Final Report on Predicting Undergraduate's Academic Success in College

Aditya Vinodh,
Sanjana Vakacherla
STSCI 1380

Introduction

Undergraduate students enrolled in universities around the world face strenuous academic challenges in addition to the already-demanding task of maintaining a balanced social life and sustainable mental health. According to sources, 40% of students in American colleges drop out every year. It's no doubt that universities are yet to take effective measures to by providing better resources for students to cope with the everyday rigor of college life.

To tackle this critical issue, the first question we aimed to answer was if there were any valuable pre-existing associations within variables that could broaden our understanding of how students from different backgrounds can be equally equipped to successfully navigate college coursework. We then examined all academic, social, economic, and household factors together, and built a model to determine which ones in particular may be most predictive of an individual student's failure to graduate. Finally, we planned to depict how drop out rate proportions varied across the most directly influential of statistics on everyday student life. These were chosen from post-modelling analysis of the coefficients derived and their significance values.

As a whole, detailed analysis of this data would provide universities with insights on where to allocate funds and supportive measures for improving student life on campus. Several of these questions cannot be easily addressed as they require holistic analysis of large scale data at multivariate levels. In addition, numerous hidden confounding factors need to be correctly identified, and then appropriately accounted for - in order for strongly predictive conclusions to be arrived at.

To do this, we obtained a dataset from Kaggle, a service that provides open source files for data science research relevant to a wide variety of fields. The raw dataset used consists of 35 columns describing demographic data, socioeconomic factors, and academic performance for 4424 observations. The unit of observation here is 1 student - either listed under enrolled, graduated, or dropped out status (serving as the response variable), each with 34 potential predictor variables. The student's enrollment status was collected under a 'Target' column, which was **originally** distributed as 50% having graduated, 32% having dropped out, and 18% still enrolled.

We began preprocessing the original file by filtering out individuals who were currently enrolled as there was no definitive response that could be modeled from this. Any rows with missing data and columns that had sufficient categorical descriptions (such as Application mode, Application order, Previous qualification) were also removed. The new sample of students after removal of empty data and the ‘Enrolled’ students turned out to consist of 1421 dropped out students and 2209 graduated ones. Following this, the distribution of all numerical variables were summarized in terms of central spread measures, and key factors were chosen in bivariate pairs to explore interactions with each other. A logistic regression model was built to predict the log odds of any student dropping out based on input of 31 fields. We also went 1 step further and optimized 9 different confusion matrices to determine the decision threshold rendering the highest accuracy of prediction.

A primary finding we obtained was that students in debt who were given scholarship opportunities were less likely to drop out, in comparison to those who weren’t scholarship holders. Based on comparison of p-values and coefficients from the logistic regression model, the most predictive factor turned out to be approved 1st and 2nd semester credits and their respective grades. The most optimal cut-off value for classification of students as Drop Out or Graduate outcomes was 0.5, suggesting that any probability output above this is indicative that the student would be unsuccessful in completing their undergraduate degree. We discuss all these statistical results in detail through the paper, and how they can be used as relevant reference for organizations planning to address this critical nation-wide issue.

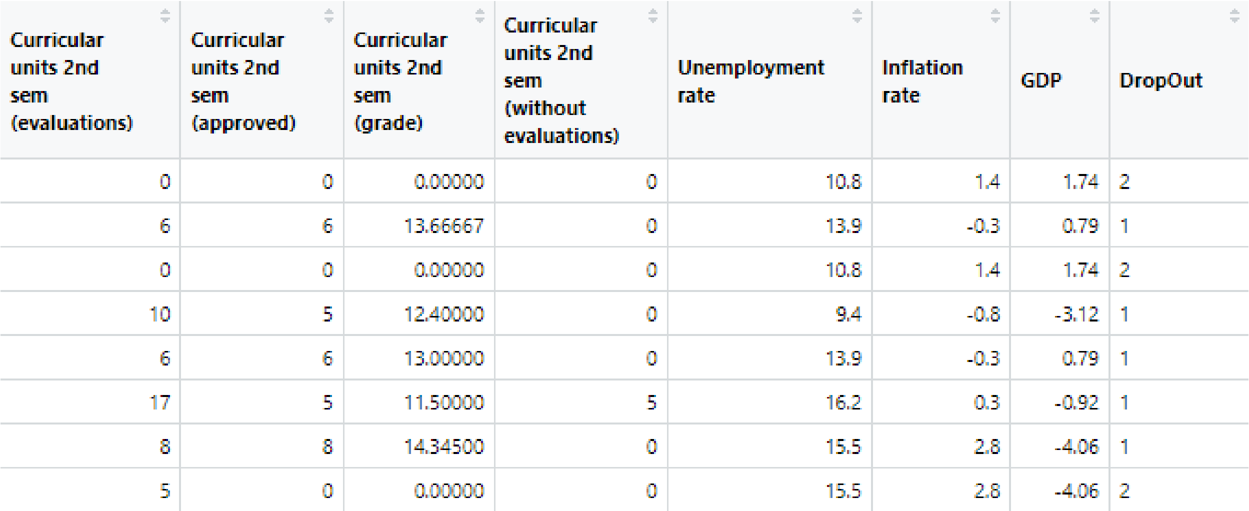
One limitation of our report was that several of the variables were accompanied by insufficient descriptions due to nominal categorization into levels, and therefore could not be used for analysis. Furthermore, there was certainly a lack of representation, in terms of the different types of students, affecting the external validity of results obtained (especially for International, Special Need-students).

Data Description and Preprocessing

The dataset that we used is a description of different demographic and academic factors attributing to a sample of students, each of which may cause them to drop out of college. Our dataset contained 35 different columns and 4,424 rows indicating 4,424 observations of students. Our columns contained both categorical and numerical information. The 35 columns represented statistics such as a student’s marital status, application method, application order, course, whether the student attended class in the morning or evening hours, whether the student had any previous qualification, their nationality, the occupation and qualifications of the parents of a student, whether the student is displaced, needs special needs, is a debtor, whether their tuition fees are up to date, their gender, whether they posses a scholarship or not, age, and whether they’re an international student or not. These were all the categorical variables that were included. Numerical data includes the credits credited, enrolled in, evaluated, and approved for each semester, along with the grade obtained in a semester. The dataset also includes the current unemployment rate, inflation rate, and gross domestic product (GDP) at the time of application. The unemployment rate, inflation rate, and GDP are calculated using each given student’s home country’s standard methods of calculation of these economic metrics at these given points in time when each student applied.

We wanted to only look at students that had graduated or dropped out, and as a result of this, during our preprocessing of our dataset, we filtered out all students who are currently enrolled in university. In order to perform efficient analysis on the dataset, we also created a new column using the mutate function, which was a binary representation of whether a student graduated or not. If a student has graduated, we set this value to 0, and if a student dropped out we set this to 1. We then used the select function in order to remove our “Target” column, which contained whether a student was enrolled, graduated, or had already dropped out. This column was no longer necessary as we now have a binary representation of this column present, making numerical analysis of this table much easier, and making the presence of this column redundant information. As a preliminary step, we calculated the correlation between each of the remaining columns against the “Dropout” column, in order to understand the correlation between each variable and whether a student dropped out or not. We also created another column representing the average number of credits graded between the two semesters, by simply taking the average of semester 1 and semester 2’s graded credits columns, and storing this value in a new column.

From our preliminary visualizations of the data, we can see that certain variables are heavily skewed left or right. For example in the Age at Enrollment plot in figure 4, it can be seen that ages mostly fall within the [0, 20] bar, indicating that perhaps most students are college aged or younger. This definitely was something to keep in mind whilst looking at the dropout rates. We also had much more enrolled students than students who had dropped out, as can be seen in figure 5. To reduce the variability that the enrolled students’ data would have caused in our analysis, we had filtered out all enrolled students, as part of our preprocessing for this dataset.



Curricular units 2nd sem (evaluations)	Curricular units 2nd sem (approved)	Curricular units 2nd sem (grade)	Curricular units 2nd sem (without evaluations)	Unemployment rate	Inflation rate	GDP	DropOut
0	0	0.00000	0	10.8	1.4	1.74	2
6	6	13.66667	0	13.9	-0.3	0.79	1
0	0	0.00000	0	10.8	1.4	1.74	2
10	5	12.40000	0	9.4	-0.8	-3.12	1
6	6	13.00000	0	13.9	-0.3	0.79	1
17	5	11.50000	5	16.2	0.3	-0.92	1
8	8	14.34500	0	15.5	2.8	-4.06	1
5	0	0.00000	0	15.5	2.8	-4.06	2

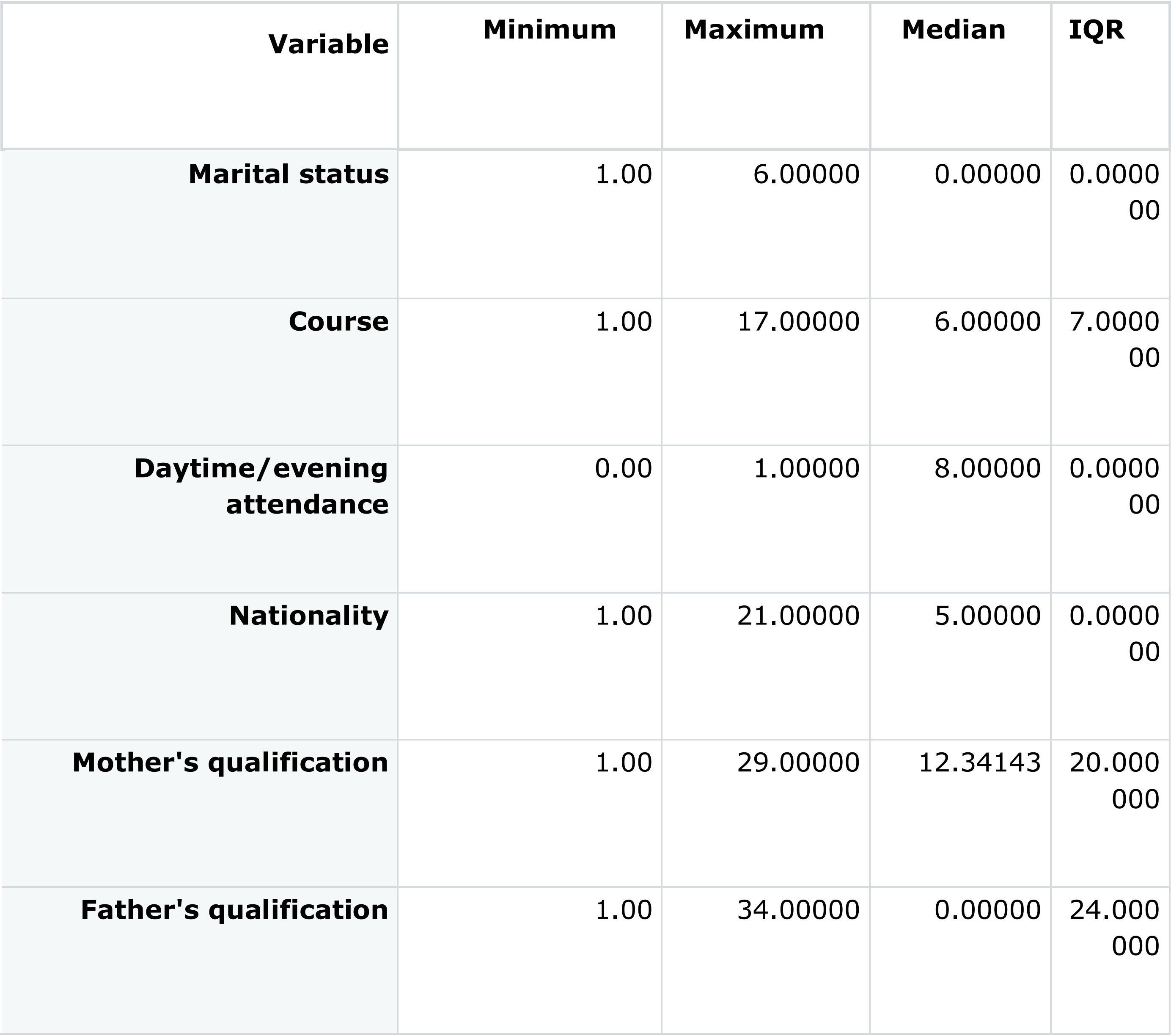
Figure 1: A snapshot of rows and columns from the dataset

Variable No.	Variable Name	Categorical/ Numerical	Variable Definition
1	Marital Status	Cat	The marital status of the student

2	Course	Cat	The course taken by the student
3	Daytime/evening attendance	Cat	Whether the student attends class in the morning or evening
4	Nationality	Cat	The nationality of the student
5	Mother's qualification	Cat	The highest qualification of the student’s mother
6	Father's qualification	Cat	The highest qualification of the student’s mother
7	Mother's occupation	Cat	The occupation of the student’s mother
8	Father's occupation	Cat	The occupation of the student’s mother
9	Displaced	Cat	Whether the student is a displaced student or not
10	Educational special needs	Cat	Whether the student is need of any educational special needs
11	Debtor	Cat	Whether the student is a debtor
12	Tuition fees up to date	Cat	Whether the student’s fees are up to date
13	Gender	Cat	The gender of the student
14	Scholarship holder	Cat	Whether the student holds a scholarship or not
15	Age at enrollment	Num	The age of the student upon enrollment
16	International	Cat	Whether the student is considered an international student or not
17	Curricular units 1st sem (credited)	Num	The total units the student was credited for in in the first semester
18	Curricular units 1st sem (enrolled)	Num	The total units the student enrolled in in the first semester

19	Curricular units 1st sem (evaluations)	Num	The total curricular units taken by the student that the student received evaluations for in the first semester
20	Curricular units 1st sem (approved)	Num	The total approved credits the student completed in the first semester
21	Curricular units 1st sem (grade)	Num	The total units the student received a grade for in in the first semester
22	Curricular units 1st sem (without evaluations)	Num	The total curricular units taken by the student that the student did not receive evaluations for in the first semester
23	Curricular units 2nd sem (credited)	Num	The total units the student was credited for in in the second semester
24	Curricular units 2nd sem (enrolled)	Num	The total units the student enrolled in in the second semester
25	Curricular units 2nd sem (evaluations)	Num	The total curricular units taken by the student that the student received evaluations for in the second semester
26	Curricular units 2nd sem (approved)	Num	The total approved credits the student completed in the second semester
27	Curricular units 2nd sem (grade)	Num	The total units the student received a grade for in in the second semester
28	Curricular units 2nd sem (without evaluations)	Num	The total curricular units taken by the student that the student did not receive evaluations for in the second semester
29	Unemployment rate	Num	The unemployment rate in the region from which the student applied from at the time of application
30	Inflation rate	Num	The inflation rate at the time of application
31	GDP	Num	The GDP at time of application
32	AvgCreditsGrade	Num	The average credits that the student received a grade for between semester 1 and semester 2

Figure 2: A data table containing the names, variable types, and descriptions of each variable in the dataset.



Variable	Minimum	Maximum	Median	IQR
Marital status	1.00	6.00000	0.00000	0.000000
Course	1.00	17.00000	6.00000	7.000000
Daytime/evening attendance	0.00	1.00000	8.00000	0.000000
Nationality	1.00	21.00000	5.00000	0.000000
Mother's qualification	1.00	29.00000	12.34143	20.000000
Father's qualification	1.00	34.00000	0.00000	24.000000

Mother's occupation	1.00	32.00000	0.00000	5.000000
Father's occupation	1.00	46.00000	6.00000	5.000000
Educational special needs	0.00	1.00000	8.00000	0.000000
Debtor	0.00	1.00000	5.00000	0.000000
Tuition fees up to date	0.00	1.00000	12.33333	0.000000
Scholarship holder	0.00	1.00000	0.00000	1.000000
Age at enrollment	17.00	70.00000	11.10000	6.000000
Curricular units 1st sem (credited)	0.00	20.00000	1.40000	0.000000
Curricular units 1st sem (enrolled)	0.00	26.00000	0.32000	2.000000

Curricular units 1st sem (evaluations)	0.00	45.00000	0.00000	4.000000
Curricular units 1st sem (approved)	0.00	26.00000	6.00000	3.000000
Curricular units 1st sem (grade)	0.00	18.87500	8.00000	2.500000
Curricular units 1st sem (without evaluations)	0.00	12.00000	5.00000	0.000000
Curricular units 2nd sem (credited)	0.00	19.00000	12.34143	0.000000
Curricular units 2nd sem (enrolled)	0.00	23.00000	0.00000	2.000000
Curricular units 2nd sem (evaluations)	0.00	33.00000	0.00000	4.000000
Curricular units 2nd sem (approved)	0.00	20.00000	6.00000	4.000000
Curricular units 2nd sem (grade)	0.00	18.57143	8.00000	2.982143

Curricular units 2nd sem (without evaluations)	0.00	12.00000	5.00000	0.000000
Unemployment rate	7.60	16.20000	12.33333	4.500000
Inflation rate	-0.80	3.70000	0.00000	2.300000
GDP	-4.06	3.51000	11.10000	3.490000
AvgCreditsGrade	0.00	18.28365	1.40000	2.600000

Figure 3: A table containing the summary statistics for each variable in the dataset

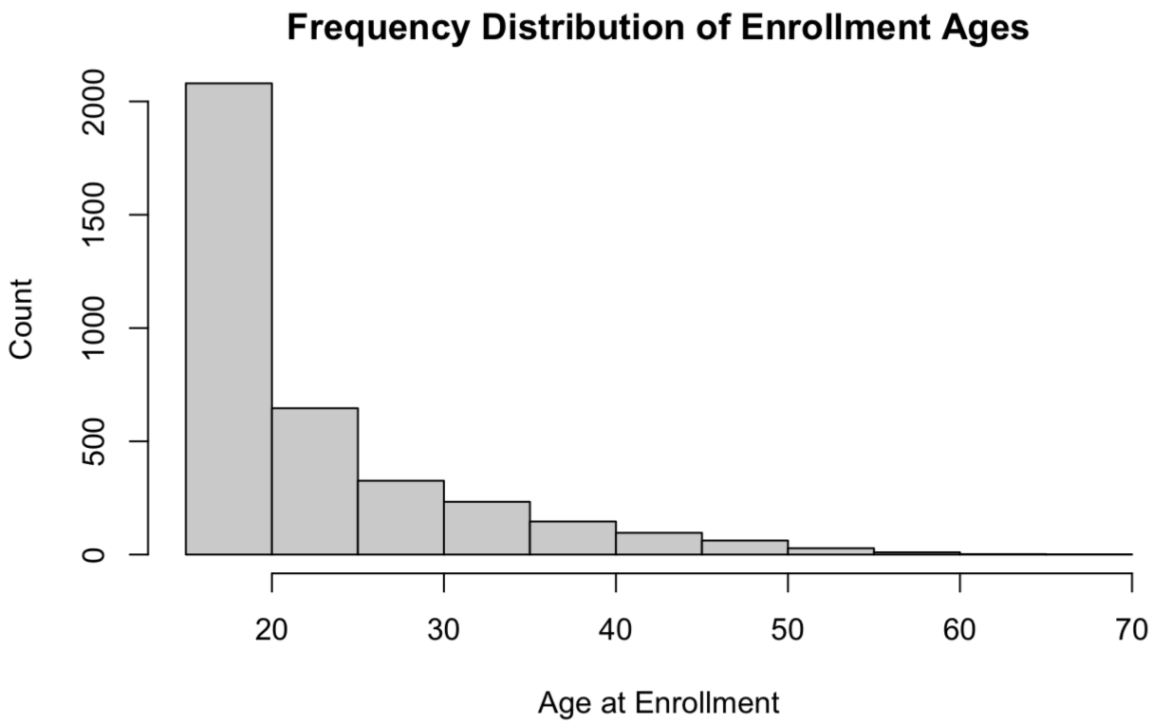


Figure 4: A histogram displaying the ranges for which the ages of students fall within. On the vertical axis we have the number of occurrences of a certain age, and on the horizontal axis we have the corresponding age. From this figure we can see that the ages are heavily skewed right, with the median falling in the [0, 20] bar.

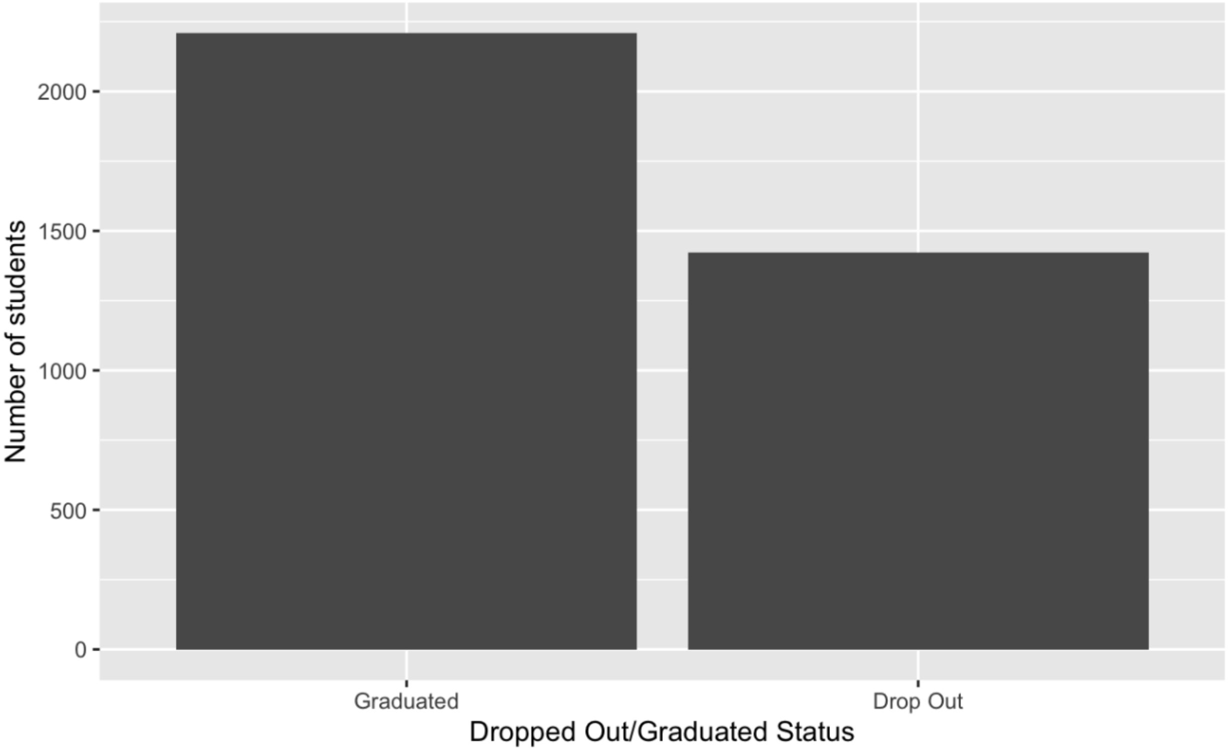


Figure 5: A bar plot displaying the number of students who have dropped out. It's clear from this figure that the dataset contains more graduated students than students who have dropped out.

Methods

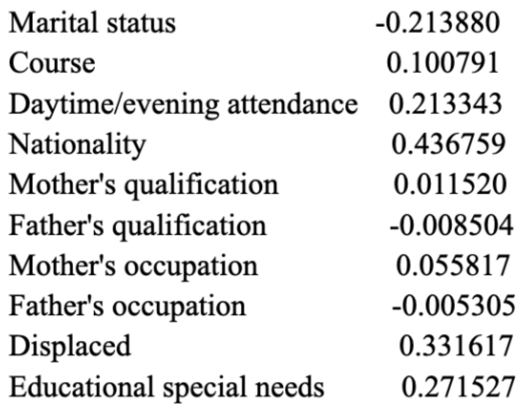
We constructed a model that would use information about students' background and current circumstances to predict whether they are likely to drop out in the future. To do so, a logistic regression technique was selected as the response variable to be predicted was categorical and final classification of output was needed. Additionally, this method reduces the chance of over-fitting, can accommodate large scale data, and provides a summary of feature importance and significance values for each variable. We used a total of 31 variables as inputs for the model, out of which 16 were demographic/socioeconomic-related, 11 academic-related and 3 described the current economic situation of the country.

The derived model calculates the log-odds of an outcome of dropping out, which can be expressed in terms of probability as:

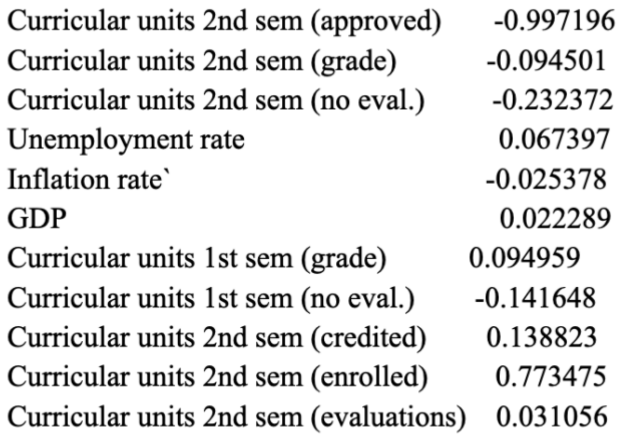
$$\log\,(\text{odds}) = \log\,(\frac{p}{1-p})$$

These 2 mathematical values share a directly proportional relationship implying that as the probability of an event occurring increases, so do its log odds. The summary of our regression model presents the following coefficients for each variable of interest:

Debtor	1.269432
Tuition fees up to date	-2.617603
Gender	0.406775
Scholarship holder	-0.734589
Age at enrollment	0.038544
International	-6.041077
Curricular units 1st sem (credited)	0.236514
Curricular units 1st sem (enrolled)	0.261976
Curricular units 1st sem (evaluations)	-0.003202
Curricular units 1st sem (approved)	-0.678168



Marital status	-0.213880
Course	0.100791
Daytime/evening attendance	0.213343
Nationality	0.436759
Mother's qualification	0.011520
Father's qualification	-0.008504
Mother's occupation	0.055817
Father's occupation	-0.005305
Displaced	0.331617
Educational special needs	0.271527



Curricular units 2nd sem (approved)	-0.997196
Curricular units 2nd sem (grade)	-0.094501
Curricular units 2nd sem (no eval.)	-0.232372
Unemployment rate	0.067397
Inflation rate`	-0.025378
GDP	0.022289
Curricular units 1st sem (grade)	0.094959
Curricular units 1st sem (no eval.)	-0.141648
Curricular units 2nd sem (credited)	0.138823
Curricular units 2nd sem (enrolled)	0.773475
Curricular units 2nd sem (evaluations)	0.031056

Figure 1: Coefficients for each variable in relation to our target (whether a student has dropped out or not) calculated as a result of logistic regression.

While it is difficult to present a compiled mathematical equation representing the relationship of ‘dropout’ logodds to the 31 predictor variables here, one can simply multiply each coefficient of a variable by its input value (whether in numerical or categorical form), and take the sum of these added along with the log odds intercept.

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Figure 2: The Log Odds Function

Here, the first term ‘β_0’ represents the y-intercept i.e. what the logodds output would be when each variable’s input is 0. ‘B_k’ denotes the coefficient of each variable, and ‘X_k’, the data provided for a particular student based on this variable ‘k’. A positive coefficient suggests that the effect of a variable increases the chances of observing a drop out while a negative one decreases these chances. Each coefficient here represents the change in log odds of observing a drop out in each student’s case when all other variables are kept constant.

We also evaluated the effectiveness of our model in its capability of determining a specific student’s probability of dropping out. First, a classification rule was used that converted the log odds into a binary outcome of either’ Graduating’ or ‘Dropping Out’. However, the ‘i’ criteria used in this needed to be set according to a logical rule. To do this, we optimized a decision threshold that outputted the highest accuracy on test data. This was done by constructing several confusion matrices for thresholds ranging from 0.1 to 0.9 using a for-loop in the R program, and taking the threshold with the greatest proportion of true positives and true negatives. This allowed us to gain insight as to what minimum level of probability a university should consider investing efforts towards a student, instead of taking an arbitrary threshold or finding one through trial error that may have led to the use of an unreliable classification technique (making inaccurate predictions. The model’s ability to correctly identify true drop-out predictions amongst falsely predicted ones, and discount students who already presented strong likelihood of graduation was then by calculating its sensitivity and specificity.

There are a few limitations with our model, however. For instance, the prominence of categorical variables and lack of sufficiently distributed data led to a number of the variables being labeled as ‘insignificant’ for their coefficient values. As seen in figure, there is more data focusing on students who graduated than ones who dropped out - revealing a potential bias in the model’s analysis of the provided data and how it may unreliably classify new data. The absence of a data distinction step into training and test sets also makes it difficult to cross-validate the prediction model derived.

Results

We observed that 17 out of 31 of the predictor variables inputted into the model are statistically significant and have a p-value of less than 0.05. 6 out of the 11 academic variables were characterized as having feature importance, while 10/16 of the demographic type and only 1 economic variable turned out to be so. Figure 1 depicts these results in detail and was used to determine which associations could be further examined, which will be discussed later.

Variable Name	Coefficient	p-value (if < 0.05)
Course	0.100791	1.45 e-08
Nationality	0.436759	4.03e-05
Mother’s Occupation	-0.055817	0.02868

Displaced	0.331617	0.0261
Debtor	1.269432	7.80e-08
Tuition Fees Up to Date	-2.617603	<2e-16
Gender	0.406775	0.00346
Scholarship Holder	-0.734589	9.19e-06
Age at Enrollment	0.035844	0.00109
International	-6.041077	5.67e-06
Credited Curricular Units (1st)	0.236514	0.02163
Approved Curricular Units (1st)	-0.678168	2e-16
Graded Curricular Units (1st)	0.094959	0.04508
Enrolled Curricular Units (2nd)	0.773475	1.14e-07
Approved Curricular Units (2nd)	-0.997196	<2e-16
Graded Curricular Units (2nd)	-0.094501	0.04671
Unemployment Rate	0.067397	0.01376

Figure 3: A table representing the logistic regression coefficients and p values corresponding to the variables that were significant at a p value cutoff of 0.05.

After our logistic regression, we were able to obtain the individual correlation between each individual variable and our target - which was whether a student dropped out or graduated. This helped us cross validate whether the coefficients from the model accurately represented how significant and in which direction the predictors had an effect on the response variable.
This is shown below:

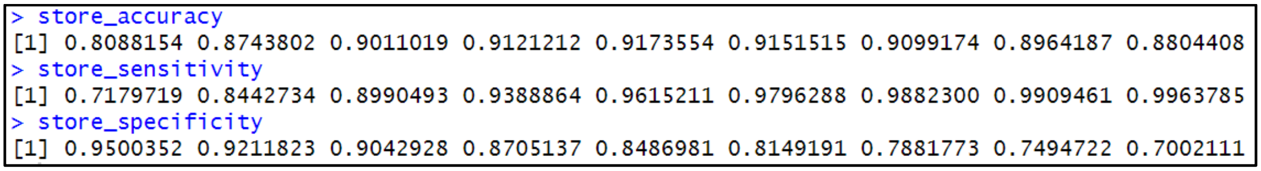
Variable	Correlation with Dropout Variable
Marital Status	0.100479066
Course	-0.006814300
Daytime/Evening Attendance	-0.084495936

Nationality	0.003822831
Mother's Qualification	0.048458563
Father's qualification	0.003849914
Mother's occupation	-0.064195026
Father's occupation	-0.073238260
Displaced	-0.126113035
Educational special needs	0.007253654
Debtor	0.267207199
Tuition fees up to date	-0.442137576
Gender	0.251954812
Scholarship holder	-0.313017663
Age at enrollment	0.267229383
International	-0.006181262
Curricular units 1st sem (credited)	-0.046900017
Curricular units 1st sem (enrolled)	-0.161073516
Curricular units 1st sem (evaluations)	-0.059786259
Curricular units 1st sem (approved)	-0.554880857
Curricular units 1st sem (grade)	-0.519927094
Curricular units 1st sem (without evaluations)	0.074642260
Curricular units 2nd sem (credited)	-0.052401971
Curricular units 2nd sem (enrolled)	-0.182896541
Curricular units 2nd sem (evaluations)	-0.119238767
Curricular units 2nd sem (approved)	-0.653995246
Curricular units 2nd sem (grade)	-0.605350126

Curricular units 2nd sem (without evaluations)	0.102686829
Unemployment rate	-0.004198105
Inflation Rate	0.030325866
GDP	-0.050260147
Average Graded Credits	-0.587407939

Figure 4: This table shows the correlation between each of our variables and whether or not a student dropped out. We know that the closer the magnitude of the correlation is to 1, the stronger the association is between that variable and dropout. The sign simply refers to whether this is a negative or positive correlation.

After determining the accuracy, sensitivity, and specificity of our model for a set of decision thresholds, we find that our accuracy is maximized for a classification of 0.5. The sensitivity continually increases as the set threshold increases, while our specificity remains the same at all levels of classification. This is seen by observing the outputted arrays containing these values which is included below. We will choose the 0.5 classification level to maximize our total accuracy.



```
> store_accuracy
[1] 0.8088154 0.8743802 0.9011019 0.9121212 0.9173554 0.9151515 0.9099174 0.8964187 0.8804408
> store_sensitivity
[1] 0.7179719 0.8442734 0.8990493 0.9388864 0.9615211 0.9796288 0.9882300 0.9909461 0.9963785
> store_specificity
[1] 0.9500352 0.9211823 0.9042928 0.8705137 0.8486981 0.8149191 0.7881773 0.7494722 0.7002111
```

Figure 5: These lists contain the calculated accuracies, specificities, and sensitivities for our model at different decision thresholds (or classification thresholds), ranging from [0.1, 0.9]. We can see that the peak accuracy is achieved at a classification level of 0.5. We will choose that as our cutoff whilst looking at the other values. At a classification level of 0.5, our accuracy, specificity, and sensitivity will be 91.7%, 96%, and 85% respectively.

At the 0.5 classification level, our model predicts drop out outcomes in students with an accuracy of 92%, a specificity of 85% and sensitivity of 96%. Therefore, we can infer that the model picks out students who would drop out 96% of the time, while it correctly classifies only 85% of all the students in cases where they graduate. This comparatively low sensitivity means that the model is better at predicting students' chances of dropping out than graduating - due to a lack of equal representation between Drop Out and Graduated students in the provided dataset itself. However, these high rates of Accuracy, Specificity, and Sensitivity from our confusion matrix serve as strong support for the model in its ability to foresee academic failure in college.

Predictor	0	1
“DropOut”	2124	215
“Graduate”	85	1206

Figure 6: The confusion matrix that shows the values used for calculating our accuracy, specificity, and sensitivity. A 0 value in our predictor indicates that a student is considered a dropout, and a 1 value indicates a student has graduated. For our predictor, we’ve assigned a 1 value to any predictor with a < 0.5 value, and a 0 to any predictor with a > 0.5 value.

To further our understanding of these associations and how college funding departments/student-support centers can help increase student retainment in their undergraduate career, we went a step ahead and attempted to investigate 3 potential associations. To ensure that each of these associations were worthwhile to study, the variables were selected based on significance levels calculated from the regression model. 2 bivariate questions were raised, 1 of which was based on a Numerical-Categorical association and the other on a Categorical-Categorical association. Finally, a multivariate question was raised to combine all 3 findings and arrive at a pattern in the data that could be used for student welfare betterment strategies. All of these potential questions considered the Dropout rates as a response/dependent variable and involved predictors that displayed the highest feature importance from each of the demographic, academic, and economic factors.

The first part delved into the association between the Average credits graded for (from Semester 1 and 2) and the dropout rates among students. The rationale behind this was to determine whether a heavy academic course load may have an effect on students’ performance, and if taking up a greater number of credits serves as a motivating or demotivating factor to continue their degree. This information can be used by college advisors and professors to accordingly mentor students in the future towards more balanced semester-planning. Since the association was of Numerical-Categorical nature, this can be summarized by comparing the mean grade for each group and visualizing the results as a side-by-side boxplot. We will first compare the mean credits student dropouts took between their 1st and 2nd semesters, vs students who graduated.

DropOut	Mean Credits Graded
0	12.670465
1	6.5777997

Figure 7: A table displaying the average graded credits per student who have dropped out (when dropout = 1), and students who’ve graduated (when dropout = 0).

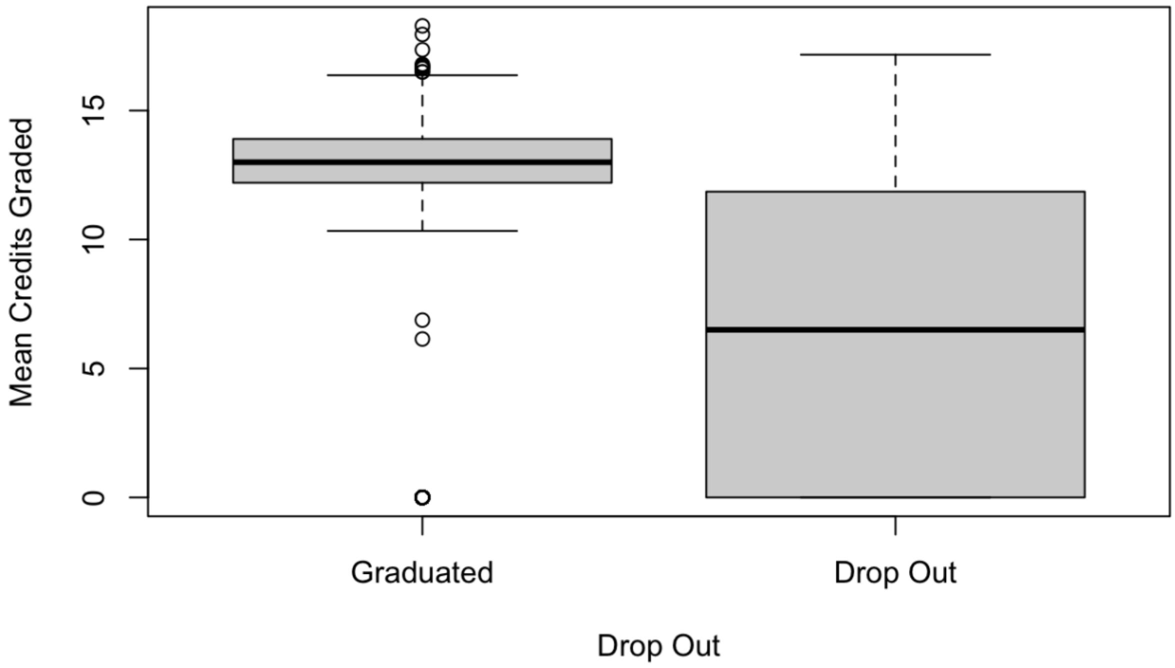


Figure 8: A box plot displaying the distribution of average graded credits amongst dropouts and graduates.

The results of this analysis show a surprising pattern - that those who drop out are more likely to be graded for fewer credits than those who have graduated. This difference is clearly shown in the table summarizing the mean credits for each group (12.67 versus 6.58) and additionally in the boxplot where the median (center of the boxplot) is significantly lower for those who dropped out. While this association may suggest that undertaking a higher number of credits pushes students to work harder and graduate successfully, it may also be linked to the role of confounding factors. For example, those who enroll in more credits for grades may also simply be more academically focused instead of this having a direct effect on drop out chances.

We then looked into the association between the status of a student as being Displaced, and dropout rates amongst them. Findings from this could be used to suggest whether additional resources need to be provided for students displaced from their home, and whether they would be more likely to succeed in college with the right support . This information can be used by admission officers while evaluating a student's socioeconomic fit from a wider pool of applicants, and by non-profit organizations while implementing ways to foster better academic/social inclusion. Since the association was of Categorical-Categorical nature, this can be summarized by comparing the proportion of students in each group and visualizing the results as a barplot.

Displaced	DropOut	Number of Students
0	0	885
0	1	752
1	0	1324
1	1	669

Figure 9: A table showing the number of students who've dropped out in relation to whether they're considered displaced or not.

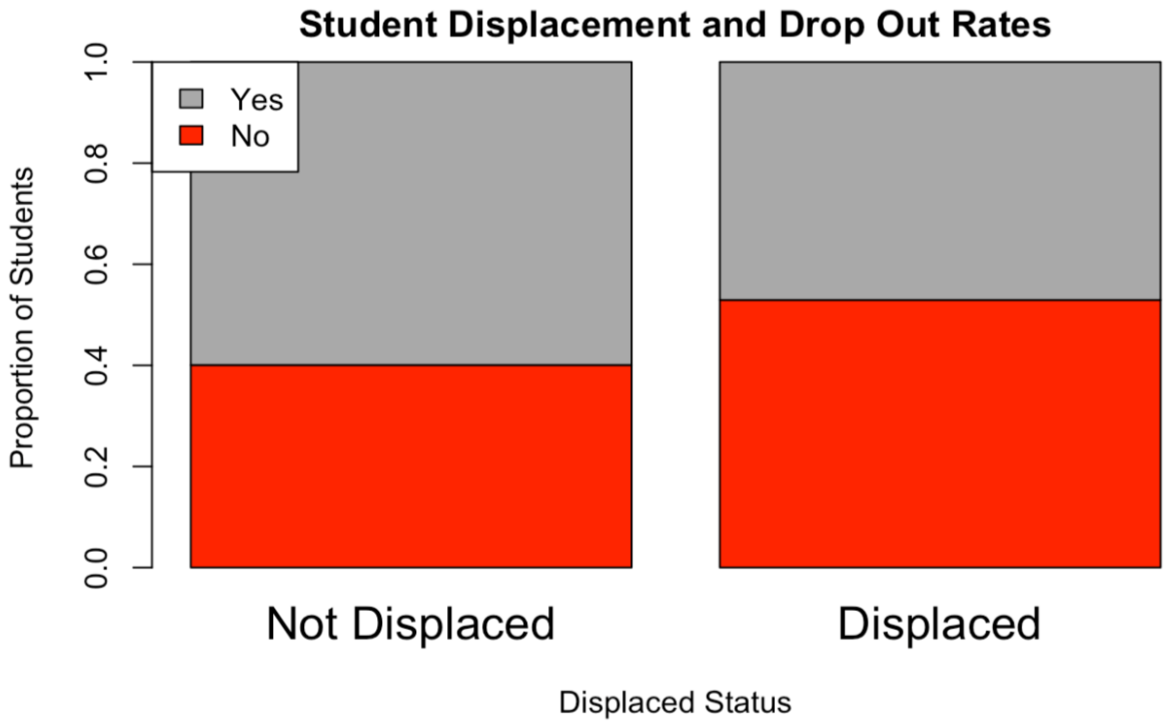


Figure 10: A stacked bar plot showing the effect of displacement on dropout rates.

The data plotted above shows an interesting finding regarding the effect of Student Displaced/Non-Displaced status on their drop out probability. That is, we see a significant difference in the proportion of dropped out students between the 'Displaced' and Non-Displaced groups. In comparison to those who are not displaced, a higher proportion of students seem to graduate successfully in the displaced category, while a lower proportion of them report drop out outcomes. This is contrary to what one would expect, as individuals evicted/removed from their residence would typically lack the same financial and educational aid as those from normal homes would. However, this is strong evidence that further opportunities should be given to such underprivileged students, and that the current program is successful in encouraging ambition and academic drive in them.

Lastly, we probed into the association between Debtor status and Dropout Rates, and how these were affected by whether one was a Scholarship holder or not. Conclusions from this analysis are reflective of the effectiveness of financial investment towards Scholarship grants, towards students struggling with tuition payment themselves. Therefore, funding departments can decide how to allocate grants and funds according to how strong the effect of scholarship is in encouraging undergraduate success. This association was of Categorical - Categorical - Categorical nature, and could be summarized through a 8x4 table of proportions under each category, while being visualized through a side-by-side barplot.

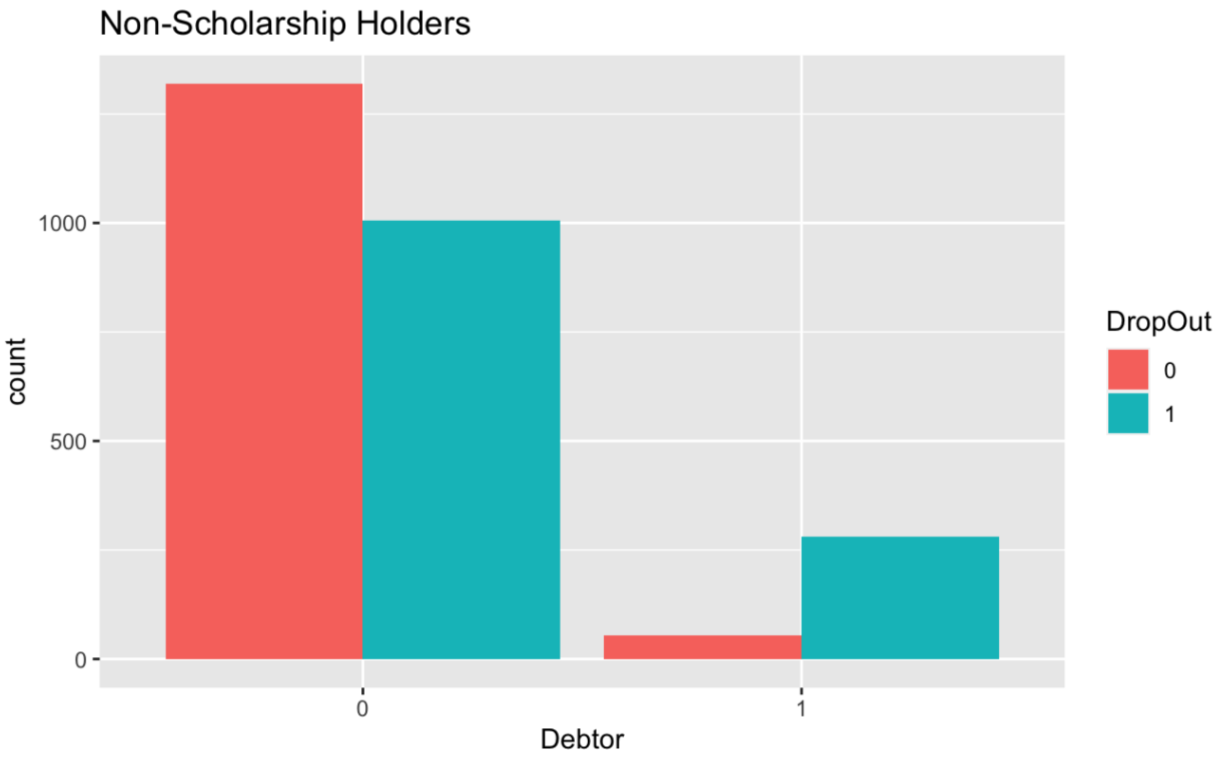


Figure 11: A grouped bar plot showing the effect of debtor status on dropout rates for non scholarship holders. The category labeled as 1 on the x-axis indicates Debtors and the 0’s represent Non-debtors.

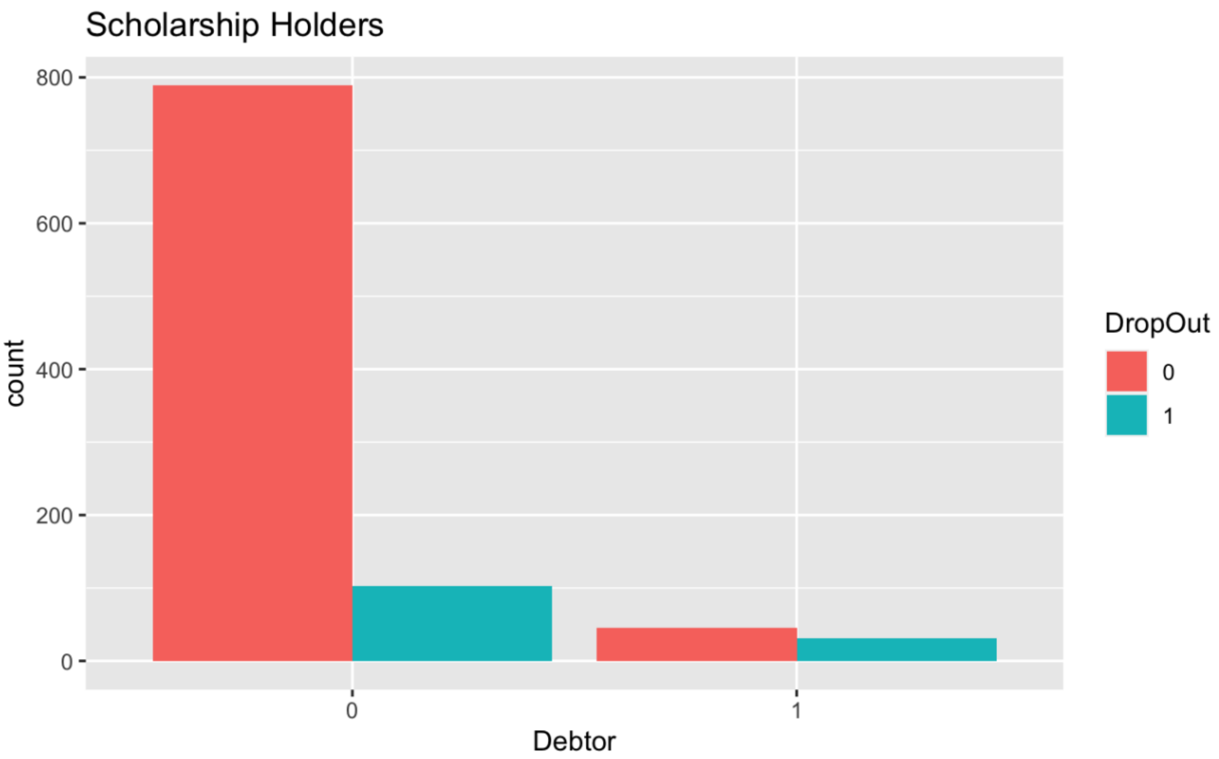


Figure 12: A grouped bar plot showing the effect of debtor status on dropout rates for scholarship holders. The category labeled as 1 on the x-axis indicates Debtors and the 0's represent Non-debtors.

The 2 color-coded figures above illustrate a distinct difference between how the proportion of Drop Out students varies across Debtor and Non-debtor categories, according to whether they are scholarship holders or not. While the first depicts how being a debtor without scholarship increases the likelihood of students dropping out over graduating, the second portrays a remarkable finding - that is, those who are scholarship holders don't show the same pattern. Instead, the ratio of Drop Out vs. Graduated students is much lower in those who are Debtors but Scholarship holders, as compared to those who are Debtors but Non-scholarship holders. This outcome of our analysis serves as strong evidence that the allocation of Scholarship to students who are financially disadvantaged motivates them to succeed in their college careers. This observed impact that merit based-financial aid has on increasing graduation rates should demonstrate that more such services need to be offered in high school or graduate-level studies for those who show promising potential to succeed.

These 3 extensions of the analysis further affirm that 4 of the 19 predictor variables labeled as significant by the proposed model can indeed be used as factors to arrive at valuable trends between each other. A similar accuracy in feature importance can also be drawn for the other variables, all according to the degree of statistical significance denoted by their p-values. Therefore, other associations can also be raised to delve deeper into underlying trends that motivate a student to work through the rigor of college coursework.

Conclusion

We developed a logistic regression model that allows us to predict the relationship between different properties of students and whether or not they are more likely to drop out of college or to graduate. By calculating the accuracy of our model in the above section of this report, we've shown that our model can accurately predict the factors that play the biggest role in whether a student drops out of college or not around 92% of the time. This model also shows for each of our 32 variables, the relationship between that variable and whether that student dropped out. If the p value outputted by the logistic regression was below 0.05, we could conclude that that variable had a significant impact on whether a student dropped out of college or not. We included all of these variables in a table above - variables that we concluded that had a significant effect on the dropout rates included: nationality, whether a student was displaced or not, whether they had a scholarship or not, and the amount of classes they took per semester. At a classification level of 0.5 our specificity and sensitivity were 96% and 85% respectively. In the context of our topic and our model, this means that for every student who is likely to drop out, our model can predict whether they are going to drop out at a success rate of 85%, and for every student who will graduate, our model can correctly predict this at a success rate of 96%. The variables that we found to be most indicative of dropping out included nationality, displacement, credits taken for a grade during semesters, and scholarship status, among other variables. We also addressed 3 associations that seemed to have the most relevance to dropout rates by visualizing associations between 4 of the most significant predictors.

With regards to the evaluation of our model, one limitation of the data used that potentially caused inaccuracies is due to the fact that only students who had either dropped out or graduated were considered (since currently enrolled students were filtered out). This meant we had to discard around 800 observations from our original 4200. This was a substantially large portion of our initial set (20%), and upon removal of these points we also had around 2200 students who had graduated and only around 1400 who had dropped out. All these imbalances could've caused skews in our data that would shift the results. Our accuracy, sensitivity, and specificity, all showed strong results, although our specificity was slightly lower than our accuracy and sensitivity. This meant that our model was better at correctly identifying whether a student would drop out than picking out those who would graduate. To increase our specificity, one future improvement could be to include a larger sample size of greater range of economic, demographic and academic factor representation. In addition, other prediction models could be looked into and compared against the logistic regression such as the Decision Tree, Classification Model, or Forecast Model. In addition, the dataset can be split into training and test data cross-validated for several samples to ensure that the regression method can be accurately generalized to other students and different contexts.

Despite these limitations and the ones mentioned in prior sections, our model serves as an effective predictor of academic success in college students. Not only does it serve as a point of reference for organizations concerned with student well being in college, but also for college officials to determine what financial support measures can be taken to reduce drop out outcomes. Additionally, this can provide valuable insight for academic advisors when foreseeing whether a student is likely to burn out or lose motivation during the course of their college education. Due to this wide range of benefits and the relevance of such a prediction model, it is important for data scientists to build future extensions based on this analysis. By being aware of such contributing factors, a country's higher education system can be vastly improved and students can be better prepared to tackle the challenges of college life with less stress. As undergraduate students ourselves, we can affirm that efforts such as these play a critical role in how individuals perceive their own abilities and the drive to take up an enriching career in the future.

References

Dataset used:
Valentim Realinho, Jorge Machado, Luís Baptista, & Mónica V. Martins. (2021). Predict students' dropout and academic success (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5777340