**Testing different sampling techniques to improve precision/recall**

We are trying to predict whether borrows will default on their loans in the next period. We have many predictors, such as the loan amount, the interest rate, monthly payment amount, annual income, etc. (there are about 85 variables when removing the outcome, loan status).

With no sampling the accuracy of the model(how correct it is at identifying outcomes) is 0.995, the average precision (the measure of how reliable a positive classification is) of the model is 0.99, though 1.00 for predicting low risk, and only 0.01 for predicting high risk. Recall (the ability of the classifier to find all the positive samples, or sensitivity) is only 0.70 for high risk, worse for low risk (0.57). The F1 score (a weighted average of the true positive rate (recall) and precision, where the best score is 1.0 and the worst is 0.0) is very small for high risk, which is ultimately what we are trying to predict (who is most likely to default). We should conclude that the model doesn’t fit the data particularly well. One reason it might not fit the data well is that there are very few high risk outcomes in the data (defaulting is a rare event). One method to try to improve the fit of the model is to change the underlying data, by under, or over sampling the minority class (high risk), the majority class (low risk), or both.

**Over sampling**

In naïve random oversampling, instances of the minority class are randomly selected and added to the training set until the majority and minority classes are balanced.

The accuracy is less (0.661) than for no sampling, the precision is somewhat better (0.67) but the f1 score for high risk is only 0.02 (not very good).

For SMOTE, or synthetic minority oversampling technique, the accuracy is somewhat improved over naïve random oversampling (0.630) but the recall and f1 scores are not improved.

**Under sampling**

In random under sampling, randomly selected instances from the majority class are removed until the size of the majority class is reduced, typically to that of the minority class.

When under sampling, the accuracy is the same (0.630) and recall (0.61 for high risk) and f1 score (0.02 for high risk) are similar to over sampling.

Another technique is to combine under and over sampling (over sample the minority class while under sampling the majority class). SMOTEENN combines the SMOTE and Edited Nearest Neighbors (ENN) algorithms.

With SMOTEENN, the accuracy is somewhat less (0.590) and the precision is still poor (0.01 for high risk) but the recall is somewhat better (0.70 for high risk). The f1 score is still poor (0.02).

Conclusion:

No form of sampling (over, under, combined) significantly improves the ability to predict high risk borrowers.

**## No Sampling**

Accuracy: 0.995

Confusion matrix:

[ 61, 26]

[7294, 9824]

Imbalanced classification report:

pre rec spe f1 geo iba sup

0 0.01 0.70 0.57 0.02 0.63 0.41 87

1 1.00 0.57 0.70 0.73 0.63 0.40 17118

avg / total 0.99 0.57 0.70 0.72 0.63 0.40 17205

**## Oversampling**

### ### Naive Random Oversampling

Accuracy: 0.661

Confusion matrix:

[ 57, 30]

[ 5694, 11424]

Imbalanced classification report:

pre rec spe f1 geo iba sup

0 0.01 0.66 0.67 0.02 0.66 0.44 87

1 1.00 0.67 0.66 0.80 0.66 0.44 17118

avg / total 0.99 0.67 0.66 0.80 0.66 0.44 17205

### SMOTE Oversampling

Accuracy: 0.630

Confusion matrix:

[ 54, 33]

[ 6163, 10955]

Imbalanced classification report:

pre rec spe f1 geo iba sup

0 0.01 0.62 0.64 0.02 0.63 0.40 87

1 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

# ## Undersampling

Accuracy: 0.630

Confusion matrix:

[ 53, 34]

[7336, 9782]

Imbalanced classification report:

pre rec spe f1 geo iba sup

0 0.01 0.61 0.57 0.01 0.59 0.35 87

1 1.00 0.57 0.61 0.73 0.59 0.35 17118

avg / total 0.99 0.57 0.61 0.72 0.59 0.35 17205

# ## Combination (Over and Under) Sampling

Accuracy: 0.590

Confusion matrix:

[ 61, 26]

[7294, 9824]

Imbalanced classification report:

pre rec spe f1 geo iba sup

0 0.01 0.70 0.57 0.02 0.63 0.41 87

1 1.00 0.57 0.70 0.73 0.63 0.40 17118

avg / total 0.99 0.57 0.70 0.72 0.63 0.40 17205