Lead Scoring Case Study

Problem Statement:

An education company named X Education sells online courses to industry professionals. The X Education company is looking for the leads who are most likely to get converted 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 a higher lead score have a higher conversion chance and the customers with a 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%.

Solutions:

1. Reading and Understanding the Data:

Reading the data using CSV file and analyzing the data.

2. Data Cleaning: In Data cleaning We dropped the records which contain high Null values which are greater than 40% . In this step we also fill up the required missing values , for numeric variables we take median and for categorical variables we take classification variables. The unwanted data and outliers are removed.

3. Data Analysis: The main step for analysing the data starts from here, We have checked both Univariate Analysis and Bivariate Analysis with the target variable(Converted) . For few variables we have only one value as the highest so dropped such kind of variables.
After basic analysis we can see there are columns which does not provide any data so we have drop ed those values.

4. Data Preparation:

4.1 Creating dummy variables for categorical variables: We have created dummy variables for all the categorical variables.

4.2 Test Train Split: Later, We need to split the dataset into test and train sets, whereas test test should contain 0.3% values.

4.3 Feature Rescaling: Using Standard scaling technique , we complete the statistical view of all the parameters in out model.

4.4 Feature Scaling using RFE: Using RFE(Recursive Feature Elimination) method we have selected 20 features, and build the model.

- Once the first model is build we can observe all the coef values, p-values,and we also calculate the VIF values.

- Based on the p-values which are greater than 0.5 and VIF values greater than 5 we will be dropping one by one by considering the values presented.
- Till we get significant values we keep on dropping the values.
- Finally we look until the p-values and VIF are found good.
- We then 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 the above assumption, we derived the Confusion Metrics and calculated the overall Accuracy of the model.
- We then 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 the above assumption, we derived the Confusion Metrics and calculated the overall Accuracy of the model.

5. ROC Curve

Then we plotted the ROC curve for the features and the curve came out be decent with an area coverage of 89%(0.89) .

6. Optimal Cutoff Value:

Then we plotted a probability graph for finding the best cutoff point. Cutoff value to be found as 0.36

Based on the new value we could observe that close to 80% values were rightly predicted by the model.
We could also observe the new values of the 'accuracy=81%, sensitivity=80.4% (0.804), specificity=82.1% (0.821)

7. Model Evaluation - Precision and Recall: Using the confusion matrix we have calculated the precision and Recall for the training dataset .
Precision is 73.8%(0.738) and Recall is 80.4%(0.804) respectively.

8. Making Prediction's on Test dataset:

Then we started the learning's to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 80.0%(0.809), Sensitivity=79.4%(0.794) and Specificity= 81.8%(0.808).