CSC8631 Report

Himika Mishra

2022-11-18

Introduction

This project is based on the raw data set of a cyber security course's video streaming, which was analyzed through exploratory data analysis. The insights provided by the data pertain to how students around the world handled the course. Due to the numerous scenarios that can be beneficial from a business perspective, quite a lot of possibilities can be generated from the data. The dataset must be cleaned and processed unambiguously for it to be useful in the business world. As a result, major business questions were taken into account and were further selected to be worked upon. The course covers three major topics that are further broken down into distinct sections.

Because the project offers a business solution, it is absolutely necessary for it to be quick, dependable, reproducible and flexible. To achieve the same objective, this project employs the Crisp DM model. Business comprehension, data understanding, data preparation, modeling, evaluation, and deployment were all taken into account. Cross-industry standard for data mining is Crisp Dm. Since this project adheres to the crisp DM process, all of the steps were taken into account during the two cycle runs, and the alignment of the objectives with the data was rethought and reconsidered. The quality of the data is improved because the process of understanding what takes place and why it is being considered is taken into account.

Investigation Ideology

The Future Learn data set will be used for this project and will be its main focus. .In this case, three data sets—question response, step activity, and enrollments—will be taken into consideration. The idea behind these is that the data will show which countries have the potential for growth and which aspects of the course should be improved to benefit these countries. Future Learn will benefit from these optimizations that will be made by fitting the model which will result in increasing enrollment in the following cycle. It has been thought about how long a step should be to help students accept the course because, if it is longer, students have a tendency to zone out and miss important information, while if it is shorter,

students may break their concentration in between and do not stay with the course for long. As a result, taking this into account as an optimization issue, it will suggest the optimal length for the course content, which cannot be divided into sections. This will benefit Future Learn because the students will stay with the course for a longer period of time and will develop a greater affinity for it. It will be beneficial for the firm as the students will be encouraged to choose other Future Learn courses. Next, we're thinking about providing assistance to students who don't understand the questions with the lowest accuracy. As a result, the course's acceptability will increase if the content is revised for this; however, it cannot be changed in the middle of the term. Any additional content related to this can be provided through a discussion forum and a research paper; the appropriate question is when to release this. Therefore, it will be answered by the number of students who enroll in a particular month. Since the majority of students might miss this material when enrollment is high, it should be made available when enrollment is low.The vast majority of students gain from this.Future learn here will be productive as substantially more understudies will remain longer for the course and will generally acknowledge this type of learning.They will, in fact, be able to adopt the Future Learn method of learning, and as a result, they will prefer Future Learn over other platforms due to their superior accuracy after completing this course. The most time spent by a student in a month proposes which is the best opportunity to begin advancement for next course which may hold any importance with the students.Future Learn can start advertising based on how many people sign up and when people spend the most time on the portal.

Data understanding:

The Future Learn Data Set is essentially video streaming data from a cyber security course. It consists of six files that were recorded for seven distinct iterations for a number of students. The following files make up the data set for a particular iteration:

1. Enrolments:which contains information like the learner's id, enrollment date, unenrolled date, gender, country, age range, and type of employment. Here, it was mostly noticed that some attributes have very little data. There were a great deal of obscure passages. Therefore, it was deduced that data must be considered biased, meaning that only attributes with the greatest amount of data should be taken into account.

2. Question response: The attributes of the question, such as the type of question, the step involved, the response for the question submission date, and whether the question was answered correctly, largely influenced the question response data. Also, a question must be answered multiple times, and if a person answers incorrectly, a false entry is added to their ID. When they give a correct response to the previous question, the next one appears. The most entries in this data were unknown.

3. Step activity: The time spent on a particular step during a particular week is basically shown by this data.

4. Weekly sentiment survey response: This data set had very little data and was only available for the sixth and seventh iterations. The majority of the data focused on user sentiments, including how they feel about the course and how they are evaluating their learning.

5. Archetype survey responses :The learner's archetype and the day they responded on were used to classify the responses to this survey.

6. Leaving survey responses: This information reveals when and why students left the course. Although there were a few blank entries in this data, it was easy to figure out why and when the student realized that the course wasn't for them.

Data preparation:

After comprehending the provided attributes and their roles in the scenario, the data were taken into consideration for analysis. Question response, step activity, and enrollments are the files that have been taken into consideration.

Data cleaning:

Because some entries in the response file were missing, data cleaning was absolutely necessary, so all rows with missing entries were omitted. The attributes taken into account for this are learner id, as well as the week number, step, and question numbers and whether or not the question is correct.

Enrolments: Three attributes were taken into consideration because the other one lacked data and did not match our business perspective. The characteristics considered were student id, enrolled at and distinguished country.

Step activity: Three aspects of the step activity were also taken into consideration because it indicated the learner's starting and ending times for each step. Thus, the time spent on the step could be determined. Therefore, learner id, first visited at, and last completed were selected.

Data transformation:

On the basis of the learner id, the data from enrolments and step activity were combined, and all entries that were missing were omitted. Lubridate and the Dplyr package were used to modify the date columns because the data was in character form. Additionally, to count the number of enrollments per month and a nation's enrollments, the count function was employed. By grouping the data by month and dividing the learner's data by the number of learners, a SQL query was used to calculate the average amount of time spent per month. This was accomplished by running a SQL query and grouping the data by country, where the total number of learners was calculated and the total time spent by them as well as the

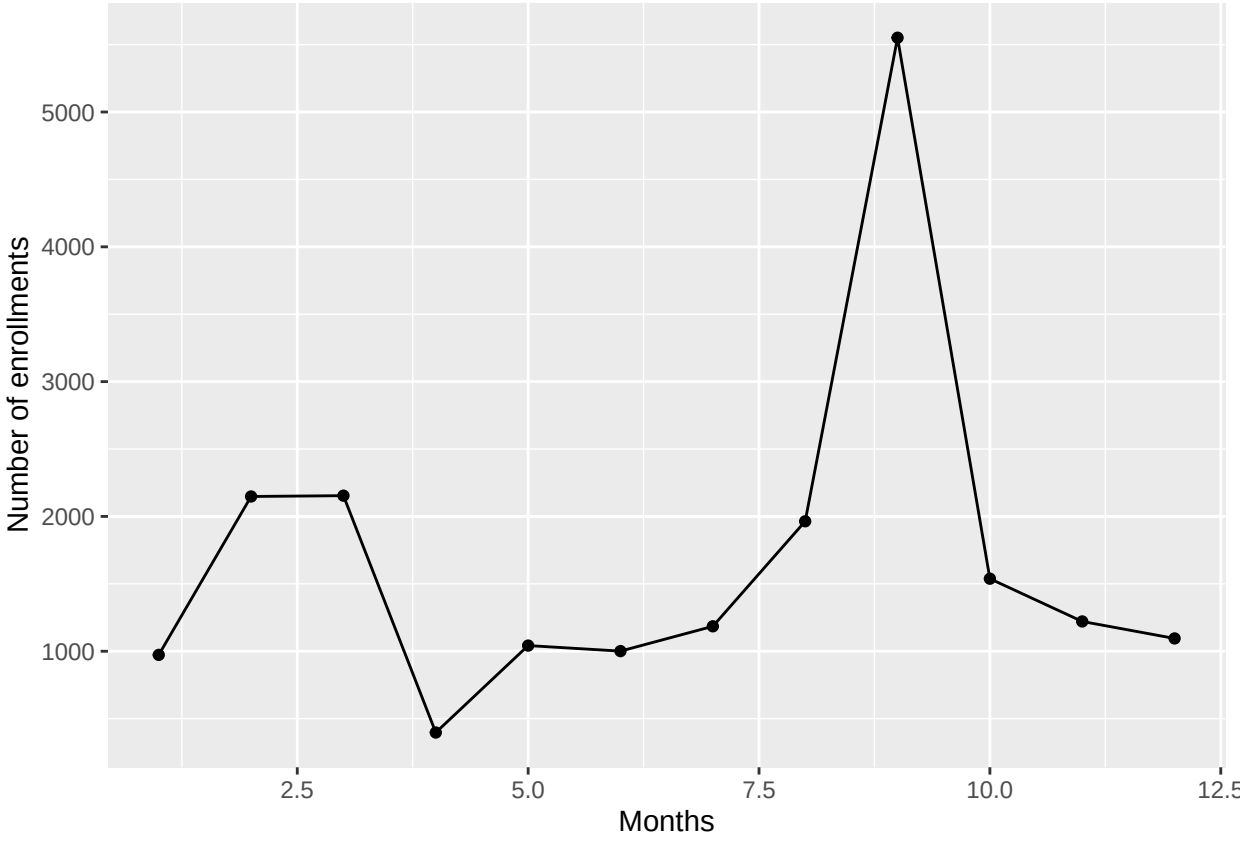
total time spent on the steps by the entire country. The entire set of data was arranged in descending order. First, the character was converted into numerical data, with categories of true and false set to 0 and 1. This transformed the data's response to the question. Also, running multiple queries on the data to first figure out a learner's accuracy on a particular question in a particular week. Then setting a new average for each question and week. That tells a person how accurate a question was overall in a given particular week.

Methodology

Importing enrollments and step activity from the data set served as the foundation for the exploratory data analysis. Based on the learner ID, which is unique to each and every learner, these two tables were combined to provide consolidated information about the country of origin, enrollment date, first and last step visits, and many more. The step has been chosen because it indicates a student's level of concentration needed to complete a task in a given time.

Cycle 1:

Objective: Enrolments observed in a particular month



Analysis and Evaluation:

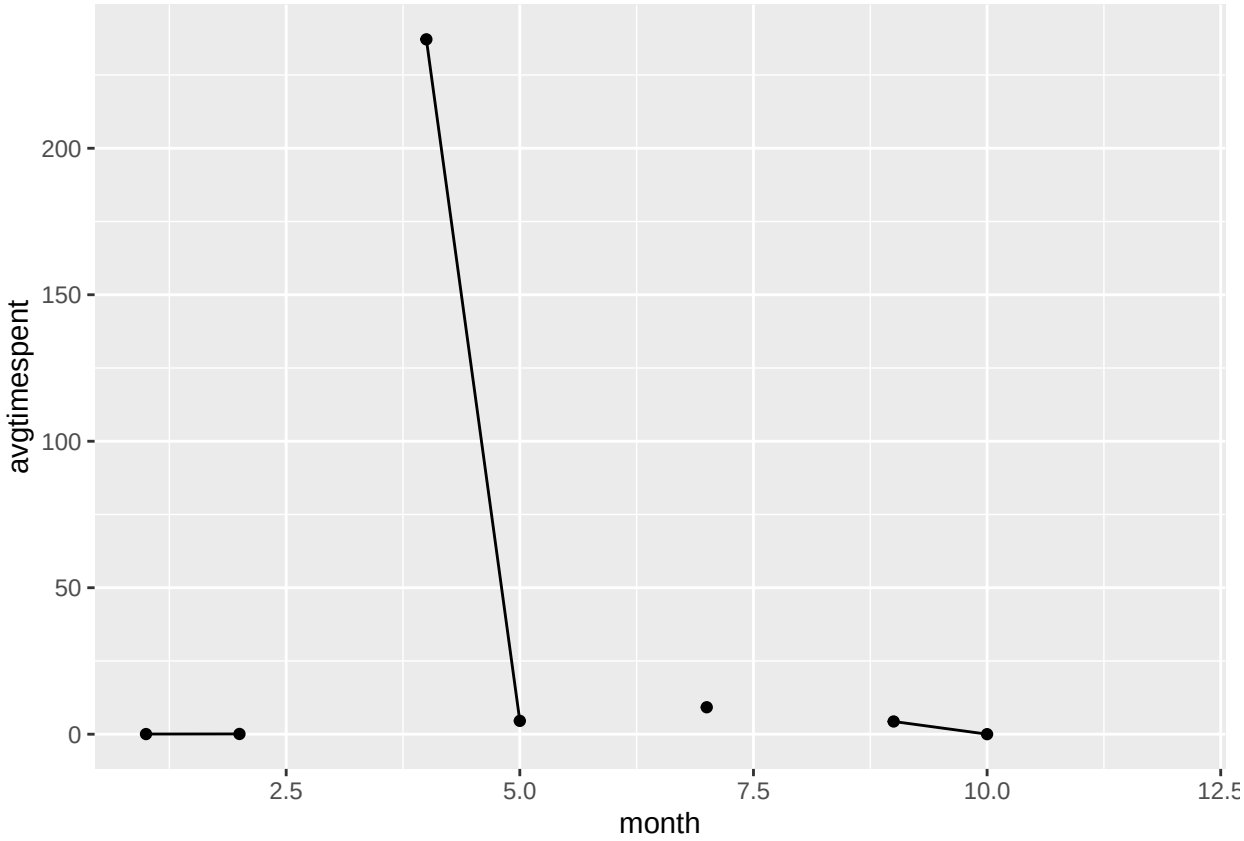
Except when enrollment is high, all additional course material should be added in the months, so that concentration on nature of the students isn't quite impacted, and they can learn in adaptability of their hours all over the planet. Additionally, they do not miss out on any additional course material, such as optional research papers or topics that are not required. It is evident that the majority of enrollments occur in September, with approximately half occurring in October. Therefore, if a course's content needs to be changed or added, November to June may be the best time to do so. As a result, the new course material ought to be researched and developed prior to the majority of students enrolling. As a result, the additional course material—research papers and non-mandatory topics—should be added in July or August. Because the number of enrollees is relatively low and the majority of new enrollees and newcomers will be able to access these materials between August and November.

Cycle 2:

Objective: Average time spent by learners in a particular month

```
## Warning: Removed 5 rows containing missing values (‘geom_point()‘).
```

```
## Warning: Removed 2 rows containing missing values (‘geom_line()‘).
```

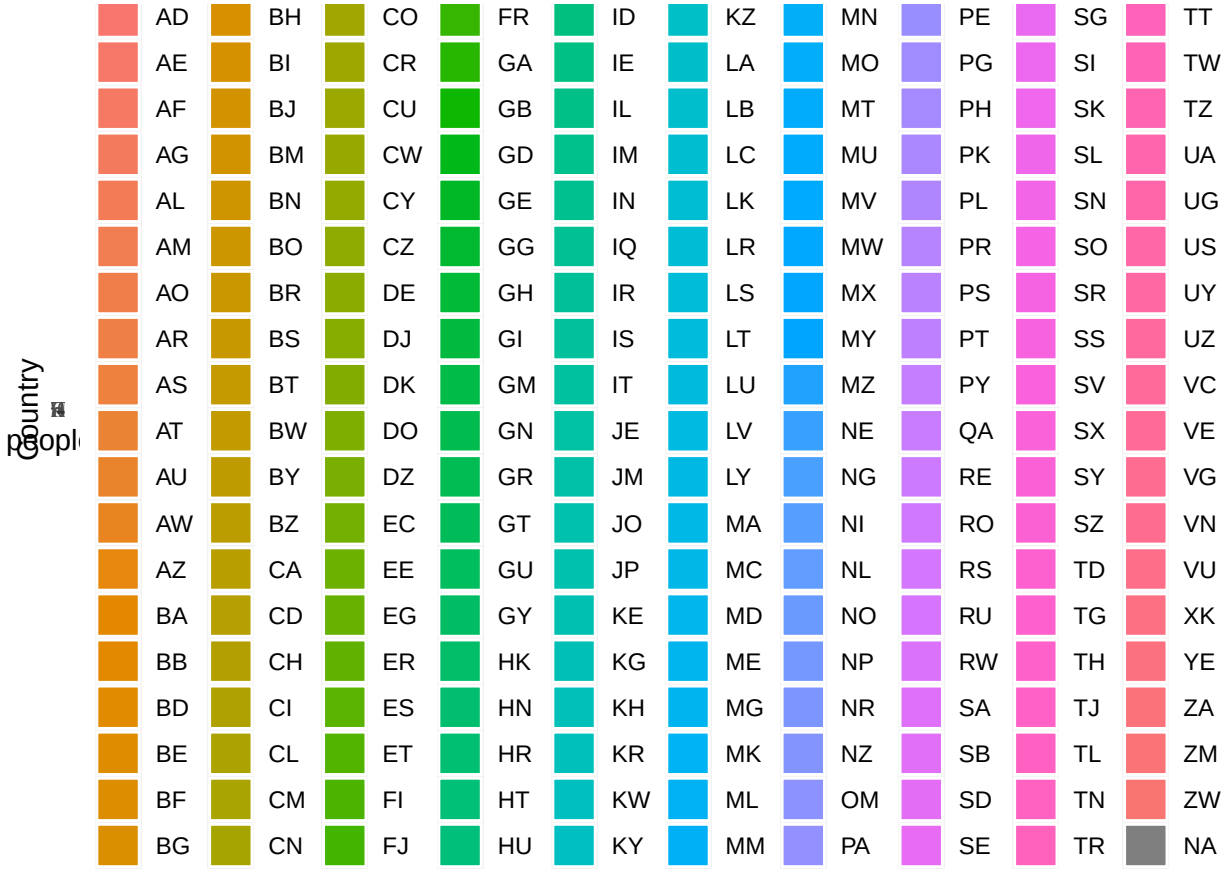


Analysis and Evaluation:

The best times to suggest related courses and release the most engaging content that keeps the learner engaged for longer periods of time should be during the months with the greatest concentration of students.As a result, students spend significantly more time in June and July than they do in other months.This indicates that the advertising budget for other courses should be used in these two months to attract users' attention, and despite the fact that September has the most enrollments, the average amount of time spent by users in that month is still a significant distance behind.The data file titled "leaving survey of learners" can be used to determine that major enrollments dropped out of the course in September due to the shorter duration.That is outside the scope of this report, but it could be taken into account if a more in-depth analysis is required.

Cycle 1:

Objective: Enrolments from a particular country



Analysis and Evaluation:

Since the course material is well-liked in these nations, more marketing campaigns should be directed at the countries with the highest enrollment rates. To increase sales, certain features should be explicitly added, like captions and notes for their particular language. The majority of enrollments come from Great Britain, which accounts for roughly 40% of all enrollments. Out of 104 countries, 60 only have one enrollment. In order to boost course sales, the remaining 40% of countries should be the focus of most marketing campaigns, and subtitles that support their language should be made available in these countries. Additionally, countries with lower enrollment rates ought to be taken into consideration in order to identify the underlying cause of the problem. By doing so, the program could be improved to make it much more palatable.

Cycle 2:

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

```
##    detected_country totalpeople average_time_spent
## 1                GB        6641       3.463334e-03
## 2                IN        1927       2.179554e-01
## 3                US        1076       3.288104e-01
## 4                SA         930       6.929510e-03
## 5                NG         555       4.175225e+00
## 6                AU         531       8.662900e-02
## 7                MX         520       5.910521e+01
## 8                PK         298       1.035475e-01
## 9                ES         285       1.543860e-01
## 10               CA         242       6.445708e+02
## 11               NL         223       4.889088e+01
## 12               FR         221       2.310860e+01
## 13               ZA         212       1.320755e-01
## 14               DE         207       4.932415e+01
## 15               GH         188       3.553191e+00
## 16               IE         186       4.027243e+02
## 17               BR         150       1.607900e+02
## 18               KE         142       9.662113e+02
## 19               TR         131       1.857044e+02
## 20               BD         129       1.859008e+02
```

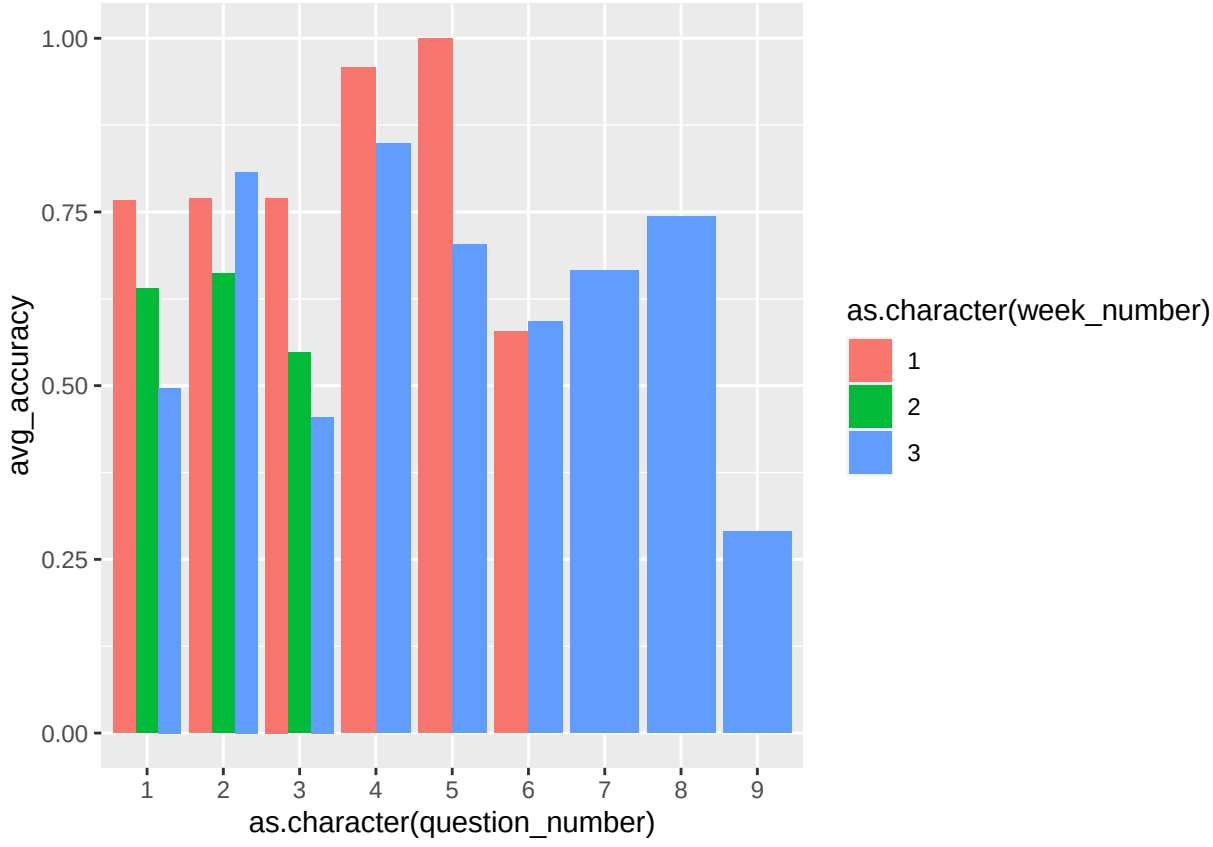
Analysis and Evaluation:

We will be able to make an informed decision regarding the content of future courses based on the average amount of time spent on a step by learners and total people from that nation, as the amount of time spent on a step indicates that learners are willing to devote this much time to the region with complete concentration.Since that is roughly typical, the

step's length could be kept in mind for additional content.For instance, if Indians typically prefer 20-minute videos, they are more likely to complete the course in one sitting if the step duration is one hour. This breaks the momentum, so the course should be optimized in this way.Additionally, the speed options can be optimized in accordance with the language's acceptability and time.For instance, Great Britain has the lowest average time spent on the course, so speed increase/decrease features could be more adapted to the amount of time a learner can spend in that area. They might miss out on some parts if they skip them because the speed is so slow.Three of the top ten countries, namely "IN," "AU," and "KE," have an exponentially higher average amount of time spent compared to other countries.Therefore, subtitle support could be made available to these nations so that they can interact with one another more quickly and at a speed that is equal to or less than their time.

Cycle 2:

objective: Accuracy in a question with respect to weeks.



Analysis and Evaluation:

A learner's comprehension of the subject matters in determining a question's accuracy.Therefore, the questions with the lowest accuracy ought to be taken into consideration,

and the material in the course that is related to that question ought to be revised to improve comprehension of that section.Additionally, assistance for these sections may be provided halfway through via discussion forums or additional reading.As can be seen, the accuracy of questions 9 and 6 is lower, so additional help or supporting material in the discussion forum should be made available for these questions to help students understand them better.It can likewise be seen that exactness has diminished over the course of the weeks which can be closed as that the substance connected with those questions should be in prior advances.Therefore, providing a summary of the preceding section in the subsequent run of the course material can improve content retention.

Conclusion

As a result, examining the analysis reveals a plethora of factors that influence a course's future.Because the data is so vast and can be viewed from as many perspectives as possible, it would be incorrect to claim that these are the only possibilities. However, there are certain factors that may contribute to the improvement of the current course and the development of a better course.