Speed dating

문기태 최윤슬 김나형 안수빈

1 데이터소개 & 분석목적

2 데이터탐색 & 정제

3 모형구축 & 설명

4 분석결과 & 결론

CONTENTS

[데이터 소개]

Speed Dating?

독신 여자,남자가 **애인을 찾을 수 있도록**

짧은 시간동안 여러 사람들을 돌아가며

만나보는 일종의 소개팅

[분석 목적]

매칭 성공에 영향을 주는 요인들을 찾기

종속변수 match를 맞춰라 !!

[데이터 탐색 및 정제]

데이터 구성

196 variables

6702 obs

iid	id	gender	idg	condtn	wave	round	position	positin1	order	partner	pid	match	int_corr	samerace
70	5	1	10	1	3	10	6	NA	5	5	60	0	0.45	0
351	11	0	21	2	14	18	11	11	7	10	370	0	0.15	0
345	5	0	9	2	14	18	17	17	13	8	368	1	0.32	0
351	11	0	21	2	14	18	11	11	3	17	377	0	-0.40	1
488	8	1	16	2	19	15	5	14	8	4	469	0	0.31	1
359	19	0	37	2	14	18	15	15	1	14	374	0	-0.07	0
8	8	0	15	1	1	10	10	NA	2	10	20	1	-0.21	1
145	4	0	7	2	7	16	14	14	12	3	160	0	-0.18	0
370	10	1	20	2	14	20	12	12	8	12	352	0	0.16	0
286	14	1	28	2	11	21	16	16	3	20	271	0	-0.22	1
382	4	0	7	2	15	19	17	17	9	3	399	0	0.24	1
301	8	0	15	2	12	14	3	3	4	10	317	0	0.67	1
171	14	1	28	2	7	16	3	1	3	6	147	0	0.54	0
523	15	0	29	2	21	22	5	5	15	7	537	0	0.14	1
177	4	0	7	1	8	10	4	4	9	7	190	0	-0.03	1
474	9	0	17	2	19	15	8	9	1	7	487	0	-0.21	0
173	16	1	32	2	7	16	10	16	11	2	143	0	-0.03	1
162	5	1	10	2	7	16	7	14	10	15	156	0	0.45	0
106	13	1	26	2	4	18	3	NA	1	14	89	0	0.38	0

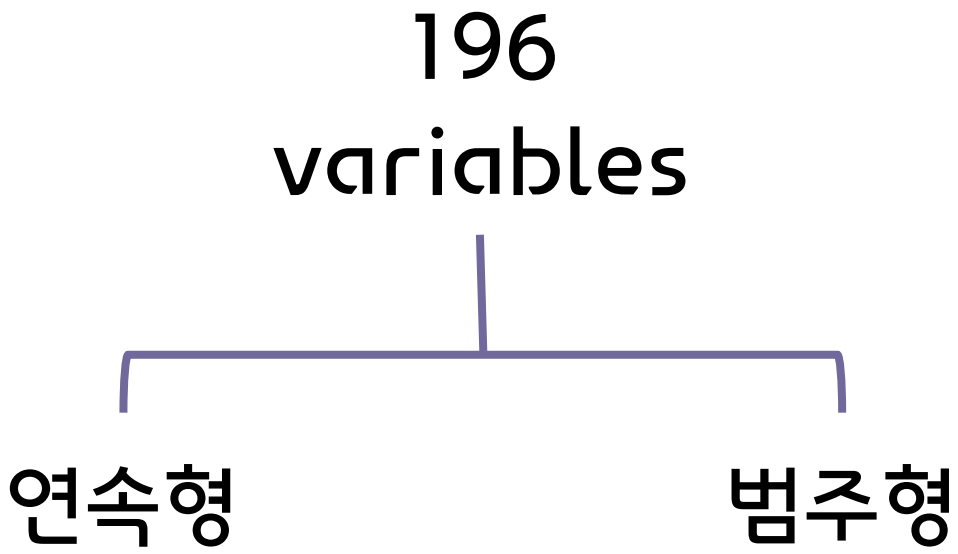
주요 변수 소개

변수 이름	변수 설명	범주 개수 / 범위
round	그 날 스피드 데이트에서 만난 파트너 수	5~22
wave	스피드 데이트 조 번호	1~22
pf_o_attr	6개 항목에 파트너가 사전에 매긴 점수	1~10
from	태어나서 자란 나라	270개
career_c	직업 코드	17개
exphappy	소개팅에 기대하는 정도	1~10점
like / like_o	이 사람이 마음에 든 정도 (_o는 상대방이 나에게 매긴 점수)	1~10점

주요 변수 소개

변수 이름	변수 설명	범주 개수 / 범위
match_es	예상 매칭 성공 수	5~22
satis_2	만난 사람에 대한 만족 정도	1~10
length	미팅의 전반적인 만족 정도	3
numdat_2	그 날 스피드 데이트 수의 적당한 정도	3
int_corr	사전조사가 파트너와 비슷한 정도	-1~1
samerace	인종이 같은지 여부	0/1
prob / prob_o	상대방이 나를 선택할 확률의 정도 (_o는 상대방이 나에게 매긴 점수)	1~10점

변수 선택



변수 선택

연속형

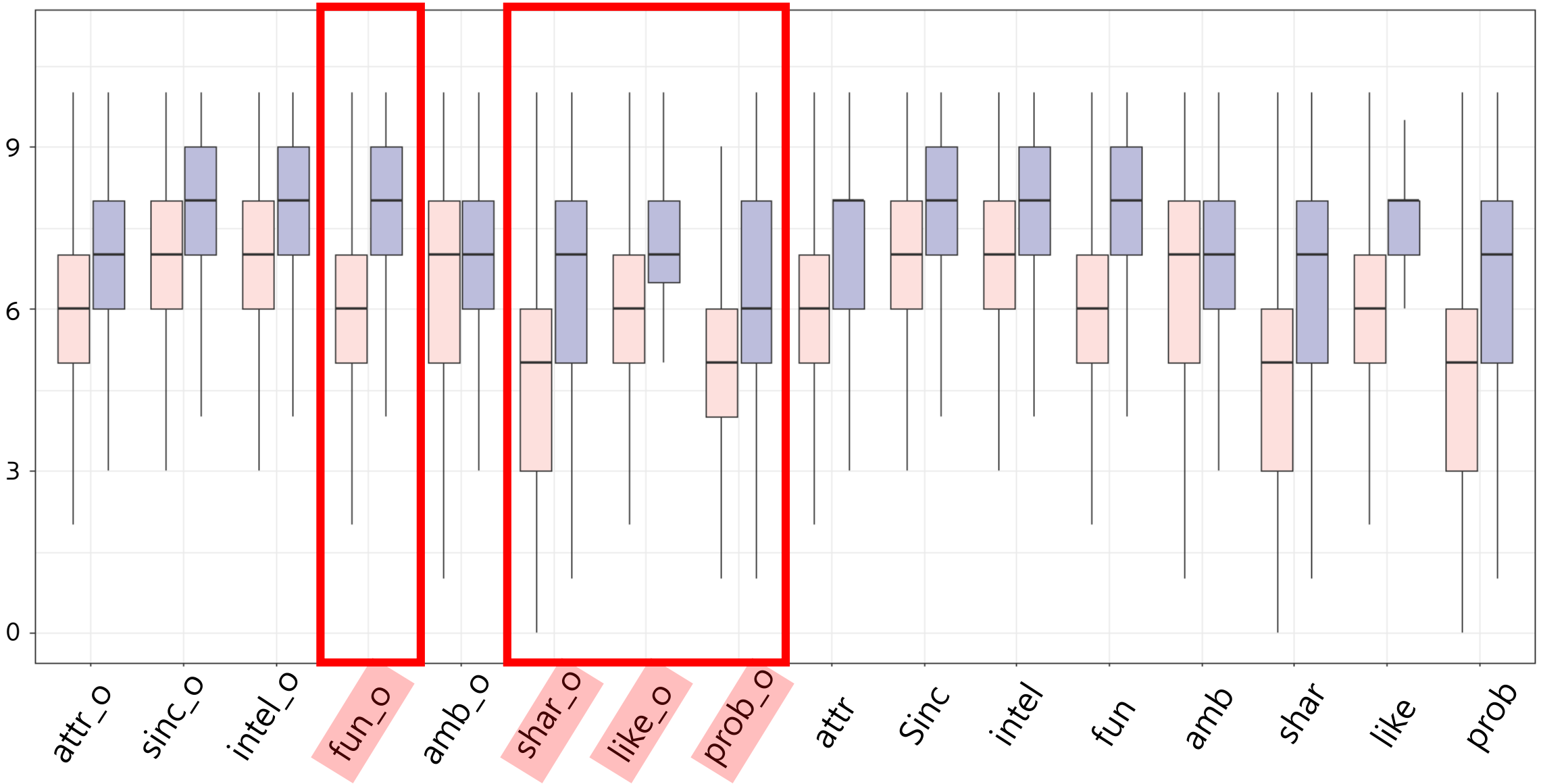
종속변수인 match와
Boxplot을 그려보아
유의미해 보이는
변수들을 선택

match = 0　　match = 1

각 변수마다 **match**별로
차이 비교

Boxplot 그리기

Boxplot 그리기


9
6
3
0
attr_o
sinc_o
intel_o
fun_o
amb_o
shar_o
like_o
prob_o
attr
Sinc
intel
fun
amb
shar
like
prob

전처리-연속형

중요도를 점수로 부여하는 변수들

ex) attr, attr1_1...

- ✔ NA → 0으로 대체
  : 점수를 매기지 않은 것은 중요하지 않다는 것!
- ✔ 1부터 10까지인데 10보다 큰 경우 → 10으로 대체
- ✔ 합산하여 100보다 큰 경우 → 합산한 점수로 나눠 합이 100이 되도록 다시 입력

전처리-연속형

- *pf_o_fun, pf_o_amb, pf_o_sha* 가 *NA*

  ⇒나머지 항목들을 더해 100점이 되므로 0 할당

- *attr_o ~ shar_o* 가 *NA*

  ⇒ 위와 같은 이유로 0 할당

  ⇒ *attr_o*가 10.5인 값, 10으로 할당

- *fun1_1, amb1_1, shar1_1* 가 *NA*

  ⇒ 위와 같은 이유로 0 할당

전처리-연속형

- *like_o, prob_o 가 NA*

  ⇒ *attr_o + sinc_o + … + shar_o = total_o*

  total_o 범위에 따라 *like_o*를 채움 (10단위로 끊어서)

- *attr ~ shar은 평가하지 않았는데 like값은 결측이 아닌 사람 - 4명*

  ⇒ *attr_shar* 각각의 평균을 넣어줌

- *사전 조사 안 한 사람들 (int_corr이 NA 값인 사람들)*

  ⇒ 사전 조사가 결과에 중요하다고 생각하여 삭제

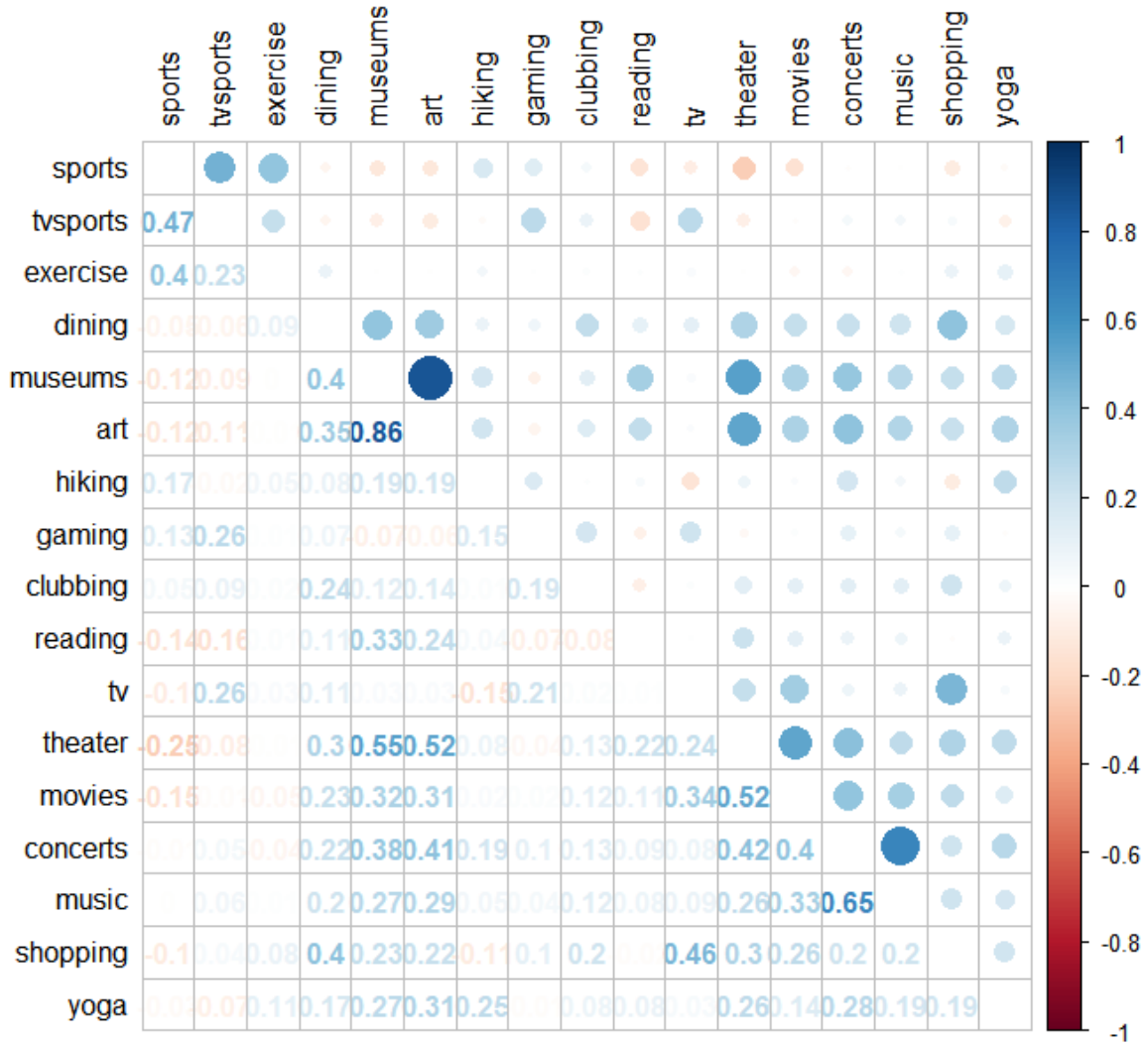
변수 선택

범주형

1 서로 상관계수가 높게 나오거나 의미없는 범주가 많은 **변수통합**

2 종속변수인 match와 도수분포표를 그리고 **카이제곱 테스트**를 하여 유의미한 변수만 선택

전처리-범주형



1. 취미 변수

: 상관계수 높은 변수끼리 묶기

museum art music reading theater movies concert

= ***cult*** : 문화생활

dining shopping clubbing

= ***outside*** : 외출활동

•
•
•

전처리-범주형

2. career_c 가 NA 인 사람들

⇒ career를 보고 범주를 확인해 넣어줌

career	career_c
Lawyer & Law	1
Economist	2
Tech Professional	5
NA	15

3. dec / dec_o 변수 제거

⇒ 제거하지 않으면
match가 그대로 예측이 됨

⇒ dec X dec_o = match

4. 카이제곱 검정을 통한 변수 제거

예) numdat_2

match / numdat_2	match X	match O
1	455	95
2	2728	509
3	2259	484

p-value = 0.1284 > 0.05

numdat_2 제거

전처리-범주형

- *match_es* 가 *NA*

  ⇒ 만난 사람의 수 (*round*)와 밀접한 관련이 있을 것이라 생각,

  *table*을 그린 후 **최빈값**을 넣는다.

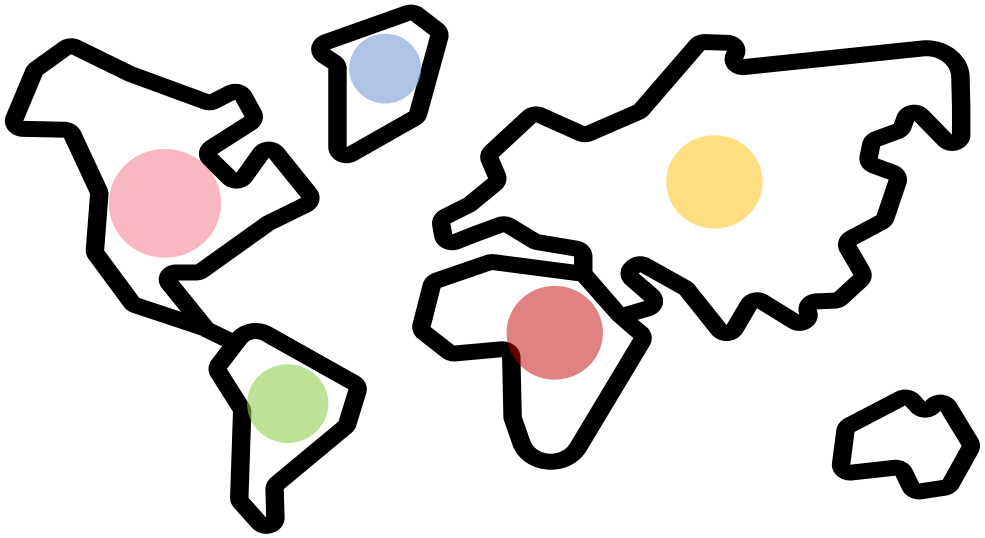
  + (1.5 , 2.5인 사람 -> 각각 2,3 / 3.4인 사람 -> 3)

- *match_ess = match_es / round*

  ⇒ 예상 매칭 성공 수를 뜻하는 *match_es*는 *wave*별로 인원이 다르기 때문에 상이.

  인원을 뜻하는 *round*로 나눠서 척도를 맞춰준다.

- from 출신 국가 및 지역 변수



270개의 범주
→ 대륙 기준 6개의 범주
(미국은 동부/서부)

전처리-범주형

→ 1) 94115 -> *San Francisco*

2) 미국 -> 동부, 서부로 나눔 *(East/West)*

3) 유럽 *(Europe)*

4) 북아메리카 *(North America)*

5) 아시아 *(Asia)*

6) 남아메리카 *(South America)*

전처리 결과

iid condtn round
pid order pf_o_sha
⋮

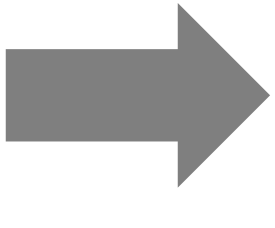
196개의 변수

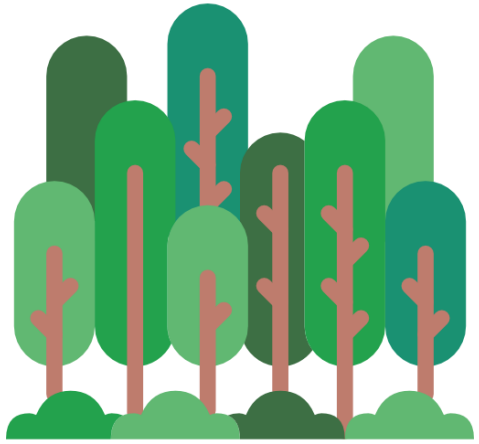
~~iid~~ condtn round
~~pid~~ order pf_o_sha
⋮

30개의 변수

[모형 구축]

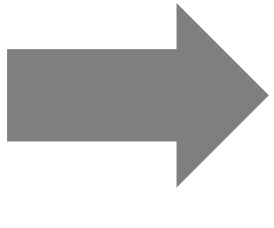
분석 1. 랜덤포레스트





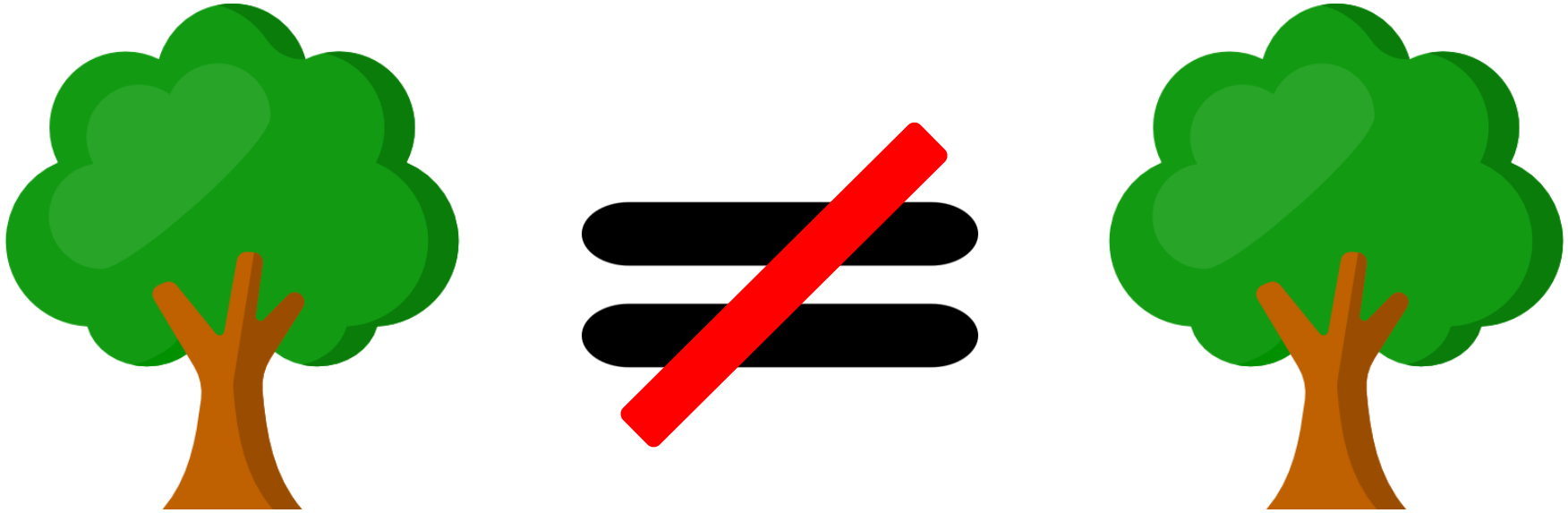
분산 감소

편향 유지



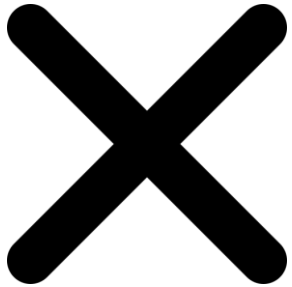
각각의 결정

분석 1. 랜덤포레스트



분석 1. 랜덤포레스트

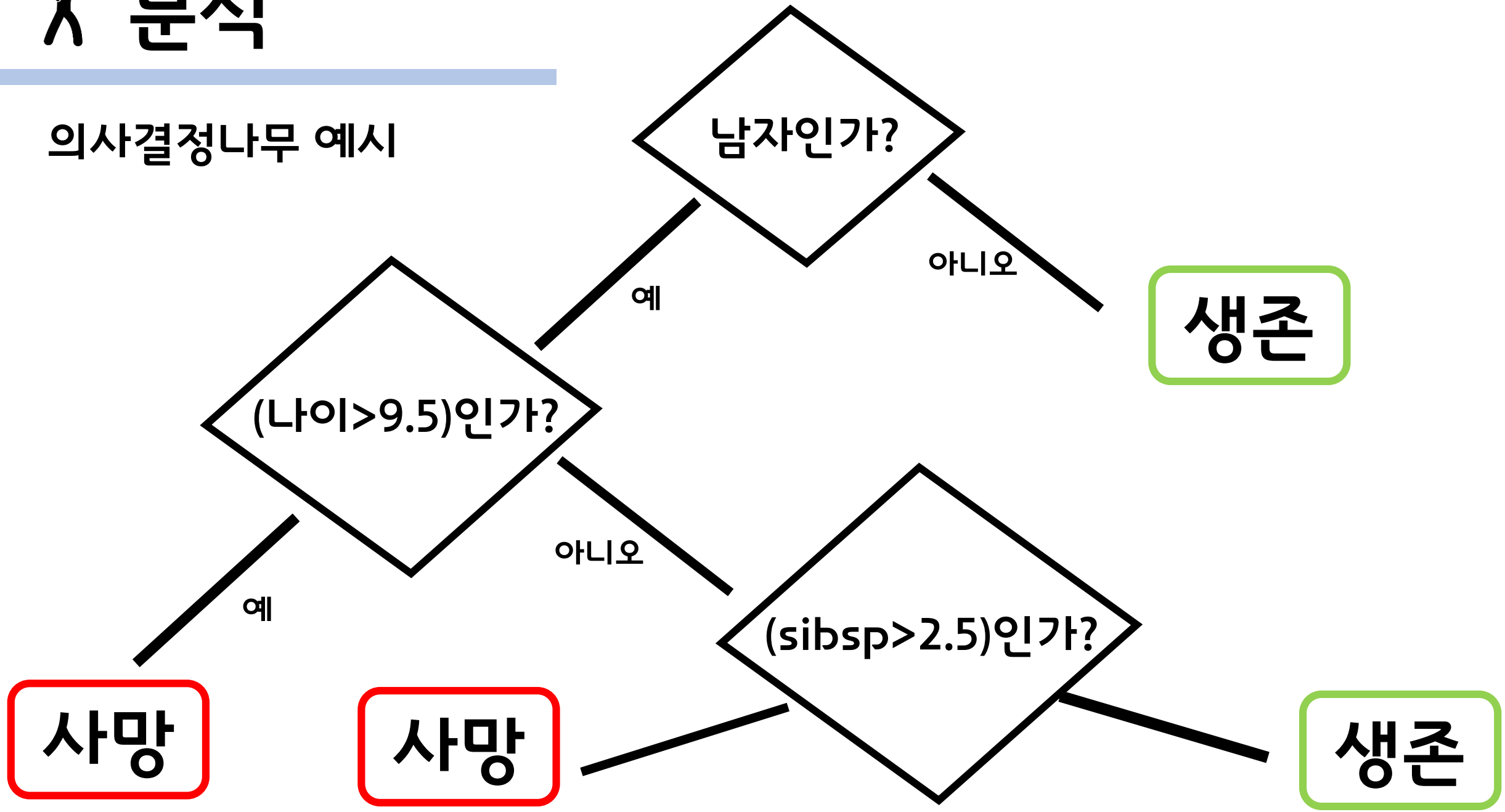
Mtry= 사용할 변수



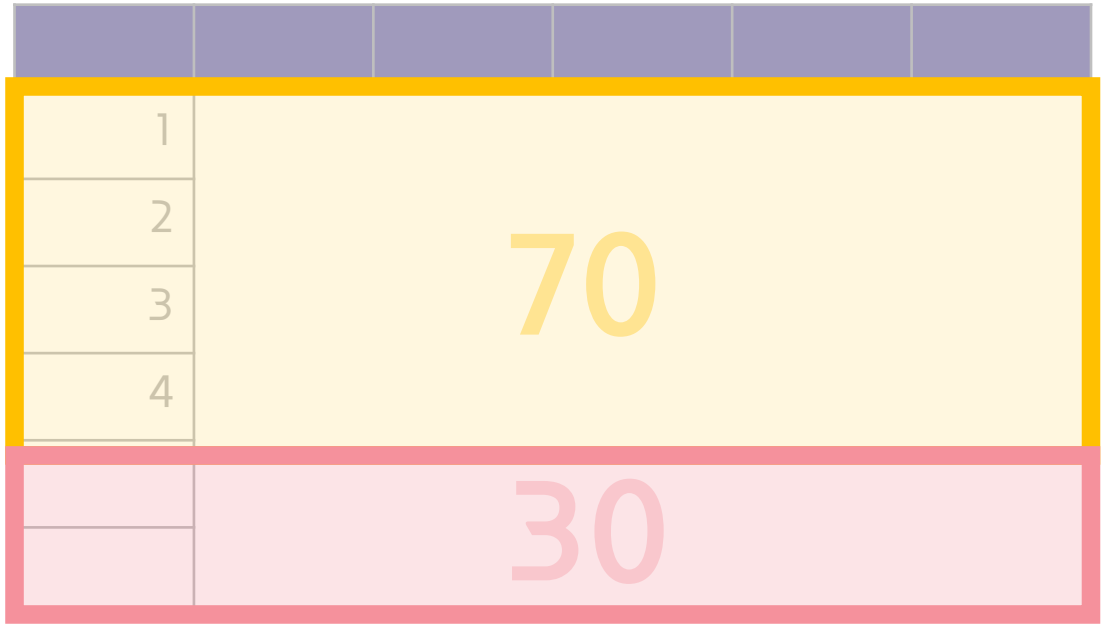
ntree

분석

의사결정나무 예시



분석 1. 랜덤포레스트



Train 데이터

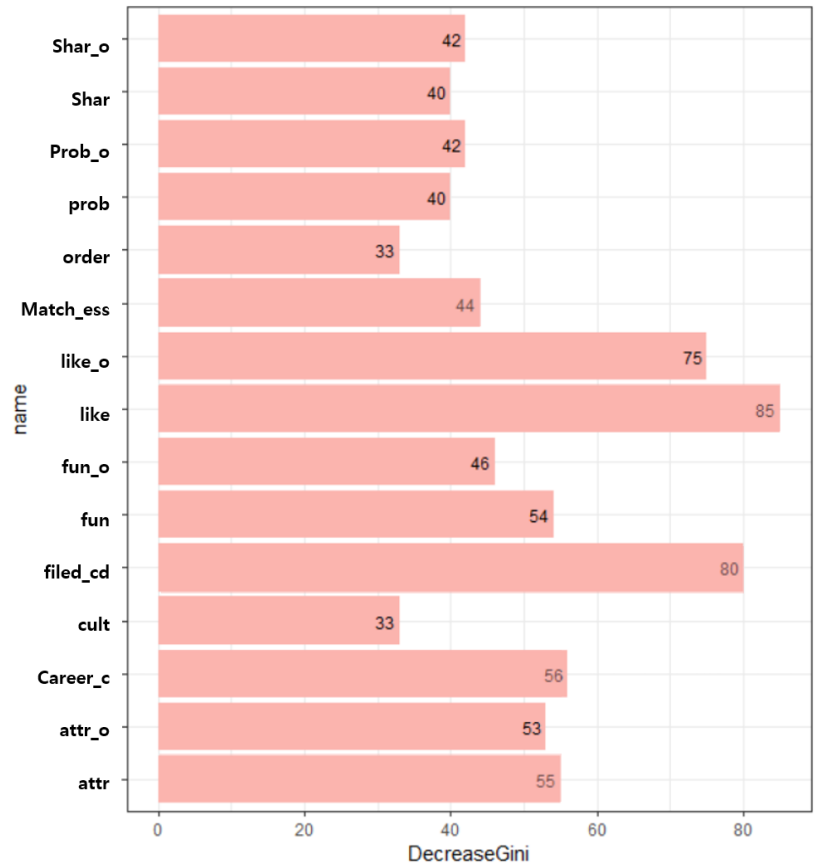
Train70 : 모형 설계

randomForest 사용

Test30 : 모형 확인

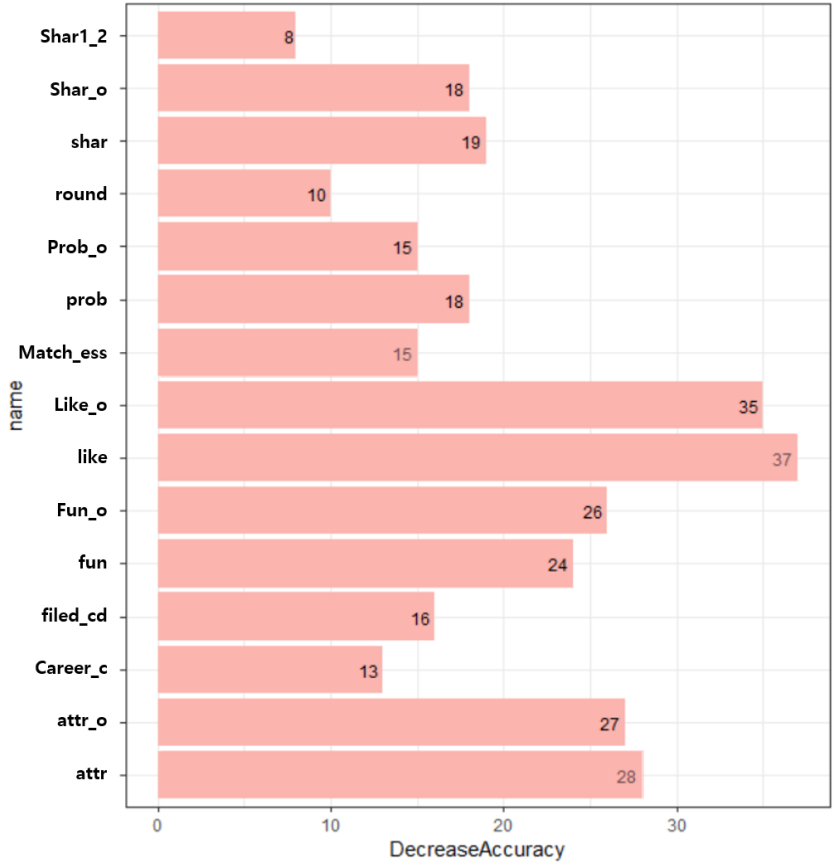
OOB 및 cofunsionMatrix 사용

분석 1. 랜덤포레스트

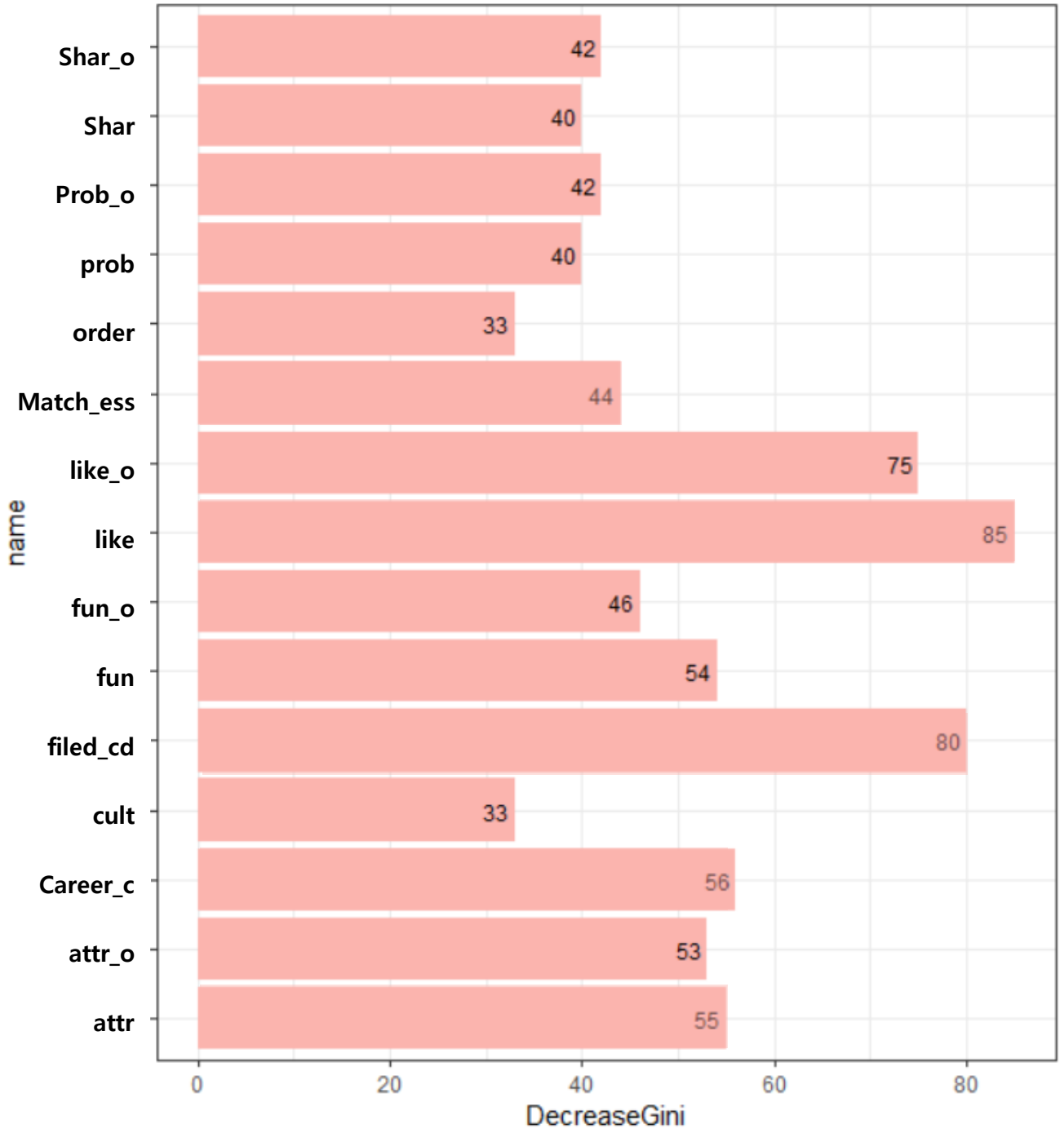


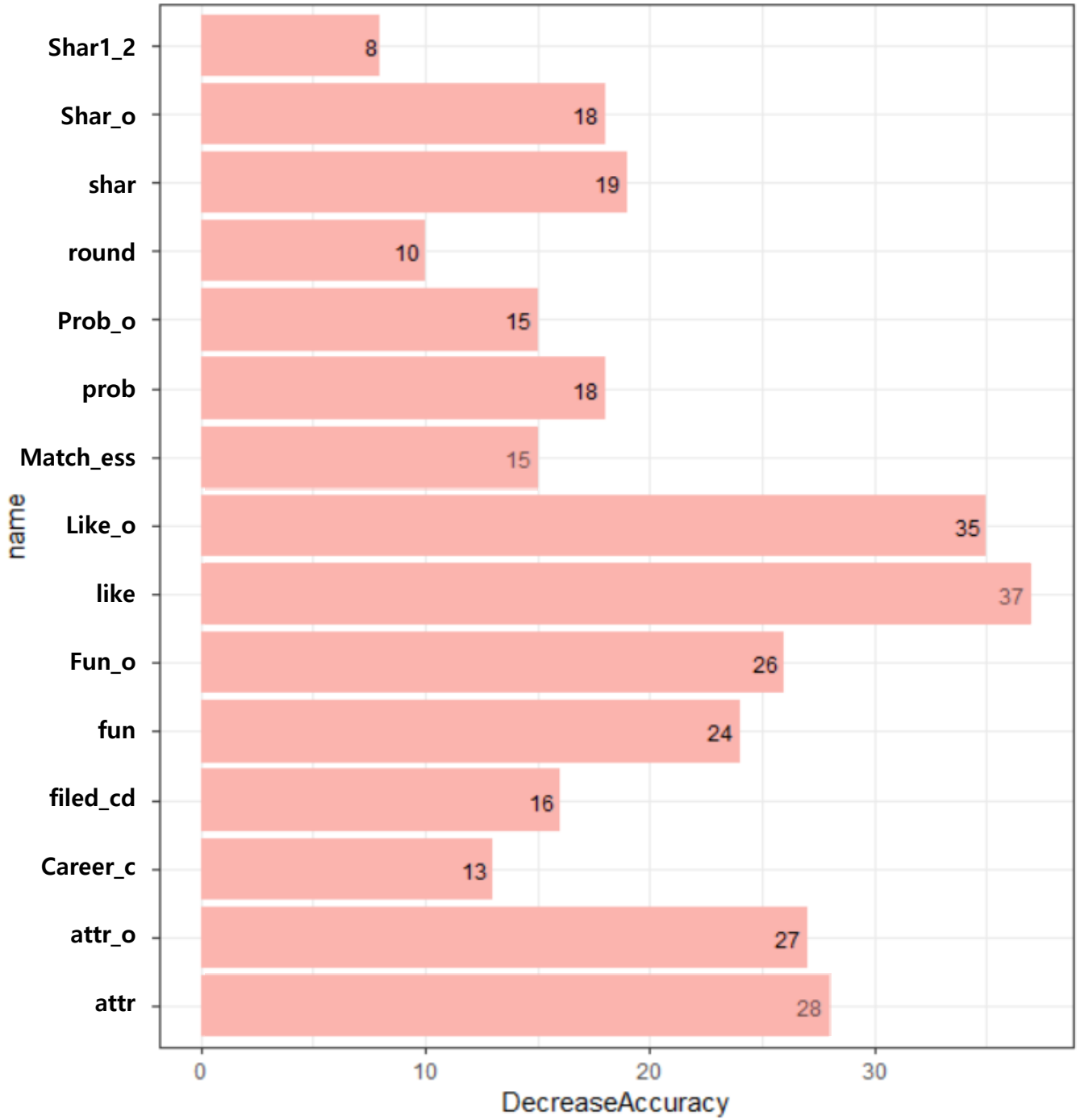
like attr match_ess

field_cd career_c



like attr match_ess


Shar_o
Shar
Prob_o
prob
order
Match_ess
like_o
like
fun_o
fun
filed_cd
cult
Career_c
attr_o
attr
42
40
42
40
33
44
75
85
46
54
80
33
56
53
55
name
0
20
40
60
80
DecreaseGini


Shar1_2
Shar_o
shar
round
Prob_o
prob
Match_ess
Like_o
like
Fun_o
fun
filed_cd
Career_c
attr_o
attr
name
8
18
19
10
15
18
15
35
37
26
24
16
13
27
28
0
10
20
30
DecreaseAccuracy

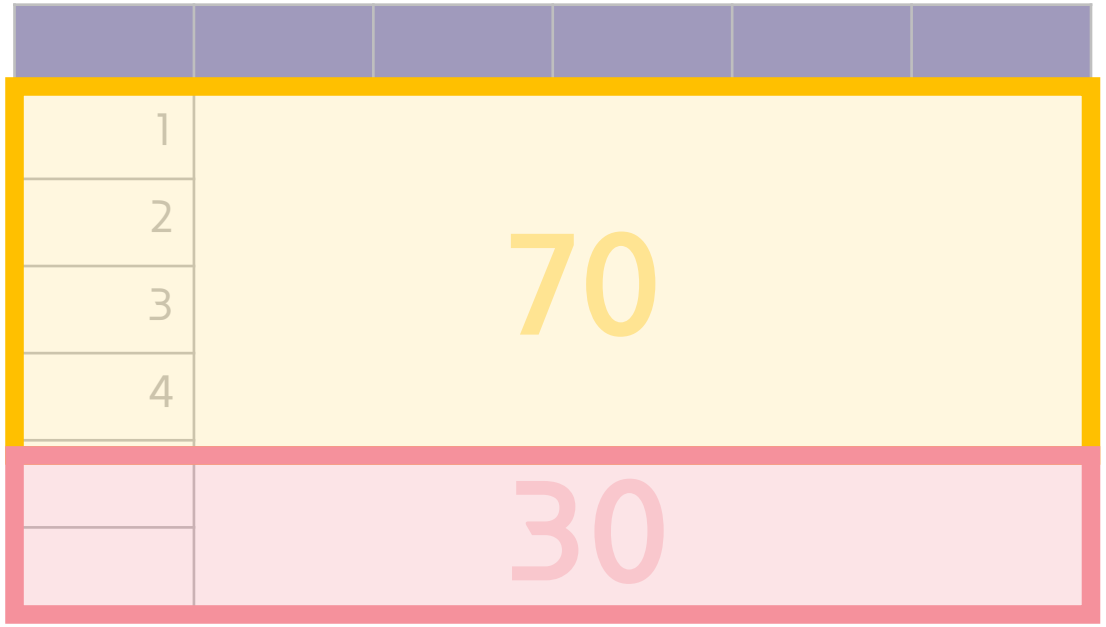
분석 1. 랜덤포레스트

prediction \ match	0	1
0	1569	240
1	50	100

Confusion Matrix
test set 적용

✔accuracy = 0.855

분석 2. 로지스틱



Train 데이터

Train70 : 모형 설계

glm을 활용한
로지스틱 회귀 모형

Test30 : 모형 확인

roc곡선, auc,
confusionMatrix

4번 반복하여 비교

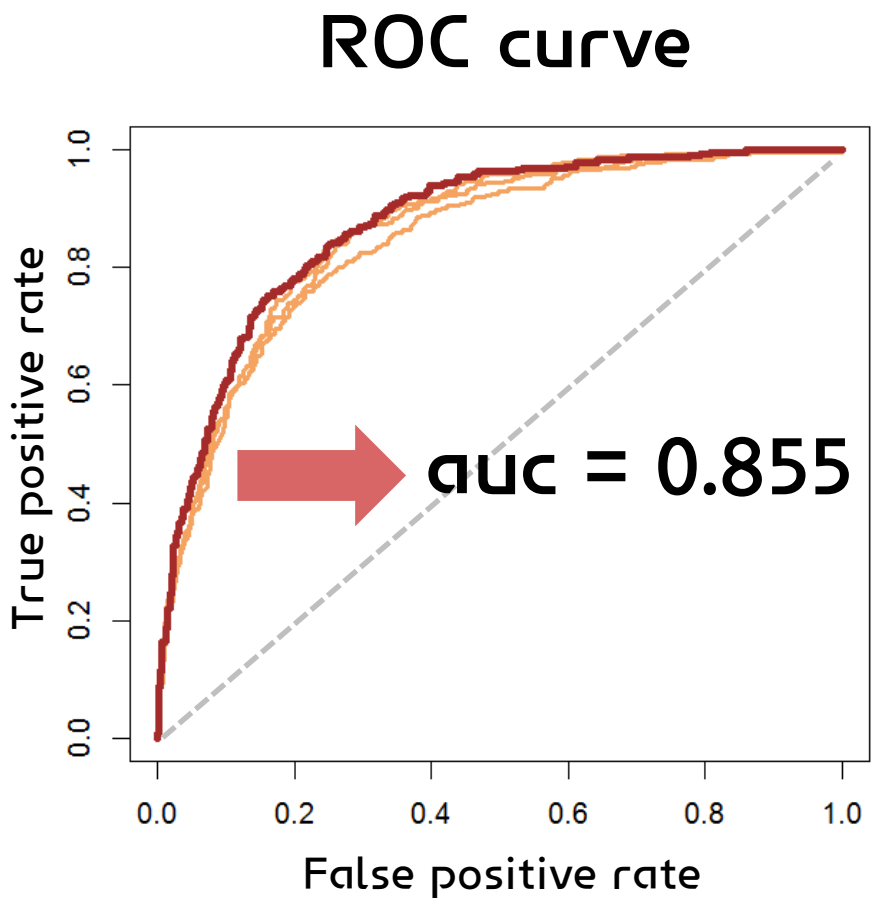
분석 2. 로지스틱

match 예측하는 로지스틱 회귀 모형 fitting

=> 유의하지 않은 변수, 다중 공선성 문제

- 변수 내 무의미한 범주 통합
- stepwise로 변수 제거

분석 2. 로지스틱



		match 0	match 1
prediction	0	1569	203
	1	63	123

accuracy = 0.864

다음을 갖는 모형을 최종적으로 선택

분석 2. 로지스틱

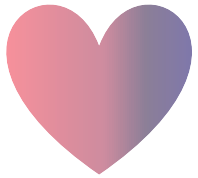
	Odds ratio	Std.error	Pr(>|z|)
attr_o	1.38	0.03	<0.05
like	1.27	0.05	<0.05
attr	1.23	0.04	<0.05
prob_o	1.20	0.03	<0.05
….	….	….	….
shar_o	1.04	0.02	0.11
attr3_2	0.81	0.04	<0.05
likedif	0.76	0.04	<0.05
condtn2	0.31	0.22	<0.05

Summary 결과

shar_o, outside변수 제외
모든 변수 유의

Odds ratio를 통해
match와의 관계 파악

[분석 결과]



높을수록 매칭이 성공하는 요인

shar / shar_o

관심사를 공유하는 정도

match_ess

예상 매칭수

like

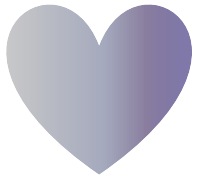
상대가 마음에 든 정도

재미 점수

fun / fun_o

매력도 점수

attr / attr_o



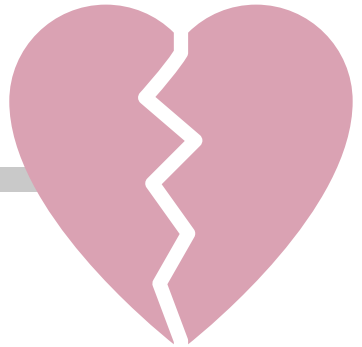
높을수록 매칭이 실패하는 요인

attr3_2

자신을 매력적으로 생각하는 정도

likedif

상대와 나의 like점수차



지성

intel / intel_o

의욕적인 정도

amb / amb_o

감사합니다