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

The following approach is followed to solve this case study.

1. **Data Understanding**:
   First step is to load the dataset and check the head of the dataset.
   Next is to check the shape of the dataset to see we had 9240 rows and 37 columns are present in the dataset. Finally find the type of datatypes present in the dataset and convert the datatypes if necessary.
2. **Data cleaning and EDA:**
   Since the dataset had 9000 plus rows, we did look for duplicate values with prospect id and lead number. Then we saw that many of the columns had null values and we dropped those columns where the null value percentage was more than or equal to 40%. Then we did some imputation of the null values for categorical variables with maximum number of occurrences in the column and did a value count check with help of count plot. Then we further dropped 15 columns which were highly skewed and didn't add much value to the model. We also did the numeric variable analysis to check if there were some multicollinearity among the variables using heatmap. We also did some outlier treatment for the numeric columns by capping the outlier data to 95% value to remove the outliers also plotted box plot to check if we can get some inferences to improve the lead conversion rate.
3. **Data Preparation:**
   We changed the categorical variables having yes and no values to 1's and 0's by doing binary map. Then we created dummy variables for the categorical columns and later dropped repetitive columns from the data frame.

4. **Train-Test Split:**
   The data frame had now 60 columns and 9103 rows and we split the data into train and test set in 70: 30 ratio and then performed scaling of 3 numeric columns using standard scaler method.

5. **Model Building:**

We first ran our train model and then did feature selection using RFE by selecting top 15 variables that add value to our model. then we ran the model with 4 iterations until we got low p-value (<0.05) and removed few columns in process which had higher p values and then checked for lower VIF's (<5) which showed there is very less multicollinearity among the variables.

An arbitrary value of 0.5 was selected as the probability that leads could get converted on the final train model. Then we calculated the confusion matrix, and metrics such as accuracy, sensitivity, precision and Recall which gave good results with AOC coming to be around 96%. we then found the converging point as the optimal cut off point to be 0.3 for the model.

Then the model was run again with 0.3 cut-off to get a very good results for all the above metrics and also calculated the precision - Recall trade-off which came to be around 0.38.

This trade off value was used as the cut -off for making predictions on the test set. We then found the accuracy, sensitivity, specificity, precision, recall values for the test set and they came out to be within a very good range.

Then finally we assigned the lead score values to the test set as to indicate if the leads were hot or not.

6. **Final Findings:**

The model we choose has low p-values<0.05 and had very low VIF's <5 meaning very less multicollinearity among the variables.

The overall results on the Test sets seems to be good with accuracy of 91% , Sensitivity: 88% ,Specificity: 93% ,precision :89% ,Recall : 88%

Based on the model we have shortlisted the variables that contribute towards potential lead conversion rate having high positive coefficients:

1.Tags Closed by Horizon.
2.Tags Will revert after reading the email.
3.Tags Busy.