Supplementary material

Glossary:

Data Loading and Summary Statistics:

- **Readtable:** A function used to import data from a CSV file and organize it into a structured table format.
- **Size:** A function that determines the number of rows and columns in the dataset.
- **Summary:** Provides an overview of basic statistical measures for individual columns within the dataset.
- **Missing Values:** Identifies and quantifies any missing or null values present in the dataset.
- **Duplicates:** Identifies and counts rows that are exact duplicates of one another within the dataset.

SMOTE (Synthetic Minority Over-sampling Technique):

- **Class Imbalance:** The imbalance between different classes within the dataset, often addressed through techniques like SMOTE.
- **SMOTE Algorithm:** Details the synthetic oversampling algorithm utilized to balance the minority class by generating artificial samples.
- **Visualization:** Graphical representations illustrating the class distribution before and after applying SMOTE.

Feature Selection and Analysis:

- **Correlation Analysis:** The method used to compute the relationships between features in the dataset.
- **Top Positively and Negatively Correlated Features:** Features exhibiting the strongest positive and negative relationships with the target variable, respectively.
- **Feature Statistics:** Statistical measurements like minimum, maximum, standard deviation, and mean calculated for selected features.

Model Training and Evaluation:

- **Logistic Regression and Random Forest:** Explanation of the process of training these models, including cross-validation techniques and performance evaluation using metrics like accuracy, precision, recall, and F1-score.
- **K-Fold Cross-Validation:** A technique that partitions data into k subsets to train and validate models iteratively.

Model Metrics and Evaluation:

- **Average Metrics:** Explanation of how performance metrics averaged across multiple folds were computed for robust model evaluation.
- **Best Model Metrics:** Details regarding the performance metrics achieved by the best-performing Logistic Regression and Random Forest models.
- **Confusion Matrix:** Description of the matrix displaying true positives, true negatives, false positives, and false negatives for model evaluation.

File Saving:

- **CSV File Generation:** Describes the process of saving training and testing data into CSV (Comma-Separated Values) files for future use or reference.

Intermediate Results:¶

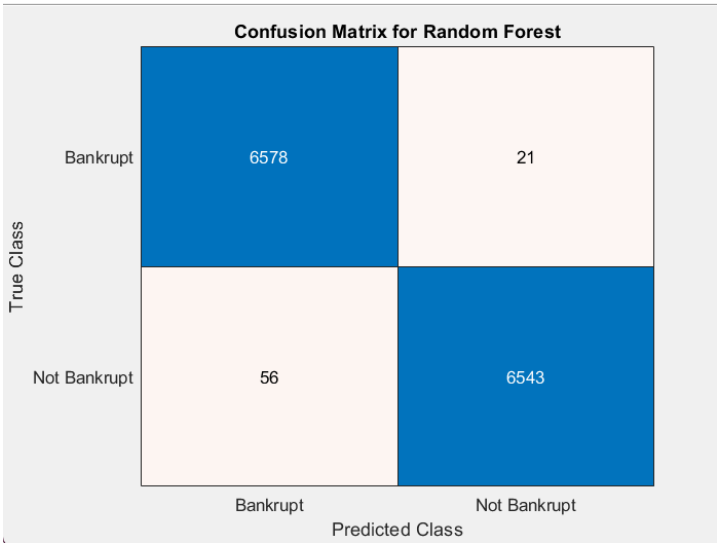
- **Size of data** : 6819 rows x 96 columns
- **Missing Values Summary**: No missing values in the data.
- **Duplicate Rows**: No duplicate rows in the data .
- **Class Counts Before SMOTE**:
  - **Class 0** : 6599
  - **Class 1 :** 220
- **Class Counts After SMOTE**:
  - **Class 0** : 6599
  - **Class 1 :** 6599
- **Selected Features**:
  - NetIncomeFlag
  - DebtRatio_
  - CurrentLiabilityToAssets
  - CurrentLiabilityToCurrentAssets
  - TotalExpense_Assets
  - BorrowingDependency
  - FixedAssetsTurnoverFrequency
  - LiabilityToEquity
  - EquityToLong_termLiability
  - CurrentLiabilities_Equity
  - NetWorth_Assets
  - ROA_C_BeforeInterestAndDepreciationBeforeInterest
  - PersistentEPSInTheLastFourSeasons
  - ROA_B_BeforeInterestAndDepreciationAfterTax
  - NetProfitBeforeTax_Paid_inCapital
  - ROA_A_BeforeInterestAnd_AfterTax
  - PerShareNetProfitBeforeTax_Yuan__
  - NetIncomeToTotalAssets
  - NetValuePerShare_B_
  - NetValuePerShare_A_
- **Statistics for Selected Features**:

Column_Name	Min	Max	Standard_Deviation	Mean
{'NetIncomeFlag' }	0	1	0.077769	0.46355
{'DebtRatio_' }	0	1	0.093626	0.51032
{'CurrentLiabilityToAssets' }	0	1	0.083494	0.50949
{'CurrentLiabilityToCurrentAssets' }	0	1	0.030781	0.17604
{'TotalExpense_Assets' }	0	1	0.030891	0.17597
{'BorrowingDependency' }	0	1	0.036189	0.20947
{'FixedAssetsTurnoverFrequency' }	0	1	0.03425	0.16679
{'LiabilityToEquity' }	0	1	0.06293	0.14859
{'EquityToLong_termLiability' }	0	1	0.06293	0.85141
{'CurrentLiabilities_Equity' }	0	1	0.045051	0.38235
{'NetWorth_Assets' }	0	1	0.032462	0.16586
{'ROA_C_BeforeInterestAndDepreciationBeforeInterest'}	0	9.99e+09	2.8127e+09	1.4587e+09
{'PersistentEPSInTheLastFourSeasons' }	0	1	0.057224	0.11655
{'ROA_B_BeforeInterestAndDepreciationAfterTax' }	0	1	0.039873	0.33693
{'NetProfitBeforeTax_Paid_inCapital' }	0	1	0.040747	0.03925
{'ROA_A_BeforeInterestAnd_AfterTax' }	0	1	0.050646	0.12294
{'PerShareNetProfitBeforeTax_Yuan__' }	0	1	0.037287	0.044863
{'NetIncomeToTotalAssets' }	0	1	0.065943	0.77471
{'NetValuePerShare_B_' }	0	1	0.042408	0.2868
{'NetValuePerShare_A_' }	1	1	0	1

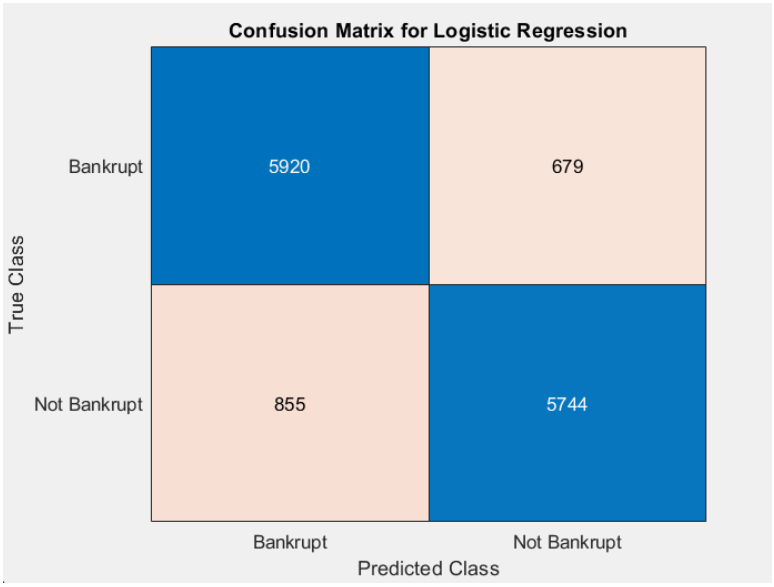
- **Performance Metrics**:
  - Average Metrics across Folds for Logistic Regression:

        - Average Accuracy: 0.88938
        - Average Precision: 0.88267
        - Average Recall: 0.89794
        - Average F1 Score: 0.89023
    - Average Metrics across Folds for Random Forest Classifier:
        - Average Accuracy: 0.96386
        - Average Precision: 0.94534
        - Average Recall: 0.98472
        - Average F1 Score: 0.9646
- **Best Model Metrics**:
    - Metrics of the Best Logistic Regression Model:
        - Accuracy: 0.90417
        - Precision: 0.90097
        - Recall: 0.90977
        - F1 Score: 0.90535
    - Metrics of the Best Random Forest Model:
        - Accuracy: 0.96591
        - Precision: 0.95322
        - Recall: 0.98045
        - F1 Score: 0.96664

- **Confusion Matrices**:
    - Confusion matrix for Random Forest Classifier:



- Confusion Matrix for Logistic Regression:



Negative Results or Considerations:

- The code doesn't include hyperparameter tuning, which could impact model performance.
- It assumes k=5 for nearest neighbors and 100 trees for the Random Forest without exploring the optimal values.
- There might be limitations due to the assumed threshold of 0.5 for rounding predictions to binary classes.
- The code saves CSV files but doesn't provide insights into the data split for training and testing.

Implementation Choices:

- SMOTE and Imbalanced Data Handling: Employing KNN for SMOTE to address class imbalance.
- Feature Selection: Using correlation analysis for feature selection.
- Model Selection: Choosing Logistic Regression and Random Forest for classification tasks.