# Academic Success Prediction Project Report

**Introduction**

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).

Knowing the success rate has multiple benefits, to name a few:

* 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.
* 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.
* 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.

**“Wouldn't it be great if we could predict whether someone will drop out or complete their higher education based on their academic and demographic data?”**

**Objective**

The objective of this project is to develop a predictive model using over 76,500 synthetic data points. The model aims 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 created from a higher education institution (acquired from several disjoint databases) is related to students enrolled in different undergraduate degrees, such as agronomy, design, education, nursing, journalism, management, social service, and technologies. 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. The data is used to build classification models to predict students' dropout and academic success. The problem is formulated as a three-category classification task, in which there is a strong imbalance towards one of the classes.

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

**Method**

Data Wrangling & Cleaning:

The data is loaded from CSV files, with separate sets for training and testing. Initial data cleaning involves removing unnecessary identifiers and correcting data types for categorical features.

18 columns of our raw data, inherently categorical, were represented as integers in the dataset because they corresponded to coded numbers. I investigated to identify and correctly convert these to categorical data types.

Statistical Analysis and EDA:

I employed both visual and statistical tools to understand the data distributions and relationships between features. This includes:

* Pie charts to visualize the distribution of the target classes.
* Histograms, box plots, and density plots for continuous variables.
* Chi-square tests explore associations between categorical features and the target variable, determining the significance of these features in predicting the outcome.

A diagram of a distribution of a class

Description automatically generated

A blue rectangular object with black text

Description automatically generatedA green graph on a white background

Description automatically generatedA green and black graph

Description automatically generatedA close-up of a graph

Description automatically generated

In this stage, I ran a Chi-square test to check the association between each categorical feature and the target value. With the Null hypothesis being defined as “There is no association between the two variables”, the features with less than 0.05 p-value have a statistically significant association with the target values. This means that these features likely provide some information about the target categories. They are listed below:

Marital Status,

Application Mode,

Course,

Daytime/Evening Attendance (though this is not zero, it's a very small number, indicating significance),

Previous Qualification,

Nationality,

Mother's Qualification,

Father's Qualification,

Mother's Occupation,

Father's Occupation,

Displaced,

Debtor,

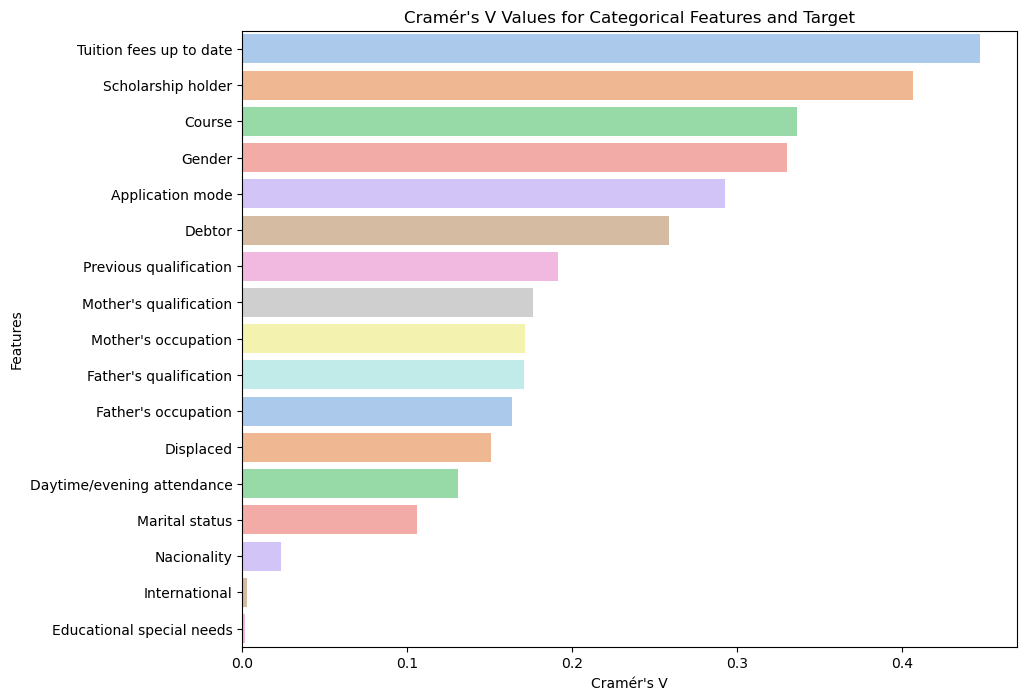
Tuition Fees Up to Date, Gender, Scholarship Holder.

Then I calculated the Cramer’s V, which measures the strength of association between two categorical variables. Cramér's V provides a value between 0 and 1, where 0 indicates no association and 1 indicates a perfect association. The result was fascinating. Below is a list ranking each feature's predictive strength concerning the target value, from the strongest to the weakest.

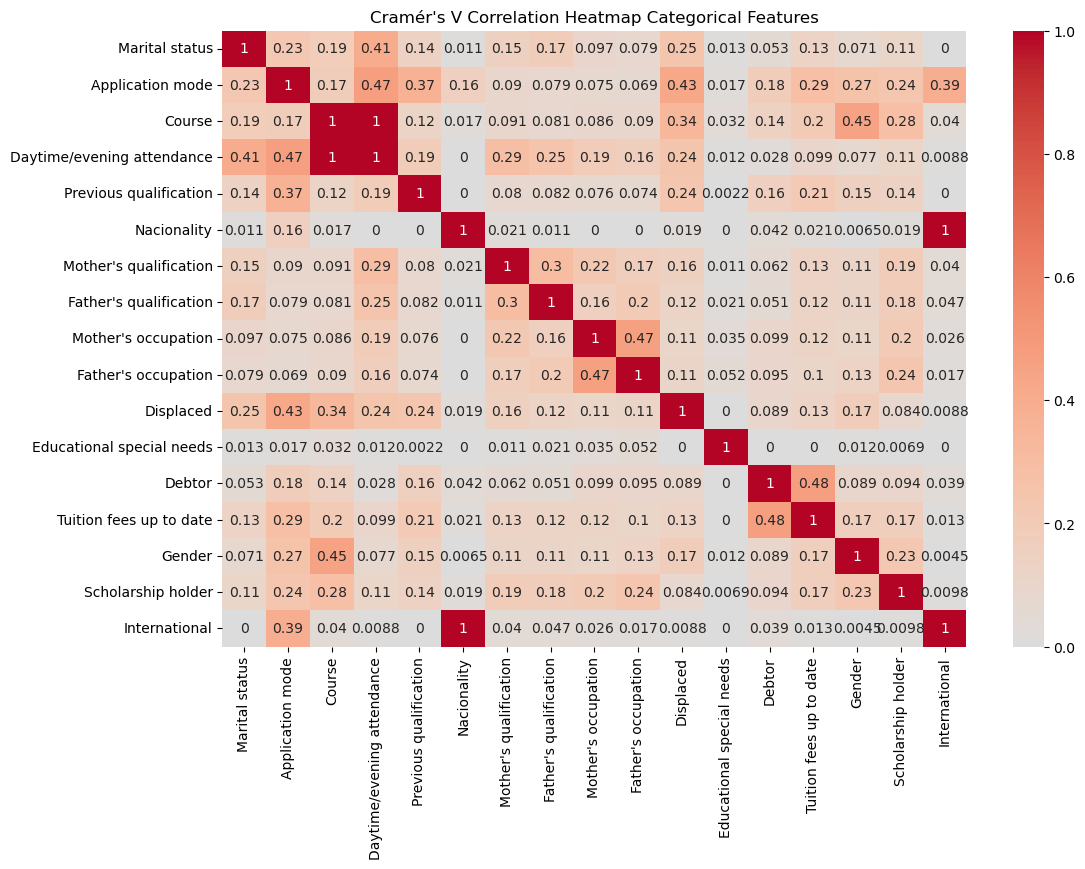
Tuition Fees Up to Date: 0.4472 - The strongest predictor, showing significant influence on the target.

Scholarship Holder: 0.4064 - Also demonstrates a strong association, indicating substantial relevance to the target.

Course: 0.3364 - Another potent indicator, reflecting its importance in predictions.

Gender: 0.3302 - This shows a meaningful association and is useful in forecasting outcomes.  
 

The same tool is used to investigate the correlation between categorical features together and the result has been shown in a heatmap below:



Features highly correlated:

Daytime/evening attendance & course Nationality & International. So, we can use either of those in our modeling.

**Preprocessing and Feature Engineering**

Categorical Data Handling: Various categorical features are transformed using techniques like binning and the creation of dummy variables to better capture the informational value in a format suitable for modeling. To name a few:

This feature ranging from 0 (first choice) to 9 (last choice), offers useful insights into the preferences or priorities associated with each application. Given the distribution and characteristics of application order, we can see that most of our applicants are in the top three choices in our dataset. Only 3 people were the first choice and only one person was the last choice. The question is what can we do with this so that we can improve the predictive power of our model?

Solution:

Binning: 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. We can also get rid of application order 9 since there is not much data on that, it can be considered an outlier.

Creation of New Features: New features are engineered from existing data to highlight differences and trends over time, such as the difference in academic performance across semesters.

**Modeling**

Model Selection: A CatBoostClassifier model was chosen due to its robust handling of categorical features and its efficiency with large datasets.

Parameter Optimization: Hyperparameters are optimized using Optuna to enhance model performance.

Cross-Validation: Stratified K-Fold cross-validation is used to ensure the model's generalizability across different subsets of the dataset.

Model Training: The final model is trained on the entire training dataset using the best parameters identified during optimization.

Making Predictions: The model is used to predict the academic status of students in the test dataset.



**Results**

Model Performance: The model's performance is assessed based on the cross-validated scores and its ability to generalize on unseen data. Key performance indicators include accuracy, precision, recall, and F1-score.

Feature Importance: An analysis of feature importance provides insights into which variables are most influential in predicting a student's academic status, aiding in understanding the underlying patterns recognized by the model.

**Future Improvements**

* Feature Selection and Engineering: Further analysis could refine the feature engineering process to extract more nuanced information and potentially remove redundant or irrelevant features.
* More data can be added to enrich and improve our model predictions, specifically regarding enrolled students.
* Algorithm Experimentation: Additional modeling techniques, such as ensemble methods or deep learning, could be explored to compare performance and robustness.
* Data Augmentation: Techniques to synthetically expand the training dataset might be considered to improve the model's ability to learn from a more diverse set of examples.
* Features with lower Eta Squared values may require additional context or could be combined with other data to improve their predictive power.

**Enhancements on the "Application Order" Feature**

The "Application Order" feature, which ranges from 0 (first choice) to 9 (last choice), reveals significant insights into the preferences or priorities associated with each application. Our analysis indicates that the majority of applicants fall within the top three choices. Notably, only three individuals selected the first choice, and a single person opted for the last choice.

**Future Enhancements Include:**

**Rank Transformation**: Another approach involves converting the application order into a rank percentage, normalizing it from 0 to 1. This transformation aids in stabilizing learning across different scales and can be particularly effective if integrated with other numerical features, enhancing the model's ability to discern nuanced differences in application priorities.

These enhancements are designed to improve the predictive power of our model by refining how we interpret and utilize the "Application Order" feature. In subsequent phases of this project, we will evaluate the impact of these strategies and determine which method optimally enhances model accuracy.