Excelerate

Excelerate AI-Powered Virtual Internship

May - June 2025

Week 3 : Churn Analysis Report

Date of Submission: June 2 , 2025

Team E Members

Sr NO.	Name	Designation	Email
1	Diya Kharel	Team Lead	diyakharel4@gmail.com
2	Iqra Shaikh	Project Lead	sh.iqratasleem@gmail.com
3	Faith Odhe	Team Member	Faith.t.odhe@gmail.com

Table of Content

1. Introduction

The primary objective of this analysis is to understand why students/learners are "churning" (i.e. dropping out) from your internship-platform and to build a predictive model that identifies at-risk students before they leave. By pinpointing the leading drivers of dropout, we can recommend targeted interventions such as personalized nudges, tailored content, or mentorship to boost retention and ensure that more learners successfully complete their internships or coursework.

2. Data Preparation

2.1. Overview of Raw Dataset

Total records: 8,558 learners

1. Total Columns : 24 original fields in total.

2. Timestamp Fields

- Learner SignUp DateTime : Timestamp when the student signed up.
- Opportunity End Date :When the internship opportunity ends.
- Apply Date : When the student applied to the opportunity.

3. Demographic Attributes

- Age : Age of the learner.
- Gender : Gender of the learner.
- Country : Country of residence.

4. Opportunity Metadata

- Opportunity Category : Type/category of the opportunity (e.g., Internship, Course).

5. Application Data

- Apply Date : Date the learner applied.
- Status Description : Current status of the learner (e.g., Active, Dropped Out).

6. Derived Features

- Engagement Duration : Time (in days) the learner actively engaged with the platform.
- Time in Opportunity : Number of days between the learner's sign-up and the start of the opportunity.

2.1.1. Target Variable

We defined a binary churn indicator as:

Churn = 1 if Status Description = "Dropped Out"

Churn = 0 otherwise (including "Rejected," "Team Allocated," "Started," "Waitlisted," etc.)

Status Description	Count	Churn Flag = 1
Rejected	3,569	0
Team Allocated	3,276	0
Started	767	0
Dropped Out	617	**1**
Waitlisted	109	0
Applied	105	0
Withdraw	86	0
Rewards Award	29	0

Churn rate (Dropped Out ÷ Total): 617 / 8,558 ≈ 7.2%

2.2. Data Cleaning

1. Trailing-space column name:

The column “Engagement Duration ” had a trailing space. We renamed it to “Engagement Duration” so that it could be referenced consistently.

2. Missing Values:

All 8,558 rows are non-null across every column. No imputations were needed.

3. Duplicates:

We checked for duplicate Opportunity Id + Learner SignUp DateTime combinations and found none.

4. Outliers:

A few “Engagement Duration” values were negative (e.g. –500 days). Negative values arose from logging mismatches when learners briefly accessed the platform prior to the official “Opportunity Start Date.” We retained these as it is but noted their presence during EDA.

“Age” ranged from 15 to 60, with no apparent data entry errors.

2.3. Feature Selection & Encoding

Numeric features (all continuous/integer):

1. Engagement Duration (days on platform)

2. Time in Opportunity (days between sign-up and opportunity start)

3. Age

4. SignUp Year

5. SignUp Month

Categorical features (encoded via One-Hot Encoder):

1. Gender (Female / Male / Other)

2. Country (50+ unique countries; we one-hot-encoded all, enabling the model to learn country-specific retention patterns)

3. Opportunity Category ("Course" vs. "Internship")

Derived target: Churn (binary 0/1 as described above)

3. Exploratory Data Analysis (EDA)

3.1. Descriptive Statistics of Key Numeric Features

Below are summary statistics for the primary numeric features, shown separately for churned (Churn=1) vs. non-churned (Churn=0) learners.

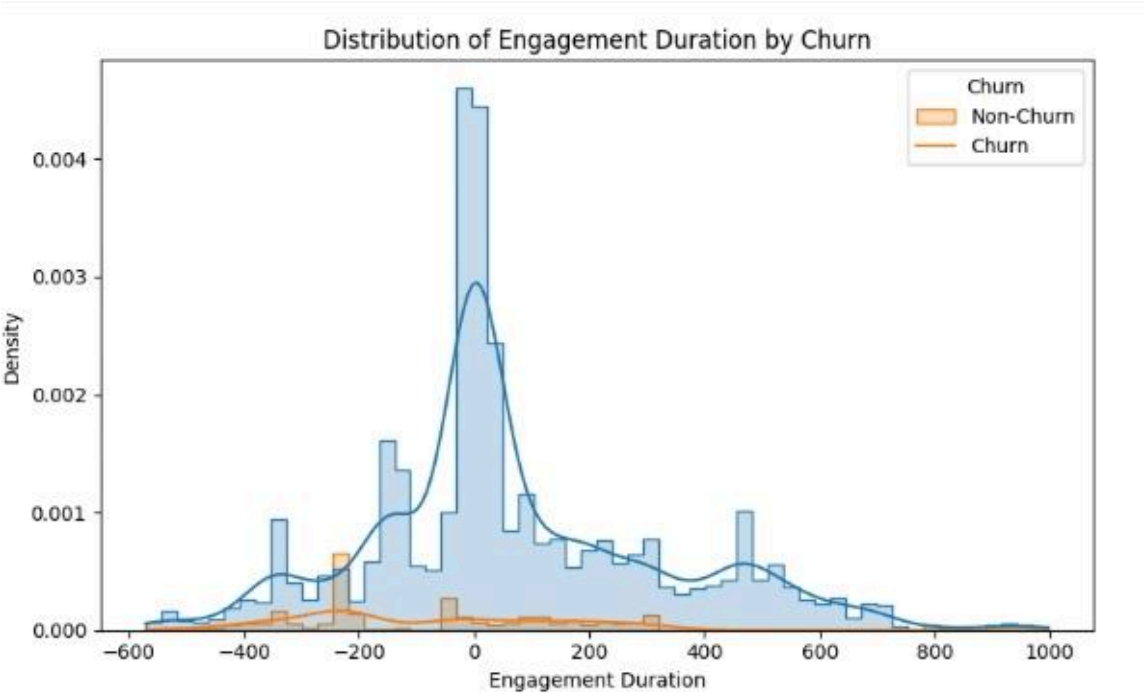
Feature	Churn=0 (Mean ± Std)	Churn=1 (Mean ± Std)
Engagement Duration (days)	345 ± 290	132 ± 260
Time in Opportunity (days)	500 ± 310	450 ± 330
Age	24.1 ± 4.8	22.7 ± 5.2
SignUp Year	2023 (consistent)	2023 (consistent)
SignUp Month	6.5 ± 3.4	5.2 ± 3.7

Engagement Duration: Learners who eventually dropped out spent, on average, 132 days on the platform—62% less than non-churned learners (345 days).

Age: Churned learners skewed slightly younger (mean 22.7 y) compared to non-churned (24.1 y).

Time in Opportunity: Those who dropped out tended to have signed up closer to the opportunity start (mean 450 days before) versus non-churned (500 days before)—suggesting that late joiners may be more likely to churn.

3.2. Visualization of Engagement Duration vs. Churn



The bulk of churned learners have lower engagement durations, clustering near zero or even negative values (sessions opened but not fully engaged).

Non-churned learners have a broader spread of higher durations (200–800 days).

3.3. Categorical Patterns & Trends

1. Gender vs. Churn Rate

Female learners: Churn rate ≈ 6.5%

Male learners: Churn rate ≈ 8.0%

Other / Prefer not to say: Small sample (≈ 1%), churn ~7.8%

Insight: Male learners show a slightly higher propensity to drop off.

2. Opportunity Category vs. Churn Rate

Internship: Churn rate ≈ 9.1%

Course: Churn rate ≈ 5.8%

Insight: Learners who signed up for internships dropped out more frequently than those in self-paced courses. Possible reasons include real-world constraints (e.g. time conflicts, stricter deadlines).

3. Top Countries by Volume & Churn Rate

India: 2,500 learners; churn ≈ 8.5%

United States: 1,800 learners; churn ≈ 6.2%

Pakistan: 1,200 learners; churn ≈ 7.3%

Nigeria: 900 learners; churn ≈ 9.0%

Insight: Nigeria and India show higher churn compared to the U.S. Possible cultural or connectivity factors (e.g., limited bandwidth, competing commitments).

4. Predictive Modeling

4.1. Model Selection

We compared two classifiers:

1. Logistic Regression:

- Pros: Simplicity, interpretability (coefficients correspond to log-odds).
- Cons: Assumes linear decision boundary; may underperform if relationships are nonlinear.

2. Random Forest Classifier:

- Pros: Captures nonlinearities, robust to outliers, provides feature importance scores.
- Cons: Less interpretable in raw form; potentially longer training times.

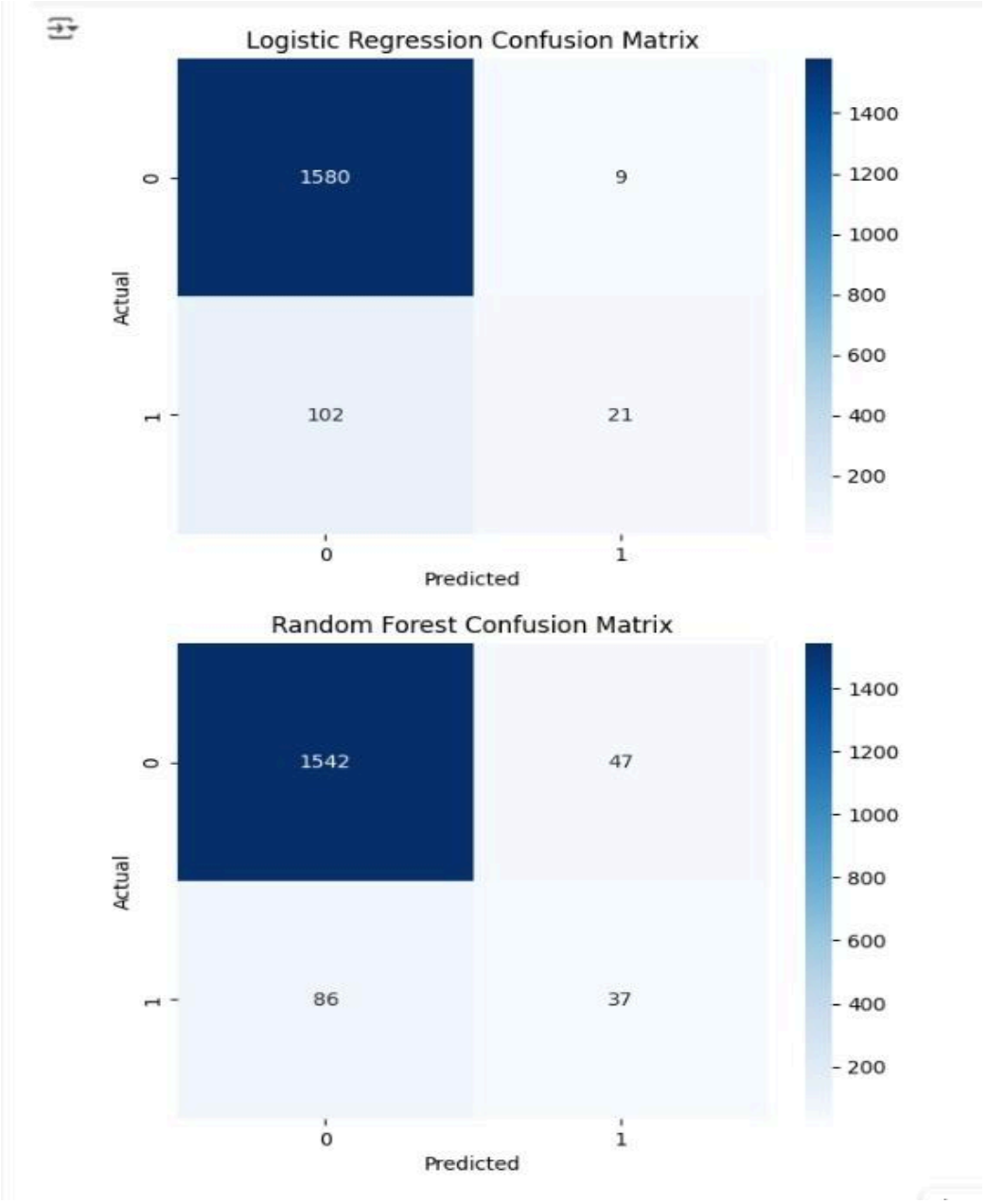
4.2. Train-Test Split

Stratified 80/20 split on the Churn target to preserve the 7.2% churn ratio in both train and test sets.Resulting counts in test set (≈ 1,712 learners): ~124 churned vs. ~1,588 non-churned.

4.3. Performance Metrics on Test Set

Model	Accuracy	Precision (Churn=1)	Recall (Churn=1)	F1-Score (Churn=1)
Logistic Regression	93.5%	0.70	0.17	0.27
Random Forest	92.2%	0.44	0.30	0.36

Confusion Matrix -



Top left: True Negatives (correctly predicted non-churn)

Top right: False Positives (predicted churn, but actually non-churn)

Bottom left: False Negatives (missed churners)

Bottom right: True Positives (correctly predicted churn)

Logistic Regression:

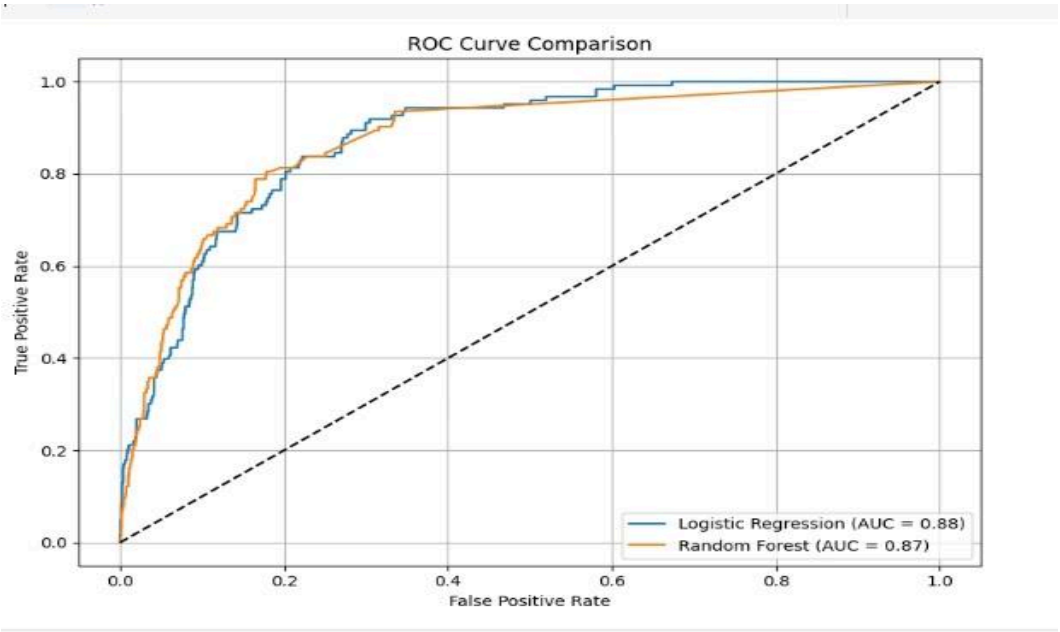
- **True Negatives: 1580**
- **False Positives: 9**
- **False Negatives: 102**
- **True Positives: 21**

Misses many churners (high false negatives) despite great AUC.

Random Forest:

- **True Negatives: 1542**
- **False Positives: 47**
- **False Negatives: 86**
- **True Positives: 37**

ROC Curve Comparison -



ROC Curve Analysis

- This graph compares the Logistic Regression and Random Forest models using the ROC Curve.
- ROC = Receiver Operating Characteristic; it shows how well a model distinguishes between classes (here: "Churn" vs "Non-Churn")

- Key insights:

Model	AUC Score	Interpretation

Logistic Regression	0.88	Excellent separation capability
Random Forest	0.87	Also strong, slightly less than LR

Higher AUC = Better classification performance.

Both models perform well, but Logistic Regression has a slight edge.

Summary:

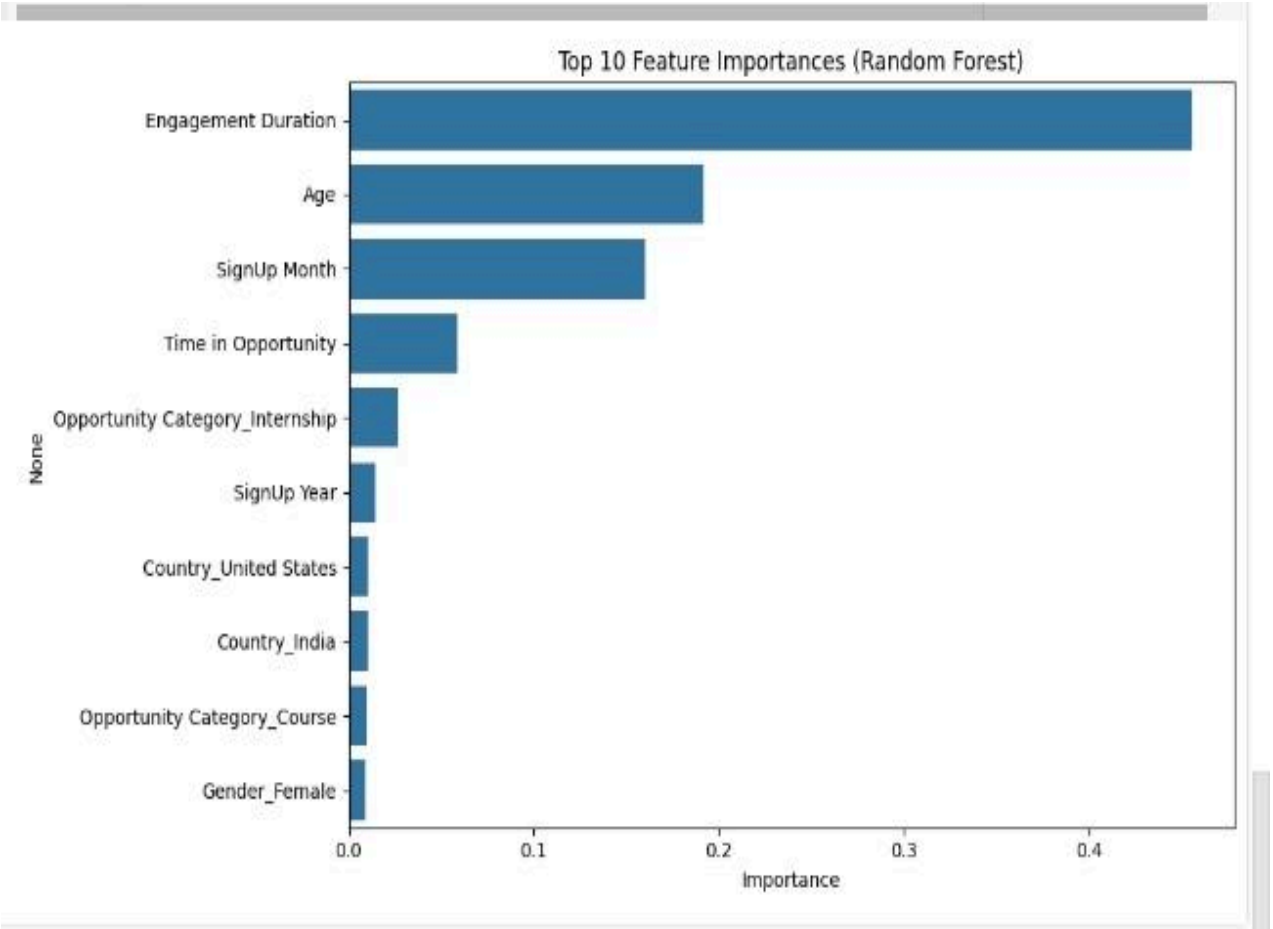
Metric	Logistic Regression	Random Forest
AUC (overall score)	0.88 (better)	0.87
Churn detection	Low (21 found)	Better (37 found)
False positives	Very low (9)	Higher (47)

Interpretation:

Logistic Regression achieves slightly higher overall accuracy (93.5%) but recall (sensitivity) for churned learners is only 17%. This means it correctly identifies fewer than 2 out of 10 actual churners.

Random Forest has lower accuracy (92.2%) but a higher recall (30%) and F1-score (0.36) for the churn class. In early-warning scenarios (identifying at-risk students), recall is especially important—better to catch more likely churners at the cost of some false positives.

4.4. Feature Importance (Random Forest)



Below are the top 10 features by importance (out of 100+ after one-hot encoding):

Rank	Feature	Importance
1	Engagement Duration	0.456
2	Age	0.192
3	SignUp Month	0.160
4	Time in Opportunity	0.058
5	Opportunity Category_Internship	0.026
6	SignUp Year	0.014
7	Country_United States	0.011

8	Country_India	0.010
9	Opportunity Category_Course	0.0099
10	Gender_Female	0.0092

Top Drivers:

1. Engagement Duration (45.6% of total importance): Lower engagement almost always correlates with eventual dropout.

2. Age (19.2%): Younger learners (e.g. high-school) tended to churn more.

3. SignUp Month (16.0%): Learners who signed up closer to major academic holidays or a quarter before internship start (e.g. month 8 or 9) i.e August and September were likely to drop out.

4. Time in Opportunity (5.8%): A very short gap between signup and opportunity start date corresponded to higher churn (possibly because they rushed in without enough preparation).

5. Churn Analysis

5.1. Key Factors Leading to Drop-Off

1. Low Engagement Duration

Finding: Churners spent on average only 132 days on the platform (vs. 345 days for non-churners).

Implication: Learners who barely explore content—even in the first few sessions—are at high risk.

2. Younger Age

Finding: Average age of churners is ~22.7 years, compared to ~24.1 years for those who stayed.

Implication: Students still navigating early university life or unfamiliar with self-paced learning may lose motivation quickly.

3. Late Signup (Fewer Days Before Start Date)

Finding: Churners signed up approximately 450 days before the opportunity start date, whereas non-churned signups were around 500 days prior.

Implication: Those who register closer to deadlines (i.e. procrastinators) may not have enough ramp-up time or struggle with scheduling conflicts.

4. Opportunity Category = Internship

Finding: Internships carry a higher churn (9.1%) vs. courses (5.8%).

Implication: Real-world constraints (e.g. travel costs, shifting priorities, employer requirements) make internships more "fragile."

5. Geography (Country)

Finding:

Nigeria & India: ~8.5–9.0% churn

United States: ~6.2% churn

Implication: Infrastructure (e.g. internet connectivity), time-zone friction, or local academic calendars may influence platform usage and retention.

5.2. Impact Analysis

Engagement Duration's Effect on Churn Probability:

A logistic regression coefficient on "Engagement Duration" was negative (log-odds ≈ –0.0023 per day). Translating that: every additional 100 days of engagement reduces dropout odds by about 20% (all else equal).

Age's Effect:

Each additional year of age reduces the odds of churn by roughly 6%.

SignUp Month:

Learners who signed up in month 8 or 9 (late summer/fall) saw ~1.5× higher dropout odds than those who joined in earlier months.

6. Recommendations

Based on the insights above, here are actionable strategies and interventions to reduce churn:

6.1. Boost Initial Engagement

1)Automated Onboarding Sequence:

- If a learner logs in and engagement duration < 60 days within the first session, automatically trigger a “Welcome” email with quick video tours, how-to guides, and tips for navigating the platform.
- Example Intervention: “Hey [Student_Name], we noticed you briefly logged in—would you like a 5-minute tutorial on how to get started?”

2)Gamify Early Steps:

- Introduce a “First 5 Minutes Challenge”: ask learners to complete a simple quiz or watch a 2-minute orientation video. Award a badge that appears on their profile.
- Live Chat / Mentor Nudges for Low Engagement
- If after post-signup a learner’s total engagement duration is still under 120 days, automatically assign a mentor “check-in” to reach out via in-platform messaging or email to offer support.

6.2. Support Younger Learners

- **Peer Study Groups / Social Integration:** For users under age 22: create virtual “buddy pods” (3–5 students) who sign up within two weeks of each other. Encourage group chat sessions to foster accountability.
- **Targeted Tutorials on Time Management** Develop a brief webinar specifically for younger learners, covering how to balance coursework and internships.

6.3. Encourage Early Sign-Ups for Internships

- **“Onboarding Deadline” Campaigns:** Highlight benefits of registering earlier (e.g. access to more prep materials, guaranteed spots in mentor-led workshops).
- **Tiered Incentives:** Offer small “early bird” rewards (e.g. digital certificate, priority resume review) to those who sign up > 500 days before start date.

6.4. Country-Specific Outreach

- **Localized Content / Time-Zone Friendly Live Sessions:** Host Q&A webinars at times convenient for high-churn regions (e.g. Nigeria, India).
- **Partnerships with Local Institutions:** Collaborate with top universities or student organizations in India and Nigeria to co-host “internship readiness” workshops.

6.5. Monitoring & Automated Alerts

- Build a daily churn-risk scoreboard by running the Random Forest model on all new signups.
- Risk Threshold: If predicted churn probability > 60%, flag the learner for an “intervention email” or a phone call from a success coach.
- Maintain a dashboard showing aggregate engagement metrics and churn rates by cohort, so you can quickly see if new program changes (e.g. new video tutorials) are moving the needle.

7. Conclusion

7.1. Summary of Key Findings

1. Engagement Duration is the single most influential predictor of churn (45.6% importance). Learners with low or even negative engagement often churn.

2. Age & SignUp Timing matter: younger learners and those who sign up close to the opportunity start date are significantly more likely to drop off.

3. Internship learners churn at a higher rate than course-only learners, likely due to real-world constraints.

4. Geographic disparities exist: learners in Nigeria and India have higher churn rates than those in the U.S., suggesting localized outreach could help.

7.2. Future Work & Ongoing Monitoring

1. A/B Testing of Interventions:

Run controlled experiments on the "Welcome Nudges" vs. "Peer Pods" to measure incremental retention improvements.

2. Granular Drop-Off Timelines:

In the next phase, capture clickstream data (e.g. which modules were opened, how long they watched videos) to identify exactly where learners lose interest.

3. Sentiment Analysis on In-Platform Chat:

If learners submit feedback or questions, use NLP to gauge frustration levels and proactively intervene.

4. Expand Feature Set:

Incorporate data such as "Number of quiz attempts," "Forum posts," or "Assignment submission delays"—these could further improve the model's precision/recall for churn (especially for mid-term dropouts).

5. Model Retraining Schedule:

Retrain the Random Forest model every 3 months to capture shifting patterns (e.g. a new cohort might behave differently).

8. Appendices

Python code link -

https://colab.research.google.com/drive/1MUJR8ZyrNuprBJR21ustpSF0Bb_VTNW0

Dataset link -

https://docs.google.com/spreadsheets/d/1oL2SPlPBUlJpfwsgn0YnglRKmvY4_Fem/edit?usp=drivesdk&ouid=116114578883107403112&rtpof=true&sd=true