SUMMARY REPORT - Lead Scoring Case Study

Objective:

This Case Study was focused on solving a strategic problem for the organization X Education. This company tries to attract more customers pursue the online courses offered by them through various promotional channels like online advertisements, chats, promotions etc. Based on the activities of people who view these ads, the company approaches the customers who they sense potential targets for conversion into actual customers who would buy their course and pursue further. The key objective of this Case Study was to identify some key factors that would help us maximize the success rate and classify the Hot Leads accurately so that the company's conversion rate goes high.

Solution Approach:

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.
- **Exploratory Data Analysis(Univariate analysis):** Performed univariate analysis of categorical and numerical columns with respect to converted variable and visualise all the important features
- **Removing outliers from data**- As we can see our data is heavily affected by the outliers so in the first step we have capped 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 Standard-Scaler 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 7th iteration we can see we are getting descent P-values and our VIF values are also in control.
- **Selecting the valid Threshold**- 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. Whereas 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.

The prediction was made on the test set and we got the following result for train set and test set:

Train Set:

- Sensitivity: 84%
- Overall accuracy: 80%
- Specificity: 77%

Test Set:

- Overall accuracy: 81.5%
- Sensitivity: 77%
- Specificity: 85%