**Summary:**

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

The following are the steps used:

### Importing Data: Loaded the data into jupyter notebook

### Inspecting the Dataframe

### Data Preparation: Performed missing value treatment for all the columns and outlier treatment on numeric columns after dropping the columns which had more than 50% missing value. Few of the null values were changed to ‘not available’ so as to not lose much data. We dropped high data imbalanced columns.

### EDA: A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seems good and no outliers were found.

1. Dummy Variables: The dummy variables were created and later on the dummies with ‘not available’ elements were removed.

### Test-Train Split: Performed train, test split using sklearn library. The split was done at 70% and 30% for train and test data respectively.

### Feature Scaling: Scaled the train data using Standard Scaler

### Looking at Correlations: The correlations was not very clear so we could not use it much.

### Model Building: Used statsmodels GLM method to perform Logistic Regression.

### Feature Selection Using RFE: RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept). Performed 2 iterations after dropping high p-value features one by one.

### Plotting the ROC Curve : A confusion matrix was made. Based on this we created the confusion matrix and checked for specificity (77%) and sensitivity (83%). Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each. Our model covered 88% area under the curve.

### Finding Optimal Cutoff Point

### Making predictions on the test set: Prediction was done on the test data frame and with an optimum cut off as 0.3 with accuracy, sensitivity and specificity of 75%.

1. Precision – Recall: This method was also used to recheck and a cut off of 0.41 was found with Precision around 79% and recall around 65% on the test data frame.
2. Calculating the lead score
3. We determined that our model is working equally well on unseen data
4. Then we created lead conversion score = (conversion probability \* 100) to give a score between 0 to 100 where higher the value means the lead is “hot” and there is high possibility that the lead can be converted.

There are many learning we gathered from this assignment.

These are as below:

1. How to handle missing value and outliers in a data set.

2. How to create dummy variables/labels on categorical columns.

3. How to use python libraries to perform logistic regression on selected features.

4. How to choose best model based on balanced sensitivity and specificity.

5. Finally, how to solve a problem with team effort.