Week 1: Data Cleaning and Feature Engineering Report

Intern's Name: Zaid Shabir

Date of Submission: 17-Feb-2025

Submitted To:- Excelerate Team

Week-1 Report Overview

Contents

Introduction

Purpose:

The purpose of this report is to document the data cleaning, validation, and feature engineering processes applied to a dataset containing learner and opportunity information. During Week 1, the focus was on preparing the data for future analysis by handling missing values, correcting inconsistencies, and creating new features that provide deeper insights for predictive modeling.

Data Description:

The dataset consists of 8,558 records and 16 columns, including personal and opportunity-related information. Key columns include:

- First Name, Gender, Country, Date of Birth
- Opportunity Id, Opportunity Name, Opportunity Start Date, Opportunity End Date
- Status Description, Status Code

The dataset also contains several date fields that are crucial for feature engineering.

Data Cleaning Process

Cleaning Steps:

1. **Date Conversion:** Date fields such as Learner SignUp DateTime, Opportunity Start Date, Opportunity End Date, and Date of Birth were converted to datetime format for easier manipulation.
2. **Duplicate Removal:** The dataset was checked for duplicate entries, and any found were removed to ensure data accuracy.
3. **Missing Values Handling:**
4. **Institution Name:** Missing values were replaced with "Unknown."
5. **Opportunity End Date:** Missing values were replaced with the current date.
6. **Opportunity Start Date:** Missing values were filled using Opportunity End Date for completeness.
7. **Inconsistent Data:** Some columns, especially date-related ones, had inconsistencies or missing data that were handled through logical imputation.

-

Issues Encountered

Some fields, such as Opportunity Start Date, had a large proportion of missing values, which required filling based on related fields (e.g., Opportunity End Date).

Contributions:

- **Zaid Shabir and Akshaf Mehoob:** Performed the data cleaning and validation using Python scripts**.**
- **Anjelo Laroza:** Contributed to the data cleaning using a spreadsheet.
- **Haruna Muhammad and Eddy Timana:** Also worked collaboratively on the dataset.

```python
# 1. Data Cleaning and Validation

# Convert date columns to datetime format
date_columns = ['Learner SignUp DateTime', 'Opportunity End Date', 'Date of Birth',
                'Entry created at', 'Apply Date', 'Opportunity Start Date']
for col in date_columns:
    df[col] = pd.to_datetime(df[col], errors='coerce')

# Remove duplicates if any
df_cleaned = df.drop_duplicates()

# Check for missing values
missing_values = df_cleaned.isnull().sum()
print("Missing values in each column:\n", missing_values)

# Handle missing values
# Drop 'Institution Name' missing entries if needed (since there are only 4)
df_cleaned['Institution Name'].fillna('Unknown', inplace=True)

# For columns with larger missing values, decide to impute or remove them based on business logic
# Fill missing 'Opportunity End Date' and 'Opportunity Start Date' where necessary
df_cleaned['Opportunity End Date'].fillna(pd.Timestamp('today'), inplace=True)
df_cleaned['Opportunity Start Date'].fillna(df_cleaned['Opportunity End Date'], inplace=True)
```

Feature Engineering

New Features:

1. **Age:** Calculated from Date of Birth. This feature was created to understand the age of each learner based on the current year.
    - Rationale: Age is a critical demographic feature that could influence how learners engage with opportunities.
2. **Opportunity Duration (days):** Calculated by subtracting Opportunity Start Date from Opportunity End Date. Missing values in duration were filled with -1 as a placeholder.
    - Rationale: The duration of an opportunity may affect how learners are able to engage with and complete tasks, making it an important feature for future analysis.

Contributions:

- **Zaid Shabir and Akshaf Mehoob:** Performed the data cleaning and validation using Python scripts.
- **Anjelo Laroza:** Contributed to the data cleaning using a spreadsheet.
- **Haruna Muhammad and Eddy Timana:** Also worked collaboratively on the dataset.

Feature Examples

Age Calculation: Using Python:

- df_cleaned['Age'] = current_date.year - df_cleaned['Date of Birth'].year
- This transformation calculates the learner's current age by subtracting the year of birth from the current year.
- **Opportunity Duration:**
- df_cleaned['Opportunity Duration (days)'] = (df_cleaned['Opportunity End Date'] - df_cleaned['Opportunity Start Date']).dt.days

This computation assesses the duration of each opportunity in days, allowing for deeper analysis of time-related features.

```
# 2. Feature Engineering

# Feature 1: Calculate Age from Date of Birth
current_date = pd.Timestamp('today')
df_cleaned['Age'] = df_cleaned['Date of Birth'].apply(lambda dob: current_date.year - dob.year)

# Feature 2: Calculate Opportunity Duration (in days)
df_cleaned['Opportunity Duration (days)'] = (df_cleaned['Opportunity End Date'] - df_cleaned['Opportunity Start Date']).dt.days

# Fill missing 'Opportunity Duration' with a placeholder value, e.g., -1 for unknown durations
df_cleaned['Opportunity Duration (days)'].fillna(-1, inplace=True)

# Preview the new features
print(df_cleaned[['Age', 'Opportunity Duration (days)']].head())
```

Code For Feature Engineering

Data Validation

Validation Summary:

The dataset was validated through:

1. Duplicate Checks: Ensuring no redundant records were present after cleaning.
2. Missing Value Checks: Reviewed the missing value counts for each column before and after cleaning.
3. Date Validity: Confirmed that all date fields were converted correctly and that no invalid date formats remained after conversion.
4. Feature Checks: Verified the accuracy of new features such as Age and Opportunity Duration by performing spot checks on sample records.

Outcomes:

- Missing values were appropriately handled.
- No duplicates were present after the cleaning process.
- New features were verified for accuracy and consistency with the original data.

Contributions:

- **Zaid Shabir and Akshaf Mehoob:** Carried out the data validation using Python scripts.

```python
# 1. Data Cleaning and Validation

# Convert date columns to datetime format
date_columns = ['Learner SignUp DateTime', 'Opportunity End Date', 'Date of Birth',
                'Entry created at', 'Apply Date', 'Opportunity Start Date']
for col in date_columns:
    df[col] = pd.to_datetime(df[col], errors='coerce')

# Remove duplicates if any
df_cleaned = df.drop_duplicates()

# Check for missing values
missing_values = df_cleaned.isnull().sum()
print("Missing values in each column:\n", missing_values)

# Handle missing values
# Drop 'Institution Name' missing entries if needed (since there are only 4)
df_cleaned['Institution Name'].fillna('Unknown', inplace=True)

# For columns with larger missing values, decide to impute or remove them based on business logic
# Fill missing 'Opportunity End Date' and 'Opportunity Start Date' where necessary
df_cleaned['Opportunity End Date'].fillna(pd.Timestamp('today'), inplace=True)
df_cleaned['Opportunity Start Date'].fillna(df_cleaned['Opportunity End Date'], inplace=True)
```

Code For Cleaning As Well As Validation

Conclusion

Validation Summary: In Week 1, we successfully completed data cleaning, validation, and feature engineering using Python. By leveraging the powerful capabilities of libraries like pandas, we ensured that the dataset is clean, validated, and enriched with meaningful features that will support further analysis and predictive modeling.

Feature Checks: Verified the accuracy of new features such as Age and Opportunity Duration by performing spot checks on sample records. Outcomes: Missing values were appropriately handled. No duplicates were present after the cleaning process. New features were verified for accuracy and consistency with the original data.

Next Steps:

In Week 2, we will focus on exploratory data analysis (EDA) to uncover patterns and insights in the dataset.

Following EDA, we will begin building initial predictive models based on the cleaned and engineered dataset.