FutureLearn MOOC Learning Analytics: Engagement and Completion (Runs 6-7)

Tejjus Suraj Bhat

16 January 2026

Contents

1 Introduction

MOOCs or Massive Open Online Courses are one of the latest prominent mechanisms for widening access to education, they are large-scale and flexible learning opportunities to diverse audiences across the world. In spite of their potential, MOOCs are consistently comprised by a low rate of completion of the full course, majority of the enrolled learners do not fully engage with the course material and disengage early in the course lifecycle. Understanding patterns of learner engagement and identifying the indicators of course completion is the question this report aims to answer.

This report presents an exploratory and predictive analysis of learner behaviour within a FutureLearn MOOC by Newcastle University on the topic "Cyber Security: Safety at Home, Online and in Life". The focus is mainly on the relationship between early engagement and eventual course completion. Using detailed enrolment records and step-level activity logs, the analysis seeks to quantify engagement patterns, assess data quality issues present in the MOOC data and evaluate whether early learner activity provides meaningful predictive signal for completion outcomes.

The analytical workflow is structured according to the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology. CRISP-DM provides a systematic and iterative framework for data-driven projects. It emphasizes clear transitions between business understanding, data understanding, data preparation, modeling/ analysis and evaluation. This structure is perfect for the analysis of educational data where the data complexity, imbalance and evolving analytical goals lead to repeated refinement across stages.

The extreme class imbalance present in MOOC datasets with completers representing only a fraction of the enrolled learners introduces a lot of statistical and interpretive challenges especially when applying standard modeling techniques. This analysis treats this imbalance as a substantial finding as it reflects the structural features of open online learning environments.

The specific research question that will be addressed is:

How does engagement affect the retention of learners in the MOOC?

The primary goal of this report is to develop a transparent and well-documented analytical narrative that illustrates how learner engagement data can be explored, transformed and modeled within the CRISP-DM framework. There is little focus on making a predictive model rather the focus is on interpretability, reproducibility and critical analysis of the data.

2 Methodology

This investigation is based on the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework for the analytical workflow. CRISP-DM is a framework which provides a systematic and iterative method to data analysis. It comprises of the following phases:

1. **Business Understanding**
    - Define the stakeholder, objectives, and success criteria.
2. **Data Understanding**

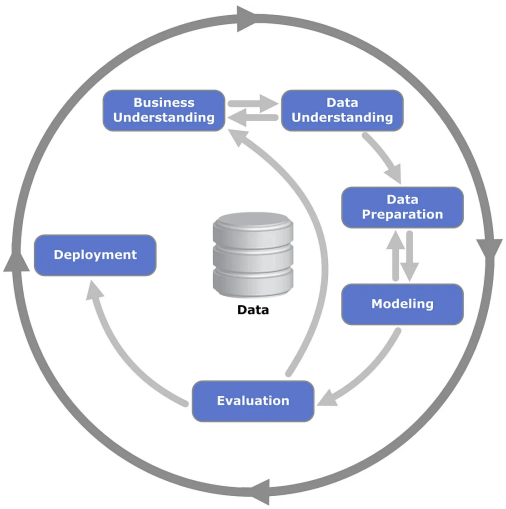

Figure 1: CRISP-DM Cycle

- Describe the available datasets, their structure, and data quality issues.

3. **Data Preparation**
   - Clean, standardise, and transform the data into analysis-ready form.
4. **Analysis / Modeling**
   - Perform exploratory analysis and (where appropriate) fit simple models to investigate the research question.
5. **Evaluation**
   - Summarise findings, relate them back to success criteria, and identify limitations and next steps.

It is cyclical in nature, meaning it allows insights from one phase to inform refinements in subsequent phases, this makes it suitable for exploratory analysis where the main focus is around understanding of the data and using it to gain value. In this report, CRISP-DM is applied across two consecutive cycles to refine the investigation of learner engagement and course completion.

2.1 Data Source

The data used in the analysis was collected from FutureLearn platform for the MOOC Cyber Security: Safety at Home, Online and in Life by Newcastle University. The dataset comprises multiple runs of the course, this analysis is restricted to runs 6 and 7. Each run has the same course content but a different selection of learners in their own cohort.

Two primary datasets are used. The enrolments dataset contains learner-level information like the unique identifier, enrolment timestamps, demographic attributes and a platform-defined indicator of course completion. Completion is recorded by the platform based on full participation criteria providing a consistent definition of the outcome variable that was used throughout the analysis.

The second dataset captures step level activity and records how the learner interacts with different sections or steps of the module. This dataset includes information on if or when the learners visit a specific step along with contextual variables such as week number and step number. This data is aggregated at the learner-step level leading to a sparse representation of the learner behaviour.

2.2 Data Preparation

Data preparation was implemented using a reproducible pipeline for consistency. Raw data files were ingested without manual modification, all the preprocessing steps are applied programmatically. Variable names are snake cased explaining exactly what the variable contains, the data types were standardised and the timestamp field was converted into consistent formats for ease of aggregation.

There was a presence of learner identifiers recurring across the two runs and the step structure. A run identifier was therefore retained throughout the analysis to prevent unintended aggregation of data in distinct course offerings.

Several data quality issues persisted in the data which were identified during preprocessing. There were learners with no step activity records which meant a substantial portion of learners recorded zero engagement. There were some timestamp fields with missing or malformed values which were handled during preprocessing. These values were not removed as they were important for the analysis but rather they were fixed programmatically.

Engagement features were derived by aggregating step-level activity data to the learner-run level. These features include counts of steps visited and completed, these are indicators of whether learners engaged with any course content. Some features were defined consistently in the analysis for comparison between the engagement behaviour. All preparation steps were aimed to answer the business question of the CRISP-DM cycle.

3 CRISP-DM Cycle 1

3.1 Business Understanding

In the first CRISP-DM cycle, the objective is to develop a basic understanding of the learner engagement behaviour and how it relates to retention and course completion within the MOOC. The stakeholder assumed in the investigation is that of the course provider and the learning analytics team, the party who would benefit from the insights gained from the description of engagement behaviour and its relationship to course outcomes. The primary goal of the first cycle is therefore descriptive, to establish a baseline for the engagement patterns, quantify non-engagement and assess whether there is any correlation between engagement behaviours from step activity to learners who complete the course.

The specific research question addressed in the first cycle is therefore:

"How does overall learner engagement relate to course completion and disengagement in the MOOC?"

Success criteria for Cycle 1 is as follows:

- Engagement measures are constructed in a transparent and reproducible way from step-level activity data.
- The analysis of learners and the patterns in their engagement with respect to their background.
- The distribution of learners who have zero recorded activity.
- Providing basis and results to define a refined question for Cycle 2.

3.2 Data Understanding

The first CRISP-DM cycle starts with the examination of the structure, scale and quality of the data used in the analysis. Two datasets are considered for this, the enrolments dataset and the step activity dataset, linked through learner and run identifiers. For Cycle 1, engagement will be defined over the full duration of the course using the step-level activity for the runs.

```
enrolments_clean |>
  dplyr::mutate(completed = ifelse(completed, "Completed", "Not completed")) |>
  dplyr::count(completed) |>
  ggplot(aes(x = completed, y = n)) +
  geom_col() +
  labs(
    x = NULL,
    y = "Number of learners"
  ) +
  theme_minimal()
```

The enrolments data provides a view of the population from the learner-level such as enrolment timestamps, learner demographic and platform-defined indicator of course completion. The first look of the data as shown in Figure 2, shows that the number of enrolled learners is large relative to the number of completers meaning there is a large imbalance in the outcome variable. Only a small fraction of learners actually completed the course which is consistent with MOOC participants.

The step activity data records learners' interaction with the course content at the step level of the course. THis dataset is larger than the enrolments data as it contains multiple entries of learners interacting with the course content. Exploratory analysis reveals that engagement measures show that the data is highly skewed, with the majority of the learners only interacting with a few modules of the course.

```
cycle1_analysis_table |>
  dplyr::mutate(activity = ifelse(any_step_activity, "Any activity", "Zero activity")) |>
  dplyr::count(activity) |>
  ggplot(aes(x = activity, y = n)) +
  geom_col() +
  labs(
    x = NULL,
    y = "Number of learners"
```

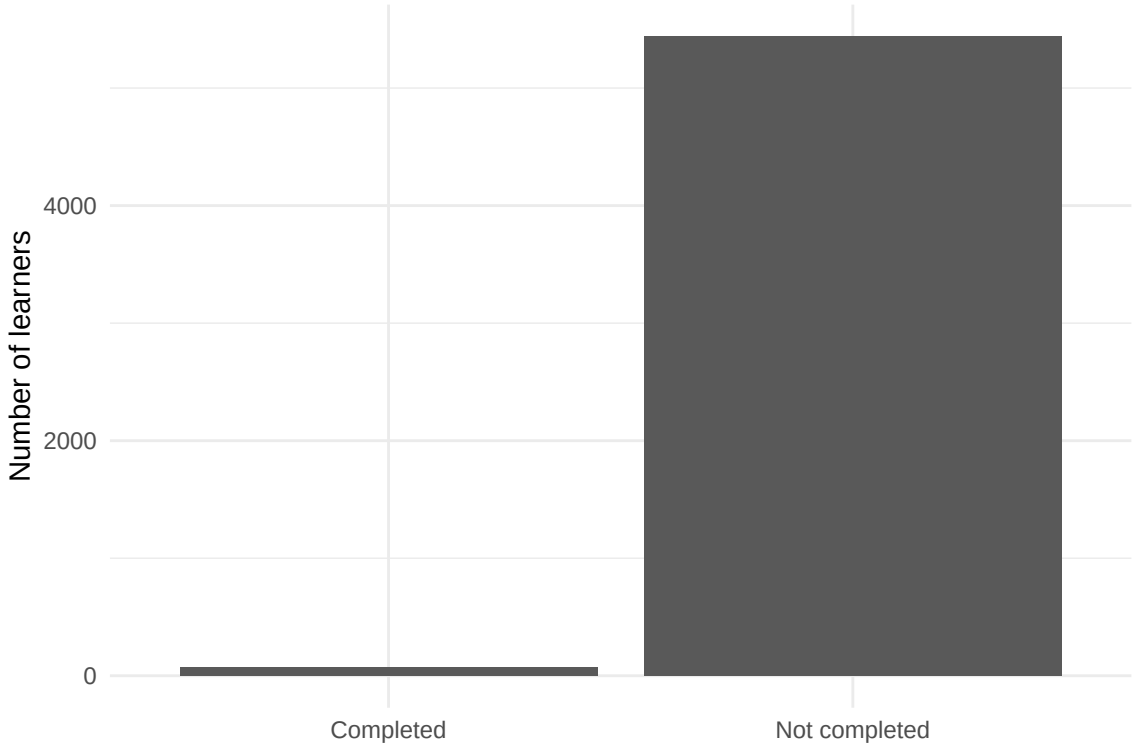

Figure 2: Completion imbalance (runs 6-7)

```
  ) +
  theme_minimal()
```

A data quality issue that was identified was the presence of learners shown in figure 3, with no recorded step activity despite being enrolled in the course, meaning they enrolled but never engaged with any of the content. These zero-activity learners represent a non-negligible portion of the population. These records were treated as valid observations and not missing values as they provide key insight into the behaviour of learners engaging with the course content or the lack thereof. Retaining these learners is important for accurately measuring the retention and engagement patterns.

Additional data quality issue was the presence of missing or unsupported timestamp values in some activity records. This was handled using a preprocessing function to convert the malformed timestamp into R readable timestamp. There was presence repeated learners across runs 6 and 7, meaning people retook the course or the same identifier was used to define learners across different runs, in either case when joining the two datasets a run identifier was introduced to act as a join primary key for the resultant dataset

```
enrolments_clean |>
  dplyr::mutate(employment_status = dplyr::if_else(
    is.na(employment_status) | employment_status == "",
    "Unknown",
    employment_status
  )) |>
```

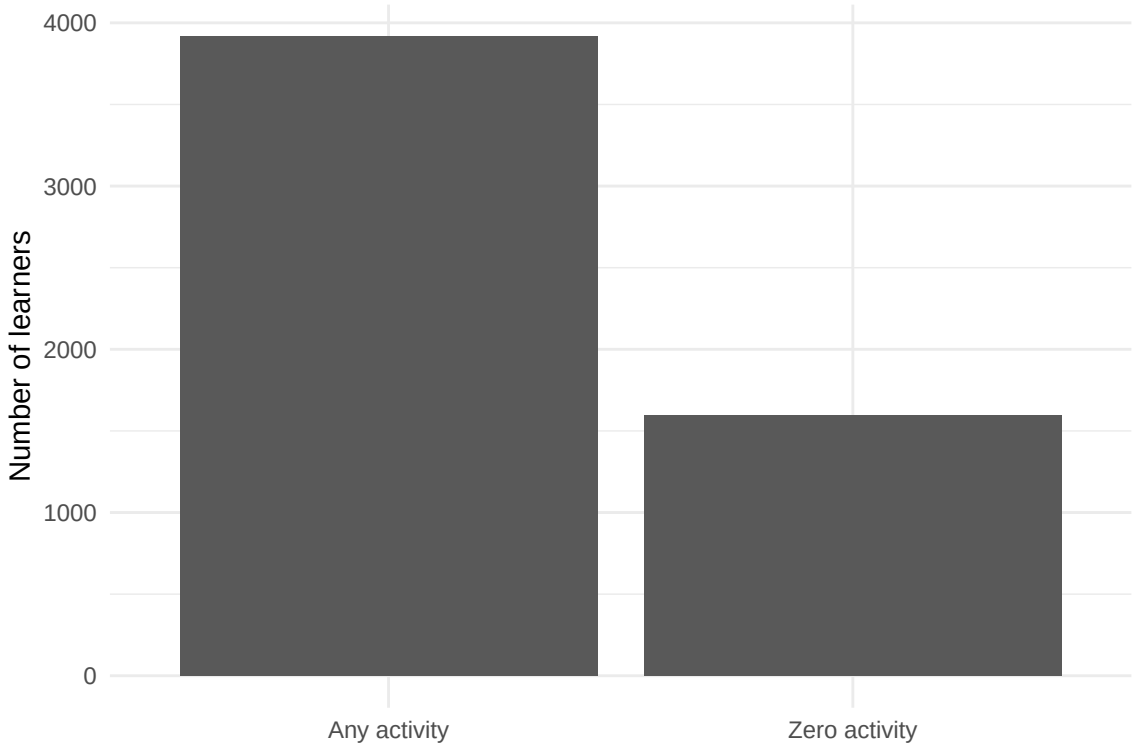

Figure 3: Zero-activity vs any-activity learners (runs 6-7)

```
dplyr::group_by(employment_status) |>
dplyr::summarise(
  n = dplyr::n(),
  completion_rate = mean(completed, na.rm = TRUE),
  .groups = "drop"
) |>
dplyr::filter(n >= 30) |>
ggplot(aes(x = reorder(employment_status, completion_rate), y = completion_rate)) +
geom_col() +
coord_flip() +
scale_y_continuous(labels = scales::percent_format(accuracy = 0.1)) +
labs(
  x = NULL,
  y = "Completion rate"
) +
theme_minimal()
```

The distribution of learners by employment status was also examined to better understand the demographic composition of the course population. Figure 4 shows the relationship between employment status and the engagement of the groups. While the majority of the learners fall into working categories the engagement overall remains low.

Overall, the data used in Cycle 1 is characterised by large scale, sparsity and extreme class im-

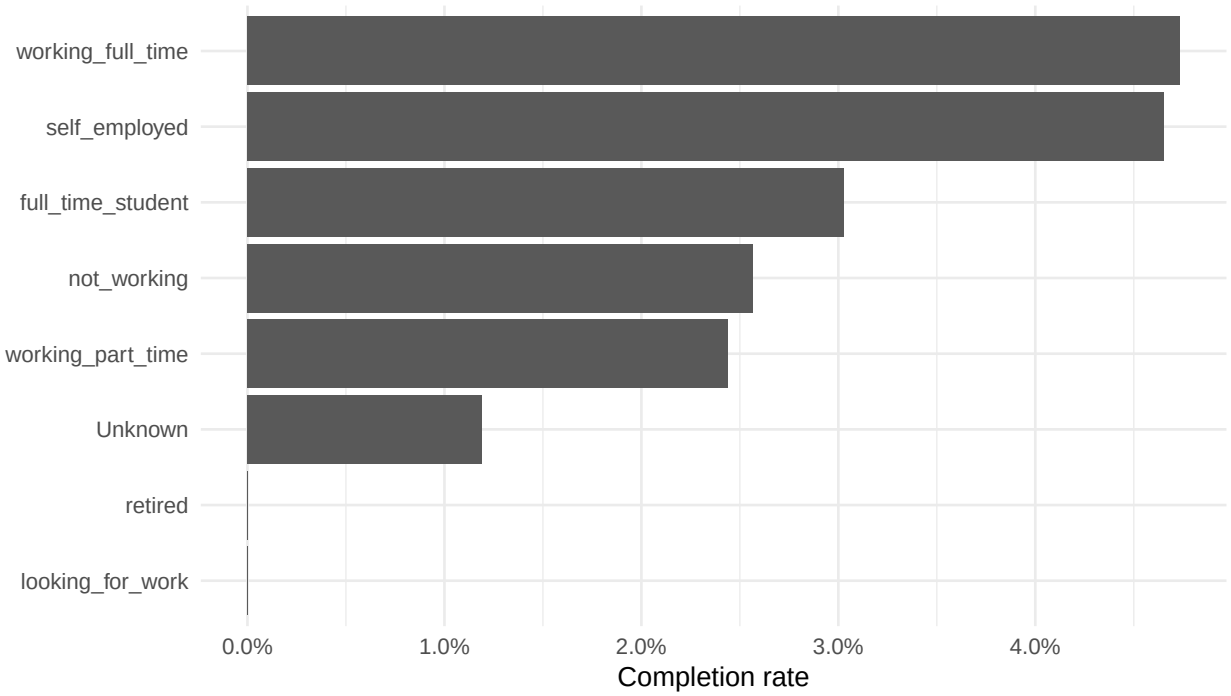

Figure 4: Completion rate by employment status

balance. These properties result in the analytical choices made in subsequent phases, favouring descriptive analysis over predictive modelling at this stage.

3.3 Data Preparation

For cycle 1, the step-level data was aggregated to the learner-run level to create the overall engagement features such as steps visited and completed as well as indicators capturing whether a learner engaged with any course content. Learners with no recorder step activity were kept as zero engagement learners for non-participants to reflect in the data. All preparation steps were implemented in the munge folder programmatically for convenience and consistency in analysis.

3.4 Analysis

```
cycle1_analysis_table |>
  dplyr::mutate(
    steps_band = dplyr::case_when(
      steps_visited == 0 ~ "0",
      steps_visited %in% 1:2 ~ "1-2",
      steps_visited %in% 3:5 ~ "3-5",
      steps_visited %in% 6:10 ~ "6-10",
      steps_visited > 10 ~ "11+"
    )
```

```
) |>
dplyr::group_by(steps_band) |>
dplyr::summarise(
  n = dplyr::n(),
  completion_rate = mean(completed, na.rm = TRUE),
  .groups = "drop"
) |>
ggplot(aes(x = steps_band, y = completion_rate)) +
geom_col() +
scale_y_continuous(labels = scales::percent_format(accuracy = 0.1)) +
labs(
  x = "Steps visited band",
  y = "Completion rate"
) +
theme_minimal()
```

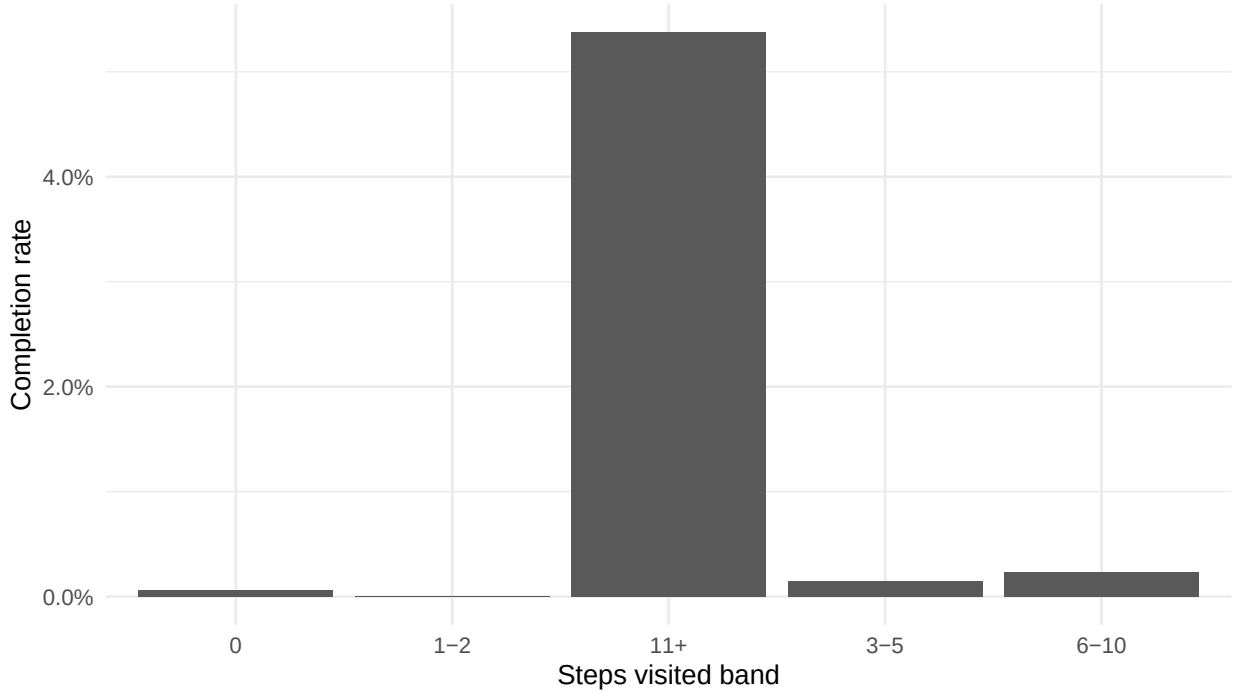

Figure 5: Completion rate by overall engagement band (runs 6-7)

Figure 5 shows that learners who completed the course generally visited more steps than non-completers. While distributions overlap, higher engagement level can be directly associated with the likelihood of completion. Further analysis shows that completion rate increases with higher engagement even though overall engagement is low. This suggests that there is some relation between early engagement and completion.

3.5 Evaluation

Cycle 1 provides a basic understanding of engagement and completion in runs 6 and 7 of the Futurelearn MOOC. The data shows extreme class imbalance with completion being rare and engagement measures having skewness with a substantial proportion of learners recording zero activity. Even with the limitations the aggregated engagement indicators from step activity show the difference of completers and non-completers, with learners completing the course showing high levels of engagement.

The Cycle 1 success criteria was met in that engagement measures were constructed in a transparent and reproducible way from step-level activity data. The analysis of learners and the patterns in their engagement with respect to their background and a refined question is now formed for cycle 2 which helps provide a more actionable signal for identifying learners at risk of disengagement.

4 CRISP-DM Cycle 2

4.1 Business Understanding

Cycle 1 established that overall engagement is associated with completion but those measures are retrospective to a large extent and therefore provide limited capability to predict user retention. Cycle 2 refines the investigation to focus on whether early engagement provides a more actionable signal related to completion outcomes. The stakeholder motivation at this stage is to find patterns that could support early intervention for learner retention.

The specific research question addressed in this cycle is:

"Does early engagement in the first weeks of the course differentiate learners who complete the MOOC from those who do not?"

Success criteria for Cycle 2 is as follows:

- Early-engagement features are defined from step activity data.
- The relationship between early engagement and completion is visualised.
- A model is fitted to assess the association between early engagement features and completion.

4.2 Data Understanding

Cycle 2 shifts focus onto early engagement by building on Cycle 1 and its focus on overall engagement. At this stage, the key data understanding objective is to construct early engagement features consistently for comparability to results from Cycle 1. Early engagement is defined using step activity within the initial weeks of the course which is aggregated at the learner level.

Exploration of this derived dataset shows that a substantial percentage of learners show some kind of early activity, while non-negligible percentage of people show no early engagement. The cross tabulation of completion by early activity shows that early activity is a potential differentiator but since the data is highly imbalanced it cannot be said for certain so it is treated as a potential rather than a cause-effect.

```
## # A tibble: 1 x 8
##   n_rows completion_rate share_any_early_activity share_zero_early_activity
##    <int>           <dbl>                    <dbl>                      <dbl>
## 1   5512          0.0134                    0.709                      0.291
## # i 4 more variables: median_steps_visited <dbl>, median_steps_completed <dbl>,
## #   median_weeks_active <dbl>, median_active_days <dbl>
```

```
## # A tibble: 2 x 2
##   any_early_activity     n
##   <lgl>              <int>
## 1 FALSE               1604
## 2 TRUE                3908
```

```
## # A tibble: 4 x 3
##   completed any_early_activity     n
##   <lgl>     <lgl>              <int>
## 1 FALSE     FALSE               1603
## 2 FALSE     TRUE                3835
## 3 TRUE      FALSE                  1
## 4 TRUE      TRUE                  73
```

```
analysis_table |>
  dplyr::group_by(any_early_activity) |>
  dplyr::summarise(
    n = dplyr::n(),
    completion_rate = mean(completed, na.rm = TRUE),
    .groups = "drop"
  ) |>
  dplyr::mutate(any_early_activity = ifelse(any_early_activity, "Any early activity", "No early
  ggplot(aes(x = any_early_activity, y = completion_rate)) +
  geom_col() +
  scale_y_continuous(labels = scales::percent_format(accuracy = 0.1)) +
  labs(
    x = NULL,
    y = "Completion rate"
  ) +
  theme_minimal()
```

Figure 6 summarises completion rates for learners with and without early engagement. The completion rate is low overall but it is higher among learners who record early activity. This supports the hypothesis of early engagement being a potentially actionable signal.

4.3 Data Preparation

For this cycle data preparation process refines the originally constructed features in the first cycle restricting attention to initial weeks of the course. The step activity was limited to the initial weeks

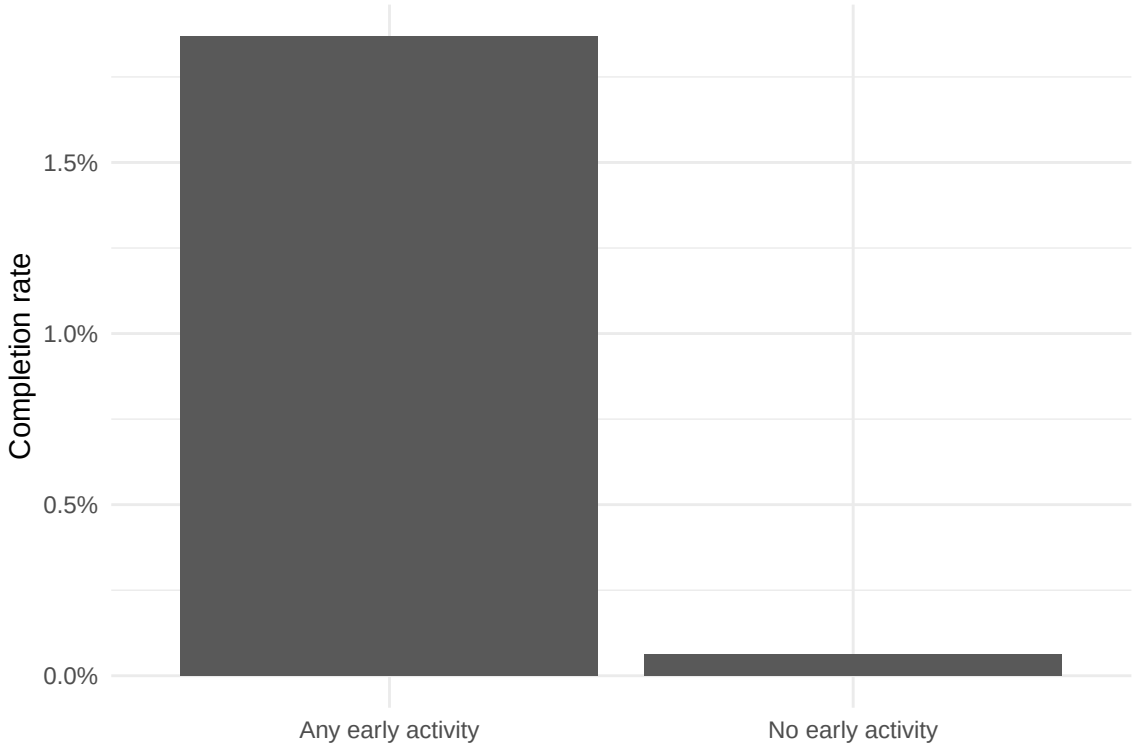

Figure 6: Completion rate by early activity (runs 6–7)

of the course and aggregated at a learner level. From this aggregation, a binary indicator of early engagement was created to capture if there was early engagement.

Learners with no early activity were retained as zero early engagement. This preserves the full population and shows the non-participation behaviour of the learners. All transformations were implemented programmatically through the munge process to maintain the reproducibility.

4.4 Analysis and Modelling

This cycle analyses the findings by fitting an exploratory model to assess the association between early engagement and course completion. Given the binary outcome variable, it was decided to fit a logit function to the model and create a logistic regression model. The purpose of this model is to evaluate statistical significance of early engagement rather than to produce a predictive model.

Due to the imbalance in the outcome variable the model results should be interpreted with caution. The estimated probabilities should reflect on the limitations rather than to make strong predictions. Model diagnostics are included to assess whether the fitted model is consistent with the observed data.

```
##
## Call:
## glm(formula = completed ~ any_early_activity, family = binomial(link = "logit"),
##     data = analysis_table)
```

```
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -7.3796     0.9991  -7.386 1.51e-13 ***
## any_early_activityTRUE    3.4182     1.0060   3.398  0.00068 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 784.97  on 5511  degrees of freedom
## Residual deviance: 742.51  on 5510  degrees of freedom
## AIC: 746.51
##
## Number of Fisher Scoring iterations: 9
```

```
analysis_table |>
  dplyr::mutate(
    predicted = predict(glm_c2, type = "response"),
    residual = completed - predicted,
    bin = dplyr::ntile(predicted, 10)
  ) |>
  dplyr::group_by(bin) |>
  dplyr::summarise(
    mean_pred = mean(predicted),
    mean_resid = mean(residual),
    .groups = "drop"
  ) |>
  ggplot(aes(x = mean_pred, y = mean_resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(
    x = "Mean predicted probability",
    y = "Mean residual (observed - predicted)"
  ) +
  theme_minimal()
```

Figure 7 shows a binned residual diagnostic for the logistic regression model fitted in Cycle 2. Residuals are centred around zero across bins of predicted probability, indicating no strong systematic miscalibration. However since the residuals are so highly variable the results must be interpreted with a grain of salt as it shows the rarity of the event.

4.5 Evaluation

Cycle 2 demonstrates that early engagement has statistically significant association with course completion with almost all completers exhibiting some form of early activity. The logit GLM analysis reinforces this pattern for learners with early engagement. These results suggest that early

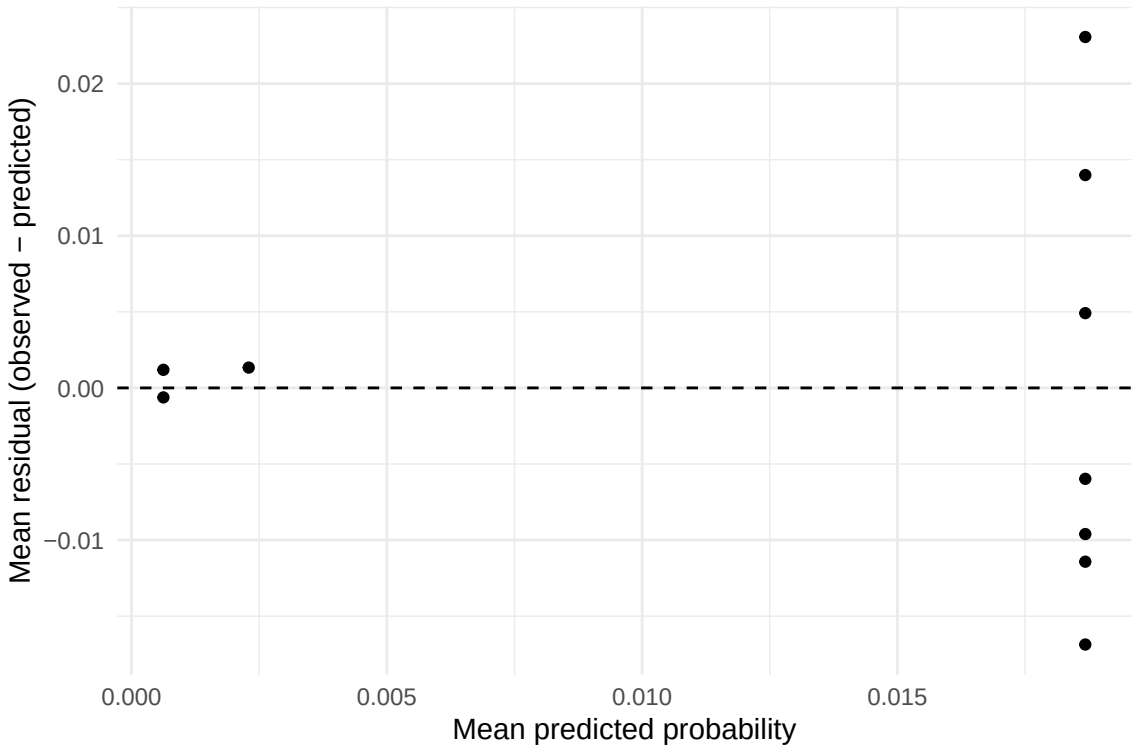

Figure 7: Binned residuals for early engagement logistic regression

engagement may serve as a useful indicator for identifying at-risk learners but should be interpreted cautiously with additional behavioural or contextual information.

5 Conclusion

This report applied a CRSIP-DM approach to investigate learner engagement and completion in a FutureLearn MOOC across two cycles . The first established a baseline showing engagement patterns and their association with completion. The second cycle built on this idea by understanding whether early engagement can be a determining factor to predict deserters early.

The analysis brings to light that the completion rate of the MOOC is low overall with severe class imbalance but early engagement holds statistical significance to show whether a learner will complete the MOOC or not. Future work could incorporate behavioural features, additional contextual data such as results of tests or surveys to better understand learner retention and provide more targeted support mechanisms.

6 References

[1] Z. Abedjan, Ł. Golab, and F. Naumann, "Profiling relational data: A survey," *The VLDB Journal*, vol. 24, no. 4, pp. 557–581, 2015, doi: 10.1007/s00778-015-0389-y.

[2] A. Khan and J. Bentham, "Chapter 11: Generalised linear models," in *Graduate Foundations of Statistics and Data Science*, 2025.

[3] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc., 2000.

[4] M. P. J. van der Loo and E. de Jonge, "Data validation infrastructure for R," *Journal of Statistical Software*, vol. 97, no. 10, 2021, doi: 10.18637/jss.v097.i10.