Lead Scoring Case Study

- By Ramanathan Rahul & Manan Juneja

Problem Statement

- X Education sells online courses to industry professionals.
- They get a lot of leads but their lead conversion rate is very poor. For every 100 leads they get in a day, they are able to convert only 30 leads.
- To be more efficient, the company wants to identify the most potential leads or ‘Hot Leads’.
- **Business Objective:**
  - X Education wants to know the promising leads.
  - They want to build a model for identifying hot leads.
  - Deployment of the model for future use.

Approach

- Data cleaning and manipulation:
  - Drop columns that have no meaning or might affect analysis
  - Check and handle NA and missing values after imputing the ‘**Select**’ option as null
  - Imputation of values if necessary
- Data transformation and removing redundant columns
- EDA: Univariate and Bivariate Analysis and Handling Outliers
- Creating dummy variables
- Feature scaling
- Model building: Logistic Regression model for making predictions
- Model validation
- Model testing and presentation
- Conclusions and recommendations

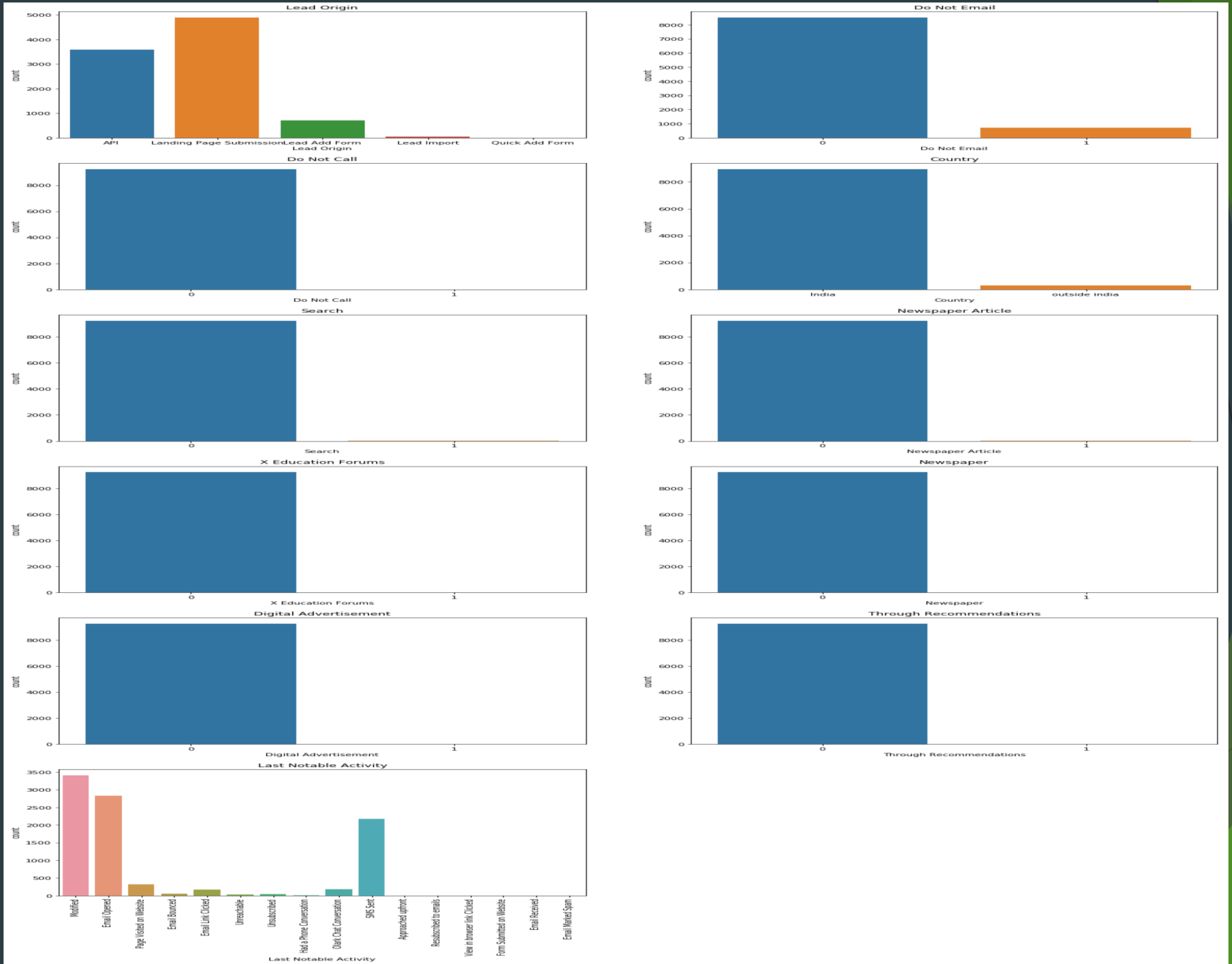
Data Manipulation

- There are totally 9240 rows and 37 columns in the dataset.
- Removed columns like Prospect ID, Lead Number, Asymmetrique Activity Index, Asymmetrique Profile Index, Asymmetrique Activity Score, Asymmetrique Profile Score, I agree to pay the amount through cheque, A free copy of Mastering The Interview, Tags, Lead Quality, How did you hear about X Education, City.
- Replaced fields having ‘Select’ as null values.
- Dropped columns having more than 30% missing values like ‘Specialization’ and ‘Lead Profile’.
- Imputing the remaining columns with the most common value in that column.

Exploratory Data Analysis

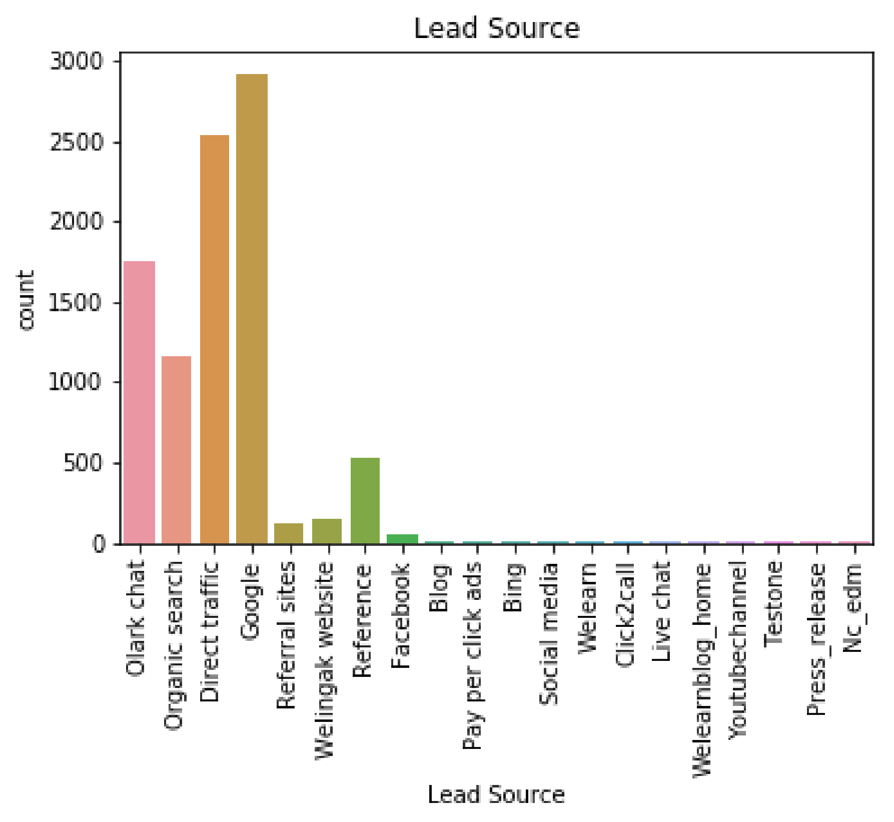
Visualizing the Distribution of Variables

- In summary, we can infer the following points from the plots (next slide):
  - Most of the leads origin is 'Landing page submission’
  - Most leads have given the permit to 'Email' them about the course
  - Very few number of people are from outside of India.
  - Very few number of people have known about this course through newspaper articles and search. Most of them may have found out about this course through other sources.
  - Most of the student's last notable activity is a modification, the next in the list being 'Email Opened' and 'SMS sent'.


Lead Origin
count
0
1000
2000
3000
4000
5000
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Origin
Do Not Email
count
0
1000
2000
3000
4000
5000
6000
7000
8000
0
1
Do Not Email
Do Not Call
count
0
2000
4000
6000
8000
0
1
Do Not Call
Country
count
0
2000
4000
6000
8000
India
outside india
Country
Search
count
0
2000
4000
6000
8000
0
1
Search
Newspaper Article
count
0
2000
4000
6000
8000
0
1
Newspaper Article
X Education Forums
count
0
2000
4000
6000
8000
0
1
X Education Forums
Newspaper
count
0
2000
4000
6000
8000
0
1
Newspaper
Digital Advertisement
count
0
2000
4000
6000
8000
0
1
Digital Advertisement
Through Recommendations
count
0
2000
4000
6000
8000
0
1
Through Recommendations
Last Notable Activity
count
0
500
1000
1500
2000
2500
3000
3500
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

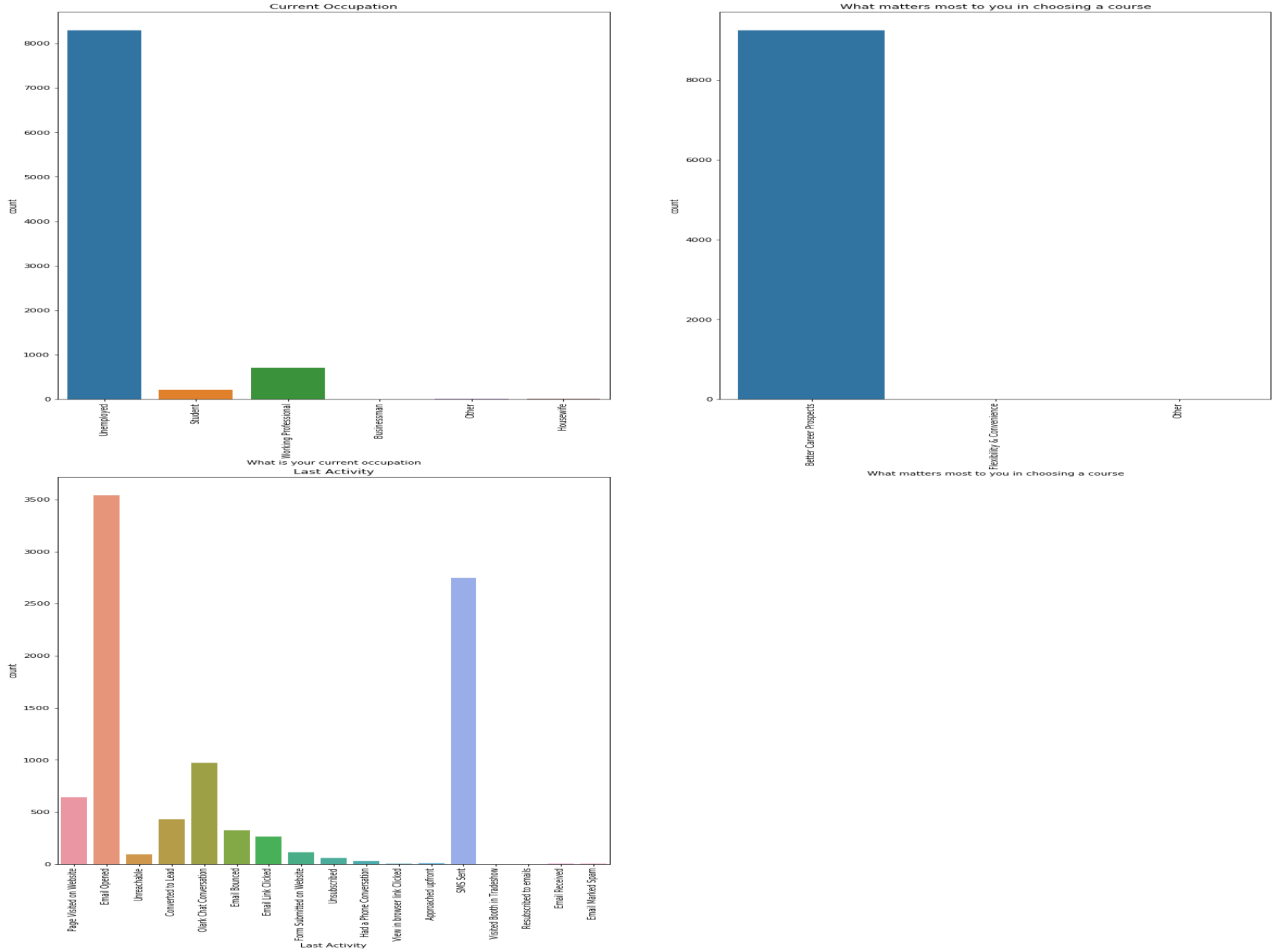
Visualizing the Distribution of Variables

- Inference from the plot for lead source:
  - We can see that most of the leads have been found through Google



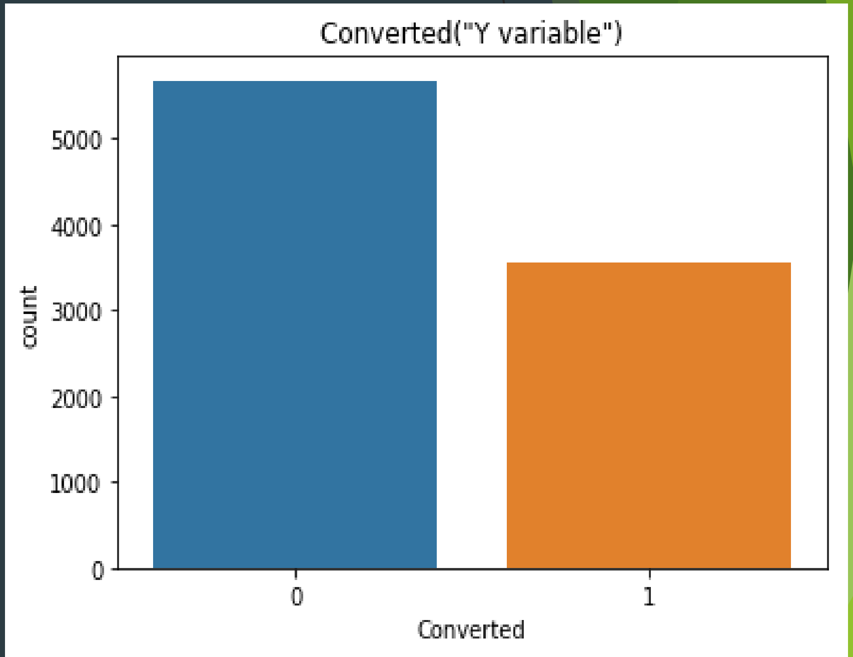
Visualizing the Distribution of Variables

- Inference from the plots (next slide):
  - We see that most of the people who are interested in this course are unemployed and those who are businessmen are not interested at all.
  - Maximum number of people are interested in "Better Career Prospects" which is understandable as most people are unemployed and there are a few students and working professionals as well.
  - The Last Activity of most of the applicants is "Email Opened" and next one is "SMS Sent".


Current Occupation
count
0
1000
2000
3000
4000
5000
6000
7000
8000
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation
What matters most to you in choosing a course
count
0
2000
4000
6000
8000
Better Career Prospects
Flexibility & Convenience
Other
What matters most to you in choosing a course
Last Activity
count
0
500
1000
1500
2000
2500
3000
3500
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

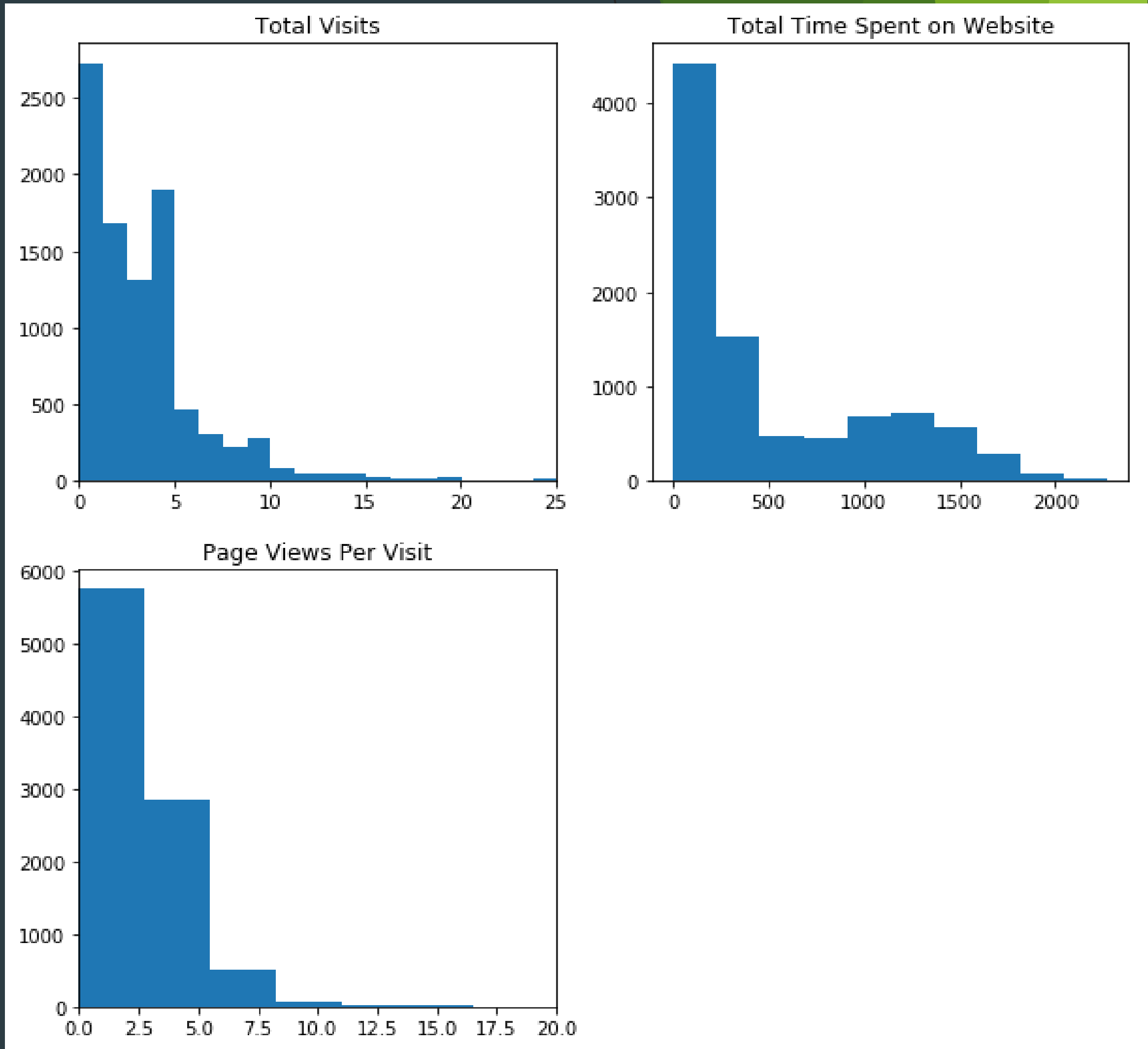
Visualizing the Distribution of Variables

- Inference from the plot for the target variable:
  - Most of the targeted leads haven't converted.



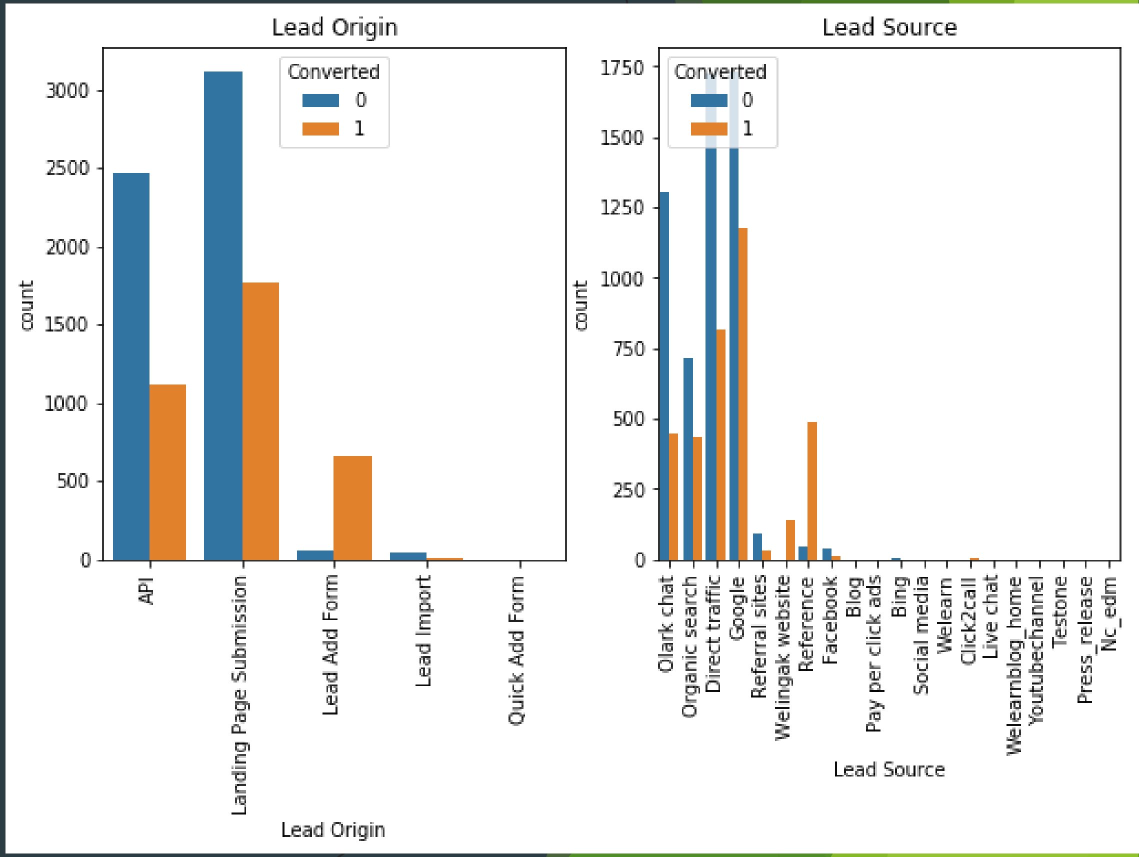
Visualizing the Distribution of Variables

- Inference from the plot:
  - Maximum number of people have visited 0 to 5 times and very few have visited the website more than 5 times.
  - The maximum time spent by the customers on the website is between 0 & 500, not many people have spent more time than that.
  - The highest number of people have average number of pages visited on the website between 0 & 2.5, there are a reasonable number between 2.5 & 5 too but very few people have and average above 5.



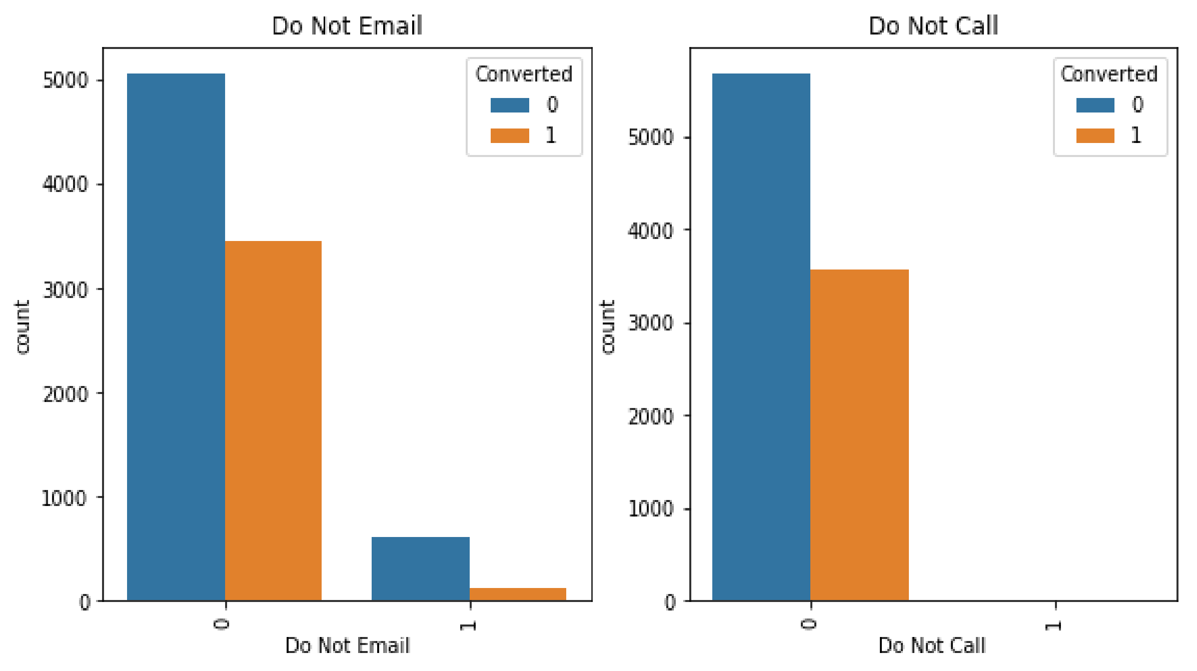
Relating the variables to the target (Converted)

- Inference from the plot:
  - The maximum number of non conversions are from "Landing Page Submission" but it has the maximum number of conversions too from the Lead Origins.
  - The "Lead Add Form" has more conversions than non conversions.
  - The Lead Source of "Direct Traffic" and "Google" have the maximum number of non conversions. But "Google" has the maximum number of converted too.



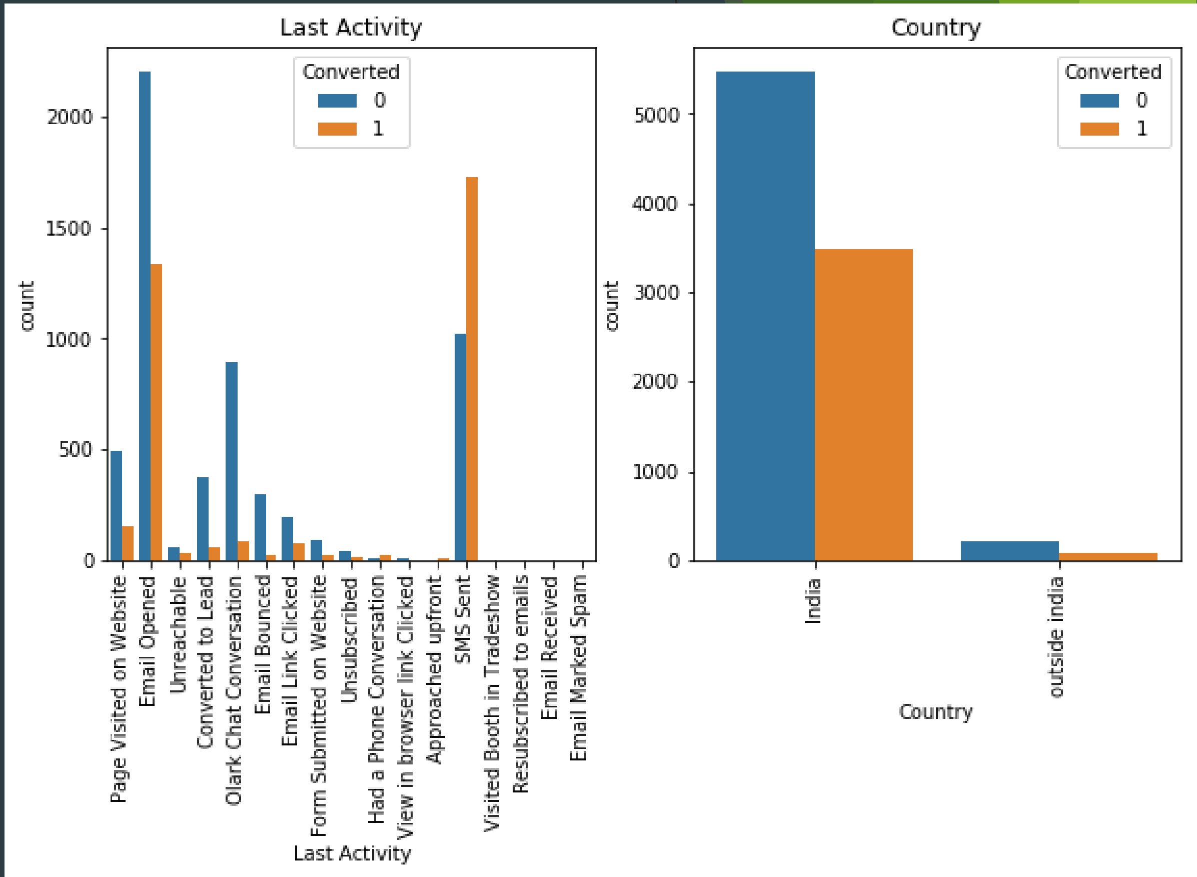
Relating the variables to the target (Converted)

- Inference from the plot:
  - Most of the people have told not to send email about the course and most of them have not converted. There are more numbers of converted also from this category compared to those who have told to mail them about the course.
  - None of the people have given permit to call them regarding the course. There are more people who haven't converted compared to those who have converted.



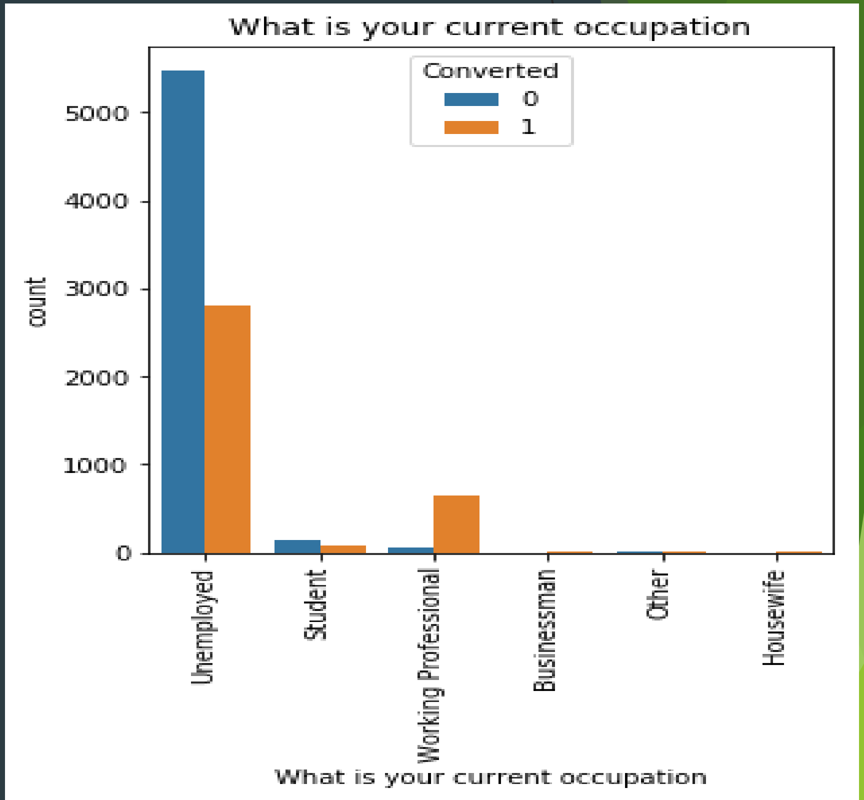
Relating the variables to the target (Converted)

- Inference from the plot:
  - The maximum number of non conversions are from those whose last activity is "Email Opened". This category has a good number of conversions too.
  - The maximum number of conversions are from those whose last activity is "SMS Sent".
  - There are many applicants from India compared to other countries and they have more number of non conversions too.



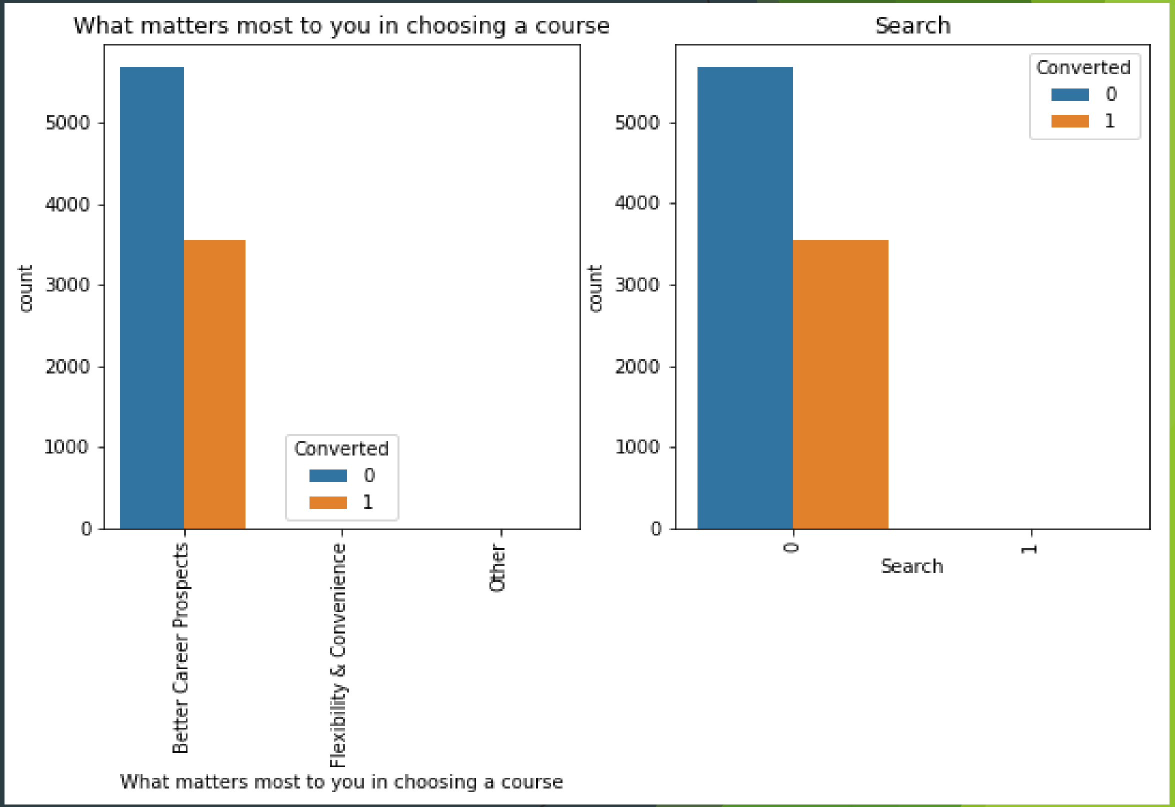
Relating the variables to the target (Converted)

- Inference from the plot:
  - Most of the leads are unemployed and a high number of them haven't converted.
  - There are few working professionals who are leads and most of them have converted.



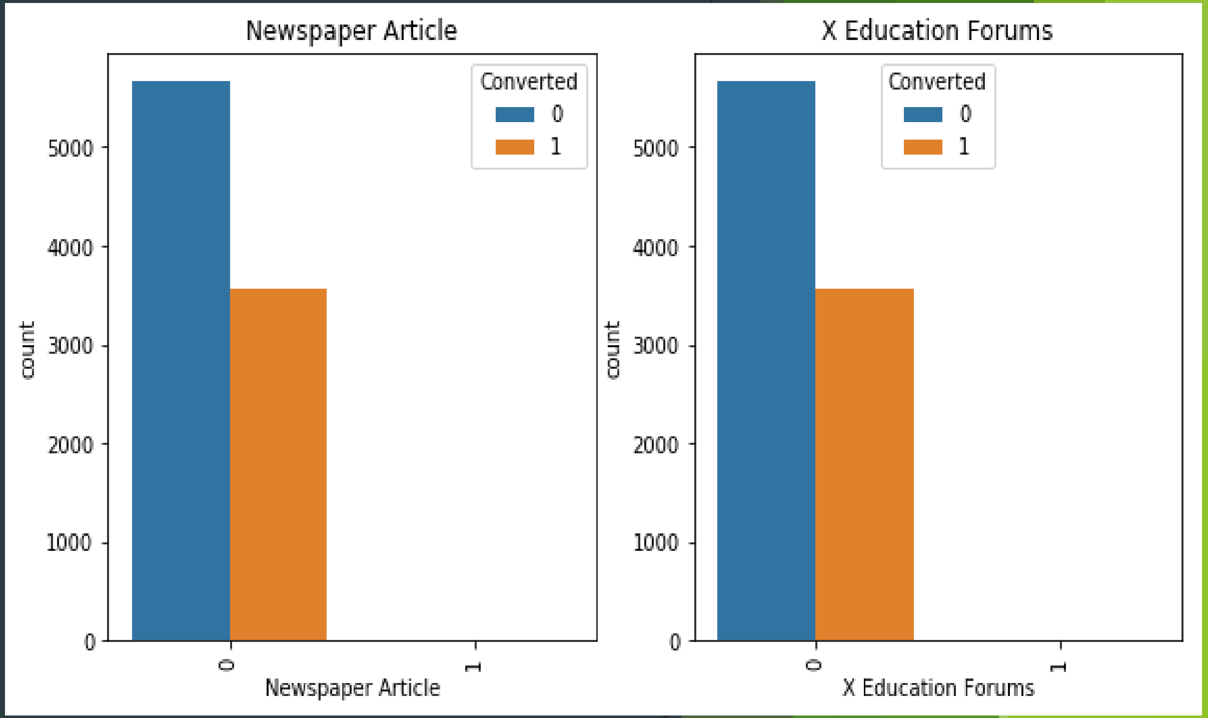
Relating the variables to the target (Converted)

- Inference from the plot:
  - All of the leads are looking for "Better Career Prospects" but most of them haven't converted.
  - None of the leads have come through a search and many of them haven't converted too.



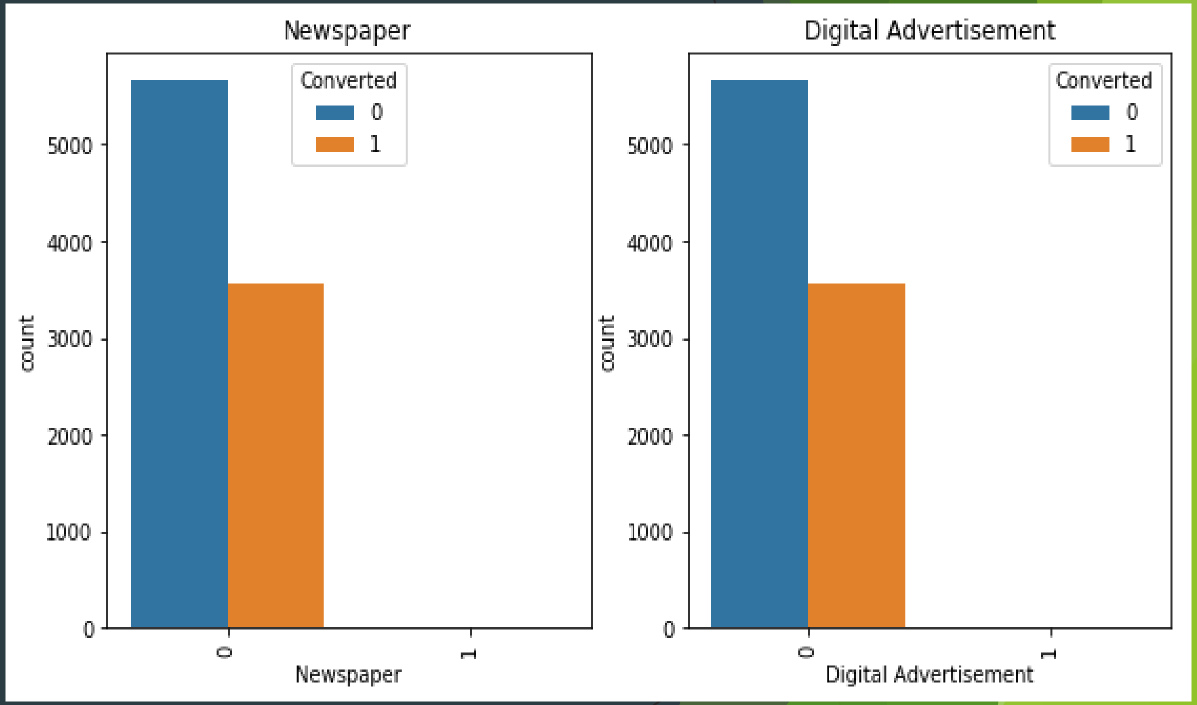
Relating the variables to the target (Converted)

- Inference from the plot:
  - None of the leads have come from newspaper articles and X Education forums.
  - Most of the leads haven't converted.



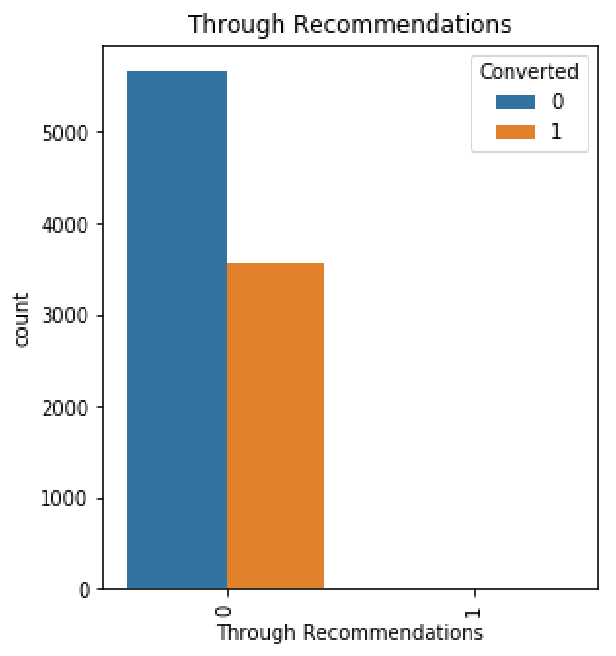
Relating the variables to the target (Converted)

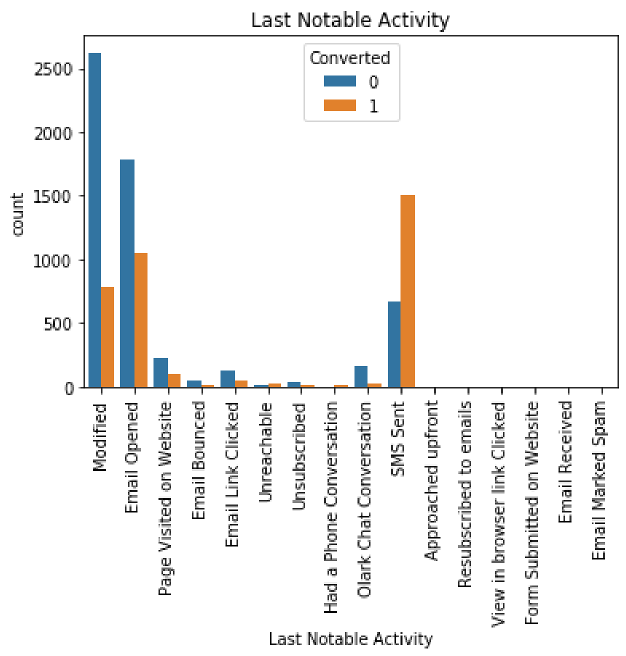
- Inference from the plot:
  - None of the leads have come from newspapers or digital advertisements.
  - Most of them haven't converted too.



Relating the variables to the target (Converted)

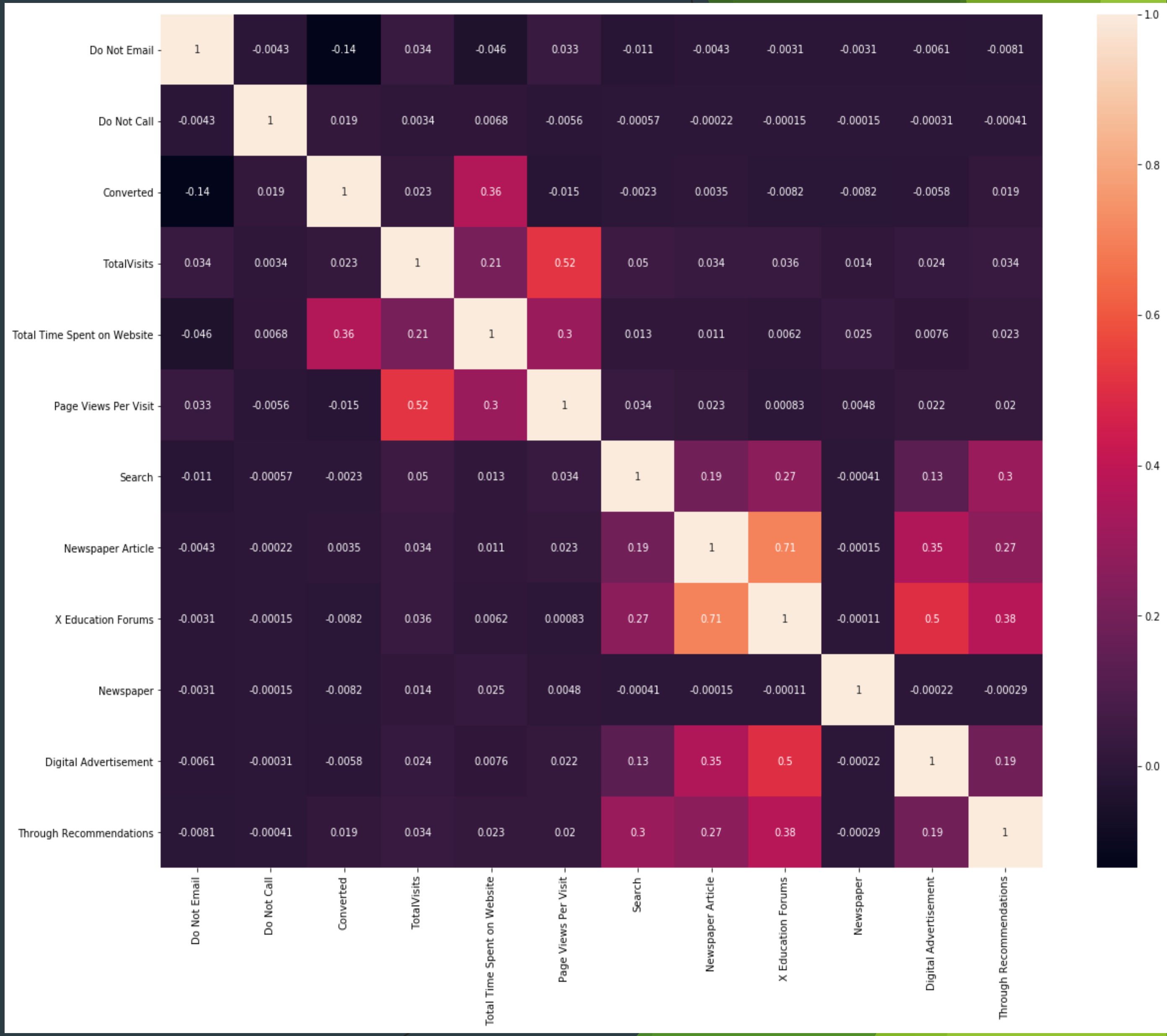
- Inference from the plot:
  - None of them have come through recommendations and most haven't converted.
  - The last notable activity of the leads is a modification and most of them haven't converted.
  - The number of leads with last notable activity as "SMS Sent" is is less but the number of converted is more.





Visualizing the Correlation of Variables

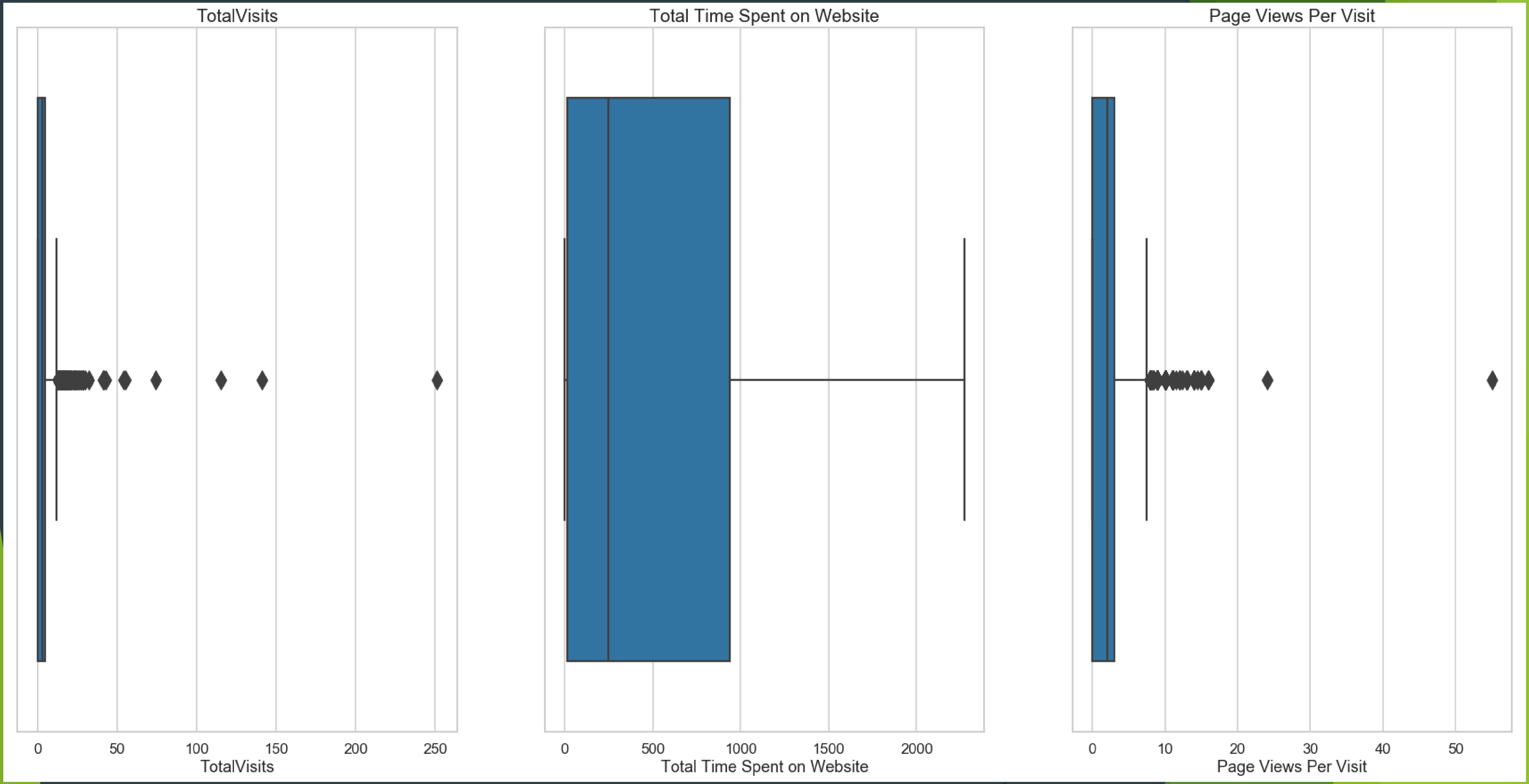
- Inferences from the plot:
  - We can see that there are no highly correlated variables in our dataset. All are very mildly correlated, whether it be a positive or negative correlation.
  - The highest correlation we can see is between 'Page views per visit' and 'TotalVisits' which also is not a very high value (0.52).



	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Search	Newspaper Article	X Education Forums	Newspaper	Digital Advertisement	Through Recommendations
Do Not Email	1	-0.0043	-0.14	0.034	-0.046	0.033	-0.011	-0.0043	-0.0031	-0.0031	-0.0061	-0.0081
Do Not Call	-0.0043	1	0.019	0.0034	0.0068	-0.0056	-0.00057	-0.00022	-0.00015	-0.00015	-0.00031	-0.00041
Converted	-0.14	0.019	1	0.023	0.36	-0.015	-0.0023	0.0035	-0.0082	-0.0082	-0.0058	0.019
TotalVisits	0.034	0.0034	0.023	1	0.21	0.52	0.05	0.034	0.036	0.014	0.024	0.034
Total Time Spent on Website	-0.046	0.0068	0.36	0.21	1	0.3	0.013	0.011	0.0062	0.025	0.0076	0.023
Page Views Per Visit	0.033	-0.0056	-0.015	0.52	0.3	1	0.034	0.023	0.00083	0.0048	0.022	0.02
Search	-0.011	-0.00057	-0.0023	0.05	0.013	0.034	1	0.19	0.27	-0.00041	0.13	0.3
Newspaper Article	-0.0043	-0.00022	0.0035	0.034	0.011	0.023	0.19	1	0.71	-0.00015	0.35	0.27
X Education Forums	-0.0031	-0.00015	-0.0082	0.036	0.0062	0.00083	0.27	0.71	1	-0.00011	0.5	0.38
Newspaper	-0.0031	-0.00015	-0.0082	0.014	0.025	0.0048	-0.00041	-0.00015	-0.00011	1	-0.00022	-0.00029
Digital Advertisement	-0.0061	-0.00031	-0.0058	0.024	0.0076	0.022	0.13	0.35	0.5	-0.00022	1	0.19
Through Recommendations	-0.0081	-0.00041	0.019	0.034	0.023	0.02	0.3	0.27	0.38	-0.00029	0.19	1

Outlier Analysis

- Inference from the plot:
  - From the above boxplots we can now confirm that we have two outlier variables in our dataset ('TotalVisits' and 'Page Views Per Visit'). Now as per business requirement we cannot drop these outliers because it may impact our analysis/model so we will create bins for these two outliers.



TotalVisits

Total Time Spent on Website

Page Views Per Visit

Data Conversion

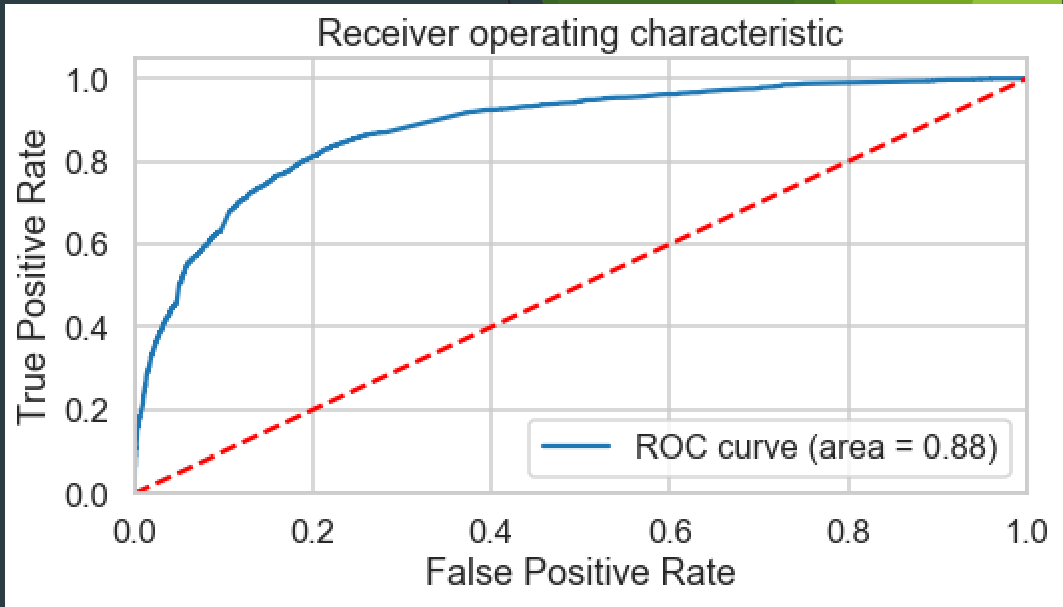
- Created dummy variables for ‘object’ data type variables.
- Created bins to handle outliers.
- Numerical variables are normalized.
- Total rows: 9240
- Total columns: 86

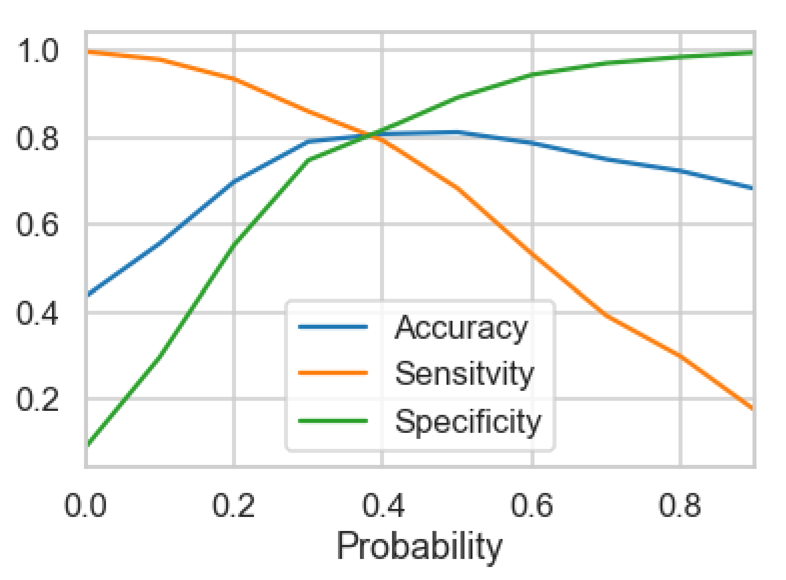
Model Building

- Split the data into train and test sets in a 70:30 ratio.
- Used RFE for feature selection, ran with 15 variables as output.
- Built model by removing variables with p-value more than 0.05 and VIF more than 5.
- Predictions done on train and test set data.
- Train set:
  - Chose a cutoff probability value of 0.4.
  - Precision and recall are 72.70% and 79.28% respectively.
- Test set:
  - Overall accuracy = 81.96%.
  - Precision and recall are 75.81% and 79.81% respectively.

ROC Curve

- Inference from the plot:
  - The curve is closer to the left side of the border and hence we can say that our model is having a very good accuracy.
  - The area under the curve is 88% of the total area.

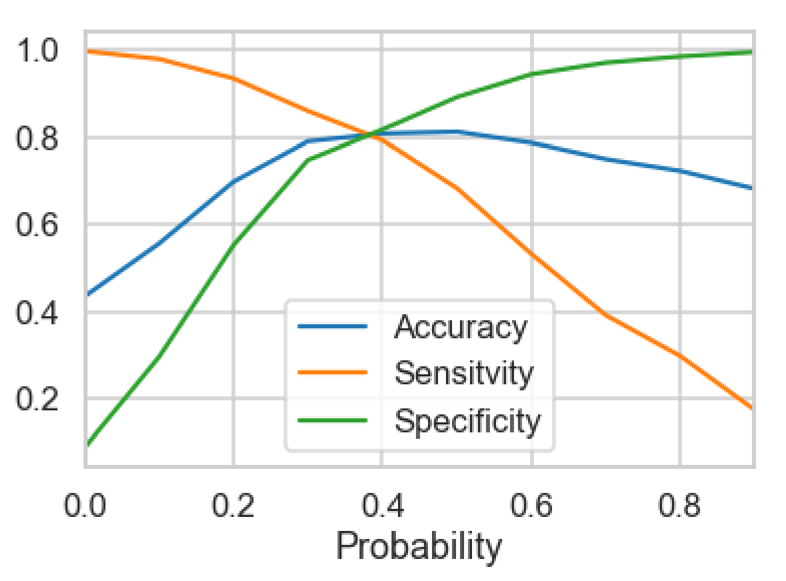




Optimal Probability Cut-off

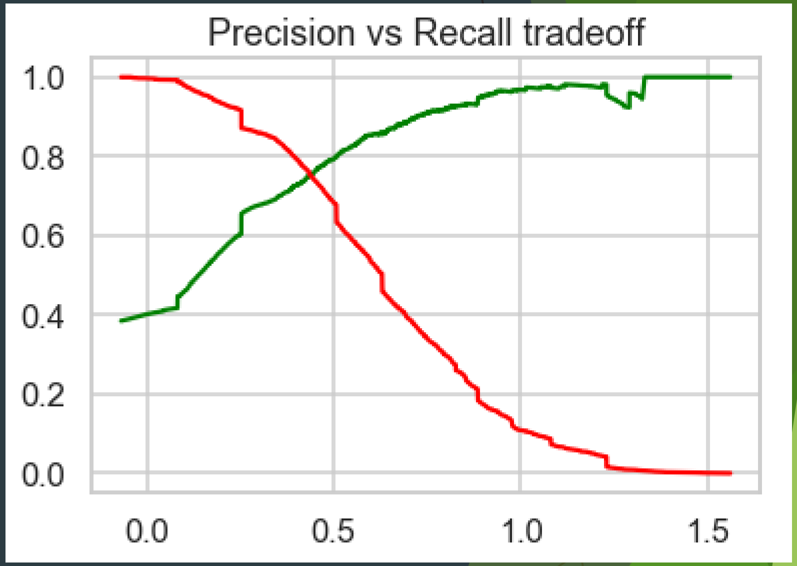
- Inference from the plot:

  - From the above curve, 0.4 is the optimum point for taking probability cutoff. The meeting point is slightly before from 0.4 and hence we can choose this as our final cutoff. Also, we can see that there is a trade off between sensitivity and specificity.



Precision vs Recall Tradeoff

- Inference from the plot:
  - We can see that there is a trade off between Precision and Recall and the meeting point is nearly at 0.5.



Conclusion

Inferences/Insights from the model:

1. The Accuracy, Precision and Recall score we got from test set are in an aceptable range.
2. We have high recall score than precision score which is what we were looking for.
3. The model is in a stable state which means that this model has an ability to adjust with the company’s requirements in the coming future.
4. Important features responsible for good conversion rate or the ones which contributes more towards the probability of a lead getting converted are:
   a) **Lead Origin_Lead Add Form**
   b) **What is your current occupation_Working Professional**
   c) **Last Notable Activity_Had a Phone Conversation**

Thank you