**Matthew Fikes**

*DSC630*

*Milestone 5: Final Paper*

**EXECUTIVE SUMMARY**

This goal of this study was predicting college enrollment habits for students taking dual credit education at my institution. While students can transfer credits to many accredited institutions, it is our hope that they continue their education with us.

This task is split into two parts – first a classification of students by whether or not they would continue their schooling with my institution or not. The second task is an attempt to identify commonalities in these students via a clustering algorithm.

The results of the classification show an accuracy of ~66%, with the random forest and ~67% with the individual decision tree, putting them close in performance. The decision tree model is slightly more accurate with predicting students who will not continue (~91%) than who will (~23%).

The clustering so far has been inconclusive. Several methods have been tested so far with results that do not appear consistent. Additional research will be necessary to make meaningful clusters of students for targeted marketing or admissions contacts.

**TECHNICAL REPORT**

**Background**

For the past 10 years, the State University of New York (SUNY) system has faced enrollment declines (Giacomelli, 2020). In hopes of stemming the losses many schools are trying to reach new student populations. These efforts include new geographic regions for targeted marketing and expanding adult and corporate education. While these efforts meet mixed success, another group has been filling in the space in the enrollment gap. This group is students taking for-credit classes while still enrolled in high school. This is commonly referred to as dual enrollment and is a partnership between school districts and colleges to enroll and earn transferable credit (EdTrust, 2019).

Between 2002-03 and 2010-11 the number of dual credit students increased 68% (Field, 2021). Internal data shows an increase of nearly 250% over the last decade. While this has been good for enrollment numbers, it raises complicated issues involving how institutions are paid for these students. While we are paid by the state that same year for traditional credit enrollments, dual enrollments are paid a percentage amount over three years. This leads to added complexities when creating budgets, especially as we are more reliant on that population.

The best outcome with dual enrollment students is them entering the college system after graduating high school as matriculated students in one of our programs. Currently somewhere around 20% of students taking credits end up at the college. If we were able to predict which students were unlikely to matriculate, we might find ways of appealing to them.

**Problem Statements**

*What dual-credit students are unlikely to matriculate?*

Students who do matriculate are known quantities to us and there is more information available about them to help guide them successfully. If we can predict which students we provide credits for who don’t enroll with us we may be able to plan intervention methods.

*Do these non-matriculating students fall into groups with common features?*

Once we know the students who won’t return, we may be able to cluster them into common groups with features that could inform recruiting efforts.

**METHOD**

**Preliminary Requirements**

Before any data is processed all personally identifiable information (PII) needs to be removed or altered in accordance with the Family Educational Rights and Privacy Act (FERPA) (20 U.S.C. § 1232g; 34 CFR Part 99). ID number and high school ID have been modified to protect the privacy of the student prior to any analysis. For privacy reasons, details about the extraction of student records are not available for this report.

**Data Sources**

All the data was retrieved from an Oracle database containing the backend for our internal student data system. Data for dual credit students is reported by individual school districts and includes only basic demographic information about the student.

The data also includes the courses taken by dual enrollment students in their first semester, identified by an internal coding of “H” in their student term file.

Data collected for students in Fall and Spring semesters covers the period between Fall 2010 and Spring 2020. Summer and Winter semesters are not included because they have reduced course availability and no dual enrollment students.

**Feature Preparation/Creation**

Student age was created as a calculated feature based on the original date of birth on record compared to the year the student was first classified as dual enrollment, based on internal coding. The feature ‘Num\_dual\_terms’ was compiled by counting the terms a student was encoded as a dual enrollment student.

The courses were handled differently to create features workable for a machine learning model. The data shows 194 possible courses taken by dual enrollment students. These are all formatted with a two-letter code to indicate subject and three digits representing the course. Columns were made for each possible course and one-hot encoded for each student based on their first semester courses.

The categorical features (HS, Ethnicity, StartTerm) were also one-hot encoded for use with the analytical models. HS is an encoded ID number referring to the origin school district of the student. Ethnicity is a two-character code identifying ethnicity as reported by the district, and StartTerm identifies whether the student started in Spring or Fall.

**Technical Approach**

This project utilized several classification models to predict whether a student would continue their education with us or not. Decision trees and random forests were my choice for testing, as the bulk of the data represents courses selected by students and hopefully will be accurately predicted with one or more decision trees. The classifications were all performed in Python.

Cluster analysis was also be performed on students tagged as not matriculating. K-means, PAM, and DBSCAN were the methods chosen for clustering the students into groups. Cluster analysis was performed in R.

**Feature Analysis**

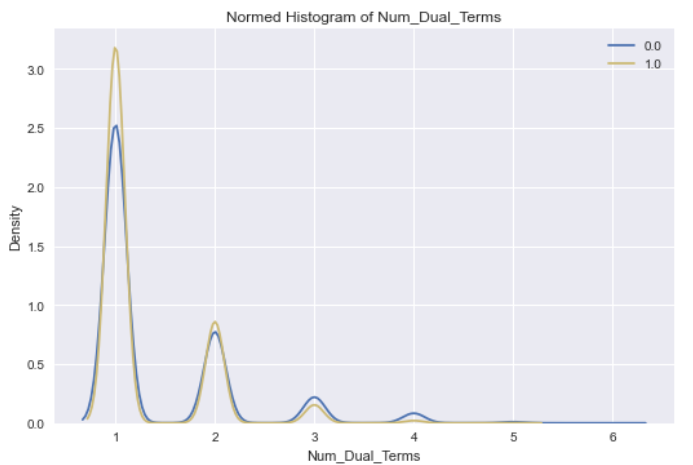
The features were all explored prior to working into a data model. No missing values are found in the dataset, but the Ethnicity feature has a ‘UN’ code to indicate unknown race. Empty fields were introduced when creating columns for each course but then filled with 0 when not flagged with a 1 indicating course enrollment.

Tables 1 and 2 show the results of a correlation test using Matric as the correlation target. Table 1 lists the top 10 results of a **corrwith()** function in Python. Table 2 lists the bottom 10, covering the range of strongest positive and negative Pearson correlations in the dataset. Age and StartYear have nearly opposite coefficient values which makes sense because students would have more time to eventually matriculate, especially if they started taking dual enrollment credits at a younger age than their peers.

Table

Overall the results of the coefficient testing do not show any strong linear correlations in the data. There may be some weak effects based on certain courses or school districts. This lack of linear correlation suggests a nonlinear approach to classification may prove to be more successful.

Table

The features were also explored with the Autoviz Python package. The package automatically plots features against one another for comparison. There were a few insights that may be useful in understanding what students will matriculate.

Figure

Figure 1 shows a plot of number of dual terms for both label values. While only one term is the most common, it appears that the more that are taken the less likely the student will stay with the school. This makes sense because the highest awards are only Associate’s degrees and a student can make significant progress by taking multiple semesters of college credit prior to graduation.

There are also indications from the Autoviz plots about some of the courses taken. While most plots are flat above 0.0 for the x-axis, a few show a rise at 1.0. These are all binary features, so seeing them visually made these several apparent. Figure 2 shows two courses that were taken more often by students who matriculated. Figure 3 is the reverse, with students who took it enrolling in a program of study less frequently. In both examples 1 represents matriculating students and 0 is those who did not continue with the college.

Figure



Figure

Taking into account the need for a good model that works well with nonlinear data and can output results that can help understand student decisions is a decision tree. I also selected a Random Forest ensemble model, which builds multiple decision trees and uses an averaging of the outputs.

**Model Deployment**

For both classification models, data was split into testing and training sets using **test\_train\_split()** from the Python sklearn library with 60% of the data for training and the remaining 40% held out for testing. The training set contains 12,085 samples and the validation set contains 8,058 samples. The models themselves are also from sklearn, **DecisionTreeClassifier()** from the tree package and **RandomForestClassifier()** from the ensemble package.

The clustering was instead performed in an R environment. Methods include K-means, PAM, and DBSCAN clustering types.

**Testing and Evaluation**

The results were checked with a confusion matrix plotted with the yellowbrick Python library. The classification models were tuned with hyperparameter testing using the GridSearchCV function of the sklearn Python library to ensure optimum accuracy.

**RESULTS**

Only a moderate amount of accuracy is expected given the paucity of information available for this student group for the model to consider. The cluster analysis may provide more actionable data depending on factors identified. High-performing students for example are likely to continue to 4-year colleges higher in their list of preferences. There is some self-selection bias introduced in the data because students who take extra college credit while in high school already tend to be more motivated.

**PROJECT MANAGEMENT/EXECUTION**

**Project Plan**

The two weeks leading up to the Milestone 3 deadline will consist of a preliminary analysis. The first week I have obtained clearance for use of the data and have started to pull the raw data from the student database. I have it clean enough for some initial analysis but still need to encode the target variable. The second week will include additional testing and writing of the preliminary analysis document for submission.

An additional three weeks for Milestone 4 allows time for plenty of additional modeling refinement while also coming up with a Powerpoint covering the work so far. I have professional experience delivering presentations in executive meetings so completing this is not a concern.

By the time three additional weeks have gone by I hope to have good results to present for my analysis. My master document will be refined enough for submission and a short presentation should be no problem to write. There is plenty of material to cover and I already enjoy talking about predictive work I’ve done with colleagues at data conferences so discussing my methods and results will be a pleasure.

**Project Risk**

Risks primarily involve the chance of an ineffective model due to a lack of useful features. There is a risk the institutional research board will not approve the use of data but the board has been notified in advance. There is a risk of violating FERPA guidelines if the data is not properly stripped of personally identifiable information prior to transferring from the school network.

**References**

Giacomelli, E. (2020, October 11). *SUNY-wide enrollment down for 10th year, but figures 'far better' than expected amid pandemic*. NNY360. https://www.nny360.com/communitynews/education/suny-wide-enrollment-down-for-10th-year-but-figures-far-better-than-expected-amid-pandemic/article\_a26f2cb6-5aad-5328-a769-f67d30704a0c.html.

Field, K. (2021, January 5). *The Rise of Dual Credit*. Education Next. https://www.educationnext.org/rise-dual-credit-more-students-take-college-classes-high-school-degree-attainment-rigor/.

Palmer, P. B., & O'Connell, D. G. (2009). Regression analysis for prediction: understanding the process. *Cardiopulmonary physical therapy journal*, *20*(3), 23–26.

EdTrust. (2019, January 24). *Dual enrollment*. The Education Trust. https://edtrust.org/issue/dual-enrollment/.