Leads Scoring Case Study

A brief 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. Reading and Understanding the Data: 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 null values.

c. Removed columns having more than 50% null values

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

2. Data Cleaning:

a. Checking the number of unique values per column and removing the column with only one unique value.

b. Checked the Lead Quality and imputing NaN value with Not sure and dropping it because the Not sure variables is occurring at least 63 % and checking the same for other variables like City, Specialization, Tag, What matters most to you in choosing a course etc.

c. Removed all the redundant and repeated columns.

3. Exploratory Data Analysis:

a. Outlier Treatment

b. Univariate Analysis.

c. Bivariate Analysis.

d. Checking correlation between variables to check multi collinearity.

4. Data Preparation:

a. Converting categorical variables into dummy variables.

b. Dropping the original columns and merging the dummies onto the original data frame.

c. Performing Train-Test Split.

d. Scaling numerical variables by Standard Scaling**.**

e. Removing highly corelated variables from the tailing and test dataset.

5. Modelling:

a. Creating Logistic Regression Model.

b. Checking the p-values and VIF for all the variable and using RFE for feature selection with 15 variables.

c. Assessing the model with StatsModels again and again till we find the most significant model.

d. Generating predicted values on the training set.

e. Checking the sensitivity and specificity.

f. Plotting the ROC Curve and also finding the optimal Cut off point.

g. Making predictions on the test set.

h. Calculating the sensitivity, specificity and accuracy on test set.

6. Conclusion:

Learning gathered are below:

- Test set is having accuracy, recall/sensitivity in an acceptable range.
- In business terms, our model is having a stable accuracy with constantly changing requirements of the company in the upcoming future.
- Top feature for good conversion rate are:
    1. Lead Source_Welingak website
    2. Lead Source_Refernce

3. Last Activity_Had a Phone Conversion