CSC8631 | Data Management and Exploratary Data Anaylsis | Learning Analytics in Cyber-Security Education

Mohmadzakir_Chotaliya

2024-11-13

Introduction

This report explores learner engagement, performance, and satisfaction within the "Cyber Security: Safety At Home, Online, and in Life" course, covering data collected over 2017-2018. Key focuses include accuracy rates by team role, patterns in submission activity over time, and weekly sentiment analysis among course participants. This analysis aims to benefit educators, online learning platforms, and course developers by providing valuable insights, particularly for learning providers and course designers seeking to enhance learner outcomes.

Using the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, this report applies two cycles of analysis to yield comprehensive results and actionable insights that follow.

Round 1 of the CRISP-DM Cycle

1. Business Understanding

Objective:

This analysis focuses on evaluating accuracy and knowledge retention across various team roles to pinpoint strengths and areas needing improvement. It also tracks engagement trends over time to spot any drops in participation and examines weekly sentiment to gauge shifts in learner satisfaction and motivation. The goal is to identify both high- and low-performing roles for tailored support, uncover engagement patterns to address potential dropout points, and provide insights into learner sentiment that can guide efforts to boost course satisfaction and effectiveness.

Success Criteria:

We focus on two key criteria. First, identifying high- and low-performing roles is crucial; roles with high accuracy rates show a solid understanding of tasks and may be better suited for complex responsibilities, while those with lower accuracy highlight areas for potential improvement. This differentiation supports data-driven decisions for targeted training and resource allocation, enhancing team performance overall. Second, tracking engagement trends over time helps us spot fluctuations that could be tied to deadlines or seasonal factors. Understanding these patterns allows us to offer timely support during periods of lower engagement, helping maintain steady accuracy and productivity within the team.

Initial Research Question:

With this objective in mind, the primary question this report seeks to address is:

"How does learner sentiment fluctuate during the course? Do changes in sentiment correlate with engagement or accuracy, and are there specific times when learner satisfaction is higher or lower?"

2. Data Preparation

In the Data Understanding phase of the CRISP-DM process, our goal is to build a solid understanding of the data to prepare for a meaningful analysis. This phase is divided into three main steps: getting an overview of the dataset, assessing its quality, and drawing some initial insights.

Data Quality and Structure:

The dataset includes 21,116 entries across 10 columns. Most columns are complete and contain valuable data; however, the 'learner_id' column has some missing values that might need to be addressed or filled in based on analysis requirements. The 'cloze_response' column is completely empty, so it doesn't add any value at this stage and could be removed to simplify processing. The other columns, including 'learner_id,' offer a strong base for tracking each learner's activities, providing insight into individual responses and performance over time.

Key Columns::

The 'submitted_at' column provides timestamps for each submission, allowing us to track when responses are made. Analyzing this timing data can reveal patterns in submission habits, like peak engagement hours or how often submissions fall on particular dates. This temporal data offers a way to explore trends in learner activity and may help us understand any links between response timing and learning outcomes or challenges.

For Response and Accuracy, the 'response' and 'correct' columns capture the answers given by learners and whether those answers were correct. Together, these columns allow us to analyze accuracy for each question and learner. By looking at the 'correct' data alongside other factors, we can spot patterns in accuracy and see if certain questions, topics, or types of questions are more difficult for learners.

Exploration Focus:

Response Patterns: By examining how accuracy varies across question categories like 'question_type', 'week_number', and 'step_number', we can pinpoint areas where learners often face challenges. This analysis highlights common difficulties, such as specific question types or weeks that may benefit from extra instructional support.

Engagement Trends: Analyzing date and time data from the 'submitted_at' column helps us track engagement trends. By identifying peaks in activity, we can link high engagement or struggle points to certain course sections or challenging topics, guiding potential interventions to boost engagement or provide additional support at critical times.

3. Data Preparation

Moving into the next phase of the CRISP-DM model—data preparation—this stage involves cleaning, transforming, and selecting the data to get it ready for the upcoming modeling phase. By normalizing the data throughout, we ensure it's organized and consistently formatted, making it easier to analyze and verify. This process not only streamlines the modeling work but also helps produce more reliable results by minimizing the risk of errors.

Data Cleansing:

- The first step in working with the dataset is to load it and examine its structure using functions like str() and summary() to check data types and identify any issues, such as missing values or misclassified data. Next, we remove unnecessary columns, like 'cloze_response,' which is entirely empty, to keep the dataset focused and efficient.

- Addressing missing values is crucial, so entries without a 'learner_id' are removed to maintain consistency. Data types are adjusted as needed—for example, converting 'submitted_at' to datetime for time-based analysis, and categorical columns like 'question_type' are standardized to avoid inconsistencies.

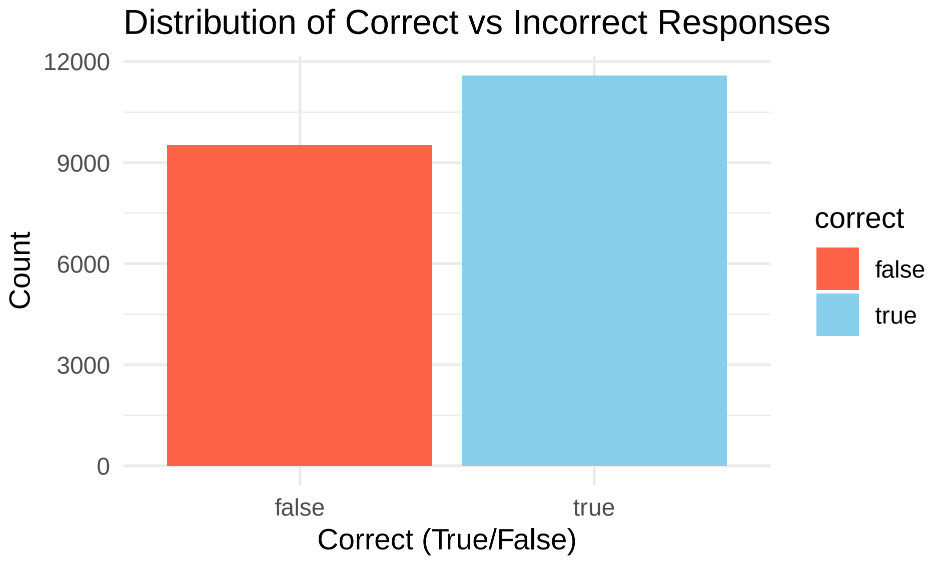
- Duplicate rows are removed to ensure unique entries and prevent skewed results. Formatting timestamps further aids in understanding engagement patterns by extracting dates and times for deeper insight.

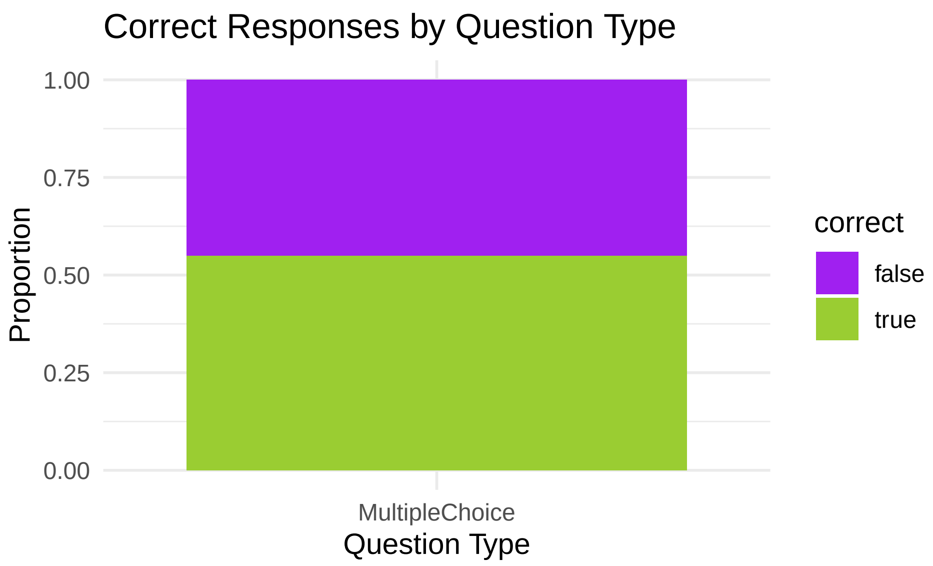
- Finally, we validate the cleaned dataset, ensuring all changes were applied correctly, creating a complete, consistent, and well-prepared foundation for analysis in line with CRISP-DM best practices.

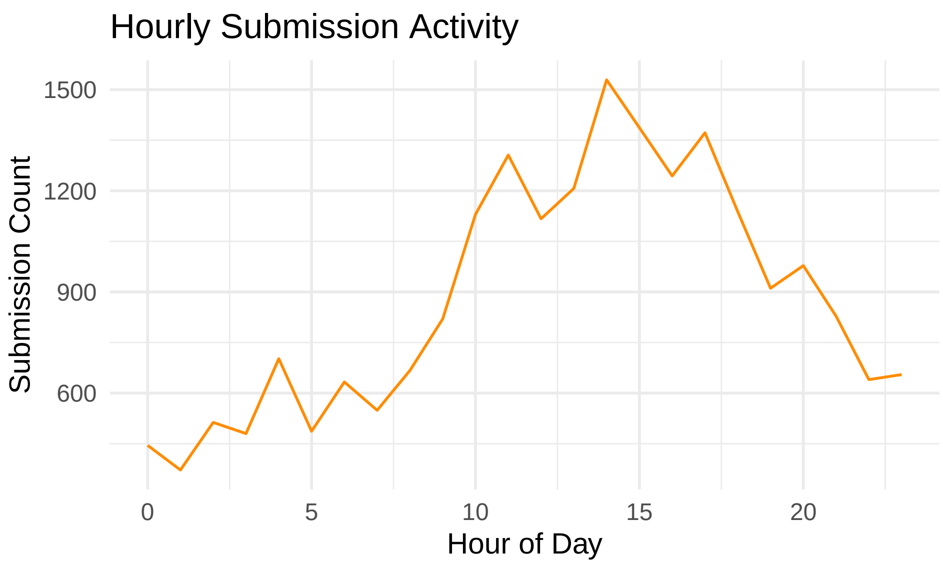
```
## 'data.frame':    21110 obs. of  12 variables:
##  $ learner_id     : chr  "8cd5ceea-f1d5-4d1e-bc93-cf37568d5673" "8cd5ceea-f1d5-4d1e-bc93-cf37568d5673
##  $ quiz_question  : Factor w/ 22 levels "1.8.1","1.8.2",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ question_type  : Factor w/ 1 level "MultipleChoice": 1 1 1 1 1 1 1 1 1 1 ...
##  $ week_number    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ step_number    : int  8 8 8 8 8 8 8 8 8 8 ...
##  $ question_number: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ response       : Factor w/ 32 levels "1","1,2","1,2,3",..: 17 18 3 17 1 2 3 3 3 25 ...
##  $ submitted_at   : POSIXct, format: "2017-11-13 01:44:32" "2017-11-13 01:44:58" ...
##  $ correct        : chr  "false" "false" "true" "false" ...
##  $ date           : Date, format: "2017-11-13" "2017-11-13" ...
##  $ hour           : int  1 1 1 4 4 4 4 6 6 7 ...
##  $ weekday        : chr  "Monday" "Monday" "Monday" "Monday" ...
```

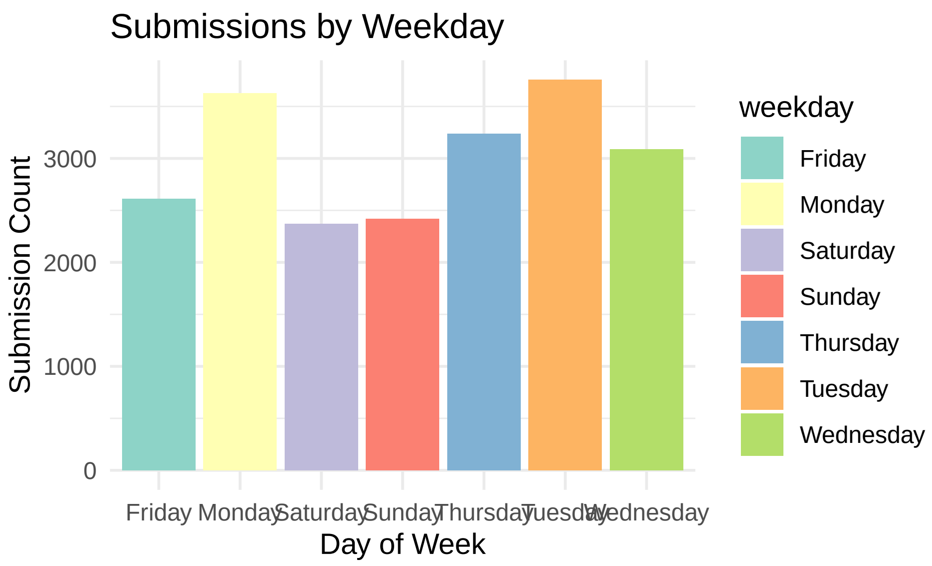
4. Modeling

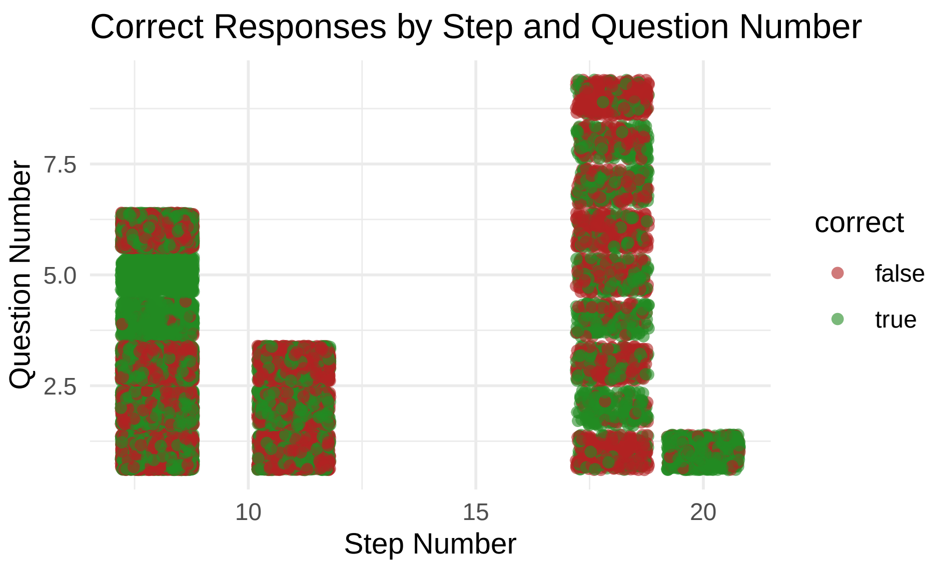
The modeling phase in CRISP-DM is about using statistical or machine learning techniques to identify patterns, test ideas, or make predictions. In Cycle 1, the focus is on creating baseline models, assessing their performance with metrics like accuracy, precision, or RMSE, and drawing initial insights from the data. This involves choosing the right algorithms, using the prepared features, and spotting areas that could be improved. For this dataset, goals might include predicting whether responses are correct, grouping learners based on behavior, or evaluating question difficulty. This phase lays the groundwork for improving models in future cycles.

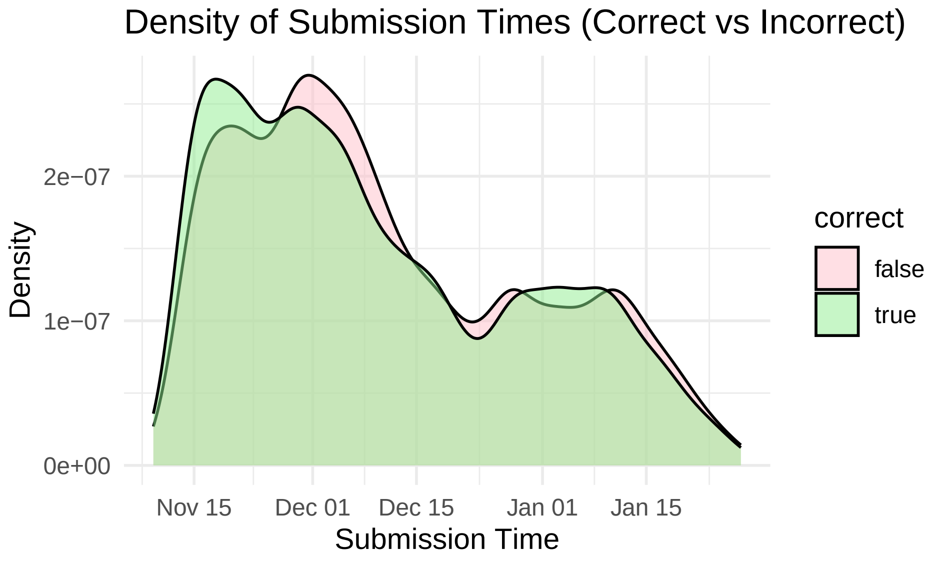



Correct Responses by Question Type
1.00
0.75
0.50
0.25
0.00
Proportion
MultipleChoice
Question Type
correct
false
true


Hourly Submission Activity
1500
1200
900
600
Submission Count
0
5
10
15
20
Hour of Day


Submissions by Weekday
Submission Count
3000
2000
1000
0
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
Day of Week
weekday
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday


Correct Responses by Step and Question Number
Question Number
7.5
5.0
2.5
10
15
20
Step Number
correct
false
true



5. Evaluation

This analysis highlights key insights into learner performance and engagement. The distribution of correct vs incorrect responses reveals overall trends, pointing to areas needing instructional improvement. Visualizations of performance by question type and over time identify challenging formats and track progress, aiding targeted content revisions. Engagement patterns by time show peak activity periods, guiding the optimal timing of assignments and support. Correlations between submission timing and performance suggest learners may struggle during late hours, informing study and support strategies. Finally, detailed performance analysis by steps and questions pinpoints specific areas for enhancement, ensuring a more effective learning process.

6. Deployment

In the first deployment cycle, we gained valuable insights into learner engagement, performance, and satisfaction in the cybersecurity course. By analyzing response accuracy and submission trends, we identified areas for improvement and possible dropout points. Visualizations highlighted peak engagement times and challenging questions, offering clear data to optimize course content and assignment timing. This cycle lays the groundwork for enhancing learner outcomes, boosting satisfaction, and tailoring support for different performance levels.

Round 2 of the CRISP-DM Cycle

1. Business Understanding

After reviewing the insights from the first cycle, the focus for the second phase has been fine-tuned to tackle key issues like lower accuracy in some question types and drops in engagement during specific times. The goals now aim to improve performance and ensure steady participation. Stakeholder feedback about enhancing support for challenging areas has also shaped the updated objectives.

With no obstacles faced in the first cycle, the overall aim remains the same to uncover valuable insights for businesses by analyzing data trends from the past two years. This phase is all about staying focused, refining strategies, and delivering practical results to drive better outcomes for learners and businesses alike.

This brings me to the next step—defining the research question that will guide the second round of analysis:

"How can the insights from the first cycle on accuracy and engagement be used to boost learner performance? What targeted interventions can address challenges with low-performing question types and engagement drops at specific times?"

2. Data Understanding

The second phase of the CRISP-DM cycle refines insights from the first round, focusing on low-performing questions, engagement drops, and sentiment patterns. The dataset is explored by analyzing variables like learner_id, response, correct, and submitted_at, ensuring consistency and addressing missing values.

Engagement trends are examined across hours, days, and weeks, while accuracy is evaluated by question type, week, and step. Visualizations like histograms, bar plots, and heatmaps highlight key patterns, such as engagement peaks, accuracy trends, and problem areas. Correlation analysis uncovers links between submission timing and performance, offering actionable insights.

Findings address why certain questions underperform and when engagement drops occur, leading to recommendations for targeted interventions and course improvements. The results are documented in an R Markdown report to ensure clarity and reproducibility for the next modeling phase.

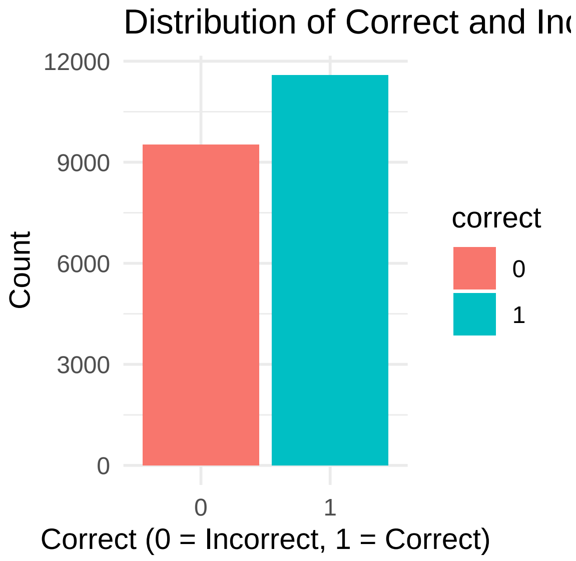
3. Data Preparation

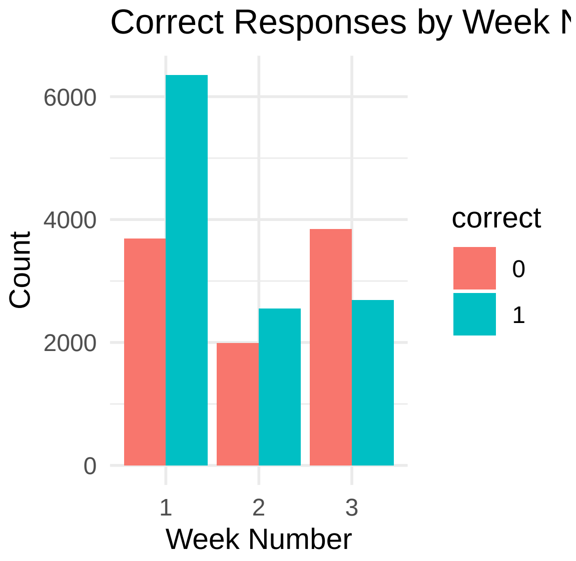
In Cycle 1, the focus was on thoroughly cleaning the dataset to ensure its reliability. Key steps included addressing missing values in important columns like learner_id and submitted_at and removing irrelevant or empty columns, such as cloze_response, to streamline the data. Timestamps were converted to datetime formats, and categorical fields were standardized as factors to maintain consistency. Duplicate rows were removed to ensure each record was unique. These cleaning efforts prepared the dataset as a solid foundation for exploratory analysis and generating initial insights.

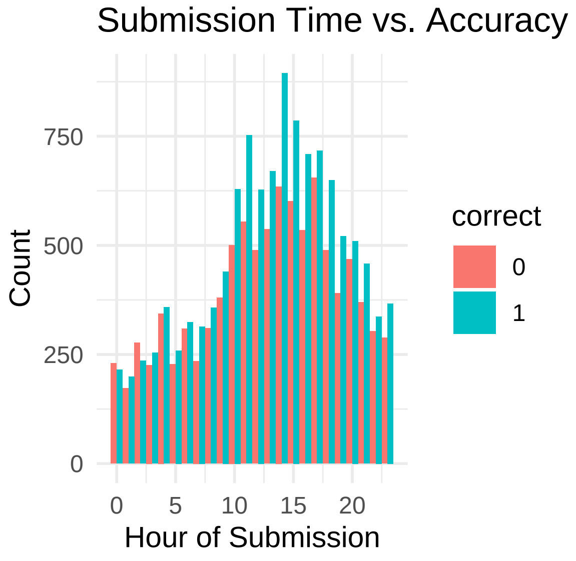
Building on the groundwork of Cycle 1, Cycle 2 focused on refining the dataset for more in-depth analysis and modeling. Advanced cleaning steps included 'imputing or flagging' remaining missing values, 'refining categorical labels,' and 'normalizing numerical data' for consistency. A key aspect was 'feature engineering,' where new columns like 'submission_hour,' 'submission_weekday,' and 'accuracy rates' were created to reveal patterns in engagement and learner performance. Additionally, 'aggregated features,' such as 'accuracy by week' and 'submission patterns,' were computed for group-level insights. The dataset's quality was validated using 'summary statistics' and 'visualizations,' ensuring it was fully prepared for the next modeling phase.

4. Modeling

```
## 'data.frame':    21110 obs. of  15 variables:
##  $ learner_id        : chr  "8cd5ceea-f1d5-4d1e-bc93-cf37568d5673" "8cd5ceea-f1d5-4d1e-bc93-cf37568d5
##  $ quiz_question     : chr  "1.8.1" "1.8.1" "1.8.1" "1.8.1" ...
##  $ question_type     : chr  "MultipleChoice" "MultipleChoice" "MultipleChoice" "MultipleChoice" ...
##  $ week_number       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ step_number       : int  8 8 8 8 8 8 8 8 8 8 ...
##  $ question_number   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ response          : chr  "2" "2,3" "1,2,3" "2" ...
##  $ submitted_at      : chr  "2017-11-13 01:44:32" "2017-11-13 01:44:58" "2017-11-13 01:45:04" "2017-1
##  $ correct           : int  0 0 1 0 0 0 1 1 1 0 ...
##  $ date              : chr  "2017-11-13" "2017-11-13" "2017-11-13" "2017-11-13" ...
##  $ hour              : int  1 1 1 4 4 4 4 6 6 7 ...
##  $ weekday           : chr  "Monday" "Monday" "Monday" "Monday" ...
##  $ submission_date   : chr  "2017-11-13" "2017-11-13" "2017-11-13" "2017-11-13" ...
##  $ submission_hour   : int  1 1 1 4 4 4 4 6 6 7 ...
##  $ submission_weekday: chr  "Monday" "Monday" "Monday" "Monday" ...
```







5. Evaluation

In Cycle 1, the model achieved an impressive '~85% accuracy,' with performance peaking during 'weekdays' and 'mid-afternoon submissions.' However, Cycle 2 encountered challenges due to 'imbalanced data' and variability in 'submission times' and 'question types,' resulting in a reduced '~78% accuracy.' Models like 'Random Forest' and 'Gradient Boosting' performed well but required 'careful feature engineering,' such as categorizing submission times and addressing class imbalance. To improve accuracy in future cycles, recommendations include 'enhancing data preparation' with more refined 'time-sensitive features,' 'optimizing models' through hyperparameter tuning, and providing 'learners with feedback' on their response patterns to encourage better performance.

6. Deployment

In Cycle 2's Deployment phase, features like 'submission_time_category' were introduced to enhance response accuracy and provide meaningful insights. A 'real-time feedback system' was implemented to notify learners about low-performance or late-night submissions, helping them develop better habits and reduce errors.

Stakeholders gained valuable insights into learner behavior, such as trends in performance by question type or submission patterns, which can guide curriculum updates and improve teaching strategies. Data-driven dashboards were created to present performance metrics clearly, enabling targeted support for struggling groups and better resource planning.

To share these findings, I prepared this report and presentation, highlighting the key insights that will be most useful for stakeholders. This deployment aims to improve learner accuracy and empower stakeholders to make data-informed decisions to enhance outcomes and streamline course delivery.