Cyber Security Course Data Analysis Summary

SATISH PILLA-B210472095

5/30/2022

Introduction

Data management is an administrative process that includes acquiring, validating, storing, protecting and processing required data to ensure the accessibility, reliability, and timeliness of the data for its users. Exploratory data analysis is an approach to analyzing data sets to summarize their main characters, often with visual methods. The Data Analysis life cycle can be complex at times, thus numerous tools and approaches can be used to manage it effectively, such as ProjectTemplate for code management and repeatability, Git for version control, and CRISP-DM for data management.

CRoss Industry Standard Process for Data Mining (CRISP-DM)

Throughout the course work we have learned various data science process models. The main process model we used in our course work is CRISP-DM Methodology. It consists primarily of the six steps listed below:

1. Business understandingWhat does the business require in terms of understanding?
2. Data understanding - What data do we have and what data do we require? Is it sanitary?
3. Data preparation — How will the data be organised for modelling?
4. Modeling — How should we approach different modelling technqiue?
5. Evaluation – Which model best matches the company's goals?
6. Deployment - How will stakeholders be able to see the results?

GIT and GitHub

Git and Github was used throughout the project to get backup of the work or changes made in the code to be saved, which is an industry standard.

ProjectTemplate and R markdown

The entire project was managed using a project template, which was useful for organising data, library files, and data pre-processing utilizing the munge folder and src folder. The experience gained from completing this project has given me the confidence and knowledge to handle tasks with massive data files and Big data.

R Markdown: It enables us to generate reports directly from the analysis environment, reducing the time and effort required to create separate reports or documentation. The project is built on R markdown numerical and graphical summaries. This project has provided me expertise with packages like ggplot for plot generation and dplyr for data transformation and cleaning. R-markdown has proven to be useful in producing PDF documents in a timely manner.

Analysis Summary

The main objective of the project is to find out some meaningful insight from the cyber-security dataset, so that we can help in developing good business model for next course run for the futureLearn company. FutureLearn provides data for the Cyber Security Course, with the goal of targeting the correct audience and improving the course. We used the data in various structures for the analysis, such as keeping all files separate as provided by FutureLearn to compare variation in enrollments across different iterations of the course, and we also merged data from the same genre file, for example, combining enrollment data row by row from all iterations.

Analysis on First Assumption: First, we plotted a heat map to see from which area of the world the highest number of students are enrolling. The bulk of students are from the United Kingdom, India, and the United States, with a few from Australia, Saudi Arabia, Nigeria, Mexico, and Russia rounding out the list. Then we plotted the top 10 nations with more than 500 enrollments, and we discovered that the United Kingdom enrolled over 100,000 students, followed by India and the United States with 3538 and 2117 students, respectively. As a result of the given information, the course provider may choose to focus on certain nations in order to attract more students and develop a solid business model there.

Analysis on Second Assumption: The second analysis is used to evaluate the course. First, we determine how students or learners access the course. We discovered that more than 75% of students use a desktop to access the course. After that, we determine which delivery mode is preferred by learners, and we discover that either video or articles are the most preferred methods. The creator of the course content can concentrate on making the video more engaging for students.

Conclusion

The course material is good, according to the multiple assessments done in this study. The course provider may choose to focus on the nations with the largest enrollments in order to improve the course. According to the data, people of all ages, from the IT and education industries, and working full-time are among the people who the course provider may target as the ideal audience. The course can be separated into many courses or the quiz questions can be reduced so that serious learners can take the assessments whenever they want and casual learners can read the content and learn from the course without having to spend a lot of time on the exam. Finally, the course material producer can concentrate on making the videos and articles more engaging.

Personal Reflection

I had a great time working on this project and learned a lot. This module provided me with a great deal of knowledge. The CRISP-DM methodology for data management projects, as well as the project template and ggplot, were very useful and fascinating. The purpose of this project was to develop a set of analyses that might be used to improve student engagement and achievement by testing and proving concepts using FutureLearn data.

Future Scope

While the research conducted for this paper yielded some encouraging outcomes, there is still need for more research to be done. Furthermore, if the enrollment file had not been so sparse, the entire study of un-enrollments in particular would have portrayed the most precise intuitions. Another proposed option is to create a dash board using R shiny that displays details of student activity based on real-time feed data, allowing module course module leaders to better understand the students' actions, perhaps lowering enrollments and increasing student engagement with the course.