Project Report

Data Management and Exploratory Data Analysis

Stanly Wilson Palathingal
190586573

The report follows the CRISP-DM methodology. It has 6 steps and they are explained briefly.

1. BUSINESS UNDERSTANDING AND OBJECTIVES

The study is oriented around an online course named “Cyber Security: Safety at Home, Online in Life”, conducted by Newcastle University.

1.1 Assessment of the Situation

Though initially, the course attracted many learners. The situation changed after the starting year where there is a decline in enrollment for the course.

1.2 Data Mining Goals

The purpose of the analysis is to understand the performance of the course, collecting insights from the available data and propose ways to overcome the drawback if any.

1.3 Project Plans

I am trying to understand the overall performance of the course during the seven years.

2. DATA UNDERSTANDING

2.1 Collect Initial Data

For any analysis, there needs to be some data to study and to explore. In this case data is provided by the concerning department. To perform analysis and suggestions, a collection of 53 files containing various parameters and seven documents to give some idea of the module taught.

2.2 Describe Data

The data is a collection of seven years during the period of 2011 – 2017. They contain details like enrollment, question response, step activity, team members, video stat, weekly sentiment survey, archetypes and leaving survey response for these seven years.

2.3 Explore and Verify Data

The data contain various fields, and many relate to a common factor which is the learners' id. It is a primary key to connect with five files in each year. They are archetype, enrollment, leaving response, question response and step activity. The other files are independent without having a connection with each other.

The data has various fields having unknown values. In Some variables most of the values are not mentioned. It affects the analysis. Consider the attributes of the enrollment file like employment status, highest education or the age range, where most of the values are marked as unknown. When the majority is not mentioned any analysis made using those columns could generate unwanted analysis.

3. DATA PREPARATION

3.1 Selection of the Data

For the part of the analysis, I have taken only several data and not all the data. I have not taken the files of video stats, weekly sentiment surveys, leaving survey response and team members. All the other files of all the seven years are used in one way or the other.

3.2 Cleaning and Constructing Data

No data cleaning or scaling for any purpose. It is due to the face that most of the attributes are having values that do not give any detail even to make some assumptions. Consider the example of the field age group where the values are given as unknown. And so, the values are taken as it is given. For the analysis, no construction of new data or values, but have created some data frames from existing data.

3.3 Integration and Formatting the Data

The analysis is the outcome of the extensive formatting of the data and have integrated some of them. Not all the available data is used for the analysis. From the available data, some fields were extracted and made data frames to suit my purposes. Formatting is done to many files of all the years and then integrated them to form some new data frame. The data from these files are used to created joins and produce a new one.

4. MODELING

No complex modeling techniques for the analysis nor any test design. The analysis was completely depended on what assumptions one could make from the data and how it could be used to explain to a second person. The data is collected from different files and made a count of the same. It is a simple way

of extracting data from multiples files, create new data frames, analyze them, and if possible, generate a plot.

5. EVALUATION

For the analysis, I have made 4 different reports which follow one to the other. Though I may not have made a great discovery I tried to make it connected in some manner.

5.1 Evaluation and Review Process

For the evaluation, R programming is used as a tool with its various options and packages. There are four reports came as the outcome of the analysis. The first report is generated from that of taking archetypes as the input files of the seven years. The purpose of the analysis to get the count of archetypes and then compare them with other fields. The analysis did not make much impact since the number of archetypes was nowhere closer to that of people who enrolled or undertook the course.

A second analysis was made with the intention of analyzing how many people start the course and reach till the end of the course. For that I first analyzed the enrollment in each year to see whether all the enrolled candidates start and complete the course or not. The enrollment chart made it clear that there is a sudden decline (around 50%) in the interest of the course after the first year and it continued ever since. The analysis is also made to see the candidates who move from one stage to the other. The course has a duration of three weeks. And so, an attempt is made to check what was the progress from week 1 – 3. There too except for the year 2011, all the other years, only a third of candidates who start the course reach the up to week 3.

Inspired from the second analysis, the analysis is taken into still deeper. There are different steps to work each week. For example, in week 1 there are 22, week 2 has 19 and week 3 has 21 step works. So, the purpose of the analysis is to investigate these step numbers and see where the fall happens. It was observed that the decline is gradual till the end except for week 3 step number 18. Except for year 2011, all the other years this step has a steep decline to a very low number. Looking into the documentation part, it was understood that this step is ‘Test’.

An attempt is made to understand any character of those who completed week 3 step 18. In order to do so, they are compared to two files and they are archetypes and enrollment. After comparing it with the archetype, it did not make much progress since there is only one-tenth of the people found matching there. Further analysis was depended on the enrollment dataset. There too much of the fields are marked as unknown. Any analysis of those would have resulted in improper results. There were two fields that have all the values and they are role and detected country. The analysis of the detected country based on the IP address may not be a good one to make the analysis. So only role is used for the analysis. Among the candidates who went through the section of ‘test’ were ‘learners’ and in each year there was one from ‘organization admin’

6. DEPLOYMENT

For the deployment of the analysis, R markdown is used as well as one simple shiny app. The link for the shiny app is https://stanlypalathingal.shinyapps.io/EDA_1/

In the shiny app, only the initial report is available, and the rest are there in the reports. There are four reports generated using the R markdown files which contain the analysis and the plots.

The project files are available in GitHub and it is publicly available. The git address is https://github.com/stanlypalathingal/EDA_Project

REFLECTION

CRISP-DM methodology has provided a good background to approach the problem. It is having an introspective character where the evaluation if not satisfiable, permits one to go back to the business understanding and try to see the problem in a different angle. So, no analysis is unacceptable, but enable one to proceed further from the lessons learned and restart with initial steps.