Lead Scoring Case Study

Problem Statement:

An education company named X Education sells online courses to industry professionals. Now, although X Education gets a lot of leads, its lead conversion rate is very poor.

We have been asked to build a model wherein we need to assign a lead score 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 expected rate of conversion after building the model is around 80%.

Technical Approach:

We followed the below steps in order to build a model which would identify the key factors that would lead to lead conversion:

1. Data Inspection and Cleaning – Null value treatment, dropping irrelevant columns, removing outliers etc.
2. EDA
3. Model building – Data preparation, modeling
4. Model Evaluation
5. Model Testing

The dataset provided to us had over 9000 rows and 37 columns to begin with. Here, we observed that quite a few columns are almost entirely biased towards one value – thus not providing any intelligent insight to our study. So, we dropped these columns.

Few others had a large percentage of null values which were either blank or with the option ‘Select’ denoting that no value had been selected for the particular attribute, so we dropped those columns as well. The columns with lesser percentage of null values were treated appropriately by replacing with mode, custom values etc.

Outliers were treated by using the IQR method.

We then performed Exploratory Data Analysis on the data set to obtain trends and relationships between the variables. We observed that the dataset was biased almost entirely towards Indian clientele, so we decided to ignore the Country variable and use the City variable for our analysis instead which contained more relevant data. Unemployed people showed good conversion rate followed by working professionals and students.

Post EDA, we proceeded to create dummy variables for all the categorical attributes in the dataset and dropped the original columns. We then split the data into train and test in 7:3 ratio. Standardization of the variables was done using **StandardScaler**.

The data set had over 100 variables after creating dummy variables, so we used RFE to choose the most relevant of them from the available columns. We created multiple models (3) choosing different number of variables each (25, 20 and 15 respectively). We checked for the p-value (<0.05) and VIF score (<5) in each of the models.

We then began with evaluation of the different models. We computed the confusion matrix and calculated accuracy, sensitivity, specificity, etc. for each model. After comparison, we observed that specificity and sensitivity is almost similar for all 3 models while converting the continuous value of the predicted variable to a categorical variable.

Result analysis for cut-off=0.5 for classifying the continuous variables was obtained as below:

- Sensitivity and True Positive rates of the models are in decreasing order:

    1. Model with 25 variables: 84%
    2. Model with 20 variables: 81%
    3. Model with 15 variables: 73%

- Specificity is almost the same for all 3 models, in slightly increasing order: 96%, 97%, 98%.
- False positive rates of the models are in decreasing order:

    1. Model with 25 variables: 3.2
    2. Model with 20 variables: 2.3
    3. Model with 15 variables: 1.2

On plotting the **ROC curve,** we observed the area under the curve was 0.95 or above in each case which confirmed our approach.

We checked for the **optimal cut-off value** in the range of 0.0 to 1 and we observed that the most optimal value for accuracy, specificity and sensitivity for the three models were 0.3,0.3 and 0.2.

After evaluation, we concluded that model-25 variables and model-20 variables ended up being the best models for prediction but since the number of variables were lesser in **Model-20 variables, we selected this model.** This was based on the fact that the model would consume less time in production and would reduce the chances of overfitting (if there is even the slightest of the chances) since the number of variables are lesser compared to Model-25 variables.

Running this model on the train data, the values obtained were:

Confusion matrix:

Actual/Predicted	0	1
0	3653	349
1	161	2305

*Accuracy:**0.92***

*Sensitivity:**0.93***

*Specificity:**0.91***

And this model on the test data gave us below values:

Confusion matrix:

Actual/Predicted	0	1
0	1521	156
1	53	1042

*Accuracy:**0.92***

*Sensitivity:**0.95***

*Specificity:**0.90***

Conclusion:

After testing we concluded that the **model - 20 variables** will not only take care to not miss the Leads having the potential to get Converted, but will also save money by not predicting falsely - the Leads which do not have the potential to be Converted.

Inferences:

The variables in our model which contribute most towards the probability of a lead getting converted are as below (in ascending order i.e. the strongest predictor first)

- 0.6311 (Tags_Lost to EINS)
- 0.5537 (Tags_Closed by Horizon)
- -0.5211 (Tags_wrong number given)
- -0.5098 (Tags_switched off)
- -0.4968 (Tags_Ringing)
- -0.4938 (Tags_number not provided)
- -0.4914 (Tags_invalid number)

- -0.4413 (Tags_Diploma holder (Not Eligible))
- 0.4294 (Tags_Will revert after reading the email)
- -0.4158 (Tags_opp hangup)
- -0.4052 (Tags_Not doing further education)
- -0.3951 (Tags_Interested in full time MBA)
- -0.3765 (Tags_Already a student)
- -0.3637 (Tags_Interested in other courses)
- -0.3497 (Tags_Graduation in progress)
- 0.3117 (What is your current occupation_Unemployed)
- 0.3053 (Lead Source_Welingak Website)
- 0.2995 (What is your current occupation_Working Professional)
- 0.1572 (Last Activity_SMS Sent)
- -0.1066 (Last Notable Activity_Modified)