Lead Scoring Case Study Summary

1. Data Cleaning:

a. First step to clean the dataset and remove the redundant variables/features.

b. After removing the redundant columns, we found that some columns are having label as ‘Select’ Hence, we changed those labels from ‘Select’ to null values/Not Provided

c. Removed columns having more than 35% null values

d. For remaining missing values, we have imputed values with Median and Mode

e. We found for one column is having two identical label names in different format (capital letter and small letter). We fixed this issue by changes the labels names into one format.

2. EDA :

a. A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seems good and no outliers were found.

3. Data Preparation:

a. Changed the multicategory labels into dummy variables and binary

b. Variables into ‘0’ and ‘1’. Checked the outliers and created bins for them. Removed all the redundant and repeated columns.

c. Split the dataset into train and test dataset and scaled the dataset.

d. After this, we plot a heatman to check the correlations among the variables.

4. Model Building:

a. Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).
b. For our final model we checked the optimal probability cutoff by finding points and checking the accuracy, sensitivity and specificity.
c. We found one convergent points and we chose that point for cutoff and predicted our final outcomes.
d. We checked the precision and recall with accuracy, sensitivity and specificity for our final model and the tradeoffs.
e. Prediction made now in test set and predicted value was recoded.
f. We did model evaluation on the test set like checking the accuracy, recall/sensitivity to find how the model is.
g. We found the score of accuracy and sensitivity from our final test model is in acceptable range.