**Overview of the Analysis**

In this analysis, we aimed to build machine learning models to predict loan statuses based on financial information. The dataset provided us with several features, including loan size, interest rate, borrower income, debt-to-income ratio, number of accounts, derogatory marks, and total debt. The objective was to classify loans into two categories: healthy loans (0) and high-risk loans (1). We performed the analysis using two different approaches: one with the original data and another with resampled data.

First, we divided the data into training and testing sets. The training set was used to fit a logistic regression model, while the testing set was used to evaluate the model's performance. The logistic regression model is a widely used classification algorithm that estimates the probability of an instance belonging to a specific class.

In the first approach, we trained the logistic regression model with the original data. The model was fitted using the training data, and predictions were made on the testing data. We evaluated the model's

performance using several metrics, including the balanced accuracy score, which provides an overall measure of how well the model predicts both classes. Additionally, we generated a confusion matrix to analyze the number of correct and incorrect predictions, and a classification report to examine precision, recall, and F1-score for each class.

In the second approach, we addressed the class imbalance issue in the dataset by resampling the data using the RandomOverSampler module from the imbalanced-learn library. This module generates synthetic samples for the minority class (high-risk loans) to balance the distribution of classes. The logistic regression model was then trained on the resampled data and evaluated using the same metrics as in the first approach.

**Results**

Machine Learning Model 1: Logistic Regression with Original Data

* Balanced accuracy score: 0.9520479254722232
* Precision for high-risk loans (1): 0.85

• Recall for high-risk loans (1): 0.91  
Machine Learning Model 2: Logistic Regression with Resampled Data

* Balanced accuracy score: 0.9945026387334079
* Precision for high-risk loans (1): 0.99
* Recall for high-risk loans (1): 0.99

**Summary**

Upon comparing the two machine learning models, it is evident that the logistic regression model trained with the resampled data (Model 2) outperforms the model trained with the original data (Model 1) in terms of accuracy, precision, and recall for high-risk loans. Model 2 achieved a significantly higher balanced accuracy score of approximately 99%, indicating its superior ability to predict loan statuses.

Considering the nature of the problem, where correctly identifying high- risk loans (1) is of utmost importance, the higher precision and recall values achieved by Model 2 make it the preferred choice. This model

demonstrates a remarkable capability to accurately classify both healthy and high-risk loans.

Based on these results, we recommend utilizing the logistic regression model trained with the resampled data (Model 2) for predicting loan statuses. Its excellent performance in identifying high-risk loans suggests that it would be an effective tool for assessing loan risk and making informed decisions.

It's important to note that the choice of the model ultimately depends on the specific requirements and priorities of the lending institution or the organization utilizing the predictions. However, the results clearly indicate that Model 2 offers superior performance in predicting loan statuses, making it a valuable solution for managing loan portfolios and minimizing the risk of defaults.