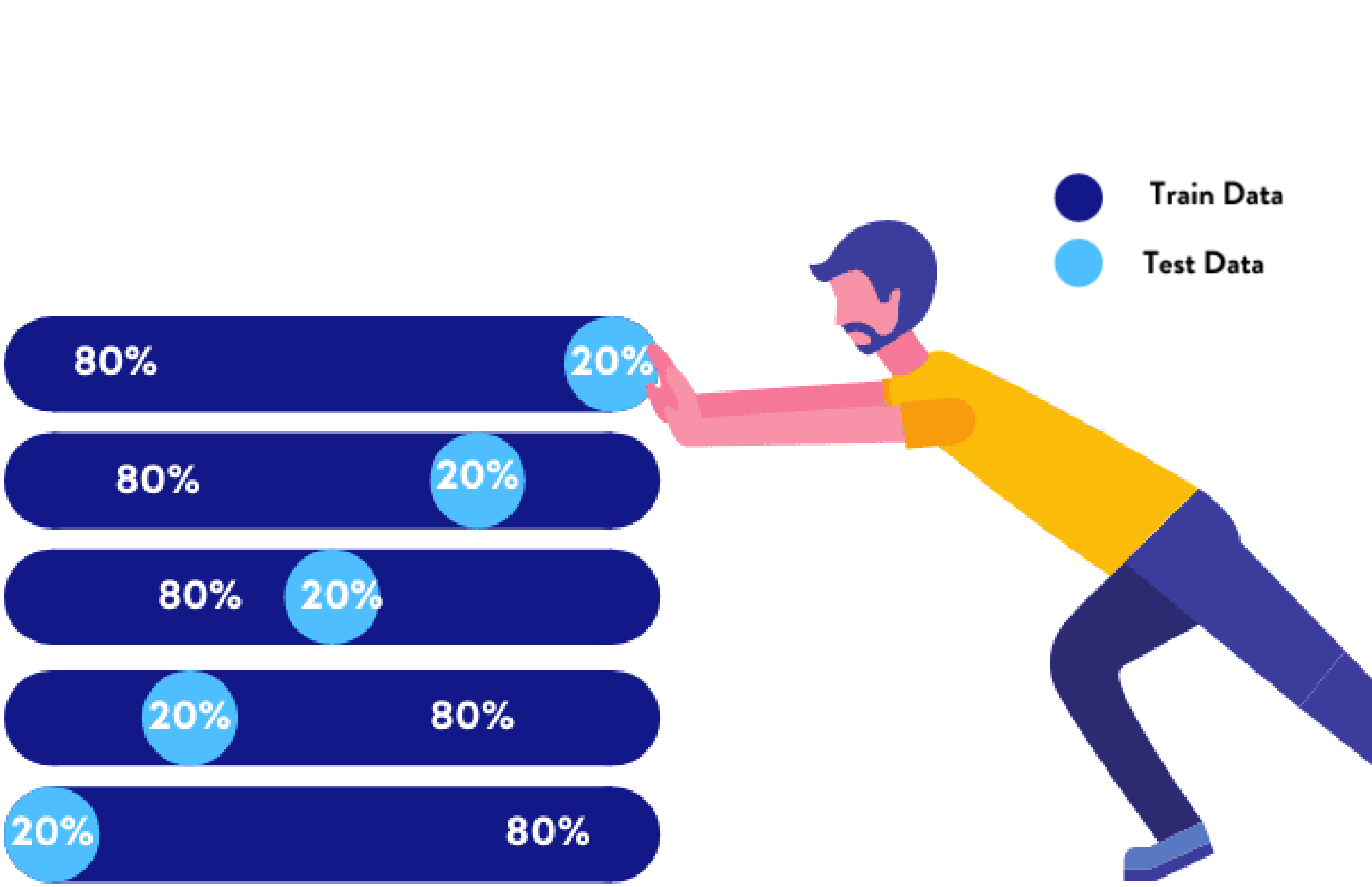
Analytics Vidhya

Cross-Validation and it's Types



Analytics Vidhya

Do you know why we split our training data into train and test data?

The training set is where you build your model, while the test set is used to see how well your model performs in the real world.

Why is this important, you ask?

Analytics Vidhya

Well, because there's always a chance that your model will work great on training data but then totally flop when it's put to the test in the actual world. And nobody wants that!

That’s where cross-validation comes to our rescue!

Cross-Validation

It is a technique for assessing Machine Learning models that involves training numerous more Machine Learning models on subsets of the available input data set and evaluating them on the subgroup.

Types of Cross-Validations

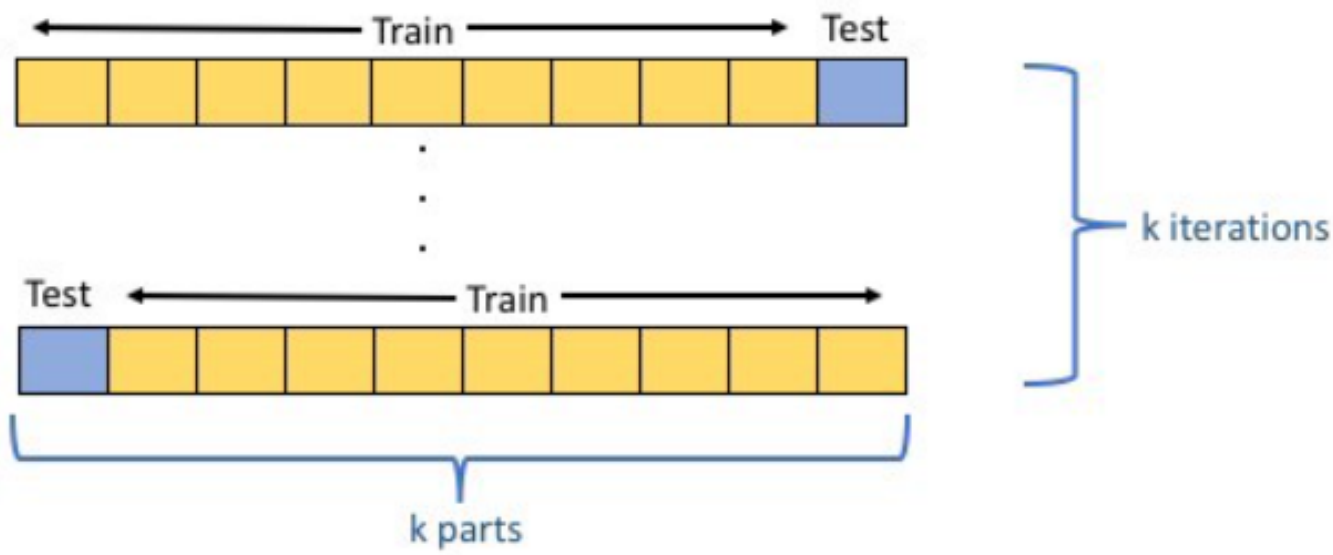
- K-Fold Cross-Validation
- Holdout Validation
- Stratified K-Fold Cross-Validation
- Leave-One_out Cross Validation

K-Fold Cross-Validation

It is a technique used in ML to evaluate the performance of a model. It involves dividing the dataset into **K no. of subsets or folds** of approximately equal size **randomly**. The model is then trained and evaluated K times, each time using a different fold as the validation set and the remaining folds as the training set.

K Folds Cross Validation Method

1. Divide the sample data into k parts.
2. Use k-1 of the parts for training, and 1 for testing.
3. Repeat the procedure k times, rotating the test set.
4. Determine an expected performance metric (mean square error, misclassification error rate, confidence interval, or other appropriate metric) based on the results across the iterations



Pros

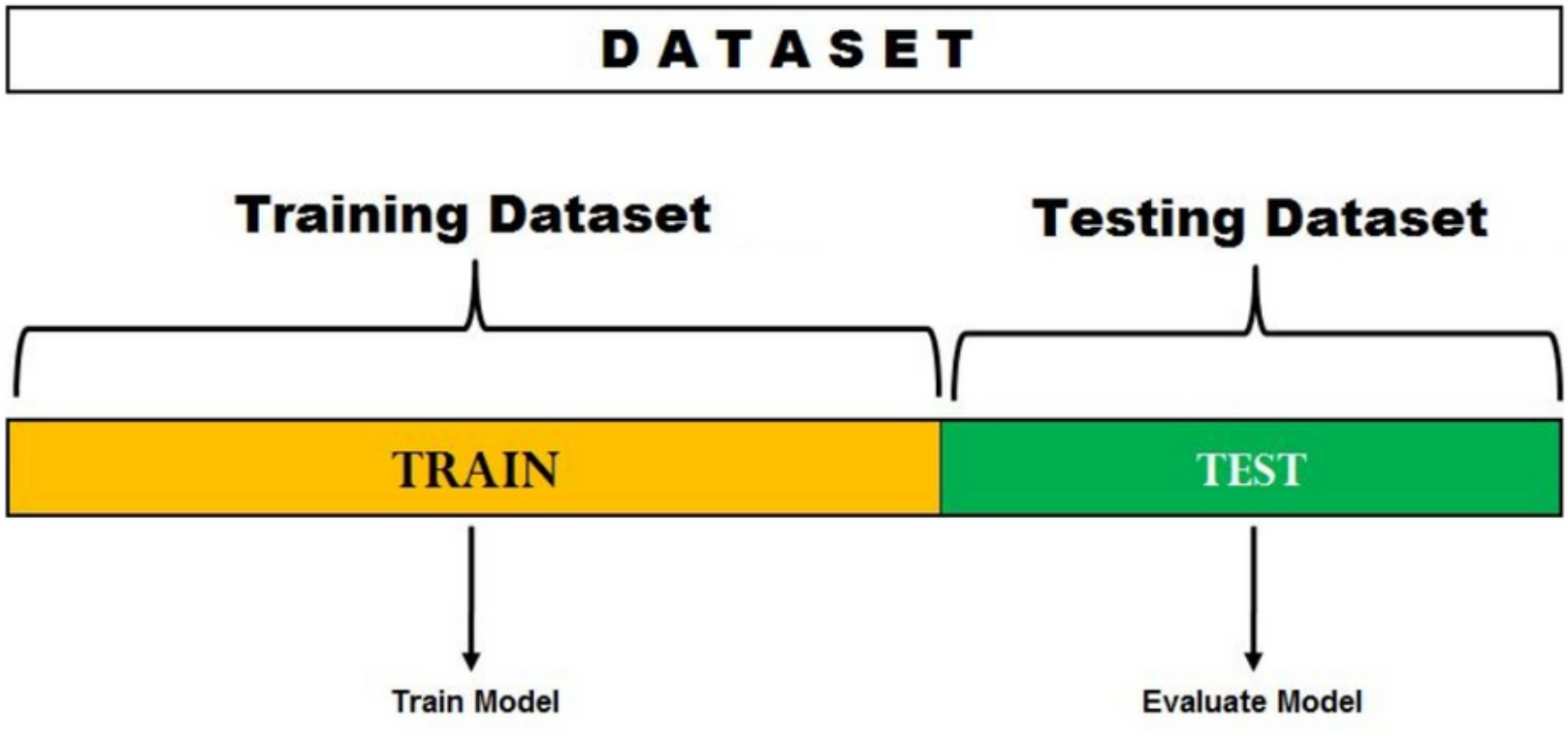
- This will aid in resolving the computing power issue.
- Models may be unaffected by the presence of an outlier in the data.
- It assists us in overcoming the issue of unpredictability.

Cons

- Incorrectly balanced data sets will affect our model.

Holdout Validation

Also known as the **train-test split**, is a technique used in ML to assess the performance of a model. This approach generates predictions by splitting the dataset into a training and test set. Train the model on the training set to learn patterns. Evaluate performance on the test set, containing unseen data, to gauge real-world effectiveness.



Pros

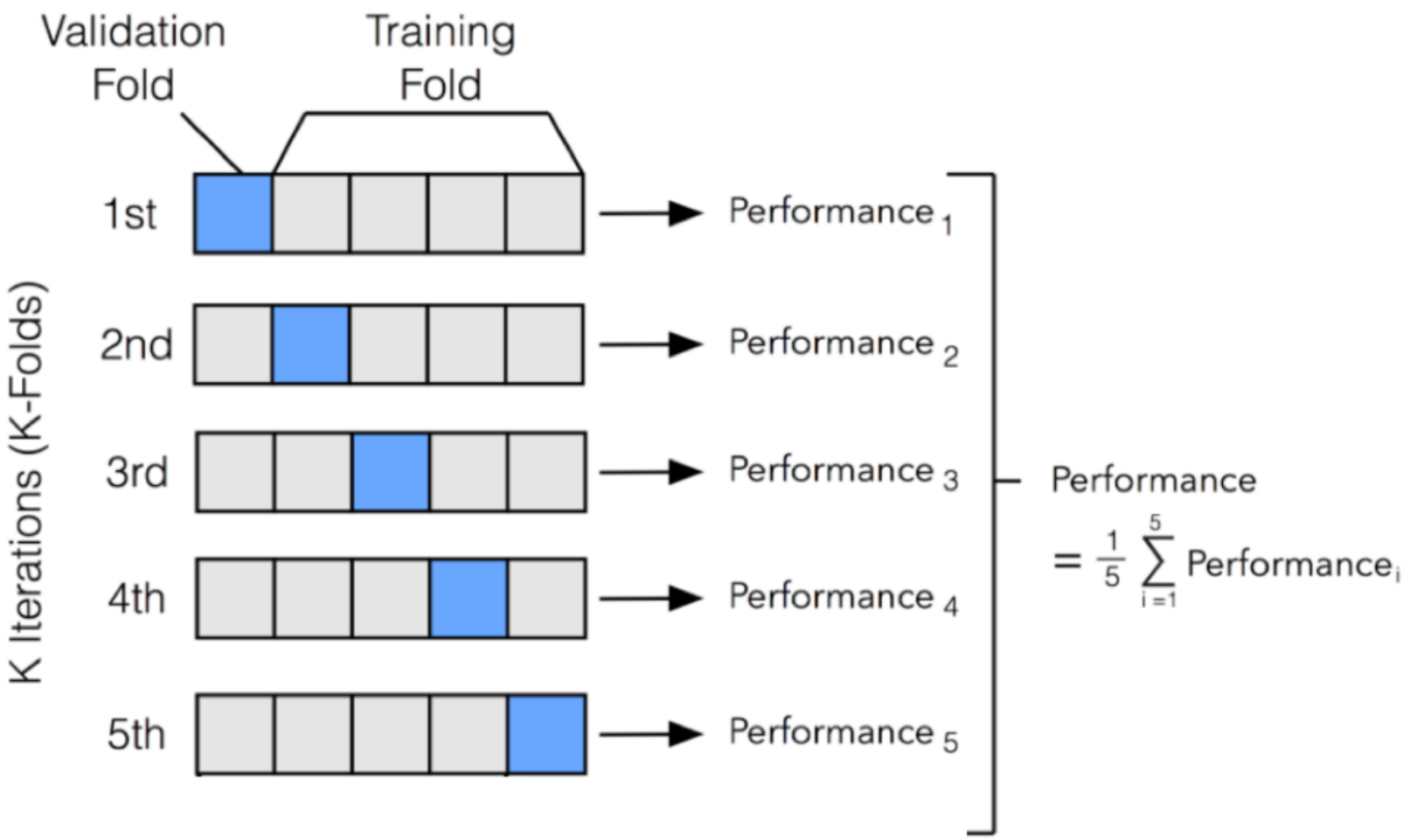
- This method is entirely data-independent.
- This method requires only one execution, which results in cheaper computing costs.

Cons

- Due to the lower amount of data, the performance is more variable.

Stratified K-Fold Cross-Validation

It is a technique used when dealing with classification problem of imbalance class distribution. It is the same as K Fold Cross Validation but does **stratified sampling** instead of random sampling. Stratified sampling ensures that each fold contains a representative sample of each class.



Pros

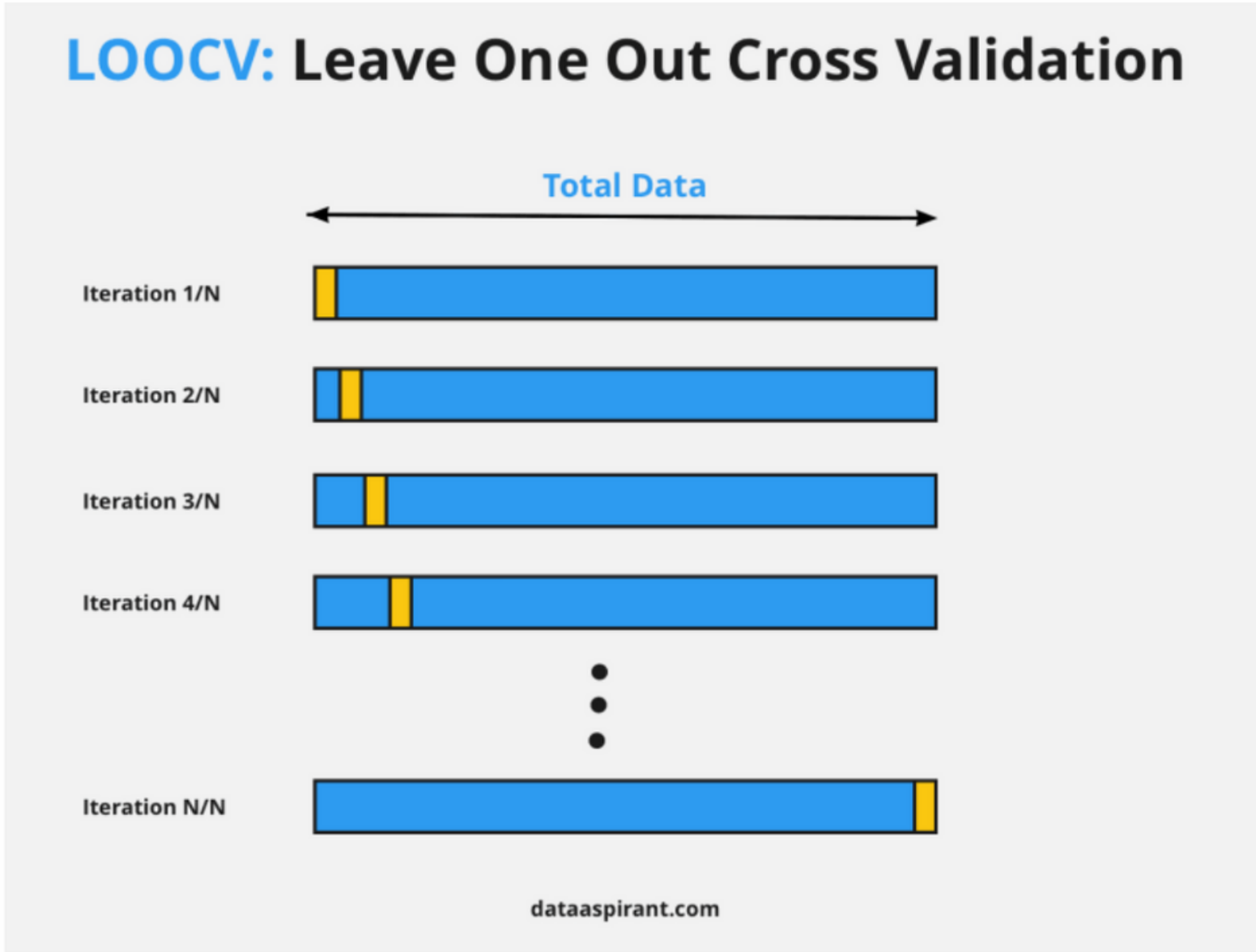
- It may enhance many models through hyper-parameter adjustment.
- Assists us in comparing models.
- It contributes to the reduction of both bias and variance.

Cons

- Execution is expensive.

Leave-One_out Cross

It is a technique used in machine learning to evaluate the performance of a model by **iteratively leaving out one data point** as the validation set and using the remaining data points as the training set.



Pros

- It contains no randomness.
- Bias will be reduced.

Cons

- This is a complete procedure that is computationally expensive.

Analytics Vidhya

That's a wrap. Was this post Helpful?

Follow us for more!

Analytics Vidhya