Lead Scoring Case Study

DSC59

(GROUP MEMBERS - Sushree Rutayanee, Prasanth Yeddu & Tejender Singh Sandhu)

Problem Statement:

There is an education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. When people browse to the website, they might go through the courses or fill up a form for the course if they are interested in it 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. The company also gets leads from 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%.

So X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires wants us to build a model where we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO has asked us to reach a lead conversion rate of 80%.

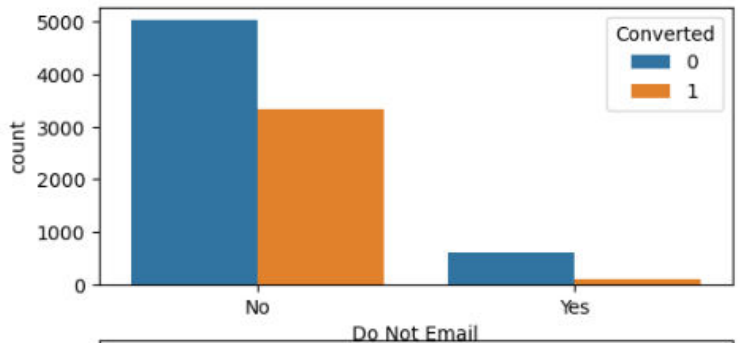
Analysis approach:

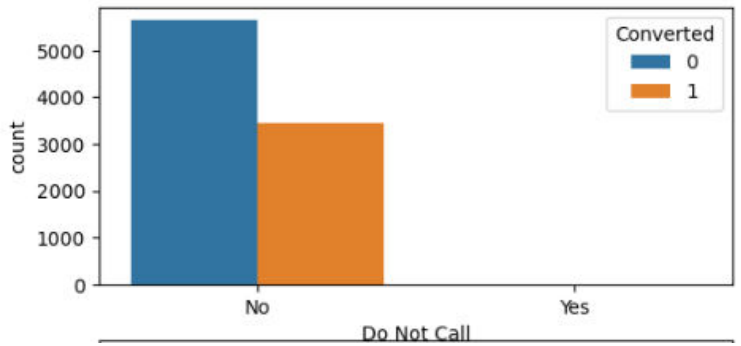
The target column which we have here is a categorical variable i.e. converted or not converted and since we have the labels and it is a classification problem we went for a supervised machine learning algorithm – Logistic regression.

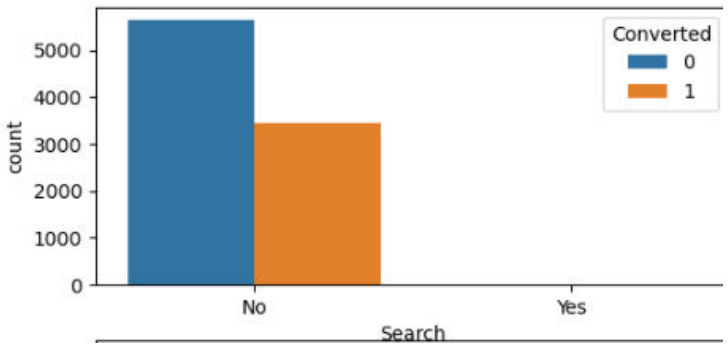
We performed the data cleaning, data preparation, train-test split, scaling and finally model building where we iteratively performed checking and came to an optimal model and in the evaluation we checked all the important metrics like accuracy, sensitivity, specificity, recall, etc.

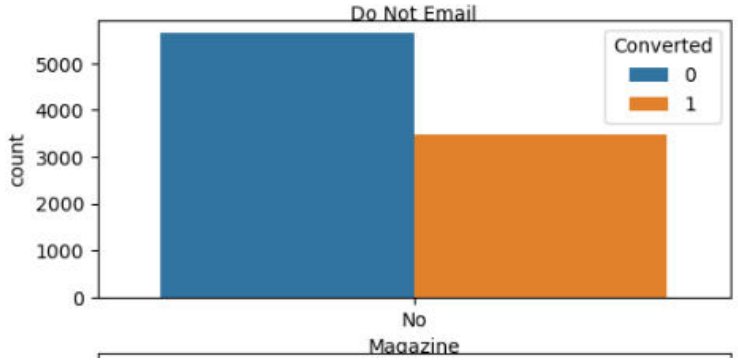
Insights acquired from the case study through EDA:

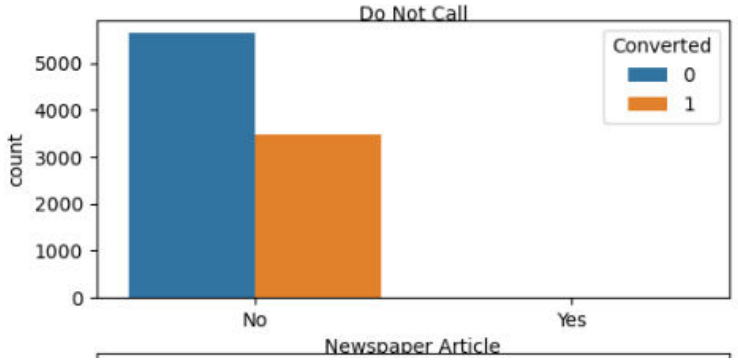
- We plotted some count plots for every categorical variables which have 2 levels V/S converted to see the how much they are contributing to conversion:

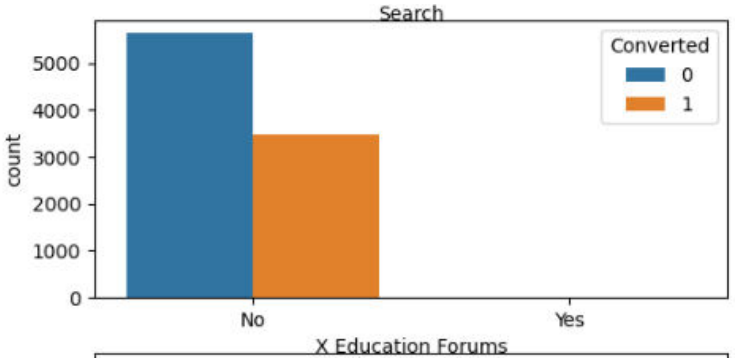


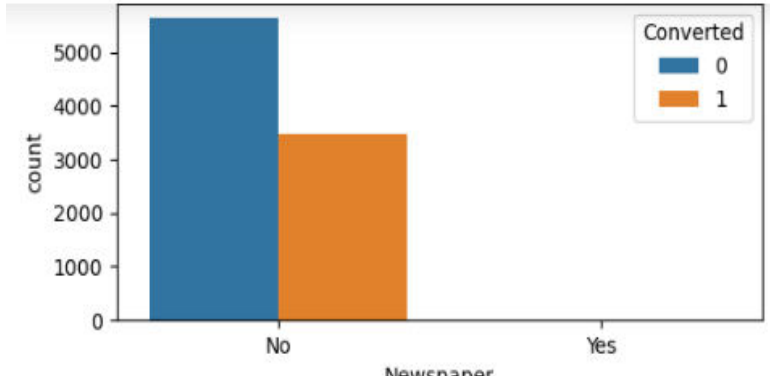




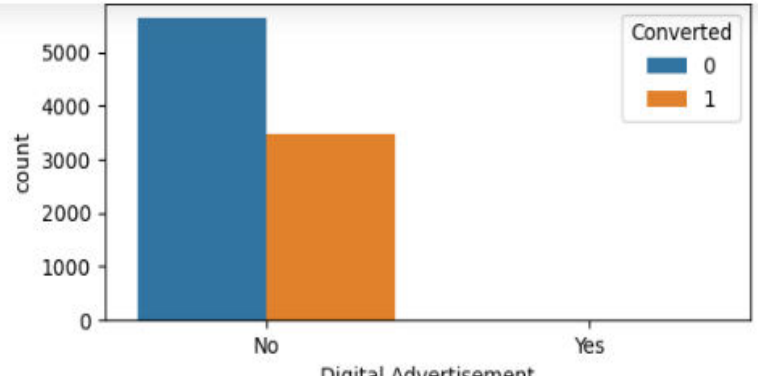




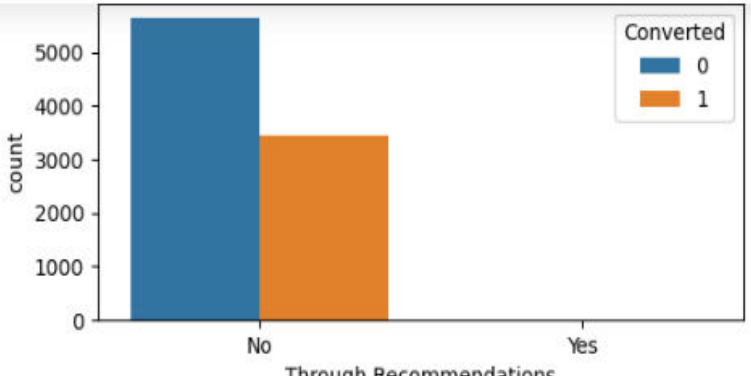




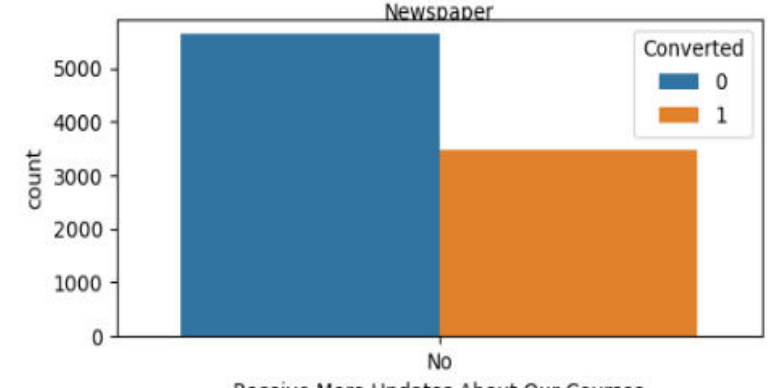
Newspaper



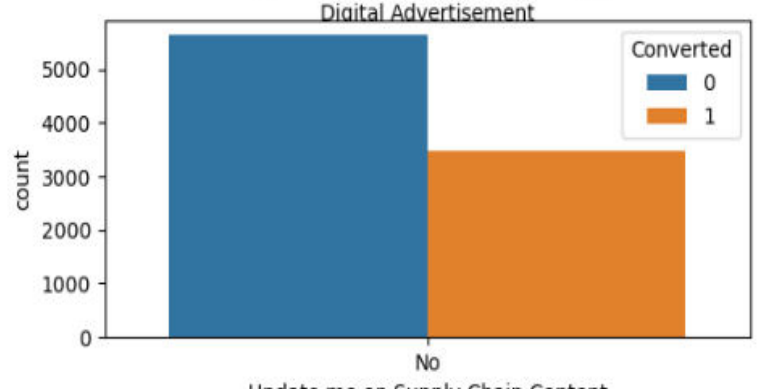
Digital Advertisement



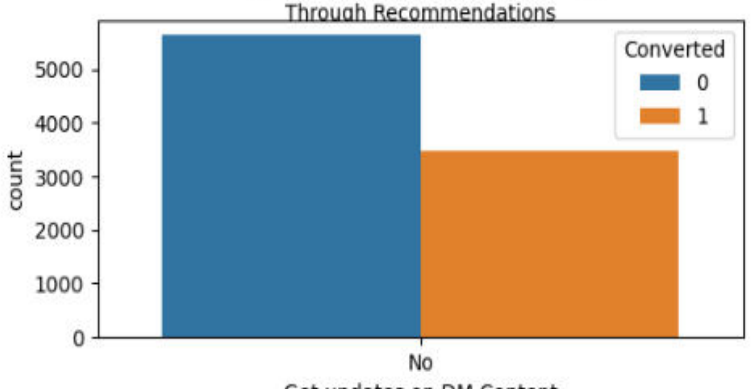
Through Recommendations



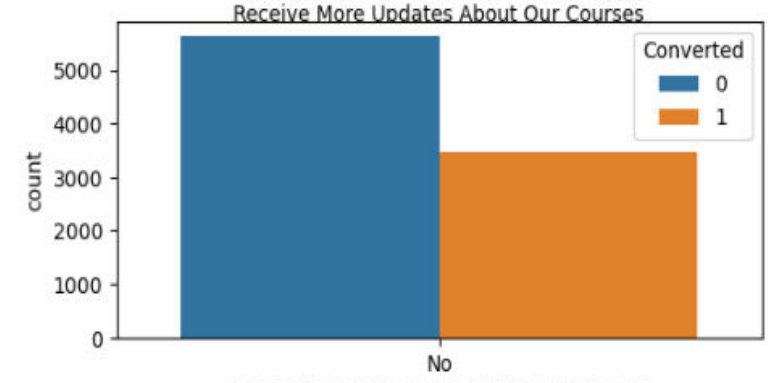
Receive More Updates About Our Courses



Update me on Supply Chain Content



Get updates on DM Content



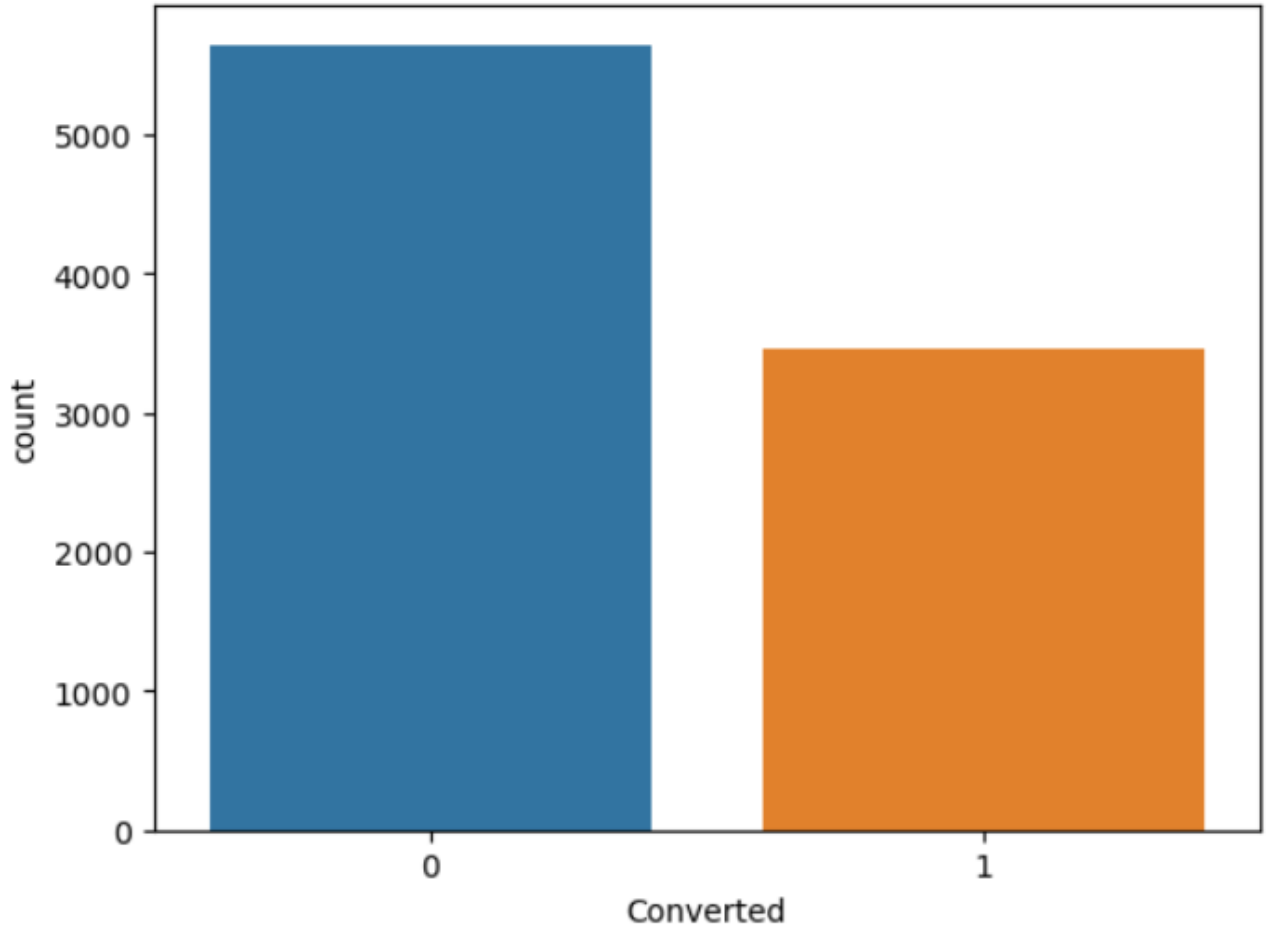
I agree to pay the amount through cheque

A free copy of Mastering The Interview

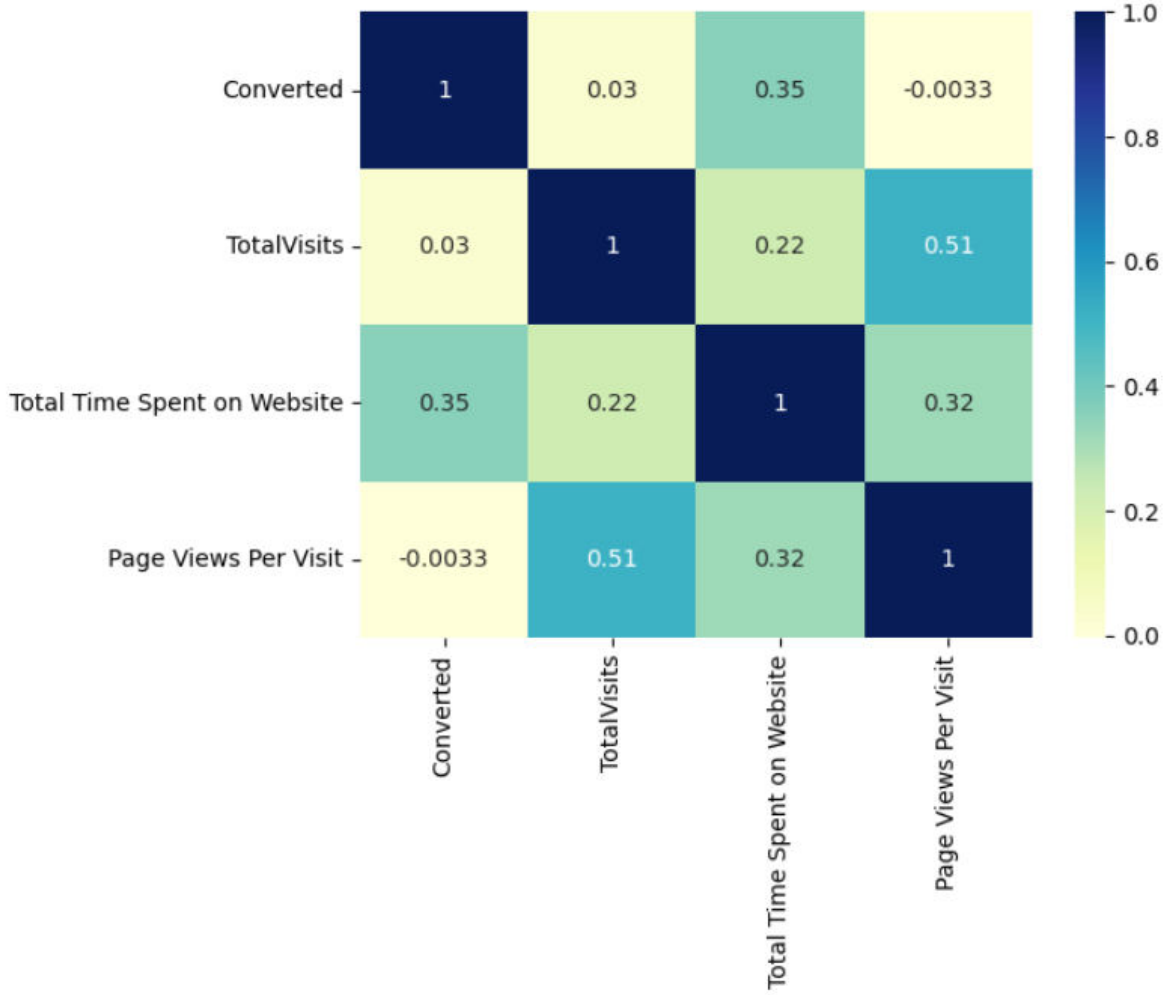
- From the plots which we saw in the previous 2 slides we observed - columns like 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement’, 'Through Recommendations', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content’, 'Get updates on DM Content', 'I agree to pay the amount through cheque', 'Do Not Call' have highly skewed data so these wont help much in the analysis so we decided to drop them.

- Coming to 'A free copy of Mastering The Interview' conversion rate is very low so no point of considering it in the analysis hence we dropped it.

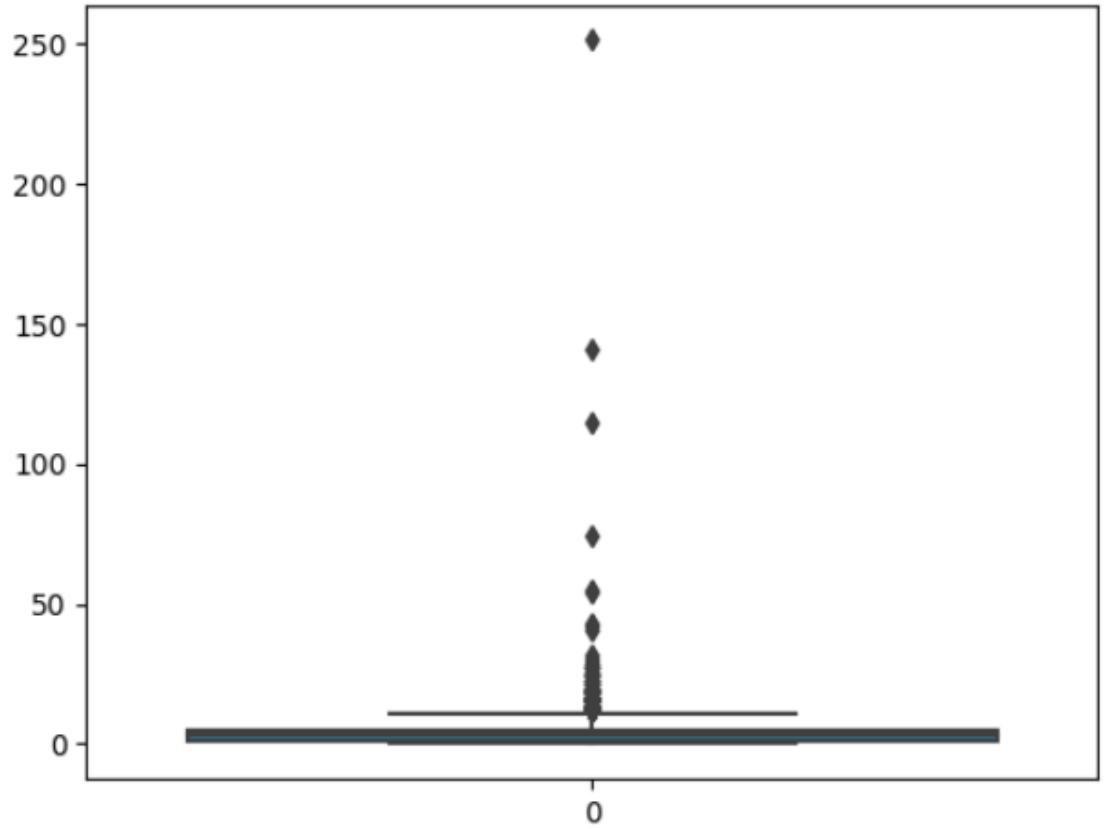
- Next we plotted the target variable i.e. “Converted” and observed the current conversion rate came out to be around 38% which is not so great.

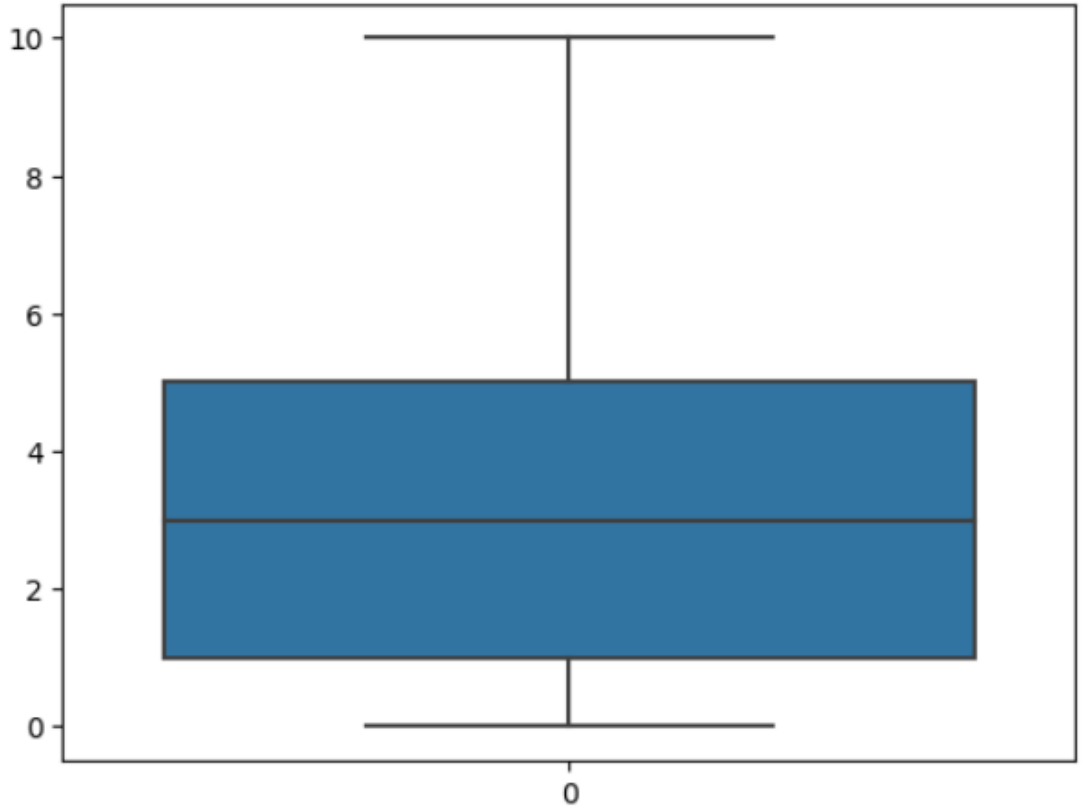


- Next we plotted a heatmap to check the correlation between all the numerical variables and we didn’t observe any heavy correlations as such.

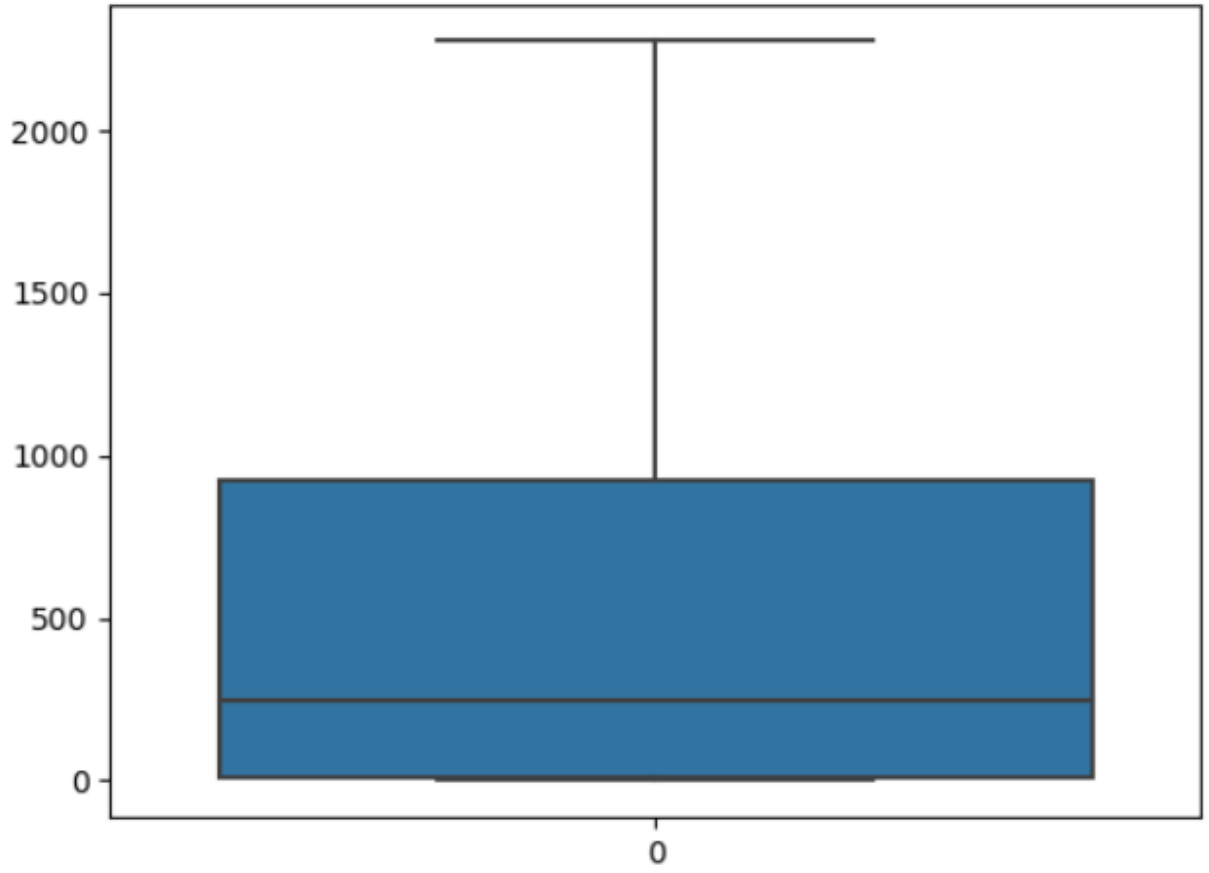


- We plotted a box plot for all the numerical variables one by one i.e. first we checked for “TotalVisits” and observed lot of outliers so then we performed capping and fixed it.

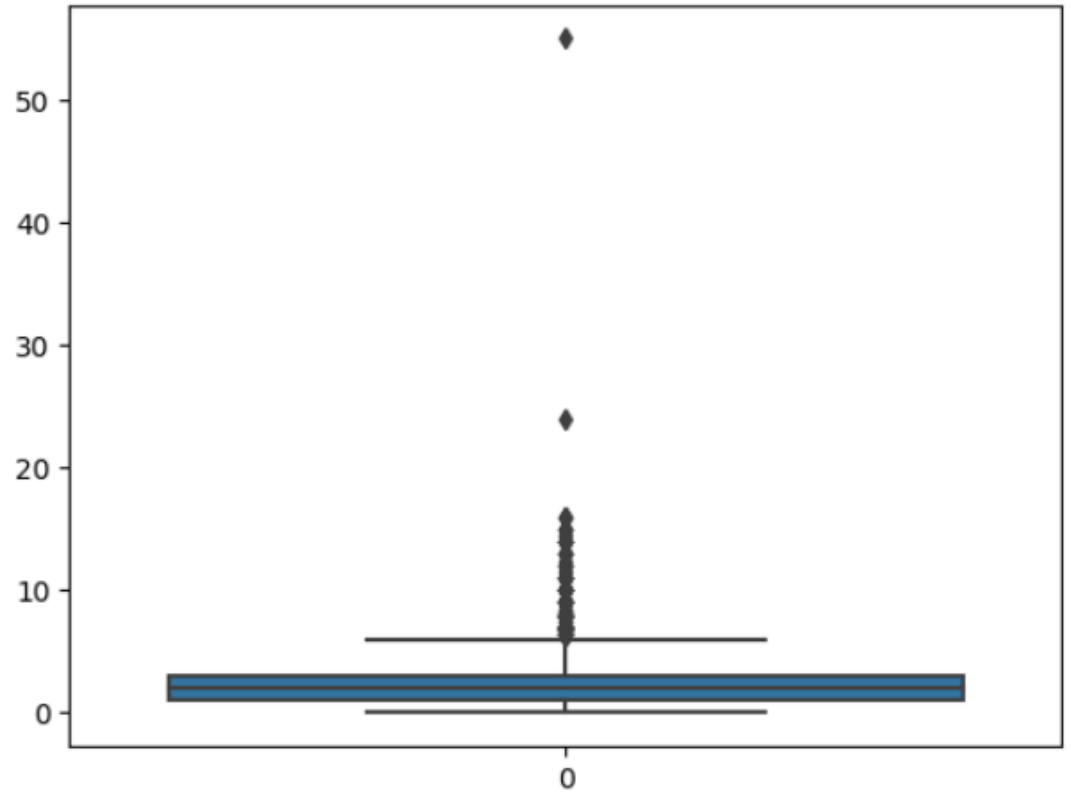


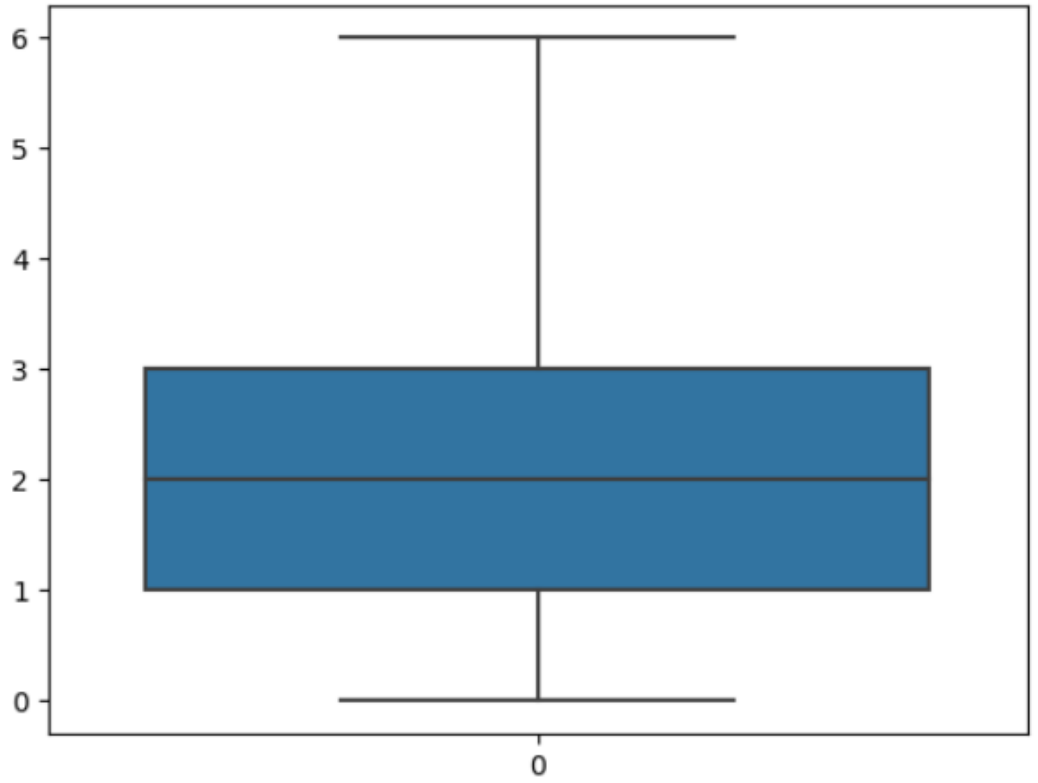


- We plotted a box plot for “Total Time Spent on Website” and observed there were no outliers so we didn’t perform any treatment on it.

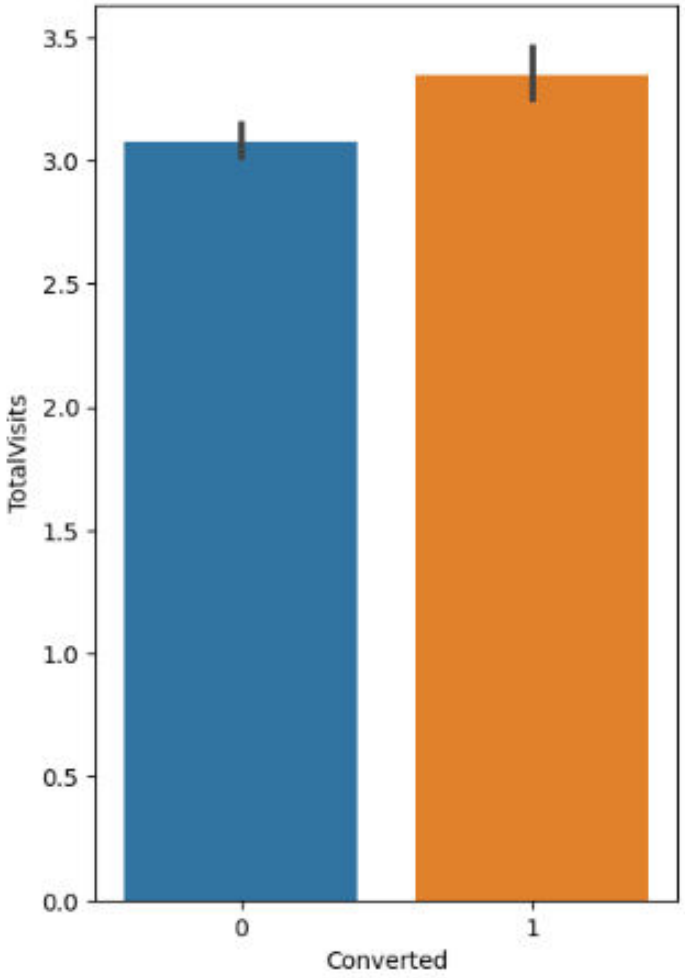


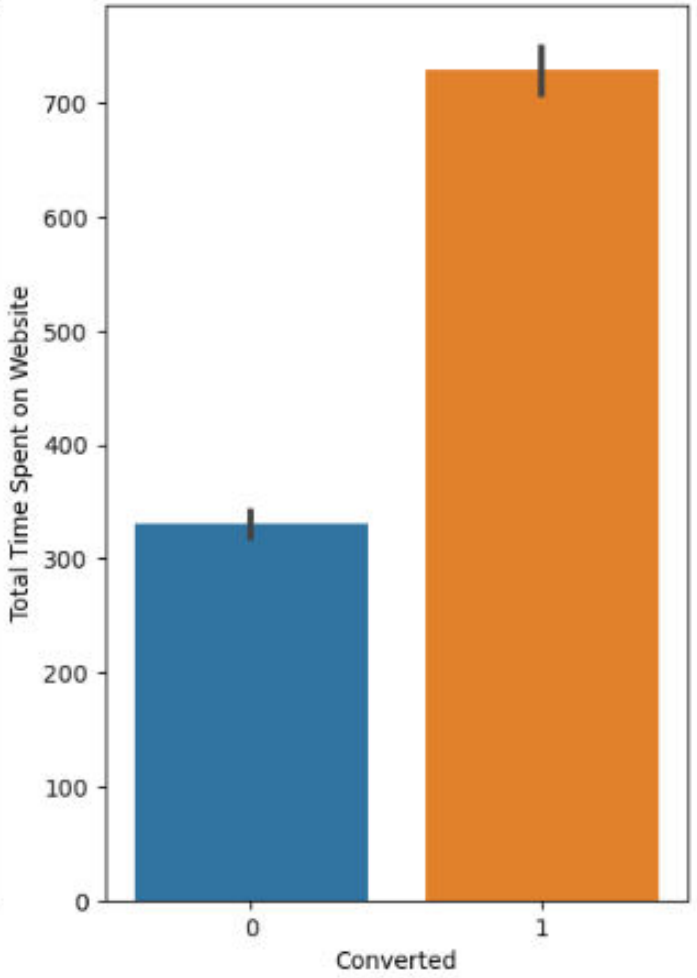
- We plotted a box plot “Page views per visit” and observed lot of outliers so then we performed capping and fixed it.

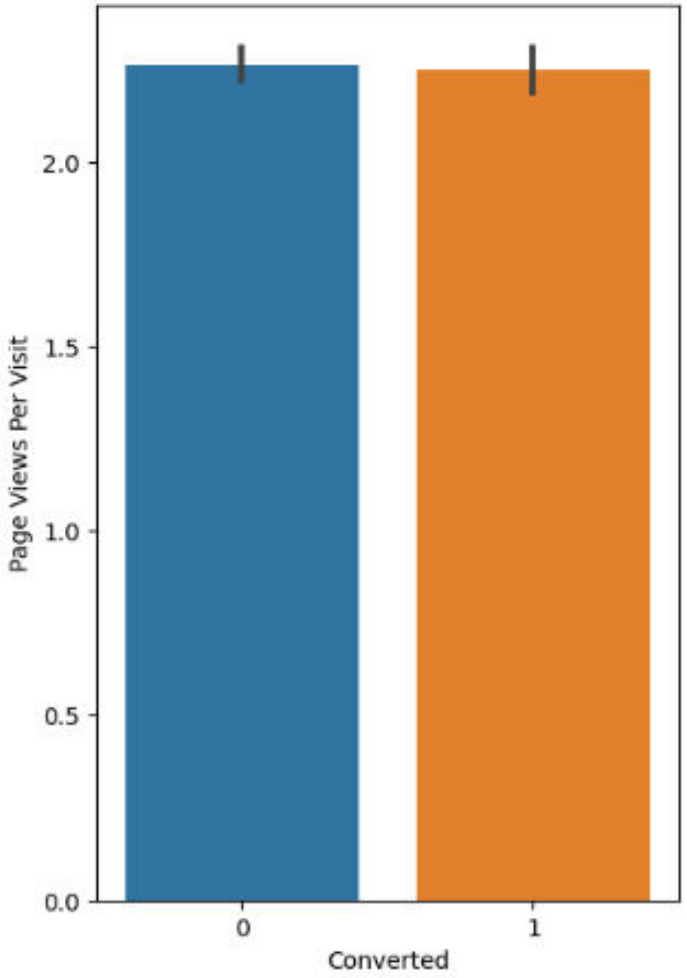




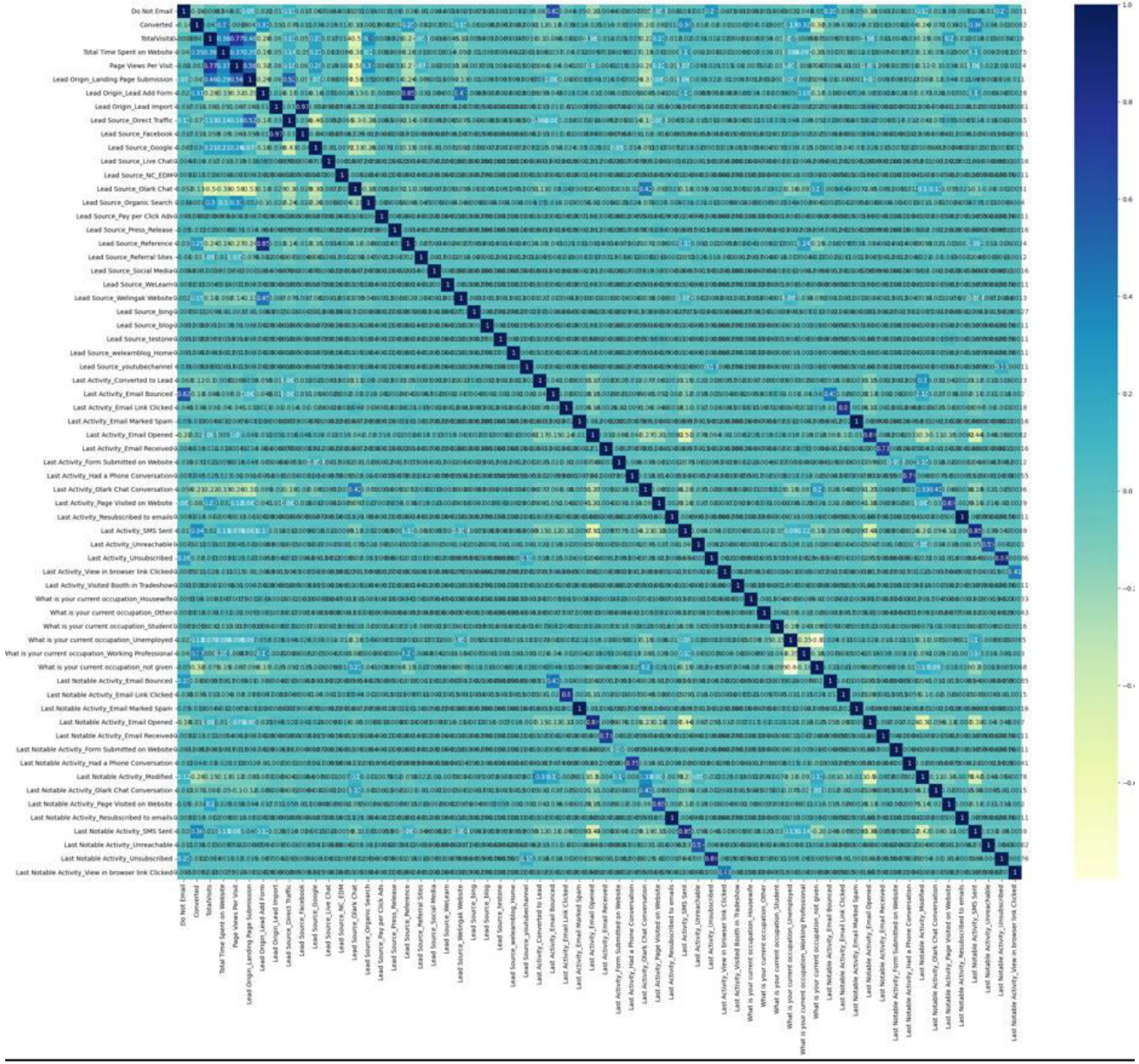
- Next we plotted few bar plots between the numerical variables and target variable to see how good the conversion rate is and we observed “Total Visits” and “Total time spent on website” are contributing towards a good conversion rate.







- Post dummification, we again plotted a heatmap to check the correlation and removed the columns that were highly correlated.

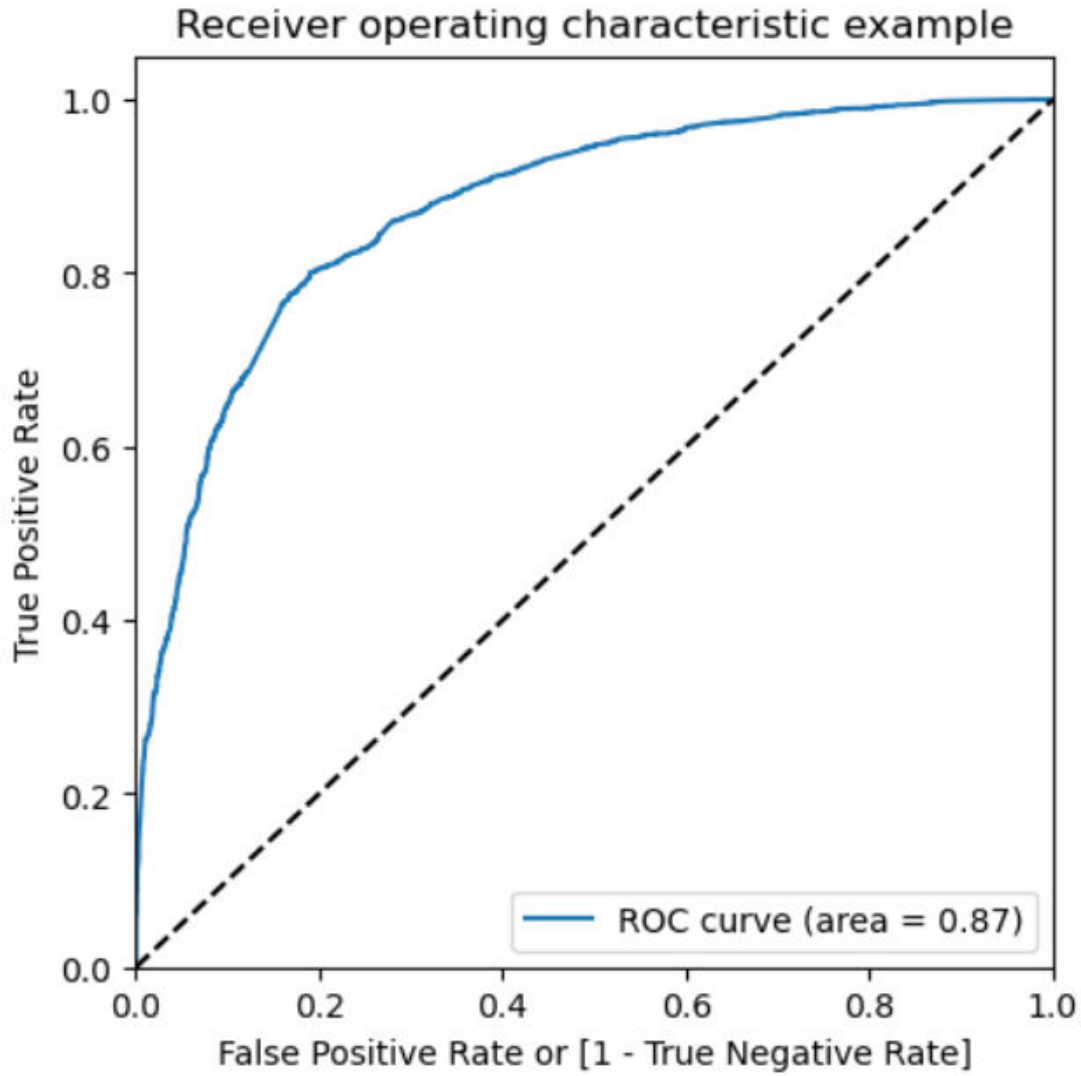


- Post checking p-value and VIF, below is the final model we got after optimizing everything:

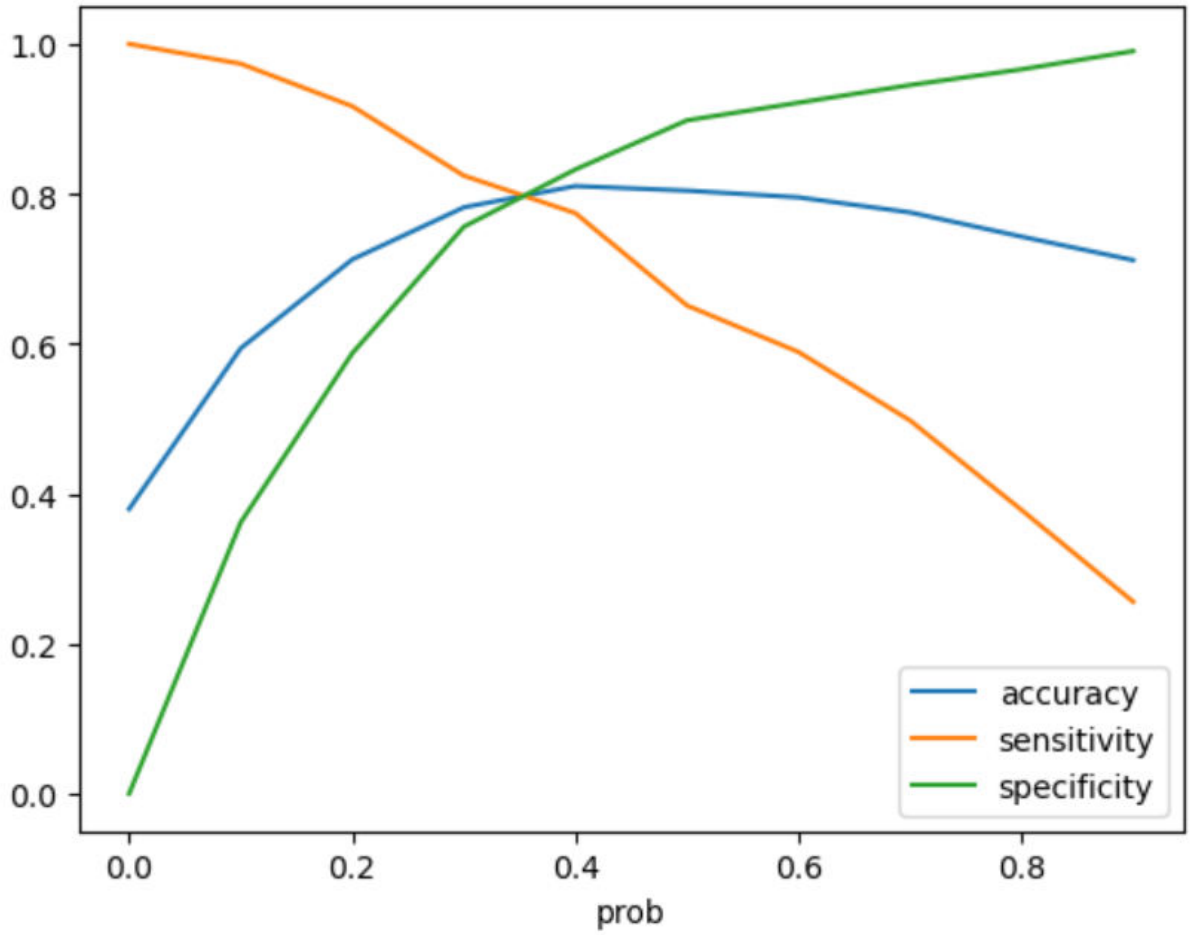
Dep. Variable:	Converted	**No. Observations:**	6372
Model:	GLM	**Df Residuals:**	6362
Model Family:	Binomial	**Df Model:**	9
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2728.8
Date:	Sun, 18 Feb 2024	**Deviance:**	5457.7
Time:	10:56:58	**Pearson chi2:**	6.34e+03
No. Iterations:	6	**Pseudo R-squ. (CS):**	0.3758
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4816	0.048	-10.012	0.000	-0.576	-0.387
Total Time Spent on Website	1.1091	0.039	28.261	0.000	1.032	1.186
Lead Origin_Lead Add Form	3.9983	0.210	19.069	0.000	3.587	4.409
Lead Source_Olark Chat	1.2014	0.102	11.747	0.000	1.001	1.402
Last Activity_Converted to Lead	-1.2062	0.221	-5.466	0.000	-1.639	-0.774
Last Activity_Email Bounced	-2.0971	0.316	-6.644	0.000	-2.716	-1.478
Last Activity_Olark Chat Conversation	-1.3781	0.167	-8.231	0.000	-1.706	-1.050
What is your current occupation_Working Professional	2.5415	0.185	13.769	0.000	2.180	2.903
What is your current occupation_not given	-1.2108	0.085	-14.226	0.000	-1.378	-1.044
Last Notable Activity_Modified	-0.7035	0.084	-8.418	0.000	-0.867	-0.540

- The ROC (Receiver Operating Characteristic) curve is as follows:



- For getting the optimal cutoff point where all three accuracy, sensitivity and specificity have a good value, we plotted the below graph and concluded the 0.3 would be a good one.



CONCLUSION:

Post Model building and evaluation, below are the variables on which the X Education business should focus more in order to increase the conversion rate:

- Total Visits
- Total time spent on website
- When the lead origin is “Lead add form”
- When the lead source is “Olark chat”
- When the current working occupation is “working professional”