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:

- "How successful was the online course; has it improved over new runs of the 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` as it enables a way to identify participants' understanding of the course by answering questions from quizzes and tests. 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:

Table 1: Data types of the columns from `cyber.security.1_question.response` dataset (excluded other runs, all of which have the same columns, for visualisation purposes)

	Data Types
learner_id	character
quiz_question	character
question_type	character
week_number	integer
step_number	integer
question_number	integer
response	character
cloze_response	logical
submitted_at	character
correct	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 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 will help 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.

Table 2: `correct` column only contains the values "true" and "false"

`correct` column values for Run 1	`correct` column values for Run 2	`correct` column values for Run 3	`correct` column values for Run 4	`correct` column values for Run 5	`correct` column values for Run 6	`correct` column values for Run 7
false	false	false	false	true	true	true
true	true	true	true	false	false	false

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.

Table 3: `cloze_response` column only contains `NA` values (`cloze_response` is abbreviated to `cl_response` for visualisation purposes)

`cl_response` column values for Run 1	`cl_response` column values for Run 2	`cl_response` column values for Run 3	`cl_response` column values for Run 4	`cl_response` column values for Run 5	`cl_response` column values for Run 6	`cl_response` column values for Run 7
NA	NA	NA	NA	NA	NA	NA

II Data Quality Report - Missing or Inconsistent Data

Table 4: Example of Run 1 containing empty `learner_id` values but still have values across the other columns (excludes `question_type`, `cloze_respond` and rest of the rows for visualisation purposes)

learner_id	quiz_question	week_num	step_num	question_num	response	submitted_at	correct
	1.7.1	1	7	1	1,2,3	2016-09-05 10:06:07 UTC	true
	1.7.1	1	7	1	1,3	2016-09-05 22:04:27 UTC	false
	1.7.1	1	7	1	1,2,3	2016-09-06 10:39:21 UTC	true
	1.7.1	1	7	1	1,2,3	2016-09-06 19:20:14 UTC	true

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.

Table 5: Total of empty `learner_id` per run

Total number of empty `learner_id` in Run 1	Total number of empty `learner_id` in Run 2	Total number of empty `learner_id` in Run 3	Total number of empty `learner_id` in Run 4	Total number of empty `learner_id` in Run 5	Total number of empty `learner_id` in Run 6	Total number of empty `learner_id` in Run 7
401	45	15	115	177	52	150

Moreover, `learner_id` have duplicates on the same `quiz_question`. This is due to participants attempting to submit new answers to the same questions. On the other hand, some individual `learner_id` have not completed all of the `quiz_question`.

Table 6: An example of a `learner_id` in Run 1 with duplicates on the same `quiz_question` and not completed all the course `quiz_question` (excluded other columns for visualisation purposes)

learner_id	quiz_question	submitted_at
398a7b88-be48-4b29-9464-f41c7e475bfa	1.7.1	2016-09-05 07:15:57 UTC
398a7b88-be48-4b29-9464-f41c7e475bfa	1.7.1	2016-09-05 07:16:19 UTC
398a7b88-be48-4b29-9464-f41c7e475bfa	1.7.2	2016-09-05 07:16:49 UTC
398a7b88-be48-4b29-9464-f41c7e475bfa	1.7.2	2016-09-05 07:16:53 UTC
398a7b88-be48-4b29-9464-f41c7e475bfa	1.7.3	2016-09-05 07:17:28 UTC
398a7b88-be48-4b29-9464-f41c7e475bfa	1.7.3	2016-09-05 07:17:31 UTC
398a7b88-be48-4b29-9464-f41c7e475bfa	1.7.4	2016-09-05 07:17:58 UTC
398a7b88-be48-4b29-9464-f41c7e475bfa	1.7.5	2016-09-05 07:18:08 UTC
398a7b88-be48-4b29-9464-f41c7e475bfa	1.7.6	2016-09-05 07:18:31 UTC

```
## [1] "All of the `quiz_question` in Run 1"
```

```
##  [1] "1.7.1"  "1.7.2"  "1.7.3"  "1.7.4"  "1.7.5"  "1.7.6"  "3.11.1" "3.11.2"
##  [9] "3.11.3" "2.8.1"  "2.8.2"  "2.8.3"  "2.19.1" "3.18.1" "3.18.2" "3.18.3"
## [17] "3.18.4" "3.18.5" "3.18.6" "3.18.7" "3.18.8" "3.18.9"
```

A solution for the duplicates could be taking the assumption that the first submission of an individual `learner_id` is the honest attempt whereas the others are simply re-attempts to get the right answer. For the individual `learner_id` who have not attempting all of the `quiz_question`, removing those `learner_id` as it is not complete overview of those participants performance, thus cannot be used for assessing the success of the online courses

Data Preparation

Dataset: Dataset Description

The datasets of focus will be on `question_response` for each run, that is `cyber.security.1_question.response` to `cyber.security.7_question.response`. As discuss in the data quality section, these are the most complete and most related to the business' success criteria.

Selecting Data: Rationale for inclusion / exclusion

The datasets all have the same columns, each column has been discussed under the section data quality report, however, the useful columns are the following:

- `learner_id` - to identify the participant for each of the question.
- `quiz_question` - to distinguish each question answered for each student.
- `week_number`, `step_number` and `question_number` - to subset the data to identify section, step number and question number for modelling.
- `correct` - to see if a participant has answered the question correctly or not.

These columns help to calculate a numerical census of the overall percentage of participants answering the question correctly or not. It can also help to find the lowest and highest percentages for questions answered correctly.

The rest of the columns are either redundant or not useful:

- `question_type` contains only one value which is "MultipleChoice".
- `response` are what the participants have choose as the answers, which is not particulary useful as `correct` shows if the answers are correct anyways.
- `cloze_response` contain only the value `NA`
- `submitted_at` is not relevant as it only shows when the participant have submitted the question.

The changes are made under the file `munge/01-A.R`, the datasets are going to be called `run1_qr` to `run1_qr` (`qr` stands for `question_response`).

Cleaning Data: Data Cleaning Report

As discuss under the data quality report there are missing `learner_id` values which still contain values across the other columns. However, due to the time constraints for this data mining report, the data would simply be deleted instead.

Furthermore, there are duplicates and some individual `learner_id` who have not completed all of the `quiz_question`. To fix that, remove duplicates by assuming that the first attempt is the honest attempt and the others are re-attempts to get the right answer, then remove any `learner_id` individuals who have not completed all of course's `quiz_question`.

The changes are made under the files `munge/01-B.R` and `munge/02-A.R`.

Integrating Data: Merged Data

There will be no merging data, as keeping the datasets separate for each run makes it easy to do analysis in the modelling section.

Formating Data: Reformatted Data

After excluding the irrelevant columns, the rest of the columns data types are character types or numerical data types. `learner_id` and `quiz_question` data types will be changed into factor data type as these columns are basically categorical data; for the `correct` as explained in data quality report only contains "false" and "true" thus it should reflect that by changing it to a logical data type.

Table 7: Data types of the columns from `run1_qr` dataset before data reformatting (excluded other runs, all of which have the same columns, for visualisation purposes)

	Data Types
learner_id	character
quiz_question	character
week_number	integer
step_number	integer
question_number	integer
correct	character

The changes are made under the file `munge/02-B.R`.

Data Understanding - after data preprocessing

Exploring Data

Exploring the datasets after data preprocessing has shown that the first run has the widest margin on total size of rows, totaling 29,854. This can impact the accuracy of comparing runs to achieve business' success criteria as the first run has the most data, whereas the other course runs are more scarce compare to it - there more participants completing the course in run 1 rather any other run.

Table 8: Total rows of each run, `Run 1` is by far the largest out of all rows

Course Runs	Total rows of each run
Run 1	29,854
Run 2	550
Run 3	902
Run 4	2,926
Run 5	440
Run 6	484
Run 7	814

Modelling

Selecting Model Technique

Following the business success criteria to represent the retrieved data after data preparation, a model of a barplot representing each run overall percentage of `correct` and `not correct` answers as stack bars can help to easily interpret the runs' overall participants performance on understanding the material and answering

the question corretly. In addition to the barplot, a lineplot is an option to see the line for the percentage of `correct` answers across course runs, helping to answering the success criteria "... has it improved over new runs of the course?"

Building Models: Models and Models Descriptions

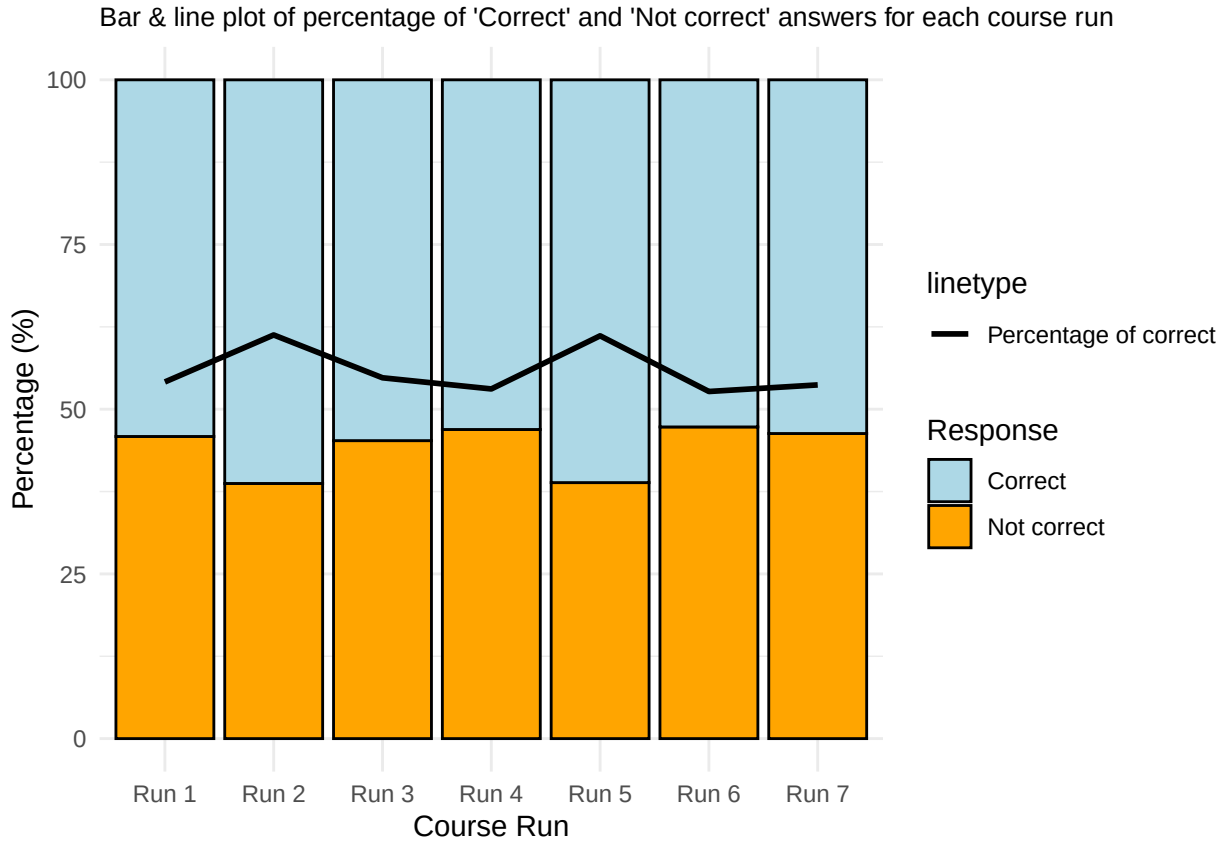

Figure 1: Percentage of 'Correct' and 'Not correct' answers for each course run

Table 9: Overiew of `correct` and `not correct` values in each run

Course Run	Total of not correct	Total of correct	Percentage of not correct	Percentage of correct
Run 1	13692	16162	45.86	54.14
Run 2	213	337	38.73	61.27
Run 3	408	494	45.23	54.77
Run 4	1373	1553	46.92	53.08
Run 5	171	269	38.86	61.14
Run 6	229	255	47.31	52.69
Run 7	377	437	46.31	53.69

Table 10: Quantiles of `correct` answers on the entire course runs

Min Correct	Q1 Correct	Average Correct	Q3 Correct	Max Correct
52.69	53.385	54.14	57.955	61.27

Figure 1 retrieves the overall total of all participants answers in each course run and extracts the percentage of `correct` and `not correct` answers, then makes a stack bar plot. This results in an easier way to interpret course run performance individually.

After that, the line plot on top of the barplot shows the trend of percentage of `correct` answers over the course runs. This helps to see visually if there the trend improving or worsening across course runs.

Table 9 is a summary table that accompanies Figure 1 by showing the exact values that is being represented in the figure, it is also show the number of total of `correct` and `not correct` answers made in each run.

Table 10 is the quantiles for percentage of `correct` answers for all of runs in total. This helps to see the least, quartiles, average and largest values of the percentage of `correct`.

Assessing Models

In Figure 1, the run with highest correct answers is "Run 2" with a 61.27% whereas the lowest is "Run 6" with a 52.69%. For the lineplot the trend seem to fluctuate:

- "Run 1" to "Run 2" there is an increase of +7.13%
- "Run 2" to "Run 4" there is a continuous decrease of -8.19%
- "Run 4" to "Run 5" there is an increase of +8.06%
- "Run 5" to "Run 7" there is a slight continuous decrease of -7.45%

Evaluation

Evaluating Results

Determining the most successful course run from dataset available, "Run 2" has the overall best percentage of `correct` answers with 61.27% and the least successful is "Run 6" with worst percentage of `correct` answers with 52.69%. However, as seen in the exploring data, after data preparation where only the participants that was retrieve are those who completed every `quiz_question` for each course run to maintain an accurate and consistent dataset, it has costed the overall total rows of data where "Run 1" has a total of 29,854 compared to other runs with much lower rows like 440. This effects the reliability of the dataset as the only run with a reasonable size of data is "Run 1" and perhaps "Run 2" with 550.

For finding if the course runs have improved overtime, it is the same idea where the dataset for other course runs are significantly low to reach conclusion on, but from the dataset available it would seem that there is not any improvements nor degradation as the percentage of `correct` fluctuate around the average mean of 54.14%.

When considering this limited dataset, it would seem that the course is slightly successful with an average percentage of `correct` being 54.14% showing that there more participants answering questions correctly than not. But as it is explained earlier, the dataset is limited to reach accurate conclusions.

Determining Next Steps

To disregard this dataset that was made from data preparation is going to be a waste, therefore, it is better solution to combine the course runs into one to determine other insights for the business needs. A possible route is determining for all the participants who have participated on answering all the questions in the courses runs whether the having participants getting more `correct` answer encourages the participants to

buy a certificate from the course or there is no correlation to it. This would help to give the business an insight to whether improve the course runs to help the understanding of the participants to successfully answer the question more `correctly` or rather even encourage more participants to answer all quizzes and test in the course to helps sales on course certificates.

Cycle 2

Business Understanding

Determining Business Objectives: Business Objectives

Following the first cycle, a better use of the data was to combine the course runs into one to determine better insights. One solution, as mentioned in determining next steps section, is to find correlation if the participants answering quizzes and tests more correctly would encourage the participants to buy course certificates.

FutureLearn is a company that might seek more revenue through selling certificates for the course, therefore, correlation on this can let the company to focus more helping students understand the course better and encouraging the participants to engage on answering quizzes and questions.

Determining Business Objectives: Business Success Criteria

- "Is there a correlation where if participants are answering quizzes and tests more correctly, it would likely encourage the participants to buy a course certificates?"

Data Understanding

...

Data Preparation

...

Modelling

...

Evaluation

...

Deployment

...