Lead Scoring Case Study Summary

Problem Statement: X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance

Steps:

Step1: Reading and Understanding Data. Read and analyse the data.

Step2: Data Cleaning:

- removing columns with null value> 30%, columns which does not show any variations, city and country
- variables lead profile, how did you hear about x education and specialization have select values. dropping other 2 columns based on its value counts showing very high select value compared to others
- cap the outliers to 95% in page views per visit

Step3: Data Analysis Then we started with the Exploratory Data Analysis of the data set to get a feel of how the data is oriented. In this step, there were around 3 variables that were identified to have only one value in all rows. These variables were dropped

Step 4: PREPARING DATA FOR MODEL

- Binary mapping of data and adding dummy columns for categorical columns
- Scaling of numeric features for normalizing independent variables using fit transform

Step 5: MODEL BUILDING

- After splitting test train, using stats models glm top 15 RFE variables are identified and used for building model
- dropping 'What is your current occupation Housewife' as its P>|z| is 0.999 again building the model with dropped data
- checking VIF for multicolinearity and found everything to be below <5
- We created the data frame having the converted probability values and we had an initial assumption that a probability value of more than 0.5 means 1 else 0. Based on the above assumption, we derived the Confusion Metrics and calculated the overall Accuracy of the model.
- We also calculated the 'Sensitivity' and the 'Specificity' matrices to understand how reliable the model is. Step8: Plotting the ROC Curve We then tried plotting the ROC curve for the features and the curve came out be pretty decent with an area coverage of 89% which further solidified the of the mode

Step6: Finding the Optimal Cut-off Point Then we plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different probability values. The cut-off point was found out to be 0.38

Step7: Making Predictions on Test Set Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out a Specificity –.78, Sensitivity -.83, ACCURACY -.81