SUMMARY REPORT

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. Although X Education gets a lot of leads, its lead conversion rate is very poor. X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not

Our Whole solving strategy involves the following steps.

- **Importing and cleaning data** - Before we jump into the actual model building, we first need to clean and prepare your data. We will import the dataset and do basic checks on our dataset, checking the dataset for the amount of nulls present. After checking the columns for the variance explained some columns needs to be dropped as it is explaining nearly no variance. Also we would drop the columns with more than 30% Nulls. Also there are certain columns which need to be imputed.

- **Removing outliers from data**- As we can see our data is heavily affected by the outliers so in the first step we have removed the outliers from the dataset then our primary goal is to deal with the categorical feature in this case we are using one hot encoding to create the dummy columns so that we can pass this created columns in our model.
  After removing the outliers we can see our data is not normally distributed which will be best for analysis.

- **Splitting the data into train and test**- We split our dataset into train and test dataset so that whatever model we build on train dataset we can test it on our test dataset.

- **Scaling the data**- We would next be scaling our data. We have applied MinMaxScaler our data will lie between 0 and 1.

- **Perform RFE & GLM to the test data**- Now we would proceed with Feature Selection using RFE. After 7^{th} iteration we can see we are getting descent P-values and our VIF values are also in control.

- **Selecting the valid cutoff**- The most important part is to check all the validation metrics which can tell you the performance of your linear model. Metrics that we have tested

includes Confusion Metrics, Sensitivity Specificity, False Positive Rate, Precision and Recall. Where as we can see a tradeoff between sensitivity and specificity we can easily identify the cutoff, For our model we have selected 0.3 as our cutoff.

- **Apply the learning to test dataset-** After applying all the leanings on the test data we can see though our accuracy has decreased but our main target is to increase the sensitivity .Our model is doing quite a good job to identifying 82% of the hot leads.