Summary [Lead Scoring Case Study]
After understanding the Business Problem and Business Objective. We got a clear understanding of our goals for the case study.
We performed the following steps :
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, the 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 depend on the data-cleaning step. Hence it must be followed thoroughly.
a. “Select” value is replaced with NAN.
b. Calculate missing values for each column and drop Score and Activity variables.
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 fewer percentage values into the "Others" category.
f. Imputing the column with the percentage of the least missing value.
g. Finally Check 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, creating the dummy variables.
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 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 the Accuracy, Sensitivity, and Specificity curve and calculated

the 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 the model, we can conclude the following points:

- The customers/leads who fill out the form are the potential leads.
- We must majorly focus on working professionals.
- We must majorly focus on leads whose last activity is an 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 they sent mail bounced back.
- If the lead source is a referral, 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 the right people
to target. So, it's better to focus less on such cases.

9. Recommendations

- 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.

- Along with phone calls, it's good to mail the leads and also to keep them reminded as email is as
powerful as cold calling.
- 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.
- Focusing on Hot Leads will increase the chances of obtaining more value for the business as the
number of people we contact are less but the conversion rate is high.