Credit Risk Analysis

Using data from LendingClub, a peer-to-peer lending services company, several models were built and evaluated to assess credit risk. The models that were used include naïve random oversampling, SMOTE oversampling, cluster centroids undersampling, and combination SMOTEENN sampling. The performance of each model are as follows:

Naïve Random Oversampling

The accuracy score for this model is 63.6%. Precision is 1% for high risk classification and 100% for low risk classification. The recall score is 63% and 64% for high risk and low risk classifications, respectively.

SMOTE Oversampling

The accuracy score for this model is 64.8%. Precision is 1% for high risk classification and 100% for low risk classification. The recall score is 62% and 68% for high risk and low risk classifications, respectively.

Cluster Centroids Undersampling

The accuracy score for this model is 51.2%. Precision is 1% for high risk classification and 100% for low risk classification. The recall score is 63% and 39% for high risk and low risk classifications, respectively.

Combination SMOTEENN Sampling

The accuracy score for this model is 60.4%. Precision is 1% for high risk classification and 100% for low risk classification. The recall score is 67% and 54% for high risk and low risk classifications, respectively.

Recommendation

In terms of accuracy scores, all models did not perform well with the highest balanced accuracy score being 64.8% for the SMOTE oversampling model and the lowest score being 51.2% for the cluster centroids undersampling model. Nonetheless, in terms of predicting low credit risk, all models performed exceptionally well with all models having 100% precision scores, indicating that applicants that were assessed with low risk are more than likely to have been classified correctly. However, since the models are meant to assess credit risk, it is important that high risk applicants are identified and unfortunately, all models scored 1% in precision for high risk.

Since all models did not have exceptional accuracy scores and were not precise in predicting high credit risk, it is recommended that the data be further analyzed to validate the models. The data for high risk applicants should be studied to determine if the data is diverse enough for a model to learn from. A model that is recommended to be focused on for further analysis would be the combination SMOTEENN sampling model as it score the highest (67%) in recall for high risk classification, indicating that the amount of who are high risk were correctly identified.