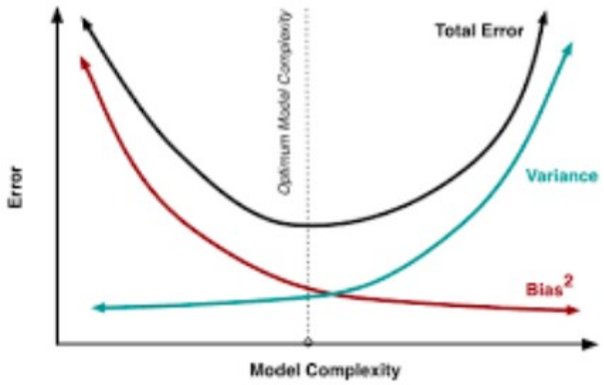
| | |
|---|---|
| **Bias-Variance Trade-off** | The error emerging from any model can be broken down into components mathematically.<br>$$Err(x) = \left(E[\hat{f}\,(x)] - f(x)\right)^2 + E\left[\hat{f}\,(x) - E[\hat{f}\,(x)]\right]^2 + \sigma_e^2$$<br>$$Err(x) = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$<br>Following are these component :<br>1. **Bias error** is useful to quantify how much on an average are the predicted values different from the actual value<br>2. **Variance** on the other side quantifies how are the prediction made on same observation different from each other<br>  <br>A high bias error means we have a under-performing model which keeps on missing important trends. A high variance model will over-fit on your training population and perform badly on any observation beyond training. In order to have a perfect fit in the model, the bias and variance should be balanced which is bias variance trade off. |
| **Big Data** | Big data is a term that describes the large volume of data – both structured and unstructured. But it's not the amount of data that's important. It's how organizations use this large amount of data to generate insights. Companies use various tools, techniques and resources to make sense of this data to derive effective business strategies. |
| **Binary Variable** | Binary variables are those variables which can have only two unique values. For example, a variable "Smoking Habit" can contain only two values like "Yes" and "No". |

| | |
|---|---|
| **Binomial Distribution** | Binomial Distribution is applied only on discrete random variables. It is a method of calculating probabilities for experiments having fixed number of trials.<br><br>Binomial distribution has following properties:<br><br>1. The experiment should have finite number of trials<br>2. There should be two outcomes in a trial: success and failure<br>3. Trials are independent<br>4. Probability of success (p) remains constant<br><br>For a distribution to qualifying as binomial, all of the properties must be satisfied.<br><br>So, which kind of distributions would be considered binomial? Let's answer it using few examples:<br><br>1. Suppose, you need to find the probability of scoring bull's eye on a dart. Can it be called as binomial distribution? No, because the number of trials isn't fixed. I could hit the bull's eye on the 1st attempt or 3rd attempt or I might not be able to hit it at all. Therefore, trials aren't fixed.<br>2. A football match can have resulted in 3 ways: Win, Lose or Draw. Thus, if we are asked to find the probability of winning in this case, binomial distribution cannot be used because there are more than two outcomes.<br>3. Tossing a fair coin 20 times is a case of binomial distribution as here we have finite number of trials 20 with only two outcomes "Head" or "Tail". These trials are independent and probability of success is 1/2 across all trials.<br><br>The formula to calculate probability using Binomial Distribution is:<br><br>$$P ( X = r ) = nCr (p^r)* (1-p) * (n-r)$$<br><br>where:<br>n : No. of trials<br>r : No. of success<br>p : the probability of success<br>1 – p : Probability of failure<br>nCr : binomial coefficient given by n!/k!(n-k)! |
| **Boosting** | Boosting is a sequential process, where each subsequent model attempts to correct the errors of the previous model. The succeeding models are dependent on the previous model. Some of the boosting algorithms are:<br><br>• AdaBoost<br>• GBM<br>• XGBM<br>• LightGBM<br>• CatBoost |