Risk Factors for College Drop-Out

Abstract

University enrollment is one measure of success of secondary education programs – but how do students perform once in college? In this paper, we examine Portuguese higher education data on students' demographics, prior education, and university outcomes to create a regression model that universities can easily implement to identify students at risk of dropping out, and ultimately provide them with extra support. We compare eight models created by the stepwise procedure for probit and logistic regression with AIC and BIC criterion at two thresholds. In our comparison, we prioritize sensitivity because we want to ensure that all at-risk students are identified. We select a final model with seven significant features to predict if a student is at risk of dropping out: the mother's previous education, if they're in debt, their gender, admission grade, if tuition fees are up to date, and if they hold a scholarship.

1. Background & Significance

University enrollment is one measure of success of secondary education programs – but how do students perform once in college? To help ensure that all university students are given the support they need to stay enrolled and graduate from college, we aim to create a regression model to identify students at-risk of dropping out. Specifically, we ask: “Can we use a regression model to accurately predict if a student is likely to drop out, if additional support is not provided?” While researchers have previously used our same dataset to create a machine learning model with the same ultimate goal of identifying at-risk students, a regression model is valuable because it would be more easily understood by a wider audience and would be easier for universities to implement themselves.

2. Data

2.1 Data Description

The dataset, “Predict students' dropout and academic success,” was posted to the UC Irvine Machine Learning Repository on December 12, 2021 by Martins et al. from their paper, "Early prediction of student’s performance in higher education: a case study.” It was funded by the program SATDAP - Capacitação da Administração Pública under the Portuguese grant POCI-05-5762-FSE-000191. The dataset includes higher education students’ background information, academic performance, and outcome university status, which is recorded to have been measured at the end of the normal duration of the program. It contains 4424 rows (i.e. students) and 36 predictors.

2.2 Data Cleaning

From the original dataset, we dropped 22 rows and 13 columns. We dropped rows that had the level “frequency of higher education” for at least one of the “qualification” variables that recorded the mother’s father’s, or student’s prior education levels. We excluded this category because, even after online research and trying to contact the original researchers, we were unable to understand it. To ensure generalizability of our model, we dropped the “Course” column because its values seemed specific to the particular university. We dropped 12 columns related to students’ “Curricular units” because these columns created a time-series issue (as they reported data for each student’s first and second semester) and had lots of “0” values, even though the data supposedly had no missing data, which seemed suspicious. Also, this kind of data would be unknown before students entered college, thus making them unhelpful for identifying at-risk students before they entered college, and several of the columns were correlated. After initial cleaning, our dataset had 4402 rows and 23 predictors.

Finally, we combined levels of categorical variables. Several predictors had many levels (e.g. “Father’s qualification” had 29), so we did this to avoid overfitting. Although our response variable only had three categories (“enrolled,” “graduated,” and “dropped out”), we also combined its levels “enrolled” and “graduated” because we were concerned about whether students dropped out. Importantly, we kept “unknown” categories, rather than imputing values, because we suspected that these values were missing not at random (e.g. “Mother’s qualification” might be missing because the mother had little to no formal education).

3. Methods & Results

3.1 Variable Selection

After data cleaning, we pursued reducing our total number of predictors (23) to a lower-dimensional feature space due to concerns of multicollinearity and available degrees of freedom. Due to our level combination for the “Nationality” predictor, there was perfect linear dependency between “International” and “Nationality (cleaned)” predictors. Thus, we dropped the latter. There was no further multicollinearity within the quantitative and categorical variables

from VIF screening (threshold of 0.5; Figure A5) and Cramer's M (using a threshold of correlation > 0.6), respectively. We did not use principal component analysis in the interest of interpretability.

3.2 Model Fit & Comparison

Using the train-test procedure (70-30 split), we performed both-direction stepwise procedure on probit and logistic regressions, comparing with AIC and BIC. This produced four models: probit AIC, probit BIC, logistic AIC, and logistic BIC. Because our target variable was not well-balanced (approximately 1:2 ratio between students who did not and did drop out), we performed threshold tuning to optimize the models. To choose thresholds for each model, we focused on "sensitivity" and "F-measure." We prioritized sensitivity because we want to ensure that all at-risk students are identified (i.e. we thought that it would be better to give extra support to students who did not need it than to miss people who did need it). However, we also considered F-measure – a trade-off between sensitivity and precision (in this case, cost) – because providing extra support to all students is costly and inequitable.

In the threshold tuning, we compared sensitivity and F-measures of all four models run with the thresholds 0.2, 0.3, 0.4, 0.5, and 0.6 (Table A1). For all models, sensitivity was maximized at the threshold 0.2, but F-measure was optimized at threshold 0.3. Specifically, between the 0.3 and 0.2 threshold, there was approximately a 0.15 increase in sensitivity and a 0.03 decrease in F-measure. From this, we decided to continue with eight models: each model – probit AIC, probit BIC, logistic AIC, and logistic BIC – at thresholds 0.2 and 0.3.

Based on the performance metrics of all eight models (Table A2), we chose the logistic BIC model at the 0.2 threshold. Across all four regression types, models at the 0.3 threshold performed better on almost every metric except sensitivity. However, we felt that the about 0.15-greater sensitivity in the 0.2 threshold models was significant enough to focus on the models with threshold 0.2. Of those, probit BIC and logistic BIC had the highest sensitivity values (0.833 and 0.829, respectively) and were parsimonious – each had 10 terms, much fewer than the probit AIC's 30 and logistic AIC's 26 terms. While these models had the same terms but with slightly different coefficient estimates (Figure A6), we chose the logistic BIC model at the 0.2 threshold because of its interpretability.

We fit a delta deviance plot to identify any possible influential points but found none, so we did not remove any points (Figure A7). Therefore, our final model was the logistic BIC model at the 0.2 threshold, refit on the full dataset to improve coefficient estimates (Figure A9).

3.3 Final Model

The final model includes these terms: mother's qualification (with levels of "less than secondary," "secondary," "specialized," and "unknown" and baseline category "Bachelor's or higher"), admission grade, debtor, tuition fees up to date, gender, scholarship holder, and age at enrollment. Regression diagnostics showed slightly heavier tails than a normal distribution and identified six outliers, although none were influential (Figure A8). We recalculated the performance metrics and plotted the ROC curve (Table A3; Figure A10).

Because we optimized sensitivity, our model correctly predicts 85.6% who actually dropped out as being at risk. However, this comes at a cost: our precision is 45.4%, meaning that only 45.5% of people we predicted to drop out actually did. Our model has a 62.5% accuracy rate and a specificity of 51.6%, meaning that we misidentified 51.6% of students who did not drop out as being at-risk. The model's F-measure is 0.594 and the AUC is 0.798, meaning that our binary classifier performs fairly well.

While our model predicts in log-odds, we interpret it on the odds scale (Table A4). Our model suggests that the odds of dropping out increases for students whose mother's have less education (compared to mother's with "Bachelor's or higher" – by +18.41% for "less than secondary," +3.67% for "secondary," +43.19% for specialized, and +451.24% for "unknown"), in debt (+69.04%), and who are male (+103.40%). However, the odds of dropping out decreases for students with higher admission grades (-1.29% for each additional point), whose tuition fees are up to date (-92.62%), and who have a scholarship (-68.300%).

4. Discussion & Limitations

The predictors with the largest impact on the predicted odds of dropping out were mother's qualification being "unknown" (compared to "Bachelor's or higher"), tuition fees being up to date, and the student being male. Additionally, the predictors with intermediate impact include mother's qualification being "specialized," if they are in debt, and if they hold a scholarship.

Our results suggest that whether a student can afford higher education may play a significant role in their risk of dropping out, given that our final model includes several measures of students' financial situations (e.g. "Tuition fees up to date", "Debtor", "Scholarship holder"). They also confirm our suspicion that unknown values may not be missing at random: the term with the largest impact is if a student's mother's education is "unknown" (+451%).

There are several limitations of our analysis. First, we were unable to examine potential correlation between quantitative and qualitative variables, potentially allowing an even more parsimonious model. For example, we might find that "admission grade" and "scholarship" are collinear, as scholarships may be based on admission grade scores. Second, there may be concern about the "Tuition fees up to date" variable because it may not be relevant to students who have not yet started college. Third, we were unable to account for all relevant variables. For example, we found that being male doubles a student's chances of dropping out; this could be because families only send daughters to college if they know they can afford it, while they send sons to college regardless. Thus, a measure of a student's family's finances may be helpful. Fourth, our final model is not cost-effective. While we correctly identified 86% of students who dropped out as being at-risk, 55% of the students we thought were at-risk did not actually drop out. If colleges have limited resources, they may value precision and thus want a model based on the 0.3 threshold (which would have lower sensitivity but greater precision). Finally, given that our dataset is from Portugal, we are hesitant to recommend its direct application to American colleges. However, we believe that following similar procedures could lead other universities to finding models that fit their schools.

In future work, we recommend building a model without the "Tuition fees up to date" variable, given its limitations discussed above. Further, we recommend distinguishing between students who were "enrolled" and "graduated" at the end of the traditional length of the student's program. While not everyone will finish a program in the traditional number of years, understanding which students are more likely to not complete programs in the traditional amount of time could provide insight into how to best support these students.

References

Knorre, Alex. “Answer to ‘Association Matrix in r.’” *Stack Overflow*, 19 May 2017, https://stackoverflow.com/a/44074895.

Martins, M. V., et al. *Predict Students’ Dropout and Academic Success*. UCI Machine Learning Repository, 2021. *DOI.org (Datacite)*, https://doi.org/10.24432/C5MC89.

Appendix

Table A1: Threshold Tuning

To optimize thresholds for the four regression models, we calculate sensitivity and F-measure and continue forward with all models with threshold 0.2 and threshold 0.3.

Threshold	Probit AIC		Probit BIC		Logit AIC		Logit BIC	
	Sensitivity	F-measure	Sensitivity	F-measure	Sensitivity	F-measure	Sensitivity	F-measure
0.2	**0.816**	**0.589**	**0.833**	**0.585**	**0.792**	**0.580**	**0.829**	**0.588**
0.3	**0.676**	**0.613**	**0.684**	**0.599**	**0.657**	**0.609**	**0.684**	**0.602**
0.4	0.529	0.582	0.495	0.569	0.527	0.581	0.493	0.567
0.5	0.459	0.568	0.423	0.540	0.454	0.559	0.425	0.541
0.6	0.394	0.526	0.362	0.503	0.396	0.528	0.362	0.503

Table A2: Performance Metrics for Model Comparison

We calculate performance metrics to compare all eight models and we choose probit and logit BIC models with threshold 0.2.

	Threshold = 0.2				Threshold = 0.3			
	Probit AIC	Probit BIC	Logit AIC	Logit BIC	Probit AIC	Probit BIC	Logit AIC	Logit BIC
# Terms	30	10	26	10	30	10	26	10
Accuracy	0.643	0.630	0.640	0.636	0.733	0.713	0.735	0.717
Specificity	0.563	0.537	0.571	0.548	0.759	0.727	0.771	0.732
Sensitivity	**0.816**	**0.833**	**0.792**	**0.829**	**0.676**	**0.684**	**0.657**	**0.684**
Precision	0.460	0.451	0.457	0.456	0.561	0.533	0.567	0.538
F-measure	0.589	0.585	0.580	0.588	0.613	0.599	0.609	0.602
AUC	0.790	0.793	0.788	0.793	0.790	0.793	0.788	0.793

Table A3: Performance Metrics of Final Model
The performance metrics of the final model (logistic BIC model with threshold of 0.2) were re-calculated with the full dataset.

	Logit BIC
# Terms	10
Accuracy	0.625
Specificity	0.516
Sensitivity	**0.856**
Precision	0.454
F-measure	0.594
AUC	0.798

Table A4: Impact of Predictors on Odds for Final Model
For the final model, the predictors with highest impact on the predicted odds of dropping out include mother's qualification being "unknown", tuition fees up to date, and the student being male.

Predictor	**Coefficient**	**Interpretation: expect odds of dropping out to ...**	
Mother's qualification - less than secondary	0.169	+ 18.41%	compared to someone whose mother has a bachelor's or higher
Mother's qualification - secondary	0.036	+ 3.67%	
Mother's qualification - specialized	0.359	+ 43.19%	
Mother's qualification - unknown	**1.707**	**+ 451.24%**	
Admission grade	-0.013	- 1.29%	
Debtor - yes	0.525	+ 69.04%, compared to someone who is not in debt	
Tuition fees up to date - yes	**-2.606**	**- 92.62%, compared to someone not up to date**	
Gender - male	**0.710**	**+ 103.40%, compared to someone who is female**	
Scholarship holder - yes	-1.149	- 68.30%, compared to someone without a scholarship	
Age at enrollment	0.040	+ 4.08%	

Figure A5: Multicollinearity for Quantitative Variables
From VIF screening with threshold 5, there are no multicollinear quantitative variables to report.

```
> vifstep(x=cleanv2.quant, th=5)
No variable from the 6 input variables has collinearity problem.

The linear correlation coefficients ranges between:
min correlation ( GDP ~ Admission.grade ):  -0.01960722
max correlation ( Admission.grade ~ Previous.qualification..grade. ):  0.5819925

---------- VIFs of the remained variables --------
                       Variables      VIF
1 Previous.qualification..grade. 1.540951
2                Admission.grade 1.518643
3             Age.at.enrollment 1.020520
4             Unemployment.rate 1.135811
5                Inflation.rate 1.020997
6                           GDP 1.154923
```

Figure A6: Probit and Logistic BIC Model on Train Dataset
Probit BIC model (left) and logistic BIC model (right) have the same terms and direction of association with slightly different coefficient estimates.

```
> summary(probit_BIC)

Call:
glm(formula = Target ~ Mother.s.qualification + Admission.grade +
    Debtor + Tuition.fees.up.to.date + Gender + Scholarship.holder +
    Age.at.enrollment, family = binomial(link = "probit"), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4035  -0.7118  -0.5453   0.5457   2.4975

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                           1.056580   0.277242   3.811 0.000138 ***
Mother.s.qualificationless than secondary  0.024236   0.078707   0.308 0.758136
Mother.s.qualificationsecondary          -0.030867   0.087387  -0.353 0.723920
Mother.s.qualificationspecialized         0.094759   0.332465   0.285 0.775630
Mother.s.qualificationunknown             0.908814   0.171445   5.301 1.15e-07 ***
Admission.grade                          -0.007148   0.001848  -3.868 0.000110 ***
Debtor1                                   0.294598   0.088364   3.334 0.000856 ***
Tuition.fees.up.to.date1                 -1.578208   0.096084 -16.425  < 2e-16 ***
Gender1                                   0.442568   0.054802   8.076 6.71e-16 ***
Scholarship.holder1                      -0.577378   0.070499  -8.190 2.61e-16 ***
Age.at.enrollment                         0.027567   0.003662   7.528 5.17e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> summary(logit_BIC)

Call:
glm(formula = Target ~ Mother.s.qualification + Admission.grade +
    Debtor + Tuition.fees.up.to.date + Gender + Scholarship.holder +
    Age.at.enrollment, family = binomial(link = "logit"), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3588  -0.7087  -0.5425   0.5385   2.4451

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                           1.895440   0.478676   3.960 7.50e-05 ***
Mother.s.qualificationless than secondary  0.046536   0.135034   0.345 0.730379
Mother.s.qualificationsecondary          -0.058973   0.150684  -0.391 0.695524
Mother.s.qualificationspecialized         0.141629   0.562281   0.252 0.801131
Mother.s.qualificationunknown             1.508758   0.288828   5.224 1.75e-07 ***
Admission.grade                          -0.012652   0.003183  -3.975 7.04e-05 ***
Debtor1                                   0.516650   0.150598   3.431 0.000602 ***
Tuition.fees.up.to.date1                 -2.652312   0.173323 -15.303  < 2e-16 ***
Gender1                                   0.765455   0.093916   8.150 3.63e-16 ***
Scholarship.holder1                      -1.040471   0.130559  -7.969 1.60e-15 ***
Age.at.enrollment                         0.045144   0.006162   7.326 2.36e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure A7: Delta Deviance Plot for Logistic BIC
From the delta deviance plot for the logistic BIC model (i.e. the final model), no points were influential.

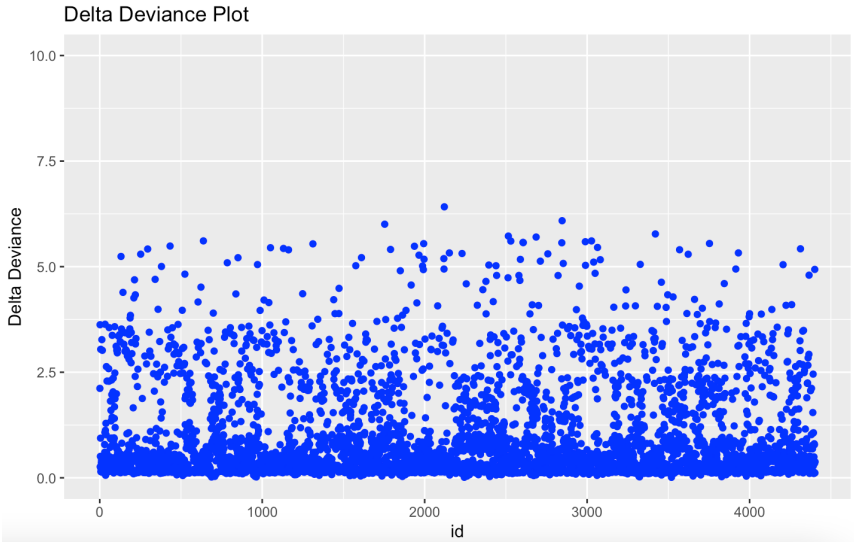

Figure A8: Regression Diagnostics for Logistic BIC
Regression diagnostics for the logistic BIC model (i.e. the final model) shows heavier tails than normal and identifies outliers (which do not end up being influential).

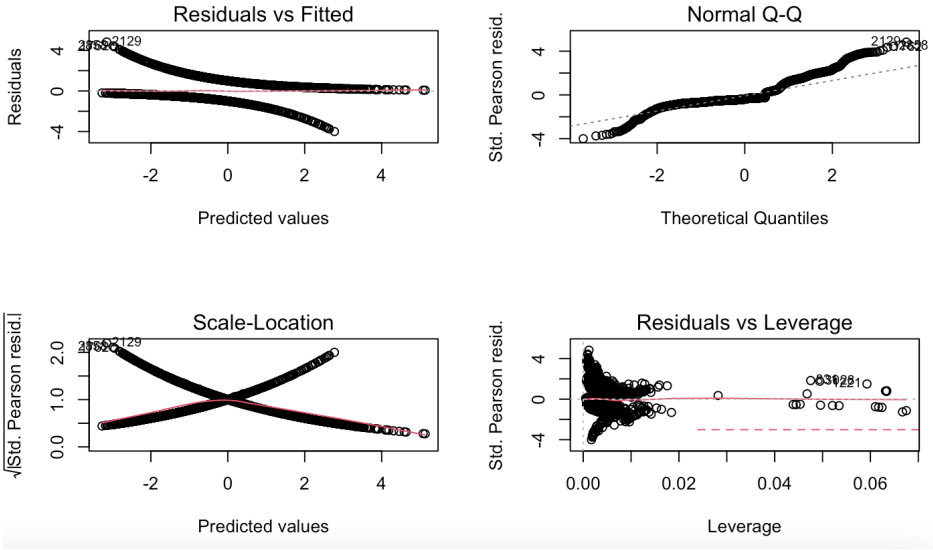

Figure A9: Final Model

The summary output for the chosen final model after refitting on the full dataset is shown below.

```
> summary(best_model)

Call:
glm(formula = Target ~ Mother.s.qualification + Admission.grade +
    Debtor + Tuition.fees.up.to.date + Gender + Scholarship.holder +
    Age.at.enrollment, family = binomial(link = "logit"), data = cleanv3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3800  -0.7240  -0.5232   0.5359   2.5262

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                           1.900110   0.394012   4.822 1.42e-06 ***
Mother.s.qualificationless than secondary  0.169034   0.114616   1.475    0.140
Mother.s.qualificationsecondary       0.035623   0.128525   0.277    0.782
Mother.s.qualificationspecialized     0.358901   0.520142   0.690    0.490
Mother.s.qualificationunknown         1.706567   0.246787   6.915 4.67e-12 ***
Admission.grade                      -0.012572   0.002643  -4.756 1.97e-06 ***
Debtor1                               0.524900   0.127652   4.112 3.92e-05 ***
Tuition.fees.up.to.date1             -2.606476   0.144853 -17.994  < 2e-16 ***
Gender1                               0.710368   0.078431   9.057  < 2e-16 ***
Scholarship.holder1                  -1.148716   0.110326 -10.412  < 2e-16 ***
Age.at.enrollment                     0.039932   0.004988   8.006 1.18e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5521.8  on 4402  degrees of freedom
Residual deviance: 4242.4  on 4392  degrees of freedom
AIC: 4264.4

Number of Fisher Scoring iterations: 5
```

Figure A10: ROC Curve for Final Model

The ROC curve for the final model after refitting on the full dataset is shown below.