Leads Scoring Case Study

A summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.

Answer:

Below are the steps how we have proceeded with our assignments:

1. Data Cleaning:

a. We chose to delete/remove the unnecessary variables/features to get a clean dataset.

b. Pursuant to the removal of the redundant columns, we noticed that customers have not chosen to answer this question and hence some columns were labelled as "Select". Since the customer have not opted any option and accordingly the ideal value given is null value. Hence, we changed those labels from 'Select' to null values.

c. Columns having more than 30% null values were removed

d. Maximum number of occurrences for a column were imputed in the remaining missing values,.

e. We noticed that one column is having two identical label names in different format i.e. capital letter and small letter. This issue was fixed by changing the labels names into one format.

2. Data Transformation:

a. Multicategory labels were changed into dummy variables and binary variables into '0' and '1'.

b. Outliers were checked, and bins were created for them.

c. Removed all the repeated and unnecessary columns.

3. Data Preparation:

a. Split the dataset into train and test dataset and scaled the dataset.

b. Heatmap was plotted to check the correlations among the variables.

c. Found some correlations and they were dropped.

4. Model Building:

a. We created our model with RFE count 15 and compared the model evaluation score like AUC .

b. For our model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity, and specificity.

c. We found one convergent points and we chose that point for cut off and predicted our final outcomes.

d. We checked the precision and recall with accuracy, sensitivity and specificity for our final model and the trade-offs.

e. Prediction made now in test set and predicted value was recoded.

f. We did model evaluation on the test set like checking the accuracy, recall/sensitivity to find how the model is look.

g. We found the score of accuracy and sensitivity from our final test model is inacceptable range.

h. We have given lead score to the test dataset for indication that high lead score are hot leads and low lead score are not hot leads.

5. Conclusion:

Learning is gathered below:

i. Test set is having accuracy, recall/sensitivity in an acceptable range.
ii. In business terms, our model is having stability an accuracy with adaptive environment skills. Means it will adjust with the company's requirement changes made in coming future.
iii. Top features for good conversion rate in descending order:
   1. Total Visits
   2. Page Views Per Visit
   3. Total Time Spent on Website
   4. Lead Origin_Lead Add Form