FutureLearn Cyber Security Retention Analysis

Angeline Aurel Efendy

2026-01-10

Introduction

This report examines the implementation of the Cross-Industry Standard Process for Data Mining (CRISP-DM) on learner data from the Future Learn Cyber Security course (Run 6 & 7). Two cycles of the CRISP-DM process were conducted to identify patterns of learner engagement, retention, and dropout which are informative for course stakeholders.

The stakeholders for this analysis comprise the Future Learn platform team, course educators and learning designers, as well as educational institutions that utilize learning analytics to evaluate and enhance the quality of online courses. Engagement and retention of learners are crucial in online education, as high enrollment counts do not inherently result in excellent learning outcomes. Courses characterized by high early dropout rates may not fulfill their educational goals, despite strong initial learner engagement.

Understanding when learners disengage and the factors associated with early dropout is essential for enhancing course design. This report aims to provide practical insights into learners' experiences of the Cyber Security course by assessing behavioral engagement data in combination with survey-based feedback, and examining how these experiences correlate with patterns of retention and disengagement.

The specific research question to be examined are:

> **"Cycle 1: What are the patterns of participant engagement and retention in the Future Learn Cyber Security course (Runs 6 & 7), and at which stage or week do most participants drop out?"**

> **"Cycle 2: What factors are associated with early dropout, based on learners' stated reasons in the leaving survey (individual level), and how do weekly sentiment patterns correlate with dropout rates at the aggregate level (Runs 6 & 7)?"**

This is of interest because identifying when learners disengage and the factors associated with early dropout, can help stakeholders improve course design and learner support. This CRISP-DM analysis is successful if it pinpoints the key dropout weeks and the main reasons/sentiment patterns linked to early dropout, or shows that dropout is evenly spread and not clearly associated with these factors.

CRISP-DM Cycle 1 - Engagement & Retention

Business Understanding

The stakeholders for this analysis are the FutureLearn platform team and the Cyber Security course educators (and institutions using the course data for evaluation). This investigation helps them by showing where

learners disengage and drop out, enabling improvements to onboarding, workload, and course design in order to increase retention.

The primary goal of this first CRISP-DM cycle is to measure patterns of learner engagement and retention and to identify the course weeks with the largest drop-off. The insights from this cycle provide the foundation for a second CRISP-DM cycle, which explores potential explanatory factors associated with early dropout.

Data Understanding

This project uses learning analytics data from the Future Learn Cyber Security course (Runs 6 & 7). A "run" refers to a single delivery of the course to a cohort of learners at a specific point in time. While the course structure is largely consistent across runs, each run contains a distinct cohort of learners.In addition, learner identifiers are only unique within a given run, which requires analyses to distinguish between runs when combining data across course deliveries.

The datasets include enrollments (learner demographics and enrollment information), step activity (logs of when learners first visited and completed learning steps), question responses (quiz performance), video statistics (video viewing behavior), and two optional feedback sources: the leaving survey (reasons for leaving) and the weekly sentiment survey (weekly experience ratings). The data are longitudinal, meaning they track learner activity over time, and several tables contain multiple rows per learner because each learner can interact with many steps, quizzes, or videos.

There are some expected data quality limitations. Survey datasets are incomplete by design because participation is optional, so many learners have missing leaving reasons or sentiment responses. Some timestamp fields may be missing for certain activities, and demographic variables may include missing values or "Unknown" entries. In addition, the raw data include non-learner accounts (e.g., educators/mentors listed in the team-members file) that must be filtered out to avoid bias when estimating engagement and retention patterns.

Data Preparation

Data preparation was carried out using a series of scripts in the munge/ directory to ensure that the analysis was reproducible. All raw datasets from Runs 6 and 7 were first loaded from the data/ folder and combined, with an additional variable indicating the course run. Timestamp variables in the activity logs were converted to appropriate date–time formats where required. Because the raw data include accounts belonging to course staff (such as educators and mentors), these non-learner accounts were identified using the team-members data and removed on a per-run basis to avoid bias in engagement and retention estimates.

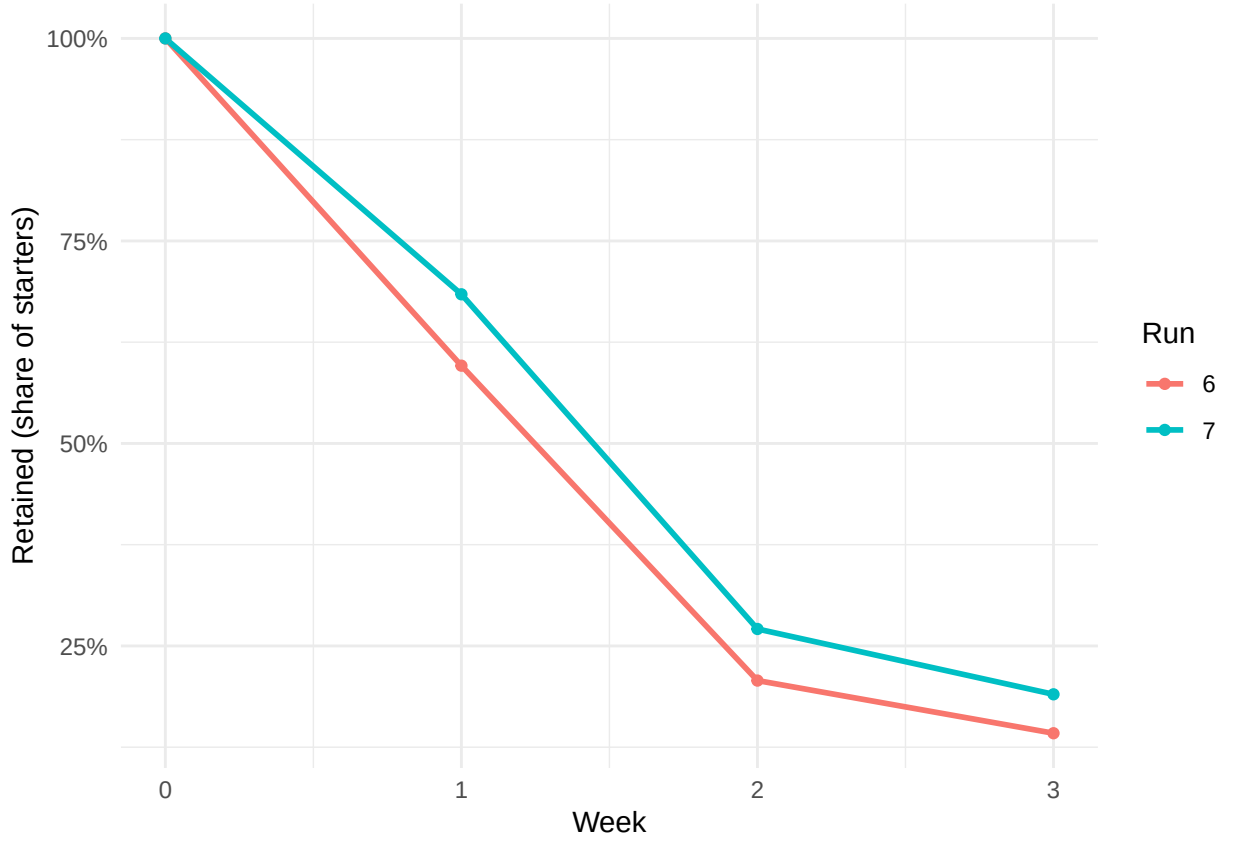
Learner activity data were then aggregated to the learner level within each run.. From the step activity logs, several derived variables were created, including whether a learner started the course, the total number of learning steps completed, and the last week in which any step was completed. These derived features were combined with cleaned enrolment data to produce a single record per learner. Finally, a weekly drop-off summary was created by counting the number of learners whose last completed activity occurred in each week of the course. These transformations produced clean, analysis-ready datasets that support the investigation of engagement, retention, and dropout patterns.

Table 1: Key engagement and retention metrics (Runs: 6, 7)

run	total_learners	learners_started	completion_rate	early_dropout_rate
6	3171	2292	10.3%	79.3%
7	2340	1624	13.2%	72.9%
Overall	5511	3916	12%	77%

Modelling

A retention model is applied to the learner activity data to estimate the proportion of learners who remain active across course weeks. The model allows retention trajectories to be compared between Runs 6 and 7 of the FutureLearn Cyber Security course. The resulting retention curves are shown in Figure 1.



The retention pattern across course weeks is shown in Figure 1, which reports the proportion of learners who started the course and remain active through each week, plotted separately for Runs 6 and 7. In both runs, retention declines sharply during the first week, indicating that disengagement is heavily concentrated at the beginning of the course. After this initial drop, the curves continue to decline but at a slower rate, suggesting that learners who persist beyond the early stage are more likely to continue. While the overall pattern is consistent across runs, Run 7 shows slightly higher retention than Run 6 at each subsequent week, indicating modest variation in early retention between cohorts.

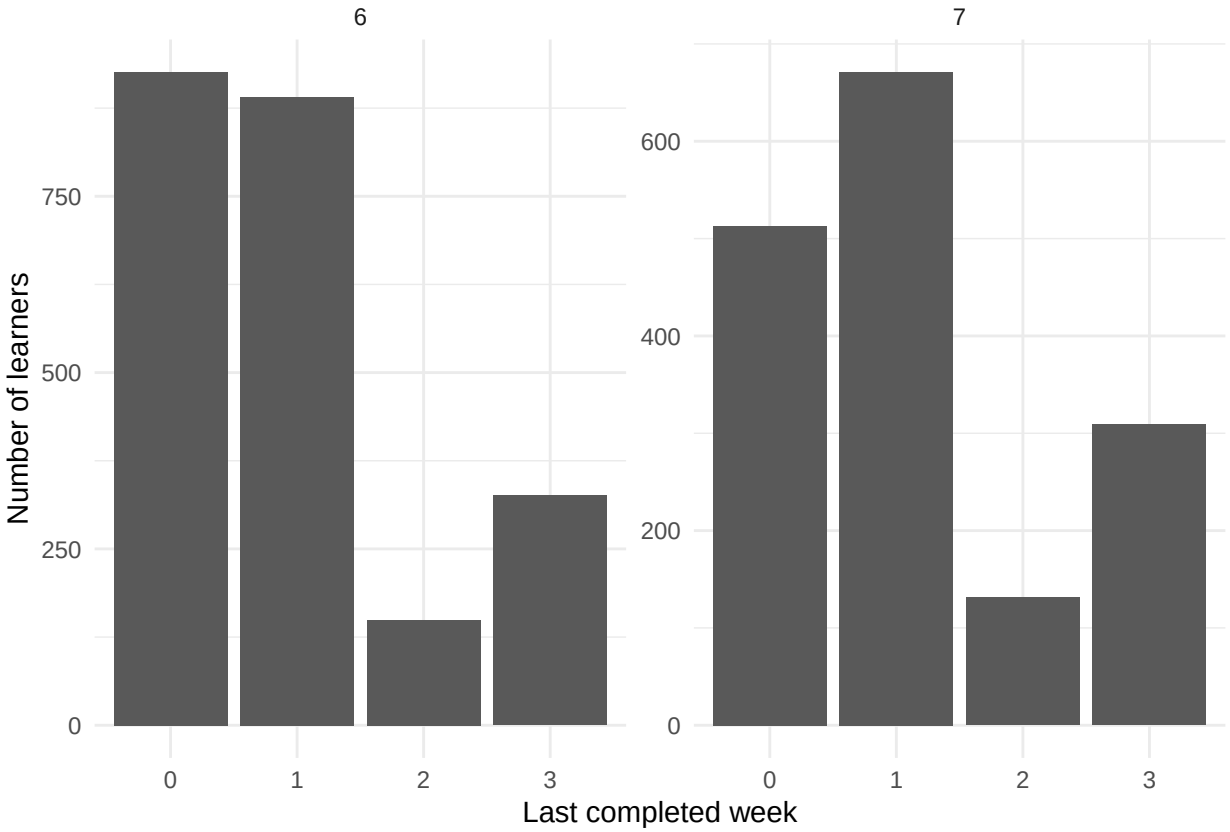
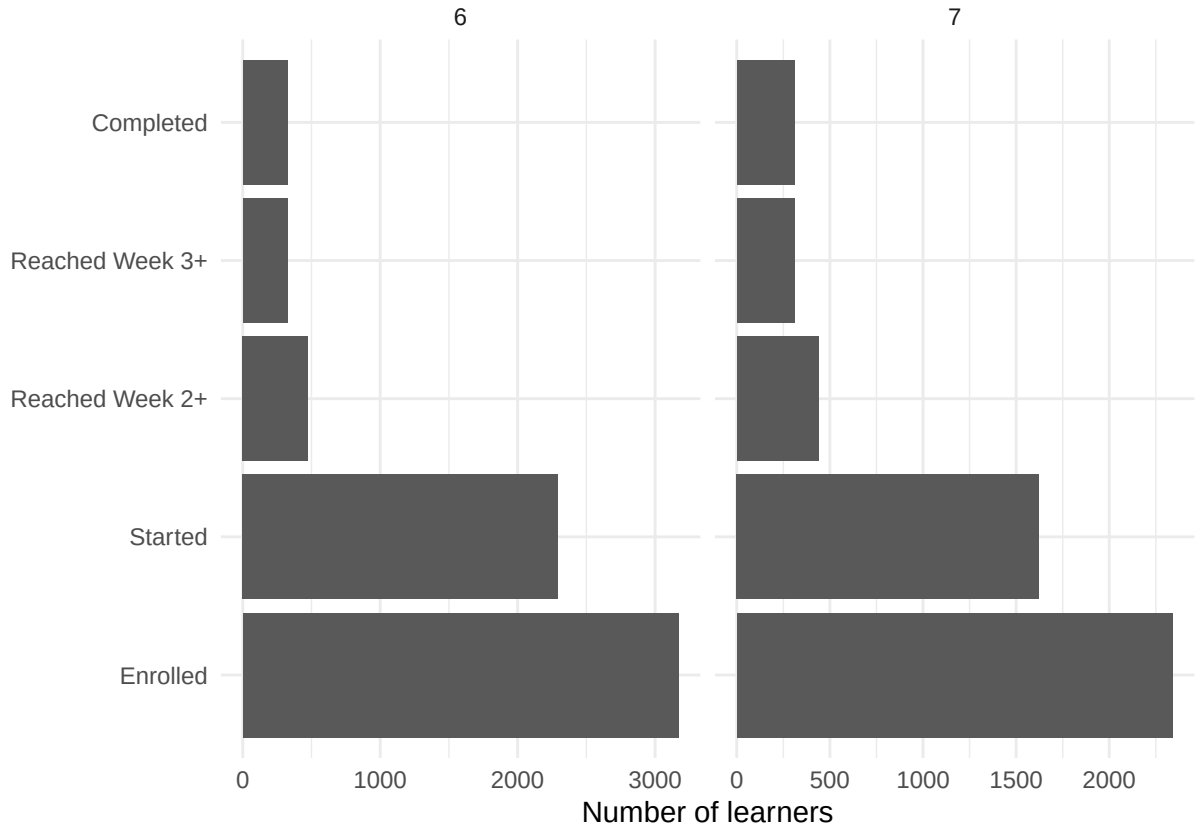
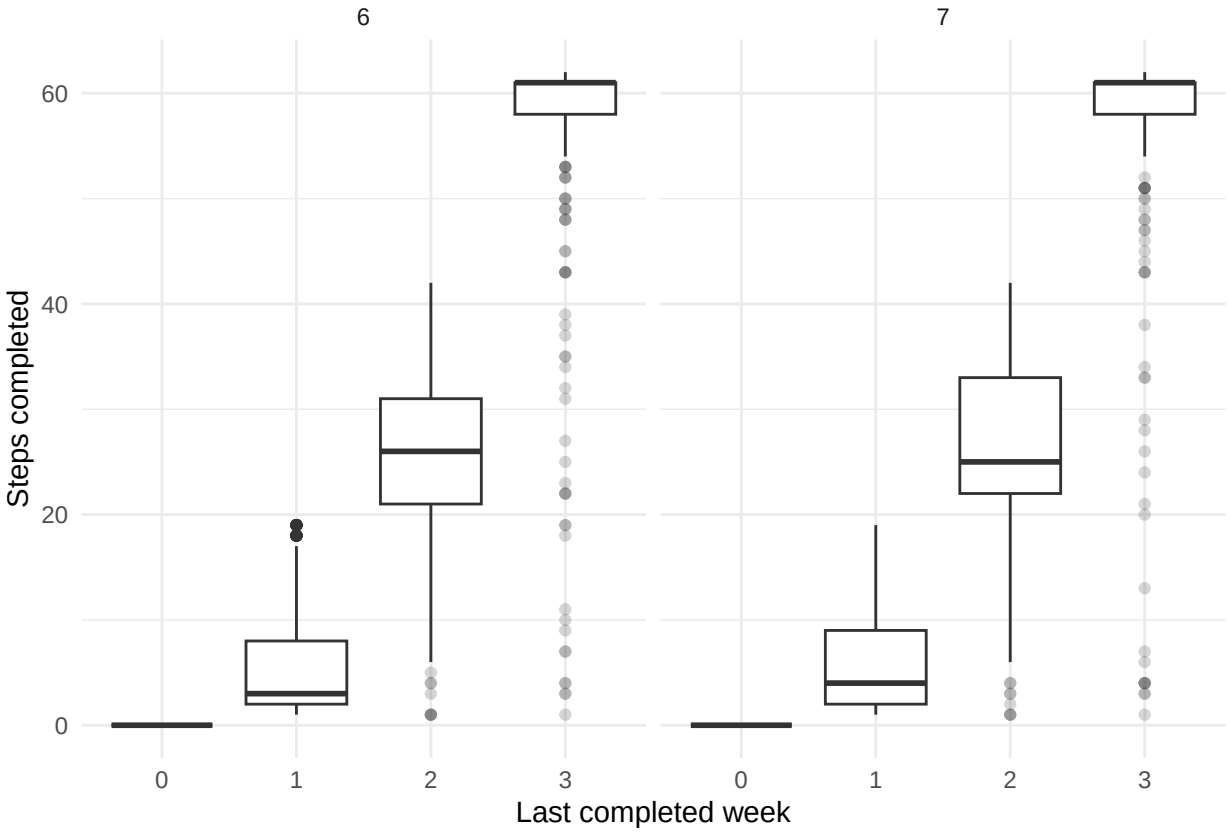


Figure 2 summarises drop-off by identifying each learner's last completed week, shown separately for Runs 6 and 7. In both runs, disengagement is heavily concentrated in the earliest stage of the course (Weeks 0–1), with substantially fewer learners progressing to later weeks. The modal drop-off week differs slightly by cohort: in Run 6 the largest drop-off occurs at Week 0, whereas in Run 7 it occurs at Week 1. This highlights the early weeks as the key "critical period" that should be prioritised for course redesign and learner support.



Idea: show the learner "flow" from *Enrolled → Started → Reached Week 2 → Reached Week 3 → Completed* This **is more engaging** and easier for non-technical audiences to understand.

Figure 3 presents a learner funnel showing how participation decreases across key milestones, plotted separately for Runs 6 and 7. In both runs, the largest loss occurs at the earliest transition from Enrolled to Started, indicating that many learners register but never begin engaging with course content. Among those who do start, participation declines sharply by the Week 2 milestone, reinforcing that the main retention challenge occurs very early in the learner journey. In relation to the research question, this strengthens the finding that disengagement is concentrated at the beginning of the course rather than gradually accumulating toward the end, suggesting that improvements to onboarding and Week 1 design may yield the greatest retention gains.



Idea: for each last_week, examine distribution of steps_completed.
This shows "what engagement looks like at a given week".

Figure 4 compares the distribution of completed learning steps across learners' last completed week, shown separately for Runs 6 and 7. In both runs, learners who drop out earlier complete substantially fewer steps, while learners who persist to later weeks exhibit markedly higher levels of engagement. The median number of completed steps increases monotonically with the last completed week, indicating a strong association between engagement intensity and retention. In relation to the research question, this reinforces the finding that disengagement emerges very early in the course and that higher engagement is closely linked to continued participation.

> The concentration of drop-off in the earliest weeks suggests that the main retention problem occurs early in the learner journey. This motivates the second CRISP-DM cycle, which explores potential explanations for early dropout using leaving survey responses (individual-level) and weekly sentiment ratings (aggregate-level).

The retention curve in Figure 1 shows a steep early decline: only 20.7%, 27.1% of learners who started remain active up to Week 2, after which the curve flattens, suggesting that learners who pass the initial stage are more likely to continue. Consistent with this, the drop-off distribution in Figure 2 indicates that disengagement is most concentrated at Week 0, which represents 24% of all drop-outs among starters. Engagement also differs sharply by drop-off timing (Figure4): learners who leave by Week 1 complete a median of 1 steps, compared with 59 steps for those who persist beyond Week 1.

Evaluation

The exploratory modelling shows that engagement and retention decline sharply in the early weeks of the Cyber Security course, with the highest drop-off occurring within Week 0–1. This analysis therefore answers the Cycle 1 question by identifying the critical dropout stage, but it does not explain why learners leave, which motivates Cycle 2.

CRISP-DM Cycle 2

Business Understanding

The business understanding remains largely unchanged from the first CRISP-DM cycle. The same stakeholders namely the FutureLearn platform team and Cyber Security course educators require insight into learner engagement and retention in order to improve course design and learner support. Building directly on the findings from Cycle 1, which showed that learner dropout is heavily concentrated in the earliest weeks of the course across both Runs 6 and 7, the focus of Cycle 2 is to understand why learners disengage early. The objective of this cycle is therefore to identify factors associated with early dropout, using learner-reported reasons for leaving and weekly sentiment feedback, in order to generate actionable insights that can inform targeted interventions during the early stages of the course.

Data Understanding

This CRISP-DM cycle uses the same core learner activity data as Cycle 1, supplemented with two feedback datasets: the leaving survey (individual-level) and the weekly sentiment survey (aggregate-level), for Runs 6 and 7.

Not all learners completed the leaving survey, so missing values in survey-related variables are expected and reflect the optional nature of the feedback. Weekly sentiment data do not include learner identifiers and are therefore analysed at the weekly aggregate level rather than at the individual level. As data quality and structure were already assessed in Cycle 1, no additional exploratory data analysis is required before proceeding to modelling.

Data Preparation

Data preparation for the second CRISP-DM cycle builds directly on the derived variables created in Cycle 1. Learners were classified into early and later dropout groups based on the last week in which they completed any course activity.

Leaving survey responses were then joined to the learner-level retention data and summarised by dropout group to explore differences in stated reasons for leaving. Weekly sentiment responses were aggregated by week and combined with weekly dropout counts to examine whether changes in learner sentiment align with observed dropout patterns across Runs 6 and 7. No additional data cleaning or transformation was required beyond these steps.

Modelling

In the second CRISP-DM cycle, descriptive modelling is applied to explore factors associated with early dropout in the Cyber Security course. Two complementary approaches are used: (1) analysis of learner-reported leaving reasons at the individual level, and (2) analysis of weekly sentiment trends in relation to dropout patterns at the aggregate level. These models build directly on the findings from Cycle 1, which identified that dropout is heavily concentrated in the early weeks of the course.

Table 2: Leaving reasons by dropout group (Early vs Later), Runs 6 and 7

early_dropout	leaving_reason	n
Early	I don't have enough time	23
Early	The course wasn't what I expected	11
Later	Other	11
Early	Other	9
Early	I prefer not to say	7
Early	The course required more time than I realised	6
Later	I don't have enough time	6
Later	I prefer not to say	5
Early	The course won't help me reach my goals	3
Early	The course was too easy	2
Later	The course required more time than I realised	2
Later	The course won't help me reach my goals	2
Early	The course was too hard	1
Later	The course was too easy	1
Later	The course wasn't what I expected	1

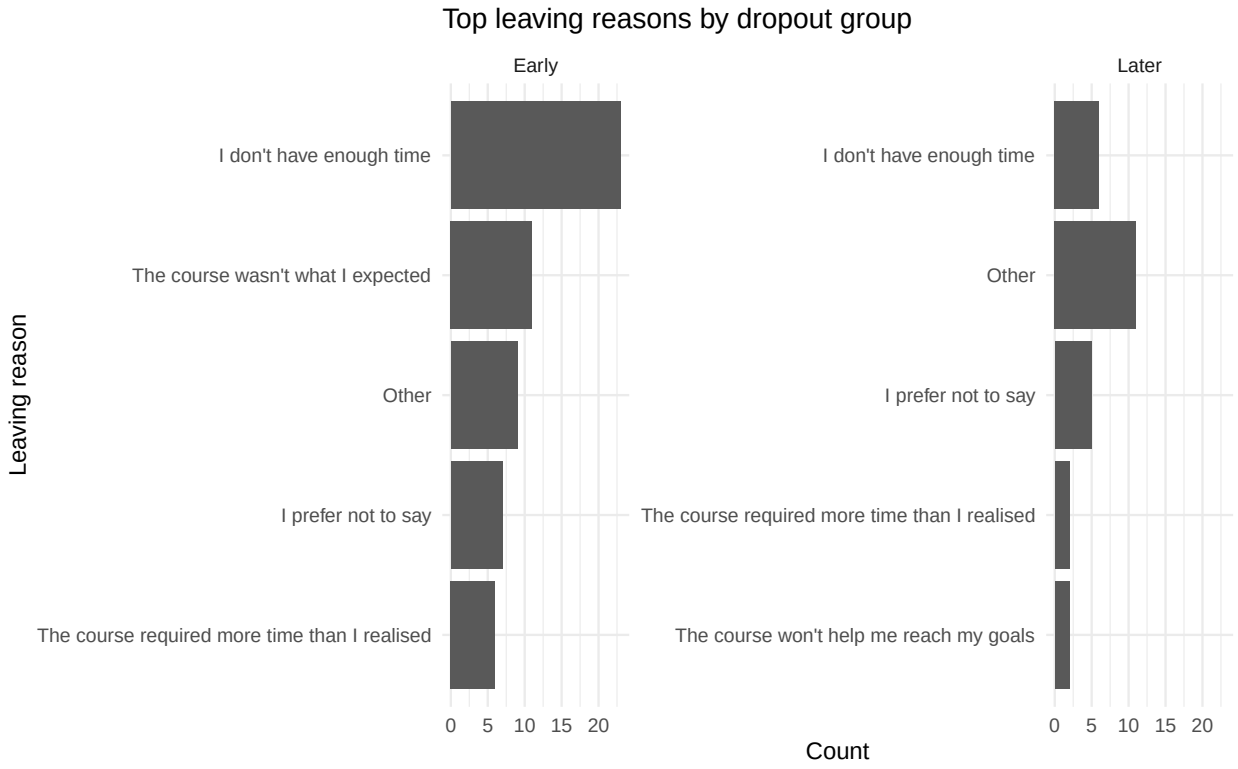


Figure 5 presents the most common leaving reasons reported by learners, split between early dropouts (Weeks 0–1) and later dropouts (after Week 1). For early dropouts, the dominant reason is "I don't have enough time", followed by expectations mismatch and underestimation of the time required. This suggests that early disengagement is strongly linked to workload expectations and time constraints rather than dissatisfaction with specific course content.

In contrast, learners who dropped out later report a more diverse set of reasons, including general or unspecified factors ("Other") and reluctance to disclose a specific reason. This indicates that later dropout may

be driven by more heterogeneous or personal factors, whereas early dropout is more consistently associated with time-related pressures. Overall, this supports the interpretation that early dropout is closely tied to onboarding and initial workload expectations.

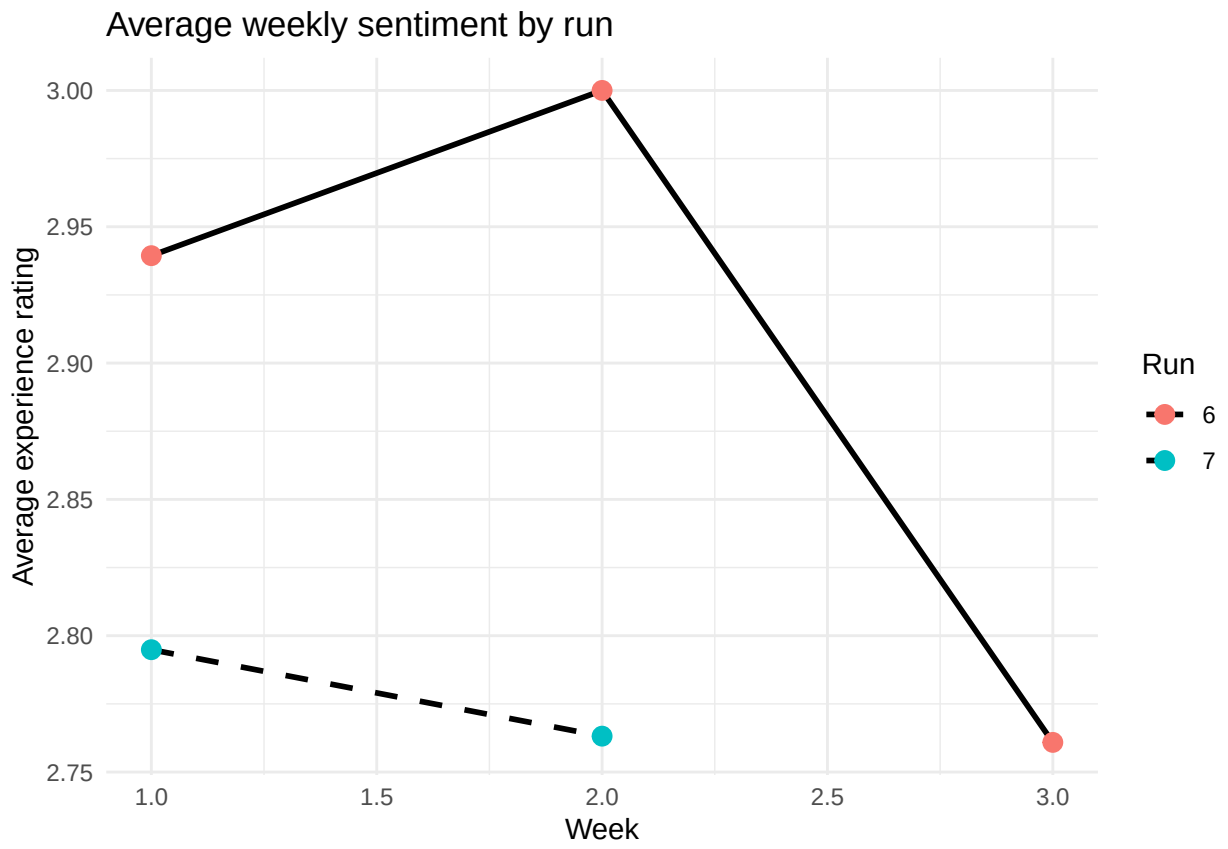


Figure 6 shows the average weekly experience rating (sentiment) for Runs 6 and 7. In both runs, sentiment levels are relatively stable but show some variation across weeks. Run 6 displays a slight increase in sentiment from Week 1 to Week 2, followed by a decline in Week 3. In contrast, Run 7 shows a modest downward trend across the observed weeks.

Although sentiment differences between runs are not large, the pattern suggests that learner experience may deteriorate slightly as the course progresses, particularly after the early weeks. This aligns with the retention findings from Cycle 1, where substantial dropout occurs before later weeks are reached.

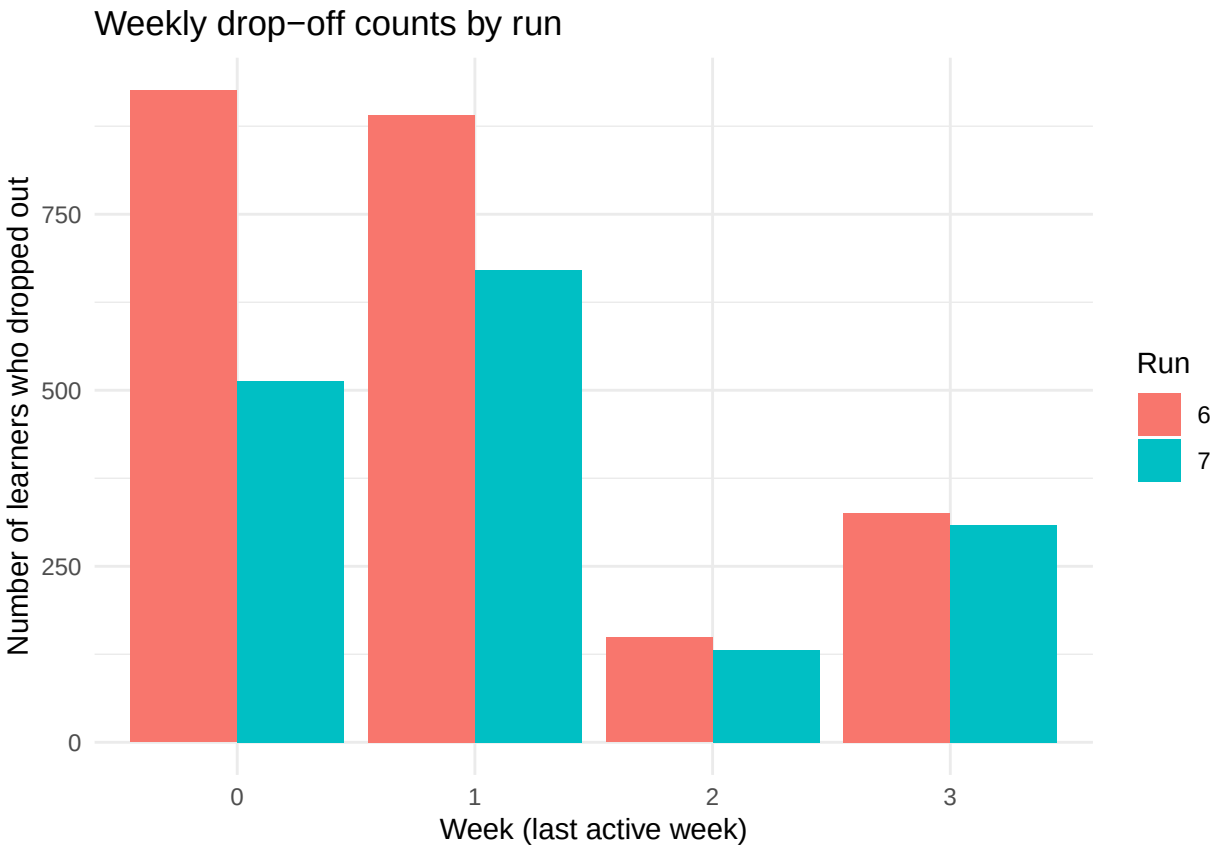


Figure 7 illustrates the absolute number of learners who dropped out in each week, shown separately for Runs 6 and 7. In both runs, the highest number of drop-offs occurs in Weeks 0 and 1, with substantially fewer learners leaving in later weeks. While Run 6 shows a higher overall volume of early dropouts, the general shape of the distribution is similar across runs. This reinforces the conclusion that dropout is heavily front-loaded and that the early stages of the course represent a critical risk period for learner disengagement.

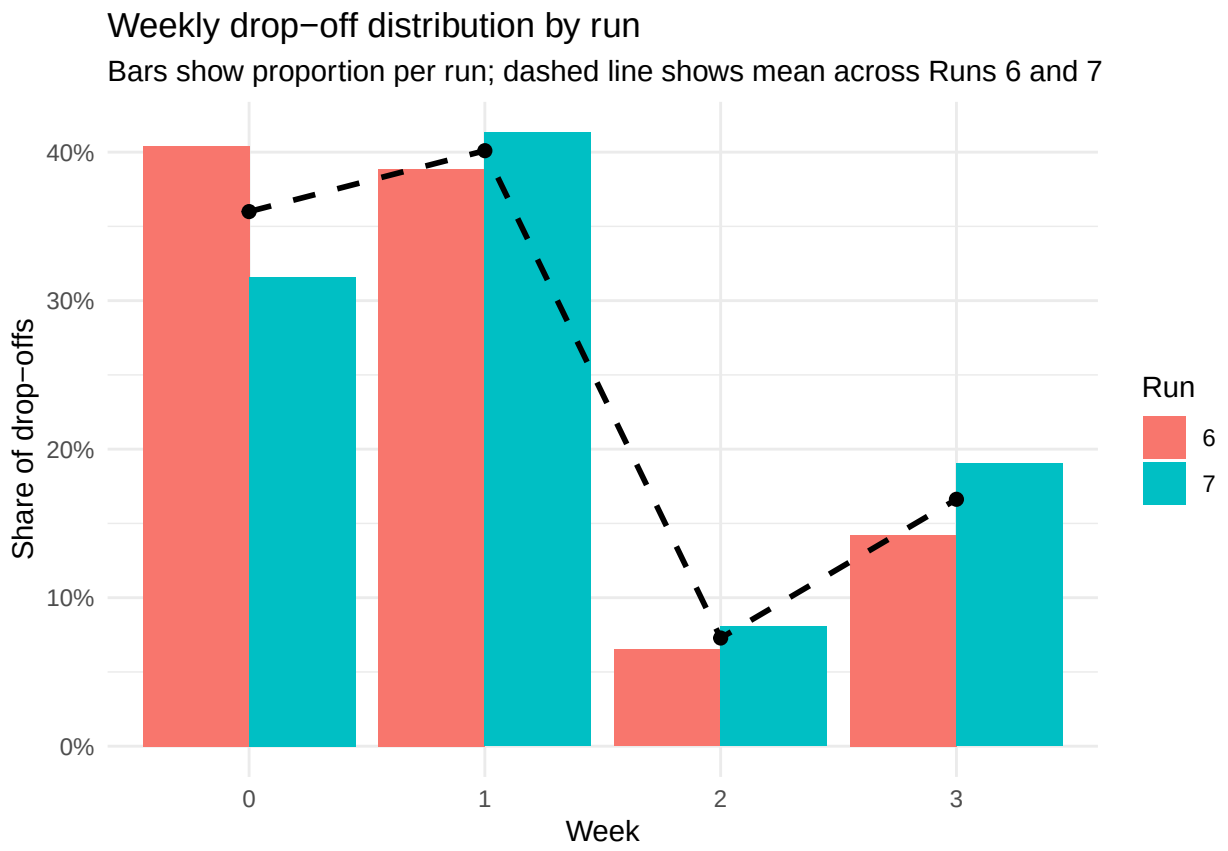


Figure 8 presents the proportional distribution of weekly drop-offs by run, with a dashed line indicating the mean proportion across Runs 6 and 7. The figure highlights that a large share of all dropouts occurs in the first two weeks, even after accounting for differences in cohort size. The mean trend shows a sharp decline after Week 1, followed by a smaller increase in later weeks.

Taken together with the sentiment results, this suggests that early negative or unmet expectations may coincide with the period in which most learners decide to disengage, even though sentiment data are only available for a subset of weeks.

Evaluation

After completing the second CRISP-DM cycle, the research question has been addressed. The analysis shows that early dropout is strongly associated with learner-reported time constraints and mismatched expectations, as revealed by the leaving survey at the individual level. At the aggregate level, weekly sentiment patterns and drop-off distributions consistently indicate that disengagement is concentrated in the earliest weeks of the course. While sentiment differences across runs are modest, the combined evidence confirms that the early course experience plays a central role in learner retention.

Deployment

The main takeaway from this project is that learner dropout in the Cyber Security course is heavily concentrated in the earliest weeks (Weeks 0–1) across Runs 6 and 7. At the individual level, leaving survey responses indicate that early dropout is most commonly associated with time constraints and workload expectations. At the aggregate level, weekly sentiment shows small changes across weeks, but the overall pattern is consistent with the early drop-off concentration identified in Cycle 1.

These findings should be communicated to FutureLearn stakeholders and course designers to support targeted interventions in the early learner journey, such as stronger onboarding guidance, clearer messaging about expected weekly workload before learners start, and additional support during Week 1 (e.g., reminders, scaffolding, or simplified early activities). Implementing and monitoring these changes in future runs could help improve retention and learner experience in large-scale online courses.

Final conclusions

The analysis has been successful in answering the research questions through two CRISP-DM cycles using FutureLearn Cyber Security data from Runs 6 and 7. Cycle 1 showed that learner disengagement is heavily concentrated in the earliest stage of the course (Weeks 0–1), with only a small proportion of learners progressing to later weeks. Building on this, Cycle 2 provided explanatory insight into early dropout: leaving survey responses indicate that early dropouts most commonly cite time constraints and workload expectations, while weekly sentiment patterns show modest variation across weeks that is broadly consistent with the early drop-off concentration. Further work should test targeted interventions aimed at the Week 1 learner journey (e.g., clearer workload messaging, onboarding support, pacing guidance) and evaluate their impact in future runs. Where possible, richer learner-level predictors (e.g., step completion intensity, quiz attempts, and video engagement in the first week) could be incorporated to better identify at-risk learners and support early retention.