WEEK 1 : DATA CLEANING AND FEATURE ENGINEERING REPORT

NAME : Syed Hur Abbas Naqvi
DATE : 23RD JUNE,2025

INTRODUCTION

Purpose: The purpose of this report is to present a clear and accurate overview of user data from Excelerate by thoroughly cleaning and validating the dataset using various tools.

During Week 1, the primary focus was on data cleaning, which involved handling missing values, addressing outliers, standardizing formats, correcting errors, removing duplicates, and resolving inconsistencies in categorical data. In addition to this, data validation was also performed to confirm that all entries met predefined standards and rules.

Feature engineering was introduced as well, which helps in creating new features or modifying existing ones from a raw dataset to enhance the performance of machine learning models. The goal is to extract relevant information and transform it into a format that can improve model accuracy and predictive power.

Data Description: The dataset draws on anonymized details from every individual who has registered on the Excelerate platform. It reveals notable patterns in user demographics and academic focus. Male users, numbering over 5000, outnumber their female counterparts, who exceed 3000, reflecting a gender distribution that is relatively balanced but still tilted toward males. This suggests that while the platform appeals to both genders, it has a stronger pull among males.

Academic interests are concentrated in technology-driven fields, with Information Systems emerging as the predominant major, closely followed by Computer Science. Most users were born between 2000 and 2002, with 2001

being the most common birth year. This indicates that the typical user is a young adult, likely at the beginning of their higher education or early in their professional journey.

Geographically, the user base is primarily concentrated in the United States, closely followed by India. There is also representation from Africa, Australia, and various Asian countries, though these numbers are comparatively lesser. Collectively, the data illustrates Excelerate's appeal to a youthful and tech-oriented audience.

DATA CLEANING PROCESS

1. Since the only missing cells were all in the "Opportunity Start Data" column, those rows were just removed since users were mostly just rejected.
2. There are no outliers since all columns are categorical and not really quantitative-focused.
3. All dates were ensured to be in the format MM/DD/YYYY HH:MM:SS
4. All student names were cleaned (converted foreign characters into normal letters, removed last names) and deleted rows with names that were not convertible.
5. All universities were cleaned (removed duplicates, combined similar names for the same university, grouped students who were not going to college into one category "None", and removed students who just put "students").
6. All majors were cleaned (removed duplicates, removed all items that were not majors, and removed the degree (i.e. bachelor/masters) to keep just the major)
7. All columns were left - aligned for consistency purposes.

FEATURE ENGINEERING

The purpose of feature engineering the cleaned dataset is to improve the analytical insights and model performance by transforming existing raw data into more meaningful and interpretable attributes.The new features created are:

1. **Learner/ Intern's Age Calculation:**
   Excel functions were used to calculate the age from the Date of Birth. For example, the formula =DATEDIF(G2, TODAY(), "Y") where G2 is the cell containing the Date of Birth. This feature helps understand age distribution among users as well analysing behavioural patterns across different age categories.
2. **Opportunity/ Engagement Duration:**
   The duration of each opportunity was computed by subtracting the Opportunity Start Date from the Opportunity End Date. Use the formula =E2 - P2 where E2 is the Opportunity End Date and P2 is the Opportunity Start Date.This was computed to find the number of days between Apply Date and Opportunity Start Date.
   This feature provides insights into the length or number of days between a learner or interns opportunities and their impact on engagement.
3. **Application Frequency:**
   This new feature describes how many opportunities each learner or intern has applied to overtime.
   This feature measures the activity level of the applicant and their engagement with the platform.

Feature Engineering Techniques

Examples of feature engineering techniques useful in transforming this data are temporal analysis and binning.|

Temporal Analysis is useful in identifying patterns in engagement based on time and can help in scheduling opportunities and targeting specific periods for enhanced participation. For example, the pattern in the application days, months and start date from the application date could be analysed to detect patterns.

Binning is useful for converting continuous numerical variables into categories that are labelled. Ages of interns or learners could be grouped into Teenagers, Young Adults and Adults based on a specific age range.

DATA VALIDATION

The cleaned dataset was validated using a Python script to ensure accuracy and consistency, employing several checks to identify data quality issues. The validation process is used to prepare the dataset for reliable analysis by addressing inconsistencies, missing values, and logical errors in key columns. The following checks were performed to assess data integrity, with outcomes indicating areas for improvement.

Validation Checks and Outcomes:
The validation process included the following checks:

• Invalid Values: Key columns, including name, gender, country, institution, status code, and application date, all were examined for null entries.Invalid values were detected in some records.

• Gender: Each Gender Value was verified against a predefined list of acceptable categories: 'Male', 'Female', or 'Other' (case-insensitive). Eleven records contained the invalid entry 'Don't want to specify', which shows a need for standardization or mapping to a valid category. No missing values were found in this column.

• Date of Birth: Entries were validated for correct date formats and completeness. All records contained valid, properly formatted date values.There were no missing entries, which shows high data quality in this column.

• Country: The column was assessed for missing values and consistency in naming conventions. There were no missing values but some inconsistencies were noted, like variations in country names.

• Status Code: Entries were checked to ensure they all were numeric and matched expected values (1070, 1080, 1110). All records contained valid numeric status codes.There were no missing values which shows consistent data entry for this field.

Ensuring Accuracy and Consistency Through Data Validation:

Data validation ensures the accuracy and consistency of a dataset by checking whether the values are correct, complete, and properly formatted. It helps identify and fix errors such as missing data, invalid entries, or incorrect formats. This process ensures accuracy by making sure the data truly represents what it is supposed to. At the same time, validation enforces consistency by standardizing categories, correcting naming variations (like country or gender), and making sure similar data is always entered in the same way. As a result, the dataset becomes more reliable and easier to analyze.

CONCLUSION

Summary: The key outcomes of this week's work include successfully cleaning and validating the Excelerate user dataset to ensure accuracy and consistency. Missing values were identified and addressed, such as filling in incomplete names or removing records lacking essential information. Outliers were detected and managed to prevent a skewed analysis.

Typographical errors and inconsistent categorical entries, such as unusual or misspelled words were corrected. Data validation checks confirmed that entries met the required standards. Additionally, initial feature engineering was performed, such as creating new variables to better capture user engagement patterns. These steps established a reliable foundation for deeper analysis of the dataset.

Next Steps: In Week 2, the focus will shift to Exploratory Data Analysis (EDA) using the cleaned dataset. We will use various techniques to understand the data's structure, identify important variables, and uncover patterns that could inform further analysis or predictive modelling.

The cleaned dataset is essential for further analysis because it ensures accuracy, consistency, and reliability in results. Once the data has been cleaned, it can confidently be used for Exploratory Data Analysis (EDA).