Bias-Variance Trade Off

Overfitting versus Underfitting

Overfitting and Underfitting

We just saw that a higher order polynomial model performed significantly better than a standard linear regression model.

But how can we choose the optimal degree for the polynomial?

Overfitting and Underfitting

In general, increasing model complexity in order to improve performance leads to a **Bias-Variance trade-off**.

We want to have a model that can generalize well to new unseen data.

Extreme bias (Underfit) or extreme variance (Overfit) both lead to bad models.

Overfitting and Underfitting

- **Overfitting is when the model fits too much to the noise or variance from the data.**
- **This often results in low error on training sets, but high error on test or validation sets. This is why overfitting can sometimes be a little hard to catch because you may think your model is performing really well when in fact it's only performing well on the training set instead of performing well to unseen data.**
- **So it's overfitting, meaning it has too much variance.**

Overfitting and Underfitting

Data

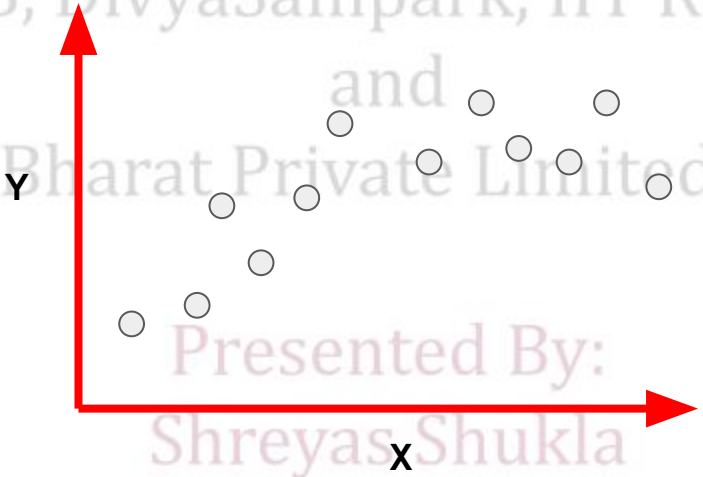
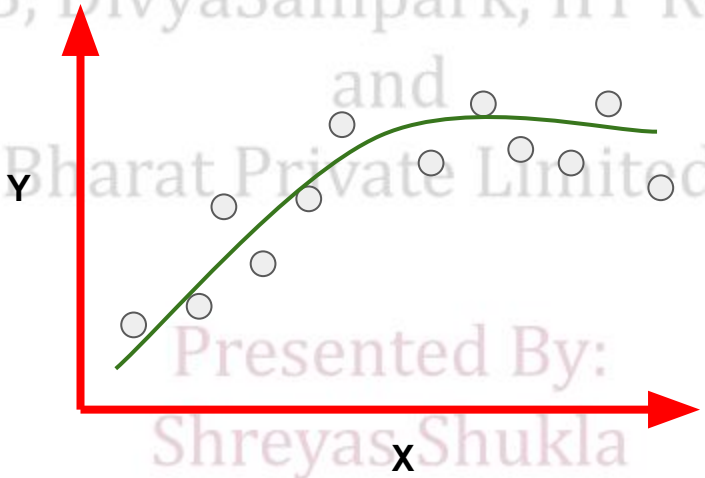


Fig Ref: Pierian Data

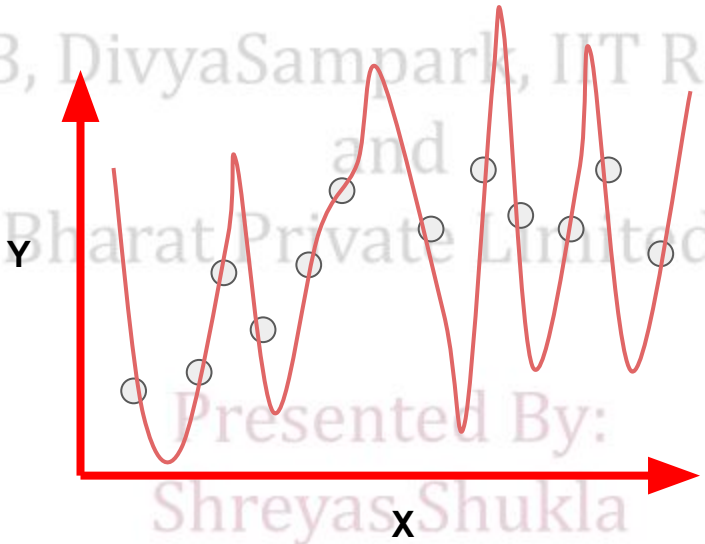
Overfitting and Underfitting

Good Model



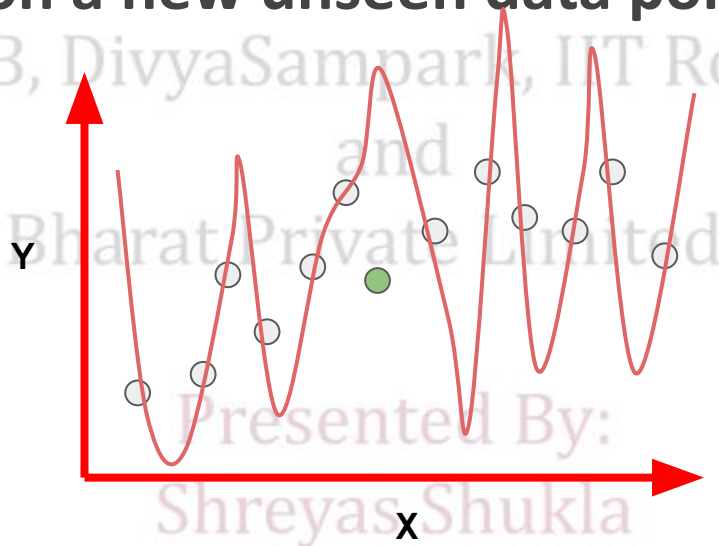
Overfitting and Underfitting

Overfitting



Overfitting and Underfitting

But what about on a new unseen data point?



Overfitting and Underfitting

Overfitting can cause large test errors!

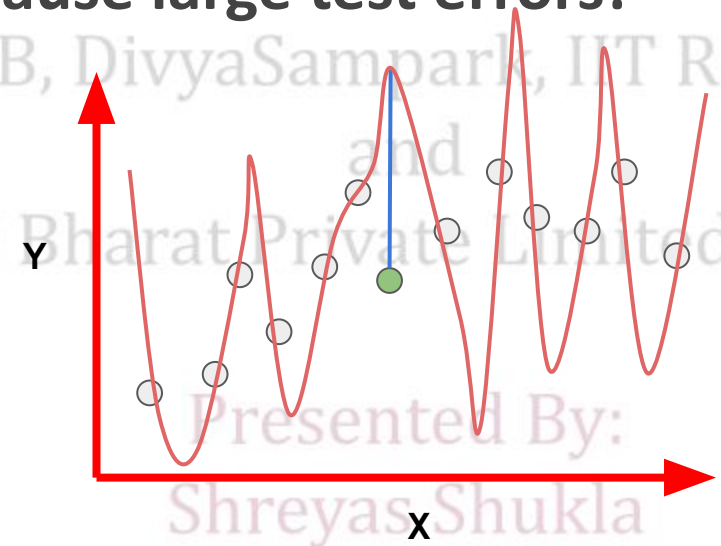


Fig Ref: Pierian Data

Overfitting and Underfitting

Overfitting

- Model is fitting too much to noise and variance in the training data.
- Model will perform very well on training data, but have poor performance on new unseen data.

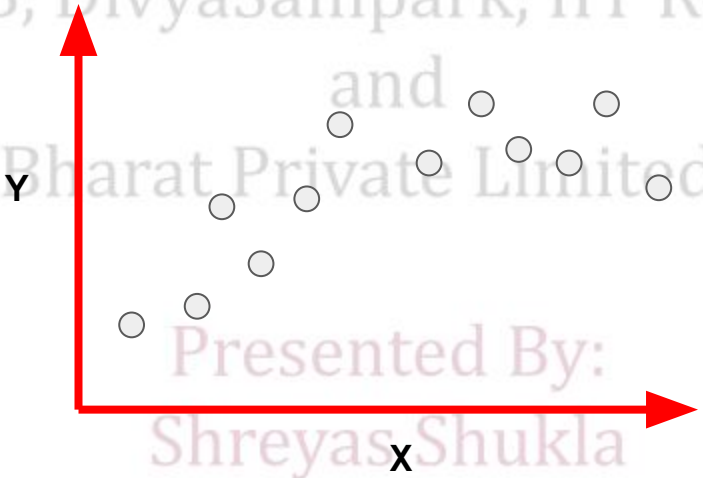
Fig Ref: Pierian Data

Overfitting and Underfitting

- **Underfitting**
  - Model does not capture the underlying trend of the data and does not fit the data well enough.
  - Low variance but high bias.

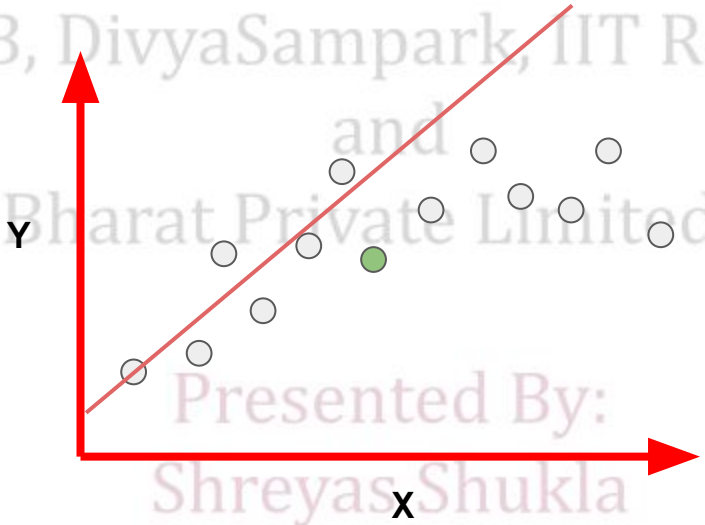
Overfitting and Underfitting

Data



Overfitting and Underfitting

Underfitting



Overfitting and Underfitting

Underfitting

- Model has high bias and is generalizing too much.
- Underfitting can lead to poor performance in both training and testing data sets.

Overfitting and Underfitting

Overfitting versus Underfitting

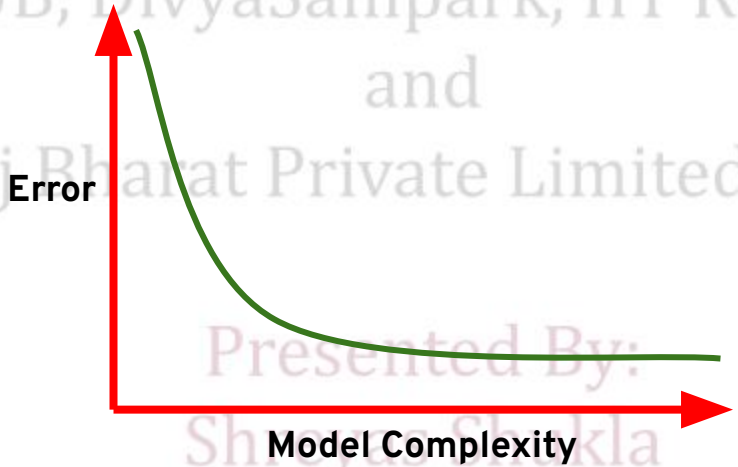
- Overfitting can be harder to detect, since good performance on training data could lead to a model that appears to be performing well.

Overfitting and Underfitting

How can we see underfitting and overfitting when dealing with multi dimensional data sets?

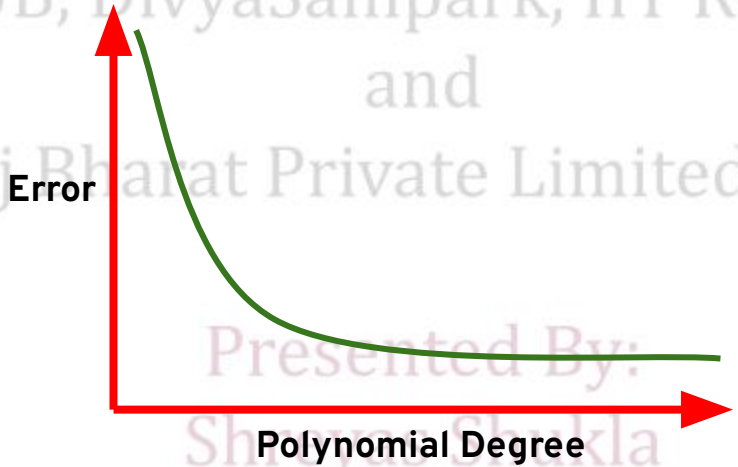
Overfitting and Underfitting

Good Model



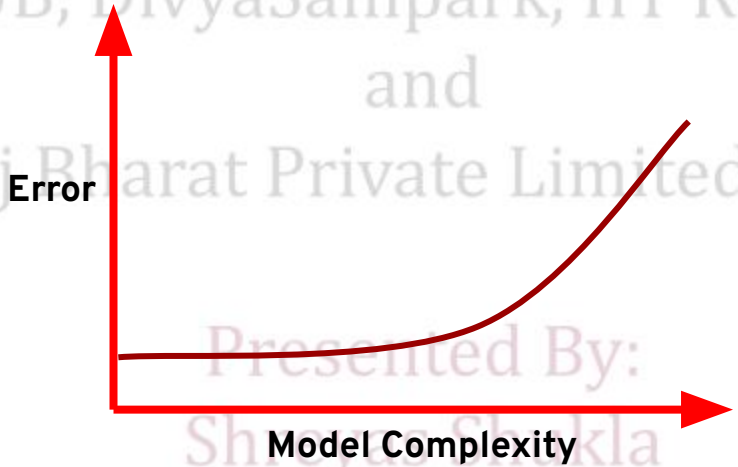
Overfitting and Underfitting

Good Model



Overfitting and Underfitting

Bad Model



Overfitting and Underfitting

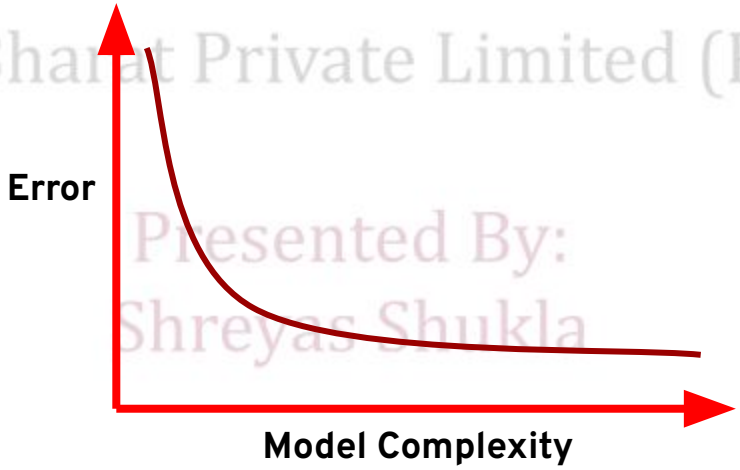
keep in mind the relationship of model performance on the training set versus the test set.

Fig Ref: Pierian Data

Overfitting and Underfitting

Imagine we split our data into a training set and a test set.

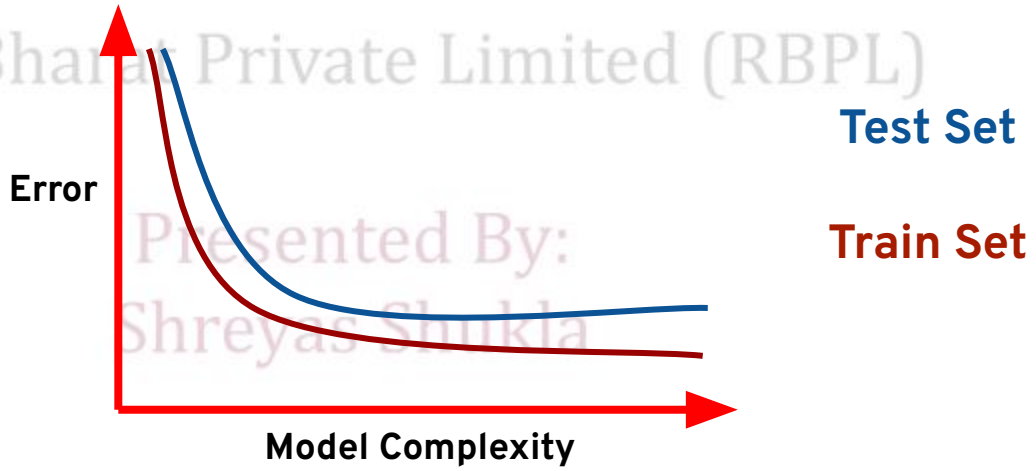
We first see performance on the training set



Overfitting and Underfitting

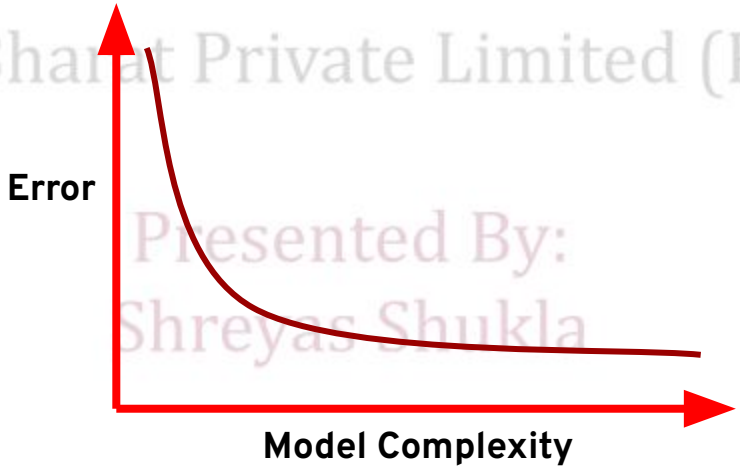
Check performance on the test set

Ideally the model would perform well and behave similar on both



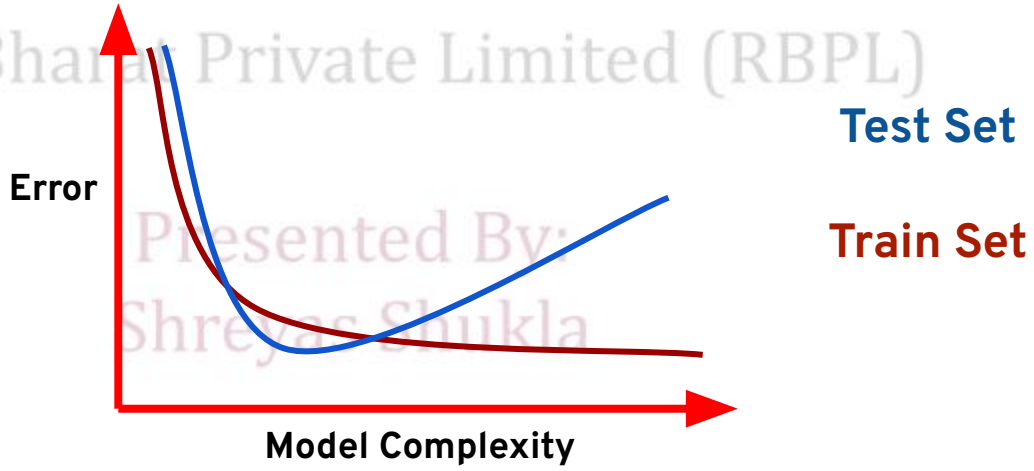
Overfitting and Underfitting

If we overfit on training data? That means we would perform poorly on new test data!



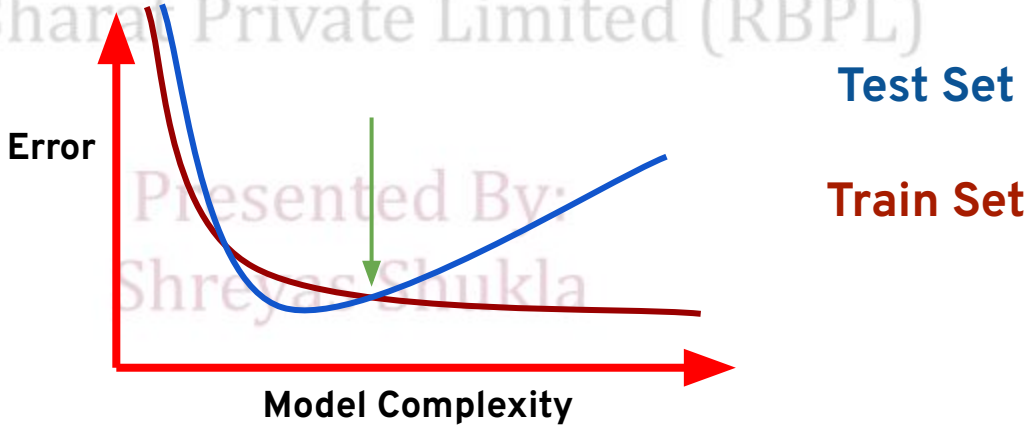
Overfitting and Underfitting

Check for overfitting



Overfitting and Underfitting

This is a good indication too much complexity.



Overfitting and Underfitting

When deciding optimal model complexity **and** wanting to fairly evaluate our model's performance, we can consider both the train error and test error to select an ideal complexity.

In the case of Polynomial Regression, **complexity directly relates to degree of the polynomial**

Overfitting and Underfitting

That's it in Overfitting and Underfitting!