Massive Open Online Course (MOOC) ‘Cyber Security: Safety At Home, Online And in Life’ Analysis: 2016 - 2018

Tejas Satish Navalkhe

2023-11-05

Introduction

The purpose of this report is to offer insights into the Massive Open Online Course (MOOC) titled “Cyber Security: Safety at Home, Online and in life” created by Newcastle university. This course was made accessible to the public through the online skills provider FutureLearn from 2016 to 2018. This report is likely to offer value to any businesses or professionals offering the same online course or considering the launch of such a course. The results presented in this report were generated using the Cross Industry Standard Process for Data Mining (CRISP-DM) framework, with two model cycles employed to produce the subsequent findings.

First CRISP-DM cycle

Business Understanding

The initial phase of this analysis involves defining business objectives, identifying stakeholders, establishing clear goals for outcomes, and setting success criteria for evaluation. Subsequently, a plan has been developed to align with the CRISP-DM framework, taking into account the criteria outlined in this initial phase of the process.

Objective

This report aims to offer insights into the Massive Open Online Course (MOOC) titled “Cyber Security: Safety at Home, Online, and in Life,” developed by Newcastle University. The goal is to identify any noteworthy trends or patterns that might be valuable for professionals or businesses involved in offering a similar online course or contemplating its launch. The intention is for stakeholders to leverage these results to inform and shape their business decisions. Specifically, the data presented in this report is likely to be of significant interest to professionals or businesses akin to those on which this research is centered-those providing a comparable online cyber security course or considering its introduction. These findings can serve as a reference point for maximizing profits and achieving success in the current era of online courses.

Success Criteria

Initially, this research should employ data that is as precise as possible to generate results that can be actual for publication and relied upon to inform business decisions effectively. The outcomes must align explicitly with the set objective, addressing the research questions and presented in a format easily accessible to stakeholders, facilitating easy interpretation with minimal exertion.

Initial Research Question

With the goal in focus, the initial question that this report seeks to address is:

> **What is the employment status of learners enrolled in this course?**

Data Understanding

Data Understanding is crucial for maximizing the likelihood of achieving the objectives and generating valuable results. This is second phase of the process, I evaluate the data requirements in alignment with the objectives, assess the availability of data to fulfill these requirements, and evaluate the reliability of the selected data source. Data source is in this case is FutureLearn itself. Following the collection of the initial dataset, a review of business objectives is undertaken to ensure that the set goals are realistic in consideration of the available data.

Data Collection

The data utilized for this research was obtained from FutureLearn as participants advanced through the course. FutureLearn serves as the platform for online skill provision. The reports were accessible in CSV format covering the years 2016 to 2018. The data is collected from 7 runs meaning we have 7 different files of data.

Exploring the data

Once the dataset is collected, the subsequent step included in the exploration of the data. This involved identifying potential quality issues, scrutinizing the variables and their types, and assessing disparities between the data requirements and its availability. This initial screening process aimed to minimize the risk of investing time in a dataset unsuitable for analysis needs, offering a systematic approach to confirm its feasibility and capacity to meet the specified requirements.

Within this dataset, there exist various files containing information about learners enrolled in the Massive Open Online Course (MOOC) "Cyber Security: Safety At Home, Online, And in Life." This report will make use of these files. During the data exploration process, it became evident that diverse variables within these files contain information about the learners. Our dataset contains numerous unknown values, necessitating interpolation before proceeding further.

In summary, the raw data comprised the following sets of variables for 7 runs of enrollment files:

- learner_id
- enrolled_at
- unenrolled_at
- role
- fully_participated_at
- purchased_statement_at
- gender
- country
- age_range
- highest_education_level
- employment_status
- employment_area
- detected_country

Following a thorough exploration and assessment of the data's alignment with the analysis requirements, it was determined that there were abundant intriguing and informative data points available. These findings indicate that the analysis can be conducted within the defined objectives and success criteria. I delved deeper into the data and initiated the formulation of steps for the data preparation phase.

Data Preparation

Moving on to the subsequent stage of the CRISP-DM model, we enter the data preparation phase. This segment encompasses the cleansing, transforming, and selection of data in readiness for the upcoming modeling phase. This process contributes to the generation of more reliable results by minimizing the risk of errors. Additionally, the normalization of data across the board ensures a uniform and orderly format as we transition into the modeling phase, facilitating easier analysis for both myself and anyone seeking to validate the results in the future.

Data Cleansing

Data cleaning step is crucial for quality data. There are five main characteristics of data which are `Validity`, `Accuracy`, `Completeness`, `Consistency`, and `Uniformity`. The first step involved a glimpse of our dataset. As there are 7 different files of enrollments, we looked at each files separately. The munge file, utilized to execute the subsequent steps, is available under '01-A.R.' In this phase of the process, I noticed that all columns initially had a character data type; subsequently, I converted certain variables into a factor type to reflect their ordinal categorical nature. Later, I found that there are no duplicates in our dataset. All numeric variables were rounded to 3.d.p to ensure a consistent degree of accuracy throughout.

Data Wrangling

After finishing data cleaning stage, I started working on the transformation phase. The procedures delineated in this phase are encapsulated in the munge file labeled '01-B.R'. During this stage of the CRISP-DM process, I employed helper functions to detect unknown values, select and filter only required columns to form a new dataset. This enabled the re-usability of these functions for all the 7 runs / files that lacked the specified parameter. Adopting this strategy enhances the code development efficiency for this task. It facilitates seamless adjustments in the future.

In this phase of the CRISP-DM process I utilised `dplyr` library along with the pipe operator (%>%) to exclude any entries with unknown values. This library also improves the readability of the code for anyone evaluating the analysis and code in future. In the next step for this phase was to filter only required and necessary columns (variables) have been chosen for the sake of analysis convenience. Lastly, I consolidated all data containing learners information from enrolment dataset / data frame for this analysis report. I merged these data sets into single 'Master' data frame for ease of use in munge folder under '01-C.R'.

Modelling

After concluding the data preparation, I revisited the previous phases to confirm adherence to the plan and assess potential impacts on the upcoming analysis output. Finding no issues, I advanced to the most captivating phase of the CRISP-DM process: conducting exploratory data analysis on the now well-organized data frame.

Exploratory Analysis

In this stage, I systematically examine the parameters of the data to diagnose its fundamental characteristics. This method enables me to unearth valuable pieces of information that warrant further exploration in subsequent stages of more detailed analysis. Additionally, these findings contribute to the second iteration of the cycle.

Number of Employment Status

Let's compile an overview of the overall employment status by generating a table to promptly observe any notable trends.

S. No.	Employment Status	Count	Percentage (%)
1	Working Full Time	1352	42.839
2	Self Employed	375	11.882
3	Working Part Time	349	11.058
4	Looking For Work	293	9.284
5	Retired	291	9.221
6	Full Time Student	277	8.777
7	Not Working	114	3.612
8	Unemployed	105	3.327

Table 1: Number of Employment Status

Upon examining this table, it became evident that the highest three categories of enrolled learners were those classified as `Working Full Time`, `Self Employed`, and `Working Part Time`. The visualization representing this analysis (Percentage of Employment Status) in the form of a Pie chart looks as follows:

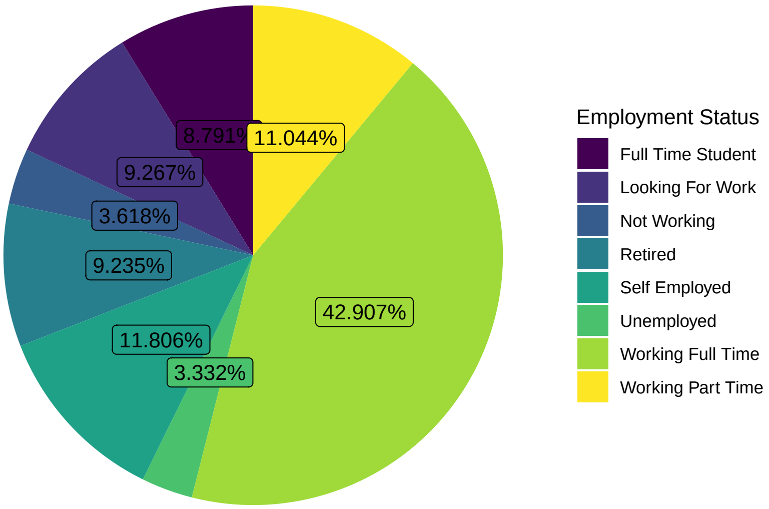

Figure 1: Percentage of Employment Status

The figure above makes it evident that the majority of learners are Full-Time Employees, followed by those who are Self-Employed, Part-Time Employees, and finally, individuals actively seeking employment.

Cycle 1: Evaluation

This leads us to the concluding component of the initial phase – a comprehensive review of the progress made thus far. This phase involves assessing the outcomes in relation to their alignment with the initially set objectives. It entails identifying whether the success criteria have been fulfilled, pinpointing areas for potential improvement in the analysis or data, and determining the direction for subsequent rounds of investigation.

Up to this point, the conducted analysis has successfully addressed the primary objective of answering the initial research question. It unveiled the three predominant categories of enrolled learners: Full-Time Employees ranked first, succeeded by Self-Employed individuals, Part-Time Employees, and, lastly, individuals actively seeking employment.

Regarding the quality of the data, I observed that the data employed has provided a satisfactory overview of the market trends for the course. Providing results that can be relied on to convey the general narrative of the online sector for companies at the top end of the market or Considering the launch of the same course. The outcomes of the analysis are presented in a format that is both accessible and straightforward to interpret.

To enhance course sales and profitability, stakeholders should communicate the idea of targeting marketing advertisements specifically towards individuals who are Full-Time Employees, Self-Employed, Part-Time Employees, and those actively seeking employment to the Marketing Team.

Ultimately, a comprehensive discussion of the subsequent steps in the process unfolds in the upcoming second CRISP-DM cycle, commencing with a thorough review of the business objectives.

Second CRISP-DM cycle

Business Understanding

Based on the outcomes of cycle 1, since there were no hindrances in conducting the research as initially planned, the overarching objectives of this study will persist. The primary aim is to extract valuable insights to assist businesses in this sector in achieving greater profitability, achieved through the observation of the trend and patterns in the data.

This brings me to formulating the subsequent research question to direct the analysis in the second cycle:

From which employment sectors do the majority of our learners originate?

This second research question is fitted in light of the findings from the initial round upon grouping dataset by Employment Status, we observed the highest number of learners enrolled in this course. Likewise, understanding the potential impact this might have on their Employment Area/Sector is of significant interest to both myself and the stakeholders outlined in this report.

Given that the success criteria and objectives have been satisfactorily achieved up to this point, and a research question has been formulated within the established parameters, there is no requirement to alter the scope of this analysis at this stage. Therefore, I proceed with the current scope and move onto review the data.

Data Understanding Review

In this subsequent stage, I evaluate the data quality in light of the outcomes from the initial round. Additionally, I assess how the data requirements have evolved in response to the newly articulated research question, contemplating adjustments in the data derivation process or the necessity of incorporating new sources to enhance accuracy and expand the potential avenues for meeting the business objective.

Following a comprehensive examination of the data during the initial analysis round, coupled with results that effectively fulfill the objectives, the quality of the current data can be confidently deemed suitable for the requirements of this upcoming analysis round. Further, there is the guarantee of the accuracy of the data because the course provider and the data provider both are same which is FutureLearn. Contemplating the newly designed research question, the existing data is useful for this analysis of the question. Having concluded that, I deemed the data suitable for the second cycle and moving ahead with the analysis. Since, the data is suitable for this analysis, we skip Data Preparation step, while moving ahead to Modelling phase.

Modelling

Number of Employment Area / Sector

Let's summarize the overall employment area by creating a table for a quick overview of any significant trends.

S. No.	Employment Status	Count	Percentage (%)
1	IT And Information Services	753	23.859
2	Teaching And Education	476	15.082
3	Engineering And Manufacturing	240	7.605
4	Health And Social Care	206	6.527
5	Public Sector	205	6.496
6	Business Consulting And Management	182	5.767

S. No.	Employment Status	Count	Percentage (%)
7	Accountancy Banking And Finance	136	4.309
8	Charities And Voluntary Work	131	4.151
9	Law	98	3.422
10	Creative Arts And Culture	97	3.074
11	Retail And Sales	88	2.788
12	Marketing Advertising And PR	78	2.471
13	Science And Pharmaceuticals	71	2.250
14	Media And Publishing	69	2.186
15	Hospitality Tourism And Sport	58	1.838
16	Armed Forces And Emergency Services	55	1.743
17	Transport And Logistics	54	1.711
18	Energy And Utilities	44	1.394
19	Environment And Agriculture	41	1.299
20	Property And Construction	41	1.299
21	Recruitment And PR	23	0.729
##	Total	3187	100.00

Table 2: Number of Employment Area / Sector

After examining this table, it became clear that the top five categories of enrolled learners are associated with the following areas/sectors: `IT And Information Services`, `Teaching And Education`, `Engineering And Manufacturing`, `Health And Social Care`, and `Public Sector`.

Cycle 2: Evaluation

That marks the completion of the second and concluding cycle of the CRISP-DM process in the analysis of the Massive Open Online Course (MOOC) “Cyber Security: Safety At Home, Online And in Life” Course. In the below assessment, I assess the effectiveness of the analysis.

This analysis proves valuable for businesses and professionals currently offering a similar course or contemplating the launch of such a program. The findings can be effectively employed by the organization’s Marketing Team to attract more enrollees. Utilizing online marketing platforms such as Facebook and YouTube in conjunction with this analysis can contribute to the enrollment drive.

To boost enrolments in the Cyber Security Course, the organization should target individuals in key sectors like ‘IT And Information Services,’ ‘Teaching And Education,’ ‘Engineering And Manufacturing,’ ‘Health And Social Care,’ and the ‘Public Sector,’ specifically focusing on ‘Full-Time Employees,’ ‘Self-Employed,’ ‘Part-Time Employees,’ and individuals actively ‘seeking employment.’ The identified employment types and sectors emerge as crucial insights from this report for enhancing sales.

Taking into account the success criteria, the findings of this analysis have been showcased in a format that is both easily accessible and comprehensible. This format was refined after the initial round to enhance the clarity of interpretation for the utilized graphics and the presented results. I am confident that the outcomes of this report can be regarded as successful when measured against the predefined criteria and objectives.

The outcomes of this report have produced valuable insights that can be confidently utilized and shared among the intended stakeholders and professionals. The inclusion of graphical evidence supports the assertions, presented in a format that is both simple and informative. With all these considerations, I am of the belief that this analysis can be deemed successful, delivering value to those who engage with the narrative embedded within the data of the FutureLearn Massive Open Online Course (MOOC) “Cyber Security: Safety At Home, Online And in Life” Course.

Final Phase: Deployment

The ultimate stage in this CRISP-DM process involves implementing my discoveries to ensure that the intended stakeholders can access the results. To achieve this, I have generated both a report and a presentation to convey the insights gained from this analysis. These materials emphasize the key aspects identified in the data that, in my belief, will offer the most valuable insights.

The elements selected for emphasis in the presentation comprise the Percentage of Employment Status and a table showcasing the foremost five employment areas/sectors.

I opted to incorporate these visuals because I believe they most effectively convey the overarching narrative within this data—information that will be of great interest and benefit to the target stakeholders. These info-graphics serve as the most efficient means to encode this information, offering a simplified summary of the comprehensive picture contained in the data. This approach facilitates easy decoding and interpretation of the information, enabling stakeholders to dedicate more time to contemplating the implications rather than struggling to derive meaning from the visuals.

References

- CRISP-DM Cycle
- Replace Column Value
- Lesson 6. Add Images to an R Markdown Report - Earth Data Science
- Count the observations in each group - dplyr Tidyverse
- Data Cleaning - Tableau
- Data Transformation
- Data Visualisation