Lily Guizatoullina 05/09/2024

Prof. Agnellini DIDA325

Predicting Student's Academic Outcomes

Introduction

This project aims to identify which students are more likely to graduate, with the ultimate goal of helping universities figure out how to allocate resources to ensure students are able to finish their coursework in a timely manner and prevent students from dropping out. This paper discusses the importance of this goal, and how a dataset with student demographic information and outcomes is analyzed to identify what factors influence a student's likelihood of dropping out before the end of their course. It covers some of the results of the analysis, and how those results suggest various ways that universities can provide more support to their students.

Background Info

At the individual level, students that drop out are more likely to be associated with emotions concerning inadequacy/self-doubts/not belonging, and will often end up wasting personal resources, time, and sometimes money (Larsen et al., 2013). Many students also go on to face financial struggles after graduation, with expected earnings of $21,000 less than their graduate counterparts per year (ThinkImpact, 2021). While dropping out of college may be the advantageous choice for some students, and success stories such as Mark Zukerberg exist, the majority of dropouts face many negative consequences suggesting that measures to prevent students from dropping out can be beneficial.

Not only can dropping out of university be a devastating event for an individual student, but it will often negatively impact the university from an economic perspective because of how a

dropout can interrupt funding schemes (Jadrić et al., 2010). Therefore, because of economic and societal loss that students dropping out can incur, many universities have been searching for promising measures and programs to identify and help students at risk of dropping out (Behr et al., 2021). This project aims to see how machine learning models can be used to identify which students are more likely to drop out, which can help universities narrow down what approaches should be used to help these students.

Description of your Data and Methods

I used a dataset from the UC Irvine Machine Learning Repository that looked at student data from the Polytechnic Institute of Portalegre and used several different sources,[1] that looked at demographic and sociological data related to the students, and macroeconomic data from the time period they graduated from. The response variable that I examined was "Target," which indicated the status of a student at the end of the standard duration of their course. In order to predict a students outcome, I used a logistic regression model since I was primarily working with categorical variables, and a random forest model since many of the predictor variables were hierarchical in nature. There were three outcomes in this dataset, which were "Graduated," "Enrolled," and "Dropout." Unlike all other categorical variables in the dataset, the "Target" column was not factored, so part of the cleaning process included changing the column to consist of numerical values.

[1] Data sources included (i) the Academic Management System (AMS) of the institution, (ii) the Support System for the Teaching Activity of the institution (developed internally and called PAE), (iii) the annual data from the General Directorate of Higher Education (DGES) regarding admission through the National Competition for Access to Higher Education (CNAES), and (iv) the Contemporary Portugal Database (PORDATA) regarding macroeconomic data

I made several bar graphs to visualize the data to try and spot trends within the data. There were several trends that stood out, and creating a correlation matrix revealed that these variables were strongly correlated with a student's outcome. The graphs and correlation matrix helped me decide which variables to use for the final models. During the visualization process it also became apparent that the "Enrollment" category of students tended to behave in a more unpredictable manner and it was difficult to spot any trends in that category compared to the "Graduated" and "Dropout" categories.

After visualizing the data, I wanted to factor the "Target" so that it could be used in machine learning models. During this cleaning process, I tried removing any rows with the "Enrolled" outcome because their outcomes were ambiguous, as it could not be known if those students graduated late or went on to drop out. After testing out the models with both the full dataset, and the dataset without the "Enrolled" students, I found that removing those rows made the models more accurate, so the final models do not use that data. I also tried grouping the "Enrolled" and "Dropout" students into one category, but found that removing the "Enrolled" students altogether had the best results.

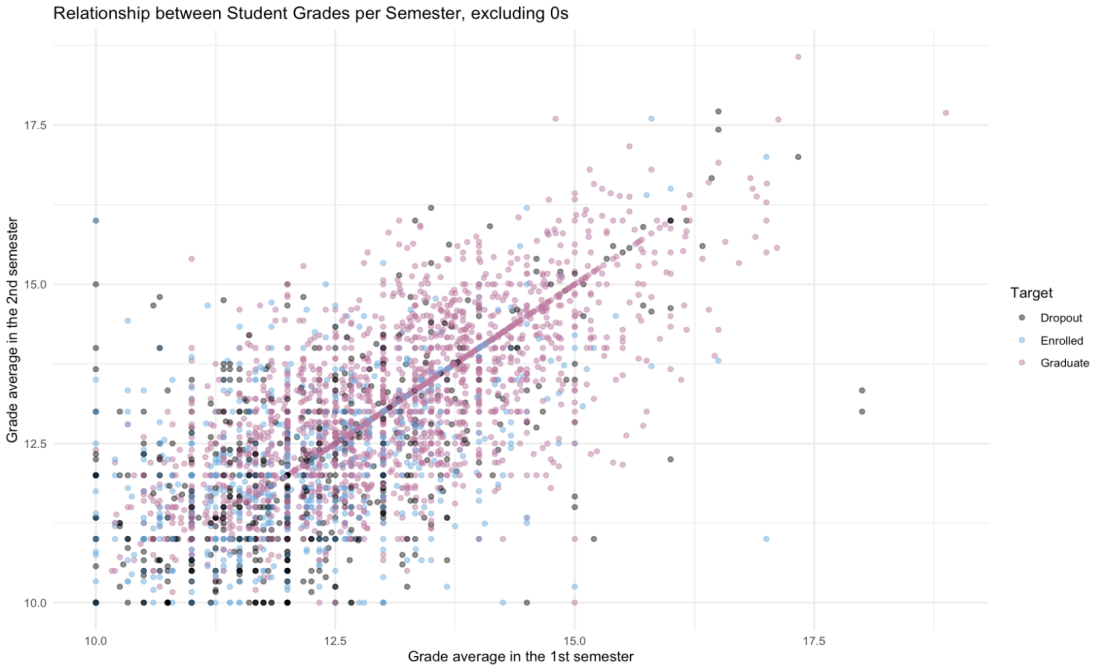
I also tried limiting which predictor variables were used, and found that the accuracy of both models slightly improved when only using predictor variables that had a strong correlation with the "Target" variable[2]. I also incorporated the macroeconomic data because it improved the accuracy of the logistic regression model, but it did not affect the random forest model. Because the dataset small (3630 students were analyzed after removing students that were under the

[2] The predictor variables used in the final model are grades for both semesters, number curricular units approved and without evaluations for both semesters, students' debtor status, if tuition had been paid for that semester, gender, age at enrollment, educational special needs, application order, attendance, admission grade, displacement status, previous qualification and macroeconomic data (GDP, inflation rate and unemployment rate).

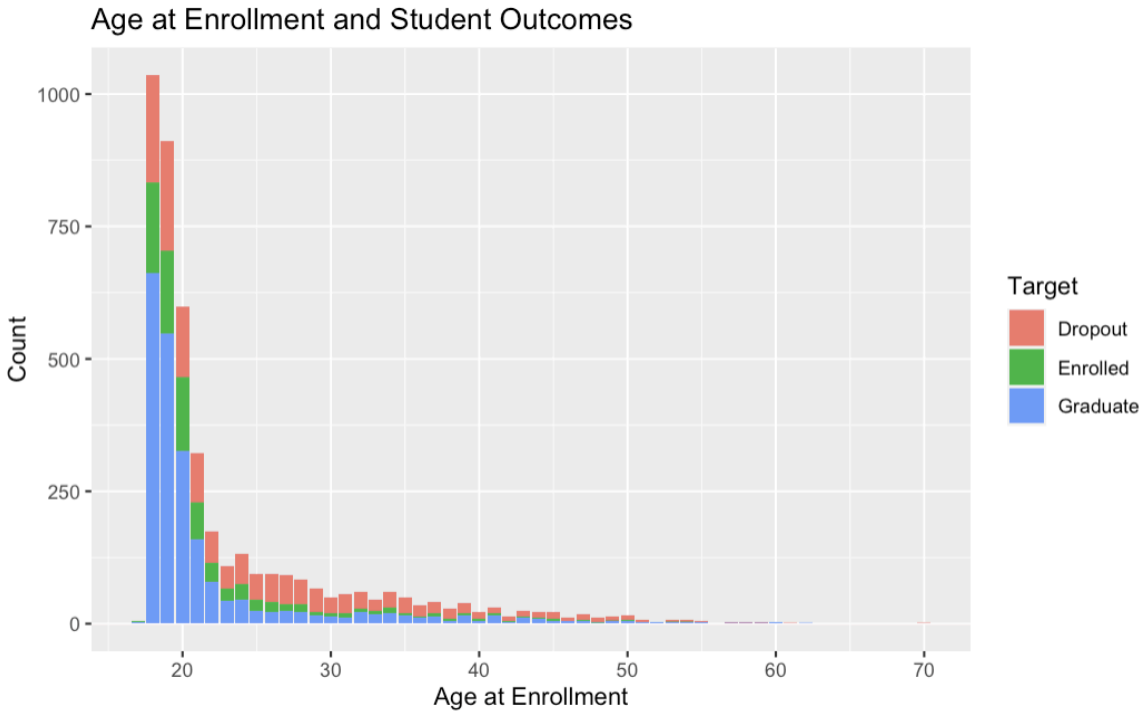
"Enrolled" category), this may have been a result of overfitting, which suggests that it could be valuable to use a similar process on a larger dataset to determine if these variables are helpful and compare the performance of both models.

Results

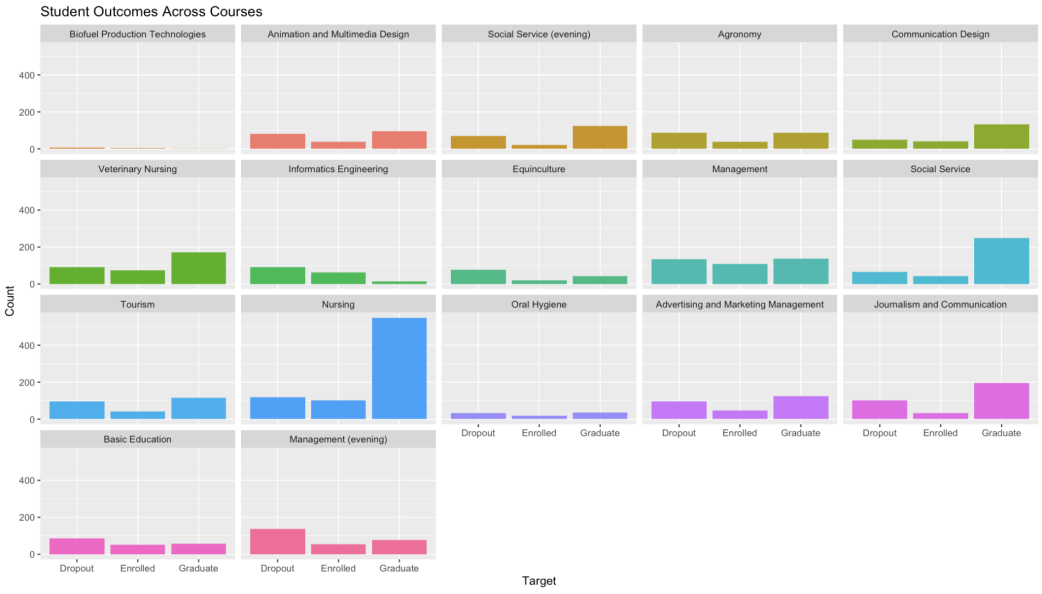
It was observed that student's credits were strongly correlated with each other, and some of those categories (such as approved credits, credits before evaluation and grades) showed strong correlation with student outcomes. It appears that most students did not see a drastic change between semesters, so students who earned high marks in the first semester would generally perform well during the second, as shown below. It is also seen that students who graduate tend to receive higher grades, but there is not a noticeable difference between students who dropped out or remain enrolled.



Additionally it was found that students who began college younger tend to make up a larger portion of each class, and make up an even larger option of the graduating class.



Students in different courses (which refers to the type of degree they are studying) also show different outcomes.



These variables, along with others that showed strong correlation and the macroeconomic with the response variable were used to create two models. The logistic regression model performed with an accuracy of 91.74% and the random forest model performed with an accuracy rate of 88.35%. For the logistic regression model, the residuals were skewed towards the left implying that the model had a tendency to make false negatives where it predicts students that end up dropping out as graduates.

Conclusions

Although the logistic regression model was able to predict student outcomes with 91.74% accuracy, it runs into the problem of over-estimating the amount of graduates, which can be problematic if this model is to attempt to actually identify which students may need more resources or support. Future improvements should include looking at larger datasets and correcting for this issue. It is also worth noting that the best variables at predicting student outcomes in this dataset had to do with student's grades and credits (enrolled without evaluations and enrolled with approval), and while this can help identify which students may need more support, they are less helpful compared to other variables in identifying what types of support students may need.

This analysis shows that nursing students are composed of a large number of graduates, which may be explained by the difference in course length (a typical nursing course at this university lasts four years, unlike other courses which only last three). However, this could also be caused by a large proportion of nursing students in this dataset being female, and more female students representing graduates overall. It could also be caused by external factors such as students who are more likely to graduate also being more likely to choose nursing as their course,

or the nursing course at this specific university providing substantially more support to students compared to other courses. Despite this, expanding more courses to be longer should be considered for more courses that have high dropout rates, such as Management and Biofuel Production Technologies. It may be the case that students will be less likely to drop out of a program if they are given extra time, even if it means the course becomes harder or more expensive.

It is also worth investigating how finances impact the likelihood of a student graduating. Whether or not a student has a scholarship, is in debt, or pays their tuition on time all influence whether a student will end up graduating or not, so financial support needs to be evaluated as another potentially valuable resource to support students. Since high dropout rates can negatively affect a university from an economic perspective, providing more financial support has the potential of increasing profits, and should be considered as an option. For example, a university could provide additional scholarships targeted at displaced students, or in courses with high dropout rates to decrease the likelihood of those students dropping out.

References

Jadrić, M., Garača, Z., Ćukušić, M.: Student dropout analysis with application of data mining methods. Manag. J. Contemp. Manag. Issues 15(1), 31–46 (2010)

ThinkImpact. (2021). College Dropout Rates. https://www.thinkimpact.com/college-dropout-rates/

Larsen, Malene Rode. Sommersel, Hanna Bjørnøy. Larsen, Michael Søgaard.: Evidence on Dropout Phenomena at Universities. Danish Clearinghouse for Educational Research (2013)

Behr, Andreas. Giese, Marco. Kamdjou, Herve D. Teguim. Theune, Katja.: Motives for dropping out from higher education—An analysis of bachelor's degree students in Germany. European Journal of Education. Volume 56 Issue 2 (2021)