Mental Health In Company

라이언조

경영정보학과 이건우

산업경영공학과 이상인

컴퓨터 공학과 김하은

CONTENTS

01 데이터 분석 방법

02 2차 데이터 탐색

01

데이터 분석 방법

01. 데이터 분석 방법

▶ Random Forest

예측 정확도 측정

변수 별 중요도 측정

▶ Association Analysis

treatment에 영향을 주는 변수를 찾고자 함

설문조사변수 간 교호작용을 확인하고자 함

▶ Logistic Regression

유의미한 변수를 선별하고, 유의미하지 않은 변수들을 제거

treatment를 종속변수로 선정하고 독립 변수들과의 유의미한 관계 추정

02

2차 데이터 탐색

02. Random Forest

- state 변수를 제외한 모든 변수를 가지고 RF한 결과

```
##         OOB estimate of  error rate: 52%
## Confusion matrix:
##      No Yes class.error
## No  185  10  0.05128205
## Yes 289  91  0.76052632
```

no의 잘못 예측 비율 : 약 5%
yes의 잘못 예측 비율 : 약 76%

→ yes의 예측 정확도가 너무 저조

02. Random Forest

- tuneRF 함수 이용하여 오류확률이 최소가 되는 최적의 parameter값 찾기

```
## mtry = 4  OOB error = 46.96%
## Searching left ...
## mtry = 8     OOB error = 47.65%
## -0.01481481 0.05
## Searching right ...
## mtry = 2     OOB error = 45.04%
## 0.04074074 0.05
```

→ mtry = 2, ntreeTry = 50 일 때 오류 확률 45.04%로 최저

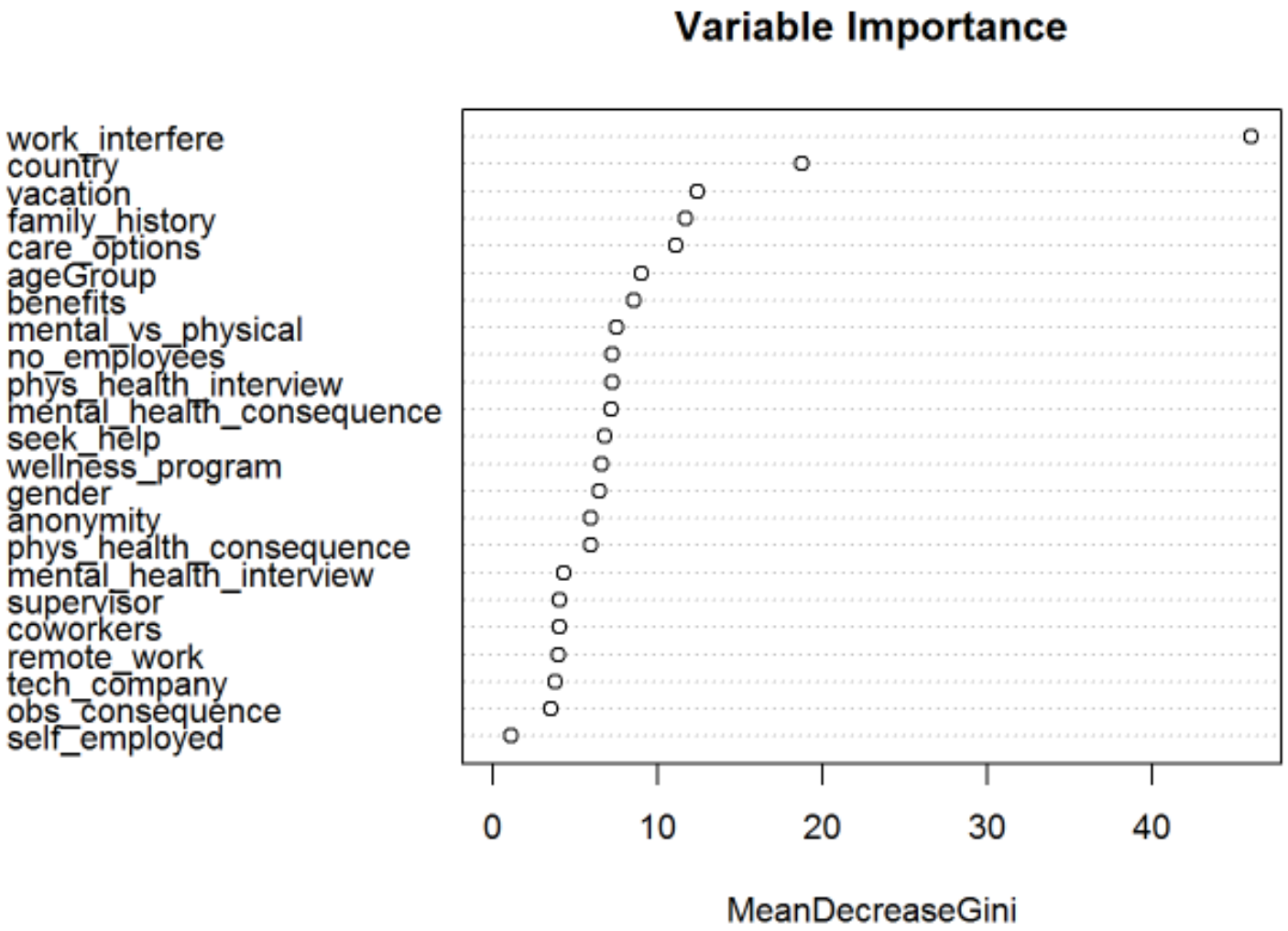
```
##         OOB estimate of  error rate: 35.3%
## Confusion matrix:
##      No Yes class.error
## No  153  42   0.2153846
## Yes 161 219   0.4236842
```

↓

mtry = 2, ntreeTry = 50로 다시 RF 한 결과,
오류 확률이 35.3%로 전보다 많이 낮아짐

02. Random Forest

- 변수 별 중요도



work_interfere가 가장 상위에 있고, 그 다음으로 country, vacation,family_history 등

02. Association Rules

- 두 결과를 미루어 보았을 때

1. 작업 환경 (work_interfere)가 정신 건강 치료법을 찾는데 영향을 많이 주고
2. 작업환경이 부정적인 영향을 끼칠 때 동료들과 그것에 대해 공유를 하여 본인 상태에 대한 확신을 얻는다고 볼 수 있음

- RHS를 'treatment = No' 로 고정했을 때 결과

```
> inspect(asso0[1:5])
    lhs                                                rhs             support   confidence lift
[1] {family_history=No,work_interfere=Never}        => {treatment=No} 0.1529284 0.9038462  2.495048
[2] {country=United States,work_interfere=Never}    => {treatment=No} 0.1030369 0.8715596  2.405922
[3] {work_interfere=Never,anonymity=Don't know}     => {treatment=No} 0.1301518 0.8695652  2.400417
[4] {work_interfere=Never,wellness_program=No}      => {treatment=No} 0.1149675 0.8760331  2.418271
[5] {work_interfere=Never,supervisor=Yes}           => {treatment=No} 0.1344902 0.8794326  2.427655
```

전반적으로 작업환경에 있어 정신 건강 상태가 방해가 된다고 느끼지 않을 때(work_interfere = Never), 정신 건강에 대한 치료법을 찾지 않음

특히 confidence가 약 90%인 confidence 값을 보았을 때 가족 병력이 없을 때 그 경향이 더 강하게 나타남을 알 수 있음

- RHS를 'treatment = Yes' 로 고정했을 때 결과

```
> inspect(asso1[1:5])
    lhs                                                   rhs              support   confidence lift
[1] {work_interfere=Often,coworkers=Yes}               => {treatment=Yes} 0.1008677 0.8691589  1.362865
[2] {work_interfere=Often,mental_health_interview=No}  => {treatment=Yes} 0.1019523 0.8623853  1.352244
[3] {coworkers=Yes,obs_consequence=Yes}                => {treatment=Yes} 0.1073753 0.8048780  1.262071
[4] {mental_health_interview=No,obs_consequence=Yes}   => {treatment=Yes} 0.1138829 0.7608696  1.193064
[5] {self_employed=No,obs_consequence=Yes}             => {treatment=Yes} 0.1008677 0.7440000  1.166612
```

작업 환경에 정신 건강 상태가 꽤 방해된다고 생각할 때 (work_interfere = Often) 정신 건강에 대한 치료법을 찾으려 함

특히 정신 건강 문제에 대해 논의할 수 록 그 경향이 강하게 나타났고 주변에서 부정적인 결과에 대해 들었을 때 치료를 찾으려고 함

02. 2차 데이터 탐색

Association Rules

country가 US 일 때와 아닐 때 두 그룹으로 나눔
변수의 levels 비율이 일정하지 않아 support를 0.05로 낮춰 분석을 진행함

```
> inspect(asso_usy_0_2[1:5])
    lhs                                                          rhs             support    confidence lift
[1] {family_history=No,mental_health_interview=Maybe}         => {treatment=No} 0.05026930 0.6363636  1.947552
[2] {supervisor=Yes,mental_health_interview=Maybe}            => {treatment=No} 0.05385996 0.4411765  1.350194
[3] {phys_health_consequence=No,mental_health_interview=Maybe} => {treatment=No} 0.05026930 0.4516129  1.382134
[4] {gender=Male,mental_health_interview=Maybe}               => {treatment=No} 0.05385996 0.4838710  1.480858
[5] {coworkers=Yes,mental_health_interview=Maybe}             => {treatment=No} 0.05565530 0.4428571  1.355338
```

가족 병력이 없고, 잠재적 고용주와 면접에서 정신건강에 대해 논의하는데 확실하지 않은 (mental_health_interview=Maybe)사람은 정신 건강 치료법을 찾지 않음

```
> inspect(asso_usy_1_2[1:10])
     lhs                                                  rhs              support    confidence lift
[1]  {tech_company=Yes,vacation=Very difficult}        => {treatment=Yes} 0.05026930 0.7567568  1.124036
[2]  {vacation=Very difficult,mental_health_interview=No} => {treatment=Yes} 0.05745063 0.8000000  1.188267
[3]  {state=WA,family_history=Yes}                     => {treatment=Yes} 0.05026930 0.8484848  1.260283
[4]  {state=WA,benefits=Yes}                           => {treatment=Yes} 0.05206463 0.8285714  1.230705
[5]  {state=WA,remote_work=No}                         => {treatment=Yes} 0.05206463 0.7073171  1.050602
[6]  {state=WA,coworkers=Yes}                          => {treatment=Yes} 0.05745063 0.6956522  1.033275
[7]  {state=WA,tech_company=Yes}                       => {treatment=Yes} 0.06283662 0.7000000  1.039733
[8]  {state=WA,mental_health_interview=No}             => {treatment=Yes} 0.06463196 0.7200000  1.069440
[9]  {state=WA,self_employed=No}                       => {treatment=Yes} 0.06463196 0.6923077  1.028308
[10] {work_interfere=Sometimes,vacation=Somewhat difficult} => {treatment=Yes} 0.05565530 0.9117647  1.354275
```

IT 회사에 있고 정신 건강 문제로 병가를 내기 어려운 사람들은 treatment를 찾았다. 워싱턴 주에 있는 사람들은 가족이 정신건강 경력이 있거나 고용주가 정신 건강 혜택을 제공함에도 정신 건강 치료법을 찾으려는 노력을 보이는 것으로 확인하였음

02. 2차 데이터 탐색

Association Rules

- country가 US 일 때와 아닐 때 두 그룹으로 나누어 보았다

```
> inspect(asso_usn_0_2[1:10])
     lhs                                                                rhs             support    confidence lift
[1]  {family_history=No,work_interfere=Rarely}                       => {treatment=No} 0.05205479 0.5277778  1.2673611
[2]  {self_employed=No,work_interfere=Rarely}                        => {treatment=No} 0.06027397 0.4313725  1.0358617
[3]  {gender=Male,work_interfere=Rarely}                             => {treatment=No} 0.05205479 0.4130435  0.9918478
[4]  {tech_company=No,obs_consequence=No}                            => {treatment=No} 0.05479452 0.4347826  1.0440503
[5]  {mental_health_consequence=No,mental_health_interview=Maybe}    => {treatment=No} 0.06027397 0.4680851  1.1240202
[6]  {mental_health_interview=Maybe,phys_health_interview=Maybe}     => {treatment=No} 0.06027397 0.5238095  1.2578321
[7]  {family_history=No,mental_health_interview=Maybe}               => {treatment=No} 0.07671233 0.6829268  1.6399230
[8]  {anonymity=Don't know,mental_health_interview=Maybe}            => {treatment=No} 0.05479452 0.4545455  1.0915072
[9]  {seek_help=No,mental_health_interview=Maybe}                    => {treatment=No} 0.05479452 0.4878049  1.1713736
[10] {supervisor=Yes,mental_health_interview=Maybe}                  => {treatment=No} 0.08767123 0.4848485  1.1642743
```

가족 병력이 없고, 잠재적 고용주와 면접에서 정신건강에 대해 논의하는데 확실하지 않은 (mental_health_interview=Maybe) 사람은 정신 건강 치료법을 찾지 않음

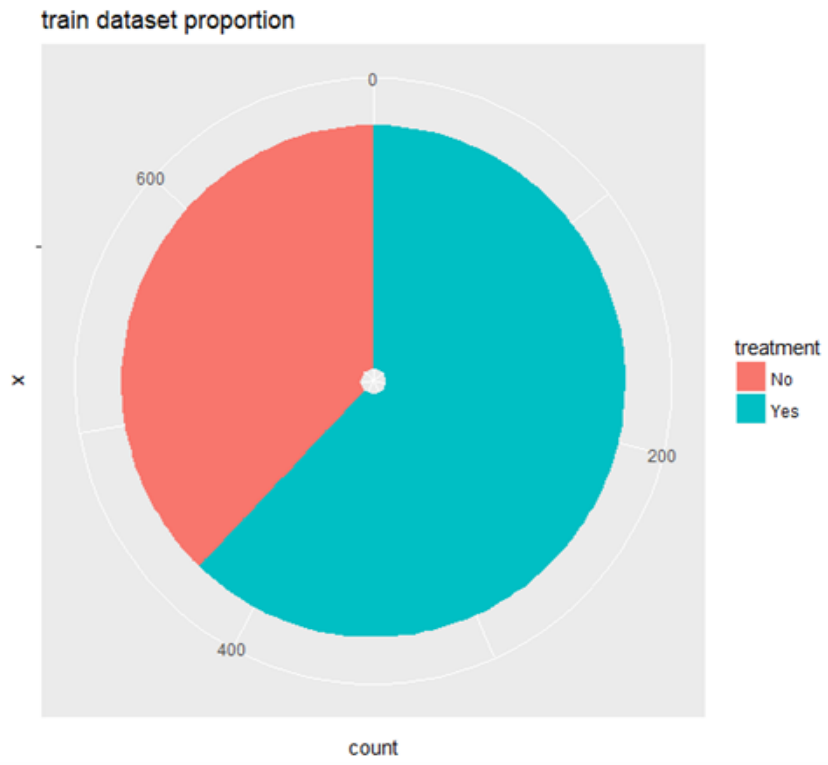
```
> inspect(asso_usn_1_2[1:10])
     lhs                                                                rhs              support    confidence lift
[1]  {anonymity=No,mental_health_consequence=Yes}                    => {treatment=Yes} 0.05205479 0.7307692  1.2522571
[2]  {seek_help=No,anonymity=No}                                     => {treatment=Yes} 0.05205479 0.5937500  1.0174589
[3]  {wellness_program=No,anonymity=No}                              => {treatment=Yes} 0.05205479 0.5588235  0.9576084
[4]  {anonymity=No,mental_health_interview=No}                       => {treatment=Yes} 0.05753425 0.6363636  1.0904823
[5]  {wellness_program=Don't know,coworkers=Yes}                     => {treatment=Yes} 0.05205479 0.5588235  0.9576084
[6]  {vacation=Very difficult,mental_vs_physical=No}                 => {treatment=Yes} 0.05479452 0.6451613  1.1055581
[7]  {benefits=No,vacation=Very difficult}                           => {treatment=Yes} 0.05753425 0.6363636  1.0904823
[8]  {seek_help=No,vacation=Very difficult}                          => {treatment=Yes} 0.06849315 0.6410256  1.0984712
[9]  {remote_work=No,vacation=Very difficult}                        => {treatment=Yes} 0.05753425 0.7241379  1.2408936
[10] {wellness_program=No,vacation=Very difficult}                   => {treatment=Yes} 0.06575342 0.6315789  1.0822832
```

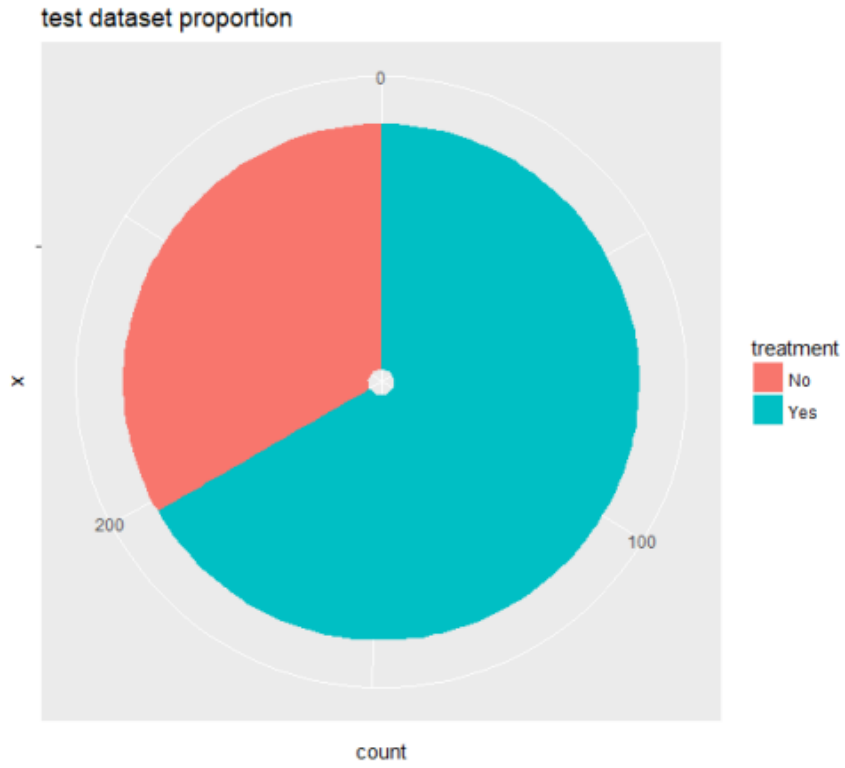
IT 회사에 있고 정신 건강 문제로 병가를 내기 어려운 사람들은 treatment를 찾았다. 워싱턴 주에 있는 사람들은 가족이 정신건강 경력이 있거나 고용주가 정신 건강 혜택을 제공함에도 정신 건강 치료법을 찾으려는 노력을 보이는 것으로 확인하였음

02. Logistic Regression

- train data와 test data로 분류 (7:3)

```
smp_size <- floor(0.7 * nrow(data1))
train_ind <- sample(seq_len(nrow(data1)), size = smp_size)
train_data<-data1[train_ind,]
test_data<-data1[-train_ind,]
```





02. Logistic Regression

- Stepwise 방법을 통해 로지스틱 회귀분석 실시하고,
  Stepwise를 통해 선별된 변수들로 다시 회귀분석과 ANOVA 분석 실시

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: treatment
##
## Terms added sequentially (first to last)
##
##
##                      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                   390     488.51
## family_history        1   34.989       389     453.52 3.316e-09 ***
## work_interfere        3  120.443       386     333.08 < 2.2e-16 ***
## no_employees          2    1.948       384     331.13  0.377625
## care_options          2   20.065       382     311.06 4.394e-05 ***
## wellness_program      2    4.417       380     306.65  0.109878
## anonymity             2    9.253       378     297.39  0.009788 **
## coworkers             1    3.227       377     294.17  0.072429 .
## phys_health_interview 2    4.193       375     289.97  0.122909
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

no_employees, wellness_program, phys_health_interview의 p-value가 유의하지 않음을 확인

해당 변수들을 제거하고, 다시 회귀분석과 ANOVA 분석 시행

02. Logistic Regression

- Reduced Model 회귀분석 및 ANOVA

```
## Call:
## glm(formula = treatment ~ family_history + work_interfere + care_options +
##     anonymity, family = binomial, data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6218  -0.3536   0.3829   0.6423   2.4666
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.2479     0.2430  -1.020 0.307589
## family_historyYes      1.0923     0.2363   4.623 3.78e-06 ***
## work_interfere.L       1.8595     0.2411   7.711 1.25e-14 ***
## work_interfere.Q      -1.8637     0.2805  -6.643 3.07e-11 ***
## work_interfere.C       1.4032     0.3194   4.394 1.12e-05 ***
## care_optionsNot sure  -0.2522     0.2877  -0.877 0.380731
## care_optionsYes        1.1025     0.2907   3.793 0.000149 ***
## anonymityNo           -0.4515     0.6843  -0.660 0.509356
## anonymityYes           0.5041     0.2800   1.800 0.071791 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 740.63  on 586  degrees of freedom
## Residual deviance: 485.25  on 578  degrees of freedom
## AIC: 503.25
##
## Number of Fisher Scoring iterations: 5
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: treatment
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                              586     740.63
## family_history  1   58.011       585     682.62 2.606e-14 ***
## work_interfere  3  158.681       582     523.94 < 2.2e-16 ***
## care_options    2   34.627       580     489.31 3.026e-08 ***
## anonymity       2    4.060       578     485.25    0.1313
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

02. Logistic Regression

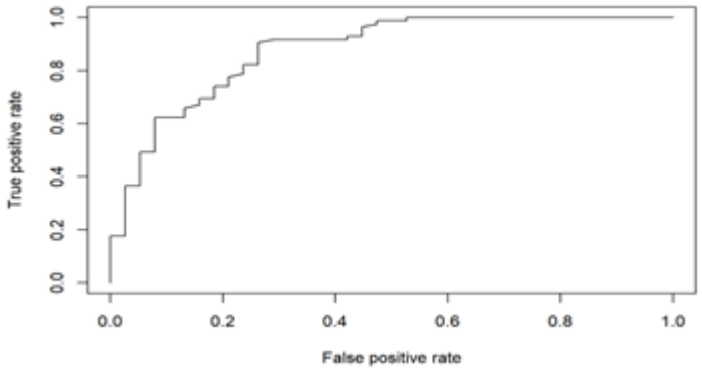
- McFadden R스퀘어로 해당 모델의 Fit 확인

```
pR2(reduced_model2)
```

```
##          llh      llhNull           G2     McFadden          r2ML
## -242.6255660 -370.3153168  255.3795016    0.3448136    0.3527734
##          r2CU
##    0.4921271
```

R스퀘어 값이 0.49로
약 49%의 설명력이 있음을 확인

- ROC 그래프를 통해 모델 평가



- AUC를 통해 모델 평가

```
auc <- performance(pr, measure="auc")
auc <- auc@y.values[[1]]
auc
```

```
## [1] 0.8780186
```

1에 가까운 0.87이라는 값이 나왔으므로
해당 모델이 결과를 잘 예측한다고 볼 수 있음

02. Logistic Regression

- 결론

결론적으로 모델에서 유의한 변수로 선택된 family_history, work_interfere, care_options,anonymity가 treatment에 주요한 영향을 끼치는 것으로 파악됨

즉 가족 중에 정신질환을 앓은 병력이 있거나 직장에서 정신질환이 업무에 방해를 준다고 느낄수록 정신치료를 받기 위한 탐색 작업을 갖는 것을 알 수 있음

익명성 보장 여부와 care option여부에서 Yes인 경우에도 treatment가 Yes가 나오는 걸 보면, 표면적으론 익명성 보장이 이루어지나 직원들이 이에 대해 신뢰하고 있지 않을 수 있다고 추측할 수 있음

또한 직장 내에서 이루어지는 care option이 별다른 효과가 없을 수도 있다는 점 역시 추측해 볼 수 있음

Thank you

라이언 조