Summary [Lead Scoring Case Study]

Problem Statement:

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers. The company needs a model wherein a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Solution Summary:

1. Data Sourcing: Importing the required libraries
2. Data Reading & Understanding: Reading the dataset “Leads.csv” and Understanding it as follows :
    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 a 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 the "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, We performed Data Preprocessing, the dummy variables are created. Performed train test data split and scaled the numerical columns.
7. Data Modelling & Model Evaluation:
    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 the 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 the model was around 80% and the conversion rate increased from 38% to 73%.

8. Conclusion: From model, we can conclude following points:
    a. The customer/leads who fills the form are the potential leads.
    b. We must majorly focus on working professionals.
    c. We must majorly focus on leads whose last activity is SMS sent or Email opened.
    d. It’s always good to focus on customers, who have spent significant time on our website.
    e. It’s better to focus least on customers to whom they sent mail is bounced back.
    f. If the lead source is a referral, he/she may not be the potential lead.
    g. If the lead didn’t fill specialization, he/she may not know what to study and are not right people
    h. to target. So, it’s better to focus less on such cases.
9. Recommendations:
    a. It’s good to collect data often and run the model and get updated with the potential leads. There is a belief that the best time to call your potential leads is within a few hours after the lead shows interest in the courses.
    b. Along with phone calls, it’s good to mail the leads also to keep them reminding as email is as powerful as cold calling.
    c. Reducing the number of call attempts to 2-4 and increasing the frequency of usage of other media like advertisements in Google, or via emails to keep in touch with the lead will save a lot of time.
    d. Focusing on Hot Leads will increase the chances of obtaining more value to the business as the numbers of people we contact are less but the conversion rate is high.