The analysis presented here focuses on evaluating the performance of two machine learning models, specifically logistic regression models, using the provided data. Going through the process and summarize the results.

Purpose of the Analysis:

The purpose of the analysis is to assess the effectiveness of logistics regression models in predicting loan statuses, specifically identifying healthy loans (labeled as 0) and high-risk loans (labeled as 1).

Financial Information and Prediction Task:

The data provided includes information about loans, and the goal is to predict whether a loan is healthy or high-risk based on the available features.

Variable Information:

The variable loan status serves as the target variable, with two distinct values: 0 and 1. Here's a summary of the target variable's distribution.

- Count of healthy loans (0): 75036
- Count of high-risk loans (1): 2500

Stages of the Machine Learning Process:

1. Splitting the Data: the dataset is divided into training and testing sets using the train_test_split function from scikit-learn.
2. Logistics Regression with Original data:
    a. The logistics regression model is fitted using the training data.
    b. Predictions are made on the testing data.
    c. Performance is evaluated using metrics such as balanced accuracy score, confusion matrix, and classification report.
3. Logistics Regression with Resampled Data:
    a. Random oversampling using the RandomOversampler module from the imbalanced-learn library is applied to the training data to address class imbalance.
    b. The logistic regression model is fitted using the resampled training data.
    c. Predictions are made on the testing data.

Results:

Machine Learning Model 1: Logistic Regression with Originals Data

- Balanced accuracy score: 95.20%
- Confusion Matrix: 18663 was determined as healthy loans with low risk and was the actual results, 102 loans were predicted as low risk but were vice-versa high

risks. 563 loans were predicted correctly as non-healthy loans, as opposed to the remaining 56 loans.

- Classification Report: were all in the 90% percentile.

Machine Learning Model 2: Logistics Regression with Resampled data

- Balanced accuracy score: 99.37%
- Confusion matrix: 18649 was predicted correctly as healthy loans and 4 incorrectly. 615 predicted correctly as non-healthy loans, and 116 incorrectly.
- Classification report: like model 1, all categories was in the 90% percentile.

Summary:

Based on the results obtained, we can make the following observation:

- Machine Learning Model 1, which uses original data, yielded the following performance:
  - Accuracy: 95%
  - Precision: 100% and 85%
  - Recall: 99% and 91%
- Machine Learning model 2, which utilized the resampled data, produced the following performance:
  - Accuracy: 99%
  - Precision: 100% and 84%
  - Recall: 99% for both

Considering the performance metrics, it appears that model 2 performs better in terms of accuracy, precision, and recall. However, the choice of the model depends on the problems at hand. If the goal is to prioritize identifying high-risk loans (1), then focusing on the recall score for high-risk label would be crucial. Alternatively, if balanced accuracy is the primary concern, a comprehensive evaluation of precision and recall for both labels is important.

In summary, model 2 is recommended due to its superior performance in the given metrics. However, the ultimate choice of the model should be aligned with the specific objectives and priorities of the problem being addressed.