CSC8631 report

Parth Sagar

26/11/2021

Introduction

This project is based on the data set of video streaming of a cyber security course which has been provided as a raw data and exploratory data analysis has been done on the same. The data is the insights about the course handling by the students in the different parts of the world. There are quite a lot of possibilities that can be made out of the data as there are multiple scenarios that can be helpful from the business perspective. So it was thought to select major business questions and work on the selected ones. The course covers 3 major topics which are further divided in certain steps.

As the project provides a business solution, it is quite necessary that the project should be quick, reliable and reproducible. This project follows the model of Crisp DM to attain the same goal. there were certain steps taken into account those are Business understanding, Data understanding, Data preparation, modelling, evaluation and Deployment.

Crisp Dm stands for cross industry standard for data mining. As this project follows crisp dm process. There were 2 cycle runs and all the steps were reflected and the objectives alignment with the data was reconsidered. As understanding of what happens and why is it is being considered is taken into account the data understanding and preparation is well thought and quality is hence enhanced.

Investigation Ideology

This project will be targeted on future learn data set. Here three data sets will be considered that is question response, step activity and enrollments. Ideology for considering these is that the data will indicate towards the countries where level of growth is possible and what optimization should be done to the course so that these countries are benefited. As these optimization will benefit Future Learn by increasing enrollments in the next cycle. It has been considered to learn what is the best duration of a step that should be considered that might help learners to accept the course as if it is of longer duration a human being tends to zone out and miss on important stuff or if the length is shorter learner might break his concentration in between and hence does not stay for long with the course. So considering this

as an optimization problem this will suggest the best length with respect to course content as an important that cannot be divided into sections cannot be considered for this. Future learn will be benefited through this as the learners would stay long with this course and their affinity towards Future learn will increase. So, that they might opt for other courses from Future Learn too. The next we are considering help for the questions with least accuracy as these are the questions that are not understood by the students. So if the content is revised for this then acceptability of the course will increase but the shortcoming is that it cannot be changed mid term, so then any additional content related to this can be provided through discussion forum and research paper but the question comes when to release this. So it will be answered by enrollments in a particular month as when the enrollments are high most of the learners might miss this content hence this should be released when enrollments are low. So that, most learners are benefited. Future learn here will be profitable as much more students will stay longer for the course and will tend to accept this form of learning. That indeed will help them indulge into future learn way of learning and hence they will prefer Future learn over other platforms because they are having the best accuracy after this course. The most time spent by a learner in a month suggests which is the best time to start promotion for next course which might be of interest to the learners. Future Learn can start promotion on the basis of the enrollments and when people are spending most time on the portal.

Data understanding:

Future learn data set is basically a video streaming data of a cyber security course which consists of 6 files that have been recorded for 7 different iterations for a number of learners. The data set for a particular iteration is consisting of files that is:

1. Enrolments: which has attributes like learner id, enrolment date, unenrolled date, gender, country, age range and employment status and type. Here it was majorly observed that some attributes have really less data available. There were a lot unknown entries. So, it was deducible that data has to be considered bias that is only those attributes should be considered where maximum amount of data is available.

2. Question response: question response data was basically dependent on the attributes like type of question, step involved, response for the question submission date and whether the question was correctly answered. So one more thing is that a question needs to be answered multiple times and each time the person is wrong the entry of false is made on their id. Next question pops up whenever they correctly answer the previous question. This data had least unknown entries.

3. Step activity: This data basically displays the time spent on a certain step in specific week.

4. Weekly sentiment survey response: this was majorly empty data set and was available for just 6th and 7th iteration. Majorly the data was centric towards the sentiments of

the user and what they feel about the course and how they are reflecting towards their learning.

5. Archetype survey responses: this survey responses were categorizes of the learner's archetype and the day they had responded.

6. Leaving survey responses: This data indicates when and why the learners left the course. This data had some empty entries but quite easily signified the reason and when the learner realized that the course wasn't meant for them.

Data preparation:

The data after understanding the well provided attributes and their functions in the scenario were considered for the analysis of the data. The files that have been considered are question response, step activity and enrollments.

Data cleaning:

Question response file had certain missing entries hence data cleaning was really required so all the missing entries rows were omitted. The attributes considered for this learner id, week number, step number, question number and if the question is correct or not.

Enrolments: three attributes were considered in this, as other attribute had a lot of missing data and didn't align to our business perspective. The attributes taken into account were learner id, enrolled at and detected country.

Step activity: three attributes of step activity were also considered as it gave the starting and ending time of a particular step by a certain learner. So, the time spent on the step could be calculated. Hence attributes chosen were learner id, first visited at and last completed at.

Data transformation:

The data from enrolments and step activity were merged on the basis of learner id and all the missing entries were omitted. As the data was in character form the date columns were mutated with Lubridate and Dplyr package. Also, to count number of enrolments in a month and number of enrolments from a country. Count function was used.

Average time spent in a month was achieved by using SQL query where the data was grouped by month and the learner's data was divided by the count of learners.

Total time spent on the steps by the total number of people in whole country this was achieved by running SQL query, it was achieved by grouping the data on the basis of the country, where the total number of learners were calculated and the total time spent by them. The whole data was the ordered in descending order.

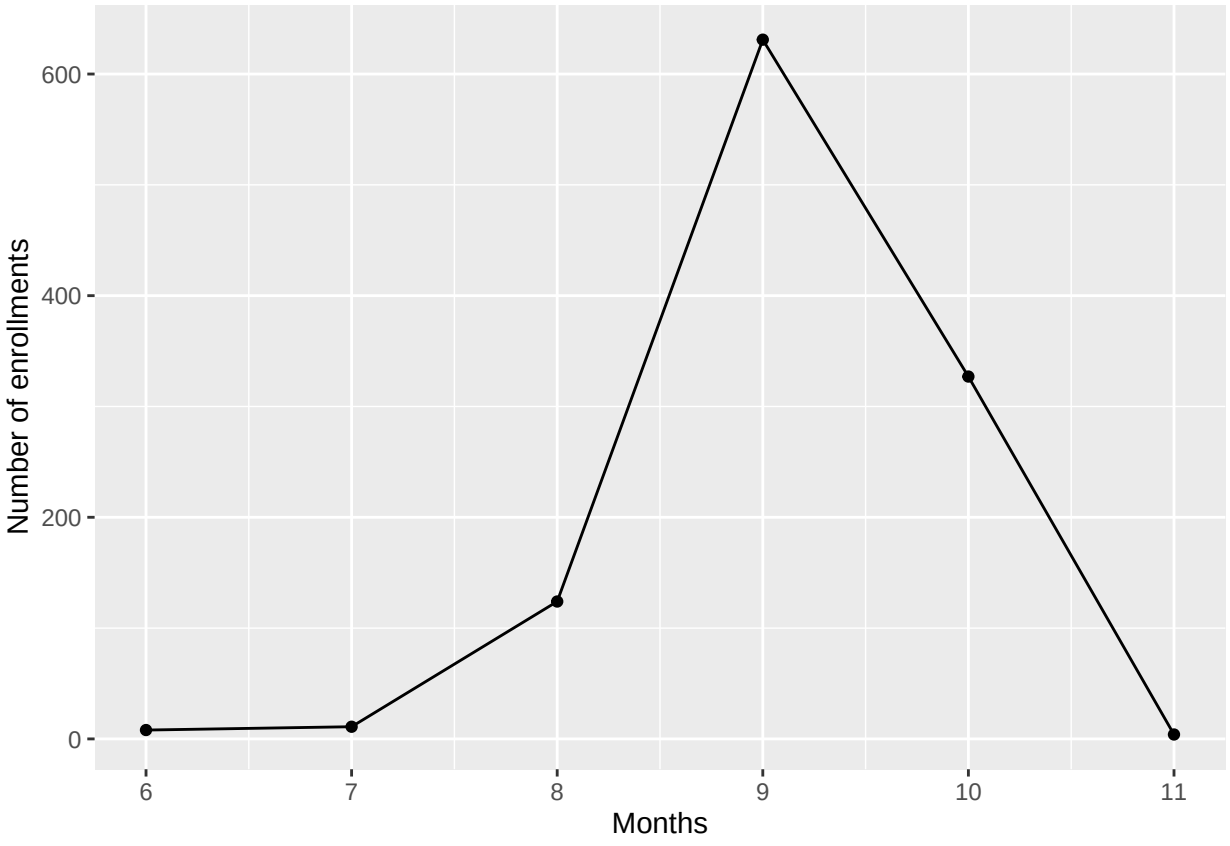
The question response for the data was transformed by first turning the character to numerical data that is categories of true and false to 0 and 1. Also running multiple queries on data, first to calculate accuracy on the basis of particular learner in a certain question in certain week. Then averaging it again for question and week. That provides the overall accuracy achieved for a particular question in certain week.

Methodology

The exploratory data analysis was started with importing data set enrolments and step activity. These 2 tables were merged on the basis of learner Id which is unique to each learner hence providing consolidated data about the country they belong to , when the people enrolled, when the first visited the step and when they last visited the step. The step has been chosen because that specifies the concentraion level of a learner to complete a task.

Cycle 1:

Objective: Enrolments observed in a particular month

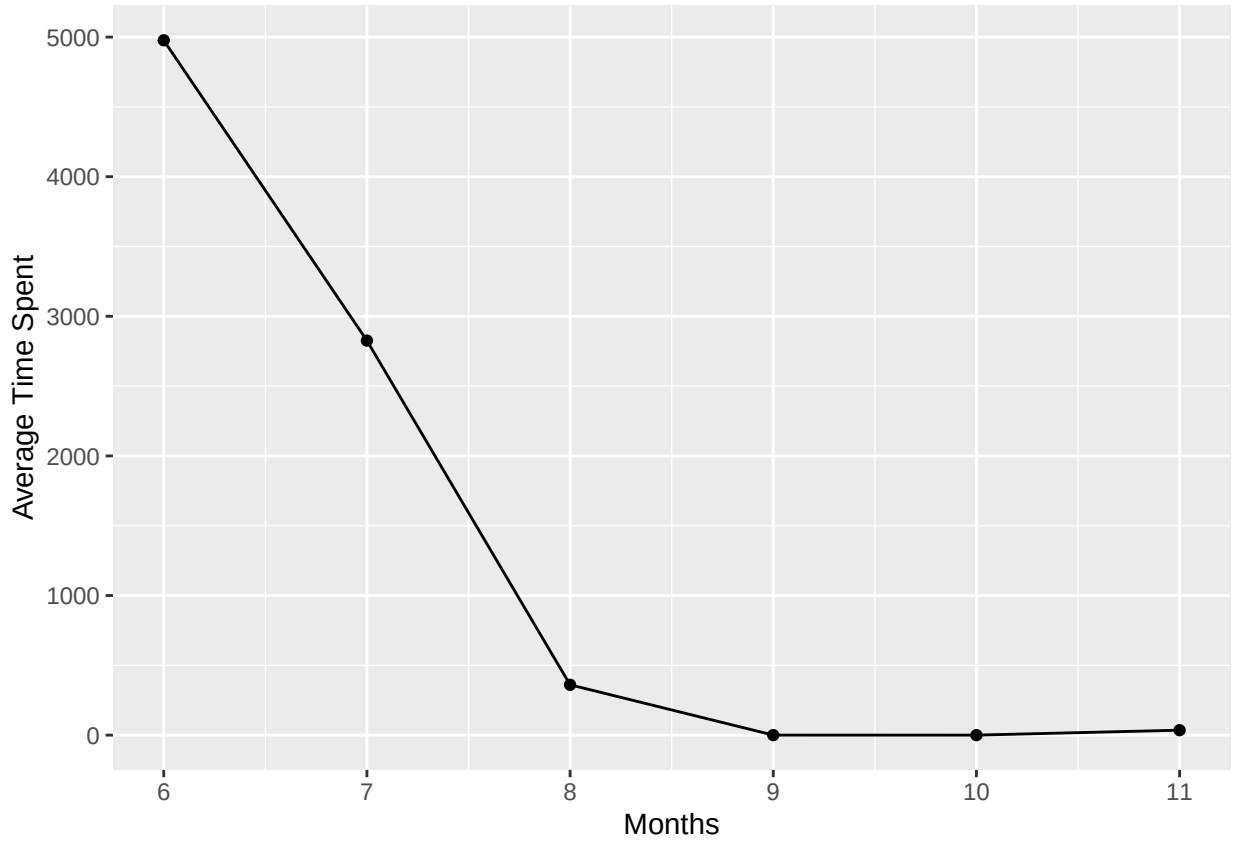


Analysis and Evaluation:

All the additional course content should be added in the months except when the enrolments are high. So that study quality of the learners isn't much affected, and they can learn in flexibility of their hours around the world. Also, they don't miss out on the additional course content added, for example optional research papers, not compulsory topics for the course. It can be observed that majority of enrolments are in month of September with approximately half in October. So, it can be assumed that if a course content needs modification or addition December to June can bethe optimum time for implementation. Hence, the new course content should be research and developed when the majority of people are enrolling. So, the additional course content (research papers and not compulsory topics) should be added in June or July as the number of enrolments is quite low hence the new and majority of the joiners will be able to access them in months of august to November.

Cycle 2:

Objective: Average time spent by learners in a particular month

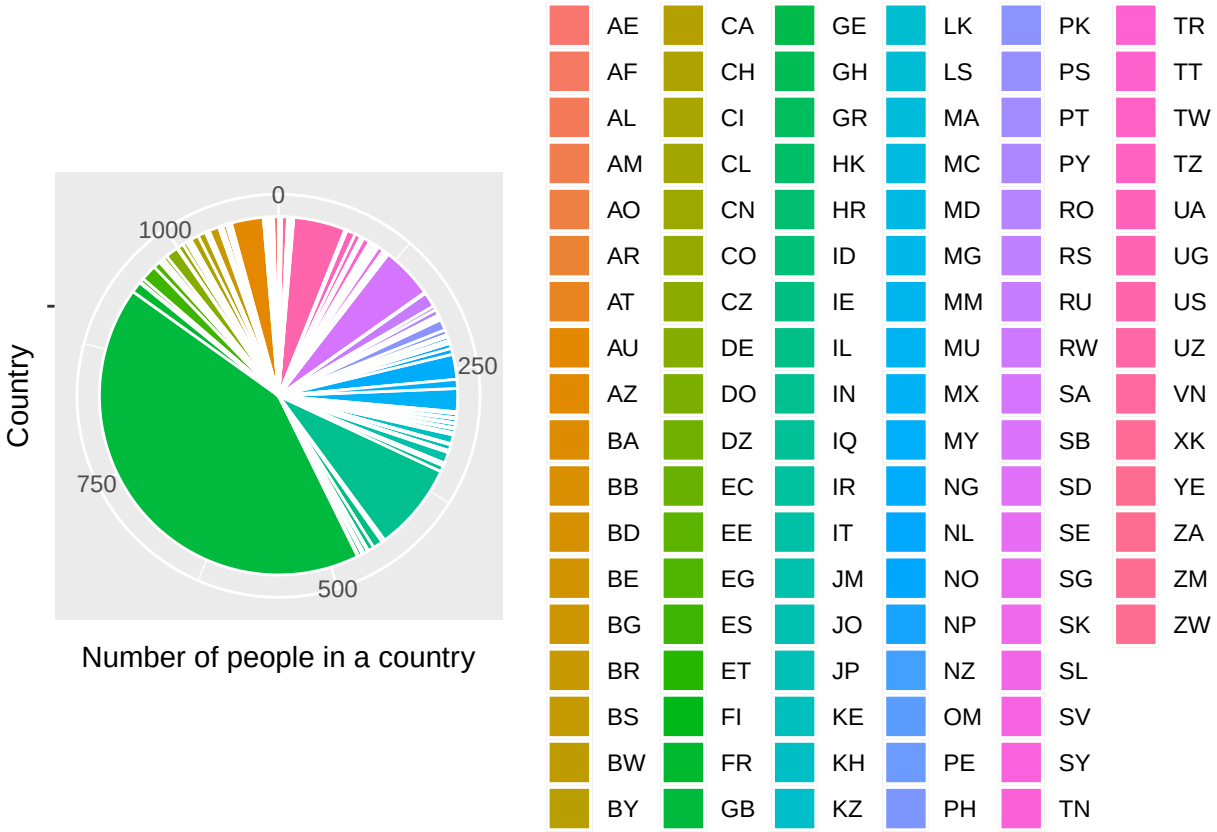


Analysis and Evaluation:

The months which have the most concentration of learners, those months should be the best time to be suggestive of related courses and releasing the most engaging content that engages the learner for more duration. So, the average time spent in a month by the learners is much more in June and July as compared to other months. This signifies the ad budget for other courses should be run in these 2 months to get users attention and although September has the most enrolments still it is way behind in terms of the average time spent by the users in that month. For the reason of the less amount of time spent, it can be observed that major enrolments in September left the course, for this it can be achieved through the data file "leaving survey of learners". That is not in scope of this report but can be considered if detailed analysis has to be accounted.

Cycle 1:

Objective: Enrolments from a particular country



Analysis and Evaluation:

The country in which there are more enrolments should be targeted for more marketing ads to increase the number of enrolments and as the course content is quite acceptable in these countries. Explicitly some features should be added to increase the sales, like captions/notes for their particular language. There are most enrolments from GB that contains about 40 percent of the total enrolments. Approximately 60 out of 104 countries have only 1 enrolment. So, the rest 40 percent countries should be targeted for majority of marketing campaigns and the subtitles support for their language should be given in these countries to increase sales of the course. Also, the countries with lower enrolments should be considered to find the root cause of less enrolments, so that program could be refined in a way that it is much more acceptable.

Cycle 2:

Objective: Average time spent by top 10 countries with highest number of learners.

```
##    detected_country totalpeople average_time_spent
## 1                GB         466          0.2989986
## 2                IN          88       2878.2859848
## 3                US          52          0.2307692
## 4                SA          52         10.1875000
## 5                AU          32        406.6093750
## 6                NG          25          3.2742857
## 7                MX          23         54.6521739
## 8                ES          16         63.6188525
## 9                RU          15         13.8666667
## 10               DE          13        785.3921816
```

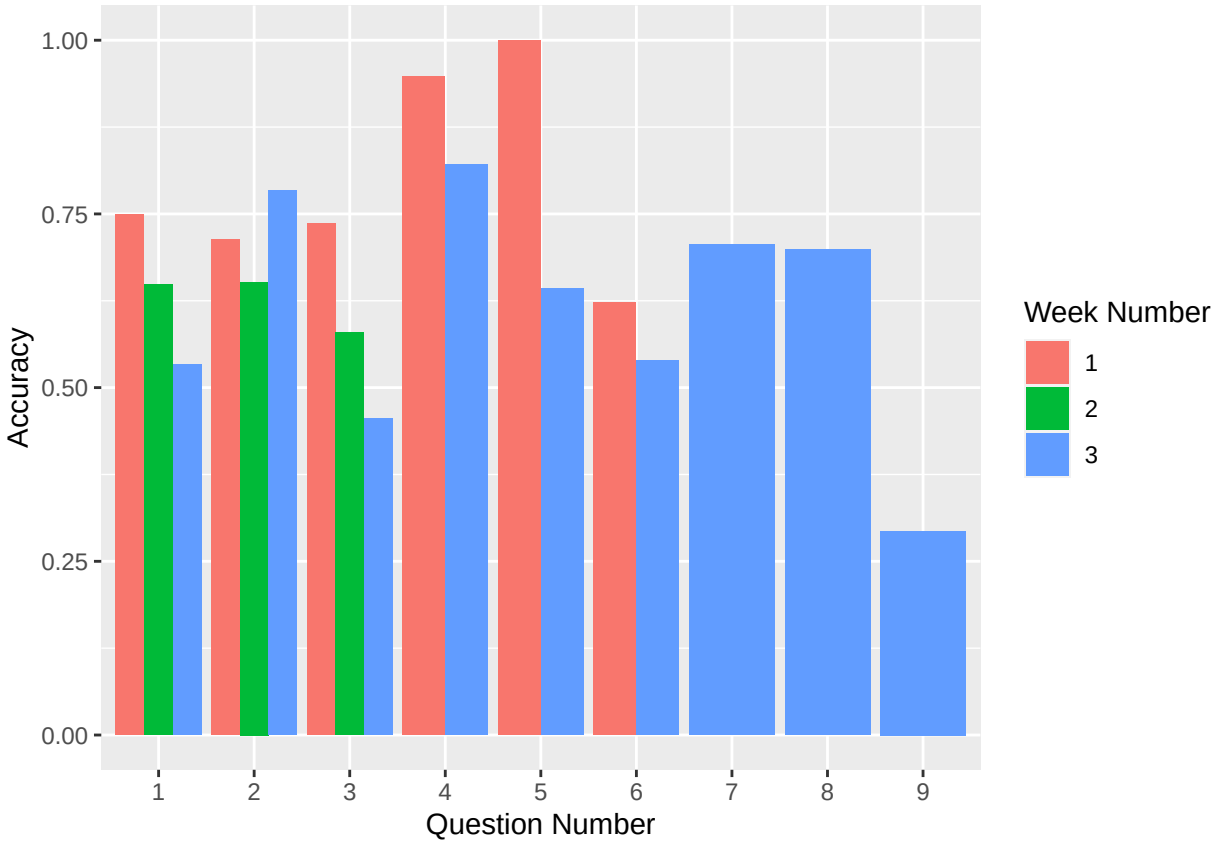
Analysis and Evaluation:

Average time spent on a step by the learners and total people from that country will allow us to make an informed decision of future course content as the time spent on a step means that learner is willing to spend this much time with full concentration in that region. That is around the average time, so for the further content length of the step could be kept in mind. For example, if people from India usually like videos of 20 minutes, then if the step duration is of 1 hour, they are most likely to not complete in a particular sitting hence the

momentum is broken so the course should be optimized like that. Also the speed options can be optimized according to the time and acceptability of the language. For example, Great Britain has the least average time spent on the course and they might miss if they skip some parts in it because of the speed is quite slow, so speed increase/decrease features could be more adjusted as per the time a learner can spend in that region. It can be observed that 3 out of top ten countries namely "IN", "AU" and "DE" have exponentially average spending time as compared to others. So for these countries subtitle support could be introduced so that they can get along faster and speed that matches there time that can be 0.5x or lesser could be introduced.

Cycle 2:

objective: Accuracy in a question with respect to weeks.



Analysis and Evaluation:

As a learner the accuracy of a question depends on the understanding of the topic. So, the questions with the least accuracy should be considered and the course content related to that question should be revised so that understanding of that section is better. Also, a help for these section through discussion forums or additional reading could be provided mid run.

As it can be observed the accuracy is less in question number 9 and 6 so additional help or supporting content in discussion forum should be released for these questions for better understanding of the learners. It can also be observed that accuracy has decreased over the weeks which can be concluded as that the content related to those questions must be in earlier steps. So, the retention of the content can be increased by providing a recap of previous part in the next run of course content.

Conclusion

So after going through the analysis it can be observed taht there are so many factors that contribute towards the future of a course. Therefore, these can be certain factors that can contribute towards the development of the better course and enhancing the experience of the current course but it would be wrong if we say these are the only possibilities because the data is so vast and can be viewed from as many angles as possible.