SUMMARY

This analysis is done for the company X Education and to find ways to get more industry professionals to join their courses. The dataset provided gave us the information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate. The existing system has utmost 30% Conversion rate and our target is to identify HOT LEADS with conversion rate of 80% through our model.

We had followed the following steps for modelling:

1. **Data Understanding and Cleaning:** We have dropped certain columns like Prospect ID and Lead Number as they are just for identification purpose. Few columns have data with 'select' which is replaced by Nan. Checked the columns with missing values of over 45% and dropped those parameters. In the Column Country we imputed Nan with 'India' and so on.

2. **Exploratory Data Analysis:** We tried to visualize the data for the target variable on different plots and identified the outliers and cleaned them either by dropping the columns or imputing the values. Specialization with Management in them has higher number of leads as well as leads converted. Working professionals going for the courses have high number of chances to convert compared to other occupations. Unemployed leads are the most in terms of absolute numbers. Certain tags like 'Will revert after reading the email', 'Not Specified', 'Lost to EINS', 'Busy' and 'Closely by Horizon' are named with 'Other tags' for ease of analysis as their count is nominal.

3. **Numerical attribute Analysis:** We have checked the correlation off numeric variables such as 'Total visits' , 'Total time spent on website' , 'Page views per visit' etc and removed the outliers if present. Median for converted and not converted leads is the close. Nothing conclusive can be said on the basis of Total Visits Leads spending more time on the website are more likely to be converted. Website should be made more engaging to make leads spend more time.

4. **Dummy Variables and Test-Train Splitting:** Dummy variables are created and Yes is assigned with 1 and No with 0. Original Columns are dropped after the dummies are created. The split was done at 70% and 30% for train and test data respectively. Scaled numeric columns by 'StandardScaler'.

5. **Model Building and Evaluation:** Three different models are built using stats model and RFE. RFE was done to get top 28 variables. Out of the three models we adopted Model 3 as the p values are all low and VIF inline. A confusion matrix was made. The ROC curve has a value of 0.97, which is very good. Accuracy : 92.29% Sensitivity : 91.70% Specificity : 92.66%

Observations:

Observations on the Test Data after the model is executed
Accuracy: 92.78%
Sensitivity: 91.98%
Specificity: 93.26%

Final Observation:
Train data observations vs Test data observations

	Accuracy	Specificity	Sensitivity
Train Data	92.29%	92.66 %	91.70%
Test Data	92.78%	93.26%	91.98%

The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model

Tags, Total Time spent on website and Lead score are the variables which contributed the most in finding the potential leads(Hot Leads)