

BIAS and VARIANCE

Today, we will look into bias and variance!

look at belowm picyures carefully

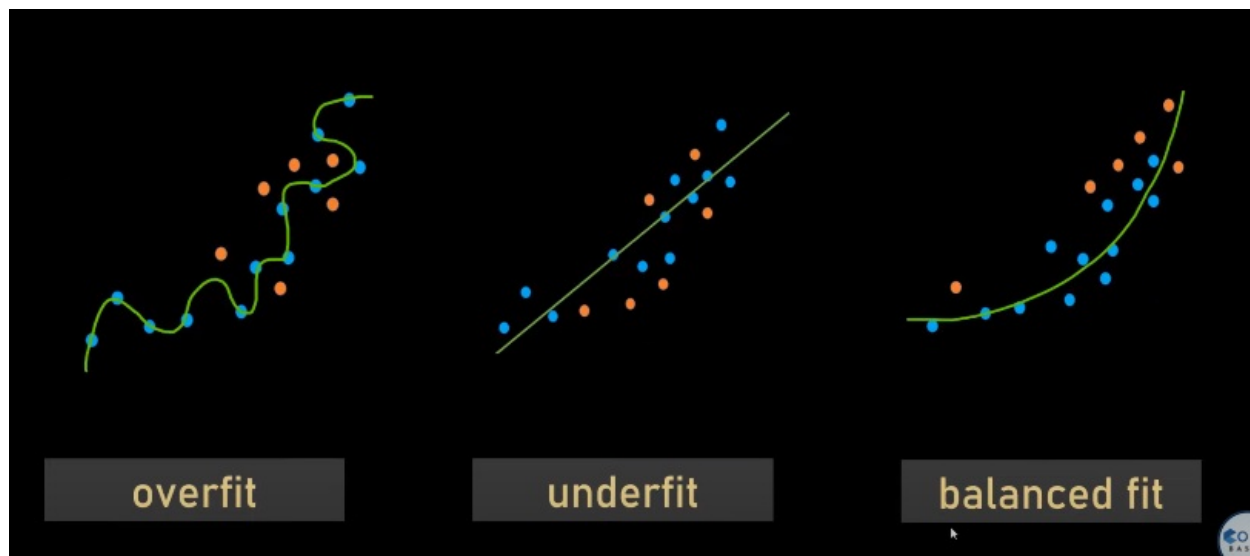


1. The first pic shows a Tshirt which seems to be too tight, means too fit, basically mean **Overfit**, so if the person gained or loses weight it will not fit
2. Second pic is **Underfit** since the Tshirt seems to be too loose
3. The third pic is however a perfect fit, where even if the person grow or lose weight, it will still be fit



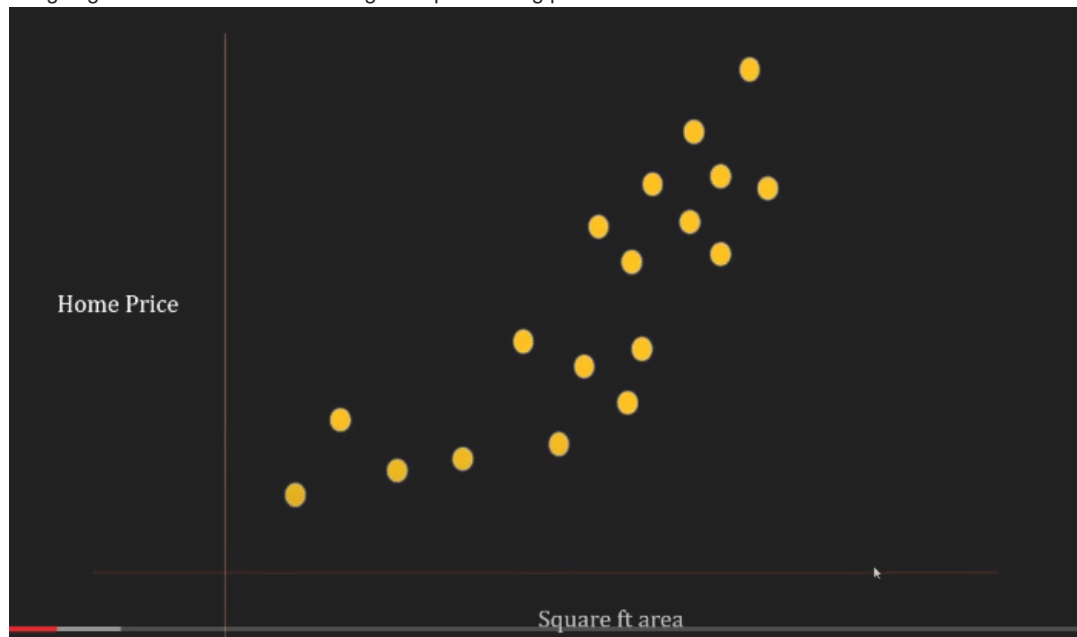
the same concept is applied in the ML World where ur model can be Underfit, Overfit or Balance fit

So,



We

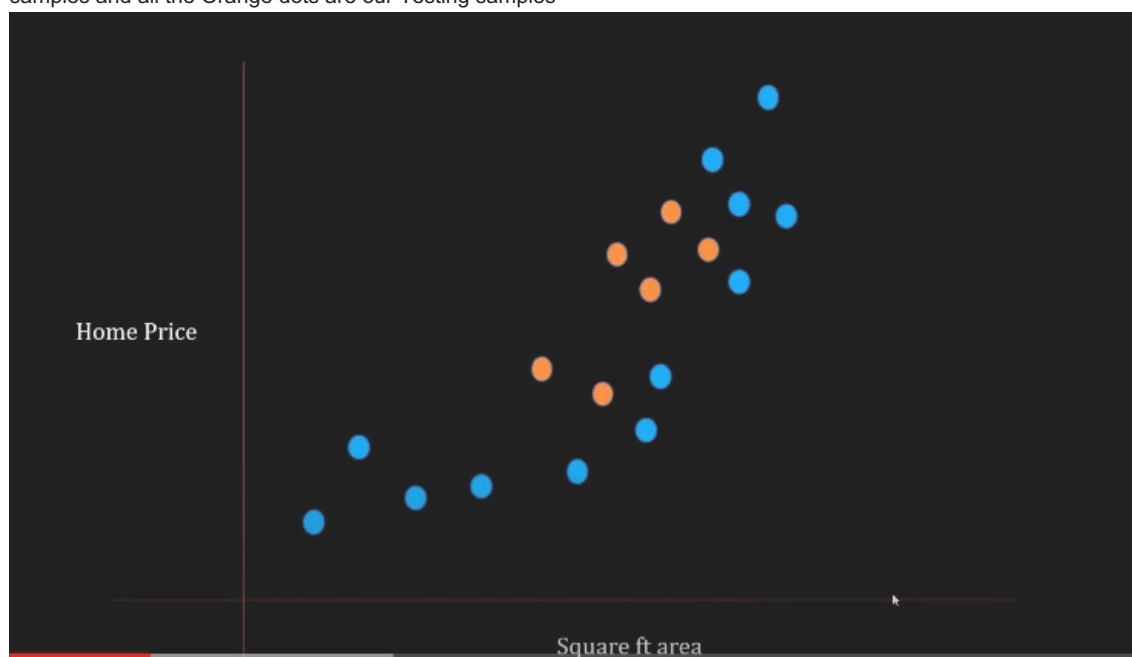
are going to look at this scenario using a simple housing price dataset



We can see, we have

a sqrt foot area and based on that we are trying to build an ML Model that can predict the house price.

In this scatter plot, what we do is, we split our data into Training and Testing samples, lets say all below Blue dots are our Training samples and all the Orange dots are our Testing samples

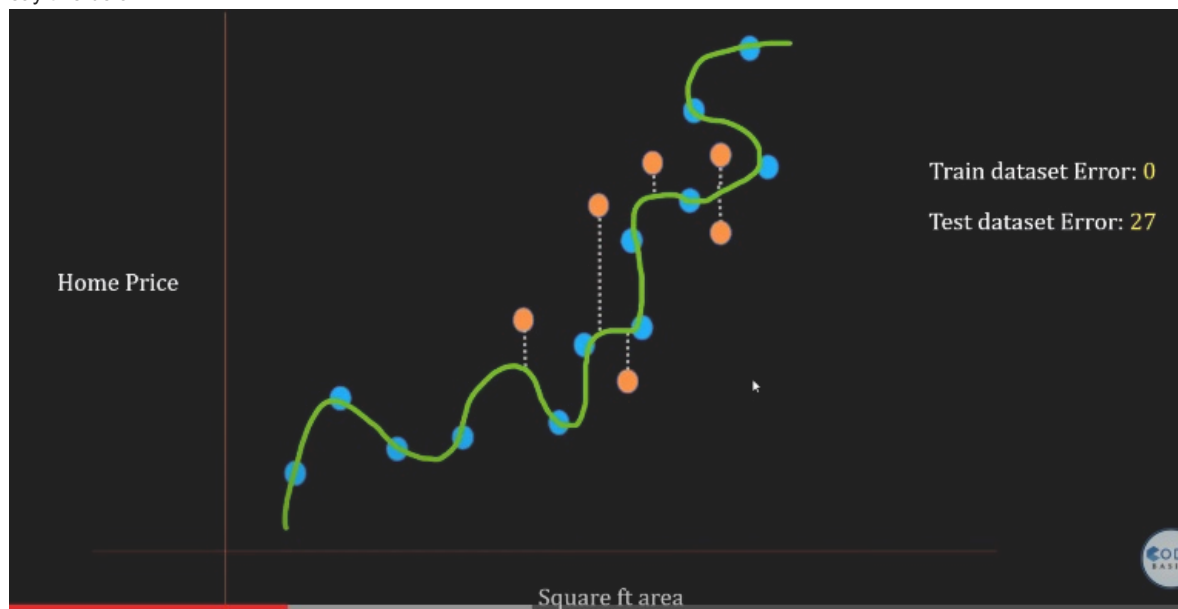


We can train our

model that fit to these blue dots, Lets say our we end up in an Overfit model, An overit model tried to fit exactly to the Training samples where your Training error becomes close to 0 or 0, but the problem here is, the Test error can become high, because once u have your model trained which is the green line in below diagram, then now if u want to figure out the Error for a particular datapoint and the Error will be the grey dotted line and u can measure your Error for aLL ur test dataset and average it, Lets assume u get the Test Error as 100



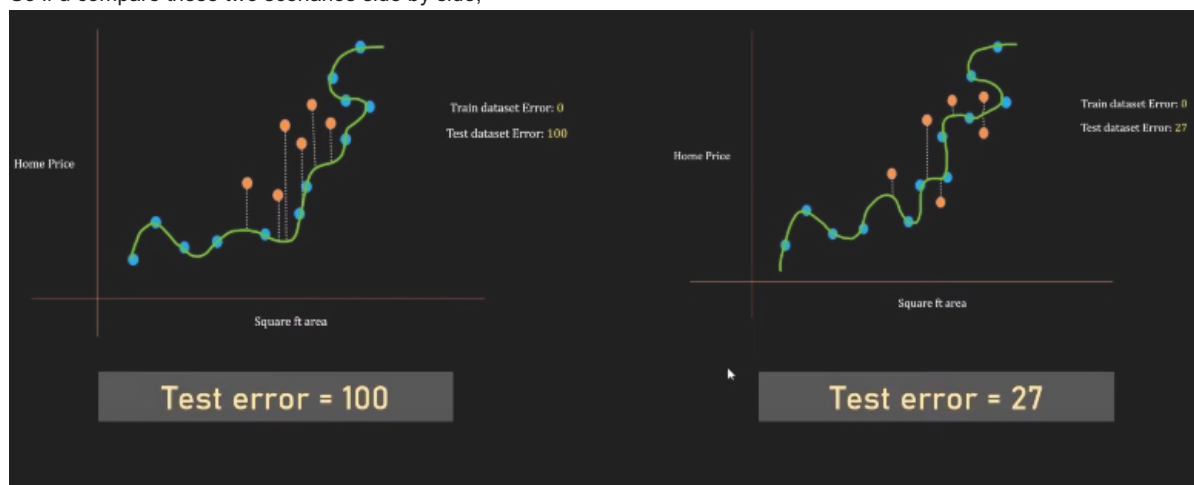
Remember one thing, when u have a dataset and when u pick your Training samples, you pick it up at Random, you have Train Test split lets say 80/20 percent, YOU PICK YOUR TRAINING TEST AT RANDOM! So some other people might chose different samples for Training lets say this below



Due to this, their model might looks different, as seen above pic, they are using the same model, same methodology, so your Training dataset Error is still 0 because u both are trying to overfit the Model, But the problem that happened here is your Test dataset Error, lets say 27 as seen in above pic,

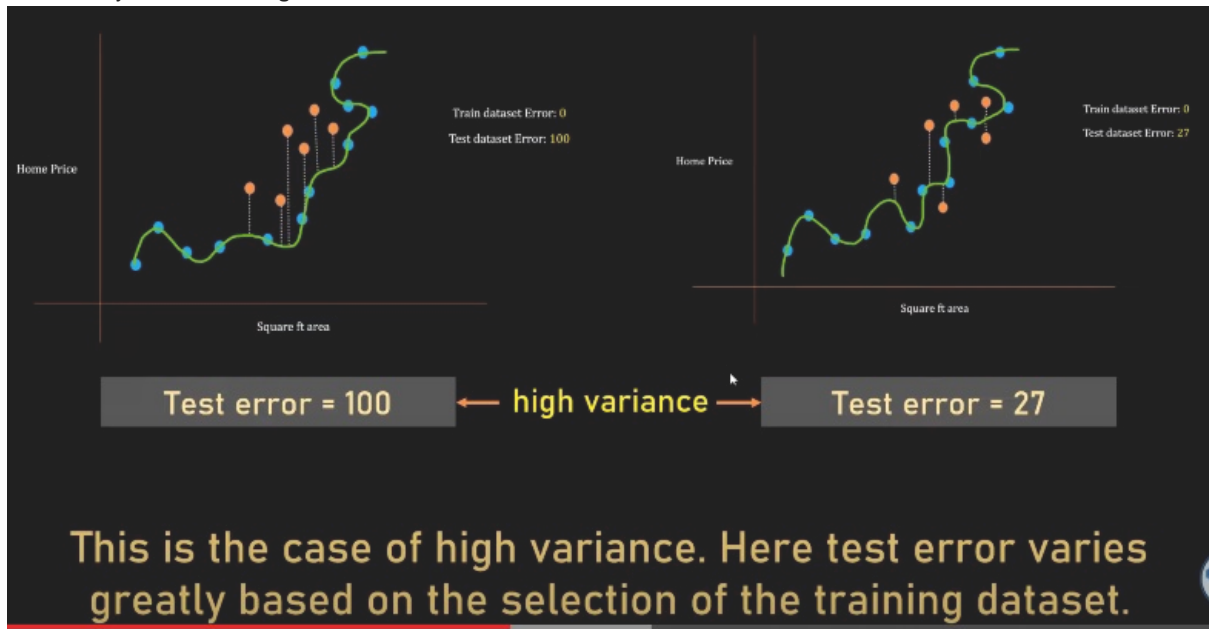
You can compare the first one and the second set and can easily see that the Test dataset error in the first one is much higher than the Second one.

So if u compare these two scenarios side by side,

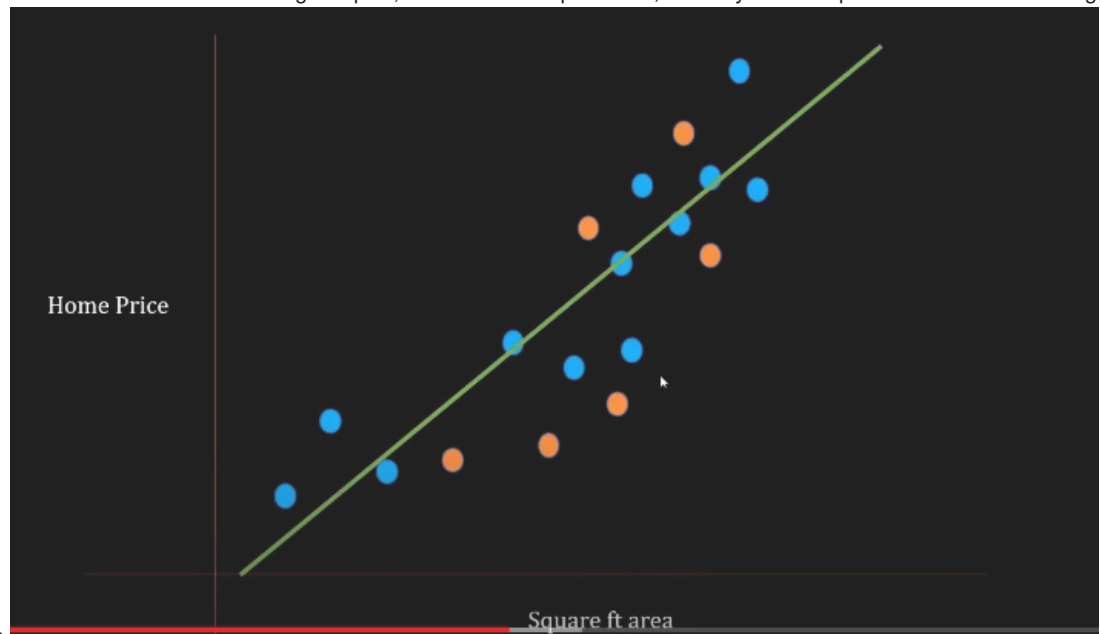


in one case ur Test error is 100 and in second its 27 which means that this Errors varies greatly based on your selection of Training dataset, and this is called **High Variance** cuz there is a high variability in the Test error based on the kind of Training sample u choose, like u are

selecting Training samples at random, so your Test error varies randomly which is not good and this is the Common issue with overfit model, They tend to have **High Variance**

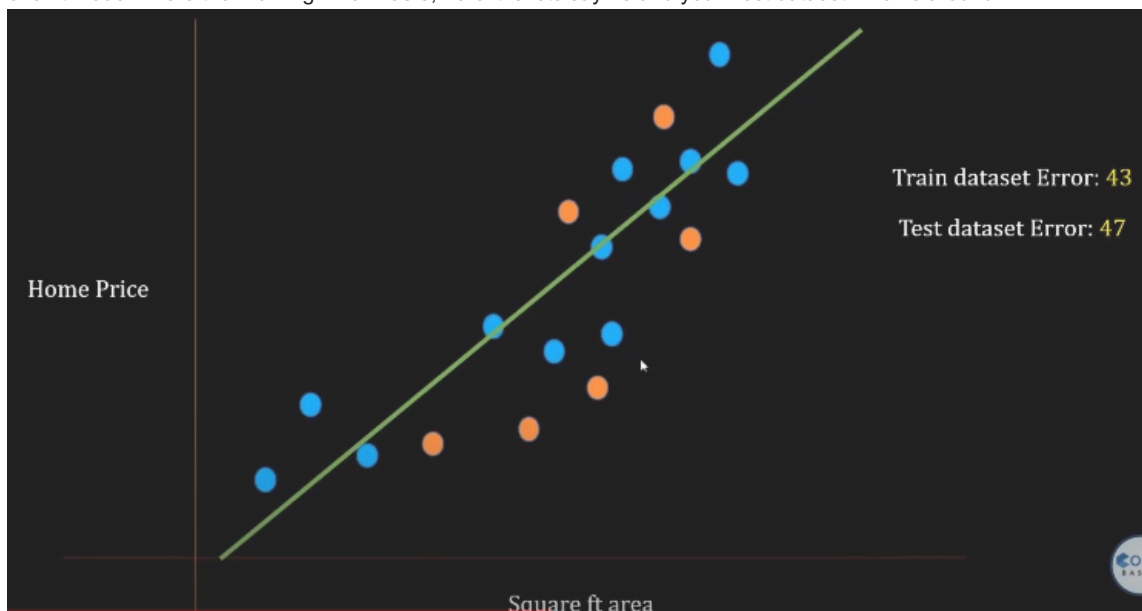


Now, let's look at another Scenario where we have the same dataset and we have split into Train and Test samples and this time instead of having a complex model that overfits our Training samples, We will do a simple model, Let's say Linear Equation which is underfitting



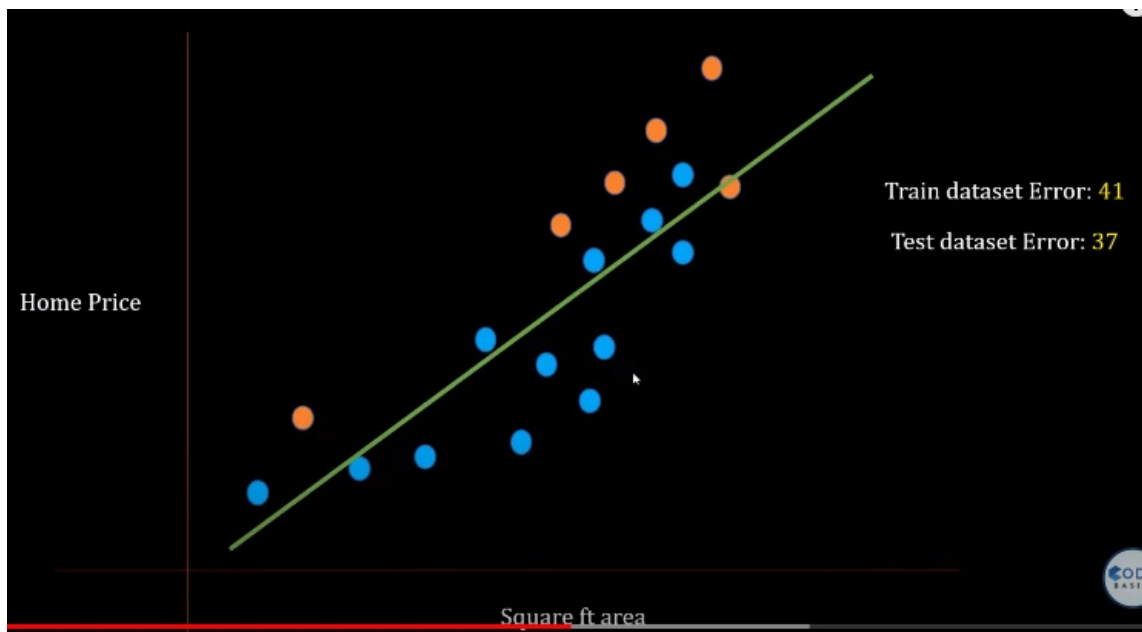
our Training sample

Because Linear Equation is a straight equation, it can't truly capture the pattern in your Training samples hence, it's a straight line so it can't pass through all the blue dots. So it's a simple model where your Training Error is actually high unlike the previous case in the overfit model where the Training Error was 0, here it is let's say 43 and your Test dataset Error is around 47

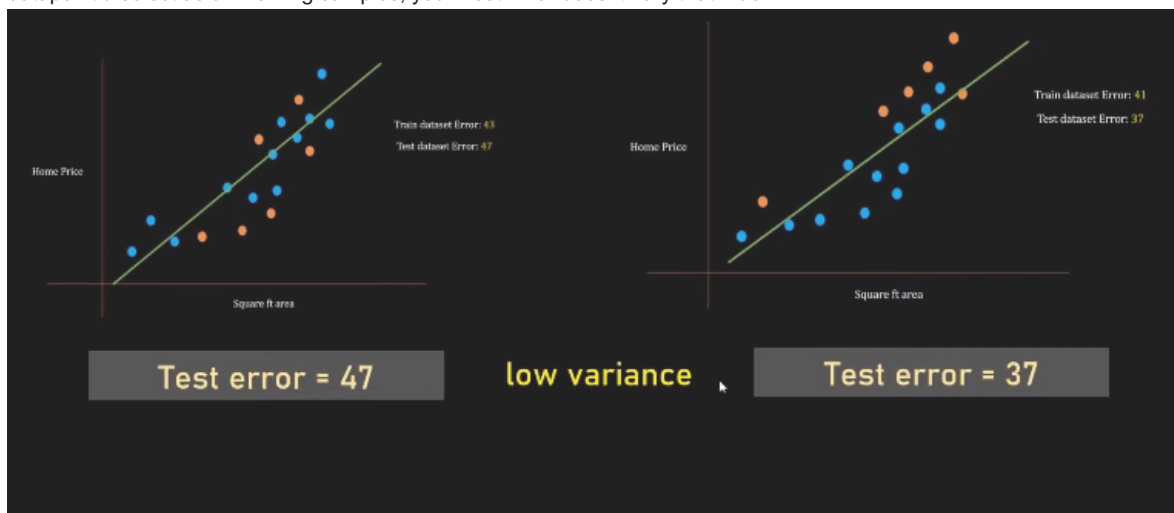


select a different set of Training datapoints, you can see below

Now, when you

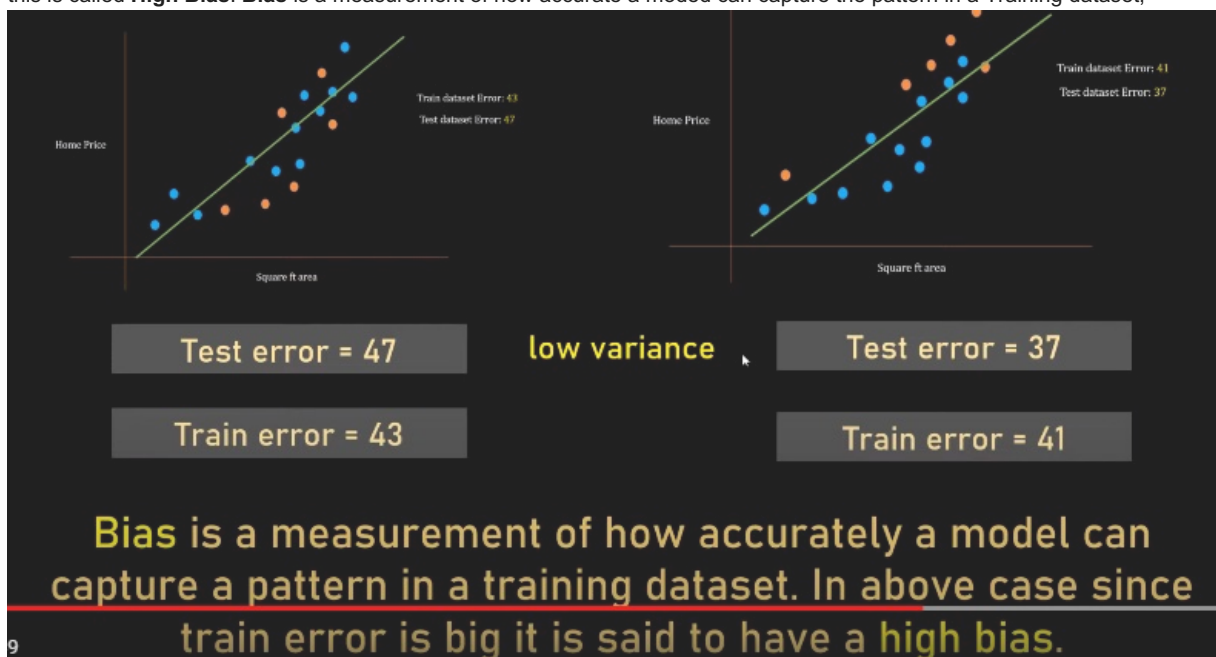


In both the cases, the line will be different but your Training and Testing dataset Error is still kind of the same, like in the first case the Test error was 47 and in the second it is 37, so its not very different, the gap is not too high, hence, this is called **Low Variance** cuz based on what datapoint u select as ur Training samples, your Test Error doesnt vary that much



Bias

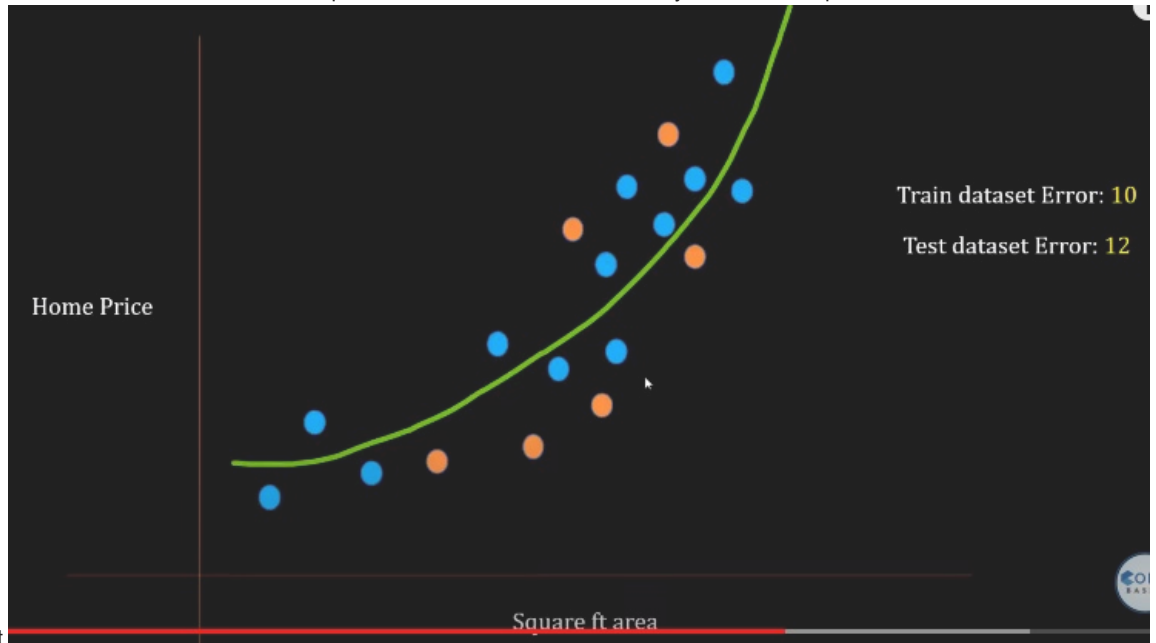
But now, if you look at your Train Error, previously in the overfit model, the Train Error was 0, now here we have some big Train Error and this is called **High Bias**. **Bias** is a measurement of how accurate a model can capture the pattern in a Training dataset,



So when ur thinking about **Bias**, you are always thinking of Train Error and when you are thinking of **Variance** you are always thinking of Test Error.

Previously when we overfit the model, the **Bias** was low cuz our Train Error was close to 0, So **higher the Train Error, Higher the Bias**

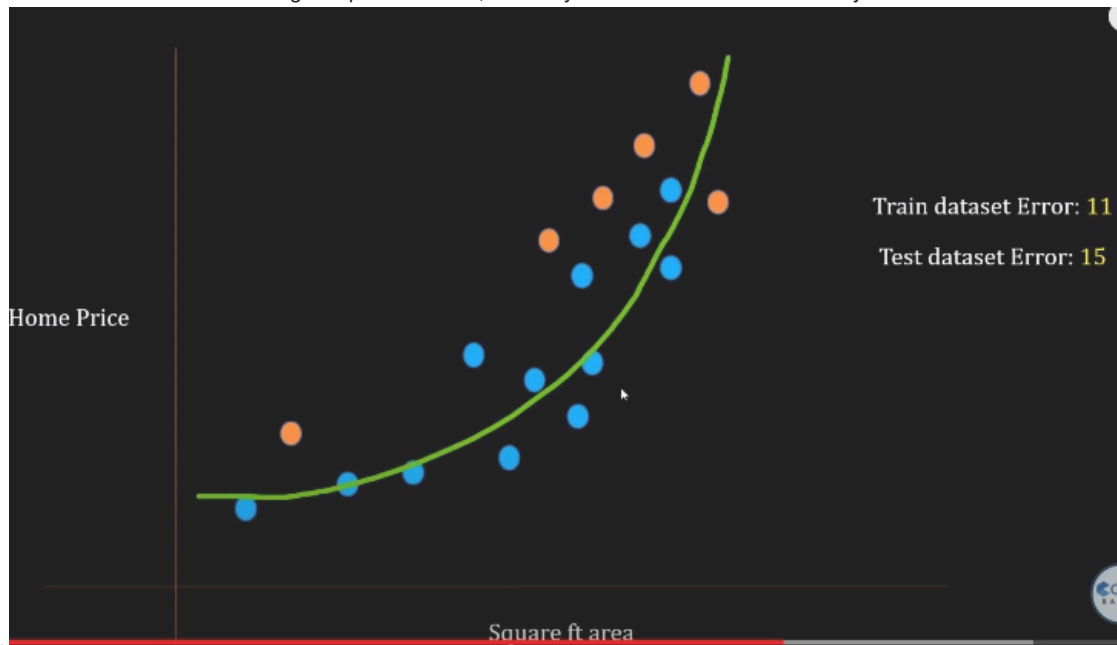
Now let's look at an ideal scenario where we come up with a model that kind of accurately describes the pattern that is available in the



Training dataset

Above, our Training Error and Test Error both are low!

Even if you select different set of Training samples as below, but still your model select is such that your Train and Test Error both are



kind of low

And in

this case, it's called **Low Variance** and **Low Bias** model cuz our Test error doesn't vary too much based on what Training samples we select. Also our Train error in general is low, hence it's called **Low Bias**



1. So whenever you have an Overfit model, it's likely that you will get **High Variance**

2. When u have Underfit model, its likely that u will get **High Bias**
3. And when u have balanced fit, u will get **Low Variance, Low Bias**

As a Data Scientist, u want to come up with a model that has a Balanced fit

Bull's Eye Diagram

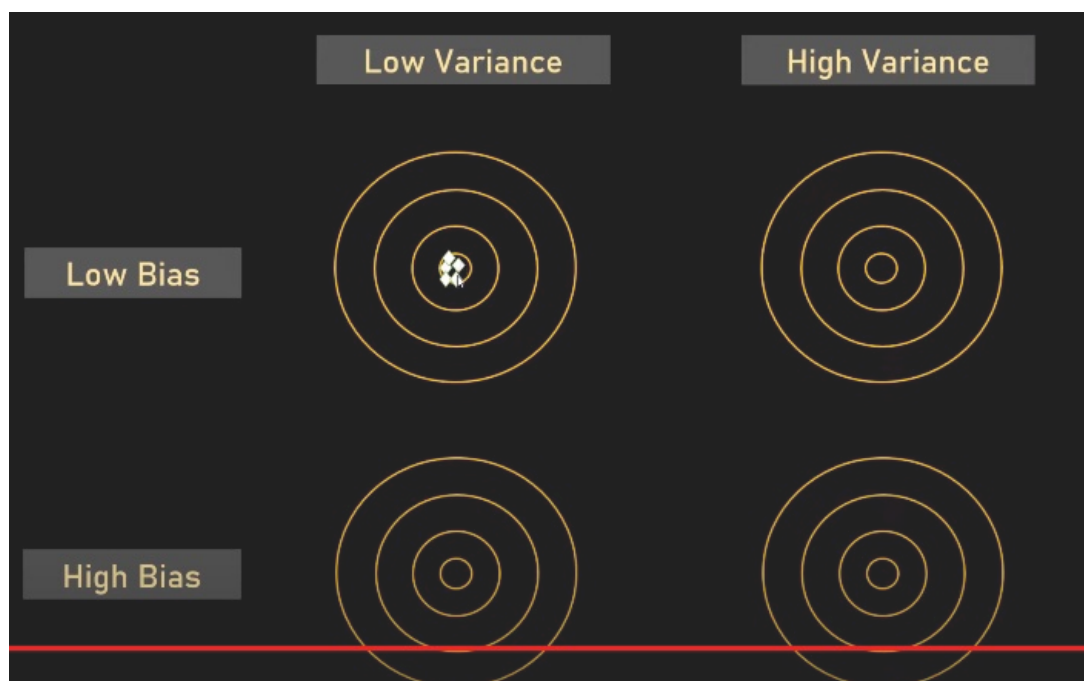
Now, there is a **Bull's eye diagram** as shown below, where the Central most circle represent the Truth, so the smallest middle most circle is your Truth sample



And when ur

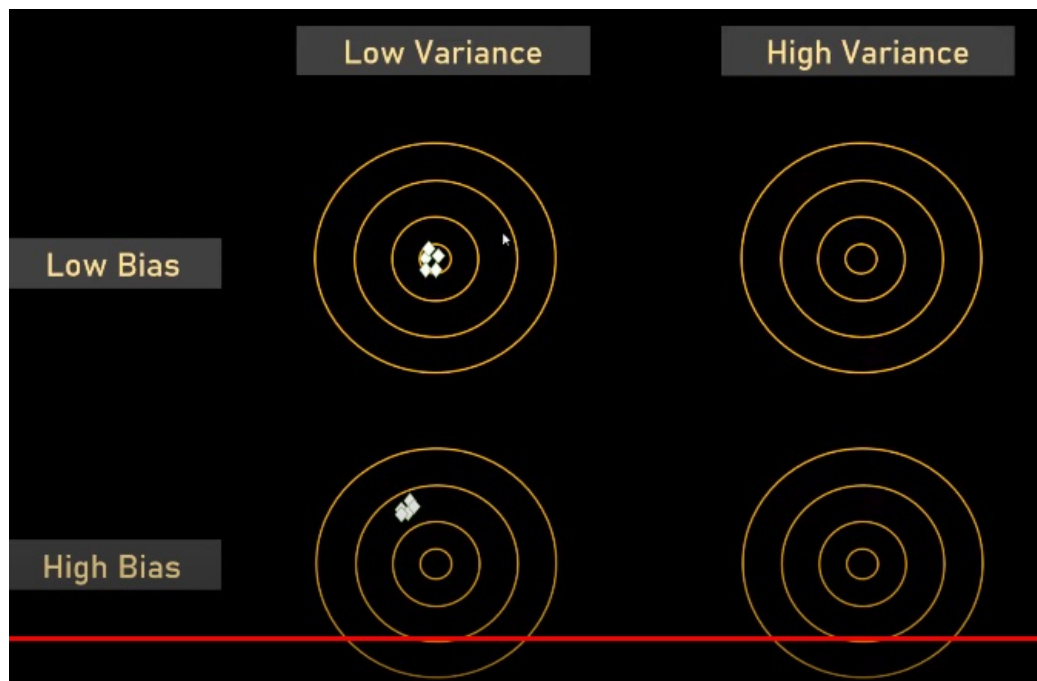
predicted values, these white diamonds are your predicted values,

1. When they are near to the inner most circle, which is the Truth, it is called **Low Bias**. **Bias** is all about how close u are to the Truth, so if ur closer to the inner most circle u have **Low Bias**

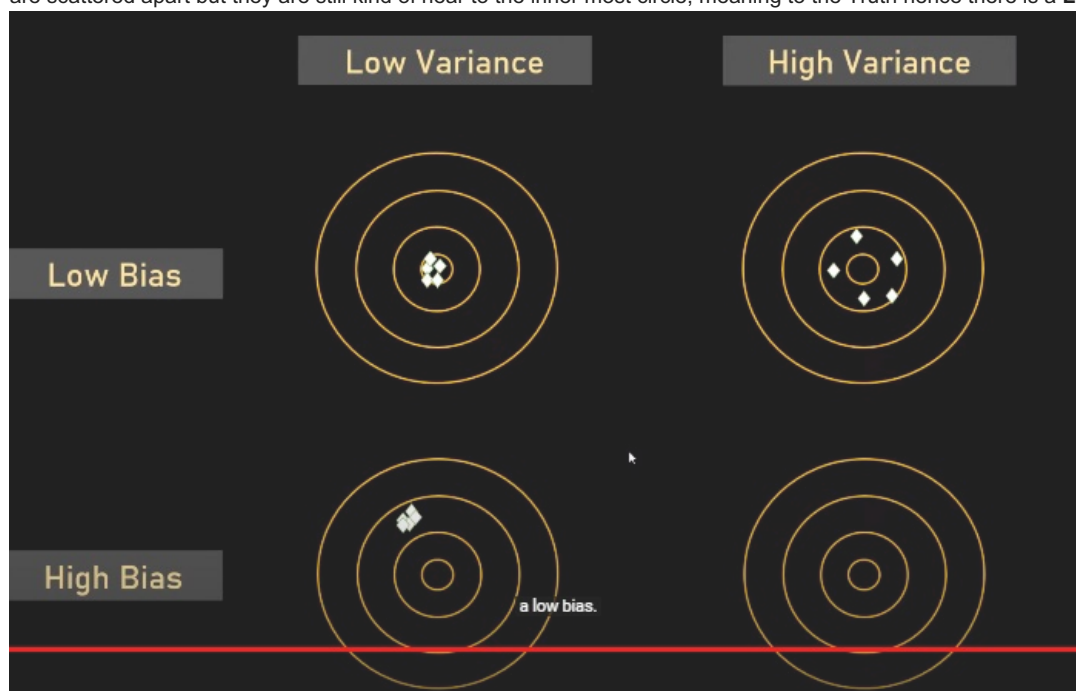


2. If you are far away

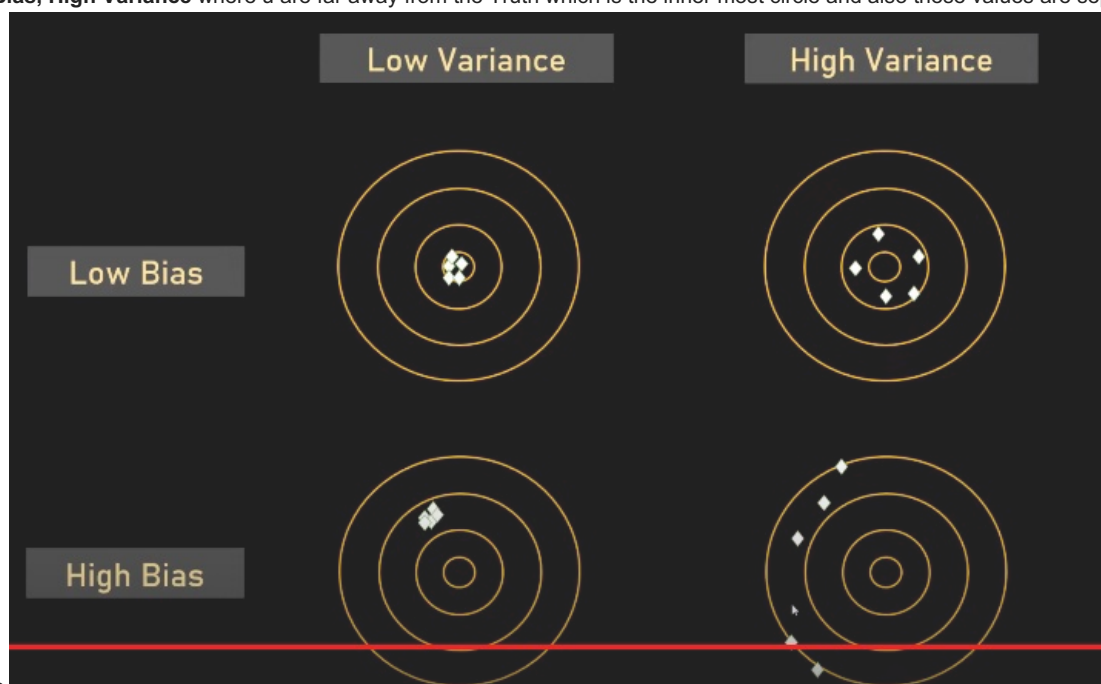
from the inner most circle, you have **High Bias** but in both 1 and 2 cases, these diamonds are clustered together, means they are close together, so whenever they are close together its called **Low Variance**



3. In **High Variance** they are scattered apart but they are still kind of near to the inner most circle, meaning to the Truth hence there is a **Low Bias**



4. Worst case scenario is **High Bias, High Variance** where u are far away from the Truth which is the inner most circle and also these values are separated out



too much


WAYS TO GET BALANCED FIT MODEL

There are several techniques, the first one is

1. **Cross Validation**
2. **Regularization**
3. **Dimensionality Reduction**
4. **Ensemble Techniques**

Ways to get balanced fit model

Cross Validation



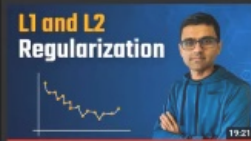
K Fold Cross validation

Machine Learning Tutorial Python 12 - K Fold Cross Validation
155K views • 2 years ago
codebasics

Many times we get in a dilemma of which machine learning model should we use for a given

25:20

Regularization



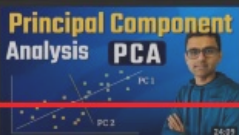
L1 and L2 Regularization

Machine Learning Tutorial Python - 17: L1 and L2 Regularization Regression
53K views • 10 months ago
codebasics

In this python machine learning tutorial for beginners we will look into, 1) What is overfitting

19:21

Dimensionality Reduction



Principal Component Analysis PCA

Machine Learning Tutorial Python - 19: Principal Component Analysis Python Code
7.9K views • 6 days ago
codebasics

PCA or principal component analysis is a dimensionality reduction technique that can help us reduce

14:41

Ensemble Techniques

Bagging, Boosting

Loading [MathJax]/extensions/Safe.js