Leads Scoring Case Study Summary

Steps followed:

1. Data Understanding:

- I checked for null values and duplicates. I found no duplicates, but some columns had null values.

2. Data Cleaning:

- Some columns were labeled 'Select,' meaning the customer refused to answer. We changed those labels to null values.
- Removed columns having more than 45% null values.
- Imputed value's mode () for Numerical columns having missing values.
- I analyzed the missing values in the categorical features and decided to drop the columns that had skewed data, like "City," "Country," "Search," and "Do not call." I also dropped "Tags" because it had values that seemed to be created by the sales team based on the current status of the lead.
- Removed columns like Prospect_ID, Lead_Number, and Last Notable_Activity that aren't needed for the model

3. EDA:

- Performed data imbalance check on target variable- "Converted". It had lower converted (38.5%) records as compared to those which were not converted (61.5%).
- Performed Univariate, Bivariate analysis to draw insights.Categorical columns e.g. Lead Source and Specialization were having various values with low counts, so they were merged to form "Others" category to avoid handling multiple categories during modelling.

4. Data Transformation:

- Removed "Free copy" redundant column identified during EDA.
- Performed Outlier treatment.
- Changed the multi-category labels into dummy variables and binary variables into '0' and '1'.

5. Data Preparation:

- Split the dataset into train and test.
- Performed Feature Scaling using StandardScalar().

6. Model Building:

- Created Logistic Regression model using RFE for 15 count, followed by manual feature reduction to reach at 13 variables by checking VIF, P-Value (VIF<5 and p-value <0.05). Also checked Information Value and negative coefficient to drop three columns to reach our final model of 10 variables

7. Model Validation:

- Performed probability predication.
- Checked the optimal probability cutoff by finding points and checking the accuracy, sensitivity and specificity and found one convergent points (at 0.34).
- Checked confusion matrix, Accuracy, Sensitivity, and Specificity ranged in 80% (acceptable range). ROC curve (0.87 area under the curve)
- Performed Precision-Recall Trade off that gave cut off 0.404 which reduced Accuracy, Sensitivity, Specificity etc. to 75% range, so we decided to use 0.34 cut off.
- Assigned Lead Score on the training data.

8. Making Predictions:

- Performed Scaling and Performed prediction using final model

9. Model Evaluation/ Recommendation:

- Created Confusion matrix, ROC curve on Test Model. Test set is also having accuracy, recall/sensitivity in an acceptable range of 80%.
- Performed Lead Score Assignment on test data.
- Top Predicators with good conversion rate:
    - Lead Origin_Lead Add Form
    - Lead Source_Welingak Website
    - Working Professionals

- Recommendations:
    - Leads from the "Lead Add Form" emerged as a most significant attribute for Hot Leads, so we should give more importance to customers you came it.
    - Working professionals have higher chance (around 90%) to convert.
    - Welingak Website has around 98% lead conversion rate. More ads should be given on this website to cater the leads from it.
    - Leads that came through a reference has over 90% conversion rate. We should encourage and incentivize existing members to bring more referrals.