Summary for Lead Scoring Case Study

1. Reading and Understanding the Data:

When checked and found 9240 records in leads.csv file and it has 37 columns which include 30 categorical and 7 numerical columns.

2. Basic Data Clean Up:

- 'Select' is not a valid class, we found that the Select might be the default value set in the form dropdowns. So, we replaced 'Select' with NaN.
- We created a function that Dropped the columns having more than 40% missing value.
- We can drop the city column as it has 39% missing value and most values are Mumbai.
- We can't drop specialisation as it as many different values spread out, so we can create additional category called "OTHERS"
- Country can be removed as most of them are Indians.

3. Visualizing Data & Data Preparation:

  - After checking the uniqueness of the data i.e Skewnes , We dropped columns which are not helpful.
  - We dropped some columns which will not add any value to Model and can create the Biased or inaccurate estimations.
  - Tool Visits and Page Views Per Visit have outliers. So, we capped them.
- We can see that conversion rate is of 38.5% i.e only 38.5% has converted to lead (Hot) and 61.5% is not converted (Cold)
- In Categorical Univariate Analysis, we determine the value count percentages for each variable, giving us insight into the distribution of values within each column.
- Current Occupation: 90% of customers are Unemployed.
- Do Not Email: 92% of people have opted out of receiving emails about the course.
- Lead Source: 58% of leads come from Google and Direct Traffic combined.
- Last Activity: 68% of customer activities are attributed to SMS Sent and Email Opened.

4. Model Building:

- By understanding statsmodel "Current_occupation_Housewife" column will be removed from model due to high p-value of 0.999, which is above the accepted threshold of 0.05 for statistical significance.
- "Lead Source_Facebook" column will be removed from model due to high p-value of 0.187, which is above the accepted threshold of 0.05 for statistical significance.

5. CONCLUSION:

The variables which decide the probability of the lead getting converted are as follows:

- Total time spent on website
- Lead source
- Current occupation (if working professional or not)