"Bias-Variance" Tradeoff :-

It is very important to know the definitions of Bias (&) Variance as a Data Scientist because it plays a vital role in assessing the quality of a Model.

Bias :- It is a phenomenon that skews the result of an algorithm in favour (or) against an idea.

Explaination :-

Consider an idea as → [Training data], if the model (or) algorithm was able to capture the pattern (or) data points accurately, it implies that the model has $\boxed{\text{Low Bias.}}$. If the model was unable to capture the pattern (or) data points from the training data accuratly, it implies that the model has $\boxed{\text{high Bias.}}$

Hence, Bias is a measurement of how accurately a model can capture a pattern in a Training Dataset.

Variance :- Variance refers to the change in the model when using different portions (OR) subsets of the training (OR) test datasets.

Explaination :-

Consider we have a dataset with (N) datapoints and out of that we have selected a subset of training data (S1) and the left over samples for testing the Model i.e (S_1').

S1 → Train Data
S_1' → Test Data
→ Test Error (T_1)

Consider, we trained the model using S1, and tested the model on S_1' and obtained the test Error as (T_1).

Similarily, Next time we have selected different training set (S2) and different test set ($S2'$) and obtained the test Error as (T_2).

S2 → Train data
$S2'$ → Test data
→ Test Error (T_2)

If the difference between different test Errors i.e (T_1) and (T_2) is high, it implies High Variance (or) high variability in the Test Error. Else, if the difference between different test Errors i.e (T_1) & (T_2) is low, it implies Low Variance.

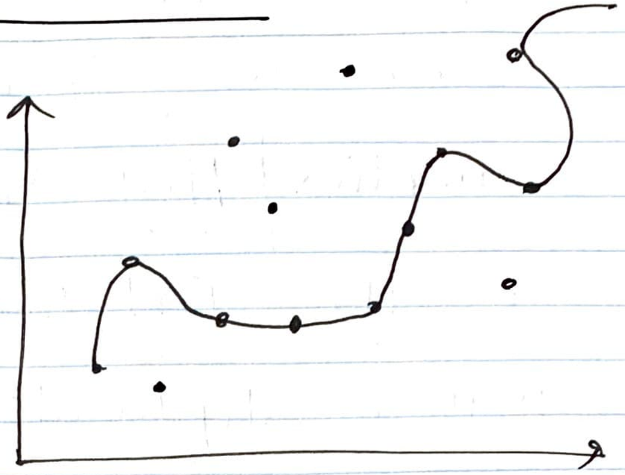
Thereby, High variance implies that the test Error varies greatly based on the Selection of the Training dataset.

//ly

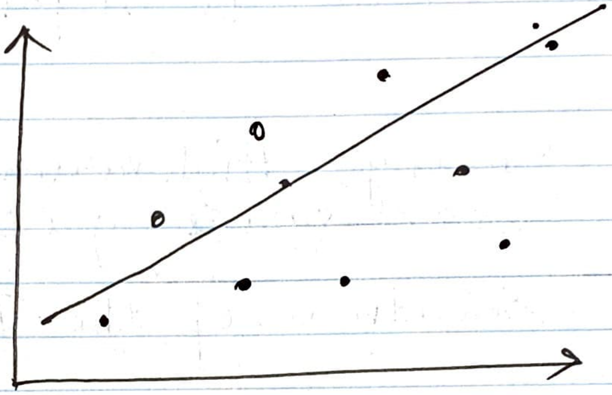
if you select different training samples, and found that the difference between the test Errors is almost same (or) low. means Low Variance.

Consider we have three models:-

Overfit :-



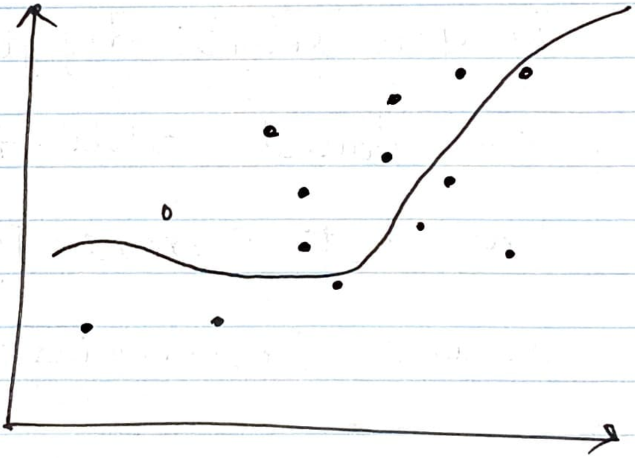
Underfit :-



Generalized model

(Or)

Balanced fit :-



* In the Overfit model, the model captures all the training data points precisely but fails to predict the output if the model is tested on unseen data.

Therefore, the training Error = 0 since, it captures all the data points in training set precisely. ⇒ Low Bias

In Overfitted model the test Error varies greatly based on the selection of the training dataset. Therefore High Variance (or) high variability in test error.

Overfit → Low Bias & High Variance.

* In the Underfit model, the model fails to capture the pattern from the training dataset, this implies high training Error ⇒ High Bias.

In Underfit model, even if you select different training samples to train the model, the difference between test Errors will be almost same i.e

Low Variance.

∴ Underfit → High Bias & Low Variance.

From above, it is clear that as a Data Scientist we should always try to build a model which has Low Bias & Low Variance i.e Perfect Model (or) Generalized Model that can capture most of the datapoints in the training dataset and ~~Also~~ should also have low test Error for different training samples.

Like this we can assess the quality of a Machine Learning Model using "Bias-Variance" Tradeoff.