Lead Scoring Case Study
Summary

1. **Data Reading**: Importing libraries and the required dataset.

2. **Data Understanding**:
   a. Sanity Checks: No. of rows, columns, data type of each column. Numeric Summary (Mean. Median, Percentiles, etc.)
   b. Proper Duplicate Checks

3. **Data Cleaning**: In this case study, Data cleaning plays a very crucial role. The quality and efficiency of the model depends on the data cleaning step. Hence it must be followed thoroughly.

   a. Dropped the columns '**Prospect ID**' and '**Lead Number**' as it has unique values and are just indicative of the ID number of the Contacted People. So they can be dropped.
   b. 'Select' which means that the customer has chosen not to answer the question. The ideal value to replace this label would be null value. Hence replaced 'Select' with Nan.
   c. Checked the percentage of **null values** for each column. And **dropped** the columns having missing values **greater** than **45%**.

4. **EDA:** In EDA, Univariate and Bi-Variate analysis was done on both categorical and numerical variables. Conversion rates for each column was checked.

5. **Outlier Treatment:** We form soft capping of upper range outlier values for Total Visits and Page View Per Visit.

6. **Data Preparation:** In this step, the dummy variables are created.

7. **Data Modelling:**
   a. Performed train test data split and scaled the numerical columns.
   b. Initially we had 35 columns. Then we used RFE feature selection method to get the final list of columns. In between the most insignificant, highly correlated columns are dropped and at last we had 15 columns in our final model
   c. We dropped the columns with higher VIF value. And at last we left with 13 columns in our final model.
   d. Set the cut-off probability initially as 0.5 and check the accuracy, sensitivity and precision scores. Also plotted ROC curve.
   e. For our final model, we chose 0.3 optimal probability cutoff by finding points and checking for accuracy, sensitivity and specificity.
   f. Use this cut-off on train data as well and predict the lead conversion

g. Generate the lead score on train and test data using converted probabilities based on cut-off set.

h. After running the model on the Test Data these are the figures we obtain:
   - Accuracy : 92.78%
   - Sensitivity: 91.98%
   - Specificity: 93.26%

i. **Conversion rate** increases from 38% to 88%.

Conclusions:

- ❖ The leads who fills the form are the potential leads.
- ❖ We must majorly focus on housewives and working professionals
- ❖ We must majorly focus on leads whose last activity is SMS sent or Email opened.
- ❖ It’s always good to focus on customers, who have spent significant time on our website.
- ❖ It’s better to focus least on customers to whom the sent mail is bounced back.
- ❖ If the lead source is referral sites, he/she may not be the potential lead.
- ❖ If the lead didn’t fill specialization, he/she may not know what to study and are not right people to target. So, it’s better to focus less on such cases.