Lead Scoring

CASE STUDY AND ASSIGNMENT

Ankur Singh & Pratik Chouhan | Machine Learning I | 12 July 2022

Summary

Problem Statement

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, but its lead conversion rate is very poor, around acquired 100 leads in a day, and only about 30 of them are converts.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads' whose conversion rate should be higher than 80%.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business objective

- X education wants to know the most promising leads.
- For that, they want to build a model which identifies the hot leads.
- Deployment of the model for future use.
- The model should have an accuracy of more than 80%.

Goal

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 Involved to achieve the goal:

Step 1: Reading and Understanding Data.

Read and analyze the data.

Step 2: Data Cleaning.

We analyze the variables that had a high percentage of NULL values in them. We dropped the columns having only one unique value and columns with highly imbalanced data. Further imputed missing values proportionately or median and created new categories/classifications for categorical columns. The outliers were identified and capped at the 99 percentiles.

Step 3: Exploratory Data Analysis

After Analysis following insights were made:

- Unemployed leads are the highest chance to get converted.
- Most leads are generated in Mumbai and Thane and have a higher conversion rate.
- Management courses are the better option for lead conversion. Financial, marketing, and HR management courses have the highest lead generation and conversion.
- Last Activity Email and SMS have the highest conversion rate.
- Reference has the highest generated to the converted ratio in Lead Source

Step 4: Data Preparation

- Numerical Variables are Normalized using binary mapping to 0 and 1 and MinMaxScaler is used.
- Dummy Variables are created for categorical variables.
- Regression is performing a train-test split, we have chosen a 70-30 ratio with a random state of 100 after splitting data into train and test datasets.
- RFE (Recursive Feature Elimination) is used for Feature Selection with 15 variables.

Step 5: Model Building and Evaluation

- Model is built by removing the variable whose p-value is greater than 0.05 and VIF value is greater than 5.
- After Choosing an arbitrary cut-off probability point of 0.5, we plotted the ROC curve for the features and the area covered under the curve is 89% which solidified the model.
- We plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' to find the optimal probability cut-off point. The cutoff point was found to be 0.36. Based on this, the Accuracy: 80.67%, Sensitivity: 78.63%, and Specificity: 81.93%.
- The Precision and Recall metrics values came out to be 78.88% and 69.67% respectively on the train data set. Based on the Precision and Recall tradeoff, we got a cut-off value of approximately 0.41.

Step 6: Making Predictions on the Test Set

- We implemented the learnings to the test model with an optimal Cut-off of 0.36 and found the following value:
    - Accuracy: 81.13%
    - Sensitivity: 79.45%
    - Specificity: 82.23%
    - False Positive Rate: 17.76%
    - Positive Predictive Value: 74.49%
    - Negative predictive Value: 85.97%

Conclusion

- Sample Final Table

	lead_number	Converted	prob_lead_conversion	final_Prediction	lead_score
0	4269	1	0.544179	1	54
1	2376	1	0.948081	1	95
2	7766	1	0.925650	1	93
3	9199	0	0.116594	0	12
4	4359	1	0.834938	1	83

- Final Model Summary and VIF

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6456
Model Family:	Binomial	Df Model:	11
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2664.2
Date:	Sun, 12 Jun 2022	Deviance:	5328.4
Time:	13:08:05	Pearson chi2:	6.85e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.4183	0.093	-26.090	0.000	-2.600	-2.237
do_not_email	-1.3753	0.165	-8.324	0.000	-1.699	-1.051
total_visits	1.0050	0.205	4.905	0.000	0.603	1.407
time_on_website	4.4952	0.164	27.350	0.000	4.173	4.817
lead_origin_Lead Add Form	4.0393	0.199	20.258	0.000	3.649	4.430
lead_source_Olark Chat	1.5757	0.116	13.632	0.000	1.349	1.802
last_activity_Olark Chat Conversation	-1.1825	0.167	-7.074	0.000	-1.510	-0.855
last_activity_Others	-1.5809	0.439	-3.603	0.000	-2.441	-0.721
last_activity_SMS Sent	1.2837	0.074	17.383	0.000	1.139	1.428
occupation_Working Professional	2.5234	0.187	13.511	0.000	2.157	2.889
reason_for_course_selection_not provided	-1.1784	0.086	-13.667	0.000	-1.347	-1.009
last_notable_activity_Unreachable	2.0033	0.539	3.716	0.000	0.947	3.060

	Features	VIF
2	time_on_website	1.89
1	total_visits	1.88
4	lead_source_Olark Chat	1.55
7	last_activity_SMS Sent	1.46
9	reason_for_course_selection_not provided	1.45
5	last_activity_Olark Chat Conversation	1.42
3	lead_origin_Lead Add Form	1.28
8	occupation_Working Professional	1.17
6	last_activity_Others	1.16
0	do_not_email	1.07
10	last_notable_activity_Unreachable	1.00

- Final model Features

Features	Co-efficient	Impact
time_on_website	4.495224	Positive
lead_origin_Lead Add Form	4.039318	Positive
occupation_Working Professional	2.523387	Positive
last_notable_activity_Unreachable	2.003294	Positive
lead_source_Olark Chat	1.575666	Positive
last_activity_SMS Sent	1.283716	Positive
total_visits	1.005008	Positive
reason_for_course_selection_not provided	-1.17837	Negative
last_activity_Olark Chat Conversation	-1.18247	Negative
do_not_email	-1.37531	Negative
last_activity_Others	-1.58092	Negative
const	-2.41828	Negative

- **Final Formula for Logistic Regression Model is**

  ln (p/(1-p)) = -2.418279 + time_on_website*4.495224 + lead_origin_Lead Add Form*4.039318 + occupation_Working Professional*2.523387 + last_notable_activity_Unreachable*2.003294 + lead_source_Olark Chat*1.575666 + last_activity_SMS Sent*1.283716 + total_visits*1.005008 + reason_for_course_selection_not provided*-1.178366 + last_activity_Olark Chat Conversation*-1.182474 + do_not_email*-1.375312 + last_activity_Others*-1.580917

- The top 3 variables that contribute to lead getting converted in the model are:
  - Total time spent on the website
  - Lead Add Form from Lead Origin
  - Having occupation as Working Professional

- The Model has an accuracy of >80%, which 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 to get a higher lead conversion rate of 80%.