Week 1: Data Cleaning and Feature Engineering Report

Prepared By	Team D
Date	5/18/2025

1. Introduction

Purpose

The primary objective of this report is to document the data cleaning and feature engineering efforts completed during Week 1. These tasks aimed to ensure data integrity, enable insightful analysis, and prepare the dataset for machine learning applications, particularly for modeling learner engagement with various opportunities. The goal is to prepare the dataset for meaningful analysis that can answer questions such as:

- Which opportunities attract the most sign-ups?
- Which sign-ups are successfully completed most?
- What patterns lead to drop-offs?
- Are there seasonal trends in user engagement?

Data Description

The dataset includes sign-up information of learners applying to various opportunities. It contains the following key columns:

- **Learner details:** First Name, Date of Birth, Gender, Country, Institution Name, Major
- **Opportunity details:** Opportunity Name, Category, Start Date, End Date
- **Engagement tracking:** SignUp Date, Apply Date, Entry Created, Status Description, Status Code
- **New Features**: Age, Opportunity Duration, Standardized/Normalized metrics.

2. Data Cleaning Process

Tool Used for Data Cleaning

Microsoft Excel was utilized for data cleaning, preprocessing, and feature engineering tasks due to its flexibility with tabular data and built-in functions for text, date, and numeric transformations.

Cleaning Steps and Rationale

Column	Cleaning Action	Reason
Learner SignUp DateTime	Split into standardized **date** and **time**	To analyze engagement patterns and compute time lags between signup and program engagement.
Opportunity Category	One-hot encoded into multiple binary columns (Internship, Event, etc.)	To facilitate modeling and understand category-wise user preferences.
Opportunity End Date	Split into standardized **date** and **time**	To calculate duration of engagement for each user.
First Name	Standardized using the "Proper" text format	To clean outliers and ensure consistency in text-based analysis.
Date of Birth	Derived **Age** feature from it	To understand age group trends and patterns in opportunity engagement.
Gender	One-hot encoded	To compare gender-wise participation trends and integrate in future modeling.
Institution Name	Standardized text formatting	For easier grouping and analysis later.
Entry Created At	Split into **date** and **time**, standardized	To calculate response and engagement timeframes.
Status Description	One-hot encoded dropout entries	To detect dropout patterns and feed into predictive models.
Apply Date	Split and standardized	To compute delay from signup to application.
Opportunity Start Date	Split and standardized	To calculate lag time between signup and opportunity start.

3. Feature Engineering

New Features Created

Feature Name	Description	Rationale
Age	Calculated from Date of Birth	To analyze age-based participation trends.
Engagement Duration (Days)	Days between Learner SignUp and Entry Created	To measure how long users stay engaged.
Opportunity Duration (Days)	Days between Opportunity Start and Opportunity End	To understand duration-related impact on engagement.
SignUp Month / Year	Extracted from Learner SignUp Date	To assess seasonality in user engagement.
Age (Standardized/Normalized)	Standardized and scaled version of Age	For improved model performance.
Engagement Duration (Std/Norm)	Standardized and normalized versions of Engagement Duration	To prepare features for modeling and ensure uniform scale.
Opportunity Duration (Std/Norm)	Standardized and normalized versions of Opportunity Duration	To normalize the contribution of each feature in scoring models.
Combined Score	Composite metric derived from normalized features	To develop an aggregated metric for ranking user engagement or opportunity quality.

Feature Engineering Examples

1. **Age Calculation**:
    - Formula: Age = Current Year - Year of Birth
    - Helps analyze what age ranges are most active or successful.
2. **Engagement Duration**:

- Formula: Engagement Duration = Entry Created Date - Learner Sign Up Date
- Enables tracking user behavior from signup to meaningful interaction.

4. Data Validation

Validation Summary

Validation Checks

1. **Missing Values**: Confirmed no null entries post-cleaning.
2. **Duplicates**: Verified removal of redundant rows.
3. **DateTime Consistency**: Validated split columns for logical ranges (e.g., no future dates).
4. **Normalization Integrity**: Checked z-score distributions for new features.

Outcome

- Cleaned dataset with more than 23 columns, ready for exploratory analysis and modeling.

5. Conclusion

Summary

In Week 1, we successfully cleaned and transformed the original dataset to prepare it for analysis. Key tasks included:

- Handling and standardizing date/time formats.
- Creating new features to better capture user behavior and opportunity characteristics.
- Encoding categorical data for modeling purposes.
- Normalizing key features to enhance model interpretability and performance.

6. Link to the Clean Data:

https://docs.google.com/spreadsheets/d/1vS18ms_UIcig8lxH5T01aBGQJ8X5CTRfYnIwfEVFi0Y/edit?usp=drivesdk

Next Steps

In Week 2, we will:

- **Exploratory Data Analysis (EDA)**: Answer key questions (e.g., most popular opportunities, drop-off trends).
- **Predictive Modeling**: Use features like *Combined Score* to forecast engagement outcomes.
- **Seasonal Analysis**: Investigate time-based patterns using *SignUp Month/Year*.