LEAD SCORING CASE STUDY

BY

VARSHITHA M

&

SIDDHESH K

PROBLEM STATEMENT

X Education sells online courses to industry professionals.
The typical lead conversion rate at X education is around 30%.
We are required to optimize this lead conversion rate to more than 80%.
What strategies can we adopt to achieve this outcome?

- We will be building a logistic regression model to identify the most positively or negatively influencing factors which affect our lead conversion rate across out sample dataset.

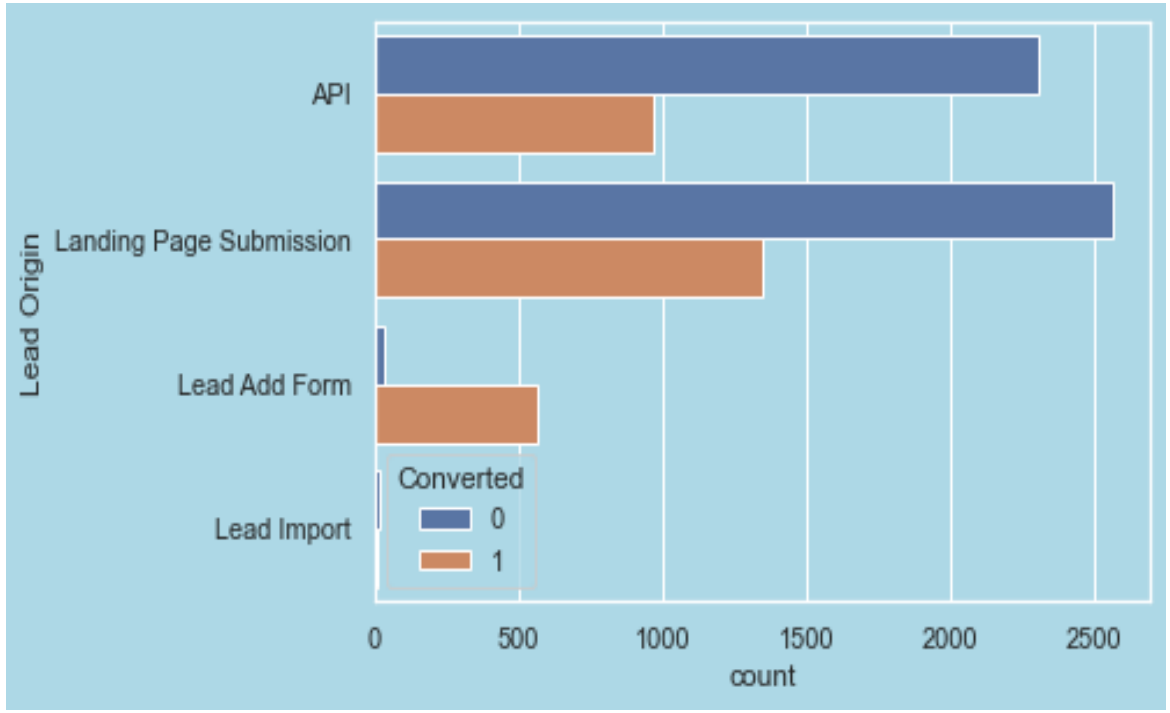
METHODOLOGY

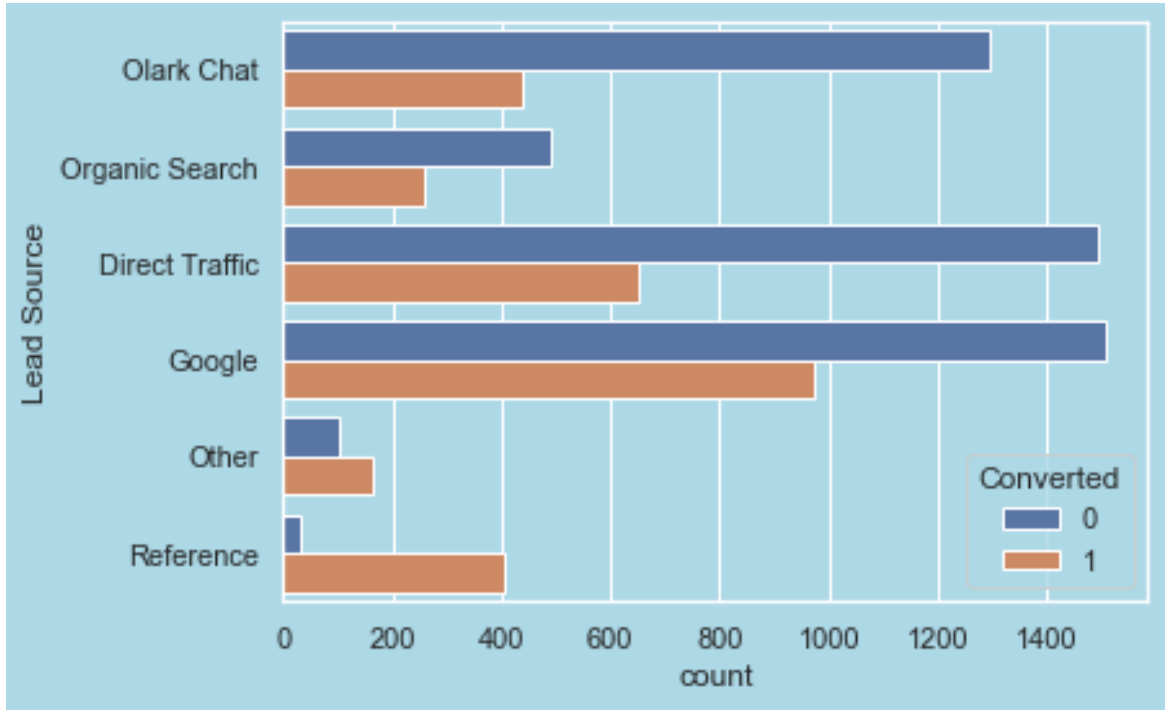
- Data Cleaning & Manipulation
- EDA
- Feature Scaling
- Dummy Variables
- Classification Technique in Logistic Regression
- Validation of the model
- Conclusions & Recommendations

DATA MANIPULATION

- The Dataset had 37 columns & over 9000 rows initially.
- We removed columns which had more than 4000 missing values.
- Removed remaining redundant columns which had more than 90% similar
- Assigned Blank columns

EDA

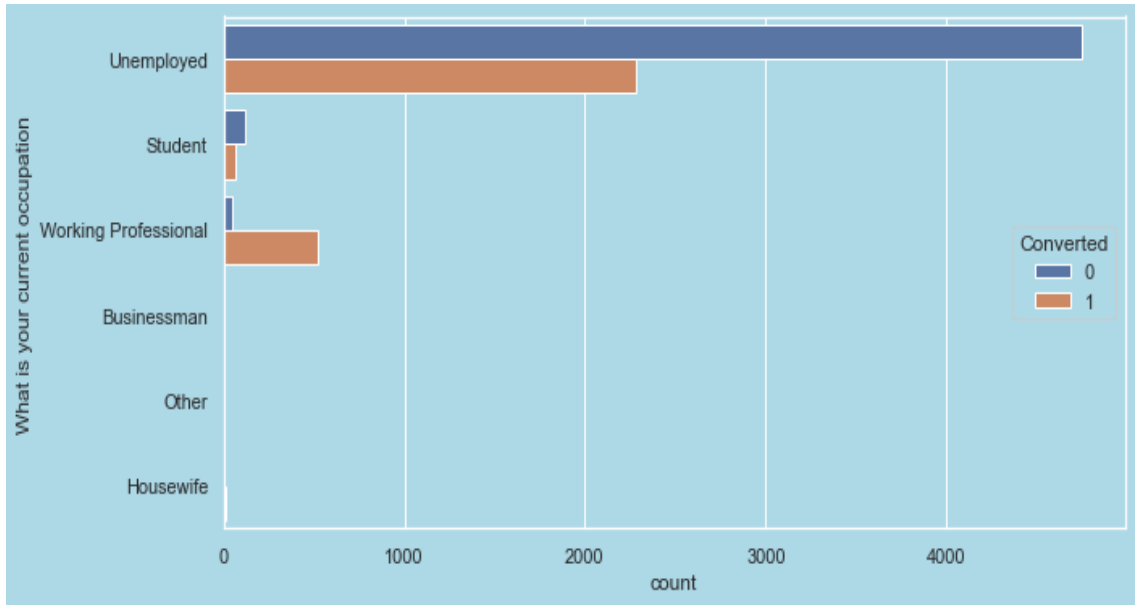




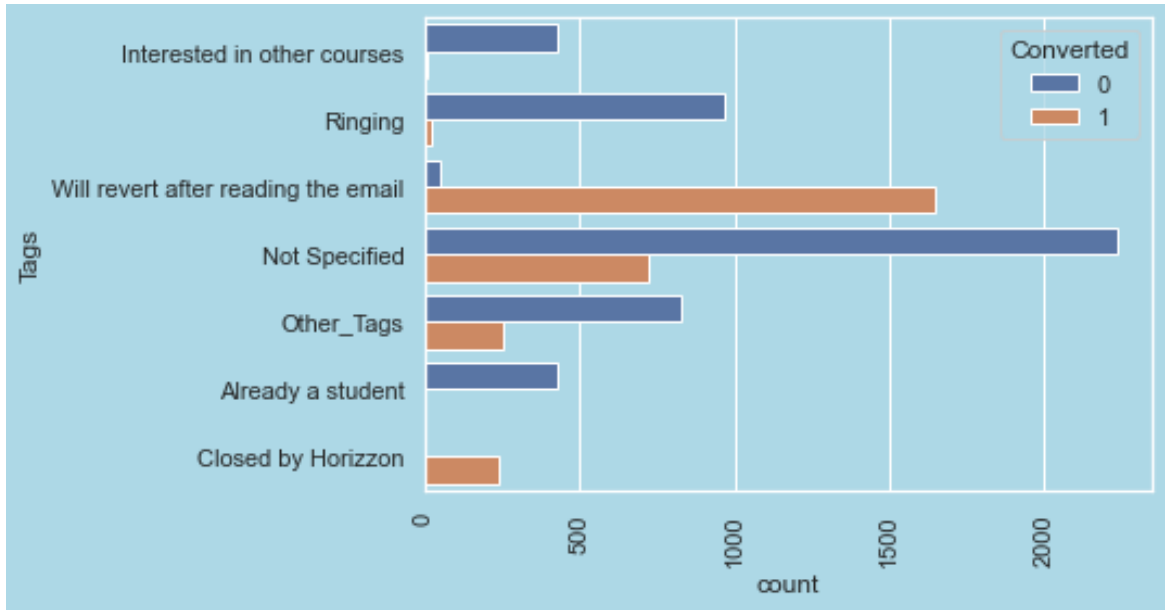
- Lead Add Form seems to have a higher conversion rate.
- Google as a lead source seems to bring a lot of promising leads.

EDA

- Working Professionals seem to have a higher conversion rate.

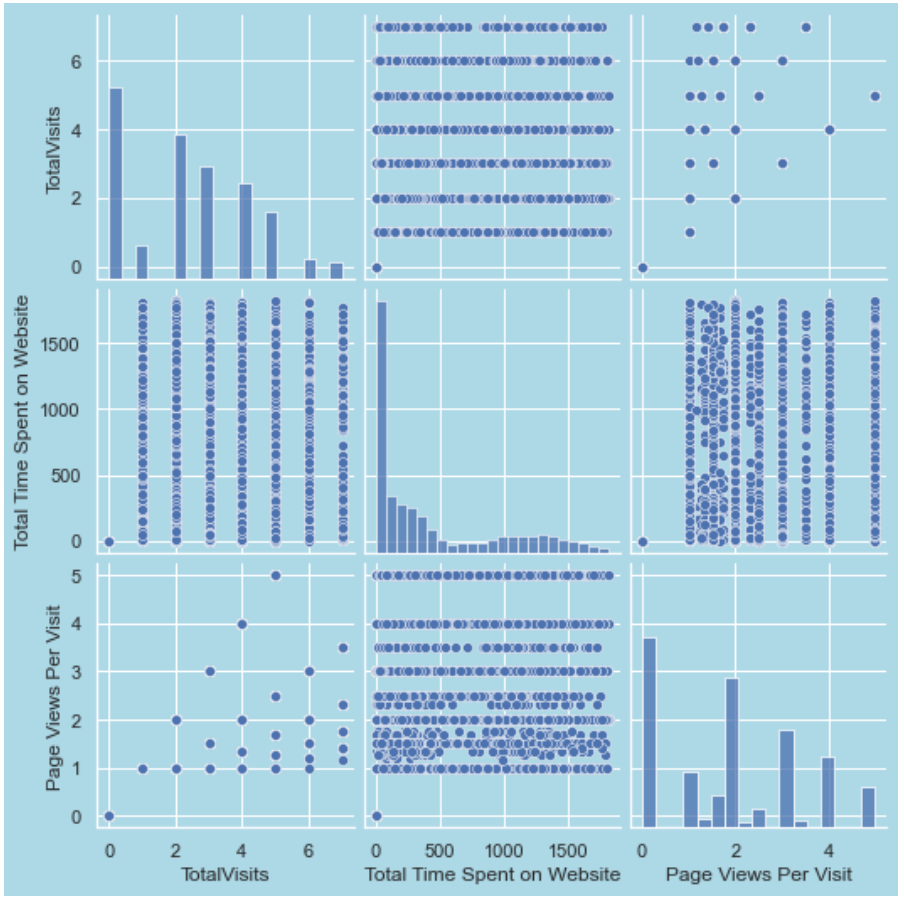


- Closed By horizzon & people who have reverted after reading email seem to have a good conversion rate.



EDA

- This graph shows us all the pair plots of all the numerical columns.
- Total Visits & page per views Visit are highly correlated which will bring in multicollinearity.



FEATURE SCALING & DUMMY VARIABLES

- 'Total Visits', ' Total Time Spent on Website' & 'Page Views Per Visit' were our numerical variables.
- For the following numerical variables we normalized them.
- Dummy variables were created for object type variables

MODEL BUILDING

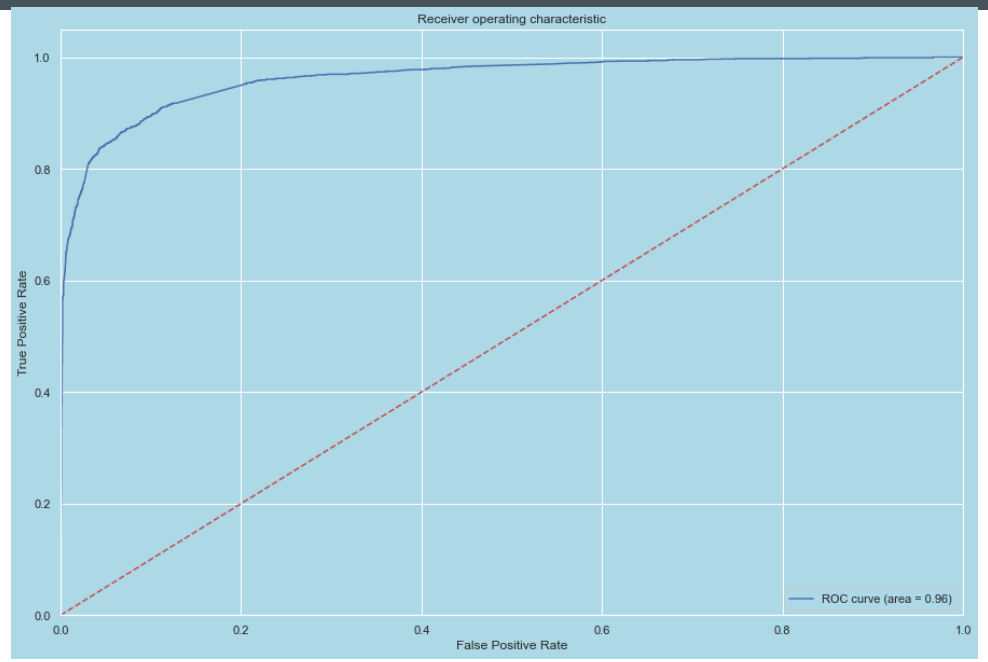
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5391	0.114	-4.738	0.000	-0.762	-0.316
Do Not Email	-1.3597	0.233	-5.838	0.000	-1.816	-0.903
Total Time Spent on Website	1.1123	0.061	18.354	0.000	0.994	1.231
Lead Origin_Lead Add Form	4.2790	0.547	7.827	0.000	3.207	5.350
What is your current occupation_Working Professional	1.0681	0.370	2.889	0.004	0.343	1.793
Lead Source_Direct Traffic	-1.6256	0.165	-9.863	0.000	-1.949	-1.303
Lead Source_Google	-1.3492	0.155	-8.686	0.000	-1.654	-1.045
Lead Source_Organic Search	-1.7560	0.220	-7.996	0.000	-2.186	-1.326
Lead Source_Reference	-3.1739	0.682	-4.653	0.000	-4.511	-1.837
Tags_Already a student	-3.4079	0.731	-4.662	0.000	-4.841	-1.975
Tags_Closed by Horizzon	6.0669	1.018	5.959	0.000	4.072	8.062
Tags_Interested in other courses	-2.4786	0.428	-5.791	0.000	-3.318	-1.640
Tags_Ringing	-3.5974	0.275	-13.097	0.000	-4.136	-3.059
Tags_Will revert after reading the email	4.1465	0.193	21.489	0.000	3.768	4.525
Last Activity_Olark Chat Conversation	-1.4456	0.216	-6.682	0.000	-1.870	-1.022
Last Notable Activity_SMS Sent	2.0981	0.125	16.732	0.000	1.852	2.344

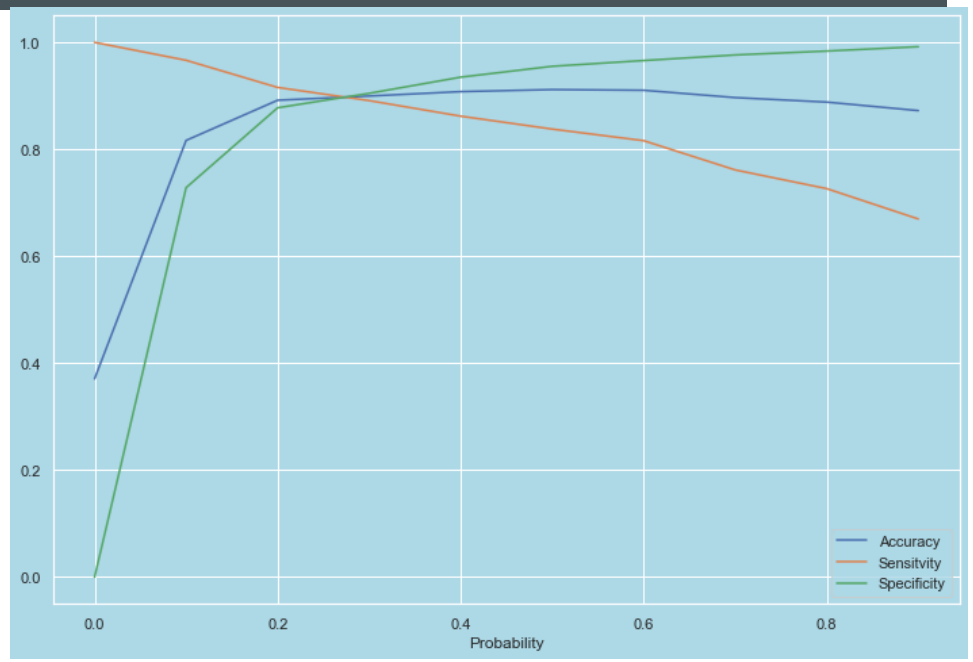
- We used Logistic regression model for given dataset to identify the following parameters which are mostly influencing our population.
- We also kept the VIF under 4 and p-value under 0.05.

	Features	VIF
7	Lead Source_Reference	3.73
2	Lead Origin_Lead Add Form	3.72
12	Tags_Will revert after reading the email	1.96
14	Last Notable Activity_SMS Sent	1.52
4	Lead Source_Direct Traffic	1.51
5	Lead Source_Google	1.50
3	What is your current occupation_Working Profes...	1.35
11	Tags_Ringing	1.32
1	Total Time Spent on Website	1.31
9	Tags_Closed by Horizzon	1.25
6	Lead Source_Organic Search	1.14
8	Tags_Already a student	1.13
0	Do Not Email	1.12
10	Tags_Interested in other courses	1.12
13	Last Activity_Olark Chat Conversation	1.12

ROC CURVE & OPTIMAL CUT-OFF POINT

- Roc area under the curve is close to **1**.
- Optimal cut-off point tends to approximately 0.3





CONCLUSIONS

Tags_Closed by Horizzon	6.066858
Lead Origin_Lead Add Form	4.278953
Tags_Will revert after reading the email	4.146513
Last Notable Activity_SMS Sent	2.098084
Total Time Spent on Website	1.112282
What is your current occupation_Working Professional	1.068061
Lead Source_Google	-1.349208
Do Not Email	-1.359653
Last Activity_Olark Chat Conversation	-1.445589
Lead Source_Direct Traffic	-1.625563
Lead Source_Organic Search	-1.755966
Tags_Interested in other courses	-2.478624
Lead Source_Reference	-3.173851
Tags_Already a student	-3.407906
Tags_Ringing	-3.597406

The following parameters are the major influencers in our analysis.

The accuracy of the model is 89.96%.

Train Data Set metrics:Test Data Set metrics

Sensitivity: 89.11 vs Sensitivity: 89.17

Specificity: 90.47 vs Specificity: 88.36
Precision: 84.64 vs Precision: 83.95

Recall: 89.11 vs Recall: 89.17
Accuracy: 89.96 vs Accuracy: 88.66

RECOMMENDATIONS

- Phone calls must be made to those customers that spend most time on our website as they are seem more interested in the course.
- Horizzon has proved to be a good partner. Therefore, we must provide better funding.
- Customers who revert after reading the email should be called ASAP.
- Target customers who aren’t students as they are least likely to join our program.
- Make website more appealing so people are likely to spend more time on our website.