Analysis on Cyber Security Learning course

Mayank Baraskar

26/11/2021

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.

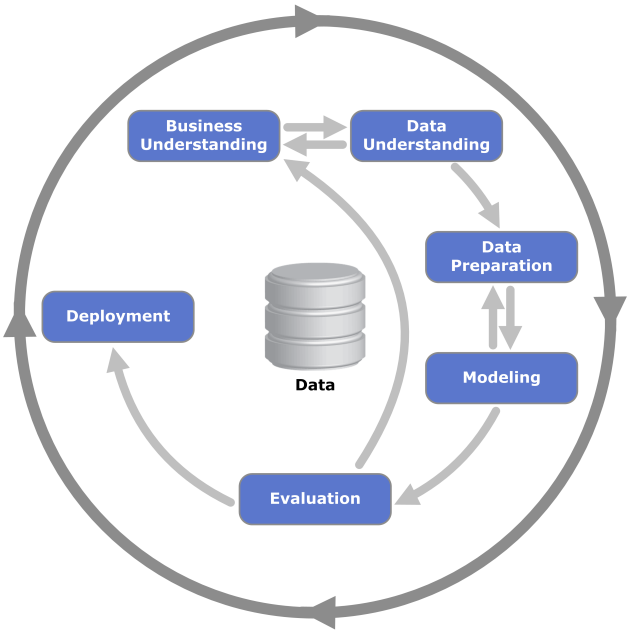

Figure 1: The CRISP-DM Model (Yaacob, Nasir, Yaacob and Sobri, 2019).

- **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:

```
summary(data_csv[[1]])
```

```
##        id           learner_id        responded_at        archetype
##  Min.   : 669320   Length:326         Length:326         Length:326
##  1st Qu.:1218994   Class :character   Class :character   Class :character
##  Median :1274975   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1280009
##  3rd Qu.:1376160
##  Max.   :1858643
```

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].