Lead Scoring Case Study

Group Members

- Gloriya Thomas
- Balasubramanian Venkatesan

Problem Statement and Business Objective

X Education sells online courses to industry professionals. They want to build a Model which identifies hot leads.

The task is to create an efficient classification model for X Education to identify the most potential leads, also known as 'Hot Leads' so that sales team can target only the potential leads rather than making calls to random individuals.

Create a model in such a way that the customers with high lead score have higher conversion chance and low lead score have lower conversion chance. The ballpark of the target lead conversion rate is around 80%. Also, the model should be able to adjust if the company's requirement changes in near future.

Solution Methodology

Data cleaning and manipulation

- handling of duplicate and unique data values.
- Missing value Treatment
- Outlier treatment

EDA

- Univariate, bivariate and multi variate analysis of both numeric and categorical variables
- Correlation Matrix and heatmap analysis

Data Preparation

- Dummy Variables Creation
- Feature Scaling (MinMaxScaler)

Train Test Split

Model building and prediction

- Classification : logistic regression - GLM

Model Validation

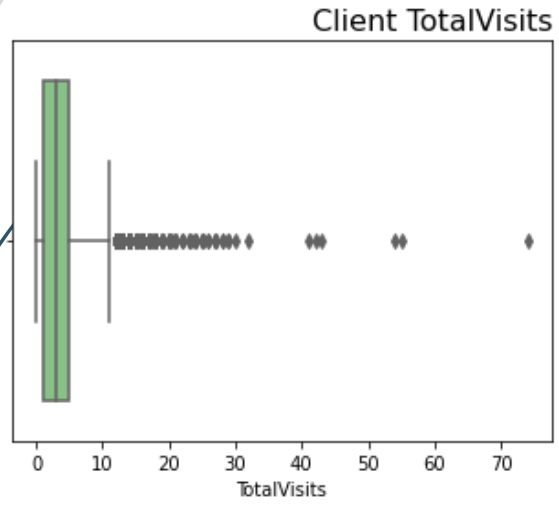
- Accuracy, Sensitivity, Specificity
- ROC and AOC
- Precision and Recall

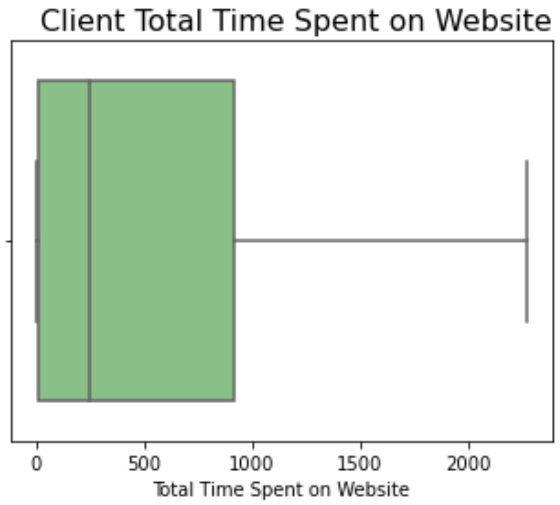
Conclusion and Recommendations

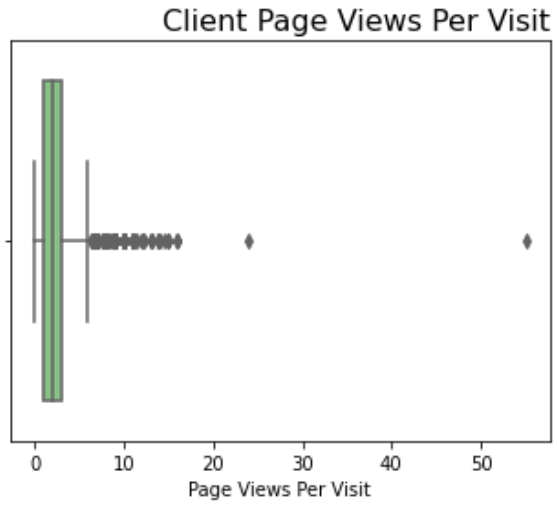
Exploratory Data Analysis

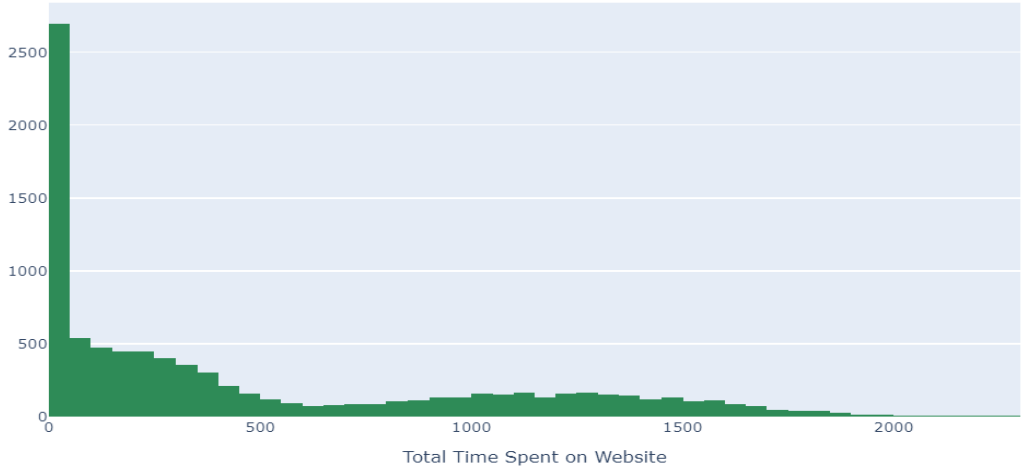
Univariate Analysis

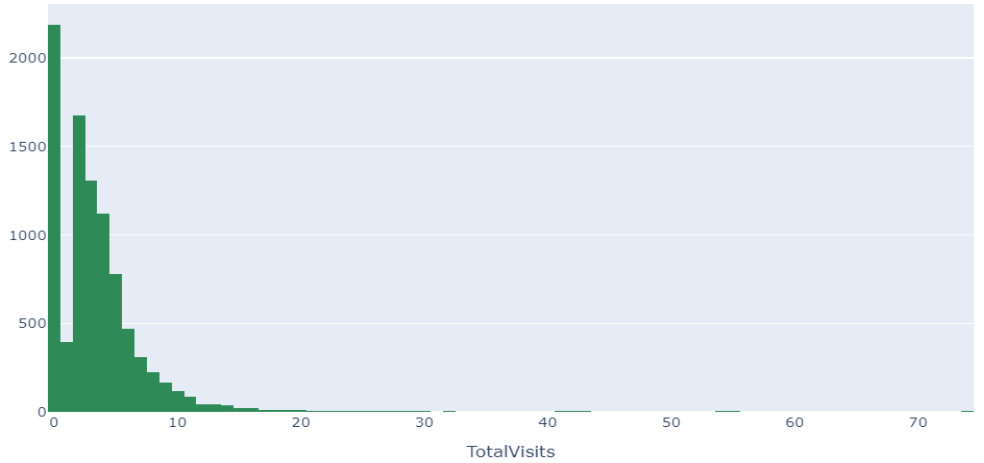
Univariate analysis of numeric variables reveals outlier issues and we have treated the outliers accordingly with appropriate capping.

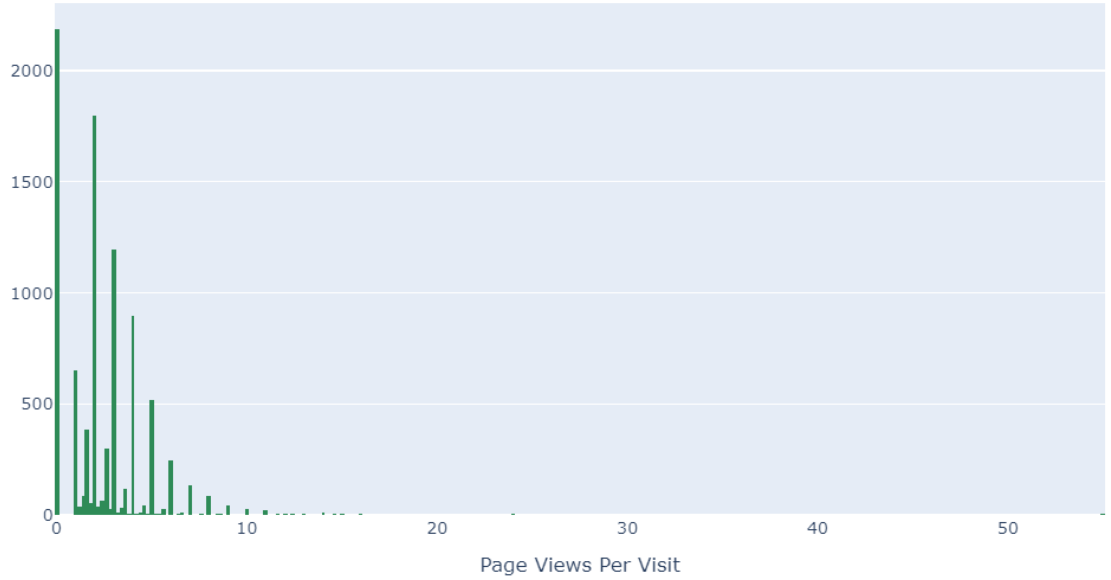








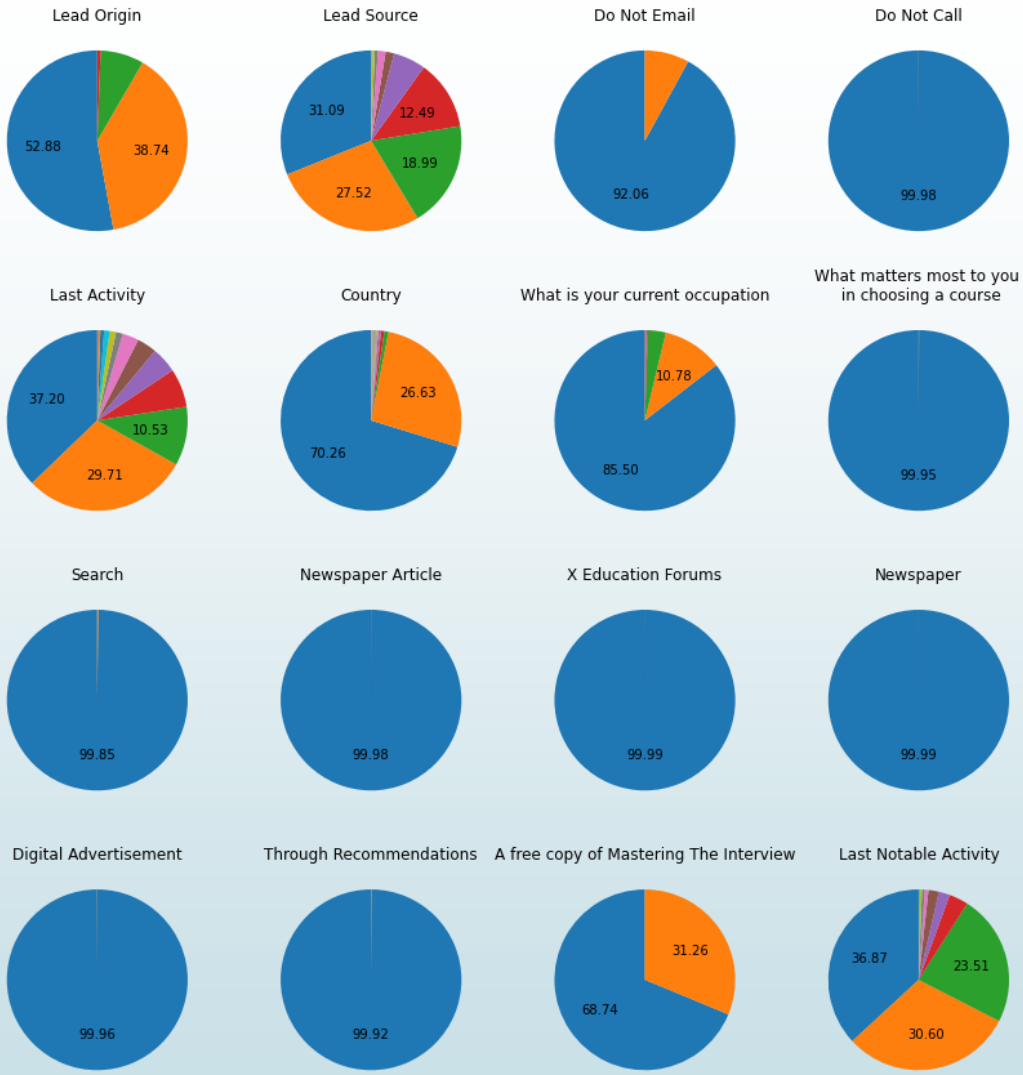


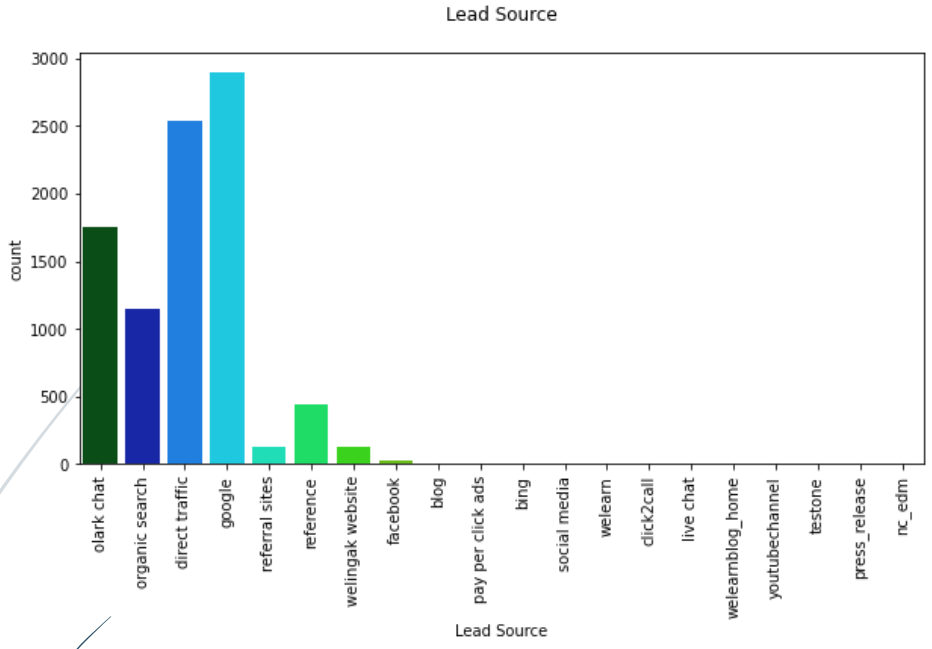


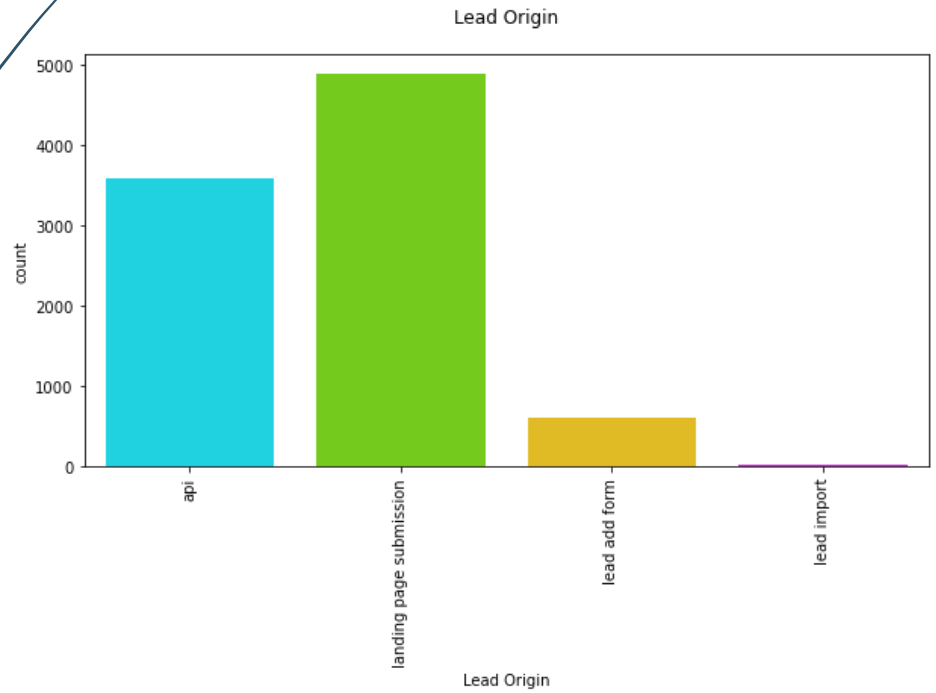
Univariate Analysis - Numeric Variables

Univariate Analysis - Categorical Variables

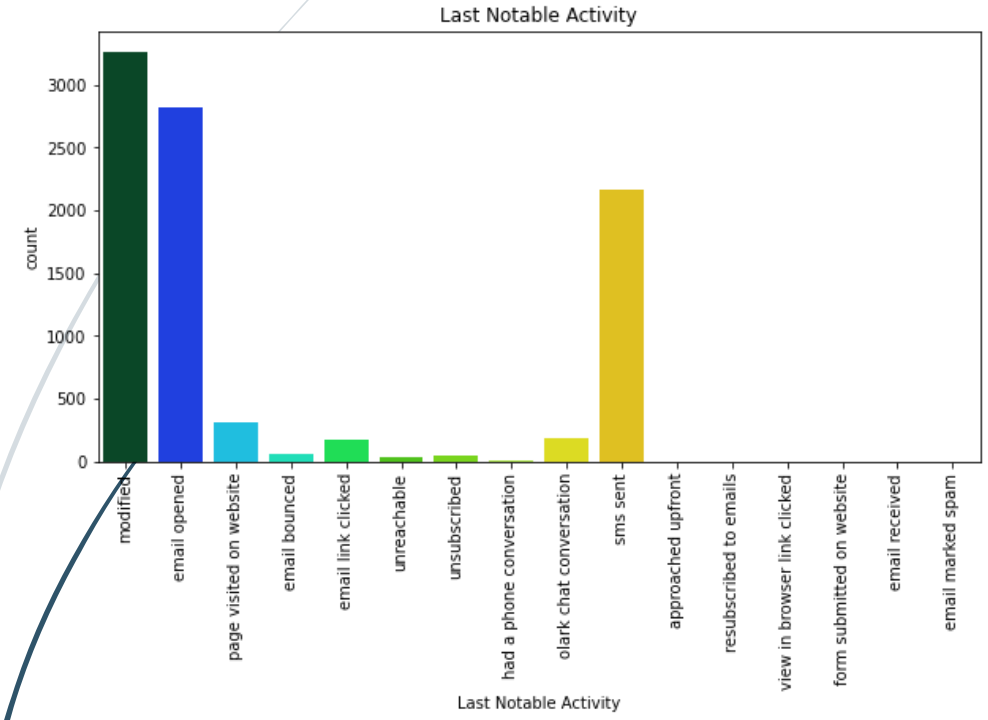
There are several features with zero or minimum variability and drop such features. There are several features with zero or minimum variability and drop such features.

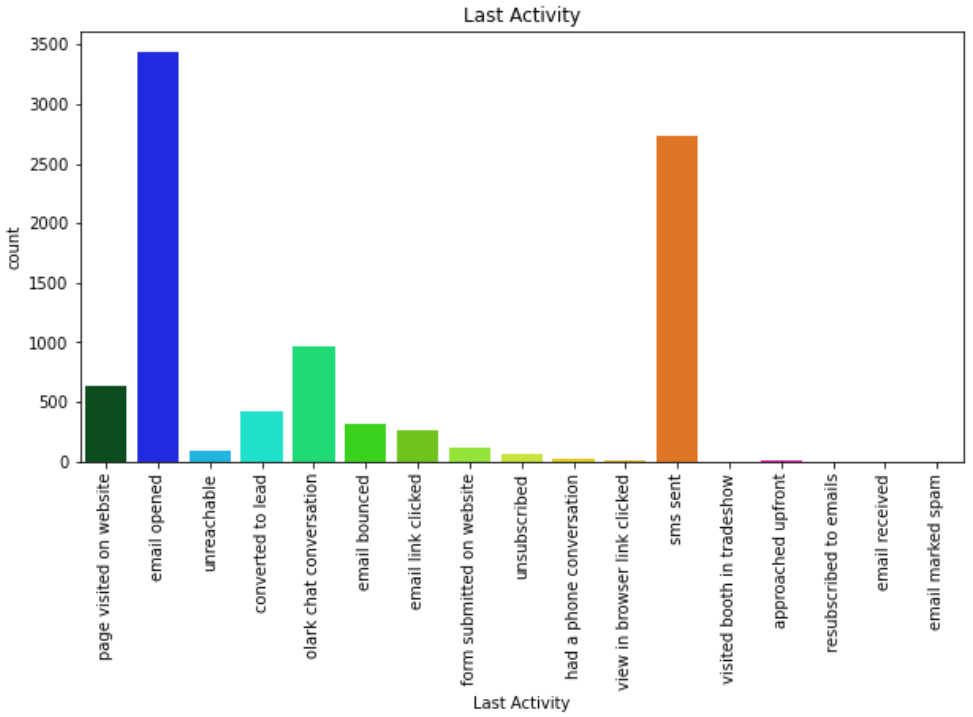


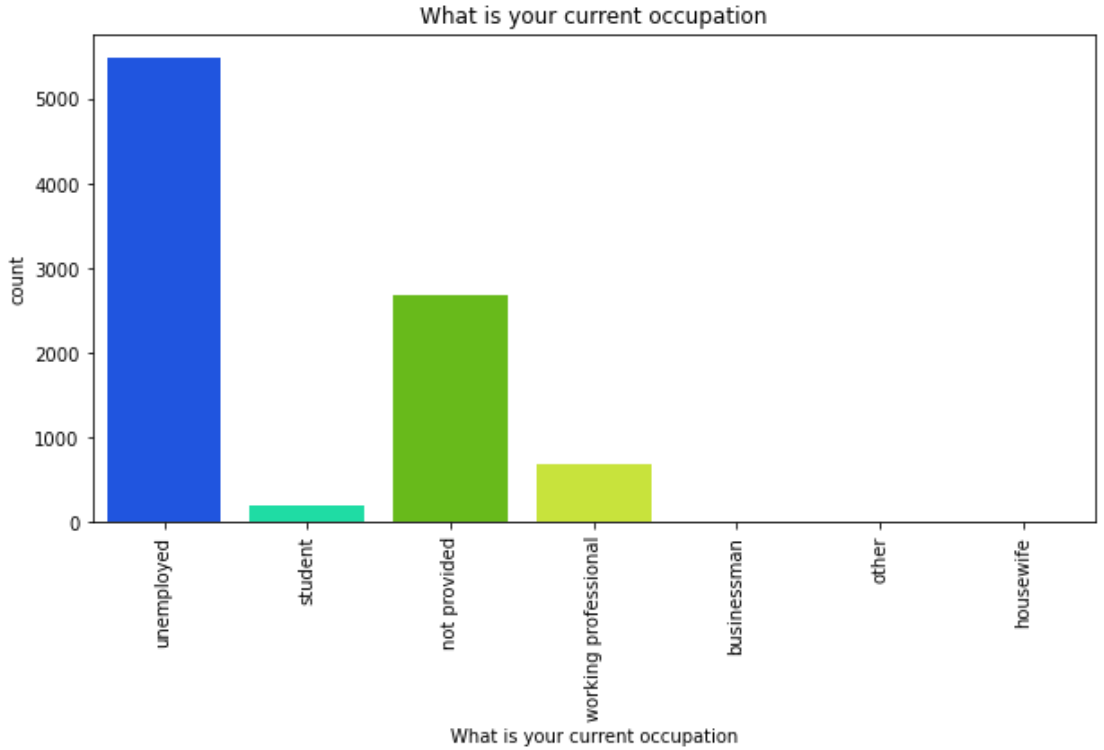


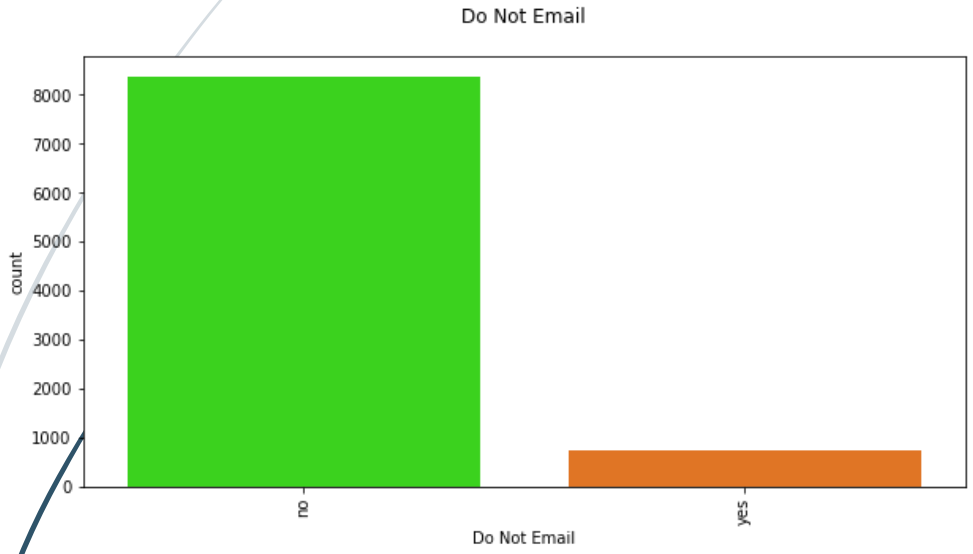


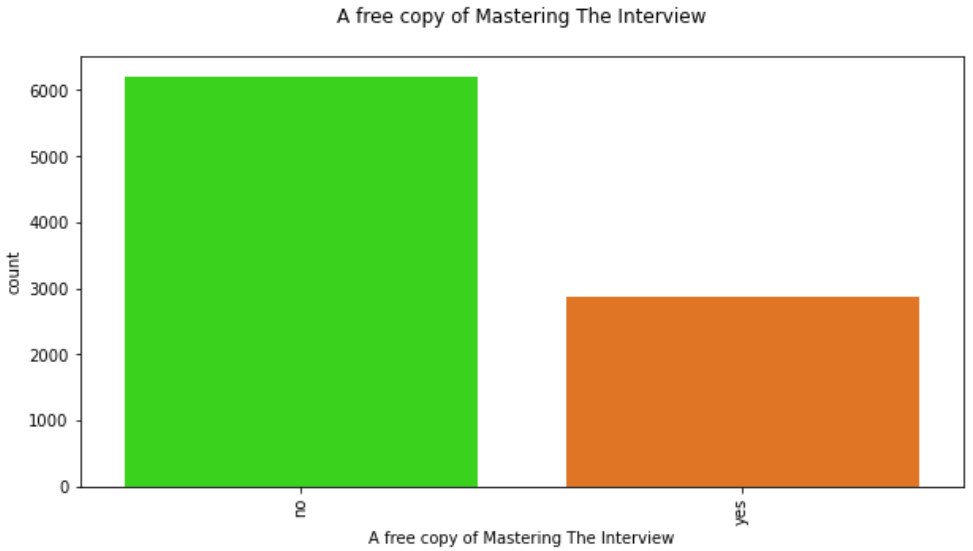
Univariate Analysis Categorical Variables


Last Notable Activity
count
3000
2500
2000
1500
1000
500
0
modified
email opened
page visited on website
email bounced
email link clicked
unreachable
unsubscribed
had a phone conversation
olark chat conversation
sms sent
approached upfront
resubscribed to emails
view in browser link clicked
form submitted on website
email received
email marked spam
Last Notable Activity

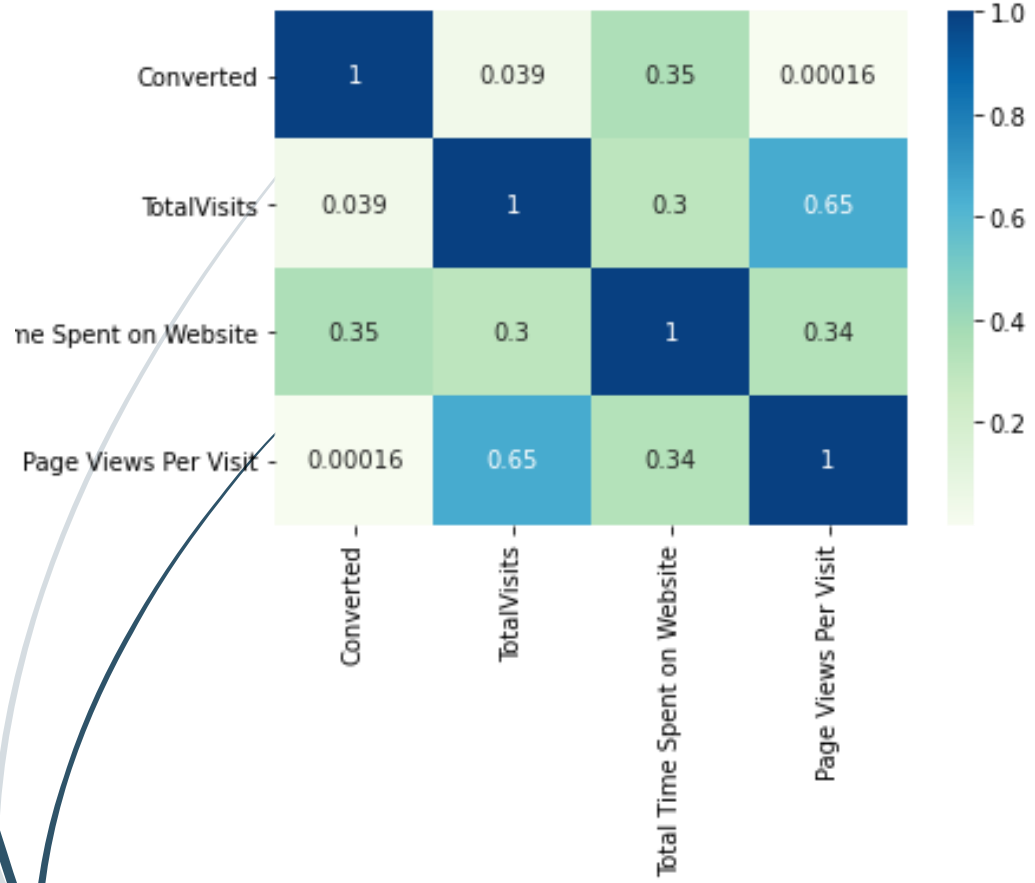

Last Activity
count
3500
3000
2500
2000
1500
1000
500
0
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
sms sent
visited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam
Last Activity


What is your current occupation
count
5000
4000
3000
2000
1000
0
unemployed
student
not provided
working professional
businessman
other
housewife
What is your current occupation


Do Not Email
8000
7000
6000
5000
4000
3000
2000
1000
0
count
no
yes
Do Not Email


A free copy of Mastering The Interview
6000
5000
4000
3000
2000
1000
0
count
no
yes
A free copy of Mastering The Interview

Bivariate Analysis



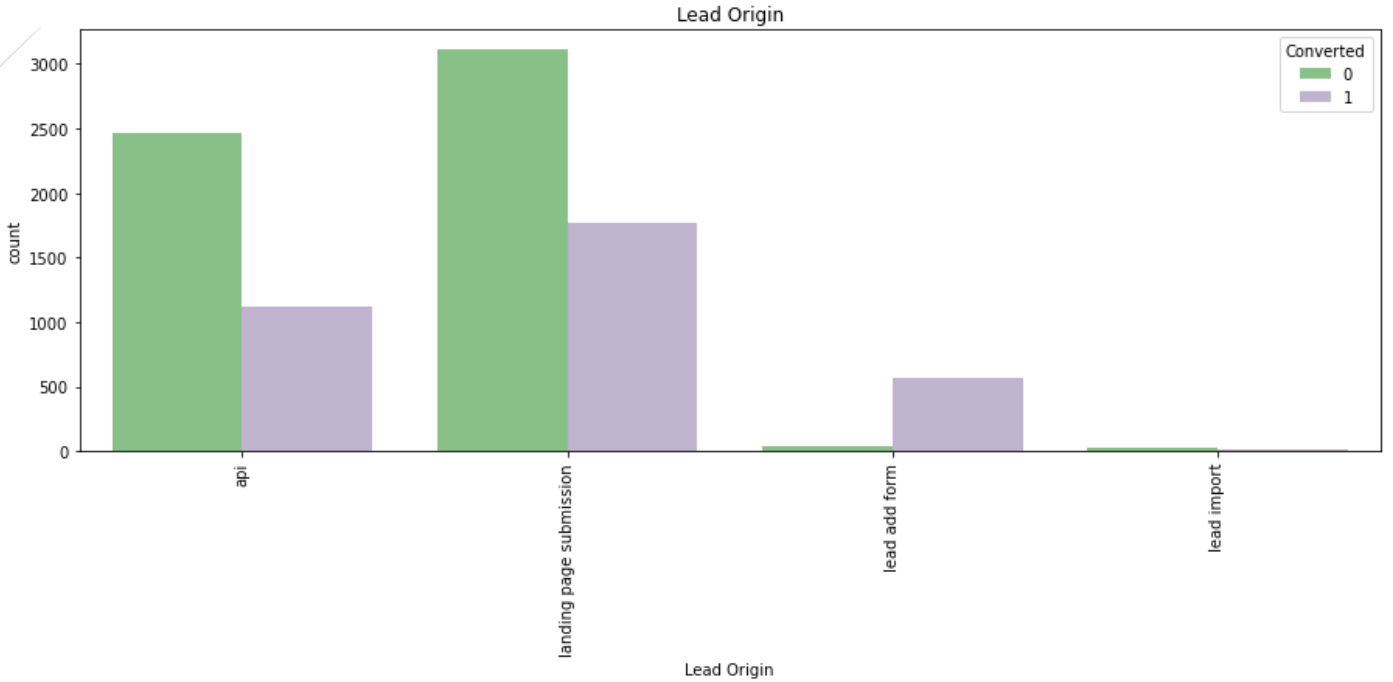
Correlation Heat Map

1. 'TotalVisits' and 'Page Views Per Visit' have a correlation of .65, indicates possible multicollinearity.

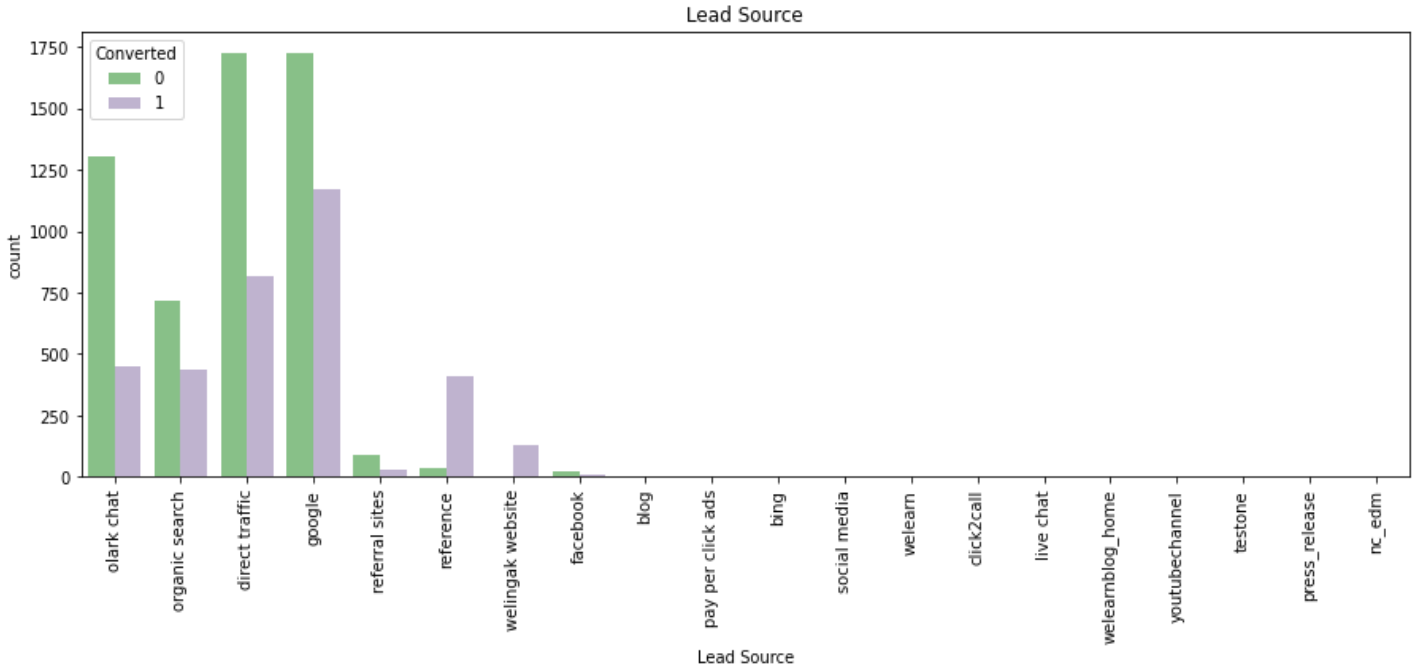
2. 'Total Time Spent on Website' and 'Converted' have a correlation of 0.35, indicates this variable could a possible predictor of successful leads.

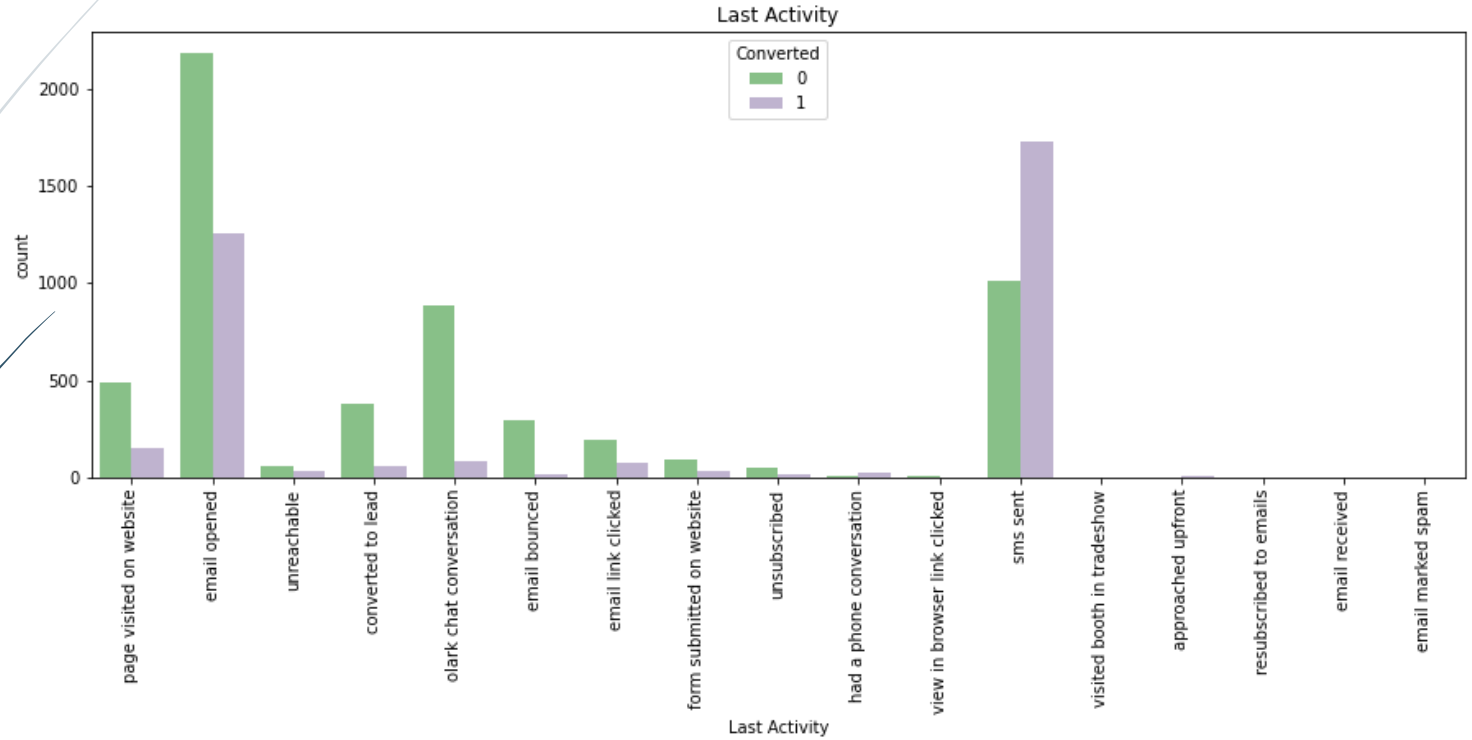
3. 'Total Time Spent on Website' and 'Page Views Per Visit' have a correlation coefficient of .34

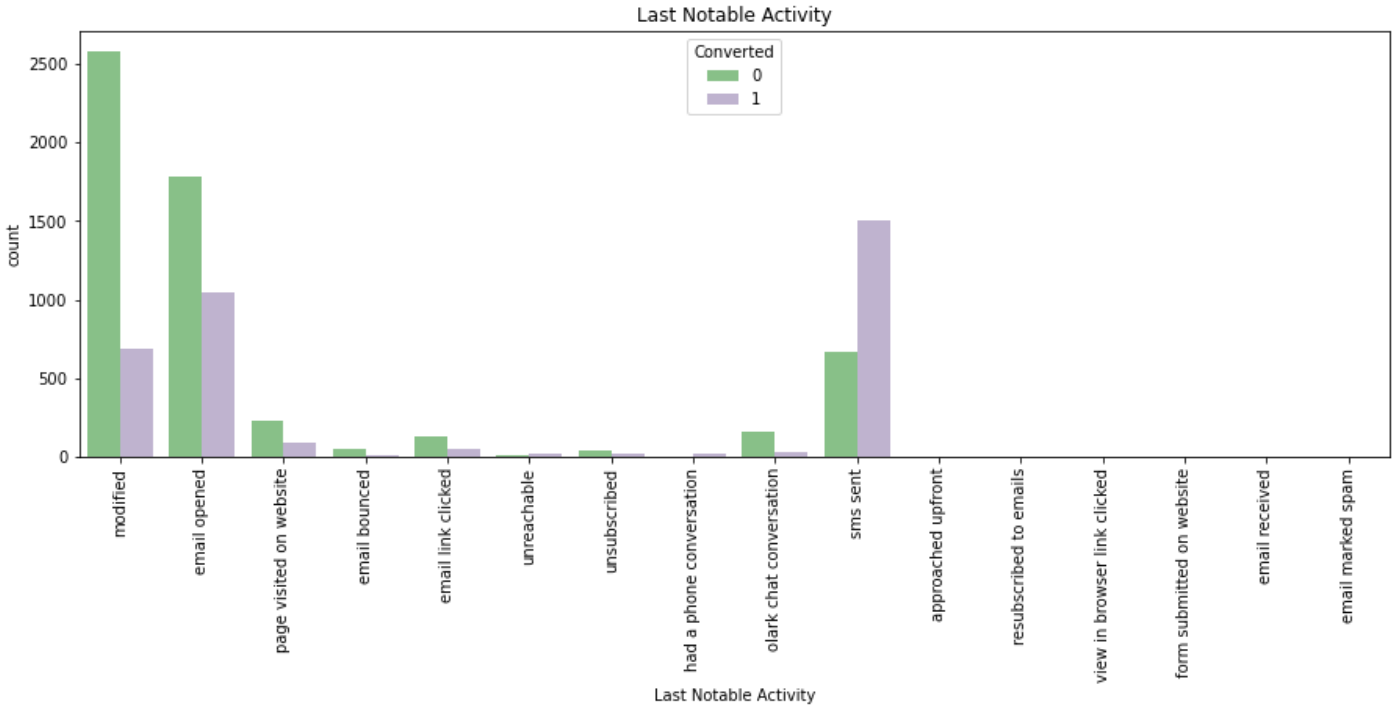
Bivariate Analysis
Categorical Variables

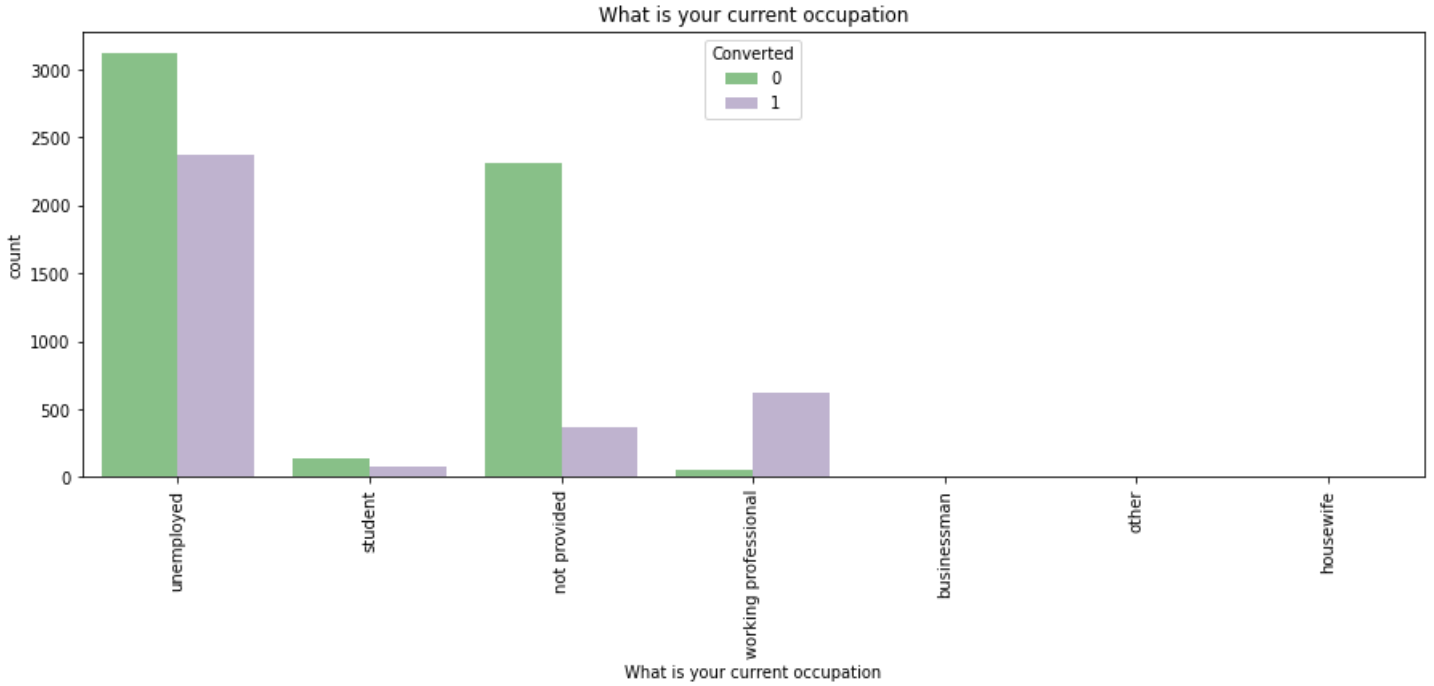


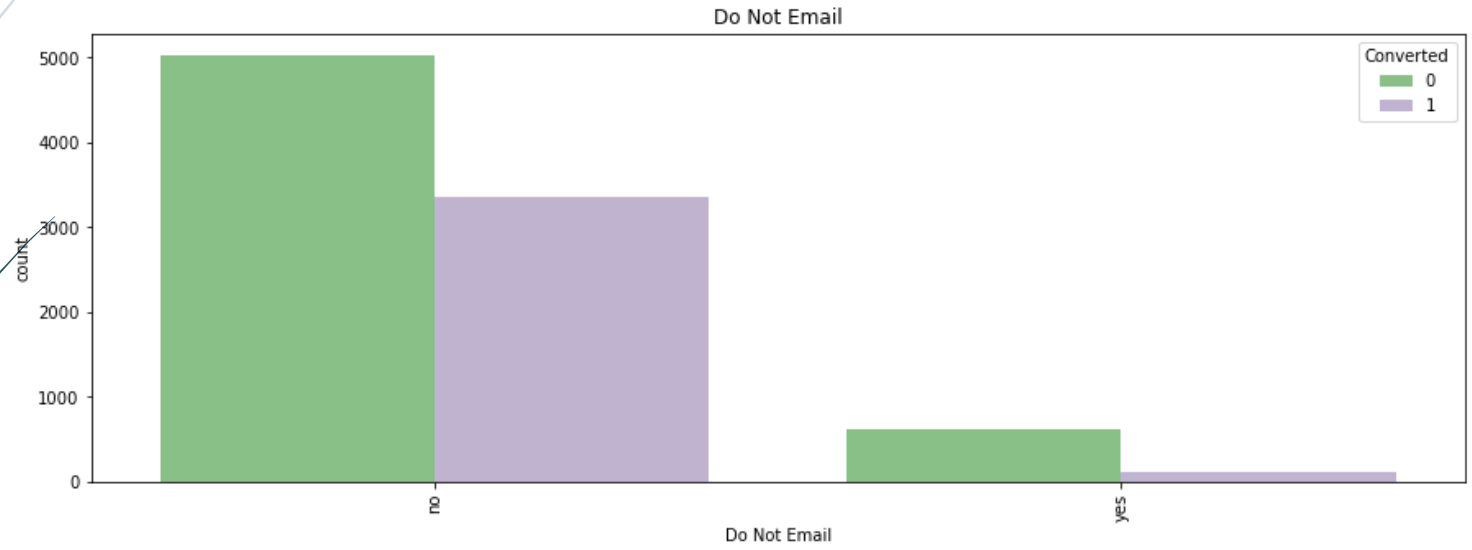
Bivariate Analysis Categorical Variable w.r.t Target

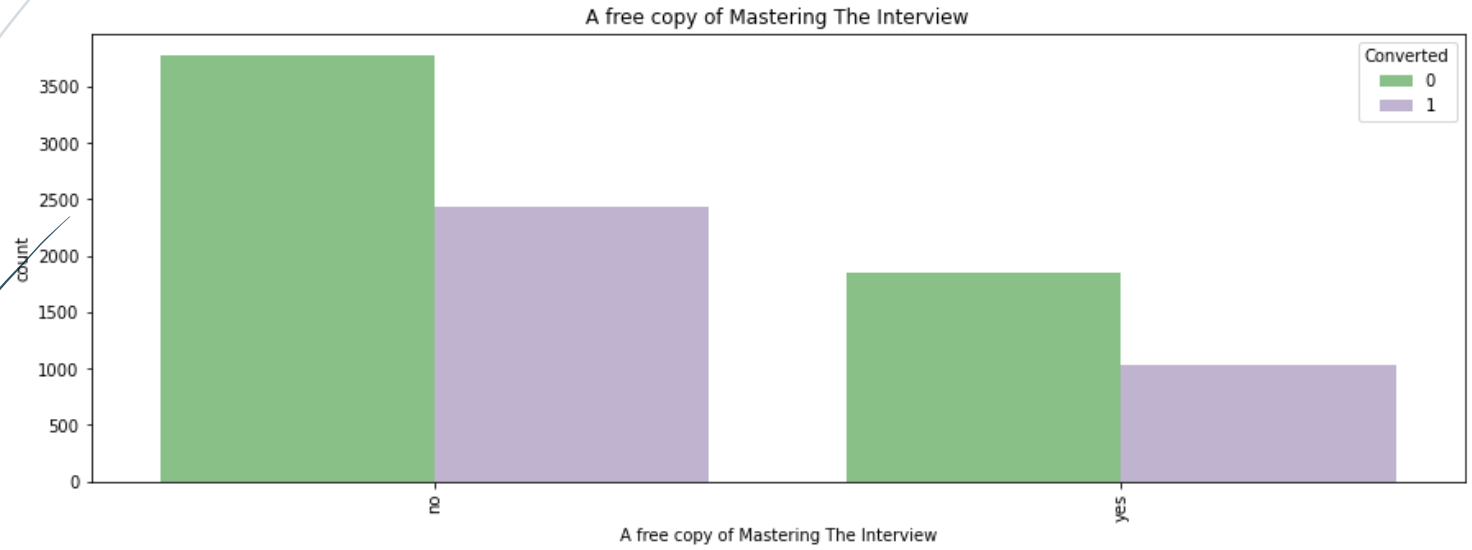

Lead Source
Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
olark chat
organic search
direct traffic
google
referral sites
reference
welingak website
facebook
blog
pay per click ads
bing
social media
welearn
click2call
live chat
welearnblog_home
youtubechannel
testone
press_release
nc_edm
Lead Source


Last Activity
Converted
0
1
count
2000
1500
1000
500
0
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
sms sent
visited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam
Last Activity

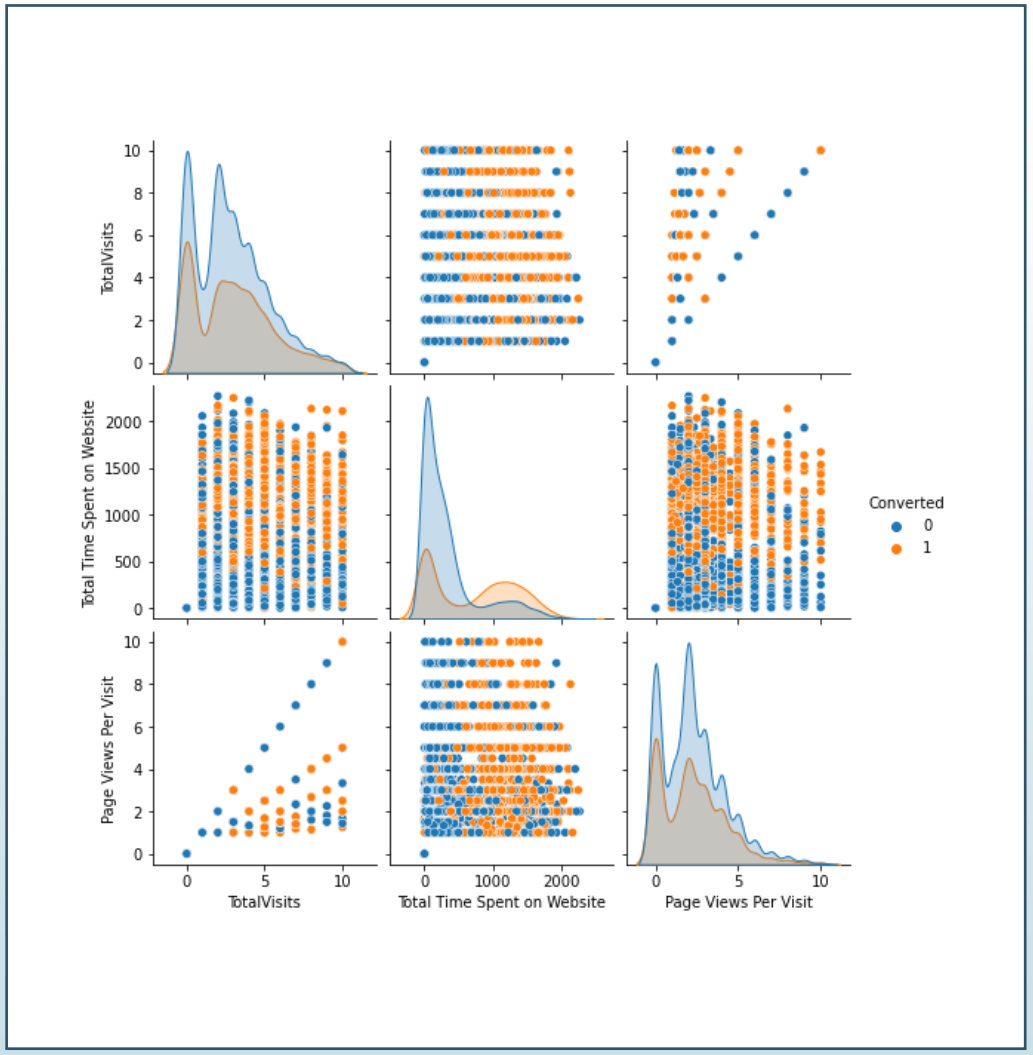

Last Notable Activity
Converted
0
1
count
2500
2000
1500
1000
500
0
modified
email opened
page visited on website
email bounced
email link clicked
unreachable
unsubscribed
had a phone conversation
olark chat conversation
sms sent
approached upfront
resubscribed to emails
view in browser link clicked
form submitted on website
email received
email marked spam
Last Notable Activity


What is your current occupation
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
unemployed
student
not provided
working professional
businessman
other
housewife
What is your current occupation


Do Not Email
Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
Do Not Email


A free copy of Mastering The Interview
Converted
0
1
count
3500
3000
2500
2000
1500
1000
500
0
no
yes
A free copy of Mastering The Interview

Multivariate Analysis



Correlation Scatter plot

'TotalVisits' and 'Page Views Per Visit' shows a positive correlation.

Lead spends more time on the web site or visit the website frequently or views more pages per visit is a good indication of successful conversion.

Data Preparation

Dummy Variable Creation and Feature Scaling

- After fixing outliers and missing values we proceed with dummy variable creation for all the categorical variable
- Next, we split the dataset into train and test set and do standardization on the features
- We Checked the correlation among the variables
- Attached heatmap is showing the correlation of all features present in the dataset.

As per the heatmap, variables which are highly correlated with Target are:

'Last Notable Activity_sms sent’

'Last Notable Activity_modified’

'What is your current occupation_working professional’

'What is your current occupation_not provided’

'Last Activity_sms sent’

Last Activity_olark chat conversation’

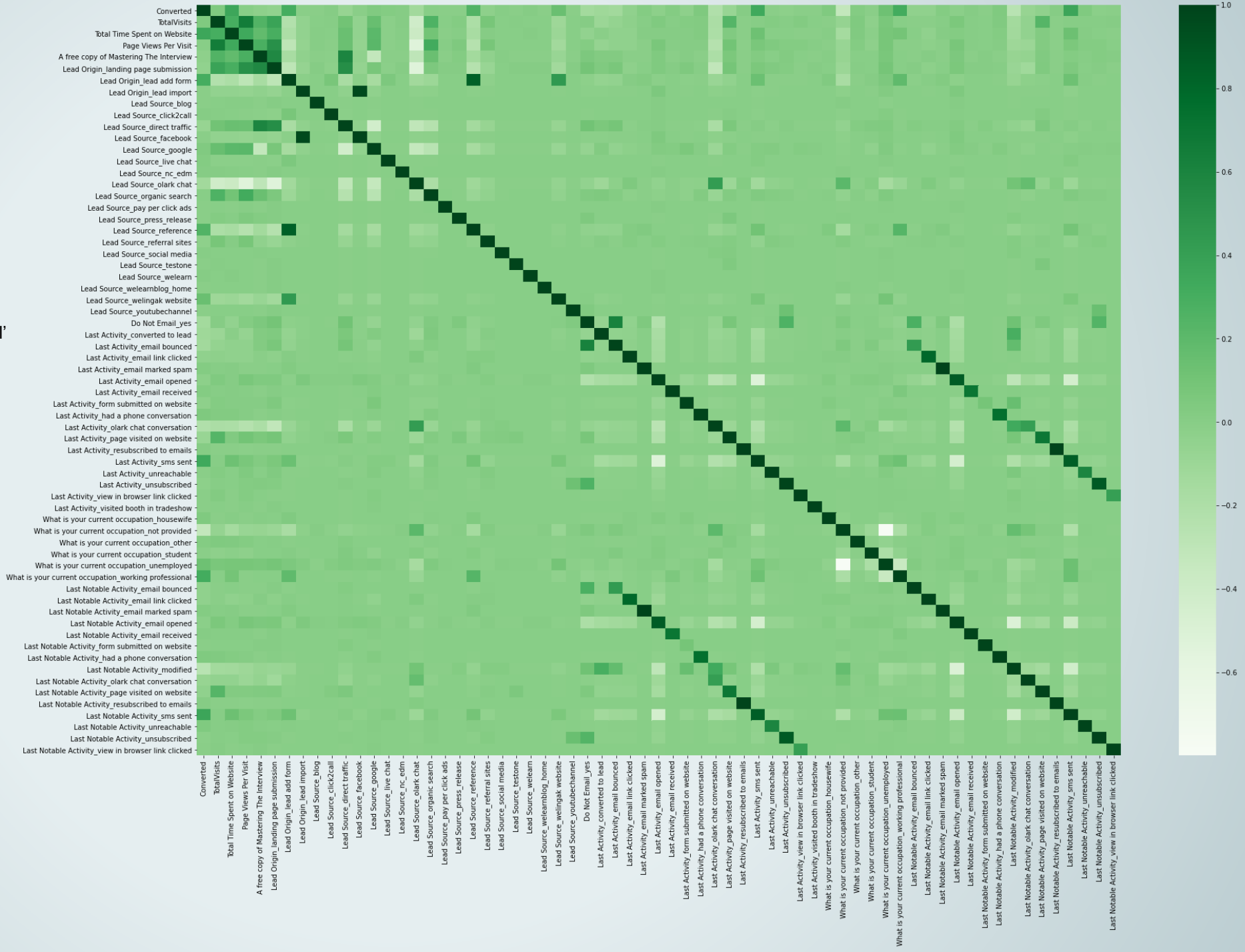
'Last Activity_page visited on website’

Total Time Spent on Website’

'Lead Origin_lead add form’

'Lead Source_olark chat’

Also, the heatmap shows several variables which are correlated to each other, we drop such variables.



Model Building using RFE

- After data cleaning and preparation, we are still left with 55 features.
- However, we need to eliminate all the insignificant features and filter the top 15 or 10 variables for our model.
- As an initial step for feature elimination, we use the scikit learn RFE class.
- We run RFE with 15 features to select.
- Further we built the model using the Generalized Linear Model (GLM) for Regression.
- After the manual elimination of all the non-significant variables, we have selected our final model with 12 most significant variables also without losing much accuracy.
- We used the default 0.5 cut off at each step of model building to maintain the model accuracy.
- We checked the precision and recall with accuracy, sensitivity and specificity for our final model and selected an optimum threshold of 0.34 for final prediction.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6363
Model:	GLM	**Df Residuals:**	6350
Model Family:	Binomial	**Df Model:**	12
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2554.9
Date:	Wed, 14 Jul 2021	**Deviance:**	5109.8
Time:	11:08:31	**Pearson chi2:**	6.40e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

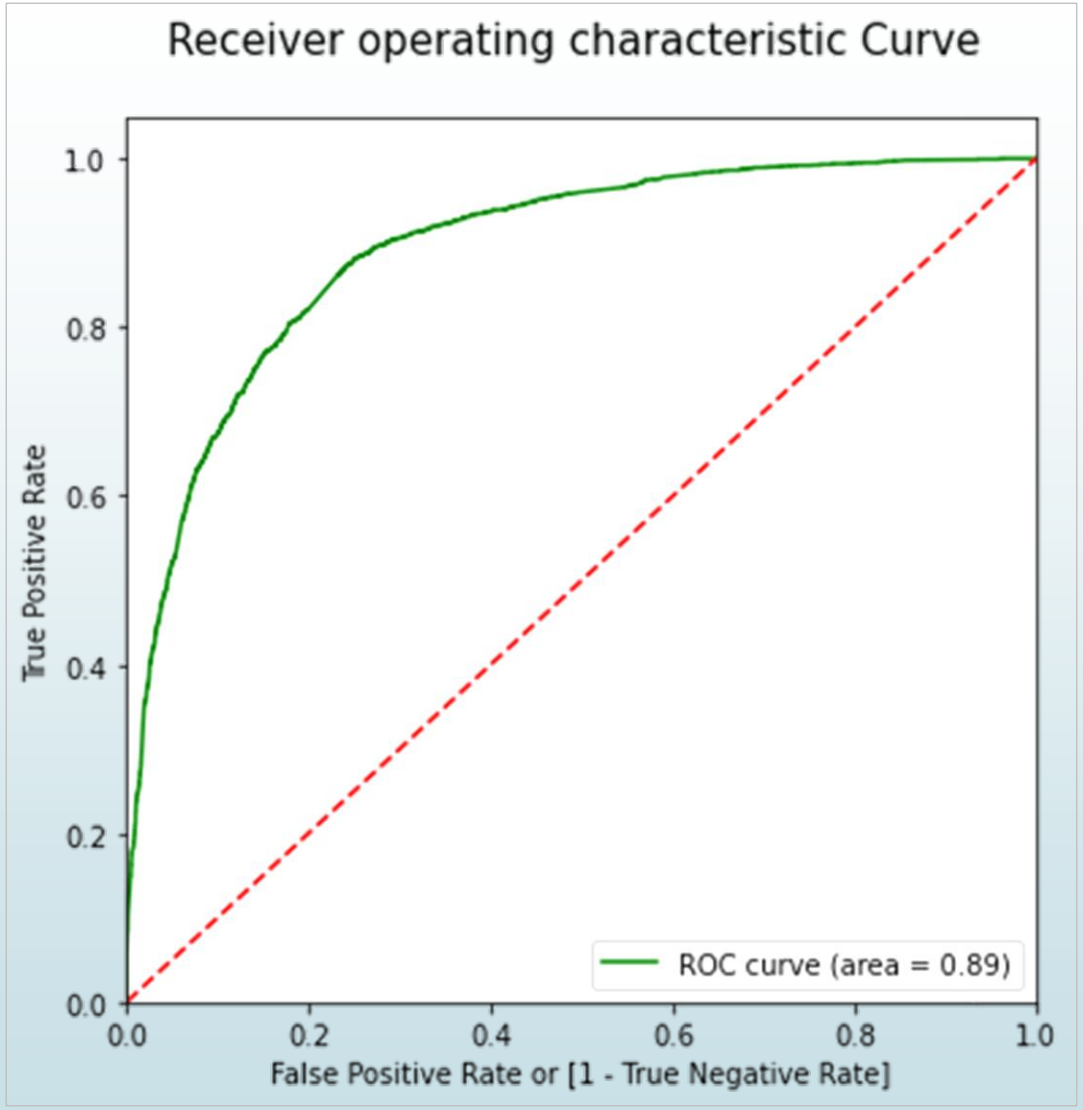
	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.1631	0.090	-24.058	0.000	-2.339	-1.987
TotalVisits	2.1493	0.790	2.720	0.007	0.601	3.698
Total Time Spent on Website	4.6187	0.169	27.313	0.000	4.287	4.950
Lead Origin_lead add form	4.1828	0.221	18.890	0.000	3.749	4.617
Lead Source_olark chat	1.4825	0.116	12.797	0.000	1.255	1.710
Do Not Email_yes	-1.4430	0.164	-8.792	0.000	-1.765	-1.121
Last Activity_converted to lead	-1.0070	0.206	-4.887	0.000	-1.411	-0.603
Last Activity_had a phone conversation	3.0894	0.868	3.560	0.000	1.388	4.790
Last Activity_olark chat conversation	-1.4940	0.169	-8.829	0.000	-1.826	-1.162
What is your current occupation_not provided	-1.2895	0.090	-14.314	0.000	-1.466	-1.113
What is your current occupation_working professional	2.4105	0.189	12.784	0.000	2.041	2.780
Last Notable Activity_sms sent	1.4569	0.081	17.985	0.000	1.298	1.616
Last Notable Activity_unreachable	1.4961	0.537	2.785	0.005	0.443	2.549

	Features	VIF
0	Total Time Spent on Website	1.84
1	TotalVisits	1.77
2	Lead Source_olark chat	1.57
3	What is your current occupation_not provided	1.45
4	Lead Origin_lead add form	1.44
5	Last Activity_olark chat conversation	1.43
6	Last Notable Activity_sms sent	1.37
7	Lead Source_welingak website	1.30
8	What is your current occupation_working profes...	1.19
9	Do Not Email_yes	1.07
10	Last Activity_converted to lead	1.04
11	Last Activity_had a phone conversation	1.01
12	Last Notable Activity_unreachable	1.01

Final model visualization with VIF

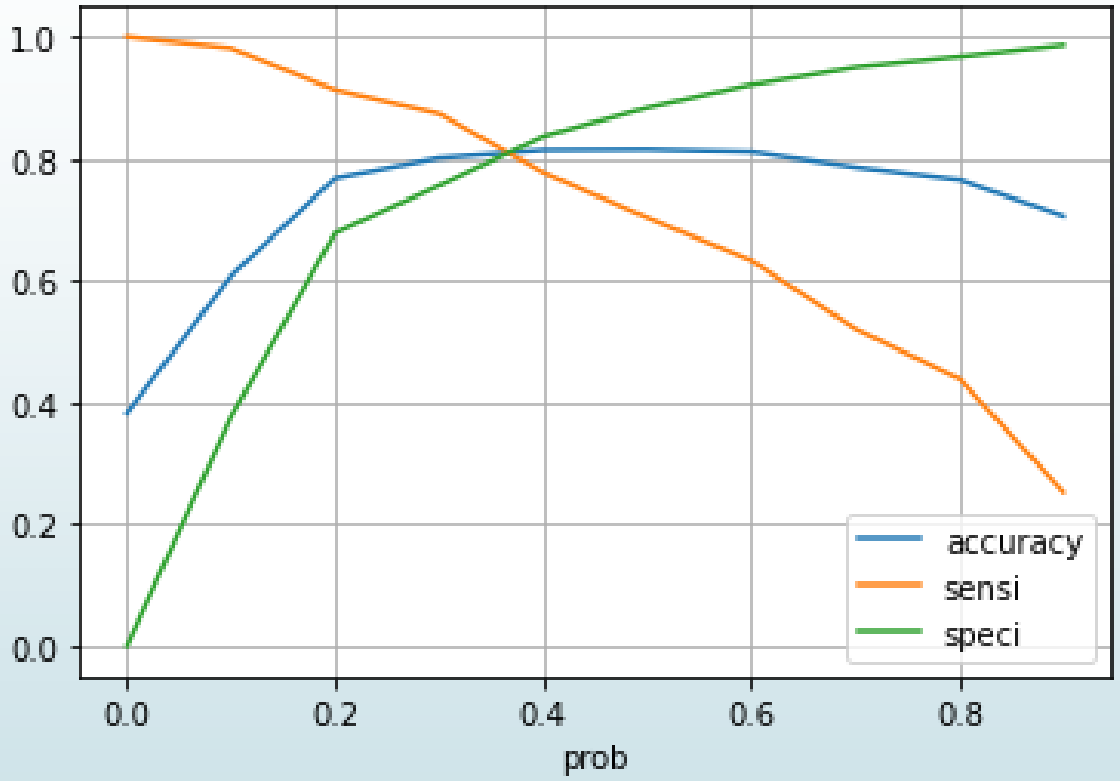
Model Evaluation

- After building the final model and making prediction on train data, we plot ROC curve (FPR Vs TPR) to find the fitness of the model.
- The curve is closer to the top left corner of the border, and this is a measure of good accuracy.
- Area under the curve for our model is 89%



Optimum cut off for Final Model

From the curve, we take a cut off at 0.34 which is an optimum threshold probability for the logistic regression's sigmoid function

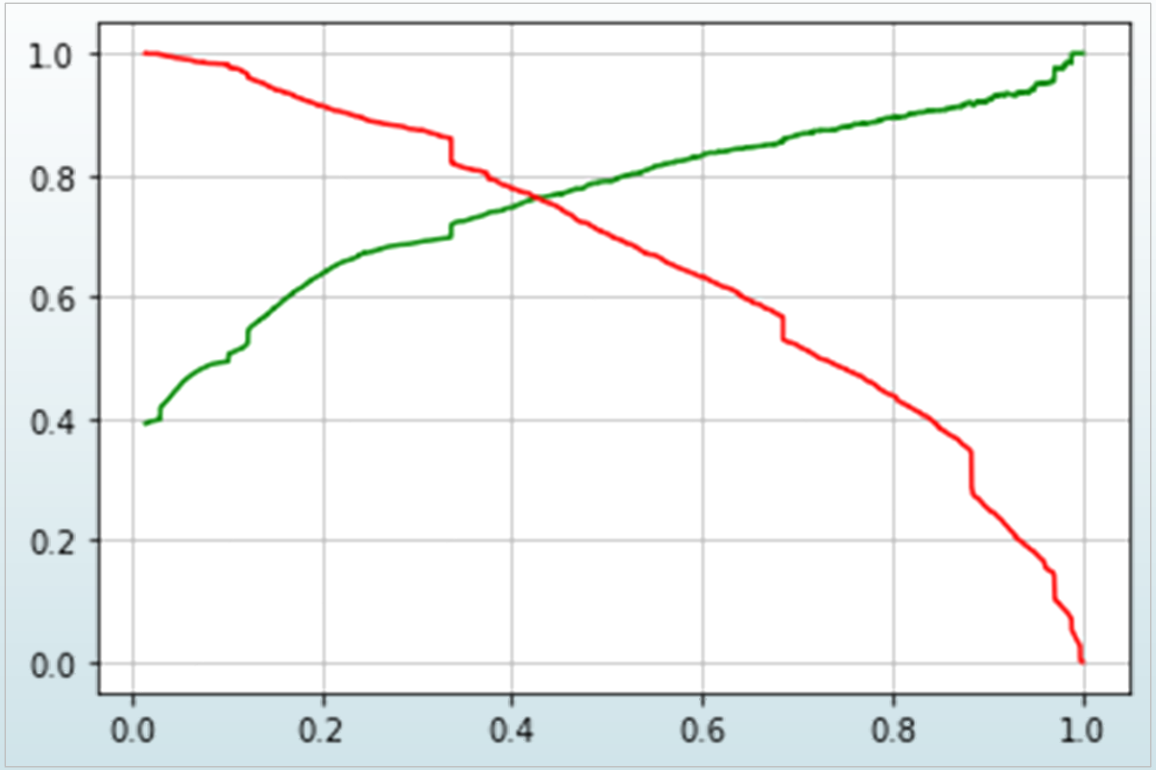


Precision and Recall

- We used 0.34 as the cutoff point for our final predictions
- We also examine another set of evaluation metrics namely, Precision and Recall
- Precision and Recall plays very important role in model building since our model is more business oriented and it also tells how our model behaves.
- Hence, we evaluated the precision and recall for this model and found the score as 72.23% for precision and 81.67% for recall.
- For the X Education business point of view , the recall percentage is more crucial since we don't want to miss any hot leads, on the other hand it is okay if our precision is low which means less hot lead customers. Hence the focus is for higher Recall than Precision.

Precision - Recall trade off

There is a trade off between Precision and Recall and the cut off point lies around 0.4



Validation

- Final predictions are done on validation data set and predicted values were recorded.
- Added a **Lead Score** for each lead in our dataset to identify the **hot leads**.
- Conducted model evaluation on validation dataset as well.
- The test prediction is having accuracy , specificity, precision and recall score in an acceptable range.
- This shows that the model is stable with good accuracy and recall/sensitivity.

SCORES FOR THE TRAINING DATA SET

Accuracy : 80.98 %

Sensitivity : 81.67 %

Specificity : 80.56 %

Precision : 72.23 %

Recall : 81.67 %

F1 score : 76.66 %

SCORES FOR THE VALIDATION DATATSET

Accuracy : 80.5 %

Sensitivity : 80.78 %

Specificity : 80.33 %

Precision : 71.2 %

Recall : 80.78 %

F1 score : 75.69 %

Evaluation metrics for Train Vs Test data

Conclusion

- Validation dataset shows promising Accuracy, Precision and Recall/Sensitivity in comparison with train data.
- We have high recall rate of 80.78% which is in line with the X Education's business requirement.
- The model has an ability to adjust with the company's requirements in coming future.
- Lead Score helps to identify the hot leads , where a higher lead score has more chance of turning into a successful conversion and vice versa.
- This model is quite stable with varying business needs.

Recommendations

To achieve maximum conversions, sales team should target customers who :

- Spend more time on X Education Website
- Fills an Application Form
- Had a phone conversation with the sales team
- Working Professionals
- Frequent Visitor of X Education Website

Thank you