**Summary**

This analysis is done for X education to find ways to get more industry professional to join their courses. From the data provided we get a lot of information about the potential leads and which variables play important role to convert the leads to paid customer. Although X education gets a lot of leads, only few of them gets converted. Conversion ratio is only 30%. Our goal is to make this conversion rate higher.

Step followed in this process.

1. **Reading and basic inspection of the data:** We import the data and and understanding the data shape, missing value etc.
2. **Cleaning the data:** After basic understanding of data we check for missing values in the data frame. There are some columns with more than 40% of missing values or Nan values. We drop them. For the columns with missing values less than 40% we impute them with most common values. By this process we get a data frame with zero missing value.
3. **EDA:** In this step we divide the columns into categorical columns and numerical columns. In some numerical columns we found outliers and removed them. For some categorical column we combined values to get a compact result and some categorical columns are dropped as they were not providing much information for further analysis.
4. **Dummy variables**: In this step we created dummy variables for categorical columns.
5. **Train - Test split:** We split the data in train and test set in ration of 7:3. We built a model using the train set and evaluate the result using test set.
6. **Feature Scaling:**  We used the StandardScaler to scale the numerical values.
7. **Model building:** We built a logical regression model using rfe. With the help of rfe we get the 15 most relevant variables. Now base on p value< 0.05 and VIF < 2 ,we created the final model by Removing those variables which do not satisfy the condition.
8. **Model Evaluation:**  Taking a initial assumption that a probability of more than 0.05 means 1 else 0 ,we created the data frame having the converted data values .Based on the initial assumption we derived confusion matices. Now we calculate overall accuracy, sensitivity and specificity of the model and evaluate the result using test set to reliability of the model.
9. **Plotting the ROC curve**: We plotted roc curve and get the are under the curve approximately 96% ,which is pretty fair.
10. **Prediction:** The optimum cut off point set at 0.3
11. **Decision**: The following are the findings.

* The more time leads spent on website the higher chances of it getting converted.
* Top three lead source are : Google,olark chat,direct traffic.
* The working professionals has higher conversion rate and unemployed generates high number of leads.
* There is high chance of conversion if last activity is sms and olark chat.