

Capstone Project – Sprint 0

The Problem Area

In my 11 years as a high school teacher, I have witnessed profound changes in British Columbia's education system including curriculum shifts, the removal of most standardized testing, a teacher shortage, and the impact of COVID-19. With these changes, it is necessary to understand the factors that predict whether a student will attend university and how these factors can be affected by schools and school districts.

The first step in this process will be to build a model that accounts for a range of factors (e.g. grades/GPA, socioeconomic status, geographic location, etc.) and classifies a student as being likely or unlikely to attend university.

The User

Broadly speaking, there are three user groups for whom this model will be informative, each of which has particular needs:

- *Students*: People with university degrees generally enjoy higher earning potential and job stability than those without; however, earning a degree often comes with a significant debt burden.
- *K-12 schools*: Each school serves students with a unique mix of cultural backgrounds, gender and socioeconomic statuses, and its geographic location may also influence the number of students who go on to attend university.
- *Universities*: There are financial implications tied to student enrolment, from tuition fees to government funding; furthermore, universities compete with each other for funding, enrolment numbers and prestige.

This project will enable each of these user groups to understand what affects university enrolment and to inform planning for coming academic years. In addition, each user group will have its own benefits:

- *Students*: The insights from this model will empower students and their families with data-driven perspectives, supporting them in making well-informed choices about their educational paths.
- *K-12 schools*: An understanding of the average likelihood of university enrolment will help schools and teachers to tailor their support strategies so that they can better address the unique needs of their students, especially in light of BC's shift from *prescribed learning outcomes* to *core and curricular competencies*.
- *Universities*: By understanding the factors and trends influencing student enrolment, universities can better predict future enrolment patterns and optimize their financial planning.

The Big Idea

The problem statement I intend to address is: "Can we use information about students and factors affecting university enrolment to predict whether a high school student in British Columbia will attend university?"

I believe a logistic regression model will be well-suited to this task as it is straightforward and capable of handling multiple variables. Additionally, I plan to explore decision tree models to account for potential non-linear relationships within the data.

The minimum viable product for this project will be a categorical model that predicts whether a student is 'Likely' or 'Not Likely' to attend university. Ideally, the model will provide a percentage likelihood, offering a more nuanced prediction rather than a simple binary output.

The Impact

My project aims to add significant societal value by providing actionable insights into the factors influencing university enrolment in British Columbia. By enabling schools, students, and universities to make more informed decisions, this project can contribute to improved educational outcomes, potentially increasing university attendance rates. If even a 5% increase in university enrolment were achieved, it could translate into millions of dollars in additional lifetime earnings for graduates, as university degree holders typically earn significantly more over their careers. Additionally, optimizing enrolment forecasting for universities could result in more efficient allocation of resources, potentially saving institutions hundreds of thousands of dollars annually in operational costs. Overall, the project could positively impact both individual financial stability and the economic strength of the region.

The Data

Several key datasets can be utilized to analyze factors influencing a student's likelihood of attending a post-secondary institution in British Columbia. Statistics Canada offers data on household income, parental education levels and demographics, essential for understanding the socioeconomic backgrounds of students. The Canadian Census (the most recent of which was from 2021) provides insights into geographic and ethnic factors that might impact university attendance. The *Longitudinal and International Study of Adults* (LISA) tracks educational attainment and labor market outcomes over time, offering valuable longitudinal data. Additionally, the BC Ministry of Education provides student performance metrics such as GPA and exam results. Together, these datasets offer a comprehensive foundation for building a predictive model on university enrolment in BC.

The Alternative

In addition to my passion for education, I am an avid runner and completed my first marathon in May 2023. The running community was stunned in February 2024 when Kelvin Kiptum, who had set a new world record with a time of 2:00:35 at the Chicago Marathon four months earlier, tragically died in a car accident. This sparked widespread speculation about when – or even if – the 2-hour barrier will ever be broken in an official marathon.

Over the years, the marathon world record has been steadily lowered; for example, it was broken eight times between 2003 and 2022. Given the continuous improvements in running shoe technology, training methods, and nutrition, I am interested in developing a machine-learning model that predicts the likelihood of breaking the 2-hour barrier in a future official marathon. This model would analyze data from thousands of elite marathon performances and incorporate relevant advancements in the sport.