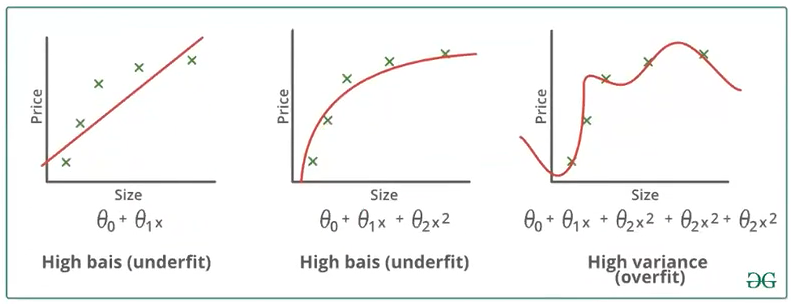
Regularization in Machine Learning

While training a machine learning model, the model can easily be overfitted or under fitted. To avoid this, we use regularization in machine learning to properly fit a model onto our test set. Regularization techniques help reduce the chance of overfitting and underfitting and help us get an optimal model.

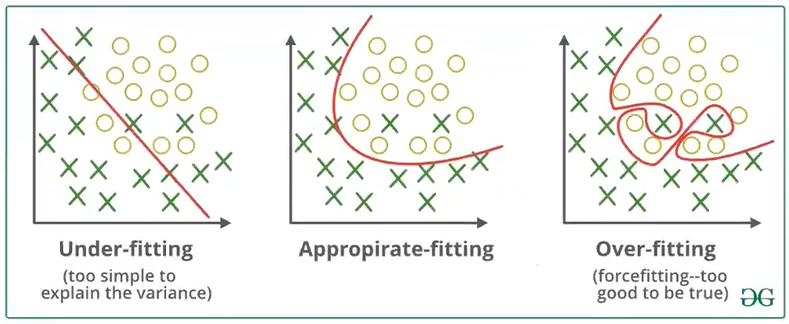
**What Are Overfitting and Underfitting?**

To train our machine learning model, we give it some data to learn from. The process of plotting a series of data points and drawing the best fit line to understand the relationship between the variables is called Data Fitting. Our model is the best fit when it can find all necessary patterns in our data and avoid the random data points and unnecessary patterns called Noise.

A scenario where a machine learning model can neither learn the relationship between variables in the testing data nor predict or classify a new data point is called Underfitting.



A scenario where the machine learning model tries to learn from the details along with the noise in the data and tries to fit each data point on the curve is called Overfitting.



**Bias and Variance**

A Bias occurs when an algorithm has limited flexibility to learn from data. Such models pay very little attention to the training data and oversimplify the model therefore the validation error or prediction error and training error follow similar trends. Such models always lead to a high error on training and test data. High Bias causes underfitting in our model.

Variance defines the algorithm's sensitivity to specific sets of data. A model with a high variance pays a lot of attention to training data and does not generalize therefore the validation error or prediction error are far apart from each other. Such models usually perform very well on training data but have high error rates on test data. High Variance causes overfitting in our model.

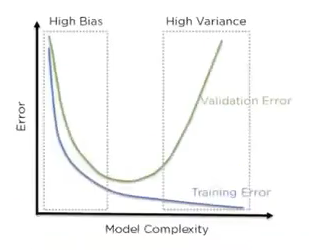
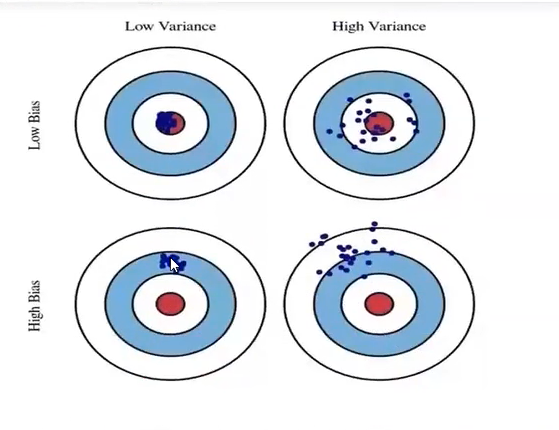
An optimal model is one in which the model is sensitive to the pattern in our model, but at the same time can generalize to new data. This happens when Bias and Variance are both optimal. We call this Bias-Variance Tradeoff and we can achieve it in over or under fitted models by using Regression.

Figure 3: Error in testing and training datasets with high bias and variance



**Regularization**

Regularization refers to techniques that are used to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting or underfitting. Using Regularization, we can fit our machine learning model appropriately on a given test set and hence reduce the errors in it.

There are two main types of regularization techniques:

Ridge Regularization and Lasso Regularization.

**MultiCollinearity with Variance Inflation Factor**

Colinearity is the state where two variables are highly correlated and contain similiar information about the variance within a given dataset.

Multicolinearity on the other hand is more troublesome to detect because it emerges when three or more variables, which are highly correlated, are included within a model.

A common R function used for testing regression assumptions and specifically multicolinearity is "VIF()" and unlike many statistical concepts, its formula is straightforward:

**V. I. F. = 1/(1 - R²).**

The Variance Inflation Factor (VIF) is a measure of colinearity among predictor variables within a multiple regression. It is a measure for the increase of the variance of the parameter estimates if an additional variable in a model.It is calculated by taking the the ratio of the variance of all a given model's betas divide by the variane of a single beta if it were fit alone.taset

One recommendation is that if VIF is greater than 5, then the explanatory variable given by exog\_idx is highly collinear with the other explanatory variables, and the parameter estimates will have large standard errors because of this.

This makes the effects of X1 on Y difficult to distinguish from the effects of X2 on Y.