Lead Scoring Case Study Summary

Data Processing and EDA

- First we try to get an overall picture of the data, the columns that are present and the data type of the columns
- We drop the columns which have only a single variable in it
- We check for duplicates, if present we remove them
- We check for the null values in the dataset, if the null value percentage is greater than 40%, we simply remove the columns
- The variables which have less than 40% missing data, we do imputation, all the columns with the entry as select goes to the no data available cells
- The numerical null values gets replaced with its mode
- We remove some columns which are not required in our analysis like city, country, prospect id and lead number
- We have plots of all the categorical columns to check for the variables present in it and its variance, the insights we generated were-
    1. Most lead originated from the landing page
    2. last noticeable activity were modified, mail opened or sms sent
    3. Data is not available at most of the places in Specialization and where did you hear about x education columns
    4. Most of the individuals applying for the course are unemployed and are looking for a better course prospect
    5. Data is not available for most of the lead profiles
- We remove the columns in which the data is highly skewed.
- We find out the correlation of the variables
- We check for data imbalance within the dependent variable and we check for this within different categorical independent variables too.
- We check for outliers in the data

Model Building

- We split the data into training and test data
- Finally we scale the variables using feature scaling, we used minmaxscaler here
- We use RFE to choose the variables to use in the model
- We build a logistic regression model
- We iterate over the model until all the variables have a p value of less than 0.05 and a VIF of less than 5

Model Prediction & Evaluation

- We find out our y predicted and evaluate it against our y train, we get an accuracy of 80%
- We further use confusion matrix to calculate specificity and sensitivity
- We make an ROC curve
- We find the optimum cut off value, it came out to be 0.35 in our case
- We again calculate the accuracy, specificity and sensitivity
- Finally we run our data on the test set and again calculate the specificity and sensitivity
- We do the precision recall tradeoff and come up with a new cut off of 0.41 and finally we calculate the specificity and sensitivity again

Model Insights

- Top Predictors to our model are
    1. TotalVisits
    2. Total Time Spent on Website
    3. Lead Origin_Lead Add form
    4. Last Notable Activity_Had a Phone Conversation
    5. What is your current occupation_Working Professional
- Negative influencers are
    1. Do not email_yes
    2. Lead origin_landing page submission