LEAD SCORING CASE STUDY

PREDICTION OF LEAD CONVERSION FOR X-EDUCATION

By Aishwarya Singh

PROBLEM STATEMENT

Case study is designed to find patterns or driving factors that dictate lead conversion probability

The variables that are chosen by the end of the study should prove to be strong indicators of default.

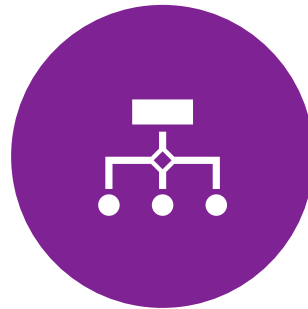This will ensure that the leads being chased have a good chance of being converted

This would at the end help X-education efficiently identify hot leads.

APPROACH TOWARDS SOLVING THE PROBLEM

IN THE COMING SLIDES WE WILL BE FOCUSING ON THE DIFFERENT ANALYSES PERFORMED ON THE GIVEN 'LEADS' DATASET.

THE PRE-PROCESSING AND CLEANING OF THE DATA HAS BEEN DONE PRIOR TO MOVING ON TO THE ANALYSIS PART.

INFERENCES DRAWN ON THE DIFFERENT ANALYSES WILL BE ADDED AS COMMENTS/NOTES.

FINAL MODEL WILL BE ANALYZED BY METRICS SUCH AS ROC, ACCURACY, SPECIFICITY, SENSITIVITY, PRECISION AND RECALL SCORES
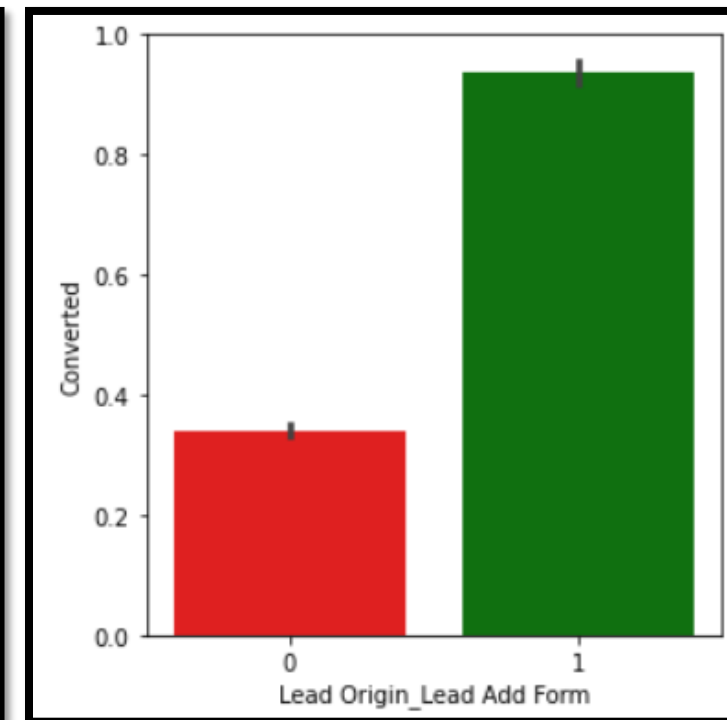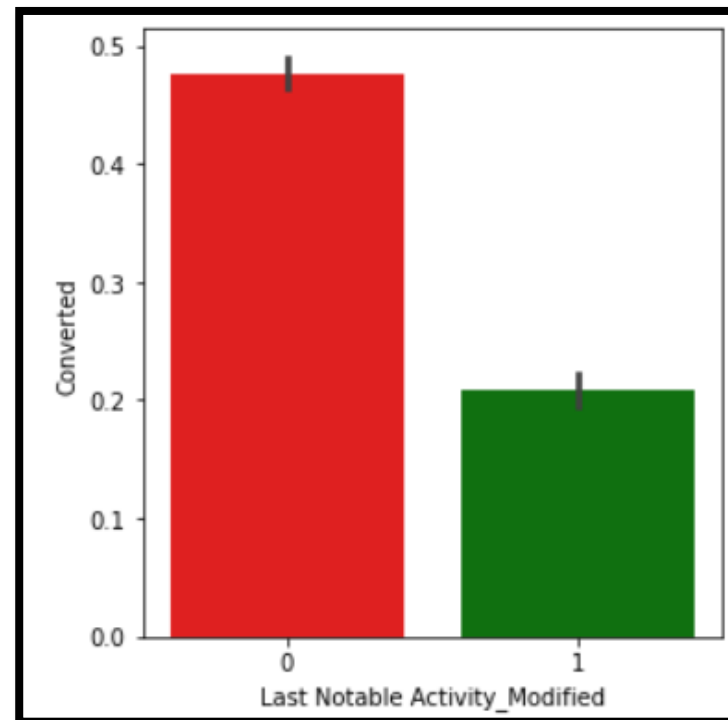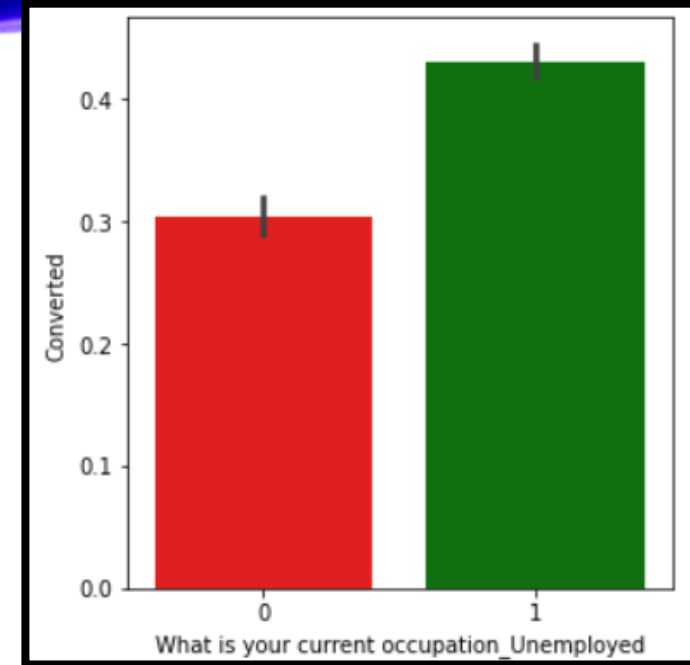
TOP FEATURES ANALYSIS
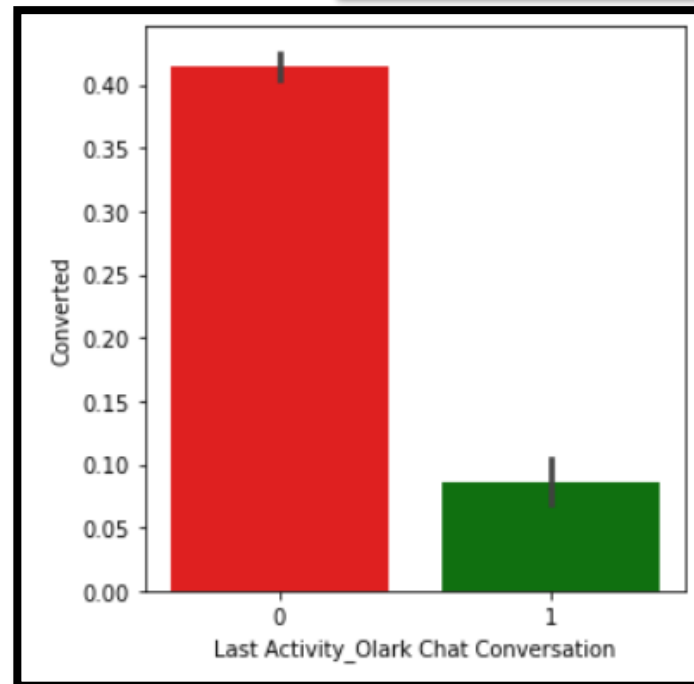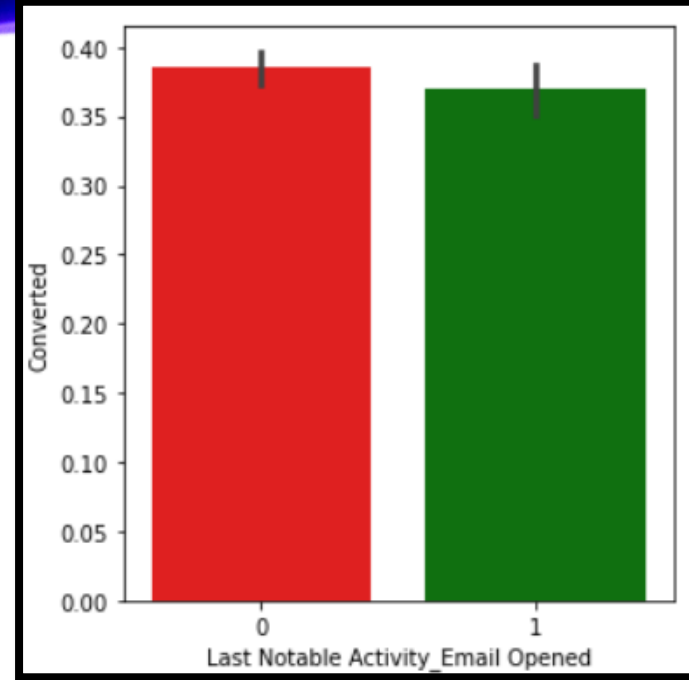## FOR CONVERTED (1) AND NOT-CONVERTED (0)

BAR PLOTS FOR THE VARIABLES CONSIDERED FOR MAKING OUR MODEL

- Bar plots for variables:
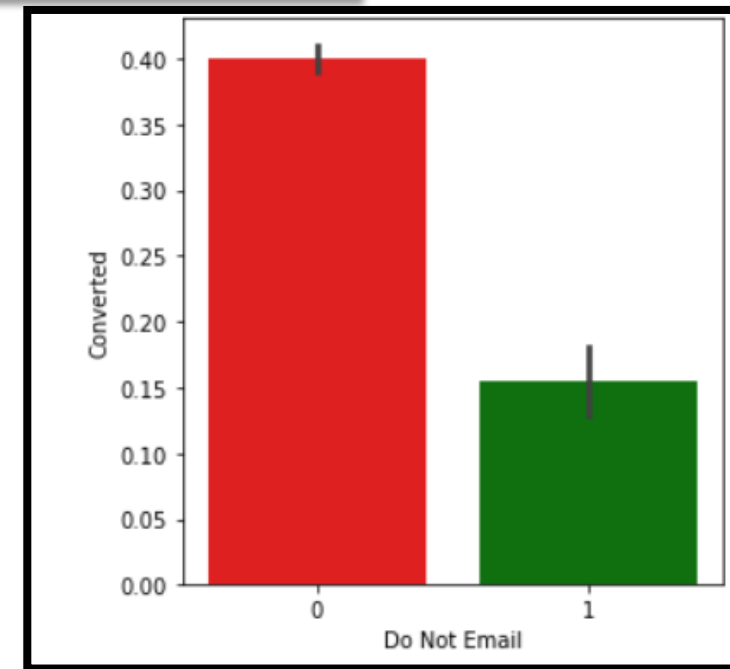 - What is your current occupation_Unemployed
  - Last Notable Activity_Modified
  - Lead Origin_Lead Add Form

- Bar plots for variables:
  - Last Notable Activity_Email Opened
  - Last Activity_Olark Chat Conversation
  - Lead Source_Welingak Website
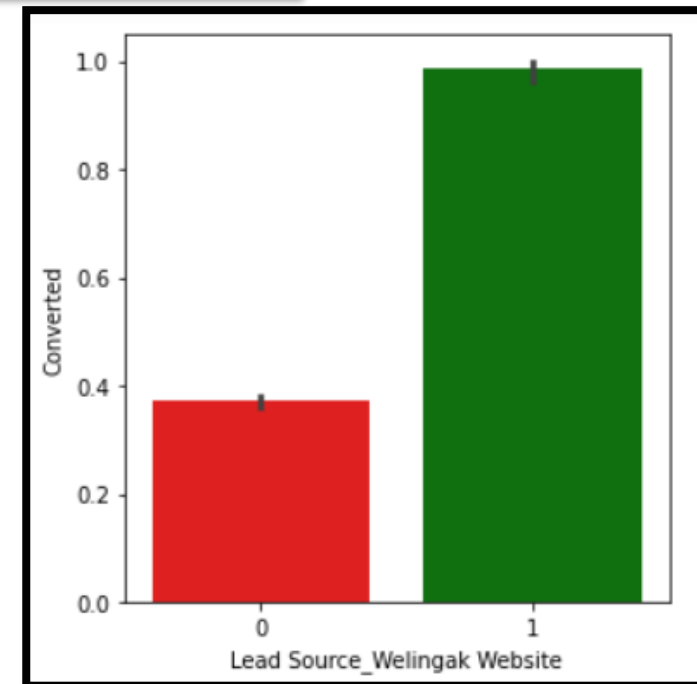
- Bar plots for variables:
  - Last Activity_Converted to Lead
  - What is your current occupation_Working Professional
  - Do Not Email
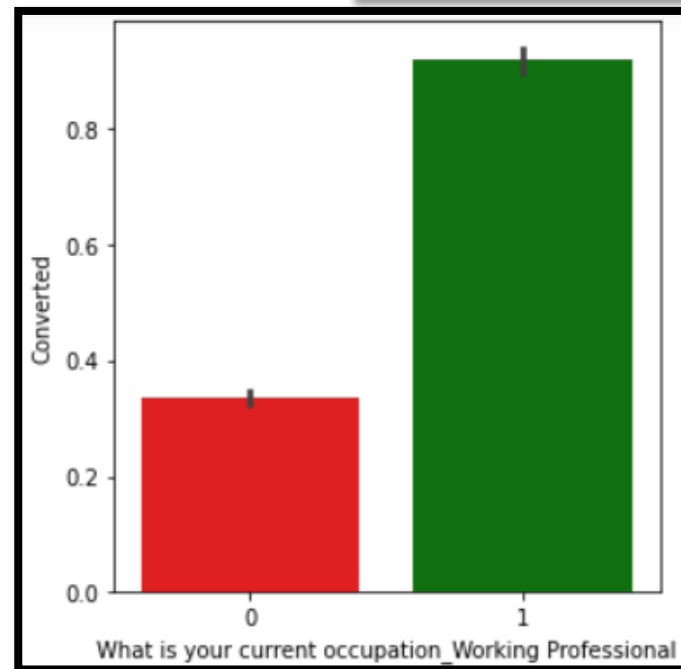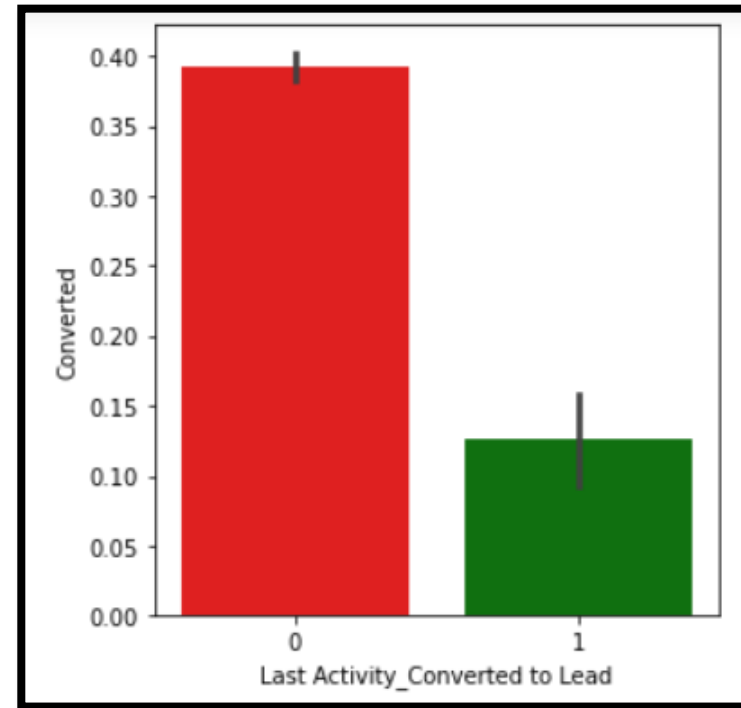
- Bar plots for variables:
  - Last Notable Activity_Page Visited on Website
  - Last Notable Activity_Email Link Clicked
  - What is your current occupation_Student
  - What is your current occupation_Other
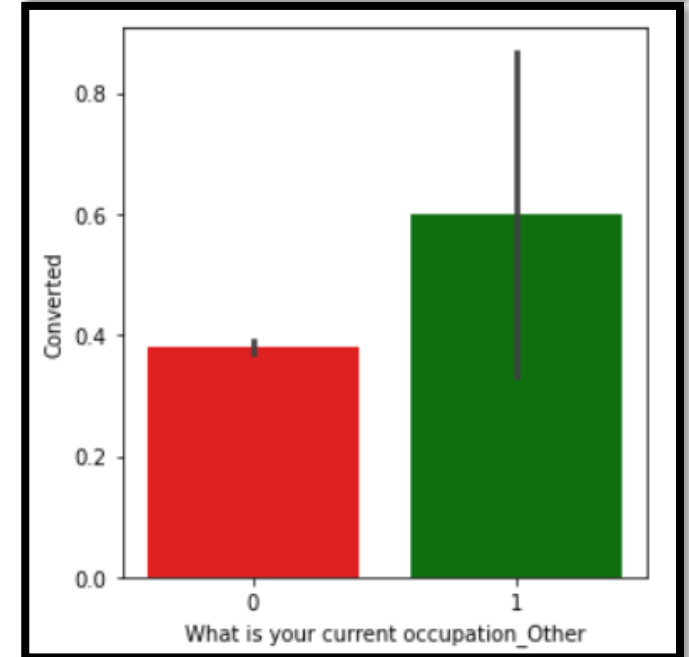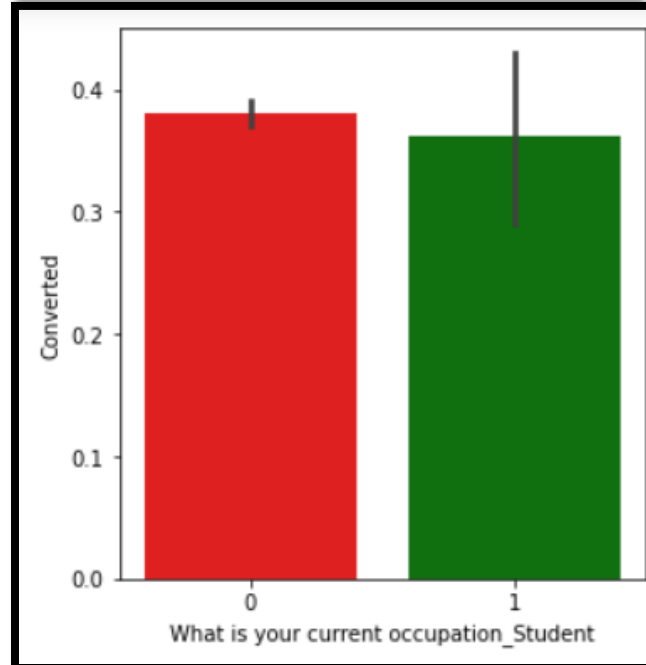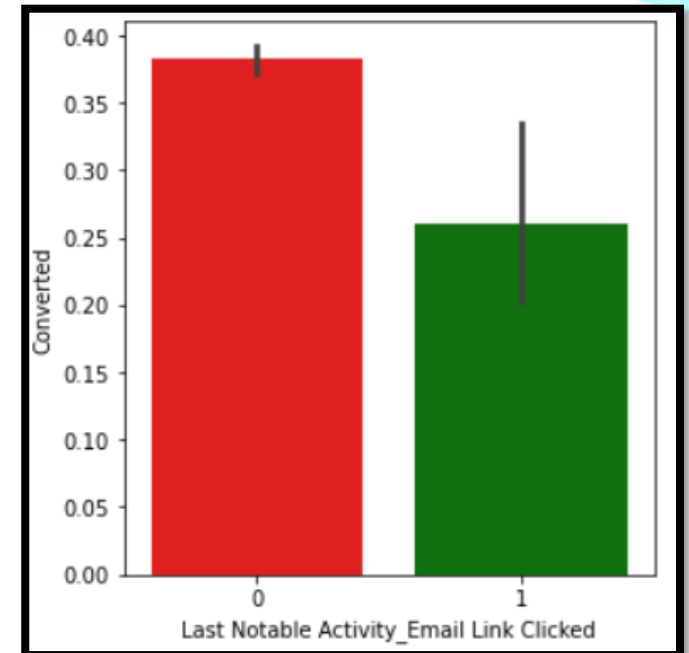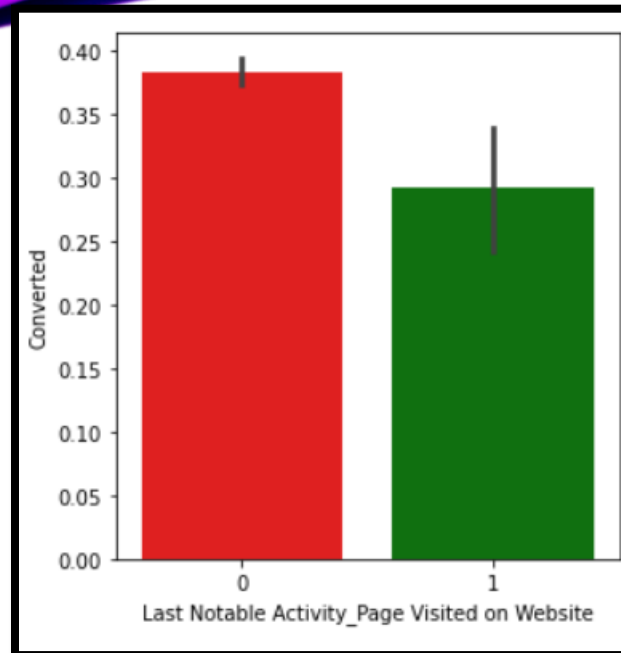
DEDUCTION FROM THE BAR PLOTS

- From the bar plot results, we can infer that, the following variables are worth exploring for getting hot leads:
  - 'What is your current occupation_Unemployed'
  - 'Lead origin_Lead Add Form'
  - 'Lead source_Welingak Website'
  - 'What is your current occupation_Working Professional'
  - 'What is your current occupation_Other'

CORRELATION ANALYSES
LEADS DATASET
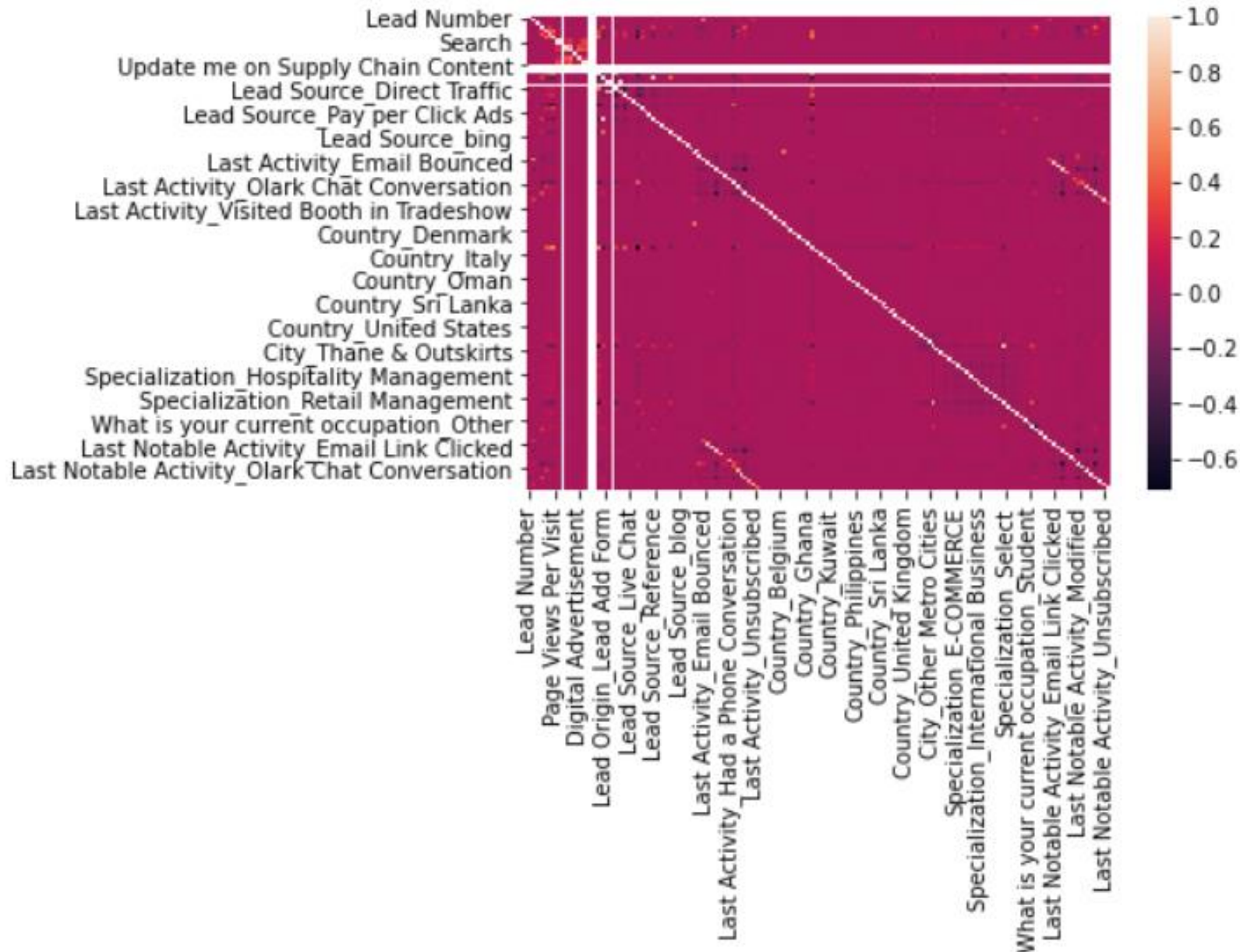
CORRELATION FOR THE VARIABLES OBTAINED AFTER CLEANING THE DATA

The correlation score between different variables from our cleaned dataset

STATISTICS AND ANALYSIS FOR THE CREATED MODEL

LOGISTIC REGRESSION MODEL STATISTICS AND IMPACTFUL FEATURES (VIF)

Dep. Variable:	Converted	R-squared:	0.312
Model:	OLS	Adj. R-squared:	0.310
Method:	Least Squares	F-statistic:	253.1
Date:	Wed, 09 Feb 2022	Prob (F-statistic):	0.00
Time:	16:11:34	Log-Likelihood:	-3705.0
No. Observations:	7282	AIC:	7438.
Df Residuals:	7268	BIC:	7534.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.4339	0.013	32.904	0.000	0.408	0.460
Do Not Email	-0.2457	0.018	-13.560	0.000	-0.281	-0.210
Lead Origin_Lead Add Form	0.3460	0.022	15.716	0.000	0.303	0.389
Lead Source_Welingak Website	0.1767	0.045	3.939	0.000	0.089	0.265
Last Activity_Converted to Lead	-0.1853	0.024	-7.637	0.000	-0.233	-0.138
Last Activity_Olark Chat Conversation	-0.2128	0.017	-12.615	0.000	-0.246	-0.180
What is your current occupation_Other	0.4235	0.108	3.918	0.000	0.212	0.635
What is your current occupation_Student	0.1668	0.034	4.968	0.000	0.101	0.233
What is your current occupation_Unemployed	0.1810	0.011	16.412	0.000	0.159	0.203
What is your current occupation_Working Professional	0.5568	0.020	27.181	0.000	0.517	0.597
Last Notable Activity_Email Link Clicked	-0.3281	0.037	-8.920	0.000	-0.400	-0.256
Last Notable Activity_Email Opened	-0.2452	0.013	-19.248	0.000	-0.270	-0.220
Last Notable Activity_Modified	-0.2612	0.013	-20.099	0.000	-0.287	-0.236
Last Notable Activity_Page Visited on Website	-0.2719	0.027	-10.087	0.000	-0.325	-0.219

Omnibus:	891.154	Durbin-Watson:	2.016
Prob(Omnibus):	0.000	Jarque-Bera (JB):	464.535
Skew:	0.467	Prob(JB):	1.34e-101
Kurtosis:	2.188	Cond. No.	29.8

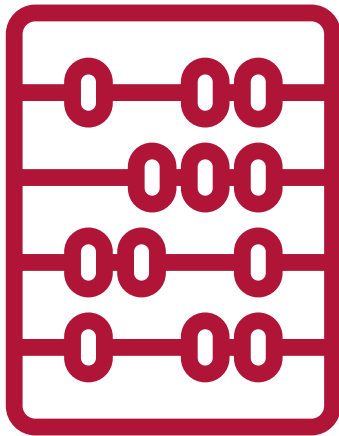|	Features	VIF
11	Last Notable Activity_Modified	1.91
7	What is your current occupation_Unemployed	1.83
1	Lead Origin_Lead Add Form	1.44
10	Last Notable Activity_Email Opened	1.42
4	Last Activity_Olark Chat Conversation	1.31
2	Lead Source_Welingak Website	1.30
3	Last Activity_Converted to Lead	1.22
8	What is your current occupation_Working Profes...	1.14
0	Do Not Email	1.13
12	Last Notable Activity_Page Visited on Website	1.05
6	What is your current occupation_Student	1.04
9	Last Notable Activity_Email Link Clicked	1.02
5	What is your current occupation_Other	1.01

Here we can see the top features according to their VIF values which was calculated for the logistic regression model

CORRELATION MATRIX FOR THE FEATURES FOR OUR MODEL

- The variables that were chosen for our model are depicted here in a correlation matrix
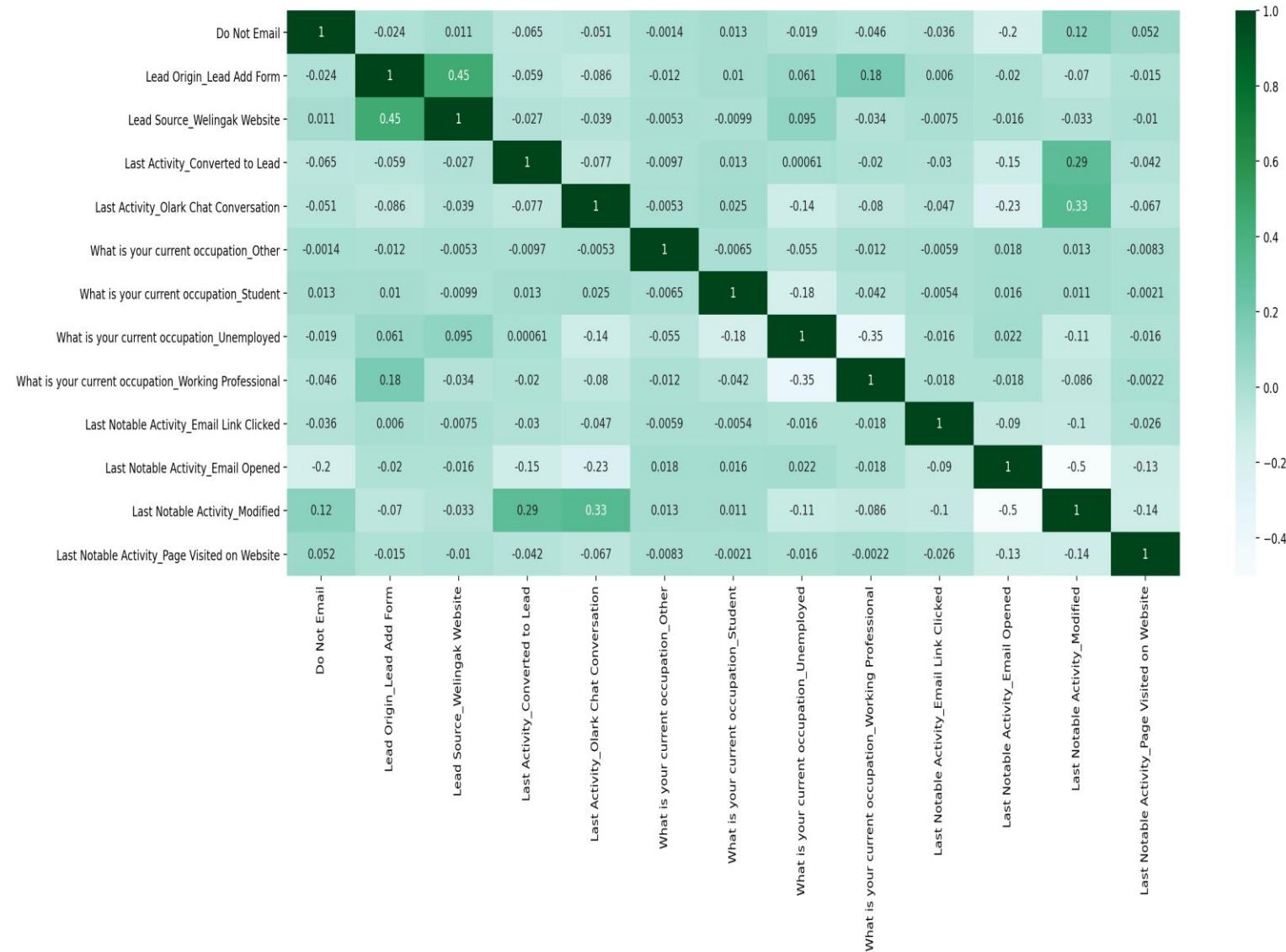
ROC CURVE AND COMPARISON BETWEEN SENSITIVITY, SPECIFICITY, AND ACCURACY
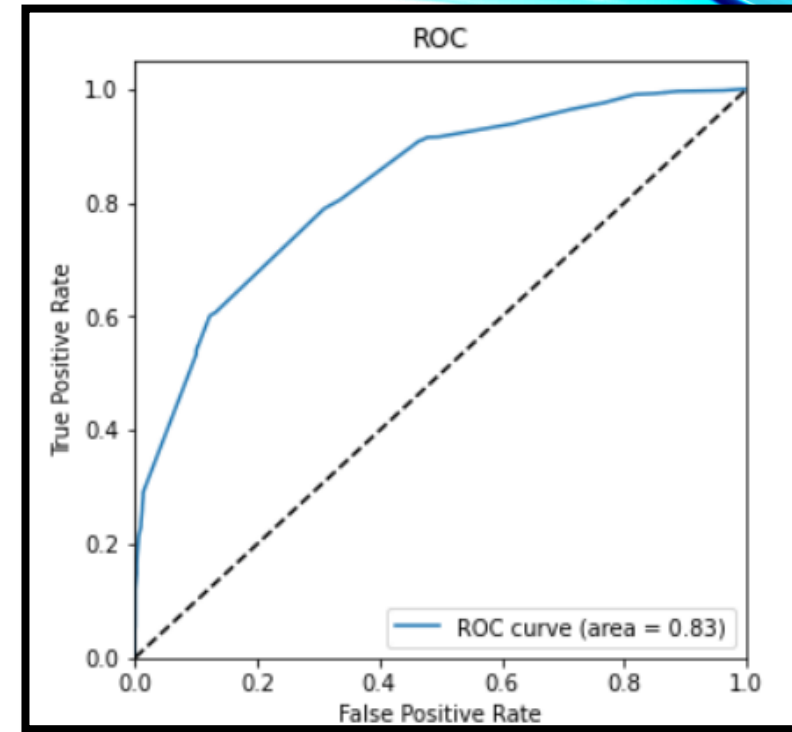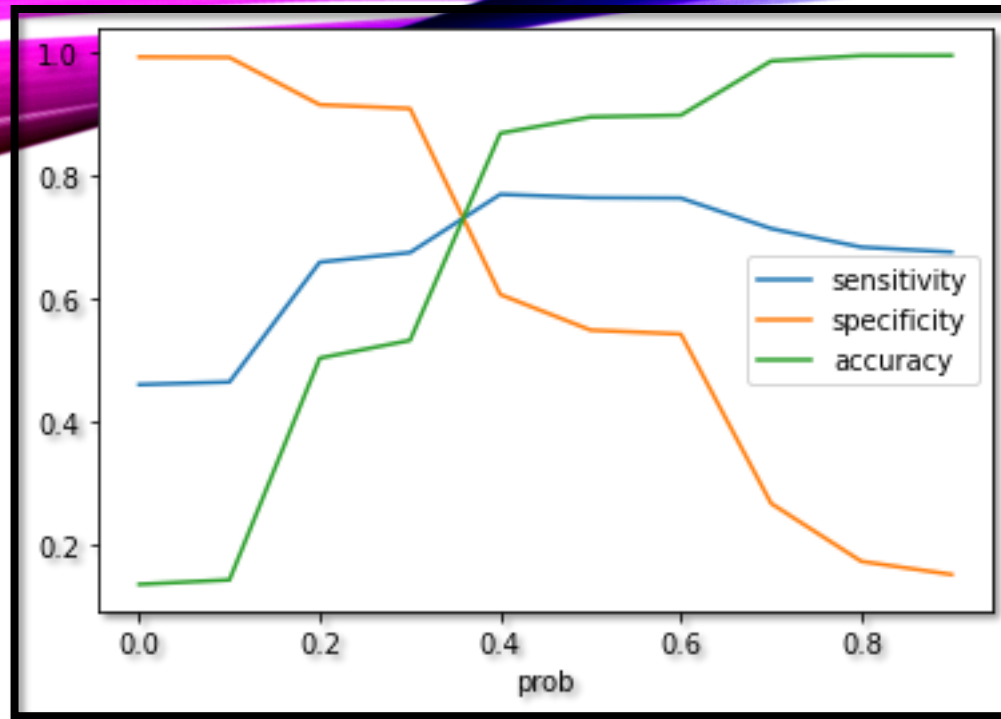
The area under the ROC curve is decent enough but when we look at the specificity, sensitivity and accuracy, the arbitrary cutoff has to be changed from 0.5 to 0.37.

ROC CURVE WHEN CUTOFF = 0.37

- Since we had deduced that the optimal cutoff is closer to 0.37 rather than 0.5

- Therefore, this is the new ROC curve

- Here AUC is 0.84, which is only slightly better than before
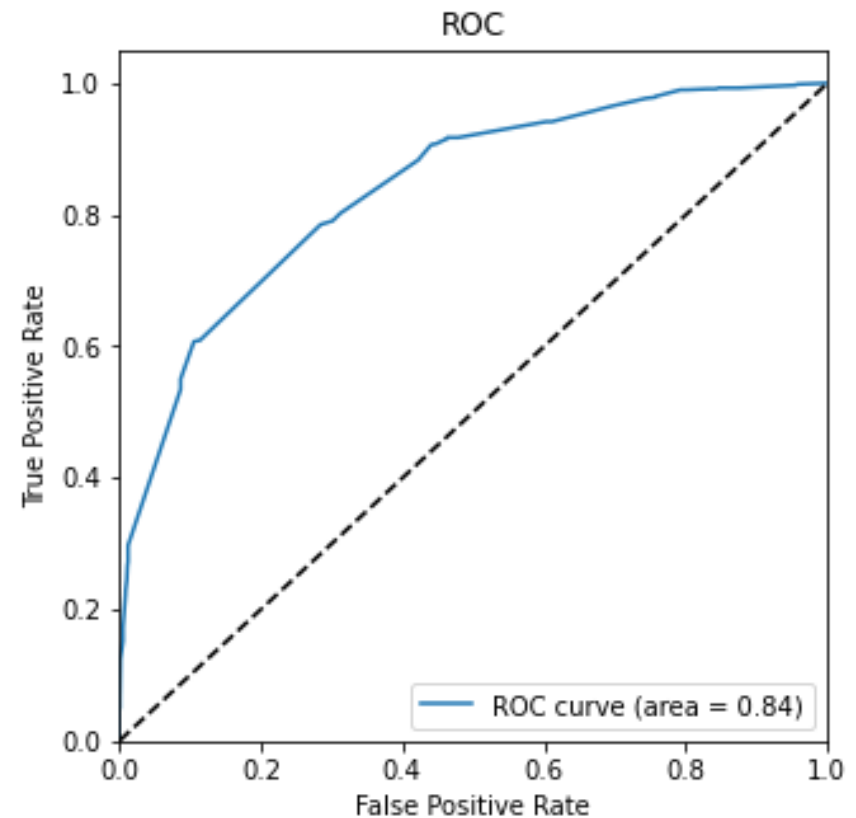
FINAL RECOMMENDATIONS

- Here we can see the absolute coefficient values for all 13 of our variables
- Looking at the values, it can be said that the top three variables in the model which contribute most towards the probability of a lead getting converted are:
  - a. 'What is your current occupation_Other'
  - b. 'What is your current occupation_Working Professional'
  - c. 'Lead Origin_Lead Add Form'
- So, the above-mentioned variables should be focused upon

	absolute_coefficients
What is your current occupation_Other	0.64
What is your current occupation_Working Professional	0.60
const	0.46
Lead Origin_Lead Add Form	0.39
Lead Source_Welingak Website	0.26
Last Notable Activity_Email Link Clicked	0.26
Last Notable Activity_Modified	0.24
What is your current occupation_Student	0.23
Last Notable Activity_Email Opened	0.22
Last Notable Activity_Page Visited on Website	0.22
Do Not Email	0.21
What is your current occupation_Unemployed	0.20
Last Activity_Olark Chat Conversation	0.18
Last Activity_Converted to Lead	0.14

CONCLUSION

- Our selected model was improved and made according to these factors:
    - Variables had p-values <0.05
    - VIF values are low to avoid multicollinearity

- Recommendations:
    - That ideal hot leads should have maximum values for the above mentioned three features. So, such leads should be focused on as much as possible.
        - This can be done by informing them about new offers, newly added course, course application deadlines, course application status, job offers, and so on.
    - The information provided to the leads would have to be designed according to the lead being followed, and for doing so the necessary data can be collected when they are being reached out to.
    - Thereon, the leads should be monitored closely for the best possible outcome.

THANK YOU