# Excelerate Data Visualization Internship

0212 DVA Team 4A


**Week 3 Report: Dashboard Creation and AI Model Training**

# Team Members:

| **Member's Name** | **Member's Email ID** |
|---|---|
| Abdullah Imran | abdullahimranarshad@gmail.com |
| Akshaya Cheruku | akshayacheruku@gmail.com |
| Nwabueze Victor | nwabuezevictor91@gmail.com |
| Chirag Pawaskar | chiragpawaskar1234@gmail.com |
| Omootemi Modupe | mariamomootemi@gmail.com |

# Contents:

# 1. Introduction

This report summarizes the work completed during **Week 3**, which focused on key activities related to data preparation, model training, and dashboard creation. The objective for this week was to build upon our existing dataset, perform exploratory data analysis, and develop predictive models while also visualizing insights effectively for better decision-making.

Our work began with a deep dive into the dataset, ensuring that it was cleaned, processed, and enriched with derived columns to enable meaningful predictions. Subsequently, we trained machine learning models tailored to specific objectives, experimenting with different algorithms to achieve optimal performance. Finally, the insights derived from the models and dataset analysis were translated into interactive dashboards, aimed at presenting data trends and model outcomes in a user-friendly and actionable format.

This structured approach—starting from data handling, progressing to model implementation, and culminating in dashboard development—formed the foundation for this week's tasks. The report outlines the methodologies, challenges, and results achieved during each phase, providing a comprehensive overview of the progress made in Week 3.

# 2. AI Model Training

The **AI Model Training and Validation Phase** represents a critical stage in the machine learning pipeline, where we leverage the processed data to develop predictive models. This phase involves defining the problem, selecting appropriate algorithms, optimizing model parameters, and evaluating performance to ensure the model's reliability and generalization capability. Below is a detailed account of the activities undertaken during this phase:

**Problem Definition and Objective Setting**

- The objective for model training was clearly defined, such as predicting specific variables (e.g., `Reward Amount` or `Opportunity Category`) or identifying patterns like user churn.
- Based on the problem type, the task was categorized into regression or classification.

**Dataset Preparation**

- **Feature Selection**: Relevant features were identified based on domain knowledge and correlation analysis to ensure they contributed meaningfully to the prediction task. For example:
    - Demographic variables like `City`, `Country`, and `Gender`.

    - Engagement metrics such as `Skill Points Earned`, `Critical Thinking`, and `Skills Count`.
- **Data Splitting**: The dataset was divided into training and testing sets (e.g., 80%-20% split) to enable model training and independent evaluation.
- **Handling Missing Values**: Rows or columns with missing data were addressed using imputation methods or exclusion to maintain data integrity.
- **Feature Engineering**: New columns, such as `Predicted Reward Amount` or encoded categorical variables, were added to enhance the dataset's predictive power.

**Model Selection and Training**

- Multiple machine learning algorithms were considered, including:
    - **Random Forest**: A robust algorithm for both classification and regression, chosen for its ability to handle non-linear relationships and importance ranking of features.
    - **Linear Regression**: Used as a baseline for regression tasks.
    - **Logistic Regression**: Tested for binary classification tasks.
- The models were trained on the training set, optimizing the parameters to minimize errors or maximize classification accuracy.

**Validation and Performance Evaluation**

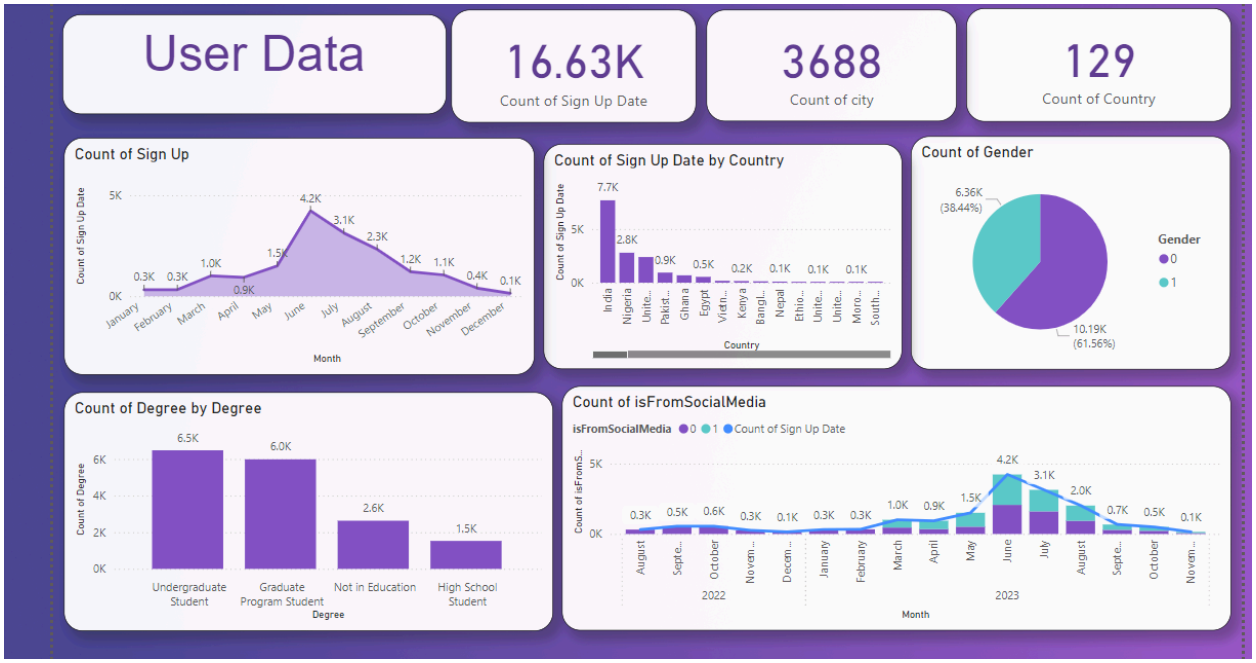
- **Cross-Validation**: Techniques like k-fold cross-validation were applied to reduce overfitting and validate the model's generalization capabilities on unseen data.
- **Evaluation Metrics**:
    - **Regression Tasks**: Metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were calculated to measure how close predictions were to the actual values.
    - **Classification Tasks**: Metrics like accuracy, precision, recall, F1-score, and the confusion matrix were used to evaluate classification performance.
- Results were analyzed to determine the strengths and weaknesses of each model, with an emphasis on improving underperforming areas.

**Model Optimization**

- **Hyperparameter Tuning**: Grid Search or Random Search techniques were used to identify the best combination of hyperparameters for the chosen algorithms.
- **Feature Importance Analysis**: Features with the most predictive power were identified and emphasized, while redundant or noisy features were excluded.

# 3. Dashboard Creation

## 3.1: User Data Dashboard




This dashboard provides a comprehensive overview of user-related data and visualizes key metrics to facilitate analysis of user engagement, demographics, and behavior patterns. It combines various charts and summaries to highlight trends and distributions across multiple dimensions of the dataset. Below is a breakdown of what the dashboard represents:

1. Overall Metrics (Top Section)

- Count of Sign-Up Dates: Displays the total number of users who signed up, indicating the scale of the dataset (16.63K users).
- Count of City: Represents the number of unique cities users are from (3,688).
- Count of Country: Shows the total number of unique countries (129), highlighting the global reach of the platform.

2. Sign-Up Trends (Left Section)

- Count of Sign-Ups by Month: A line chart showcasing the monthly distribution of user sign-ups.
    - It indicates a clear seasonal trend, with the highest number of sign-ups observed in June (4.2K) and a gradual decline afterward.

    - Sign-ups are significantly lower during the late months of the year, such as November and December.

3. Geographic Distribution (Top Center Section)

- Sign-Up Date by Country: A bar chart illustrating the number of sign-ups from each country.
    - India dominates with the highest number of sign-ups (7.7K), followed by Nigeria (2.8K), the United States (0.9K), and other countries.
    - The chart emphasizes the platform's strong presence in specific regions.

4. User Demographics (Right Section)

- Count of Gender: A pie chart representing the gender distribution of users.
    - Male users (denoted as "1") account for 61.56%, while female users ("0") make up 38.44%, indicating a slightly skewed gender distribution.

5. Degree Distribution (Bottom Left Section)

- Count of Degree by Degree: A bar chart visualizing the educational background of users.
    - Most users are Undergraduate Students (6.5K) and Graduate Program Students (6.0K), followed by those Not in Education (2.6K) and High School Students (1.5K).
    - This indicates that the platform primarily caters to higher education students.

6. Social Media Influence (Bottom Right Section)

- Count of `isFromSocialMedia`: A line chart comparing sign-up trends between users from social media and other sources over time.
    - Social media appears to have a steady but relatively small impact on user sign-ups, with consistent counts across months.
    - A significant spike in sign-ups is observed in May and June, regardless of the source.

## Insights:

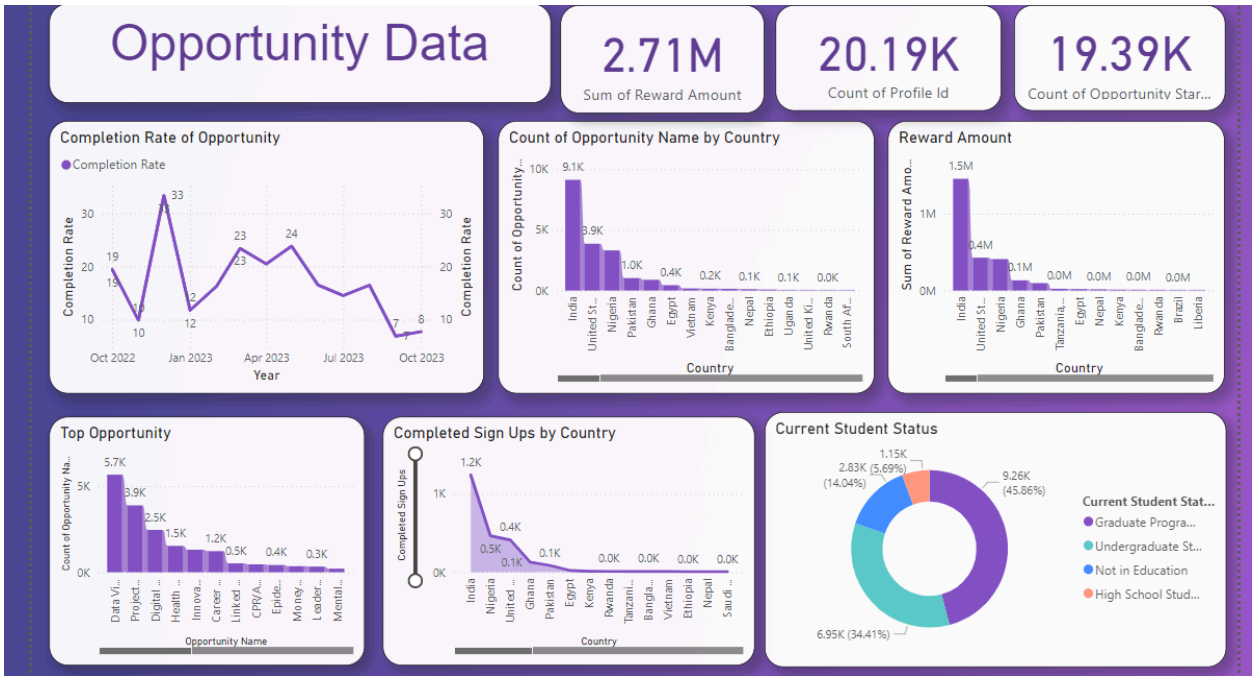
- Sign-Up Trends: Seasonal patterns suggest promotional efforts during peak months like June could drive sign-ups further.
- Geographic Reach: Focused marketing in countries with lower user counts could expand the platform's reach.
- Demographics: Tailored campaigns for undergraduate and graduate students, along with addressing the gender skew, could enhance user diversity.
- Social Media Impact: While social media contributes to sign-ups, there is room for optimizing these channels to increase their effectiveness.

Relevance to stakeholders:

This dashboard is highly relevant to stakeholders as it provides actionable insights into user engagement, demographics, and geographic reach, enabling data-driven decision-making. By identifying seasonal trends in sign-ups, stakeholders can optimize marketing efforts during peak months to maximize user acquisition. The geographic distribution highlights the platform's strengths in specific regions, such as India, while revealing untapped markets for potential expansion. The demographic breakdown allows stakeholders to tailor content and services to the predominant user base (undergraduates and graduates) while addressing gaps like the gender imbalance. Additionally, the analysis of social media influence can guide stakeholders in refining outreach strategies to enhance the platform's growth through these channels.

## 3.2: Opportunity Wise Data Dashboard

### 3.2.1: Dashboard 1




This dashboard provides a comprehensive overview of opportunity-related data and visualizes key metrics to facilitate analysis of engagement, rewards, and completion patterns across different countries and student segments. It combines various charts and summaries to highlight trends and distributions across multiple dimensions of the dataset. Below is a breakdown of what the dashboard represents:

1. Overall Metrics (Top Section)
- Sum of Reward Amount: Displays the total rewards distributed, indicating significant investment (2.71M)
- Count of Profile ID: Represents the number of unique participants (20.19K)
- Count of Opportunity Starts: Shows the total number of initiated opportunities (19.39K)
2. Completion Rate Trends (Top Left Section)
- Completion Rate of Opportunity: A line chart showcasing monthly completion rates from Oct 2022 to Oct 2023
    - Peak completion rate of 33% observed in early 2023
    - Recent trend shows decline to approximately 8% by October 2023
    - Fluctuating patterns throughout the year with notable variations
3. Geographic Distribution (Center Sections)
- Count of Opportunity Name by Country: A bar chart illustrating opportunity distribution across countries
    - India leads with approximately 9.1K opportunities
    - Followed by United States, Malaysia, and Pakistan with decreasing counts
- Completed Sign Ups by Country: Shows successful completions by region
    - India dominates with about 1.2K completed sign-ups
    - Significant gap between India and other participating countries
4. Reward Distribution (Top Right Section)
- Reward Amount by Country: Visualizes monetary distribution across regions
    - India shows highest reward amount at approximately 1.5M
    - United States and Malaysia follow with notable but lesser amounts
    - Other countries show comparatively smaller reward distributions
5. Opportunity Analysis (Bottom Left Section)
- Top Opportunity: Bar chart showing popularity of different opportunity types
    - DiFi AI L leads with about 5.7K participants
    - Project and other categories follow with decreasing participation
6. Student Demographics (Bottom Right Section)
- Current Student Status: Pie chart representing educational background of participants
    - Graduate Program students form largest segment (45.86%, 9.26K)
    - Undergraduate Students follow (34.41%, 6.95K)
    - Not in Education (14.04%, 2.83K)
    - High School Students (5.69%, 1.15K)

Insights:

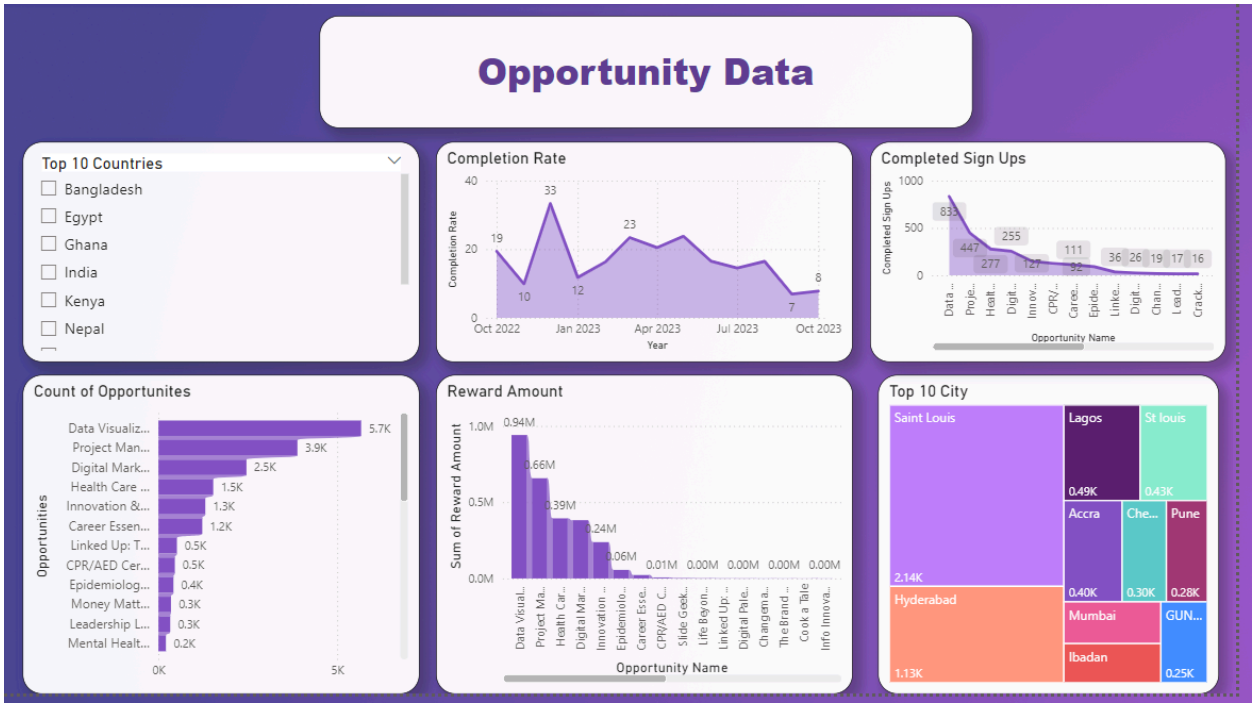
- Geographic Focus: India emerges as the primary market across all metrics, suggesting successful market penetration but potential for geographic diversification
- Completion Trends: Declining completion rates in recent months warrant investigation and potential intervention strategies
- Student Engagement: Strong participation from graduate and undergraduate students indicates effective targeting of higher education segments
- Opportunity Distribution: DiFi AI L's success could inform strategy for other opportunity types

- Reward Effectiveness: Significant reward amounts suggest substantial investment in participant incentivization, particularly in the Indian market

### Relavance to stakeholders:

Executive Leadership: Provides high-level metrics (2.71M total rewards, 20.19K profiles) that demonstrate program scale and investment returns
Program Managers: Student status distribution helps in curriculum planning and targeting specific educational segments
Financial Teams: Detailed reward distribution by country helps in budget allocation and ROI analysis
Market Expansion Teams: Clear indication of country-wise performance with India's dominance suggests opportunities in other markets
Student Success Teams: Completion rate trends help identify intervention needs and success patterns
Recruitment Teams: Understanding the current educational demographic split helps in targeted outreach

## 3.2.2: Dashboard 2




This dashboard provides a comprehensive overview of opportunity-related data and visualizes key metrics to facilitate analysis of engagement, rewards, and participation patterns across different locations and opportunity types. Below is a breakdown of what the dashboard represents:

1. Geographic Reach (Top Left Section)

- Top 10 Countries: Provides a filter list including Bangladesh, Egypt, Ghana, India, Kenya, and Nepal among others
    - Indicates the primary regions where opportunities are available
    - Suggests a focus on developing markets

2. Completion Metrics (Top Center Section)

- Completion Rate Trend: A line chart showing rates from Oct 2022 to Oct 2023
    - Peak completion rate of 33% in January 2023
    - Recent decline to approximately 7-8% by October 2023
    - Shows significant fluctuation throughout the year

3. Opportunity Performance (Top Right Section)

- Completed Sign Ups by Opportunity: Bar chart showing completion distribution
    - DiFi leads with about 833 completions
    - Followed by Project (255) and other opportunities
    - Shows a clear hierarchy in opportunity success rates

4. Opportunity Distribution (Bottom Left Section)

- Count of Opportunities: Details the variety of available programs
    - Data Visualization leads with 5.7K opportunities
    - Project Management follows with 3.9K
    - Digital Marketing (2.5K) and Health Care (1.5K) show significant presence
    - Diverse range including Innovation, Career Essentials, and specialized courses

5. Reward Structure (Bottom Center Section)

- Reward Amount by Opportunity: Shows financial distribution across programs
    - DiFi Visual shows highest reward amount at 0.94M
    - Project Management follows with 0.66M
    - Digital Marketing and Health Care maintain significant reward allocations

6. Geographic Distribution (Bottom Right Section)

- Top 10 City: Treemap showing participation across urban centers
    - Saint Louis shows significant presence
    - Lagos, Accra, Mumbai, and other cities represented
    - Suggests strong urban concentration of participants

Insights:

- Program Diversity: Wide range of opportunities from technical to healthcare sectors
- Geographic Focus: Strong presence in both African and South Asian urban centers
- Completion Challenges: Declining completion rates suggest need for intervention
- Reward Distribution: Significant investment in technical and project management tracks
- Urban Concentration: Clear focus on major metropolitan areas for opportunity distribution

This visualization highlights both the broad reach of the program and potential areas for improvement in completion rates and geographic expansion.

Relevance to stakeholders:

Operations Teams: Detailed count of opportunities by type helps in resource allocation and program management
City Planning Teams: Top 10 city distribution helps in local market strategy and resource deployment
Product Teams: Opportunity performance metrics (completion rates, sign-ups) help in program design improvements
Marketing Teams: Geographic data (Top 10 Countries) helps in targeting promotional efforts
Business Development: Reward amount distribution by opportunity type guides future program investment
Local Partners: City-level data helps in understanding regional performance and partnership opportunities

# 4. Conclusion

This report provides a comprehensive analysis of user engagement, educational impact, and operational performance, offering actionable insights for stakeholders. The findings highlight strong geographic and demographic trends, with India leading in user participation and rewards, and significant representation from undergraduate and graduate students. Seasonal trends in sign-ups and completion rates point to opportunities for improving user retention and engagement over time. Operationally, the data underscores the success of popular opportunities like DiFi AI L and DiFi Visual, while detailed geographic and program-specific metrics help identify growth areas and resource optimization opportunities. Together, these insights enable stakeholders to make data-driven decisions to enhance the program's scalability, impact, and effectiveness.