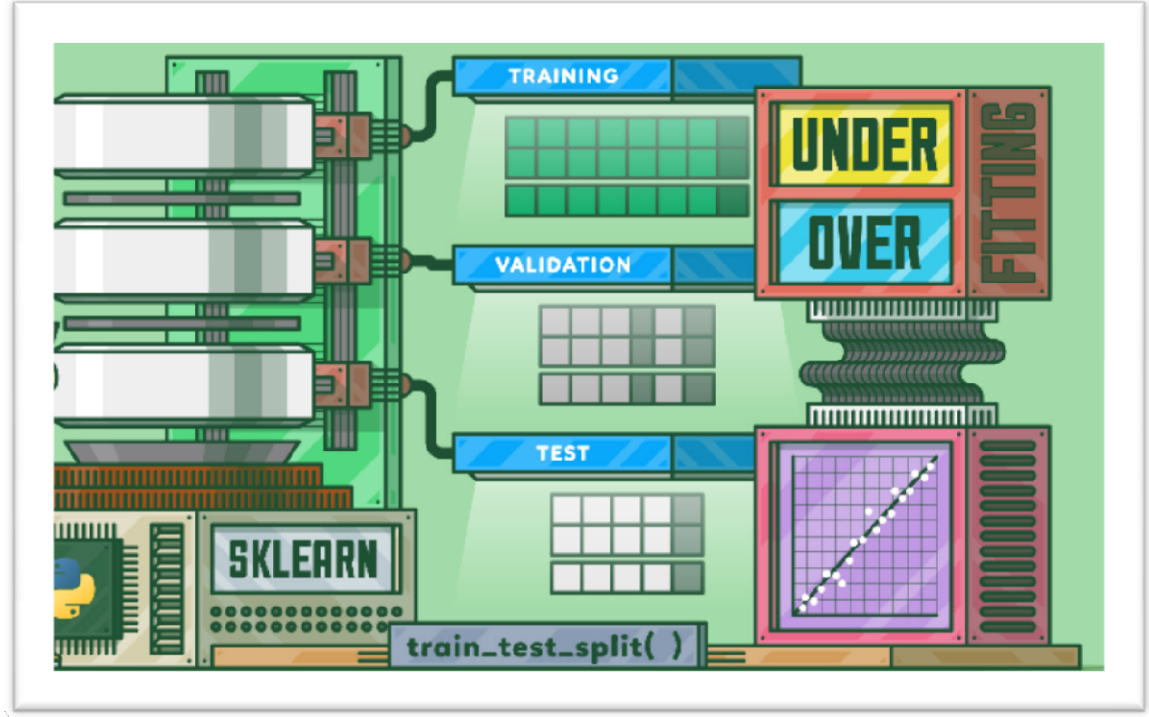
OVER FITTING
UNDER FITTING

Split Your Dataset With scikit-learn's train_test_split()



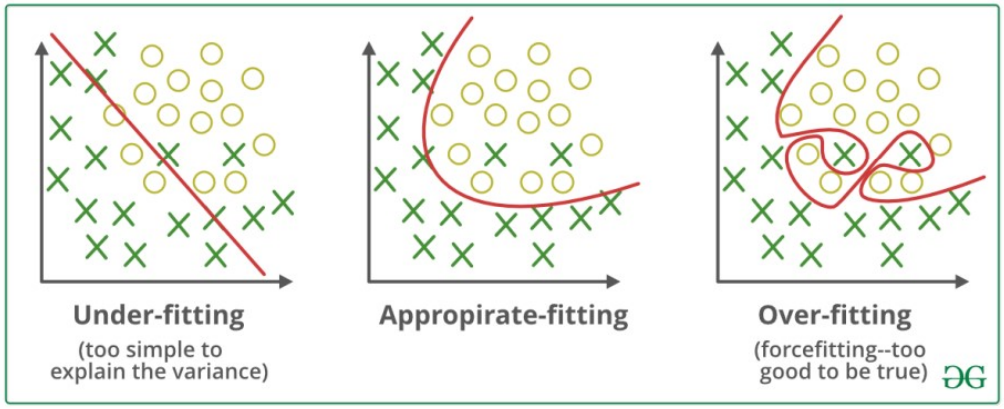
taroonreddy.com

Underfitting and Overfitting

Splitting a dataset might also be important for detecting if your model suffers from one of two very common problems, called under fitting and over fitting.

Bias:
Assumptions made by a model to make a function easier to learn.

Variance:
If you train your data on training data and obtain a very low error, upon changing the data and then training the same previous model you experience a high error, this is variance.



Underfitting:

A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data.

- Underfitting destroys the accuracy of our machine learning model.

- Its occurrence simply means that our model or the algorithm does not fit the data well enough.

- It usually happens when we have fewer data to build an accurate model and also when we try to build a linear model with fewer non-linear data.

- In such cases, the rules of the machine learning model are too easy and flexible to be applied on such minimal data and therefore the model will probably make a lot of wrong predictions.

Underfitting – High bias and low variance

Underfitting can be avoided by using more data and also reducing the features by feature selection.

Techniques to reduce underfitting:

- Increase model complexity
- Increase the number of features, performing feature engineering
- Remove noise from the data.
- Increase the number of epochs or increase the duration of training to get better results.

Overfitting:

A statistical model is said to be overfitted when we train it with a lot of data
When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set.

Then the model does not categorize the data correctly, because of too many details and noise.

The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models.

High variance and low bias

A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.

Techniques to reduce over fitting:

- Increase training data.

- Reduce model complexity.

- Early stopping during the training phase
(have an eye over the loss over the training period as soon as loss begins to increase stop training).

- Ridge Regularization and Lasso Regularization

- Use dropout for neural networks to tackle over fitting.

Best Fit Model:

Model makes the predictions with Zero error, is said to have a *Best fit* on the data.
It's achievable at a spot between overfitting and underfitting.

Performance of our model over time is learning from training dataset.

Model will keep on learning and reducing the error for the model on the training and testing data. If it will learn for too long, the model will become more prone to overfitting due to the presence of noise and less useful details. Hence the performance of our model will decrease.

In order to get a good fit, we will stop at a point just before where the error starts increasing. At this point, the model is said to have good learning skills on training datasets as well as our unseen testing dataset.