Summary Report on Lead Case Study

Here after importing the data, we started preparing it. We dropped the columns which had missing values more than 40%. Rest which had missing value less than 40%, we treated them. In case of categorical columns, in place of nan values we imputed the value which had the majority in a column or can say using mode. Later we did outlier treatment on continuous columns.

Once data was cleaned, we proceeded towards doing the EDA on the data and found inferences which would help us in getting the idea of variables which had good conversion rate. Through this we dropped the variables which were skewed or shown no inference.

Then we proceeded towards doing the Logistic regression on the data and firstly using RFE technique we found out the top significant variables and applied the logistic regression with those variables. Finally, we got the model with variables having p value less than 0.05 and low vifs. After getting the model, we found optimal cut off to be 0.4 and at that cut off our accuracy, sensitivity and specificity were above 80% for train data. We ran the same codes on test data and found accuracy, sensitivity and specificity to be above 80% with optimal cut off 0.4. In Train data, The ROC was 0.93 and for test data it was 0.92. Anything between 0.90 and 1.00 means the model is perfect. After that we found the lead score of variables for all data. Accordingly, we found the top variables to find the hot leads.

Learning: Learnt different ways of imputing the missing values, importance of EDA to understand the data better and found which variables are important and which are not. Also found cleaning and dropping of not useful variables plays important role while applying rfe and logistic regression on the model.

Challenges: If data is not cleaned properly or any wrong imputation will lead to a lot of difference in accuracy, specificity and sensitivity which was very time consuming and had to check complete code and steps again.