**Credit Classification Machine Learning Model Analysis**

**Overview of the Analysis**

The purpose of this analysis is to train a model to predict loan as healthy or high risk based on a set of features. Our initial dataset contained a population of 77,536 loans collected from a peer-to-peer lending services activity. The features of the dataset included:

* loan\_size
* interest\_rate
* borrower\_income
* debt\_to\_income
* num\_of\_accounts
* derogatory\_marks
* total\_debt
* loan\_status

In building this model, I used a Logistic Regression model as this is a common method used in binary classification. Our target in this case was the feature called “Loan\_status”, which contained a 0 or 1 to indicate a healthy or high risk loan, respectively.

Here is an overview of the steps I took

1. Preprocessing: First, I did a bit of preprocessing by removing duplicated data. As previously mentioned, our initial dataset had 77k rows. Upon removing duplicates, the dataset was significantly reduced to 5,229 rows.
2. Check balance of value counts: I wanted to see the split of healthy vs. high risk loan counts in our de-duped dataset.
3. Train/Test Split: I split the data into training vs. testing data.
4. Create Logistic Regression Model: I imported the LogisticRegression model from sklearn and ran our training data through it.
5. Test Model: I then used the testing data and ran that through the model. This yielded an accuracy score of 89.5%.
6. Resampling Training Data: Based on the balance of the value counts in step 2, we are able to see that our training data was highly imbalanced. In order to rebalance, I imported the RandomOverSampler module to add datapoints to create a more even distribution of 0s and 1s.
7. Retesting Model with Resampled Data: Once the resampled data was run through the model, it yielded an accuracy score of 91.8%.

**Results**

1. Precision: Percentage of correct positive predictions relative to total positive predictions.

2. Recall: Percentage of correct positive predictions relative to total actual positives.

3. F1 Score: A weighted harmonic mean of precision and recall. The closer to 1, the better the model.

Machine Learning Model 1:

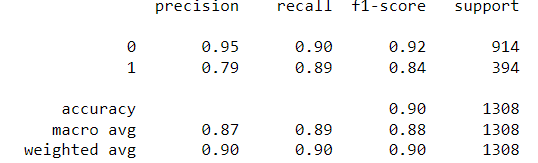
0 = Healthy

1 = High Risk

*Precision: 95% precision in correctly predicting healthy loans vs total healthy loans.*

*Recall: 90% recall in correctly predicting healthy loans vs actual healthy loans.*

*Accuracy: 90% overall accuracy*

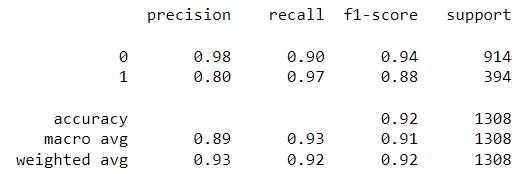


Machine Learning Model 2:

*Precision: 98% precision in correctly predicting healthy loans vs total healthy loans.*

*Recall: 90% recall in correctly predicting healthy loans vs actual healthy loans.*

*Accuracy: 92% overall accuracy*



**Summary**

As a takeaway, both variations of the model do appear to have a high level of accuracy. For a problem like this, it is probably ideal to overpredict high risk loan. Relatively speaking, those numbers are lower than the accuracy scores for predicting healthy loans.

From a data perspective, I do not feel that we had very good data to train this model. For one thing, our original dataset had a very high volume of duplications, which is very unhelpful when trying to train a model. Furthermore, once those dupes were removed, the dataset was extremely small. If you were trying to scale this for wider use, you would want to have a much higher volume of datapoints.