**SUMMARY**

This analysis is done for X Education to find ways to get most promising leads to join their courses i.e. the leads that are most likely to convert into paying customers. The basic data provided gave us a lot of information about how the potential leads visit the site, the time they spend there, how they reached the site and the conversion rate. There were a lot of leads in initial stage of analysis but only few of them come out as paying customers. The following are the steps used in the analysis:

1. **Data Cleaning & Preparation:**

The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to ‘Nan’ so as to not lose much data. Although they were later removed while making dummies.

1. **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 seem good and some outliers were found. The dummy variables were created and later on the dummies with ‘Select’ elements were removed. For numeric values we used the MinMaxScaler.

1. **Test – Train Split & Scaling:**

The split was done at 70% and 30% for train and test data respectively.

1. **Model Building:**
2. **Feature Elimination based on correlation**
3. **Checked RFE** to attain the top 10 relevant variables.
4. **Manual fine tuning using p-values & VIFs-**

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)

1. **Model Evaluation:**
2. A confusion matrix was made. Later on we determined threshold, the optimum cut off value (using ROC curve) was used to find the **Accuracy 77%, Sensitivity 68% and Specificity 83%** .
3. **Precision & Recall** –

This method was also used to recheck and a cut off of 0.4 was found with Precision around 77% and recall around **75%** on the test data frame.

1. **Prediction on the test set:** Prediction was done on the test data frame and with an optimum cut off as 0.4

with accuracy 77%, sensitivity 68 %and specificity of 83%.

The leads having current occupation as working professionals or unemployed were having more convert ratio. The most important variables were 'Total Visits' , 'Total Time Spent on Website' , 'Page Views Per Visit' which contribute most towards probability of a lead getting converted. Maintaining a list of interested leads and keep them posted with latest courses and offers, seminars, holding QA session for interested candidates, and Monitoring each lead carefully to tailor the information you send to them. Make further inquiries and appointments with leads to determine their intention to join online courses.