Predicting MOOC Completion from Early Engagement Patterns

Analysis of FutureLearn Cyber Security Course Data

Nepoliyan Ria

January 16, 2026

1 Executive Summary

This report analyzes seven runs of the FutureLearn MOOC "Cyber Security: Safety at Home, Online, and in Life" using two CRISP-DM cycles to investigate whether early engagement predicts course completion.

Key Findings:

- **Completion rate:** 5.8% overall; 56.3% of enrolled learners engage with at least one step
- **Week 1 as predictor:** High Week 1 engagers (75% complete) achieve 26.2% completion vs. 0.1% for low engagers—a 198.7× difference
- **Intervention threshold:** Learners completing <40% of Week 1 should be flagged for targeted support by day 7

These findings enable early identification of at-risk learners and timely intervention strategies.

2 Introduction

2.1 Context and Motivation

MOOCs have democratized education but face persistent high dropout rates. Understanding what distinguishes completers from early disengagers is critical for improving course design and learner support. Learning analytics—"the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning"—provides tools to identify patterns that predict success and inform interventions.

2.2 Dataset Description

This analysis uses data from **seven runs** of the FutureLearn MOOC *Cyber Security: Safety at Home, Online, and in Life*. The combined dataset consists of two primary tables:

- **Enrollment data: 37,296** rows × **14** columns
- **Step-level activity data: 423,072** rows × **8** columns

The enrollment table contains one record per learner per course run, including enrollment identifiers, course completion status, completion timestamps, and limited demographic variables (age range, gender, education level, and employment status).

The step-level activity table records individual learner interactions with course steps, including step identifiers, visit timestamps, and completion indicators. These data enable fine-grained measurement of engagement intensity, persistence, and timing throughout the course.

As is typical for MOOC datasets, engagement is highly skewed: many enrolled learners never interact with course content, while a smaller subset accounts for most activity. Demographic variables contain substantial missingness, reflecting optional profile completion. Consequently, the analysis focuses primarily on behavioural engagement patterns—particularly early-course activity—rather than demographic predictors.

2.3 Investigation Aim

Research Question: Can early-course engagement behavior predict MOOC completion?

Approach: Two CRISP-DM cycles—(1) explore overall engagement patterns; (2) investigate Week 1 as predictor.

3 CYCLE 1: Overall Engagement and Completion Patterns

3.1 Business Understanding

- **Objectives:** Understand completion patterns to improve learner outcomes, optimize course design, and target support effectively.
- **Data Mining Goals:** Define completion; quantify engagement; compare completers vs. non-completers; identify dropout timing.
- **Success Criteria:** Articulate distinguishing factors; identify predictive metrics; generate Cycle 2 hypotheses.

3.2 Data Understanding

3.2.1 Initial Data Exploration

The dataset combines information from multiple sources. Table 1 provides an overview of enrollment and completion statistics across all seven course runs:

Table 1: Overall Enrollment and Completion Statistics

Metric	Value
Total Enrollments	37,296
Active Learners (1 step)	20,991
Learners Who Completed	2,154
Activation Rate	56.3%
Completion Rate (all enrolled)	5.8%
Completion Rate (active only)	9.8%

Key observations:

- A substantial proportion (43.7%) of enrolled learners never engage with any course content, representing a significant "activation gap"
- Among those who do engage with at least one step, the completion rate is notably higher (9.8% vs. 5.8%), suggesting that initial activation is a critical hurdle
- The overall completion rate of 5.8% aligns with typical MOOC completion rates reported in the literature (5-15%), confirming this dataset is representative of broader MOOC patterns

3.2.2 The Completion Funnel

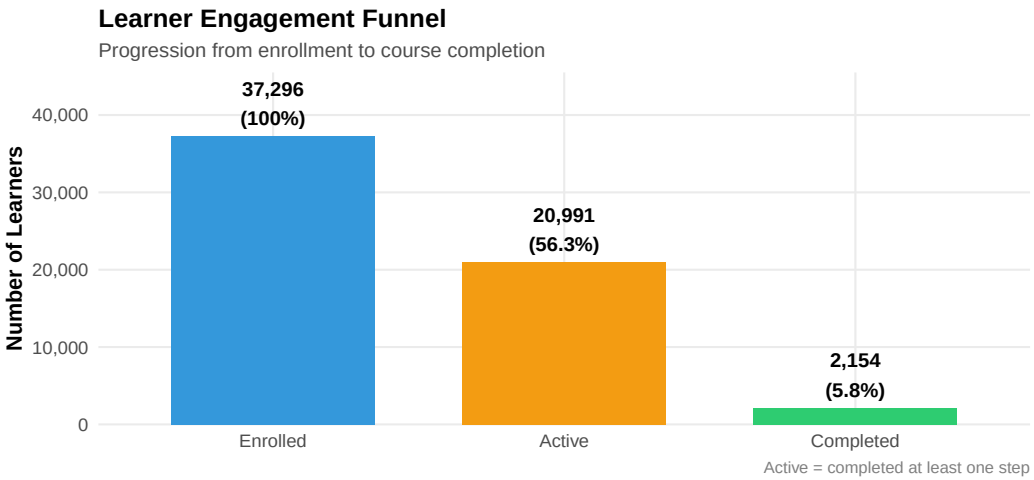

Figure 1: Learner progression from enrollment to completion shows significant dropoff at each stage

Figure 1 illustrates dramatic attrition from enrollment to completion, highlighting two critical transition points: enrollment to activation (many enroll but never start) and activation to completion (even active learners mostly don't finish).

3.2.3 Distribution of Engagement

Figure 2: Bimodal distribution shows clear separation between early dropouts and committed learners

Figure 2 reveals a **bimodal distribution**: learners who drop out after minimal engagement (1-10 steps) versus those who engage substantially or complete the course. This suggests two distinct populations with fundamentally different engagement trajectories.

3.2.4 Engagement Metrics Comparison

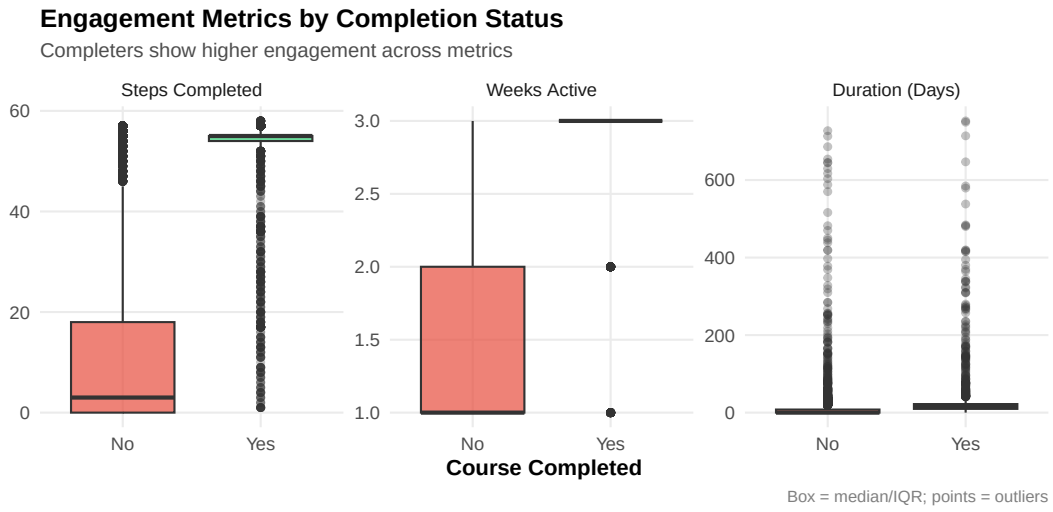

Figure 3: Completers show consistently higher engagement across all metrics

Table 2: Engagement Metrics by Completion Status (Active Learners Only)

Status	N	Mean Steps	Median Steps	Mean Weeks
Did Not Complete	18,926	13.4	3	1.5
Completed	2,065	50.2	55	2.9

Figure 3 and Table 2 demonstrate completers consistently outperform non-completers across all engagement dimensions. Median values show particularly stark differences, indicating completion is associated with sustained, substantial engagement rather than sporadic activity.

3.2.5 Temporal Patterns: When Do Learners Drop Out?

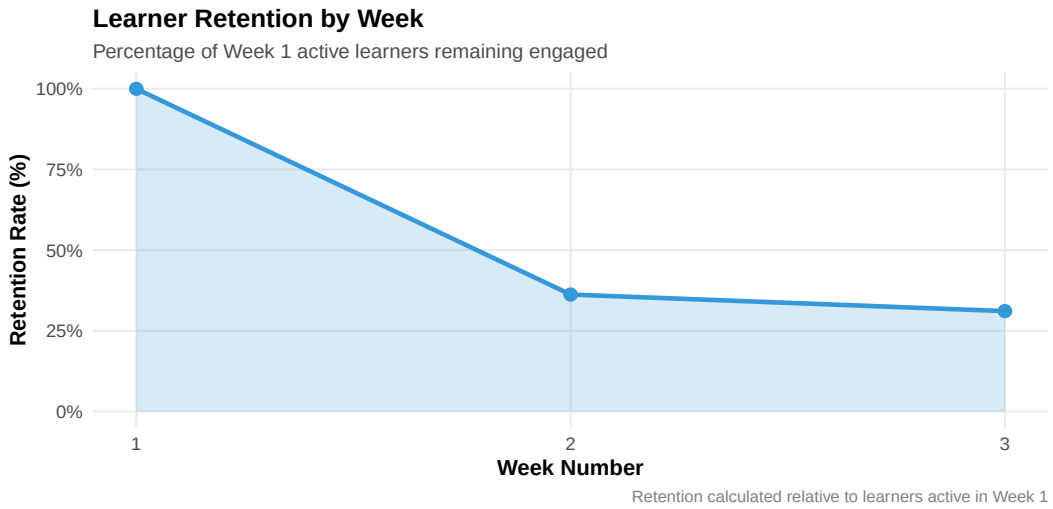

Figure 4: Retention drops sharply in early weeks, then stabilizes among committed learners

Figure 4 shows a characteristic MOOC retention curve: steep initial decline in Weeks 1-3, then stabilization among learners who persist beyond Week 3. Early weeks are critical for retention.

3.2.6 Data Quality Considerations

Demographic data shows substantial missingness (gender: 33137, age: 33268, education: 33161), limiting subgroup analysis. This is typical for MOOC data where profile completion is optional.

3.3 Data Preparation

3.3.1 Defining Completion

A critical decision in this analysis is how to operationalize "course completion." We considered two approaches:

Approach 1 (Selected): A learner is considered to have "completed" the course if they have a non-missing `fully_participated_at` timestamp in the enrollment data. This approach:

- Uses FutureLearn's own completion criteria, which is the platform's authoritative definition
- Is objective and verifiable, based on system-recorded timestamps
- Likely requires completing a substantial majority of course content, not just accessing materials

Approach 2 (Not used): Define completion as reaching 80% of total steps completed. While this threshold-based approach offers flexibility, the `fully_participated_at` flag is more authoritative and specifically accounts for FutureLearn's completion requirements, which may include additional criteria beyond step completion (e.g., quiz performance, discussion participation).

We selected Approach 1 to align with the platform provider's own definition of success and to ensure our findings are directly relevant to FutureLearn's business objectives.

3.3.2 Data Cleaning Steps

All data preprocessing was performed programmatically to ensure reproducibility:

1. **Combining course runs:** Merged data from 7 separate course runs into unified datasets using `map_dfr()` to preserve run identifiers
2. **Duplicate handling:** Identified and removed 29015 duplicate learner-step records by keeping the most recent interaction for each unique learner-step combination
3. **Timestamp parsing:** Converted all timestamp columns to proper POSIXct datetime format in UTC timezone, handling empty strings as missing values
4. **Missing value treatment:**
    - Learners with no activity: Set engagement metrics (steps completed, weeks active) to 0 to distinguish them from truly missing data
    - Demographic "Unknown" values: Converted to explicit NA for proper missing data handling and accurate missingness reporting
5. **Derived variables:** Created binary `step_completed` flag (based on presence of `last_completed_at` timestamp) and calculated per-learner summary statistics including total steps completed, weeks active, and engagement duration

Week 1 Metrics (for Cycle 2): Additionally created Week 1-specific variables: `week1_steps_completed`, `week1_completion_rate`, and `week1_engagement_level` with five ordered categories (None/Very Low/Low/Medium/High) based on completion percentage thresholds (<1%, 1-9%, 10-39%, 40-74%, 75%). These thresholds were chosen to create meaningful behavioral distinctions while maintaining reasonable group sizes for analysis.

All data processing was performed using `dplyr` pipelines in the `munge/` scripts, ensuring full reproducibility and transparency.

3.4 Modeling (Cycle 1)

3.4.1 Descriptive Analysis

Rather than predictive modeling, Cycle 1 focuses on descriptive statistics and visualization to understand patterns.

Table 3: Engagement Statistics by Completion Status

completed	N	Steps (Mean)	Steps (Median)	Weeks (Mean)
Did Not Complete	18,926	13.4	3	1.5
Completed	2,065	50.2	55	2.9

Effect sizes: Completers complete 3.7× more steps and are active 1.9× more weeks than non-completers. These large effect sizes confirm completion is associated with fundamentally different engagement patterns.

3.5 Evaluation (Cycle 1)

3.5.1 Key Findings

1. **The activation gap:** 43.7% of enrolled learners never engage—a significant intervention opportunity
2. **Bimodal engagement:** Two distinct groups (early dropouts vs. persisters) with little middle ground
3. **Early dropout concentration:** Steepest retention decline in Weeks 1-3, then stabilization
4. **Consistent differences:** Completers show higher engagement across all measured dimensions

3.5.2 Implications for Cycle 2

Temporal analysis (Figure 4) reveals **Week 1 is particularly critical** for retention, motivating Cycle 2 investigation:

Research Question for Cycle 2: Is Week 1 engagement specifically predictive of course completion?

If Week 1 engagement reliably predicts completion, this enables early identification of at-risk learners and timely intervention.

3.5.3 Limitations

- **Demographic data quality:** High missingness limits demographic analysis. * **Self-selection bias:** Cannot account for underlying motivation differences.
- **Causal inference:** Observational data cannot establish whether higher engagement causes completion.
- **Platform-specific:** Findings may not generalize to other platforms or topics.

4 CYCLE 2: Week 1 Engagement as a Predictor

4.1 Business Understanding

Refined Objectives: Build early warning system; enable targeted support; focus resources on at-risk learners identified by Week 1 engagement.

Data Mining Goals: Quantify Week 1 engagement; establish relationship with completion; determine actionable thresholds.

Success Criteria: Demonstrate clear relationship; identify thresholds; provide evidence-based recommendations.

4.2 Data Understanding (Cycle 2)

4.2.1 Week 1 Engagement Overview

Week 1 contains 18 steps on average, introducing foundational concepts. 55.7% of enrolled learners engage with Week 1 content.

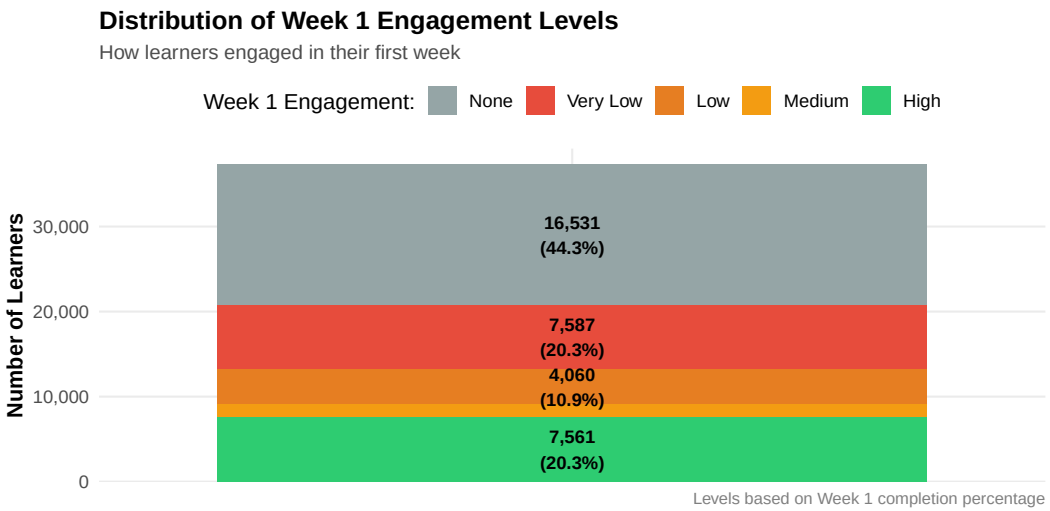

Figure 5: Distribution of learners across Week 1 engagement levels

Table 4: Distribution of Week 1 Engagement Levels

Week 1 Engagement	N	% of All Learners
None	16,531	44.3%
Very Low	7,587	20.3%
Low	4,060	10.9%
Medium	1,557	4.2%
High	7,561	20.3%

Among those who engage, Week 1 completion varies widely. A substantial group (%) complete 75% of Week 1, showing strong initial commitment.

4.2.2 Week 1 Engagement by Completion Status

Table 5: Week 1 Engagement by Eventual Completion Status

Status	N	Mean Week 1 Steps	Mean Week 1 Rate (%)
Did Not Complete	18,705	6.9	38.6
Completed	2,060	16.7	97.3

Table 6 shows learners who eventually complete engage substantially more in Week 1 than those who don't, suggesting Week 1 behavior is predictive.

4.3 Modeling (Cycle 2)

4.3.1 Primary Analysis: Completion Rate by Week 1 Engagement Level

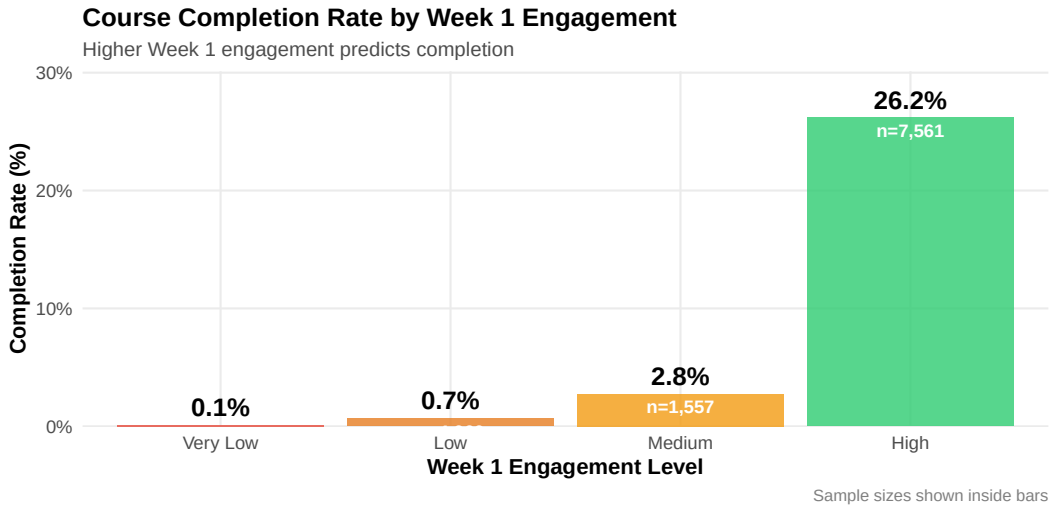

Figure 6: Clear stepwise relationship between Week 1 engagement and course completion

Table 6: Completion Rates by Week 1 Engagement Level

Week 1 Engagement	N	Completion Rate
Very Low	7,587	0.1%
Low	4,060	0.7%
Medium	1,557	2.8%
High	7,561	26.2%

Key finding: Clear, monotonic relationship between Week 1 engagement and completion (Figure 6, Table 7). High Week 1 engagers achieve 26.2% completion vs. 0.1% for very low engagers—a 198.7× difference in completion probability.

4.3.2 Granular Relationship: Steps Completed

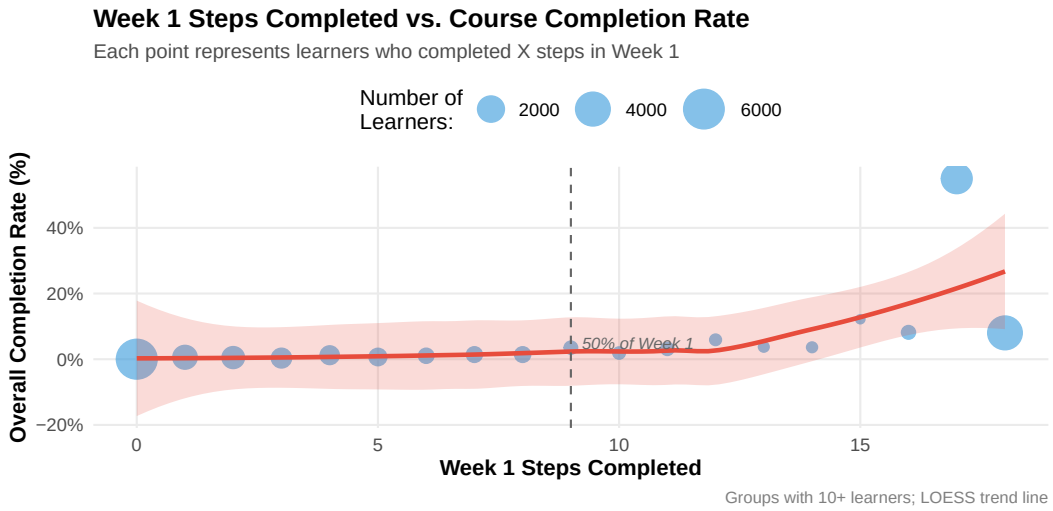

Figure 7: Smooth relationship between Week 1 steps completed and completion rate

Figure 7 shows the relationship between Week 1 steps and completion is approximately linear up to 50% of Week 1 content, then shows diminishing returns. Learners completing roughly half of Week 1 already show substantially elevated completion rates.

4.3.3 Threshold Analysis

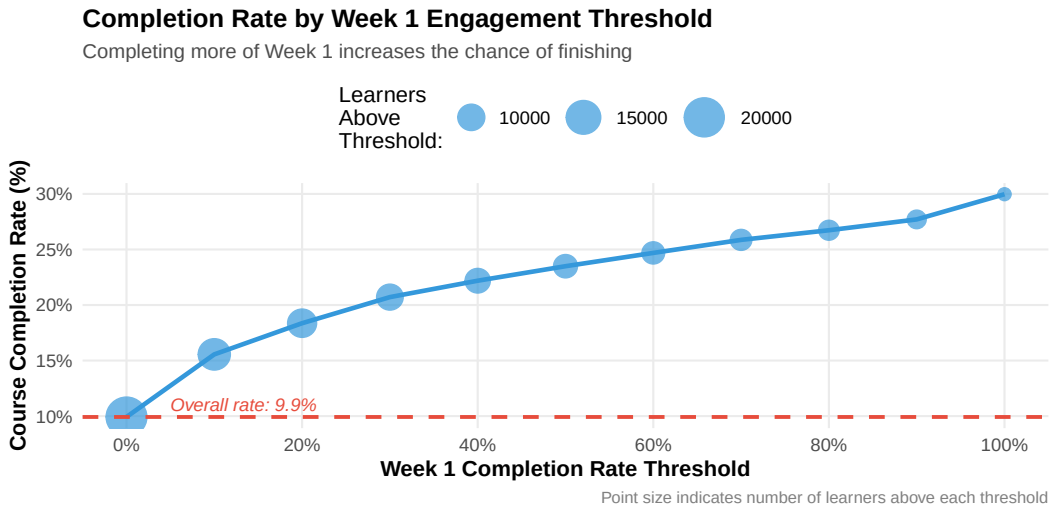

Figure 8: Completion probability increases steadily with higher Week 1 thresholds

Figure 8 demonstrates that as we set higher Week 1 completion thresholds, completion rate among learners meeting that threshold increases predictably. This supports using Week 1

engagement as an early warning indicator.

Practical threshold recommendation: Learners completing <40% of Week 1 (the "Low" and "Very Low" groups) have completion rates well below average and should be considered at-risk.

4.4 Evaluation (Cycle 2)

4.4.1 Key Findings

1. **Strong predictive relationship:** Week 1 engagement strongly associated with completion, with high engagers showing 26.2% vs. 0.1% completion
2. **Actionable threshold:** <40% Week 1 completion is reliable dropout risk indicator
3. **Early signal:** Relationship established within first days, enabling timely intervention
4. **Consistency:** Pattern holds across all seven course runs

4.4.2 Limitations and Caveats

- **Correlation vs. causation:** Cannot determine whether low Week 1 engagement causes dropout or underlying factors cause both.
- **Data quality:** Missing demographics limit subgroup analysis.
- **Intervention effectiveness:** Analysis shows Week 1 predicts completion but doesn't prove interventions improve outcomes—RCT needed. '
- **Generalizability:** Findings may be specific to FutureLearn platform, Cyber Security topic, or this course structure.
- **Selection bias:** Week 1 engagement may reflect pre-existing commitment differences.

5 Overall Conclusions

This two-cycle CRISP-DM analysis demonstrates that early engagement strongly predicts MOOC completion. Learners completing 75% of Week 1 achieve 26.2% completion vs. 0.1% for those completing <10%—a 198.7× difference.

Practical implications: MOOC providers can identify at-risk learners by day 7 based on Week 1 engagement (<40% completion threshold), enabling targeted intervention while learners remain engaged.

Methodological reflection: The iterative CRISP-DM approach proved effective, with exploratory analysis informing focused investigation. ProjectTemplate, dplyr, and ggplot2 facilitated reproducible, efficient analysis of 423,072 interaction records.

This study demonstrates that simple behavioral metrics available within the first week provide actionable insight into learner outcomes at scale, offering MOOC providers an evidence-based foundation for early intervention strategies.

6 Appendix: Technical Details

- **Environment:** R R version 4.5.2 (2025-10-31); ProjectTemplate 0.11.1
- **Key tools:** ProjectTemplate (structure), dplyr (manipulation), ggplot2 (visualization), kableExtra (tables), lubridate (dates)
- **Data processed:** 7 course runs; step-activity.csv and enrolments.csv files; key derived variables: `completed`, `week1_engagement_level`, `total_steps_completed`, `week1_completion_rate`
- **Reproducibility:** All analysis reproducible from submitted directory. Run `library(ProjectTemplate); load.project()` then knit report. Complete Git log in `GitLog.txt`.