A Study on University Students' Success and Dropout Rates

CS 216 Project Report
April 26th, 2023
Dillan Sant (dks43), Thomas Barker (jtb80), Talia Granick (tjg41), Nolan Zhong (nxz), Isabelle Xiong (yx215)

Part 1: Introduction and Research Questions

In this research project, we are examining the success and dropout rates of university students, as we would like to know more about what kinds of factors contribute positively or negatively to college students' experiences and outcomes.

We are focusing on the research question: "Which factors contribute most to an undergraduate student's propensity to drop out?" We found this research question to be particularly relevant because there is still a huge proportion of U.S. students dropping out of college, and being able to identify factors contributing to this issue could hopefully lead to this proportion decreasing.

Ultimately, we are hoping to address the problem of students dropping out of university, so we chose to focus on an in-depth analysis of which kinds of factors/combinations of factors seem to contribute most to a student graduating or dropping out.

Part 2: Data Sources

For this research project, we used two separate data sources that both came from Kaggle.

Data Source #1:

This data source is called "Predict students' dropout and academic success", and we accessed it from Kaggle. The link to the dataset can be found in the citation below. The data was collected from a higher level education institution and includes information known about the students at the time of their enrollment as well as their academic performance by the end of their first and second semesters. This data source provides one dataset, which contains several demographic variables (like gender, age, nationality, etc.) for university students, as well as whether or not they dropped out of college, graduated, or are still attending. These metrics are very valuable for our research question, as we can see how different demographic variables may affect dropout rates. As there is only one dataset from this source and the CSV is pretty clean from the start, not much intensive data wrangling was needed to use this source.

Citation:

The Devastator. (2023, January). Predict students' dropout and academic success. Retrieved March 5, 2023 from https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention

Data Source #2:

The data source is called "College Completion Rates and Efficiency Measures for US", and we accessed it from Kaggle. The link to the dataset can be found in the citation below. This source contains data from the National Center for Education Statistics' Integrated Postsecondary Education System (IPEDS) and the

Voluntary System of Accountability's Student Success and Progress Rate. The data source has information on college completion rates and associated efficiency measures in the United States and provides us with four datasets: institution_details, institution_grads, state_sector_details, and state_sector_grads. We decided to focus on the state_sector_grads dataset, as we felt the variables present in this dataset are most important to our research question. Through this specific dataset, we analyzed factors such as state, race, and graduation rate to identify potential predictors of student dropout rates.

Citation:

The Devastator. (2023, January). College Completion and Efficiency Measures for US. Retrieved March 5, 2023 from https://www.kaggle.com/datasets/thedevastator/college-completion-and-efficiency-measures-for-u?select=cc_state_sector_grads.csv

Part 3: What Modules are You Using?

Aside from modules 1-3, we also used concepts from modules 5, 8, and 9 in our research.

Module 5: Statistical Inference:

After we make our hypotheses, statistical inference will help solidify whether our intuitions regarding correlations impacting success and drop out rates at colleges are correct or not. Our justification is that it will help us understand which variables in our datasets have significant effects on dropout rates, allowing us to test our hypotheses. We use concepts like the central limit theorem and confidence intervals in our testing. Statistical inference is a large portion of our data analysis stage, and it helps us determine the significance of a number of variables in this stage. More specifically, this module was most useful in conducting confidence intervals and hypothesis tests on different explanatory variables such as age and institution type, in order to tell us if those variables may have a significant impact on a student's outcome.

Module 8: Visualization:

Visualizations help us show how a particular variable affects student outcome/dropout rates. Our justification is that visualizations help in conveying information well and clearly. Concepts like bar plots aid us in seeing the differences among variables, and separating visualizations by certain metrics or demographic variables help us see different relationships among many variables with regard to dropout rates. Visualizations are present in our data analysis stage to deepen a reader's understanding of how student outcomes and dropout rates differ based on certain explanatory variables such as race, gender, and more.

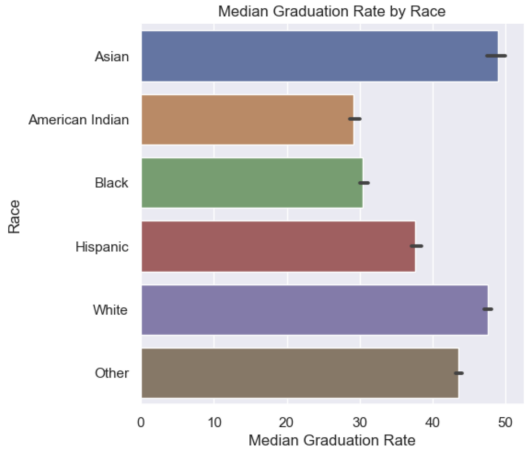
Module 9: Prediction & Supervised Machine Learning:

Supervised machine learning is essential for us in seeing which variables help best predict whether or not a student would drop out. Our justification is that comparing the accuracies of different predictive models lets us see which kinds of variables are impacting student outcomes. Concepts like logistic regression provide us with a model to predict a student's outcome. Supervised machine learning is present in our data analysis and final report stages, as they can show readers the extent to which different factors are impacting a student's propensity to drop out.

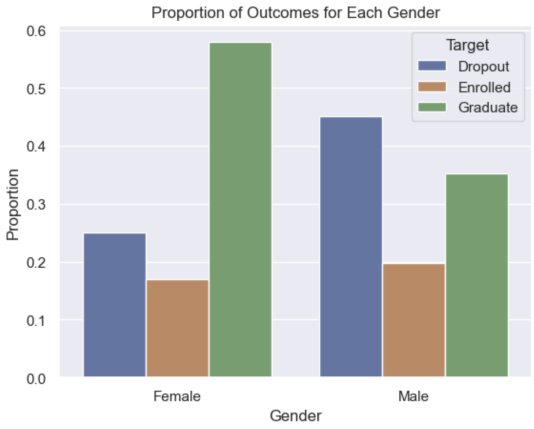
Part 4: Results and Methods

To start off, for the data sources we chose, not much data cleaning was needed as the datasets from Kaggle were generally clean and pre-processed. Since our two different datasets came from different sources and have no linking variables, each is analyzed in a separate setting. As there were 10s of different potential explanatory variables to analyze, we were not able to conduct tests on all of them, but we at least produced relevant visualizations for the variables we thought could impact students' outcomes and were able to go more in-depth with tests and regressions for some of these variables.

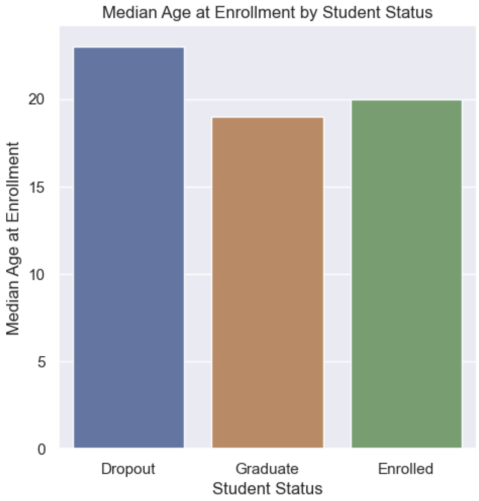
The first demographic variable we chose to look at was a student's race. To get a good idea of how a student's race may impact their graduation rate, we plotted the median graduation rate based on race (both taken from the 'cc_state_sector_grads' dataset) by using a Seaborn barchart with the median as the estimator. As we can see in the below graph, the Asian and White groups have the highest median graduation rates while the Black and Native American groups have the lowest median graduation rates.



Also, for individual demographics, we looked at the proportion of each outcome experienced by each gender. To do this, we first grouped the 'student_dropout_academic_success' data frame by gender and outcome and got the counts of each variable. Then, we grouped the resulting data frame by gender and applied a lambda function to get the proportion of each outcome. As we can see in the below graph, there is a higher proportion of dropouts among males than females. The effect of a student's gender on their outcome is also further explored later in this section with logistic regression.



For explanatory variables that we explored more in depth, we first wanted to take a look at students' age at enrollment and its potential effects on their outcomes by looking at our first dataset, 'student_dropout_academic_success'. To do this, we grouped the data by whether a student was a dropout, enrolled, or a graduate and calculated their mean age at enrollment. This bar graph was made using the seaborn catplot function, with the x-variable set to student status, and the y-variable set to age at enrollment with median as the estimator.

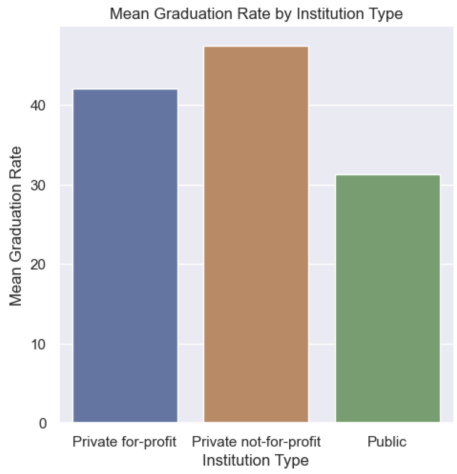


As shown by our above graph, it seems as though students who dropout generally have a higher median age at enrollment around 24 while those who graduate have a lower mean age at enrollment at around 18. To take a look at the significance of these age differences, we conducted a 95% confidence interval for the median enrollment ages of graduated, enrolled, and dropout students.

```
Confidence Interval for Median Starting Age of Graduated Students:  (18.720730603181103, 19.279269396818897)
Confidence Interval for Median Starting Age of Enrolled Students:  (19.560937918814982, 20.439062081185018)
Confidence Interval for Median Starting Age of Dropout Students:  (22.54705919212147, 23.45294080787853)
```

As shown by the lack of overlap of our confidence intervals, at the 95% confidence level, there is a significant difference in the starting ages of students based on their outcome. A potential cause of this could be that those who enroll at older ages may have a higher desire to dropout as they are already at an age where their peers are much more likely to be working.

A second explanatory variable that we explored more in-depth was the type of institution a student was attending, and how that may have affected their graduation rate, by looking at our second dataset, 'state_sector_grads'. We first visualized this variable by creating another bar graph using seaborn's catplot function, with the x-variable set to be the institution type ('Public', 'Private not-for-profit', and 'Private for-profit') and the y-variable set to be the graduation rate with the mean as the estimator.

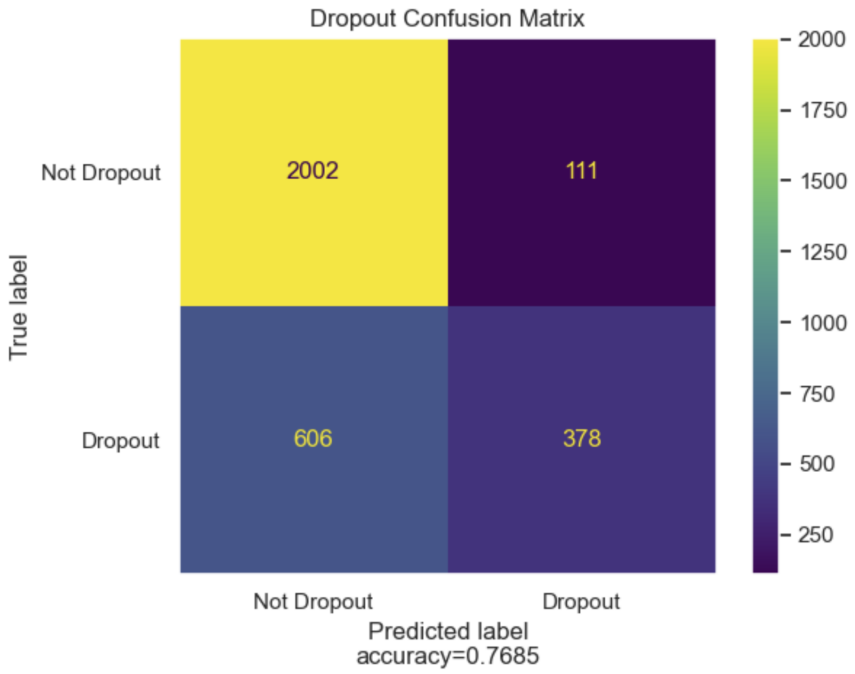


As shown by our above graph, it appears that private not-for-profit institutions have the highest mean graduation rate at around 46%, and public institutions have the lowest mean graduation rate at around 31%. To look at the significance of these differences, we chose to conduct a t-test of independence between the mean public school graduation rate and mean private school graduation rate (we grouped together for-profit and not-for-profit private institutions).

```
Mean Graduation Rate of Public Schools:  31.261984422621065
Mean Graduation Rate of Private Schools:  38.24154050618282
P-value:  0.0
```

As shown in our results, with a p-value of approximately 0.0, we can conclude with significance that the mean graduation rate is higher for private institutions than public institutions.

Finally, we wanted to revisit our first dataset, ‘student_dropout_academic_success’, and look at which combinations of factors would best predict a student’s outcome. To do this, we performed a logistic regression with the following x/explanatory variables: 'Gender', 'Scholarship holder', 'Debtor', 'Tuition fees up to date', and ‘Age at enrollment’, and a variable called ‘Dropout?’ as the response variable. It is worth noting that the first four explanatory variables ('Gender', 'Scholarship holder', 'Debtor', 'Tuition fees up to date') were one hot encoded, and that the explanatory variable was created by assigning a 1 to students who dropped out and a 0 to students who did not (either graduated or still enrolled). We split our data into training and testing sets, trained the logistic regression model on the training data, and then created a confusion matrix and checked the accuracy of our model using the test data. We found that this combination of explanatory variables produced the highest accuracy in our model at around 76.85%, indicating that perhaps these are some of the most important factors that influence a student’s decision to drop out of university.



Lastly, we wanted to see how these variables were affecting student outcomes using the coefficient method of the logistic regression model. As seen by the variables with positive coefficients from the logistic regression (in the table below), we found that factors like being male, not having a scholarship, being a debtor, not having your tuition up to date, and being older all positively contributed to a student dropping out/made them more likely to drop out.

Feature	Coefficient
Gender_Female	-0.233886
Gender_Male	0.232787
Scholarship holder_No Scholarship	0.658032
Scholarship holder_Scholarship	-0.659131
Debtor_Debtor	0.152930
Debtor_Not Debtor	-0.154029
Tuition fees up to date_Not up to date	1.305195
Tuition fees up to date_Up to date	-1.306295
Age at enrollment	0.043187

Repository Link to Code: https://duke.box.com/s/mzcpm1jhlghhusg6oj5k4brrdtzjozwv

Part 5: Limitations and Future Work

While we were able to make many meaningful conclusions about what factors contribute to a student's ability to succeed and graduate in school, our data did have a few limitations. Especially in the second

data source we used, “College Completion Rates and Efficiency Measures for US,” many of the variables were redundant or hard to use in a statistical modeling context. Some of the data was not numerical, while other parts of it were hard to understand, which led us to only utilize one of the four data sets within the entire second source. However, our group felt that we were ultimately able to make enough substantial claims about the data, even though the second source was not as applicable as we thought it would be prior to working with the data itself.

In the future, it would be interesting to further assess and analyze how race contributes to graduation and drop out rates. In our data, we were able to see the breakdown of graduation rates by race, however our conclusions about these differences are only inferences. By further exploring trends within races, like whether certain races are more likely to attend private or public school or racial familial expectations, our group could make more concrete, meaningful conclusions about how race impacts a student’s success in school. Another interesting topic to explore would be to what extent career interests contribute to graduation rates. Perhaps going into college as “pre-law” or “pre-med” would positively or negatively impact the longevity of undergraduate university students.

Part 6: Conclusion

Our analysis of individual and university level data of students across our two datasets helped us gain insight into which factors contribute to a student’s outcome in higher-level education. Through visualizations, confidence intervals, hypothesis testing, and logistic regression, we found that variables such as race, age at enrollment, gender, and type of institution all play a role in a student’s propensity to drop out. Specifically, we found that being male, not holding a scholarship, being in debt, not having tuition fees up to date, and being older all positively contributed to a student’s likelihood of dropping out. Moreover, we identified a significant difference in graduation rates between private and public institutions, with private institutions having a higher mean graduation rate.

Overall, our research sheds light on some of the key factors that influence a student’s decision to drop out of higher-level education, which can be helpful for educational institutions and policymakers aiming to reduce dropout rates. Future research could explore alternative factors or further investigate the relationships between these factors and dropout rates, as well as examine the effectiveness of intervention strategies based on the identified predictors.