**Report/Summary Lead Score Case Study**

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.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. Around 30% of the leads get converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.

 We have been provided with a leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity. The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.

EDA:

We started by looking at the null values. Found that many columns had values as ‘Select’ which meant that the person has chosen not to fill that value. So these values are similar to null values. Columns with a high percentage of null values was dropped. For other columns like ‘Current Occupation’, ‘What matters most when choosing a course’ we imputed null values with mode.

Many columns like ‘Last Activity’, ‘Lead Source’ had values having low frequencies which were later clubbed as one category. This was done in order to avoid skewness. For the null values a new Category called ‘Others’ was created.

Outlier Treatment:

We removed the low percentile values for columns Page Views Per Visit, TotalVisits.

Dummy Variable Creation:

Labelled values for 'A free copy of Mastering The Interview', 'Do Not Email' to 0,1.

For other columns like ‘Specialization’, 'Lead Source', 'Tags' etc. we created dummy variables which could be fed to the machine learning model.

Model Building using Logistic Regression:

Divided the test and train data in the ratio 70:30 . Used Standard Scalar to scale numerical columns.

We then selected top 18 columns using RFE technique and created a machine leering model using GLM. VIF was computed for the columns in order to eliminate columns with high VIF.

A cutoff probability of 0.5 was chosen such that a probability higher than 0.5 ,meant converted and less than that meant not converted. A confusion matrix was drawn with accuracy 92.8%,

In order to find the accuracy of the test, ROC curve was drawn which shows a tradeoff between Sensitivity and Specificity. The curve was following left-hand border and then the top border of the ROC space thus the test came out as more accurate.

Also in order to find the optimal probability, accuracy, sensitivity, specificity was plotted for various probabilities. We selected 0.3 as the optimal probability from the graph.

Predictions on the test set:

The overall accuracy of the model on test set came out to be as 92.5%, Sensitivity: 91.1%, Specificity: 93.4%, Recall 91.1%