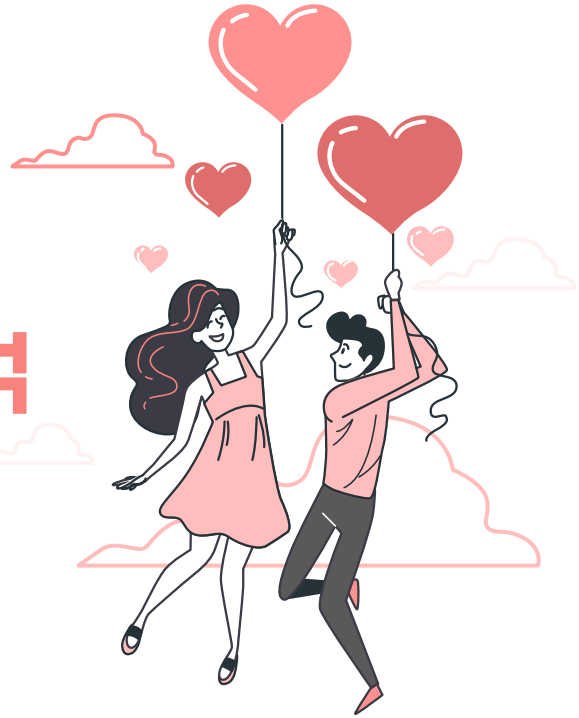
# 스피드 데이트 성공요인 분석

2019110485 통계학과 정다혜

# CONTENTS

# 01

# 데이터 소개 및 분석목적



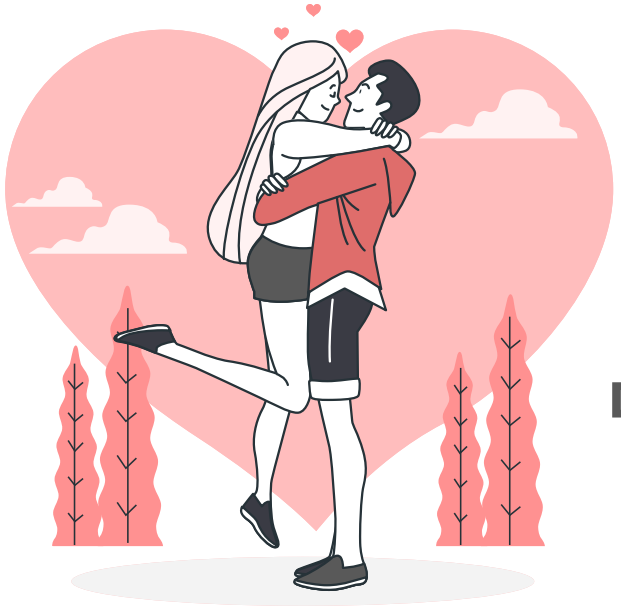
# 데이터 소개

## 스피드 데이팅이란?

자신의 짝을 찾기 위해
짧은 시간동안 여러사람을
만나는 소개팅

**dataset** 콜롬비아 경영대학 교수의 연구에서 사용된 dataset

# 분석 목적



스피드 데이트 후에 다음
만남여부를 예측하고
매칭 성공에 영향을 주는 요소를
찾기

# 변수 소개

## round

그 날 데이트에서 만난 파트너 수

## Match_es

예상 매칭 성공 수

## prob

상대방이 나를 선택할 가능성 (본인이 작성)

## impreling

같은 종교인 것이 얼만큼 중요한지 (0-10)

## exhappy

소개팅에 기대하는 정도

## Int_corr

사전조사가 파트너와 비슷한 정도 (-1~1)

## goal

참가 목적

## match

Target변수, 매칭성공여부

# 변수 소개

## Attr, sinc, intel, fun, amb, shar (0-10)

1

내가 이성을 볼 때

2

이성이 중요하다고 생각할 것 같은 것

3

내가 생각하는 나

4

지인이 이성을 볼 때 중요하다고 생각할 것 같은 것

5

지인이 보는 나

_o

파트너의 응답

**Ex) attr2는 이성이 매력을 어느정도 중요하다고 생각할지를 0부터 10까지 나타냄**

# 02

# 데이터 전처리 및 탐색

# 변수 제거

## 변수 120-195

매치된 사람만
작성한 변수

## 결측치 50% 이상 변수

expnum, attr1_s 등
12개 변수 제거

## 범주형 변수

from, undergrad 등
범주와 결측치가
많은 변수 제거

## 호감도에 영향이 없다고 판단한 변수

position - 만난위치
positin1 – 처음위치
앞서 6가지 특성 변수
중 2,4,5 번

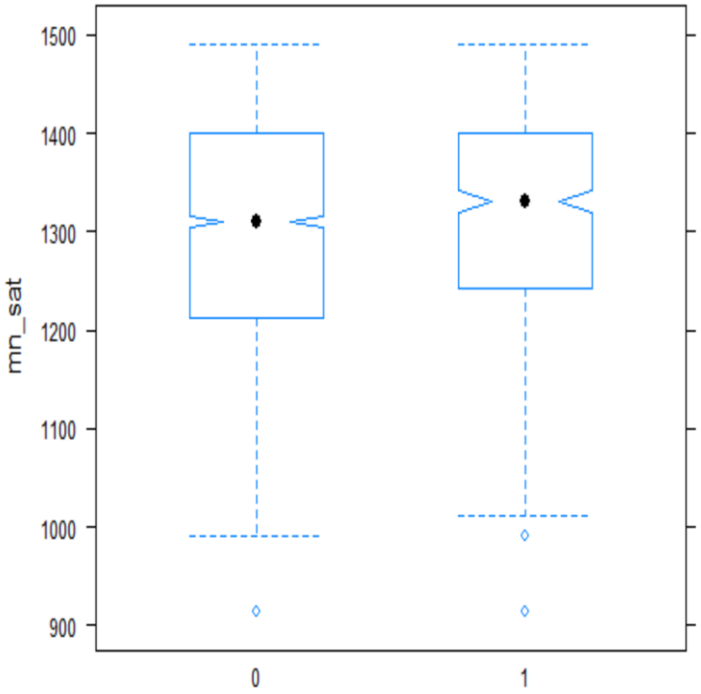
## 대체가능한 변수가 존재

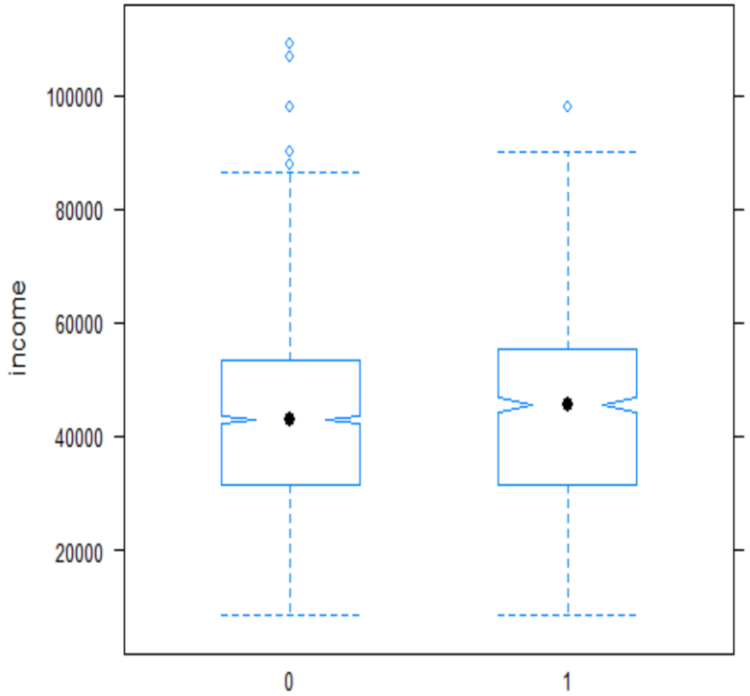
같은인종인지 여부인
samerace변수 존재
직업, 전공을 범주화한
코드변수 존재

## 불필요한 변수

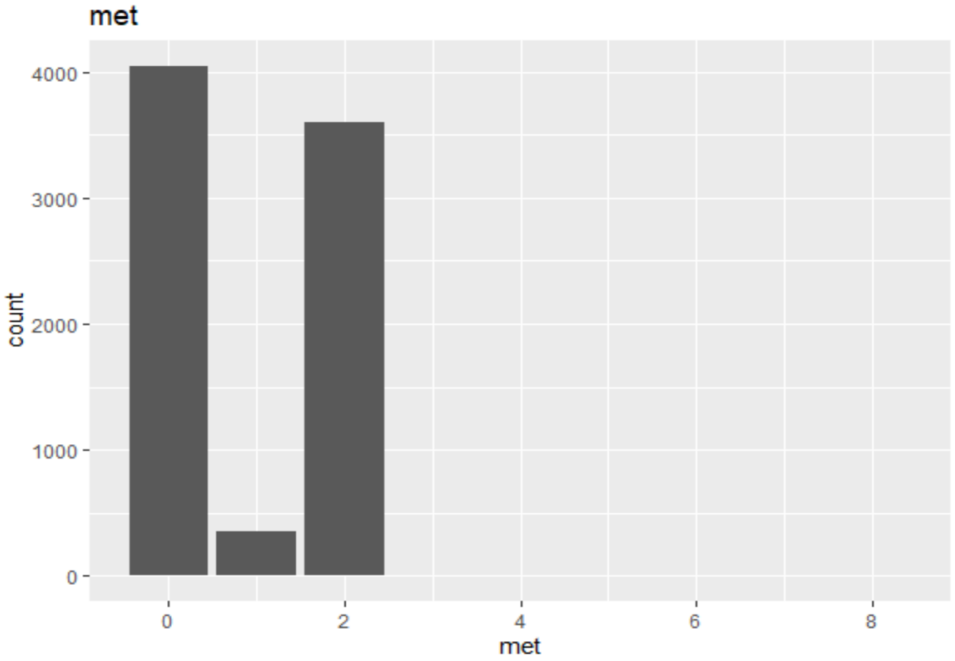
데이트 조 관련
wave, condtn
id 관련
id, partner, ldg, iid, pid

# 변수 제거


mn_sat
1500
1400
1300
1200
1100
1000
900
0
1


income
100000
80000
60000
40000
20000
0
1

# 이상치 처리




## met

1or 2 값을
갖는 변수에서
0값이 반 이상
-> 제거

상대방의 응답인 met_o변수만 사용

# 이상치 처리

attr_o, fun_o에서 10 이상인 값 -> 10으로

예상 매칭 성공 수(match_es)는 정수 -> 소수점 반올림

전에 만난 적 여부(met_o) 는 1 또는 2 -> 2 이상
값을 2로

# 결측값 처리

결측값을 3개 초과로 가지고 있는 행 637개 제거

나머지 변수들 & 결측값 많지 않음
-> 해당 행 삭제

0~10 과 같이 범위가 있는 변수들
-> 중앙값으로 대체

# 변수생성

activ

exercise + sports +hiking + yoga +clubbing

culture

museums + art + theater + concerts + movies + music +shopping

static

dining + gaming + tvsports + reading + tv



age_diff

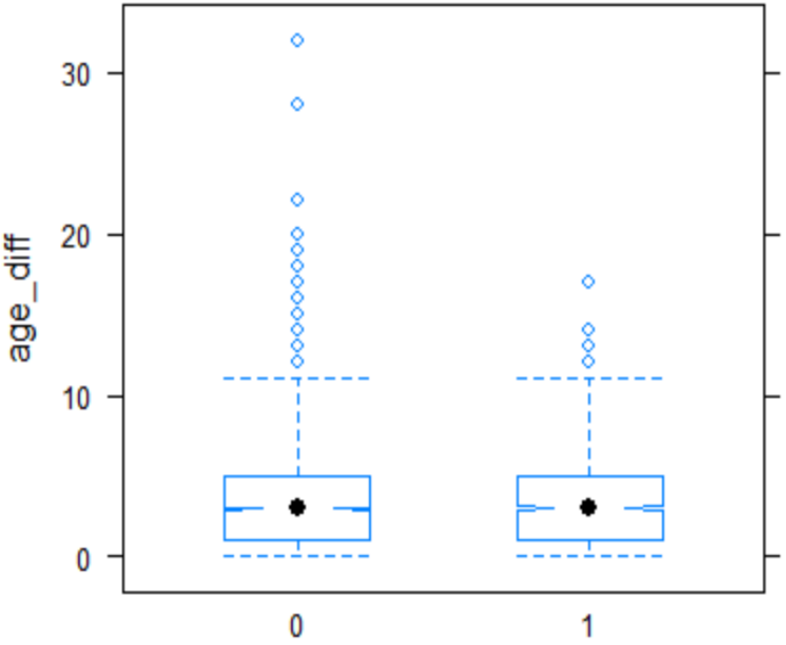
abs(age - age_o)

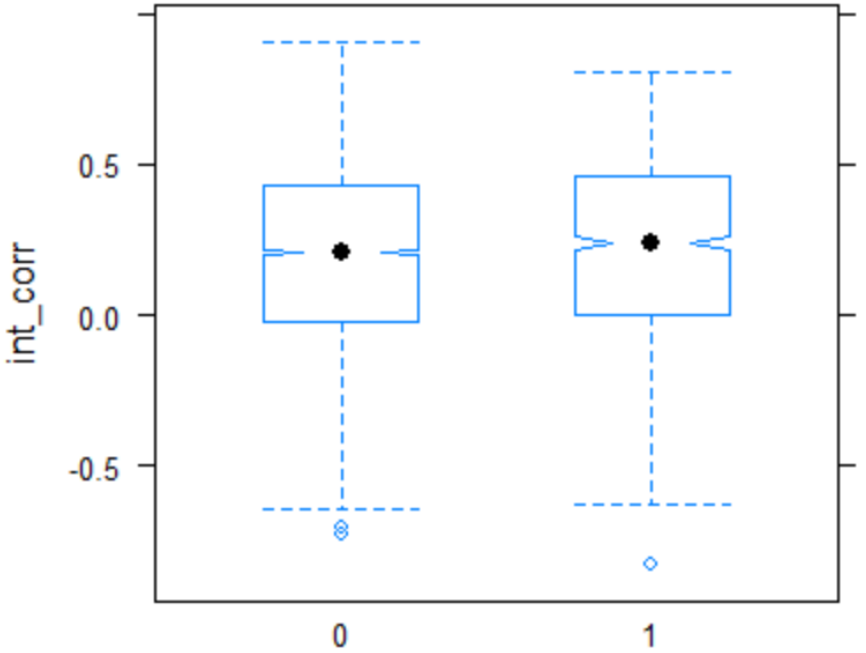


match_mean

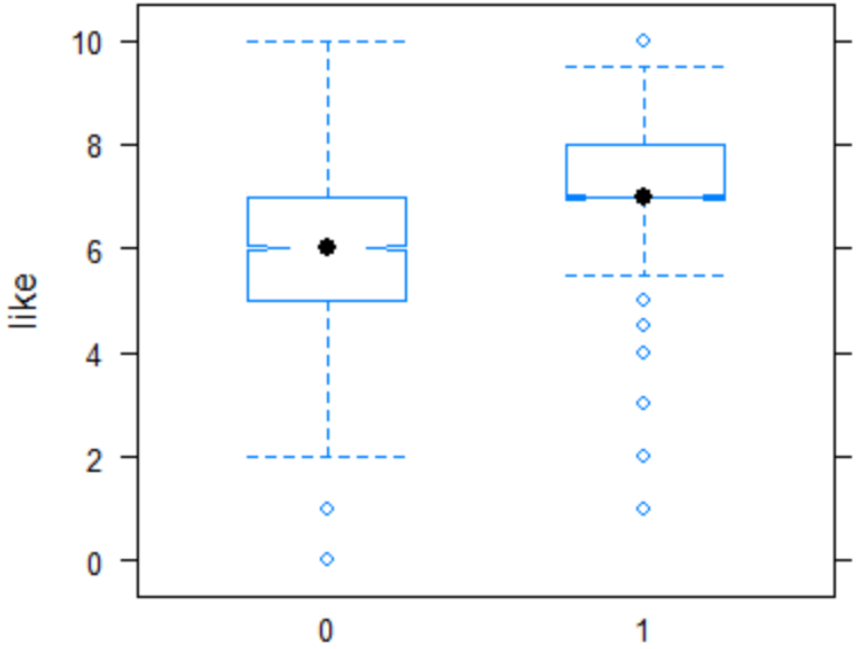
match_es/round

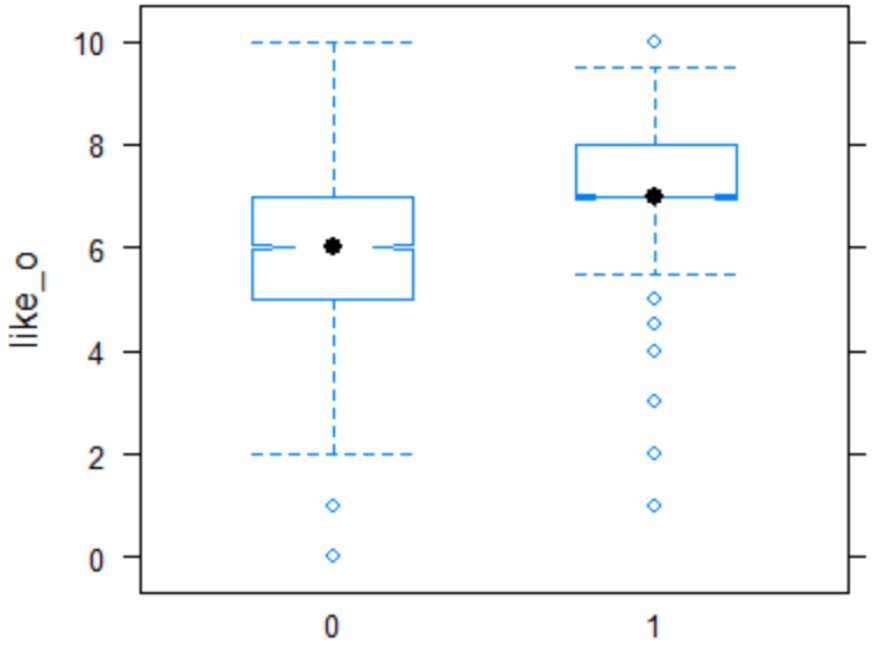
# 데이터 EDA



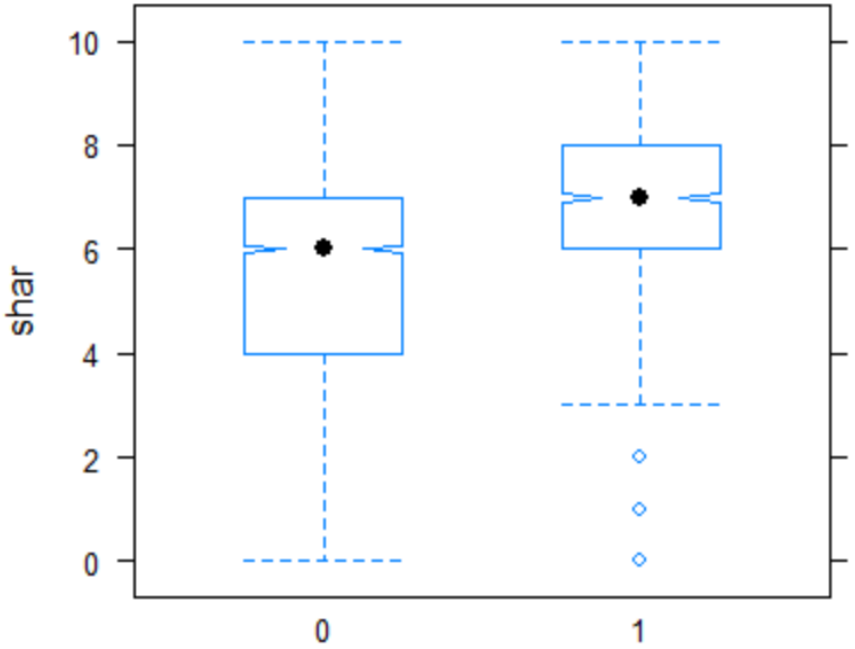



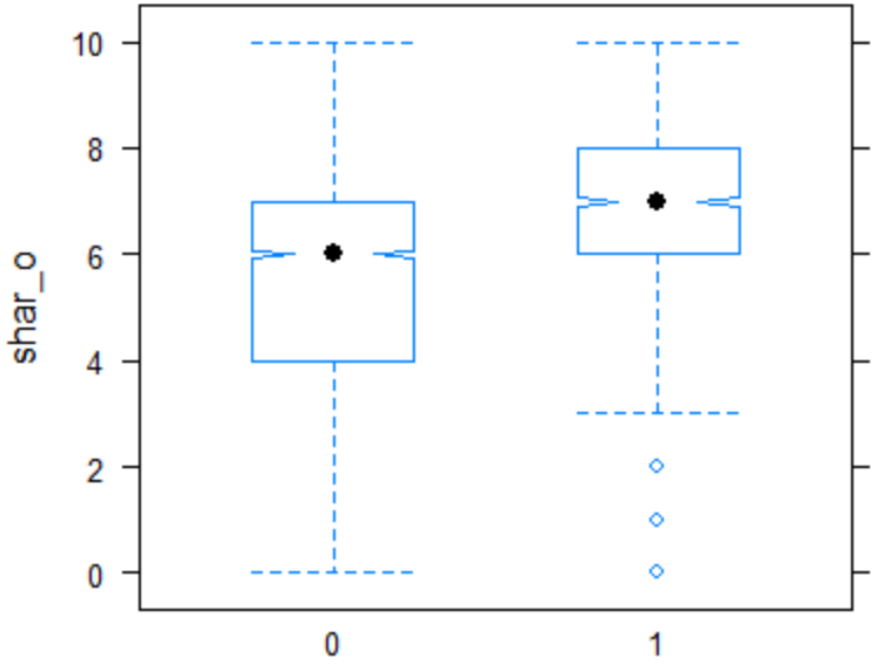
# 데이터 EDA



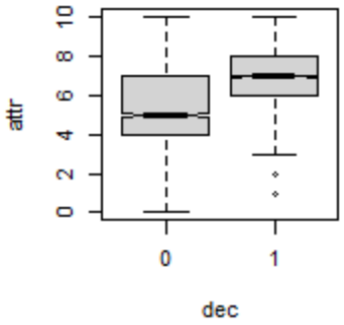



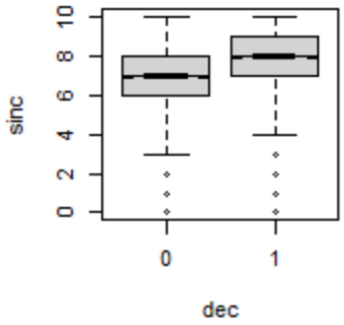
# 데이터 EDA

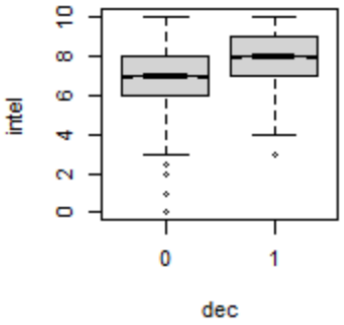


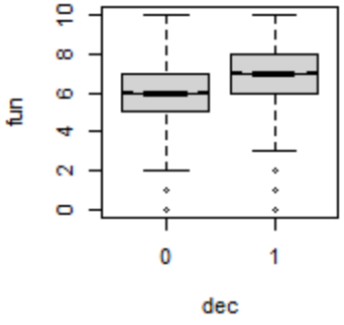



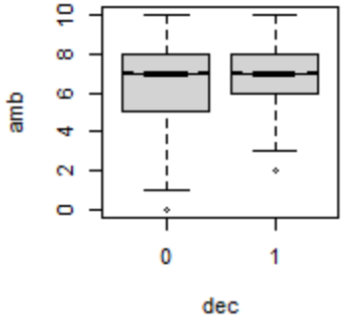
# 데이터 EDA

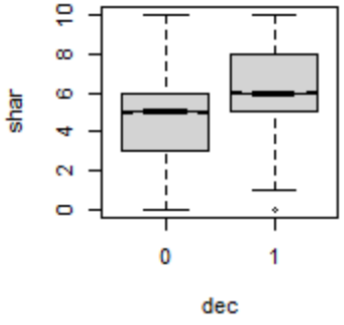

attr
0 2 4 6 8 10
0
1
dec


sinc
0 2 4 6 8 10
0
1
dec

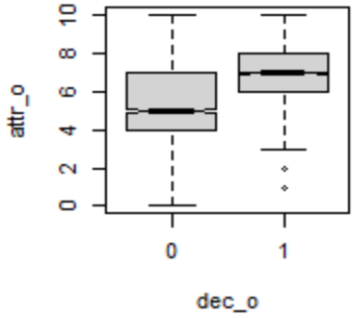

intel
0 2 4 6 8 10
0
1
dec

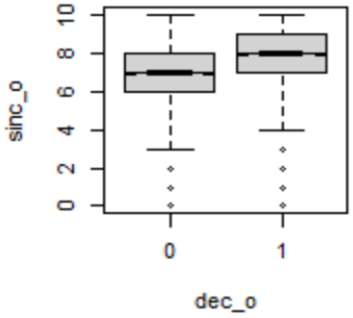

fun
0 2 4 6 8 10
0
1
dec

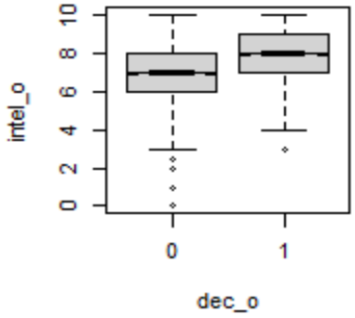

amb
0 2 4 6 8 10
0
1
dec

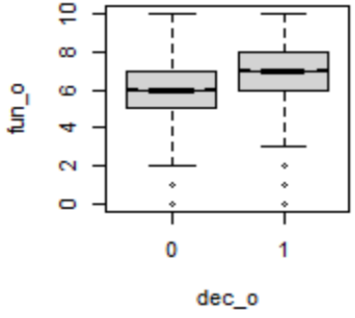

shar
0 2 4 6 8 10
0
1
dec

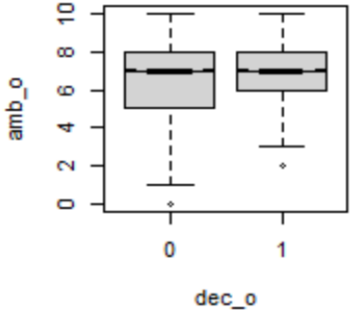
# 데이터 EDA

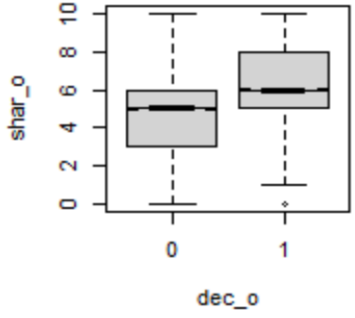

attr_o
0 2 4 6 8 10
0
1
dec_o


sinc_o
0 2 4 6 8 10
0
1
dec_o


intel_o
0 2 4 6 8 10
0
1
dec_o

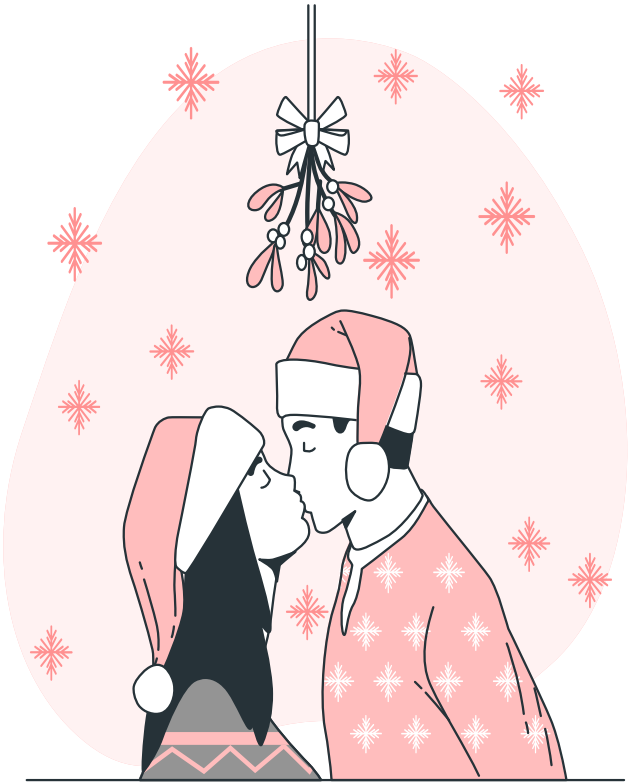

fun_o
0 2 4 6 8 10
0
1
dec_o


amb_o
0 2 4 6 8 10
0
1
dec_o


shar_o
0 2 4 6 8 10
0
1
dec_o

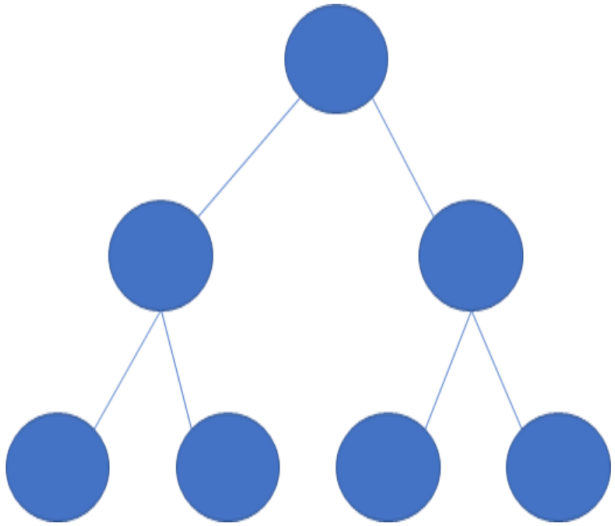
# 03

# 모델링



# 사용한 모델

## 의사결정 나무



데이터를
“뿌리” 로부터
말단의 “잎” 까지
순차적으로
분류하는 모델

모델 선정 이유
- 해석력
- 빠르고 간단함
- 정확도보다는 해석에 중점을 두었기 때문에 선택
- 회귀, 분류 모두 가능

# Train / Test set 나누기

Train set

70%

Test set

30%

# Parameter tuning

C5.0Control

## Winnowing (True/False)

입력 필드에 대해서 사전에 필드가 유용한지 측정한 다음 유용하지 않는 경우 배제하고 모델링

## pruning severity (0~1)

지역적 가지치기의 강도를 조정

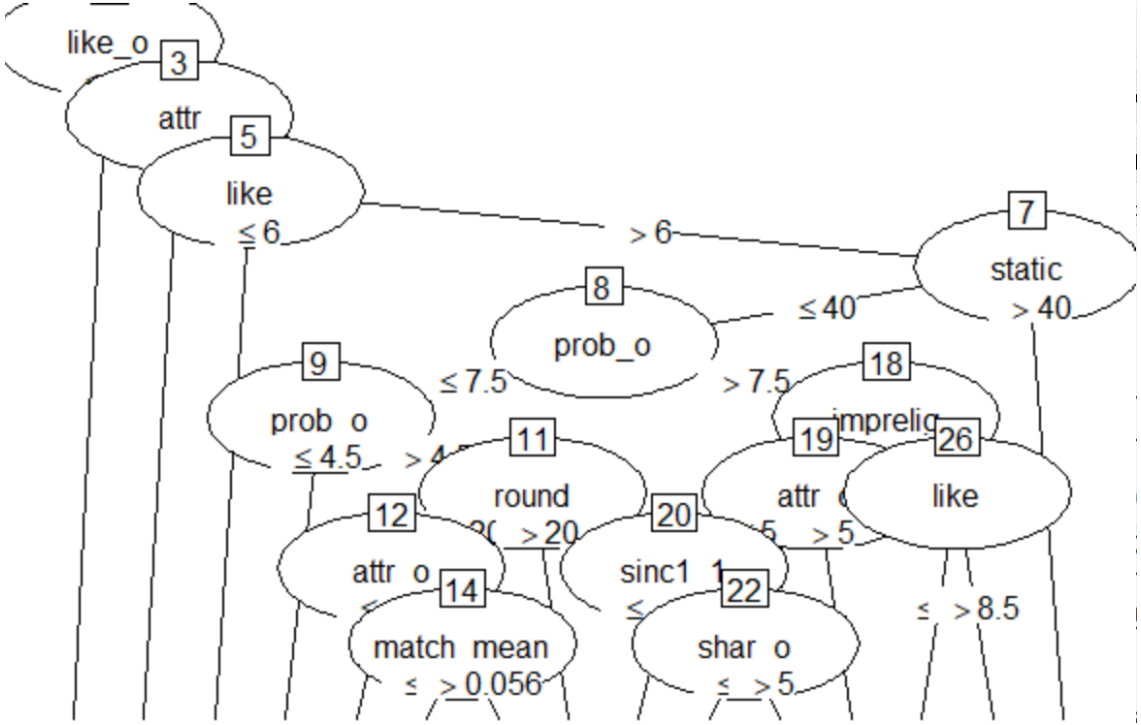
클수록 overfitting

## global Pruning (True/False)

전역적 가지치기 여부를 결정

**변수가 많으므로 winnowing=True, overfitting이 일어나지 않도록 cf=0.005, 가지치기 진행(True)을 parameter로 설정**

# 모델 plot



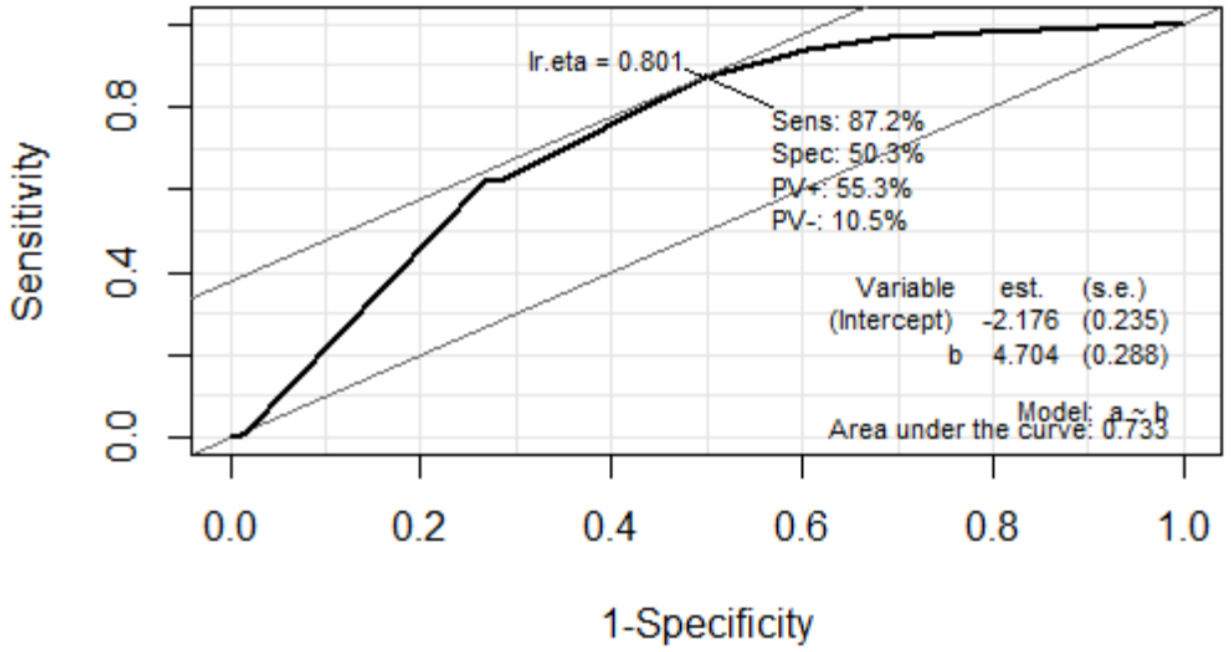

변수가 많아 간단한 tree를 구하는데 어려움이 있음

# 모델 summary

```
Read 5228 cases (54 attributes) from undefined.data

5 attributes winnowed
Estimated importance of remaining attributes:

      14%  attr
      12%  like_o
       5%  prob
       3%  attr_o
       1%  pf_o_att
       1%  pf_o_fun
       1%  shar
```

```
Attribute usage:

100.00% like_o
 45.64% attr
 22.19% like
 17.20% static
 16.68% prob_o
 12.55% attr_o
  9.91% round
  4.50% imprelig
  3.98% match_mean
  1.61% pf_o_int
  1.26% shar1_1
  0.54% sinc1_1
  0.40% shar_o
  0.34% intel
  0.27% int_corr
  0.27% shar
```

# ROC curve


lr.eta = 0.801
Sens: 87.2%
Spec: 50.3%
PV+: 55.3%
PV-: 10.5%
Variable est. (s.e.)
(Intercept) -2.176 (0.235)
b 4.704 (0.288)
Model: a ~ b
Area under the curve: 0.733
Sensitivity
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
1-Specificity

# 04

# 결과 및 한계점

# CrossTable

| act\pred | 0 | 1 | Row Total |
|---|---|---|---|
| 0 | 1799<br>0.803 | 59<br>0.026 | 1858 |
| 1 | 264<br>0.118 | 118<br>0.053 | 382 |
| Column Total | 2063 | 177 | 2240 |

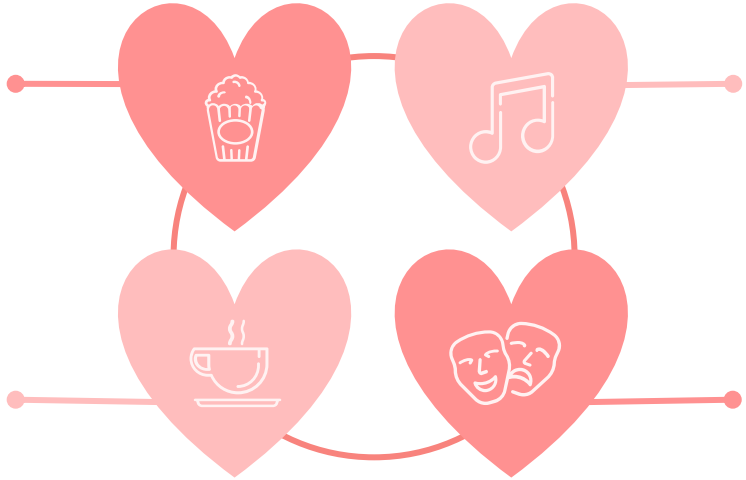
# 모델 해석

## Like_o

상대가 나에 대한
호감이 높을 수록

## attr

자신이 매력적이라
생각할 수록



## prob

상대가 나를 선택할
가능성을 높게 볼 수록

## Attr_o

상대가 나를 매력적이라
생각할 수록

# 매칭확률 up!

# 한계점

- 데이터 만으로는 현장의 분위기를 자세히 알 수 없음
- 많은 이상치와 결측값으로 인한 정보손실
- 주관적 데이터
- 많은 변수로 해석에 용이한 plot을 그리지 못함

# Thanks

CREDITS: This presentation template was created by **Slidesgo,** including icons by **Flaticon**, infographics & images by **Freepik** and illustrations by **Storyset**