**Overview of the Analysis - Credit Risk Classification**

In this Challenge, various Supervised Machine Learning techniques were used to train and evaluate a model based on loan risk. The dataset used consisted of historical lending activity from a peer-to-peer lending services company, where the task was to build a model that can identify the creditworthiness of borrowers.

In supervised learning, a set of known answers called labels (y)are used to fit a model with a set of features (inputs/X) that correspond to the labels. These models are called supervised learners.

The purpose of this report is to identify the credit worthiness of the dataset and to determine if the model created is accurate enough to predict healthy loan or high risk loan status for applicants.

The data used to represent the y-variable, also referred to as the labels, was “loan status”

The features/x-variable was based on “loan size”, “interest rate”, “borrower income”, “debt to income”, “number of accounts”, “derogatory marks” & “total debt”.

The balance of the labels was found the using value\_counts() function where approximately 75% of the data set represented ‘healthy loan’ [0] and the remaining 25% represented high risk loan [1]

The stages of the machine learning process included:

**Preprocess** – Cleaning the data such as assigning the labels and features, as well as splitting into testing and training data

**Train** – Using training data to define classification by teaching the model to recognise patterns

**Validate** – Using the test data to validate the trained model.

**Predict** – Using the model to predict the outcome for unclassified data

To build the model the data was split into training and testing data once train\_test\_split was imported from sklearn.model\_selection. The variables used were X\_train, X\_test, y\_train, y\_test and the random\_state of 1 was assigned which will ensures that the same randomization is used each time before splitting.

The Logistic Regression Model was model used as this algorithm is able to predict which category the data points should be assigned too using binary classification.

Alongside Logistic Regression, the RandomOverSampler module from the imbalanced-learn library was imported to resample the data. Random oversampling duplicates examples from the minority class in the training dataset

**Results**

\* Machine Learning Model 1: Logistic Regression

The model has an overall accuracy of 99%, with a recall & f1-score of 100% for healthy loans, however the recall and f1-score predicting high risk is substantially lower at 89% & 88%. This would suggest this model would not be the best model to predict credit risk based on the High risk loan results.

\* Machine Learning Model 2: Random OverSampler

The logistic regression model, fit with oversampled data worked very well with an overall as the accuracy of 99%. In addition the recall and f1 score for both healthy loan & high-risk loan was calculated at 99%, showing this model to be a more reliable model to use for credit risk predictions.

**Summary**

I believe that, predicting both, healthy loan and high risk loan as just as important a s the other as they would both need to have an accurate state of prediction to manage the results and ensuring there’s not bias in either case.

I would recommend the RandomOver Sampler model as the results were more consistent with a higher accuracy, precision and recall that was closer to 100%