Exploratory data analysis of the survey for Future Learn's Cyber Security course

Shubham Rajesh

03/12/2021

Abstract

This report deals with Exploratory data analysis of the Survey for the Course Cyber Security offered by Future Learn using CRISP-DM methodology. The primary objective was to understand why enrollments in the course had lowered by investigating their behaviours as well as the course completion rate. Data from students who enrolled in the Cyber Security online course in late 2017 and 2018 was used for the research. The goal of this study was to use Exploratory Data Analysis to better identify the gaps in the course and reduce the course dropout rate. This report focuses on the reason for leaving, the last step completed by the student before leaving, and the week in which the students left the course after enrolling. The data of approximately 400 students were analysed by looking for patterns and determining when and why they left. To acquire a better understanding of these business questions, visualization was used.

Introduction

Candidates all around the world are becoming more interested in mastering cyber security as a way to further their careers. Because of the increased use of the Internet daily and the unprecedented nature of the times, most universities have opted to make their courses available via online sessions. Because the course is provided entirely online, there may be certain challenges and issues that both students and instructors dealing with the modules may encounter. This report examines a Massive Open Online Course (MOOC) on Cyber Security for late 2017 and 2018 using the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology and R programming. These courses have traditionally been used in the context of teaching and learning which also have recently gained popularity because of the COVID-19 outbreak. The primary objective of this study is to understand why dropping out in the course has increased by analysing their behaviours as well as the course retention and completion rate. By doing this research, the reason why students drop out of classes would be clear and, as a result, enhance the course process for the interest of students. It is universally acknowledged that different learners use learning content in various ways. Some learners, for example, choose to complete their course work, but others make an effort to learn the most fundamental elements from the lengthy coursework and then abandon the rest. On the other hand, this research could be used to uncover the most common student learning patterns and reasons for dropping out. Future Learn gathered information from students who enrolled in the Cyber Security online course in late 2017 and 2018, with students who dropped out in the middle of the course being required to complete the survey. The course was offered in four different intakes: November, February, June, and September. With the help of visualizations, the focus is to enhance the course plan for future intakes so that students can acquire better quality material and presentation of the entire course work with a lower leaving rate. This report focuses on the reason for leaving, the last step the participant completed before leaving, and the week the student left the course after enrolling. The data of around 400 students were examined for trends and to determine when and why they left. Data Visualization was employed to have a better understanding of the business problem.

About CRISP-DM Methodology

CRISP DM abbreviated name for CRoss Industry Standard Process for Data Mining (CRISP-DM). The CRISP-DM process is a technique for planning a data mining project in an organised manner and is also well-known for its dependable and well-tested technique. CRISP-DM involves a huge amount of repetition. It's crucial to remember that going through the process without finding a solution isn't always a failure. The data science team learns more about the data and the whole project upon each run through the process. CRISP-DM methodology is divided into six stages as shown in Figure: 1.1 CRISP-DM Model.

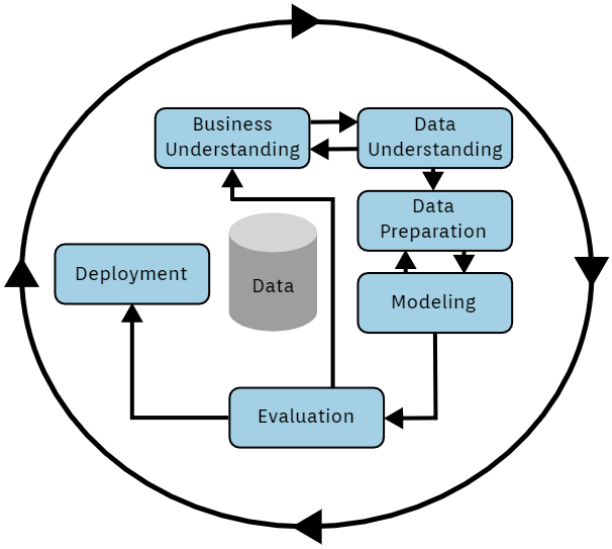

Figure 1: The CRISP-DM Model

Business Understanding:

Business Understanding defines that identifying the project objectives and needs from a business standpoint, and then incorporating that knowledge into a data mining problem statement and a preliminary plan to achieve the objectives. The purpose of this stage of the process is to identify key aspects that could have an impact on the project's outcome.

Expected Steps:

- Identify the business Objective
- Plan project strategy
- Determine business success criteria for business problem

Data Understanding:

The data science team should engage in data exploration at this point to determine what data is already available and how clean it is. Another potential obstacle at this stage is whether the data is reliable for the target variable.

Expected Steps:

- Describe data
- Explore data
- Verify data quality
- Data quality report

Data Preparation:

Data must be evaluated to see if it can be used to train a predictive model at this stage. It is essential to choose which features are most suited to predicting our target variable and which should be excluded from the model. At this stage, data must be examined to see if it can be used to train a predictive model. It's very critical to determine which features are best for forecasting our targeted variable and which should be left out of the equation.

Expected Steps:

- Conversion of data into a tabular format.
- Missing values treatment.
- Conversion of data in proper data type
- Normalizing the data

Modelling

In the modelling stage, select the modelling process. While you may have already chosen a tool during the business understanding stage, you will now choose one modelling technique, such as decision tree, Support Vector Machine, or neural network generation, at this point. If you're using multiple techniques, apply this task for each one separately.

Expected Steps:

- Test design generation
- Model building
- Model Assessment

Evaluation

Throughout this stage, the model needs to be examined how well the model fulfils the business objectives also try to figure out if there's a business reason why it's not working. If time and budget permit, another alternative is to test the model(s) on test applications in real-world applications. In addition to assessing any other data mining results created, the assessment step entails assessing any other data mining outcomes generated.

Expected steps:

- Review Model and Process
- Determine next steps

Deployment

In the deployment step, you'll take the conclusions of your evaluation and devise a rollout strategy. If a generic technique for creating the required model(s) has been determined, it is described here for later deployment. It makes it useful to think about deployment options at the business understanding phase as well because deployment is vital for a successful project.

Expected Steps:

- Plan for monitoring and maintenance.
- Make a final report
- Reviewing project

Data Description

Future Learn provided the following data for the analysis of the increase in the dropout rate for their Cyber Security course, Archetype survey response, enrolments information, leaving survey response, students' questions response, students' step activity, and team members managed the course, and video statistics. Only the leaving survey response data is used for analysis for this report. It is assumed that there were four intakes (i.e., November, February, June, September) for this course in the 'period of late 2017 and 2018. This data is from the students who registered in the Cyber Security online course were asked to fill out a survey for dropping out in the midst of the course. Around 400 students' records were analysed for patterns and to establish when and why they dropped out.

More information about Leaving response survey data:

- Id- Unique id for the records. (e.g. 212)
- Learner Id- Unique id for Students.
- Leaving reason- Reason selected by the student while filling out the exit survey. (e.g. I don't have enough time)
- Last completed step at- The timestamp when the student completed the last step before leaving the course. (e.g. 2017-11-25 12:33:05 UTC)
- Last completed week number- The week in which the student left the course. (e.g. 3)
- Last completed step number- Last completed step by the student before leaving the course. (e.g. 3.2)

```
> head(Final_data)
      id                              learner_id                  left_at
4   2831 5cbc8827-23de-415c-aa41-b3f2c9f28c0f 2017-11-27 17:51:03 UTC
7   5010 173449ed-285f-4b8c-8459-e270fd14eb5d 2017-11-29 18:00:23 UTC
8   5012 f5cb04e7-e09a-4cce-8f28-4a04b203f27a 2017-11-29 18:02:17 UTC
9   5459 5b4373c7-72cf-4675-8af6-c5589497d392 2017-11-30 01:17:18 UTC
14  7493 02fbf753-9bbe-4236-a0a6-a5d25f3d65f8 2017-12-01 13:33:27 UTC
20 11930 1139938f-3772-40c7-bee2-53a77f1cb878 2017-12-04 11:40:31 UTC
                            leaving_reason  last_completed_step_at
4        The course wasn't what I expected 2017-11-20 18:01:15 UTC
7                                    Other 2017-11-29 17:57:41 UTC
8  The course won't help me reach my goals 2017-11-28 21:01:21 UTC
9        The course wasn't what I expected 2017-11-19 22:02:09 UTC
14       The course wasn't what I expected 2017-12-01 13:31:56 UTC
20                                   Other 2017-11-17 18:28:42 UTC
   last_completed_step last_completed_week_number
4                 1.16                         1
7                 3.20                         3
8                 1.15                         1
9                 1.10                         1
14                3.20                         3
20                1.20                         1
   last_completed_step_number
4                          16
7                          20
8                          15
9                           1
14                         20
20                          2
>
```

Figure 2: Head of Final Data

The basic structure of the head is extracted using the head function.

Business Problems considered in this report

1. What were the most and least common reasons for students dropping out of the Cyber Security course? Week wise analysis for the most common leaving reason?

Business Understanding:

Here the business objective for the Future Learn team is to use the insights to figure out why students have dropped out in the midst of the course. This might help Future learn to understand the gap in the course, also they could be able to improve the content delivery of that course.

Data Understanding and Cleaning:

Data was provided in .csv files so, it could be open in MS excel, so the first level of understanding the data was performed in MS Excel. Data used for this analysis is "cyber-security-4_leaving-survey-responses.csv", "cyber-security-5_leaving-survey-responses.csv", "cyber-security-6_leaving-survey-responses.csv" and "cyber-security-7_leaving-survey-responses.csv". The Data was divided into four different files based on the intake, so it was required to merge the data into one single file. Also, the data needed to be cleaned, as some fields weren't useful for analysis. (For example: Leaving Reason - "I prefer not to say") So those records where the leaving reason was "I prefer not to say" required to be omitted. Later, there were certain fields in the data where the text was not in an acceptable format, so it had to be updated. (For example: "I don't have enough time")

The Pie chart to compare most and least common leaving reasons for students to have left the course. [Figure 3]

- To make the Pie-chart, the data was transformed from a table to a data frame, and the percentage of the Leaving reason variable was determined. For plotting the Pie chart "ggplot2" library was used.
- The pie chart [Figure 3] shows that the majority of individuals chose "Other" as their reason for leaving the course (i.e. 28.57 %), but no conclusions can be drawn from this. So, the next most selected reason "I don't have enough time" which was chosen by 24.40 % of students.

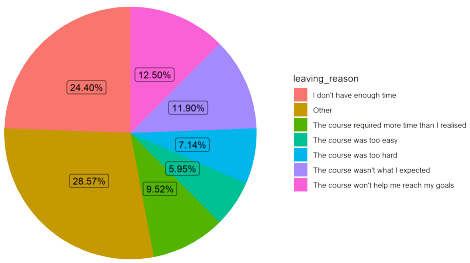

Figure 3: Pie chart for comparing the most common leaving reasons

The Bar-plot for comparison of leaving reason provided by students in 1st week of their course. [Figure 4]

- In order to compare the dropping out reasons given by students in the first week of their course, the data was filtered by the Last completed week number equal to 1, then the data was converted to Data frame to plot the Bar-plot using ggplot library.
- By looking at the [Figure 3] Bar-plot, it can be inferred that the most popular reason for leaving the course was "I don't have enough time," while "The course was too easy" was the least chosen reason. Also, the second most reason is "The course wasn't what I expected" so it can be interpreted that the most learner didn't like the content of the course and decided to drop out.

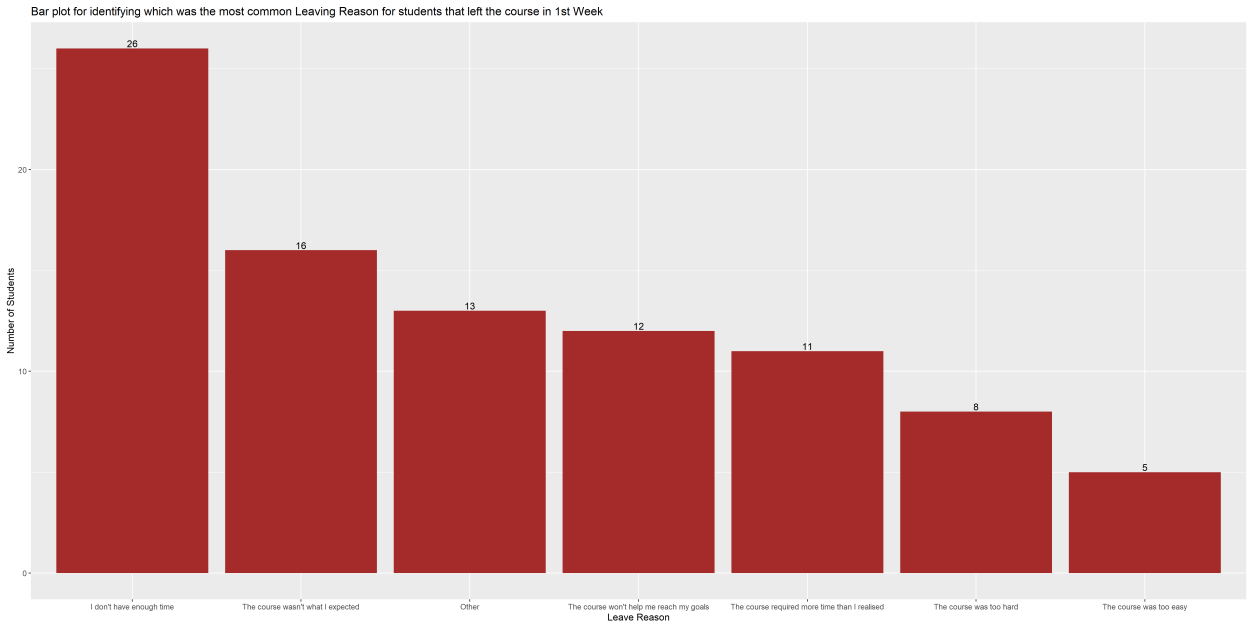

Figure 4: Comparing the leaving reason provided by students in 1st week of their course

The Bar-plot for comparison of leaving reason provided by students in 2nd week of their course.[Figure 5]

- In order to compare the leaving reasons selected by learners in the second week of their course, the data was filtered by the Last completed week number equal to 2, then the data was converted to Data frame to plot the Bar-plot using ggplot library.
- By looking at the [Figure 4] Bar-plot, it can be inferred that the most popular reason for leaving the course was "I don't have enough time," while "The course wasn't what I expected" was the least chosen reason.

The Bar-plot for comparison of leaving reason provided by students in 3rd week of their course. [Figure 6]

- The data was filtered by keeping the Last completed week number equal to 3, then transformed to the Data frame to plot the Bar-plot using the ggplot package in order to compare the reasons chosen by learners in the third week of their course.
- The most popular reason for leaving the course was "Other" according to the [Figure 5] Bar-plot, but there is no proper explanation about "Other", so it can not be concluded.

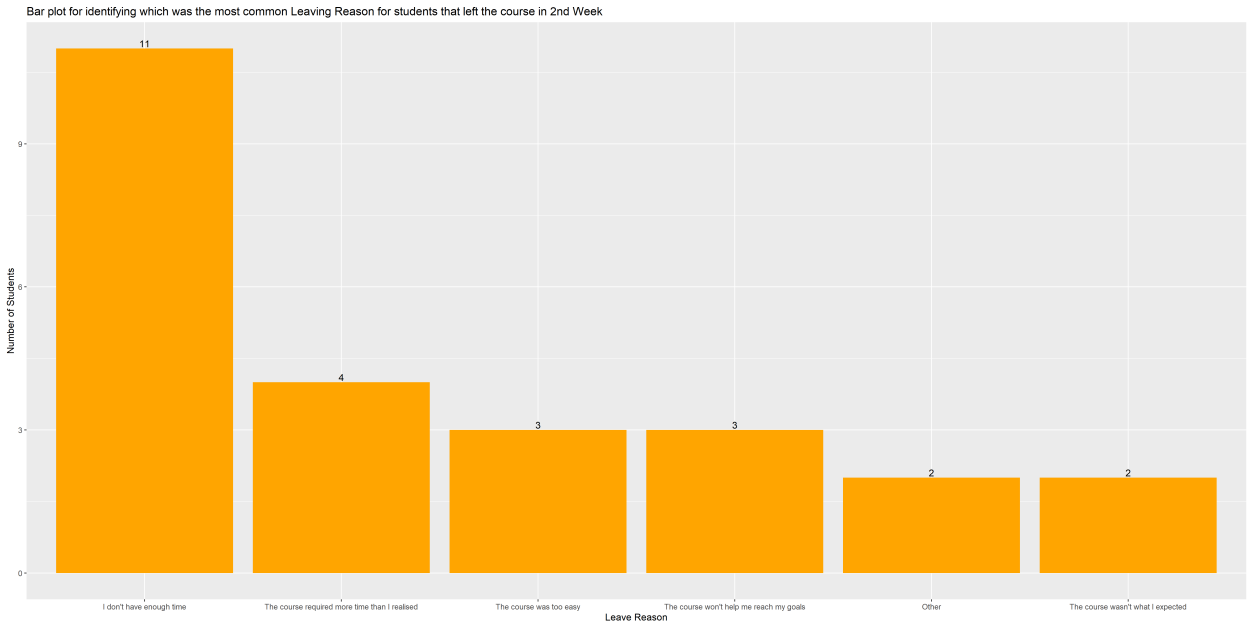

Figure 5: Comparing the leaving reason provided by students in 2nd week of their course

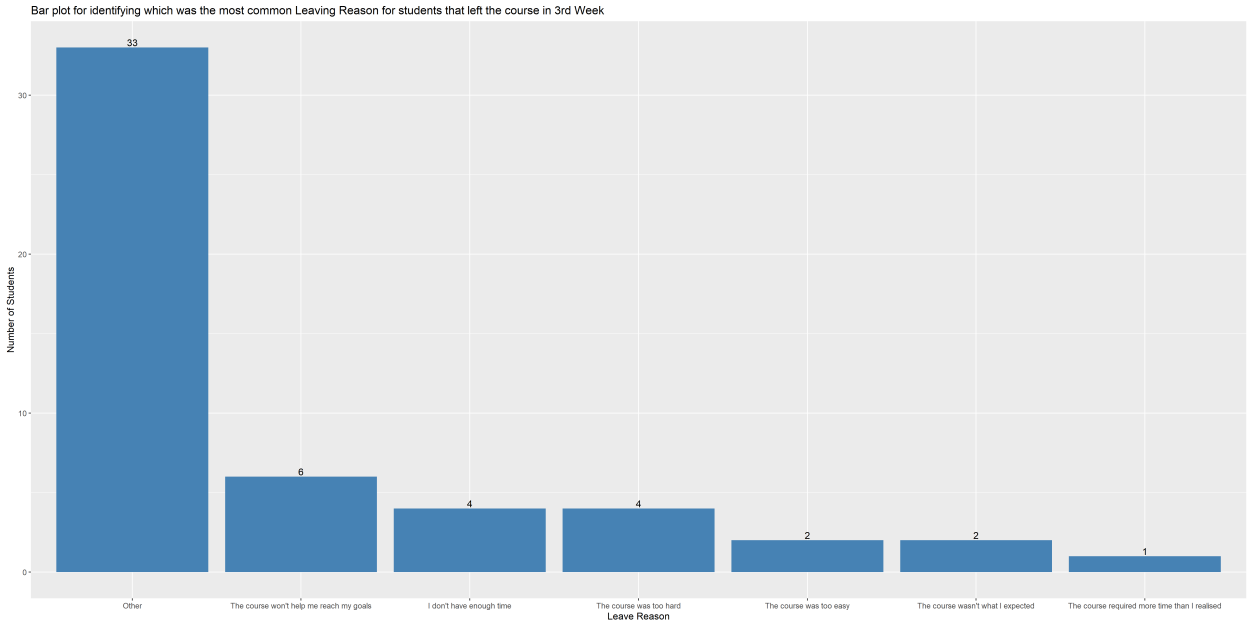

Figure 6: Comparing the leaving reason provided by students in 3rd week of their course

Data Preparation

Treatment of missing values: There were Null values in the data, that had to be omitted to keep the analysis biased.

```
> summary(Final_data)
       id            learner_id          left_at          leaving_reason     last_completed_step_at
 Min.   :   212   Length:403         Length:403         Length:403         Length:403
 1st Qu.: 50185   Class :character   Class :character   Class :character   Class :character
 Median : 93773   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :107492
 3rd Qu.:163906
 Max.   :221447

 last_completed_step last_completed_week_number last_completed_step_number
 Min.   :1.100       Min.   :1.00               Min.   : 1.00
 1st Qu.:1.200       1st Qu.:1.00               1st Qu.: 4.25
 Median :1.800       Median :1.00               Median :13.00
 Mean   :2.062       Mean   :1.78               Mean   :12.05
 3rd Qu.:3.188       3rd Qu.:3.00               3rd Qu.:20.00
 Max.   :3.900       Max.   :3.00               Max.   :23.00
 NA's   :217         NA's   :217                NA's   :217
>
```

Figure 7: Comparing the leaving reason provided by students in 2nd week of their course

Modelling

- Machine learning and statistics are mostly used in the modelling step. The first stage in the modelling process is to select approaches, which might include unsupervised algorithms. The model is examined in this part to produce correct predictions. Because the focus of this study is solely on data analysis, this stage is bypassed.

Evaluation

- At the evaluation stage, the entire data is split into training and test sets, the accuracy is computed. A training set will be used to install the model, and a test set will be used to put it through its paces. The evaluation may be used to determine whether these models meet the business objective. This stage is skipped because the focus of this investigation is only on data analysis.

Deployment

- The deployment approach includes an exploratory data analysis report and a final presentation for insights.

2. Which week did the most and least participants leave the Cyber Security course? Also, what was the most last common step completed by students before dropping out of the course?

Business Understanding

Future learn would be able to figure out when did most of their learners loose interest in the course or decided to drop out. Analyzing the last completed step can help Future learn to pinpoint the step where the most of users have left the course.

Data Understanding and Cleaning

Data was provided in .csv files so, it could be open in MS excel, so the first level of understanding the data was performed in MS Excel. Later, the summary function was used in R studio to view the summary of single files. Data used for this analysis is "cyber-security-4_leaving-survey-responses.csv", "cyber-security-5_leaving-survey-responses.csv", "cyber-security-6_leaving-survey-responses.csv" and "cyber-security-7_leaving-survey-responses.csv". The Data was divided into four different files based on the intake, so it was required to merge the data into one single file. Also, the data needed to be cleaned, as some fields weren't useful for analysis. (For example: Leaving Reason - "I prefer not to say") So those records where the leaving reason was "I prefer not to say" required to be omitted. Later, there were certain fields in the data where the text was not in an acceptable format, so it had to be updated. (For example: "I don't have enough time")

Bar-plot for comparing in which week did the most student left the course [Figure 8]

- Only the last completed week number data was filtered and, then transformed to a Data frame to plot the Bar-plot using the ggplot package in order to compare the time when learners left the course.
- It can be interpreted from [Figure 8] that most of the students have left the course in the third and final week, whereas the least students left in the second week.

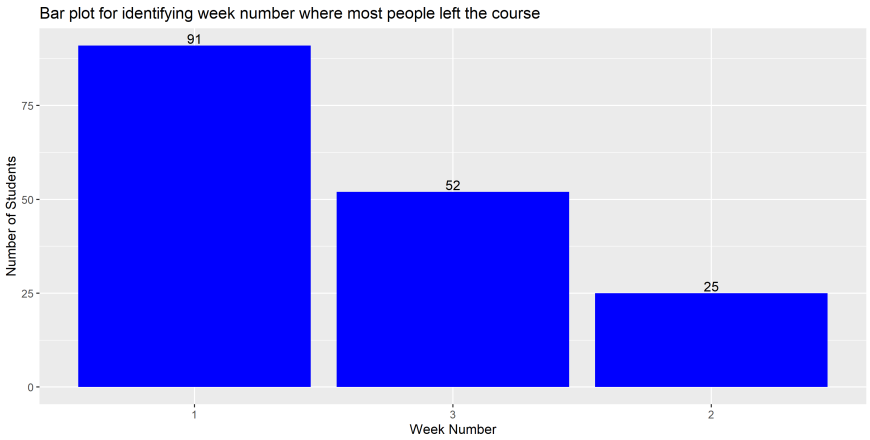

Figure 8: Identifying the week where most students left the course

Bar-plot for comparing Which was the most common Last step completed by students that left the course [Figure 9]

- Only the last completed step number data was filtered and, then transformed to Data frame to plot the Bar-plot using the ggplot package in order to compare at which step did most of the learners left the course.
- It can be interpreted from [Figure 9] that around 40 students have left the course at Step 3.20 (i.e.Wrapping up ARTICLE) and that was the final step of the course. Also, the second highest last completed step is 1.2(i.e Why are you here?) So it can be interpreted that at this step, most people understood the course outcomes and then decided not to continue with the course.

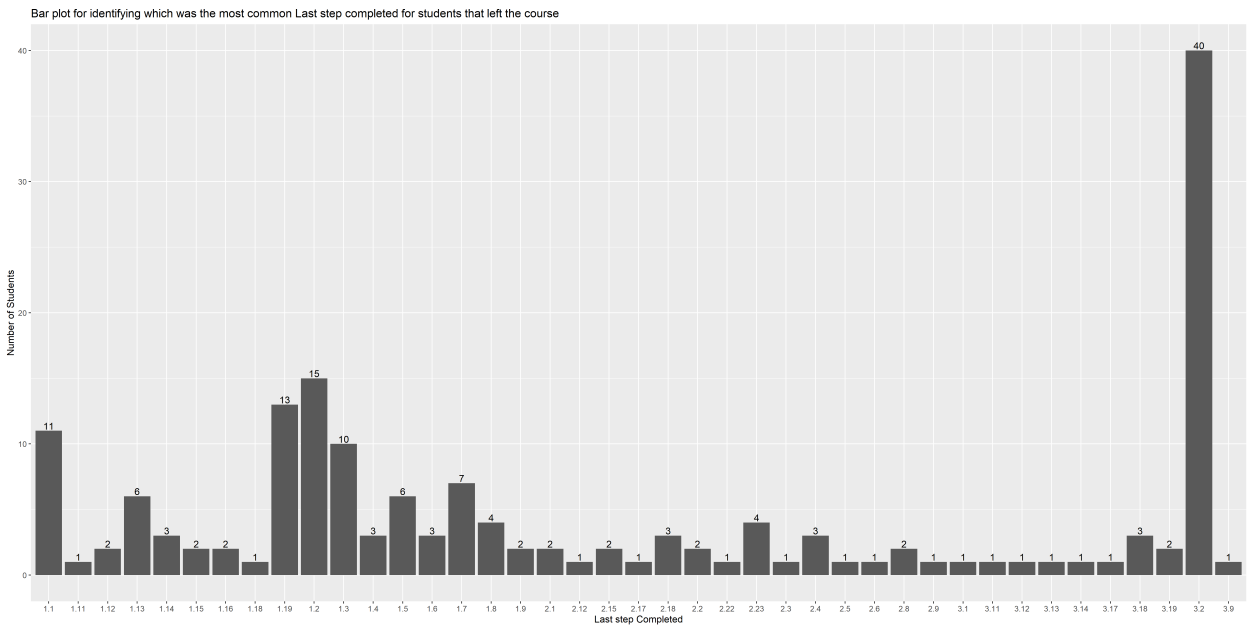

Figure 9: Identifying most common last step completed by students

Data Preparation

- Treatment of missing values: There were Null values in the data, that had to be omitted to keep the analysis biased.

Modelling

- In the modelling stage mostly machine learning and statistics are involved. The initial step in the modelling process is to identify techniques, which might include algorithms like regression and classification. To make accurate predictions, the model is analyzed in this section. This stage is skipped because the focus of this investigation is only on data analysis.

Evaluation

- The accuracy is calculated during the evaluation step, which divides the total data set into training and test sets. The model will be installed on a training set, then it will be put through its paces on a test set. It's possible to utilise evaluation to see if these models achieve the business goal. Because the focus of this research is solely on data analysis, this stage is bypassed.

Deployment

- This exploratory data analysis report and final presentation for insights is part of the deployment procedure.

Conclusion

- Based on the findings of both analyses, it can be inferred that the majority of participants completed steps up to 3.20 i.e. the final step, but for some reason did not complete the course. "Other" was the

most popular answer among those who dropped out of the course, according to the statistics. As a result, it's reasonable to presume that the "Other" cause is that the learners are failing to pass an quiz after completing all of the video lectures and steps.

- Therefore, they've chosen "Other" as a reason to leave. Also, the second highest last completed steps is 1.2(i.e Why are you here?) So it can be interpreted that at this step, most people understood the course outcomes and then decided not to continue with the course. Therefore in huge number of drop out rate in first week.
- So we can speculate that the reason for so many students dropping out of the course was the difficulty in quiz. It's possible that the students dropped out because the quiz was too difficult for them, but that's just speculation. An analysis was undertaken using data from the leaving survey responses, and it was observed that the majority of students chose "Other", and the second most popular answer was "I don't have enough time" so it means students didn't felt they should invest their time for this course. Also by observing the last completed steps analysis, it can speculated that most people completed the last step (i.e. Step 3.2) but didn't complete the course, so "Other" reason can't be neglected. Therefore, it can be concluded that learners found the quiz hard and could not clear it.

Reproducibility

- The all four datasets are merged as a single data frame for analysis, thus if any news files need to be reported, they must be manually modified in the pre-process file to merge the data in the munge subdirectory.
- The R-markdown is used to pre-process the data, compute the necessary functions, and then generate the report with the help of Project Template. If a new file is needed to produce the report, it should be placed in the data directory and include the same variable names as those indicated in the report to prevent future mistakes.
- The report's analysis is not replicable since certain manual interventions are necessary to update variable names and link the datasets together in this single data frame when a new data file emerges.
- Aside from that, when the report is set up to run from the R markdown report file in the Report subdirectory, the report will be automatically created.
- For analysis the datasets are combined as a single data frame and hence if any news files should be reported, it needs to be manually updated to the pre-process file to combine the data which would be in the munge subdirectory.

References

- Yaacob, W., Nasir, S., Yaacob, W. and Sobri, N., 2019. Supervised data mining approach for predicting student performance. Indonesian Journal of Electrical Engineering and Computer Science.
- Chen-Hsiang Yu, Jungpin Wu and An-Chi Liu 2019. Predicting Learning Outcomes with MOOC Clickstreams. https://www.mdpi.com/2227-7102/9/2/104/htm
- Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) and Rudiger Wirth (DaimlerChrysler) - CRISP-DM 1.0 https://the-modeling-agency.com/crisp-dm.pdf
- Data Science Process Alliance. 2021. CRISP-DM - Data Science Process Alliance. https://www.datascience-pm.com/crisp-dm-2/
- Tatiana Likhomanenkoa, Alexey Rogozhnikova,Alexander Baranova, Egor Khairullina mikari, Andrey Ustyuzhanin -Improving reproducibility of data science experiments https://indico.ijclab.in2p3.fr/event/2914/contributions/6476/subcontributions/168/attachments/6033/7158/Likhomanenko_REP_paper.pdf

- Brillinger, D., Preisler, H., Ager, A. and Kie, J., 2004. An exploratory data analysis (EDA) of the paths of moving animals. Journal of Statistical Planning and Inference, https://www.sciencedirect.com/science/article/abs/pii/S0378375803002404
- Taylor & Francis. 2021. An overview of learning analytics. [online] Available at: https://doi.org/10.1080/13562517.2013.827653
- Chatti, M., Dyckhoff, A., Schroeder, U. and Thus, H., 2012. A reference model for learning analytics. International Journal of Technology Enhanced Learning.
- What is the CRISP-DM methodology? - https://www.sv-europe.com/crisp-dm-methodology/
- The Data Mining Process (CRISP-DM) - https://www.nimblecoding.com/data-mining-process-crisp-dm/