Predicting Successful Graduation from Performance in Mathematics Courses

A data science project using a dataset from Iowa State University to predict the likelihood a student will complete an undergraduate degree based on their success in mathematics courses.

The Dataset

The original dataset consisted of the mathematics courses taken by 13,065 students in the College of Liberal Arts and Sciences at Iowa State University during a 10-year timeframe, from Spring 2014 to Summer 2024. We know if a student got below a C or not, if/when they graduated, when they took the course, and when they enrolled at the university.

After cleaning the data set, we had 9,181 undergraduate students who took a total of 28 different courses, were enrolled between 2008 and before 2021, and had a graduation rate of 49%, where graduation was binarily indicated if the student matriculated within 9.5 semesters (the Summer Session was counted as 0.5 semesters). Performance in each course was also measured binarily, and the cumulative performance of each student across each of the 9.5 semesters was tallied.

Modeling Approach

As graduation is a binary target, we will use binary classification as our modeling approach. After an initial exploration with several different methods, we focused our efforts on four models:

- *Logistic Regression* – this will be our baseline model.
- *Support Vector Classifier* – this will capture any non-linearity that is missed by logistic regression.
- *XGBoost Classifer* – the use of gradient boosting will help this model learn wit hteh subtler aspects of the data that may have been missed by the previous two
- *Custom Stacked Classifer* – this is a customized ensemble classifier which uses prediction probabilities from our previously fitted SVC and XGB classifiers as features for a logistic regression model.

Results

Each model was trained and cross-validated on 5 splits of our dataset. Here is a summary of the mean accuracies across all 5 splits:

Model	*CV Accuracy*	*Test Accuracy*	*Accuracy Change*
Logistic	63.08	62.44	-1.01
SVC	64.86	64.53	-0.5
XGBoost	66.5	65.7	-1.19
Stacked	69.47	65.78	-5.27

The accuracy scores are overall not great, but there does not appear to be an issue with overfitting or underfitting since the change in accuracy from train to test sets is relatively small.

As for feature importance, the logistic models seemed to prefer the cumulative performance of a student as opposed to performance in individual courses; specifically, the cumulative columns were the top 16 coefficients by absolute value. The XGBoost models were quite different; the most significant features were the individual courses with the top 3 being MATH 104 (Introduction to Probability), MATH 143 (Preparation for Calculus), and MATH 166 (Calculus II).

Concluding Analysis and Recommendations

Ultimately, there is an intrinsic limit to the predictive power of our dataset. The theoretical minimum error for any binary classifier is given by the Bayes Error Rate and ours is 25.3%, meaning that there is an upper bound of 74.7% on the accuracy of any model we can apply. This is primarily since there are only 2580 distinct observations in our data set and of those only 578 have feature values of size greater than 1. Regardless, our models only incurred 10% more error than the theoretical least possible error, so relative to this it is reasonable to conclude that our models performed reasonably well.

The bulk of the error is caused by the insufficient ability of the features (courses) to separate the classes. This limitation may only be overcome by a having more granular data for students, indicating that future studies will likely either need to be performed by employees of institution at which the students are enrolled or by using detailed data that has been rigorously anonymized to protect student privacy. Given access to such a dataset, this analysis can be applied with greater effectiveness and the ability to provide course-specific guidance.