**Overview of Analysis**

The purpose of this analysis is to create a model that predicts whether a loan recipient is at risk. The model uses variables loan size, interest rates, borrower income, debt-to-income ratio, number of accounts, derogatory marks, and total debt as independent variables that are used to explain loan status. A score of zero indicates the loan is healthy. A score of one indicates the loan is at risk.

In the first part of the analysis, a Logistics Regression Model is used to predict loan risk. This is a statistical method used for predicting binary outcomes from data. Once the data is processed, it follows the steps: train, validate, and predict.

1. Train: The independent and dependent variables are split into training and testing data. The training data is used to teach the model to recognize classification patterns. The loan model is fitted to the training data.
2. Validate: Using a subset of the data, the model is tested to see how well it can predict labels.
3. Predict: The model uses the independent variables in the loan model to predict the dependent variable labels.

In the second part of the analysis uses RandomOverSampler in the Logistics Regression Model. This is a technique that is used when one class (in this case high-risk) has significantly fewer instances. This method fixes the instances issue by randomly duplicating rows until there is an equal number of value counts. It follows the same train, validate, and predicting steps above.

**Results**

Logistics Regression Model with Original Data:

* Accuracy = 99%
* Classification Results:
  + Healthy
    - Precision = 100%
    - Recall = 99%
  + High-Risk
    - Precision = 86%
    - Recall = 95%

Logistics Regression Model with Reclassified Data:

* Accuracy = 99%
* Classification Results:
  + Healthy
    - Precision = 99%
    - Recall = 99%
  + High-Risk
    - Precision = 99%
    - Recall = 99%

The accuracy score is how often the model is correct. It is the ratio of correctly predicted observations to the total number of observations. In both cases, the model performs very well in predicting classifications. Precision in the ratio of correctly predicted observations to the total predicted positive observations. In the LRM with Original data, precision falls for high-risk observations. Finally, recall is the ratio of correctly predicted positive observations to all predicted observations for that class. Similarly to precision, recall falls for high-risk in the original data model.

**Summary**

The Reclassified Data LRM performs best with its high accuracy, precision, and recall scores. The original data is skewed with a high number of data points for high-risk than healthy. This is having negative effects on the precision and recall of the model.

When it comes to predicting healthy loans, the model performs very well across all summary statistics. This is important in minimizing the number of healthy loans that are classified as high risk which could waste time for loan officers and banks. On the other hand, the model performs slightly worse on classifying high-risk loans. This could increase risk for a portfolio if high-risk loans are classified as healthy.