Lead Scoring Case Study

Summary:

The Lead Scoring analysis is in alignment with X Education's requirement and the objective was to find out which attributes lead to get more industry professionals to join their courses. The data used is historical which gives a lot of information about the customers, their behaviour, and conversions.

Below mentioned are the steps used to perform the analysis:

1. Cleaning data:

- Treating NULL values: The most important step of data cleaning is finding NULL values and imputing them, or dropping them, the same has been applied on the provided data set.
- The next step taken was to understand the business description behind each column and drop if any unnecessary columns
- Categorical data with high level of distinct values were grouped together

2. EDA:

- An Exploratory Data Analysis was done to analyse the data using visual techniques and how they react to target field (Converted). It was used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.
- It was found that a lot of attributes in the categorical variables were irrelevant.
- The numeric values had outliers which were handled.

After validating Bar plots for all the categorical values:

- ➢ Lead Source: google and Google can be merged
- ➢ Country can be segregated as India, NA and US and Others, there are many similar fields
- ➢ Field with one Unique value can be dropped as no one is interested for these offers and this might make the model skewed

Observations:

- ➢ Most converted leads are originated from Lead add and landing page submission
- ➢ Most Conversion has happened from Google, Direct traffic and reference
- ➢ Leads don't want calls or mails
- ➢ Most converted are unemployed and looking for better career prospects
- ➢ Most leads didn't see any adds in search, magazine, Education forum, newspaper or digital forum

3. One Hot encoding and Scaling:

The dummy variables were created and NA were removed
For numeric values we used the MinMaxScaler.

4. Train-Test split of Data:

The split was done at 70% and 30% for train and test data respectively.

5. Model Building:

- Firstly, RFE was done to attain the top 15 relevant variables.
- Secondly VIF values and p-value were used to remove few more variable which were creating influence (The variables with VIF < 5 and p-value < 0.05 were kept).

6. Model Evaluation:

- Confusion matrix and ROC was created for train Data, then calculation was made to find out the accuracy, sensitivity and specificity which came to be around

Train set:
Accuracy: 0.8630179344465059
Sensitivity: 0.7821939586645469
Specificity: 0.9144736842105263

Test set:
Accuracy: 0.8647186147186147
Sensitivity: 0.7923444976076555
Specificity: 0.9085118702953098

7. Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.6 with accuracy ~86%.

8. Conclusion:

From EDA the attributes which is of great interest

- Most converted leads are originated from Lead add and landing page submission
- Most Conversion has happened from Google, Direct traffic and reference
- Leads don't want calls or mails
- Most converted are unemployed and looking for better career prospects
- Most leads didn't see any adds in search, magazine , Education forum, newspaper or digital forum

It was found that the attributes that mattered the most in prediction are as follows

- Total Time Spent on Website
- Last Notable Activity Modified
- Last Activity_SMS Sent
- Interested in other courses
- Closed by Horizzon
- Ringing
- Already a student
- Lost to EINS
- Welingak Website
- Not doing further education
- Interested in full time MBA

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.