Predicting Student Graduation at a Private College

Theo Jongerius, Dr. Schon
Mathematics and Physics, Northwestern College

Introduction

The purpose of this project was to accurately predict which students attending Northwestern will graduate based on information we collect on incoming students, as well as information we gain from students during their time at NWC. Through use of the programming language Python and its machine learning libraries, I was able to develop a machine learning model that can accurately predict if a student would graduate 85% of the time.

Method

A dataset with information on 1130 students was provided for testing predictive models on. Microsoft Excel was used to clean the data. The programming language R was used to perform statistical tests and analyses. Machine learning algorithms included in Python's scikit learn library allowed this project to successfully achieve its goal.

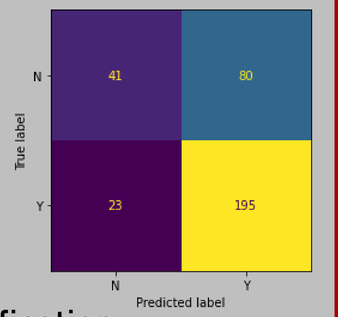
Three models were developed to predict students' graduation status. All three models were trained in order to accurately classify a student as 'graduated' or 'did not graduate'. The first model used a K-Nearest Neighbors algorithm. The second model used Logistic Classification. The third and final model used a Decision Tree algorithm for classification.

Results

Of the 1130 students provided, 750 (66.4%) graduated.

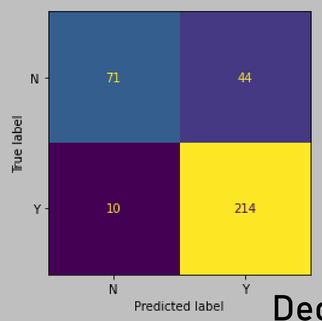
K-Nearest Neighbors

Using variables from students' 1st and 2nd years at NWC, the model was able to accurately predict if a student would graduate 84% of the time.



Logistic Classification

Using only 2 variables, FYS grade and Christian Story I grade, this model was able to accurately predict if a student would graduate 85% of the time.



Decision Tree

Allowing the Decision Tree model three levels of decisions, the model chose 3 variables: 1st semester GPA, 1st year GPA, and 1st year Earned hours. This was the best-performing instance of the model, with an accuracy score of 83%.

Conclusion

Of the 3 machine learning models developed during this project, the Logistic Classification model is the most promising, predicting if a student would graduate with 85% accuracy. This result was reached using only 2 variables, both of which are often available as soon as a student's first year. Variables from prior to enrollment at NWC were not useful, while information from students' early college careers was useful.

Although the initial goal of aiding the admissions department in determining which students are most likely to succeed at NWC was not reached, the models may still be useful to NWC's academic support team

Future Directions

One of the most powerful and popular machine learning algorithms commonly used today is the XGBoost algorithm, or Extreme Gradient Boosted Trees algorithm. XGBoost works in a similar way to the Decision Tree algorithm, albeit with more processing power. Applying that model to this project has the potential to yield over 90% accuracy, while decreasing the previous models' tendency toward false positives, which would increase NWC's effectiveness at determining students at-risk of not graduating and focusing academic resources to those students who have the best chances of success with additional aid.