# AI Engagement Data Analyst Internship
# Excelerate – Team 7


**Sumaiya**
Team Lead
sumaiyaa.tasnim.18@gmail.com

**Saniya**
Project Scribe
dantalsb04@gmail.com

**Zaheer**
Project Manager
mianzaheer4195@gmail.com

**Devadharshini**
Project Lead
dharshinideva83@gmail.com

**Peter**
Team Member
mpeter778@gmail.com

**Hamsini**
Team Member
bokkahamsini@gmail.com

# The Mission

Assigned Raw Dataset: "SLU_Opportunity_Wise_Dataset.csv"

## Week 1

### Data Cleaning

Handling missing values, outliers, inconsistencies & Feature Engineering

## Week 2

### EDA

Data exploration, visualization, and analysis

## Week 3

### Modeling

Predictive modeling and feature importance

## Week 4

### Insights

Actionable recommendations and solutions

# Week 1 Data Cleaning in Action

| Issue Type | Before | After |
| --- | --- | --- |
| **Missing Dates** | Empty cells, corrupted formats (Excel serial numbers) | Filled with most frequent date, standardized datetime format |
| **Inconsistent Names** | "st. louis", "Nwihs", "xxxhhyy" | "Saint Louis University", "Northwest Institute of Health Sciences", removed invalid entries |
| **Corrupted Formats** | Mixed date formats, inconsistent text cases | Uniform datetime64 format, standardized text cases |
| **Duplicates** | Duplicate rows, duplicate Opportunity Id + Learner SignUp DateTime | Removed duplicates safely while keeping first occurrence |

**Result: A clean, reliable dataset for analysis**

# Week 1 Feature Engineering

## Age

Calculated from Date of Birth to analyze demographic patterns and engagement across age groups

## Engagement Days

Time lag between application and opportunity start to identify quick vs delayed engagement

## Engagement Score

Weighted composite score combining duration, age, and interaction features for holistic measurement

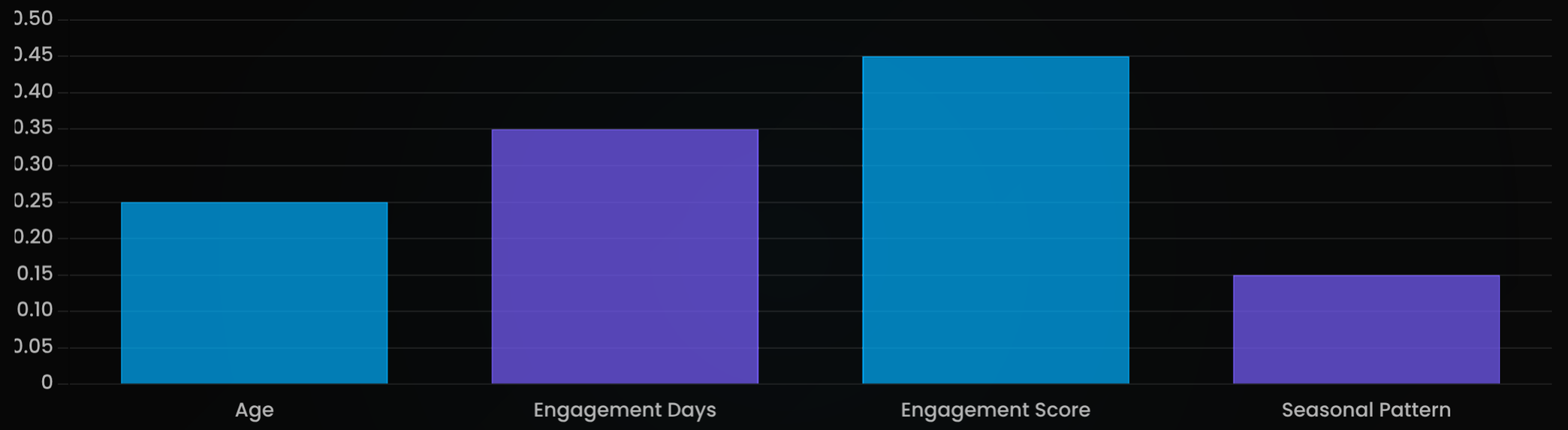
## Seasonal Pattern

Grouped months into seasons to identify seasonal peaks/troughs in engagement

We also created **additional features** to enhance dataset analysis

# Week 1: Engineering Predictive Features

## Feature Impact Analysis




### Problem

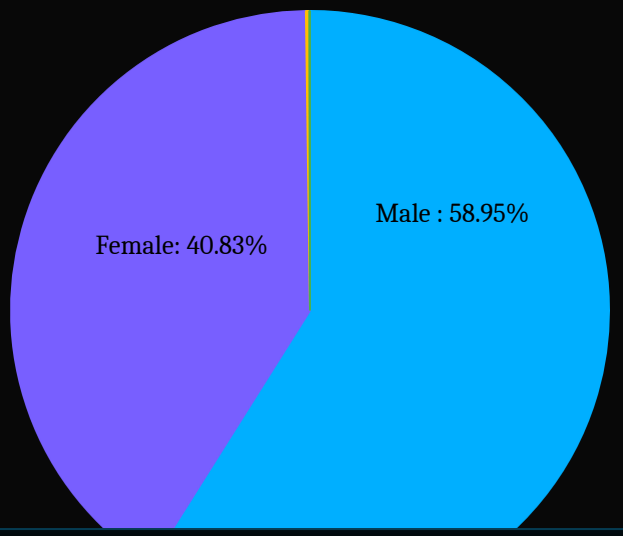
Limited feature set restricts insights

### Solution

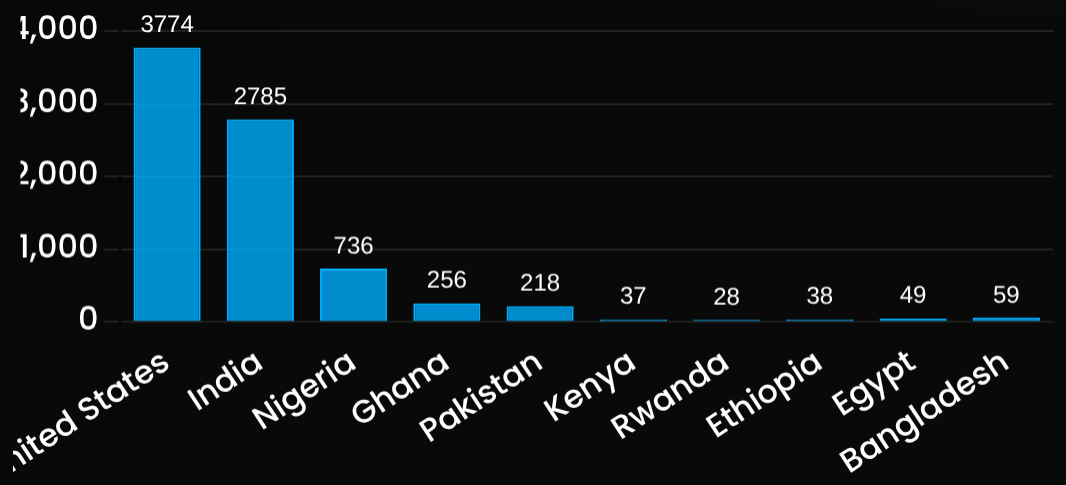
Engineered features like Engagement Score uncover hidden learner behavior drivers

# Week 2 EDA Overview (Who Are Our Learners?)

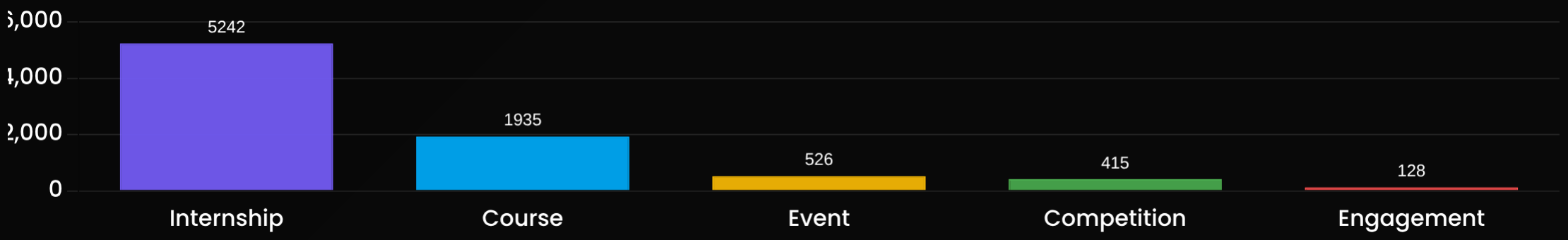
## Gender Distribution




## Top 10 Countries by Count




## Opportunity by Category Counts



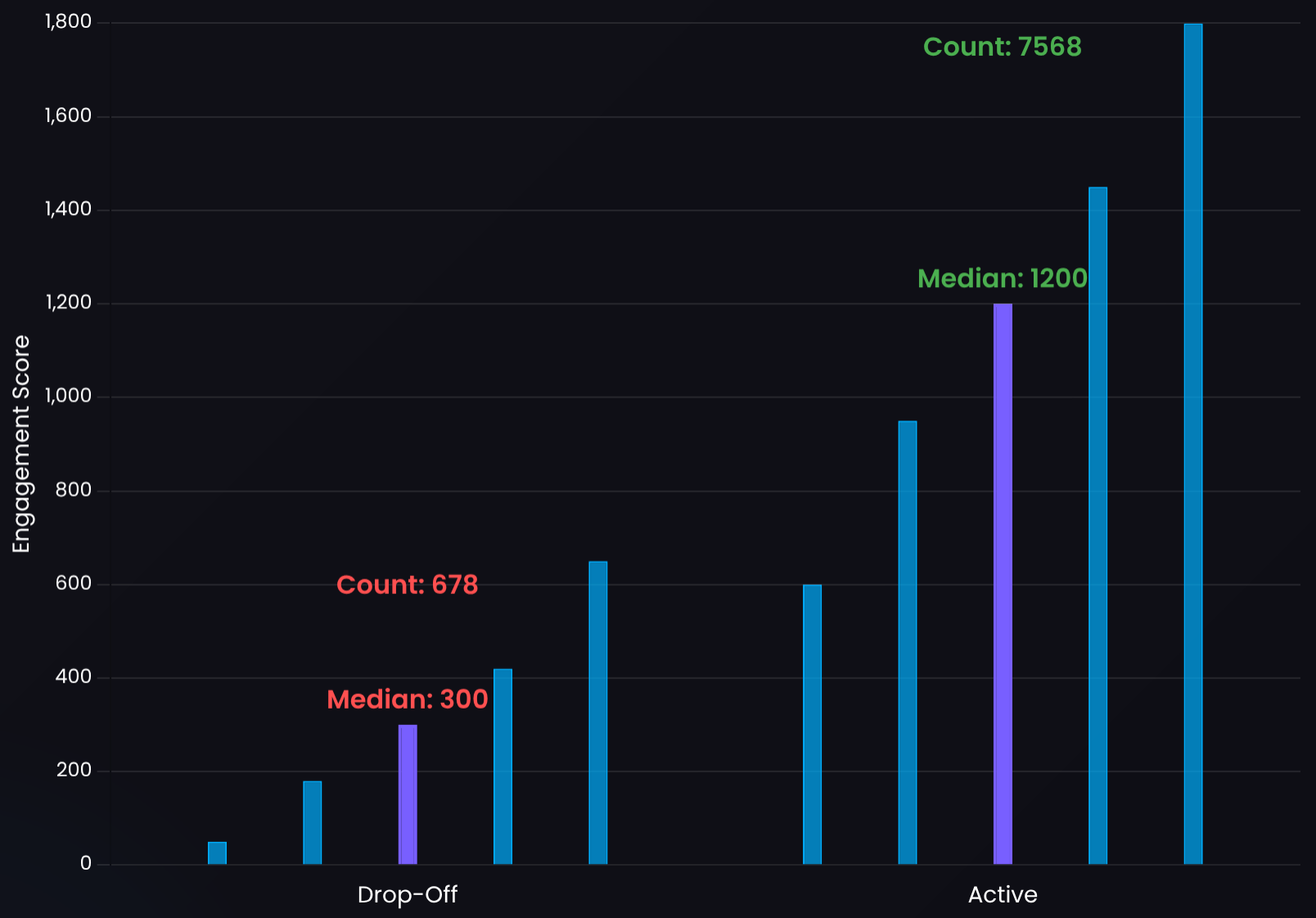

## Business Insights

Learners want jobs (63% seek internships). Our top countries are key growth markets. Focus on internships & courses; use events to attract users.

# Week 2 Engagement Story

Engagement distribution based on Week 2 dataset

## Engagement Score by Status Description




### Key Insights

"Dropped Out" = Low engagement
"Active" = High engagement

**300** Drop-off median score

**1200** Active median score

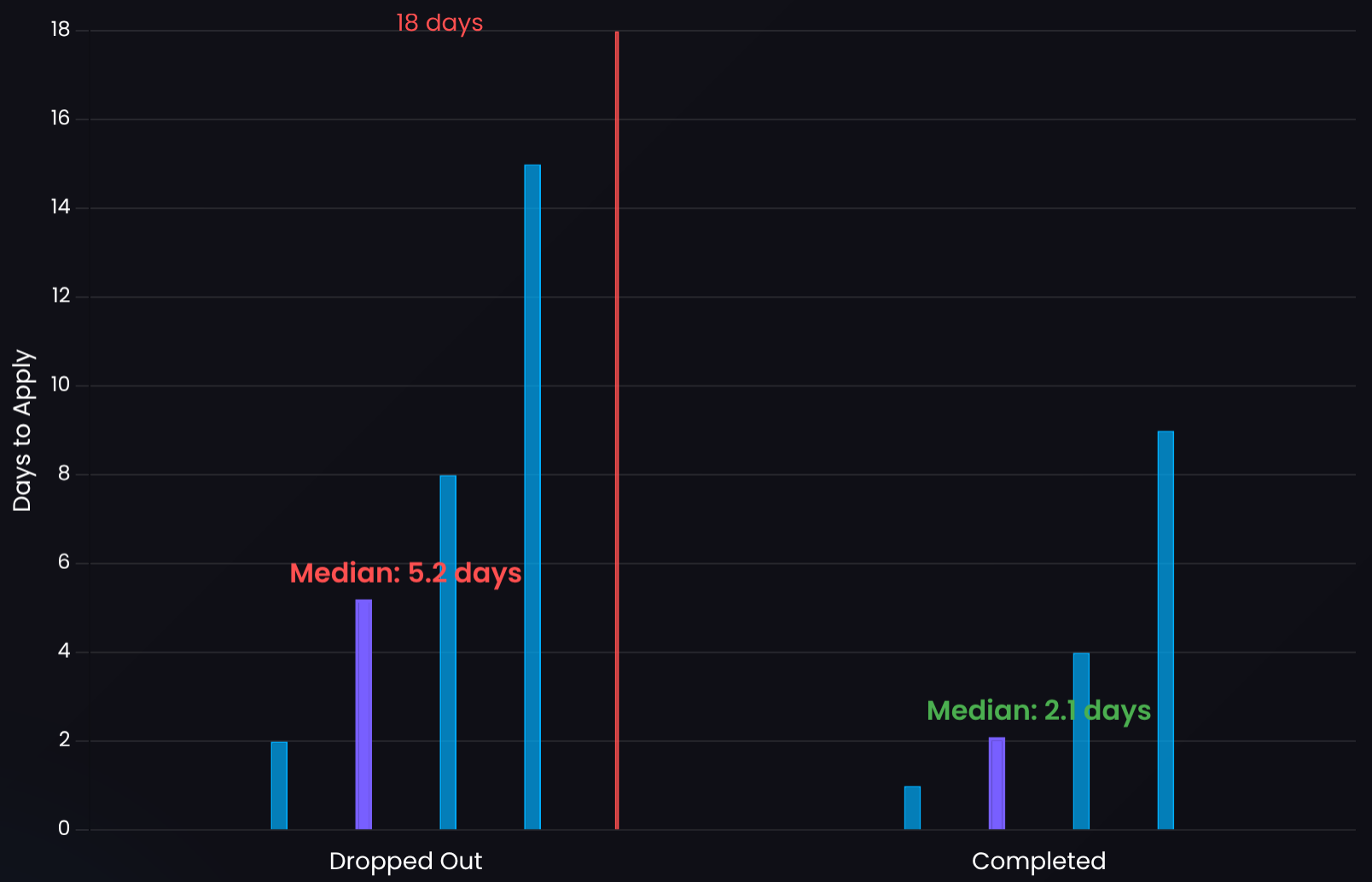
### Status Distribution

**Active / Non-Drop-Off:** 7,568 learners (91.8%)
**Drop-Off:** 678 learners (8.2%)

Includes 596 Dropped Out + 82 Withdrawn

# Week 2 Power of Timing

## Time to Apply by Completion Status




### Key Finding

Distributions clustered near **0 days**

**5.2** Drop-off median days

**2.1** Completion median days

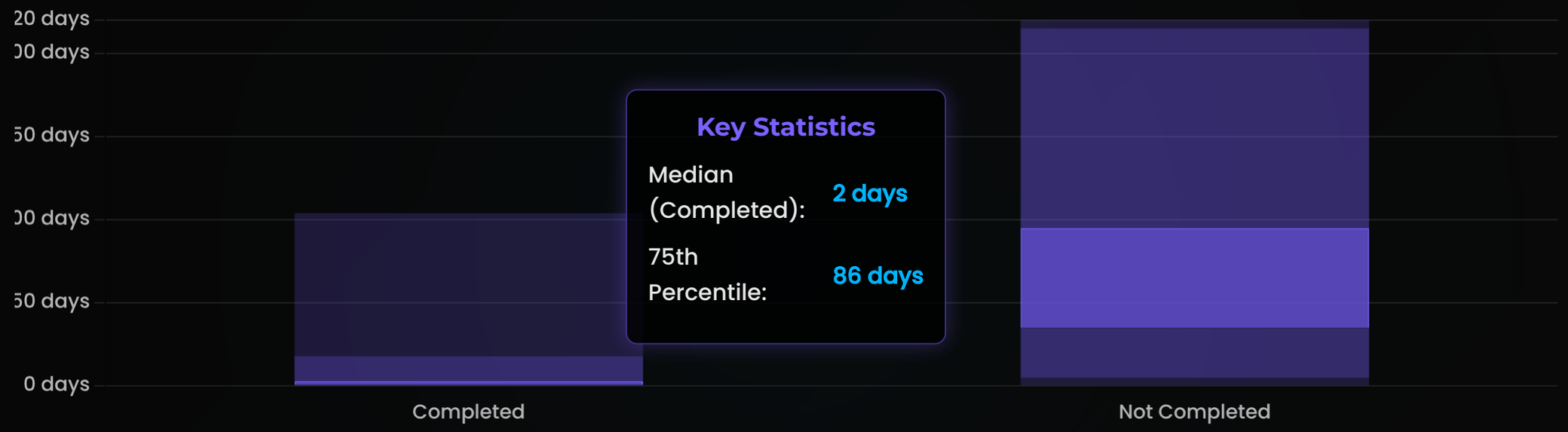
### Outliers

Significant outliers in drop-off group
Some applicants waited **30+ days** before applying
Successful applicants rarely exceeded **10 days**

**Late applicants** drop out — **early action** drives success

# Week 2: The Critical Application Window

## Time to Apply vs. Completion Status




### Problem

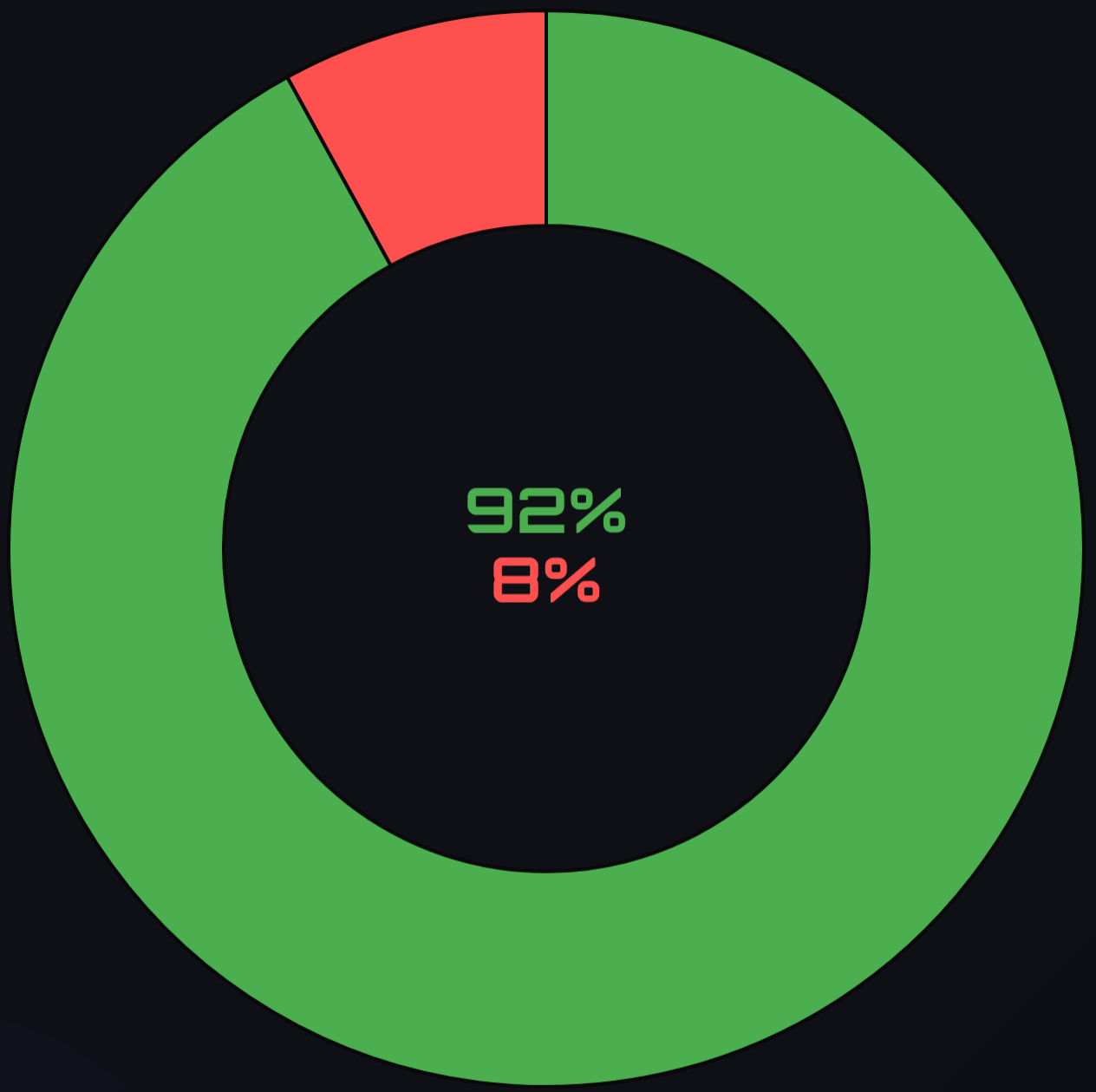
Delayed applications correlate with drop-outs

### Solution

Encourage immediate application to boost completion rates

# Week 3 Prediction Challenge

## Completion vs Drop-off Distribution




## ⚠ Class Imbalance Challenge

Dataset shows significant **imbalance** between completion and drop-off rates
Standard ML models may **bias** toward majority class
Need specialized techniques to **balance** the dataset

92% Complete

8% Drop-off

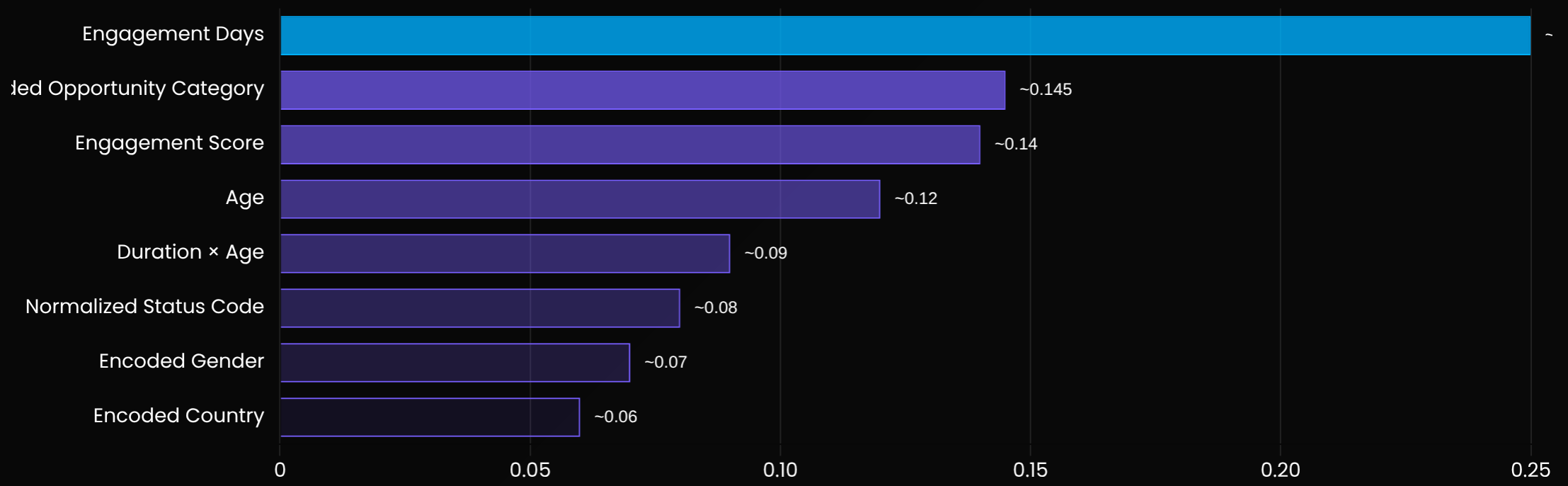
## Solution Approach

Applying **SMOTE** (Synthetic Minority Over-sampling Technique)
Creates synthetic samples of minority class
Balances dataset without losing information

**SMOTE**

# Week 3 Why Students Leave

## Feature Importance (Random Forest + SMOTE Model)




### Engagement Days

~0.25

### Encoded Opportunity Category

~0.145

### Engagement Score

~0.14

**Engagement Days** is the most critical factor in predicting drop-offs, followed by **Opportunity Category** and **Engagement Score**

# Week 3: Model Showdown

## Model Performance Comparison (Recall %)

| Model | Recall (%) |
|---|---|
| Logistic Regression | 72.1% |
| Random Forest | 30.1% |
| Random Forest + SMOTE | 46.3% |
| Random Forest + SMOTE + Tuning Best Model | 98.1% |
| Decision Tree | 38.24% |

**Problem**

Low recall for identifying drop-offs

**Solution**

Tuned model achieves 98.1% recall for at-risk learners

# Week 3 High-Risk Radar

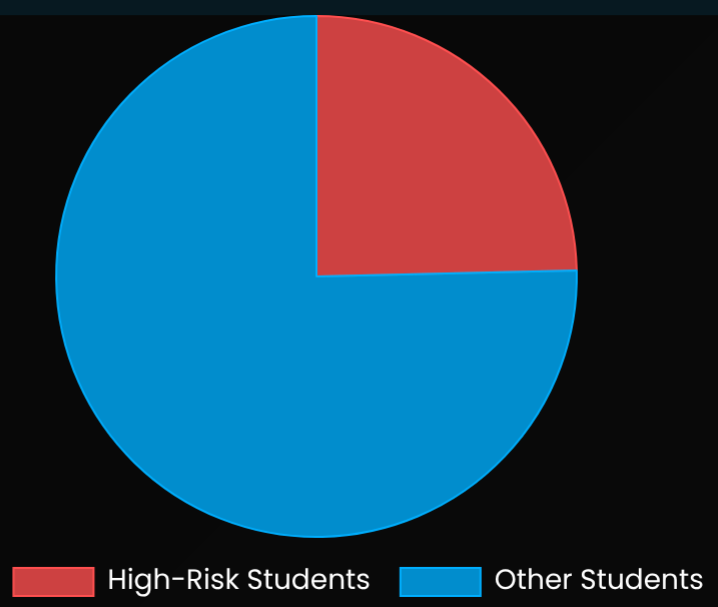
## High-Risk Students

Number of High-Risk Students

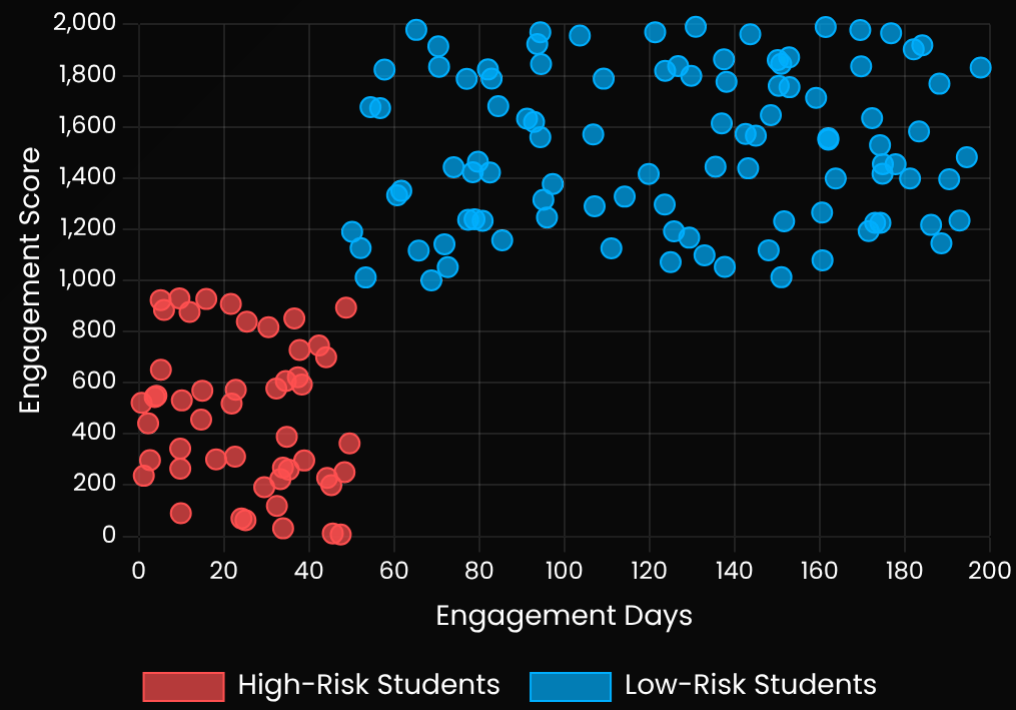
**2,032**

Proportion of Total

**24.64%**




## Risk Thresholds




**Engagement Days**

< 50 days

**Engagement Score**

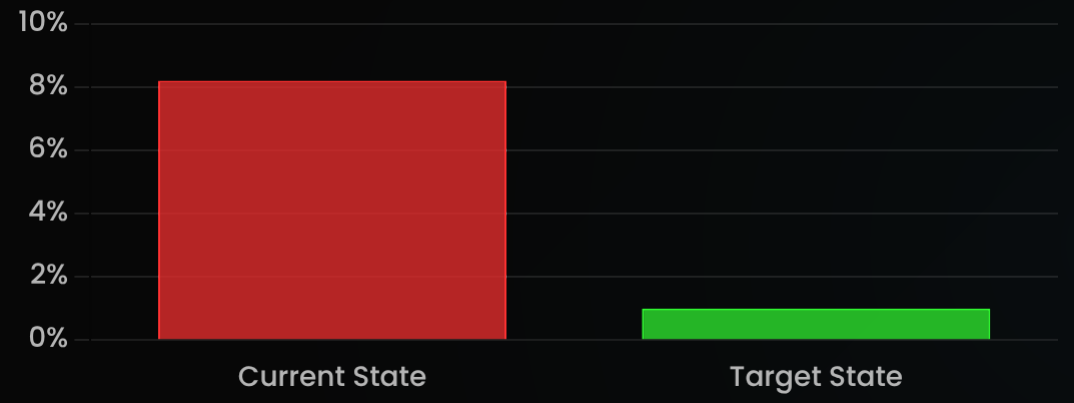
< 1000

**Students with low engagement days and low engagement scores are at significantly higher risk of dropping out, representing nearly 25% of all learners**

# Week 4: The Strategic Blueprint for Retention

## From Insights to Action: A 4-Point Plan to Reduce Drop-Offs

### Drop-Off Rate: Current vs Target




**88% Reduction Potential** ↘

**Problem**

High drop-off rate threatens success

**Solution**

Engagement boost could reduce drop-offs by 88%

### Real-Time Alerts & Dashboard

**Problem:** At-risk students are invisible

**Solution:** Live dashboard flags students below thresholds

### Proactive Early Outreach

**Problem:** Late applicants drop out

**Solution:** Automated comms within first week

### Optimized Opportunity Catalog

**Problem:** "Internship" category has 3,447 rejections

**Solution:** Refine requirements; promote high-completion opportunities

### Personalized Support Pathways

**Problem:** Generic support is ineffective

**Solution:** Tailored resources for high-risk learners

# Conclusion

Our project turned raw engagement data into **actionable insights** to improve learner retention.
Using data cleaning, feature engineering, EDA, and predictive modeling, we identified key drop-off factors.

The tuned Random Forest model with SMOTE achieved **98.1% recall**, enabling early detection of at-risk learners.

Our four-point strategy—focused on early outreach, interactive visualizations, optimized opportunities, and personalized support—can reduce drop-offs by up to **88%**.

This internship strengthened our ability to apply data science for practical, real-world solutions.

# AI Engagement Data Analyst Internship – Excelerate

We deeply appreciate the opportunity to contribute to the **Excelerate AI Engagement Internship**. The experience refined our data analytics skills and taught us how to translate data into **actionable strategies**.

Data Analysis

Predictive Modeling

Strategic Solutions

# Questions?