Progetto di Modelli e metodi per l'inferenza statistica

Pietro Masini, Giulia Riccardi, Sofia Sannino, Alessandro Wiget

5 Giugno 2024

Obiettivo

Vogliamo costruire un modello in grado di prevedere la probabilità di dropout di uno studente di ingegneria matematica al termine del primo semestre del primo anno.

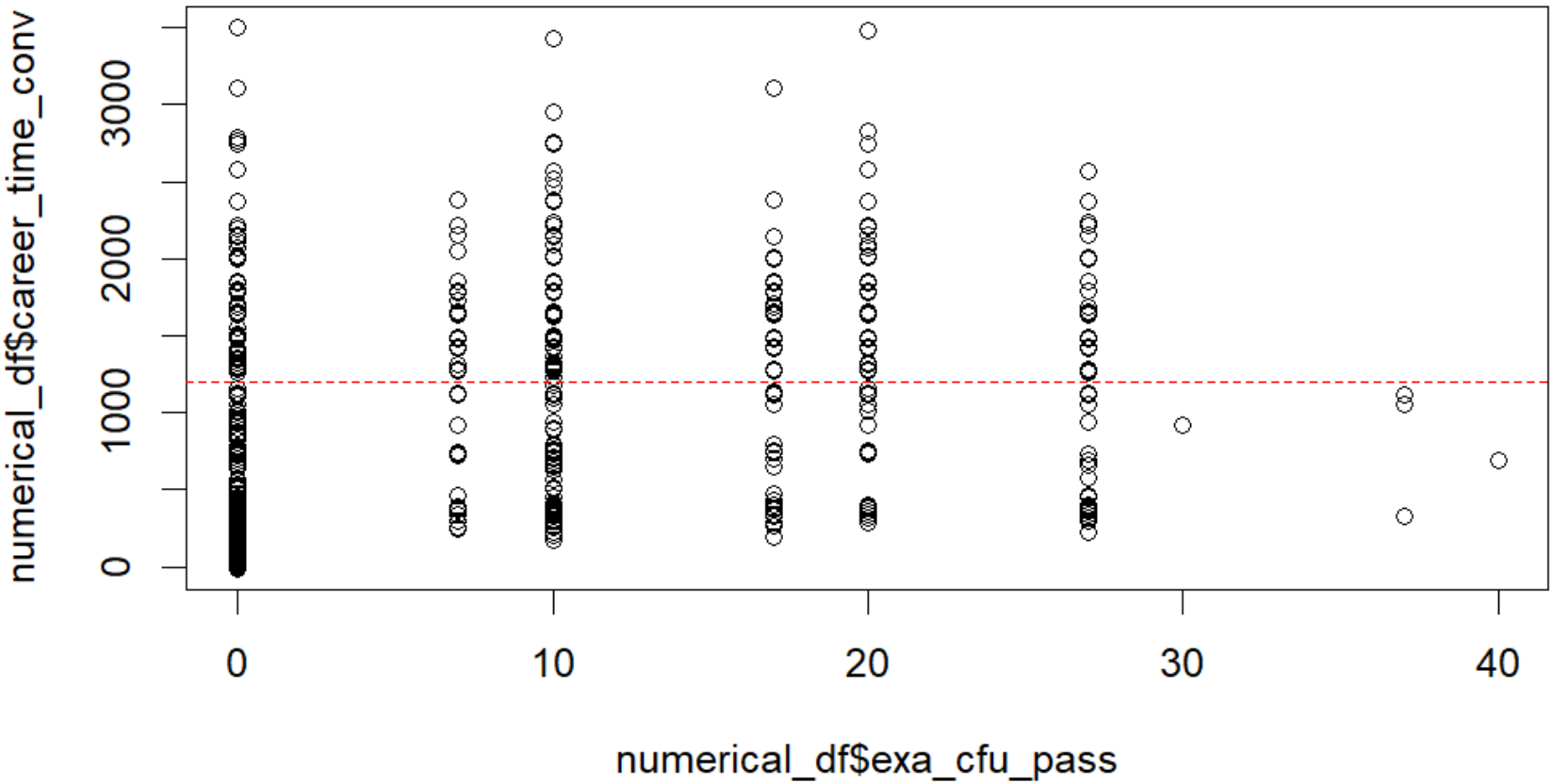
Il modello

Scelta di un modello di regressione logistico:
$Y_i \sim Be(p_i)$
$\log\left(\frac{p_i}{1-p_i}\right) = Z\underline{b}$
dove Z è la matrice disegno e $\underline{b}$ è il vettore dei parametri.

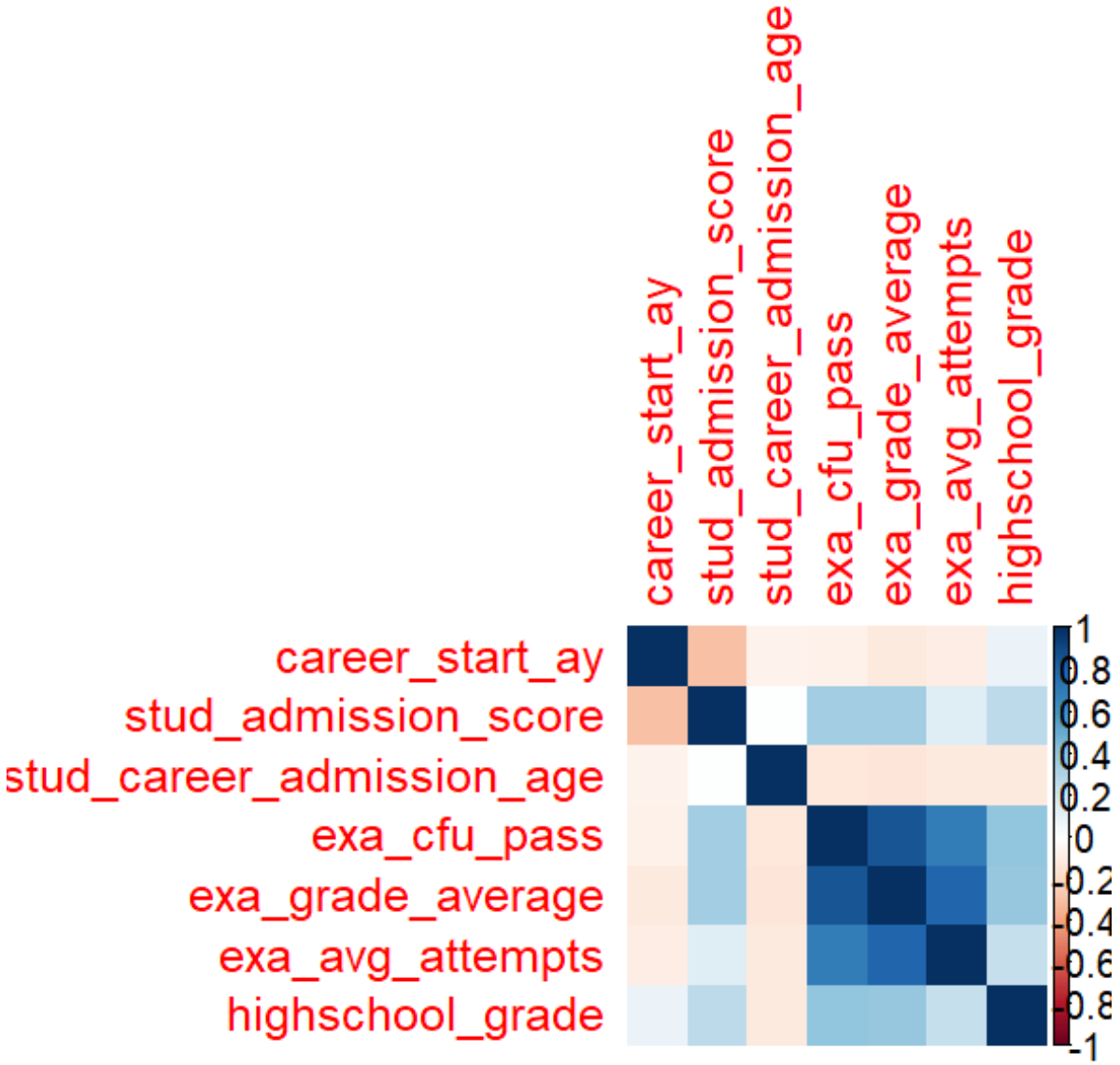
Analisi esplorativa

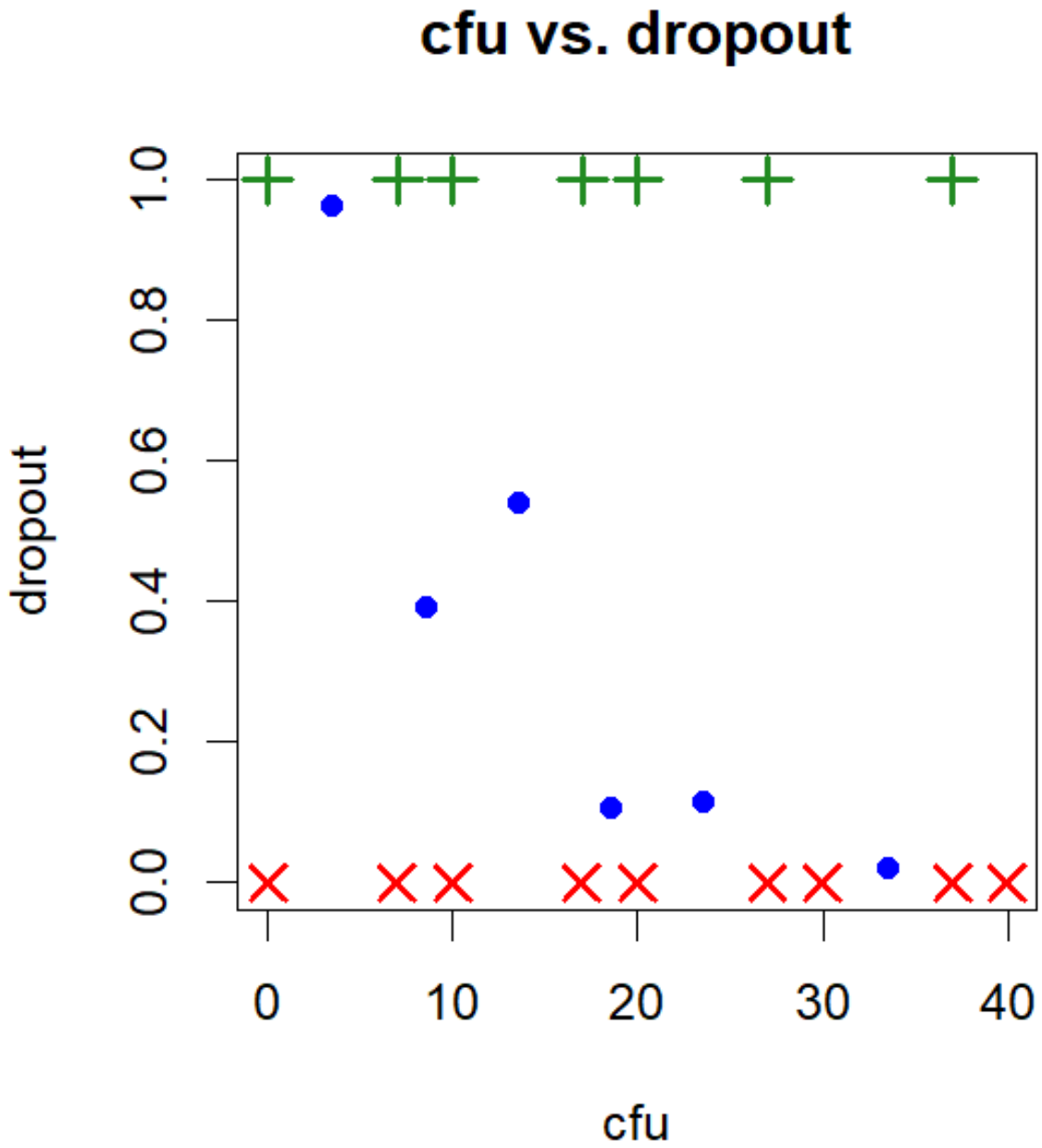
Il nostro dataset è del Politecnico di Milano e presenta covariate sia numeriche che categoriche.
Notiamo che ci sono persone che non hanno iniziato il corso e in più ci sono persone che sono iscritte da 3 anni e mezzo o più.

Analisi esplorativa e selezione degli studenti



Correlazioni tra le covariate numeriche




cfu vs. dropout
dropout
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
cfu

Prima regressione logistica


```
Call:
glm(formula = formula_num, family = binomial(link = logit), data = numerical_df)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               34.755964  66.394957   0.523    0.601
career_start_ay           -0.015488   0.032665  -0.474    0.635
stud_admission_score      -0.002402   0.010171  -0.236    0.813
stud_career_admission_age  0.098931   0.148400   0.667    0.505
exa_cfu_pass              -0.168512   0.014925 -11.290   < 2e-16 ***
exa_grade_average         -0.115155   0.019016  -6.056   1.4e-09 ***
exa_avg_attempts           0.310379   0.238548   1.301    0.193
highschool_grade          -0.019135   0.009362  -2.044    0.041 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2107.76  on 1879  degrees of freedom
Residual deviance:  815.99  on 1872  degrees of freedom
AIC: 831.99

Number of Fisher Scoring iterations: 6
```

Odds Ratio delle covariate più significative

```
exa_cfu_pass
   0.1854219
exa_grade_average
         0.8912281
```

Se uno studente acquisisce 10 cfu in più,
il rischio di dropout diminuisce dell'80%.

Inoltre, il rischio di dropout diminuisce del 10%
all'aumentare di un punto di media.

Selezione del modello tramite backward selection


```
Model:
dropout ~ exa_cfu_pass + exa_grade_average + highschool_grade
                  Df Deviance    AIC     LRT  Pr(>Chi)
<none>                 818.49 826.49
exa_cfu_pass       1   967.13 973.13 148.641 < 2.2e-16 ***
exa_grade_average  1   862.97 868.97  44.479 2.571e-11 ***
highschool_grade   1   823.67 829.67   5.182   0.02283 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] 826.4862
Single term deletions

Model:
dropout ~ exa_cfu_pass + exa_grade_average
                  Df Deviance    AIC    LRT  Pr(>Chi)
<none>                 823.67 829.67
exa_cfu_pass       1   986.72 990.72 163.05 < 2.2e-16 ***
exa_grade_average  1   870.03 874.03  46.36 9.841e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] 826.4862
```

Modello finale con covariate numeriche

```
Call:
glm(formula = "dropout~1+exa_cfu_pass+exa_grade_average+highschool_grade",
    family = binomial, data = numerical_df)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        5.473043   0.838854   6.524 6.83e-11 ***
exa_cfu_pass      -0.167997   0.014626 -11.487  < 2e-16 ***
exa_grade_average -0.104008   0.017211  -6.043 1.51e-09 ***
highschool_grade  -0.021203   0.009246  -2.293   0.0218 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2107.76  on 1879  degrees of freedom
Residual deviance:  818.49  on 1876  degrees of freedom
AIC: 826.49

Number of Fisher Scoring iterations: 6
```

Introduzione interazione tra le covariate numeriche

```
Call:
glm(formula = dropout ~ exa_cfu_pass + exa_grade_average + highschool_grade +
    exa_cfu_pass:exa_grade_average, family = binomial, data = numerical_df)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   4.674151   0.859873   5.436 5.45e-08 ***
exa_cfu_pass                 -0.037988   0.048964  -0.776   0.4378
exa_grade_average            -0.087164   0.016953  -5.141 2.73e-07 ***
highschool_grade             -0.015864   0.009292  -1.707   0.0878 .
exa_cfu_pass:exa_grade_average -0.005669   0.002091  -2.712   0.0067 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2107.76  on 1879  degrees of freedom
Residual deviance:  810.78  on 1875  degrees of freedom
AIC: 820.78

Number of Fisher Scoring iterations: 6
```

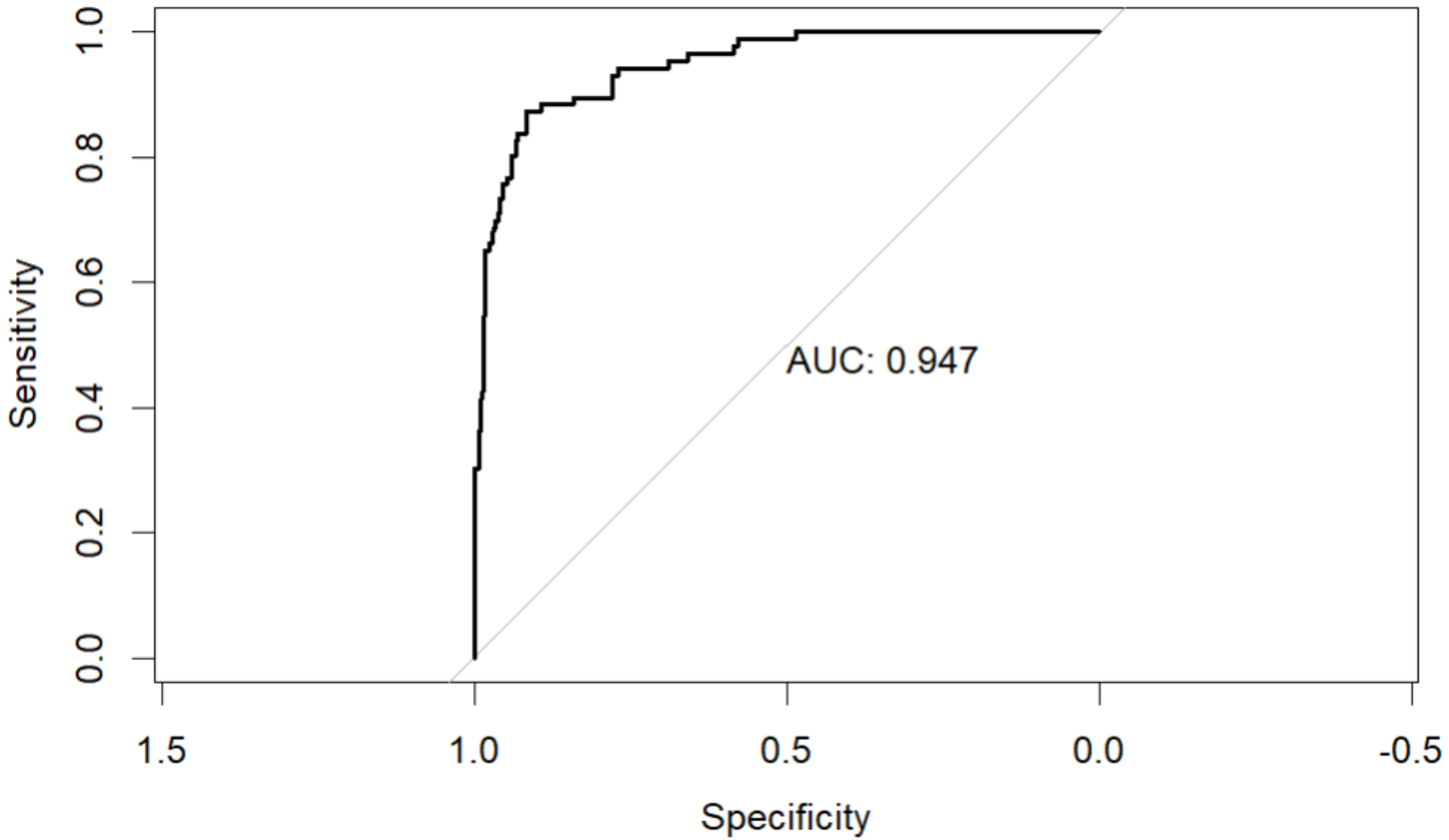
Introduzione variabili categoriche


```
Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                               4.811048   1.031176   4.666 3.08e-06
***
exa_cfu_pass                             -0.166832   0.014937 -11.169  < 2e-16
***
exa_grade_average                        -0.111726   0.017680  -6.320 2.62e-10
***
stud_genderM                              0.302170   0.202790   1.490   0.1362
previousStudiesOthers                     1.334128   0.657691   2.029   0.0425 *
previousStudiesScientifica                0.355375   0.371562   0.956   0.3389
previousStudiesTecnica                    1.184662   0.565604   2.095   0.0362 *
originsForeigner                         -0.396477   1.195581  -0.332   0.7402
originsMilanese                          -0.224974   0.219894  -1.023   0.3063
originsOffsite                           -0.779675   0.640482  -1.217   0.2235
income_bracket_normalized_on4fascia bassa  0.083811   0.269375   0.311   0.7557
income_bracket_normalized_on4fascia media  0.176529   0.227462   0.776   0.4377
income_bracket_normalized_on4LS          -0.531922   0.355726  -1.495   0.1348
highschool_grade                         -0.017799   0.009916  -1.795   0.0727 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2107.76  on 1879  degrees of freedom
Residual deviance:  801.76  on 1866  degrees of freedom
AIC: 829.76
```

Backward selection

Tramite una backward selection, sono state selezionate le covariate significative.
E' emerso che nessuna covariata categorica è significativa.
Il modello ottimale rimane quello con le sole numeriche.

Classificazione, curva ROC, confusion matrix



Soglia ottimale: 0.2181341

Classificazione, curva ROC, confusion matrix


```
          Reference
Prediction   0   1
         0 266  11
         1  24  75

               Accuracy : 0.9069
                 95% CI : (0.8729, 0.9343)
    No Information Rate : 0.7713
    P-Value [Acc > NIR] : 5.045e-12

                  Kappa : 0.7495

 Mcnemar's Test P-Value : 0.04252

            Sensitivity : 0.9172
            Specificity : 0.8721
         Pos Pred Value : 0.9603
         Neg Pred Value : 0.7576
             Prevalence : 0.7713
         Detection Rate : 0.7074
   Detection Prevalence : 0.7367
      Balanced Accuracy : 0.8947
```

Predizione sugli studenti in corso

Usando il valore soglia arrontondato e il modello ottimale costruito, la probabilità predetta sugli studenti attivi è circa del 37%.

Fonti