Lead Scoring Case Study
Ankit Chand
Sohini Chaudhuri

Data Quality checks and handling missing values:

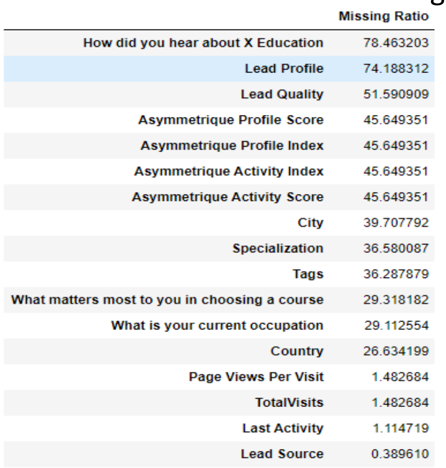
Data set were provided

- Leads Data

We have considered Leads data set and performed necessary actions as explained below.

To find out the missing values and handle it :

- After converting label 'select' into nan values. We will again check all the missing value.



	Missing Ratio
How did you hear about X Education	78.463203
Lead Profile	74.188312
Lead Quality	51.590909
Asymmetrique Profile Score	45.649351
Asymmetrique Profile Index	45.649351
Asymmetrique Activity Index	45.649351
Asymmetrique Activity Score	45.649351
City	39.707792
Specialization	36.580087
Tags	36.287879
What matters most to you in choosing a course	29.318182
What is your current occupation	29.112554
Country	26.634199
Page Views Per Visit	1.482684
TotalVisits	1.482684
Last Activity	1.114719
Lead Source	0.389610

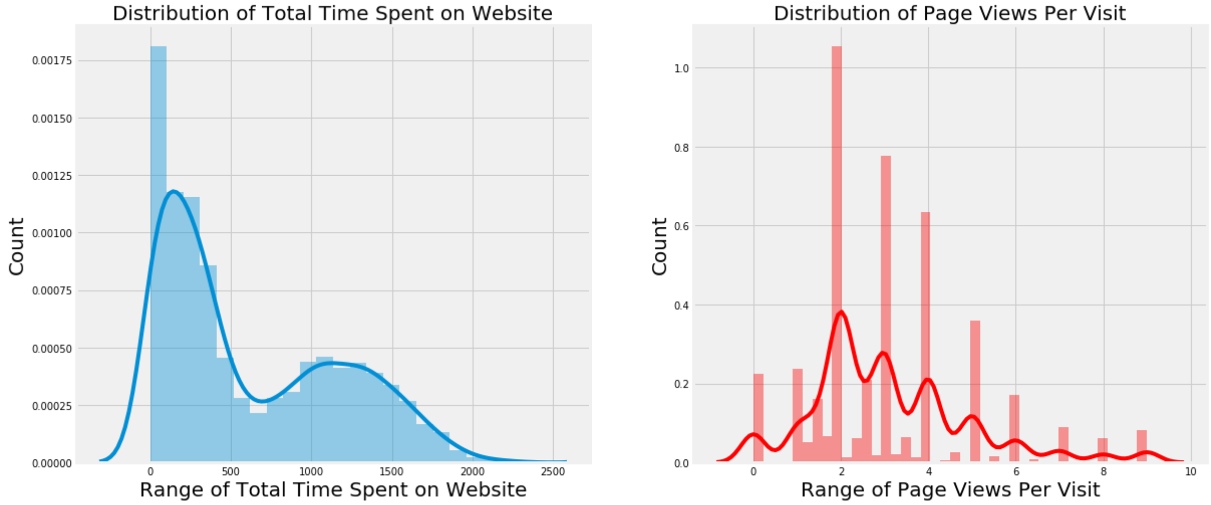
- And dropped all columns from data frame for which missing values % is more than 40.
- We will check the skewness which is very high and drop them also.

Data Preparation:

While understanding the data, we get to know that many categories with a very little percentage of rows, we combine into a new category name 'Other'. After cleaning, the retained percentage of rows is 58% of the rows.

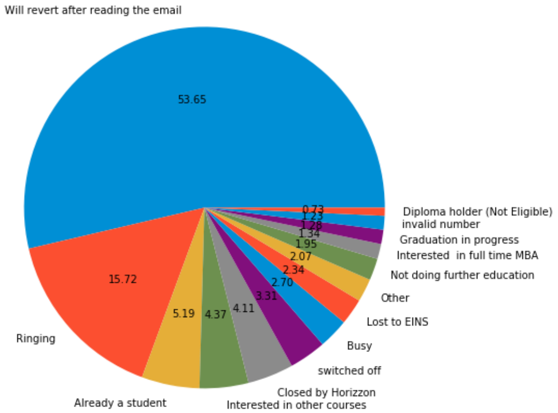
Data Visualization:

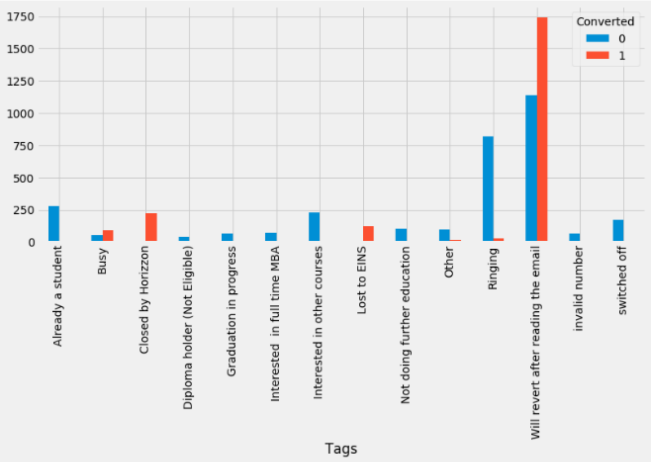
Total Time Spent on Website vs Page Views Per Visit



we can infer one thing that at max Page views per visit is 3.8 and total time spent on website is 0.12. Means 2000 Customer likely to spent time on website where as average number of page view per visit is 2.

Tags

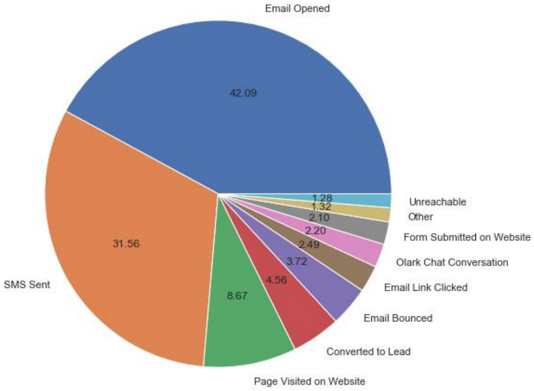


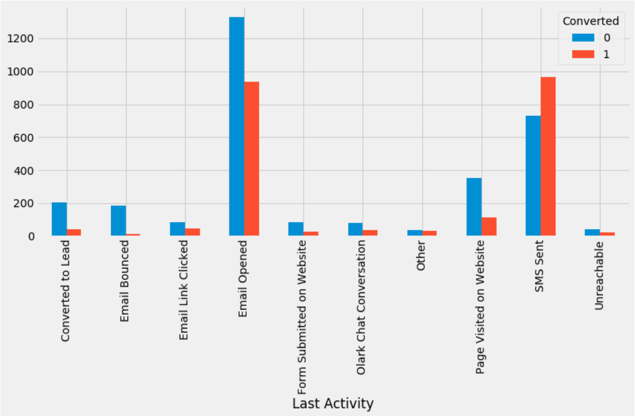


We can see that majority of customer in conversion process are 53.65% were those who will revert after reading the email and 15.72% of ringing. And also converted leads were high in will revert after reading the email

From this we can say closed by horizon and will revert after reading the email are our potential leads

Last Activity



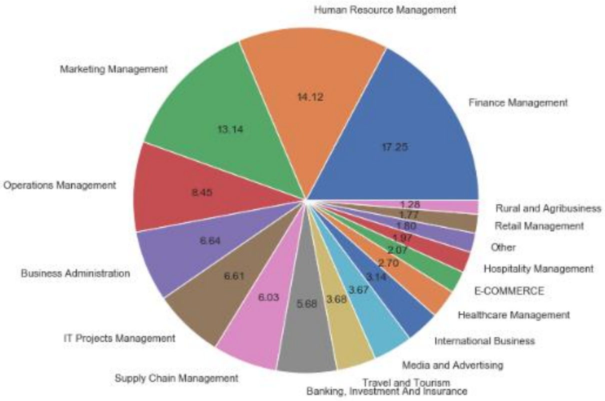


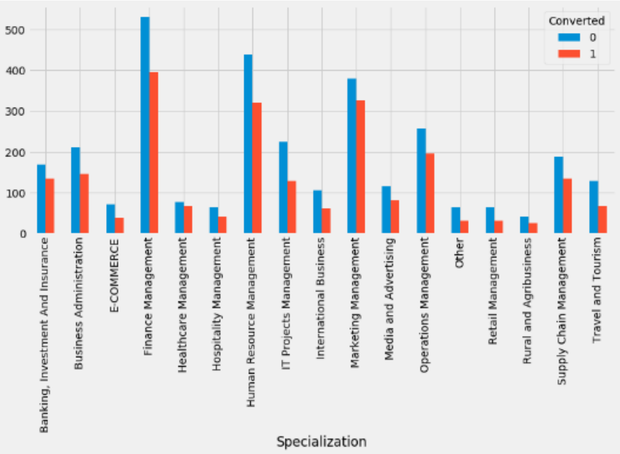
We see that majority of customer in conversion process are 42.09% were those who will open the email and 31.56% of sent through SMS.

Out of non-conversion lead, x-education able to achieved 950 conversion lead

We say SMS sent and open the email are our potential leads

Specialization

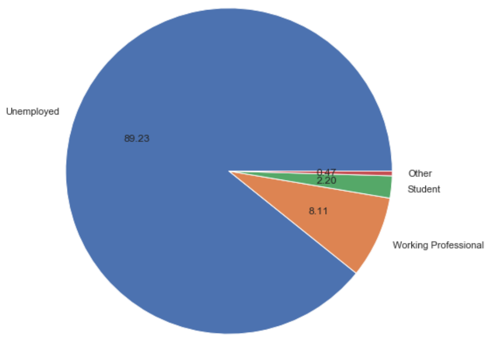


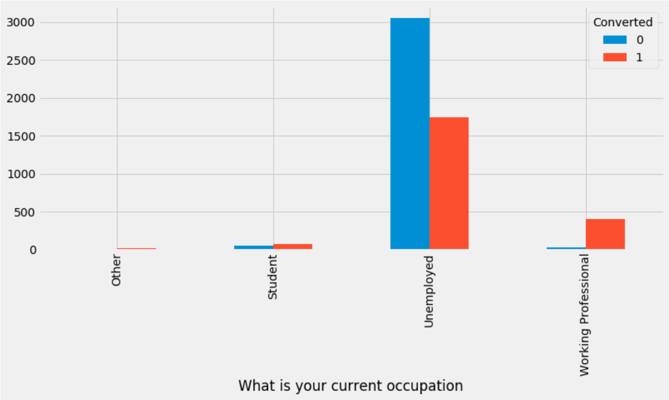


We see that majority of customer in conversion process are 17.25% were those who are in Finance

Management, 14.12% of Human Resource Management and 13.14 of Marketing Management.
Out of non-conversion lead, x-education able to achieved 370 conversion lead
So, we say open the email and SMS sent are our potential leads

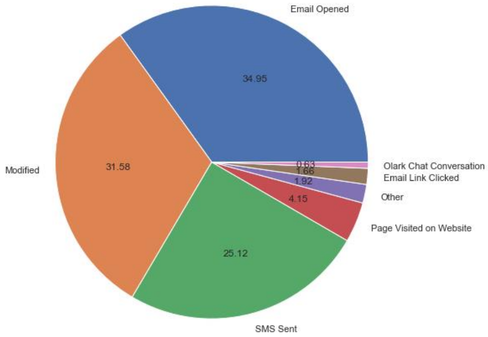
What is your current occupation

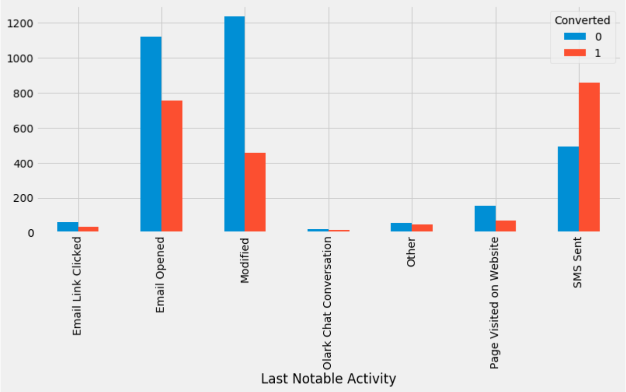




We see that majority of customer in conversion process are 89.23% were Unemployed and 8.11% of Working Professional. And highly conversion rate from Working Professional and some from student. And highly non-conversion rate from Unemployed but still x-education able to achieved the lead to convert from Unemployed around 1700.
Thus, Working Professional and student are our potential leads

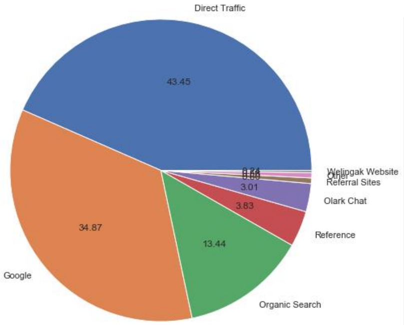
Last Notable Activity

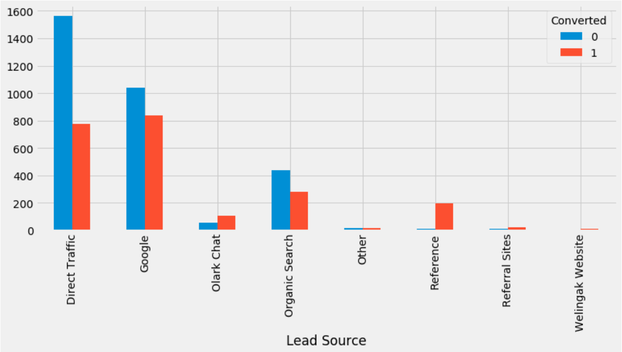




We see that majority of customer in conversion process are 34.95% were those who will open the email, 31.58% of modified and 25.12 through SMS Sent. Means those who check the SMS are our potential leads

Lead Source

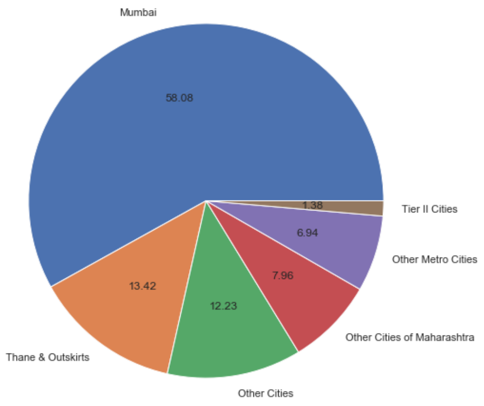


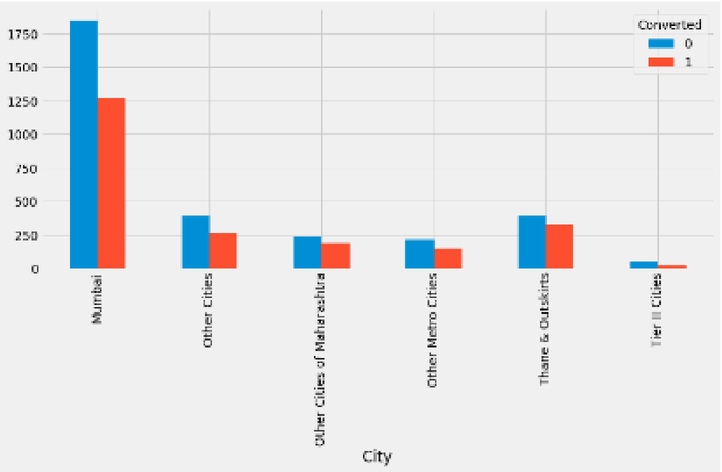


We see that majority of customer in conversion process were from 43.45% of direct traffic, 34.87% from Google and 13.44 from Organic Search

Therefore, most of the customers that are in reference are converted to full time and through Google also 830 is achieved

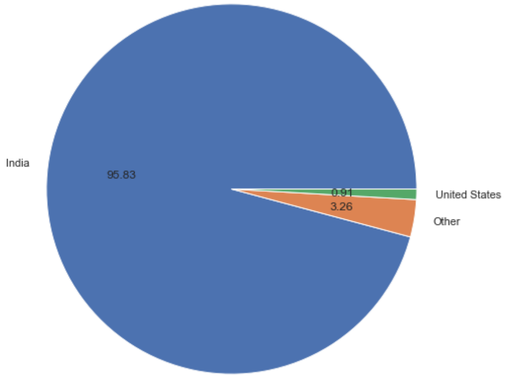
City

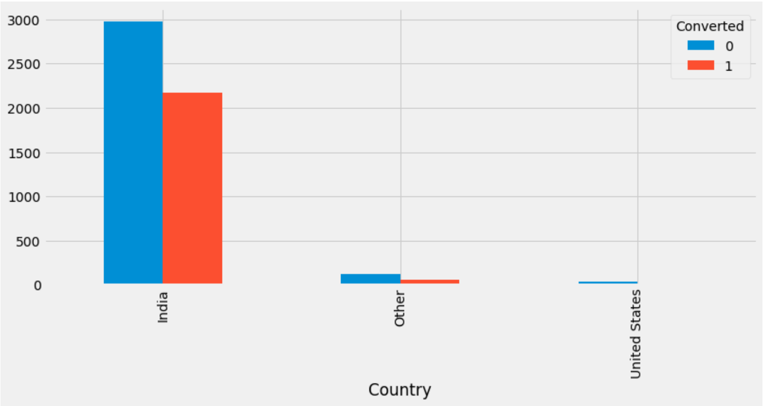




The majority of customer in conversion process were from 58.08% of Mumbai, 13.42% of Thane & Outskirts and 12.23% of Other Cities. Means most of the customers were from Maharashtra. Out of non-conversion rate, customer from Mumbai were able get converted up to 1255 and from Thane & Outskirts is 260

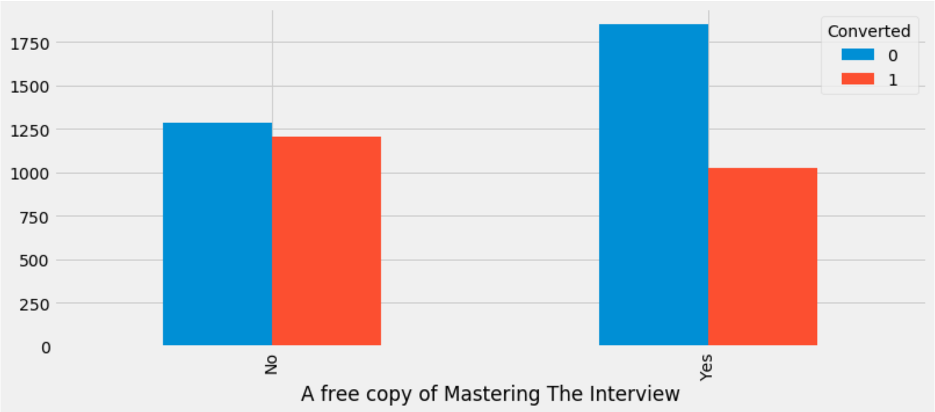
Country





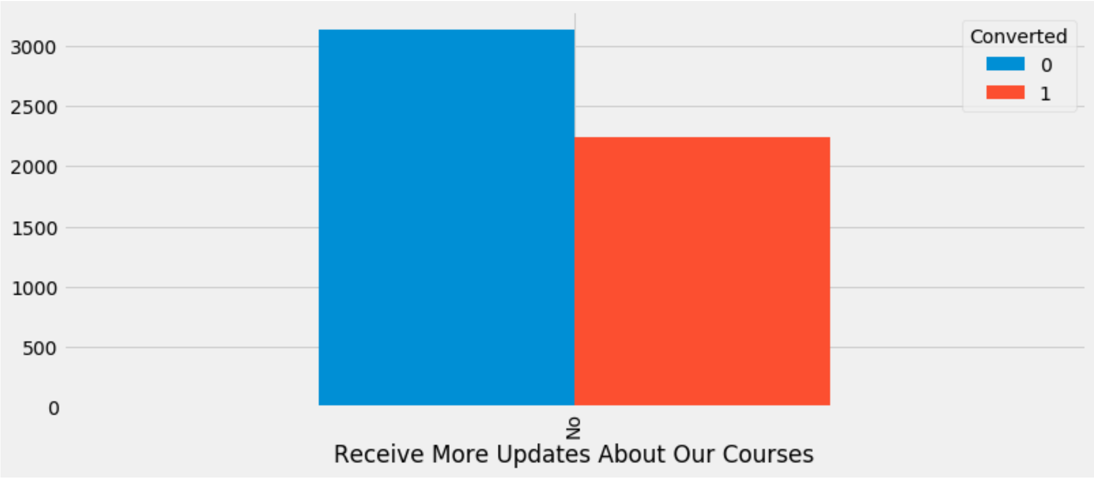
Definitely most of the customers were from India but 0.91% were from United States. X-Education targating customer more from India rather than others. And out of 95%, 80% is converted

A free copy of Mastering the Interview



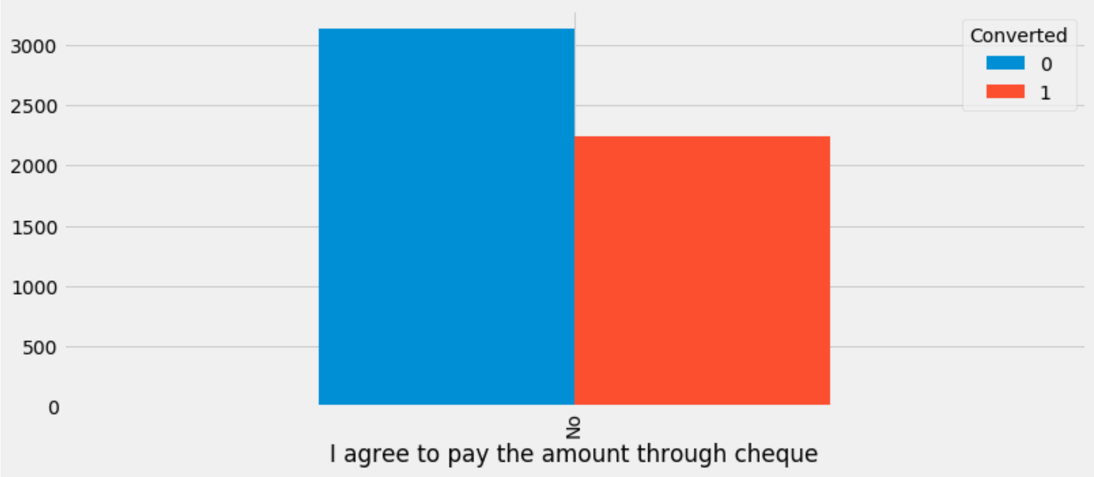
Most of the customers who wants a free copy of 'Mastering the Interview' has not converted into full time whereas most of the customers who don't want a free copy of 'Mastering the Interview' has nearly converted into full time

Receive More Updates About Our Courses



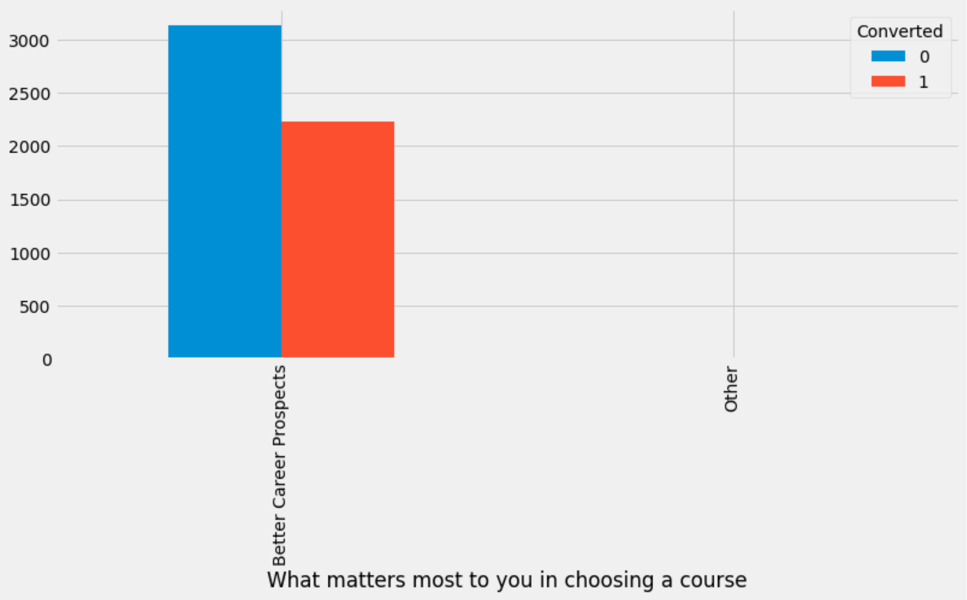
Those customers don’t want to choose to receive more updates about the courses were converted more into full time

I agree to pay the amount through cheque



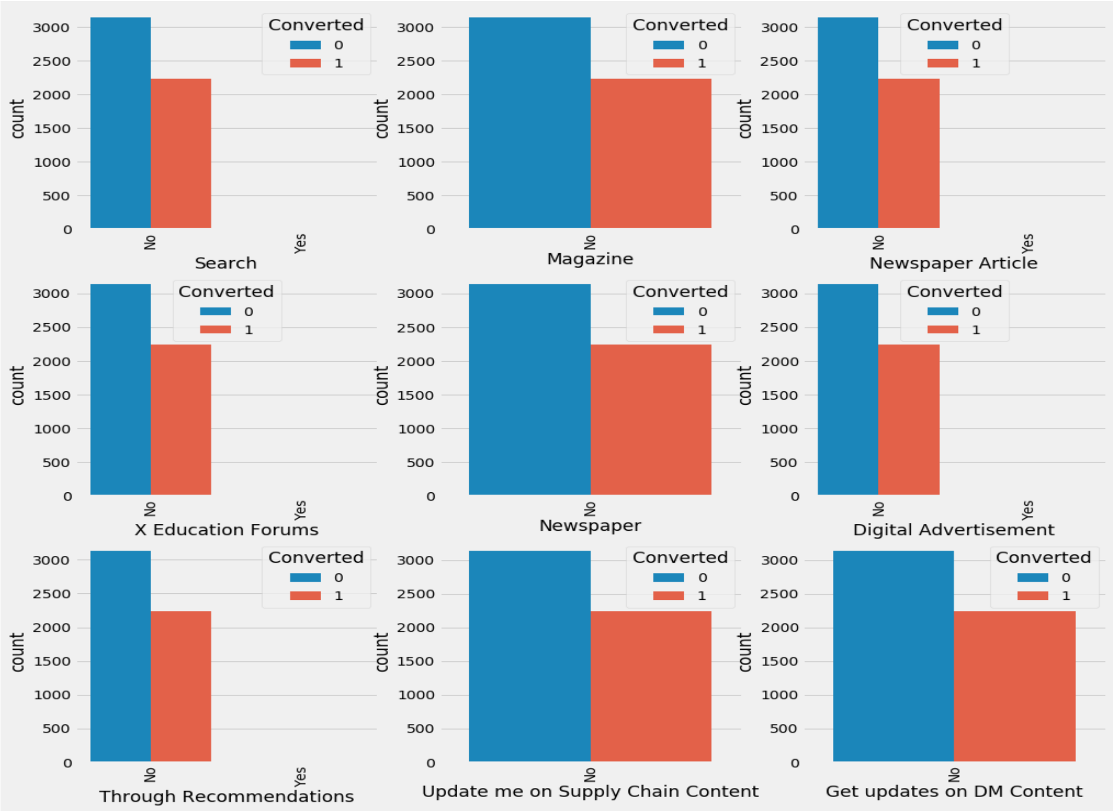
We clearly see, customers don’t want to pay the amount through cheque but preferred more digitally.

What matters most to you in choosing a course



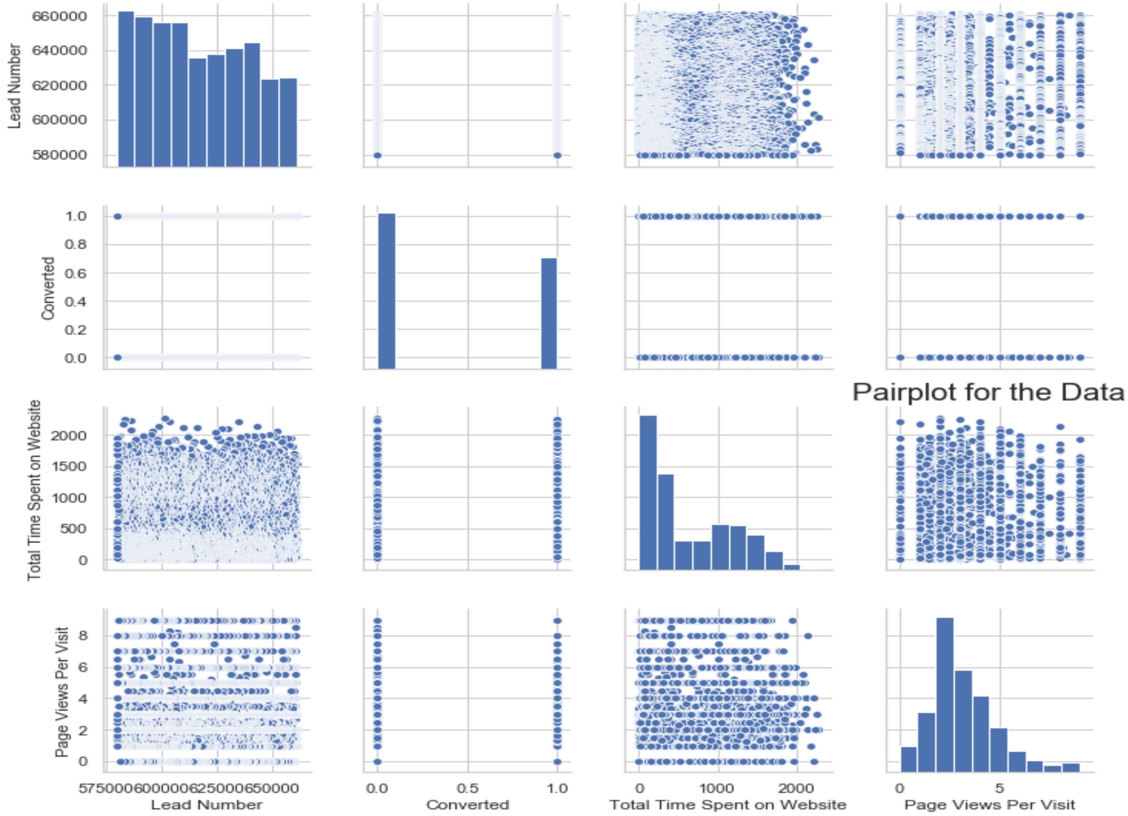
Of course, doing courses will be better for career prospects, still some of customer were not interested to convert

Count Plot



we can infer that most of them were in lower lead conversion. But out of lower lead still customers has been converted into full time up to 2000

Pair Plot

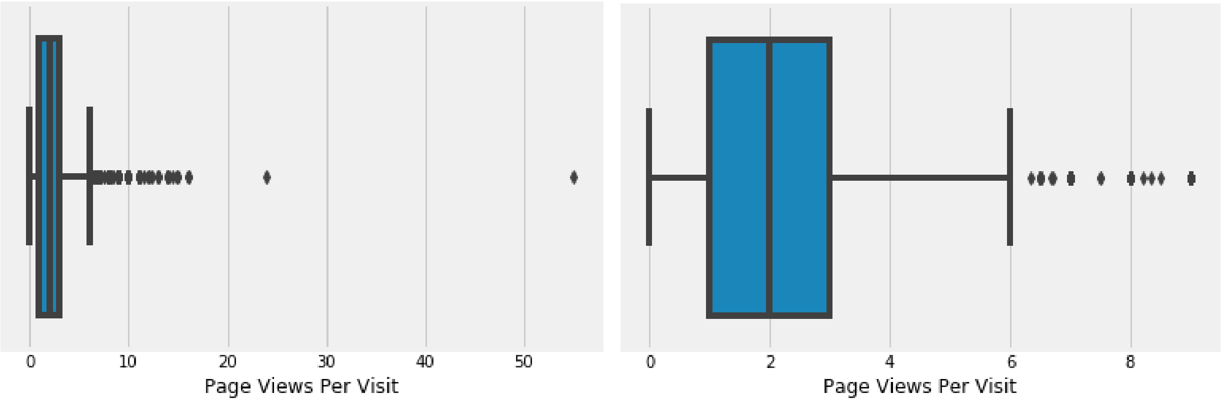


By pair plot we can able to see the pattern of highest correlation with the target variable

Outlier Treatment

Analyzing and Interpreting *outlier through* Skewness Values *by plo*tting boxplot

Page Views Per Visit Column



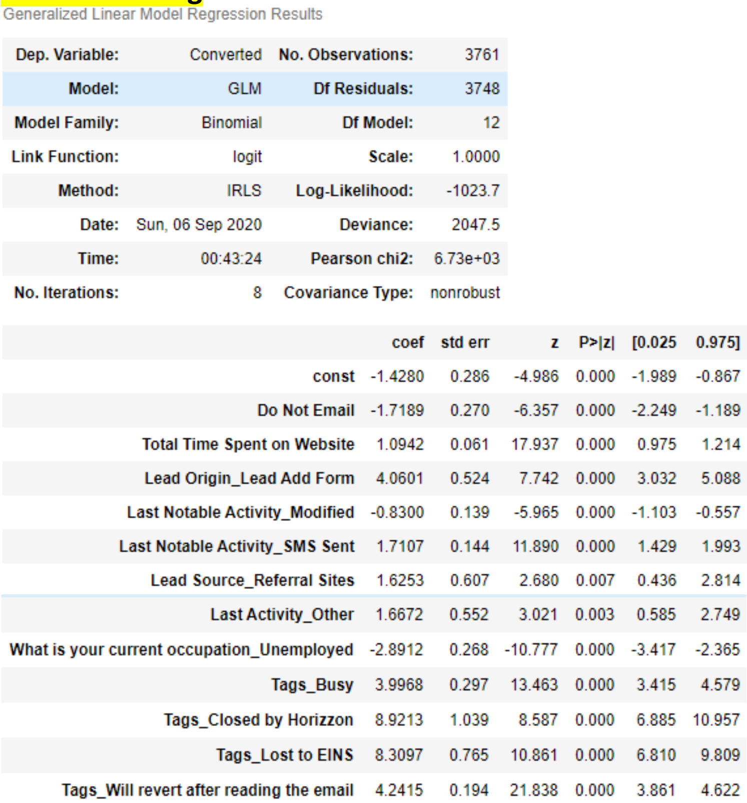
After Looking at the Box plot, it become clear that there are Outliers, and these outliers are important to be treated for a better predictive model. We do cap because if we drop them then we may lose students who want to converted.

So, skewness gets overcome (like in page views per visit column earlier skewness was 2.87132 and now it came down to 0.91362)

For now, we can leave those values as it is. And good to go

Scaling: Scaling is performed mostly during model building processes to bring everything to the same scale. Standardized scaling, on the other hand, brings all the data points in a normal distribution with mean zero and standard deviation one. It can improve the efficiency of model. So, it becomes an essential step before model building.

Model Building



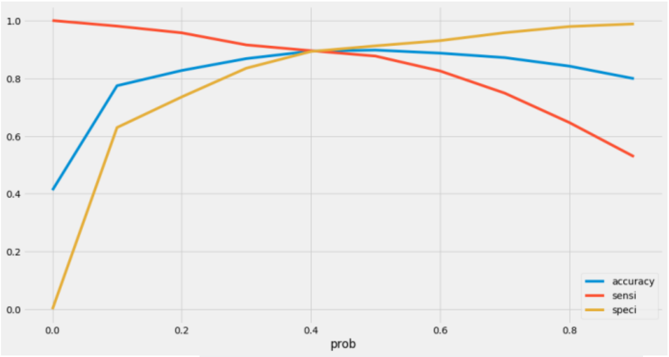
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	3761
Model:	GLM	Df Residuals:	3748
Model Family:	Binomial	Df Model:	12
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1023.7
Date:	Sun, 06 Sep 2020	Deviance:	2047.5
Time:	00:43:24	Pearson chi2:	6.73e+03
No. Iterations:	8	Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.4280	0.286	-4.986	0.000	-1.989	-0.867
Do Not Email	-1.7189	0.270	-6.357	0.000	-2.249	-1.189
Total Time Spent on Website	1.0942	0.061	17.937	0.000	0.975	1.214
Lead Origin_Lead Add Form	4.0601	0.524	7.742	0.000	3.032	5.088
Last Notable Activity_Modified	-0.8300	0.139	-5.965	0.000	-1.103	-0.557
Last Notable Activity_SMS Sent	1.7107	0.144	11.890	0.000	1.429	1.993
Lead Source_Referral Sites	1.6253	0.607	2.680	0.007	0.436	2.814
Last Activity_Other	1.6672	0.552	3.021	0.003	0.585	2.749
What is your current occupation_Unemployed	-2.8912	0.268	-10.777	0.000	-3.417	-2.365
Tags_Busy	3.9968	0.297	13.463	0.000	3.415	4.579
Tags_Closed by Horizzon	8.9213	1.039	8.587	0.000	6.885	10.957
Tags_Lost to EINS	8.3097	0.765	10.861	0.000	6.810	9.809
Tags_Will revert after reading the email	4.2415	0.194	21.838	0.000	3.861	4.622

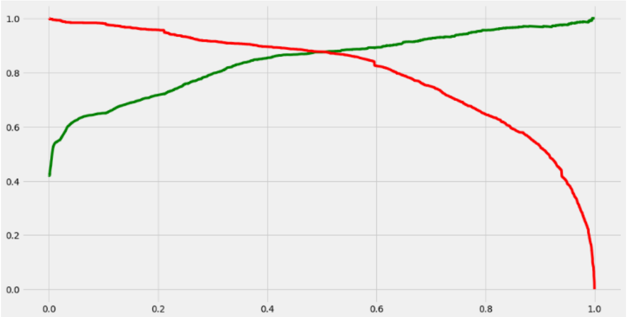
we use the recursive feature elimination technique to choose 25 variables, and using the manual approach we drop some variables based on p-value and VIF, and the final model we have 12 variables

Let’s find out Number of Optimal cutoff point:



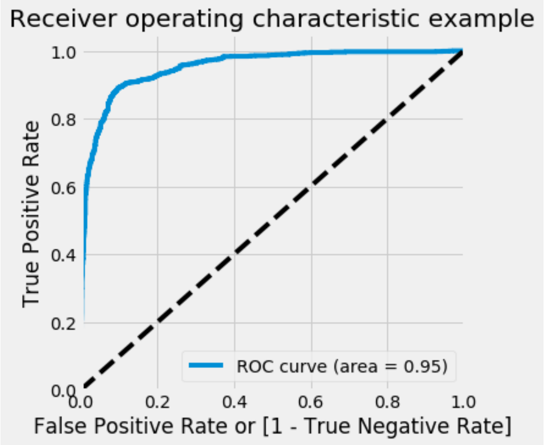
At 0.4 accuracy, sensitivity and specificity are 0.893911, 0.895551 and 0.892760 almost equal

Precision and Recall tradeoff



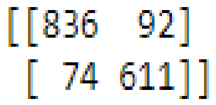
We get to see high recall/sensitivity and high precision. From this, we get to know that model is a very good means we don't have unwanted variables (noise) in our model.

Plotting the ROC Curve

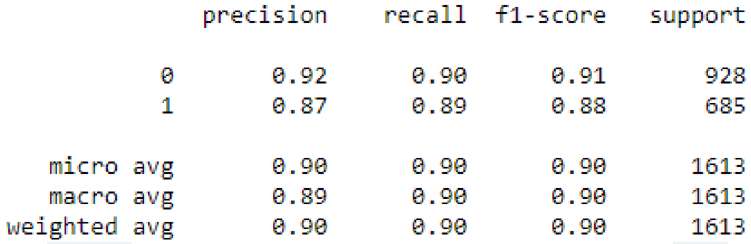


The above model says that there is a 95% chance that the model can distinguish between positive and negative class

As per my Confusion Matrix, Classification Report is generated



```
[[836  92]
 [ 74 611]]
```



```
             precision    recall  f1-score   support

           0       0.92      0.90      0.91       928
           1       0.87      0.89      0.88       685

   micro avg       0.90      0.90      0.90      1613
   macro avg       0.89      0.90      0.90      1613
weighted avg       0.90      0.90      0.90      1613
```

By optimal cutoff 0.4: class 0 (student not get converted) Sensitivity is 90%. And class 1 (student get converted) Sensitivity is 89%. Out of 1613 student only 685 students are able to converted into full time student. And rest 928 students are not converted into full time student.

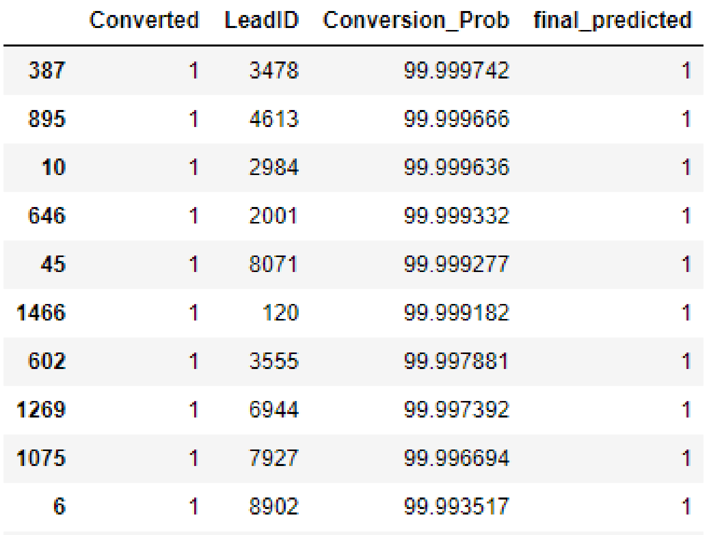
Conclusion

So, from this analysis by sensitivity we can say that, probability of student getting converted into full time student is 89%.

On Training: sensitivity = 89.5% and precision = 85.4%
On Testing: sensitivity = 89% and precision = 86.9%
Which conclude Model is very good and can able to classify correctly.

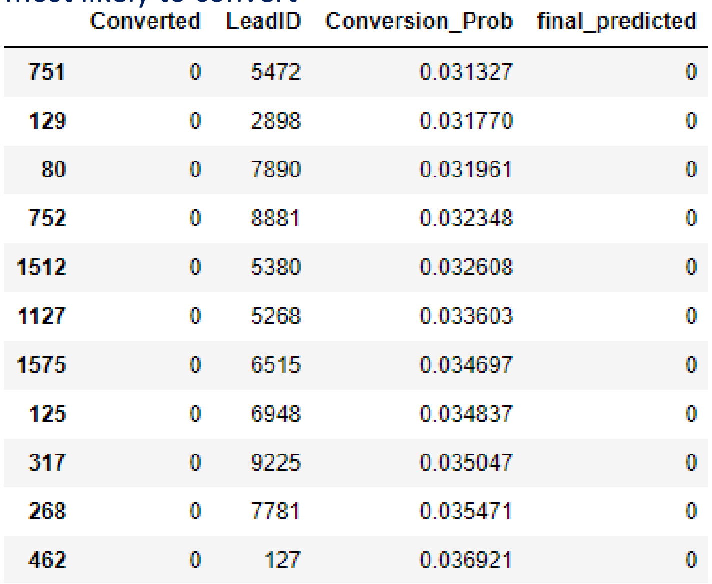
As per business requirement, the potential leads are those on which at a certain threshold the recall/sensitivity more. And in the future also the company can handle the model smoothly by changing the threshold according to their requirements.

A lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads are:



	Converted	LeadID	Conversion_Prob	final_predicted
387	1	3478	99.999742	1
895	1	4613	99.999666	1
10	1	2984	99.999636	1
646	1	2001	99.999332	1
45	1	8071	99.999277	1
1466	1	120	99.999182	1
602	1	3555	99.997881	1
1269	1	6944	99.997392	1
1075	1	7927	99.996694	1
6	1	8902	99.993517	1

According to business requirement, we need to identify a higher score would mean that the lead is hot, i.e. is most likely to convert



	Converted	LeadID	Conversion_Prob	final_predicted
751	0	5472	0.031327	0
129	0	2898	0.031770	0
80	0	7890	0.031961	0
752	0	8881	0.032348	0
1512	0	5380	0.032608	0
1127	0	5268	0.033603	0
1575	0	6515	0.034697	0
125	0	6948	0.034837	0
317	0	9225	0.035047	0
268	0	7781	0.035471	0
462	0	127	0.036921	0

A lower score would mean that the lead is cold and will mostly not get converted