LEAD SCORE CASE STUDY

- MAHESH BABU R & CHITRA D

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. The typical lead conversion rate at X education is around 30%.
- 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.

CASE STUDY OBJECTIVE

- The company requires us to build a model wherein we 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%.

STEPS INVOLVED IN PROBLEM SOLVING

- Understanding problem statement & Business Objective
- Understanding of Data
- Data Cleansing and Outlier Treatment
- Feature Scaling & Data Split
- Model Building
- Evaluating the model on Test Data

BUSINESS OBJECTIVE & DATA UNDERSTANDING

- Business objective is to find out the potentially hot leads to increase the lead conversion rate. Also help the company to understand the factors which are most critical for lead conversion
- Data Understanding :
  - Dataset provided to us has around 9000 data points.
  - It has various attributes such as Lead source, Total time spent on website, Total Visits, Last activity done etc.. Which can be used to predict whether a lead will be converted or not.
  - Target variable in this case is 'Converted' which tells us based on the past data whether the lead was converted or not

DATA CLEANSING & OUTLIER TREATMENT

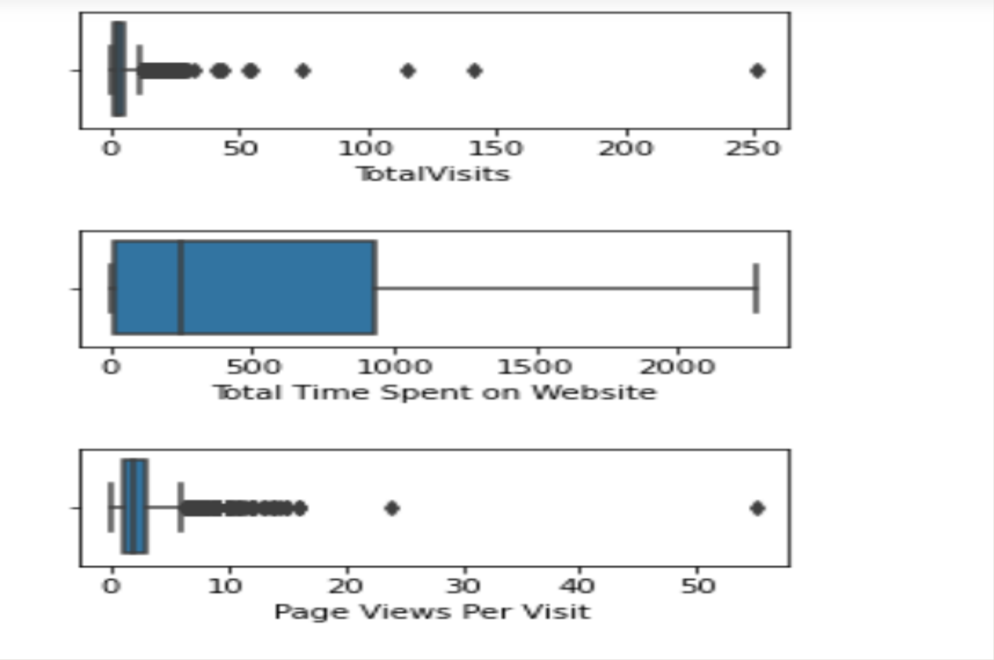
- Initially the data is imported and basic dataset shape,describe,data types are checked.
- Missing value Treatment : Here the % of null counts for each column is checked. All the Select values are replaced by null and overall percentage of each column is checked. Cut off of 30 % is taken and columns having more than 30 % nulls is dropped.
- For few columns based on the datatype and null value percent ; categorical columns null values are replaced with mode and numerical null values are replaced with the median .
- For few categorical columns the null percent data is huge, so by replacing such nulls with mode data becomes biased , hence it is filled as not provided.
- For outlier treatment 99 percentile of the columns is considered and the datapoints having less than the 99 percentile values is considered for our analysis

DATA CLEANSING & OUTLIER TREATMENT (CONT..)

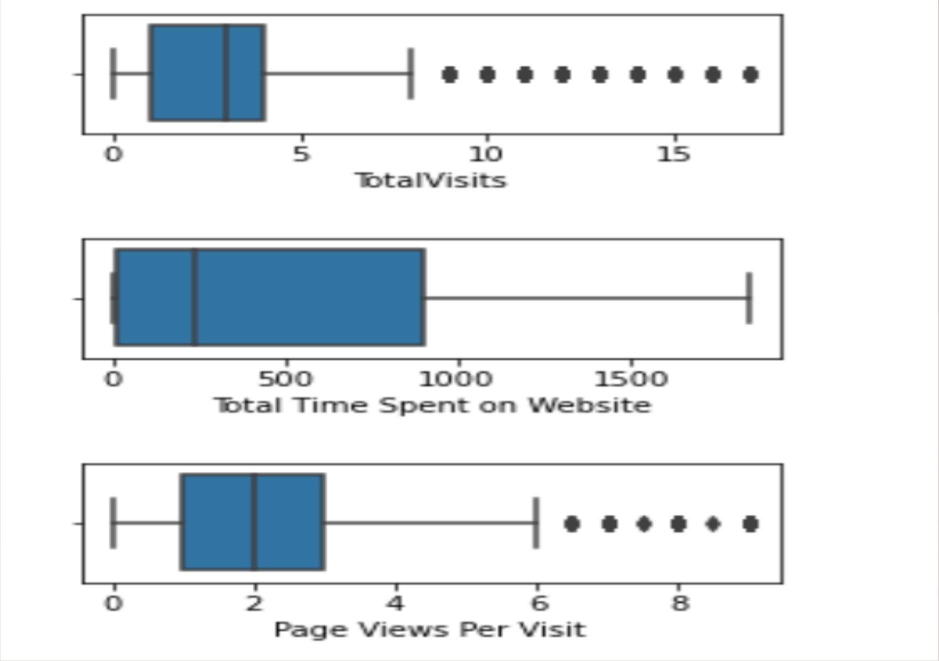
- After null handling and outlier treatment we are left with 98% of the datapoints.
- Univariate Analysis : It is performed on all the categorical columns and the columns which have the majority of data as a single value cannot be useful sed for our analysis
  - Eg : 'Do Not Email','Do Not Call','Search','Magazine','Newspaper Article','X Education Forums','Newspaper','Through Recommendations','Digital Advertisement','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 are dropped , as majority of the data in these columns is having only singlevalue.

BEFORE & AFTER OUTLIER TREATMENT

BEFORE OUTLIER TREATMENT

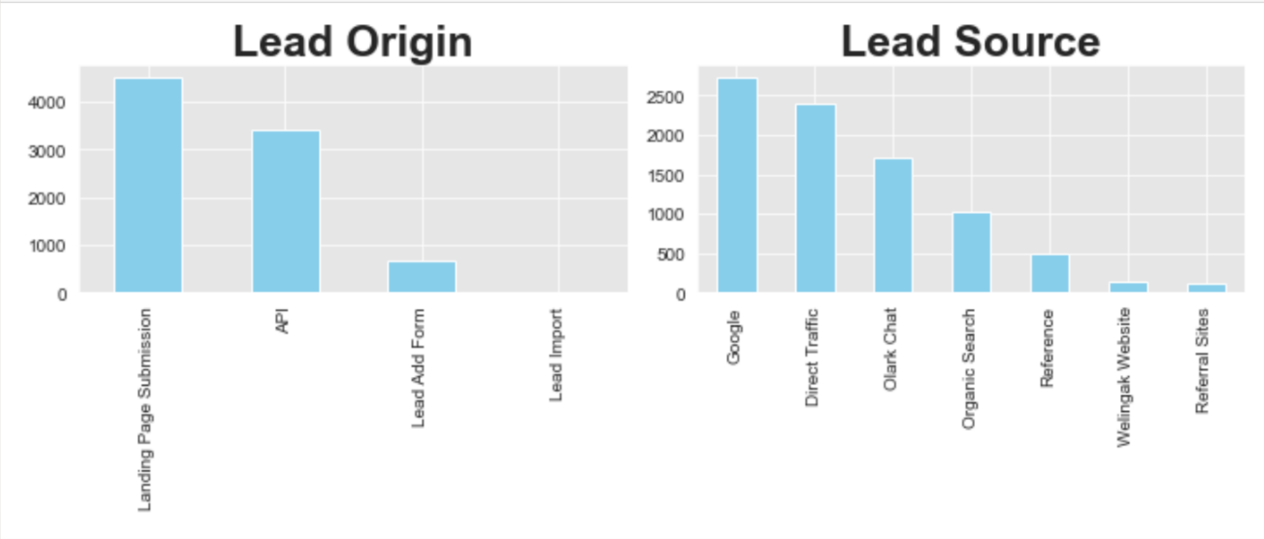


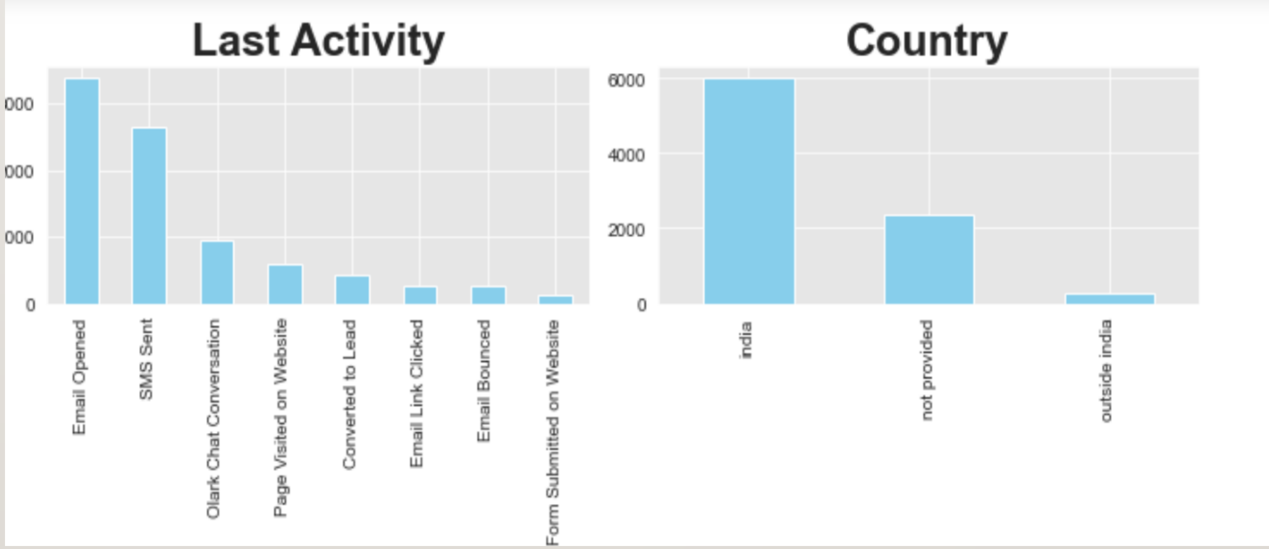
AFTER OUTLIER TREATMENT



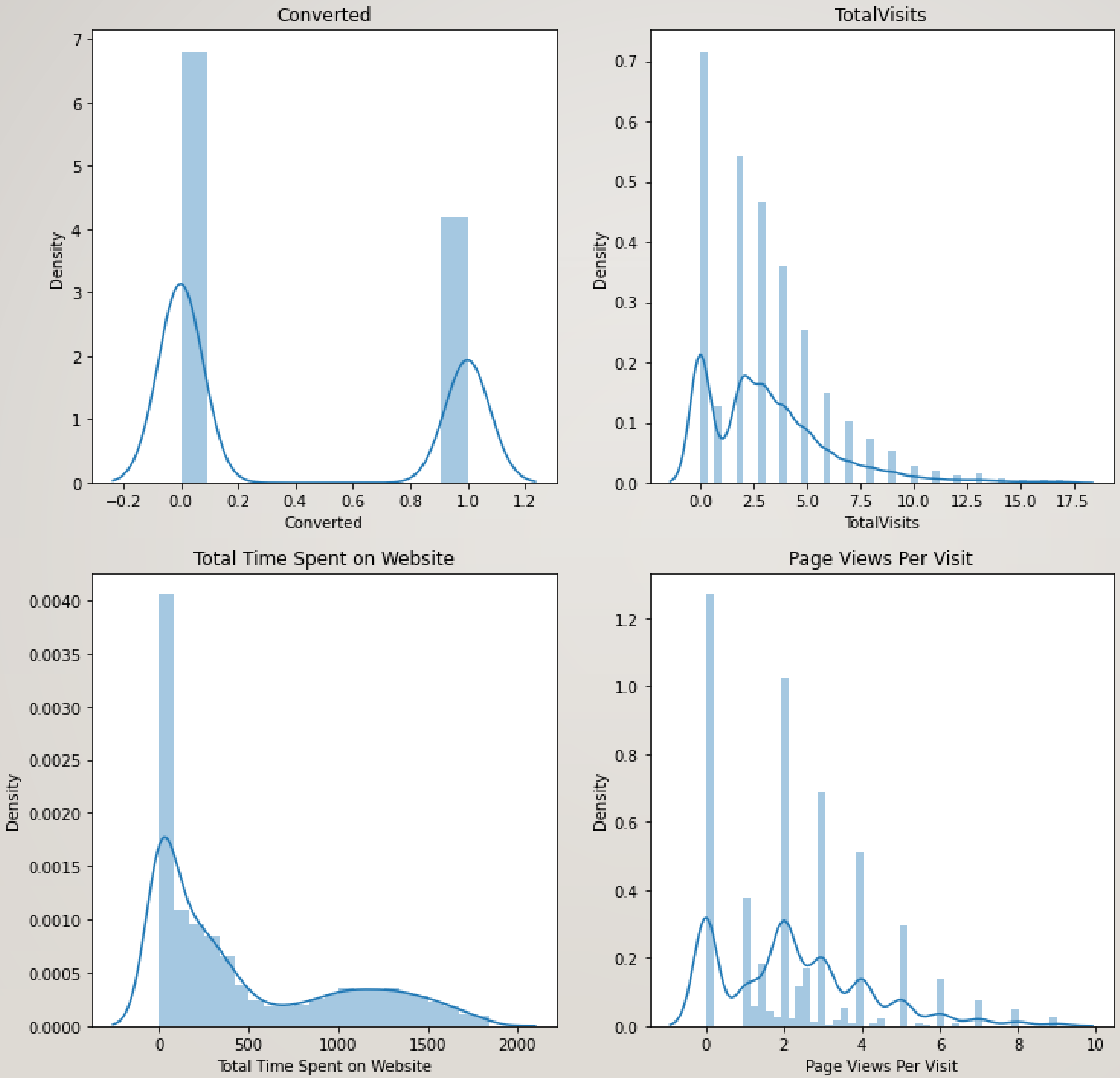
UNIVARIATE ANALYSIS

- Most of the Converted Leads Lead Origin is from Landing Page Submission.
- Most of the Converted Lead is from Google.
- Last Activity done was Email Opening and converted leads are from India.

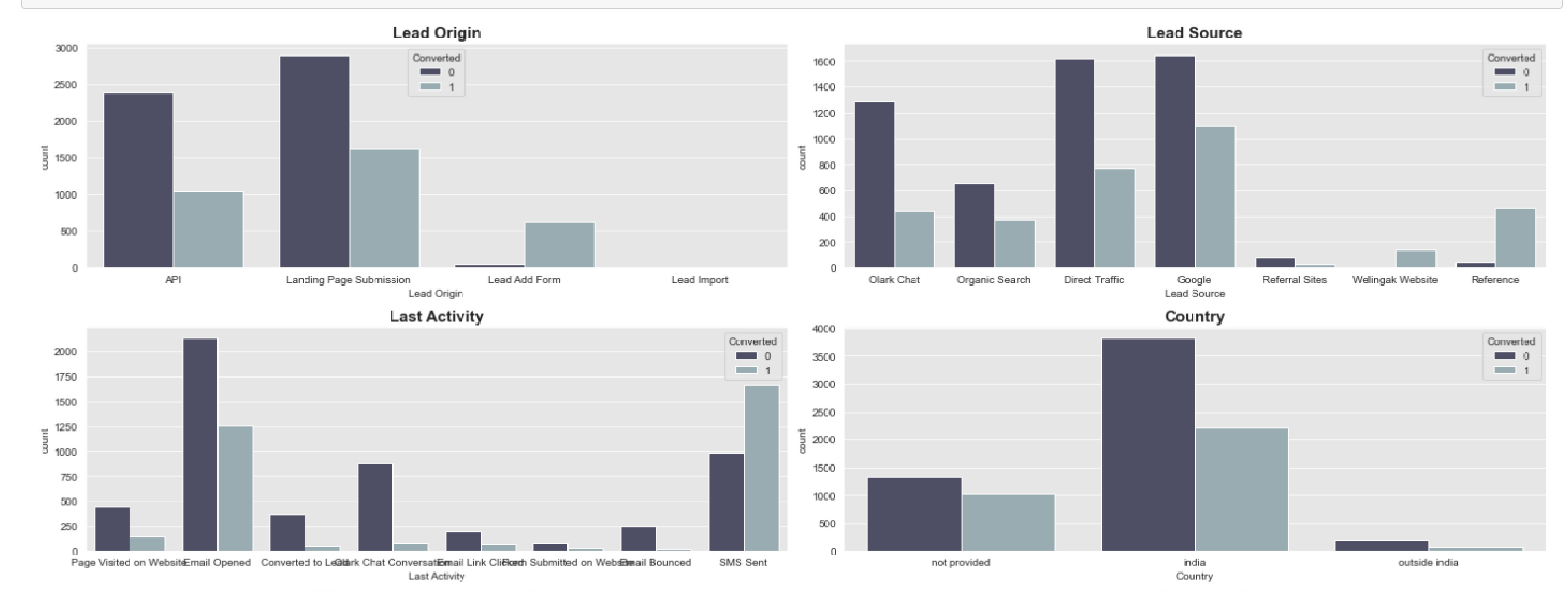




Distribution of Continuous data variables



BIVARIATE ANALYSIS WITH TARGET VARIABLE CONVERTED



Lead Origin

Lead Source

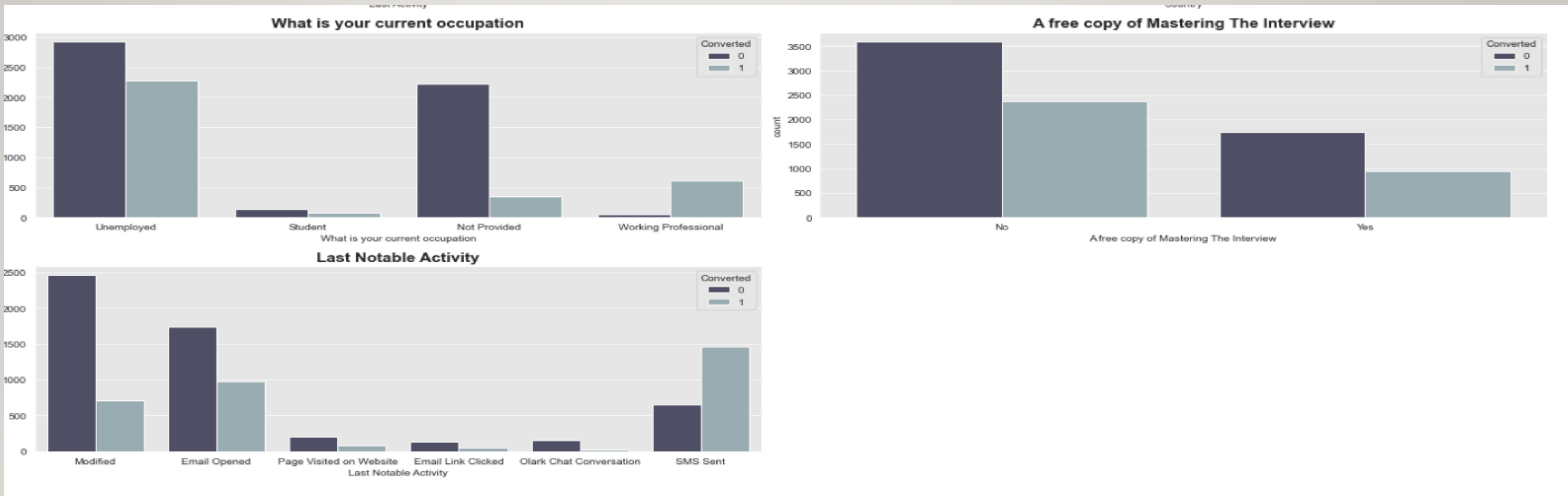
Last Activity

Country

Most of the Converted leads are from Lead Source Google ,Direct Traffic. And having Lead Origin as Landing Page Submission

Last Activity done by the Converted leads was mail opening, SMS Sent. Most of the Converted Leads are from India .

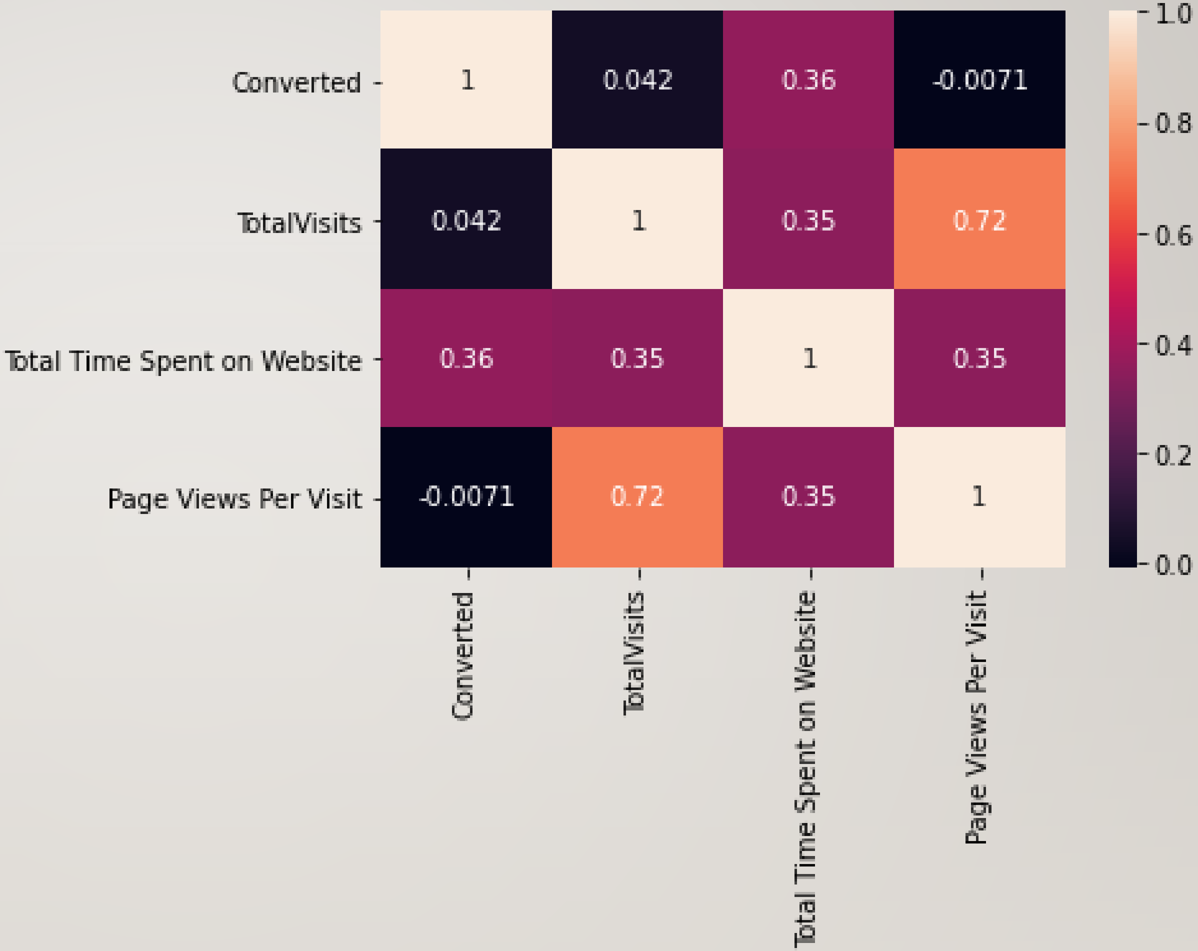
BIVARIATE ANALYSIS WITH TARGET VARIABLE CONVERTED (CONT...)



Most of the converted leads are unemployed and working professionals

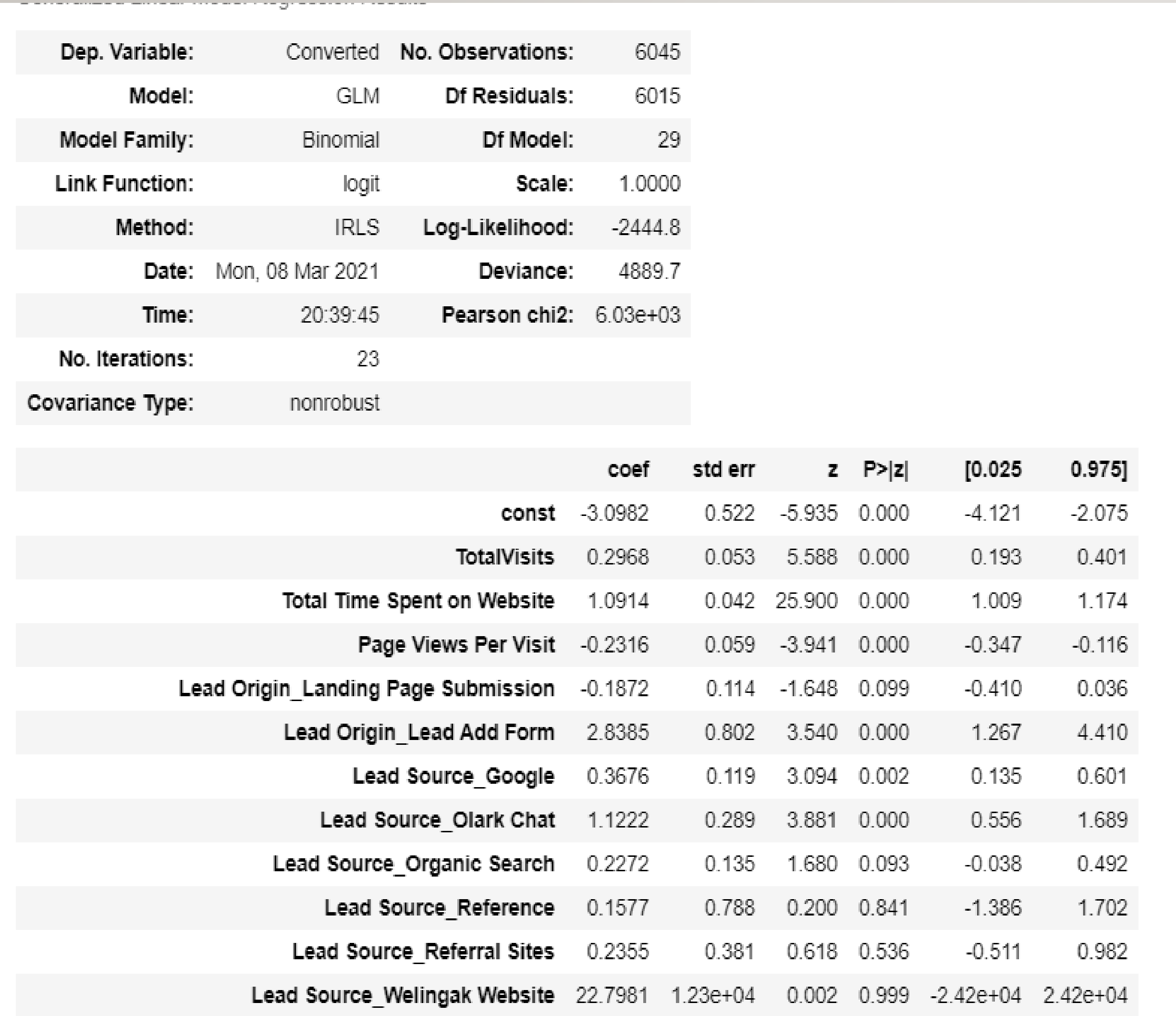
Last Notable Activity done by converted leads was mail opening and sending SMS.

HEAT MAP TO PREDICT CORRELATION B/W NUMERICAL VARIABLES



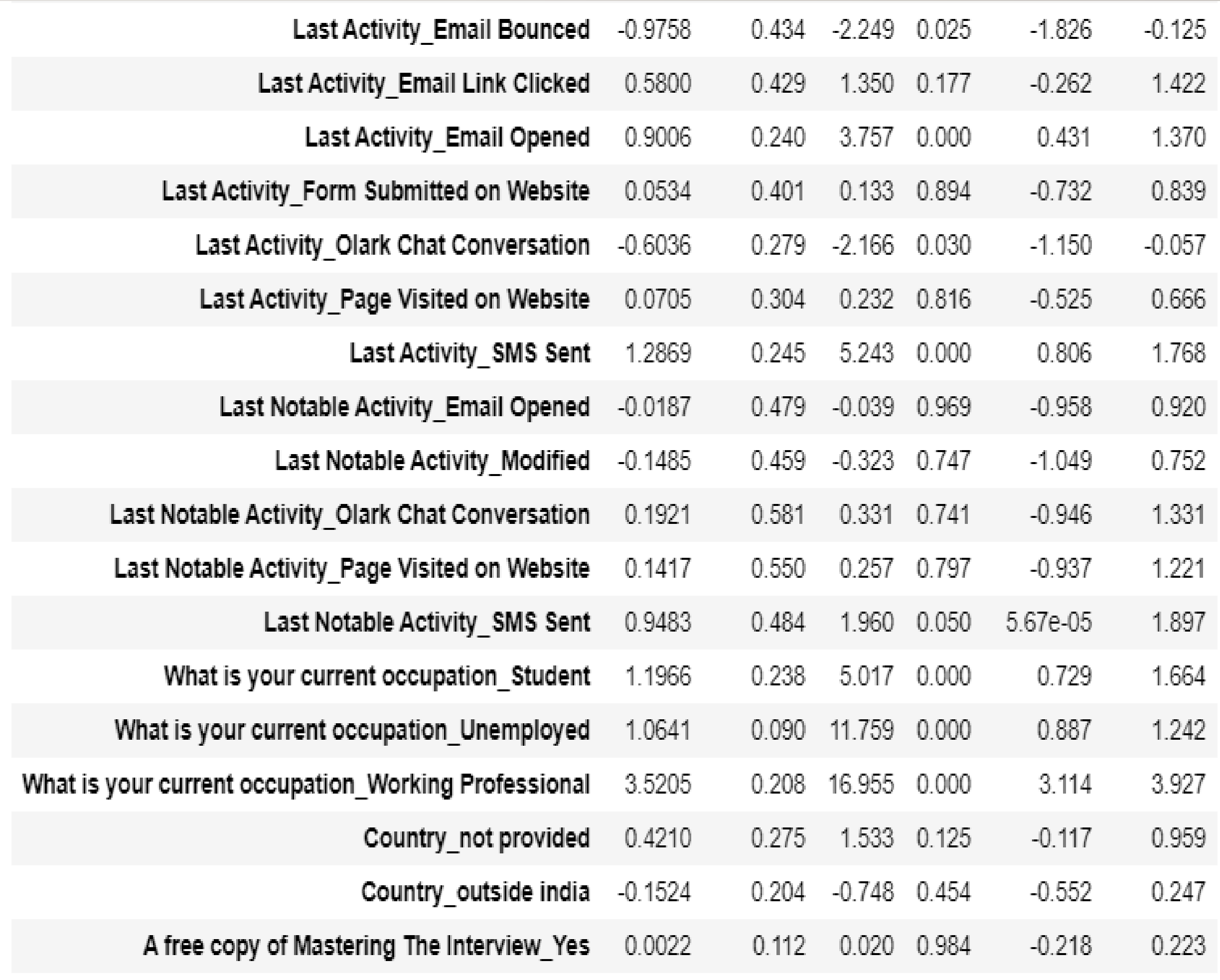
- ✓ From the heat map we can predict the Total Visits and Page Views per Visit are highly correlated.

Initial Model built with all the Variables after Data Cleansing



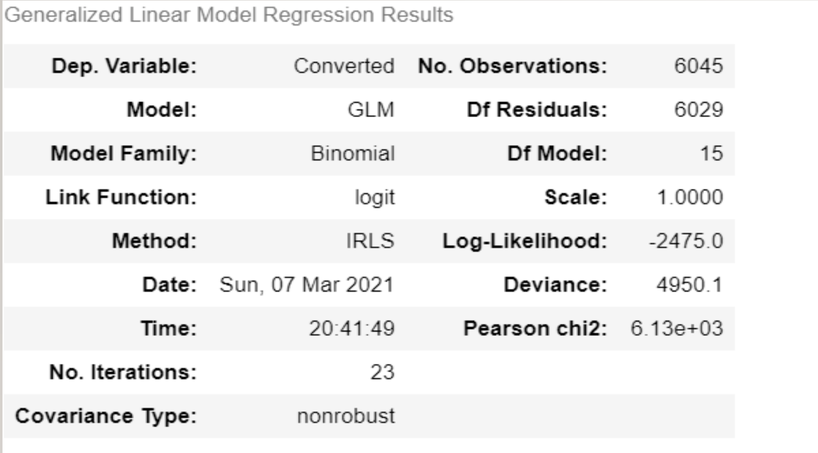
Dep. Variable:	Converted	**No. Observations:**	6045
Model:	GLM	**Df Residuals:**	6015
Model Family:	Binomial	**Df Model:**	29
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2444.8
Date:	Mon, 08 Mar 2021	**Deviance:**	4889.7
Time:	20:39:45	**Pearson chi2:**	6.03e+03
No. Iterations:	23		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.0982	0.522	-5.935	0.000	-4.121	-2.075
TotalVisits	0.2968	0.053	5.588	0.000	0.193	0.401
Total Time Spent on Website	1.0914	0.042	25.900	0.000	1.009	1.174
Page Views Per Visit	-0.2316	0.059	-3.941	0.000	-0.347	-0.116
Lead Origin_Landing Page Submission	-0.1872	0.114	-1.648	0.099	-0.410	0.036
Lead Origin_Lead Add Form	2.8385	0.802	3.540	0.000	1.267	4.410
Lead Source_Google	0.3676	0.119	3.094	0.002	0.135	0.601
Lead Source_Olark Chat	1.1222	0.289	3.881	0.000	0.556	1.689
Lead Source_Organic Search	0.2272	0.135	1.680	0.093	-0.038	0.492
Lead Source_Reference	0.1577	0.788	0.200	0.841	-1.386	1.702
Lead Source_Referral Sites	0.2355	0.381	0.618	0.536	-0.511	0.982
Lead Source_Welingak Website	22.7981	1.23e+04	0.002	0.999	-2.42e+04	2.42e+04



Last Activity_Email Bounced	-0.9758	0.434	-2.249	0.025	-1.826	-0.125
Last Activity_Email Link Clicked	0.5800	0.429	1.350	0.177	-0.262	1.422
Last Activity_Email Opened	0.9006	0.240	3.757	0.000	0.431	1.370
Last Activity_Form Submitted on Website	0.0534	0.401	0.133	0.894	-0.732	0.839
Last Activity_Olark Chat Conversation	-0.6036	0.279	-2.166	0.030	-1.150	-0.057
Last Activity_Page Visited on Website	0.0705	0.304	0.232	0.816	-0.525	0.666
Last Activity_SMS Sent	1.2869	0.245	5.243	0.000	0.806	1.768
Last Notable Activity_Email Opened	-0.0187	0.479	-0.039	0.969	-0.958	0.920
Last Notable Activity_Modified	-0.1485	0.459	-0.323	0.747	-1.049	0.752
Last Notable Activity_Olark Chat Conversation	0.1921	0.581	0.331	0.741	-0.946	1.331
Last Notable Activity_Page Visited on Website	0.1417	0.550	0.257	0.797	-0.937	1.221
Last Notable Activity_SMS Sent	0.9483	0.484	1.960	0.050	5.67e-05	1.897
What is your current occupation_Student	1.1966	0.238	5.017	0.000	0.729	1.664
What is your current occupation_Unemployed	1.0641	0.090	11.759	0.000	0.887	1.242
What is your current occupation_Working Professional	3.5205	0.208	16.955	0.000	3.114	3.927
Country_not provided	0.4210	0.275	1.533	0.125	-0.117	0.959
Country_outside india	-0.1524	0.204	-0.748	0.454	-0.552	0.247
A free copy of Mastering The Interview_Yes	0.0022	0.112	0.020	0.984	-0.218	0.223

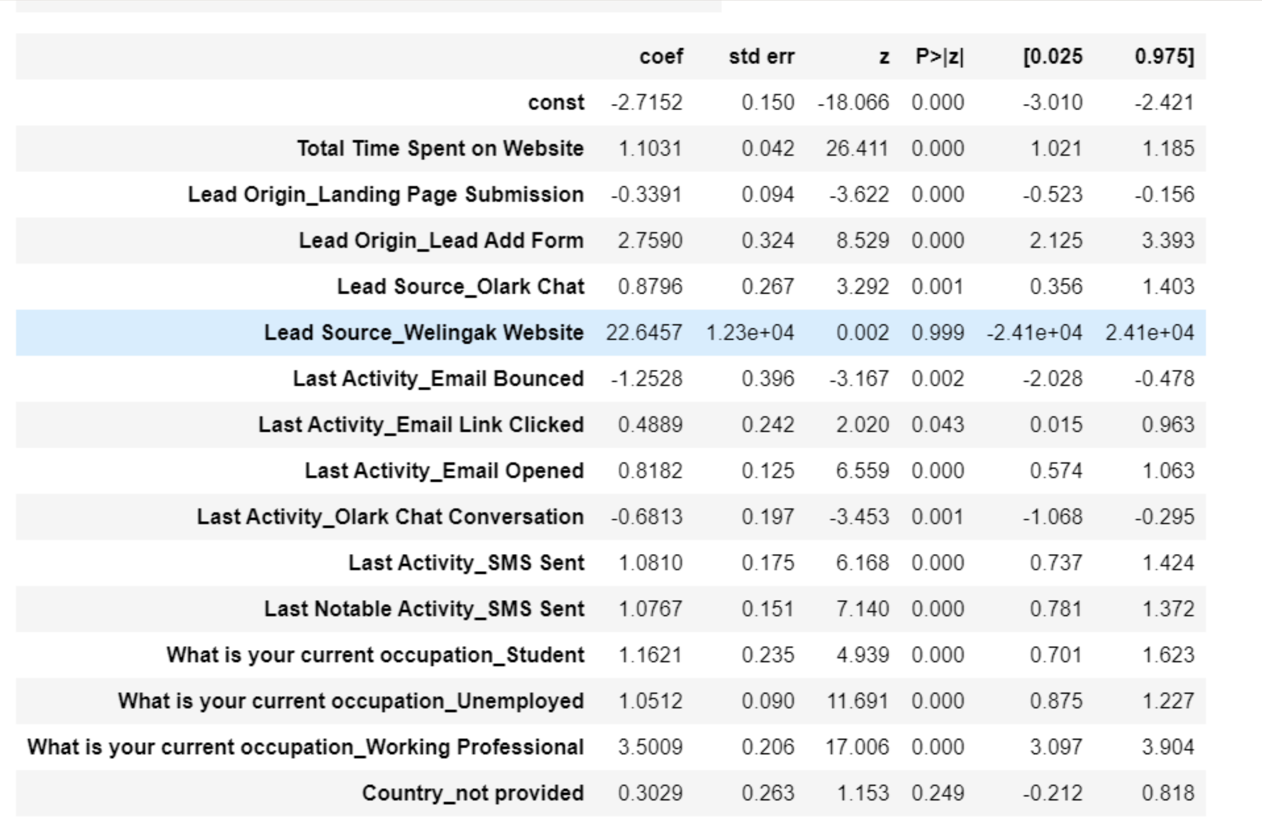
MODEL BUILDING (FEATURE SELECTION OF 15 FEATURES USING RFE)



Generalized Linear Model Regression Results

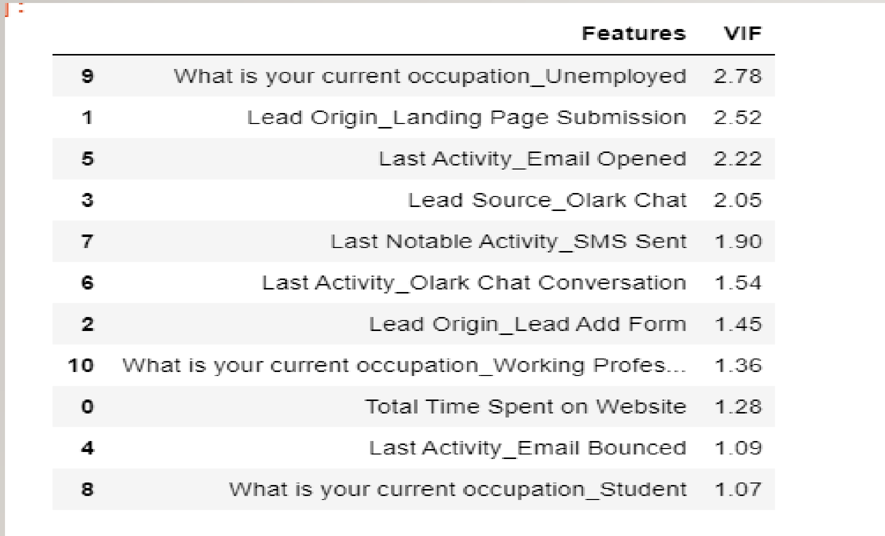
Dep. Variable:	Converted	**No. Observations:**	6045
Model:	GLM	**Df Residuals:**	6029
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2475.0
Date:	Sun, 07 Mar 2021	**Deviance:**	4950.1
Time:	20:41:49	**Pearson chi2:**	6.13e+03
No. Iterations:	23		
Covariance Type:	nonrobust		

- ➢ Model was built initially with all the remaining variables after data cleansing.
- ➢ After this the model was build by selecting top 15 features required to build the model

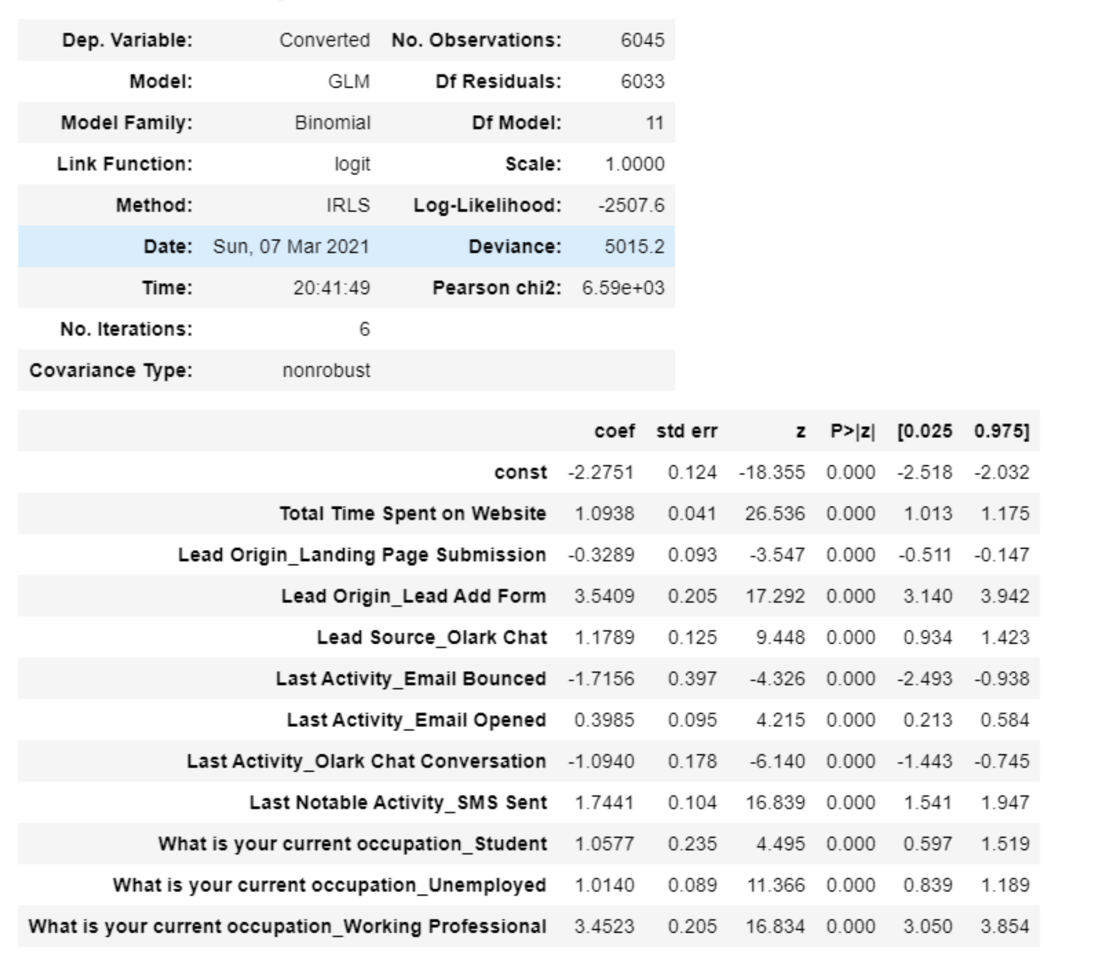


	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.7152	0.150	-18.066	0.000	-3.010	-2.421
Total Time Spent on Website	1.1031	0.042	26.411	0.000	1.021	1.185
Lead Origin_Landing Page Submission	-0.3391	0.094	-3.622	0.000	-0.523	-0.156
Lead Origin_Lead Add Form	2.7590	0.324	8.529	0.000	2.125	3.393
Lead Source_Olark Chat	0.8796	0.267	3.292	0.001	0.356	1.403
Lead Source_Welingak Website	22.6457	1.23e+04	0.002	0.999	-2.41e+04	2.41e+04
Last Activity_Email Bounced	-1.2528	0.396	-3.167	0.002	-2.028	-0.478
Last Activity_Email Link Clicked	0.4889	0.242	2.020	0.043	0.015	0.963
Last Activity_Email Opened	0.8182	0.125	6.559	0.000	0.574	1.063
Last Activity_Olark Chat Conversation	-0.6813	0.197	-3.453	0.001	-1.068	-0.295
Last Activity_SMS Sent	1.0810	0.175	6.168	0.000	0.737	1.424
Last Notable Activity_SMS Sent	1.0767	0.151	7.140	0.000	0.781	1.372
What is your current occupation_Student	1.1621	0.235	4.939	0.000	0.701	1.623
What is your current occupation_Unemployed	1.0512	0.090	11.691	0.000	0.875	1.227
What is your current occupation_Working Professional	3.5009	0.206	17.006	0.000	3.097	3.904
Country_not provided	0.3029	0.263	1.153	0.249	-0.212	0.818

REBUILDING THE MODEL ITERATIVELY TO RETAIN MOST USEFUL VARIABLES



	Features	VIF
9	What is your current occupation_Unemployed	2.78
1	Lead Origin_Landing Page Submission	2.52
5	Last Activity_Email Opened	2.22
3	Lead Source_Olark Chat	2.05
7	Last Notable Activity_SMS Sent	1.90
6	Last Activity_Olark Chat Conversation	1.54
2	Lead Origin_Lead Add Form	1.45
10	What is your current occupation_Working Profes...	1.36
0	Total Time Spent on Website	1.28
4	Last Activity_Email Bounced	1.09
8	What is your current occupation_Student	1.07



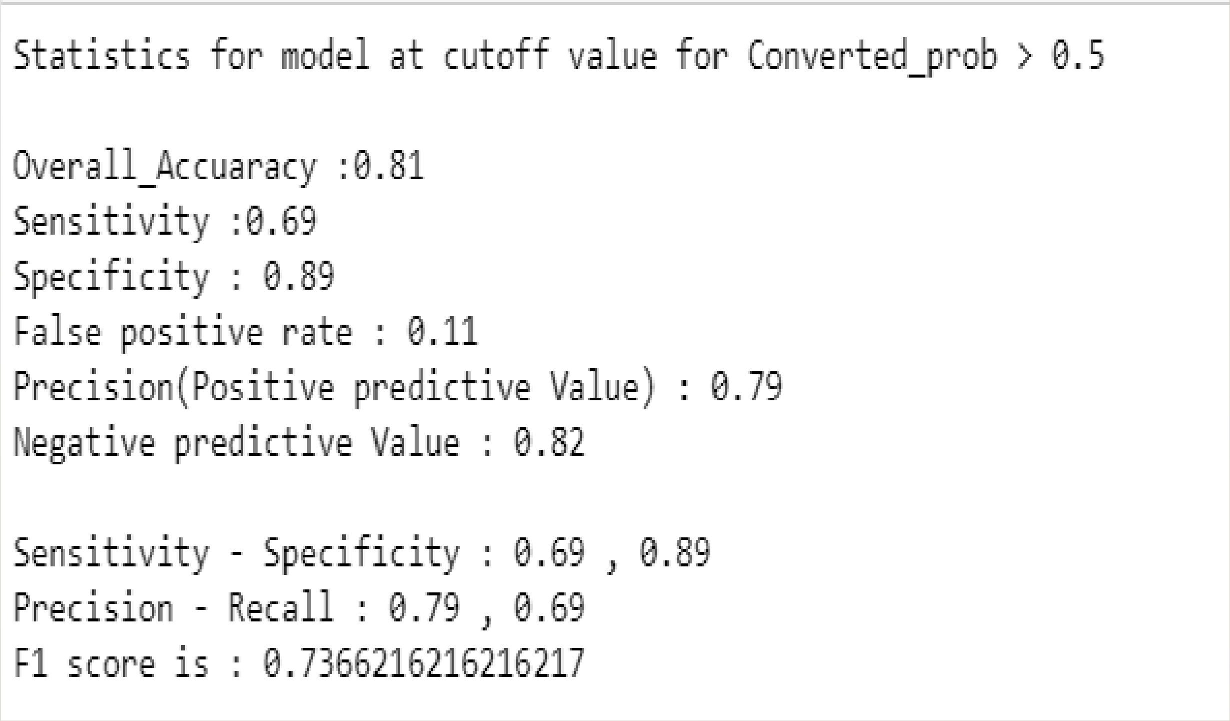
Dep. Variable:	Converted	**No. Observations:**	6045
Model:	GLM	**Df Residuals:**	6033
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2507.6
Date:	Sun, 07 Mar 2021	**Deviance:**	5015.2
Time:	20:41:49	**Pearson chi2:**	6.59e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.2751	0.124	-18.355	0.000	-2.518	-2.032
Total Time Spent on Website	1.0938	0.041	26.536	0.000	1.013	1.175
Lead Origin_Landing Page Submission	-0.3289	0.093	-3.547	0.000	-0.511	-0.147
Lead Origin_Lead Add Form	3.5409	0.205	17.292	0.000	3.140	3.942
Lead Source_Olark Chat	1.1789	0.125	9.448	0.000	0.934	1.423
Last Activity_Email Bounced	-1.7156	0.397	-4.326	0.000	-2.493	-0.938
Last Activity_Email Opened	0.3985	0.095	4.215	0.000	0.213	0.584
Last Activity_Olark Chat Conversation	-1.0940	0.178	-6.140	0.000	-1.443	-0.745
Last Notable Activity_SMS Sent	1.7441	0.104	16.839	0.000	1.541	1.947
What is your current occupation_Student	1.0577	0.235	4.495	0.000	0.597	1.519
What is your current occupation_Unemployed	1.0140	0.089	11.366	0.000	0.839	1.189
What is your current occupation_Working Professional	3.4523	0.205	16.834	0.000	3.050	3.854

- ✓ Model was built iteratively by dropping the variables in the below order :
- ✓ High p value (Rechecking the model again with the p values and VIF)
- ✓ Weighing both p values and VIF model is built again and again till the p values is below 0.05 and VIF factor is less than 5 for all columns.

CHECKING ACCURACY,SENSITIVITY,SPECIFICITY WHEN THE PROBABILITY THRESHOLD WAS TAKEN AS 0.5

- ✓ Model Accuracy came around 81 %
- ✓ Sensitivity and Specificity are around 69 and 89 percent respectively.
- ✓ We can infer here the Sensitivity is bit low for the model with threshold considered as 0.5



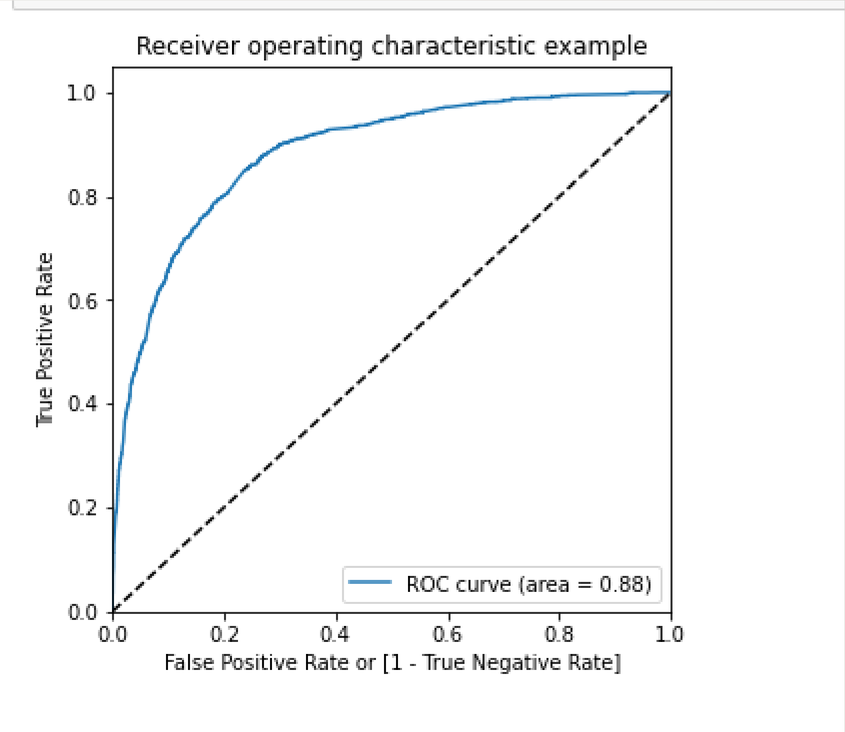
```
Statistics for model at cutoff value for Converted_prob > 0.5

Overall_Accuaracy :0.81
Sensitivity :0.69
Specificity : 0.89
False positive rate : 0.11
Precision(Positive predictive Value) : 0.79
Negative predictive Value : 0.82

Sensitivity - Specificity : 0.69 , 0.89
Precision - Recall : 0.79 , 0.69
F1 score is : 0.7366216216216217
```

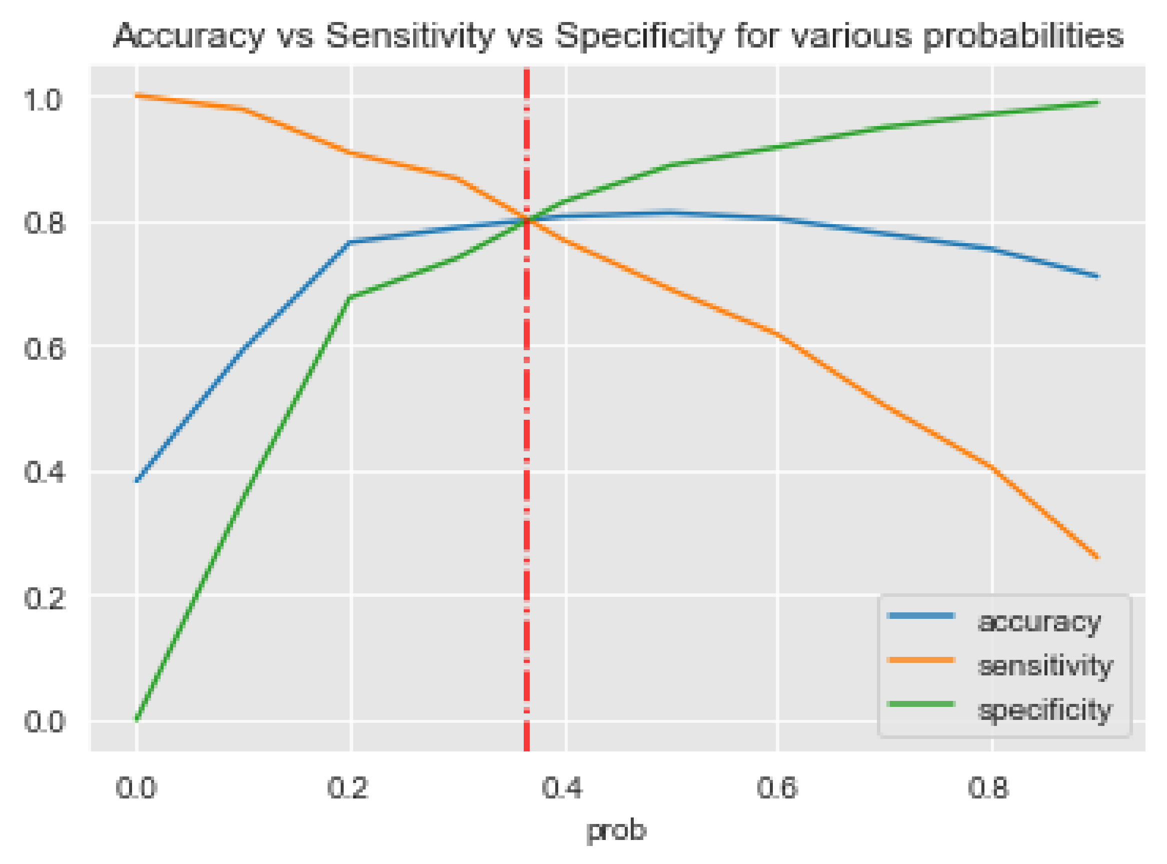
ROC CURVE

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



FINDING OPTIMAL CUF OFF PROBABILITY POINT

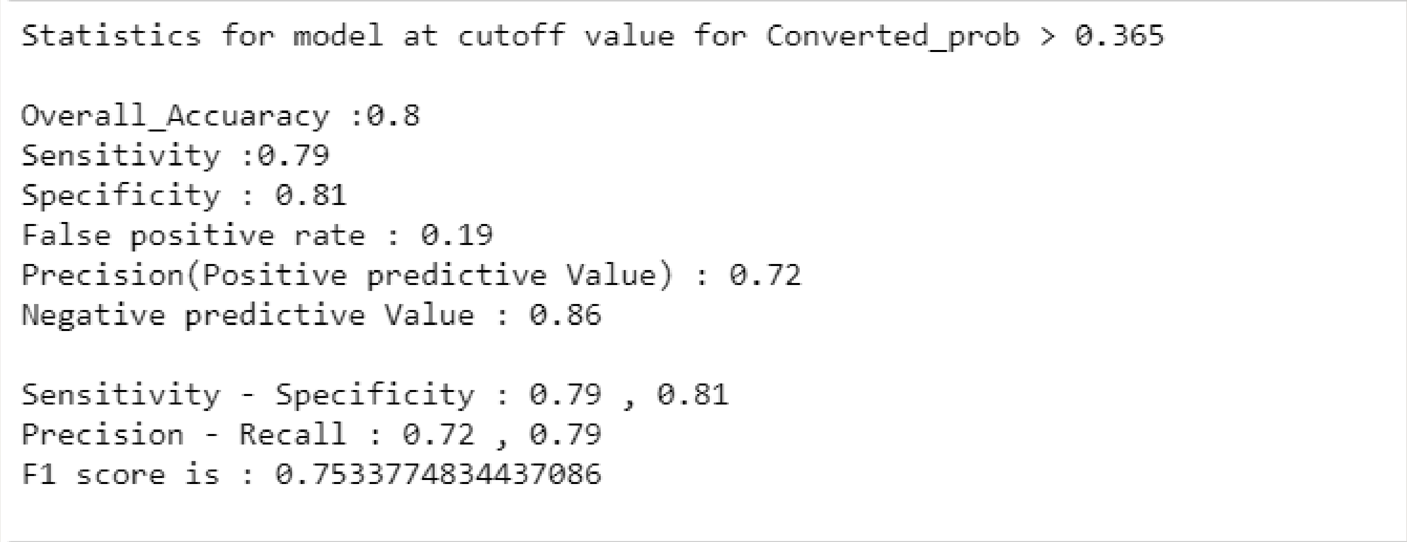
- ✓ From the Plot we were able to infer 0.365 as the optimal threshold



RECALCULATING MODEL ACCURACY,SENSITIVITY , SPECIFICITY WITH OPTIMAL THRESHOLD VALUE

- ✓ Final Predicted conversion rate is close ~= 80 %

- ✓ Model Accuracy is pretty good around 80 %
- ✓ Also the Sensitivity and Specificity are 79 % and 81 % respectively.
- ✓ So, we can live with this threshold value.



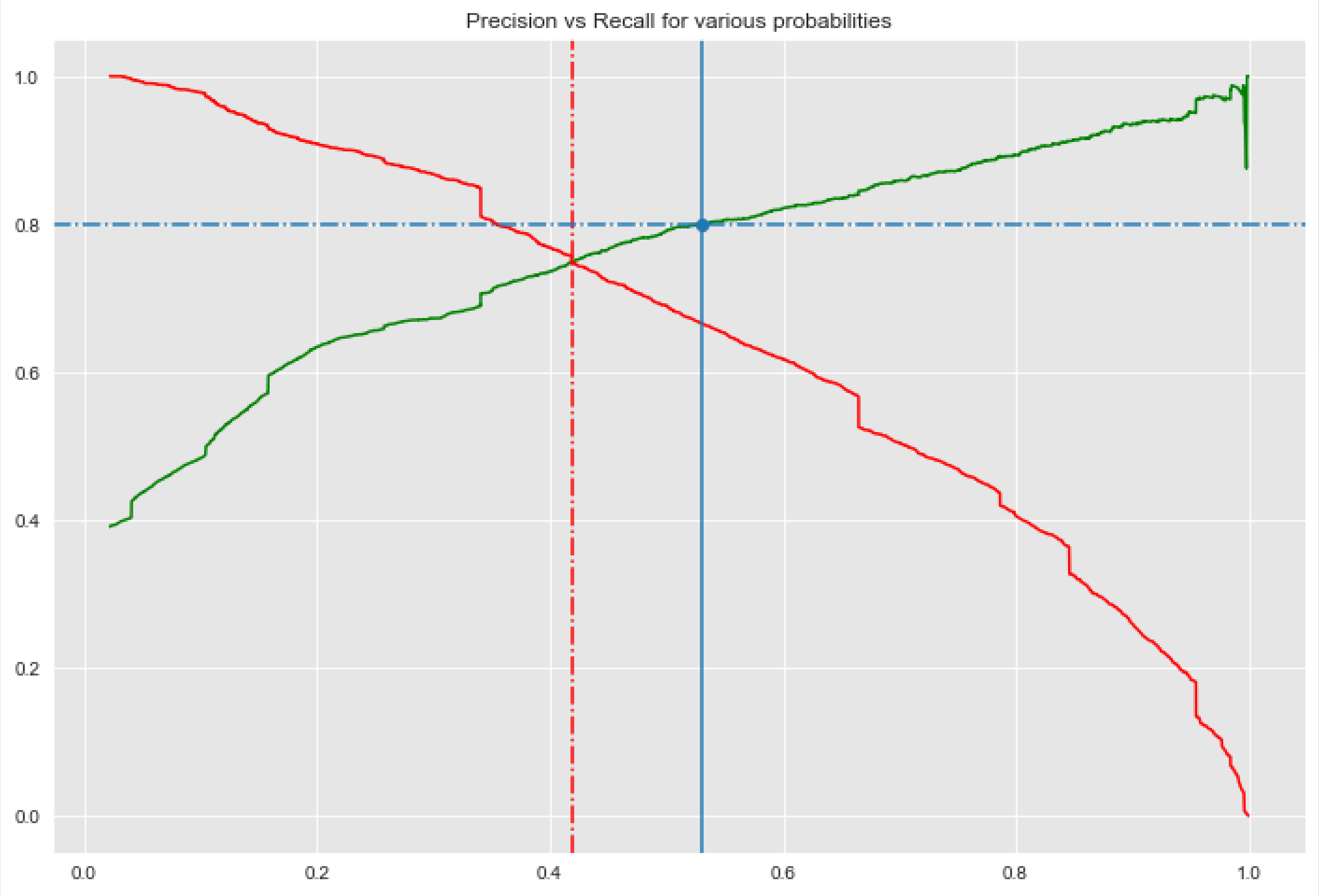
```
Statistics for model at cutoff value for Converted_prob > 0.365

Overall_Accuaracy :0.8
Sensitivity :0.79
Specificity : 0.81
False positive rate : 0.19
Precision(Positive predictive Value) : 0.72
Negative predictive Value : 0.86

Sensitivity - Specificity : 0.79 , 0.81
Precision - Recall : 0.72 , 0.79
F1 score is : 0.7533774834437086
```

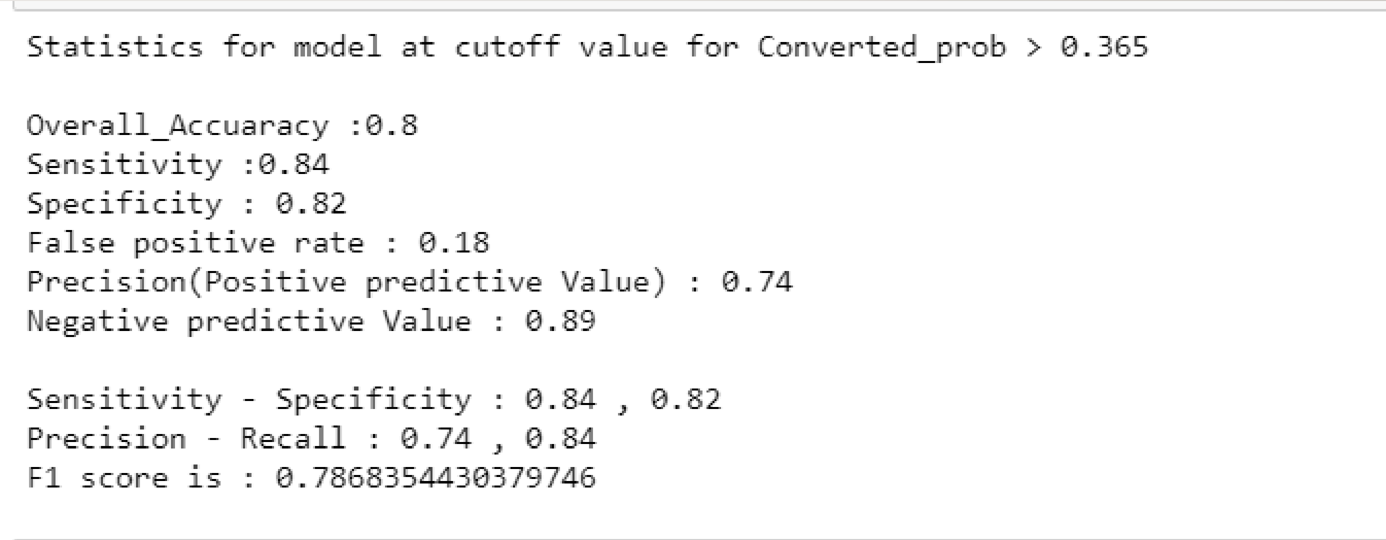
PRECISION & RECALL TRADE OFF

- ✓ Got decent Precision and Recall score as well with the considered threshold 0.42



MODEL EVALUATION ON TEST DATA .

- We can observe that Final predicted conversion is around 84 % (More than 80 % which is required)
- Accuracy , Sensitivity and specificity are around 80,84 and 82 % respectively which are pretty good.
- So, we can infer that model designed is working good on test data set as well

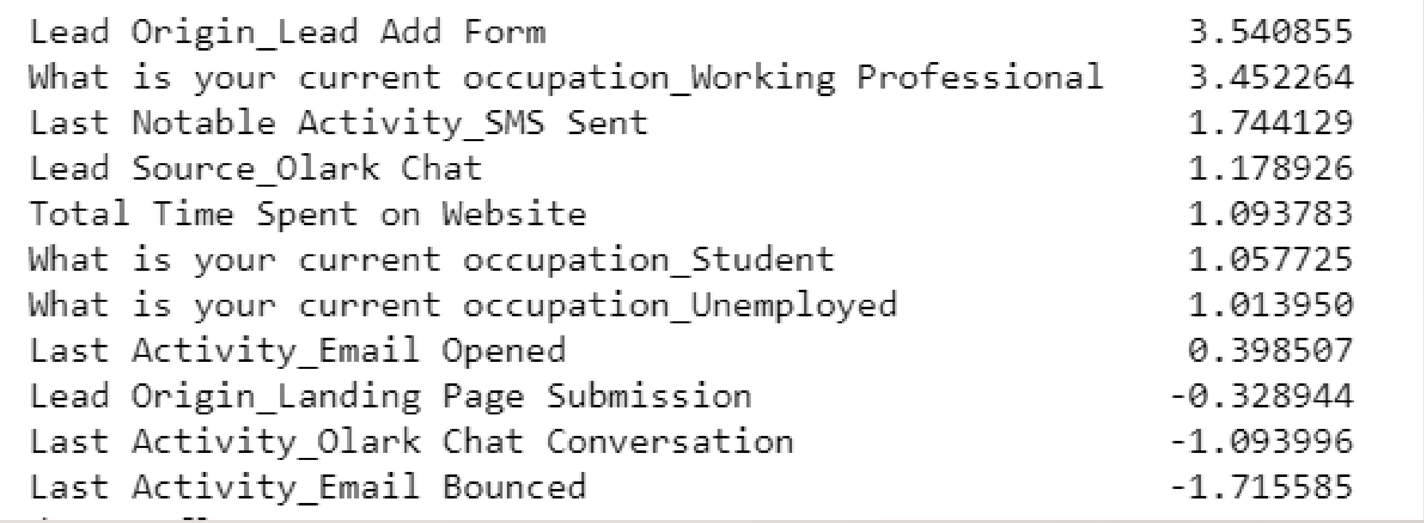


```
Statistics for model at cutoff value for Converted_prob > 0.365

Overall_Accuaracy :0.8
Sensitivity :0.84
Specificity : 0.82
False positive rate : 0.18
Precision(Positive predictive Value) : 0.74
Negative predictive Value : 0.89

Sensitivity - Specificity : 0.84 , 0.82
Precision - Recall : 0.74 , 0.84
F1 score is : 0.7868354430379746
```

MOST IMPORTANT PARAMETERS FOR LEAD CONVERSION



```
Lead Origin_Lead Add Form                               3.540855
What is your current occupation_Working Professional    3.452264
Last Notable Activity_SMS Sent                          1.744129
Lead Source_Olark Chat                                  1.178926
Total Time Spent on Website                             1.093783
What is your current occupation_Student                 1.057725
What is your current occupation_Unemployed              1.013950
Last Activity_Email Opened                              0.398507
Lead Origin_Landing Page Submission                    -0.328944
Last Activity_Olark Chat Conversation                  -1.093996
Last Activity_Email Bounced                            -1.715585
```

Category Variables Lead Origin, Current Occupation , Lead Sources and Last Notable Activity are most important variables affecting the model.

Particularly Working professionals are having more converted leads.

CONCLUSION

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- - Accuracy, Sensitivity and Specificity values of test set are around 83%, 83% and 82% which are approximately closer to the respective values calculated using trained set.
- - Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 79% and test data set is around 83 % which is close to the expected rate by CEO (80 %)
- The most important Categorical Variables Affecting the lead conversion is : Lead Origin , Lead Source, Current Occupation and last notable Activity.
- In the Numerical Variables the Total Time spent on website is the most variable affecting the lead conversion.

THANK YOU