Lead Scoring – Group Case Study: Subjective Questions

- By Anindo Mazumdar & Mukundan AP

Question 1:

Which are the top three variables in your model which contribute most towards the probability of a lead getting converted?

Answer –
As per our analysis, the following 3 fields contribute the most towards probability of a lead getting converted –

1. *Lead Origin*
2. *Lead Activity*
3. *Current Occupation*

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	5959
Model:	GLM	**Df Residuals:**	5945
Model Family:	Binomial	**Df Model:**	13
Link Function:	logit	**Scale:**	1
Method:	IRLS	**Log-Likelihood:**	-2388.9
Date:	Sun, 03 Mar 2019	**Deviance:**	4777.9
Time:	18:40:22	**Pearson chi2:**	5.80E+03
No. Iterations:	7		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.9979	0.063	-15.786	0	-1.122	-0.874
Do Not Email	-1.5784	0.195	-8.083	0	-1.961	-1.196
Total Time Spent on Website	1.1052	0.042	26.171	0	1.022	1.188
Lead Origin_Lead Add Form	3.6946	0.227	16.304	0	3.25	4.139
Lead Source_Olark Chat	1.4364	0.11	13.094	0	1.221	1.651
Lead Source_Welingak Website	2.4763	1.034	2.396	0.017	0.45	4.502
Last Activity_Had a Phone Conversation	3.3549	1.374	2.441	0.015	0.662	6.048
Last Activity_Olark Chat Conversation	-1.1486	0.178	-6.453	0	-1.497	-0.8
Last Activity_SMS Sent	1.2953	0.079	16.376	0	1.14	1.45
What is your current occupation_Other	-1.18	0.09	-13.107	0	-1.356	-1.004
What is your current occupation_Working Professional	2.5096	0.197	12.764	0	2.124	2.895
Last Notable Activity_Modified	-0.691	0.083	-8.277	0	-0.855	-0.527
Last Notable Activity_Unreachable	1.7599	0.598	2.945	0.003	0.589	2.931
Last Notable Activity_Unsubscribed	1.4167	0.512	2.765	0.006	0.413	2.421

A) **Lead Origin** – Specifically the instances where the lead has filled a form to get more details about the program

B) **Last Activity** – We observed that instances where the lead was engaged in telephonic conversation with the sales team, it resulted in better conversion rate. Also, interesting to note that potential leads who enquired about the program using online chat had negative correlation with the lead conversion, most probably because of the fact that these could be leads just looking at various courses available online and wanted to get some more details about this specific program.

C) **Current Occupation** – We observed that people who are currently working were better leads. This could be due to the flexible nature of online education and time constraint being a working professional, which makes the idea of online course seem better for enhanced future prospects in respective jobs.

Question 2:

What are the top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion?

Answer –

As per our analysis, below fields were marked as top 3 categorical/dummy variables –

A) Lead Origin_Lead Add Form (dummy variable)

B) Last Activity_Had a Phone Conversation (dummy variable)

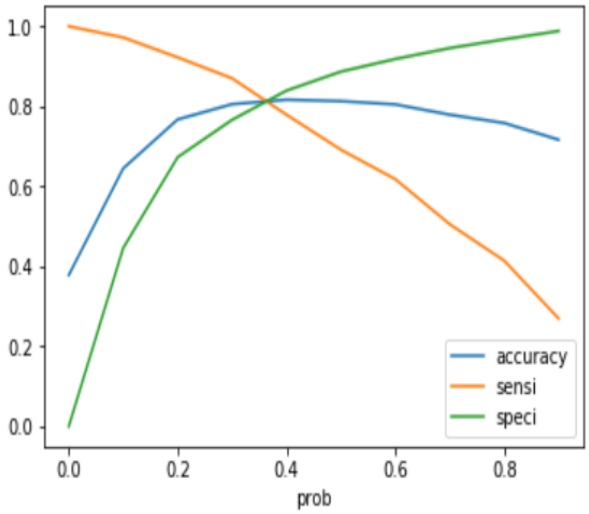
C) What is your current occupation_Working Professional (dummy variable)

Question 3:

X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.

Answer – During our analysis, we came across a few variables that contribute heavily on lead conversion (both positively and negatively). To deploy a strategy for aggressive marketing, suggestion would be to also look at variables which are marked as an important variable for lead conversion but has lower coefficient i.e. lower contribution to the lead conversion model.
Example of such case would be variables like Total time spend on the website, Last Activity SMS sent, Lead source Olark char. All these variables positively contribute to the lead conversion model, however have lower impact on the final conversion.
Bottom line – Also focus on leads with smaller positive coefficients.

Since the company wants to make phone calls as much as possible, so in this case even if we identify some leads which are not going to convert as hot lead that won't make any difference since the company is trying to reach maximum leads as possible. So the company need to increase the cut-off for the model they have built for predicting the hot leads. In this case the model needs to have less false negative count and higher true positive count. So ideally the company should focus on the sensitivity of the model and would prefer a higher sensitivity. The sensitivity will increase with increasing cut-off value. We had the below plot for sensitivity, accuracy and specificity earlier:



prob	accuracy	sensi	speci
0.0	0.378084	1.000000	0.000000
0.1	0.644739	0.972037	0.445764
0.2	0.767075	0.922326	0.672693
0.3	0.805336	0.869951	0.766055
0.4	0.816412	0.778961	0.839180
0.5	0.812888	0.691522	0.886670
0.6	0.804665	0.617843	0.918241
0.7	0.778990	0.505104	0.945494
0.8	0.758181	0.414115	0.967350
0.9	0.716395	0.269419	0.988127

From the above table and plot we can see that for cut-off value around 0.3 we are getting sensitivity of around 87% without compromising accuracy (80%). So the company can predict based on this and this will give them a lot of leads whom they can now contact for trying to convert them to paying customers.

Question 4:

Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage.

Answer - As mentioned in Answer 3, we focused on all the variables with lower positive coefficients for aggressive marketing. On similar lines, in case we want to be conservative about marketing resources, we'll only focus on heavy influencers in the model i.e. focus on people with higher positive correlated variables like working professional etc. and completely exclude people having high negatively correlated variables in scope like do not email etc.

- This would mean that we are significantly reducing the leads by only focusing on high conversion cases and completely disregarding cases where we see high negative correlation in variables for a lead.

- In this case the company doesn't want to make useless phone calls , so the model which is used to predict leads needs to have low FPR(false positivity rate).

- The cut-off values should be high so that false positives are avoided. Since FPR = 1 - Specificity, the Specificity of the model needs to be high in this scenario.

- The specificity of the model will increase with increasing in cut-off value.

For cut-off value 0.6, the specificity is around 0.91 and accuracy is 0.8. So based on this cut-off the company can pursue the identified hot leads (converted value 1) and it will ensure that the company is making as less phone calls as possible.