**1. Data Preprocessing for Loan Default Prediction**

The dataset was cleaned and prepared for modeling through the following steps:

Missing data was handled by removing rows and columns with excessive missing values.

Categorical variables were one-hot encoded, and percentage columns were converted to numerical values.

Data types were corrected, and missing numerical values were imputed based on skewness.

Duplicate entries were removed.

**2. EDA and Feature Selection for Loan Default Prediction**

The analysis focused on performing exploratory data analysis (EDA) and feature selection for a loan default prediction dataset. Key steps included:

Target Variable Distribution: A class imbalance was found in the target variable, bad\_flag, with more non-default loans, requiring potential techniques like oversampling for model balance.

Numerical Feature Distribution: Key features like loan amount and annual income showed right-skewness, indicating the presence of outliers.

Correlation Analysis: A heatmap revealed relationships between numerical features, helping to identify redundant variables.

Feature Importance: A Random Forest model was used to rank features, identifying the top 15 most important ones for prediction.

Feature Selection Comparison: The Random Forest method and Chi-squared test both selected similar top features, validating their importance.

Anomaly Detection: The Isolation Forest algorithm identified a small number of anomalies, which need to be addressed before modeling.

**3. XGBoost Reference Model for Loan Default Prediction**

The XGBoost model was developed for predicting loan defaults, addressing class imbalance with SMOTEENN, which balances the dataset by generating synthetic minority samples. Stratified K-Fold cross-validation was used to ensure fair evaluation. Key performance metrics included F1 score, AUC score, and ROC curve. After training, the model was tested with a 0.6 threshold for predictions. Class distribution was checked to ensure balanced predictions. The model shows good generalization and performance despite the class imbalance.

**4. Neural Network Model for Loan Default Prediction**

*NN Model with 1 hidden layer*

A neural network model was built using PyTorch for loan default prediction. The architecture includes an input layer, a hidden layer with ReLU activation, dropout for regularization, and a Sigmoid output for binary classification. It was trained using Binary Cross-Entropy loss and the Adam optimizer for 50 epochs.

Performance was evaluated using F1 score and ROC AUC, with Stratified K-Fold Cross-Validation ensuring good generalization. The model made predictions on test data using a threshold of 0.75. The model showed strong performance and generalization in predicting loan defaults.

*NN Model with 2 hidden layers*

The neural network for predicting loan defaults used a binary classification approach with the following key components:

**Architecture:**

Input Layer: Size based on dataset features.

Hidden Layers: Two layers with tunable neurons and ReLU activation.

Output Layer: One neuron with Sigmoid activation for binary output.

Dropout: Applied to prevent overfitting.

**Training:**

Loss Function: Binary Cross-Entropy.

Optimizer: Adam for efficient training.

Epochs: Trained for 30 epochs with full-batch processing.

Hyperparameter Tuning: Optimized neuron count and learning rate via grid search.

**Evaluation:**

Cross-Validation: Stratified K-Fold with 5 splits.

Metrics: F1 score (0.7802) and AUC score (0.7948).

Test Inference: Predicted outcomes with a threshold of 0.85, maintaining class distribution similar to training data.

The model achieved strong performance with high F1 and AUC scores, generalizing well to unseen data, indicating its effectiveness for loan default prediction.