Lead Scoring Case Study

- *PRESENTED BY – Chetan Kadam, Kenneth Lobo & Rahul Game*

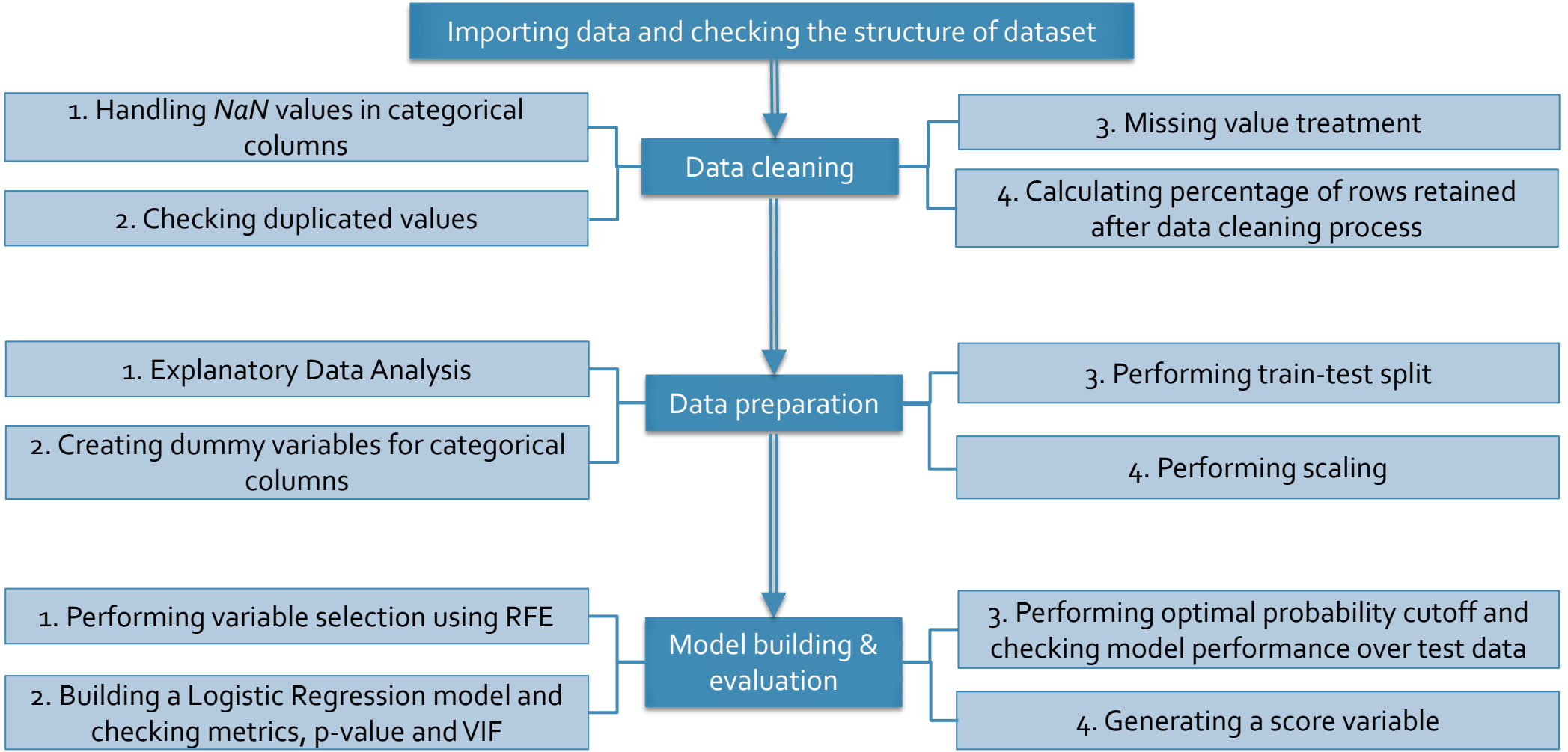
PROBLEM STATEMENT

- X Education is an education company which sells online courses to industry professionals.
- 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.
- Although X Education gets a lot of leads, its lead conversion rate is very poor.

Business Target *:*

- Selecting the most promising leads
- Building a model and deploying it for future usage
- Assigning a lead score to each of the leads and identifying 'Hot Leads'

ANALYSIS APPROACH



DATA CLEANING

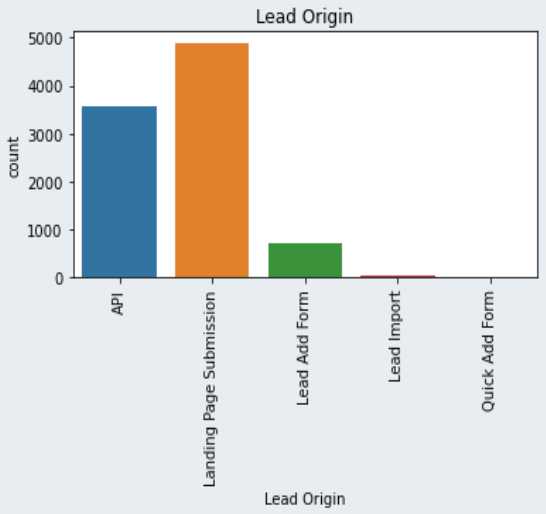
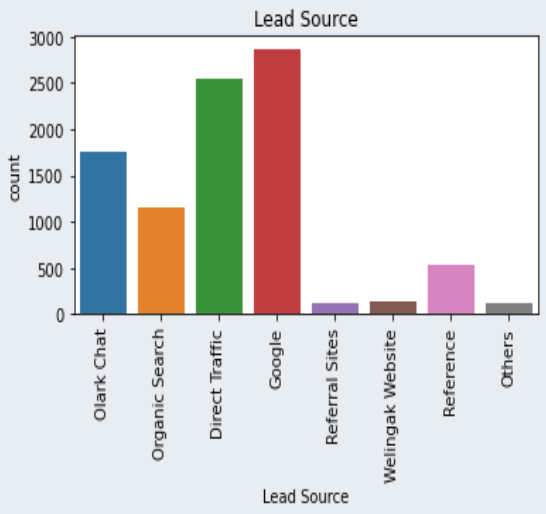
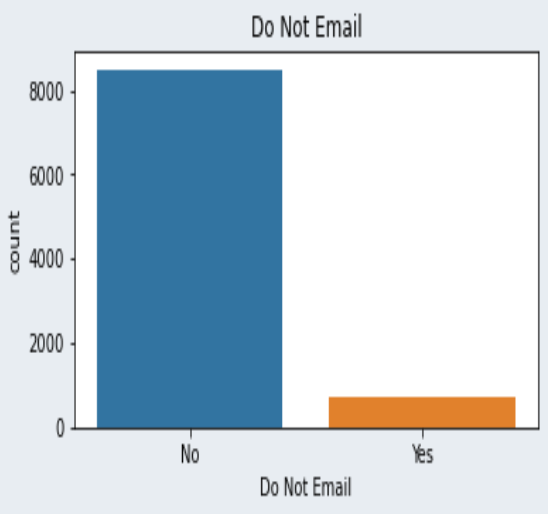
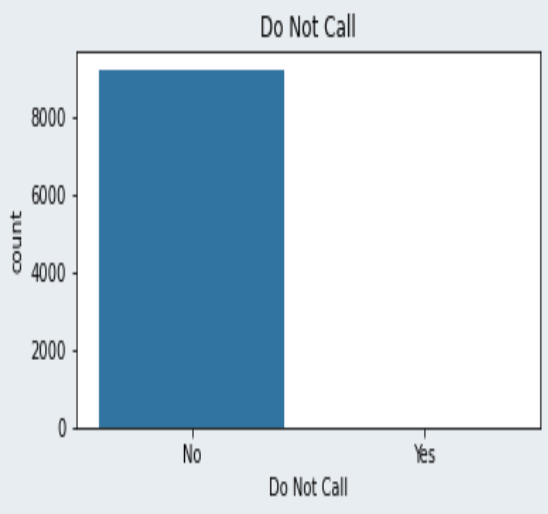
- Categorical columns having more than 40% missing values have been removed. These are –

Variable	Missing Value
How did you hear about X Education	78.46%
Lead Profile	74.19%
Lead Quality	51.59%
Asymmetrique Activity Index	45.65%
Asymmetrique Profile Index	45.65%
Asymmetrique Activity Score	45.65%
Asymmetrique Profile Score	45.65%

- The heavily skewed categorical columns have also been removed. These are –
  - *Country*
  - *What matters most to you in choosing a course*
- The missing values in the other categorical columns have been imputed and merged with the low frequency variable values.
- The missing values in the numerical columns have been imputed with the respective medians.

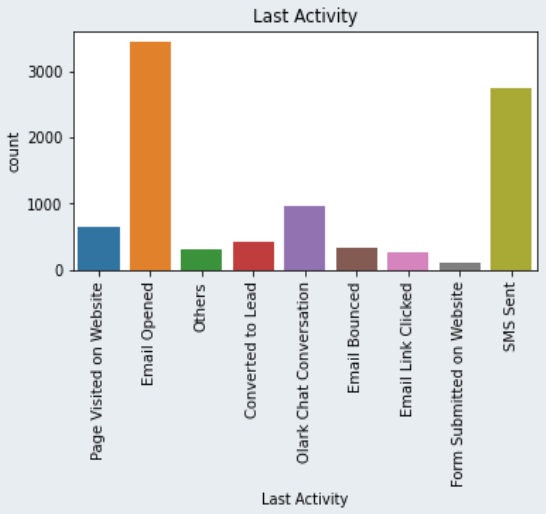
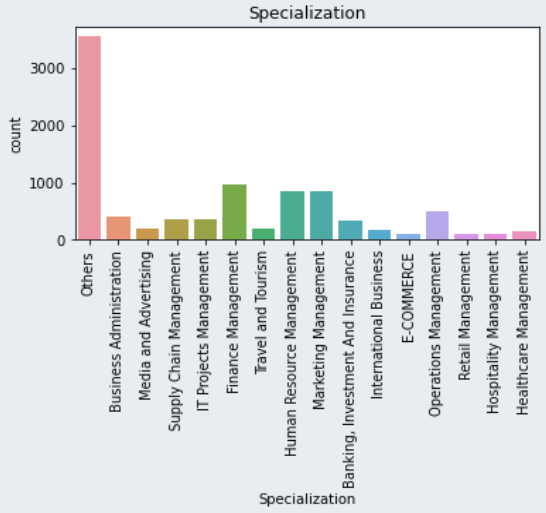
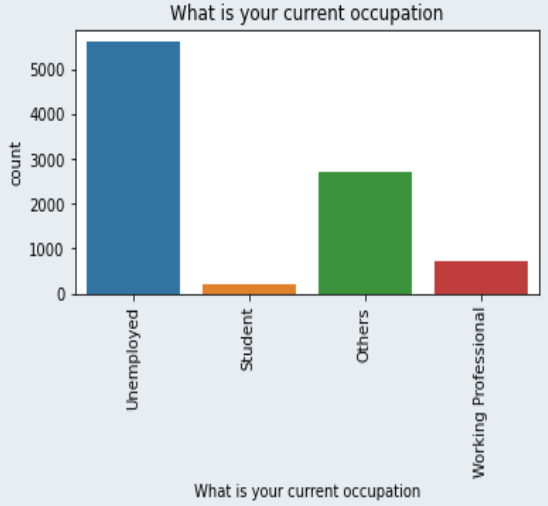
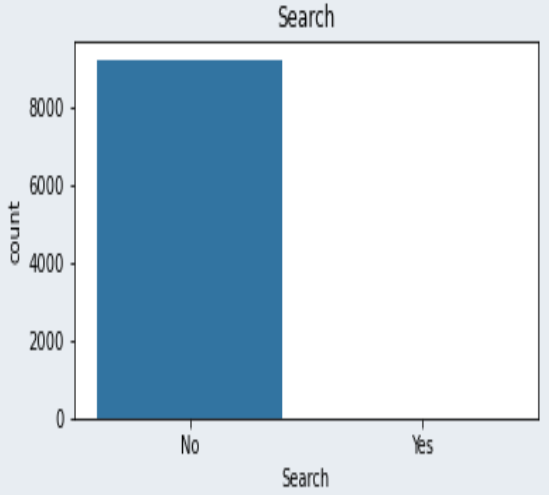
EXPLANATORY DATA ANALYSIS

Univariate Analysis of Categorical Columns

Customers are identified to be leads mainly through *Landing Page Submission*, *API* and *Lead Add Form*. The contributions of *Lead Import* and *Quick Add Form* are the least.	Most leads are contributed through *Google*, followed by *Direct Traffic* and *Olark Chat*.	Most customers prefer not to be emailed about the course.	Most customers prefer not to be called to discuss about the course. The count for customers who opt to discuss about the course over phone is significantly less. This will add no valuable information to the model and so we drop this variable.
			

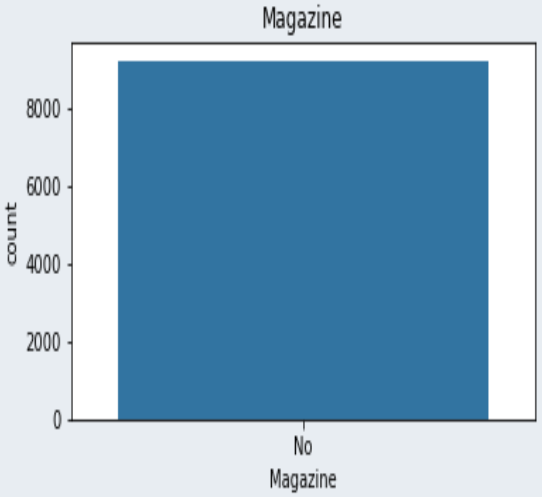
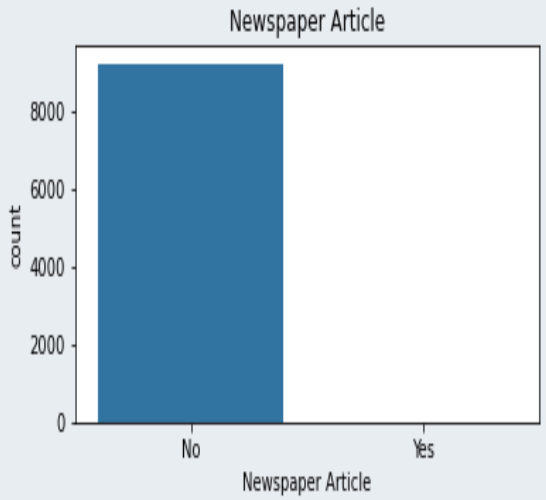
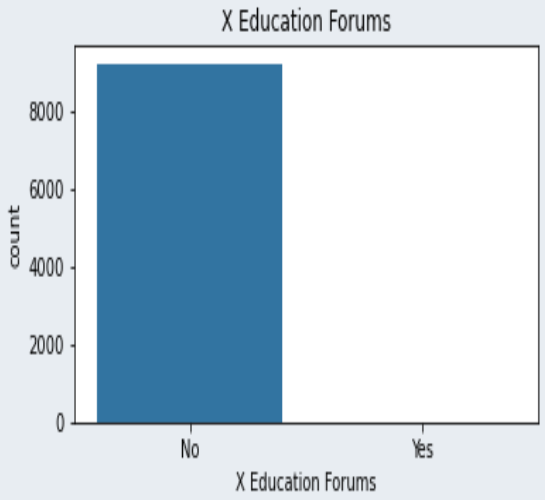
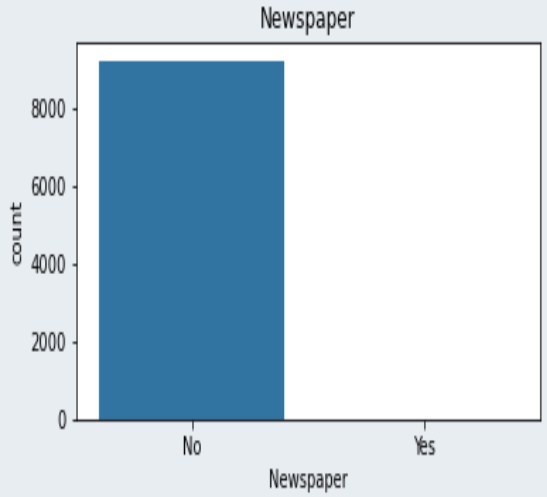
EXPLANATORY DATA ANALYSIS

Univariate Analysis of Categorical Columns

Most customers have sent *Email* and *SMS* while some interacted through *Olark Chat Conversation* or visited *Page on Website*.	Most customers did not select the specialization which has been grouped into *Others*.	Maximum customers are *Unemployed* while *Working Professionals* are interested than *Students*.	Most customers have not seen the advertisement in *Search*. The count for customers have seen is negligible. This will add no valuable information to the model and so we drop this variable.
			

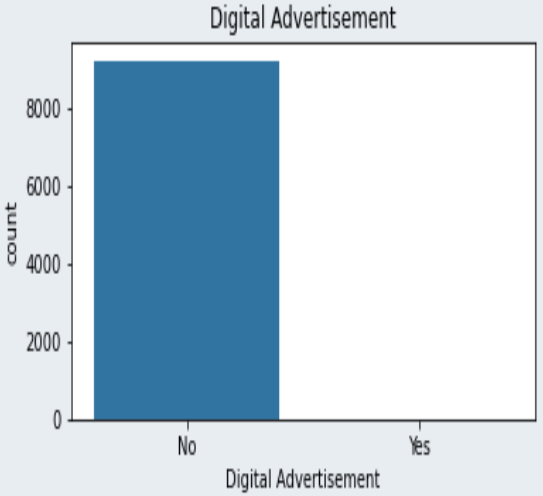
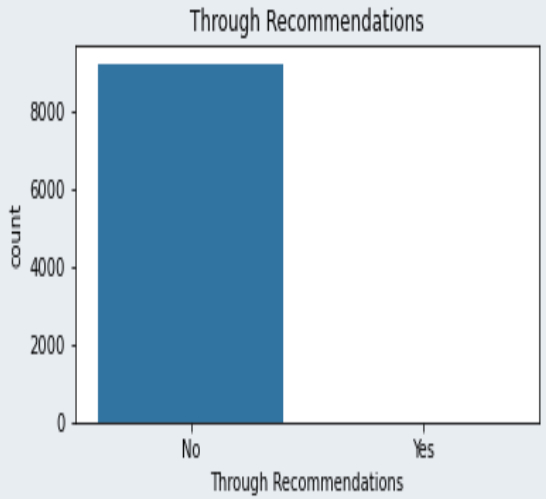
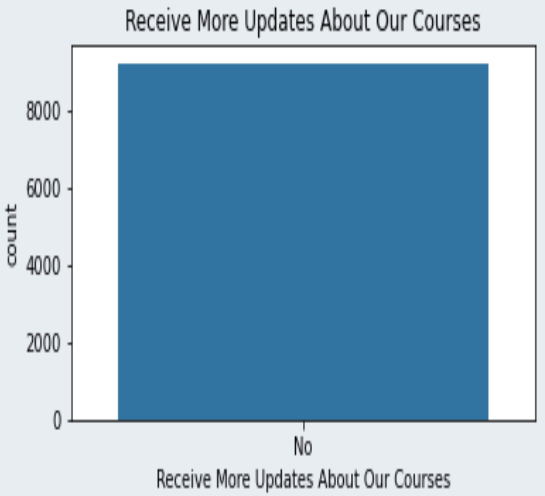
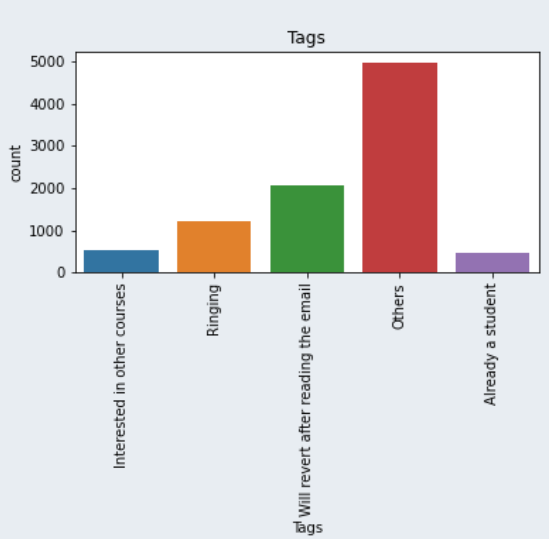
EXPLANATORY DATA ANALYSIS

Univariate Analysis of Categorical Columns

Most customers have not seen the advertisement in *Magazine*. The count for customers have seen is negligible. This will add no valuable information to the model and so we drop this variable.	Most customers have not seen the advertisement in *Newspaper Article*. As the count for customers have seen is negligible, this will add no valuable information to the model and so we drop this variable.	Most customers have not seen the advertisement in *X Education Forums*. This will add no valuable information to the model and so we drop this variable.	Most customers have not seen the advertisement in *Newspaper*. As the count for customers have seen is negligible, this will add no valuable information to the model and so we drop this variable.
			

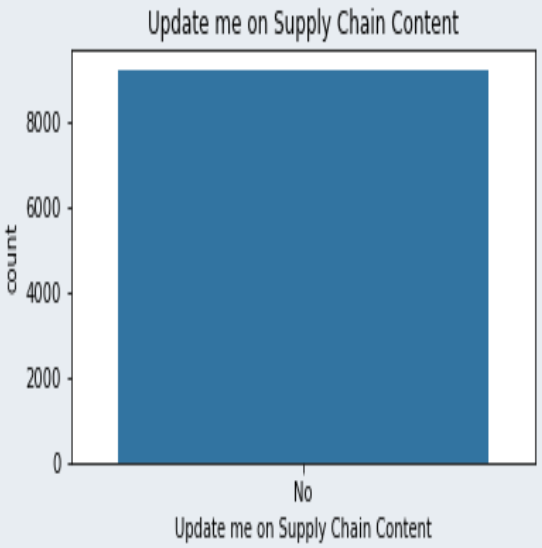
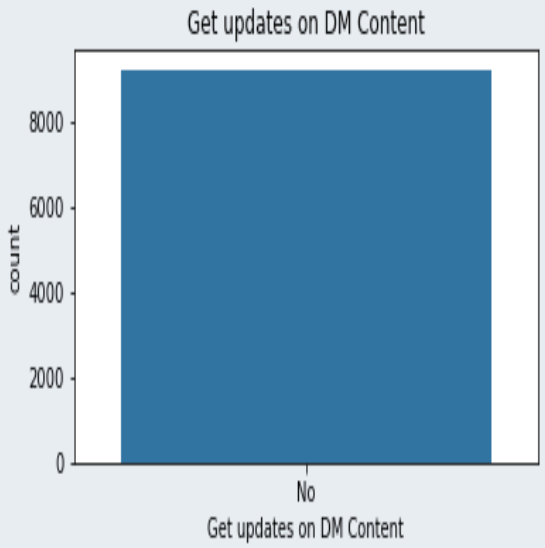
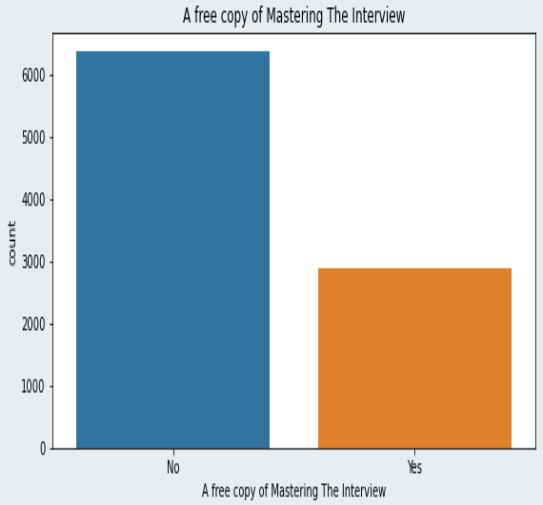
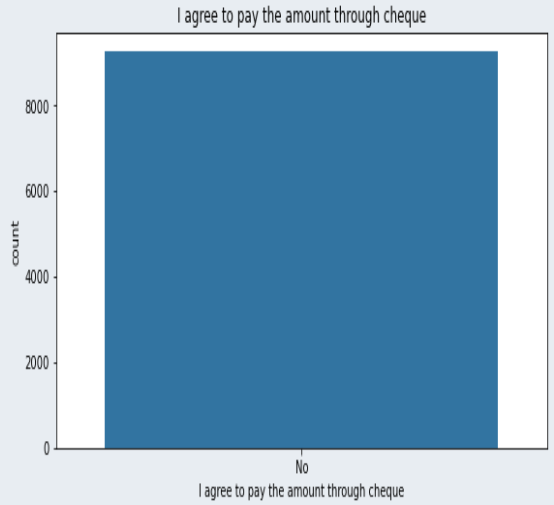
EXPLANATORY DATA ANALYSIS

Univariate Analysis of Categorical Columns

Most customers have not seen the *Digital Advertisement*. As this will add no valuable information to the model and so we drop this variable.	The majority of customers did not come *Through Recommendations*. As this will add no valuable information to the model and so we drop this variable.	No customer wants to *Receive More Updates About Our Course*. So we drop this variable.	The current status of leads have not been assigned in most cases. However, some leads *revert after reading the email*.
			

EXPLANATORY DATA ANALYSIS

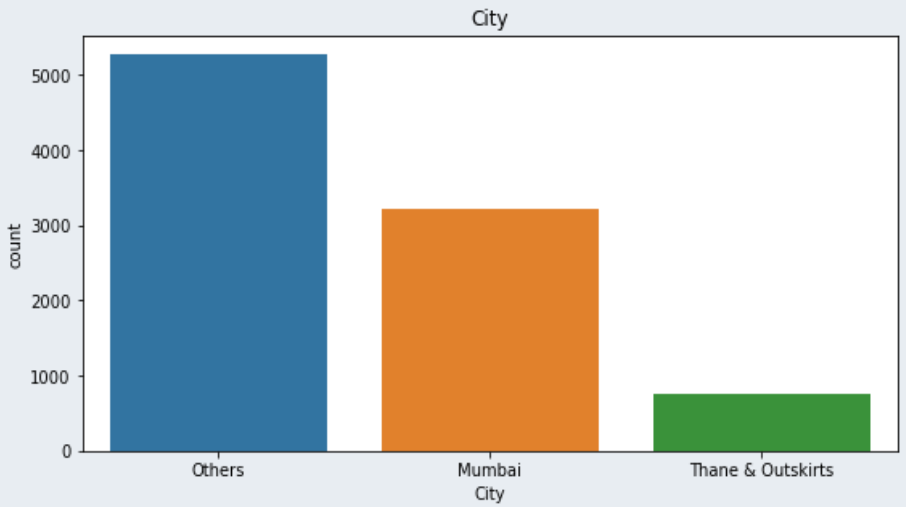
Univariate Analysis of Categorical Columns

No customer wants to *Update on Supply Chain Content*. So we drop this variable.	No customer wants to *Get updates on DM Content*. So we drop this variable.	Majority of the customers do not want *a free copy of 'Mastering The Interview'*. Although some customers demand it. So, we will retain this variable and further analyze it.	No customer *agrees to pay the amount through cheque*. So this variable adds no importance to the model and hence we drop it.
			

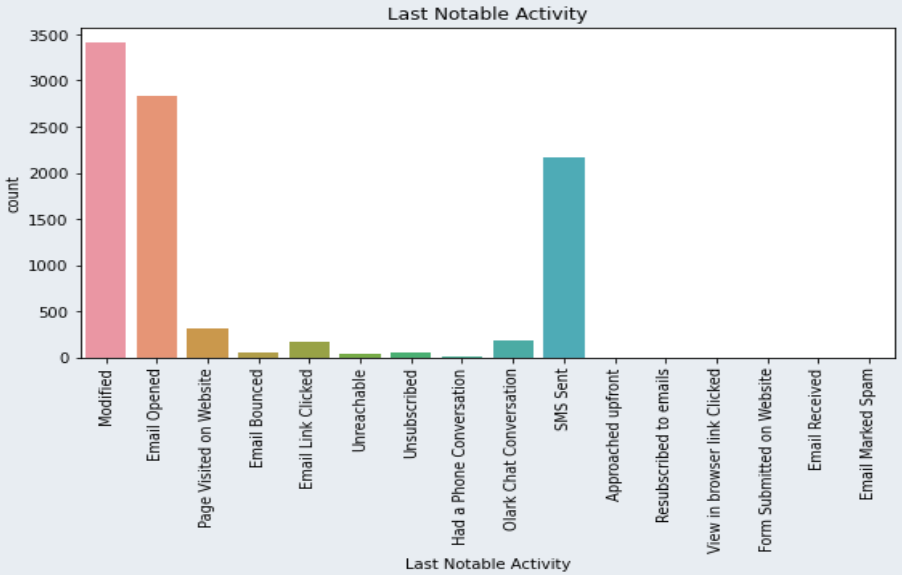
EXPLANATORY DATA ANALYSIS

Univariate Analysis of Categorical Columns

Many customers belong to *Mumbai City*. So *Mumbai*, *Thane & Outskirts* contribute to many leads converting to customers.



Most customers have sent *Email* and *SMS*.

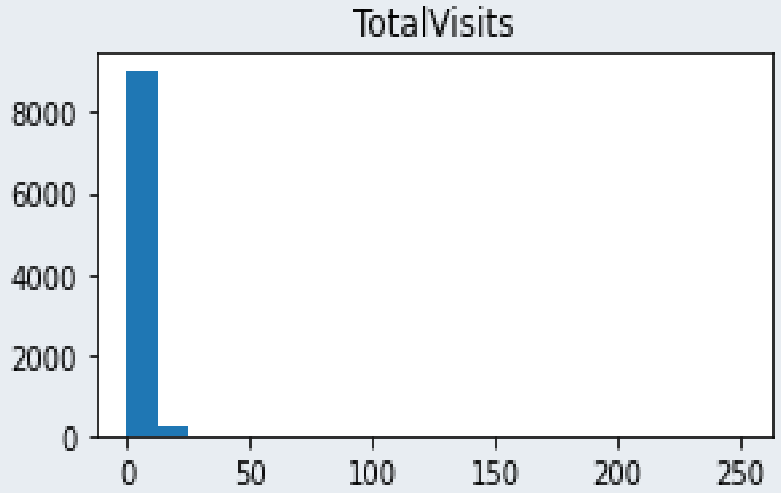


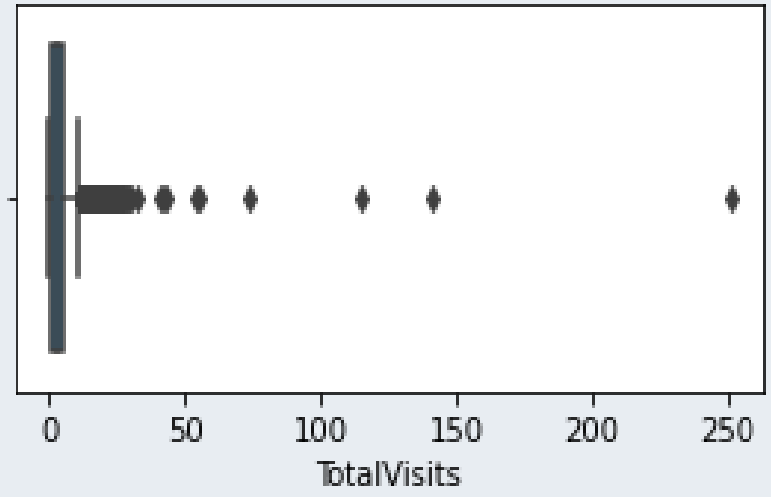
EXPLANATORY DATA ANALYSIS

Univariate Analysis of Numerical Columns

The histogram plot of *TotalVisits* data has a high peak and appears skewed. So we have to check for outliers.

There are many outliers in the *TotalVisits* data. So this variable needs outlier treatment.

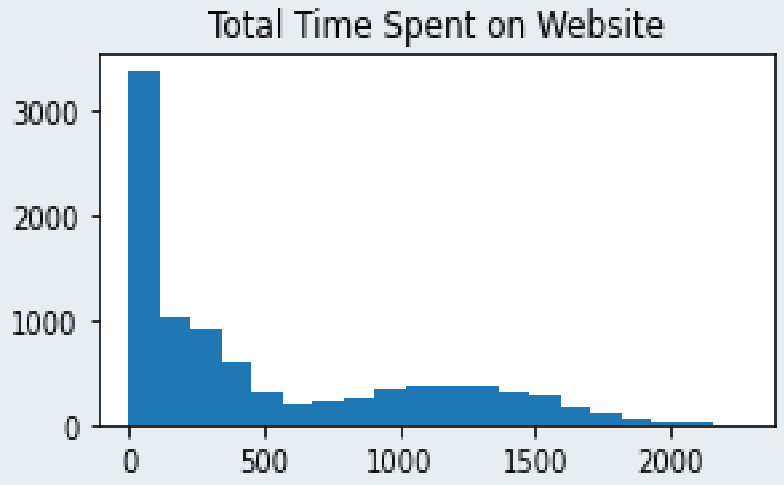




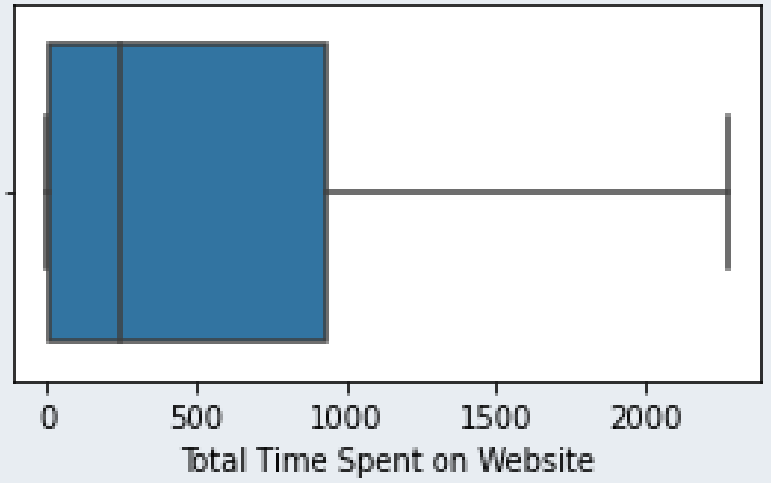
EXPLANATORY DATA ANALYSIS

Univariate Analysis of Numerical Columns

The histogram plot of *Total Time Spent on Website* variable does not show high skewness.



The box plot of *Total Time Spent on Website* variable does not seem to have outliers.

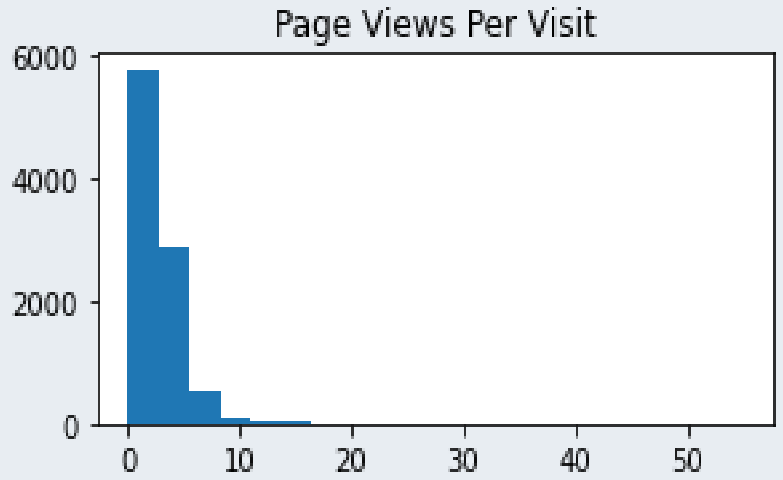


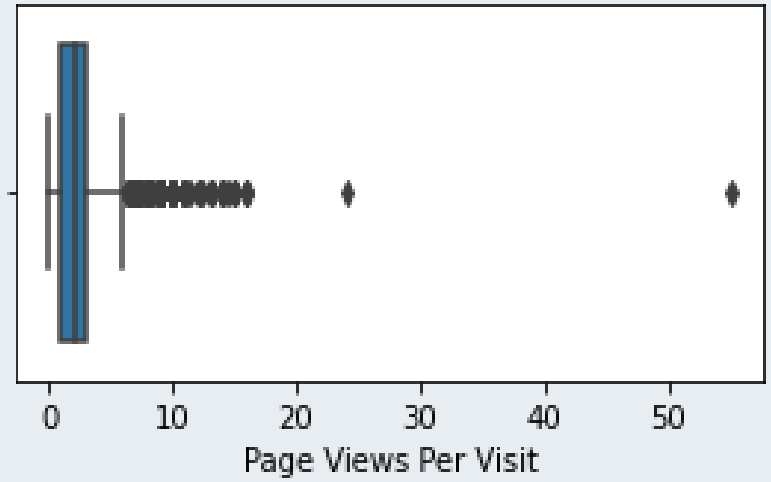
EXPLANATORY DATA ANALYSIS

Univariate Analysis of Numerical Columns

The *Page Views Per Visit* data seems to have high peak and appears skewed. So we have to check for outliers.

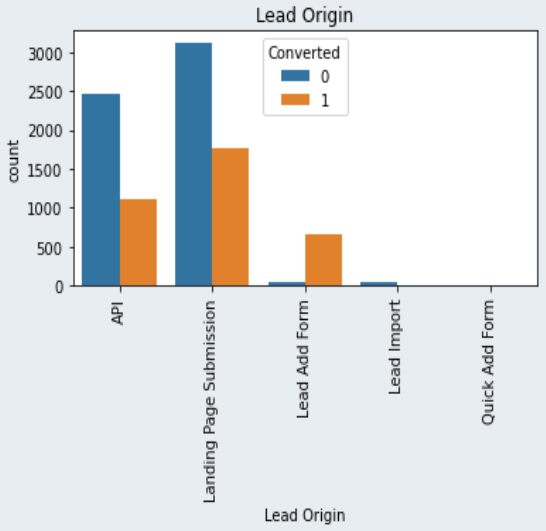
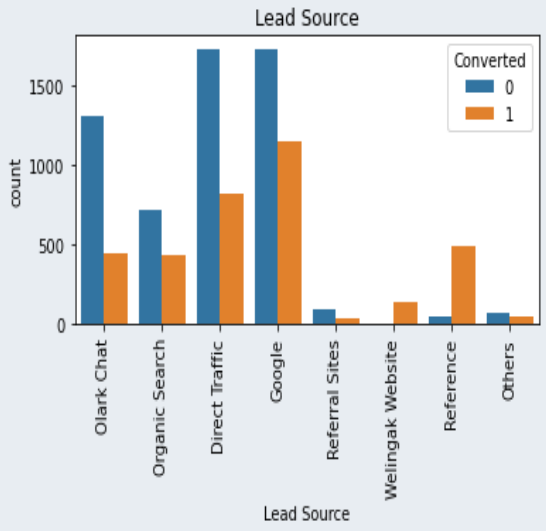
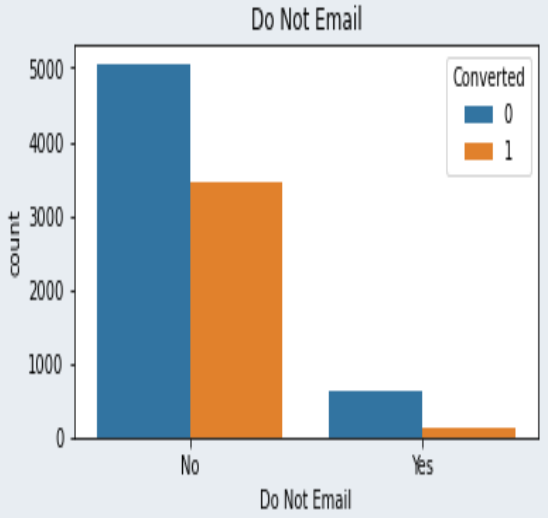
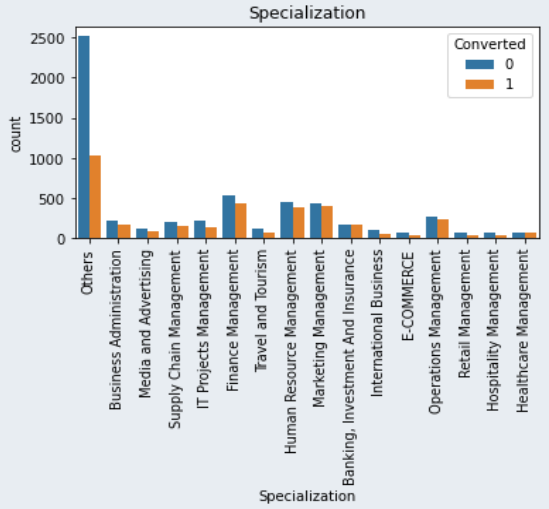
There are many outliers in the *Page Views Per Visit* data. So this variable needs outlier treatment.





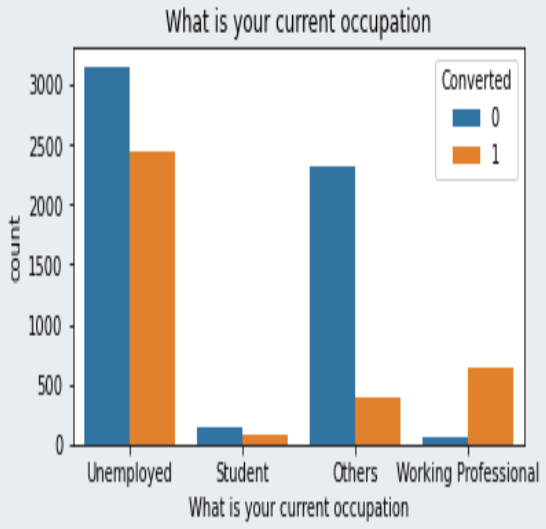
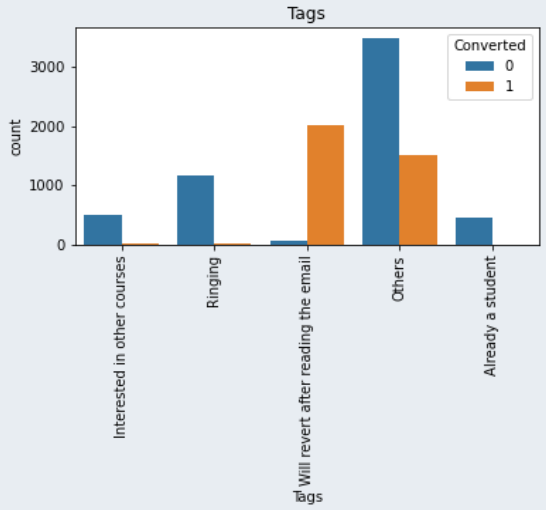
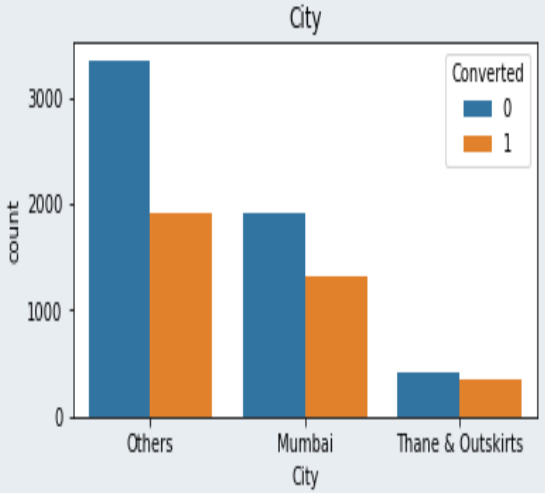
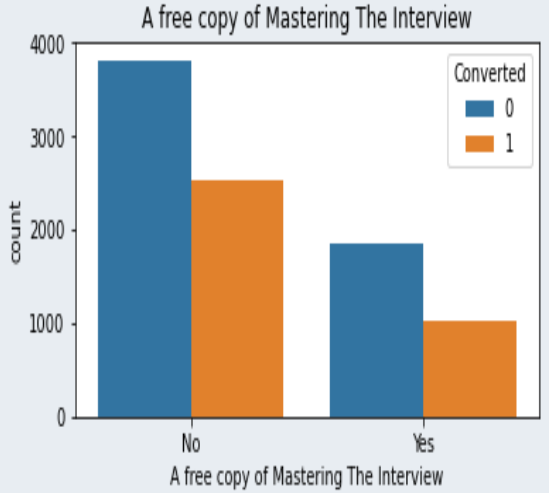
EXPLANATORY DATA ANALYSIS

Bivariate Analysis of Categorical Columns

Leads originating from *Lead Add Form* have a very high conversion rate compared to those from *API* or *Landing Page Submission*.	Leads contributed through *Reference*, *Welingak Website*, *Google*, *Direct Traffic* and *Olark Chat* have high conversion rate.	Although most customers prefer not to be emailed about the course, but they have a significant conversion rate.	Leads having background in *Finance*, *Human Resource*, *Marketing* and *Operations Management* and who are from *Banking, Investment And Insurance* have a high rate of conversion.
			

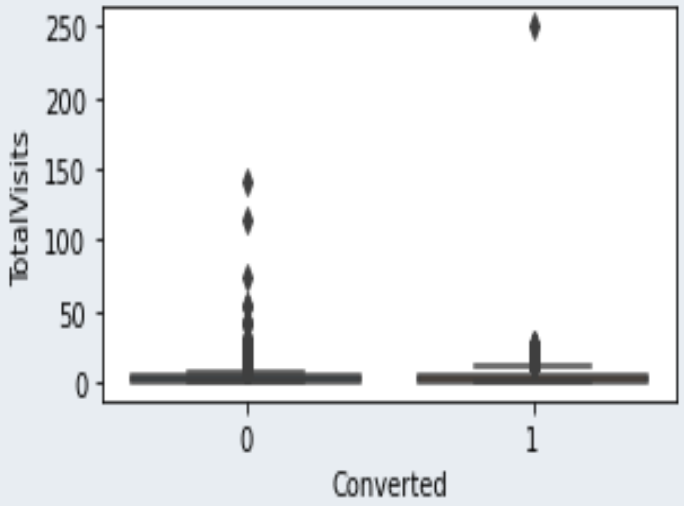
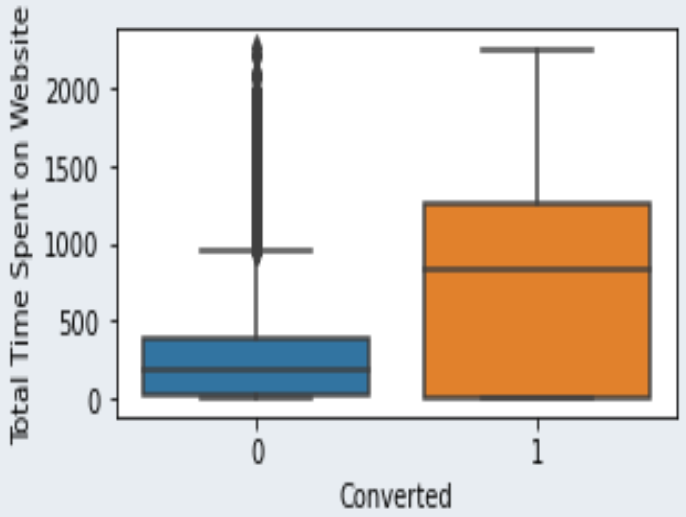
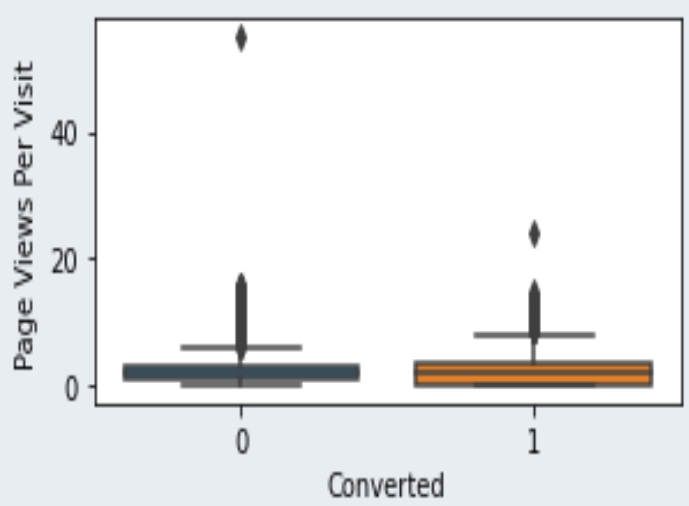
EXPLANATORY DATA ANALYSIS

Bivariate Analysis of Categorical Columns

Working Professionals are the most potential leads who can be converted. They are followed by *Student* and *Unemployed* leads. So *Working Professionals* should be targeted for conversion.	Conversion rate of leads who opt to *revert after reading the email* is very high. So, sending the course details over email is a very workable method of converting leads into potential customers.	*Mumbai*, *Thane & Outskirts* have a significant conversion rate. So leads in that region should be targeted for conversion.	Leads who have either asked for *a free copy of 'Mastering Interview'* or not, have converted.
			

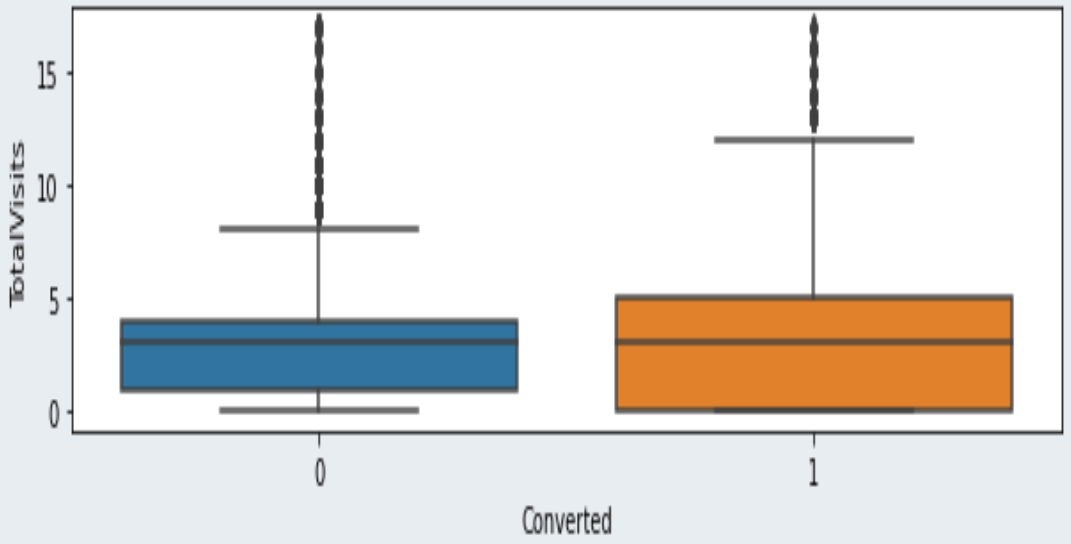
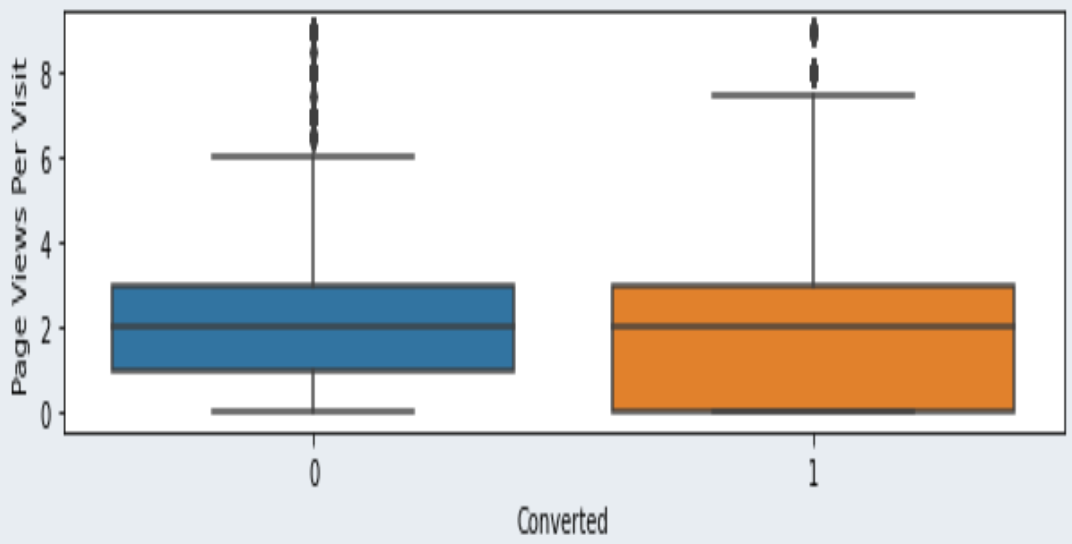
EXPLANATORY DATA ANALYSIS

Bivariate Analysis of Numerical Columns

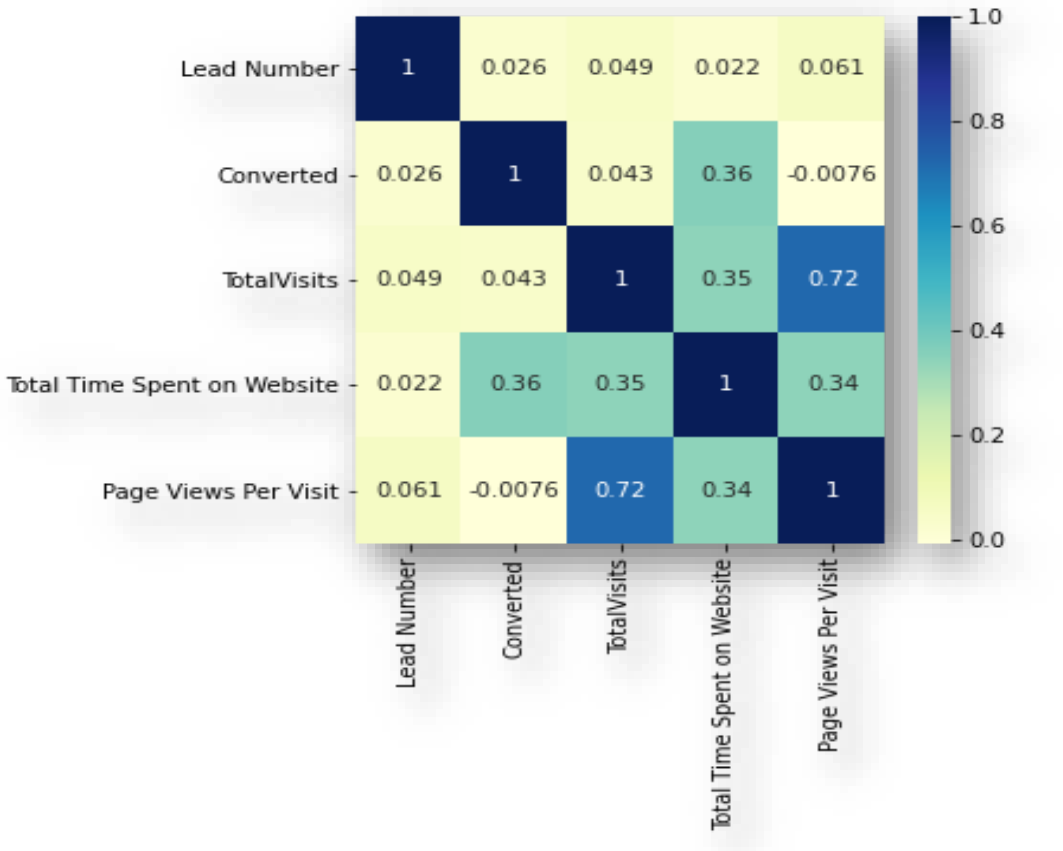
The box plot of the *TotalVisits* data of converted leads have outliers. So this variable needs outlier treatment.	The box plot of the *Total Time Spent on Website* data of converted leads do not have outliers. So we can use this data.	The box plot of the *Page Views Per Visit* data of converted leads have outliers. So this variable needs outlier treatment.
		

EXPLANATORY DATA ANALYSIS

Outlier Treatment

The top and bottom 1% of the outlier values have been removed and the boxplot of the *TotalVisits* data has been plotted.	The top and bottom 1% of the outlier values have been removed and the boxplot of the *Page Views Per Visit* data has been plotted.
	

EXPLANATORY DATA ANALYSIS



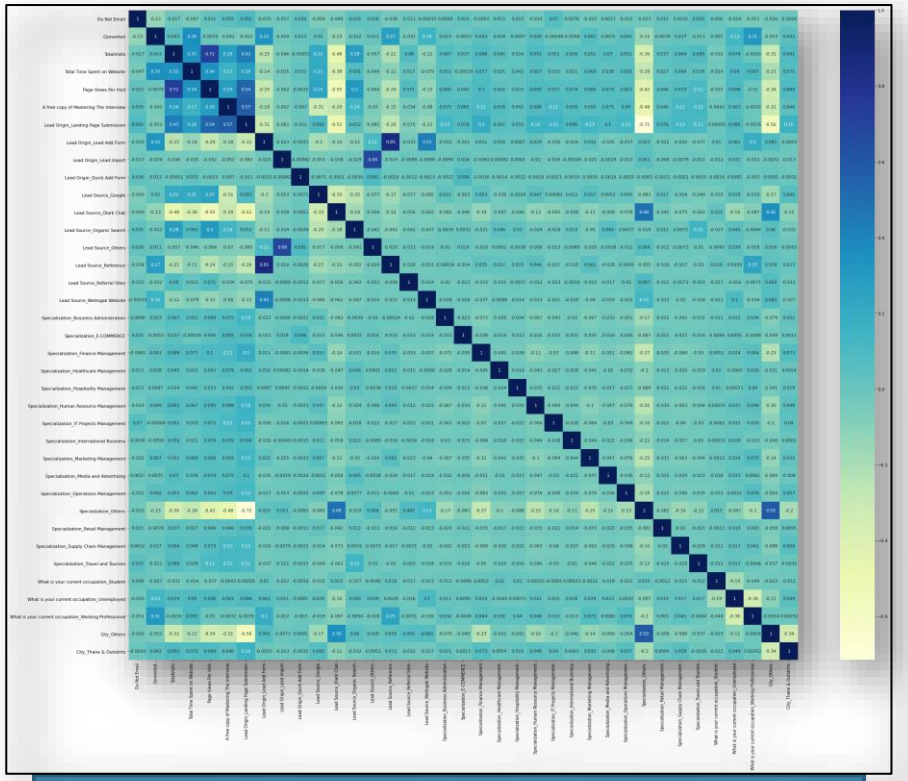
Heatmap showing correlation between the numerical columns

Correlation

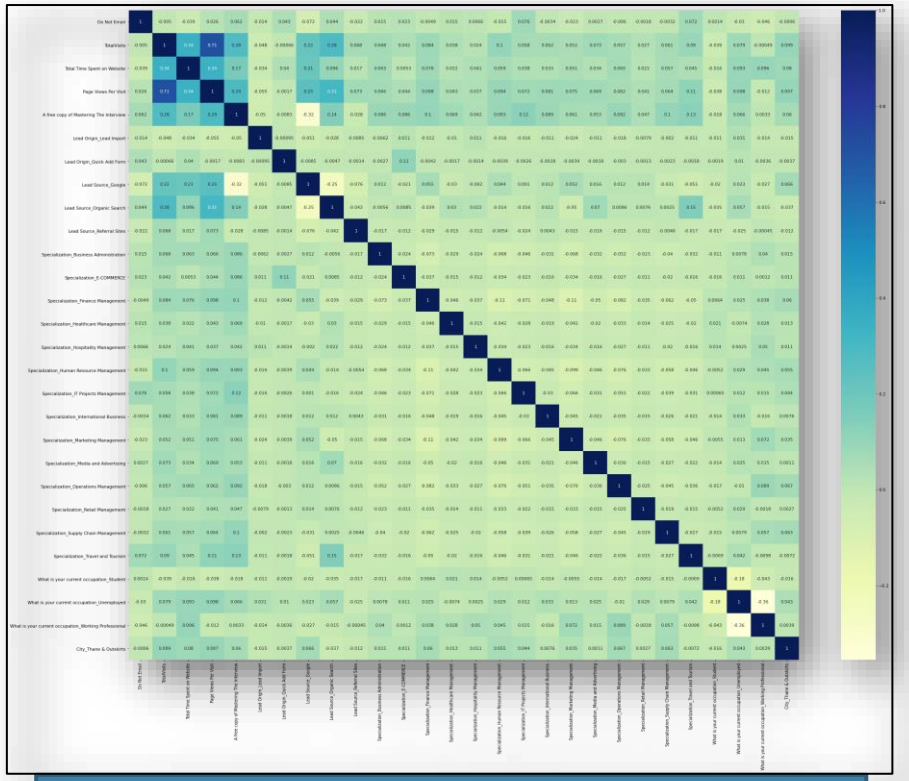
- The heatmap shows that *TotalVisits* have a strong correlation with *Page Views Per Visit*.
- The correlation between *TotalVisits* and *Total Time Spent on Website* is also good.

TRAIN-TEST SPLIT & SCALING

- After creating the dummy variables for the categorical columns, the data is split into train (70%) and test (30%) data.
- The data is then normalized using MinMax scaler and the conversion rate is found to be 38%.
- The correlation is plotted in a heatmap and the highly correlated (r >0.4) dummy variables are dropped from the data.



Heatmap showing correlations before dropping dummies



Heatmap showing correlations after dropping dummies

MODEL BUILDING & EVALUATION

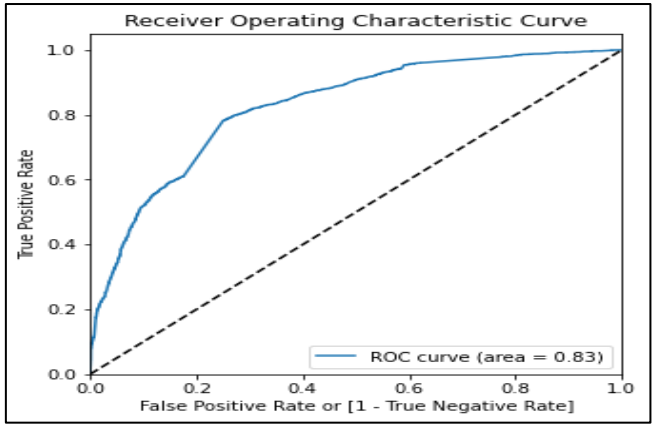
- A logistic regression model (with 75.5% accuracy) is build using RFE and after repeated iterations, the following variables with insignificant p-values (> 0.05) or high VIF (> 5) values are dropped –
  - *Specialization_E-COMMERCE*
  - *Specialization_Retail Management*
  - *Specialization_Hospitality Management*
  - *Page Views Per Visit*
  - *TotalVisits*
- The model has no multicollinear variables and therefore we will make predictions using this model.

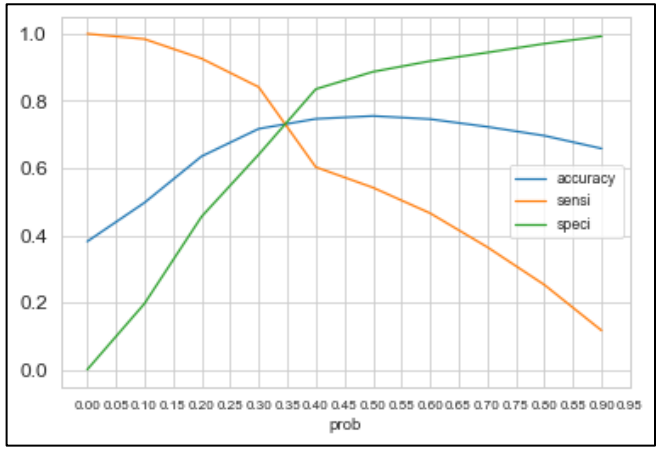
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.9813	0.081	-24.578	0.000	-2.139	-1.823
Do Not Email	-1.2549	0.147	-8.521	0.000	-1.544	-0.966
Total Time Spent on Website	3.8217	0.147	26.066	0.000	3.534	4.109
A free copy of Mastering The Interview	-0.8673	0.077	-11.228	0.000	-1.019	-0.716
Lead Origin_Lead Import	-1.7064	0.520	-3.281	0.001	-2.726	-0.687
Lead Source_Google	-0.7388	0.079	-9.359	0.000	-0.894	-0.584
Lead Source_Organic Search	-0.6411	0.102	-6.295	0.000	-0.841	-0.441
Lead Source_Referral Sites	-1.0235	0.312	-3.278	0.001	-1.635	-0.412
What is your current occupation_Student	1.0978	0.205	5.364	0.000	0.697	1.499
What is your current occupation_Unemployed	1.5259	0.079	19.311	0.000	1.371	1.681
What is your current occupation_Working Professional	4.1625	0.183	22.784	0.000	3.804	4.521

	Features	VIF
1	Total Time Spent on Website	2.07
8	What is your current occupation_Unemployed	1.93
4	Lead Source_Google	1.62
2	A free copy of Mastering The Interview	1.56
5	Lead Source_Organic Search	1.22
9	What is your current occupation_Working Profes...	1.13
0	Do Not Email	1.06
6	Lead Source_Referral Sites	1.02
7	What is your current occupation_Student	1.02
3	Lead Origin_Lead Import	1.01

MODEL BUILDING & EVALUATION

An ROC curve, with 0.83 area under curve (AUC), is plotted to visualize the tradeoff between specificity (89%) and sensitivity (54%).

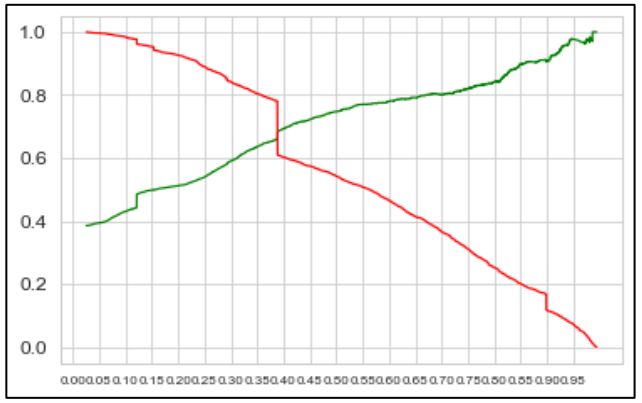




The intersection point of accuracy, sensitivity and specificity gives the probability of balanced sensitivity and specificity *(optimal cutoff probability).*

- *On setting the optimal threshold as 0.35, the sensitivity calculates to 80% and specificity to 72%.*

The precision-recall tradeoff curve is used to obtain the optimised threshold of 0.38.

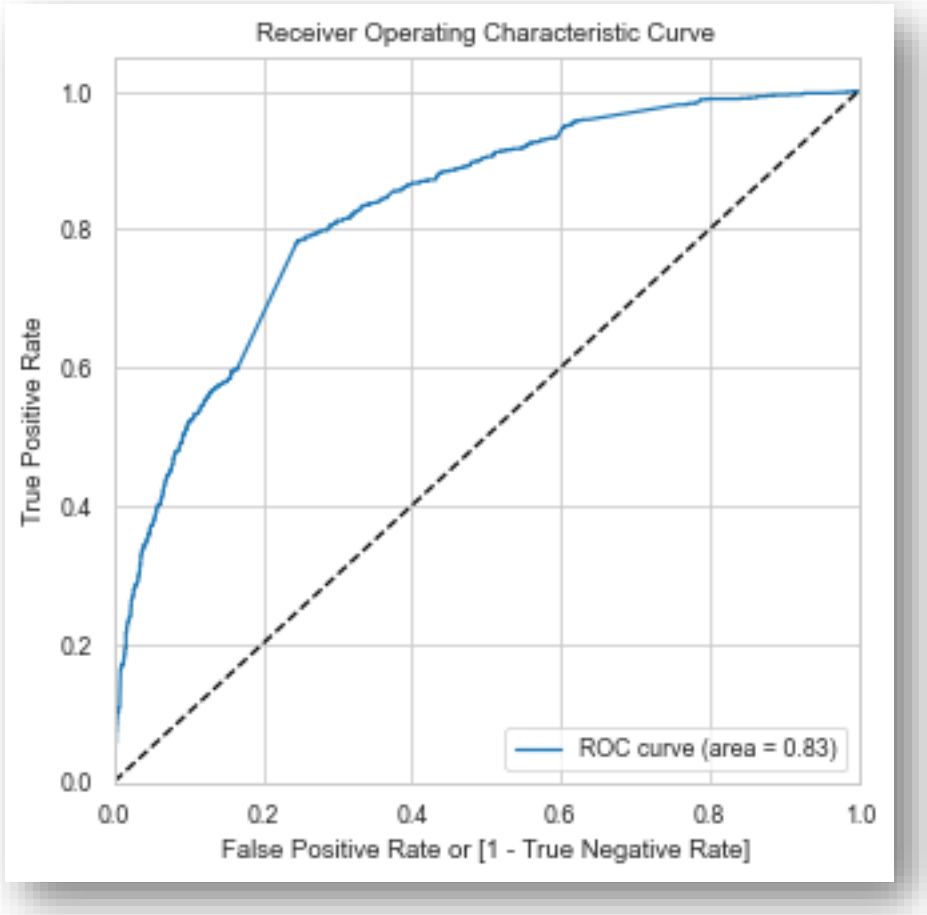


MODEL BUILDING & EVALUATION

Predictive Accuracy

- The model has predictive test ROC-AUC of 0.83.

- This explains that the model has good accuracy and stability.



MODEL BUILDING & EVALUATION

Final Train Data Metrics	Final Test Data Metrics
• Accuracy = 76.02%	• Accuracy = 76.42%
• Sensitivity = 78.58%	• Sensitivity = 78.26%
• Specificity = 74.44%	• Specificity = 75.24%
• Precision = 65.48%	• Precision = 67.02%
• Recall = 78.58%	• Recall = 78.26%

The model seems to predict the conversion rate pretty well. This can give confidence for business calls.

MODEL BUILDING & EVALUATION

Score Variable Generation

- The leads with lead score of 40 and above are 'Warm Leads' as they have a high chance of getting converted.
- The leads having lead score between 90 and 100 have a very high chance of getting converted. So, they are termed as 'Hot Leads'.

Lead Score

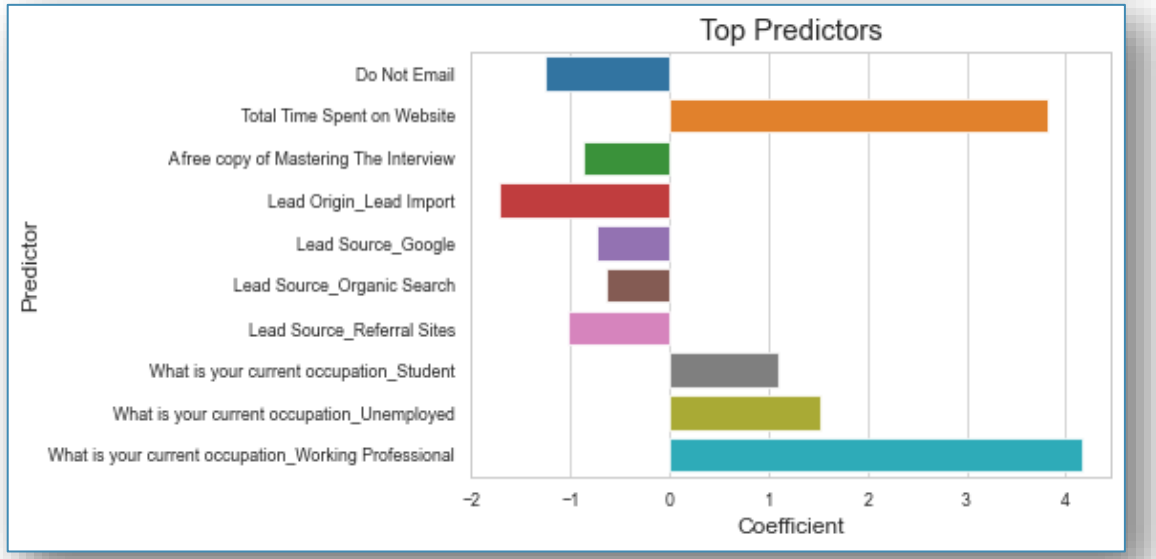
	Converted	Prospect ID	Converted_Prob	final_predicted	lead_score
1601	1	3428	0.990197	1	99
2662	1	3478	0.993432	1	99
1651	1	7167	0.990099	1	99
842	1	7420	0.990113	1	99
2632	1	5019	0.990106	1	99
...	...	...	...	...	...
1816	0	8947	0.019679	0	1
170	0	6734	0.019009	0	1
2673	0	4716	0.012542	0	1
2079	0	8586	0.015380	0	1
962	0	3107	0.009707	0	0

2727 rows × 5 columns

MODEL BUILDING & EVALUATION

Top Predictors

- The top 3 features that highly contribute towards the probability of getting a lead converted are as follows :
  - *What is your current occupation_Working Professional*
  - *Total Time Spent on Website*
  - *Lead Origin_Lead Import*



	Predictors	Coefficient
0	Do Not Email	-1.25
1	Total Time Spent on Website	3.82
2	A free copy of Mastering The Interview	-0.87
3	Lead Origin_Lead Import	-1.71
4	Lead Source_Google	-0.74
5	Lead Source_Organic Search	-0.64
6	Lead Source_Referral Sites	-1.02
7	What is your current occupation_Student	1.10
8	What is your current occupation_Unemployed	1.53
9	What is your current occupation_Working Profes...	4.16

CONCLUSION

Insights on Business Target *:*

✓ *Selecting the most promising leads*

- The following are the most promising leads according to our model –
  - Working professionals
  - Leads who spend more time on the website and therefore the website should be upgraded with exciting offers and interesting topics to attract them

✓ *Building a model and deploying it for future usage*

- Our model has a good predictive accuracy (ROC-AUC = 0.83) and so it can be deployed in future business decision making

✓ *Assigning a lead score to each of the leads and identifying 'Hot Leads'*

- Leads with lead score above 90 ('Hot Leads') have a very high chance of getting converted