CSC8631: Critical Reflection

Morgan Frodsham

03/12/2021

Critical Reflection on CSC8631 Coursework: FutureLearn_Assessment

Introduction

CSC8631 is a module focused on data management and exploratory data analysis. This coursework assessment is designed to give experience of building a data analytics pipeline to process, query and gain insights into data sets from Newcastle University's FutureLearn course called "Cyber Security: Safety at Home, Online, in Life". The purpose of this short document is to reflect on my experience of the tools and techniques introduced on CSC8631 in completing the coursework assessment. It includes a summary of the analysis undertaken, a reflection on the technical elements of the coursework, and a reflection on CRISP-DM as a methodology.

Analysis

The exploratory analysis undertaken on this project aims to better understand the learning effectiveness of Newcastle University's Future Learn Cyber Security course. I focused on the concept of learning effectiveness after researching FutureLearn as a business, which positioned the concept as an advantage for potential learners. After some further research into learning effectiveness, I realised that my analysis should explore at least one of the following concepts: learner engagement; learner success rate; or learner completion time.

My exploration began by examining at the data sets provided by Newcastle University; there were 53 for seven runs of the course. I spent time looking at the different data field and conducted some basic visual analysis of the data using PowerBi. With the aforementioned scope in mind, I decided to explore the relationship between learner engagement and learner success rate in the enrolments spreadsheets. I chose the enrolments data sets because they seem to contain the most optional demographic data for learners to share with FutureLearn. I was curious whether asking for learners' demographic data was akin to making sure everyone speakings at the beginning of a meeting by introducing themselves - a simple method to begin creating an inclusive environment. This thought allowed me to hone in on the question: "Do learners have a higher chance of success the more data they share with FutureLearn?"

To address this question, my exploratory data analysis focused on nine questions:

1. What is the gross success rate of learners?
2. How many learners provide any demographic data?
3. What is the success rate of learners who provide demographic data?
4. What is the success rate of learners who do not provide demographic data?
5. Is there any correlation between learners' success and the provision of demographic data?

6. Which genders declared by learners have a higher chance of success?
7. Which ages declared by learners have a higher chance of success?
8. Which levels of education declared by the learners have a higher chance of success?
9. Which employment statuses declared by learners have a higher chance of success?

These questions arose in iterations as I learnt more R, became more comfortable with RStudio, and felt more confident that I could conduct the analysis necessary. The questions address learners success rate and learners sharing demographic data as a proxy for learner engagement. I think the findings of my exploratory analysis outlined in the other report do produce useful insights into the relationship between learner engagement and learner success rate, and I'm disappointed that I do not have more time to explore these insights further. There is a clear correlation between these indicators of learning effectiveness, and it does appear that learners who engage by providing demographic data are more likely to succeed.

Technical

The technical tools and techniques were definitely the most challenging element for me; I am fundamentally impressed with what I have achieved through this coursework. This section is structured by the technique suite I have used:

- R. I have a substantial amount to learn about this programming language, especially as it is the first time I've done a significant amount of coding. I like that R comes across as applied mathematics or statistics and, once I remember what different R terms mean more readily, I'll find it easier to be able to solve issues that arise. A huge thank you to those who have patiently answered my very simply questions as what I read online still feels like a foreign language.
- RStduio. After learning how to actually load a data set into RStudio and do something to it (with the help of a very kind CSC8631 demonstrator), I have come to appreciate that RStudio is somewhat insinuative although I do wish that the errors offered more explanation.
- RMarkdown for R. RMarkdown is a rather unforgiving tool (as most R packages seem to be), but it was invaluable in being able to do 'literate programming' and document the exploratory data analysis conducted throughout this coursework. I'm looking forward to learning more about it.
- ProjectTemplate for R. I found this R package really useful once I had this set up, had overcome my multitude of working directory issues (I did not realise how many times one needs to reset it), and learnt what all the different folders meant. It felt like a tool that forces you to go slow in order to be able to go faster later down the line. It really helped me improve the reproducibility of my work, and I don't think I would have achieved any reproducibility without ProjectTemplate.
- Git. I found Git to be the easiest technical element of this coursework, and I think that's because it was so thoroughly explained in the labs with lots of opportunity for me to ask questions. I found it particularly useful that I regularly pushed my work to GitHub as halfway through the coursework my laptop died. I can see it's real utility as a version control system, and look forward to exploring this.

Methodology

I found CRISP-DM to be very similar to Agile methodologies, which I appreciate and enjoy. I was, however, really struck by how much CRISP-DM came across as a technologist's methodology. For example, I have restructured the "Business Understanding" section in my report as I didn't find their order of "Assess Situation" section particularly logical. I also appreciate why the "Data Preparation" comes after the "Data Understanding", especially for building large-scale data models, but I really wanted to move between both those sections with more iteration. Overall, I see the benefits of CRISP-DM as a methodology.