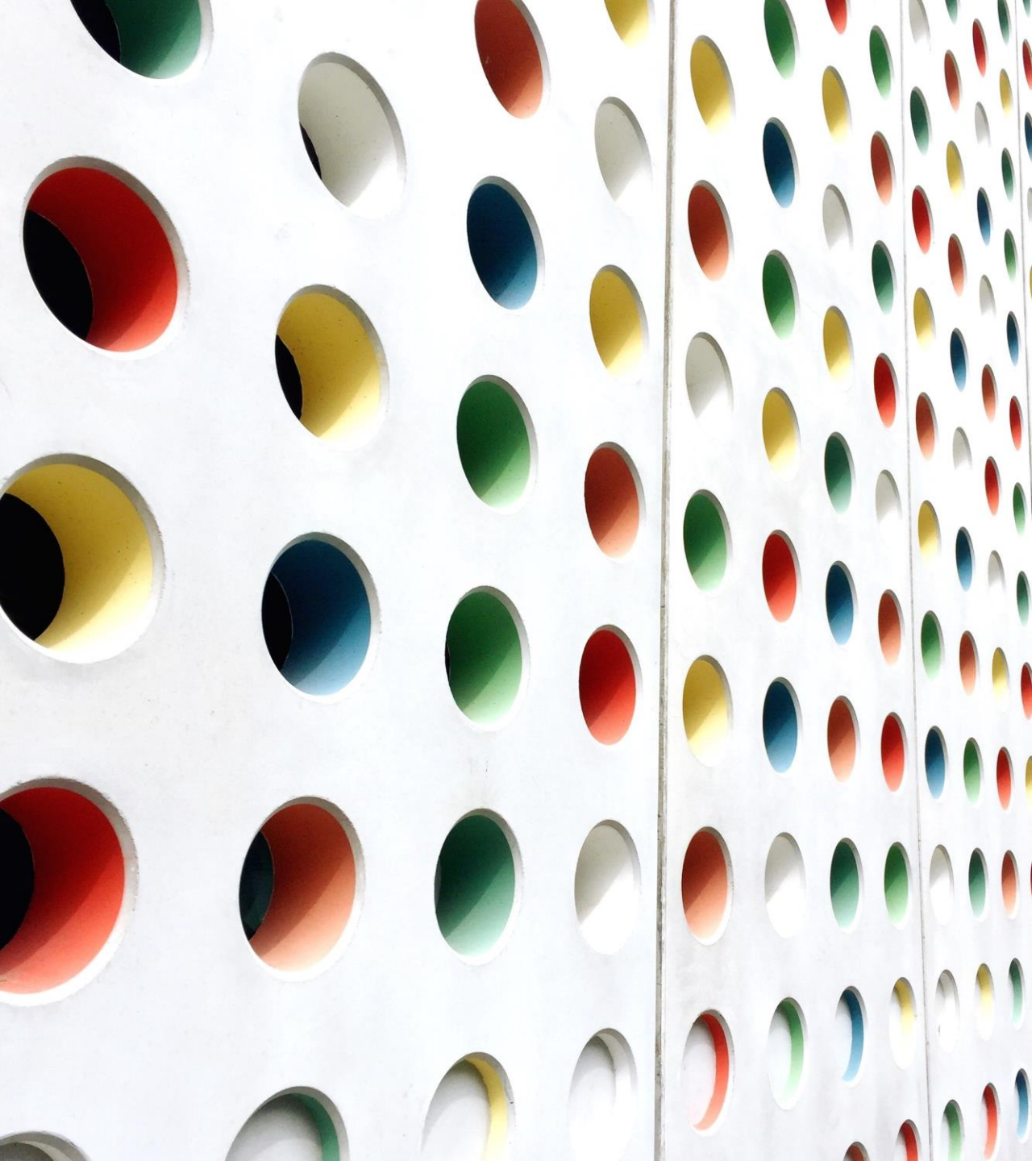


Lead Scoring Case Study

Alka T James

Business Problem Statement

Business Problem Statement:

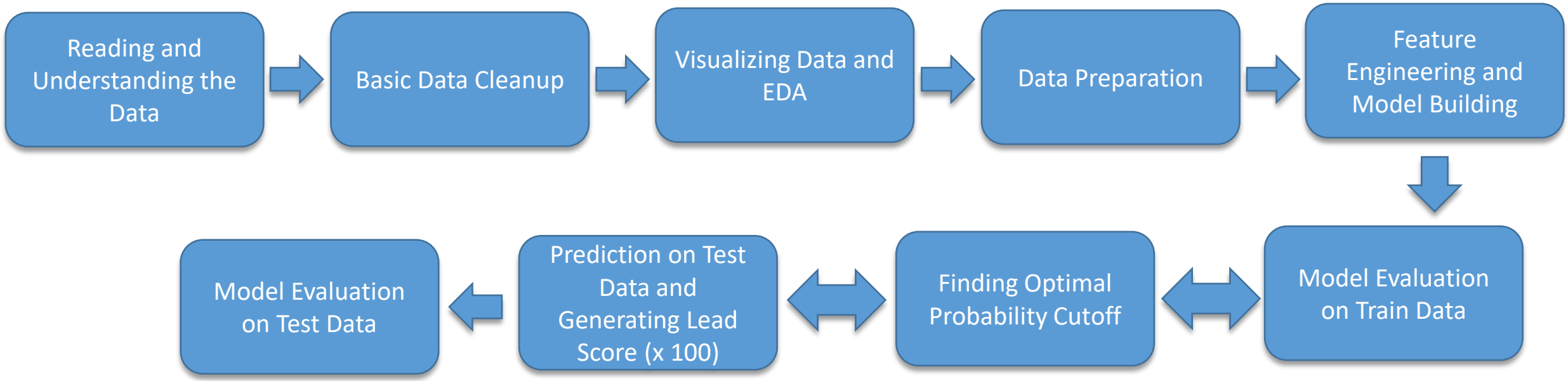
An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

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%.

Goal:

1. To identify the features that contributes to predict Lead Conversion.
2. Identifying Hot Leads by generating Lead Score for all leads, so that leads having higher Lead Scores can be contacted with priority for achieving Higher Lead Conversion Rate.

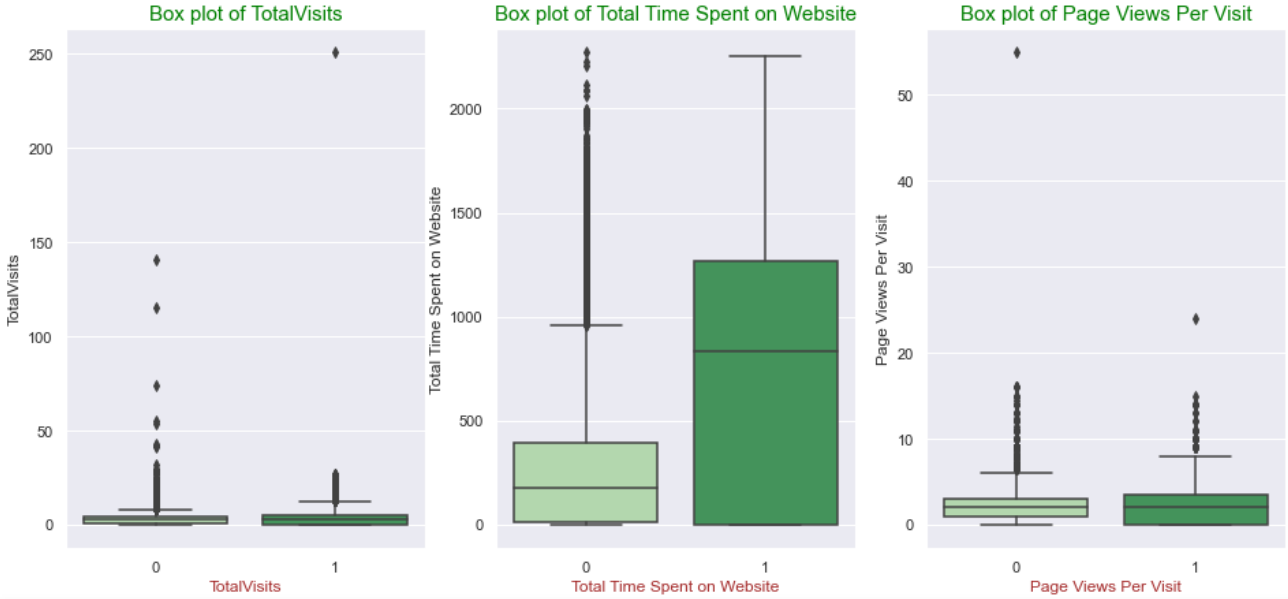
Overall Approach

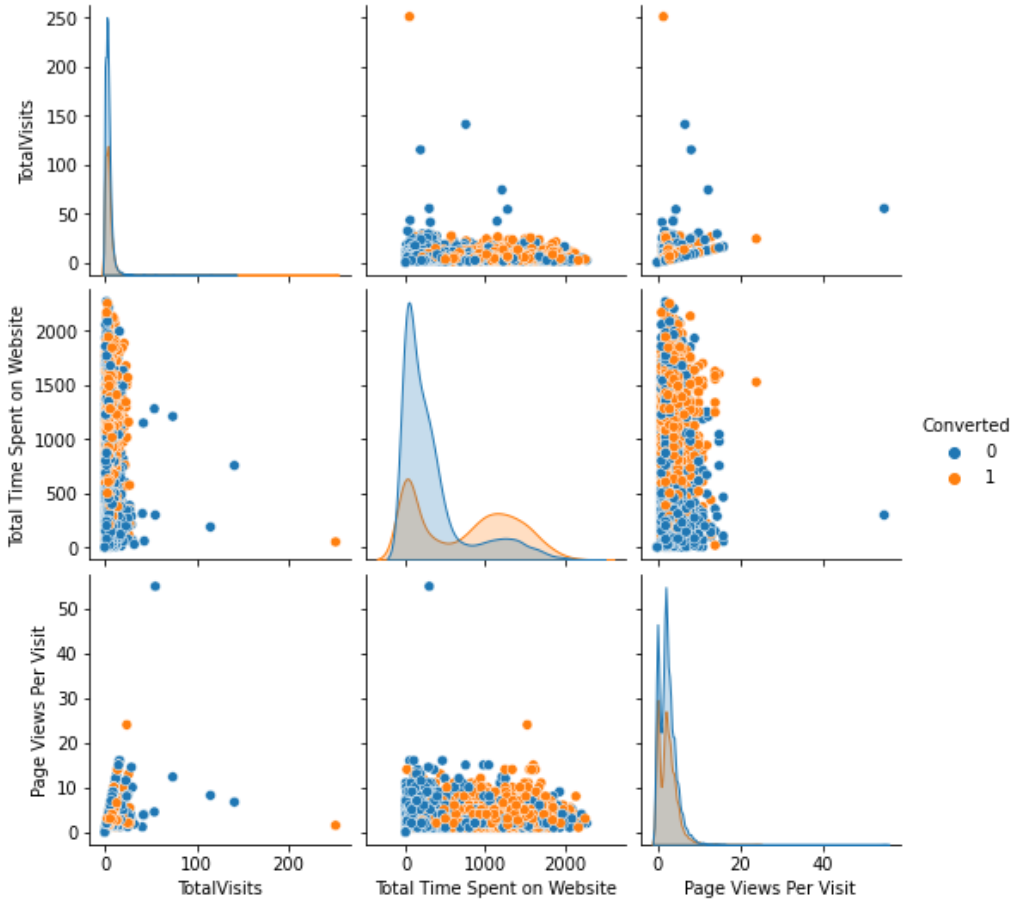


Understanding the Data & Basic Data Cleanup

- There are 37 columns (30 categorical and 7 Numeric) and 9240 observations in the dataset.
- Select is present as a class in different columns like:
  - Specialization
  - How did you hear about X Education
  - Lead Profile
  - City
- As Select is not a valid class, we can conclude that the Select might be the default value set in the form dropdown and if the user has not selected any option from the dropdown, then the value remained as Select. We replaced Select with NaN.
- **Magazine, 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** - These columns have no missing data and have only one unique value. So, these columns have no variance and not helpful for our EDA or model building, hence we dropped these columns.
- **How did you hear about X Education, Lead Profile, Lead Quality, Asymmetrique Activity Index, Asymmetrique Profile Index , Asymmetrique Activity Score, Asymmetrique Profile Score** – These columns have more than 40% missing value. So, we have dropped these columns from our EDA and model building.
- There is no datapoint/ observation (rows) in our dataset having more than 70% missing values.
- We have created new buckets/bins for the categorical variables having very high numbers of classes with few datapoints: **Lead Origin, Lead Source, Last Activity**, **Last Notable Activity**, **Country**, **Specialization**, **What is your current occupation.**
- Performed missing value treatment using **Business Understanding.** For **Specialization** and **Occupation** NaN values are replaced with a new category **Not Disclosed.**
- We renamed **What is your current occupation** column to **Occupation** and **What matters most to you in choosing a course** to **Reason_choosing** for our convenience during EDA and Model building .

Visualizing Data and EDA : Numerical variables





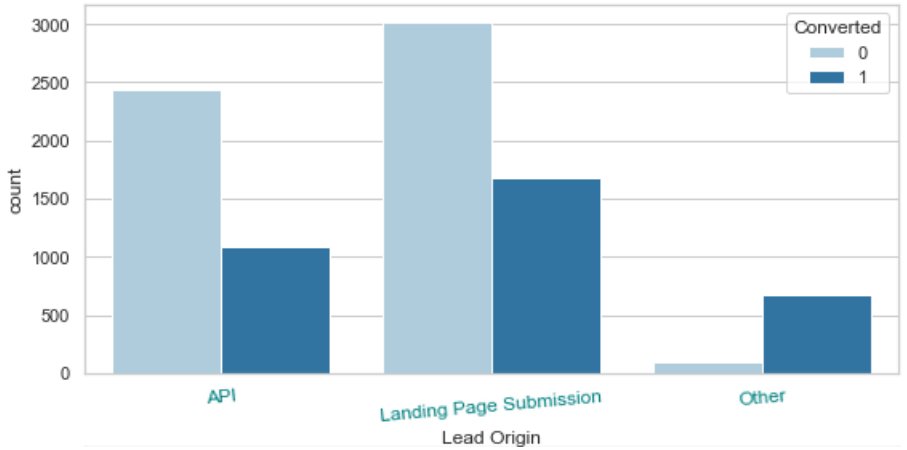
Inferences:

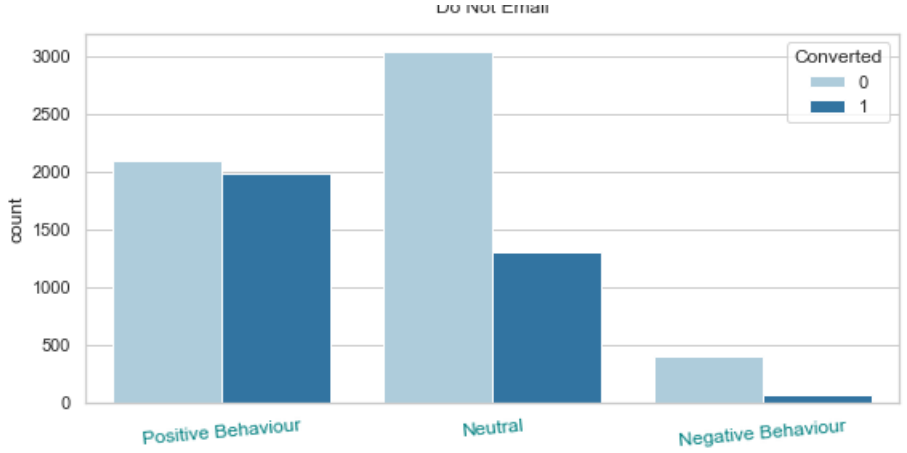
1. Median value of **Total Time Spent on Website** for converted Leads are considerably higher than the other group. Team should target those customers who are spending higher time on website. Those leads have higher odds of getting converted.

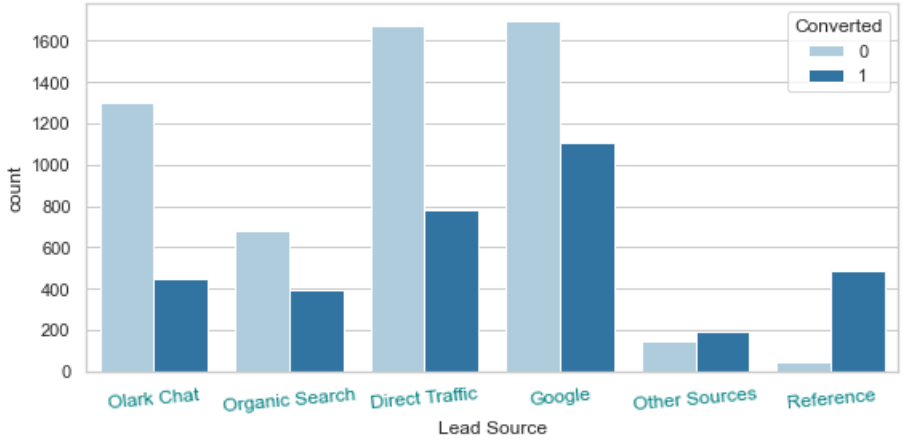
2. We can see high number of outliers in **Total Visit** for Converted = 0. A good number of customers visiting the website higher number of times, but they are not opting for the course. Team should investigate the reason behind this. It could be because of financial problems, they are searching for other courses that is currently not offered by X Education, Getting other good options by competitors etc.

3. There are many outliers in **Total Time Spent on Website** column for Converted= 0.

Visualizing Data and EDA : Categorical Variables I







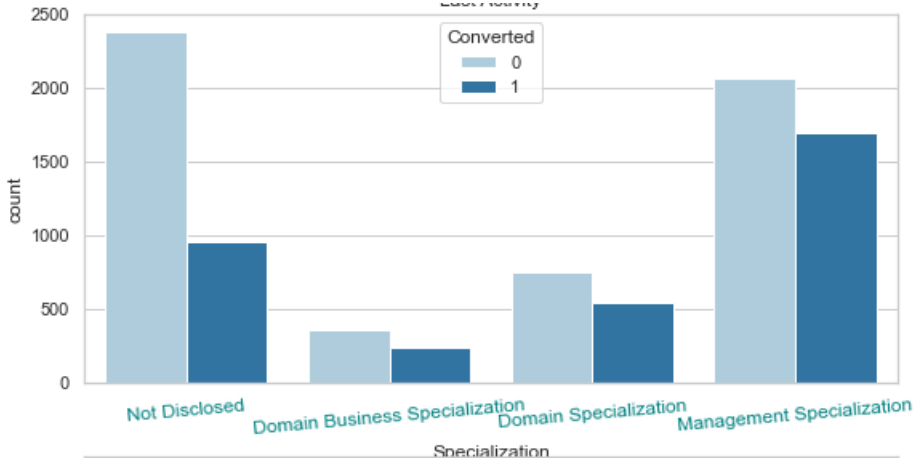
Inferences:

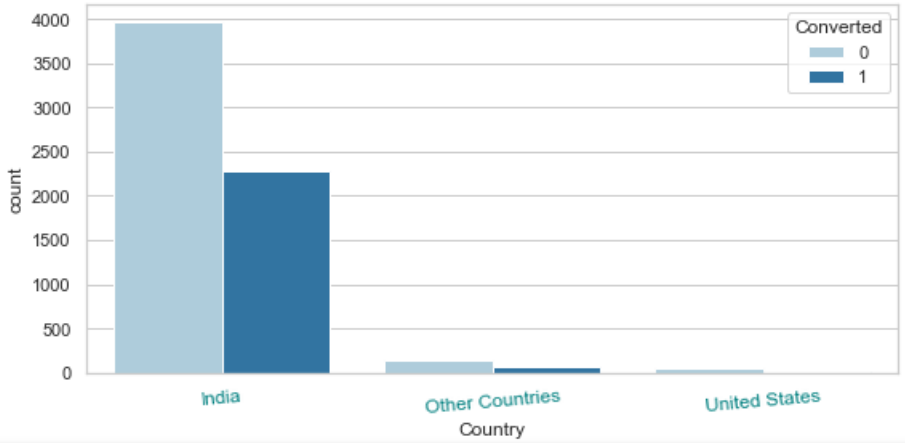
4. For **Other type Lead Origin** there is a very high chance that the lead will be converted successfully.

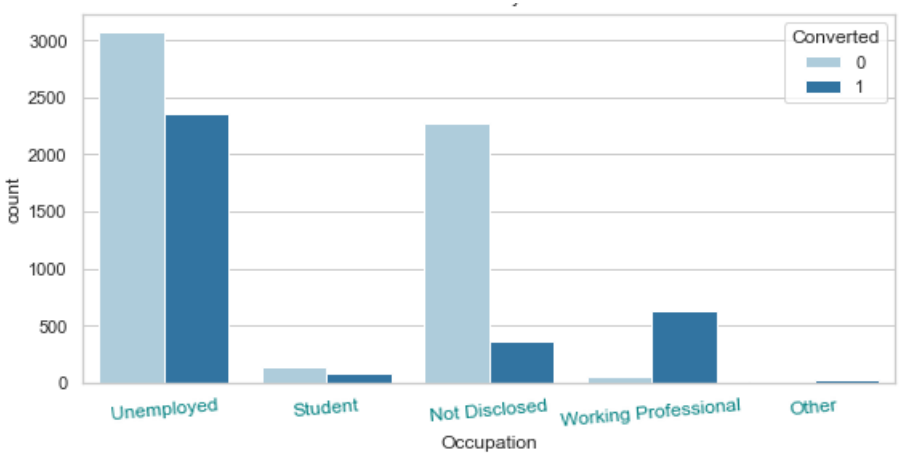
5. Reference type of Lead Source has very high success rate. Team should consider referenced customers on priority. **Other Sources** contributes very a smaller number of customers but has very good conversion rate. Customers coming through **Organic Search** have also considerably higher chance of getting successfully converted.

6. Customers who have shown **Positive Behaviour in Last Activity** have considerably higher chance of getting converted successfully.

Visualizing Data and EDA : Categorical Variables II







Inferences:

7. Most of the customers are from India and having **Management Specializations**.

8. **People who have mentioned their specializations** in the form have higher chance of opting for the course.

9. **Most of the customers who have shown interest in the campaigns are Unemployed. Working Professional** very high odds of successful conversion. Sales team should launch a campaign to reach more Working professionals.

10. **People mentioned employment status** during filling up the forms have higher chance of successful conversion.

Data Preparation

Outlier Treatment

Identified 2.8% of total data (< 5%) as outliers and removed those rows

```
# Checking percentile values for 'Page Views Per Visit'
leads_df0['Page Views Per Visit'].quantile([.01,.05,.5,.75,.9,.95,.99,1]).values
array([ 0.,  0.,  2.,  3.,  5.,  6.,  9., 55.])

# Removing observations having Page Views Per Visit > 9
leads_df0= leads_df0[~(leads_df0['Page Views Per Visit'] > 9)]

# Checking percentile values for 'TotalVisits'
leads_df0['TotalVisits'].quantile([.01,.05,.5,.75,.9,.95,.99,1]).values
array([  0.  ,   0.  ,   3.  ,   5.  ,   7.  ,   9.  ,  16.73, 251.  ])

# Removing bservations having TotalVisits > 16.93
leads_df0= leads_df0[~(leads_df0['TotalVisits'] > 16.93)]

# Checking percentile values for 'Total Time Spent on Website'
leads_df0['Total Time Spent on Website'].quantile([.01,.05,.5,.75,.9,.95,.99,1]).values
array([   0.  ,    0.  ,  246.  ,  930.  , 1378.  , 1558.35, 1840.27,
       2272.  ])

# Removing observations having 'Total Time Spent on Website' > 1840.27
leads_df0= leads_df0[~(leads_df0['Total Time Spent on Website'] > 1840.27)]
```

Train-Test Split

Dataset has been splitted into Train and Test in 70:30 ratio. Train dataset is used to train the model and Test dataset to evaluate the model.

Missing Value Imputation (Statistical Imputation)

1. Calculated median, mode on Train dataset.
2. Used that value to impute missing values in Train and Test Dataset.
3. Statistical Imputation is done as below:
   - A. Nominal Categorical Columns: Mode Imputation
   - B. Numeric Columns: Median Imputation

Categorical Variables Encoding

1. Below Columns having Yes and No values. Yes have been replaced by 1 and No with 0.
Do Not Email, Do Not Call, Search, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, A free copy of Mastering The Interview

2. Dummy variables have been created for below columns and dropped original variable and First Dummy Variable for each column from the data frame:
Lead Origin, Lead Source, Country, Specialization, Reason_choosing, Occupation, City

MinMax Scaling on Train Data

Performed MinMax Scaling (fit and transform) on Train data on all numeric predictors: TotalVisits, Total Time Spent on Website, Page Views Per Visit

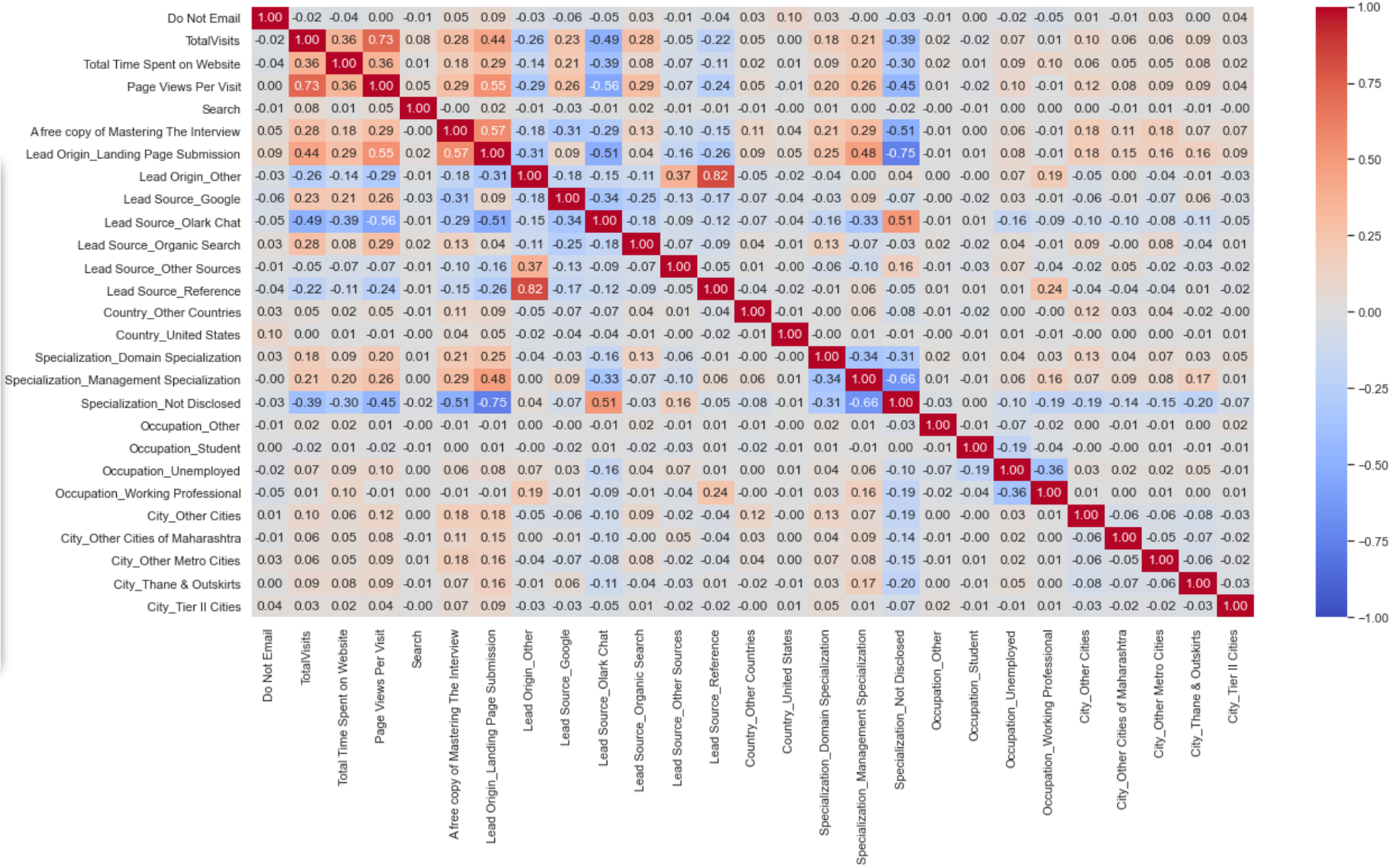
Variance Thresholding

Performed Variance Thresholding and removing columns having lower variance than threshold= .001
Removed Columns: Do Not Call, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, Reason_choosing_Flexibility & Convenience

Data Preparation: Pairwise Correlation

Lead Origin_Other has very high correlation (.82) with **Lead Source_Reference**. Column **Lead Source_Reference** has been dropped.

Lead_Origin_Landing Page Submission has very high correlation (.75) with **Specialization_Not Disclosed**. **Specialization_Not Disclosed** column has been dropped.



Model Building : Approach

1. Recursive Feature Elimination (RFE) has been used to get top 16 features.

- **Do Not Email :** An indicator variable selected by the customer wherein they select whether of not they want to be emailed about the course or not.
- **TotalVisits:** The total number of visits made by the customer on the website.
- **Total Time Spent on Website:** The total time spent by the customer on the website.
- **Page Views Per Visit:** Average number of pages on the website viewed during the visits.
- **Lead Origin_Landing Page Submission:** Dummy variable for Landing Page category of the origin identifier with which the customer was identified to be a lead.
- **Lead Origin_Other:** Dummy variable for the Other category of the origin identifier with which the customer was identified to be a lead.
- **Lead Source_Olark Chat:** Dummy variable for the Olark Chat category of the source of the lead.
- **Lead Source_Other Sources:** Dummy variable for the Other category (other than Google, Direct Traffic, Olark Chat, Organic Search, Reference) of the source of the lead.
- **Country_Other Countries:** Dummy variable for the Other category (other than India and United States) of the country of the customer.
- **Specialization_Domain Specialization**: Dummy variable for Domain Specialization bin of Specialization variable.
- **Specialization_Management Specialization:** Dummy variable for Management Specialization bin of Specialization variable.
- **Occupation_Other:** Dummy variable for 'Other' category of customer's occupation.
- **Occupation_Student:** Dummy variable for 'Student' category of customer's occupation.
- **Occupation_Unemployed:** Dummy variable for 'Unemployed' category of customer's occupation.
- **Occupation_Working Professional:** Dummy variable for 'Working Professional' category of customer's occupation.
- **City_Tier II Cities:** Dummy variable for 'Tier II Cities' category of customer's city.

2. We built first Logistic Regression model using GLM (Generalized Linear Model) in statsmodels with these 16 features.

3. Then manually fine tuned the model to get statistically significant features (by checking the p-values) and removed multicollinearity (By checking Variance Inflation Factors) simultaneously. Accepted p-value is lower than .05 and accepted VIF is lower than 5.

4. Total 7 models were built and after each model building p-values of all beta coefficients and VIFs have been checked and identified feature has been removed in next model building. We have also checked Overall model accuracy and Confusion Matrix after each new model, to understand how the new model is performing in compared to the previous one.

Model Building : Model 1

Features for Logistic Regression Model 1 (16 Features):

Do Not Email,
TotalVisits,
Total Time Spent on Website,
Page Views Per Visit,
Lead Origin_Landing Page Submission,
Lead Origin_Other,
Lead Source_Olark Chat,
Lead Source_Other Sources,
Country_Other Countries,
Specialization_Domain Specialization,
Specialization_Management Specialization,
Occupation_Other,
Occupation_Student,
Occupation_Unemployed,
Occupation_Working Professional,
City_Tier II Cities

Model Summary (Coefficients with p value)

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.2281	0.134	-24.165	0.000	-3.490	-2.966
Do Not Email	-1.1754	0.159	-7.380	0.000	-1.488	-0.863
TotalVisits	1.2198	0.274	4.448	0.000	0.682	1.757
Total Time Spent on Website	3.9556	0.139	28.536	0.000	3.684	4.227
Page Views Per Visit	-0.5423	0.265	-2.043	0.041	-1.063	-0.022
Lead Origin_Landing Page Submission	-0.4422	0.108	-4.088	0.000	-0.654	-0.230
Lead Origin_Other	3.4569	0.198	17.489	0.000	3.070	3.844
Lead Source_Olark Chat	1.1691	0.131	8.911	0.000	0.912	1.426
Lead Source_Other Sources	-0.5101	0.218	-2.343	0.019	-0.937	-0.083
Country_Other Countries	-0.4444	0.235	-1.891	0.059	-0.905	0.016
Specialization_Domain Specialization	0.4011	0.123	3.261	0.001	0.160	0.642
Specialization_Management Specialization	0.4068	0.099	4.118	0.000	0.213	0.600
Occupation_Other	1.4708	0.521	2.825	0.005	0.450	2.491
Occupation_Student	1.1282	0.213	5.289	0.000	0.710	1.546
Occupation_Unemployed	1.3191	0.086	15.383	0.000	1.151	1.487
Occupation_Working Professional	3.7367	0.194	19.244	0.000	3.356	4.117
City_Tier II Cities	0.4399	0.375	1.174	0.240	-0.294	1.174

VIF

Features	VIF
Page Views Per Visit	6.06
Lead Origin_Landing Page Submission	5.04
TotalVisits	4.78
Specialization_Management Specialization	3.68
Occupation_Unemployed	2.82
Total Time Spent on Website	2.20
Specialization_Domain Specialization	1.92
Lead Origin_Other	1.66
Occupation_Working Professional	1.43
Lead Source_Other Sources	1.25
Lead Source_Olark Chat	1.21
Do Not Email	1.09
Occupation_Student	1.07
Country_Other Countries	1.04
City_Tier II Cities	1.02
Occupation_Other	1.01

Observations:
'City_Tier II Cities' and 'Country_Other Countries' have p value higher than .05. So, their coefficient value is not statistically significant.
'Page Views Per Visit' and 'Lead Origin_Landing Page Submission' have bit higher VIF value but it's not too high. So, in next model 'City_Tier II Cities' feature has been excluded from the predictors.

```
Confusion Matrix:
True Negative: 3458      False Positive: 433
False Negative: 828      True Positive: 1569

Overall model accuracy: 0.7994592875318066
```

Model Building : Model 2

Features for Logistic Regression Model 2 (15 Features):

Do Not Email,
TotalVisits,
Total Time Spent on Website,
Page Views Per Visit,
Lead Origin_Landing Page Submission,
Lead Origin_Other,
Lead Source_Olark Chat,
Lead Source_Other Sources,
Country_Other Countries,
Specialization_Domain Specialization,
Specialization_Management Specialization,
Occupation_Other,
Occupation_Student,
Occupation_Unemployed,
Occupation_Working Professional,

Model Summary (Coefficients with p value)

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.2248	0.133	-24.157	0.000	-3.486	-2.963
Do Not Email	-1.1715	0.159	-7.353	0.000	-1.484	-0.859
TotalVisits	1.2183	0.274	4.443	0.000	0.681	1.756
Total Time Spent on Website	3.9544	0.139	28.534	0.000	3.683	4.226
Page Views Per Visit	-0.5462	0.265	-2.058	0.040	-1.066	-0.026
Lead Origin_Landing Page Submission	-0.4344	0.108	-4.025	0.000	-0.646	-0.223
Lead Origin_Other	3.4556	0.198	17.486	0.000	3.068	3.843
Lead Source_Olark Chat	1.1671	0.131	8.900	0.000	0.910	1.424
Lead Source_Other Sources	-0.5107	0.218	-2.346	0.019	-0.937	-0.084
Country_Other Countries	-0.4466	0.235	-1.899	0.058	-0.907	0.014
Specialization_Domain Specialization	0.4017	0.123	3.267	0.001	0.161	0.643
Specialization_Management Specialization	0.4046	0.099	4.099	0.000	0.211	0.598
Occupation_Other	1.4784	0.518	2.852	0.004	0.462	2.495
Occupation_Student	1.1233	0.213	5.268	0.000	0.705	1.541
Occupation_Unemployed	1.3176	0.086	15.370	0.000	1.150	1.486
Occupation_Working Professional	3.7371	0.194	19.252	0.000	3.357	4.118

VIF

Features	VIF
Page Views Per Visit	6.06
Lead Origin_Landing Page Submission	5.01
TotalVisits	4.78
Specialization_Management Specialization	3.68
Occupation_Unemployed	2.82
Total Time Spent on Website	2.20
Specialization_Domain Specialization	1.92
Lead Origin_Other	1.66
Occupation_Working Professional	1.43
Lead Source_Other Sources	1.25
Lead Source_Olark Chat	1.21
Do Not Email	1.09
Occupation_Student	1.07
Country_Other Countries	1.04
Occupation_Other	1.01

Observations:
There is no change in Confusion matrix and as well as in overall model accuracy even after removing ‘City_Tier II Cities.
Still 'Page Views Per Visit' and 'Lead Origin_Landing Page Submission' have bit higher VIF value but it's not too high.
So, in next model 'Country_Other Countries' feature has been excluded from the predictors (as it has higher p value).

```
Confusion Matrix:
True Negative: 3458      False Positive: 433
False Negative: 828      True Positive: 1569

Overall model accuracy: 0.7994592875318066
```

Model Building : Model 3

Features for Logistic Regression Model 3 (14 Features):

Do Not Email,
TotalVisits,
Total Time Spent on Website,
Page Views Per Visit,
Lead Origin_Landing Page Submission,
Lead Origin_Other,
Lead Source_Olark Chat,
Lead Source_Other Sources,
Specialization_Domain Specialization,
Specialization_Management Specialization,
Occupation_Other,
Occupation_Student,
Occupation_Unemployed,
Occupation_Working Professional,

Model Summary (Coefficients with p value)

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.2340	0.133	-24.235	0.000	-3.496	-2.972
Do Not Email	-1.1759	0.159	-7.378	0.000	-1.488	-0.864
TotalVisits	1.2173	0.274	4.443	0.000	0.680	1.754
Total Time Spent on Website	3.9536	0.138	28.554	0.000	3.682	4.225
Page Views Per Visit	-0.5397	0.265	-2.034	0.042	-1.060	-0.020
Lead Origin_Landing Page Submission	-0.4413	0.108	-4.093	0.000	-0.653	-0.230
Lead Origin_Other	3.4676	0.198	17.548	0.000	3.080	3.855
Lead Source_Olark Chat	1.1738	0.131	8.955	0.000	0.917	1.431
Lead Source_Other Sources	-0.5186	0.218	-2.384	0.017	-0.945	-0.092
Specialization_Domain Specialization	0.4031	0.123	3.279	0.001	0.162	0.644
Specialization_Management Specialization	0.4018	0.099	4.073	0.000	0.208	0.595
Occupation_Other	1.4896	0.518	2.873	0.004	0.473	2.506
Occupation_Student	1.1340	0.213	5.319	0.000	0.716	1.552
Occupation_Unemployed	1.3190	0.086	15.387	0.000	1.151	1.487
Occupation_Working Professional	3.7332	0.194	19.238	0.000	3.353	4.114

VIF

Features	VIF
Page Views Per Visit	6.06
Lead Origin_Landing Page Submission	5.00
TotalVisits	4.78
Specialization_Management Specialization	3.68
Occupation_Unemployed	2.82
Total Time Spent on Website	2.20
Specialization_Domain Specialization	1.92
Lead Origin_Other	1.66
Occupation_Working Professional	1.43
Lead Source_Other Sources	1.25
Lead Source_Olark Chat	1.21
Do Not Email	1.09
Occupation_Student	1.07
Occupation_Other	1.01

Observations:

If we compare this model with our model 2 then we can see there is very small change in confusion matrix, TP increased a bit and TN reduced by same amount. There is no significant reduction in overall model accuracy.
'Page Views Per Visit' still has bit higher VIF value, so it has been excluded I our next model.

```
Confusion Matrix:
True Negative: 3454      False Positive: 437
False Negative: 824      True Positive: 1573

Overall model accuracy: 0.7994592875318066
```

Model Building : Model 4

Features for Logistic Regression Model 4 (13 Features):

Do Not Email,
TotalVisits,
Total Time Spent on Website,
Lead Origin_Landing Page Submission,
Lead Origin_Other,
Lead Source_Olark Chat,
Lead Source_Other Sources,
Specialization_Domain Specialization,
Specialization_Management Specialization,
Occupation_Other,
Occupation_Student,
Occupation_Unemployed,
Occupation_Working Professional,

Model Summary (Coefficients with p value)

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.3113	0.128	-25.805	0.000	-3.563	-3.060
Do Not Email	-1.1706	0.159	-7.344	0.000	-1.483	-0.858
TotalVisits	0.9315	0.236	3.953	0.000	0.470	1.393
Total Time Spent on Website	3.9502	0.138	28.543	0.000	3.679	4.221
Lead Origin_Landing Page Submission	-0.4672	0.107	-4.364	0.000	-0.677	-0.257
Lead Origin_Other	3.5518	0.194	18.345	0.000	3.172	3.931
Lead Source_Olark Chat	1.2526	0.126	9.975	0.000	1.006	1.499
Lead Source_Other Sources	-0.5321	0.218	-2.443	0.015	-0.959	-0.105
Specialization_Domain Specialization	0.3897	0.123	3.176	0.001	0.149	0.630
Specialization_Management Specialization	0.3960	0.099	4.017	0.000	0.203	0.589
Occupation_Other	1.4862	0.518	2.871	0.004	0.472	2.501
Occupation_Student	1.1348	0.213	5.316	0.000	0.716	1.553
Occupation_Unemployed	1.3129	0.086	15.337	0.000	1.145	1.481
Occupation_Working Professional	3.7285	0.194	19.218	0.000	3.348	4.109

VIF

	Features	VIF
3	Lead Origin_Landing Page Submission	4.69
8	Specialization_Management Specialization	3.67
1	TotalVisits	2.87
11	Occupation_Unemployed	2.74
2	Total Time Spent on Website	2.18
7	Specialization_Domain Specialization	1.92
4	Lead Origin_Other	1.64
12	Occupation_Working Professional	1.43
6	Lead Source_Other Sources	1.24
5	Lead Source_Olark Chat	1.20
0	Do Not Email	1.09
10	Occupation_Student	1.07
9	Occupation_Other	1.01

Observations:

Model accuracy is still almost same as in Model 3. In Model 4, p values of all beta coefficients looks good. We can say that all predictor coefficients are statistically significant. 'Lead Origin_Landing Page Submission' has bit higher VIF value (though it's lesser than 5).
If after dropping 'Lead Origin_Landing Page Submission' there is no significant reduction in overall model accuracy, then we can exclude this feature. In next model 'Lead Origin_Landing Page Submission' has been dropped to test that.

```
Confusion Matrix:
True Negative: 3455      False Positive: 436
False Negative: 826      True Positive: 1571

Overall model accuracy: 0.7993002544529262
```

Model Building : Model 5

Features for Logistic Regression Model 5 (12 Features):

Do Not Email,
TotalVisits,
Total Time Spent on Website,
Lead Origin_Other,
Lead Source_Olark Chat,
Lead Source_Other Sources,
Specialization_Domain Specialization,
Specialization_Management Specialization,
Occupation_Other,
Occupation_Student,
Occupation_Unemployed,
Occupation_Working Professional,

Model Summary (Coefficients with p value)

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.4895	0.123	-28.302	0.000	-3.731	-3.248
Do Not Email	-1.2071	0.159	-7.590	0.000	-1.519	-0.895
TotalVisits	0.8951	0.235	3.807	0.000	0.434	1.356
Total Time Spent on Website	3.9656	0.138	28.680	0.000	3.695	4.237
Lead Origin_Other	3.7946	0.185	20.498	0.000	3.432	4.157
Lead Source_Olark Chat	1.4571	0.118	12.395	0.000	1.227	1.687
Lead Source_Other Sources	-0.5210	0.220	-2.364	0.018	-0.953	-0.089
Specialization_Domain Specialization	0.1536	0.109	1.405	0.160	-0.061	0.368
Specialization_Management Specialization	0.1615	0.082	1.967	0.049	0.001	0.322
Occupation_Other	1.5617	0.512	3.048	0.002	0.558	2.566
Occupation_Student	1.1274	0.213	5.305	0.000	0.711	1.544
Occupation_Unemployed	1.3233	0.085	15.482	0.000	1.156	1.491
Occupation_Working Professional	3.7670	0.193	19.497	0.000	3.388	4.146

VIF

Features	VIF
TotalVisits	2.70
Occupation_Unemployed	2.68
Specialization_Management Specialization	2.27
Total Time Spent on Website	2.16
Lead Origin_Other	1.50
Specialization_Domain Specialization	1.46
Occupation_Working Professional	1.42
Lead Source_Other Sources	1.24
Lead Source_Olark Chat	1.17
Do Not Email	1.06
Occupation_Student	1.06
Occupation_Other	1.01

Observations:

Model 5 has almost same overall accuracy as our previous model. After removing Lead 'Origin_Landing Page Submission', VIF value of 'Specialization_Management Specialization' has also significantly decreased. So, there is no significant multicollinearity anymore in our model. But p value of 'Specialization_Domain Specialization' beta coefficient is now higher than .05. 'Specialization_Domain Specialization' has been excluded in our next model.

```
Confusion Matrix:
True Negative: 3445      False Positive: 446
False Negative: 825      True Positive: 1572

Overall model accuracy: 0.7978689567430025
```

Model Building : Model 6

Features for Logistic Regression Model 6 (11 Features):

Do Not Email,
TotalVisits,
Total Time Spent on Website,
Lead Origin_Other,
Lead Source_Olark Chat,
Lead Source_Other Sources,
Specialization_Management Specialization,
Occupation_Other,
Occupation_Student,
Occupation_Unemployed,
Occupation_Working Professional,

Model Summary (Coefficients with p value)

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.4541	0.120	-28.687	0.000	-3.690	-3.218
Do Not Email	-1.2053	0.159	-7.569	0.000	-1.517	-0.893
TotalVisits	0.9402	0.233	4.038	0.000	0.484	1.397
Total Time Spent on Website	3.9709	0.138	28.724	0.000	3.700	4.242
Lead Origin_Other	3.7902	0.185	20.481	0.000	3.427	4.153
Lead Source_Olark Chat	1.4257	0.115	12.379	0.000	1.200	1.651
Lead Source_Other Sources	-0.5546	0.219	-2.531	0.011	-0.984	-0.125
Specialization_Management Specialization	0.1079	0.073	1.488	0.137	-0.034	0.250
Occupation_Other	1.5788	0.510	3.093	0.002	0.578	2.579
Occupation_Student	1.1398	0.212	5.379	0.000	0.724	1.555
Occupation_Unemployed	1.3308	0.085	15.604	0.000	1.164	1.498
Occupation_Working Professional	3.7871	0.193	19.663	0.000	3.410	4.165

VIF

Features	VIF
Occupation_Unemployed	2.55
TotalVisits	2.48
Total Time Spent on Website	2.15
Specialization_Management Specialization	1.88
Lead Origin_Other	1.49
Occupation_Working Professional	1.39
Lead Source_Other Sources	1.23
Lead Source_Olark Chat	1.16
Do Not Email	1.06
Occupation_Student	1.06
Occupation_Other	1.01

Observations:

Even after droping 'Specialization_Management Specialization' there is no significance change in our overall model accuracy.
Now, we can see that beta coefficient of 'Specialization_Management Specialization' has higher p value.
Multicollinearity is not present in Model 6. 'Specialization_Management Specialization' has been dropped in our next model as its beta coefficient not statistically significant.

```
Confusion Matrix:
True Negative: 3445      False Positive: 446
False Negative: 828      True Positive: 1569

Overall model accuracy: 0.7973918575063613
```

Model Building : Model 7

Features for Logistic Regression Model 7 (10 Features):

Do Not Email,
TotalVisits,
Total Time Spent on Website,
Lead Origin_Other,
Lead Source_Olark Chat,
Lead Source_Other Sources,
Occupation_Other,
Occupation_Student,
Occupation_Unemployed,
Occupation_Working Professional,

Model Summary (Coefficients with p value)

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.4116	0.117	-29.245	0.000	-3.640	-3.183
Do Not Email	-1.2128	0.159	-7.608	0.000	-1.525	-0.900
TotalVisits	0.9544	0.233	4.102	0.000	0.498	1.410
Total Time Spent on Website	3.9791	0.138	28.804	0.000	3.708	4.250
Lead Origin_Other	3.7822	0.185	20.457	0.000	3.420	4.145
Lead Source_Olark Chat	1.3881	0.112	12.385	0.000	1.168	1.608
Lead Source_Other Sources	-0.5853	0.218	-2.683	0.007	-1.013	-0.158
Occupation_Other	1.5865	0.511	3.107	0.002	0.586	2.587
Occupation_Student	1.1444	0.212	5.406	0.000	0.730	1.559
Occupation_Unemployed	1.3392	0.085	15.739	0.000	1.172	1.506
Occupation_Working Professional	3.8140	0.192	19.913	0.000	3.439	4.189

VIF

Features	VIF
Occupation_Unemployed	2.39
TotalVisits	2.34
Total Time Spent on Website	2.11
Lead Origin_Other	1.49
Occupation_Working Professional	1.32
Lead Source_Other Sources	1.22
Lead Source_Olark Chat	1.16
Do Not Email	1.05
Occupation_Student	1.05
Occupation_Other	1.01

Observations:

After droping 'Specialization_Management Specialization' our model accuracy has been improved a bit.
After droping 'Specialization_Management Specialization', we can see that all the beta coefficients are now statistically significant also there is no multicolinearity present in Model 7.

```
Confusion Matrix:
True Negative: 3453      False Positive: 438
False Negative: 826      True Positive: 1571

Overall model accuracy: 0.7989821882951654
```

Prediction & Model Evaluation : (on Training data - cutoff .5)

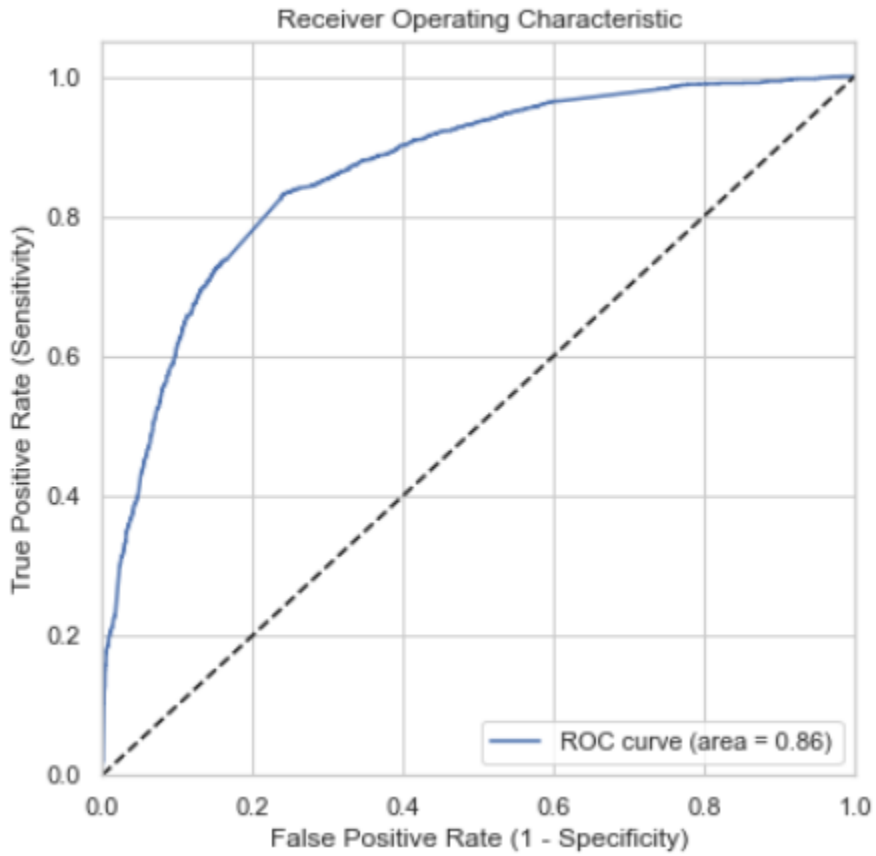
1. We have used Model 7 to predict the probability for all the observations in out training dataset and then used .5 as probability cut off to calculate our target variable ‘Converted’. So, if probability is > .5 then ‘Converted’= 1 (Yes) otherwise 0 (No).

2. After predicting the target on our training data set, we calculated different evaluation metrics as below:

```
Overall model accuracy: 0.7989821882951654
Sensitivity / Recall:  0.6554025865665415
Specificity:  0.887432536622976
False Positive Rate:  0.1125674633770239
Positive Predictive Value:  0.7819810851169736
Positive Predictive Value:  0.8069642439822389
```

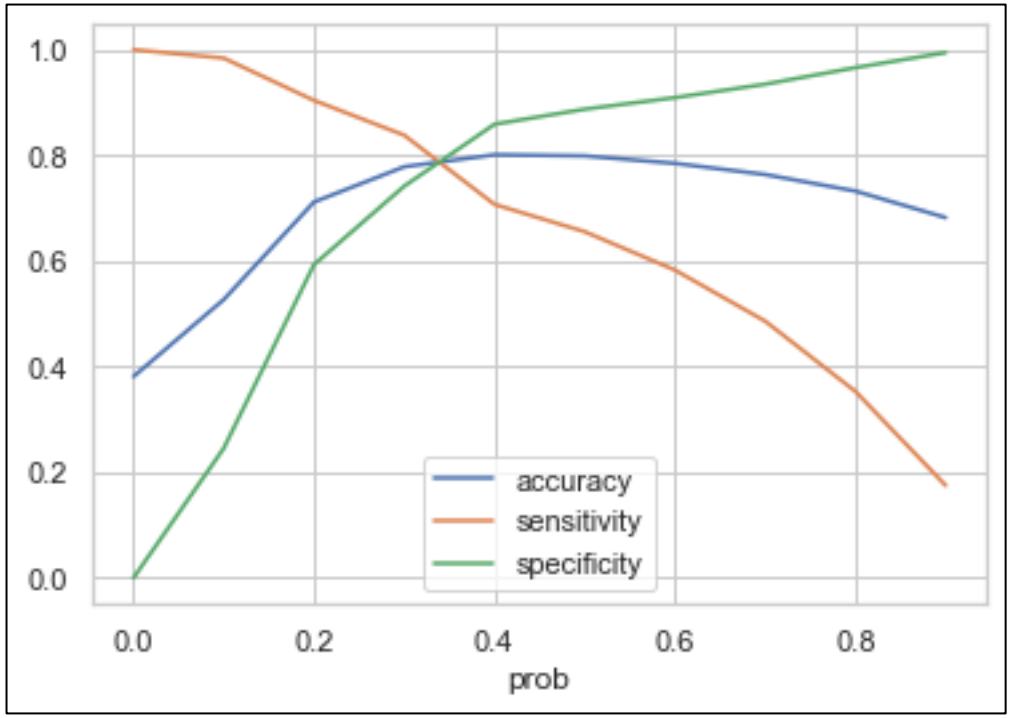
```
Confusion Matrix:
True Negative: 3453      False Positive: 438
False Negative: 826      True Positive: 1571

Overall model accuracy: 0.7989821882951654
```



3. Our model has bad Sensitivity (with probability cut-off .5). By doing trade-off between Sensitivity-Specificity optimal probability cut-off value has been calculated.

Finding Optimal Probability cutoff & Evaluating on Train Data



In above plot, it's visible that 0.32 is the optimal point to set as cutoff probability for our model.

Model Evaluation Metrics on Train dataset with probability cutoff .32

```
Model Evaluation Metrics on Train dataset
#####################################################
Confusion Matrix:
True Negative: 2946      False Positive: 945
False Negative: 402      True Positive: 1995

Overall model accuracy: 0.7857824427480916
Sensitivity / Recall:  0.8322903629536921
Specificity:  0.7571318427139553
False Positive Rate:  0.24286815728604472
Positive Predictive Value:  0.6785714285714286
Positive Predictive Value:  0.8799283154121864
```

Observations:

Sensitivity of our model has been increased without any significant reduction in overall accuracy. New Specificity is also in well accepted range.

Prediction & Generating Lead Score (Business Requirement)

Generating Lead Score on Train Dataset

Generating Lead Score on Train dataset

```
# Adding Lead Number from initial dataframe for better understanding

y_train_pred= y_train_pred.merge(leads_df['Lead Number'], how= 'left', left_index= True, right_index= True)

# Generating Lead Score on Train dataset
y_train_pred['Lead Score']= y_train_pred.prob * 100
y_train_pred= y_train_pred[['Lead Number', 'Converted', 'pred_Converted', 'prob','Lead Score']].sort_values(
    'Lead Score', ascending= False)
y_train_pred.head(10)
```

	Lead Number	Converted	pred_Converted	prob	Lead Score
818	651812	1	1	0.999804	99.980378
2656	634047	1	1	0.999693	99.969269
3478	627106	1	1	0.999669	99.966924
6383	600952	1	1	0.999649	99.964935
5921	604411	1	1	0.999364	99.936442
7579	591536	1	1	0.999325	99.932475
6751	598055	1	1	0.999257	99.925736
8081	588013	1	1	0.999058	99.905798
9015	581257	1	1	0.998964	99.896399
120	659283	1	1	0.998851	99.885074

Generating Lead Score on Test Dataset

```
# Generating Lead Score on Test dataset
y_test_pred['Lead Score']= y_test_pred.prob * 100
y_test_pred= y_test_pred[['Lead Number', 'Converted', 'pred_Converted', 'prob','Lead Score']].sort_values(
    'Lead Score', ascending= False)
y_test_pred.head(10)
```

	Lead Number	Converted	pred_Converted	prob	Lead Score
8074	588037	1	1	0.999642	99.964168
3428	627462	1	1	0.999444	99.944417
8063	588075	1	1	0.999068	99.906811
4613	615524	1	1	0.998992	99.899175
2984	631268	1	1	0.998894	99.889355
7187	594369	1	1	0.998119	99.811861
8057	588097	0	1	0.997273	99.727347
79	659710	1	1	0.997238	99.723794
2978	631318	1	1	0.997207	99.720731
739	652787	1	1	0.996684	99.668376

Using Model 7 we calculated the probability on Test dataset and used cutoff =.32 to predict the pred_Converted (0,1). As per business requirement we have created a column Lead Score (between 0 to 100) of the leads. A higher score means hot lead (most likely to convert), lower score implies cold lead (mostly not get converted). We have multiplied the probability (pred_Converted) with 100 to generate the Lead Score.

Model Evaluation : (on Test data) & Interpretation

Model Evaluation Metrics on Test dataset with probability cutoff .32

```
Model Evaluation Metrics on Test dataset
#####################################################
Confusion Matrix:
True Negative: 1258      False Positive: 402
False Negative: 203      True Positive: 832

Overall model accuracy: 0.7755102040816326
Sensitivity / Recall:  0.8038647342995169
Specificity:  0.7578313253012048
False Positive Rate:  0.2421686746987952
Positive Predictive Value:  0.6742301458670988
Positive Predictive Value:  0.8610540725530459
```

Model is performing well on test data with Sensitivity= 80%, Specificity= 76% and overall accuracy: 78%,

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.4116	0.117	-29.245	0.000	-3.640	-3.183
Do Not Email	-1.2128	0.159	-7.608	0.000	-1.525	-0.900
TotalVisits	0.9544	0.233	4.102	0.000	0.498	1.410
Total Time Spent on Website	3.9791	0.138	28.804	0.000	3.708	4.250
Lead Origin_Other	3.7822	0.185	20.457	0.000	3.420	4.145
Lead Source_Olark Chat	1.3881	0.112	12.385	0.000	1.168	1.608
Lead Source_Other Sources	-0.5853	0.218	-2.683	0.007	-1.013	-0.158
Occupation_Other	1.5865	0.511	3.107	0.002	0.586	2.587
Occupation_Student	1.1444	0.212	5.406	0.000	0.730	1.559
Occupation_Unemployed	1.3392	0.085	15.739	0.000	1.172	1.506
Occupation_Working Professional	3.8140	0.192	19.913	0.000	3.439	4.189

Top 3 variables which contribute most towards the probability of a lead getting converted:

- **Total Time Spent on Website**
- **What is your current occupation (Working Professional)**
- **Lead origin (Other)**

Conclusion and Recommendations

1. As per business requirement Lead Score (between 0 to 100) of the leads have been calculated by using this Logistic Regression model. A higher score means hot lead (most likely to convert), lower score implies cold lead (mostly not get converted).
2. This Lead Score would help to identify the hot leads faster and efficiently, that would result in **decrease in lead conversion time** and **increase in lead conversion rate.** Leads should be sorted in descending order according to their Lead Scores.
3. Phone calls or contact should be made to the leads having higher Lead Score first. Some special attentions should be provided to these hot leads (may be assigning a dedicated support SPOC for a small batch of hot leads that have higher Lead Scores), as there is a very high chance of Lead conversion.
4. Leads having medium Lead Score are also potentially good candidates for Lead conversion. They also should be contacted, and right questions should be asked, so that business can understand their requirements and problem areas and can take necessary actions. Few to mention: some changes in existing courses, introducing new courses, some change in class schedules, introducing easy financial options for fees etc. may help to successfully convert these leads.
5. Cold Leads should be contacted after business gets very good Conversion rate with the leads having High and Medium Lead scores. As chance of conversion is very less here, they could be part of company’s aggressive marketing policy.