**Analysis Writeup  
Banking Course-End Project 2**

**Objective**

The objective of this project is to analyze the factors influencing vehicle loan defaults and to develop a predictive model that identifies potential defaulters. This study is crucial for financial institutions to mitigate risks associated with vehicle loan defaults, improve underwriting decisions, and optimize loan approval processes.

**Approach and Findings**

**1. Data Preliminary Analysis**

* **Data Structure and Inspection**:
  + The dataset contains 41 attributes, capturing demographic, financial, and loan-related details.
  + Initial inspection identified:
    - Presence of missing values in critical variables.
    - Duplicate entries that need removal.
    - Inconsistencies in variable naming conventions, which were standardized for compatibility with Python's syntax.
* **Strategies for Handling Missing Values**:
  + Numeric variables with missing values were imputed using statistical measures like the mean or median.
  + Categorical variables were filled using the mode or based on patterns observed in the data.

**2. Exploratory Data Analysis (EDA)**

* **Target Variable Analysis**:
  + The distribution of defaulters and non-defaulters was studied. The overall default rate provided a baseline for understanding the severity of defaults in the dataset.
* **Categorical Variable Insights**:
  + **Branch, City, and State**:
    - Certain geographic regions exhibited higher default rates, indicating potential location-based risk factors.
  + **Supplier and Manufacturer**:
    - Relationships between specific manufacturers or suppliers and default probability were analyzed.
* **Employment Type Analysis**:
  + The dataset categorized customers by employment types (e.g., salaried, self-employed). Pie charts highlighted significant differences in default rates across employment categories. Missing employment data was imputed based on other variables like income and loan history.
* **Age Distribution**:
  + Age was found to correlate with default probability, with younger borrowers showing higher default rates. Histograms showed clear distribution differences between defaulters and non-defaulters.
* **Customer Identification Proofs**:
  + Most customers presented a specific type of ID proof (e.g., Aadhar, PAN). Variability in the type of ID presented did not show a strong correlation with defaults.

**3. Key Factors Affecting Loan Defaults**

* **Credit Bureau Scores**:
  + Defaulters generally had lower credit bureau scores, with a skewed distribution compared to non-defaulters. Box plots and density plots captured these differences.
* **Loan History**:
  + Variables such as the number of loans defaulted in the past six months, time since the first loan, and new loans in the last six months significantly impacted default probability.
* **Loan Disbursement vs. Sanction**:
  + The gap between sanctioned and disbursed loan amounts was larger for defaulters, as observed through statistical summaries and visualizations.
* **Inquiry Numbers**:
  + Borrowers with higher numbers of credit inquiries were more likely to default, indicating a potential signal of financial stress.

**4. Modeling**

* **Logistic Regression**:
  + A logistic regression model was employed to predict defaulters. Key predictors included credit bureau scores, loan history variables, age, and employment type.
* **Model Validation**:
  + The model's performance was evaluated using a confusion matrix. Metrics such as accuracy, precision, recall, and F1-score were calculated to assess predictive performance.

**Conclusions and Recommendations**

1. **Key Determinants of Loan Defaults**:
   * Low credit scores, younger age, frequent loan inquiries, and specific employment types were found to be significant predictors of default.
2. **Model Effectiveness**:
   * The logistic regression model demonstrated satisfactory predictive performance, providing actionable insights for loan approval and risk mitigation strategies.
3. **Recommendations for Financial Institutions**:
   * Tighten eligibility criteria based on credit bureau scores and loan history.
   * Enhance customer profiling by incorporating geographic and demographic factors.
   * Monitor and limit the number of credit inquiries for high-risk customers.

This analysis provides a robust foundation for improving vehicle loan underwriting processes and developing more effective credit risk models.