Lead Scoring Case Study Summary

Summary:

1.Reading and Understanding Data.

Read and analyze the data.

2.Data Cleaning:

The data cleaning has been done to drop the variables that had high percentage of NULL values in them. The imputing of the missing values. Creation of new classification variables in case of categorical variables. The outliers were identified and removed.

3. Data Analysis

The Exploratory Data Analysis of the data is done. In this step, there were around 3 variables that were identified to have only one value in all rows. These variables were dropped.

4. Creating Dummy Variables

The dummy data for the categorical variables are created.

5.Test Train Split:

The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

6. Feature selection using RFE:

The Recursive Feature Elimination was used to select the 20 top important features. Using the statistics generated and looked at the P-values in order to select the most significant values and dropped the insignificant values. At last the 15 most significant variables are selected. The VIF's for these variables were also found to be good. Then created the data frame having the converted probability values and had an initial assumption that a probability value of more than 0.5 means 1 else 0. Based on the above assumption the Confusion Metrics are derived and calculated the overall Accuracy of the model. Then calculated the 'Sensitivity' and the 'Specificity' matrices to understand how reliable the model is.

7. Plotting the ROC Curve

The ROC curve is plotted for the features and area coverage of 89% was found which further solidified the of the model.

8. Finding the Optimal Cutoff Point

The 'Accuracy', 'Sensitivity', and 'Specificity' for different are found. The intersecting point of the graphs was considered as the optimal probability cutoff point. The cutoff point was found out to be 0.37 Based on the new value we could observe that close to 80% values were rightly predicted by the model. The new values of the 'accuracy=92.21%, 'sensitivity=91.9%', 'specificity=92.4%' are found.

9. Computing the Precision and Recall metrics

The Precision and Recall metrics values are found to be 88.13% and 91.9% respectively on the train data set.

10. Observation

After the test and train data are computed analyse of results should be done.