1. **If the linear regression coefficient of a predictor is 0.54 then what does it mean?**

Regression coefficients represent the mean change in the response variable for one unit of change in the predictor variable, with constant other predictor. Coefficient of predictor isolates the role of one variable from all of the others in model. This is also called as slope coefficients.

If we will take the sales quantity and revenue in a graph, and x-axis holds the sales quantity. Then one unit increase or decrease in sales quantity the revenue will increase or decrease by .54 units.

**2. How would you deal a data with Target class imbalance problem?**

Class imbalance problem occurs when total number of +ve class of data is far less than the total number if –ve class of data. This problem is very common and can be observed in disciplines like fraud detection, anomaly detection, medical diagnosis, etc.

So this is a problem because the error rate is very less and the model learns to predict the false cases as true as most of the results are true.

We can take various approach to get rid of this problem. For that we have apply various model on the data like:

* Cost function based approach
* Sampling based approach
* Generate Synthetic samples
* Various algorithms (Tree based)

The best resulting algorithm can solve the class imbalance problem.

Exp:

If we have to detect the fraud transaction in an ecommerce web. Then out of 1000 there will be 10 fraud transactions, in this case model will learn to mark most of the transactions as good instead of fraud which will lead to great loss. In this case we have to understand the accuracy in True negative part.

**3. You have built a classification model with 90% accuracy but your client is not happy because False Positive rate was very high then what will you do?**

False positive rate was higher that means the model is biased towards positive class and so there are many –ve points classified as positive. To overcome from this situation we can increase –ve points in the training data or use over sampling to balance the dataset.

We can use following procedures to gain more accuracy:

* Missing Value analysis
* Feature engineering
* Normalization and Standardization.
* Ensambling
* Synthesizing more data
* Cross validation.

Exp:

Let’s take an example of disease detection model. According to my data the model is 90% correct but sometimes the model detected a disease in patient which is in real it’s not in him/her. So that creates a huge problem. To avoid that if we will follow the above steps this issue may decrease.

**4. Does multicollinearity effects in Naïve Bayes? If yes/no then why?**

The Naïve Bayes algorithm makes assumption that feature used are independent. Chi squire selects the most independent features out of all features. Correlation doesn’t give more information about the variable is independent or not. So Naïve Bayes performs poorly when there is multicollinearity.

**5. If we do not define number of trees to be built in random forest then how many trees random forest internally creates?**

There is no specific number of trees RF algorithm will create by default. Because the next tree RM will create is dependent upon the error observed in previous tree. If the error is same then it will stop creating trees, else it will create more trees to reduce the error.