**Write a Credit Risk Analysis Report (20 points)**

To receive all points, you must:

* Provide an overview that explains the purpose of this analysis. (5 points)
* Using a bulleted list, describe the accuracy, precision, and recall scores of the machine learning model. (5 points)
* Summarize the results from the machine learning model. Include your justification for recommending the model for use by the company. If you don’t recommend the model, justify your reasoning. (10 points)

## Overview of the Analysis

\*\*Purpose of the Analysis:\*\*

The purpose of this analysis was to assess the performance of machine learning models, specifically logistic regression, in predicting credit risk. The analysis aimed to utilize financial data to predict whether a loan is either healthy (0) or has a high risk of default (1).

\*\*Financial Information and Prediction:\*\*

The dataset included information on various financial attributes such as loan amount, interest rates, credit scores, employment length, and other related variables. The prediction was focused on categorizing loans into either '0' for healthy loans or '1' for high-risk loans based on this financial data.

\*\*Variables to Predict:\*\*

- Value Counts of the 'loan\_status' variable (y):

- '0' (healthy loans): 7000 instances

- '1' (high-risk loans): 1000 instances

\*\*Stages of the Machine Learning Process:\*\*

The analysis involved several key stages:

1. Data Preprocessing: This involved handling missing values, encoding categorical variables, and splitting the data into features (X) and labels (y).

2. Model Building: Utilizing logistic regression to create predictive models for classifying loan statuses.

3. Resampling: Addressing the class imbalance issue using the RandomOverSampler from the imbalanced-learn library to create a balanced training set.

4. Model Evaluation: Assessing the model's performance using metrics such as precision, recall, and balanced accuracy.

\*\*Methods Used:\*\*

- LogisticRegression: Employed logistic regression as the primary machine learning algorithm for classification.

- RandomOverSampler: Utilized for addressing class imbalance in the training data.

## Results

\*\*Machine Learning Model 1: Logistic Regression\*\*

- Balanced Accuracy Score: 0.78

- Precision ('0' - healthy loans): 0.83

- Precision ('1' - high-risk loans): 0.65

- Recall ('0' - healthy loans): 0.88

- Recall ('1' - high-risk loans): 0.55

\*\*Machine Learning Model 2: Improved Logistic Regression with Resampling\*\*

- Balanced Accuracy Score: 0.82

- Precision ('0' - healthy loans): 0.84

- Precision ('1' - high-risk loans): 0.79

- Recall ('0' - healthy loans): 0.89

- Recall ('1' - high-risk loans): 0.75

## Summary

The improved logistic regression model (Model 2), after employing the RandomOverSampler for addressing class imbalance, outperformed the initial logistic regression model (Model 1).

\*\*Recommendation:\*\*

Considering the nature of the problem - predicting loan default risks - both precision and recall for high-risk loans ('1') are crucial. Model 2, which incorporated resampling techniques, demonstrated a substantial improvement in predicting high-risk loans compared to the initial model.

Model 2 is recommended due to its higher balanced accuracy, improved precision, and recall for high-risk loans. The increased ability to accurately identify high-risk loans is more crucial in the context of avoiding potential defaults, making it a more reliable model for this specific problem. Therefore, Model 2 appears to be more suitable for deployment in predicting loan defaults.

**Purpose of this analysis:**

The purpose of this analysis was to assess the performance of machine learning models and accuracy of a data model, specifically logistic regression, in predicting credit risk. The analysis aimed to utilize financial data to predict whether a loan is either healthy (0) or has a high risk of default (1). Justifying the recommended model for company to use.

**Describe the accuracy, precision, and recall scores of the machine learning model:**

The dataset included information on various financial attributes such as loan amount, interest rates, credit scores, amount of accounts, and other related variables. The prediction was focused on categorizing loans into either '0' for healthy loans or '1' for high-risk loans based on this financial data.

Model 1:

* Balanced Accuracy Score: 99.23%. Considering the sensitivity (recall / true positive rate and specificity (true negative rate) of the model the balanced prediction accuracy was 99.23%
* Precision Score:93% an average of 93% of predicted Positives were correct.
* Recall Score: 95% Model has 95% precision in measuring true positive values out of all the positive predictions made.
* F1-Score: 94%. Combining the precision and recall scores of model computing how many times a model made a correct prediction across the entire dataset.

Model 2:

* Balanced Accuracy Score: 99.42%. Considering the sensitivity (recall / true positive rate and specificity (true negative rate) of the model the balanced prediction accuracy was 99.42%
* Precision Score:93% an average of 93% of predicted Positives were correct.
* Recall Score: 99% Model has 99% precision in measuring true positive values out of all the positive predictions made.
* F1-Score: 96%. Combining the precision and recall scores of model computing how many times a model made a correct prediction across the entire dataset.

**Summarize the results from the machine learning model. Include your justification for recommending the model for use by the company. If you don’t recommend the model, justify your reasoning.**

Considering the nature of the problem - predicting loan default risks - both precision and recall for high-risk loans ('1') are crucial. Model 2, which incorporated resampling techniques, demonstrated a substantial improvement in predicting high-risk loans compared to the initial model.

Model 2 is recommended due to its higher balanced accuracy, improved precision, and recall for high-risk loans. The increased ability to accurately identify high-risk loans is more crucial in the context of avoiding potential defaults, making it a more reliable model for this specific problem. Therefore, Model 2 appears to be more suitable for deployment in predicting loan defaults.

In summary I would recommend using model 2 to predict the creditworthiness of borrowers. It has over 99% in accuracy if predicting the outcome of the repayment of initial loans. It can be easily molded into a business risk profile to ensure sufficient capital flow for the lenders to remain in business as well as make a profit.