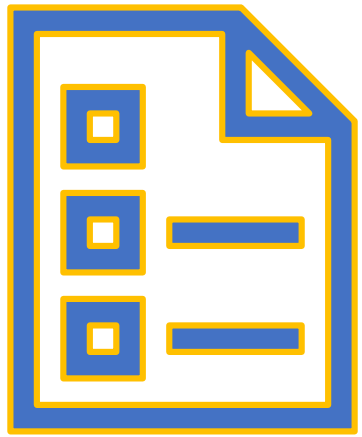


Lead Scoring- Case Study

BY : PHANI TEJA, GANESH SHINDE, MALLESH HG

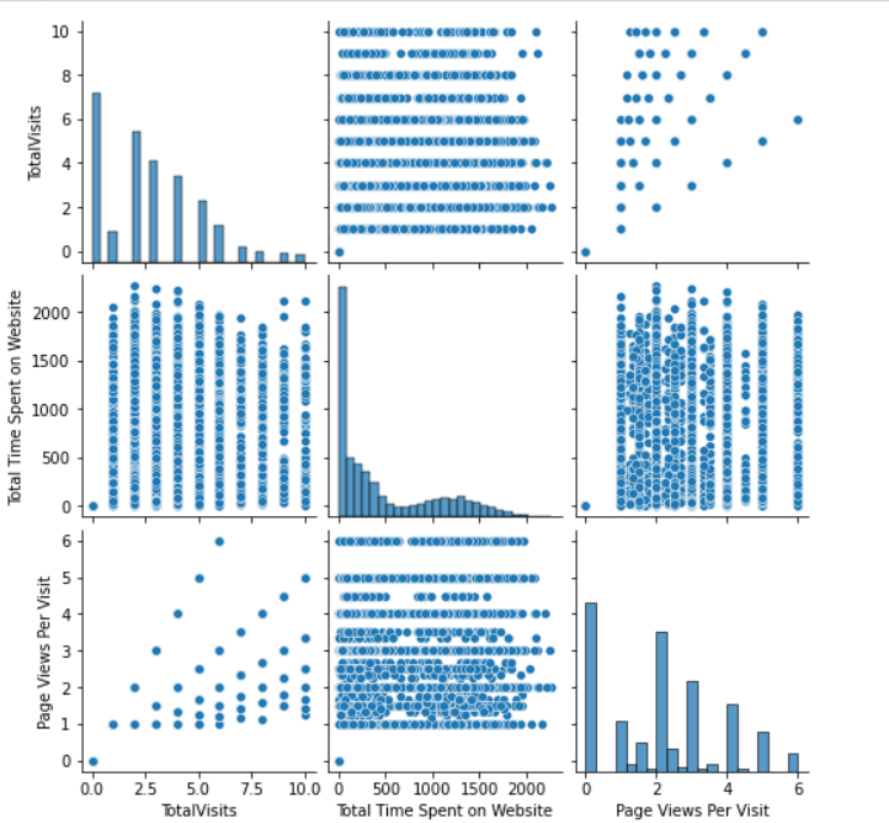
Contents

Problem Statement

Lead Conversion Process - Demonstrated as a funnel

The X Education company requires you to build a logistic regression 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.

Bi Variate/Multi Variate Analysis



- Except for Total Visits and Page Views Per Visit there is no proper correlation among others.
- As the number of Total Visits to the website increases, the maximum number of pages viewed is increased but the minimum number of pages viewd remains same between 0-2.

Road Map

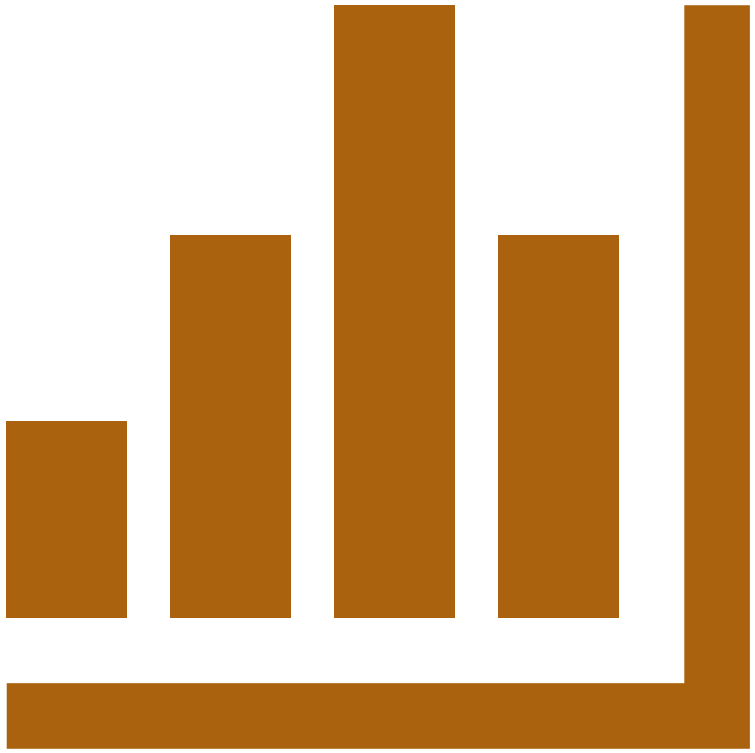
- ✓Created train and test set by splitting the original cleaned data set after treating missing values.
- ✓Selected 15 features using Recursive Feature Elimination (RFE) after creating dummy variables and scaling the data.
- ✓Applied Logistic Regression algorithm to build a model and more than 92% accuracy and 87% sensitivity.
- ✓Indentified the optimal probability cutoff from the accuracy, sensitivity and specificity.
- ✓Applied the model on the test data to identify the conversion probability. (accuracy 92%, sensitivity 87%)
- ✓Based on the calculated predicted probability, and optimal probability cutoff, all the leads are assigned with a lead score value (lead score = predicted probability x 100)

Finalized Model

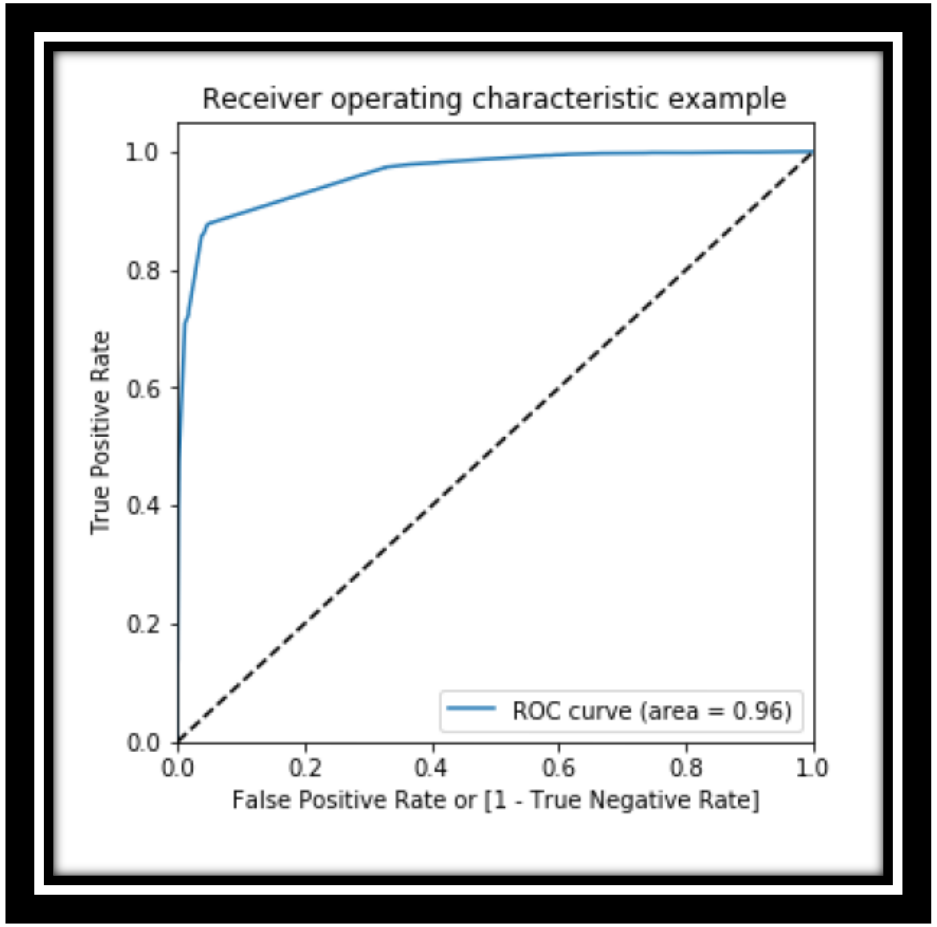
	Features	VIF
11	Tags_Will revert after reading the email	2.88
4	Lead Quality_Might be	2.56
13	Last Notable Activity_SMS Sent	1.58
1	Lead Origin_Lead Add Form	1.53
2	Lead Source_Welingak Website	1.32
10	Tags_Ringing	1.28
3	Last Activity_Olark Chat Conversation	1.23
7	Tags_Closed by Horizzon	1.11
8	Tags_Interested in other courses	1.11
0	Do Not Email	1.10
5	Lead Quality_Worst	1.08
12	Tags_switched off	1.05
6	Tags_Busy	1.03
9	Tags_Lost to EINS	1.03

Dep. Variable:	Converted	**No. Observations:**	5931
Model:	GLM	**Df Residuals:**	5916
Model Family:	Binomial	**Df Model:**	14
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1337.4
Date:	Sun, 22 Jan 2023	**Deviance:**	2674.8
Time:	16:07:44	**Pearson chi2:**	3.35e+04
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.5822
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.9056	0.282	-10.321	0.000	-3.457	-2.354
Do Not Email	-1.3505	0.226	-5.963	0.000	-1.794	-0.907
Lead Origin_Lead Add Form	2.1095	0.444	4.750	0.000	1.239	2.980
Lead Source_Welingak Website	2.7323	1.111	2.459	0.014	0.554	4.910
Last Activity_Olark Chat Conversation	-1.5272	0.217	-7.053	0.000	-1.952	-1.103
Lead Quality_Might be	-3.9731	0.154	-25.819	0.000	-4.275	-3.671
Lead Quality_Worst	-2.3138	0.742	-3.120	0.002	-3.767	-0.860
Tags_Busy	2.3640	0.359	6.586	0.000	1.660	3.068
Tags_Closed by Horizzon	9.9483	1.094	9.096	0.000	7.805	12.092
Tags_Interested in other courses	-0.3922	0.547	-0.717	0.473	-1.464	0.679
Tags_Lost to EINS	9.8695	0.681	14.489	0.000	8.534	11.205
Tags_Ringing	-1.7729	0.378	-4.687	0.000	-2.514	-1.031
Tags_Will revert after reading the email	5.4331	0.310	17.509	0.000	4.825	6.041
Tags_switched off	-2.6663	0.773	-3.451	0.001	-4.181	-1.152
Last Notable Activity_SMS Sent	2.4261	0.124	19.603	0.000	2.184	2.669

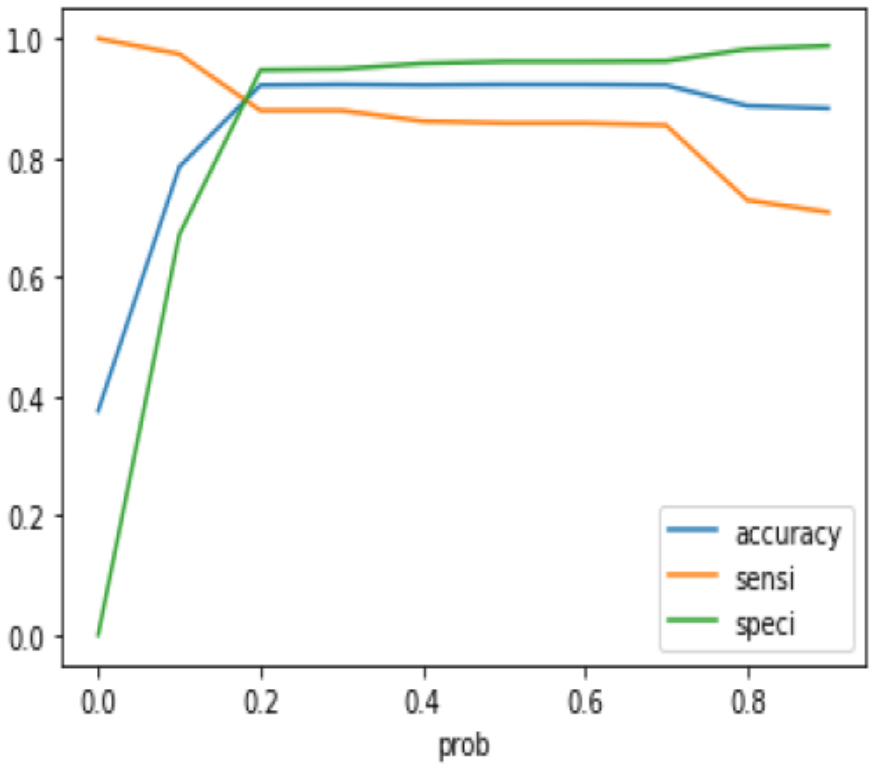


ROC Curve & Optimal Probability Cutoff



ROC Curve

- 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.
- The ROC curve shows that the 96% of the area is under the curve.
- The classification probability of lead conversion (1/0) is very high by the model.



Optimal probability cutoff

- ✓Optimal probability cutoff is identified as 0.2 for better accuracy of the classification of lead conversion.
- ✓With 0.2 cutoff the model has
  - –Accuracy : 92%
  - –Sensitivity : 87%
  - –Specificity : 94%

Confusion matrix on Test data

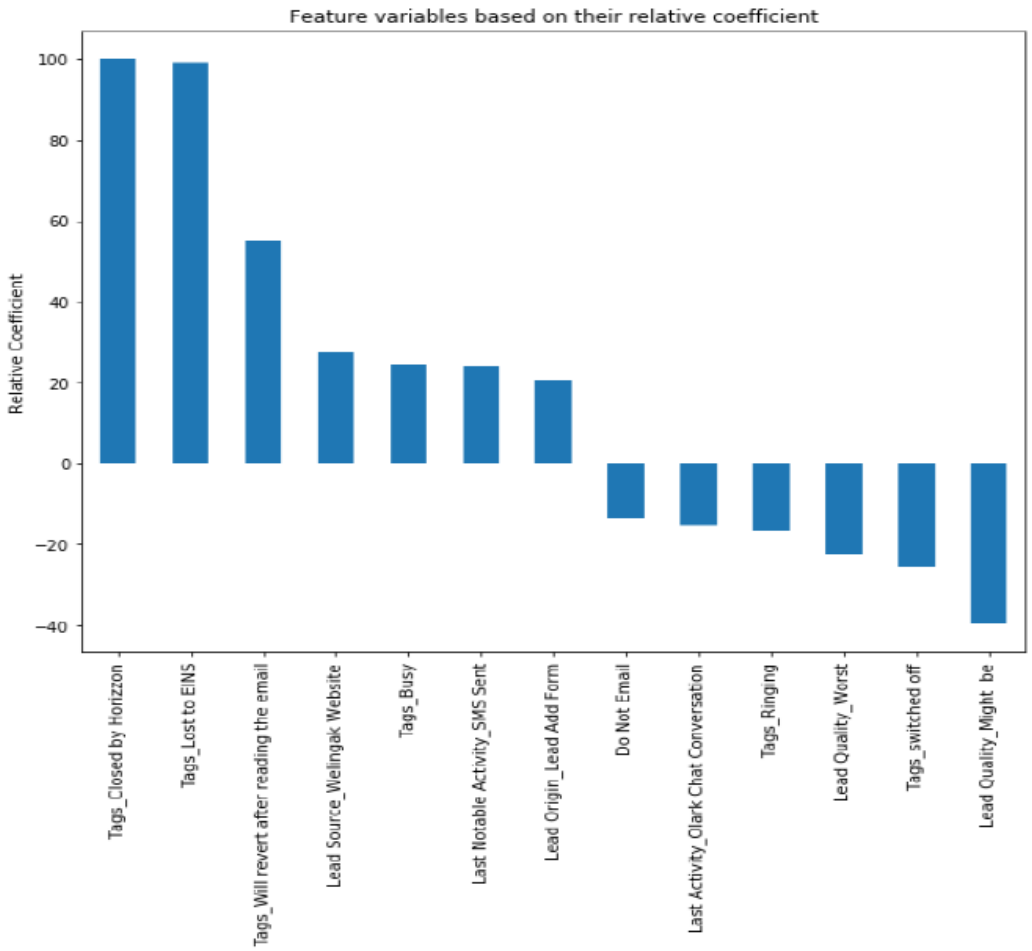
Actual/Predicted	Not Converted	Converted
Not Converted	1498	82
Converted	121	842

Accuracy : 92% | Sensitivity : 87% | Specificity : 94%

The model can predict if a lead can be converted or not with 92% accuracy on unseen data. This will help the company to predict the probability of ‘hot’ leads with 92% accuracy.

Also, the model can predict the probability of a lead which are actually converted over total converted lead with 87% chances.

The model’s prediction of a lead not getting converted is also very high (94% over unseen data). This means that the X education company will save lot of time and resources by discarding low scoring leads.



Important Features

- Top 3 variables that contributing to convert a lead are:
  - –Tags_Closed by Horizzon
  - –Tags_Lost to EINS
  - –Tag_We will revert after reading the email
- Top 3 variables that need improvement to convert a lead are:
  - –Lead Quality_ Might Be
  - –Tag_switchedoff
  - –Lead Quality_Worst

Situation 1: Company has interns for 2 months. They wish to make lead conversion more aggressive. They want almost all of the potential leads to be converted and hence, want to make phone calls to as much of such people as possible.

Solution:

- $\boldsymbol{Sensitivity = TruePositives / (TruePositives + FalseNegatives)}$

- Sensitivity can be defined as the number of actual conversions predicted correctly out of total number of actual conversions. As we saw earlier, sensitivity decreases as the threshold increases.

- High sensitivity implies that our model will correctly predict almost all leads who are likely to convert. At the same time, it may overestimate and misclassify some of the non-conversions as conversions.

- As the company has extra man-power for two months and wants to make the lead conversion more aggressive, it is a good strategy to go for high sensitivity. To achieve high sensitivity, we need to choose a low threshold value.

Situation 2:

At times, the company reaches its target for a quarter before the deadline. It wants the sales team to focus on some new work. So during this time, the company's aim is to not make phone calls unless it's extremely necessary.

Solution:

- $\boldsymbol{Specificity = TrueNegatives / (TrueNegatives + FalsePositives)}$
- Specificity can be defined as the number of actual non-conversions predicted correctly out of total number of actual non-conversions. It increases as the threshold increases.
- High specificity implies that our model will correctly predict almost all leads who are not likely to convert. At the same time, it may misclassify some of the conversions as non-conversions.
- As the company has already reached its target for a quarter and doesn't want to make unnecessary phone calls, it is a good strategy to go for high specificity.
- It will ensure that the phone calls are only made to customers who have a very high probability of conversion. To achieve high specificity, we need to choose a high threshold value.

Recommendations

- The leads which have high score can be treated as “hot” leads and sales team need to follow up as there is high possibility to convert those leads.
- Leads who have applied for ‘Do Not Email’ already does not needs to be attended again.
- Based on the previous chat conversations if the lead is classified as ‘Might be’ or ‘Worst’ then those leads can be ignored.

Thank You

- Team:
  - Phani Teja.D
  - Ganesh Shinde
  - Mallesh HG