Interim report:

10 Academy: Artificial Intelligence Mastery

Week 8&9 Challenge Document

Date: 05 Feb - 18 Feb 2025

Improved detection of fraud cases for e-commerce and bank transactions

By
Simone Asnake

Date: Feb 08/2025

Version 0.01

Contents

List of Acronyms

KAIM: Kifya Artificial Intelligence Mastery

CDR: Challenge document report

Min: Minimum

Max: Maximum

Std: standard deviation

EDA: Exploratory Data Analysis

M: male

F: female

SEO: search engine optimization

Freq: frequency

IP: internet protocol

Executive summary

❖ Data Overview:

- The dataset contains 151,112 entries with 11 columns.
- Column data types include integers, objects (strings), and floats.
- Columns include user_id, signup_time, purchase_time, purchase_value, device_id, source, browser, sex, age, ip_address, and class.
- RangeIndex: 151112 entries, 0 to 151111: This line indicates that the DataFrame has a total of 151112 rows, with row indices ranging from 0 to 151111.
- Data columns (total 11 columns): This signifies that the DataFrame consists of 11 columns.

❖ purchase_value column statistics:

- Count: 151,112
- Mean: 36.94
- Standard Deviation: 18.32
- Min: 9, Max: 154

❖ age column statistics:

- Count: 151,112
- Mean: 33.14
- Standard Deviation: 8.62
- Min: 18, Max: 76

❖ sex column breakdown:

- Two unique values: 'M' and 'F'
- Most frequent value: 'M' with a frequency of 88,293

❖ Column Descriptions:

- signup_time: Most common value is '2015-02-24 22:55:49'.
- purchase_time: Most common value is '2015-06-08 09:42:04'.
- Source: Most common source is 'SEO' with a frequency of 60,615.

❖ Data Quality Check:

- There are no missing values in any of the columns.

Project objective

- Utilize the dataset to identify patterns or anomalies that could indicate fraudulent activities.
- Build models to predict and prevent potential fraudulent transactions based on user behaviour and purchase patterns.

Tools and packages/libraries

- pandas
- numpy
- matplotlib.pyplot
- matplotlib
- seaborn
- scipy.stats
- zscore
- from mpl_toolkits.mplot3d import Axes3D

Methodologies

- **Data Loading and Inspection**: The notebook starts by loading the dataset and inspecting its structure and content.
- **Descriptive Statistics**: The notebook calculates and explains descriptive statistics for various columns.
- **Data Visualization**: Histograms and other visualizations are used to explore the distribution of numerical and categorical variables.
- **Missing Value Analysis**: The notebook checks for and confirms the absence of missing values in the dataset.
- **Exploratory Data Analysis (EDA)**: The notebook provides a comprehensive overview of the dataset, including the distribution of key variables and the relationship between different features.

Findings

Null values for all columns:

- **RangeIndex: 151112 entries, 0 to 151111**: This line indicates that the DataFrame has a total of 151112 rows, with row indices ranging from 0 to 151111.
- **Data columns (total 11 columns)**: This signifies that the DataFrame consists of 11 columns.
- **Column details**:
    1. **user_id**: It contains 151112 non-null integer values.
    2. **signup_time**: It contains 151112 non-null date time values represented as objects.

3. **purchase_time**: It contains 151112 non-null date time values represented as objects.
4. **purchase_value**: It contains 151112 non-null integer values.
5. **device_id**: It contains 151112 non-null values represented as objects.
6. **Source**: It contains 151112 non-null values represented as objects.
7. **Browser**: It contains 151112 non-null values represented as objects.
8. **Sex**: It contains 151112 non-null values represented as objects.
9. **Age**: It contains 151112 non-null integer values.
10. **ip_address**: It contains 151112 non-null float values.
11. **Class**: It contains 151112 non-null integer values.

Sex column:

- **Count**: This represents the total number of non-null values in the "sex" column, which is **151112.**
- **Unique**: It indicates the number of unique values present in the "sex" column. In this case, there **are 2 unique values** (**'M' and 'F'**).
- **Top**: This shows the most frequently occurring value in the "sex" **column, which is 'M'.**
- **Freq**: It displays the frequency of the top value 'M' in the "sex" column, which is **88293**. This means 'M' appears **88293 times** in the dataset.

"purchase_value" column:

- **count**: This represents the total number of non-null values in the "purchase_value" column, which is 151112.
- **mean**: The mean (average) purchase value in the "purchase_value" column is approximately 36.94.
- **std**: The standard deviation of the values in the "purchase_value" column is around 18.32. This indicates the variability or dispersion of values around the mean.
- **Min**: The minimum purchase value in the column is 9. This is the smallest value present in the dataset.
- **25%**: This is the first quartile (Q1) value, which means that 25% of the values in the "purchase_value" column are below 22.
- **50%**: This is the second quartile or median value. It indicates that 50% of the values in the "purchase_value" column are below 35.
- **75%**: This is the third quartile (Q3) value, showing that 75% of the values in the "purchase_value" column are below 49.
- **Max**: The maximum purchase value in the column is 154. This is the largest value present in the dataset.

The "purchase_time" column:

- **Count**: This represents the total number of non-null values in the "purchase_time" column, which is 151112.
- **Unique**: It indicates the number of unique values present in the "purchase_time" column, which is 150679. This means that there are repeated timestamps in the column.
- **Top**: This shows the most frequently occurring value in the "purchase_time" column, which is '2015-06-08 09:42:04'.
- **Freq**: It displays the frequency of the top value '2015-06-08 09:42:04' in the "purchase_time" column, which is 3. This means that '2015-06-08 09:42:04' appears three times in the column, making it the most common timestamp.

The "source" column:

- **Count**: This indicates the total number of **non-null values** in the "source" column, which is **151112.**
- **Unique**: It represents the number of unique values present in the "source" column, which is 3. This suggests that there are **only 3 unique sources in the dataset**.
- **Top**: The "top" value signifies the most frequently occurring value in the **"source" column, which is 'SEO'.**
- **Freq**: This shows the frequency of the top value **'SEO'** in the "source" column, which is **60615**. This means that **'SEO' appears 60615** times in the column, making it the most common source.

Next action points

- Additional EDA for credit and IP dataset
- Feature Engineering Implement
- Default estimator and WoE binning
- Modelling