SUMMARY- LEAD SCORING CASE STUDY (DSC 47)

SUBMITTED BY : RICHA RASHMI /HARSHADKUMAR CHAVDA/TEJ VAMSI

Problem Statement : The analysis is done for X Education to find out how we can improve a lead getting converted to enrol into the course.

The dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which has been further analysed through various steps of logistic regression and most impactful variables has been found to improve the conversion rate to 80%.

Step 1 : Date Inspection, Cleaning and preparation

Detailed data inspection was done understanding the business meaning of all the variable. Data consists of 37 variables in total consisting of continuous as well as categorical variables. A lot of variables were not having null records while some of the categorical variables consisted of 'select' in the records. This is equivalent to null, and was handled in the next steps accordingly.

Step 2 : Exploratory Data visualisation

- Handling missing values : Replacing 'select' values in data with np.NaN and then handling the null values were done. Dropped columns where more than 40% null values were found.

- Handling categorical variables : Univariate and Bivariate analysis was done between the categorical variables and continuous variables and Columns with categories having few data are grouped into Others or similar groups

- Handling imbalance data fields : Columns with only few unique category data can be removed
- Handling Outliers : Removed top 1% from **Total Visits** and **Page Views Per Visit** columns
- Mapping categorical variables to integers :Columns with only binary data has been mapped with the help of binary map

Step 3 : Dummy variables creation :

After step 2, below is the list of categorical variables for dummy variable creation:

Lead Origin, Lead Source, Do Not Email', A free copy of Mastering The Interview, Specialization, What is your current occupation

Step 4 : Train-Test Split :A test-train split was done after this step, A 70 - 30 split was done with 70% train data and 30% test data.

Step 5 : Scaling : Standard scaling was performed on continuous variables

Step 6 : Feature Elimination : First using corelation matrix then selecting top 15 using RFE and then Manually (using p-values and VIFs) using RFE

Step 7 : Model Building
Arrived at final model after various iterations, where VIF <3 and p value <0.05

Step 8 : Model Evaluation -

- With 0.5 as cut-off: Accuracy is 79.6 %, Sensitivity and Specificity scores are 64.5% and 88.01 % respectively
- Optimal cut-off using ROC curve : 85% of data is defined under the curve which is significantly good.
  It can be inferred from graph that the threshold value optimised at 0.3 and scores were at Accuracy : 76.9 % , Sensitivity : 76.6 % and Specificity : 77.05 %

Step 9 : Predictions on the test set

1. With threshold of 0.3 on test set we are getting : Accuracy= 77.3%, Sensitivity = 75.02 % and Specificity = 78.6 %

Step 10 : Precision and Recall

2. Threshold with precision and Recall Trade-off is approximated to 0.35 and :Accuracy: 78.74 %, Specificity = 82.02 %, Precision = 71.9 %, Sensitivity/ Recall = 73.5 %

Result : The optimum cut-off was observed to be at 0.35. On train data we achieved an accuracy of 78.74%, Sensitivity of 73.5% and Specificity of 82.02%. For test data we achieved an accuracy of 78.99%, Sensitivity of 71.28% and Specificity of 83.39%. This indicates that the model is able to predict the dependent variable with around ~80% accuracy which meets the required criteria.

Conclusion :

1. The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.
2. The company should make calls to the leads who are the "working professionals" as they are more likely to get converted.
3. The company should make calls to the leads who spent "more time on the websites" as these are more likely to get converted.
4. The company should make calls to the leads whose current occupation was Student, Unemployed, Other as they are more likely to get converted.
5. The company should not make calls to the leads whose Lead Source was "Referral Sites" as they are not likely to get converted.
6. The company should not make calls to the leads who opted for "Do Not Email" as they are not likely to get converted.
7. The company should not make calls to the leads whose Lead Origin is "Lead Origin_Landing Page Submission" as they are not likely to get converted.