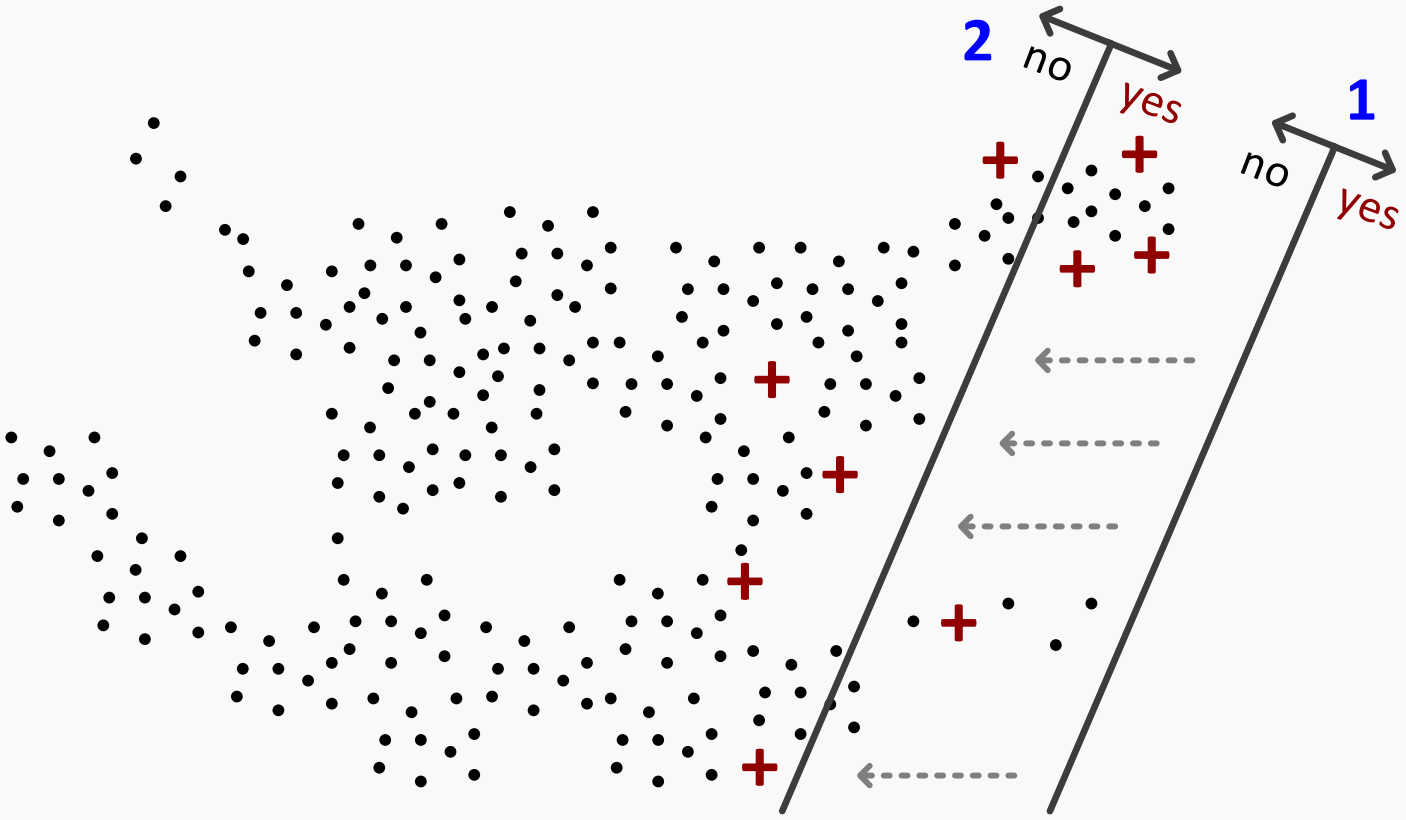
imbalanced data

The following illustration represents an imbalanced data set and a classification model applied:



The **Classification Model 1** represents a likely model the initial data will produce. It is noted that the predictions **Classifier 1** will make is **"no"** or **0** for every example in the dataset. This appears to be highly flawed, however the resulting model is **99% accurate** considering that only **1%** of the examples exhibit a **"yes"** or **1** response.

Regardless of the accuracy, the model is degenerative or trivial in nature; the model is essentially meaningless because it cannot interpret data to predict anything of value for future observations.

The alternate **Classification Model 2** will obviously have a much higher **Misclassification Rate** and predict a large number of **False Positives (+)**. However, the ability for **Classifier 2** to capture the existence of a positive result is much more valuable than the loss of accuracy in predicting negatives. The scarcity of the positive examples in the dataset prove the value of the latter.

The way to achieve **Classification Model 2** from the resulting dataset is to weight the value of **positive (+)** results appropriately. Weighting the algorithm is achieved through several methods of which **Regularization** is the most widely utilized:

$$\frac{1}{n}\sum_{i=1}^{n}\ell\left(y_i f(x_i)\right) + \text{Regulatization}(f)$$

The above **objective function** will treat the **positives (+)** equally to that of the **negatives (-)** by segregating the two types of labels as follows:

$$\frac{1}{n}\left(\boldsymbol{C} \sum_{\substack{i \text{ positives} \\ i \text{ where } y_i = 1}}^{n} \ell\left(y_i f(x_i)\right) + \sum_{\substack{k \text{ negatives} \\ k \text{ where } y_k = 1}}^{n} \ell\left(y_k f(x_k)\right) \right) + \text{Regulatization}(f)$$

The above notation dictates each **positive (+)** example is weighted as $\boldsymbol{C}$ × **a negative (-)** example.

The Classifier will now favor the **Classification Model 2** as opposed to **Classification Model 1**.

In summary, instead of relying on accuracy as an absolute evaluation metric, adjust **the imbalance parameter** $\boldsymbol{C}$ to obrain an ideal balance between **True Positives (+)** and **False Positives** $= \frac{\mathbf{TP}}{\mathbf{FP}}$.

Additionally to using **the imbalance parameter** $\boldsymbol{C}$ to adjust the model interpretation, the parameter C can equally be used to evaluate the performance of an entire algorithm itself **(discussed later)**.