Wasay Khan

DSC 345

Project Write-Up

Abstract

A dataset of undergraduate students from the Polytechnic Institute of Portalegre (Portugal) is used with predictive models to determine if a student will dropout or gradate. The goal of this study is to determine critical features related to dropping out/graduating to drive action by administration in preventing more student dropouts. Variations of the dataset are also tested, using feature removal and feature engineering. Gradient boosting was observed as the best model for this dataset, with an accuracy of 0.91 (+/- 0.01) and AUC of 0.96 (+/- 0.01). This was consistent with performance for both modifications of the dataset. Feature selection across all datasets leads to the conclusion that academic performance and financial situation are the two main categories that lead to student dropouts. These two categories are represented across 18 of the 36 features in the dataset. To enact change, college administration must invest in their students by providing educational and financial assistance. This is through accessible, supplemental assistance outside the classroom and grants/loans/scholarships.

Introduction

Over the years, pursuing higher education has continued to be a route that many take. It has proved to be a great way to students to invest in themselves, allowing them to kickstart their careers across a variety of disciplines. However, there are often many setbacks that prevent students from completing their degree, which leads to lots of time and money wasted. In fact, only 60% of students that attend college graduate [1]. Depending on where the education is received, this can mean spending years of one's life without a degree to show for it or thousands of dollars wasted, which only gets worse with potential loans. It is only recently (post Covid era) that the value of a college education is beginning to be questioned, which means one's ability to earn a diploma influences the decision of attending college. This shaped the research question pursued in this paper, which was to understand the factors that influence a student's ability to dropout or graduate. This way, college boards/administration can better prepare their students for success within their universities by identifying those who may be more susceptible to failure. Better understanding of these factors will also allow students to understand the factors in their own life that can work for/against them, leading to a more calculated decision regarding attending college.

The dataset selected for this study was collected from the Polytechnic Institute of Portalegre (Portugal). Data regarding higher education statistics is minimal in Europe and the OECD, which makes working with this dataset more impactful. Rather than there being a single factor that leads to a high risk of dropout, it is rather a group of many factors together [2]. This makes the 37 total features captured in the dataset more useful since there is not necessarily a single answer needed. The dataset includes over 4,000 instances (each instance is a student) including various demographic, financial, and academic factors that influence a student's ability to graduate. Each student is also classified as having dropped out or graduated. It was sourced from the UC Irvine Machine Learning Repository, which is beneficial since this removes the headache of having to completely reformat the data and ensures that the dataset is beneficial for

solving the problem. Many of the features have been transformed to be represented as integers to assist with correct formatting for machine learning methods.

Literature Review

College attrition is an issue that has already undergone substantial research. This includes general research covering factors and trends in higher education as well as employing different machine learning methods to provide data-driven conclusions. Barbera et al. provides an overview of trends across different papers to provide a holistic understanding of what impacts college attrition. While this paper does not employ machine learning methods directly, it acknowledges the value of utilizing them as predictors due to the many factors that affect if a student will graduate. This paper notes that academic ability (GPA, test scores, academic rigor) plays the largest role in determining student success and will overshadow negative socioeconomic and demographic factors. This is primarily due to scholarships and loans alleviating some of the impact that negative factors have [1].

Another study by Aulck et al. aims to identify the leading features that correlate with dropping out on a University of Washington student dataset. This time, researchers employ different machine learning methods such as random forests, k-nearest neighbors, and logistic regression to predict if a student drops out. Logistic regression marginally performed the best, with all methods having very similar accuracy and AUC scores. The most interesting part of this study was the findings of features that had the highest predictive performance. Many of the listed features had to do with GPA and had an accuracy and AUC in the low-mid fifties. Because of this, we can conclude that GPA (aka academic performance) will have an impact on attrition but is not all that useful on its own. It must be used in conjunction with other features to paint a better picture [3].

Cardona et al. provides an analysis on publications regarding student attrition. This includes publications that employ all types of machine learning methods. All methodologies have varying degrees of effectiveness depending on the dataset used. However, neural networks were a standout in terms of use in publications. The most common methodologies were decision trees, neural networks, and support vector machines, all of which showed competitive model accuracy. There is a trend of using ensemble methods, but it must be tested further to deem if it has merit [4].

Lagman et al.'s paper was analyzed to investigate an effective use of ensemble methods. The same problem of predicting if a student graduates is solved, but a focus on boosting methods is employed to compare results to other methods. Once again, this paper analyzes common machine learning methods such as decision trees, neural networks, and the naive bayes algorithm, all of which perform similarly. A standout in logistic regression slightly outperforms the other 3 listed methods. The logistic regression method is further refined by using bagging, which is concluded as the best method in terms of accuracy. This example shows that ensemble methods have merit, which was introduced in Cardona et al.'s paper [5].

Adnan et al. takes a more in-depth approach to predicting student attrition. The authors use feature engineering to mark performance at different time periods of a course for an online learning program. This includes demographic data, clickstream data (online engagement with material), and assessment scores. There are also separate phases of testing where the researchers focus on different categories of features (Phase 1: demographic, Phase 2: demographic and

clickstream data, Phase 3: demographics, clickstream, and assessment scores). This way, the researchers can clearly identify the impact that different categories of features have on model accuracy. While this paper focused on a single course rather than dropout rate, it still provides a useful approach to preprocessing techniques. This paper also lacks the detail on specific features, as the focus is the effectiveness of feature engineering [6].

Methodology

Since this dataset came from the UC Irvine Machine Learning Repository, problems like missing data and discretizing categorical values were already addressed. This removed a lot of the work needed to convert many of the features into values that could be interpreted by the machine learning models. However, initially the dataset was a multiclass classification problem (dropout, enrolled, graduated) which needed to be converted into a binary classification problem (dropout and graduated). This was done by removing all entries with the "enrolled" class, which was not a problem since there was initially a large class imbalance that is addressed by removing this class.

Due to findings from literature, initially a holistic approach was taken for testing different machine learning methods to discover which method was best for this dataset. This included using decision trees, random forests, neural networks, different boosting methods, and SVMs. Decision tree testing included using both the gini and entropy criterion across 3-50 folds, both of which provided similar results. Random forests were tested using a 5-fold cross validation and a 10-fold cross validation, across 5-500 trees. Wrapper based feature selection was also conducted to keep track of selected features of importance. Neural network models were tested using all 3 solvers (sgd, adam, and lbfgs) with and without feature selection. The gradient boosting model was tested by changing maxdepth from 3-10, as well as the number of n_estimators from 50-200. This was the same case for Ada boosting. Both gradient boosting and Ada boosting were tested using 5-fold cross validation. The SVM model testing used all three kernel types (linear, sigmoid, rbf) with both wrapper-based feature selection and linear stepwise backwards removal.

Different feature selection methods were used in conjunction with these models including stepwise recursive backward feature removal and wrapper-based feature selection. WB FS proved to be the most effective. Through all findings, there were specific features regarding academic performance that would appear multiple times. I was curious as to how the model would perform without these features as it seemed redundant to discover that students who do well in school tend to graduate while those who do poorly fail. At this point, the accuracy and AUC scores recorded were consistent and great across the board (will be elaborated upon later), so I decided to run the all-model tests again on an edited dataset which removes all features that had to do with academic performance in college. This removed a total of 12 features, and the results were still acceptable.

While conducting the testing without curricular features, it raised the question of if feature engineering could be used to simplify all curricular features into one. This would still provide some usability of academic features without them being overrepresented in the feature selection process. This was done by consolidating 12 features regarding academic achievement into 1. The most selected features across all models were curricular units (enrolled), curricular units (approved), and curricular units (grade). These 3 features were listed for both first and second semester totaling 6 features. The other 6 curricular features provided were not included in any wrapper-based feature selection and didn't seem like a useful metrics, so they were deleted.

A new feature called “academic score” was created comparing credits enrolled versus credits earned for both semesters and it checks if a student is above the average curricular unit (grade). The feature holds a value of either 0 or 1, which was opted for over holding a continuous value due to not knowing what determines a passing grade. Once again, testing across all models was done with this dataset.

Results

Results across all variations of datasets were very good, however the most extensive testing occurred on the first dataset as this was done to explore all variations of models. As expected, decision trees performed the worst yet still maintained an accuracy of accuracy of 0.85 (+/- 0.01) and AUC of 0.85 (+/- 0.01). RFs, NNs, SVMs, and boosting methods were all close in terms of score, but there were a few drawbacks to some methods. For example, SVMs maintained an accuracy of 0.91 (+/- 0.01) and AUC of 0.95 (+/- 0.01) , but this was under the linear kernel type which led to extremely long runtime of 236.73s. This was remedied with wrapper-based feature selection, which provided the exact same scores with a far shorter runtime. However, there were still other methods that outperformed this.

Neural networks were investigated next, which was a commonly selected method in literature. As expected, neural networks without feature selection provided lackluster results, accuracy of 0.61 (+/- 0.01) and AUC of 0.49 (+/-0.07), failing to capture a lot of the data. However, once WB FS is used, the adam solver provides impressive results with an accuracy of 0.90 (+/- 0.02) and AUC of 0.94 (+/- 0.02) having a runtime of 3.97s. This was the only model to have significantly higher scores when using feature selection. This validates the findings from literature, but there were still slightly better models for this dataset.

Random forests with 10-fold cv and 20 trees yielded the highest scores for RFs with an accuracy of 0.91 (+/- 0.02) and AUC of 0.95 (+/- 0.02). When WB FS was conducted, we see marginally worse scores. The best model used was gradient boosting without feature selection with a max depth of 3. This yielded an accuracy of 0.91 (+/- 0.01) and AUC of 0.96 (+/- 0.01). Runtime for this was 3.24s. This model ended up being the best by a small margin, and, like RFs, feature selection was slightly worse.

Using the findings from the first dataset, I focused specifically on configurations of models that provided the best results when testing the dataset without academic performance features and the dataset using feature engineering. After all testing was completed, gradient boosting without feature selection was among the best models in terms of performance for dataset 2 and 3. It should be noted that dataset 3 had a tighter spread of standard deviation for gradient boosting with feature selection. However, actual scores were the same.

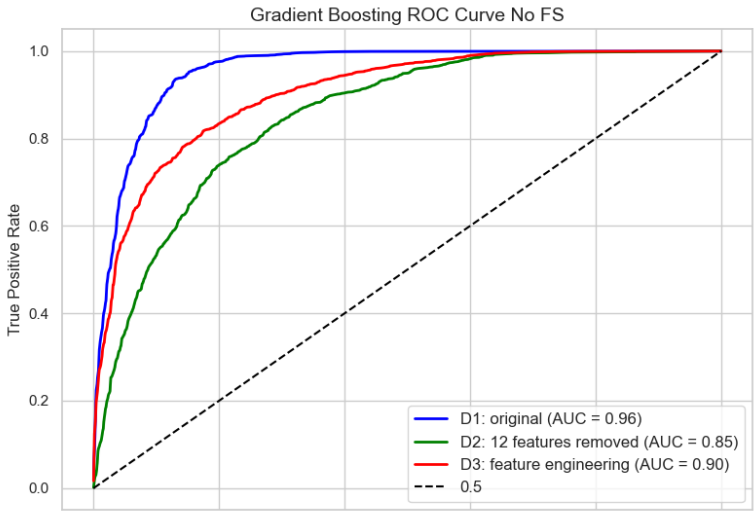

Figure 1 – Gradient Boosting ROC Curves

Interpreting figure 1, we can see that the raw dataset performed the best compared to the dataset with academic performance features

removed and the feature engineering dataset. However, I believe all datasets still have merit to accurately predict dropouts/graduates. The findings of this chart are consistent with beliefs that academic achievement and financial status features are important for predictions. It emphasizes how important the academic metrics are, since the dataset lacking them perform the worst, and the dataset that has a single engineered academic metric sits in the middle.

Commonly Selected Features	
Dataset 1	curricular units 1st sem (approved), curricular units 2nd sem (approved), curricular units 1st sem (grade), curricular units 2nd sem (grade), Curricular units 1st stem (enrolled), curricular units 2nd sem (enrolled), Tuition fes up to date, Course (major)
Dataset 2	Course (major), Tuition fes up to date, Debtor, Scholarship holder, Age at enrollment
Dataset 3	Course (major), Tuition fes up to date, Debtor, Scholarship holder, Age at enrollment, Academic Score (feature engineered)

Figure 2 - Commonly selected features across 3 datasets

Referencing figure 2, results from feature selection in the first dataset is what initially opted for creating other datasets to see what else influences dropouts. The commonly selected features would have been dominated by variations of the curricular units feature, which is useful to know but more conclusions should be drawn. When these features are removed, features regarding finances appear, proving that this also impacts dropout rate. In our feature engineering dataset, financial features as well as the engineered feature appear. This proves that the engineered feature has merit for predicting dropouts/graduates, and further emphasizes the importance of both, academic features and financial features. Course (major) also appears in all 3 datasets, suggesting that there is an imbalance among dropouts across majors.

Discussion

Depending on which dataset we look at, we can see varying degrees of success. Across the board, all models provided scores comparable to each other for each of the 3 datasets. Specifically for the first model, we see the highest accuracy and AUC among all 3 datasets, primarily due to it including academic features that are either removed or modified for the other two datasets. Feature selection also informs us that academic features hold a lot of weight when it comes to predicting if a student will graduate or dropout. The downside here is that this is obvious, and all this does is confirm what can be assumed. The second dataset, which lacks all forms of academic performance features, has the worst model scores compared to the other two datasets while still providing results that tell us that the models have merit for predicting if a student will dropout. The most important part of testing this dataset were the features extracted from feature selection, which paints a better picture for the factors that impact student dropouts. This dataset shows that financial situation and course (major) rigor also plays a role in dropout rates. Tuition fees up to date, debtor, scholarship holder, and course are all features that are selected among the different models. This is important since they are more actionable metrics

that can be investigated rather than a student's performance. "Course" refers to the major selected, which makes sense given some majors are harder than others. This can lead to investigating if some majors are too difficult for the undergraduate level.

The third dataset, which removes all academic performance features and replaces them with the engineered feature of "academic score", is what I think worked the best among all datasets. While this dataset does not outperform the first dataset with all features, the accuracy and AUC scores earned through the inclusion of an engineered feature speaks to the importance of academic performance while also providing color for other features that impact dropout rates. The engineered feature was included in feature selection for some of the models, proving that academic success is one of the most crucial factors. There was room for improvement in the engineering of this "academic score" feature since it simplified many continuous features into a single binary one. Despite this, this feature was still selected while contributing to scores that were better than the second dataset, proving the success of this feature. Repeats of other features regarding financial status and course rigor appear again, solidifying their importance.

One aspect of the modeling that did not work was univariate feature selection. At first there were performance issues that would freeze my computer which I assume is due to the large number of features and dataset. There was also the issue of having negative values in inflation rate and GDP which prevents this type of feature selection from running. The dataset already had preprocessing conducted on it, and I was confused on if re-normalizing inflation rate and GDP would have unforeseen consequences since the rest of the dataset is normalized. This method of feature selection would have provided a ranked list of features which could further confirm the existing findings.

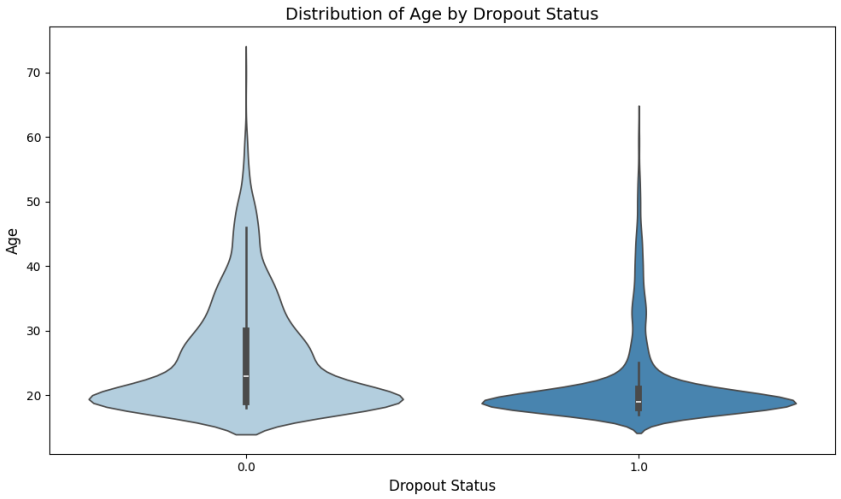

Figure 3 - Distribution of Age by Dropout Status

An aspect of the dataset that could be investigated further can be how age impacts dropping out. In figure 3, the left violin plot maps dropouts while the right is for graduates. We can see that graduates skew much younger and there are far more dropouts that are above the traditional age of university. It would be interesting to partition the data for non-traditional students and see which features are significant. This way, initiatives for non-traditional students can created, since chances are they face a separate set of problems preventing them from graduating since they are underrepresented on the graduate side. Perhaps features regarding familial financial assistance would be interesting to include since it can be assumed that the younger people graduating are having some help.

Conclusions and Future Work

Academic success and financial situation are the two largest factors that impact a student's ability to graduate. The former was proved through analysis on the raw dataset while the latter was made clear by manipulating the dataset to remove academic performance features.

This approach was taken as it did not seem interesting enough to me to say that those who performed well in their college courses graduate while those who don't dropout. Especially, giving the little amount of context the dataset provided for how curricular units are evaluated. Removing 12 academic success features highlighted how important financial status is, as this affects a student's ability to attend and succeed in school. If a student is working full time, they are not able to dedicate the same amount of time as a student who has a better financial status, meaning that this impacts chances of dropping out. The feature engineered dataset reinforced the ideas provided by the first dataset, since selected features included those relating to financial status and the engineered feature, relating to academic performance.

Next steps include comparing findings from this dataset to others in Portugal or another part of the world with a similar economic and higher education system. This is because higher education differs worldwide, and analysis must be completed with caution to ensure that accurate comparisons are being made [2]. For example, comparing financial situations in Portugal versus the United States would need to be done very carefully due to the extreme cost of education in the U.S. and the loan systems present. An introduction of studying psychological factors was mentioned in Barbera et al.'s paper due to the perceived mental health of the current generation [1]. I think it would be more interesting to first understand what causes those returning to school to never complete it, or leading factors of dropping out for non-traditional students.

To act on the results found requires institutional investment. This includes support for remedial education or more scholarships/grants to relieve the financial burden on struggling students. How much money college administrators spend on their students directly impacts dropout rates [1], yet each university will have a different approach to this that can cloud overall understanding. This is often a metric not provided in datasets but should be included since it seems like common sense that if a college invests more resources into students, they tend to do well. In fact, this is part of the reason I even attended DePaul. There was less of a financial burden by attending DePaul which allows me to not have to work full time while being a student. To say this benefits me is an understatement, as I would have 40 extra hours a week to dedicate to schooling.

Overall, academic performance features and financial features hold the largest impact on determining the success of a student for this dataset. College administrations can use this information to invest in their student's success, ensuring that the student has sufficient educational and financial support. This can be done through accessible, extra assistance outside of the classroom and scholarships/loans/grants.

References

[1] S. A. Barbera, "Review of Undergraduate Student Retiton and Graduation since 2010: Patterns, Predictions, and Recommendations for 2020," SAGE, 2020.

[2] ,. D. N. 1. Iván Sandoval-Palis 1, "Early Dropout Prediction Model: A Case Study of University Leveling Course Students," MDPI, 2020.

[3] L. Aulck, "Predicting Student Dropout in Higher Education," ICML, Seattle, 2016.

[4] T. Cardona, "Data Mining and Machine Learning Retention Models in Higher Education," MAGE, St. Louis, 2023.

[5] A. C. Lagman, "Classification Algorithm Accuracy Improvement for Student Graduation Prediciton Using Ensemble Model," International Journal of Information and Education Technology, Manila, 2020.

[6] M. Adnan, "Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models," IEEE Access, Kohat, 2021.