LEADING SCORE CASE STUDY
IDENTIFYING HOT LEADS

Maneesh Dhyani

Problem Statement

- 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.

Business Objective

- X education wants to know most promising leads
- For that they want to build a Model which identifies the hot leads.
- Help the sales team to divert their focus on potential leads & avoid them from making useless phone calls.
- **To build a model wherein a lead score is assigned** 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
- Deployment of the model for the future use.

Solution Methodology

1. Data cleaning and data manipulation.

   - Check and handle duplicate data.
   - Check and handle NA values and missing values.
   - Drop columns, if it contains large amount of missing values and not useful for the analysis.
   - Imputation of the values, if necessary.
   - Check and handle outliers in data.

2. EDA
   - Univariate data analysis: value count, distribution of variable etc.
   - Bivariate data analysis: correlation coefficients and pattern between the variables etc.

3. Feature Scaling & Dummy Variables and encoding of the data

4. Classification technique: logistic regression used for the model making and prediction

5. Validation of the model.

6. Model Presentation

7. Conclusions and recommendations.

Data Manipulation

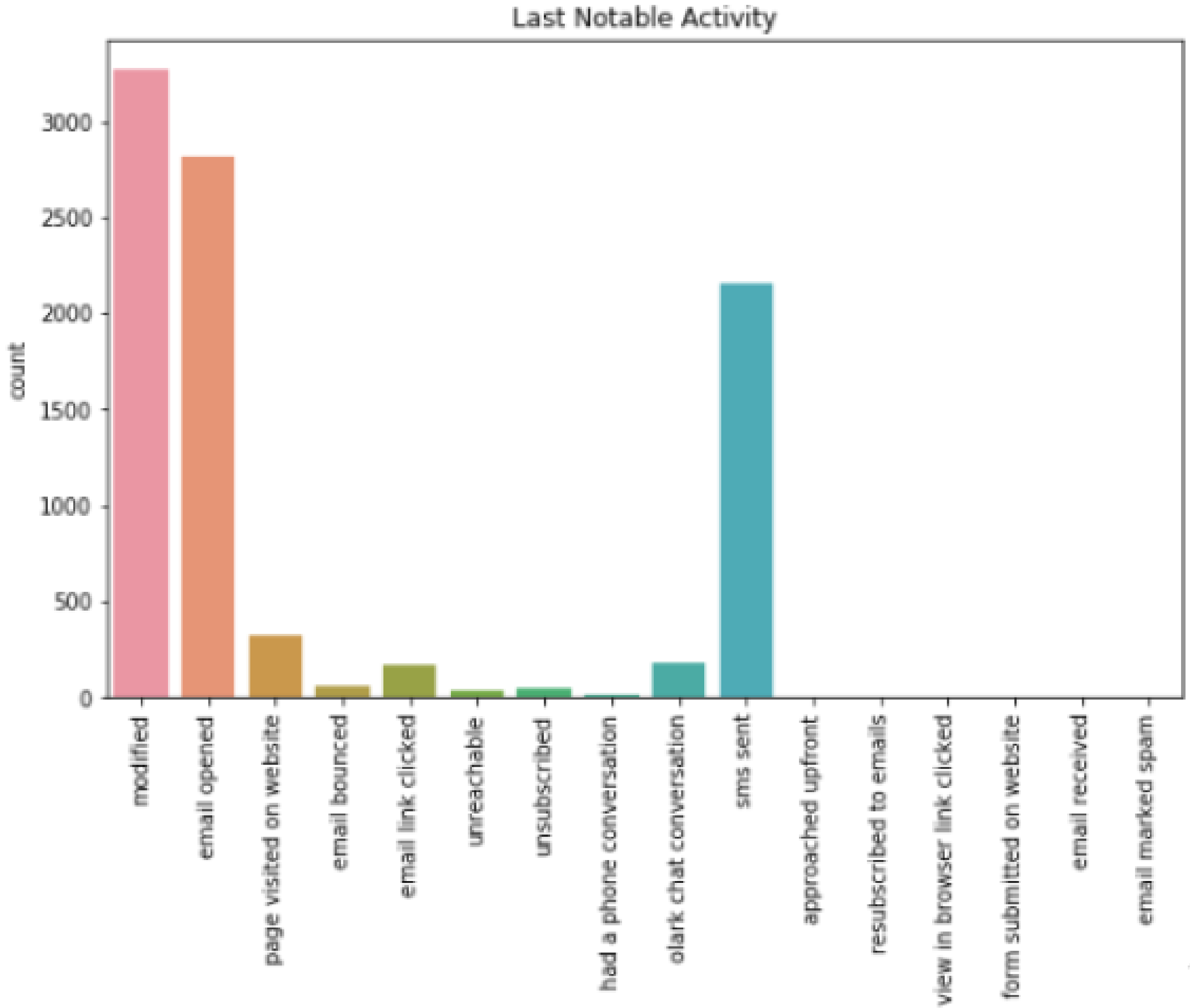
- Total Number of Rows =37, Total Number of Columns =9240.

- Single value features like “Magazine”, “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” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.

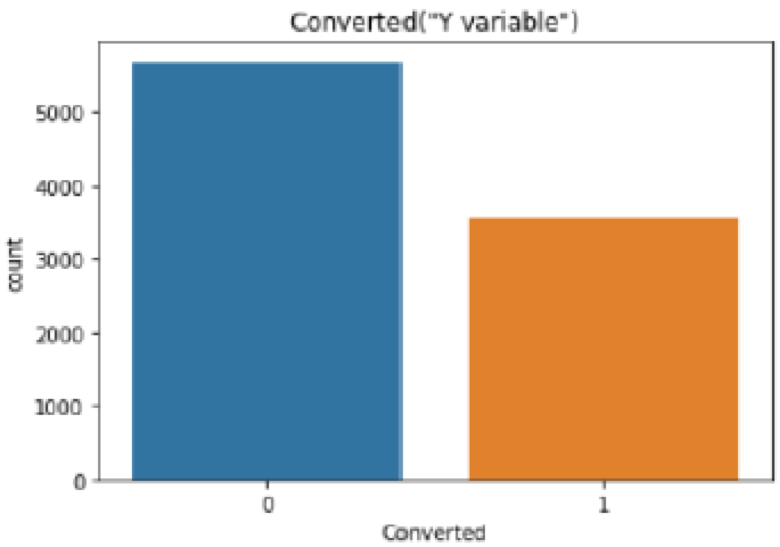
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.

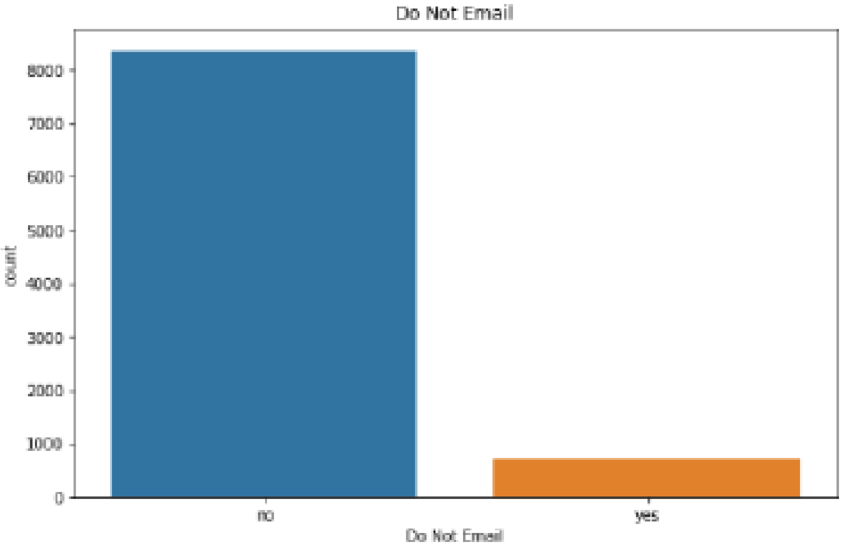
- Dropping the columns having more than 35% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

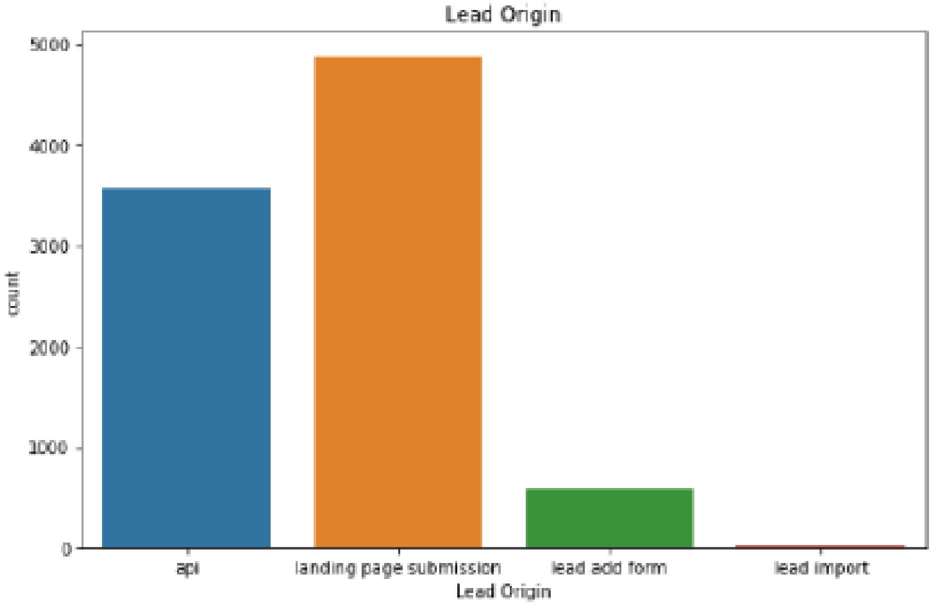
EDA



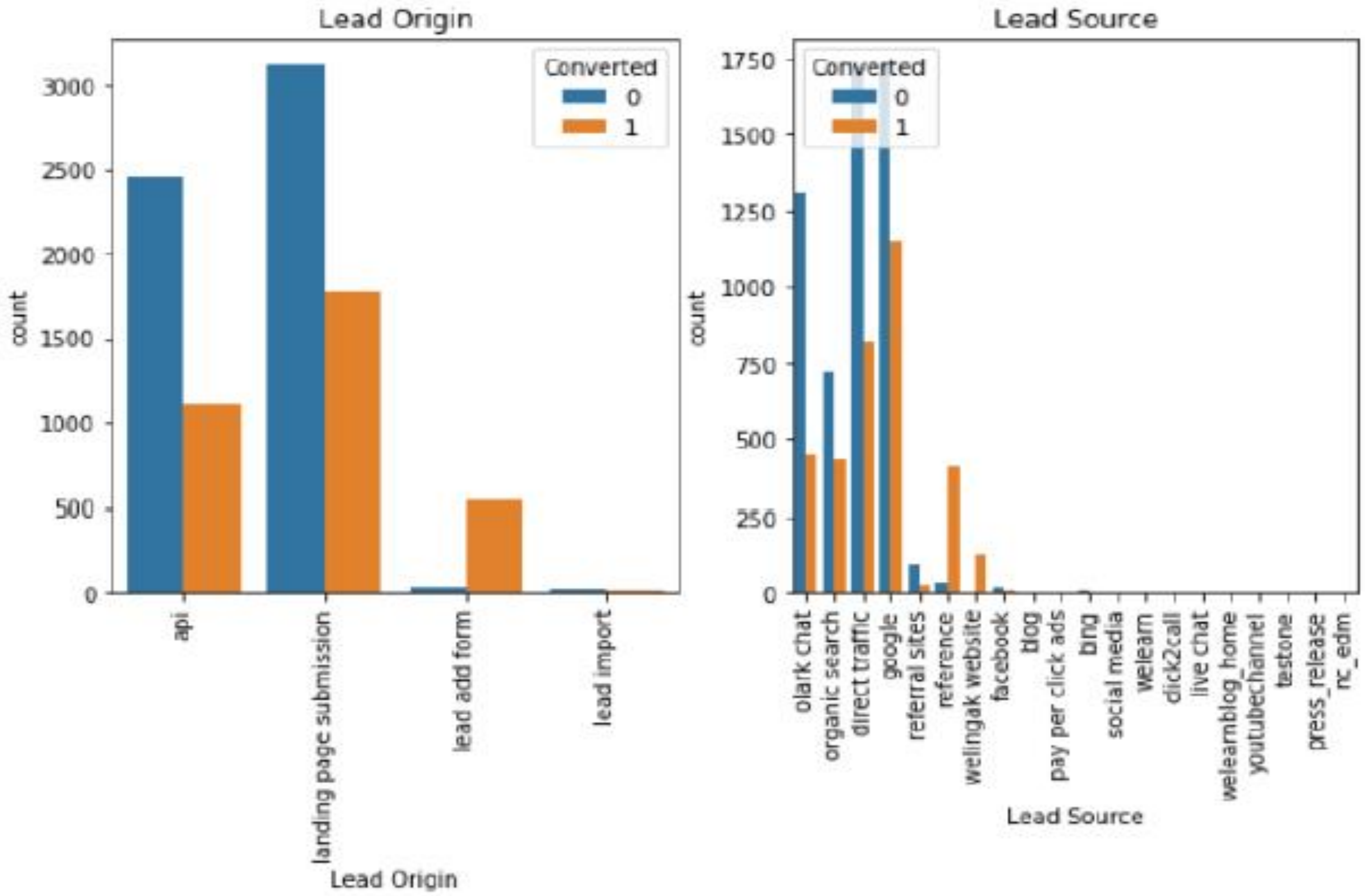
Last Notable Activity
count
0
500
1000
1500
2000
2500
3000
modified
email opened
page visited on website
email bounced
email link clicked
unreachable
unsubscribed
had a phone conversation
olark chat conversation
sms sent
approached upfront
resubscribed to emails
view in browser link clicked
form submitted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

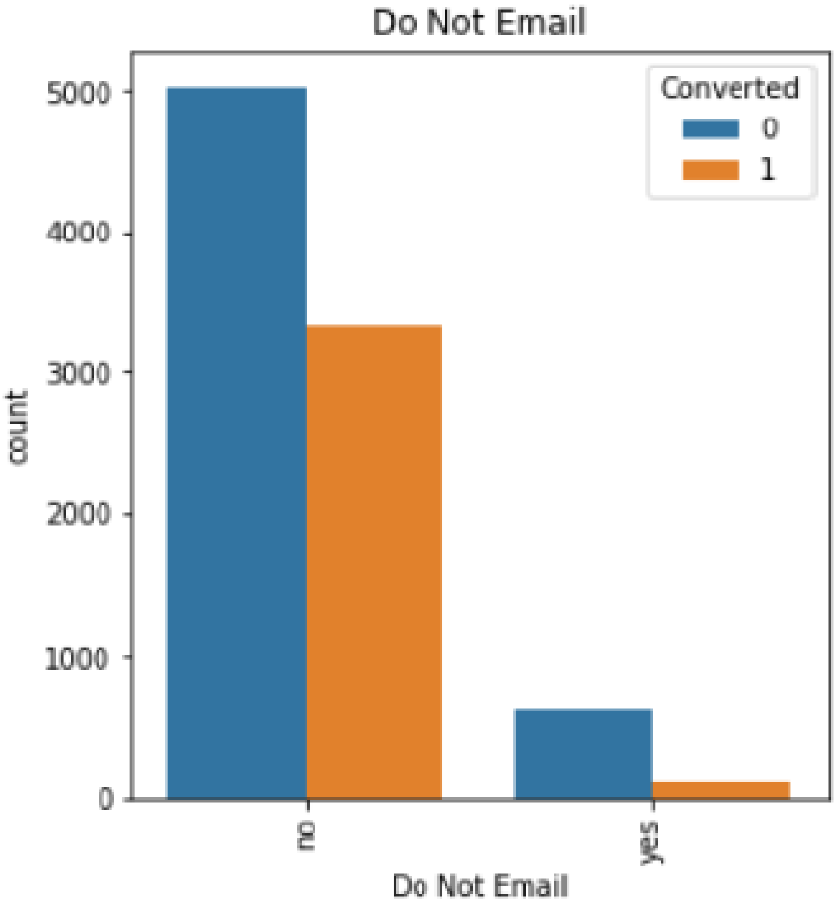

Converted("Y variable")
count
5000
4000
3000
2000
1000
0
0
1
Converted


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

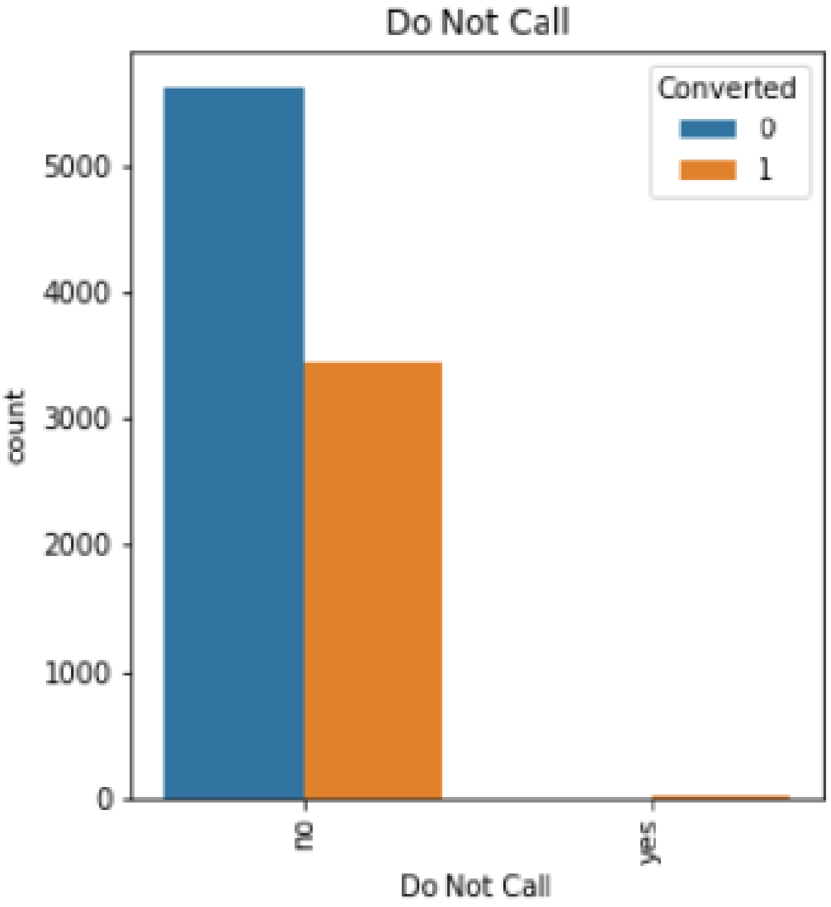

Lead Origin
count
5000
4000
3000
2000
1000
0
api
landing page submission
lead add form
lead import
Lead Origin

Categorical Variable Relation

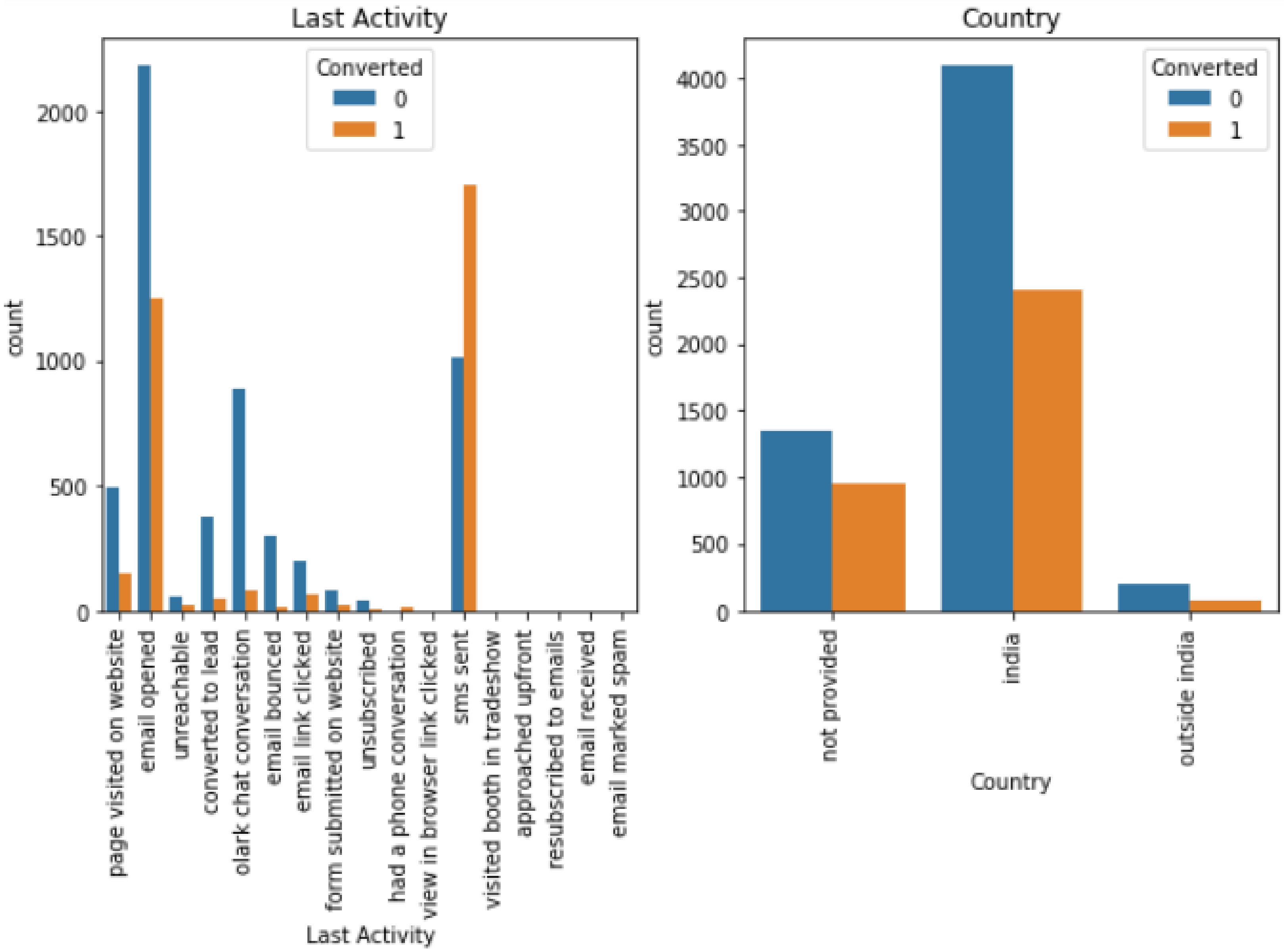




Do Not Email



Do Not Call

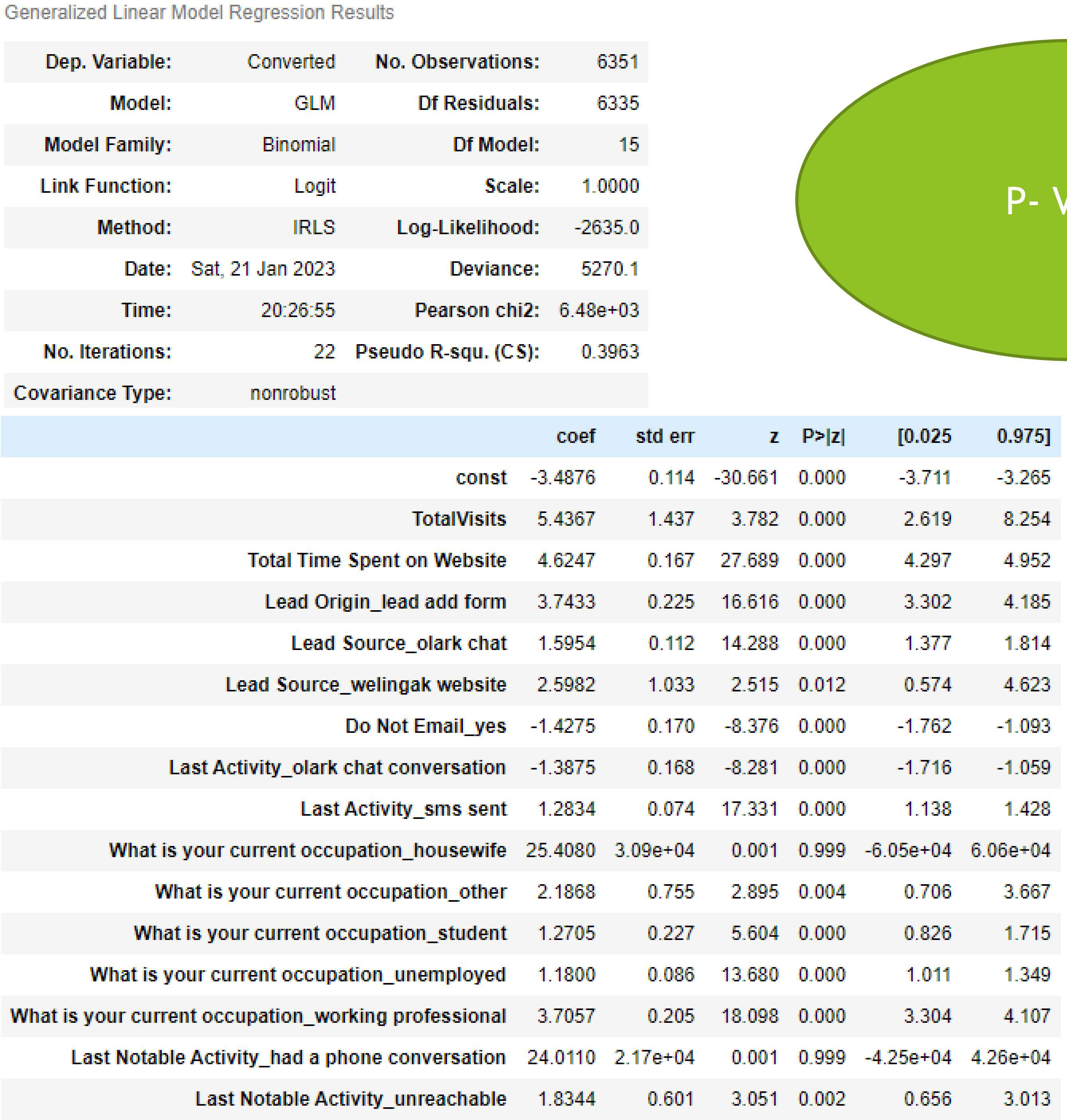

Last Activity
Converted
0
1
count
0
500
1000
1500
2000
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
sms sent
visited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam
Last Activity
Country
Converted
0
1
count
0
500
1000
1500
2000
2500
3000
3500
4000
not provided
india
outside india
Country

Data Conversion

- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 8792
- Total Columns for Analysis: 43

Model Building

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5
- Predictions on test data set
- Overall accuracy 81%



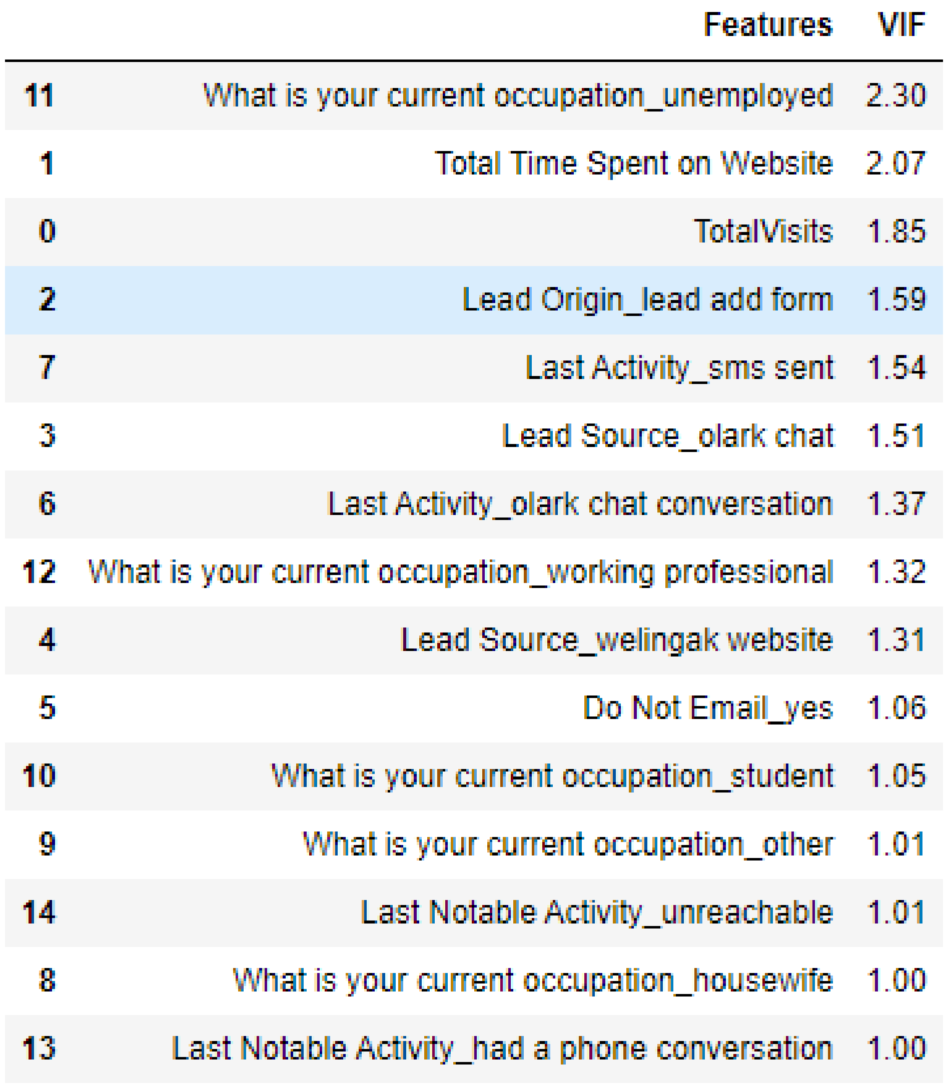
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6335
Model Family:	Binomial	**Df Model:**	15
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2635.0
Date:	Sat, 21 Jan 2023	**Deviance:**	5270.1
Time:	20:26:55	**Pearson chi2:**	6.48e+03
No. Iterations:	22	**Pseudo R-squ. (CS):**	0.3963
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.4876	0.114	-30.661	0.000	-3.711	-3.265
TotalVisits	5.4367	1.437	3.782	0.000	2.619	8.254
Total Time Spent on Website	4.6247	0.167	27.689	0.000	4.297	4.952
Lead Origin_lead add form	3.7433	0.225	16.616	0.000	3.302	4.185
Lead Source_olark chat	1.5954	0.112	14.288	0.000	1.377	1.814
Lead Source_welingak website	2.5982	1.033	2.515	0.012	0.574	4.623
Do Not Email_yes	-1.4275	0.170	-8.376	0.000	-1.762	-1.093
Last Activity_olark chat conversation	-1.3875	0.168	-8.281	0.000	-1.716	-1.059
Last Activity_sms sent	1.2834	0.074	17.331	0.000	1.138	1.428
What is your current occupation_housewife	25.4080	3.09e+04	0.001	0.999	-6.05e+04	6.06e+04
What is your current occupation_other	2.1868	0.755	2.895	0.004	0.706	3.667
What is your current occupation_student	1.2705	0.227	5.604	0.000	0.826	1.715
What is your current occupation_unemployed	1.1800	0.086	13.680	0.000	1.011	1.349
What is your current occupation_working professional	3.7057	0.205	18.098	0.000	3.304	4.107
Last Notable Activity_had a phone conversation	24.0110	2.17e+04	0.001	0.999	-4.25e+04	4.26e+04
Last Notable Activity_unreachable	1.8344	0.601	3.051	0.002	0.656	3.013

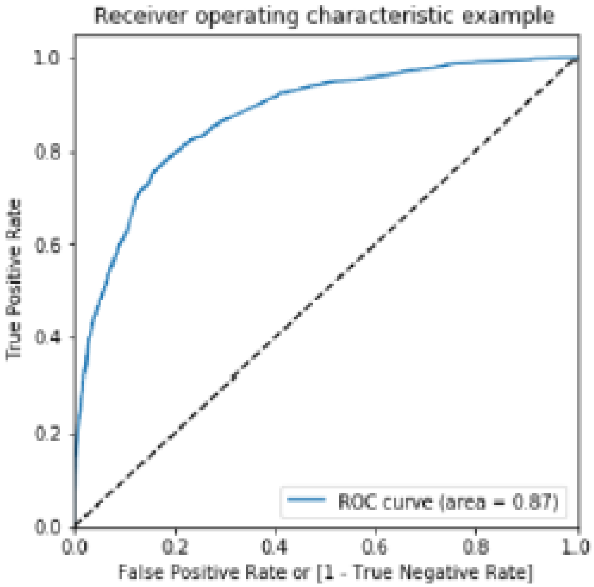
P- Value

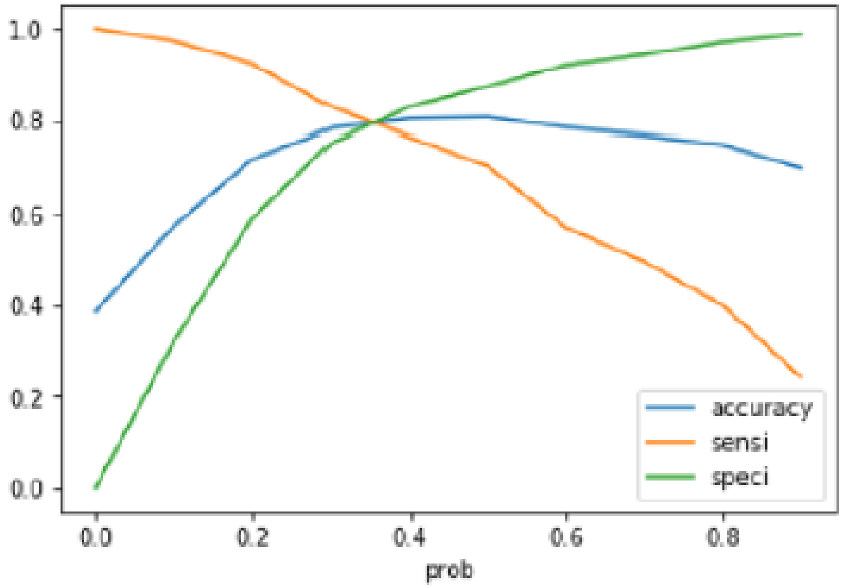
VIF Calcaulation



	Features	VIF
11	What is your current occupation_unemployed	2.30
1	Total Time Spent on Website	2.07
0	TotalVisits	1.85
2	Lead Origin_lead add form	1.59
7	Last Activity_sms sent	1.54
3	Lead Source_olark chat	1.51
6	Last Activity_olark chat conversation	1.37
12	What is your current occupation_working professional	1.32
4	Lead Source_welingak website	1.31
5	Do Not Email_yes	1.06
10	What is your current occupation_student	1.05
9	What is your current occupation_other	1.01
14	Last Notable Activity_unreachable	1.01
8	What is your current occupation_housewife	1.00
13	Last Notable Activity_had a phone conversation	1.00

ROC Curve





- **Finding Optimal Cut off Point**
- Optimal cut off probability is that
- probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.35.

Conclusion

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

- The total time spend on the Website.
- Total number of visits.
- When the lead source was:
  a. Google
  b. Direct traffic
  c. Organic search
  d. Welingak website
- When the last activity was:
  a. SMS
  b. Olark chat conversation
- When the lead origin is Lead add format.
- When their current occupation is as a working professional.
  Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.