What factors will affect your grades?

2014024007 김민기
2014024028 이준호
2016111719 전혜린

Content

01

데이터 소개

Content

데이터 소개

포르투갈 중등학교의 수학 및 언어 강좌를 대상으로 한 설문조사에서 얻은 데이터로, 학생들에 대한 많은 흥미로운 사회적, 성별, 그리고 공부에 대한 정보를 포함한다.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	G3	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	fai
2	6	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	
3	6	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	
4	10	GP	F													
5	15	GP	F													
6	10	GP	F													
7	15	GP	M	16	U	LE3	T									
8	11	GP	M	16	U	LE3	T									
9	6	GP	F	17	U	GT3	A									
10	19	GP	M	15	U	LE3	A									
11	15	GP	M	15	U	GT3	T									
12	9	GP	F	15	U	GT3	T									
13	12	GP	F	15	U	GT3	T	2	1	services	other	reputatior	father	3	3	
14	14	GP	M	15	U	LE3	T	4	4	health	services	course	father	1	1	
15	11	GP	M	15	U	GT3	T	4	3	teacher	other	course	mother	2	2	
16	16	GP	M	15	U	GT3	A	2	2	other	other	home	other	1	3	

성적

상위 30%(A학점) 안에 드는 성적은 1
그 외 0을 부여하여
이항형 변수로 변환

데이터 출처 : https://www.kaggle.com/

Content

데이터 소개

변수명	변수뜻
Famsize	가족 구성원 수 / LE3 (3 이하) I GT3 (3 초과)
Studytime	한 주의 공부 시간 / 1 (~15분) ~ 4 (1시간 이상)
Failures	유급 횟수 / 1,2,3,4
Romantic	솔로 or 커플 / yes or no
Dalc	주중 알코올 섭취량 / 1 (very low) ~5 (very high)
Walc	주말 알코올 섭취량 / 1(very low) ~ 5 (very high)
Higher	더 높은 수준의 교육을 원하는지 / yes or no
Famrel	가족간의 화목 관계 / 1 (very bad) ~ 5 (excellent)
Pstatus	부모님과 함께 사는지 / T (living together) or A (living apart)

Content

데이터 소개

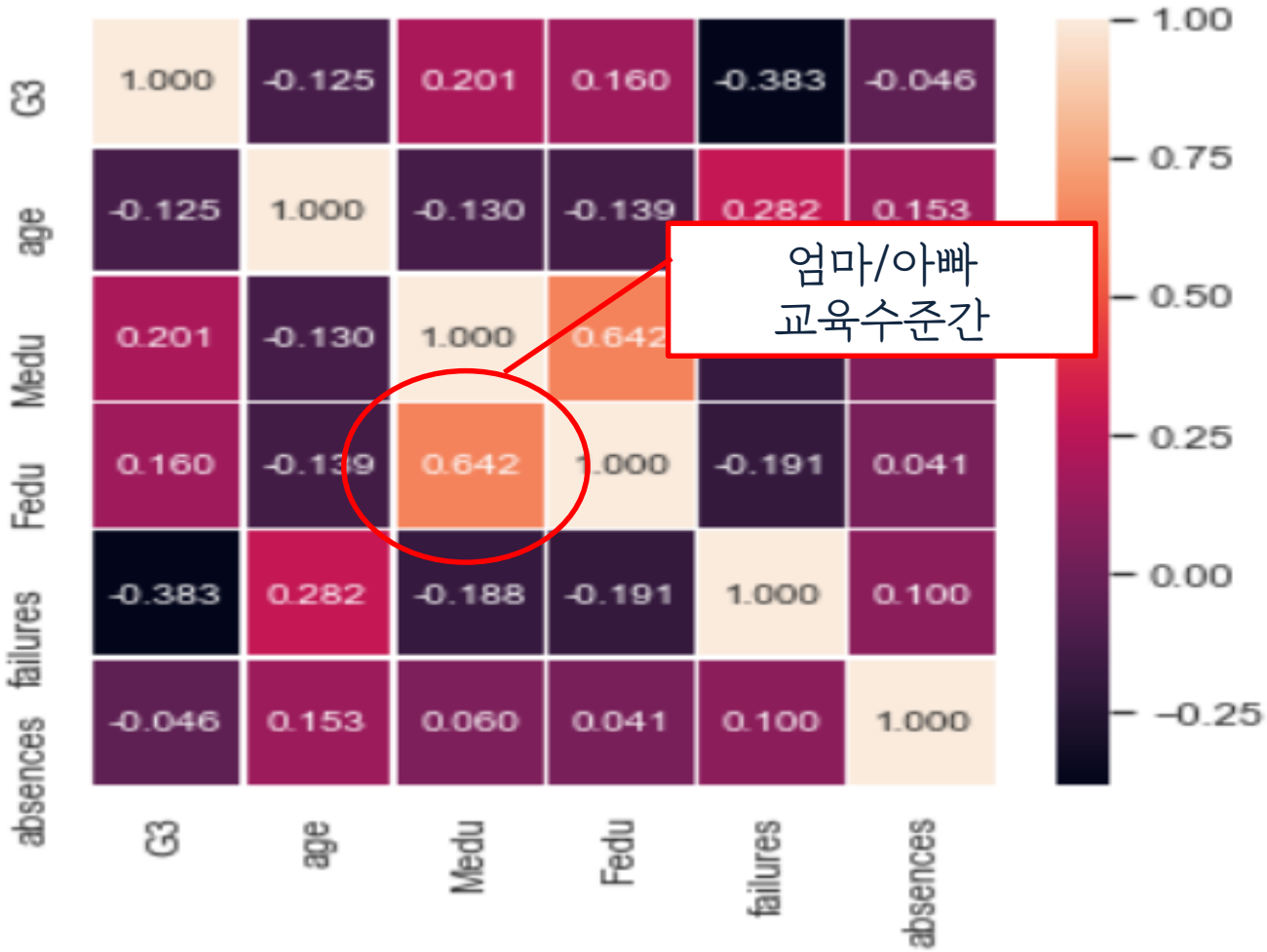
기초통계량

```
> summary(data)
       G3          school   sex         age          address famsize   Pstatus      Medu            Fedu            Mjob    
 Min.   : 0.00   GP:772   F:591   Min.   :15.00   R:285   GT3:738   A:121   Min.   :0.000   Min.   :0.000   at_home :194  
 1st Qu.:10.00   MS:272   M:453   1st Qu.:16.00   U:759   LE3:306   T:923   1st Qu.:2.000   1st Qu.:1.000   health  : 82  
 Median :11.00                   Median :17.00                             Median :3.000   Median :2.000   other   :399  
 Mean   :11.34                   Mean   :16.73                             Mean   :2.603   Mean   :2.388   services:239  
 3rd Qu.:14.00                   3rd Qu.:18.00                             3rd Qu.:4.000   3rd Qu.:3.000   teacher :130  
 Max.   :20.00                   Max.   :22.00                             Max.   :4.000   Max.   :4.000                 
      Fjob           reason      guardian    traveltime      studytime        failures      schoolsup famsup     paid     activities
 at_home : 62   course     :430   father:243   Min.   :1.000   Min.   :1.00   Min.   :0.0000   no :925   no :404   no :824   no :528   
 health  : 41   home       :258   mother:728   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:0.0000   yes:119   yes:640   yes:220   yes:516   
 other   :584   other      :108   other : 73   Median :1.000   Median :2.00   Median :0.0000                                         
 services:292   reputation:248                Mean   :1.523   Mean   :1.97   Mean   :0.2644                                         
 teacher : 65                                 3rd Qu.:2.000   3rd Qu.:2.00   3rd Qu.:0.0000                                         
                                              Max.   :4.000   Max.   :4.00   Max.   :3.0000                                         
 nursery   higher    internet  romantic      famrel         freetime        goout           Dalc            Walc      
 no :209   no : 89   no :217   no :673   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 yes:835   yes:955   yes:827   yes:371   1st Qu.:4.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000  
                                       Median :4.000   Median :3.000   Median :3.000   Median :1.000   Median :2.000  
                                       Mean   :3.936   Mean   :3.201   Mean   :3.156   Mean   :1.494   Mean   :2.284  
                                       3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000  
                                       Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000  
     health         absences     
 Min.   :1.000   Min.   : 0.000  
 1st Qu.:3.000   1st Qu.: 0.000  
 Median :4.000   Median : 2.000  
 Mean   :3.543   Mean   : 4.435  
 3rd Qu.:5.000   3rd Qu.: 6.000  
 Max.   :5.000   Max.   :75.000
```

Content

데이터 소개

상관관계



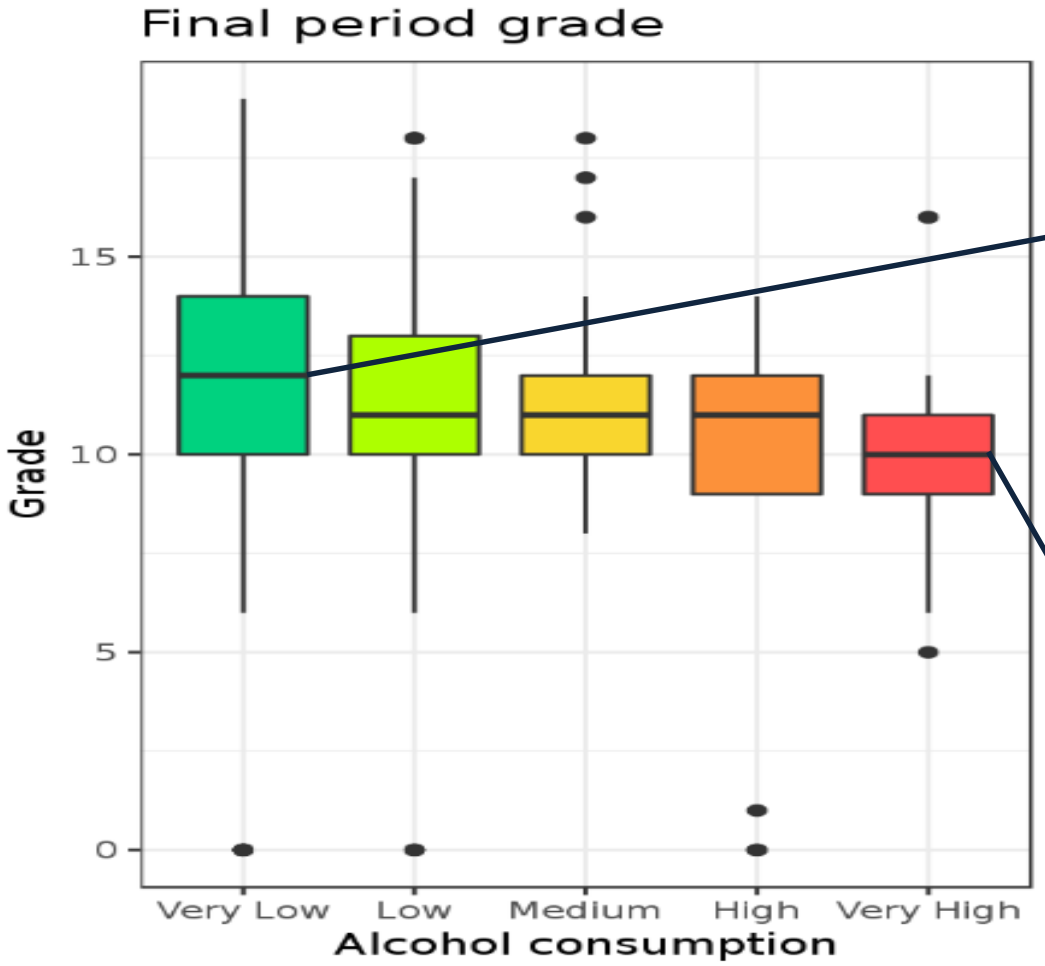
02

분석 목표

Content

분석 목표

주중 알코올 섭취량 ~ 성적



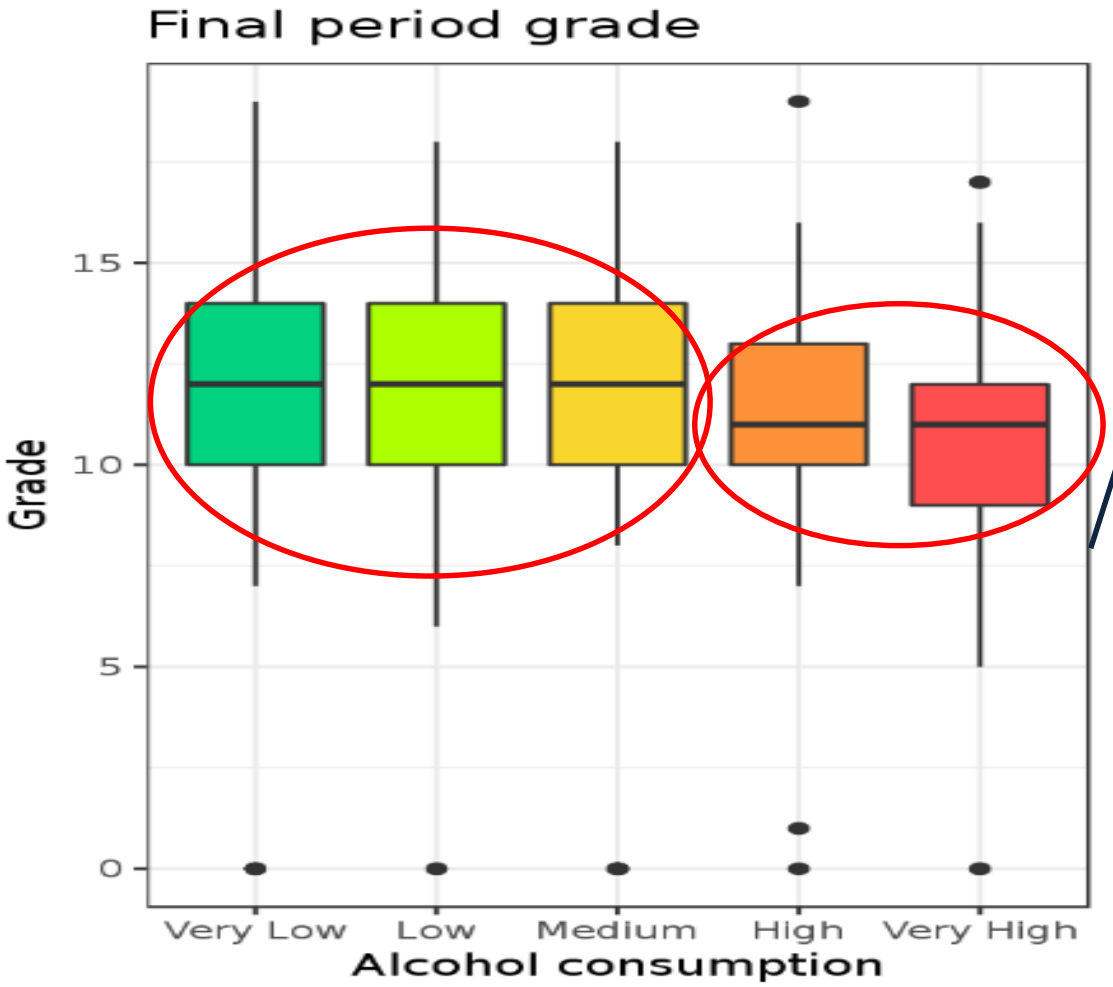
평균 성적이
눈에 띄게 높다.

평균 성적이
눈에 띄게 낮다.

Content

분석 목표

주말 알코올 섭취량 ~ 성적

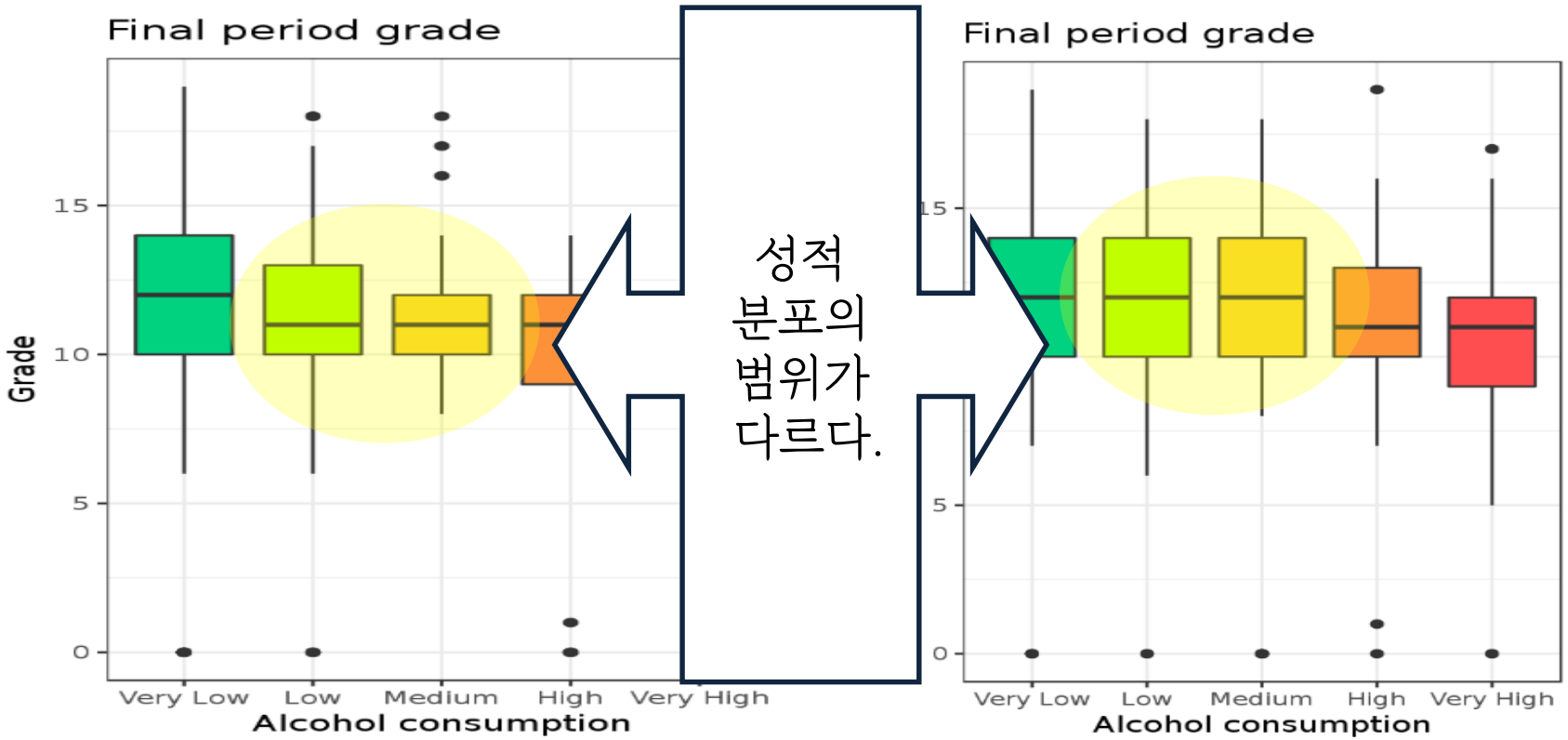


두 그룹의 평균 성적 차이가
두드러지게 나는 것을
알 수 있다.

Content

분석 목표

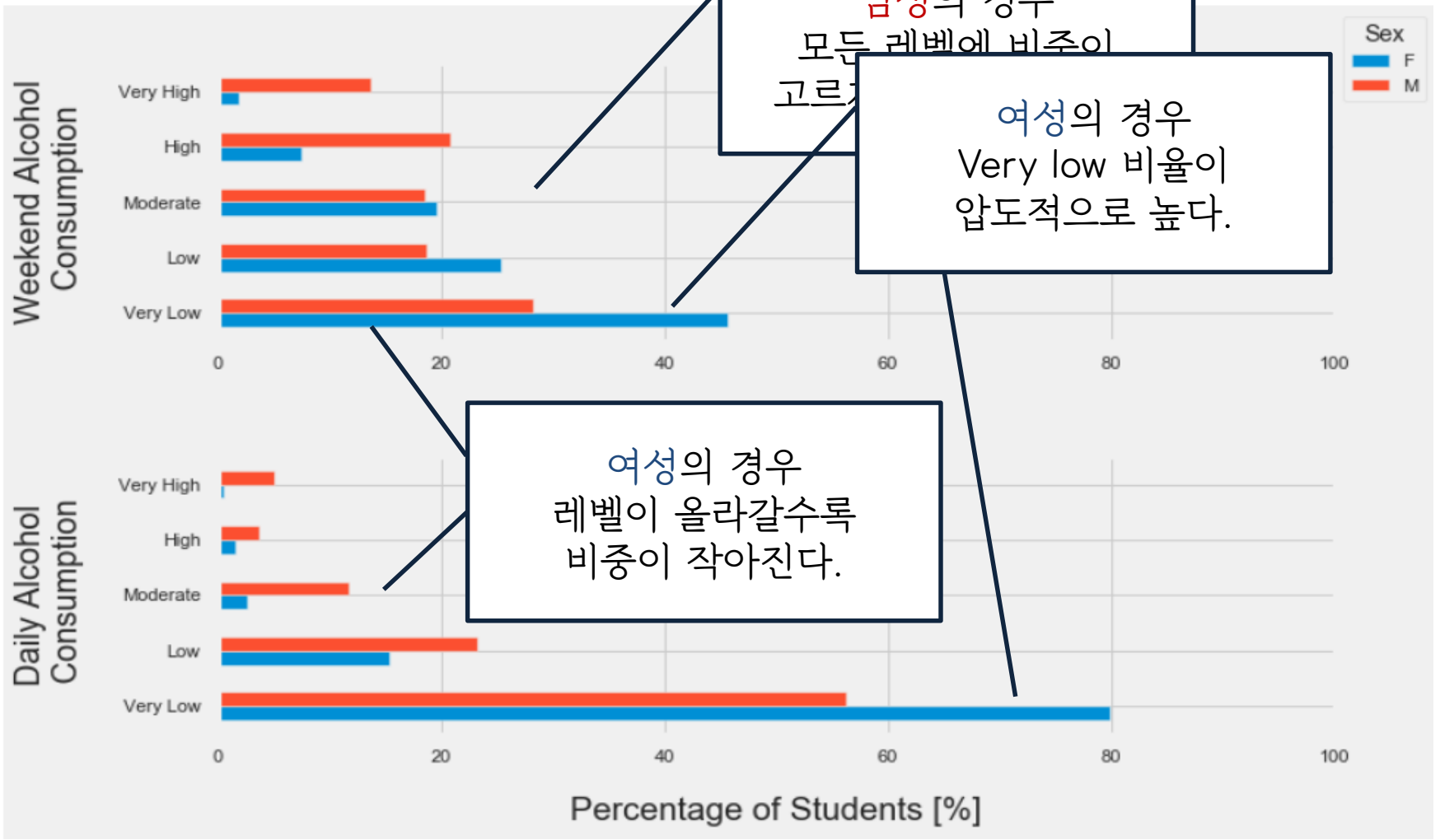
알코올 섭취량 ~ 성적



Content

분석 목표

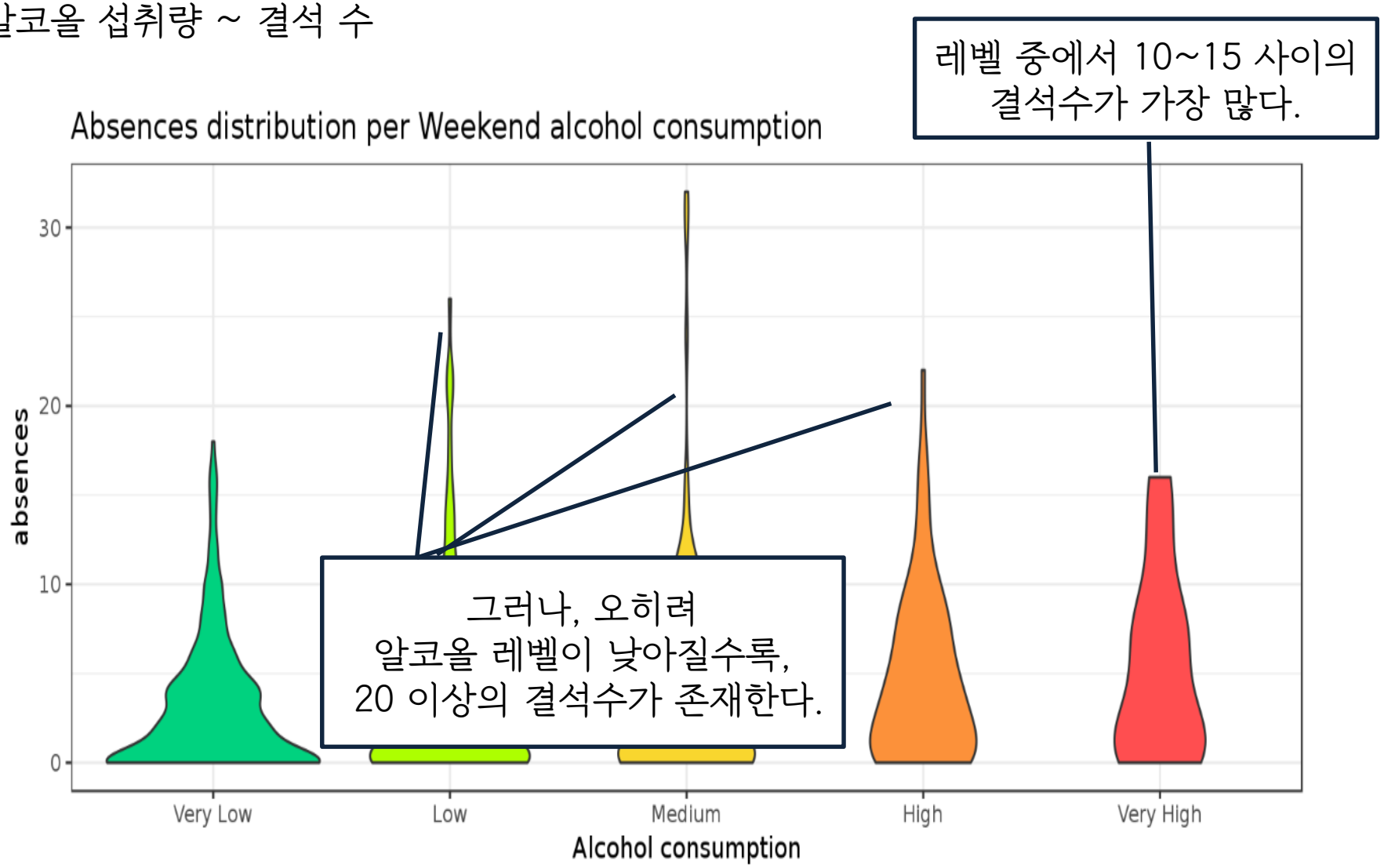
성별 ~ 알코올 섭취량



Content

분석 목표

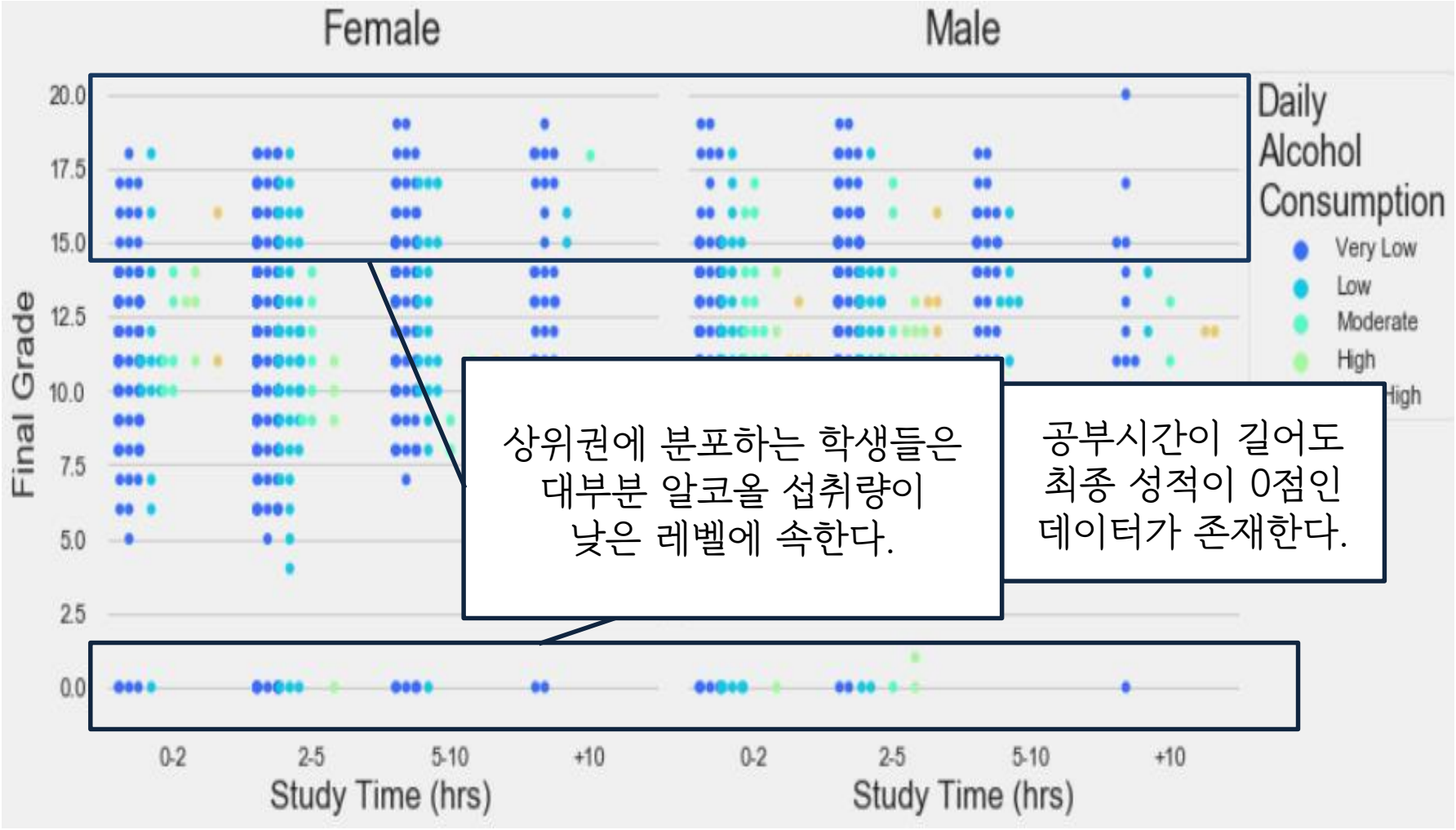
알코올 섭취량 ~ 결석 수



Content

분석 목표

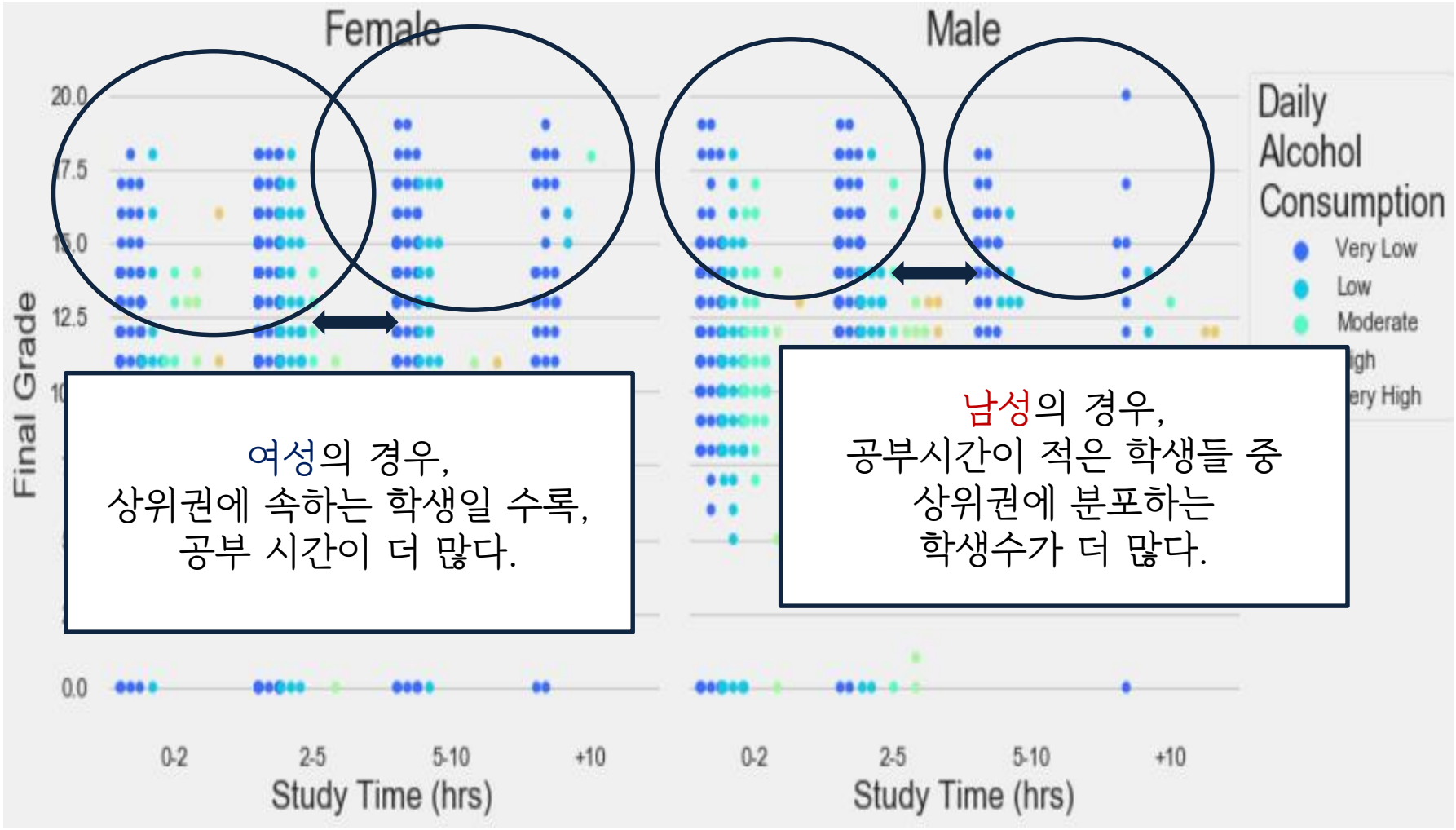
공부 시간 ~ 최종 성적 ~ 알코올 섭취량



Content

분석 목표

공부 시간 ~ 최종 성적 ~ 알코올 섭취량



Content

분석 목표

여러 분석 기법들을 이용하여
가장 좋은 분석 기법을 선정하여
성적 A등급에 영향을 미치는 변수를 알아보고
예측률 구하기

03

데이터 분석

Content

데이터 분석

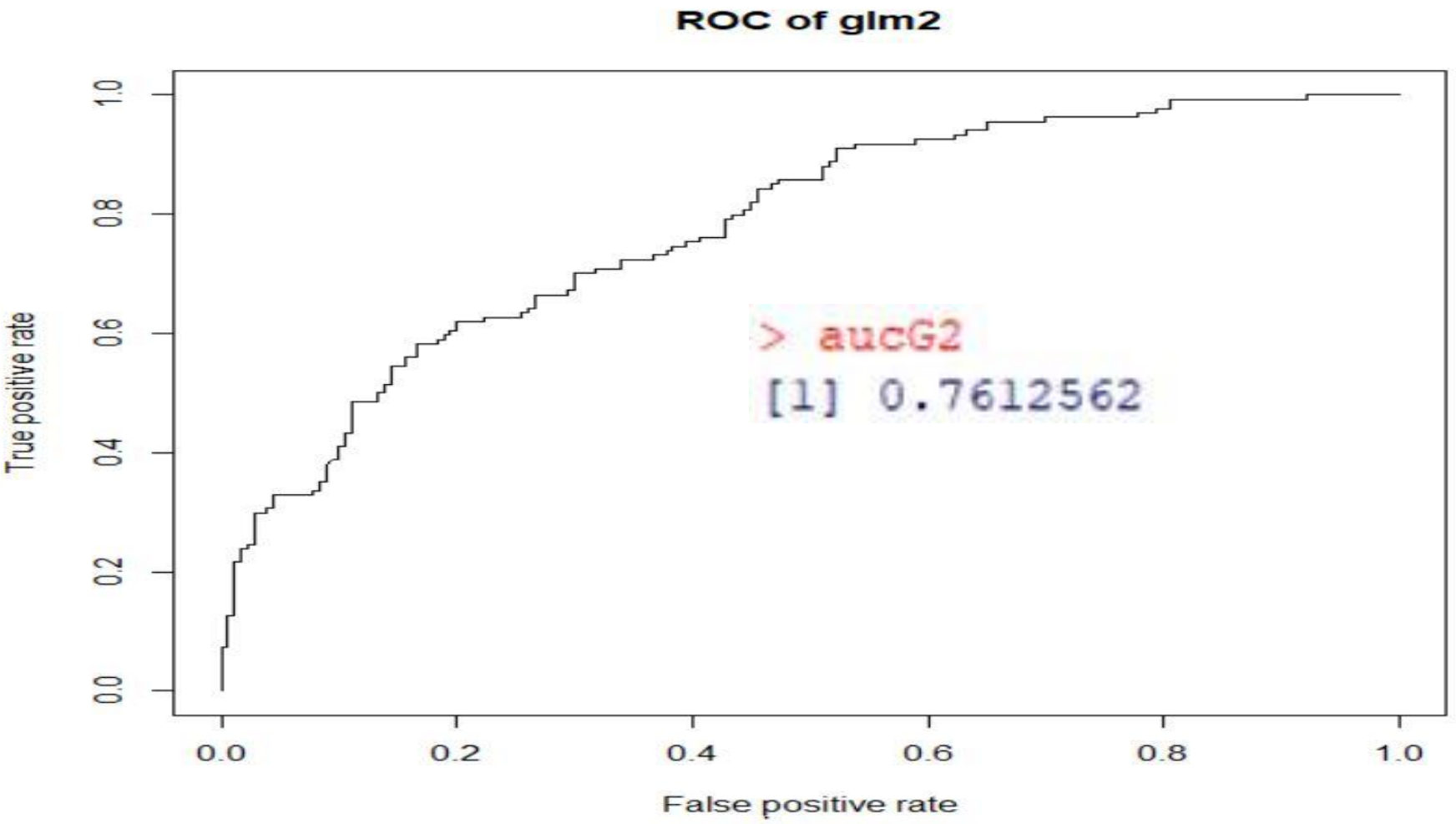
GLM

```
Call:
glm(formula = G3 ~ school + age + address + Fedu + Fjob + studytime +
    failures + schoolsup + paid + higher + internet + romantic +
    goout + health + absences, family = binomial(link = "logit"),
    data = data[train, ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0529  -0.8736  -0.3455   0.8951   2.9379

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.62315    1.76756  -3.181 0.001466 **
schoolMS     -0.50149    0.24794  -2.023 0.043111 *
age           0.16318    0.08906   1.832 0.066933 .
addressU      0.63051    0.23224   2.715 0.006630 **
Fedu.L        1.64628    0.80132   2.054 0.039930 *
Fedu.Q       -0.91872    0.67826  -1.355 0.175570
Fedu.C        0.61535    0.43355   1.419 0.155803
Fedu^4        0.07777    0.23387   0.333 0.739471
Fjobhealth    0.45709    0.63931   0.715 0.474627
Fjobother     0.05508    0.42559   0.129 0.897030
Fjobservices -0.35426    0.43812  -0.809 0.418755
Fjobteacher   1.06083    0.56082   1.892 0.058548 .
studytime.L   0.69140    0.28603   2.417 0.015641 *
studytime.Q  -0.23108    0.24471  -0.944 0.345023
studytime.C  -0.38522    0.20184  -1.909 0.056316 .
failures     -1.33979    0.29873  -4.485 7.29e-06 ***
schoolsupyes -1.40090    0.31573  -4.437 9.12e-06 ***
paidyes      -0.78624    0.23508  -3.345 0.000824 ***
higheryes     2.59263    0.76142   3.405 0.000662 ***
internetyes   0.35779    0.24875   1.438 0.150336
romanticyes  -0.36010    0.19892  -1.810 0.070256 .
goout.L      -0.43988    0.29493  -1.491 0.135835
goout.Q      -0.05088    0.26663  -0.191 0.848664
goout.C       0.37783    0.22505   1.679 0.093177 .
goout^4      -0.09477    0.18017  -0.526 0.598891
health.L     -0.73566    0.21814  -3.372 0.000745 ***
health.Q      0.39908    0.21733   1.836 0.066322 .
health.C     -0.30763    0.23424  -1.313 0.189070
health^4     -0.46303    0.22509  -2.057 0.039673 *
absences     -0.05423    0.02017  -2.688 0.007180 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
    Null deviance: 965.11  on 729  degrees of freedom
Residual deviance: 736.04  on 700  degrees of freedom
AIC: 796.04

Number of Fisher Scoring iterations: 6
```

데이터 분석

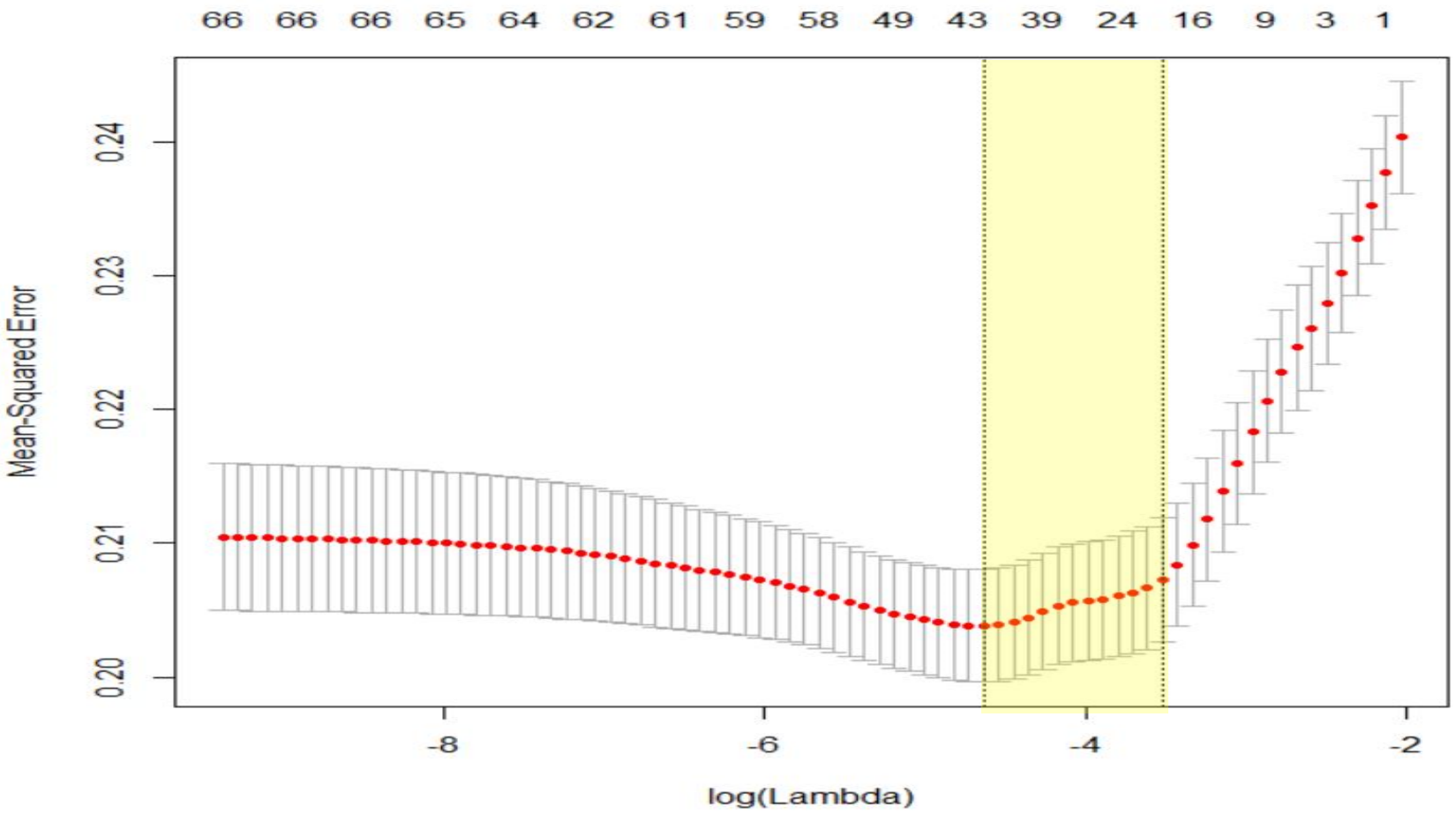
GLM2 ROC Curve



Content

데이터 분석

랏쏘(lambda 값에 따른 MSE)



데이터 분석

랏쏘(분석 결과)

```
> cv.lasso$lambda.min
[1] 0.009718963
```

⇨ MSE 최소일 때 Lasso 값

```
> cv.lasso$lambda.1se
[1] 0.02968031
```

⇨ MSE 최소일 때 1 표준편차 내의 좋은 Lasso 값

Content

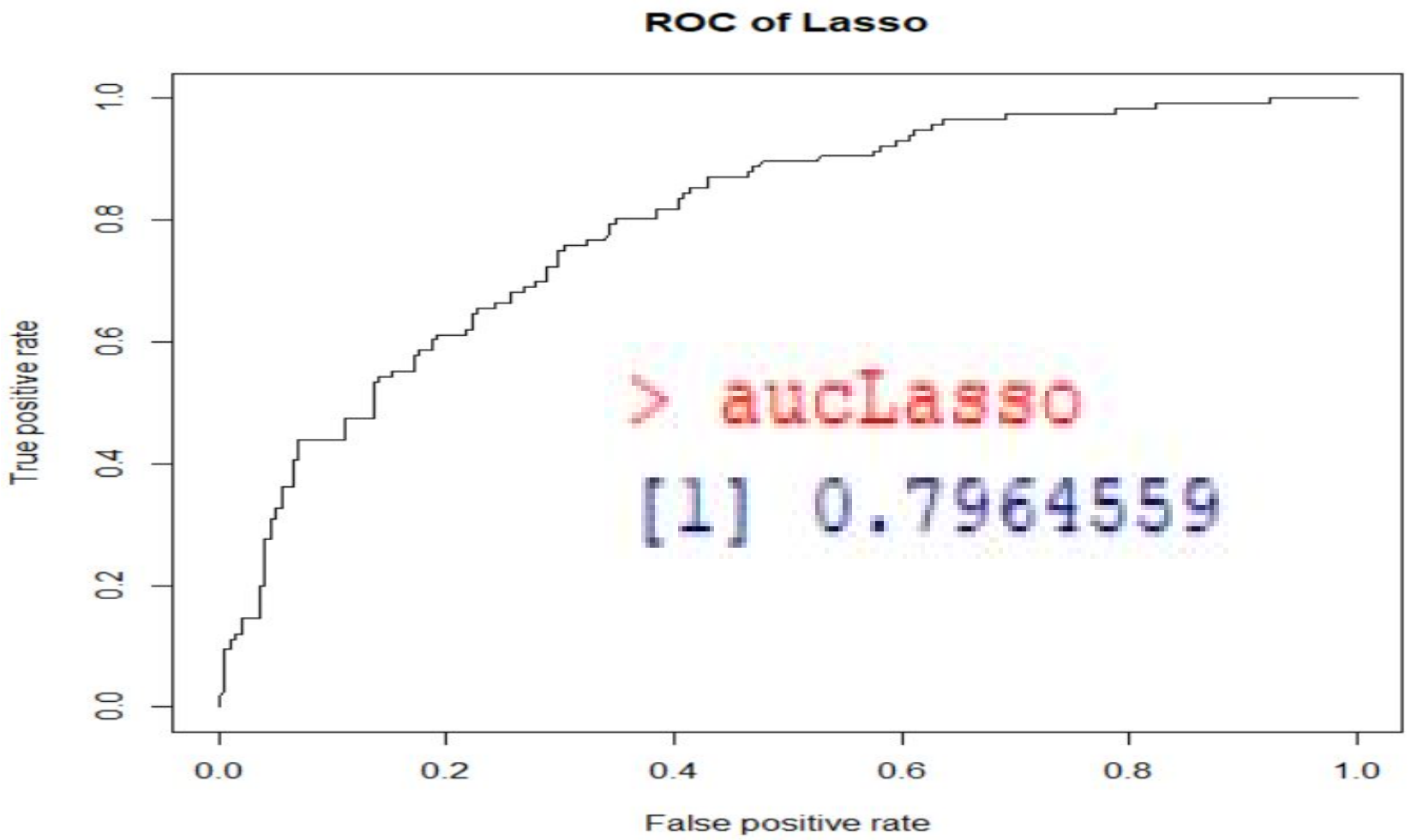
데이터 분석

랏쏘(계수)

```
> small.lambda.betas
      schoolMS          sexM       addressU       PstatusT
-0.0611695834 -0.0207068068  0.0739111003 -0.0023005669
        Medu.L        Medu.Q        Fedu.C        Fedu^4
 0.0297139734  0.0822434498  0.0190047763  0.0142015920
   Mjobhealth  Mjobservices  Fjobservices   Fjobteacher
 0.0903625642  0.0190964231 -0.0325048412  0.0483597338
    reasonhome   reasonother reasonreputation guardianmother
 0.0221524467 -0.0220897530  0.0201779810 -0.0077494592
guardianother  traveltime.L   studytime.L   studytime.Q
-0.0201449545 -0.0009600448  0.0682399347 -0.0484830698
  studytime.C      failures   schoolsupyes      famsupyes
-0.0558761687 -0.0730386531 -0.2201363074 -0.0401174177
       paidyes     higheryes    internetyes    romanticyes
-0.1178130628  0.2246767217  0.0827362933 -0.0375434425
      famrel.L      famrel^4     freetime.C     freetime^4
 0.0554380203 -0.0259741273  0.0447481662 -0.0372353142
       goout.L       goout.C        Dalc.L        Dalc.Q
-0.0318928843  0.0152445400 -0.0313724761  0.0304539114
        Walc.L        Walc.C      health.L      health.Q
-0.0353537972  0.0154275116 -0.0887177423  0.0333212957
      health.C      health^4
-0.0587615141 -0.0578282659
```

데이터 분석

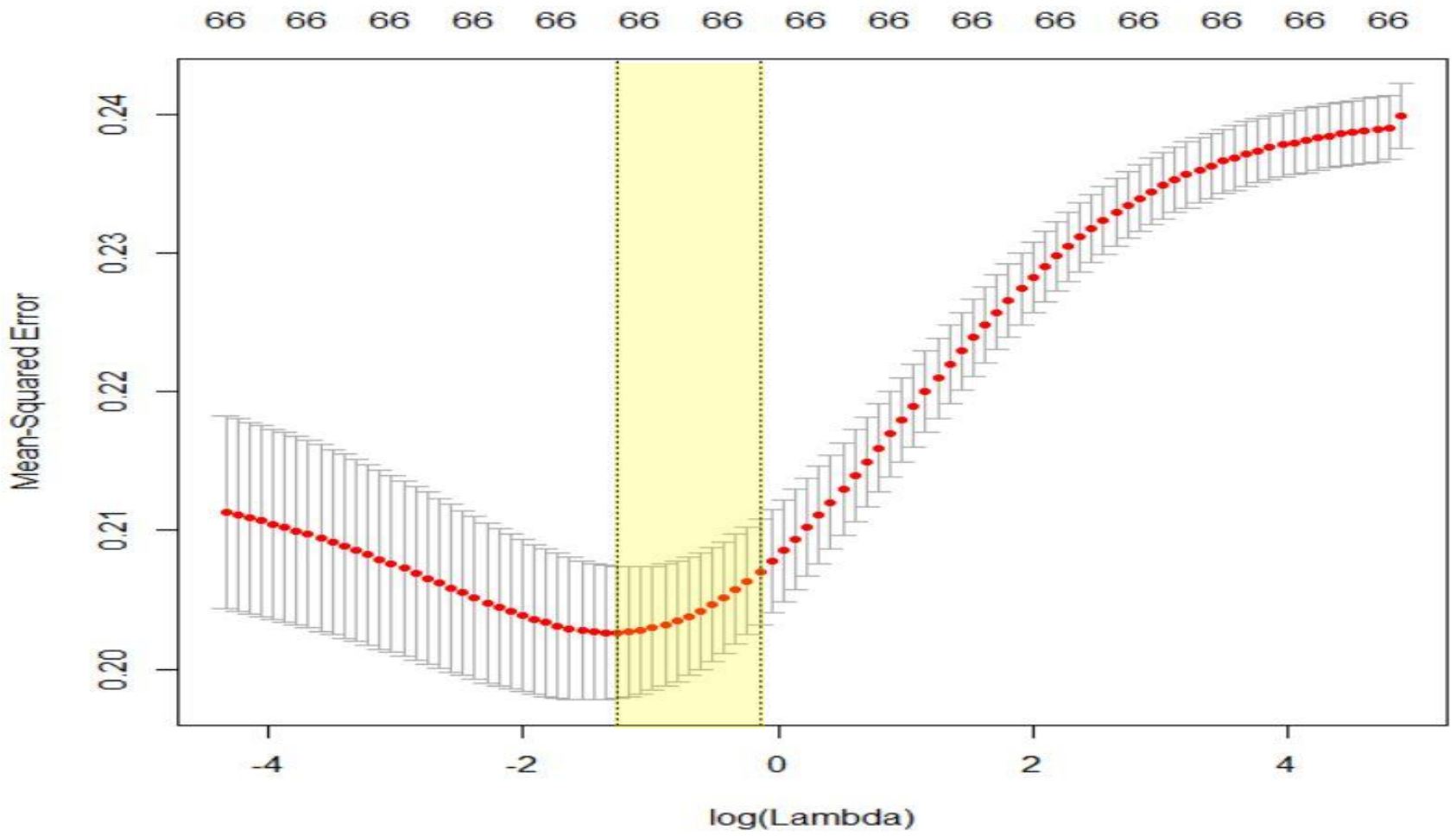
랏쏘(ROC Curve)



Content

데이터 분석

릿지(lambda 값에 따른 MSE)



데이터 분석

릿지(분석 결과)

```
> cv.ridge$lambda.min
[1] 0.2833129
```

MSE 최소일 때
Lasso 값

```
> cv.ridge$lambda.lse
[1] 0.8651968
```

MSE 최소일 때
1 표준편차 내의 좋은
Lasso 값

Content

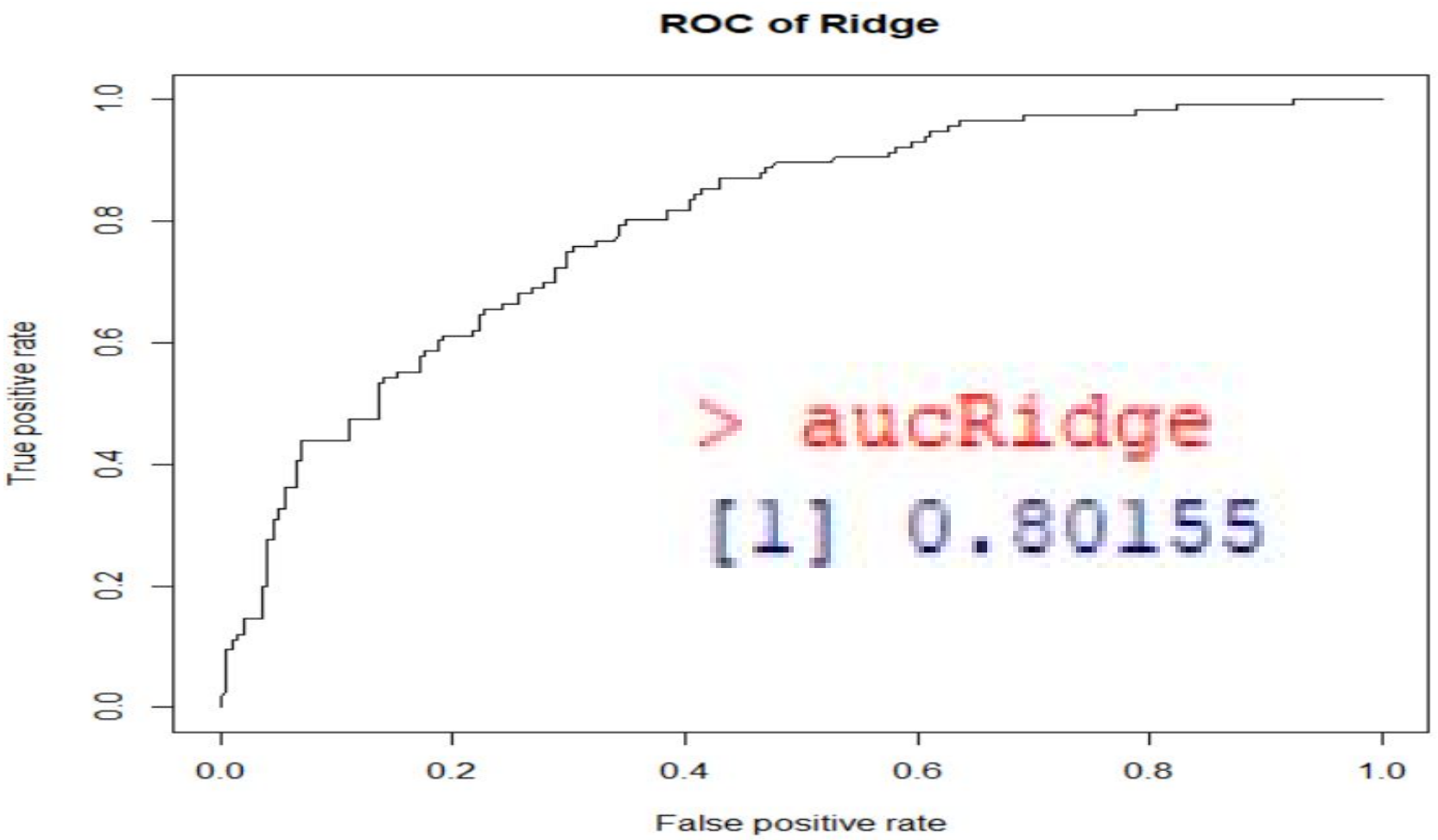
데이터 분석

릿지(계수)

```
> small.lambda.betas2
          schoolMS              sexM               age           addressU
-0.049031282 -0.024328565  0.004141615  0.060267569
   famsizeLE3      PstatusT       Medu.L       Medu.Q
 0.010742559 -0.013448871  0.054317968  0.063390250
       Medu.C       Medu^4       Fedu.L       Fedu.Q
-0.005840991 -0.009421968  0.019997452  0.001201631
       Fedu.C       Fedu^4   Mjobhealth    Mjobother
 0.026059440  0.021903744  0.083123967  0.006820016
  Mjobservices   Mjobteacher   Fjobhealth    Fjobother
 0.024625099  0.000458634  0.017011511  0.003318210
  Fjobservices   Fjobteacher   reasonhome  reasonother
-0.030318466  0.054947893  0.029232633 -0.036467898
reasonreputation guardianmother guardianother traveltime.L
 0.030197694 -0.021541250 -0.060393865 -0.029829837
  traveltime.Q   traveltime.C   studytime.L   studytime.Q
-0.013471499 -0.003478155  0.063152308 -0.040810156
   studytime.C       failures   schoolsupyes    famsupyes
-0.047092260 -0.054074914 -0.158731326 -0.041365319
      paidyes  activitiesyes   nurseryyes    higheryes
-0.084942833 -0.005570567  0.011668411  0.162233793
   internetyes   romanticyes       famrel.L       famrel.Q
 0.072485474 -0.035207913  0.056496168 -0.005284976
       famrel.C       famrel^4   freetime.L   freetime.Q
 0.010070378 -0.035245288 -0.007945014  0.002812037
   freetime.C   freetime^4       goout.L       goout.Q
 0.037491744 -0.035354046 -0.030636161 -0.009116357
       goout.C       goout^4       Dalc.L       Dalc.Q
 0.023121931 -0.004733521 -0.050091207  0.033681536
       Dalc.C       Dalc^4       Walc.L       Walc.Q
-0.008112868 -0.003852408 -0.032343912  0.010779218
       Walc.C       Walc^4       health.L     health.Q
 0.033042832  0.013175142 -0.068509187  0.029783626
     health.C     health^4
-0.054856476 -0.052632635
```

데이터 분석

릿지(ROC Curve)



Content

데이터 분석

GAM

```
Call: gam(formula = G3 ~ ns(failures) + ns(age) + ns(absences) + .,
    data = data[train, ])
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-1.00080 -0.35166 -0.07142  0.37734  0.97141

(Dispersion Parameter for gaussian family taken to be 0.1912)

    Null Deviance: 173.289 on 729 degrees of freedom
Residual Deviance: 126.5852 on 662 degrees of freedom
AIC: 930.596

Number of Local Scoring Iterations: 2

Anova for Parametric Effects
                Df  Sum Sq Mean Sq F value    Pr(>F)
ns(failures)     1  11.308 11.3078 59.1361 5.342e-14 ***
ns(age)          1   0.008  0.0078  0.0408 0.8399285
ns(absences)     1   1.326  1.3261  6.9352 0.0086489 **
school           1   3.453  3.4532 18.0594 2.449e-05 ***
sex              1   0.755  0.7551  3.9489 0.0473144 *
address          1   1.213  1.2127  6.3420 0.0120261 *
famsize          1   0.100  0.0998  0.5217 0.4703747
Pstatus          1   0.250  0.2495  1.3049 0.2537262
Medu             4   4.563  1.1408  5.9659 0.0001016 ***
Fedu             4   0.471  0.1178  0.6162 0.6511406
Mjob             4   0.729  0.1823  0.9535 0.4325065
Fjob             4   0.581  0.1453  0.7598 0.5516707
reason           3   1.495  0.4983  2.6062 0.0508175 .
guardian         2   0.250  0.1248  0.6527 0.5209713
traveltime       3   1.022  0.3408  1.7822 0.1492253
studytime        3   1.315  0.4383  2.2922 0.0769581 .
schoolsup        1   3.882  3.8820 20.3017 7.819e-06 ***
famsup           1   0.466  0.4662  2.4380 0.1189056
paid             1   1.274  1.2737  6.6611 0.0100683 *
activities       1   0.010  0.0099  0.0516 0.8204433
nursery          1   0.189  0.1891  0.9889 0.3203688
higher           1   1.228  1.2283  6.4236 0.0114909 *
internet         1   0.499  0.4993  2.6114 0.1065761
romantic         1   0.812  0.8118  4.2452 0.0397513 *
famrel           4   0.498  0.1246  0.6515 0.6259421
freetime         4   0.474  0.1185  0.6199 0.6484297
goout            4   1.350  0.3375  1.7649 0.1341743
Dalc             4   0.881  0.2203  1.1522 0.3308726
Walc             4   2.103  0.5258  2.7497 0.0274233 *
health           4   4.198  1.0495  5.4886 0.0002374 ***
Residuals      662 126.585  0.1912
```

데이터 분석

GAM 2

```
> summary(gam2)

Call: gam(formula = G3 ~ ns(failures) + ns(absences) + school + sex +
    address + Medu + reason + studytime + schoolsup + paid +
    higher + famrel + Walc + health, data = data[train, ])
Deviance Residuals:
    Min      1Q  Median      3Q     Max
-0.8348 -0.3698 -0.1069  0.4161  1.1134

(Dispersion Parameter for gaussian family taken to be 0.1928)

    Null Deviance: 173.289 on 729 degrees of freedom
Residual Deviance: 134.7355 on 699 degrees of freedom
AIC: 902.1464

Number of Local Scoring Iterations: 2

Anova for Parametric Effects
               Df  Sum Sq Mean Sq F value    Pr(>F)
ns(failures)    1  11.308 11.3078 58.6641 6.247e-14 ***
ns(absences)    1   1.323  1.3235  6.8660  0.008976 **
school          1   3.326  3.3262 17.2564 3.670e-05 ***
sex             1   0.793  0.7931  4.1146  0.042894 *
address         1   1.218  1.2182  6.3198  0.012163 *
Medu            4   4.643  1.1609  6.0225 9.097e-05 ***
reason          3   1.687  0.5622  2.9167  0.033528 *
studytime       3   1.069  0.3562  1.8479  0.137112
schoolsup       1   3.915  3.9148 20.3100 7.719e-06 ***
paid            1   1.722  1.7218  8.9324  0.002900 **
higher          1   1.226  1.2263  6.3618  0.011882 *
famrel          4   0.533  0.1331  0.6907  0.598510
Walc            4   2.422  0.6055  3.1414  0.014166 *
health          4   3.369  0.8422  4.3691  0.001701 **
Residuals     699 134.735  0.1928
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

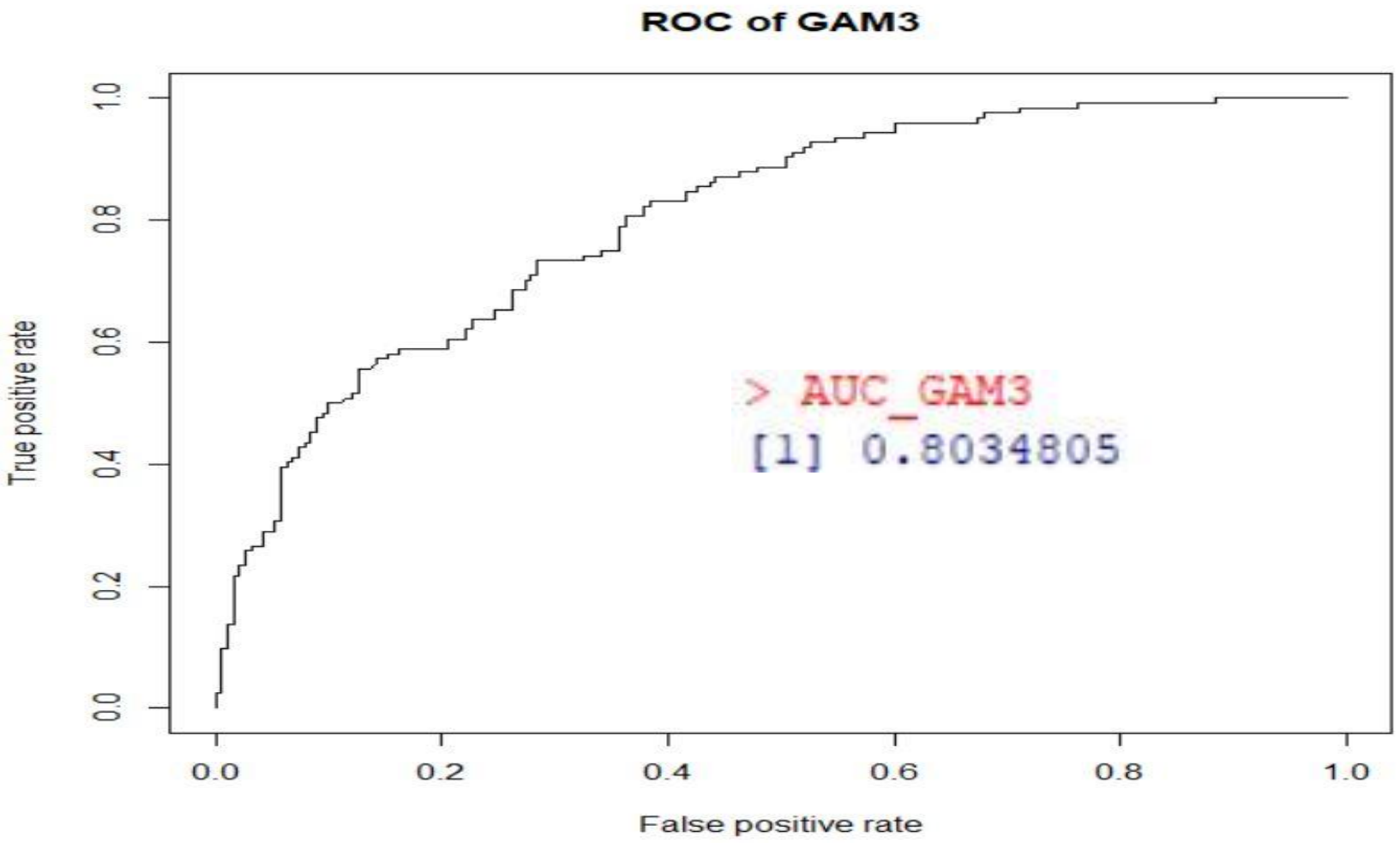
데이터 분석

GAM 3

```
> summary(gam3)

Call: gam(formula = G3 ~ ns(failures) + ns(absences) + school + sex +
    address + Medu + reason + schoolsup + paid + higher + Walc +
    health, data = data[train, ])
Deviance Residuals:
    Min      1Q  Median      3Q     Max
-0.8007 -0.3740 -0.1024  0.4274  1.1272

(Dispersion Parameter for gaussian family taken to be 0.1931)

    Null Deviance: 173.289 on 729 degrees of freedom
Residual Deviance: 136.3175 on 706 degrees of freedom
AIC: 896.6681

Number of Local Scoring Iterations: 2

Anova for Parametric Effects
               Df  Sum Sq Mean Sq F value    Pr(>F)
ns(failures)    1  11.308 11.3078 58.5640 6.474e-14 ***
ns(absences)    1   1.323  1.3235  6.8543  0.009032 **
school          1   3.326  3.3262 17.2269 3.722e-05 ***
sex             1   0.793  0.7931  4.1075  0.043068 *
address         1   1.218  1.2182  6.3090  0.012235 *
Medu            4   4.643  1.1609  6.0122 9.250e-05 ***
reason          3   1.687  0.5622  2.9117  0.033745 *
schoolsup       1   4.010  4.0098 20.7672 6.111e-06 ***
paid            1   1.516  1.5160  7.8516  0.005217 **
higher          1   1.401  1.4007  7.2543  0.007241 **
Walc            4   2.889  0.7221  3.7401  0.005079 **
health          4   2.858  0.7144  3.6999  0.005443 **
Residuals     706 136.318  0.1931
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

데이터 분석

GAM3 ROC Curve



Content

데이터 분석

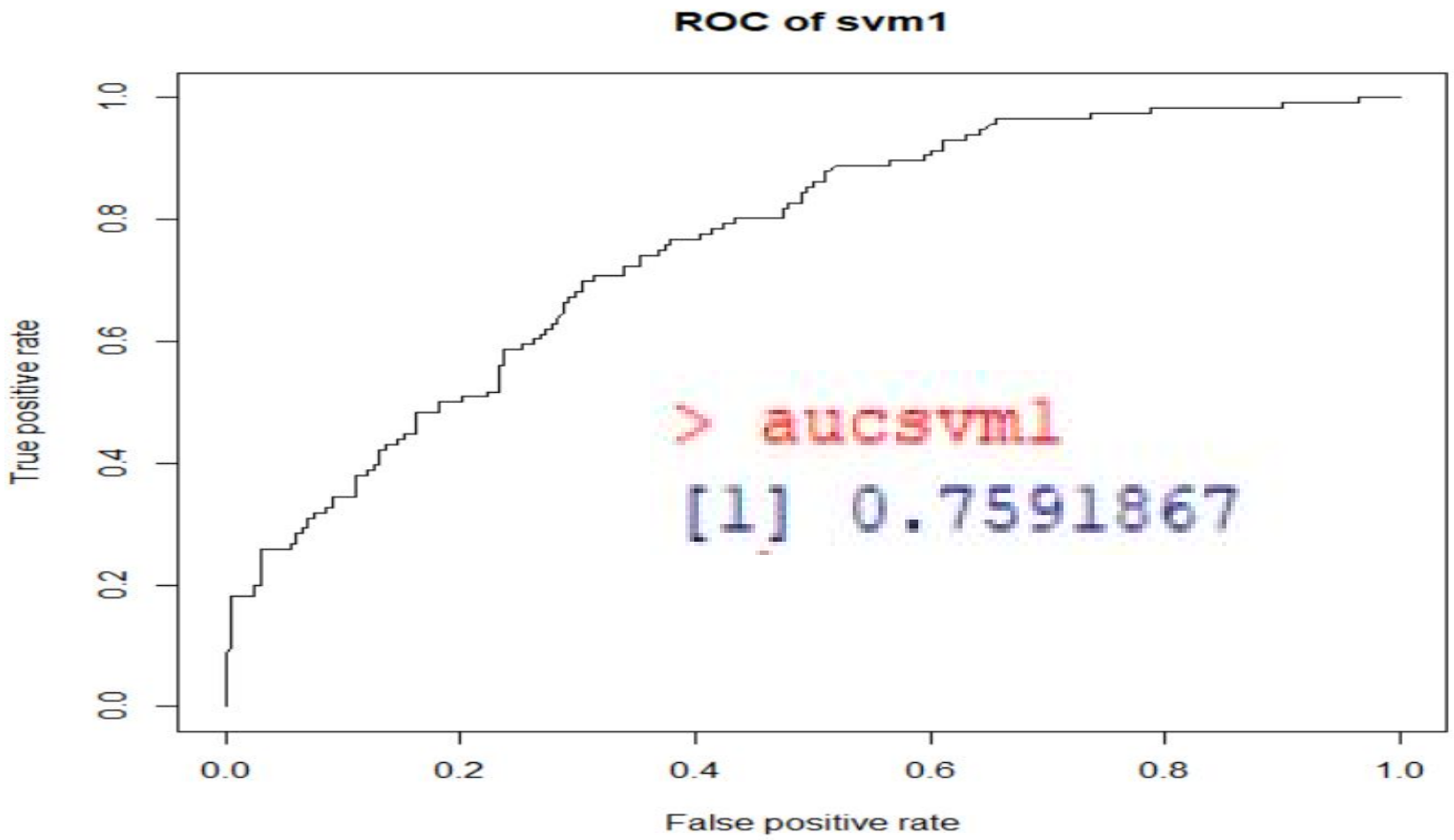
SVM (kernel=radial,linear)

```
> summary(svm_tune)

Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 cost gamma
    5   0.1

- best performance: 0.3246575

- Detailed performance results:
        cost gamma     error dispersion
1    0.00100   0.1 0.3876712 0.06956543
2    0.01001   0.1 0.3876712 0.06956543
3    0.01000   0.1 0.3876712 0.06956543
4    1.00000   0.1 0.3479452 0.06098925
5    5.00000   0.1 0.3246575 0.04834579
6   10.00000   0.1 0.3246575 0.04834579
7    0.00100   0.5 0.3876712 0.06956543
8    0.01001   0.5 0.3876712 0.06956543
9    0.01000   0.5 0.3876712 0.06956543
10   1.00000   0.5 0.3575342 0.06038802
11   5.00000   0.5 0.3589041 0.05981561
12  10.00000   0.5 0.3589041 0.05981561
13   0.00100   1.0 0.3876712 0.06956543
14   0.01001   1.0 0.3876712 0.06956543
15   0.01000   1.0 0.3876712 0.06956543
16   1.00000   1.0 0.3575342 0.06038802
17   5.00000   1.0 0.3575342 0.06038802
18  10.00000   1.0 0.3575342 0.06038802
19   0.00100   5.0 0.3876712 0.06956543
20   0.01001   5.0 0.3876712 0.06956543
21   0.01000   5.0 0.3876712 0.06956543
22   1.00000   5.0 0.3575342 0.06038802
23   5.00000   5.0 0.3575342 0.06038802
24  10.00000   5.0 0.3575342 0.06038802
25   0.00100  10.0 0.3876712 0.06956543
26   0.01001  10.0 0.3876712 0.06956543
27   0.01000  10.0 0.3876712 0.06956543
28   1.00000  10.0 0.3575342 0.06038802
29   5.00000  10.0 0.3575342 0.06038802
30  10.00000  10.0 0.3575342 0.06038802
```

```
> summary(svm_tune2)

Parameter tuning of 'svm':

- sampling method: 10-fold cross validati

- best parameters:
 cost gamma
    1   0.1

- best performance: 0.3342466

- Detailed performance results:
        cost gamma     error dispersion
1    0.00100   0.1 0.3876712 0.06956543
2    0.01001   0.1 0.3356164 0.06872108
3    0.01000   0.1 0.3356164 0.06872108
4    1.00000   0.1 0.3342466 0.05890234
5    5.00000   0.1 0.3369863 0.06098925
6   10.00000   0.1 0.3383562 0.05920246
7    0.00100   0.5 0.3876712 0.06956543
8    0.01001   0.5 0.3356164 0.06872108
9    0.01000   0.5 0.3356164 0.06872108
10   1.00000   0.5 0.3342466 0.05890234
11   5.00000   0.5 0.3369863 0.06098925
12  10.00000   0.5 0.3383562 0.05920246
13   0.00100   1.0 0.3876712 0.06956543
14   0.01001   1.0 0.3356164 0.06872108
15   0.01000   1.0 0.3356164 0.06872108
16   1.00000   1.0 0.3342466 0.05890234
17   5.00000   1.0 0.3369863 0.06098925
18  10.00000   1.0 0.3383562 0.05920246
19   0.00100   5.0 0.3876712 0.06956543
20   0.01001   5.0 0.3356164 0.06872108
21   0.01000   5.0 0.3356164 0.06872108
22   1.00000   5.0 0.3342466 0.05890234
23   5.00000   5.0 0.3369863 0.06098925
24  10.00000   5.0 0.3383562 0.05920246
25   0.00100  10.0 0.3876712 0.06956543
26   0.01001  10.0 0.3356164 0.06872108
27   0.01000  10.0 0.3356164 0.06872108
28   1.00000  10.0 0.3342466 0.05890234
29   5.00000  10.0 0.3369863 0.06098925
30  10.00000  10.0 0.3383562 0.05920246
```

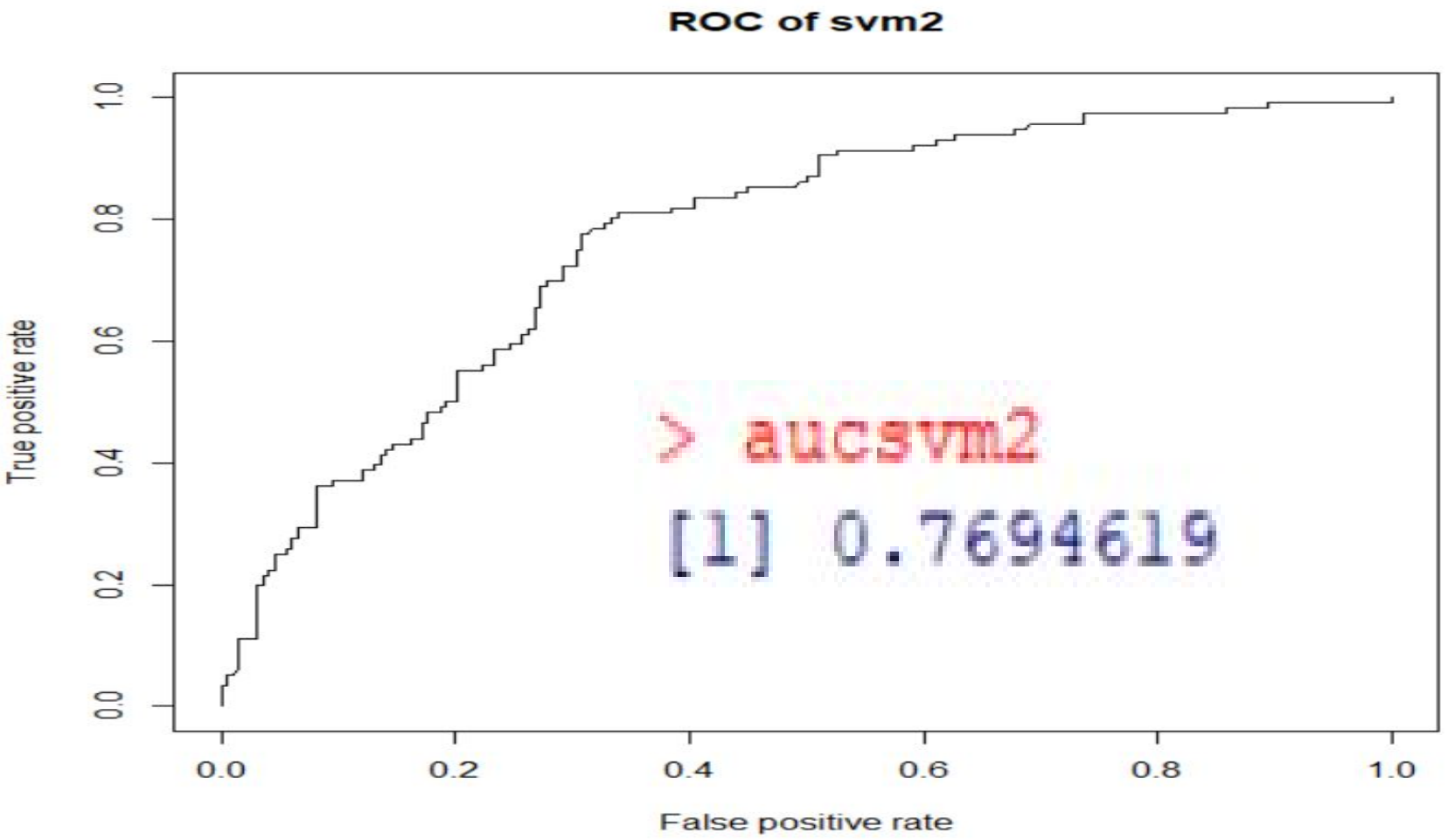
데이터 분석

SVM1(kernel=radial) ROC Curve



데이터 분석

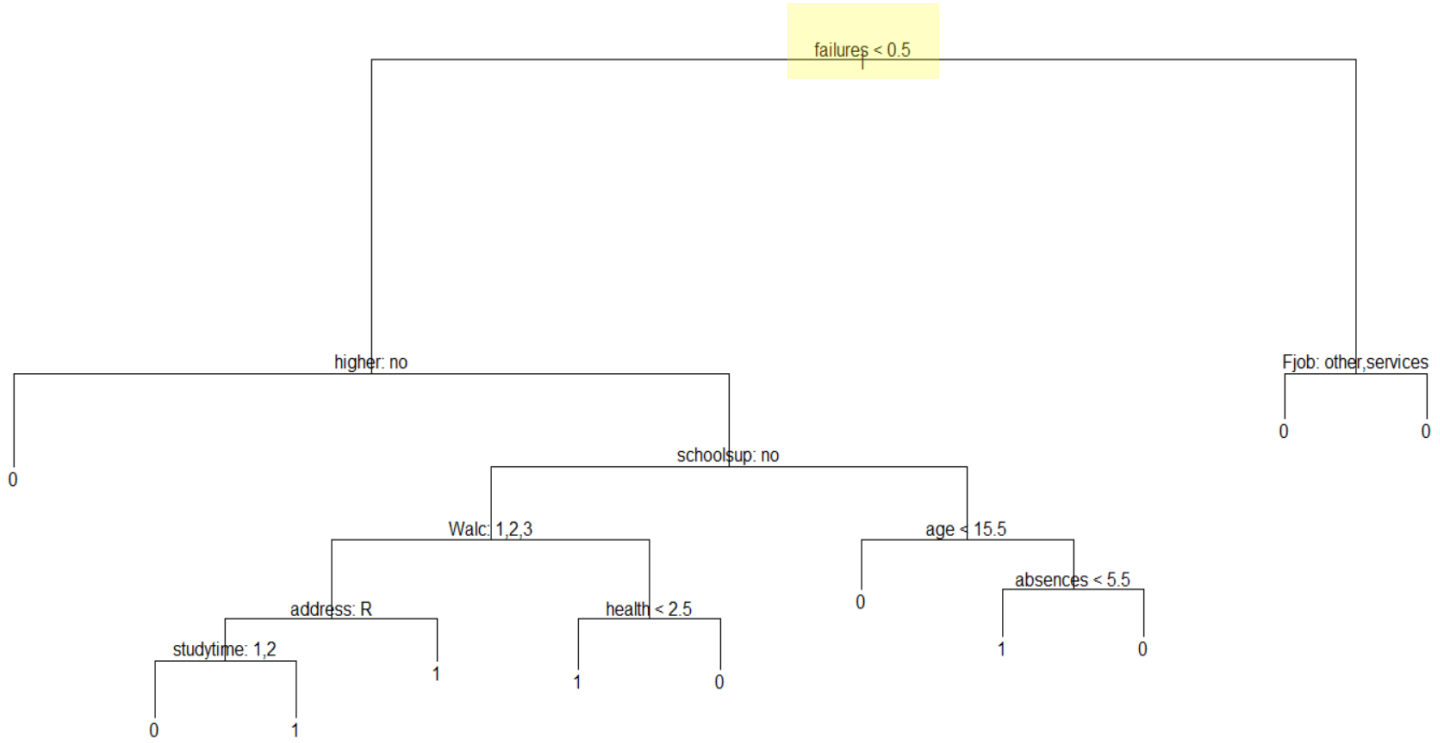
SVM2(kernel=linear) ROC Curve



Content

데이터 분석

Decision Tree (Pruning 전)

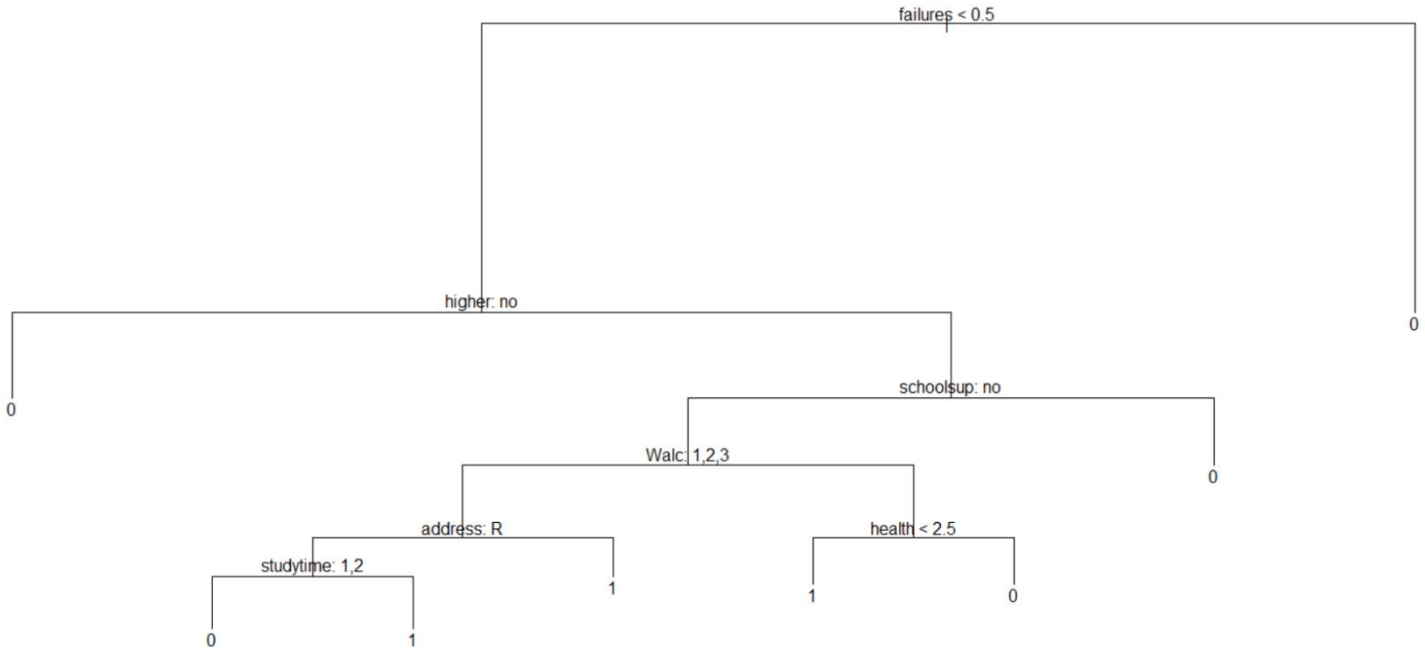



```
Classification tree:
tree(formula = G3 ~ ., data = weekday, subset = train)
Variables actually used in tree construction:
 [1] "failures"  "higher"    "schoolsup" "Walc"      "address"   "studytime" "health"    "age"       "absences"  "Fjob"
Number of terminal nodes:  11
Residual mean deviance:  1.071 = 772 / 721
Misclassification error rate: 0.2746 = 201 / 732
```

Content

데이터 분석

Decision Tree (Pruning 후)




```
Classification tree:
snip.tree(tree = tree.mod, nodes = c(3L, 11L))
Variables actually used in tree construction:
[1] "failures"  "higher"    "schoolsup" "Walc"      "address"   "studytime" "health"
Number of terminal nodes:  8
Residual mean deviance:  1.114 = 806.2 / 724
Misclassification error rate: 0.2773 = 203 / 732
```

데이터 분석

Decision Tree (Pruning 전/후 비교)

Pruning 전
Test data 적합 결과

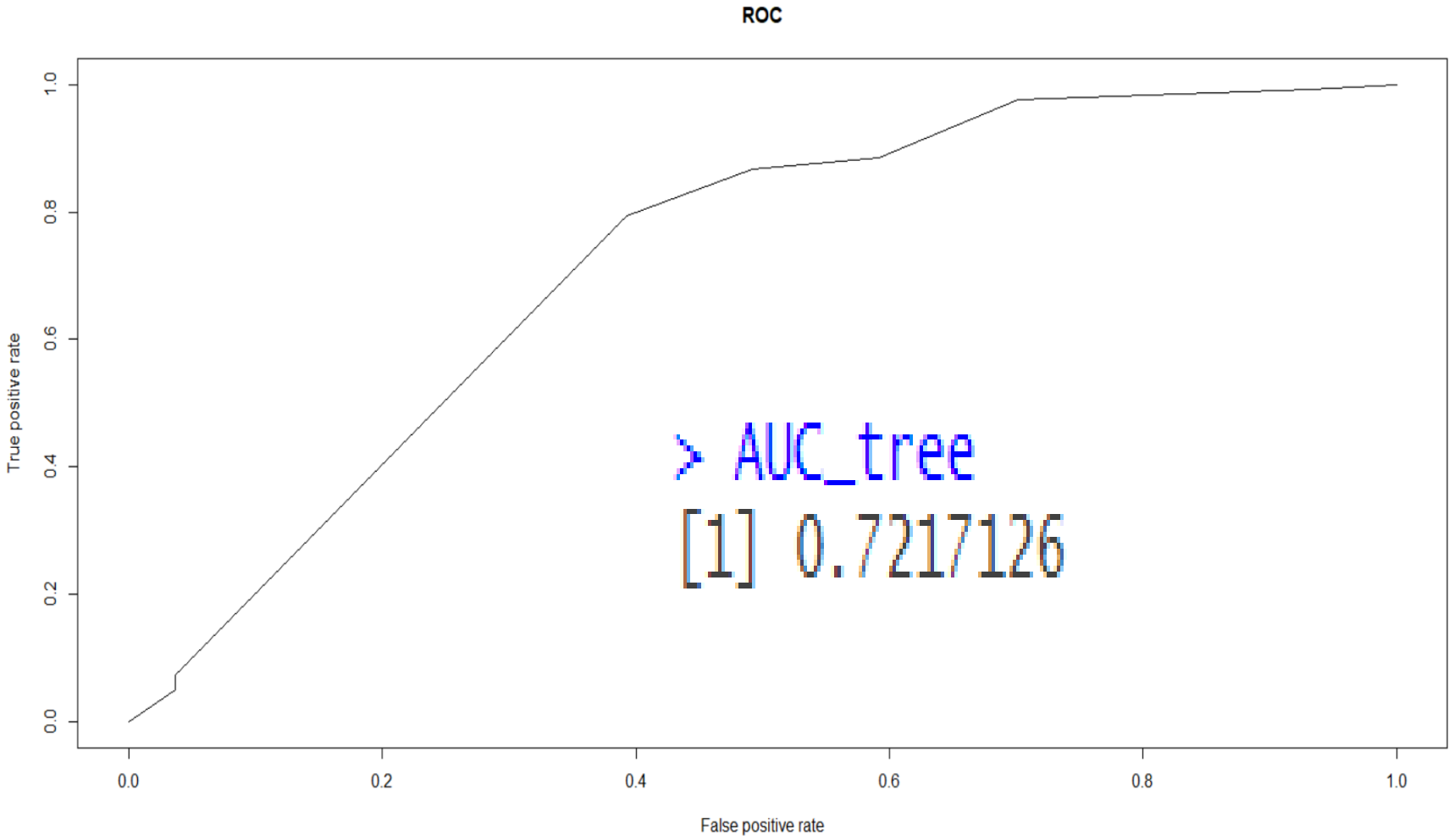
```
> table(tree.pred,y.test)
         y.test
tree.pred   0   1
        0 109  23
        1  82  98
> (23+82)/(23+82+109+98)
[1] 0.3365385
```

Pruning 후
Test data 적합 결과

```
> table(tree.pred1,y.test)
          y.test
tree.pred1   0   1
         0 116  25
         1  75  96
> (25+75)/(25+75+116+96)
[1] 0.3205128
```

데이터 분석

Decision Tree (ROC Curve)



데이터 분석

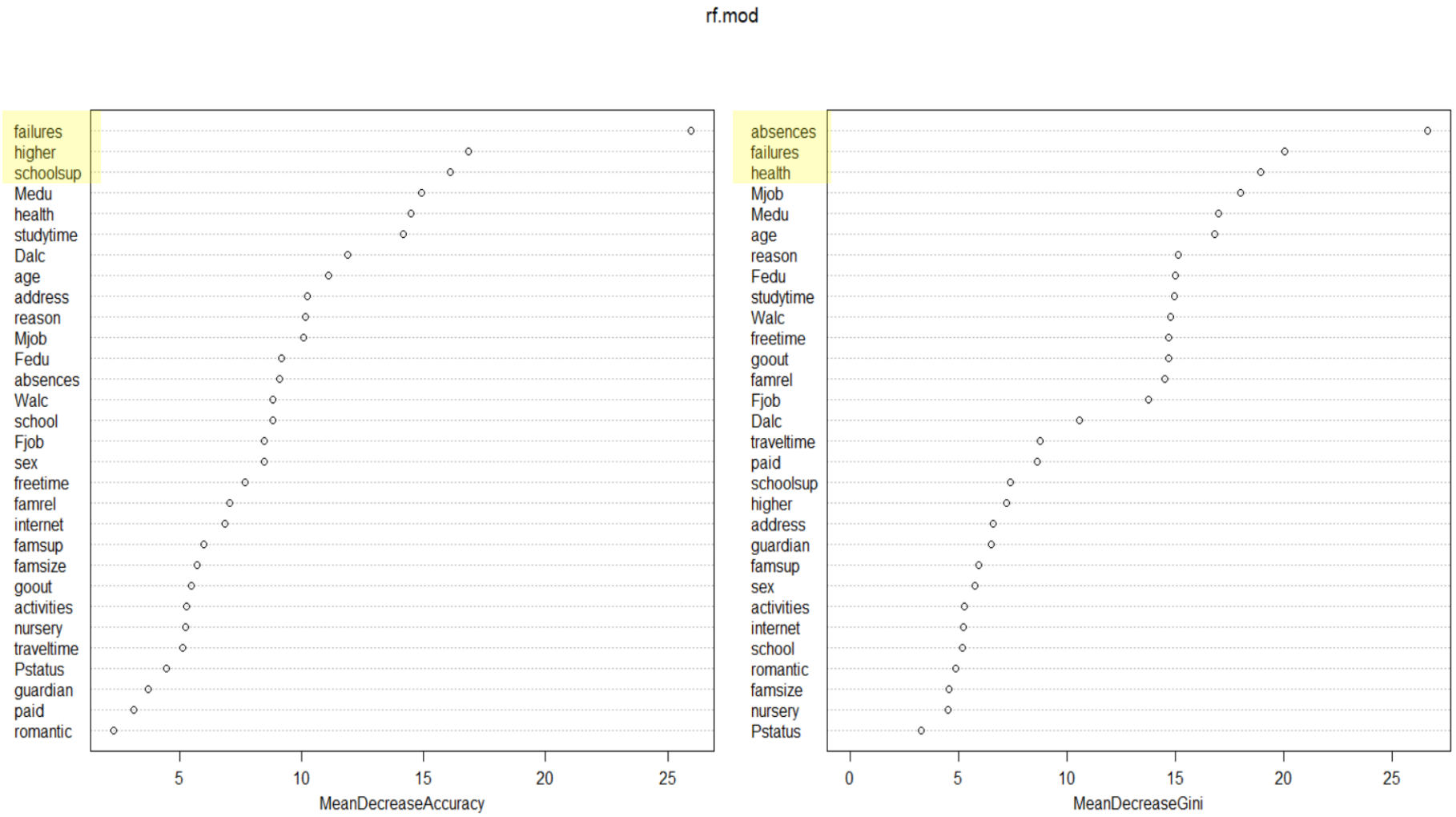
Random Forest

```
> #randomforest
> rf.mod<-randomForest(G3~.,data=weekday,subset=train,mtry=6,importance=T)
> rf.mod

Call:
 randomForest(formula = G3 ~ ., data = weekday, mtry = 6, importance = T,      subset = train)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 6

        OOB estimate of  error rate: 29.1%
Confusion matrix:
    0   1 class.error
0 355  92   0.2058166
1 121 164   0.4245614
```

Content

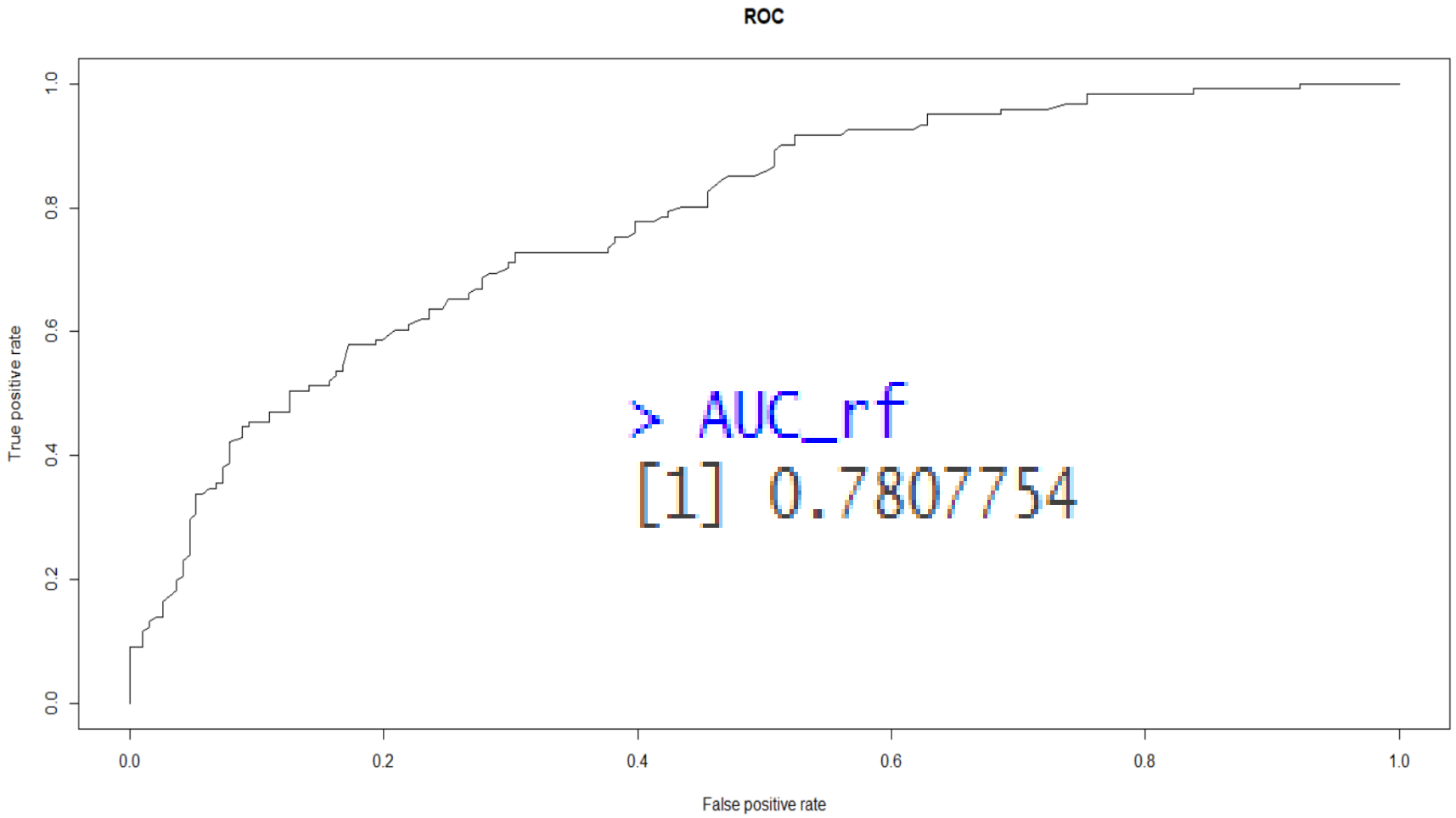
데이터 분석

Random Forest



데이터 분석

Random Forest (ROC Curve)



데이터 분석

Bagging

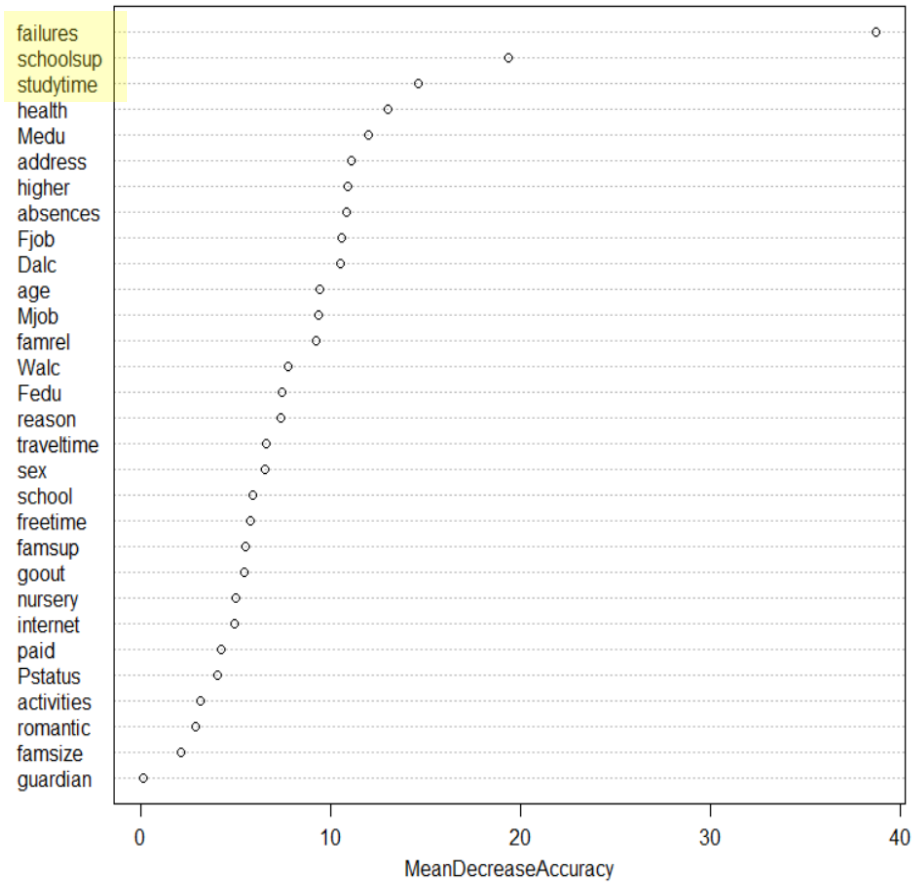
```
> #tree bagging
> bag.mod<-randomForest(G3~.,data=weekday,subset=train,mtry=30,importance=T)
> bag.mod

Call:
 randomForest(formula = G3 ~ ., data = weekday, mtry = 30, importance = T,      subset = train)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 30

        OOB estimate of  error rate: 29.23%
Confusion matrix:
    0   1 class.error
0 349  98   0.2192394
1 116 169   0.4070175
```

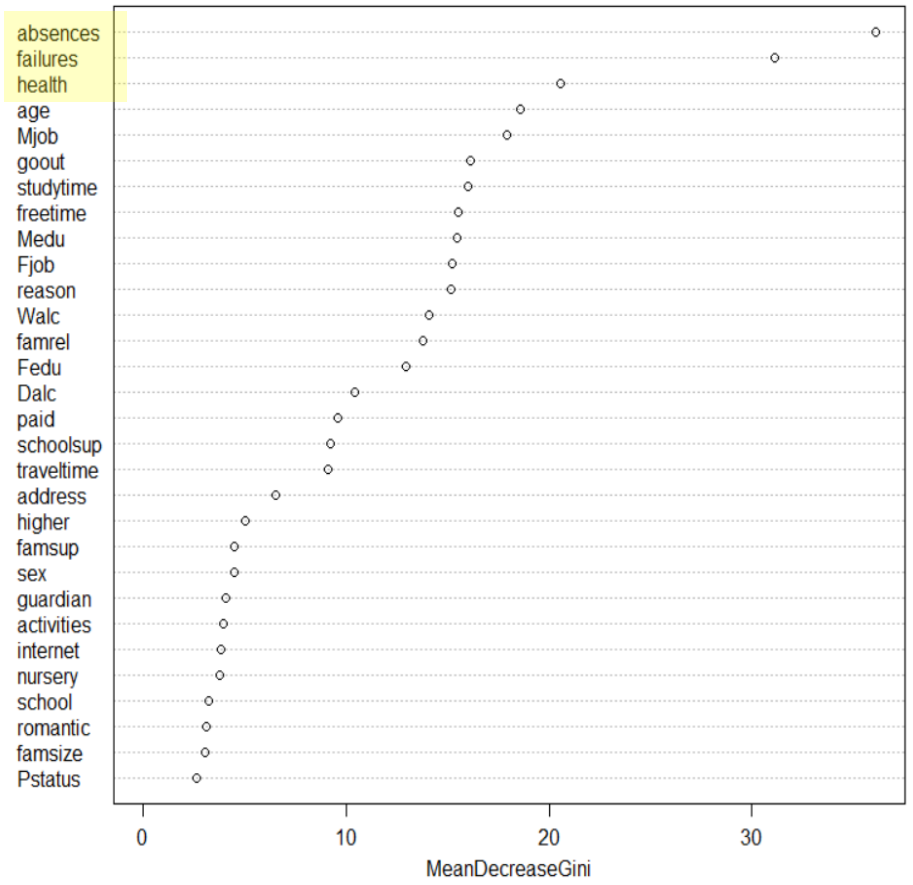
Content

데이터 분석

Bagging

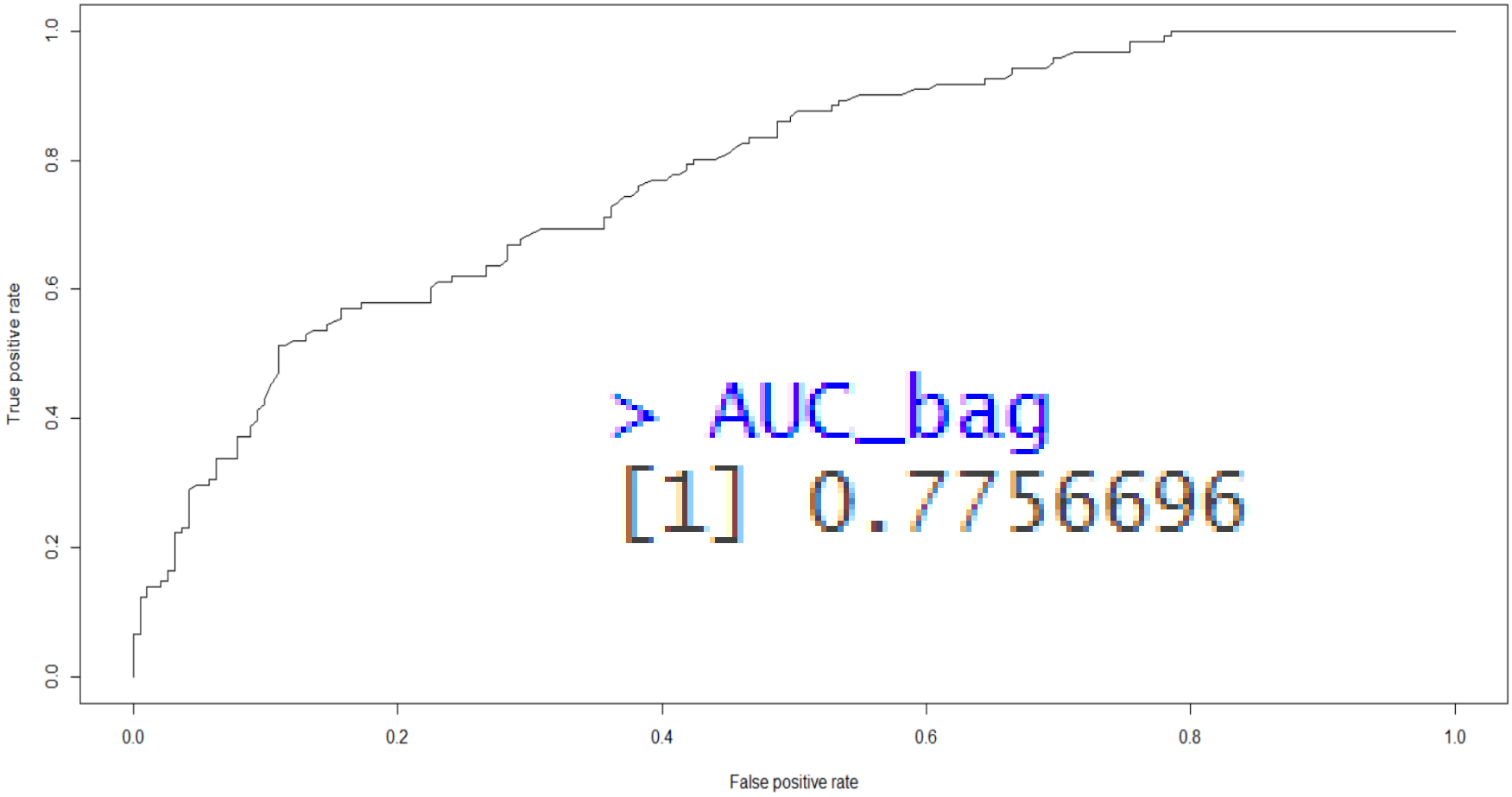
bag.mod





데이터 분석

Bagging(ROC Curve)



데이터 분석

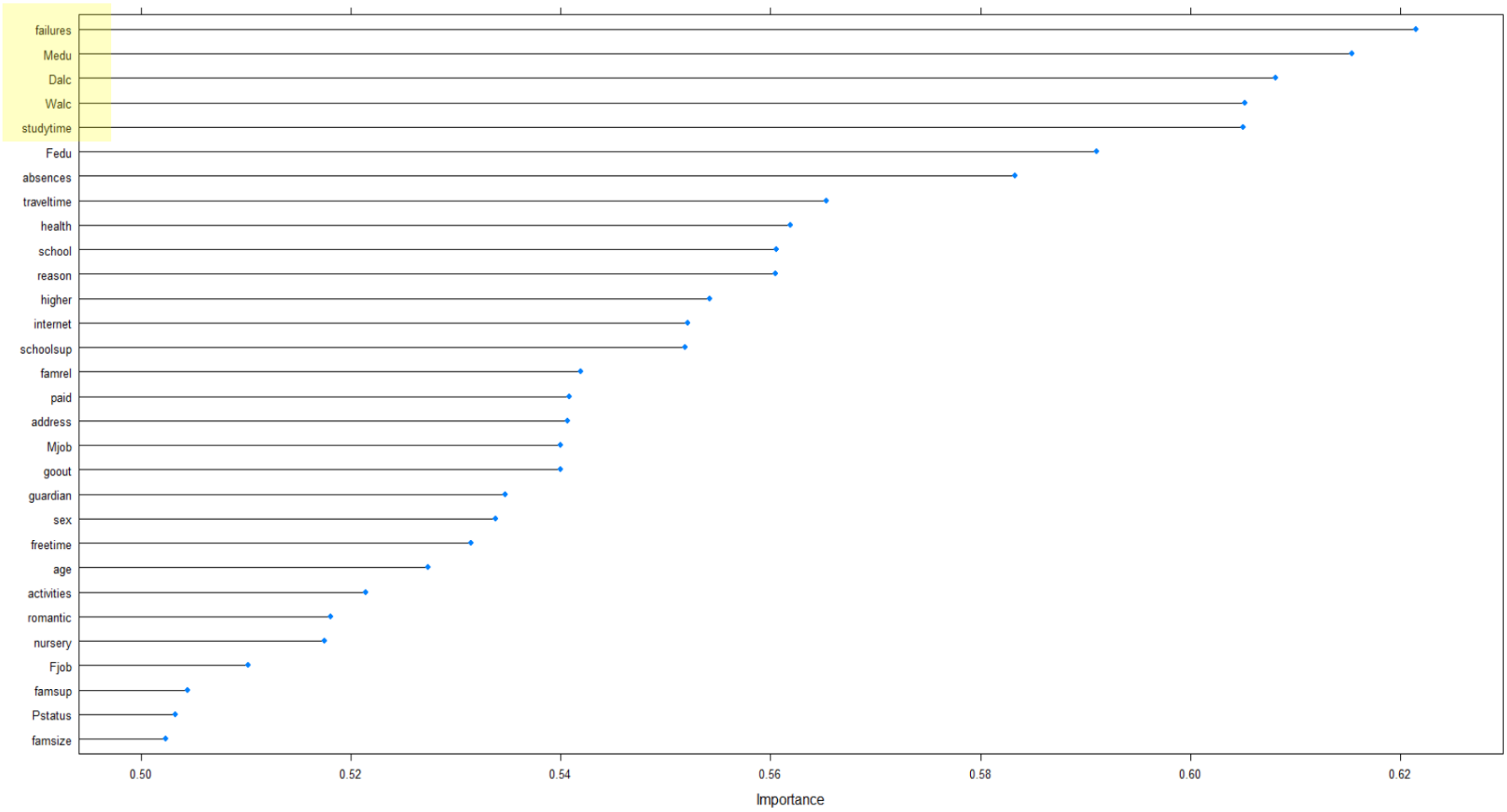
Gradient Boosting

```
Boosted Tree

732 samples
 30 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 659, 659, 658, 659, 659, 659, ...
Resampling results across tuning parameters:

  maxdepth  mstop  Accuracy   Kappa
  1          50    0.6188449  0.03280066
  1         100    0.6188449  0.03280066
  1         150    0.6188449  0.03280066
  2          50    0.6202332  0.05553533
  2         100    0.6202332  0.05553533
  2         150    0.6202332  0.05553533
  3          50    0.6257127  0.08639621
  3         100    0.6257127  0.08639621
  3         150    0.6257127  0.08639621

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were mstop = 50 and maxdepth = 3.
```

Content

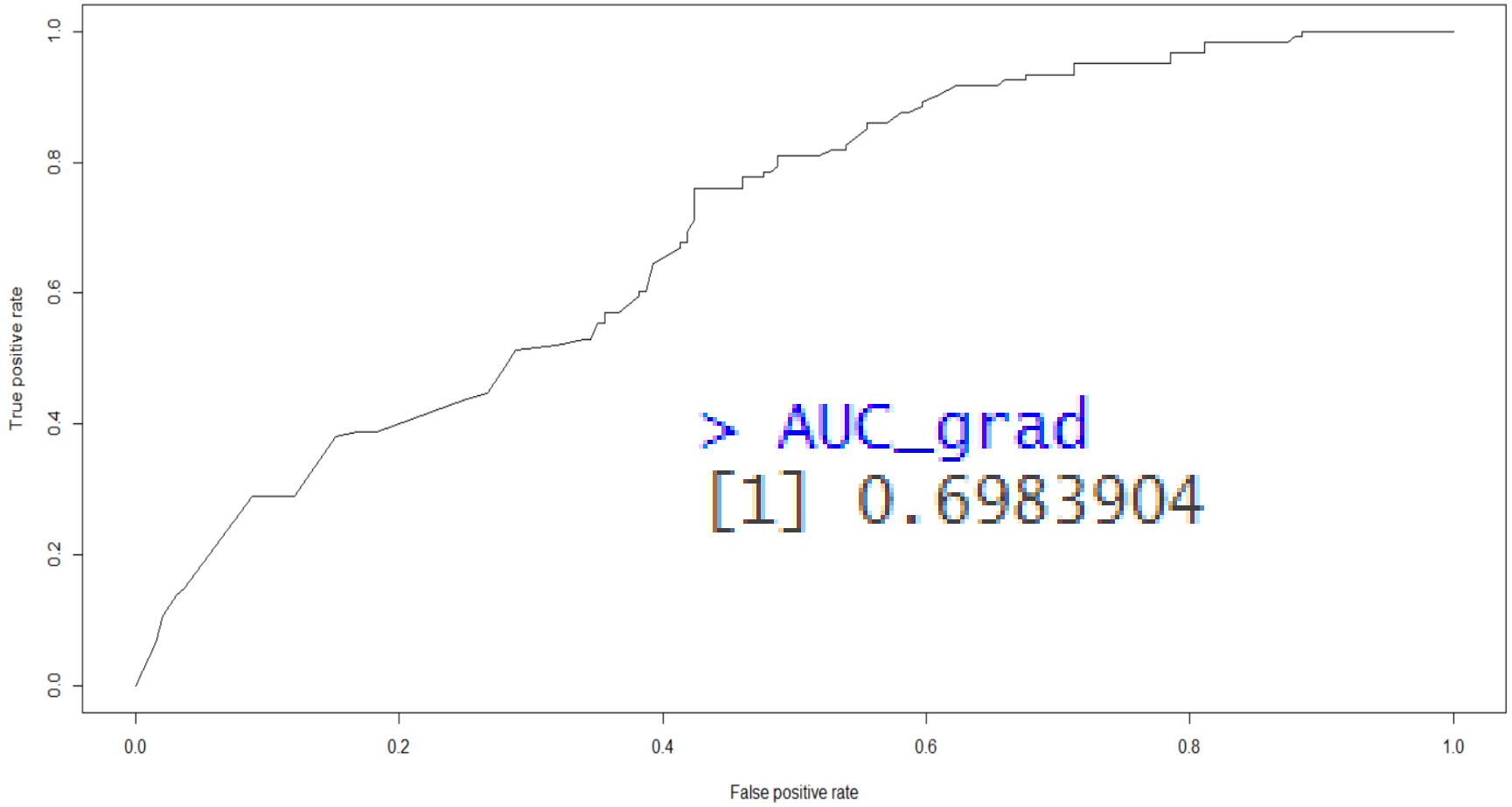
데이터 분석

Gradient Boosting(importance plot)



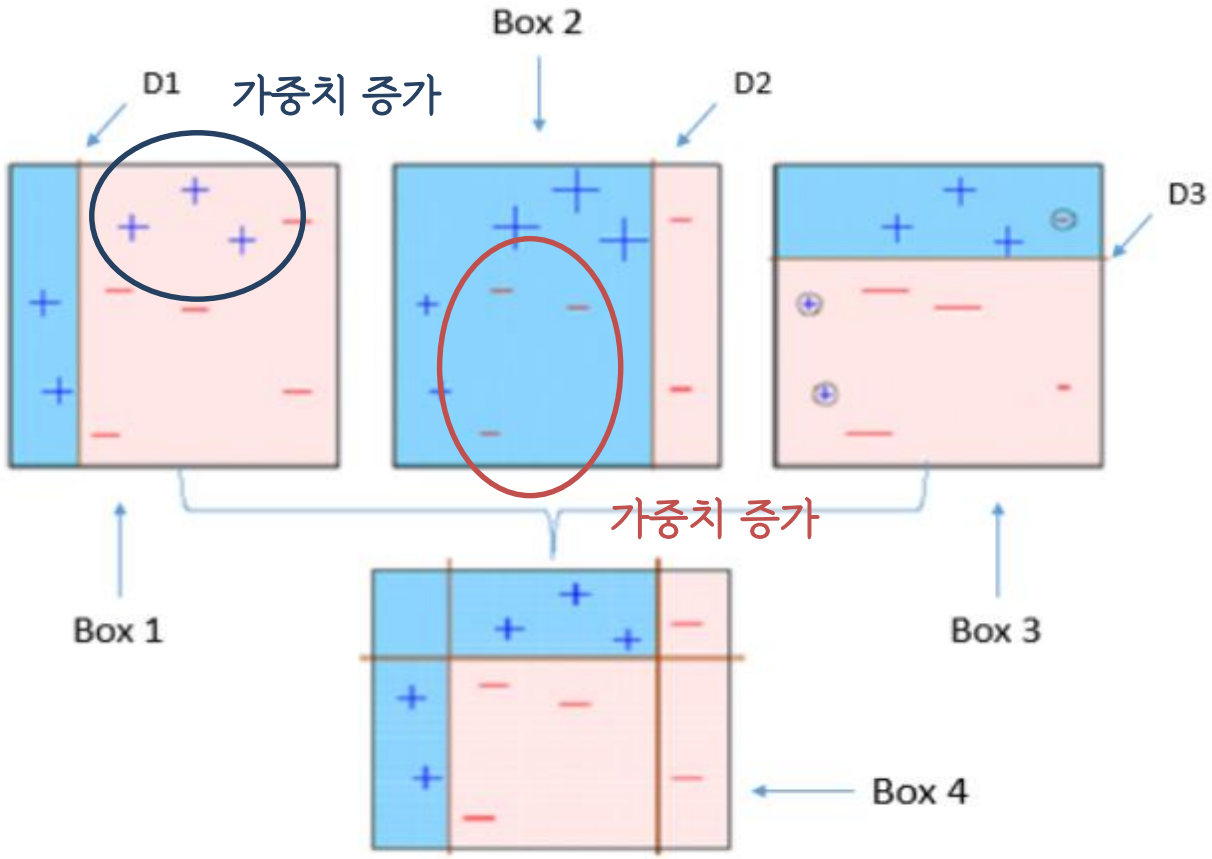
데이터 분석

Gradient Boosting ROC Curve



데이터 분석

Ada Boosting 이란?



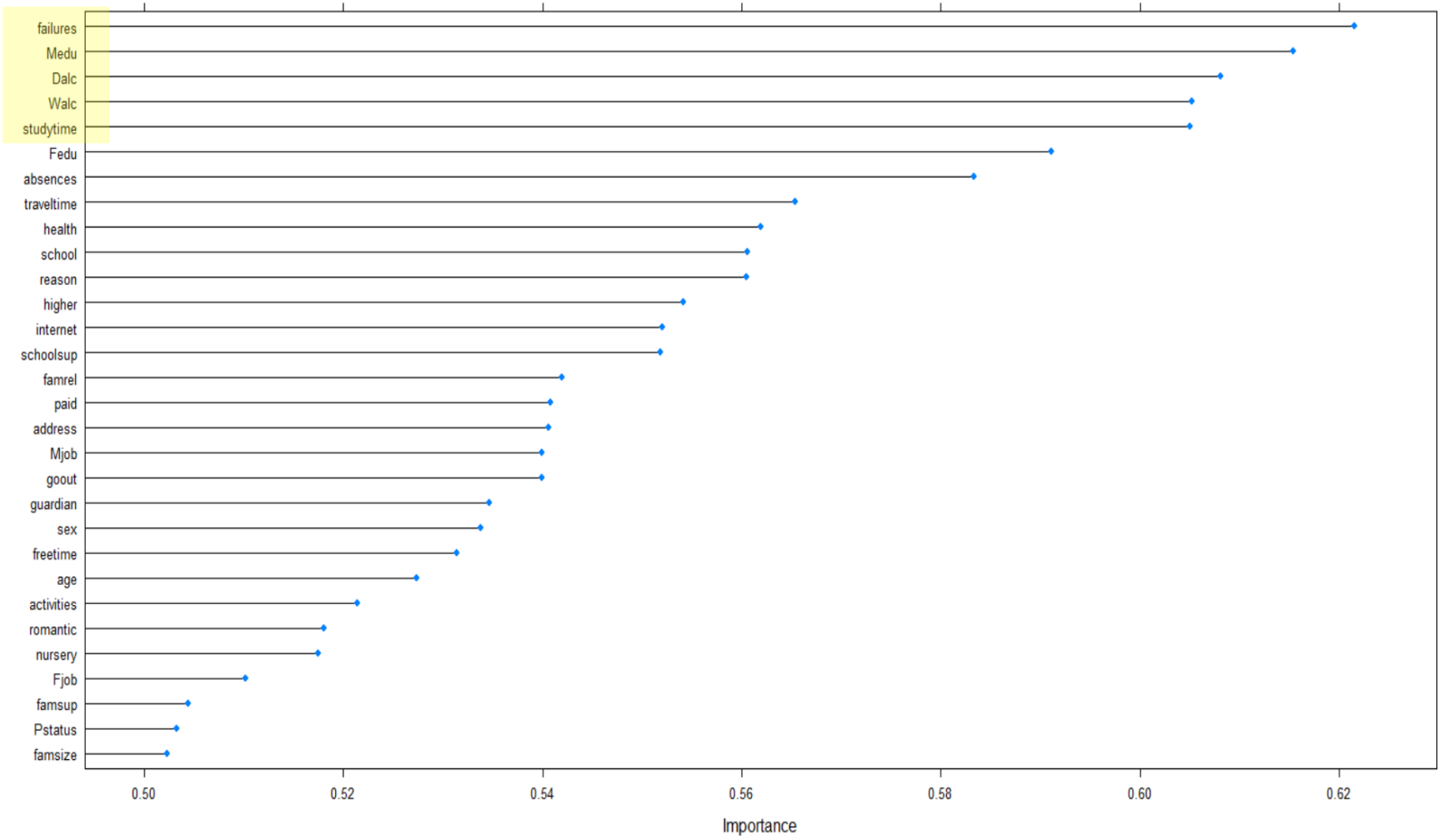
데이터 분석

Ada Boosting

```
AdaBoost Classification Trees

732 samples
 30 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 658, 658, 658, 659, 659, 658, ...
Resampling results across tuning parameters:

  nIter  method         Accuracy   Kappa
   50    Adaboost.M1    0.7008490  0.3718562
   50    Real adaboost  0.7091072  0.3665765
  100    Adaboost.M1    0.6995357  0.3701838
  100    Real adaboost  0.7010192  0.3551543
  150    Adaboost.M1    0.7022569  0.3732978
  150    Real adaboost  0.7023890  0.3568748

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were nIter = 50 and method = Real adaboost.
```

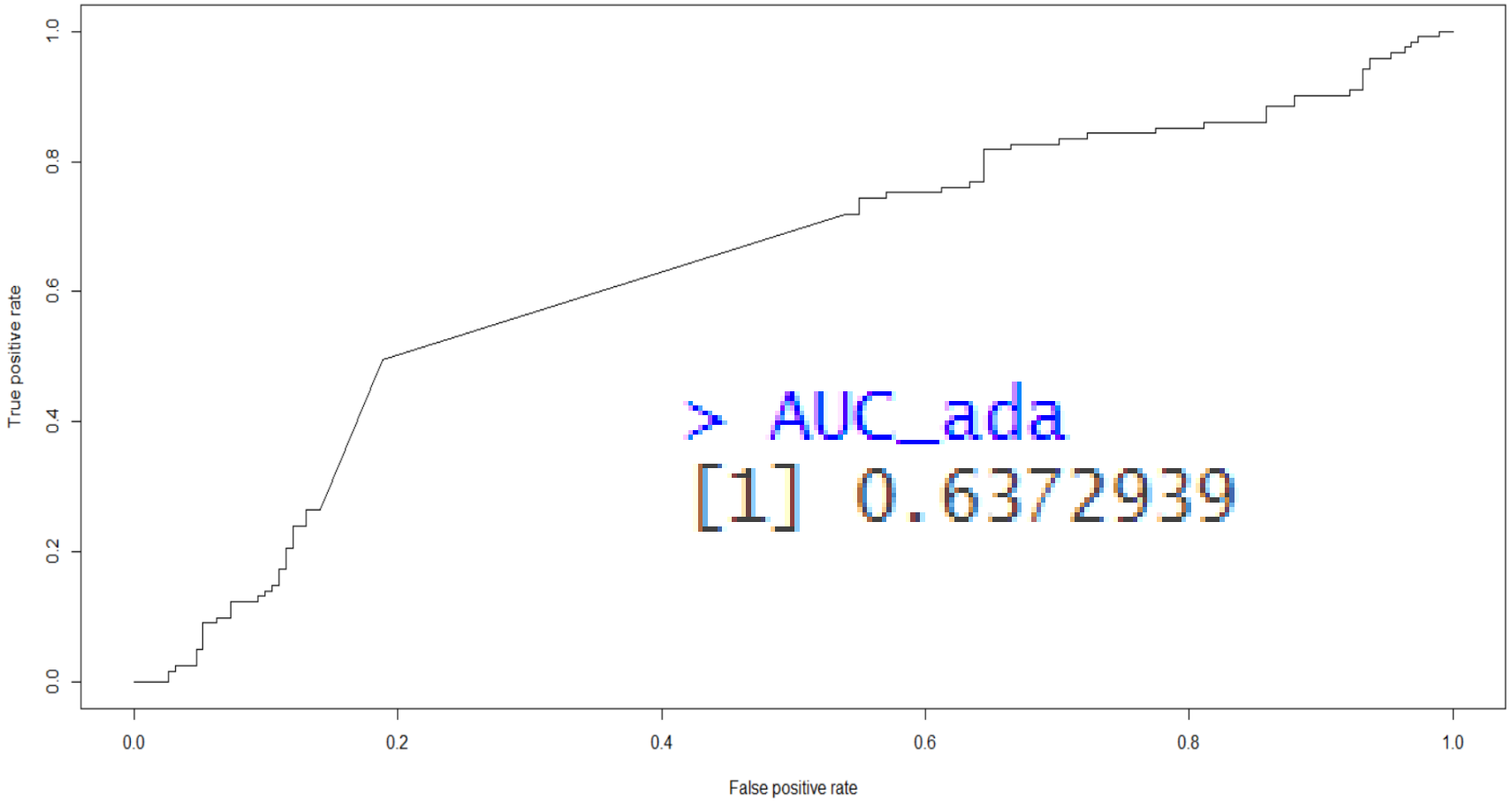
데이터 분석

Ada Boosting(importance plot)



데이터 분석

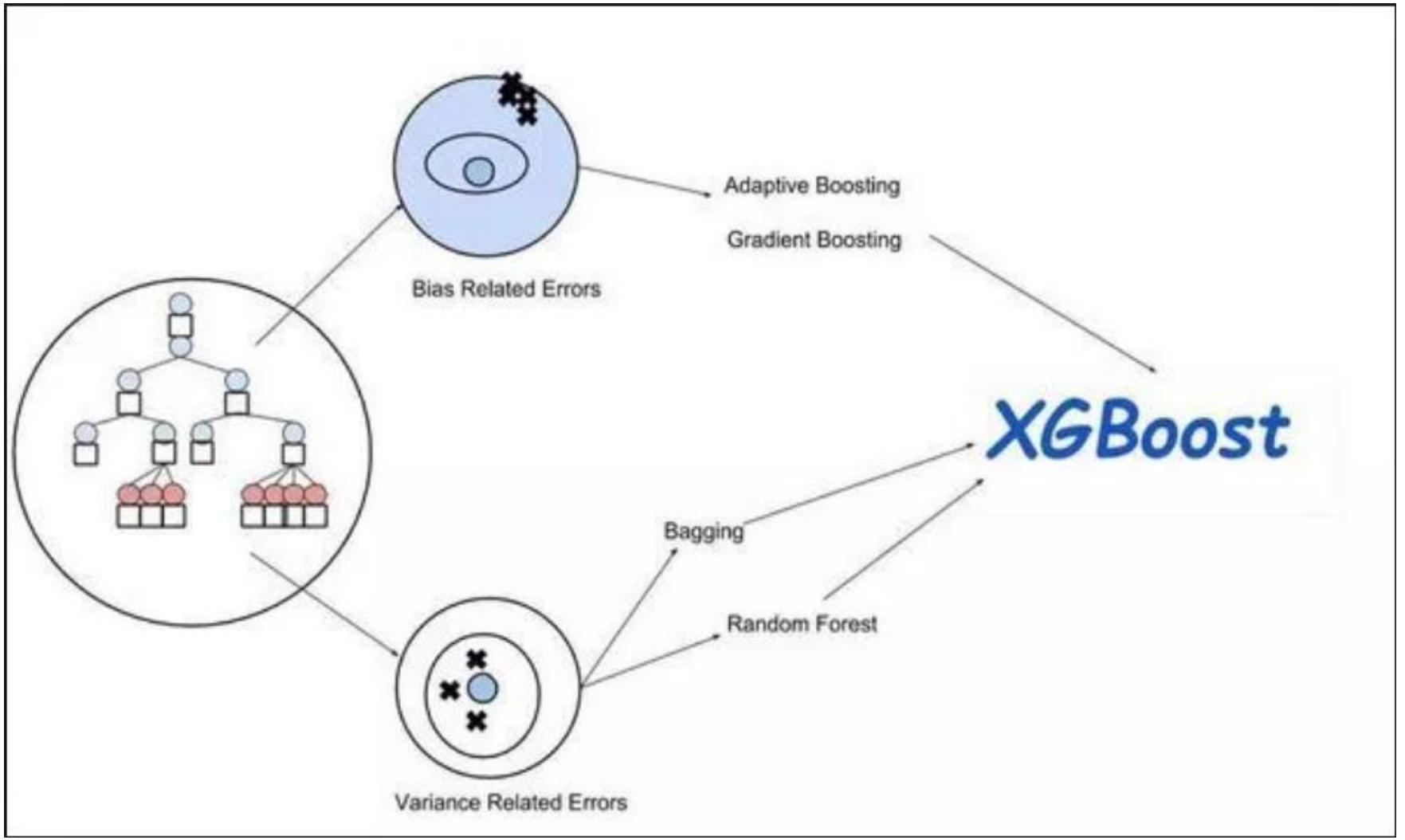
Ada Boosting ROC Curve



Content

데이터 분석

XG Boosting 이란?



Content

데이터 분석

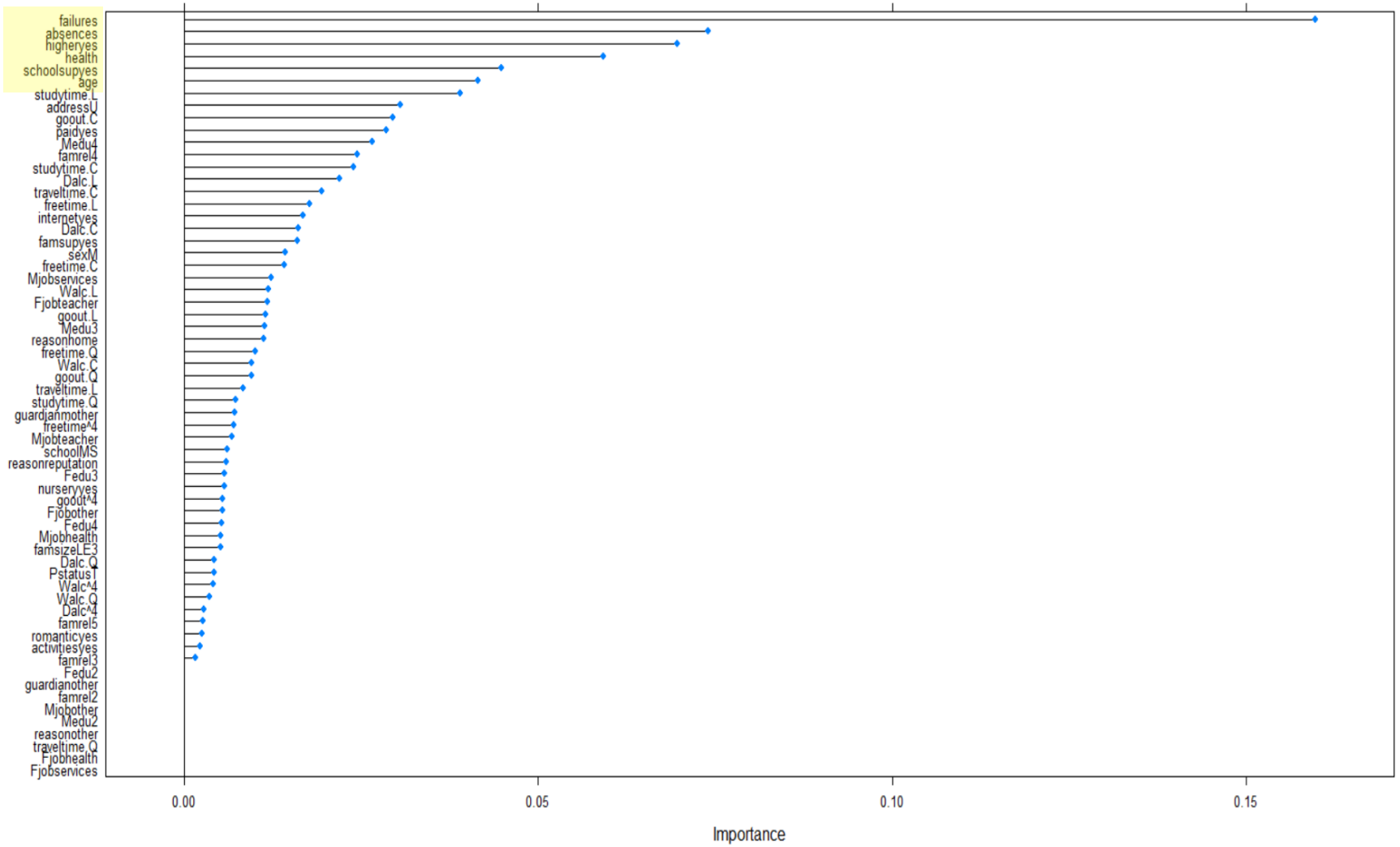
XG Boosting

```
eXtreme Gradient Boosting

732 samples
 30 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 659, 658, 659, 659, 659, 659, ...
Resampling results across tuning parameters:

  eta   max_depth  nrounds  Accuracy   Kappa
  0.01   5          100     0.6911884  0.3308697
  0.01   5          500     0.7089226  0.3676179
  0.01   5         1000     0.7048130  0.3610192
  0.01  10          100     0.6980193  0.3524828
  0.01  10          500     0.7199185  0.3949203
  0.01  10         1000     0.7226583  0.4022171
  0.10   5          100     0.7294521  0.4180702
  0.10   5          500     0.7294521  0.4202171
  0.10   5         1000     0.7308034  0.4232504
  0.10  10          100     0.7321733  0.4229163
  0.10  10          500     0.7240466  0.4044971
  0.10  10         1000     0.7253795  0.4070225

Tuning parameter 'gamma' was held constant at a value of 5
Tuning parameter 'colsample_bytree' was held constant at a value of
 0.7
Tuning parameter 'min_child_weight' was held constant at a value of 1
Tuning parameter 'subsample' was held constant at
 a value of 1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were nrounds = 100, max_depth = 10, eta = 0.1, gamma = 5, colsample_bytree =
 0.7, min_child_weight = 1 and subsample = 1.
```

Content

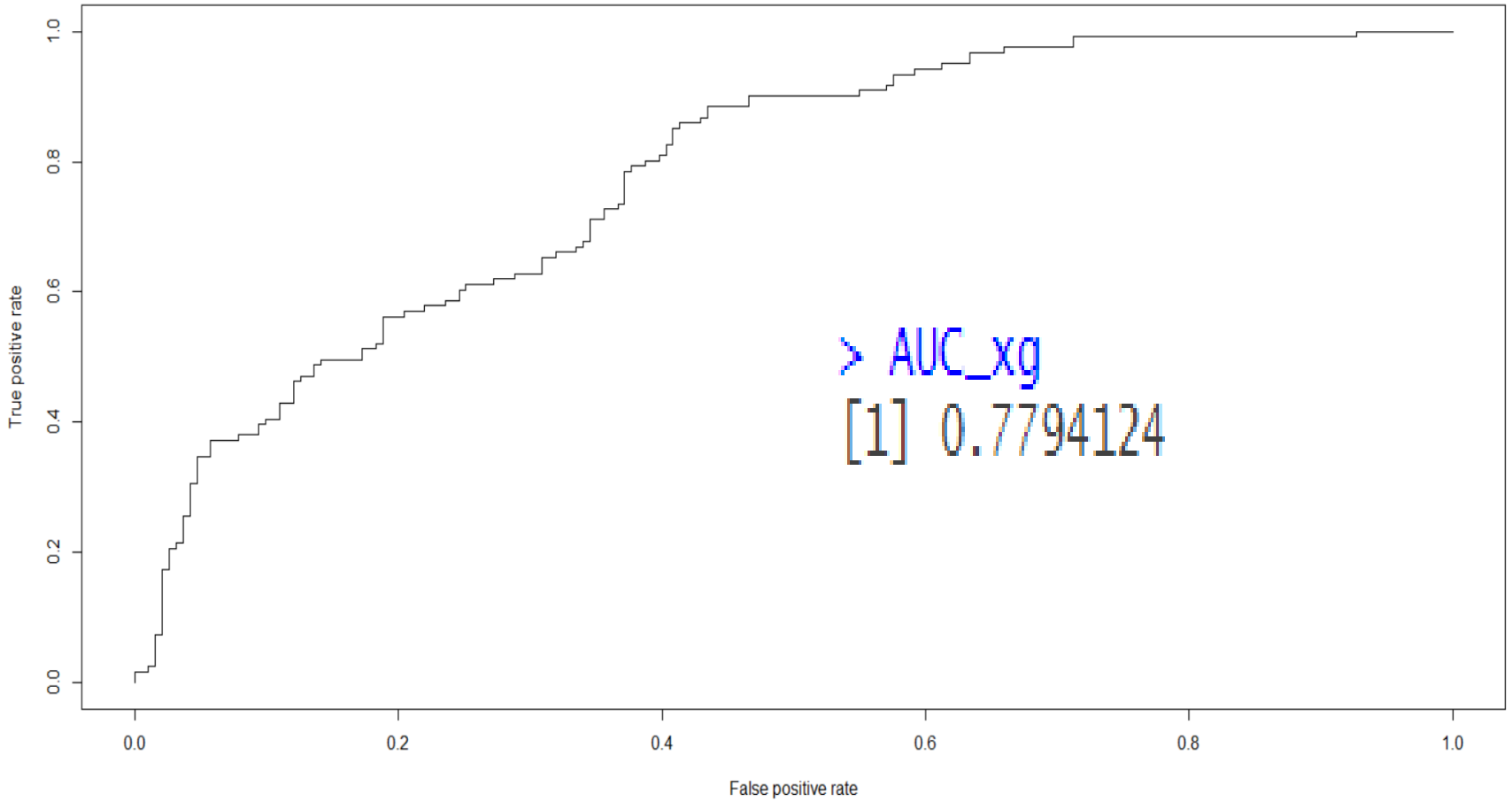
데이터 분석

XG Boosting(importance plot)



데이터 분석

XG Boosting ROC Curve



04

결론

Content

결론

모델 선정 기준

1. 같은 모델에 Seed를 다르게 적용

```
> set.seed(721)
> train=createDataPartition(weekday$G3,p=.7,list=F)
> traindata<-weekday[train,]
> testdata<-weekday[-train,]
> x.test=weekday[-train,-1]
> x.train=weekday[train,-1]
> y.test=weekday$G3[-train]
> y.train<-weekday$G3[train]
> #tree
> tree.mod<-tree(G3~.,data=weekday,subset=train)
> plot(tree.mod)
> text(tree.mod,pretty=0)
```

```
> set.seed(723)
> train=createDataPartition(weekday$G3,p=.7,list=F)
> traindata<-weekday[train,]
> testdata<-weekday[-train,]
> x.test=weekday[-train,-1]
> x.train=weekday[train,-1]
> y.test=weekday$G3[-train]
> y.train<-weekday$G3[train]
> #tree
> tree.mod<-tree(G3~.,data=weekday,subset=train)
> plot(tree.mod)
> text(tree.mod,pretty=0)
```

결론

모델 선정 기준

2. 서로 다른 결과 도출

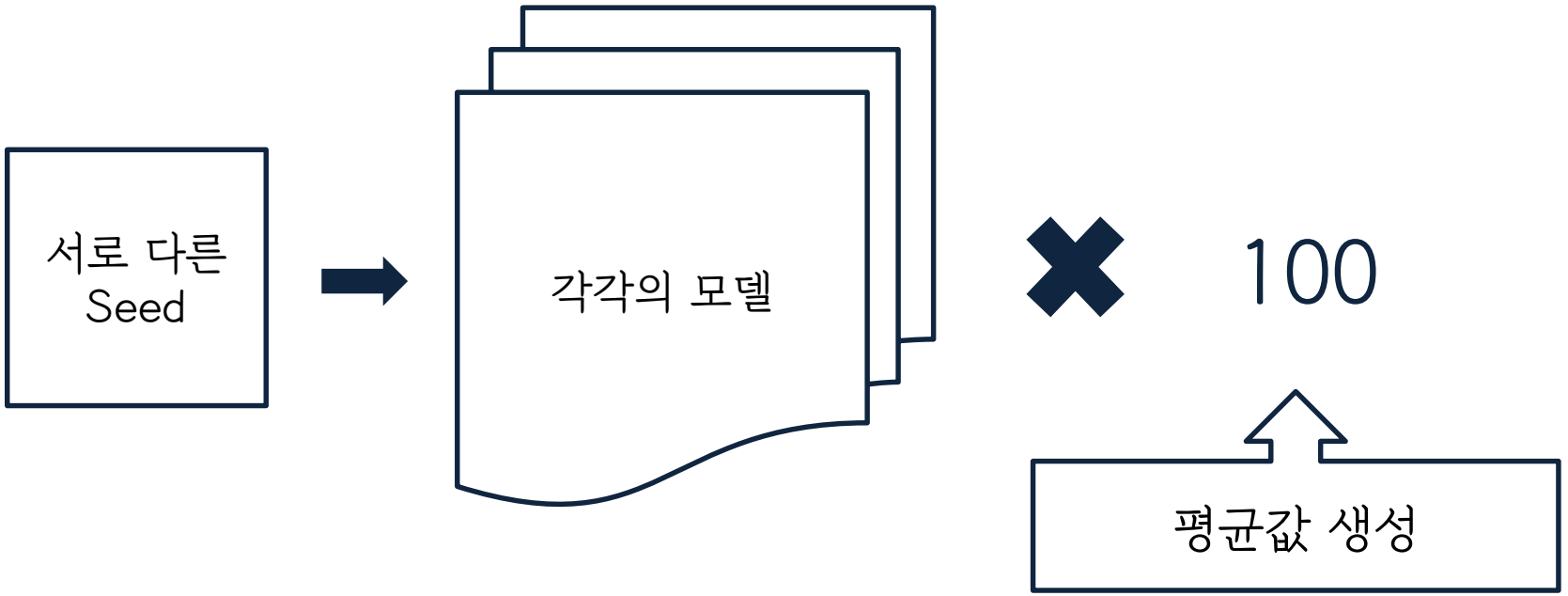
모든 모델에서 같은 현상 발생.

Content

결론

모델 선정 기준

3. 100번 실행 후, 각 결과의 평균값 생성



Content

결론

최종 결과

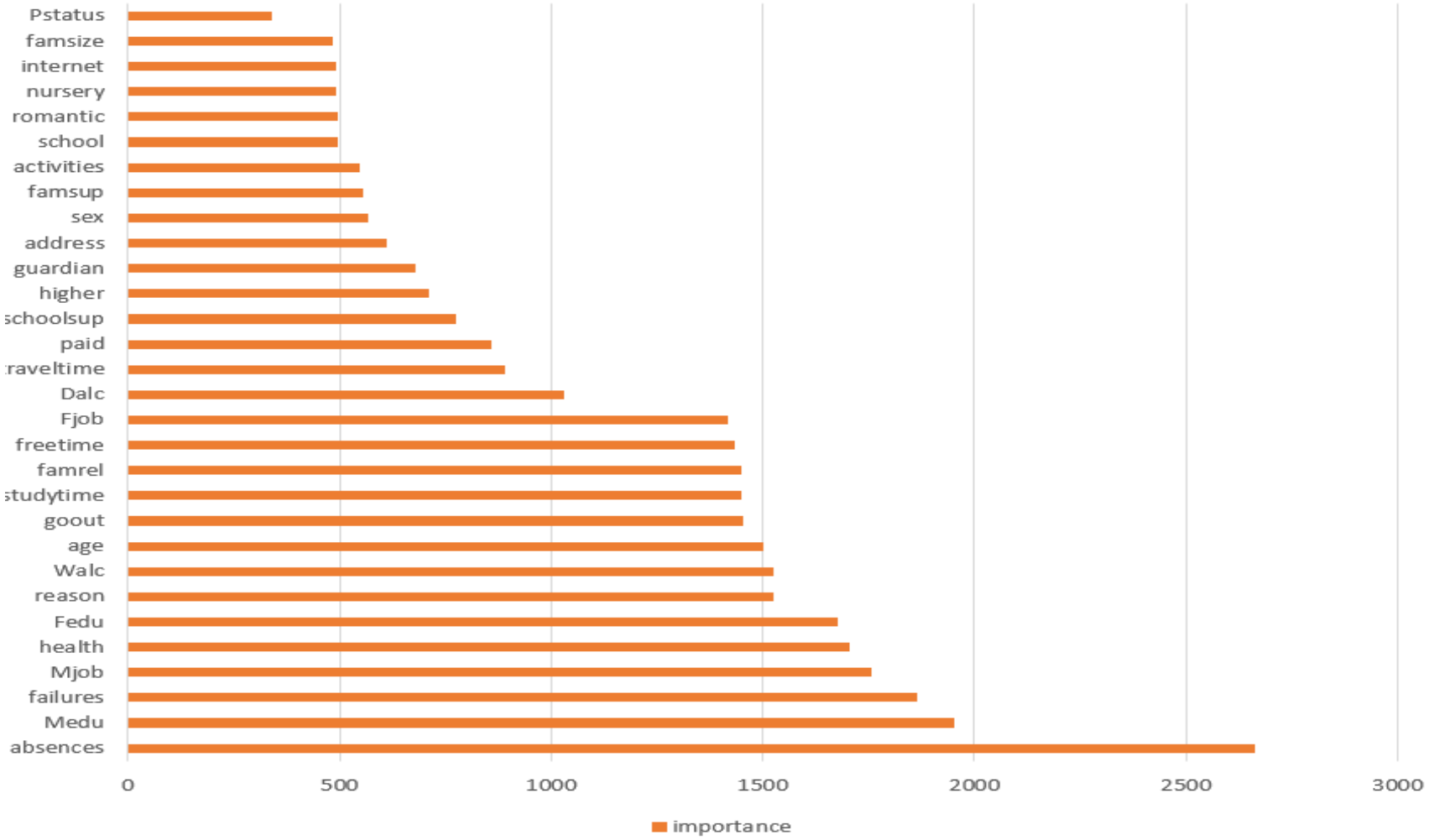
```
> summary(D100)
       X                GLM              LASSO            RIDGE
 Min.   :  1.00   Min.   :0.7314   Min.   :0.6882   Min.   :0.6907
 1st Qu.: 25.75   1st Qu.:0.7619   1st Qu.:0.7429   1st Qu.:0.7443
 Median : 50.50   Median :0.7739   Median :0.7559   Median :0.7601
 Mean   : 50.50   Mean   :0.7760   Mean   :0.7566   Mean   :0.7598
 3rd Qu.: 75.25   3rd Qu.:0.7890   3rd Qu.:0.7685   3rd Qu.:0.7739
 Max.   :100.00   Max.   :0.8355   Max.   :0.8271   Max.   :0.8276
      SVM1             SVM2             TREE             RF
 Min.   :0.5580   Min.   :0.6187   Min.   :0.6418   Min.   :0.7252
 1st Qu.:0.6115   1st Qu.:0.6551   1st Qu.:0.6814   1st Qu.:0.7631
 Median :0.6284   Median :0.6681   Median :0.7015   Median :0.7773
 Mean   :0.6256   Mean   :0.6700   Mean   :0.7007   Mean   :0.7764
 3rd Qu.:0.6424   3rd Qu.:0.6867   3rd Qu.:0.7171   3rd Qu.:0.7898
 Max.   :0.6725   Max.   :0.7277   Max.   :0.7685   Max.   :0.8309
      BAG              GRAD             ADA              XG
 Min.   :0.7254   Min.   :0.6769   Min.   :0.5226   Min.   :0.7160
 1st Qu.:0.7559   1st Qu.:0.7303   1st Qu.:0.6084   1st Qu.:0.7571
 Median :0.7723   Median :0.7443   Median :0.7185   Median :0.7719
 Mean   :0.7730   Mean   :0.7443   Mean   :0.6869   Mean   :0.7711
 3rd Qu.:0.7868   3rd Qu.:0.7598   3rd Qu.:0.7622   3rd Qu.:0.7835
 Max.   :0.8303   Max.   :0.8053   Max.   :0.8267   Max.   :0.8332
      GAM
 Min.   :0.7134
 1st Qu.:0.7504
 Median :0.7635
 Mean   :0.7657
 3rd Qu.:0.7800
 Max.   :0.8326
```

```
> sd(D100$GLM)
[1] 0.02261577
> sd(D100$LASSO)
[1] 0.02352366
> sd(D100$RIDGE)
[1] 0.02426985
> sd(D100$SVM1)
[1] 0.02388096
> sd(D100$SVM2)
[1] 0.02438459
> sd(D100$TREE)
[1] 0.02882484
> sd(D100$RF)
[1] 0.02244188
> sd(D100$BAG)
[1] 0.02354799
> sd(D100$GRAD)
[1] 0.02420008
> sd(D100$ADA)
[1] 0.08597267
> sd(D100$XG)
[1] 0.02301971
> sd(D100$GAM)
[1] 0.02439146
```

Content

결론

Random Forest



Content

결론

성적에 영향을 미치는 변수

변수명	변수뜻
Absences	학교 결석 횟수 / 0~93
Medu	엄마 교육 수준 / 0~4 (범주형)
Failures	유급 횟수 / 1,2,3,4
Mjob	엄마 직업 / (명목형)
Health	현재 건강 상태 / 1~5

결론

예측률 =

1 - OOB estimate of error rate

1-0.290=0.7095

```
> a
  [1] 0.2900214 0.3022003 0.2868131 0.2832900 0.2859828 0.2870196 0.2674457
  [8] 0.2928882 0.2892694 0.3151085 0.2845714 0.2740202 0.2757679 0.2716835
 [15] 0.2700449 0.2798561 0.3263465 0.3175461 0.3059599 0.2694809 0.2859459
 [22] 0.2814147 0.2834630 0.2582830 0.2788785 0.2962141 0.2821212 0.2826145
 [29] 0.2886231 0.2828560 0.2888859 0.2830916 0.2684733 0.2947004 0.2842625
 [36] 0.2803426 0.2597035 0.3176245 0.3084419 0.2899705 0.2914973 0.2659305
 [43] 0.2849908 0.2969023 0.2977577 0.2758451 0.3044585 0.2870819 0.2855498
 [50] 0.2808531 0.3075540 0.2664794 0.3017232 0.2887940 0.2857886 0.3038022
 [57] 0.3067207 0.2987703 0.3012601 0.3036260 0.2720153 0.3048761 0.2907873
 [64] 0.2865045 0.3079503 0.2911529 0.2979346 0.3021688 0.2692074 0.2943105
 [71] 0.2926618 0.3222360 0.2858152 0.2981877 0.3045573 0.2960455 0.2793609
 [78] 0.2790743 0.2756167 0.2825970 0.3200858 0.3015566 0.2944343 0.2932969
 [85] 0.2804242 0.2779717 0.3103177 0.2743625 0.2862056 0.2972997 0.2940042
 [92] 0.3112492 0.3043877 0.2913241 0.2808432 0.2946583 0.2963884 0.3120032
 [99] 0.2855825 0.2719764
> mean(a)
[1] 0.2904585
```

감사합니다.