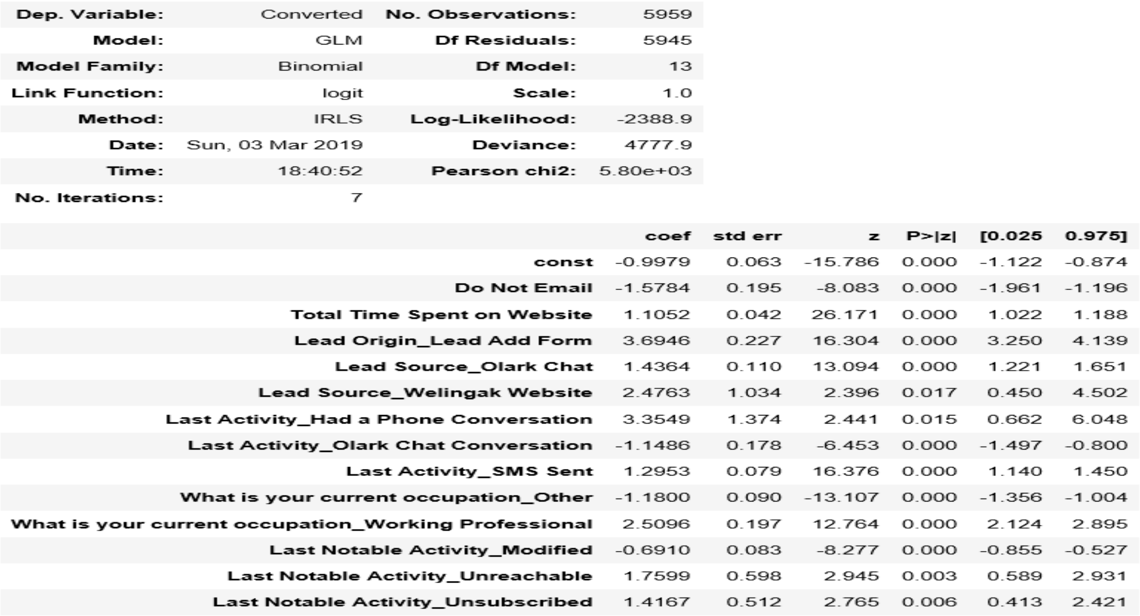
Lead Score Case Study

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



Dep. Variable:	Converted	No. Observations:	5959
Model:	GLM	Df Residuals:	5945
Model Family:	Binomial	Df Model:	13
Link Function:	logit	Scale:	1.0
Method:	IRLS	Log-Likelihood:	-2388.9
Date:	Sun, 03 Mar 2019	Deviance:	4777.9
Time:	18:40:52	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	0.000	-1.122	-0.874
Do Not Email	-1.5784	0.195	-8.083	0.000	-1.961	-1.196
Total Time Spent on Website	1.1052	0.042	26.171	0.000	1.022	1.188
Lead Origin_Lead Add Form	3.6946	0.227	16.304	0.000	3.250	4.139
Lead Source_Olark Chat	1.4364	0.110	13.094	0.000	1.221	1.651
Lead Source_Welingak Website	2.4763	1.034	2.396	0.017	0.450	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	0.000	-1.497	-0.800
Last Activity_SMS Sent	1.2953	0.079	16.376	0.000	1.140	1.450
What is your current occupation_Other	-1.1800	0.090	-13.107	0.000	-1.356	-1.004
What is your current occupation_Working Professional	2.5096	0.197	12.764	0.000	2.124	2.895
Last Notable Activity_Modified	-0.6910	0.083	-8.277	0.000	-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

After application of the logistic regression model on the dataset, the analysis report (as shown above) depicts that the top three variables which contribute the most towards the probability of a lead getting converted are: Lead Origin , Last Activity and Current Occupation of the lead because, these three variables have the highest coefficient.

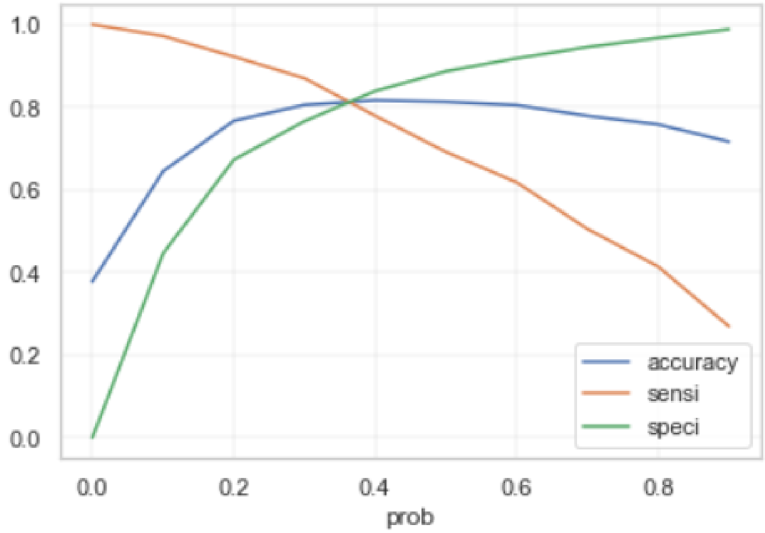
2. **What are the top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion?**

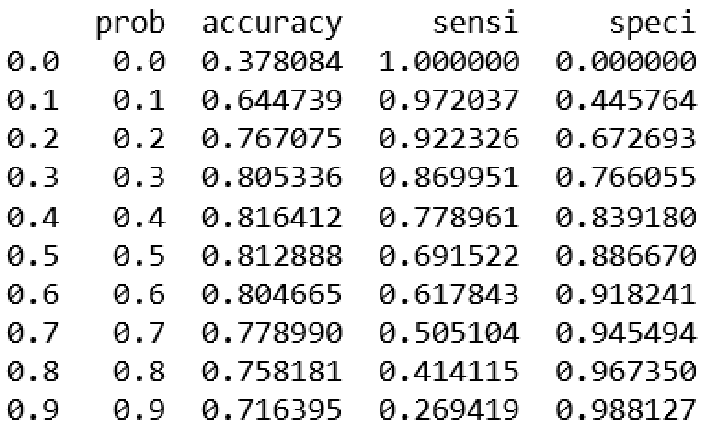
The dummy variables which impact the most on the conversion rate are 'Lead Origin_Lead Add Form' , 'Last Activity_Had a Phone Conversation' and 'What is your current occupation_Working Professional', which are the categorical variables having the highest coefficient.

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.

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 cutoff 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 cutoff value. We had the below plot for sensitivity, accuracy and specificity earlier:





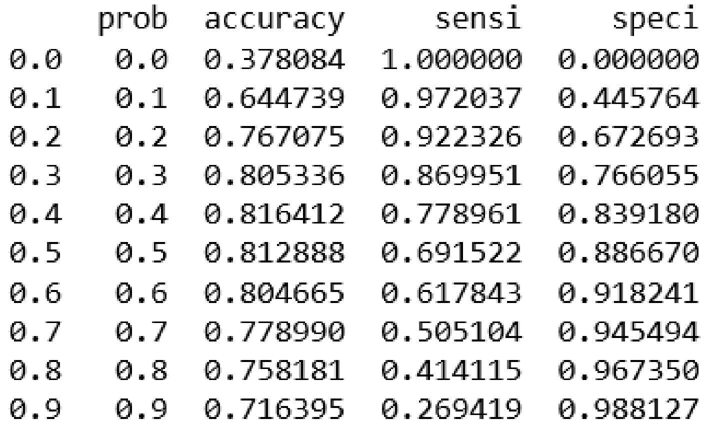
```
     prob  accuracy     sensi     speci
0.0   0.0  0.378084  1.000000  0.000000
0.1   0.1  0.644739  0.972037  0.445764
0.2   0.2  0.767075  0.922326  0.672693
0.3   0.3  0.805336  0.869951  0.766055
0.4   0.4  0.816412  0.778961  0.839180
0.5   0.5  0.812888  0.691522  0.886670
0.6   0.6  0.804665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

From the above table and plot we can see that for cutoff 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.

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.

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 cutoff 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 cutoff value.



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

For cutoff value 0.6, the specificity is around 0.91 and accuracy is 0.8. So based on this cutoff 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.