LEAD SCORING CASE STUDY

SUBMISSION

by,
Rahamtullah Noorbasha
Bhakti kumar Bengani

Objective

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

Actions Performed

- Read the dataset
- Remove Null Values
- Data Preparation
- TEST-TRAIN Split
- Model Building
- Assessing Model
- Checking Accuracy, VIFs
- Metrices beyond simple accuracy
- Plotting ROC curve
- Plotting Accuracy, Sensitivity and Specificity
- Make predictions on Test
- Append Train and Test dataset with Lead Probability
- Create new dataset and classify lead into Hot, Cold and Medium labels

Model Building

- Created Dummy variables for categorical variables and removoed one column. Instead of removing first column removed the column(cateogory) that has very less values(significance).
- Selected features using RFE and build model.
- Validated the model using statsmodel
- In first fit itself we got all the P values and VIF values are with in the range of good model

Statsmodel stats

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.4603	0.494	2.955	0.003	0.492	2.429
Do Not Email	-1.5767	0.204	-7.743	0.000	-1.976	-1.178
Total Time Spent on Website	1.1503	0.040	28.973	0.000	1.072	1.228
LeadOrigin_API	-1.0782	0.496	-2.176	0.030	-2.050	-0.107
LeadOrigin_Landing Page Submission	-1.2225	0.492	-2.484	0.013	-2.187	-0.258
LeadOrigin_Lead Add Form	3.1684	0.534	5.937	0.000	2.122	4.214
LeadSource_Olark Chat	1.1370	0.117	9.682	0.000	0.907	1.367
LeadSource_Welingak Website	2.2692	1.038	2.187	0.029	0.235	4.303
LastActivity_Email Bounced	-1.5928	0.473	-3.365	0.001	-2.521	-0.665
LastActivity_Had a Phone Conversation	1.4794	0.641	2.306	0.021	0.222	2.737
LastActivity_Olark Chat Conversation	-1.4026	0.190	-7.392	0.000	-1.774	-1.031
LastNotableActivity_Email Link Clicked	-2.0074	0.262	-7.669	0.000	-2.520	-1.494
LastNotableActivity_Email Opened	-1.3967	0.086	-16.326	0.000	-1.564	-1.229
LastNotableActivity_Modified	-1.7568	0.094	-18.714	0.000	-1.941	-1.573
LastNotableActivity_Olark Chat Conversation	-1.6667	0.369	-4.512	0.000	-2.391	-0.943
LastNotableActivity_Page Visited on Website	-1.5918	0.203	-7.845	0.000	-1.989	-1.194

P values are below .05

VIF

	Features	VIF
2	LeadOrigin_API	3.46
12	LastNotableActivity_Modified	2.84
3	LeadOrigin_Landing Page Submission	2.81
5	LeadSource_Olark Chat	2.43
11	LastNotableActivity_Email Opened	2.31
9	LastActivity_Olark Chat Conversation	1.97
0	Do Not Email	1.90
7	LastActivity_Email Bounced	1.80
4	LeadOrigin_Lead Add Form	1.44
13	LastNotableActivity_Olark Chat Conversation	1.39
1	Total Time Spent on Website	1.31
6	LeadSource_Welingak Website	1.30
14	LastNotableActivity_Page Visited on Website	1.13
10	LastNotableActivity_Email Link Clicked	1.09
8	LastActivity_Had a Phone Conversation	1.01

Even the VIF values are below and well around 3

Lead Probability on train and test set

Train Data

LeadNumber	Lead_Prob
631712	0.621835
618416	0.554525
641905	0.798643
591151	0.030149
616910	0.397511

Test Data

LeadNumber	Lead_Prob
604797	0.067416
656356	0.648546
638583	0.067416
654471	0.891262
609314	0.067416

Lead Score for entire dataset

LeadNumber	Lead_Prob	Lead_score	Lead_type
631712	0.621835	62	Medium
618416	0.554525	55	Medium
641905	0.798643	80	Hot
591151	0.030149	3	Cold
616910	0.397511	40	Cold

- Merged both test and train lead probability data and calcualted scores and classified as hot, medium and cold

```
Cold      5375
Hot       1781
Medium    1670
Name: Lead_type, dtype: int64
```

- Total number of Hot ,Medium or Cold customers are calculated as shown which can be used to take necessary action plan.

Conclusion

- Out of 8826 leads, below is the classification :
  - Cold 5375 (~61 %)
  - Medium 1670 (~19 %)
  - Hot 1781 (~20%)