Assessment Report

Jack Westmoreland

2024-11-01

idea: - check how each archetype interacts with videos, questions, steps etc. - Are question results a good indicator of the sutudents overall performance? - compare video watch time to question response accuracy *** - which archetypes werte best at quizzes?

Introduction

Newcastle University have completed seven runs of Cyber Security: Safety At Home, Online, and in Life, A MOOC (massive open online course) teaching cyber security concepts to the public. This report aims to explore the data collected from these runs to be able to provide important analytics and valuable insights into online learning environments and further optimise online teaching. To achieve this, two cycles of CRISP-DM will be completed hopefully providing valuable insight into the provided data set.

CRISP-DM Cycle 1

Business Understanding

This section of CRISP-DM entails defining a problem we would like to solve for our data set, then setting the tasks and success criteria, which must be completed for the problem to be solved.

Objective

With the prevalence of online learning increasing in recent years (Cambridge Home School Online 2024), being able to predict learning outcomes of students from the data we can collect could prove very useful for online educators. Being able to predict learner outcome from their interactions with online courses could allow for educators to "check up" on students who have poor predicted performance; providing further guidance and support, increasing learner outcomes. Therefore being able to find some predictors of learner outcomes would be beneficial for Newcastle University's online education programs.

Success Criteria

For this EDA to succeed a strong predictor of learning outcomes should be identified from the provided data set. This predictor must be measurable for students before the completion of the course to allow for intervention. Furthermore, ideally the data should be measurable for each student individually so that each students predicted performance can be personalized to them. In a simple sentence the goal of this CRISP-DM cycle can be:

"Can we predict individual student performance from their data?"

Data Understanding

In this phase of CRISP-DM I will evaluate the data we have been given, considering its usefulness in completing the task set about from the Business Understanding step. The data will have its reliability considered from

what we know about how it was collected. A close analysis will also be conducted to check what data is avialable, considering how it could be used to perform a successful CRISP-DM cycle.

Data Collection

The FutureLearn MOOC data set has been provided by Newcastle University. It consists of several CSV files containing data on student performance and interaction with the online material for each of the seven runs. Each run consists of 5+ CSVs each containing data on ways students have interacted with the program.

Each run has near identical data being recorded such as x_enrollments, where x is the run of the program. These data sets exists seperately for each 7 runs and contains specific student enrollment data such as their gender and the date they enrolled on the couse . However, earlier runs do not have some of the data that was collected later on. All this data has some relevance to my goal of predicting student performance from data collected about their interactions. Therefore, each of these data sets will be considered for exploratory analysis later in this report.

Data Exploration

Using R we can see some simple information about each of our CSV files (which have been loaded as data frames using ProjectTemplate) Below is a table containing each types of data recorded for the seven runs, alongside a short description and what runs the data was recorded for.

Data set	Short Description	Recorded for Runs
archetype.survey.respnses	Survey results which place each student into one of 8 catagories of learning "archetypes"	3,4,5,6,7
enrolments	Enrollment data for each student on the course.	1,2,3,4,5,6,7
leaving.survey.responses	Survey responses kept from a questionnaire given to students who decided to leave the course.	4,5,6,7
question.response	Saved responses for each student for any quizzes they have completed throughout the course.	1,2,3,4,5,6,7
step.activity	The start and completion date and time for each student for each step in the program.	1,2,3,4,5,6,7
weekly.sentiment.survey.responses	Responses to a weekly survay containing a quantitative 1-3 rating and qualitative general feedback.	5,6,7
team.members	little information can be extracted from this, likely has something to do with team building exercises.	2,3,4,5,6,7
video.stats	Data on how students as a whole interacted with videos. Such as how long each video was watched, what devices on, etc.	3,4,5,6,7

From further checking, although some csv files exist they actually don't contain any data. I check has been created to look for this and any empty data frames are not present in the above table's runs, although they technically exist. Furthermore it although many data sets from different runs follow the same name I have checked weather they actually contain the same types of data. This was done my ensuring that all data frame variables shared the same column names. This check is important as later we I likely be merging table rows from different runs to expand on the data I can use to see what are good markers of student performance.

```
## [1] "All data frames ( video ) have the same columns."
## [1] "All data frames ( team ) have the same columns."
## [1] "All data frames ( sentiment ) have the same columns."
## [1] "All data frames ( activity ) have the same columns."
## [1] "All data frames ( question ) have the same columns."
## [1] "All data frames ( leaving ) have the same columns."
## [1] "All data frames ( enrolments ) have the same columns."
## [1] "All data frames ( archetype ) have the same columns."
```

From this check we can see that all the 8 collections of data from each runs share the same column names. This gives confidence in the health of the data set and that they can be merged later on. Furthermore, from manual inspection, using R's view() function, it is clear that some of the rows for these data frames contain missing values. This is especially true for survey response data where not all students have taken the time to respond.

References

Cambridge Home School Online. 2024. "The Rise of UK Online Learning: Trends and Statistics." https://www.chsonline.org.uk/blog/the-rise-of-uk-online-learning-trends-and-statistics.