LEAD SCORING CASE STUDY

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 a higher lead score have a higher conversion chance and the customers with a 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%.

STEPS FOR ANALYSIS

1) READING AND UNDERSTANDING THE DATA
2) CLEANING THE DATA AND PERFORMING EDA
3) MODEL PREPARATION
4) LOGISTIC REGRESSION MODEL BUILDING AND EVALUATION
5) PREDICTION ON TEST DATA
6) RECOMMENDATIONS

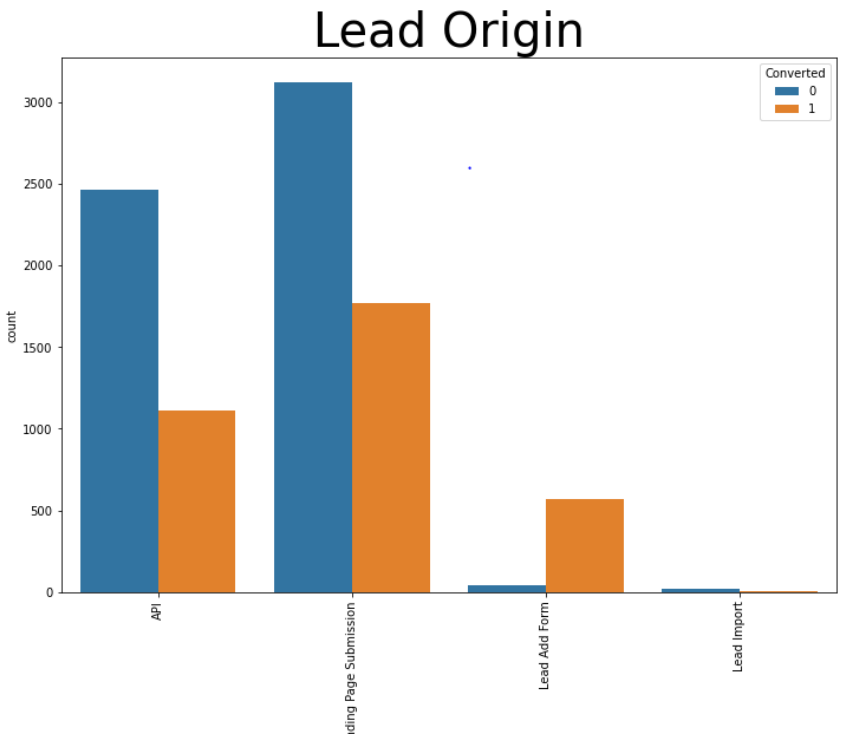
1) READING AND UNDERSTANDING THE DATA

- There are 9240 rows and 37 columns present in the data set
- The column named ‘Converted’ is the target variable which describes Whether a lead has been successfully converted or not.

2) CLEANING THE DATA AND PERFORMING EDA

- All the ‘Select’ values of ‘Specialization’ column are replaced by null values
- The columns containing null values greater than 45% are dropped. The columns ‘Prospect ID’ and ‘Lead Number’ describes the unique ID and lead number assigned to each customer are dropped. Adding these columns into analysis will not give further insights.
- The columns containing null values less than 45% are imputed with most occurred value from that Particular column.
- The outliers present in the numerical columns are removed.

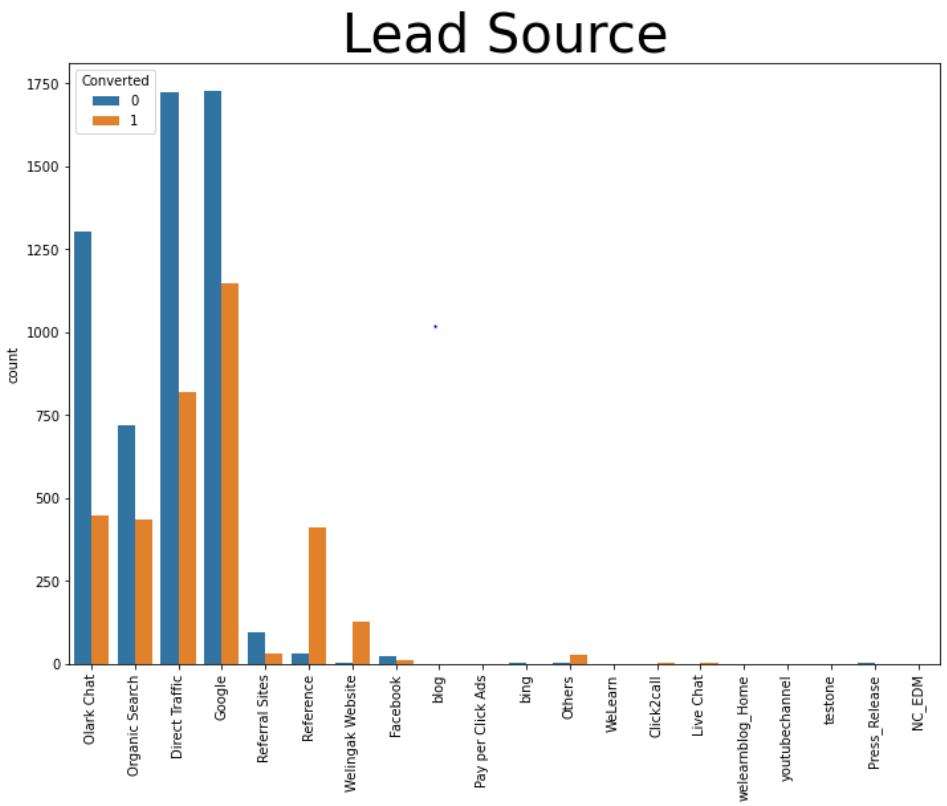
2.1) INSIGHTS ON CATEGORICAL COLUMNS



- Column ‘Lead Origin’

1. The column represents the origin identifier with which the customer was identified as a lead.

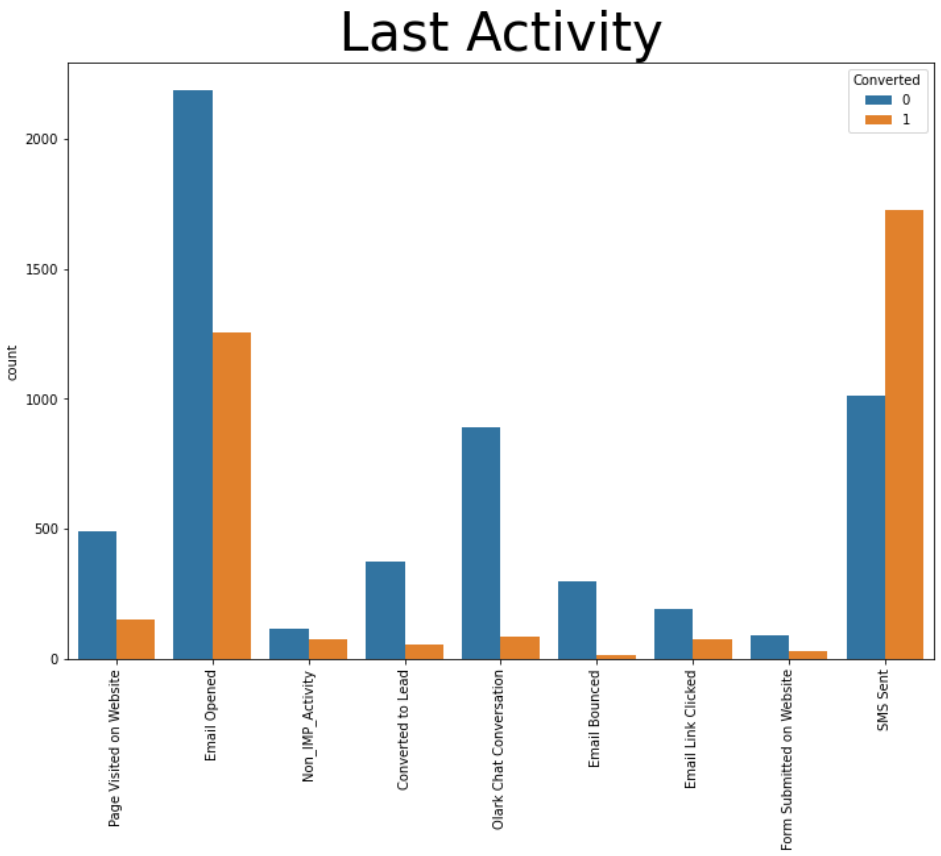
2. Most of the leads originated from landing page submission and API out of which the conversion rate of landing page submission is greater than API.



- Column ‘Lead Source’

1. The column describes the source of lead.

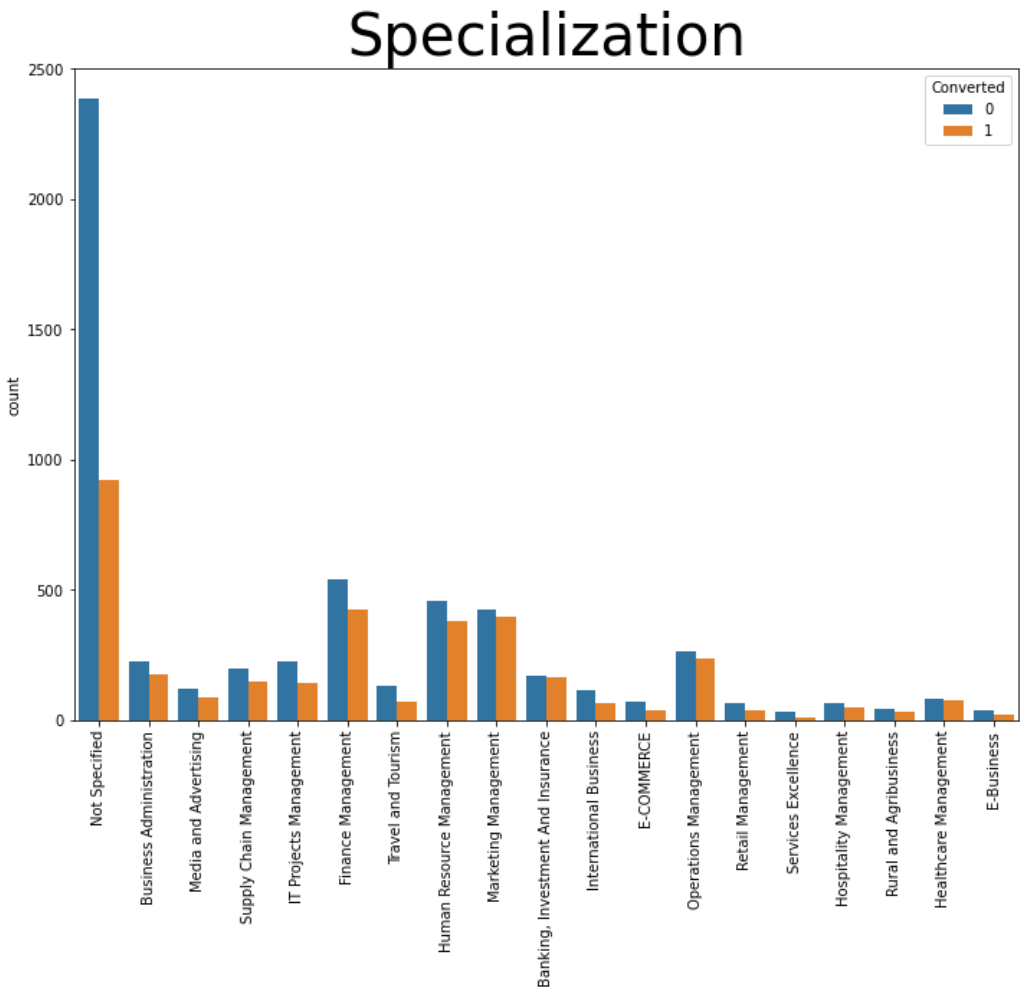
2. Olark chat, organic search, direct traffic, google these lead sources can create and able to convert maximum number of leads.



- Column ‘Last Activity’

1. The column describes the last activity performed by the customer.

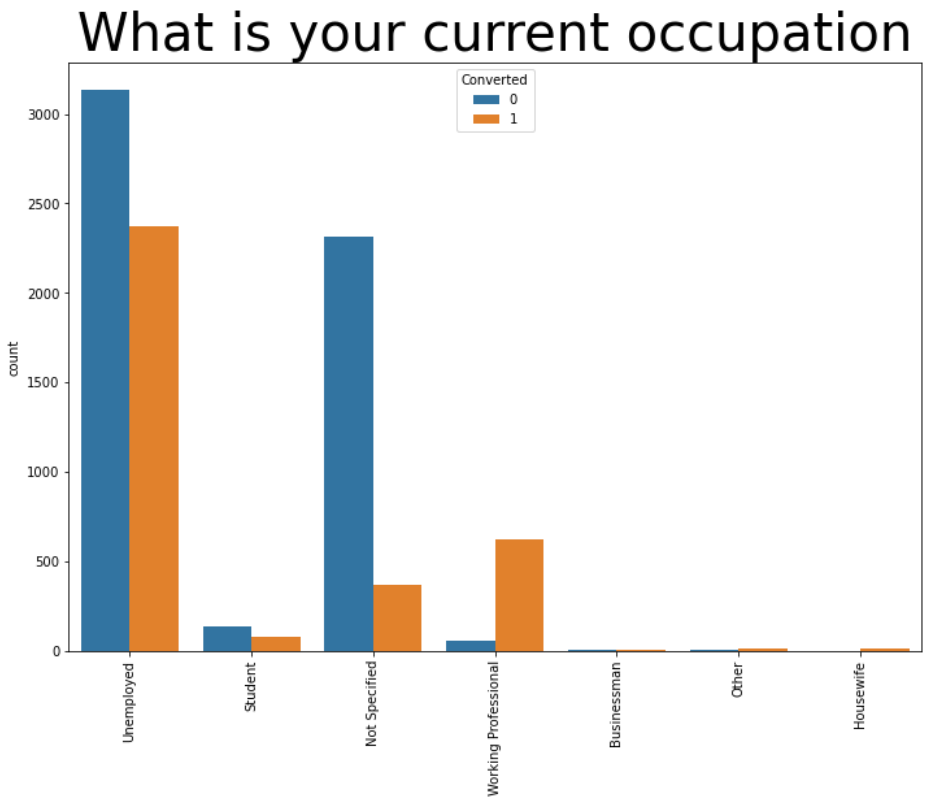
2. The last activity done by the customer i.e. email opened, SMS sent create more number of leads and able to convert these leads to take the course. Focusing more on these last activities of the customer will help to convert the leads.



- Column ‘Specialization’

1. The column describes the industry domain in which customer worked before.

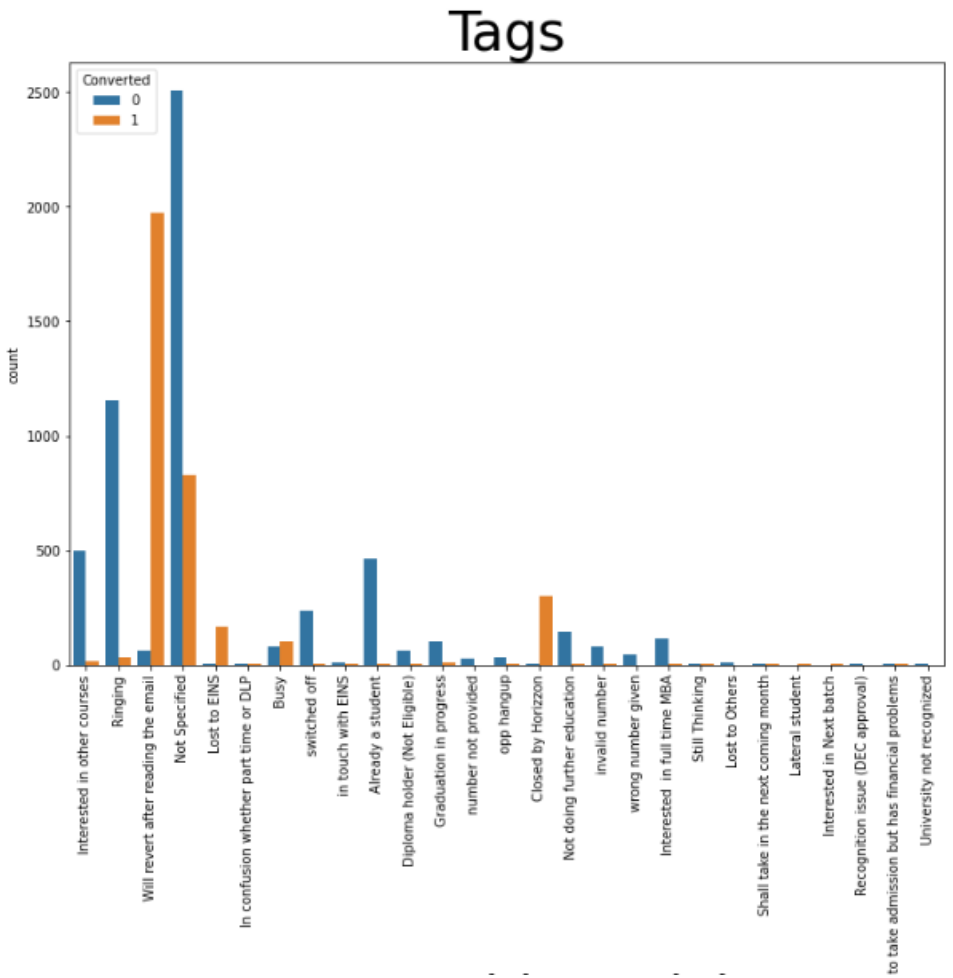
2. The maximum number of leads created and converted have not specified their specialization. Although, the managment give better results in terms of converting the leads to take the course.



- Column ‘What is your current occupation’

1. The column indicates whether the customer student, unemployed or businessman.

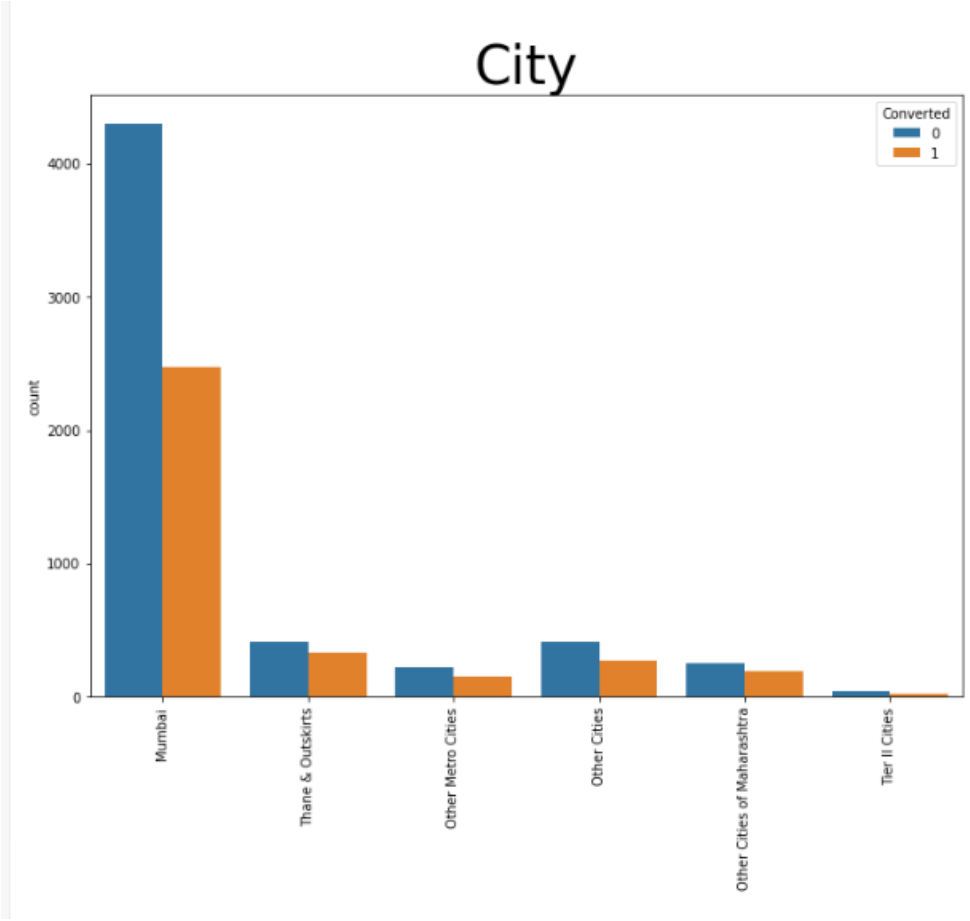
2. Most of the leads converted are unemployed. Focusing more on student and working professionals will give better results.



- Column ‘Tags’

1. The column indicates the current status of lead.

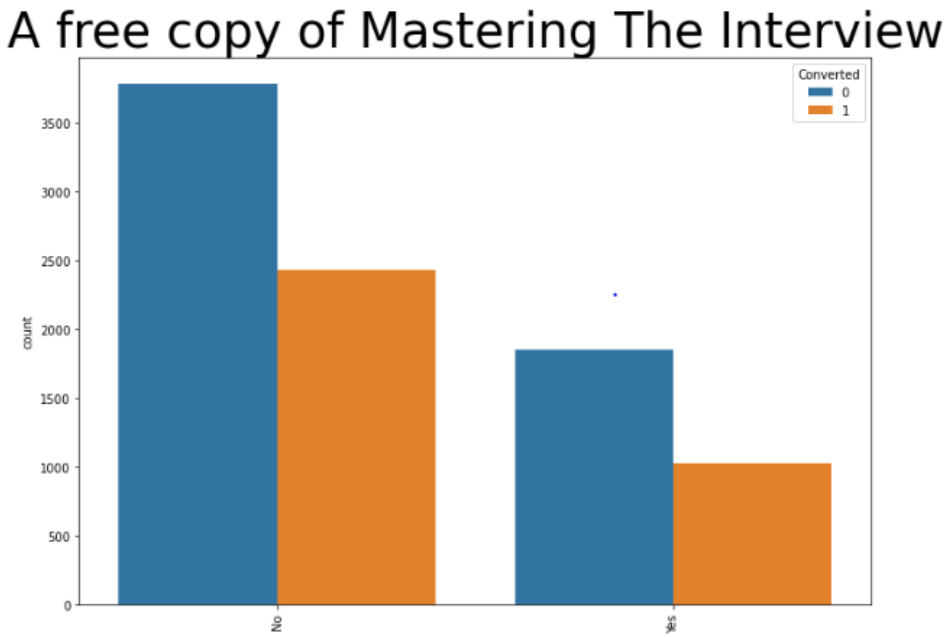
2. The current status of the lead 'will revert after reading the mail' has more lead conversion rate but not able to create more number of leads. Most of the leads created have not specified their current status.



- Column ‘City’

1. The column indicates the city of the customer.

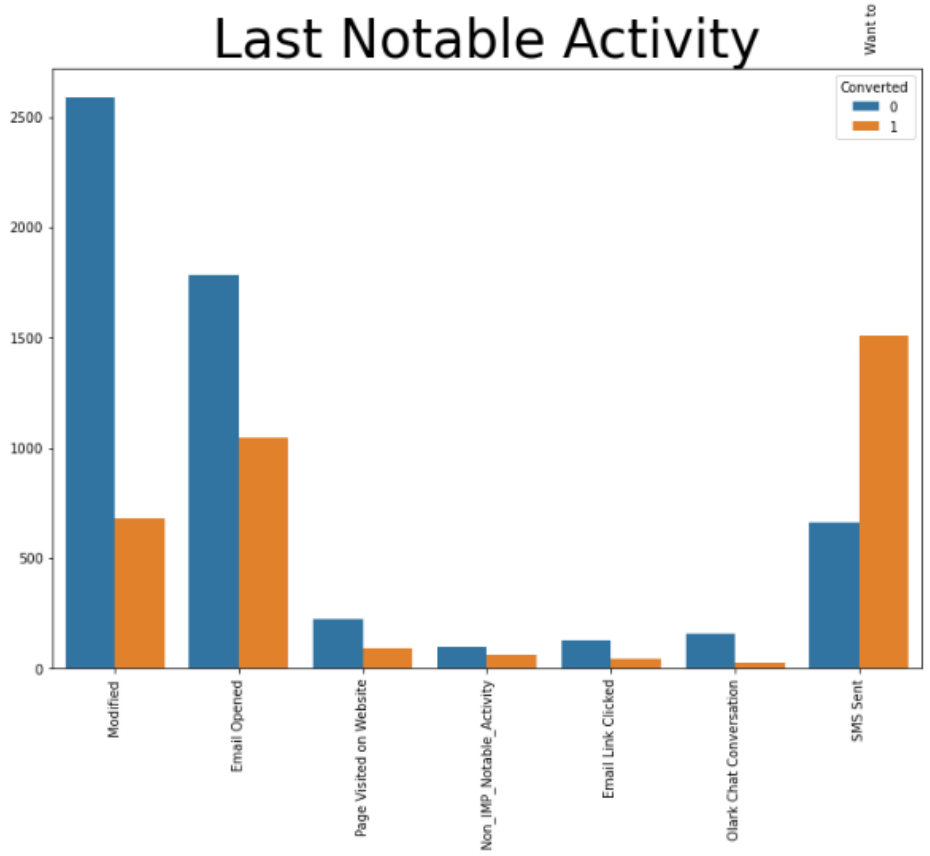
2. Mumbai city will create maximum number of leads and able to convert them to take a course followed by Thane and other metro cities.



- Column ‘A free copy of mastering the interview’

1. The column indicates whether the customer wants a free copy of 'Mastering the Interview' or not.

2. The maximum number of leads converted do not want a free copy of mastering the interview.

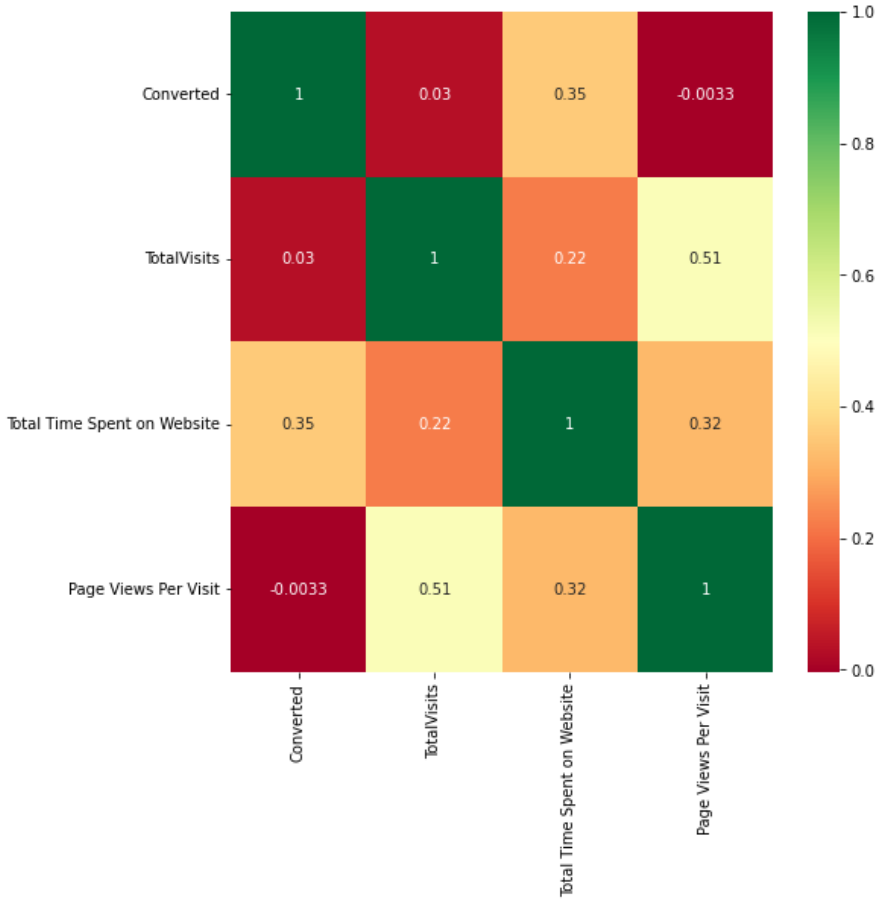


- Column ‘Last Notable Activity’

1. The column indicates the last notable activity performed by the customer.

2. There is a high correlation between last activity and last notable activity.

2.2) INSIGHTS ON NUMERICAL COLUMNS



- From the heat map we can see that, there is a high correlation between total number of visits made by the customer on the website and average number of pages on the website viewed during the visits.

3) MODEL PREPARATION

- Dummy variables are created for all categorical variables and concatenated these dummies with our master data frame ‘leads’.
- The original categorical columns for which dummies are created are dropped from master data set.
- We left with 90 columns in our data with no categorical variable present in the data.
- This data is used to create Logistic Regression model.

4) LOGISTIC REGRESSION MODEL BUILDING AND EVALUATION

- We splitted the data into Target and Independent variables and with the help of train_test_split from sklearn we splitted it further into training and testing data sets.
- Numeric variables from training data set are scaled to brought in equal range.
- The RFE algorithm is used with initial 20 variables for training data set.
- The output of the RFE algorithm is combined with the columns of training data along with ranking.
- We selected the columns that are ranked 1 by RFE and all other columns not supported by RFE are dropped.
- Constant is added and Generalized Linear Model is used to fit the training data.

4.1) LOGISTIC REGRESSION MODEL RESULTS

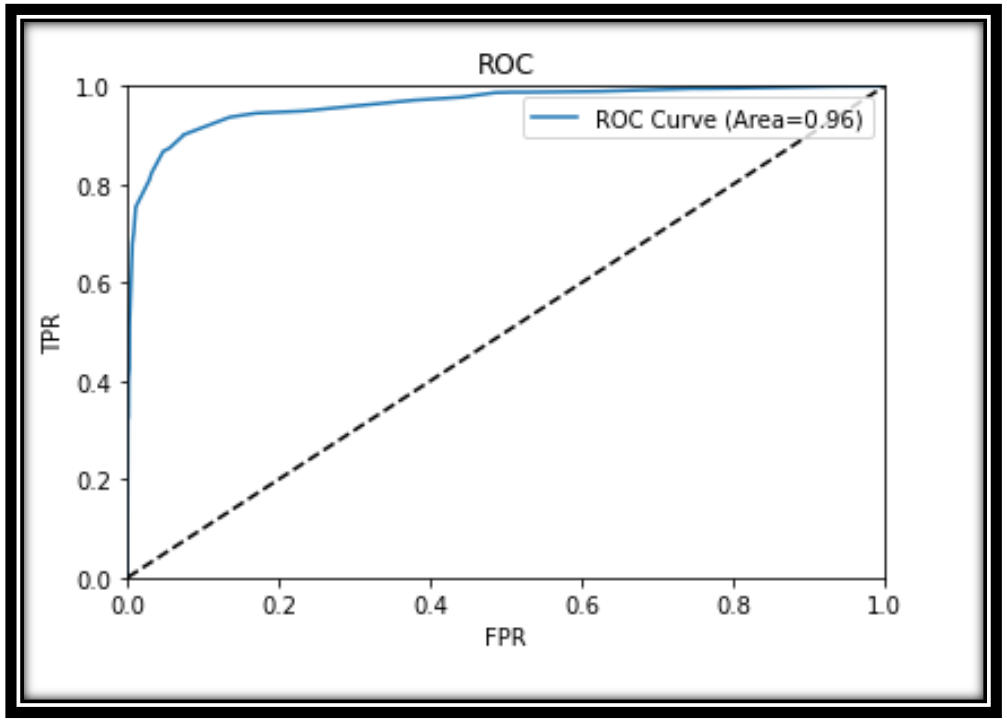
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6356
Model:	GLM	**Df Residuals:**	6338
Model Family:	Binomial	**Df Model:**	17
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1399.7
Date:	Mon, 17 Apr 2023	**Deviance:**	2799.4
Time:	03:07:36	**Pearson chi2:**	1.02e+04
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.2045	0.103	-21.429	0.000	-2.406	-2.003
Lead Origin_Lead Add Form	0.8431	0.414	2.039	0.041	0.033	1.654
Lead Source_Olark Chat	0.2855	0.119	2.400	0.016	0.052	0.519
Lead Source_Welingak Website	3.0841	0.846	3.644	0.000	1.425	4.743
Last Activity_SMS Sent	2.1995	0.110	20.086	0.000	1.985	2.414
What is your current occupation_Unemployed	1.2395	0.108	11.447	0.000	1.027	1.452
What is your current occupation_Working Professional	1.3330	0.358	3.724	0.000	0.632	2.034
Tags_Closed by Horizzon	6.4806	0.731	8.861	0.000	5.047	7.914
Tags_Diploma holder (Not Eligible)	-2.6855	1.033	-2.600	0.009	-4.710	-0.661
Tags_Interested in full time MBA	-2.9045	1.014	-2.865	0.004	-4.891	-0.917
Tags_Interested in other courses	-2.0154	0.343	-5.875	0.000	-2.688	-1.343
Tags_Lost to EINS	6.2540	0.603	10.366	0.000	5.072	7.436
Tags_Not doing further education	-3.0378	1.022	-2.973	0.003	-5.040	-1.035
Tags_Ringing	-3.8303	0.233	-16.406	0.000	-4.288	-3.373
Tags_Will revert after reading the email	4.5888	0.211	21.786	0.000	4.176	5.002
Tags_opp hangup	-1.5894	0.708	-2.244	0.025	-2.978	-0.201
Tags_switched off	-4.3991	0.597	-7.372	0.000	-5.569	-3.230
Last Notable Activity_Modified	-1.6993	0.124	-13.759	0.000	-1.941	-1.457

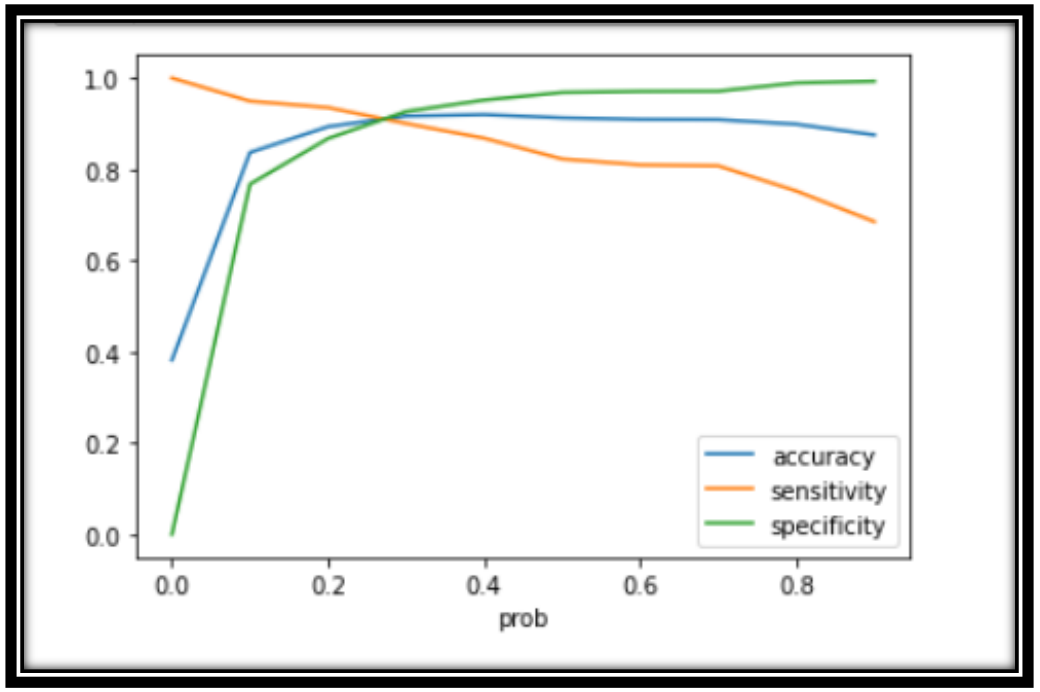
- All the variables have P value less than 0.05 which indicates that these variables are good enough for prediction.
- There is no redundant variable present in the data as the VIF of all variables are less.
- Prediction is done on the training set.

4.2) LOGISTIC REGRESSION MODEL EVALUATION



- ROC curve shows the trade off between sensitivity and specificity.
- Good ROC curve touches the upper left corner of the graph as we can see above. Higher the area under the curve of ROC better is the model. We got area = 0.96 indicating a good predictive model.

CONTINUED..



- The graph shows the 'probability' vs. 'accuracy', 'sensitivity', 'specificity' for various probability cutoffs.
- From this graph, the optimal probability cutoff point is selected as 0.27.
- The performance of the model on training data is :
- accuracy 89.33%
- Sensitivity 93.53%
- specificity 86.73%

5) PREDICTION ON THE TEST DATA

- Numeric variables from test data set are scaled to brought in equal range.
- The columns of training data set are selected for test data set.
- Constant is added to test data and prediction of target variable is done.
- Performance of the model on test data set is :
- accuracy 89.14%
- Sensitivity 92.7%
- specificity 86.98%

6) RECOMMENDATION

- The model seems to predict the conversion rate very well.
- Focusing more on the sources of leads like Direct Traffic, last notable activity like Olark chat conversation or paying attention on the current status of the customer will provide better results