Lead Scoring Case Study

- By
Anant Patil and Sakshi Chopkar

Problem Statement

- To assist X Education in choosing the "hot leads"—the most promising leads—that are most likely to become paying clients.
- Design a logistic regression model to give each lead a lead score between 0 and 100, with higher lead scores indicating a better chance of conversion and lower lead scores indicating a lower likelihood of conversion.
- Determine the driver characteristics, which are reliable predictors of lead conversion, and comprehend their importance.
- If there are any outliers in the dataset, identify them and explain why.
- While developing the model, take both the technical and business considerations into account.
- Utilizing evaluation metrics like accuracy, sensitivity, specificity, and precision, to summarise the conversion projections.

Business Objectives

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

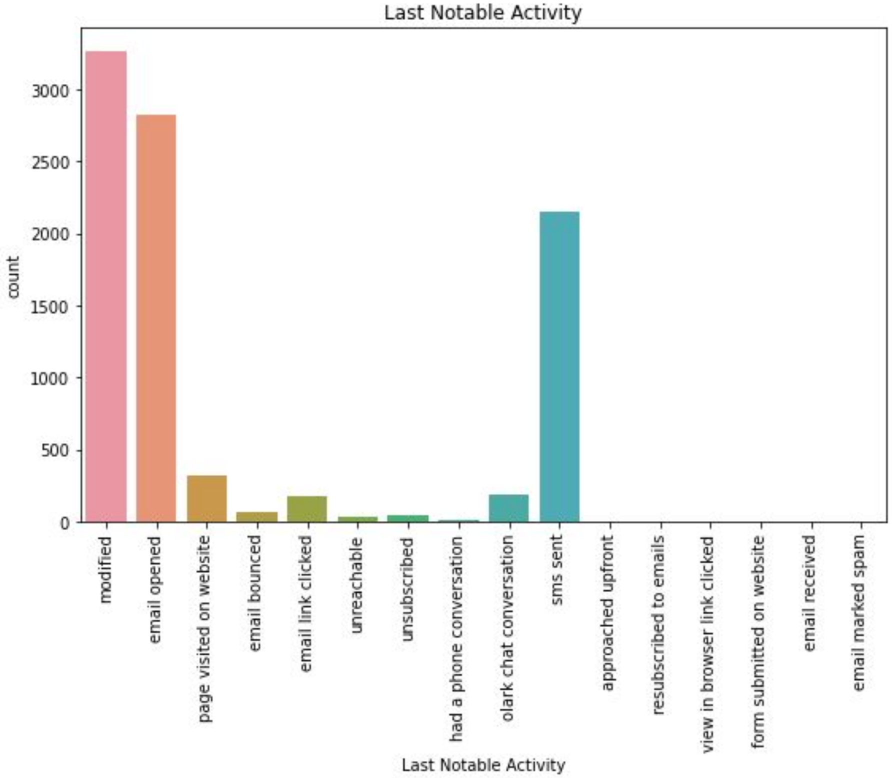
Approach

1. Data Cleaning
2. Exploratory Data Analysis
3. Assigning Dummy Variables to Categorical variables
4. Scaling
5. Train-Test Split
6. Model Building
7. Model Evaluation
8. Prediction

Data Manipulation

- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”
- Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- Dropping the columns having more than 35% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

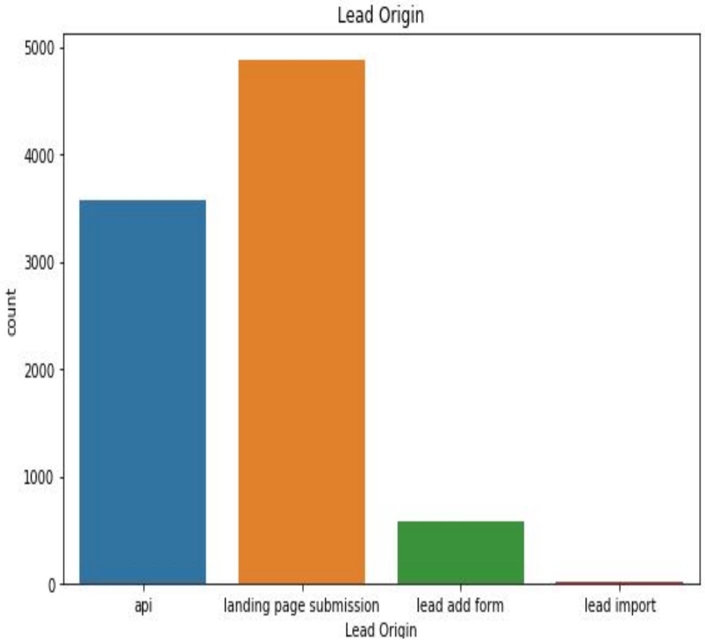
EDA

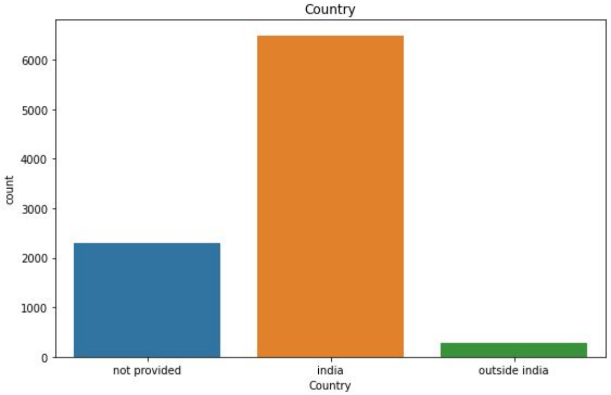

Last Notable Activity
count
0
500
1000
1500
2000
2500
3000
modified
email opened
page visited on website
email bounced
email link clicked
unreachable
unsubscribed
had a phone conversation
olark chat conversation
sms sent
approached upfront
resubscribed to emails
view in browser link clicked
form submitted on website
email received
email marked spam
Last Notable Activity

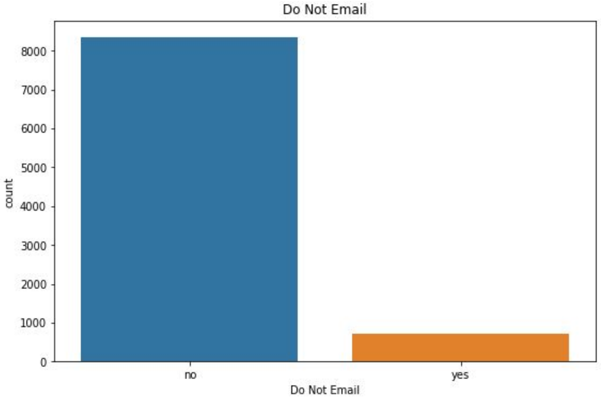
Univariate Analysis

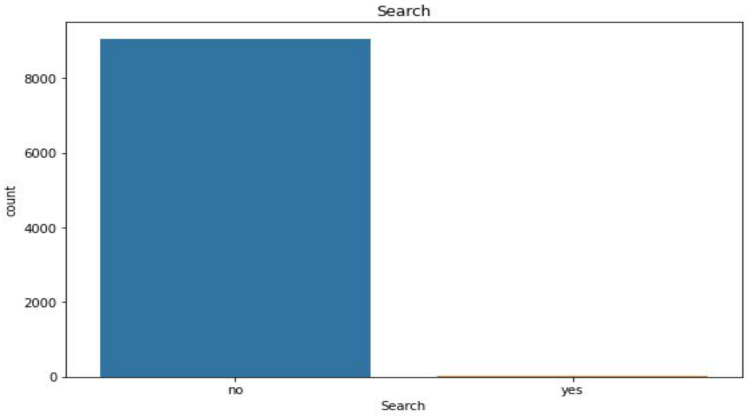
- Data distribution and outliers in the "Leads" data were discovered via univariate analysis.
- Outliers were found in the following key columns:
  - Total Visits
  - Page Views Per Visit
  - Asymmetrique Activity Score
  - Asymmetrique Profile Score
- Outliers in the data have been handled using the Inter Quantile Range (IQR) approach.
- Due to the high percentage (9%), the decision has been made to leave all outliers in place.
- To make sure this has no effect on the score, we will analyse the final model.

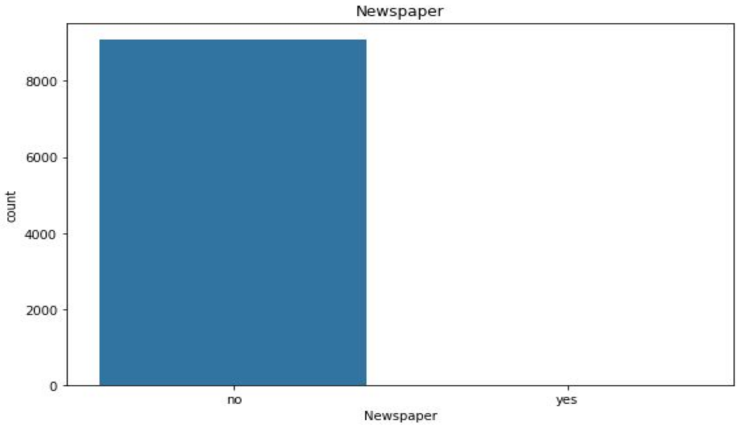
Categorical Variables

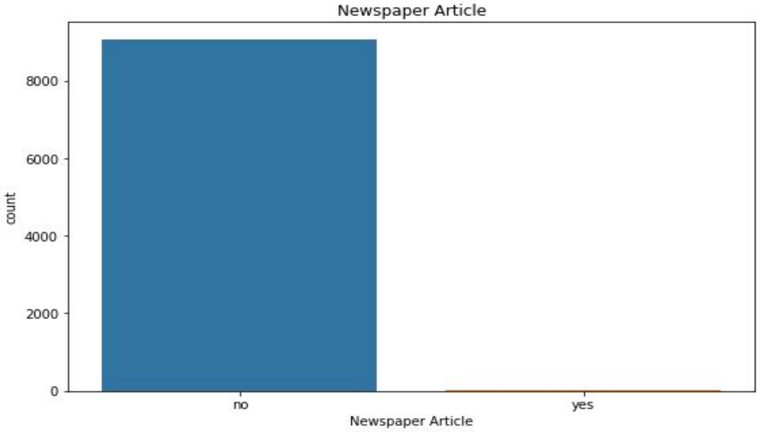


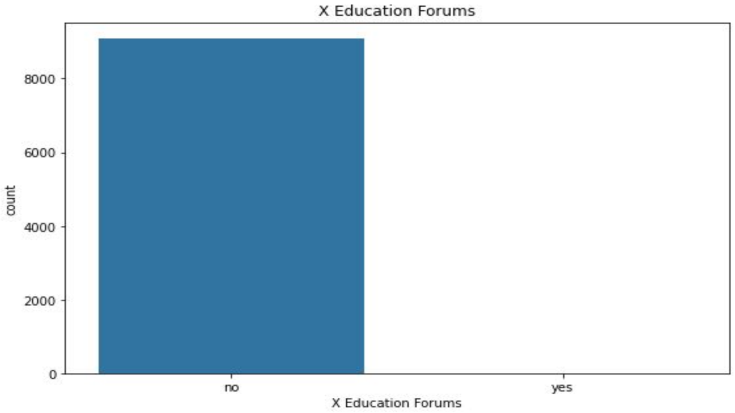


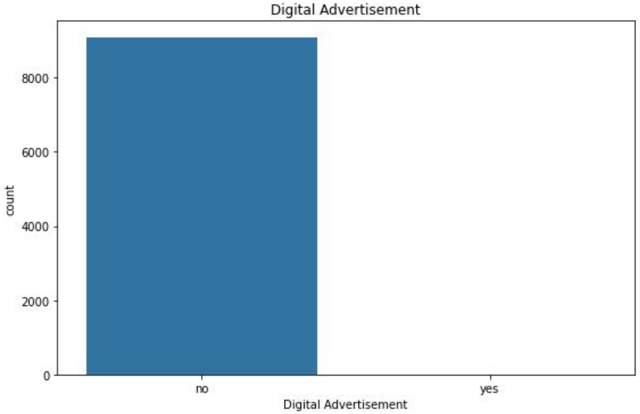


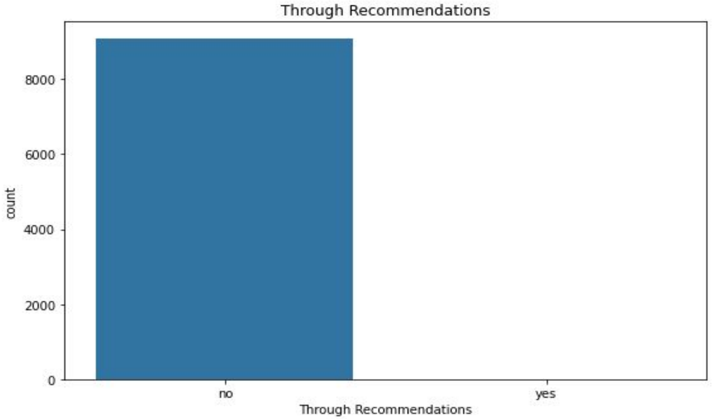

Search
count
8000
6000
4000
2000
0
no
yes
Search

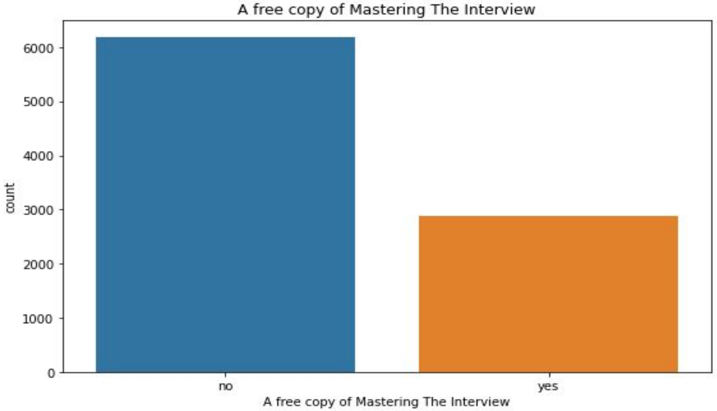

Newspaper
count
8000
6000
4000
2000
0
no
yes
Newspaper


Newspaper Article
count
8000
6000
4000
2000
0
no
yes
Newspaper Article


X Education Forums
count
8000
6000
4000
2000
0
no
yes
X Education Forums


Digital Advertisement
8000
6000
4000
2000
0
count
no
yes
Digital Advertisement


Through Recommendations
8000
6000
4000
2000
0
count
no
yes
Through Recommendations


A free copy of Mastering The Interview
6000
5000
4000
3000
2000
1000
0
count
no
yes
A free copy of Mastering The Interview

Bivariate Analysis

The target variable has been set to the "Converted" column. Therefore, bivariate analysis of significant variables with regard to the target variable has been carried out.

- Visitors who are interested in the upcoming batch and lateral students have a greater possibility of being converted.
- Lead quality with the "High in Relevance" tag has historically had high conversion rates.
- The likelihood of converting a lead generated by a "Lead Add Form" or "Quick Add Form" is high.
- More leads are converted via the Welingak website, WeLearn, live chat, and NC EDM than by any other source.

Checking Correlation

The columns listed below have a strong positive correlation with one another:

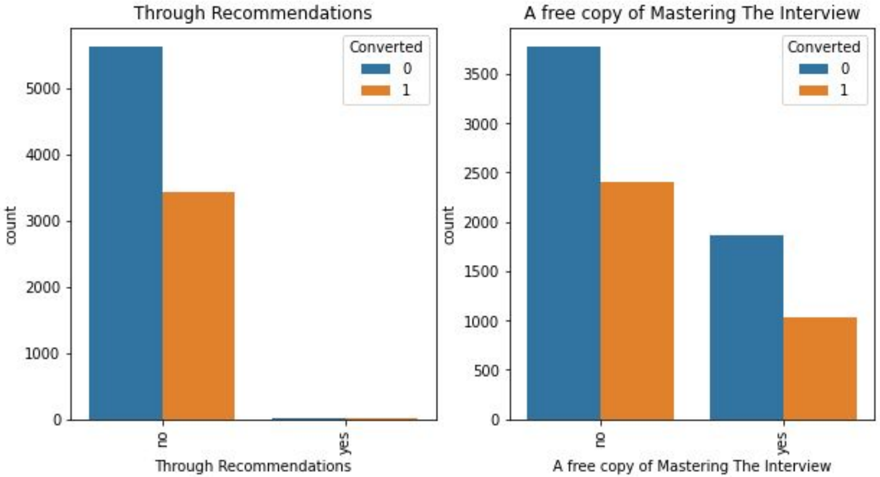
1. Digital Advertisement
2. Newspaper Article
3. Through Recommendations
4. X Education
5. Search

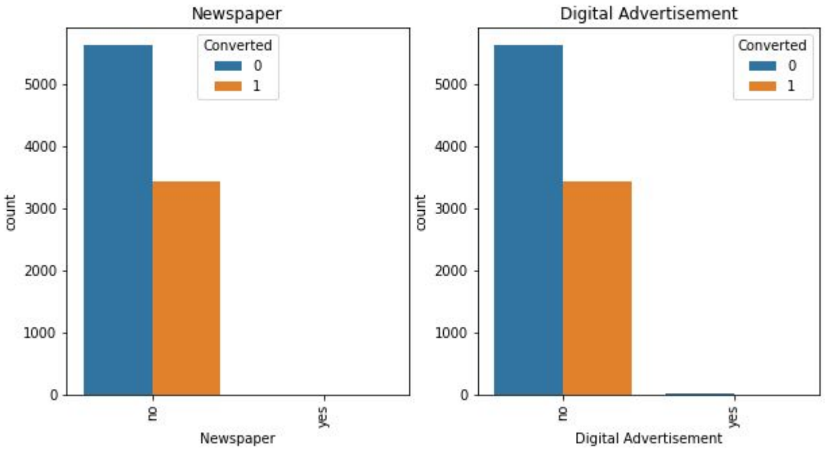
Additionally, a different set of columns have a strong positive correlation with one another.

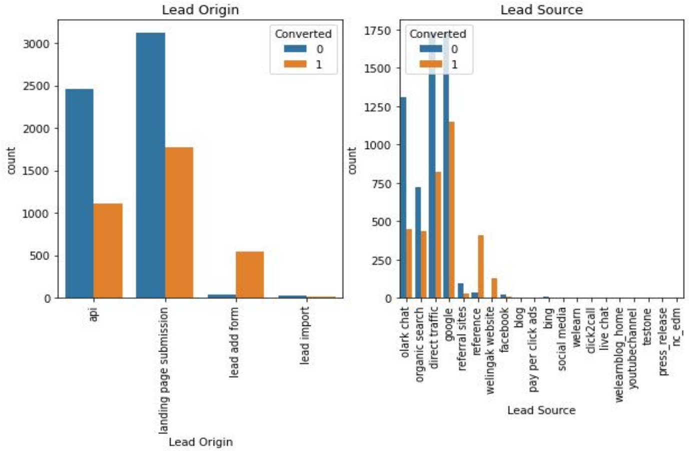
1. Total Visits
2. Total Time Spent on Website
3. Page reviews per Visit

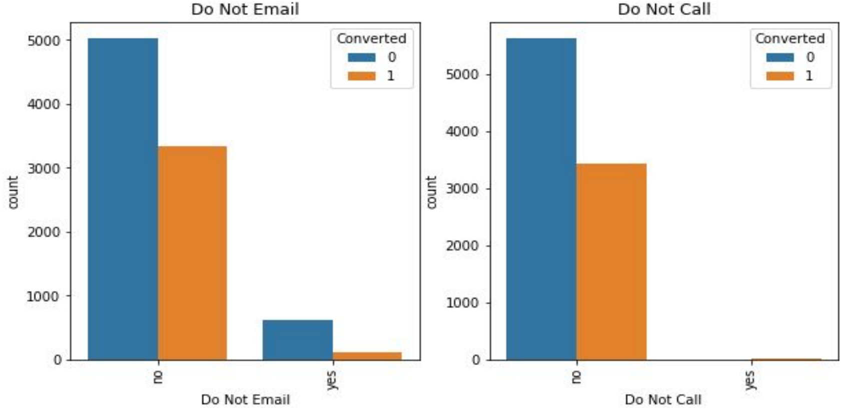
Asymmetrique Activity Index and Asymmetrique Profile Index have a significant positive association.

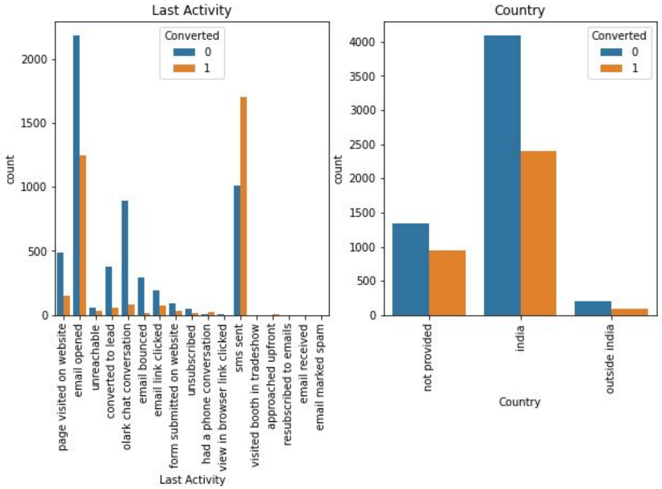
Relating all the categorical Variables to converted

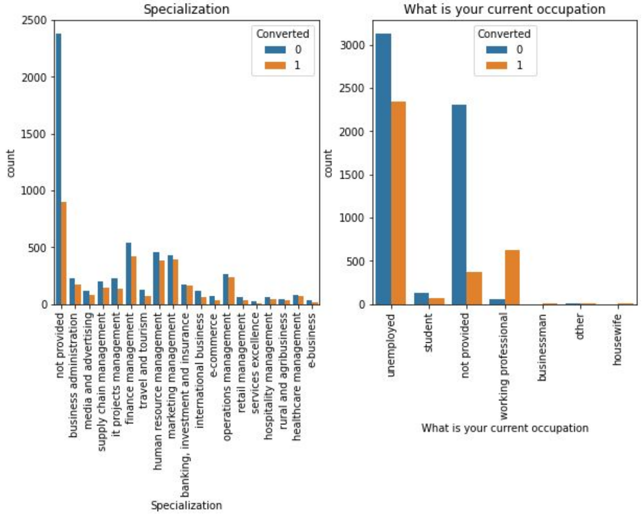


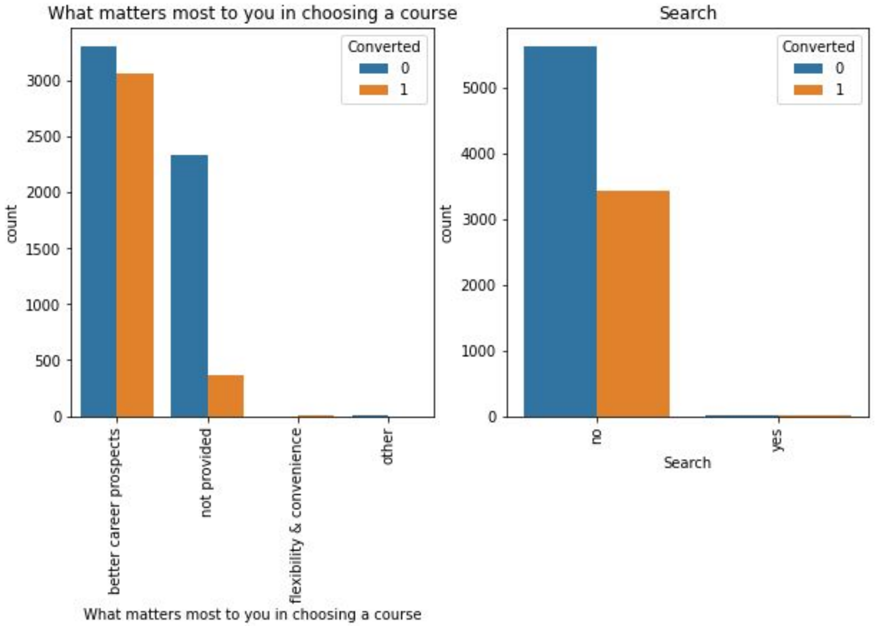



Lead Origin
Converted
0
1
count
0
500
1000
1500
2000
2500
3000
api
landing page submission
lead add form
lead import
Lead Origin
Lead Source
1750
1500
1250
1000
750
500
250
0
olark chat
organic search
direct traffic
google
referral sites
reference
welingak website
facebook
blog
pay per click ads
bing
social media
welearn
click2call
live chat
welearnblog_home
youtubechannel
testone
press_release
nc_edm
Lead Source

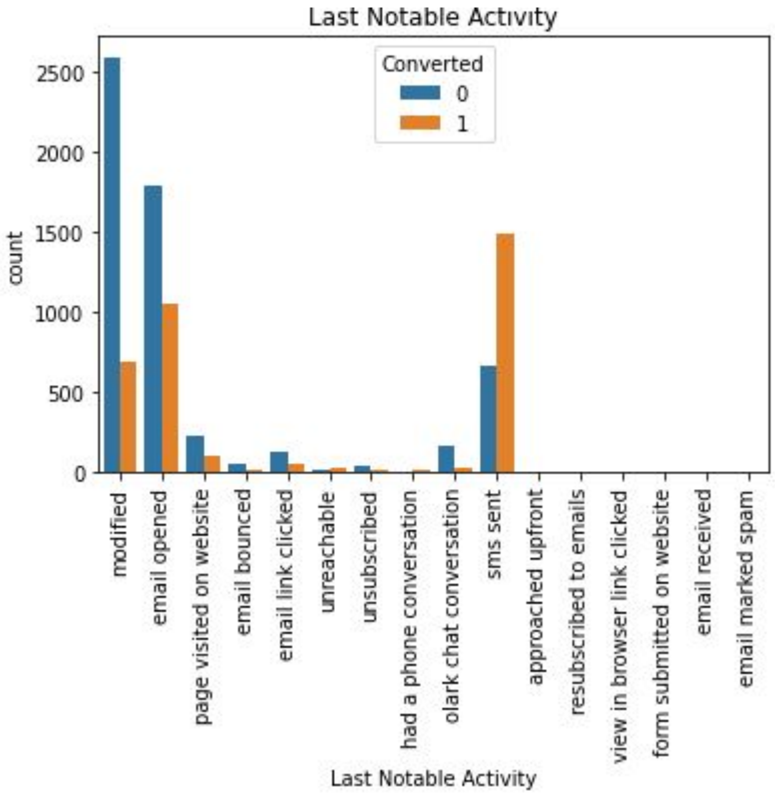

Do Not Email
Converted
0
1
count
0
1000
2000
3000
4000
5000
no
yes
Do Not Email
Do Not Call
Converted
0
1
count
0
1000
2000
3000
4000
5000
no
yes
Do Not Call


Last Activity
Converted
0
1
count
0
500
1000
1500
2000
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
sms sent
visited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam
Last Activity
Country
Converted
0
1
count
0
500
1000
1500
2000
2500
3000
3500
4000
not provided
india
outside india
Country


Specialization
Converted
0
1
count
0
500
1000
1500
2000
2500
not provided
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
What is your current occupation
Converted
0
1
count
0
500
1000
1500
2000
2500
3000
unemployed
student
not provided
working professional
businessman
other
housewife
What is your current occupation

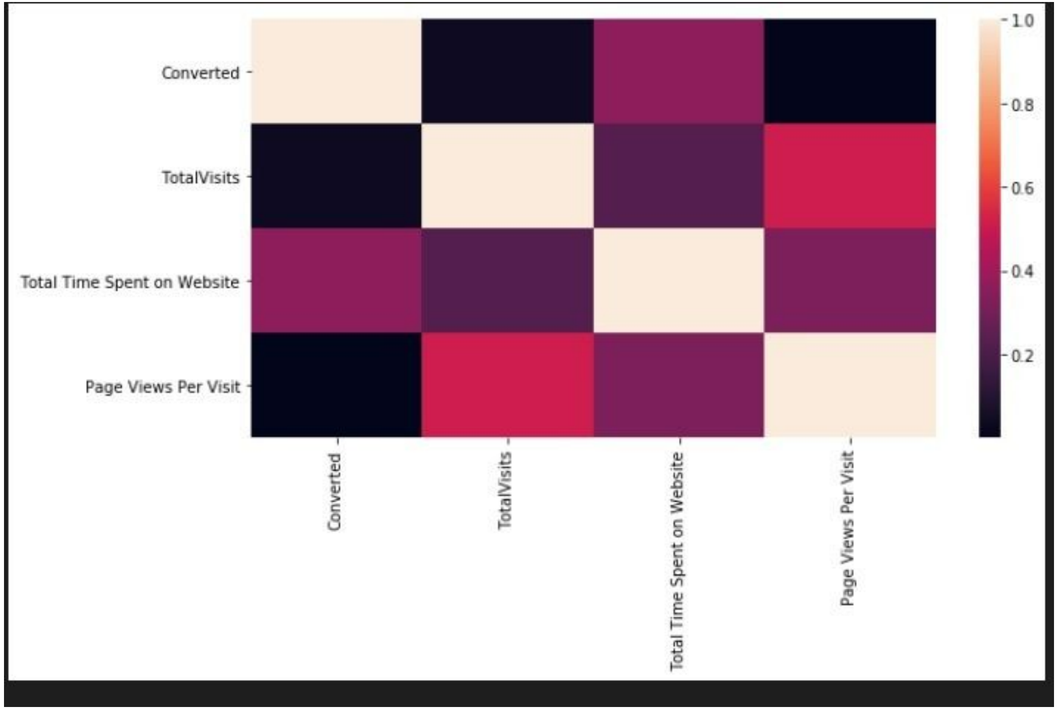


What matters most to you in choosing a course
Converted
0
1
count
0
500
1000
1500
2000
2500
3000
better career prospects
not provided
flexibility & convenience
other
What matters most to you in choosing a course
Search
Converted
0
1
count
0
1000
2000
3000
4000
5000
no
yes
Search



Last Notable Activity
Converted
0
1
count
0
500
1000
1500
2000
2500
modified
email opened
page visited on website
email bounced
email link clicked
unreachable
unsubscribed
had a phone conversation
olark chat conversation
sms sent
approached upfront
resubscribed to emails
view in browser link clicked
form submitted on website
email received
email marked spam
Last Notable Activity

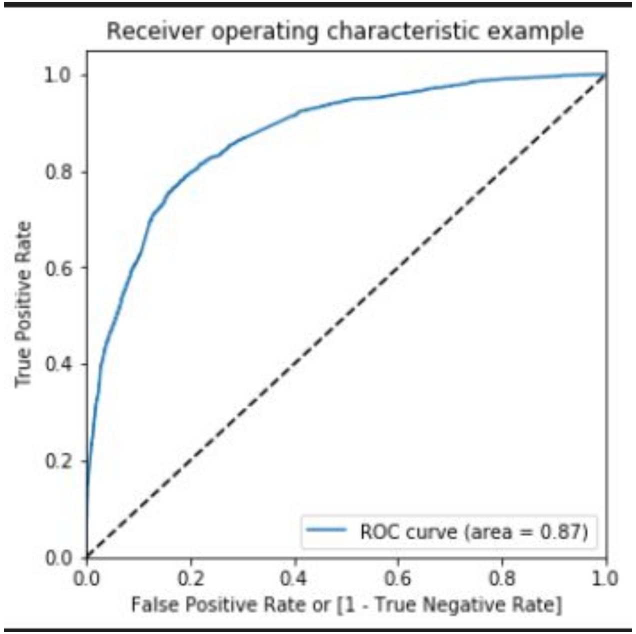
Correlation Between Variables



Model Building

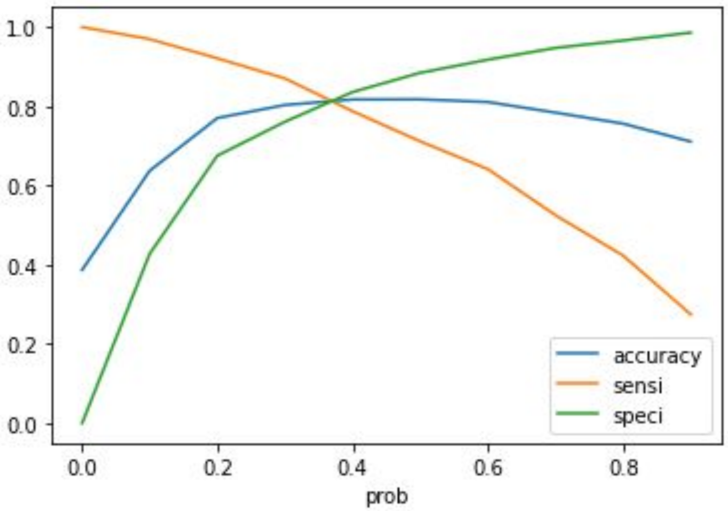
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5
- Predictions on test data set
- Overall accuracy 81%

ROC Curve



Area under curve = 0.87

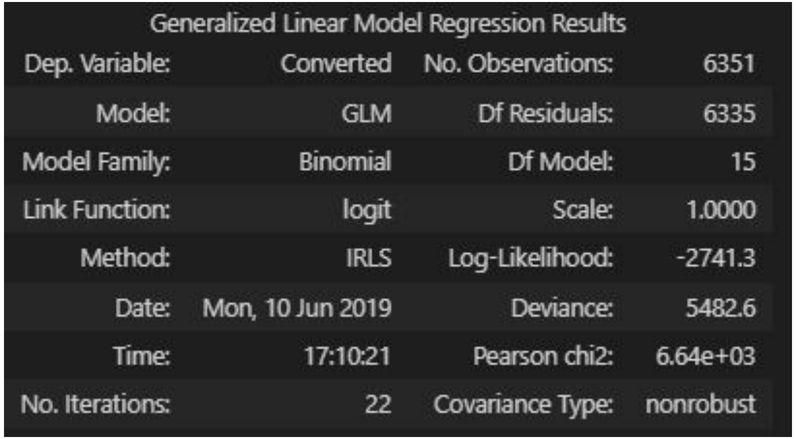
Optimal Threshold



Graph showing changes in Specificity, Accuracy and Sensitivity with changes in the probability threshold values.

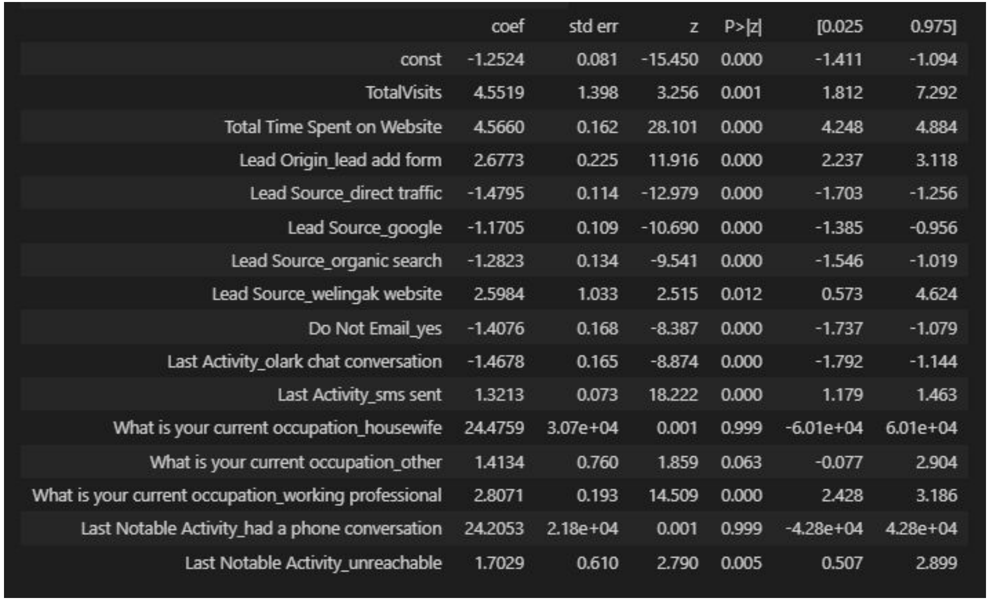
Optimal cutoff = 0.35

Model Summary: P-Values are zero



Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6335
Model Family:	Binomial	Df Model:	15
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2741.3
Date:	Mon, 10 Jun 2019	Deviance:	5482.6
Time:	17:10:21	Pearson chi2:	6.64e+03
No. Iterations:	22	Covariance Type:	nonrobust



	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2524	0.081	-15.450	0.000	-1.411	-1.094
TotalVisits	4.5519	1.398	3.256	0.001	1.812	7.292
Total Time Spent on Website	4.5660	0.162	28.101	0.000	4.248	4.884
Lead Origin_lead add form	2.6773	0.225	11.916	0.000	2.237	3.118
Lead Source_direct traffic	-1.4795	0.114	-12.979	0.000	-1.703	-1.256
Lead Source_google	-1.1705	0.109	-10.690	0.000	-1.385	-0.956
Lead Source_organic search	-1.2823	0.134	-9.541	0.000	-1.546	-1.019
Lead Source_welingak website	2.5984	1.033	2.515	0.012	0.573	4.624
Do Not Email_yes	-1.4076	0.168	-8.387	0.000	-1.737	-1.079
Last Activity_olark chat conversation	-1.4678	0.165	-8.874	0.000	-1.792	-1.144
Last Activity_sms sent	1.3213	0.073	18.222	0.000	1.179	1.463
What is your current occupation_housewife	24.4759	3.07e+04	0.001	0.999	-6.01e+04	6.01e+04
What is your current occupation_other	1.4134	0.760	1.859	0.063	-0.077	2.904
What is your current occupation_working professional	2.8071	0.193	14.509	0.000	2.428	3.186
Last Notable Activity_had a phone conversation	24.2053	2.18e+04	0.001	0.999	-4.28e+04	4.28e+04
Last Notable Activity_unreachable	1.7029	0.610	2.790	0.005	0.507	2.899

Evaluation Result

Comparing the values obtained for Train & Test:

Train Data:-

- Accuracy : 81.7%, Sensitivity : 71.1 %, Specificity : 88.4 %

Test Data:-

- Accuracy : 80.9 %, Sensitivity : 84.4 %, Specificity : 78.9 %

Thus, target lead conversion rate using this model is around 80%.

This Model seems to predict the Conversion Rate as desired and decision shall be made in making good calls to get a higher lead conversion rate of 80% from roughly 38% in the raw data

Inferences

- In decreasing order of impact, the following three factors are those that have the most influence on the likelihood of a lead conversion:
  - Tags_Lost to EINS
  - Tags_Closed by Horizon
  - Tags_Will revert after reading the email
- Each of these three increases the likelihood that a lead will be converted.
- The categorical variable Tags was used to produce these dummy features.
- These findings suggest that the business should pay closer attention to the leads with these three tags.

Recommendations

- Focus on the following by using the data visualisations:
    - Enhancing the categories' conversion rates to produce more leads and
    - Increasing lead generation for industries with high conversion rates.
- Pay close attention to the relative weighting of the model's attributes and how they affect the likelihood of conversion, either favourably or unfavourably.
- Change the probability threshold value for detecting possible leads in accordance with changing company demands.