Assessment Report

Jack Westmoreland

2024-11-01

Introduction

Newcastle University have completed seven runs of Cyber Security: Safety At Home, Online, and in Life, A MOOC (massive open online course) teaching cyber security concepts to the public. This report aims to explore the data collected from these runs to be able to provide important analytics and valuable insights into online learning environments and further optimise online teaching. To achieve this, two cycles of CRISP-DM will be completed hopefully providing valuable insight into the provided data set.

CRISP-DM Cycle 1

Business Understanding

This section of CRISP-DM entails defining a problem we would like to solve for our data set, then setting the tasks and success criteria, which must be completed for the problem to be solved.

Objective

With the prevalence of online learning increasing in recent years (Cambridge Home School Online 2024), being able to predict learning outcomes of students from the data we can collect could prove very useful for online educators. Being able to predict learner outcome from their interactions with online courses could allow for educators to "check up" on students who have poor predicted performance; providing further guidance and support, increasing learner outcomes. Therefore being able to find some predictors of learner outcomes would be beneficial for Newcastle University's online education programs.

Success Criteria

For this EDA to succeed a strong predictor of learning outcomes should be identified from the provided data set. This predictor must be measurable for students before the completion of the course to allow for intervention. Furthermore, ideally the data should be measurable for each student individually so that each students predicted performance can be personalized to them. In a simple sentence the goal of this CRISP-DM cycle can be:

"Can we predict individual student performance from their data?"

Data Understanding

In this phase of CRISP-DM I will evaluate the data we have been given, considering its usefulness in completing the task set about from the Business Understanding step. The data will have its reliability considered from what we know about how it was collected. A close analysis will also be conducted to check what data is available, considering how it could be used to perform a successful CRISP-DM cycle.

Data Collection

The FutureLearn MOOC data set has been provided by Newcastle University. It consists of several CSV files containing data on student performance and interaction with the online material for each of the seven runs. Each run consists of 5+ CSVs each containing data on ways students have interacted with the program.

Each run has near identical data being recorded such as x_enrollments, where x is the run of the program. These data sets exists separately for each 7 runs and contains specific student enrollment data such as their gender and the date they enrolled on the course . However, earlier runs do not have some of the data that was collected later on. All this data has some relevance to my goal of predicting student performance from data collected about their interactions. Therefore, each of these data sets will be considered for exploratory analysis later in this report.

Data Exploration

Using R we can see some simple information about each of our CSV files (which have been loaded as data frames using ProjectTemplate) Below is a table containing each types of data recorded for the seven runs, alongside a short description and what runs the data was recorded for.

Data set	Short Description	Recorded for Runs
archetype.survey.responses	Survey results which place each student into one of 8 categories of learning "archetypes".	3,4,5,6,7
enrollments	Enrollment data for each student on the course.	1,2,3,4,5,6,7
leaving.survey.responses	Survey responses kept from a questionnaire given to students who decided to leave the course.	4,5,6,7
question.response	Saved responses for each student for any quizzes they have completed throughout the course.	1,2,3,4,5,6,7
step.activity	The start and completion date and time for each student for each step in the program.	1,2,3,4,5,6,7
weekly.sentiment.survey.responses	Responses to a weekly survey containing a quantitative 1-3 rating and qualitative general feedback.	5,6,7
team.members	little information can be extracted from this, likely has something to do with team building exercises.	2,3,4,5,6,7
video.stats	Data on how students as a whole interacted with videos. Such as how long each video was watched, what devices on, etc.	3,4,5,6,7

From further examination, although some csv files exist they actually don't contain any data. A check has been created to look for this and any empty data frames are not present in the above table's runs, although they technically exist. Furthermore although many data sets from different runs follow the same name conventions I have checked weather they actually contain the same types of data. This was done by ensuring that all data frame variables shared the same column names. This check is important as later I will likely be merging table rows from different runs to expand on the data I can use to see what are good markers of student performance.

```
## [1] "All data frames ( video ) have the same columns."
```

```
## [1] "All data frames ( team ) have the same columns."
## [1] "All data frames ( sentiment ) have the same columns."
## [1] "All data frames ( activity ) have the same columns."
## [1] "The data frames ( question ) do not all have the same columns."
## [1] "All data frames ( leaving ) have the same columns."
## [1] "All data frames ( enrolments ) have the same columns."
## [1] "The data frames ( archetype ) do not all have the same columns."
```

From this check we can see that all the 8 collections of data from each runs share the same column names. This gives confidence in the health of the data set and that they can be merged later on. Furthermore, from manual inspection, using R's view() function, it is clear that some of the rows for these data frames contain missing values. This is especially true for survey response data where not all students have taken the time to respond.

Unfortunately, the provided data set does not appear to contain any grades for each of the students. Therefore to actually gauge student performance another measure must be used. For this quiz data will be used from the x_question.response data frames. This data frame contains the following attributes:

- **learner_id:** id of the student partaking in the quiz.
- **quiz_question:** id of the question being answered.
- **question_type:** categorical data on the type of question being asked.
- **week number:** the week of the program the quiz is from.
- **step_number:** the step of the program the quiz is from.
- **question_number:** the location of the question in the quiz.
- **response:** the student's response to the question.
- **cloze_response:** N/A columns, no data.
- **submitted_at:** data the answer was submitted.
- **correct:** weather the response was correct.

One possible predictor of student performance could be the "learning archetype" that student falls under. It would be a safe bet to guess that certain archetypes of students would perform differently on the program compared to others. For example, "Advancers" who are highly self-motivated and ambitious may perform better more hands off online learning than other archetypes; therefore needing less support to perform well on the course. The data recording each students archetype is as follows:

- **id:** id of the survey response.
- **learner_id:** id of the student partaking in the quiz.
- **responded_at:** date the student responded to the survey.
- **archetype:** categorical data of the archetype that student falls under.

Data Preparation

Onto the next step in this CRISP-DM cycle I began work on data preparation. This step involves transforming the raw data set into something more useful. This involves shaping the data, renaming variables, and merging tables to get access to the important data I have identified during data understanding. This is vital as data preparation will allow for the later modelling step of this cycle to be done far more easier.

Data Transformations

Before ensuring high data quality I first transformed the data so that all runs of the program will have their data frames combined. This was done to ensure that there is a large enough data set to analyse so that I can

ensure that outliers in the data set do not "throw off" my results when modelling. For archetype prediction testing only rune 3 - 7 will be combined as there is no data for this in the first 2 runs of the program. This decision was also applied to the question.response data so that we can later match students IDs in this set to their archetype. The scripts to achieve this `01-Combine-questions` and `02-Combine archetypes` can both be found in this projects munge file. These merges result in two data frames being created:

- `merged_question.response`
- `merged_archetype_survey.response`

Using R I checked for duplicate archetype survey responses which are present from students filling out the survey multiple times. These duplicated have been removed in the R script `03-Remove-duplicate-archetypes`. I have only kept the most recent survey responses in the case of duplicates, this was decided as it would keep the most recent data on the type of learner that student is.

To associate each student with an their learner archetype we can then merge these two data frames. This column merge, done in `04-Merge-question-archetype`, adds each students respective archetype to the question data frame, using their **learner_id** as the key for the transformation. This was done with an inner join so that only students with documented archetypes were present in the resulting table.The resulting data frame can then be used for modelling to check for any relationships between student performance and their learning archetype.

After creating the `question_archetype_df` data frame some datatype conversion was needed for later pre-processing steps to work. Students had weather they answered each question correctly or not saved along the 'correct' column of the `question.responses` data frames. This was originally recorded as a character type, being either "true" or "false", and needed to be converted to Boolean values for a grading value to be derived from. This was completed in the `05-Correct-boolean-conversion` R script.

Next, to obtain some sort of grade for each student from the combined `question_archetype_df` I have created a new data frame `student_grades` in `06-Student-grades`. This data frame, which is derived from `question_archetype_df`, contains each student and their archetype alongside a new correct_percent column which is the % of questions that student answered correctly across the course.

Modeling

Now that the data has been suitably prepared it can be modeled. This step of the CRISP-DM process involves interacting with the prepared data set to extract information that we can use to answer the research question outlined in business understanding. During this cycle this will involve summarizing and plotting interesting relationships present in the `student_grades` data frame, this will be done using both dplyr and ggplot2 libraries.

Archetype Distributions

Since the archetypes variable in `student_grades` is categorical it would follow that counting the amount of instances of each archetype is present for further analysis. Having a larger amount of each student archetype in my data set would allow for stronger claims to be made from my modelling, as the chance of outliers significantly effecting analysis would diminish greatly as the number of observations in the sample data increases. To count how many of each archetype group is present, dplyr's `group_by()` and `n()` functions can be utilized. This data can then be plotted as a column chart for easy comparison amongst each archetype.

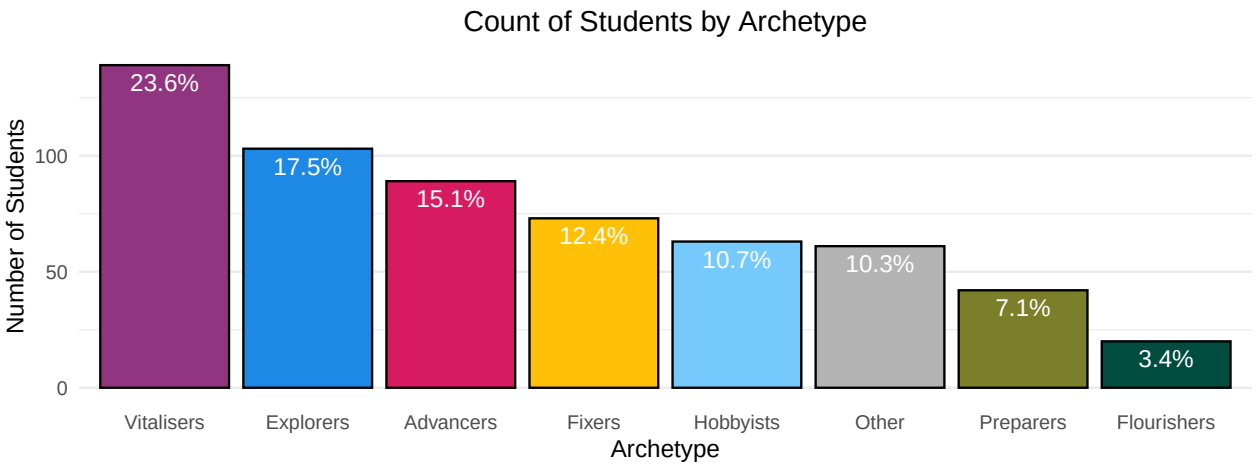

Figure 1: Columns Count Chart for Student Archetypes

From the column chart produced above in Fig 1 we can see the distribution of how many of each student archetype is in the data set. Notably it is clear that there is not an even distribution of these archetypes in my data. Instead groups such as vitalizes are highly represented by 23.6% of my data, whilst archetypes like flourishers and preparers are seen far less. This will have to be considered when making predictions about these groups, as their will be less data to support any claims. Furthermore around 10.3% of students did not belong to any of the pre-defined archetypes. If this scales similarly to the wider population of students they will not be able to have their performance predicted with this measure.

Archetype Score Distributions

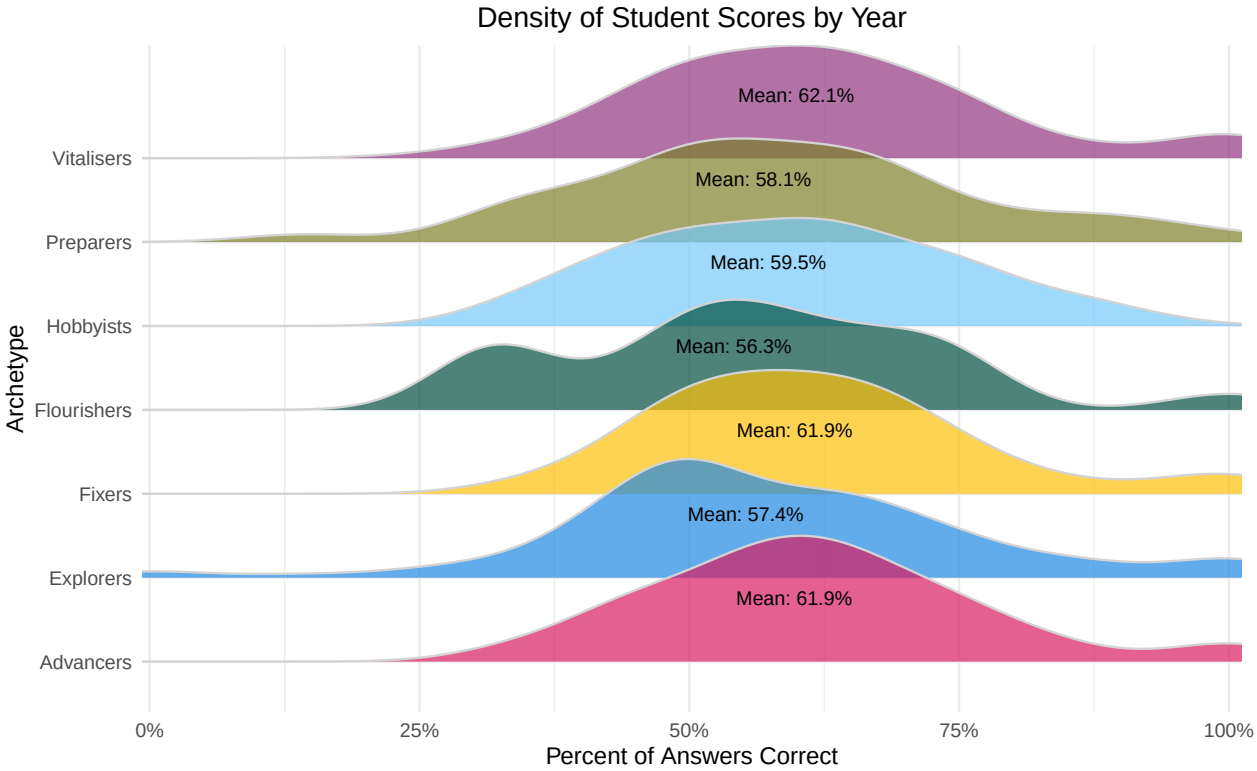

Figure 2: Density Plot of Archetype Scores and their Means

To see if learner archetypes can be a good predictor of a students overall performance I ridge plot has been created using ggridges. This plot shows the density of score distributions for each archetype, whilst separating each archetype on a new line and different colour to allow for easy comparison between groups. As I am only interested in the 8 leaner archetypes students who did not fall into one of these, and instead received "other", have been removed from the plot.

From Fig 2 we can see how each of the eight archetypes are performing on the program quizzes. I have also added the mean correct answer percents to each archetype to allow for further comparison between groups. From the plot we can see that there are some noticeable differences in performance across difference archetypes. For example vitalisers and advancers seem to be performing the best on these quizzes, having the highest means and more right-sided distributions. However, when examining the flourishers archetype (which is the lowest scores) care needs to be used as this archetype has the smallest sample set. This is apparent when looking at the bumps in their distribution, which are more pronounced than others, signifying that outliers are having a dramatic effect on the group.

From the distribution graph we can get some key takeaways. First, most groups average around a 60% success rate in answering questions, with there being little delineation away from this between groups. We can also see that some groups are less spread than others, for example advances tend to group around the distribution mean when compared to explorers. We can also see the amount of people from each group perfectly answering all their quizzes and therefore having a 100% success rate. Although, this looks great it likely signifies some level of cheating from these students, which does differ slightly between archetype groups.

Archetype Step Heat Map

As we can see no large differences between student performance and their archetypes so far It may be worth investigating deeper into our questions to see if there are any differences being missed due to just grouping all each student's quiz answers together. It could be that different archetypes are performing differently depending on the stage of the course, one archetype could be better at content relating to online payments than others. Therefore analyzing how each group performs on questions for each stage of the program could reveal some insights. To allow for heat map to be made the `question_archetype_df` data frame has been grouped by archetypes and step_number. A percent_correct column has then been made storing how each of the archetypes has performed for each step. Throughout the seven runs the program material (which can be seen in docs) has not changed so therefore all data could be combined for this.

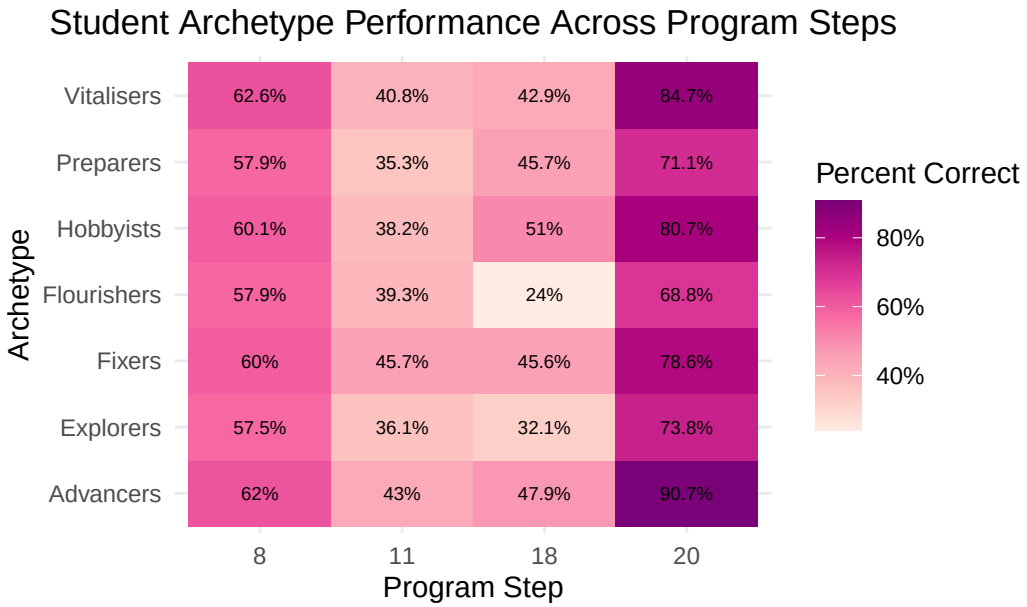

Figure 3: Archetype Performance Across Program Steps Heat Map

From the heat map shows in figure 3 we can how each archetype has performed over each step in the program. Firstly we can see a large difference in performance for all students between steps 8 & 20 vs steps 11 & 18. Perhaps indicating that the quizzes in the middle of the program are easier. We can also see that there appears to be some noticeable differences in performance between archetypes for specific steps. For example advancers are largely outperforming other archetypes during step 20. This relationship does not appear to be as simple as advancers always outperforming other archetypes as in earlier analysis this difference is not as apparent. Furthermore, we can now see that for each step a different archetype is outperforming others. Step 11 has fixers outperforming others and vitalisers are performing best during step 8, this was not apparent in earlier analysis.

Evaluation

Now that analysis on my first CRISP-DM cycle has concluded I can move one to the evaluation. After business understanding and data understanding I believe the modelling step of this cycle has lead to some interesting insights into the data set. Firstly the data seems to be relatively healthy, the provided CSVs across each run are all fully compatible and there doesn't appear to be a large amount of missing data. The main issue found during this cycle was multiple responses to the archetype survey, however this was quickly remedied by only keeping each student's most recent submission.

Some interesting data on the amount of each archetype present in the data set has also been found. With specifically vitalisers being a very common archetype for students partaking in this online course. Focusing more on the initial research question the distribution graph from figure two appears to show little difference between archetype performance for the program overall. However, after producing a heat map of archetype performance across each step we can see that performance is not as similar as previously thought. It seems that archetype performance does change significantly depending on the step of the program student's are being quizzed on.

This leads to the conclusion that archetype performance could be a decent indicator of student performance when taking into account the step the student is in and not just looking at overall performance. Therefore, I believe that this first cycle of CRISP-DM has provided reasonable evidence that student archetype could be a good indicator of student performance. However, the differences in performance are too small for intervention of one specific archetype to be useful as they tend to differ by only around 10%. It may be the case that multiple predictors for each student would need to be used to predict how they will perform on the course.

CRISP-DM Cycle 2

Business Understanding

Now that my first cycle of CRISP-DM has been completed I will now begin work on cycle 2. This cycle will be used to investigate other aspects of the data set which would be relevant to shareholders. After evaluating cycle 1 I believe some fairly interesting relationships have been found. Mainly that archetype, although having an affect, is not that strong of a predictor of student performance. This is likely due to student performance being a complex variable with a lot of student data effecting it. Perhaps making this task too hard to complete for now. Therefore for my second cycle i will be shifting my data analysis goals.

This cycle will be used to analyze how student performance and archetype effect weather a participant is likely to purchase a certificate at the end of the program. This analysis could help shareholders make business decisions about running future MOOC programs, and come up with ways they can increase the number of certificate purchases. If more learners were willing to purchase certificates at the end of the program this could provide valuable funding for the program. Therefore, the question I will be proposing for this cycle is:

"What makes students likely to purchase a certificate?"

Data Understanding

For this analysis similar data from the fist cycle will be used, with the addition of `x_enrolments` data. This data frame contains enrollment information on every student who participated in the program. Crucially it records pruchased_statement_at which records the date each student has purchased a certificate at. This column is left blank when no purchase was made, making it simple to see weather each student ended up purchasing a certificate after they were done with the program. Since I am still interested in if archetype data effects purchase probability, data from runs 3-7 will be used so that students can be tied to their archetype.

Data Preparation

For analysis to occur on student purchase chance a new data frame must be created which records each learner, their archetype, percent_correct and weather they purchased a certificate from the program. The first step for this was to merge enrollment data from runs 3-7, which has been done in the `07-Combine-enrolments` R script. interestingly this lead to the merged_enrolments data frame having multiple instances of the the same learner_ids, this has likely happened from the same learners repeating the course throughout it's runs. This data was kept so that learners who repeated the course and purchased certificates at different runs would have their data retained for analysis.

This was then merged with the previously created `student_grades` data frame to create student_grades_certificates, in the `8-Create-student-certificate` R script. The resulting data frame stores grade, student id and archetype data as well as their certificate purchase date. The decision was made to merge these tables using an inner join, so that the students in the merged frame have all the necessary data for modelling. Finally so that it was easy to determine weather a student made a certificate purchase or not a new Boolean column (certificate_purchased) was added using dplyr's mutate() function.

Modelling

I will now begin the modelling stage of my second CRISP-DM cycle. The modelling performed in this section is done with the aim of exposing relationships between the variables found in the `student_certificate_data` data frame and certificate purchases. With the goal of discorvering what makes learners more likely to purchase certificates at the end of their course.

Initial Exploration

Table 2: Certificate purchase count

Certificate Purchased	Count	Percent of total
FALSE	623	95.7
TRUE	28	4.3

As we can see from the above table students as a whole are quite unlikely to purchase a certificate from the program, with less than 95.7% of students not making a purchase. Whilst this will partly be due to students dropping out of the course below completion this is still quite a low figure. Furthermore, this will make finding correlations for purchasing certificates harder as there will be a smaller sample set to make determinations about.

References

Cambridge Home School Online. 2024. "The Rise of UK Online Learning: Trends and Statistics." https://www.chsonline.org.uk/blog/the-rise-of-uk-online-learning-trends-and-statistics.