Exploratory Data Analysis (EDA)

Introduction

Exploratory Data Analysis (EDA) is the first step in understanding a dataset. It involves summarizing key characteristics, identifying patterns, and detecting anomalies. This notebook aimed to uncover insights that would guide further modeling decisions.

Data Cleaning and Preprocessing

- **Handling Missing Values**: Strategies such as mean imputation and forward-fill were used.
- **Outlier Detection**: Boxplots and Z-score analysis helped identify extreme values that could distort analysis.
- **Feature Transformation**: Skewed distributions were normalized using log transformations where necessary.

Descriptive Statistics and Visualization

- **Univariate Analysis**: Histograms, boxplots, and frequency distributions were used to study individual variables.

- **Bivariate and Multivariate Analysis**: Correlation heatmaps and scatter plots helped in understanding relationships between variables.
- **Categorical Data Analysis**: Count plots and bar charts displayed the distribution of categorical features.

Insights and Findings

- Strong correlations between certain features indicated potential multicollinearity issues for predictive modeling.
- Significant patterns and trends were identified, forming a foundation for further analysis.
- Key business insights were extracted, such as customer purchasing behaviors or fraud detection indicators.

Conclusion and Future Work

- The clustering analysis successfully segmented the data into meaningful groups, aiding in strategic decision-making.

- The lookalike modeling approach proved effective for identifying similar users, demonstrating its value for targeted marketing.
- The EDA phase provided deep insights into the dataset, setting a strong foundation for predictive modeling.
- Future work could explore advanced deep learning models, automated feature engineering, and real-time deployment of these methodologies.

Exploratory Data Analysis (Kumar_Sahil_EDA.ipynb)

This notebook focuses on EDA, analyzing dataset characteristics using statistical summaries, visualizations, and correlation analysis. It includes data cleaning steps like handling missing values, outlier detection, and transformation techniques. The notebook uses histograms, boxplots, and scatter plots to understand variable distributions and relationships. Advanced analysis might involve feature engineering and hypothesis testing to uncover patterns in the data. The conclusions provide a strong foundation for further modeling or decision-making based on dataset insights.