Problem Statement:

- An education company named X Education markets and sells its courses on several websites and search engines. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their contact details, they are classified to be a lead. Lead can also be got through past referrals
- Once these leads are acquired, employees from the sales team contacts them through various mediums, so the leads get converted.
- The typical lead conversion rate at X education is around 30%.
- X Education wants us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

Objective of the assignment –

We have to prepare a logistic regression model to predict whether a lead will convert or not, here the target lead conversion rate need to be around 80%.

Input Data Sets – Leads .csv

Solution –

- First we have to read, understand and clean the dataset.

  In this,1st task we did is to replace**"Select"** with nan(Null value), post this we have Dropped the columns with higher missing values (40%), then we have imputed missing values of columns types object with the mode and numerical with mean respectively . After this we checked highly skewed columns and dropped them, more than 15 columns we dropped in this step and columns with values that are in very low percentage in occurrence and those values can be merged and mapped to **Others (**Lead Origin, Lead Source, Last Activity, What is your current occupation,Last Notable Activity). After that, we checked for outlier of numerical columns and capped var TotalVisits. Finally we created dummy variables for (Lead Origin, Lead Source, Last Activity, What is your current occupation, City, Last Notable Activity, A free copy of Mastering The Interview). After this we plotted heatmap to reove highly correlated variables.

  We then divided the the dataset into Train & Test.

- Modeliing
  In this step first we scaled all the values with StandardScaler technique. Then We build the model with RFE , we are now left with 18 columns . We did manual feature removal method by checking p-value and VIF values , We stopped at 13 clos left here we have all p-values below 0 and VIFs below 5.
- Making Predictions
  Here we checked all the parameters like accuracy, sensitivity and specificity, confusion matix for the dataset for prob 0.5 first . Then we plotted ROC curve and calculated accuracy sensitivity and specificity for various probability cutoffs.Finally we came to conculsion to use optimal probability cutoff of 0.3. With this prob we calculated Assign Lead Score by simply multiplying the prdectited probability with 100. We got sensitivity as 84.25 and specificity as 76.67 for train dataset.
  Then we Predicting on test data after scaling the continous variables and got sensitivity as 83 and specificity as 76 for train dataset.

Final conclusions of the model-

- Top 5 variables that contribute the most in this model are:
  1. Lead Origin_Others - Lead Origin of type Other
  2. What is your current occupation-Working professionals
  3. Last Activity_SMS sent - Leads to whom SMS is sent
  4. Lead Source_Olark Chat
  5. Total Time Spent on Website.

- Final Suggestions from the model:
- ➢ Call leads whose origin are from Lead Add Form, Lead Import or Quick Add Form.
- ➢ Leads to whom SMS was sent and are Working Professional must be chosen.
- ➢ Consider leads whose source is from Olark Chat and have highest time spent on website.
- ➢ Employees may focus on sending email as it is seen leads who are opening their emails and reading the offer are also contributing highly in conversion.
- ➢ Making UI/UX of the website will also help in the converting the leads as this improves the trust of users on the course and the company.
- ➢ Optimizing SEO also helps heavily in conversion rate. It is seen that leads that are coming from Google are also getting converted into customer at a higher rate as it is considered that these kinds of users they know what they are looking for and more often or not these users also know the company they want buy from.