Impact of learner archetype on study completion and study approach

Nicha Wilanan

Semester 1 2024/2025

Introduction

This report analyses the **impact of archetype on study completion and approach** to answer the questions of the stakeholder, a researcher studying the learning strategies for each archetype for the best approach. The online course data and survey are the material used to answer the question of how the learner archetype and the study method can impact the completion of studying online courses. The procedure to construct the research is using two cycles of Cross-Industry Standard Process for Data Mining (CRISP-DM). The first cycle will be conducted to examine whether the archetype has an impact on course completion. Then, the second cycle will built on top of the first cycle, depending on the result. If the result suggests that some archetypes successfully dominate the course completion outcome, the second cycle will focus on the behaviour of those archetypes. On the contrary, if the result shows that the archetypes do not contribute to the success, the behaviour of the learners of each archetype who completed the course will be investigated to show the study pattern to finish the online course by different types of learners.

CRISP-DM Round 1

1. Business Understandings

This phase will focus on understanding the stakeholders' business objectives, assessing the situation based on the data provided, and planning the project steps.

1.1 Determine Business Objective

The overall business objective of this report is to provide informative research on how archetypes can affect study outcomes and the study method that the successful learner type has. This report will benefit researchers in the education field because they can use the research to motivate and help learners with different archetypes succeed in their study.

The first cycle of this report aims to find the impact of the learner archetype on course completion. This question is the main topic of this research that will shape the following cycle's question. By getting to know whether or not the archetype affects learners' completion rate, the stakeholders can have a general idea of what archetype the learners should have to have a higher chance to complete the course. If there is an archetype that completes the course more than others, the researcher may advise learners to approach the study material with that particular mindset. On the other hand, if the archetype is not a significant factor in course completion, each archetype has a fair chance of completing the course, and it will be investigated further in the next cycle on the study pattern of the archetype.

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 participated in this online course 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 archetype is the goal and mindset that the learners have when approaching the study. It groups learners with the same goal 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 Determine Data Mining Goals

A solid completion rate pattern for individual archetypes on every run must be found to determine whether the learners' archetype affects the online course completion rate. If the completion rate in each archetype is random, it cannot be deduced that the archetype affects the course completion.

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 the 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 Produce Project Plan

After come up with a main research topic, the data is going to be cleaned, explored and analysed. Then, the findings will be evaluated. In this research, R language is used to extract the insights through statistical summary and visualization.

2. Data Understanding

In this phase, the credibility and data quality will be checked. The table that will be used will be mentioned, and the data will be explored to gain an overall understanding of the course enrolment and learner archetype.

2.1 Collect Initial Data

Future-Learn, a UK-based online learning platform that partners with many global universities, collected and provided the data for this report. Based on the company's credibility, the data can be considered valid and safe. For this CRISP-DM cycle, three groups of datasets will be utilized:

1. The data from *archetype survey responses* to gather the learner's archetype.
2. The *course enrolment* data to know how many people enrolled in each course.
3. The *course activity step* to determine 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** provided by Newcastle University. 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 enrolment, and course activity are kept for each run.

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

1. Enrolment: 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

Enrolment and archetype data will be explored to get the overall picture before going further in later phase. Based on summary on *Table 1*, there is a vast difference in each run's enrolment number and finisher. In this report, the learners are considered to complete the course if they finish more than 80% of the module. 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%.

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 %
Run 6	2,969	287	9.67 %
Run 7	2,231	278	12.46 %

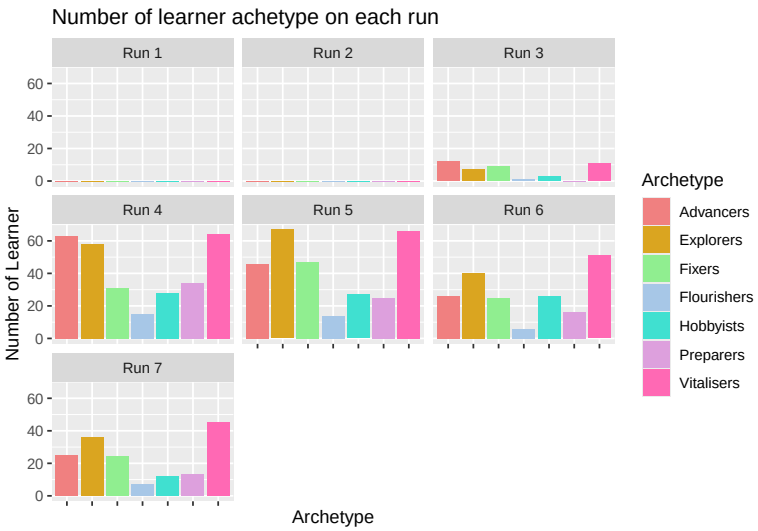

Figure 1: Number of learners by archetype in each run

For a number of learner archetypes, *Figure 1* shows that runs 1 and 2 do not have archetype data response, and the number of each archetype does not distribute evenly among learners on runs 3 to 7. As runs 1 and 2 do not have the archetype data, which is our primary research focus, run 1 and run 2 will be ignored for the analysis that needs to use an archetype. The top 4 most common archetypes are *Vitalisers*, *Advancers*, *Explorers* and *Fixers*.

2.4 Verify Data Quality

The data quality of this round of analysis is quite good, and there are no missing data. Although the archetype survey responses of Run 1 and Run 2 were not kept, the report can still analyse the link between the archetype and learner finishing rate from Run 3 to Run 7. The only concern is that the survey response rate is low compared to the enrolment, making it hard to deduce the result of the archetype effect from a whole enrolment 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	43	1.51 %
Run 4	3,428	293	8.55 %
Run 5	3,134	292	9.32 %
Run 6	2,969	190	6.4 %
Run 7	2,231	162	7.26 %

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 columns being used, learner_id and archetype.

3.2 Clean Data

The data selected is clean, so we will clean it just by selecting the fields needed and keeping only the unique rows. Also, as there are main 7 archetypes in this report, the archetype name *Other* will be removed from the report to prevent any confusion.

3.3 Construct Data

After selecting the field, a new table is created from attributes in the activity step data set. This new table is called *learner_progress*, which contains learner progress and the archetype. Each run has its own *learner_progress* table. To construct the new table, count the number of steps each learner finished.

Then, compare the number of each learner's finished steps to the number of all steps and put it in the *finish_percentage* field. Later, add the *finish* field that stores the 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*, and 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 the course completion rate?**

4.1 Archetype of learners who completed the course

Table 3: Archetype of the learners who complete the course

Run	Advancers	Explorers	Fixers	Flourishers	Hobbyists	Preparers	Vitalisers	NA
3	3	2	1	1	1	0	2	457
4	14	22	10	1	14	15	19	485
5	9	16	18	3	15	5	30	490
6	6	1	5	2	7	4	11	251
7	11	9	3	4	4	0	12	235

Let's start by taking a look at *Table 1* on the archetype of the learner who completes the course. As mentioned before, only a few people responded to the survey of the archetype, so most learner archetypes are unknown. If we ignore the unknown type, depending on each run, some archetypes perform better than others.

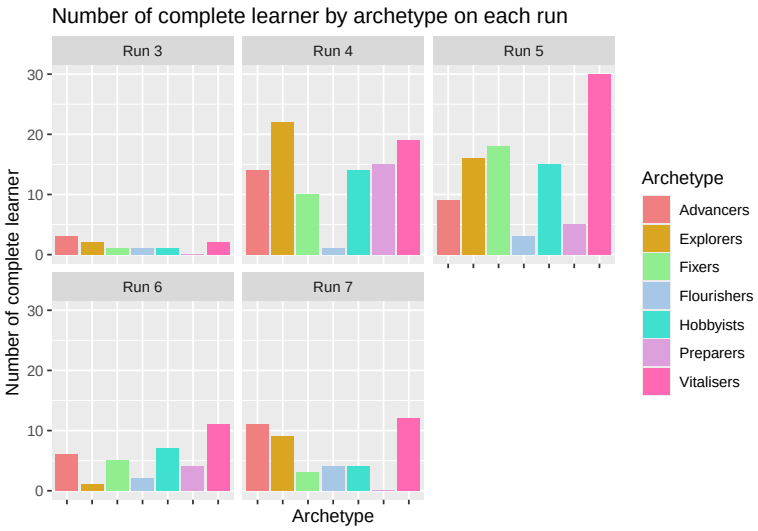

Figure 2: Bar chart shows the number of learners who complete the course in each run

Figure 2 is a graphical version of the *Table 3*. From the chart, it is evident that the *vitalisers* archetype performs best in most runs compared to other archetypes followed by *advancers* and *explorers*, where there are fewer completers in some runs. However, it needs to be tackled whether *vitalisers* have more completer just because most people are *vitalisers* or whether the archetype really affects course completion. To check the performance, each archetype's success rate will be compared.

4.2 Success rate for each learners' archetype

For this part, let us determine the ratio of learners with the same archetype who complete or do not complete the course. If the success rate within the archetype is high, it can be deduced that the archetype affects the completion of the online course.

Table 4: Course completion rate by archetype

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

By checking the completion percentage on each archetype, the result is low and random among different learner types in each run. Although the *flourishers* archetype has a high completion rate in runs 3,4 and 7, the rate is low in the other two. The low success rate among each archetype is not surprising because the overall finish rate, as shown in *Table 1* is also low.

5. Evaluation

5.1 Evaluate Results

After thoroughly exploring the data on archetypes and course completion rates, the archetypes have no distinctive effect on completion. At first glance, some archetypes stand out as having more complete learners, but that is happening because there are more learners in that archetype, 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 archetypes to significantly impact the course completion. Although the learner's completion rate with the known archetype is higher than the overall completion rate, there must be a more solid pattern among the archetypes in each run to conclude that the archetype impacts course completion.

5.2 Review Process

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

5.3 Determine next steps

As the archetype is not the factor of course completion, the stakeholders would like to know more about each archetype's strategy to finish the online course successfully. For the next phase, the research will investigate the study pattern that can lead to success in online course completion for each archetype.

CRISP-DM Round 2

1. Business Understanding

In this round, we will investigate the user success rate through users' actions in the course and compare it with the archetype.

1.1 Determine Business Objective

The stakeholder wants to know how each learner archetype approaches the course that leads to completion. As archetypes alone cannot contribute directly to the completion rate, other factors that lead to success among archetypes will be explored further.

1.2 Assess Situation

In this phase, the primary data still rely on the user survey response, which is limited, and the archetype distribution is not symmetrical, so the result may not represent the entire archetype population perfectly. As new data is needed to see the action of each learner archetype, new data is added to this scope to calculate the time it takes for each learner to finish the weekly module of the course. This report will assume that the data were collected accurately and can be used immediately.

1.3 Determine Data Mining Goals

One action made by the learners when studying online courses will be explored in this cycle to see the pattern of the successful learner with different archetypes. The action chosen to be determined is the time spent finishing a week of online course material. As the only way to complete the online course is to finish each small step, seeing how long the learner spent finishing each weekly module is a good start to knowing the characteristics of the learner archetype approach to finish the online material.

Therefore, the main research topic for the second cycle is: **Does the successful learner in each archetype have a unique pattern in time spent on learning weekly material?**

1.4 Produce Project Plan

The project plan is similar to the previous phase. The data will be cleaned, explored, analysed, and evaluated using R language. The insights are gained through statistical summary and visualisation.

2. Data Understanding

Two datasets as the previous cycle are used: the archetype survey responses and the course activity step. However, the data details used in the activity step will differ as we will examine a new question with a new perspective.

2.1 Verify data quality

From the dataset, many rows spend too little time to complete the weekly course. For example, the data in *Table 5* shows that the learner only takes 1 minute or less to complete each step. With this small amount of time, the learner may skim through the course without finishing it, which does not show the actual complete

date as the data aims to show. However, there is no method to distinguish whether the learner completed the course or just skimmed through the course, so this report will proceed with the data given.

Table 5: Lower than usual duration to finish the course

step	step_number	first_visited_at	last_completed_at
1.1	1	2017-10-20 13:51:26 UTC	2017-10-20 13:52:28 UTC
1.2	2	2017-10-20 13:52:31 UTC	2017-10-20 13:52:44 UTC
1.3	3	2017-10-20 13:52:47 UTC	2017-10-20 13:52:55 UTC

3. Data Preparation

In this phase, the steps of data preparation will be shown. As the CRISP-DM cycle 2 is built on top of cycle 1, some data in the first cycle will also be used in this round.

3.1 Select data

These are the list of tables and the fields used in this cycle:

1. Activity step: This data will be used to find the learner's duration to complete each weekly module. The fields that will be used are *learner_id*, *week_number*, *first_visited_at*, and *last_completed_at*.
2. Learner progress archetype: The data constructed in the previous round will be utilized as it already contains the field *finish*, which indicates whether the learner completes the course, along with the field *archetype* that tells the learner type.

3.2 Clean data

This report believes that the data provided by the Future-Learn is clean and valid. The process of cleaning data in this round is to filter out the data that is not valid for usage, such as the data that has no *last_completed_at*.

3.3 Construct data

After selecting and filtering out the rows, the duration in days is calculated by grouping the *learner_id* and *week_number*, then finding the *first_visited_at* of that entire week and subtracting with the last date for *last_completed_at* in the same week. This new table is called *weekly_module_time_spent_for_finisher*. This step is done on the data of runs 3 to 7. Also, only the successful learner will be modelled to see the time the complete learner takes, so the data will only filter to have the learner with *finish* = True.

3.4 Integrate data

Then, the five new tables *weekly_module_time_spent_with_arch* for runs 3 to 7 are formed by merging *learner progress archetype* to the *weekly_module_time_spent_for_finisher* to get the learner archetype. Lastly, the five tables are combined by row to create the big table that contains *learner_id*, *week_number*, *duration_in_days*, and *archetype* of all five runs together.

4. Modelling

With all data prepared, the duration graph is ready to plot, but before going into the archetype, let's explore the general duration it takes for the learner who completes the course.

4.1 Duration the learner who complete the course spend on weekly material

This part will explore the time to complete the weekly module without looking at the archetype. *Figure 3* calculated the days the successful learner, the learner with more than 80% course completion rate, took to complete the weekly material in runs 3 to 7. Run 1 and 2 are ignored because they don't have survey responses on archetypes that will be used to analyse later on.

Ignoring the outlier from the plot, *Figure 3* shows the trend that most learners who complete the course finished the weekly module within a week. Additionally, most people spend less time as the week module number progresses, meaning the week 3 module is completed faster than the week 1. This exploration is promising in that it shows the pattern that most course completers generally finish the course within 1 week, the recommended time frame of the online course, which can boost the online course completion rate.

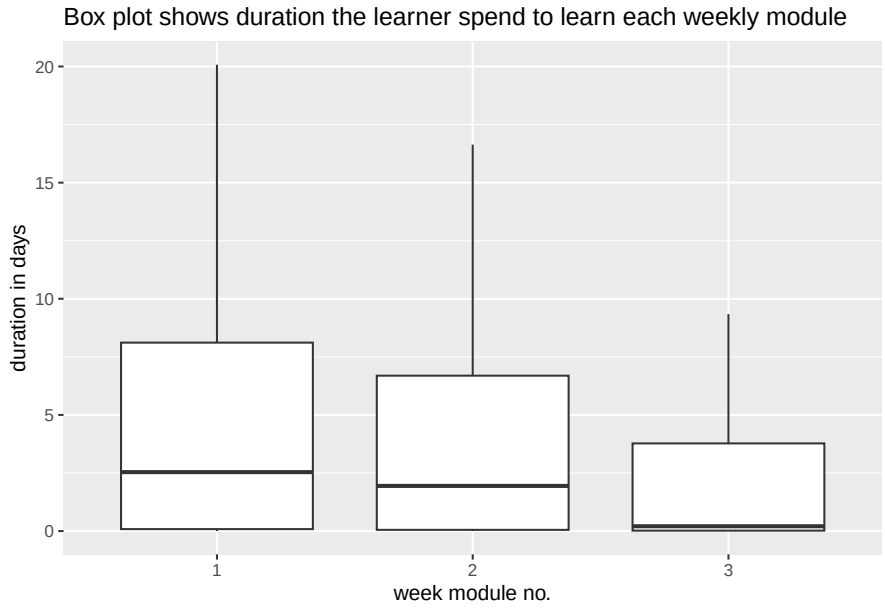

Figure 3: Box plot shows duration the complete learner spend to learn each weekly module of the course

4.2 Duration the each learner archetype who complete the course spend on weekly material

Based on the plot in *Figure 4*, each archetype has a distinct duration to finish the weekly module. Learners from all archetypes mostly take less than 10 days to complete the weekly material. Three archetypes that take less time to study are *flourishers*, *preparers*, *fixers* in which most people took less than 10 days to complete a weekly module. On the contrary, other archetypes that are not mentioned have more variability in the time learners took to complete the module. For example, some of the *explorers* may use a day or a few to complete the course, but some *explorers* also took a month to complete the course.

For the trend that learners spend less time as the week module number progresses that we have explored in *Figure 3*, this trend is still valid in most of the archetypes but not for the *vitalisers* and *fixers* where the learner tends to spend the same amount of time for each weekly module.

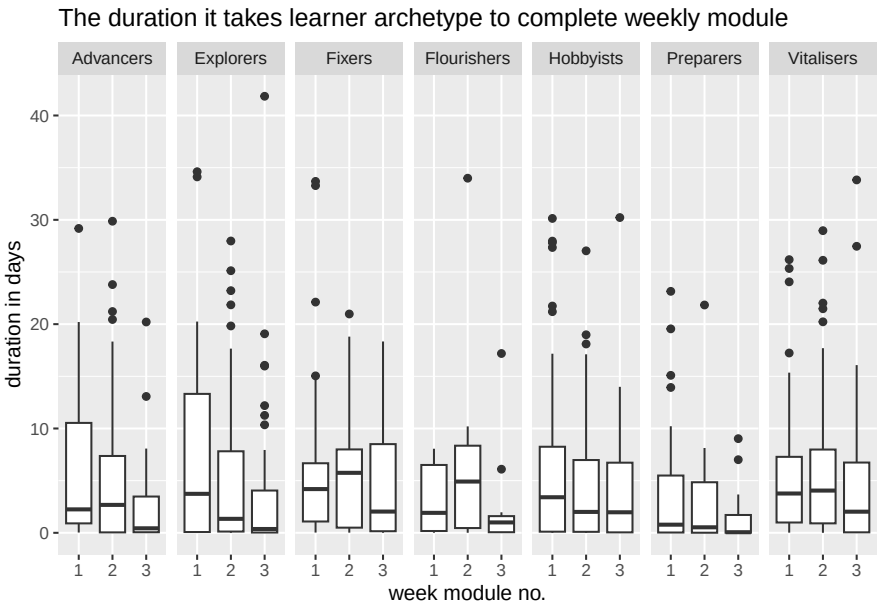

Figure 4: The duration it takes the complete learner to complete weekly module by archetype

5. Evaluation

That concludes the analysis for cycle 2 of the impact of archetype on the time spent on the weekly material. The box plot in *Figure 4* shows that each archetype has its study methodology for the time spent studying the weekly module. The pattern we see can be linked with the definition and characteristics of the learner type. Such as, *fixers* learn to manage the current aspects of their life, so learners of this type tend to finish the course fast. Also, the *vitalisers*, who learn as a hobby and for the love of learning, can finish the course quickly and without a rush toward the end. On the contrary, the *explorers* who learn to decide what to do next tend to take longer to finish a course and rush to finish the course towards the end.

The model and exploratory analysis in this cycle can answer the business objective of the impact of archetypes on the study methods. The graph shown in *Figure 4* is self-explanatory on how the behaviour of learners in each archetype varies.

6. Deployment

To conclude, the first cycle of the CRISP-DM can help answer the questions on the impact of the archetype on the study completion. The second cycle also uses the study duration to show each archetype's learning approach. For the deployment phase, the report and presentation will be delivered to the stakeholders to help them understand the findings of this report analysis.

The presentation will include a recap of the business understanding, the data usage and the graphical representation that shows the insight behind the data provided. Lastly, it will end with a summary that will leave users with the key takeaway of the analysis so that the stakeholders can use it to work on their research on learning strategies for the archetype.