SUMMARY

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 wants its leads generation rate to be increased. For which they need to identify what are the potential leads aka Hot Leads.

For this company wants to create a model where we can assign lead score to each of the identified leads for higher chances of conversion. Target for the company is 80 percent.

Goals to Achieve as per the assignment

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.

Steps to follow to build a Model

1. Reading and Understanding the data
2. Data Cleaning
3. Data Visualization using EDA
4. Data Preparation for Modelling
5. Model Building
6. Model Evaluation

Reading and Understanding the data

We import all the necessary libraries for e.g. NumPy, pandas, matplotlib seaborn etc., import all the warnings.

We read the data and check the no. of rows and columns. We also, check if there are any missing/ null values or not. Afterwards, we see the statistical summary of the data.

Data Cleaning

1. We saw there were few columns with high percentage of null values, so we decided to drop those columns.
2. Few columns had null values but the columns were important for analysis so we replaced all null values with 'Not Provided'.
3. Few of the columns had values as 'Select' so we replaced it by 'NaN'
4. Few columns were having outliers and the treatment of outliers was performed.

Data Visualization using EDA

For data visualization we performed exploratory data analysis (EDA). We first did the univariate analysis of categorical and continuous data.

Moved ahead with the bi-variate analysis of categorical and continuous data.

We also found in country data most of the records were from India and few were from outside India and we classified as same.

We found correlations between variables by using different plots.

Data Preparation for Modelling

Data preparation for multiple linear regression involves handling the categorical variables first and then performing dummy encoding.

We then performed the train and test split using 70%-30% rule and then performed the scaling of variables. Since scaling of variables is an important step, we may have different variables of different scales. So, it's important to have everything on the same scale for the model to be easily interpretable.

Therefore, we used **MinMaxScaler** for the same.

Model Building

We follow the bottom up approach for this, i.e. we start by building the model with just one variable. Hence, the choice of variables becomes very crucial.

RFE was done to attain the top 15 relevant variables. First, we will look at the significance of variables and based on the significance.

We checked VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept). In our model for all the feature VIF < 5 and p-value < 0.05.

Model Evaluation

We first created confusion matrix to find the TP, TN, FP and FN.

We calculated the sensitivity and specificity.

We plotted the graph of sensitivity, accuracy and specificity for each level of probability. We found that 0.335 was the cut-off point.

We then used the cut-off point 0.335 to select the person and see if he would be converted or not then We again created the confusion matrix to calculate TP,TN,FP & FN calculated the sensitivity and specificity and got 0.81 and 0.78 respectively.

We performed the Precision – Recall method to recheck and a cut off of 0.42 was found with Precision around 75% and recall around 75% on the test data frame.

And finally the ROC curve for the Logistic Regression model was aligned towards the TPR which showed high sensitivity and thus is indicative of a good model'.

Finally, It was found that the variables that mattered the most in the potential buyers are:

1. Total Visits

2. Total time spent on Website

3. Lead origin is Lead Add Form