Credit Card Default Prediction Low-Level Document

1. Data Exploration:

1.1 Display Top and Last Rows:

- **Objective:** Understand the structure and content of the dataset.
- **Explanation:** Displaying the top and last rows helps to observe the initial and final entries, ensuring data import and structure are as expected.

1.2 Dataset Shape:

- **Objective:** Understand the size of the dataset.
- **Explanation:** Identifying the number of rows and columns provides a basic understanding of the dataset's scale.

1.3 Dataset Information:

- **Objective:** Obtain detailed information about columns, data types, and memory usage.
- **Explanation:** Dataset information provides insights into column data types, non-null counts, and memory requirements, aiding in subsequent data processing decisions.

1.4 Null Values Check:

- **Objective:** Identify missing values in the dataset.
- **Explanation:** Checking for null values is crucial for addressing data quality issues, guiding the imputation or removal of missing data.

2. Data Preprocessing:

2.1 Feature Matrix and Response Vector:

- **Objective:** Prepare the data for model training by separating features and the target variable.
- **Explanation:** Creating a feature matrix (X) and response vector (Y) establishes the input variables and the variable to predict.

2.2 Train-Test Split:

- **Objective:** Divide the dataset into training and test sets for model evaluation.

- **Explanation:** Splitting the data ensures independent datasets for training and assessing model performance.

3. Handling Imbalanced Dataset:

3.1 Undersampling:

- **Objective:** Address class imbalance by reducing the majority class instances.
- **Explanation:** Undersampling mitigates the impact of class imbalance, improving model training on the minority class.

3.2 Oversampling:

- **Objective:** Address class imbalance by increasing the minority class instances.
- **Explanation:** Oversampling ensures sufficient representation of the minority class, enhancing the model's ability to learn from it.

4. Model Development:

4.1 Logistic Regression:

- **Objective:** Develop a logistic regression model for credit card default prediction.
- **Explanation:** Logistic regression is chosen for its simplicity and interpretability, making it suitable for binary classification tasks.

4.2 Decision Tree Classifier:

- **Objective:** Implement a decision tree classifier.
- **Explanation:** Decision trees capture non-linear relationships and are interpretable. Suitable for understanding feature importance.

4.3 Random Forest Classifier:

- **Objective:** Utilize a random forest classifier.
- **Explanation:** Random forests improve upon decision trees by aggregating multiple trees, enhancing predictive performance.

5. Model Evaluation:

5.1 Performance Metrics:

- **Objective:** Assess model performance using accuracy, precision, recall, and F1-score.

- **Explanation:** Performance metrics provide a comprehensive evaluation of the model's ability to predict credit card defaults.

6. Save the Model:

6.1 Save Logistic Regression Model:

- **Objective:** Persist the trained logistic regression model for future use or deployment.
- **Explanation:** Saving the model allows for easy integration into production systems without the need for retraining.

6.2 Save Decision Tree Model:

- **Objective:** Save the decision tree model.
- **Explanation:** Saving models ensures consistency across different environments and enables reproducibility.

6.3 Save Random Forest Model:

- **Objective:** Save the random forest model.
- **Explanation:** Similar to other models, saving the random forest model facilitates deployment and sharing across different platforms.