Lead Case Study Summary of the Reports

This case study talks about the X-Education group wanting to convert their leads to a full time student. The problem exists in the current lead is the only 30% conversion rate. X-Education group what to increase the conversion rate to high as possible based on the information present in the leads.

Steps:

1. The very first steps for any task is to understand the data and check the null values inside the data.
2. So in the lead data there was huge columns were having null values so there are multiple approaches taken to clean the data, like more then 40% column having null values, can be removed, then select values in the columns, then maximum data fall into on category also cleaned, after that mark category as Other if very less data assign to it. Where this cleaning of the data was very difficult is where I have learned a lot of new techniques of cleaning the data.
3. After the visualization the data to check the correlation between columns
4. Once I understand the data using visualization, then create dummy variables for categorical columns and remove the columns whose dummy variables created.
5. Once the final columns we have to perform the model analysis, pass those columns to the train test split method with 70% and 30% ratio. Also perform the StandardScaler for the numeric columns to scale them.
6. Next I am starting to find the best 25 variables among all the variables using RFE approche. Out of the best 25 columns I perform the manual approaches to remove the less contribution to the model and lead conversion. These columns are manually removed via checking the high p-values and VIF in the model. In each model trying to find the accuracy_score to assure that the model note reduces drastically after removing the particular column.
7. After running the 11 models finally we got the best columns and categories are participating to convert the lead to a full time student and its overall accuracy_score also ~80%. Also we calculated the sensitivity, specificity, which are perfectly positive.
8. Perform the ROC curve and find the optimal cutoff to convert probability of the leads
9. After all these steps, run a model on the test data which is matching with the train accuracy_score, specificity, sensitivity.
10. So at the conclusion we can say the model 11 has the best conversion rate around ~80%