<https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18>

**1. Change the performance metric**

* Confusion Matrix: a table showing correct predictions and types of incorrect predictions.
* Precision: the number of true positives divided by all positive predictions. Precision is also called Positive Predictive Value. It is a measure of a classifier’s exactness. Low precision indicates a high number of false positives.
* Recall: the number of true positives divided by the number of positive values in the test data. Recall is also called Sensitivity or the True Positive Rate. It is a measure of a classifier’s completeness. Low recall indicates a high number of false negatives.
* F1: Score: the weighted average of precision and recall.

**2. Change the algorithm**

* Decision trees frequently perform well on imbalanced data. They work by learning a hierarchy of if/else questions and this can force both classes to be addressed.

**3. Resampling Techniques — Oversample minority class**

* Oversampling can be defined as adding more copies of the minority class. Randomly replicate samples from the minority class.
* Oversampling can be a good choice when you don’t have a ton of data to work with.
* Always split into test and train sets BEFORE trying oversampling techniques!

**4. Resampling techniques — Undersample majority class**

* Undersampling can be defined as removing some observations of the majority class. Randomly remove samples from the majority class.
* Undersampling can be a good choice when you have a ton of data -think millions of rows. But a drawback is that we are removing information that may be valuable. This could lead to underfitting and poor generalization to the test set.

**5. Generate synthetic samples**

* A technique similar to upsampling is to create synthetic samples.
* For example, [imblearn’s](https://imbalanced-learn.readthedocs.io/en/stable/index.html?source=post_page---------------------------) SMOTE or Synthetic Minority Oversampling Technique. SMOTE uses a nearest neighbors algorithm to generate new and synthetic data we can use for training our model.
* Again, it’s important to generate the new samples only in the training set to ensure our model generalizes well to unseen data.
* <http://rikunert.com/SMOTE_explained>