**Credit Risk Analysis**

**Overview:**

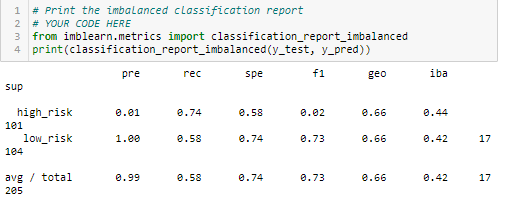
Credit risk is an inherently unbalanced classification problem; therefore, I will employ different techniques to train and evaluate models with use of “imbalanced-learn” and “scikit-learn” libraries to build and evaluate models using resampling.

Using the credit card credit dataset from LendingClub, a peer-to-peer lending services company, I will oversample the data using the “RandomOverSampler” and “SMOTE” algorithms, and undersample the data using the “ClusterCentroids” algorithm. Then, will use a combinatorial approach of over- and undersampling using the “SMOTEENN” algorithm. Next, will compare two new machine learning models that reduce bias, “BalancedRandomForestClassifier” and “EasyEnsembleClassifier” to predict credit risk. After predicting the credit risk, will evaluate the performance of those model and making a recommendation on whether the model should be used to predict risk.

**Results:**

**Random Oversampling**

* Balanced accuracy score: 0.66
* Precision: is the measure of how reliable a positive classification is. From our results, the precision for high risk credit applications is 0.01 and for low risk is 1.00. the avg total for 255 application is 0.99
* Recall avg total: 0.58. Recall is the ability of the classifier to find all the positive samples. 0.74 were high risk credit application and 0.58 were the low risk credit applications.

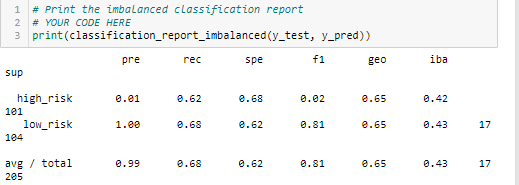


**SMOTE Oversampling**

Balanced accuracy score: 0.65

Precision avg total: 0.99. Precision is the measure of reliability of positive classification. The high-risk credit precision score is 0.01 and low risk precision is 1.00.

Recall avg total: 0.68. Recall is the ability of the classifier to find all the positive samples. The recall score for high risk is 0.62 and low risk recall is 0.68.

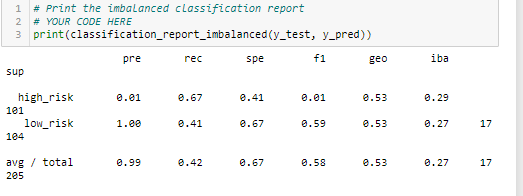


**ClusterCentroids Undersampling**

Balanced accuracy score: 0.54

Precision avg total: 0.99- the precision score for high risk credit is 0.01 and the precision score for low risk credit is 1.00

Recall avg total: 0.42. the recall score for high risk is 0.67 and for low risk is .41

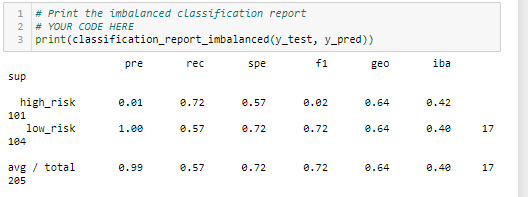


**SMOTEENN Under and Over Sampling**

Balanced accuracy score: 0.64

Precision avg total: 0.99. the precision score for high risk credit is 0.01 and low risk credit is 1.00

Recall avg total: 0.57. the recall score for high risk is 0.72 and low risk is 0.57.

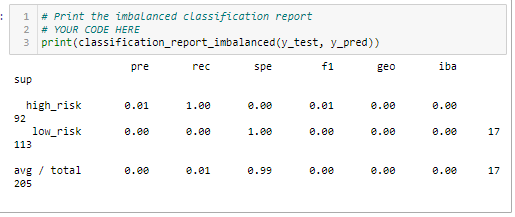


**Balanced Random Forest Classifier**

Balanced accuracy score: 0.5

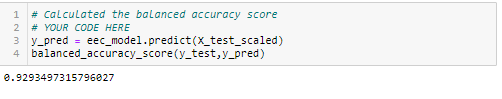
Precision avg total: 0.00. the high-risk precision score is 0.01 and low risk is 0.00.

Recall avg total: 0.01. the recall score for high risk is 1.00 and low risk is 0.00.



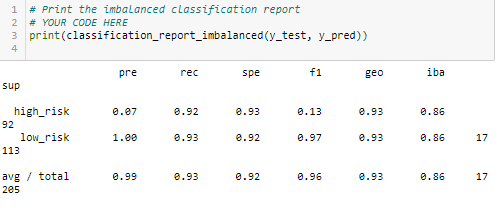
**Easy Ensemble AdaBoost Classifier**

Balanced accuracy score: 0.93



Precision avg total: 0.99. the precision score for high risk is 0.07 and low risk is 1.00.

Recall avg total: 0.93. the recall score for high risk is 0.93 and low risk is 0.93.



**Summary:**

Random Oversampling may not be the best model for identify the low and high risk credit applications because the model accuracy is 0.66, low and the precision and recall scores are not good enough to state that the model will be good to classify the risk for credit applications. The model may need more data, more cleaning or another model parameter to make predictions.

SMOTE Oversampling is not the perfect model for classify the high and low risk credit applications because of low accuracy test, which is low 0.65. the precision and recall scores are not good alone to make the model identify the high and low risk credit applications.

ClusterCentroids Undersampling, SMOTEENN Under and Over Sampling and Balanced Random Forest Classifier also are not a good fit model because of the low accuracy score, which is 0.54, 0.64 and 0.5. the precision average score in two models are 0.99 but the Balanced Random Forest classifier has 0 precision score. The recall scores are low in all models

The Easy Ensemble AdaBoost Classifier may be the best model to identify the low and high credit risk because of the high accuracy score, which is 0.93. the precision and recall average scores 0.99 and 0.93 which is playing an important role to identify the low and high risk credit applications.

I would recommend the Easy Ensemble Adaboost Classifier model to predict the low and high risk for this project.

In summary, this model may not be the best one for preventing fraudulent loan applications because the model's accuracy, 0.552, is low, and the precision and recall are not good enough to state that the model will be good at classifying fraudulent loan applications. Modeling is an iterative process: you may need more data, more cleaning, another model parameter, or a different model. It's also important to have a goal that's been agreed upon, so that you know when the model is good enough.