종속변수 match가 0,1의 범주형 변수이기 때문에 로지스틱 회귀 모형으로 매칭 성공 여부를 예측하는 모형을 구축.

앞선 랜덤 포레스트와 마찬가지로 train데이터를 70:30으로 랜덤하게 뽑아 70은 모형구축, 나머지 30은 모형확인.

4세트의 train, test를 통해 auc값이 1에 가깝고 confusionmatrix으로 나온 예측력이 가장 높은 모형으로 선택.

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

-stepwise : 모형에 변수가 유의하면 넣고, 유의하지 않으면 빼는 방식으로 AIC(올바른 모형인지 판단하는 값)가 작은 것을 선택해줌

-stepwise결과 다중 공선성이 해결되지 않은 경우 서로 연관이 높다고 나온 변수 중 뺐을 때 다중공선성이 더 낮은 변수를 빼주었음.

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.

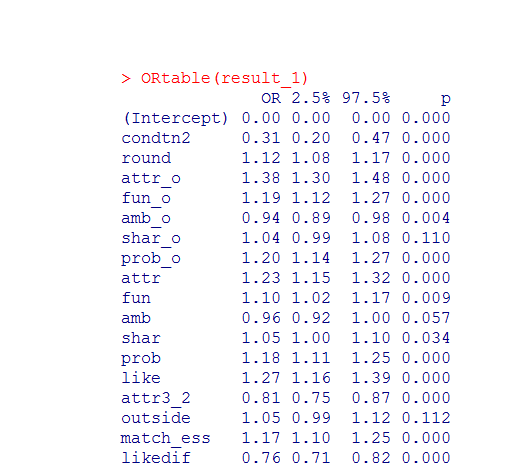
이 중 빨간 roc 곡선을 가질 때 auc=0.855로 가장 큰 값이 나옴. 이 때의 confusionmatrix를 구해본 결과 정확도는 0.864가 나왔음. 본래 0의 비율 83%보다 높은 정확도가 나옴.

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.

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