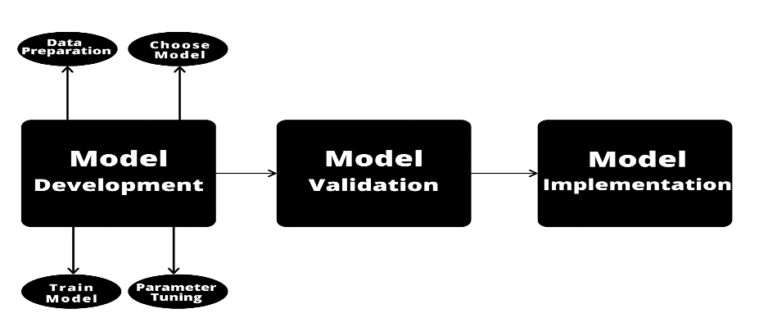
# ***MODEL VALIDATION APPROACH***

# What is Model Validation and Why is it Important?

We all have pursued enough articles about Machine Learning, and the first notion we often come up with is ‘Machine Learning is about making predictions.’

Yes, it is somewhat convincing, but these predictions come up after assorted processes like Data Preparation, Choosing a Model, Training the Model, Parameter Tuning, Model Validation, etc. So, only after carrying out the aforementioned operations, a Machine Learning Model (Regression or Classification) is efficient to make predictions.

Let’s have a look below to have a better understanding.



What is Model Validation?

So, as the name suggests ‘Model Validation’, we can perceive that the model is seeking some validation, but what’s that validation all about? Let’s try to answer it.

**Model validation** is the process that is carried out after **Model Training** where the trained model is evaluated with a testing data set. The testing data may or may not be a chunk of the same data set from which the training set is procured.

To know things better, we can note that the two types of Model Validation techniques are namely,

* In-sample validation – testing data from the same dataset that is used to build the model.
* Out-of-sample validation – testing data from a new dataset that isn’t used to build the model

Conclusion alert! Model validation refers to the process of confirming that the model achieves its intended purpose i.e., how effective our model is.

But how is it achieved? Take a look below.

The ultimate goal for any machine learning model is to learn from examples in such a manner that the model is capable of generalizing the learning to new instances which it has not yet seen. So, when we approach a problem with a dataset in hand, it is very important that we find the right machine learning algorithm to create our model. Every model has its own strengths and weaknesses. For instance, some algorithms have a higher tolerance for small datasets, while others may be good with large amounts of data. For this reason, two different models using similar data can predict different results with different degrees of accuracy and hence model validation is required.

**Following is the chronology for Model Validation-**

-Choose a machine learning algorithm.

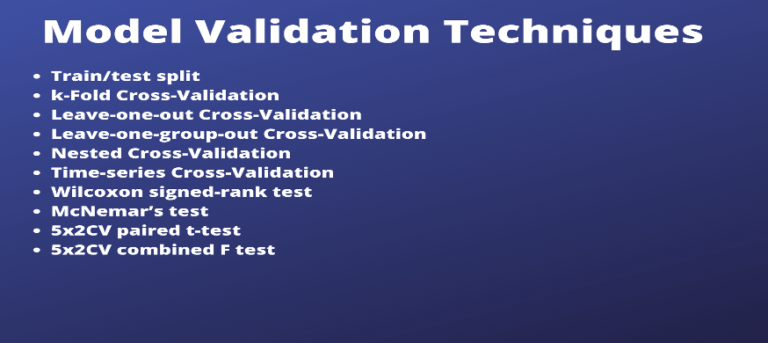
-Choose hyperparameters for the model.

-Fit the model to the training data.

-Use the model to predict labels for new data.

Note- In machine learning, we use the term **parameters** to refer to something that can be learned by the algorithm during training and **hyperparameters** to refer to something that is passed to the algorithm.

Then the accuracy score for the model is calculated and if in any case, this accuracy score is low, we change the value of the hyperparameters used in the model, and retest it until we get a decent accuracy score.



There are various ways of validating a model among which the two most famous methods are Cross Validation and Bootstrapping but there is no single validation method that works in all scenarios. Therefore, it is important to understand the type of data we are working with.

Although you can read more compositions to learn these techniques better.

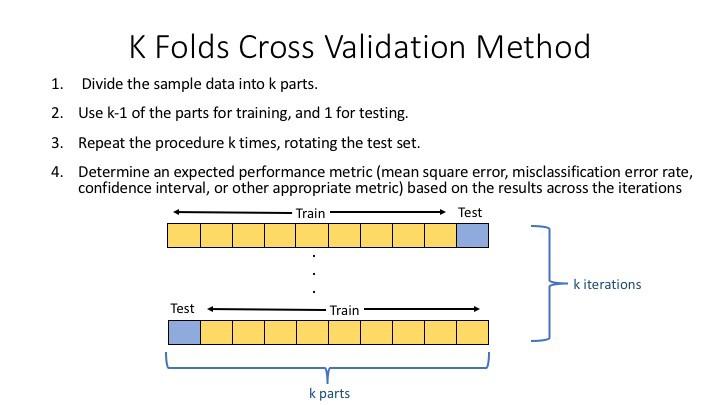
##### **Importance of Model Validation**

Now after having a glimpse of Model Validation, we all can imagine how important a component it is of the entire Model development process. Validating the machine learning model outputs are important to ensure its accuracy. When a machine learning model is trained, a huge amount of training data is used and the main aim of checking the model validation provides an opportunity for machine learning engineers to improve the data quality and quantity. As it happens, without checking and validating the model it is not right to rely on its prediction. And in sensitive areas like healthcare and self-driven vehicles, any kind of mistake in object detection can lead to major fatalities due to wrong decisions taken by the machine in real-life predictions. And validating the ML model at the training and development stage helps to make the model make the right predictions. Some added advantages of Model Validation are as follows.

* Scalability and flexibility
* Reduce the costs.
* Enhance the model quality.
* Discovering more errors
* Prevents the model from [overfitting and underfitting](https://www.educative.io/edpresso/overfitting-and-underfitting).

It is extremely important that data scientists validate machine learning models that are under training for accuracy and stability as it needs to be ensured that the model picks up on most of the trends and patterns in the data without incurring too much noise.

Now we are clear with the fact that building the machine learning model is not just enough to rely on its predictions, we need to check the accuracy and validate the same to ensure the precision of results given by the model and make it usable in real-life applications.



**What is data splitting in modelling?**

Data splitting is the process of splitting data into 3 sets:

Data which we use to design our models (Training set)

Data which we use to refine our models (Validation set)

Data which we use to test our models (Testing set)

If we do not split our data, we might test our model with the same data that we use to train our model.

**Example**

If the model is a trading strategy specifically designed for Apple stock in 2008, and we test its effectiveness on Apple stock in 2008, of course it is going to do well.

We need to test it on 2009’s data. Thus, 2008 is our training set and 2009 is our testing set.

To recap what are training, validation and testing sets…

**What is a Training Set?**

The training set is the set of data we analyse (train on) to design the rules in the model.

A training set is also known as the in-sample data or training data.

**What is a Validation Set?**

The validation set is a set of data that we did not use when training our model that we use to assess how well these rules perform on new data.

It is also a set we use to tune parameters and input features for our model so that it gives us what we think is the best performance possible for new data.

**What is a Test Set?**

The test set is a set of data we did not use to train our model or use in the validation set to inform our choice of parameters/input features.

We will use it as a final test once we have decided on our final model, to get the best possible estimate of how successful our model will be when used on entirely new data.

A test set is also known as the out-of-sample data or test data.

**Why do we need to split our data?**

To prevent look-ahead bias, overfitting and underfitting.

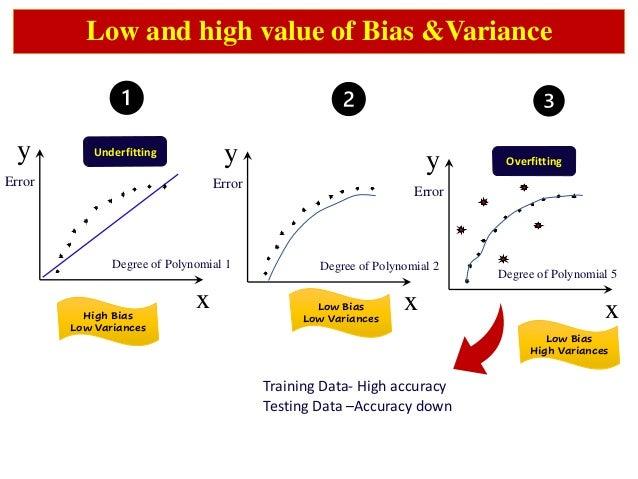
**Look-ahead bias**: Building a model based on data that is not supposed to be known.

**Overfitting**: This is the process of designing a model that adapts so closely to historical data that it becomes ineffective in the future.

**Underfitting**: This is the process of designing a model that adapts so loosely to historical data that it becomes ineffective in the future.

**BIAS VARIANCE DICHOTOMY**

The bias–variance dilemma or bias–variance problem is **the conflict in trying to simultaneously minimize these two sources of error that prevent supervised learning algorithms from generalizing beyond their training set**: The bias error is an error from erroneous assumptions in the learning algorithm.



Machine learning is a branch of Artificial Intelligence, which allows machines to perform data analysis and make predictions. However, if the machine learning model is not accurate, it can make predictions errors, and these prediction errors are usually known as Bias and Variance. In machine learning, these errors will always be present as there is always a slight difference between the model predictions and actual predictions. The main aim of ML/data science analysts is to reduce these errors in order to get more accurate results. In this topic, we are going to discuss bias and variance, Bias-variance trade-off, Underfitting and Overfitting. But before starting, let's first understand what errors in Machine learning are?

In machine learning, an error is a measure of how accurately an algorithm can make predictions for the previously unknown dataset. On the basis of these errors, the machine learning model is selected that can perform best on the particular dataset. There are mainly two types of errors in machine learning, which are:

**Reducible errors:** These errors can be reduced to improve the model accuracy. Such errors can further be classified into bias and Variance.

**Irreducible errors:** These errors will always be present in the model

regardless of which algorithm has been used. The cause of these errors is unknown variables whose value can't be reduced.

**What is Bias?**

In general, a machine learning model analyses the data, find patterns in it and make predictions. While training, the model learns these patterns in the dataset and applies them to test data for prediction. While making predictions, a difference occurs between prediction values made by the model and actual values/expected values, and this difference is known as bias errors or Errors due to bias. It can be defined as an inability of machine learning algorithms such as Linear Regression to capture the true relationship between the data points. Each algorithm begins with some amount of bias because bias occurs from assumptions in the model, which makes the target function simple to learn. A model has either:

x

**Low Bias:** A low bias model will make fewer assumptions about the form of the target function.

**High Bias**: A model with a high bias makes more assumptions, and the model becomes unable to capture the important features of our dataset. A high bias model also cannot perform well on new data.

Generally, a linear algorithm has a high bias, as it makes them learn fast. The simpler the algorithm, the higher the bias it has likely to be introduced. Whereas a nonlinear algorithm often has low bias.

Some examples of machine learning algorithms with low bias are Decision Trees, k-Nearest Neighbours and Support Vector Machines. At the same time, an algorithm with high bias is Linear Regression, Linear Discriminant Analysis and Logistic Regression.

**Ways to reduce High Bias:**

High bias mainly occurs due to a much simple model. Below are some ways to reduce the high bias:

Increase the input features as the model is underfitted.

Decrease the regularization term.

Use more complex models, such as including some polynomial features.

**What is a Variance Error?**

The variance would specify the amount of variation in the prediction if the different training data was used. In simple words, variance tells that how much a random variable is different from its expected value. Ideally, a model should not vary too much from one training dataset to another, which means the algorithm should be good in understanding the hidden mapping between inputs and output variables. Variance errors are either of low variance or high variance.

Low variance means there is a small variation in the prediction of the target function with changes in the training data set. At the same time, High variance shows a large variation in the prediction of the target function with changes in the training dataset.

A model that shows high variance learns a lot and perform well with the training dataset, and does not generalize well with the unseen dataset. As a result, such a model gives good results with the training dataset but shows high error rates on the test dataset.

Since, with high variance, the model learns too much from the dataset, it leads to overfitting of the model. A model with high variance has the below problems:

A high variance model leads to overfitting.

Increase model complexities.

Usually, nonlinear algorithms have a lot of flexibility to fit the model, have high variance.

Some examples of machine learning algorithms with low variance are, Linear Regression, Logistic Regression, and Linear discriminant analysis. At the same time, algorithms with high variance are decision tree, Support Vector Machine, and K-nearest neighbours.

**Ways to Reduce High Variance:**

Reduce the input features or number of parameters as a model is overfitted.

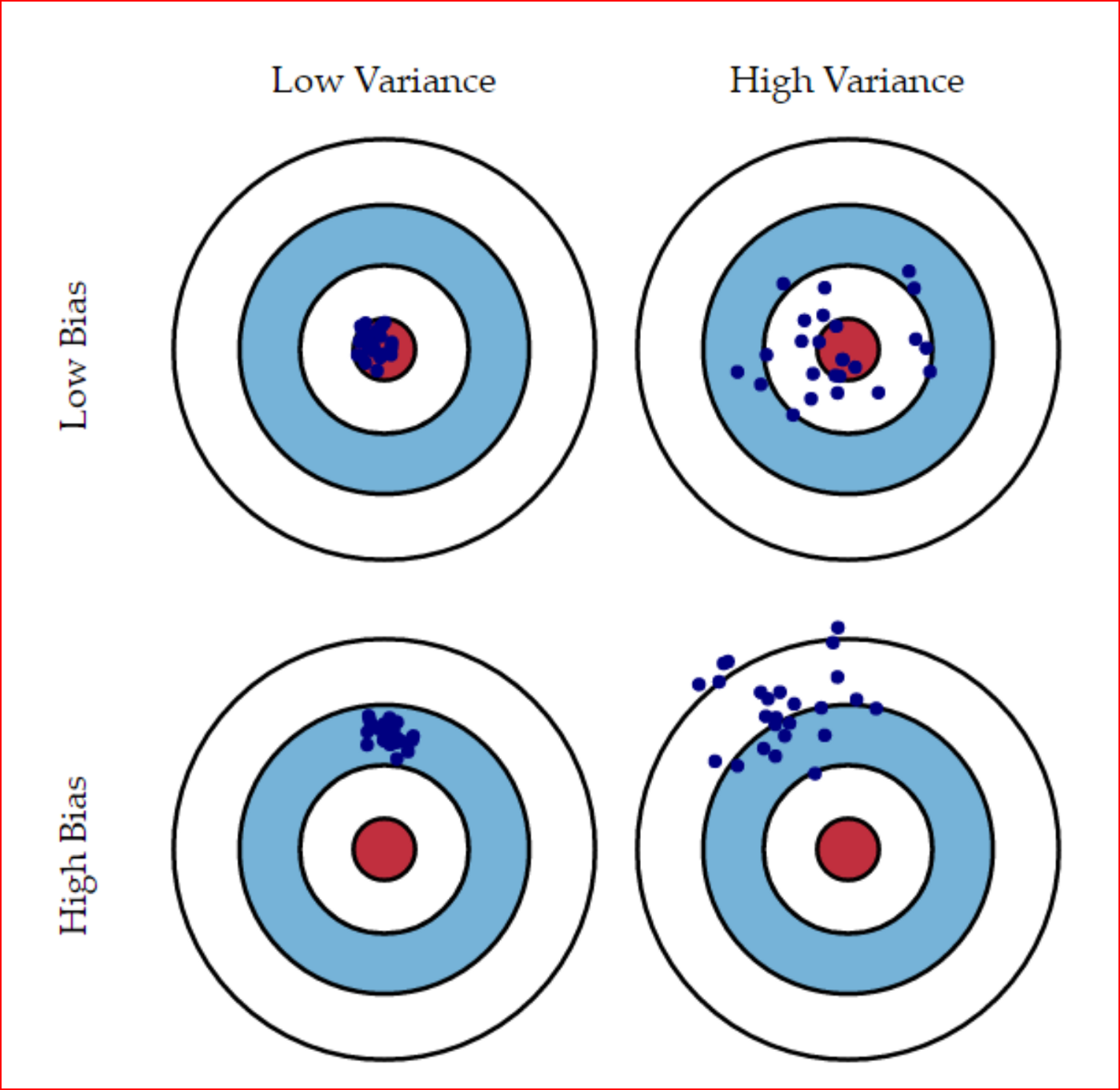
Do not use a much complex model.

Increase the training data.

Increase the Regularization term.

**Different Combinations of Bias-Variance**

There are four possible combinations of bias and variances, which are represented by the below diagram:



**Low-Bias, Low-Variance:**

The combination of low bias and low variance shows an ideal machine learning model. However, it is not possible practically.

**Low-Bias, High-Variance:** With low bias and high variance, model predictions are inconsistent and accurate on average. This case occurs when the model learns with a large number of parameters and hence leads to an overfitting.

**High-Bias, Low-Variance:** With High bias and low variance, predictions are consistent but inaccurate on average. This case occurs when a model does not learn well with the training dataset or uses few numbers of the parameter. It leads to underfitting problems in the model.

**High-Bias, High-Variance:**

With high bias and high variance, predictions are inconsistent and also inaccurate on average.

**How to identify High variance or High Bias?**

High variance can be identified if the model has:

Low training error and high test error.

High Bias can be identified if the model has:

High training error and the test error is almost similar to training error.

**Bias-Variance Trade-Off**

While building the machine learning model, it is really important to take care of bias and variance in order to avoid overfitting and underfitting in the model. If the model is very simple with fewer parameters, it may have low variance and high bias. Whereas, if the model has a large number of parameters, it will have high variance and low bias. So, it is required to make a balance between bias and variance errors, and this balance between the bias error and variance error is known as the Bias-Variance trade-off.