Analysis_Report

Nisarga

2026-01-16

Executive Summary

Massive Open Online Courses (MOOCs) commonly have high participation but concurrently high attrition rates. This report examines the re-engagement of participants in a Cyber Security course on the FutureLearn platform.

Two CRISP-DM cycles succeed one another:

CRISP-DM 1 assesses the decay in engagement following enrolment and determines the point at which the learners become inactive.

CRISP-DM 2 measures the probability of learner re-engagement after inactivity, focusing on completers versus non-completers

Key findings:

- It peaks strongly on Day 0, then declines dramatically during the first week.
- Numerous students achieve persistent inactivity relatively early.
- The probability of re-engagement becomes lower with longer intervals of inactivity
- Completers have a greater chance to re-engage as compared to non-completers.
- There appears to be a 'recoverability window' after the initial onset of inactivity, beyond which the chance of recovery becomes highly improbable
- Engagement drops sharply between Day 0 and Day 7, and re-engagement probability approaches zero after ~36 inactive days.

1. Introduction

MOOCs are scalable platforms for learning but are plagued with issues of retention of learners. High dropout rates are exhibited by a significant number of learners immediately after enrollment. A study of engagement patterns after enrollment is important for online course providers since it helps in the early identification of at-risk individuals and informs the design of timely interventions, such as reminders and adaptive support.

This report is based on the "FutureLearn Cyber Security" online course and aims to examine the following aspects are:

1. When learners tend to become inactive.
2. Whether students re-engage (or recover) from non-participation, as opposed to non-completers.

2. Research Questions

RQ1: What are the typical engagement patterns of learners after enrolment, and at what points do learners most commonly become inactive (defined as having no recorded course activity for >= 7 consecutive days)?

RQ2: How does the probability of re-engagement change with inactivity length, and how does it differ for completers vs non-completers?

3. Data Description

3.1 Raw datasets

These datasets were selected because the enrolment timestamps allow alignment of learners by start date, while step activity logs provide observable engagement events needed to define inactivity streaks and re-engagement behaviour.

The analysis uses two main FutureLearn datasets (for seven runs):

1. Enrolments data (*_enrolments.csv)
    a. Contains enrolment timestamps (enrolled_at).
    b. Contains completion status (fully_participated_at).
2. Step activity data (*_step-activity.csv)
    a. Contains learner activity timestamps (first_visited_at)
    b. Represents interactions with course steps.

Data structure and granularity

- Enrolments data is **learner-level** (one row per learner per course run).
- Step activity data is **event-level** (multiple activity rows per learner).
- Runs analysed: **Run 1**–**Run 7**.

3.2 Operational definitions

a. Engagement: learner performs at least one step activity on a given date.
b. Inactive: learner has no recorded activity for >= 7 consecutive days.
c. Re-engagement: learner returns to activity after an inactive period.
d. Completer: fully_participated_at is not missing.

4. Data Preparation

4.1 List CSV files

- Used list.files("data", pattern="\.csv$") to locate all csv files
- Applied str_detect() filters to extract:
    a. enrolment files: "enrolments.csv"
    b. step activity files: "step-activity.csv" or "step_activity.csv"
- Extracted run number from file name using regex (\d+) to create a run column
- Performed sanity checks to stop if files were missing (pipeline integrity)

I started by listing all the enrollment and step activity csv files available from the various course runs to ensure I knew which data sets were available for use.

4.2 Combine enrolments

1. Read each enrolments CSV using read_csv()
2. Standardised key ID fields:
    - learner_id cast to character
3. Parsed timestamps:
    - enrolled_at → enrolment_time using ymd_hms()
4. Created derived time variable:
    - enrolment_date = as.Date(enrolment_time)
5. Created completion indicator:
    - completer = !is.na(fully_participated_at)
6. Appended run identifier for each file
7. Combined all runs using map_dfr() into one dataset (master_enrolments)

I combined all the enrolment data into a master data table (master_enrolments), cleaned enrolment dates, developed key variables such as 'run', 'learner_id,' and 'completer,' among others.

4.3 Combine step activity

1. Read each step-activity file with read_csv()
2. Standardised:
   - learner_id as character
3. Parsed timestamp:
   - first_visited_at → activity_time
4. Derived date:
   - activity_date = as.Date(activity_time)
5. Removed invalid entries:
   - filtered out missing timestamps (filter(!is.na(activity_time)))
6. Added run identifier
7. Combined all runs using map_dfr() into (master_step_activity)

I combined the step activity data into a single table called master_step_activity, preprocessed the activity timelines, and pulled the dates of the activities to reflect interaction events.

4.4 Daily engagement summary

1. Joining enrolment and step activity Process Used:
   a. Joined activity to enrolments using:
      - keys: (run, learner_id)
   b. Ensured engagement happens after enrolment:
      - days_after_enrol = activity_date - enrolment_date
      - filtered days_after_enrol >= 0
   c. This created your unified learner activity table: master_dataset
2. Daily engagement aggregation
   a. Grouped by:
      - (run, days_after_enrol)
   b. Computed:
      - active_learners = n_distinct(learner_id) (unique learners)
      - total_events = n() (total step visits)
   c. Saved result to: master_daily_engagement

I transformed activity data from a step level to create a daily engagement data set (master_daily_engagement) based on the number of learners who were engaged on each day following enrollment. This helped to identify decays and inactivity trends in the data.

4.5 Data Quality & Validation Checks

To ensure inactivity and re-engagement are valid, the following checks were applied:

1. Missing timestamps removed
   - step records with missing first_visited_at were removed
   - enrolment records without valid enrolled_at cannot be used for alignment
2. Timeline validity check
   - step activities occurring before enrolment were excluded
   - ensures inactivity streaks aren't miscalculated
3. Consistency of identifiers
   - learner_id converted to character to prevent numeric/integer mismatch during joins
4. Duplicates handled
   - multiple activities on the same day were reduced later into 1 engagement unit per learner-day (important for inactivity measures)

5.CRISP-DM 1: Engagement & Inactivity

5.1 Business Understanding

MOOCs is that of disengagement where most of the learners quickly become disengaged as soon as they enroll in the course, and very few learners complete the courses. This creates a challenge for learning outcomes as well as the value for the MOOC platform where the learners who are disengaged may fail to benefit from the learning content. The first task in this CRISP-DM process is to identify the points of learner engagement behavior post-enrollment for which most learners tend to drift away.

5.1.1 Why engagement + inactivity matters

Engagements following enrollment show the greatest correlation to whether the individual will persist in the educational process. When inactive early, the individual may never reach meaningful educational outcomes. From a learning analytics perspective, observations of the typical patterns of inactivity versus engagements help the service providers in the following ways:

1. To discover early indicators of disengagement.
2. Design interventions on a timely.
3. Enhance rates of course completion and retention.
4. enable scaled personal support.
5. Platform providers care because improving retention improves platform value and course completion.

This inactivity may not always be indicative of a dropout and therefore it may be very important to determine the point where inactivity occurs in order to estimate where it may be recovered (Cycle 2).

5.1.2 The main parties who can benefit from the above analysis are:

1. Providers of the FutureLearn/MOOC platform (retention strategies, product design, intervention automation).
2. Course faculty members and institutions (Learning Impact, Satisfaction, Course Development Decisions).
3. EdTech learning analytics groups (dashboards, early warning systems, behavior models)

These stakeholders have an interest in this research because disengagement at an early stage negatively hampers learning impact and completion rates, thus negatively influencing platform value and, in turn, reducing course effectiveness. Estimating this inactivity point will help design strategies for preserving learners to improve completion rates.

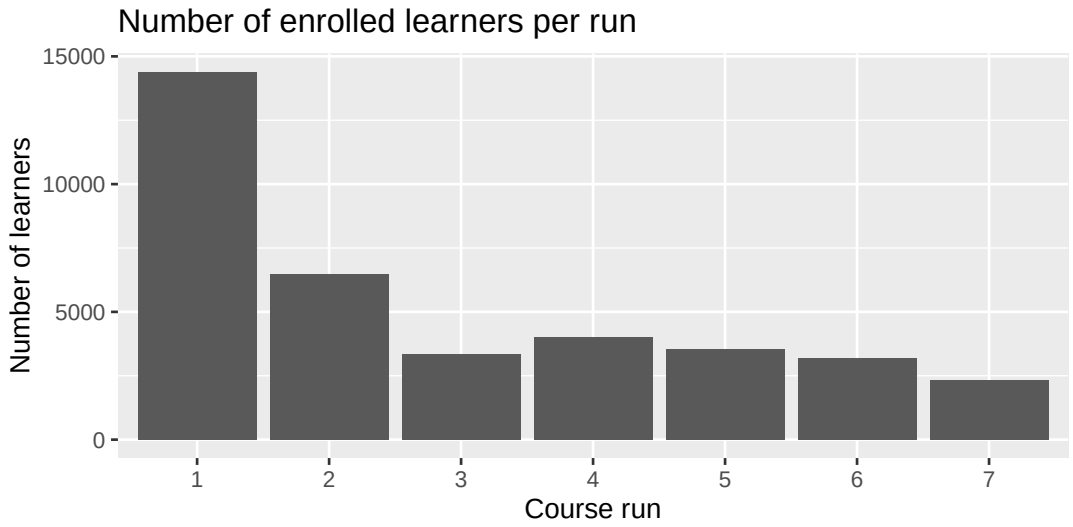

Figure 1: Number of enrolled learners per run

The datasets mainly contain timestamp variables ('enrolled_at', 'first_visited_at'), learner identifiers ('learner_id', 'run'), and derived numeric variables such as `days_after_enrol` and 'inactivity_days'.

RQ1: What are the typical engagement patterns of learners after enrolment, and at what points do learners most commonly become inactive (defined as having no recorded course activity for >= 7 consecutive days)?

It highlights the critical dropout period (e.g., Day 0–Week 1) and helps explain how rapidly learners disengage from the MOOC.

5.1.3 Success criteria

- A clear engagement curve showing activity decline over time after enrolment.
- A measurable distribution of inactivity onset points.
- Evidence-based identification of the most common timing of inactivity.
- A defensible inactivity definition (>=7 consecutive days without activity), aligned with engagement decline patterns.

-From a business point of view, the success criteria would be if the analytics outcome is the point in time at which the student is at the risk of dropping out the most where intervention would create the most value.

These results are automatically applicable to the second cycle of CRISP-DM, identifying when disengagement happens.

5.2 Data Understanding

The enrolments dataset provides enrolment timestamp and completion outcome. The step activity dataset records the first time a learner visited a step ('first_visited_at'), which we treat as engagement events. Timestamps that were missing were removed to prevent incorrect inactivity measurement.

Table 1: Dataset Overview

Dataset	Rows	Columns
Enrolments	37296	17
Step activity	423072	9
Master joined dataset	422971	12

5.2.1 Engagement after enrolment (first 30 days)

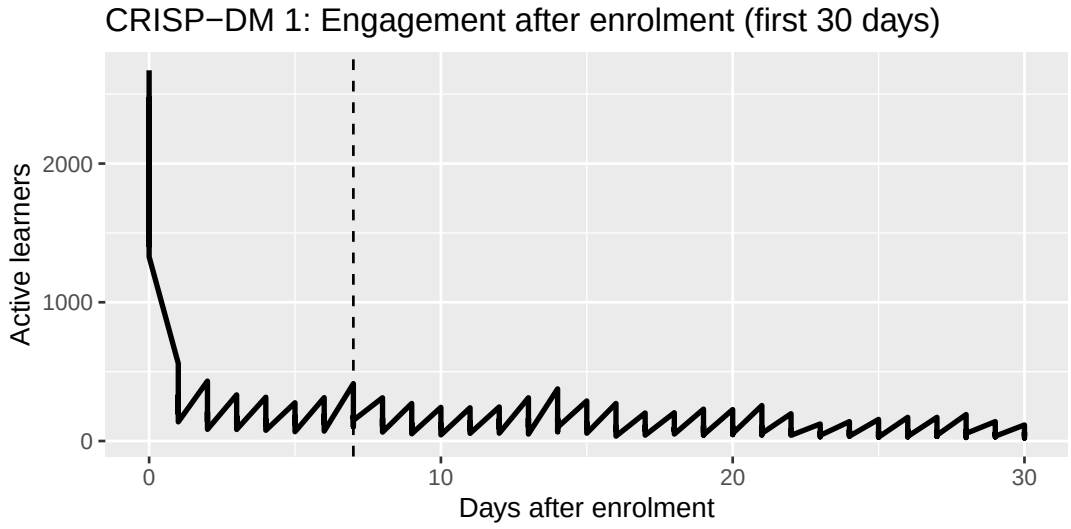

Figure 2: Engagement after enrolment (first 30 days)

Interpretation : Engagement peaks on Day 0 and declines sharply in the first week. The weekly oscillation likely reflects course weekly release structure.

Numerical Evidence: Active learners drop substantially from Day 0 to Day 7, confirming that most attrition occurs within the first week after enrolment.

Table 2: Table: Active learner counts at key time points (Day 0, 7, 14, 30).

day0	day7	day14	day30
2672	413	376	116
1409	167	165	36
1812	175	138	41
2472	158	137	63
2324	184	161	50
1989	103	67	21
1328	149	99	26

This confirms that engagement falls sharply after enrolment, with most dropout occurring within the first week.

5.2.2 Detecting inactivity (>= 7 consecutive days)

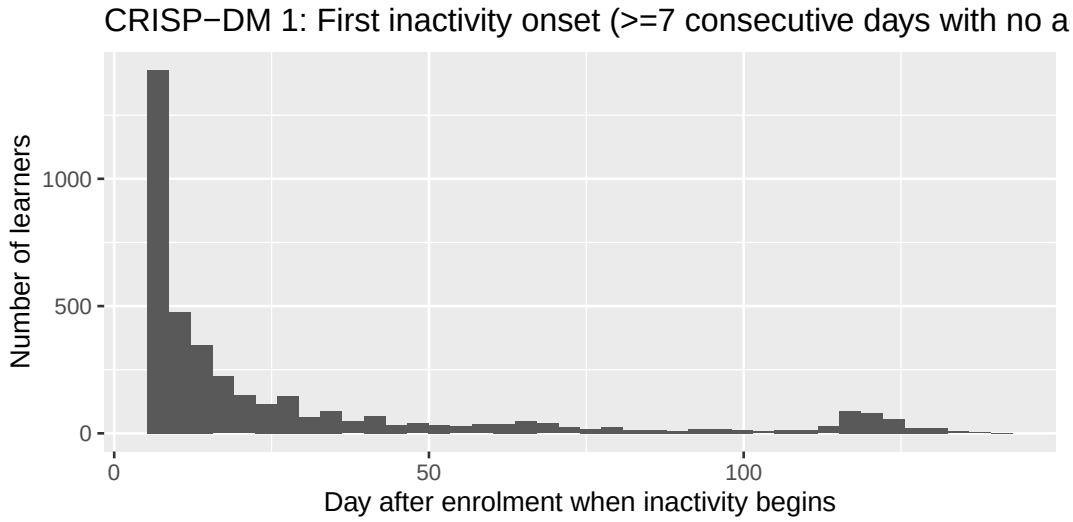

Figure 3: When learners first become inactive Despite differences in enrolment size

Interpretation: Engagement shows the same pattern across all runs—high activity immediately after enrolment followed by rapid decline—indicating this attrition is a consistent MOOC behaviour rather than a single-run anomaly.

Most inactivity onset happens within the first 1–2 weeks, indicating that learners are most vulnerable to disengagement during early course participation.

Interpretation: Most learners become inactive within the first 1–2 weeks (>=7 days no activity), confirming that early course participation—especially the first week—is the key retention period.

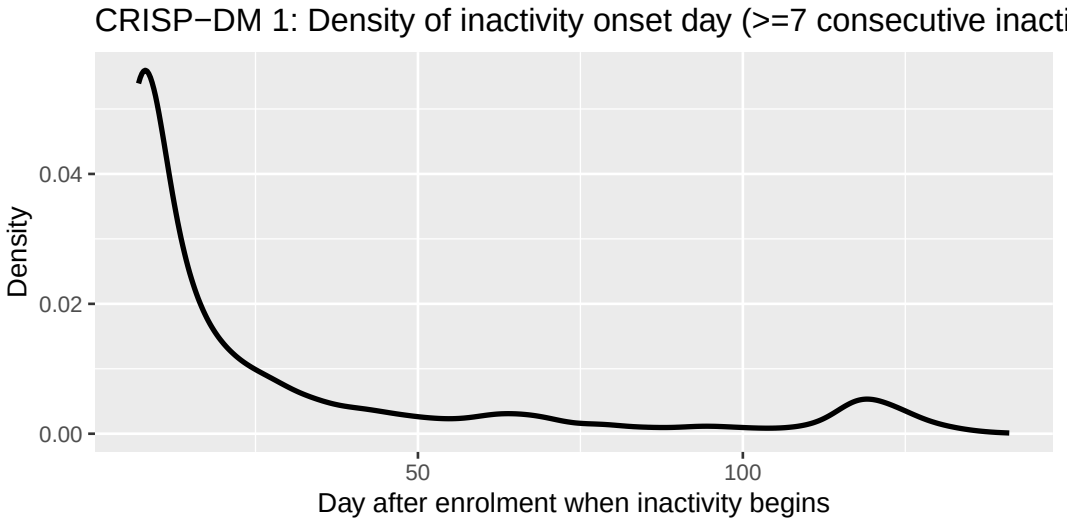

Figure 4: Density of inactivity

Table 3: Table: Summary of inactivity onset day (>=7 inactive days).

median_day	q1	q3	n_learners
13	7	32.5	3935

Interpretation : The summary confirms that inactivity begins early for most learners, and those who stay engaged initially are less likely to become inactive later.

5.5 Evaluation

Cycle 1 pinpoints where the point of dropout (critical dropout) is important (Days 1-7), while Cycle 2 is required in order to estimate recoverability by considering inactivity length effects on re-engagement probabilities for converters and non-converters.

Success Criteria Link: Cycle 1 satisfies its success criteria by creating a distinct decay engagement plot with measurable inactivity distribution. The findings indicate that learners become inactive early, helping justify the frequent identification of drop-out point at cycle one.

Next Step: One step that could follow is the implementation of early retention nudges (welcome, remind, progress) during Days 1–7, which has been indicated as the period of highest dropout risk from the previous Cycle 1.

6. CRISP-DM 2: Re-engagement probability

6.1 Business Understanding

6.1.1 Goal

Though CRISP-DM Cycle 1 points out when learners become inactive, it is important for MOOC providers to know whether these learners can be regained or whether they have stopped learning for good. Some learners have paused for a short term and resumed, while some have left for good. The aim of Cycle 2 is to measure:

- The likelihood of a learner returning after a period of inactivity how re-engagement probability varies with increasing durations of inactivity.
- Whether this behaviour varies between completers & non-completers.

RQ2: How does the probability of re-engagement change with inactivity length, and how does it differ for completers vs non-completers?

It defines a recoverability window — the time period where interventions are still likely to succeed.

6.1.1.1 Stakeholders

The key stakeholders include the providers of the MOOC platforms (FutureLearn retention teams), the staff members responsible for the courses being offered, and the learning analytics teams. The reason why these stakeholders benefit from the above analysis is because the intervention point for the inactive students is identified by the analytical process.

6.1.2 Future Aspect

The output of Cycle 2 can be directly implemented in the learning analytics system:

- A set of develop automated risk indicators for inactive learners
- define intervention triggers at 7 days, 14 days, etc.
- prioritize support resources efficiently
- lower the cost of interventions by excluding cases which are irrecoverable.

6.1.3 Success criteria

This cycle is successful if it produces:

1. A re-engagement probability curve as inactivity increases.
2. A comparison curve between completers and non-completers.
3. Identification of a practical threshold where re-engagement becomes highly unlikely.
4. Interpretable findings that can be translated into intervention timing policies
5. From a deployment point of view, the goal is met if the recoverability window can be identified that would trigger the intervention for recovery (e.g., 7-14 days of inactivity).

6.2 Data Understanding

This ensures re-engagement probabilities are calculated from the same validated engagement timelines used to detect inactivity onset in Cycle 1.

CRISP-DM Cycle 2 uses the same cleaned master datasets created in the preparation stage, particularly:

- 'master_dataset': step-level activity events aligned to enrolment time
- 'master_daily_timeline': daily learner engagement timeline used to calculate inactivity streak length
- 'completer' flag derived from `fully_participated_at` in enrolments data

Cycle 2 uses the same cleaned datasets as Cycle 1 (Table 1), containing timestamp variables, learner identifiers, and derived inactivity measures.

Re-engagement is considered an identifiable event within the activity log, where the learner visits any course activity recorded after a period of inactivity. As this is a log-based analysis, the metrics for this study focus on the interaction aspect instead of the learning quality.

6.3 Data Preparation

1. It is most likely that re-engagement will happen within the first 1–7 days of inactivity and becomes progressively unlikely after longer gaps.

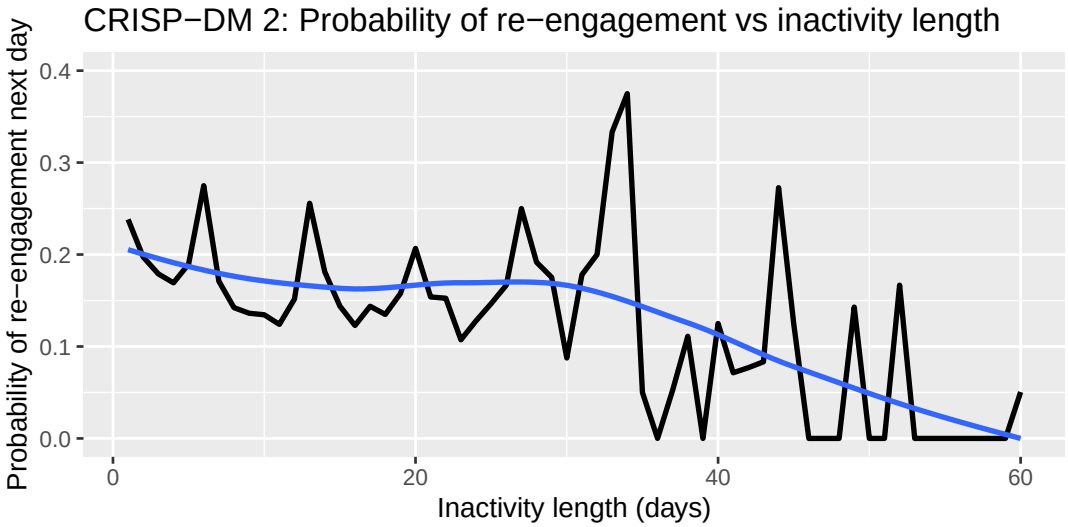

Figure 5: overall curve

Interpretation : Re-engagement probability is high after short inactivity periods but decreases as inactivity length increases, showing that early inactivity is recoverable while long inactivity indicates likely dropout.

Table 4: Table: Re-engagement probability at key inactivity durations.

inactivity_days	n	reengage_prob
1	18370	0.2382145
7	4503	0.1709971
14	1324	0.1812689
30	80	0.0875000
36	19	0.0000000

2. Re-engagement is most likely in the first 1–7 days of inactivity, and becomes progressively unlikely after longer gaps.

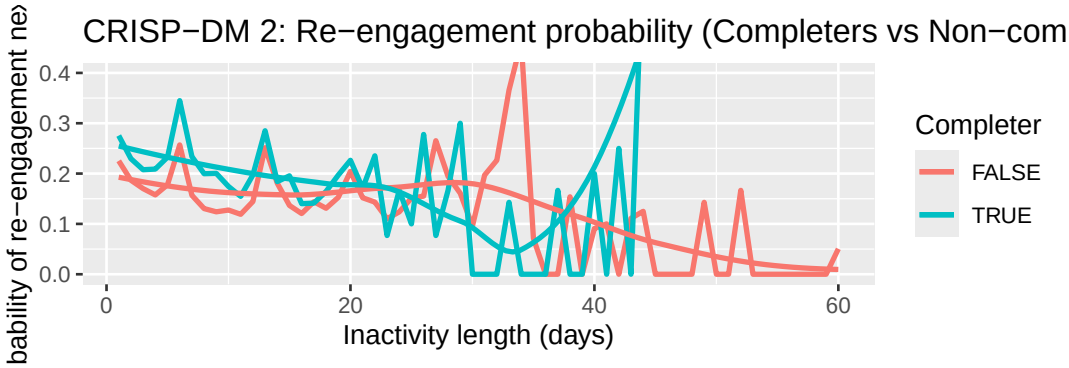

Figure 6: Re-engagement probability (Completers vs Non-completers)

Interpretation : Re-engagement is most likely after short inactivity, but declines as inactive days increase, highlighting a clear recoverability window for timely interventions.

Numeric Evidence : Completers show consistently higher re-engagement probability across all inactivity lengths, while non-completers decline faster, indicating stronger persistence among completers.

Table 5: Table: Re-engagement probability by completion status at key inactivity durations.

completer	inactivity_days	n	reengage_prob
FALSE	7	3673	0.1568200
FALSE	14	1161	0.1808786
FALSE	30	73	0.0958904
TRUE	7	830	0.2337349
TRUE	14	163	0.1840491
TRUE	30	7	0.0000000

Across all inactivity durations, completers retain consistently higher re-engagement probabilities than non-completers, suggesting stronger persistence behaviours.

6.4.1 Recoverability threshold

After ~36 consecutive inactive days, re-engagement becomes extremely unlikely.

Table 6: Table: First inactivity length where re-engagement probability reaches zero.

inactivity_days	n	reengage_prob
36	19	0

Interpretation: The re-engagement probability approaches zero at about 36 days of inactivity indicating that after this period, inactivity cannot be recovered from through mere nudges on the platform.

Numeric Evidence : After around 36 inactive days, re-engagement drops to 0, meaning learners inactive for over a month are unlikely to recover, so interventions should happen earlier.

6.3.1 Convert raw clickstream events into daily engagement

Each learner has at most 1 row per day showing:

- active = 1 if they did anything that day.

6.3.2 Create a full calendar timeline per learner

For each learner you now generate ALL days between:

- first activity date
- last activity date

Even if they didn't do anything.

6.3.3 Fill missing days as inactive

Now every learner-day is labelled:

- active = 1
- active = 0

6.3.4 Calculate inactivity streak length

For each inactive day:

- inactivity_days = 1, 2, 3 … until activity returns.

Example:active days: 1 1 0 0 0 1,

inactivity_days: 0 0 1 2 3 0

This directly answers "length of inactivity".

6.3.5 Define re-engagement

Re-engagement is a return event:

- learner inactive today.
- learner active tomorrow.

This becomes your dependent variable.

6.5 Evaluation

The CRISP-DM Cycle 2 model Appropriately addresses RQ2 since it shows that "the trend of re-engagement probability follows a decreasing trend with regards to time inactive, whereby a distinction was made between task completers and non-task completers, including the first 7 to 14 days of inactivity.

Success Criteria Link: The success criteria for Cycle 2 are met because Cycle 2 produces a total probability of re-engagement and a plot of completers versus non-completers. Cycle 2 also points out a threshold beyond which there would be little chance of re-engagement. That threshold is about 36 inactive days.

Next Step: A next step could be to set up triggers for automated interventions (such as after 7 and 14 days of inactivity) and start ceasing low-value attempts after approximately 36 days of inactivity since recovery at this point is very unlikely.

7. Discussion: Link between CRISP-DM 1 and CRISP-DM 2

7.1.1 "why the drop happens"

Engagement drops fast because MOOCs allow flexible learning, so many learners register but do not persist after initial curiosity.

7.1.2 "limitation in Cycle 2 probability"

Re-engagement probability may be underestimated because some learners may continue offline (e.g., watching videos without activity logs).

CRISP-DM 1 determines the time points where learners go inactive (>=7 days). But inactivity does not necessarily translate to permanent dropout. CRISP-DM 2 expands on the above results by identifying the probability of recovery, as well as the recovery window.

Practical implication: Cycle 1 provides evidence on which day inactivity starts, that is, early dropout risk. Cycle 2 informs on how recoverable those learners are. In combination, this sets an evidence-informed intervention schedule: reminders at 7 days inactivity and stronger escalation by 14 days, before learners enter the irrecoverable zone (~36 days).

These results enable FutureLearn to launch targeted retention efforts. For instance, the point where inactivity is prevalent is on Day 7; hence efforts to regain them should be started early (Days 3 to 7). Taking into consideration that recoverability is almost zero after 36 days of inactivity, resources aimed at supporting learners should be directed at learners only within this period.

Taken together, the two CRISP-DM instances offer an actionable story:

1. Learners Dropout considerably during the first period (Cycle 1)

2. Recovery is possible, though with reduced likelihood based on the time spent inactive (Cycle 2)

8. Industry Usefulness

This analysis can be deployed as an early-warning retention system:

1. **Monitoring component**

- Daily check learner inactivity streak length.

2. **Risk classification**

- Low risk: inactive 1–3 days.
- Medium risk: inactive 4–6 days.
- High risk: inactive >=7 days.
- Unrecoverable: inactive >=36 days.

3. **Automated interventions**

- Personalised email nudges.
- Recommended "easy re-entry" steps.
- Encouragement messages.
- Progress summaries.

4. **Future work**

- Include survey sentiment as predictor.
- Predict dropout using logistic regression/survival analysis.
- Evaluate intervention effectiveness via A/B testing.

9. Limitations and Ethics

Limitations

a. Activity logs measure platform interaction, not learning quality
b. Some inactivity may be external (work, health, time).
c. Small sample size for long inactivity days reduces stability.

Ethics

a. Learner IDs are anonymised.

b.Analysis is aggregated and does not profile individuals.

10. Conclusion

This report applied two CRISP-DM cycles to examine MOOC engagement and recovery behaviour. Engagement sharply declines after enrolment and many learners become inactive early. Re-engagement probability decreases with inactivity length and is higher for completers than non-completers. These findings strongly support early retention interventions, especially within the first week after enrolment. This would suggest that interventions on FutureLearn are best targeted in the week immediately following enrollment and within a fortnight of a learner becoming inactive. Beyond around 36 days, the interventions are unlikely to succeed; therefore, it would be beneficial to target resources toward learners with a higher potential to recover. This meets the success criteria in identifying the highest dropout-risk period and a practical recoverability window for intervention, but results are based on platform activity logs and cannot measure learning quality or external factors like personal time constraints.