LEAD SCORING CASE STUDY

TEJASWINI KAMATH

RANJIV SUKUMARAN

THE PROBLEM STATEMENT

What is the 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. Now, although X Education gets a lot of leads, the lead conversion rate is very poor.

How will I know this problem has been solved?

- Hence X Education has appointed us to build a model to help them select the most promising leads, i.e., the leads that are most likely to convert into paying customers.
- 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%.

WORKABLE SOLUTION

Part #1

- Reading and understanding data
- Data Preparation
  1. Variable treatment
  2. Missing Value imputation
  3. Dummy variable creation
  4. Outlier Treatment
  5. Bivariate and Multivariate analysis

Part #2

- Test/Train Split
- Feature Scaling
- Model Building
- Feature selection using RFE and p-value/VIF.
- Determining model Accuracy, other metrics using confusion matrix and ROC Curve.

Part #3

- Optimal threshold determination using Sensitivity/Specificity and Precision/Recall curve. Rebuild model as per optimal threshold.
- Decide whether to go for precision/recall or sensitivity/specificity as per metrics observed.
- Running model on test data. Determine Metrics.
- Add Lead Score to Leads

DATA VISUALIZATION AND PREPARATION

1. For columns like Update me on Supply Chain Content, Get updates on DM Content, I agree to pay the amount through cheque, Magazine, Receive More Updates About Our Courses :
    - All leads have same value. Hence, using these fields will not impact the final model.
    - As a result, we will be dropping these columns.

2. For Columns with a value 'Select' like 'Specialization', 'How did you hear about X Education', 'Lead Profile'' City' :
    - This 'Select value will be converted to Np.nan(null values)
    - The figures on the right show the differences in the null values in the columns before and after conversion.
    - As you an see the column 'How did you hear about X Education', 'Lead Profile' have had a major change.

3. We drop the 'Prospect Number' column since we need to determine Lead Scores related to 'Lead Number'.
    - Hence 'Lead Number' column which will be retained.

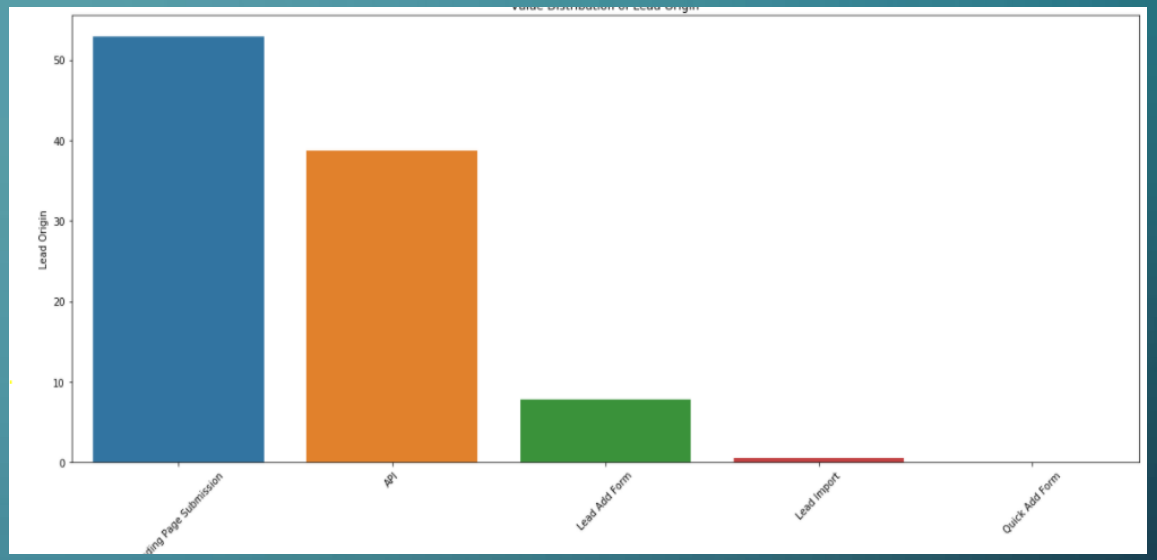
	Total	%
Lead Quality	4767	51.590909
Asymmetrique Profile Score	4218	45.649351
Asymmetrique Activity Score	4218	45.649351
Asymmetrique Profile Index	4218	45.649351
Asymmetrique Activity Index	4218	45.649351

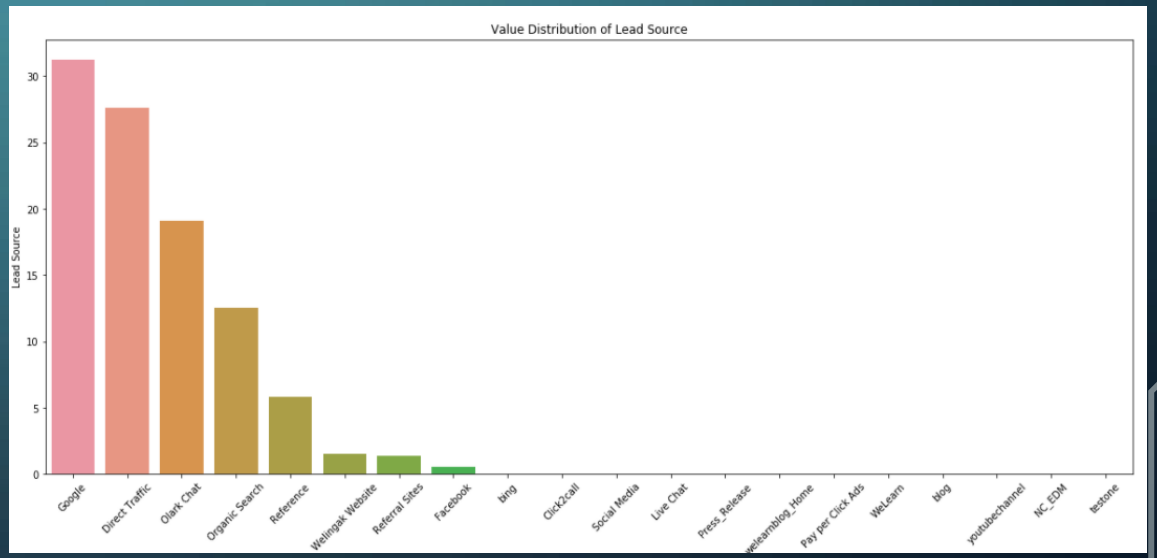
	Total	%
How did you hear about X Education	7250	78.463203
Lead Profile	6855	74.188312
Lead Quality	4767	51.590909
Asymmetrique Profile Score	4218	45.649351
Asymmetrique Activity Score	4218	45.649351
Asymmetrique Profile Index	4218	45.649351
Asymmetrique Activity Index	4218	45.649351

DATA VISUALIZATION AND PREPARATION

Categorical Columns Analysis:

- We have checked the columns in case the values are skewed. In case it's skewed the columns shall be dropped.
- Combining minimal category values to 'Others' where applicable.
- For Lead Origin, while 'Lead Add Form' is a very low percentage of 'Lead Origin', a very high percentage are conversions.
  Here it would be a good idea to spend more on advertising using Lead Form
- Similarly, for Lead Source, 'Welingak Website' and 'Reference' constitute a low percentage of Lead Source but have very high conversion rate. It might be worth looking into advertising on 'Welingak Website' and ask satisfied customers to provide references.

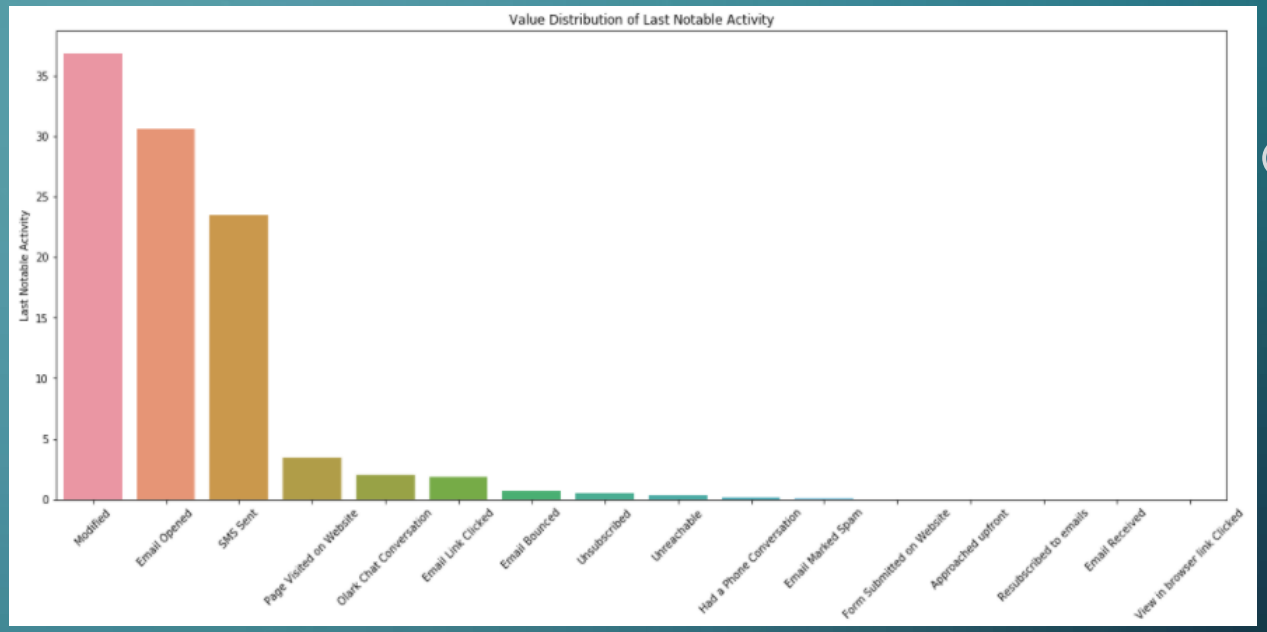


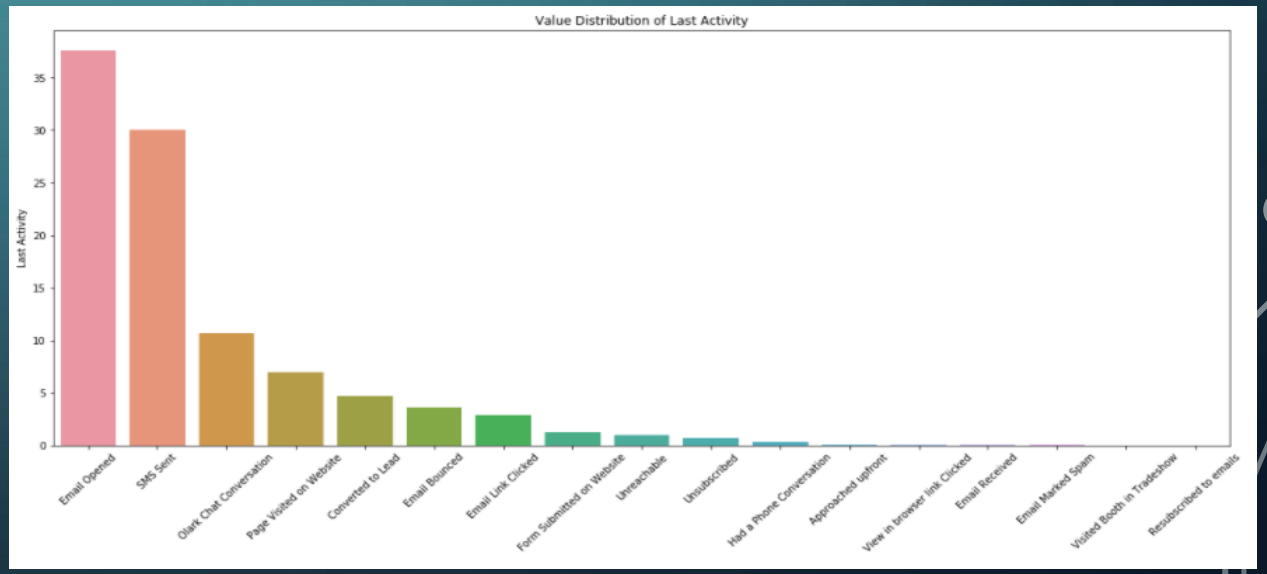


DATA VISUALIZATION AND PREPARATION

Categorical Columns Analysis:

- Looking at similar high conversion factors for other variables:
  - Last Activity -> SMS Sent
  - Last Notable Activity -> SMS Sent
    It might be a good idea to follow up via SMS
  - What is your current Occupation -> Working Professional
    Concentrating on Working Professionals might lead to more conversions
  - Tags -> Will revert after reading the email, Closed by Horizon

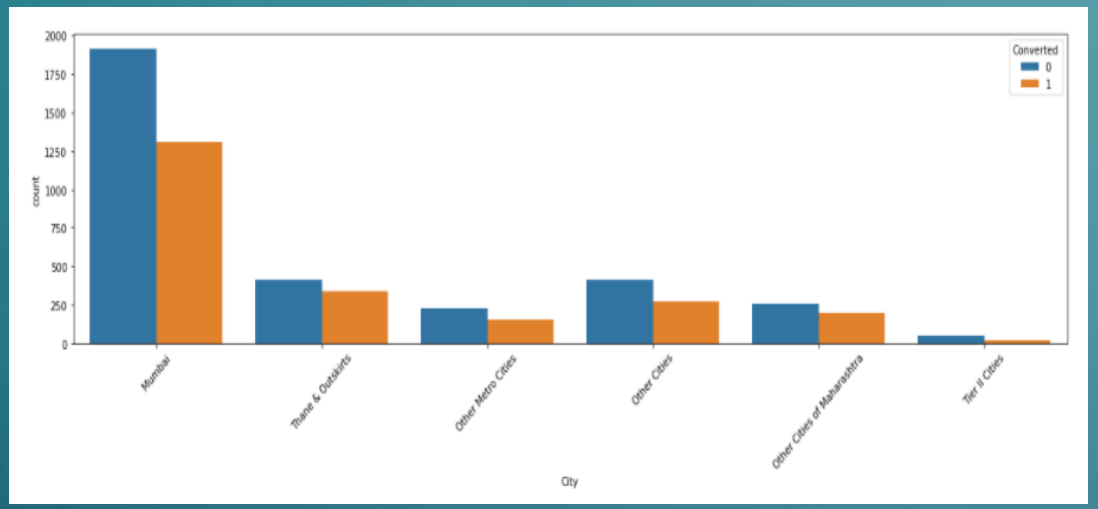


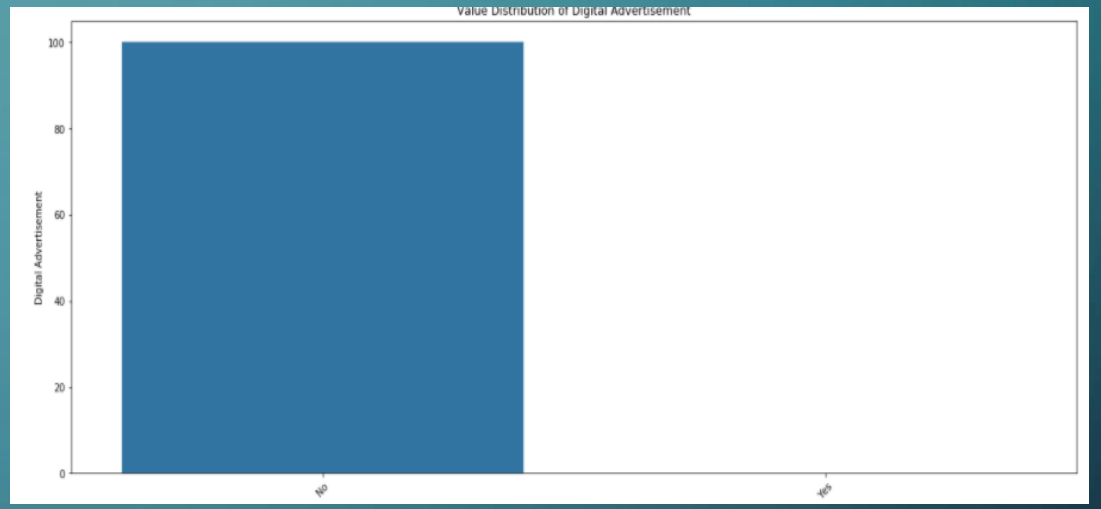


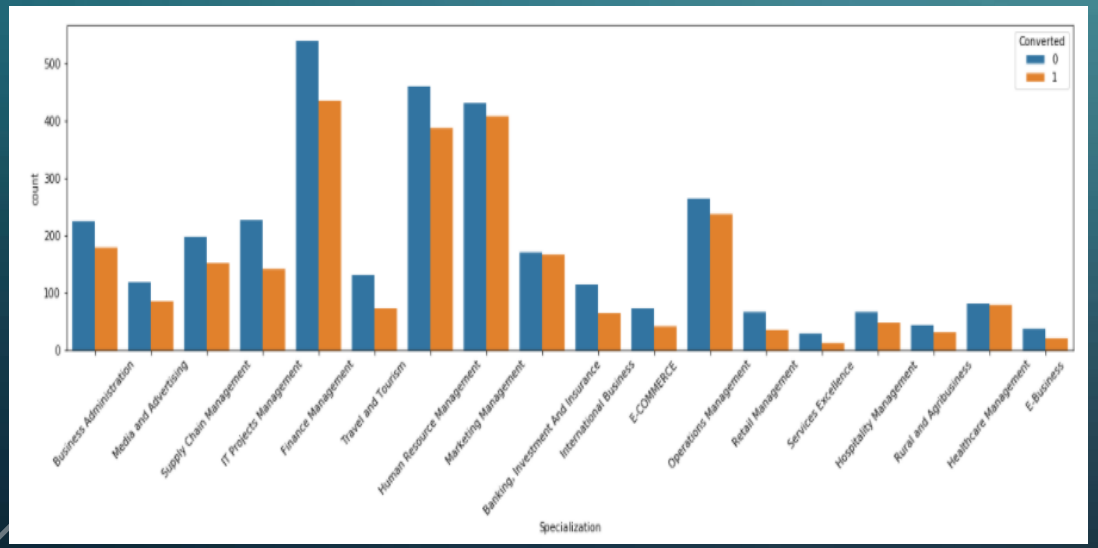
DATA VISUALIZATION AND PREPARATION

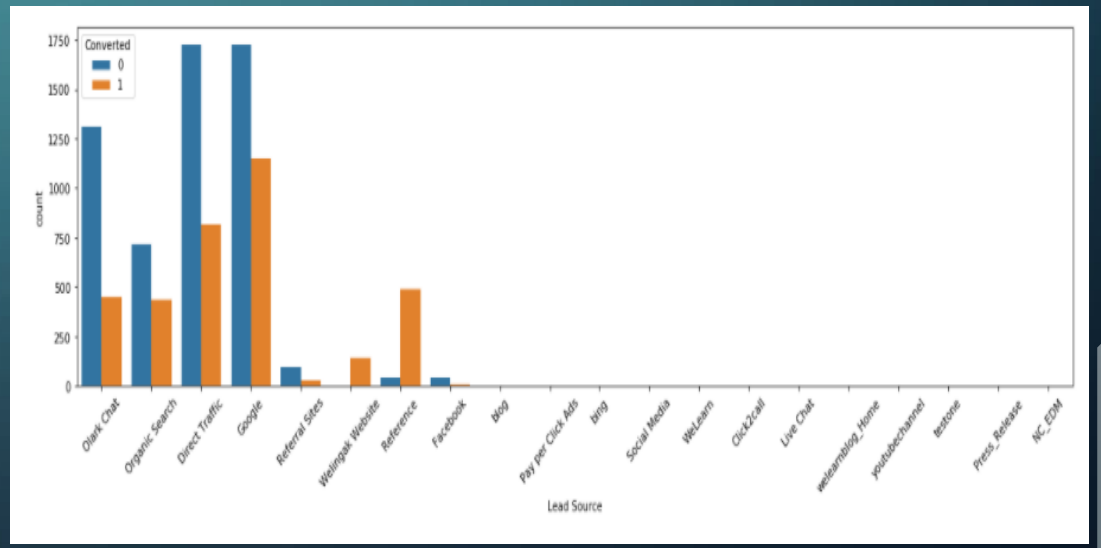
Categorical Columns Analysis (Combining minimal category values) :

- We have checked the columns in case the values are skewed. In case it's skewed the columns shall be dropped.
- Combining minimal category values to 'Others' where applicable





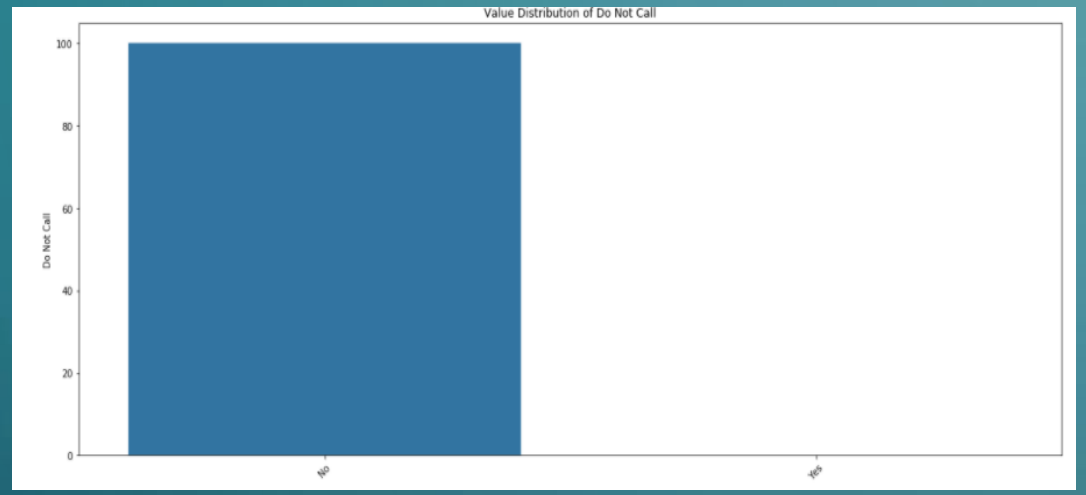


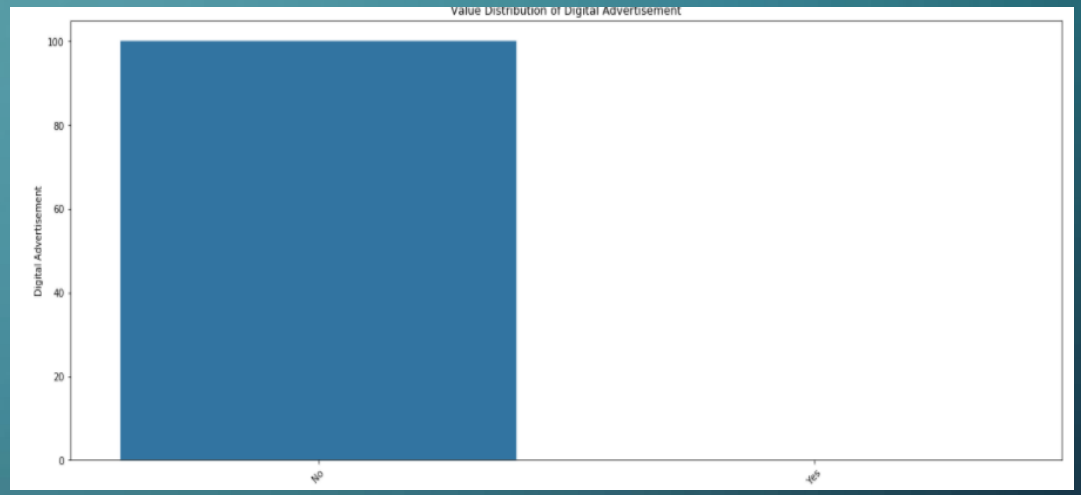


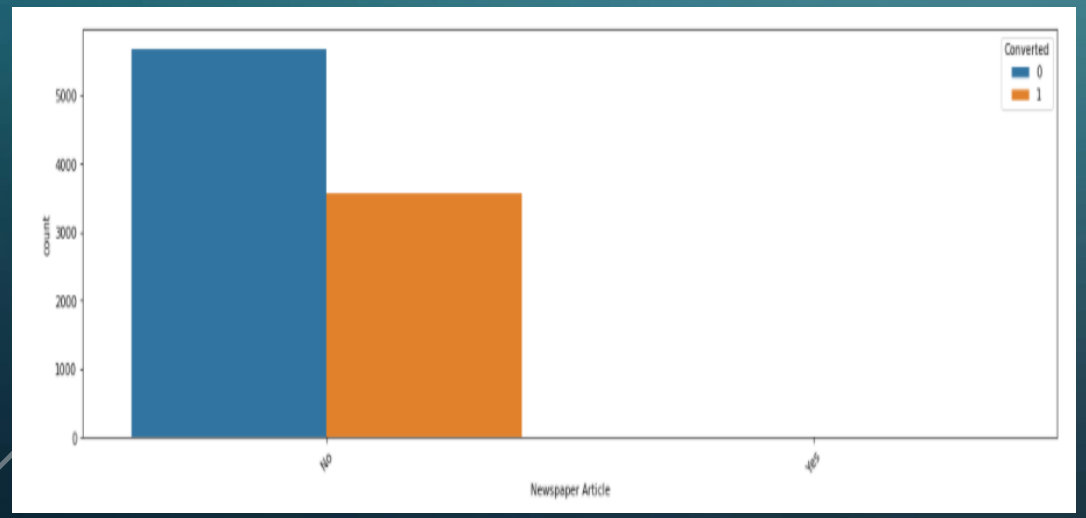
DATA VISUALIZATION AND PREPARATION

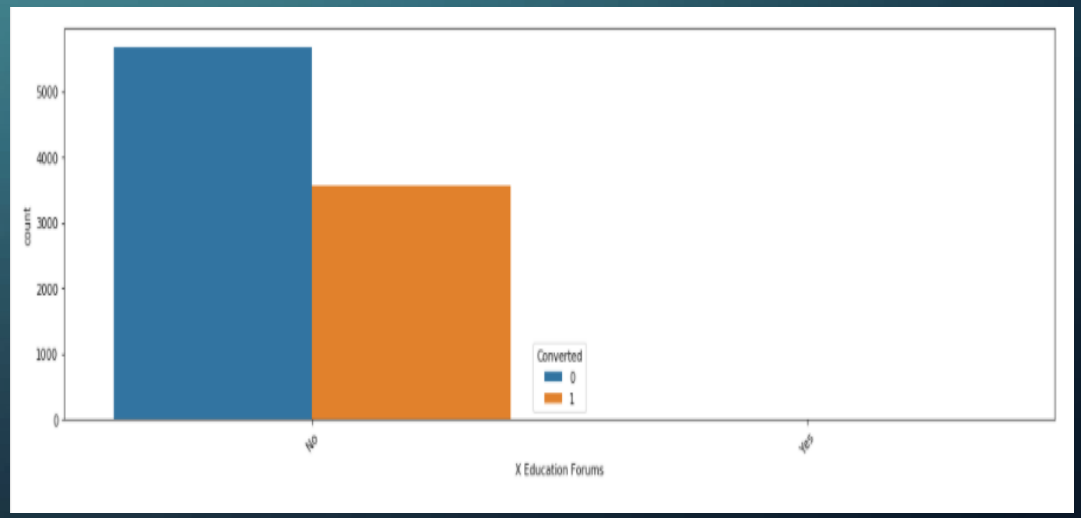
Categorical Columns Analysis (Dropping of Skewed Data) :

Columns with highly skewed data i.e. >85% in one value have been dropped. i.e. columns like 'Do Not Email', 'Do Not Call', 'Country', 'What is your current occupation', 'What matters most to you in choosing a course', 'Search', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations'. Few Sample columns shown below.





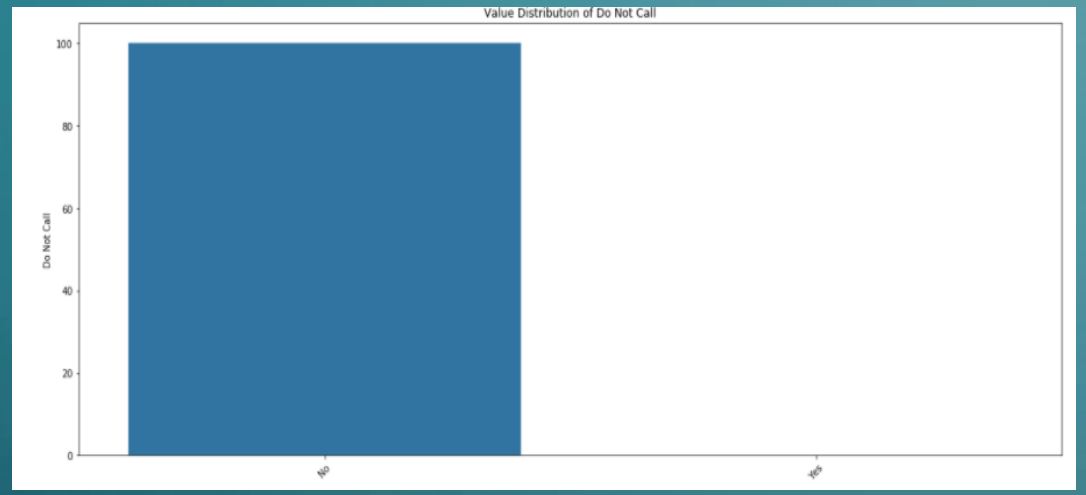


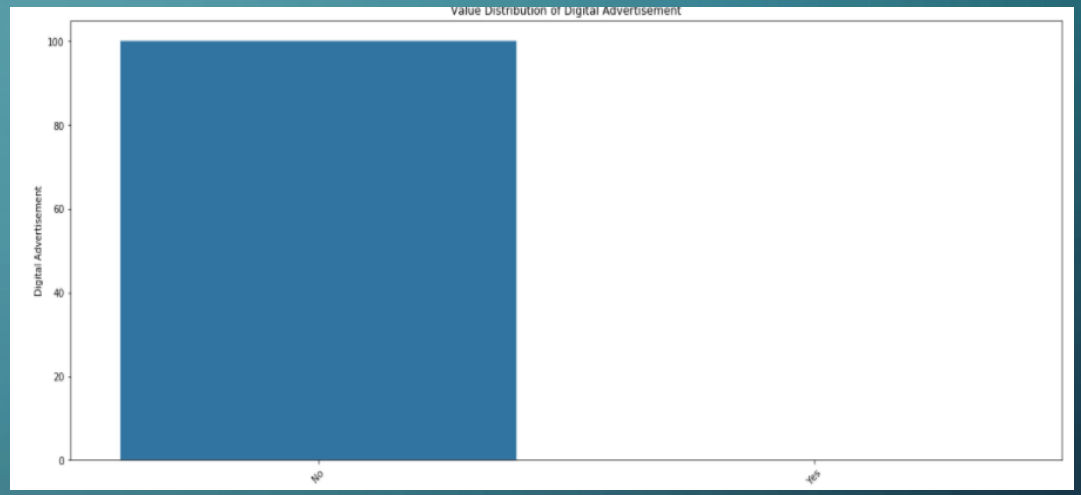


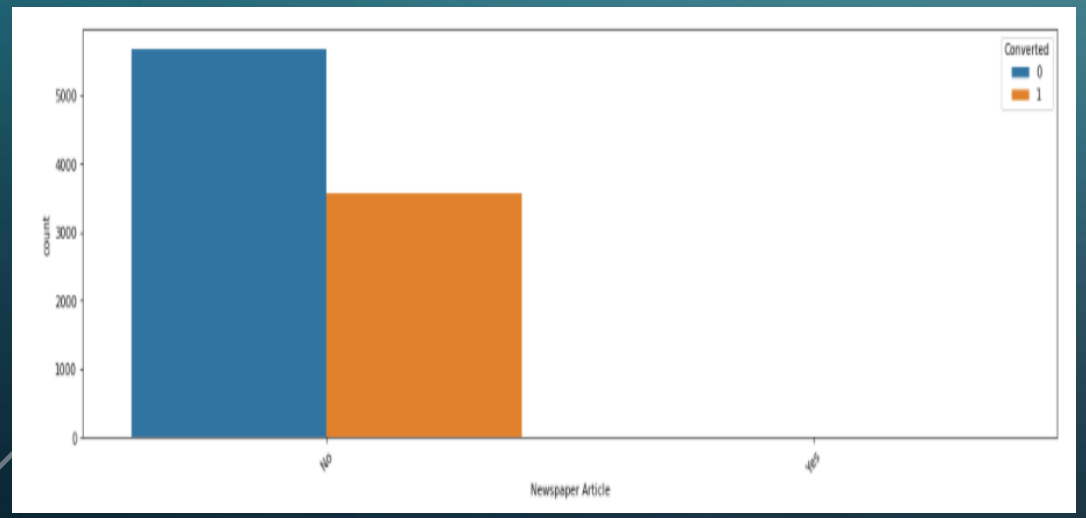
DATA VISUALIZATION AND PREPARATION

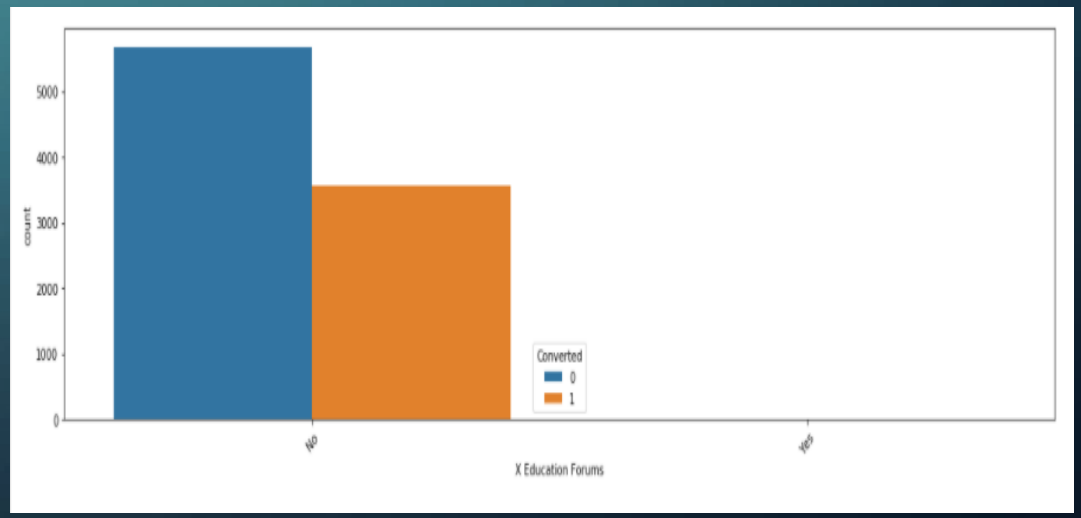
Categorical Columns Analysis (Dropping of Skewed Data) :

Columns with highly skewed data i.e. >85% in one value have been dropped. i.e. columns like 'Do Not Email', 'Do Not Call', 'Country', 'What is your current occupation', 'What matters most to you in choosing a course', 'Search', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations'. Few Sample columns shown below.









DATA VISUALIZATION AND PREPARATION

Categorical Columns Analysis (Imputing Data) :

- Imputing values for numeric and categorical columns where missing data is less than 2% by median and mode values respectively. Columns like Lead Source, Last Activity, Page Views Per Visit, TotalVisits
- Post this activity we checked for nulls again and dropped columns with high % nulls.

	Total	%
Asymmetrique Profile Score	4218	45.649351
Asymmetrique Activity Score	4218	45.649351
Asymmetrique Profile Index	4218	45.649351
Asymmetrique Activity Index	4218	45.649351
City	3669	39.707792
Specialization	3380	36.580087
Tags	3353	36.287879
Last Notable Activity	0	0.000000
A free copy of Mastering The Interview	0	0.000000

Categorical Columns Analysis (Conversion to Binary Data) :

- For Better analysis Columns with Binary Values like 'Yes/No' have been converted into 1/0 respectively.

Lead Number	Lead Origin	Lead Source	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Last Activity	Specialization	Tags	City	A free copy of Mastering The Interview	Last Notable Activity
660737	API	Olark Chat	0	0.0	0	0.0	Others	NaN	Others	NaN	0	Modified
660728	API	Organic Search	0	5.0	674	2.5	Email Opened	NaN	Ringing	NaN	0	Email Opened
660727	Landing Page Submission	Direct Traffic	1	2.0	1532	2.0	Email Opened	Others	Will revert after reading the email	Mumbai	1	Email Opened
660719	Landing Page Submission	Direct Traffic	0	1.0	305	1.0	Others	Others	Ringing	Mumbai	0	Modified
660681	Landing Page Submission	Google	1	2.0	1428	1.0	Others	NaN	Will revert after reading the email	Mumbai	0	Modified

DATA VISUALIZATION AND PREPARATION

Categorical Columns Analysis (Creating Dummies) :

- Each of these fields have 35%+ missing values. Imputing median values can skew data here.
- # One way to impute missing values would be to mark the missing values as 'Missing' and drop the first dummy created.
- # By default, no dummy variable is created for NaN values. Hence, we are leaving the NaN values as NaN and not dropping any columns. Columns for which this has been done 'Specialization','Tags','City'
- For columns like 'Last Activity', 'Lead Origin', 'Lead Source', 'Last Notable Activity' we are creating the dummies and dropping the first one.
- Once the dummies have been created, we will be dropping the original columns

```
1  dummy1.columns
```

```
Index(['Specialization_Finance Management',
       'Specialization_Human Resource Management',
       'Specialization_Marketing Management', 'Specialization_Others',
       'Tags_Others', 'Tags_Ringing',
       'Tags_Will revert after reading the email', 'City_Mumbai',
       'City_Other Cities', 'City_Others', 'City_Thane & Outskirts'],
      dtype='object')
```

```
1  dummy2.columns
```

```
Index(['Last Activity_Olark Chat Conversation', 'Last Activity_Others',
       'Last Activity_SMS Sent', 'Lead Origin_Landing Page Submission',
       'Lead Origin_Others', 'Lead Source_Google', 'Lead Source_Olark Chat',
       'Lead Source_Organic Search', 'Lead Source_Others',
       'Last Notable Activity_Modified', 'Last Notable Activity_Others',
       'Last Notable Activity_SMS Sent'],
      dtype='object')
```

```
1  leads.columns
```

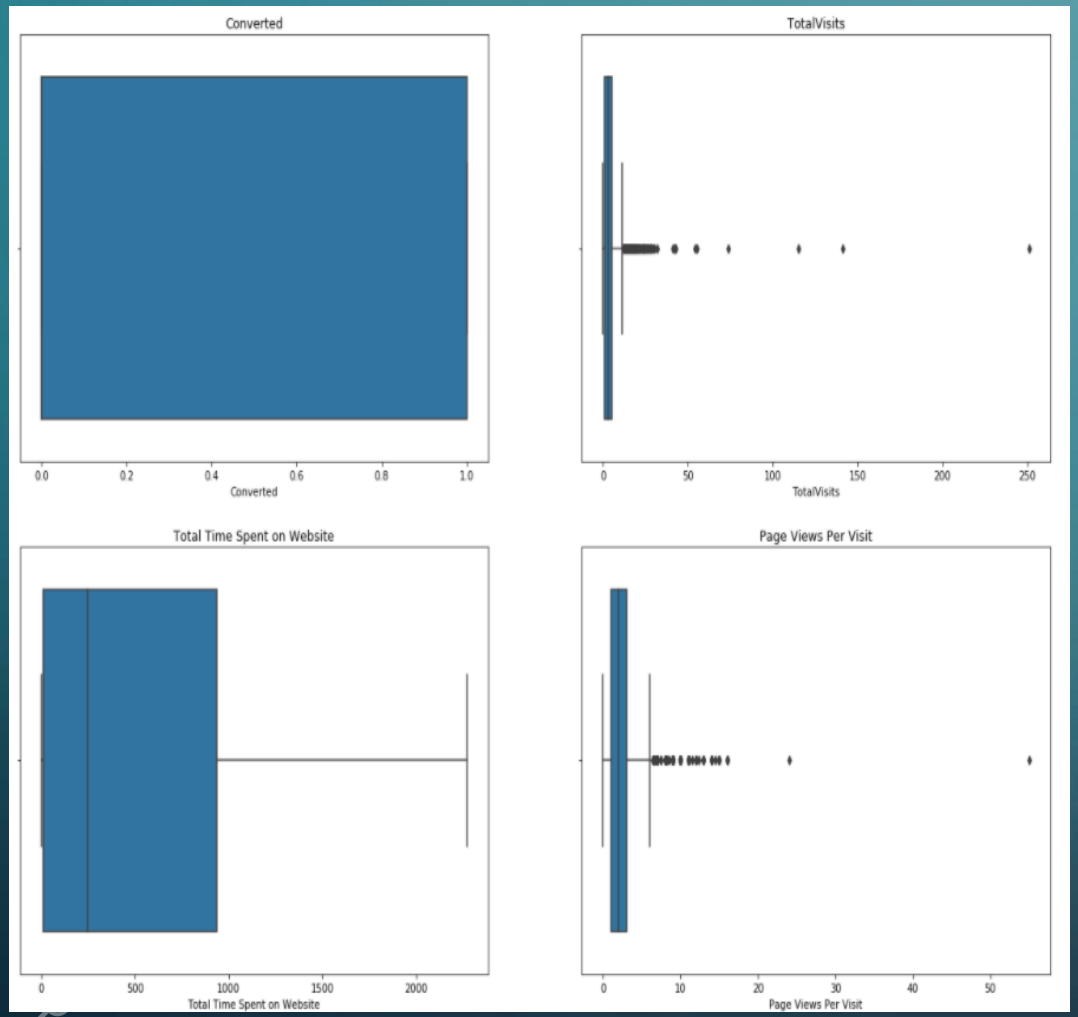
```
Index(['Converted', 'TotalVisits', 'Total Time Spent on Website',
       'Page Views Per Visit', 'A free copy of Mastering The Interview',
       'Specialization_Finance Management',
       'Specialization_Human Resource Management',
       'Specialization_Marketing Management', 'Specialization_Others',
       'Tags_Others', 'Tags_Ringing',
       'Tags_Will revert after reading the email', 'City_Mumbai',
       'City_Other Cities', 'City_Others', 'City_Thane & Outskirts',
       'Last Activity_Olark Chat Conversation', 'Last Activity_Others',
       'Last Activity_SMS Sent', 'Lead Origin_Landing Page Submission',
       'Lead Origin_Others', 'Lead Source_Google', 'Lead Source_Olark Chat',
       'Lead Source_Organic Search', 'Lead Source_Others',
       'Last Notable Activity_Modified', 'Last Notable Activity_Others',
       'Last Notable Activity_SMS Sent'],
      dtype='object')
```

DATA VISUALIZATION AND PREPARATION

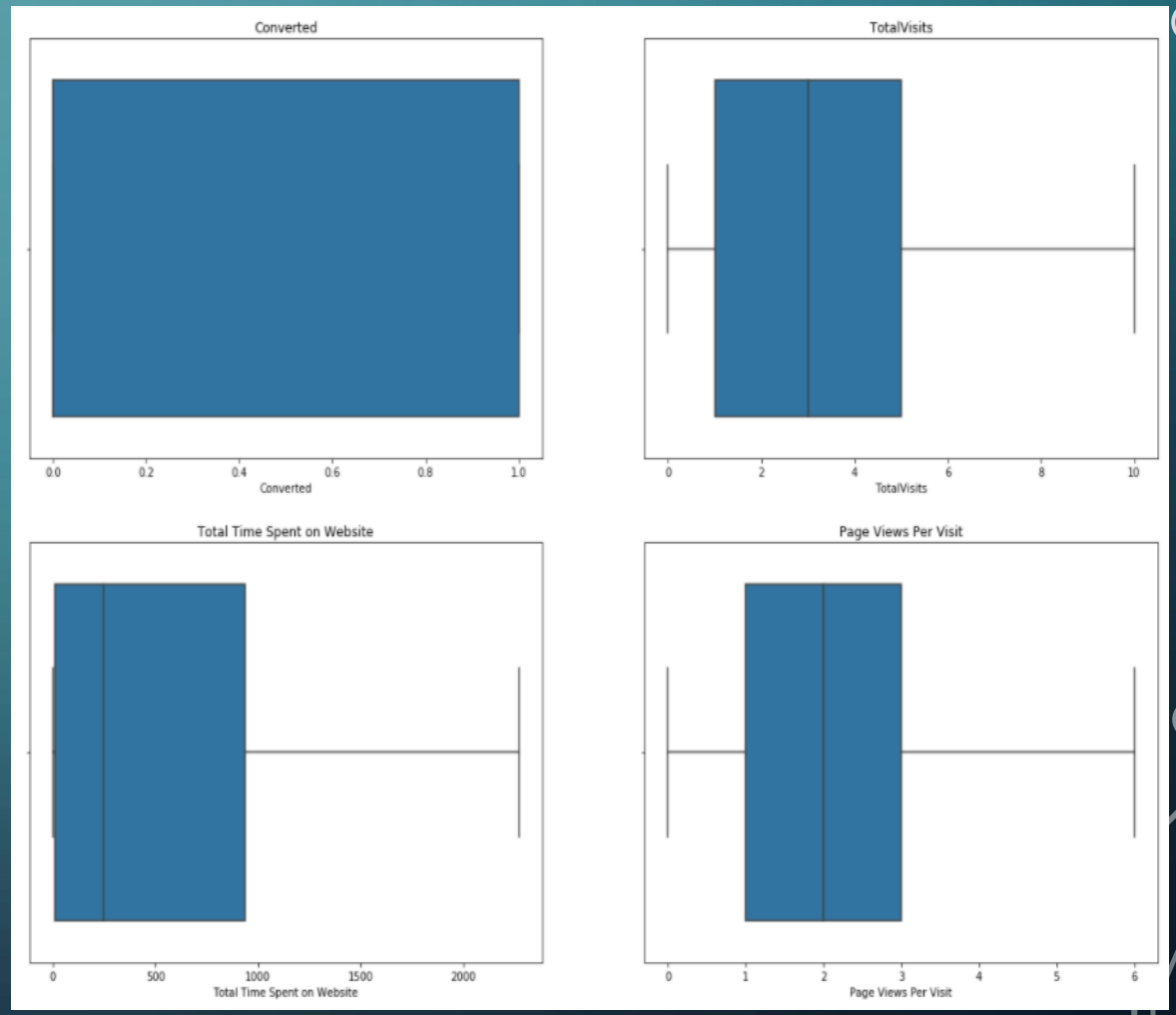
Outliers Analysis and Treatment:

- Since there appear to be a lot of outliers on the upper side, capping the values for 'TotalVisits' and 'Page Views Per Visit' to 95%

Before Treatment



After Capping Treatment

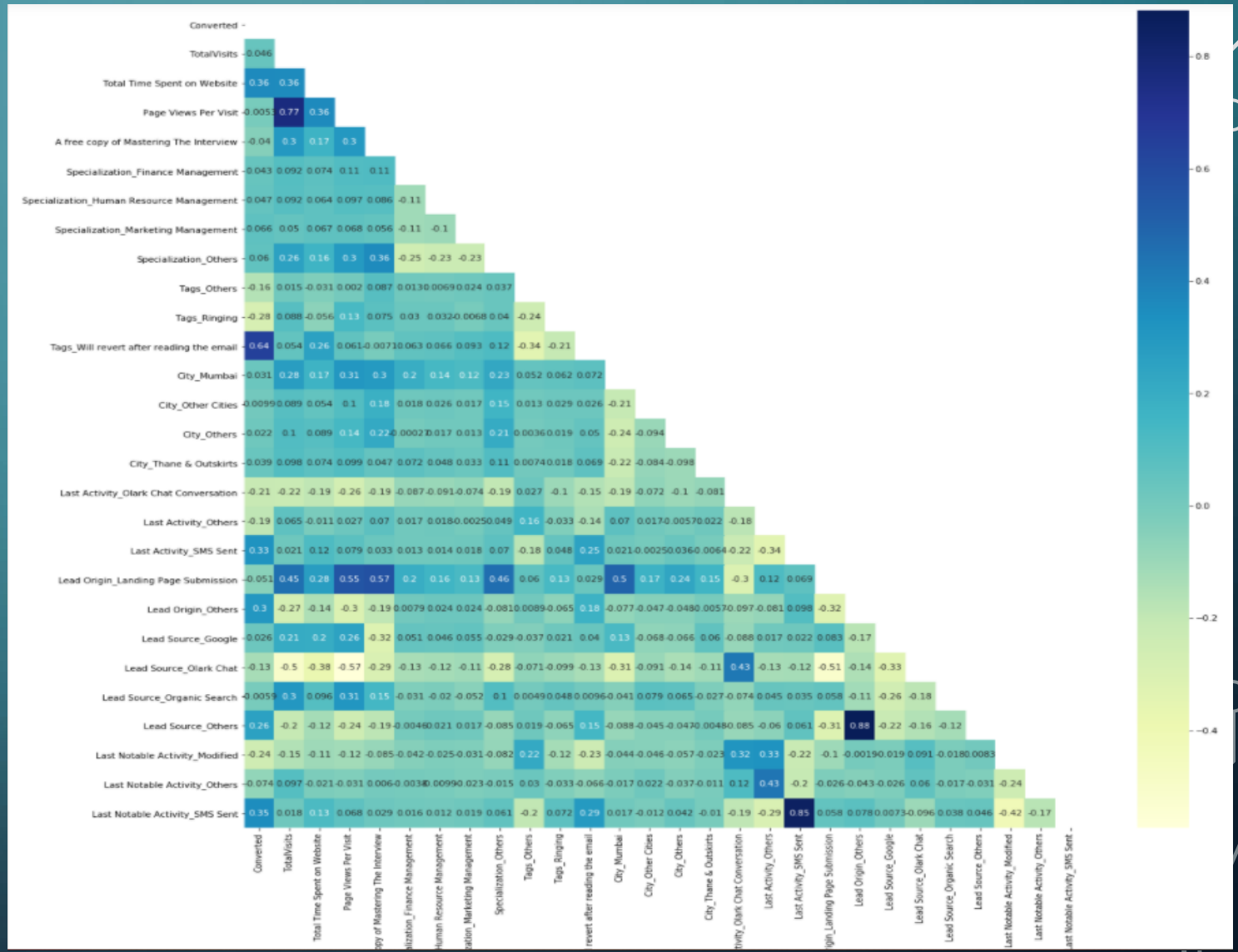


DATA VISUALIZATION AND PREPARATION

Bivariate Analysis and Correlations using Heatmap:

- It appears that the columns with highest correlation amongst themselves are:
  - Lead origin_Others and Lead Source_Others: 0.88
  - Last Notable Activity_SMS Sent and Last Activity_SMS Sent: 0.85
  - Page Views Per Visit and TotaVisits: .77

This might lead to multicollinearity. We will look at this using VIFs post defining model

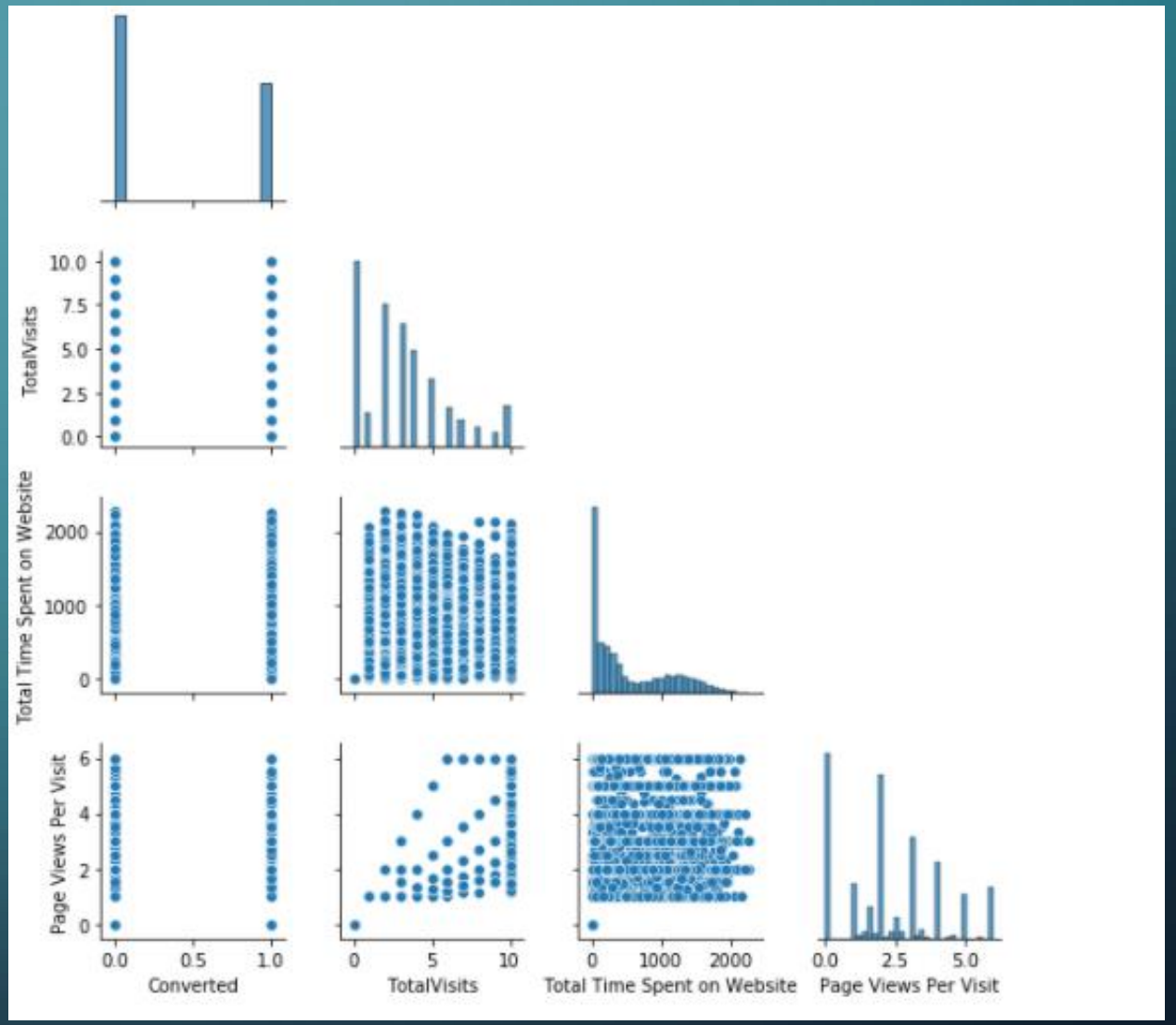


DATA VISUALIZATION AND PREPARATION

Bivariate Analysis and Correlations using Pairplots:

- It appears that the columns with highest correlation amongst themselves are:
  - Lead origin_Others and Lead Source_Others: 0.88
  - Last Notable Activity_SMS Sent and Last Activity_SMS Sent: 0.85
  - Page Views Per Visit and TotaVisits: .77

This might lead to multicollinearity. We will look at this using VIFs post defining model



SCALING

- We will be using the Standard Scaler for scaling the feature variables. Variables being scaled are :
  - TotalVisits
  - Total Time Spent on Website
  - Page Views Per Visit

Prior to Scaling we will doing the Train/Test Split in a 70/30 Ratio and keeping the Converted column as the Target Variable.

Before scaling

	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	9240.000000	9240.000000	9240.000000	9240.000000
mean	0.385390	3.179221	487.698268	2.255105
std	0.486714	2.761219	548.021466	1.779471
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	1.000000	12.000000	1.000000
50%	0.000000	3.000000	248.000000	2.000000
75%	1.000000	5.000000	936.000000	3.000000
max	1.000000	10.000000	2272.000000	6.000000

After Scaling

	TotalVisits	Total Time Spent on Website	Page Views Per Visit	6
count	6.468000e+03	6.468000e+03	6.468000e+03	
mean	-6.206813e-17	-1.609377e-16	-9.142003e-17	
std	1.000077e+00	1.000077e+00	1.000077e+00	
min	-1.149699e+00	-8.853708e-01	-1.266675e+00	
25%	-7.873438e-01	-8.634138e-01	-7.025878e-01	
50%	-6.263344e-02	-4.352528e-01	-1.385005e-01	
75%	6.620769e-01	8.098906e-01	4.255868e-01	
max	2.473853e+00	3.271816e+00	2.117849e+00	

MODEL BUILDING

Initially we have checked the p-values for the columns. View the results on the right for **p-values**

Then we proceed to use **Feature Selection using RFE** to select the **15** most significant features. List of columns mentioned below.

```
1 list(zip(X_train.columns, rfe.support_, rfe.ranking_))
```

```
[('TotalVisits', True, 1),
 ('Total Time Spent on Website', True, 1),
 ('Page Views Per Visit', True, 1),
 ('A free copy of Mastering The Interview', False, 11),
 ('Specialization_Finance Management', True, 1),
 ('Specialization_Human Resource Management', False, 2),
 ('Specialization_Marketing Management', True, 1),
 ('Specialization_Others', True, 1),
 ('Tags_Others', False, 13),
 ('Tags_Ringing', True, 1),
 ('Tags_Will revert after reading the email', True, 1),
 ('City_Mumbai', False, 5),
 ('City_Other Cities', False, 6),
 ('City_Others', False, 3),
 ('City_Thane & Outskirts', False, 4),
 ('Last Activity_Olark Chat Conversation', True, 1),
 ('Last Activity_Others', True, 1),
 ('Last Activity_SMS Sent', False, 9),
 ('Lead Origin_Landing Page Submission', True, 1),
 ('Lead Origin_Others', True, 1),
 ('Lead Source_Google', False, 8),
 ('Lead Source_Olark Chat', True, 1),
 ('Lead Source_Organic Search', False, 12),
 ('Lead Source_Others', False, 7),
 ('Last Notable Activity_Modified', False, 10),
 ('Last Notable Activity_Others', True, 1),
 ('Last Notable Activity_SMS Sent', True, 1)]
```

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6440
Model Family:	Binomial	**Df Model:**	27
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1789.2
Date:	Sun, 07 Feb 2021	**Deviance:**	3578.4
Time:	13:51:08	**Pearson chi2:**	1.42e+04
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.7279	0.176	-9.825	0.000	-2.073	-1.383
TotalVisits	0.4880	0.066	7.411	0.000	0.359	0.617
Total Time Spent on Website	1.0694	0.050	21.470	0.000	0.972	1.167
Page Views Per Visit	-0.4901	0.079	-6.226	0.000	-0.644	-0.336
A free copy of Mastering The Interview	-0.0473	0.137	-0.345	0.730	-0.316	0.221
Specialization_Finance Management	0.9884	0.242	4.079	0.000	0.513	1.463
Specialization_Human Resource Management	0.6010	0.246	2.443	0.015	0.119	1.083
Specialization_Marketing Management	0.6636	0.238	2.789	0.005	0.197	1.130
Specialization_Others	0.6432	0.213	3.024	0.002	0.226	1.060
Tags_Others	-0.0043	0.099	-0.043	0.966	-0.198	0.189
Tags_Ringing	-3.4791	0.239	-14.569	0.000	-3.947	-3.011
Tags_Will revert after reading the email	4.3226	0.184	23.484	0.000	3.962	4.683
City_Mumbai	-0.5438	0.224	-2.424	0.015	-0.983	-0.104
City_Other Cities	-0.4669	0.264	-1.771	0.077	-0.984	0.050
City_Others	-0.6891	0.252	-2.733	0.006	-1.183	-0.195
City_Thane & Outskirts	-0.5928	0.262	-2.264	0.024	-1.106	-0.080
Last Activity_Olark Chat Conversation	-1.6817	0.236	-7.130	0.000	-2.144	-1.219
Last Activity_Others	-1.0226	0.185	-5.521	0.000	-1.386	-0.660
Last Activity_SMS Sent	0.2620	0.204	1.287	0.198	-0.137	0.661
Lead Origin_Landing Page Submission	-0.2537	0.193	-1.317	0.188	-0.631	0.124
Lead Origin_Others	3.4543	0.354	9.770	0.000	2.761	4.147
Lead Source_Google	0.2787	0.144	1.942	0.052	-0.003	0.560
Lead Source_Olark Chat	1.1030	0.201	5.499	0.000	0.710	1.496
Lead Source_Organic Search	0.0420	0.169	0.249	0.804	-0.289	0.373
Lead Source_Others	0.2485	0.324	0.767	0.443	-0.387	0.884
Last Notable Activity_Modified	-0.0763	0.157	-0.488	0.626	-0.383	0.231
Last Notable Activity_Others	0.5466	0.239	2.287	0.022	0.078	1.015
Last Notable Activity_SMS Sent	1.5524	0.235	6.592	0.000	1.091	2.014

MODEL BUILDING

Post Feature Selection we assess the model using the 15 features and check for p-values again

We also check the VIF scores for them.

Our process/order to drop the columns will be as follows :

- High p-value, high VIF
- High p-value, low VIF
- Low p-value, high VIF

	Features	VIF
0	const	7.18
3	Page Views Per Visit	3.18
1	TotalVisits	2.60
11	Lead Origin_Landing Page Submission	2.56
13	Lead Source_Olark Chat	2.48
6	Specialization_Others	1.93
12	Lead Origin_Others	1.71
10	Last Activity_Others	1.53
4	Specialization_Finance Management	1.42
9	Last Activity_Olark Chat Conversation	1.40
14	Last Notable Activity_Others	1.36
2	Total Time Spent on Website	1.35
8	Tags_Will revert after reading the email	1.34
5	Specialization_Marketing Management	1.33
15	Last Notable Activity_SMS Sent	1.26
7	Tags_Ringing	1.12

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6452
Model Family:	Binomial	Df Model:	15
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1798.9
Date:	Sun, 07 Feb 2021	Deviance:	3597.9
Time:	13:51:09	Pearson chi2:	1.26e+04
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.5027	0.108	-13.952	0.000	-1.714	-1.292
TotalVisits	0.4735	0.065	7.333	0.000	0.347	0.600
Total Time Spent on Website	1.0691	0.049	21.672	0.000	0.972	1.166
Page Views Per Visit	-0.4738	0.077	-6.186	0.000	-0.624	-0.324
Specialization_Finance Management	0.5006	0.171	2.933	0.003	0.166	0.835
Specialization_Marketing Management	0.2237	0.177	1.261	0.207	-0.124	0.571
Specialization_Others	0.1411	0.129	1.093	0.275	-0.112	0.394
Tags_Ringing	-3.5026	0.235	-14.930	0.000	-3.962	-3.043
Tags_Will revert after reading the email	4.3092	0.173	24.929	0.000	3.970	4.648
Last Activity_Olark Chat Conversation	-1.7655	0.201	-8.770	0.000	-2.160	-1.371
Last Activity_Others	-1.1080	0.138	-8.036	0.000	-1.378	-0.838
Lead Origin_Landing Page Submission	-0.4392	0.140	-3.139	0.002	-0.713	-0.165
Lead Origin_Others	3.4889	0.215	16.210	0.000	3.067	3.911
Lead Source_Olark Chat	0.8912	0.167	5.350	0.000	0.565	1.218
Last Notable Activity_Others	0.6328	0.181	3.503	0.000	0.279	0.987
Last Notable Activity_SMS Sent	1.7952	0.112	15.994	0.000	1.575	2.015

MODEL BUILDING

During the iterations, the following columns were dropped:

- Specialization_Others (highest p-value despite low VIF)
- Specialization_Marketing (highest p-value despite low VIF)

Our process/order to drop the columns will be as follows :

- High p-value, high VIF
- High p-value, low VIF
- Low p-value, high VIF

The p-values and VIF after dropping the columns are shown in the snapshot.

As you can see both p-values and VIF for remaining features are in acceptable values (p-value<0.05 and VIF<5). Hence proceeding with these features for model evaluation

	Features	VIF
0	const	7.00
3	Page Views Per Visit	3.17
1	TotalVisits	2.60
11	Lead Source_Olark Chat	2.48
9	Lead Origin_Landing Page Submission	1.89
10	Lead Origin_Others	1.68
8	Last Activity_Others	1.53
7	Last Activity_Olark Chat Conversation	1.40
12	Last Notable Activity_Others	1.36
2	Total Time Spent on Website	1.34
6	Tags_Will revert after reading the email	1.30
13	Last Notable Activity_SMS Sent	1.26
5	Tags_Ringing	1.12
4	Specialization_Finance Management	1.05

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6454
Model Family:	Binomial	Df Model:	13
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1799.9
Date:	Sun, 07 Feb 2021	Deviance:	3599.8
Time:	13:51:09	Pearson chi2:	1.26e+04
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.4829	0.107	-13.917	0.000	-1.692	-1.274
TotalVisits	0.4770	0.064	7.401	0.000	0.351	0.603
Total Time Spent on Website	1.0707	0.049	21.714	0.000	0.974	1.167
Page Views Per Visit	-0.4699	0.077	-6.142	0.000	-0.620	-0.320
Specialization_Finance Management	0.3890	0.146	2.657	0.008	0.102	0.676
Tags_Ringing	-3.5121	0.235	-14.949	0.000	-3.973	-3.052
Tags_Will revert after reading the email	4.3297	0.172	25.154	0.000	3.992	4.667
Last Activity_Olark Chat Conversation	-1.7627	0.201	-8.789	0.000	-2.156	-1.370
Last Activity_Others	-1.1000	0.138	-7.994	0.000	-1.370	-0.830
Lead Origin_Landing Page Submission	-0.3418	0.117	-2.922	0.003	-0.571	-0.113
Lead Origin_Others	3.5263	0.213	16.567	0.000	3.109	3.943
Lead Source_Olark Chat	0.8963	0.166	5.391	0.000	0.570	1.222
Last Notable Activity_Others	0.6224	0.180	3.453	0.001	0.269	0.976
Last Notable Activity_SMS Sent	1.7960	0.112	15.986	0.000	1.576	2.016

MODEL EVALUATION

Prediction :

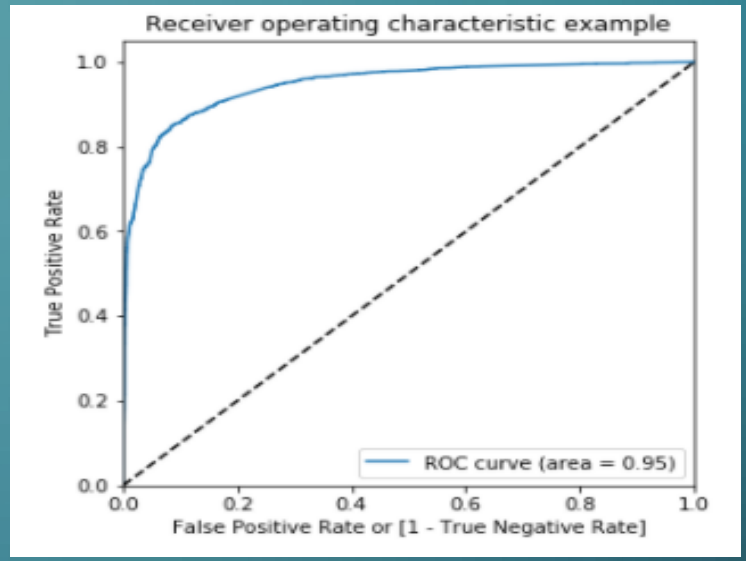
- Creating a dataframe with the actual Converted flag and the predicted probabilities
- We will taking all probabilities values greater than 0.5 as 1 i.e., 'Converted' and less than 0.5 as 'Not Converted'
- The Metrics calculated are as follows:

	Metric	Value
0	accuracy	0.891466
1	sensi/TPR/Recall/HitRate	0.816707
2	speci/TNR	0.937531
3	FPR	0.062469
4	FNR	0.183293
5	PositivePredictiveValue/Precision	0.889576
6	NegativePredictiveValue	0.892483

ROC CURVE AND OPTIMAL THRESHOLD PROBABILITY

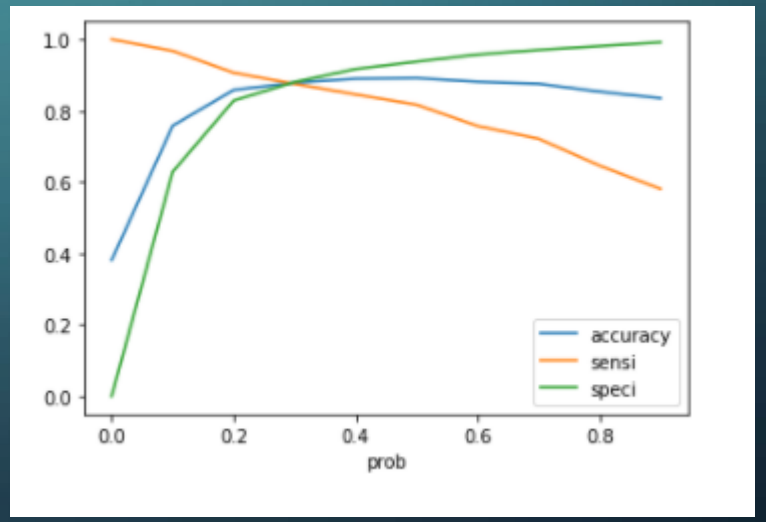
ROC Curve:

- The Maximum area for the ROC curve is at 0.95



Optimal threshold:

- From the accuracy/sensitivity/specificity curve, 0.3 is the optimum threshold.



PREDICTION ON TEST DATA

We scale the features similarly on the test data set as well and using our model predict the Test set.

Post prediction the Metrics for various probability thresholds are calculated

Result : It's observed, metrics for test and train models are very similar. Hence, model can predict quite well.

Hence, we go ahead assign the Lead scores predicted to the Dataframe. This is added as a new column called 'Lead Score'

Train Set Metrics

	Metric	Value
0	accuracy	0.878324
1	sensi/TPR/Recall/HitRate	0.875101
2	speci/TNR	0.880310
3	FPR	0.119690
4	FNR	0.124899
5	PositivePredictiveValue/Precision	0.818354
6	NegativePredictiveValue	0.919603

Test Set Metrics

	Metric	Value
0	accuracy	0.874459
1	sensi/TPR/Recall/HitRate	0.874886
2	speci/TNR	0.874180
3	FPR	0.125820
4	FNR	0.125114
5	PositivePredictiveValue/Precision	0.819504
6	NegativePredictiveValue	0.914535

DataFrame with Lead Score

	Lead Number	Converted	Converted_Prob	final_predicted	Lead Score
0	619003	1	0.765825	1	77.0
1	636884	1	0.999303	1	100.0
2	590281	1	0.754440	1	75.0
3	579892	0	0.037309	0	4.0
4	617929	1	0.995814	1	100.0

Higher the Lead Score, greater the chance of conversion. Hence, while making calls, it is better to start with the highest lead scores

RECOMMENDATIONS

- The Final model created has the column Lead Score which indicates which Leads have a higher conversion Rate.
- As per the ask it targets to highlight 80% of the leads which have a higher chance of converting into customers.
- The Sales team can prioritize using these High Lead scores and accordingly create and achieve targets and increase the overall revenue for the X-Education group.

REQUESTED STRATEGY

1. During the 2 month period when interns are hired to make additional calls and leads need to be chased aggressively, the following strategy needs to be applied
   a. The Sales team along with the interns should target the Lead scores from highest to Lowest.
   b. The higher the Lead score the higher is the chance for conversion.
   c. In this manner they will making proper use of the time and resources at hand to increase revenue during the 2 months where the Leads volume is at the highest.

2. In scenarios where target is already reached and only necessary calls need to be made so that sales team can concentrate on new work, the following strategy can be applied.
   a. The Sales team can check for the high Lead scores e.g., 95-100 which can be targeted first and calls be made for checking for conversion into Sales/Revenue for X-Education.
   b. Once the quarter target is achieved, they can concentrate on the new work.