# 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 lots of information about how the potential leads visit the site, the time spent there and their conversion rate.**

**The following steps are followed for EDA and Model building:**

1) **Cleaning Data:**

After looking at the dataframe closely following things have been carried out encoding categorical variables, missing value handling, convert the Select into Nan and then Dropping columns that are having 70 % null value. Handling the 'Select' level that is present in many of the categorical variables. We observe that there are 'Select' values in many columns. It may be because the customer did not select any option from the list, hence it shows 'Select'. 'Select' values are as good as NULL. So we can convert these values to null values.

2) **EDA:**

We have retained 98% of the row after cleaning the data. Then we had performed EDA. We had no duplicated data. Taking the help of seaborn we have plotted various countplots and taken inferences out of it. Also, we have handled the outliers with a cap of 95% for the analysis.

3) **Dummy Variables:**

The dummy variables were created for categorical features. Then concatenated the dummy data with lead data.

4) **Train-Test Split:**

The split was done 70% and 30% for the train and test data respectively. Also we have scaled the data using standardscaler.

5) **Model Building:**

Feature selection was done using RFE with 20 variables as output. Then rest of the variable were removed based on the VIF values and p-values (VIF<5 and p-values <0.05 were kept). Model 9 was our final model with 12 variables.

6) **Model Evaluation:**

A confusion matrix was made. The optimum cut-off value(using ROC curve) was used to find the accuracy, specificity and sensitivity , which came around to be 80%.

**7) Prediction:**

Prediction was done on test data frame and with an optimum cut-off value of 0.34 with the accuracy, specificity and sensitivity of 80%.

**8) Precision and Recall:**

This method also used to recheck and cut-off of 0.41(on curve) was with Precision 79% and Recall 70% on the train data-frame.

**9) Comparing the values obtained for Train & Test:**

a) Train Data:
# Accuracy : 81.0 %
# Sensitivity : 81.7 %
# Specificity : 80.6 %

b) Test Data:
# Accuracy : 80.4 %
# Sensitivity : 80.4 %
# Specificity : 80.5 %

Thus we have achieved our goal of getting a ballpark of the target lead conversion rate to be around 80% .
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 to get a higher lead conversion rate of 80%.
The customers which should be contacted are the customers whose "Lead Score" is equal to or greater than 85.
They can be termed as 'Hot Leads'.