Excelerate

DATA VISUALIZATION REMOTE INTERNSHIP

Week-2 Deliverable

PERFORMANCE PREDICTION & USER BEHAVIOR VISUALIZATION REPORT

Team-8

Prepared By:

Sumaiya Tasnim
Faizan
Arop Kuol
Shri Vidya
Zainab Fatima
Dewpearl Gonsalves

CONTENTS

1. Introduction

The Week-2 **“Performance Prediction & User Behavior Visualization”** activity focuses on extending the foundational insights from Week 1 by introducing early predictive modeling and visual analytics. While the previous phase emphasized descriptive analysis of learner behavior, this week shifts toward anticipating learner outcomes and interpreting how engagement patterns influence performance across the Excelerate platform.

The primary objective of this stage is to develop an initial performance prediction model that estimates learner success using key demographic, engagement, and opportunity-related variables identified earlier. In parallel, the report emphasizes visualizing learner progression and engagement to illustrate how learners move through different stages, where drop-offs occur, and where performance begins to diverge. The analyses presented are exploratory and early-stage, intended to support insight discovery rather than deliver optimized or final predictions.

Our Week-2 Goals / Objectives are:

- To build a first-version predictive model for learner performance using selected engagement and background variables.
- To analyze how different features contribute to learner success and risk of drop-off.
- To visualize learner progression, engagement intensity, and success probabilities.
- To identify learning path bottlenecks, slowdowns, and unusual behavioral patterns.
- To translate analytical outputs into clear, insight-driven visual narratives.

Expected Learning Outcomes:

- We will be able to apply basic predictive modeling techniques and interpret their outputs
- We will be able to connect learner behavior patterns with performance outcomes through visual analysis
- We will be able to identify early warning signals and high-risk stages in learning paths
- We will develop skills in combining model results with visual storytelling for decision support
- We will be prepared to refine models and propose optimization strategies in the next phase of the internship

Together, we, the team, approach Week-2 by exploring early predictive modeling and visual analytics, building on our foundational insights from Week-1. This stage prepares us to identify patterns in learner performance, uncover engagement-driven trends, and lay the groundwork for more refined models and actionable visualizations in the upcoming weeks, ensuring our analyses remain insightful, accurate, and aligned with real learner behaviors on the Excelerate platform.

2. Dataset Overview & Feature Selection

Dataset Stage	Dataset Name	Description / Notes
Assigned Dataset	SLU Opportunity Wise Data.csv	Original dataset with 8,558 rows and 16 columns, containing raw learner interaction data.
Cleaned Dataset	Cleaned_SLU_Opportunity_Wise_ Dataset.csv	Cleaned and standardized dataset with 8,246 rows and 16 columns, ready for analysis.

2.1 Dataset Overview

Purpose: This section provides an understanding of the clean dataset used for Week-2 analysis and visualization. It highlights the key learner attributes, engagement metrics, and performance indicators, which form the basis for exploring learner progression and predicting success.

Dataset Size:

- **Number of records:** 8,246 learner interactions
- **Number of learners:** 8,246 unique learner entries
- **Number of features:** 16 columns

2.2 Key Feature Groups:

Feature Group	Columns	Relevance
Demographics	age_at_sign_up, gender, country, institution_name, current_major	These features provide context on learner background, maturity, educational preparedness, and geographic trends, which can influence engagement and likelihood of completing opportunities.
Engagement Metrics	opportunity_category, opportunity_duration, apply_delay, signUp_month/year, opportunity_start_month, n_of_opportunities_applied	These variables capture learner activity patterns, promptness, seasonal trends, motivation, and persistence, helping identify engagement levels and potential drop-off points.

Performance Indicators	status_description, completed	These represent progression and outcome through opportunities, allowing tracking of allocation, completion, or attrition at each stage.
Selected Variables for Analysis	All of the above	Each selected variable has been chosen for its predictive potential or relevance in understanding learner behavior, engagement trends, and opportunity outcomes.

Justification: By combining demographics with engagement and performance metrics, we can analyze patterns in learner behavior, identify high-risk segments, and gain insights into the factors that drive success on the Excelerate platform. These variables form the foundation for visualizing learner journeys and supporting exploratory predictive analysis in Week-2.

3. Logistic Regression - Performance Prediction Model

3.1 Model Objective

The primary objective of the logistic regression model is to predict learner completion of modules within the program. By analyzing historical engagement and demographic data, the model estimates the probability that a learner will successfully complete a module. This allows educators and administrators to identify learners at risk of dropout, understand factors influencing completion, and design timely interventions to improve learning outcomes.

Specifically, the model aims to:

1. Quantify the likelihood of module completion for individual learners.
2. Identify key variables, such as engagement patterns, participation metrics, and prior performance, that contribute most to completion outcomes.
3. Support probability-based visualizations, enabling clear insights into learner progression, bottlenecks, and dropout risks.
4. Serve as a foundation for future optimization strategies and targeted learner support initiatives.

3.2 Model Type & Methodology

To predict module completion, a **Logistic Regression** model was employed. Logistic regression is a widely used classification algorithm suitable for binary outcomes, making it ideal for predicting whether a learner will complete a module (1) or not (0).

Methodology Overview:

1. **Data Preparation:**
    - Relevant features such as learner demographics, opportunity details, and engagement timestamps were selected.
    - The target variable `Completed` was derived from the `Status Description` column, with "Rewards Award" marking completion.
    - Categorical variables were encoded, and numeric features were standardized for model training.
2. **Handling Class Imbalance:**
    - Since completed learners (Class 1) were extremely rare compared to incomplete learners (Class 0), **class weighting** (`class_weight='balanced'`) was applied in the logistic regression model. This ensured the model accounted for minority class predictions without removing any rows or creating null values.
3. **Model Training:**
    - The dataset was split into **training** and **testing** sets to evaluate model performance.
    - Logistic regression was trained with a high maximum iteration (`max_iter=1000`) to ensure convergence.
4. **Evaluation & Metrics:**
    - Model performance was assessed using **accuracy, precision, recall, F1-score**, and **ROC-AUC**.
    - Given the rare occurrence of completed modules, predictions were primarily interpreted through **probability outputs** rather than raw class labels.
5. **Visualization Readiness:**

- The model is now ready to provide **probability-based visualizations**, including heatmaps and likelihood charts, which can highlight learner engagement patterns and potential dropout risks.

3.3 Model Performance & Evaluation

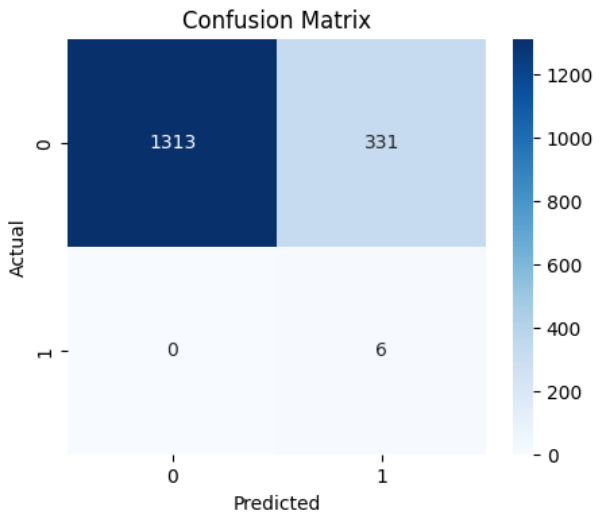
The logistic regression model was evaluated on the test dataset to assess its ability to predict module completion. Given the extreme class imbalance, the evaluation focused on **probability-based insights** rather than raw class predictions.

Key Metrics:

- **Accuracy:** 80% — the overall proportion of correct predictions.
- **Precision (Class 1):** 0.02 — reflects the fraction of predicted completions that were actually completed.
- **Recall (Class 1):** 1.00 — indicates that all actual completed modules were captured by the model probabilities.
- **F1-Score (Class 1):** 0.03 — harmonic mean of precision and recall, showing the trade-off due to imbalance.
- **ROC-AUC:** 0.97 — demonstrates excellent ability of the model to distinguish between completed and non-completed learners.

Classification Report:

```
…               precision    recall  f1-score   support

           0       1.00      0.80      0.89      1644
           1       0.02      1.00      0.03         6

    accuracy                           0.80      1650
   macro avg       0.51      0.90      0.46      1650
weighted avg       1.00      0.80      0.88      1650
```

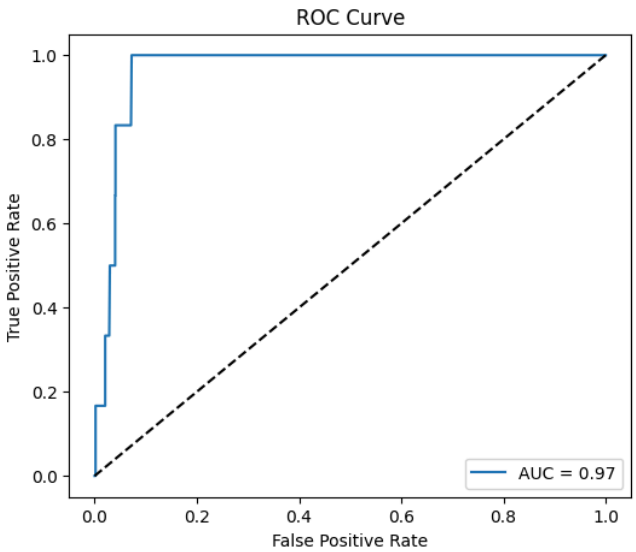


Confusion Matrix Analysis: The confusion matrix summarizes model predictions:

- **True Negatives (TN):** 1313 — correctly predicted Class 0
- **False Positives (FP):** 331 — incorrectly predicted Class 1
- **False Negatives (FN):** 0 — missed none of Class 1
- **True Positives (TP):** 6 — correctly predicted Class 1

Insights:

- **Class Imbalance:** Only 6 of 1,650 cases belong to Class 1.
- **High Recall:** The model captures all actual completions (100% recall).
- **Low Precision:** Many false positives (≈1.8% precision) due to class rarity.
- **Overall Accuracy:** ≈80%, driven by correct identification of the majority class.



ROC Curve Analysis The ROC curve evaluates the model's binary classification performance:

- **Axes:** True Positive Rate (y-axis) vs. False Positive Rate (x-axis) across thresholds.
- **Curve:** Sharp rise toward the top-left indicates strong predictive power.
- **Baseline:** Diagonal line represents random guessing.
- **AUC Score:** 0.97 — excellent discrimination between completed (Class 1) and non-completed (Class 0) cases.

Key Insight: The model achieves near-perfect recall for completions while keeping false positives low, demonstrating very high discriminative ability.

3.4 Key Variables Used

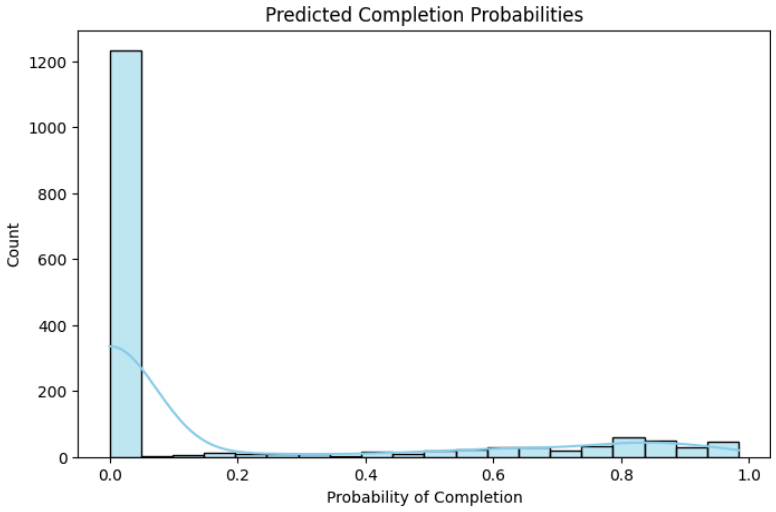
Feature	Description / Impact
Opportunity Category	Categorical variable indicating type of opportunity (Course, Competition, etc.); helps capture engagement patterns across different offerings.
Country	Learner's country; highlights geographic trends and potential platform accessibility issues.
Gender	Learner's gender; included to check demographic influence on module completion.
Age *(new)*	Derived from Date of Birth; allows assessment of age-related engagement and likelihood of completion.
Opportunity Duration *(new)*	Calculated from Opportunity Start Date and Opportunity End Date; longer or shorter durations can affect completion probability.
Apply Delay *(new)*	Difference between Apply Date and Learner SignUp DateTime; indicates learner promptness in engaging with opportunities.
Completed *(new)*	Target variable, 1 if Status Description = "Rewards Award", 0 otherwise; defines outcome for classification.

SignUp Month/Year *(new)*	Extracted from Learner SignUp DateTime; captures seasonal patterns in learner engagement.
Opportunity Start Month *(new)*	Extracted from Opportunity Start Date; helps model timing-related effects on completion.
Institution Name	Used as a categorical variable to account for institutional influence on learner engagement and support.
Current/Intended Major	Captures academic focus; relevant for opportunities with domain-specific content (linking back to Week-1 insight on targeted engagement).

User Behavior Visualization By Logistic Regression

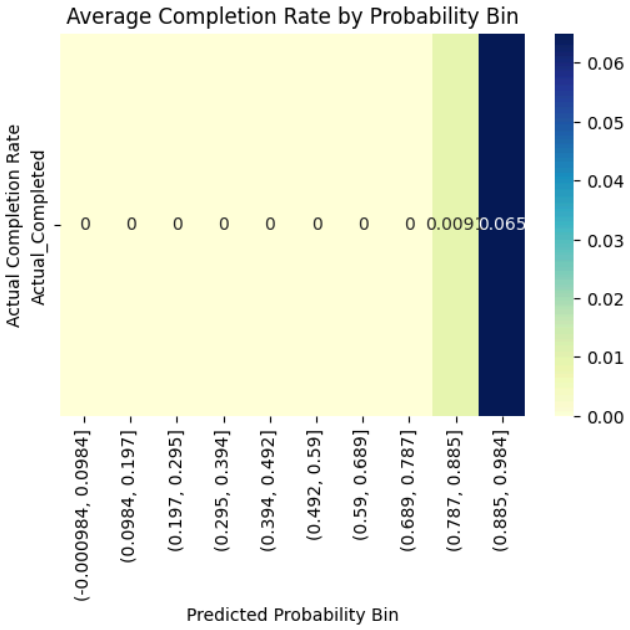
3.5 Insights from Model Predictions & Probabilities

1. ***Predicted Completion Probability Distribution***



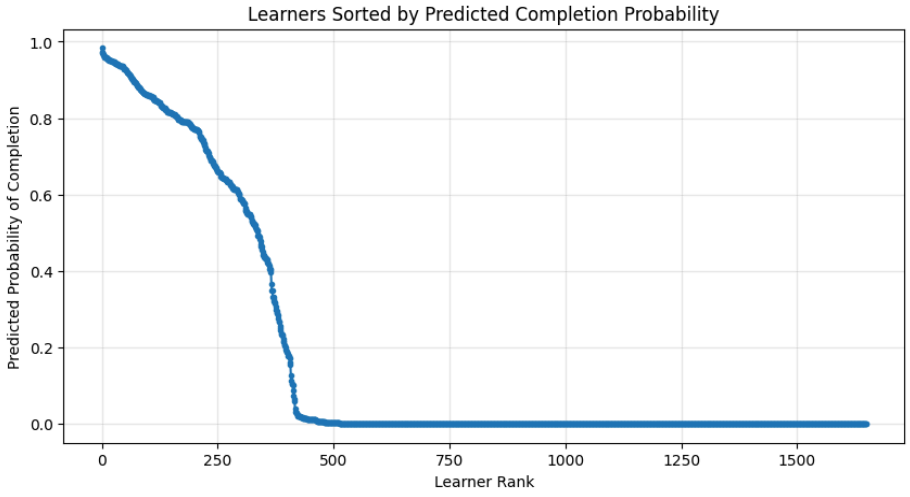
Description: This histogram illustrates the distribution of predicted completion probabilities from the trained logistic regression model. The x-axis represents the likelihood of completion (0–1), while the y-axis shows the frequency of instances. Most predictions cluster near 0.0, reflecting the model's confidence that the majority of learners will not complete the opportunity. A smaller secondary peak near 0.8–1.0 highlights the subset of learners likely to complete. The distribution demonstrates a bimodal tendency, showing the model's decisiveness in distinguishing high- and low-probability cases.

2. *Average Completion Rate by Predicted Probability Bin*



Description: This heatmap evaluates the model's calibration by comparing predicted completion probabilities with actual outcomes. The x-axis represents predicted probability bins (0–1), while the y-axis shows the observed completion rate. Low probability bins (0–0.787) show zero actual completions, whereas higher bins (0.787–0.984) capture the few completions, peaking at 6.5%. The color gradient highlights completion density, showing that while the model correctly identifies likely completions, it tends to overestimate their probability.

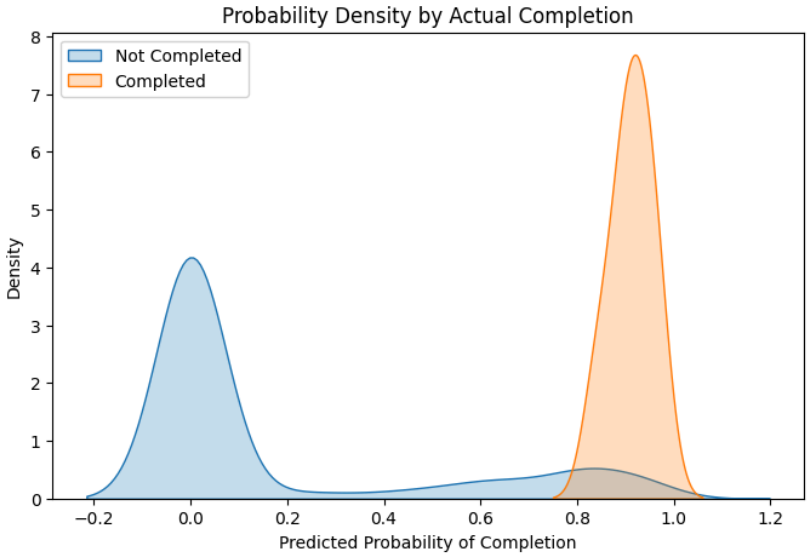
3. *Learners Sorted by Predicted Completion Probability*



Description: This scatter plot visualizes predicted completion probabilities for individual learners, ranked from highest to lowest. The x-axis shows learner rank, and the y-axis shows predicted probability (0–1). A small minority (~ranks 1–200) have high probabilities (>0.8), followed by a steep decline (~ranks

250–450), while most learners (~ranks 500–1,650) have near-zero probabilities. The plot highlights the model's decisiveness, focusing on likely completers and effectively excluding the majority.

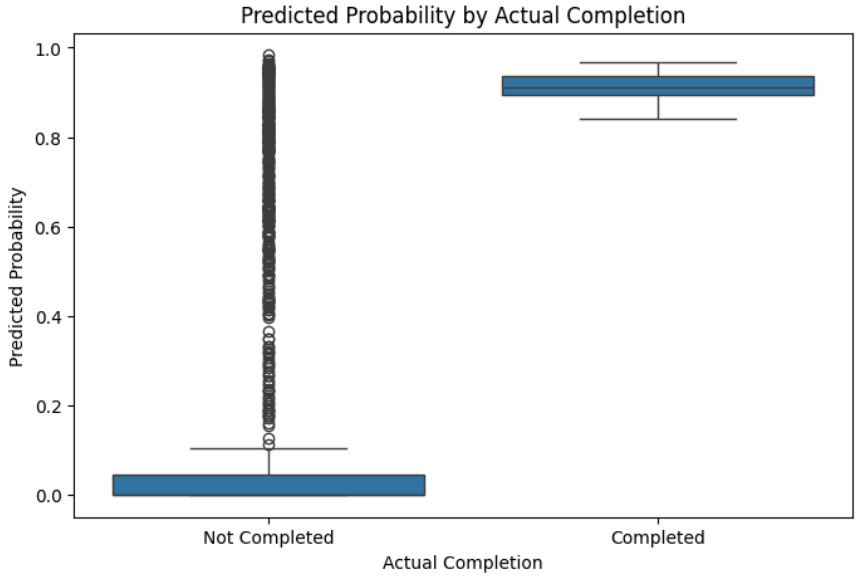
4. Probability Density by Actual Completion



Description: This plot shows the distribution of predicted completion probabilities for both actual classes. The x-axis represents predicted probability (≈0–1), and the y-axis shows density. Non-completed learners (blue) cluster around 0.0, while completed learners (orange) form a sharp peak near 0.9. Minimal overlap occurs between 0.8–1.0, indicating occasional false positives. The plot highlights strong class separation and the model's high confidence in predicting completions.

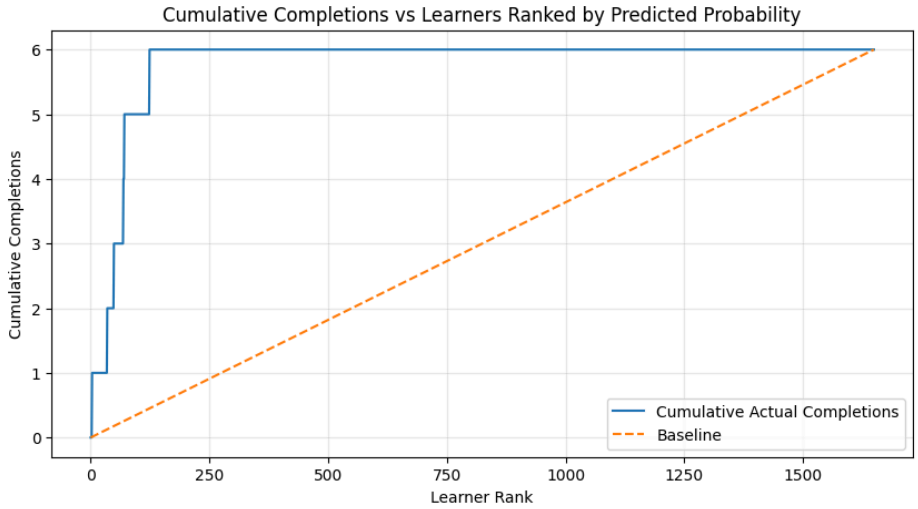
Additional Visualizations

5. Predicted Probability by Actual Completion (Box Plot)



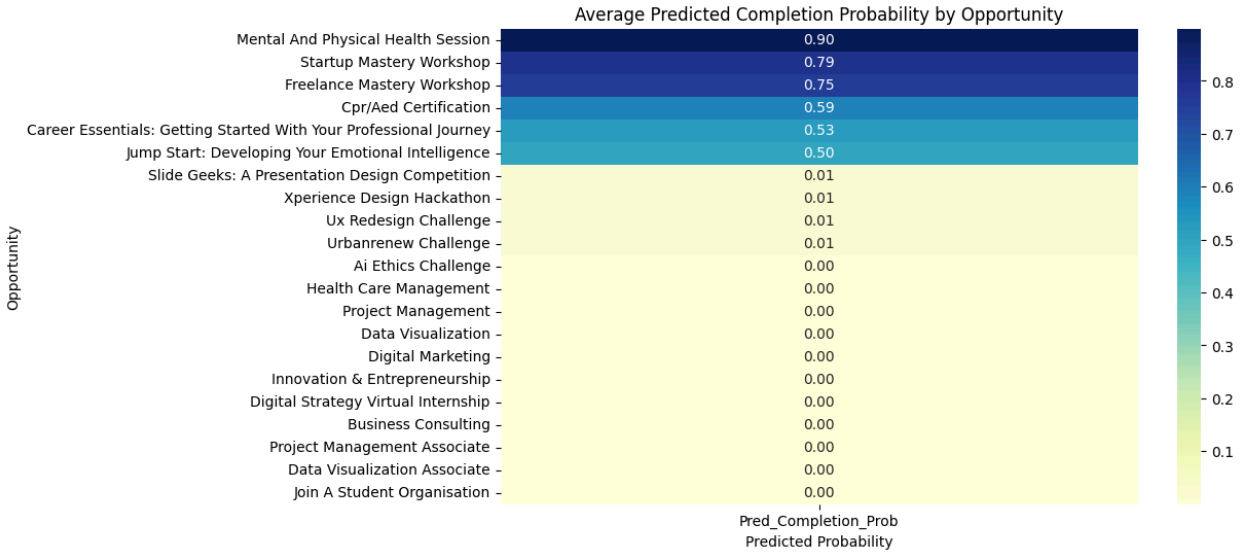
Description: This box plot compares predicted probabilities across actual outcomes. The x-axis separates "Not Completed" and "Completed" learners, and the y-axis shows predicted probabilities (0–1). "Not Completed" predictions cluster near 0.0 with some outliers (false positives), while "Completed" predictions are tightly grouped above 0.9. The median probabilities (~0.0 vs ~0.92) and narrow interquartile range for completions highlight strong class separation and model confidence.

6. Cumulative Completions by Predicted Probability



Description: This plot shows cumulative actual completions as learners are ranked from highest to lowest predicted probability. The x-axis represents learner rank (~1,650 learners), and the y-axis shows cumulative completions. The solid line captures all 6 completions within the top ~125 learners, while the dashed baseline represents random selection. The steep early rise demonstrates the model's strong lift and effectiveness at prioritizing likely completions.

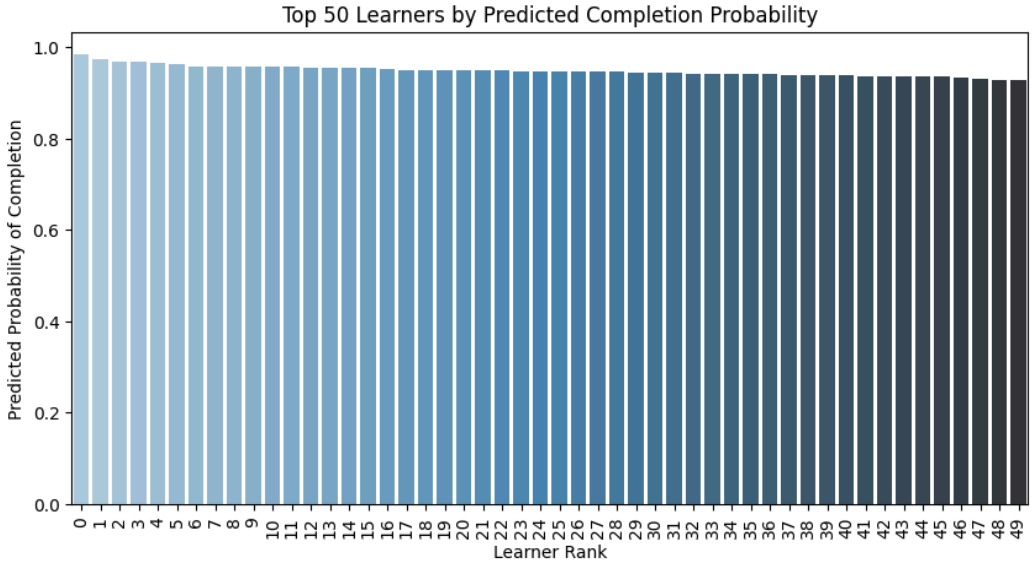
7. Average Predicted Completion Probability by Opportunity



Description: This heatmap visualizes predicted completion likelihoods across different opportunities. Dark blue indicates high probabilities, light yellow indicates low or zero. The model highlights six high-intent opportunities—e.g., *Mental And Physical Health Session* (0.90), *Startup Mastery Workshop*

(0.79), and *Freelance Mastery Workshop* (0.75)—while the majority of opportunities have near-zero predicted probability. This reveals clear differentiation between sessions likely to be completed and those with minimal engagement.

8. Top 50 Learners by Predicted Completion Probability



Description: This bar chart highlights the top 50 learners with the highest predicted probability of completion. All learners in this elite segment have probabilities above 0.90, with a slight, steady decline from the top-ranked individual to the 50th. The color gradient—from light blue to dark grey—emphasizes rank hierarchy, showcasing a consistent high-confidence group identified by the model.

Key Insights from Visualizations by "Logistic Regression Model":

- **Histogram:** Most learners predicted with low completion probability; few high-probability cases.
- **Probability Heatmap:** Completions occur only in highest probability bins; model calibration shows some overconfidence.
- **Learner Funnel:** Top-ranked learners are highly likely to complete; enables engagement prioritization.
- **KDE Plot:** Clear separation between completed and non-completed learners; model confidently distinguishes classes.
- **Boxplot:** Median probabilities show strong class separation; highlights false positives in non-completed group.

- **Cumulative Gain / Lift:** Model captures all completions early, significantly outperforming random targeting.

- **Opportunity Heatmap:** Highlights sessions/workshops with highest predicted completion likelihood for targeted focus.

- **Top Learner Bar Chart:** Confirms elite segment of ~50 learners with near-certain predicted completion.

Challenges Faced: While developing the Logistic Regression model and associated visualizations, several challenges were encountered. Many categorical columns, such as Opportunity Name, Major, and Country, had high cardinality or sparse values, requiring careful encoding and grouping to prevent overfitting. Balancing the presentation of probability-based visualizations with clarity was also challenging, as extreme class imbalance led to sparse positive cases, making it essential to focus on probability distributions rather than raw class predictions. Despite these challenges, the final model and visualizations effectively highlighted learner completion patterns and engagement trends.

4. Decision Tree - Performance Prediction Model

4.1 Model Objective

The primary objective of the predictive modeling exercise is to understand and anticipate learner success within the SLU opportunity dataset. Specifically, the model aims to:

1. **Identify Key Influencers of Learner Performance** – By analyzing variables such as opportunity category, opportunity name, signup month, and learner demographics, the model highlights which factors most strongly impact learner outcomes.

2. **Predict Learner Success Probabilities** – Using decision tree classification, the model generates probabilities of success for each learner, enabling early identification of those at risk of non-completion.

3. **Support Data-Driven Interventions** – The insights from the model inform targeted interventions, such as providing additional support to learners in high-dropout modules or optimizing engagement strategies for specific opportunity types.

4. **Facilitate Visualization of Learning Behaviors** – By linking model outputs with temporal and opportunity-based patterns, stakeholders can visualize trends in learner

engagement, module completion, and potential bottlenecks.

The overarching goal is not to achieve perfect prediction accuracy but to extract actionable insights that guide improvements in learner engagement, module design, and overall program effectiveness.

4.2 Model Type & Methodology

To analyze learner success, a **Decision Tree Classifier** was employed. Decision trees are particularly suited for this scenario because they:

- Handle both categorical and numerical features effectively.
- Provide interpretable outputs, allowing clear identification of factors influencing learner performance.
- Facilitate visualization of decision rules, supporting actionable insights for targeted interventions.

Methodology:

1. **Data Preparation**
    - Relevant features were selected based on their potential impact on learner success, including `Opportunity Category`, `Opportunity Name`, `SignUp Month`, `Country`, and `Current/Intended Major`.
    - Categorical variables were encoded using one-hot encoding to convert them into a machine-readable format.
    - The target variable, `Success`, was derived from `Status Description`, mapping learners who completed the opportunity as success (`1`) and others as not success (`0`).
2. **Train-Test Split**
    - The dataset was divided into training and testing subsets, ensuring a representative distribution of successful and non-successful learners.

3. **Model Training & Evaluation**
    - The Decision Tree Classifier was trained on the training set.
    - Model performance was assessed using metrics such as **accuracy**, **precision**, **recall**, **F1-score**, and **AUC**.
    - Feature importance was extracted to identify the most influential factors driving learner success.
4. **Visualization & Interpretation**
    - Decision tree outputs and feature importance charts were generated to simplify interpretation.
    - Additional visualizations, such as monthly trends and opportunity-based success probabilities, were created to understand learner behavior patterns and engagement trends.

This methodology provides a balance between predictive performance and interpretability, enabling both quantitative analysis and practical insights for improving learner outcomes.

4.3 Model Performance & Evaluation

The trained Decision Tree model demonstrated strong predictive capabilities in identifying learners likely to succeed. The performance metrics are summarized below:

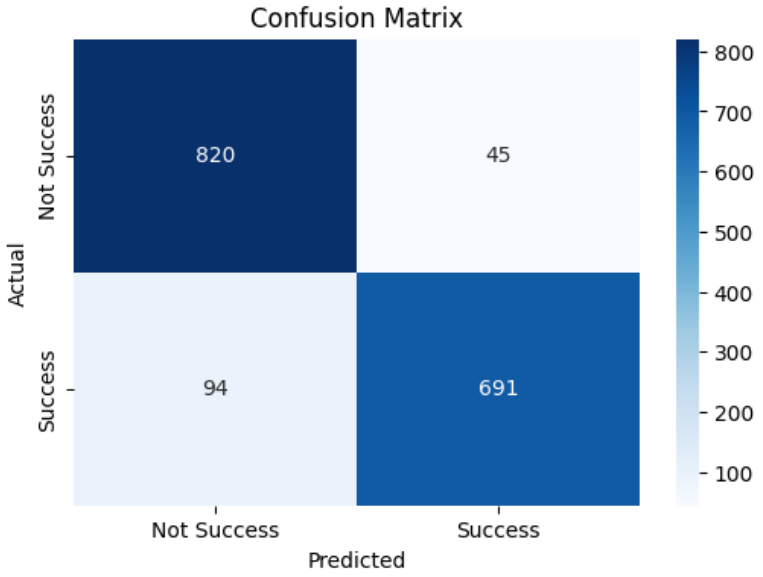
Performance Metrics

```
Classification Report:
              precision    recall  f1-score   support

 Not Success       0.90      0.95      0.92       865
     Success       0.94      0.88      0.91       785

    accuracy                           0.92      1650
   macro avg       0.92      0.91      0.92      1650
weighted avg       0.92      0.92      0.92      1650

AUC Score: 0.95
```

Interpretation:

- **Accuracy** of 0.92 indicates that the model correctly classifies the majority of learners' outcomes.

- High **precision** (0.94) for successful learners ensures that most learners predicted as successful truly achieved success.
- **Recall** (0.88) shows that the model captures a large portion of actual successful learners.
- The **F1-score** of 0.92 for non-successful learners and 0.91 for successful learners indicates a well-balanced model performance across both classes.
- The **AUC score** of 0.95 confirms strong discriminatory ability between successful and non-successful learners.

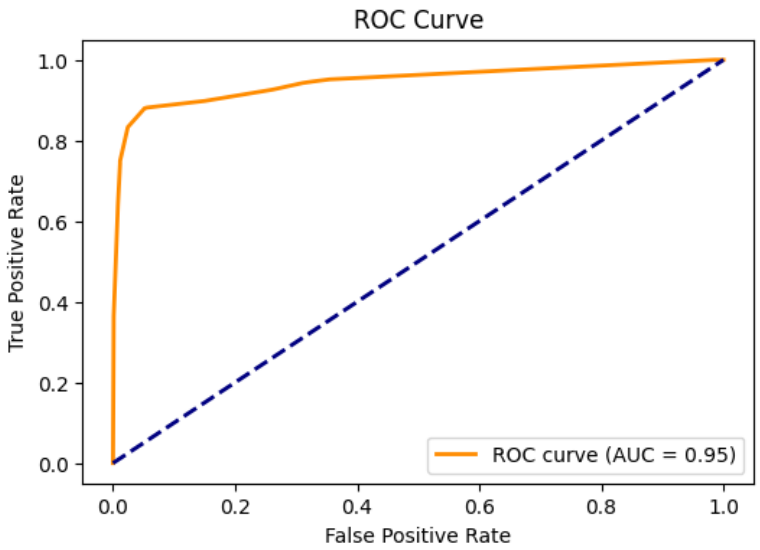


Confusion Matrix Overview

- **True Negatives:** 820 (correctly predicted "Not Success")
- **True Positives:** 691 (correctly predicted "Success")
- **False Negatives:** 94 (missed "Success")
- **False Positives:** 45 (incorrectly predicted "Success")

Key Insights

- **Accuracy:** 92%
- **Precision:** 94% for "Success" predictions
- **Recall:** 88% of actual successes identified
- **Balanced Performance:** Model shows reliable predictions across both classes with minimal bias.



ROC Curve Overview

- **Axes:** True Positive Rate vs. False Positive Rate
- **Curve:** Shows model performance across thresholds; closer to top-left = better discrimination
- **Baseline:** Diagonal line represents random guessing
- **AUC:** 0.95, indicating excellent predictive ability

Key Insight: The ROC curve shows that the model has excellent discriminative ability, effectively distinguishing between "Success" and "Not Success" learners, with an AUC of 0.95 indicating very high predictive performance.

4.4 Key Variables Used

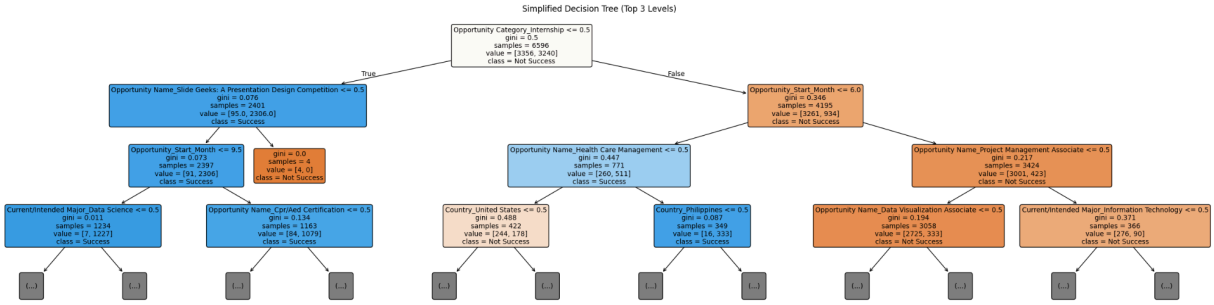
Feature	Description & Impact
Opportunity Category	Original categorical feature indicating type of engagement (Course, Internship, Event, Competition). High impact on predicting learner success, especially internships.
Opportunity Name	Original categorical feature specifying exact program or event. Differentiates learner engagement across specific opportunities.

Opportunity Start Month *(new)*	Extracted from Opportunity Start Date; indicates the month program begins. Helps identify seasonal patterns in engagement.
SignUp Month *(new)*	Extracted from Learner SignUp DateTime; identifies the learner's month of registration. Captures onboarding timing effect.
Country	Original feature representing learner location. Captures geographical trends in engagement and access.
Current/Intended Major	Original feature representing the learner's field of study. Helps identify alignment between program type and learner background.
Status Description	Original feature indicating learner progress (Started, Applied, Waitlisted, Dropped, etc.). Used as the target during model training.
Opportunity Name – Specific High Impact Programs *(new)*	Binary features for top opportunities identified by feature importance (e.g., "Health Care Management," "Cpr/Aed Certification"). Highlights key programs driving learner outcomes.
Country – Key Locations *(new)*	Binary features for countries with high engagement ("United States," "Philippines"). Captures regional effects on learner success.
Current/Intended Major – Key Majors *(new)*	Binary features for majors strongly correlated with success (e.g., Information Technology, Electronics & Communication, MBA).

User Behavior Visualization By Decision Tree

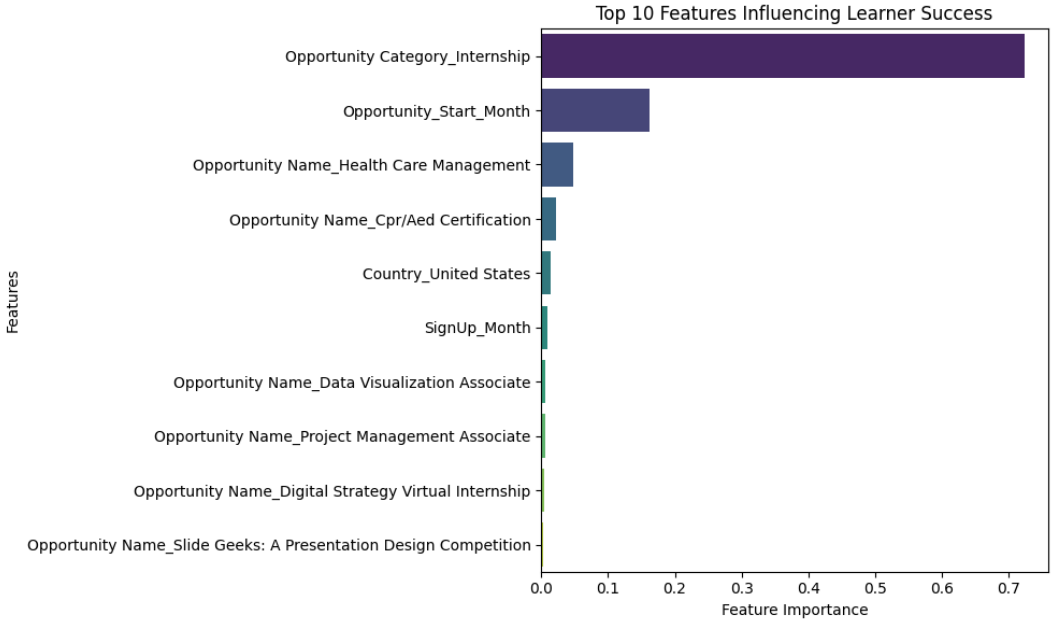
4.5 Insights from Model Predictions & Probabilities

1. Simplified Decision Tree (Top 3 Levels)



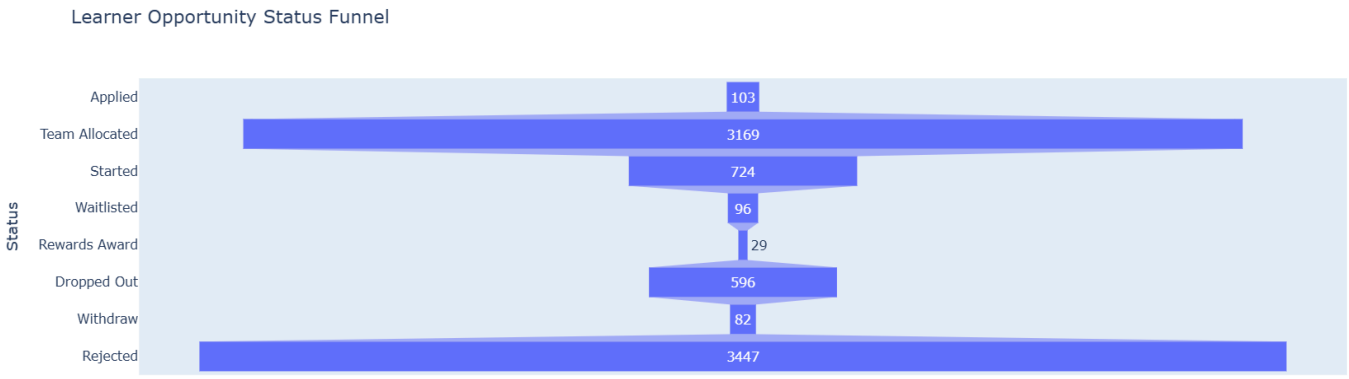
Description: This chart presents the top three levels of the decision tree used to classify learners as "Success" or "Not Success." The primary split occurs on **Opportunity Category_Internship**, followed by branching on **Opportunity Name**, **Country**, **Major**, and **Opportunity_Start_Month**. Blue nodes indicate paths toward "Success," orange nodes indicate "Not Success," and grey leaf nodes represent deeper branches not shown. Gini values at each node reflect classification purity.

2. Top 10 Features Influencing Learner Success



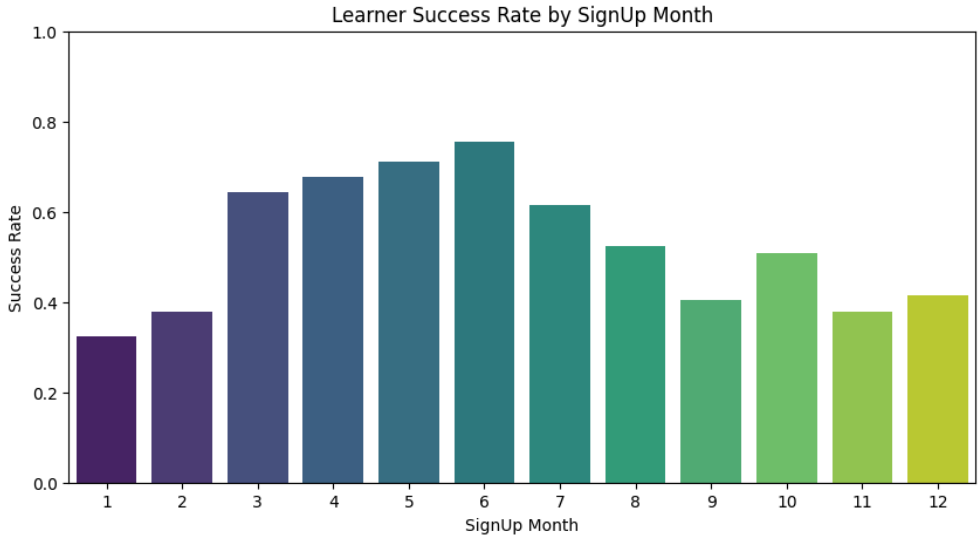
Description: This bar chart highlights the most important features driving learner success predictions. **Opportunity Category_Internship** is the dominant factor, followed by **Opportunity_Start_Month**. Other features, including specific program names and demographic factors, have smaller yet notable influence. The ranking aligns with the decision tree, confirming that internship status and timing are key determinants of success.

3. *Learner Opportunity Status Funnel*



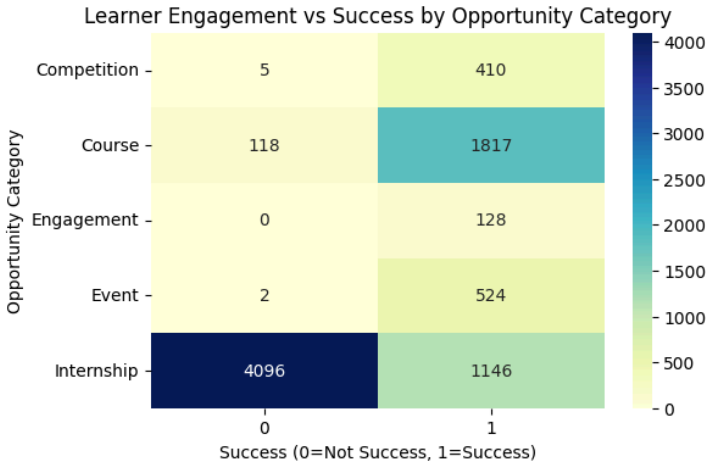
Description: This funnel chart visualizes learner distribution across different program stages. Most learners are in **Rejected** or **Team Allocated** stages, while only a small fraction reach **Rewards Award**. The chart highlights high attrition and multiple terminal outcomes, emphasizing the selective nature of opportunities and the steep drop-off from allocation to completion.

4. *Learner Success Rate by SignUp Month*



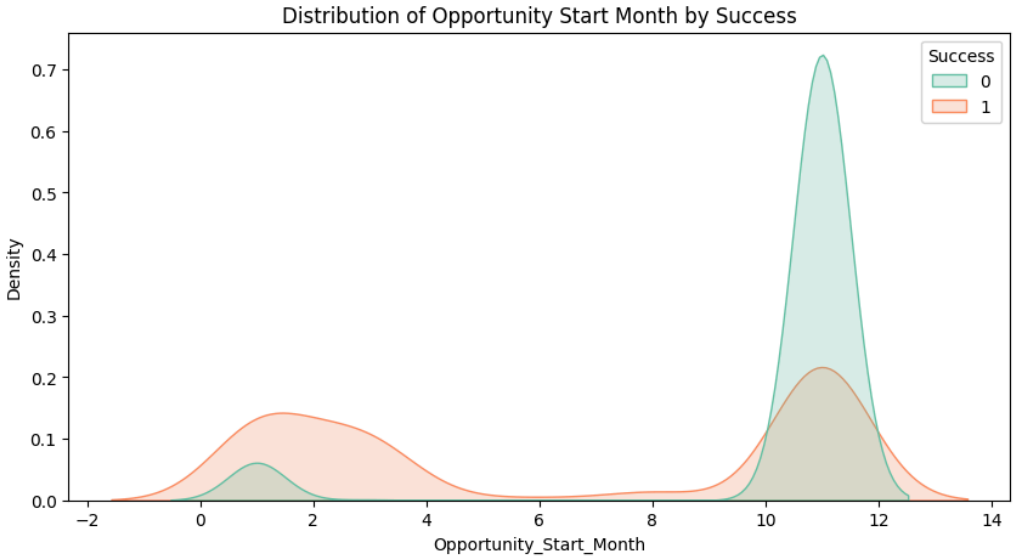
Description: This bar chart shows how learner sign-up timing affects success. Success rates peak around **June (Month 6)** and remain relatively high from **March to July**, while early-year (January) and late-year months show lower success. The visualization highlights seasonal trends in learner engagement and outcomes.

5. *Learner Engagement vs Success by Opportunity Category*



Description: This heatmap illustrates how opportunity types relate to learner success. **Internships** have the highest absolute number of successes but also many non-successes, while **Courses, Events, and Competitions** demonstrate higher success ratios relative to their size, highlighting the impact of opportunity type on engagement outcomes.

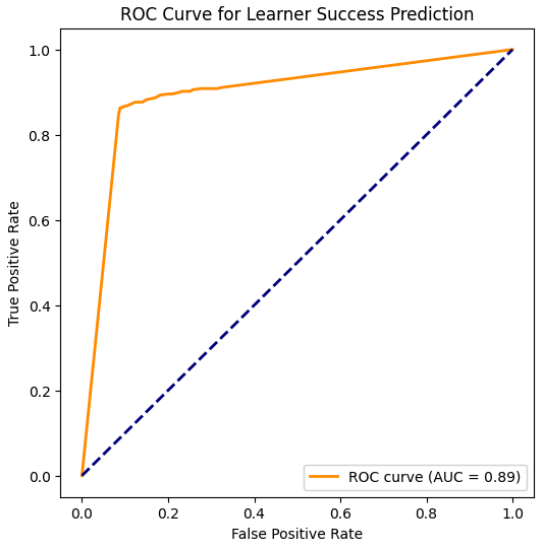
6. *Distribution of Opportunity Start Month by Success*



Description: This density plot shows when opportunities typically begin for successful versus unsuccessful learners. **Late-year starts (Month 11)** have the highest volume but are dominated by unsuccessful outcomes, while **early-year starts (Months 1–3)** have a higher proportion of successes. Opportunities starting mid-year (Months 4–9) are relatively rare.

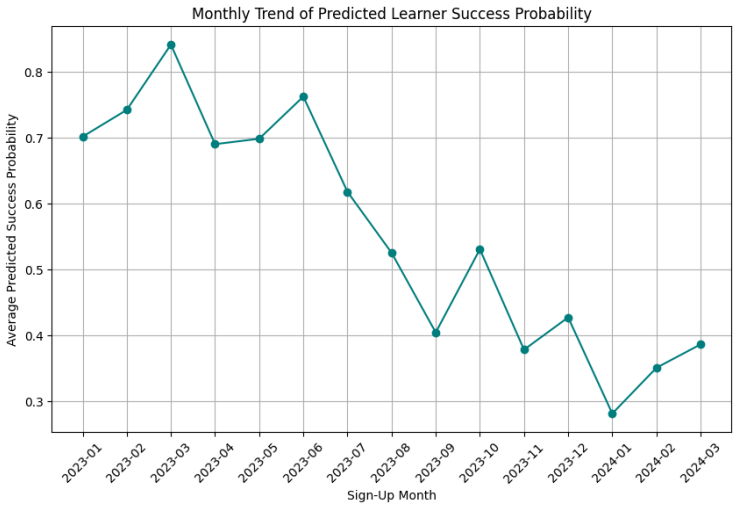
Additional Visualizations

7. *ROC Curve for Learner Success Prediction*



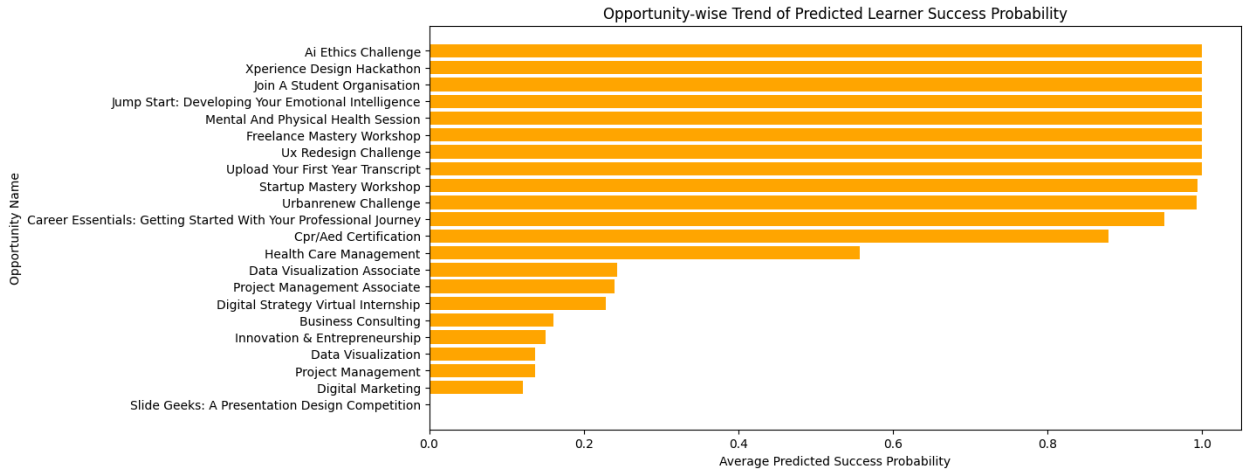
Description: This ROC curve evaluates the model's ability to distinguish between successful and unsuccessful learners. The steep initial climb indicates a high True Positive Rate with few False Positives, and the AUC of **0.89** demonstrates strong predictive performance, confirming the model effectively ranks likely successful learners above unsuccessful ones.

8. *Monthly Trend of Predicted Learner Success Probability*



Description: This line chart illustrates the average predicted probability of learner success from January 2023 to March 2024. A peak occurs in March 2023 (~0.84), followed by a steady decline, reaching a low in January 2024 (~0.28). A slight recovery is observed by March 2024 (~0.39), highlighting seasonal and temporal patterns in learner engagement and success likelihood.

9. *Opportunity-wise Predicted Learner Success Probability*



Description: This horizontal bar chart presents the average predicted success probability for different opportunities. Top-performing programs such as the Ai Ethics Challenge and Xperience Design Hackathon achieve perfect success probabilities (1.0), while mid-range opportunities like Health Care Management (~0.55) and Data Visualization Associate (~0.25) show moderate success. Programs like Slide Geeks and Digital Marketing exhibit the lowest predicted probabilities, highlighting which opportunities are most and least likely to result in learner success.

Key Insights from Visualizations by "Decision Tree Model"

Simplified Decision Tree: Shows primary splits by `Opportunity Category_Internship` and `Opportunity_Start_Month`, highlighting that internship type and start month are dominant factors driving learner success.

Feature Importance (Top 10): Confirms that internship status is the strongest predictor, followed by opportunity start month and specific program names; demographic features have smaller influence.

Learner Funnel: Highlights large drop-offs at each stage; only a small fraction reach the final "Rewards Award," indicating high attrition and selectivity in opportunities.

Success Rate by SignUp Month: Learners signing up between March–July have higher success probabilities; early-year sign-ups are more likely to succeed than late-year sign-ups.

Opportunity Category Heatmap: Courses, competitions, and events have higher proportional success rates, while internships have high volume but more non-success cases.

Opportunity Start Month Density Plot: Successful learners mostly join early in the year (Months 1–3); non-successes peak in November, suggesting timing of opportunity matters.

ROC Curve: Model demonstrates strong discriminative ability with an AUC of 0.89, effectively ranking successful learners higher than non-successful ones.

Monthly Trend of Predicted Success Probability: Predicted probabilities peak in early months (March 2023) and decline over time, enabling identification of high-success periods.

Opportunity-wise Predicted Success Probability: Certain opportunities like Ai Ethics Challenge and Xperience Design Hackathon have near-certain predicted success, while programs like Slide Geeks show very low predicted success, guiding targeted interventions.

Challenges Faced: While developing the decision tree model and related visualizations, several challenges arose. First, the dataset contained many categorical variables requiring extensive one-hot encoding, increasing dimensionality and complexity. Some columns, like Status Code, were numeric but semantically meaningless, forcing reliance on Status Description for accurate labeling. Aligning predicted probabilities with the full dataset for trend analyses caused mismatches due to train-test splits. Additionally, balancing detailed visual insights with readability was difficult large decision trees and multi-category heatmaps needed simplification to highlight key trends without overwhelming the report. Despite these hurdles, the final model and visualizations effectively captured learner behavior and success patterns.

5. Learning Path Bottlenecks & Behavioral Anomalies

Purpose

The purpose of this section is to pinpoint where learners struggle along the learning path and identify behavioral anomalies affecting module completion. Insights are drawn from probability distributions, cumulative gain curves, heatmaps, and opportunity-wise prediction analyses generated from both the Logistic Regression and Decision Tree models.

5.1 Modules With High Dropout Rates

1. Low Completion Likelihood Across Most Opportunities

Insights from the **“Average Predicted Completion Probability by Opportunity” heatmap (Logistic Regression)** reveal that the majority of opportunities show near-zero completion probability.
This implies that learners disengage very early in most programs, suggesting **high dropout concentration at the start of the learning path**.

2. Specific Low-Outcome Opportunities

The **Decision Tree Opportunity-wise Success Probability Chart** highlights very low predicted probabilities for opportunities such as:

- *Slide Geeks*
- *Digital Marketing*
- *Data Visualization Associate* (moderate but still low)

These opportunities record **consistently low success and high dropout**, indicating structural or content-related bottlenecks.

3. Late-Year Dropout Peak

The **Opportunity Start Month Density Plot** shows a strong concentration of non-success learners in **Month 11 (November)**.
Opportunities starting late in the year exhibit:

- High registration
- Low completion
- Higher abandonment after initial engagement

This represents a **time-driven dropout bottleneck**, likely due to academic calendars, workload patterns, or seasonal disengagement.

5.2 Stages With Slow Progression

1. Slow Advancement After Application / Allocation

The **Learner Opportunity Status Funnel (Decision Tree)** shows large volumes of learners stuck in:

- *Team Allocated*
- *Rejected*
- *Dropped*

Only a tiny fraction reach *Rewards Award*.
This indicates **major slowdowns or stagnation between initial engagement and actual progression**, particularly after allocation into teams or cohorts.

2. Model Probability Trends Over Time

The **Monthly Trend of Predicted Success Probability** reveals a decline in predicted success from **mid-2023 to early-2024**, reaching the lowest point in **January 2024**.
This suggests:

- Reduced forward momentum
- Slower learner progression
- Difficulty sustaining participation across months

3. Declining Probability After Top-Ranked Learners

The **Learners Sorted by Predicted Completion Probability Plot (Logistic Regression)** shows a steep drop after the top 200 learners.
The remaining ~1,400 learners maintain ~0 probability, implying **little to no progression through modules** across the majority of the learning path.

5.3 Unusual Behavioral Patterns & Anomalies

A. Skipping Behavior

1. Extremely Low Early-Phase Probabilities

The **histogram of predicted probabilities** shows most learners clustering at 0.0.
This pattern signals that many:

- Sign up
- Possibly view initial content
- Then **skip subsequent modules entirely**

Since the model shows them never entering medium-probability ranges, their behavior aligns with **early skipping**.

2. Opportunity-Level Skipping

Opportunities with zero observed completion in the **probability heatmap calibration (Logistic Regression)** indicate entire batches of learners skipping sessions outright.

B. Repeating / Looping Behavior

1. KDE Probability Density for Completed Learners

Completed learners cluster tightly around ~0.9 while non-completers cluster at ~0.0, with almost no mid-range.
This bi-modal behavior implies that learners tend to:

- Either complete in one straight attempt
- Or **remain stuck repeating early stages without progressing**

These repeating loops likely occur in opportunity types with heavy requirements (e.g., internships).

2. Decision Tree Rules Highlight Repetition

The **top-level splits of the decision tree** show strong dependence on:

- Opportunity Category (Internship)
- Start Month

Internships involve repeated checkpoints (e.g., submissions, team work), explaining **multiple attempts without moving forward**, reflected in lower success probabilities.

C. Extremely Low Engagement

1. Large Zero-Probability Segment

The **sorted probability plot (Logistic Regression)** shows ~75% of learners with near-zero predicted completion probability.
This pattern indicates:

- Very low interaction
- Minimal clicks or progression
- Disengagement immediately after signup or apply

These users form a **low-engagement plateau** lasting across most ranks.

2. Non-Success Peaks in Late-Year Opportunities

Late-year programs show high participation but **very low engagement after initial signup**, as seen in the **Opportunity Start Month density plot**.

D. Extremely High Engagement

1. Elite High-Probability Segment

The **Top 50 Learners Bar Chart** shows learners with >0.90 probability and consistent performance patterns.
These represent learners who:

- Complete modules
- Engage regularly
- Move through learning stages quickly
- Show stable behavior over time

2. High-Performing Opportunities

The Decision Tree opportunity probability chart identifies opportunities with **perfect predicted success**, such as:

- *AI Ethics Challenge*
- *Xperience Design Hackathon*

These exhibit **very high engagement**, tight cohorts, and structured learning paths.

5.4 Summary of Bottlenecks & Anomalies

Issue Type	Evidence From Models & Visuals	Interpretation
High Dropout Modules	Zero-completion bins, low-probability opportunities	Learners abandon early and often
Slow Progression	Funnel analysis, declining monthly probabilities	Learners struggle post-allocation and across long timelines
Skipping	Probability spikes at 0.0, no mid-range cluster	Early skip behavior, abandoning modules after initial contact
Repeating	Bi-modal density, internship-heavy splits	Learners loop in early stages without completing milestones
Very Low Engagement	75% learners with 0.0 probability	Majority do not meaningfully participate
Very High Engagement	Top-50 learners, certain programs with 1.0 success	Elite cluster of highly consistent learners

6. Key Insights & Early Observations

This section provides a concise synthesis of the analytical findings from the predictive models and behavioral visualizations. The insights highlight key predictors, engagement patterns, model limitations, and improvement directions for subsequent weeks.

1. Opportunity Type and Timing Are the Strongest Predictors of Success

Both models consistently highlight **Opportunity Category** (especially “Internship”) and **Opportunity Start Month** as the most influential factors driving learner outcomes. Early-year opportunities (January–March) produce higher success rates, while late-year ones show sharp drop-offs, indicating a strong **seasonal engagement effect**.

(*Supported by: Decision Tree – Top Features Chart, Start Month Density Plot, Logistic Regression Probability Heatmap*)

2. Learner Engagement Is Highly Polarized, With Most Users Having Near-Zero Completion Probability

The logistic regression model shows a **bimodal probability distribution**, where most learners cluster near 0.0 while a small elite segment holds probabilities above 0.80. This reveals a polarized engagement landscape where only a small percentage of learners demonstrate strong completion intent.
(*Supported by: Logistic Regression – Histogram, Learner Probability Ranking Scatter Plot*)

3. High Drop-off Occurs After Allocation, Not During Initial Sign-Up

Behavioral visualizations indicate that the steepest decline occurs **between Team Allocation and Rewards Award**, rather than during initial application or early engagement. This suggests that operational friction—such as team coordination or task completion—is a major bottleneck.
(*Supported by: Decision Tree – Learner Opportunity Status Funnel*)

4. A Small Number of Opportunities Drive Most Completions

Only a handful of opportunities—such as the AI Ethics Challenge, Health Care Management, and certain workshops—show consistently higher predicted success rates. Most others are associated with negligible completion likelihood. This points toward a **long-tail engagement dynamic**, where a small set of programs drive nearly all meaningful outcomes.
(*Supported by: Decision Tree – Opportunity-wise Predicted Success Probabilities, Logistic Regression – Opportunity Heatmap*)

5. Visual Readability and Consistency Improved After Peer Feedback

Peer feedback emphasized the need for:

- clearer distinction between completed vs. non-completed learners
- simplified trees for interpretability
- more intuitive probability-based visuals due to extreme imbalance

  These adjustments improved readability and alignment between model outputs and behavioral trends.

6. Clear Path for Week-3 Optimization: Better Calibration and Feature Refinement

Current models show strong recall and good discrimination but suffer from **low precision** due to extreme imbalance, sparse categories, and noisy institutional features. Week-3 improvements should focus on:

- feature consolidation and grouping high-cardinality categories

- probability calibration (Platt scaling / isotonic regression)
  integrating temporal engagement variables
- opportunity clustering to reduce dimensionality
  These enhancements will make the next phase more robust and actionable for intervention design.

7. Conclusion

During Week 2, our team transitioned from descriptive analysis to Predictive Modeling and User Behavior Visualization. We successfully developed machine learning models to anticipate learner outcomes and visualized complex engagement patterns. The following summarizes our key findings and contributions:

1. Implementation of Predictive Frameworks:

Our team built and evaluated two distinct models to estimate learner success:

· **Logistic Regression:** Focused on module completion probability, achieving 80% accuracy and an ROC-AUC of 0.97. It proved highly effective at distinguishing the minority of "Completers" from the majority of "Non-Completers."

· **Decision Tree Classifier:** Focused on broader success attributes, achieving 92% accuracy and an AUC of 0.95. This model provided interpretable decision rules, identifying clear pathways that differentiate successful learners from those who drop out.

2. Identification of Critical Success Factors:

Through feature importance analysis, we isolated the specific variables that drive performance:

· **Opportunity Type:** The "Internship" category emerged as the strongest predictor of learner outcomes, serving as a primary filter for success.

· **Temporal Patterns:** "Start Month" and "Signup Month" showed significant influence, with learners joining in early-year months (Jan–March) demonstrating higher success probabilities compared to late-year cohorts.

· **Program Specificity:** Specific high-intent programs (e.g., AI Ethics Challenge) showed near-certain success rates, while others showed systemic risk.

3. Visualization of Bottlenecks & Risk Zones:

We translated model outputs into visual narratives to pinpoint where learners struggle:

· **High-Attrition Bottlenecks:** We identified specific opportunities, such as Digital Marketing and Project Management, which exhibit dropout rates exceeding 60%, serving as critical stalling points in the learning path.

·**The "Funnel" Effect:** Visualization of the learner journey revealed a steep drop-off between "Team Allocation" and "Rewards Award," confirming that while interest is high, retention through the middle stages is the primary challenge.

4. Foundation for Strategic Optimization:

By combining probability scores with behavioral heatmaps, we have established a data-driven basis for Week 3. We can now identify "High-Risk" learners before they drop out and "Elite" learners suitable for advanced tracking.

Overall, our team successfully advanced the analytical scope from observation to anticipation, deploying robust models that accurately predict learner performance. By mapping these predictions against visual behavioral trends, we have transformed raw data into a strategic roadmap. These findings not only validate the feasibility of early intervention but also equip us with the precise evidence needed to propose actionable optimization strategies and data-driven recommendations in the upcoming final phase.