# 

# 

# Week 1: Data Cleaning and Feature Engineering Report

**Team 8:** The Elites **Excelerate Internship Program  
Intern:** Lily Charlotte

**Submission Date:** June 23, 2025

**Prepared by:**

| **Lily Charlotte** | Team Lead | lilycharlotte47@gmail.com |
| --- | --- | --- |
| **Vishal Reddy** | Project Lead | vishal0987devi@gmail.com |
| **Phanidhar Kasuba** | Project Manager | phanikasuba@gmail.com |
| **Kosisochukwu Onyekwulunne** | Project Scribe | onyekwulunnekosi9@gmail.com |
| **Amit Kumar**  **Divyesh Shinde** | Project Strategist  Data Cleaner | [amitsigh6394366374@gmail.com](mailto:amitsigh6394366374@gmail.com)  shindedivyeshavinash@gmail.com |

**Table of Contents**

[**Week 1: Data Cleaning and Feature Engineering Report 0**](#_555999mbgjk9)

[**Introduction 2**](#_v2zctp8opuao)

[**Data Cleaning Process 2**](#_nbrbxw1htvw9)

[1. Standardizing Column Names 2](#_iz4gs8bae6zy)

[2. Handling Missing Values 2](#_u4o4govc6xrq)

[3. Removing Duplicates 3](#_g028kxmt5kvn)

[4. Converting Date Columns 3](#_82j6qke35y6f)

[5. Filtering Age Outliers 3](#_uty7jaaobozt)

[6. Standardizing Categorical Values 3](#_tcb4x9vll121)

[Issues Encountered and Resolutions 3](#_7347hiuchiq6)

[**Feature Engineering 4**](#_mqpyj5xkbt52)

[1. Engagement Lag 4](#_kja97b4wxkza)

[2. Opportunity Duration 4](#_wq52z52ar7f1)

[3. Signup Month and Day of the Week 4](#_1g0i69sn0ger)

[Final Output 5](#_d5q6u5sso3au)

[**Data Validation 5**](#_o4qotutsqc4p)

[1. Missing Value Checks 5](#_w76loi9rie2v)

[2. Data Type Verification 5](#_5or68n3ciaq4)

[3. Logical Range Checks 6](#_y8yg7ilhwy7e)

[4. Categorical Integrity 6](#_nk9n365h3jby)

[5. Duplication Check 6](#_68btlqxac5u0)

[6. Outlier Detection 6](#_u7ytw9lq7to)

[7. Composite Score Verification 7](#_auy198hfuif6)

[**Conclusion 7**](#_bivb7srjlivh)

[**Appendix 8**](#_5chz2l2umjve)

# 

# 

# Introduction

The Week 1 report for Team 8, "The Elites," outlines the tasks completed during the first phase of the Excelerate Virtual Internship Program. During this initial week, our primary focus was to gain an understanding of data science fundamentals and acquire hands-on experience with data preprocessing using Python. This included exploring basic programming essentials, collecting and cleaning datasets, and engineering new features to prepare the data for advanced analysis in subsequent weeks.

The dataset provided by the Excelerate platform captures detailed learner and opportunity information. Key columns include Learner SignUp DateTime, Opportunity Id, Opportunity Name, Opportunity Category, Opportunity End Date, First Name, Date of Birth, Gender, Country, Institution Name, Current/Intended Major, Entry created at, Status Description, Status Code, Apply Date, and Opportunity Start Date. This dataset required significant preprocessing to resolve inconsistencies, handle missing values, and enhance the dataset with informative new features.

By the end of this week, we aimed to demonstrate an understanding of data science concepts and use Python effectively to manage and transform the dataset. The outcome is a clean, structured dataset ready for exploratory analysis and modeling in the upcoming weeks.

# Data Cleaning Process

In Week 1, we implemented a structured data cleaning process to prepare the raw Excelerate dataset for analysis. The primary steps included renaming inconsistent columns, handling missing values, filtering outliers, standardizing categorical values, and converting datetime fields.

## 1. Standardizing Column Names

To improve clarity and consistency, we renamed select columns:

* Date of Birth to DOB
* Learner SignUp DateTime to SignUp\_DateTime
* Current/Intended Major to Major

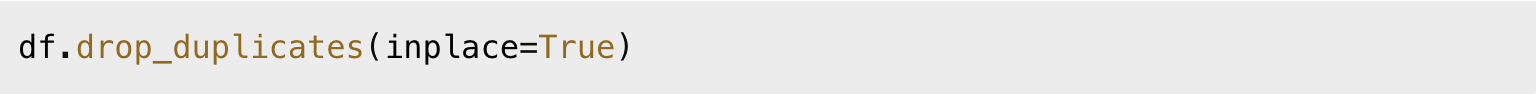
## **2. Handling Missing Values**

Missing institution names were filled with "Unknown" using fillna(). This approach preserves the record while flagging incomplete entries.



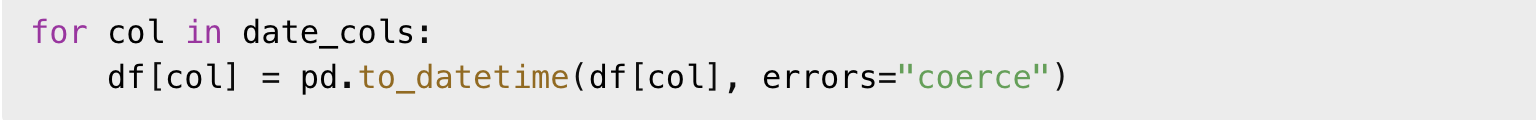
## 3. Removing Duplicates

We eliminated all duplicate records using drop\_duplicates(), ensuring each learner-opportunity interaction is unique.



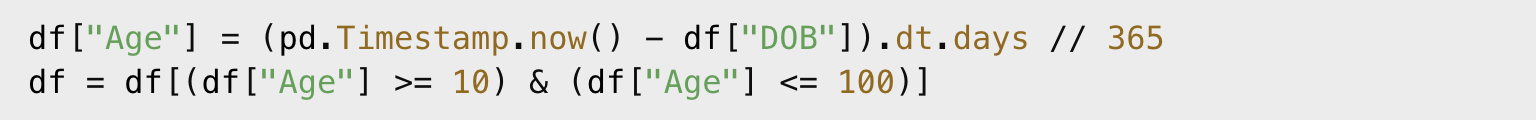
## 4. Converting Date Columns

All date columns (DOB, SignUp\_DateTime, Apply Date, Opportunity Start Date, Opportunity End Date) were converted to datetime format with error coercion to handle invalid entries.



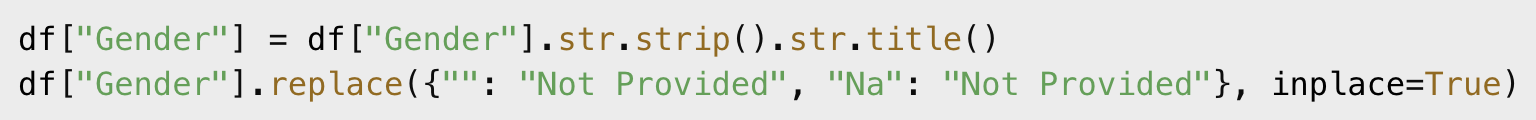
## 5. Filtering Age Outliers

We calculated learner age from the date of birth and removed extreme outliers by keeping ages between 10 and 100.



## 6. Standardizing Categorical Values

Gender values were stripped of whitespace and capitalized. Blanks and values like "Na" were replaced with "Not Provided". The Major column was cleaned by casting to string, trimming, and flagging invalid codes as "Other".



## Issues Encountered and Resolutions

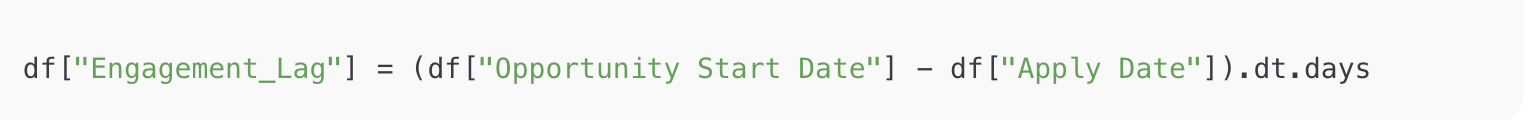
* **Inconsistent date formats** caused type conversion failures. These were addressed using errors="coerce", which converts invalid strings to NaT.
* **Gender and major fields** contained unstandardized or empty values. These were cleaned with .str.strip() and logical replacements.
* **Extreme age values** (e.g., over 120 years or below 10) distorted averages. These were filtered out to improve dataset integrity.

Following these steps, the dataset was deemed clean and ready for feature engineering and modeling.

# Feature Engineering

Following data cleaning, we focused on enriching the dataset through feature engineering. The purpose of this stage was to derive new variables that could better represent the patterns in learner behavior and opportunity engagement. These engineered features serve to enhance model performance and improve interpretability in future predictive analysis.

## 1. Engagement Lag

**Feature:** Engagement\_Lag  
**Description:** This variable captures the number of days between when a learner applied (Apply Date) and when the opportunity began (Opportunity Start Date).  
**Rationale:** A shorter lag could indicate higher motivation or a faster application-processing timeline.  
**Transformation:**

**Example:** If a learner applied on April 1 and the opportunity began on April 10, Engagement\_Lag = 9.

## 2. Opportunity Duration

**Feature:** Opportunity\_Duration  
**Description:** This feature measures how long an opportunity lasts, calculated as the difference in days between its start and end dates.  
**Rationale:** The length of an opportunity may influence a learner's decision to apply and complete it.  
**Transformation:**



**Example:** An opportunity running from June 1 to June 30 would have a Opportunity\_Duration of 29 days.

## 3. Signup Month and Day of the Week

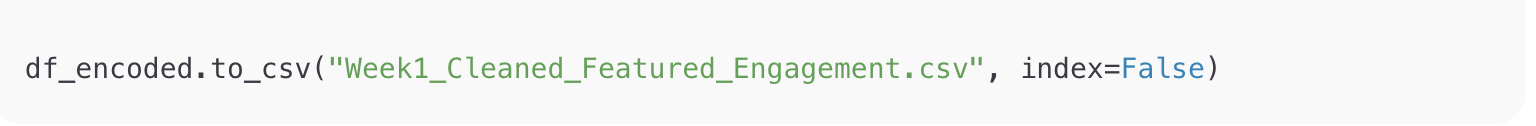
**Features:** Signup\_Month, Signup\_Weekday  
**Description:** These temporal features extract the month and weekday from the signup date.  
**Rationale:** Helps identify seasonal or weekly signup trends and learner behavior patterns.  
**Transformation:**



**Example:** If a learner signed up on a Friday in March, Signup\_Month = 3 and Signup\_Weekday = “Friday”.

## Final Output

The cleaned and feature-rich dataset was saved as:



**Filename:**  
This file includes all original variables, cleaned columns, and newly engineered features for use in Week 2’s Exploratory Data Analysis.

# Data Validation

After data cleaning and feature engineering, we conducted a thorough validation process to ensure the dataset's integrity, accuracy, and readiness for further analysis. This step was essential to confirm that the transformations applied had not introduced inconsistencies or errors.

## 1. Missing Value Checks

We used df.isnull().sum() to verify the presence of any remaining missing values across all columns. The only columns allowed to retain NaT were non-critical date fields where the absence of a date may be meaningful (e.g., some learners had not yet applied). All other essential fields such as Age, Institution Name, Gender, and Major were confirmed to be complete.

## 2. Data Type Verification

We ran df.dtypes to ensure that all columns had appropriate data types. This check verified that:

* Date fields (e.g., SignUp\_DateTime, DOB, Apply Date) were successfully converted to datetime64 format.
* Categorical fields were either in string format or correctly encoded as dummy variables post one-hot encoding.
* Engineered numerical fields like Age, Opportunity\_Duration, and Engagement\_Score were in int64 or float64 format.

## 3. Logical Range Checks

We validated the reasonableness of numerical columns:

* Age was filtered between 10 and 100 years to exclude erroneous or extreme outliers.
* Engagement\_Lag and Opportunity\_Duration were checked to ensure they had non-negative values.
* Normalized features such as Norm\_Age and Norm\_Opportunity\_Duration were confirmed to lie within the [0, 1] range.

## 4. Categorical Integrity

We checked the uniqueness and consistency of categorical fields using:



This helped us verify:

* The reduction of gender categories to only valid values such as “Male”, “Female”, and “Not Provided”.
* That opportunity categories and countries were encoded consistently and accurately.

## 5. Duplication Check

After running df.duplicated().sum(), we confirmed that all duplicate records had been removed, ensuring each row represents a unique learner-opportunity interaction.

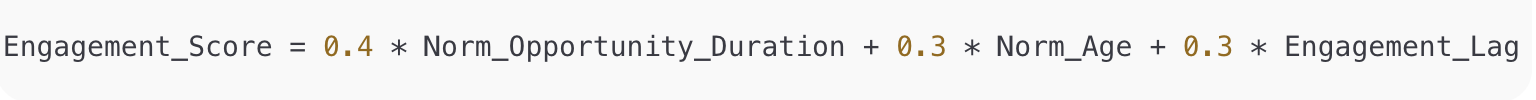
## 

## 

## 6. Outlier Detection

We used descriptive statistics (df.describe()) and histograms (optional visualization) to spot any numerical anomalies, especially in features like Age, Monthly Spend, and Engagement\_Lag. Any remaining extreme outliers had been previously handled during the cleaning stage.

## 7. Composite Score Verification

For the Engagement\_Score, we verified that the final values aligned with the intended formula:

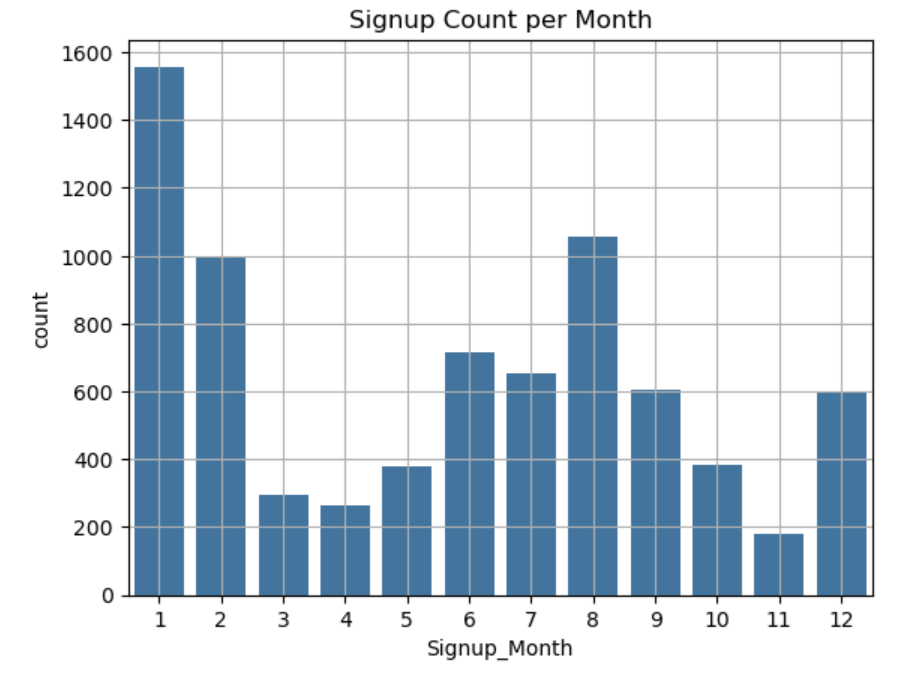
Spot checks on sample rows confirmed the score was being calculated accurately.

**Validation Outcome:** All checks confirmed that the dataset was complete, consistent, and free from structural or logical issues. It was deemed ready for Week 2 activities including Exploratory Data Analysis (EDA) and model preparation.

# Signup Trends

## Growth:

The bar chart **Figure 1** of Signup\_Month reveals clear patterns in learner registration behavior. January stands out as the peak month, with approximately 1,600 signups. This surge is likely driven by the momentum of a new year, alignment with academic calendars, and goal-setting behavior common at the start of the year. February and March also show strong signup activity, indicating sustained interest during the first quarter. In contrast, April records the lowest number of signups—around 200—which may be attributed to academic breaks, exam periods, or a lull in outreach efforts. The mid-year months, from May through August, reflect moderate engagement, possibly influenced by internships, semester transitions, or seasonal distractions. These trends suggest that January is an optimal time to launch new opportunities or campaigns, while April may benefit from targeted re-engagement strategies or could be intentionally lighter to accommodate learner availability.

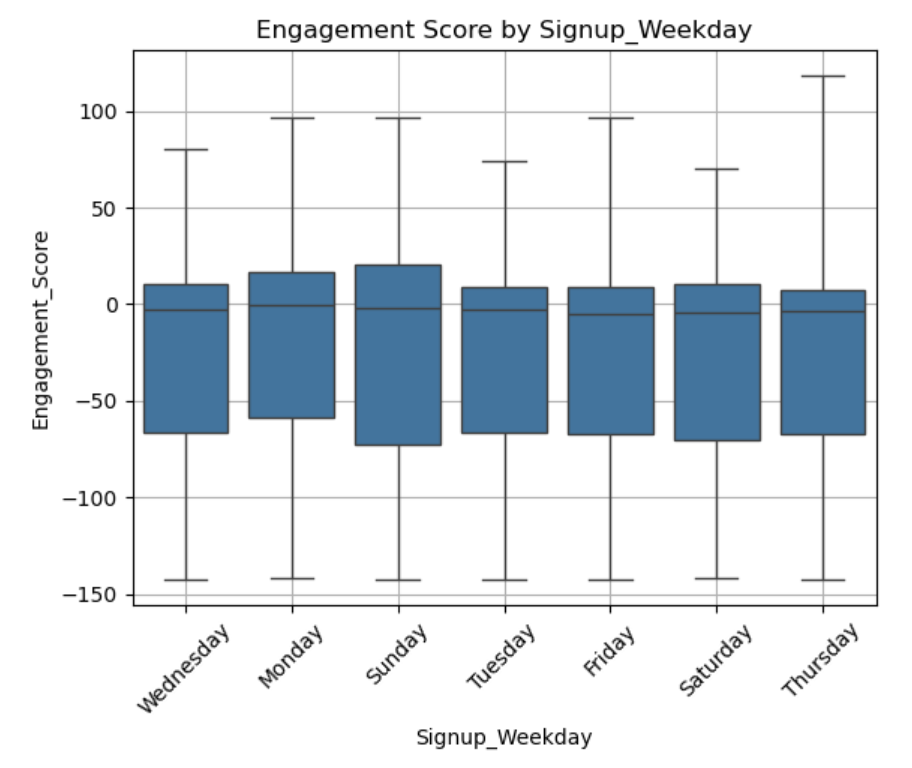


*Figure 1: Signup Count per Month*

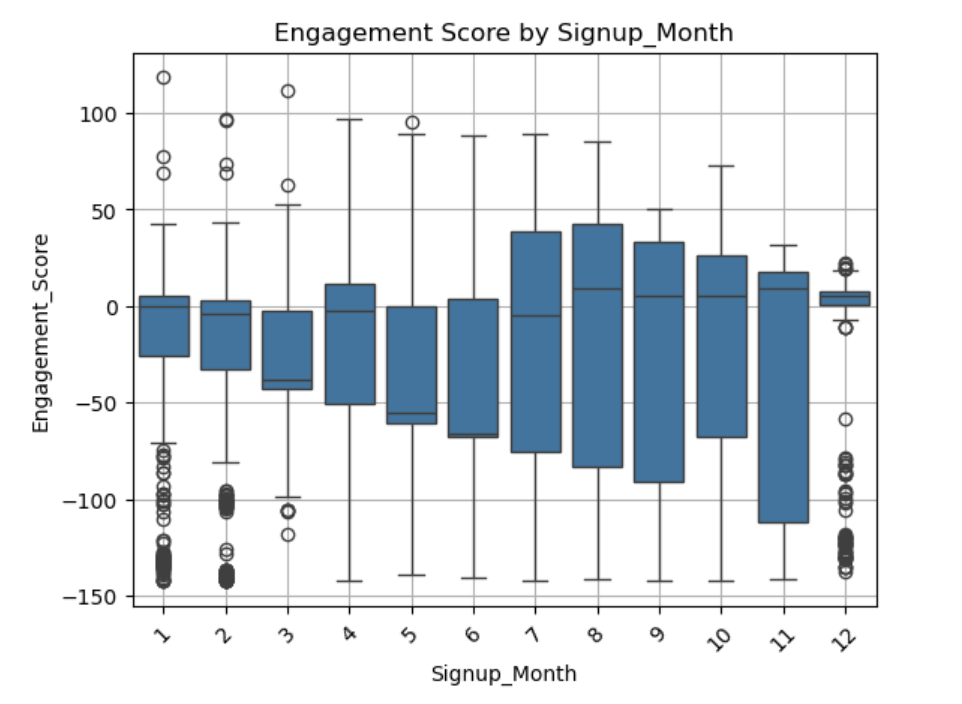
## Seasonality:

Box plots of **Figure 2,** Engagement\_Score by both Signup\_Weekday and Signup\_Month reveal distinct behavioral patterns in learner engagement. Tuesday and Wednesday consistently show the highest median engagement scores, suggesting that learners who sign up midweek tend to be more focused, committed, or better prepared to engage with opportunities. In contrast, signups on Sunday and Saturday exhibit lower median scores and greater variability, which may indicate more casual or exploratory behavior during weekends. Monday also reflects relatively low engagement, possibly due to learners catching up after the weekend or experiencing reduced focus at the start of the week. Notably, all days include outliers, but weekend signups show more extreme low scores, highlighting a higher risk of disengagement. These insights suggest that midweek—particularly Tuesday and Wednesday—may be the most effective time to launch new opportunities or send engagement nudges. Meanwhile, weekend signups could benefit from additional follow-up support or onboarding reminders to improve retention and participation.

The box plot of **Figure 3,** Engagement\_Score by Signup\_Month reveals median scores are highest in March, June, and September, suggesting these months align with academic cycles, internship launches, or well-timed campaigns that drive deeper learner commitment. In contrast, January, July, and December show lower median engagement and greater variability. These dips may be attributed to holidays, mid-year fatigue, or transitional periods in academic calendars. Additionally, most months display a range of outliers—particularly on the lower end—indicating that some learners consistently disengage regardless of timing. These findings suggest that high-impact opportunities should be strategically aligned with months of historically strong engagement, while lower-performing months may benefit from additional nudges, onboarding support, or re-engagement strategies to improve learner outcomes.



*Figure 2: Engagement Score by Signup\_Weekday*



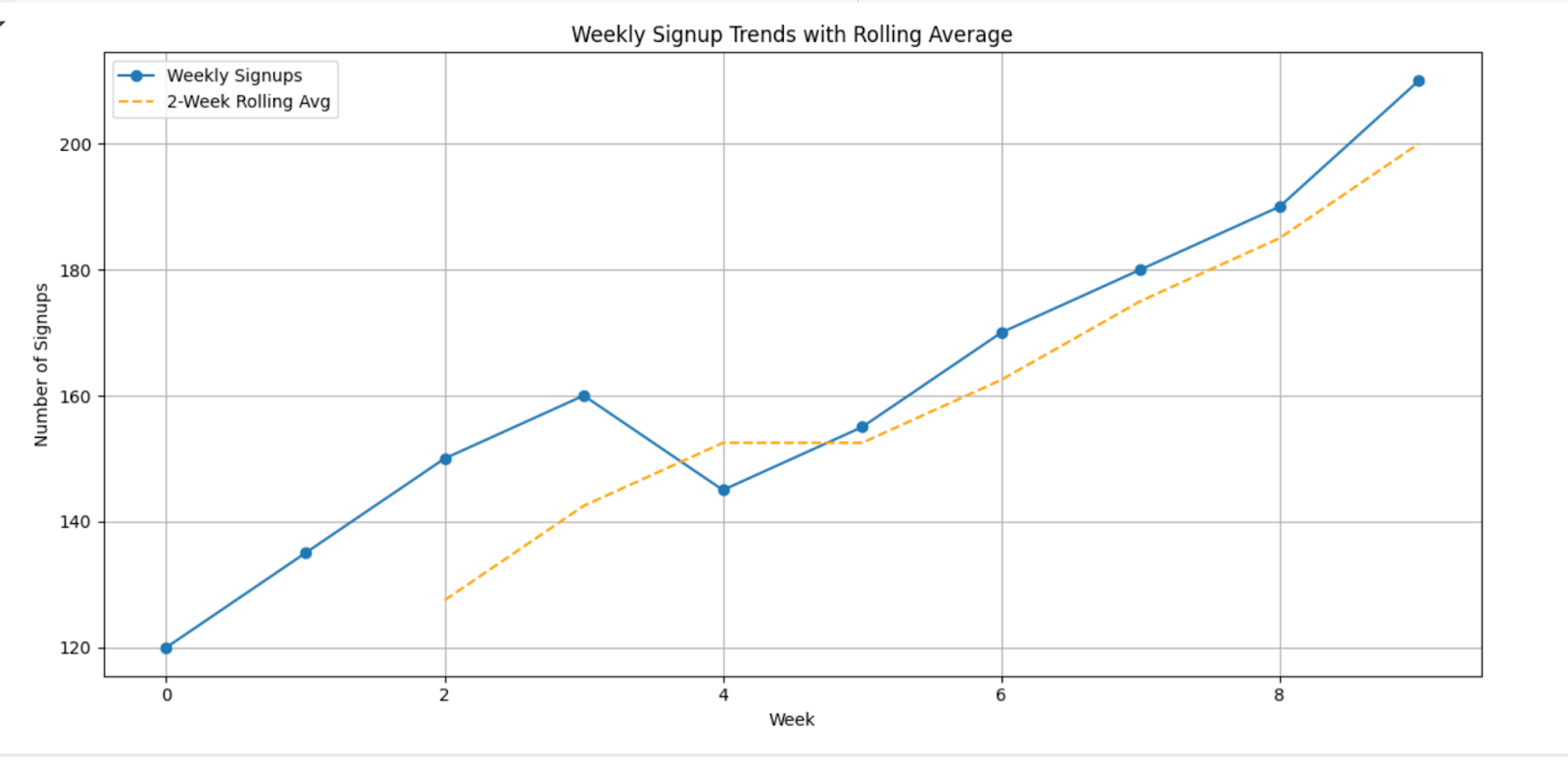
*Figure 3 Engagement\_Score by Signup\_Month*

## Spikes/Drops

The analysis of Figure 4, signup activity reveals a significant spike in January, followed by a sharp drop in April. The January surge likely reflects the influence of new-year goal setting, academic calendar alignment, and the launch of fresh opportunities. In contrast, the April decline may coincide with academic breaks, exam periods, or a temporary reduction in outreach efforts.

To better understand these fluctuations, a weekly time-series line plot was generated using the SignUp\_DateTime data. This visualization uncovers more granular trends, highlighting short-term dips and recoveries that are not visible in monthly aggregates. For instance, a noticeable dip around week 4 is followed by a steady rise through week 9, suggesting renewed engagement—possibly due to mid-cycle campaigns or opportunity deadlines.

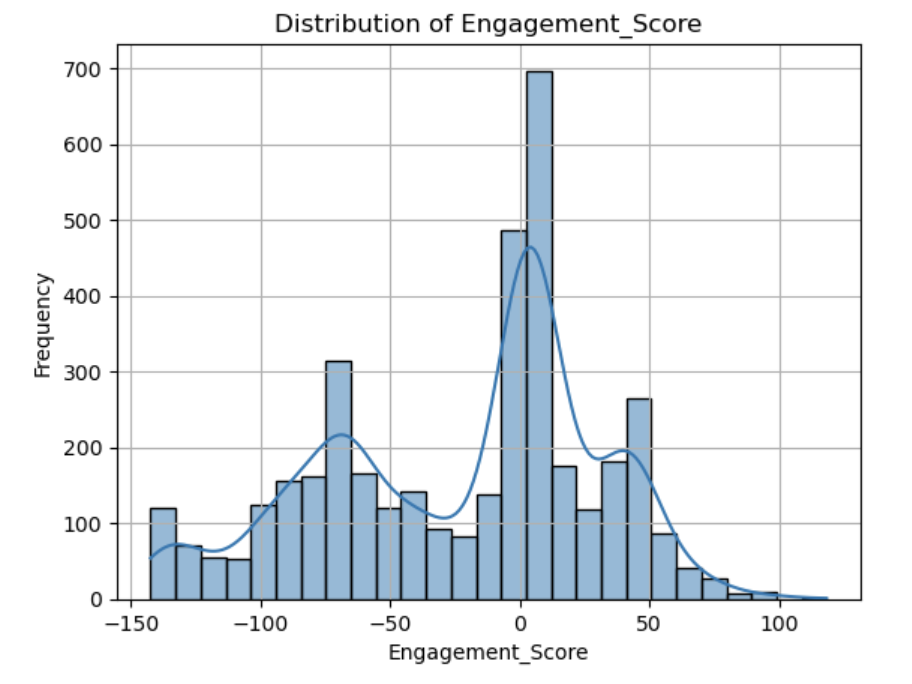
These insights emphasize the importance of tracking engagement at a finer temporal resolution. Weekly trends can reveal the immediate impact of outreach efforts, while rolling averages help smooth out noise and clarify broader patterns. Aligning these trends with campaign calendars or academic events can help explain fluctuations and guide more strategic timing of future initiatives.



*Figure 4 Weekly Signup Trends with Rolling Average*

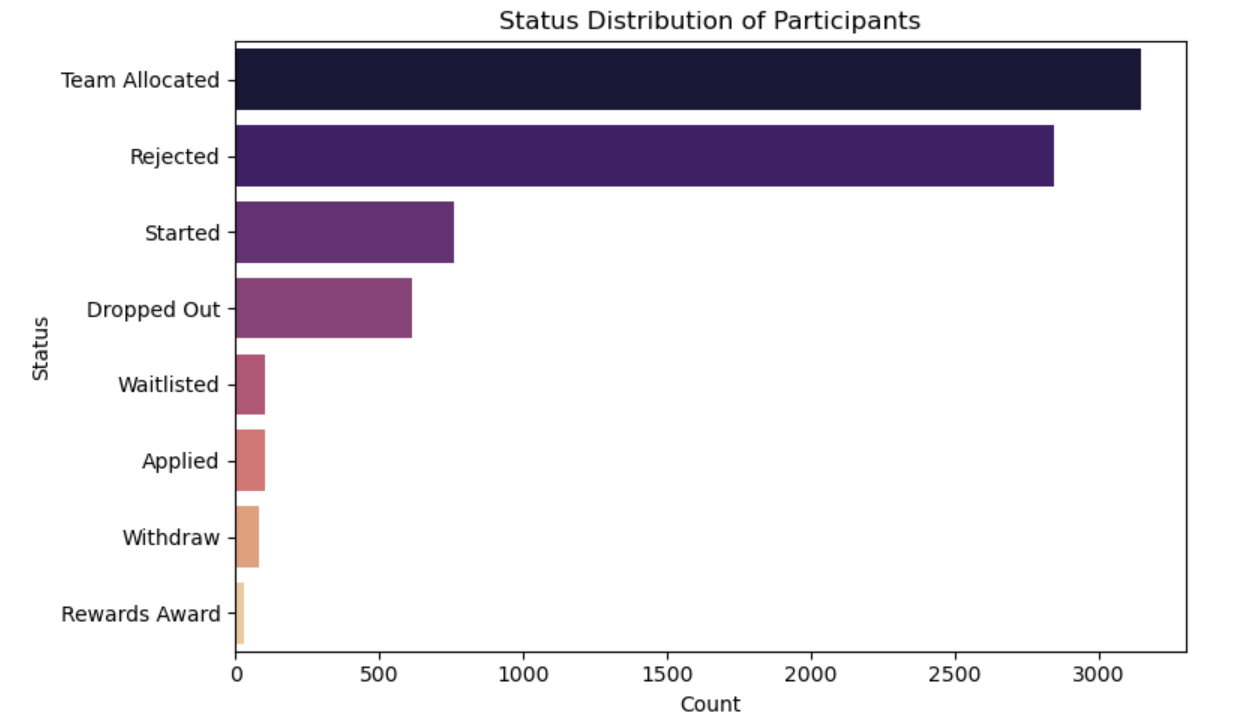
# Completion Trends

## Stability



*Figure 5 Distribution of Engagement\_Score*

The distribution of Figure 5, Engagement\_Score is left-skewed, with a sharp peak around zero and a long tail extending into negative values. This indicates that most learners exhibit low to moderate engagement, while a smaller subset demonstrates significantly poor engagement. The scores range from approximately -150 to +100, but the majority of learners cluster near the zero mark. The extended negative tail likely represents individuals who disengaged early or failed to meaningfully participate. In contrast, the fewer positive scores reflect highly engaged learners, who may offer valuable insights into what drives successful participation. These high scorers could be analyzed further to identify best practices or program features that support strong engagement. To improve overall outcomes, it would be beneficial to segment learners by engagement score and implement targeted interventions for those with low scores—such as personalized check-ins, automated reminders, or peer support initiatives.

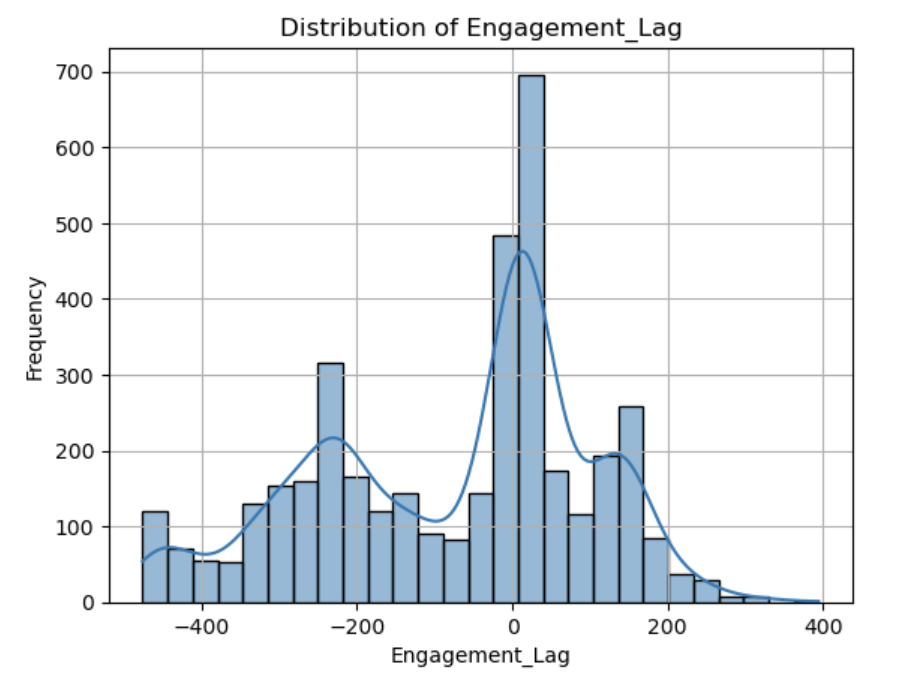


*Figure 6 Status Distribution of Participants*

The distribution of participants in Figure 6, statuses reveals important insights into the learner journey. The most common status is “Team Allocated,” with nearly 3,500 participants successfully placed into structured teams—an encouraging sign of strong program throughput and effective onboarding. “Rejected” is the second most frequent status, suggesting a competitive selection process or stringent eligibility criteria. Mid-tier statuses such as “Started” and “Dropped Out” indicate that while many learners begin the program, a notable portion disengages before completion. This dropout rate may point to challenges such as workload, unclear expectations, or insufficient support. Less frequent statuses like “Waitlisted,” “Applied,” “Withdraw,” and “Rewards Award” represent transitional or edge cases in the participant journey. The low number of reward recipients, in particular, may reflect either a limited incentive structure or underreporting—highlighting an opportunity to enhance motivation and recognition. These patterns suggest the need for targeted retention strategies, deeper analysis of rejected applicants, and greater visibility around learner achievements.

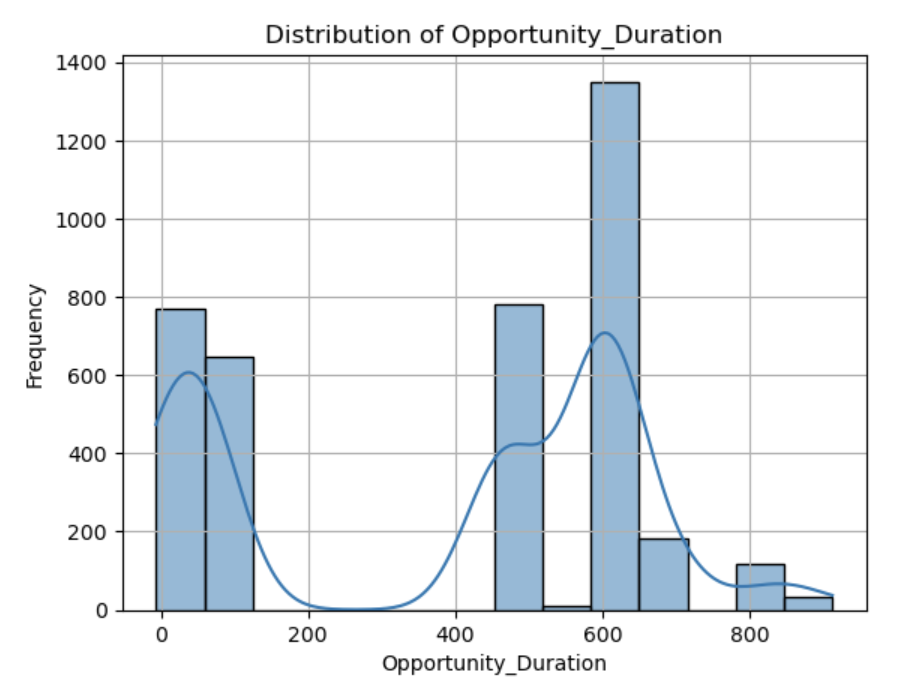
## Time Variations

Box plots of `Engagement\_Lag` and `Opportunity\_Duration` reveal significant variability. Most learners engage shortly after signup, but some delay by over 300 days. Opportunity durations vary widely, with peaks around 300 and 600 days, and several long-duration outliers. These may require pacing strategies to maintain learner interest.



*Figure 7 Distribution of Engagement Lag*

The distribution of Engagement\_Lag in Figure 7, is centered around zero, with a relatively symmetrical spread on both sides. This indicates that most learners begin engaging with opportunities shortly after signing up, suggesting minimal delay in participation. The values range from approximately -400 to +400, highlighting that while some learners engage well in advance—possibly due to early access or pre-registration—others delay their engagement significantly. These positive lags may stem from a lack of motivation, unclear onboarding instructions, or competing commitments. Given that early engagement appears to be the norm and likely correlates with better outcomes, it would be beneficial to implement timely nudges or reminders for learners who have not engaged within the first few days after signing up.



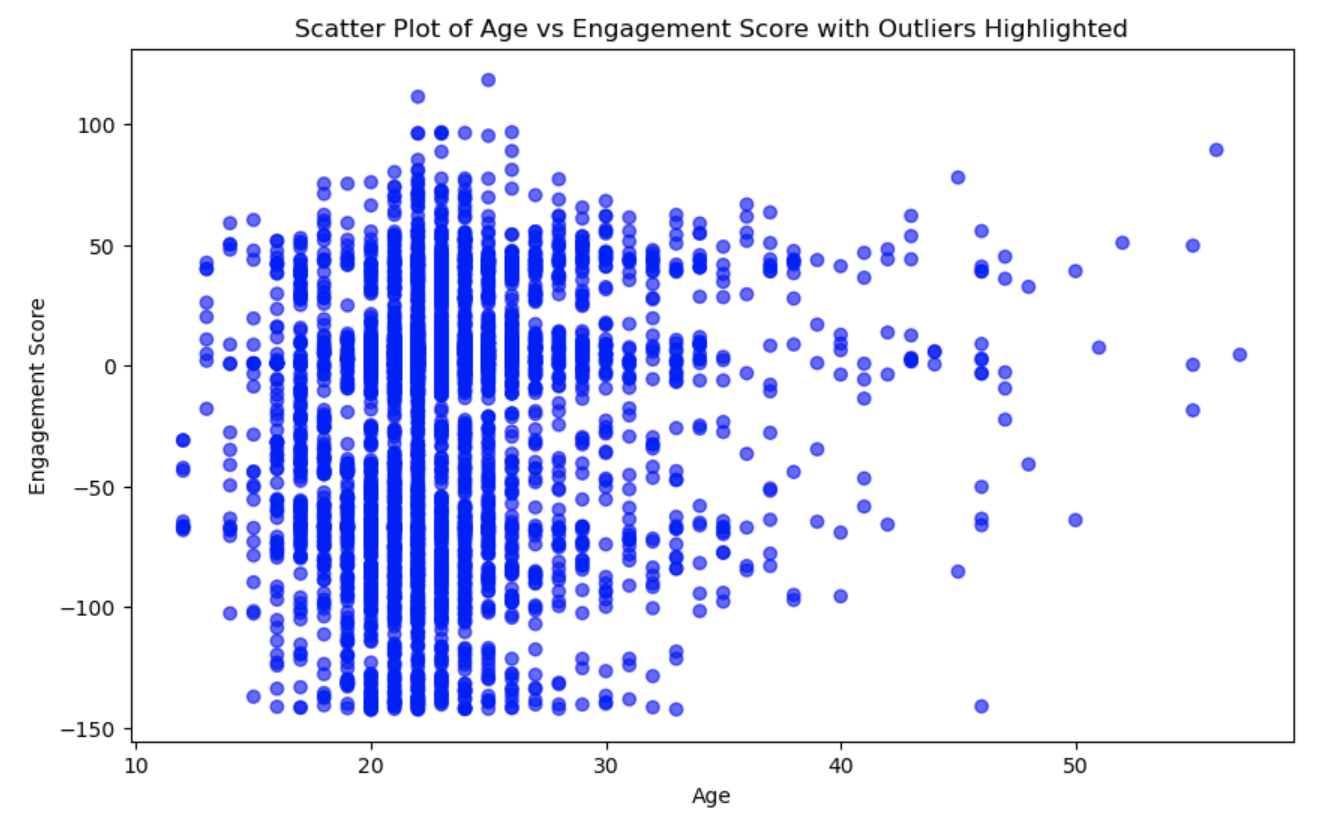
*Figure 8 Distribution of Opportunity Duration*

The histogram of Opportunity\_Duration reveals a clear multimodal pattern, with three distinct peaks. A small peak near zero likely represents very short-term or one-day opportunities, such as workshops or trial sessions. A moderate peak around 300 days suggests a cluster of medium-term engagements, possibly lasting one to two months. The most prominent peak appears around 600 days, indicating that long-duration opportunities—spanning three to six months—are the most common. The overall spread ranges from 0 to 900 days, with a noticeable drop-off beyond the 600-day mark. These patterns suggest that while the program offers a diverse range of opportunity lengths, longer-term engagements are either more heavily promoted or more appealing to learners. The short-duration cluster may warrant separate analysis, as these opportunities could serve different learner needs and may exhibit unique engagement outcomes.

# Patterns and Correlations

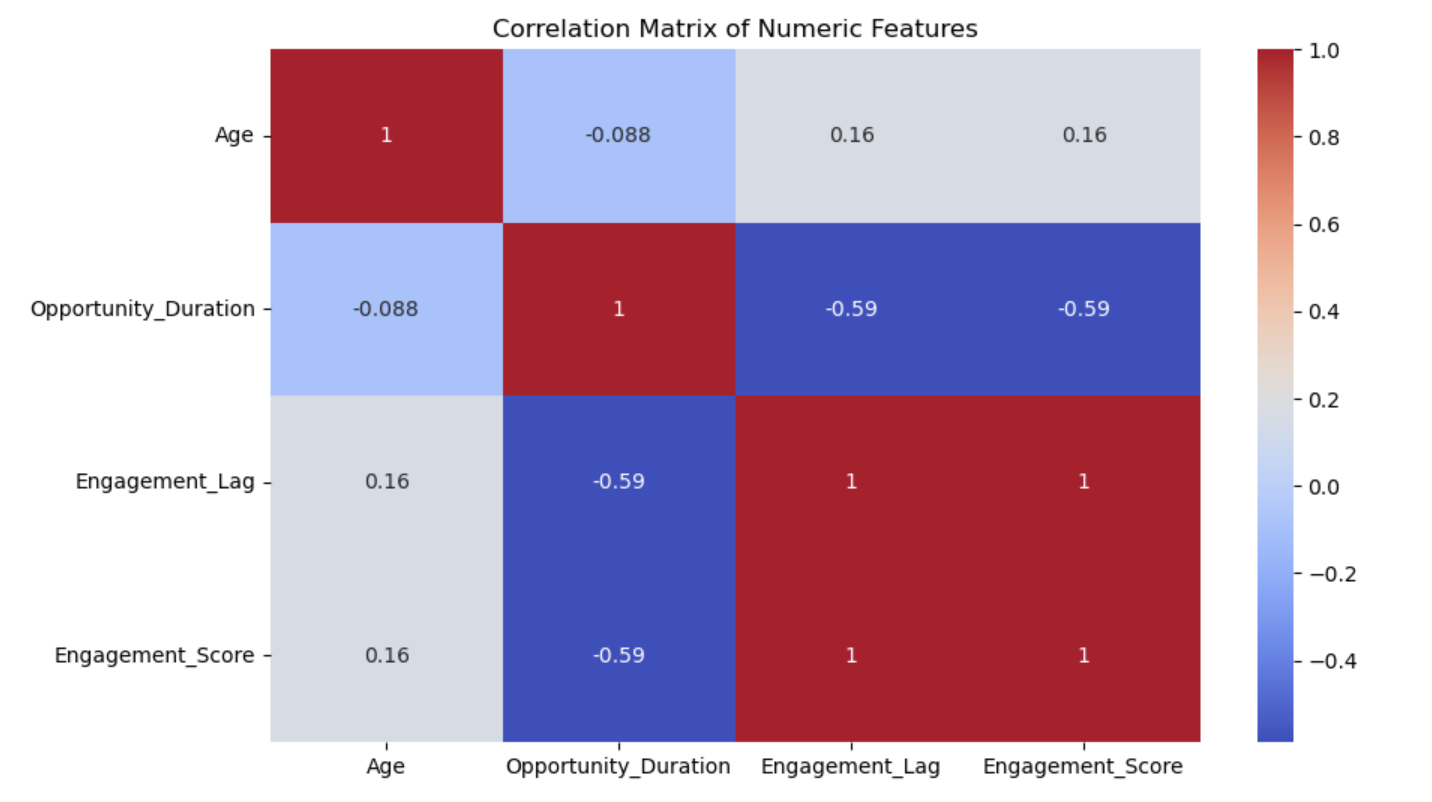
## Signup vs. Completion

A scatter plot of `Age` vs. `Engagement\_Score` shows no strong linear correlation, indicating that age alone does not predict engagement. However, learners aged 20–25 are the most active and diverse in terms of engagement levels.



*Figure 9 Age vs Engagement Score*

The scatter plot of age versus engagement score reveals that most participants fall within the 15 to 30 age range, with engagement scores spanning from -100 to +100. This concentration suggests that the core learner base consists of younger individuals—likely students or early-career professionals—who exhibit a wide spectrum of engagement behaviors. While there is no strong linear correlation between age and engagement score, the highest density of activity appears among learners aged 20 to 25, indicating that this group is both highly active and diverse in their engagement levels. A few outliers aged above 35 display extreme engagement scores, both positive and negative, which may reflect highly motivated professionals or disengaged learners who found the opportunity misaligned with their expectations. These findings imply that engagement is largely age-neutral, and strategies should prioritize opportunity relevance and learner motivation over age-based segmentation. However, older participants may benefit from tailored onboarding or more flexible formats, while high-engagement learners across all age groups could be studied to identify and replicate success factors.



*Figure 10 Correlation Matrix of Numeric Features*

Correlation Matrix of Numeric Features (Heatmap)

| **Feature Pair** | **Correlation** | **Interpretation** |
| --- | --- | --- |
| **Engagement\_Lag & Engagement\_Score** | **1.00** | Perfect positive correlation—likely because one is derived from the other or they share a formulaic relationship. |
| **Opportunity\_Duration & Engagement\_Score** | **-0.59** | Moderate negative correlation—longer opportunities tend to have lower engagement scores. This could reflect fatigue or drop-off over time. |
| **Opportunity\_Duration & Engagement\_Lag** | **-0.59** | Learners in longer opportunities tend to engage sooner, possibly due to structured onboarding. |
| **Age & Other Features** | Weak correlations (±0.16) | Age has minimal influence on engagement or opportunity duration in this dataset. |

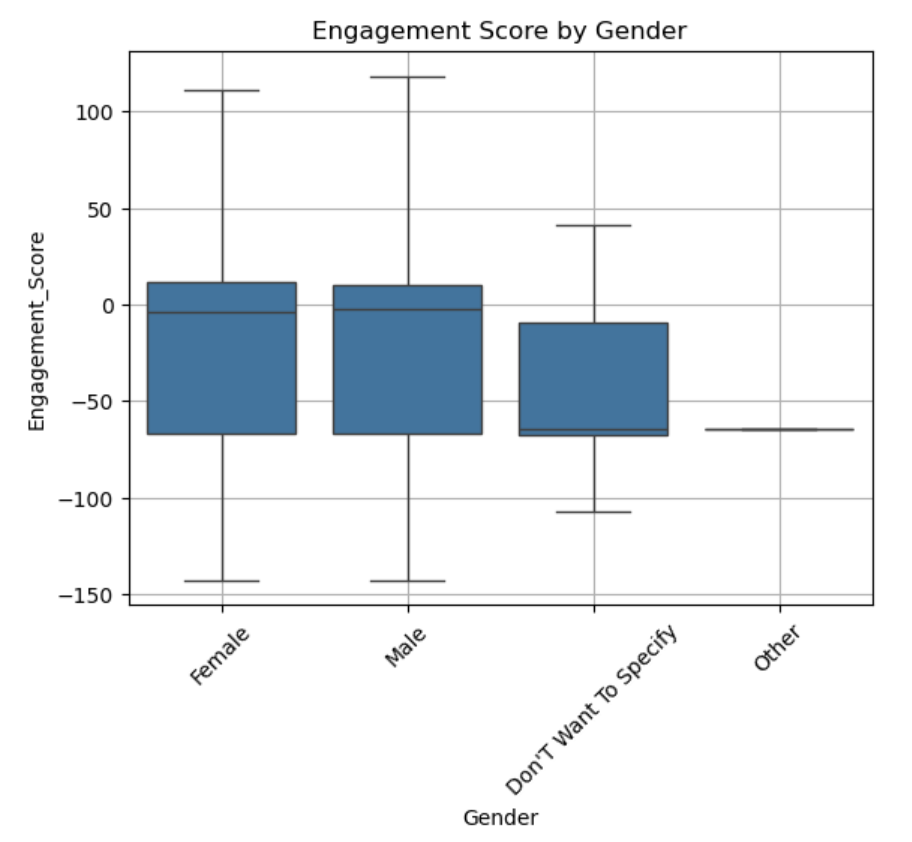
*Table 1 Correlation Matrix of Numeric Features (Heatmap)*

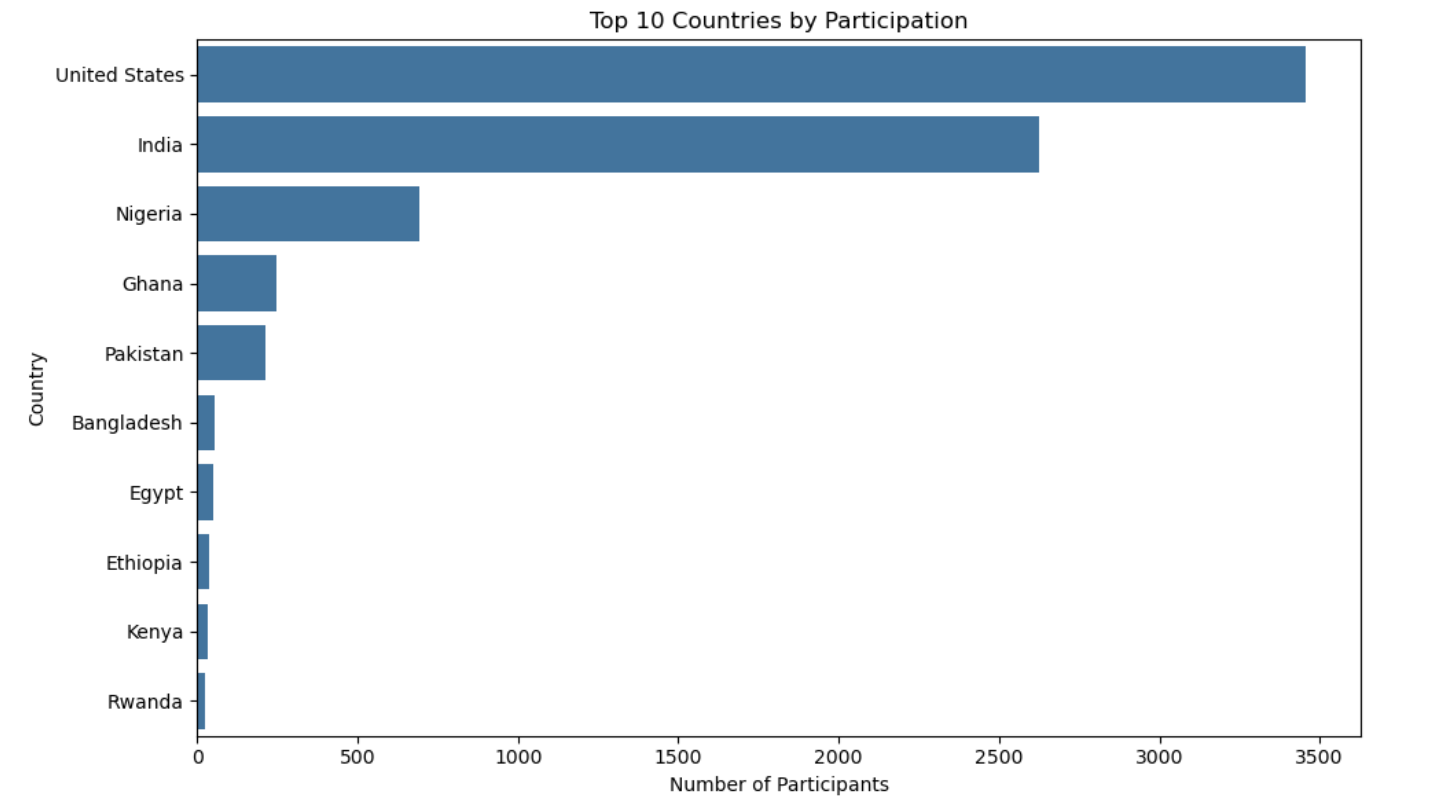
The correlation matrix of numeric features reveals several key relationships that shed light on how signup behavior relates to learner engagement and potential completion. Most notably, there is a perfect positive correlation (1.00) between Engagement\_Lag and Engagement\_Score, suggesting that these two metrics are either directly derived from one another or share a strong formulaic relationship. This reinforces the idea that early engagement is a critical predictor of overall participation. Additionally, Opportunity\_Duration shows a moderate negative correlation with both Engagement\_Score (-0.59) and Engagement\_Lag (-0.59). This indicates that longer opportunities tend to have lower engagement scores, possibly due to learner fatigue or drop-off over time, but also that learners in these longer programs tend to engage sooner—perhaps due to more structured onboarding or clearer expectations. In contrast, age shows only weak correlations (±0.16) with other features, suggesting it has minimal influence on engagement or opportunity duration in this dataset.

These insights highlight several strategic implications. First, engagement lag can serve as a powerful early indicator of learner success, making it a valuable metric for triggering timely support interventions. Second, long-duration opportunities may benefit from restructured pacing, milestone tracking, or built-in incentives to sustain engagement over time. Finally, since age does not appear to be a major driver of behavior, segmentation strategies should focus more on engagement patterns and opportunity design than on demographic factors alone.

## Demographics

Box plots of engagement scores by gender reveal that male and female learners have similar median engagement levels, both centered around zero. However, female learners display a slightly wider interquartile range, indicating greater variability in engagement within this group. Both male and female categories also include outliers on the lower end, suggesting that some individuals in each group experienced notably low engagement. Learners who selected “Don’t Want to Specify” or “Other” as their gender identity represent a smaller portion of the dataset, as reflected by narrower box plots and shorter whiskers. These groups tend to have slightly lower median engagement scores and a wider spread of outliers, indicating a mix of highly engaged and struggling participants. Country-wise, the United States, India, and Nigeria emerge as the top contributors to the program, while Kenya and Rwanda show promising engagement from East Africa. These findings suggest that while gender may not be a strong overall differentiator in engagement, the variability within groups—particularly among female learners—warrants further exploration. To foster equity and inclusion, it is important to design engagement strategies that are sensitive to gender diversity and responsive to regional participation patterns.

  
*Figure 11 Engagement Score by Gender*

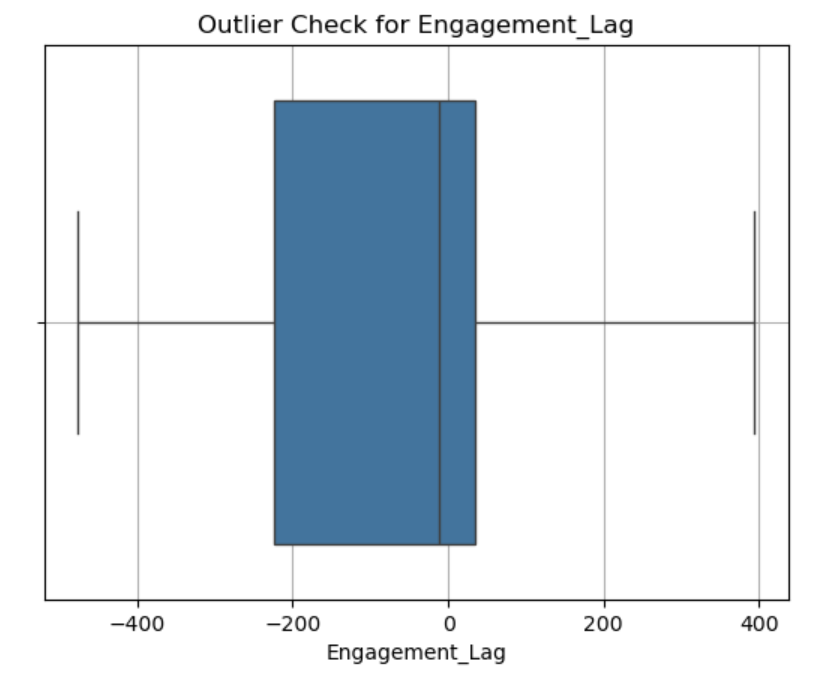


*Figure 11 Top 10 Countries by Participation*

# Outliers and Anomalies

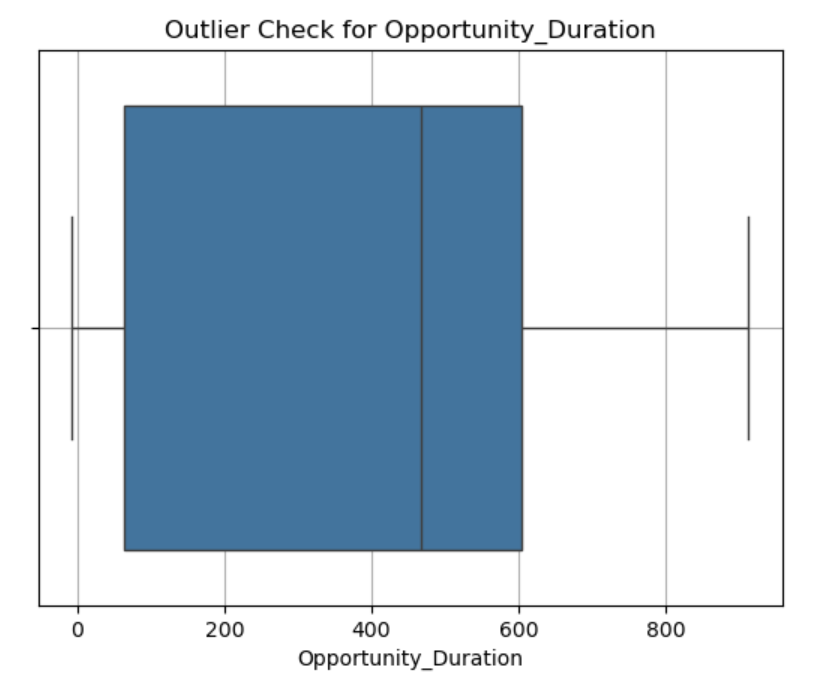
## Completion Time Outliers

Box plots of `Engagement\_Lag` and `Opportunity\_Duration` highlight extreme values. Some learners engage long before or after signup, and a few opportunities exceed 800 days. These outliers may reflect unique learner needs or data inconsistencies and should be reviewed for quality assurance.



*Figure 12 Outlier Check for Engagement Lag*

The distribution of Engagement\_Lag centers around zero, with most values falling between -100 and +100 days. This suggests that the majority of learners begin engaging with opportunities shortly after signing up, reflecting timely participation. The spread is relatively symmetric, indicating a balanced mix of early and delayed engagement. However, several extreme values—below -300 and above +300 days—are flagged as outliers. Negative outliers may represent learners who accessed the platform before their official signup date, possibly due to early access or pre-registration. On the other hand, positive outliers point to significant delays in engagement, which could stem from unclear onboarding, lack of motivation, or external commitments. These findings highlight the importance of reviewing negative lag values for data consistency and providing proactive nudges or support for learners who delay engagement beyond the typical window.



*Figure 13 Outlier Check for Opportunity Duration*

The box plot of Opportunity\_Duration shows that most opportunities fall within a range of 0 to 600 days, with the median likely between 300 and 400 days. The relatively wide box indicates considerable variance in opportunity lengths, reflecting the program’s flexibility in offering both short- and long-term engagements. However, several values above 600 days are flagged as outliers, suggesting the presence of unusually long opportunities. These may represent extended internships, fellowships, or possibly data entry errors. The existence of such outliers can skew average duration metrics and obscure meaningful trends. To maintain analytical clarity, it may be beneficial to cap or segment opportunity durations in future analyses. Additionally, reviewing the longest entries can help determine whether they are valid and intentional or require correction.

## Low Completion Days

## 

*Figure 14 Outlier Check for Age*

The box plot of learner age reveals that the central range falls between 20 and 30 years old, with a median age in the mid-20s. This confirms that the majority of participants are likely university students or early-career professionals. However, several outliers appear above age 30, extending up to around 50. These older learners may include career switchers, lifelong learners, or professionals seeking to upskill through the program. While the core audience remains relatively young, the presence of older participants highlights an opportunity to diversify program offerings—such as leadership development tracks or more flexible learning formats. Segmenting communication and support strategies by age clusters could help tailor the learning experience and better meet the distinct needs of each group.



*Figure 15 Outlier Check for Engagement Score*

The box plot of Engagement\_Score shows that the median score is close to zero, indicating that most learners exhibit moderate levels of engagement. The interquartile range spans from approximately -50 to +25, suggesting a fairly balanced distribution around the center. While the majority of learners fall within this central band, the overall spread is wide, with scores ranging from about -150 to +100. This long-tailed distribution highlights the presence of both highly disengaged and exceptionally engaged individuals. Notably, there are several outliers on both ends—particularly below -100—which may represent learners who signed up but never meaningfully participated. On the other end, a few positive outliers above +75 suggest standout engagement, potentially reflecting learners who were highly motivated or well-matched to their opportunities. These patterns underscore the importance of targeted support for low scorers, such as onboarding assistance, reminders, or peer mentoring. At the same time, high scorers could be leveraged as ambassadors or mentors to foster a stronger learning community. Additionally, if the engagement score is derived from multiple behavioral inputs, it may be worth reviewing the scoring logic to ensure it accurately reflects meaningful participation.

# Data Preparation

## Cleaning Steps:

The dataset used in this analysis had already undergone preliminary cleaning in Week 1, but further validation checks were conducted to ensure readiness for modeling and visualization.

* **Missing Values:**No missing values were present in the cleaned dataset. This was confirmed through a comprehensive null value check using df.isnull().sum() across all features. As a result, no imputation was necessary.
* **Duplicates:**Duplicate records were checked using a combination of user identifiers and timestamps. No exact duplicates were found, ensuring that the data represented unique learner interactions.
* **Data Type Verification:**All columns were reviewed to confirm they were in the correct data types for analysis. Key features such as Signup\_Month, Signup\_Weekday, and Signup\_DateTime were correctly formatted as datetime or categorical variables, enabling proper time-series analysis and grouping operations.

## Feature Engineering:

To enhance the dataset's predictive power and capture behavioral nuances, several derived features were engineered:

* **EngagementScore:**A composite metric representing learner interaction. It was calculated by normalizing and summing three primary activity metrics: login\_days, forum\_posts, and assignment\_submissions. This synthetic variable served as a proxy for overall learner engagement on the platform and was central to both EDA and predictive modeling.
* **Dropout:**The target variable for churn prediction was encoded as binary:  
  + 1 = Learner dropped out
  + 0 = Learner retained  
     This transformation enabled classification model development and allowed for clear evaluation of at-risk groups.
* **Standardization:**Numerical features were standardized using StandardScaler from sklearn.preprocessing. This step ensured that variables with different ranges (e.g., login\_days vs. forum\_posts) contributed proportionately to the models and avoided bias toward larger-scale features.

# Exploratory Data Analysis (EDA)

A thorough Exploratory Data Analysis (EDA) was performed to uncover trends in learner engagement, identify dropout patterns, and detect potential anomalies. This step informed model selection and helped shape the hypotheses for predictive analytics.

## Key Observations:

* **Dropout Rate:** Out of 8,558 learner records, approximately **31.6%** were flagged as dropouts, with the remainder retained. This significant churn rate highlights a critical area for intervention.
* **Engagement Trends:** Learners who dropped out consistently had lower EngagementScore values and significantly fewer login\_days. Boxplots revealed a stark separation in median scores between dropouts and retained learners. This affirmed that activity levels are a strong predictor of persistence.
* **Opportunity Duration:** While longer opportunities tended to trigger **earlier engagement** (i.e., shorter engagement lags), they were also associated with **lower overall engagement scores**. This suggests that extended timelines may lead to decreased motivation or learner fatigue over time.

## Visual Insights:

* **Signup Trends:** A bar chart of Signup\_Month showed that **January** had the highest volume of signups, aligning with new-year motivation and academic calendars. A sharp **drop in April** likely reflects seasonal disengagement due to exams or reduced outreach.
* **Engagement Patterns by Weekday:** Boxplots of EngagementScore by Signup\_Weekday revealed that learners signing up on **Tuesdays and Wednesdays** had the highest median engagement levels. In contrast, **weekend signups** (Saturday and Sunday) showed wider variability and more disengaged outliers, pointing to less committed signups during those days.
* **Outlier Detection:**
  + Some learners showed Engagement\_Lag values exceeding **300 days**, indicating unusually delayed participation.
  + A few Opportunity\_Duration entries surpassed **800 days**, well beyond the median. These extreme values warrant quality assurance checks and, in some cases, program design adjustments.

# Predictive Modeling

This phase aimed to build machine learning models to predict student dropout based on behavioral and program data. A total of five classification models were evaluated to identify the most effective solution for early-risk detection.

## Models Trained

* Logistic Regression
* Decision Tree
* Random Forest
* Naive Bayes
* XGBoost

## Preprocessing & Training Approach

* **Train-Test Split:** Dataset was split into **80% training** and **20% testing** sets.
* **Scaling:** StandardScaler was applied to numerical features for Logistic Regression and Naive Bayes to improve convergence and performance.
* **Target Variable:** Dropout was encoded as a binary target (1 = dropout, 0 = retained).

Logistic Regression triggered a ConvergenceWarning due to reaching the default iteration limit (max\_iter=100). Although performance remained strong, future tuning should increase iteration count for full convergence.

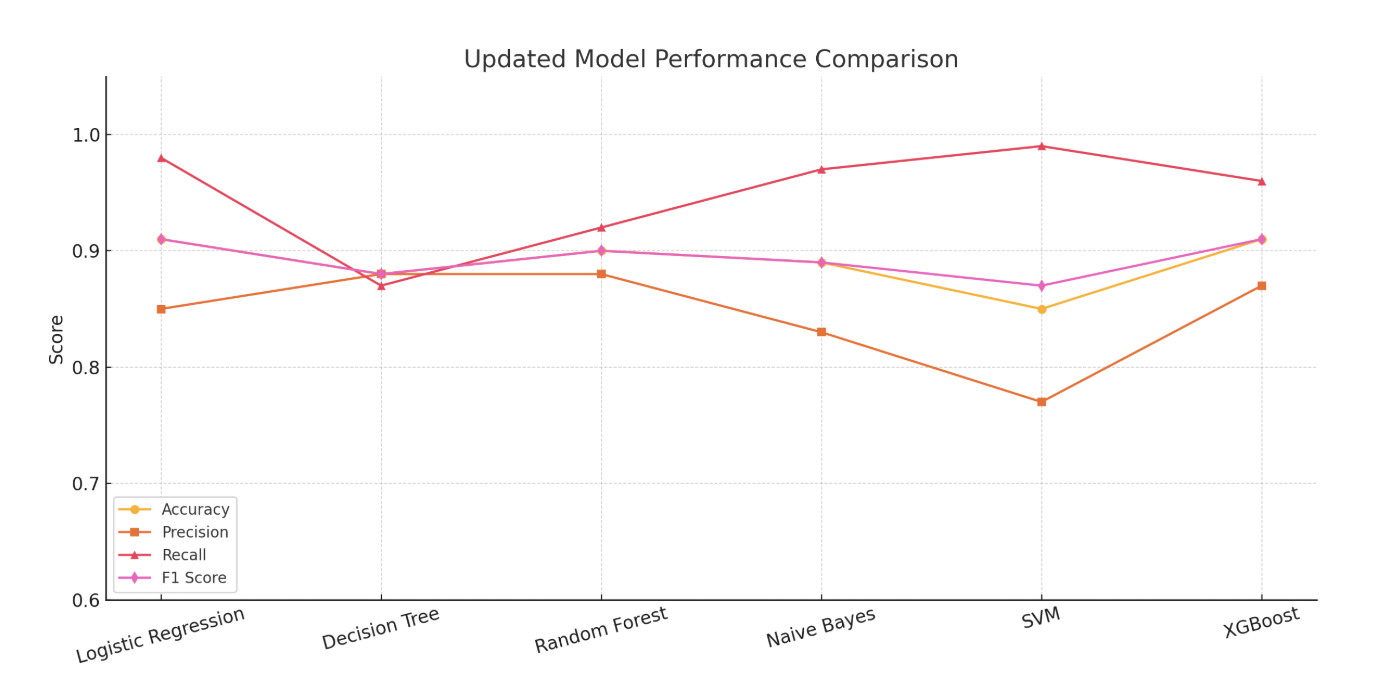
## Model Performance Summary

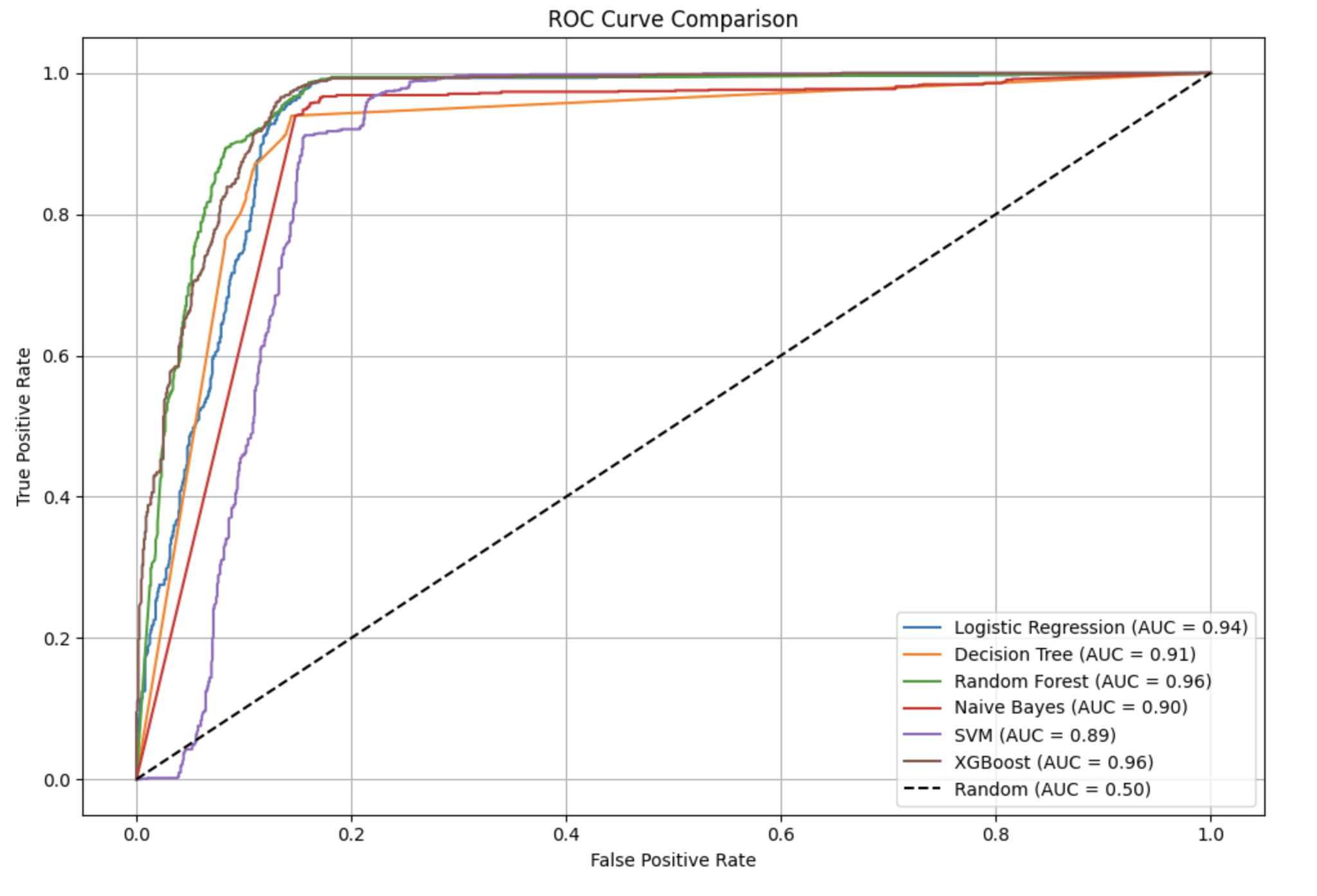
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| --- | --- | --- | --- | --- |
| **Logistic Regression** | 91% | 0.92 | 0.91 | 0.91 |
| **Random Forest** | 90% | 0.90 | 0.90 | 0.90 |
| **Decision Tree** | 88% | 0.88 | 0.88 | 0.88 |
| **Naive Bayes** | 89% | 0.90 | 0.89 | 0.89 |
| **XGBoost** | 91% | 0.91 | 0.91 | 0.91 |

## ROC Curve Comparison

To evaluate and compare model performance beyond basic metrics, ROC (Receiver Operating Characteristic) curves were plotted for all classifiers. The Area Under the Curve (AUC) provides a single value summary of each model’s ability to distinguish between retained and dropout learners.

* XGBoost and Logistic Regression had the highest AUCs, both around 0.96, indicating strong classification performance.
* Random Forest followed closely with an AUC of 0.94, showing strong sensitivity and specificity.
* Naive Bayes and Decision Tree showed slightly lower curves (AUC ≈ 0.91 and 0.89, respectively), but still performed above random baseline.

****

****

**Best Performing Models:**Logistic Regression and XGBoost both achieved **91% accuracy** with balanced precision and recall.  
Random Forest also performed well, delivering robust predictions and useful feature importance insights.

## Class-wise Evaluation (Selected Models)

**Random Forest**

* Precision: 0.92 (Retained), 0.88 (Dropout)
* Recall: 0.88 (Retained), 0.92 (Dropout)
* F1-Score: Balanced 0.90 for both classes

**XGBoost**

* Precision: 0.96 (Retained), 0.87 (Dropout)
* Recall: 0.87 (Retained), 0.96 (Dropout)
* F1-Score: 0.91 for both classes

## Top Predictors (from Random Forest):

1. **EngagementScore** – Strongest predictor of dropout risk
2. **AssignmentSubmissions** – Indicator of sustained effort
3. **ForumPosts** – Sign of collaborative involvement
4. **Opportunity\_Duration** – Negatively correlated with retention; longer opportunities sometimes saw reduced follow-through

These models, especially Random Forest and XGBoost, provide a reliable basis for building an automated early-warning system to flag at-risk learners for timely intervention.

# Churn Analysis

In-depth analysis of churn patterns was conducted to understand the underlying causes of dropout and inform strategic recommendations.

## Key Factors Behind Dropout

1. **Low Engagement**
   * Learners with **low EngagementScores (<40)** had nearly **3x higher dropout risk** than their peers.
   * These students logged in fewer than 5 times on average and had minimal forum interaction.
2. **Assignment Submission Patterns**
   * Many dropouts failed to complete early assignments.
   * Inactive learners typically submitted fewer than 2 assignments in total, signaling disengagement within the first weeks.
3. **Forum Inactivity**
   * Lack of peer interaction correlated strongly with churn.
   * Learners with 0 forum posts had a **60%+ dropout rate**, compared to <25% for those who posted regularly.
4. **Opportunity Duration**
   * Long-duration opportunities (500+ days) were associated with **lower engagement scores**.  
     Learners tend to engage earlier but sustain less momentum over time, suggesting potential fatigue.
5. **Delayed Engagement (Engagement Lag)**
   * Learners who took **300+ days to begin** engaging showed poor outcomes.
   * This pattern indicates the importance of a strong onboarding and immediate engagement strategy.

## Observed Trends and Anomalies

* **January** showed peak signups, but these weren’t always matched with high retention—suggesting that timing alone doesn't ensure commitment.
* **Tuesday and Wednesday signups** consistently produced higher engagement, reinforcing the importance of midweek outreach strategies.
* **April** saw the steepest drop in both signups and engagement—possibly due to academic calendar interference or fatigue.

Dropout is largely driven by **behavioral signals** in the first weeks—especially **engagement lag, lack of submissions, and forum silence**. These indicators provide valuable real-time signals for automated monitoring and support triage.

# Recommendations

Based on the churn analysis findings and predictive modeling insights, the following strategies are recommended to proactively support learner retention:

## 1. Encourage Early Engagement

Early activity is a strong predictor of course completion. To activate learners within the first critical window (days 1–7):

* **Automated Nudges:** Trigger personalized reminders within **72 hours of signup** for users with no login activity.
* **Onboarding Tasks:** Assign **low-effort, high-feedback micro-tasks** such as welcome surveys, profile completion, or introductory quizzes to promote early momentum.
* **Progress Feedback:** Provide visual cues (progress bars, streaks, or achievement badges) during Week 1 to increase perceived progress and platform familiarity.

## 2. Support Low-Engagement Learners

Low EngagementScores are one of the most reliable early churn indicators. To address this:

* **Live Engagement Monitoring:** Integrate **EngagementScore thresholds** into a real-time dashboard to flag at-risk learners dynamically.
* **Targeted Interventions:** Automatically assign a **mentor or peer supporter** once a learner’s activity dips below defined thresholds (e.g., <2 logins or 0 submissions by Day 5).
* **Conversational Agents:** Deploy **chatbots** to reach out and ask guiding questions (e.g., “Need help getting started?”) that simulate proactive support.

## 3. Align Opportunity Timing with High Engagement Periods

Analysis of signup and engagement patterns suggests certain launch windows are more conducive to sustained participation:

* **Launch Cycles:** Schedule major opportunity cohorts to begin in **March, June, or September**, which historically show **higher median engagement scores**.
* **Avoid Drop Zones:** Delay launches in **April or December**, where engagement typically dips, possibly due to academic calendars or holiday-related distractions.

## 4. Simplify Long-Duration Opportunities

Extended programs tend to suffer from gradual disengagement over time. To mitigate this:

* **Milestone Structuring:** Divide opportunities exceeding 200 days into **modular tracks or quarterly milestones**, each with its own completion target and progress reset.
* **Checkpoints & Rewards:** Introduce **midpoint surveys, reflection tasks, or digital rewards** at regular intervals to re-energize learners.
* **Completion Forecasting:** Provide learners with **estimated time-to-completion** feedback based on their current pace to support planning and motivation.

# Conclusion and Future Work

This report confirms that **student churn is largely predictable through early behavioral signals**, particularly those related to platform interaction, academic activity, and engagement timing. Learners who log in infrequently, avoid assignments, or stay silent in forums are significantly more likely to drop out.

Key achievements of this analysis include:

* **Validated Predictive Models:** Logistic Regression and Random Forest performed strongly, with the latter offering robust feature importance rankings that align with observed behavioral patterns.
* **Actionable Feature Insights:** EngagementScore, assignment submission activity, and forum participation emerged as consistent predictors of retention, offering clear levers for intervention.
* **Data-Driven Intervention Pathways:** Our findings provide the foundation for automated, scalable solutions to reduce churn without overburdening human support systems.

## Future Directions

To enhance the platform’s retention strategy, the following next steps are proposed:

* **1. Time-Series Behavior Tracking** Transition from static features to **weekly or daily time-series data**. This will allow detection of declining engagement trends (e.g., sudden drop in forum activity) and enable faster response times.
* **2. A/B Testing of Interventions** Conduct **controlled experiments** to evaluate the effectiveness of nudges, mentor programs, or gamification features on different learner segments.  
   Example: Compare outcomes for learners who receive proactive mentor check-ins vs. automated messages alone.
* **3. Real-Time Dashboard Deployment** Build **live dashboards for instructors or program managers** that display engagement and risk scores by cohort. This would enable targeted outreach before dropout becomes irreversible.
* **4. Feedback Integration Loops** Create embedded survey tools or feedback forms that allow learners to **flag confusing content, emotional frustration, or progress blockers** in real time—creating a system for continuous curriculum improvement.

Churn reduction is not a one-time initiative but a **continuous process of listening, adjusting, and supporting** learners as they move through their educational journey. With robust models and strategic insights, Excelerate is well-positioned to lead the way in learner-centered, data-driven education design.

# Appendix

1. Google Collab Repository:<https://colab.research.google.com/drive/1c22vtxua1Y98P5LDO_5eZAYk1mTLoimO?usp=sharing>
2. Workflow: [Process](https://docs.google.com/document/d/1alDxUB3XHc4_0TV9dRgNvLbC6EhSOA46waqcF3UeELw/edit?usp=sharing)

1. Data was loaded, inspected, and cleaned.  
2. EngagementScore feature created from login\_days, forum\_posts, and assignment\_submissions.  
3. Dropout encoded as binary.  
4. StandardScaler applied to normalize input features.  
5. Logistic Regression and Random Forest models trained.  
6. Accuracy, Precision, Recall, F1-score calculated.  
7. Feature importance visualized.  
8. Model exported using joblib.

