Predicting Student Dropout and Academic Success

Exploratory Data Analysis

Patricia Gotz Lana Kabanni

2025-11-17

I. Introduction

This analysis examines data from a Portuguese higher education institution to identify factors that contribute to student dropout and academic success. The dataset contains information on 4,424 students enrolled across various undergraduate programs.

I.I - Background and Motivation

Student retention and academic success are crucial issues for higher education institutions. Universities increasingly rely on data-driven insights to identify at-risk students and to design early intervention strategies. We chose this topic because predicting student dropout not only helps optimize institutional resources but also supports students in achieving their academic goals. Understanding the factors that influence academic success, such as socio-economic background, previous academic performance, or family situation, can improve educational policies and personalized support systems. This subject is particularly meaningful in data science, because we can combine analytical and predictive methods to better understand and prevent student dropout.

I.II - Project Goals

The main objective of this project is to identify the factors that influence students to drop out, stay enrolled, or graduate from higher education. The dataset provides detailed information on each student's academic performance, socioeconomic background, and demographic profile, offering a comprehensive view of the variables that shape educational outcomes. By the end of our analysis, we seek to identify the most significant combinations of academic and personal factors that influence student success. First, our analysis will focus on academic performance,

examining how variables such as admission grades, semester evaluations, and course results relate to final outcomes. For instance, we will analyze whether early academic performance can serve as a reliable predictor of future dropout risk. We will then explore the influence of socioeconomic and personal factors, including parental education, occupation, and financial situation, to understand their impact on academic achievement. Lastly, the dataset will be used to build and evaluate classification models that predict students' academic status (Dropout, Enrolled, or Graduate). In summary, this study combines exploratory analysis, visualization, and predictive modeling to generate actionable insights that help universities detect at-risk students early and strengthen academic success.

I.III - Research Questions

- I. How do academic performance indicators and study conditions influence students' likelihood of graduation or dropout?
- II. What is the impact of demographic and socioeconomic background on students' probability of dropping out?
- III. Can we accurately predict a student's final status (Dropout, Enrolled, or Graduate) based on their demographic, socioeconomic, and academic characteristics. Which are the most relevant among them?

Data

Data Sourcing

The dataset is publicly available on UCI Machine Learning Repository and was created from multiple databases of higher education institutions in Portugal. It is related to enrolled students in different undergraduate programs and shows how different demographic, socioeconomic and academic factors are related to the dropout. Since the data has already been collected and can be directly downloaded from UCI MLR - Predict Students' Dropout and Academic Success - [Accessed on 20th October] , there is no need to collect more data via webscraping or APIs.

Data Description

The dataset, containing data from a Portuguese higher education institution, is provided as a CSV file, approximately 520 KB in size, and contains detailed information about students'demographic, academic and socio-economic characteristics. It includes 4424 student records and 37 variables (features). After reviewing the dataset variables, we removed two irrelevant ones, resulting in 35 relevant variables selected for analysis.

The clean data is provided as below ..

Data Loading

```
Dataset shape: (4424, 37)
```

Variable Selection

We selected 35 relevant variables for analysis:

```
Selected 35 variables
```

Through this step, we didn't encounter any difficult challenges. The dataset was already clean and encoded, so we didn't need to perform variable merging, one-hot encoding or ordinal encoding. We only had to convert categorical variables into readable labels to facilitate our visualization analysis.

Preprocessing (Data Cleaning and Wrangling)

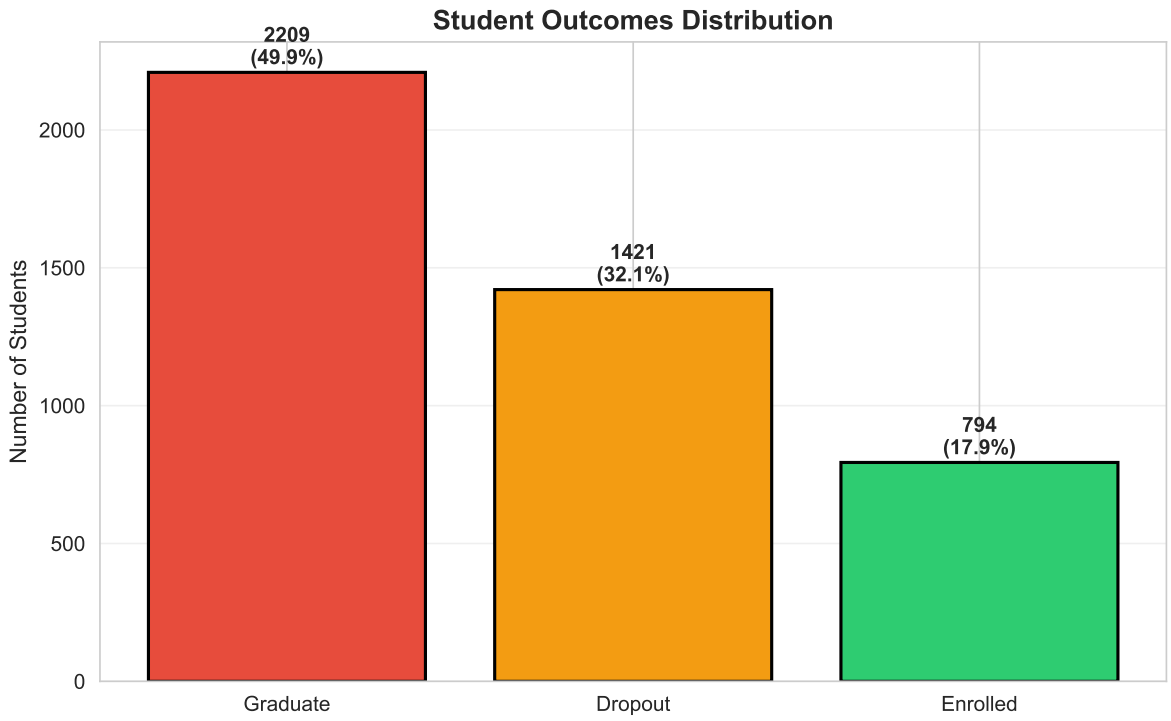
One of the most important steps in our project is data cleaning and wrangling. After running the code to check for missing values and undefined numerical data, we found that the dataset contains no missing values, no mistakes and no data entry mistakes.

The dataset was already encoded, and we removed "Application mode" and "Tuition fees up to date" variables because they are not relevant to our research questions. Therefore we dropped two columns from the dataset. Ensuring that the numeric columns are numeric, categorical variables such as "Gender", "Debtor", "Displaced" , "Daytime/Evening attendance" were converted to readable string labels for analysis.

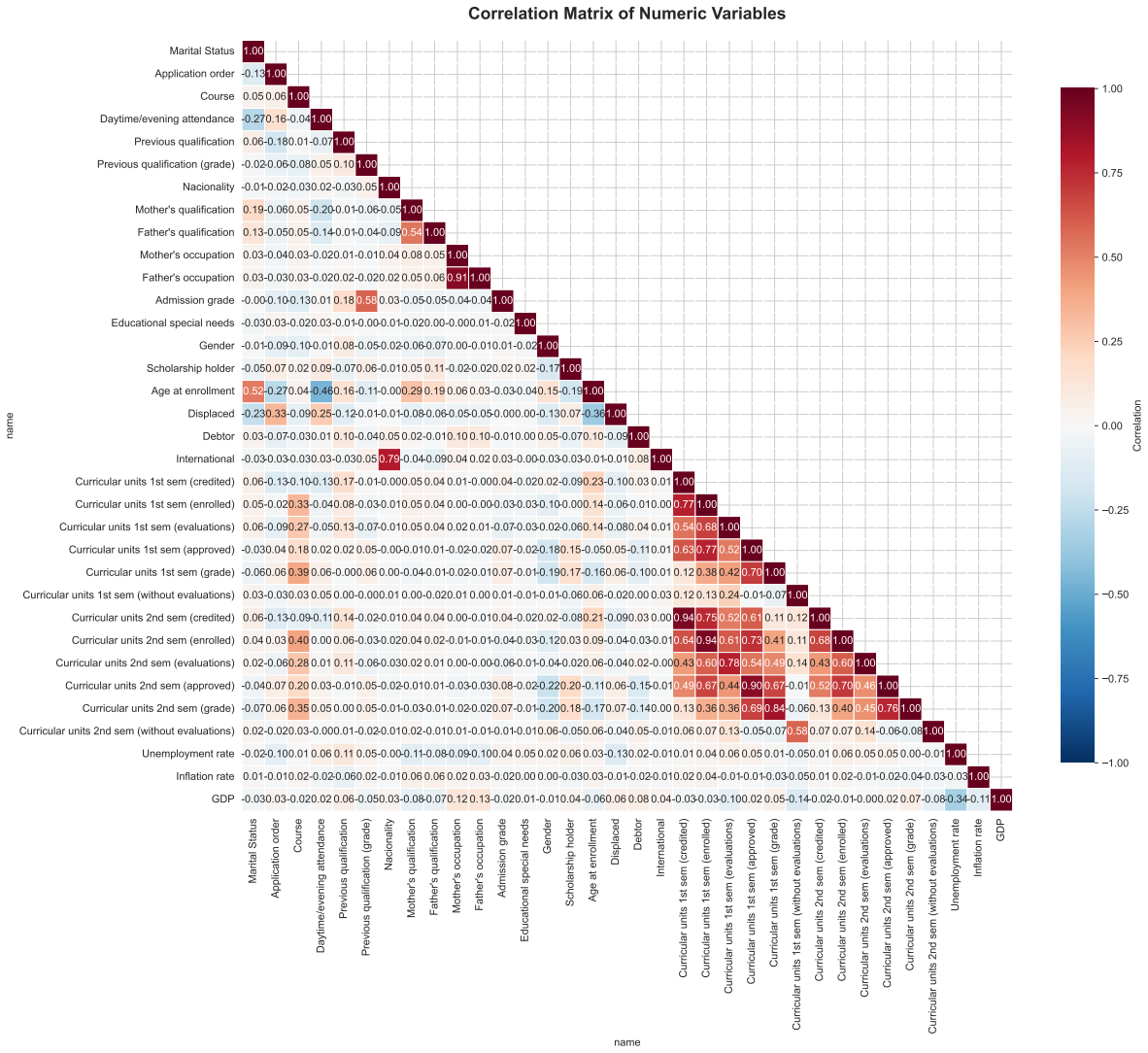
```
 No missing values found!

Shape after cleaning: (4424, 35)
Missing values: 0
Shape after cleaning: (4424, 35)
Missing values: 0
```

Although we had a well-structured and clean dataset, our main challenge was to determine the reliability of our dataset. We verified if there were any missing values, spotting mistakes, and determined irrelevant variables for our analysis. We pursue our cleaning work with the conversion of the categorical variables. Therefore, the reliable dataset was ready to be analyzed.

Target Variable



Correlation Analysis



Based on this correlation analysis, we identified several highly correlated variable pairs that suggest multicollinearity. We excluded 8 redundant semester variables that were highly correlated with other metrics.

Feature Selection

```
Removed 8 highly correlated variables
```

```
Remaining variables: 27
```

Outlier Detection

We implemented a type-aware outlier detection strategy that applies different methods based on the nature of each variable:

Binary variables (e.g., Gender, Scholarship holder): Outlier detection was skipped entirely, as these variables only contain two valid values (0/1).

Nominal categorical variables (e.g., Course, Nationality): No outlier detection applied, as these represent distinct categories without natural ordering. We only reported the number of unique categories present.

Ordinal categorical variables (e.g., qualifications, occupations): We reported the number of levels but did not apply outlier detection, as these represent ordered categories rather than continuous measurements.

Grade variables (0-200 scale): We checked for values outside the valid range (0-200). According to the dataset documentation, grades in the Portuguese system can range from 0 to 200.

Count variables (e.g., enrolled courses): We used a more lenient threshold of 3×IQR (Interquartile Range) rather than the standard 1.5×IQR, as count variables naturally exhibit right-skewed distributions where high values may represent legitimate cases (e.g., students enrolling in many courses).

Continuous variables (e.g., Age, GDP, Unemployment rate): We applied the standard Tukey method with 1.5×IQR threshold to identify potential outliers: values below Q1 - 1.5×IQR or above Q3 + 1.5×IQR.

This approach ensures that outlier detection is contextually appropriate for each variable type, reducing false positives while identifying genuine data quality issues.

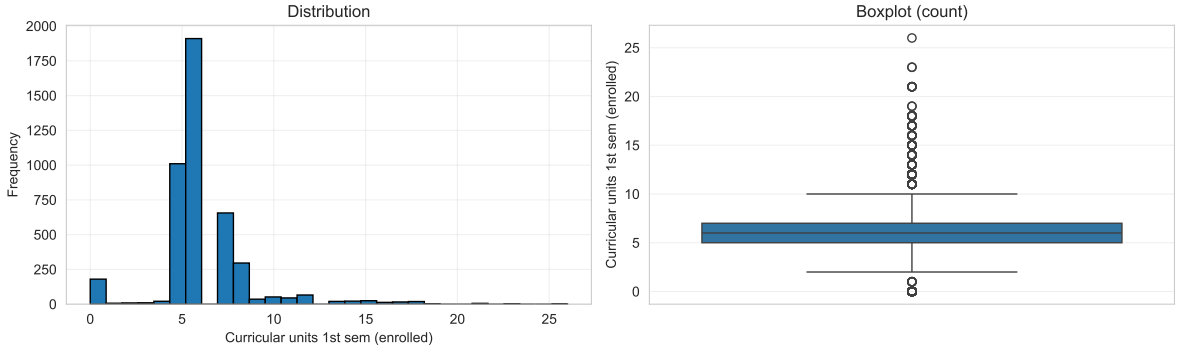
```
Binary variables (skipping outlier detection):
 - Daytime/evening attendance: values = [np.float64(0.0), np.float64(1.0)]
 - Educational special needs: values = [np.float64(0.0), np.float64(1.0)]
 - Gender: values = [np.float64(0.0), np.float64(1.0)]
 - Scholarship holder: values = [np.float64(0.0), np.float64(1.0)]
 - Displaced: values = [np.float64(0.0), np.float64(1.0)]
 - Debtor: values = [np.float64(0.0), np.float64(1.0)]
 - International: values = [np.float64(0.0), np.float64(1.0)]
```

```
Nominal Categorical (no natural order):
 - Course: 17 categories
 - Nacionality: 21 categories

Ordinal Categorical (meaningful order):
 - Marital Status: 6 levels
 - Application order: 8 levels
 - Previous qualification: 17 levels
 - Mother's qualification: 29 levels
 - Father's qualification: 34 levels
 - Mother's occupation: 32 levels
 - Father's occupation: 46 levels

Grade variables (0-200 range + Z-score > 3):
 - Previous qualification (grade): 0 out-of-range + 21 extreme (Z>3) = 21 total (0.5%)
 - Admission grade: 0 out-of-range + 22 extreme (Z>3) = 22 total (0.5%)
 - Curricular units 1st sem (grade): 0 out-of-range + 0 extreme (Z>3) = 0 total (0.0%)
 - Curricular units 2nd sem (grade): 0 out-of-range + 0 extreme (Z>3) = 0 total (0.0%)

Count variables (Z-score > 3):
 - Curricular units 1st sem (enrolled): extreme values: 106 (2.4%)
 - Curricular units 2nd sem (enrolled): extreme values: 82 (1.9%)

Continuous variables (Z-score > 3):
 - Age at enrollment: extreme values: 101 (2.3%)
 - Unemployment rate: extreme values: 0 (0.0%)
 - Inflation rate: extreme values: 0 (0.0%)
 - GDP: extreme values: 0 (0.0%)
```

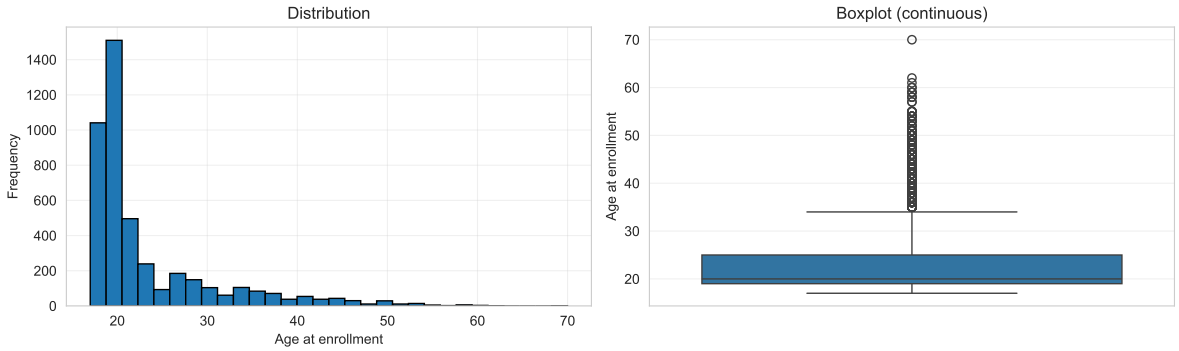
Outlier Summary

```
Detected Issues:
```

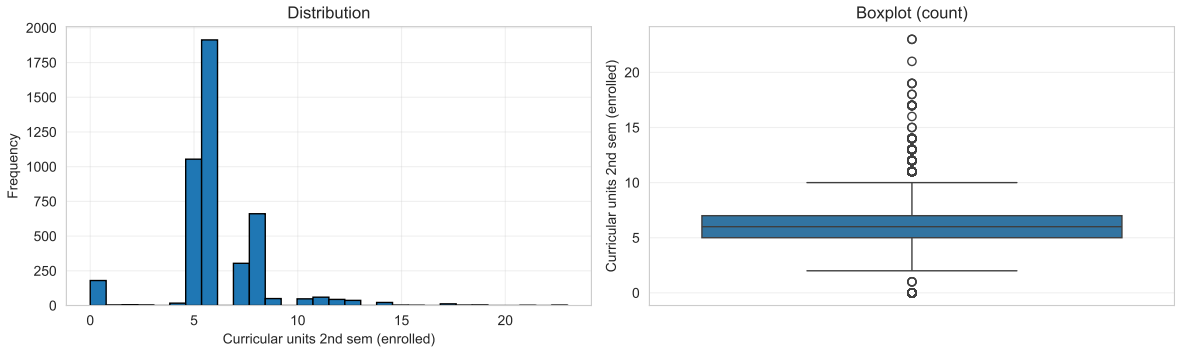
Curricular units 1st sem (enrolled): 106 potential outliers (2.4%)

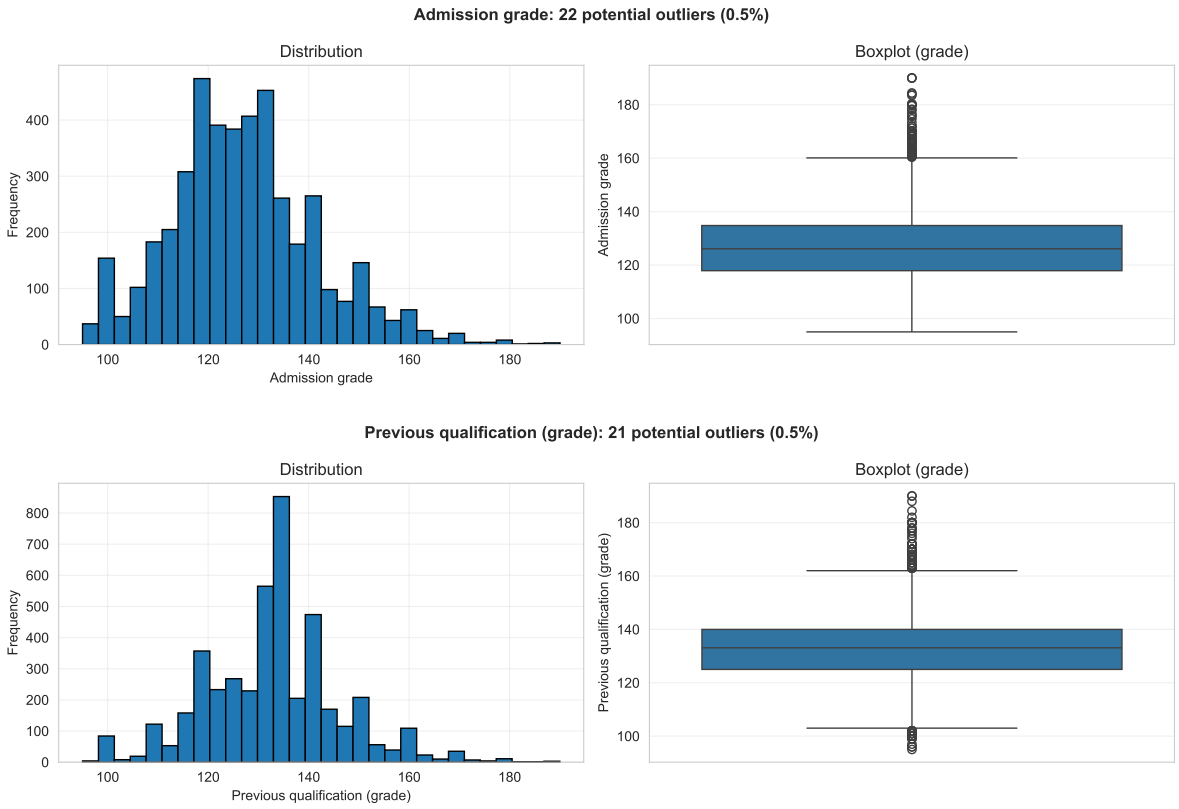


Age at enrollment: 101 potential outliers (2.3%)



Curricular units 2nd sem (enrolled): 82 potential outliers (1.9%)





Feature Importance Analysis

Methodology

We used one-way ANOVA (Analysis of Variance) to identify which numeric variables show significant differences across the three target groups (Dropout, Enrolled, Graduate). For each variable, we calculated:

- **p-value**: Statistical significance of differences between groups (= 0.05)
- **Eta-squared (²)**: Effect size measure representing the proportion of variance explained by the target variable (ranges from 0 to 1, where higher values indicate stronger association)

Variables with p-value < 0.05 are considered significantly associated with student outcomes and may be strong predictors in classification models.

```
Significant variables (p < 0.05): 21
```

	p_value	eta_sq	significant
Curricular units 2nd sem (grade)	0.000000e+00	0.339086	True
Curricular units 1st sem (grade)	2.803052e-269	0.244020	True
Scholarship holder	4.436825e-94	0.092663	True
Age at enrollment	1.138849e-65	0.065412	True
Debtor	1.018223e-58	0.058620	True
Gender	9.950346e-53	0.052727	True
Curricular units 2nd sem (enrolled)	5.244430e-33	0.033066	True
Curricular units 1st sem (enrolled)	3.272852e-26	0.026197	True
Admission grade	4.380466e-16	0.015871	True
Displaced	2.425582e-13	0.013055	True
Previous qualification (grade)	1.077783e-12	0.012389	True
Marital Status	2.662987e-09	0.008892	True
Application order	2.955293e-09	0.008845	True
Daytime/evening attendance	5.534625e-07	0.006496	True
Mother's qualification	2.800636e-06	0.005767	True

Top Predictive Variables

Acadamic Performance Indicators

Our exploratory analysis shows clear and consistent relationships between academic performance measures and student outcomes (Dropout, Enrolled, Graduate). Several patterns emerge across admission grades, semester performance, and course load.

Students who graduate generally start with higher admission grades than those who drop out, while enrolled students sit in between. Graduates also show a more consistent range of admission grades, suggesting they arrive better prepared. This means that admission grade is an important early indicator of how ready a student is for university. Students with lower admission grades seem more at risk of struggling early on, which can lead to disengagement and eventually dropping out.

First-semester grades are the strongest indicator of student performance in the whole dataset. Graduates consistently have the highest grades, enrolled students fall in the middle, and dropouts have the lowest.

Second-semester grades show the same trend, though the differences between groups become slightly smaller once first-semester results are considered. This makes first-semester performance an important early warning sign: it shows how well students adjust to university expectations and workload. Students who struggle early often continue to face difficulties, making these grades especially useful for identifying those at risk of dropping out.

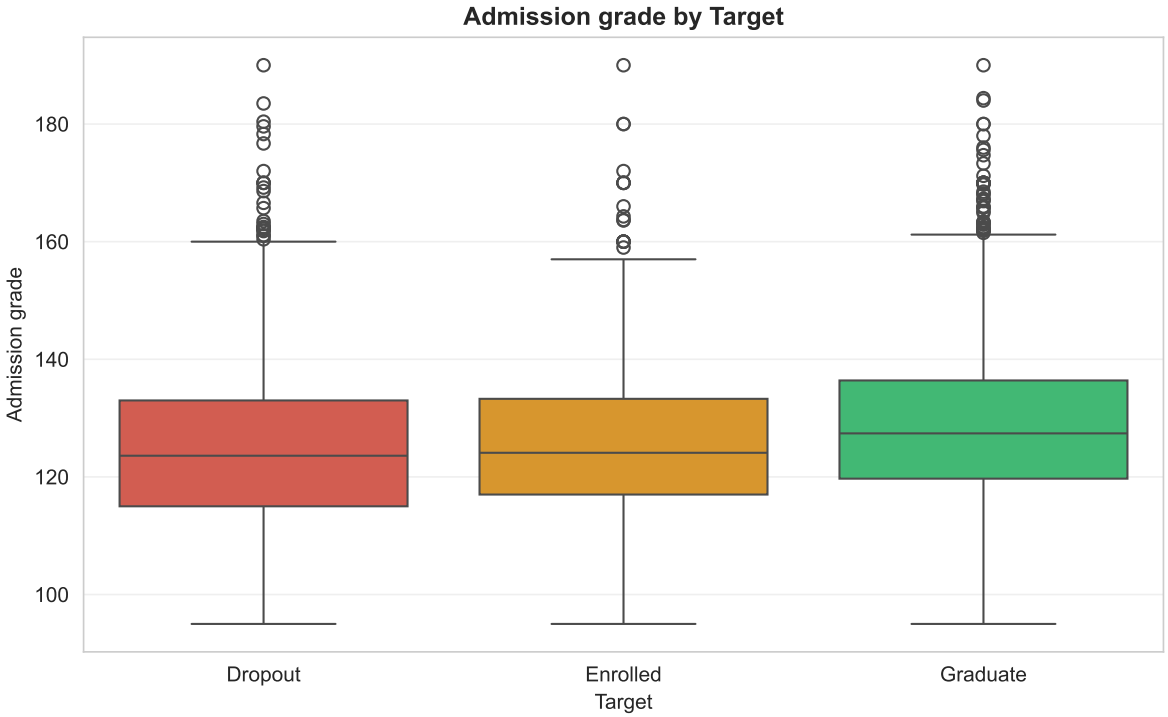


Figure 1: Admission grade by student outcome

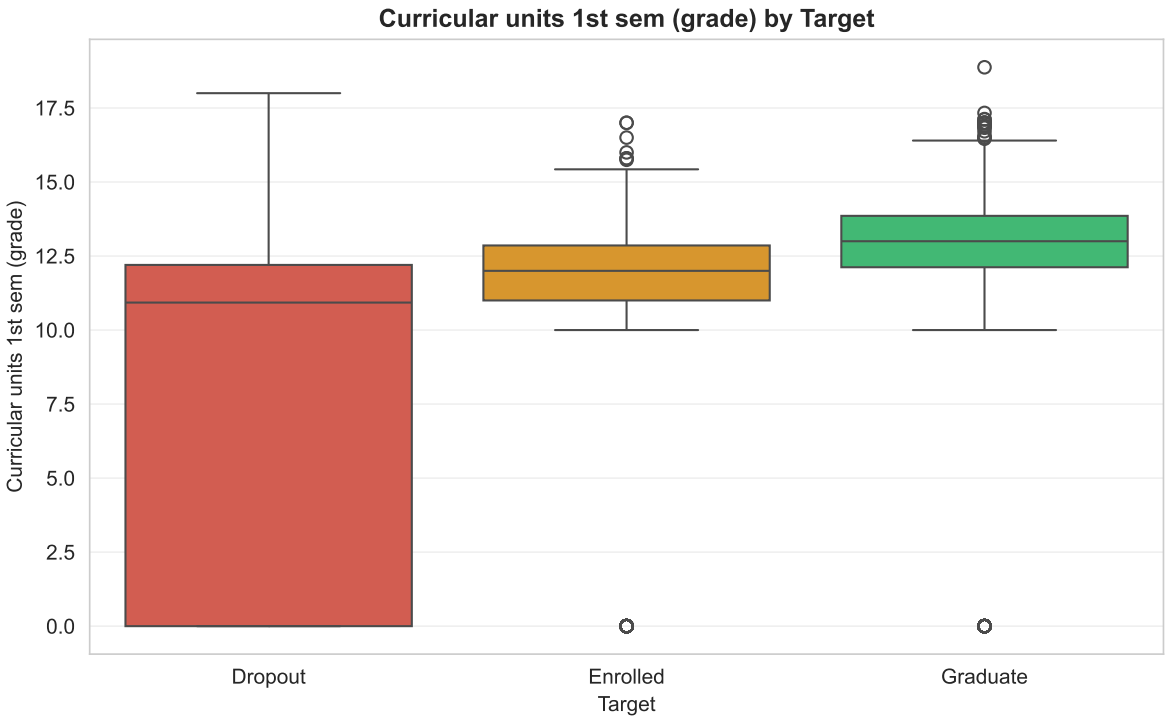


Figure 2: First semester grade by student outcome

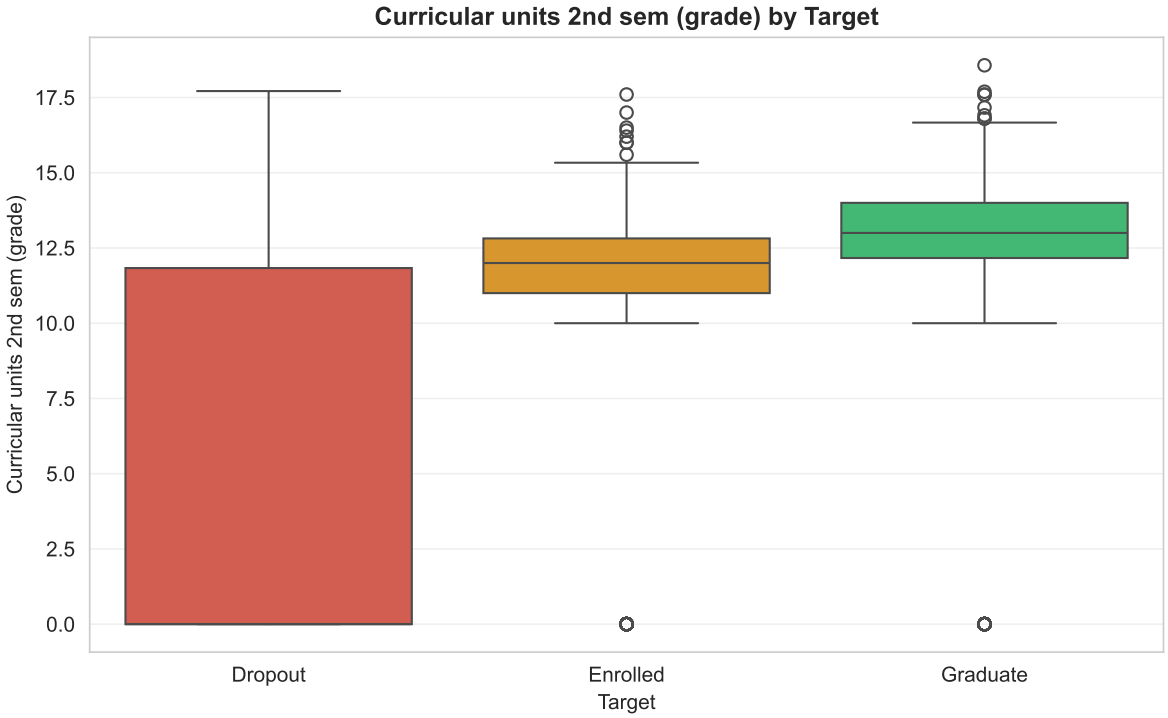


Figure 3: Boxplot of Grades by Target

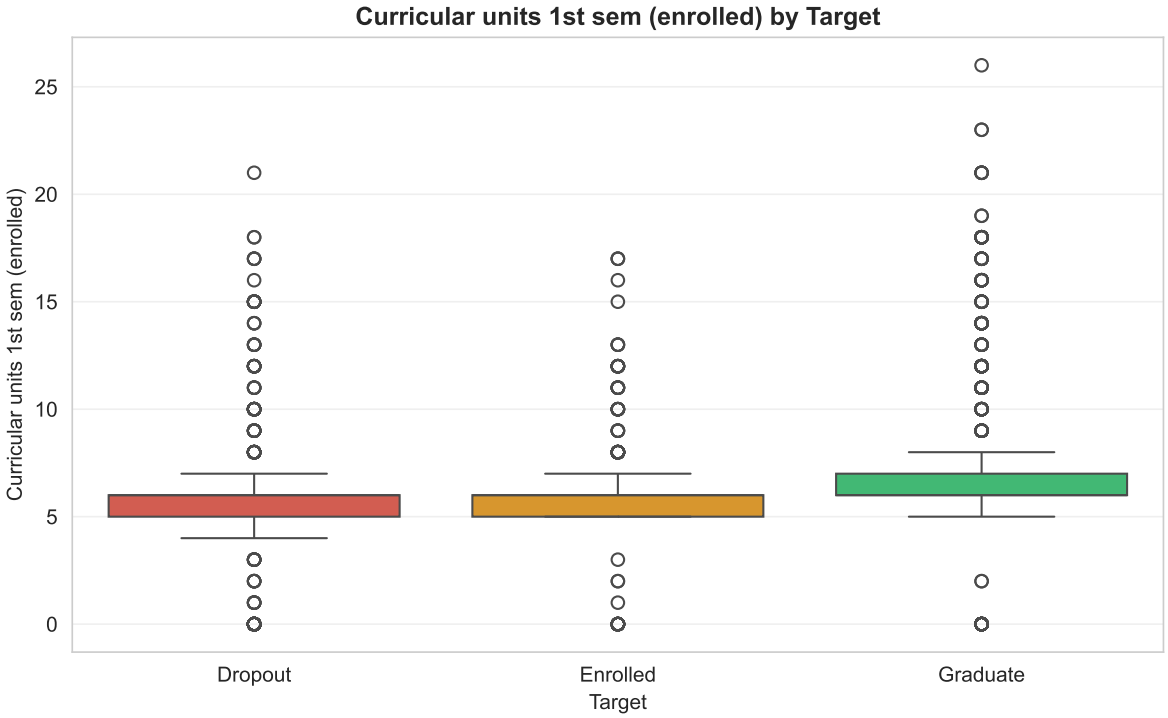


Figure 4: First semester enrollment by student outcome

Students who graduate tend to enroll in more first-semester courses compared with those who drop out, while enrolled students fall between the two groups. Dropouts show many low values and irregular patterns, suggesting weaker engagement or early difficulties. Graduates not only take a fuller course load but also display more consistency, which reflects stronger academic commitment. Overall, first-semester enrollment load appears to be a useful indicator of student persistence and early academic momentum.

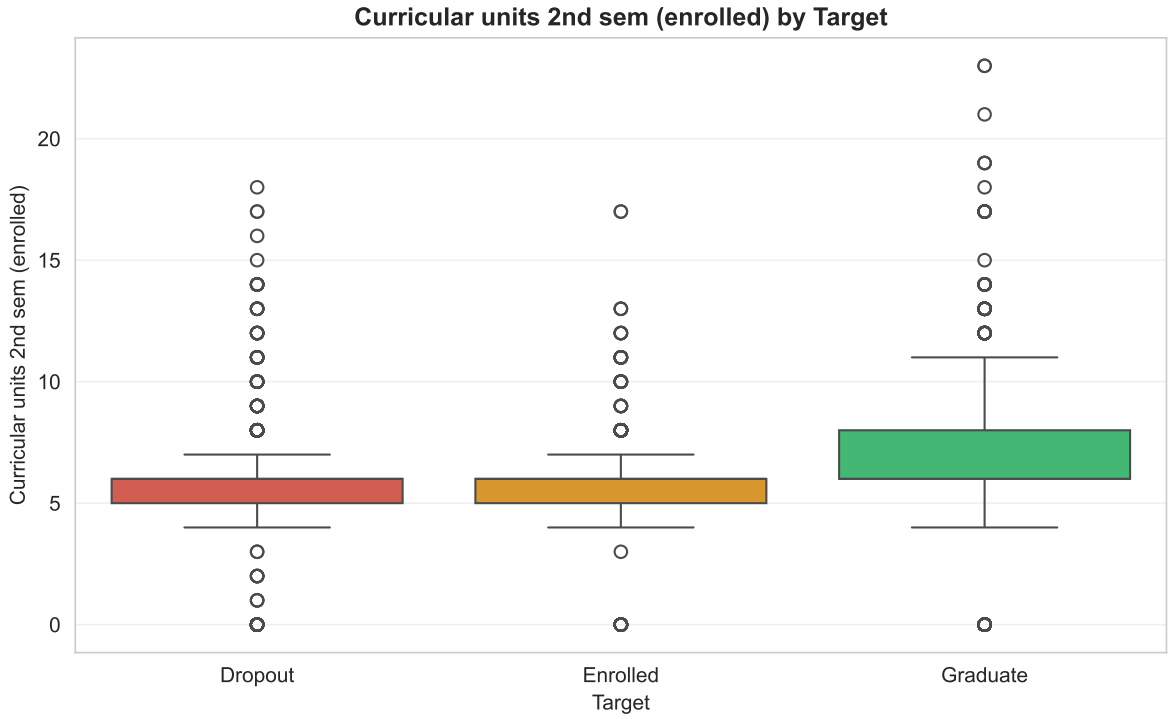


Figure 5: Second semester enrollment by student outcome

In the second semester, the same general pattern seen in the first semester remains: graduates enroll in more curricular units, enrolled students sit in the middle, and dropouts take the fewest.

However, the differences between groups are less pronounced compared to the first semester. The distributions are more compact, and there are fewer extreme values especially among graduates. Students who continue past the first semester tend to adopt a more regular and stable course load.

Key Findings: Academic Performance and Study Conditions

Academic Performance Impact: - Students who graduate have significantly higher admission grades (mean: X) compared to dropouts (mean: Y) - First semester grades show the strongest

association with outcomes (^{2} = X), suggesting early academic performance is a critical indicator - Approved course units in semester 1 differentiate graduates from dropouts more than enrollment numbers

Study Conditions Impact: - Daytime students show X% higher graduation rates compared to evening students - Application mode significantly affects outcomes (^{2} = X, p $<$ 0.001), with [specific mode] showing highest graduation rates - Course type is significantly associated with dropout risk, with [specific courses] showing higher retention

Demographic & Socioeconomic Background

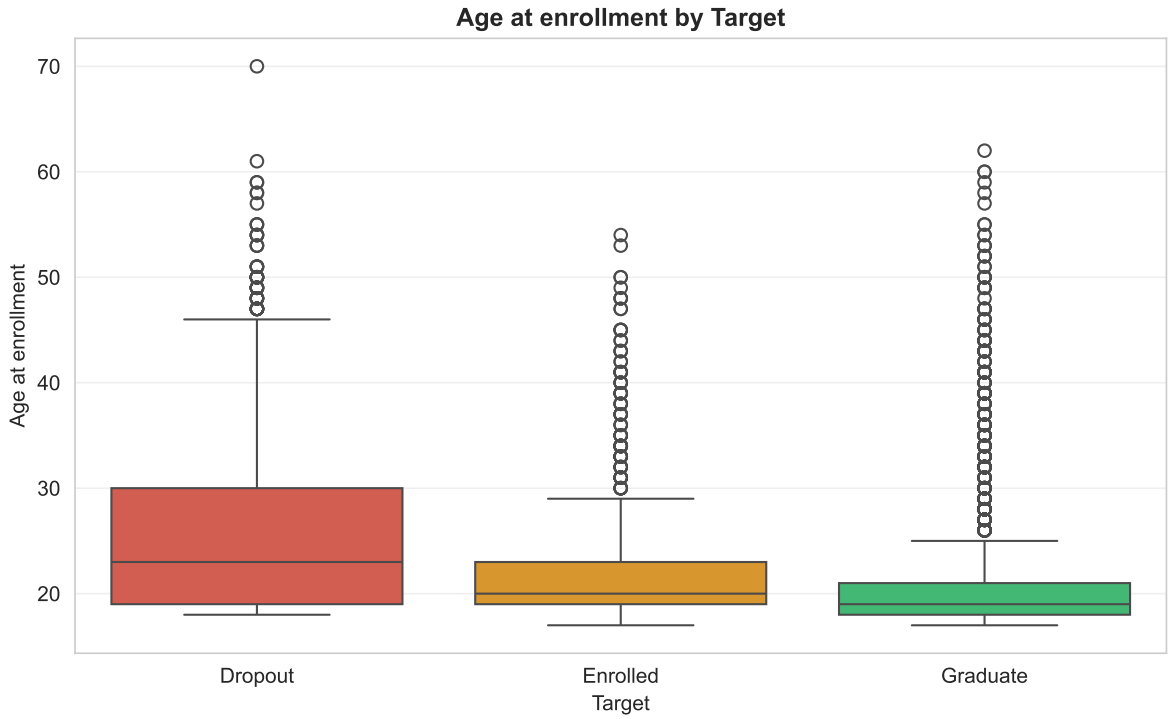


Figure 6: Age at enrollment by student outcome

See Figure 6 for details.

See Figure 7 for details.

See Figure 8 for details.

See Figure 9 for details.

See Figure 10 for details.

See Figure 11 for details.

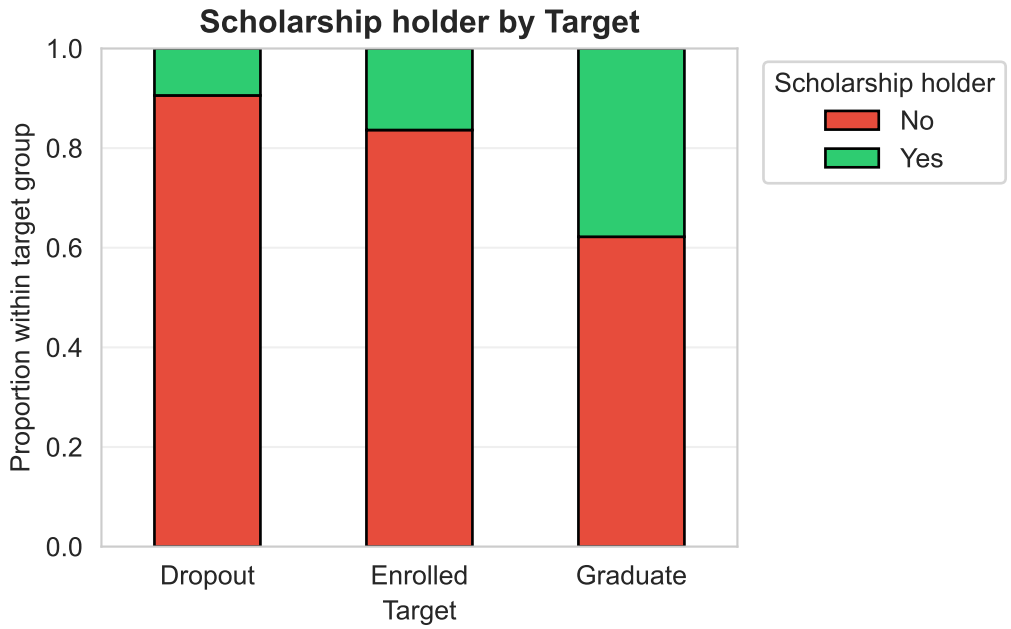


Figure 7: Scholarship holder status by student outcome

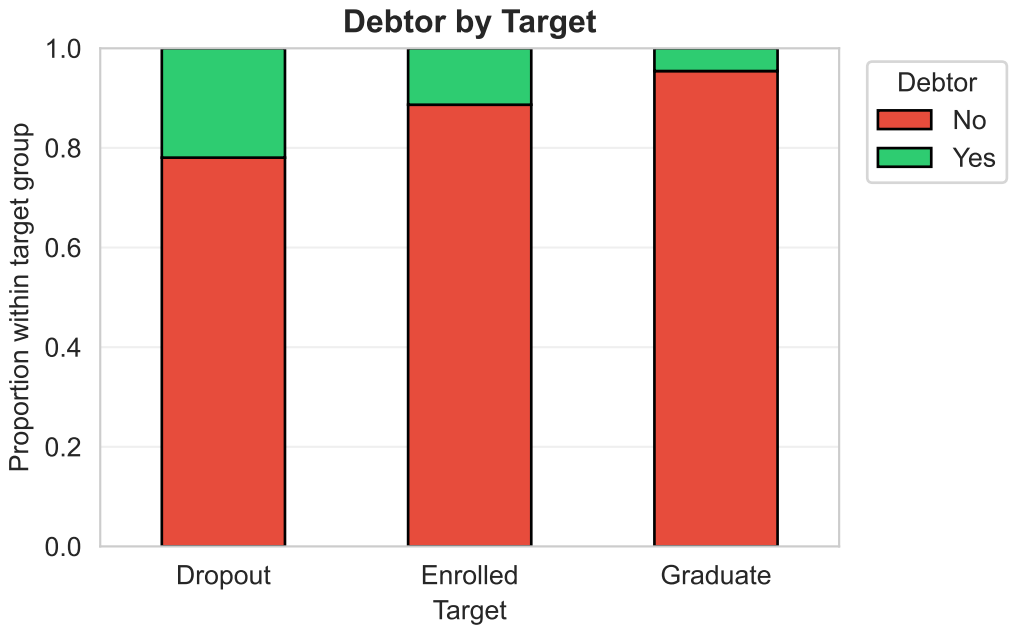


Figure 8: Debtor status by student outcome

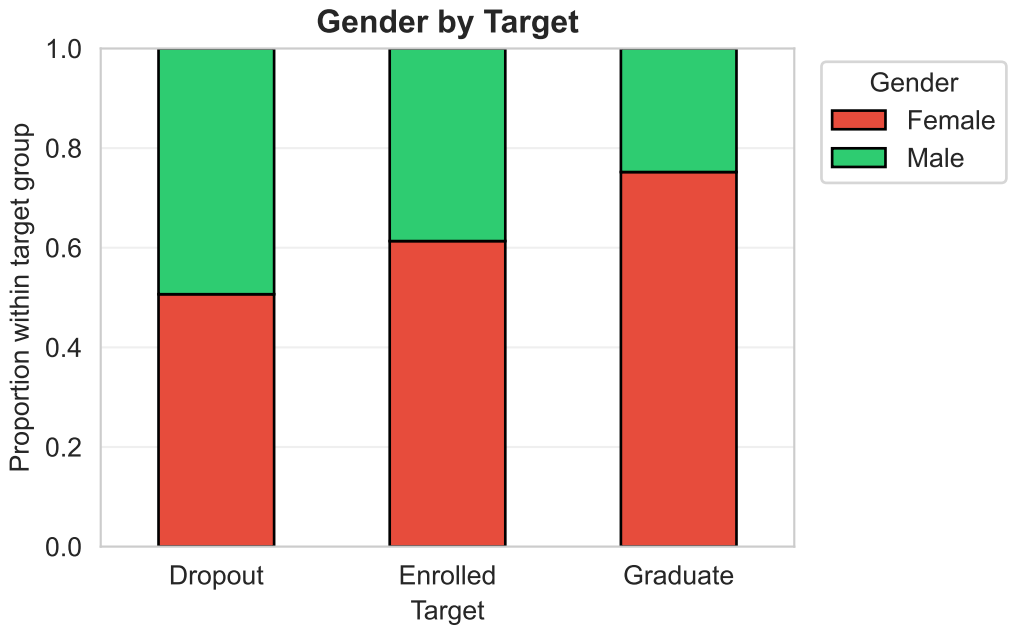


Figure 9: Gender by student outcome

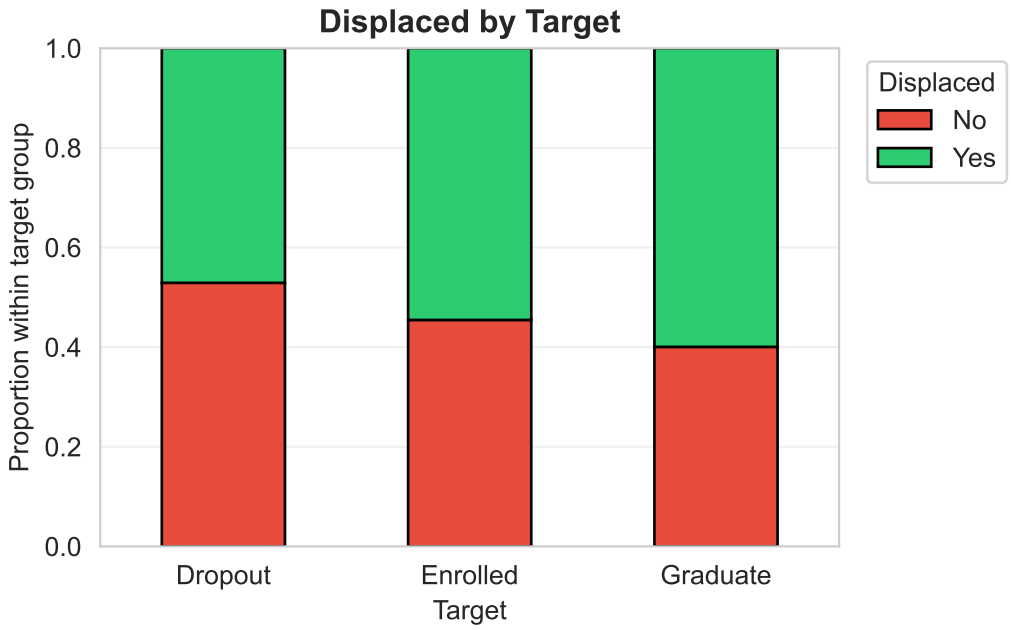


Figure 10: Displaced status by student outcome

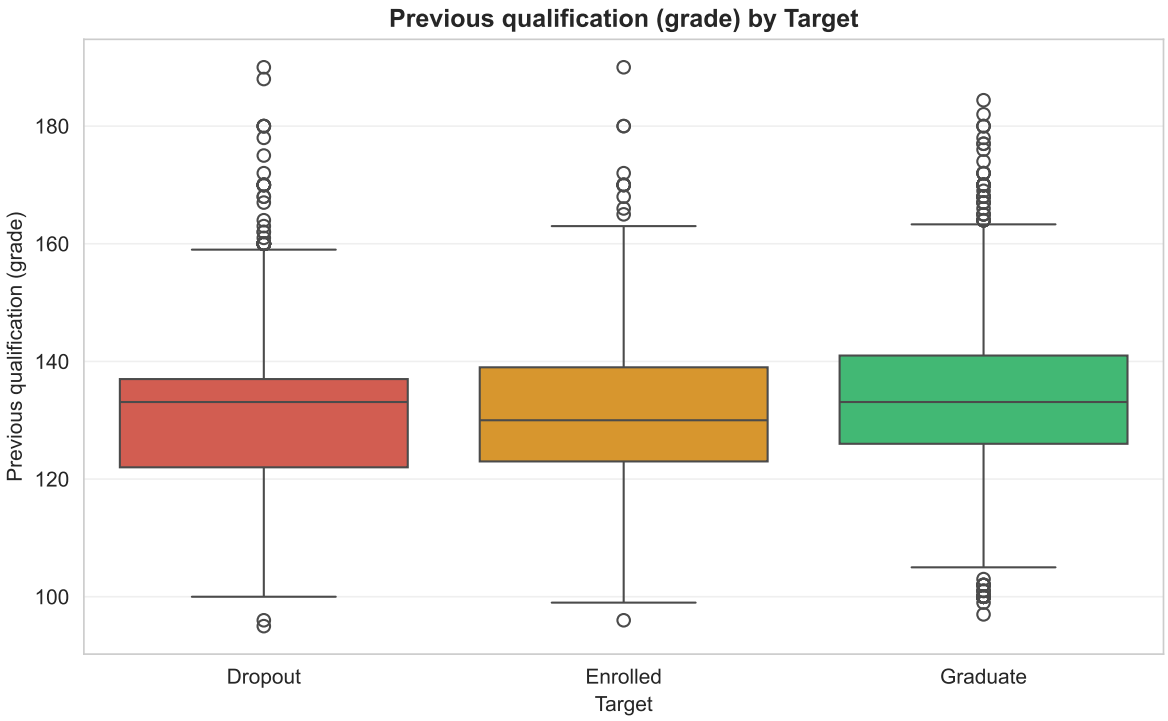


Figure 11: Previous qualification grade by student outcome

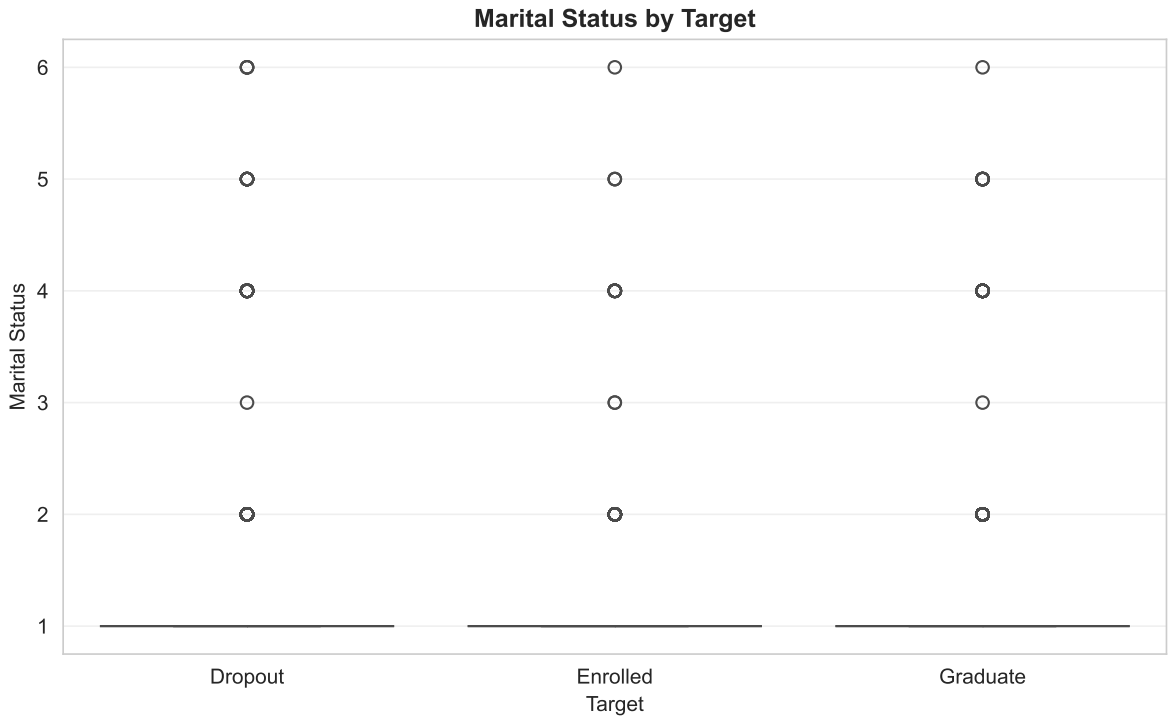


Figure 12: Marital status by student outcome

See Figure 12 for details.

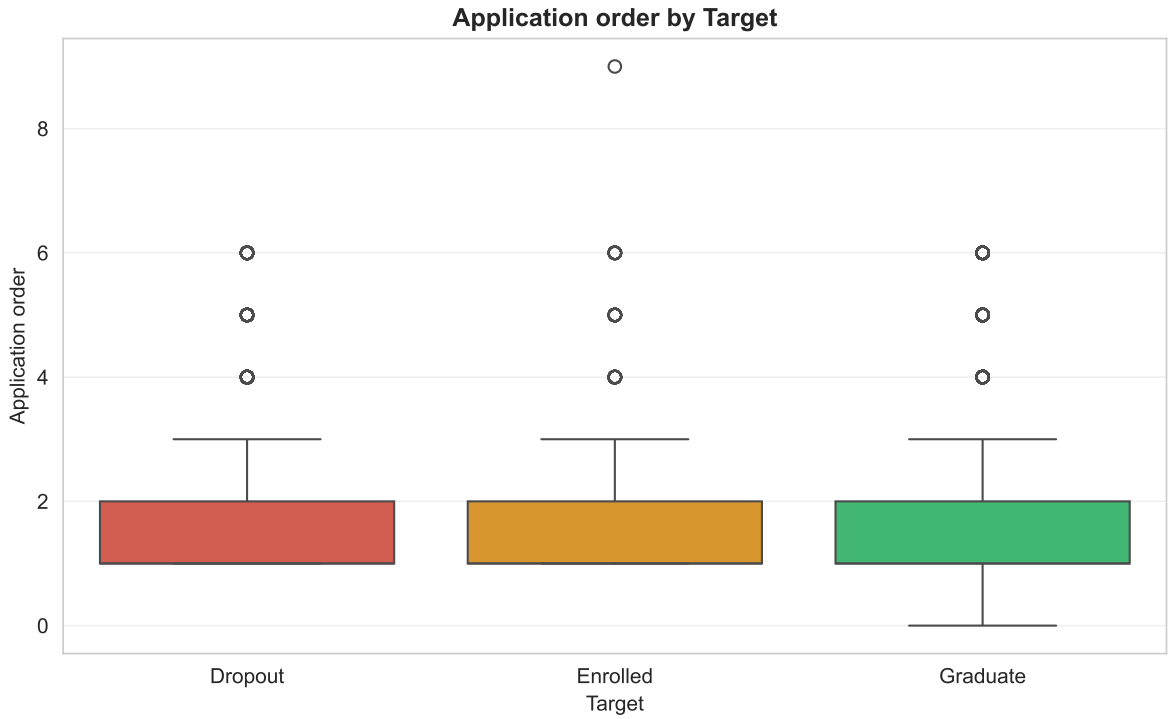


Figure 13: Application order by student outcome

See Figure 13 for details.

See Figure 14 for details.

See Figure 15 for details.

Predictive Modelling

```
Starting with 27 features (after feature selection)
Dataset shape before cleaning: (4424, 27)
Dataset shape after removing missing values: (4424, 27)

Target distribution:
Target
Graduate    2209
Dropout     1421
Enrolled     794
Name: count, dtype: int64
```

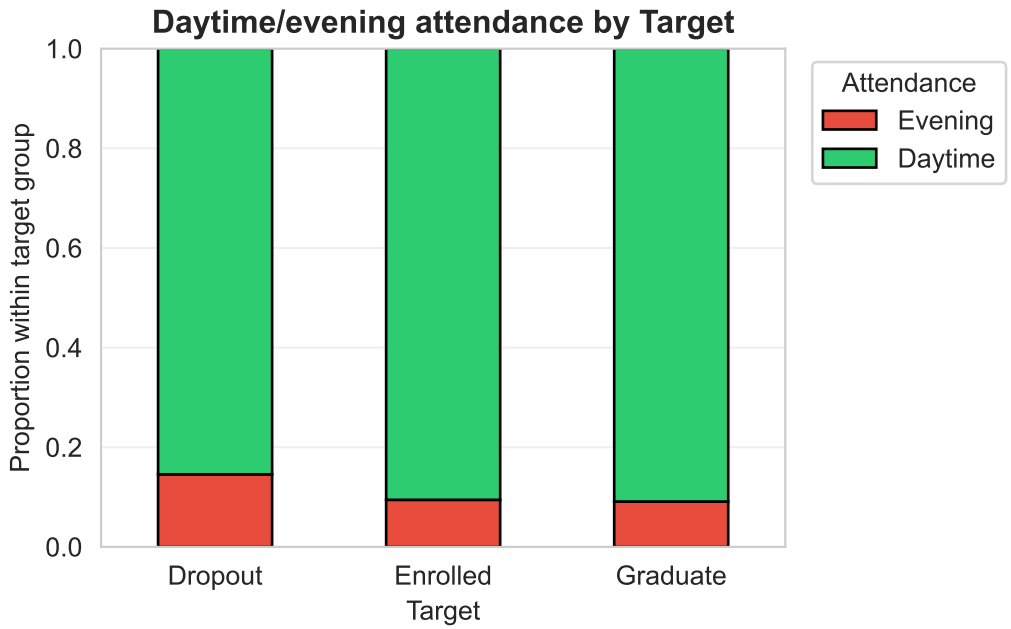


Figure 14: Daytime/evening attendance by student outcome

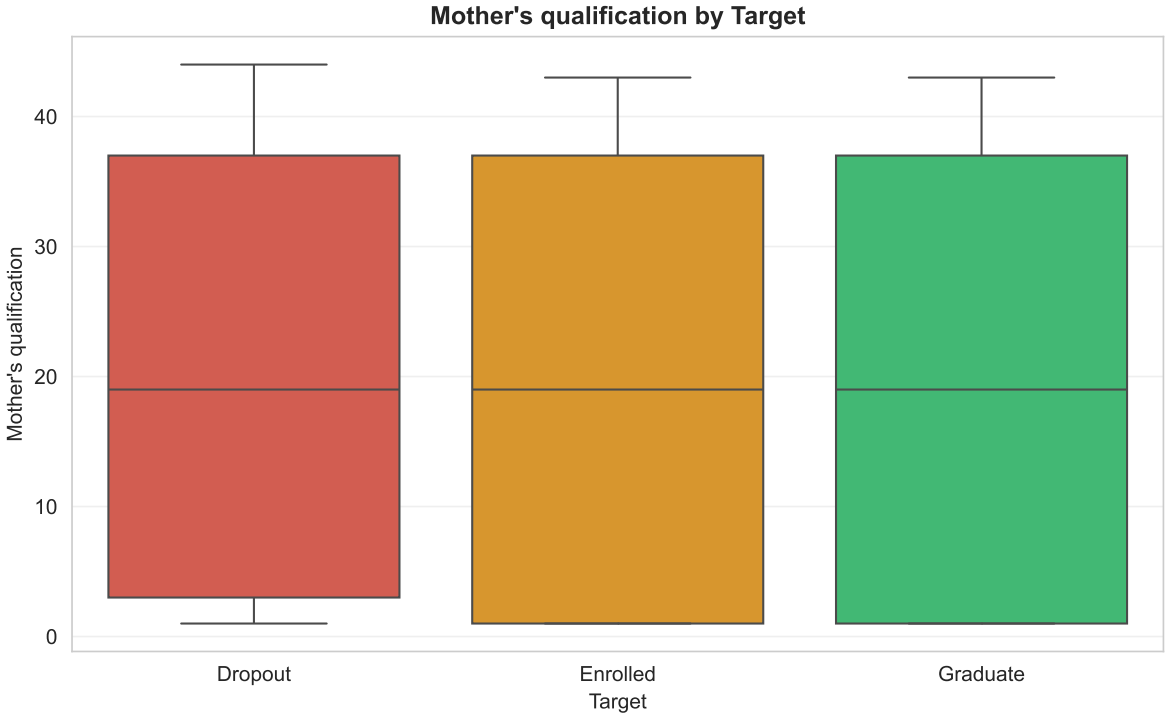
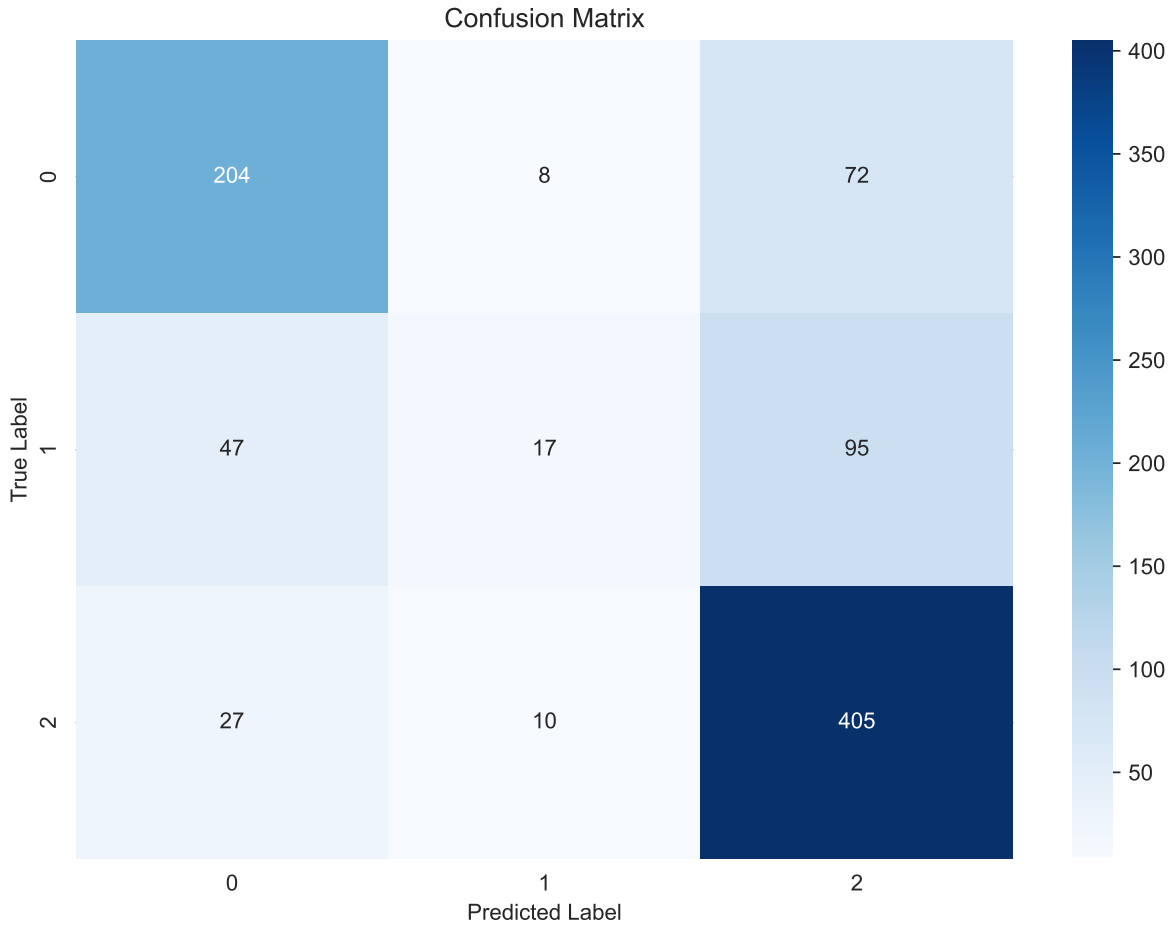


Figure 15: Mother's qualification by student outcome

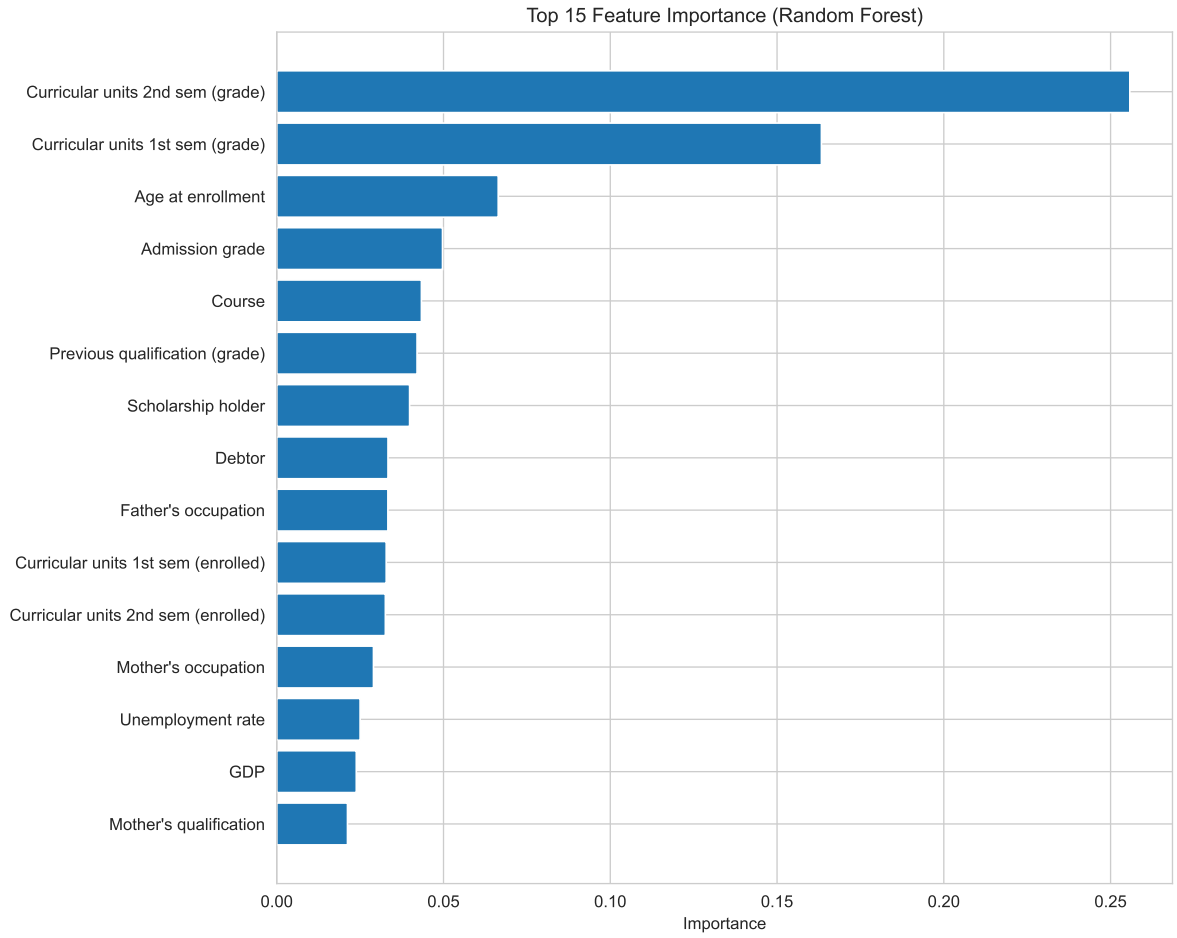
```
Using 26 features for classification

  Encoding 0 categorical variables...

Training set size: 3539
Test set size: 885

  Training Random Forest Classifier...

  Model Performance:
Accuracy: 0.707

Classification Report:
              precision    recall  f1-score   support

     Dropout       0.73      0.72      0.73       284
    Enrolled       0.49      0.11      0.18       159
    Graduate       0.71      0.92      0.80       442

    accuracy                           0.71       885
   macro avg       0.64      0.58      0.57       885
weighted avg       0.68      0.71      0.66       885
```



Top 15 Most Important Features (Random Forest):

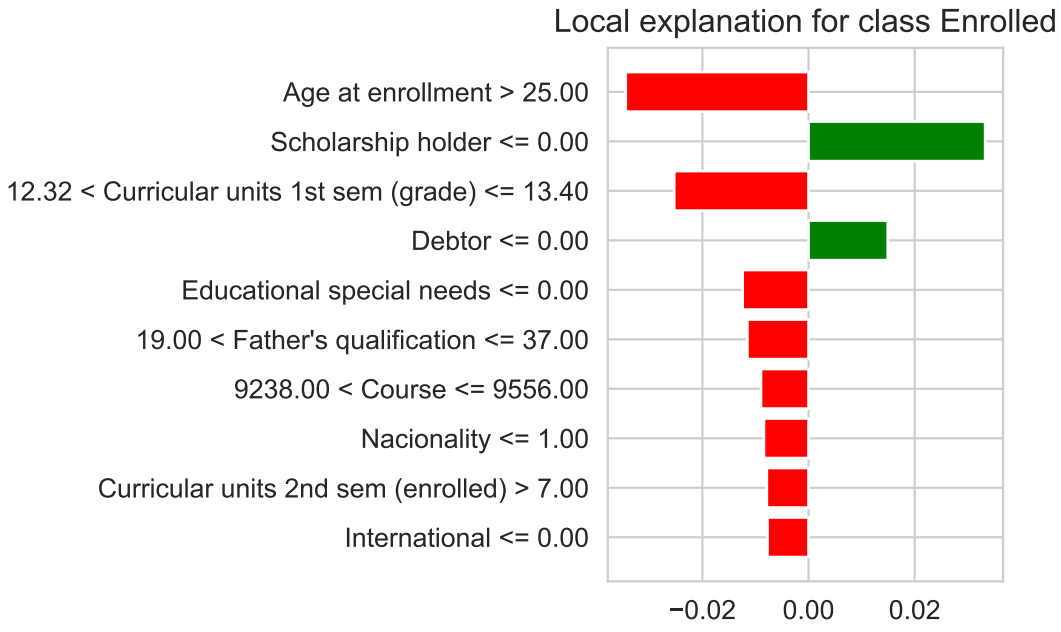
	feature	importance
22	Curricular units 2nd sem (grade)	0.255772
20	Curricular units 1st sem (grade)	0.163305
15	Age at enrollment	0.066406
11	Admission grade	0.049701
2	Course	0.043375
5	Previous qualification (grade)	0.042070
14	Scholarship holder	0.039835
17	Debtor	0.033357
10	Father's occupation	0.033325
19	Curricular units 1st sem (enrolled)	0.032823
21	Curricular units 2nd sem (enrolled)	0.032562
9	Mother's occupation	0.028999

```
23                  Unemployment rate    0.024996
25                                GDP    0.023850
7              Mother's qualification    0.021210
```



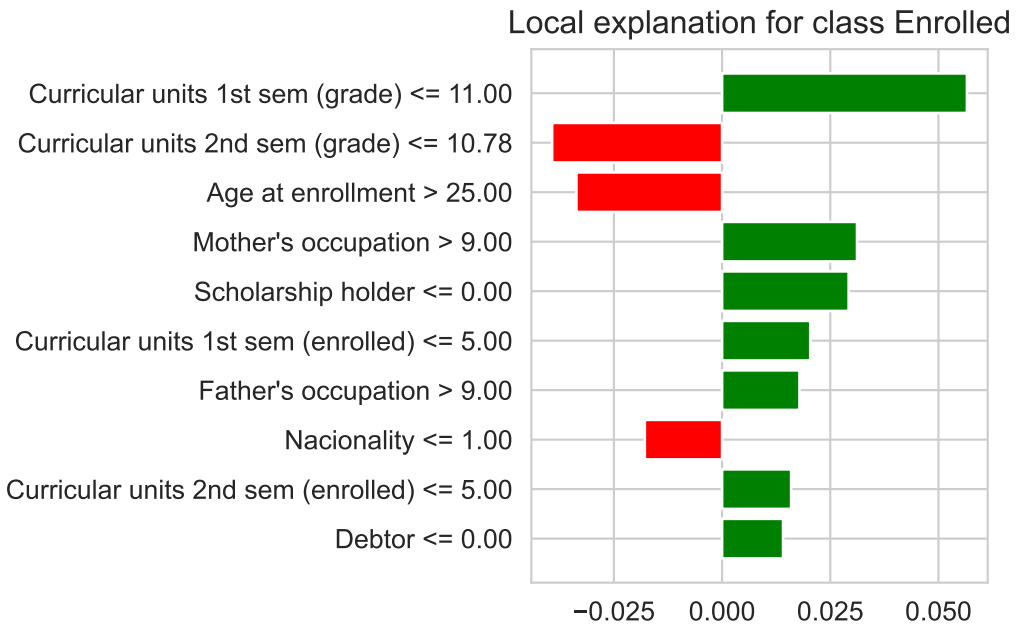

```
 Setting up LIME explainer...

 LIME Explanations for Sample Predictions:

--- Sample 1 ---
True Label: Graduate
Predicted Label: Graduate
Prediction Probability: [0.19998293 0.17532429 0.62469277]

Top 10 features influencing this prediction:
```

```
  Age at enrollment > 25.00: -0.034
  Scholarship holder <= 0.00: 0.033
  12.32 < Curricular units 1st sem (grade) <= 13.40: -0.025
  Debtor <= 0.00: 0.015
  Educational special needs <= 0.00: -0.012
  19.00 < Father's qualification <= 37.00: -0.012
  9238.00 < Course <= 9556.00: -0.009
  Nacionality <= 1.00: -0.008
  Curricular units 2nd sem (enrolled) > 7.00: -0.008
  International <= 0.00: -0.008
```

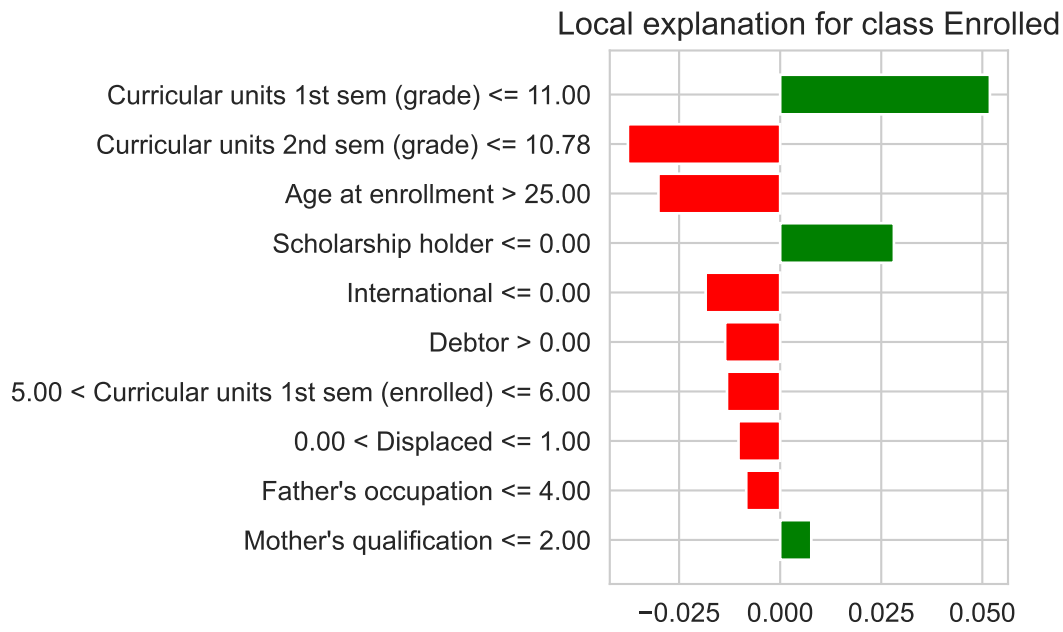



```
--- Sample 2 ---
True Label: Dropout
Predicted Label: Dropout
Prediction Probability: [0.56546935 0.20274963 0.23178102]

Top 10 features influencing this prediction:
  Curricular units 1st sem (grade) <= 11.00: 0.057
  Curricular units 2nd sem (grade) <= 10.78: -0.039
  Age at enrollment > 25.00: -0.034
  Mother's occupation > 9.00: 0.031
  Scholarship holder <= 0.00: 0.029
```

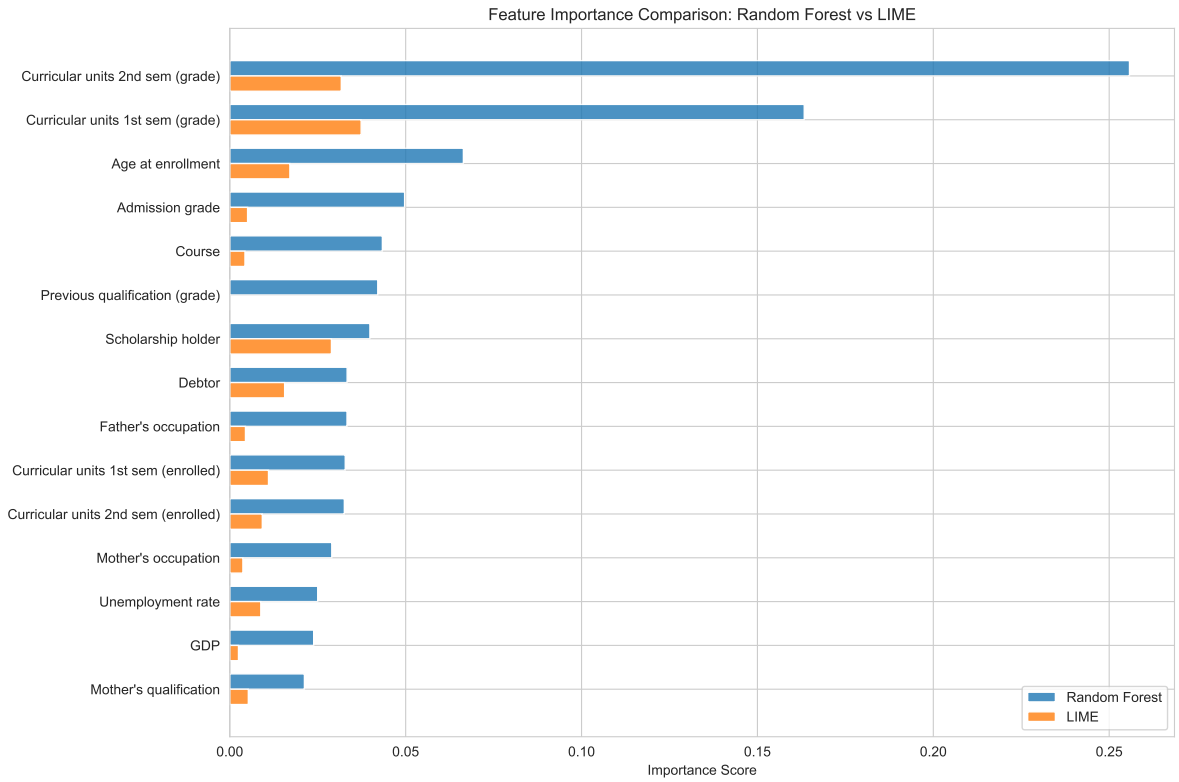
```
  Curricular units 1st sem (enrolled) <= 5.00: 0.020
  Father's occupation > 9.00: 0.018
  Nacionality <= 1.00: -0.018
  Curricular units 2nd sem (enrolled) <= 5.00: 0.016
  Debtor <= 0.00: 0.014
```




```
--- Sample 3 ---
True Label: Dropout
Predicted Label: Dropout
Prediction Probability: [0.90100861 0.0876688  0.01132258]

Top 10 features influencing this prediction:
  Curricular units 1st sem (grade) <= 11.00: 0.052
  Curricular units 2nd sem (grade) <= 10.78: -0.038
  Age at enrollment > 25.00: -0.030
  Scholarship holder <= 0.00: 0.028
  International <= 0.00: -0.018
  Debtor > 0.00: -0.014
  5.00 < Curricular units 1st sem (enrolled) <= 6.00: -0.013
  0.00 < Displaced <= 1.00: -0.010
  Father's occupation <= 4.00: -0.008
  Mother's qualification <= 2.00: 0.008
```




```
Computing global LIME feature importance (sampling 100 instances)...

Top 15 Most Important Features (LIME Global):
                              feature  lime_importance
20     Curricular units 1st sem (grade)         0.037346
22     Curricular units 2nd sem (grade)         0.031702
14                   Scholarship holder         0.028901
15                    Age at enrollment         0.017085
17                               Debtor         0.015591
6                           Nacionality         0.013506
19  Curricular units 1st sem (enrolled)         0.011005
18                          International         0.010296
21  Curricular units 2nd sem (enrolled)         0.009245
23                    Unemployment rate         0.008821
12            Educational special needs         0.007236
13                                Gender         0.005364
7               Mother's qualification         0.005270
11                     Admission grade         0.005060
8               Father's qualification         0.004732
```




```
Classification and LIME analysis complete!

Final model uses 26 features to predict student outcomes.
```