LEAD SCORING CASE STUDY

Submitted by

Kalla Pavan ajay

Swapnali Patil

Raju Nagineni

Problem Statement:

- X Education sells online courses to industry professionals on their websites.
- X Education marketing through search engines & generating leads which has very less conversion rate of 30%.
- To make the process more efficient the company wants us to identify hot & promising leads.
- If we success in identifying the hot leads, it will help the company with marketing them & improve the leads conversion rate.

Business Objective:

- X Educations wants to know which leads are promising & which leads are less chances to get converted.
- The company need us to build a logistic regression model to identify the potential leads by assigning the lead score to sort the hot leads from 0-100.
- We should build a model in a way such that in future it should the handle these as well.

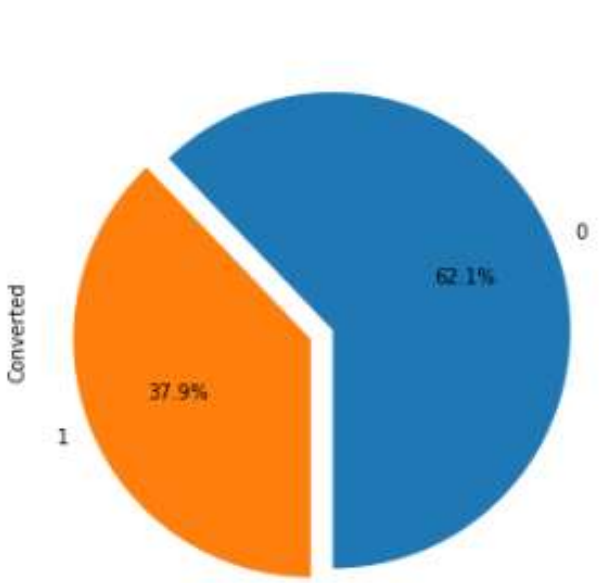
Following approach to the case study:

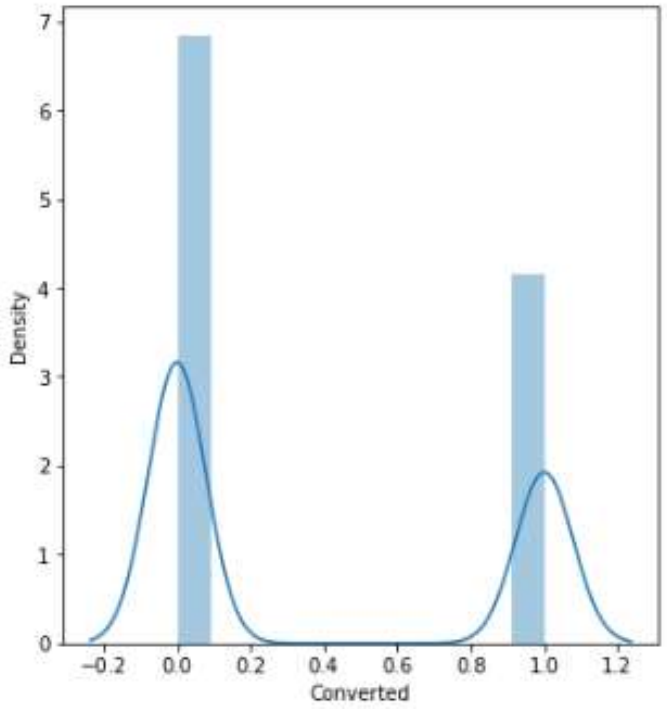
- ❑ Understanding Problem statement
- ❑ Understanding Data
- ❑ Data Cleaning by handling missing values and unique variables.
- ❑ Exploratory Data Analysis by performing Univariate , Bivariate and Multi-variate analysis
- ❑ Data interpretation
- ❑ Data preparation for Modelling
- ❑ Logistic regression Model building
- ❑ Model Evaluation
- ❑ Conclusion

Data cleaning & Manipulation

- Identified the unique values in the columns of data & dropped which are less important for the analysis.
- Some of the category columns have values as SELECT which are as good as nulls were replaced by Not Available.
- Columns having the null values are identify those having >35% nulls were dropped to make the data precise.
- The rows that having less % of nulls were dropped.

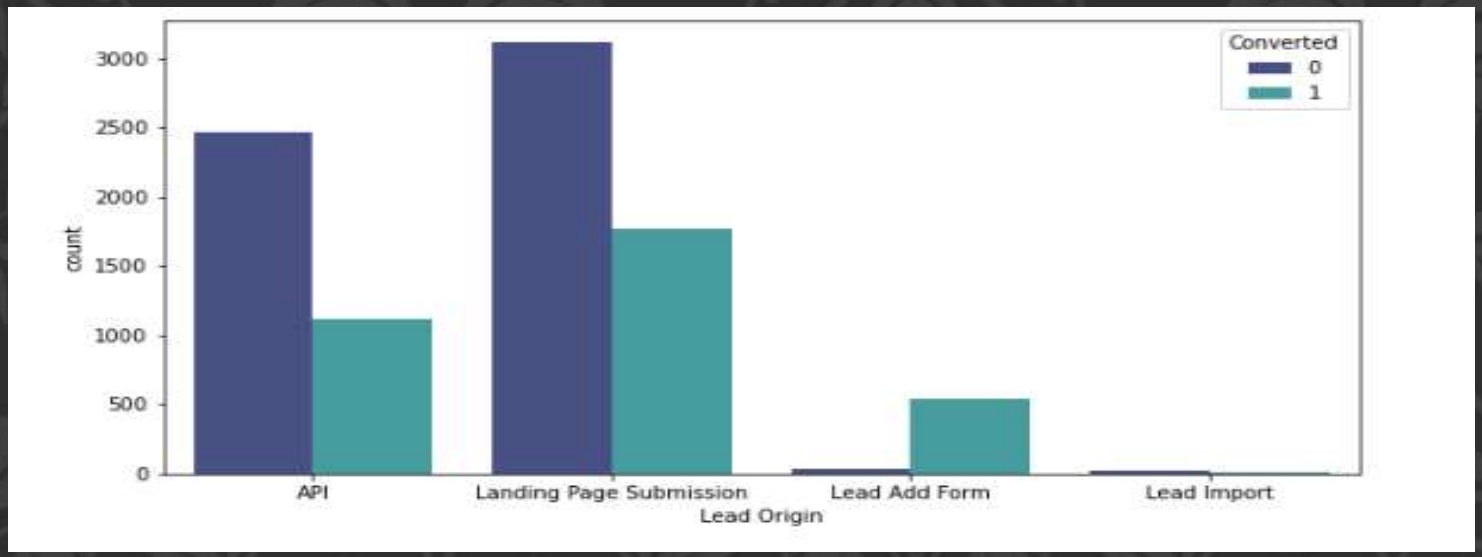
EXPLORATORY DATA ANALYSIS

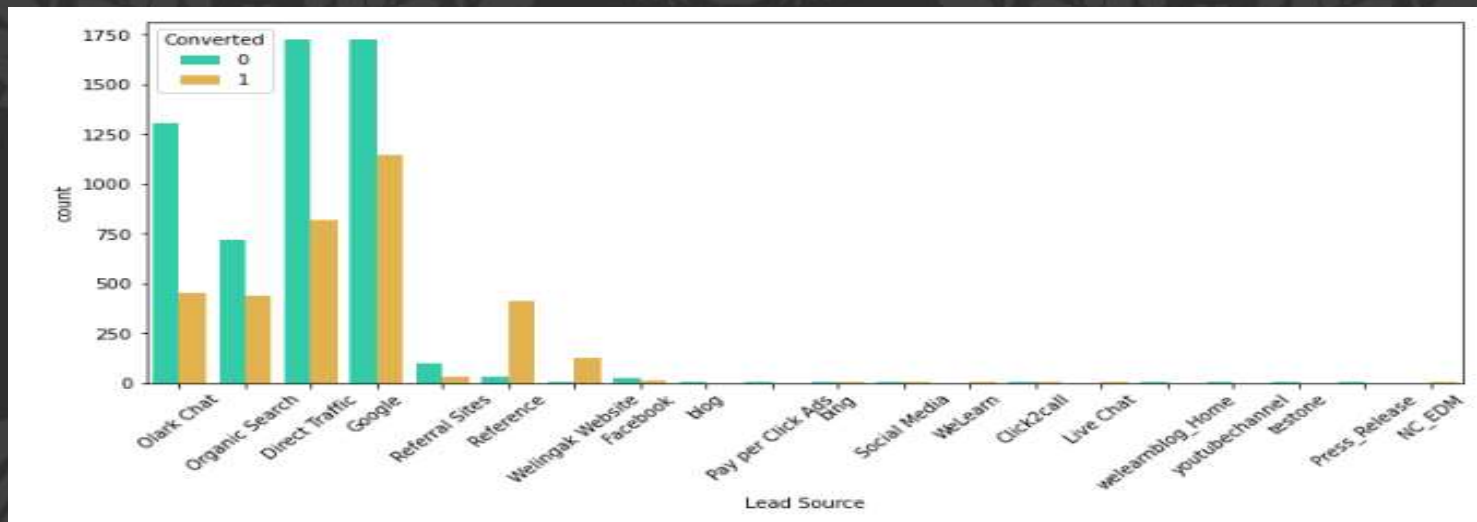




From the plot we can observe that 37.9% are converted & rest are not converted

Categorical variable analysis

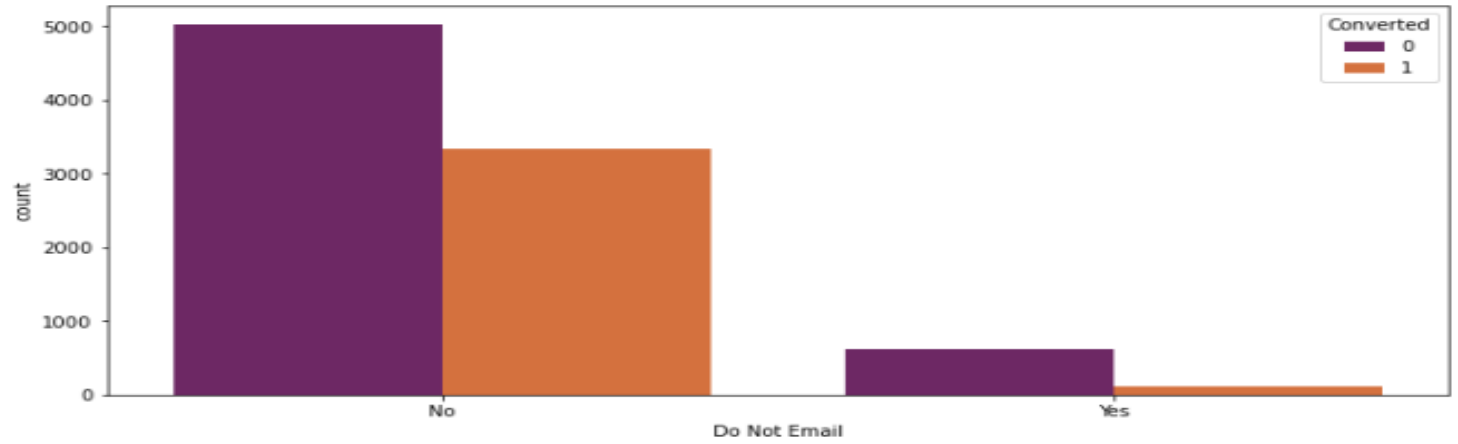




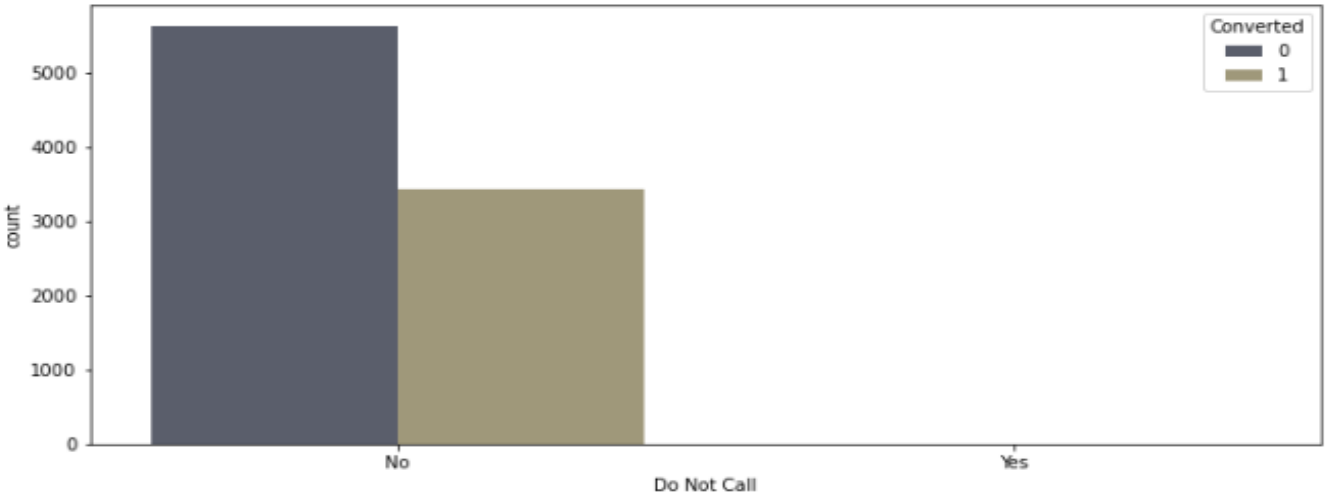
From the plot we can analyse that Direct traffic & google shows more count of lead sources

Lead conversion rate is high for Reference & Wellingak Website

Categorical variable analysis

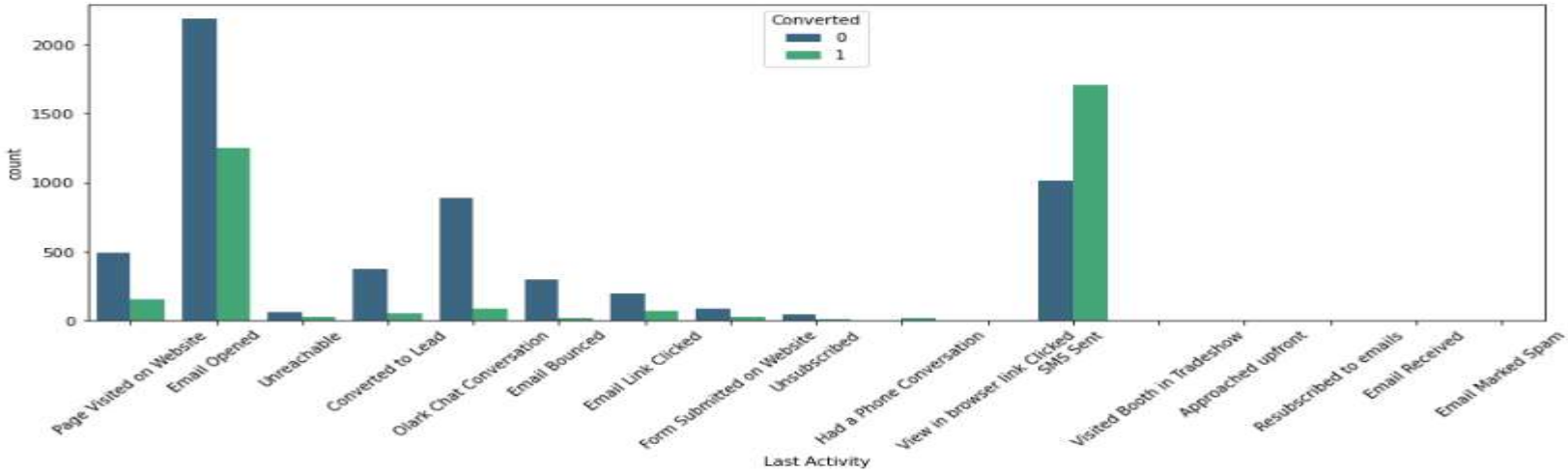


From the plot we can say that majority of the leads preffered mail them & shows more conversion rate



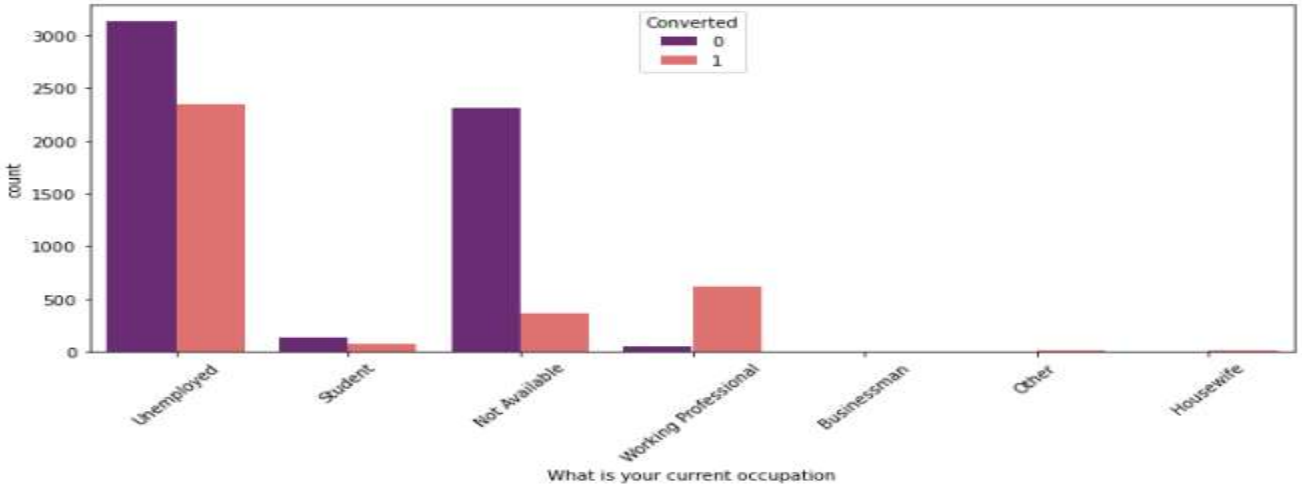
from the plot we can see the leads generated through calls & shows some decent conversion rate

Categorical variable analysis



More than 2000 leads are who opened email

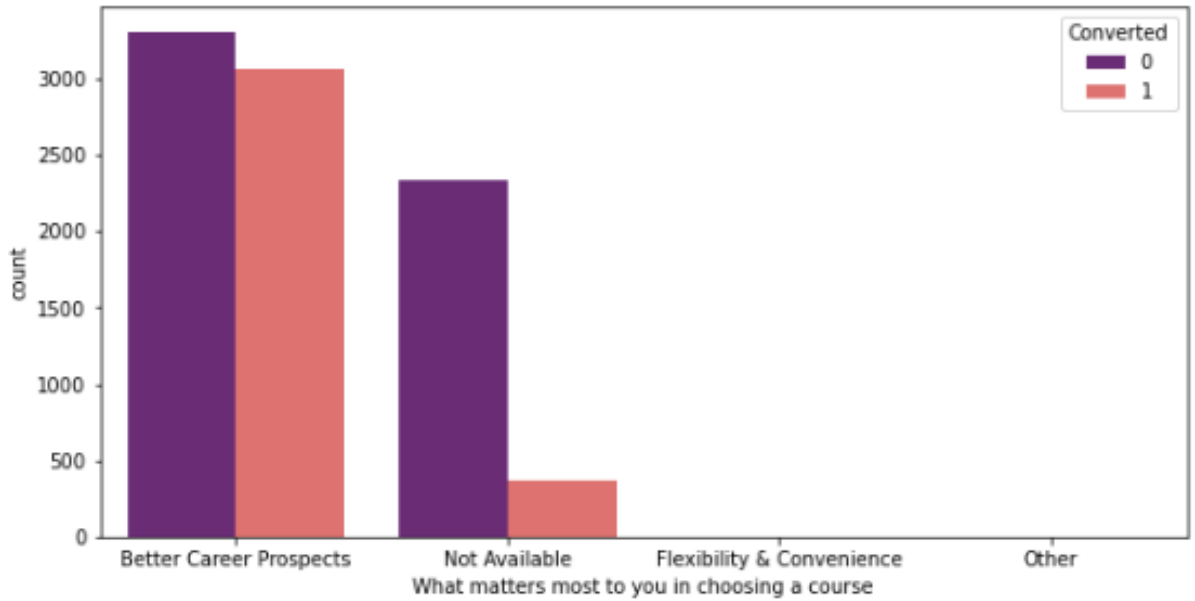
For leads to whom sms sent show very high conversion rate of arround ~1600



From the plot we can see Unemployed category more no of leads & shows good conversion rate

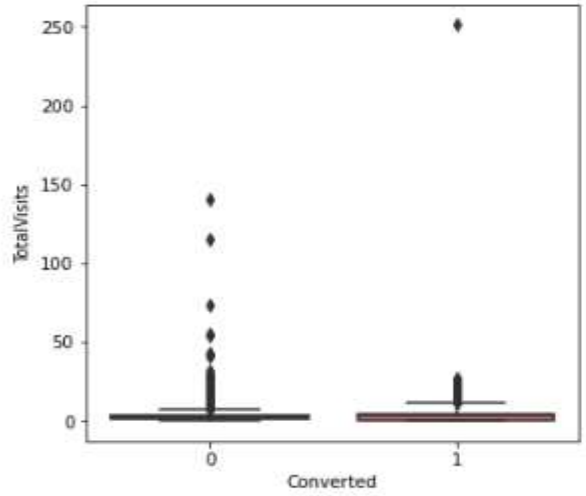
But majority of the conversion rate shown by Working professionals

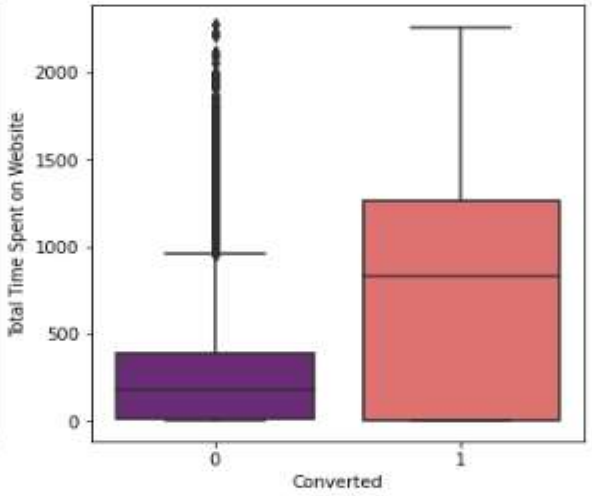
Categorical variable analysis

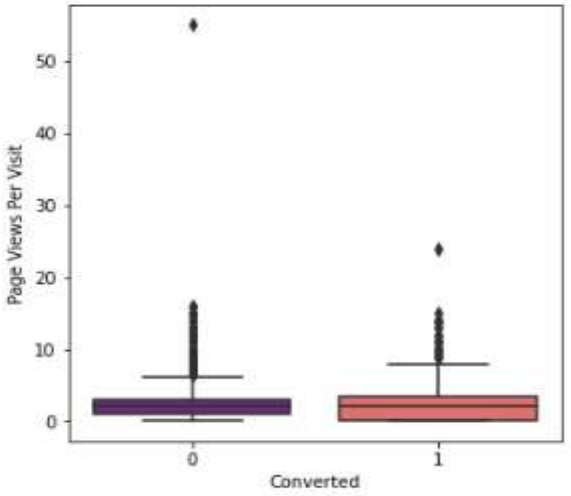


From the plot we can see Better career prospects shows more leads & good conversion rate

Numerical variable analysis

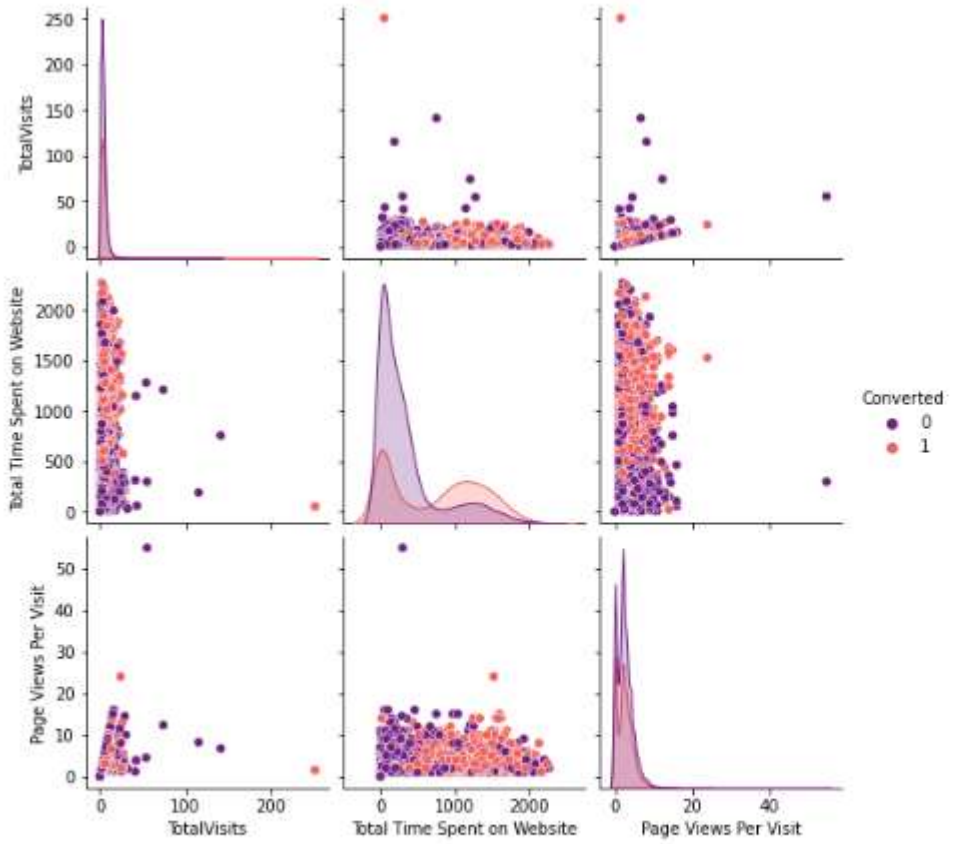






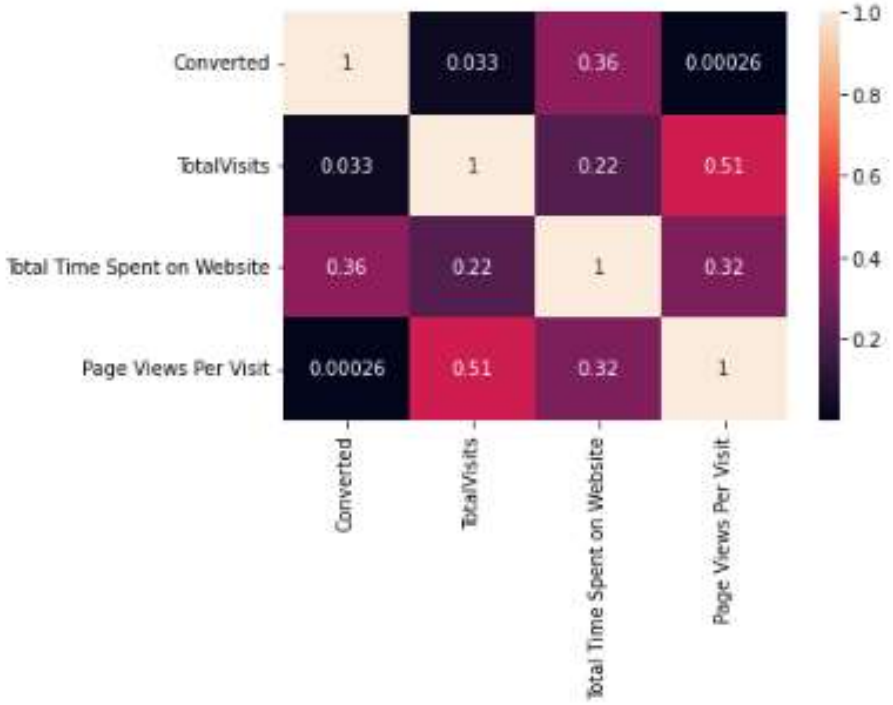
For a range of ~1000 average hours spent on website showing good conversion rate

Pair plots on numerical variables



From the plot we can see Total time spent of > 1000 hours shows more conversion rate with average page visits being 10

Correlation matrix



From the heat map above we can say Total visits shows high correlation with pages per visit

Data Preparation

- Dummy variables were created for all Object type datatypes.
- Total Rows for Analysis: 9074
- Total Columns for Analysis: 20

Model building

- Train - Test split was done at 70% and 30% respectively.
- RFE was done to attain the top 15 relevant variables.
- Removed variables manually depending on the VIF values and p-value.
- Predictions on test data set
- Overall accuracy 81%

Final logistic regression Model

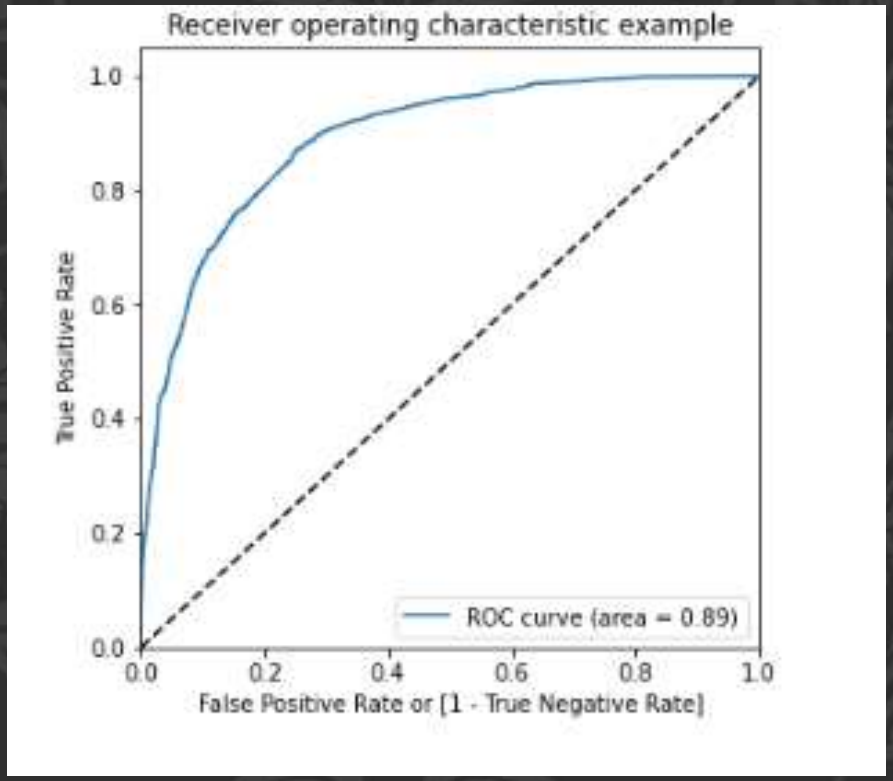
	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.0170	0.191	-21.072	0.000	-4.391	-3.643
Total Time Spent on Website	1.1357	0.041	27.825	0.000	1.056	1.216
Lead Origin_Lead Add Form	3.5894	0.221	16.272	0.000	3.157	4.022
Lead Source_Olark Chat	1.4255	0.107	13.359	0.000	1.216	1.635
Lead Source_Welingak Website	1.9765	0.751	2.631	0.009	0.504	3.449
Do Not Email_No	1.6427	0.172	9.572	0.000	1.306	1.979
Last Activity_Converted to Lead	-1.2969	0.225	-5.764	0.000	-1.738	-0.856
Last Activity_Olark Chat Conversation	-1.3554	0.164	-8.264	0.000	-1.677	-1.034
Last Activity_SMS Sent	1.2274	0.075	16.282	0.000	1.080	1.375
What is your current occupation_Working Professional	2.5096	0.189	13.266	0.000	2.139	2.880
What matters most to you in choosing a course_Better Career Prospects	1.3146	0.088	14.916	0.000	1.142	1.487
Last Notable Activity_Had a Phone Conversation	3.5674	1.119	3.189	0.001	1.375	5.760
Last Notable Activity_Unreachable	2.0132	0.495	4.063	0.000	1.042	2.984

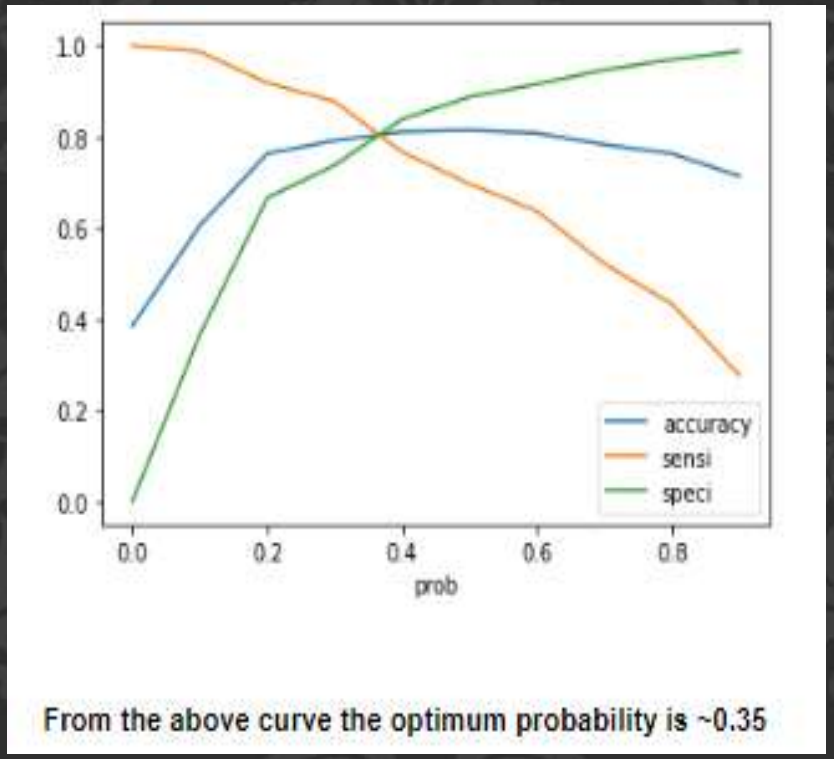
VIF scores

	Features	VIF
4	Do Not Email_No	4.08
9	What matters most to you in choosing a course_...	3.35
2	Lead Source_Olark Chat	1.76
1	Lead Origin_Lead Add Form	1.63
7	Last Activity_SMS Sent	1.60
6	Last Activity_Olark Chat Conversation	1.43
3	Lead Source_Welingak Website	1.33
0	Total Time Spent on Website	1.31
8	What is your current occupation_Working Profes...	1.20
5	Last Activity_Converted to Lead	1.10
11	Last Notable Activity_Unreachable	1.01
10	Last Notable Activity_Had a Phone Conversation	1.00

We have good VIF score & p value so the model is fit

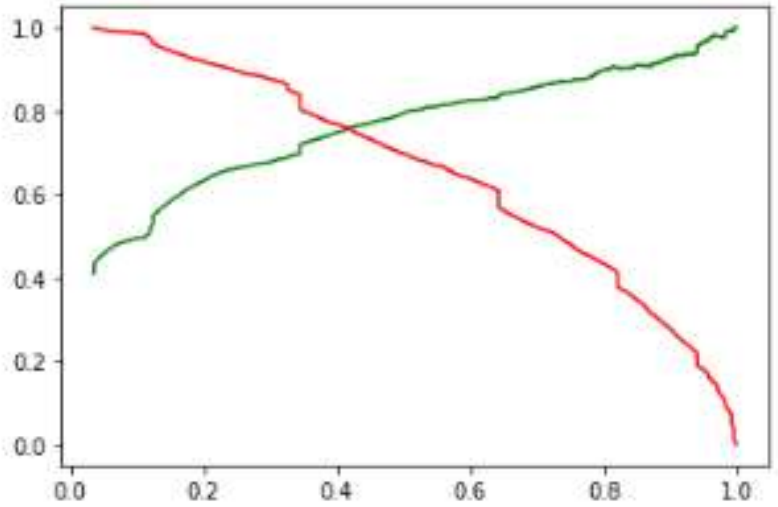
ROC Curve





Cut-off Probability we found is ~0.35

Precision & Recall Curve



From the plot we got Precision 79% & Recall 69% at a cuttoff probability as 0.41

CONCLUSION

1) From the given data the conversion rate was 39% .

2) When 'Lead Origin' is 'Lead add form', the probability of lead getting converted is high.

3) Google and direct traffic generating majority of leads, but referral sites shows high Conversion rate.

4) Most of the leads are opting for mailing them.

5) Conversion rate is higher when the information is sent through SMS .

6) Unemployed people have more conversion rate as well as more count.

7) Our model gives accuracy of 81% with our selected cut-off of 0.5.

8) An optimal cut-off comes out to be 0.35 which gives the accuracy of 80% with sensitivity and specificity at 79%
and 80%.

9) Lead Origin Lead Add Form, Last Notable Activity Had a Phone Conversation, What is your current occupation, Working Professional are the features shows high promising leads.

10) Last Activity Converted to Lead, Last Activity_Olark Chat Conversation shows less lead conversion rate.