Cyber security course exploratory data analysis

-(Parth Sagar 210431117)

Introduction

Exploratory data analysis is one of the tedious tasks to carry around with the data. While handling unstructured and structured data it often leads to clumsy solutions, where reflecting one's own work sometimes seem difficult and the one who is doing analysis feels lost. So, for the management of this hassle some solutions are introduced which helps organise the process. Some of the solutions are project template, CRISP DM and Git for version control. These solutions have been used in our project and hence are discussed in this report.

Project Template

Project template provides a way to organise and automate significant part of the code so that functions and reports can be easily be reproduced. The pre-processing code is saved in munge dataset, SRC defines all codes of Exploratory data analysis and the reports are kept in reports folder.

Crisp DM

Cross Industry standard process for data mining, it is a waterfall-based approach. Which includes 6 major steps:

1. Business Understanding: How the business model is running and what are the requirements of the business.
2. Data understanding: what all files are present and how they are changing with respect to time.
3. Data preparation: what is our relevant data and how it needs to be transformed for modelling
4. Modelling: what all models would be a great fit for the data and how can be the future values be predicted with it.
5. Evaluation: does what all the model is predicting is true with our data.
6. Deployment: when it is deployed, all the customers and business stakeholders can access that.

Git

It is used for version control to maintain all the different versions made during the project and store the progress of the project too. So, we can move freely towards our previous versions of the project too and can share that with other developers.

R markdown

It provides a one-step solution to hold the r code and the data that has to be added to explain in report. It generates html and pdf formats for reports when knit. It reduces the pain of documentation.

Project summary

The project was about future learn dataset and we had to be found a solution that was effective from business perspective so researching about the MOOC courses was a bit of necessity and to know what challenges they face. Then data understanding was important because with that knowledge then only we can what was the way forward. The future learn dataset had 6 files for each iteration and there were 7 iterations in total. I observed that we had quite a lot of data missing in certain columns and hence had to optimise the data with respect to that. So, we can find good insights in it. I chose 3 data sets those are enrolments, question response

and step activity. These were quite insightful so I prepared them according to crisp DM and went to data preparation phase and chose 2 to 3 columns from all 3 data files separately which had the most values. Then I started exploratory data analysis and displayed all the insightful solutions I could get hands on. Aligned them to business perspective and then evaluate them against all the odds. All this data was regularly committed on git hence the version control was maintained. The project was made on project template so the modularity was maintained. There were 2 cycles taken into consideration. So, in 2nd cycle I got an idea what more could be added to the EDA and tried to take a look from a fresh perspective.

Evaluation:

I felt like that project was quite provocative in the sense that it enhances the thought procedure by the data set provided. As it has so much information to be interpreted. There are n number of possibilities but due to time constraint we had to limit ourselves.

Bring out emotions

The journey quite emotional because this was the first time, I took the 2nd cycle to reflect on my work and thinking from a fresh perspective again. I just felt that the whole process being a waterfall model is quite old and there are new amendments that can be made to the process. As now a days even software development cycle has moved onto spiral model and devops features are being considered. So that would be really helpful.

Review in the light of previous experience

As I have worked on git before when I was working in Accenture so it filled me with some memories. Also, over the last 2 modules I have got a hands-on experience with exploratory data analysis and still didn't how to put a constraint on one's thought If there are so many options to choose from.

Identify lessons learned

Thought smartification is important. We need to weigh upon our options, how modularity comes in hand for reproducibility. How reflective process enhances your thought procedure.

Follow up actions

I discussed the business ideas with my classmates and took their advices on how these things will add up. Then to get an experienced opinion I asked my professors to think as a business stakeholder and analyse if this would be things they want. To my luck they had some great insights about the business ideas I had and motivated me to think how can they help me go for cycle 2 analysis.

Feedback

I really like what the project template made me enable to modularity and git version control has always been a great practice. I intend to use these in future too. About CRISP DM, I really liked working on it but if there would be somewhat new practice that is introduced for data mining, I would like to go for that.

Shortcomings

1. There were a lot of unknown values so the dataset considered is biased.
2. Detected country was used instead of country provided as it had many empty values, so if the detection algorithm was fed some VPN data, the data might indicate false results.