# Data Analysis Report of Cyber Security Course

James Terence White

23/11/2020

# 1 Business Understanding

## 1.1 Business Objectives

This report is an investigation into a Cyber Security Course hosted by Future Learn and delivered by Newcastle University. Future Learn is an online platform which has partnered with numerous world leading universities and organizations to delivers a wide range of courses.

Since Future Learn is an educational site, their interests will fall within anything that will enhance learning and increase the student's interaction with the site, and ultimately sign up for more online courses. Thus, we can say a positive outcome from someone taking their course would be to have gain skills or knowledge related to the course they signed up for, and the course was delivered in a way that was stimulating and engaging for the student. Currently, the government are addressing issues, such as "identifying at-risk students" – presented in "From Bricks to Clicks". However, what this report will investigate is "Can we identify a student that require additional support based around their interaction and ability to answer question on a online course". By doing so we can ideally increase the success of the course and reduce the number of students that potentially drop out.

## 1.2 Assess Situation

For this project all data analyses will be partaken on ProjectTemplate, and reports compiled on RMarkDown. Regular Gitlog version control will be used to enhance the reproducibility of the project. The project lifespan is 4 week, and scheduled to be completed by 4^th of December 2020. With regards to the legality of the data, the assumption is that we have full consent from the data owner and is provided by CSC8631 - Data Management and Exploratory Data at Newcastle University.

### 1.2.1 Data Assessment

We are presented a online CyberSecurity course Big Data set over the course that has run 7 times. The data presented is comprised into numerous csv files that can be summarised into the following;

- Survey questions
  - Archetype - relating to to the users psychological traits
  - Weekly sentimental - students feedback on the course
  - Leaving
- Stats
  - Enrollment
  - Step Activity
  - Question Responses
  - Video (>run2)
  - Team members (>run1)

The course content is assumed to be delivered in the form of videos and notes, of which are separated into steps, i.e. chapters to the Cyber Security course. Within these "steps" are sub-sections, and is what is referred to throughout the data set table headings. However, it is important to note that the video data is only present after run two, it is assumed that videos were provided to the student, but the assimilation of this data was not yet available. Furthermore, there is a Team member file available after run one which contains data on any rolls that were allocated within the course, such as mentors, or course organisers.

Cyber Security ran seven time from 2016 to 2018, for a period of three weeks - see details below Tab. @ref(tab:dates).

Table 1: Summary of start and end dates for each run

| Run | StartDate | EndDate |
|---|---|---|
| 1 | 05/09/2016 | 26/09/2016 |
| 2 | 20/03/2017 | 10/04/2017 |
| 3 | 18/09/2017 | 09/10/2017 |
| 4 | 13/11/2017 | 04/12/2017 |
| 5 | 05/02/2018 | 26/02/2018 |
| 6 | 11/06/2018 | 02/07/2018 |
| 7 | 10/09/2018 | 01/10/2018 |

We are provided with numerous csv file and pdf to interrogate, however, due to the limited lifespan of the project only a set few csv files will be interrogates through exploratory data analysis. Namely, the files with the extension 'question.responses'.

## 1.3 Data Mining Goals

Throughout this data mining process the goal is to explore the data set surrounding the posed question of "identifying students that may need additional support" - is there any evidence within the data provided that support this. A successful outcome of this analysis would be to find a correlation.

## 1.4 Project Plan

To achieve these goals an analysis will be partaken on the quiz.responses files, where trends will hopefully present themselves. The stages of the project are as follows.

- The Cyber Security Course data will be analysed this will given an idea of what how we can answer a question that would be of some use to the business
- Once the question is posed, the data will be cleaned, prepped for some sort of analysis
- The cleaned data will then modeled and evaluated, where if anything of interest is presented a deeper anaysis will be diverged into that section
- More data will be cleaned as a supplement answer the question depending on the result obtained.
- Final conclusions on the findings will be presented.

### The Question Asked - Time

Time has a profound effect on our day to day life and it also influences our behavior, be that irrational or not. What I aim to investigate is if timing is correlated to a students ability to answer questions and at what stages of the course are they doing so. Thus, for will be a informal analysis of quiz response timing and if there is a correlation in the timing of these events.

# 2 Data Understannding

# 3 Data Preperation

## 3.1 Select Data

This report is using the Cyber Security Course data set located within the "data" folder of the cyber-security-course project template, however, we will only be analysing the following files; `question.responses` and `archetype` for runs 1:7 of the course. This is because, the question asked is related to the detecting a student that is at risk based upon their interaction with the questions asked. Therefore, it is not essential for this analysis to make used of the the video.stat files, nor the enrollment or files related around the survey questions.

## 3.2 Clean Data

This analysis will consider the the question.responses as well as archetype data. From the analysis of the response.question data there is allot of repeated columns that are essentially stating the same thing, more specifically, referring to the question number answered - the head of this data frame is shown below. Again the long cumbersome heading, making for error prone coding and more difficult representation when plotting data frames. because of this, they were amended. Finally the date was accounted for and put into seconds - a more manageable format. However, since all the course were starting at different time the start date of the course was subtracted - see @ref(tab:dates). All this was cleaned using the `cleanQuizData(quiz, courseStartDate)` which took two arguments; firstly, `cyber.security.X_question.response` data frame and secondly, date `YYYY-MM-DD` that the course started. By doing so, we could standardise the course around the start date. The function was carried out as part of the pre processing in the munging section, where all resultant data frames were stored in the cache. Annoyingly (but necessary), the learner id was incredibly long, but for cross-referencing purposes this was kept. The resultant cleaning from this function resulted in 7 data frames that were in the following format.

Table 2: First two rows of quizStat1 data frame

| id | qq | wn | sn | qn | r | t | ans |
|---|---|---|---|---|---|---|---|
| 77454a73-6b8b-46a2-8dee-35f36b6c4fc1 | 1.7.1 | 1 | 7 | 1 | 1,2 | -5232175 | false |
| 77454a73-6b8b-46a2-8dee-35f36b6c4fc1 | 1.7.1 | 1 | 7 | 1 | 1,2,3 | -5230975 | true |

As seen in the above data frame the data frame names have been made more compact;

- id - the students id
- qq - quiz question
- wn - week number
- sn - section number
- r - responses
- t - time relative to the start of the course
- ans - was the response correct

It may be worth mentioning that the time is negative because they completed the quiz questions prior to the course even starting.

### 3.3 Construct Data

To put the data into a more suitable format, a second function was run as part of the munging. This function constructed the data into a format which we could then start exploring to find relationships. It was determined from the data within the quiz question it was reasonable to put this into the following format.

Table 3: First two rows of quizStatClean1 data frame

| id | numAns | numQues | numCorr | ft | st | dt | acc | scr | tot |
|---|---|---|---|---|---|---|---|---|---|
| 77454a73-6b8b-46a2-8dee-35f36b6c4fc1 | 39 | 17 | 18 | 7831 | -5232175 | 5240006 | 0.4615385 | 0.4358974 | 0.7727273 |
| a4fa6f89-a596-4d00-9397-420a348c398d | 40 | 19 | 19 | 7859 | -4805410 | 4813269 | 0.4750000 | 0.4750000 | 0.8636364 |

It is clear that these number appear to be more user friendly for a data analyst and correspond to the following;

- id - the students id
- numAns - the total number attempts at all questions
- numQues - the total number of questions answered
- numCorr - the number of correct answers provided
- ft - final time a question was answered
- st - first time a question was answered
- dt - the change in time between the first and last question
- acc - ratio of correct to false answers given
- scr - ratio of correct answers against to different questions

All this was compiled in the `quizStatClean()` which simply takes the argument of the result from the constructed data function `cleanQuizData()`. This function was run seven times on each run producing a platform to start modeling data.

### 3.4 Reformatting Data

No reformatting was required of this analysis.

## 4 Modelling

### 4.1 Selecting Modelling Technique

#Evaluation

#Deployment