In this challenge, we are evaluating several machine learning models to access credit risk using data from LendingClub.

Since Credit risk is an inherently unbalanced classification problem, as the number of good loans easily outnumber the number of risky loans, we will use different techniques to train and evaluate models with unbalanced classes of low risk (68470) & high risk (347). Here we use imbalanced-learn and scikit-learn libraries to build and evaluate models using resampling to evaluate the performance of these models and make a recommendation on whether these models should be used to predict credit risk or not.

First, we will oversample the data using the RandomOverSampler and SMOTE algorithms.

For RandomOverSampler, LogisticRegression model shows only 50% of accuracy and negligible precision and recall for high risk:

* Balance accuracy score 0.5
* Confusion matrix:

[ 0, 87],

[ 0, 17118]

* Classification Report:

pre rec spe f1 geo iba sup

high\_risk 0.00 0.00 1.00 0.00 0.00 0.00 87

low\_risk 0.99 1.00 0.00 1.00 0.00 0.00 17118

avg / total 0.99 0.99 0.01 0.99 0.00 0.00 17205

whereas for SMOTE Oversampling, the LogisticRegression model shows model shows only 63% of accuracy, negligible precision (0.01) and fairly good recall (0.62) for high risk:

* Balance accuracy score increased to 0.6303
* Confusion matrix:

[ 54, 33],

[ 6163, 10955]

* Classification Report:

pre rec spe f1 geo iba sup

high\_risk 0.01 0.62 0.64 0.02 0.63 0.40 87

low\_risk 1.00 0.64 0.62 0.78 0.63 0.40 17118

avg / total 0.99 0.64 0.62 0.78 0.63 0.40 17205

At the same time, Undersample the data using the cluster centroids algorithm shows only 50% of accuracy. Further metrics shown below:

* Balance accuracy score 0.509
* Confusion matrix:

[ 55, 32],

[10489, 6629]

* Classification Report:

pre rec spe f1 geo iba sup

high\_risk 0.01 0.63 0.39 0.01 0.49 0.25 87

low\_risk 1.00 0.39 0.63 0.56 0.49 0.24 17118

avg / total 0.99 0.39 0.63 0.55 0.49 0.24 17205

With a combination approach using SMOTEENN algorithm shows slightly similar accuracy score as SMOTE model with highest recall score (0.71) for high risk:

* Balance accuracy score 0.632
* Confusion matrix:

[ 62, 25],

[ 7666, 9452]

* Classification Report:

pre rec spe f1 geo iba sup

high\_risk 0.01 0.71 0.55 0.02 0.63 0.40 87

low\_risk 1.00 0.55 0.71 0.71 0.63 0.39 17118

avg / total 0.99 0.55 0.71 0.71 0.63 0.39 17205

Based on above analysis we can observe that the SMOTE Oversampling and SMOTEENN models predict nearly similar accuracy scores and average precision for both low risk and high-risk classifications. Both of these models also have good F1 scores but since SMOOTEENN model has better predictions with high risk classification I would not recommend using any of these models as best to prevent fraudulent loan applications because the models’ accuracy doesn’t exceed 63% and F1 score beyond 0.71.

### Additionally, further analysis using Ensemble model shows better metrics with accuracy score of 93% and average precision 99% and F1 score 0.97:

* Balance accuracy score 0.931
* Confusion matrix:

[ 93, 8],

[ 983, 16121]

* Classification Report:

### pre rec spe f1 geo iba sup

### high\_risk 0.09 0.92 0.94 0.16 0.93 0.87 101

### low\_risk 1.00 0.94 0.92 0.97 0.93 0.87 17104

### avg / total 0.99 0.94 0.92 0.97 0.93 0.87 17205

### Hence, I would recommend Ensemble model would best suit the model predictions for this dataset.