Classification Accuracy Paradox

Recently, I worked on a diabetic readmission patient data. The dataset was taken from UCI Machine Learning Website. It is comprised of diabetic clinical care patient data collected over a period of ten years (1998 – 2008) from hospitals across the United States. The objective is to identify the key factors associated with readmission and to predict the probability of readmission.

I created a classification model and achieved an accuracy of 90%. My first reaction was Fantastic Model!!!

I dive a little deeper and discovered that 90% of the data belongs to one class. Damm!

This is a typical case of imbalanced data and the frustrating result it can cause. Most classification algorithms calculate accuracy based on the percentage of observations correctly classified. However, with imbalanced data, the results are highly deceiving since minority classes hold minimum effect on overall accuracy. The most frequently used metrics are Accuracy & Error Rate.

Accuracy: (TP + TN)/(TP+TN+FP+FN) –(i)

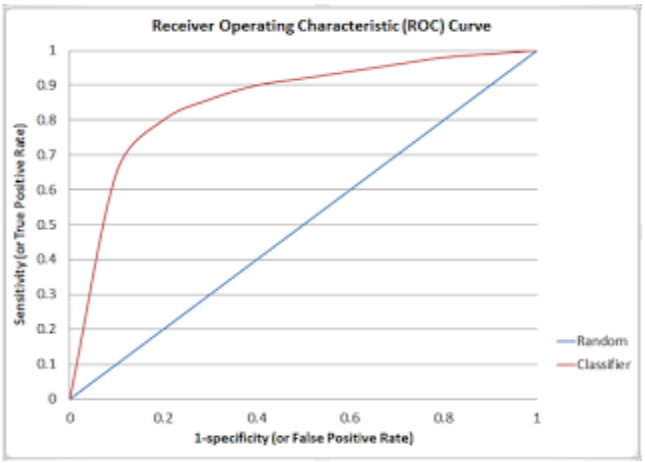
Where TP= True Positive, TN=True Negative, FP=False positive, FN=False Negative
Error Rate = 1 -Accuracy

Using equation (i), I achieved an accuracy of ~ 90% but it is not a right indicator as my objective is to classify the true positive correctly. The true positive rate in the metric below is only 6%.

		Prediction	
		0(Negative)	1(Positive)
Actual	0(Negative)	17,364(TN)	34(FP)
	1(Positive)	2077(FN)	137(TP)
Accuracy	89.2%		
Specificity	99%		

The above metric is based on subset of original UCI machine learning data

These metrics (Accuracy and Specificity) may provide deceiving results and are highly sensitive to changes in data. Further, various metrics can be derived from the confusion matrix. I used a ROC (Receiver Operating Characteristics) curve to measure the accuracy of a classification prediction. I examined ROC Curve by plotting TP rate (Sensitivity) and FP rate (Specificity) and AUC under ROC curve was only 60%, (an indication that my initial model was not good enough). In this case, two ratios we should look at are Recall(Sensitivity) and Precision.



Recall/ Sensitivity/True positive rate: It is a measure of actual observations which are labeled (predicted) correctly i.e. how many observations of the **positive class are labeled correctly**. It is also known as 'Sensitivity'. Recall = TP / (TP + FN) **= Σ True** positive/Σ Condition Positive

Precision or Positive predictive value (PPV) = TP/(TP+FP)= ΣTrue positive/Σ Test outcome positive
True negative rate(TNR), Specificity (SPC) = Σ True negative/Σ Condition negative

		Prediction			
		0(Negative)	1(Positive)		
Actual	0(Negative)	17,364(TN)	34(FP) Type I error	Specificity (TN/TN+FP)	100%
	1(Positive)	2077(FN) Type II error	137(TP)	Recall or sensitivity (TP/TP+FN)	6.6%
			Precision TP/(TP+FP)		
			80.1%		

The Precision of a patient classifier is the fraction of the patient in the test set it labeled as positive that really are positive whereas Recall is the percentage of all Patients readmission in the test set that it correctly labeled as a positive. In other words, Precision is the probability that a retrieved observation is relevant, and Recall is the probability that a relevant observation is retrieved. There is often a tradeoff between having high precision and high recall.

Having a single-number evaluation metric speeds up our ability to decide when we are selecting among a large number of classifiers. It gives a clear preference ranking among all of them, and therefore a clear direction for progress. We may combine, Precision and Recall, to a single number. For example, we could take the average of precision and recall, to end up with a single number. Or we can calculate the "geometric mean" between Precision and Recall, and is calculated as 2/((1/Precision) +(1/Recall)) also known as the F1 score.

If true negative is not much valuable to the problem or negative examples are abundant. Then, PR-curve is typically more appropriate. For example, if the class is highly imbalanced and positive samples are very rare, then we should use PR curve. For further reading,

please see http://pages.cs.wisc.edu/~jdavis/davisgoadrichcamera2.pdf