CSC8631 Project: Learning Anaylitcs

Abdullah Turki H Alshadadi, 190582184

Introduction

This report is a data mining report exploring the online course of "Cybersecurity: Safety At Home, Online, and in life" - which was created by Newcastle University published in FutureLearn educational platform.

It will be structured in the method of CRISP-DM (Cross-Industry Standard Process for Data Mining), where it will discuss the Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment from the given data.

Due to following the CRISP-DM, there will be cycles around the previously discussed headings because of better built understanding of the business intentions with the current focused data that needs more insight, or because limitation of data to meet the business needs, or even because of a better understanding on the data that shows significant importance to the business that the business intentions are changed to meet align more to the data.

Cycle 1

Business Understanding

Determining Business Objectives: Background

This online course is about the discipline Cybersecurity made by Newcastle University, a high education provider that seeks to make the course publicly accessible by individuals through the educational online platform FutureLearn.

FutureLearn provides the platform to host the online course with the benefit of providing Newcastle University with learning data of participating individuals on their course.

Using the provided raw data from FutureLearn, Newcastle University will likely want to derive insights on how "successful" is the learning design of the course, did most individuals pass the online course and did it increase over the life time of the course?

Determining Business Objectives: Business Objectives

FutureLearn's raw data of the online course needs insights to be able to quantify measurements of participating individuals engagement on the course for Newcastle University.

Through data analysis, Newcastle University will likely want to know the how "successful" is the engagement of the course, in terms whether the the participants have pass or not. This is important because it will show the effectiveness of their course and determine the best run in the online course.

Determining Business Objectives: Business Success Criteria

The following is the possible success criteria of Newcastle University:

- Give insights on how "successful", is the overall percentage of participants have passed or not, such that the University will be able to view the effectiveness of the course and derive the best run in the online course.

Assessing Situation: Inventory of Resources

The given data is 7 runs of the online course from September 2016 to February 2018, where it is not expected to have more data from the FutureLearn as currently Newcastle University stopped the online course. Therefore, management of resources are much easier and manageable as the only data to be concern with is the provided 7 runs data.

Assessing Situation: Requirements, assumptions, and constraints

For requirements, there is a deadline of submission for this report, limiting the chance of deep data exploration. Thus, the report will focus on the basic vital needs to meet the Newcastle Univesity's success criteria.

For assumptions, the models for the data should be presented in way where it can show trends and variance in the data, so that it will be possible to explain to a novice statistician or user from Newcastle University's online course designers' to understand the success of their course while retaining solid statistical foundation.

Lastly, there is no constraints on the use of the given raw data from FutureLearn, but the only issue is that this is the only data that will be provided.

Assessing Situation: Risks and Contingencies

There is the risk of data not showing results of that convey passed or failed participants, but rather show other factors such as the participants satisfaction or engagement in the sense that they have partaken in the activities of the course. For that, the plan is to adjust to these types of data factors and readjust the Newcastle University's success criteria to reflect better to the data and the business needs.

Furthermore, there is risk of the data to be lost or corrupted, thus a backup is made for the data away from the ProjectTemplate `data` folder just in case if a mistake is made, and everything is stored in the cloud by OneDrive provided by Newcastle University.

Assessing Situation: Terminology

There is no set of business terminologies to be aware of, therefore, there should not be any special terms that contain exclusive meanings in the data by Newcastle University or FutureLearn.

Assessing Situation: Cost and Benefits

There is no cost on this project, in terms of financial cost but instead it is a workflow cost, where this data mining project is heavily limited by time. However, the benefits are that Newcastle University would have a better understanding on the performance of their online course.

Determining Data Mining Goals: Data Mining Goals

Gain insights on the data to determine what constitutes as "success" in the online course by concept description and classify those who have passed or those who have not passed.

Determining Data Mining Goals: Data Mining Success Criteria

Models that can identify the percentage of passed or not passed participants using the "success" data to show and view the change of "success" over the 7 runs of the course to see if it has decreasing over time or not, and what run was that has the most "success". This would likely help to show how effective is the course for Newcastle University over the span of 7 runs and identify the best run in the online course.

Data Understanding

Describing Data: Data Description Report

Before describing the data, there is the need to describe how the online course is structured. The online course duration is 3 weeks, therefore it consists of 3 main sections which are "Exploring personal privacy

online", "Online payment security" and "Security in the future home".

Within each section, there is what the Newcastle University and FutureLearn called "steps", which is the main subsection activities to do in the online course, and it also contains a "step number" that shows the participant the total activities to do per sections.

These step activities are catogrise into ARTICLE, DISCUSSION, EXERCISE, QUIZ, TEST and VIDEO.

For first run, it is structured like so:

- Section 1 "Exploring personal privacy online"
  - Steps 1.1 to 1.18, which has the step number of 18
  - It contains 9 ARTICLE, 2 DISCUSSION, 1 EXCERCISE, 1 QUIZ and 5 VIDEO
- Section 2 "Online payment security"
  - Steps 2.1 to 2.21, which has the step number of 21
  - It contains 13 ARTICLE, 2 DISCUSSION, 1 EXCERCISE, 2 QUIZ and 3 VIDEO
- Section 3 "Security in the future home"
  - Steps 3.1 to 3.21, which has the step number of 21
  - It contains 10 ARTICLE, 4 DISCUSSION, 1 EXCERCISE, 1 QUIZ, 1 TEST and 4 VIDEO

For the second run, there are steps been added to Section 2 and Section 3:

- Section 1 "Exploring personal privacy online"
  - Step 1.2 "Why are you here? DISCUSSION" is the new step, making the steps number to become 19, which means the steps are 1.1 to 1.19
  - It added 1 ARTICLE
- Section 2 "Online payment security"
  - Step 2.11 "Exploring vulnerabilities in online payments VIDEO (05:12)" and step 2.22 "Auditing your Mobile App permissions ARTICLE" are the new step, making the steps number to become 23, which means the steps are 2.1 to 1.22
  - It added 1 ARTICLE and 1 VIDEO

Then for the third run a step has been removed:

- Section 3 "Security in the future home"
  - Step 3.21 "Glossary and references ARTICLE" is the removed step, making the steps number to become 20, which means the steps are 3.1 to 3.21
  - It removed 1 ARTICLE

After understanding the online course terminology, the description of data will be discussed.

For the first run of the online course, the data from FutureLearn is split into 6 data frames:

1. `cyber.security.1_archetype.survey.responses`

2. `cyber.security.1_enrolments`

3. `cyber.security.1_leaving.survey.responses`

4. `cyber.security.1_question.response`

5. `cyber.security.1_step.activity`

6. `cyber.security.1_weekly.sentiment.survey.responses`

Then in the second run of the online course, the data from FutureLearn has expanded to include:

7. `cyber.security.2_team.members`

Lastly, the third run of the online course, the data from FutureLearn has once more expanded to include:

8. `cyber.security.3_video.stats`

1. Archetypes are list of categorical data that describes the behaviour and personality of the participants of the course, and these archetypes are Advancers, Explorers, Fixers, Flourishers, Hobbyists, Preparers and Vitalisers (there is also the options to pick Other).

However, this data set would unlikely be useful as it shows personal character traits which is more appropriate for marketing purpose especially if the business (that is Newcastle University) would like to have more participants to the course it would target archetypes of users - with more supporting data - that has the most engagement (for example, successfully completing the whole course) to market for.

In contrast, from Business Understanding section, Newcastle University would prefer to understand if their online course was simply "successful", in the sense that the participants have passed the online course, on the 7 runs of the course.

2. The data set that keeps track of the participants of the online course. Where the most interesting columns are `learner_id` that provides a unique code for referencing individual participants and `fully_participated_at` that verify that a participant have fully completed the course.

3. Data set that includes participants that have left the course without completing it. It contains data that verifies the time the participants have left, their reasoning, last completed step in the course, last completed week of the course and last completed step number.

This could be useful for another in depth analysis to pinpoint why did the participant not complete the course and what step or step number they were in before they left, to infer need of improvements for the sections that has the most leaving response.

4. This data keeps track of the quiz or test subsections responses from participants in the online course. It shows what step number the quiz or test was that might be useful to link it back to the leaving response to figure out if the quiz and test were the discouraging problem that participants had, or link it back to the step activity that keeps track on the step the participants have done which is useful to determine why some participants have performed better than others.

It also contains a column, `correct`, which shows if the participant have answered the question of a quiz or test correctly or not. This is especially useful to determine the total passed participants in the course to meet Newcastle University's success criteria.

5. For this data, it keeps track of what step activity of the online course is the participant is in, when did the participant started it and when did the participant have finished the step activity.

This could also be useful to keep track of the engagement the participants and can be useful to also link it back to the leave response to determine what was the most discouraging section out of the online course.

6. This data keeps track of the weekly responses on the online course to determine the participants experience for what week section there are in. It contains rating system, called `experience_rating`, and a response, `reason`, of the picked rating system.

7. This data lacks enough information to understand what is trying to represent. An educated guess that it could be for representing the Newcastle University staff who keeps track of the online course and perhaps questions or feedback from the participants. This is because in the column `team_role` it classifies the individuals in the data frame as `host`, `lead_educator`, `educator`, `mentor`, `reviewer` and `facilitator`.

8. This data keeps track of the video statistics. It contains a lot of informative numerical statistics about the videos.

It shows the videos duration in the online course in seconds - this was verified by document image shot provided by Newcastle University of how the online course looked like, for example the first video in the

online course contains the number format of "01:39" which is usually representing "minutes:seconds", and 1 minute and 39 seconds in seconds are in total 99 seconds.

From the other stats, the most interesting are the following

- `total_views`
- `viewed_five_percent`
- `viewed_ten_percent`
- `viewed_twentyfive_percent`
- `viewed_fifty_percent`
- `viewed_seventyfive_percent`
- `viewed_ninetyfive_percent`
- `viewed_onehundred_percent`

This could be useful to gain insight on how engaging was the videos are, however, it is limited because there is no unique `learner_id` to link back to the other data frames such as `cyber.security.1_leaving.survey.responses` to perhaps determine that the video might not been that informative or understandable to the participant to complete (or even not that engaging to watch), or link it to the `cyber.security.1_step.activity` to determine how useful were the videos to be able to pass the quizzes or tests in `cyber.security.1_question.response`.

Verifying Data Quality: Data Quality Report

The data frame that is the most complete and could be most useful to infer an data insight to answer the Newcastle University's success criteria is `cyber.security.1_question.response`. Thus, for now, the focus will be on it.

In this section, it first discusses the quality of the data types compare to the data its representing; second, it will discuss any missing or inconsistent data.

I Data Quality Report - Data Types

Viewing the data types of each column in the data frame, it is divided in like the following:

```
##       learner_id   quiz_question   question_type     week_number     step_number
##      "character"     "character"     "character"       "integer"       "integer"
## question_number        response   cloze_response    submitted_at         correct
##        "integer"     "character"       "logical"     "character"     "character"
```

All of the columns, beside `submitted_at` and `correct`, should be the data type of factor. Factor is categorical (or commonly called in software development, enumerated type), which is more appropriate to these columns because the data is attempting to be represented categorically:

- `learner_id` - unique id to identify individual participants.
- `quiz_question` - the quiz or test section number.
- `question_type` - the type of the question.
- `week_number` - the week of the online course (this will always be 1 up to 3 because that is how long the online course duration takes)
- `step_number` - the step number of the question in the section of the `question_type`.
- `question_number` - the individual question number in the `quiz_question`.
- `response` - the chosen answers for the `question_number`.

The column `submitted_at` is the representing the time the answer was submitted, therefore, the data type of character does not fully captures the data. Converting it to POSIX date time it helps to capture the time series of the data for data modelling and exploration.

For the column `correct`, even though the data type is represented as character, the only 2 sets of the data are "false" or "true", which is better represented as a logical data type.

```
# The Different Data in the Column `correct`
cyber.security.1_question.response %>% count(correct)
```

```
## # A tibble: 2 x 2
##   correct     n
##   <chr>   <int>
## 1 false   34817
## 2 true    42185
```

Lastly, the column `cloze_response` is completely empty. Therefore, it cannot be used and it does not seem to be as important as the rest of the data frame's data.

```
# The Lack of Data in the Column `cloze_response`
cyber.security.1_question.response %>% count(cloze_response)
```

```
## # A tibble: 1 x 2
##   cloze_response     n
##   <lgl>          <int>
## 1 NA             77002
```

II Data Quality Report - Missing or Inconsistent Data

```
sprintf("Total of Empty `learner_id`: %s", nrow(cyber.security.1_question.response %>%
  filter(learner_id == "") %>%
  select(learner_id)))
```

```
## [1] "Total of Empty `learner_id`: 401"
```

In the column of `learner_id` there is missing data but yet it shows that those empty has data in other columns. There is no simple solution to re-populate the data, thus, due to the limited given time for this data mining project, the missing `learner_id` will be just removed.

Data Preparation

Dataset: Dataset Description

. . .

Selecting Data: Rationale for inclusion / exclusion

. . .

Cleaning Data: Data Cleaning Report

. . .

Constructing Data: Derived Attributes

. . .

Constructing Data: Generated Records

. . .

Integrating Data: Merged Data

. . .

Formating Data: Reformatted Data

. . .

Modelling

. . .

Evaluation

. . .

Deployment

. . .