Future Learn - Analysis Report

Nicha Wilanan

2024-10-31

Introduction

TODO:

CRISP-DM Round 1

1. Business Understandings

1.1 Determine Business Objective

This report aims to gather information for an author of self-development books who would like to share valuable tips for people who aspire to do self-learning in their free time. In the world's current trends, there are so many concepts to learn with less time. The stakeholder's objective is to help people achieve their goals by extracting the essence of success factors to finish self-paced learning. The stakeholder aims to dedicate one chapter of the book to tell anecdotes of how some people can finish the online course while others don't as an exemplar for the readers to follow. Therefore, the criteria for the success finding in this report is to find the common characteristics of those who successfully finish the online course, which is unique to the people who do not complete the course.

1.2 Assess Situation

The risk for the report is that it relies mainly on the learner's survey response, so the answer may not be accurate in case the learners do not answer truthfully. Moreover, the analysis may not reflect all learners' method of study as there are only minority of learners complete the survey.

There is one terminology in this phase, which is the word *learner archetype*. The learner archetype groups learners with the same behavior together. In this report, there are 7 archetypes - advancers, explorers, fixers, flourishers, hobbyists, preparers and vitalisers. If you want to learn more about each archetype, you can go through this link.

1.3 Produce Project Plan

The analysis will follow the CRISP-DM methodologies for two cycles in order to get the key factors leads to course completion. Each cycle will come up with different factors that potentially lead to finish the course. After exploring and analyse data, each factor will be evaluated. In this research, R language is used to extract the insights through statistical summary and visualization.

Initially, the research question for the first cycle is:

Do the learner archetypes affect online course success rate?

1.4 Determine Data Mining Goals

If archetype is significant to the MOOC finish rate, there should be a huge difference in the archetype of finisher and those who unable to finish the course. However, if the archetype is quite similar in both groups, it cannot be deduce that the archetype has an effect on the course completion.

2. Data Understanding

2.1 Collect Initial Data

Future-Learn collected and provided the data that will be used in this report. For this CRISP-DM cycle, three groups of datasets will be utilized. First is the data of archetype survey responses, which will be used to gather the learner's archetype. Second is the course enrollment data that tells how many people enrolled in each course. Lastly, the course activity determines how far the learner goes through the steps in the course.

2.2 Describe Data

The data that will be explored is the raw data of the course named **Cyber Security: Safety at Home, Online, in Life.** This course is designed to be finished in three weeks. Each week consists of multiple steps, including articles, videos, quizzes, and discussions. From September 2016 to September 2018, the course had been run for 7 times. Throughout the runs, steps are added to enhance learner understandings. From 60 steps in the first run, the run was enhanced to 62 steps in the seventh run. The datasets of archetype, course enrollment, and course activity are kept for each run.

These are the list of table for this analysis and its field name:

1. Activity Step: learner_id, week_number, step_number
2. Enrollment: 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
3. Archetype Survey Response: learner_id, archetype, run

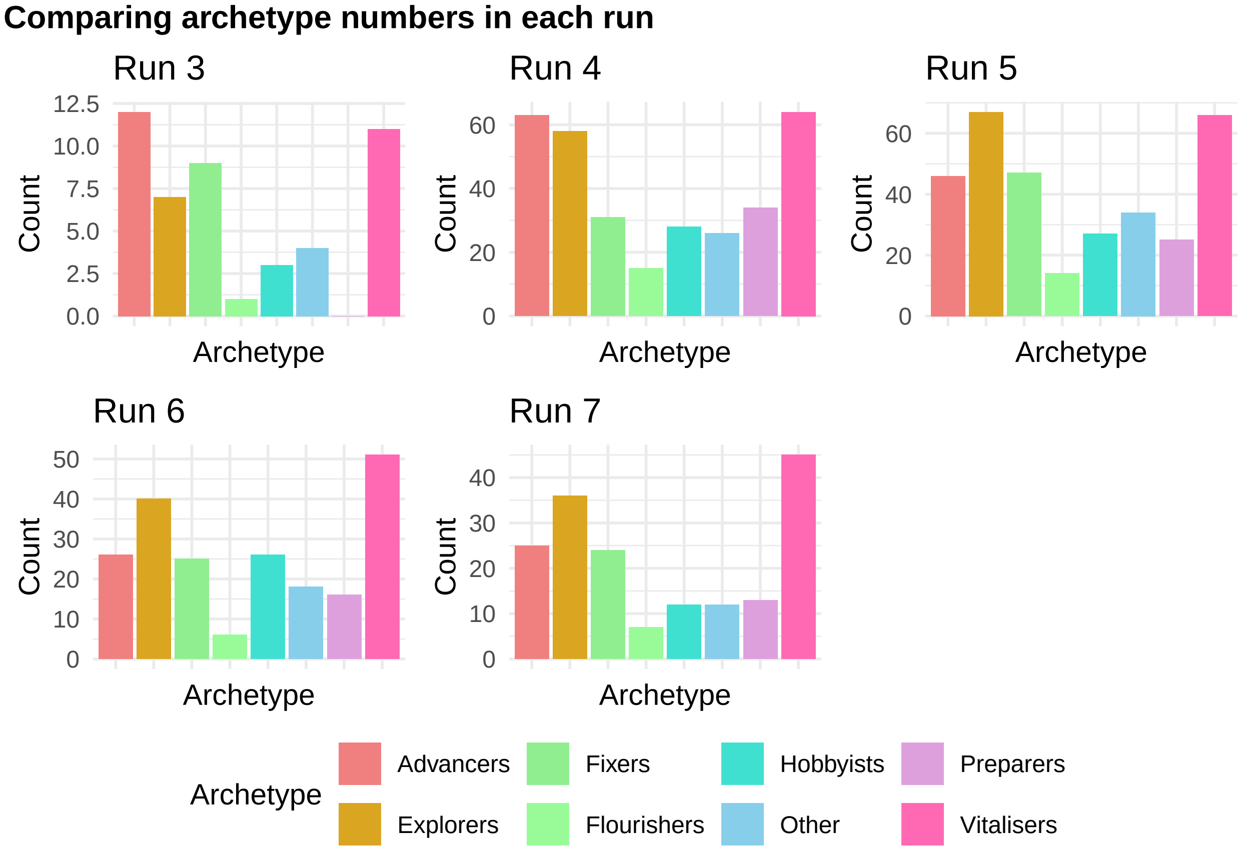
The field that are picked for the analysis will be mentioned in part 3.1.

2.3 Explore Data

	Enrollment	Finisher	Finisher_Percentage
Run 1	13,169	2,654	20.15 %
Run 2	5,279	1,089	20.63 %
Run 3	2,857	801	28.04 %
Run 4	3,428	897	26.17 %
Run 5	3,134	869	27.73 %
Run 6	2,969	472	15.9 %
Run 7	2,231	467	20.93 %

The table above shows a vast difference in the enrollment number and the finisher in each run. In this report, the learners are considered to finish the course if they finish more than or equal to 80% of the activities.

There is a clear pattern that most people do not finish the course, with the lowest number of finishers compared to the enroller in Run 6 at approximately 10% and the highest number in Run 5 at almost 20%.



When comparing the number of archetype in each run, it is obvious that the number of each archetype does not distribute evenly among the learners. The top 4 archetype posesses by the learner are Vitalisers, Advancers, Explorers and Fixers.

2.4 Verify Data Quality

The data quality of this round analysis is quite good and there are no missing data in the available table. Although archetype survey response of Run 1 and Run 2 were not kept, the report can still analyse the link between archetype and learner finishing rate from other Runs. The only concern is that the survey response rate is low comparing to the enrollment, which will be hard to deduce the result of archetype effect from a whole enrollment population.

	Enrollment	Response_Survey	Response_Percentage
Run 1	13,169	0	0 %
Run 2	5,279	0	0 %
Run 3	2,857	47	1.65 %
Run 4	3,428	319	9.31 %
Run 5	3,134	326	10.4 %
Run 6	2,969	208	7.01 %
Run 7	2,231	174	7.8 %

3. Data Preparation

3.1 Select Data

1. Activity Step This data is going to be used as a criteria on finding successful learner. As mentioned before, the learner is considered to complete the course if they completed more than 80% of the steps in the module. To check that, the learner_id, week_number, step_number, and last_completed_at will be used.
2. Enrollment The enrollment data is used just for the exploratory purpose and to understand learners more. There is no direct use to answer the main questions on the archetype for this round.
3. Archetype This table is another main table that will be used along with the activity step. It is needed to know the learner archetype. There are two column being used, learner_id and archetype.

3.2 Clean Data

The data selected is quite clean, so we are going to clean it just by select the fields needed and keep only the unique rows

3.3 Construct Data

After selected the field, new table is created from attributes in the activity step data set. This new table calls *learner_progress* contains learner progress and the archetype. Firstly, it is done by counting the number of steps each learner finished. Then, the percentage of the steps done is calculated and store in a new field. Later, there is an attribute called *finish* that stores boolean of whether the steps are complete more than 80%, which is the threshold for the completion of the course in this report.

3.4 Integrate Data

Afterwards, the *learner_progress* table is merged with the archetype table by *learner_id* to show the type of each learner if any. This new table is called *learner_progress_archetype*

Modelling