LEAD SCORE CASE STUDY

Indranil Kundu
Divya Dindorkar

Problem Statement

- X Education sells online courses to industry professionals.
- 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 Objective:

- 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.

SOLUTION APPROACH

- Data cleaning and Data Manipulation:

  1. Check and handle missing values.

  2. Drop columns, if it contains large number of missing values and not useful for analysis.

  3. Imputation of the values, if necessary.

  4. Check and handle outliers in data.

- EDA:

  1. Univariate Data Analysis: using value counts, checking distribution of variable etc.

- Feature Scaling and creation of dummy variables for the data.
- Classification technique: Logistic Regression was used for model building and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

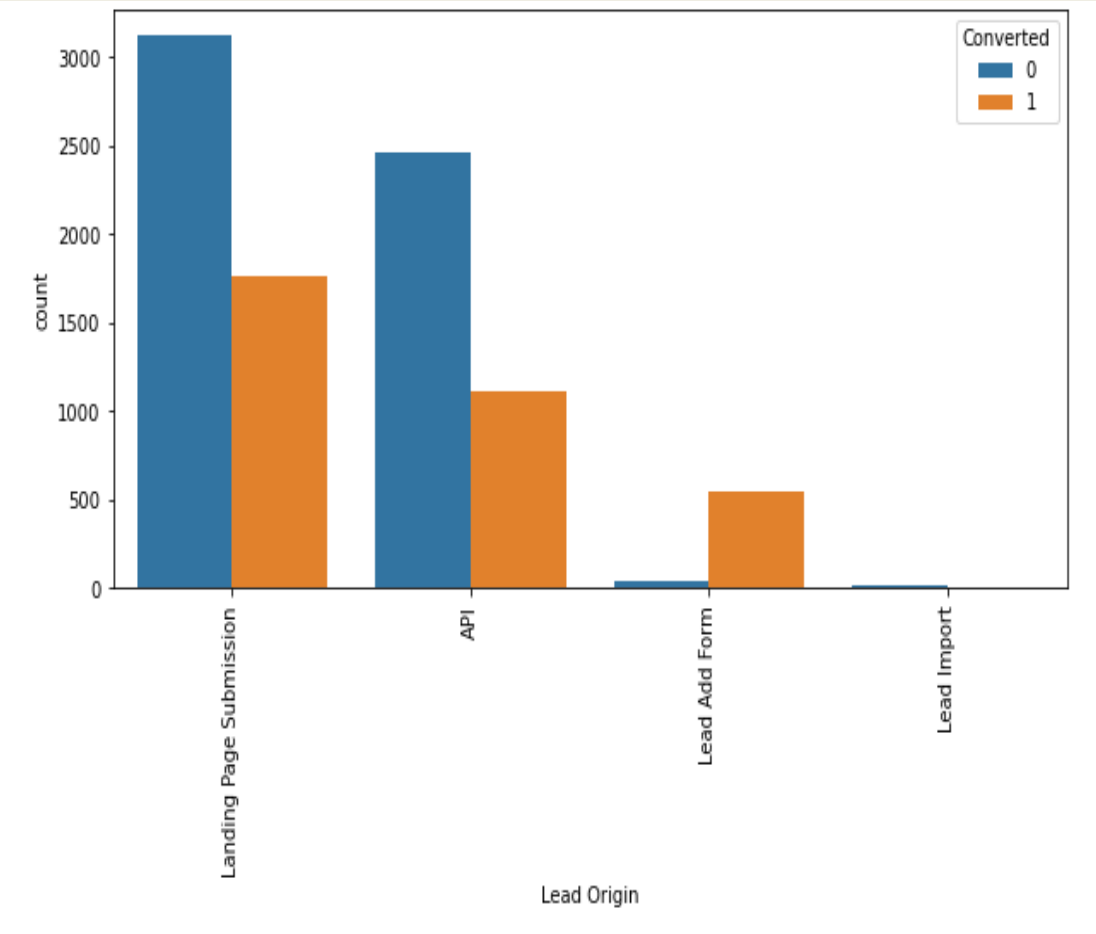
Data Cleaning and Data Manipulation

- There were total of 9240 rows and 37 columns in original dataset.
- 'Prospect ID', 'Lead Number‘ columns were dropped as they had distinct and unique values which is not needed for analysis.
- All the columns with more than 40% of missing data present in it were dropped.
- Sales team generated columns such as ‘Tags’, ‘Lead Quality’, ‘Last Activity’, ‘Last Notable Activity’, ‘Lead Profile’, ‘Asymmetrique Activity Index’, ‘Asymmetrique Profile Index’, ‘Asymmetrique Activity Score’ and ‘Asymmetrique Profile Score’ were dropped.
- After removing the columns with more than 40% missing values, few columns still had large number of missing values. For those, we went through the columns one by one, found out the maximum value for that specific column and imputed the null values with that value. Some null values were changed to ‘Others’ for better understanding of the data.
- Later only the rows with missing values less than 2% were deleted.

EDA

UNIVARIATE ANALYSIS

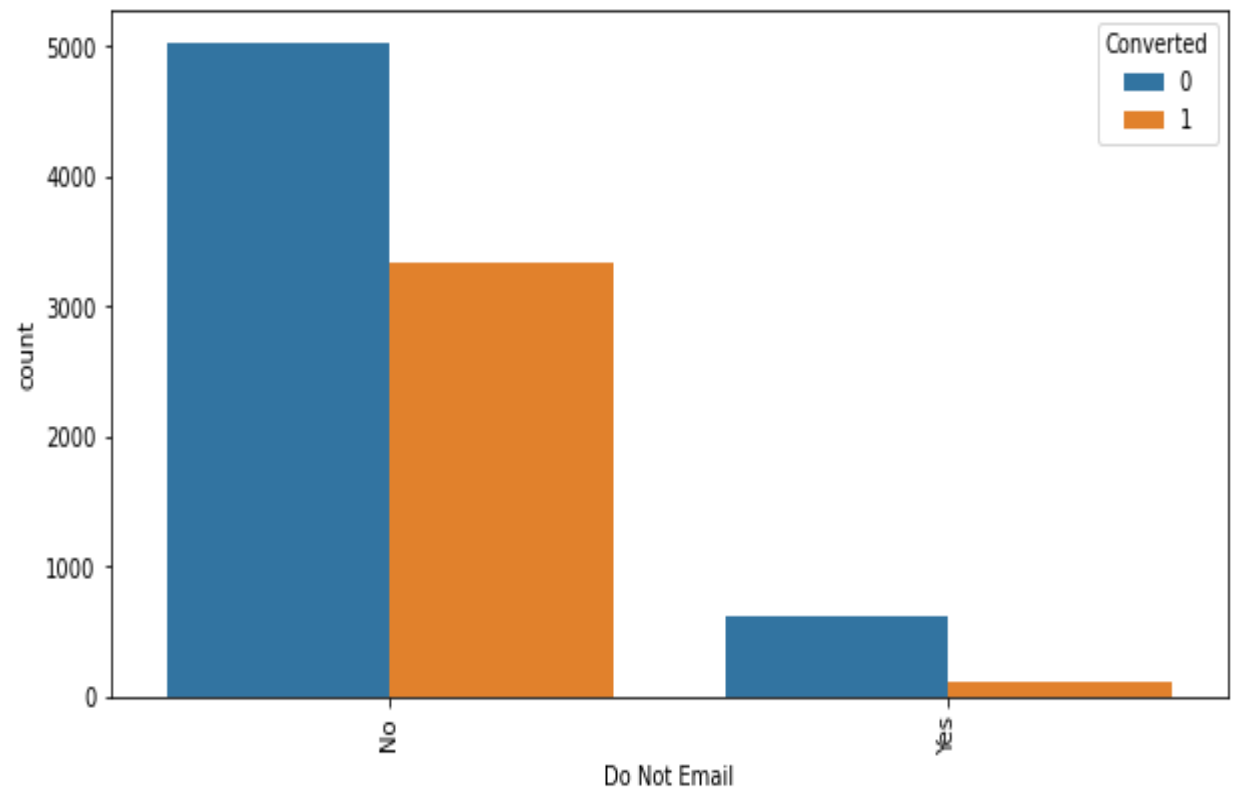
Analysis of ‘Lead Origin’ Column



Inference:

- 'API' and 'Landing Page Submission' generate maximum number of leads.
- ‘Lead Add Form' has a great conversion rate but the count of lead is not very high.
- 'Lead Import' is very less in count.
- To improve conversion rate, company should focus on improving lead conversion of 'API', 'Landing Page Submission' and also generate more leads from 'Lead Add Form' and 'Lead Import'.

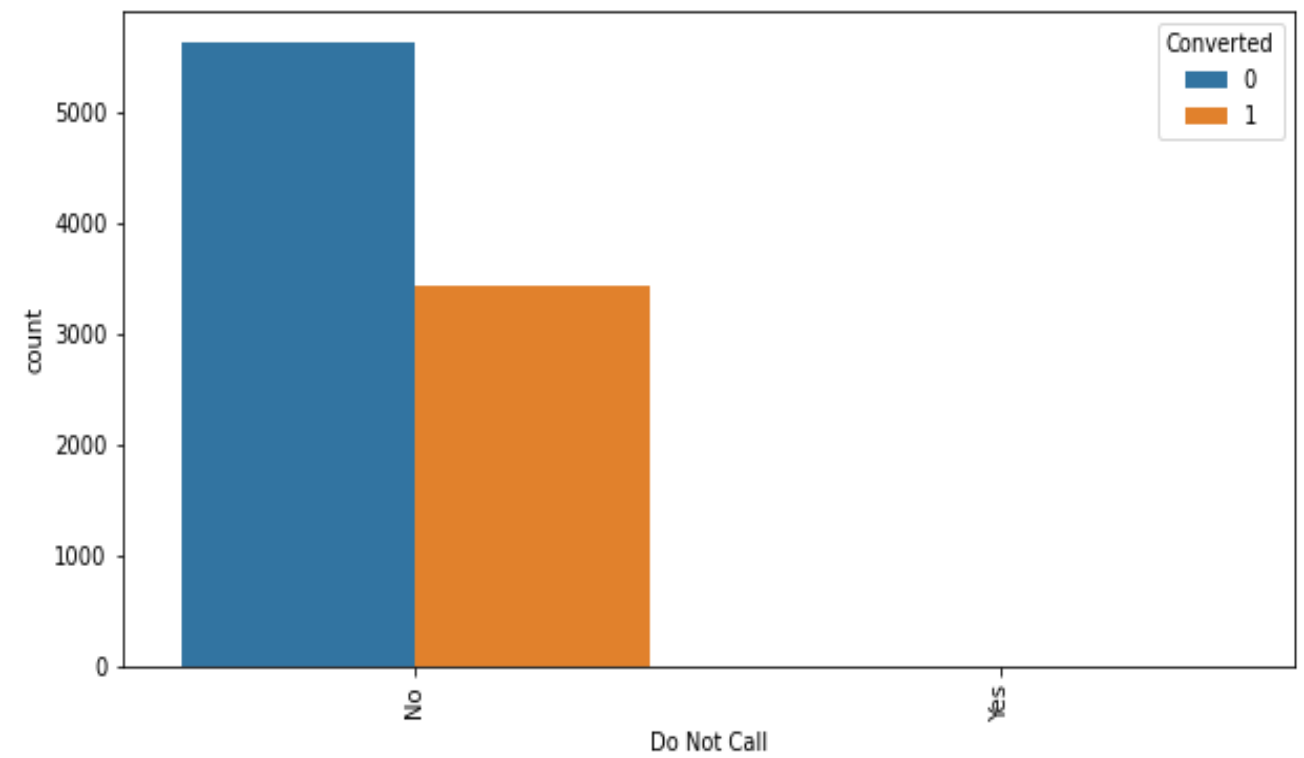
Analysis of ‘Do Not Email’ Column



Inference:

- Customers to whom the course details were emailed had a high chance of converting than the customers to whom the course details were not emailed.
- To improve conversion rate, the company should send the course details as an email to the customers.

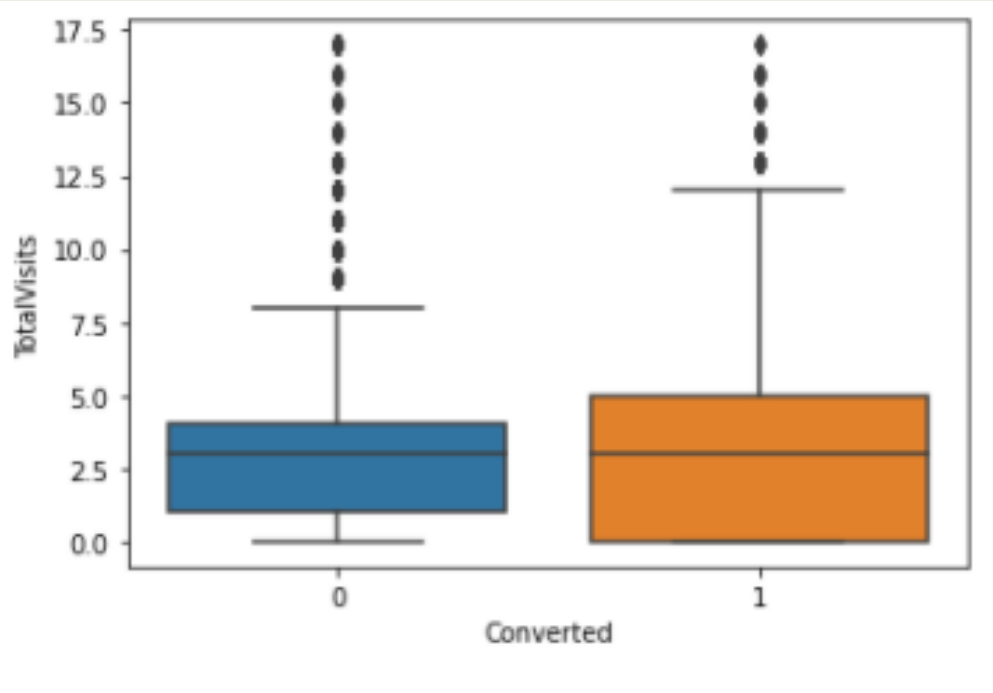
Analysis of ‘Do Not Call’ Column



Inference:

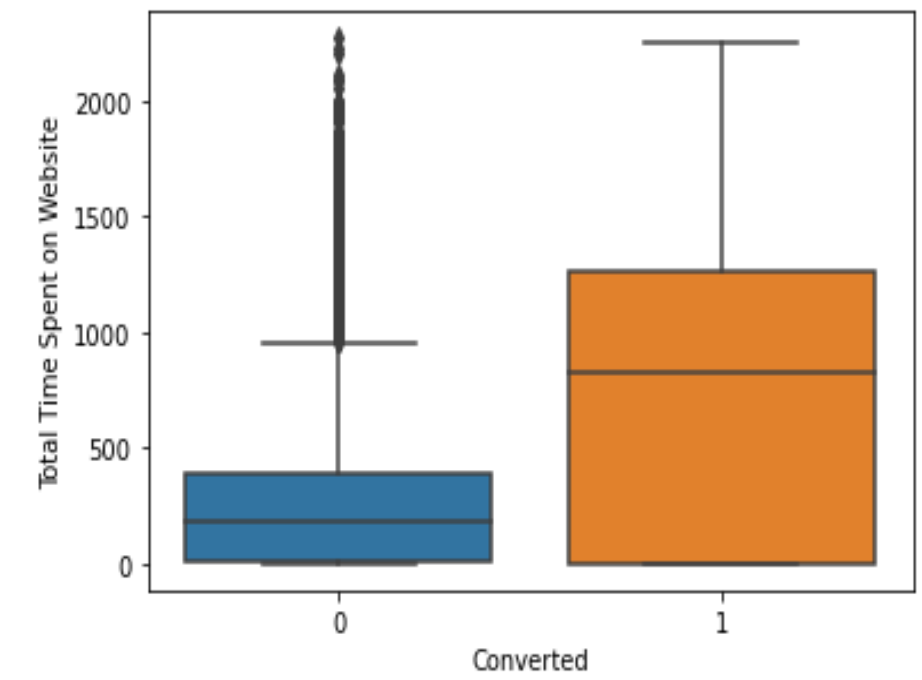
- Similar trend is also seen here. Customers who had received a call about the course curriculum had a lot more chance of converting.
- To improve conversion rate, the company should call the customers and explain the course curriculum.

Analysis of ‘TotalVisits’ Column



Inference:
Median is same for Converted and Non-converted.
Nothing conclusive can be said.

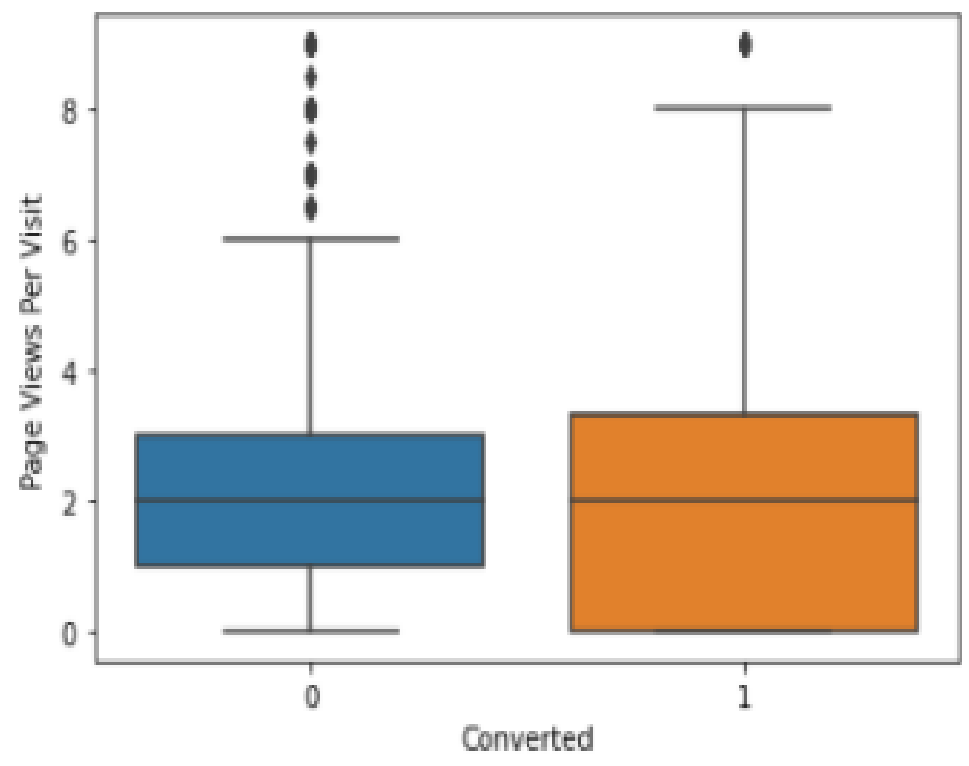
Analysis of ‘Total Time Spent on Website’ Column



Inference:

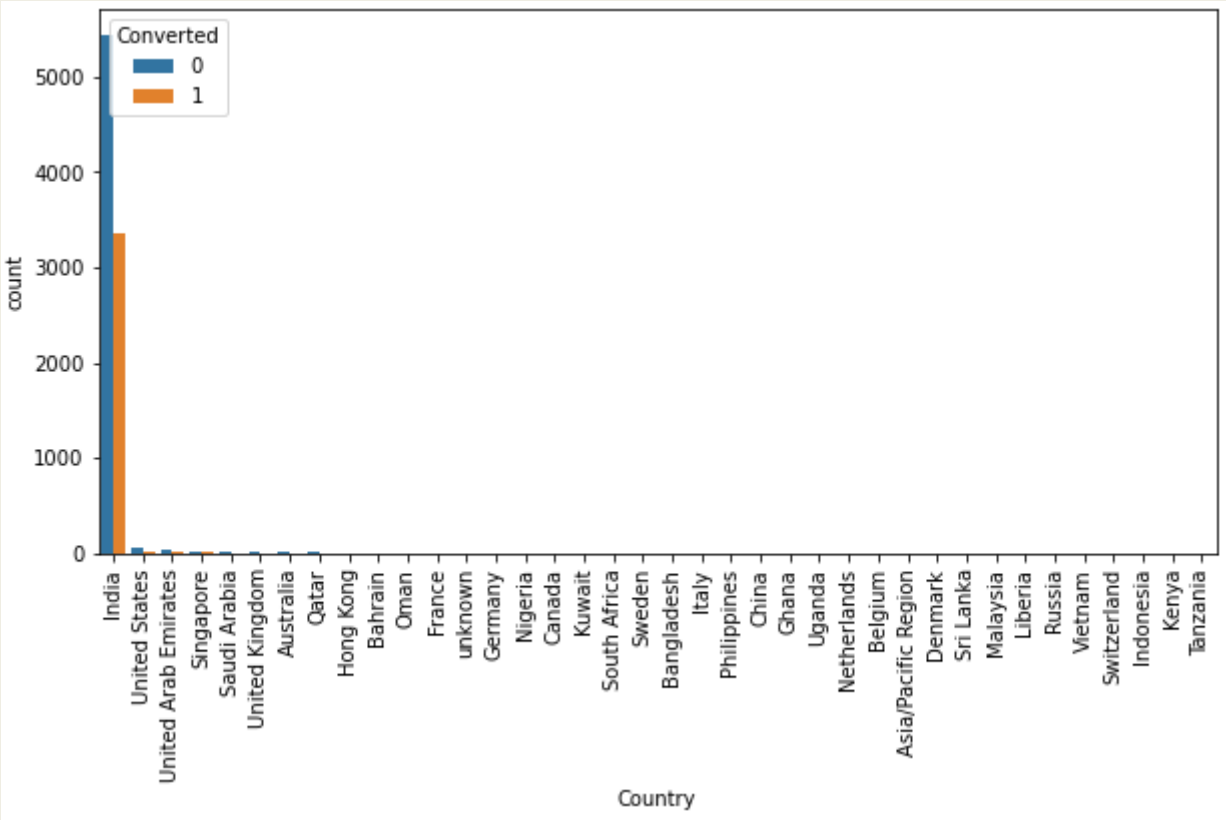
- Customers spending more time on the website are likely to be converted.
- Companies should thus try and make the websites more engaging so that the customers spend more time.

Analysis of ‘Page Views Per Visit’ Column



Inference:
Median is same for Converted and Non-converted. Nothing conclusive can be said.

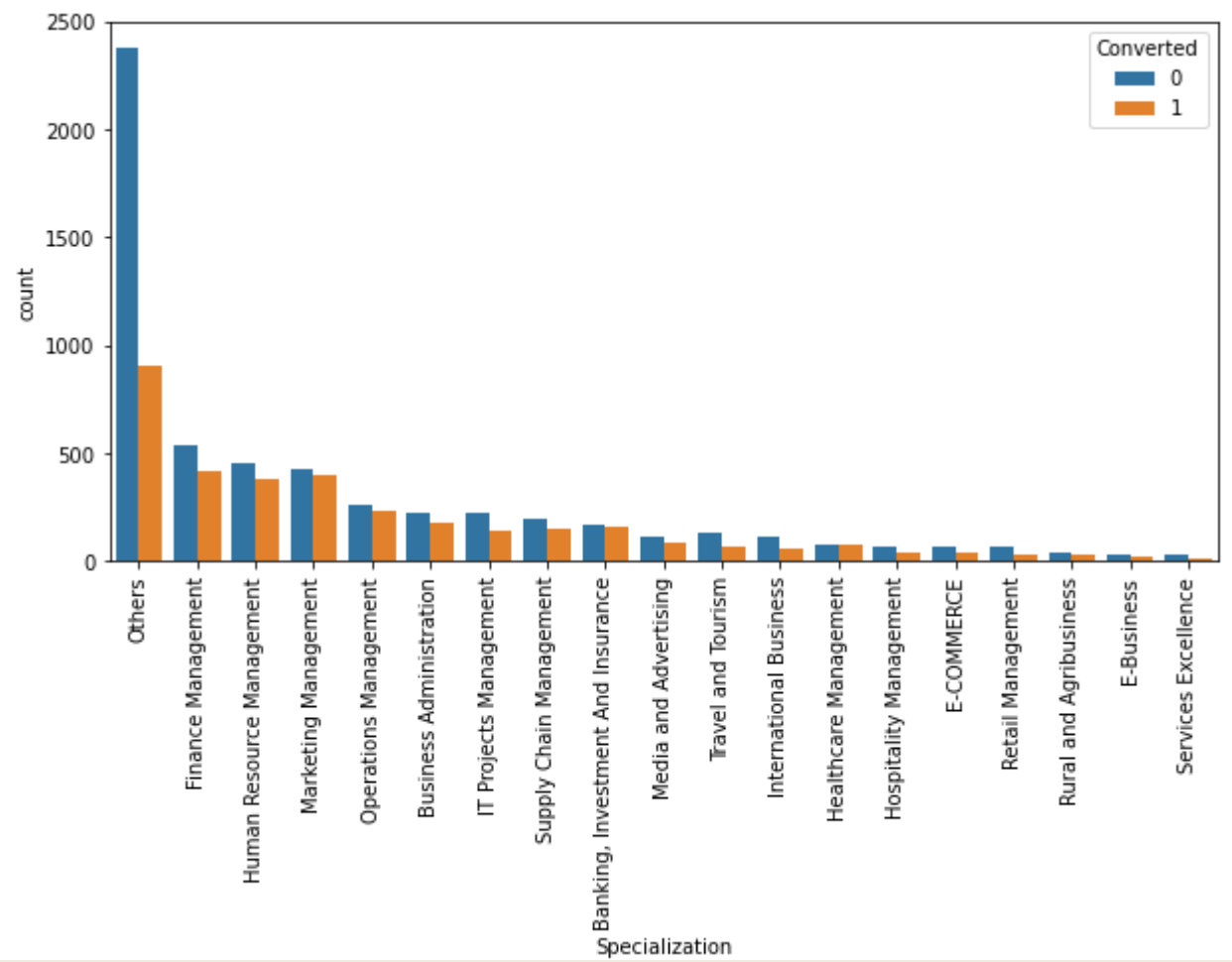
Analysis of ‘Country’ Column



Inference:

Most values are from 'India'.

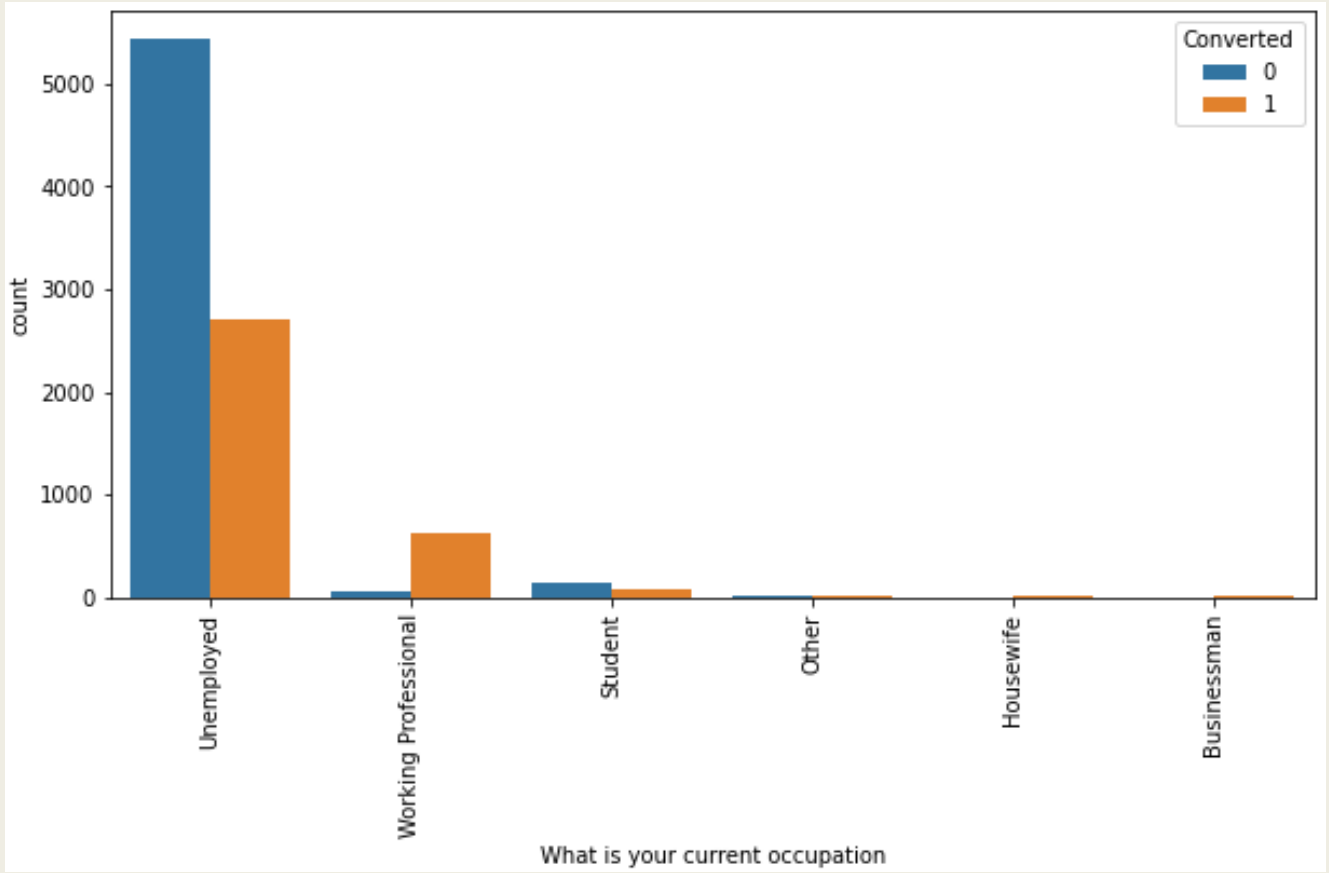
Analysis of ‘Specialization’ Column



Inference:

- Specializations such as 'Finance Management', 'HR Management', 'Marketing Management', 'Operations Management', 'Business Administration' etc have higher conversion rate.
- Companies should thus focus on specializations which have higher conversion rate.

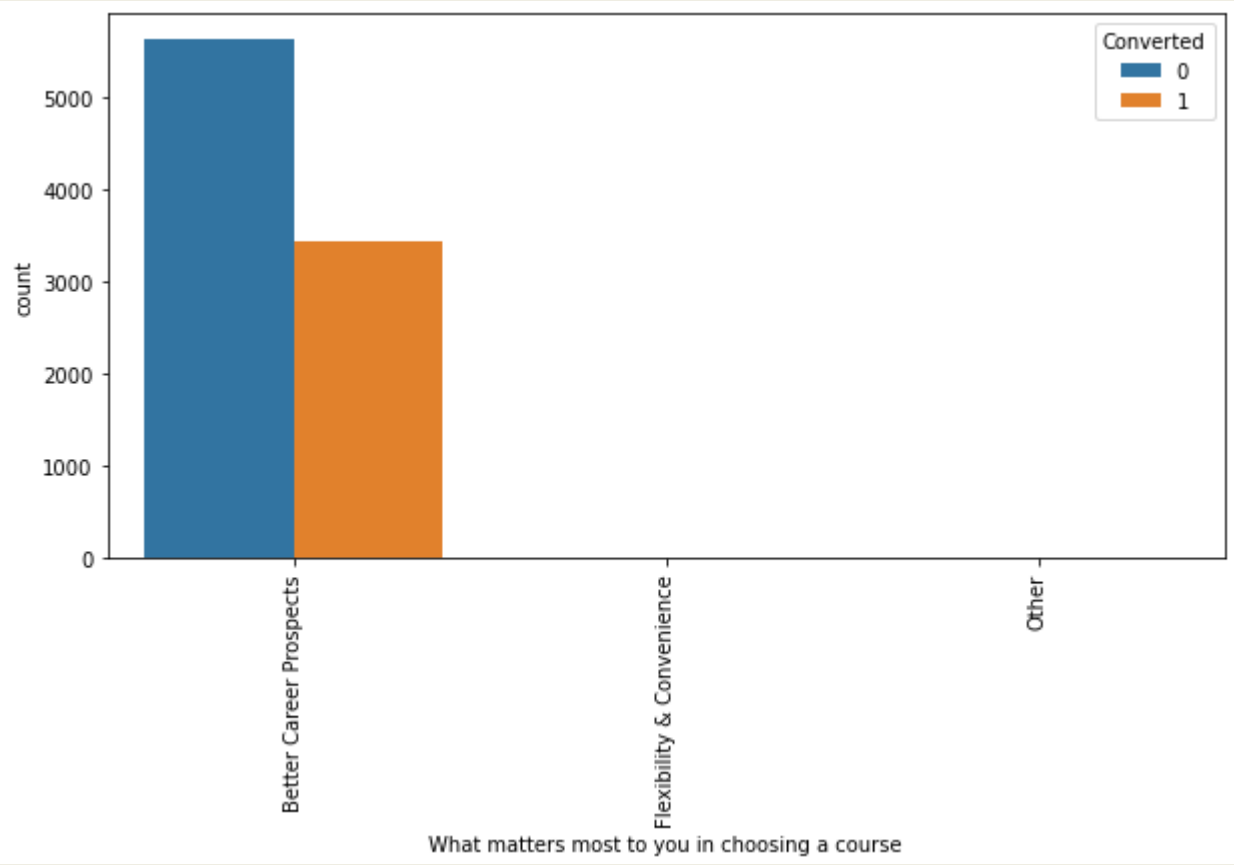
Analysis of ‘What is your current occupation’ Column



Inference:

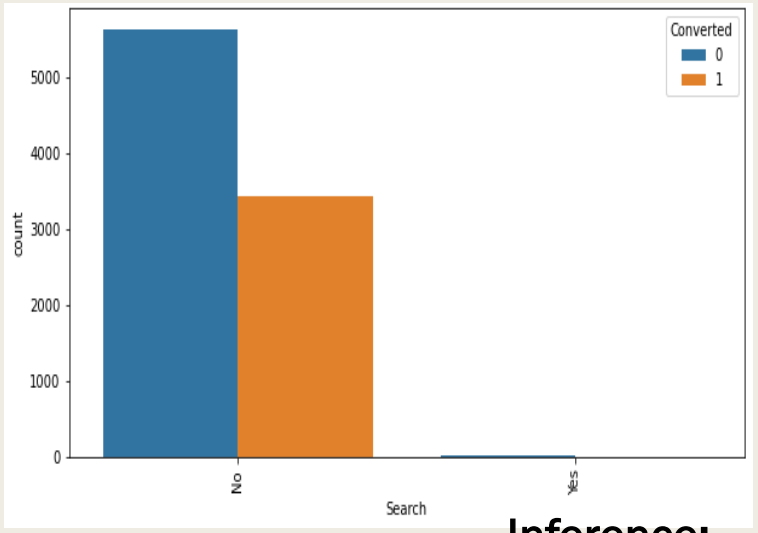
- 'Unemployed' people are more in numbers, but have poor conversion rate.
- 'Working Professional' people opting for the course have a higher chance of converting.
- Company should thus focus more on 'Working Professional' and 'Unemployed' people.

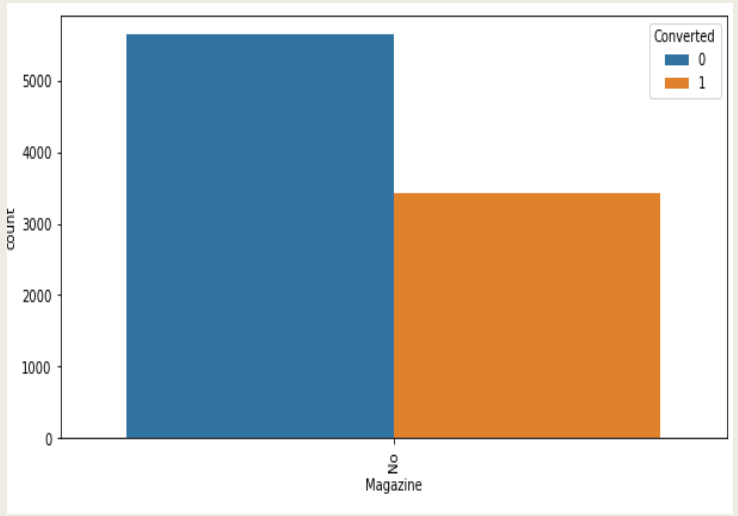
Analysis of ‘What matters most to you in choosing a course’ Column

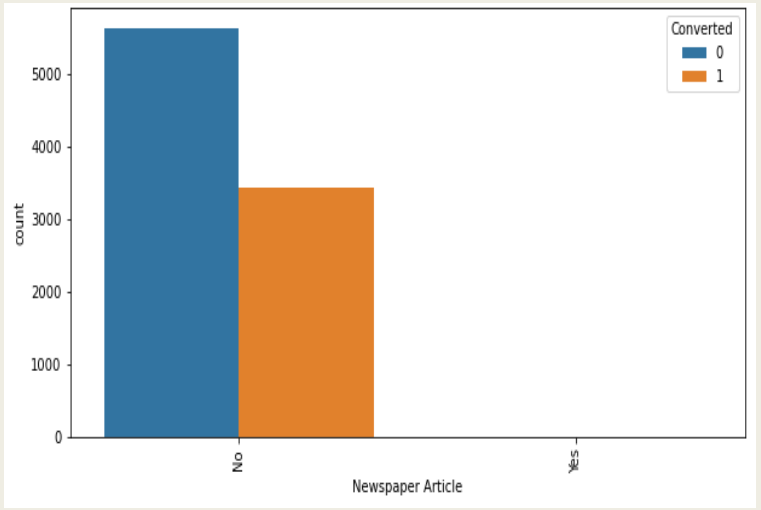


Inference:
Everyone wants to enroll for the course in order to have 'Better Career Prospects'.

Analysis of ‘Search’, ’Magazine’ ,’X Education Forums’, ‘Newspaper’ , ‘ Digital Advertisement’ and ‘Newspaper Article’ Column

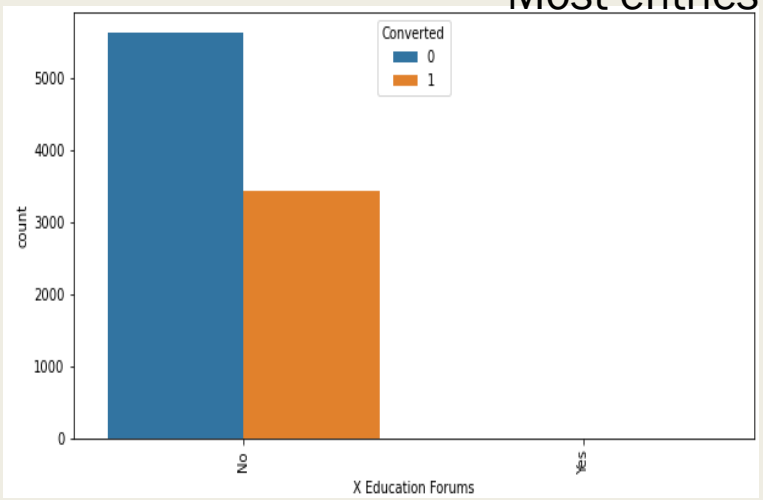


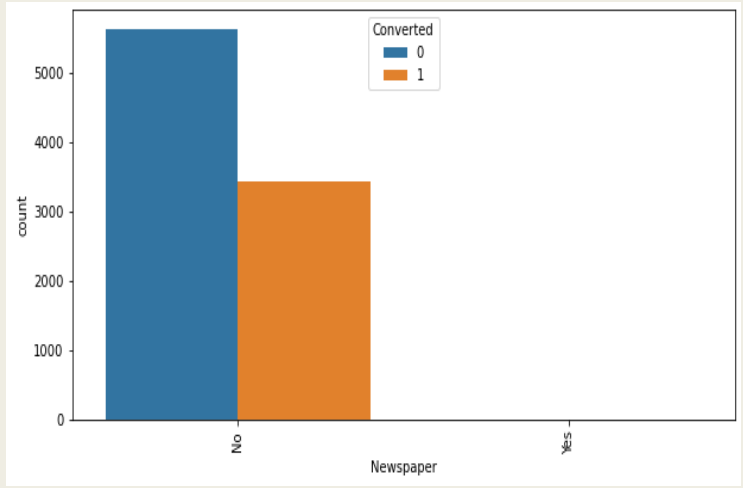


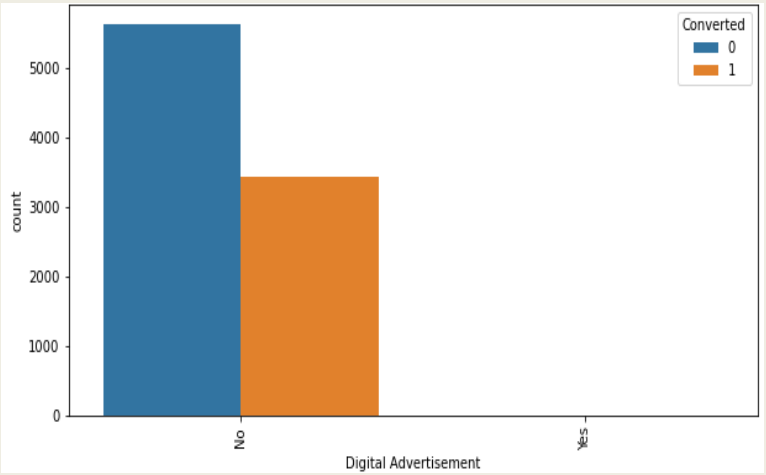


Inference:

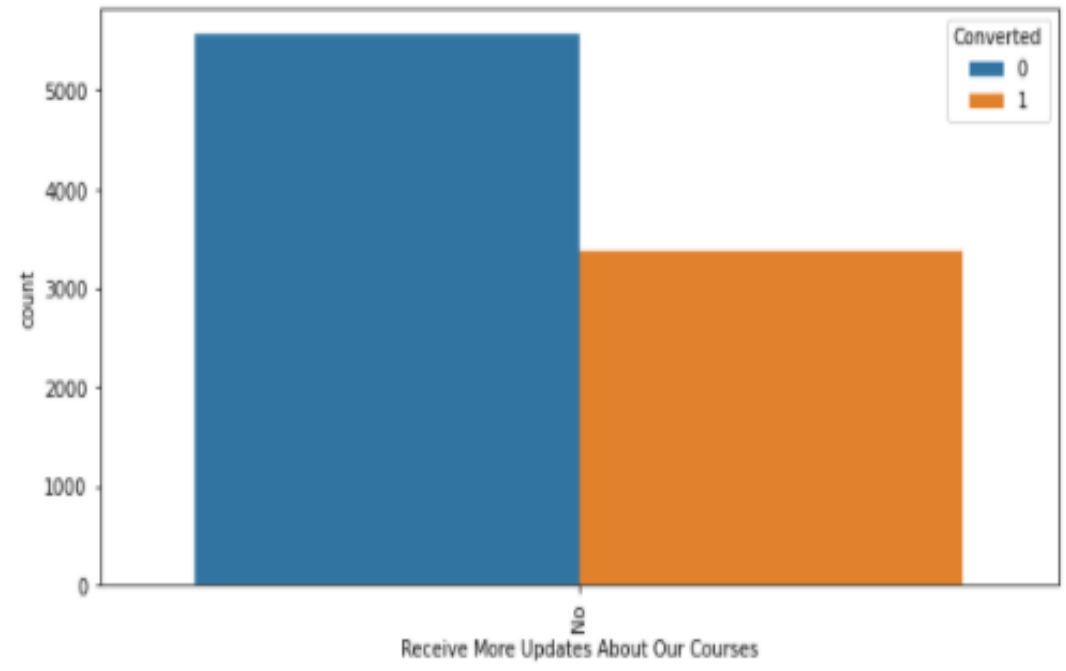
Most entries are 'No'. No conclusive inference can be drawn.





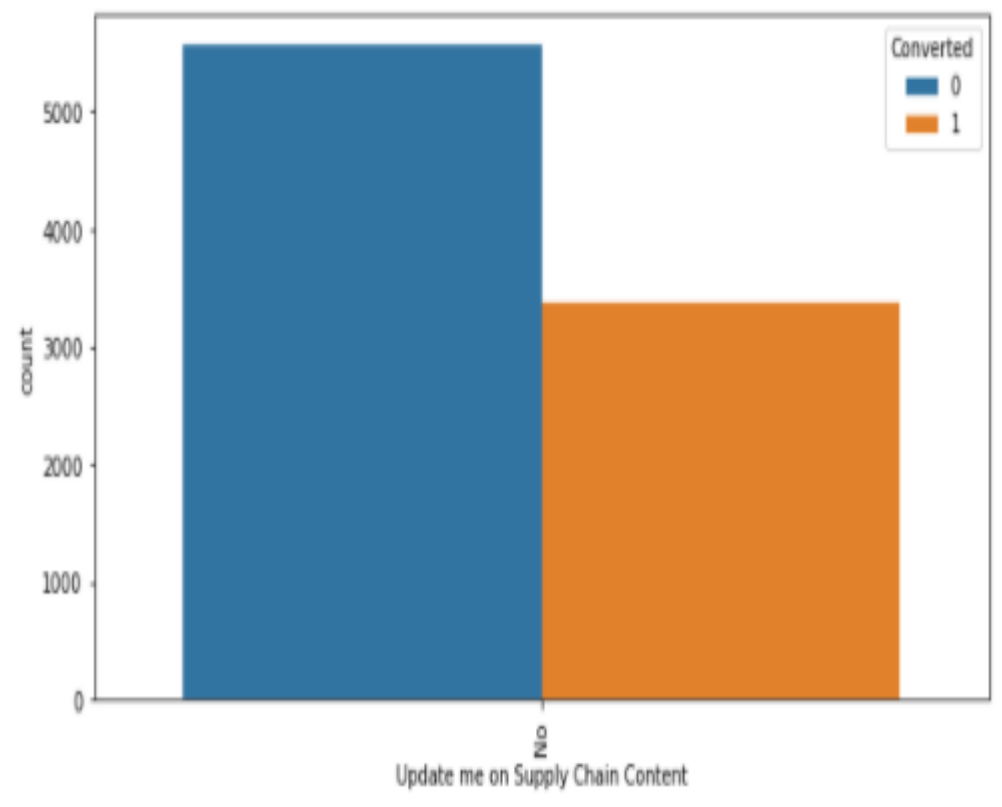


Analysis of ‘Receive More Updates About Our Courses’ Column



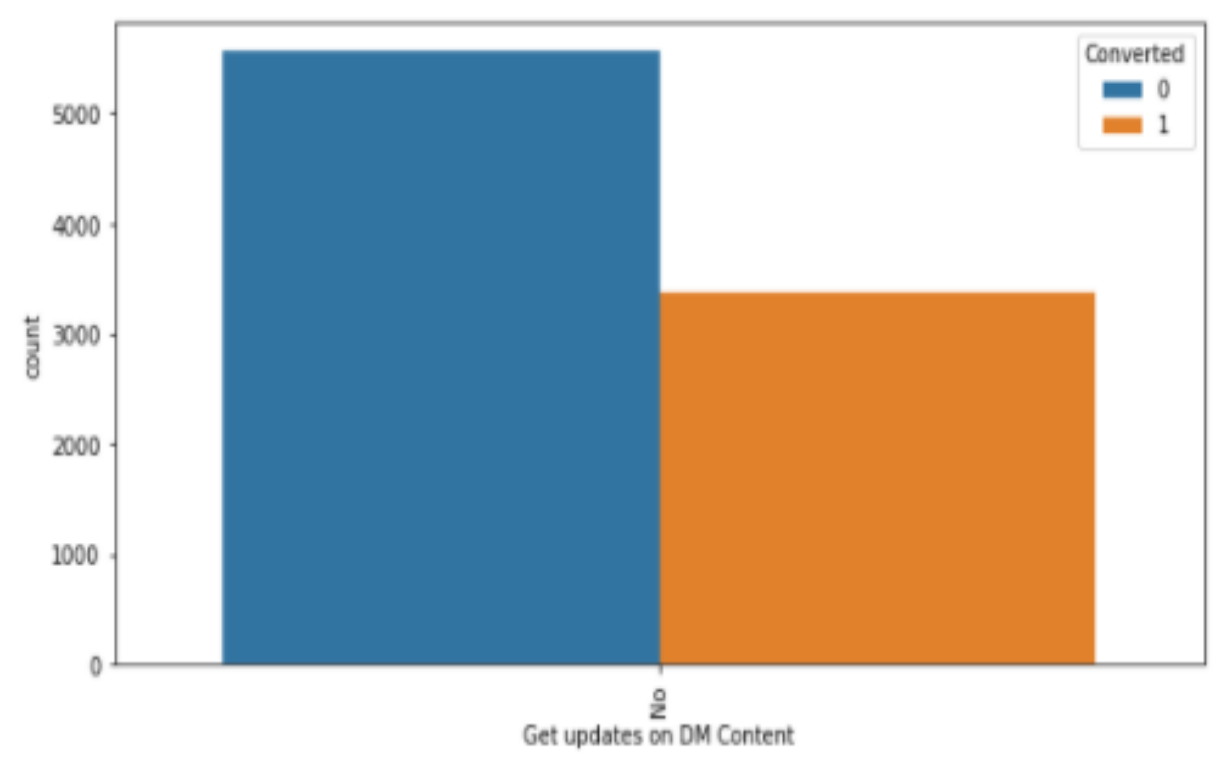
Inference:
Most entries are 'No'. No conclusive inference can be drawn.

Analysis of ‘Update me on Supply Chain Content’ Column



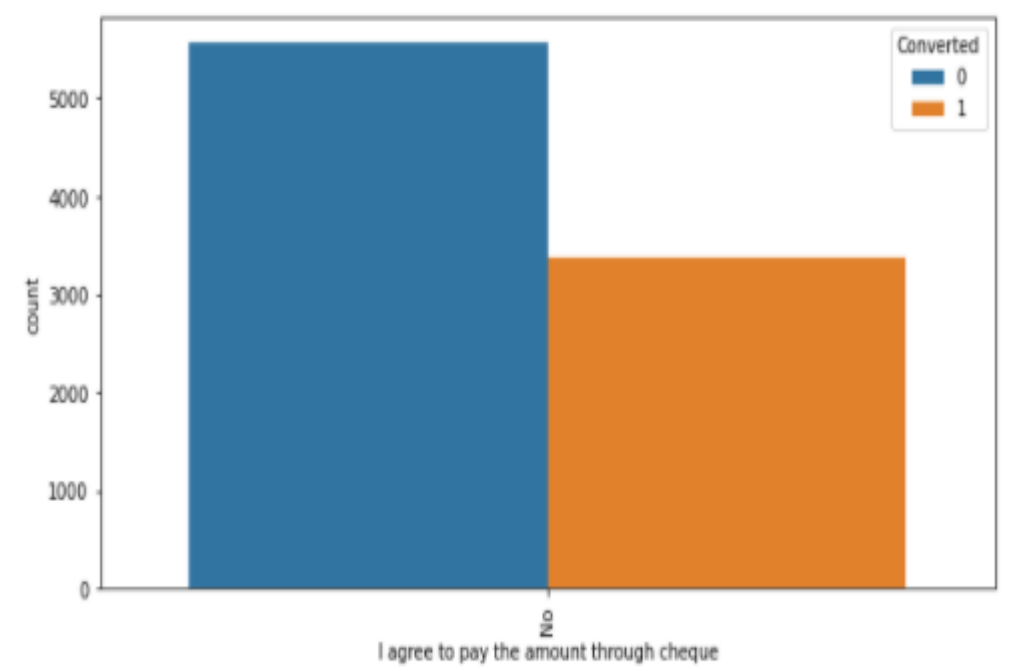
Inference:
Most entries are 'No'. No conclusive inference can be drawn.

Analysis of 'Get updates on DM Content‘ Column



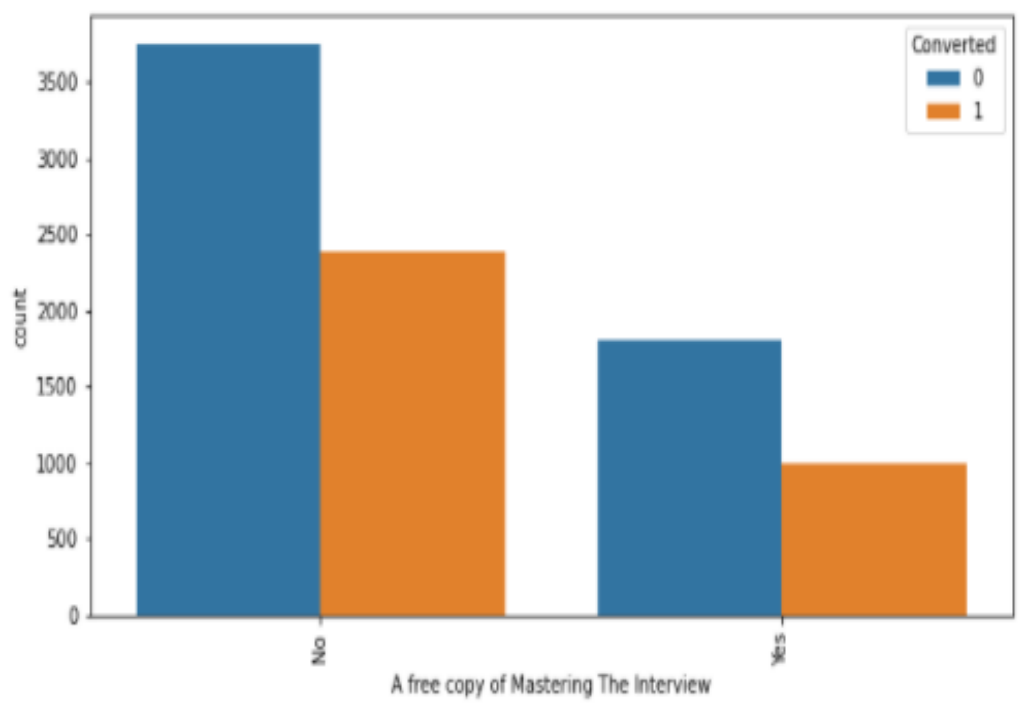
Inference:
Most entries are 'No'. No conclusive inference can be drawn.

Analysis of 'I agree to pay the amount through cheque‘ Column



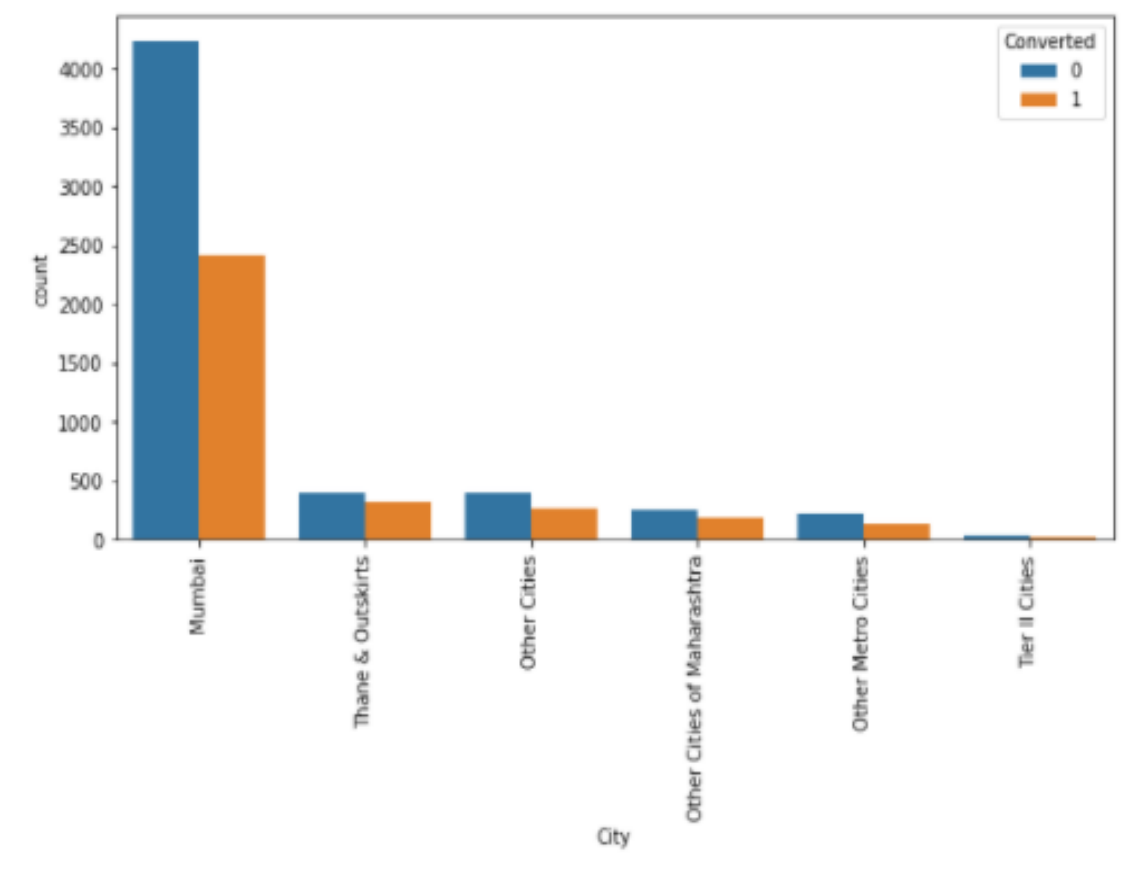
Inference:
Most entries are 'No'. No conclusive inference can be drawn.

Analysis of ‘A free copy of Mastering The Interview’ Column



Inference:
Most entries are 'No'. No conclusive inference can be drawn.

Analysis of ‘City’ Column



Inference:

Most of the customers belong to Mumbai

MODEL BUILDING

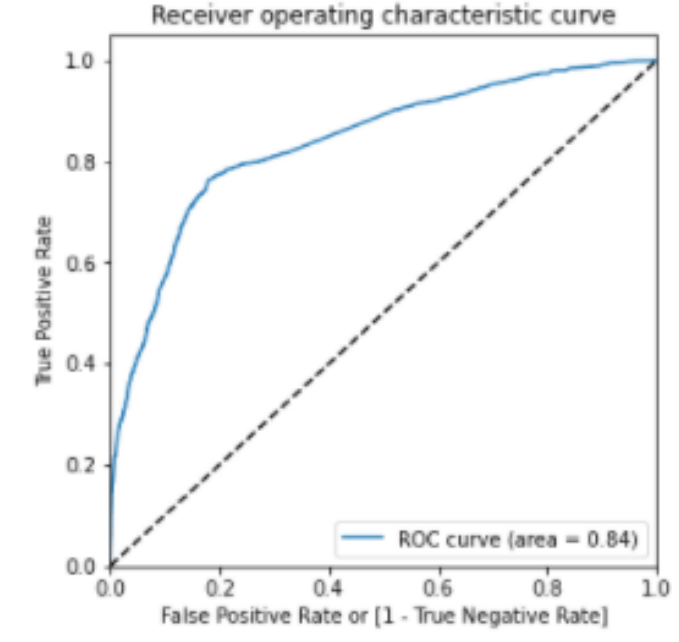
Dep. Variable:	Converted	No. Observations:	6246
Model:	GLM	Df Residuals:	6235
Model Family:	Binomial	Df Model:	10
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2949.5
Date:	Sun, 16 May 2021	Deviance:	5898.9
Time:	21:41:20	Pearson chi2:	6.49e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

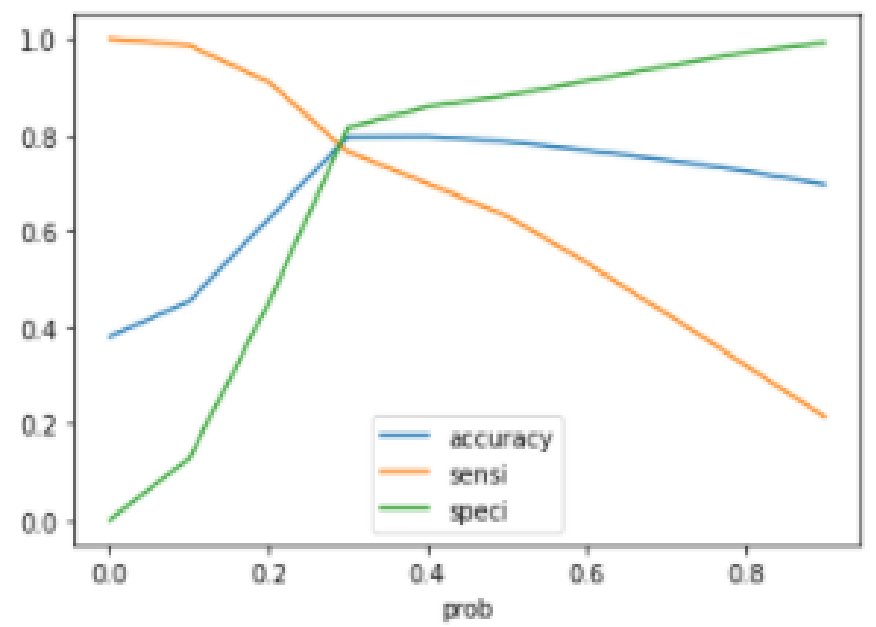
	coef	std err	z	P>|z|	[0.025	0.975]
const	1.7483	0.173	10.113	0.000	1.409	2.087
Do Not Email	-1.5081	0.166	-9.099	0.000	-1.833	-1.183
Total Time Spent on Website	1.1152	0.038	29.338	0.000	1.041	1.190
Lead Source_Olark Chat	1.0556	0.104	10.105	0.000	0.851	1.260
Lead Source_Reference	4.2085	0.231	18.208	0.000	3.755	4.662
Lead Source_Welingak Website	6.2591	0.728	8.598	0.000	4.832	7.686
Specialization_Hospitality Management	-0.9059	0.321	-2.823	0.005	-1.535	-0.277
Specialization_Others	-0.5108	0.081	-6.285	0.000	-0.670	-0.352
What is your current occupation_Other	-2.4266	0.709	-3.425	0.001	-3.815	-1.038
What is your current occupation_Student	-2.6872	0.276	-9.730	0.000	-3.228	-2.146
What is your current occupation_Unemployed	-2.6429	0.177	-14.963	0.000	-2.989	-2.297

- Splitting the Data into Train and Test datasets.
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection.
- Ran RFE with 15 variables as output.
- Build Logistic Regression model using the variables chosen by RFE and then manually removed variables whose p-value was greater than 0.05 and VIF value was greater than 5.
- Predictions was then made on test data set using the model built.

	Features	VIF
6	Specialization_Others	2.33
2	Lead Source_Olark Chat	1.92
9	What is your current occupation_Unemployed	1.90
1	Total Time Spent on Website	1.24
0	Do Not Email	1.09
3	Lead Source_Reference	1.08
4	Lead Source_Welingak Website	1.06
8	What is your current occupation_Student	1.03
5	Specialization_Hospitality Management	1.02
7	What is your current occupation_Other	1.00

ROC Curve





Finding Optimal Cut off Point

Optimal cut off probability is that probability where we get balanced accuracy, sensitivity and specificity.

From the second graph it is visible that the optimal cut off is at ‘0.3’.

Metrics Comparison

Metrics	Train Set	Test Set
Accuracy	79.66%	79.08%
Sensitivity	76.50%	76.55%
Specificity	81.60%	80.58%

Conclusion

- Customers whose source of lead was ‘Welingak Website’ have the highest possibility of converting.
- Customers whose source of lead was ‘Reference’ also have the highest possibility of converting.
- Customers spending major time on website should be targeted.
- Customers whose source of lead was ‘Olark Chat’ also have high probability of converting.

THANK YOU