Summary of lead score Logistic regression model creation:

According to the problem statement we have to find the hot lead to increase the converted to 80%, to achieve that we have created a model in logistic regression model with the past history data (given lead data). By the help of data dictionary found the importance of each column for example the country and city column would not make any difference in lead so we neglected that columns and also columns which looks not important for the model.

Later the data cleaning process there were multiple columns which had more than 40% null value but there were some importance columns like as Asymmetrique profile score and Asymmetrique activity score to secure the columns we had to delete the rows which had null values though it was huge number this step had to be done. We had deleted the rest of the unimportant columns.

Outlier treatment had done with the help of boxplot and percentile values of the columns. Dropped the column which had more value than 99^{th} percentile to keep the uniformity of the data. Next for the data preparation part we did scaling with the help of standard scaler from sklearn library and dummy variable were created with the help of pandas dummy variable. Mapped all classification columns to numeric (yes to 1 and no to 0).

Train test split was 70-30 split. Checked for correlation to avoid multicollinearity, we identified that there were multiple columns only with the value of 0, removed those columns also there were columns with more correlation removed that also.

Finally, the data was ready with the balance of 37 and 73 percentage respectively. Logistic regression is available in both statsmodel and sklearn library, first we proceeded with all columns in statesmodel but recognised p value was more in most of the case.

So, with the help of RFE in states model we found the top important columns. First tries top 20 columns but that also had some more p values, moved with top 15 important columns which helped to build a good model also took the help from VIF to make it more perfect. Taken boundary as VIF<5 and p<0.05.

Model had calculated the probability of lead conversion the main part in the model building was to find the cutoff of probability, first tried with arbitrary

cutoff of 0.5 and checked the model accuracy which was 79.8%. ROC curve was also good it is towards y axis, if it is hugging the x axis and has area which is greater than 0.5 means the model is decent. Our model has the area of 0.88 it also indicates the model is a good model.

There are some techniques to find the cutoff probability, found 'probability', 'accuracy', 'sensitivity', 'specificity' of the model with the help of confusion matrix, tried finding all of the value for the cutoff 0.1 to 0.9 which shows 0.4 as good cutoff, plotted curve for accuracy sensitivity and specificity for various probabilities which also indicated 0.38 was good cut off, proceeded with both the cutoff and fond the accuracy and all other values.

Precision and Recall values had calculated for multiple probabilities and the cure with the value also gave cutoff close to 0.4.

The final train model values,

Accuracy – 80.3%

Sensitivity – 78.1%

Specificity – 81.59%

False positive rate – 18.4%

Positive predictive value – 71.9%

negative predictive value – 86%

precession – 71.9%

Recall – 78.1%

The final values of test model,

Accuracy – 79.6%

Sensitivity – 74.1%

Specificity – 82.8%

False positive rate – 17.17%

Positive predictive value – 71.32%

negative predictive value – 84.72%

both train and test set of data has decent values, so the model we built was a good model with the help of our model we can find the hot lead , once we get the hot lead we can put effort in converting them to achieve the given target.