Student no. 170339779

Short Report

Business Understanding	We analysed what we knew of the data and confirmed objectives with FutureLearn over email (Matthew Forshaw). After understanding the business objectives, we continued to work on defining engagement as the opposite of disengagement and thus retention time and full participation are the natural metrics to measure engagement. After understanding these engagement metrics we decided to use the latest two runs of the course to have the most relevant and new information whilst maintaining a large enough sample size.
Data Understanding	We analysed the data and identified potential data quality issues such as the step numbers not a being string values and instead being float values making step 1.1 and 1.10 values the same in this column of the data provided. We instead used week number and step number (separate columns) to determine the real step number and stored this as a string. We understood that lots of students did not have a fully participated at date nor a unenrolment date and so we developed the na_pass column which considers all students that did not have a fully participated at data as not having fully participated in the course.
Data Preparation	We prepared the dates necessary as numeric values as a distance from 1/1/1970 in seconds and then days when necessary. We prepared the last step completed variable for each student along with dates of last completed. We then calculated retention time by subtracting the enrolment numeric date from the a last completed step date. Additionally, we prepared video statistic data (percentage watching at percentage of video completion), data from the question answers dataset (accuracy in answered questions and how many questions answered) and country data for each student (based on detected country). We constructed the uni_ids dataframe to hold the majority of this data. We also collected repeat student ids as biproduct which were then later used to develop an understanding of what occurred with those students. We also collected feedback from unenrolment questions for single attempt and repeat students.
Modelling/ Deployment/ Evaluation	We created graphs and information outputs at for all the data collected including a logistic regression to demonstrate the difference in probability of purchasing a certificate if fully participating or retained for 21 or more days. We suggest improvements to the FutureLearn course throughout based on what the data appears to be suggesting and suggest the need for further investigation when the outcome of our analysis is limited. We also comment on the usefulness of the route we investigated for example we comment that the question data and analysis was not broadly useful for our purposes.

Reflection

This experience has taught me two things, in principle, which were broadly derived from my inexperience with ProjectTemplate and the CRISP-DM method. The first thing that I have learnt is

that segmentation of code leads to an easier review process. To elaborate, when running bug fixing or editing my code it was often necessary to run the whole pre-processing and graphing stages to check on my changes to see if I was successful. However, had I segmented my code from the start not only would it have been easier to find specific processes, but it would have also made it easier to run small segments of code. You can see me realise this as I sporadically flip to sourcing 2 graphs from external code in the graphs folder. The second thing that I learnt from this process is to have a clear line of investigation at the start of the project. Whilst I did have a more general idea of what I wanted to achieve at the start of this process I believe it would have been more in line with the CRISP-DM method to have ended the business understanding stage with a better idea of what I wanted to investigate. This may be slightly more difficult to observe in the report itself as it was necessary to "back write" a lot of the narrative so that it made a more cohesive story.

I had tried not to assume much of the data and note when assumptions are necessary in the report such as detected country not actually being a sure way of determining country of origin for students. I have assumed that run 6 and run 7 were the most relevant and disregarded much of the data as a result it may have been beneficial to investigate the excluded data in more detail. I have also assumed that 21 days was a good cut-off point for the logistic regression. I also decided on an arbitrary number of students (n=25 or more) to be the cut-off for inclusion in the last two figures of my report. Finally, we could have shown the result of our logistic regression where retention time in days was not converted to a Boolean with a cut-off of 21 and instead explored how additional days affect purchasing probability. In general, I think the ProjectTemplate format is a really good way to maintain reproducibility across projects. Although I feel comfortable with it know it will most likely take me more time to understand all that it has to offer. To conclude my project suffers from me "learning on the job" throughout as I got to grips with ProjecTemplate and CRISP-DM but the experience will likely make my next project better than this one. Following CRISP-DM rigorously hinders some flexibility and makes it harder to consider previously unconsidered routes until the end of the cycle.