# 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 4ˆth 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. **??**

Table 1: An example table caption.

| 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 Determine Data Mining Goals

### 1.4 Project Plan

#### 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

This report is using the Cyber Security Course dataset 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.

```
quiz <- head(cyber.security.1_question.response)
quiz
```

```
## # A tibble: 6 x 10
##   learner_id quiz_question question_type week_number step_number question_number
##   <chr>      <chr>         <chr>                <int>       <int>           <int>
## 1 77454a73-~ 1.7.1         MultipleChoi~            1           7               1
## 2 77454a73-~ 1.7.1         MultipleChoi~            1           7               1
## 3 a4fa6f89-~ 1.7.1         MultipleChoi~            1           7               1
## 4 a4fa6f89-~ 1.7.1         MultipleChoi~            1           7               1
```

```
## 5 a4fa6f89-~ 1.7.1         MultipleChoi~           1           7               1
## 6 f27eec8c-~ 1.7.1         MultipleChoi~           1           7               1
## # ... with 4 more variables: response <chr>, cloze_response <lgl>,
## #   submitted_at <chr>, correct <chr>
```

#Data Preperation

#Modeling

#Evaluation

#Deployment