Lead Scoring Case Study

Deepthi Yalavarthi | Muthuraja Sivanantham | March 08, 2021

Problem Statement:

- An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted

Objective:

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

Analysis Approach:

The following are the steps that is taken to achieve the above mentioned ask by the Customer (X Education)

Step 1: Data Collection – Collecting the dataset in **csv** format, read and understand it and found there are **9240 rows** and **37 columns**

Step 2: Data cleaning – Performed following steps as part of Data Cleaning activities

- Dropping Variables having more missing values (**>=30%**)
- Imputing Variables having less missing values (**<=30%**)
- Dropping Skewed variables
- Checking null values after dropping and imputing the values in the dataset
- Verifying the Percentage of null values in the variables at last
- Finally we have cleared null values in all the variables and removed the skewed, unique identifier variables as well. After doing data cleaning we are ended up with 13 variables

Step 3: Outlier Analysis and Treatment – Have done Outlier Analysis using **'boxplot'** for all the variables and observed the following. Then the Outliers in the variables have been treated using soft capping technique

- **Observations:**
  - ✓ We don't have any outliers in **'Total Time Spent on Website'** variable but we have in **'TotalVisits'**. We have removed the outliers using soft capping method
  - ✓ Predominantly the outliers are there in the upper range for **'TotalVisits'** variable
  - ✓ After applying the soft capping method, the outliers which is there in **'TotalVisits'** variable has been reduced drastically

Step 4: Data Pre-Processing – The following steps were performed as part of Data Pre-Processing

- ✓ Creating Dummy Variables
    1. Identified the columns with 'Yes' or 'No' values and converted to **1's and 0's**
    2. Checked the other categorical columns and converted to **1's and 0's** based on its values
- ✓ Train-Test Split
    1. Assigned all the independent variables to X and dependent variable (Converted) to y
    2. Have done Train-Test split with train set as **70%** and test set as **30%** and random state as **100**
- ✓ Scaling
    1. Have used Standard Scaler for scaling as it will transform the data such that its distribution will have a mean value 0 and standard deviation of 1
    2. Verified the converted rate and it is **38%**.
- ✓ Looking at Correlation
    1. Based on the heatmap we found that the high correlation was between the following variables and dropped those:
        - Last Notable Activity_SMS Sent
        - Last Notable Activity_Page Visited on Website
        - Lead Origin_Lead Add Form
    2. After dropping the high correlated variables, we had all the variables are correlated as expected. So went ahead with these variables for further processing

Step 5: Model Building:

- ✓ We have performed initial model building and noticed that Significance (p-values) is more for many of the variables
- ✓ Have considered 15 variables for RFE and done the coarse tuning as part of feature selection
- ✓ Done manual tuning (using p-value and VIF), and selected the right features for the model
- ✓ Finally all the variables had a VIF and p-values are within a range

✓ We came up with the below model for further evaluation and testing

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6340
Model Family:	Binomial	**Df Model:**	10
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2742.2
Date:	Mon, 08 Mar 2021	**Deviance:**	5484.4
Time:	16:10:00	**Pearson chi2:**	6.55e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.8938	0.084	-10.617	0.000	-1.059	-0.729
Do Not Email	-1.4108	0.165	-8.540	0.000	-1.735	-1.087
Total Time Spent on Website	1.0946	0.040	27.616	0.000	1.017	1.172
Lead Origin_Landing Page Submission	-0.3690	0.088	-4.216	0.000	-0.540	-0.197
Lead Source_Olark Chat	0.9463	0.119	7.942	0.000	0.713	1.180
Lead Source_Others	1.8674	0.169	11.068	0.000	1.537	2.198
Lead Source_Reference	3.6793	0.238	15.480	0.000	3.213	4.145
Current Occupation_Working Professional	2.7358	0.188	14.543	0.000	2.367	3.105
Last Activity_Olark Chat Conversation	-1.0946	0.167	-6.569	0.000	-1.421	-0.768
Last Activity_SMS Sent	1.2929	0.073	17.792	0.000	1.150	1.435
Last Notable Activity_Modified	-0.9113	0.078	-11.644	0.000	-1.065	-0.758

	Features	VIF
9	Last Notable Activity_Modified	1.65
2	Lead Origin_Landing Page Submission	1.63
3	Lead Source_Olark Chat	1.60
7	Last Activity_Olark Chat Conversation	1.55
8	Last Activity_SMS Sent	1.45
1	Total Time Spent on Website	1.29
5	Lead Source_Reference	1.23
6	Current Occupation_Working Professional	1.18
0	Do Not Email	1.13
4	Lead Source_Others	1.04

✓ We have calculated Accuracy score (**80.69%**), Sensitivity (**68.88%**), Specificity (**88.09%**), False Positive Rate (**11.90%**), Positive Predicted Value (**78.37%**) and Negative Predicted Value (**81.88%**) for train set

Step 6: Plotting ROC Curve – We have plotted ROC and the Area under curve is **0.87**

Step 7: Finding optimal probability cut-off (Lead Score):

- ✓ Optimal cut-off point based on the graph would be **0.38**. So leads having probability conversion more than **38%** are considered as **hot leads**
- ✓ We have also done the Trade-off between Precision and Recall
- ✓ We have again calculated Accuracy score (**79.05%**), Sensitivity (**83.89%**), Specificity (**76.03%**), Precision (**78.37%**) for train set after ROC and Optimal Cut-off

Step 8: Model Evaluation and Model Performance:

- ✓ Made the predictions on the test set
- ✓ Created new column 'Final_predicted' with 1 if probabilities predicted by the model > 0.4 else 0
- ✓ We have then calculated Accuracy score (**80.35%**), Sensitivity (**76.13%**), Specificity (**82.75%**), Precision (**71.57%**) for test set and found that it is almost close to train set

Step 9: Final Observation:

	Train Data	**Test Data**
Accuracy	**79.05%**	**80.35%**
Sensitivity	**83.89%**	**76.13%**
Specificity	**76.03%**	**82.75%**
Precision	**78.37%**	**71.57%**

- **Final Conclusion:** The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model performance