**READ ME**

Imbalanced-learn and scikit-learn libraries are used in this project to build and evaluate models using resampling.

**Random Oversampling**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Classification Report: Random Oversampling | | | | | | | |
|  |
|  | pre | rec | spe | f1 | geo | iba | sup |  |
| high\_risk | 0.01 | 0.74 | 0.58 | 0.02 | 0.66 | 0.44 | 101 |  |
| low\_risk | 1 | 0.58 | 0.74 | 0.73 | 0.66 | 0.42 | 17104 |  |
| avg/total | 0.99 | 0.58 | 0.74 | 0.73 | 0.66 | 0.42 | 17205 |  |

The first technique we used for. From this method, we made sure

We start with resampling is the Random Oversampling method. We understand that the sizes of both high and low credit risk samples are equal.

We can see from the below classification table that the precision and recall scores from this model are fairly high.

99% of predictions were accurate, and 61% of positive samples were correctly identified.

The balanced accuracy score for this model is 65%. It is a little bit higher than that of the recall score, however still significant.

**SMOTE Oversampling**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Classification Report: SMOTE Oversampling | | | | | | | |
|  |
|  | pre | rec | spe | f1 | geo | iba | sup |  |
| high\_risk | 0.01 | 0.62 | 0.68 | 0.02 | 0.65 | 0.42 | 101 |  |
| low\_risk | 1 | 0.68 | 0.62 | 0.81 | 0.65 | 0.43 | 17104 |  |
| avgtotal | 0.99 | 0.68 | 0.62 | 0.81 | 0.65 | 0.43 | 17205 |  |

I used smote oversampling. From the classification table below, the precision and recall scores are also relatively high. With 99% precision score, 69% recall score, and 81% F1 score, these metrics are slightly improved over the metrics of random oversampling. Though the percentage of precisions were about the same for both oversampling techniques, SMOTE oversampling shows a higher accuracy rate.

The balanced accuracy score for this model is 66%. This percentage is slightly higher that the balanced accuracy of the naive random oversampling model. Given that the F1 score of this model is also higher that that of the other oversampling model, we can say that this is a more reliable model at the moment. We will next explore undersampling and combination sampling to test whether these result is better performance compared to the oversampling models.

**Undersampling**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Classification Report: SMOTE Oversampling | | | | | | | |
|  |
|  | pre | rec | spe | f1 | geo | iba | sup |  |
| high\_risk | 0.01 | 0.62 | 0.68 | 0.02 | 0.65 | 0.42 | 101 |  |
| low\_risk | 1 | 0.68 | 0.62 | 0.81 | 0.65 | 0.43 | 17104 |  |
| avg/total | 0.99 | 0.68 | 0.62 | 0.81 | 0.65 | 0.43 | 17205 |  |

I used undersampling algorithms to determine if this results in the best performance compared to the oversampling algorithms

The classification table shows high precision (99%), though a low recall score of 41%. This means that a lesser number of positive samples were correctly identified. The F1 score is also a low and is not good enough to state that the model will be good at classifying loan status.

The balanced accuracy score of this model is also fairly low at 55%, meaning this is not a reliable model for our classifiers.

**Combination Sampling**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Classification Report: Combination Sampling | | | | | | | |
|  |
|  | pre | rec | spe | f1 | geo | iba | sup |  |
| high\_risk | 0.01 | 0.7 | 0.57 | 0.02 | 0.63 | 0.41 | 101 |  |
| low\_risk | 1 | 0.57 | 0.7 | 0.73 | 0.63 | 0.4 | 17104 |  |
| avg/total | 0.99 | 0.57 | 0.7 | 0.72 | 0.63 | 0.4 | 17205 |  |

 I used combination sampling to determine if the algorithm results in the best performance compared to the other sampling algorithms above.

We can see that the precision rate is still 99%, which is good. The recall score is higher at 58%, compared to the undersampling algorithm, however still low compared to the oversampling models. The F1 score is 73% is still higher than undersampling, but not oversampling.

Looking at the balanced accuracy score, this reads at 65%, which makes sense given the recal rate is also higher.

When we look at the classification reports, we can easily say that Smote Oversampling algorithm is the best model to use for resampling this data.