Data Management_EDA_CSC8631_Report

Jyotsna Verma

12/11/2021

INtroduction

Data Understanding:

We have data of 7 batches which was run in a span of 2 years from 2016 to 2018.

We have 7 sets of static as well as fluid data for each batch which gives us the details about the student records, and the students' interaction with the virtual course. There different datasets for all the seven batches are:

1) Archetype survey response – it gives the details about the learners and their archetype, that is how they intend to use this course for.
2) Enrolments – this gives the students' details and information, and their enrolment details and the full participation in the course (course completion, that is purchased the upgraded course certificates).
3) Leaving survey response – this data set gives us an idea of the learners who left the course when (the timestamp as well as the last step at which they left the course) and the reason of leaving.
4) Question response – this dataset gives us the details of the assessment in the course.
5) Step activity – this dataset gives us the details of the steps(topics) and if the learners visited and completed that topic (or step)
6) Team members – this dataset gives us the knowledge of the role of the user in the course and their role in their team.
7) Video stats – this dataset gives us the details of the video modules in the course. It gives us the description of the video duration of the topics, the total no of views, number of people who downloaded the video and other details how people accessed these video modules. It also gives us the information about the percentage length of each video viewed. It provides us with the information of the devices which were used to access the videos and the information about the views from all the continents.
8) Weekly sentiment survey – this dataset gives the details about the feedback from the learners for the weekly modules.

We have similar types of files for all the seven batches. The number 1,2,3... and so on in the file names represent the batch number.

The data for enrolments, question response and step activity are provided for all the seven batches. The archetype survey response data and the data set containing the details of the video stats is not provided for batch number 1 and 2, but is available for all the other batches. The leaving survey response data is provided only for the batches 4,5,6 & 7. The data for team members is provided for all the batches except

for the first batch. The weekly sentiment survey dataset is only available for the batches 5, 6 and 7 where batch 5 contains the response from just one learner.

After examining the basic properties, summary and the details of the data we discovered that a lot of the entries in the data set are "Unknown" or have missing values.

So, we have done our analysis and visualizations on the data that is provided to us excluding the unknown values (that constitutes a significant amount of data). So, this analysis may or may not give exact results and the insights.

In the enrolment's dataset, we have country columns and the detected country columns, There are a lot of missing values in the country column, so we have analysed out data based on the detected country.

Data Preparation:

The data from the same type of dataset from all the seven batches are merged to get an overview of the kind of learners and for other analysis of the course.

The enrolment dataset and the archetype dataset are also merged to get an idea about the learners that how they intend to use this course for.

The exploratory textual analysis of the weekly sentiment's dataset is done by word cloud. It provides an excellent way to analyze the text data through visualization and helps to find important words that can help in extracting insights from the data through which we can communicate the most salient points in the reporting. For this a vector is created containing only the text and then the text data is loaded as a corpus. Then the data was cleaned by removing the special characters, numbers or punctuation from the text, as well as removing common stop words and stripping the white spaces. Then a data frame is created containing each word in the first column and their frequency in the second column which was used to create our desired visualization.

Evaluation/ Data Analysis:

Here we are checking the trend of the learners who enroll in the course and the ones who fully participated in the course.

Fig 1. this gives the trend of all the learners from all the 7 batches who enrolled for the course as well as who completed the course.

Here we can see highest number of learners enrolling in the first batch and then the no. Of participants decreases significantly. We can see that there is an increase in the number of participants for the 4th batch but after that it has shown a decreasing trend. We observe that there is significant difference in the number of learners who have enrolled and the ones who have completed the course. This can be analyzed further with leaving survey response dataset which contains the reason for participants leaving the course.

Fig 2. gives a visual analysis of the leaving response by the learners.

This shows that most of the learners who left the course have mentioned the reason of not having enough time. The reason about course requiring more time than expected is also chosen by a lot of learners, which depicts the same reason of learners struggling with time. Here we can further check what kind of learneres are involved and how can this issue be resolved.

Now, here fig 3 shows the Combined overview of learners who enrolled for the course and the ones who fully participated in the program (purchased the upgraded course certificates) .

Now this graph, fig 3 shows the different characteristics of the learners who enrolled and the learners who fully participated in the course, i.e., they completed the course.

Here when we see the employment area of the learners, we see that people from various domain have enrolled for the course but we can see that most of the learners are from IT and Information Services background and thus they are much interested in the course. The second dominating employment area is Teaching and education. May be there is a possibility that learners from these areas are pursuing this course for career advancements in their field. The line graph represents the learners who have fully participated (completed

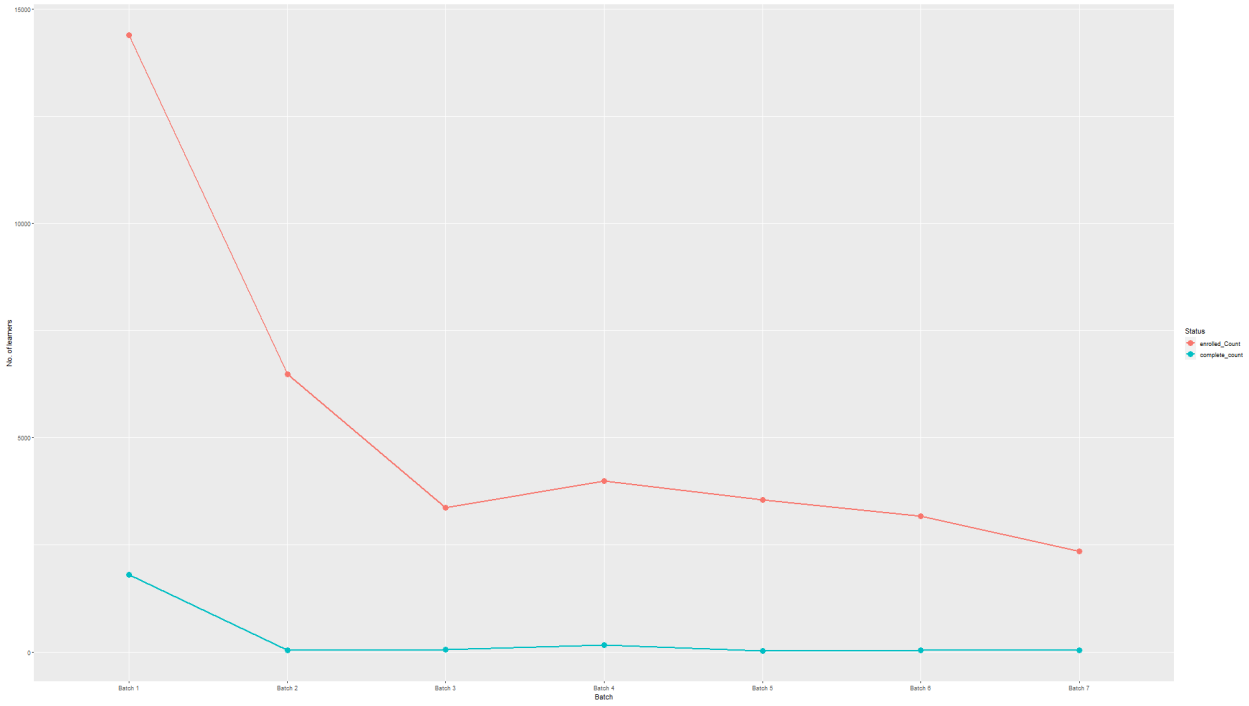

Figure 1: Trend of learners who enrolled vs who completed across all batches

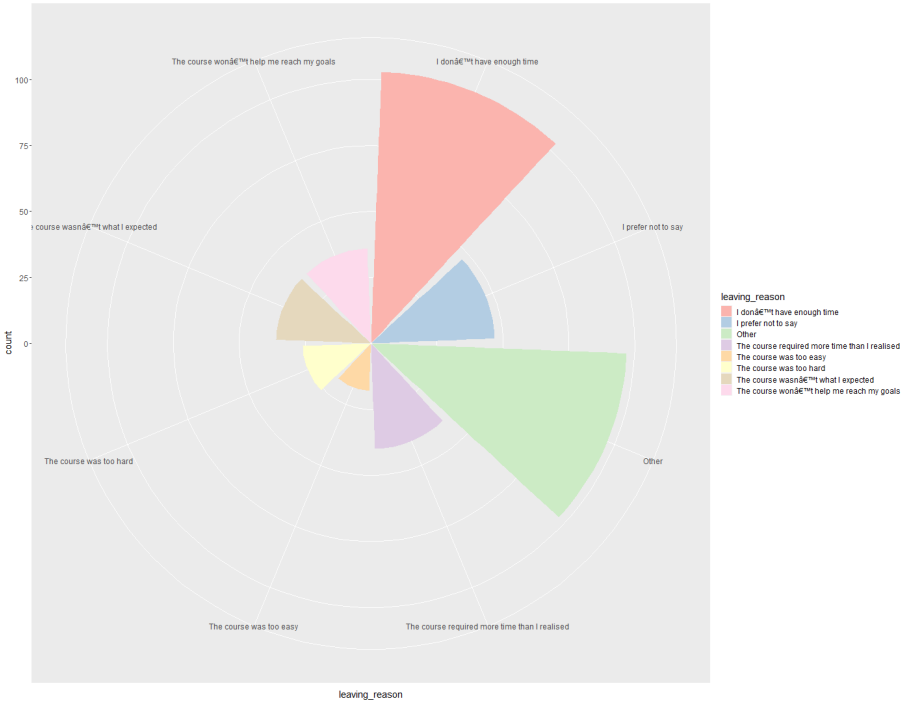

Figure 2: Leaving response of the learners

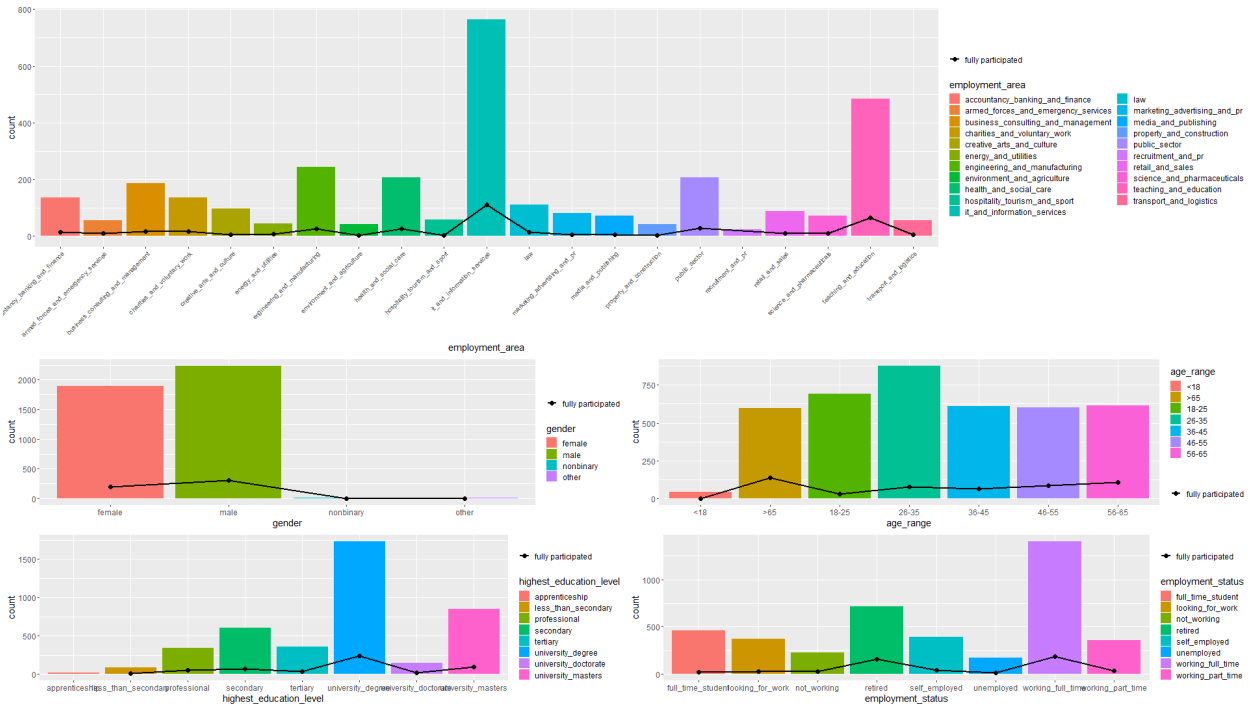

Figure 3: Enrolled and fully participated over different variables

the course, i.e, that is purchased the upgraded course certificates). Here we can see that they mostly follow the same trend as enrolments in this case. We notice that from the employment area, Recruitment and PR, none of the learners have completed the course. This can be further checked and can be a future scope of this project to see why this particular category doesn't have any learner who completed the course.

The course offered, that is, Cyber Security (Cyber Security: Safety at home, online, in Life) teaches about the cyber threats and the security measures to be followed and is kind of related to the IT domain-based course, so we can see that most of the learners are from that area and thus are our target audience. Now, when we are aiming the other group of learners, despite the course (as seen from the modules and the topics) explains about the basic cyber safety rules that can be implemented in daily lives, people from other backgrounds tend to be skeptical about the course thinking that it might be on the very technical side and thus not enrolling for the same. In this case we can focus on tailoring our course in a way that it's understandable and approachable from learners from varied backgrounds. One way can be introducing this course in two levels as 1) Beginners and 2) Advanced. With this there is a possibility that learners from different background will now be tempted and will look forward to joining the Beginner level of the course to get familiar with the topic and if they build a foundation, they will be interested in enrolling for the next level of the course to get the knowledge of the topic on whole. This can increase the reach and customer base for our business. There will be a tradeoff with effort and timing but it can be compensated with the profit made.

When we see the higher education level, we see that most of the learners who enrolled, have completed their university degree. They are our serious learners and our target audience. The approach that we used previously to target learners from various other domains can be beneficial here as well because then learners from any educational level background will be comfortable in enrolling for the course. Here, we can observe the learners who are on apprenticeship never fully participated for the program.

Now when we are analyzing the age range, we can see that most of the learners are from the age range 26-35. The second and third highest no of learners are from the age range 18-25 and more than 65 years respectively. Here we see that learners who are in the age range of more than 65 yrs tend to fully participate in the course the most as compared to others. Now, simultaneously if we see the employment status, we

can see the highest number of learners enrolled are working full time and then the next highest number of learners are from the retired category. Here again we see that learners from retired category tend to have fully participated in the course the highest as compared to others. This analysis shows that learners who are retired (generally they belong in the age range >65 years tend to fully participate in the course (I.e., compete the course and purchase the upgraded certificate). This can be due to two possible reasons.