Data Science and Decision Making:

Forecasting Enrollments for Bucknell Courses

Jake Luther, Class of 2023, Computer Engineering

Professor Sam Gutekunst, Departments of Computer Science and of Mathematics

Proposed 9 Weeks of Research

Research Topic

Before every semester, the Registrar and individual department's at Bucknell University must predict the number of sections of each class that should be offered. For large courses with multiple sections, miscalculations force the schedules of both professors and students to change. Right now, this process is done manually, but some universities have developed models that use previous enrollment data to automate and improve this process [1-3].

Creating a model that can accurately predict the number of course sections to offer in a semester is a complex task. This is particularly true at Bucknell as a result of the small class sizes, changing major requirements, and the liberal arts curriculum. Further complicating the problem is the impact of COVID-19 on the reliability of historical enrollment data, as well as transfer students, and student performance in classes. Solving this problem requires developing and implementing algorithms that use techniques from data science, statistics, and machine learning. Using these techniques, Professor Gutekunst and I will create and test algorithms that the Registrar's Office at Bucknell can use. Broadly, this involves developing and programming algorithms for enrollment prediction, comparing and evaluating these algorithms, and ultimately producing code and recommendations that can be used by the Registrar's Office and Bucknell's data analytics team. We will start with predicting MATH 201 (Calculus 1) enrollments, since there are 15 or more sections a year and it is a required credit for many majors. We will develop an algorithm that helps the registrar better estimate the number of sections to offer each semester.

Research Value and Interest

University course scheduling is an important task for faculty and students. By creating an algorithm that can predict future enrollment of students, the reliability of course schedules can be increased for both students and professors. On a personal level, I have a strong interest in data science so the experience I would gain from this project would be invaluable. Specifically, this project would enable me to develop models and algorithms that may actually be implemented here at Bucknell, while learning new data science techniques. Working on this project in coordination with the Registrar's office and with people on campus such as Ken Flerlage, the Associate Director of Data Analytics, and Tim Kracker, the Associate Registrar, is a tremendous opportunity that I have much to learn from.

Research Approach

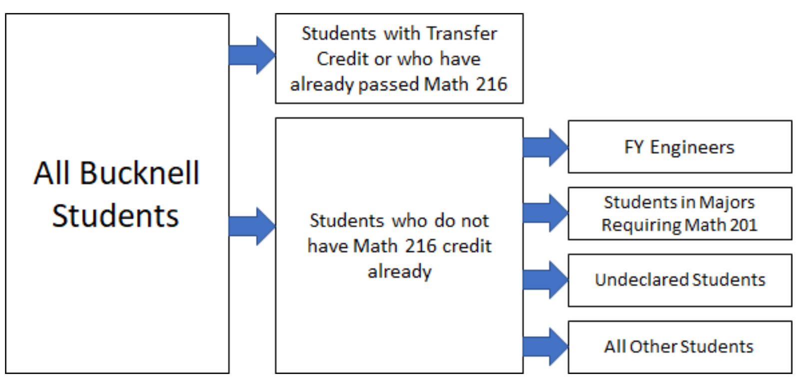

Figure 1: One way of organizing all Bucknell students who do not already have Math 201 credit into groups of students who may take MATH 201 at similar times.

In developing an approach to this problem, Professor Gutekunst and I have come up with several specific goals. First, we aim to focus on designing and testing several algorithms used to predict student enrollment in MATH 201. One such algorithm involves splitting Bucknell students who have not yet earned credit for MATH 201 into different groups, and estimating the proportion of students in each group that will enroll in MATH 201 that upcoming semester. These estimates can be applied to current enrollment data to predict the number of students who will enroll in Math 201, and determine the number of sections to offer in turn. See Figure 1.

These groups would be formed by looking at historical enrollment data specific for MATH 201, and would group students into core audiences for MATH 201. For example, these might include groups such as first-year engineering majors (who all take MATH 201 in the fall, unless they have transfer credit), or other students whose majors require MATH 201 at similar points in time. We will use historical data to experiment with and evaluate different grouping strategies. Then, this algorithm will be benchmarked against general prediction algorithms such as random forests, a machine learning technique. After obtaining an algorithm for student enrollment in MATH 201, our second goal is to extend these algorithms to all calculus classes. This adds another element of complexity to the project, and will incorporate student's performance in MATH 201. Our third goal, if time permits, is to see if our algorithms can extend beyond calculus classes.

While working on the project, we will be programming in Python, a programming language used by Ken Flerlage's team. A key library of Python we will use is Pandas, which enables us to manipulate the large amount of historical enrollment data supplied by the Registrar's office. The data we will start with contains information about past students and when (and if) they took MATH 201. It has already been provided by the registrar's office. The data for each student looks like Table 1, and traces out their trajectory in Math 201. This sample table shows a single student who came into Bucknell in Fall 2015 as an undeclared BA, took and passed MATH 201 in sophomore fall, and then declared a Biology major.

Student[1]	Academic Period	College	Degree	Major	Class	Year	Math 201 Status
12345678	FA 15	A&S	BA	Undeclared	2019	First-year	No Record
12345678	SP 16	A&S	BA	Undeclared	2019	First-year	No Record
12345678	FA 16	A&S	BA	Undeclared	2019	Sophomore	Enrolled
12345678	SP 17	A&S	BA	Biology	2019	Sophomore	Earned Credit
12345678	FA 17	A&S	BA	Biology	2019	Junior	Earned Credit
12345678	SP 18	A&S	BA	Biology	2019	Junior	Earned Credit
12345678	FA 18	A&S	BA	Biology	2019	Senior	Earned Credit
12345678	SP 19	A&S	BA	Biology	2019	Senior	Earned Credit

Table 1: Example data showing a single student's MATH 201 trajectory. Our data has these trajectories for each student at Bucknell.

Data in Table 1 allows us to determine things like the percent of students who come in undeclared and take MATH 201 at any point in their studies, the percent of engineers who took MATH 201 in their freshman fall, etc. Thus we can compare our algorithms by training them on

[1] Note: these ID numbers allow us to keep track of an individual student's trajectory, but are not the same as student ID numbers. All personally identifiable information is masked by the Registrar's office.

data from previous semesters (like Fall 2017-2020) and seeing how well they predict future semesters (like Fall 2021).

This project incorporates many different skills I hope to gain. For one, the project requires me to write code that is easy for other people to understand and use, which will challenge me as a designer. I will also have to develop new algorithms as well as implement existing classification algorithms which are vital skills in data science. Furthermore, it is important to be able to evaluate and compare different algorithms which is another aspect of this project. Finally, I will learn how to make a professional research report and guide for clients to use when running the model.

Research Deliverables

These skills will help me produce three products as a result of my research. The first of which is working, documented, and commented code that is understandable for Ken Flerlage's team. The second product is a technical report analyzing the algorithms used and making recommendations on those algorithms. This report will discuss the different approaches evaluated, the groupings used, and whether the same approach works for non calculus courses. It will also include a literature review and an instruction guide for using the model. Finally I will produce a research poster with the intention of presenting at the Joint Math Meeting AMS - PME Student Poster Session.

Educational Goals and Preparation

My intrigue for data science has grown as I have explored the field and its potential. Presenting my own research at a national conference is something that I have always wanted to do during my undergraduate career, and I am very excited about the opportunity to do so. More generally, this project will enable me to explore research for the first time in a field that I am very passionate about. My goal is to be able to continue research beyond this project and potentially attend graduate school. Thus, one of the things I would appreciate about this experience would be meeting other Emerging Scholars also pursuing research and graduate school.

I am confident that my prior coursework here at Bucknell has prepared me to take on this research project. The software engineering courses I have taken (especially CSCI 205) have given me the base knowledge I need to program the algorithms and models involved with this project. Additionally, my statistics course (MATH 216) will be very helpful in analyzing the enrollment data in the project.

Summary of Prior Research Projects

I have not worked on any research projects before. However, I am currently working on an independent study with Professor Gutekunst focused on exploring the same data that would be used as a part of this project. This has introduced me to some of the core concepts that I will use in the project such as coding in Pandas, working with large datasets, and implementing various techniques such as clustering. Because I will have already gained familiarity with the data and how to manipulate it, I will be much more efficient working with the more complex aspects of the project that I have not yet explored.

Mentoring Relationship

Throughout the summer Professor Gutekunst and I plan on having daily check-ins, as well as formal progress meetings twice a week. I am looking forward to working with Professor Gutekunst on this project, who has taught a number of my courses here at Bucknell. Lastly, I have already been in communication with both Ken Flerlage and Tim Kracker, who will be very helpful in communicating with the Registrar's Office and working with enrollment data.

References

[1] Kraft, Christine R., and James P. Jarvis. An adaptive model for predicting course enrollment. No. 2005-11. Technical Report, 2005.

[2] Lee, Dianne. 2020. A Classy Affair: Modeling Course Enrollment Prediction. Bachelor's thesis, Harvard College.

[3] Bowman, R. Alan. "Student Trajectories for Enrollment Forecasting, Management, and Planning. The AIR Professional File, Spring 2021. Article 153." Association for Institutional Research (2021).