# Credit Risk Classification Analysis Report

## Overview of the Analysis

The purpose of the analysis is to build a predictive model using historical lending activity dataset that can determine the creditworthiness of borrowers.

The datasets comprise of features or attributes about the borrower’s

* loan amount,
* interest rate,
* income,
* debt to income ratio,
* number of accounts
* derogatory marks
* total debt

that can be used to predict the loan status of the borrowers. There are total of 77536 loan records with 75036 loans classified as healthy loan and 2500 as the high-risk loans.

In order to build the model, logistic regression modeling is leveraged to predict the binary outcomes of the loan status, whether a borrower’s loan should be classified or categorized as healthy loan or high risk loan.

Here are the steps taken in implement the logistic regression modeling:

Preprocessing:

By reading in the lending data in a csv format using the pandas data frame module/dependencies. And separated the data into label and features. The label (y/dependent variable) is the loan status that would like the model to predict and the features (x/independent variables) are the remaining columns/features in the data frame that are fitted in model to determine what is loan status (y variable).

Two datasets are created from the original dataset to perform the hold out validation that is applied using the test\_train\_split function: by splitting the data into two subsets for training and testing the model. The train to test split ratio is 75% or 3:4 which means 75% of the data is leveraged train the model and 25% of the data is used to measures the performance of our model and parameter selections.

Training & validate the model:

* The logistic model regression is generated using sklearn packages.
* Instantiate the model and fit the training data (75%) into the model
* Made predictions of the loan status using the test subset of data (25%)

Evaluate the model performance through:

* Calculating the accuracy score of the model
* Generating a confusion matrix and classification report

Using sklearn packages for each step in evaluating how good is the model in predicting the loan status.

To optimize the performance of the model based on the accuracy score and classification report, the training data is resampled using the RandomOverSampler to achieve a balanced class distribution between healthy and high risk loan record within the training data subset.

## Results

Machine Learning Model 1 without resampling:

For healthy loan model prediction performance

* Overall accuracy score = 99%
* Overall precision score = 100%
* Overall recall score = 100%

For high-risk loan model prediction performance

* Overall accuracy score = 99%
* Overall precision score = 87%
* Overall recall score = 89%

Machine Learning Model 2 with resampling:

For healthy loan model prediction performance

* Overall accuracy score = 99%
* Overall precision score = 100%
* Overall recall score = 100%

For high-risk loan model prediction performance

* Overall accuracy score = 99%
* Overall precision score = 87%
* Overall recall score = 100%

## Summary