LEAD SCORE CASE STUDY SUMMARY

Problem Statement

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

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 have a lower conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

Solution Summary

Step 1: Reading and Understanding Data
Reading the data from the given Leads.csv file into the jupyter notebook as Pandas Dataframe.

Step 2: Data Cleaning and Analyse
By checking and analysing the variables and dropping those which exceeds 30% null value threshold.
The remaining low threshold null values are to be imputed depending on numerical and categorical we use mean, median and mode respectively.
In this step, there were few other variables that are remove based on significance of the variable to the data.
The outliers were also identified and were dealt with.
The analysis was done with taking converted variable as target, and input variables as one to one and one to many.

Step 3: Creating Dummy variables
This step involves creating dummy variables for categorical features for which we can maintain the data as binary state (0 or 1).
This step is crucial as we have to analyse the correlation between the target and the other variables.

Step 4: Test Train Split and Scaling
This step used to split the data set into test and train sections with a proportion of 30% and 70% respectively.
After the split is done, we can do scaling based on the requirement we are looking for. In this case, we need to build a model on train data split, and the values on numerical features are to be scaled transform.

Step 5: Model building using Recursive Feature Elimination (RFE)
Using RFE to take important features automatically. And by using statistics generated from the the logistic regression class, we try to remove the feature which has high p-value and this step is done recursively, one by one, to make sure, we get the model with significant features and their standard error.
Used Binomial feature selection for the above process. And Finally, we concluded at 12 most important features for which the Variance inflation Factor (VIF) of all the features are under 5% cut-off threshold.
Then by creating a data frame having the converted probability values and with initial assumption that a probability value of more than 0.5 means 1 or else 0.

Based on above assumption, we derived confusion metrics and calculated overall accuracy of the model.
By calculating Sensitivity , specificity and recall, we made sure that the model is reliable and understandable.

Step 6: Plotting The ROC curve
By using ROC formula to plot the features and the curve came out to be decent with an average area coverage of 88% which further solidified our model accuracy and the optimal cut off was founded at 0.33.

Step 7: Finding the Optimal Cut-off Point
By plotting probability graph for accuracy, sensitivity and specificity for different cut off values. The intersecting point was founded at 0.33.
Based on new value, we observed that close to 80% values were predicted by the model.
We also calculated the accuracy, sensitivity and specificity along with recall to be having in the range of 79 to 82 % which is pretty good for our model.

Conclusion

The final learning on the model, the precision and recall metrics values came out to be at 79 and 80% respectively on the train data set. So moving ahead with test data set, we implemented the model and found that the conversion rate based on sensitivity and specificity along with recall metrics are found out to be at 80.5% accuracy hit.