Undergraduate Dropout Predictor Model

Katia Albito
Anum Yaseen

Table of contents

Objective

Recent statistics suggest that Canadian colleges have a dropout rate of approximately 33% (https://www.collegefactual.com/colleges/canada-college/academic-life/graduation-and-retention/#drop_outs). Studies also suggest that college dropouts make an average of 35% less income compared to their graduate peers and they are 20% more likely to be unemployed than any degree holder (https://educationdata.org/college-dropout-rates).

With the use of data, we aim to identify the risk factors for university dropout and use these findings to create a model that pinpoints at-risk students. Our goal is to help educators identify students at risk of dropping out so they can receive the necessary resources and support on time to successfully complete their higher education.

Data Preparation

The data for this model was obtained in Kaggle. The data corresponds to records of students from 17 different undergraduate degrees, and it includes demographic data, socioeconomic and macroeconomic data, and academic data on enrollment and academic performance.
The students in this dataset were enrolled between the academic years 2008/2009 to 2018/2019. The data collected corresponds to the student's final year of study.

The dataset consists of 4424 records with 35 attributes. In the data exploration phase, we noticed the data had no null or duplicate values and contained no missing values.

The 35 attributes can be classified into categorical (18) and continuous (17). The 'Target' attribute is categorical and has values: Dropout, Enrolled, Graduate. For the purpose of this model, we will exclude records in the Enrolled category since we only want to predict the likelihood of students dropping out.

Some of the challenges that we faced during the data preparation phase were the following:

- The current dataset contains only 2 semester-worth information on each admission intake which limits us in assessing overall student performance. Access to additional semester information would have helped us provide more accurate results
- The data set had a mix of both categorical & continuous information. In order to deal with the two different data types, we had to use different analysis methods at every stage of this project
- Some variables had a huge number of repetitive categories which we had to regroup in order to provide cleaner results

The number of categories among the categorical variables ranges from 2 all the way up to 46 categories. In cases where there were a large number of categories we grouped similar categories together in order to make the data more concise and understandable by our model.

This was the case for: ‘Mother/Father Occupation’, ‘Mother/Father Qualification’, ‘Previous qualification’ and ‘Course’.

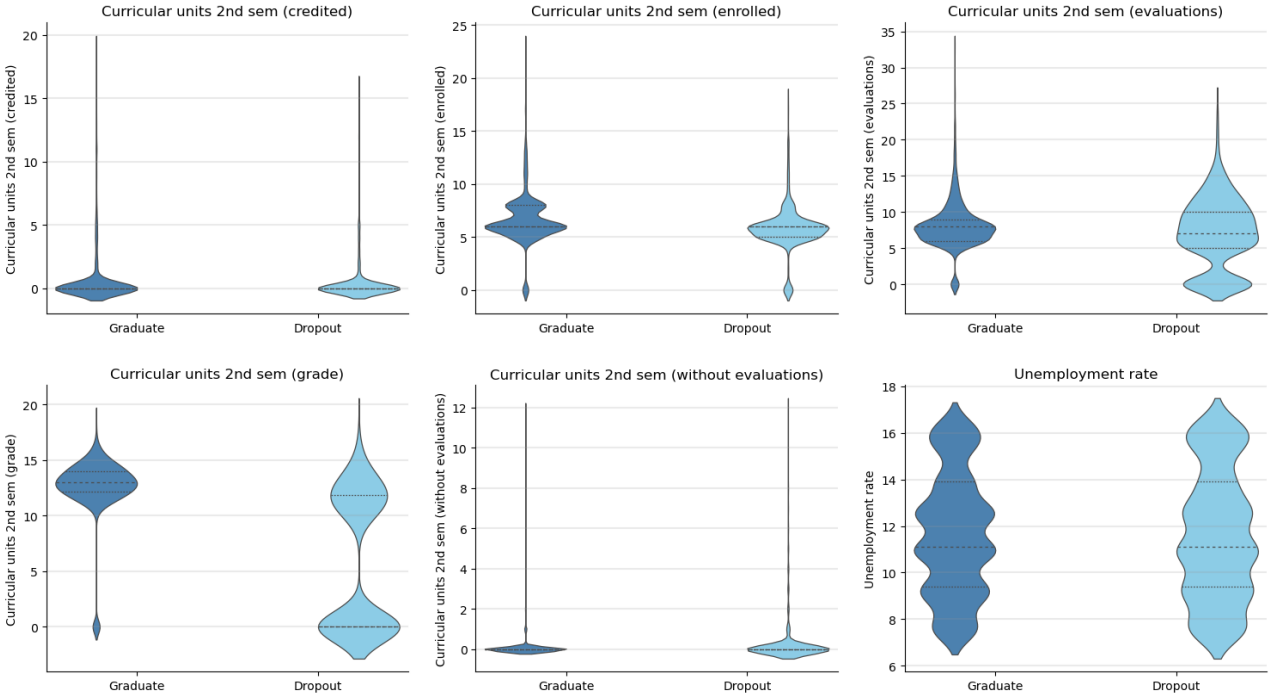
Our goal is to identify which factors impact the student’s decision to drop out and use these to build a prediction model.

Data Visualization

Since the data contains a mix of categorical and continuous variables, each variable type will be plotted using a different type of plot to get a better idea of how each factor impacts student retention and whether we can use those attributes for predictions. Continuous variables will be visualized with violin plots and categorical variables with bar plots.

Continuous Independent Variables

For the continuous variables, a violin plot was selected for visualization. This type of plot provides information on the distribution of the independent variable across each Target class. Below are some of the plots produced (see annex A for all).

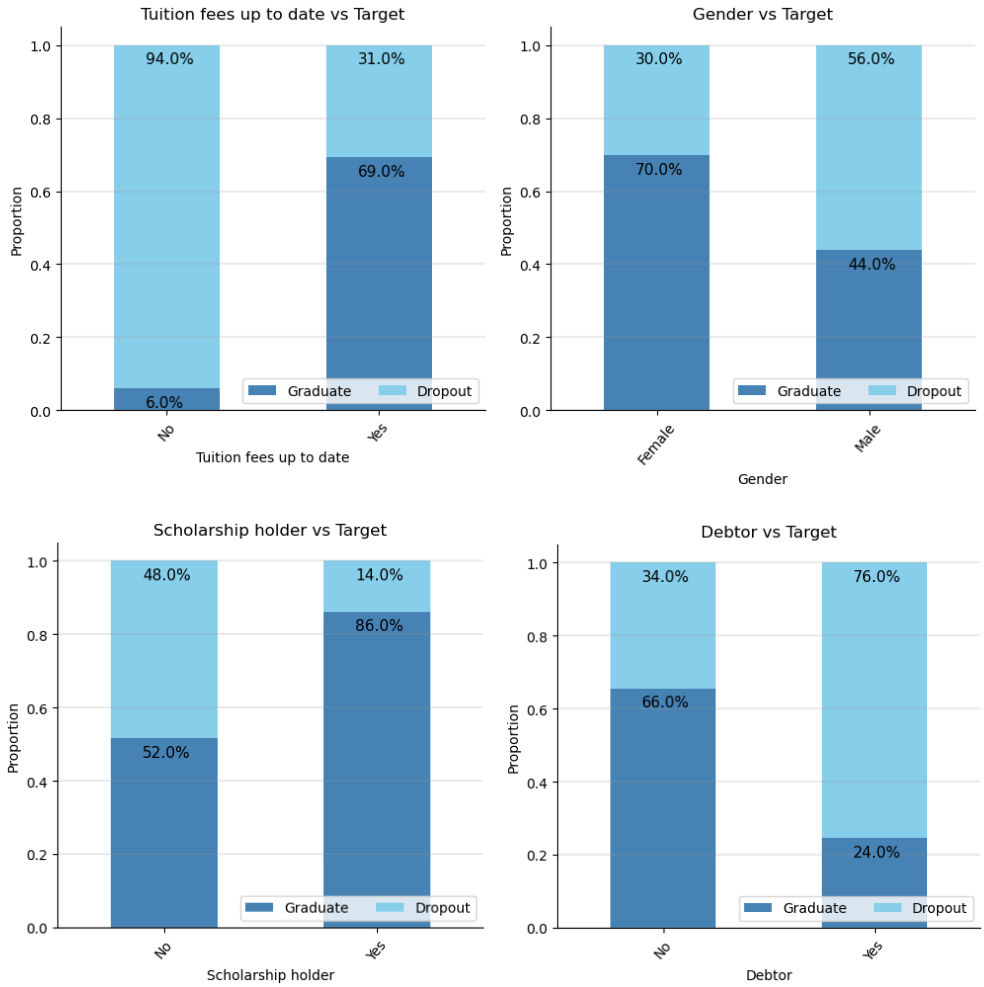


From the plots we are able to identify which variables have a normal-like distribution and which ones do not. For instance, the Unemployment Rate distribution in both Target groups is not normally distributed. We can also see both distributions have similar shapes. This leads us to believe Unemployment Rate is not a variable impacting the Target value. A similar reasoning can be applied to other variables such as Inflation Rate or GDP.

On the other hand, we see variables like Curricular units 2nd sem (grade) with distributions showing two clusters. The Graduate group has an almost normal-like distribution with most of the values around 12 and a very small cluster around 0 that can be considered as outliers. The cluster around 0 in the Dropout group is a lot larger and thus it cannot be discarded as outliers. These distributions seem to indicate a possible relationship between the grades in the 2nd semester and Dropout likeliness.

Categorical Independent Variables

For the categorical variables, we decided to use bar plots illustrating the proportion (%) of a category of students that had graduated vs dropped out.



Looking at the categorical plots (see annex A for all) we can see some factors that immediately standout:

- Students who don't have their tuition fees up to date are 63% more likely to drop out than those who have their accounts paid.
- 56% of male students tend to dropout vs 30% female students

- Students without scholarships are more likely to dropout ~ 48% of non scholarship students drop out
- Students with debts are more likely to dropout~ 76% of students who are under debt drop out

Looking at the above plots, we were able to identify the attributes that seemed to have the most impact on the target variable. In order to confirm that, we will test the correlation using 2 types of tests since our data consists of both categorical & continuous data. A T-test for continuous variables & a Chi-square test for the categorical variables. We are using a confidence interval of 99%.

Model Building

To select the variables that can predict student dropout, we first performed a series of statistical tests that dictated which variables were statistically related to the Target. We performed a second round of tests to build a correlation matrix that we used to select the final variables based on their association strength with the target and among each other for collinearity.

Statistical Tests

Different statistical tests were applied to determine the relationship between independent variables and the Target value. Continuous variables were tested using a t-test while categorical variables were tested using a chi-square test.

A Cramer's V correlation matrix was chosen to identify the strength of correlations between the variables and the target and among the variables in order to prevent collinearity.

Continuous Variables

A two-sided t-test was applied to decide if there exists a significant difference between the means of the Graduate group and the Dropout group for each continuous independent variable. Even though the data does not meet the t-test normality check in some of the variables, the test is still valid since the sample size is large enough thanks to the Central Limit Theorem.

Test setup:

- Null hypothesis (H0): The difference of means is zero, i.e. the two groups have the same mean and therefore the variable does not have an effect on the Target outcome.
- Alternative hypothesis (H1): The difference of means is nonzero, i.e. the two groups have different means and therefore the variable does have an effect on the Target outcome.

- Variance: we splitted the variables in two groups based on their variances: equal variance and different variance. For each variable, if the difference in variances (absolute value) between Dropout and Graduate was less than 1, we said their variance is the same.
- To perform the t-test we made use of the package `stats` from the `scipy` Python library.

Test for equal variances:

```
stats.ttest_ind(group1, group2, axis=0, equal_var=True, alternative='two-sided')
```

Test for unequal variances:

```
stats.ttest_ind(group1, group2, axis=0, equal_var=False, alternative='two-sided')
```

In this test, a p-value that is less than or equal to our significance level of '0.01' indicates we can reject the null hypothesis. In other words, there is sufficient evidence to conclude that the variable has an effect on the Target outcome.

From our t-test results, we were able to conclude, 2 out of 17 continuous variables are not correlated to the target variable. We were able to strike out the following variables from our model (see Annex C for all test results):

1. Inflation Rate
2. Unemployment Rate

Categorical Variables

We used a Chi-square test to find a correlation between Categorical variables vs Target variable which is also a categorical variable.
The Chi-square test is a hypothesis test for independence which is used to test whether there is independence between two categorical variables. The goal is to analyze if the values of the first variable are affected by the values of the second variable, and vice versa

Chi-square test finds the probability of a Null hypothesis(H0):

- Assumption(H0): The two columns are NOT related to each other
- Result of Chi-Sq Test: The Probability of H0 being True

In this test, a p-value that is less than or equal to our significance level of '0.01' indicates there is sufficient evidence to conclude that a relationship exists between the categorical variable and target variable.
From our chi-square test results, we were able to conclude, 3 out of 17 categorical variables are not correlated to the target variable. Thus we were able to strike out the following variables from our model (see Annex C for all test results):

1. Educational special needs
2. International
3. Nationality

Identifying the final Variables using Cramer's V Correlation Matrix

Our next step was to find the strongest variables amongst the resulting 14 categorical variables and 15 continuous variables that showed a correlation to the Target variable and to identify any collinearity between those variables so we could remove those from our model.

We used a correlation matrix that uses the Cramer's V association method to measure the strength of the relationship between the variables. It ranges from 0 to 1, where 0 indicates no association and 1 indicates a perfect association between the two variables.

Formula:

$$V = \sqrt{\frac{\chi^2 \setminus n}{min(r-1, c-1)}}$$

Key:

χ^2 = chi square statistic
n = sample size
r = number of rows in the contingency table
c = number of columns in the contingency table

Ranges of association strength:

- Weak = [0 - 0.3)
- Moderate = [0.3 - 0.6)
- Strong = [0.6 - 1]

For our purpose, we will be using any absolute association value >=0.3 to decide if there is a strong correlation.

Using the above method we narrowed down the independent variables to the followings:

	Curricular units 1st sem (approved)	**Curricular units 1st sem (grade)**	**Curricular units 2nd sem (approved)**	**Curricular units 2nd sem (grade)**
Curricular units 1st sem (approved)	1.000	0.710	0.916	0.709
Curricular units 1st sem (grade)	0.710	1.000	0.6919	0.8459
Curricular units 2nd sem	0.916	0.6919	1.000	0.7868

(approved)				
Curricular units 2nd sem (grade)	0.709	0.8459	0.7868	1.000
Target	-0.5549	-0.5199	-0.6540	-0.6054

Table 1: Cramer's V correlation matrix of continuous variables with strength >= 0.3

	Tuition fees up to date	**Scholarship holder**
Tuition fees up to date	1.000	0.1696
Scholarship holder	0.1696	1.000
Target	-0.4421	-0.3130

Table 2: Cramer's V correlation matrix of categorical variables with strength >= 0.3

The next step was to identify and remove collinearity between the variables. Table A suggests collinearity between all continuous variables. Table B suggests no collinearity exists between the categorical variables.
In order to remove collinearity we must select one of the collinear variables as the sole independent variable. Although Curricular units 2nd sem (approved) has the strongest relationship with the target (-0.65), selecting a 1st semester independent variable is more coherent with the model's goal of identifying and preventing student dropout.

Out of the 1st semester variables, the number of approved courses has the strongest relationship with the Target (-0.55). Nevertheless, the number of approved courses is not an accurate enough metric. For instance, a student approving 2 out of 2 courses should not be evaluated the same way as a student taking 3 courses and only approving 2. The number of curricular units approved needs to be used with the context of how many courses the student was enrolled in.

To resolve this, we created a new continuous variable to represent the ratio of courses approved and courses enrolled and proceeded to perform statistical tests to verify the validity and usability of this new variable.

$$Curricular\ units\ 1st\ sem\ (approved/enrolled)\ =\ \frac{Curricular\ units\ 1st\ sem\ (approved)}{Curricular\ units\ 1st\ sem\ (enrolled)}$$

Two-sided t-test:

- H0: The Dropout and Graduate groups have the same mean and therefore the ratio of approved units does not have an effect on the Target outcome.

- H1: The Dropout and Graduate groups do not have the same mean and therefore the ratio of approved units does have an effect on the Target outcome.
- Significance level = 0.01.
- Dropout and Graduate groups have equal variances.

```
stats.ttest_ind(group1, group2, axis=0, equal_var=True, alternative='two-sided')
```

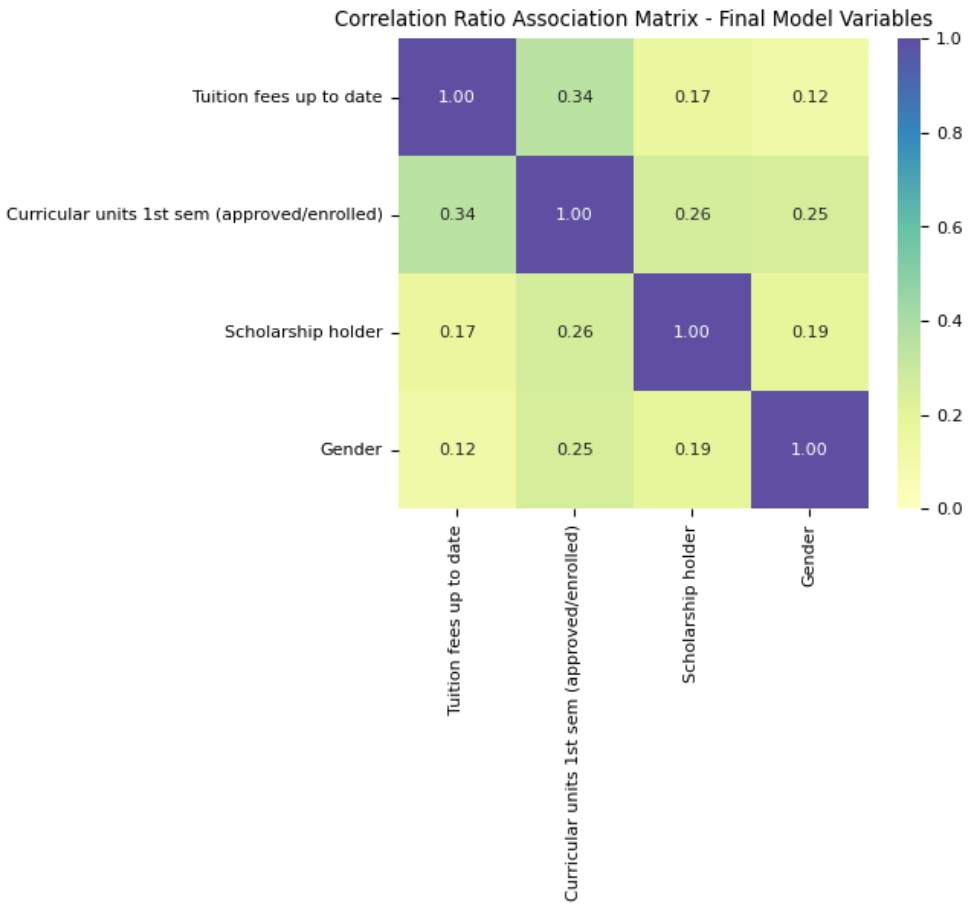
The resulting p-value is less than 0.01 and therefore we reject the null hypothesis and can conclude that there is evidence that the new variable has an effect on the Target outcome.

The next step is to add the variable to the Cramer's V correlation matrix. From the table below (Table 3) we can see the new variable has an even stronger relationship with the Target than any of the previous continuous variables.

	Curricular units 1st sem (approved/enrolled)	**Curricular units 1st sem (approved)**	**Curricular units 1st sem (grade)**	**Curricular units 2nd sem (approved)**	**Curricular units 2nd sem (grade)**
Curricular units 1st sem (approved/enrolled)	1.000	0.8069	0.8847	0.7855	0.8632
Curricular units 1st sem (approved)	0.8069	1.000	0.710	0.916	0.709
Curricular units 1st sem (grade)	0.8847	0.710	1.000	0.6919	0.8459
Curricular units 2nd sem (approved)	0.7855	0.916	0.6919	1.000	0.7868
Curricular units 2nd sem (grade)	0.8632	0.709	0.8459	0.7868	1.000
Target	-0.6792	-0.5549	-0.5199	-0.6540	-0.6054

Table 3: Cramer's V correlation matrix of continuous variables with strength >= 0.3

Now that we have found the relevant continuous and categorical variables, we will perform a final check for collinearity between all variables.



As we sought to incorporate a demographic variable into our model, we included Gender in our final list of independent variables. Although Gender only showed a modest relationship in the correlation matrix, the chi-square test did indicate a statistical effect. In the table above we can see there is no evidence of collinearity between our variables.

Model

Now that we have determined the final variables we want to include in the model:

1. Curricular units 1st sem (approved/enrolled)
2. Tuition Fees up to date
3. Scholarship holder
4. Gender

Our next step is to determine the best model to use and find the optimal model parameters. Considering our response variable (Target) is a binary variable, we continued to assess if our data meets the requirements of a Logistic regression model:

- ☑ The response variable is binary
- ☑ The observations are independent

- ☑ Little or no collinearity between the explanatory variables
- ☑ No extreme outliers
- ☑ There is a Linear Relationship Between Explanatory Variables and the Logit of the Response Variable
- ☑ The sample is sufficiently large (3,630)

Our choice of model: **Binary logistic regression model**

Formula:

$$p(x_1,\ldots,x_n) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\ldots+\beta_n x_n)}}$$

Key:

β_0= vertical intercept

β_n= coefficient for variable *n*

While setting up this model, we standardized the data to ensure all the variables have the same scale in order to improve model performance. After testing the algorithm with various train vs. test splits, we concluded that using a 70-30 split for training and testing respectively was the optimal split for the best predictions. We also assessed that including an intercept did not improve the accuracy of our model, hence, we excluded it.

We utilized the `sklearn` Python library to build our model.

Model Results

Using the logistic regression model in `sklearn`, we derived the subsequent results:

- Train score: 87.6%
- Test score: 86.6%
- Accuracy: 86.6%

Using the confusion matrix results, we determined the following:

- # of True negatives (TN) = 621
- # of False Negatives (FN) = 109
- # of False Positives (FP) = 37
- # of True Positives (TP) = 322

Based on the numbers above, we calculated our model has a 75% recall, which means it correctly identifies 75% of all drop outs. Furthermore, it has a 90% precision in predicting the drop outs which in other words means when our model predicts a dropout, it's correct 90% of the time.

Moreover, below are the coefficient values obtained in the regression model.

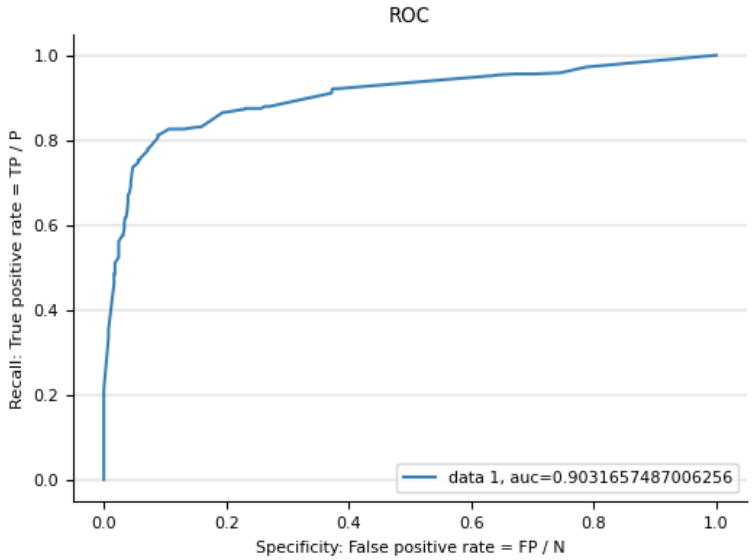
Variable	Regression Coefficient	Dropout % effect
Tuition fees up to date	- 1.196692	30.22
Scholarship holder	-0.461859	63.01
Gender	0.188952	20.80
Curricular units 1st sem (approved/enrolled)	-1.861610	15.54

- Students who aren't up to date on tuition payments are 30% more likely to dropout
- Students who don't have a scholarship are 63% more likely to dropout
- For every course not passed, a student is 15% more likely to dropout
- Being a male increases the likelihood of dropping out by 20%

Classification Report:

	Precision	**Recall**	**F1-score**	**Support**
0.0	0.85	0.94	0.89	658
1.0	0.90	0.75	0.82	431
Accuracy			0.87	1089
Macro Avg	0.87	0.85	0.86	1089
Weighted Avg	0.87	0.87	0.86	1089

Receiver Operating Characteristic (ROC) plot



Finally we examined the ROC curve produced. Our ROC curve lies above the 45 degree line and it is very close to the top left corner of the graph indicating a high True Positive rate and a low False Positive rate. The area under the curve demonstrates excellent model performance. And lastly, a high F1 score of 0.82 further emphasizes on the well balanced model performance.

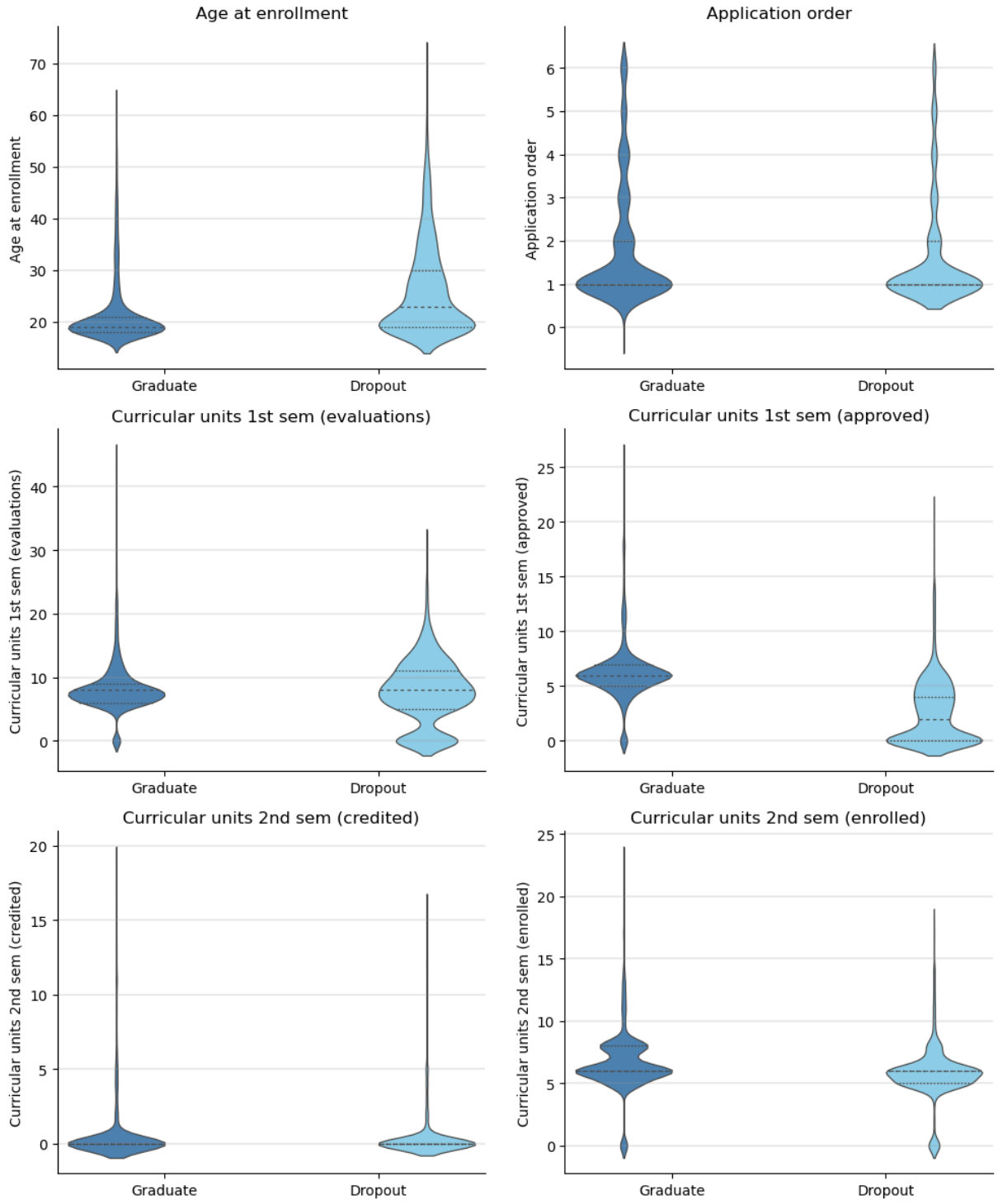
Conclusions

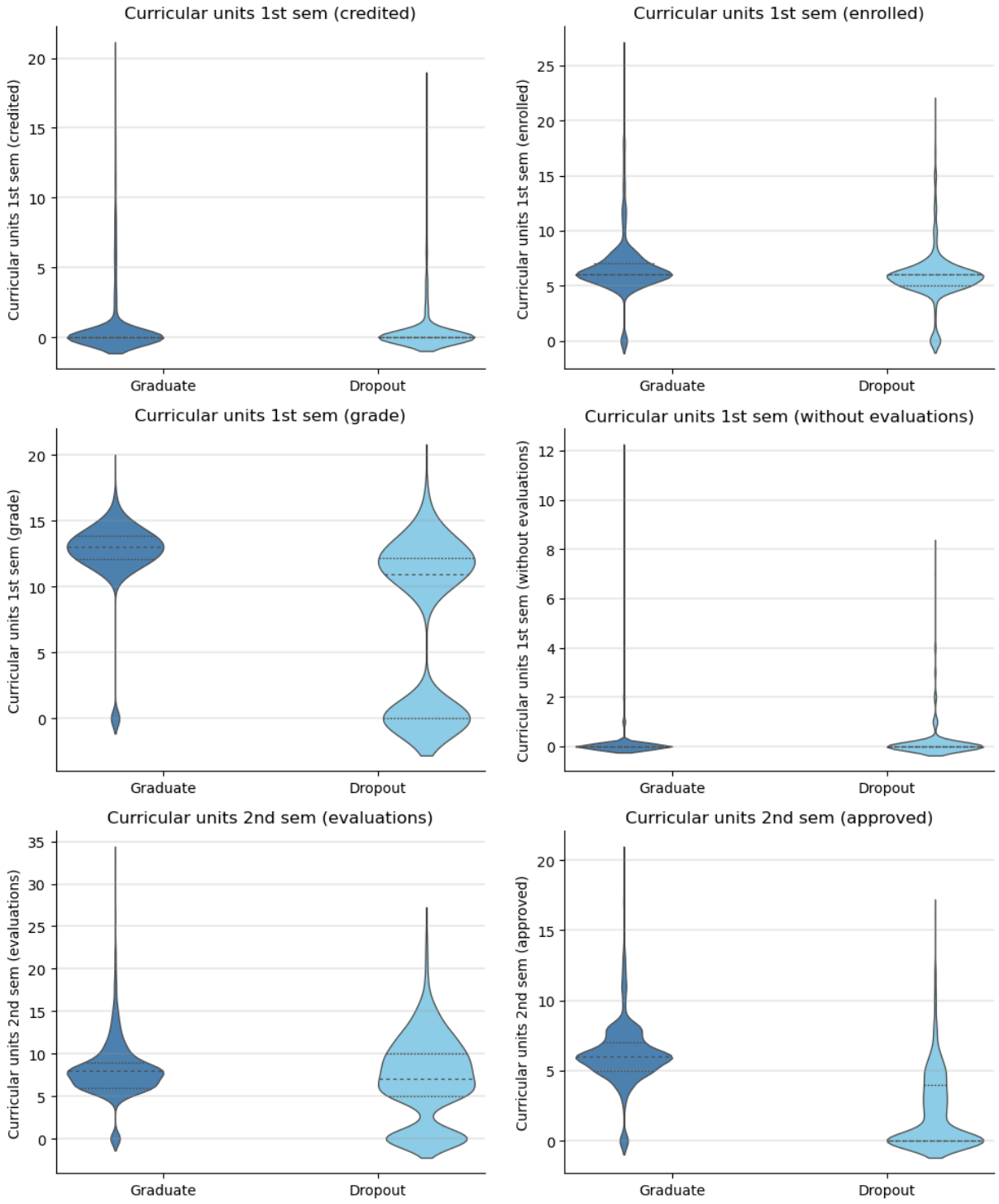
The goal of this project was to identify the factors impacting a student's decision to drop out and to use these factors to build a prediction model. After carefully following a step by step process of data cleaning and statistical tests, we successfully identified the 4 variables with greatest impact and used them to build our prediction model, which yielded an 86.6% accuracy. After examining the outcomes from the classification report and the ROC curve above, we can safely conclude that our model performs exceptionally well and can be used for predicting student dropouts accurately.

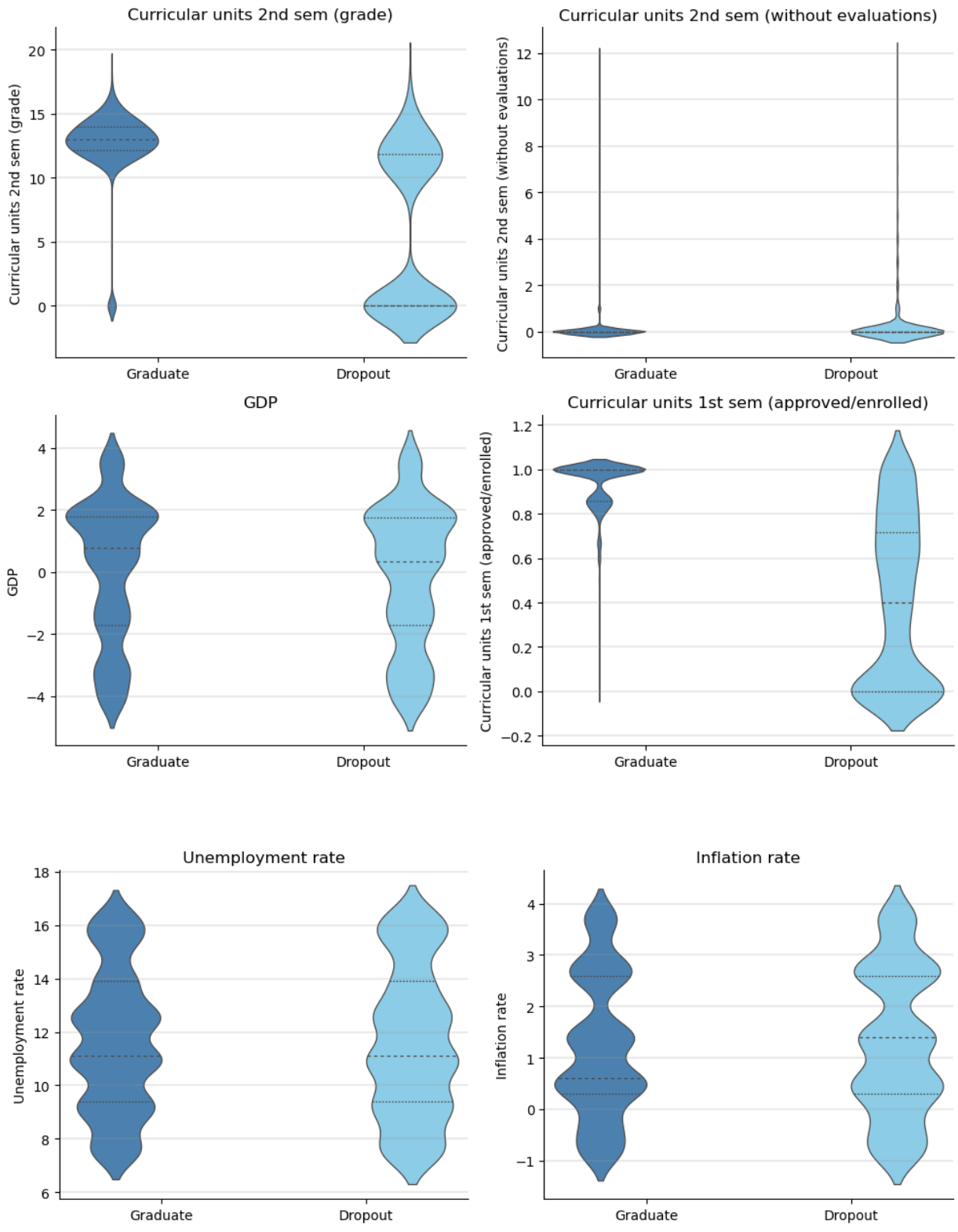
Through this project, we came across some expected and unexpected findings. For instance, we anticipated economic factors to have an impact on student dropout, which proved true. Surprisingly, parents' level of qualification and occupation did not have a great effect on student dropout. We revealed that gender had a much greater impact predicting dropout rates than any other socio demographic factor.

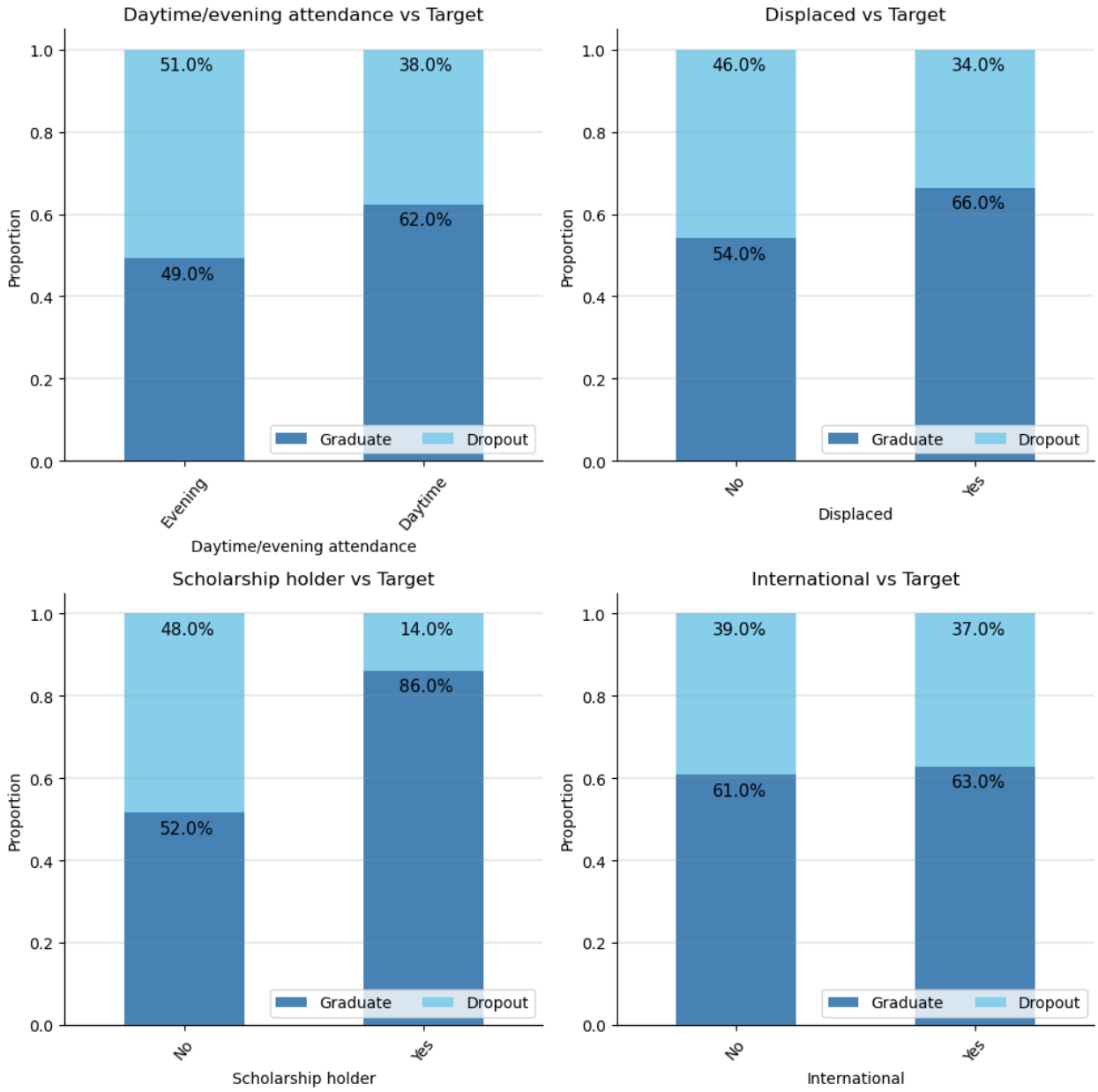
To further expand on this analysis and prediction model, we suggest an analysis covering more years of the students' program, and more extensive socio economic and demographic data such as household income, employment or residency.

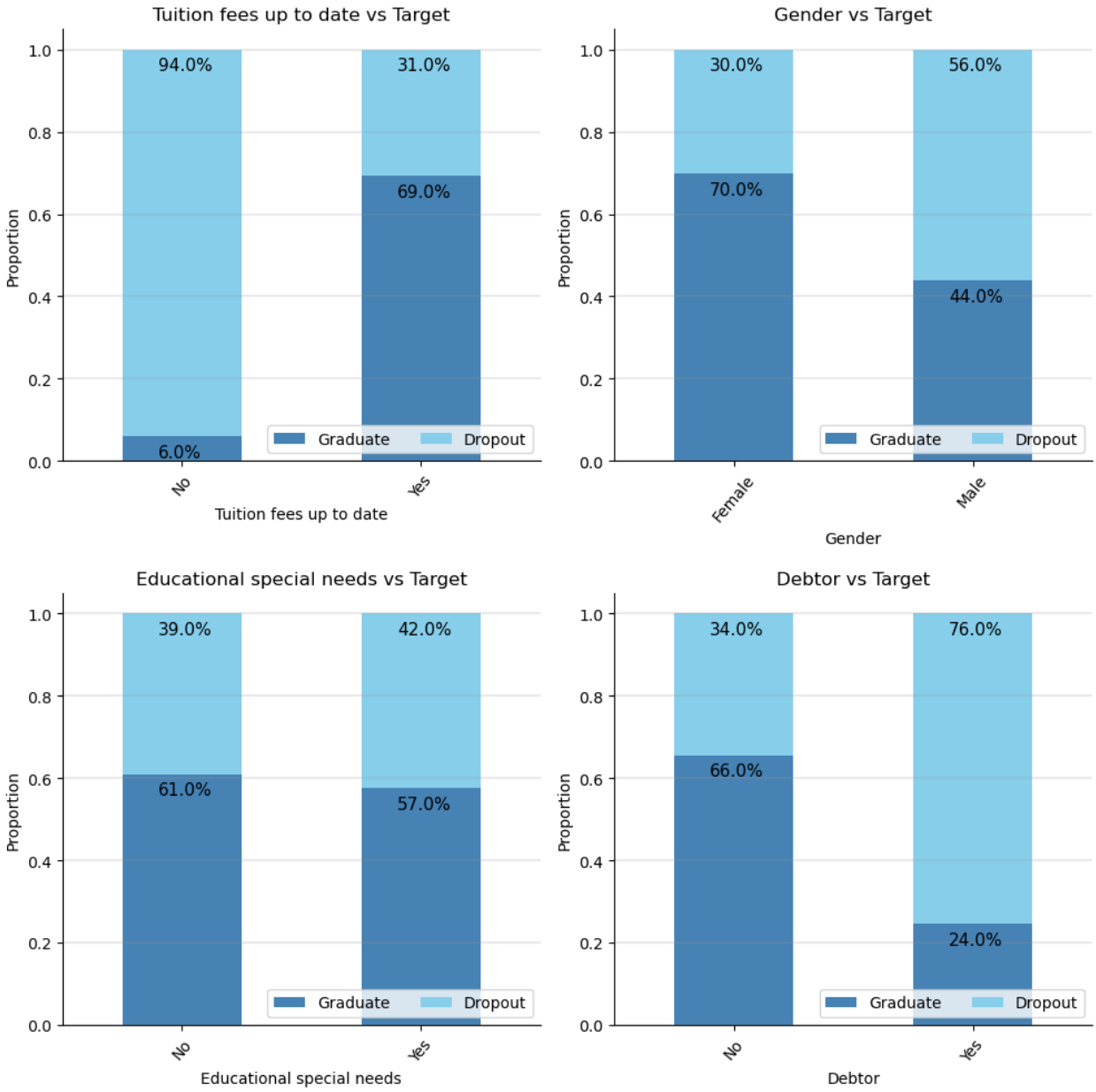
Annex A: Independent Variables Plots

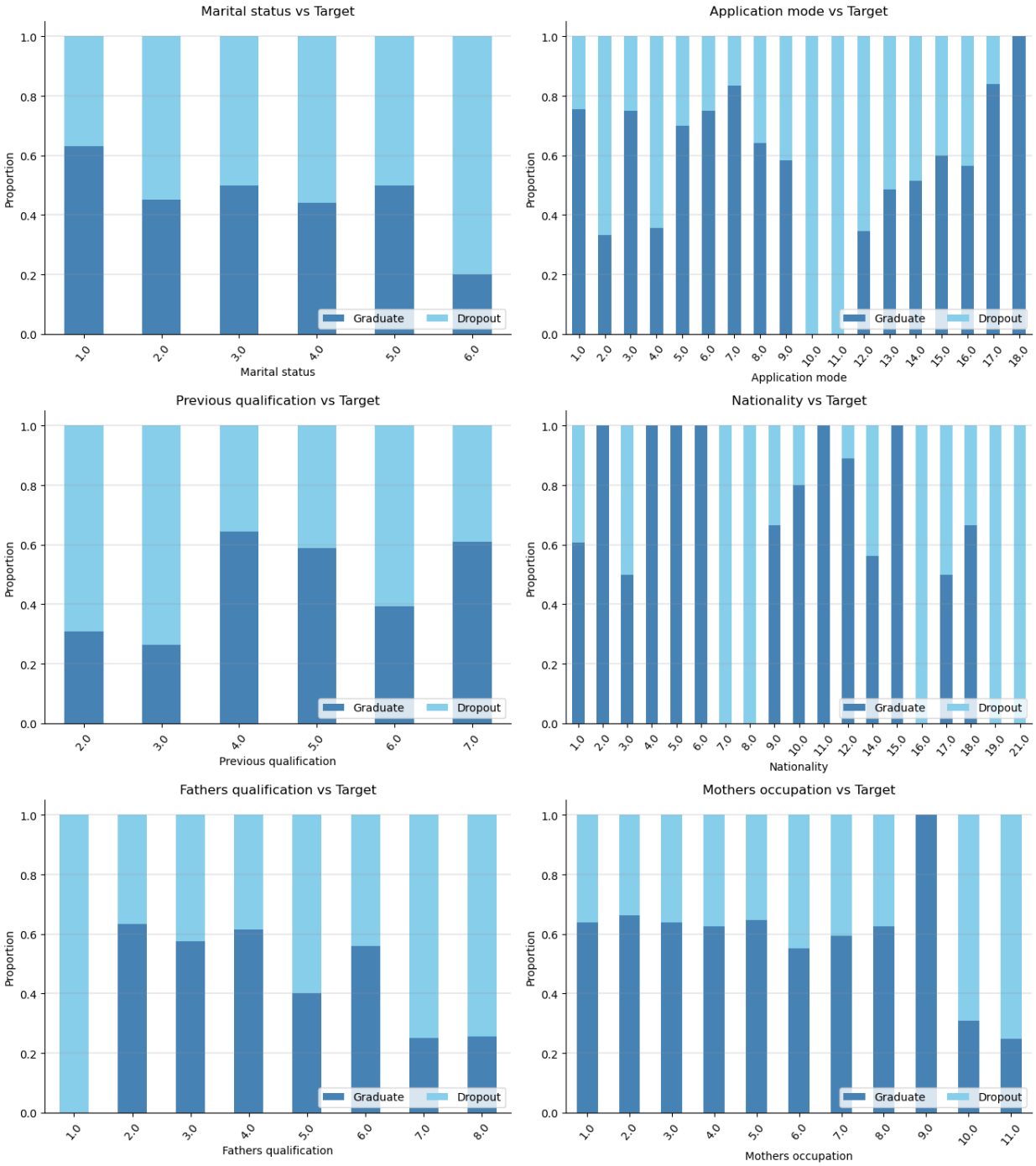


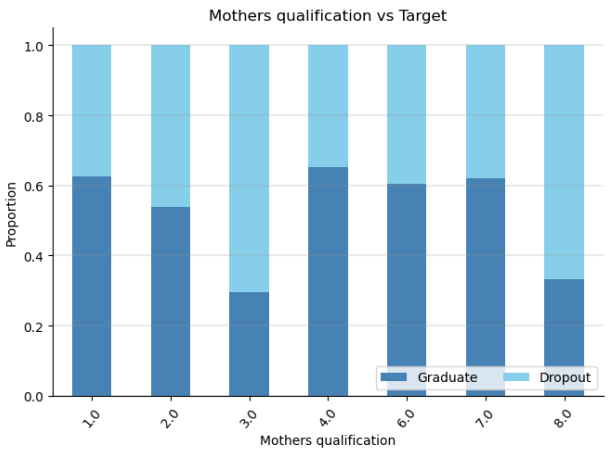

Curricular units 1st sem (credited)
Curricular units 1st sem (credited)
Curricular units 1st sem (enrolled)
Curricular units 1st sem (enrolled)
Curricular units 1st sem (grade)
Curricular units 1st sem (grade)
Curricular units 1st sem (without evaluations)
Curricular units 1st sem (without evaluations)
Curricular units 2nd sem (evaluations)
Curricular units 2nd sem (evaluations)
Curricular units 2nd sem (approved)
Curricular units 2nd sem (approved)
Graduate
Dropout

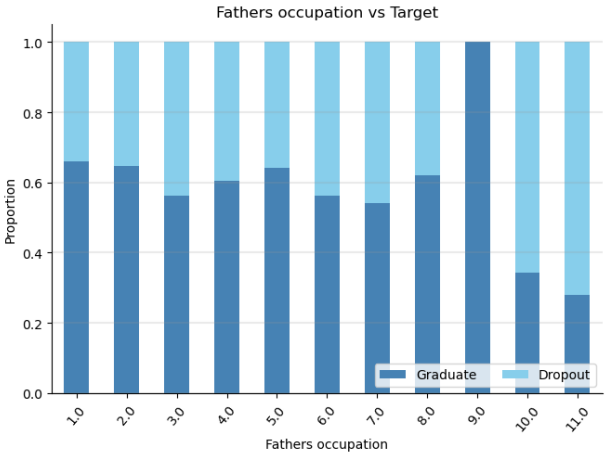

Curricular units 2nd sem (grade)
Curricular units 2nd sem (grade)
Graduate
Dropout
Curricular units 2nd sem (without evaluations)
Curricular units 2nd sem (without evaluations)
Graduate
Dropout
GDP
GDP
Graduate
Dropout
Curricular units 1st sem (approved/enrolled)
Curricular units 1st sem (approved/enrolled)
Graduate
Dropout
Unemployment rate
Unemployment rate
Graduate
Dropout
Inflation rate
Inflation rate
Graduate
Dropout

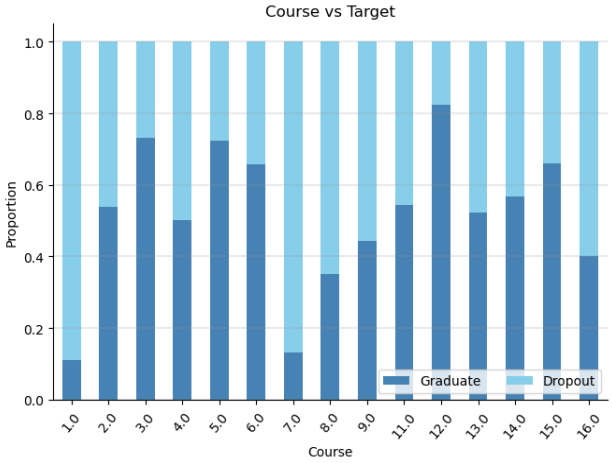

Daytime/evening attendance vs Target
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
51.0%
49.0%
38.0%
62.0%
Graduate
Dropout
Evening
Daytime
Daytime/evening attendance
Displaced vs Target
46.0%
54.0%
34.0%
66.0%
No
Yes
Displaced
Scholarship holder vs Target
48.0%
52.0%
14.0%
86.0%
Scholarship holder
International vs Target
39.0%
61.0%
37.0%
63.0%
International


Tuition fees up to date vs Target
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
94.0%
6.0%
31.0%
69.0%
Graduate
Dropout
No
Yes
Tuition fees up to date
Gender vs Target
30.0%
70.0%
56.0%
44.0%
Female
Male
Gender
Educational special needs vs Target
39.0%
61.0%
42.0%
57.0%
Educational special needs
Debtor vs Target
34.0%
66.0%
76.0%
24.0%
Debtor

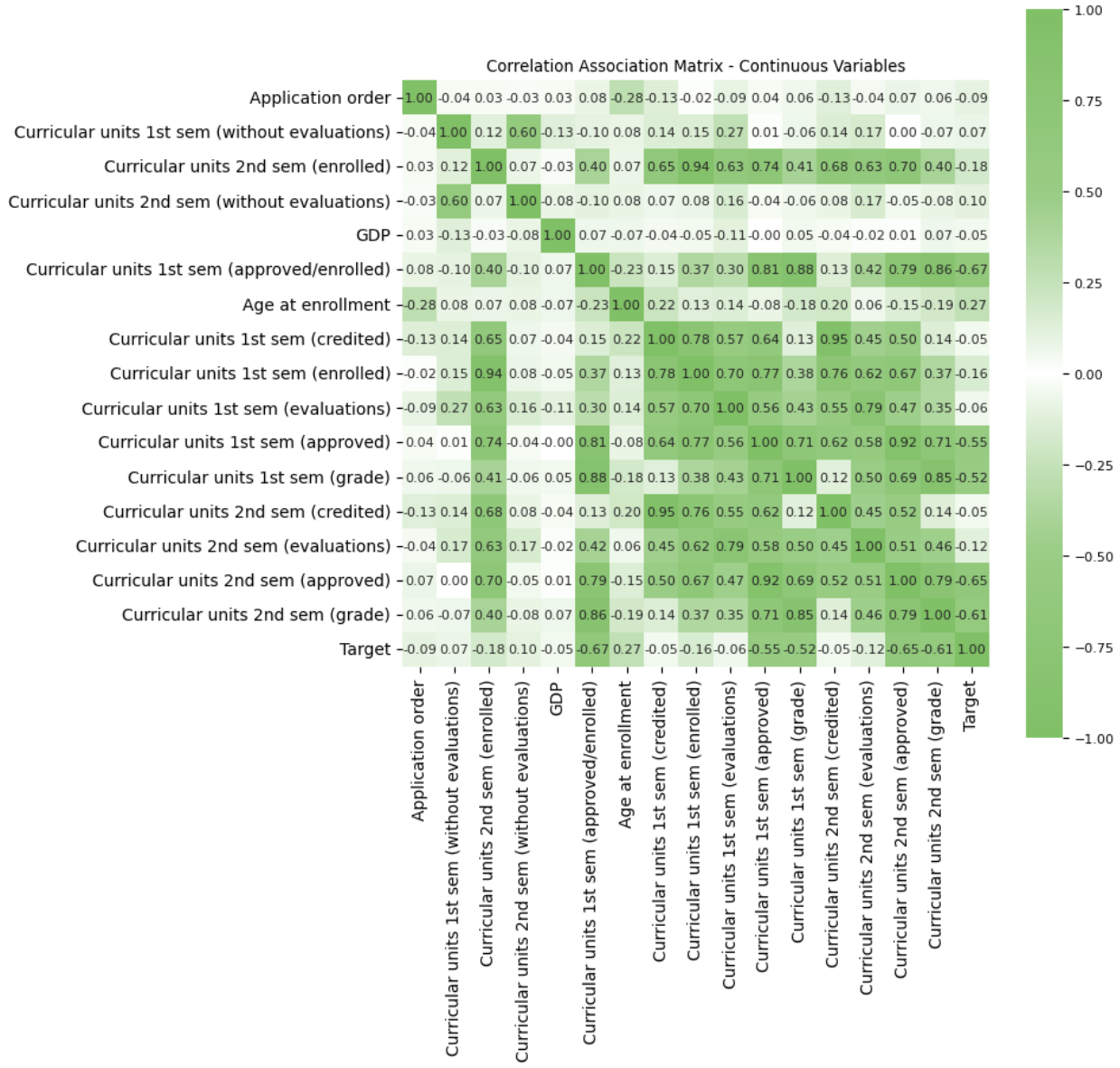

Marital status vs Target
Proportion
Marital status
Graduate
Dropout
Application mode vs Target
Application mode
Previous qualification vs Target
Previous qualification
Nationality vs Target
Nationality
Fathers qualification vs Target
Fathers qualification
Mothers occupation vs Target
Mothers occupation

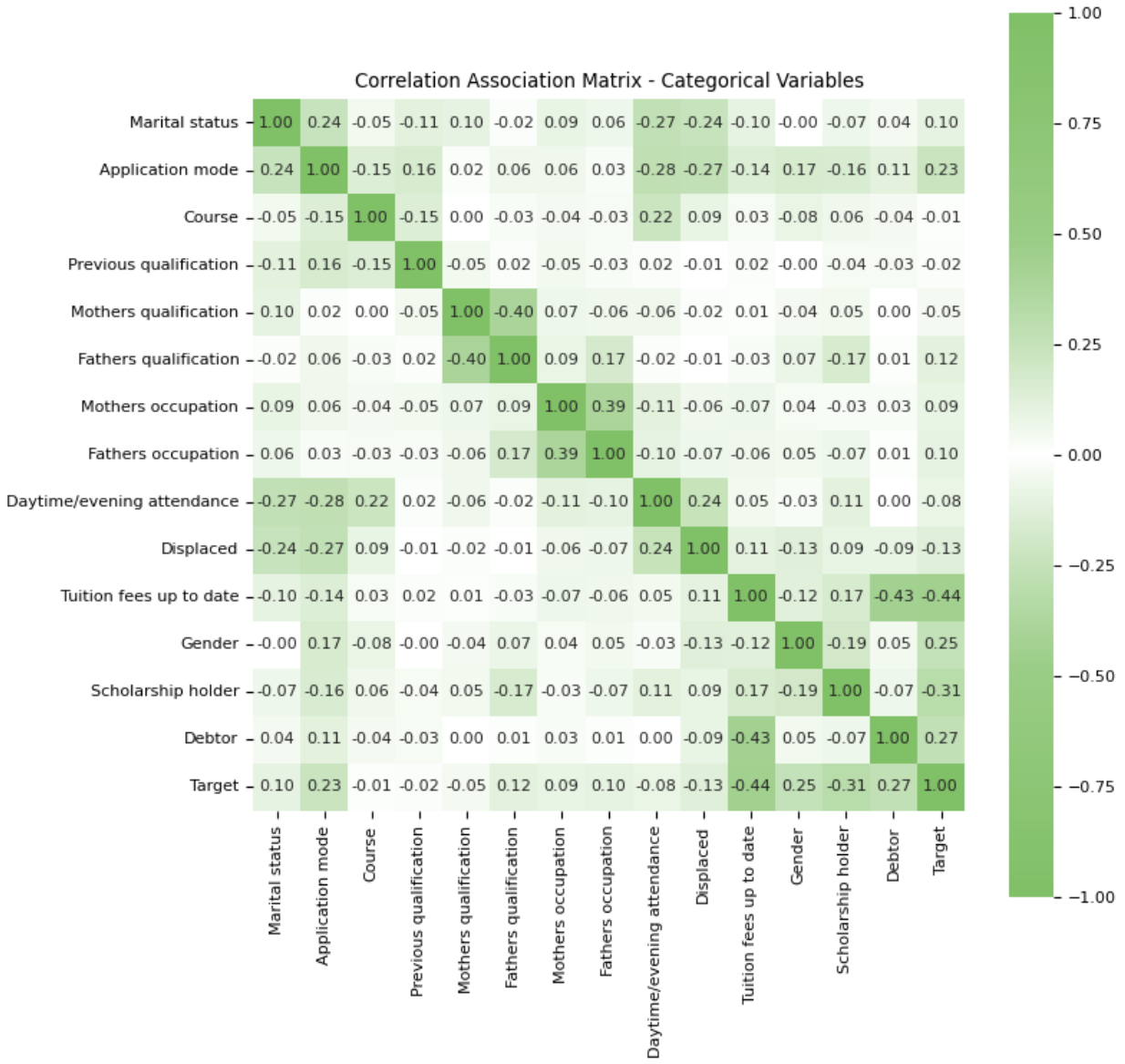

Mothers qualification vs Target
Proportion
Mothers qualification
Graduate
Dropout

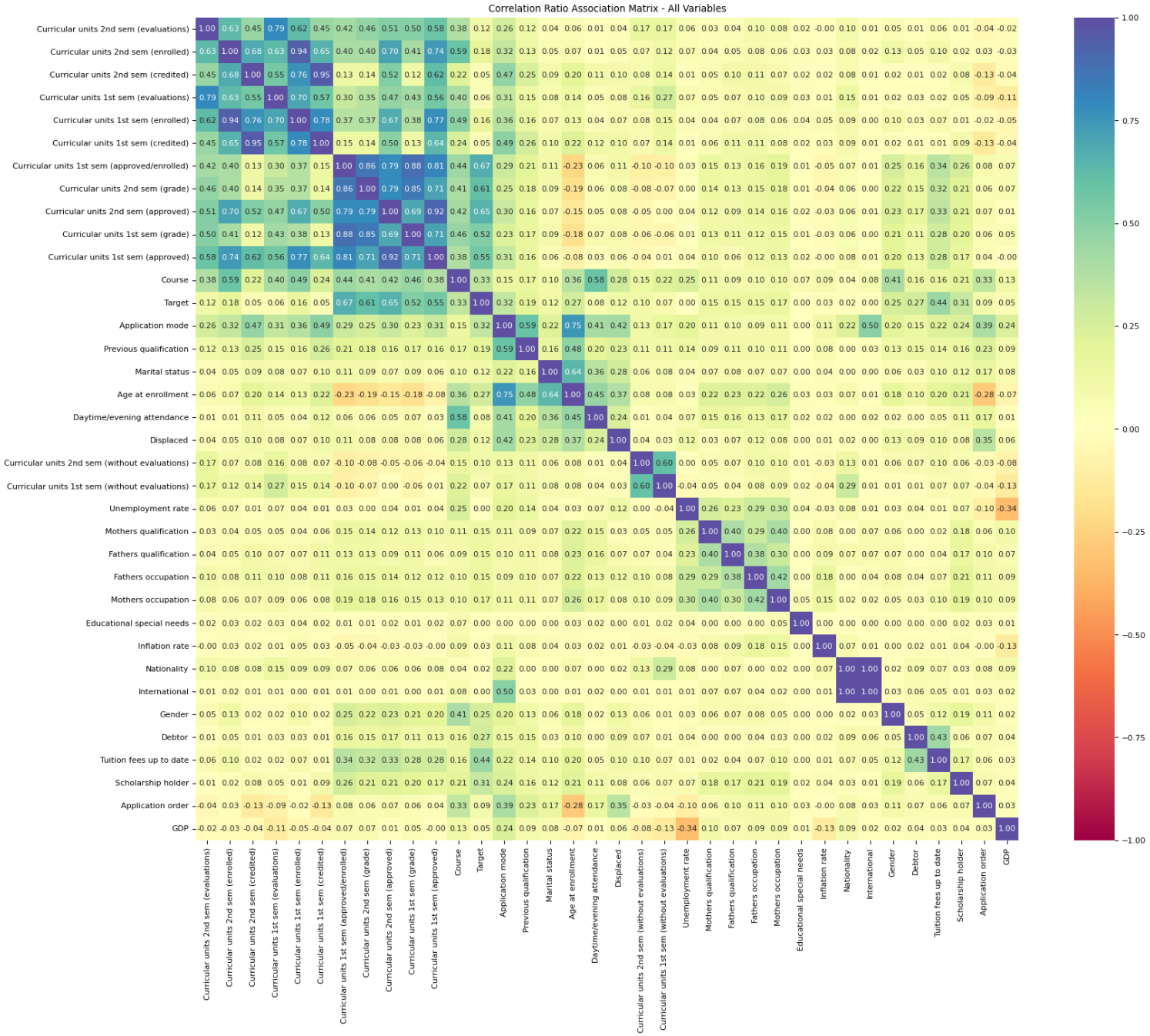

Fathers occupation vs Target
Proportion
Fathers occupation
Graduate
Dropout


Course vs Target
Proportion
Course
Graduate
Dropout

Annex B: Correlation Matrices




Correlation Association Matrix - Categorical Variables
Marital status 1.00 0.24 -0.05 -0.11 0.10 -0.02 0.09 0.06 -0.27 -0.24 -0.10 -0.00 -0.07 0.04 0.10
Application mode 0.24 1.00 -0.15 0.16 0.02 0.06 0.06 0.03 -0.28 -0.27 -0.14 0.17 -0.16 0.11 0.23
Course -0.05 -0.15 1.00 -0.15 0.00 -0.03 -0.04 -0.03 0.22 0.09 0.03 -0.08 0.06 -0.04 -0.01
Previous qualification -0.11 0.16 -0.15 1.00 -0.05 0.02 -0.05 -0.03 0.02 -0.01 0.02 -0.00 -0.04 -0.03 -0.02
Mothers qualification 0.10 0.02 0.00 -0.05 1.00 -0.40 0.07 -0.06 -0.06 -0.02 0.01 -0.04 0.05 0.00 -0.05
Fathers qualification -0.02 0.06 -0.03 0.02 -0.40 1.00 0.09 0.17 -0.02 -0.01 -0.03 0.07 -0.17 0.01 0.12
Mothers occupation 0.09 0.06 -0.04 -0.05 0.07 0.09 1.00 0.39 -0.11 -0.06 -0.07 0.04 -0.03 0.03 0.09
Fathers occupation 0.06 0.03 -0.03 -0.03 -0.06 0.17 0.39 1.00 -0.10 -0.07 -0.06 0.05 -0.07 0.01 0.10
Daytime/evening attendance -0.27 -0.28 0.22 0.02 -0.06 -0.02 -0.11 -0.10 1.00 0.24 0.05 -0.03 0.11 0.00 -0.08
Displaced -0.24 -0.27 0.09 -0.01 -0.02 -0.01 -0.06 -0.07 0.24 1.00 0.11 -0.13 0.09 -0.09 -0.13
Tuition fees up to date -0.10 -0.14 0.03 0.02 0.01 -0.03 -0.07 -0.06 0.05 0.11 1.00 -0.12 0.17 -0.43 -0.44
Gender -0.00 0.17 -0.08 -0.00 -0.04 0.07 0.04 0.05 -0.03 -0.13 -0.12 1.00 -0.19 0.05 0.25
Scholarship holder -0.07 -0.16 0.06 -0.04 0.05 -0.17 -0.03 -0.07 0.11 0.09 0.17 -0.19 1.00 -0.07 -0.31
Debtor 0.04 0.11 -0.04 -0.03 0.00 0.01 0.03 0.01 0.00 -0.09 -0.43 0.05 -0.07 1.00 0.27
Target 0.10 0.23 -0.01 -0.02 -0.05 0.12 0.09 0.10 -0.08 -0.13 -0.44 0.25 -0.31 0.27 1.00
Marital status
Application mode
Course
Previous qualification
Mothers qualification
Fathers qualification
Mothers occupation
Fathers occupation
Daytime/evening attendance
Displaced
Tuition fees up to date
Gender
Scholarship holder
Debtor
Target
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Correlation Ratio Association Matrix - All Variables
Curricular units 2nd sem (evaluations)
Curricular units 2nd sem (enrolled)
Curricular units 2nd sem (credited)
Curricular units 1st sem (evaluations)
Curricular units 1st sem (enrolled)
Curricular units 1st sem (credited)
Curricular units 1st sem (approved/enrolled)
Curricular units 2nd sem (grade)
Curricular units 2nd sem (approved)
Curricular units 1st sem (grade)
Curricular units 1st sem (approved)
Course
Target
Application mode
Previous qualification
Marital status
Age at enrollment
Daytime/evening attendance
Displaced
Curricular units 2nd sem (without evaluations)
Curricular units 1st sem (without evaluations)
Unemployment rate
Mothers qualification
Fathers qualification
Fathers occupation
Mothers occupation
Educational special needs
Inflation rate
Nationality
International
Gender
Debtor
Tuition fees up to date
Scholarship holder
Application order
GDP
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Annex C: Statistical Test Results

variable	n1	n2	statistic	pvalue	test_result	t-abs
Curricular units 1st sem (approved/enrolled)	1421	2209	-54.396784	0.000000e+00	1.0	54.396784
Curricular units 2nd sem (approved)	1421	2209	-50.671268	0.000000e+00	1.0	50.671268
Curricular units 2nd sem (grade)	1421	2209	-39.504220	1.768346e-245	1.0	39.504220
Curricular units 1st sem (approved)	1421	2209	-39.306474	1.307655e-269	1.0	39.306474
Curricular units 1st sem (grade)	1421	2209	-31.690690	2.399283e-175	1.0	31.690690
Age at enrollment	1421	2209	15.796570	1.252287e-53	1.0	15.796570
Curricular units 2nd sem (enrolled)	1421	2209	-11.205396	1.128426e-28	1.0	11.205396
Curricular units 1st sem (enrolled)	1421	2209	-10.122656	9.676921e-24	1.0	10.122656
Curricular units 2nd sem (evaluations)	1421	2209	-6.666051	3.295788e-11	1.0	6.666051
Curricular units 2nd sem (without evaluations)	1421	2209	6.217994	5.607154e-10	1.0	6.217994
Application order	1421	2209	-5.708720	1.229608e-08	1.0	5.708720
Curricular units 1st sem (without evaluations)	1421	2209	4.508495	6.735200e-06	1.0	4.508495
Curricular units 1st sem (evaluations)	1421	2209	-3.416323	6.449183e-04	1.0	3.416323
Curricular units 2nd sem (credited)	1421	2209	-3.350630	8.147541e-04	1.0	3.350630
GDP	1421	2209	-3.031144	2.453491e-03	1.0	3.031144
Curricular units 1st sem (credited)	1421	2209	-2.978949	2.912313e-03	1.0	2.978949
Inflation rate	1421	2209	1.827455	6.771352e-02	0.0	1.827455
Unemployment rate	1421	2209	-0.252866	8.003859e-01	0.0	0.252866

variable	pvalue	test_result
Marital status	3.160761e-10	1.0
Application mode	8.435153e-73	1.0
Course	7.389961e-81	1.0
Previous qualification	1.678957e-27	1.0
Nationality	3.412815e-01	0.0
Mothers qualification	2.877852e-16	1.0
Fathers qualification	2.362734e-15	1.0
Mothers occupation	1.202485e-19	1.0
Fathers occupation	2.138908e-16	1.0
Daytime/evening attendance	4.728747e-07	1.0
Displaced	3.906373e-14	1.0
Tuition fees up to date	9.189124e-156	1.0
Gender	8.255974e-52	1.0
Scholarship holder	5.109075e-79	1.0
International	7.943867e-01	0.0
Educational special needs	7.839673e-01	0.0
Debtor	6.141424e-58	1.0