REFLECTIVE LOG

CRISP-DM method is followed in the analysis of open online courses (MOOC) at Newcastle University. There are three cycles and the output of each one is fed into the next cycle and generated some useful information and insights on the provided data. Since the analysis followed all the CRISP-DM steps, the reproducibility and replicability of the study can be achieved.

The following are the advantages of CRISP-DM methods over other methods

- Flexibility- The model can be imperfect at the beginning but with each iteration, it is improved
- Long-term effect- The model can be developed from a simple structure if new data is obtained with time.
- Cross-Industry Standards- This methodology can be implemented in any DS project and any domain. This is the main advantage of this technique.

Implementing this method in this analysis provided the following information/outputs.

Based on the results of the first two cycles, the number of candidates working as full-time employees in the age range 26-35 which is also the majority age group is determined. The last cycle deals with the year-wise analysis of the output of the second cycle which indicates that the number of candidates is decreasing each year from 2016 to 2018.

The importance of this exploratory analysis is that it shows some insights into the existing trends of this online learning platform regarding its business growth. By reading the output of this design method, a business analyst or any stakeholder can take immediate action.

The graphical reports generated by these models can be easily reported using any data visualization tools like Power BI, Tableau etc. Since the method follows an iteration or cycle-like strategy, information with more clarity than the previous one is obtained which is one of the advantages of the CRISP-DM methodology. Since CRISP-DM serves as the foundation for all of these processes, analysts can also begin from scratch and improve their exploration, data mining, analysis and exploration phases as soon as possible.

FEEDBACK ON DATA

The model that we used in this project is a simple exploratory analysis. Even though it provided some useful insights into the data, more questions could have been considered. But, coming to a conclusion without having enough data can be dangerous. The online learning platform generated mainly three data sets. They are,

- cyber-security-x_question-response
- cyber-security-x_enrolments
- cyber-security-x_step-activity

All these data contain 7 sub-data for each run. The problem with the remaining data is that it is less in number so analysis may not be that much useful. That is the reason, the enrolment data has been chosen since it is a common form of data and enough records can be seen in it.

In the data preparation step, all the unnecessary and ambiguous values were removed or ignored which helps in increasing the efficiency of the model.

FUTURE IMPROVEMENTS

The observations in cyber-security-x_question-response and cyber-security-x_step-activity can also be considered which gives us different information. Apart from it, data sets like cyber-security-x_archetype-survey-responses can give additional information in future if more observation is obtained.

Also, some of the responsive variables can be predicted effectively with ML algorithms like decision trees and random forests, and with each cycle, prediction accuracy will improve.

If the predictive variable in the data is a continuous type, then Principle Component Analysis, Linear Discriminant Analysis and Regression model can be implemented which will result in depth mining from these data.

DISADVANTAGE OF THE CRISP-DM METHOD

- Lack of clear communication: Improper understanding of the business Requirement, will result in unwanted conclusions to the business which affect the time, resources and efficiency of the model.
- Irrational rework: This happens when there is a discrepancy between the business requirement and analytic results due to a lack of clarity on the objectives. This causes deviation from main objectives, and sometimes the team aim for a new model or data to solve this mismatch.
- Lack of Model updation: The ageing model must be maintained and monitored with time otherwise the model predicts irrelevant analysis.

Regardless of these disadvantages, the CRISP-DM model can generate the best results and it is very helpful in the data analysis field.