Summary of CSC8631: An Exploratory Data Analysis of Newcastle University's Future Learn Course in Cyber Security

By Ambreen Masud

03 December, 2021

Understanding clients and evaluating business operation is imperative in managing a business effectively and data analytics provides the opportunity for this. Education is a business and therefore must operate as such. Newcastle University has a renowned reputation as a global educator and this is not limited by the walls of the buildings in Newcastle. The university offers courses online to a wider audience, using Future Learn to do so.

This project was commissioned by Newcastle University to provide analytical insights into one of their online Future Learn courses entitled *Cyber Security:Safety at Home, Online, in Life*. In developing the pipeline for the analysis, this report took a CRISP-DM approach as it is iterative and allows for analysis to be considered in more than one cycle. The cycle of CRISP-DM is developing a *business understanding*, then creating a *data understanding* of the properties of the data, cleaning, constructing and formatting the data through *data preparation* and then *modeling* the results through *deployment*. However, these stages are not set and there can be crossover between the stages, for example, data understanding and data preparation work together.

For the purposes of this project, an understanding of the business was imperative in creating and answering the business questions. Recognising that Newcastle University is a business and that the students are their clients allowed for a focus on the clients. This led to the rationale of focusing on the *enrolments* data as it could provide insights into the clientele. Through assessing and understanding the data, it was possible to then go back to the business understanding section and produce business questions. The business questions were as follows:

1. How many students enrolled per run?
2. Where are the majority of students situated?
3. What are the student demographics and who should be targeted for future marketing perspectives?

A number of assumptions was also made at this stage:

- The data will display how many students enrolled on the course and will likely show that more students enrolled as each run progressed.
- The data will show where students are based geographically, which will provide insights into which countries to target in future course promotional activities.
- Analysing the student demographics will provide insights into understanding who to market future course promotional activities to.

After creating the questions, the data was assessed in further detail, particularly in understanding which columns should be used and which ones would not provide sufficient data, which then allowed moving onto the next stage of data preparation. For example, in answering question 2, *"Where are the majority of students situated?"*, the data was assessed in greater detail as two columns could have answered this question: *country*

and *detected_country*. It was decided that *detected_country* would be best as it had less missing values. However, it was presumed that this data was not from students themselves, but from the analytics provided by Future Learn. This then led onto the data preparation stage. Each question was assessed in the same way; understanding the data, preparing and cleaning the data and then producing the outputs before moving onto the next question.

This was actually a very effective method in organising the analysis. However, things did not always run as smoothly. For example, the project may have been in the *deployment* stage when a small error would be noticed and there would be a need to go back to *data preparation* and *modeling* to make edits before outputting the results once again. Throughout this project, version control was used to document the project processes and a copy of the git log file can be found in the files of this project.