Projecting the Graduates – and Dropouts – for Early Intervention

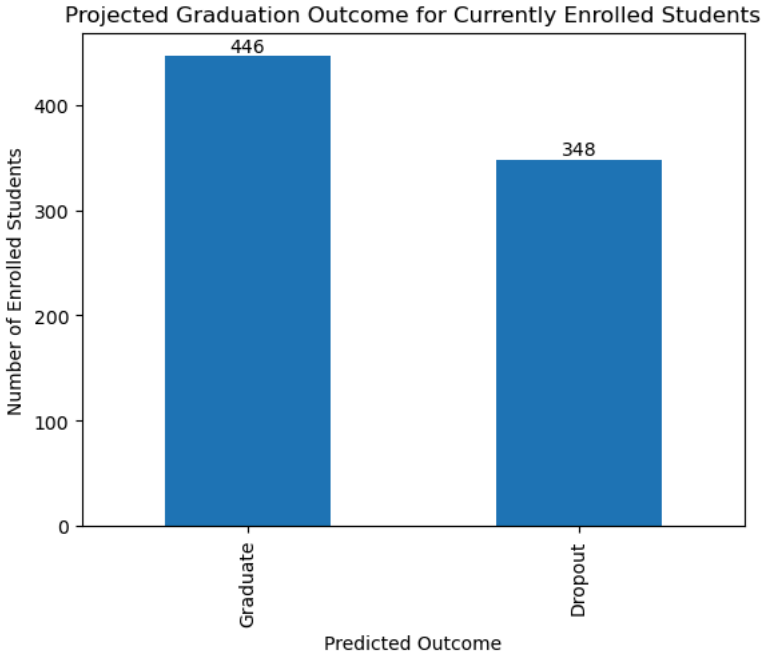
Problem Statement

What are the primary indicators of student drop out in colleges and universities so that incoming scholars can be screened and dropout rate can be reduced over the next four years? Given that we have a lot of data about our past scholars, we can use this data to model and predict the future success of scholars and intentionally intervene with potential dropouts before they leave our institution.

Overall Findings

After studying the data of our past graduates and dropouts and testing five classification models, the best model is a Logistic Regression predicting 56.2% of our currently enrolled scholars to graduate (shown in the bar chart below and right). In studying the factors that lead to dropouts through these models, there are several actions to take to course correct and work with the other 43.8% to move scholars out the dropout track and towards graduation. The suggestions are:

1. additional support for scholars who are not up-to-date on tuition and don't have a scholarship as these two financial elements were in the top five most important factors in drop out rate.
2. increased academic mentorship and support for freshman students as successful completion of the first two semesters of classes were the two most important factors in determining graduation versus dropout.
3. a further analysis in the rates of completion for the various course paths at the school. There is a massive gap in graduation rate between different programs of study that needs to be fully explained.



Process

Exploratory Data Analysis

The data in the set was wide-ranging and incorporated many features and details of each scholar's life in and out of college. There were several application methods and qualifications that were represented by less than 1% of the scholars in the data set. To account for this, qualification was simplified to No High School, Some High School, High School and College Degree to better group the scholars based on their previous qualifications. The same was applied to the academic qualifications of their parents. 88% of scholars reported being single, so this data was binarized to be either "single" or "other." The application modes were simplified to "1st phase," "2nd phase" and "other."

After this simplification, simple bar charts were analyzed to determine if any features needed to be explored in more detail. This is where the Nursing program stood out, along with the financial obligations of the scholars (scholarship, debtor and tuition up-to-date all showed a gap between graduating and not). Gender and age also showed gaps in graduation versus dropping out. The curricular units earned also showed a significant gap, which would play out significantly later on in the process.

The correlations between variables were also useful in determining some key features early on. The heat map confirmed the connection between the variables of interest and dropout rates as shown in our bar charts. This also showed that family occupation and family qualification were highly related, so family qualification could be dropped from the data set to avoid doubling up on that information.

Pre-Processing

The pre-processing for this data set required work to ensure that our many categorical features in the data were binarized. This included displaced status, special needs, debtor, tuition fees up-to-date, gender, scholarship holder and international students. These elements could all be applied to our regression once binarized. The data on marital status, application mode, application order, course of study, family qualification and previous educational experience were binarized using dummy variables, meaning each scholar was 0 or 1 for each of the possible outcomes for those categories. Finally, the curricular units, unemployment rate, inflation rate and GDP were normalized to a 0 to 1 scale so that all variables would be weighted equally for use in the regression modeling.

With all of this data binarized and scaled, the scholars where we knew the dropout or graduate result were put into a 75/25 train/test split to model. The still-enrolled scholars were separated into their own set to use the model on for projections. These projections will be used to determine where the most support is needed for our currently enrolled scholars.

Modeling

Five classification models were built using the training data and then tested. By using these five methods, the data was passed through the majority of the most common ways to classify data and determine efficacy in identifying labels.

The first model was a Logistic Regression model. This model was fitted with the base parameters for logistic regression of C=1.0 and a penalty of “l2” on the training data before being tested on our test set.

The second model was a Random Forest model. The hyperparameters for the Random Forest model were analyzed and determined to be 1400 n_estimators, a min_sample_split of 5, a min_sample_leaf of 1, max_features of square root and a max_depth of 80. Using these parameters, a model was formed on the training set and tested on the test set.

The third model was a K-Nearest Neighbors model. The hyperparameter tuning showed that 8 was the optimal number of neighbors with this model. A model was built with 8 neighbors on the training set and tested on the test set.

The fourth model was a Decision Tree with dimension reduction through PCA. The elbow method was used to determine that 4 principal components were the best number to use. This model was also generated using the training set and then tested with the test set.

The fifth model was an SVM model. The SVM model was trained and tested in the same way as the others.

All five models were tested with 5-fold cross validation to ensure that the models were replicable and that the particular versions modeled were not outliers.

Final Model Comparisons & Choosing Logistic Regression

In analyzing the data on the five models. It was clear that two stood out from the rest, with Logistic Regression and SVM being the strongest in terms of dropout precision and recall. Random Forest classification identified graduates well, but did not perform at the main task of identifying dropouts at the level of the two top models. Logistic Regression performed slightly better in identifying the dropouts from our test list as shown by the 0.2 advantage in recall versus SVM, which was the ultimate decision maker in choosing the Logistic Regression model over the SVM model, though both performed very similarly in the process (Confusion Matrices for the two are in Appendix A).

Figure: Performance Statistics on the Models

Model	Dropout Precision	Dropout Recall	Dropout F1	Overall Accuracy
Logistic Regression	0.92	0.84	0.88	91%
Random Forest	0.92	0.81	0.86	90%
KNN	0.87	0.64	0.74	83%
PCA w/ Decision Tree	0.72	0.74	0.73	80%
SVM	0.94	0.82	0.88	91%

Applying the Model & Features of Importance

In studying the importance for the individual features in our regression model. The full list of feature importance is in Appendix B. There were five that stood out as the most predictive of future success or lack thereof.

- Curricular Units 2nd Semester (approved)
- Curricular Units 1st Semester (approved)
- Tuition Fees Up to Date
- Curricular Units 2nd Semester (grade)
- Scholarship Holder

The approved curricular units were far and away the most important data points. This is somewhat intuitive, but passing classes in the first year of college sets scholars up for future success. This group of scholars was not falling behind on degree requirements and were seeing success throughout their careers. On the converse, the

enrolled curricular units were at the bottom of the list of features. Just enrolling was not enough, passing those first year classes is what made a tremendous difference. The grade for the 2nd semester also has an impact on future success. This indicates that passing the 1st semester is vital, but passing with better grades was important in the 2nd semester.

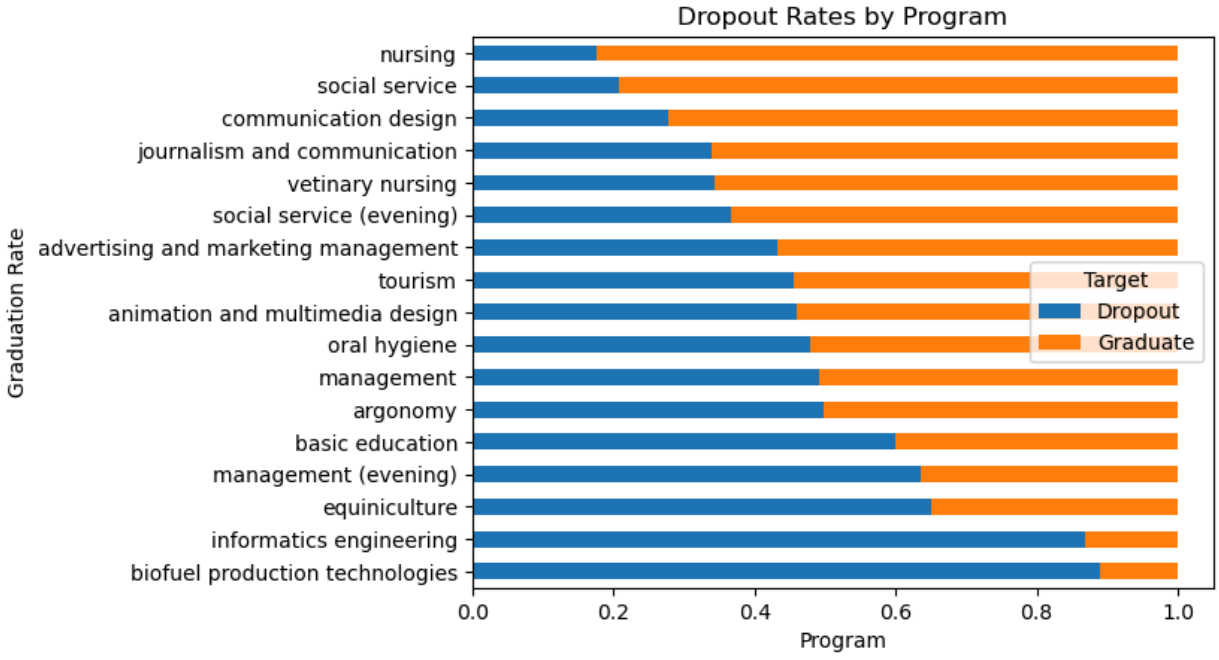
The other two main features identified in the model were about the financial situation of the scholar. If tuition fees are up to date and a scholar is a scholarship holder, that scholar has a much better chance of successfully graduating from our school. These financial elements play a clear role in the ability for the scholar to have success. This could be as simple as the scholar doesn't also have to go to work to pay for college while in college or come from more complex issues of not being able to pay for college and being forced to drop out solely for financial reasons. The global variables (like Inflation Rate and GDP) played less of a role than the individual financial situation of each scholar. Regardless of the reason, scholarships and tuition matter significantly to future success.

Once the modeling was applied, the data for all of the currently enrolled scholars can be found in this spreadsheet. The information here will allow our academic advisors to identify scholars projected to drop out and add additional support. With those additional supports in place, we can analyze if an increase in targeted support for the most at-risk scholars will lead to an increase in graduation rate, which is projected at 56.2% for our currently enrolled scholars based on the model.

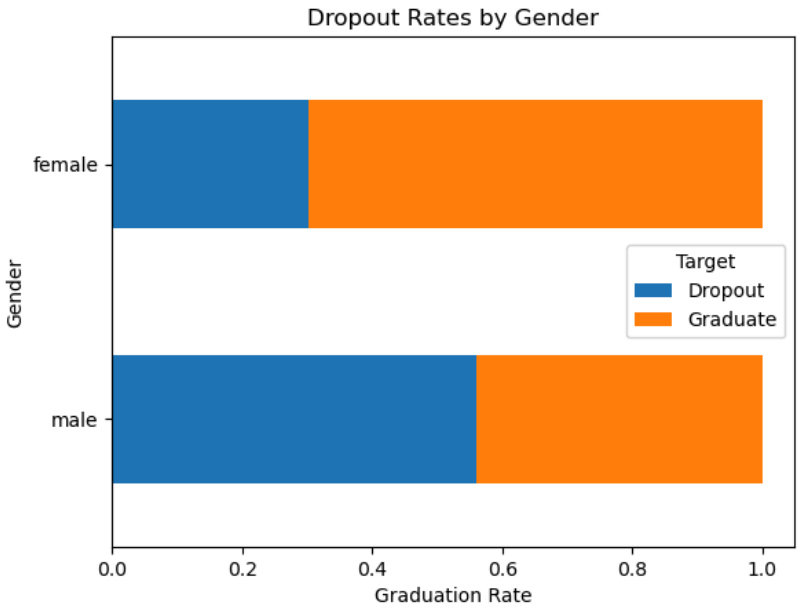
Features to Pursue Further

There were three features of the data which were included in the model, but warrant further exploration individually. This was the program and of study, gender and debt status for our students.

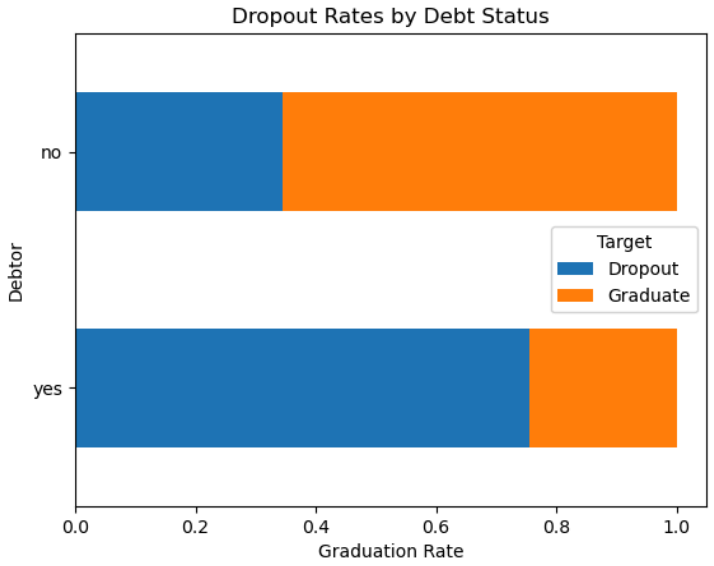
When broken down by program of study, as shown below, individual programs have a significant difference in graduation rate, with the nursing program graduating over 80% of its students and informatics and biofuels production with graduation rates below 20%. Each program needs to be examined to determine the reasons behind the gaps shown program to program across our campus.



There is also a significant gap between male and female scholars in terms of drop out rate. Over half of all male scholars who enter our school dropout, but two thirds of female scholars go on to graduate. Given this disparity, there needs to be further study into the reasons behind the male dropout rate. These scholars were able enter our college, but are underrepresented in terms of actually graduating. There could be many reasons for this, but more data needs to be collected to better analyze this gap.



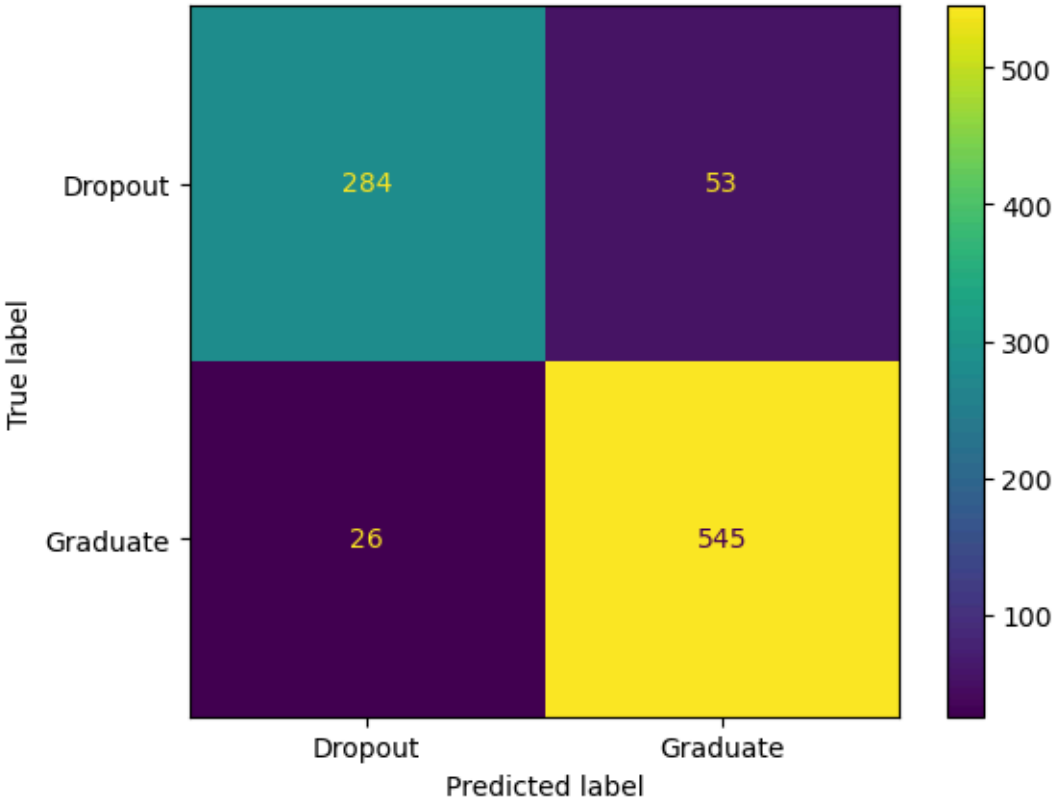
Finally, the scholars who are in debt at our school are only graduating at about a rate of 1 in 4. Almost two-thirds of our scholars without debt at our school are graduating, so there is a strong link between the individual finances of a scholar and if they will graduate or not. More data needs to be collected, but there is something about the personal expense incurred by scholars that is making graduation difficult to attain for this group of scholars.



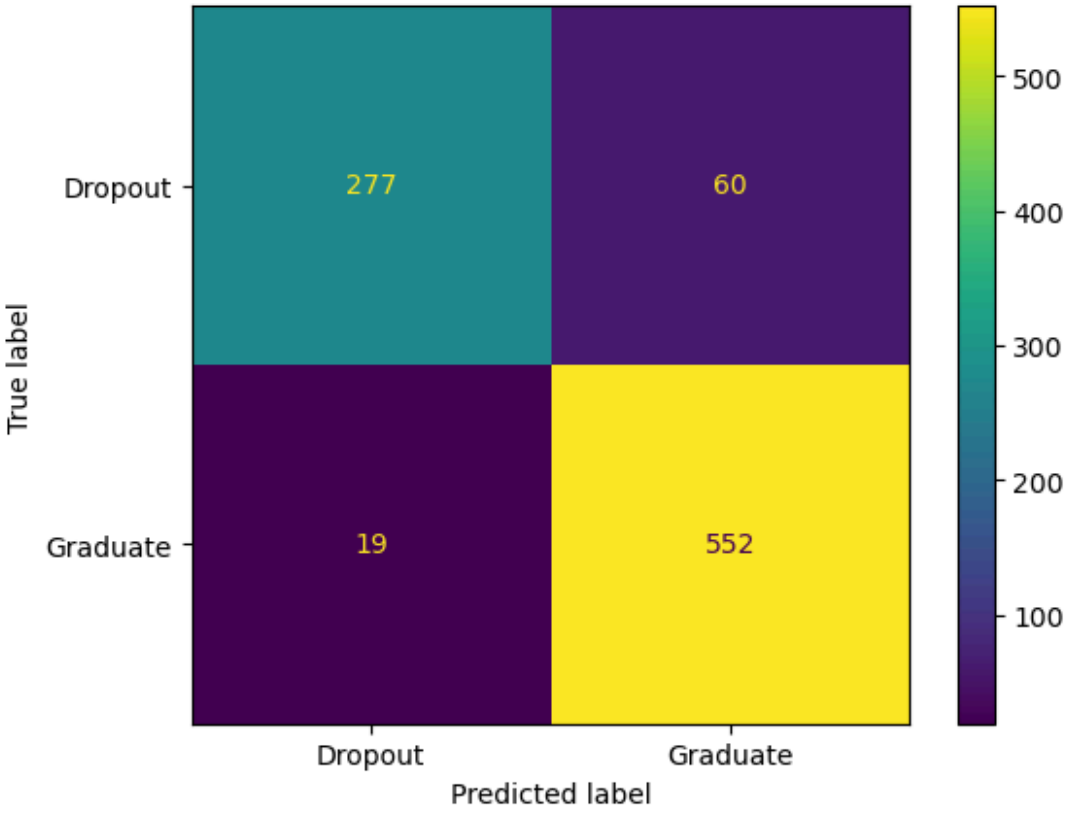
Conclusion

After reviewing, processing and modeling the data, it is clear that there are links between first-year success and individual finances on the eventual graduation from our college. By analyzing these two issues, we can identify scholars who need additional support and work to help them overcome these issues early in their career and go on to matriculate from our program. This data also shines light on disparities between program, gender and debt that must be further addressed. Going forward, we now have identified potential dropouts amongst are enrolled students and can start moving to change their outcome to the one that we want for all of our scholars.

Appendix A: Confusion Matrix for Logistic Regression Model shows 26 mislabeled Graduates and 53 mislabeled Dropouts.



Confusion Matrix for SVM Model shows 19 mislabeled Graduates and 60 mislabeled Dropouts.



Appendix B: Feature Importance in Logistic Regression