# Module 12 Report Template

**## Overview of the Analysis**

In this section, describe the analysis you completed for the machine learning models used in this Challenge. This might include:

\* Explain the purpose of the analysis.

We will use Machine Learning to create a model that could predict the creditworthiness of borrowers who would default (one class) or not default (a second class) on their loan.

\* Explain what financial information the data was on and what you needed to predict.

A machine learning model will predict which loans are healthy (low risk) or high-risk based on the loan status data provided by the lending company.

\* Provide basic information about the variables you were trying to predict (e.g., `value counts`).

1. My known target class (loan status) is imbalanced. 75,036 out of

77,536 equals 97% non-default loans, while 3% (2,500/77,536 =.03) loans are on default. Most of the data belongs to the class label, not on default. We have more data on non-default than default, so we can better predict healthy loans (non-default).

2. The data was split on 58,152 records under train data, while 19,384 were on test.

\* Describe the stages of the machine learning process you went through as part of this analysis.

1. Check the balance of the target class

2. Split the data using train\_test\_split

3. Created Logistic Regression and AdaBoost models

\* Briefly touch on any methods you used (e.g., `Logistic Regression` or any resampling method).

Logistic Regression Model:

1. Fit a logistic regression model by using x\_train and y\_train to identify patterns

2. Generate two sets of predictions that allow us to check for overfitting

3. Evaluate the model for probability by creating a confusion matrix and a classification report

AdaBoost Model:

1. Fit an AdaBoost model by using x\_train and y\_train

2. Generate two sets of predictions that allows us to check for overfitting

3. Evaluate the model for probability by creating a confusion matrix and the classification report

**## Results**

Using bulleted lists, describe the balanced accuracy scores and the precision and recall scores of all machine learning models.

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

Description of Model 1 Accuracy, Precision, and Recall scores.

· The precision, recall, and accuracy scores of both train and test metrics are about the same.

· Recall score is between 90 and 91.

· Resulting in no sign of overfitting.

· False Positive numbers are a bit high for both metrics (277/1881 = 15%, 102/619 = 16%).

· False Negative numbers are also high (181/1881 = 9%, 56/619 = 9%)

\* **Machine Learning Model 2**: AdaBoost

\* Description of Model 2 Accuracy, Precision, and Recall scores.

· The precision, recall, and accuracy scores of both train and test metrics are almost the same.

· Recall score is between 99 and 100.

· Resulting in no sign of overfitting.

· Though there are 115 False Positive on test

· There are only 4 False Negative

**## Summary**

Summarize the results of the machine learning models, and include a recommendation on the model to use, if any.

The good news is that most people do not default on their loans. There are a lot of patterns in predicting a healthy loan or no loan default. Though both logistic regression and AdaBoost models had an accuracy of 99%, there is a big difference in the recall scores, 90 to 99, respectively. I recommend the AdaBoost model as it is better than logistic regression in predicting the probability of high-risk loans.

\* Which one seems to perform best? How do you know it performs best?

The AdaBoost performed the best because it reduced the False Negative tremendously. It will also predict loan statuses as high-risk at 99% better than logistic regression predicting high-risk at 91%. So, AdaBoost is better than logistic regression in predicting the probability of a target variable in classification problems.

\* Does performance depend on the problem we are trying to solve? (For example, is it more important to predict the `1`'s, or predict the `0`'s?)

Yes. The most important thing should be predicting the 1s high-risk loans rather than the 0s healthy loans. High-risk loans are more costly for a lending company since they might cause more loss of funds than the loss of healthy loans or borrowers.