Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. 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 email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert

The goal is:

1. To build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
2. To adjust to if the company's requirement changes in the future so you will need to handle these as well.

Step1: Read and Understand Data

- The file data has got 9240 entries, 37 columns. out of which 4 variables are of float, 3 integers and 30 objects.

Step2: Data Cleaning

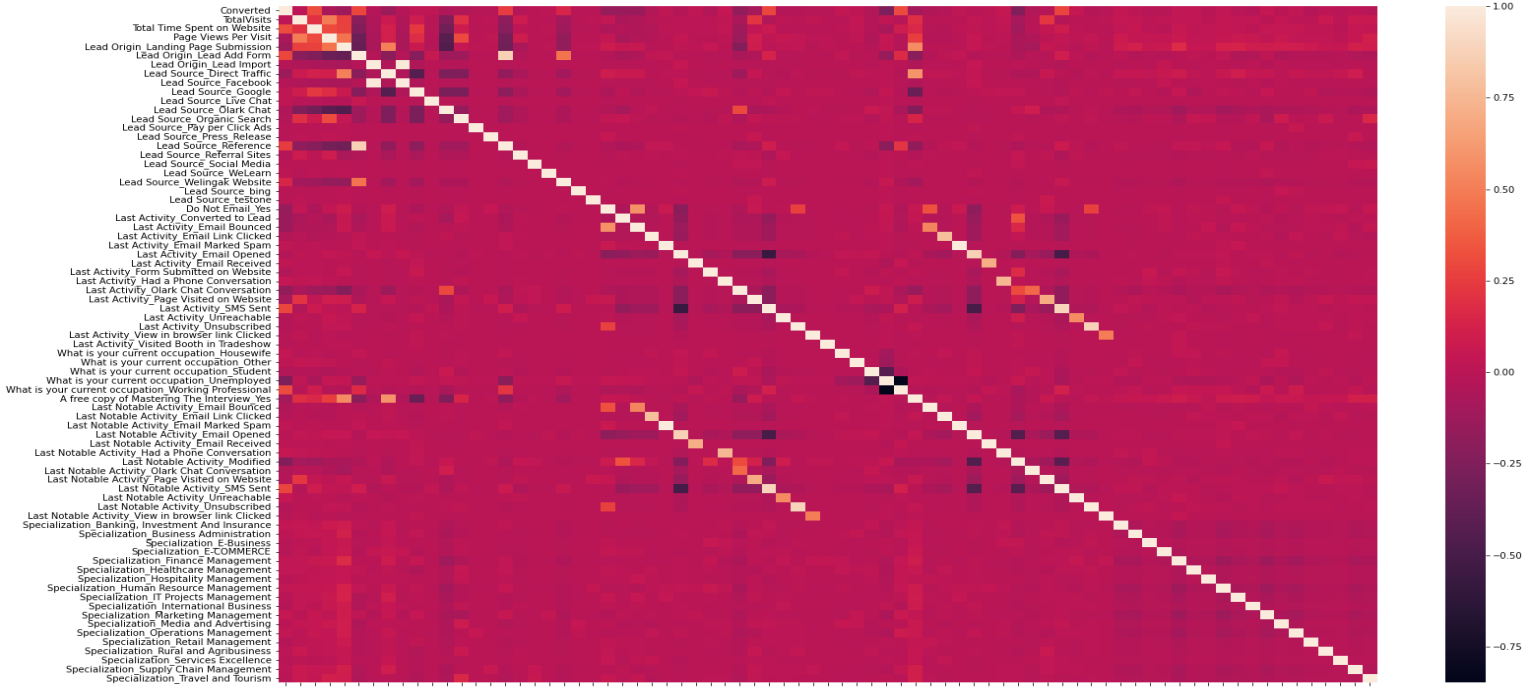
- **1. Missing Data Treatment**
- Dropping columns with greater than 35% missing values, we get the shape of the new dataframe as (9240, 31)
- **2. Dropping columns that are of no significance to analysis**
- the columns Lead Profile and How did you hear about X Education have a Select% greater than 40% which is of no use to the analysis so it's best that we drop them.
- when you got the value counts of all the columns, there were a few columns in which only one value was majorly present for all the data points. These include Do Not Call, Search, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, Receive More Updates About Our Courses, Update me on Supply Chain Content, Get updates on DM Content, I agree to pay the amount through cheque. Since practically all of the values for these variables are No, it's best that we drop these columns as they won't help with our analysis.
- What matters most to you in choosing a course has the level Better Career Prospects 6528 times while the other two levels appear once twice and once respectively. So we should drop this column as well.
- **New shape (9240, 14)**
- Dropping the missing rows from What is your current occupation, Lead Source, TotalVisits, Page Views Per Visit, Last Activity and Specialization, we are left with (6373, 14)
- Dropping Prospect ID and Lead Number won't be of any use in the analysis, so it's best that we drop these two variables.

Step3: Data Preparation

- **Treating the Categorical Data using one-hot encoding**
- 'Lead Origin', 'Lead Source', 'Do Not Email', 'Last Activity', 'Specialization', 'What is your current occupation', 'A free copy of Mastering The Interview', 'Last Notable Activity']
- **Scaling the Numeric features using min-max scaler**
- num_var = ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website']

Correlation of Independent Variables with Target

- The correlation heatmap is not well readable as the number of features are too big.



Step4: Model Building

- After importing necessary model for Logistic Regression, and doing the test train split in the Data Preparation phase, we have directly used Recursive Feature Elimination Technique and no of features selected are 15 out of 74.
- As we are interested in the Statistical Inferences of p-value and VIF to check the effectiveness of the parameters and multicollinearity among the variables, we have proceeded with the Stats model.

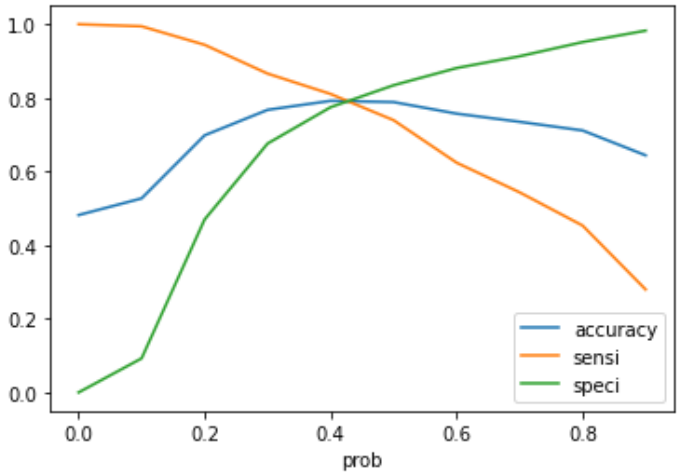
- **Inferences:**
- VIFs seem to be in a decent range except for three variables, Lead Origin_Lead Add Form, Lead Source_Reference and Lead Source_Welingak Website.
- But before treating VIFs, we would first drop the feature having p-value > 0.05 and high VIF >5, whhich is Lead Source_Reference

- The VIFs of all the features look good below 5.
- We have then proceeded to evaluate the features having higher p-values and drop them one at a step.
- The features dropped during the process are: Last Notable Activity_Had a Phone Conversation, What is your current occupation_Housewife, What is your current occupation_Working Professional.

Step5: Model Evaluation

- Using Confusion Matrix, at a random cut-off of 0.5,
- Accuracy on the Train set = 0.788
- sensitivity on the Train set = 0.739413680781759
- specificity on the Train set = 0.834

- Searching Optimum Cut-off using Sensitivity-Specificiy and Accuracy, 0.42 is found to be the best.



Step8: Making Predictions on the Test Set

- With threshold at 0.42, the train and test set seem to be showing a stable value for Accuracy, sensitivity and specificity with the values as below:.
- **Train Set parameters:**
- Accuracy: 0.7908
- sensitivity = 0.793392275476966
- specificity = 0.7884

- **Test Set with the same columns chosen in the train set and the optimal cutoff at 0.42**,
- Accuracy: 0.784
- sensitivity = 0.779
- specificity = 0.789

- **List of Final Parameters chosen by the model and their coefficients are:**

- const 0.204037
- TotalVisits: 11.148912
- Total Time Spent on Website: 4.422291
- Lead Origin_Lead Add Form: 4.205123
- Lead Source_Olark Chat: 1.452589
- Lead Source_Welingak Website: 2.152559
- Do Not Email_Yes: -1.503680
- Last Activity_Had a Phone Conversation: 2.755220
- Last Activity_SMS Sent: 1.185594
- What is your current occupation_Student: -2.357784
- What is your current occupation_Unemployed: -2.544455
- Last Notable Activity_Unreachable: 2.784594
- dtype: float64