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

Answer:

In [183]:
```
# Top 3 feature contributing the most
pd.DataFrame(feature_importance).reset_index().sort_values(by=0,ascending=False).head(3)
```

Out[183]:

	index	0
8	Tags_Closed by Horizzon	100.00
11	Tags_Will revert after reading the email	80.12
4	Lead Source_Welingak Website	77.06

In [184]:
```
# Summary for the selected logistic regression model (Model - 7)
res.summary()
```

Out[184]:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6363
Model:	GLM	Df Residuals:	6348
Model Family:	Binomial	Df Model:	14
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1533.4
Date:	Sun, 11 Apr 2021	Deviance:	3066.9
Time:	22:29:42	Pearson chi2:	7.01e+03
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.6715	0.309	-15.097	0.000	-5.278	-4.065
Do Not Email	-1.3573	0.223	-6.074	0.000	-1.795	-0.919
Total Time Spent on Website	1.0909	0.054	20.279	0.000	0.985	1.196
Lead Source_Olark Chat	1.1193	0.132	8.460	0.000	0.860	1.379
Lead Source_Reference	2.0237	0.356	5.689	0.000	1.326	2.721
Lead Source_Welingak Website	6.8043	1.027	6.626	0.000	4.792	8.817
Last Activity_Email Opened	0.7321	0.114	6.445	0.000	0.509	0.955
Specialization_Travel and Tourism	-0.9068	0.401	-2.260	0.024	-1.693	-0.120
What is your current occupation_Unemployed	-0.7761	0.244	-3.184	0.001	-1.254	-0.298
Tags_Closed by Horizzon	8.8298	0.751	11.757	0.000	7.358	10.302
Tags_Not Mentioned	3.0398	0.198	15.357	0.000	2.652	3.428
Tags_Other_Tags	3.0496	0.208	14.684	0.000	2.643	3.457
Tags_Will revert after reading the email	7.0742	.246	28.728	0.000	6.592	7.557
Last Notable Activity_Other_Notable_Activity	1.3747	0.413	3.325	0.001	0.564	2.185
Last Notable Activity_SMS Sent	2.5763	0.129	20.012	0.000	2.324	2.829

Based on observation we have found the following 3 variables contributed the most towards the probability of lead getting converted:

- Tags_Closed by Horizzon
- Tags_Will revert after reading the email
- Lead Source_Welingak Website

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:
   Based on the coefficient values from the screen shot above in question 1, the following are the top three categorical/dummy variables that should be focused the most in order to increase the probability of lead conversion:

   - Tags_Closed by Horizzon (from Tags)
   - Tags_Will revert after reading the email (from Tags)
   - Lead Source_Welingak Website (from Lead Source)

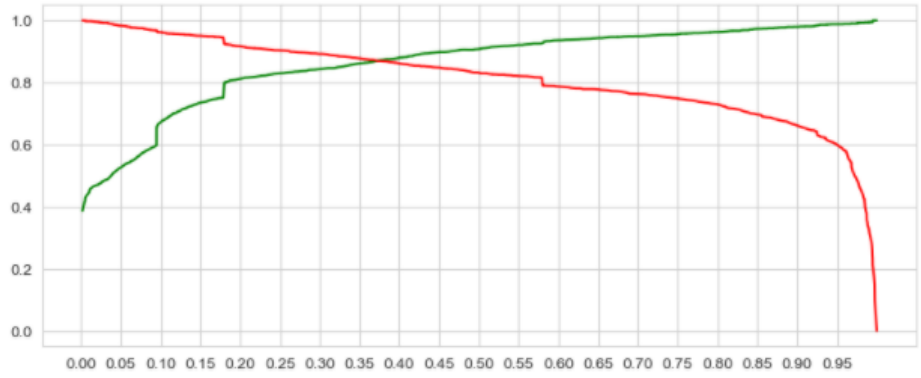
3. **X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So, during this phase, they wish to make the lead conversion more aggressive. So, they want almost all of the potential leads (i.e. the customers 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:

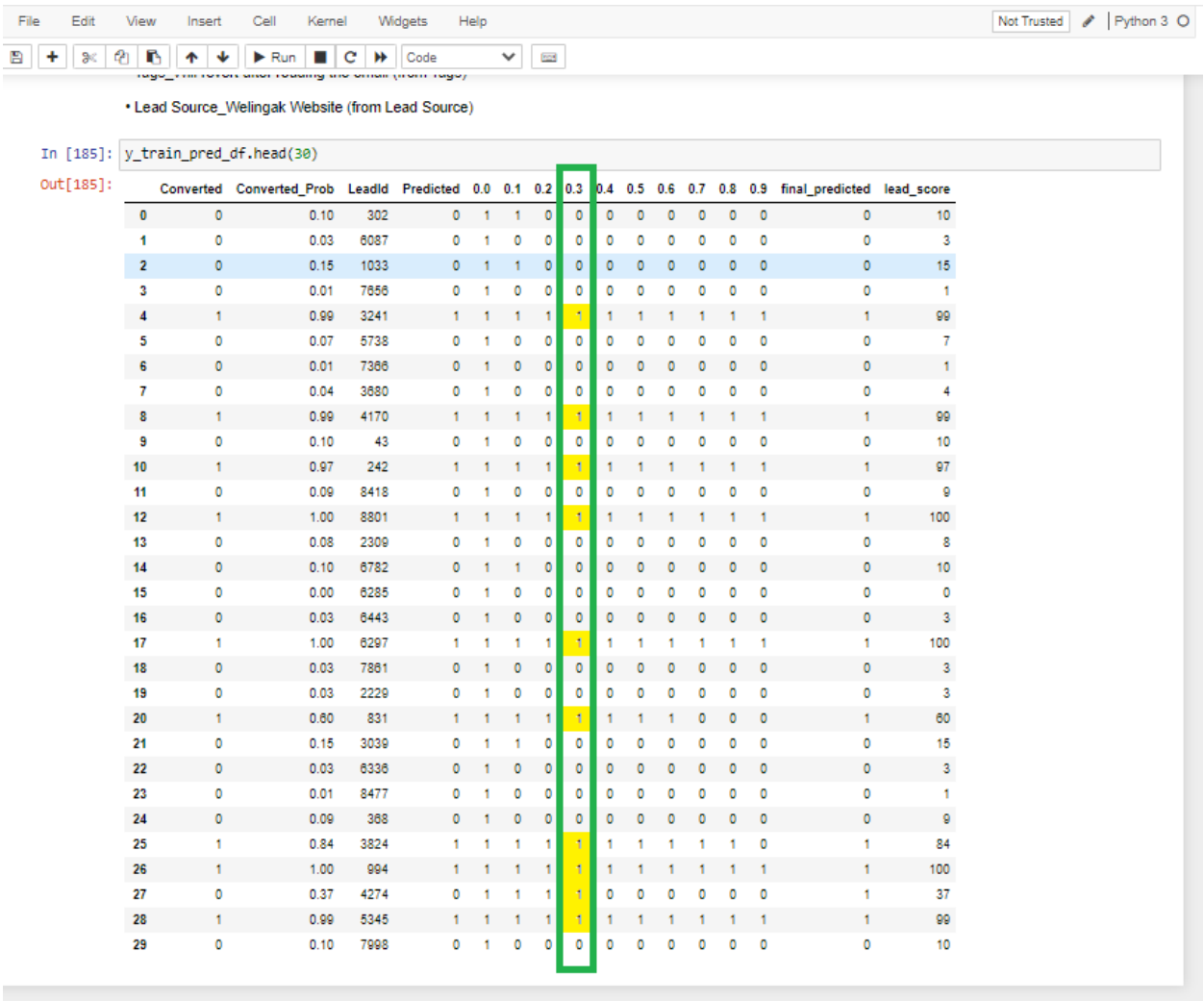
Precision and recall tradeoff

```
In [152]: # y_train_pred_df.Converted, y_train_pred_df.Predicted
          p, r, thresholds = precision_recall_curve(y_train_pred_df.Converted, y_train_pred_df.Converted_Prob)
```

```
In [153]: plt.figure(figsize=(10, 4), dpi=100, facecolor='w', edgecolor='k', frameon='True')
          plt.plot(thresholds, p[:-1], "g-")
          plt.plot(thresholds, r[:-1], "r-")
          plt.xticks(np.arange(0, 1, step=0.05))
          plt.show()
```



From the precision-recall graph above, we get the optimal threshold value as close to 0.37. However our business requirement here is to have Lead Conversion Rate around 80%. This is already achieved with our earlier threshold value of 0.3. So we will stick to this value.



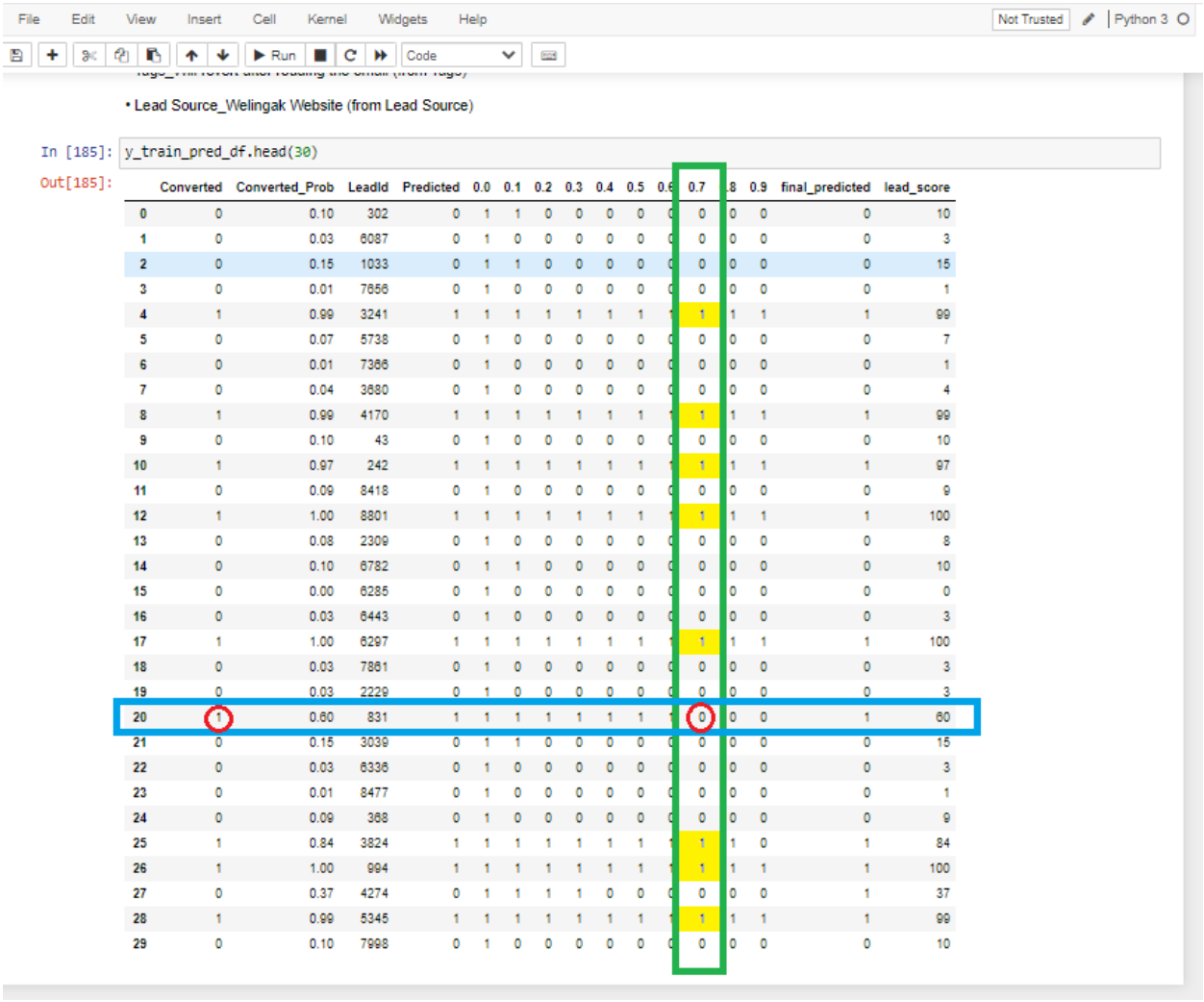
	Converted	Converted_Prob	LeadId	Predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	final_predicted	lead_score
0	0	0.10	302	0	1	1	0	0	0	0	0	0	0	0	0	10
1	0	0.03	6087	0	1	0	0	0	0	0	0	0	0	0	0	3
2	0	0.15	1033	0	1	1	0	0	0	0	0	0	0	0	0	15
3	0	0.01	7656	0	1	0	0	0	0	0	0	0	0	0	0	1
4	1	0.99	3241	1	1	1	1	1	1	1	1	1	1	1	1	99
5	0	0.07	5738	0	1	0	0	0	0	0	0	0	0	0	0	7
6	0	0.01	7366	0	1	0	0	0	0	0	0	0	0	0	0	1
7	0	0.04	3680	0	1	0	0	0	0	0	0	0	0	0	0	4
8	1	0.99	4170	1	1	1	1	1	1	1	1	1	1	1	1	99
9	0	0.10	43	0	1	0	0	0	0	0	0	0	0	0	0	10
10	1	0.97	242	1	1	1	1	1	1	1	1	1	1	1	1	97
11	0	0.09	8418	0	1	0	0	0	0	0	0	0	0	0	0	9
12	1	1.00	8801	1	1	1	1	1	1	1	1	1	1	1	1	100
13	0	0.08	2309	0	1	0	0	0	0	0	0	0	0	0	0	8
14	0	0.10	6782	0	1	1	0	0	0	0	0	0	0	0	0	10
15	0	0.00	6285	0	1	0	0	0	0	0	0	0	0	0	0	0
16	0	0.03	6443	0	1	0	0	0	0	0	0	0	0	0	0	3
17	1	1.00	6297	1	1	1	1	1	1	1	1	1	1	1	1	100
18	0	0.03	7861	0	1	0	0	0	0	0	0	0	0	0	0	3
19	0	0.03	2229	0	1	0	0	0	0	0	0	0	0	0	0	3
20	1	0.60	831	1	1	1	1	1	1	1	1	0	0	0	1	60
21	0	0.15	3039	0	1	1	0	0	0	0	0	0	0	0	0	15
22	0	0.03	6336	0	1	0	0	0	0	0	0	0	0	0	0	3
23	0	0.01	8477	0	1	0	0	0	0	0	0	0	0	0	0	1
24	0	0.09	368	0	1	0	0	0	0	0	0	0	0	0	0	9
25	1	0.84	3824	1	1	1	1	1	1	1	1	1	1	0	1	84
26	1	1.00	994	1	1	1	1	1	1	1	1	1	1	1	1	100
27	0	0.37	4274	0	1	1	1	1	0	0	0	0	0	0	1	37
28	1	0.99	5345	1	1	1	1	1	1	1	1	1	1	1	1	99
29	0	0.10	7998	0	1	0	0	0	0	0	0	0	0	0	0	10

In the above image, the final prediction is calculated based on optimal cut off value of **0.37**.

In order to make the sales aggressive, the company may contact all the leads which have a conversion probability value equals to 1 and under a cut off 0.37 (column 0.3 highlighted in green-yellow).

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:



File Edit View Insert Cell Kernel Widgets Help — Not Trusted — Python 3

Run — Code

• Lead Source_Welingak Website (from Lead Source)

```
In [185]: y_train_pred_df.head(30)
```

Out[185]:

	Converted	Converted_Prob	LeadId	Predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	final_predicted	lead_score
0	0	0.10	302	0	1	1	0	0	0	0	0	0	0	0	0	10
1	0	0.03	6087	0	1	0	0	0	0	0	0	0	0	0	0	3
2	0	0.15	1033	0	1	1	0	0	0	0	0	0	0	0	0	15
3	0	0.01	7656	0	1	0	0	0	0	0	0	0	0	0	0	1
4	1	0.99	3241	1	1	1	1	1	1	1	1	1	1	1	1	99
5	0	0.07	5738	0	1	0	0	0	0	0	0	0	0	0	0	7
6	0	0.01	7366	0	1	0	0	0	0	0	0	0	0	0	0	1
7	0	0.04	3680	0	1	0	0	0	0	0	0	0	0	0	0	4
8	1	0.99	4170	1	1	1	1	1	1	1	1	1	1	1	1	99
9	0	0.10	43	0	1	0	0	0	0	0	0	0	0	0	0	10
10	1	0.97	242	1	1	1	1	1	1	1	1	1	1	1	1	97
11	0	0.09	8418	0	1	0	0	0	0	0	0	0	0	0	0	9
12	1	1.00	8801	1	1	1	1	1	1	1	1	1	1	1	1	100
13	0	0.08	2309	0	1	0	0	0	0	0	0	0	0	0	0	8
14	0	0.10	6782	0	1	1	0	0	0	0	0	0	0	0	0	10
15	0	0.00	6285	0	1	0	0	0	0	0	0	0	0	0	0	0
16	0	0.03	6443	0	1	0	0	0	0	0	0	0	0	0	0	3
17	1	1.00	6297	1	1	1	1	1	1	1	1	1	1	1	1	100
18	0	0.03	7861	0	1	0	0	0	0	0	0	0	0	0	0	3
19	0	0.03	2229	0	1	0	0	0	0	0	0	0	0	0	0	3
20	1	0.60	831	1	1	1	1	1	1	1	1	0	0	0	1	60
21	0	0.15	3039	0	1	1	0	0	0	0	0	0	0	0	0	15
22	0	0.03	6336	0	1	0	0	0	0	0	0	0	0	0	0	3
23	0	0.01	8477	0	1	0	0	0	0	0	0	0	0	0	0	1
24	0	0.09	368	0	1	0	0	0	0	0	0	0	0	0	0	9
25	1	0.84	3824	1	1	1	1	1	1	1	1	1	1	0	1	84
26	1	1.00	994	1	1	1	1	1	1	1	1	1	1	1	1	100
27	0	0.37	4274	0	1	1	1	1	0	0	0	0	0	0	1	37
28	1	0.99	5345	1	1	1	1	1	1	1	1	1	1	1	1	99
29	0	0.10	7998	0	1	0	0	0	0	0	0	0	0	0	0	10

In order to minimize the rate of useless phone calls, the company may contact all the leads which have a conversion probability (value = 1 highlighted in green–yellow color) under column 0.7. However, they may miss out on those leads that are actually **converted** but then the model wrongly predicted them as **not converted**. (See red highlights in the image above). This should not be a major cause for concern as the target has already been achieved.