Smote and cross validation

31 May 2023 20:54

When do we say data is Balanced?

In diabetes data, if we have equal number of observations related to diabetic and non diabetic, then we say data is Balanced.

Eg: Diabetic - 150
Non Diabetic - 150.

Imbalanced data: → If data has unequal number of observations related to diabetic or non diabetic

Diabetic → 800
Non Diabetic → 200

what happens if data is Imbalanced?

Diabetic → 800 → Majority class

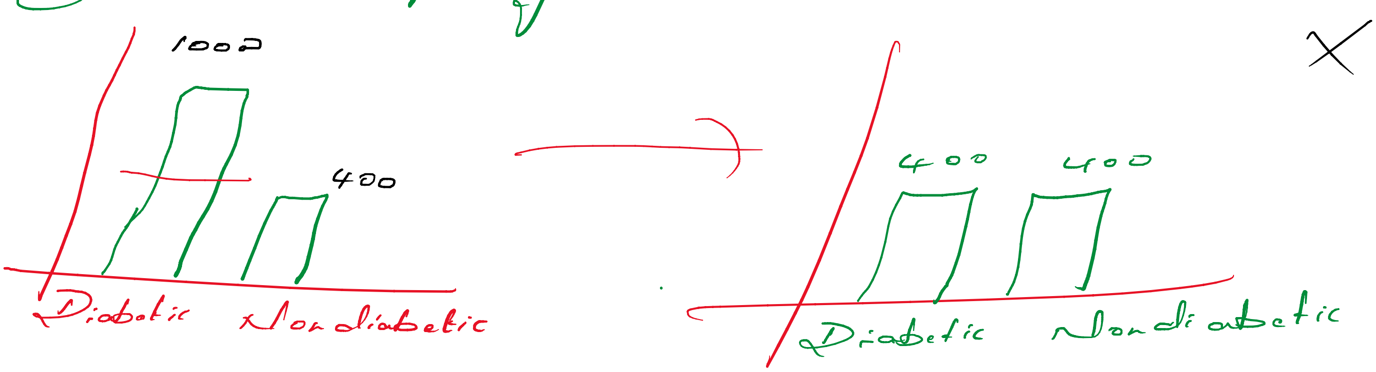
Non Diabetic → 200 → Minority class.

* when data is Imbalanced the model learns more about the majority class and learn less about minority class since it has very less information.

* Because of this issue model makes wrong predictions.

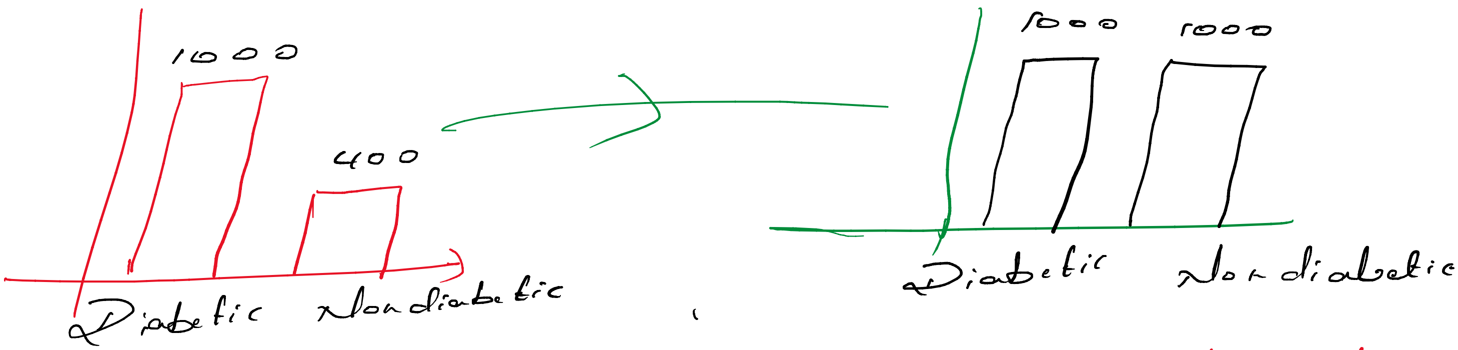
How to Balance data?

① Undersampling



undersampling is not a good option because data can't be lost.

② Oversampling



* Oversampling means increasing number of observations in minority class. by duplicating the records.
* Oversampling fails because it increases the duplicates.

Smote

It is the very popular method used to balance data.
SMOTE :- Synthetic minority oversampling technique.

* Smote is used to create synthetic/artificial data from the original data.
* It uses KNN and interpolation to create new data.

Note

Balancing should always be done on training data.

Cross Validation

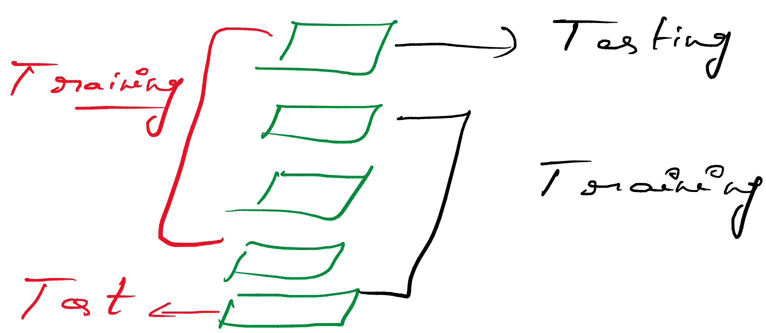
It is one of the resampling technique which divides data for training and testing

K-Fold Cross Validation

* It divides entire data into k parts of equal size.

$K \rightarrow$ No of Folds.
$K = 5$ or 10

$K = 5$



* The first set is selected for Testing and remaining n-1 sets used for training

* The first set is selected for Testing and remaining k-1 sets used for training.
* It will calculate the error made by the model.
* The above process repeats multiple times.

$k = 5 \longrightarrow 5$ Error.

Imp
Model is good only if it gives similar scores for all $\forall k$.

Note
using Cross validation we are validating model using different possible datasets using k-Fold.