Cyber Security Course: Future Learn
Exploratory Data Analysis

-(Himika Mishra 220068080)

Introduction:

One of the tedious data-handling tasks is exploratory data analysis. When handling both unstructured and structured data, it frequently results in clumsy solutions, making it difficult to reflect on one's own work and leaving the analyst feeling lost. As a result, some solutions have been introduced to help organize the process and manage this issue. Project template, CRISP DM, R Markdown are a few of the options. This report discusses these solutions because they were utilized in our project.

Project Template:

A project template makes it simple to duplicate functions and reports by organizing and automating a significant portion of the code. The code for the pre-processing is saved in the munge dataset, the codes for the exploratory data analysis are defined by SRC, and the reports are kept in the reports folder.

Crisp DM:

A waterfall-based approach to data mining is the industry standard. Which consists of six major steps:
1. Understanding the business: What the business's needs are and how the business model operates.
2. Understanding the data: How each file is changing over time and how many are present.
3. Information readiness: Which of our relevant data needs to be transformed for modeling, and how?
4. Modelling: Which models are ideal for the data, and how can future values be predicted using them?
5. Evaluation: Does our data confirm everything the model predicts?
6. Deployment: That will be accessible to all customers and business stakeholders when it is implemented.

R markdown:

It offers a one-step method for storing the r code and the data that must be added to the report for explanation. When knit, it produces reports in the formats of html and pdf. It eases the burden of documenting.

Project summary:

We had to find a solution that was effective from a business perspective for the future learn dataset that was the subject of the project. As a result, researching MOOC courses was a bit of a necessity to learn about the obstacles they face. Understanding the data was crucial then because only with this information can we determine the path forward. There were seven iterations in all, and the future learn dataset contained six files for each one. I noticed that there was a lot of data missing from some columns, so the data had to be optimized for that. As a result, it provides useful insights. I selected enrollments, question response, and step activity as my three data sets. These were very insightful, so I prepared them in accordance with clear DM and proceeded to the data preparation phase, selecting the two to three columns from each of the three data files that contained the most values individually. After that, I began exploratory data analysis and presented every insightful solution I could find. They were aligned with the business perspective, and then they were evaluated against all odds. Because all of this data was consistently committed to git, version control was maintained. The project was constructed using a project template to maintain modularity. There were two cycles that were taken into account. So, during the second cycle, I tried to see things from a different angle and came up with a list of things that could be added to the EDA.

Evaluation:

Because the data set provided enhances the thought process, I thought that project was quite provocative. because there is so much information to interpret in it. There are many options, but we had to limit ourselves due to time constraints.

Bring out emotions:

The journey was very emotional because this was the first time I did the second cycle to think about my work from a new angle and reflect on it. I just thought that the waterfall model of the process as a whole was quite old and that new changes could be made to it. Even the software development cycle has changed to a spiral model these days, and devops features are being thought about. Therefore, that would be extremely helpful.

Review in the light of previous experience:

Because I had previously worked with git while I was employed by Accenture, it brought back some fond memories. In addition, although I have gained practical experience with exploratory data analysis over the past two modules, I have not yet learned how to restrict one's thoughts in the face of so many choices.

Identify lessons learned:

Smartification of the mind is crucial. We need to weigh our options and consider the relationship between modularity and reproducibility. How the reflective process helps you think more clearly.

Follow up actions:

My classmates and I discussed my business ideas and sought their guidance on how these things will add up. Then, to get an expert opinion, I asked my professors to consider this from the perspective of a business stakeholder and determine whether they would like it. Lucky for me, they had insightful thoughts about my business concepts, which inspired me to consider how they could assist me in undergoing cycle 2 analysis.

Feedback:

The project template helped me become more modular, and git version control has always been a great practice. I also plan to use these in the future. Regarding CRISP DM, I thoroughly enjoyed working on it, but I would prefer to adopt a novel data mining method if it were available.

Shortcomings:

1. The considered dataset is biased due to the large number of unknown values.
2. Since the country that was provided had a lot of empty values, the detected country was used instead of the country that was given. If the detection algorithm was fed some VPN data, the data might show false results.