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 predictive and a good interpretable model wherein a lead score is assigned to each of the leads, such that the customers with higher leads score have a higher conversion chance and the customers with lower score have a lower conversion chance.

The CEO, in particular, has given a ballpark of target lead conversion rate to be around 80%.

SOLUTION SUMMARY:

STEP1: Reading and Understanding Data

Read and inspected the data for variables and study the data dictionary.

Step2: Data Cleaning and Transformation

a. First step to clean the dataset we chose was to drop the records with duplicate values.

b. Then, there are few columns with the value 'Select' which means the leads did not choose any given option. We changed those values to 'NAN' Values for missing treatment.

c. We iterated over the entire dataset and removed columns which had only 1 value as it will not help in model building.

d. In an iterative fashion, We dropped the columns having NULL values greater than 35%.

e. We did impute the missing values as and where required with median values in case of numerical variables. In categorical variables where the data was null, we imputed with "Unknown category" for the categorical variables.

Step3: EDA

a. Did basic EDA and Identified very interesting patterns in data.
b. Performed bivariate analysis on categorical columns to see how they vary.
c. Performed bivariate analysis on numerical columns by plotting box plot.
d. The outliers were identified and removed for 3 numerical columns with 99 percentile and above.
e. We did multivariate analysis of all the variables with the target variable. Few categorical variables we removed some categories which had very less data with the Others category. We dropped many categorical columns which had majority of the values as No in the dataset as it will not contribute to model building.
f. Changed the binary no yes variables into '0' and '1'.

Step 5: Dummy Variables Creation:

a. We created dummy variables for categorical variables.
b. For the last time Removed all the repeated and redundant variables.

Step 6: Test Train Split

The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

Step7: Numeric Feature Rescaling

We used the MinMax scaling to scale the original numerical variables. We used fittransform for train data and just fit for test data.

Step8: Model Building

a. Using the RFE method, we went ahead and selected the 20 top important features.
b. We trained the logistic regression model and did the statistical analysis.. Drop the variable which has a high pvalue in the overall features and then train the model again by removing that feature
c. In the subsequent iterations we found that all features have pvalue less than 0.05. hence we started looking at multi collinearity with the VIF method.
d. We dropped the highest value VIN column above 5 and did few more iterations.
e. In the 6th iteration of the model training we found that all features pvalue is less than 0.05 and all features VIF is also below 5.
f. We did the predictions on the training data with a default cut off of 0.5
g. We calculated the confusion matrix to realize that the precision is good at 80%, but the recall is very less around 70%
h. We plotted the ROC curve and came out with a good AUC score of 0.89 which means that the model is good but needs to be balanced for Precision and Recall.
i. We calculated the accuracy metrics(confusion matrix, precision, recall, sensitivity and specificity) for all the probabilities from the range 0.1 – 0.2- 03 and so on till 0.9 to get a sense of when there is a balance between precision and recall.
j. With the above data we calculated the plot for accuracy, sensitivity and specificity and found out that the probability threshold is coming as 0.36
k. We again calculated the accuracy metrics on the training data with probability cut off of 0.36 and found out that the model is having good results for sensitivity and specificity around 80% which is good.
l. We did the predictions on the test set with the conclusion that the model is holding good on training data.
m. We calculated the accuracy metrics on the test data and found out that the model is holding good with the test data as well with both sensitivity and specificity above 81%.
n. We also figured out that the conversion rate with both training and test data is above 80% which meets the objectives set by the Xeducation CEO

Step9: Conclusion

a. The lead - score calculated in the test set of data shows the conversion rate of 81% on final predicted model.
b. Good value of recall of our model will help to select the most promising leads.
c. The model is found good and can be deployed to get the prediction of the hot leads based on the data obtained for each and every lead.
d. The lead score generated for each lead will help the sales team decide whether to pursue and put more efforts on the specific lead, helping the sales team to optimize their efforts and also meet their targets of 80% conversion.
e. Top 3 Features which contribute more towards the probability of a lead getting converted are:
   i. Lead Origin- Lead Add Form
   ii. Current occupation - Working Professional
   iii. Total Time Spent On Website