1. 안녕하십니까? Speed dating 발표를 맡은 통계학과 김나형이라고 합니다. (짝짝짝)
2. 저희의 발표 구성은 데이터소개, 분석목적을 확인하고, 데이터를 탐색, 정재한 후 모형 구축과 설명을 드린 후, 마지막으로 분석결과를 보고 결론을 내리겠습니다.
3. 첫번째, 데이터 소개입니다.
4. Speed datin이란, 독신 여자, 남자가 애인을 찾을 수 있도록 여러 사람들을 돌아가며 잠깐씩 만나보게 하는 행사입니다. 약 4분정도 만난다고 생각하시면 됩니다.
5. 다음으로 분석목적입니다.
6. 분석 목적은 매칭 성공에 영향을 주는 요인을 찾아 종속변수인 match를 맞추는 것 입니다.
7. 다음은 데이터 탐색 및 정재입니다.
8. 주어진 데이터 셋은 196개의 변수들과 6702개의 관측치를 가지고 있습니다. 변수와 결측치가 상당히 많은 것을 알 수 있습니다.
9. 192개의 변수들 중 다음과 같은 주요 변수들이 있습니다.
10. 일단, 196개의 변수들을 연속형과 범주형 변수로 나누었습니다..
11. 연속형 변수들은 종속변수인 match와 박스플랏을 그려 유의미해 보이는 변수들을 선택하였습니다.
12. 여긴 1~100점 척도를 갖는 변수들 중 boxplot을 다 그려보아 그 중 유의미해 보이는 것들을 추려 놓은 그래프입니다. 보시는 바와 같이, pf\_o\_sha , shar1\_1, shar1\_2 변수들이 종속변수 (match)에 따라 차이가 큰 것을 볼 수 있습니다.
13. 여긴 1~10점 척도를 갖는 변수들 중 boxplot을 다 그려보아 그 중 유의미해 보이는 것들을 추려 놓은 그래프입니다. 보시는 바와 같이, fun\_o, sharing\_o, like\_o 변수들이 종속변수에 따라 큰 값의 차이를 갖는 것으로 보입니다.
14. 이성을 볼 때 중요하다고 생각하여 중요도를 점수로 부여하는 변수들의 결측치는 0으로 대체하였고 1~10사의 값만 존재하여야 하지만 10보다 큰 값이 있는 경우 10으로 대체하였습니다. 또한, 합산하여 100보다 큰 경우 합산한 점수로 나누어 합이 100이 되도록 대체하였습니다.
15. Pf\_o~~인 변수들은 모두 합하여 100 이 되어야 하는 변수들인데 결측치를 제외한 나머지 항목을 더하여 100이 되므로 0을 할당하였습니다. 마찬가지로 ~~\_o(상대가 나를 평가한 점수)변수들, fun1\_1, amb1\_1, shar1\_1 인 변수들의 결측치도 0을 할당하였습니다.
16. ~~~\_o 인 변수들은 상대가 자신을 평가한 변수인데 각 항목을 모두 더하여 total 변수를 만들어 total\_o의 범위를 10단위로 끊어 그 범위내 평균으로 대체 하였습니다. Attr~shar 변수들은 내가 상대를 평가한 변수로 각 항목을 평가하지 않았는데 like는 결측이 아닌 사람을 보니 4명이었고 이 사람들에게 각 변수의 평균을 넣어 결측치를 대체하였습니다.또한, 사전조사를 하지 않아 int\_corr이 결측인 사람들이 존재했는데 저희 조는 사전조사가 결과에 중요한 영향을 미친다고 생각하여 삭제했습니다.
17. 다음은 범주형 변수입니다. 서로 상관계수가 높게 나오거나 의미 없는 범주가 많은 변수들은 통합하고, 종속변수인 match와 도수분포표를 그리고 카이제곱 테스트를 하여 유의미한 변수만 선택하였습니다.
18. 취미를 나타내는 변수들은 개 수가 많이 상관계수가 높은 변수끼리 묶어 새로운 변수 cult와 outside 등을 만들었습니다.
19. 첫번째 직업코드를 뜻하는 career\_cd를 채우기 위해 직업명인 Career값을 보았고, 보시는 대로 채웠습니다. Dec와 dec\_o는 제거하지 않으면 match를 그대로 예측 할 수 있기 떄문에 제거하였습니다.
20. Match\_es 는 예상매칭 성공수를 뜻하는 변수인데 만난 사람의 수와 관련이 있을 것이라고 생각하여 round와 table을 그린 후 최빈값을 넣었습니다. 그리고 wave 별로 인원이 다르기 때문에 wave내 인원인 round변수로 나누어 척도를 맞춰주었습니다.
21. From 변수의 관측치는 약 270개의 범주로 범주가 너무 많아 다음과 같이 대륙별로 범주화 하였습니다.
22. 위와 같은 과정을 거치며 196개의 변수를 30개로 줄였습니다.
23. 다음은 모형 구축입니다.
24. 랜덤포레스트란 여러 개의 의사결정트리를 만들고, 투표를 시켜 다수결로 결과를 결정하는 방법입니다. 의사결정트리는 새로운 데이터에 대한 일반화 성능이 좋지 않게 오버피팅(Overfitting)되기 쉽다는 것입니다. 오버피팅을 방지하기 위해 대표적인 방법 중 하나인 랜덤 포레스트를 사용하였습니다. 랜덤포래스트는 분산은 감소하지만 편향은 유지 되기 떄문에 여러 개의 트리를 통해 투표를 해서 오버피팅이 생길 경우에 대비할 수 있습니다.
25. 효율성 측면에서 가장 좋은 나무 하나를 n개 만들어서 랜덤 포레스트를 구성하는게 가장 좋다고 생각하는데 실제로는 나무 하나하나가 다 같게 생성되지 않습니다.
26. Mtry 의사결정나무에 사용하는 변수이며 이 변수들이 많을 수록 의사결정 나무는 더 깊어집니다. Ntree는 생성할 나무의 개 수 이며 의사결정나무 n개가 모여 랜덤 포레스트를 형성합니다.
27. 먼저 랜덤포레스트를 시작하기전에 train데이터를 70:30으로 랜덤하게 뽑아 70은 모형구축, 나머지 30은 모형확인을 했습니다.
28. 우리가 만든 랜덤 포레스트 모델에서의 변수 중요도를 보면 다음과 같습니다. 첫번쨰로 지니 불순도를 보면 이렇다. 지니 불순도가 높다는 것은 데이터가 1:1로 나누어졌다는 것으로 의미가 없다는 뜻 따라서 불순도를 낮추어주는 변수가 의미있는 것 입니다.
29. 이 변수들이 지니불순도를 낮추어주는 변수입니다.
30. 정확도 측면에서 보면 다음과 같은 변수들이 중요하다는 것을 볼 수 있습니다. // 두 방식으로 변수 중요도를 평가 해본 결과, like attr match\_ess 가 중요한 변수임을 알 수 있었습니다.
31. 마지막으로 confusion matrix를 사용하여 test set을 예측한 결과 정확도는 0.855가 나왔습니다.
32. 종속변수 match가 0,1의 범주형 변수이기 때문에 로지스틱 회귀 모형으로 매칭 성공 여부를 예측하는 모형을 구축했습니다. 앞선 랜덤 포레스트와 마찬가지로 train데이터를 70:30으로 랜덤하게 뽑아 70은 모형구축, 나머지 30은 모형확인을 했습니다. 4세트의 train, test를 통해 auc값이 1에 가깝고 confusionmatrix으로 나온 정확도가 가장 높은 모형으로 선택.
33. 첫번째 train세트를 활용하여 총 30개의 변수로 모형을 구축한 결과, from이나 career\_c, field\_cd와 같은 범주가 많은 변수에서 유의하지 않은 범주가 많이 나타나고 다중공선성의 문제가 발견됨. 이를 해결하기 위해 stepwise방법을 통해 불필요한 변수를 제거하여 모형의 설명력을 더 높이기로 함
34. 생성한 모형을 평가하기 위해 먼저 roc곡선을 그려봄.

-roc곡선 : tpr(true positive rate : match=1을 모형에서 1이라고 맞게 예측한 비율)과 fpr(false positive rate : match=0을 모형에서는 1이라고 잘못 예측한 비율)을 비교하는 그래프.

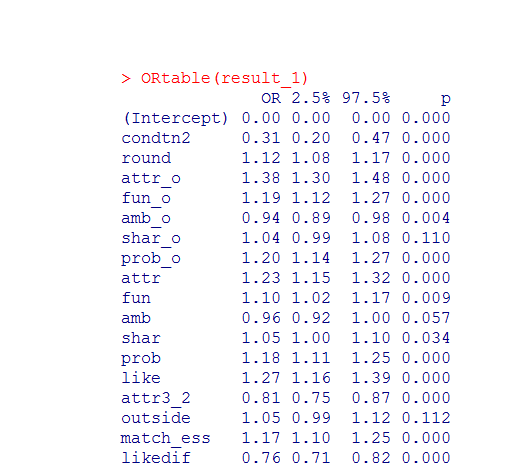
y=x 그래프일 때가 두 비율이 같을 때이기 때문에 그 대각선 위로 그래프가 그려질 경우 tpr>fpr이라는 것이고 맞게 예측한 비율이 더 크기 때문에 예측이 잘 되었다는 것을 의미함.

이 중 빨간 roc 곡선을 가질 때 auc=0.855로 가장 큰 값이 나옴. 이 때의 confusionmatrix를 구해본 결과 정확도는 0.864가 나왔음. 본래 0의 비율 83%보다 높은 정확도가 나옴.

1. 이 모형의 coefficient estimate(각 변수들의 계수 추정값)를 통해 각 요인에 따른 오즈비를 구해보았고 그것으로 match와의 관계를 파악. Shar\_o outside변수를 제외하고 모두 유의한 17개의 변수로 이루어진 모형 구축.
2. 마지막으로 분석결과입니다.
3. 감사합니다.

이 표는 OR부분을 보면 됨. OR=Odds Ratio. 오즈비인데 쉽게 말해서 이라고 생각. 가능성이라고도 해석하는데 확률이라고만 안하면 됌.

범주형변수는 기준이 되는 변수가 있음(예:from의 범주가 1부터 6까지이고 만약에 기준이 from=1이면, from=2는 1을 기준으로 했을 때 몇 배의 오즈비를 갖는다 이런식으로 해석. 다행히 우리는 condtn말고는 범주형 모두 사라짐..)



해석은 ppt에 넣은 상위 4개 하위 4개만 해주면 될 듯…

-다른것보다 이야기 해주고 넘어가면 좋을 것들-

Attr\_o : 파트너가 나에게 준 점수가 1점 올라갈 때마다 매칭 성공 가능성(오즈비)이 1.38배가 된다.

Like : 파트너가 얼마나 마음에 들었는지 매긴 점수가 1점 올라갈 때마다 매칭 성공 가능성(오즈비)이 1.27배가 된다.

Attr : 파트너의 매력도 점수에 1점 더 평가할수록 매칭 성공 가능성(오즈비)이 1.23배가 된다.

Prob\_o

Attr3\_2 : 소개팅이 끝나고 자신의 매력도 1점 더 평가한 사람일수록 매칭 성공 가능성(오즈비)이 0.81배가 된다 - 감소

Likedif : 파트너와 내가 서로 마음에 들었다고 매긴 점수의 차이가 1점 높아질수록 매칭 성공 가능성(오즈비)가 0.76배가 된다. – 점수 차이가 적을수록 매칭이 잘된다는 뜻