LEAD SCORING CASE STUDY – SUMMARY REPORT:

The Problem is approached by six steps:

1. Reading and Understanding the Data:
   After loading the dataset, The Sales team's imputed columns were removed initially. There were 'Select' text which were present in each column by default and those were converted to null values. Dropping of the columns were done for which had null values more than 35%. Dropped all the highly skewed or completely skewed columns i.e., more than 99% to a single category. Combined the categories in which columns had value counts less than 10% and named as others. After the above steps, the rows with null values were removed as it was 1.48% and thereby, Utilized 98.52% rows for modelling. Outliers in two columns were handled by soft capping i.e., by substituting the outliers with 99th percentile values
2. Data Preparation:
   All the Categorical columns were changed to dummies keeping the drop first= true and the dummies were concatenated to the main dataset. Lead Identifier and Categorical(as the data is moved to dummies) columns were removed. After these procedures we were left with 17 columns including the Target variable 'Converted'.
3. Splitting the datasets into train and test dataset:
   The datasets were split into two with 70% and 30% for train and test respectively. Normalized scale was made to fit_transform on the numerical variables. From the correlation plot, we could see the time spent on the website and Other occupation were more positively correlated and na_chosen was more negatively correlated with converted values. Thereby, these variables should help more in determining the potential leads.
4. Creation of the model:
   After popping out the Converted column for the y_train. We were left with 16 variables. We used RFE to remove one variable 'India' from the country dummies. Seven iterations were done using Statmodel-GLM- Binomial method and each time a column was removed based on the higher p-value(>0.1) and higher VIF value(>5). By which, we were left with 8 variables which contribute in determining the probability of getting the potential leads. We moved the Converted and Predicted probabilities to a new dataset for further analysis.
5. Model Evaluation:
   In this step, we worked towards getting the right threshold and we decided that by Optimal probability cut-off method. We found that the cut-off point by plotting the sensitivity, specificity and accuracy of probabilities at the maximum interval of 0.1 between 0.0-1.0. Also from the plot, we got the threshold by the intersection of the above mentioned parameters approximately around 0.342886. By this threshold, we got the recall score of train dataset as 81.39%.
6. Prediction on test dataset:
   The trained scale was applied, transformed on test dataset and used the model to predict the probabilities in test dataset. From the optimal cut-off obtained in model evaluation 0.342886, we got the recall score of 80.71%. Score variable generated on the test output by getting the predicted probability in percentages.