**Credit Risk Analysis**

**Overview**

Credit risk is an inherently unbalanced classification problem as good loans easily outnumber risky loans. Due to this risk, we are to employ the imbalanced-learn and scikit-learn libraries to build and evaluate models using resampling.

In this analysis, we will perform the following to evaluate our models:

* Oversample the data using RandomOverSampler and SMOTE algorithms
* Undersample the data using the ClusterCentroids algorithm
* Use a combinational approach of over and undersampling using the SMOTEENN algorithm
* Compare BalancedRandomForestClassifier and EasyEnsembleClassifier (machine learning models used to reduce risk. We will compare the two models’ ability to predict credit risk)

**Results**

* Naïve Random Oversampling: The balanced accuracy score is 65%. The high-risk precision has a low positivity at 1% and the recall is at 61%, with F1 score of 2%.

Table

Description automatically generated with medium confidence

* SMOTE Oversampling: The accuracy score for the SMOTE algorithm was 62%, with a high-risk precision of very low positivity at 1%, average recall is at 64%.

Table

Description automatically generated

* Cluster Centroids Resampling: The balanced accuracy score is at 51%, high risk precision of 1%, overall recall score of 43%

Table

Description automatically generated

* SMOTEENN Combination Sampling: The balanced accuracy score is at 62%, high risk precision of 1% and overall recall score of 54%

Table

Description automatically generated

* Balanced Random Forest Classifier: The balanced accuracy score for this algorithm is at 79%, high risk precision is at 4% and average recall is at 91%

Table

Description automatically generated

* Easy Ensemble Classifier: The balanced accuracy is at 93%, high risk prediction is 7%, with high risk recall at 91% and F1 score of 14%

Table

Description automatically generated

**Summary**

In the first deliverable, we calculated the accuracy score, a confusion matrix was generated, and an imbalanced classification report was generated for the three algorithms used.

In the second deliverable, we used the SMOTEENN algorithm to resample the data from the first deliverable and predict Credit risk. And in the third deliverable, we used the BalancedRandomForestClassifier and EasyEnsemble Classifier to resample the data and predict Credit Risk. For each algorithm used, we calculated the accuracy score, confusion matric and imbalanced generated the imbalanced classification report.

Of all the algorithms used to predict the credit risk for this analysis, the Easy Ensemble AdaBoost Classifier had the highest balanced accuracy score of 93%, low positivity high risk score of 7%, and a recall score of 91%. Therefore, I would recommend the Easy Ensemble Classifier algorithm due to its accuracy score and its ability to detect high risk credit.