LEAD SCORE CASE STUDY

Report

Reading Data

Data is loaded , read and stored in dataframe.

Data Preparation

Identification of the null values of each and every record and it resulted in null values ranging from (20-52)% .

Data type of each variable is identified and treated.

Variables which have high null values > 40% we dropped them knowing that those attributes doesn't give that much value.

Now we checked the count of each and every distinct record and we started figuring out each and every variable and we imputed the null values as:

- The Categorical attributes which has null values we have imputed them to "Others", "Not Sure","Unknown".
- The Categorical attributes which have "Select" as the records we have considered them as null value and imputed them to "Others" etc.
- For some categorical attribute eg : "Tags" we have replaced redundant and not useful records to "Phone Issue".
- For some of the categorical attribute we have replaced the null values with the major outstanding record in that variable.

Finding Outliers

Outliers were found by plotting the box plot and the count plot to check the variables with high outliners and variance.
For numerical attributes we have checked them by **describe()** and checking for a large variance in [0.1,0.25,0.5,0.75,0.9] percentile and is found that we don't have that much variance in them.

Data Exploration

- We have started creating dummy variable for the categorical attributes which have more than two distinct records.
- For the attributes which have only two distinct records we have replaced them with 1 or 0.
- After creating the dummy variables we have dropped the original variables.

Train_Test Split

We have divided the dataset into train and test set and did analysis on train first so that after analysing it and building the model we can test out model on test data set to get the accuracy of the model.

Scaling

We have scaled the numerical attributes to bring them in range such that the mean is 0 and standard of deviation is 1.
Scaled both the train data set and test data .

Model Building

- Primarily the model was built upon Stats Model methodology.
- Initially the input to model had 164 columns or attributes.
- RFE was applied for top 15 features.
- Running the model again with 15 features.
- Applied VIF to check multicollinearity.
- Further 1 more feature was dropped having VIF value greater than 5.
- Finally, Model was built on 14 features.

Model Evaluation

- Confusion Matrix was used to evaluate the model.
- The Model was evaluated on the parameters of sensitivity, specificity, positive predictive value and negative predictive value.
- The outputs of test data for these parameters validates the robustness of the Model.

specificity	0.83
sensitivity	0.74
positive predictive value	0.74
Negative predictive value	0.84