WEEK 1: DATA CLEANING AND FEATURE ENGINEERING REPORT

TEAM NAME: RIT 1410AI Team 5A

DATE: 21-10-2024

Contents

Team and Delegated Tasks

Team Member Name	Email ID
Abdullah Imran	abdullahimranarshad@gmail.com
Matthew Ojo	ojoaisosamatthew@gmail.com
Krishin Tharani	Krishintharani1+internships@gmail.com
Emani Likhita	likhi.m9363@gmail.com
Sangeeta Sahoo	sahoo1107.sangeeta02@gmail.com
John syllah	johnsyllah2003@gmail.com
Tracy Reson	tracyreson@gmail.com
Afra Falakh	afrafalakh16@gmail.com
Mathew OJo	ojoaisosamatthew@gmail.com
Eluit Cruz	ej.cruz.sant@gmail.com

1. Introduction

The purpose of this report is to document the data cleaning and feature extraction tasks completed during Week 1 to ensure that the dataset is in optimal condition for analysis.

While our dataset which came from an Excelerate platform contained an assortment of data on the activity and involvement of our customers, it needed to be carefully refined before it could be used for analytical procedures.

We used an efficient data processing technology built to effectively manage extensive data cleaning and analysis operations to clean our dataset. Python platforms with libraries like Pandas, NumPy, and Matplotlib, are typically preferred because of their flexibility and strong data manipulation capabilities. We have collectively decided to employ the Python platform to complete our given tasks.

We started by addressing missing values, a prevalent problem in datasets that can provide skewed or deceptive conclusions if ignored. We improved the completeness of our dataset by making sure that all required data points were included using meticulous identification. We also removed columns with very few rows of data.

The dataset was streamlined and its processing performance increased by eliminating features that were unnecessary to our analysis. Additionally, we used the data that already existed to build derived features.

In conclusion, during the first week, we prioritized data cleansing and feature extraction. We started by finding and addressing missing information and confirming that all data entries were full, ensuring data completeness and extracting relevant features to improve data quality and usability.

These steps were vital for laying the foundation for further analysis. These activities also guaranteed the dataset was clean, consistent, and suitable for analysis and modeling.

2. Data Description

We were provided with the collection which contains non-identifying information about every user who has ever created an Excelerate account. The Source was from the Excelerate. All users, regardless of whether they have seized certain offers, are included in the entire data.

Here’s a brief description of each of those key columns:

- **Profile Id**: A unique identifier assigned to each learner's profile.
- **Learner SignUp DateTime**: The date and time when the learner signed up.
- **Opportunity Id**: A unique identifier for each opportunity.
- **Opportunity Name**: The name of the opportunity.
- **Opportunity Category**: The category to which the opportunity belongs.
- **Opportunity End Date**: The date when the opportunity concludes.
- **First Name**: The learner's first name.
- **Last Name**: The learner's last name.
- **Date of Birth**: The learner's birth date.
- **Gender**: The learner's gender.
- **Mobile Number**: The learner's contact number.
- **Address Line 1**: The first line of the learner's address.
- **Address Line 2**: The second line of the learner's address (if any).
- **City**: The city where the learner resides.
- **State**: The state where the learner resides.
- **Country**: The country where the learner resides.
- **Zip Code**: The postal code of the learner's address.
- **Institution Name**: The name of the learner's educational institution.
- **Graduation Date**: The expected or actual graduation date of the learner.
- **Current Student Status**: Indicates whether the learner is currently enrolled as a student.
- **Current/Intended Major**: The learner's current or intended field of study.
- **Entry created at** The date and time when the entry was created.
- **Status Description**: Describes the current status of the opportunity.
- **Status Code**: A code representing the current status of the opportunity.
- **Apply Date**: The date when the learner applied for the opportunity.
- **Opportunity Start Date**: The start date of the opportunity.
- **Reward Awarded Date**: The date when a reward was awarded to the learner.
- **Completion Date**: The date when the learner completed the opportunity.
- **Reward Amount**: The amount of reward given to the learner.
- **Badge ID**: A unique identifier for each badge earned.
- **Badge Name**: The name of the badge earned.

- **Skill Points Earned**: The number of skill points earned by the learner.
- **Skills Earned**: The skills acquired by the learner.

A unique person is represented by each row, and the dataset offers a thorough picture of the user database.

3. Data Cleaning Process

The first step (for week 1) in the data cleaning process was creating a clean and usable version of the raw dataset which Excelerate provided. Making sure that the processed data was error-free, consistent, and didn't have any missing numbers was the aim of providing a strong basis for further analysis. We had to manage missing data, eliminate duplicate entries, fix formatting problems, and deal with any outliers or incorrect values were among the main actions performed and overcame during the data cleaning process.

3.1: Data Cleaning Steps:

1. **Handling Missing Values:** Substitution strategies, such as filling with mean/median values or eliminating rows with excessive missing values, were used to identify and manage missing data.
2. **Removing Duplicates:** To make sure the dataset had unique records, duplicate items were found using key columns and eliminated.
3. **Fixing Formatting Issues:** To ensure consistency throughout the dataset, inconsistent formats (such as date formats and inconsistent units) were standardized.
4. **Addressing Outliers and Inaccurate Values:** Inaccurate values were rectified to assure correctness, and outliers were examined and either eliminated or fixed.
5. **Data Type Standardization:** To guarantee consistency throughout the dataset, inconsistent data types were translated to the appropriate forms (e.g., numerical, categorical).
6. **Correcting Data Type:** This ensures accurate calculations and corrections, verifying and correcting data type of their appropriate formats (e.g, converting strings to dates)
7. **Conduct a final review and Documentation:** Review a cleaned data sheet to ensure all cleaning steps were correctly applied. Document each cleaning step taken, including decisions made and the rationale behind them for transparency and reproducibility.

3.2: Data Cleaning Specifics:

Dropping Missing Values: Missing values from the columns such as 'Address Line 1', 'City', 'State', 'Zip Code', 'Institution Name', and 'Current/Intended Major', ‘Graduation Date’ was dropped to minimize the missing data of the rows.

Dropping Address Line 2: The address Line 2 column had a lot of missing data since many people preferred to only provide information about the first address line. This was important since we dropped the data of missing values based on the specific values of this column, it would have resulted in data loss of many rows, which is what we are trying to avoid in this step.

Data Type Conversion: The data type of Zip Code, Date of Birth, Graduation Date, Opportunity End Date, Apply Date, and Opportunity Start Date was converted into datetime64[ns] format to handle time and date in the same column, which is crucial for data analysis process somewhere down the line.

4. Data Validation

Zip Code Validation: Dropping all Zip Codes that were non-numeric since generally, Zip Codes are numeric in real-world data, including the value 0 itself.

Missing Data: Dropping all of the columns in the dataset that had any missing values.

Name Validation: Dropping all of the rows where the names are non-alphabetical, including the first and the last names.

Mobile Number Validation: Since the Mobile Number on the website’s backend was passed as a string to the database, and then from the database to the sheet, it was enclosed in single quotation marks, which was not ideal, and hence these quotation marks were dropped.

Opportunity and its generated columns: This lies outside the scope of the validation since not all of the people who were selected for the internship opportunity availed of that offer and joined the team. Dropping the data based solely on this column would have resulted in a huge loss of data, hence, this step was skipped.

5. Feature Engineering

The website mentioned a few of the feature engineering techniques and steps to use, and hence those were preferred in our case. Following are some of the applied feature engineering techniques and steps:

5.1 Creating new features:

In this step, we create new features based on the existing dataset through some calculations and other techniques.

5.1.1 Age of Learner:

Based on the column “Date of Birth” which was already given in the dataset, it was first seen if the birthday of the user had already occurred in the ongoing year or not. Based on that, a new column “age of learner” was added using the calculated logic:

Age of Learner = Current Year−Birth Year−(I(Current Month<Birth Month)+I((Current Month=Birth Month) $\wedge$ (Current Day<Birth Day)))

- Where “I” is an indicator function that returns true or false.

5.1.2 Engagement Duration:

This calculates the engagement duration based on the date the difference between the date they applied and the date they started that opportunity. Those who never started have empty entries.

Engagement Duration (in days) = Opportunity Start Date − Apply Date.

5.2 Transforming the Existing Features:

5.2.1 Normalization of Engagement Metrics

This script normalizes the columns 'Age of Learner' and 'Engagement Duration' using *Min-Max Scaling* from the *sklearn.preprocessing* library. This is important and comes in handy during the machine learning tasks when we need to have data in a proper format.

X_scaled = (X - Xmin) / (Xmax - Xmin)

5.2.2 Encoding Categorical Data:

Converting the categorical columns into one-hot encoded variables, turning each unique category into a separate binary column using *pd.get_dummies(),* creating n-1 new binary columns.

Ci = {1 if C = ci; 0 otherwise}

5.3 Extracting Useful Components:

5.3.1 Date-Based Features

Creates a new column called SignUp Month that contains the month extracted from the Learner SignUp DateTime. The result is stored in the new column SignUp Month, which will have values between 1 (January) and 12 (December), representing the month the learner signed up.

5.3.2 Opportunity Engagement:

Calculates the duration of each opportunity in terms of days based on the start and end dates of the opportunity; converts both the Opportunity Start Date and Opportunity End Date columns to a proper DateTime format to enable date calculations. Then it computes the time difference between the Opportunity End Date and the Opportunity Start Date for each opportunity. And finally, it converts the time delta object into the total number of days to make the duration easier to interpret.

5.4 Combing Features

5.4.1 Interaction Features:

Introduces an interaction feature by multiplying two existing features: 'Age of Learner' and 'Time in Opportunity (Days)'. Interaction features are useful for capturing relationships between variables that may not be apparent when considered individually.

It first computes the age of the learner based on their date of birth. Then it multiplies 'Age of Learner' by 'Time in Opportunity (Days)' to create a new interaction feature.

Age of Learner = (Today's date - Date of Birth) / 365

Interaction feature = Age of learner x Time in Opportunity = Age of Learner × Time in Opportunity (Days)

5.4.2 Engagement Scores:

This calculates an Engagement Score for each student based on their 'Time in Opportunity' and 'Age of Learner'. First, it converts the 'Time in Opportunity' column to a format that represents total days. Then, it calculates an engagement score based on the 'Time in Opportunity' and 'Age of Learner'.

Time in Opportunity (Days) = Time Opportunity is seconds / 86400

Engagement Score = 0.4 × Time in Opportunity (Days) + 0.3 × Age of Learner

5.5: Temporal Analysis

5.5.1: Seasonal Patterns

This analyzes seasonal patterns in the Engagement Score of learners based on the month of sign-up and the day of the week. First, it creates new columns that capture the month and day of the week from the learner's sign-up date, and then it calculates the average engagement score for each month to identify seasonal trends.

average_monthly_engagement = students.groupby(’SignUp Month’)[’Engagement Score’].mean().reset_index()

```
weekly_engagement = students.groupby('SignUp Day of Week')['Engagement
Score'].mean().reset_index()
```

5.6: Miscellaneous Feature Engineering Steps:

5.6.1: Time Since Last Engagement:

This calculates the number of days since the last engagement for each student based on their sign-up date and the last engagement date derived from the opportunity start dates.

Days Since Last Engagement = Learner SignUp DateTime − Last Engagement Date

5.6.2: Target Variable Creation

This creates a target variable called 'High Engagement' based on the 'Engagement Score of each student to be able to establish a threshold to classify engagement levels so as to classify students into two categories based on whether their engagement score is above or below the threshold.

Engagement Threshold = median(Engagement Score)

High Engagement = {1 if Engagement Score > engagement_threshold; 0 otherwise}

6. Conclusion

Our first week of data preparation was productive and demanding, setting the foundation for a solid and trustworthy dataset. We prepared the groundwork for perceptive analysis by carefully cleaning the data to guarantee that each record was correct, comprehensive, and consistent.
Our dataset is now a clean, consistent, and potent resource prepared to deliver significant insights and choices owing to the thorough methodology implemented during Week 1.

In the following week, our approach will concentrate on delving into data visualization and exploratory data analysis (EDA) and in terms of visualization, we will map the data to identify patterns and trends.

By Applying various visualization techniques we aim to reveal key insights that can inform strategic planning and decision-making processes.

Moreover, we recognize that effective Data cleaning is an ongoing process that requires regular updates and continuous improvement. This commitment to maintaining data integrity will enable us to adapt to new challenges and opportunities as they arise, ensuring our dataset remains a valuable asset for future analyses.

As we move forward our goal is to transform the cleaned data into actionable insights that will enhance user experience and drive growth for the Excelerate platform. We look forward to the next stages of our analysis and potential discoveries that lie ahead.