Machine Learning Report: Loan Default Prediction

1. Data Preparation and Exploration

- **Assumptions:**
  - Target variable is 'fpd_15'
  - The data represents a snapshot of loan details and outcome (fpd_15) at a particular point in time.
- **Dataset Overview**: The dataset, initially saved as a CSV but tab-delimited, was loaded into a pandas DataFrame. The unnamed index column was dropped.
- **Exploratory Data Analysis**: Using the `sweetviz` module, extensive exploratory data analysis was conducted. Key observations generated include: bar charts of values of each attribute, centrality and spread measures, and correlation matrix.
- **Missing Data Analysis**:
  - Approximately 50% of the attributes have missing values, ranging from 4% to 82%.
  - High missing rates were found in attributes from salary and previous application data.
  - Correlation between missing values was significant, indicating data from the same sources were often missing together.
  - Attributes with more than 30% missing data were dropped after creating indicator columns.

2. Data Cleaning and Preprocessing

- **Date Handling**: Converted date features (CreationDate and DateOfBirth) into numerical values representing loan age and customer age in months.
- **Data Cleaning**: Standardized state names to lowercase, reducing unique states from 45 to 37.
- **Handling Missing Data**:
  - Imputation was applied for numerical features with 4%-11% missing data using mean imputation.
  - For categorical features, missing data was imputed after target encoding.
- **Outlier Removal**: Used Local Outlier Factor to remove extreme values in income, electricity spend, and other features, reducing the dataset from 3621 to 3584 observations.
- **Feature Encoding**: Categorical features were target encoded to numerical values to facilitate model training.
- **Class Imbalance**: Addressed using random oversampling, balancing defaulters and non-defaulters to 50% each.
- **Feature Selection**:
  - Features were screened for correlation with the target variable and among themselves.
  - Removed features with less than 2% correlation with the target and highly correlated features.

3. Model Building and Evaluation

- **Model Selection**: Various models were evaluated using 4-fold cross-validation:
  - **Random Forest (RF)** showed the highest performance with an accuracy of 81.77%.
- **Hyperparameter Optimization**: Optimized RF hyperparameters using cross-validation, resulting in the best model configuration. Selected Random Forest with parameters: {'bootstrap': True, 'class_weight': {0: 1, 1: 1.1}, 'criterion': 'entropy', 'max_depth': 4, 'max_features': 2, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 80}.
- **Recursive Feature Elimination**: Identified that 8 features were optimal for the RF model after hyperparameter optimisation, improving performance and reducing redundancy.
- **Model Testing**:
  - **Accuracy**: 78% on the test set.
  - **Confusion Matrix**: Shows a slight tendency to misclassify non-defaulters as defaulters.
  - **Classification Report**:
    - Precision: 0.73 (defaulters), 0.84 (non-defaulters)
    - Recall: 0.82 (defaulters), 0.75 (non-defaulters)
    - F1-Score: 0.77 (defaulters), 0.79 (non-defaulters)
  - **ROC AUC Score**: 0.79, indicating good model performance in distinguishing between classes.

4. Conclusion

The Random Forest model demonstrates strong performance in predicting loan defaults with an accuracy of 78% and a balanced confusion matrix. Improvements can be made by increasing both the amount of data available and also the number of features utilized concurrently.