CROSS VALIDATION METHODS

BY:

M. Ashish Reddy

DSWP -> Batch-5

What is Cross-validation?

- Cross-validation is a **resampling procedure used to evaluate machine learning models on a limited data sample**.
- It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods.

Types of Cross-validation Methods:

Holdout Method

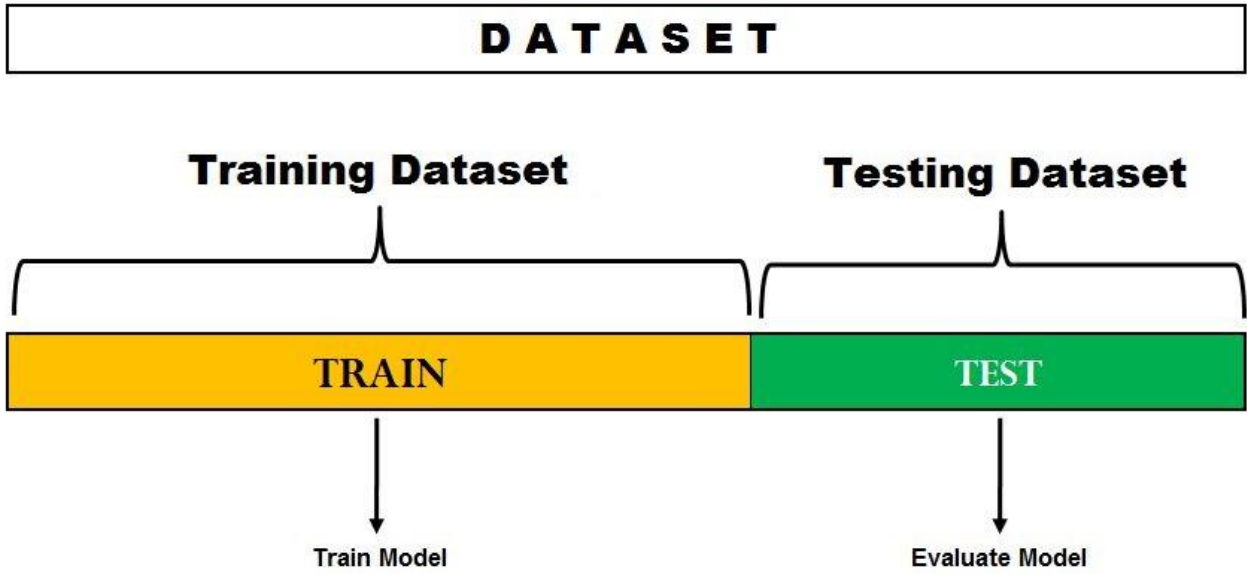
K-Fold Cross Validation

Stratified K-Fold Cross Validation

Leave-P-Out Cross Validation

Holdout Method

- Holdout Method is the **simplest sort of method to evaluate a classifier**. In this method, the data set (a collection of data items or examples) is separated into two sets, called the Training set and Test set. A classifier performs function of assigning data items in a given collection to a target category or class.



- **Example** –

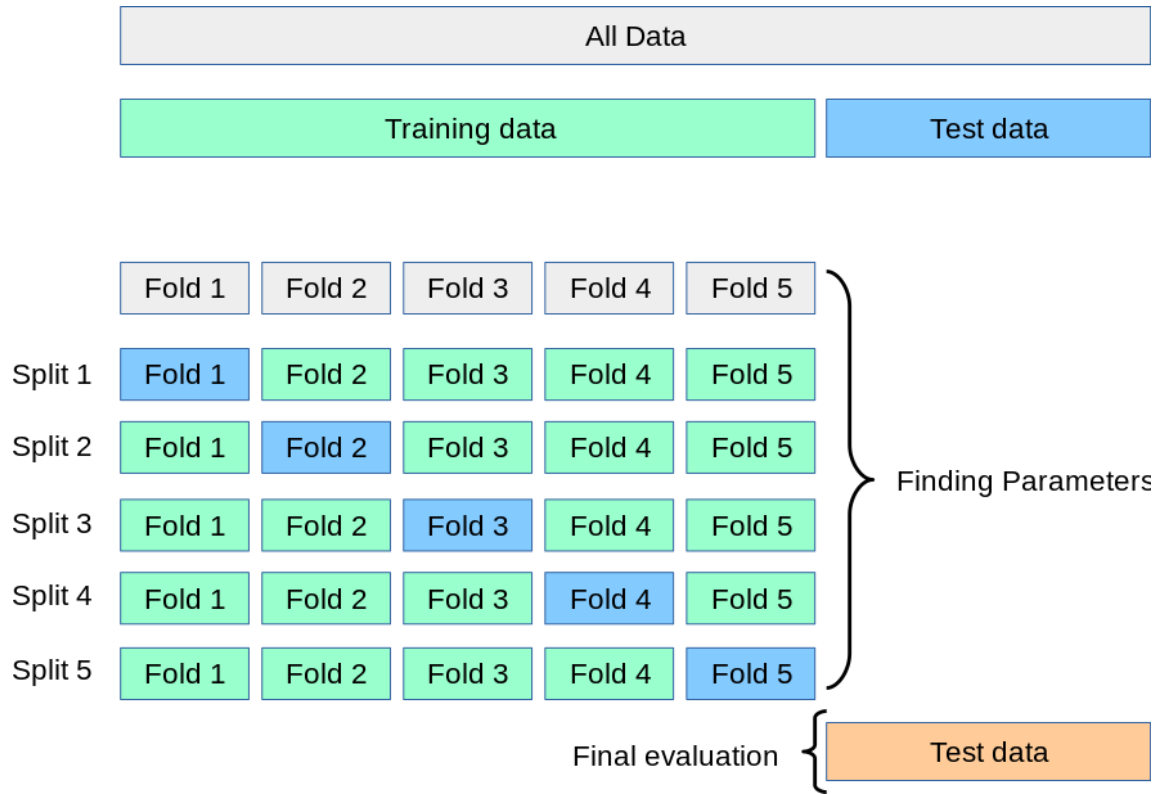
- If there are 20 data items present, 12 are placed in training set and remaining 8 are placed in test set.
- After partitioning data set into two sets, training set is used to build a model/classifier.
- After construction of classifier, we use data items in test set, to test accuracy, error rate and error estimate of model/classifier.
- However, it is vital to remember two statements with regard to holdout method. These are :
- If maximum possible data items are placed in training set for construction of model/classifier, classifier's error rates and estimates would be very low and accuracy would be high. This is sign of a good classifier/model.

When to use Holdout Method ?

- The hold-out method is good to use when you have a very large dataset, you’re on a time crunch, or you are starting to build an initial model in your data science project

K-Fold Cross Validation

- In **K Fold cross validation**, the data is divided into k subsets. Now the holdout method is repeated k times, such that ***each time, one of the k subsets is used as the test set/ validation set and the other k-1 subsets are put together to form a training set***.

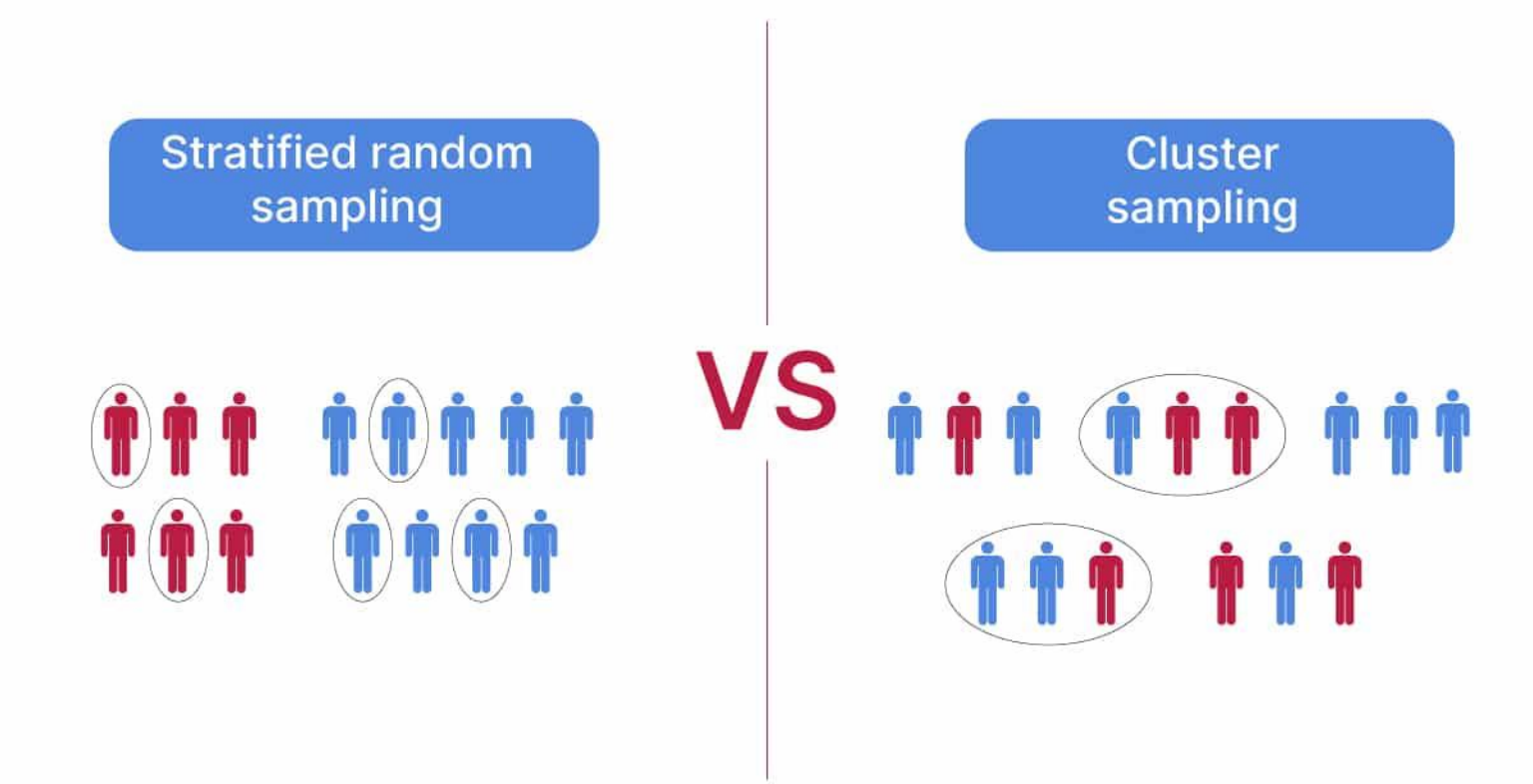


Advantages of K-Fold validation

- Computation time is reduced as we repeated the process only 10 times when the value of k is 10.
- Reduced bias.
- Every data points get to be tested exactly once and is used in training k-1 times.
- The variance of the resulting estimate is reduced as k increases.

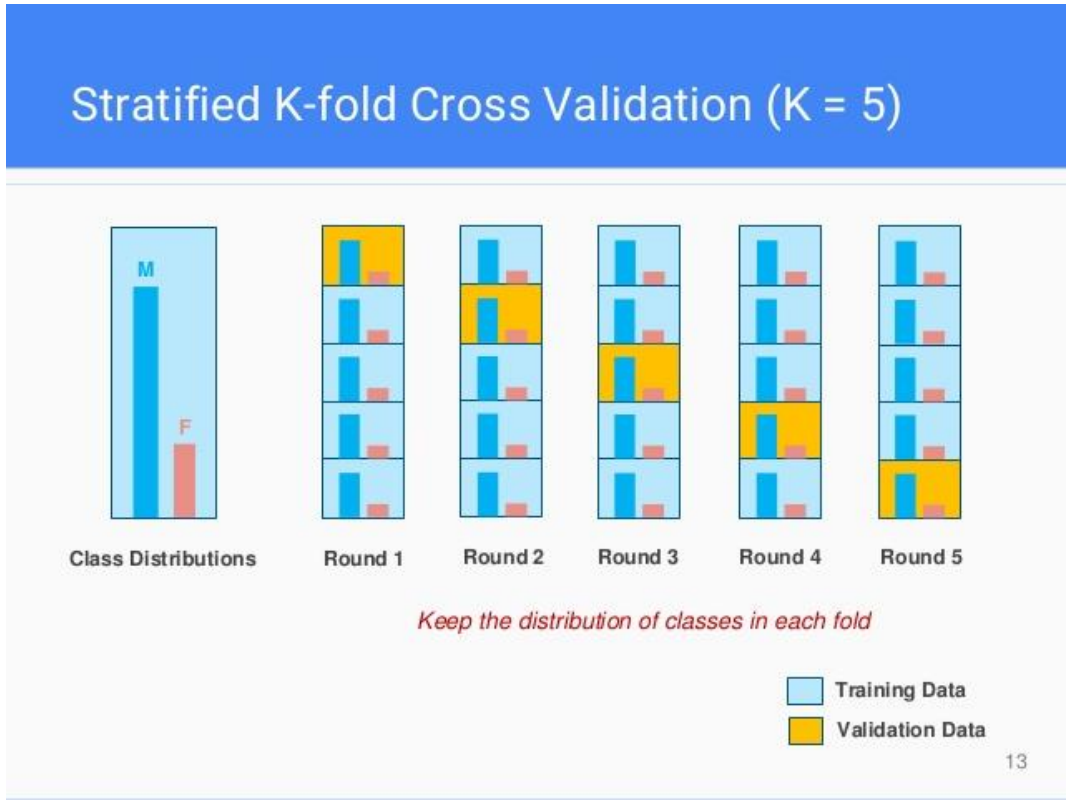
Stratified K-Fold Cross Validation

- To understand this concept we must first know what is the difference between Stratified Sampling and Cluster Sampling(Random Sampling)



What is Stratified K-Fold Cross Validation?

- Stratified k-fold cross-validation is same as just k-fold cross-validation, But in Stratified k-fold cross-validation, it does stratified sampling instead of random sampling.



Advantages of Stratified K-Fold Cross Validation

- Cross-validation implemented using stratified sampling ensures that the proportion of the feature of interest is the same across the original data, training set and the test set. This ensures that no value is over/under-represented in the training and test sets, which gives a more accurate estimate of performance/error.


THANK YOU