Lead Scoring

Input Data

Provided with a leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

Goal

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
2. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

   Steps taken to build a logistic regression model:

   1) Imported python libraries for processing, visualization and customization
   2) Data Preparation -
      By removing duplicates(this step is not required as there are no duplicate records in input data)
   3) Data cleaning -
      - As we can observe that there are select values for many column. This is because customer did not select any option from the list, hence it shows select. So replaced "Select" with null value
      - Dropped the columns containing more than 45% null values one by one
      - For remaining columns having considerable missing values replaced null values with respective highest occurring value
      - For columns containing missing values under 2% so we can drop these rows.
      - Created Good data- "Leads_Cleaned" for model building
   4) Exploratory Data Analytics-
      Done univariate analysis for the "Converted" target variable, Indicates whether a lead has been successfully converted (1) or not (0).
      Based on the univariate analysis we have seen that many columns are not adding any information to the model, hence we dropped them for further analysis
   5) Data Preparation –
      - Converting some binary variables (Yes/No) to 1/0
      - For categorical variables with multiple levels, create dummy features (one-hot encoded)

    - Adding the results to the master dataframe
    - Splitting the data into train and test

6) Feature scaling –
    - Checking the Churn Rate
7) Model Building –
    - Run the first training model
    - Feature selection using REF
    - Assessing the model with StatsModel
    - Generalized Linear Model Regression Results
8) Getting the predicted values on train dataset –
    - Created a dataframe with the actual churn flag and the predicted probabilities
    - Creating new column 'predicted' with 1 if Churn_Prob > 0.5 else 0
    - Calculated the VIF, Accuracy, sensitivity and specificity
9) Plotted the ROC Curve
10) Finding Optimal Cutoff Point – balanced sensitivity and specificity
11) Precision and Recall tradeoff
12) Making predication on the train set