CSC8631 Report

Selina So

06/11/2021

Business Objectives

Background

Learning Analytics is a study of the "*measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the information system in which it occurs*(Shi, 2018)". Learning analytics will provide insights to the factors which influences learners retention. This will therefore enable course designers from educational institutes and MOOC (Massive open online course) providers to make informed decisions on the design and improvements of their courses. Consequently improving the learning environment for learners and drive more influx of learners enrolling.

FutureLearn is an MOOC provider, which collaborates with universities globally to offer online courses. Since their launch in 2013, they have attracted over seven million learners across the world (www.futurelearn.com). With the insights driven from Learning Analytics, it will help FutureLearn on how to improve the retention rate and learners' engagement.

Business objective

There are many factors which could influence the learners' retention rate. In this study, the focus will be on the video lectures provider by FutureLearn, which are generally used to form part of a course. (why important from a business perspective, what are the problems to solve - financially strained, etc)

This study will examine the Cyber Security online course, which is divided into three weel;y blocks of study. The course consist of a combination of videos, articles, exercises, discussions, quizzes and tests.

There are a number of steps for each weekly block to complete. The first week block contains 18 steps, and the second and third week blocks contains 21 steps. (Shi, 2018)

Inventory of Resources

The CRISP-DM methodology (Cross-Industrie Standard Process for Data Mining) will be applied to achieve the objective of this study (link the CRISP-DM guide). The key areas of focus from the process are Business Understanding, Data Understanding, Data Preparation and Evaluation.

Data Mining Goals

For this study, we will investigate the videos from the course to answer the following questions:

1. Does the duration of the videos have an impact on the viewing rates across different continents?
2. Does the content of the videos have an impact on the viewing rates across different continents?
3. Is there a correlation between duration of videos and drop out rate

to insert reference use this notation [@RN22] - this is not working, try later.

Data Understanding

The raw data was provided by FutureLearn on their Cyber Security course. There are seven runs of data, each run of data were measured several months apart from each other. There were no descriptions for the data, therefore assumptions will be made as to what the data means.

As the study is on the video content, therefore the datafiles with the title containing 'video-stats' would be used. There are only 5 (out of 7) runs, which contains the video-stats datafiles. Therefore runs 1 and 2 will be eliminated from this study as no data are available.

Below is the list of column names in the data.

```
names(run3)
```

```
##  [1] "step_position"                  "title"
##  [3] "video_duration"                 "total_views"
##  [5] "total_downloads"                "total_caption_views"
##  [7] "total_transcript_views"         "viewed_hd"
##  [9] "viewed_five_percent"            "viewed_ten_percent"
## [11] "viewed_twentyfive_percent"      "viewed_fifty_percent"
## [13] "viewed_seventyfive_percent"     "viewed_ninetyfive_percent"
## [15] "viewed_onehundred_percent"      "console_device_percentage"
## [17] "desktop_device_percentage"      "mobile_device_percentage"
## [19] "tv_device_percentage"           "tablet_device_percentage"
## [21] "unknown_device_percentage"      "europe_views_percentage"
## [23] "oceania_views_percentage"       "asia_views_percentage"
## [25] "north_america_views_percentage" "south_america_views_percentage"
## [27] "africa_views_percentage"        "antarctica_views_percentage"
```

There are 13 rows for each datafile, one row corresponding to each video content throughout the course.

There are 28 columns, and a combination of columns will be selected for particular analysis.

The dataset is very complete with no visible missing data.

The data are mostly continuous data, other than for the first two columns, which shows the step of where the video content is located at and the title of the video content.

Currently no columns will be removed as all columns may contain relevant information for the study.

Number of viewers per video

Figure XX shows the number of viewers per video, throughout the course. Fix the labels and maybe titles

Finding pairs of relationships

A scatterplot matrix will be used to visualize any pairs of relationships amongst the data.

Percentage viewed whole duration of videos

We considered just the number of learners watching each video for the whole duration of the video, using the earliest video dataset (based on other datasets of run 3, this appears to roughly cover the time period between Jul 2017- Nov 2017).Figure XX shows the number of learners who have watched the videos for the whole duration.

Data Preparation

The 'step_position' shows the step within the course, of the location of the video. This column is listed as numerical, therefore will have to be changed to character.

The data pre-processing codes are located in the 'munge' folder.

Recommendation

May have to do some more analysis to compare with other MOOC courses

References