나에게도 봄이 올까?

김영석, 양원직, 김희수, 오수현

목차

분석 목적

분석 목적

스피드 데이트 참가자들의 데이터들을 통해

데이트 후 애프터를 할 의향이 있는지 예측

데이터 및 변수소개

데이터 소개

	iid	id	...	match	...	amb5_3
1	70	5	...	0	...	NA
2	351	11	...	0	...	NA
3	345	5	...	1	...	8
4	351	11	...	0	...	NA
...	...	...	...	...	...	...
6702	94	1	...	0	...	NA

변수: 195개

관측치: 6702개

변수 소개

★: attr, sinc, intel, fun, amb, shar
(매력, 정직, 총명, 재미, 야망, 공감)

인적사항 관련 변수들(34)

gender, age, filed, field_cd, undergra, mn_sat, tuition, race, imprace, imprelig, from, zipcode, income, go out, career, career_c, sports, tvsports, excersice, dining, museums, art, Hiking, gaimg, clubbing, reading, tv, theater, movies, concerts, music, shopping, yoga, date

본인이 생각하는 본인의 가치관과 관련된 변수들(54)

★3_1, like, ★1_s, ★3_s, satis_2, length, numdat_2, ★7_2,★1_2, ★1_3, ★7_3, ★3_3, ★3_2

본인

타인

본인 & 타인

기타

변수 소개

★: attr, sinc, intel, fun, amb, shar
(매력, 정직, 총명, 재미, 야망, 공감)

- 상대방과 관련된 변수들(18)

age_o, race_o,
pf_o_★, dec_o,★_o,
like_o, prob_o, met_o

- 타인의 가치관과 관련된 변수들(24)

★4_2, ★2_2,
★4_3, ★2_3

본인

타인

본인 & 타인

기타

변수 소개

★: attr, sinc, intel, fun, amb
(매력, 정직, 총명, 재미, 야망)

본인

타인

본인 & 타인

기타

- 타인이 생각하는 본인과 관련된 변수들(18)

prob,match_es, expnum, ★5_1, ★5_2, ★5_3

- 본인이 생각하는 타인에 대한 변수들(19)

★1_1, ★4_1, ★2_1, dec, attr, sinc, intel, fun, amb, shar

변수 소개

ID관련 변수들(5)

iid, id, pid,
partner, idg

경험과 관련된 변수들(6)

you_call, them_cal,
date_3, numdat_3,
num_in_3, met

기타 변수들(11)

condtn, wave, round,
position, positin1, order,
match, int_corr, samerace,
goal, exphappy

본인

타인

본인 & 타인

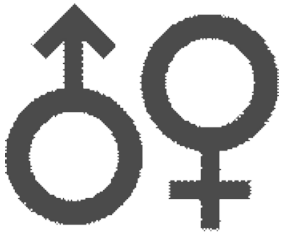
기타

데이터 전처리

데이터 전처리

WAVE

1~5웨이브, 6~9웨이브, 10~21웨이브
별로 설문항목에 대한 척도 및 결측치가 다름



성별

여자, 남자 별로
특성이 다를 것이라고 생각

성별과 WAVE

총 6개로 나눔

데이터 전처리 - 성별, wave별로 나누기 전

- 대부분 변수에서 결측값인 행 68개
- 해석이 불가능한 변수
  예시) int_corr
- 결측치가 많은 변수
  예시) undergra, goal, tuition 등
- 다른 변수로 대체 가능한 변수
  예시) field, career 등

데이터 제거

데이터 정리

새로운 변수 생성

데이터 재범주화

데이터 전처리 - 성별, wave별로 나누기 전

- 최저점수를 1점으로 만들어 주기
  예시) fun_o, imparce, museums 등
- 최고점수를 10점으로 만들어 주기
  예시) gaming, reading
- 소수점 반올림 해주기
  예시) match_es

데이터 제거

데이터 정리

새로운 변수 생성

데이터 재범주화

데이터 전처리 - 성별, wave별로 나누기 전

- 총합 100점이 되는 변수 6개를 각 항목의 합으로 나눔
  예시) pf_o_att~, attr1_1~, attr2_1~ 등

- From 변수 나라 이름으로 변경
  => Country 변수 생성

데이터 제거

데이터 정리

새로운 변수 생성

데이터 재범주화

데이터 전처리 - 성별, wave별로 나누기 전

- field_cd
: 문과, 이과, 공과,
예술, 사범, 기타

- career_c
: 사회, 과학, 예술/문화,
무직, 기타

- Country
: 6대륙

- date_3
: NA-> 0, 나머지 -> 1

- met_o
: 3,5,6,7,8 -> 1

데이터 제거

데이터 정리

새로운 변수 생성

데이터 재범주화

데이터 전처리 - 성별, wave별로 나눈 후

결측치 처리된 값들 처리

- 나이 (소수점) -> 올림
- 설문 항목 점수들 (소수점) -> 반올림
- 합이 10이 되어야하는 항목들 -> 다시 비율 맞춰줌

데이터 정리

변수 병합

변수 제거

데이터 전처리 - 성별, wave별로 나눈 후

- 여가생활 항목들 묶어서 더해줌
  예시) sports.s = sports +tvsports + exercise + hiking + yoga
  art.s = museums + art + theater + movies + concerts+music

- 같은 종류의 설문조사 항목들끼리 더해줌
  예시) attr.s = attr1_1 + attr1_2 + attr2_1 ...

- 상대방의 의한 평가 항목끼리 더해줌
  예시) interests = attr_o + fun_o + intel_o ...

데이터 전처리 - 성별, wave별로 나눈 후

- 이미 나누어주어 필요 없는 변수
  예시) iid, id, idg, partner, pid

- 논리적으로 필요 없을 것 같은 변수
  예시) gender, wave

- 분산이 0인 변수
  예시) met_o

데이터 정리

변수 병합

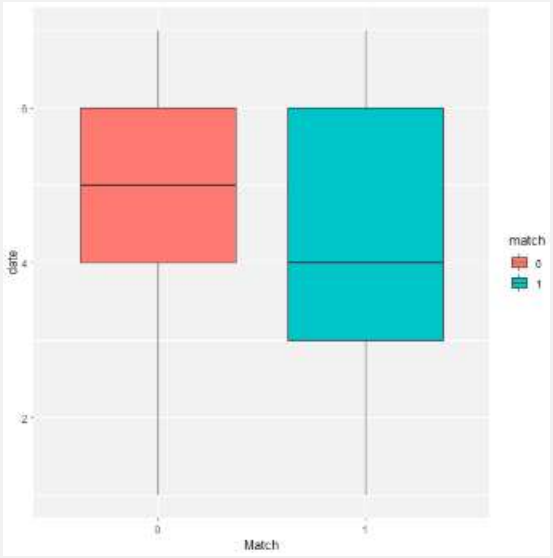
변수 제거

시각화

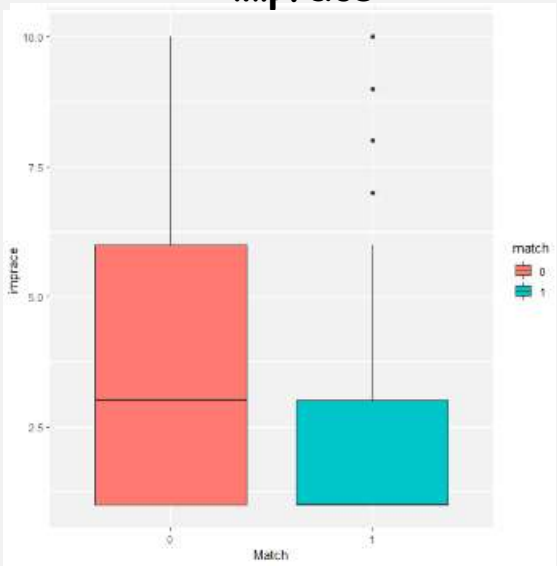
시각화

<연속형 변수들 boxplot>

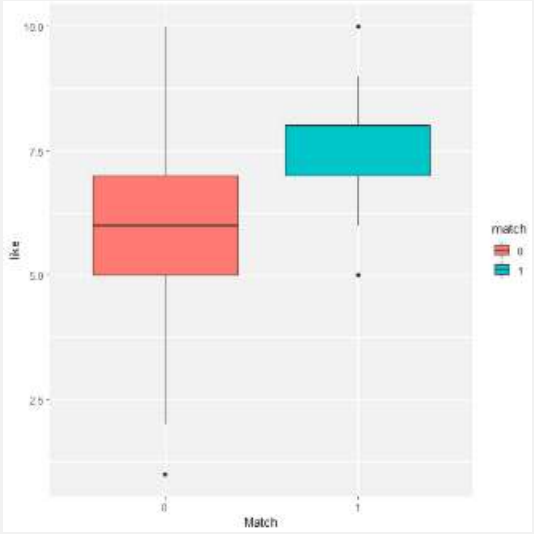
date



imprace



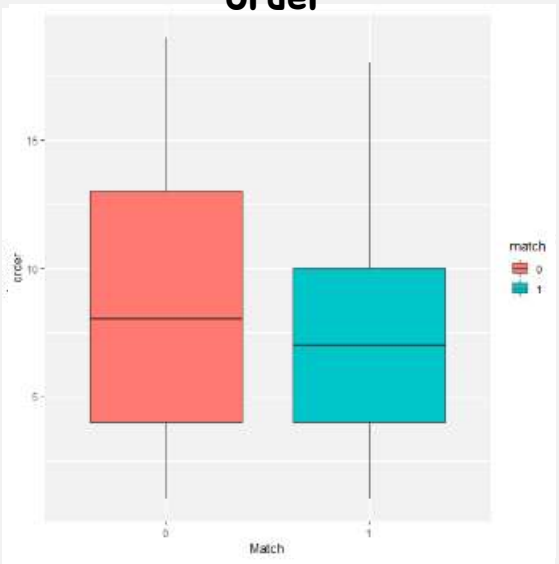
like



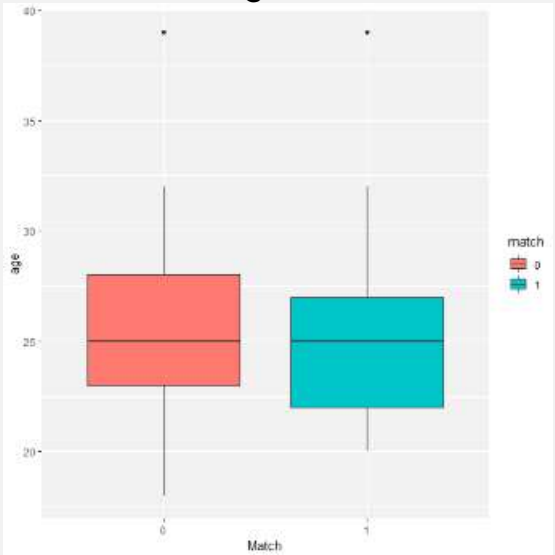
시각화

<연속형 변수들 boxplot>

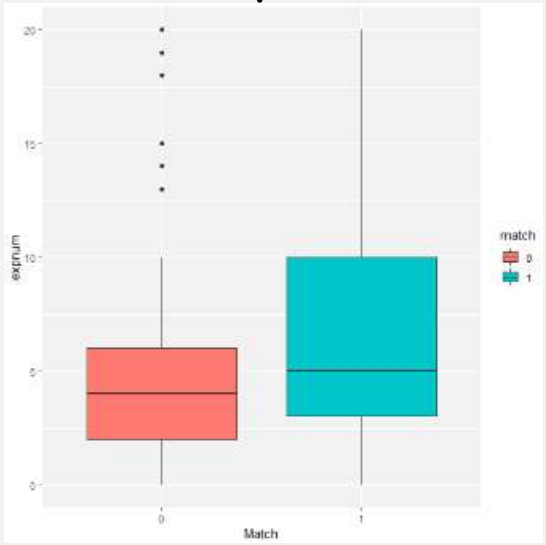
order



age



expnum



시각화

시각화 후 차이가 없어보이는 변수들

↓

hotelling T 제곱 검정

↓

결과가 유의하다고 나온 경우

↓

각각 T 검정을 실시해 변수 선택

연속형 변수

범주형 변수

시각화

table과 카이스퀘어 검정 확인

타겟변수인 match와
종속인 변수들 선택

연속형 변수

범주형 변수

모델링

모델링

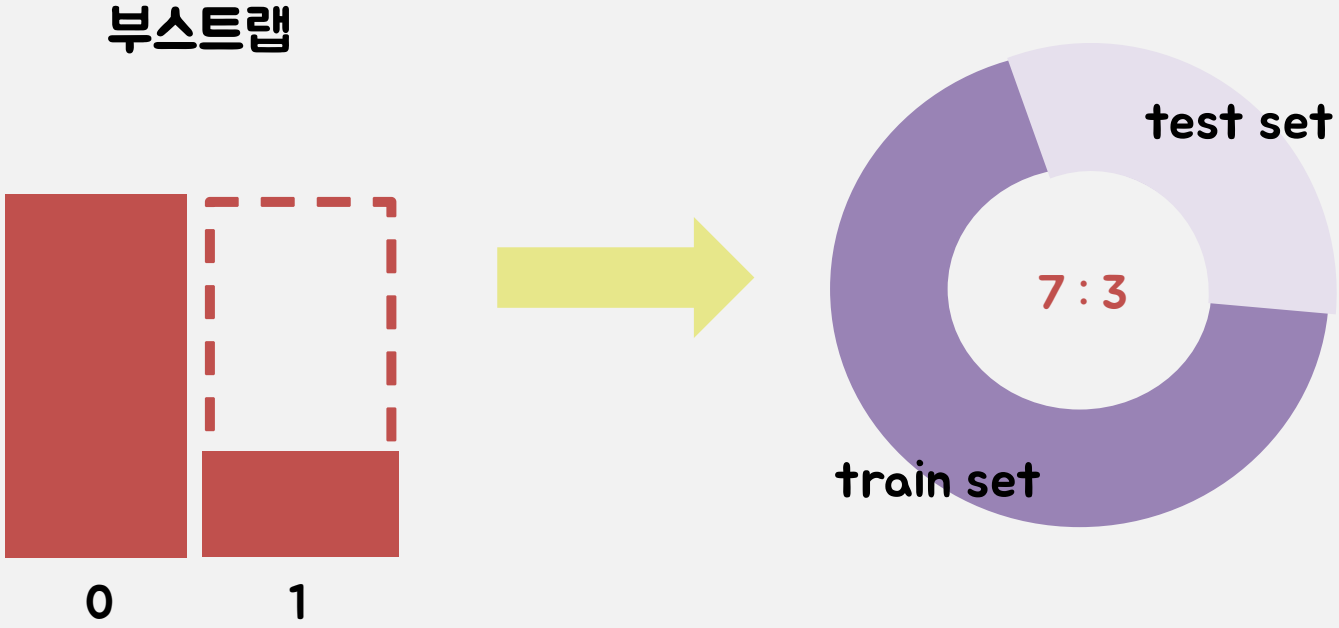
부스트랩

회귀 모형

로지스틱

랜덤포레스트

최종 모델



모델링

앞서 선택한 변수들로
회귀 모형 생성해 다중공선성 확인

condtn 과 round
다중공선성 발생

해석 불가능한
condtn 제거

부스트랩

회귀 모형

로지스틱

랜덤포레스트

최종 모델

모델링

변수 match에 대해 이항 로지스틱 모형 생성

- dec, dec_o -> 영향력이 너무 커서 제거
- position -> 많은 범주 때문에 과적합 문제 생성해 제거

부스트랩

회귀 모형

로지스틱

랜덤포레스트

최종 모델

모델링

로지스틱 모형의 VIF값 확인 후
Stepwise를 이용하여 변수 선택

그 결과 남은 변수(23)

order + race_o + age + field_cd + race +
goal + go_out + reading + expnum + like + prob + match_es +
satis_2 + length + sinc.s + fun.s + amb.s + interests + sports.s +
art.s + amusement.s + malling.s + met_o

모델링

Type 3 sum of squares를 통해
다시 한번 유의한 변수 선택

최종 변수(19)

order + race_o + age + field_cd + race +
goal + expnum + exphappy + like + prob + match_es +
sinc.s + fun.s + amb.s + interests + sports.s +
art.s + amusement.s + malling.s

모델링

앞서 나온 최종 변수로 나온 로지스틱 예측률

〈train set〉

	0	1
0	330	52
1	81	376

약 84.14%

〈test set〉

	0	1
0	145	19
1	38	147

약 83.66%

모델링

랜덤 포레스트를 통한 예측 => overfitting

〈train set〉

	0	1
0	382	29
1	5	423

약 95.94%

〈test set〉

	0	1
0	172	2
1	11	164

약 96.27%

모델링

overfitting을 해결하기 위해

로지스틱 모형에서 선택된 변수로
랜덤포레스트 모형 생성

트리 갯수 확인 후 수정

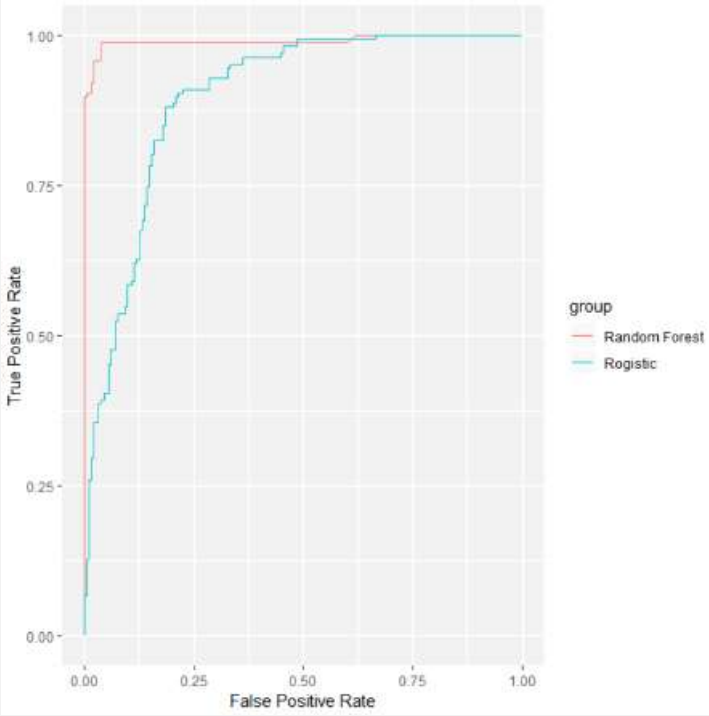
랜덤포레스트 최종 모형

랜덤포레스트 최종 모형
예측률

	0	1
0	1371	125
1	237	1518

약 88.49%

모델링



〈roc 커브 그래프〉

로지스틱과 랜덤포레스트 비교

랜덤포레스트 채택

부스트랩

회귀 모형

로지스틱

랜덤포레스트

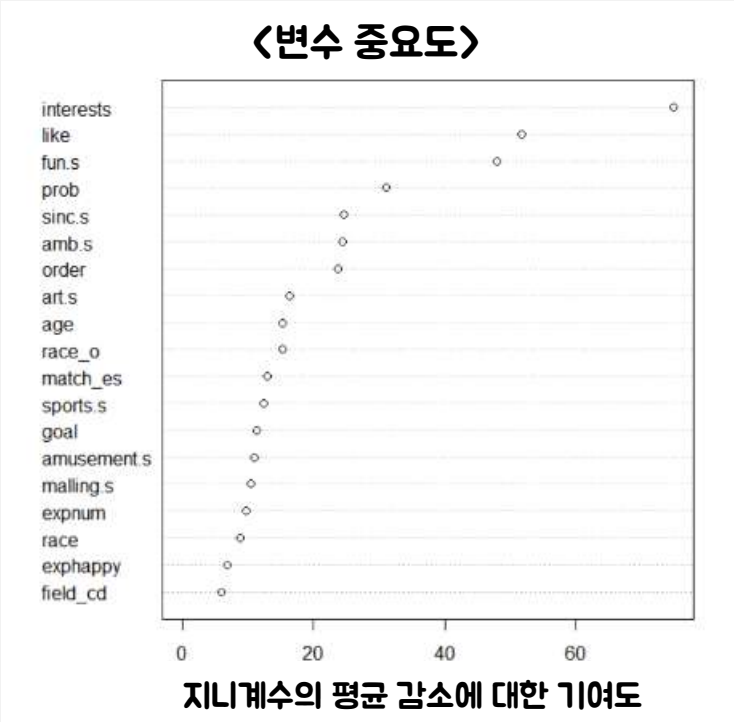
최종 모델

최종 결론

최종 결론

최종 랜덤포레스트 모형
88.49%

"interests, like, fun.s"
애프터에 중요한 요소!



결론

아쉬운 점

최종 결론

우리가 갖고있던 데이터로 예측했을때

서포트벡터머신 예측률 97~100%

하지만, test 데이터로 돌렸을시 굉장히 낮은 예측률.

overfitting 해결하지 못함

감사합니다.

Q&A