Questions

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

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?

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.

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.

Answers

Ans. 1- Below is the final model summary statistics.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.0109	0.099	-20.269	0.000	-2.205	-1.816
Do Not Email	-1.3702	0.162	-8.475	0.000	-1.687	-1.053
TotalVisits	0.7342	0.249	2.944	0.003	0.245	1.223
Total Time Spent on Website	4.5904	0.164	27.972	0.000	4.269	4.912
Lead_Origin_Lead Add Form	4.4073	0.216	20.420	0.000	3.984	4.830
Lead_Source_Olark Chat	1.7114	0.126	13.564	0.000	1.464	1.959
Lead_Source_Welingak Website	2.5057	1.030	2.433	0.015	0.487	4.525
Last_Activity_Olark Chat Conversation	-1.1268	0.173	-6.514	0.000	-1.466	-0.788
Last_Activity_SMS Sent	1.2390	0.074	16.852	0.000	1.095	1.383
Specialization_Unknown	-0.6704	0.089	-7.532	0.000	-0.845	-0.496
Last_Notable_Activity_Email Link Clicked	-0.5741	0.248	-2.319	0.020	-1.059	-0.089
Last_Notable_Activity_Modified	-0.8076	0.079	-10.230	0.000	-0.962	-0.653
Last_Notable_Activity_Page Visited on Website	-0.4502	0.197	-2.284	0.022	-0.837	-0.064
Last_Notable_Activity_Unreachable	1.7645	0.498	3.545	0.000	0.789	2.740

To determine the top three most contributing variables in the logistic regression model, we can examine the absolute values of the coefficients. The variables with larger absolute coefficient values have a stronger influence on the predicted outcome. Here are the variables with their corresponding coefficients:

1. Total Time Spent on Website: 4.5904
2. Lead_Origin_Lead Add Form: 4.4073
3. Lead_Source_Welingak Website: 2.5057

Ans. 2- As per the summary statistics of the final model 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 are:

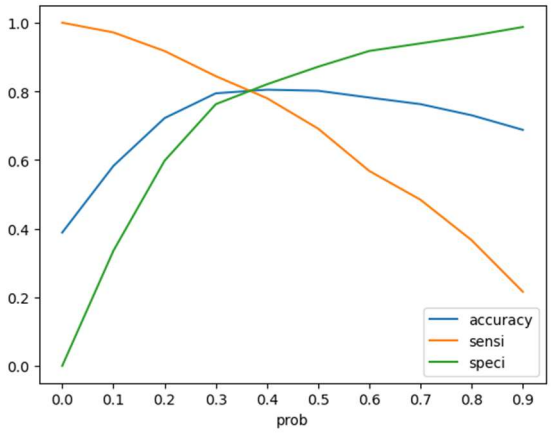
Lead_Origin_Lead Add Form (Coefficient: 4.4073)
As per data dictionary, these are the leads that have filled the application form. Hence, leads for which the application form is filled should be considered a good potential lead.

Lead_Source_Welingak Website (Coefficient: 2.5057)
As per data dictionary, these are the leads that have come in from Welingak website. Hence, leads that come in from this website should be considered as a good potential cause apparently this website could be where people are looking for courses.

Lead_Source_Olark Chat (Coefficient: 1.7114)
As per data dictionary, these are leads with Olak chat as their native source. Hence, leads that come in from Olark chat should be catered as good potential. It could be that people are using Olark chat more than any other source to look for such courses.

Ans. 3- The logistic regression model is used to predict the probability of a lead conversion. We then use this probability with a dedicated cut-off to call a lead as converted.
As per analysis of sensitivity, specificity and accuracy of the model for each value of probability cut-off we had the below output.

	prob	accuracy	sensi	speci
0.0	0.0	0.388560	1.000000	0.000000
0.1	0.1	0.582681	0.971628	0.335511
0.2	0.2	0.722000	0.917352	0.597857
0.3	0.3	0.794216	0.844161	0.762477
0.4	0.4	0.804761	0.780016	0.820486
0.5	0.5	0.801566	0.691201	0.871701
0.6	0.6	0.781754	0.567845	0.917690
0.7	0.7	0.762582	0.483964	0.939639
0.8	0.8	0.730149	0.365954	0.961589
0.9	0.9	0.687650	0.215872	0.987458



Now for the 2-month period with interns, i.e., extra workforce.

1. Tune the model such that we have a higher sensitivity = decrease cut-off probability. This will get more hot leads for the sales team to work with since they now have extra workforce. The accuracy will take a hit but should be compensated by the extra workforce.
2. Focus on the factors in a lead as follows:
    a. Total Time Spent on Website - 4.5904
    b. Lead_Origin_Lead Add Form - 4.4073
    c. Last_Activity_SMS Sent - 1.2390
    d. Lead_Source_Olark Chat - 1.7114
    e. Last_Notable_Activity_Unreachable - 1.7645
    f. Last_Notable_Activity_Modified - 0.8076
    g. Lead_Source_Welingak Website - 2.5057
    h. Last_Activity_Olark Chat Conversation - 1.1268
    i. Last_Notable_Activity_Email Link Clicked - 0.5741
    j. Last_Notable_Activity_Page Visited on Website - 0.4502
    k. Specialization_Unknown - 0.6704
    l. Do Not Email - 1.3702
    These are the feature variables for the model building along with their coefficients. As per this, attention should be paid on leads with higher than average time spent on the website, and those who have filled the form, and the

ones that have come in through Welingak website. As per the top 3 features of the model.

3. With the added workforce we can also have prompt follow-ups with leads as well. The preferred communication way as per model features in order of preference are
    a. SMS
    b. Olark Chat conversation
    c. Email
4. Sales team should keep in mind that as per the data, almost all of the leads prefer not to be called. So, there should be an aggressive push on the above-mentioned means of communication as well in the given order of priority to maximize conversion.

Ans. 4- The period wherein the sales team wants to avoid phone calls, there are other measures communication that the team can resort to. This was analyzed at the time of feature variable list selection using RFE.
The means of communication that the team should resort to in order of priority are as follows:

- SMS
- Olark Chat
- Email