**X Education – Lead Scoring Case Study**

First, we import the required libraries and then we read the csv file. Then we check the data frame shape and then check the null values in the given data set. If the column contains the null values, then do the data cleaning. If the column contains more than 30% of null values then we have to take a look at column whether we need to remove it or not. After data cleaning and visualization, we prepare the data for model building. First, we check the numerical variables and check the distribution of the variables using boxplot. For very high outliers we remove the outliers in the column. And for categorical variables we convert the binary variables (Yes/No) into 1/0. For categorical variables with multiple levels, we create dummy features. Now we split the data set into test train data set with 70% train set and 30% test set using test\_train\_split algorithm from sklearn. Then we use feature scaling on numerical data in order for machine learning models to interpret these features on the same scale. Here we used Standardization method from StandardScaler.

X = (x -mean(x))/sd(x)

In this data set we standardized the 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit'. Now we start building our first model using LogisticRegression algorithm from sklearn. Now we use RFE (Recursive Feature Elimination) for selecting the features automatically. And now we build our model again. Now we use manual feature elimination based on p-values and VIF (Variance Inflation Factor). After repeating this model until all our features meet the conditions (p-value < 0.05, VIF value < 2). And we use confusion metrics to calculate the accuracy, sensitivity, specificity of our model. Now we do the model evaluation using ROC curve. ROC curve area determines the goodness of the model higher the curve area the better the model is. Actually, we have use 0.5 as a cutoff for classification of the data set for using predicted y values. So, now we determine the optimal cutoff point for our model using two methods confusion metrics with different cutoff values or Precision and Recall trade off

Or both. Now we get the optimal cutoff point. Now we make prediction on the test set. And after making prediction we again use confusion metrics for evaluating the model accuracy, sensitivity, specificity.

From this assignment, I have learned about:

1. How to build a logistic regression on a real-world data.
2. Making a presentation for Business Understanding.
3. EDA (Exploratory Data Analysis) on a real-world data.