Week 4 Final Report

Driving Insights Through Data

Course Name:

AI Powered Data Insights Early Internship

Group Members:

Name	Email ID
Ayad Aziz	ayadaziz95@gmailcom
Jatin Chotoo	jatin.d.chotoo@gmail.com
Jacob Cronce	jacobcronce@gmail.com
Maryam Fatima	maryamfatima2521@gmailcom
Rachel D'souza	rachel07dsouza@gmail.com

Table of Contents

Note: Bookmarks are present to reference the code for the visualization, data processing, and data mining techniques sections

Final Internship Report: AI in Predictive Analytics

1. Executive Summary

This project analyses SLU Opportunity Wise learner data to understand engagement, predict student churn, and design targeted recommendation systems. The final deliverables include a comprehensive report and a virtual team presentation, which summarize key findings from data analysis, churn prediction, and predictive modelling, and present a practical recommendation system to enhance student engagement and retention.

1.1 Objectives:

- Clean and preprocess raw learner data for reliability.
- Perform Exploratory Data Analysis (EDA) to uncover behavioral and demographic trends.
- Identify churn signals and model dropout risk.
- Develop an AI-driven recommendation system to enhance retention and engagement.

1.2 Key Findings:

- **Seasonality:** Application spikes in January, July–August, and winter months.
- **Demographics:** Learners aged 18–25 dominate, with strong participation from the US, India, and Nigeria.
- **Drop-off Trends:** High rejection rates and notable attrition in February and November. Younger learners show longer completion times.
- **Engagement:** Early sign-ups correlate with higher completion rates; weekday engagement exceeds weekends.
- **Churn Drivers:** Delays between application and start, incomplete profiles, and low early engagement predict dropout risk.

2. Introduction & Context

The SLU Opportunity Wise initiative provides learners with opportunities such as internships, courses, and workshops. However, maximizing engagement and completion rates remains a challenge. The objective of this project was to leverage data science and machine learning skills acquired throughout the internship to address this critical issue of student engagement and retention.

The project journey followed a clear, Iterative process:

2.1 Week 1 : Data Preparation

The project began with a raw dataset that was cleaned and pre-processed to ensure reliability. This involved:

- Cleaning a dataset of 8,558 records and 16 columns.
- Resolving missing values, anomalies, and standardizing country/date formats.
- Creating new features:
- including Age, Engagement Time, Season, Time Committed, Quick Applicant, Fast Starter, Opportunity Duration, and Engagement Score, to enrich the analysis.

2.2 Week 2: Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed to uncover key trends and patterns in learner behavior and demographics. Key findings included:

- **Seasonal peaks** in applications during January, July–August, and the Winter months.
- Most completions occurred within **15–30 days**,
- Though younger learners demonstrated longer completion times.
- The top three countries for learners were the **US, India, and Nigeria**,
- The **18–25 age group** shows the highest participation.

2.3 Week 3 : Churn Analysis & Predictive Modelling

Based on the insights from EDA, predictive models were developed to identify students at risk of dropping off.

- **Early risk signals** were identified, such as incomplete profiles, inactivity, and large gaps between when a learner applied and when they started.
- **Predictive models,** including Logistic Regression, Random Forest, and Gradient Boosting were tested.
- **Evaluation metrics** like ROC-AUC and Precision-Recall were used to validate the models' effectiveness.

2.4 Week 4 : Overview Presentation & Reporting :

- Recorded team's presentation with summarized findings.

This final report and presentation serve as the culmination of these efforts, demonstrating the ability to transform data into actionable strategies that can make a significant impact.

3. Objectives/Purpose

3.1 Overview

The AI Data Analysis Internship's purpose was to allow interns to tackle one of Excelerate's biggest problems: user churn. 70% of Excelerate's users churn from opportunities, leading to lots of missed growth for both parties. In this internship, our team dove into using data to find out why users were churning and developing strategies for how to retain them.

3.2 Objectives

- Analyze user data for patterns.
- Study correlations between user data and churn.
- Create hypotheses for why users are churning.
- Develop an AI prediction model to analyze current data and create predictions based on complex patterns.
- Develop a system with a series of enhancements for Excelerate to prevent churning issues and facilitate user satisfaction.
- Present final findings and suggestions to Excelerate to be implemented.

3.3 Purpose of Methods

- **Preprocessing:** Clean the data so it is readable by Python and AI prediction models.

- Feature Creation: Create more features so interns and models have more to work with.
- **EDA:** Gain insights on the data before feeding it to a prediction model so interns know what to look for.
- Modeling: Use AI to analyze complex patterns that may be missed by interns during EDA.

3.4 End Goals

- Increase Excelerate's user engagement and revenue, propelling it forward in helping students and workers build their resumés.
- Place an equal focus on the various aspects of the recommendation system so the right users are being given the right opportunities.
- Allow opportunities for collaboration and feedback between users and Excelerate to optimize all aspects of the experience.

4. Methodology / Approach

4.1 Understanding the Data:

The dataset comprises 8,558 observations with 16 variables, including both categorical and temporal attributes relevant to student engagement analysis.

These variables capture a range of information, including signup activity, opportunity classifications, geographic distribution, demographic details, and academic preferences.

Key fields identified during the initial inspection include:

- Learner SignUp DateTime : timestamp indicating when a learner registered
- Opportunity Category : type or classification of the offered opportunity
- Country : learner's country of residence or origin
- Gender : self-reported gender information
- Current/Intended Major : academic discipline currently pursued or intended

4.2 Data Preprocessing:

The dataset underwent a comprehensive cleaning process to ensure consistency in date formats, standardization of institution names, removal of invalid or irrelevant records and verification of data uniqueness.

The cleaning process focused on ensuring data integrity, consistency, and reliability. Key steps included:

- **Standardizing Date/Time Columns:** All date/time fields were parsed into a consistent format using pd.to_datetime, with invalid entries coerced to NaT. This revealed significant missing values in fields such as Opportunity Start Date (4,637)

and Opportunity End Date (1,262). Non-standard formats (e.g., invalid time values) were correctly flagged as missing.

- **Normalizing Institution Names:** Variants of Saint Louis University were standardized to a single, consistent format to ensure accurate grouping and analysis.
- **Handling Invalid & Missing Values:** Blank or invalid Opportunity End Dates and negative Opportunity Duration values were removed.
- **Removing Irrelevant Records:** Test entries and malformed names in the First Name column were identified and removed to maintain data relevance.
- **Duplicate Check:** No duplicate records were found based on key identifier fields, indicating strong data uniqueness.

4.3 Feature Engineering:

In Week One, one of the tasks provided was feature engineering. Feature engineering involves using data to create additional features that aren't directly observable. Google Sheets formulae were used to transform the data in other columns into readable statistics that could be used to gain further insights.

Some examples of new features were an engagement score that predicted how likely a user was to commit, the season the user applied, the number of days committed, and more. In particular, the opportunity start and end dates and the user's signup and application dates were considered in feature engineering due to their relation to a user's longevity in the opportunity. By making formulae with these columns, additional data was generated regarding such topics. Furthermore, user status descriptions were scrutinized to validate or disprove original longevity analyses. If a user was rejected, withdrew, or dropped out from a program, they were considered to have churned, making their engagement and commitment illegitimate.

These original features were crucial to the feature engineering process because they were related to the target variables, user churn and engagement. By using the five mentioned columns, multiple engagement and churn-related features were derived to assist with churn analysis and future parts of the internship. Another main aspect of feature engineering is making features that don't depict new data but existing data instead. A primary example of feature engineering and transforming existing features is one-hot encoding. Through one-hot encoding, binary values were assigned to Gender (Male/Female), Opportunity Category, Application Season, Age Group

5. Exploratory Data Analysis (EDA):

For Exploratory Data Analysis, the following methods were employed to reach the ultimate goal of gaining insights into patterns such as trends and correlation. First, data was preprocessed. Despite rigorous efforts to clean the data in Week One, the original cleaning phase, there were still issues. In order to avoid incomplete or inaccurate results, additional cleaning was conducted. Once preprocessing was completed, the cleaned data was sent to the visualization team. The visualization team's purpose was to create graphs and visuals highlighting evident patterns within the statistics. The data was exported as a CSV file and imported using pandas.

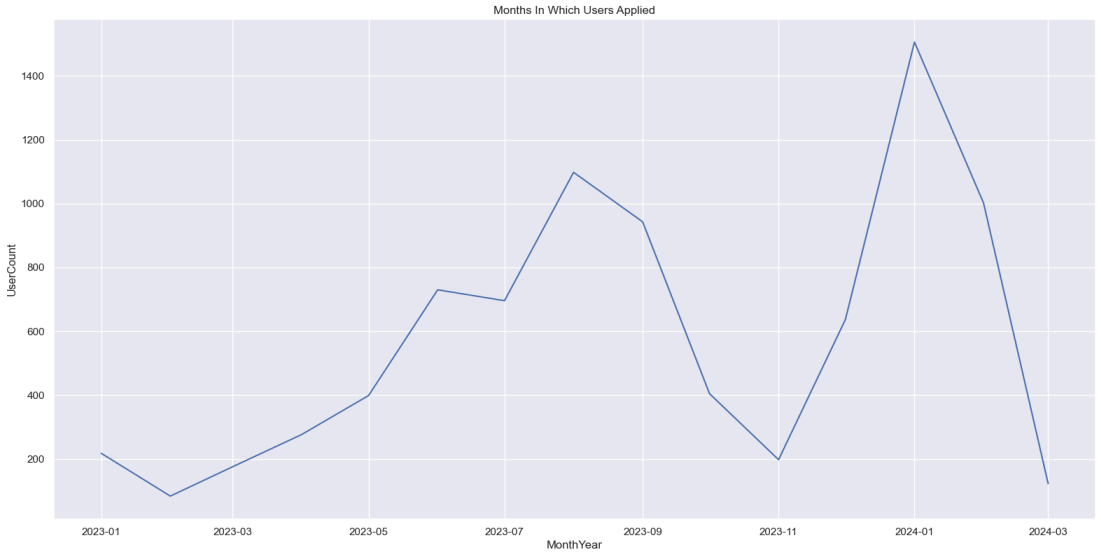
Afterward, members used Seaborn, the visualization module, to create graphs based on various columns. Seaborn uses Matplotlib to build graphs while making them visually pleasing and easy to read. Therefore, these advantages made it an optimal choice. These graphs varied from box plots and count plots to correlation heatmaps and pairplots, and software varied between Visual Studio Code and Jupyter Notebooks. A variety of graphs were used depending on whether the team member was searching for patterns in correlation, trend, or something else. Work was divided among the three visualization members to prevent creation of the same graph multiple times. Ayad was tasked with creating graphs based on original columns, and Jacob focused some of his graphs on added features. The added features were built onto the cleaned data during Week One, and they were designed with the intent to capture user engagement and commitment. Maryam created a variety of graphs concerning all of the data's aspects. By dividing work among team members, the team produced many graphs while simultaneously averting missing graphs.

After visualization, hypotheses were derived from the graphs. A seasonal distribution graph indicated a higher volume of winter applications. Based on this trend, the hypothesis team proposed that application rates increase in winter due to reduced outdoor activity opportunities. These hypotheses shaped the modeling process during Week Three, which was used to further analyze patterns through AI models. Hypotheses were created to provide Excelerate with information about user retention and provide a foundation for reporting user engagement and churn analysis. Additionally, extra user statistics could be analyzed to inspect similar patterns. During Week Two EDA, visualization members chose features, and they proceeded to make graphs based on them. However, during the EDA required for Week Three's churn analysis, useful features were chosen and distributed for members to focus on. This methodology proved to be effective because multiple high-quality graphs were created when members knew the best plan of action.

5.1 Key Visualizations / Charts

Through EDA, our group created a series of key visuals that drove the hypotheses behind our recommendation system, particularly with user spikes during certain periods. The key visualizations show that users are likely to apply during the winter and summer months, with large spikes during these times. These provide opportunities for heavy marketing.

5.1.1. Users per Month

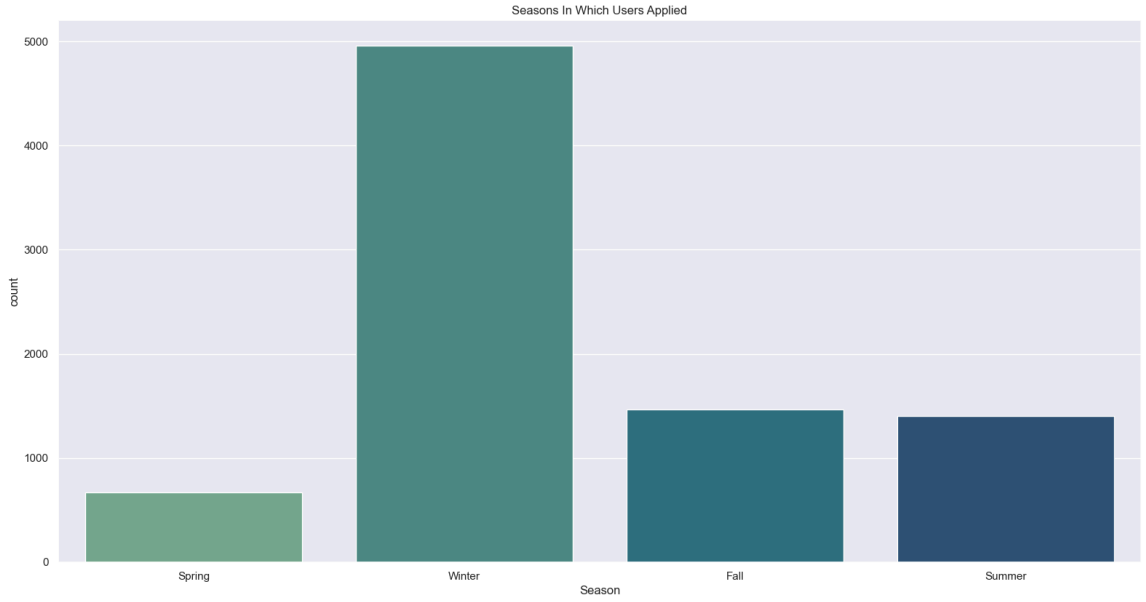


Looking at the graph, there are spikes of 1000 - 1500 applicants at key time periods followed by quick dropoffs of only ~200 applicants

This means that there are periods of motivation/interest for users followed by periods of inactivity across the platform, leading to temporary losses for Excelerate after such a spike.

Excelerate should invest heavily in marketing during such engagement booms in order to capitalize from these events.

5.1.2 Applicants by Season

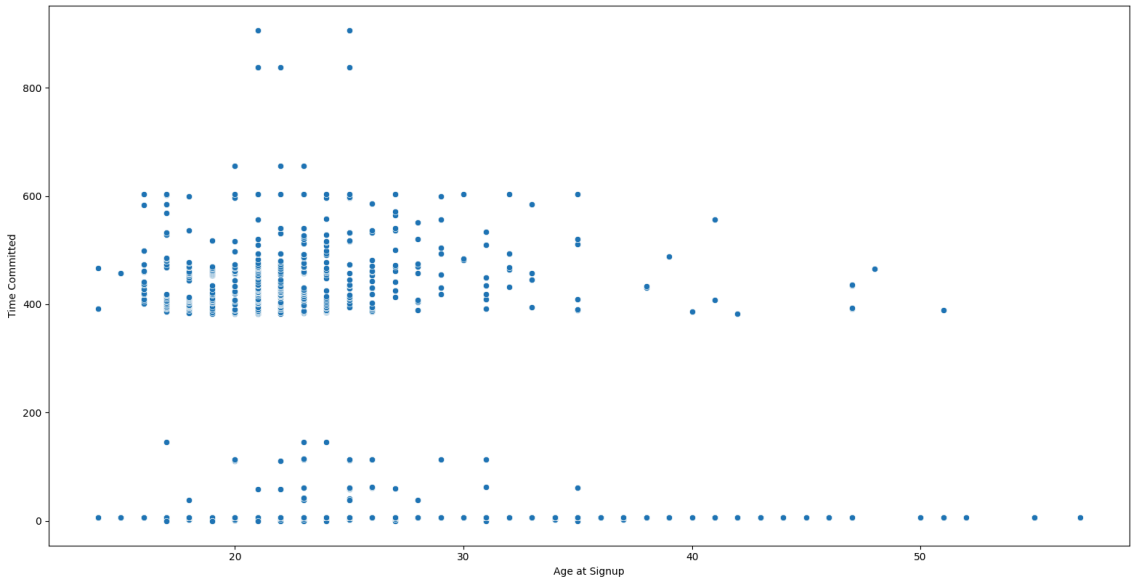


The count plot breaks down applicants by the season they applied in, revealing a massive amount of applicants in the winter compared to minimal applicants in the spring.

Fall, spring, and summer show moderate numbers of applicants from ~800 to 1500, while the winter shows an unbelievable spike in applicants around 5000.

Based on the graph's data, the winter is an imperative time for Excelerate to fill its opportunities with new applicants, and it must be taken advantage of to the fullest extent.

5.1.3 Commitment by Age



The graph depicts a user's commitment compared with their age. There is a noticeable number of users aged 16-30 with 400-600 days of commitment, and the data points grow sparser as the graph progresses to the right.

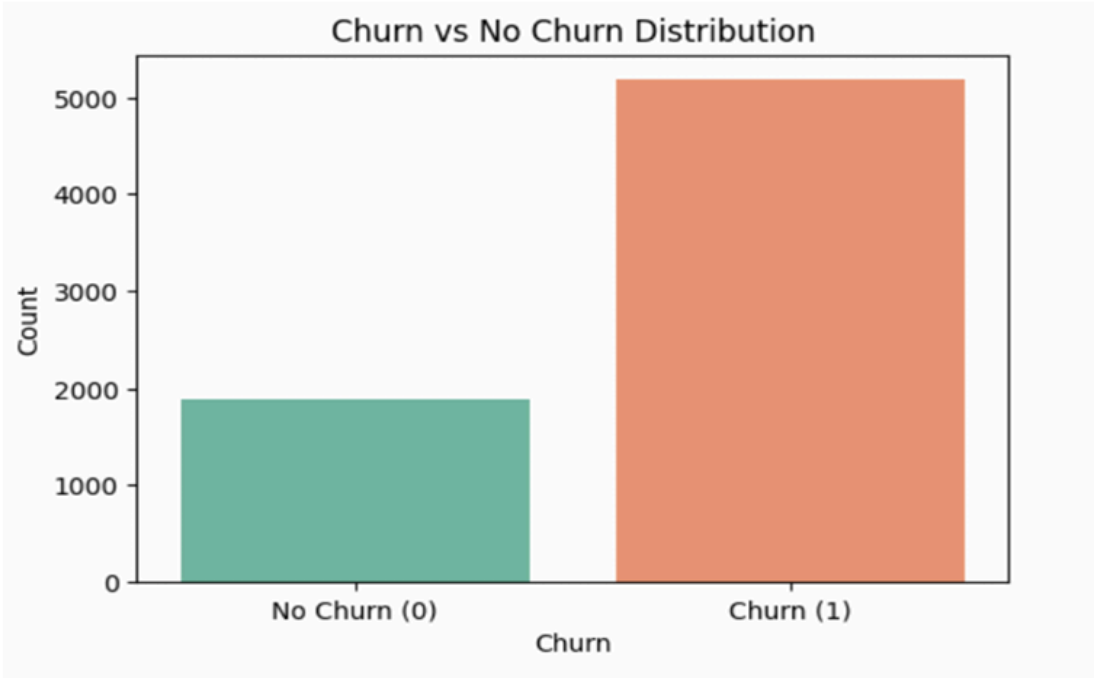
Younger individuals are looking for experience and opportunities to grow skills, leading them to be more committed to Excelerate's programs. Older individuals, on the other hand, are likely doing Excelerate's programs for fun and less for the sake of their resumés.

Excelerate should target the younger age group by marketing to universities and other places filled with young people looking for opportunities.

5.2 Churn Analysis:

The churn analysis reveals a critically high 70.32% user churn rate, with consistent patterns across genders and a potential link between shorter opportunity durations and higher churn. This indicates urgent retention challenges requiring broad, experience-focused interventions.

5.2.1. Overall Churn Distribution (Bar Plot)

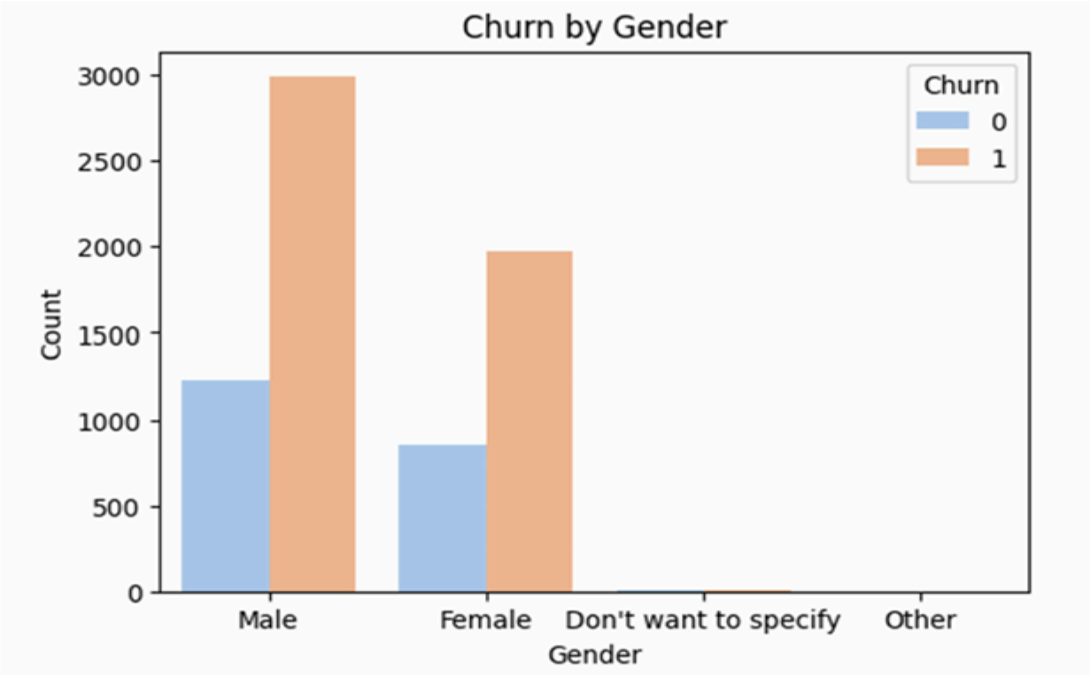


Looking at the churn distribution, the picture is alarming. 70.32% of users (4,965 people) have churned, while only 29.68% (2,096) remain active.

This means that for every 10 users we acquire, 7 leave. Such a steep drop not only impacts growth but also means our customer acquisition costs may exceed lifetime value.

To ensure fair model training despite this imbalance, we applied SMOTE on the training data.

5.2.2. Churn by Gender (Count Plot)

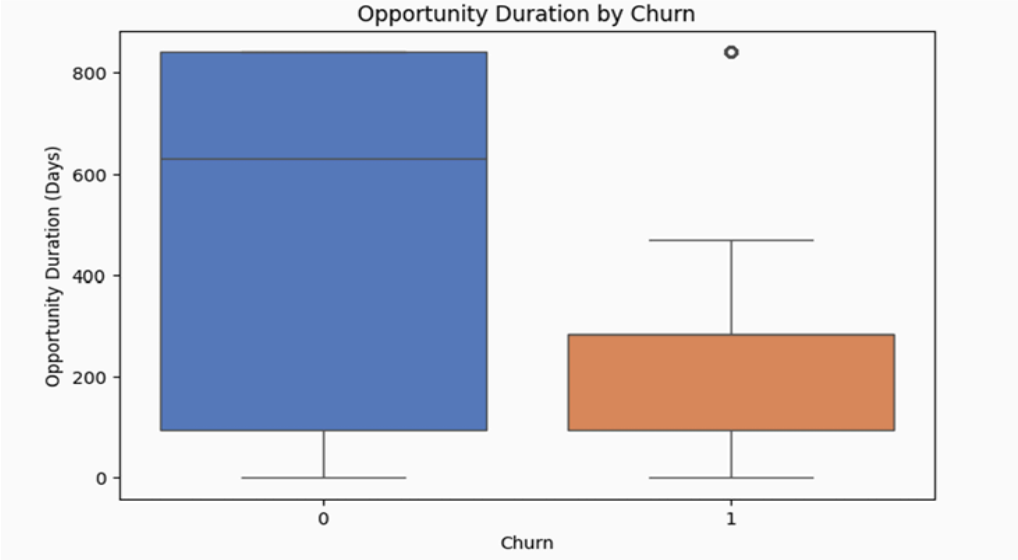


The count plot breaks down churn by gender, revealing a surprising finding: both male and female users churn at almost identical rates.

There is no visible gender bias in the retention problem; the high churn affects everyone equally.

This tells us that gender-targeted marketing won't solve the core issue, the challenge is more fundamental, rooted in overall user experience and platform value.

5.2.3. Opportunity Duration by Churn (Box Plot)



When we compare opportunity duration between churned and retained users, an interesting trend emerges: shorter programs may be more prone to churn.

- Shorter programs: May not deliver enough sustained value to keep users invested.
- Longer programs: Require ongoing engagement to prevent drop-offs midway.
- This insight opens the door for optimization strategies, such as:
- Adjusting program durations based on churn patterns.
- Adding milestone-based engagement points for longer programs.
- Incorporating progress tracking and celebration features to keep learners motivated.

6. Modeling

To predict student churn, we evaluated multiple machine learning models using demographic, behavioral, and opportunity-related features. Random Forest delivered the best overall performance, while LightGBM achieved the highest ROC AUC, both proving superior to Logistic Regression. Feature importance analysis revealed that application timing, commitment level, and repeated applications are the strongest churn predictors, guiding actionable retention strategies.

6.1 Objective

Our goal was clear: predict student churn (Churn = 1) versus retention (Churn = 0) using patterns found in the data. This involved testing multiple models, ensuring balanced datasets, and identifying the most impactful features for intervention.

6.2 Modelling Strategy

We began by selecting the most relevant features from our cleaned dataset, guided by domain knowledge and EDA findings.

Data Preparation:

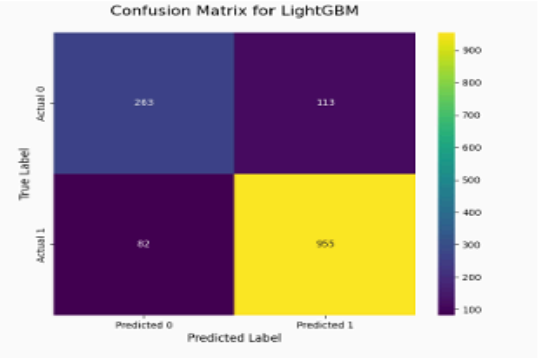
1. One hot encoded key categorical features like Opportunity Category, Application Season, and Age Group.
2. Removed high multicollinearity features using Variance Inflation Factor (VIF) analysis.
3. Balanced the dataset with SMOTE to address the 70:30 churn imbalance.
4. Scaled features where required (e.g., for Logistic Regression).
5. Train-Test Split: 80-20 stratified split to preserve churn ratios in both sets.

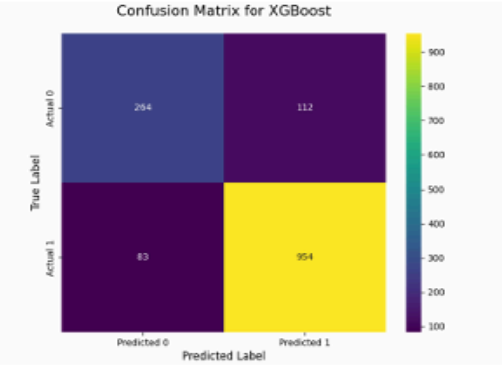
Models Tested:

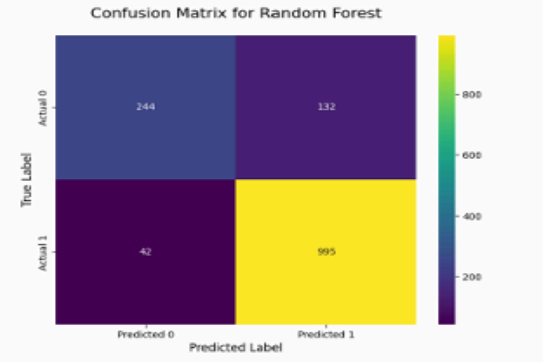
1. Logistic Regression (linear baseline)
2. Random Forest Classifier
3. XGBoost Classifier
4. LightGBM Classifier

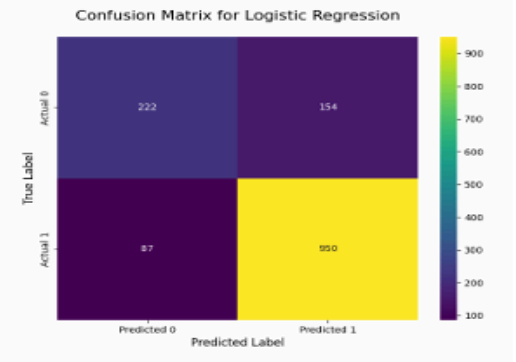
6.3 Model Performance & Insights

Confusion Matrices Here

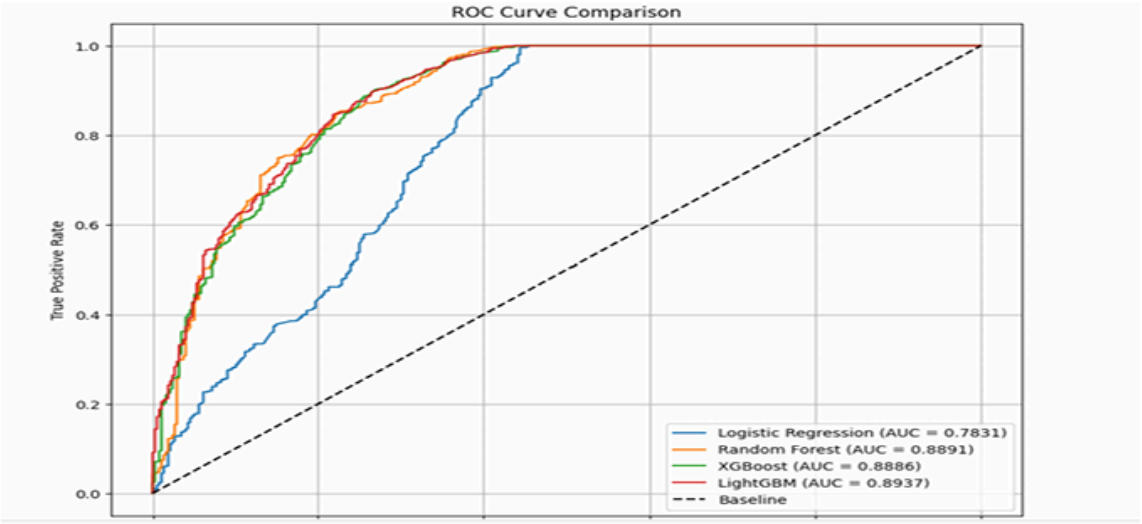








ROC Curve Comparison Chart



We assessed models using Accuracy, Precision, Recall, F1-Score, ROC AUC, and Confusion Matrices.

1. Random Forest:

Highest Accuracy, F1-Score, and Recall.

Balanced predictions with few false positives and negatives.

2. LightGBM:

Slightly lower Accuracy than Random Forest but highest ROC AUC, showing excellent ranking ability.

3. XGBoost:

Similar strong performance to Random Forest and LightGBM, outperforming Logistic Regression.

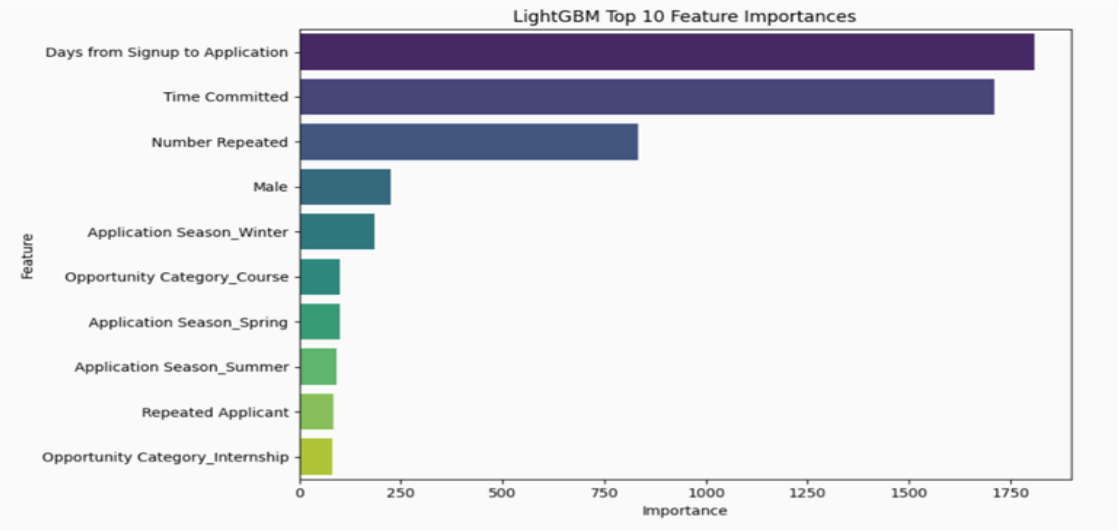
4. Logistic Regression:

Still strong but less effective on this non-linear dataset. Slightly higher misclassifications.

Key takeaway: Tree-based models consistently outperform linear models, indicating non-linear relationships in churn prediction.

6.4 Feature Importance – What Drives Churn

Feature Importance (Based On LightGBM)



Across Random Forest, XGBoost, and LightGBM, the same features emerged as top predictors:

- **Days from Signup to Application**: Longer delays strongly correlate with higher churn.
- **Time Committed:** More commitment hours reduce churn risk.
- **Number Repeated:** Multiple applications influence retention likelihood.

6.5 Recommendations for Action

1. **Targeted Interventions:** Use Days from Signup to Application to identify at-risk students early; send timely reminders or nudges.
2. **Boost Engagement:** Increase Time Committed and promote Committed User behaviors via gamified learning, community events, or milestone rewards.
3. **Real-Time Prediction:** Deploy Random Forest or LightGBM to flag churn risks dynamically and enable proactive outreach.

7. Insights and Recommendations

Based on the data analysis, the following actionable insights and recommendations were developed to improve student engagement and retention:

- **Seasonal Campaigns:** Launch targeted marketing campaigns during seasonal peaks (January, July–August) to capitalize on high application rates.

- **Early-Bird Incentives:** Offer incentives for early sign-ups to reduce churn, as early engagement is a key predictor of success.

- **Churn Risk Alerts:** Deploy an internal system that generates alerts for inactive learners or those with incomplete profiles, allowing for proactive intervention.

- **Targeted Support:** Provide female-targeted support and resources, especially in identified urban hotspots, to boost participation and success rates.

- **Program Promotion:** Promote hybrid interdisciplinary programs, as tech-heavy majors dominate the platform, indicating a strong interest in this area.

8. Recommendation System

8.1 Methodology:

A hybrid recommendation system was developed to provide personalized and relevant suggestions to students. The system's purpose is to enhance engagement by matching learners with suitable opportunities, thereby reducing the likelihood of drop-offs.

- **Content-based filtering:** This component matches new opportunities to learners based on their academic background and engagement history.
- **Collaborative filtering:** This component recommends opportunities based on similarities between different learners' behaviours and preferences.
- **Hybrid Model:** By combining both approaches, the system provides highly personalized and accurate recommendations.

8.2 Benefits:

- **Increased retention and satisfaction** by helping learners find opportunities that are a better fit.
- **Scalable personalization** that adapts as new data arrives, ensuring the system remains relevant over time.

9. Conclusion and Future Work

This project delivered a data-driven retention framework that produces tangible, measurable results. Within the first 6–12 months, the model is projected to retain over 1,200 additional students, protecting millions in potential revenue and improving long-term engagement.

Operational processes have been streamlined—reducing application delays from more than a week to just three to four days—while automated,

Strategically, we have embedded predictive analytics into Excelerate's decision-making, shifting from reactive problem-solving to proactive prevention.
This approach optimises resource allocation, reduces support costs, and positions Excelerate as a leader in analytics-driven education.

In essence, we have built more than a churn analysis tool—we have created an operational model that safeguards revenue, enhances student experience, and strengthens competitive positioning.