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 has given a ballpark of the target lead conversion rate to be around 80%

Solution Summary:

Step1: Reading and Understanding Data.

Data id read and analysed.

Step2: Data Cleaning:

Columns with high percentage of NULL values are dropped. The select data ais replaced with NaN. Missing values are imputed with respective median or mode. The outliers were identified and removed. Converted binary vars to 1/0

Step3: Data Analysis

Highly imbalanced with 62 % not convert and 38% convert.

Dropped the columns that as they don't add any much information to model after analysis.

Step4: Data Modelling:

Converted categorical variables to **dummy** variables.

Splitting data into train and test with proportion of 70%-30%.

Rescaling of variables using Min Max Scaling.

Checked **correlation** and dropped variables with correlation higher than 85%

RFE -Using RFE selected 20 top important features.

P-Value-Recursively looked P-Value using statistics to get most significant values. and arrived at 17 variables.

VIF- Calculated VIF variables with higher VIF are dropped.

Final Model with 13 variables with P-values of all variables is 0 and VIF values are low.

Step 5: Prediction:

Created the data frame having the converted probability values and we had an initial assumption that a probability value of more than 0.5 means 1 else 0.

Based on that, derived the Confusion Metrics and calculated Accuracy :80%, Sensitivity :68%, Specificity :88%to understand how reliable the model is.

ROC- Plotted the ROC Curve and the curve came out be pretty decent with an area coverage of 87%.

Optimal Cut-off Point -Plotted the probability graph for the Accuracy, Sensitivity, and Specificity'for different probability values. The intersecting point of the graphs, the optimal probability cut-off point was found out to be 0.37.

Based on the last prediction of conversions have a target of 80% conversion as per the X Educations CEO's requirement. Hence this is a good model. We could also observe the new values of the accuracy=79%, sensitivity=68.8%, specificity=88%'.

Precision and Recall -- Precision and Recall metrics values came out to be 78% and 68.3% respectively on the train data set.

Precision and Recall trade off, - we got a cut off value of approximately 0.42.

Step11: Prediction on Test Set:

Implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 80.6%; Sensitivity=68.35%; Specificity= 87.9%.