Summary

Problem Statement: An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. The typical lead conversion rate at X education is around 30%.

Goal:

1. Build a logistic regression model to identify top factors affecting lead conversion.
2. Identify the most potential leads, also known as 'Hot Leads' to increase lead conversion rate to around 80%.

Data:

9240 rows, 37 columns

Approach:

1. We started by loading the libraries, loading the data
2. We examined the data for shape, descriptive stats, dtypes etc.
3. We calculated the null values to remove columns with more than 35% null values
4. We replaced 'select' value with null values
5. We recalculated null values and again removed columns with more than 40% null values
6. We identified columns with only one unique value and removed them as they had limited variance and information.
7. We checked variables for outliers and removed the extreme values
8. We conducted univariate, bivariate and multivariate analysis to understand the relationship between variables.
9. We checked the relationship of each of the variables with the target variable to check which category of leads are showing maximum lead conversion.
10. We checked for columns having similar data and removed them
11. We obtained dummy variables for all the categorical columns
12. With the final list of variables we started model building.
13. Divided the data into train-test data.
14. Scaled the numerical features to bring them in range with dummy variables.
15. Ran the first model with statsmodel library using all variables
16. Conducted RFE to select top 15 features
17. Made the first iteration of the model with 15 features
18. Checked the p-vales and VIF values of the variables to eliminate the insignificant variables and variables showing collinearity, one at a time.

19. Simultaneously checked the accuracy of each model to ensure no drastic drop in accuracy
20. Arrived at the final model with good p-values (less than 0.05) and VIF values (less than 5)
21. Calculated the training evaluation metrics with cutoff at 0.5
22. Plotted the ROC curve to check robustness of model.
23. Calculated the optimum probability cutoff point with accuracy, specificity and sensitivity.
24. Recalculated the training evaluation metrics with new cutoff value (0.35).
25. Calculated the precision and recall metrics with cutoff at 0.5
26. Calculated the optimum probability cutoff point with precision and recall (0.41).
27. Scaled the test data.
28. Made predictions on the test data with these new cutoff values.
29. Compared evaluation metrics of test and train data to check if any significant change in accuracy which might indicate overfitting / underfitting.
30. Tabled the lead conversion probability percentage values.
31. Identified hot leads as leads with conversion probability percentage more than 80%.