Leads Scoring Case Study

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

1. Data Cleaning:

a. First step to clean the dataset we choose was to remove the redundant variables/features.

b. After removing the redundant columns, we found that some columns are having label as 'Select' which means the customer has chosen not to answer this question. The ideal value to replace this label would be null value as the customer has not opted any option. Hence, we changed those labels from 'Select' to Unknown values.

c. Removed columns having more than 30% null values

d. For remaining missing values, we have imputed values with maximum number of occurrences for a column.

e. We found for one column is having two identical label names in different format (capital letter and small letter). We fixed this issue by changes the labels names into one format.

2. Data Transformation:

a. Changed the multicategory labels into dummy variables and binary variables into '0' and '1'.

b. Checked the outliers.

c. Removed all the redundant and repeated columns.

3. Data Preparation:

a. Split the dataset into train and test dataset as 70% & 30% respectively and scaled the dataset.

b. After this, we plot a heatmap to check the correlations among the variables.

4. Model Building:

a. Firstly, RFE was done to attain the top 15 relevant variables.

b. . Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

5. Model Evaluation:

a. A confusion matrix was made and was used to find the accuracy, sensitivity and specificity which came to be

- Accuracy = 79.20%
- Sensitivity= 81.11%
- Specificity= 78.00%
- Precision = 69.78%
- Recall = 81.11%

b. We found one convergent point of 0.35 and we chose that point as cutoff and predicted our final outcomes.

6. Prediction:

a. Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of

- Accuracy = 79.14%
- Sensitivity= 79.77%
- Specificity= 78.77%
- Precision = 68.19%
- Recall = 79.77%

b. We found the score of accuracy and sensitivity from our final test model is in acceptable 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.

7. Conclusion:

Learning gathered are 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:
1] Last Notable Activity_Had a Phone Conversation
2] Lead Source_Welingak website
3] Lead Origin_Lead Add Form
4] Last Notable Activity_Unreachable
5] Last Notable Activity_SMS Sent