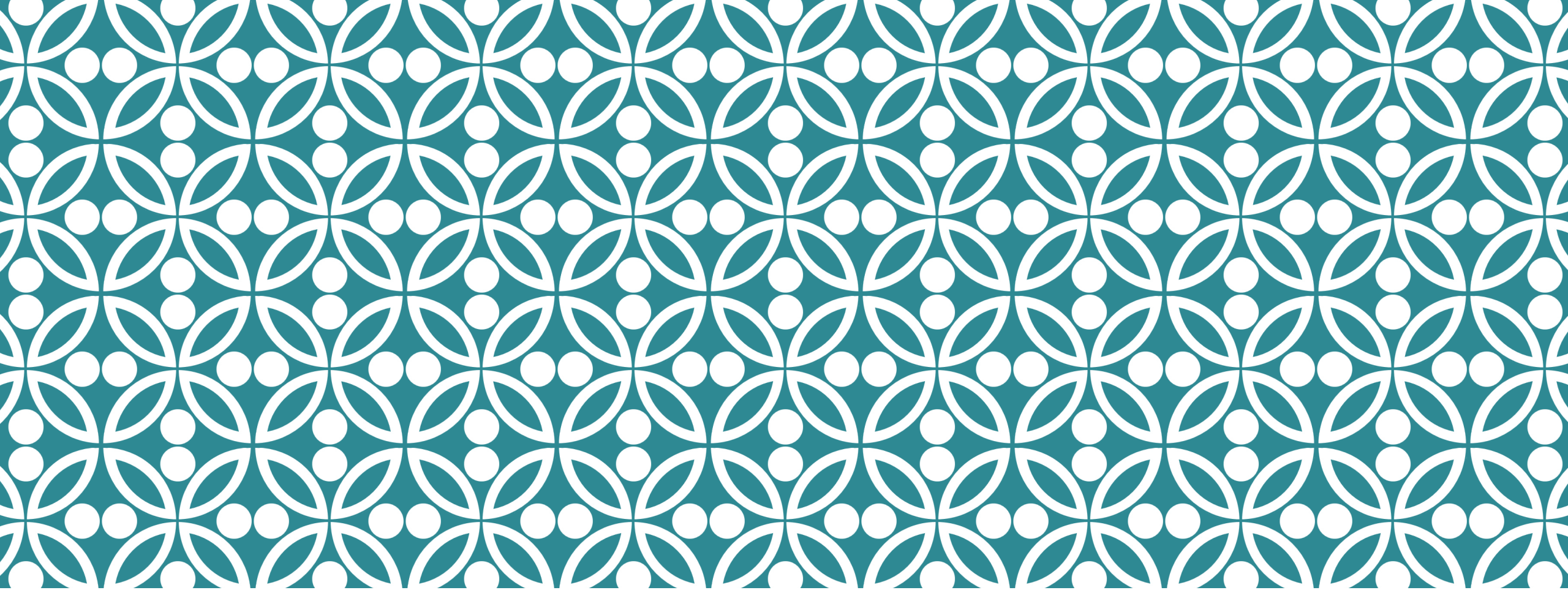


LEAD SCORING CASE STUDY

Analysed By:
Ratik Khanna
Subhanjan Roy

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%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you 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%.

OVERALL APPROACH

- Reading and Understanding Data
- Checking The Missing Value
- Exploratory Data Analysis(EDA)
- Checking Outlier
- Outlier Treatment
- Dummy Variable
- Train Test Split
- Model Building
- Model Evaluation
- ROC
- Optimal Cut-off Point
- Assign Lead Score to Train
- Prediction on Test set
- Assigning Lead Score on Test Data
- Conclusion

READING AND UNDERSTANDING DATA
CHECKING THE MISSING VALUE
DATA CLEANING

Reading and Understanding Data

- **Checking the shape, Information and Parameters of the Data frame.**
- **Data Preparation for Analysis**

Checking The Missing Value

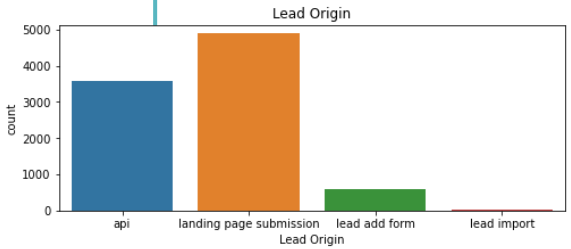
- **Checking the percentage of missing values**
- **Replacing 'Select' with NaN (Since it means no option is selected)**

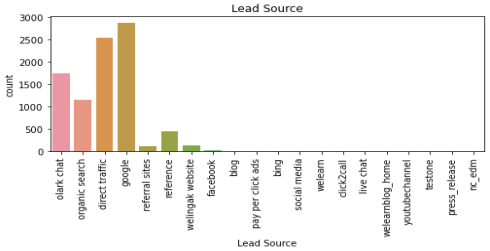
Data Cleaning

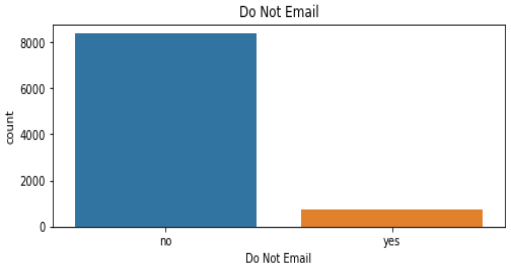
- **Removing all the columns that are not required and have 30% null values**
- **Checking if there are columns with one unique value since it won't affect our analysis**
- **Dropping unique value columns**

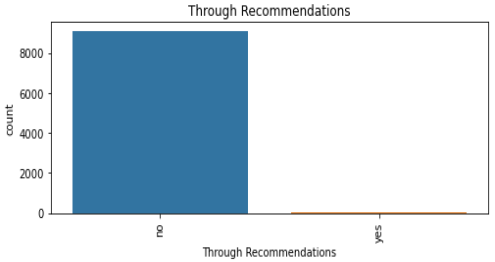
EXPLORATORY DATA ANALYSIS(EDA)

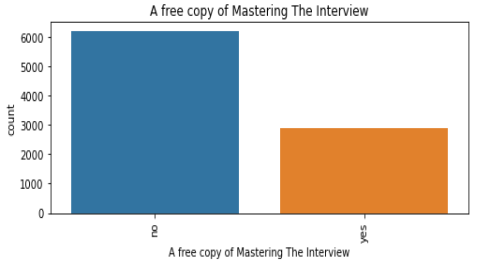
Univariate Analysis(Removing Features with 95% Constant Value)

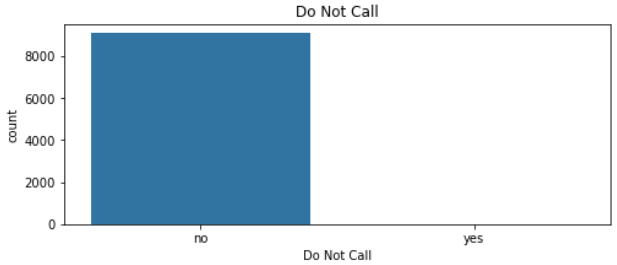


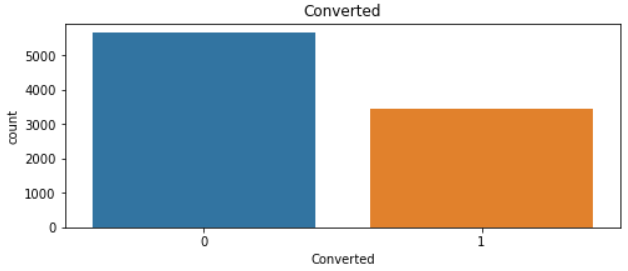


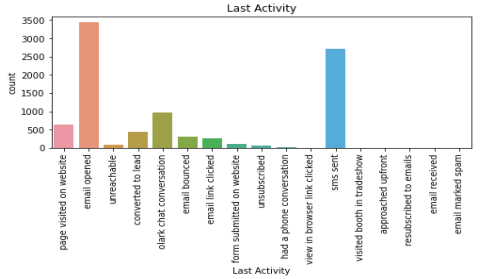


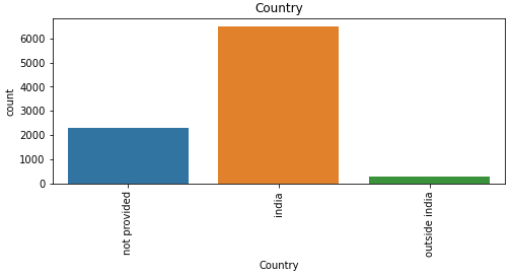


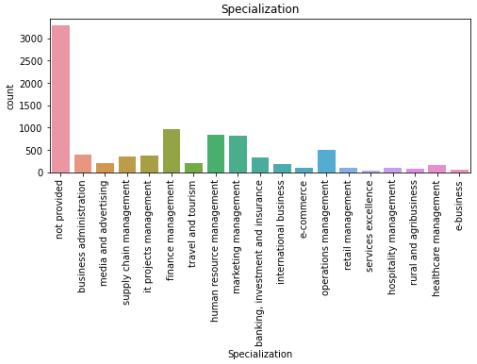


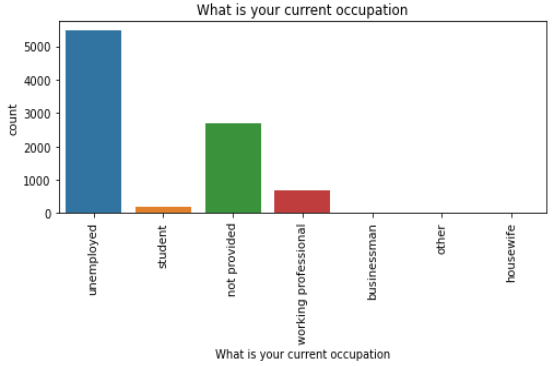


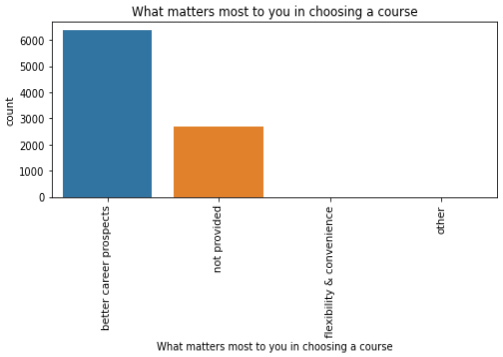


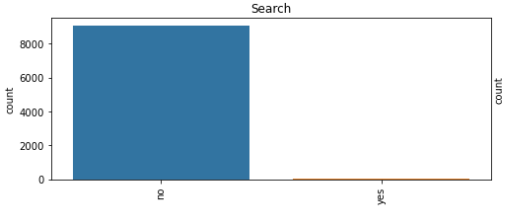


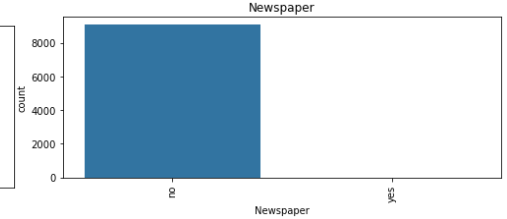


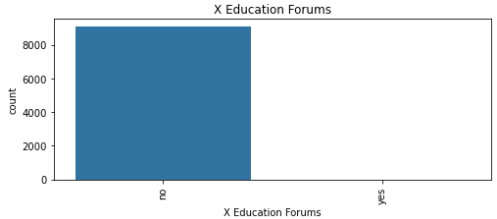


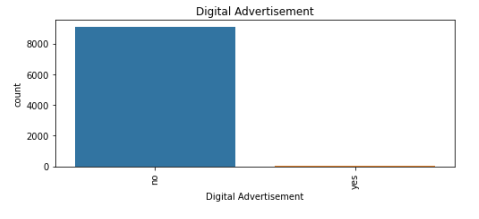




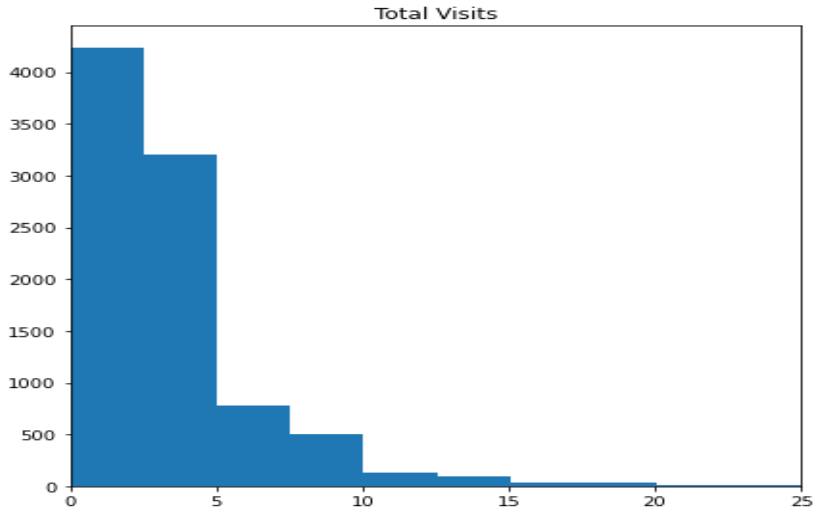


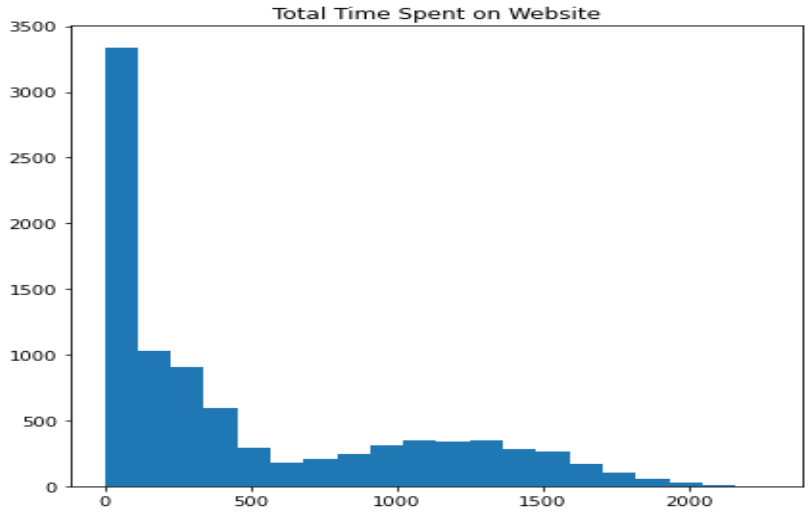


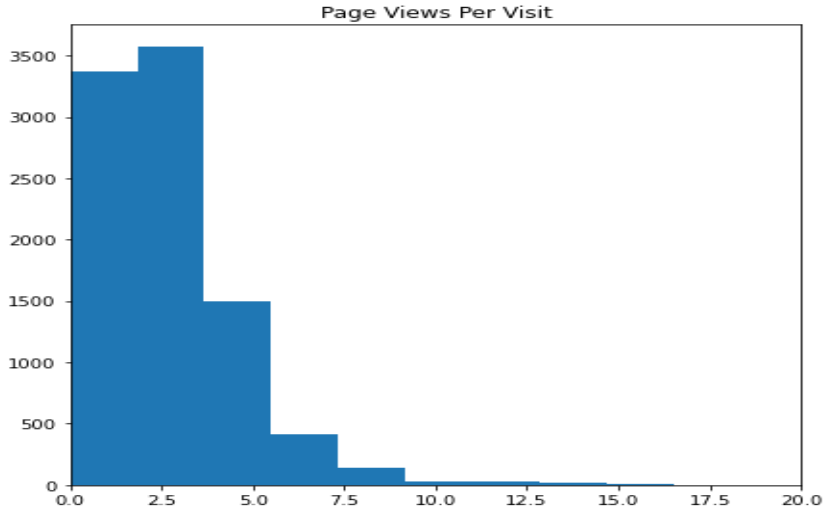




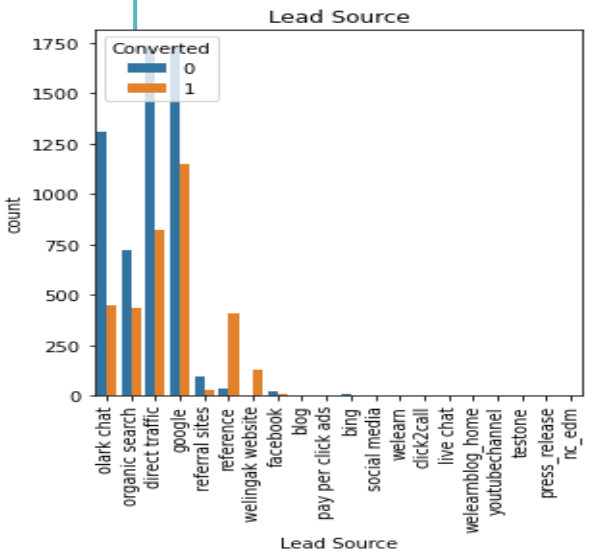
VISUALIZING NUMERICAL VARIABLES

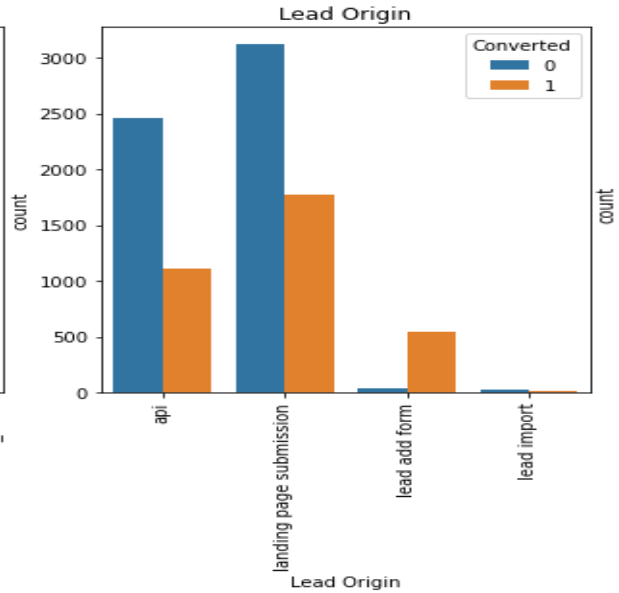


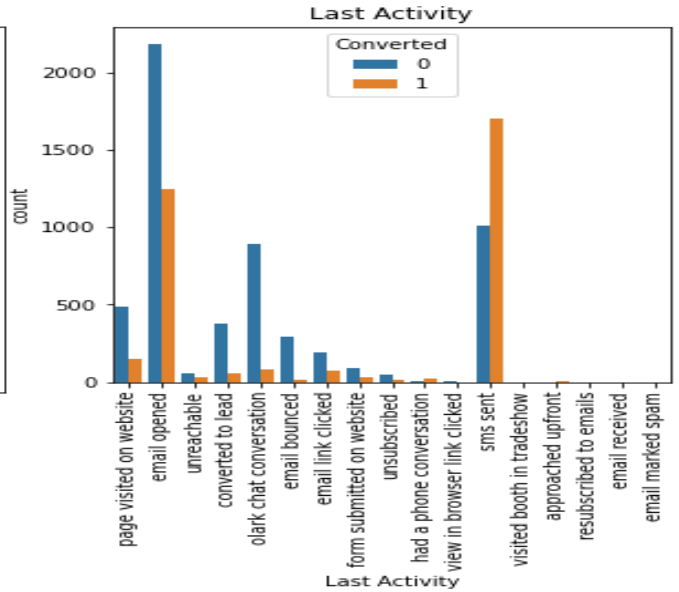


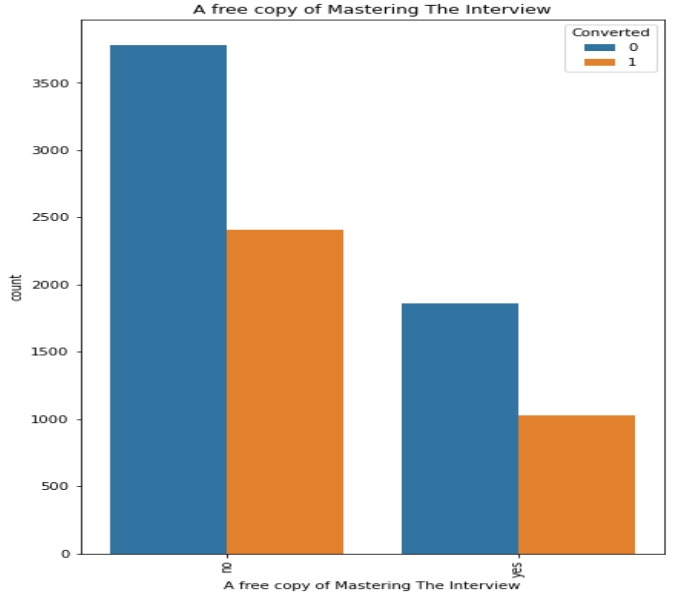


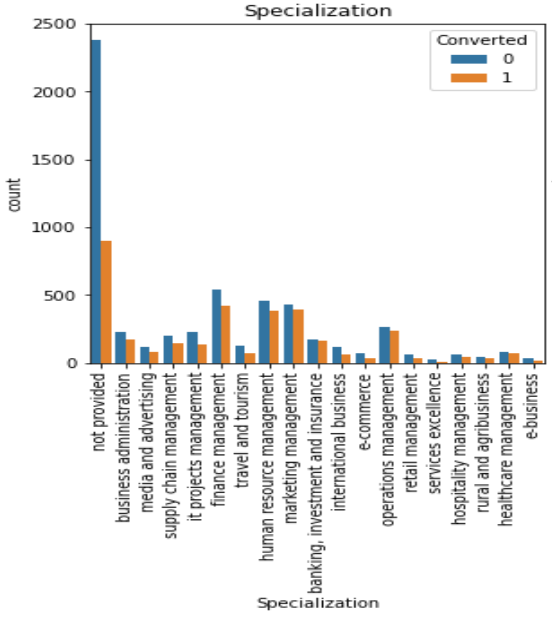
VISUALIZING CATEGORICAL VARIABLES

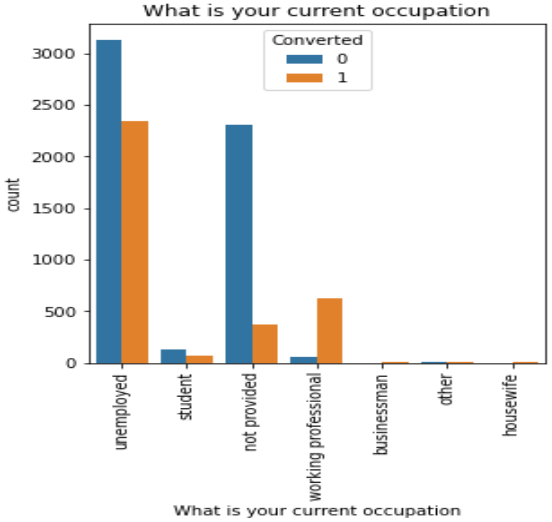


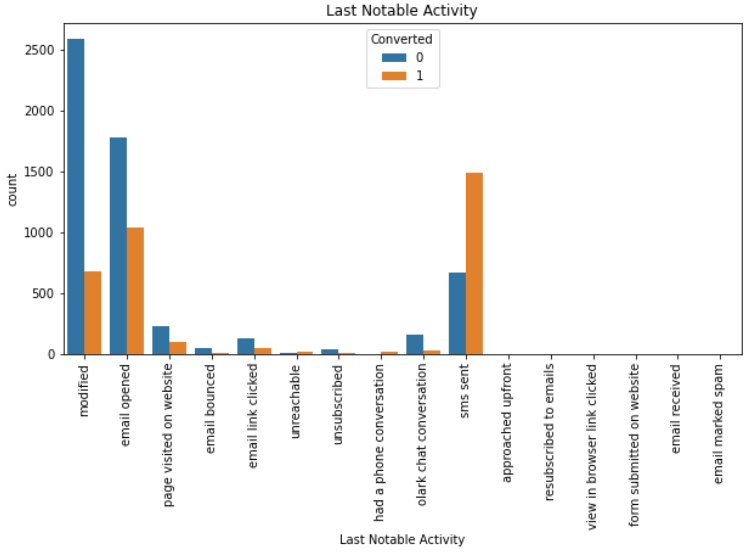






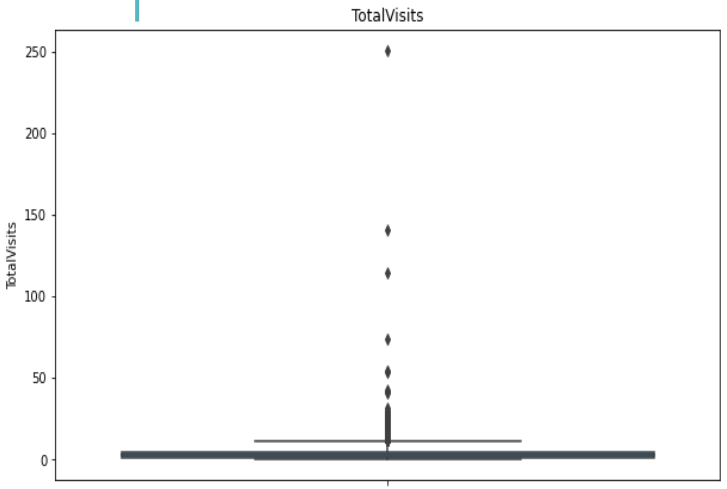


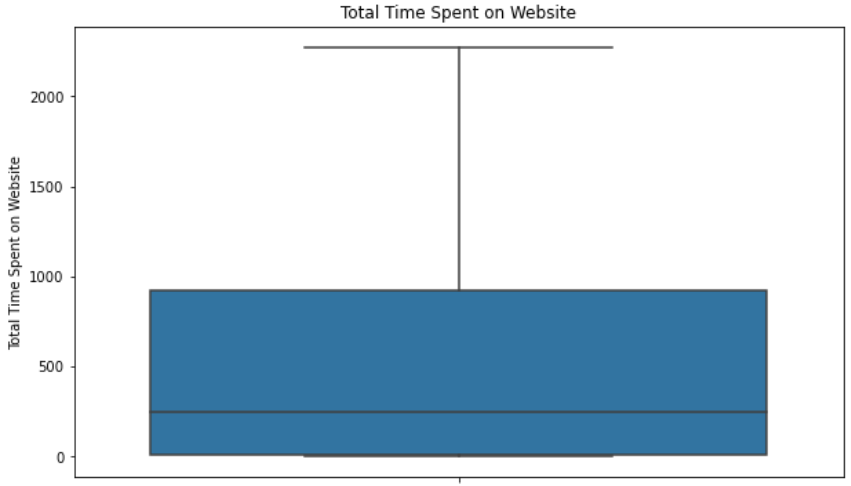


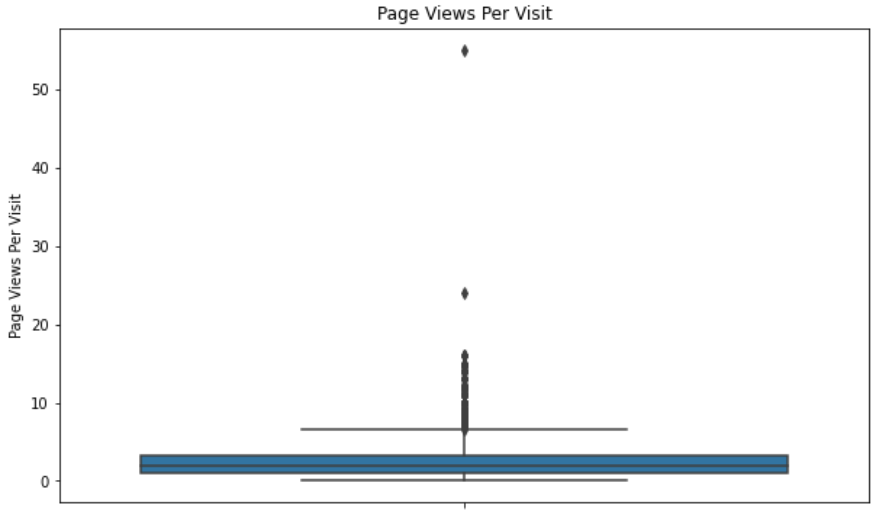


CHECKING OUTLIER

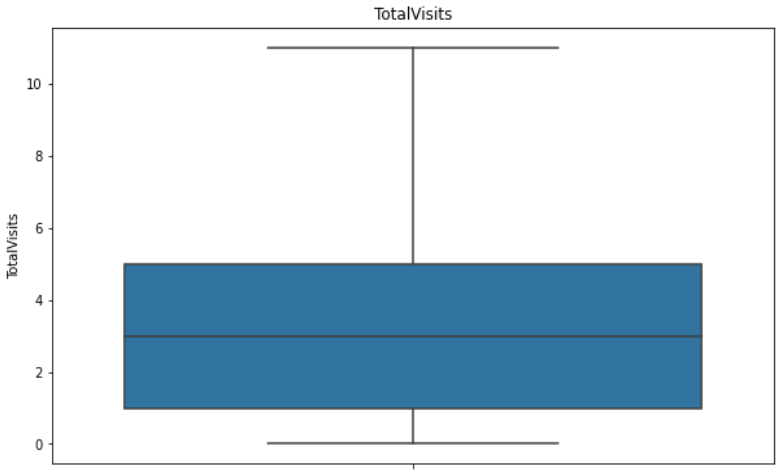
Outliers Analysis

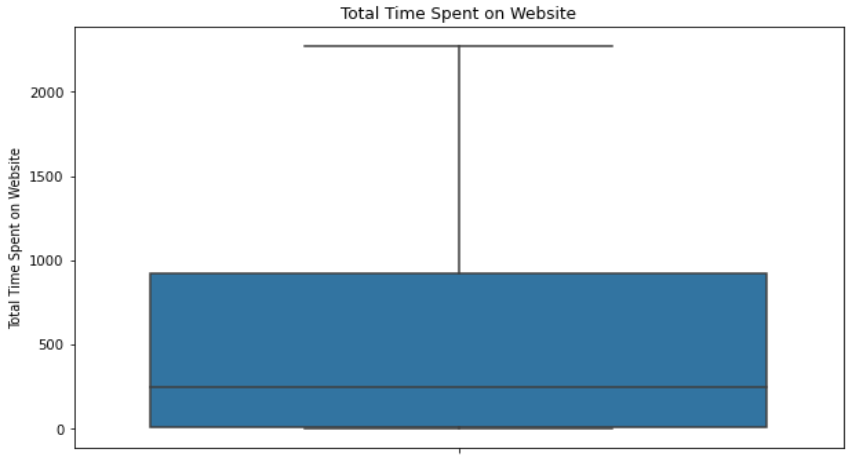


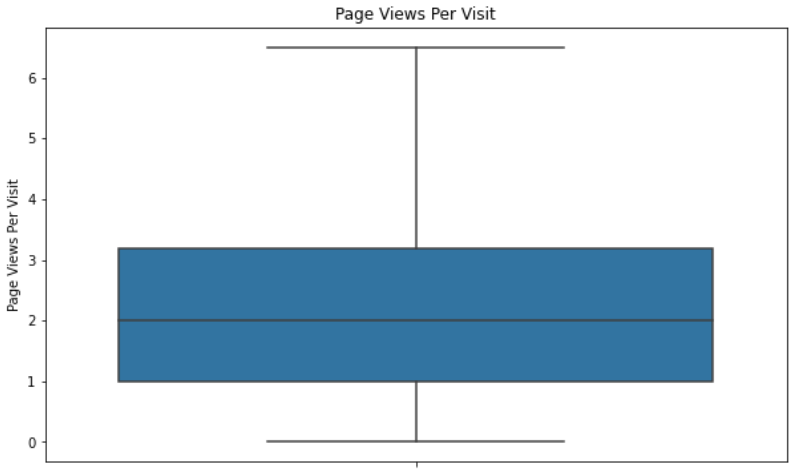




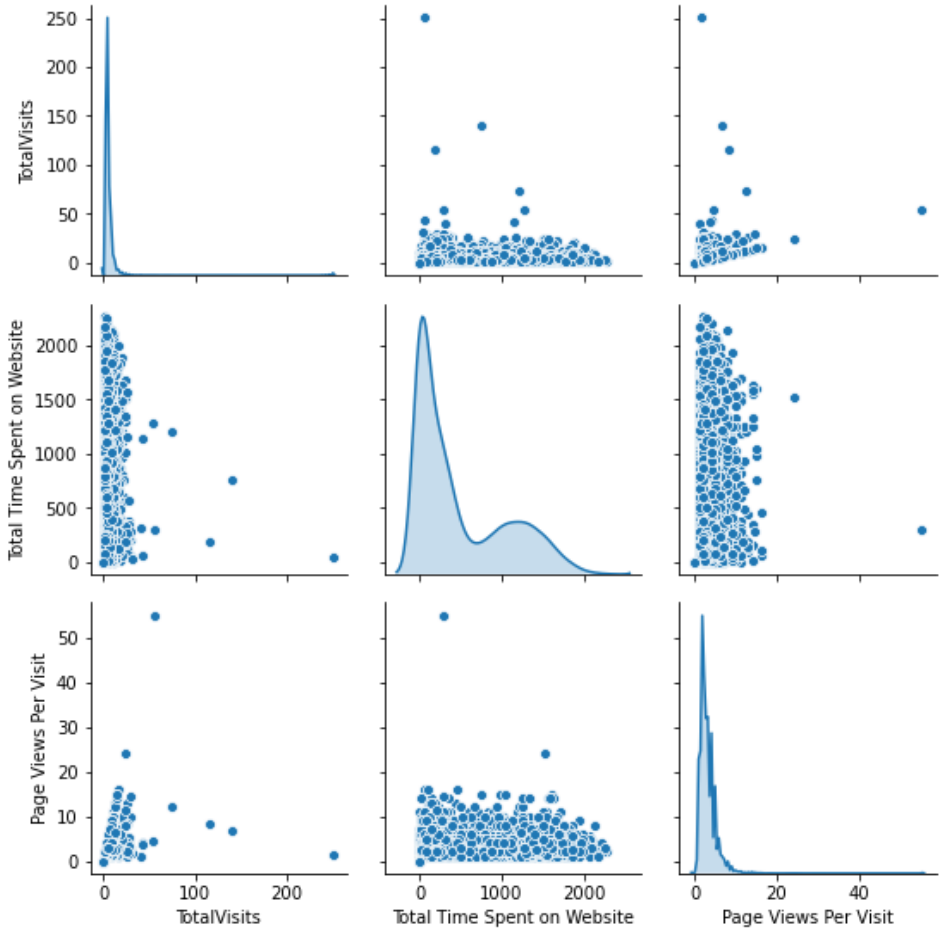
Outlier Treatment

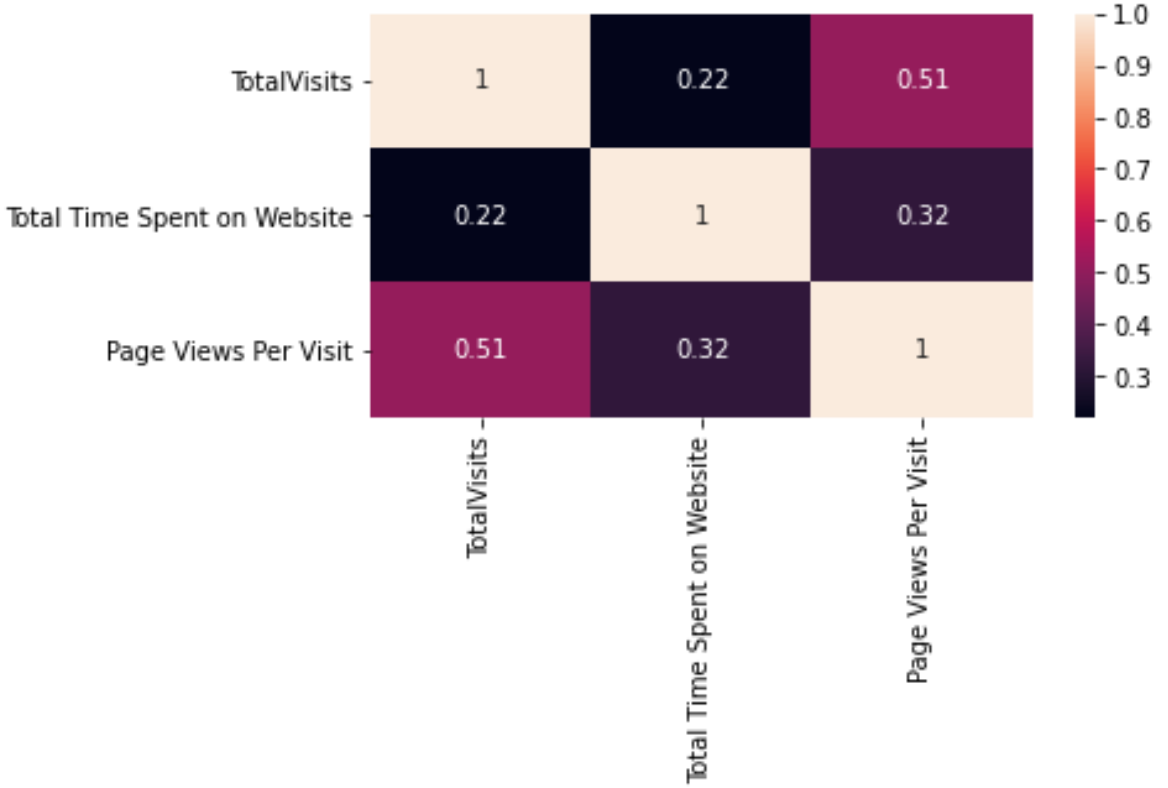






PAIRPLOT AND CORRELATION





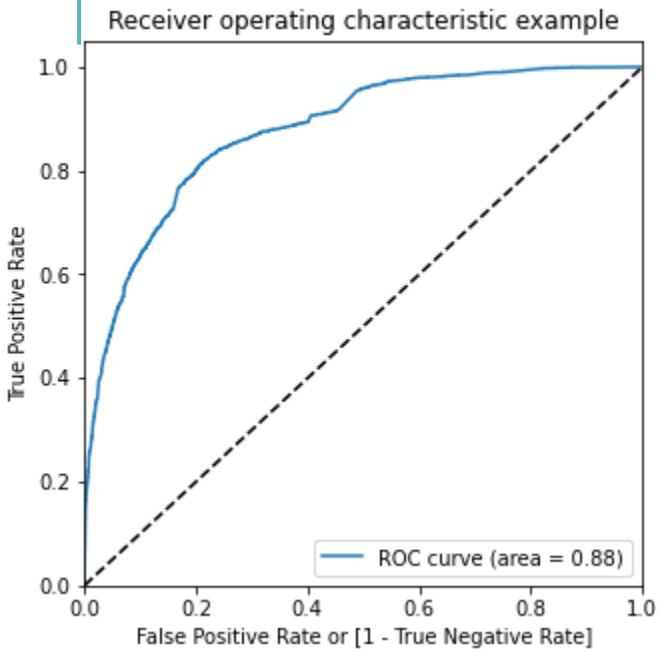
MODEL BUILDING

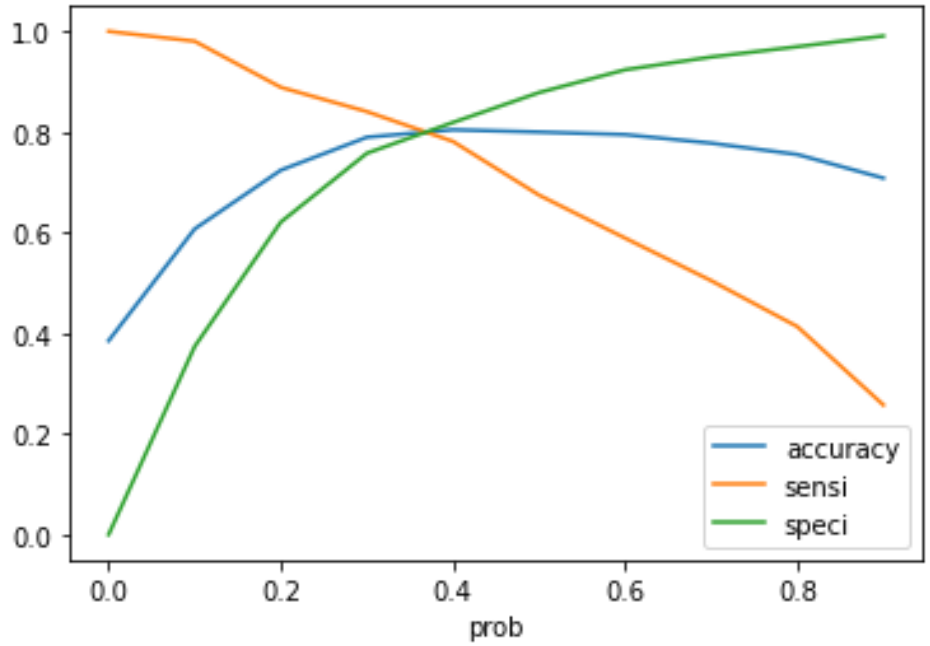
Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6337
Model Family:	Binomial	**Df Model:**	13
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2674.2
Date:	Mon, 11 Jan 2021	**Deviance:**	5348.5
Time:	18:25:40	**Pearson chi2:**	6.12e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.7493	0.094	-29.310	0.000	-2.933	-2.565
Total Time Spent on Website	3.8357	0.145	26.497	0.000	3.552	4.119
Lead Origin_lead add form	3.1143	0.219	14.251	0.000	2.686	3.543
Lead Source_direct traffic	-0.5780	0.077	-7.468	0.000	-0.730	-0.426
Lead Source_welingak website	1.9092	0.751	2.543	0.011	0.438	3.381
Do Not Email_yes	-1.5819	0.170	-9.325	0.000	-1.914	-1.249
Last Activity_had a phone conversation	2.3772	0.729	3.262	0.001	0.949	3.806
Last Activity_olark chat conversation	-0.9436	0.162	-5.821	0.000	-1.261	-0.626
Last Activity_sms sent	1.2938	0.073	17.638	0.000	1.150	1.438
What is your current occupation_other	1.9781	0.713	2.773	0.006	0.580	3.376
What is your current occupation_student	1.4172	0.229	6.192	0.000	0.969	1.866
What is your current occupation_unemployed	1.1723	0.086	13.642	0.000	1.004	1.341
What is your current occupation_working professional	3.5992	0.195	18.412	0.000	3.216	3.982
Last Notable Activity_unreachable	1.9155	0.495	3.873	0.000	0.946	2.885

- Splitting The Data into Train and Test Set
- The Ratio for Train-Test Split is 70:30
- Use RFE for Feature Selection and run them
- Building Mode by removing all the variable where P Value is high and VIF is high

MODEL EVALUATION, ROC
OPTIMAL CUT-OFF POINT FOR TRAIN DATASET





Outcome:

- In ROC Curve, The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
- In ROC Curve, The curve seems to be good for our model.
- Optimal Cut-off Point, 0.34 seems to be the optimum point to take it as a cutoff probability.

- Accuracy of 80%
- Sensitivity of 80%
- Specificity of 79%

All our good values (approx. 80%) and consistent with each other suggests that the model is good.

LEAD SCORE TO TRAIN DATASET- THE FOLLOWING ARE THE TOP LEADS HAVING HIGH LEAD SCORE- POTENTIAL LEAD LIST

	Converted	Conversion_Prob	Predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	final_predicted	Lead Score
4773	1.0	0.995447	1	1	1	1	1	1	1	1	1	1	1	1	100
5871	1.0	0.998822	1	1	1	1	1	1	1	1	1	1	1	1	100
5043	1.0	0.995260	1	1	1	1	1	1	1	1	1	1	1	1	100
2790	1.0	0.996201	1	1	1	1	1	1	1	1	1	1	1	1	100
3687	1.0	0.996997	1	1	1	1	1	1	1	1	1	1	1	1	100
5971	1.0	0.998732	1	1	1	1	1	1	1	1	1	1	1	1	100
2980	1.0	0.995044	1	1	1	1	1	1	1	1	1	1	1	1	100
4242	1.0	0.996243	1	1	1	1	1	1	1	1	1	1	1	1	100
1543	1.0	0.995878	1	1	1	1	1	1	1	1	1	1	1	1	100
1335	1.0	0.995186	1	1	1	1	1	1	1	1	1	1	1	1	100
5664	1.0	0.995767	1	1	1	1	1	1	1	1	1	1	1	1	100
2463	1.0	0.996090	1	1	1	1	1	1	1	1	1	1	1	1	100
1568	1.0	0.988891	1	1	1	1	1	1	1	1	1	1	1	1	99
4077	1.0	0.991339	1	1	1	1	1	1	1	1	1	1	1	1	99
4062	1.0	0.991339	1	1	1	1	1	1	1	1	1	1	1	1	99
2693	1.0	0.990901	1	1	1	1	1	1	1	1	1	1	1	1	99
1561	1.0	0.991339	1	1	1	1	1	1	1	1	1	1	1	1	99
1985	1.0	0.994821	1	1	1	1	1	1	1	1	1	1	1	1	99
1506	1.0	0.994821	1	1	1	1	1	1	1	1	1	1	1	1	99
1637	1.0	0.994821	1	1	1	1	1	1	1	1	1	1	1	1	99

PREDICTION ON TEST SET

After Successful Prediction model:

- Accuracy of 81%
- Sensitivity of 80%
- Specificity of 80%

All our good values (approx. 80%) and consistent with the trained model. Suggesting the model was well trained

LEAD SCORE TO TEST DATASET- THE FOLLOWING ARE THE TOP LEADS HAVING HIGH LEAD SCORE- POTENTIAL LEAD LIST

	Lead Number	Converted	Converted_Prob	predicted	Lead Score
33	4062	1.0	0.996108	1	100
1617	8521	1.0	0.995339	1	100
1786	2011	1.0	0.995102	1	100
1422	4613	1.0	0.997363	1	100
505	8213	1.0	0.995526	1	100
868	2495	1.0	0.996981	1	100
2491	6944	1.0	0.985614	1	99
2310	3723	1.0	0.994821	1	99
1526	3248	1.0	0.991363	1	99
1055	2674	1.0	0.994821	1	99
678	7036	1.0	0.987903	1	99
1582	2158	1.0	0.993992	1	99
1239	5808	1.0	0.994821	1	99
1585	3355	1.0	0.991339	1	99
619	6784	1.0	0.991823	1	99
2686	8052	1.0	0.994821	1	99
2510	2673	1.0	0.994821	1	99
2117	2354	1.0	0.991339	1	99
70	3542	1.0	0.991339	1	99
738	6999	1.0	0.991339	1	99

TOP 5 VARIABLES WHICH CONTRIBUTE MOST TOWARDS THE PROBABILITY OF A LEAD GETTING CONVERTED

Features	Coeff
Total Time Spent on Website	3.8357
What is your current occupation_working profes...	3.5992
Lead Origin_lead add form	3.1143
Last Activity_had a phone conversation	2.3772
What is your current occupation_other	1.9781

CONCLUSION

From our Analysis, we come to a conclusion that the following factors are very important for X-Education for Lead conversion:

- Time Spent on Website
- Lead Origin
- Lead Source
- Last Activity
- What is your current Occupation
- Last Notable activity

And we have also generated the lead score for all the leads we have in the dataset. So our focus should be more on those leads where the lead score are high. Where the conversion will be higher.