**Module 12 Report – Nandini Tivakaran**

**Credit Risk Classification Analysis**

**Overview**

**Purpose of the Analysis**

The purpose of the analysis was to train and evaluate a model based on loan risk. We were given a dataset of historical lending activity from a peer-to-peer lending services company to build a model that can identify the creditworthiness of borrowers.

**About the Data**

The data included several predictors of creditworthiness, including: the size of the loan, interest rate, borrower’s income, the debt to income ratio, the borrower’s number of accounts, derogatory marks, and the borrower’s total debt. These were used to predict the status of a loan: whether it would be paid off or defaulted on.

**Stages of the Machine Learning Process and Methods Used**

First, I split the data into training and testing sets. I separated the data into labels and features, and checked the balance of the labels. I instantiated a logistic regression model and fit the model using the training data. Finally, I evaluated the model’s performance by calculating the accuracy score of the model, generating a confusion matrix, and printing the classification report.

These stages were repeated to predict a logistic regression model with resampled training data. I used the RandomOverSampler module from the imbalanced-learn library to resample the data and ensure that the labels had an equal number of data points. I fit the model with the resampled data and made predictions as in the previous steps. Finally, the performance of the second model was evaluated by the accuracy score, confusion matrix, and classification report.

I compared these three (accuracy score, confusion matrix, and classification report) measures of the model’s performance between both logistic regression models to determine which model would be better for banks to use to reduce their risk in lending money.

**Results**

**Confusion Matrix Analysis**

**ML Model 1: Logistic Regression Model**

* True negatives: 18,679 were healthy loans, and the model predicted correctly.
* False positives: 80 were healthy loans, but the model predicted they would be high-risk loans.
* False negatives: 67 were high-risk loans, but the model predicted they would be healthy loans.
* True positives: 558 were high-risk loans, and the model predicted correctly.

**ML Model 2: Resampled Logistic Regression Model**

* True negatives: 18,668 were healthy loans, and the model predicted correctly.
* False positives: 91 were healthy loans, but the model predicted they would be high-risk loans.
* False negatives: 2 were high-risk loans, but the model predicted they would be healthy loans.
* True positives: 623 were high-risk loans, and the model predicted correctly.

**Analysis of Accuracy, Precision, and Recall Scores**

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

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

* Accuracy
  + The accuracy score is 0.992, meaning that the model correctly predicted 99.2% of the instances in the actual dataset.
* Precision:
  + For the Prediction class, precision is 1.00, meaning that all the positive predictions were correct. There were no false positives.
  + For the Actual class, precision is 0.87, meaning that 87% of the negative predictions were correct. There were some false positives.
* Recall
  + For the Prediction class, recall is 1.00, meaning that all positive predictions were correct. There were no false negatives.
  + For the Actual class, recall is 0.89, meaning that 89% of actual negatives were correctly predicted as negative. This means that there were some false negatives in the Actual class.

**Machine Learning Model 2: Logistic Regression Model with Resampled Data**

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

* Accuracy
  + The model has an accuracy score of 0.995, meaning that the model correctly predicted 99.5% of the instances in the actual dataset.
* Precision
  + For the Resampled Prediction class, precision is 1.00 meaning that all positive predictions were correct, and there were no false positives.
  + For the Actual class, precision is 0.87, meaning that 87% of the negative predictions were correct, and there were some false positives.
* Recall Scores
  + For the Resampled Prediction class, recall is 1.00, meaning that all actual positive instances were predicted correctly. There were no false negatives.
  + For the Actual class, recall is 1.00, meaning that all actual negative instances in the dataset were correctly predicted as negative. There were no false positives.

**Summary**

The classification reports indicate that both Model 1 and Model 2 perform well with high precision and recall. Accuracy is 0.3% higher in Model 2 (Resampled Logistic Regression). Model 1 had more false predictions than Model 2 (Resampled Logistic Regression Model). However, it had fewer false positives than Model 2. Model 2 had much fewer false negatives and more false positives. From the company’s perspective, using Model 2 would be safer and more cautious as it is safer for the company to predict that a borrower is riskier than for the model to predict that a loan is healthy when it is actually high-risk. However, Model 2 would be unfavorable from the borrower’s perspective because it would make it more difficult to get approved for a loan if the model falsely identifies a borrower as high-risk.

Both models are highly accurate, and it is unlikely that there would be a significant difference in the practical use of one model or the other. Therefore, I would recommend either model.