STATS 503 GROUP PROJECT

SETTASIT CHITPHENTOM, TINGXIAO FU, NATHAN ROSENFELD

1. Introduction

With higher education becoming more important in today's world, it is essential to know what makes students stay in school. A student who fails to complete a degree program represents, not only a loss for him or herself - who is stuck with the debt and time expended with no degree to show for it - but for the University which may have allocated that place for him or her and whose funding may depend on completion rates.

The idea of the project is whether there is some way to accurately predict which students are "at risk" of dropping out. If so, perhaps such students could be singled out for intervention to prevent that from happening. Thus, the goal of this project is to develop machine learning classification models that can predict whether a student is likely to drop out of higher education.

We use a dataset that includes various predictors, such as demographics, students' academic performances, and financial aid. We use both decision tree model and the logistic regression model to predict whether a student will dropout. We also get some hints for the motivations for students to stay in school or abandon their studies for various disciplines.

This report is organized as following: Section 2 is about the dataset and exploratory data analysis; Section 3 for the model; Section 4 is the conclusion and Section 5 contains potential errors and problems.

2. Data

2.1. **About the Data:** The data was obtained from Zenodo [1].

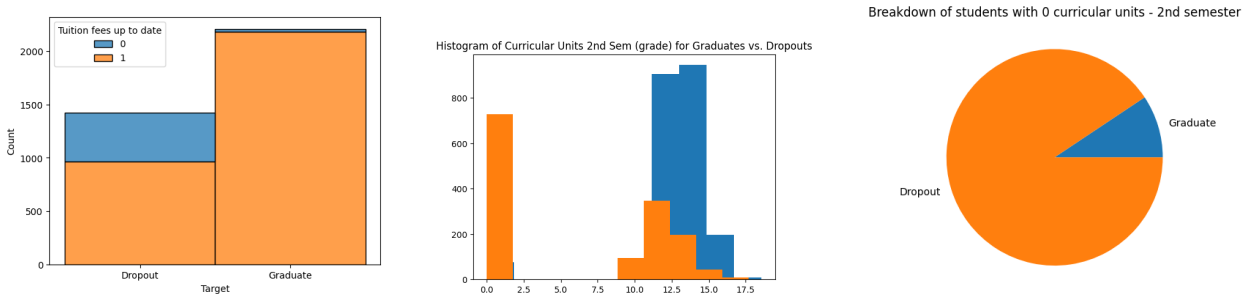
It contains data - both quantitative and categorical - regarding 4424 college students in Portugal. Each student is given one of three outcomes: `Graduate, Enrolled,` and `Dropout`. For this study, we concentrate on the 3630 who were classified either `Graduate` or `Dropout`. See Table 2 in Appendix A for the complete list of variables. For a description of the various category values in the categorical variables, see [1]

In order to make the data more useful, the categories relating to the student's and their parents' educational attainments were consolidated. As provided, the variable `Previous qualification` had 17 categories, and the variables `Mother's qualification` and `Father's qualification` each had 34 categories. Given that the data were categorical and couldn't be treated as quantitative, it was decided to consolidate each of the variables into one of the following categories:

- No educational background (only applies to parents)
- Didn't complete high school
- Completed high school
- Some college or professional training
- Completed a bachelor's degree or more.

One minor detail: The variable `International` was dropped for this analysis because `Nationality` is already one of the categories. No student classified as `International` had Portuguese nationality. Conversely, no non-international student had any nationality other than Portuguese. So, the `International` category was redundant for this analysis.

2.2. **Some idiosyncrasies:** Some variables had a very clear relationship with the likelihood of a student being classified `Dropout`. Students whose tuition is not up-to-date are dramatically more likely to be classified `Dropout`. Students who have 0 credited classes from the 2^{nd} semester are overwhelmingly more likely to dropout than to graduate (Figure 1).



((A)) Stacked bar of `Tuition up to date`

((B)) Histogram of Unit in 2nd term

((C)) Pie chart of students with 0 units in 2nd term

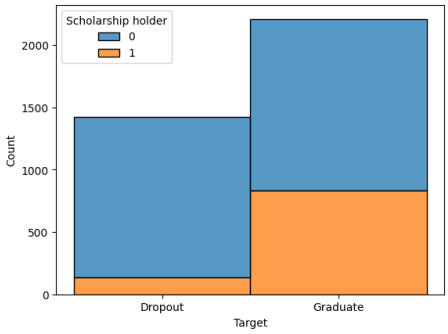
FIGURE 1. Figure for Some Idiosyncrasies

The dramatic scale of this relationship raises the question of whether we should include these variables in the model at all. A student who was credited with no credits in the previous semester may have already dropped out (in a de-facto way) during the previous semester. There may be some "lag-time" before a student is formally recorded as having dropped out. Similarly, a student who is interested in registering for the next semester's classes needs, as a prerequisite, to get caught up in tuition payments. A student who is not re-registering doesn't have this impetus to get caught up.

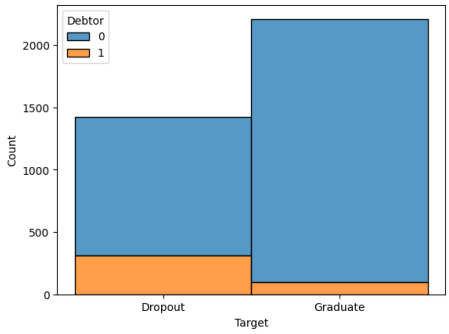
This raises the question: If we are interested in predicting a student's likelihood of dropping out, presumably in order to target him/her for extra assistance to encourage him/her to complete the program, a predictor which means that the student has already de-facto dropped out may not be very useful. For this reason, it seems reasonable to create a model which leaves out these predictors.

2.3. **Some clear relationships:** Some variables, upon EDA, showed a clear relationship with the `Target` outcome.

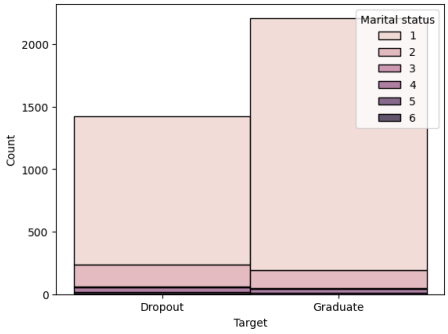
- As shown in Figure 2 (a), scholarship recipients are much less likely to drop out. This may be because of selection-effects, that better students are more likely to be scholarship recipients. Or, it may be because the reduced financial strain makes them better able to finish.
- Students who are in debt (Figure 2 (b)) or `Married` or `Widowed` are more likely to drop out (Figure 2 (c)). This seems intuitive, as students who are more financially stressed may not be able to complete the program.
- Students who have more credits, both attempted and earned, in previous semesters are more likely to graduate (Figure 2 (d)).



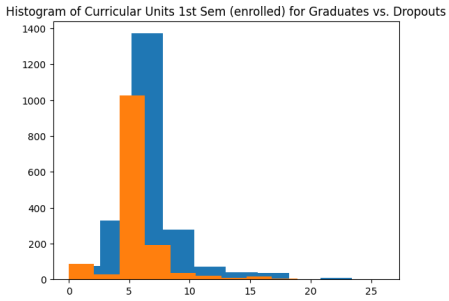
((A)) Stacked bar of Scholarship



((B)) Stacked bar of Debator



((C)) Stacked bar of Marital Status



((D)) Histogram of Curricular Units 1st sem

FIGURE 2. Figures of some Clear Relationships

2.4. **Variable correlation:** The correlation matrix (Figure 3) shows:

- The predictors based on the student's previous academics are highly correlated.
- The macroeconomic predictors `GDP` and `Unemployment` are negatively correlated (understandably so).
- Otherwise, there was little correlation amongst the various demographic variables and between them and the academic predictors.

3. MODEL

3.1. **Model Considerations.** The considerations for this study are:

- There are many categorical predictors. Some of them have many categories. Using one-hot encoding, the model will be highly multi-dimensional.
- A model should capture any interaction between the predictors, but we have no premonition regarding how this interaction might be modeled.
- Given the above considerations, a tree-based model would seem like the best way to find significant predictors.
- To allow for interpretability, we can use a logistical regression model - consisting of those predictors which seemed significant in the tree-based model.

The dataset is divided into training and evaluation set using 75% and 25% random split.

3.2. **Random Forest.** Preprocessing:

- As mentioned in section 2, the predictor `Tuition up to date` was dropped, as it may not be helpful for the purpose of the study.

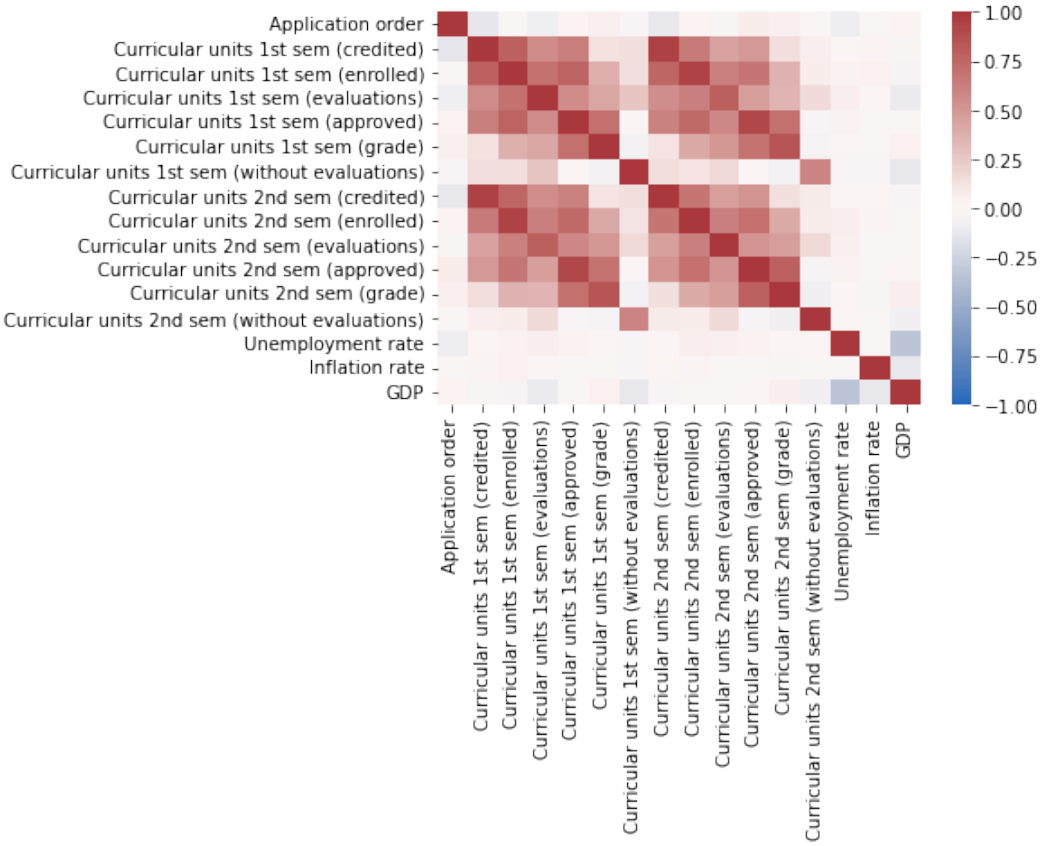
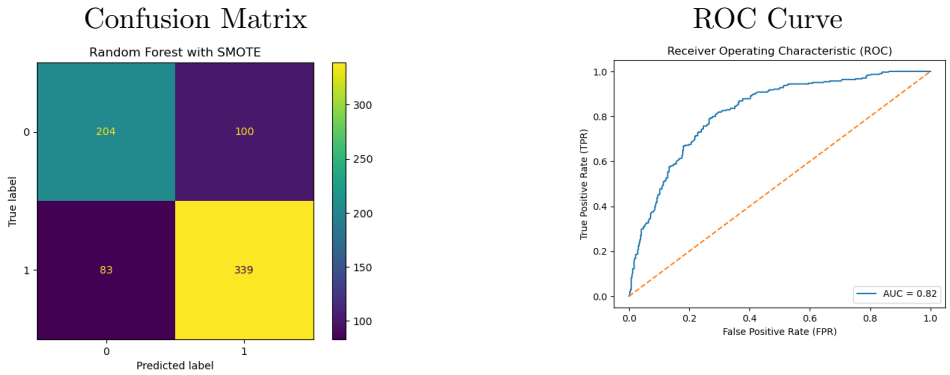


FIGURE 3. Correlation Matrix

- Many correlated variables involving academic record were dropped.
- Since the data are imbalanced, an oversampling algorithm was used to better balance the errors.

FIGURE 4. Random Forest Model



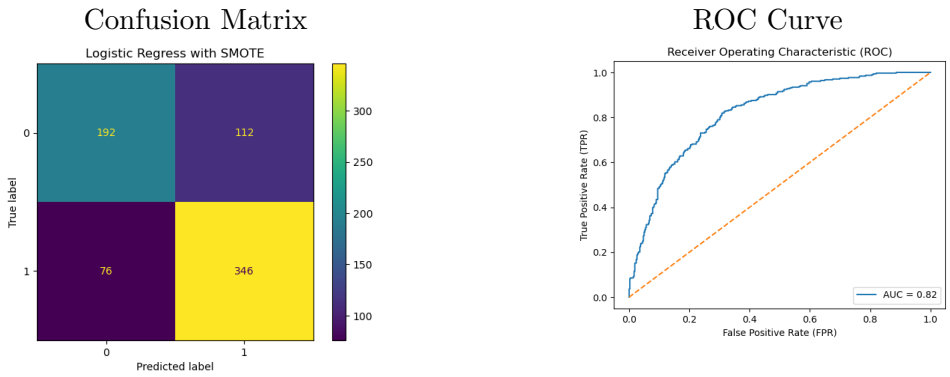
Where that was done and a random-forest algorithm applied, it yielded the results shown in Figure 4. The overall accuracy rating in that iteration was 0.737.

3.3. **Logistic Regression.** Using the features that were listed as one of the 10 most significant in Random Forest using mean increase in impurity and mean decrease in accuracy (total of 14 features), we apply logistic regression to add some interpretation to the previous results. The variables used were: `Age at enrollment`, `Application mode`, `Course`, `Curricular units 1st sem (enrolled)`, `Curricular units 2nd sem (enrolled)`, `Debtor`, `Father's occupation`, `GDP`, `Gender`, `Inflation rate`, `Mother's occupation`, `Scholarship holder`, `Unemployment rate`, and `Target` as output. Of which, for the categorical predictors, the regression was done relative to Application: **1st phase—general contingent**, Course: **Biofuel Production Technologies**, **NO** Debt, Parents' Occupation: **students**, and **NO** Scholarship.

To tune the hyper-parameters, `GridSearchCV` is used with 5-fold cross validation to find optimal `C` and `Solver` which the optimal hypter-parameters are `C = 6.5` and `Solver = 'sag'`

The results are as shown in Figure 5. Using the test-data left aside, the model had a total accuracy rate of .741 and a balanced accuracy score of .726.

FIGURE 5. Logistical Regression Model



Under this model, the most significant coefficients (having greatest absolute value) are shown in Table 1. Since this regression was done against the likelihood if a student is going to graduate. The higher β implies more likely for a student to graduate.

TABLE 1. Most negative and most positive coefficients

Variable	Coefficient
Debtor-True	-1.982093
Application mode: Other Institution	-1.305551
Application mode: Holders of other higher courses	-1.229976
Application mode: Short cycle diploma holders	-0.943523
Application mode: Over 23 years old	-0.940549
Application mode: Transfer	-0.904800
Mother's occupation: (blank)	-0.882914
Application mode: Ordinance No. 612/93	-0.661791
Application mode: Change in course	-0.626255
Application mode: 3rd phase—general contingent	-0.545460

Variable	Coefficient
Mother's occupation: Other administrative support staff	5.047525
Course: Social Service (evening)	4.200172
Course: Social Service	3.788586
Course: Veterinary Nursing	3.431535
Mother's occupation:Hotel, catering, trade, and other services directors	3.398663
Course: Nursing	3.385272
Course: Communication Design	3.327599
Father's occupation: Other administrative support staff	3.275534
Course: Agronomy	3.238632
Course: Animation and Multimedia Design	3.221708

This analysis seems to reveal several predictors that weren't apparent in the EDA. Besides confirming the importance of `Debtor` as a predictor of dropping out and the significance of having a scholarship and completing more previous coursework as predictors of graduating, factors which were clear from exploratory analysis, this model shows some other significant factors. Specifically, that students with some application modes were more likely to drop out. Conversely, there are certain courses aaociated with a greater likelihood of completion.

4. Conclusion

The model shows that some significant predictors of a student's graduating are `Debtor`, `Age at enrollment`, `Scholarship holder` and any of the variables indicating previous credits earned.

It also shows that students enrolled in certain courses are less likely to dropout than others. For instance, students whose course is describes as `Social service, nursing` or `Veterinary nursing` are less likely to drop out. Some application modes are associated with greater likelihood of dropping out, such as those described as `Holders of other higher courses` or `Transfer`.

Subject to the caveats described in Section 5, the analysis seems to show how earlier academic achievement, along with certain economic distress factors, can be used to better identify students at-risk of dropping out. This analysis can be further sharpened with knowledge of the student's course of study and sociological background. Obviously, the socioeconomic dynamics may differ from country to country, and any actionable model would be best developed in the country for which it is intended.

5. Potential Errors and Improvements

Probably the biggest potential flaw in this analysis would be the result of its many variables of many categories each. The variable `Application mode` has 18 different categories. Some of those categories have a count of only 1. The variable for `Father's occupation` has 46 different categories. Any rigorous attempt to use either of these features as predictors would need to deal with some of the categories which include very few individuals. For example, the `Application mode` 11, which has the second lowest coefficient, represents a single individual. To continue this analysis and provide actionable knowledge, it would be necessary to either find some way to consolidate some of these categories in a way that fits the sociological realities being modelled, or, if that's not possible, to calculate the p-values associated with these coefficients and limit the findings to those categories whose coefficients are statistically significant.

References

[1] Realinho, V.; Machado, J.; Baptista, L.; Martins, M.V. *Predicting Student Dropout and Academic Success.* Data 2022, 7, 146. 10.3390/data7110146.ZECjJOyZPrB.

6. Appendix A - Tables

Table 2. Variables in the Dataset

Class of Attribute	**Attribute**	**Type**
Demographic Data	Marital status	Numeric/discrete
	Nationality	Numeric/discrete
	Displaced	Numeric/binary
	Gender	Numeric/binary
	Age at enrollment	Numeric/discrete
	International	Numeric/binary
Socioeconomic Data	Mother's qualification	Numeric/discrete
	Father's qualification	Numeric/discrete
	Mother's occupation	Numeric/discrete
	Father's occupation	Numeric/discrete
	Educational special needs	Numeric/binary
	Debtor	Numeric/binary
	Tuition fees up to date	Numeric/binary
	Scholarship holder	Numeric/binary
Macroeconomic Data	Unemployment rate	Numeric/continuous
	Inflation rate	Numeric/continuous
	GDP	Numeric/continuous
Academic data at enrollment	Application mode	Numeric/discrete
	Application order	Numeric/ordinal
	Course	Numeric/discrete
	Daytime/evening attendance	Numeric/binary
	Previous qualification	Numeric/discrete
Academic data at end of 1st semester	Curricular units 1st sem (credited)	Numeric/discrete
	Curricular units 1st sem (enrolled)	Numeric/discrete
	Curricular units 1st sem (evaluations)	Numeric/discrete
	Curricular units 1st sem (approved)	Numeric/discrete
	Curricular units 1st sem (grade)	Numeric/continuous
	Curricular units 1st sem (without evaluations)	Numeric/discrete
Academic data at end of 2nd semester	Curricular units 2nd sem (credited)	Numeric/discrete
	Curricular units 2nd sem (enrolled)	Numeric/discrete
	Curricular units 2nd sem (evaluations)	Numeric/discrete
	Curricular units 2nd sem (approved)	Numeric/discrete
	Curricular units 2nd sem (grade)	Numeric/continuous
	Curricular units 2nd sem (without evaluations)	Numeric/discrete
Outcome	Target	Categorical