Handling Imbalance in the Data

The sample design data has 5820 records, out of which, 1169 belong to positive class (sales conversion) and remaining 4651 are negative class (no conversion). The imbalance ratio is:

$$\text{Imbalance Ratio} = \frac{\text{Number of majority class records}}{\text{Number of minority class records}} = \frac{4651}{1169} = 3.97$$

Imbalance ratio of more than 3 can result in lower accuracy in classification. From all the models that we developed so far; we can notice that the sensitivity was much lower than specificity. The AUC was more than 85% for all the models. To overcome the problem posed by imbalance in the data, the following sampling strategies can be used:

Oversampling (upsampling) Minority class: In this strategy, we duplicate the minority class to reduce the imbalance.

Undersampling (downsampling) Majority Class: In this strategy, we remove majority class records using random sampling to improve the imbalance ratio.

Synthetic Minority Oversampling Technique (SMOTE): In SMOTE, synthetic records are created using k nearest neighbors (KNN) technique to improve imbalance.

We will demonstrate oversampling (upsampling) minority class; the new data set after oversampling has 6990 records in which 2338 belong to positive class and the remaining 4652 belong to negative class. The imbalance ratio now is 1.98.

TN-Table 8 shows the classification table of the logistic regression model with oversampling. From the table, we can observe that the sensitivity has increased to 66% from 47.5% (**TN-Table 2**). The ROC and AUC are shown in **TN-Figure 14**. AUC has improved slightly to 87.8 from 87.3 (**TN-Figure 6**).

Copy of Pos