X-EDUCATION LEAD SCORE CASE STUDY

Identification of Hot Leads to focus more on them and thus enhancing the conversion ratio for X Education

By,
Madhumitha R

Back Ground of X Education

- X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.
- 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.

Back Ground of X Education

- 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%.

Problem Statement

- Although X Education gets a lot of leads, its lead conversion rate is very poor.
- 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.

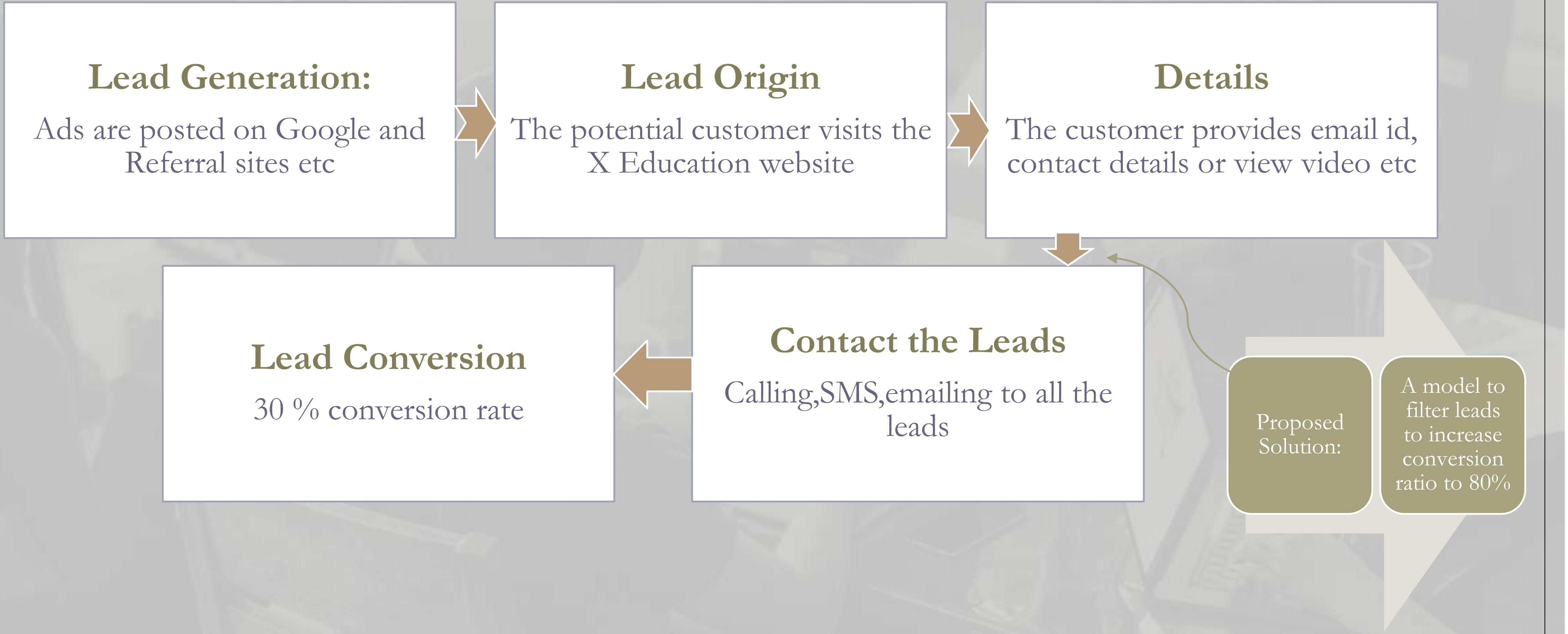
Problem Statement

- The company requires us to build a model wherein we need to assign a lead score 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.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Objective

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

LEAD CONVERSION PROCESS



Lead Generation:

Ads are posted on Google and Referral sites etc

Lead Origin

The potential customer visits the X Education website

Details

The customer provides email id, contact details or view video etc

Contact the Leads

Calling,SMS,emailing to all the leads

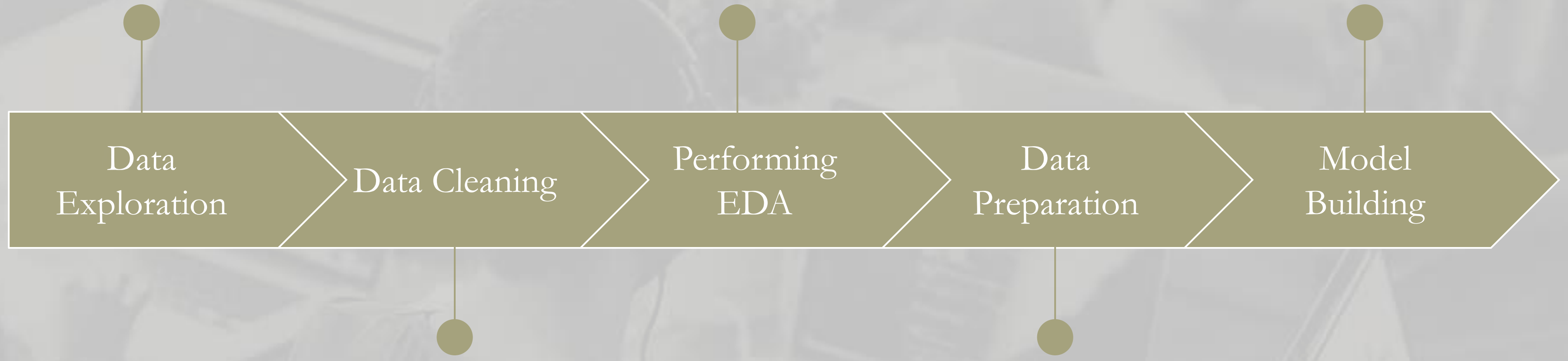
Lead Conversion

30 % conversion rate

Proposed Solution:

A model to filter leads to increase conversion ratio to 80%

Implementation



Data Exploration

Loading & Observing the past data provided by the Company

Data Cleaning

Duplicate removal, null value treatment, unnecessary column elimination, etc.

Performing EDA

Univariate, Bivariate, and Heatmap for numerical and categorical columns

Data Preparation

Outlier Treatment, Feature-Standardization

Model Building

Build an optimal model of Logistic Regression using automatic(RFE) and manual (p value and VIF) feature selection

Data Exploration

- Leads.csv contains all the information about the leads generated from various sources.
  - The file contains 9240 rows and 37 columns.
  - Out of 37 columns 7 are numeric and 30 are categorical columns.
  - The current conversion rate of leads is 38.5% .

- Leads Data Dictionary excel file is a data dictionary that describes about the attributes in the Leads.csv file.
- The following columns have missing value greater than 45%.Hence we proceed with dropping of these columns:
  1. How did you hear about X Education
  2. Lead Quality
  3. Lead profile
  4. Asymmetrique Activity Index
  5. Asymmetrique Profile Index
  6. Asymmetrique Activity Score
  7. Asymmetrique Profile Score

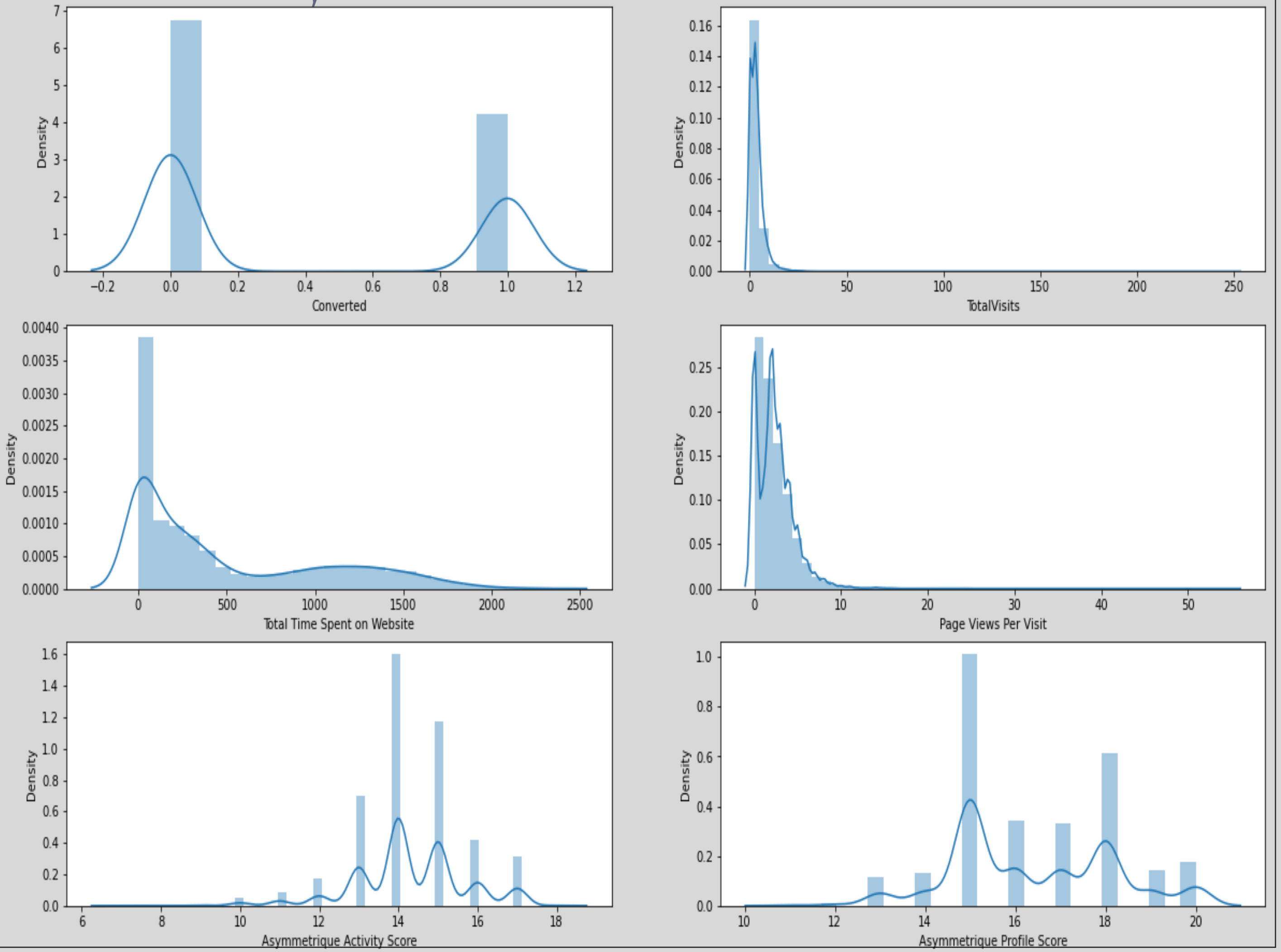
Data Cleaning

- The following columns have Select as value. Replaced the Select with null value.
  - Lead Profile
  - City
  - Specialization
  - How did you hear about X Education .

- Replaced the null values and field category having low count in each columns with Others/Not Specified.
  - Lead Source
  - Last Activity
  - Country
  - Specialization
  - Tags
- Replaced now the columns having fewer missing values with median for numerical columns and mode for categorical columns(What matters most to you in choosing a course, What is your current occupation ,City , Country)

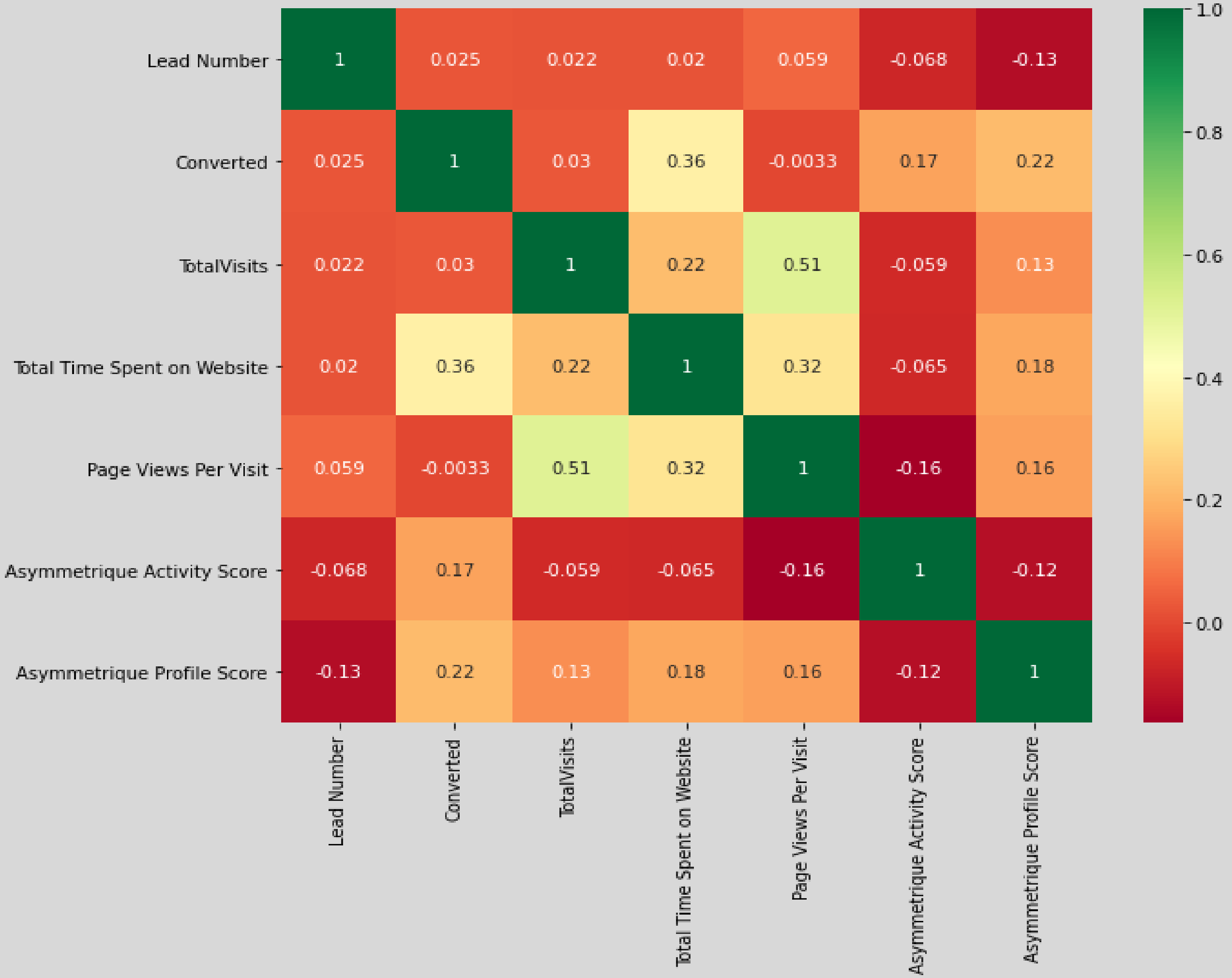
EDA - Univariate Analysis

- From the above plots we can conclude that the converted value is high for 0 than 1.It means most of the leads remains not converted to business.
- The Total Time Spent on Website is right skewed and is less than 500.It has a peak between 1000 and 1500.
- The Asymmetrique Activity Score is left skewed and the value is highest for 14 and then followed by 15.
- The Total visits is right skewed and the peak is between 0-20 .
- The Page Views per Visit is between 0 and 10.The values peaks at 0 and slowly decreases later on .It is right skewed
- The Asymmetrique Profile is left skewed and the value is highest for 15 and then followed by 18.



Heatmap-Correlation

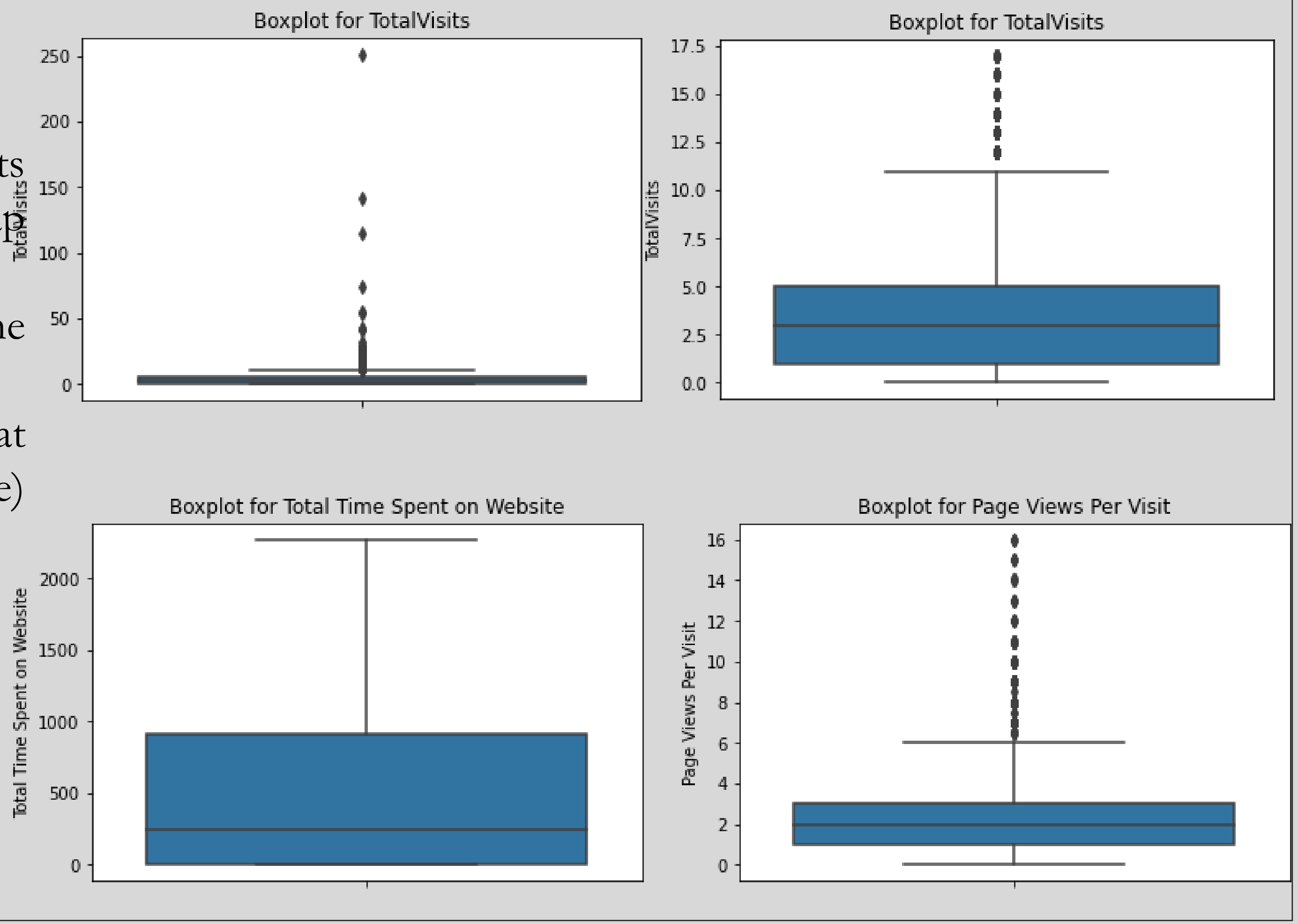
- From the plot nothing much could be inferred.
- There are no two columns that have high correlation between them.
- The columns Total Visits and Total time spent on Website are positively correlated .



	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Asymmetrique Activity Score	Asymmetrique Profile Score
Lead Number	1	0.025	0.022	0.02	0.059	-0.068	-0.13
Converted	0.025	1	0.03	0.36	-0.0033	0.17	0.22
TotalVisits	0.022	0.03	1	0.22	0.51	-0.059	0.13
Total Time Spent on Website	0.02	0.36	0.22	1	0.32	-0.065	0.18
Page Views Per Visit	0.059	-0.0033	0.51	0.32	1	-0.16	0.16
Asymmetrique Activity Score	-0.068	0.17	-0.059	-0.065	-0.16	1	-0.12
Asymmetrique Profile Score	-0.13	0.22	0.13	0.18	0.16	-0.12	1

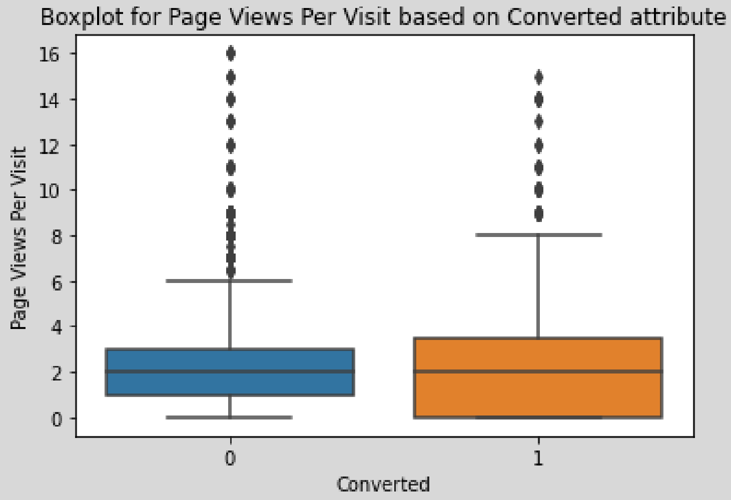
Outliers-Boxplot

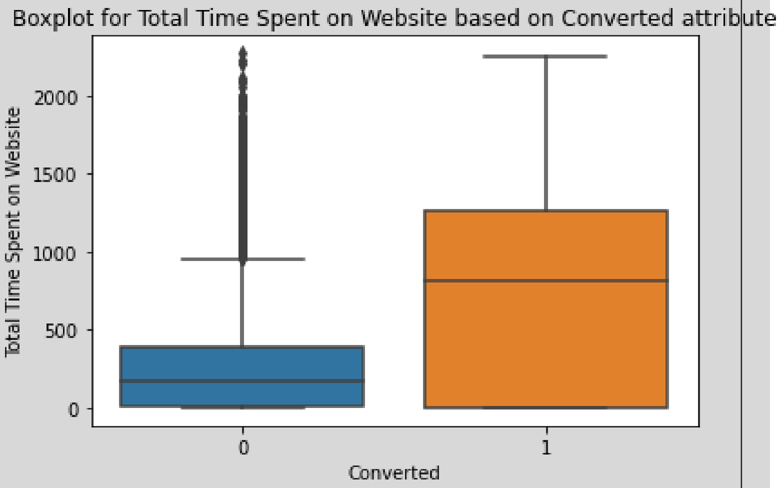
- We can see that the boxplot for TotalVisits have outliers .Hence we will use capping and cap the lower limit at 0.01 and upper limit at 0.99 .
- There is no outlier for the column 'Total Time Spent on Website’. Hence we will leave it as it is.
- We can see there are outliers .It is possible that the page views per visit be 16(maximum value) .Hence we ignore those outliers.

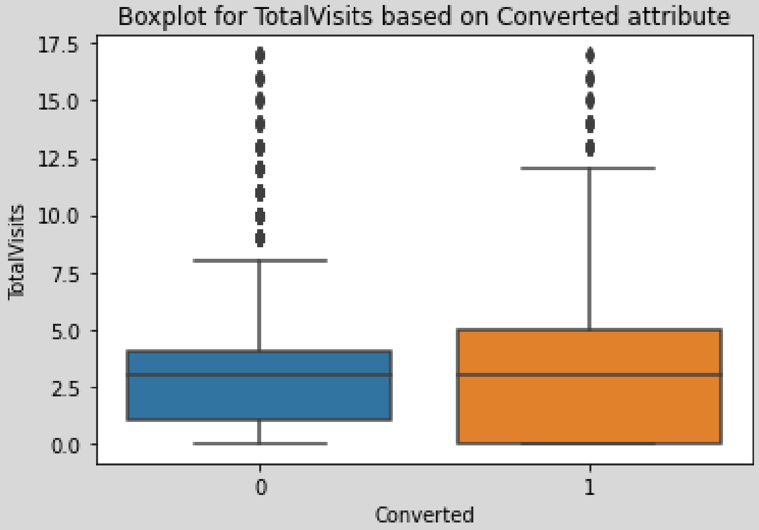


Bivariate Analysis

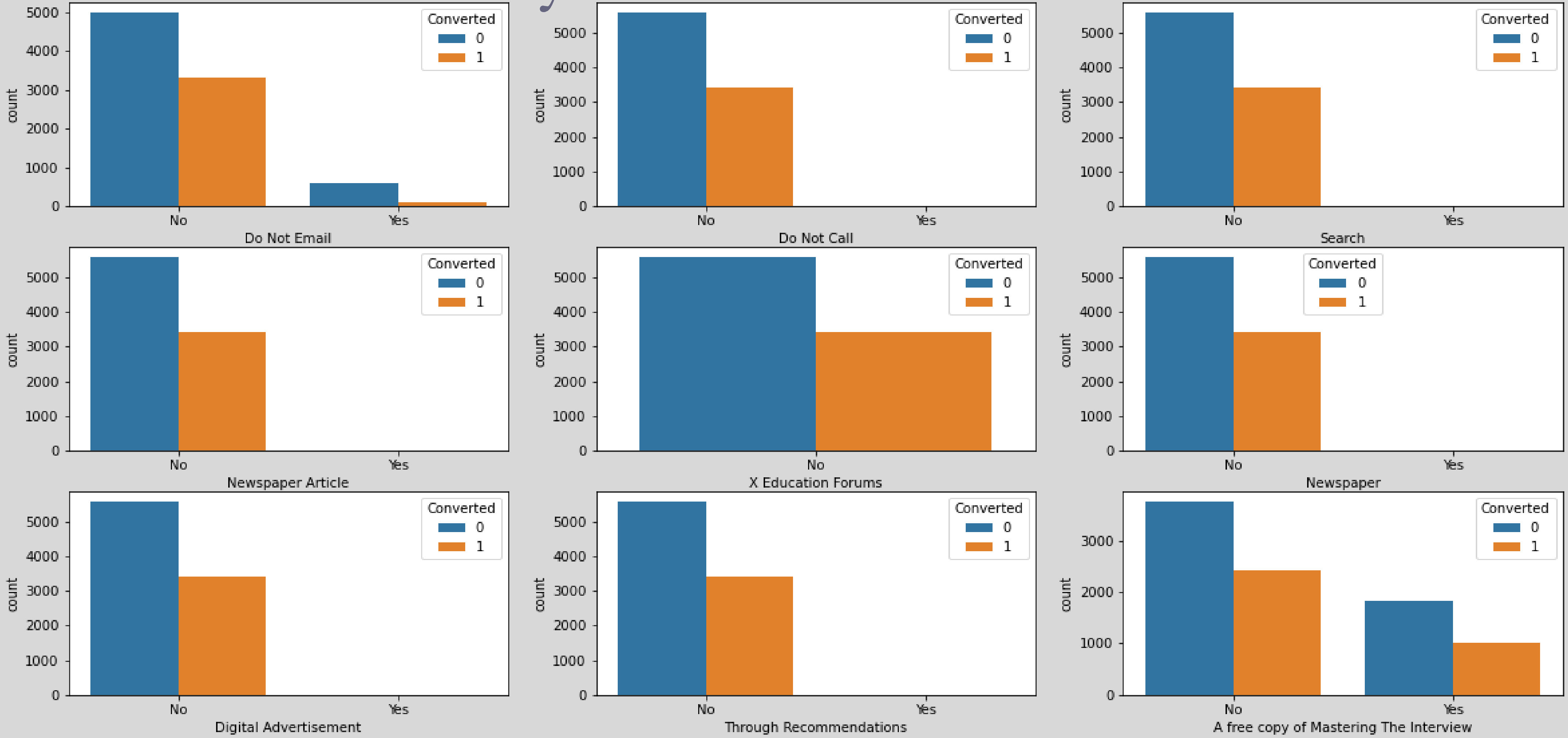
- Nothing much can be inferred from the above chart of Page Views Per Visit.The median is same for converted and not converted
- The Total Time Spent on Website is high for the data converted when compared to Not converted.So customers tends to spend more time on website before accepting the course
- From the above chart it is clearly evident that when TotalVisits count is high from the 75% for getting converted.But the medain is same for both


Boxplot for Page Views Per Visit based on Converted attribute


Boxplot for Total Time Spent on Website based on Converted attribute


Boxplot for TotalVisits based on Converted attribute

Bivariate Analysis



Bivariate Analysis

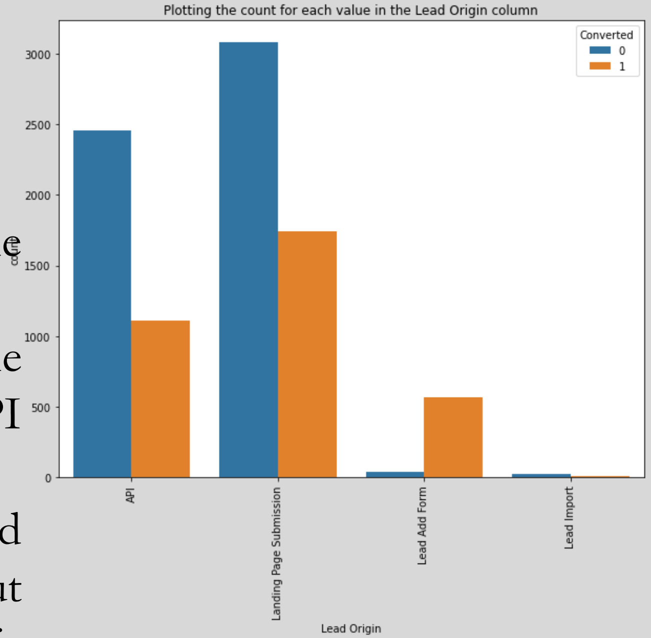
From the above subplot we can infer the following:

- The customer who have wished to not receive any emails tend to be not converted.
- Most of the customer prefer calls
- From the above plot it can be concluded that whether the customer had seen the ad in any of the listed items:'Search', 'Newspaper Article','X Education Forums','Newspaper','Digital Advertisement','Through Recommendations' are very low.
- Not much can be inferred from the column 'A free copy of Mastering the Interview' as we can see from the graph that either they opt for it or not around half the percentage of each group tends to be converted.

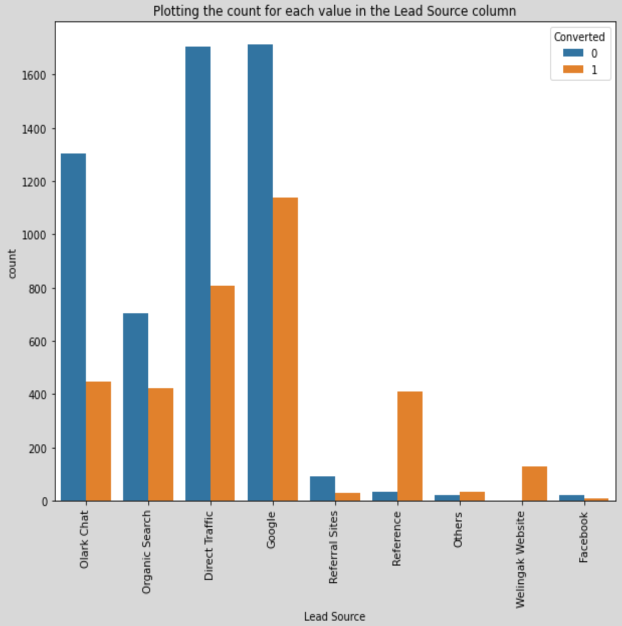
Bivariate Analysis

From the above plot we can infer the following:

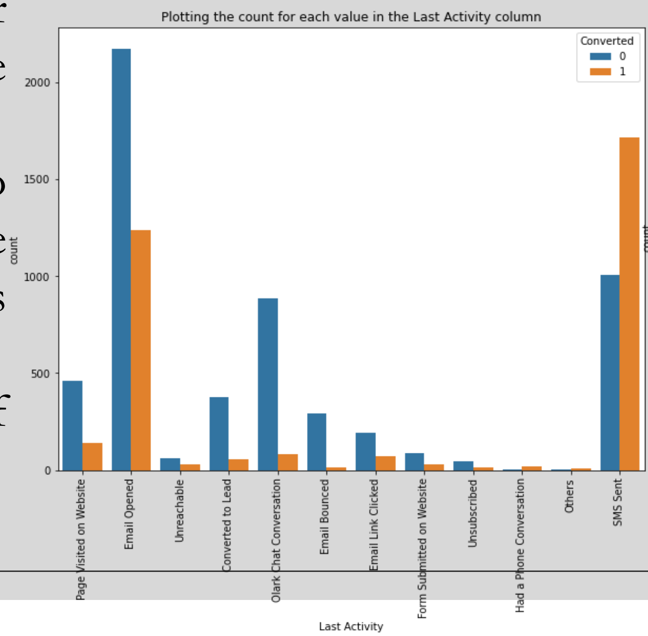
- The Landing Page Submission has the high conversion value followed by API and Lead Add Form.
- The converted count are high for Lead Source of Google and Direct Traffic.But when the customers are 'Reference' their chance of getting converted to leads are high compared to not getting converted.
- The chance of getting converted to potential leads is high when the customer sents SMS,When the Email is opened.
- From The chart we can see majority of the clients are from India.



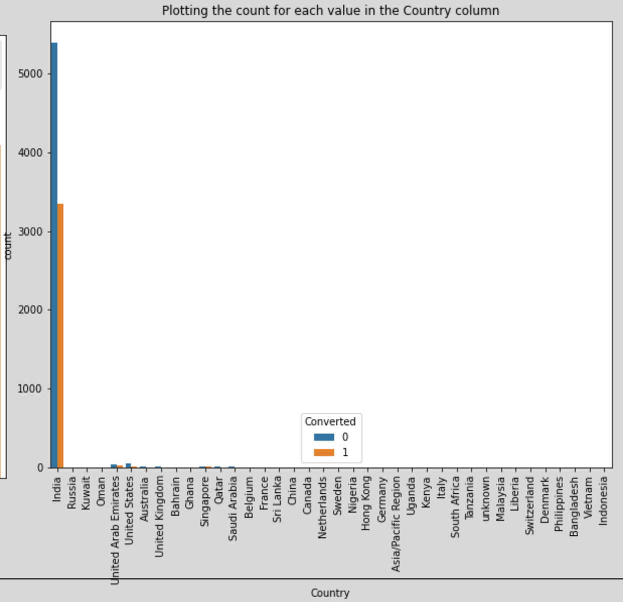
Plotting the count for each value in the Lead Origin column



Plotting the count for each value in the Lead Source column



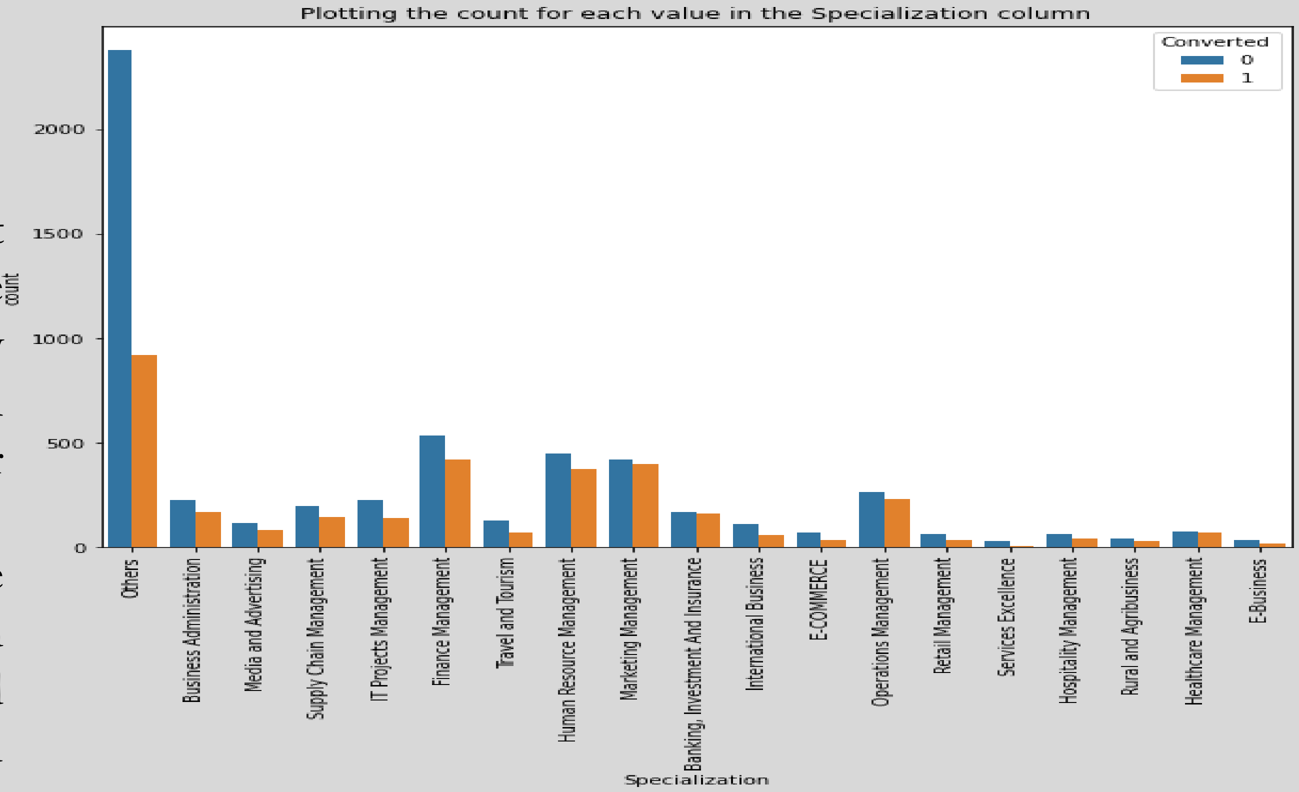
Plotting the count for each value in the Last Activity column

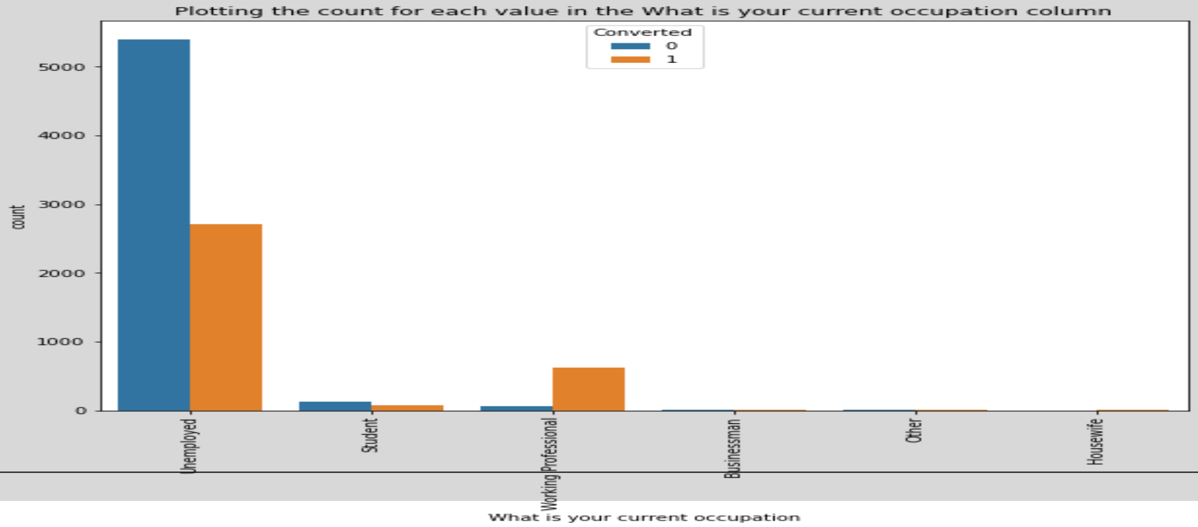


Plotting the count for each value in the Country column

Bivariate Analysis

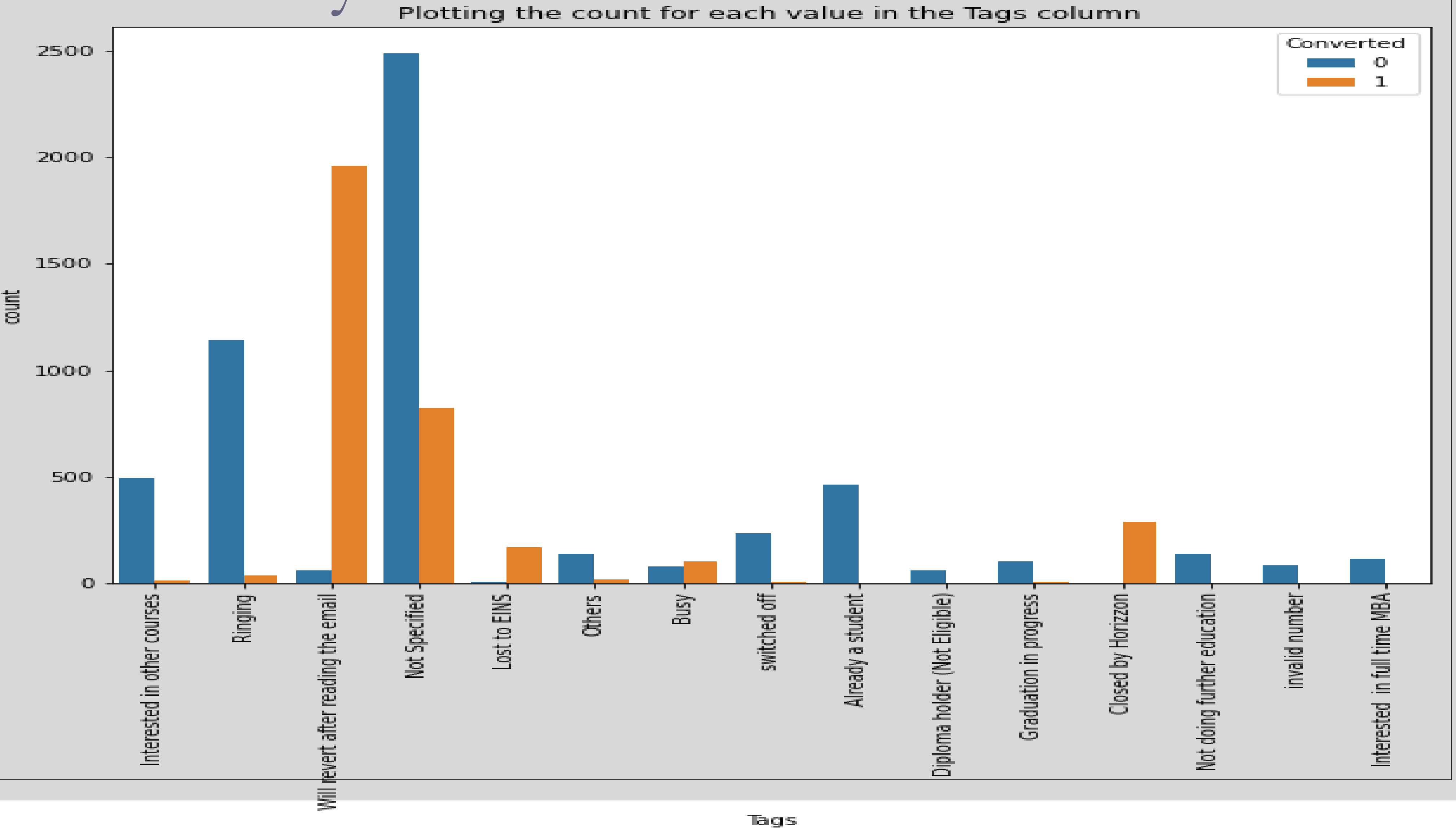
- Majority of the customers have not mentioned their specialization. The Finance management followed by Marketing Management and Human Resource management have the higher chance of converting to lead.
- The Working Professional have the higher chance of converting to leads and joining the course. The unemployed have the highest count of converted leads.
- From the chart above it is clearly evident that the category 'Will Revert after reading the email' has the highest chance of getting converted to leads. Followed by Lost to EINS and Closed by Horizzon.





Bivariate Analysis

- From the chart above it is clearly evident that the category 'Will Revert after reading the email' has the highest chance of getting converted to leads. Followed by Lost to EINS and Closed by Horizzon.



Data Preparation

- **Binary variables Encoding**

  Variables which have binary(yes/no) values are encoded with 1 and 0. 1 denotes Yes and 0 denotes No .

- **Create Dummy Variables**

  Dummy variables allows easy interpretation and calculation of odds ratio which increases the stability and significance of the coefficients. Dummy variables are created for the following columns:

  1. Lead Origin
  2. Specialization
  3. Tags
  4. Last Notable Activity
  5. Lead Source
  6. Last Activity
  7. Country
  8. City
  9. What is your current occupation
  10. What matters most to you in choosing a course

Data Preparation

- **Train-Test Split**

The 'Leads' dataset is now split into Training and Testing set in 70:30 ratio.The Training dataset is used to the train the model whereas the testing dataset is used for prediction and evaluation of model.

- **Feature Scaling**

It is important to have all the variables on the same scale in order to avoid the dominance of variables with high magnitude in the model.

'StandardScaler' function has been used to scale the data for modelling which brings all the data points into a standard normal distribution with mean at 0 and standard deviation at 1.

Model Building

Generalised Linear Model(GLM) from statsmodels library has been used to build a logistic Regression model. Initially the model had 99 columns present in the X_train dataset. Most of the features are insignificant ,we need to use feature selection technique.

- **Feature Selection using Recursive Feature Elimination(RFE)**

RFE is an optimization technique for finding the best performing subset of features. It is based on the idea of repeatedly constructing a model and choosing either the best (based on coefficients), setting the feature aside and then repeating the process with the rest of the features. This process is applied until all the features in the dataset are exhausted. Features are then ranked according to when they were eliminated.

We ran RFE to identify top 15 features for further model building process. Insignificant features were dropped one by one after checking the P-value and Variance Inflation Factor (VIF). **Accepted P-value should be kept below 0.05 and VIF should be less than 5.**

Model Building

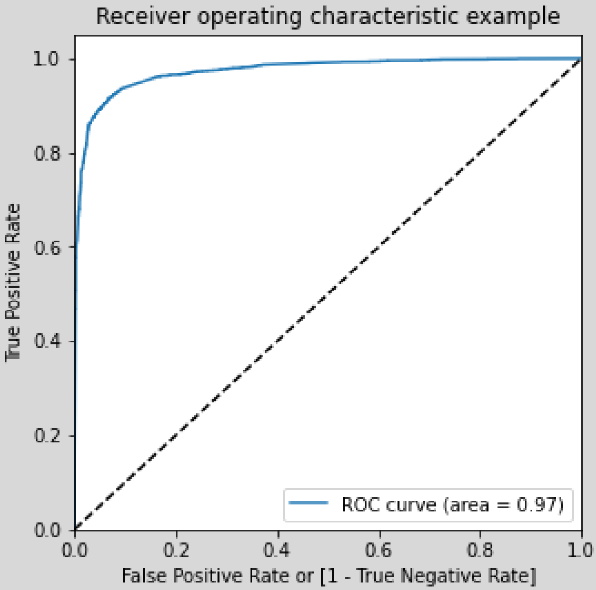
Final Model and Evaluation Metrics

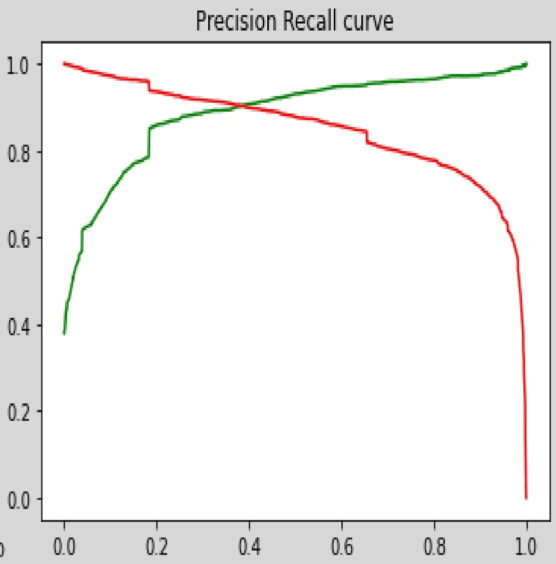
- Final model contains 14 most important features which satisfy all the selection criteria.
- Lead score having conversion probability greater than 0.5 are being predicted as “Converted”.
- Using this probability threshold value (0.5), the leads from the test dataset have been predicted whether they would get converted or not.
- Confusion matrix with cut-off 0.5 has been created to calculate evaluation metrics.
- Confusion matrix:
  ```
  [[3782 155]
   [291 2086]]
  ```
- Metrics
  - Accuracy : 92.93%
  - Sensitivity : 87.75%
  - Specificity: 96.06%

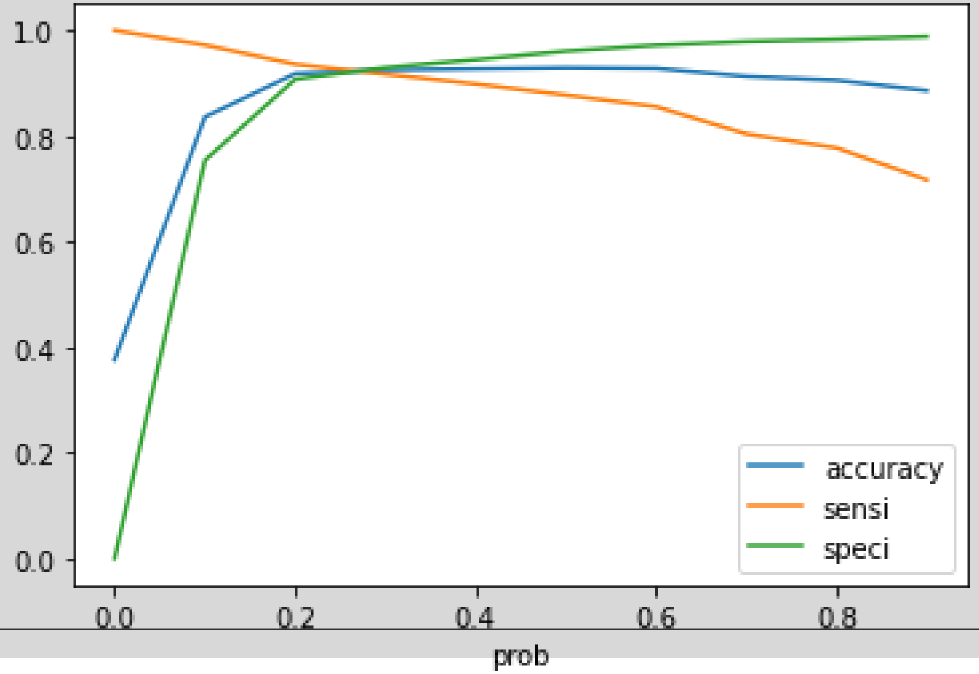
Model Building

Evaluation Metrics

- Receiver Operating Characteristics (ROC) Curve:
  - By determining the Area Under the Curve (AUC) of the ROC curve, the goodness of the model is determined.
  - Since the ROC curve is close to the upper left part of the graph, it means this model is a very good model.
  - The value of AUC for our model is 0.97.
  - Plot accuracy sensitivity and specificity: Tradeoff between sensitivity and accuracy can be observed (cutoff = 0.3).
  - Precision and Recall plot: Ideal cutoff of 0.4 is observed from recall and precision plot.







Model Building

Evaluation Metrics

- Lead score having conversion probability greater than 0.3 (optimal cut off) are being predicted as "Converted".
- Confusion matrix with cut-off 0.3 has been created to calculate evaluation metrics.
- Training set

```
Confusion matrix: [[3660 277]
                   [197 2180]]
```

Metrics

- Accuracy : 92.49%
- Sensitivity : 91.71%
- Specificity: 92.96%

Model Building

Evaluation Metrics

- Using this probability threshold value (0.3), the leads from the test dataset have been predicted whether they would get converted or not.
- Testing set

  Confusion matrix: [[1524 134]
  [84 964]]

  Metrics

  - Accuracy : 91.94%
  - Sensitivity : 91.98%
  - Specificity: 91.91%

- The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model.

Conclusion and Recommendations

After completeing the Model building the following are my recommendations :

- From the coeff values of the model it is clearly evident that the following variables should be focussed the most inorder to increase the probability of lead conversion
  a) Tags_Closed by Horizzon
  b) Tags_Lost to EINS
  c) Tags_Will revert after reading the email
- And the 3 variables that influence the lead conversion are-Tags,Lead Source ,Last Activity . Then followed by Lead Origin,Last Notable Activity,Specialization and Total Time Spent on Website

Conclusion and Recommendations

Apart from this from the EDA we conclude that :

- The customer who have wished to not receive any emails tend to be not converted.So its better to target only people who opt to receive emails from the organization
- The customer had seen the ad in any of the listed items:'Search', 'Newspaper Article','X Education Forums','Newspaper','Digital Advertisement' are very low.
- The converted count are high for Lead Source of Google and Direct Traffic.Hence it is suggested the company promotes their ads on the Google page .
- The Landing Page Submission has the high conversion value. So when the people have submitted the forms their chance of converting is high .Hence we can target those clients through calls ,emails and sms.
- Also when the customer sents us sms or when they have the email is opened their chance of turning to lead is high.
- The unemployed,Working professionals,student have higher chance of converting to leads .Hence we have to target these category of clients.

THANK YOU!