Lead Scoring Case Study

Submitted By

Vasu Kaul

Problem Statement

- An education company named X Education sells online courses to industry professionals.
- Although X Education gets a lot of leads, the lead conversion rate is very poor.
- The CEO wants to increase the conversion rate to 80%.

Objective

- To build a model wherein a ‘lead score’ is assigned 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.
- A high score will mean that the lead is ‘HOT’, i.e. is most likely to be converted whereas lower score will mean that the lead is not likely to be converted.

Data cleaning and data manipulation.

- There are no duplicate values present in the dataframe.
- Missing values have been handled with mode values where the null values are less than 40%.
- Columns with more than 40% missing data have been dropped.
- Outliers have been removed from “Page Views Per Visit”.
- “Prospect ID” & “Lead number” columns have also been removed as they don’t help with our analysis.

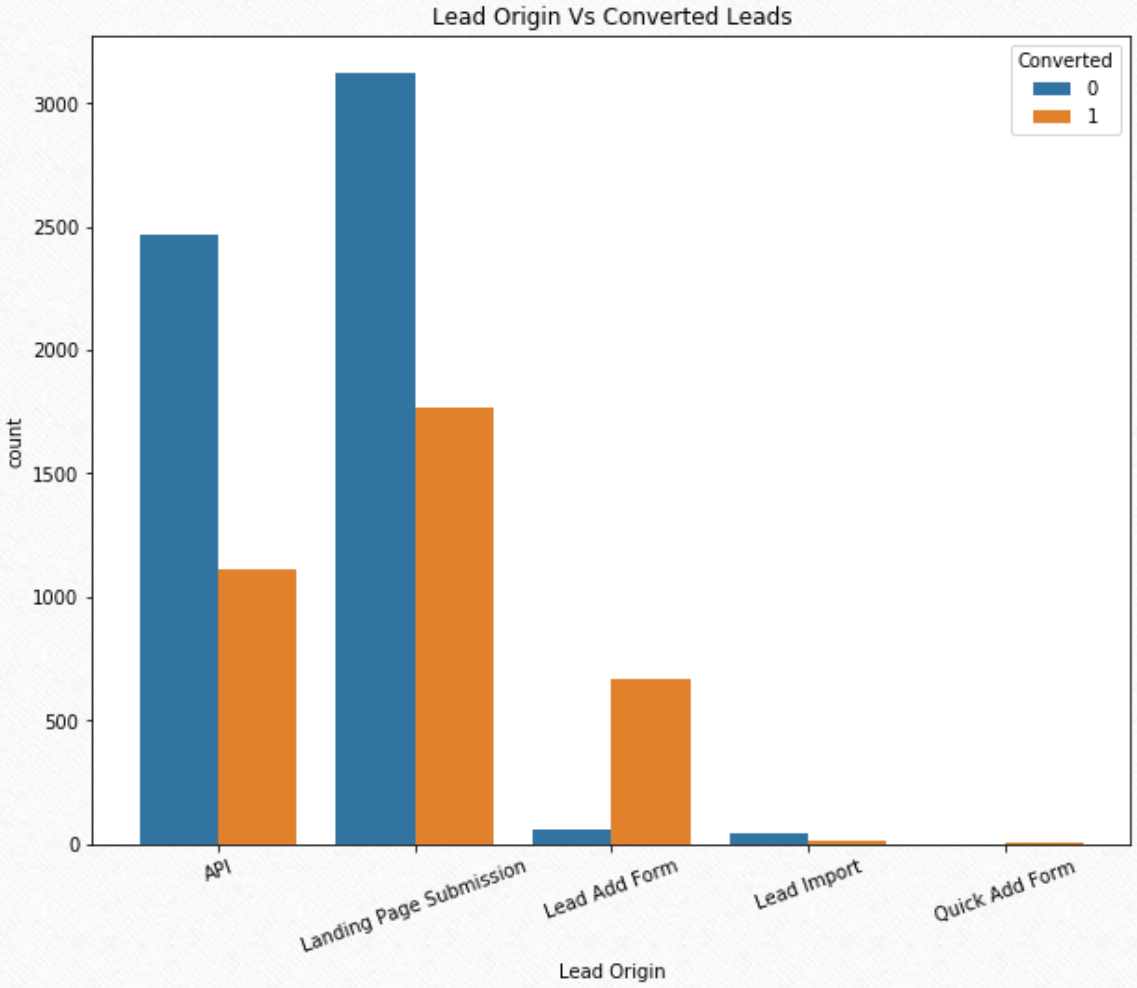
Exploratory Data Analysis

Pair plot Inference

- Asymmetrique Activity Score & Asymmetrique Profile Score are evenly distributed in terms of conversion and non-conversion ,for lower values and however Asymmetrique Activity Score shows an even distribution from mid to high range as converted but an even distribution can be seen for Asymmetrique Profile Score for lower to higher values. And with respect to each other show an almost even distribution across all almost all values. Whereas with respect to Total visits both show a distribution across lower values and almost no observations are present for higher values.
- Total Visits show lower to mid range values as not converted while very low and very high values as converted. Total Visits do not vary very much with respect to Total time spent on websites, increase in Total time spent on websites does not increase the Total Visits. Total visits vs Page views per visit tend to be concentrated in the lower ranges, page views do not show increase in value with increase in number of visits.
- Page View per visit show an even distribution for low to mid range with respect to converting or not converting. Page per view does not increase with increase in total time spent on the website.
- Total time spent on websites show an even distribution of converted and non-converted leads

Numerical Columns

API and Landing Page Submission are the highest contributors in terms of bringing in new leads, also Landing page Submission has more percentage of conversions than API.

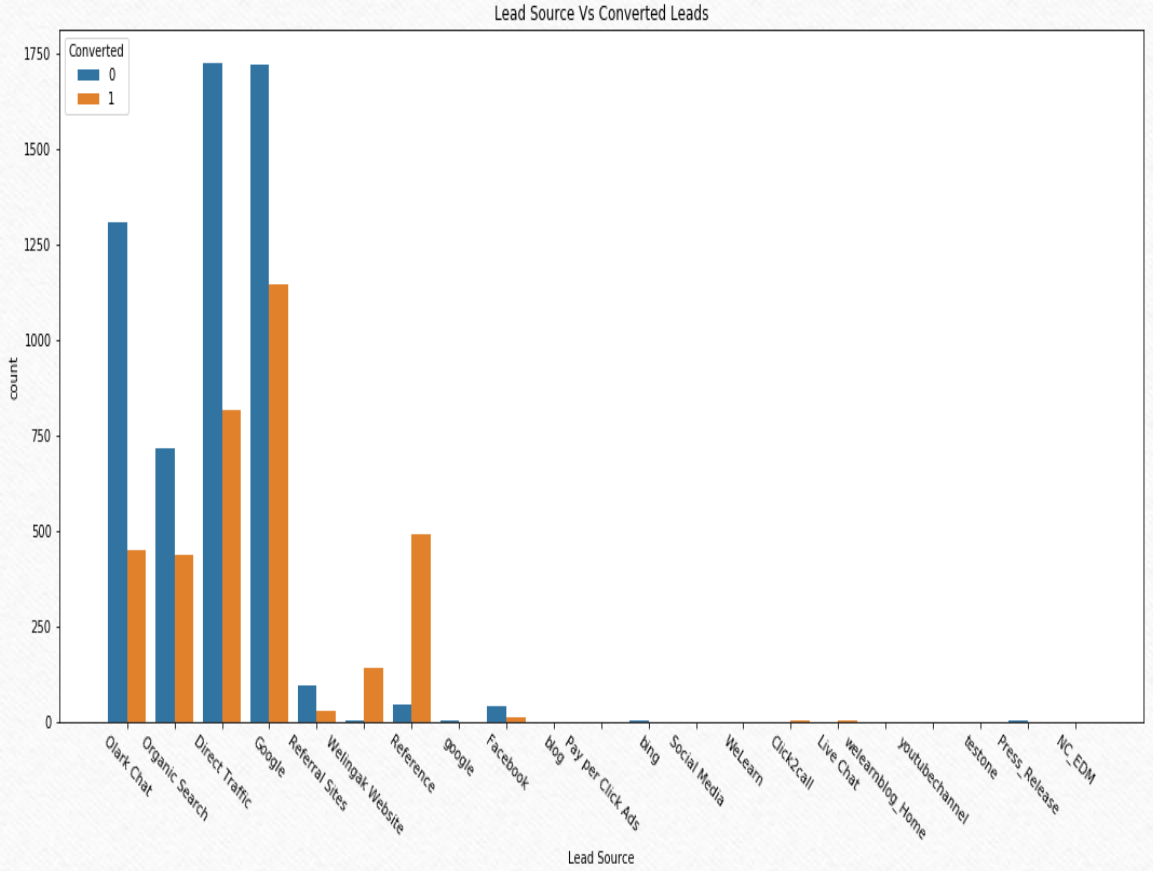


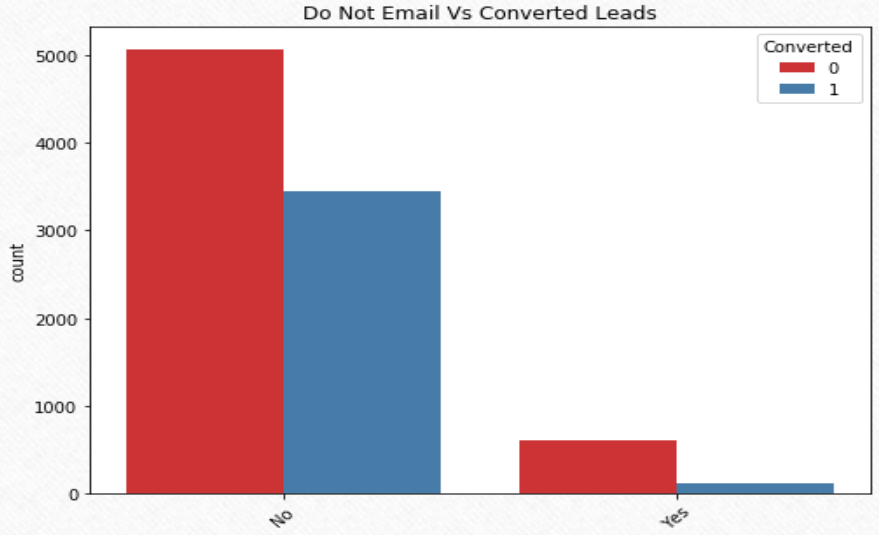
Lead Source Vs Converted Leads

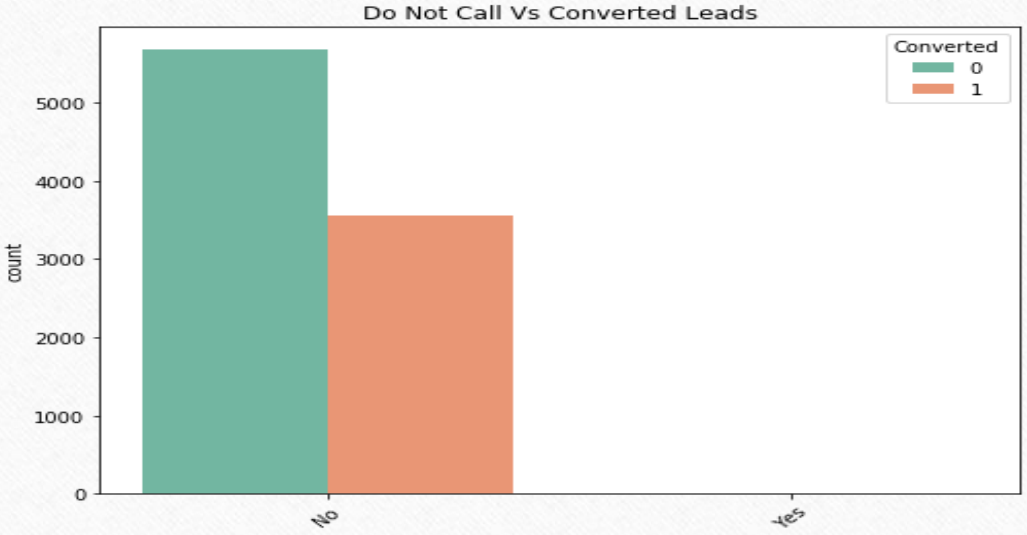
In the above plot it can be seen there are multiple sources of leads and some have contributed next to negligible, like testone, youtubechannel and Welearn.

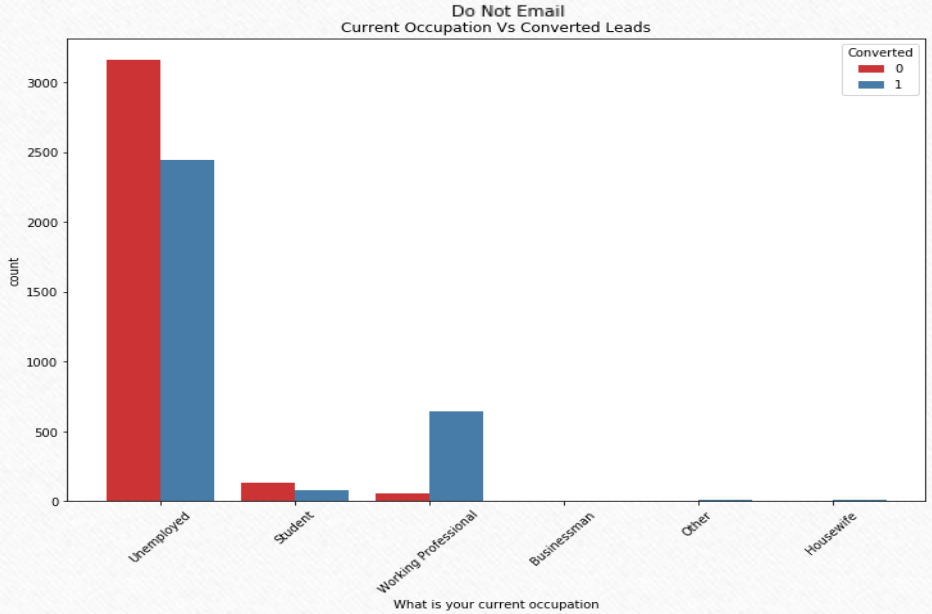
However olark Chat , Organic Search, Direct Traffic and Google have contributed to the most number of leads, converted & non-converted together.

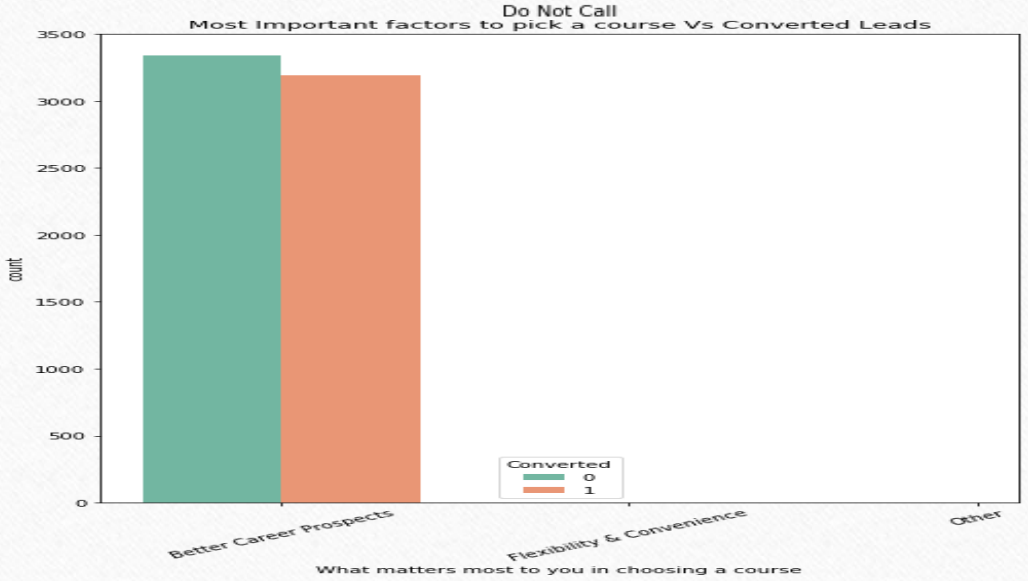
Google has the maximum number of leads converted compared to other lead sources.

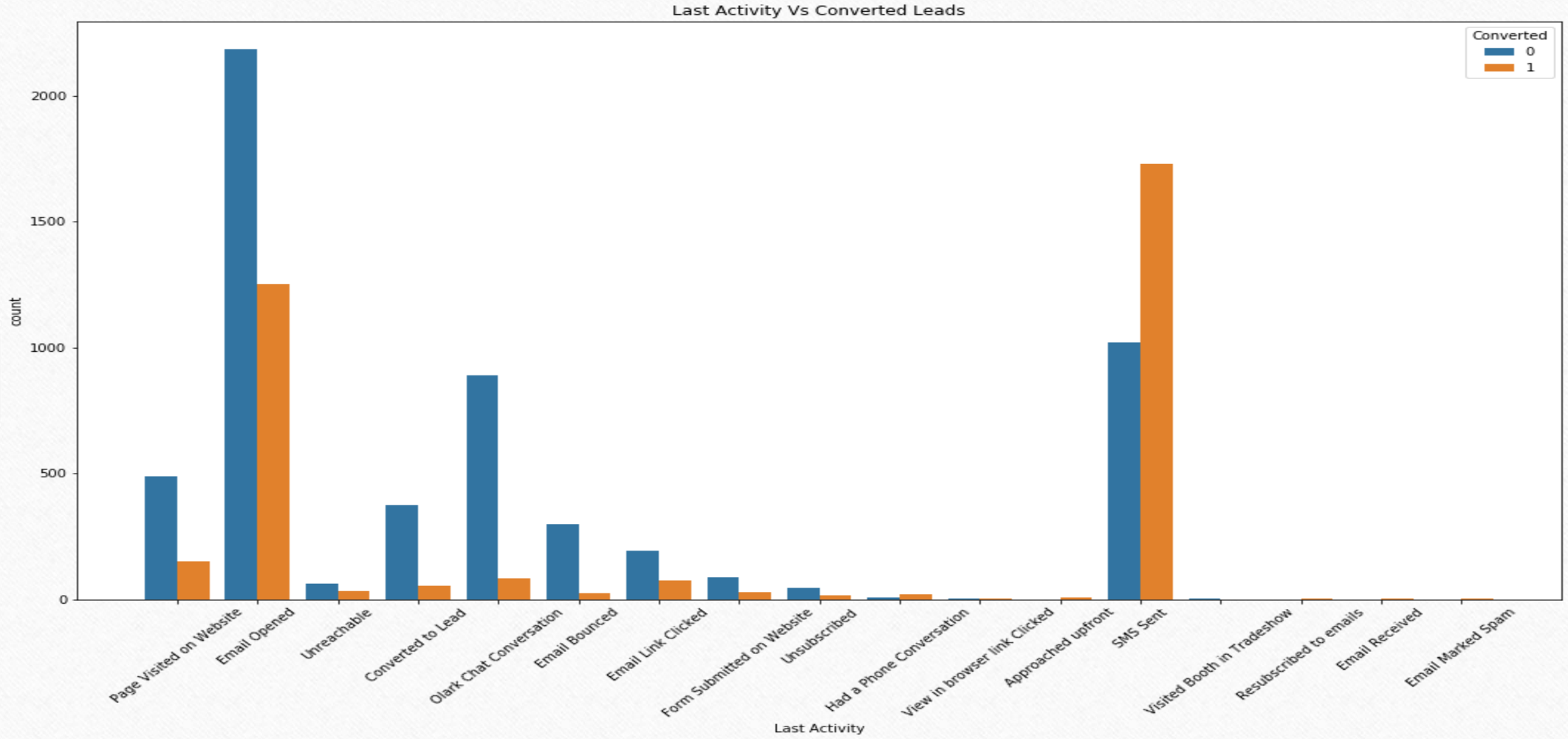


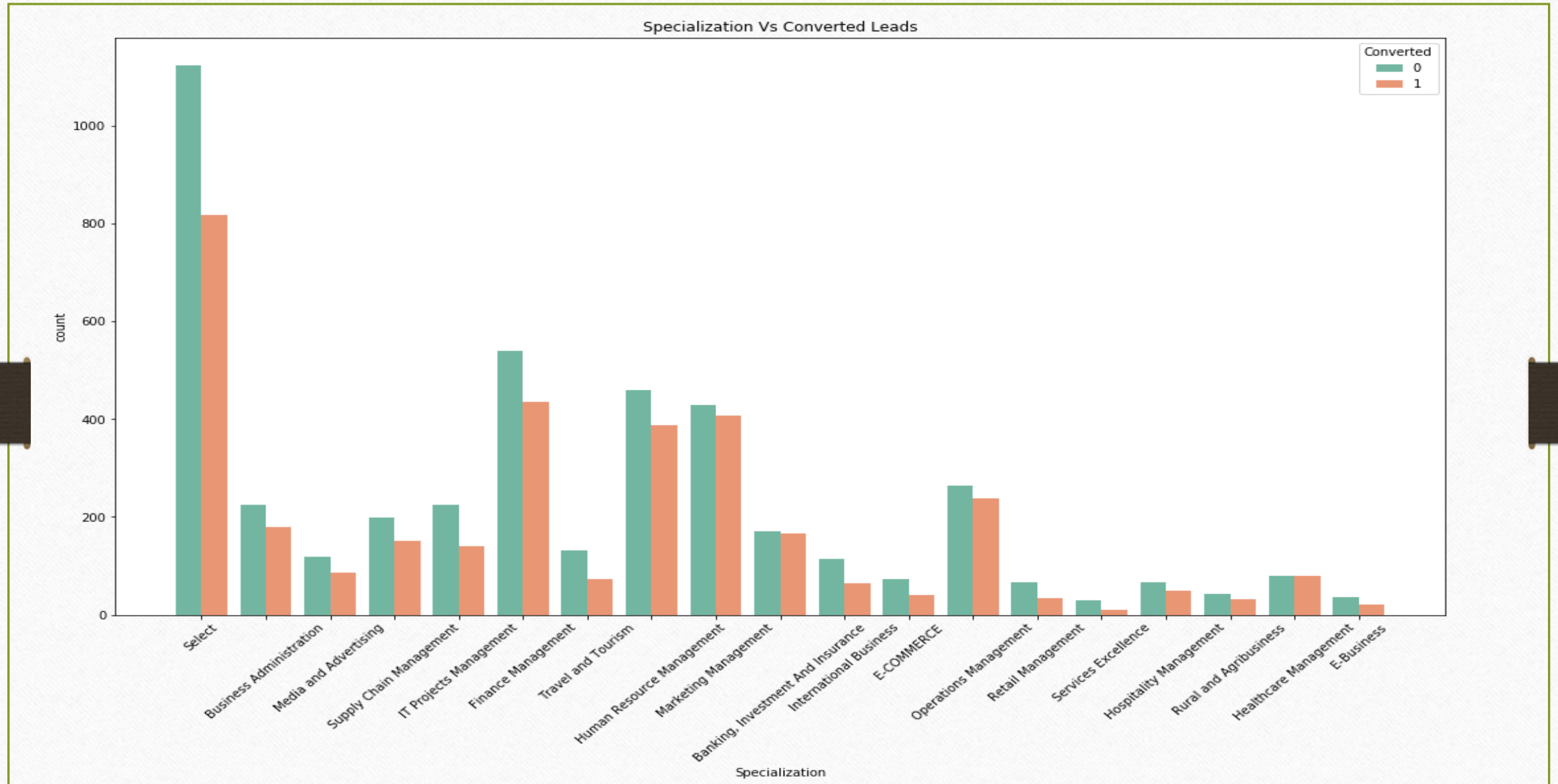

Do Not Email Vs Converted Leads
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes


Do Not Call Vs Converted Leads
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes


Do Not Email
Current Occupation Vs Converted Leads
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation


Do Not Call
Most Important factors to pick a course Vs Converted Leads
Converted
0
1
count
3500
3000
2500
2000
1500
1000
500
0
Better Career Prospects
Flexibility & Convenience
Other
What matters most to you in choosing a course


Last Activity Vs Converted Leads
Converted
0
1
count
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
Approached upfront
SMS Sent
Visited Booth in Tradeshow
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity


Specialization Vs Converted Leads
Converted
0
1
count
0
200
400
600
800
1000
Select
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization

Data Conversion

- Numerical Variables are standardised.
- Dummy Variables are created for object type variables
- Data is standardised using standard scaler.
- Total Rows for Analysis: 8844
- Total Columns for Analysis: 39

Model Building

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variables where p-value is greater than 0.05 and VIF value is greater than 5
- Predictions on test data set
- Overall accuracy 77%

Running RFE with the Top 15 Variables

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6190
Model:	GLM	**Df Residuals:**	6174
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2880.0
Date:	Sun, 10 Jan 2021	**Deviance:**	5760.0
Time:	17:07:53	**Pearson chi2:**	8.30e+03
No. Iterations:	21		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.7140	1.030	0.693	0.488	-1.305	2.733
Total Time Spent on Website	1.0786	0.038	28.253	0.000	1.004	1.153
Lead Origin_Landing Page Submission	-1.1018	0.124	-8.852	0.000	-1.346	-0.858
Lead Origin_Lead Add Form	3.4317	0.206	16.662	0.000	3.028	3.835
Lead Source_Olark Chat	0.6673	0.113	5.890	0.000	0.445	0.889
Lead Source_Others	-0.8828	1.042	-0.847	0.397	-2.926	1.160
Lead Source_Referral Sites	-0.5415	0.303	-1.784	0.074	-1.136	0.053
Lead Source_Welingak Website	3.1772	1.024	3.103	0.002	1.171	5.184
Specialization_Others	-1.2491	0.118	-10.568	0.000	-1.481	-1.017
Specialization_Retail Management	-0.5504	0.334	-1.648	0.099	-1.205	0.104
Specialization_Travel and Tourism	-0.3674	0.245	-1.502	0.133	-0.847	0.112
What is your current occupation_Housewife	21.8083	1.31e+04	0.002	0.999	-2.56e+04	2.56e+04
What is your current occupation_Other	-1.7515	1.246	-1.406	0.160	-4.193	0.691
What is your current occupation_Student	-0.5636	1.049	-0.537	0.591	-2.619	1.492
What is your current occupation_Unemployed	-0.7566	1.028	-0.736	0.462	-2.771	1.258
What is your current occupation_Working Professional	1.6729	1.042	1.605	0.108	-0.370	3.715

Linear Regression Model Results

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	2654
Model:	GLM	**Df Residuals:**	2646
Model Family:	Binomial	**Df Model:**	7
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1358.1
Date:	Sun, 10 Jan 2021	**Deviance:**	2716.2
Time:	09:07:05	**Pearson chi2:**	2.76e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0967	0.160	0.603	0.547	-0.218	0.411
Total Time Spent on Website	1.0540	0.056	18.692	0.000	0.943	1.164
Lead Origin_Landing Page Submission	-0.9414	0.169	-5.555	0.000	-1.274	-0.609
Lead Origin_Lead Add Form	3.2630	0.299	10.928	0.000	2.678	3.848
Lead Source_Olark Chat	1.0647	0.165	6.445	0.000	0.741	1.388
Lead Source_Welingak Website	2.6186	1.052	2.489	0.013	0.557	4.680
Specialization_Hospitality Management	-0.9342	0.422	-2.216	0.027	-1.760	-0.108
Specialization_Others	-1.4986	0.165	-9.097	0.000	-1.821	-1.176

Evaluating The Model on Train Dataset

Confusion Matrix

```
[1266  350]
[255   783]
```

Optimal Cut-off

0.3

TP = confusion2[1,1] # true positives
TN = confusion2[0,0] # true negatives
FP = confusion2[0,1] # false positives
FN = confusion2[1,0] # false negatives

Sensitivity of the Logistic Regression Model

TP / (TP+FN) = 0.75

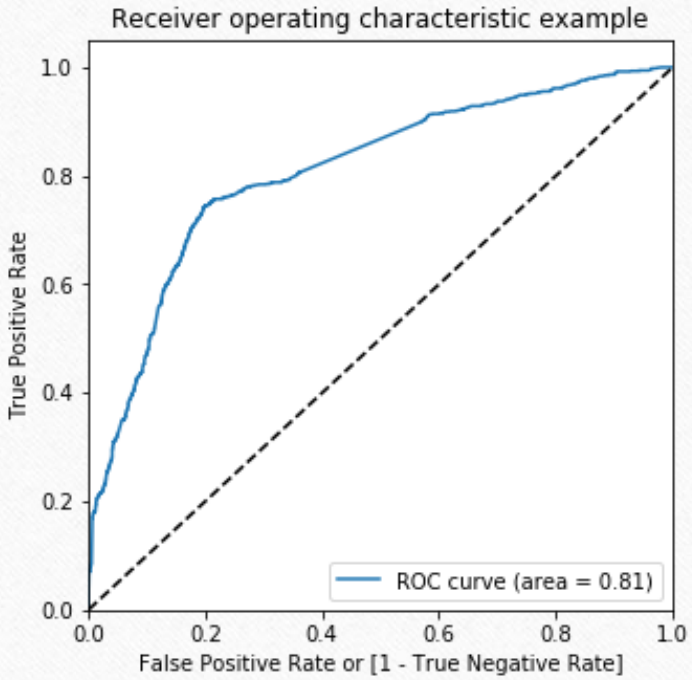
Specificity of the Logistic Regression Model

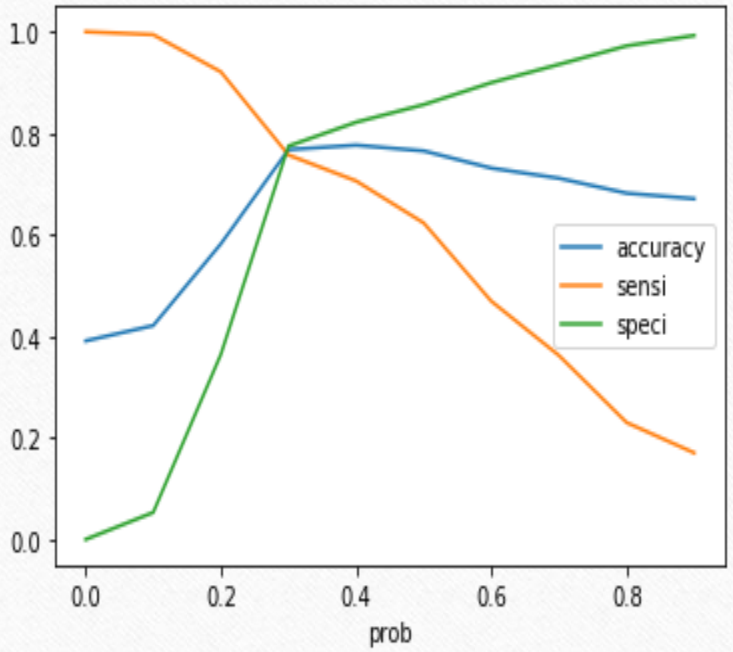
TN / (TN+FP) = 0.78

Final Model Features

	Features	VIF
3	Lead Source_Olark Chat	1.83
6	Specialization_Others	1.76
2	Lead Origin_Lead Add Form	1.29
4	Lead Source_Welingak Website	1.28
0	Total Time Spent on Website	1.21
1	Lead Origin_Landing Page Submission	1.05
5	Specialization_Hospitality Management	1.02

ROC Curve


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.81)
0.0
0.2
0.4
0.6
0.8
1.0


accuracy
sensi
speci
prob
0.0
0.2
0.4
0.6
0.8
1.0

ROC Curve

- Finding Optimal Cut off Point.
- Optimal cut off probability is that.
- Probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.3

Lead Score

- The train & test dataset is concatenated to get the entire ;list of leads available.
- The conversion probability is multiplied by 100 to obtain the “Lead Score” for each lead.
- Higher the “Lead Score”, higher is the probability of getting the lead converted.

	Prospect ID	Lead_Prob
0	8891	0.789645
1	5928	0.174797
2	4779	0.946230
3	3295	0.145205
4	42	0.135670

Prospect ID	Lead Score
8891	78.96
5928	17.48
4779	94.62
3295	14.52
42	13.57

Conclusion

Most Important variables for the potential buyers are (In descending order) :

- The total time spend on the Website.
- Total number of visits.
- The lead source is: a. Google b. Direct traffic c. Organic search d. Welingak website
- The last activity is: a. SMS b. Olark chat conversation
- The lead origin is Lead add format.
- When the current occupation is as a working professional.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.