Name: Syed Mohib Raza
Student ID: 200740241
Newcastle University
CSC8631 - Data Management and Exploratory Data Analysis

Executive Summary

Methodology used: CRISP-DM

Business Requirement:

Since 2020, there has been significant rise in use of online education system due to the unfortunate impact of COVID-19 Pandemic engulfing the entire globe. The dataset provided is from the students enrolled in online courses offered by university. The dataset can be analyzed to evaluate and garner insights to further enhance the online offering to offer better solutions amidst the pandemic to every learner.

Due to the free hand approach of the research there was no interactions to understand how I could be able to assist efficiently to solve a certain problem, hence the business requirements were assumed and generalized while an interaction could've helped to personalize on the requirements.

Data Understanding

According to CRISP-DM, following points are constitute for Data Understanding.

CRISP-DM Summary:

- Initial data collection
- Familiarity with data
- Identify data quality
- Discover first Insights

The above points have been achieved partially, as the data was already provided. However, the familiarity with data have been achieved using basic R functions and quality been assessed for missing values.

Data Preparation

The Project contains 3 datasets, two of which can be immediately used, however for the dataset containing "enrollment" data, some data wrangling has been done to remove unknown values to obtain a refined sample.

The difficulties rise while acquiring data as it brings ethical and societal implications, and the legitimacy of data is also a concern as the data mined or provided may contain unknown or incorrect values.

Modelling and Evaluation:

In some specific data science workflows, modelling is not required and statistical summary and relations between variables for the given data set is sufficient to take informed decisions. Modelling techniques are required only in specific case basis.

The Evaluation is dependent on modelling where performance of the chosen model is evaluated before final deployment. However, as mentioned above, if there is no sophisticated modelling, there is no evaluation.

Deployment

The target audience is the university officials responsible for online course offerings to enhance the course structure and understand valuable insights on the performance of current courses hence a report has been generated using R markdown and Project Template to have better reproducibility, all the necessary files are included in the Project Template. A presentation has also been presented summarizing every analysis to stakeholders.