LEAD SCORE CASE STUDY

SUJITH M TOM
SHARIE R NATH

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.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.

BUSINESS UNDERSTANDING

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.

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%.

BUSINESS UNDERSTANDING contd..

- 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.

BUSINESS GOAL

- X Education requires to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company requires to build a model wherein you 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, has given a ballpark of the target lead conversion rate to be around 80%.

GOALS OF CASE STUDY

- 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.
- There are some more problems presented by the company which our model should be able to adjust to if the company's requirement changes in the future so we will need to handle these as welL

DATA UNDERSTANDING

- Provided with a leads dataset from the past with around 9000 data points.
- This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.

TARGET VARIABLE

- The **target variable**, in this case, is the column '**Converted**' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted

Exploratory Data Analysis Approach

The dataset is being analysed using following steps:

- Data Sourcing
- Data Cleansing
- Binary Mapping
- Dummy variable Creation
- Segmented Univariate Analysis
- Bivariate Analysis
- Arriving at Insights

Data Sourcing

- ***Leads.csv*** Provided with a leads dataset from the past with around 9000 data points.

Data Cleansing

- Fixing Rows and Columns
- Dealing with missing values
- Detecting Outliers
- Lead Conversion Rate

Fixing Rows and Columns

- Duplicate Rows and Columns are removed
- Insignificant columns are removed

Dealing with missing values

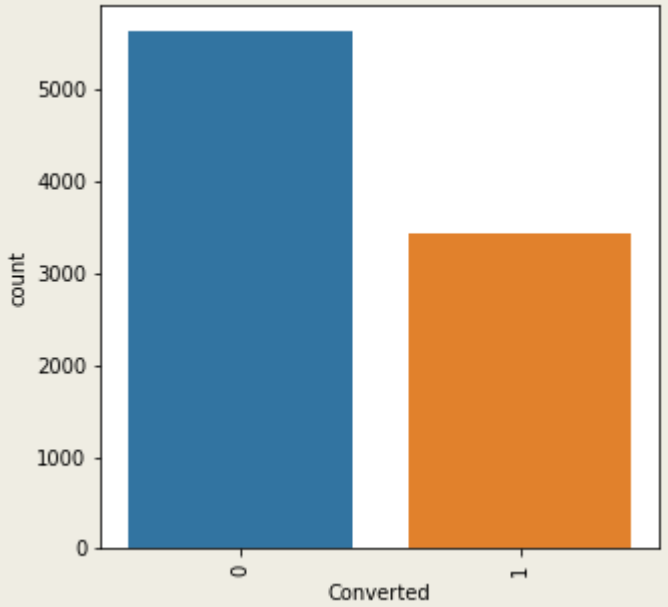
- ‘Select’ values in certain columns could be considered as NULL values
- Columns with more than 70% missing values identified and removed
- Null values in categorical columns replaced with appropriate values

Detecting Outliers

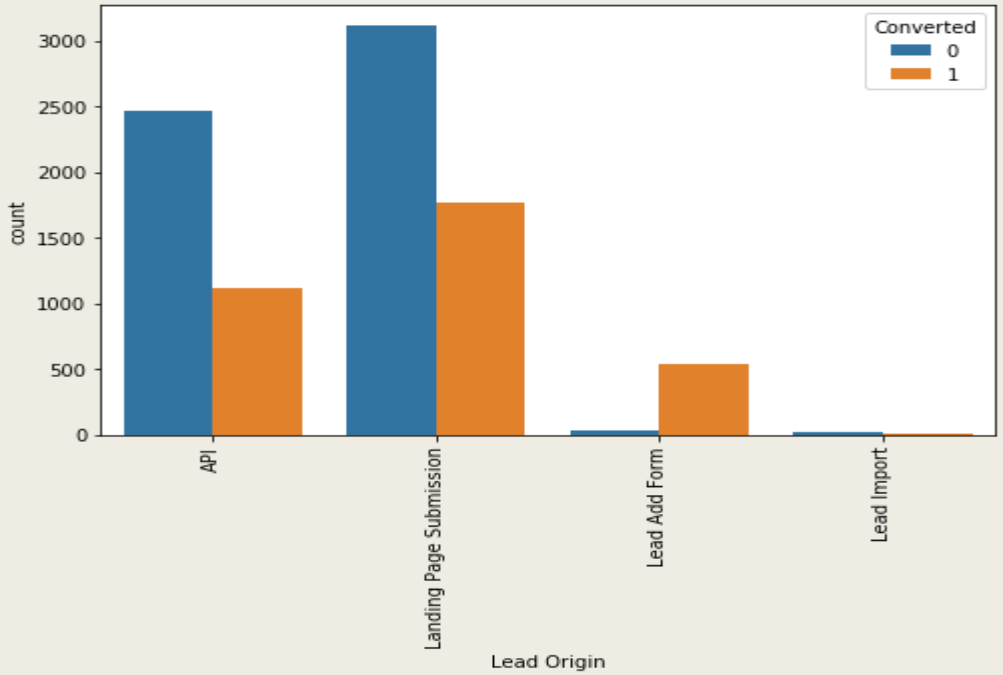
- Numerical columns are analysed for identifying outliers using **Box plots**
- Outlier Datapoints are detected using **Quantile value method**
- Datapoints capped to an upper quantile of 0.95 and lower quantile of 0.05

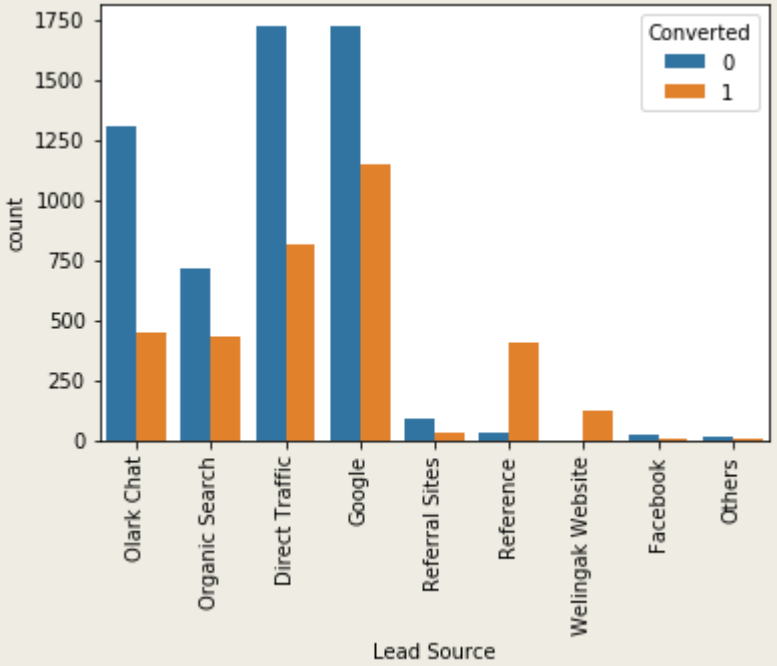
Lead Conversion Rate

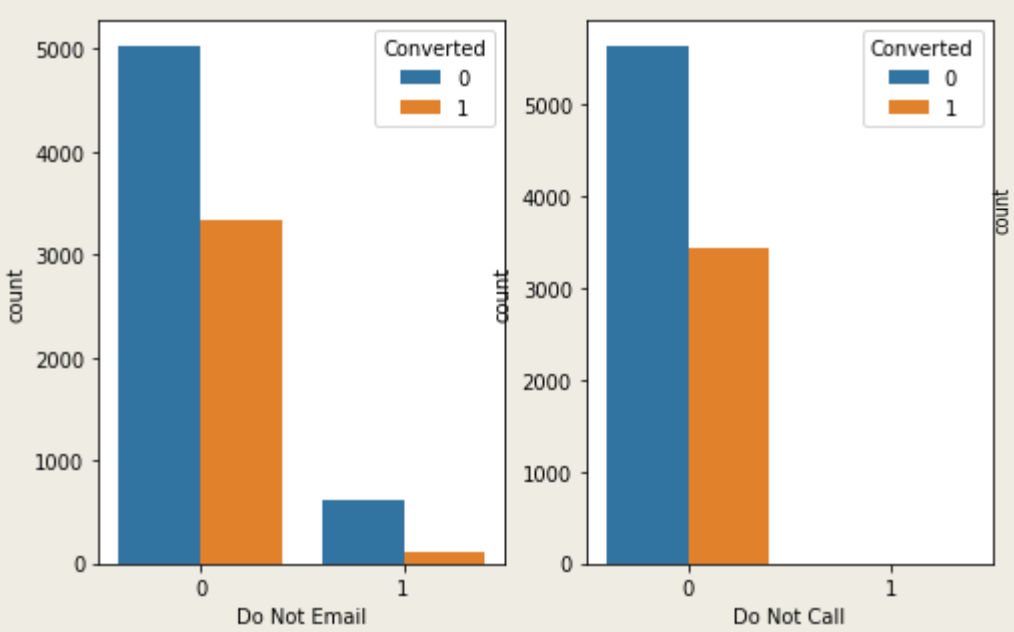
Lead Conversion Rate as per the dataset= 37.86 %

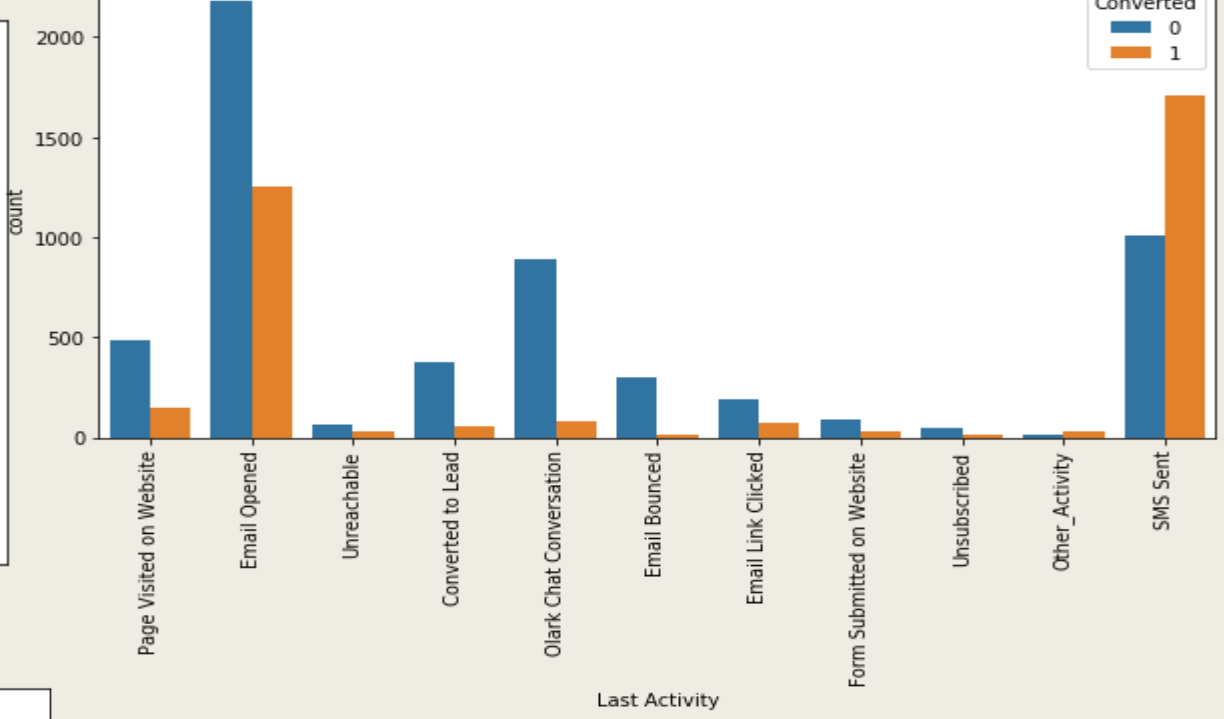


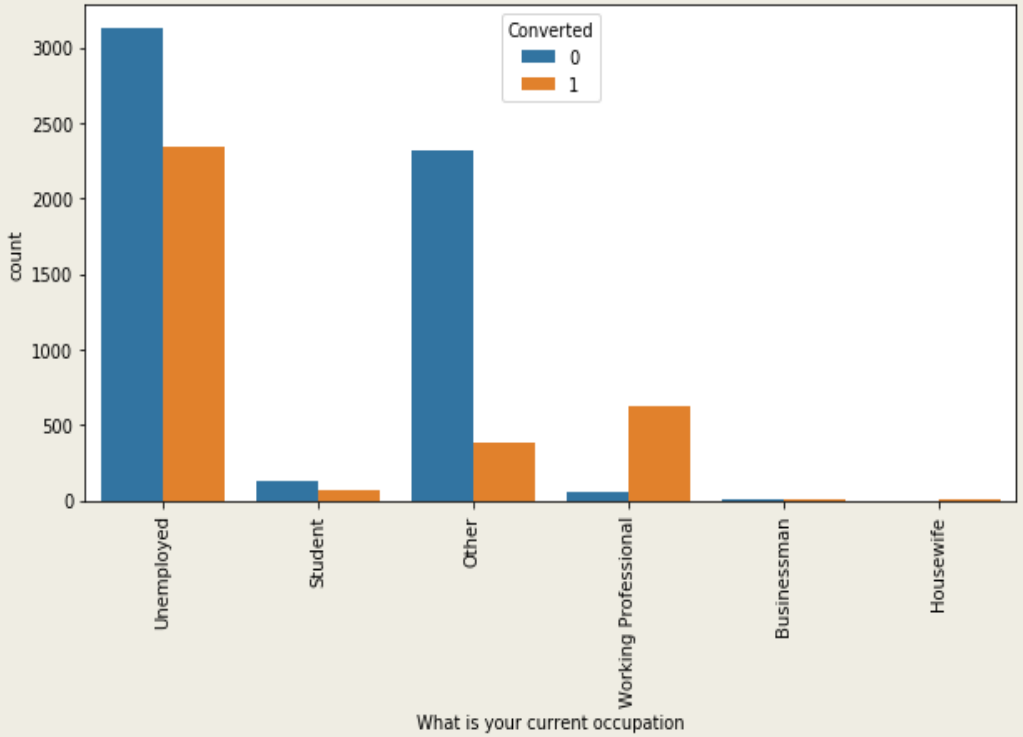
Segemented Univariate Analysis

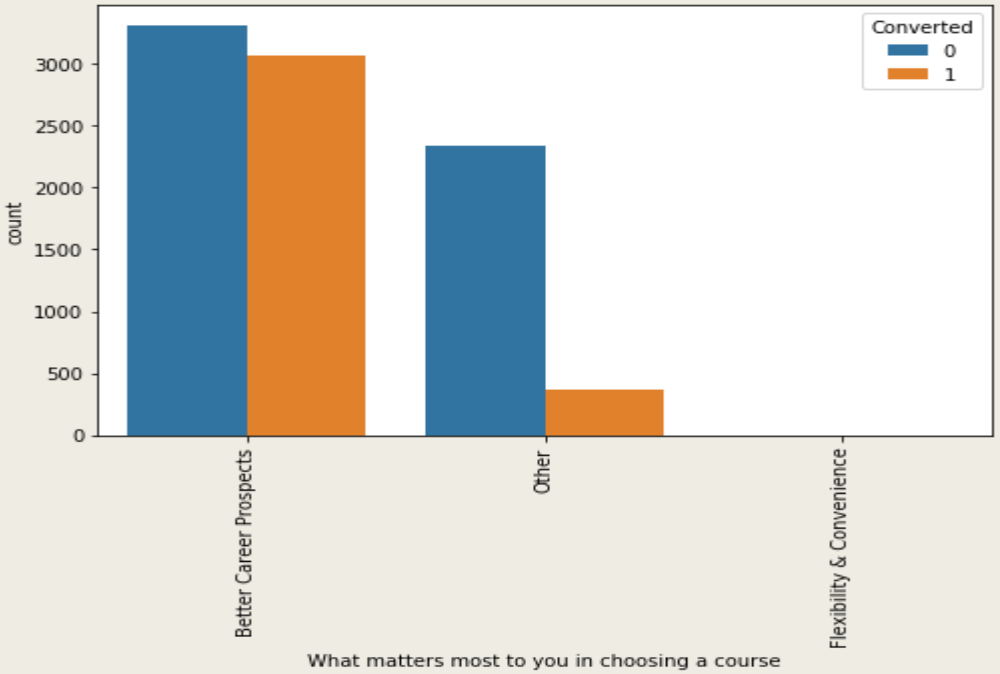


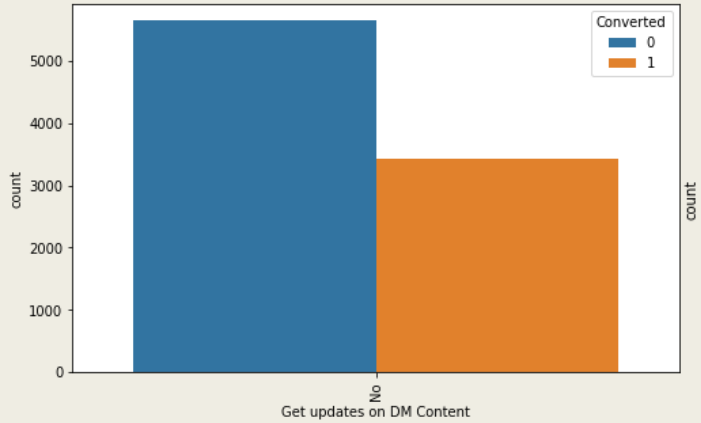


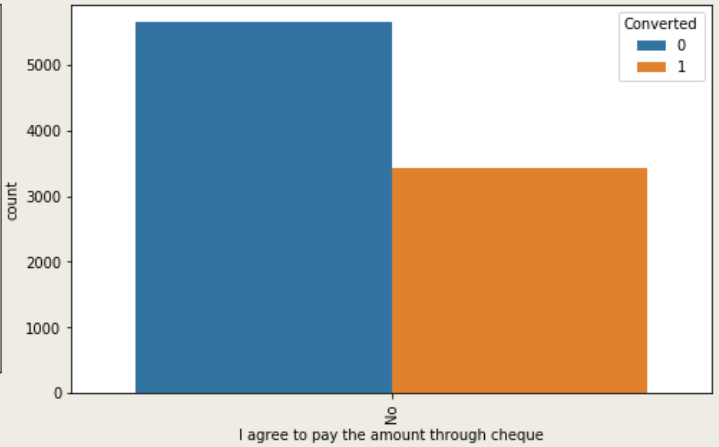

Converted
0
1
count
Do Not Email
Do Not Call

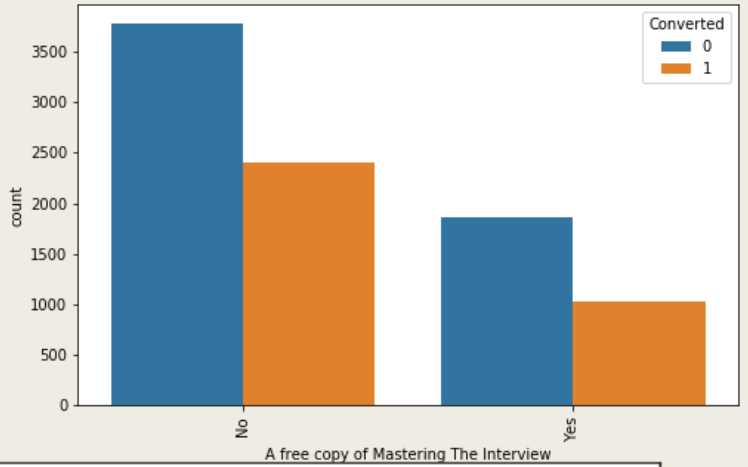

Converted
0
1
count
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Other_Activity
SMS Sent
Last Activity

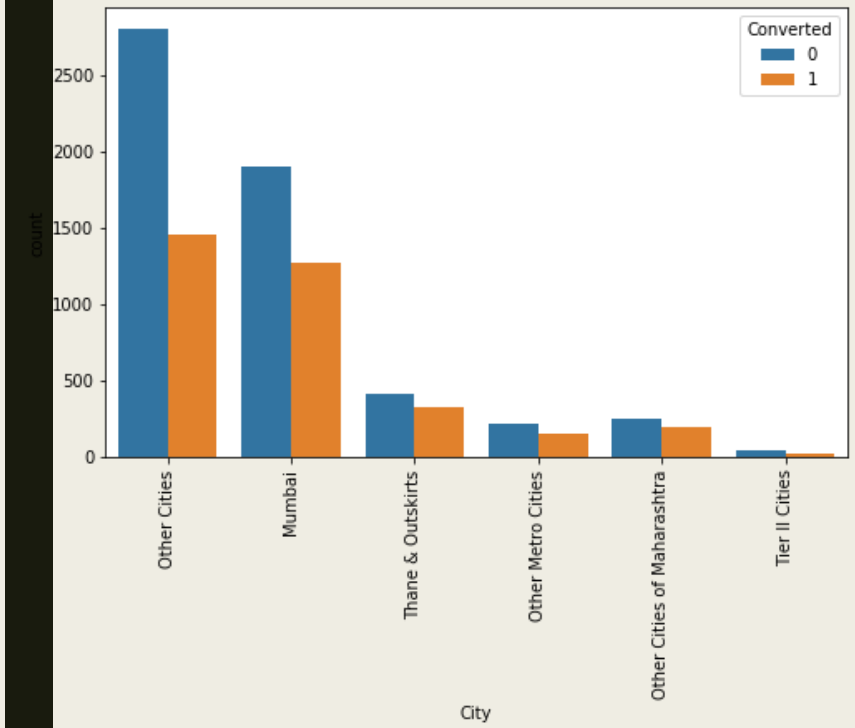

Converted
0
1
count
Unemployed
Student
Other
Working Professional
Businessman
Housewife
What is your current occupation

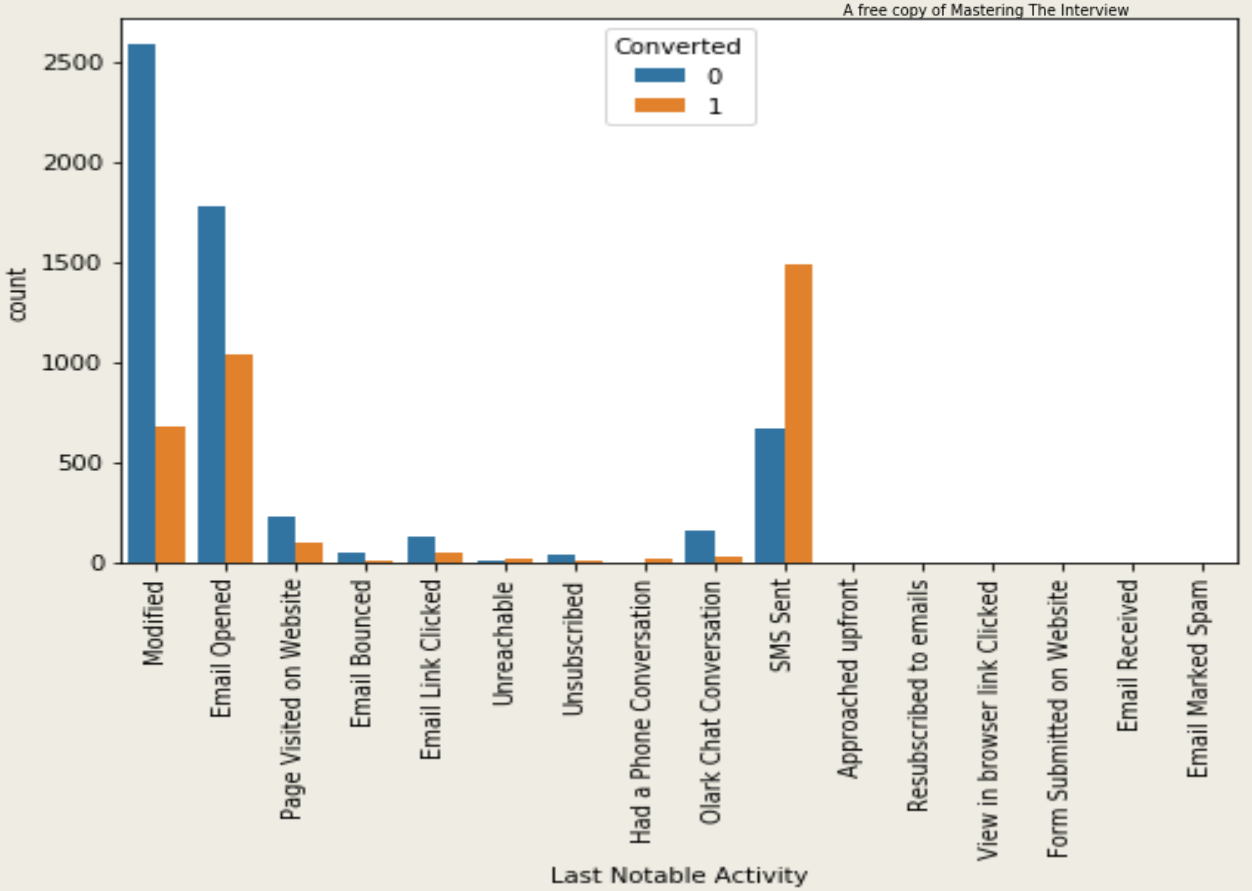

Converted
0
1
count
Better Career Prospects
Other
Flexibility & Convenience
What matters most to you in choosing a course

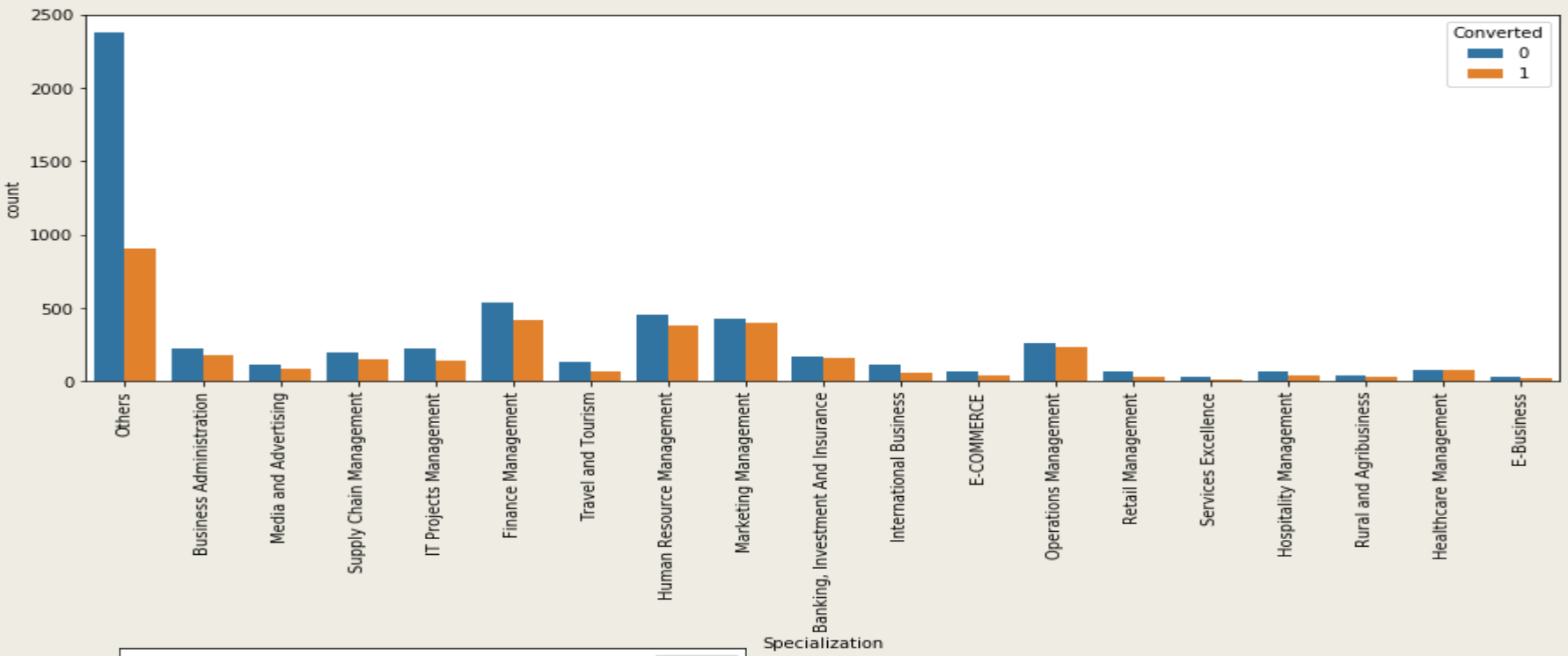

Converted
0
1
count
Get updates on DM Content
No

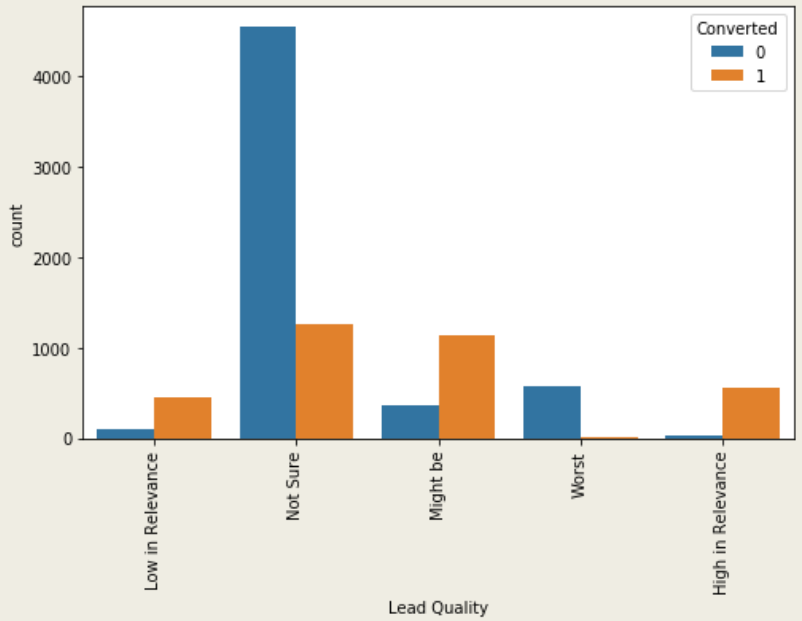

Converted
0
1
count
I agree to pay the amount through cheque
No


Converted
0
1
count
A free copy of Mastering The Interview
No
Yes

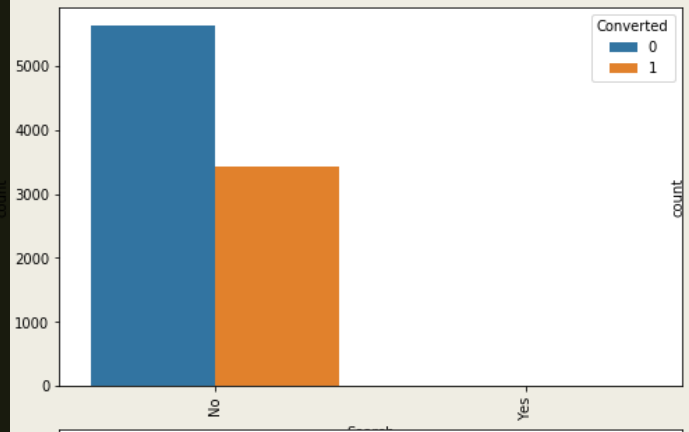

Converted
0
1
count
City
Other Cities
Mumbai
Thane & Outskirts
Other Metro Cities
Other Cities of Maharashtra
Tier II Cities

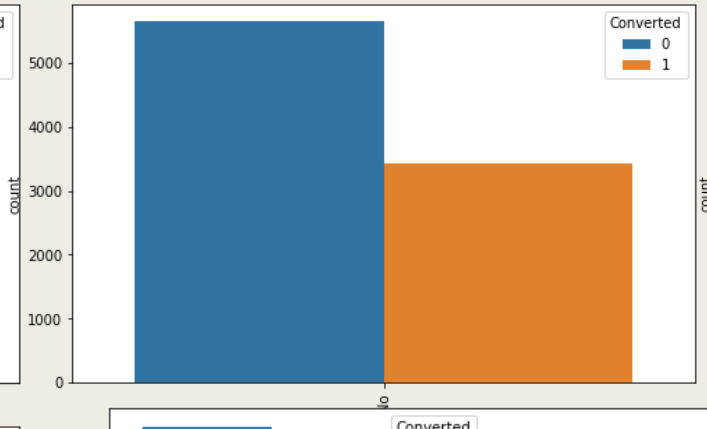

Converted
0
1
count
Last Notable Activity
Modified
Email Opened
Page Visited on Website
Email Bounced
Email Link Clicked
Unreachable
Unsubscribed
Had a Phone Conversation
Olark Chat Conversation
SMS Sent
Approached upfront
Resubscribed to emails
View in browser link Clicked
Form Submitted on Website
Email Received
Email Marked Spam

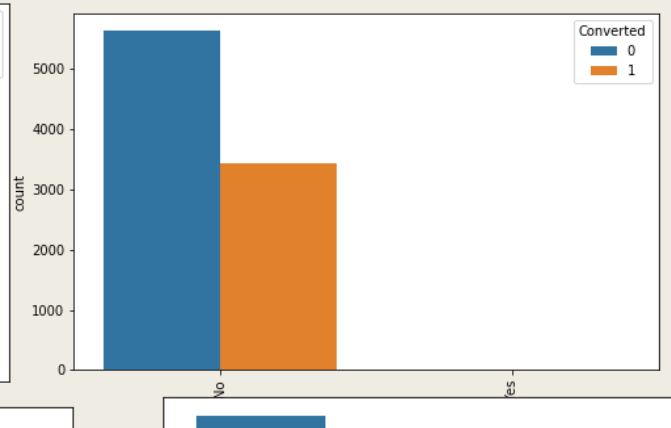

Converted
0
1
count
2500
2000
1500
1000
500
0
Others
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization

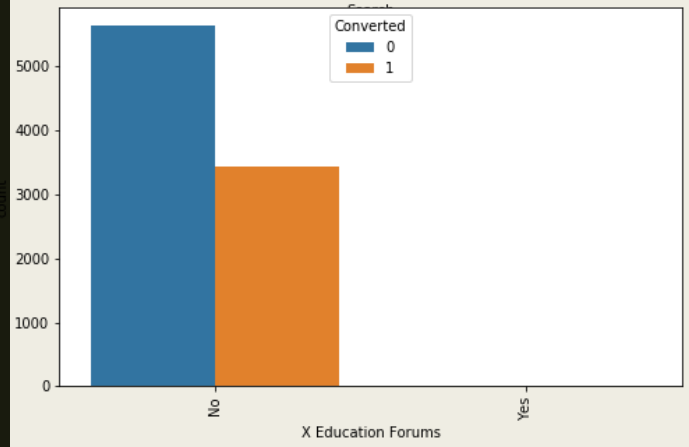

Converted
0
1
count
4000
3000
2000
1000
0
Low in Relevance
Not Sure
Might be
Worst
High in Relevance
Lead Quality

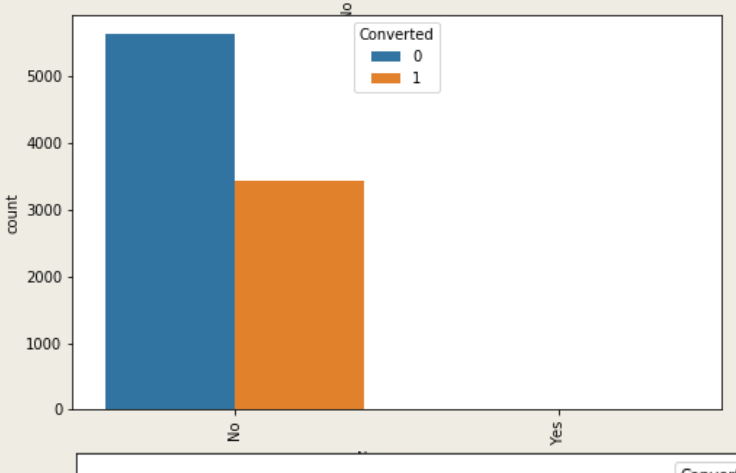
Converted
0
1
count
2500
2000
1500
1000
500
0
Interested in other courses
Ringing
Will revert after reading the email
Others
Lost to EINS
In confusion whether part time or DLP
Busy
switched off
in touch with EINS
Already a student
Diploma holder (Not Eligible)
Graduation in progress
number not provided
opp hangup
Closed by Horizzon
Not doing further education
invalid number
wrong number given
Interested in full time MBA
Still Thinking
Lost to Others
Shall take in the next coming month
Lateral student
Interested in Next batch
Recognition issue (DEC approval)
Want to take admission but has financial problems
University not recognized
Tags

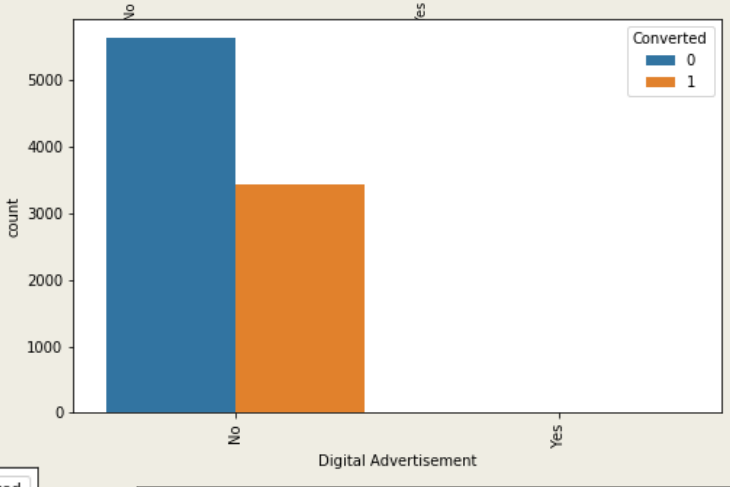

Converted
0
1
5000
4000
3000
2000
1000
0
No
Yes

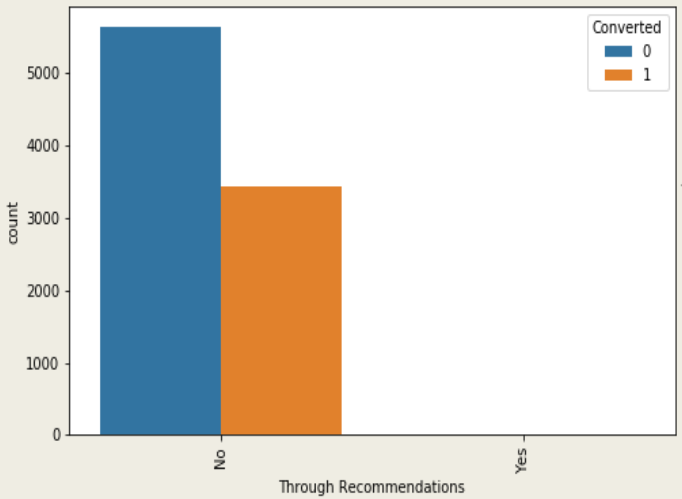

Converted
0
1
count
5000
4000
3000
2000
1000
0
No

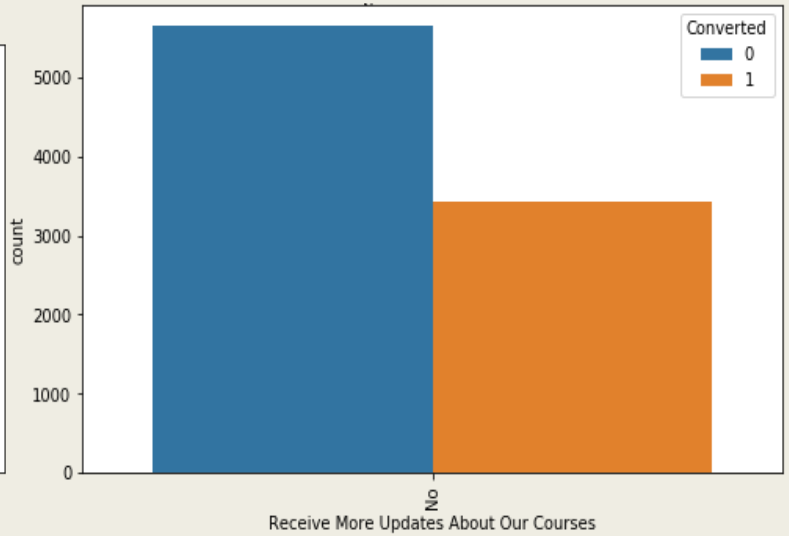

Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes

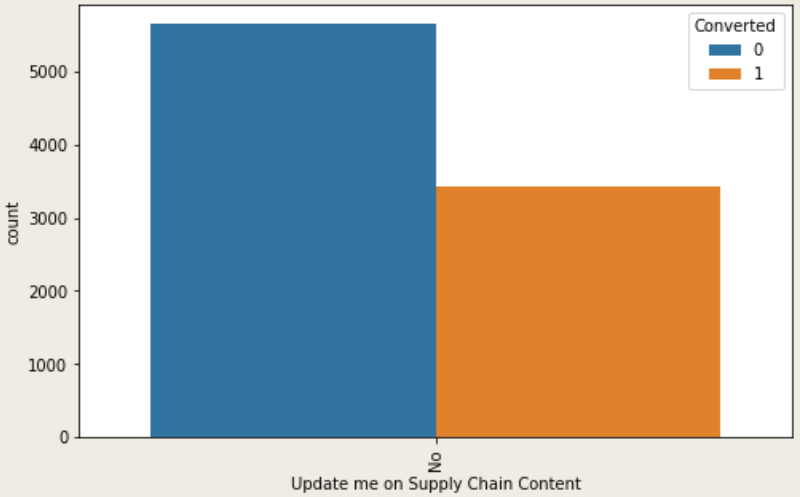

Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
X Education Forums


Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes


Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Digital Advertisement


Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Through Recommendations

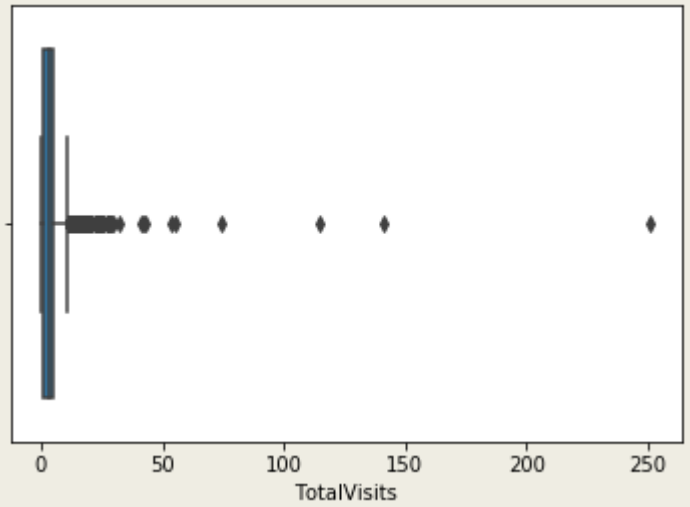

Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Receive More Updates About Our Courses

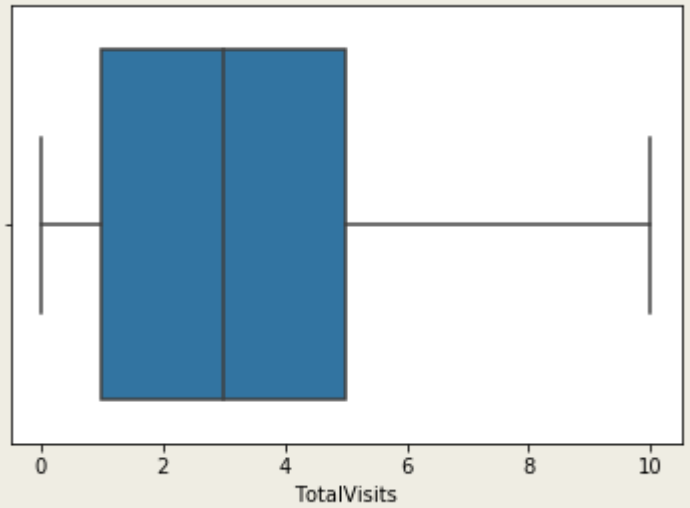

Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Update me on Supply Chain Content

Univariate Analysis

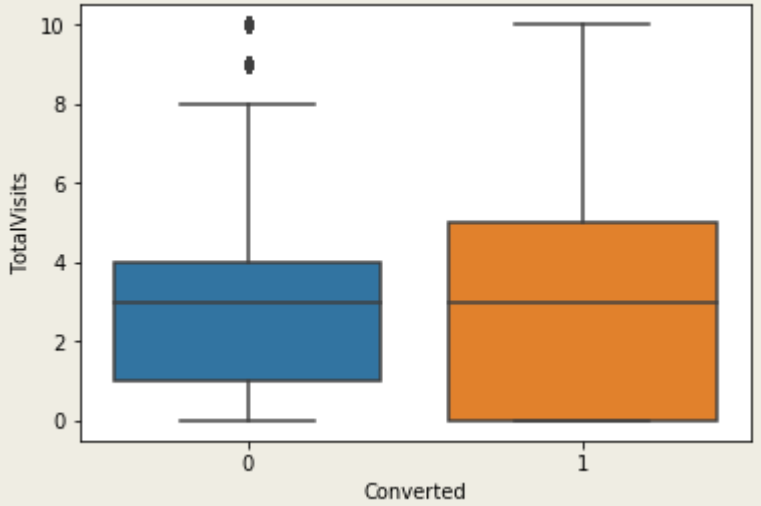
- Box plot for three numerical variables plotted
- Outlier Identified and treated
- Segmented box plots for converted variable

TOTAL VISITS

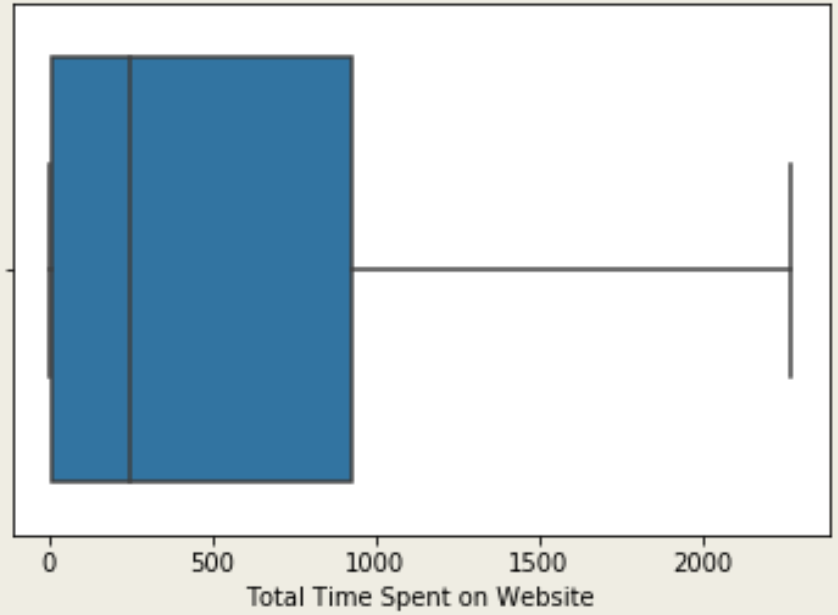


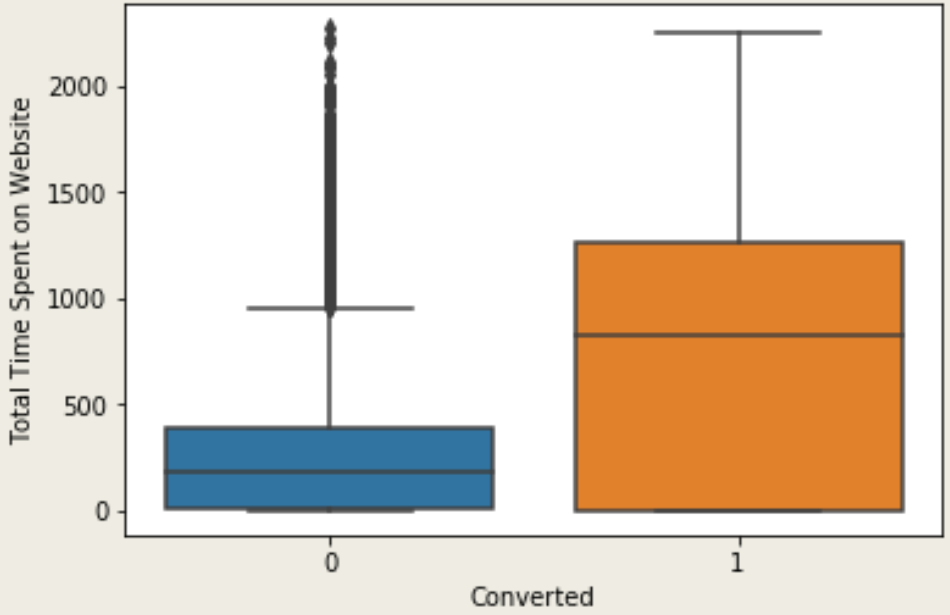


After Outlier Treatment

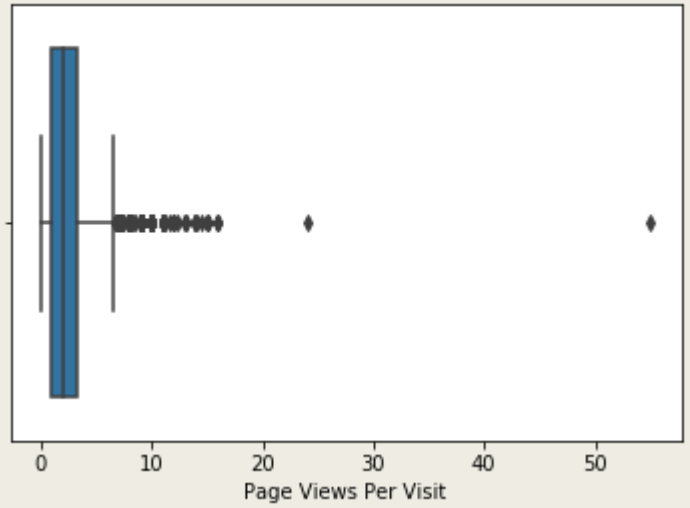


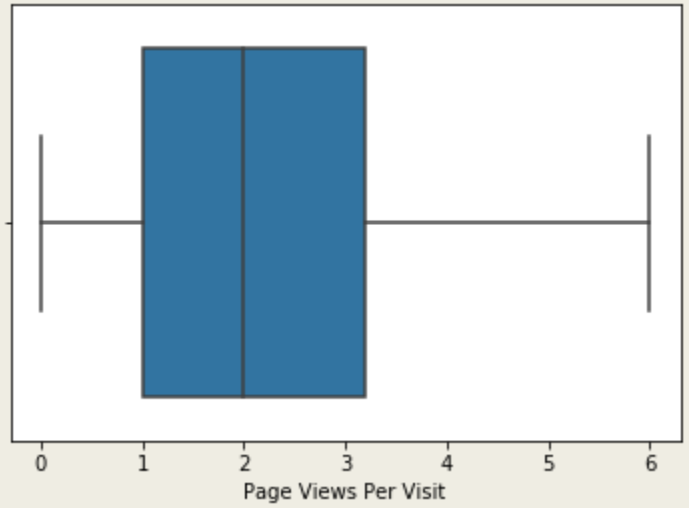
TOTAL TIME SPENT ON WEBSITE



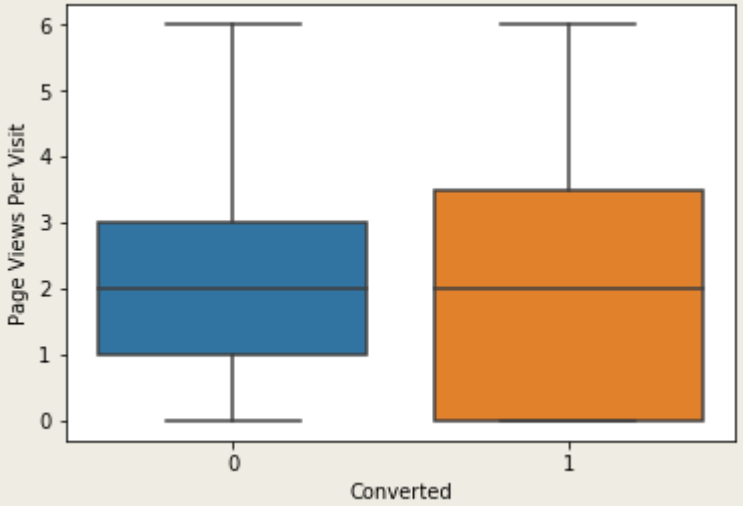


PAGE VIEWS PER VISIT

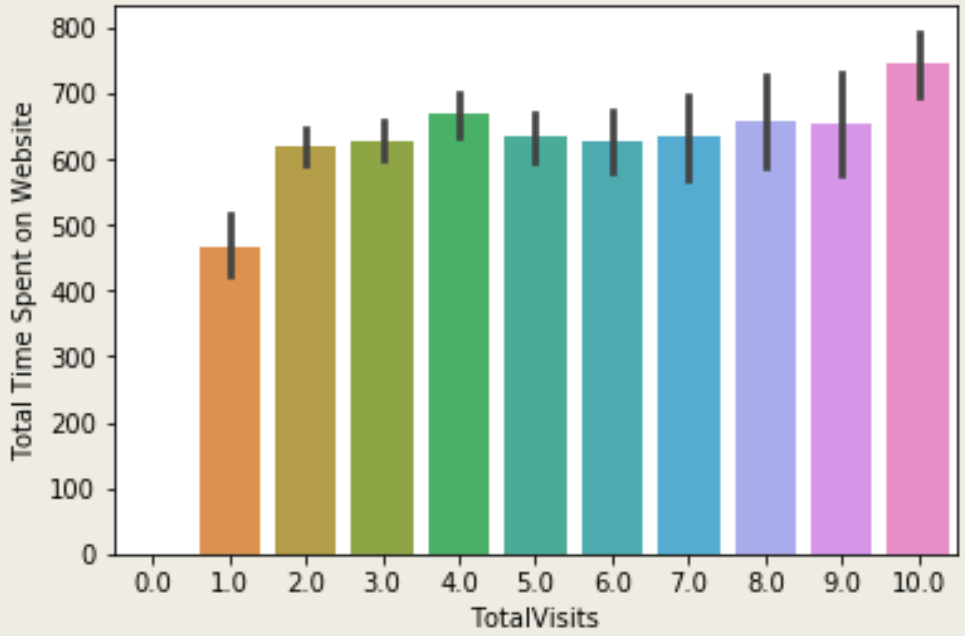




After Outlier Treatment



Bivariate Analysis



Business Insights

- To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.
- Google and Direct traffic generates higher number of leads.
- Conversion Rate of reference leads and leads through welingak website is high
- Recomended to generate more leads through Reference and Wellingak Website
- Focus on improving lead conversion through olark chat, organic search, direct traffic, and google

Business Insights contd..

- The median time spend on website for converted leads is more than that of not converted. Website should be made more engaging to make leads spend more time.
- Working Professionals going for the course have high chances of joining it
- Unemployed leads are the most in numbers and has a good conversion rate compared to others. Focus on unemployed category

MODEL BUILDING AND EVALUATION

MODEL BUILDING

- TEST-TRAIN SPLIT
- FEATURES SCALING using Standard Scalar
- Feature selection using RFE
- Run the model with RFE selected variables
- Models checked for p-values and VIF
- Final model arrived upon with reasonable VIF values and p-values
- Creating a dataframe with the actual conversion flag and the predicted probabilities
- Creating new column 'predicted' with 1 if Convert_Prob > 0.5 else 0

MODEL EVALUATION METRICS

- CONFUSION MATRIX

[[3741 164]

[324 2122]]

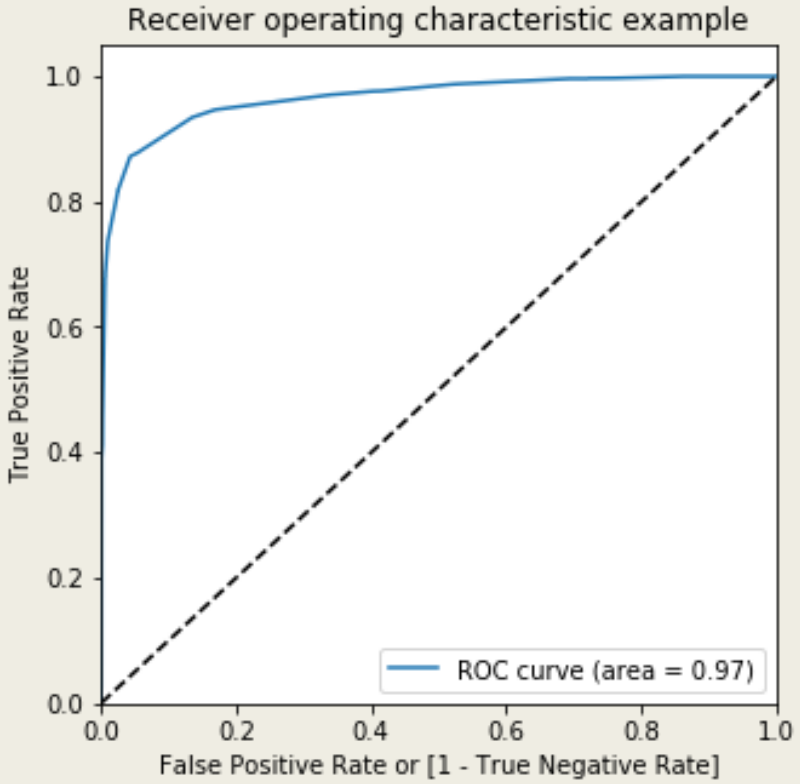
- Accuracy :- 0.923161706817824
- Sensitivity :- 0.867
- Specificity :- 0.958
- False postive rate:-0.0419
- Positive predictive value:-0.928
- Negative predictive value:-0.920

Plotting the ROC Curve

An ROC curve demonstrates following things:

- - It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- - The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- - The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

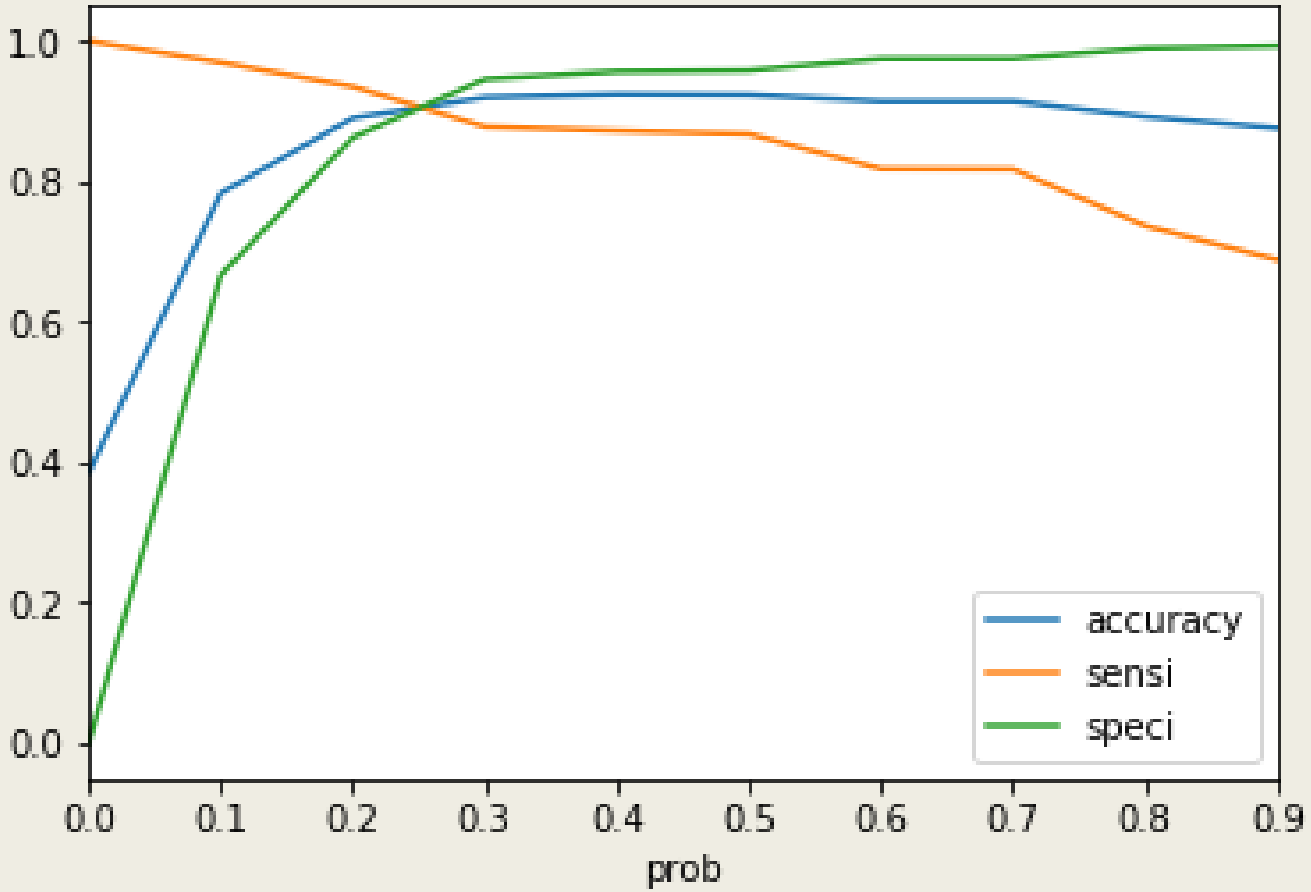
ROC CURVE


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.97)

Finding Optimal Cutoff Point

- Optimal cutoff probability is that prob where we get balanced sensitivity and specificity
- plot accuracy sensitivity and specificity for various probabilities.
- Optimal cut off obtained at probability of 0.3

Plot accuracy sensitivity and specificity for various probabilities.



MODEL EVALUATION contd..

- Recalculating column 'predicted' with 1 if Convert_Prob > 0.3 else 0
- Assigning Lead Score for each leads in a scale of 1-100
- Recalculating model evaluation metrics
- Confusion Matrix and Accuracy
- Precision and Recall Trade off
- Making predictions on the test set
- Evaluating metrics on Test predictions

MODEL EVALUATION METRICS for Cutoff of 0.3

- CONFUSION MATRIX

[[3693 212]

[299 2147]]

- Accuracy :- 0.91954
- Sensitivity :- 0.877
- Specificity :- 0.9457
- False postive rate:-0.05
- Positive predictive value:-0.910
- Negative predictive value:-0.925

PRECISON AND RECALL

- Precision :- 0.9101

TP / TP + FP

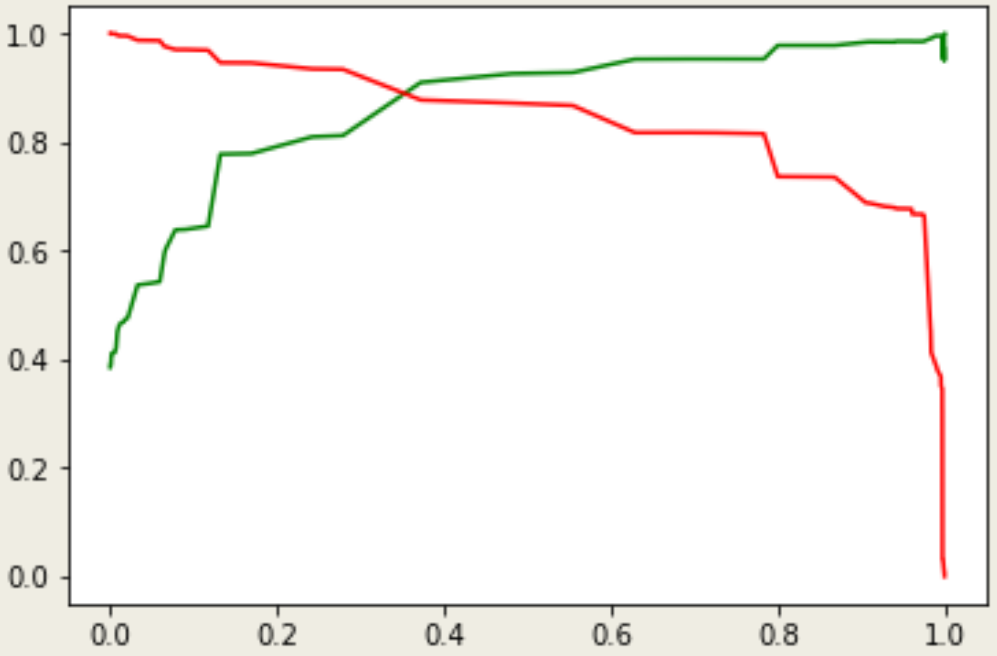
- Recall :-0.8777

TP / TP + FN

- F1-score :- 0.8936

2 * (Precision * Recall)/(Precision + Recall)

Precision and recall tradeoff



TEST DATA PREDICTION and METRICS

- CONFUSION MATRIX

[[1635 99]

 [138 851]]

- Accuracy :- 0.9129
- Sensitivity :- 0.860
- Specificity :- 0.942
- Precision:-0.895
- Recall:-0.860
- F1 score :- 0.8777

ASSIGNING LEAD SCORE TO ORIGINAL CLEANED DATSET

- ORIGINAL CLEANED DATSET and Final Predicted Lead score combined to identify the HOT LEADS SUBSET
- Leads with Leads Score Greater than 30 are considered as HOT LEADS
- Further Analysis could be done on the subset to identify most promising conversions

FINAL ANALYSIS and INSIGHTS

- Analysis could be done on the Hot leads subset to arrive at insights for increasing conversion rate
- Have to prioritize on Hot leads(Lead Score >30) with following attributes:-

1. Tags_Lost to EINS
2. Tags_Closed by Horizzon
3. Tags_Will revert after reading the email
4. Lead Source_Welingak Website
5. Last Activity_SMS Sent

THANK YOU