# 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 a book author on learning strategies to gain more insights into the characteristics of each type of learner. The stakeholder's objective is to see how each archetype is different in their study's outcome or study method. The stakeholder aims to use this research as a stepping-stone to know the differences or similarities between the archetypes. Therefore, the criteria for the success finding in this report is to figure out whether the archetypes contribute to the uniqueness of study results or study methods. It could end up that each type is unique or it can be that all types share similar methods and outcomes.

### 1.2 Assess Situation

The risk for the report is that it relies mainly on the learner's survey response on their archetype, 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 a 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 behaviour together. In this report, there are 7 archetypes - advancers, explorers, fixers, flourishers, hobbyists, preparers and vitalisers. To learn more about the archetype, go through this link.

### 1.3 Produce Project Plan

The analysis will follow the CRISP-DM methodologies for two cycles in order to know more about the archetype study result and study strategy. Each cycle will come up with a main research topic. After exploring and analyse data, the findings will be evaluated. In this research, R language is used to extract the insights through statistical summary and visualization.

Initially, the research question will aim to see the result of the course completion on each archetype. The main research topic for this cycle is:

**Does the learner archetype affect online course completion rate?**

Noted that the *completion* in this report refers to the learners who accomplished more than 80% of the course content.

## 1.4 Determine Data Mining Goals

To know that the learners' archetype affect online course completion rate, a solid pattern of completion for the archetype on every run must be found. If the completion rate in each archetypes are random, it cannot be deduce that the archetype actually affect 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 is gathered from a 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 years, steps are added to enhance learner understandings. From 60 steps in the first run, the run was gradually 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 tables for this analysis and its field name:

1. 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
2. Activity Step: learner_id, step, week_number, step_number, first_visited_at, last_completed_at
3. Archetype Survey Response: id, learner_id, responded_at, archetype

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

## 2.3 Explore Data

Table 1: Enrollment and course finisher percentage

|        | Enrollment | Finisher | Finisher_Percentage |
|--------|-----------|----------|---------------------|
| Run 1  | 13,169    | 1,581    | 12.01 %             |
| Run 2  | 5,279     | 691      | 13.09 %             |
| Run 3  | 2,857     | 467      | 16.35 %             |
| Run 4  | 3,428     | 580      | 16.92 %             |
| Run 5  | 3,134     | 586      | 18.7 %              |

|  | Enrollment | Finisher | Finisher_Percentage |
|---|---|---|---|
| Run 6 | 2,969 | 287 | 9.67 % |
| Run 7 | 2,231 | 278 | 12.46 % |

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%.
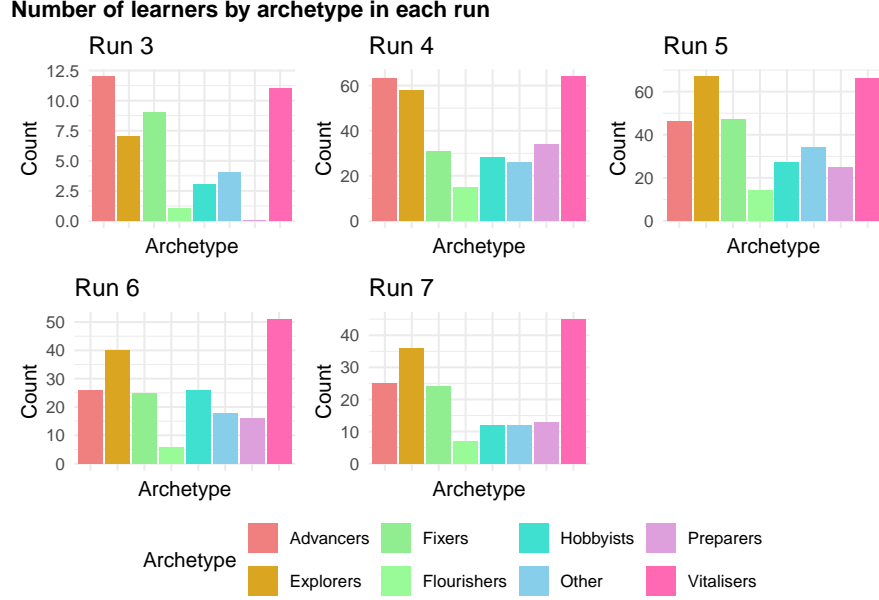
**Number of learners by archetype in each run**



Figure 1: Number of learners by archetype in each run

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 archetypes posess 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.

Table 2: Archetype response survey percentage

|  | 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 % |

|          | Enrollment | Response_Survey | Response_Percentage |
|----------|------------|-----------------|---------------------|
| Run 6    | 2,969      | 208             | 7.01 %              |
| Run 7    | 2,231      | 174             | 7.8 %               |

## 3. Data Preparation

### 3.1 Select Data

1. Enrolment: The enrolment 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.

2. Activity Step: This data is going to be used as a criteria on finding successful learner. To check that learners completed more than 80% of the steps in the module, the learner_id, week_number, step_number, and last_completed_at will be used.

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*, which contains learner progress and the archetype. Each run has its own *learner_progress* table. To construct the new table, firstly, count the number of steps each learner finished. Then, compare number of finished steps to number of all steps and put it in *finish_percentage* field. Later, add *finish* field that stores boolean value of whether the steps are complete more than 80%.

### 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*. Again, each run has its own *learner_progress_archetype* table.

## 4. Modelling

As right now we have all data prepared, we will examine our main research topic for this round, **Does the learner archetype affect online course completion rate?**

### 4.1 Archetype of learners who completed the course

Let start by take a look at the table below on the archetype of the learner who complete the course.

Table 3: Archetype of the learners who complete the course

| Run | Advancers | Explorers | Fixers | Flourishers | Hobbyists | Other | Preparers | Vitalisers | NA |
|---|---|---|---|---|---|---|---|---|---|
| 3 | 3 | 2 | 1 | 1 | 1 | 2 | 0 | 2 | 455 |
| 4 | 14 | 22 | 10 | 1 | 14 | 7 | 15 | 19 | 478 |
| 5 | 9 | 16 | 18 | 3 | 15 | 16 | 5 | 30 | 474 |
| 6 | 6 | 1 | 5 | 2 | 7 | 4 | 4 | 11 | 247 |
| 7 | 11 | 9 | 3 | 4 | 4 | 4 | 0 | 12 | 231 |

As mentioned before, not many people response to the survey of the archetype, so the majority of learner archetype is unknown. If we ignore unknown type, depends on each run, some archetype perform better than the others. To compare fairly, the data is scaled for each run, so that the number of completers can be compare properly throughout different run. Or else the priority would be given to the run with higher enrollment as higher enrolment has higher chance to have a higher finishers. The brighter color on the heatmap represents more number of successful learners in that archetype.

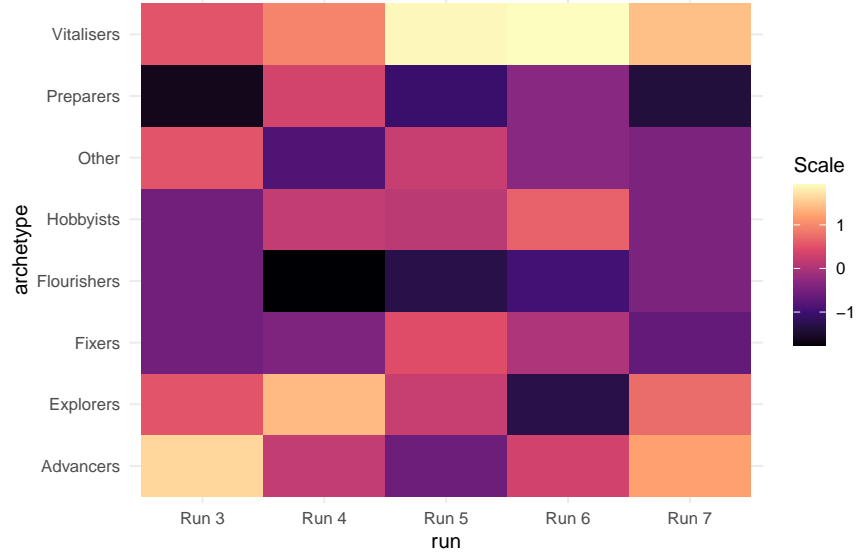**Heatmap showing scaled number of learners who complete the course in each run**



Figure 2: Heatmap showing scaled number of learners who complete the course in each run

From the heatmap, it is evident that *vitalisers* archetype perform best in all run comparing to other archetype followed by *advancers* and *explorers*, where they have less completer than the others in some runs. However, it needs to be tackle that did vitalisers perform better just because most people are *vitalisers* or it is really affect the course completion. To check the performance, the success rate vs unsuccess rate for each archetype will be compared.

## 4.2 Success rate for each learners' archetype

For this part, let we determine the ratio of learner with same archetype that complete or not complete the course. If the success rate within the archetype is high, it can be deduced that the archetype affect the online course completion.

Table 4: Course completion rate by archetype

| Run | Advancers | Explorers | Fixers | Flourishers | Hobbyists | Other | Preparers | Vitalisers |
|-----|-----------|-----------|--------|-------------|-----------|-------|-----------|------------|
| 3 | 27.27 % | 50.00 % | 14.29 % | 100.00 % | 33.33 % | 66.67 % | 0.00 % | 22.22 % |
| 4 | 28.00 % | 44.90 % | 43.48 % | 10.00 % | 66.67 % | 33.33 % | 55.56 % | 40.43 % |
| 5 | 29.03 % | 32.65 % | 46.15 % | 33.33 % | 62.50 % | 51.61 % | 31.25 % | 53.57 % |
| 6 | 27.27 % | 5.00 % | 29.41 % | 40.00 % | 36.84 % | 36.36 % | 36.36 % | 28.21 % |
| 7 | 61.11 % | 33.33 % | 25.00 % | 80.00 % | 44.44 % | 36.36 % | 0.00 % | 37.50 % |

By checking the percentage of completion on each archetype, the result is low and random among different learner's type in each run. The low success rate among each archetype is not surprising because the overall finish rate as a whole shows in *Table 1* is also low.

## 5 Evaluation

### 5.1 Evaluate Results

After explore the data on archetype and course completion rate thoroughly, the archetype has no distinctive effect on the completion. At first glance, some archetype is standing out to have more number of complete learner but actually that is happening because there are more learner in that archetype as a whole so there is a higher chance for learners in that archetype to finish the course more. When comparing the success rate for each archetype, the result is too random in all archetype to have a significant impact on the course completion. Although the rate of completion of the learner with known archetype is more than the overall completion rate, there is a lack of solid pattern among the archetype in each run. Based on the findings, the archetype does not affect whether or not learner finish the online course.

### 5.2 Review Process

Data wrangling and data exploration have been revisited. The results provided above are legit.

### 5.3 Determine next steps

It is known that the type of learner solely cannot use to determined the completion rate, so it can be concluded that the archetype cannot determine the study outcome. For the next phase, the research will go deeper into

# CRISP-DM Round 2

## Business Understanding

In this round, instead of using the survey response from users, we are going to investigate the user success rate through users' actions in the course.

## 1.1 Determine Business Objective

The business objective for this round is exactly the same as the first round. The stakeholder wants to know the common traits of the course completer.

## 1.2 Assess Situation

There is lower risks in this phase as we do not rely on the user survey response. There is a dataset that shows the first visited date and last completed date of each learners for each step in the course. This report will assume that the data were collected accurately and can be used straightaway.

## 1.3 Produce Project Plan

As we now realize that the archetype doesn't affect the result, we need to take a look at the action made by the learners. As the course is designed to finished the sub-steps within a week, we will examined that is there a higher chance for a learner to finish a course if they can follow the course schedule.

Therefore, the main research topic for the second cycle is: **Does following course schedule contribute to course completion?**

## 1.4 Determine Data Mining Goals

If keeping to the course schedule can make a learner complete the course.