Handling of Imbalanced data

DS-VI-B
Made By- Karan Choudhary 18CSU103

What is Imbalanced dataset

- **Imbalanced dataset** means instances of one of the two or more **classes** is higher than the other, in another way, the number of observations is not the same for all the **classes** in the **dataset**
- **Eg- Class B Class A**

  100 900

  Class A and B are not equally distributed

  Class A has large amount of data/observations for one class (referred to as the majority class), and much fewer observations for other class B.

Real world examples of Imbalanced data

- Credit Card fraud Detection.
- Anomaly detection
- Medical Diagnosis(brain tumor, diabetes ,heart disease)
- Market segmentation
- Emotion recognition
- Customer Churn modeling(banking)

Most of the algorithm work well when the number of instances of each class are roughly equal.But problem starts when we have large number of instances of one class(majority class).

To handle imbalanced data we will using certain techniques

How to balance Imbalanced classes ?

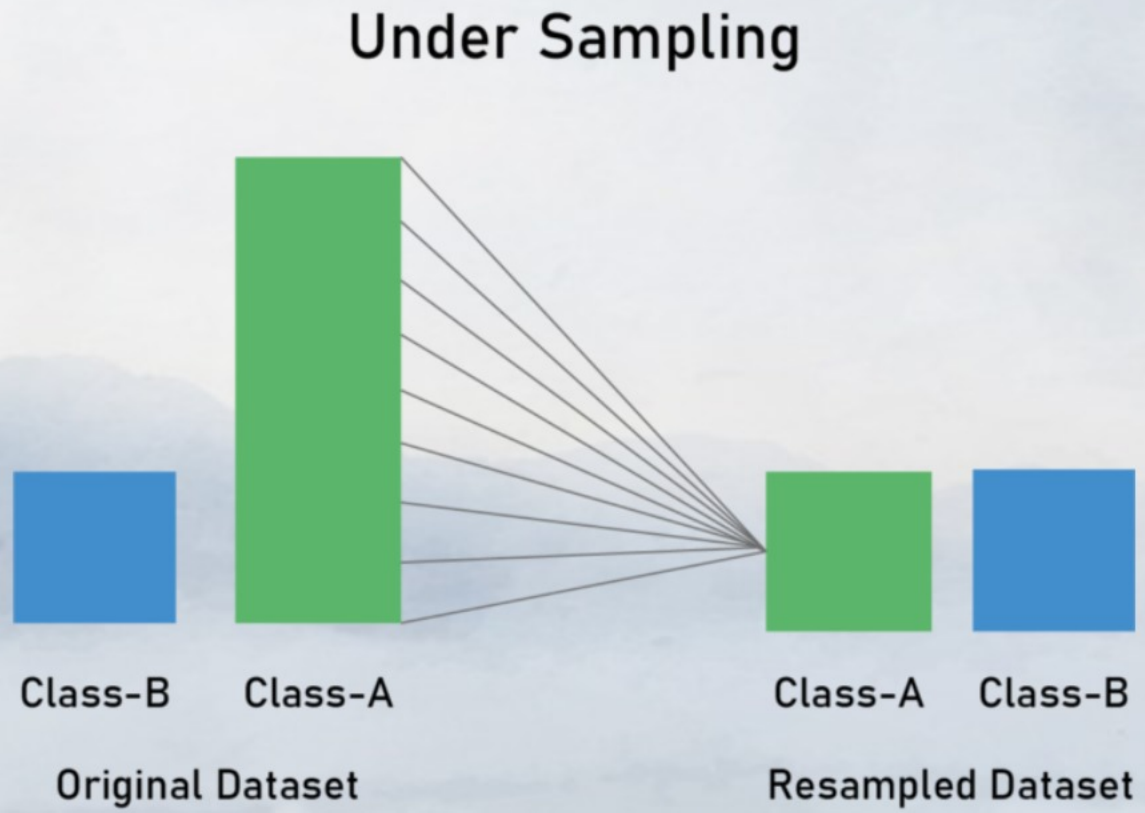
- Under Sampling
- Over Sampling
- SMOTE
- Class Weights

What are the challenges of imbalanced dataset in machine learning?

1. **What metric to choose** – It is not advisable to choose metric like accuracy with imbalanced dataset. Hence, one should choose the evaluation metric wisely we have precision,recall,F1score etc .
2. **Bias** – Classifiers are more sensitive to detect the majority class leading to biased classification output.
3. **Difficulty in getting more data** – Applications like fraud detection, faulty machine detection etc offers a unique challenge that frauds or faults occur rarely and hence less data for that particular class.

1. Under Sampling

- Method to remove instances of majority class so it has less effect on the machine learning algorithm.

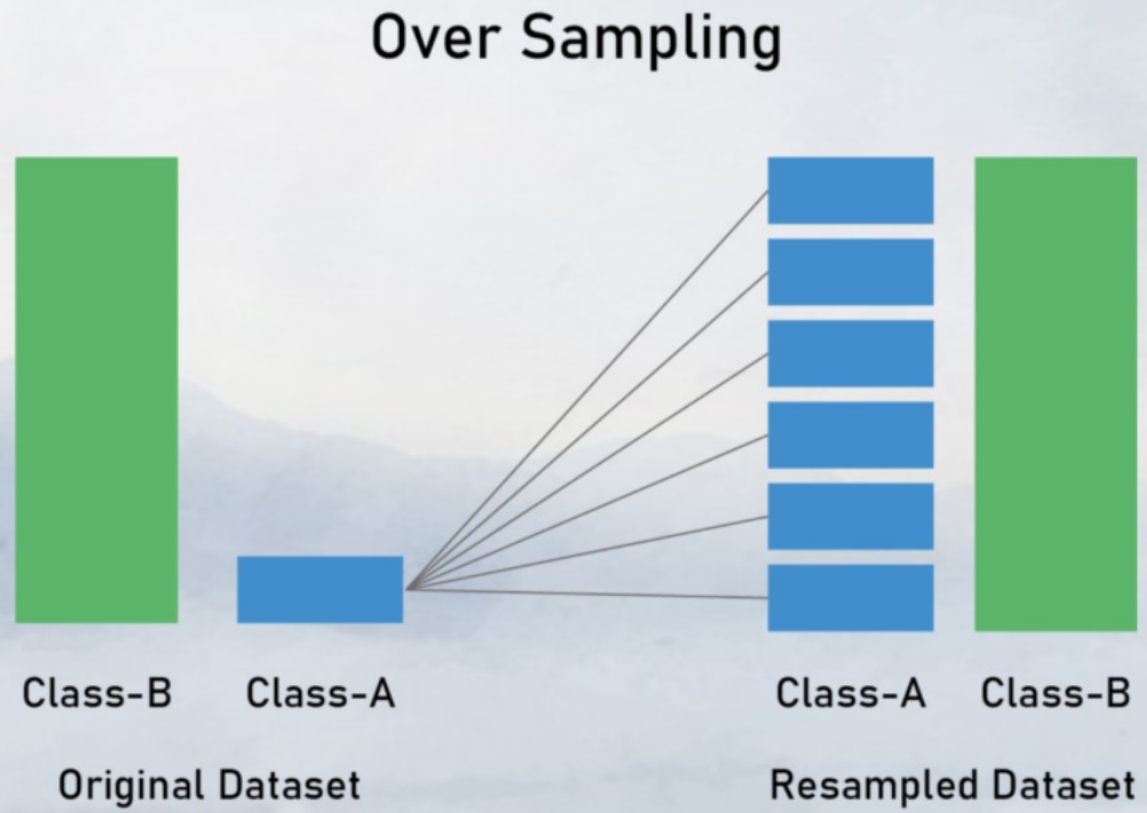


Disadvantages of Under Sampling

- It can discard potentially useful information which could be important for building rule classifiers.
- The sample chosen by random under-sampling may be a biased sample. And it will not be an accurate representation of the population. Thereby, resulting in inaccurate results with the actual test data set.

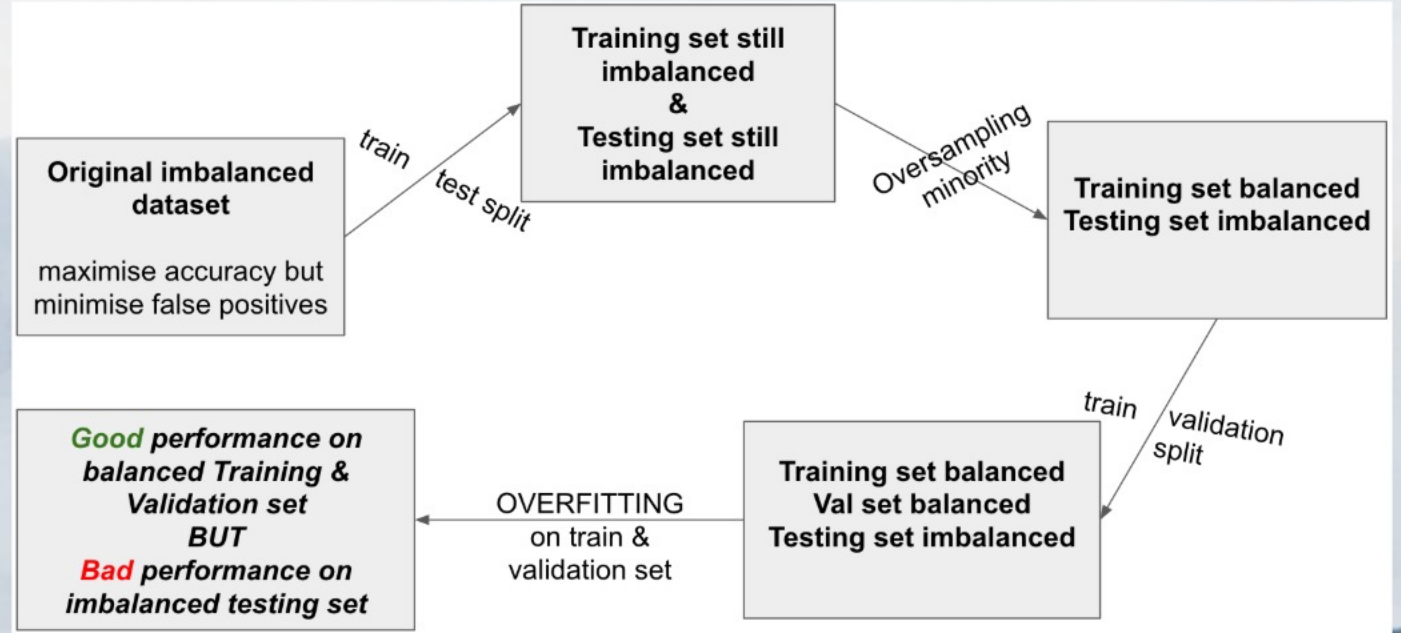
2.Over Sampling

- Sampling method duplicating the minority classes could lead the classifier to make the class equal in terms of instances.



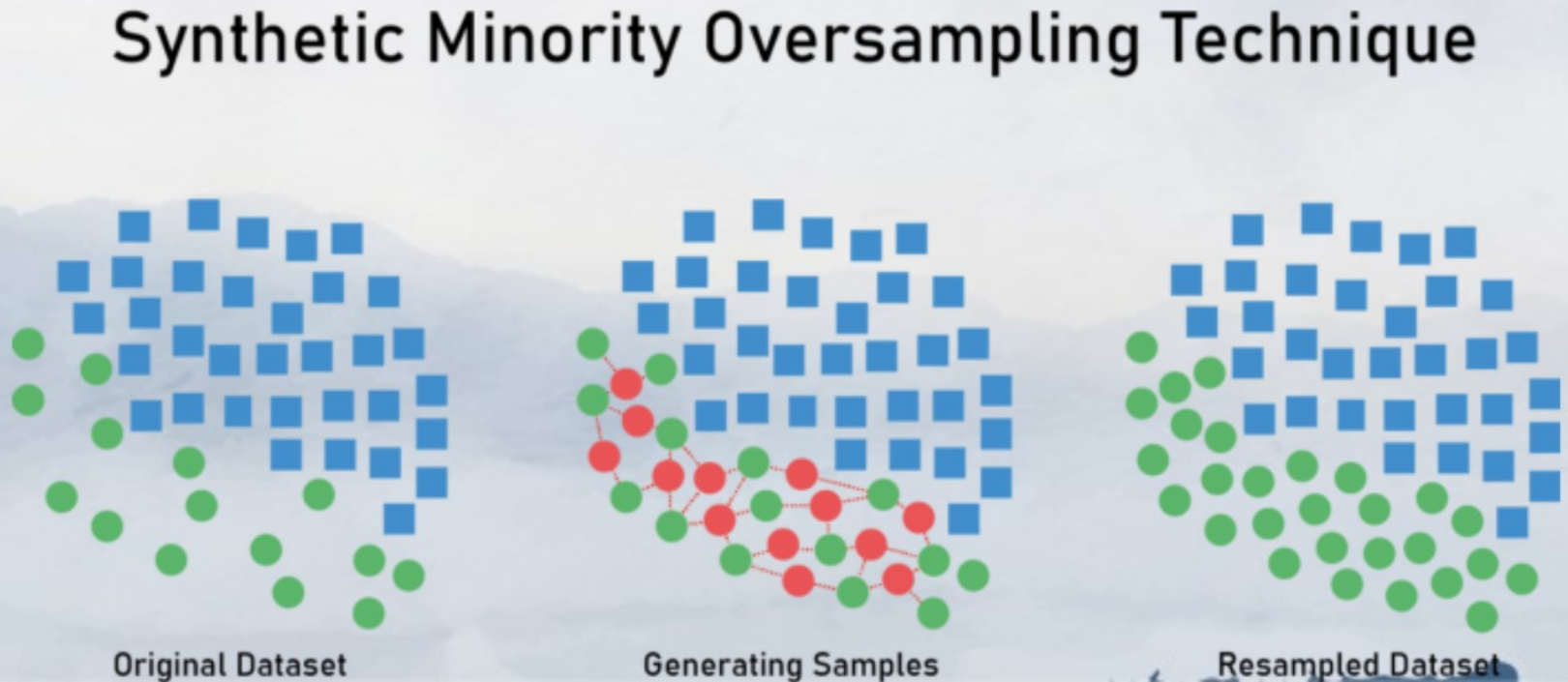
Disadvantage-

Usually, we begin with splitting entire dataset into training and testing set. Training set is further split into training and validation set. If validation split is done after oversampling has been done on the original bigger training set, it leads to overfitting and bad performance on the test set.



3.SMOTE

- SMOTE refers to Synthetic Minority Oversampling Technique.
- SMOTE works by randomly picking a point from the minority class and computing the k-nearest neighbors for this point. The **synthetic points are added** between the chosen point and its neighbors.



4.Class Weight

In this method we will increase the weights of minority class so that our model learn from the minority class having less number of instances.

Eg: class A class B

900 100

Class weights earlier -> 1:1(majority class : minority class)

After assigning weights - > 1:9(majority class : minority class)

Conclusion-Comparing between technique

- Over sampling performs good than under Sampling as a lot of meaning full information is lost during random Up Sampling.
- Smote performs better than Over Sampling due to making new points along with the minority classes.
- Smote perform better than class weights as weight increase of minority class leads to have better understanding for model but the training is not that impacted on dataset for credit card fraud detection it varies from data to data.

Thank you