Lead scoring case study summary

PROBLEM STATEMENT: -

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 has a lower conversion chance.
Analysis must be performed to find leads with higher probabilities so that they can be our fruitful customers. Historic data has been provided to us for understanding various factors affecting the conversion rate.

SUMMARY:

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

Step2: Data Cleaning:
We excluded the variables that contained a significant proportion of missing values, and we dealt with the remaining missing values by imputing median values for numerical variables or creating new categorical variables for categorical variables. We also detected and eliminated outliers.

Step2: EDA
Exploratory Data Analysis was performed to understand the data. Different graphs were plotted using categorical data analysis and numerical variable analysis and relation of the categorical variables with convert. To our observation no outliers were found but we were able to figure out various variables (columns) that were of no use to us.

Step4: 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.

Step5: Creating Dummy Variables
We went on with creating dummy data for the categorical variables.

Step6: Test Train Split:
The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

Step7: Feature Rescaling
We used the MinMaxScaler to scale the original numerical variables. Then using the stats model we created our initial model, which would give us a complete statistical view of all the parameters of our model.

Step8: Feature selection using RFE:
We utilized Recursive Feature Elimination to choose the 15 most significant features. We repeatedly examined the P-values generated by the statistics to select the most relevant values and eliminate the insignificant ones. Additionally, we calculated the 'Sensitivity' and 'Specificity' matrices to evaluate the model's reliability.

Step9: 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 under the curve of 0.88 which further enhanced our model.

Step10: Finding the Optimal Cutoff Point

We generated probability graphs for 'Accuracy', 'Sensitivity', and 'Specificity' using different probability values. The point where the graphs intersected was identified as the ideal probability cutoff, which was determined to be 0.35. With this cutoff value, we achieved an accuracy, sensitivity, and specificity of approximately 80%. We also computed the lead score and estimated that the final predicted variables yielded a target lead prediction of roughly 80%.

Step11: Computing the Precision and Recall metrics

We also found out with cutoff 0.35 the Precision and Recall metrics values came out to be around 80% and 70% respectively on the train data set. With the current cut off as 0.40 we have Precision around 75% , Recall around 77% and accuracy around 81%.

Step12: Making Predictions on Test Set

Precision around 73% , Recall around 76% and accuracy around 80%,This Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model.