Lead score case study

Problem Statement:

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score 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%.

Approach :

In order to find the hot leads i.e the leads who has higher chances of conversion we followed the logistic regression model approach as the target variable given to us was categorical. Below steps were taken to build the effective model:

Loading of the data and knowing the structure of the data.

EDA:

Data cleaning and treatment:

First it was checked if there was any duplicated rows. Then it was seen that the dataset had columns where the value was ‘select’ which was equivalent to null hence converted these values to ‘Null’. Then the null percentage value was checked and the columns have null percentage greater than 45 or equal were dropped off and the columns with null percentage less than 45 were treated. Categorical columns were imputed with the mode value and continuous columns were imputed with median.

Also for the categorical analysis if the columns values had low frequency then they were all clubbed in ‘other’ category to reduce the no of levels and make the analysis simpler. The columns which was imbalanced or where the data was not proportionate it was dropped off.

Numerical analysis was carried out by treating the outliers by drawing boxplots and identifying the outliers

Since the logistic model can be built only with numerical data, we used dummy variable creation.

Heatmap was then plotted to identify the correlation.

After the data preparation and cleaning we moved to the step of model building. We first started with label encoding process where the binary value yes/no were converted into 1's and 0's.Then splitting the dataset into train and test. Then standard scaling was performed to bring all the numerical value on the same scale. To reduce the no of columns used RFE.

Once the model was built, checked the P-value of the feature to ensure whether it's significant or not by using the thumb rule of p<0.05 i.e then we are able to reject the hypothesis .The co-linearity was then checked with the help of VIF and if VIF value was >5 then the columns needs to be dropped.

Using the model we predicted the probability and calculated the score.

Created a dataframe and with then calculated the accuracy, sensitivity, specificity, precision and recall. With the help of sensitivity and specificity curve found the optimal cut-off probability.

Conclusions:

Though we have plotted both sensitivity-specificity and precision-recall but we have considered sensitivity-specificity to calculate the final prediction.

Accuracy,sensitivity and specificity calculated on the trained dataset are 92%,91% and 92% respectively after taking the cut-off as 0.3.

Accuracy,sensitivity and specificity calculated on the trained dataset are 92%,90% and 93% respectively.

Since the values are close we can say that model built is efficient enough.

Learning:

We learnt how EDA plays a important role for creating a model. From cleaning the data, to data imputation, removing null values or treating outliers, all of these are significant in preparing the data for the model.

The creation of dummy variable is also important as the model can be only be prepared using the numerical data thus converting the categorical columns into numerical using dummy variable is significant.

Then using RFE to reduce the number of columns and VIF to observe the co-linearity is also an important step

We also learnt how the confusion matrix can be helpful and the purpose of calculating the accuracy, sensitivity and specificity.

Lastly we learnt about the sensitivity and specificity curve and how it can be used to calculate the optimal cut-off/threshold probability.