Lead Scoring Case Study Summary

Step 1: The libraries required for data processing, data visualization and finally model building and evaluation are imported.

Step 2: The dataset is loaded and inspected.

Step 3: Now Exploratory Data Analysis is carried out with the following steps—

- Firstly the 'Select' category is dealt with in certain columns by replacing it with other values.
- Next the columns with single values are dropped as they do not add any value to our data.
- Thirdly, we check the percentage of missing values in the columns. The columns with null value percentage>=45 are dropped. Other columns with missing values lesser than 45 are imputed with mode or the rows having missing values are dropped.
- Categories with less count in columns are clubbed together into a single category in certain columns.

Step 4: Further we carry out data visualization of categorical and numerical columns using countplots and boxplots. Outlier treatment is carried out for numerical columns.

Step 5: After completion of the EDA, we prepare our data for model building in the following steps.

- Binary data containing Yes and No are mapped to 1 and 0.
- Dummy variables are created for the required categorical columns and the first column is dropped.
- Dummy variables dataframe is combined with the original dataframe and the original column are dropped as they become redundant.

- The ‘Converted’ column is put into a dataframe ‘y’ and the remaining columns in X.
- The data is now divided into Train and Test data(X train, X test, y train and y test) in 70-30 ratio.

Step 6: Now we proceed with **Model Building**.

- Features are selected using Stats Model and RFE.
- After looking at the detailed summary and the p values of columns, columns with high p values are dropped one at a time.
- After dropping columns based on p value, we have a look at the VIF values and drop any more required columns.

Step 7: After Model building we proceed to making predictions on train set using the final list of columns we have. After predictions, we calculate various metrics required for model evaluation.

Step 8 : Now a ROC curve is made to find an optimal cut off point of probability.

Step 9: A dataframe with various probabilities is prepared and for each probability accuracy, sensitivity and specificity is checked.

Step 10: A curve with accuracy, sensitivity and specificity of various probabilities is plotted and the point where these curves intersect is chosen as the optimal cut off probability.

Step 11: Now final predictions are made using this point obtained above and a Lead Score is assigned and all of these is stored in a dataframe. For this dataframe, a confusion matrix is developed and scores like accuracy, Recall, Precision, TP, TN etc. are calculated.

Step 12: Finally predictions are made on the Test set to get to know how the model is performing on the unseen data. Similar process is carried out on test set and a dataframe with original converted variable, probability calculated for conversion, lead score and final prediction made by model is framed. Similar metrics like accuracy, precision, recall etc. are calculated for the above dataframe and a comparison is made with the performance of model on test data with respect to the performance of model on train data.

Conclusion: It was found out that the variables that mattered the most in the potential buyers are:

○ Total time spent on the website.

○ When lead sources are: ■ Google ■ Direct Traffic

○ When the lead origin is lead add format

○ When their current occupation is as a working professional.