Loan Default Prediction - Code Section Explanation

STEP 1: Install Required Libraries

Installs required packages:
- `xgboost`, `lightgbm`: Boosting algorithms
- `shap`: Explainable AI tools

STEP 2: Import Libraries

Imports essential libraries for data processing, visualization, modeling, and evaluation.

STEP 3: Load Dataset

Loads the HMEQ dataset. The target column is 'BAD' (1 = default, 0 = not default).

STEP 4: Handle Missing Values

Processes missing values:
- Drops rows with missing 'BAD'
- Fills numeric columns with median
- Fills categorical columns with mode
- Label encodes categorical variables

STEP 5: Data Visualization

Displays:
- Correlation matrix to show feature relationships
- Countplot for class distribution of the target variable

STEP 6: Data Splitting and Scaling

Splits data into train/test sets and normalizes features using StandardScaler.

STEP 7: Train ML Models

Trains models: Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM. Evaluates them with accuracy, precision, recall, F1, and AUC.

STEP 8: Summary Table

Creates a performance comparison table for all models.

STEP 9: Threshold Optimization & ROC/PR Curves

Plots Precision vs Recall and ROC curves to visualize threshold tuning and classification performance.

STEP 10: GridSearchCV for Decision Tree

Tunes hyperparameters (max_depth, min_samples_split) to optimize recall.

STEP 11: Feature Importance

Displays the top 10 most important features from the best Decision Tree.

STEP 12: Boxplot Analysis

Shows how a selected feature's values differ by default status using boxplot.

Error Inspection

Examines False Negatives and False Positives to better understand model misclassifications.

Manual Thresholding (XGBoost)

Adjusts decision threshold (e.g., to 0.3) to reduce false negatives.

Final Evaluation: Confusion Matrix

Visualizes the confusion matrix for XGBoost to analyze TP, TN, FP, FN.