Leads Scoring Case Study Summary

Steps followed:-

1. **Data Understanding:**
    a. Performed Null/duplicates inspection. But no duplicates found

2. Data Cleaning:

    a. Some columns were having label as 'Select' changed to null values.
    b. Removed columns having more than 45% null values.
    c. Imputed value's mode () for Numerical columns having missing values.
    d. Checked missing values in the Categorical features and decided to drop columns for which data was skewed e.g. City, Country, Search,Do not call etc..Dropped "Tags" because it had values like incorrect number, switched off etc. by which it seemed to be created by the sales team based on the current status of the lead.
    e. Deleted columns irrelevant for model building e.g. Prospect_ID, Lead_Number and Last Notable_Activity (updated by sales team).

4. EDA:

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

5. Data Transformation:

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

6. Data Preparation:

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

7. Model Building:

- **a.** Created Logistic Regression model using RFE for 15 count, followedby 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.

8. Model Validation:

- **a.** Performed probability predication.
- **b.** Checked the optimal probability cutoff by finding points and checking the accuracy, sensitivity and specificity and found oneconvergent points (at 0.34).
- **c.** Checked confusion matrix, Accuracy, Sensitivity, and Specificity ranged in 80% (acceptable range). ROC curve (0.87 area under thecurve)
- **d.** 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.
- **e.** Assigned Lead Score on the training data.

9. Making Predictions:

- **a.** Performed Scaling and Performed prediction using final model

10. Model Evaluation/ Recommendation:

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

d. Recommendations:

- Leads from the "Lead Add Form" have third highest conversions with conversion rate. Hence we should try to put Lead Add Forms on the social media websites specially on the Welingak Website and we should give more importance to customers you came through this channel.
- More focus should be given in engaging with the Working professionals because of high conversion rate
- More adds should be given on Welingak Website to cater the leads from their, as it has higher chance to conversion.
- Leads that came through a "reference" has over 90% conversion, we should encourage and incentivize existing members to bring more of their referrals.