Lead Scoring Case Study - Presentation

By,

- Rajesh Kumar
- Nitesh Rustogi
- Swathi Nizam

Background

- X Education, an education company sells online courses to industry professionals. They have a website to browse courses. People can either land here directly or through other websites and fill up form.
- Though there are lot of leads coming in and filling up form, the lead conversion rate is very low.

Problem Statement

- Even though they are spending a lot of time calling up on leads, most of them are not joining (low conversion rate).
- Here the problem is two-fold:
  - ✓You would need more salespeople to cover all the customers who filled the form (leads). This would increase the cost.
  - ✓If we are going with the existing work force, they may not be able to spend quality time and brief about the courses with the customers. Or they may choose random customers to contact which may not give fruitful results.

Solution

- A model that can help X Education understand the most potential customers based on the previous history.
- As this is binary classification problem, one of the possible solutions could be to create Logistic Regression model that outputs lead conversion rate with good metrics.
- The metrics in focus could be accuracy and specificity/Recall as we need to be mindful about the possible leads to be converted to hot leads.
- With this solution, sales team can call only the leads that have more potential to become customers. This way they can spend quality time with potential customers without increasing the head count of the sales team.

Analysis Approach

- Analyzed the data, cleaned it up by removing unwanted columns.
- Handled outliers, missing values. Visualized data through charts.
- Identified the most relevant features, built models in the recursive fashion based on VIF and pvalue.
- Identified the optimal probability and calculated sensitivity/precision, recall, specificity, accuracy etc part of model evaluation.
- Calculated lead score, based on which salesperson can identify whether to contact him or not.

Assumptions

- Columns with missing data > 40% are of not much use to the model and are dropped.
- Management specific specializations are grouped under “Management Specialization” and business specific specializations are grouped under “Business Specialization”.
- Columns with values inclined towards one particular value are dropped.
- If the number of missing values in a column are less, these missing values are replaced with the highest occurrence values in that column.
- If the number of missing values in a column are more, these missing values are categorized into “Others” section.
- Values whose representation is very less in a column (categorical) are renamed to a common value.
- While handling outliers, the upper limit to which the outliers are brought to is considered as 75% quantile+IQR*1.5.
- Train set is taken as 70% and test set as 30% from the actual dataset.

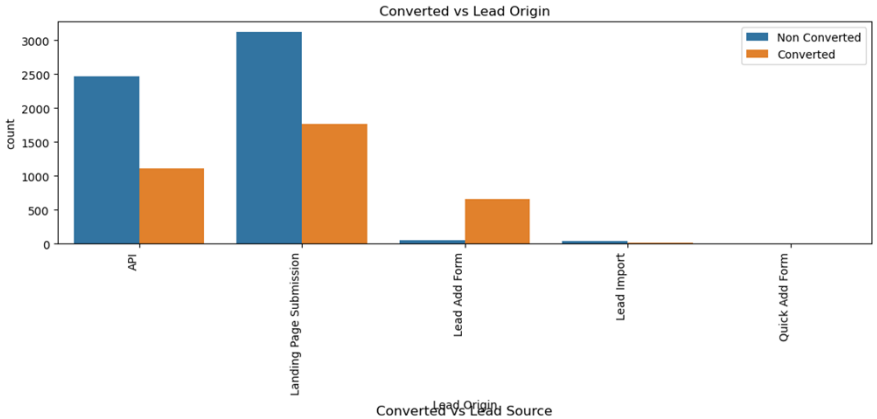
Data Distribution

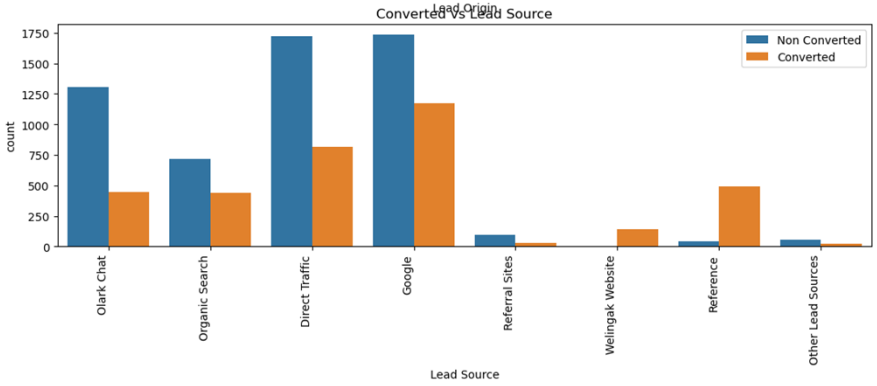
What is your current occupation	Not Converted	Converted	Convertedperc
Not Specified	2320	370	13.754647
Other	9	25	73.529412
Student	132	78	37.142857
Unemployed	3159	2441	43.589286
Working Professional	59	647	91.643059

Last Notable Activity	Not Converted	Converted	Convertedperc
Email Opened	1783	1044	36.929607
Modified	2624	783	22.982096
Other Last Notable Activity	383	133	25.775194
Page Visited on Website	225	93	29.245283
SMS Sent	664	1508	69.429098

Specialization	Not Converted	Converted	Convertedperc
Banking, Investment And Insurance	171	167	49.408284
Business Administration	224	179	44.416873
Business Specialization	192	116	37.662338
E-COMMERCE	72	40	35.714286
Management Specialization	2331	1922	45.191629
Media and Advertising	118	85	41.871921
Not Specified	2411	969	28.668639
Services Excellence	29	11	27.500000
Travel and Tourism	131	72	35.467980

Results





Inferences:

For Lead Origin:

1.In terms of numbers, "Landing Page Submission" and "API" have more conversions when compared to Lead Add Form and Lead Import.

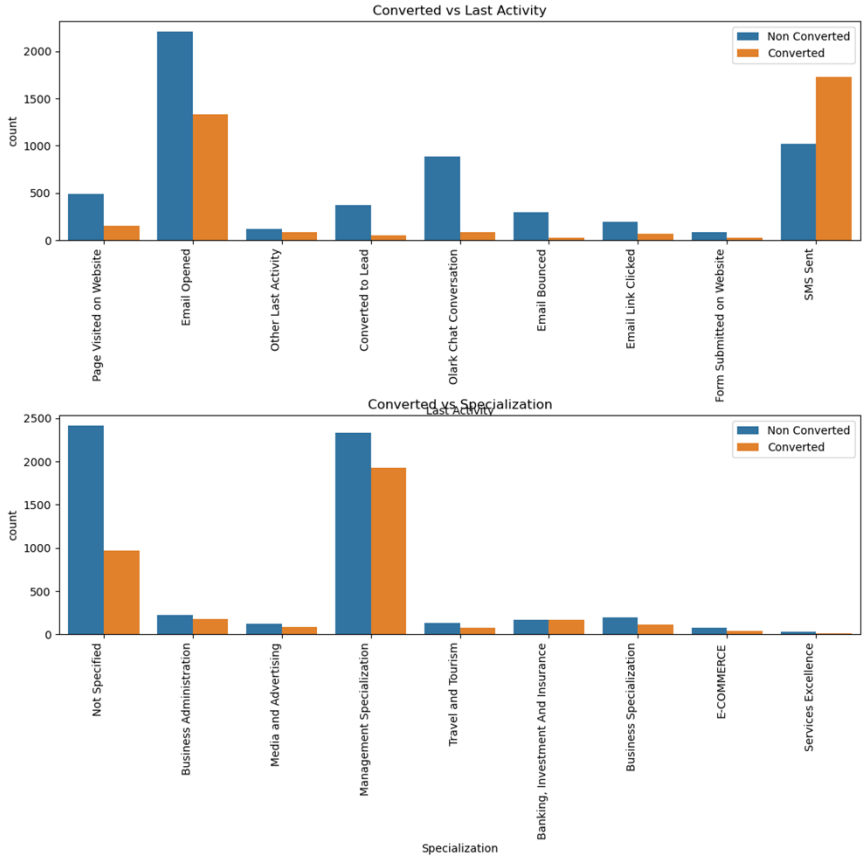
2.In terms of percentages, lot of the leads are converted for "Lead Add Form" Lead origin. So, conversion rate is high.

For Lead Source

1.It's observed that more number of leads are coming from Google, Direct Traffic and Olark Chat.

2.However, the conversion rate is more from "Reference" and "Welingak Website" Lead Source.

Results



Inferences:

For Last Activity

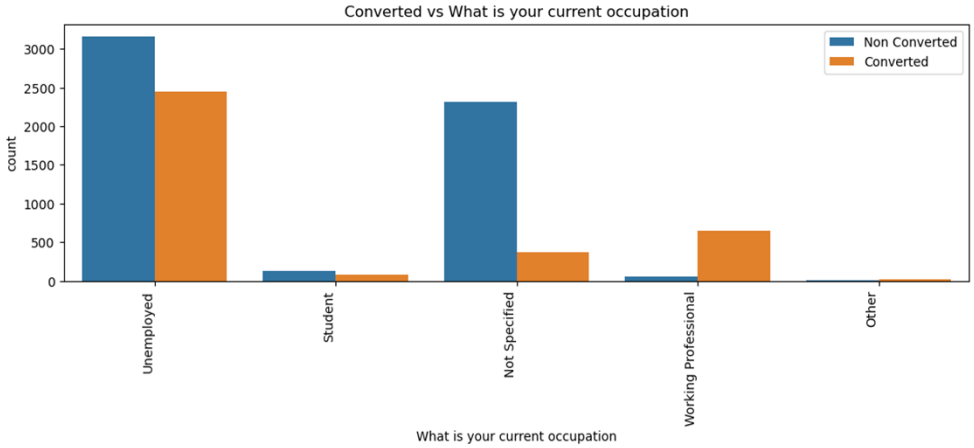
1.Most of the leads have their last activity "Email Opened".

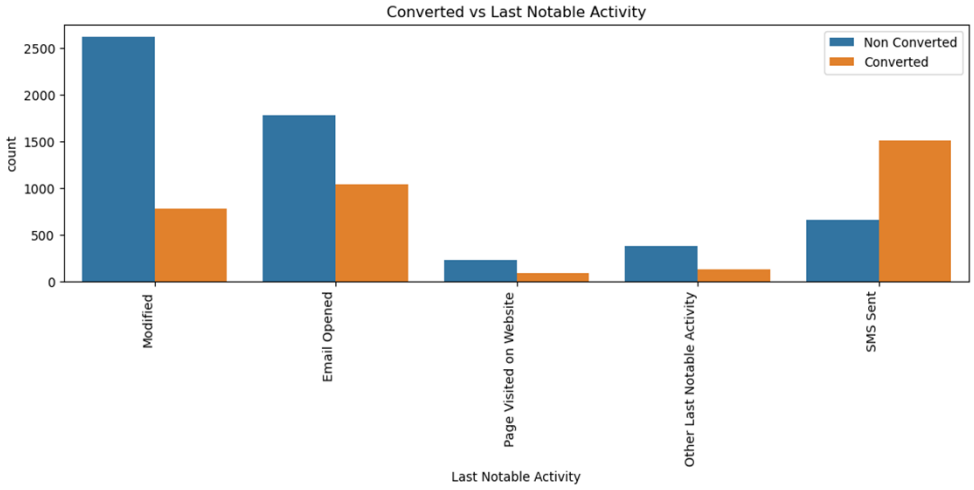
2.Most of the leads are converted when their last activity is SMS sent.

For Specialization

1.Conversion rates are good for leads specialized in "Banking, Investment And Insurance" (49%), "Management Specialization" (45%), Business Administration (44%).

Results





Inferences:

For What is your current occupation

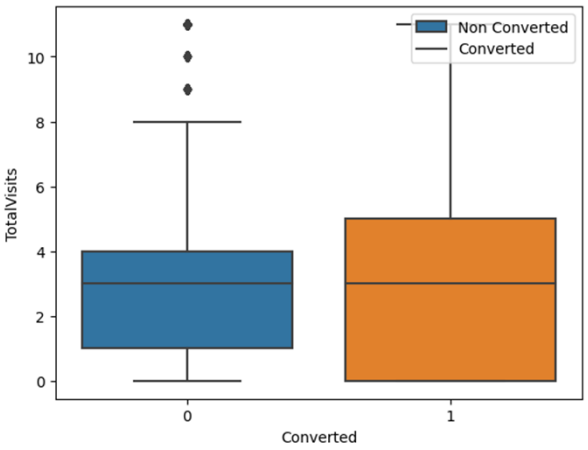
1.Most of the leads are from Unemployed occupation type. However, the conversion rate is low.

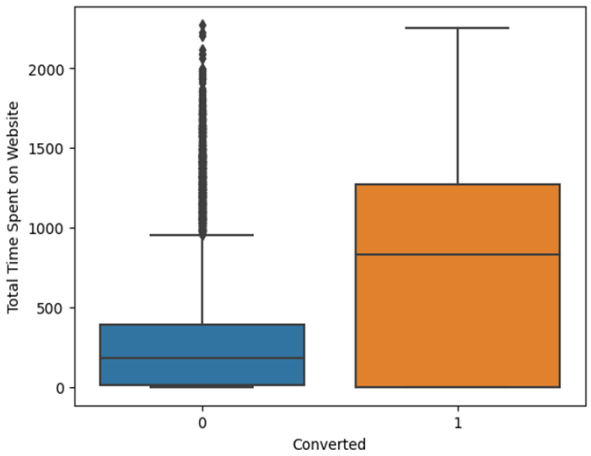
2.Many working professionals are converted may be for better career prospects (91.6%)

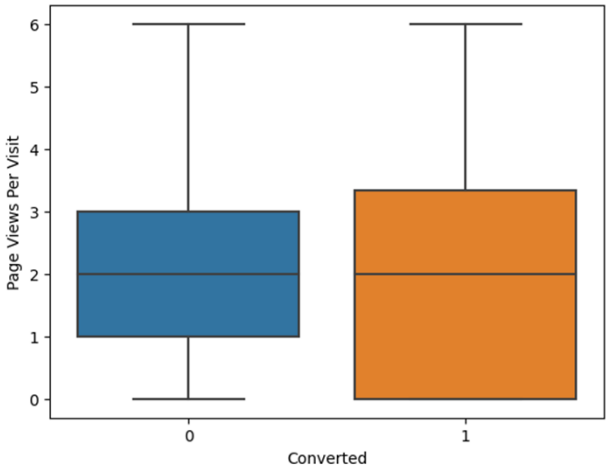
For Last Notable Activity

1.Even though most of the leads have "Modified" as their "Last Notable Activity", most of them are converted when a lead's last notable activity is "SMS Sent". Same inference can be obtained from the field "Last Activity" also.

Results







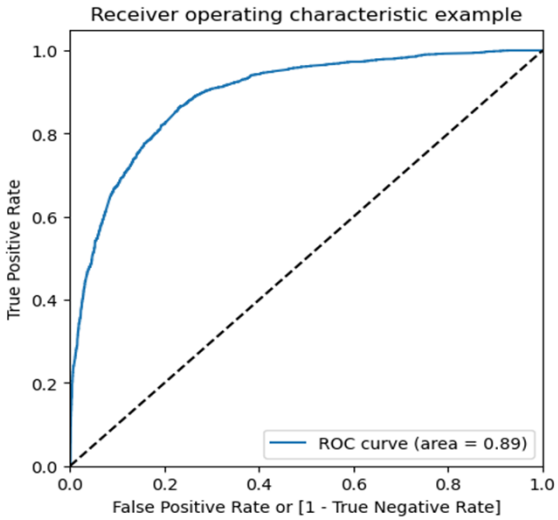
Inferences:

- ✓ More the number of visitors, more the chances of them getting converted. Median is same for both 'Converted' and ‘Not Converted' for total visits.
- ✓ If the time spent on the website is more, there are high chances of them being converted to hot leads.
- ✓ If Page Views per visit is more, there are little more chances of becoming hot leads. Median is same for both 'Converted' and ‘Not Converted' for page views per visit. In the lower quantile, we see that even though page views per visit is less, leads got converted.

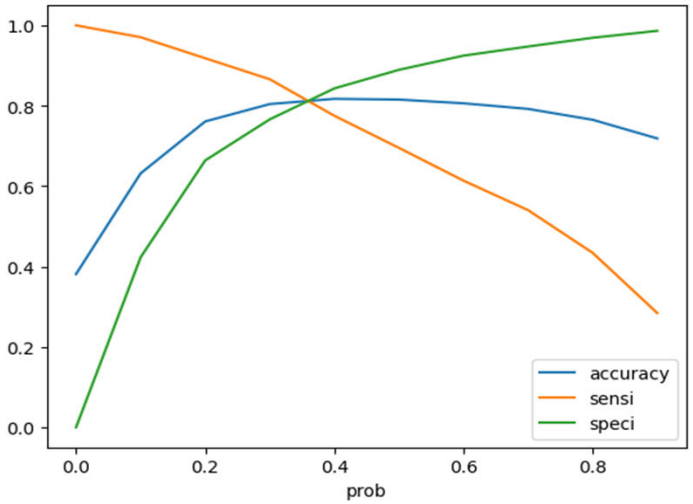
Model Building Summary

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.1985	0.138	-15.929	0.000	-2.469	-1.928
Do Not Email	-1.2034	0.175	-6.884	0.000	-1.546	-0.861
TimeSpent	1.0828	0.040	26.984	0.000	1.004	1.161
Freecopy	-0.4467	0.087	-5.164	0.000	-0.616	-0.277
Lead Origin_Lead Add Form	3.3329	0.192	17.381	0.000	2.957	3.709
Lead Source_Olark Chat	1.3143	0.115	11.446	0.000	1.089	1.539
Lead Source_Welingak Website	2.1844	0.746	2.930	0.003	0.723	3.646
Last Activity_Email Opened	0.5203	0.106	4.906	0.000	0.312	0.728
Last Activity_Other Last Activity	1.3618	0.249	5.479	0.000	0.875	1.849
Last Activity_SMS Sent	1.6503	0.107	15.374	0.000	1.440	1.861
Specialization_Not Specified	-0.4657	0.095	-4.891	0.000	-0.652	-0.279
Occupation_Other	2.1163	0.528	4.012	0.000	1.082	3.150
Occupation_Student	1.1137	0.244	4.564	0.000	0.635	1.592
Occupation_Unemployed	1.0524	0.088	11.959	0.000	0.880	1.225
Occupation_Working Professional	3.5119	0.203	17.333	0.000	3.115	3.909
Last Notable Activity_Modified	-0.7385	0.087	-8.514	0.000	-0.909	-0.569

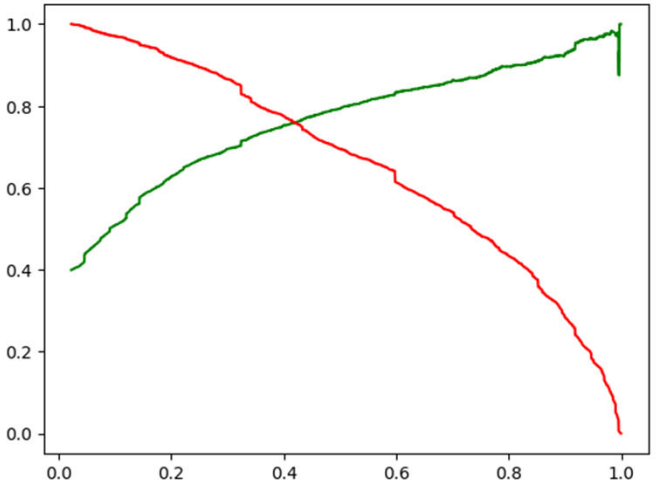
ROC



ROC Curve



Cutoff Probability using sensitivity, specificity, accuracy.



Cutoff Probability using precision, recall.

***ROC Curve**: The area under ROC curve is 0.89 which is good and the True positive rate line is hugging the vertical line (y axis)*

***Cutoff Probability using sensitivity, specificity, accuracy**: cut off is 0.35*

***Cutoff Probability using precision, recall**: cut off is 0.4*

Model Evaluation Metrics

```
####### Train Dataset Metrics:

Accuracy: 81.1%
Recall/Sensitivity:  80.5%
Specificity:  81.5%
Precision:  72.8%

ROC: 0.89

####### Test Dataset Metrics:

Accuracy: 81.6%
Recall/Sensitivity:  80.5%
Specificity:  82.3%
Precision:  74.81%


Looks like the model is behaving the good with training and the test dataset.
```

Recommendations

- Target the leads:
  - ✓Whose total visits, total time spent on website are more.
  - ✓Who is either “working professional” or “unemployed” or “Student”.
  - ✓Through “Welingak Website”, “Reference”, “Olark Chat” lead sources.
  - ✓Whose last activity or last notable activity is SMS sent.
  - ✓Working in “Banking, Investment And Insurance”, any Management Specialization or “Business Administration” specialization.
  - ✓Whose lead origin is “Lead Add Form”.

Thank You