Lead Score Group Case Study Summary

Problem Statement:

- X Education is a Ed Tech firm which sells online courses to professionals across industry. They need help in selecting the most promising leads, i.e. the leads that are most likely to become into paying customers in future.
- The CEO needs to come up with a model in which you assign a score to each of the leads such that the customers having higher lead score have a higher conversion chance and the
- customers with lower lead score have a lower conversion chance.
- The CEO, has given a ballpark of the target lead conversion rate to be ~80%.

Steps Taken:

Step 1 – *Reading and understanding the data.*
Step 2 – *Data Cleaning* : Dropping the variables which have high percentage of null values. Imputing the variables which have less percentage of null values by mean or median. We have also treated the outliers.
Step 3 – EDA (*Exploratory Data Analysis)* : Plot the various graphs to understand relationship distribution and relationship between target variable.
Step 4 – *Train Test Split* : Divide the dataset into train & test set in 70 : 30 split.
Step 5 – *Scaling* : We scale numerical variables by using MinMaxScaler to make them on same scale
Step 6 – *Feature selection using RFE* : By using Recursive Feature Elimination method we selected top 20 features. We build model with these 20 features & eliminated the features having insignificant p value and high VIF.
Finally, we got 15 most significant features. Building on this model we derived the confusion matrix & calculated metrics for model assessment such as accuracy, sensitivity and specificity. We derive the confusion matrix because it allows us to calculate the ratio of accurate predictions (both True Positive and True Negative) as well as inaccurate predictions (False Positive and False Negative)
Step 7 – *Plotting ROC curve* :We can say that our model is a good one as the plotted line is pulling more towards the top left of the plot, maximizing the area under the curve and the line is rising faster as well
Step 8 – *Optimal Cutoff point* : On plotting the graph of accuracy, sensitivity & specificity we found that all these lines meet at a common point. This point is the optimal cutoff point for us which is 0.35.
Using this cutoff point, we again calculated the accuracy, sensitivity & specificity.
After that we also calculated precision & recall metrices.
Step 9 – *Prediction on test set* : We implemented the same model on test set and calculated accuracy, sensitivity & specificity.

Conclusion:

While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.

Accuracy, Sensitivity and Specificity values of test set are around 80%, 81% and 80% which are approximately closer to the respective values calculated using trained set.

Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%

Hence through the overall process we derive a model where the objective of solving the business problem is successfully fulfilled