Credit Risk Analysis – Insights Summary

1. Dataset Overview

- The dataset originally contained **5,960 loan applicants** across **13 variables**.
- The target variable is **`BAD`**, renamed to `default_flag`, indicating loan default (`1`) vs. non-default (`0`).

---

2. Data Cleaning & Preparation

- Variables like `DEBTINC`, `MORTDUE`, `DEROG`, etc., had missing values.
- Applied a **data-aware imputation strategy**:
  - **Mean imputation** for `DEBTINC`, `CLNO`
  - **Median imputation** for skewed variables (`MORTDUE`, `VALUE`) o
    Zero imputation for `DEROG`, `DELINQ`
  - **Dropped rows** where imputation wasn't appropriate •
    Removed duplicates and renamed columns for clarity.

---

3. Exploratory Analysis & Outlier Treatment

- Outliers were treated using the **IQR method (clipping)** for all numeric predictors except the target.
- Visual distributions were inspected before imputation to ensure appropriate assumptions.

---

4. Feature Engineering

- Categorical variables (`loan_reason`, `job_type`) were encoded via **one-hot encoding**.
- A total of **4,831 cleaned records** were used for modeling after preprocessing.

---

5. Modeling Approach

- The dataset was split into **70% training** and **30% testing** using **stratified sampling** to preserve default distribution.
- Class imbalance handled using **class weights** `{0: 0.2, 1: 0.8}` in tree-based models.

Models Built:

- **Logistic Regression**
- **Decision Tree** (baseline and tuned with GridSearchCV)
- **Random Forest** (baseline and tuned)

---

6. Evaluation Metric

- Emphasis placed on **Recall for the default class (1)** to **maximize the capture of high-risk applicants**.
- Other metrics: Accuracy, Precision, and F1-score (reported for both training and test sets).

---

7. Model Comparison

Model	Accuracy (Test)	Recall (Default)	Precision (Default)
Logistic Regression	79.6%	2.0%	66.7%
Decision Tree (Base)	85.7%	57.7%	67.9%
Decision Tree (Tuned)	86.6%	71.4%	66.1%
Random Forest (Base)	90.0%	60.1%	87.3%
Random Forest (Tuned)	**90.2%**	**62.4%**	**86.1%**

---

8. Key Insights

- **Random Forest (Tuned)** offered the best **balance between accuracy and recall**, making it most suitable for detecting defaults.
- **Top features** influencing default risk include:
  - `recent_credit_inquiries`
  - `years_on_job`
  - `open_credit_lines`
  - `debt_income_ratio`

---

9. Business Recommendation

- Deploy the **tuned Random Forest model** to flag risky applicants early.
- Use **feature importances** for targeted interventions (e.g., assessing job stability or credit behavior).
- Combine this model with **credit officer judgment** for a robust loan approval process.