Centre for Machine Intelligence and Data Science (CMInDS)
Indian Institute of Technology Bombay

Programming for Machine Learning and Data Science

Assignment: Exploratory Data Analysis (EDA)

---

23 Mar, 2025

Dataset: Payment Fraud Detection

Dataset link: https://drive.google.com/file/d/1gXW4hqSThkNeLJQnQZatctvzj05bkgqV/view?usp=sharing

You have been provided with a fraud detection dataset containing various features and a target variable indicating **fraudulent** (1) or **non-fraudulent** (0) transactions. Your task is to understand and perform exploratory data analysis on this dataset and answer the following questions:

1. Create and analyze descriptive statistics for each column in the dataset. Include measures such as mean, median, standard deviation, min/max, 25th / 50th / 75th percentiles, skewness, and kurtosis values for numerical columns, and frequency counts and mode for categorical columns. Based on the descriptive statistics, summarize your findings about each feature.

2. Analyze missing values in the dataset:

   (a) Identify which features contain missing values and quantify them both in terms of rows and columns.

   (b) Create appropriate visualization(s) to clearly display the pattern of missing values across the entire dataset. (Hint: can you create a plot which contains a rectangular block representing the entire dataset, with missing values highlighted in a different colour within this block?)

   (c) Based on your observations, propose and implement appropriate handling strategies for the missing values. Justify your approach for each feature.

   (d) After implementing your strategy, report the final data size. Are there any rows or columns that were completely dropped? Explain why.

3. Explore the distribution of numerical features:

   (a) Create histograms with overlaid Kernel Density Estimation (KDE) plots for all numerical features.

   (b) Create similar plots, but this time separate ones for fraudulent and non-fraudulent transactions. What differences do you observe?

   (c) Based on your observations, what transformations would you recommend for any skewed numerical features? Justify your recommendations.

4. Investigate outliers in numerical features:

   (a) Create individual box plots for all numerical columns to identify potential outliers.

   (b) Highlight fraudulent observations in a different color within these box plots. Are outliers more common in fraudulent or non-fraudulent transactions?

   (c) Create box plots of all numerical columns on a common scale to visualize any scale differences between features. What conclusions can you draw about the relative scales and distributions?

5. The "Location" feature contains cities like Mumbai, Delhi, and Bangalore etc. Calculate the fraud rate for each location and create appropriate visualizations to show these rates. Identify if certain locations show significantly higher fraud risk. Explain how this insight could inform fraud detection strategies.

6. Data normalization and standardization:

   (a) Normalize all numerical columns and create histograms and box plots of the normalized features. How do these compare to the original distributions? Provide adequate explanations.

   (b) Standardize all numerical columns to have mean 0 and standard deviation 1. Create histograms and box plots of the standardized features. How do these compare to the original and normalized distributions? Provide adequate explanations.

   (c) Explain the differences observed and discuss which approach (original, normalized, or standardized) might be most appropriate for this dataset.

7. Based on all your analysis above, and considering the domain of fraud detection:

   (a) What strategy would you recommend for handling outliers in this dataset? Justify your approach.

   (b) Summarize the key insights you've discovered about the patterns of fraudulent transactions in this dataset.

   (c) What pre-processing steps would you recommend before building a machine learning model for fraud detection using this data?

Report Preparation and Submission Guidelines

1. **Submission due by: 30/03/2025, 11:55pm**
2. Submission process: will be conveyed ...
3. Solutions to all the above problems should be created in a single **Notebook** (eg. Colab Notebook).
4. In this Notebook all the explanations / analysis / conclusions should be done in **Markup** cells and all program segments should be in **Code** cells.
5. The first cell of the Notebook **should** contain a submission code - which will be unique to every participant. It will be shared with you by email.
   - You should ensure that your name / email / any other identifying mark **is not** present anywhere in the Notebook.

---

oooOOOooo