HANDLING IMBALANCED CLASSES

- If you have IMBALANCED CLASS distribution in your raw dataset, then your validation set should have the same distribution as you would expect in the testing set.
- Your training set should have EQUAL DISTRIBUTIONS of each of the classes, i.e. 50-50.
- If you do not have a 50-50 distribution, you should replicate the less common class.
- It is always the case that oversampling the less common class in the training set produces better results in a model.
- Sklearn has a classweights parameter which you can tweak so that while the sklearn prepares the training set it samples the less numbered class with a higher probability.