Curse of Imbalanced data

The term *imbalanced* refer to the disparity encountered in the *dependent (response) variable*. Therefore, an imbalanced classification problem is one in which the dependent variable has an imbalanced proportion of classes.

For example, we may have a 2-class (binary) classification problem with 100 observations. A total of 90 observations are labeled with Class-1 and the remaining 10 instances are labeled with Class-2.
This is an imbalanced dataset and the ratio of Class-1 to Class-2 instances is 90:10.

With imbalanced data sets, an algorithm doesn't get the necessary information about the minority class to make an accurate prediction. In this diabetes data set, we have only 11% readmissions under 30 days and about 89% readmissions were after 30 days or were no readmissions. This binary classification made data highly imbalanced.

I used four data balancing techniques using "ROSE" package to balance the data:

- Under sampling
- Oversampling
- Balanced sampling
- Synthetic Data balancing

ROSE package provides functions to deal with binary classification problems in the presence of imbalanced classes. Artificial balanced samples are generated per a smoothed bootstrap approach and allow for aiding both the phases of estimation and accuracy evaluation of a binary classifier in the presence of a rare class

1. Undersampling

This method works with majority class and decreases the number of observations from majority class to make the data set balanced. Apparently, removing observations may cause the training data to lose important information pertaining to majority class. With under-sampling AUC under ROC curve increased from 60%(balanced sample) to 62%.

2. Oversampling

This method works with minority class and repeats the observations from minority class to balance the data. A benefit of using this method is that it leads to no information loss. The disadvantage of using this method is that, since oversampling simply adds replicated observations in the original data set, it ends up adding multiple observations of several types, thus leading to overfitting. With oversampling AUC under ROC curve increased from 60%(balanced sample) to 63%.

3. Both under and over sampling-With both technique method AUC under ROC curve increased from 60%(balanced sample) to 62%.

4. Synthetic Data balancing

SMOTE algorithm creates artificial data based on feature space (rather than data space) similarities from minority samples. To generate artificial data, it uses bootstrapping and k-nearest neighbors.

Rose package	No of observations	AUC	Color
Synthetic Balancing using ROSE	9848	73%	Red
Both	9847	62%	Light red
Undersampling	2214	62%	Blue
Oversampling	17398	63%	Green
Unbalanced	19612	60%	Black

Many performance metrics assume the default decision threshold, this is inappropriate for imbalanced data. Often, the default threshold will simply classify everything as majority-class, since that gives the highest overall accuracy. If you want to see the performance of a learner on imbalanced data, you need to use the AUC, which gives performance across the whole range of decision thresholds.

There are a few implementations of the SMOTE algorithm, for example:

- In R, ROSE and DMwR package provides an implementation of SMOTE.
- In Weka, we can use the SMOTE supervised filter.
- In Python, "Unbalanced Dataset" module provides a number of implementations of SMOTE.

To deal with imbalanced data one can also try penalized models. Penalized classification such as penalized-SVM and penalized-LDA imposes an additional cost on the model for making classification mistakes on the minority class during training. Weka has a Cost-Sensitive classifier that can wrap any classifier and apply a customs penalty matrix for miss-classification.

To Learn more about SMOTE

- See the original paper published in 2002 titled SMOTE: Synthetic Minority Over-Sampling Techniquehttps://www.jair.org/media/953/live-953-2037-jair.pdf
- Haibo He, Edwardo A. Garcia, "Learning from Imbalanced Data," IEEE Transactions on Knowledge and Data Engineering, pp. 1263-1284, September 2009