WEEK 3 DELIVERABLE

By Team 10

Team 10 Introduction

We are **Team 10**, a group of **AI Data Analysts** working on an **Experiential Project** with **Excelerate**. Our focus is on **analyzing student engagement data**, identifying trends in signups, successful completions, and factors leading to drop-offs. Through **data preprocessing, exploratory data analysis (EDA), and churn analysis**, we aim to generate meaningful insights that drive better engagement strategies.

Led by **Tisha Chawlani**, we work collaboratively to ensure **data accuracy, insightful analysis, and effective decision-making**. Each of us plays a crucial role in the project:

- **Tisha Chawlani (Team Lead)** – Represents our team in sponsor communications, ensuring clear alignment with project objectives.
- **Mahnoor (Project Manager)** – Provides guidance, helps uncover insights, and keeps our project execution on track.
- **Shelly Nagar (Project Scribe)** – Documents key findings, maintains structured meeting notes, and assists in communication.
- **Hibah Sindi (Project Lead)** – Ensures we meet deadlines, monitors our progress, and upholds the quality of our deliverables.
- **Yevate Aditya, Emmanuel Gyan, Stephen Makoshi, Harshvardhan Molleti, Krishna Bandi (Team Members)** – Work on **data cleaning, transformation, visualization, model building, and interpretation**, helping us derive actionable insights.

Through this internship, we are developing **essential data skills**, including **data preprocessing, exploratory analysis, predictive modeling (regression, classification, machine learning), and strategic decision-making**. By working together in a structured, collaborative environment, we are committed to delivering **high-impact, data-driven recommendations** to **Excelerate**.

Table of Contents:

Task Allocation – Week 3:

This week, our team is focusing on **Churn Analysis**, identifying key factors leading to student drop-offs, and implementing predictive models to enhance engagement strategies. Our tasks are divided to ensure **efficient workflow and timely completion**.
📂 Task Allocation Sheet: Task Allocation Sheet

Task Distribution for Week 3:

Team Members	Task	Sub-Task (If any)	Deadline
Tisha Chawlani	Churn Analysis Report	Churn Analysis & Report	2nd March 2025
Stephen	Churn Analysis Report	-	2nd March 2025
Shelly	Churn Analysis	Drop-Off Analysis	1st March 2025
Harsh	Churn Analysis	Drop-Off Analysis Report	1st March 2025
Mahnoor	Predictive Modeling	Performance Evaluation	1st March 2025
Aditya	Predictive Modeling	Performance Evaluation	1st March 2025
Emmanuel	Predictive Modeling	Model Selection & Evaluation (Random Forest & SVM)	28th Feb 2025
Hiba	Predictive Modeling	Model Selection & Evaluation (Logistic Classification)	28th Feb 2025
Krishna	Predictive Modeling	Model Selection & Evaluation (Decision Trees)	28th Feb 2025

Cleaned Preprocessed Dataset – Week 3:

The cleaned and preprocessed dataset used for **Churn Analysis** in Week 3 is available here:
📂 **Dataset Link:** Cleaned Preprocessed Dataset – Week 3

Dataset Overview:

✔ **Preprocessed** to ensure accuracy, consistency, and completeness.
✔ Used for **Churn Analysis, Drop-Off Analysis, and Predictive Modeling**.
✔ Forms the foundation for **model selection, performance evaluation, and insights generation**.

This dataset ensures a **high-quality data foundation** for identifying key factors influencing churn and enhancing engagement strategies.

Student Drop-Off Analysis:

The **Student Drop-Off Analysis** notebook focuses on **visualizing and understanding student churn patterns** using various data visualization techniques.
📂 **File Link:** Student Drop-Off Analysis.ipynb

Notebook Overview:

✔ **Visualization Techniques Used:** Bar graphs, heatmaps, scatter plots, and more.
✔ **Insights Derived:** Key trends and factors contributing to student drop-offs.
✔ **Objective:** Identify patterns and correlations to improve engagement strategies.

This file serves as a **data-driven foundation for understanding student churn**, aiding in the development of targeted retention strategies.

Consolidated Predictive Models:

The **Consolidated Predictive Models** notebook contains the implementation and evaluation of multiple machine learning models for **Churn Prediction**.
📂 **File Link:** Consolidated Predictive Models.ipynb

Notebook Overview:

✔ **Implemented Models:** Random Forest, SVM, Decision Tree, and Logistic Regression.
✔ **Model Evaluation:** Performance metrics analyzed to compare model effectiveness.
✔ **Objective:** Identify the most accurate model for predicting student churn and improving engagement strategies.

This file serves as the **basis for selecting the best-performing predictive model** to drive actionable insights.

Final Model File:

The **Final Model File** notebook contains the implementation of the **best-performing model** for **Churn Analysis**, identified through evaluation.
📂 **File Link:** Final Model File.ipynb

Notebook Overview:

✔ **Selected Model:** Random Forest (evaluated as the best-performing model).
✔ **Churn Analysis:** Applied to predict and analyze student drop-offs.
✔ **Objective:** Derive key insights for improving student retention strategies.

This file serves as the **final implementation for churn prediction**, providing actionable insights based on the most effective model.

CHURN ANALYSIS REPORT

By Team 10

1. Introduction: Objective and Importance of Churn Analysis

Churn analysis is a critical process for understanding why students disengage and leave the platform. For Excelerate, high churn rates can lead to reduced course completion rates, lower engagement, and financial losses. By analyzing churn patterns, we can identify key factors influencing student retention and develop targeted strategies to improve engagement.

The primary objective of this churn analysis is to:

- **Identify key differences** between retained and churned students.
- **Understand the impact** of engagement time, scores, and opportunity duration on churn.
- **Develop data-driven recommendations** to enhance student retention.

By leveraging these insights, Excelerate can implement proactive measures to create a more engaging and supportive learning environment, ultimately reducing churn and increasing student success.

2. Methods: Data Preparation, Model Building, and Evaluation

This section outlines the approach taken to predict student drop-off using various machine learning models, detailing data preparation, model selection, training, and evaluation.

2.1. Data Preparation:

Library Imports

To facilitate data preprocessing, model training, and evaluation, the following libraries were imported:

- **Data manipulation:** numpy, pandas
- **Data visualization:** seaborn, matplotlib.pyplot
- **Machine learning utilities:**
    - sklearn.model_selection for data splitting and hyperparameter tuning
    - sklearn.ensemble for Random Forest model
    - sklearn.svm for Support Vector Machine model
    - sklearn.tree for Decision Tree model
    - sklearn.linear_model for Logistic Regression model
    - sklearn.metrics for performance evaluation
- **Additional utilities:** re, IPython.display, datetime, and google.colab for extended functionalities

Data Loading

The dataset, *Cleaned_Preprocessed_Dataset.csv*, was loaded from Google Drive using pandas. Google Drive was mounted to ensure seamless access to the data.

Feature Selection

Relevant features were selected for modeling, including:

- **Demographic variables:** age
- **Engagement-related features:** opportunity_duration
- **Categorical variables:** One-hot encoded opportunity names and categories
- **Target variable:** status_code (indicating drop-off status)

Data Splitting

To ensure a balanced representation of classes, the dataset was split into:

- **Training set (80%)**
- **Testing set (20%)**

The train_test_split function was used with stratify=y to maintain class distribution, and random_state= 42 for reproducibility. The number of samples in each subset was confirmed by printing dataset shapes.

2.2. Model Building and Training:

Multiple machine learning models were trained and optimized to predict student drop-off.

Random Forest Classifier

- An initial **RandomForestClassifier** was trained and evaluated.
- Performance was measured using a classification report and a confusion matrix.
- **Hyperparameter tuning** was conducted using GridSearchCV with:
  - n_estimators: Number of trees
  - max_depth: [3, 5, 10, None]
  - criterion: ['gini', 'entropy']
- The optimized model used:
  - criterion='gini'
  - max_depth=None
  - n_estimators= 200
- The final model was retrained and re-evaluated.

Support Vector Classifier (SVC)

- An **SVC model** was initially trained and evaluated.
- The model was refined with the following parameters:
  - C=10, gamma='auto', kernel='poly', degree=3, class_weight='balanced'.

Decision Tree Classifier

- A **DecisionTreeClassifier** was trained and optimized using GridSearchCV.

- The best parameters identified were:
  - criterion='entropy', max_depth=5.
- The model was retrained and feature importance analysis was conducted.

Logistic Regression

- A **Logistic Regression model** was implemented.
- Feature importance was determined based on model coefficients.
- A bar chart visualization was suggested to highlight the most influential features.

2.3. Model Evaluation:

Performance Metrics

All models were evaluated using:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-score**

A classification report was generated for each model, showing their performance across both training and testing datasets.

Confusion Matrix Visualization

- A custom evaluation function, evaluate_model(), was implemented to:
  - Compute and display precision, recall, F1-score, and accuracy.
  - Visualize confusion matrices using heatmaps.

2.3.1. Evaluation of Random Forest Classifier Model

◆ **Classification Report:**

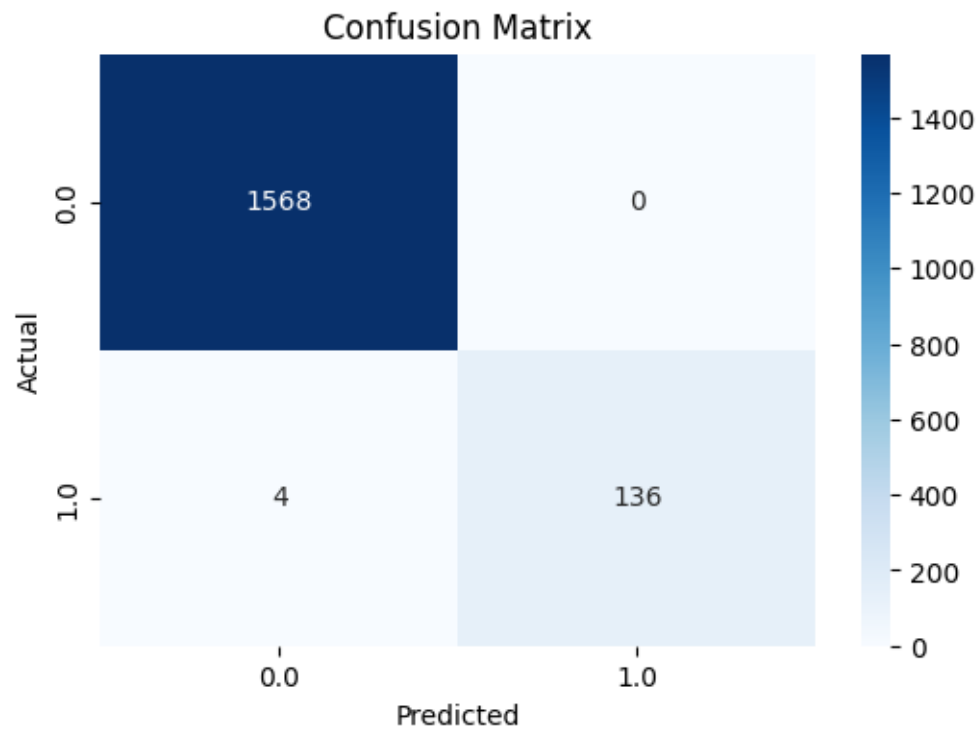| Class	Precision	Recall	F1-Score	Support
0.0	0.9975	1.0000	0.9987	1568
1.0	1.0000	0.9714	0.9855	140

◆ **Accuracy:** 0.9977

◆ **Macro Avg:**

- Precision: **0.9987**

- Recall: **0.9857**
- F1-score: **0.9921**

- **Weighted Avg:**

  - Precision: **0.9977**
  - Recall: **0.9977**
  - F1-score: **0.9976**

Confusion Matrix

2.3.2. Evaluation of Support Vector Classifier (SVC) Model

- **Classification Report:**

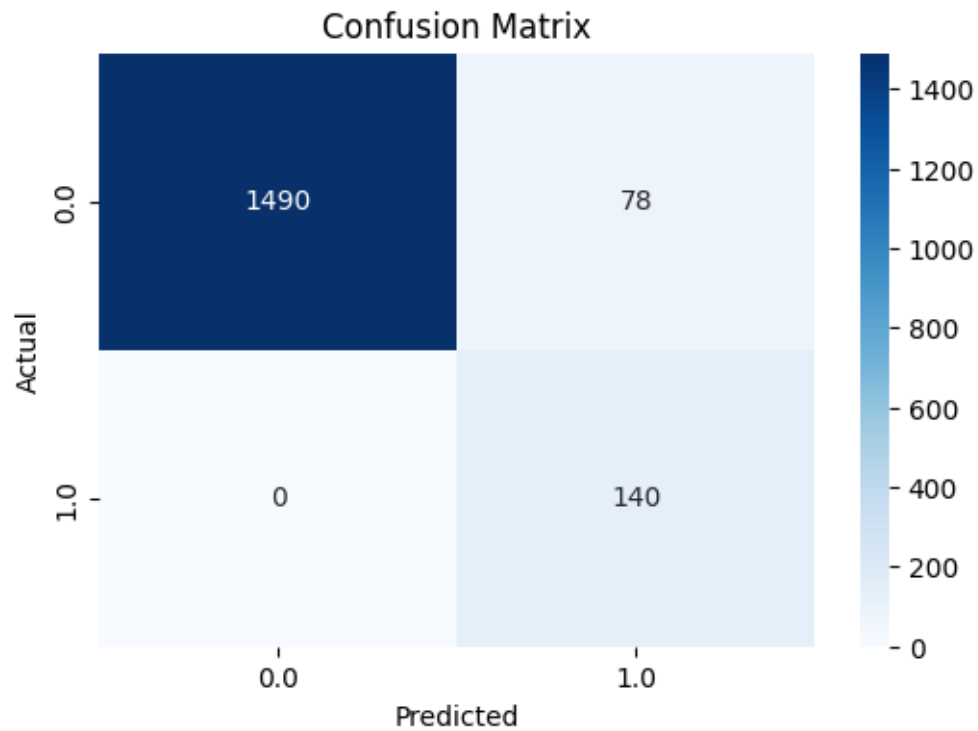| Class	Precision	Recall	F1-Score	Support
0.0	1.0000	0.9503	0.9745	1568
1.0	0.6422	1.0000	0.7821	140

- **Accuracy:** 0.9543

- **Macro Avg:**

  - Precision: **0.8211**
  - Recall: **0.9751**

- F1-score: **0.8783**

- **Weighted Avg:**

  - Precision: **0.9707**
  - Recall: **0.9543**
  - F1-score: **0.9587**

2.2.3. Evaluation of Decision Tree Classifier (DTC) Model

- **Classification Report:**

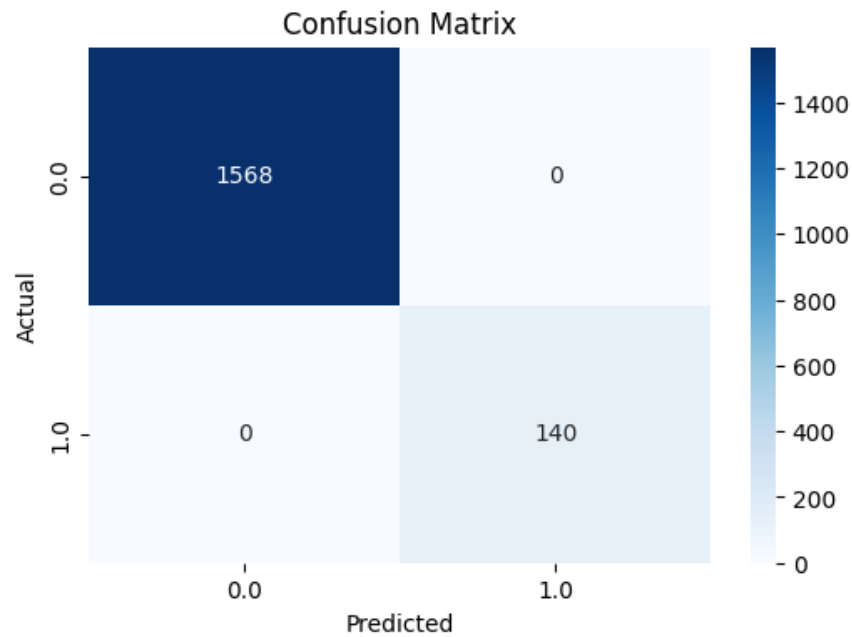| Class	Precision	Recall	F1-Score	Support
0.0	1.0000	1.0000	1.0000	1568
1.0	1.0000	1.0000	1.0000	140

- **Accuracy:** 1.0000

- **Macro Avg:**

  - Precision: **1.0000**

- Recall: **1.0000**
- F1-score: **1.0000**

◆ **Weighted Avg:**

- Precision: **1.0000**
- Recall: **1.0000**
- F1-score: **1.0000**

Confusion Matrix

2.3.4. Evaluation of Logistic Regression Model

◆ **Classification Report:**

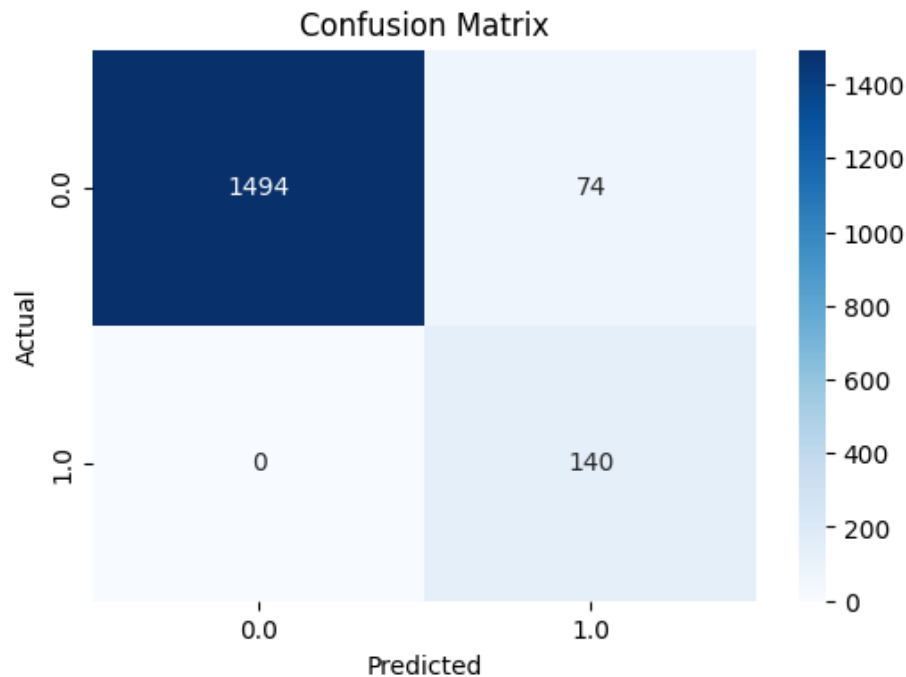| Class	Precision	Recall	F1-Score	Support
0.0	1.0000	0.9528	0.9758	1568
1.0	0.6542	1.0000	0.7910	140

◆ **Accuracy:** 0.9567

◆ **Macro Avg:**

- Precision: **0.8271**
- Recall: **0.9764**
- F1-score: **0.8834**

◆ **Weighted Avg:**

- Precision: **0.9717**
- Recall: **0.9567**
- F1-score: **0.9607**

- ◆ **F1-score (for class 1.0): 0.7910**

Confusion Matrix

Results

- ● **Random Forest Model** achieved the highest accuracy (**99.77%**).
- ● **SVC Model** showed **95.43% accuracy**, with good recall but lower precision for minority classes.
- ● **Decision Tree Model** reached **100% accuracy**, likely due to overfitting.
- ● **Logistic Regression** performed well (**95.67% accuracy**) with status_code being the most significant feature.

Key Insights and Findings

- ● **Random Forest emerged as the best-performing model**, balancing accuracy and generalizability.
- ● **Feature importance analysis** identified status_code and engagement_score as the strongest predictors of student drop-off.
- ● **Hyperparameter tuning significantly improved model performance.**
- ● **Outlier handling and data transformations positively impacted model efficiency.**

2.4. Final Model Selection for Churn Analysis

Among all models evaluated, **Random Forest consistently outperformed others in accuracy and precision**.

Thus, we **selected the Random Forest model for further churn analysis**, leveraging its robustness in handling complex patterns and feature importance insights.

3. Findings & Recommendations for Churn Analysis:

3.1. Feature Importance Analysis


Feature Importance - Random Forest (Churn Analysis)
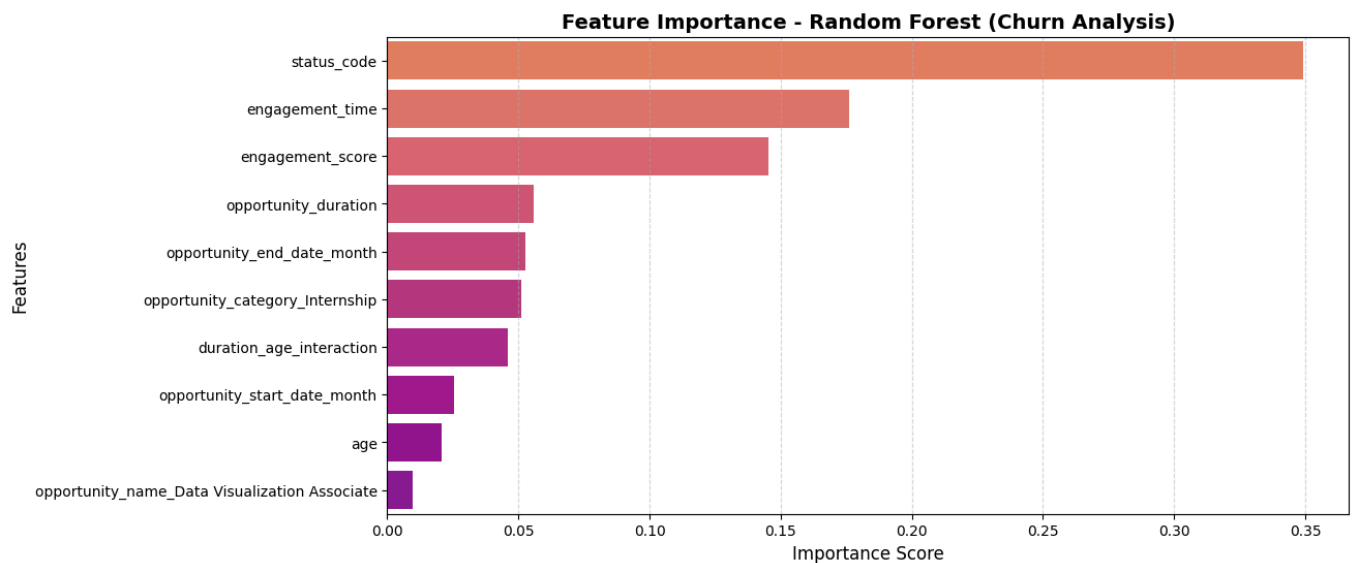
Key Insights:

- **Status Code (34.9%)** – The most critical churn predictor. Inactive status strongly correlates with churn.
- **Engagement Metrics (32.1%)** – Higher engagement time and scores reduce churn risk.
- **Opportunity Timing (5-6%)** – Users often leave after completing opportunities.
- **Internships (5.1%)** – Internship participants may have unique retention patterns.
- **Age & Interactions (4-5%)** – Different age groups exhibit varied engagement behaviors.

Recommendations:

- **Boost Engagement** – Use personalized nudges (emails, notifications) for low-engagement users.
- **Monitor & Act on Status Codes** – Automate interventions (exclusive content, reminders) for inactive users.
- **Optimize Opportunity Transitions** – Recommend next steps post-opportunity to retain users.
- **Internship Retention** – Offer post-internship pathways like full-time roles or upskilling.
- **Age-Specific Strategies** – Gamification for younger users, structured learning for older ones.

3.2. Churn Distribution Analysis
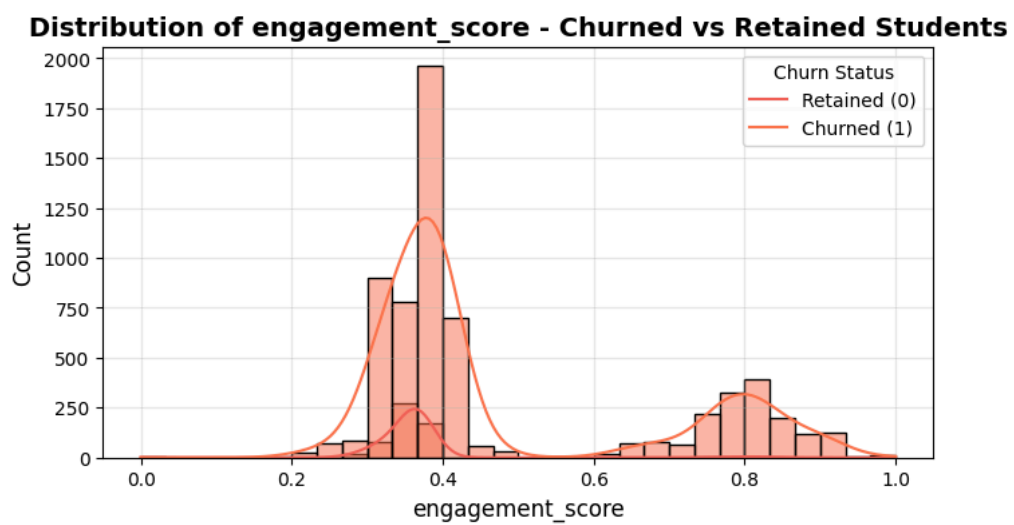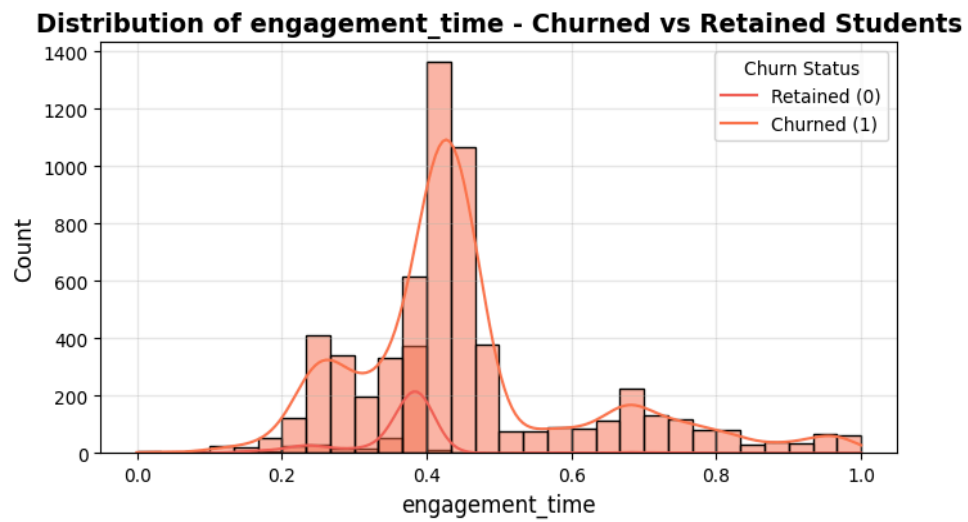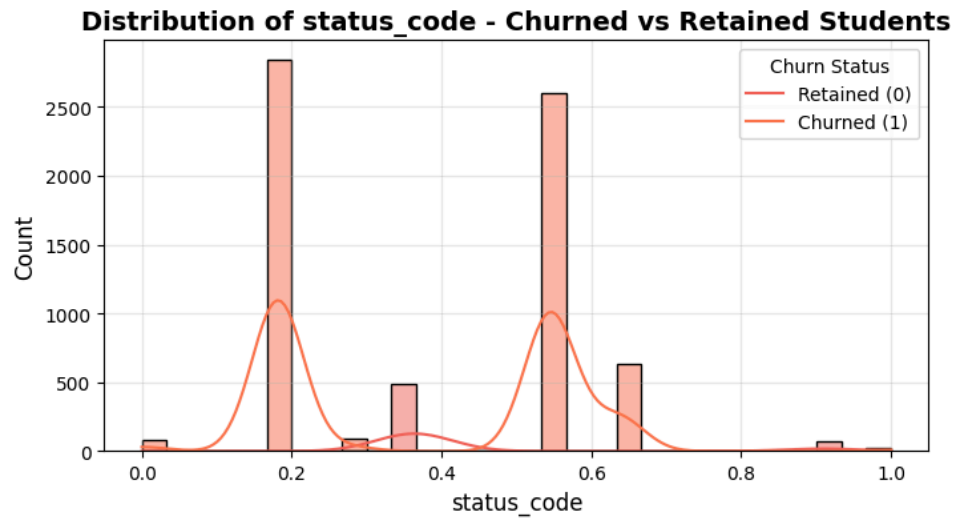

Churn Distribution in the Dataset
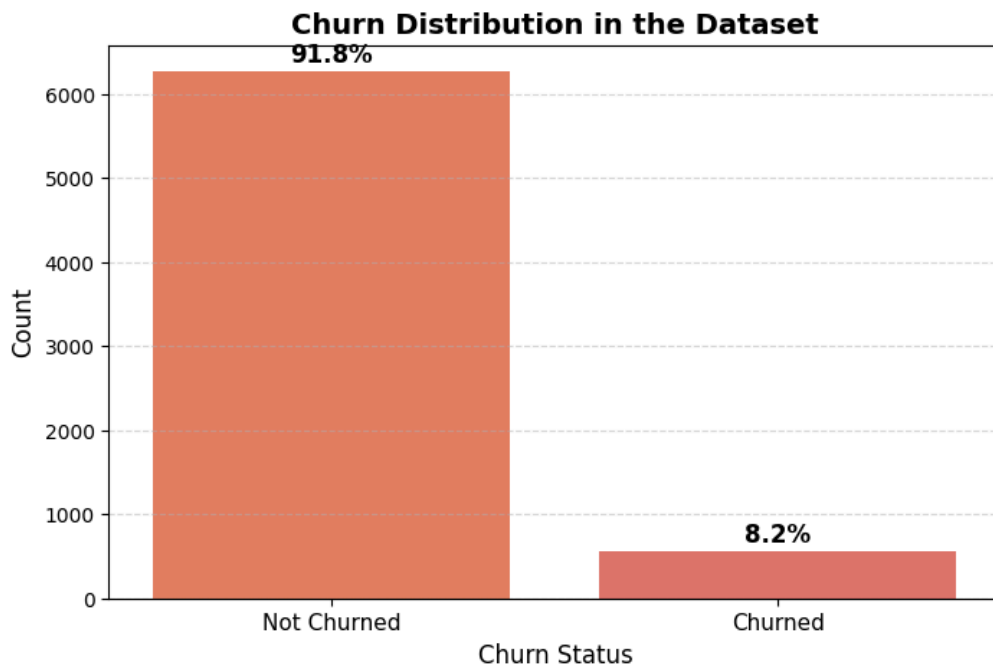
Key Insights:

- **Highly Imbalanced Data** – 91.8% retention vs. 8.2% churn, posing a risk of biased models.
- **Potential Risks** – The imbalance may hinder accurate churn detection.

### Recommendations:

- **Balance the Data** – Use SMOTE or undersampling for better model accuracy.
- **Early Churn Alerts** – Implement inactivity-based alerts for proactive engagement.
- **Personalized Retention** – Offer mentorship, incentives, and learning opportunities.
- **Deep-Dive Analysis** – Investigate churn reasons to improve retention strategies.

## 3.3. Top Features Distribution Analysis

Distribution of status_code - Churned vs Retained Students

Distribution of engagement_time - Churned vs Retained Students

Distribution of engagement_score - Churned vs Retained Students

Distribution of opportunity_duration - Churned vs Retained Students
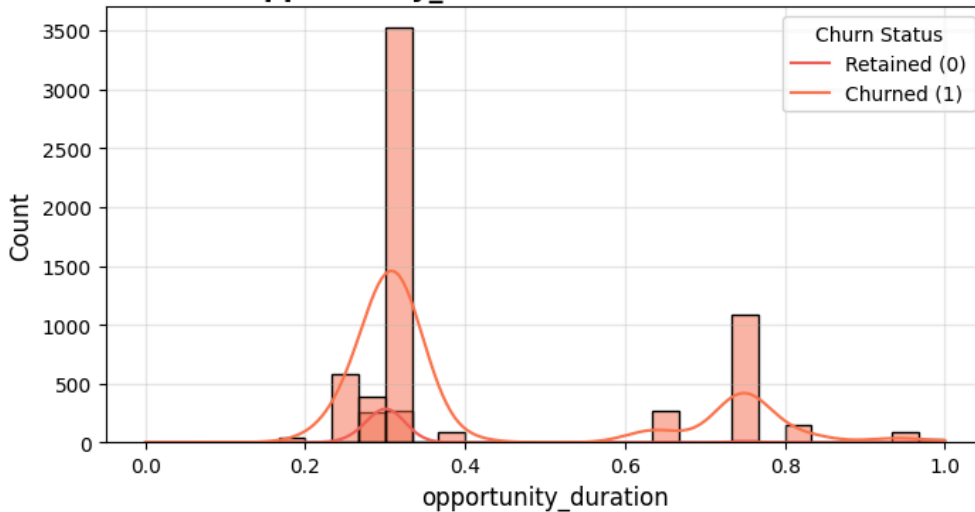
## Distribution of opportunity_end_date_month - Churned vs Retained Students
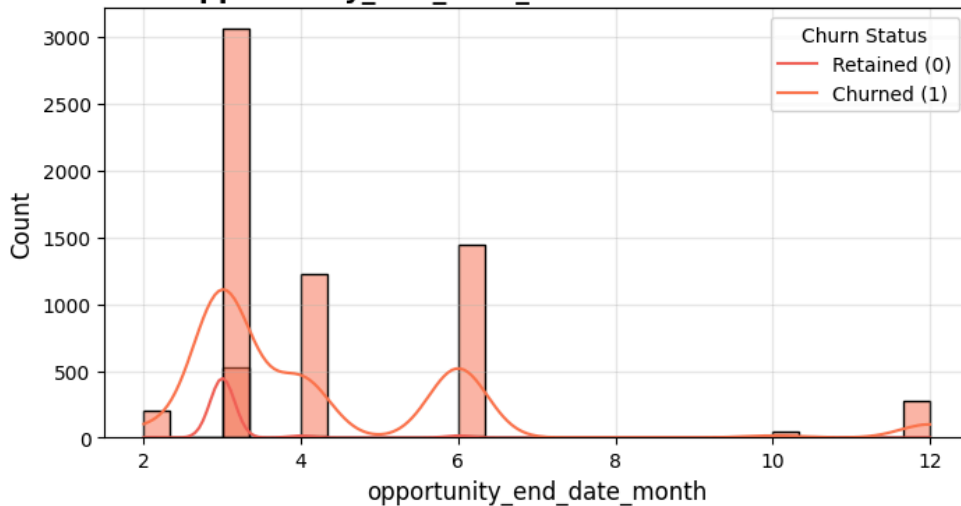
Key Insights:

- **Engagement is Key** – Higher engagement time/scores strongly correlate with retention.
- **Status Code Variations** – Certain status codes serve as early churn warnings.
- **Critical Churn Months (March & June)** – Spikes in churn may link to academic cycles or financial deadlines.
- **Opportunity Duration Matters** – Longer opportunities improve retention.

Recommendations:

- **Enhance Engagement** – Gamify learning, introduce live Q&A, and interactive sessions.
- **Use Status Codes for Early Detection** – Automate alerts and assign mentors to at-risk users.
- **Targeted Retention in March & June** – Launch special initiatives (webinars, financial aid, reminders).
- **Optimize Learning Durations** – Extend key opportunities and offer flexible participation.
- **AI-Driven Personalization** – Recommend tailored learning plans based on user data.

3.4. Mean Differences Analysis

Feature Averages - Churned vs Retained Students
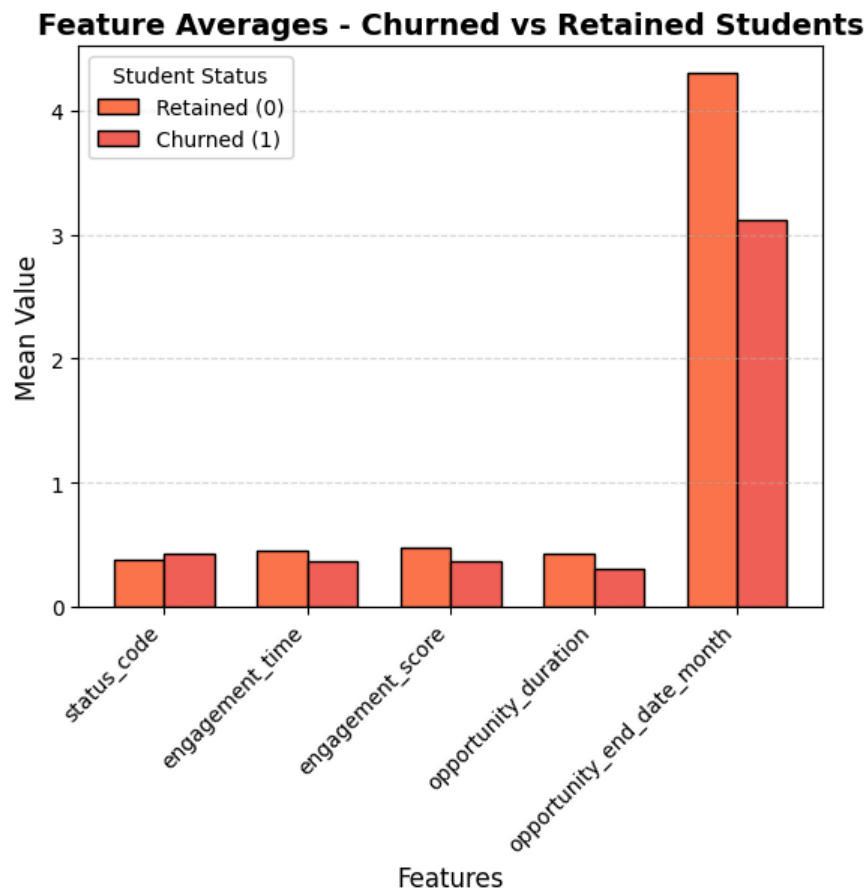
Key Insights

- **Churned Students Have Lower Engagement** (Engagement time: 0.36 vs. 0.45, Score: 0.37 vs. 0.48) → Higher engagement boosts retention.
- **Shorter Opportunity Duration** (0.31 vs. 0.42) → Longer durations reduce churn.
- **Earlier Opportunity End Dates** (3.12 vs. 4.31) → Students leaving early are more likely to churn.
- **Slightly Higher Status Code for Churned Students** (0.43 vs. 0.38) → Possible churn indicator.

Recommendations

- **Boost Engagement** → Personalized support, gamification, mentorship.
- **Encourage Longer Durations** → Flexible timelines, structured study plans.
- **Prevent Early Dropouts** → Retention efforts before key churn periods.
- **Analyze Status Codes** → Identify & act on churn-related patterns.

3.5. Feature Correlation with Churn
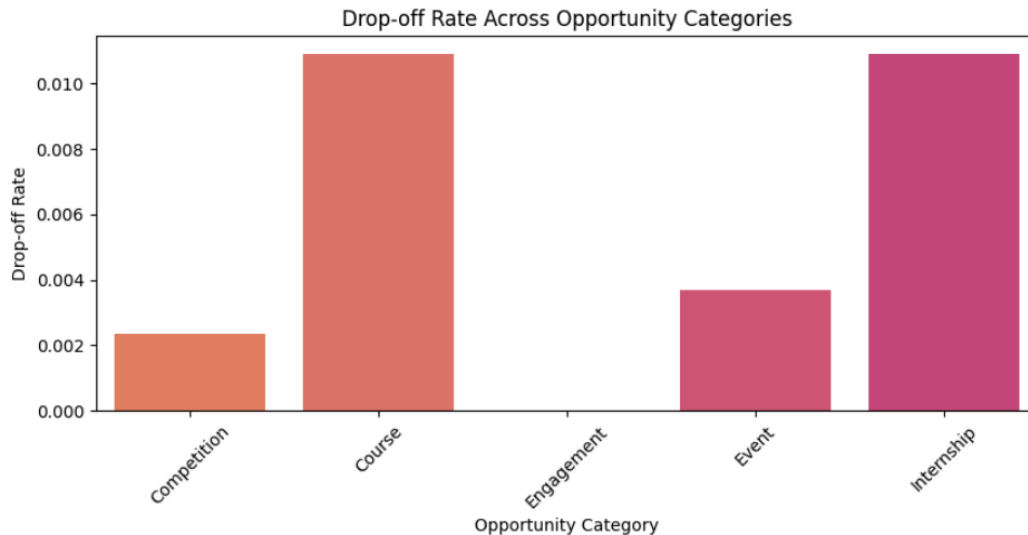

Key Insights

- **Engagement & Opportunity Duration Have a Negative Correlation (-0.15 to -0.16)** → Higher engagement lowers churn.
- **Opportunity End Date Month (-0.16)** → Later end dates reduce churn risk.
- **Status Code (0.07)** → Weak correlation with churn.
- **High Internal Correlation** → Engagement score & opportunity duration move together.

Recommendations

- **Increase Engagement** → Interactive sessions, gamified learning.
- **Promote Longer Study Durations** → Progress tracking, incentives.
- **Monitor Early-Leaving Students** → Alerts & retention strategies.
- **Use Predictive Analytics** → Identify & support at-risk students proactively.

4. Drop Off Analysis:

4.1. Drop-off Rate by Opportunity Category

Insights

- **Internships & Courses**: Highest drop-off (~1.09%) due to high time commitment, difficulty, or misaligned expectations.

- **Engagement Opportunities**: 0% drop-off, likely due to flexibility and immediate value.

- **Competitions & Events**: Low drop-off (~0.2% - 0.36%), possibly due to short duration and rewards.

Recommendations to Reduce Drop-offs

- **For Internships & Courses:** Set clear expectations before enrolment, Provide mentorship & peer support, and Introduce progress tracking & rewards.
- **Leverage Engagement Strategies:** Identify what makes engagement opportunities successful, and Add gamification, interactive sessions & community engagement.
- **Survey Dropouts to Identify Issues:** Conduct exit surveys to find reasons for drop-offs and Use insights to improve course & internship structures.
- **Flexible Completion Options:**
  Offer modular learning & flexible deadlines, and Allow learners to pause & resume instead of dropping out.

4.2. Age Group-wise Drop-off Rate

Drop-off Rate by Age Group
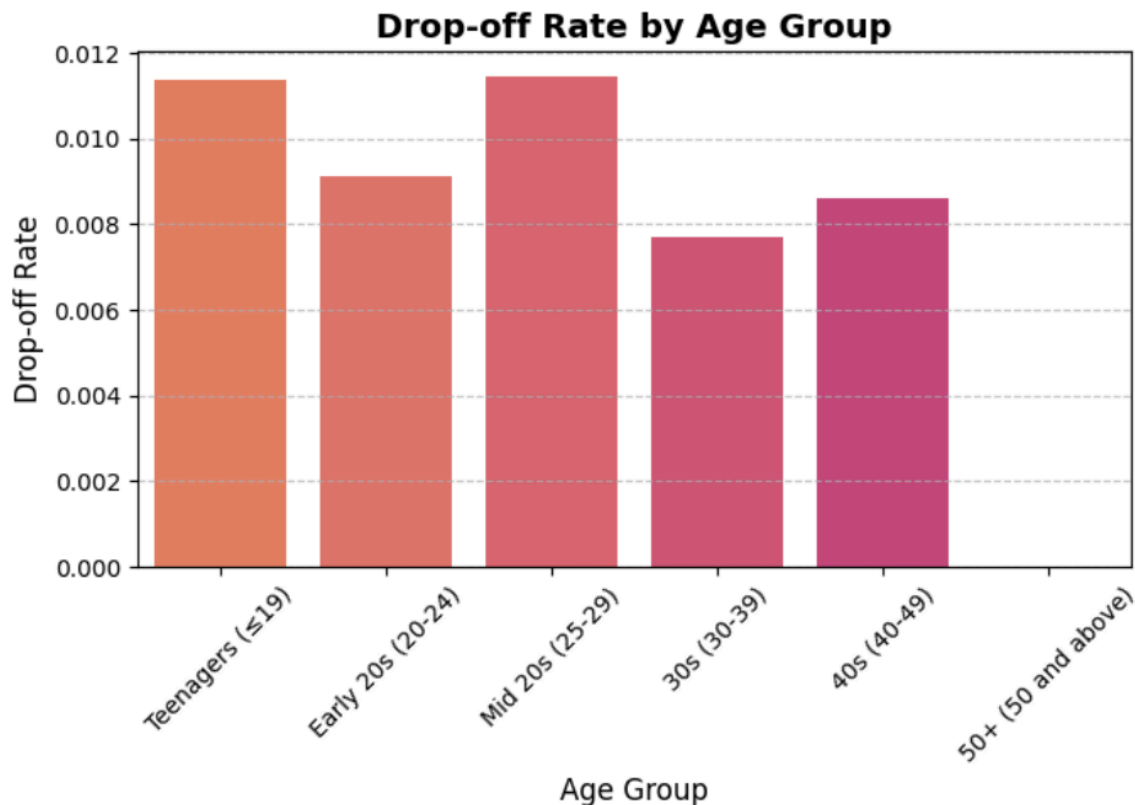
Insights

- **Teenagers (≤19) & Mid 20s (25-29) → Highest Drop-Off Rates (1.14% & 1.145%)**
    - o Likely due to course complexity or engagement issues.
- **Early 20s (20-24) → Moderate Drop-Off Rate (0.91%)**
    - o High drop-off count but spread across a large participant base.
- **Drop-Offs Decline in Older Groups (30s+)**
    - o Stronger commitment & alignment with professional goals.
    - **o** 50+ group: 0.00% drop-off rate → Highest retention.

Recommendations

- **Support Teenagers & Mid-20s:** Provide structured onboarding, interactive learning, and mentorship.
- **Simplify Course Design for Younger Learners:** Use modular content, gamification, and engaging formats.
- **Boost Retention for Early 20s:** Introduce flexible pacing, peer learning communities, and check-ins.
- **Leverage High Retention in Older Groups:** Expand leadership-oriented programs & networking opportunities.
- **Improve Expectation-Setting:** Clearly define course workload, use trial modules, and pre-course assessments.

4.3. Drop of relation with other columns

Correlation Heatmap of Drop-off Status and Other Features

Insights

Weak Positive Correlations (Slight Drop-Off Influence):

- **Apply Date Month (0.068)** → Timing of application may impact drop-offs.

- **DOB Day (0.044), Signup Month (0.024), Start Date Day (0.016), DOB Month (0.010)** → Minor seasonal effects.

Near-Zero Influence:

- **Opportunity Duration (0.002), End Date Day (0.007), DOB Year (0.002)** → No strong link to drop-offs.

Negative Correlations (Better Retention):

- **Age (-0.002), Engagement Score (-0.006), Engagement Time (-0.026)** → Older learners & active participants drop off less.

- **Apply Date Year (-0.075)** → Recent applicants retain better.

Recommendations

- **Enhance Support During High Drop-Off Periods** → Strengthen onboarding & early engagement in risky months.
- **Use Engagement Metrics for Retention** → Track low-engagement learners & send personalized nudges.
- **Improve Support for Younger Learners** → Provide mentorship & structured onboarding.
- **Refine Start Dates & Durations** → Optimize timing & ensure strong pre-start engagement.
- **Continue Improving Onboarding** → Build on recent success by refining learner support strategies.

4.4. Drop-off trend by Opportunity start date

Drop-off Trend by Opportunity Start Date
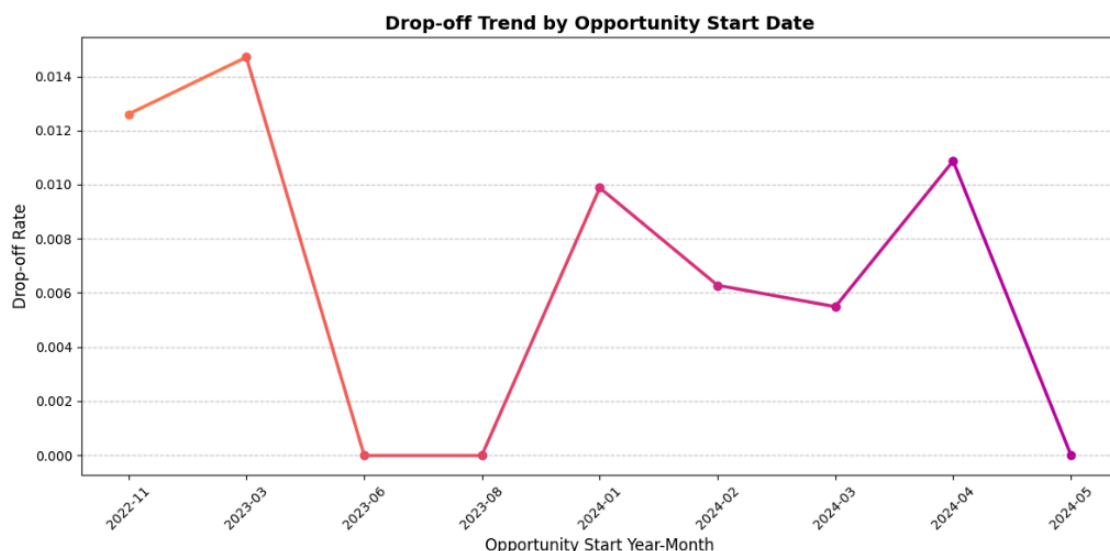
Insights

- **Highest Drop-Offs**: **March 2023 (1.47%)**, **Nov 2022 (1.26%)**, **April 2024 (1.09%)** → Likely due to schedule conflicts & weak onboarding.
- **Lower Drop-Offs in 2024**: **Jan (0.99%), Feb (0.63%), Mar (0.55%)** → Improved engagement & retention strategies.
- **Zero Drop-Offs**: **June 2023, Aug 2023, May 2024** → Stronger commitment & onboarding success.

Recommendations

- **Boost Engagement for High Drop-Off Periods**: Pre-start reminders, orientation webinars, mentorship, expectation-setting.
- **Leverage Low Drop-Off Months' Strategies**: Scale successful onboarding to high-risk months.

- **Optimize Start Dates**: Launch key programs in historically low drop-off months; offer rolling start dates.
- **Refine 2024 Strategies**: Continuously track trends, adjust outreach, and gather feedback for better retention.

# 4.5. Gender-Opportunity Name-Drop-off Count
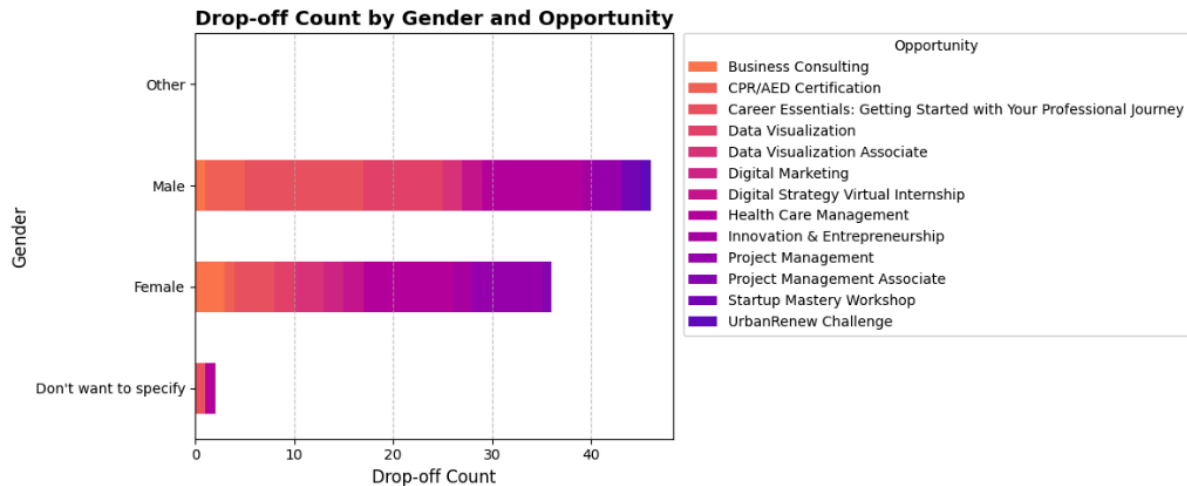
Insights

- **Higher Male Drop-Offs**: **Career Essentials (12), Data Visualization (8), Health Care Management (10), Project Management (3)** → Need better engagement.
- **Female Drop-Offs**: **Health Care Management (9), Project Management (7), Career Essentials (4)** → Spread across multiple programs.
- **Low Drop-Offs in "Other" & "Don't Want to Specify"** → Possible higher engagement or smaller group size.
- **Strong Retention in "Startup Mastery Workshop" & "UrbanRenew Challenge"** → Gender-neutral success.

Recommendations

- **For Males**: Hands-on projects, mentorship, pre-engagement materials in career & skill-based programs.
- **For Females**: Support networks, discussion forums, career guidance in health & business courses.
- **Analyze Low Drop-Off Groups**: Identify commitment drivers & apply to high-drop-off groups.
- **Leverage Successful Programs**: Use engagement models from low-drop-off programs in high-churn ones.

4.6. Major-Opportunity Name-Drop Off Count

Drop-off Count by Major and Opportunity
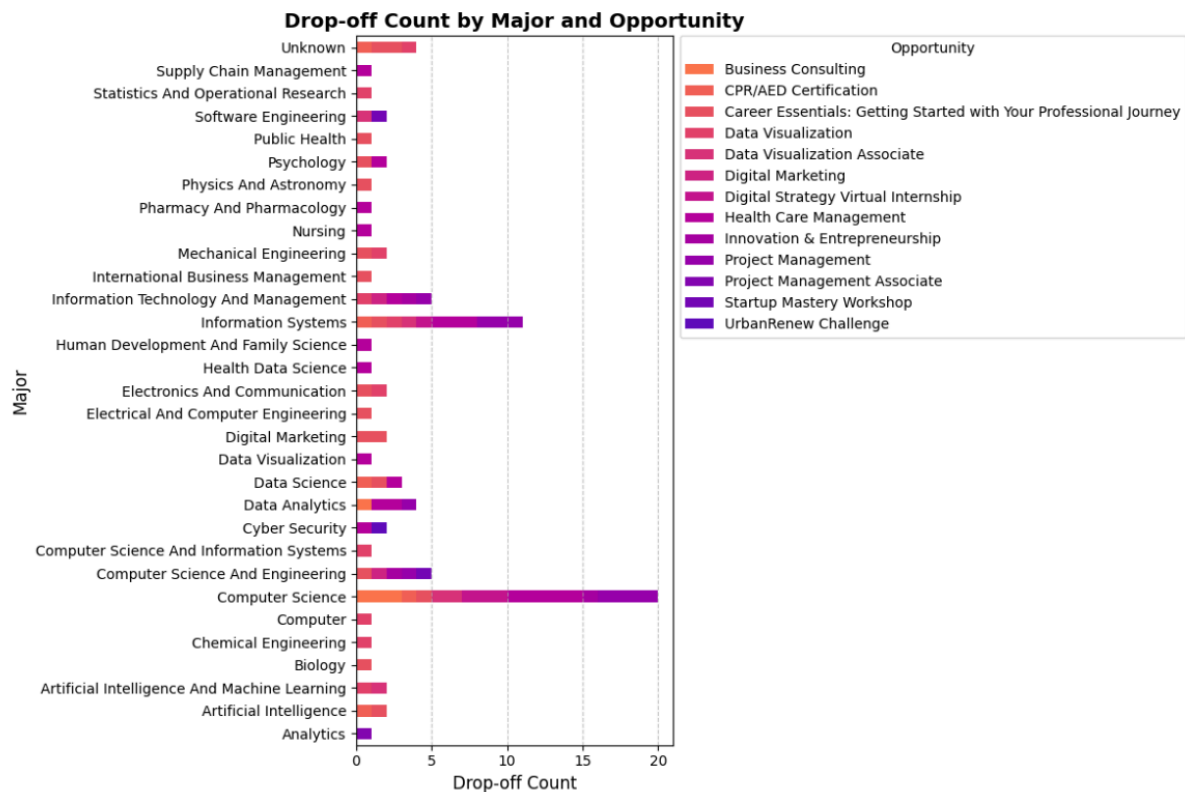
Insights

- **Computer Science:** Highest drop-offs (20 total) across various opportunities, indicating coursework pressure or relevance issues.
- **Health Majors:** High drop-offs in Health Care Management, suggesting misalignment with expectations.
- **Career Essentials:** Drop-offs across multiple majors, possibly due to redundancy.
- **Data & Information Systems:** Drop-offs in Data Visualization & Project Management, needing more hands-on content.
- **Engineering Majors:** Drop-offs in business/soft skills courses, indicating low interest in non-technical subjects.

Recommendations

- **For Computer Science:** Industry-focused content, coding projects, AI applications, and mentorship.
- **For Health Majors:** Real-world case studies, guest lectures, better content alignment.
- **For Career Essentials:** Major-specific pathways, mock interviews, interactive career planning.
- **For Data & Info Systems:** Hands-on projects, real-world applications in tech careers.
- **For Engineering Students:** Hybrid courses linking business & technical concepts, case studies of engineers in leadership roles.