Lead Score Case Study
Subjective Questions

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.

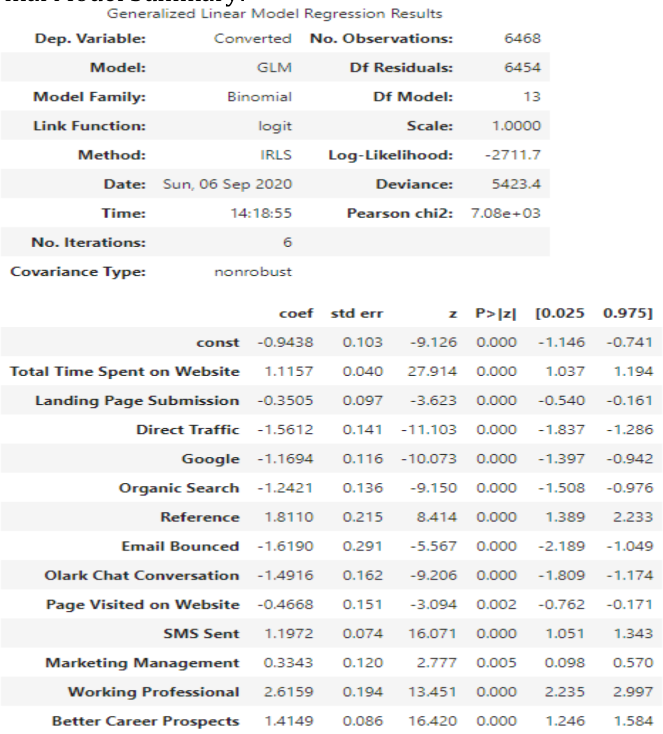
Solutions

Answer1:
The Final model is built using the combination of 'Total Time Spent on Website', 'Landing Page Submission', 'Direct Traffic', 'Google', 'Organic Search', 'Reference', 'Email Bounced', 'Olark Chat Conversation', 'Page Visited on Website', 'SMS Sent', 'Marketing Management', 'Working Professional' and 'Better Career Prospects'.

Following are few metrics which obtained using the Final Model.

- Train Set: Accuracy -> 0.799, Sensitivity -> 0.823
- Test Set: Accuracy -> 0.7832, Sensitivity -> 0.8201

Final Model Summary:



Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6454
Model Family:	Binomial	**Df Model:**	13
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2711.7
Date:	Sun, 06 Sep 2020	**Deviance:**	5423.4
Time:	14:18:55	**Pearson chi2:**	7.08e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.9438	0.103	-9.126	0.000	-1.146	-0.741
Total Time Spent on Website	1.1157	0.040	27.914	0.000	1.037	1.194
Landing Page Submission	-0.3505	0.097	-3.623	0.000	-0.540	-0.161
Direct Traffic	-1.5612	0.141	-11.103	0.000	-1.837	-1.286
Google	-1.1694	0.116	-10.073	0.000	-1.397	-0.942
Organic Search	-1.2421	0.136	-9.150	0.000	-1.508	-0.976
Reference	1.8110	0.215	8.414	0.000	1.389	2.233
Email Bounced	-1.6190	0.291	-5.567	0.000	-2.189	-1.049
Olark Chat Conversation	-1.4916	0.162	-9.206	0.000	-1.809	-1.174
Page Visited on Website	-0.4668	0.151	-3.094	0.002	-0.762	-0.171
SMS Sent	1.1972	0.074	16.071	0.000	1.051	1.343
Marketing Management	0.3343	0.120	2.777	0.005	0.098	0.570
Working Professional	2.6159	0.194	13.451	0.000	2.235	2.997
Better Career Prospects	1.4149	0.086	16.420	0.000	1.246	1.584

According to our understanding the top 3 variables which contribute most towards the probability of lead getting converted from a statistical point of view would be the ones with the highest coefficient values. Based upon this parameter the Top 3 variables are

1. Working Professional – coefficient 2.62 (approx.)
2. Reference – coefficient 1.81 (approx.)
3. Email Bounced – coefficient -1.62 (approx.)

Note: These results are based on the random state which we choose and might vary if another random state is chosen.

Answer2:

In our model, the answer is the same as the one provided for the previous questions since our top 3 variables are all categorical/dummy variables. As is evident from the model summary, our model is significantly influenced by the categorical/dummy variables with exception to the one numeric variable that is present. In short, our top 3 dummy variables would once again be: -

1. Working Professional
2. Reference
3. Email Bounced

If we were to turn our focus to the original variables from which the above dummy variables were extracted, the following would be the top 3 variables: -

1. Specialization (original categorical variable for Working Professional)
2. Lead Source (original categorical variable for Reference)
3. Last Activity (original categorical variable for Email Bounced)

Also, other dummies created from the categories of Lead Source and Last Activity like Direct Traffic (Lead Source) and SMS Sent (Last Activity) are also present in the Final Model. Hence, we can also say that these categorical variables influence the Probability of Lead Conversion.

Answer 3:

For model evaluation the most common metric that is used is Accuracy. Accuracy of a model is defined as the percentage of correct predictions which the model makes i.e.

1. 0 classified correctly as 0
2. 1 classified correctly as 1

Or in terms of Lead, how good the model is predicting if a Lead is a Lead or not.

Two other key metrics are:-

- Sensitivity - Evaluates the True positive rate or the ratio of Predicted converted Lead vs Actual converted Leads.
- Specificity - Evaluates the True negative rate or the ratio of predicted unconverted Leads vs actual unconverted Leads.

Now, if we want to increase the overall conversion rate of the Model so that once the X Educations sales team involve the additional resources for making the phone calls then we must increase the number of total Leads that are getting predicted to convert. To do this we have to increase the overall sensitivity. In the process of increasing sensitivity which will allow the model to identify a higher number of potential leads for conversion, the specificity will reduce a bit. This is an acceptable tradeoff since the aim here is to get higher conversion rates.

Based on our model, the suggested probability threshold for this requirement would be **0.25** with Sensitivity = 90% (approx) and an Accuracy >75% (approx).

Answer 4:

Once again, we will focus on the 3 metrics that were mentioned in Answer 3 – Accuracy, Sensitivity, Specificity.

Since the company's strategy is to deploy resources only for 'extremely necessary' leads, it would mean that the model should have an improved ability to identify leads that are unlikely to convert. For this strategy to work, we will focus on increasing specificity which will allow the model to better identify unlikely converts and only tag those leads that are significant enough to be followed up on.

Based on our model, the suggested probability threshold for this requirement would be **0.6** with Specificity= 91% (approx) while maintaining an Accuracy >80% (approx). Here, the accuracy of the model will tend towards identifying more insignificant leads so that only the highly likely converts are tagged for potential conversion.