**Credit Risk Challenge (Module 20)**

**Activity Analysis Report**

**Summary Description**

For the ‘Credit Risk’ challenge (Module 20), I used various techniques to train and evaluate a supervised machine-learning model based on purported loan risk. Specifically, with the initial dataset provided, I constructed two separate logistical regression models with the ultimate intention of using these models to identify the creditworthiness of future prospective borrowers.

**Data**

The given dataset **lending\_data.csv** is comprised of historical lending activity from a peer-to-peer lending services company. The 77,536 total records included individual loan recipient information such as loan size, interest rate, borrower income, debt-to-income ratio, number of accounts held, number of derogatory marks, total debt, and, lastly, loan status.

**Goal**

The challenge goal is to develop a model using the historical data in order to identify and distinguish between credit worthy and high risk loan applicants. Finally, once developed, the model’s performance is assessed for potential use on future loan applicant data.

**Process**

First, the historical data is split into training and testing sets— technically speaking, the creation of a

‘labels set’ named ‘y’, and a features dataframe named ‘x’. Note, ‘y’ is a set comprised of only the ‘loan status’ field, and the ‘x’ dataframe contains all the remaining fields, referred to as ‘features’. It is the relationship between the ‘features’ and the ‘labels’ by which a model is built to ultimately determine ‘credit worthiness’ vs ‘credit risk’.

The initial balance of the labels variable y, determined by using the value\_counts function, is as follows:

Loan Status Count

------------------------------

0 (credit-worthy) 75036

1 (credit risk) 2500

After successfully splitting the initial data into training and testing sets (using the ‘train\_test\_split’ function), I fit a logistic regression model to it, and ran prediction analysis on the testing dataset. The saved predictions on the testing labels, utilizing the ‘testing feature data’ (i.e., x\_test), yielded results in the form of a list of the ‘prediction value’ as compared to the ‘actual value’ of each record.

From here, I was able to evaluate the model’s performance. Evaluation was determined with the following 3 assessments:

1. Accuracy Score calculation
2. Confusion Matrix construction
3. Classification Report output

The aforementioned process (from splitting data into training and testing sets to the final model assessment) was then completed, in entirety, using ‘resampled training data’. Specifically, the ‘RandomOverSampler’ model was instantiated, yielding the following perfectly balanced labels set:

Loan Status Count

------------------------------

0 (credit-worthy)  75036

1 (credit risk) 75036

Ultimately, this second model ‘Linear Regression with Resampled Data’ showed improvement, as demonstrated in all 3 assessments -- accuracy score, confusion matrix and the classification report.

**Results Summary**

The results are as follows:

* **Machine Learning Model 1:**

Linear Regression on historical lending activity dataset: ‘lending\_data.csv’

* Accuracy Score: 0.9443
* Classification Report (first 2 lines featured below):

precision recall f1-score support

-----------------------------------------------

Class Purple 1.00 1.00 1.00 18759

Class Yellow 0.87 0.89 0.88 625

*Note, ‘Class Purple’ represents the predicted ‘healthy loan’ population, whereas the ‘Class Yellow’ represents the ‘unhealthy’.*

* **Machine Learning Model 2:**

Linear Regression using ‘resampled training data’

* Accuracy score: 0.99597… which when rounded up is a perfect 1.0 (100 perent-o!)
* Classification Report (first 2 lines featured below):

precision recall f1-score support

--------------------------------------------------

Class Purple 1.00 1.00 1.00 18759

Class Yellow 0.87 1.00 0.93 625

*Again, ‘Class Purple’ represents the predicted ‘healthy loan’ population (i.e., that with a strong pulse), whereas the ‘Class Yellow’ represents the ‘unhealthy’ (perhaps jaundice due to low liver function… metaphorically speaking 😉).*

**Conclusion**

So, it proves to be that in the world of big, big, and ever bigger data, the smallest of differences count. Hence between the models, the 0.05 increase in accuracy score, and the improved recall and f1-scores, lead me to conclude that the second model, **Linear Regression with resampled data**, earns my recommendation.

And yet, this is not to overlook the fact that making a balanced dataset from a severely unbalanced one runs the risk of causing one’s model to overfit the minority (in this case those with low liver function 😉), leading to what is called ‘generalization error’. I submit that a population of only 2,500 from the original 77,536 (i.e., 3.22% of the original population) qualifies as ‘severely unbalanced’. Undoubtedly, that risk is present here.

In all, this exercise was extremely valuable as an educational instrument, but may not suffice for real world application. Effective real-world application may require initial training on a larger and more balanced dataset, as well as the provision of additional relevant feature variables. With such, prediction assessments have a greater chance of being in line with the true complexity of a loan applicants financial future and their overall loan-worthiness.