SUMMARY

Problem Statement:

Company X Education sells online courses to professionals. They get many leads daily, but their lead conversion rate is low (around 30%). They want to improve their conversion rate to around 80%.

To achieve this, they want to identify the most potential leads, called 'Hot Leads.' These are the leads with a higher chance of converting into paying customers. By focusing on these high-potential leads, the company aims to increase their lead conversion rate.

Your task is to build a lead scoring model that assigns a numerical score to each lead. The higher the score, the more likely the lead is to convert. The company will use this lead scoring system to prioritize their sales efforts and communication with potential leads.

Data:

Build a lead scoring model for 9000 past leads using attributes like Lead Source, Time Spent on Website, Total Visits, etc. The target variable is 'Converted' (1 for converted, 0 for not converted).

The goal is to prioritize communication with leads likely to convert. Handle 'Select' values in categorical variables as null.

Preprocess the data, select relevant features, choose a machine learning algorithm, and predict lead scores. Segment leads based on scores to identify potential leads.

Goals of the case study:

- 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.

- There are some more problems presented by the company which our model should be able to adjust to if the company's requirement changes in the future so we will need to handle these as well.

Steps to follow to build a model:

For the given problem statement, we performed following steps on a provided data during whole lead scoring case study to build a logistic regression model:

1) **Reading & understanding the data** : First of all we import all the necessary libraries such as NumPy, Pandas, Matplotlib, etc. and import all the warnings. Then we check the number of rows and columns after that we check if there are any missing/null values or not. Afterwards, we see the statistical summary of the data.

2) **Cleaning the data** : After reading and understanding the data we saw there were few columns with high percentage of null values, so we drop these columns. Some columns had null values, but the columns were important for analysis, so we replace all null values with 'not provided'. Also, few columns had values as 'Select' so we replaced it by 'NaN'. After that we performed treatment of outliers on few columns.

3) **Data Visualization using EDA**: After data cleaning we performed exploratory data analysis. First, we did the univariate analysis of categorical and continuous data. Then we checked the percentage of null values and dropped columns with more than 45% missing values. Then we moved ahead with the bi-variate analysis of categorical and continuous data. Then we found correlations between variables by using different plots.

4) **Preparation of data for modeling**: In this stage we handled categorical variables first and after that we performed dummy encoding. We then made train and test split using 70% - 30% rule and then performed the scaling of variables. For easy interpretation we convert all the values on the same scale for our model by using MinMaxScaler.

5) **Model Building**: In this step we follow the bottom up approach for this we start by building the model with just one variable. Hence, the choice of variables becomes very crucial. RFE was done to attend the top 30 relevant variables. First we looked at the variables based on the significance. After that we checked VIF values and p-values. In our model for all the feature VIF < 5 and p-value < 0.05.

6) **Model Evaluation**: In this we first created confusion matrix to find the TP, TN, FP and FN . We calculated sensitivity and specificity. For each level of probability we plotted the graph of sensitivity, accuracy and specificity. We got the cut-off value of 0.42. Then we used this cut-off value to select the person and see if he would be converted or not. We again created the confusion matrix to calculate TP, TN, FP and FN. So, we calculated the sensitivity and specificity.

After following these steps, we got following conclusions:

We plotted the graph of sensitivity, accuracy and specificity for each level of probability. We found that 0.42 was the cut-off point. Then we used the cut-off point 0.42 to select the person and see if he would be converted or not. After that we again created the confusion matrix to calculate TP, TN, FP, FN & calculated the sensitivity and specificity and got 0.76 and 0.85 respectively.

Evaluation:

❖ For Sensitivity- Specificity Evaluation:

1) On Training data set we found that optimum cutoff was 0.35 which gave us
    i. Accuracy 81.48%

ii. Sensitivity 70.56%
iii. Specificity 88.36%.

2) Prediction on Test data set:
   i. Accuracy 80.94%
   ii. Sensitivity 84.98%
   iii. Specificity 78.66%

- ❖ For Precision – Recall Evaluation:

3) On Training data set with the cutoff of 0.42 we get the Precision & Recall of 79.27% & 70.56% respectively which gave us

   i. Accuracy 81.31%
   ii. Precision 79.27%
   iii. Recall 70.56%

4) Prediction on Test data set :
   i. Accuracy 82.26%
   ii. Precision 74.84%
   iii. Recall 76.30

<u>Conclusion:</u>

- ❖ TOP VARIABLE CONTRIBUTING TO CONVERSION ARE:

  - ➢ **Last Activity_Resubscribed to emails**: The peoples which are come to resubscribe to our emails.
  - ➢ **Last Notable Activity_Resubscribed to emails**: Last activity performed by the customer which includes Email Opened, Olark Chat Conversation, etc. & decided to resubscribe our emails.
  - ➢ **TotalVisits**: The number of visits made on our website by customers.

Hence, the company should focus on targeting the customers which are more frequently visited to our website and the customers who are decided to resubscribe again our email service. Thus, our model seems to predict the Conversion Rate very well and we should be able to give the Company confidence in making good calls.