Study of dropout of undergraduate students using a logistic regression model

Miriam Rodrigues Silvestre[1], Pedro Henrique de Aragão[1]
[1] São Paulo State University - UNESP

Abbreviated abstract: Every year new classes start their courses at public colleges. Winning a place is usually very difficult, due to the competition faced in the entrance exam. However, there are several cases of student dropout, and this ends up harming other students who could fill these vacancies and graduate. To avoid the dropout problem, it was thought to find a statistical model that would be able to detect these cases, in order to allow decisions to be taken to prevent them.

Problem, Data, Previous Works

The problem is to avoid the dropping out at Statistics Course at University Unesp, at Presidente Prudente, São Paulo, Brazil.
To solve this problem is proposed to build a model that can predict if a student is going to dropout or not.

- There is not any previous work in this area at this Course/University.
- Were collect information about the student in the Inscription System of the University.
- A model could help the Course Coordinator to identify the new students with high probability of dropout and try to avoid this action.
- There was collected 19 variables of students that were at University by 2008 to 2016, and the cut point was at 1o. Sem/2020.
- The data was divided in two datasets with about 70% for training and 30% for test.
- The training data set was used to construct the model. First, it was applied the statistical Logistic Regression Model, then after was applied stepwise procedure, that provided the best model with just 3 variables.
- The generalizability of the constructed model was evaluated in a test set.

Methods

Figure 1 – Adjusted Stepwise Model in training data

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -1.657e+03  3.115e+02  -5.319 1.04e-07 ***
Anoentrada                  8.241e-01  1.549e-01   5.320 1.04e-07 ***
Nota_Vestibular            -7.791e-02  3.153e-02  -2.471   0.0135 *
CotasNão                    2.511e+00  1.094e+00   2.295   0.0217 *
CotasPreto, pardo ou índio  2.379e-01  1.197e+00   0.199   0.8424
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 155.78  on 113  degrees of freedom
Residual deviance: 104.95  on 109  degrees of freedom
AIC: 114.95
```

Figure 2 – Deviance of Adjusted Stepwise Model

```
             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                           113     155.78
Anoentrada    1   39.726       112     116.06 2.922e-10 ***
Nota_Vestibular  1    3.247       111     112.81   0.07154 .
Cotas         2    7.862       109     104.95   0.01962 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3 – Residuals and Leverage

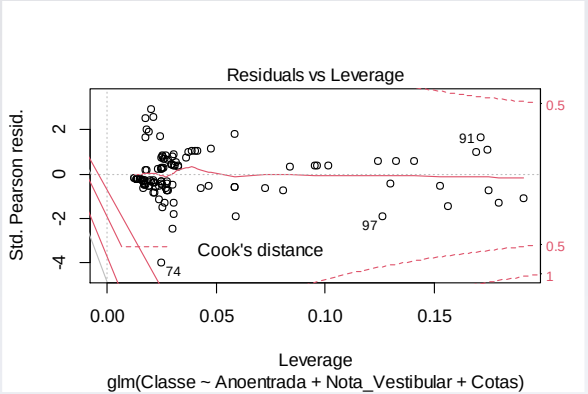
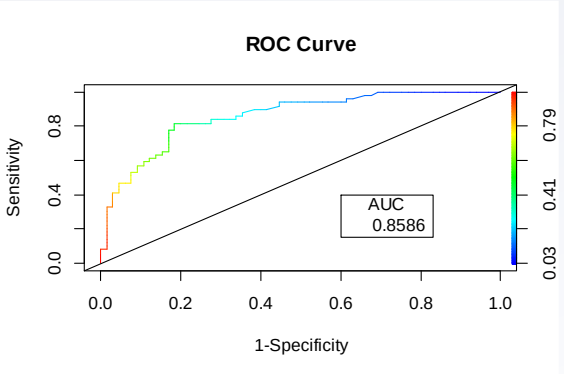


Figure 4 – The ROC Curve



Results and Conclusions

Training Data (n=114) Accuracy = 81.58 % and Classification Error = 18.42 %.
Test Data (n=50) Accuracy = 66 % and Classification Error = 34 %.

Figure 5 – Odds Ratio and Confidence Interval 95%.

	OR	2.5 %	97.5 %
(Intercept)	0.000000	0.0000000	0.0000000
Anoentrada	2.279804	1.6828339	3.0885441
Nota_Vestibular	0.925045	0.8696177	0.9840052
CotasNão	12.314936	1.4423445	105.1466216
CotasPreto, pardo ou índio	1.268590	0.1215075	13.2446249

Anoentrada: increases one year, can increases the chance of dropping out in 127,98 %.
Nota_Vestibular: increases one point, can decreases the chance of dropping out in 7,50 %.
CotasNão: chooses for not Quotas increases the chance of dropping in 1131,49 %.
CotasPPI: its not significant (Figure 1) and CI 95% includes 1.

Conclusions: The University Staff can apply the adjusted model at the beginning of a class to detect future dropouts and make decisions that can prevent this.