Academic Success Prediction

GHAZAL VAUGHAN

JUNE 2024

What happens if the academic success rate is low?

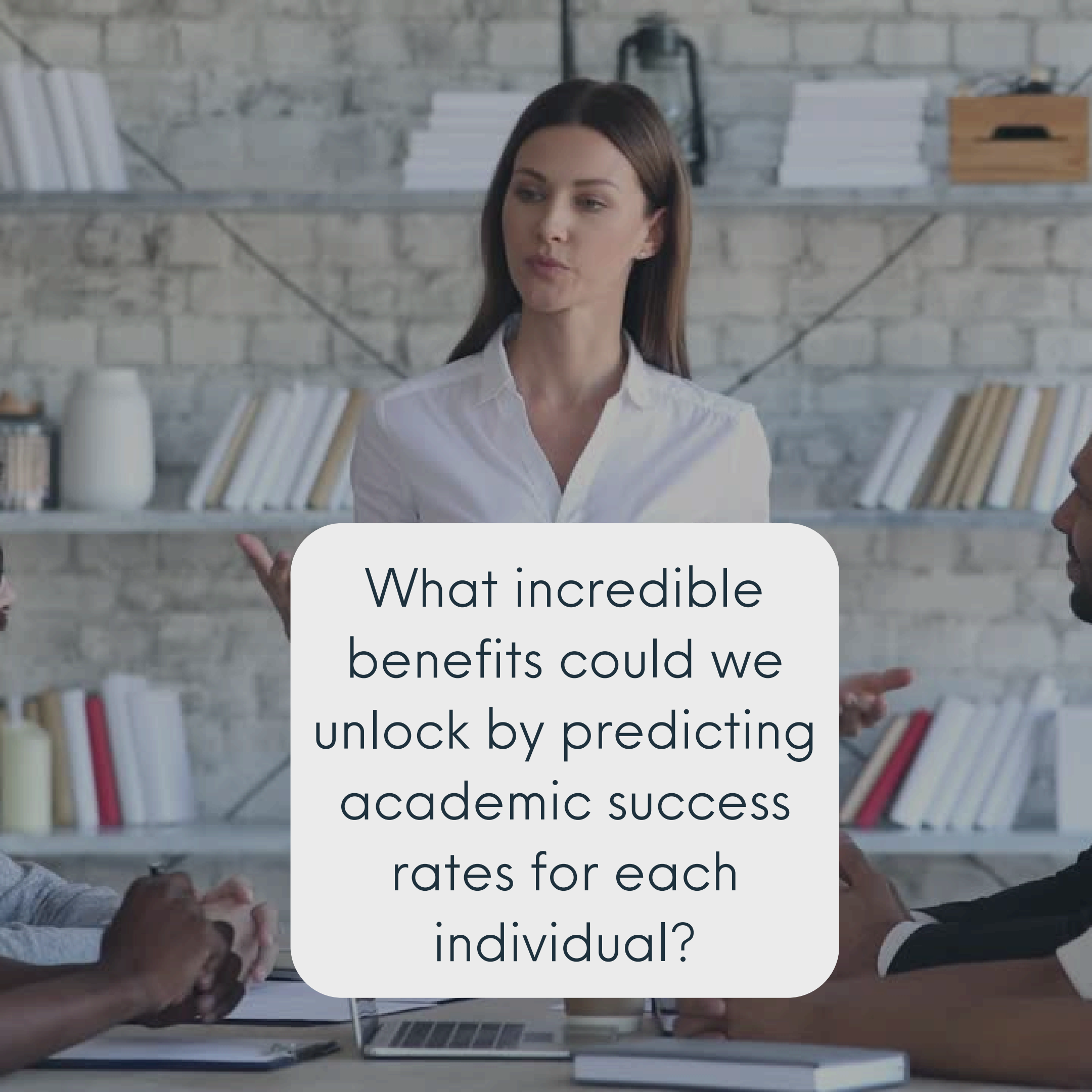
In general, demographic factors can influence the academic success rate in higher education.

High dropout rates can negatively impact educational institutions in many ways, including their Reputation, and their revenue.

Each year's class of dropouts will cost the country over $200 billion during their lifetimes in lost earnings and unrealized tax revenue (Catterall, 1985).

The estimated tax revenue loss from every male between the ages of 25 and 34 years of age who did not complete high school would be approximately $944 billion, with cost increases to public welfare and crime at $24 billion (Thorstensen, 2004).

>>> https://dropoutprevention.org <<<



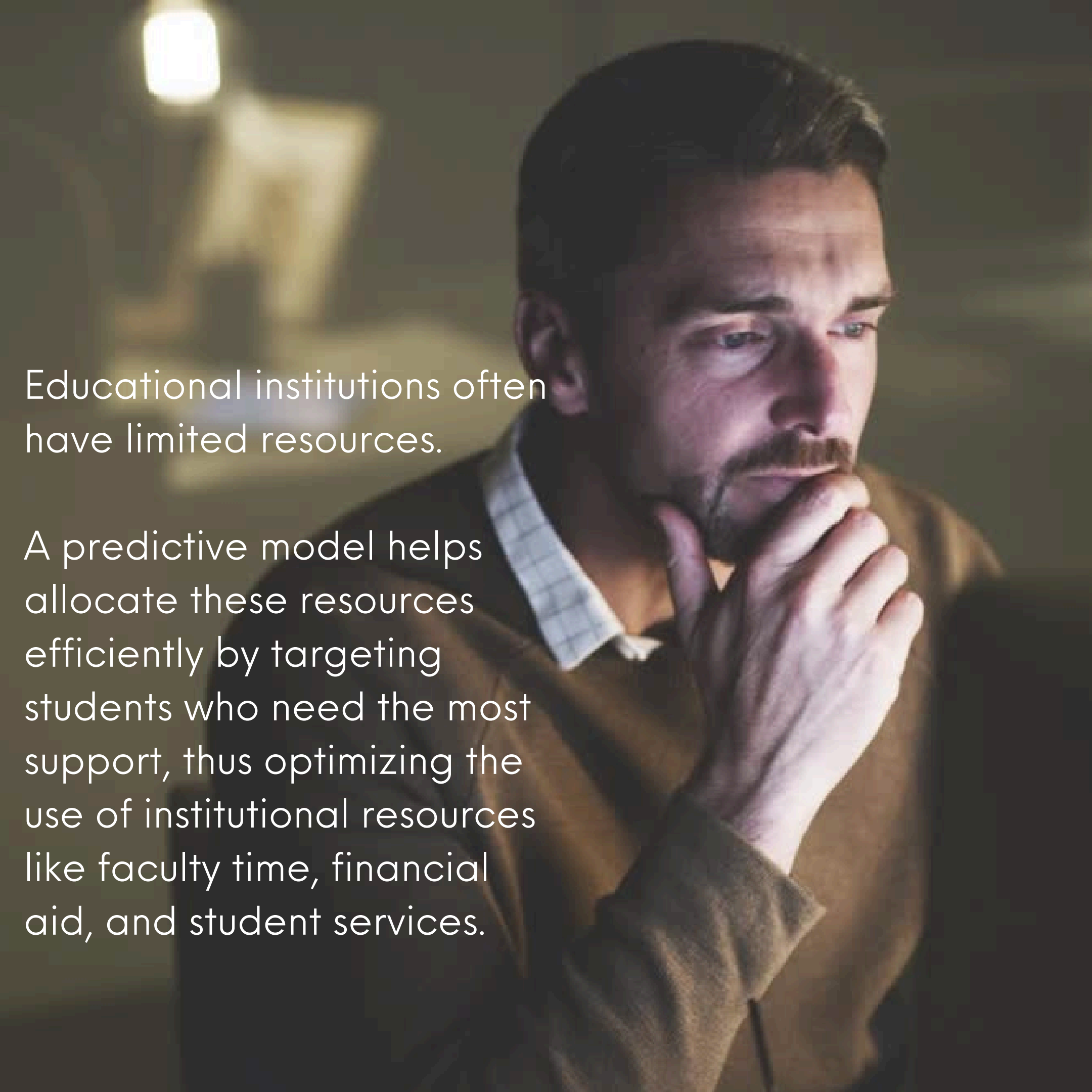
What incredible benefits could we unlock by predicting academic success rates for each individual?

Boosting Reputation

Higher retention rates are often correlated with better institutional rankings and reputations. By increasing the number of students who stay and graduate, schools enhance their standing, which can attract more applicants and potentially increase funding and resources.

Proactive Interventions

By identifying students at risk of dropping out early, institutions can intervene proactively. This could involve providing tailored support services, such as tutoring, counseling, and academic advising, specifically targeted to the needs of those students.

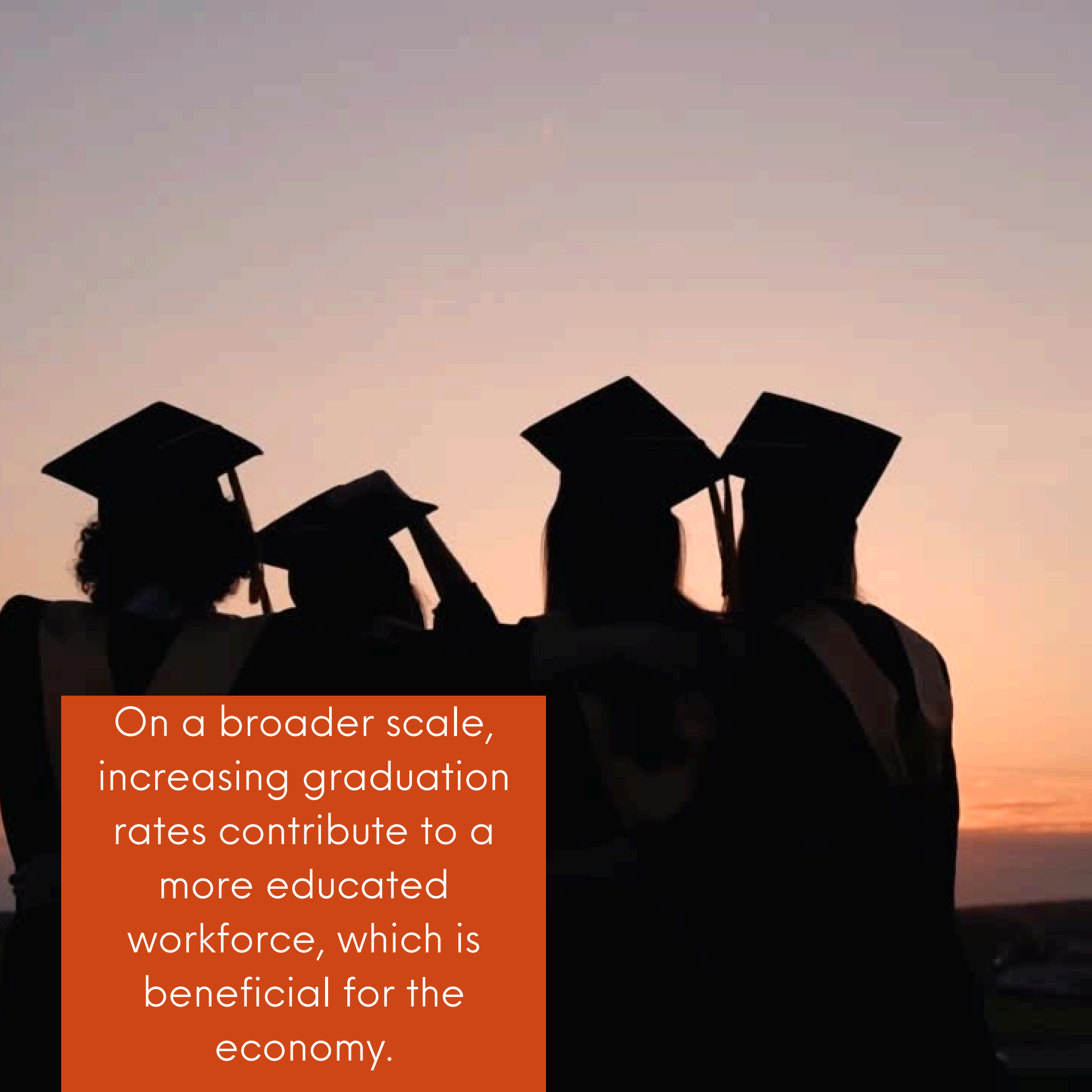


Educational institutions often have limited resources.

A predictive model helps allocate these resources efficiently by targeting students who need the most support, thus optimizing the use of institutional resources like faculty time, financial aid, and student services.

In many regions, government funding for educational institutions is tied to performance indicators such as graduation rates.

By using predictive analytics to improve these rates, schools can ensure compliance with performance benchmarks and secure or increase governmental funding



On a broader scale, increasing graduation rates contribute to a more educated workforce, which is beneficial for the economy.

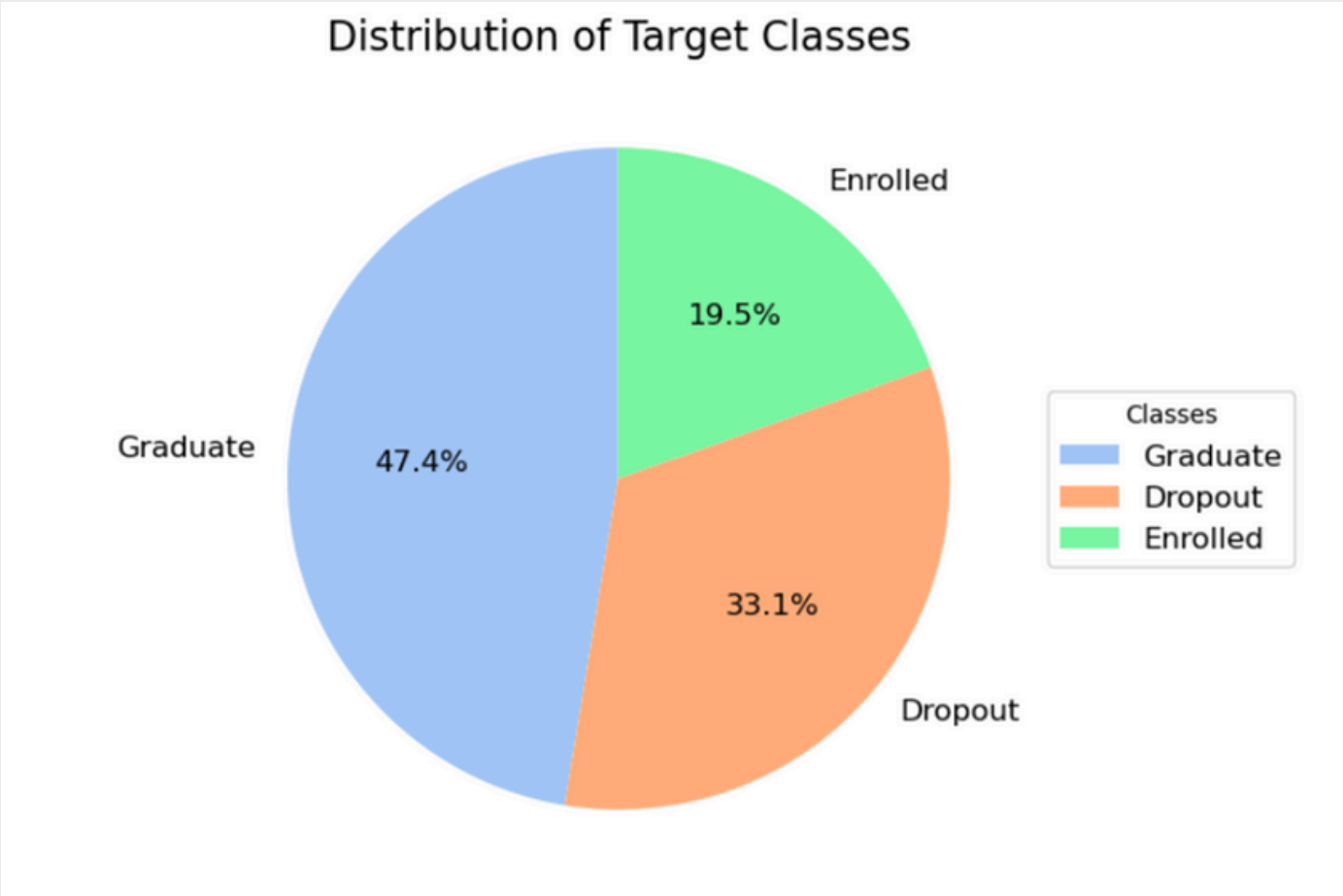
OBJECTIVE

The objective of this project is to develop a predictive model using over 76,500 synthetic data points to classify individuals into three categories based on their academic status: graduated, dropped out, or still enrolled. The prediction is made by analyzing various features related to the individual's academic and personal backgrounds.

DATA

The dataset includes information at the time of student enrollment (academic path, demographics, and socioeconomic factors) and the student's academic performance at the end of the first and second semesters.

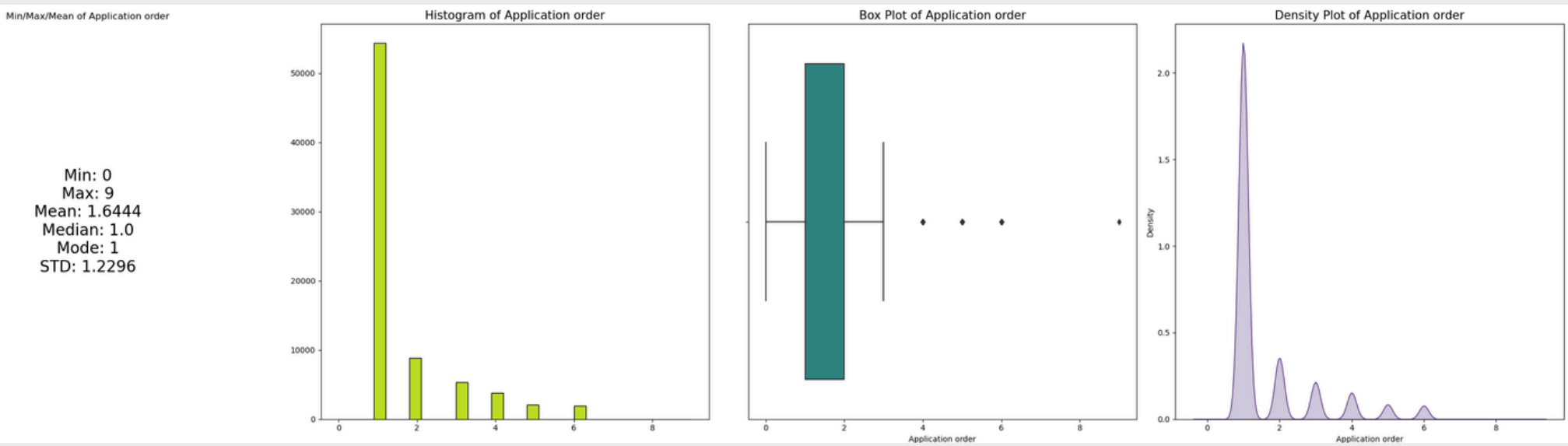
>>> https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success <<<



Synthetic Minority Over-sampling Technique (SMOTE) used to oversample the minority class shown above.

```
Original training set class distribution: Counter({'Graduate': 29023, 'Dropout': 20268, 'Enrolled': 11923})
Resampled training set class distribution: Counter({'Graduate': 29023, 'Dropout': 29023, 'Enrolled': 29023})
```

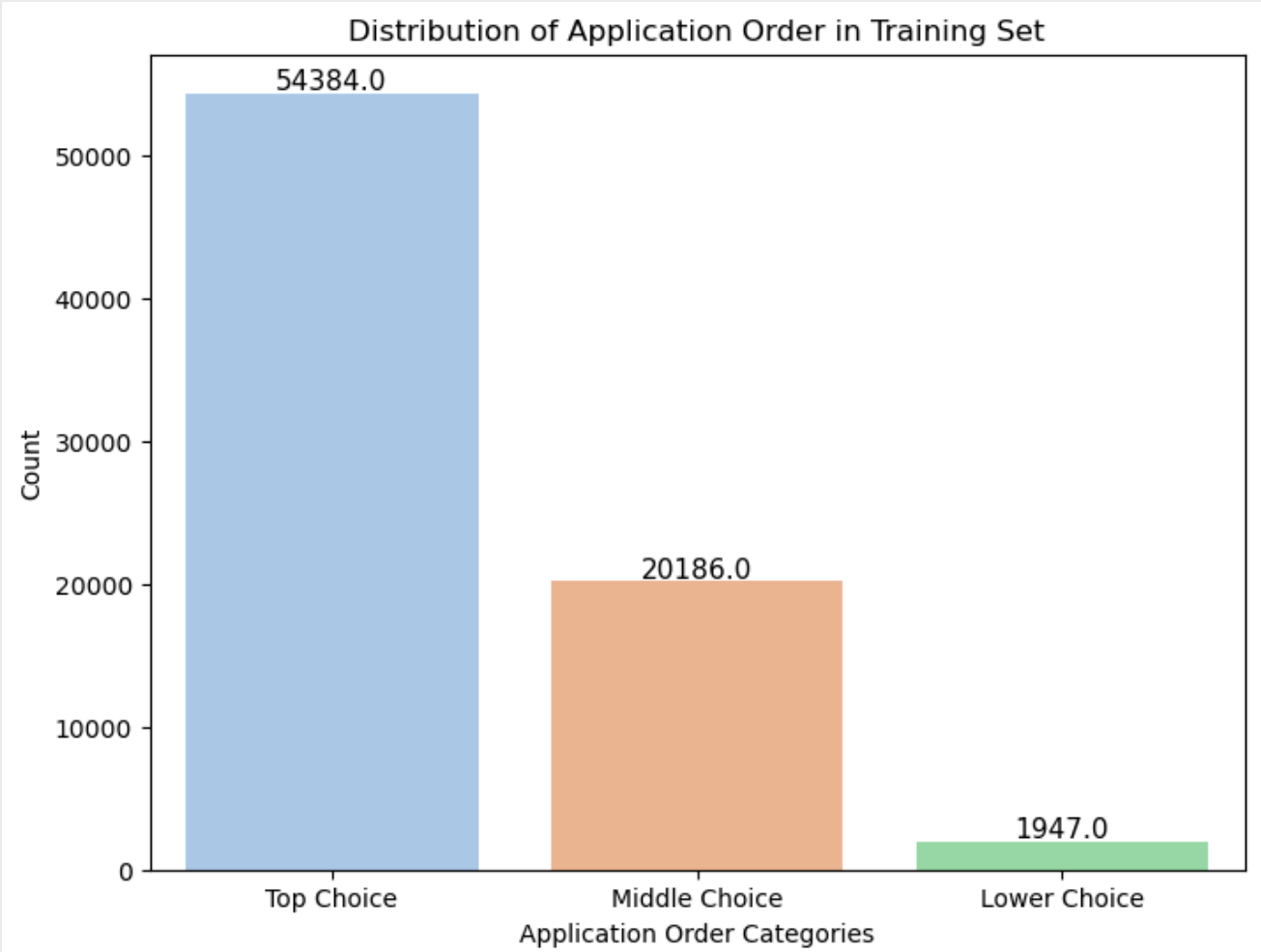
Application Order



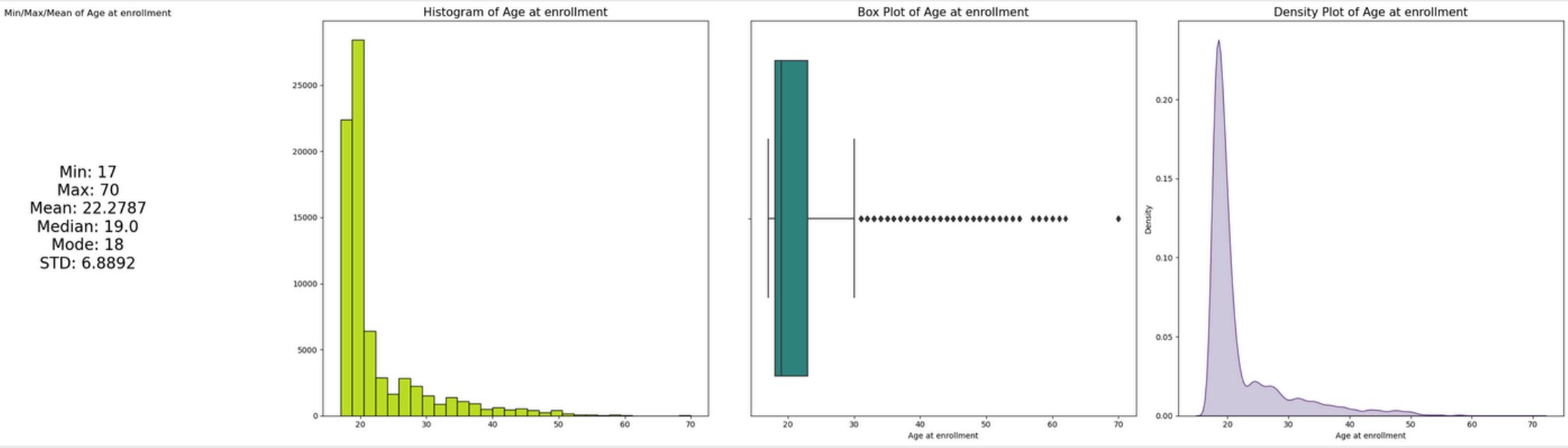
This feature ranging from 0 (first choice) to 9 (last choice), offered useful insights into the preferences or priorities associated with each application. Given the distribution and characteristics of application order shown above, the majority of our applicants are in top three choices in our dataset. Only 3 people were first choice and only one person was the last choice.

Question was what can I do with this information, so that we can improve the predictive power of my model?

I decided to transform the application order into categorical bins. For example, grouping orders into 'Top Choice' (0-1), 'Middle Choice' (2-5), and 'Lower Choice' (6-9). This can simplify the model's understanding of preference tiers.



Age at Enrollment

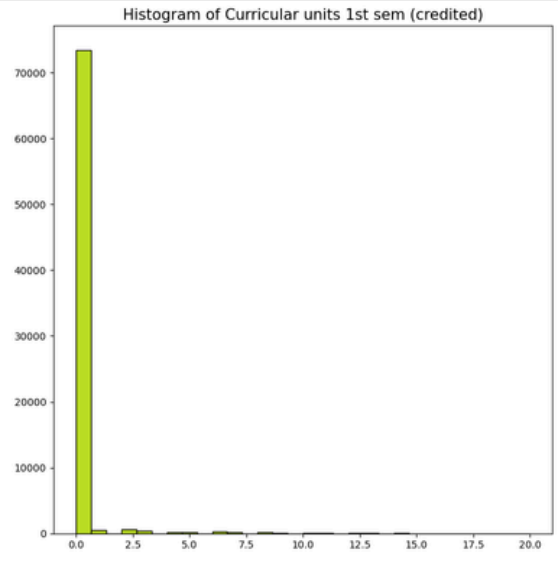


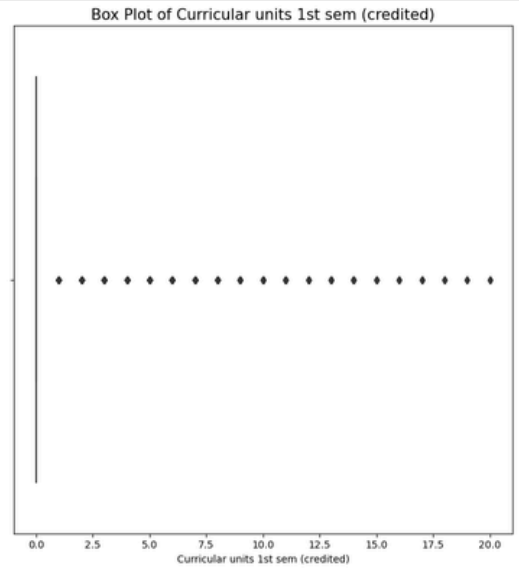
To effectively analyze and utilize the age at enrollment data for predicting outcomes such as graduation, dropout, or continued enrollment, we consider the typical duration of a bachelor's program, which is generally 3 to 4 years. Therefore, I have structured the age data into bins of four-year ranges, starting from the minimum age of 18, to facilitate a more targeted and relevant analysis of our dataset.

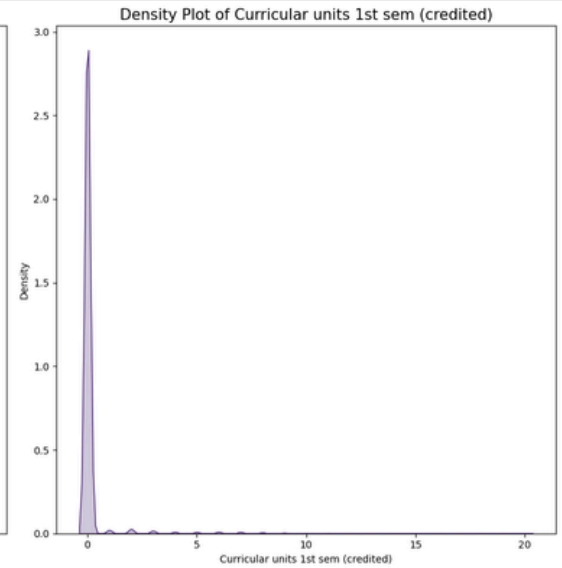
curricular units (credited)

Min/Max/Mean of Curricular units 1st sem (credited)

Min: 0
Max: 20
Mean: 0.1889
Median: 0.0
Mode: 0
STD: 1.1753

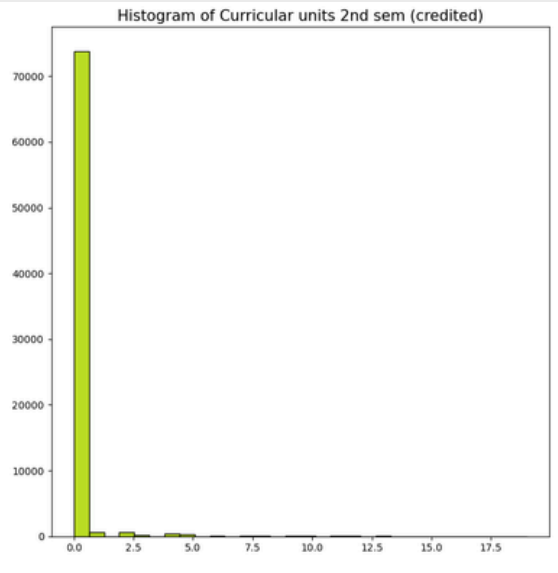


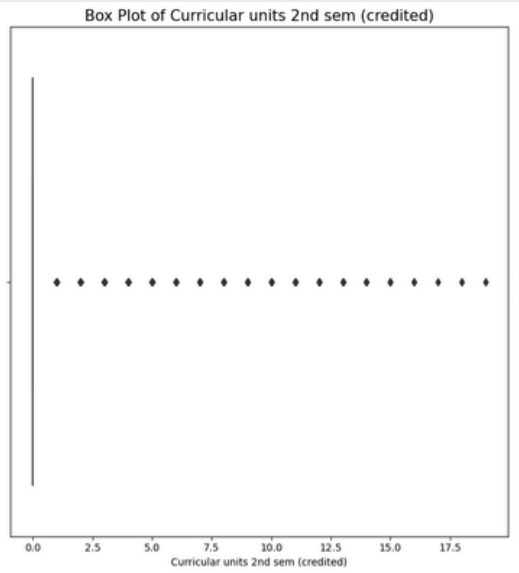


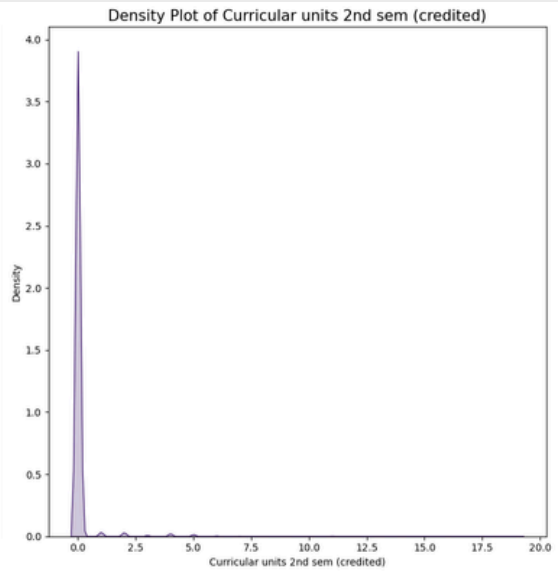


Min/Max/Mean of Curricular units 2nd sem (credited)

Min: 0
Max: 19
Mean: 0.1371
Median: 0.0
Mode: 0
STD: 0.9338



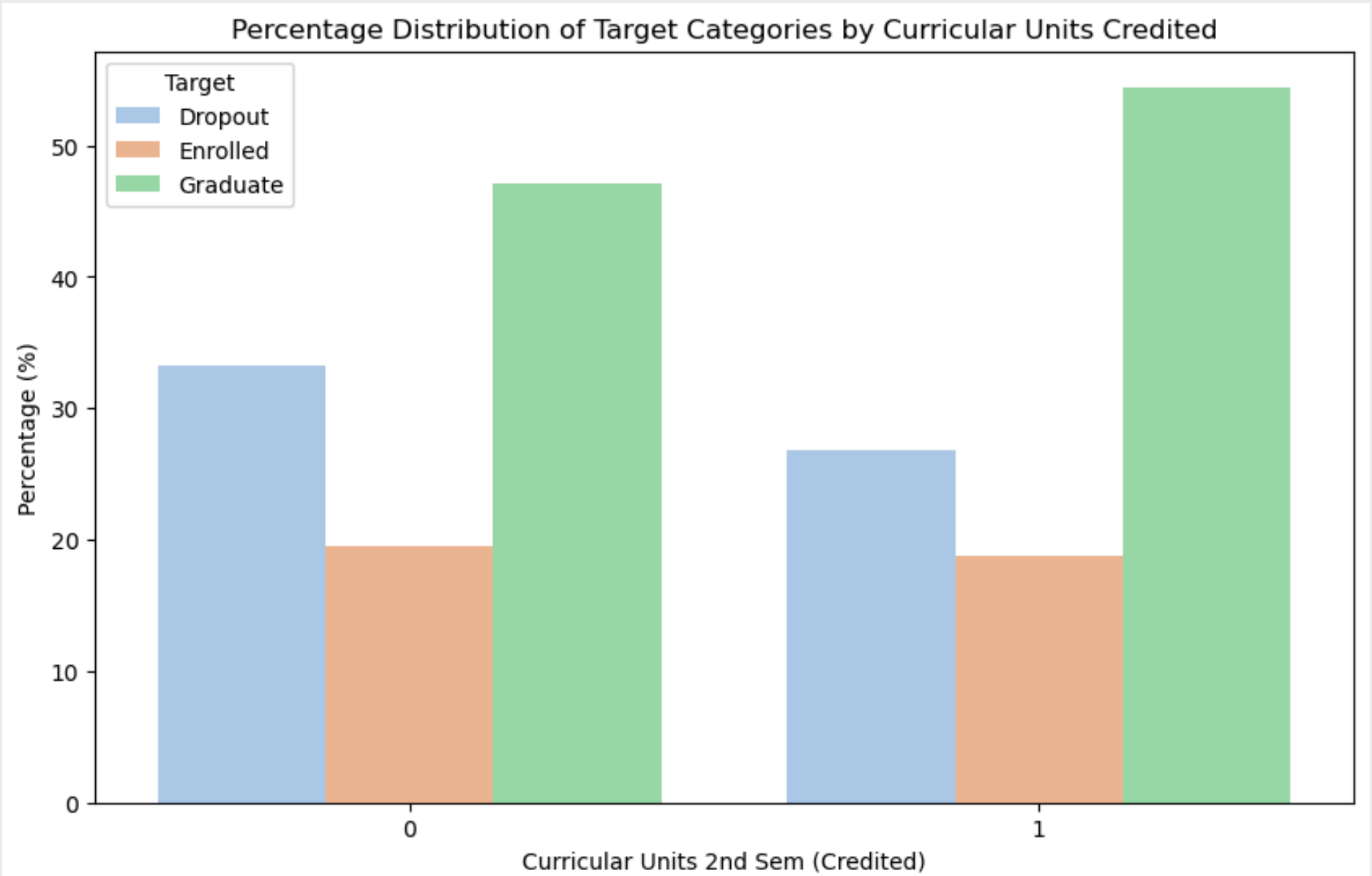




The majority of data points (students) are at zero for these features, So I created a binary feature indicating whether a student has been credited for these semester.

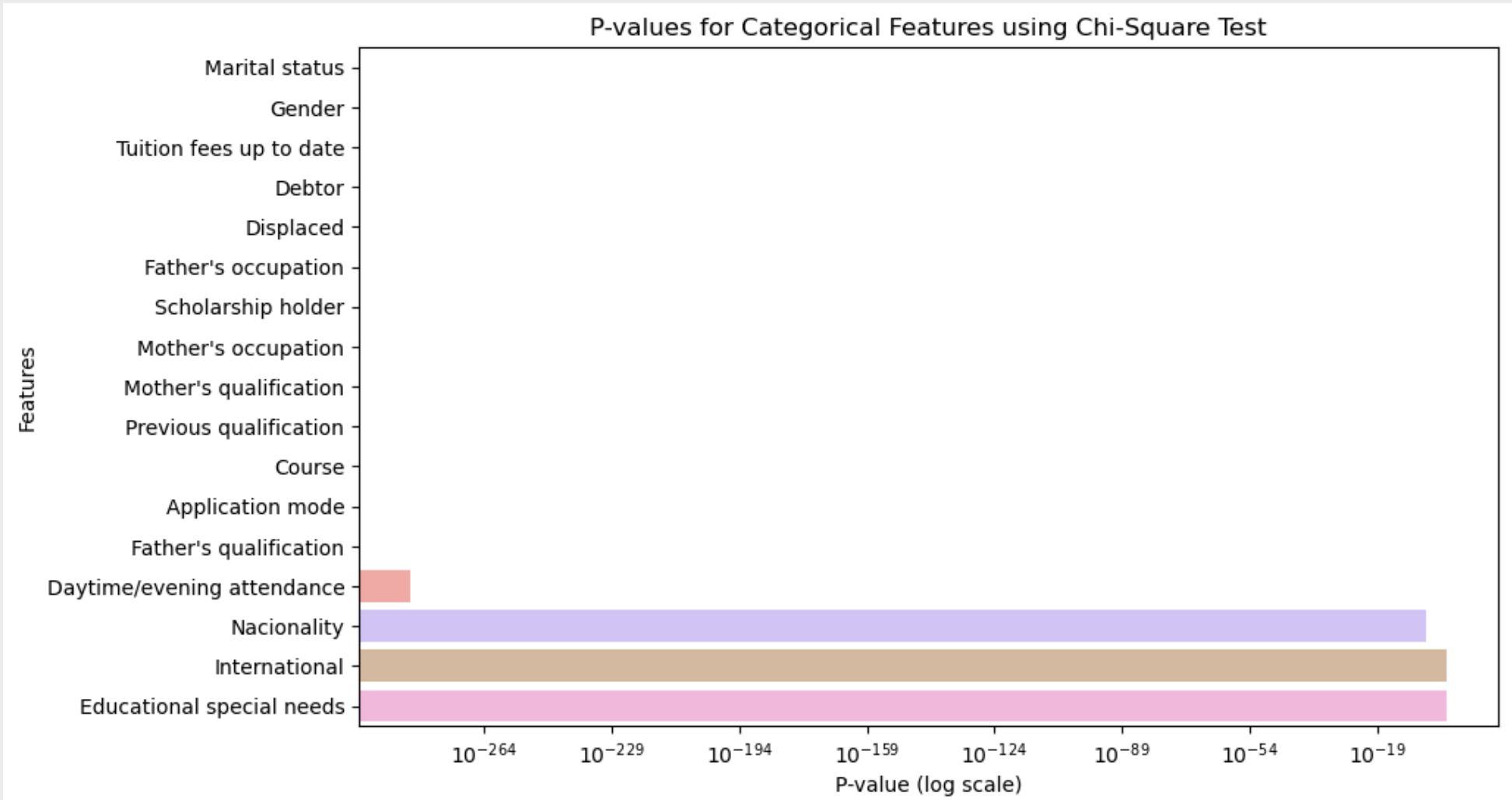
```
Credited_units_1st_sem
0    73428
1     3089
Name: count, dtype: int64
```

```
Curricular units 2nd sem (credited)
0    73808
1     2709
Name: count, dtype: int64
```


Percentage Distribution of Target Categories by Curricular Units Credited
Target
Dropout
Enrolled
Graduate
Percentage (%)
0
10
20
30
40
50
0
1
Curricular Units 2nd Sem (Credited)

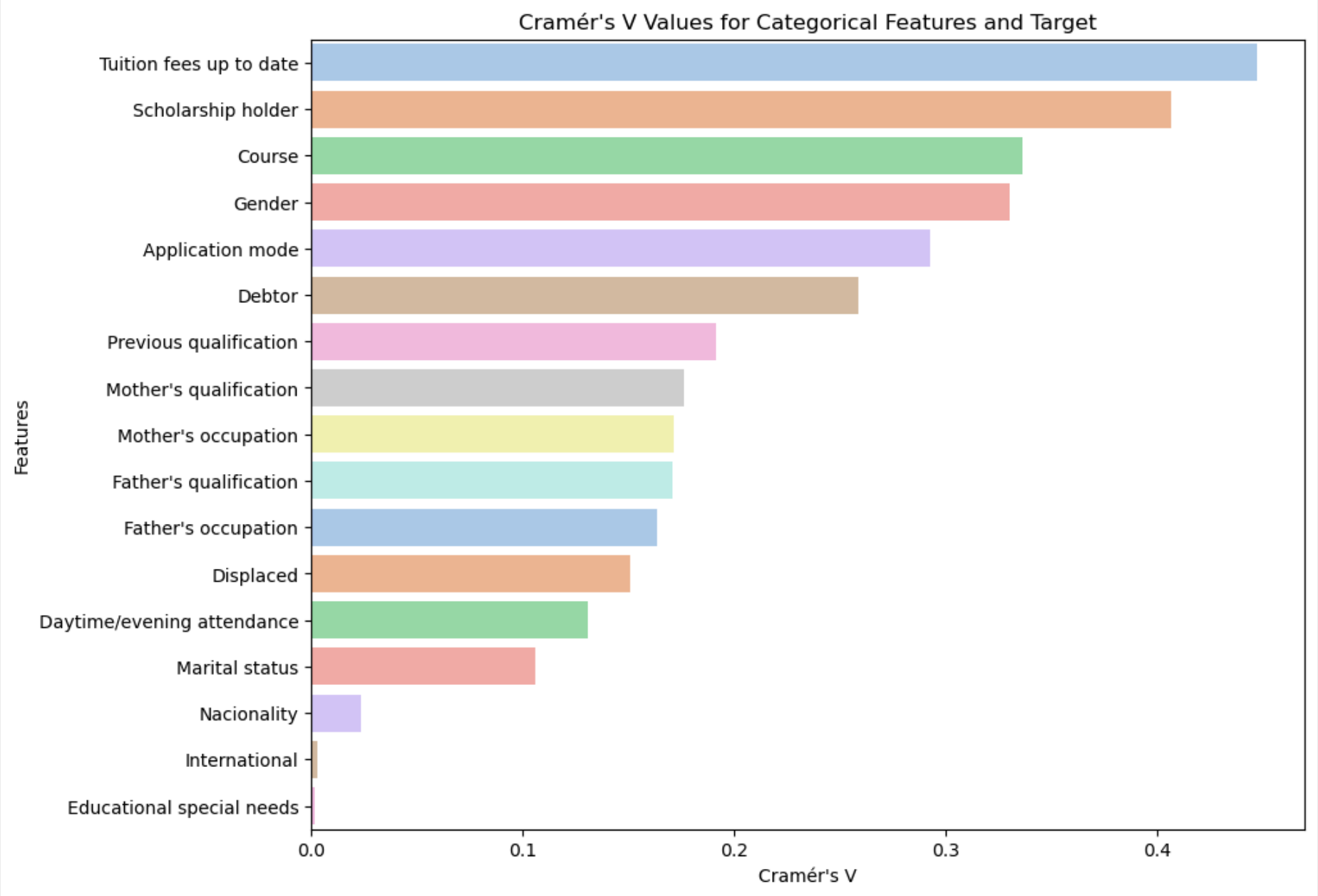
Relationship between categorical variables and target values

I implemented a Chi-Square test of independence to check the association between each categorical feature and the target.
Null hypothesis: There is no association between the two variables.
Features with Low p-values = statistically significant association

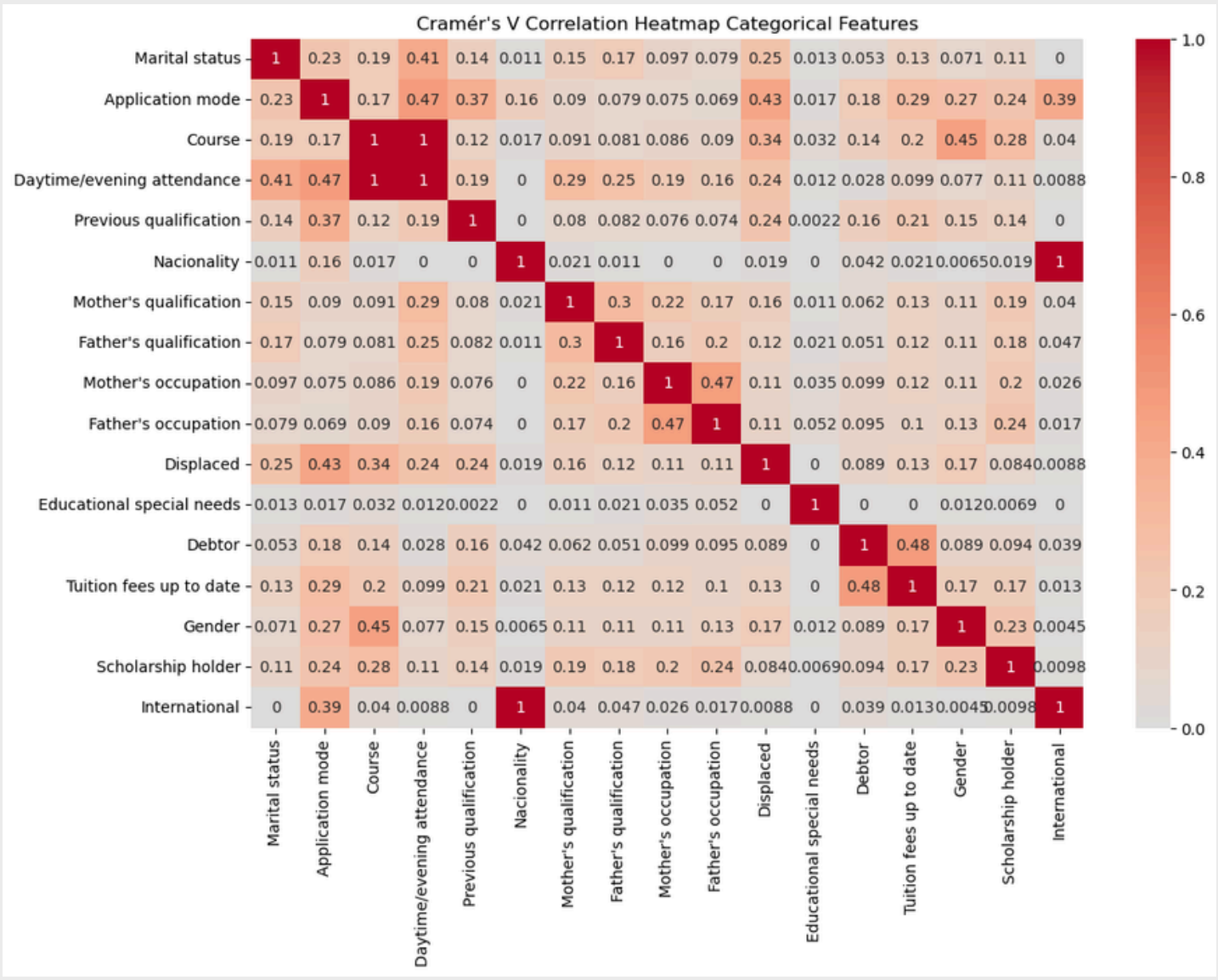


Cramér's V
Between Features and Target Value

It is a measurement of the association between two categorical variables



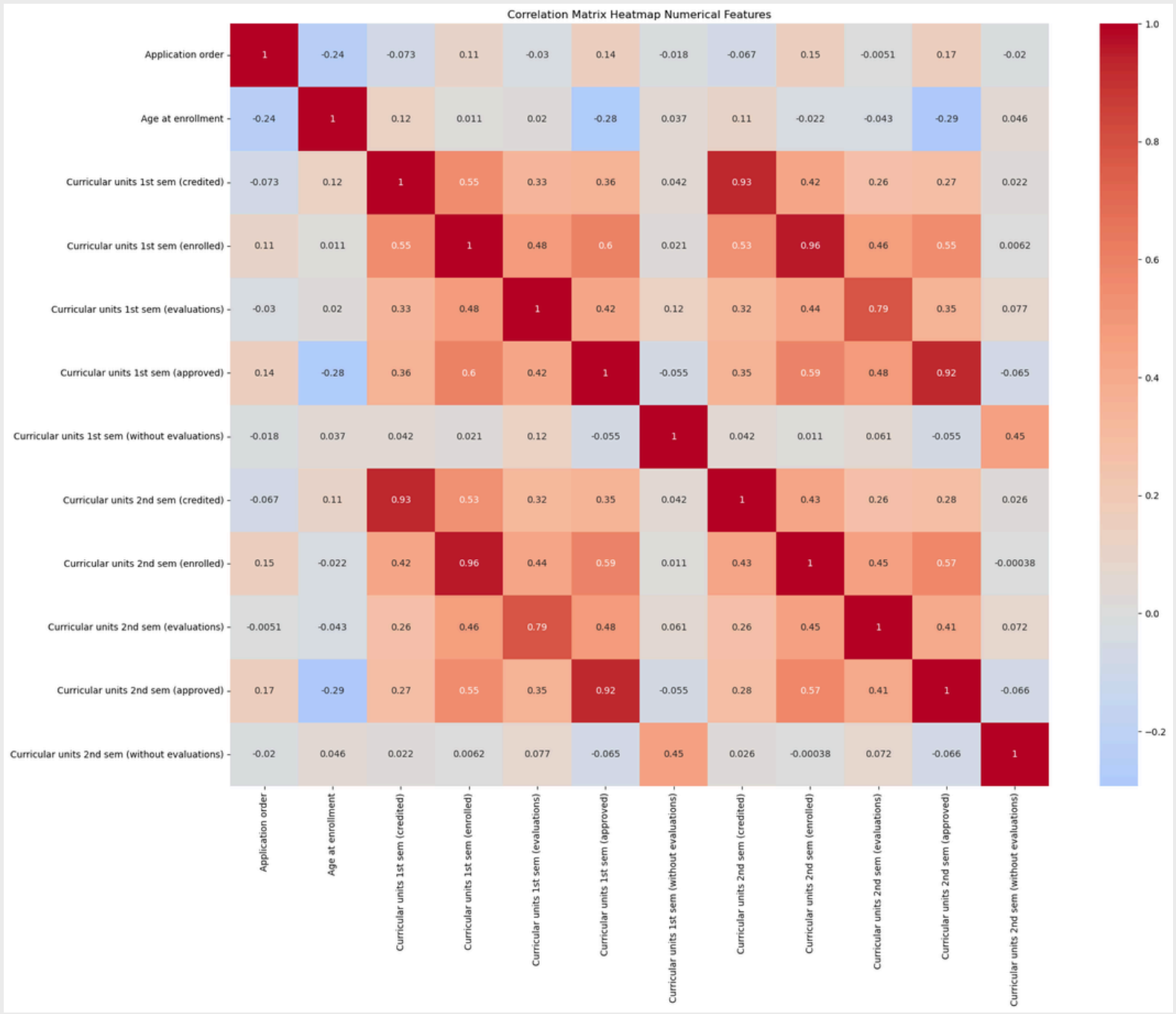
Cramér's V statistic between each 2 categorical features



Features highly correlated:

Daytime/evening attendance & course Nationality & International

Heatmap of The Correlation Matrix



Features highly correlated:

curricular unites 1st sem (credited) & curricular unites 2st sem (credited)

curricular unites 1st sem (enrolled) & curricular unites 2st sem (enrolled)

curricular unites 1st sem (approved) & curricular unites 2st sem (approved)

Modeling

I selected some of the most powerful machine learning algorithms available today—LightGBM, XGBoost, CatBoost. These models are known for their high accuracy and efficiency in handling large datasets with complex feature interactions.

CatBoost outperformed in the evaluation part.

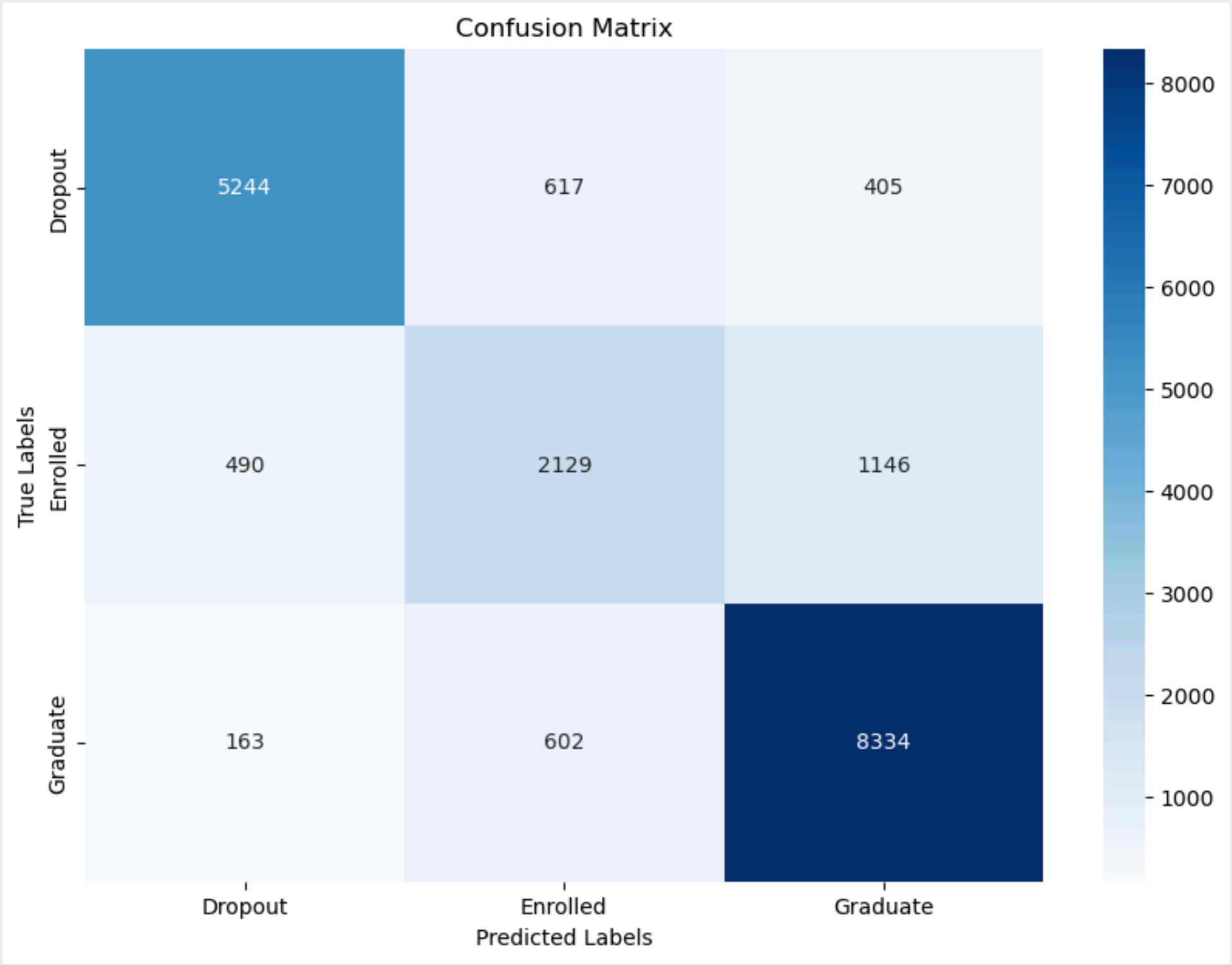
Model Evaluation

False Negatives:

Particularly concerning in educational settings, as failing to identify students at risk of dropping out (FN for 'Dropped Out') might mean missing the opportunity to intervene and provide necessary support.

False Positives:

Also critical, as incorrectly identifying a student as at risk of dropping out (FP for 'Dropped Out') could lead to unnecessary interventions, misallocation of resources, or put stress on the student.


Confusion Matrix
Dropout
Enrolled
Graduate
True Labels
5244
617
405
490
2129
1146
163
602
8334
Dropout
Enrolled
Graduate
Predicted Labels
8000
7000
6000
5000
4000
3000
2000
1000

1

The model performs best at identifying graduates, with a high number of true positives (8334) and comparatively lower false negatives. However, there is a noticeable number of graduates being incorrectly classified as enrolled. This might indicate a trend or pattern that causes the model to hesitate between these two classes.

2

There are significant false negatives (653 total), indicating that the model often fails to identify dropouts correctly, misclassifying many as either enrolled or graduated. This is a critical area for improvement, as failing to identify at-risk students could lead to missed interventions.

3

The enrolled class has a moderately high number of false negatives (1763 total), suggesting the model struggles to correctly identify students who are still enrolled. This might be improved by revisiting feature engineering or model parameters that specifically affect the differentiation between ongoing and completed academic statuses.

Incorporate more diverse datasets

Implement machine learning models that adapt over time to changes in educational strategies, student behavior, and external factors like economic shifts or technological advancements.