SUMMARY REPORT

X Education sells online courses to industry professionals. on any given day, many professionals Those who are interested in the courses visit their website and browse for the courses. Although x Education gets a lot of leads, it has a very poor lead conversion rate. X Education has appointed us Help them select the most promising leads, i.e. leads that are most likely to convert into payments Customer. This dataset includes various attributes like Lead Source, Total Time Spent on.

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 a basic check on our dataset, checking the dataset for the amount of nulls present. After checking the columns for variance explained some columns need to be dropped as it is explaining almost no variance. Also we will drop columns with more than 30% nulls. Apart from this there are some columns which need to be put in place.

- **Removing outliers from data**- As we can see our data is highly affected by 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 use one hot encoding to create dummy columns so that we can pass this created column to our model.
  After removing outliers we can see that our data is not normally distributed which would 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.