Week 3 Churn Analysis Report

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

**Submission Date:** July 7, 2025

**Prepared by:**

**Lily Charlotte** - [lilycharlotte47@gmail.com](mailto:lilycharlotte47@gmail.com)

**Phanidhar Kasuba -** [phanikasuba@gmail.com](mailto:phanikasuba@gmail.com)

**Kosisochukwu Onyekwulunne -** [onyekwulunnekosi9@gmail.com](mailto:onyekwulunnekosi9@gmail.com)

**Amit Kumar -** [amitsingh6394366374@gmail.com](mailto:amitsingh6394366374@gmail.com)

**Divyesh Shinde -** [shindedivyeshavinash@gmail.com](mailto:shindedivyeshavinash@gmail.com)

**Vishal Reddy -**

### **Table of Contents**

[Table of Contents 1](#_yk6l7xgg8u6x)

[**Introduction 2**](#_qtrkb5pdie2d)

[**Data Preparation 2**](#_qn7zwir3bnb6)

[Cleaning Steps: 2](#_pg4t4xy6yuif)

[Feature Engineering: 2](#_ikl2yyhoih03)

[**Exploratory Data Analysis (EDA) 3**](#_tcozsmrxguva)

[Key Observations: 3](#_1n2a7820b3tx)

[Visual Insights: 3](#_qd8kgspxuk37)

[**Predictive Modeling 4**](#_tfunzevqlm6o)

[Models Trained 4](#_z5pn946h7yi2)

[Preprocessing & Training Approach 4](#_8dtfcbnk7j92)

[Model Performance Summary 5](#_ixnfg7jnp8u4)

[ROC Curve Comparison 5](#_1xmjvz946xvx)

[Class-wise Evaluation (Selected Models) 7](#_xjn3p5twtz2f)

[Top Predictors (from Random Forest): 7](#_do0401o710vv)

[**Churn Analysis 7**](#_vg7tqmh4kt1f)

[Key Factors Behind Dropout 7](#_w6f7m3xz8jl0)

[Observed Trends and Anomalies 8](#_nevnaj4zftwf)

[**Recommendations 9**](#_seni442lybzl)

[1. Encourage Early Engagement 9](#_1wmkmeydl7xk)

[2. Support Low-Engagement Learners 9](#_ja89ypj98et8)

[3. Align Opportunity Timing with High Engagement Periods 9](#_wlt8ml528gpu)

[4. Simplify Long-Duration Opportunities 10](#_1cuzmzuotts)

[**Conclusion and Future Work 10**](#_vbixjvd4cd0e)

[Future Directions 10](#_3ol9yq4wmmx1)

[**Appendix 11**](#_wz5slzh7kthv)

# 

# 

# Introduction

This report aims to analyze student churn using predictive modeling techniques. The goal is to identify key drivers behind student drop-off and recommend targeted strategies for improving retention. Student churn poses a significant challenge for online learning platforms, impacting learner outcomes and program success. Through data-driven insights, we aim to equip decision-makers with actionable strategies to enhance student engagement and reduce attrition.

# 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.

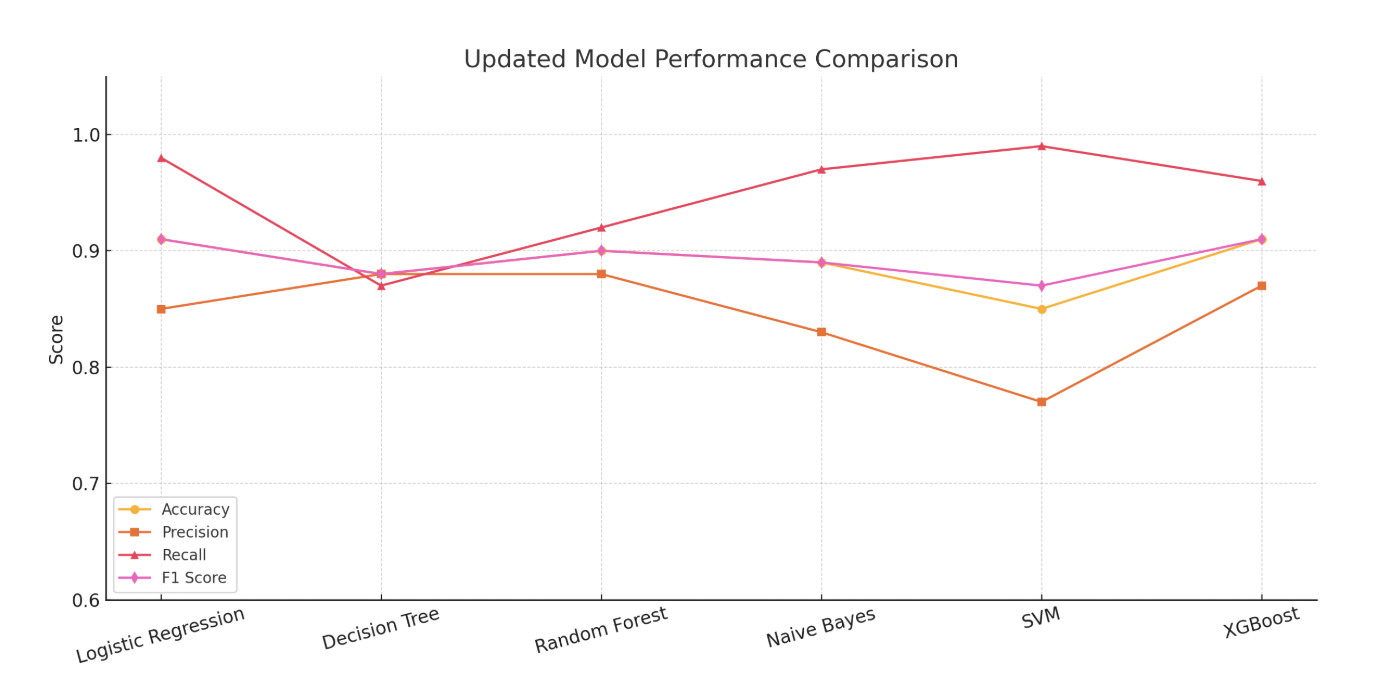
## Model Performance Summary

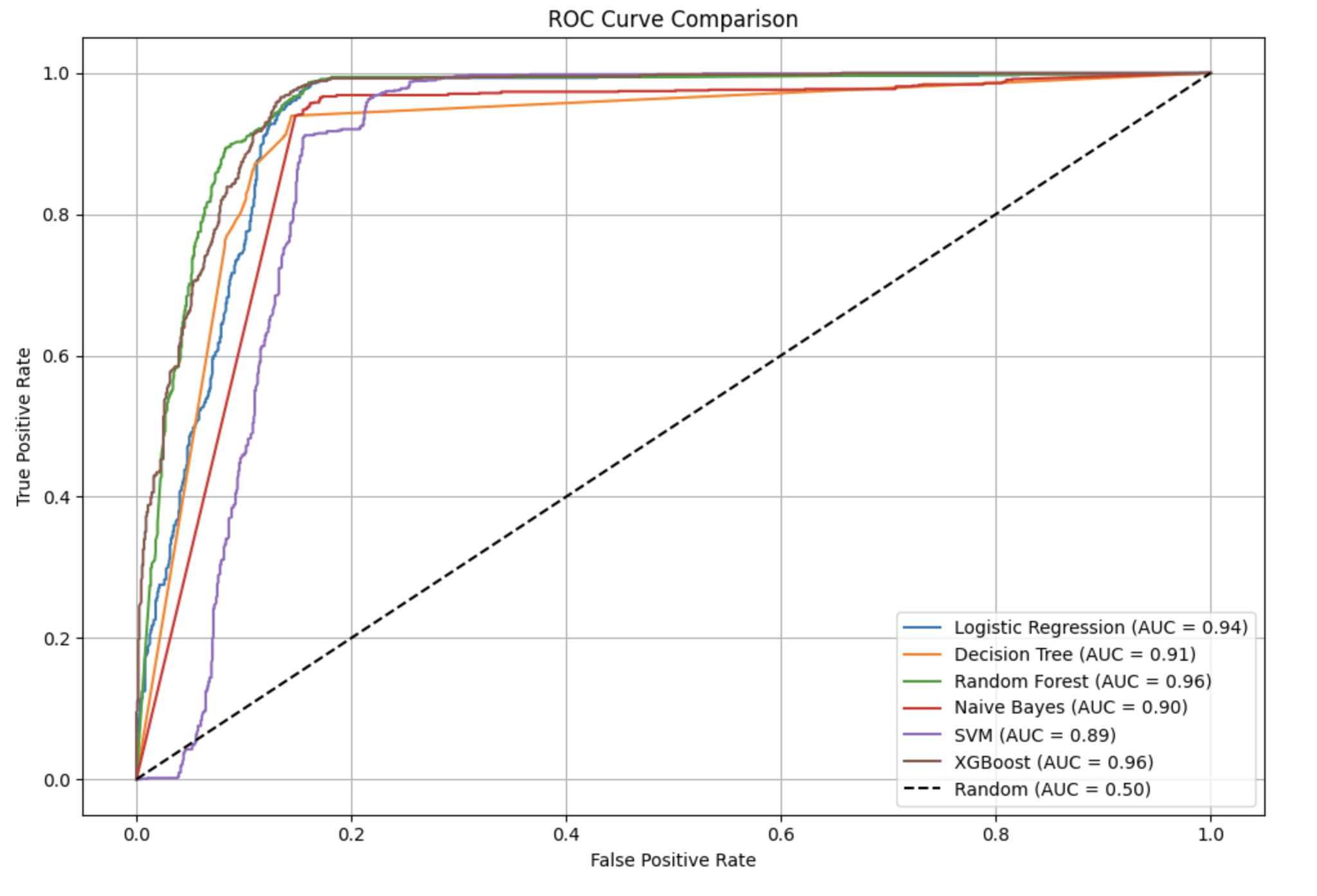
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| --- | --- | --- | --- | --- |
| **Logistic Regression** | 91% | 0.92 | 0.91 | 0.91 |
| **Random Forest** | 90% | 0.90 | 0.90 | 0.90 |
| **Decision Tree** | 88% | 0.88 | 0.88 | 0.88 |
| **Naive Bayes** | 89% | 0.90 | 0.89 | 0.89 |
| **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. Tools and Libraries: [Data Preparation, Week 1 & 2](https://github.com/Kasuba-phani/Excelerate_Intership/blob/main/Excelerate_Intership_Week_2.ipynb)  
   - pandas, numpy, matplotlib, seaborn, scikit-learn
2. Google Collab Repository:<https://colab.research.google.com/drive/1c22vtxua1Y98P5LDO_5eZAYk1mTLoimO?usp=sharing>
3. 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.

