**Module 12 Report Template**

**Overview of the Analysis**

* The purpose of this analysis was to utilize various supervised machine learning techniques to fit, train, and evaluate a model based on loan risk to predict whether future loans will be a healthy or risky loan. The financial information that the data was on were categories such as loan size, interest rate, borrower income, debt to income percentage, number of accounts, derogatory marks, total debt, and loan status that shows whether the loan is healthy or risky. Through this information, predictions on whether the borrower was a worthy borrower or a risky borrower. The variables that needed to be predicted were the categorical and numerical columns aside from the loan status column that were put in the feature variable. The loan status was put as the label variable. The value counts for the label variable were also looked at. The feature and label variables were transformed and split into training and testing data. The machine learning stages that the data went through were preparing the given data, choosing the model, training the models, making predictions, and evaluating the model through confusion matrix and classification report. For the model, logistic regression model was utilized and then the data was resampled through the random oversample model and fit into another logistic regression model.

**Results**

**Machine Learning Model 1:**

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

* With the original data in the first logistic regression model, the balanced accuracy score came out to be around 95%. The accuracy for the macro average was 92% and the weighted average accuracy was 99%. The precision for the healthy loans was 100% and the precision for the high-risk loans was 85%. The recall score for the health loans was 99% and the recall score for the high-risk loans was 91%.

**Machine Learning Model 2:**

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

* With the resampled data in the second logistic regression model, the balanced accuracy score came out to be around 99%. The accuracy for the macro average was 92% and the weighted average accuracy was 99%. The precision for the healthy loans was 100% and the precision for the high-risk loans was 85%. The recall score for the health loans was 99% and the recall score for the high-risk loans was 91%.

**Summary**

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

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

**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? )**

**If you do not recommend any of the models, please justify your reasoning.**

* I believe that the logistic model with the resampled data performs the best as it has the better balanced accuracy score. However, I would recommend the first logistic model with the original data that was not resampled because I believe that the data should not be resampled to achieve a higher accuracy score. It should be left alone so that it can get honest and authentic data results. I feel that performance does matter but not as much because predicting the high risk loans would be the most important task at hand. Predicting the high risk loans would support damage control and avoid clients who are predicted to be at high risk.