# **Assignment Name - Machine Learning Basics Problem Statement**

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

***Answer: -*** If the predictor of a Linear Regression is 0.54, means for every additional unit of the predictor, the value of **dependent variable** will increase by 0.54 unit.

We know that p-value of each term tests the null hypothesis that the coefficient is equal to zero or not. A low p-value indicates that we can reject the null hypothesis. Conversely, a larger p-value suggests that changes in the predictor are not associated with changes in the dependent variable.   
Suppose from a Data Model we received p-value for three variables A, B and C respectively 0.0,0.9 and 0.1. In this case we are supposed to eliminate the variable B, because of its high p-value (generally consider >0.5 as high p-value). Generally, we consider low p-value to be statistically significant.

Basically, we obtain p-value to decide which terms to keep and which needs to be removed from model.

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

***Answer: -*** While working with data model where the number of observations belonging to one class is significantly lower than those belonging to the other classes, is lead towards Target Class Imbalance Problem. In this situation, the predictive model developed using conventional machine learning algorithms could be biased and inaccurate. This happens because Machine Learning Algorithms are usually designed to improve accuracy by reducing the error.

We try to improve data pre-processing technique with classification algorithm to the training data before applying typical Machine Learning.

We use the K-means clustering algorithm to identify clusters in the dataset. So that each cluster is oversampled such that all clusters of the same class have an equal number of instances and all classes have the same size.

Suppose we have,

Total Observations = 100

False Observations =20

True Observations = 80

Event Rate= 4 %

*Majority Class Clusters:*

* Cluster 1: 50 Observations
* Cluster 2: 20 Observations
* Cluster 3: 12 observations

*Minority Class Clusters:*

* Cluster 1: 8 Observations
* Cluster 2: 10 Observations

After oversampling of each cluster, all clusters of the same class contain the same number of observations.

*Majority Class Clusters:*

* Cluster 1: 20 Observations
* Cluster 2: 20 Observations
* Cluster 3: 20 observations

*Minority   Class Clusters*:

* Cluster 1: 20 Observations
* Cluster 2: 20 Observations

Event Rate post cluster based oversampling sampling = 40/ (100+40) = 28.5 %

This is the clustering technique which helps overcome Imbalance problem of dependent class by decreasing the event rate.

1. ***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?***

***Answer: -*** A false positive rate is the probability of falsely rejecting the null hypothesis for a test. It considers incorrect positive prediction where as accuracy is being measured by True positive and true negative divided by total observations. Hence the accuracy value is impure as False positive rate is high. So, we should try to increase specificity.   
Here we have received accuracy as 90% which might imply that there is data imbalance in predictor class i.e. the number of one class is much higher than other class/classes. In that case we should go for other confusion matrics.

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

***Answer: -***Yes, multicollinearity effects Naive Bayes because Naive Bayes assumes that variables are independent. Multicollinearity occurs when there are high correlations among predictor variables, leading to unreliable and unstable estimates of regression coefficients. It is measured by Variance Inflected Factor i.e. 1/(1-*R2*). Where R2 is the linear regression predictor values.

Naive Bayes is estimated under the assumption that predictors are conditionally independent given the target variable. As a result, relationships between the dependent variable and the independent variable are estimated in isolation without paying attention to multicollinearity between the predictors. The Naïve Bayes is therefore not able to approximate the multivariate regression function, if multicollinearity exist it won’t give accurate analysis than a univariate analysis.

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

***Answer: -*** If we do not define the number of trees to be created for a Random Forest, it generally creates 10 trees.