Module 17 Challenge

Sheila Brear

Overview

The purpose of this analysis is to evaluate and compare two “machine learning” (data analysis) models that can be used to predict credit risk in dataset obtained from LendingClub.

Results

1. Compare balanced accuracy, precision and recall:
    - RandomOverSampler: An excellent score of 0.9, this tool accurately predicted a high credit risk of 131 people, and inaccurately scored 18 people who were high risk as low risk.
    - SMOTE: Accureately predicted 133 high risk people, and inaccurately predicted 16 people who were actually high risk as low risk. This was not as good a tool as RandomOverSampler for predicting risk.
    - SMOTEENN: I was not able to obtain this data ☹
    - ClusterCentroids: According to my data, accurately predicted 12 high risk individuals, and did not misclassify anyone who was high risk as low risk.
    - BalancedRandomForestClassifer: accurately predicted the high risk of 58 people, but inaccurately predicted a low risk of 30 people, who turned out to be high risk (false negatives). This was not a successful tool.
    - EasyEnsembleClassifier: Provided an excellent score of over 0.9, and accurately predicted the high risk of 79 people, and predicted a low risk of 8 people who turned out to be high risk (false negatives). This was a better tool than the BalancedRandomForestClasifier.

Summary

Of the tools used to predict risk, according to my data, the ClusterCentroids data did not miss any individuals who turned out to be high risk. It did misclassify 13 people who were actually low risk, as high credit risk, however, this in calculating credit risk, it is less risky for a company that is loaning money to over-assign risk status than under-assign.

I recommend using ClusterCentroids for assessing risk, based on the data presented.