Summary: Lead Scoring Case Study

Following is the approach used and steps taken to proceed with the case study as per business goal:

1. Data Cleaning/EDA:

a) Dropped all the columns with high percentage of nulls as part of missing value treatment
b) I handled the “Select” level that is present in many of the categorical variables. Replaced it with *np.nan* and dropped the columns having more than 40% “Select” as it would have converted into null values
c) For the columns with less percentages of missing values, I replaced nulls with imputation technique like median/mode or removed redundant data rows
d) Checked unique categories in each categorical columns like in “Country” which had 75% as “India” and 20% null, so I dropped such skewed columns when the data is limited to a particular category, which won’t help for good model fit as effective predictor

2. EDA/Data Transformation:

a) Mapped the binary variables into 0 or 1 and multiple category labels into dummy variables after appropriate grouping for categorical predictors having too many levels
b) Performed outlier handling
c) Removed all the redundant data as per complete exploratory data analysis

3. Data Preparation:

a) As part of data prep before modelling stage, I deleted the data collected by Sales team, which included columns like Tags, Lead Profile, Last Activity, etc.
b) Created Dummy variables for all the categorical columns
c) Performed Train–test split on complete data
d) Completed feature scaling using standard scaling approach

4. Model Building:

a) For feature selection, I used RFE followed fine tuning manual approach
b) Built a logistic regression model with good sensitivity/recall
c) Checked the p-values of all predictors to be less than 0.05 and VIF less than 5
d) Checked the optimal probability cut-off value and plotted a graph for the same
e) Checked the model performance over test data via various model evaluation metrics
f) Generated the lead score variable with respective lead numbers where a high lead score indicated a hot lead

✚ **Model Summary/Conclusion:**

a) We can see that we after our model build, the company's sales team will be able to identify *79.3%* of hot leads accurately (*accuracy score*) and *sensitivity* of *76.5%* – this will enable the company to attain a higher conversion rate based on a good identification of promising leads
b) We have arrived at a good model for the identification of hot leads with the significant variables
c) Top 3 features for good conversion rate are:
   - *Lead Origin_Lead Add Form* with the coefficient of 4.18
   - *Current Occupation_Working Professional* with the coefficient of 2.719
   - *Lead Source_Welingak Website* with the coefficient of 2.024