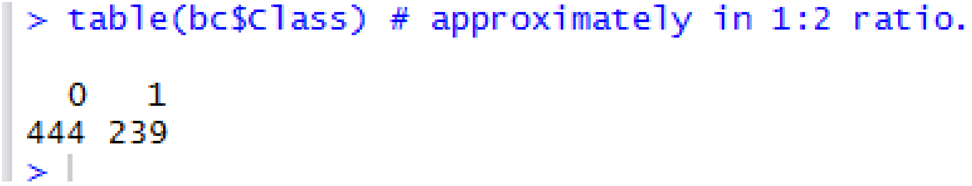
31 March 2018
20:36

How to deal with Class Imbalance?

- Before building the logistic regressor, you need to randomly split the data into training and test samples.

- Since the response variable is a binary categorical variable, you need to make sure the training data has approximately equal proportion of classes.



```
> table(bc$Class) # approximately in 1:2 ratio.

  0   1 
444 239 
> 
```

- Clearly there is a class imbalance. So, before building the logit model, you need to build the samples such that both the 1's and 0's are in approximately equal proportions.

This concern is normally handled with a couple of techniques called:

1. **Down Sampling**
2. **Up Sampling**

- In Down sampling, the majority class is randomly down sampled to be of the same size as the smaller class. That means, when creating the training dataset, the rows with the benign Class will be picked fewer times during the random sampling.

- Similarly, in UpSampling, rows from the minority class, that is, malignant is repeatedly sampled over and over till it reaches the same size as the majority class (benign).