

TEAM 10A WEEK-4 DELIVERABLE

ON

INTERNSHIP COMPREHENSIVE REPORT

Edmar Beatingo
Farwa Fayyaz
Maryam Mohammed
Vishaal Dayashanker

I. INTERNSHIP OVERVIEW

The internship is in the role of an AI Data Analyst, engaging in a comprehensive exploration of data science and machine learning to analyze student engagement data. The focus will be on understanding which opportunities attract the most signups, which are completed successfully, and identifying patterns that lead to drop-offs.

II. INTERNSHIP OBJECTIVE

1. Develop the ability to clean, transform, and prepare raw data for analysis, ensuring it is in the best format for building models.
2. Learn to explore and visualize data to uncover patterns, relationships, and insights that inform the modeling process.
3. Gain expertise in building, training, and validating models that predict outcomes based on historical data, using techniques like regression, classification, and machine learning algorithms.
4. Learn to interpret model results and data findings to generate insights that drive decisions and strategies for enhancing student engagement.

III. PROJECT CHARTER

Sponsor Company – Rochester Institute of Technology

Individual Company Contacts – Oparinde Kolawole, Associate Consultant

Roles and Responsibilities:

Edmar Beatingo - Team Lead, represents team to sponsor, via email and on calls, to minimize communication errors.

Farwa Fayyaz - Project Manager, provides guidance and draws out insight from other team members, ensures that the project execution remains on track.

Maryam Mohammed - Project Scribe, responsible to taking meeting minutes and distributing notes/assignments. Can assist Team Lead in drafting emails and communication between sponsor and group.

Vishaal Dayashanker - Project Lead, responsible for holding the group accountable for meeting deadlines and ensures that the project deliverables are being met.

Mission:

Our mission is to do a comprehensive exploration of data science and machine learning methodologies to analyze student engagement data at the Rochester Institute of Technology. We aim to understand which opportunities attract the most signups, which are completed successfully, and identify patterns that lead to student drop-offs.

Vision Objectives:

To establish a strong reputation for providing actionable insights that significantly impact educational efficiency and effectiveness, positioning our team as crucial contributors to Rochester Institute of Technology's strategic student retention program and decisions.

Core Values

1. Data-Driven Decision Making: We base our recommendations on thorough analysis and evidence.
2. Collaboration: We work together effectively, leveraging each team member's strengths to achieve our common goals.
3. Integrity: We ensure that our analysis is honest, transparent, and free from bias.
4. Accountability: We take ownership of our tasks and deliver high-quality work within agreed timelines.

IV. WEEK-1 SUMMARY**Objective:**

The objective for Week-1 is that the team should be able to submit both a cleaned and preprocessed dataset and a detailed report summarizing all work done in Week 1. This includes detailing the data structure, the new features created, and the overall data preprocessing efforts. This report will provide context and documentation for the dataset, ensuring that it is ready for analysis in the upcoming weeks.

Data Description:

The shared dataset that the team will work on encompasses non-identifying information about every user who has ever created an account on Excelerate. The data is comprehensive, covering all users, regardless of their engagement with specific opportunities. Each row represents a unique user, and the dataset provides a holistic view of the user base. The dataset is composed of 35 columns and 8559 rows, including the header.

(Copy of SLU Opportunity Wise Data-1710158595043(raw)) Raw Dataset Link:

https://docs.google.com/spreadsheets/d/1tGEFzCwSpvwideKSK70JJJWhEqJWmXJKj68SyEOsj40/edit?usp=drive_link

DATA PREPARATION:**Data Cleaning Steps:**

1. Standardizing Formats: Ensuring consistency in data formats, such as dates, units of measurement, and categorical variables.
2. Handling Inconsistent Categorical Data: Standardizing categories and resolving discrepancies in categorical variables. All values in a particular column should follow the same format
3. Correcting Errors: Identifying and correcting any errors in the dataset, including typographical errors, inaccuracies, or inconsistencies.

4. Handling Outliers: Identifying and addressing data points that deviate significantly from the majority, which could distort analysis results.
5. Dealing with Duplicates: Identifying and removing duplicate records to avoid redundancy and ensure data integrity.
6. Deleting Empty Columns: Identifying records with no value and impact to the overall data analysis results.

FEATURE ENGINEERING:

1. Creating New Features:

- 1.1 Age Calculation - calculate the age from the Date of Birth to understand age distribution among users.
- 1.2 Opportunity Duration - subtracting the Opportunity Start Date from the Opportunity End Date to provide insights into the length of opportunities and their impact on engagement.

2. Transforming Existing Features

- 2.1 Normalization - normalize numerical features by scaling them to ensure all features contribute equally to the analysis.
- 2.2 Encoding Categorical Data - convert categorical data into numerical format to transform into a format suitable for analysis.

3. Extracting Useful Components

- 3.1 Date-Based Features - extract the month or year from the Learner SignUp DateTime to identify trends related to different times of the year.
- 3.2 Opportunity Engagement - calculate features like engagement time to reveal patterns in user behavior.

4. Combining Features

- 4.1 Interaction Features - create new features by combining existing ones to provide deeper insights into how different factors interact.
- 4.2 Engagement Scores - Develop composite scores by combining several features to provide a single metric to evaluate overall engagement.

(learner_engagement.csv) Pre-processed Dataset Link:

https://drive.google.com/file/d/1pQACBXZ-F_ZOCw_O_I_FeVjTy7OB6nLC/view?usp=sharing

V. WEEK-2 SUMMARY

DATA CLEANING:

Pre-processed Dataset Source:

The learner_engagement.csv dataset appears to contain data related to learner engagement in online learning opportunities. It is derived from Excelerate online education platform that tracks user activities, such as signups, course completions, and engagement scores.

Dataset Structure:

The dataset is structured in a tabular format with rows representing individual learner records.

Key columns typically include:

1. Learner ID: Unique identifier for each learner.
2. SignUp Date: The date when each learner registered for the course.
3. Opportunity End Date: The date when the learner completed or exited the course.
4. Engagement Scores: A numerical score representing the level of engagement by the learner, possibly calculated based on factors like login frequency, module completion, or interaction levels.
5. Completion Status: A field indicating whether the course was completed.

Dataset Key Features:

1. Signup Data: Tracks when users registered, allowing for analysis of signup trends over time.
2. Completion Information: Indicates whether learners completed the courses, enabling completion rate analysis.
3. Completion Time: Derived by calculating the difference between the signup date and end date, this feature provides insights into how long learners take to complete their courses.
4. Engagement Scores: Offers a quantitative measure of user activity within the platform, helping to analyze engagement patterns.
5. Segmenting Fields: These includes gender, age, city, state, country, to explore patterns across learner demographics.

This dataset is well-suited for analyzing trends in learner engagement, completion behaviors, and for identifying opportunities to improve user retention and satisfaction.

Image No.1 Shows the heatmap of signup seasonality per month and year.

Insights:

Year 2023:

1. The highest number of signups occurred in **July** (1081).
2. The lowest number of signups occurred in **January** (219).
3. There is a steady increase in signups from January to July, followed by a decline.

Year 2024:

1. The highest number of signups occurred in **January** (1485), which is a significant increase compared to the previous year.
2. The lowest number of signups occurred in **December** (11).
3. There is a sharp decline in signups after the January peak.

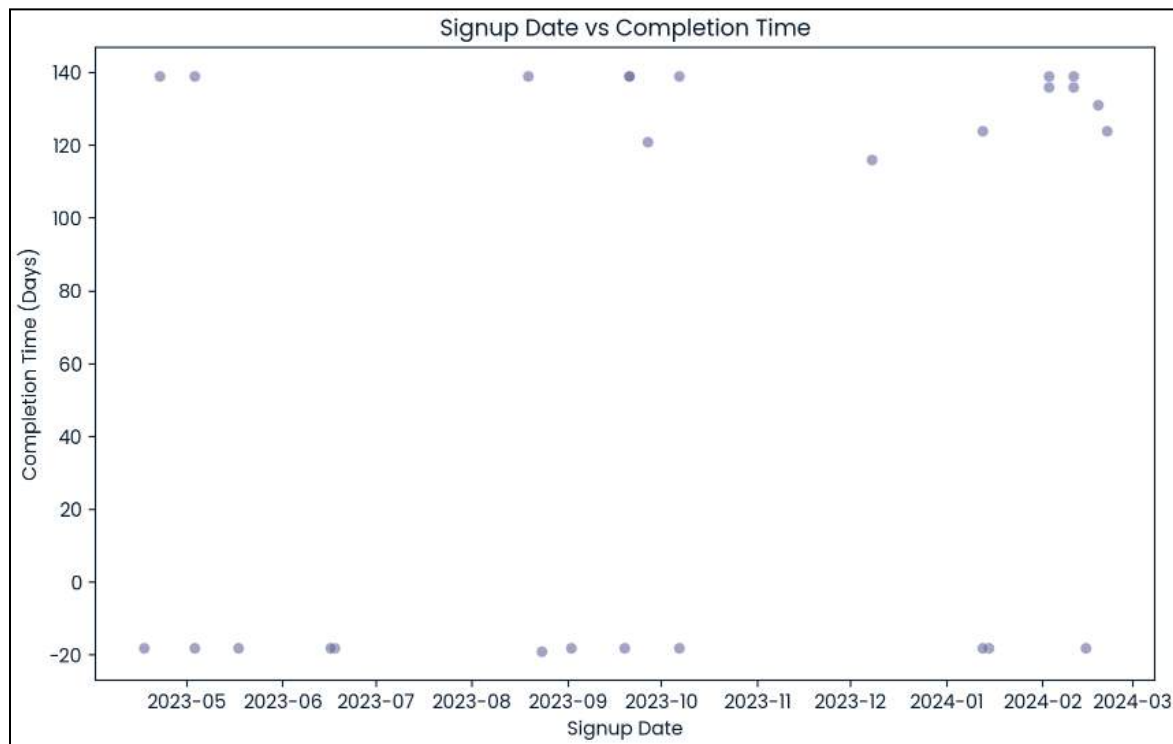


Image No. 2 Shows scatterplot correlation of the signup date and completion time.

Insights:

1. There is a cluster of data points in the middle of the plot, indicating that a significant portion of students completed the course within a similar timeframe, regardless of their signup date.

2. There are a few outliers, both in terms of early signups with longer completion times and late signups with shorter completion times. These outliers might be due to individual factors like student motivation, workload, or personal circumstances.

Key Insights:

1. **Seasonal Trends:** There are distinct seasonal trends in signups and completions, likely influenced by factors like holidays, academic calendars, and marketing campaigns.
2. **Student Engagement:** Student engagement, as measured by daily completions and overall progress, fluctuates over time.
3. **Completion Times:** Completion times vary significantly among learners, influenced by factors such as age, experience, and personal circumstances.
4. **Dropout Rates:** There is a potential for dropout, particularly among learners who experience significant delays or disengagement.

VI. WEEK-3 SUMMARY

Objectives:

1. Identify why students might leave a course or program, enabling Excelerate to develop strategies to boost retention and improve overall student success.
2. To proactively identify students who are at high risk of dropping off, allowing for early intervention.
3. To improve user experience by addressing areas that contribute to dissatisfaction or disengagement, such as difficult-to-navigate interfaces, lack of interactive content, or slow support response.
4. To ensure courses are engaging and well-structured, keeping students motivated to continue.
5. To equip instructors with insights into how their course design or teaching methods affect student engagement and retention.
6. To support students at risk of dropping out by implementing support strategies, such as personalized messages, additional tutoring, or reminders.

Further Data Cleaning in Python:

1. Duplicate removal: Removed duplicate rows using `drop_duplicates`.
2. Missing value handling: Filled categorical columns with 'Unknown' and numeric columns with median values using `fillna`.
3. Data type correction: Converted date columns ('Learner SignUp DateTime', 'Opportunity End Date', 'Opportunity Start Date') to datetime format using

pd.to_datetime.

4. Unnecessary column removal: Dropped irrelevant columns specified in columns_to_drop using drop.

(eda_engagement.csv) Further Cleaned Dataset Link:

https://drive.google.com/file/d/1CBt5ZjrW9UT_QHQFaLwLJcYKITS2E9Rx/view?usp=sharing

Feature Selection in Python:

1. Activity Duration (Days): Calculated duration between 'Opportunity Start Date' and 'Opportunity End Date'.
2. Completion Status: Binary indicator (1 = completed, 0 = not completed) based on 'Completion Date' presence.
3. Normalized Reward Engagement: Scaled 'Reward Amount' to a 0-1 range, reflecting engagement level.
4. Participation Frequency: Count of opportunities per learner (by 'Profile Id').
5. Time to Completion (Days): Days between 'Opportunity Start Date' and 'Completion Date'; defaults to 'Activity Duration' for incomplete activities.
6. Engagement Score: Weighted sum combining Completion Status (40%), Normalized Reward Engagement (30%), Time to Completion (20%) and Participation Frequency (10%).

(features_and_scores.csv) New Features Manipulation Dataset Link:

https://drive.google.com/file/d/1KoGPLYNMHLPOvzmlW9EF87f9EUTKj_uyJ/view?usp=sharing

(churn_notebook.ipynb) Python Churn Analysis Link:

https://drive.google.com/file/d/1dq8x5F0XIRaEY67ftiUk3F_Tyk74wDDO/view?usp=sharing

CHURN ANALYSIS:

Definition:

Churn analysis is the process of identifying and understanding the reasons why customers, users, or members leave or stop engaging with a product, service, or organization. In the context of educational institutions or any membership-based organization, churn analysis examines factors that influence dropout rates, aiming to reduce student or customer attrition and improve retention.

Descriptive Statistics:

The code calculates descriptive statistics for numerical columns (excluding 'Learner SignUp Year' and 'Normalized Reward Engagement') in the features_and_scores_df DataFrame.

Statistics Calculated:

1. Count: Number of non-missing observations.
2. Mean: Average value.
3. Standard Deviation (Std): Measure of data spread.
4. Minimum: Smallest value.
5. 25% (Q1): First quartile, representing 25th percentile.
6. 50% (Median or Q2): Median value, representing 50th percentile.
7. 75% (Q3): Third quartile, representing 75th percentile.
8. Maximum: Largest value.

Correlation Matrix:



Image No. 3 is a correlation matrix, which visually represents the relationships between different numerical variables.

Interpreting the Correlations:

1. Strong Positive Correlations:

Reward Amount and Completion Status: This suggests that higher reward amounts are associated with higher completion rates.

Reward Amount and Engagement Score: This indicates that higher rewards tend to lead to increased engagement.

Completion Status and Engagement Score: This implies that completed tasks are generally associated with higher engagement scores.

2. Moderate Positive Correlations:

Reward Amount and Activity Duration: This suggests that higher rewards might lead to longer activity durations.

Completion Status and Time to Completion: This indicates that completed tasks tend to take longer to finish.

3. Weak Negative Correlations:

Activity Duration and Participation Frequency: This might suggest that longer activity durations are associated with lower participation frequency.

Time to Completion and Participation Frequency: This implies that tasks taking longer to complete might have lower participation frequency.

PREDICTIVE MODELING:

Model Selection:

Three models are selected for comparison:

1. Logistic Regression: Suitable for binary classification problems, effective for linear relationships.

2. Decision Tree Classifier: Handles non-linear relationships, easy to interpret.

3. Random Forest Classifier: Ensemble method, robust to overfitting, suitable for complex relationships.

Model Training:

1. Data preprocessing: encoding categorical variables, handling missing values and scaling features.

2. Train/Test Split: Divides data into training (70%) and testing sets (30%).

3. Model instantiation and training using fit method.

Performance Metrics:

Evaluate model effectiveness using:

1. Accuracy: Proportion of correctly predicted instances.

2. Precision: $\text{True positives} / (\text{true positives} + \text{false positives})$.

3. Recall: $\text{True positives} / (\text{true positives} + \text{false negatives})$.

4. F1 Score: Harmonic mean of precision and recall.

Feature Importance:

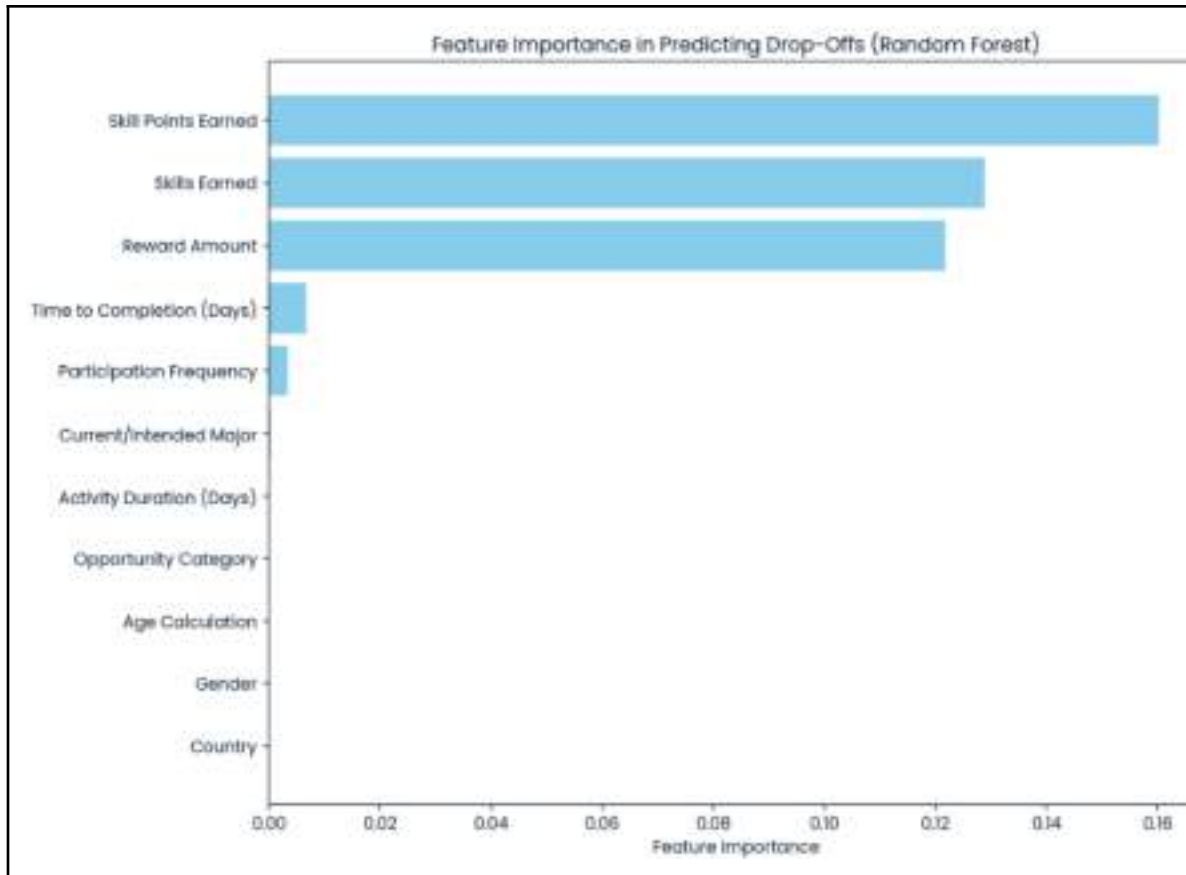


Image No. 4 shows the relative importance of different features in predicting whether a user will drop off from a particular course.

Key Observations:

1. **Most Important Features:** The top three features that have the highest impact on predicting drop-offs are:
 - Skill Points Earned: This suggests that earning skill points is strongly associated with user retention.
 - Skills Earned: Similar to skill points, acquiring new skills seems to be a significant factor in keeping users engaged.
 - Reward Amount: The amount of reward offered also plays a crucial role in influencing user retention.
2. **Moderately Important Features:** Features like "Time to Completion (Days)" and "Participation Frequency" have moderate importance. This suggests that the time it takes to complete a task and the frequency of participation might impact drop-off rates.
3. **Less Important Features:** Features like "Age Calculation," "Gender," "Country," "Opportunity Category," and "Current/Intended Major" have relatively low

importance in predicting drop-offs. This indicates that these factors may not have a strong impact on user retention.

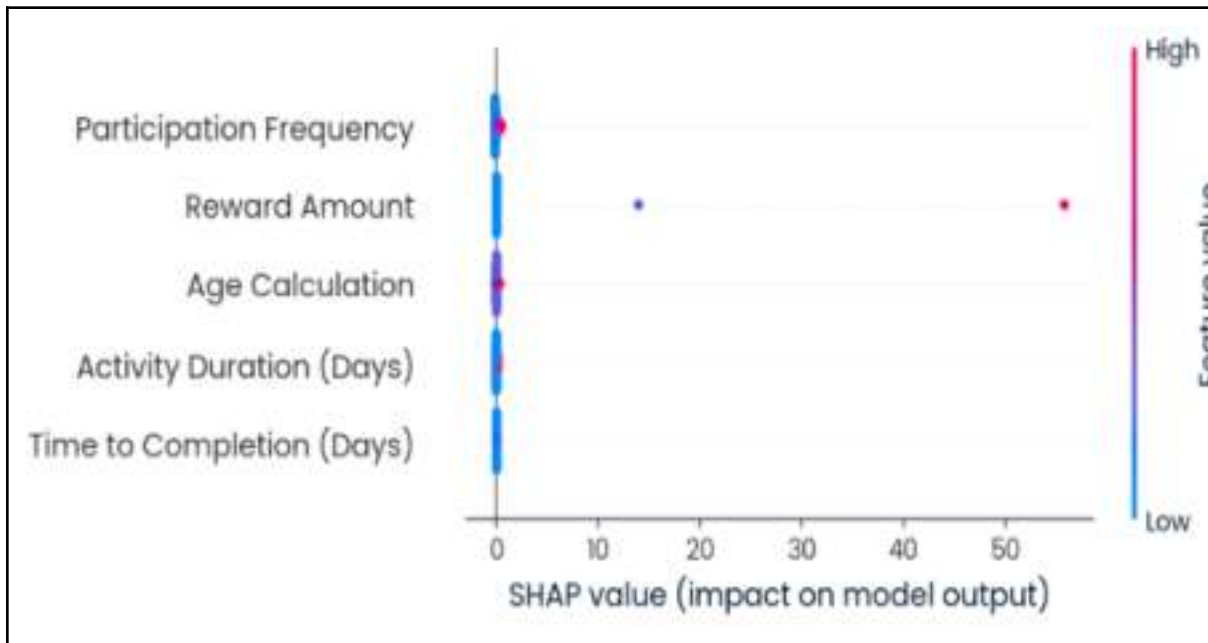


Image No. 5 shows a SHAP (SHapley Additive exPlanations) summary plot. SHAP is a technique used to understand the contribution of each feature in a machine learning model's prediction.

SHAP Values:

- The horizontal axis represents the SHAP value, which indicates the impact of a feature on the model's output.
- Positive SHAP values mean the feature increases the model's output, while negative values mean it decreases the output.

5.1 Key Factors

1. **Participation Frequency:** This feature has the highest impact on the model's output. Higher participation frequency generally leads to a higher model output.
2. **Reward Amount:** Similar to participation frequency, higher reward amounts tend to increase the model's output.
3. **Age Calculation:** Age seems to have a moderate impact, with older learners generally having a higher model output.
4. **Activity Duration (Days) and Time to Completion (Days):** These features have a mixed impact, with both positive and negative SHAP values. This suggests that the effect of these features on the model's output depends on the specific values.

5.2 Impact Analysis

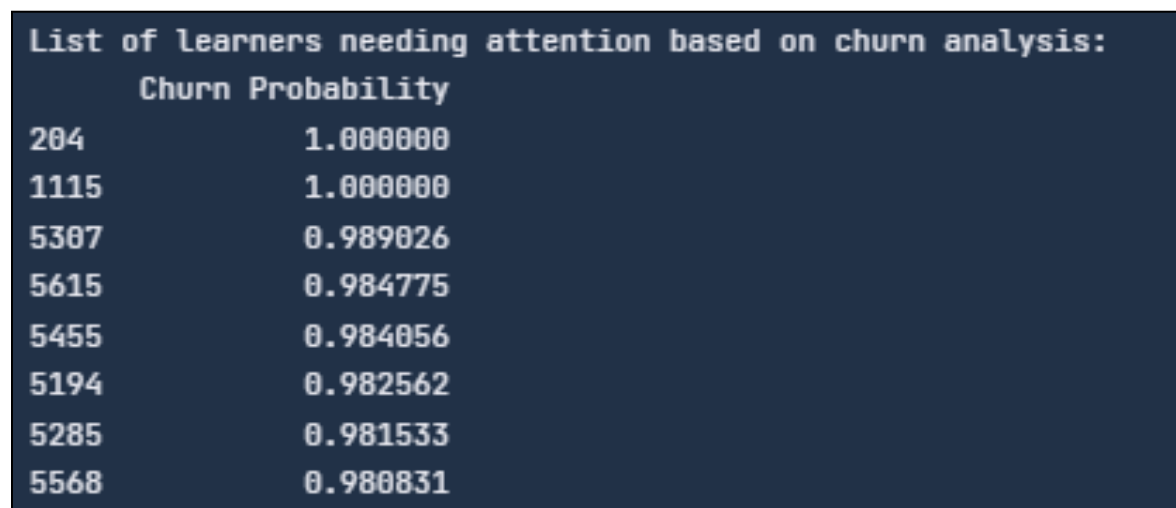
1. Participation Frequency and Reward Amount: These are the most significant factors influencing the model's predictions.
2. Age Calculation: Age also plays a role, but to a lesser extent.
3. Activity Duration and Time to Completion: The impact of these features is more nuanced and depends on the specific values.

VII. WEEK-4 HIGHLIGHTS

FILTERING CHURN:

(churn_filter_notebook.ipynb) Filtering Learners “at risk” of Churn by Probabiliy Link:

<https://drive.google.com/file/d/1XVzuof9h3VJJea9kakQxa7hwp4dTCbvT/view?usp=sharing>



List of learners needing attention based on churn analysis:	
Churn Probability	
204	1.000000
1115	1.000000
5307	0.989026
5615	0.984775
5455	0.984056
5194	0.982562
5285	0.981533
5568	0.980831

Image no. 6 shows the summary of learners that are “at risk” of churn based on probability

Age Calculation	Reward Amount	Activity Duration (Days)	Participation Frequency	Time to Completion (Days)	Churn Probability	At Risk
204	23	200	604	11	487	1 True
5455	23	50	-1	8	17	0.9840564374 True
1115	22	200	604	15	467	1 True
5568	23	50	-1	4	17	0.9808313311 True
5307	22	50	468	15	328	0.9890257044 True
5194	26	50	468	4	328	0.9825615533 True
5285	25	50	468	3	328	0.9815328083 True
5615	23	50	-1	9	17	0.9847754483 True

Image no. 7 shows the list of learners that are “at risk” of churn based on engagement features

	Profile Id	Opportunity Name	Churn Probability
204	28da8f68-3e23-4124-83b4-bbe4a2bb8f94	Career Essentials: Getting Started with Your Professional Journey	1
1115	c5f56bba-7f74-4c24-8c23-67d91daa5e82	Career Essentials: Getting Started with Your Professional Journey	1
5307	c5f56bba-7f74-4c24-8c23-67d91daa5e82	CPR/AED Certification	0.9890257044
5615	b1445feb-32ed-45d5-93a8-cd68450e7944	Startup Mastery Workshop	0.9847754483
5455	f9ce6f4c-2f5c-412e-9e29-f09d6f5a4422	Startup Mastery Workshop	0.9840564374
5194	79480b51-1ff4-43cb-82b7-ec814630b1ef	CPR/AED Certification	0.9825615533
5285	b8fb877d-539d-4c73-ad1e-d025d094d505	CPR/AED Certification	0.9815328083
5568	7f8e0b1c-bb39-4e5c-a9e2-88b3d7f3b851	Startup Mastery Workshop	0.9808313311

Image no. 8 shows the list of learners that are “at risk” of churn with specified Profile Id and Opportunity Name based on churn probability.

(at_risk_learners.csv) List of ‘at risk” learners by Churn Probability Link:

https://drive.google.com/file/d/1_DPUcTCNomGLPh9tfmGzkwJW_24NocjO/view?usp=sharing

Insights:

1. Consistent At-Risk Programs:

Certain programs like "Career Essentials: Getting Started with Your Professional Journey," "CPR/AED Certification," and "Startup Mastery Workshop" have multiple learners with a high probability of churn (close to or at 1). This suggests that these programs may face challenges in retaining learners, possibly due to content relevance, perceived value, or engagement levels.

2. Repeated Learner Profiles:

Specific profiles appear multiple times (e.g., Profile Id c5f66bba-7f74-4c24-8c23-678d19dac5e2 shows up for both "Career Essentials" and "CPR/AED Certification"). This indicates that certain learners might be struggling across multiple programs, which could signal broader issues with their engagement or motivation rather than specific program issues.

3. High-Risk Threshold:

The churn probabilities for learners in this table are consistently high, exceeding 0.98, with some reaching a probability of 1. This suggests an urgent need for intervention to prevent these learners from dropping out. Programs with such high churn probabilities may require immediate attention to address any underlying issues.

4. Opportunity for Program Enhancement:

The presence of high churn rates in certain programs could indicate areas where program design, support, or content might need adjustments. For instance, if these programs are dense or demanding, breaking down content into more digestible modules or adding interactive elements could help retain learners.

RECOMMENDATIONS:



	Feature	Average Impact on Model Output
3	Participation Frequency	0.128613
1	Reward Amount	0.082596
0	Age Calculation	0.032329
2	Activity Duration (Days)	0.025478
4	Time to Completion (Days)	0.001631

Image No. 9 provides a ranking of factors that contribute to student drop-off, as determined by a machine learning model. A higher average impact on model output indicates a stronger influence on predicting whether a student will drop out.

Based on the SHAP analysis, consider the following **strategies and interventions**:

1. Encourage Active Participation:

- Implement strategies to increase student engagement, such as:
 - Providing timely feedback and personalized guidance.
 - Creating interactive and collaborative learning experiences.
 - Offering opportunities for peer-to-peer interaction.

2. Optimize Reward Systems:

- Carefully design reward systems to incentivize participation and progress.
- Consider offering a variety of rewards, such as:
 - Tangible rewards (e.g., certificates, merchandise)
 - Intangible rewards (e.g., recognition, badges)
 - Experiential rewards (e.g., exclusive access to resources)

3. Identify and Support At-Risk Students:

- Use predictive analytics to identify students who are at risk of dropping out.
- Provide targeted support and interventions to these students, such as:
 - Personalized mentoring
 - Additional resources and tutorials
 - Flexible deadlines and scheduling

4. Monitor Student Progress and Provide Timely Feedback:

- Track student progress regularly to identify potential issues early on.
- Provide timely and constructive feedback to help students stay on track.

5. Create a Supportive Learning Environment:

- Foster a positive and inclusive learning environment.

- Encourage open communication between students and instructors.
- Offer emotional support and counseling services.

CONCLUSIONS:

Key Factors Contributing to Churn:

1. Participation Frequency: Students who participate less frequently are more likely to drop out.
2. Reward Amount: Lower reward amounts might correlate with higher dropout rates.
3. Age Calculation: Older students might be more prone to dropping out.
4. Activity Duration (Days): Longer activity durations might be associated with higher dropout rates.
5. Time to Completion (Days): Longer completion times might also contribute to higher dropout rates.

Insights:

1. Engagement is Critical: Encouraging active participation is essential to retain students.
2. Reward Systems Matter: Well-designed reward systems can motivate students and reduce churn.
3. Targeted Support: Identifying at-risk students and providing personalized support can improve retention.
4. Timely Feedback and Progress Tracking: Monitoring student progress and providing timely feedback can help keep students engaged.
5. Supportive Learning Environment: Creating a positive and inclusive learning environment is crucial for student retention.

Future Work:

1. Deeper Dive into Age and Churn: Conduct a more detailed analysis to understand the specific age groups that are most prone to dropping out. This could involve exploring factors like life stage, career goals, and learning preferences.
2. Impact of Activity Type and Difficulty: Analyze how different types of activities (e.g., quizzes, discussions, projects) and their difficulty levels influence student engagement and retention.
3. Long-Term Impact of Rewards: Evaluate the long-term effects of different reward systems on student behavior and retention.
4. Effectiveness of Support Interventions: Assess the impact of various support interventions (e.g., tutoring, mentoring, counseling) on student retention.
5. Continuous Monitoring and Adaptation: Regularly monitor student engagement metrics and adapt strategies as needed to maintain high retention rates.