Summary - Lead Scoring

Based on data provided on potential customers visiting the websites, time spend and channels taken, we have done analysis to generate more leads and having the professional enroll the courses

1. Data Cleaning :

The initial and foremost step is to clean the data such as removing null values, replacing the options select with null variables, few were changed to data “not provided” to avoid loosing major data

2. EDA:

EDA is done to check the condition of data. Few elements in categorical variables were found irrelevant, clubbed minimal lead sources into others. No outliers found.

3. Dummy Variables

Dummy variables were created. Standard scaler was used for numerical values.

4. Train-Test Split

The data was split in the ratio of 70% as train data and 30% as test data. Random state of 100% was adopted as well.

5. Model Building

RFE was used to reveal the top 15 variables/features to proceed further with the model building. The final model was decided based on the p-value and the VIF of the features. It was ensured in the final model that all the features have a p-value less that 0.05 and VIF less than 5.

6. Model Evaluation

ROC curve shows the tradeoff between sensitivity and specificity. The closer the curve follows the y-axis and then the top border of ROC space means more area under curve and the more accurate the test.

The closer the curve comes to the 45-degree diagonal of the ROC space means less area and the less accurate is the test.

7. Prediction

In sensitivity-specificity-accuracy plot 0.27 probability looks optimal. And in precision-recall curve 0.3 looks optimal.

We are taking 0.3 is the optimum point as a cutoff probability and assigning Lead Score in training data.

8. Precision Recall

The graph depicts optimal cutoff of 0.42 based on precision and recall
Precision = 79%
Recall = 65%