LEAD SCORING 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 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% . The Model seems to predict the Conversion Rate according to the requirement. The Business can be confident on using this model.

SOLUTION SUMMARY:

STEP 1: Importing data

Uploaded the data.

STEP 2: Data Inspection

Read and analyze the data.

STEP 3: Data Cleaning

This step included imputing the missing value as and where required the median values in case of numerical variables and creation of new classification variables in categorical variables. The outliers are also identified and removed. We also dropped the variables which have high percentage of null values in them.

STEP 4: Exploratory data analysis

Now we started with the exploratory data analysis of the data set to get a feel of how the data is oriented.

STEP 5: Data preparation

Creating dummy variables is a common preprocessing step in machine learning and statistical analysis, especially when dealing with categorical variables. Dummy variables are binary variables that represent different categories or levels of a categorical variable. Each level is assigned a binary value of 0 or 1, indicating its presence or absence in a particular observation. The next step was to divide the data set into test and train section with 70-30% values.

STEP 6: Feature selection using RFE

Using the recursive feature elimination we went ahead and selected the 20 top features.

Using the statistics generated, we recursively tried looking at the P values in order to select the most significant values that should be present and dropped the insignificant values.

After that we arrived at the top significant variables. The VIF's for these variables were also found to be good.

We also calculate the sensitivity and specificity metrices to understand how reliable the model is.

STEP 7: Logistic Regression Model Building

In this we create the models using logistic regressions. Model 9 with 12 variables is our final model in which features are having 0 as p-values and accepted level of VIF.

STEP 8: Making Predictions

It can be clearly seen that our model is good on Specificity(~88%) but lacks on Sensitivity (only~70%) as we have not chosen an optimal threshold.

STEP 9: **Using ROC curve to get the optimal cutoff point**

We then tried the plotting ROC curve for the features and the curve came out to be pretty decent with an area coverage which further solidified of the model.

STEP 10: Model Evaluations

Then we plotted the probability graph for the Accuracy, Sensitivity and Specificity for different probability values.

The intersecting points of the graphs was considered as the optimum probability cutoff point. The cutoff point was found out to be 0.34. Based on the new value we could observe that close to 80% values were rightly predicted by the model.

We observe the new values of Accuracy- 81.0%, Sensitivity - 81.7% , Specificity- 80.6%.

STEP11: Making predictions on Test data set

Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrices found out the accuracy value to be 80.4%.

Sensitivity-80.4% and Specificity- 80.5%