**Module 12 Report Template**

**Overview of the Analysis**

The aim of this analysis was to employ various supervised machine learning techniques to develop, train, and assess a model focused on loan risk, with the goal of predicting whether future loans will be classified as healthy or risky. The financial data encompassed variables such as loan size, interest rate, borrower income, debt-to-income ratio, number of accounts, derogatory marks, total debt, and loan status indicating whether the loan is healthy or risky. Utilizing this data, predictions were made to determine whether a borrower is deemed creditworthy or risky. The target variables included both categorical and numerical columns, excluding the loan status column, which served as the feature variable, while the loan status itself was designated as the label variable. The distribution of label variable counts was examined. Subsequently, the feature and label variables underwent transformation and were divided into training and testing datasets. The machine learning process involved data preparation, model selection, model training, prediction generation, and model evaluation using techniques such as confusion matrix and classification report. Initially, a logistic regression model was employed, followed by resampling of the data using the random oversampling method, and fitting into another logistic regression model.

**Results**

**Machine Learning Model 1:**

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

In the initial logistic regression model using the original data, the balanced accuracy score was approximately 95%. The macro average accuracy stood at 92%, while the weighted average accuracy reached 99%. Precision for healthy loans reached 100%, whereas for high-risk loans it was 85%. Furthermore, the recall score for healthy loans was 99%, and for high-risk loans, it 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.