**Lead Score Case Study Summary Report**

**Objective:**

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.

A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted

**Data Explanation:**

Shape of data: There are 37 features explaining leads to conversion rate of a customer with 9240 customers data

Target Variable: Converted: 38% of them got converted as lead

**Data Pre-processing**:

Cleaning

The Variables with high null values which have more than or nearly to 30% of null share are choose to drop

Columns with low nulls, replaced numeric columns with 0 and removed the rows which have nulls for categorical columns

Dummy variable creation

After cleaning we are left with 19 columns, with which have created dummies for categorical columns with final of 56 columns left

Splitting

Split the data to train and test sets, train set to be 80% of whole data to train the model and test 20% to evaluate the train results

Scaling

Scaled the data using MinMax Scaler, as the dataset contains most of the categorical columns with 0’s ad 1’s and also this a classification problem

**Modelling:**

As there are 56 columns, we choose to reduce the columns for better modelling by use of RFE technique. Reduced to 15 columns as output (15 is a random number)

Built 3 models using GLM, and reduced dimensions to 11 as significant variables.

By the final logistic regression model, we are going to predict probabilities of each customer

**Predictions**

From the probabilities, By choosing a cut-off of 0.5(randomly chosen), we got the specifications as

* Accuracy of 79 % Sensitivity 66% and Specificity 87%

We have further tuned the model for finding an optimal cut-off by plotting accuracy sensitivity and specificity for various probabilities

* The optimal cut-off to be 0.35

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

For The conversion rate to be 80. i.e., Sensitivity to be 80%, decrease the cut-off of the model and found the cut-off to be 0.33

ROC:

We also checked ROC i.e, area under curve which shows a model performance which we got 86% with optimal cut-off which represents a good model

Test set Results:

For a optimal cut-off of 0.35 tested results on random test set got following specifications:

Accuracy of 72 % Sensitivity 66% and Specificity 87%

Lead Score:

To find the lead score in a way to communicate for non-technical persons, the probability scores are multiplied by 100 to achieve scores between 1 to 100