Student Number: 210237793
Name: Mayank Baraskar

Short Report on Data Analysis Cyber Security Online Course

Because this is a reflective essay, I'll begin with the question "What?" We were given data from Future Learn, a MOOC (Massive open online course) platform, which included enrolments, leaving responses, quiz responses, archetype survey responses, and more for various intakes in various years. Massive open online courses (MOOCs) have been widely used in the field of education, and have recently been promoted as a result of the COVID-19 epidemic, as I can attest to because our college and university lectures were all delivered online. FutureLearn is an online learning platform that allows students to enrol in and complete any field course. Because the course materials are widely accessible online, students can learn from home or anywhere else. They are not obligated to attend regular classes. On this data set, we were requested to conduct exploratory data analysis. The purpose of this research was to analyse the data offered by FutureLearn in order to get insight into students' online learning engagement and performance. With the help of enrolling, I completed the CRISP-DM approach, leaving survey and quiz response data. The key issue that was discovered was that enrolments were declining at a large rate; hence this data was chosen. Not only that, but the completion percentage of the students that enrolled in this course was quite low. As a result, I chose to conduct a data analysis on two important questions:

- An analysis of when the majority of learners drop out of the course.
- An analysis of why the majority of learners drops out of the course.

We chose data from the year 2018 since it was consistent for runs 5, 6, and 7, all of which took place in 2018. To create the visualisation and complete the study, the data was cleansed and reorganised. The graphs were plotted when needed as part of the data analysis, and assumptions were made that are described in the main paper. The interpretations were stated and sought to fulfil the two-business objectives that were determined via visualisation. Because there were two business objectives, two CRISP-DM cycles were used. We generated a report and a presentation for the Deployment stage. R markdown was used to construct the report. We were able to integrate the code and describe the output using R markdown. Because it was markdown, we had a lot of formatting possibilities, which made it easier to develop a structured format.

Once the data analysis was complete and crisp DM steps were completed, we could come to a conclusion where the highest number of dropouts was seen in week one. It was deduced that the student used week 1 as a trial week and was likely dissatisfied with the course content for a variety of reasons. Those who choose to continue after week 1 had a considerable decrease in dropout in week 2 and a slight increase in week 3. However, because week 3 was the last week of the course, it was presumed that the majority of students had completed it. Also, it was discovered that the majority of students found the course to be too long, although the difficulty of the course was the second most common reason for leaving.

The two major methodologies used in the creation of this report and data analysis project were reproducibility and CRISP-DM. We used ProjectTemplate for reproducibility, which supports good workflow. It enables us to automate project configuration, package loading, data loading, data munging for cleaning, and data analysis. It creates a directory for us to keep our files as well as the settings we'll need to execute our project. It almost standardises setup and loads r script files in the working directory automatically. It is not totally automated, and some user involvement is required, although it is modest. It also helps us to load datafiles from data directory and automate the initial data munging when project is run without requiring us to manually execute each r scripts.

CRISP-DM: It stands for cross-industry data mining methodology. It is a very adaptable, reliable, and well-proven methodology. It's a six-step cycle strategy made up of events with various goals. We can adjust the sequence of events based on the business requirements because it is highly adaptable.

Packages such as dplyr which was used for data cleaning and wrangling which helped us throughout the data understanding and data preparation step of CRISP-DM model. Also, ggplot2 was used for visualizing the graphs and pie charts. There are lot of options in ggplot2 which made it a bit difficult to understand and use it, however, as we use it, we start to realize the power of ggplot2.

It became a lot easier to handle our project with the help of version control. Daily changes and code pushes were made, lowering the chance of data loss. Furthermore, if the prior version of the code was working and the new modifications were not, we were able to discover the problem using a diff checker and comparing the previous version of the code.