Lead Scoring Case Study

Presented by

Chaithanya Kumar Gadi

&

Harsha Gudihalam

Problem Statement

An education company named X Education sells online courses to industry professionals. When the people fill up a form providing their email address or phone number, they are classified to be a lead. Although X Education gets a lot of leads, its lead conversion rate is very poor. So, help the education company in identifying the most promising leads, i.e. the leads that are most likely to convert into paying customers by building a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

Analysis approach

- Performed all data cleaning operations, outlier analysis and feature scaling on the dataset and obtained the appropriate dataset for model building.
- Performed Test/Train split with 30%/70% on dataset, respectively to test and train the model.
- Performed both automated and manual process using RFE (Recursive Feature Elimination) and Statistics Model on the cleaned dataset, respectively.
- Performed manual feature elimination until resulted model contains appropriate p-values (where they were considered as less than 0.05 are good values) and VIF (Variance Inflation Factor) values (where they were considered as less than 5 are good values).
- Calculated metrics of the final model such as Accuracy, Sensitivity, Specificity including plotting of ROC curve and Precision-Recall curve to assess the final model.
- Created a predicted probability variable from the identified optimal cut-off point.
- Assigned a “Lead Sore” to each lead from their predicted probability value.

Logistic Regression Model with RFE returned 15 variables

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.5338	0.242	-6.332	0.000	-2.009	-1.059
Do Not Email	-1.2038	0.189	-6.381	0.000	-1.574	-0.834
Lead Origin_Lead Add Form	2.6535	0.233	11.410	0.000	2.198	3.109
Last Activity_SMS Sent	1.8446	0.093	19.931	0.000	1.663	2.026
What is your current occupation_Unemployed	-2.5115	0.188	-13.374	0.000	-2.880	-2.143
Tags_Busy	3.2776	0.321	10.224	0.000	2.649	3.906
Tags_Closed by Horizzon	8.9967	0.758	11.872	0.000	7.511	10.482
Tags_Interested in Next batch	23.9314	1.51e+04	0.002	0.999	-2.96e+04	2.97e+04
Tags_Lateral student	25.4193	2.07e+04	0.001	0.999	-4.05e+04	4.05e+04
Tags_Lost to EINS	9.0635	0.761	11.908	0.000	7.572	10.555
Tags_Ringing	-0.8959	0.358	-2.503	0.012	-1.597	-0.194
Tags_Will revert after reading the email	3.8110	0.234	16.262	0.000	3.352	4.270
Tags_in touch with EINS	3.3100	0.872	3.795	0.000	1.600	5.020
Tags_switched off	-1.9905	1.033	-1.926	0.054	-4.016	0.035
Last Notable Activity_Modified	-1.6526	0.094	-17.659	0.000	-1.836	-1.469
Last Notable Activity_Olark Chat Conversation	-1.8407	0.350	-5.263	0.000	-2.526	-1.155

Final model received after manual feature elimination using StatsModel

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	5866
Model:	GLM	**Df Residuals:**	5854
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1867.5
Date:	Thu, 27 Feb 2020	**Deviance:**	3734.9
Time:	14:02:26	**Pearson chi2:**	8.55e+03
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.8641	0.219	-8.492	0.000	-2.294	-1.434
Do Not Email	-1.1776	0.187	-6.302	0.000	-1.544	-0.811
Lead Origin_Lead Add Form	2.6176	0.227	11.543	0.000	2.173	3.062
Last Activity_SMS Sent	1.8242	0.091	20.011	0.000	1.646	2.003
What is your current occupation_Unemployed	-2.6019	0.189	-13.790	0.000	-2.972	-2.232
Tags_Busy	3.7056	0.273	13.583	0.000	3.171	4.240
Tags_Closed by Horizzon	9.3958	0.742	12.666	0.000	7.942	10.850
Tags_Lost to EINS	9.4585	0.745	12.691	0.000	7.998	10.919
Tags_Will revert after reading the email	4.2254	0.171	24.672	0.000	3.890	4.561
Tags_in touch with EINS	3.7239	0.857	4.346	0.000	2.045	5.403
Last Notable Activity_Modified	-1.6262	0.093	-17.411	0.000	-1.809	-1.443
Last Notable Activity_Olark Chat Conversation	-1.8321	0.351	-5.221	0.000	-2.520	-1.144

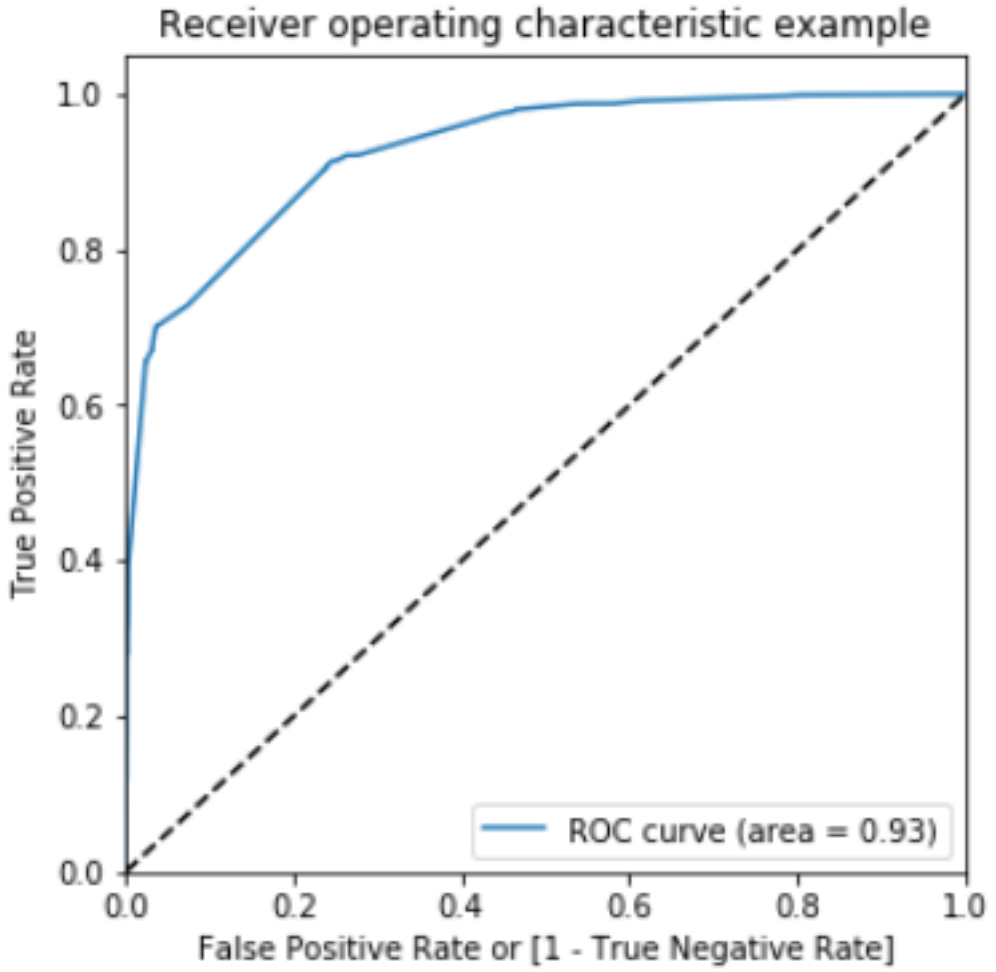
VIF (Variance Inflation Factor) of the final model variables

	Features	VIF
3	What is your current occupation_Unemployed	3.16
7	Tags_Will revert after reading the email	2.42
9	Last Notable Activity_Modified	1.71
2	Last Activity_SMS Sent	1.53
1	Lead Origin_Lead Add Form	1.30
5	Tags_Closed by Horizzon	1.27
0	Do Not Email	1.10
4	Tags_Busy	1.07
6	Tags_Lost to EINS	1.05
10	Last Notable Activity_Olark Chat Conversation	1.05
8	Tags_in touch with EINS	1.00

Plotting ROC Curve

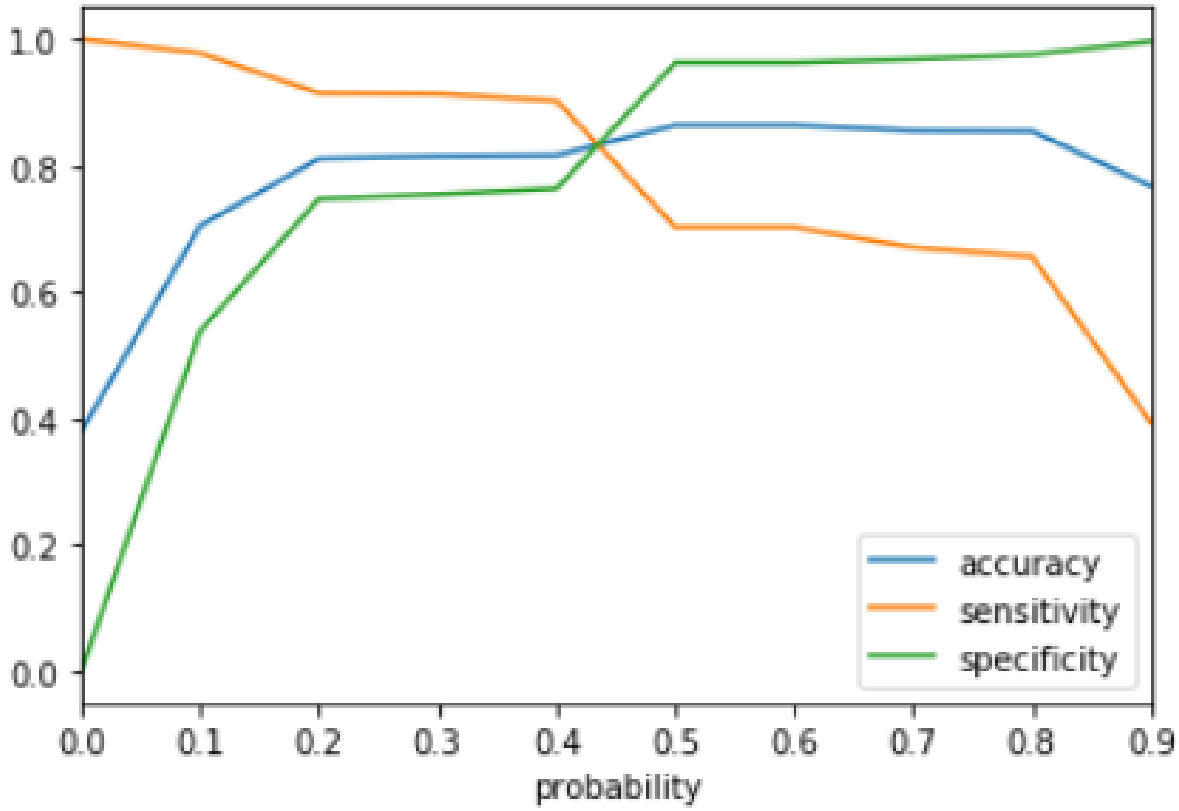
An ROC curve demonstrates follwoing:

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



Finding Optimal Cut-off point

From the curve, 0.44 is the optimum point identified as a cutoff probability which increases the lead conversion.

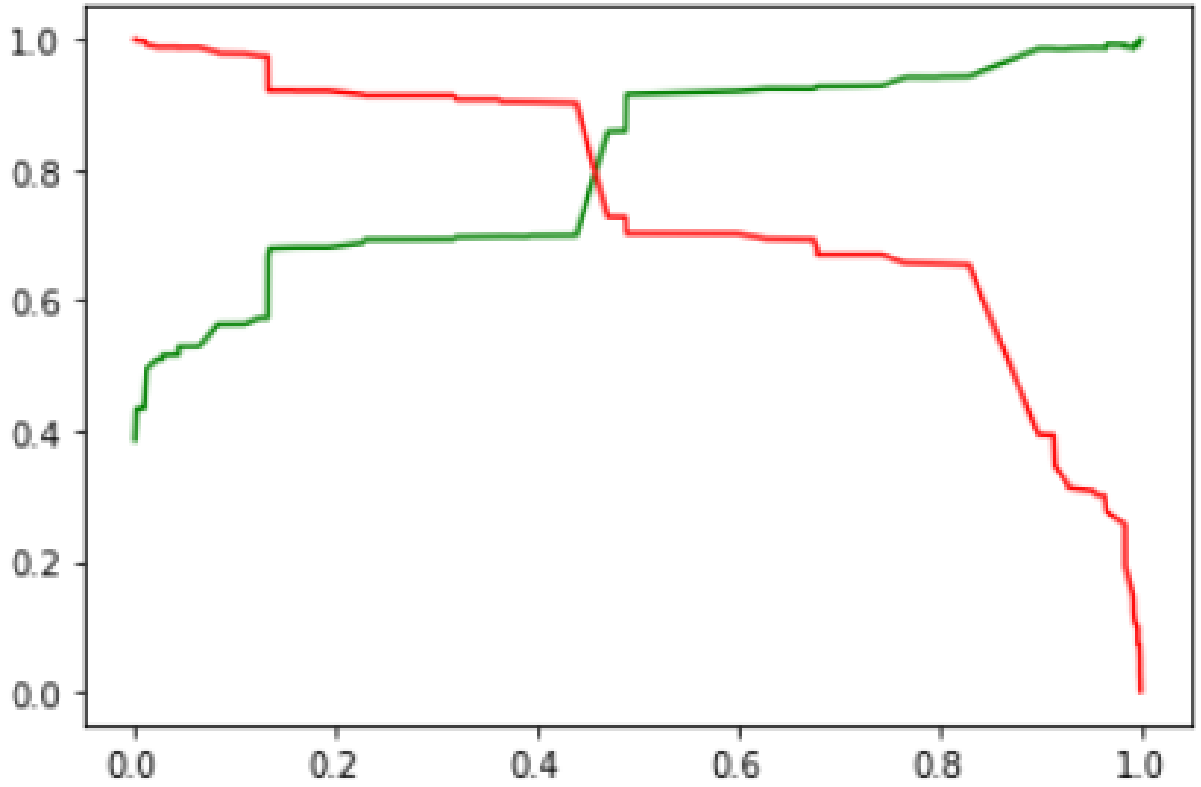


Actual converted vs Predicted with conversion probability cut-off as 0.44 (i.e., optimal cut-off) and assigned respective “Lead Score”

	Lead Number	Converted	Convert_Prob	Predicted	Lead Score
0	2930	1	0.318566	0	32.0
1	378	0	0.440161	1	44.0
2	2655	1	0.998891	1	100.0
3	3752	0	0.011363	0	1.0
4	6388	0	0.440161	1	44.0

Precision-Recall Trade-off

- The precision-recall curve shows the tradeoff between precision and recall for different threshold.
- A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate.
- High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall).
- From the graph, identified cut-off point was 0.48.




```
precision_score(y_train_pred_final.Converted, y_train_pred_final.Predicted)
```

```
0.7000348796651552
```

```
recall_score(y_train_pred_final.Converted, y_train_pred_final.Predicted)
```

```
0.9024280575539568
```

Metrics of the final model

```
# Let's check the overall accuracy.
metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.Predicted)
```

```
0.8163995908625981
```

Metrics beyond simle accuracy

```
confusion = metrics.confusion_matrix(y_train_pred_final.Converted, y_train_pred_final.Predicted )
confusion
```

```
array([[2782,  860],
       [ 217, 2007]], dtype=int64)
```

```
TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives
```

```
# Let's see the sensitivity of our logistic regression model
TP / float(TP+FN)
```

```
0.9024280575539568
```

```
# Let us calculate specificity
TN / float(TN+FP)
```

```
0.7638660076880834
```

Metrics of the final model (...continued)

```
# Calculate false postive rate - predicting conversion when customer does not have converted
print(FP/ float(TN+FP))
```

```
0.23613399231191654
```

```
# Positive predictive value
print (TP / float(TP+FP))
```

```
0.7000348796651552
```

```
# Negative predictive value
print (TN / float(TN+FN))
```

```
0.9276425475158386
```

Recommendations and solutions to business problems

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

1. 'Tags_Closed by Horizzon',
2. 'Tags_Lost to EINS',
3. 'Tags_Will revert after reading the email'

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?

1. 'Tags_Closed by Horizzon',
2. 'Tags_Lost to EINS',
3. 'Tags_Will revert after reading the email'

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.

Here, there are two things that should be implemented parallelly.

1. Interns need training
2. Make phone calls to potential leads as much as possible.

Follow below steps:

1. First week, train the interns with senior sales executives on making phone calls, gather current status of potential leads.
2. Second week onwards, whole sales team including interns should concentrate on the making phone calls and gather current status of potential leads.
3. Enquire with potential leads about their availability for discussion.
4. Call them in given time and explain them the courses availability, discounts/offers, seat limitation and demand of courses. Provide them the sales contact number for any of their queries to get resolved.
5. Keep calling them and repeat the step 4 which will definitely increase the conversion rate.

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.

From question 1 and 2, top 3 categorical/dummy variables in the model which contributed and should be focused the most on in order to increase the probability of lead conversion contains:

✓ **'Tags_Will revert after reading the email'**

- This feature tells us that most of the leads who are getting converted might be contacting back to X-Education for details and getting conversion after reading the mail sent by X-Education.
- As the company want to minimize the rate of useless phone calls, instead of calling customers use other channels such as Email and SMS with sales contact numbers and email ids.

End