Excelerate

AI POWERED DATA INSIGHTS INTERNSHIP

1108 | TEAM 3 | AI DATA INSIGHTS

Week 1: Data Cleaning and Feature Engineering Report

Link to Cleaned and Processed dataset - https://docs.google.com/spreadsheets/d/1AnDMMoeTQbm7XyMha-u-14lyr2c8SBNne01tndgwv0g/edit?usp=sharing

Introduction

Purpose:

The purpose of this report is to document the tasks completed during Week 1 of the internship. The focus was on data cleaning and feature engineering to ensure that the dataset is accurate, consistent, and enriched with new meaningful features that will be used for further analysis in the upcoming weeks.

Data Description:

The dataset used for this task was obtained from Excelerate Dashboard. It contains learner sign-up information, opportunity details, demographic information, and application records.

Key original columns include:

- Learner Sign-Up Date Time – Timestamp when a learner registered.
- Opportunity Id & Name – Unique identifiers and names of the opportunities.
- Opportunity Start/End Date – Duration of the program or course.
- Personal Information – First name, date of birth, gender, country, institution, and major.
- Application Details – Status description, status code, and application date.

Data Cleaning Process

Cleaning Steps Taken:

1. **Handling Missing Values** – Checked for missing values and imputed or removed records as needed.
2. **Removing Duplicates** – Removed duplicate rows based on Opportunity Id and Learner Sign-Up Date Time.
3. **Standardizing Formats** – Converted date columns into consistent datetime format.

4. **Data Type Corrections** – Converted numerical fields and ensured derived columns were numeric.

Issues Encountered & Resolution:

Some date columns had inconsistent formats, which were standardized. A few incomplete records were dropped to maintain dataset integrity.

New Features Created (Engineered)

Feature Name	Description	Rationale
Age	Age of learner at the time of application (from Date of Birth).	Allows age-based segmentation and demographic analysis.
Application_ Lag_Days	Days between learner signup and opportunity application.	Measures engagement, speed and promptness of learners.
Opportunity_ Duration_ Days	Duration of the opportunity in days (End Date – Start Date).	Useful for analyzing program lengths.
Region	Geographic region mapped from learner's country (e.g., Africa, Asia).	Enables regional analysis and cross-continent comparisons.
Signup_Season	Season of the year learner signed up (Spring, Summer, Fall, Winter).	Identifies seasonal signup/application patterns.
Tenure_Days	Days between learner signup and first recorded application.	Measures learner engagement lifecycle.

Age_Group	Age binned into ranges (18–22, 23–27, 28+).	Supports analysis by educational/career stage.
Apply_Year	Year extracted from Apply Date.	Enables year-over-year trend analysis.
Apply_Month	Month extracted from Apply Date.	Identifies monthly and seasonal variations.
Signup_Timing	Labels learners as Early or Late joiners based on signup relative to start date.	Provides insights into learner preparation and behavior.

Example Transformations

Application_Lag_Days

```
 df['Application_Lag_Days'] = (df['Apply Date'] - df['Learner SignUp DateTime']).dt.days
```

Calculates the number of days between signup and application submission.

Opportunity_Duration_Days

```
 df['Opportunity_Duration_Days'] = (df['Opportunity End Date'] - df['Opportunity Start Date']).dt.days
```

Finds how long each opportunity lasts, in days.

Data Validation

Validation Checks Performed:

1. Range Checks – Verified Age and Opportunity_Duration_Days values.
2. Consistency Checks – Ensured Apply Date follows Sign-Up Date Time.
3. Duplicate Checks – Confirmed no duplicates remain.

Validation Outcome:

All new features are consistent, no major anomalies remain, and the dataset is ready for analysis in Week 2.

Conclusion

During Week 1, the dataset was successfully cleaned and enhanced with ten new engineered features: Age, Application Lag, Opportunity Duration, Region, Signup Season, Days of Tenure, Age Group, Year of Application, Month of Application, and Signup Timing. This ensured consistency, removed redundancies, and prepared the dataset for analysis.

Next Steps (Week 2): *Perform Exploratory Data Analysis (EDA), begin segmentation and visualization, and prepare a dataset for predictive modeling.*