Lead Scoring Case Study – Summary

Topic: Logistic Regression

Name 1: Prerana V

Name 2: Aruna D

Summary of Case Study:

We performed the following steps in this case study:

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

2. Data Understanding:
    a. Routine Data Check: No of rows , columns, data type of each column, distribution, mean and median for all numerical columns etc.
    b. Missing value analysis
    c. Duplicate rows check.

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. “Select” value is replaced with NAN
    b. Calculation of missing values for each column and dropping Score and Activity variable.
    c. Dropping the columns with high percentage of missing values.
    d. Checking the unique category for each column.
    e. If the columns are highly skewed with one category, such columns will be dropped. Combining different categories of the columns with less percentage values into "Others" category.
    f. Imputing the column with least missing values percentage
    g. Finally Checking for the number of rows kept after performing all the above steps.

4. EDA : In EDA, Univariate and Bi-Variate analysis was done on both categorical and numerical variables.

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

6. Data Preparation : In this step, the dummy variables are created. Performed train test data split and scaled the numerical columns.

7. Data Modelling:

   a. Initially we had 35 columns. Then we used both RFE and manual feature selection methods to get the final list of columns. In between the most insignificant, highly correlated columns are dropped and at last we had 14 columns in our final model

   b. We know that the relationship between ln(odds) of 'y' and feature variable "X" is much more **intuitive** and easier to understand. The equation is:
   c. ln(odds)= -1.0565 * const + 0.1944 * TotalVisits + 1.0574* Time Spent -0.3186 * Free Copy -1.0199 * Lead Origin_Landing Page Submission + 4.4017 * Lead Origin_Lead Add Form + 1.2101 * Lead Source_Olark Chat-1.1764 * Lead Source_Reference -1.1921 * Last Activity_Email Bounced + 0.8166 * Last Activity_Email Opened -0.6859 * Last Activity_Olark Chat Conversation + 0.6463 * Last Activity_Others - 1.9097 * Last Activity_SMS Sent -1.1380 * Specialization_Not Specified + 2.6908 * Current Occupation_Working Professional

   d. We chose the cutoff probability as 0.35 from Accuracy, Sensitivity, Specificity curve and calculated lead score for all the leads. The sensitivity of model was around 80% and the conversion rate increased from 38% to 73%.

Conclusion:

From model, we can conclude following points:

i. The customer/leads who fills the form are the potential leads.
ii. We must majorly focus on working professionals
iii. We must majorly focus on leads whose last activity is SMS sent or Email opened.
iv. It's always good to focus on customers, who have spent significant time on our website.
v. It's better to focus least on customers to whom the sent mail is bounced back.
vi. If the lead source is referral, he/she may not be the potential lead.
vii. 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.