LEAD SCORE CASE STUDY

SUMMARY REPORT

1. After importing the dataset, we first replaced the select values into NaN, as they are the as good as null values and won't give any insights into the data.
2. We dropped columns having null percentage greater than 45%
3. We dropped columns like country as these columns tend to have a very high percentage of only one particular value and the remaining values have very less number of rows.
4. In some of the columns, there are a lot of categories with very less number of rows. Hence we club them all in one category called 'others'.
5. Imputation of values into the columns that have very less missing percentage with either mean, median or mode.
6. Data Visualisation (Performing EDA with various variables) to gain insight of the data and for any outliers (if any outliers present, we cap them).
7. Creating dummy variables for all the categorical columns.
8. Splitting the data into test and train sets.
9. Rescaling all variables except dummies, using MinMaxScalar so that the numerical values are similar in terms of magnitudes, units, and range.
10. Building the logistic regression model.
11. Select the top 20 features using RFE.
12. Then, using manual approach, that is VIF scores and the p-values, we remove features one by one until VIF score is less than 0.5 and p-value is between 0 to 0.05. The range can vary according to the business aspects.
13. We then find the optimal probability cut-off. For this, we calculate accuracy, sensitivity and specificity for various probability cut-offs. And we plot them in order to find the probability cut-off. The point at which all the 3 curves meet is the optimal cut-off, in our case, being 0.35.

14. We find out the final prediction value, according to the optimal probability cut-off.
15. We create new column 'Lead Score' with where lead Score= Converted_prob*100. Hence we find out what the lead score for each lead will be.