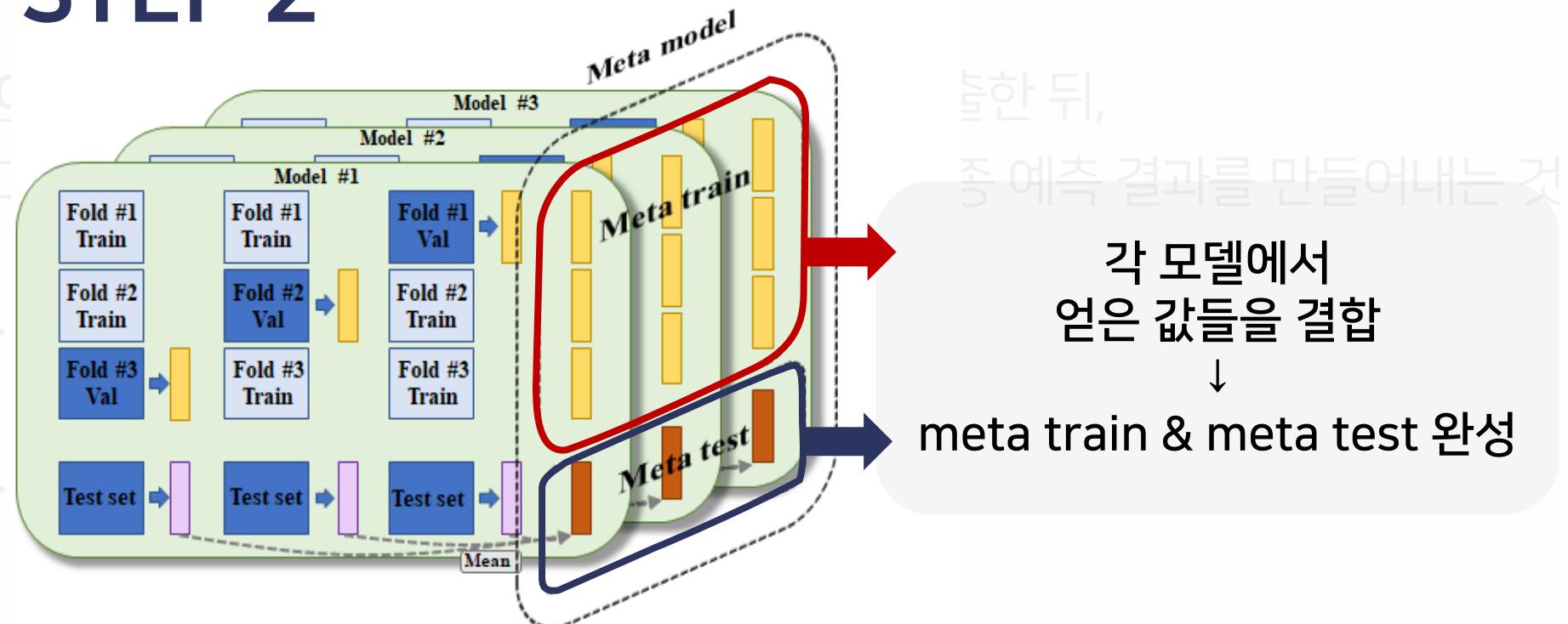
02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

## STEP 2



01.  
1주차 피드백

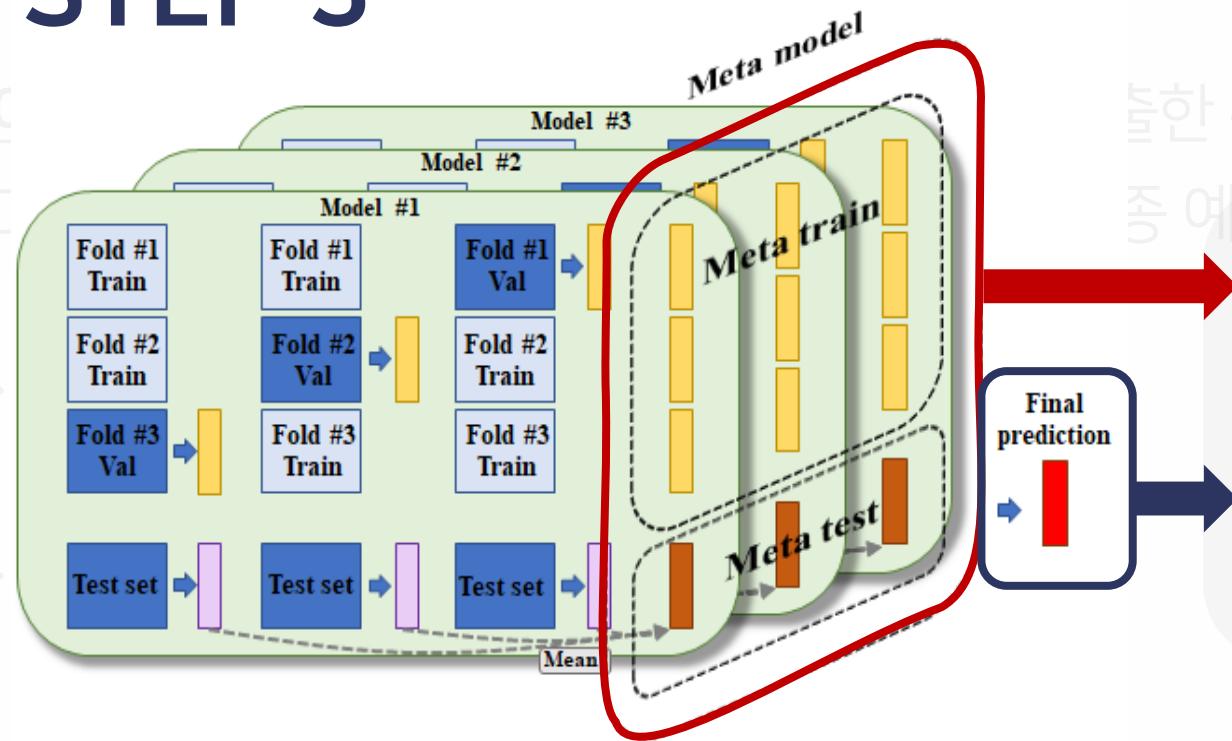
02.  
DATA 정리

03.  
모델링

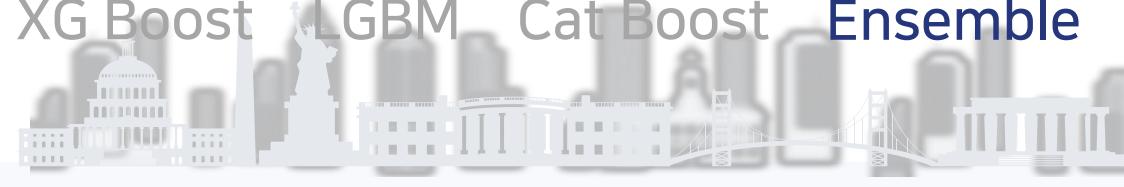
04.  
결과 해석

05.  
한계와 의의

## STEP 3



meta model로  
meta train을 훈련  
↓  
meta test에 대한  
최종 예측값 출력



# Ensemble

## Stacking Ensemble

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

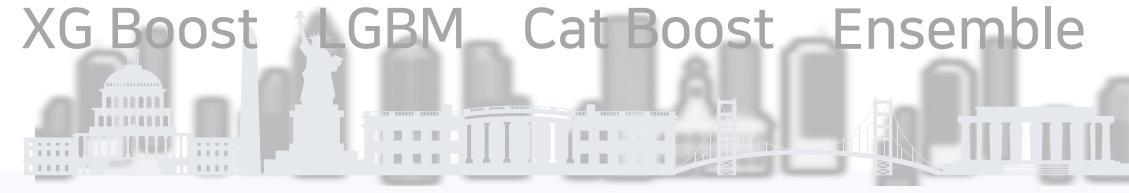
### 사용한 모델

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

meta model  
: XGBoost

Test Accuracy  
0.65517

찢어따... 앙상블...



## 총 365번의 파일 제출 결과...

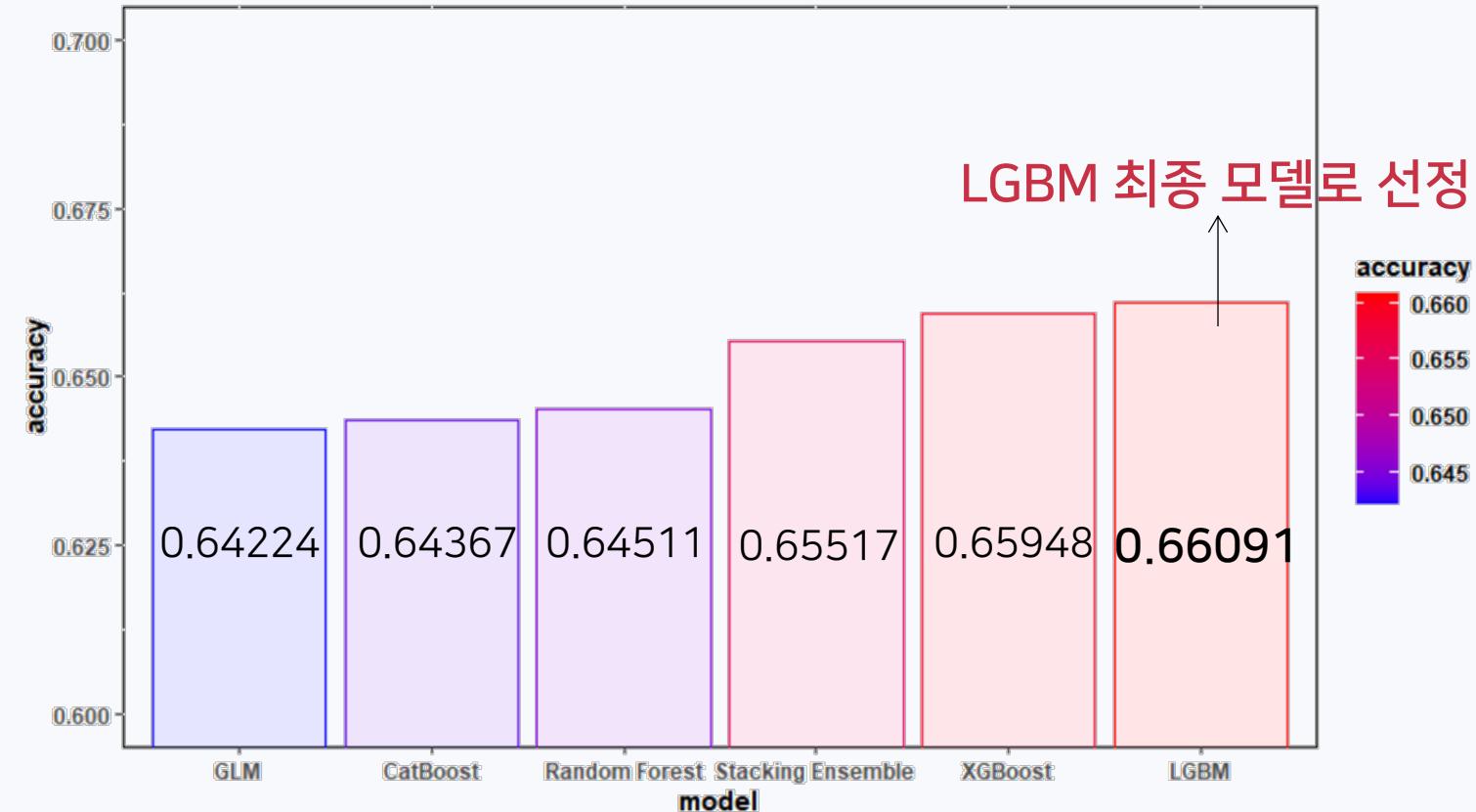
01.  
1주차 피드백

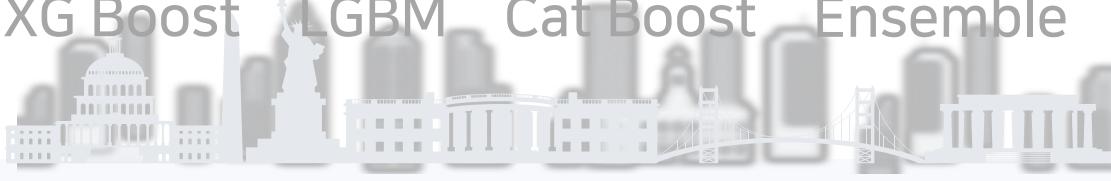
02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의





총 365번의 파일 제출 결과...



이런 우리의 K-열정으로 인해

## Error: Server Error

The server encountered a temporary error and could not complete your request.

Please try again in 30 seconds.

한때 사이트가 마비가 되었다고 한다^^ 의지의 한국인의 힘을 보여준 행복 범주

Kaggle 별거 아니네^^?

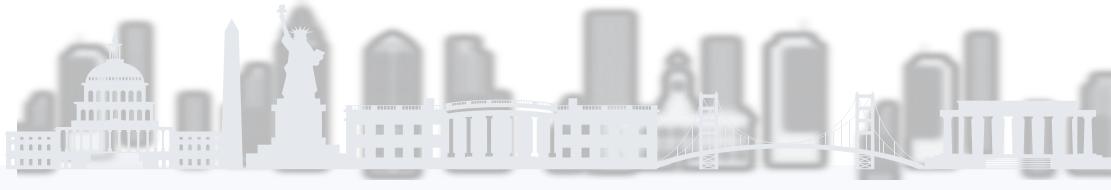


04

# 결과 해석



# 모델링 결과 해석



- What is SHAP?

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

# SHAP

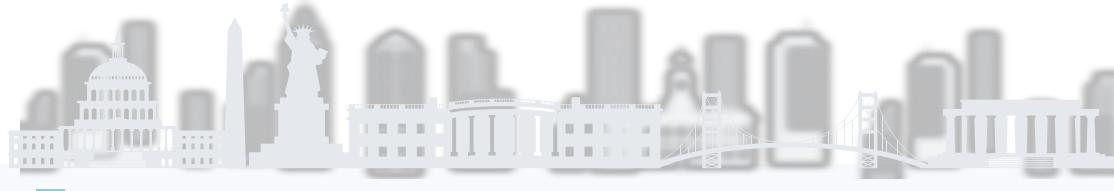
Shapley Additive exPlanations

SHAP value를 활용해 독립 변수가 타겟 변수에 어떤 영향을 미치는지

그 영향의 방향과 정도를 확인할 수 있는 방법

변수 선택 단계에서 선택한 변수에 대한 타당성을  
뒷받침하는 지표로도 활용 가능하다!

## 모델링 결과 해석



# SHAP value

## Shapley value

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

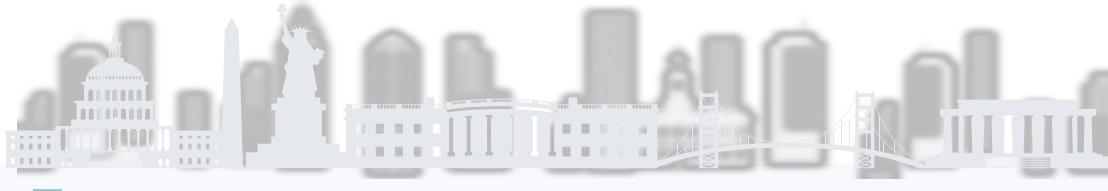
05.  
한계와 의의

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

(갓카피디아에서 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



# SHAP value

## Shapley value

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

특정 결과에 각 특성이 얼마나 영향을 미쳤는지 나타내는 수치

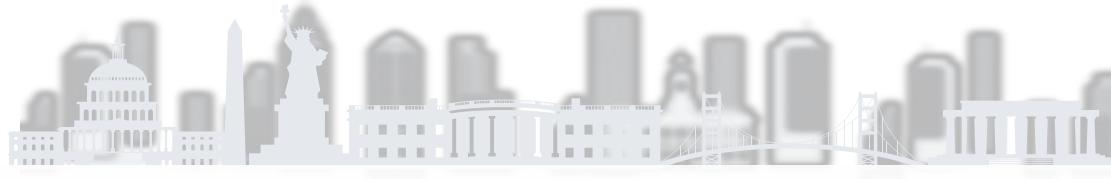
→ 게임이론에서 파게임이론이 뭔..지도 말해주고 가시죠!

어떠한 행위의 결과가 자신 뿐 아니라 다른 참가자에 의해서도 결정되는 상황에서  
자신의 최대 이익에 부합하는 행동을 추구한다는 **상호 의존적인 의사결정에 관한 수학적 이론**

(갓카피디아에서 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



Shapley value 예시까지 말하고 갈 겁니다..

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

전성기 박지성 선수와 현재 손흥민 선수 중 누가 더 레전드인가..?



(예시가 선 쌤네..?)

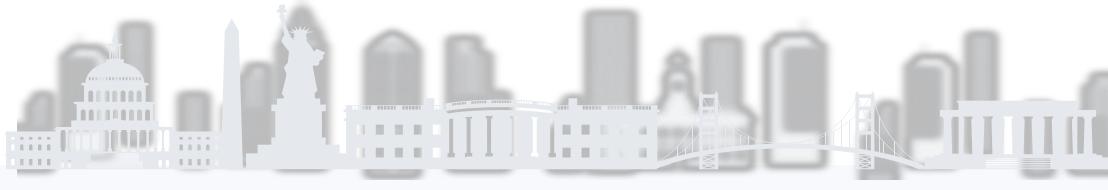
(형 나 레알가면 몰라..)

위 주제를 논할 때,

→ 소속팀에서 두 선수가 빠진다면 소속팀은 어떻게 될까?

[마감일 기준으로] 같은 대회에서 같은 결승전에 진다니깐 뺀 값

## 모델링 결과 해석



# SHAP value

## Shapley value

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

### Feature importance

모델 성능 감소에 기반한 방법

영향에 대한 방향성 설명 불가

Inconsistency한 특성을 가진다!

### SHAP value

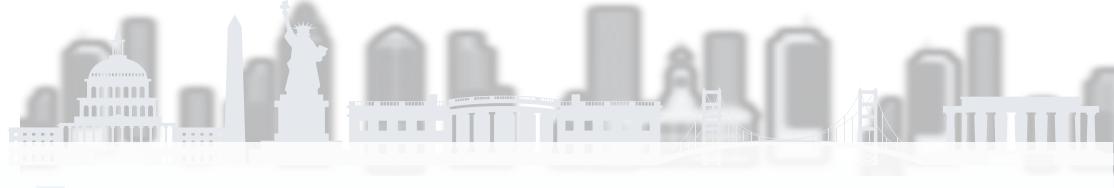
특성 기여도에 기반한 방법

영향에 대한 방향성 설명 가능

Consistency한 특성을 가진다!



# 모델링 결과 해석



01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

## Inconsistency? Consistency?

### Shapley value

#### Inconsistency

Feature importance와의 차이점

변수 중요도를 구하는 기준(Weight, Gain, Cover 등)에 따라 결과 값이 다르므로  
상대적으로 일관성이 적다!

모델 성능 감소에 기반한 방법

특성 기여도에 기반한 방법

#### Consistency

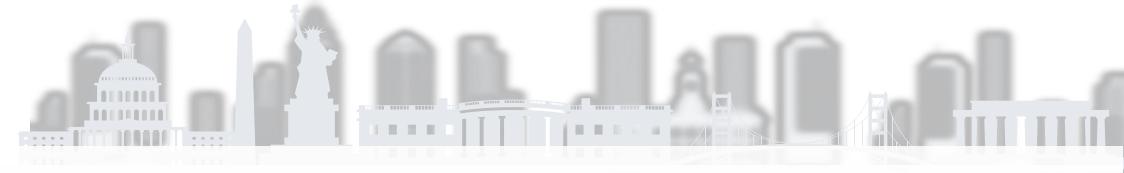
영향에 대한 방향성 설명 불가

영향에 대한 방향성 설명 가능

특성과 관련된 모든 조합에 대해 여러 번 연산을 반복해 수행한 결과의 평균 값

Inconsistency한 특성을 가진다! Consistency한 특성을 가진다!

# 모델링 결과 해석



본격적으로 해석하기 이전, 스치듯 도메인 복습..

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

## Republicans

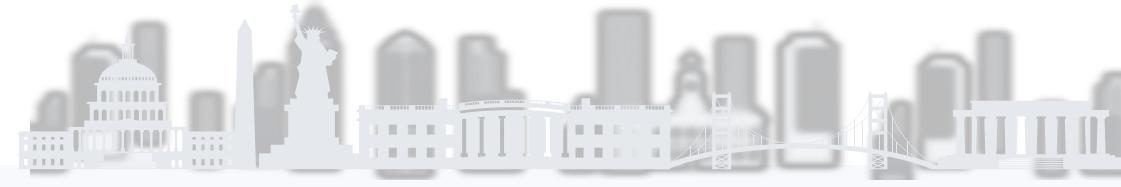
- 백인 남성 / 노인
- 자유경제
- 높은 소득 수준
- 총기 규제 완화
- 기업과 개인의 자유
- 현실적

## Democrats

- 여성 / 성소수자
- 사회경제적 평등
- 낮은 소득 수준
- 총기 규제 강화
- 노조 권리 보장
- 이상적

<미국 공화당과 민주당에 대한 일반적인 인식>

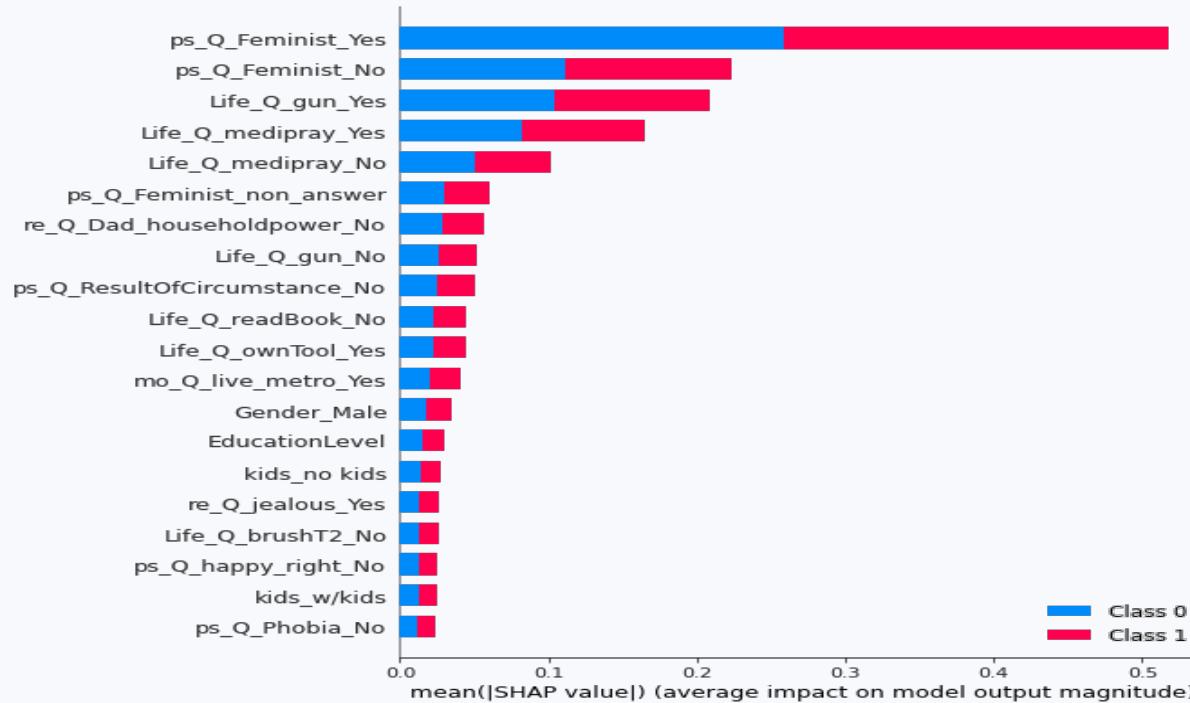
# 모델링 결과 해석



# 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

01.  
1주차 피드백

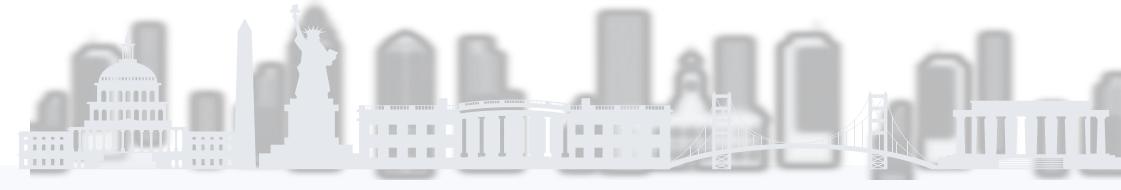
02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

# 모델링 결과 해석



# SHAP value

## Shapley value

01.  
1주차 피드백

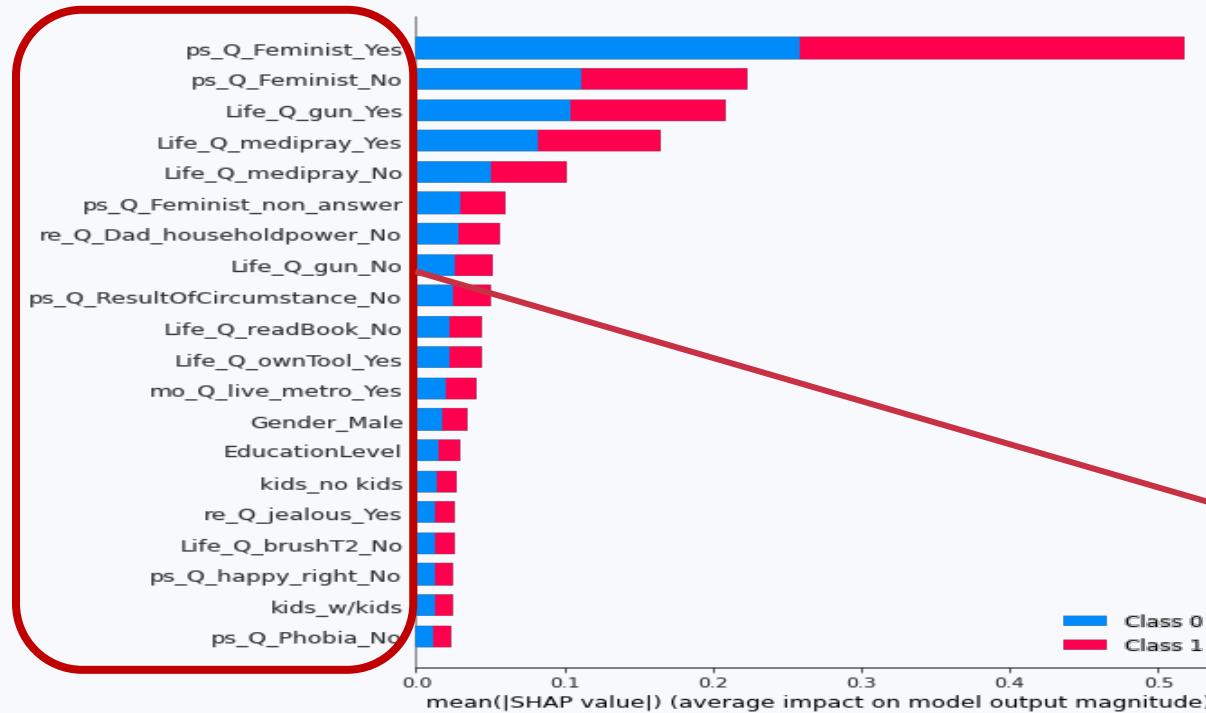
02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

What is shap.summary plot?



shap. summary plot

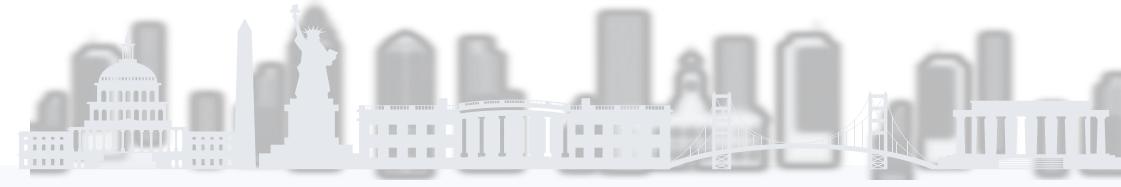
타겟 변수에 대한 X변수들의  
평균적인 영향력을 나타낸 Plot

Class 0 → Democrat

Class 1 → Republicans

Y축 : 영향력이 높은 순서대로 변수 나열

# 모델링 결과 해석



# SHAP value

## Shapley value

01.  
1주차 피드백

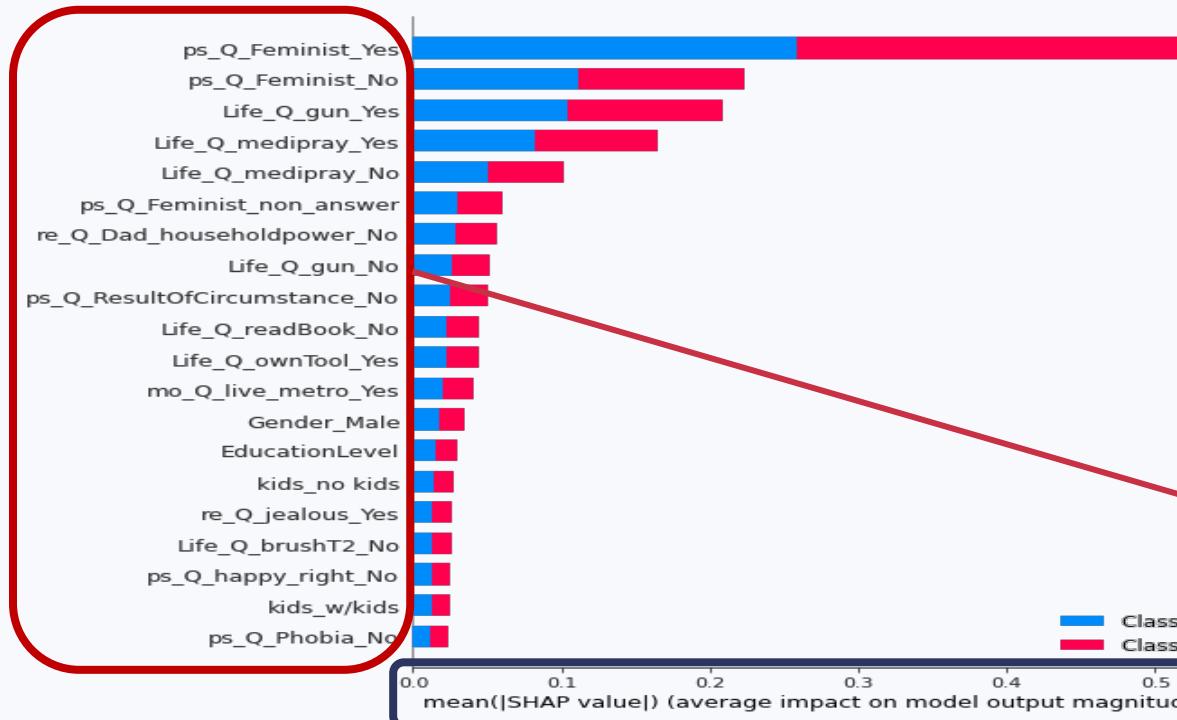
02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

What is shap.summary plot?



shap. summary plot

타겟 변수에 대한 X변수들의  
평균적인 영향력을 나타낸 Plot

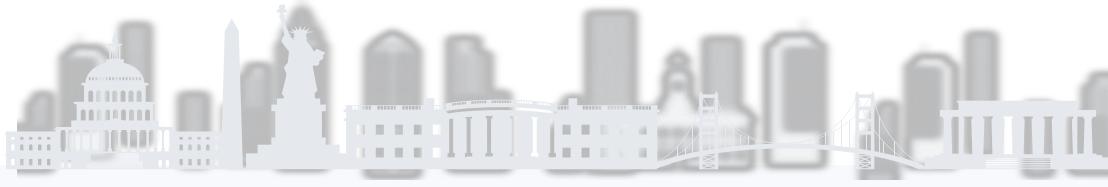
Class 0 → Democrat

Class 1 → Republicans

Y축 : 영향력이 높은 순서대로 변수 나열

X축 : mean|Shap value|

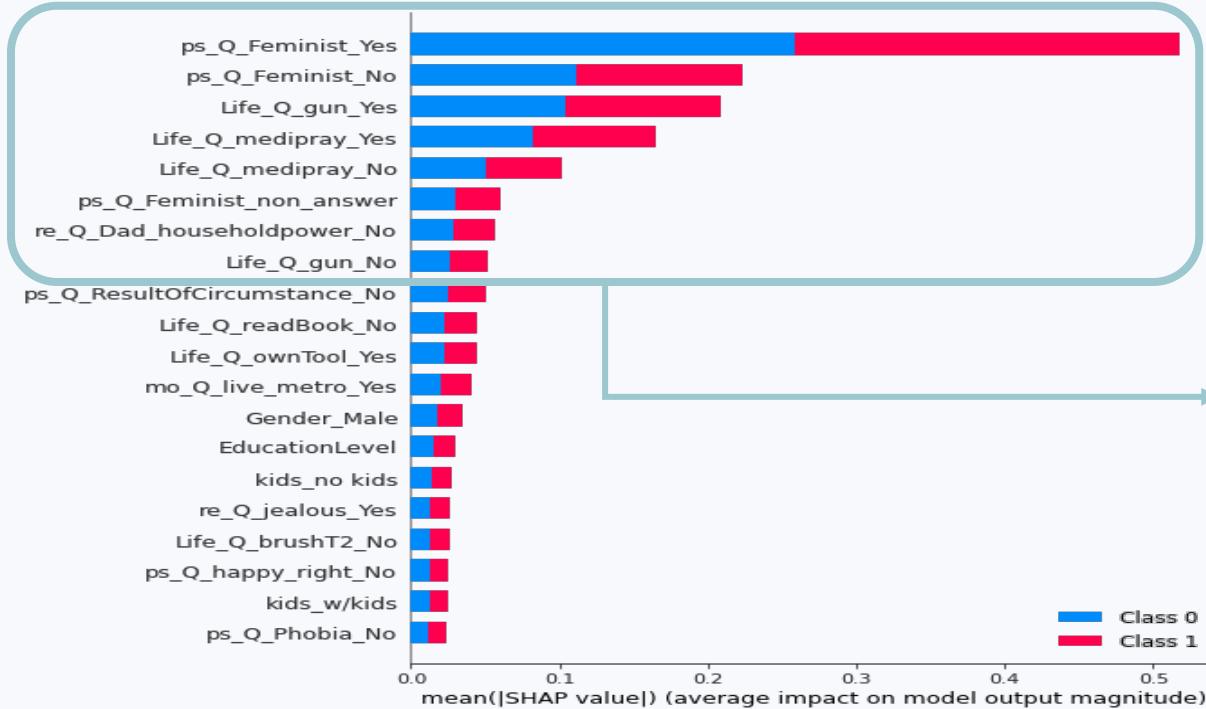
# 모델링 결과 해석



# SHAP value

## Shapley value

### shap.summary plot 해석



예측 시, 영향력이 높은 질문

ps\_Q\_Feminist

→ 당신은 페미니스트입니까?

Life\_Q\_gun

→ 총기를 소지하고 있습니까?

Life\_Q\_medipray

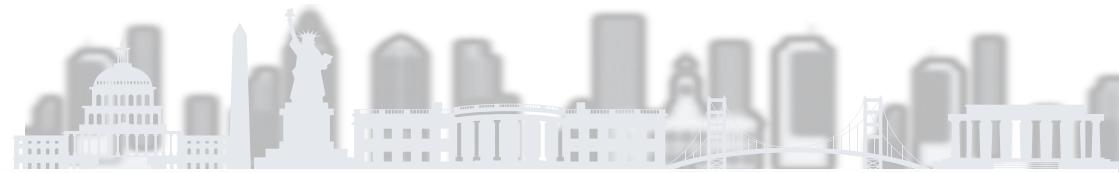
→ 주기적으로 명상이나 기도를 하십니까?

위 3개의 질문은

두 정당 모두에 대해 영향력이 높다!

즉, 정당의 예측에 있어 중요한 질문이다!

# 모델링 결과 해석



최종모델로 범주팀의 정치 성향을 예측하고 해석해보자!

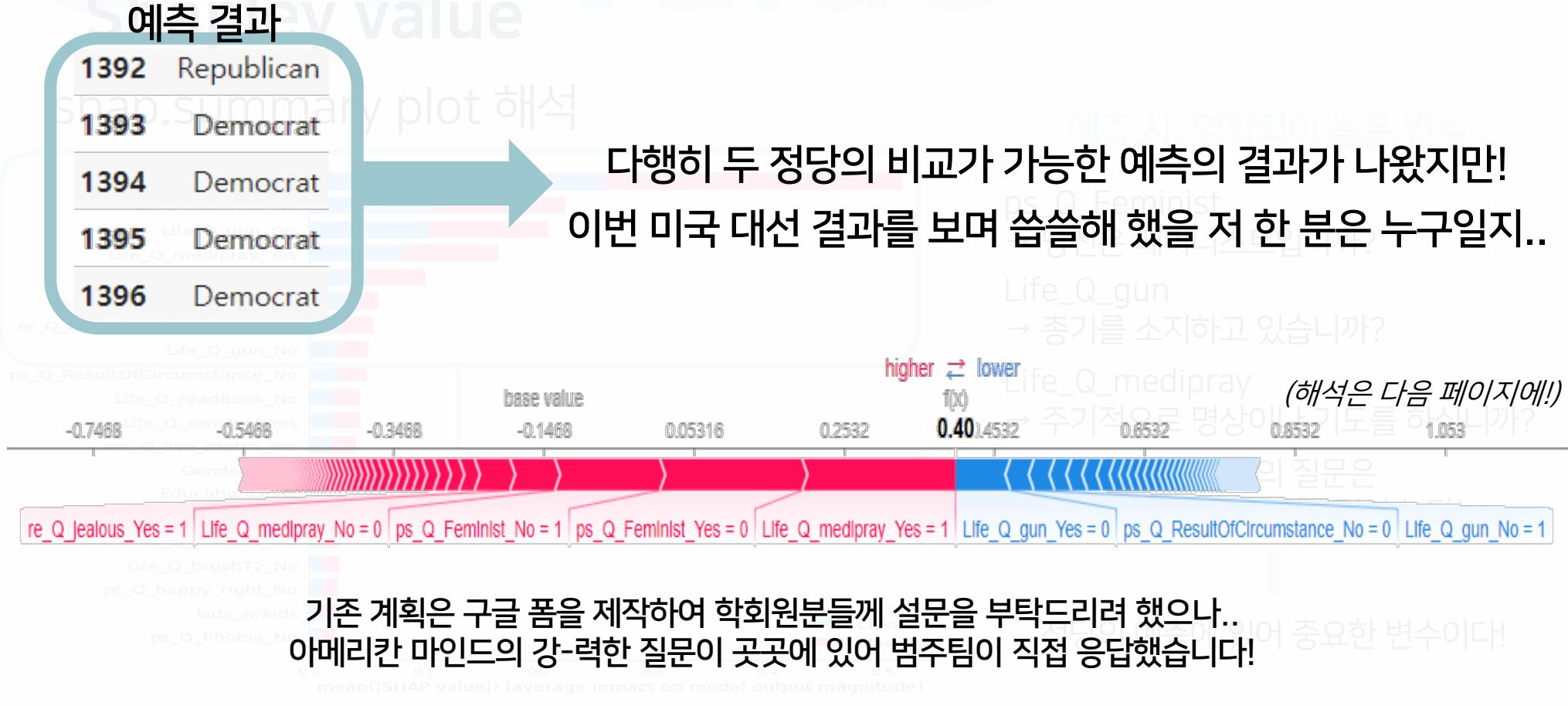
01.  
1주차 피드백

02.  
DATA 정리

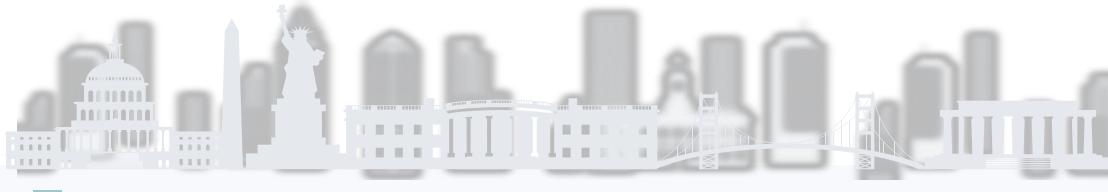
03.  
모델링

04.  
결과 해석

05.  
한계와 의의



# 모델링 결과 해석



# SHAP value

## Shapley value

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

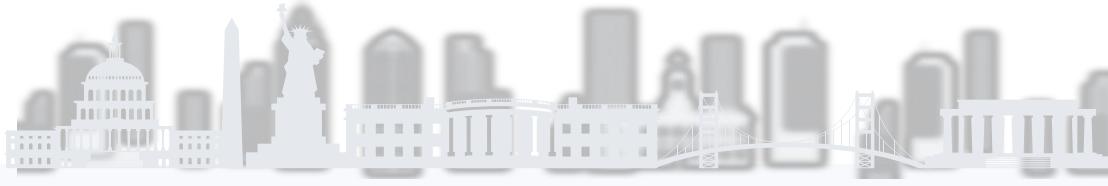
05.  
한계와 의의

What is shap.force plot?

- 각 설문 참여자의 응답결과를 shap value를 활용하여 시각화한 plot
- 각 설문 참여자들의 예측 결과에 대한 해석이 가능해진다!



# 모델링 결과 해석



# SHAP value

## Shapley value

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의



(기준 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

<미국 공화당에 대한 일반적 인식>

Republicans

- 백인 남성 / 노인
- 자유경제
- 높은 소득 수준
- 총기 규제 완화
- 기업과 개인의 자유
- 현실적

Republican으로 예측함에 있어  
양의 영향력을 가지는 질문과 대답

Q. ps\_Q\_Feminist → A: No

Republican으로 예측함에 있어  
음의 영향력을 가지는 질문과 대답

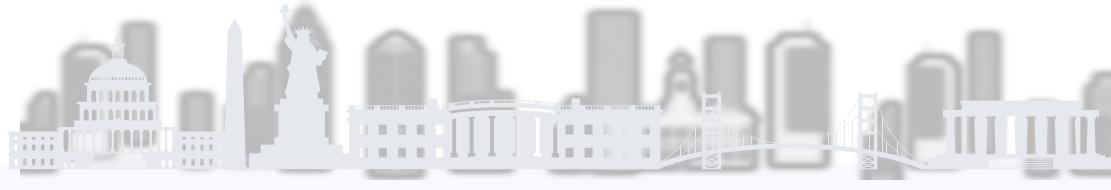
Q. Life\_Q\_gun → A: No

Q. Life\_Q\_gun → A: No

Q. ps\_Q\_ResultOfCircumstance → A: No

공화당에 대한 일반적 인식과 관련된 결과를 확인할 수 있다!

# 모델링 결과 해석



# SHAP value

## Shapley value

01.  
1주차 피드백

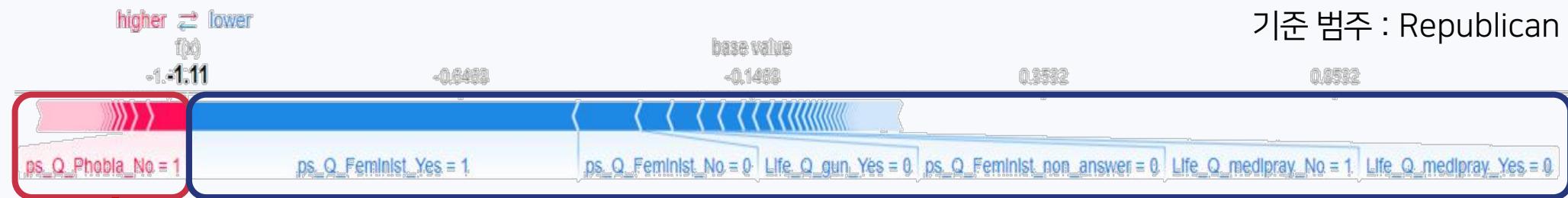
02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

shap.force plot 해석 : 민주당 지지자



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

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

# 모델링 결과 해석



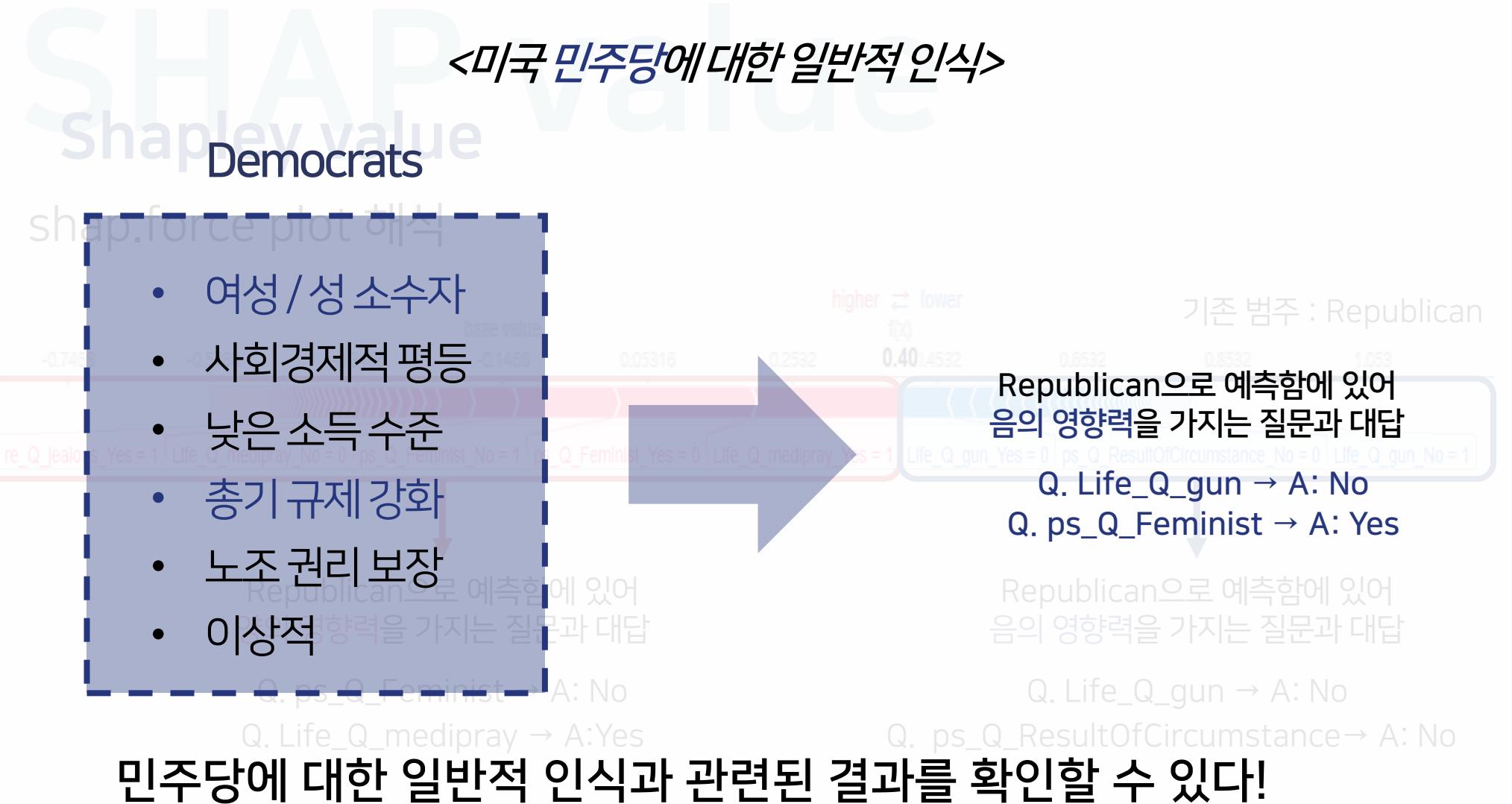
01.  
1주차 피드백

02.  
DATA 정리

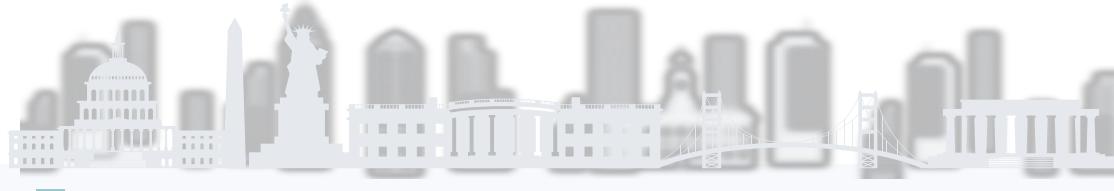
03.  
모델링

04.  
결과 해석

05.  
한계와 의의



# 모델링 결과 해석



# SHAP value

## Shapley value

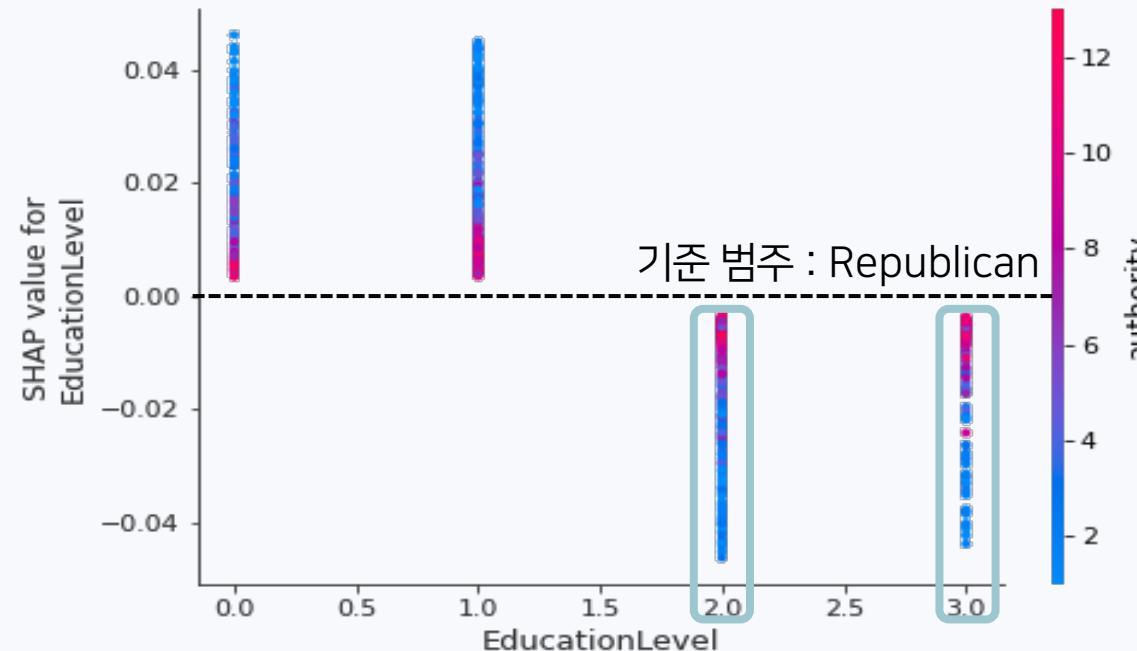
01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의



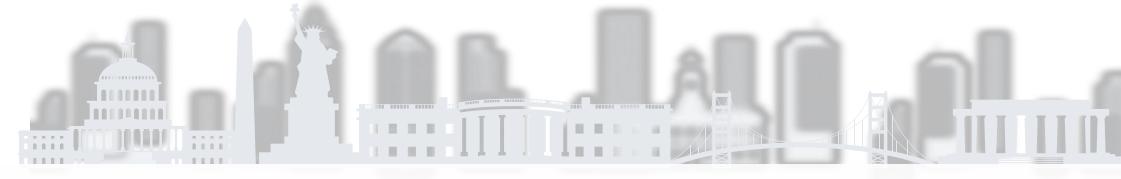
shap.dependence plot

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!)

# 모델링 결과 해석



01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

## 해석 시 유의할 점

What is shap.dependence plot? & 해석



shap value가 의미하는 영향력의 방향과 정도에  
인과관계를 투입하여 해석하는 것은 유의!

shap.dependence plot  
shap value를 기반으로 인과관계를 통해 나타낸 Plot

교육수준에 따른  
권위에 대한 복종심은

높은 교육수준은 Rep  
음의 영향력을

(즉, 민주당의 특성)



(유.의.해.)

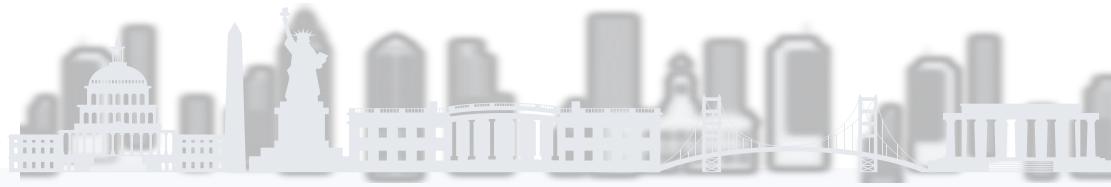


05

# 한계와 의의



# 우리의 한계와 의의는?



- **한계 - 1)**

**Show of Hands**  
asks...

Last year  
Lifestyle

37 of Hands

Do you own a gun?

**DISCUSS - 155**      **RESULTS**

Are you generally more of an optimist or a pessimist?

**DISCUSS - 36**      **RESULT**

Are you adventurous?

**DISCUSS - 26**      **RESULTS**

**Show of Hands**  
asks...

6 years ago  
Lifestyle

128

Does the "power of positive thinking" actually work?

**DISCUSS - 46**      **RESULTS**

**Show of Hands**  
asks...

4 years ago  
Lifestyle

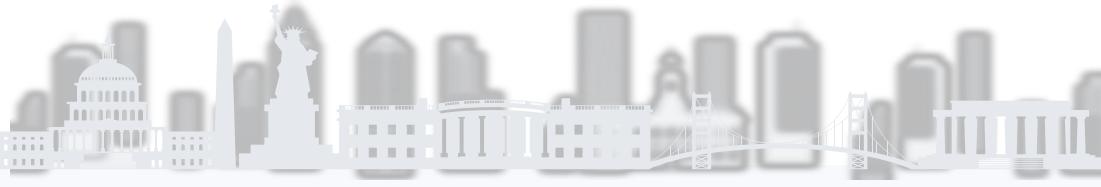
58

Are you more of a night person or a morning person?

**DISCUSS - 46**      **RESULTS**

단일 문항을 수집한 설문지 데이터 구성 방식으로 인해 많은 NA 존재

# 우리의 한계와 의의는?



- 한계 - 2)



설문지 데이터 특성 상 파생 변수 생성 과정에서  
주관이 많이 반영됨

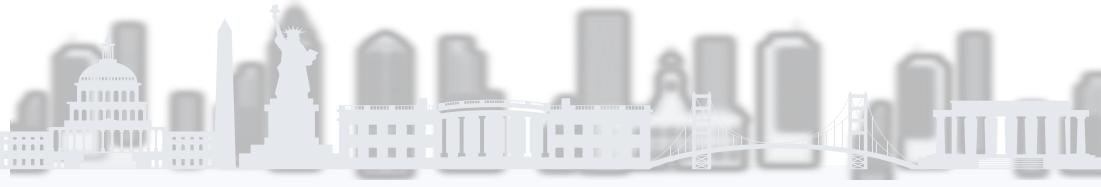
조나단씨로부터 시작해 다른 파생 변수 생성 과정까지...

- 한계 - 3)



MICE의 전체적인 flow 중  
imputation만 진행

# 우리의 한계와 의의는?



- 의의

01.  
1주차 피드백

범주가 범-주 했다  
(100개 이상의 범주형 변수 ^\_^)

02.  
DATA 정리

팀원 모두 한 개 이상의 모델을 이용한 모델링 진행!

03.  
모델링

R과 Python 모두 마스터 🔒

04.  
결과 해석

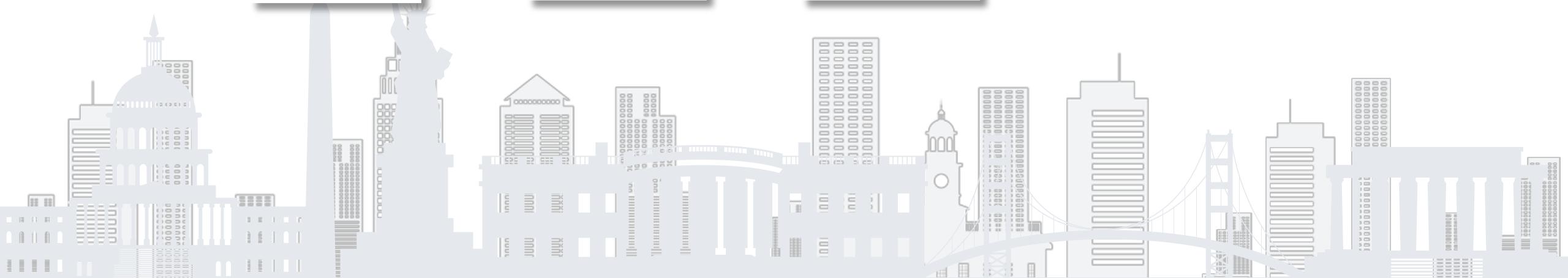
그 외 오조 오억 개

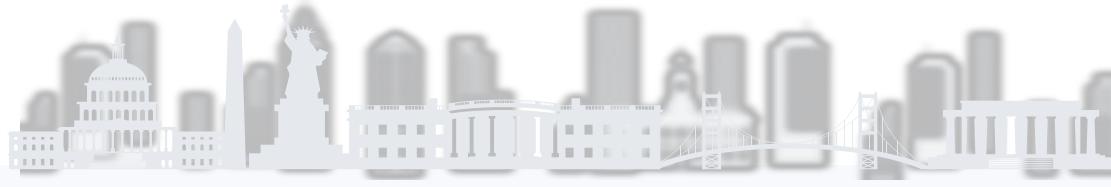
05.  
한계와 의의

워라밸 최고 행복 범-주 ❤️



스토리텔링





## 그렇다면 우리의 예측률은 몇 등이나 될까?

01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의

22	▼ 3	Izwhite		0.70689	5	4y
23	▼ 1	bouillabaisse		0.70114	14	4y
24	▲ 663	Fan Luo		0.66235	4	4y
25	▲ 411	Gabelsman		0.66235	8	4y
26	▲ 1603	AndresPerez		0.66235	11	4y
27	▲ 106	Rohan Sachdev		0.65948	24	4y
28	▲ 294	SergioAntonio		0.65804	17	4y
29	▲ 10	DagnyT		0.65804	8	4y
30	▲ 408	Caio Taniguchi		0.65804	19	4y

"0.66091"

행복범주는  
2874팀 중 27등을  
기록했다!



그런데 말입니다...

우리는 리더보드에서 이상한 점을 발견했다...



(매우 양쪽)



## 리더보드의 수상한 점

1	▲ 7	HugoSilveiradaCunha		0.92241	20	4y
2	—	@mos		0.92097	31	4y
3	▲ 1	Elie		0.91522	58	4y
4	▲ 1	Keks		0.91379	74	4y
5	▲ 1	캬 (安全保障)		0.90948	5	4y
6	▲ 5	Komaxx		0.90517	44	4y
7	▼ 6	Ohm J. Patel		0.90229	17	4y
8	▲ 1	Artur B		0.90229	47	4y
9	▼ 2	Stefan Hoglund		0.90086	18	4y
10	—	ShedrickBridgeforth		0.89793	49	4y
11	▲ 1	Wynnie		0.78160	27	4y
12	▲ 1	chrislit		0.77442	64	4y

극 상위권의 예측률이  
말도 안된다는 생각이 들어버린  
범주는 조사에 들어갔다

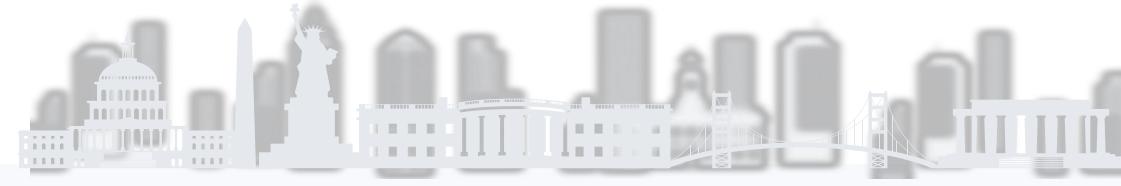
01.  
1주차 피드백

02.  
DATA 정리

03.  
모델링

04.  
결과 해석

05.  
한계와 의의



Discussion에 들어가보니 말도 안되는 결과에 잔뜩 화가 난 참가자들이  
상위권 사람들의 리더보드를 분석한 plot을 그린 것을 발견

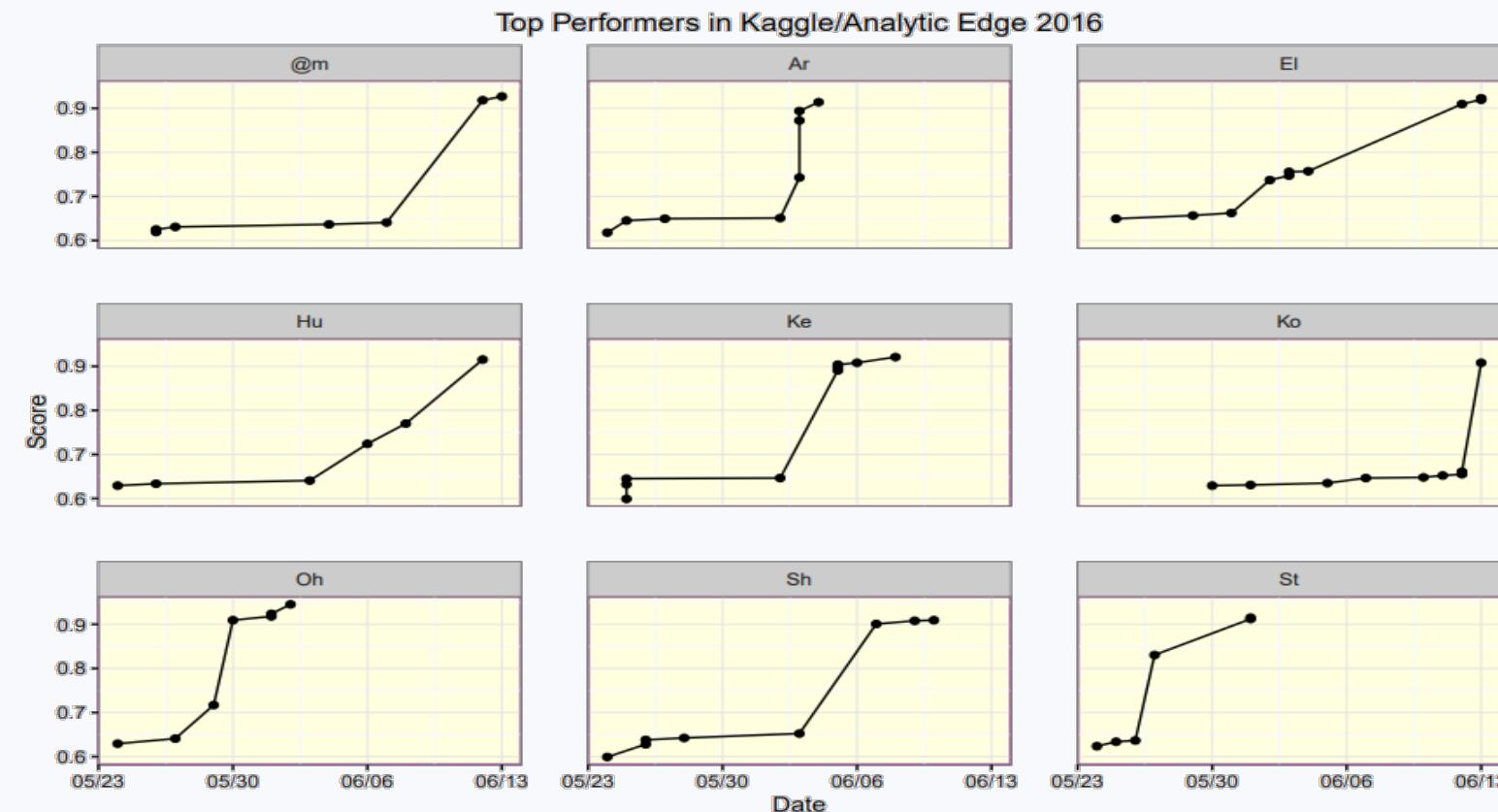
01.  
1주차 피드백

02.  
DATA 정리

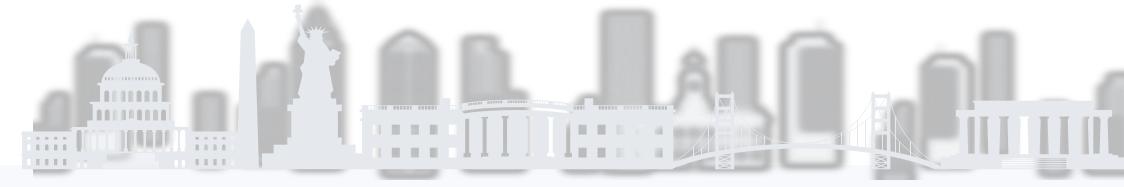
03.  
모델링

04.  
결과 해석

05.  
한계와 의의



다들 한 번에 말도 안되게 결과가 좋아졌다... 이게 말이 돼..?



크게 실망했다



알고보니 같은 데이터로 2년 전 대회가 열렸었고 이들은 이 때의 데이터를 cheating한 것...

## 01. 1주차 피드백

## 02. DATA 정리

## 03. 모델링

## 04. 결과 해석

## 05. 한계와 의의

1	▲ 1	Wynnie		0.78160	27	4y
2	▲ 1	chrislit		0.77442	64	4y
3	▲ 1	RandXie		0.77011	14	4y
14	▲ 2	A_Z_L		0.76580	49	4y
15	—	PJBu		0.76436	19	4y
16	▲ 2	William Chiu		0.74281	50	4y
17	▲ 3	SergeyP		0.73994	1	4y
18	▲ 5	KishoreKumar		0.73275	18	4y
19	▲ 2	RahulMadhavan		0.72844	9	4y
20	▼ 3	Ayush Singh		0.72701	44	4y
21	▲ 3	Cristinaatx		0.71839	32	4y
22	▼ 3	Izwhite		0.70689	5	4y
23	▼ 1	bouillabaisse		0.70114	14	4y
24	▲ 663	Fan Luo		0.66235	4	4y
25	▲ 411	Gabelsman		0.66235	8	4y
16	▲ 1603	AndresPerez		0.66235	11	4y
17	▲ 106	Rohan Sachdev		0.65948	24	4y

“0.66091”  
←  
우리 범주 17등이다!

(그렇다.. 범주는 상위 0.5%의 엘리트 집단..)  
설문조사를 통한 지지정당 예측 | 103



너 안녕에 문제 있어?

# 감사합니다

