MAS8600 Cyber Security MOOC Engagement (Runs 1–7) — CRISP-DM Report (Two Cycles)

Rishabh

2026-01-15

```
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)
knitr::opts_knit$set(root.dir = normalizePath(".."))
```

```
library(ProjectTemplate)
load.project()
```

Introduction

This report describes the application of the Cross-Industry Standard Process for Data Mining (CRISP-DM) to learner engagement data from a Cyber Security Massive Open Online Course (MOOC) delivered across seven course runs (Runs 1–7).

MOOCs provide flexible and scalable access to education, but they are often characterised by high levels of learner attrition. Understanding how learners interact with course content and identifying where disengagement occurs are therefore important for improving course design and learning outcomes.

Two cycles of the CRISP-DM process were carried out. The first cycle focuses on identifying where learners disengage during the course. The second cycle examines whether engagement with quizzes is associated with learner persistence.

The stakeholders for this analysis include course instructors, curriculum designers, and online education providers seeking to improve learner retention.

The specific research questions addressed in this report are:

Cycle 1: At which points in the course do learners disengage, and do disengagement patterns differ across runs?

Cycle 2: Is engagement with quizzes associated with increased learner persistence?

The CRISP-DM process is considered successful if meaningful disengagement points are identified (Cycle 1) and if quiz engagement is shown to be associated with persistence (Cycle 2), acknowledging that causal claims are not possible using observational data alone.

CRISP-DM Cycle 1 Business Understanding

The Cyber Security MOOC is intended to introduce learners to fundamental cyber security concepts. From the perspective of course providers, the objective is to maximise learner engagement while ensuring effective knowledge transfer.

High dropout rates reduce the overall impact of the course. Identifying where learners disengage enables stakeholders to prioritise improvements to course content and structure.

Data Understanding

Two primary datasets are used:

Step activity data: learner visits to course steps and completion status

Question response data: learner quiz submissions and correctness

Across runs, the raw datasets contain missing timestamps and variation in engagement levels. These issues were handled during preprocessing.

Data Preparation

All data preparation was performed in the munge/ directory using ProjectTemplate. Preprocessing included standardising identifiers, parsing timestamps, generating completion indicators, constructing per-run funnel tables, and caching analysis-ready datasets.

No manual file reads occur within this report.

Summary Statistics (Cycle 1)

```
funnel_step_all <- dplyr::bind_rows(lapply(1:7, function(r) {
get(paste0("funnel_step_run", r)) |>
dplyr::mutate(run = r)
}))

funnel_run <- funnel_step_all |>
dplyr::group_by(run) |>
dplyr::summarise(
total_step_visits = sum(visited, na.rm = TRUE),
mean_completion_rate = mean(completion_rate, na.rm = TRUE),
biggest_drop = ifelse(all(is.na(drop_off_to_next)),
NA_real_,
max(drop_off_to_next, na.rm = TRUE)),
.groups = "drop"
)

knitr::kable(
funnel_run,
caption = "Run-level summary of engagement metrics (Cycle 1)."
)
```

Table 1: Run-level summary of engagement metrics (Cycle 1).

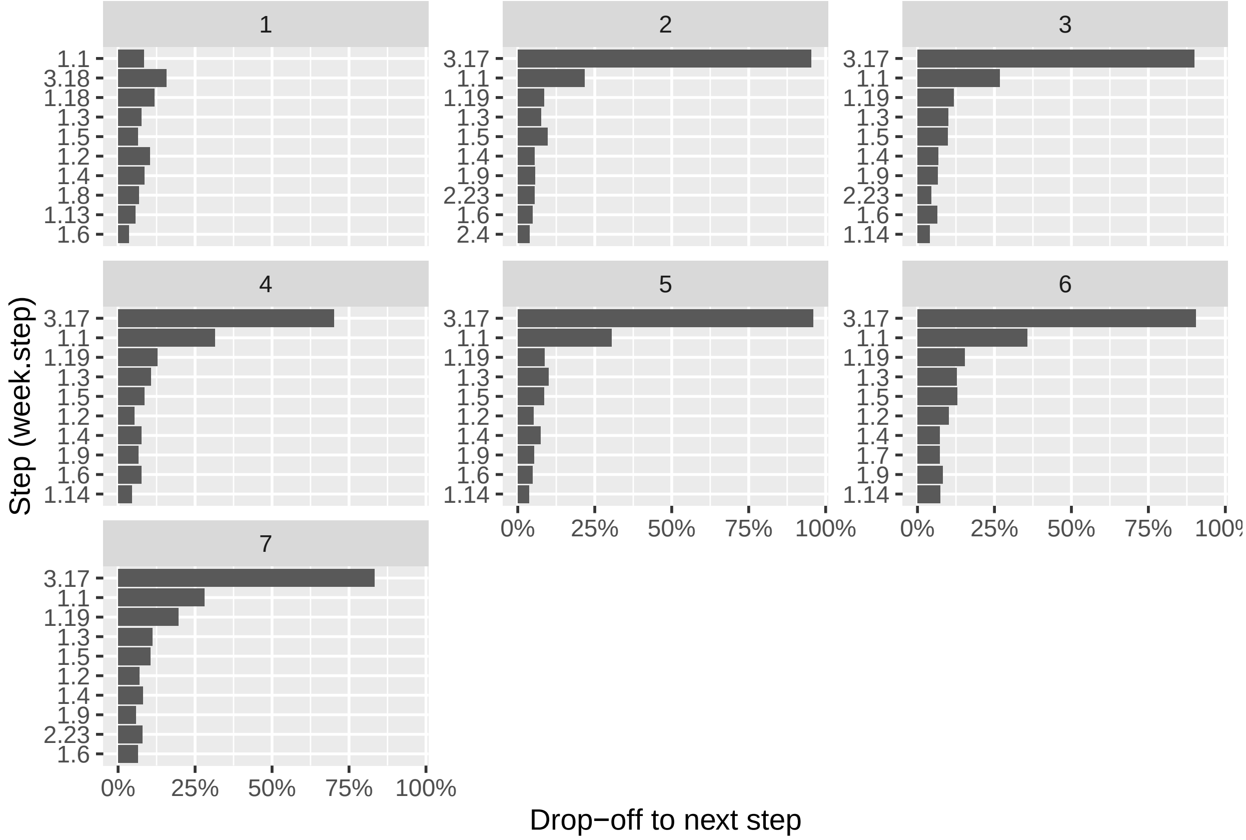
run	total_step_visits	mean_completion_rate	biggest_drop
1	143092	0.9385613	0.1575630
2	64809	0.9338104	0.9534207
3	46614	0.9285787	0.8987603
4	54524	0.9300287	0.7003367
5	54257	0.9137169	0.9601911
6	31472	0.9002498	0.9041096
7	28304	0.9101303	0.8321918

The largest observed step-to-step drop-off across all runs is r scales::percent(max(funnel_run$biggest_drop, na.rm = TRUE)).

```
top_dropoffs <- funnel_step_all |>
dplyr::filter(!is.na(drop_off_to_next)) |>
dplyr::group_by(run) |>
```

```r
dplyr::slice_max(drop_off_to_next, n = 10, with_ties = FALSE) |>
dplyr::ungroup() |>
dplyr::mutate(step_label = paste0(week_number, ".", step_number))

ggplot2::ggplot(
top_dropoffs,
ggplot2::aes(x = reorder(step_label, drop_off_to_next),
y = drop_off_to_next)
) +
ggplot2::geom_col() +
ggplot2::facet_wrap(~ run, scales = "free_y") +
ggplot2::scale_y_continuous(labels = scales::percent) +
ggplot2::coord_flip() +
ggplot2::labs(x = "Step (week.step)", y = "Drop-off to next step")
```


Step (week.step)
Drop−off to next step