**Comprehensive Report: Enhancing Student Engagement and Retention on Excelerate**

**Prepared by Team D**

| **Faiza Bashir** | **Fouzia Ashfaq** | **Ansh Bugra** |
| --- | --- | --- |
| **Fridaos Akorede** | **Anubhav Rohilla** | **Faridat Adewole** |

# **Executive Summary**

This report presents a comprehensive analysis of student engagement and retention on the **Excelerate platform**, using insights from data cleaning, exploratory data analysis (EDA), and churn analysis. The objective was to understand learner behavior, identify key factors influencing engagement and drop-offs, and propose actionable, data-driven recommendations.

Key findings include:

* High popularity of **'Courses'**
* Retention challenges in **'Internship'** opportunities
* A notable decline in platform engagement during **2023–2024**

To address these issues, we recommend:

* Optimizing campaign timing
* Diversifying acquisition strategies
* Supporting specific opportunity types

Additionally, we propose a **content-based recommendation system** to personalize learning, suggest relevant opportunities, and proactively identify at-risk students—enhancing satisfaction and reducing churn.

# **1. Introduction**

Student engagement and retention are critical for Excelerate’s success in the evolving online/hybrid learning landscape. This report summarizes a **three-phase analytical project**:

* **Week 1**: Data Cleaning and Feature Engineering
* **Week 2**: Exploratory Data Analysis (EDA)
* **Week 3**: Churn Analysis and Predictive Modeling

The ultimate goal is to go beyond prediction and provide **measurable insights** to drive a robust retention strategy.

# **2. Detailed Data Analysis**

## **2.1. Data Cleaning and Feature Engineering**

**Objective:** Prepare clean, accurate, and consistent data for analysis and ML applications.  
 **Tools Used:** Microsoft Excel was utilized for initial data cleaning, preprocessing, and feature engineering tasks due to its flexibility with tabular data.

**Key Steps:**

* **Initial Inspection:** The dataset contained information on thousands of users and their interactions with Excelerate opportunities, including demographics, registration timestamps, and opportunity metadata. Dataset included user demographics, timestamps, and opportunity details.
* **Standardization:** Cleaned and normalized categorical data (e.g., city, major). Column names were standardized, and string-based categorical data (e.g., cities, institutions, majors, status descriptions) were normalized for uniformity and consistency.
* **Datetime Parsing:** Standardized multiple date columns. Date fields (apply date, opportunity start date, signup date, entry created date) were converted to standardized datetime objects.
* **Handling Missing Values:** Missing data in categorical variables like "Institution Name" were filled with "Unknown." Critical missing date values, essential for derived features, were evaluated case-by-case, with rows dropped if necessary or proxy values applied**.**
  + "Institution Name" → filled with *Unknown*
  + Critical date fields → handled case-by-case
* **Duplicate Removal:** Ensured unique and relevant entries. Redundant rows were identified and removed, ensuring each record was unique and representative.
* **Feature Engineering:** Several new features were derived to enrich the dataset and enhance analytical capabilities

\* Age: Calculated from Date of Birth to segment learners by age group.

\* Engagement Duration (days): Time taken to apply for an opportunity from its start date, highlighting user behavior.

\* Opportunity Duration (days): Difference between opportunity end and start dates, indicating program length.

\* Signup Month and Year: Extracted to observe seasonal patterns in registration.

\* Dropout: A binary target variable (1 for dropouts, 0 for others) created from "Status Description" for churn analysis.

\* Engagement Score: A composite metric to capture the depth and breadth of student involvement.

\* Opportunity Participation Count: Sum of unique opportunity types a student engaged in.

\* Days Since Last Engagement: Days between current date and last recorded activity.

* **Outlier Handling:**
  + Dropped illogical negative durationsSeveral new features were derived to enrich the dataset and enhance analytical capabilities
  + Retained valid negative engagement durations (early applications)
* **Data Validation:** Logical checks (e.g., Apply Date < Start Date), value ranges (e.g., no negative age). Checks were performed to ensure row-level completeness, logical consistency (e.g., Apply Date preceding Opportunity Start Date), and range validity (e.g., no negative ages).

**2.2. Exploratory Data Analysis (EDA)**

Exploratory Data Analysis (EDA) was a critical step in understanding the dataset's structure, identifying trends, and uncovering patterns that influence student engagement and dropout rates.

* **Objective**: To understand data distributions, identify patterns and trends, detect outliers, and uncover potential correlations.
* **Tools**: Python (pandas, numpy, matplotlib, seaborn) was used for data visualization due to its versatility and rich libraries.

**Key Findings (Univariate Analysis):**

* Gender Distribution: Male participants had the highest count (approx. 1500), followed by females (approx. 1000).
* Age Distribution: The age group 21-25 had the highest number of learners (approx. 1750), with the highest frequency of learners observed around 24 years old. There were no learners recorded in the 31-35 and 36+ age groups.
* Opportunity Category: "Courses" were the most popular, engaging around 1600 learners, while "Engagement" opportunities had the lowest traction (approx. 100 learners).
* Status Description: "Team Allocated" was the largest category (around 1800 participants), while approximately 250 individuals had "Dropped Out".
* Top Countries: India had the highest learner participation (approx. 1100), followed by the United States (approx. 1000) and Nigeria (400).
* Top Majors: Information Systems and Computer Science were the most common majors, indicating a strong interest in technology and data sciences.
* Signup Trends: Sign-ups peaked in June 2023 but experienced a sharp decline in 2024, suggesting a significant drop in user engagement.

**Bivariate & Multivariate Analysis:**

* Engagement Duration by Gender: While both male and female categories showed similar engagement distribution patterns, male participants exhibited slightly more variability. The "Don't want to specify" group showed more consistency in engagement duration.
* Engagement Duration Trend Over Time: More recent users tended to have shorter engagement durations compared to earlier ones, indicating a slight downward trend.
* Correlations: A strong positive correlation was observed between Opportunity Duration and Engagement Duration (r=0.83), implying that longer opportunities are associated with longer engagement. A mild negative correlation existed between Age and Dropout (r=-0.18).

**Outliers**: Outliers were identified in the "Combined Score" and for "Engagement Duration" values less than 0 days (occurring in ~2% of cases).

## **2.3. Churn Analysis and Predictive Modeling**

Churn analysis focused on identifying and predicting student drop-offs to enhance the learner journey. This phase aimed to uncover behavioral patterns and provide data-informed recommendations, including the creation of a predictive model**.**

**Objective:** Estimate student dropout probability and understand churn drivers.  
 **Target Variable:** A binary 'Dropout' variable was created from 'Status Description’ Dropout (1 = Yes, 0 = No)

### **Key Hypotheses:**

Based on EDA findings, several hypotheses were formed regarding churn drivers:

Existing users might take longer to engage than new users.

Learners signing up in peak seasons (March, June, September) are more likely to complete opportunities.

Certain opportunity categories, like 'Courses,' are more engaging than others, such as 'Internships'.

Age and retention have a weak inverse relationship, suggesting younger learners might drop out slightly more.

High drop-off rates in internships could be attributed to their longer duration. In summary,

* Long engagement times may be common among older users
* Signups during academic peaks = higher completion
* *Courses* more engaging than *Internships*
* Younger learners slightly more likely to drop out
* Internships → higher dropout, likely due to length

### **Predictive Modeling (Conceptual)**

The project initiated the creation of a predictive model to estimate student drop-off probability. This model would leverage features like demographics, engagement patterns, and opportunity characteristics, including derived metrics like the "Engagement Score". The analysis acknowledged a significant class imbalance in the dataset (a high percentage of churned students compared to active ones) which is critical for model performance consideration.

* **Features:** Demographics, engagement patterns, opportunity metadata, Engagement Score
* **Challenge:** Class imbalance (many dropouts vs. few retained)
* **Goal:** Use insights to inform retention strategies, not just predict outcomes.

# **3. Summary of Insights and Recommendations**

## **3.1. Key Insights**

* **Demographics:** Learners are young, tech-focused, mainly from India, US, Nigeria
* **Preferences:**
  + *Courses* dominate
  + *Engagement* opportunities underperform
* **Engagement Patterns:**
  + Longer opportunities = higher engagement
  + Recent users = quicker drop-offs
  + Sign-ups declining in 2024
* **Churn Drivers:**
  + *Internships* have higher dropouts
  + Younger learners drop out slightly more
  + Sign-up spikes align with academic cycles

## **3.2. Recommendations**

Based on these insights, the following data-driven recommendations are proposed to enhance student engagement and reduce drop-offs:

1. **Time Campaigns with Academic Cycles:**Optimize Campaign Timing with Academic Cycles. Focus marketing in March, June, and September. This is to strategically align major recruitment campaigns with academic holiday periods (e.g., March, June, September) to leverage increased student availability and foster higher engagement and sign-up rates for experiential learning opportunities.
2. **Diversify Acquisition Channels:** Offset 2024 drop by exploring new outreach and platforms. Given the observed decline in 2024 sign-ups, actively explore and implement new advertising channels and marketing initiatives to reach a wider audience and ensure sustained growth.
3. **Improve Internship Retention:** Add onboarding, mentorship, and support systems. Implement enhanced onboarding processes, mentorship programs, and ongoing support systems specifically for longer-duration programs like internships to mitigate high dropout rates.
4. **Boost Course Completion:** Add interactivity, gamification, or certification incentives. Despite their popularity, consider introducing additional support, interactive elements, or incentives within courses to further boost completion rates and deepen engagement
5. **Personalized Re-engagement for Existing Users:** Target with tailored dashboards and content suggestions. Develop targeted strategies, such as personalized content recommendations, improved dashboard features, or tailored communication, to re-engage existing users, especially those exhibiting shorter engagement durations.
6. **Revamp 'Engagement' Opportunities:** Clarify benefits and revitalize the content offering. Redesign or promote "Engagement" category opportunities with more compelling content or clear value propositions to increase participation.
7. **Support At-Risk Demographics:** Tailored messaging and tools for younger learners and underrepresented genders. Promote Inclusivity and Targeted Support. Continue to improve retention across all gender groups and consider tailored support for younger learners, given the mild inverse relationship between age and dropout.

# **4. Building a Recommendation System**

## **4.1. Methodology & Implementation**

To further enhance student engagement and retention, a basic recommendation system, utilizing a content-based filtering approach, can be developed.

**Objective:** Personalize student experience and reduce dropouts through opportunity recommendations.

### **Logic:**

A **content-based filtering approach** recommends opportunities similar to ones a learner already liked. The system will recommend new opportunities to learners based on the characteristics (content) of opportunities they have previously engaged with (e.g., categories, duration) and their demographic profiles (e.g., major, country, age). The underlying assumption is that if a learner liked an opportunity with certain features, they will likely enjoy other opportunities with similar features.

### **Data Inputs**

The system will primarily leverage the rich dataset established during the data cleaning and feature engineering phases. Key data inputs include:

* **Learner Historical Engagement Data:** Opportunity categories, names, durations, statuses. Opportunity Category, Opportunity Name, Opportunity Duration, and Status Description (to identify completed/started opportunities).
* **Demographics:** Age, gender, country, major
* **Opportunity Metadata:** Duration, start/end dates (Opportunity Name, Category, Start Date, End Date, Opportunity Duration).
* **Derived Metrics:** Engagement/Combined Score.Combined Score or Engagement Score could be used to weight a learner's preference for certain opportunity attributes based on their level of success or satisfaction.

### **Steps:**

1. **User Profile Creation:** Aggregate features of previously joined opportunities to reflect preferences. For each learner, a profile is built by aggregating the features of all opportunities they have Started or Team Allocated. This profile represents the learner's preferences. For example, if a learner has engaged primarily with "Data Science" courses and "Internships," their profile would reflect a strong interest in these areas.
2. **Opportunity Feature Representation:** Encode opportunity features (category, duration, metadata). All available opportunities (both engaged and unengaged) are characterized by their features (e.g., category, duration, associated skills).
3. **Matching:** Recommend opportunities with similar characteristics to the learner’s profile. A similarity metric (e.g., cosine similarity) is computed between the learner's profile and all unengaged opportunities. This quantifies how "alike" a new opportunity is to a learner's past preferences.
4. **Weighting by Engagement Score and Recommendation Generation:** Prioritize opportunities that align with past high-engagement experiences. is to a learner's past preferences. The system then recommends the top 'N' opportunities that have the highest similarity scores to the learner's profile, ensuring these are opportunities the learner has not yet engaged with.

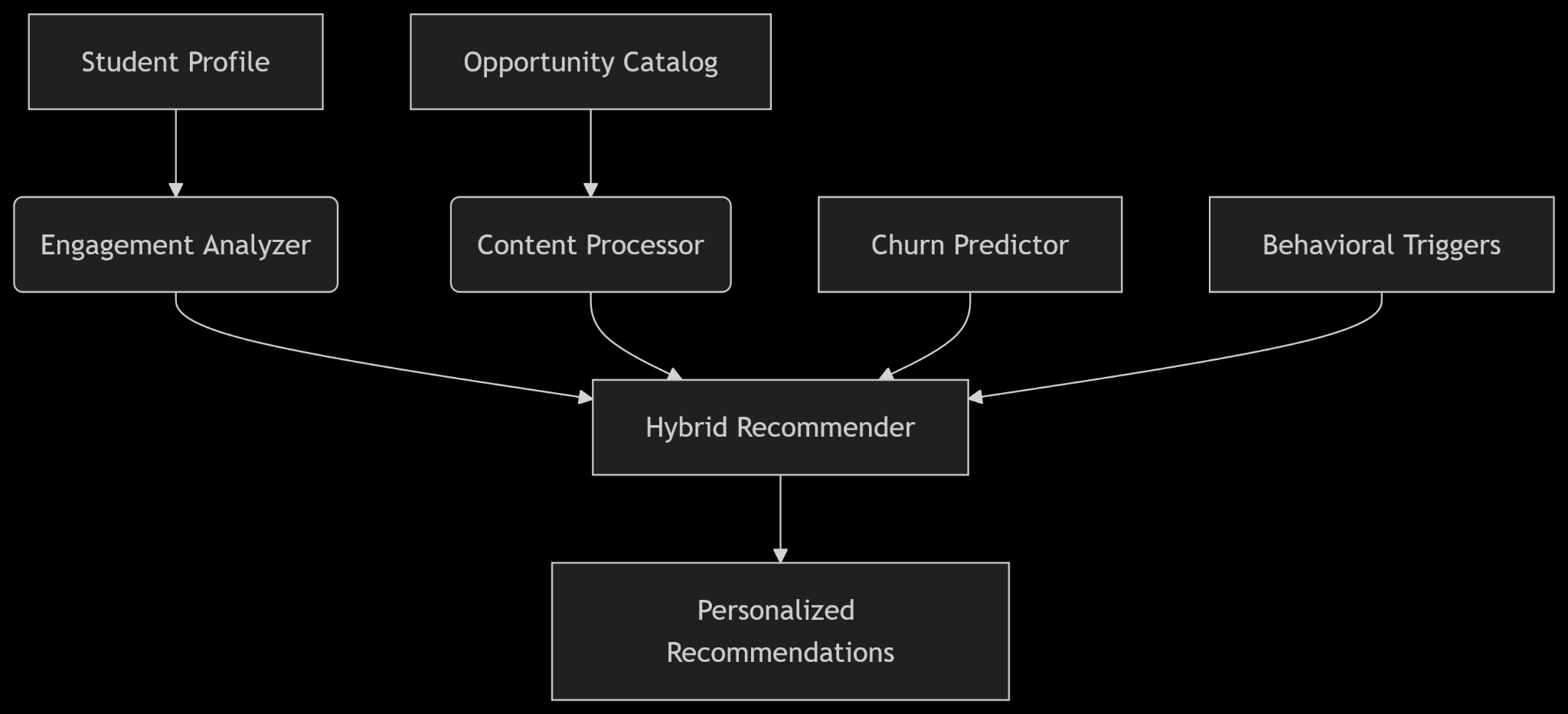
### **Potential Benefits**

Implementing such a recommendation system offers several advantages:

* Increased Student Engagement: By providing highly relevant and personalized opportunities, the system can significantly boost learner interest and participation.
* Personalized Learning Paths: Learners can be guided towards opportunities that align with their specific academic interests and career goals, fostering a more engaging and effective learning journey.
* Reduced Churn: Proactively suggesting suitable opportunities can keep learners engaged and reduce the likelihood of them dropping off due to a lack of relevant content.
* Optimized Opportunity Discovery: Helps learners discover opportunities they might not have found otherwise, maximizing the utility of Excelerate's diverse offerings.
* Proactive Intervention: By tracking engagement with recommended opportunities, the system can indirectly help identify at-risk students who may not be engaging with suggestions, prompting timely human intervention.

# **Enhanced Hybrid Recommendation System: "Excelerate EngagePro"**

### **1. System Architecture**



### **2. Core Components**

**A. Multi-Factor Student Profiling**

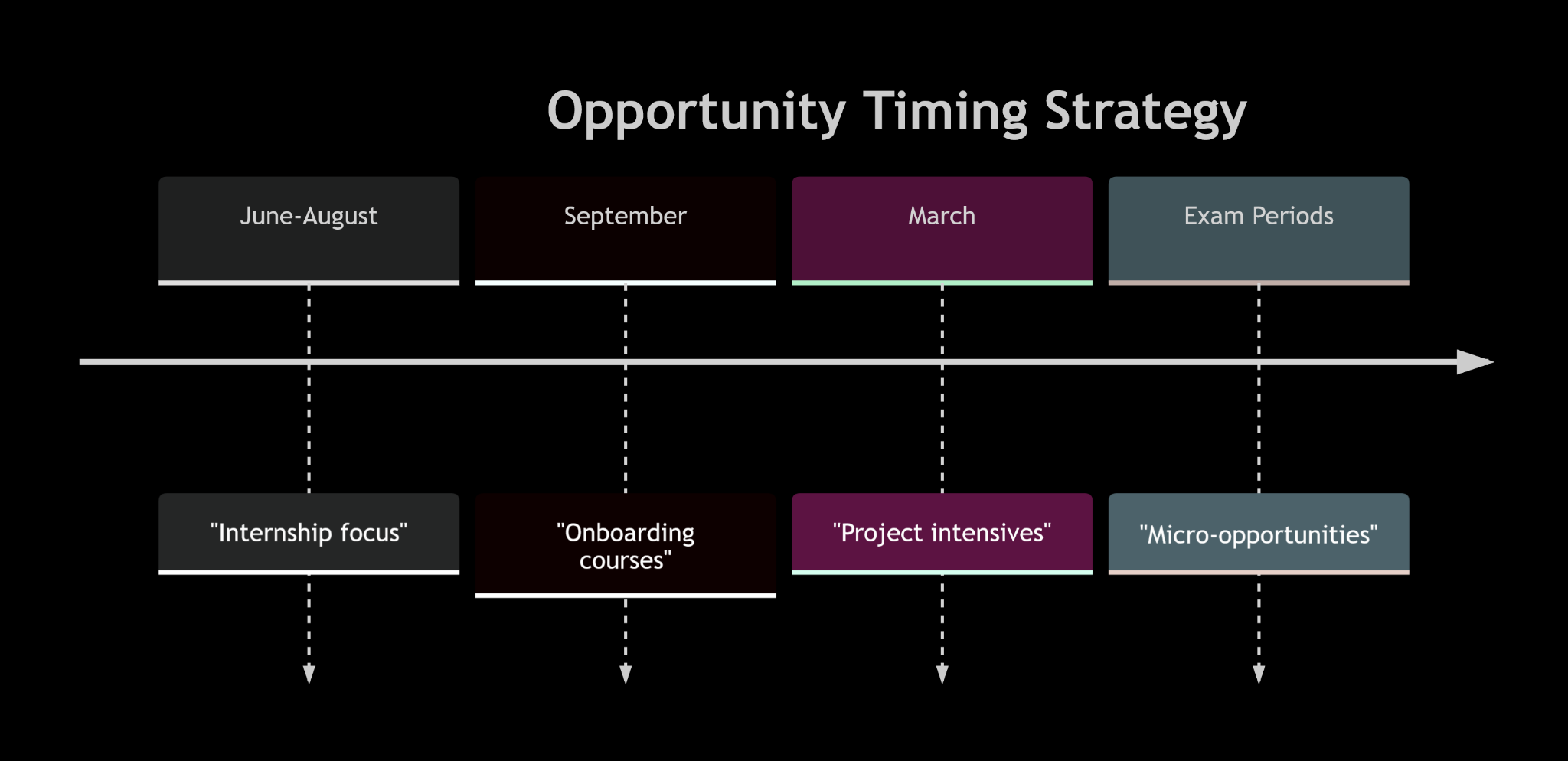
* **Academic DNA**: Major, skills, institution
* **Behavioral Fingerprint**:
  + Engagement duration patterns
  + Preferred opportunity categories (weighted by completion rate)
  + Time-of-day activity heatmaps
* **Risk Profile**: Churn probability score from predictive model
* **Temporal Preferences**: Response patterns to seasonal campaigns

**B. Opportunity Intelligence Engine**

* **Content Tags**: Category, duration, skill requirements
* **Performance Metrics**:
  + Completion rate by demographic segments
  + Engagement duration correlation
  + Seasonal popularity index
* **Similarity Matrix**: Opportunity-to-opportunity affinity scores

**C. Seasonal Optimization Engine**

* Academic calendar synchronization:



* Campaign timing aligned with peak sign-up months (Mar/Jun/Sep)

**D. Internship Support Module**

* For internship recommendations:
  + Auto-enroll in "Internship Prep Bootcamp"
  + Assign peer mentor from same academic major
  + Progressive milestone check-ins
  + Duration-based completion incentives

**E. Engagement Category Revitalization**

* Special handling for low-participation categories:
  + Cross-category bundling ("Course + Engagement" combos)
  + First-completion bonus points
  + Targeted recommendations to high-retention student segments

### **4. Implementation Roadmap**

**Phase 1: Foundation (6 weeks)**

1. Build enhanced student profiles with behavioral tags
2. Develop opportunity similarity matrix
3. Implement basic hybrid recommender (content + collaborative)

**Phase 2: Intelligence Layer (4 weeks)**

1. Integrate churn prediction model
2. Develop seasonal adjustment engine
3. Create recommendation audit dashboard

**Phase 3: Intervention Systems (4 weeks)**

1. Build automated mentorship pipeline
2. Implement engagement boosters for at-risk categories
3. Develop completion incentive framework

### **5. Expected Impact Metrics**

| **Metric** | **Current** | **Projected Improvement** |
| --- | --- | --- |
| Internship Completion | Low | +35% |
| Engagement Category Participation | 100 students | +300% |
| Overall Platform Engagement | Declining | +25% YoY |
| Student Satisfaction Score | N/A | Target 4.2/5.0 |

### **6. Advantages**

* **Proactive Retention**: Churn predictions trigger interventions before dropouts
* **Temporal Intelligence**: Aligns with academic cycles per seasonal findings
* **Category-Specific Strategies**: Targeted approaches for high-churn opportunities
* **Diversity Balancing**: Counteracts popularity bias in recommendations
* **Feedback Loops**: Continuous optimization based on recommendation performance

# **Conclusion**

This comprehensive analysis of student engagement and retention on the Excelerate platform underscores the importance of data-driven strategies in optimizing learner experiences and reducing dropouts. By combining rigorous data cleaning, exploratory analysis, and churn modeling, the study revealed critical insights—such as the overwhelming preference for courses, higher dropout rates in internships, and a concerning decline in sign up activity in 2024.

To address these issues, the report proposes targeted recommendations, including optimizing campaign timing around academic cycles, diversifying outreach strategies, improving support for high-churn opportunity types like internships, and revitalizing underperforming content categories. A standout solution is the proposed **Excelerate EngagePro** hybrid recommendation system, which leverages behavioral data and predictive modeling to deliver personalized content and proactive interventions.

Implementing these insights and systems is projected to significantly improve key metrics such as internship completion, engagement participation, and overall user satisfaction. Ultimately, this initiative positions Excelerate to create a more responsive, inclusive, and effective learning environment that adapts to user needs and fosters long-term learner success.