Loan Risk Classification Report

Author: Hemant Mogia

Approach Taken

- **Data Loading and Cleaning:**
    - Train and test data were loaded from CSV files.
    - Relevant columns were selected for model training.
    - Missing values were imputed using techniques like KNN imputation and interpolation.
    - Categorical features were encoded using LabelEncoder.
    - Numerical features were scaled using StandardScaler.
- **Feature Selection:**
    - Correlation analysis identified highly correlated features.
    - Univariate feature selection used the f-classif scoring function.
    - Feature importances were extracted from a Random Forest Classifier model.
    - Recursive Feature Elimination (RFE) was employed with a Random Forest Classifier.
    - L1 regularization with a LinearSVC model was used.
- **Model Building and Evaluation:**
    - A Random Forest Classifier model was used for loan risk classification.
    - The model was trained and evaluated using various performance metrics, including accuracy, ROC AUC score, precision, recall, and F1 score.
    - The model's performance was evaluated on different feature sets selected through the feature selection techniques.
- **TensorFlow Model Implementation:**
    - A TensorFlow Sequential model with multiple hidden layers and activation functions was built and trained.
    - Early stopping technique was used to prevent overfitting.
    - The model was evaluated on the validation data using the classification report metric.

Insights and Conclusions from Data

- Feature selection techniques helped identify a critical subset of features for predicting loan risk, potentially improving model efficiency and interpretability.
- The model achieved good performance on the training data set using various evaluation metrics.

Model Performance

Metric	Value
Accuracy	0.8623
Loss	0.3070
Validation Accuracy	0.8577
Validation Loss	0.3047

Classification Report on Validation Data:

Class	Precision	Recall	F1-score	Support
0	0.76	0.83	0.80	1995
1	0.91	0.87	0.89	4005
Accuracy			0.86	6000
Macro Avg	0.84	0.85	0.84	6000
Weighted Avg	0.86	0.86	0.86	6000

Analysis:

- The model achieved an overall accuracy of 86.23% on the validation data, indicating a good performance.
- The precision and recall for both classes are relatively balanced, suggesting a good trade-off between identifying true positives and minimizing false positives/negatives.
- The F1-score, which considers both precision and recall, is also satisfactory for both classes.
- The macro average and weighted average of these metrics further confirm the model's overall performance.