Lead Scoring Case Study Summary

Problem Statement:

X Education sells online courses to industry professionals. Professionals who are interested in the courses land on their website and browse for courses.

The Education company markets its courses on various sources like google and several other websites. Once these people land on the company website, they might browse the courses or fill form for more details by providing contact details such as Email address or phone number which are classified to lead.

Also, company gets leads through referrals.

Once these leads are acquired, sales team start contacting through calls or emails etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

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%.

Goals of the Case Study:

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot,

i.e., is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Steps used to achieve the goals of the case study.

Data Understanding:

Read and understood the data.

Checked Totals rows and columns (9240, 37)

Checked info and statistical description of data.

There are no duplicate values in Prospect ID and Lead Number.

Data Cleaning:

Clearly Prospect ID & Lead Number are two variables that are just indicative of the ID number of the Contacted People so Dropped both.

In data we saw many ‘select’ values and the reason for that is customer did not select any option from the list, so we changed those values to Nan.

Checked for null values and dropped those columns which has more than 45% null values.

Also treated the remaining null values.

Also, we dropped some imbalances columns such as Country, 'Search', 'Magazine', 'Newspaper Article’, Etc.

Because More than 95% have the same values

Treated outliers on columns like ‘Page Views Per Visit’, ‘Total Visits’ and Etc.

Exploratory data analysis:

Checked the spread of numerical values

used boxplots to check the

Spread of "Total Visits" vs Converted variable

Spread of "Total Time Spent on Website" vs Converted variable

"Page Views Per Visit" vs Converted variable

used heatmap to check the correlations of numeric values

Data Preparation:

Creating Dummy variable and Test train split:

Created the dummy variables for the categorical variables and dropped the original columns, then added the dummies to Master data frame.

splitted the data by using train_test_split function into test and train sections with a proportion of 70-30%.

Feature selection using RFE:

By using the RFE, selected the 15 top important features.

Extracted the list of RFE supported columns.

Model Building:

In our first model we dropped 'Lead Source_Referral Sites' because P- Value is High.

Created Second Model, P values are zero, check Variance Inflation Factor to see if there is any correlation, dropping 'Last Notable Activity_SMS Sent' because of high correlation between the variables.

Then our third model was good, then moved on to derive the Probabilities, Lead Score, Predictions on Train Data.

Model Evaluation:

plotted ROC curve for the features and the curve came out be pretty decent with 0.97 which is very near to 1, so it's indicating a good predictive model

Previously, we had chosen an arbitrary cut-off value of 0.5.

then we calculated accuracy sensitivity and specificity for various probability cut-offs.

after that plotted accuracy sensitivity and specificity for various probabilities.

As per the plot 0.3 is the optimum point to take it as a cut-off probability.

Then we got the following.

We have the following values for the Train Data:

Accuracy: 92.29%

Sensitivity: 91.70%

Specificity: 92.66%

We have the following values for the Test Data:

Accuracy: 92.78%

Sensitivity: 91.98%

Specificity: 93.26%

Comparison between train and test:

	Train	Test
Accuracy	92.29%	92.78%
Sensitivity	91.70%	91.98%
Specificity	92.66%	93.26%
Precision	88.47%	89.15%
Recall	91.69%	91.98%

Summary:

Logistic Regression Model is good and accurate

Education Company needs to focus on following key aspects to improve the overall conversion rate:

a. Increase user engagement on their website since this helps in higher conversion

b. Increase on sending SMS notifications since this helps in higher conversion.

c. Get TotalVisits increased by advertising etc. since this helps in higher conversion

d. Improve the Olark Chat service since this is affecting the conversion negatively.

we can make good calls based on this model