What is Over-fitting?

Supervised machine learning is best understood as approximating a target function (f) that maps input variables (X) to an output variable (Y).

Y = f(X)

The goal of a good machine learning model is to generalize well from the training data to any data from the problem domain. This allows us to make predictions in the future on data the model has never seen.

Over-fitting and under-fitting are the two biggest causes for poor performance of machine learning algorithms.

Overfitting refers to a model that models the training data too well. Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the models ability to generalize. Overfitting is more likely with nonparametric and nonlinear models that have more flexibility when learning a target function. As such, many nonparametric machine learning algorithms also include parameters or techniques to limit and constrain how much detail the model learns.

For example, decision trees are a nonparametric machine learning algorithm that is very flexible and is subject to overfitting training data. This problem can be addressed by pruning a tree after it has learned in order to remove some of the detail it has picked up.This will be talked in detail in the next Module.

Underfitting refers to a model that can neither model the training data nor generalize to new data.

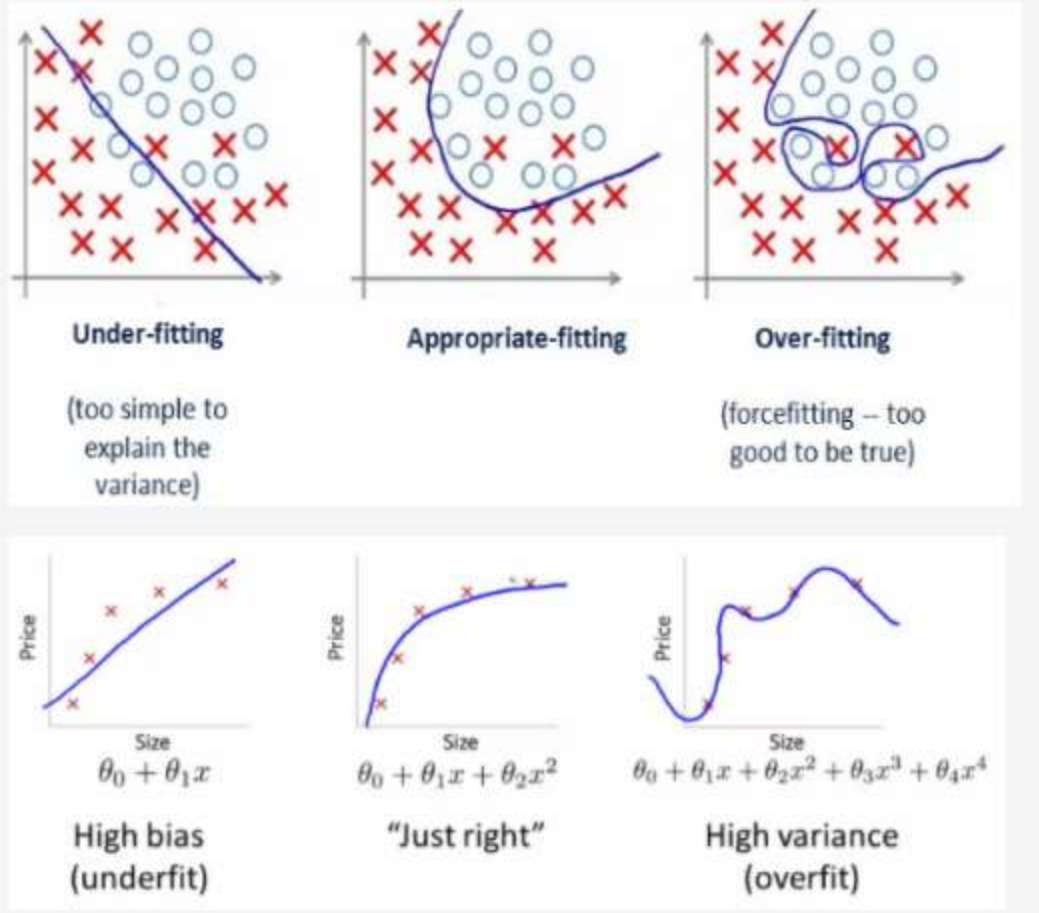
An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data.

Ideally, you want to select a model at the sweet spot between underfitting and overfitting. Both overfitting and underfitting can lead to poor model performance. But by far the most common problem in applied machine learning is overfitting.

Summary

Overfitting: Good performance on the training data, poor generliazation to other data.

Underfitting: Poor performance on the training data and poor generalization to other data



For example, it would be a big red flag if our model saw 99% accuracy on the training set but only 55% accuracy on the test set.

How to Prevent Overfitting

1) Cross validation (discussed in a later Module)

2) Train with More data- It won't work everytime, but training with more data can help algorithms detect the signal better. Of course, that's not always the case. If we just add more noisy data, this technique won't help.

3) Remove unwanted IV's- Remove unwanted features that exist in the data. This can be known by business intuition/logic or using P values (for example in Linear/Logistic Regression).

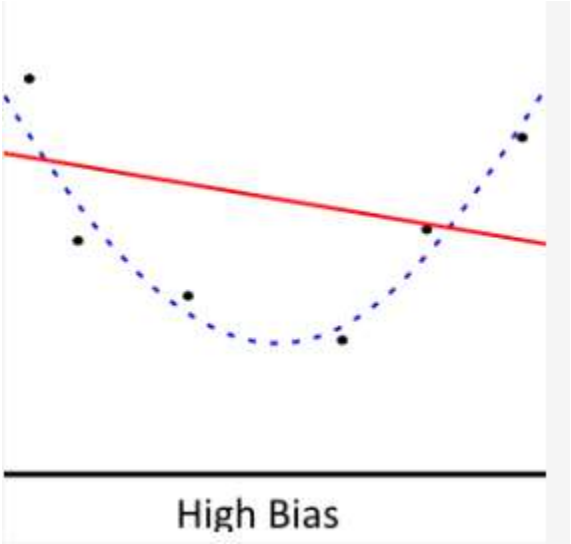
4) Early Stopping- When you're training a learning algorithm iteratively, you can measure how well each iteration of the model performs. So, keep checking the accuracy after every iteration and stop whenever you find that you are overtraining or when the test data accuracy starts to fall. (Discussed in details in Next Module)

Bias vs Variance theory

Bias:-

Bias means how far off our predictions are from real values. Generally parametric algorithms have a high bias. They are easier to understand but generally less flexible. For example a Linear Regression Model (for one IV) will always plot a straight line, irrespective of the relation between DV and IV. This is a inflexible model as it does not change but always remains a line. So, it may have high Bias.

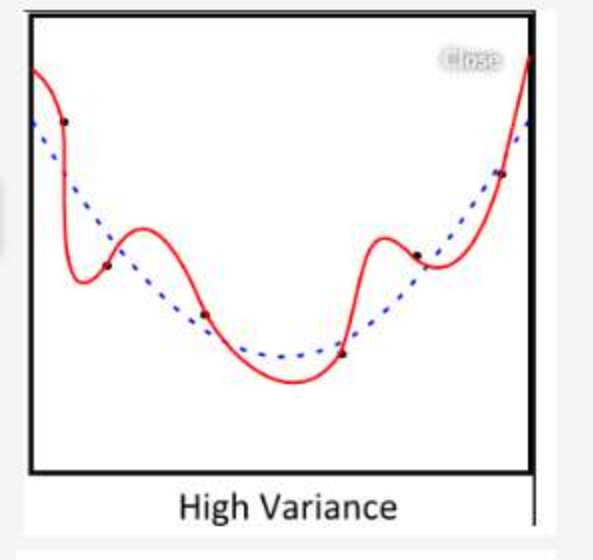
Examples of high-bias machine learning algorithms include: Linear Regression, Logistic Regression. High Bias Models usually result in Under fitting as the model is unable to capture the underlying pattern of the data.



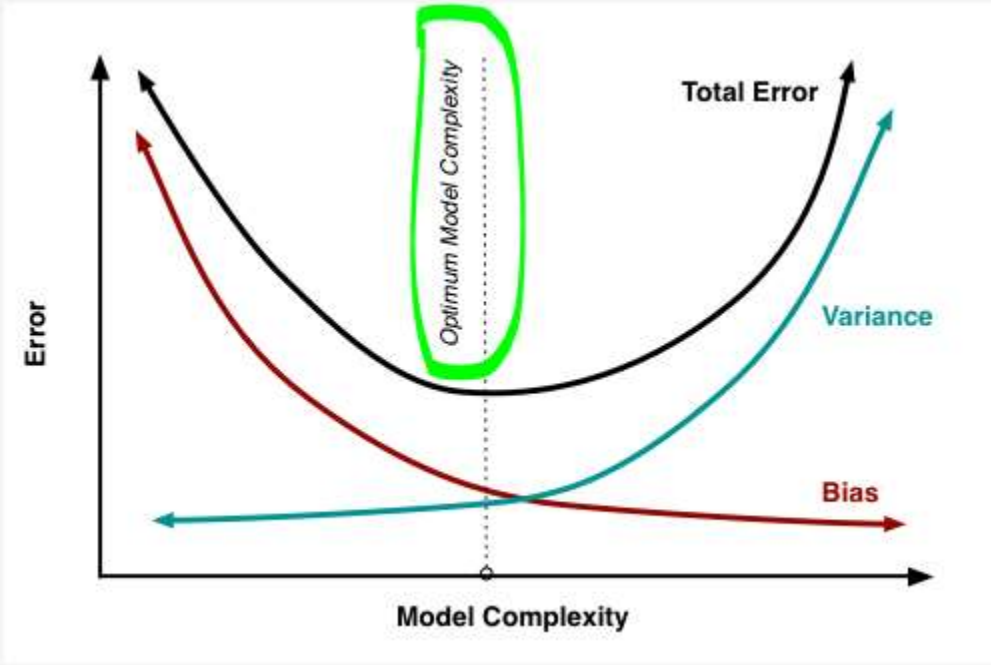
Variance:-

Change in predictions across different data sets. Again, imagine you can repeat the entire model building process multiple times. The variance is how much the predictions for a given point vary between different realizations of the model. In other words, Variance is the amount that the estimate of the target function will change if different training data (subsets from the same dataset) was used. Ideally, it should not change too much from one training dataset to the next, meaning that the algorithm is good at picking out the hidden underlying mapping between the inputs and the output variables. Generally non-parametric machine learning algorithms that have a lot of flexibility have a high variance. For example decision trees have a high variance.

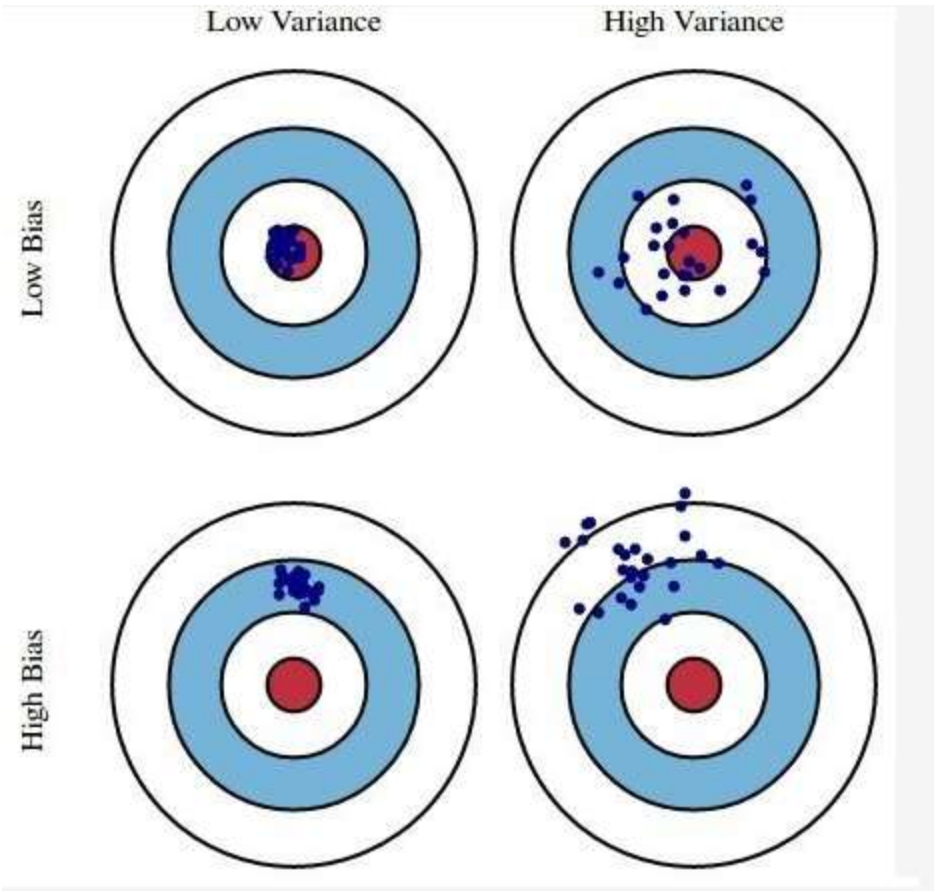
Complex Models usually have a high variance and cause overfitting.



At its root, dealing with bias and variance is really about dealing with over- and under-fitting. Bias is reduced and variance is increased in relation to model complexity. As more and more parameters are added to a model, the complexity of the model rises and variance becomes our primary concern while bias steadily falls. For example, as more polynomial terms are added to a linear regression, the greater the resulting model's complexity will be.



Bias vs Variance in ONE IMAGE


Low Variance
High Variance
Low Bias
High Bias