**Credit Risk Classification Report**

**Overview of the Analysis**

The main objective of this analysis is to develop a logistic regression model to determine whether loans are healthy or high-risk. Minimising the risk of defaulting a loan is major concern and task for financial institutions such as banks, lending organisations and financial service companies.

Machine Learning models help these organizations and institutions to identify debtors in advance and improve their capabilities to minimise the risk of defaulting a loan.

These organisations increasingly relying on technology to predict which clients are more prone to stop honouring their debts. They use methodologies such as logistic regression model to predict the status of a relevant loan, the result of which will be compared with the actual loan status. The loan status will indicate whether a loan is a healthy or high-risk.

**Stages of the Process:**

1. Read the data set from the CSV file into a Pandas dataframe.
2. Create the labels set (y) from the “loan\_status” column, and then create the features (X) DataFrame from the remaining columns. Separate data into labels
3. Check the balance of labels variable using the value\_function count
4. Split the data into training and testing datasets by using train\_test\_split.
5. Create a Logistic Regression Model with the original data
6. Fit a logistic regression model by using the training data (x-train and y\_train).
7. Save the predictions on the testing data labels by using the testing feature data (x-test)and the fitted model.
8. Evaluate the model’s performance by balanced accuracy score, a confusion matrix and a classification report for the model using sklearn.metrics.
9. Predict a logistic regression model with resampled training data
10. Use the RandomOverSampler module from the imbalanced-learn library to resample the training data so that the labels have an equal number of data points.
11. Use the logistic regression classifier and resampled data to fit the model to make prediction using the LogisticRegression module.
12. Evaluate the model’s performance by balanced accuracy score, a confusion matrix and a classification report for the model using sklearn.metrics.

**Results**

* **Machine Learning Model (1)**
* The model has a balanced accuracy of 0.949
  + Healthy loan [0]
  + It has a precision of 1 for healthy loans, means that all loans classified by the model as healthy.
  + Recall is 1 for healthy loans, that means for all are healthy loans.
  + High Risk loan [1]
* This has a precision of 0.87, which means that for all loans classified by the model as high-risk.
* The model has a recall of 0.90 for high-risk loan, means that for all high-risk loans, 90% of them are classified by the model as high-risk.
* **Machine Learning Model (2)**
* This model has a balanced accuracy of 0.996
  + Healthy loan [0]
* For this model too has a precision score of 1 for healthy loans, that means for all loans are healthy.
* Like the previous model, this too has a recall of 1, which means all are healthy loans.
  + High Risk Loan [1]
* This model has a precision score of 0.87, that means for all loans classified by the model as high-risk.
* This has a recall score of 1 for high-risk loan, means all are high-risk loans, and all of them are classified by the model as high-risk loans.

**Summary**

Machine Learning Model 2 has a higher recall for high-risk loans and a higher overall accuracy as compared to Model 1. Machine Learning Model 2 is recommended compared to the model 1 due to much higher recall for high-risk loan. Means this is more accurately classify all actual high-risk loans as high-risk. This is useful if we want to use the model to determine if a loan is potentially high-risk.

Still there could be incorrect predictions such as false positives and false negatives.