Cyber Security Course Data Analysis

Introduction

Data Analysis lifecycle can be baffling at times so, to manage it efficiently several tools and techniques can be used such as, ProjectTemplate for managing the code and reproducibility, Git for version control, and CRISP-DM for managing the data analysis life cycle. How these tools and techniques helped in our analysis will be discussed in this report.

CRISP-DM

CRoss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (***CRISP-DM***) is a project lifecycle management framework that helps us to keep check on each stage of the data analysis project. It is consist of mainly consist of 6 steps mentioned below:

1. Business understanding – What does the business need?
2. Data understanding – What data do we have / need? Is it clean?
3. Data preparation – How do we organize the data for modeling?
4. Modeling – What modeling techniques should we apply?
5. Evaluation – Which model best meets the business objectives?
6. Deployment – How do stakeholders access the results?

ProjectTemplate

ProjectTemplate is a framework which allows us to enhance the reproducibility of the code and makes life easier to manage the code base.

GIT

GIT is a version control system. It assists us to manage different versions and progress in the project. If we want to go back to some specific version we can do that too. It also helps in collaborating with other developers.

R Markdown

It allows us to generate reports right from the analysis environment which reduces the overhead of making separate reports or documentations.

Project Summary

Data was provided by FutureLearn for the Cyber Security Course and the objective was to target the right audience and improve the course. For the analysis we used the data into different structures such as keeping all files separately as it was provided by FutureLearn to compare variation in enrollments over different iterations of the course and we also merged the data from the same genre file, for example, merged enrollment data row wise from all iterations.

We started with analysing whether the course is doing well or not and realized the number of enrollments are decreasing with increase in iteration of the course.

Target Right Audience

To find the right audience for the course we begin our analysis by figuring out from which continent most learners are enrolling then we focused on specific countries. Both analyses supported each other and we found out most learners are coming from Europe and are based in the UK. Then we tried to figure out what educational qualifications they have, what's their employment status and area, their age and gender. This suggested that learners are from all age ranges and they are working full time. People from the IT and Educational sector seem to be attracted to this course a lot. So, advertisements can be tailored as per above mentioned audience.

Course Improvement

For this objective we started our analysis by looking into what type of device most learners are preferring and we figured out desktop is most widely used but there is an upward trend in usage of mobile devices. So, optimizing the course for both desktop and mobile phone is important which will improve the learning experience of the learners.

We also tried figuring out what learners feel about the course, whether they had positive or negative sentiments towards the course. We found out that in general sentiments were positive, so we decided to dig deeper and tried figuring out if their experience with the course was good, what was the reason for leaving the course and we realized most people had time constraints. So the course can either be divided into multiple courses or maybe assessments can be made easier or the number of questions can be reduced.

Furthermore, we analysed the video stats data to figure out what can be the optimal length of the video, this suggested that a video of length around 5 mins can be the optimal duration but we can not say this for certain because importance of the video content can also be a factor.

Merits of CRISP-DM

- Each step is clearly laid out which makes it easier to follow and keep track of all the stages in the project.

Demerits of CRISP-DM

- It is a very old framework and with the advancements in the data analytics industry workflow has changed a lot.

Assumptions and Design Decisions

- All the analyses are done by removing “unknown” values and hence, this may or may not represent the true population distribution so, we carried our similar analysis on different sets of data and checked whether they support each other or not. For most part similar analysis supported each other.
- We are considering “detected_country” rather than “country” from enrollment data because most of the fields in the country column are “unknown”.