Lead Scoring Case Study

BY – Amya Gupta

Khushi Sapra

Business Problem

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.

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%.

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.

The company requires 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%

DATA CLEANING

Percentage Null values in each column

How did you hear about X Education	78.463203
Lead Profile	74.188312
Lead Quality	51.590909
Asymmetrique Profile Score	45.649351
Asymmetrique Activity Score	45.649351
Asymmetrique Profile Index	45.649351
Asymmetrique Activity Index	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
TotalVisits	1.482684
Page Views Per Visit	1.482684
Last Activity	1.114719
Lead Source	0.389610
Null_Count	0.000000
Total Time Spent on Website	0.000000
Converted	0.000000
Do Not Call	0.000000
Do Not Email	0.000000
Newspaper Article	0.000000
Search	0.000000
Last Notable Activity	0.000000
X Education Forums	0.000000
Newspaper	0.000000
Digital Advertisement	0.000000
Through Recommendations	0.000000
A free copy of Mastering The Interview	0.000000
Lead Origin	0.000000

Looking at the NULL percentage, we took 45% to be the threshold for missing values and we dropped all the columns that had more than 45% of missing values.

After removing all the unnecessary columns, we looked at the remaining columns individually, removed the missing values from them, explored them for EDA and made them ready for the regression

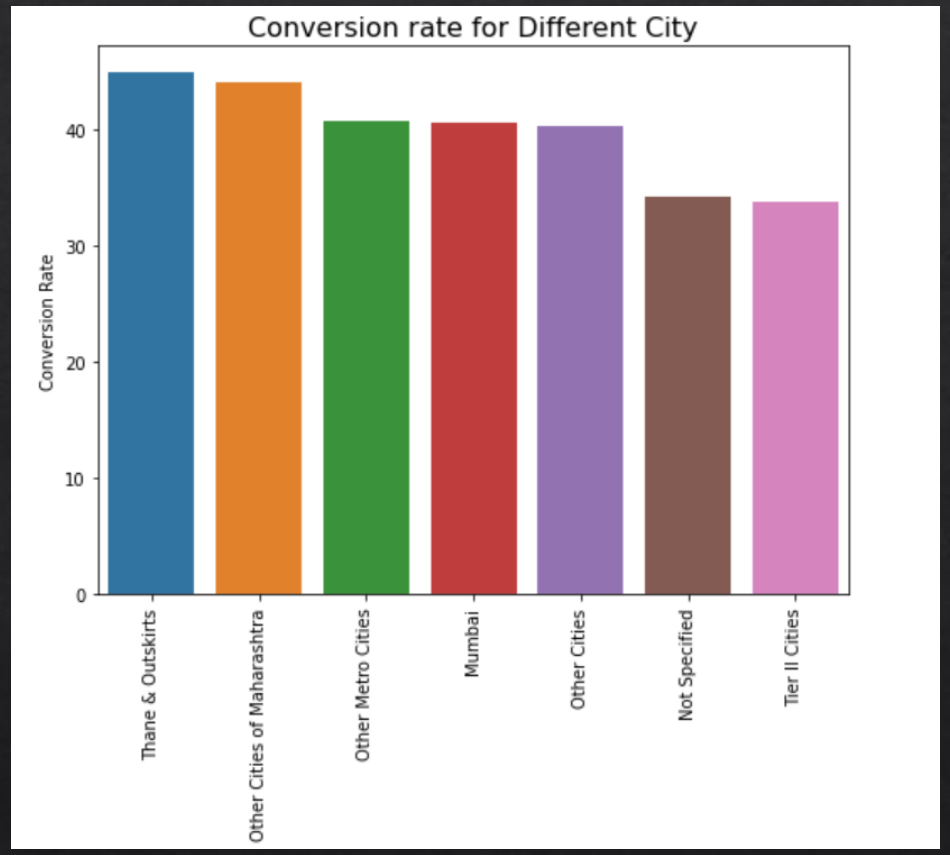
CITY

```
df["City"].value_counts(dropna = False)
```

NaN	3669
Mumbai	3222
Thane & Outskirts	752
Other Cities	686
Other Cities of Maharashtra	457
Other Metro Cities	380
Tier II Cities	74

Name: City, dtype: int64

As the number of missing values are more than the mode, it wouldn't be appropriate to replace them with the mode and hence we replaced the missing values with "Not Specified"




```
conversion_percent("City")
```

Thane & Outskirts	44.946809
Other Cities of Maharashtra	43.982495
Other Metro Cities	40.789474
Mumbai	40.626940
Other Cities	40.233236
Not Specified	34.260016
Tier II Cities	33.783784

Name: City, dtype: float64

The conversion rate for all of the cities is above 30%, however, conversion rate for Thane and Outskirts is 45 % while conversion rate for Tier 2 cities is less than 35%

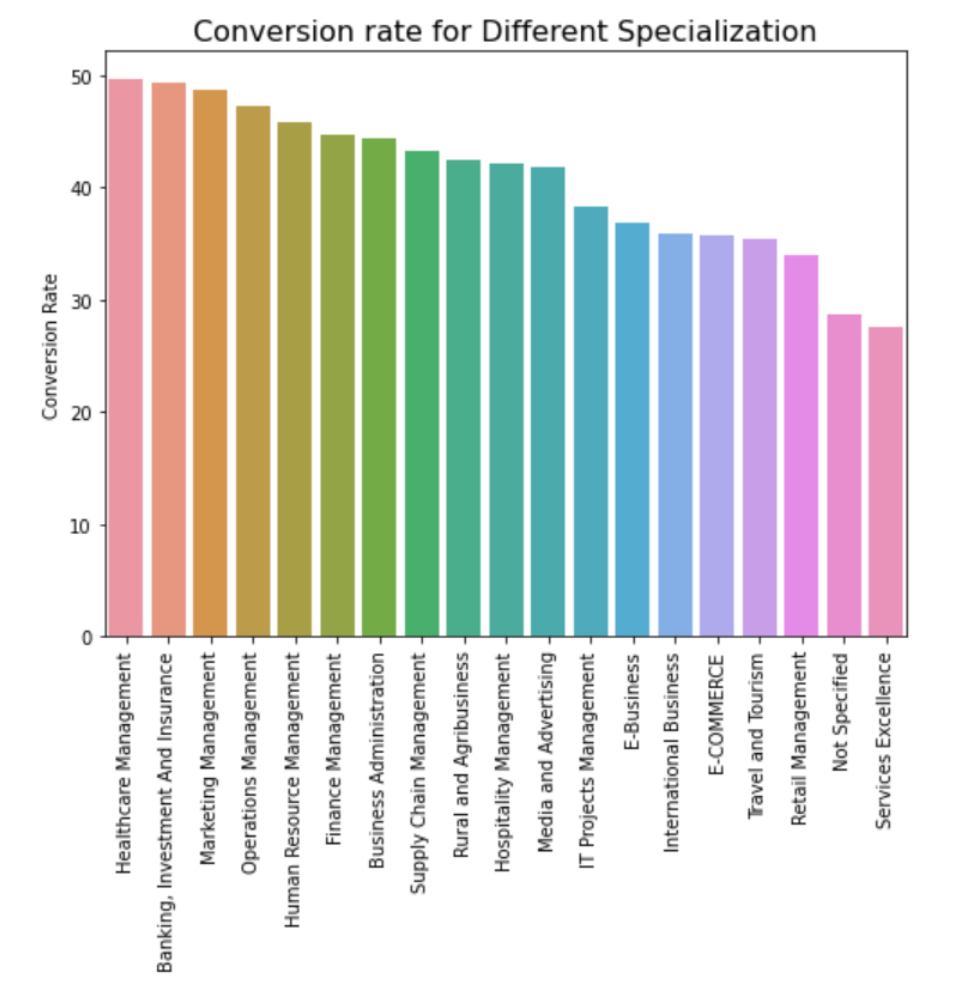
SPECIALIZATION

```
df.Specialization.value_counts(dropna = False)
```

NaN	3380
Finance Management	976
Human Resource Management	848
Marketing Management	838
Operations Management	503
Business Administration	403
IT Projects Management	366
Supply Chain Management	349
Banking, Investment And Insurance	338
Media and Advertising	203
Travel and Tourism	203
International Business	178
Healthcare Management	159
Hospitality Management	114
E-COMMERCE	112
Retail Management	100
Rural and Agribusiness	73
E-Business	57
Services Excellence	40

Name: Specialization, dtype: int64

As the number of missing values are more than the mode, it wouldn't be appropriate to replace them with the mode and hence we'll replace the missing values with "Not Specified"



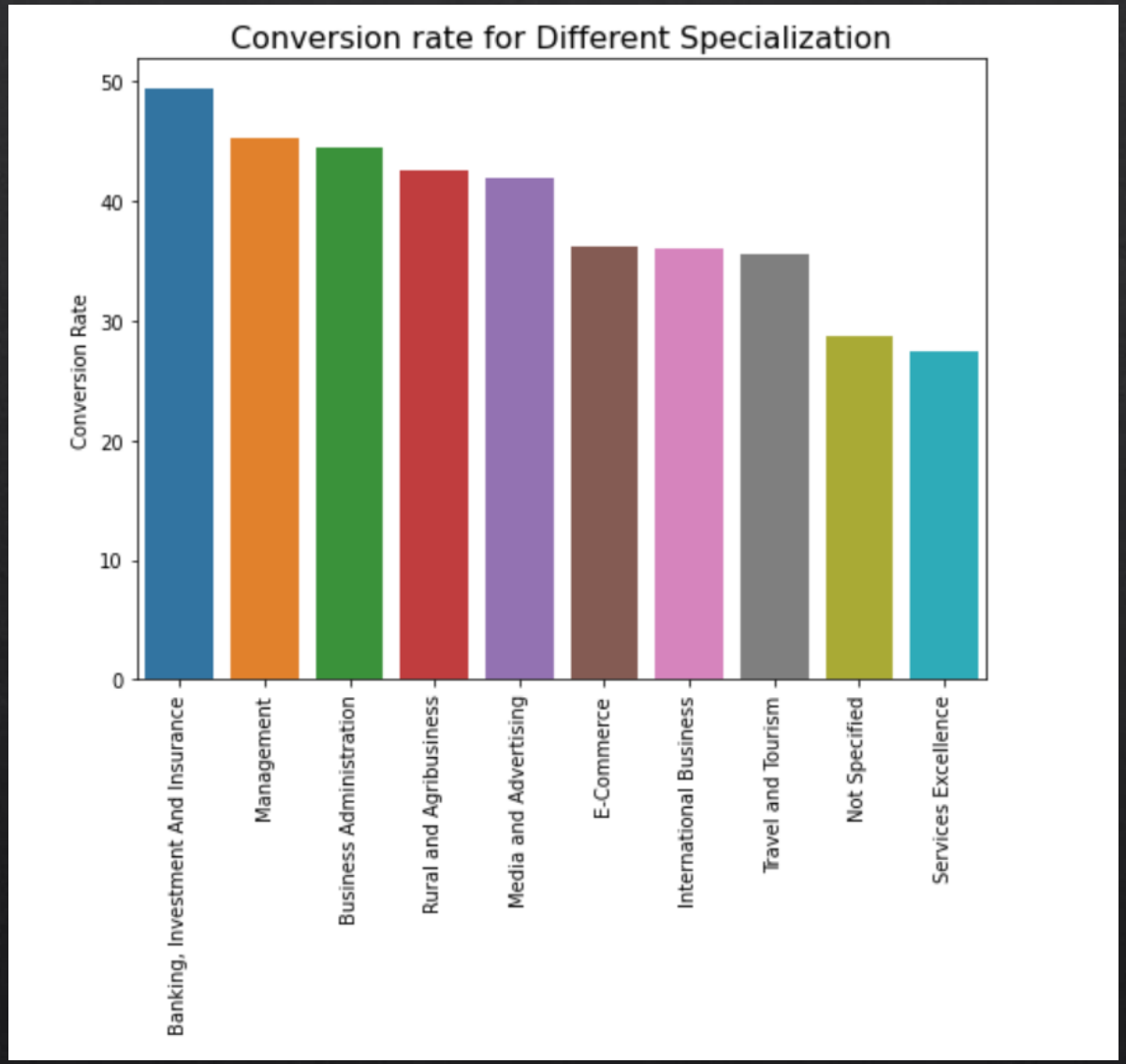

```
conversion_percent("Specialization")
```

Healthcare Management	49.685535
Banking, Investment And Insurance	49.408284
Marketing Management	48.687351
Operations Management	47.316103
Human Resource Management	45.754717
Finance Management	44.672131
Business Administration	44.416873
Supply Chain Management	43.266476
Rural and Agribusiness	42.465753
Hospitality Management	42.105263
Media and Advertising	41.871921
IT Projects Management	38.251366
E-Business	36.842105
International Business	35.955056
E-COMMERCE	35.714286
Travel and Tourism	35.467980
Retail Management	34.000000
Not Specified	28.668639
Services Excellence	27.500000

Name: Specialization, dtype: float64

So many levels make our model complex. So, we clubbed

1) Management Specializations into a single level: Management
2) E-commerce and E-business into E-commerce.



After clubbing the levels, we see a clearer picture. Banking and Management have a conversion rate of nearly 50% whereas Services Excellence has a conversion rate less than 30%.

TAGS

```
df.Tags.value_counts(dropna = False)
```

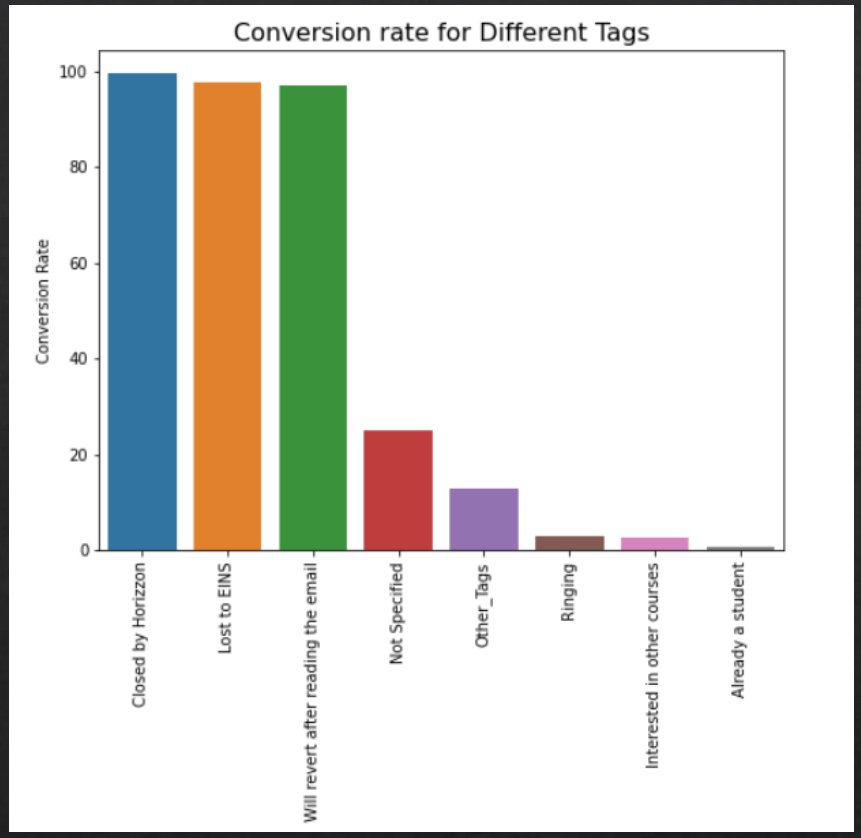
NaN	3353
Will revert after reading the email	2072
Ringing	1203
Interested in other courses	513
Already a student	465
Closed by Horizzon	358
switched off	240
Busy	186
Lost to EINS	175
Not doing further education	145
Interested in full time MBA	117
Graduation in progress	111
invalid number	83
Diploma holder (Not Eligible)	63
wrong number given	47
opp hangup	33
number not provided	27
in touch with EINS	12
Lost to Others	7
Still Thinking	6
Want to take admission but has financial problems	6
Interested in Next batch	5
In confusion whether part time or DLP	5
Lateral student	3
University not recognized	2
Shall take in the next coming month	2
Recognition issue (DEC approval)	1

Name: Tags, dtype: int64

As the number of missing values are more than the mode, it wouldn't be appropriate to replace them with the mode and hence we replaced the missing values with "Not Specified"

Will revert after reading the email	0.644307
Closed by Horizzon	0.251229
Lost to EINS	0.168937
Busy	0.052753
Interested in Next batch	0.029384
Lateral student	0.022759
Shall take in the next coming month	0.003465
Want to take admission but has financial problems	-0.002726
Recognition issue (DEC approval)	-0.008238
In confusion whether part time or DLP	-0.008863
in touch with EINS	-0.010032
Still Thinking	-0.011456
University not recognized	-0.011651
Lost to Others	-0.021804
opp hangup	-0.036225
number not provided	-0.042868
wrong number given	-0.056620
Diploma holder (Not Eligible)	-0.062908
Graduation in progress	-0.073029
invalid number	-0.073033
Interested in full time MBA	-0.083709
Not doing further education	-0.098195
switched off	-0.123718
Already a student	-0.179234
Interested in other courses	-0.179365
Not Specified	-0.210985
Ringing	-0.283895

We can see that multiple levels share similar relationship with conversions, we clubed these levels.



This looks very interesting. Three categories here have conversion rate more than 95% whereas other categories have a very low conversion. This seems like an important variable.

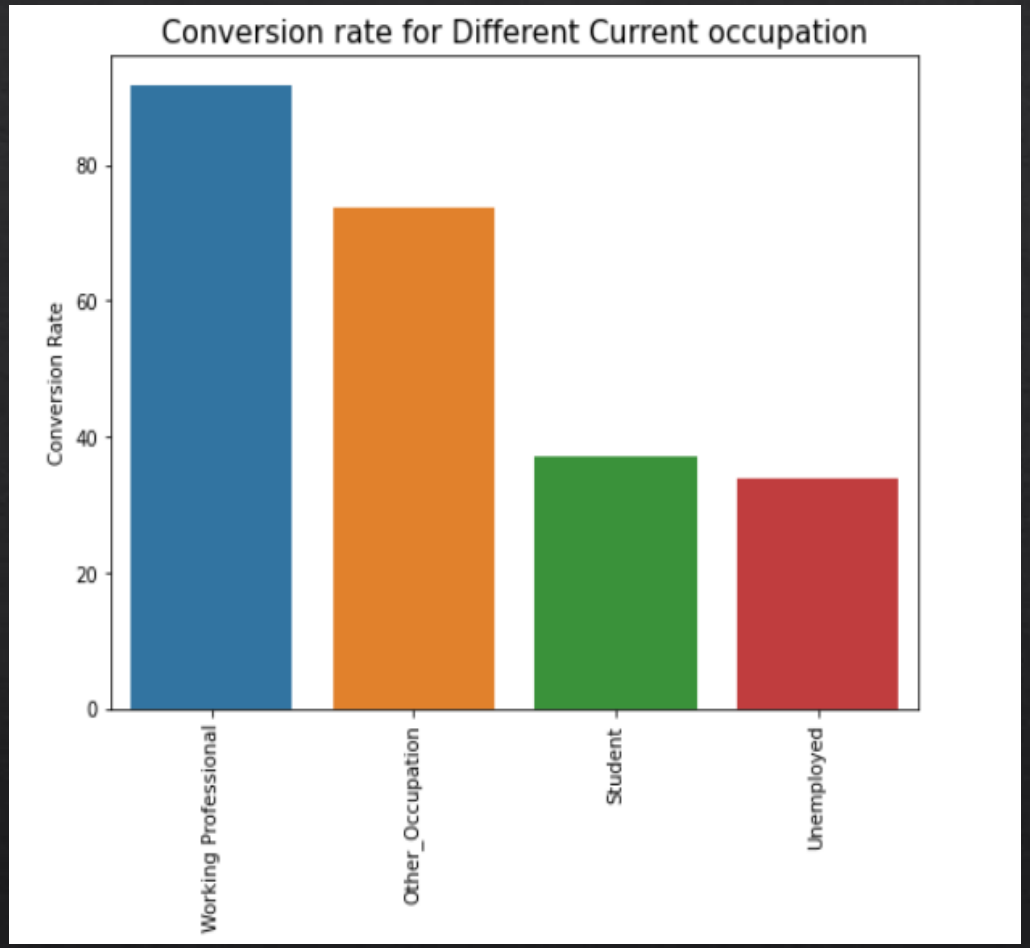
CURRENT OCCUPATION

```
df["Current occupation"].value_counts(dropna = False)
```

```
Unemployed              5600
NaN                     2690
Working Professional     706
Student                  210
Other                     16
Housewife                 10
Businessman                8
Name: Current occupation, dtype: int64
```

We imputed null values using mode, i.e. “Unemployed”.

We also grouped Housewife, other and Businessman as they have very less entries.




```
conversion_percent("Current occupation")
```

```
Working Professional    91.643059
Other_Occupation        73.529412
Student                 37.142857
Unemployed              33.908323
Name: Current occupation, dtype: float64
```

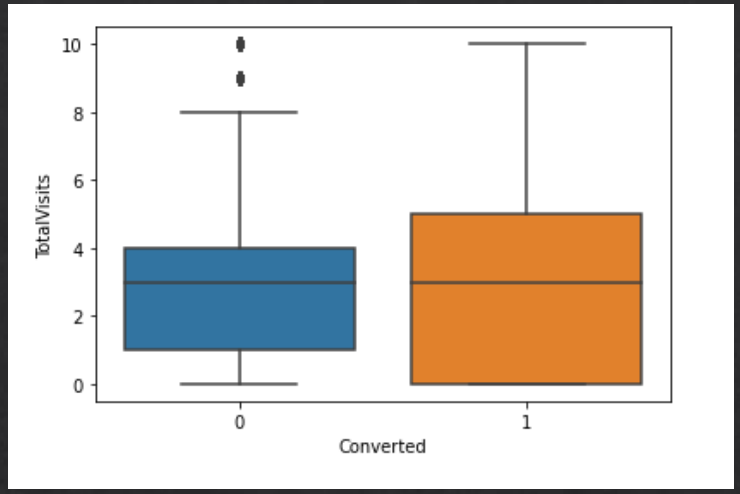
We see here that the conversion rate is as high as 91% where as it is as low as 33% for unemployed.

TOTAL VISITS

```
df.TotalVisits.describe()
```

```
count    9103.000000
mean        3.445238
std         4.854853
min         0.000000
25%         1.000000
50%         3.000000
75%         5.000000
max       251.000000
Name: TotalVisits, dtype: float64
```

Null Values → Median
Outliers → Capping(95th Percentile)



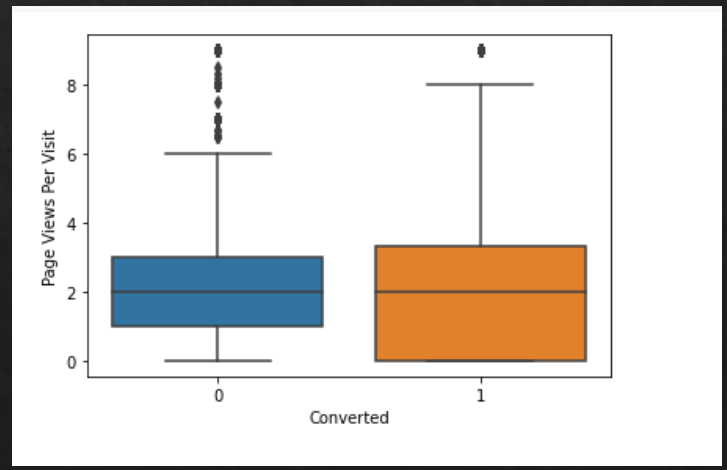
- Median almost same.
- 75th Percentile higher for Converted.

PAGE VIEWS PER PAGE

```
df['Page Views Per Visit'].describe()
```

```
count    9103.000000
mean        2.362820
std         2.161418
min         0.000000
25%         1.000000
50%         2.000000
75%         3.000000
max        55.000000
Name: Page Views Per Visit, dtype: float64
```

Null Values → Median
Outliers → Capping(99th Percentile)

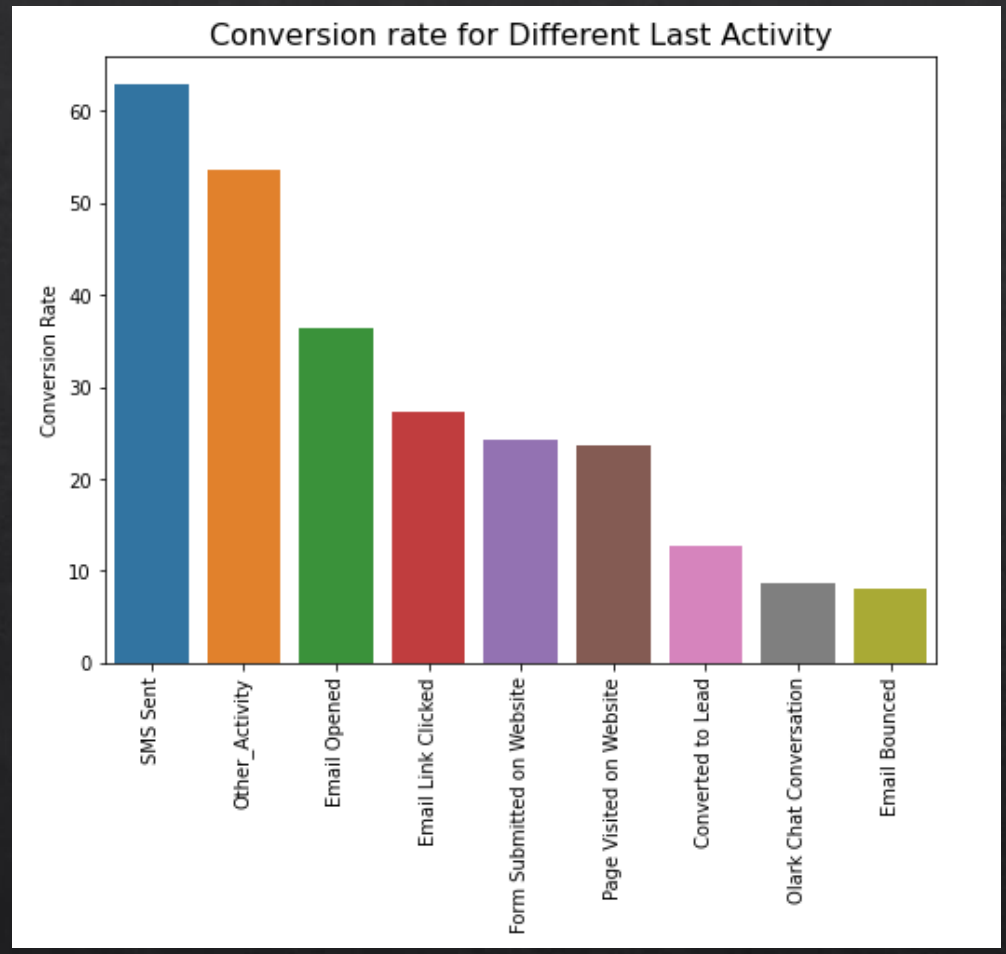


- Median almost same.
- 75th Percentile higher for Converted.

LAST ACTIVITY

Email Opened	3437
SMS Sent	2745
Olark Chat Conversation	973
Page Visited on Website	640
Converted to Lead	428
Email Bounced	326
Email Link Clicked	267
Form Submitted on Website	116
NaN	103
Unreachable	93
Unsubscribed	61
Had a Phone Conversation	30
Approached upfront	9
View in browser link Clicked	6
Email Marked Spam	2
Email Received	2
Resubscribed to emails	1
Visited Booth in Tradeshow	1

Name: Last Activity, dtype: int64



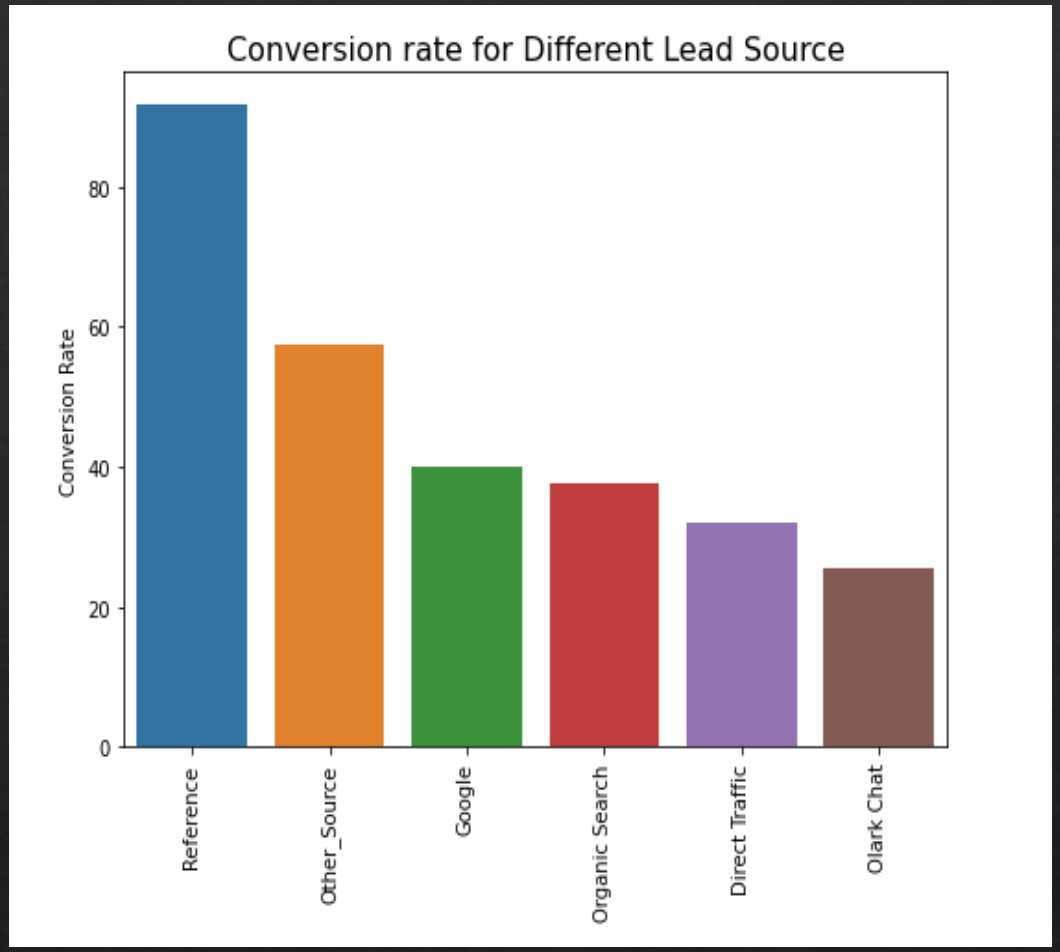
The conversion rate is max for leads with last Activity SMS sent and is significantly low for when a lead's email is bounced.

Replaced the null value and club other less frequency values in a level called "Others“.

LEAD SOURCE

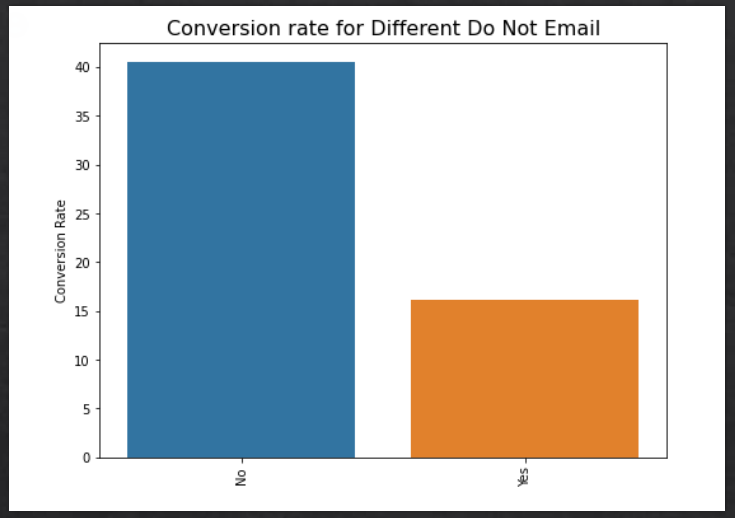
Google	2868
Direct Traffic	2543
Olark Chat	1755
Organic Search	1154
Reference	534
Welingak Website	142
Referral Sites	125
Facebook	55
NaN	36
bing	6
google	5
Click2call	4
Social Media	2
Live Chat	2
Press_Release	2
blog	1
welearnblog_Home	1
WeLearn	1
testone	1
Pay per Click Ads	1
youtubechannel	1
NC_EDM	1

Name: Lead Source, dtype: int64



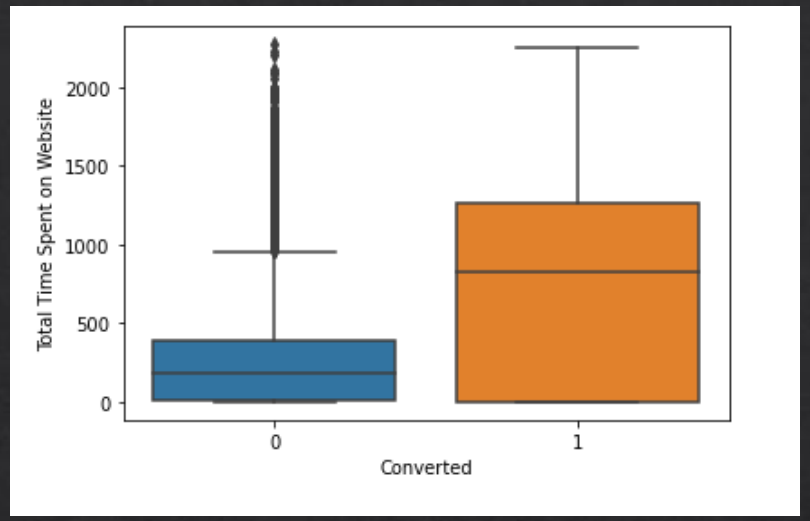
Replaced all the less frequency values with Other_Source

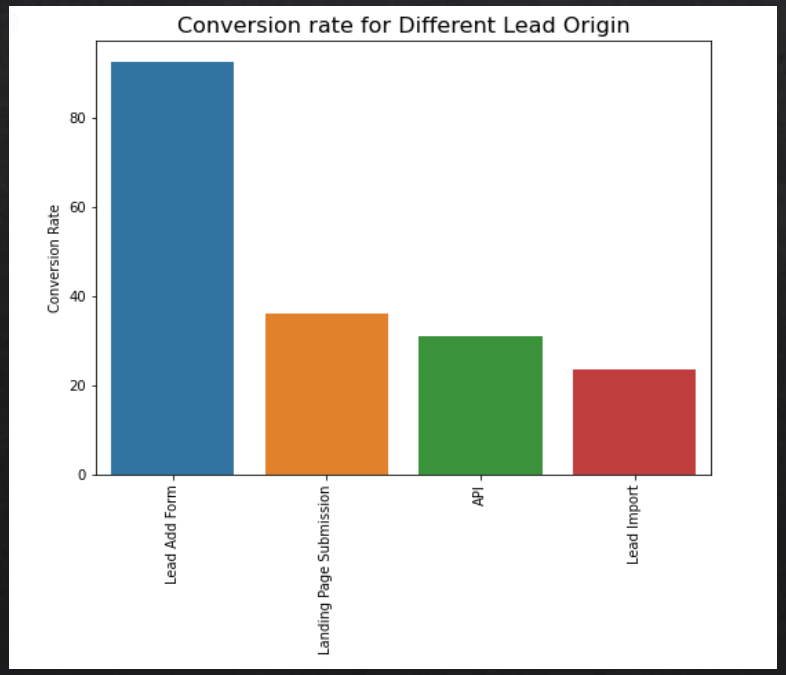
Conversion rate for people who come in through reference have a very high conversion rate and conversion rate for those who come through Olark Chat is very less.



Customers who didn't choose the option "Do not Email"

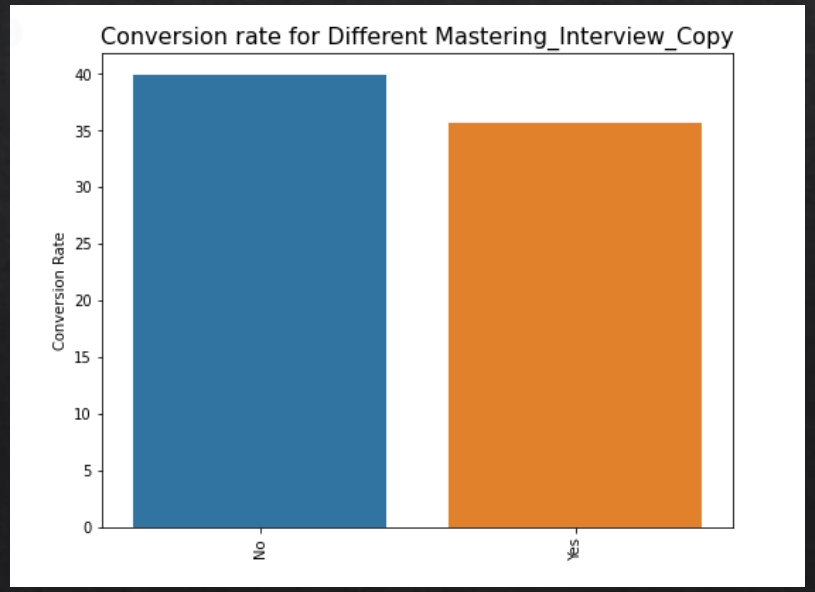
Leads spending more time on the website are more likely to be converted.





Conversion Rate is max for Lead Add Form and lowest for Lead Impact.

Customers who didn't want a free copy of mastering the interview have a higher rate of conversion which would have been expected to be other way round.



Columns which were dropped…

- Why did you choose the course? : All the leads chose better career prospects
- Country : Almost 95% of leads were from India.
- Last Notable Activity : Very similar to Last Activity
- Search
- Newspaper Article
- X Education Forums
- Digital Advertisement
- Newspaper

Majority of Responses were No

Building Our Model

(LOGISTIC REGRESSION MODEL)

STEP 1 : Used RFE to decide top 20 most important variables. Build a LOGISTIC REGRESSION model using these 20 variables.

Summary

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6467
Model:	GLM	**Df Residuals:**	6446
Model Family:	Binomial	**Df Model:**	20
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1356.2
Date:	Sun, 07 Mar 2021	**Deviance:**	2712.4
Time:	00:43:20	**Pearson chi2:**	1.05e+04
No. Iterations:	24		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2336	0.236	-5.235	0.000	-1.696	-0.772
Do Not Email	-0.9276	0.250	-3.712	0.000	-1.417	-0.438
Total Time Spent on Website	4.2924	0.238	18.071	0.000	3.827	4.758
Null_Count	-1.9010	0.313	-6.080	0.000	-2.514	-1.288
Landing Page Submission	-0.8951	0.134	-6.699	0.000	-1.157	-0.633
Lead Add Form	4.1544	0.422	9.849	0.000	3.328	4.981
Lead Import	-0.9645	0.760	-1.269	0.204	-2.454	0.525
Olark Chat	0.8970	0.166	5.396	0.000	0.571	1.223
Reference	-3.3881	0.570	-5.949	0.000	-4.504	-2.272
Converted to Lead	-1.3958	0.379	-3.682	0.000	-2.139	-0.653
Email Bounced	-0.9919	0.494	-2.010	0.044	-1.959	-0.025
Olark Chat Conversation	-1.4151	0.233	-6.068	0.000	-1.872	-0.958
SMS Sent	1.8947	0.112	16.990	0.000	1.676	2.113
Working Professional	0.7947	0.367	2.168	0.030	0.076	1.513
Already a student	-4.4932	1.020	-4.407	0.000	-6.492	-2.495
Closed by Horizzon	26.1612	1.21e+04	0.002	0.998	-2.37e+04	2.38e+04
Interested in other courses	-2.7412	0.369	-7.424	0.000	-3.465	-2.018
Lost to EINS	5.5243	0.737	7.498	0.000	4.080	6.968
Other_Tags	-1.3815	0.168	-8.203	0.000	-1.712	-1.051
Ringing	-3.5239	0.249	-14.127	0.000	-4.013	-3.035
Will revert after reading the email	3.9829	0.218	18.277	0.000	3.556	4.410

VIF

	Features	VIF
4	Lead Add Form	4.71
7	Reference	4.42
2	Null_Count	3.48
3	Landing Page Submission	2.88
6	Olark Chat	2.37
1	Total Time Spent on Website	2.33
19	Will revert after reading the email	2.27
0	Do Not Email	1.86
9	Email Bounced	1.78
11	SMS Sent	1.75
10	Olark Chat Conversation	1.47
14	Closed by Horizzon	1.43
12	Working Professional	1.36
18	Ringing	1.33
17	Other_Tags	1.31
13	Already a student	1.13
8	Converted to Lead	1.12
15	Interested in other courses	1.11
16	Lost to EINS	1.08
5	Lead Import	1.02

We decided to drop Closed by Horizzon because of it's high p-value.

STEP 2 : Built a second model after dropping “Closed by Horizzon” from Model 1.

Summary

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6467
Model:	GLM	**Df Residuals:**	6447
Model Family:	Binomial	**Df Model:**	19
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1475.4
Date:	Sun, 07 Mar 2021	**Deviance:**	2950.7
Time:	00:43:26	**Pearson chi2:**	2.01e+04
No. Iterations:	9		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.2581	0.209	-1.233	0.218	-0.668	0.152
Do Not Email	-0.9838	0.246	-4.007	0.000	-1.465	-0.503
Total Time Spent on Website	4.3126	0.228	18.953	0.000	3.867	4.759
Null_Count	-3.1492	0.284	-11.085	0.000	-3.706	-2.592
Landing Page Submission	-0.8577	0.129	-6.650	0.000	-1.110	-0.605
Lead Add Form	4.2235	0.419	10.081	0.000	3.402	5.045
Lead Import	-1.0325	0.748	-1.380	0.168	-2.499	0.434
Olark Chat	0.9602	0.162	5.932	0.000	0.643	1.278
Reference	-1.1816	0.482	-2.450	0.014	-2.127	-0.236
Converted to Lead	-1.5080	0.366	-4.120	0.000	-2.225	-0.791
Email Bounced	-0.8868	0.451	-1.968	0.049	-1.770	-0.003
Olark Chat Conversation	-1.4076	0.216	-6.520	0.000	-1.831	-0.985
SMS Sent	1.7127	0.109	15.779	0.000	1.500	1.925
Working Professional	1.5412	0.316	4.884	0.000	0.923	2.160
Already a student	-5.1777	1.022	-5.067	0.000	-7.181	-3.175
Interested in other courses	-3.6000	0.374	-9.621	0.000	-4.333	-2.867
Lost to EINS	4.9980	0.735	6.801	0.000	3.558	6.438
Other_Tags	-2.0133	0.162	-12.453	0.000	-2.330	-1.696
Ringing	-4.1944	0.249	-16.854	0.000	-4.682	-3.707
Will revert after reading the email	3.1370	0.196	16.032	0.000	2.754	3.521

VIF

	Features	VIF
4	Lead Add Form	4.59
7	Reference	4.32
2	Null_Count	3.44
3	Landing Page Submission	2.80
6	Olark Chat	2.32
1	Total Time Spent on Website	2.17
18	Will revert after reading the email	2.03
0	Do Not Email	1.86
9	Email Bounced	1.78
11	SMS Sent	1.74
10	Olark Chat Conversation	1.47
12	Working Professional	1.33
17	Ringing	1.31
16	Other_Tags	1.29
8	Converted to Lead	1.12
13	Already a student	1.12
14	Interested in other courses	1.10
15	Lost to EINS	1.06
5	Lead Import	1.02

We decided to drop Lead Import because of it’s high p-value.

STEP 3 : Built a third model after dropping “Lead Score” from Model 2.

Summary

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6467
Model:	GLM	**Df Residuals:**	6448
Model Family:	Binomial	**Df Model:**	18
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1476.4
Date:	Sun, 07 Mar 2021	**Deviance:**	2952.8
Time:	00:44:32	**Pearson chi2:**	2.00e+04
No. Iterations:	9		
Covariance Type:	nonrobust		

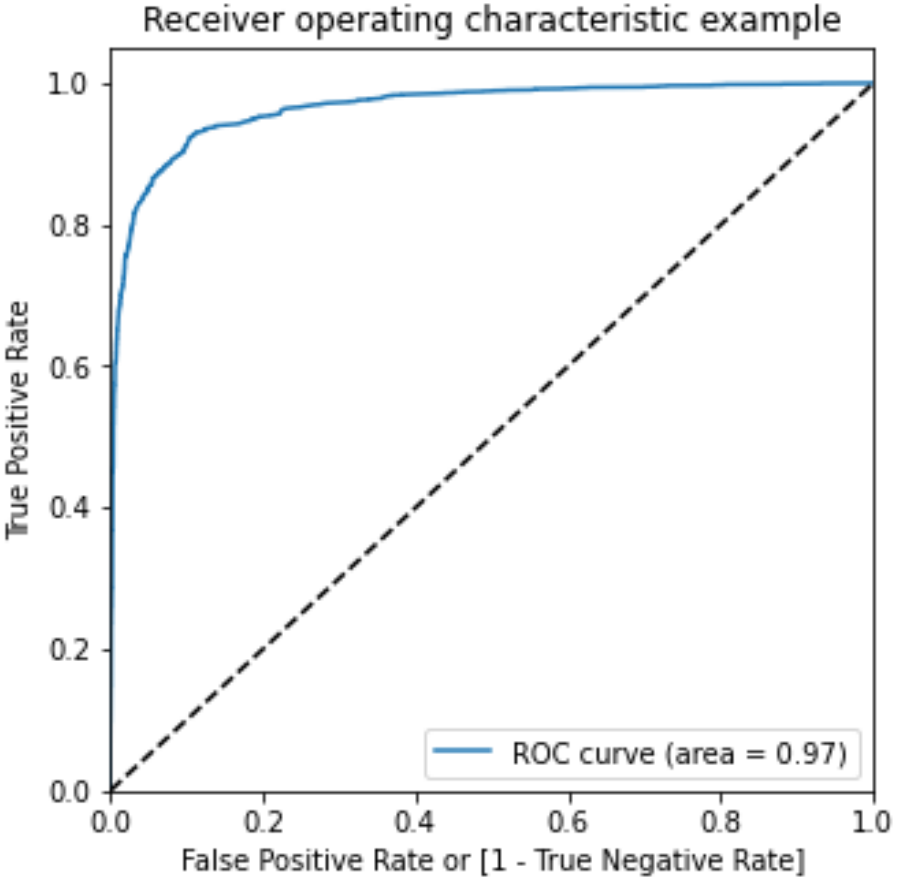
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.2913	0.208	-1.401	0.161	-0.699	0.116
Do Not Email	-0.9794	0.245	-3.989	0.000	-1.461	-0.498
Total Time Spent on Website	4.3336	0.227	19.069	0.000	3.888	4.779
Null_Count	-3.1389	0.284	-11.060	0.000	-3.695	-2.583
Landing Page Submission	-0.8344	0.128	-6.521	0.000	-1.085	-0.584
Lead Add Form	4.2499	0.419	10.153	0.000	3.430	5.070
Olark Chat	0.9865	0.161	6.128	0.000	0.671	1.302
Reference	-1.1817	0.482	-2.450	0.014	-2.127	-0.236
Converted to Lead	-1.5020	0.366	-4.103	0.000	-2.219	-0.784
Email Bounced	-0.8910	0.451	-1.977	0.048	-1.774	-0.008
Olark Chat Conversation	-1.4034	0.216	-6.504	0.000	-1.826	-0.981
SMS Sent	1.7063	0.108	15.753	0.000	1.494	1.919
Working Professional	1.5396	0.316	4.877	0.000	0.921	2.158
Already a student	-5.1770	1.022	-5.065	0.000	-7.180	-3.174
Interested in other courses	-3.5912	0.374	-9.601	0.000	-4.324	-2.858
Lost to EINS	4.9985	0.735	6.802	0.000	3.558	6.439
Other_Tags	-2.0075	0.162	-12.429	0.000	-2.324	-1.691
Ringing	-4.1901	0.249	-16.847	0.000	-4.678	-3.703
Will revert after reading the email	3.1375	0.196	16.039	0.000	2.754	3.521

VIF

	Features	VIF
4	Lead Add Form	4.59
6	Reference	4.32
2	Null_Count	3.40
3	Landing Page Submission	2.77
5	Olark Chat	2.31
1	Total Time Spent on Website	2.17
17	Will revert after reading the email	2.02
0	Do Not Email	1.86
8	Email Bounced	1.78
10	SMS Sent	1.74
9	Olark Chat Conversation	1.47
11	Working Professional	1.33
16	Ringing	1.29
15	Other_Tags	1.28
7	Converted to Lead	1.12
12	Already a student	1.12
13	Interested in other courses	1.10
14	Lost to EINS	1.06

This model looks good, as there seems to be VERY LOW Multicollinearity between the predictors and the p-values for all the predictors seems to be significant.

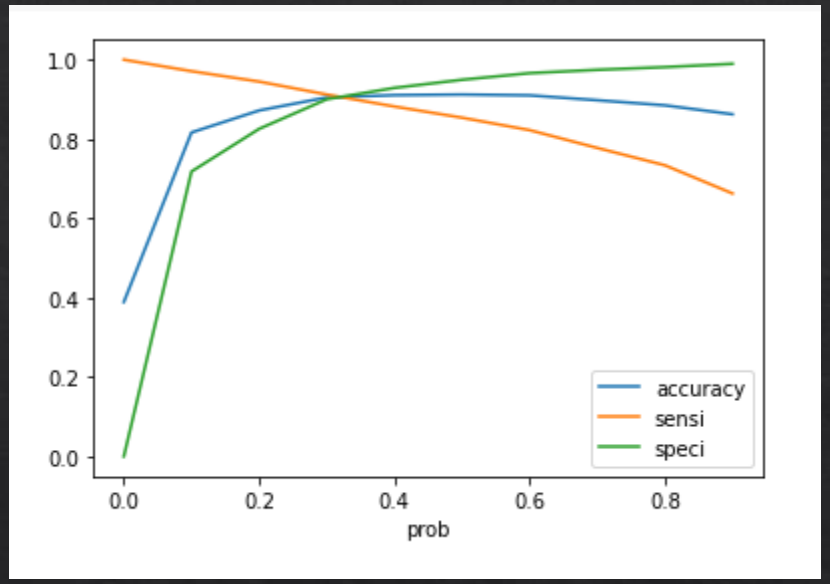
ROC Curve



The curve looks good as it is in the upper left corner, it shows that our model is a good fit. The area under the curve is 0.97 which also depicts the same.

Finding the optimal cut off probability…

	prob	accuracy	sensi	speci
0.0	0.0	0.389207	1.000000	0.000000
0.1	0.1	0.816453	0.970997	0.717975
0.2	0.2	0.871965	0.944776	0.825570
0.3	0.3	0.904902	0.912197	0.900253
0.4	0.4	0.910623	0.882002	0.928861
0.5	0.5	0.912169	0.853794	0.949367
0.6	0.6	0.910159	0.822408	0.966076
0.7	0.7	0.897789	0.777513	0.974430
0.8	0.8	0.884954	0.733810	0.981266
0.9	0.9	0.862378	0.662694	0.989620



We Calculated Specificity, Sensitivity and Accuracy for probabilities ranging from 0.1 till 0.9.

The optimal cut off probability was taken as 0.3 as a balance between specificity and sensitivity can be observed at prob = 0.3

FINAL PREDICTIONS..

	Lead Number	Converted	Conversion_Prob	final_predicted
0	608309	0.0	0.388386	1
1	601302	0.0	0.038130	0
2	597540	0.0	0.001135	0
3	591462	1.0	0.498530	1
4	589061	0.0	0.030497	0

	Lead Number	Lead Score
0	608309	38.838608
1	601302	3.813000
2	597540	0.113475
3	591462	49.852972
4	589061	3.049673
5	628456	0.150288
6	588058	99.979407
7	649586	11.584304
8	592599	12.067846
9	649186	92.652729

LEAD SCORE

Metrics for Train Data: -
Accuracy: 0.905
Sensitivity: 0.912
Specificity: 0.900

Metrics for Test Data: -
Accuracy: 0.901
Sensitivity: 0.906
Specificity: 0.898

The model accuracy, sensitivity and specificity, both for train and test data, are coming out to be nearly 90% which shows that the model is a good fit.

CONCLUSION

Variables that are influencing the conversion rate positively -

- Total Time Spent on Website: The more time a lead spends on the website, the more is the chance of conversion.
- Lead Add Form: Conversion rate is high when the lead origin is Lead Add Form.
- SMS Sent: Conversion rate is high when the last activity is sending SMS.
- Working Professional: Lead is most likely to be converted if he/she is a working professional.
- Conversion rate is high when Tag = Lost to EINS
- Conversion rate is high when the lead has received the email and will likely revert back.

Variables that are influencing the conversion rate negatively -

- An important variable is Null Count, i.e. the amount of information a lead is revealing is important to note as the missing information leads to less conversion rate.
- The conversion rate is negatively influenced if the lead source is Reference.
- Already a student - If the lead is a students, he/she is less likely to be converted. It seems like the courses are designed for working professionals.
- If the lead is interested in other courses, he is less likely to be converted.
- Ringing: If the leads isn't answering and ignoring (RUDE) the calls, they are less likely to be converted.