**HANDLING SKEWED DATASET:**

(<https://towardsdatascience.com/what-to-do-when-your-classification-dataset-is-imbalanced-6af031b12a36> )

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

* Achieve higher accuracy in the minority class, not just the overall higher accuracy on the whole dataset.
* Over sampling the minority class vs. Under sampling the majority class

**Under sampling technique:**

<https://www.youtube.com/watch?v=YMPMZmlH5Bo>

* Take minority class as it is.
* Randomly select same number of points in the majority class.
* Makes both balanced

**Disadvantage:**

* Should only be done when you have a huge dataset that has millions of records (Use over sampling if you have less data)
* Will lose some information in under sampling.

**Imblean :**

* Library has all the techniques of under sampling and up sampling.

**Over Sampling:**

<https://www.youtube.com/watch?v=OJedgzdipC0>

* Majority class will be taken as same.
* Try to create new points in minority class.

**Advantages:**

* We will not lose information by decreasing the amount of data.

**Compare both methods under sampling and oversampling and show which one has better results.**

**Journal of what we have been doing**

**STEP 1: Understanding the dataset**

Currently we have 2 label columns: Noshow, cancel late

**NoShow: 9,106 samples**

**Show: 86,116 samples**

Cancelled late: 15,221 samples (patients who cancelled late did not show up)

**Sampling from show people (NoShowFLG = 0), we should take people who cancelled late**

**STEP 2: Cleaning Data**

**NA**

NoShow: **9,106 samples** (without any na value: **8,113 samples**)

Show: **86,116 samples** (without any na value: **80,343 samples**)

---> We do not impute NA value. We just drop them

**Data Noise**

CheckintoCheckoutNBR and AppttoCheckoutNBR: have negative values that need to be dropped

1006 no-show samples and 7538 show samples have negative values in CheckintoCheckoutNBR

1,143 no-show samples and 8709 show samples have negative values in AppttoCheckoutNBR

The number of appointment in the past is smaller than the number of late arrivals in the past: okay, we keep them

Medical history: mostly 0 because the patient either does not have it or not disclosing the information yet: okay, we still keep them and let the model decide whether it is helpful in predicting the no-show of a patient.

**Data Outlier**

After cleaning the data, we have:

NoShow: 7945 **samples**

Show: 71515 **samples**

**STEP 3: Sampling Data**

**Training:**

Show: 5,500

Noshow: 5,500

**Real data: skewed. Every 10 patients, 1 will show, 9 will not show**

**Validation:**

Show: 1,500

Noshow: 9,000

**Testing:**

Show: 1,000

Noshow: 9,000

**STEP 3: Normalizing Data**

**STEP 4: Choose algorithms**

**STEP 5: Train Data**