6_3 Final report

Alexander J Malt

28/11/2019

This report outlines:

- the business understanding;
- how the exploratory data analysis (**EDA**) was performed (method, process and tools);
- the results of the EDA;
- the evaluation of the results of the EDA;
- the cost/benefits of the EDA;
- the conclusions for the business; and
- the conclusions for further data mining.

Summary

The university provided a set of data from a series of 'runs' of an online cyber security course. This project is an exploratory data analysis - there were no set business objectives or success criteria given at the outset. I analysed unenrollment data to see whether I could identify an underlying cause of students leaving the course (and whether this occured at specific points in time). The attempt was unsuccessful: the data does not seem to reveal any clear underlying causes, and the quantity of data available for analysis significantly lessens as we focus on particular areas. However, it did reveal that many students remained enrolled on the course for a while despite completing *none* of the content. On this basis I recommend the University considers a means to prompt students to complete 'bite sized chunks' of the course more regularly (splitting longer content up if necessary), thereby inducing action and making it easy to find time throughout the day to study. I also recommend the university improve the quality and quantity of data available, both by incentivising students to complete input more data and by extracting more precise information from the sources of the data provided to facilitate more detailed analysis.

Business understanding

This project is exploratory data analysis (**EDA**): the University gave no clear questions or hypotheses at the outset, i.e. there were no business objectives and no correlated success criteria.

Instead, the University provided data from multiple 'runs' of an online cyber-security course. The objective of the EDA is to examine the data, see whether any avenues of investigation suggest themselves, perform an anaylsis and report the results.

The University provided data from the course relating to the following:

- User activity;
- User personality archetypes;
- User enrollment;
- Leaving survey responses;
- Assessment data;
- Team members; and
- Video statistics.

EDA: Method, process and tools

Tools: The EDA project was conducted with the R programming language using R Studio. The following additional packages were also used:

- ProjectTemplate
- plyr
- lubridate
- ggplot2
- rmarkdown
- tinytex

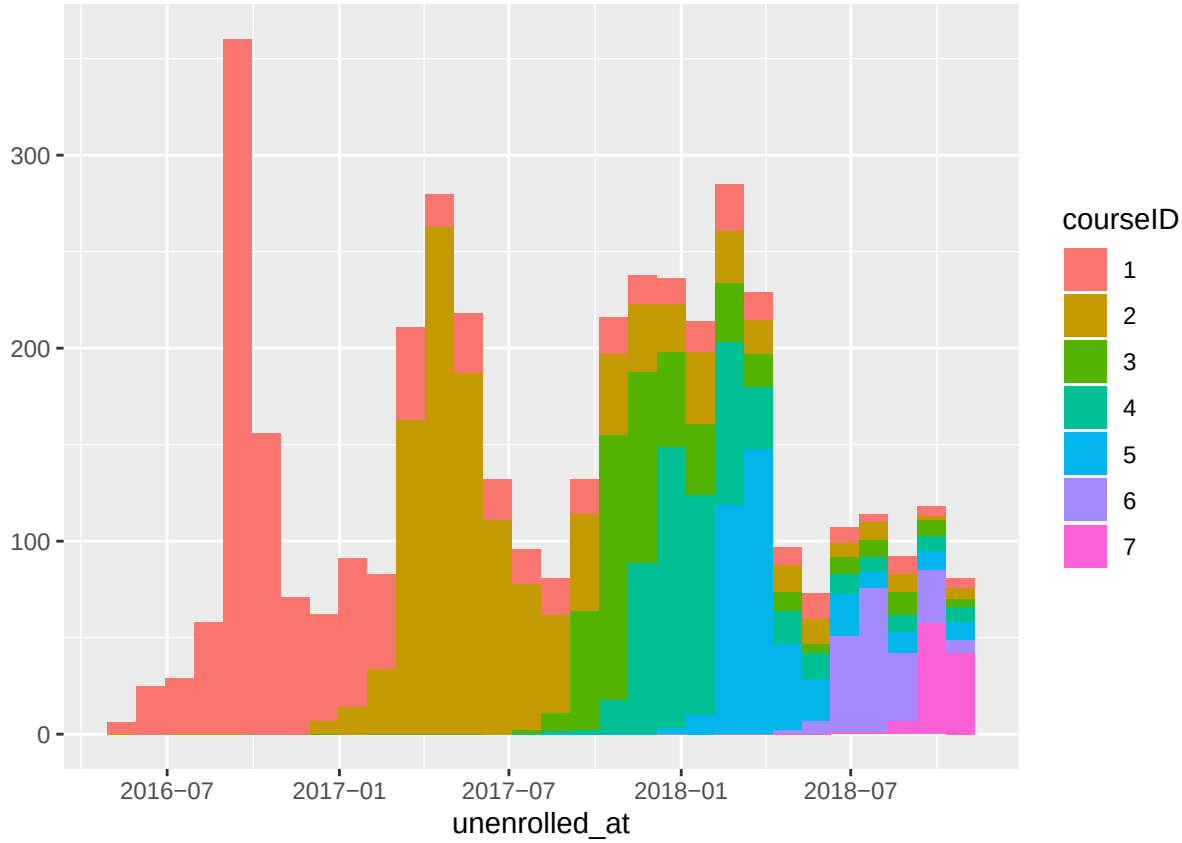
Method/process: The method used in this EDA project is an implementation of the Cross Industry Standard Process for Data Mining (**CRISP-DM**).

1. **Business Understanding:** reading the supplied materials and listening to an initial outline of the data by stakeholders from the University.
2. **Data Understanding:** Gaining an initial familiarity with the structure and content of the data; identifying which files shared common structure; determining and assuming the meaning of parituclar fields in the data; early explorations using basic plots (charts and graphs) to see what avenues of investigation suggested themselves.
3. **Data Preparation:** Merging files relating to the the same topics, i.e. all files containing data relating to the same area (e.g. all enrollment data) were merged into single 'frame' to facilitate later analysis; reformatting the data (e.g. ensuring that dates are formatted correctly) to facilitate later analysis; creating subsets of the data for latter analysis.
    - Prior to being merged, each row of each individual data file was assigned a number to ensure it could be traced to a particular run of the course, e.g. '3' would denote that the row is from the third run of the course. This enabled comparison of data across course runs.
4. **Detailed Investigation:** creating more advanced graphs and plots in order to analyse particular aspects of the data in more detail. This phase resulted in a series of plots 'slicing' the data in numerous ways (see below for the plot). *N.B. In CRISP-DM this is referred to as the 'Modelling' phase. However no formal models were created in this project. To avoid confusion I refer to this phase as Detailed Investigation.*
5. **Evaluation:** assessing the validity and utility of the results for the University, and making recommendations relating to (i) the course itself and (ii) improving data captured on the course.
6. **Presentation:** A presentation of the results was given using slides generated via R and the additional libraries. This report presents the same analysis and recommendations using the same text. *N.B. In CRISP-DM this is referred to as the 'Deployment' phase. Although this analysis aims to be reproducible, there is no model to deploy. To avoid confusion I therefore refer to this phase as 'Presentation'.*

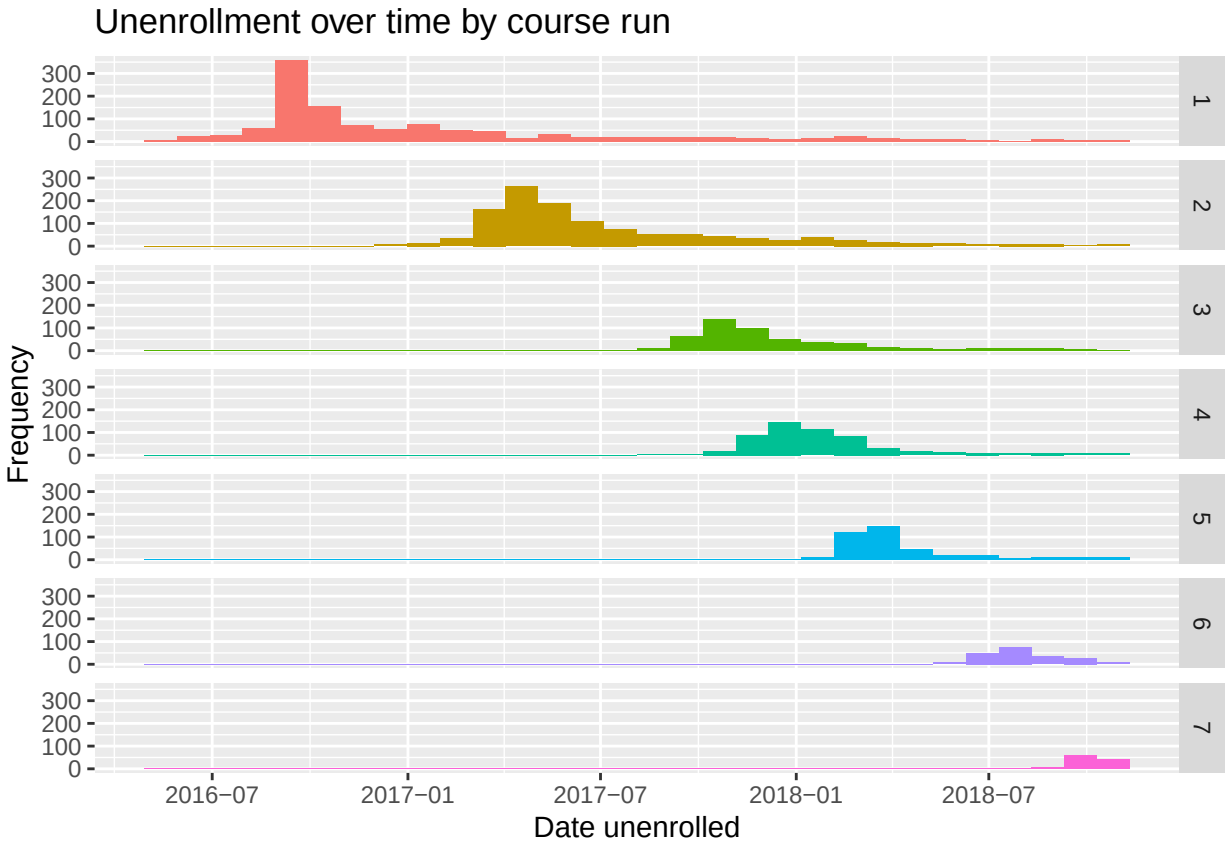
N.B. Steps 2-5 in the CRISP-DM process are iterative. The actual investigation involved many cycles and included (i) multiple 'dead-ends' during investigation and (ii) more detailed examination into particular patterns/areas. This report presents a final 'polished view' of the results and recommendations; it does not describe each cycle in detail.

EDA results

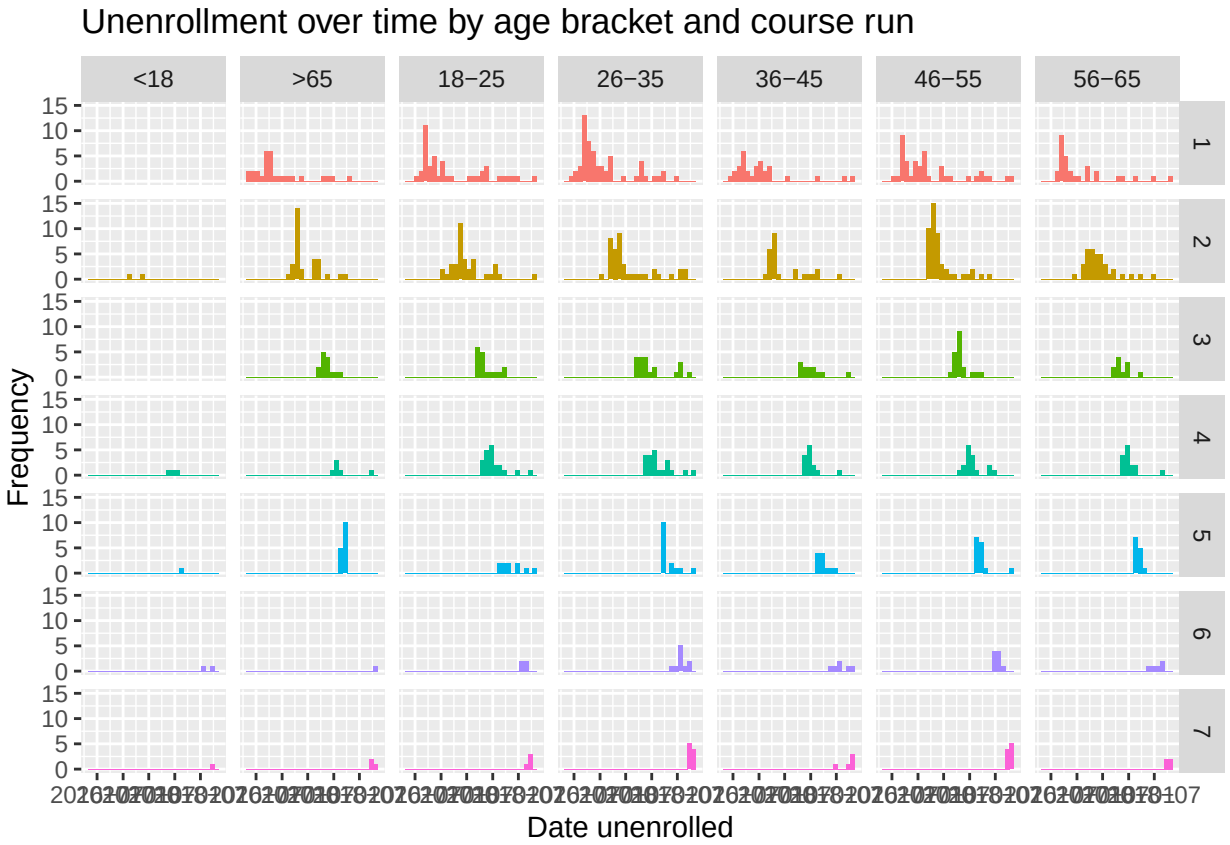
During initial Data Understanding I performed a basic plot of data concerning when students unenrolled from the course:



The pattern that suggested itself in the above plot is that there are a large number of unenrollments relatively close to the beginning of the start of a course run, with less unenrollments as time goes on. To see whether this was true I produced a more sophisticated plot, breaking the dataset down by course run:



I therefore decided to investigate the cause of the high unenrollment rate near the start of each course run. I first looked to see whether there was a visible difference in this pattern when the data was sliced according to other attibutes, e.g. age, gender, employment status, personality archetype, etc. However, none of these attributes seemed to be the main cause of the unenrollment rate. See, for example the plots showing the data sliced be age range and gender (code for other plots showing the data sliced by other attributes can be found in the relevant source code):

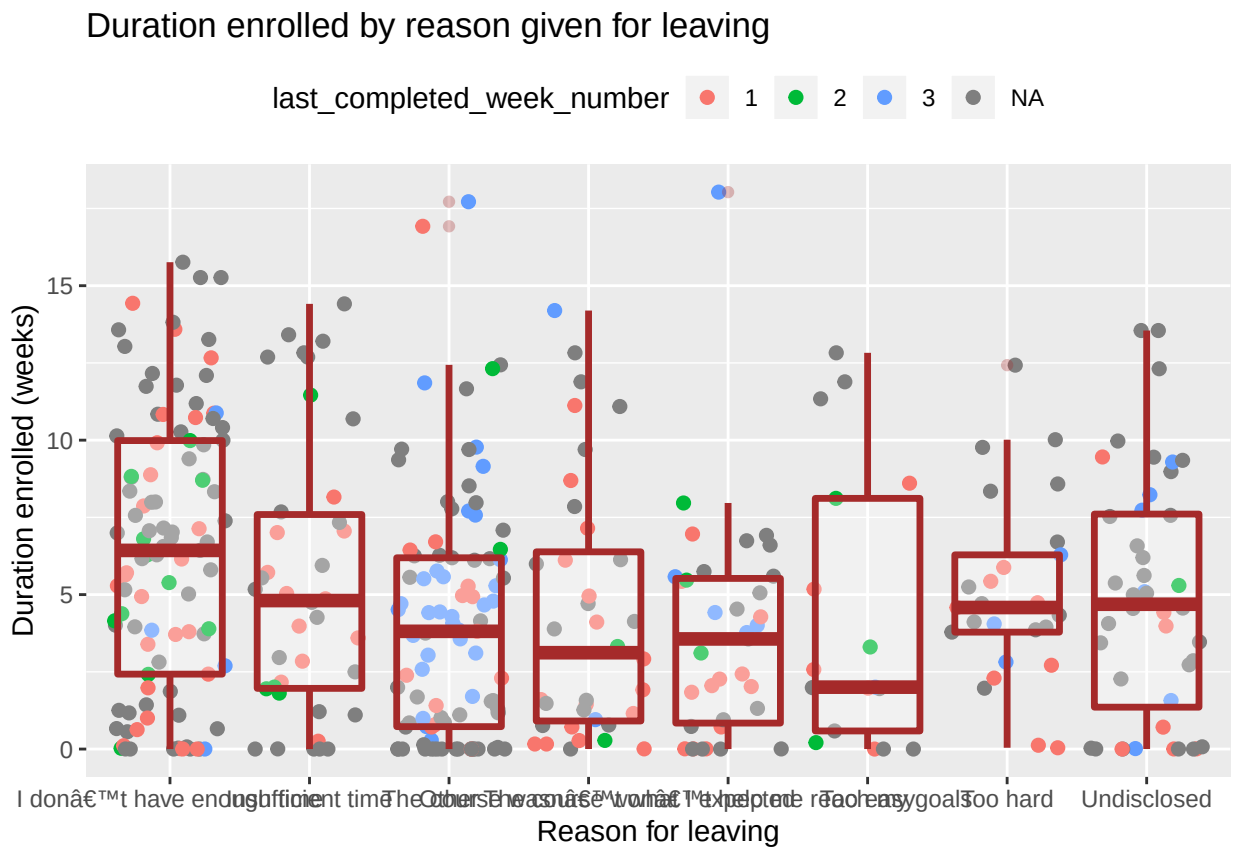

Unenrollment over time by age bracket and course run
<18
>65
18–25
26–35
36–45
46–55
56–65
1
2
3
4
5
6
7
Frequency
0
5
10
15
Date unenrolled

Without finding, e.g. a particular age range, gender, profession, etc., that could account for this pattern, I decided investigate whether there was an aspect of the course that was causing people to enroll, e.g. a particular lecture or test that people found too difficult or required too much time to become familiar with/proficiant at.

To this end, I calculated the average duration that these students were enrolled on the course (in weeks) and plotted this information broken down by course with an boxplot overlay showing the median and inter-quartile range:

However, this plot does not show, e.g. a consistent median enrollment duration. One explanation for this is could be that the course is delivered online and that sudents are able study whenever they please (which would also account for the long enrollment durations).

I therefore merged/joined (a) the data concerning unenrollment and durations enrolled with (b) results from a leaving questionnaire. I wished to see whether there was any pattern in the reasons given for unenrolling from the course (e.g. the subject matter being too complicated, the student having insufficient time to do the work, etc). For this new plot, I kept the duration on the y axis and continued to use a box plot to display the median and interquartile range. However, I clustered the points by reason for leaving and coloured the points by the student's 'last completed week' (i.e. if the student had completed all parts of the course scheduled for the first week of the course run, their last week would be week 1 - however, this does not mean that they did this in one week or that they unenrolled after a single week).



This did not reveal a particular reason for leaving correlated wth a particular point in time. However, given the large number of data points showing students who did not complete a single week's content - despite, in some cases, being enrolled for over 15 weeks and, in other cases, unenrolling immediately (not starting at all) - the plot does imply that we lack sufficient data to determine why students are unenrolling from the course.

However, in case a lot of the unenrollments happened following a particular *step* in the course, I produced another plot to show the frequency of unenrollments following a particular step. This plot shows that the overwhelming majority of unenrollments occured without a single step being completed:

Evaluation of results

The analysis unfortunately yielded little of practical value (though see recommendations in 'Conclusions for the business', below).

A principle weakness is that, as the analyses become deeper and more specific, the amount of available data goes down. For example, not everyone who unenrolled from the course also completed a leaving questionaire or a personality test. Therefore there are only very limited subsets of data available to explore these potential connections.

Costs/benefits of the EDA

The analysis yielded little benefit.

By contrast the cost was extensive (estimated at 12 days continuous work; a further 3 days being required to upskill to the required technical level).

Conclusions for the business

There are two recommendations that can be made on the basis of this report:

1. **Additional data capture:** Mechanisms to incentivise or make mandatory further data input would increase the amount of data available and enable us to draw out any patterns that might be present. This could be acheived by, e.g. offering some money back upon completion of a leaving questionnaire or making it mandatory to complete a personality test at a certain step of the course.

2. **Daily prompts:** The amount of people who unenroll after a significant amount of time but have not completed a single step of the course could imply that they are signing up then never finding the time to log into the course. Given this, it could be worth considering giving users prompts to complete parts of the course in bite-sized chunks (perhaps accomodating this by splitting up longer content, or producing content that is short duration but can easily be slotted into a busy day, e.g. a commute).

Conclusions for further data mining

As the analysis proceeded I encountered issues relating to data quality. For example, video statistics data is broken down by video, not by student. This means that it could not be merged/joined with the enrollment data to determine, e.g. whether the length of time watched of videos correlates with the propensity to unenrol from the course. It also means that we cannot see, e.g. whether video are watched repeatedly by a few people or few times by many people. Before engagin in further data mining it would be worth looking at how the quality of the data could be increased via a close examination of the systems in which this data originates.