Lead Scoring Case Study

Problem Statement:

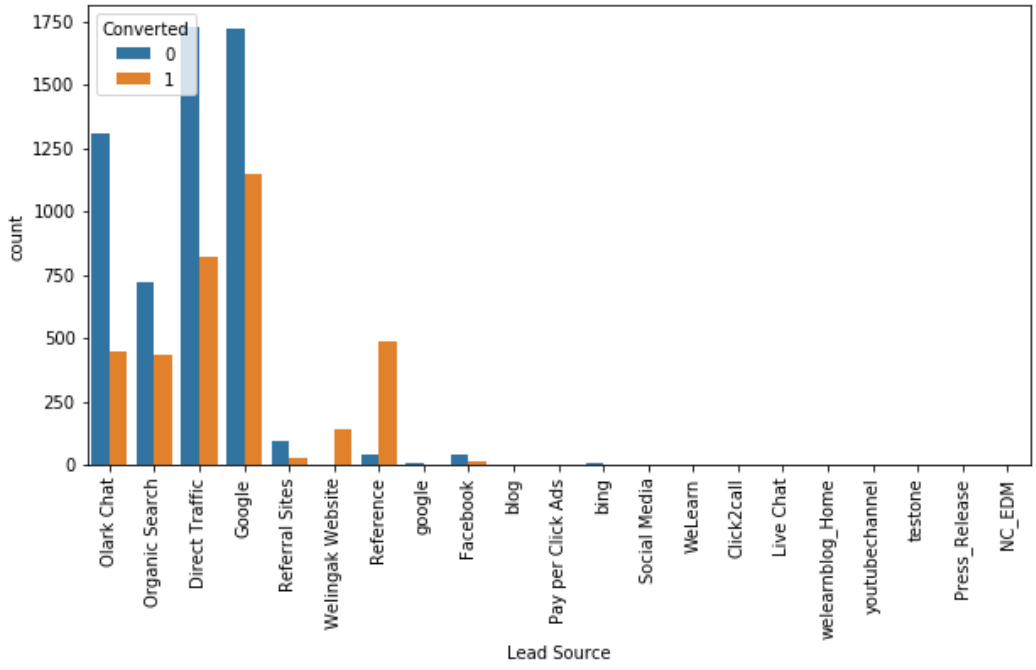
An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%. Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Solution Summary Report:

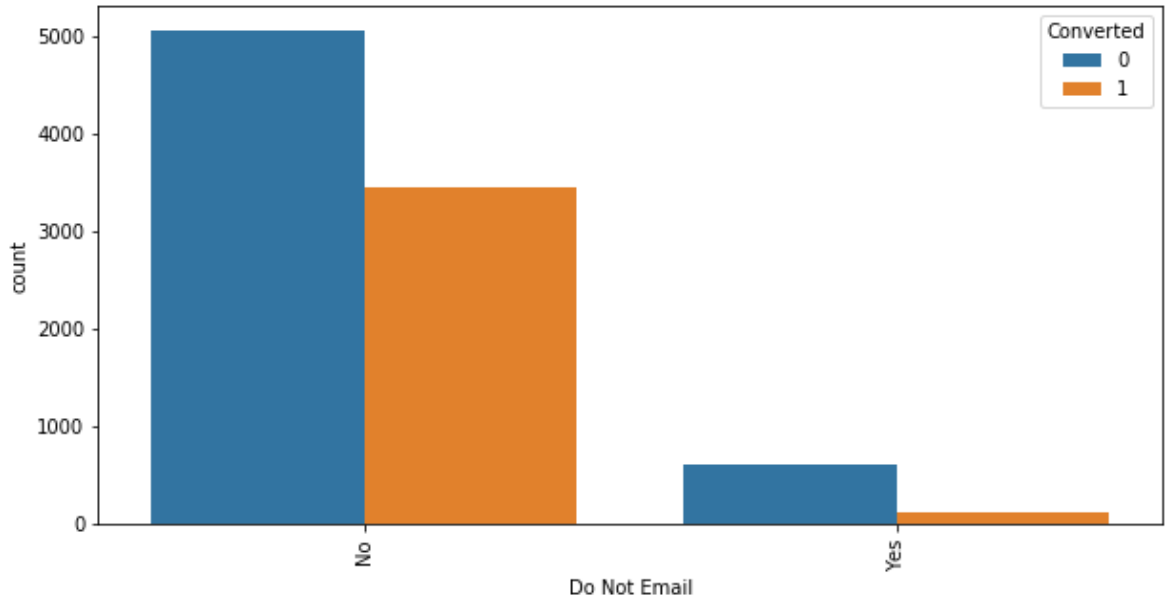
From the given data there are different columns like Lead Origin, Lead Source, Converted Leads, Total visits, time spent on website, Country, Specialization etc…

Now, performed Univariate and Bi-variate analysis w.r.t Target variable(converted). There are different insights can be observed from them like below.

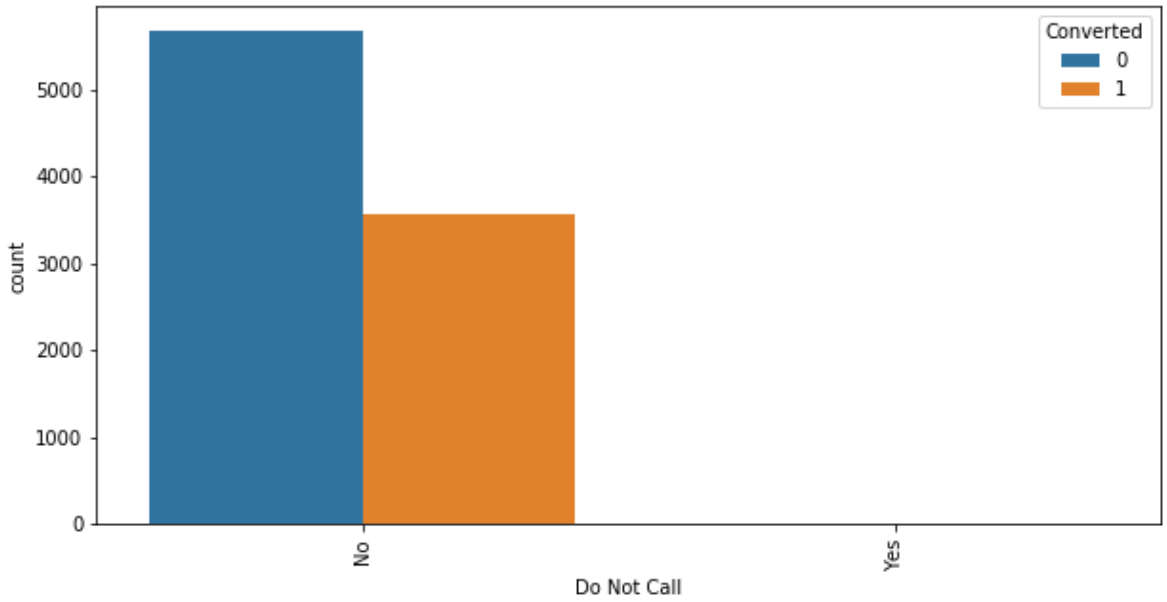
- Most of the leads are came through Google Search and Direct Traffic only



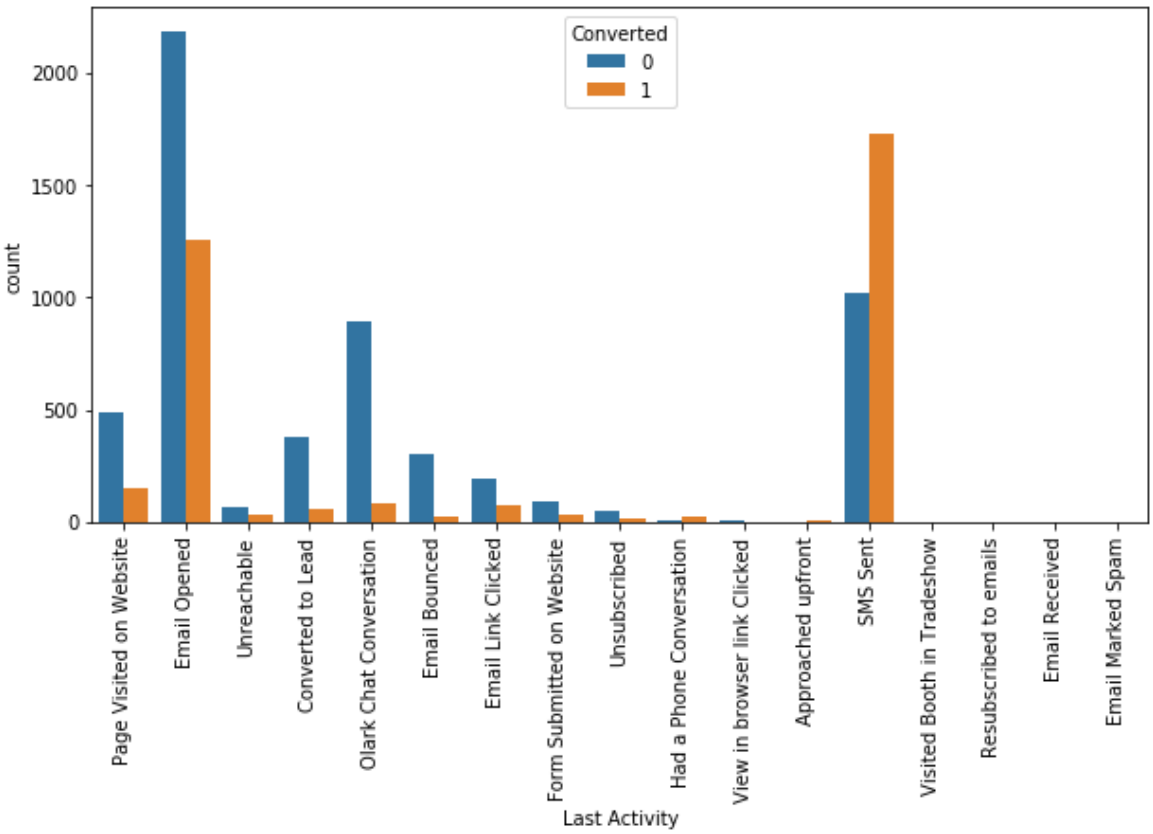
- People who opted “Do not Email” to No, they are mostly converted leads.



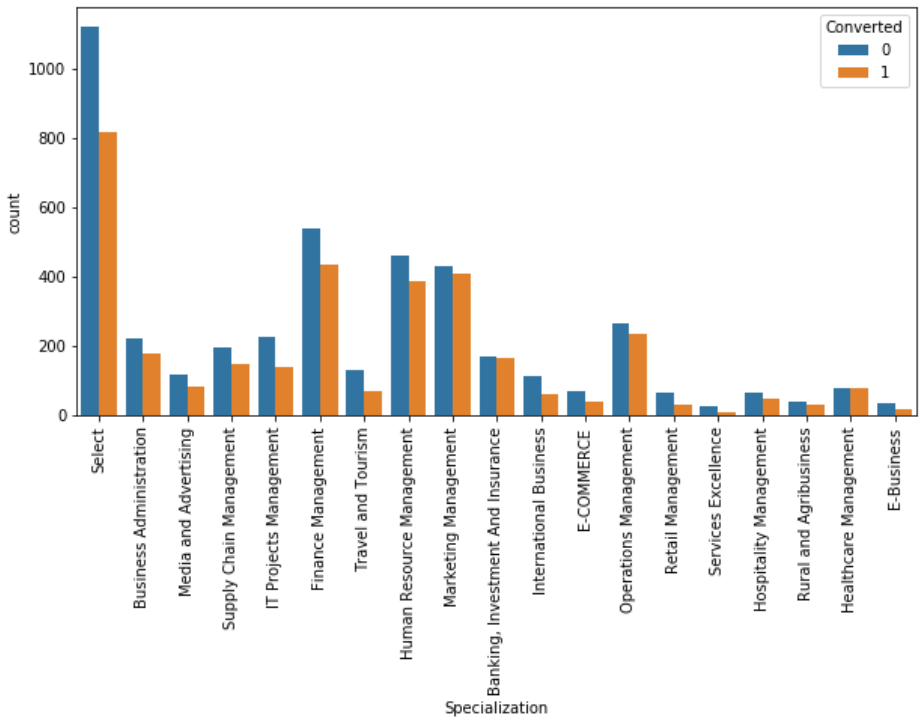
- Leads who don’t opted for Do not call option has high conversion rate.



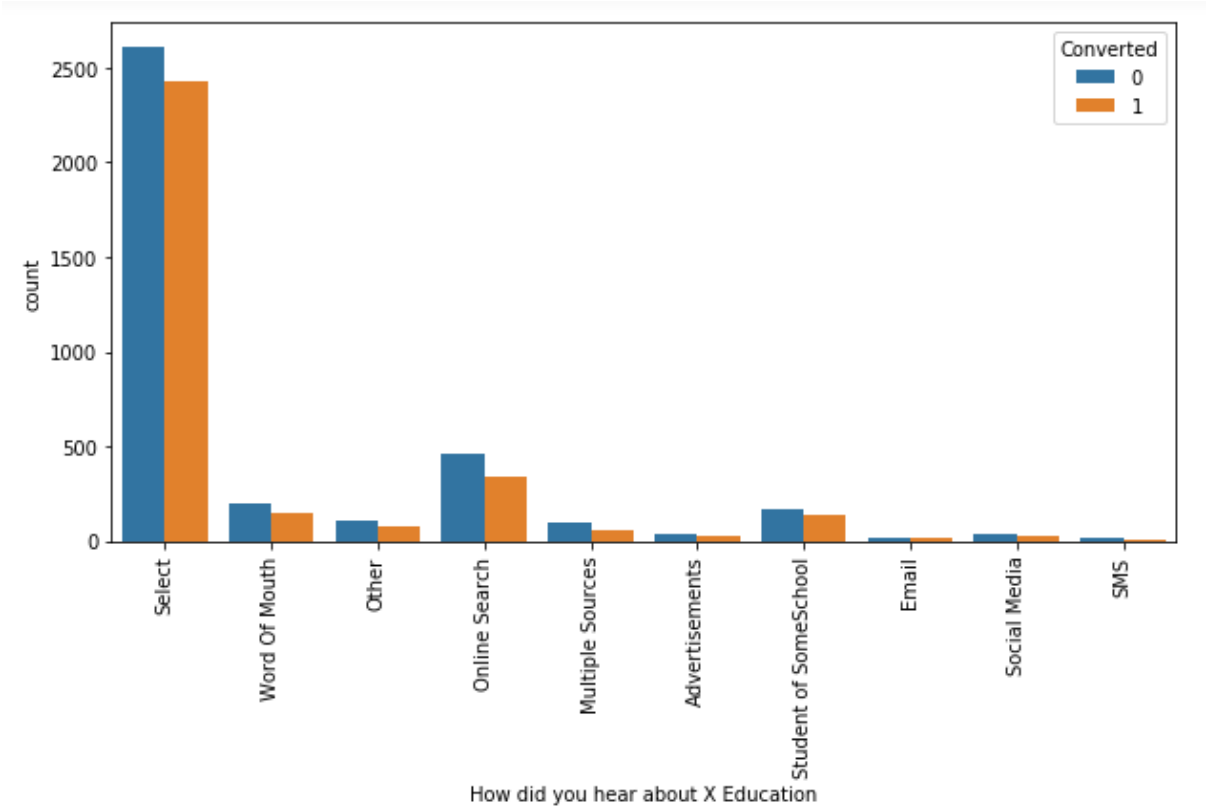
- Users who opened the email and who sent SMS has the high probability of the leads conversion rate.



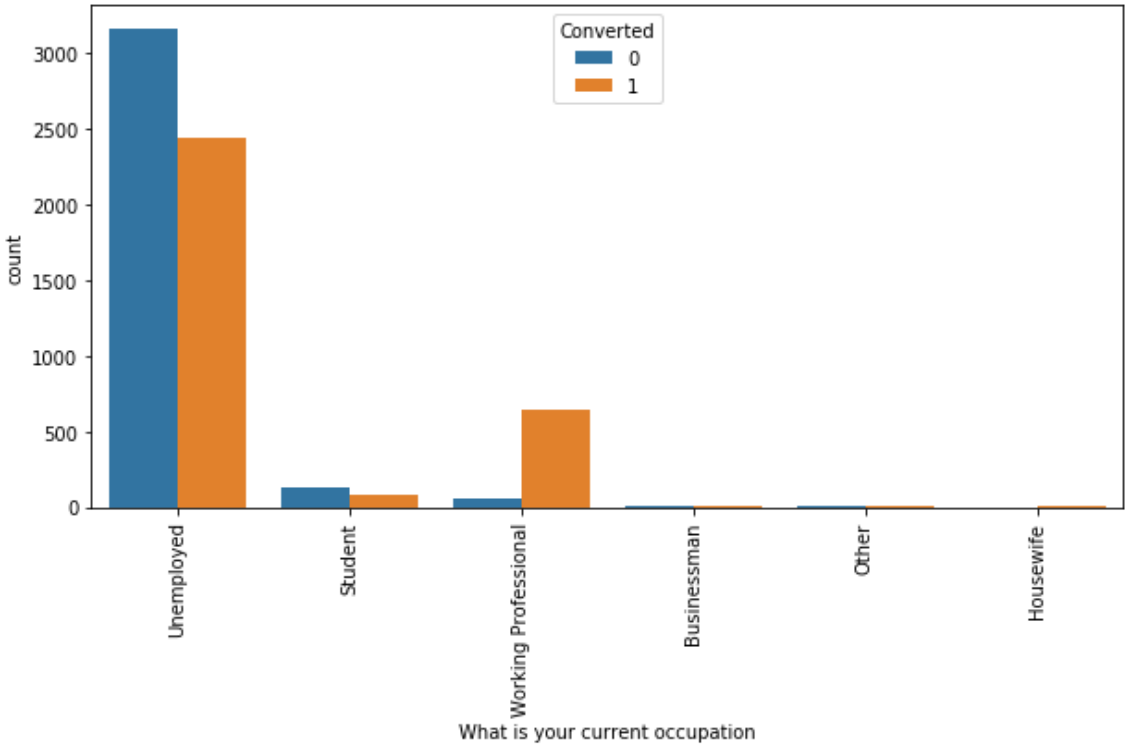
- There are multiple Specializations, in which many users are not selected the specialization option and left that select column. So many entries are coming under select category. Which needs to be treated as null value only.



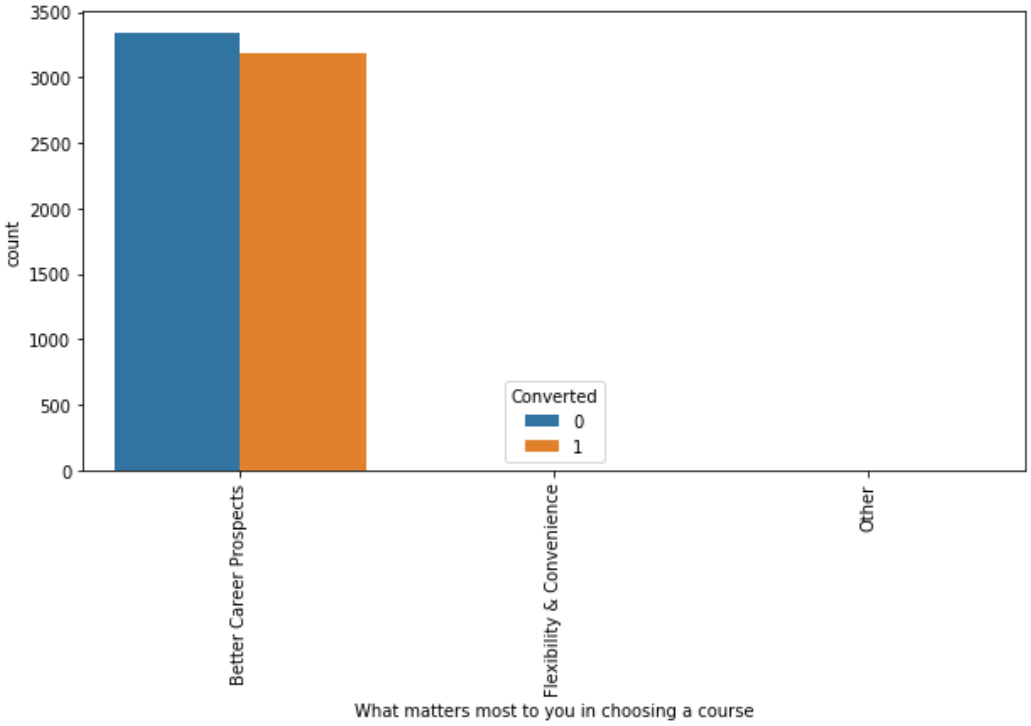
- Many users didn't mention that how did they hear about the X Education due to which many records are falling under default category **select**.



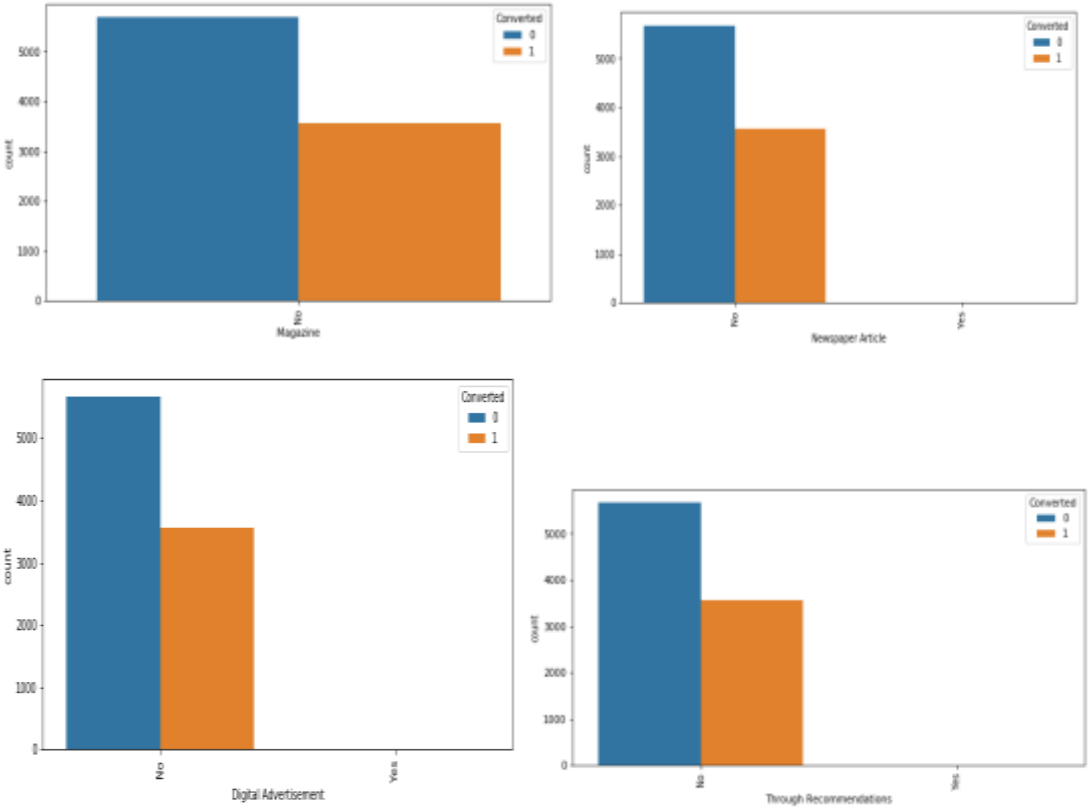
- Compare to all, users who are "Un employed" are most likely to be potential leads.



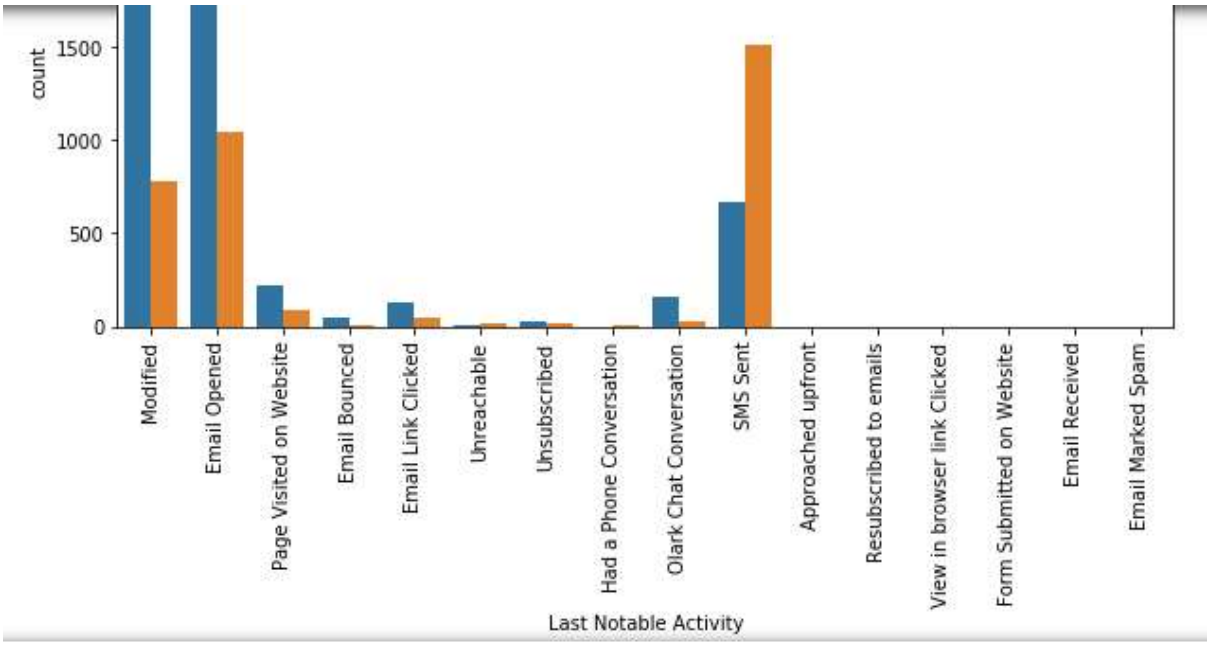
- Many users are choosing course for Better Career Opportunities.



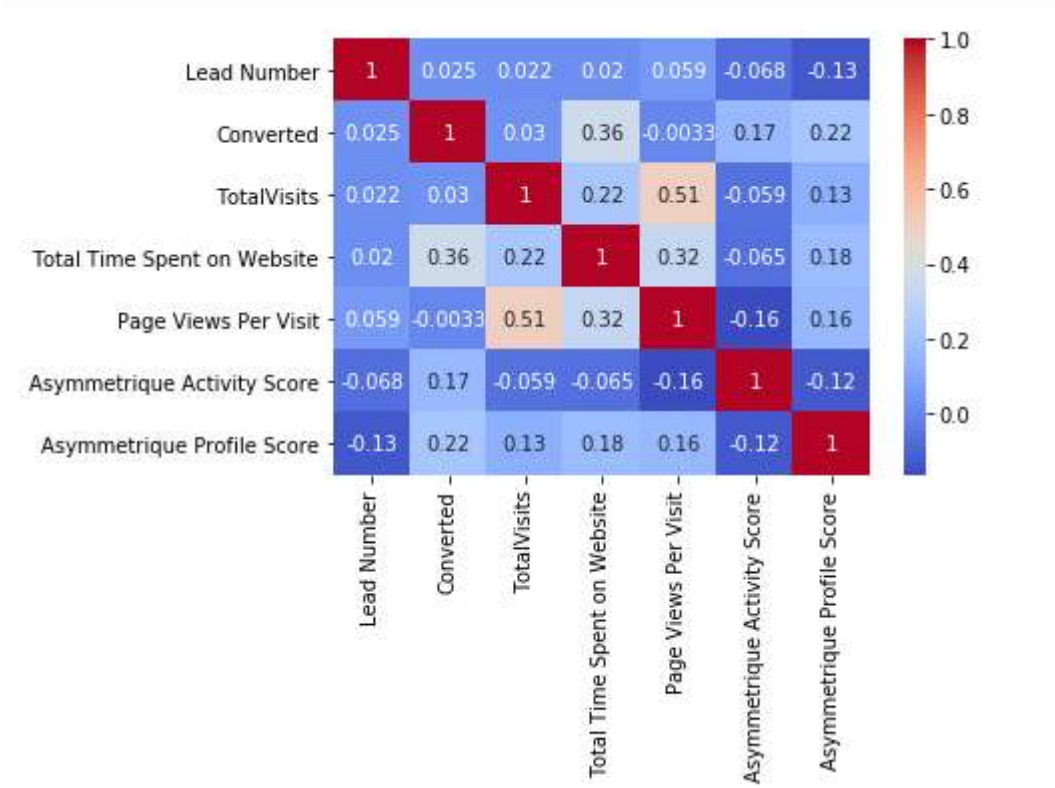
- Coming to the Advertisements from Magazine, Newspaper, Digital Advertisement, Through Recommendation has no impact on potential leads.



- From the Last Notable Activity Column, the users who SMS sent and Email Opened, Modified has higher probability for potential leads.



Correlation:

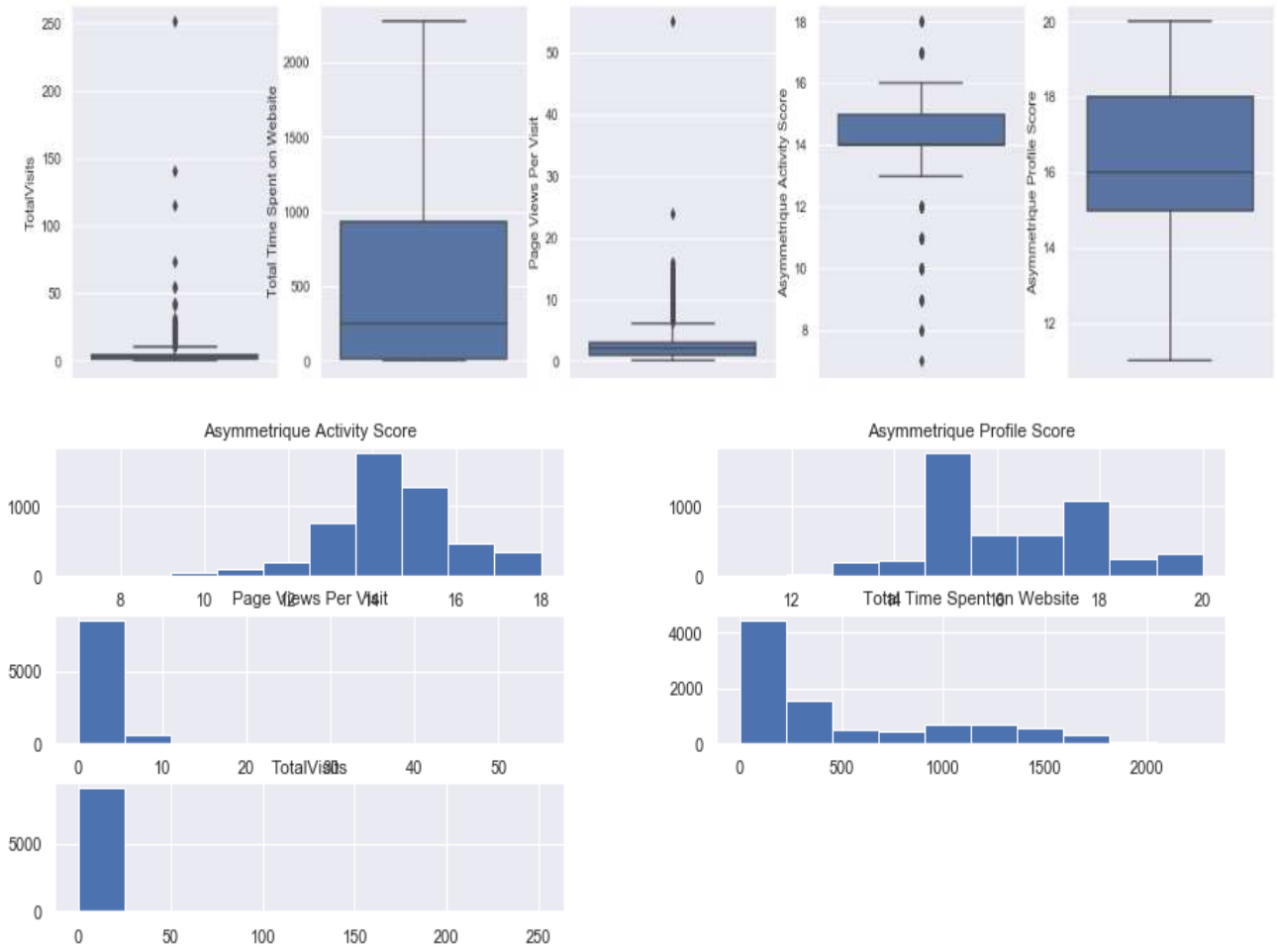


From the above correlation plot:

- "Total time spent on website", "Asymmetrique Activity Score" and "Asymmetrique Profile Score" are correlated with Target variable - "Converted"
- "Total visits" correlated with "Time spent on website", "Asymmetrique Profile Scor" and "Page views per visit"

Outlier Analysis:

From the box plot it is showing that they have outliers for 'TotalVisits','Total Time Spent on Website', Page Views Per Visit, Asymmetrique Activity Score and Asymmetrique Profile Score columns. Let's see them using histogram.



From the above graphs we could see that the datapoints are skewed not outliers. So instead of removing them we can keep them do the scaling to bring them under same distribution.

Data Cleansing:

Missing values Treatment:

Many of the columns has missing values. So missing values are replaced with median for numerical and with mode for categorical variables. Variables having more than 50% of missing values are removed from the analysis. Since anyway missing value imputation is giving bias to those variable. There are few columns having many categories, so in those columns missing values are replaced using random choice. So that it will be randomly replaced the values with all the available categories..

Page Views Per Visit	1.48
Last Activity	1.11
Country	26.63
Specialization	15.56
How did you hear about X Education	23.89
What is your current occupation	29.11
What matters most to you in choosing a course	29.32
Search	0.00
Magazine	0.00
Newspaper Article	0.00
X Education Forums	0.00
Newspaper	0.00
Digital Advertisement	0.00
Through Recommendations	0.00
Receive More Updates About Our Courses	0.00
Tags	36.29
Lead Quality	51.59
Update me on Supply Chain Content	0.00
Get updates on DM Content	0.00
Lead Profile	29.32
City	15.37
Asymmetrique Activity Index	45.65
Asymmetrique Profile Index	45.65
Asymmetrique Activity Score	45.65
Asymmetrique Profile Score	45.65

Creating Dummy variables:

Once the missing values treatment is done then creating dummy variables for the model building. There is a situation for few variables there are many categories. So creating dummy variables for those categories ended up with adding too many variables to the dataset. So instead of creating dummy variables to them in the first place, selecting the categories which are 95% contributing to the dataset and keeping rest of the categories as others. By this way we can eliminate the few columns and then we created dummy variables for the existing main categories.

Label encoding:

There are 2 variables **Asymmetrique Activity Index** and **Asymmetrique Profile Index** consists of low, medium and high categories. Converted them to label encoding to reduce the number of categories by making high-3,medium-2 and low-1.

Scaling:

There are numerical variables like Total time spend on website etc… are in different scales. Applied standard scaler to bring the under same scale

Model Building:

Implemented Logistic regression of stats model to find the coefficients and model statistics. Then implemented Logistic Regression model with RFE to select the desired number of parameters. Here we selected 20 features and calculated the

predicted accuracy, precision and Recall, ROC curve to find the optimum probability cutoff. Then find the VIF to check the correlation and removed the highest correlated variable which is having high VIF and re-iterate the model building steps above until there is no value >4 in VIF.

Output:

As per the requirement they need 0-100 score for leads. So stored the probabilities of training and test variables to new variable and joined that based on index with the original dataframe.

Learnings from the Analysis:

- **Factors that are mainly contributing to form Potential Leads:**
  - other_Lead_Origin means 'Lead Add form', 'Lead Import' and 'Quick Add Form' together from 'Lead Origin' is likely to contribute
  - 'Lost to EINS' from Tags is 24.35 times more likely to contribute
  - 'Closed by Horizzon' from Tags is 15.67 times more likely to contribute
  - 'SMS Sent' from last_notable_activity is 14.04 times more likely to contribute
  - 'Will revert after reading the email' from Tags is 11.7 times more likely to contribute
  - Busy from Tags is 3.12 time more like to contribute
  - Total Time Spent on Website is 2.8 times more like to contribute
- **Factors not creating any positive affect on Leads:**
  - 'Olark Chat Conversation' from Last notable activity is 5.97 times not likely creating any impact
  - Do not Email column is 3.82 times not likely
  - 'switched off' from Tags is 2.72 times not likely
  - Ringing from Tags is 2.60 times not likely
  - 'Interested in other courses' from Tags is 1.74 times not likely
  - 'Already a student' from Tags is 1.46 times not likely to create any impact to Potential Leads