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

Tejjus Suraj Bhat

15 January 2026

Contents

1 Introduction **2**

2 Methodology **2**
2.1 Data Source . . . 3
2.2 Data Preparation . . . 3

3 CRISP-DM Cycle 1 **4**
3.1 Business Understanding . . . 4
3.2 Data Understanding . . . 4
3.3 Data Preparation . . . 4
3.4 Analysis . . . 4
3.5 Evaluation . . . 4

4 CRISP-DM Cycle 2 **4**
4.1 Business Understanding . . . 4
4.2 Data Understanding . . . 4
4.3 Data Preparation . . . 5
4.4 Analysis and Modelling . . . 5
4.5 Evaluation . . . 5

5 Conclusion **5**

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**
   - 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**

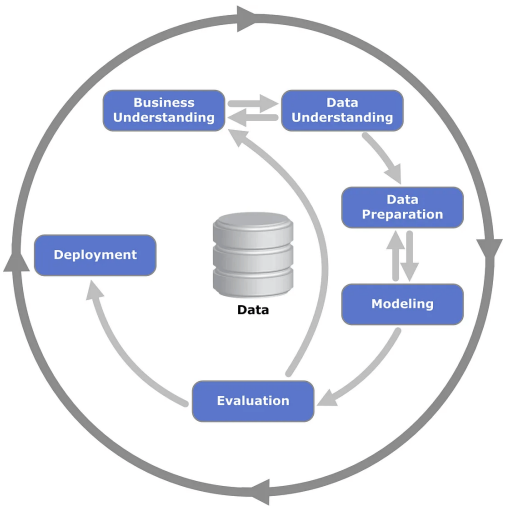

Figure 1: Figure 2.1: CRISP-DM Cycle

- 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

Describe the business understanding here.

3.2 Data Understanding

Describe the data understanding here.

3.3 Data Preparation

Describe the data preparation here.

3.4 Analysis

Describe the modeling here.

3.5 Evaluation

Describe the evaluation here.

4 CRISP-DM Cycle 2

4.1 Business Understanding

Describe the business understanding here.

4.2 Data Understanding

Describe the data understanding here.

4.3 Data Preparation

Describe the data preparation here.

4.4 Analysis and Modelling

Describe the modeling here.

4.5 Evaluation

Describe the evaluation here.

5 Conclusion