Summary:

Data understanding and cleaning

We have started with understanding the data after loading them. In the initial glimpse, we observed that there were 37 columns and 9240 rows or records. Then we observed the continuous and categorical variables present with majority of columns being the categorical variables. The target variable was ‘Converted’. There was a good distribution of 1 and 0 in the target variable. As there are many columns or variables, we explored various methods to eminate some of those columns. The first method was to check if the any variable has a constant value, then it will not add any value to our model and hence it is recommended to remove such variables. Columns like 'Magazine','Receive More Updates About Our Courses','I agree to pay the amount through cheque','Get updates on DM Content','Update me on Supply Chain Content' had a constant value and we removed them. Then we looked at the variable for missing data and we removed variables which had more that 35% of the values missing. Along with it we also removed variables which were unique to each row and does not add value to the model. Also, one important thing we understood after looking at the leads sales cycle, we realized that there were few variables which were added to data after making the call to the leads. Since, we are trying to flag the customers before making the calls, it does not make sense to add variables which would be added only after making the calls,. So we removed such variables too. Here is the list of variables removed based on the three methods just discussed : 'Last Activity','Last Notable Activity','Lead Profile','Tags','Lead Quality','Asymmetrique Profile Index','Asymmetrique Activity Index','Asymmetrique Activity Score','Asymmetrique Profile Score','Lead Number','How did you hear about X Education','City','Specialization','Prospect ID'. Next we imputed the remaining missing values from the available columns. For the continuous variable we used mean and for the categorical variable we have used mode to impute. Next we looked at variable which had highly skewed data. Columns such as 'Do Not Call', 'Country', 'What matters most to you in choosing a course', 'Search', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations' were removed.

Next we looked at the distribution of the variables split by ‘Converted’ to check the distributions and relation between variables. Also, correlation matrix was created for the continuous variables.

Model building and evaluation:

Once we finalized the variables we wanted to work with, we have dummified the categorical variables. Then we split the data into train and test dataset with the split ratio of 70 to 30. Then we used RFE to shortlist the variables and we experimented with multiple values and almost always ended up similar model. So, we shortlisted 12 variables using RFE and built the model. Couple of thing we looked for is the p value(for null hypothesis) and VIF(for multicollinearity). We have roved variables which had p value > 0.05 and rebuilt our model. Moving on to the cutoff value, we have started with 50% and observed an accuracy of 77% and then looked at accuracy, sensitivity and specificity across different cutoff values and found the 30% giving the best results. We then calculated the evaluation metrics on the test data and found the accuracy to be 77%