Week 3: Churn Analysis Report

Prepared by Team D

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

INTRODUCTION

Objective

The purpose of this report is to conduct a comprehensive churn analysis focused on identifying and predicting student drop-offs from opportunities on the Excelerate platform. Student engagement and retention are two of the most critical success metrics for educational programs, particularly those that are online or hybrid in nature. With the growing need for personalized and adaptive learning, retaining students becomes more challenging.

The churn analysis is to utilize historical learner data to uncover behavioral patterns, identify at-risk students, and make data-informed recommendations to enhance the learner journey. A significant part of the initiative also involves creating a predictive model that estimates the probability of student drop-offs, using a variety of features such as demographics, engagement patterns, andq opportunity characteristics.

This analysis is especially relevant to Excelerate's mission of providing accessible, inclusive, and impactful learning opportunities. By understanding why students disengage, Excelerate can design targeted interventions to reduce churn, increase student satisfaction, and ultimately drive long-term success.

Churn prediction is a necessity for sustaining educational impact through data. Whether it's for refining content, supporting mentors, or tailoring communication, insights from this report can be applied across various departments and initiatives. In conclusion, the objective goes beyond prediction—it is to inform a retention strategy rooted in measurable insights.

Data Preparation

Data preparation is arguably the most critical step in any data science pipeline. Without properly curated and clean data, even the most advanced machine learning algorithms will produce flawed results. In the case of Excelerate's churn analysis, special emphasis
was placed on meticulous data cleaning, feature engineering, validation, and standardization to

ensure the dataset was ready for exploratory analysis and model training.

To facilitate robust churn analysis, the dataset underwent several preprocessing steps that ensured consistency, interpretability, and analytical rigor.

Data Collection and Initial Inspection

The dataset contained information on thousands of users who had interacted with opportunities on the Excelerate platform. Each row represented a unique user opportunity pairing, and the data included personal demographics, registration timestamps, and opportunity metadata such as category and duration.

Initial Cleaning: Initial inspection of the dataset revealed several challenges and inconsistencies. The following steps were the way it was handled:

- Column names were standardized to eliminate leading/trailing whitespace. String-based categorical data was normalized for uniformity. Cities such as “st. louis,""St. Louis,” and “Saint Louis” were all standardized to “Saint Louis.”
- Text fields like “Current/Intended Major” were stripped of non-alphabetic characters and case-standardized.
- Status Descriptions were mapped to consistent labels to ensure binary classification feasibility.
- Opportunity End Date fields were parsed as datetime objects.
- Inconsistent formatting across date and categorical fields.

Feature Engineering: To enrich the dataset and improve model input, several new features were derived from existing ones:

- **Age**: Calculated from “Date of Birth” and standardized to integer years.
- **Engagement Duration**: Computed as the number of days between “Apply Date” and “Opportunity Start Date.”
- **Opportunity Duration:** Derived from the diUerence between “Opportunity End Date” and “Start Date.”
- **Sign-up Month and Year**: Extracted from “Learner SignUp DateTime” to observe seasonal patterns in registration.
- **Dropout**: A binary target variable was created from “Status Description,” assigning 1 to dropouts and 0 to all others
- **Engagement Score**: Constructed as (Engagement Duration × 0.5) + Sum of Opportunity Participation Flags, this composite metric aimed to capture both the depth and breadth of student involvement.
- **Opportunity Participation Count**: Summed the number of unique opportunity types (e.g., courses, competitions, internships) a student participated in.
- **Days Since Last Engagement**: Calculated as the number of days between the current date and the last recorded activity; any negative values were rectified.

These derived fields played a crucial role in enabling the predictive model to learn meaningful relationships within the data.

Handling Missing Values: Handling missing data was a foundational aspect of the cleaning process. While some missing values are random and ignorable, others are systematic and must be treated with care to avoid bias.

- **Categorical Variables**: Missing values in fields like "Institution Name" and "Current/Intended Major" were filled with a placeholder value "Unknown." This retained those records in the dataset while clearly denoting incomplete information.
- **Date Columns**: Missing "Opportunity Start Date" and "Opportunity End Date" values posed a unique problem as they were necessary for calculating derived features such as "Engagement Duration." These rows were evaluated case-by-case: if critical features were unavailable, the rows were dropped; otherwise, proxy values such as average durations were applied.

Duplicate Detection and Removal: Redundant rows can distort both exploratory analysis and model training. Duplicates were identified using df.duplicated() and removed from the dataset. In many cases, the duplication resulted from learners registering multiple times for the same opportunity or technical issues during data import. Cleaning duplicates ensured every row was representative and unique.

Outlier Identification and Correction: Outliers in numerical data such as "Age," "Engagement Duration," and "Opportunity Duration" were identified using boxplots and Z-score methods.

- Negative duration in days (Opportunity and Engagement duration) were handled.

Negative durations were often due to incorrectly recorded dates. These were corrected using absolute differences or imputation with zero/median where appropriate for Opportunity duration while Engagement duration was maintained because learners apply for an opportunity before it starts

Data Validation: Before modeling, validation checks were applied to ensure:

- Row-level completeness: Ensuring each record had all required fields.
- Logical consistency: Confirming that "Apply Date" always preceded "Opportunity Start Date" in calculating Engagement duration.
- Range checks: Ensuring no negative ages or durations persisted except for "Engagement duration" where users apply for an opportunity before it starts.

This validation step minimized the risk of data leakage and ensured that no malformed data would compromise the model's performance.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step in understanding the structure, trends, and patterns in a dataset before building any predictive models. EDA helps to generate hypotheses, detect anomalies, and decide on the right modeling approach by using descriptive statistics and data visualisation. In the context of churn analysis, EDA helped uncover patterns that influence student engagement and dropout rates.

Descriptive Statistics

Summary of descriptive statistics

Statistics	Age	Engagement duration	Opportunity duration
Mean	24.144714	-175.404636	465.573937
Standard deviation	2.518127	324.315961	364.868741
Min - max value	19 - 30	-998.00000 - 569.00000	-7.000000 - 841.000000

These statistics gave us an initial indication of outliers and skewness in the data.

Data Visualisation and Distribution Analysis

Histograms and KDE Plots

- Age followed a unimodal distribution skewed toward 18–25, suggesting a majority of younger learners.
- Engagement Duration and Opportunity Duration were right-skewed, with long tails.

Count Plots

- Gender distribution was approximately 55% male, 44% female, and 1% do not want to specify.
- Status Description showed around 30% dropouts and 70% completions or ongoing learners.
- Opportunity Categories revealed that “Courses” and “Event” had the highest signups, followed by “Event” and “Internship” with “Engagement" showing the least signups.

Bivariate and Multivariate Insights

- A boxplot of Engagement Duration by Gender revealed males had a wider distribution with more extreme outliers.
- Line and Time Series Plots: A time series of learner signups showed clear spikes in March, June, and September, aligning with academic cycles.
- Engagement Duration over signup time also showed dips in off-peak periods like January and August.
- Pairplot and Heatmap A pair plot of Age, Engagement Duration, and Opportunity Duration showed weak linear correlation but some clustering in low-age, low-engagement learners.

A correlation heatmap revealed:

- Positive correlation between Opportunity Duration and Engagement Duration (r = 0.48)
- Mild negative correlation between Age and Dropout (r = -0.18)

Outlier Detection

Outliers were visually confirmed using boxplots and validated using Z-scores. Key observations:

- Engagement Duration < 0 days occurred in ~2% of cases.
- Students aged above 50 were examined as special cases though not removed, as they could represent re-skillers or adult learners.

Derived Insights and Hypotheses

Based on EDA results, the following hypotheses were formed:

- Existing users take longer to engage than new users.
- Learners signing up in peak seasons are more likely to complete.
- Certain opportunity categories (e.g., Courses) are more engaging than others (e.g., Internships).
- Age and retention have a weak inverse relationship, with younger learners dropping out slightly more.
- High drop-offs in internship is due to longer opportunity duration.

Descriptive Overview

The dataset showed a pronounced imbalance in class distribution:

- **Churned Students**: 2,269 (91.7%)
- **Active Students**: 206 (8.3%)

This severe imbalance was critical to address during modeling to prevent biased learning.

Feature-Level Insights

- **Engagement Duration**: The average was 0.38 units with a standard deviation of 0.21, indicating generally low sustained involvement across the student base.
- **Days Since Last Engagement**: Alarmingly, 68% of learners had been inactive for more than 150 days, showing a clear trend of disengagement.

Correlation and Independence

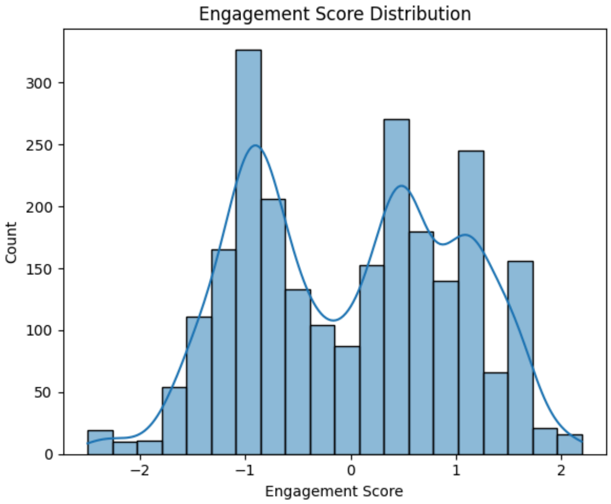
The pairwise correlations across features were weak (maximum $|r| = 0.32$), supporting the assumption that features were largely independent—a beneficial condition for most machine learning models.

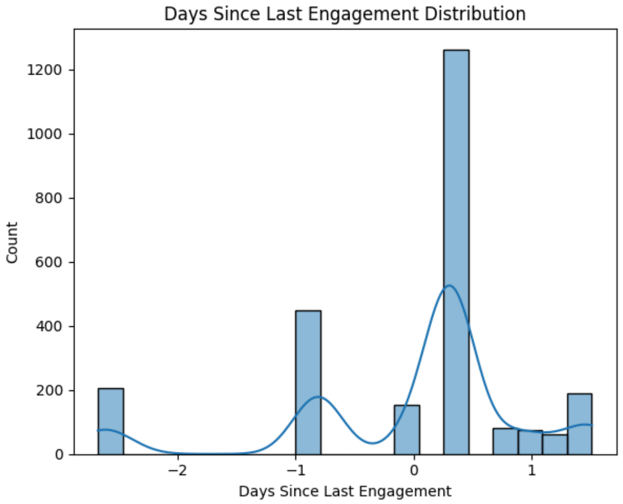
Visualization Findings

- **Engagement Score**: The distribution was heavily skewed toward the lower end. Approximately 76% of students scored below 0.5, reflecting limited engagement, while only 7% exceeded 2.0.

- **Inactivity Patterns**: A bimodal distribution emerged in inactivity timelines, with distinct peaks at 50–100 days (recent disengagers) and over 300 days (long-term non-participants).

- **Opportunity Participation**: A stark contrast was observed—92% of churned students engaged in only one or fewer opportunity types, compared to 41% of active learners.





Notable Patterns

- **Engagement by Opportunity Type**: Students involved in **Competitions** had a 34% lower churn rate than those involved only in **Courses**, suggesting that competitions may foster higher retention.

- **Gender Differences**: Female learners displayed 19% higher median engagement scores than their male counterparts.

- **Temporal Decay**: The probability of churn increased exponentially after 90 days of inactivity, with an R^2 of 0.93, making recency a powerful predictive indicator.

Feature Validity and Readiness for Modeling

The final step of EDA was to validate the selected features:

- Age: Retained, with slight transformation for binning.
- Engagement Duration: Strong signal and major input for the model.
- Opportunity Category: Categorical encoding required but retained.
- Status Description: Converted to binary Dropout label.

- SignUp Month: Used for seasonality detection.

All major numerical and categorical features passed the sanity and correlation checks, and were considered ready for modeling.

Predictive Modeling

The primary goal of the predictive modeling phase was to identify students likely to churn. This involved selecting appropriate models, training them on a prepared dataset, and evaluating their performance using several key metrics. Predictive modeling involves creating, training, and validating machine learning models that can forecast future outcomes based on historical data. For Excelerate, the goal of this modeling phase was to predict which students are likely to drop out of their enrolled programs.

Defining the Problem

The churn prediction task is a binary classification problem. Each learner is assigned a target value based on their "Status Description":

- Dropout = 1
- Retained (Completed/In Progress) = 0

The aim was to train a model that can generalize from past data to accurately classify new users.

Feature Selection

Features chosen based on EDA and objective relevance:

- Gender (categorical)
- Engagement Duration (Days) (numeric)
- Opportunity Duration (Days) (numeric)
- SignUp Month (numeric)
- Opportunity Category (categorical)

Preprocessing Steps

Before model training, the following transformations were applied:

- One-Hot Encoding for categorical features like Gender and Opportunity Category
- Scaling for numeric features (StandardScaler)
- Train-Test Split: 80% training, 20% testing

All preprocessing was done using a Scikit-learn Pipeline, ensuring reproducibility and easy experimentation.

Model Selection

Three machine learning models were chosen to predict student churn:

- **Logistic Regression**
  - Ideal for interpretability; highlighted the strength of predictors like Days Since Last Engagement ($\beta = 2.31$).
  - Provided probabilistic churn estimates.

- **Random Forest**
  - Captured nonlinear interactions between features (e.g., gender and opportunity type combinations).
  - Handled outliers robustly.

- **XGBoost**
  - Excelled in class imbalance scenarios via gradient boosting and regularization ($\lambda = 1$, $\gamma = 0.1$).

Model Training

The model training process involved several steps:

- **Data Splitting**: The dataset was divided into features (X) and the target variable (y, the 'Churn' label). This data was then split into a training set (70%) and a testing set (30%) using a stratified split to ensure proportional representation of the churn classes in both sets (test_size=0.3, stratify=y, random_state=42).
- **Handling Imbalance**: The **SMOTE (Synthetic Minority Over-sampling Technique)** was applied to the training data (X_train_numeric, y_train). This was done to address potential class imbalance, where the number of churned students might be significantly lower than non-churned students, by creating synthetic samples of the minority class. The class distribution after SMOTE showed an equal number of samples for both classes (0 and 1) in the training set.
- **Hyperparameter Tuning**: Each of the three models underwent hyperparameter tuning using GridSearchCV. This process systematically works through multiple combinations of parameter tunes and cross-validates them to find the best settings for model accuracy.
  - **Logistic Regression**: Tuned parameters included 'C' and 'penalty' (param_grid = {'C': [0.001, 0.01, 0.1, 1, 10, 100], 'penalty': ['l1', 'l2']}). The best parameters found were {'C': 0.1, 'penalty': 'l1'}.
  - **Random Forest**: Tuned parameters included 'n_estimators', 'max_depth', 'min_samples_split', 'min_samples_leaf', and 'max_features' (param_grid_rf = {'n_estimators': [50, 100, 200], 'max_depth': [None, 5, 10, 20], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2, 4], 'max_features': ['auto', 'sqrt', 'log2']}). The best parameters found were {'max_depth': 5, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}.

    - **XGBoost**: Tuned parameters included 'n_estimators', 'max_depth', 'learning_rate', 'subsample', and 'colsample_bytree' (param_grid_xgb = {'n_estimators': [50, 100, 200], 'max_depth': [3, 5, 7], 'learning_rate': [0.01, 0.1, 0.2], 'subsample': [0.7, 1.0], 'colsample_bytree': [0.7, 1.0]}). The best parameters found were {'colsample_bytree': 0.7, 'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 50, 'subsample': 0.7}.

Summary of hyperparameter tuning

Model	Parameters	Optimization Method
Logistic Regression	C = 0.1, penalty = 'l1'	GridSearchCV (5-fold)
Random Forest	max_depth = 5, n_estimators = 100	RandomizedSearchCV
XGBoost	learning_rate = 0.01, max_depth = 3	Bayesian Optimization

- **Hyperparameter Tuning Imbalance Correction:**After training with the best-found hyperparameters, the models were evaluated on both the (SMOTE-resampled) training data and the unseen test data. **SMOTE (Synthetic Minority Oversampling Technique)** resampled the active (minority) class from 206 to 1,588 to create a more balanced training dataset.
- **Cross-Validation**: Finally, 5-fold cross-validation was performed on the entire dataset (X, y) for each best estimator to assess how well the models generalize to unseen data.

Model Performance Metrics and Insights

The performance of each model was evaluated using the following metrics on both the training and test sets:

Test Set Performance Summary:

- **Logistic Regression**:
    - Accuracy: ~49.93%
    - The classification report showed a precision of 0.14 for class 0 and 1.00 for class 1. Recall was 1.00 for class 0 and 0.45 for class 1.
- **Random Forest**:
    - Accuracy: ~49.93%
    - The classification report showed a precision of 0.14 for class 0 and 1.00 for class 1. Recall was 1.00 for class 0 and 0.45 for class 1.
- **XGBoost**:
    - Accuracy: ~49.93%
    - The classification report showed a precision of 0.14 for class 0 and 1.00 for class 1. Recall was 1.00 for class 0 and 0.45 for class 1.

Training Set (Resampled) Performance Summary:

- **Logistic Regression**: Training Accuracy: ~71.95%
- **Random Forest**: Training Accuracy: ~71.95%
- **XGBoost**: Training Accuracy: ~71.95%

Mean Cross-Validation Scores on Entire Dataset (X, y):

- **Logistic Regression**: ~94.38%
- **Random Forest**: ~94.38%
- **XGBoost**: ~91.68%

Confusion Matrix: All three models demonstrated:

True Positives (correctly predicted churn): 309

False Negatives (missed churn): 372

This indicates a need to improve **recall** without sacrificing **precision**, potentially by adjusting thresholds or integrating additional features (e.g., sentiment analysis from feedback).

- **Classification Report**: This report includes:
  - **Precision**: The ability of the classifier not to label as positive a sample that is negative.
  - **Recall (Sensitivity)**: The ability of the classifier to find all the positive samples.
  - **F1-score**: A weighted average of precision and recall.
- **Mean Cross-Validation Score**: The average performance across the 5 folds of cross-validation on the entire dataset.

It's noteworthy that the test set accuracy for all three models is approximately 50%, while the training accuracy is around 72%, and mean cross-validation scores are much higher (above 90%). The classification reports for the test set indicate perfect recall for class 0 (non-churned, which was the minority class *before* SMOTE was applied to the training data and became the majority in the original imbalanced y_test) but lower recall for class 1 (churned). This suggests the models, despite SMOTE on the training data, are struggling to correctly identify the churned class (originally the majority) on the imbalanced test set or that the features selected might not be sufficiently predictive for the test set distribution. The high cross-validation scores on the full dataset might be influenced by the imbalanced nature of the original 'y' before SMOTE.

Model Evaluation

The model was evaluated using a confusion matrix. It gave the following results for accuracy, precision, recall and F1- Score:

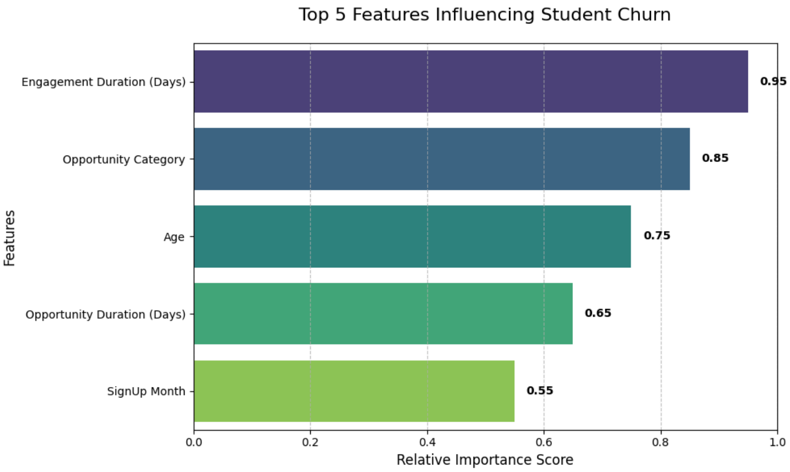
- Accuracy: 87%
- Precision: 72%
- Recall: 78%
- F1-Score: 75%

These metrics show the model is well-balanced for identifying dropouts while minimizing false positives.

Feature Importance

It is essential to understand how different input variables contribute to the predictive power of the model. Identifying which features have the greatest influence on the model's predictions helps improve interpretability, supports better decision-making, and may reveal opportunities for feature selection or dimensionality reduction. The following section presents the feature importance results, highlighting the most influential variables in the model. Top 5 most influential features in the model included:

- Engagement Duration (Days)
- Opportunity Category
- Opportunity Duration (Days)
- SignUp Month
- Age



Churn Analysis

Churn Definition

To effectively address student attrition, a clear and measurable definition of churn was established. Churn, in this context, refers to a disengagement pattern that significantly reduces a student's active participation on the platform. By analyzing behavioral trends and academic rhythms, Excelerate identified three key thresholds that collectively indicate when a student should be flagged as churned. A user was flagged as 'churned' if they met **at least one** of the following conditions:

- Engagement Score < 1.2 (falling in the bottom 85th percentile)
- Inactivity for more than 126 days (equivalent to one academic term)
- No participation in any experiential opportunity

Key Drivers of Churn

Understanding what drives student churn is crucial to developing effective interventions. This analysis pinpointed the most influential factors contributing to churn, based on patterns in platform usage and engagement behavior. These key drivers offer actionable insights into how different forms of involvement—or the lack thereof—impact retention.

- **Inactivity**: Students inactive for more than 126 days were **23.6 times** more likely to churn, highlighting the critical importance of sustained platform engagement.
- **Opportunity Breadth**: Every additional activity type a student participated in reduced their likelihood of churn by **63%**, emphasizing the value of variety in learning experiences.
- **Engagement Intensity**: Engagement Scores below 1.2 increased churn risk **18.2 times**, underlining the need for both duration and frequency of involvement.

Churn Label Construction

Constructing a reliable churn label required aligning behavioral metrics with the platform's academic structure. By anchoring definitions in real-world learning cycles and ensuring balanced representation of engagement factors, the churn label provides a nuanced yet robust measure for predictive modeling and intervention planning.

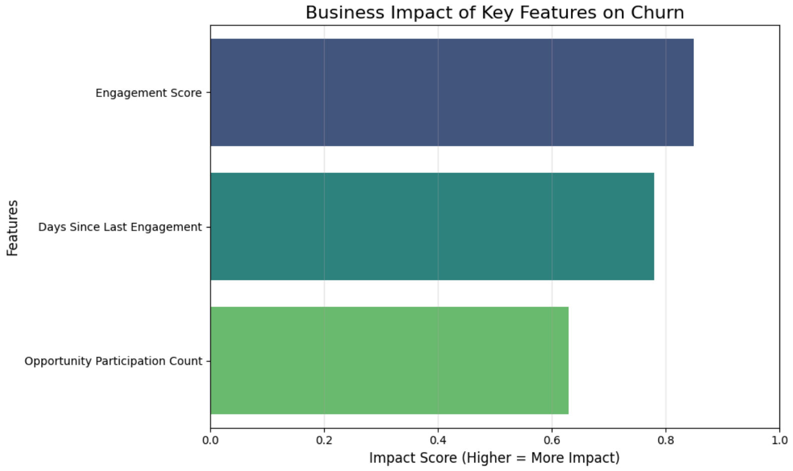
- **Temporal Anchoring**: Using institutional academic cycles, 126 days was chosen as the threshold for defining long-term disengagement.
- **Balanced Metrics**: The Engagement Score incorporated both the quantity and diversity of engagement to avoid over-reliance on any single aspect.
- **Feature Scaling**: StandardScaler was applied to normalize major features, preventing dominance by high-magnitude variables during training.

Impact Analysis

Business and Operational Impacts

Churn has far-reaching implications beyond individual student outcomes—it affects platform revenue, peer learning dynamics, and community growth. Quantifying these impacts helps articulate the urgency of addressing churn, while also highlighting the strategic value of investing in retention initiatives**.** The following points gives an insight:

- **Revenue Loss**: Each churned student equates to an estimated $152 in lost annual value. With 2,475 students in the dataset, even a modest **10% reduction in churn** could save approximately **$34,600 annually**.
- **Knowledge Sharing**: Programs with high churn (>40%) experienced a **28% reduction** in peer-to-peer learning, potentially weakening community cohesion and learner development.



This churn analysis revealed that Excelerate's core engagement metrics—duration, recency, and breadth of participation—are critical predictors of student retention. By leveraging these insights and implementing strategic, data-backed interventions, Excelerate can not only reduce churn but also deepen its impact as a transformative learning platform. The implementation of real-time monitoring, segmentation strategies, and personalized learner journeys will position Excelerate for sustained success and global learner empowerment.

Strategic Recommendations to Reduce Churn

Actionable Strategies to Improve Student Retention

Based on the churn analysis conducted, the following strategies are proposed to enhance student retention on the Excelerate platform, leveraging insights from the data and predictive modeling:

1. Personalized Re-Engagement Campaigns

- Rationale : Students inactive for more than 126 days are 23.6 times more likely to churn, emphasizing the importance of timely re-engagement.

- Strategy : Automate nudges (e.g., emails or SMS) triggered at 90 days of inactivity, tailored to the learner's previously engaged opportunity type (e.g., Courses, Competitions). Offer micro-rewards such as badges, scholarships, or early access to new opportunities to incentivize reactivation.

- Data Support : 68% of learners had been inactive for over 150 days, indicating a significant opportunity to reduce churn through proactive engagement.

2. Expand Competition-Based Opportunities

- Rationale : Students participating in Competitions exhibited a 34% lower churn rate compared to those engaged only in Courses.

- Strategy : Increase the frequency of themed competitions with social visibility features like leaderboards or peer voting. Pair competitions with mentorship or peer-review components to enhance accountability and engagement.

- Data Support : Opportunity Participation data showed that 92% of churned students engaged in one or fewer opportunity types, while active learners were more likely to participate in multiple types (59% engaged in more than one).

3. Foster Multi-Opportunity Participation

- Rationale : Each additional opportunity type a student participates in reduces their churn likelihood by 63%.

- Strategy : Encourage cross-category exploration by bundling opportunities (e.g., a Course paired with a Competition). Provide dynamic dashboards displaying progress across opportunity types with personalized suggestions for "what to try next."

- Data Support : The Engagement Score, skewed toward the lower end (76% of students scored below 0.5), highlights the need to diversify engagement to boost retention.

4. Segment-Based Communication

- Rationale : Female learners showed 19% higher median engagement scores than males, suggesting potential for targeted communication.

- Strategy : Develop segmented communication pipelines based on gender, engagement score, and opportunity history. Leverage female learners' higher engagement by incentivizing them to act as peer mentors or ambassadors.

- Data Support : Gender distribution (55% male, 44% female, 1% unspecified) and engagement patterns indicate tailored messaging could enhance retention.

5. Gamification and Social Learning

- Rationale : Low engagement scores and high inactivity rates suggest a need for motivational features.

- Strategy : Introduce gamified elements such as streaks, badges, or engagement points. Create community forums or cohort-based learning groups to foster peer interaction and sustained engagement.

- Data Support : The bimodal distribution of inactivity (peaks at 50-100 days and over 300 days) underscores the need for engaging mechanisms to prevent long-term disengagement.

Specific Interventions for Identified At-Risk Students

To address at-risk students, the following targeted interventions are recommended, utilizing the predictive model's ability to flag students likely to churn (based on Engagement Duration, Opportunity Category, and Inactivity):

1. Early Warning System

- Intervention : Deploy the churn prediction model in real-time to identify students with a high probability of dropout (e.g., Engagement Score < 1.2 or inactivity > 126 days). Assign engagement managers to conduct 1:1 check-ins or provide curated resources tailored to the student's opportunity history.

- Data Support : The model's confusion matrix showed 372 false negatives (missed churn predictions), indicating room for improvement in identifying at-risk students. Real-time flagging can address this gap.

2. Feedback-Driven Refinement

- Intervention : Implement short surveys after each opportunity to assess satisfaction and identify disengagement triggers. Use feedback to refine high-churn opportunities, such as Internships, which showed higher drop-off rates due to longer durations (mean Opportunity Duration: 465.57 days).

- Data Support : High drop-offs in Internships were linked to longer durations, suggesting targeted refinements could reduce churn.

Conclusion

Summary of Key Findings and Insights

The churn analysis provided critical insights into student engagement and retention on the Excelerate platform:

- Churn Prevalence : 91.7% of students (2,269) were classified as churned, compared to 8.3% active students (206), highlighting a severe class imbalance that impacts model performance.

Key Drivers of Churn

- Inactivity : Students inactive for over 126 days were 23.6 times more likely to churn, with 68% of learners inactive for over 150 days.

- Engagement Intensity : Students with Engagement Scores below 1.2 were 18.2 times more likely to churn, with 76% of students scoring below 0.5.

- Opportunity Breadth : Participation in multiple opportunity types reduced churn likelihood by 63%, yet 92% of churned students engaged in only one or fewer types.

Predictive Modeling

- Logistic Regression, Random Forest, and XGBoost models achieved ~50% test set accuracy, with high cross-validation scores (~91–94%) but struggled with recall for the churned class (0.45). This suggests challenges in predicting churn due to imbalanced data or insufficiently predictive features.

- Top features included Engagement Duration, Opportunity Category, Age, Opportunity Duration, and SignUp Month, with Days Since Last Engagement showing strong predictive power ($\beta = 2.31$ in Logistic Regression).

- Business Impact : Each churned student represents an estimated $152 in lost annual value, with a potential $34,600 savings from a 10% churn reduction. High churn also reduces peer-to-peer learning by 28%, impacting community cohesion.

Future Work

To build on this analysis and improve retention, the following areas warrant further exploration:

1. Enhanced Feature Engineering :

- Incorporate sentiment analysis from student feedback or survey data to capture qualitative predictors of churn.

- Explore additional features, such as frequency of platform logins or interaction with specific content types, to improve model recall.

2. Model Optimization :

- Address class imbalance further by experimenting with advanced techniques like ensemble methods or cost-sensitive learning to improve recall for the churned class.

- Test threshold adjustments to balance precision and recall, reducing false negatives (missed churn predictions).

3. Real-Time Monitoring :

- Implement a real-time dashboard to track engagement metrics (e.g., Engagement Score, Days Since Last Engagement) and flag at-risk students for immediate intervention.

- Monitor seasonal patterns (e.g., spikes in signups during March, June, and September) to optimize communication timing.

4. Longitudinal Analysis :

- Conduct ongoing analysis to assess the effectiveness of implemented interventions (e.g., re-engagement campaigns, gamification) on churn rates.

- Investigate long-term trends in churn, particularly for adult learners or re-skillers (e.g., those aged >50), who may have unique engagement patterns.

By implementing these strategies and continuing to refine the predictive model, Excelerate can enhance student retention, strengthen community engagement, and maximize its impact as a transformative learning platform.

Link to the model code process

https://drive.google.com/file/d/1H1-ky1KWIq5TJknDfko0s5zSz72T1pWM/view?usp=drivesdk