Analysis on Cyber Security Learning course

Mayank Baraskar

Abstract

This study uses the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology and R to analyze a Massive Open Online Course (MOOC) on cyber security for the year 2018. Massive open online courses (MOOCs) have been widely employed in the field of education and have lately been promoted as a result of the COVID-19 epidemic. The primary goal of this research is to determine why enrollment in the course has decreased by examining their interactions as well as the completion rate of the course work. This study intends to better understand why students drop out of class and, as a result, improve the course process for the benefit of both students and professors. It is well established that different learners use learning materials in different ways. Some students, for example, tend to complete their course work, whereas others attempt to learn crucial concepts from the entire course work and abandon the rest. This study, on the other hand, can be utilized to figure out the most prevalent student learning patterns and reasons for dropping out. The study was conducted using data from students who enrolled in the Cyber Security online course in 2018. The course was offered in three distinct intakes during the year, in February, June, and September. The goal is to improve the course work for future intakes with the help of visualization of this analysis so that students can get better quality material and presentation of total course work with a lower leaving rate.

Introduction

The use of online learning and techniques such as Massive Open Online Courses (MOOCs) has increased significantly as a result of the COVID-19 epidemic. FutureLearn is a learning platform that allows students to enroll in any field course and complete their coursework online. Students can learn from home or anywhere else because the course materials are easily available online. They are not required to attend traditional classrooms.

There are many benefits to using this platform, the most important of which is flexibility, since students may replay content and look over course work anytime they choose. MOOC courses were created to improve learner performance and learning outcomes. When a student enrolls in a course, he or she is required to do course work, which includes notes, videos, and tests. However, course analytics have revealed that, despite these benefits, many students abandon the course halfway through and, even if they finish, do not perform well. Each intake, the completion rate decreases in tandem with the number of people enrolled. Because the number of students enrolled is bigger than in traditional classrooms, and they come from all over the world, it will be difficult for professors to grasp each student's condition. Because not every student will be comfortable with the course work, material, or any other reason, students make a choice either to drop out or refuse to enroll in future intakes and courses. (Yu, Wu, and Liu, 2019).

The goal of this study is to learn about students' behavior patterns during the course, using analysis to determine when most students are likely to drop out and what the most prevalent reasons are for dropping out. The investigation will be based on data from three different intakes of FutureLearn's cyber security course, which is one of the MOOC learning platforms. The year 2018 was chosen from the entire data set since it had the leaving responses for various intakes in 2018, as opposed to 2016 and 2017, which will improve the accuracy of interpreting the learner's comments.

Research method

We will mine data using the CRISP-DM approach in this investigation. It stands for cross-industry data mining methodology. It is a very adaptable, reliable, and well-proven methodology. It's a six-step cycle strategy made up of events with various goals. We can adjust the sequence of events based on the business requirements because it is highly adaptable. As shown in Figure 1, there are primarily six steps, each of which has a distinct purpose.

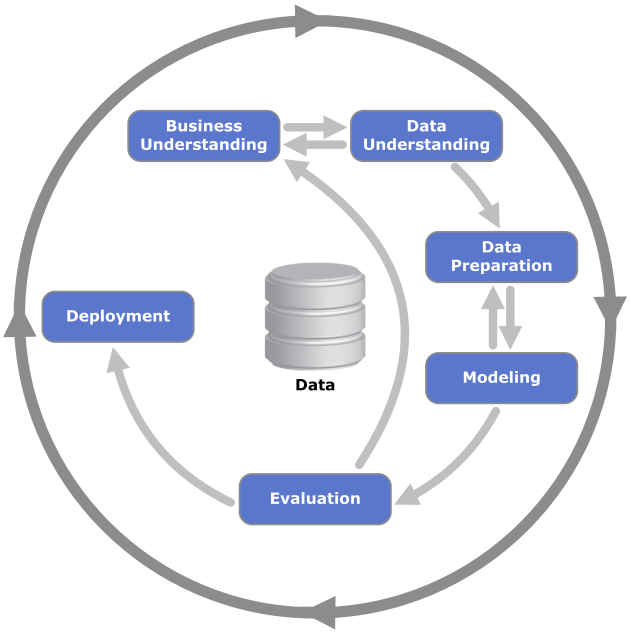

Figure 1: The CRISP-DM Model (Yaacob, Nasir, Yaacob and Sobri, 2019).

- **Business Understading**: This step focuses mostly on the business perspective and its requirements. After reviewing the situation, such as data availability, there are primarily two purposes for analysis in this study. The initial goal is to determine when the majority of students will be leaving the course. The second goal is to figure out why you dropped out of school. The overarching purpose is to do an explanatory data analysis on students and generate a report that will help institute enhance course content and increase general engagement.

- **Data understanding**: Understanding the data is the second phase in the CRISP DM process. The major goal of this stage is to identify, collect, and evaluate the data set that FutureLearn has provided to us. R was used to load the entire data set into Project Template. After the data was loaded, it was evaluated to determine the data format, number of records, and noise level in order to clean the data further. In this step, visuals for learning metrics were created, which FutureLearn can use to optimise the course process.

- **Data Preparation**: The preparation of data is the third process. Data munging is another name for it. The final data set for the visualisation challenge was prepared in this stage. Unnecessary data was removed from the whole data set. The chosen data was cleaned and assembled according to the specifications. Because the data was split into numerous files, it was combined, and joins were employed in some cases. The data was also formatted, with string numbers being converted to numeric values for mathematical calculations.
- **Modelling and Evaluation**: Machine learning and statistics are mostly involved in this step. The first phase in the modelling process is to choose methodologies, which might include algorithms such as regression, classification, and others. The model is fitted to the data in order to produce precise

predictions. The accuracy is calculated at the evaluation step, which divides the entire data set into training and test sets. The model will be installed into a training set and put through its paces on a test set. The decision can be made if these models met the business criteria via evaluation. However, because the focus of this study is on data analysis, neither of these stages will be included.

- **Deployement**: The report, which comprises the data analysis and the final presentation for the outcome, is prepared during the deployment step.(CRISP-DM - Data Science Process Alliance, 2021)

Data Description

Learners who enrolled in FutureLearn's cyber security course had their data utilised. Because there was insufficient data for analysis for the previous year, the data for the year 2018 was used, with three different intakes in the months of February, June, and September. Multiple variables were used for data analysis, and they are listed in the table below with descriptions. This definition can be used as a reference throughout the report because these attributes were employed throughout the research. This table can also show which data attributes were accessible and how they were used for analysis. The data properties included in the table are a comprehensive list; nonetheless, they are used in various data frames.

Sr No.	Attribute Name	Description
1	learner_id	Unique ID which is alloted to learner at time of enrollment
2	enrolled_at	Timestamp of when the student is enrolled
3	fully_participated_at	Timestamp for when the learner completed the course
4	left_at	Timestamp for when the learner left the course
5	leaving_reason	Reason for learner leaving the course
6	last_completed_step_at	Timestamp for when the learner completed the last step
7	fully_participated_at	Timestamp for when the learner completed the course
8	last_completed_step	Last completed step number
9	last_completed_step_week_no	The week number when learner completed the last step
10	last_completed_step_no	The step number when learner completed the last step

Analysis

As two business objectives have been specified in this article, there are two CRISP-DM cycles. The steps of CRISP-DM have been followed and discussed separately for each business aim.

Business Objective 1 - An analysis of when the majority of learners drop out of the course.

There are numerous topics that will be explored in this business purpose. The main goal is to figure out when the majority of students abandon the course halfway through. Despite the fact that MOOCs are flexible, allowing students to play content or complete course work whenever they choose, unlike traditional classrooms, students can opt to leave or unenroll from the course.
The procedure was advanced to the second level of Data understanding because our business aim for this cycle has been sorted. The data came from FutureLearn's records, which covered seven runs with various input and years. The procedure was advanced to the second level of Data understanding because our business aim for this cycle has been sorted. The data came from records provided by FutureLearn, which covered seven separate runs with different input and years. Students' enrollment, archetype survey responses, quiz responses, step activity, weekly attitude survey responses, leaving survey responses, video statistics, and team members were all part of the data for each intake. Because the goal of this business is to figure out when students choose to drop out, the focus is on enrollment data and survey responses. Unfortunately, there was no data from previous intakes for the departing survey, but there were responses

for the year 2018, thus this analysis will be done on three separate intakes from that year to gain a better understanding and learn learners behavioural patterns. The raw data of enrollments and survey responses was in the following format after collecting the data for 2018:

Enrollment Data

```
head(feb_2018_enrollments, 4)
```

```
##                           learner_id             enrolled_at
## 1 480d5ab0-b755-4ea7-8374-100c25b920c4 2018-03-26 12:41:18 UTC
## 2 46a4c71e-8819-4c5a-8164-2f786186b9fb 2018-01-02 20:57:49 UTC
## 3 afaf7031-a708-49f6-a823-37eb8e4bc721 2018-03-31 01:03:01 UTC
## 4 41ad5273-6853-43d7-b615-06f8e808cea1 2018-09-10 12:08:49 UTC
##             unenrolled_at    role fully_participated_at purchased_statement_at
## 1 2018-10-19 09:57:54 UTC learner
## 2 2018-10-19 10:31:15 UTC learner
## 3 2018-10-14 04:09:49 UTC learner
## 4                         learner
##    gender country age_range highest_education_level employment_status
## 1    male      GB     46-55                tertiary working_full_time
## 2  female      GR     26-35      university_masters full_time_student
## 3 Unknown Unknown   Unknown                 Unknown           Unknown
## 4 Unknown Unknown   Unknown                 Unknown           Unknown
##                   employment_area detected_country
## 1     it_and_information_services               GB
## 2 accountancy_banking_and_finance               GR
## 3                         Unknown               VE
## 4                         Unknown               --
```

There are 3544 observations of 13 variables in the enrollments data, as can be seen. For this data, the variables were cleaned and prepared as follows.

- **learner_id**: Because there are no null values in this variable, no cleaning or inclusion/exclusion was performed.
- **enrolled_at**: Because the values are timestamps of when the student enrolled and there are no NA or null values, no cleaning or inclusion/exclusion was necessary.
- **unenrolled_at**: Because the data will be left joined with the survey response which has left_at attribute, and unenrolled at variables appears to be inconsistent, this column was removed.
- **role**: Because the focus of this variable is on learner role rather than organisation admin, the data was filtered for role admin.
- **fully_participated_at**: This variable is needed to figure out what the completion rate was, so it will be included. However, there are NA values that can be assumed to mean that the learner did not complete the course, and if the value has a timestamp, the learner completed the course. This variable is transformed into a categorical variable with the categories "Completed" and "Not Completed."
  There are some variables that aren't needed because they're not part of our analysis for the business goal, so they've been left out. Due to unknown values and NA values, the data was also very inconsistent. They are *purchased_statement_at*, *gender*, *country*, *age_range*, *highest_education_level*, *employment_status*, *employment_area*, *detected_country*

After transforming and cleaning enrollment data, the subset looks like:

```
head(feb_2018_enrollments_final, 4)
```

```
##                             learner_id             enrolled_at    role
## 1 480d5ab0-b755-4ea7-8374-100c25b920c4 2018-03-26 12:41:18 UTC learner
## 2 46a4c71e-8819-4c5a-8164-2f786186b9fb 2018-01-02 20:57:49 UTC learner
## 3 afaf7031-a708-49f6-a823-37eb8e4bc721 2018-03-31 01:03:01 UTC learner
## 4 41ad5273-6853-43d7-b615-06f8e808cea1 2018-09-10 12:08:49 UTC learner
##   fully_participated_at
## 1         Not Completed
## 2         Not Completed
## 3         Not Completed
## 4         Not Completed
```

Leaving Survey Response Data

```
head(feb_2018_leaving_survey, 4)
```

```
##      id                             learner_id                 left_at
## 1 34003 8853543d-b930-43e0-a3cb-99de6b9ea4f3 2018-01-30 20:04:59 UTC
## 2 38604 b170480c-7ed6-434f-8827-73cfa130ece1 2018-02-05 03:01:39 UTC
## 3 39016 92b8485e-b2a6-4345-8721-3e93142b469d 2018-02-05 10:30:03 UTC
## 4 39241 f4cae359-0966-4a6d-830f-4218861f2fd9 2018-02-05 13:27:14 UTC
##                                leaving_reason last_completed_step_at
## 1                          I prefer not to say
## 2 The course required more time than I realised
## 3                                       Other
## 4                                       Other
##   last_completed_step last_completed_week_number last_completed_step_number
## 1                  NA                         NA                         NA
## 2                  NA                         NA                         NA
## 3                  NA                         NA                         NA
## 4                  NA                         NA                         NA
```

There are 173 observations of 8 variables in the enrollments data, as can be seen. For this data, the variables were cleaned and prepared as follows.

- **id**: This variable is for storing unique ID while recording leaving survey, however this will be excluded from the filtered data
- **response_id**: Because there are no null values in this variable, no cleaning or inclusion/exclusion was performed. However, this variable was used for joining enrollment and leaving survey response data
- **left at**: Because the values are timestamps of when the student left the course and there are no NA or null values, no cleaning or inclusion/exclusion was necessary.
- **leaving reason**: This variable contains the reasons for the learner's withdrawal from the course. For the purposes of accurate analysis, we will omit the complete record for some NA values. Furthermore, when the reason was recorded, the string was not converted, resulting in responses like "The course wasn't what I expected," which were transformed to "The course wasn't what I expected" for better visualisation.

For better visualisation and accuracy, we omitted complete rows for some variables that contained NA values. They are *last_completed_step_at*, *last_completed_step*, *last_completed_week_number*, *last_completed_step_number*

After transforming and cleaning enrollment data, the subset looks like:

```
head(feb_2018_leaving_survey_final, 4)
```

```
##                              learner_id                left_at
## 6  a04b5f29-c283-419f-b262-5797c9220aa9 2018-02-05 18:12:27 UTC
## 11 3ae76b85-153a-4fe6-8a24-46d795e2608c 2018-02-06 20:00:27 UTC
## 14 f6ec6af3-d021-4422-a7c7-82012145f590 2018-02-07 12:49:19 UTC
## 19 5086b814-3121-43e7-999b-926113ae98d2 2018-02-09 18:31:44 UTC
##                         leaving_reason  last_completed_step_at
## 6                                Other 2018-02-05 18:12:15 UTC
## 11 The course wasnâ\200\231t what I expected 2018-02-06 19:04:22 UTC
## 14             The course was too easy 2018-02-06 11:17:02 UTC
## 19                 I prefer not to say 2018-02-09 18:24:29 UTC
##    last_completed_step last_completed_week_number last_completed_step_number
## 6                  1.2                          1                           2
## 11                 1.1                          1                           1
## 14                 1.2                          1                           2
## 19                 1.3                          1                           3
```

The data preparation and data understanding are done simultaneously because the study is based on the CRISP-DM model for explanatory data analysis and the main objective is to visualise and explore the data rather than modelling it. Now that the final enrollment and survey response data is clean and filtered according to the business objective, the left join was used to merge the data from both tables. Because a left join was used after merging the data and not all of the students responded to the learning survey, the NA values were again omitted throughout the data. After merging the data, this were the columns which were created:

```
colnames(feb_2018_merged_data)
```

```
##  [1] "learner_id"                "enrolled_at"
##  [3] "role"                      "fully_participated_at"
##  [5] "left_at"                   "leaving_reason"
##  [7] "last_completed_step_at"    "last_completed_step"
##  [9] "last_completed_week_number" "last_completed_step_number"
## [11] "intake"
```

Because the data set for all of the intakes was quite similar, the entire data preparation and cleaning process was repeated separately for June and September intakes. Because the data set was similar for all intakes, the entire data preparation and cleaning process was repeated separately for June and September intakes. In addition, the datasets from February, June, and September were concatenated, yielding 152 observations across 11 variables. As a result, we were finished with the data understanding and data preparation and were ready to move on to the explanatory data analysis. (Ford, 2021).

References

- Preidys, S.; Sakalauskas, L. 2010. Analysis of students' study activities in virtual learning environments using data mining methods, Technological and Economic Development of Economy 16(1): 94–108.
- Yu, C., Wu, J. and Liu, A., 2019. Predicting Learning Outcomes with MOOC Clickstreams. Education Sciences, [online] 9(2), p.104. Available at: https://www.mdpi.com/2227-7102/9/2/104/htm.
- 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, 16(3), p.1584.

- Data Science Process Alliance. 2021. CRISP-DM - Data Science Process Alliance. [online] Available at: https://www.datascience-pm.com/crisp-dm-2/ [Accessed 2 December 2021].
- Ford, C., 2021. Data Wrangling in R: Combining, Merging and Reshaping Data. [online] Clayford.github.io. Available at: https://clayford.github.io/dwir/dwr_05_combine_merge_rehsape_data.html [Accessed 3 December 2021].