Predicting which college students will complete their degree with different techniques

For my thesis, I will use various data science techniques to predict which students will complete their college degree in the expected time, a longer time, or not at all. The dropout rate for all undergrad students is 32.9%, almost a third (What to Become, 2021). Despite the high dropout rate, the cost of college has only increased over time. In 2000, average tuition and fees were $8,082. in 2020, it was $13,677. That's a 69% increase (Bryant 2023). With the increasing cost of college, financial trouble has been the largest cause for students to drop out, with 51% of dropouts citing lack of money as a cause for making the decision to drop out (ThinkImpact, 2021). Some students will delay their graduation either by taking a gap year/semester or by taking only a few classes for a semester and working additional hours or getting a second job to save up money to pay for college/life expenses. While this may be necessary in some situations, delaying graduation increases the likelihood of a student dropping out. 6 years after the starting semester, the dropout rate for students at a 4-year college jumps to 56% (What to Become, 2021). Being in a program for 1.5x the advertised length can be demoralizing so it should be a priority to get students to graduation in 4-5 years. Recognizing students who might be at higher risk of dropping out and reaching out to them to offer extra help might be a good way to reduce dropout rates.

The availability of data on this subject is limited since specific data about the students who do drop out is confidential and finding data that isn't synthesized for educational purposes is difficult. Luckily, I've found a few detailed datasets that can be used to train models. One is very detailed but is small and only has data for students in a Math or Portuguese class so it could be biased (Data source 1). Another focuses on the institution and not so much the students but is very large so it could be used to see if the state and type of institution makes a difference (Data source 2). Finally, the last one is large and has plenty of detail on the socioeconomic factors of the student (Data source 3). This last data set will be used the most in my analysis but I plan to use a combination of all three in my analysis as well as any others I find throughout the course of working on this project.

For the techniques, I plan to use a variety of different classification techniques such as Logistic Regression, Random Forest, and Neural Networks. I plan to lump the predictions into three categories: Finished in the expected time, finished in more than the expected time, and did not finish at the time of data collection. This way the model's prediction can have some nuance without spreading the categories too thin. Since the intended use case is to be used as a tool to identify the students at risk of dropping out and reaching out to them and it would probably be better to reach out to students who don't need it than to miss some students who do need help, intentionally biasing our results to make it more likely to predict a student is at risk could be better for that specific use case even though our models will technically be less accurate.

In addition to training these models, I plan to code an interactive dashboard with different options to change the data being graphed to allow others to explore the data and try to find any trends I missed in my analysis for future research. Normally with prediction models like these, I'd want to add a section where you can input your own variables but I'm hesitant to do that in this

specific case since it would be discouraging to input your own personal values and be told you're likely to drop out though it would help administrators tell if a person could be at risk.

**Resources:**

What To Become. (2021). Everything You Need to Know About the College Dropout Rate. https://whattobecome.com/blog/college-dropout-rate/

Bryant, Jessica. (2023). "Cost of College over Time: BestColleges." *BestColleges.Com*, Best Colleges, www.bestcolleges.com/research/college-costs-over-time/

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

**Data Sources:**

1. Learning, UCI Machine. "Student Alcohol Consumption." Kaggle, 19 Oct. 2016, www.kaggle.com/datasets/uciml/student-alcohol-consumption

2. "College Completion Dataset." Kaggle, www.kaggle.com/datasets/thedevastator/boost-student-success-with-college-completion-da

3. "Predict Students' Dropout and Academic Success." Kaggle, www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention