LEAD SCORING CASE STUDY

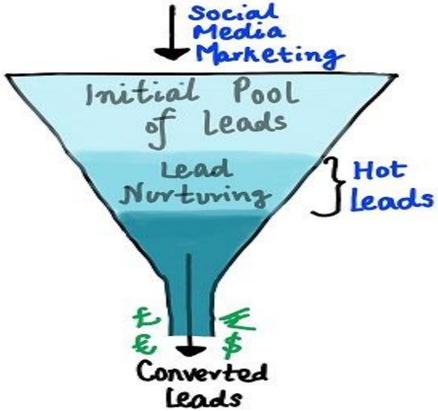
Submitted by :
Shubham Singh
Garima Sharma

Batch : PGDDS 2020

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.



As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, we need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

Plan of Action

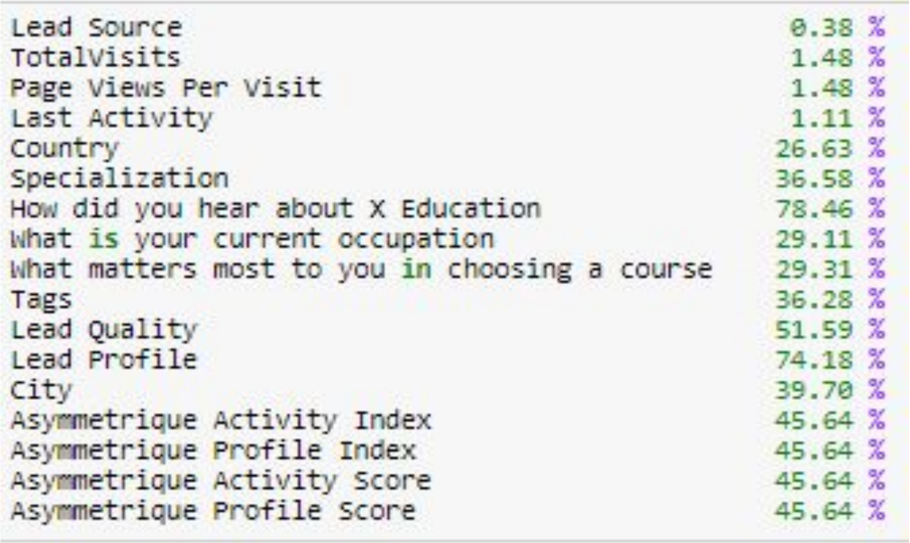
Building a Logistic Regression model to figure out if a lead has high probability of sales conversion.

Steps Involved:

1. Data Cleaning
2. EDA & Visualizations
3. Preparing the data for Model building
4. Creating the model
5. Business Evaluation

DATA INSPECTING AND DATA CLEANING

1.The dataset provided has 9240 rows and 37 columns. Out of 37 columns 20 had missing values and few columns had ‘Select’ category which was nothing but null values. After imputing ‘Select’ as np.nan the data has following percentages of missing values.



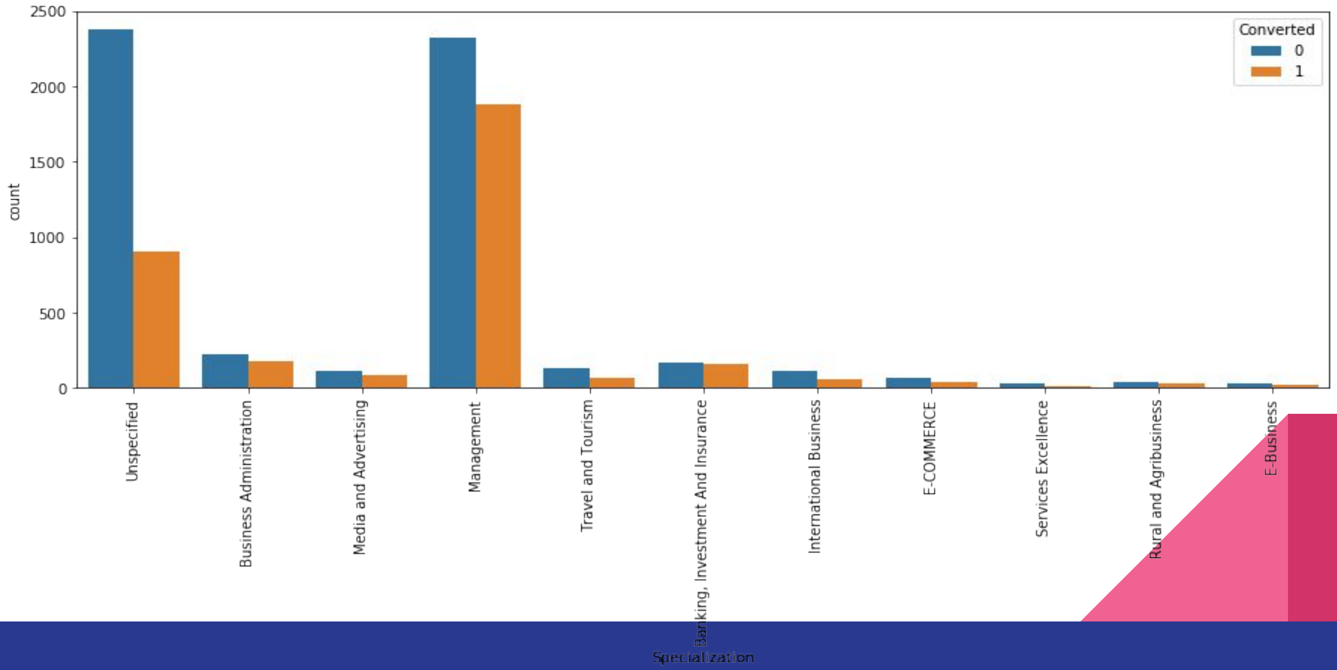
```
Lead Source                                  0.38 %
TotalVisits                                  1.48 %
Page Views Per Visit                         1.48 %
Last Activity                                1.11 %
Country                                     26.63 %
Specialization                              36.58 %
How did you hear about X Education          78.46 %
What is your current occupation             29.11 %
What matters most to you in choosing a course   29.31 %
Tags                                        36.28 %
Lead Quality                                51.59 %
Lead Profile                                74.18 %
City                                        39.70 %
Asymmetrique Activity Index                 45.64 %
Asymmetrique Profile Index                  45.64 %
Asymmetrique Activity Score                 45.64 %
Asymmetrique Profile Score                  45.64 %
```

- Removed all the columns where missing values percentage was more than 45%.
- **Lead Source , TotalVisits , Page Views Per Visit & Last Activity** columns had missing values less than 2% , since we had approx. 9000 rows removing these rows won't impact much , so **removed all the rows where these columns had missing values.**
- **Country , Specialization , What is your current occupation , What matters most to you in choosing a course , Tags & City** - these columns had significant amount of missing values , so handled these columns by **mode** imputation.
- **[‘Country’ , 'What matters most to you in choosing a course' ,'Do Not Email' , 'Do Not Call' , 'Search' , 'Magazine' , 'Newspaper Article' , 'X Education Forums' , 'Newspaper' , 'Digital Advertisement' , 'Through Recommendations' , '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']** All these columns had skewed distribution so simply removed these columns.

EDA & VISUALIZATIONS

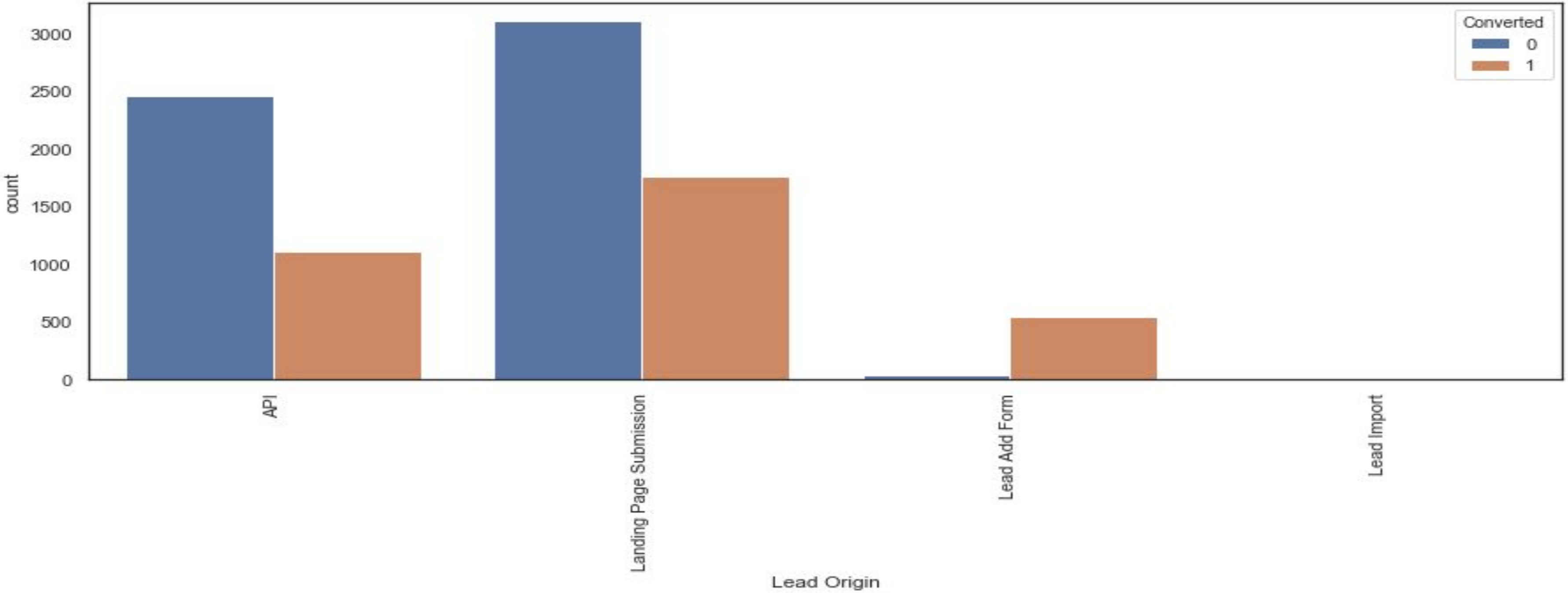
1. Visualizing Specialization and Conversion trend

After clubbing all the management related specializations as ‘Management’ and specializations which were recorded very few as ‘Unspecified’ , it was observed that the ‘Management’ has maximum sales conversion rate.



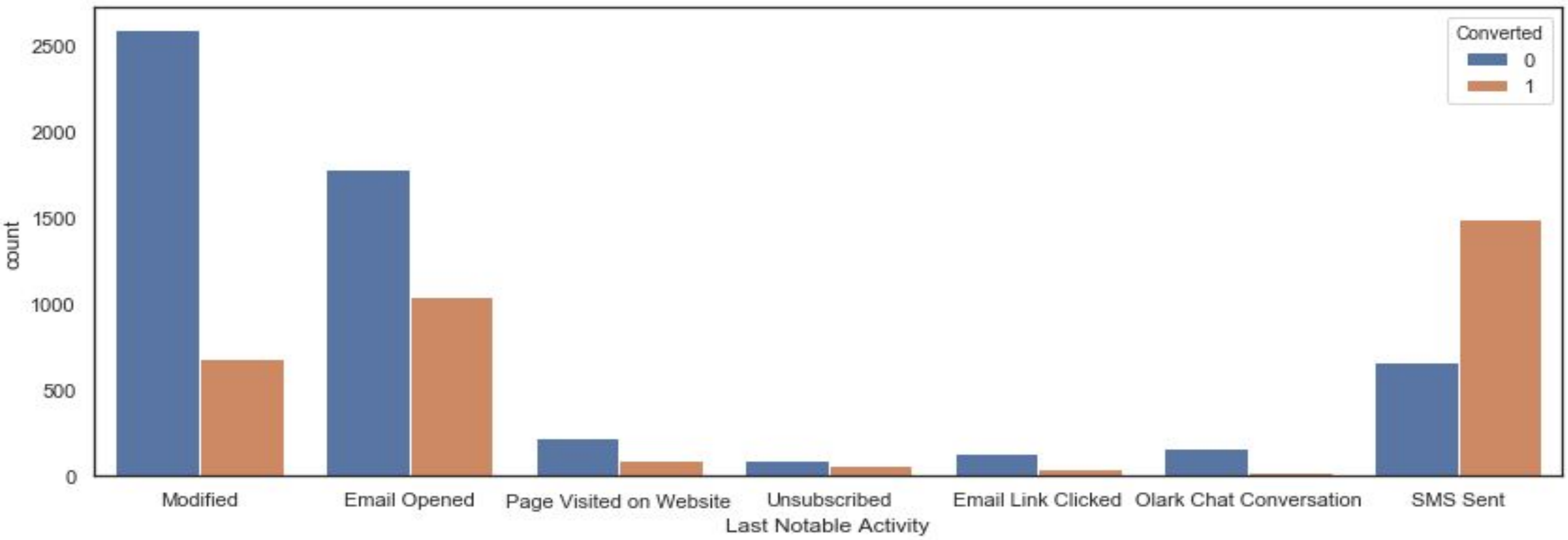
2. ‘Lead origin’ and ‘Converted’ trend

On visualizing the ‘Lead Origin’ column vs. ‘Converted’ column it was observed that the highest number of lead generated and highest sales conversion is through ‘Landing Page Submission’ followed by ‘API’. But if we closely observe we can also see that highest leads converted with respect to leads generated is through ‘Lead Add Form’.



3. Last Notable Activity & Converted value

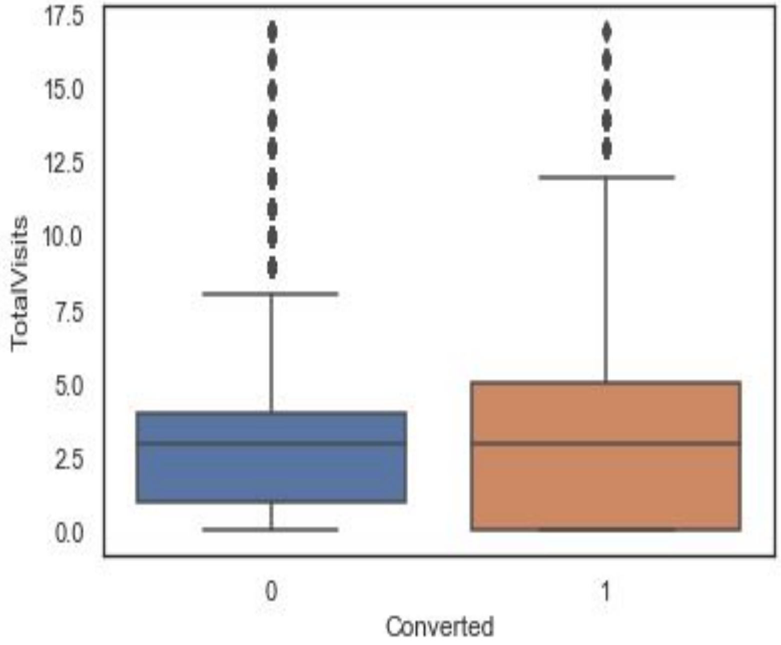
Here we can observe that those leads who Sent SMS has the highest sales conversion rate and the two most frequent last activities noted are ‘ Modified ‘ and ‘Email Opened’.



"Total Visits" and “Converted variable”

After removing the outliers from ‘Total Visits’ , It was observed on using boxplot with ‘Converted’ that median values of both Converted and Not converted are same.

No Inference can be made on the basis of ‘Total Visits’.



"Total Time Spent on Website" vs “Converted” variable

After removing the outliers from ‘Total Time Spent on Website’

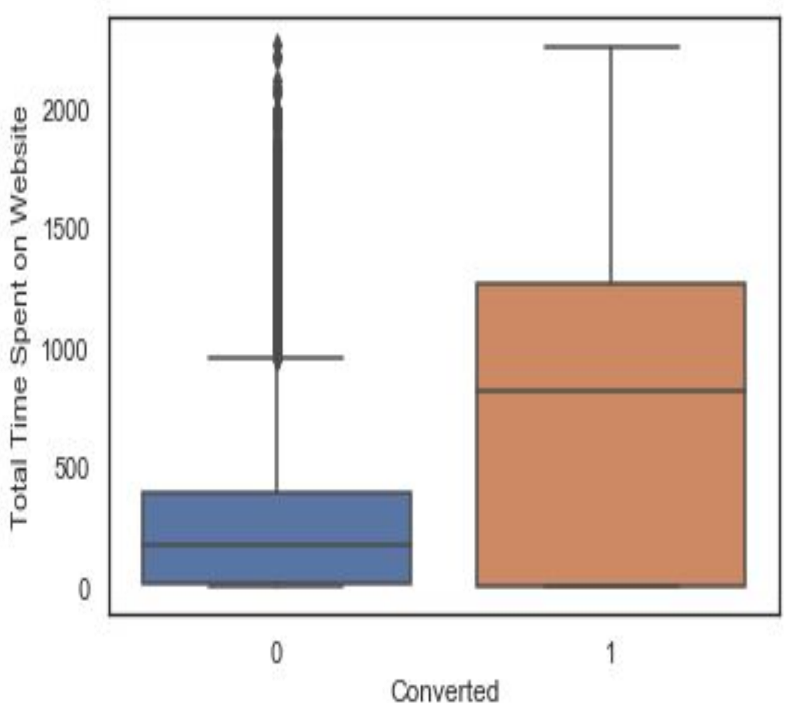
, It was observed on using boxplot with ‘Converted’ that leads

Which got converted spent way much more time as compared

to the leads which were not converted.From this we can

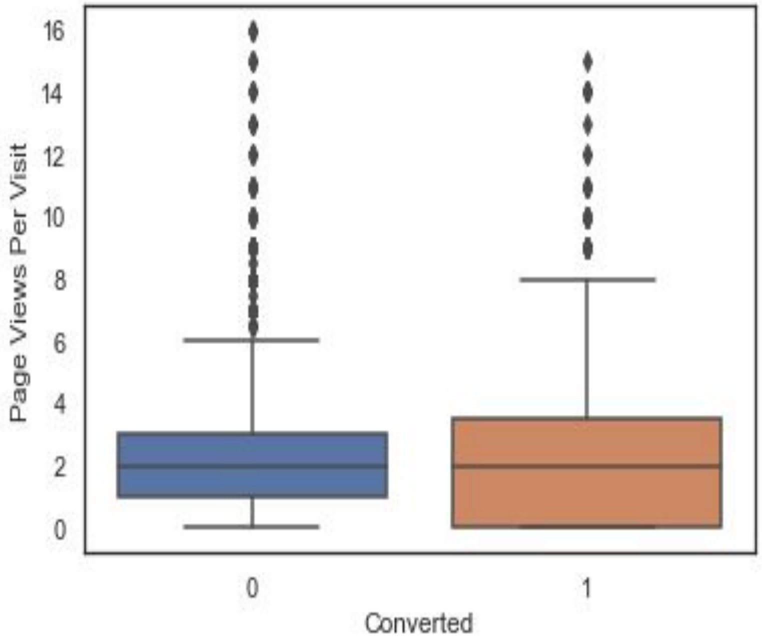
make an inference that total time spent on website

has a direct effect on Sales conversion.



"Page Views Per Visit" vs 'Converted' variable

By using boxplot it was observed that median values for both ‘Converted’ & ‘Not converted’ are same , which means that Lleads which got Converted and Not converted both browsed equal number of pages per visit on average. We cannot make any inference from this column.



Preparing the data for modelling

- After cleaning the data and removing all the columns which had highly skewed distribution now it's time to prepare the data for modelling

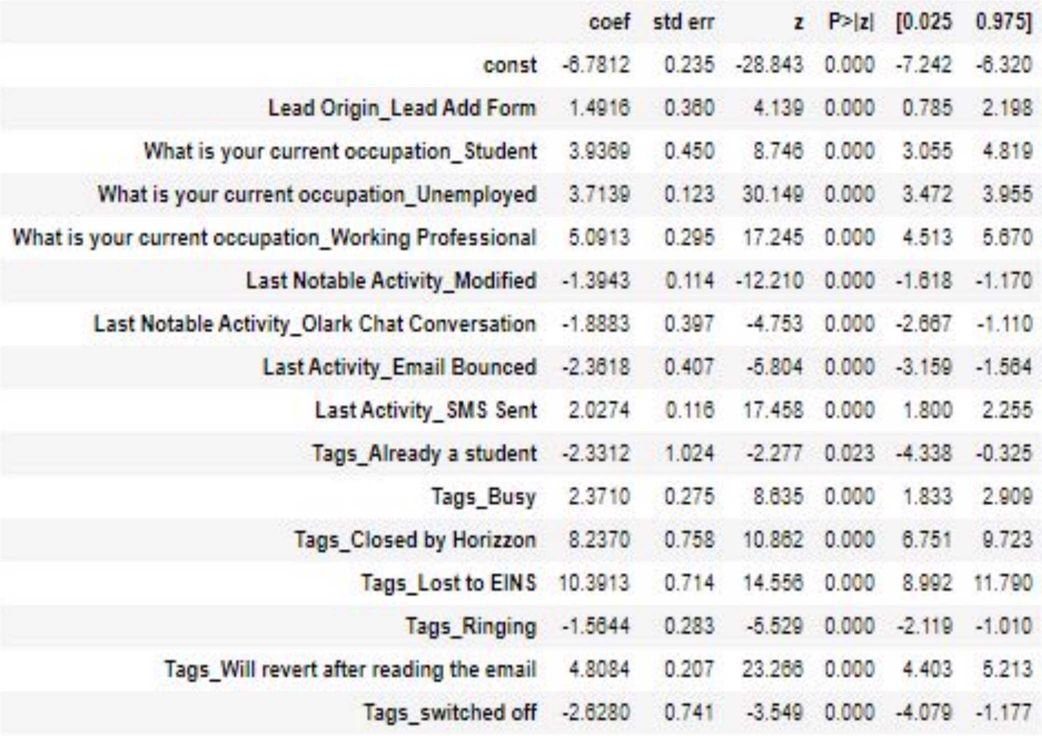
  Steps involved :

1. Dummy variables creation: For categorical columns dummy variables were created and original columns were dropped by using pandas.
2. Splitting the data : The dataset was splitted into training set(size=70%) and testing set(30%).
3. Scaling : All the numerical columns were scaled by using Standard scaling method.

Model building on Training data set

After performing data cleaning , data preparing(Dummy variable creation , Scaling) and splitting the data into training and testing datasets , proceeded with the model building using RFE & VIF using 15 output variables on training dataset.

In the first attempt it was observed that p-value for all the 15 variables was less than 0.05 which was great.



	coef	std err	z	P>|z|	[0.025	0.975]
const	-6.7812	0.235	-28.843	0.000	-7.242	-6.320
Lead Origin_Lead Add Form	1.4916	0.360	4.139	0.000	0.785	2.198
What is your current occupation_Student	3.9369	0.450	8.746	0.000	3.055	4.819
What is your current occupation_Unemployed	3.7139	0.123	30.149	0.000	3.472	3.955
What is your current occupation_Working Professional	5.0913	0.295	17.245	0.000	4.513	5.670
Last Notable Activity_Modified	-1.3943	0.114	-12.210	0.000	-1.618	-1.170
Last Notable Activity_Olark Chat Conversation	-1.8883	0.397	-4.753	0.000	-2.667	-1.110
Last Activity_Email Bounced	-2.3618	0.407	-5.804	0.000	-3.159	-1.564
Last Activity_SMS Sent	2.0274	0.116	17.458	0.000	1.800	2.255
Tags_Already a student	-2.3312	1.024	-2.277	0.023	-4.338	-0.325
Tags_Busy	2.3710	0.275	8.635	0.000	1.833	2.909
Tags_Closed by Horizzon	8.2370	0.758	10.862	0.000	6.751	9.723
Tags_Lost to EINS	10.3913	0.714	14.556	0.000	8.992	11.790
Tags_Ringing	-1.5644	0.283	-5.529	0.000	-2.119	-1.010
Tags_Will revert after reading the email	4.8084	0.207	23.266	0.000	4.403	5.213
Tags_switched off	-2.6280	0.741	-3.549	0.000	-4.079	-1.177

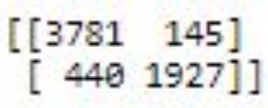
Checking VIF values

On checking the variation inflation factor it was observed that all 15 variables had vif less than 5 , which is great again.

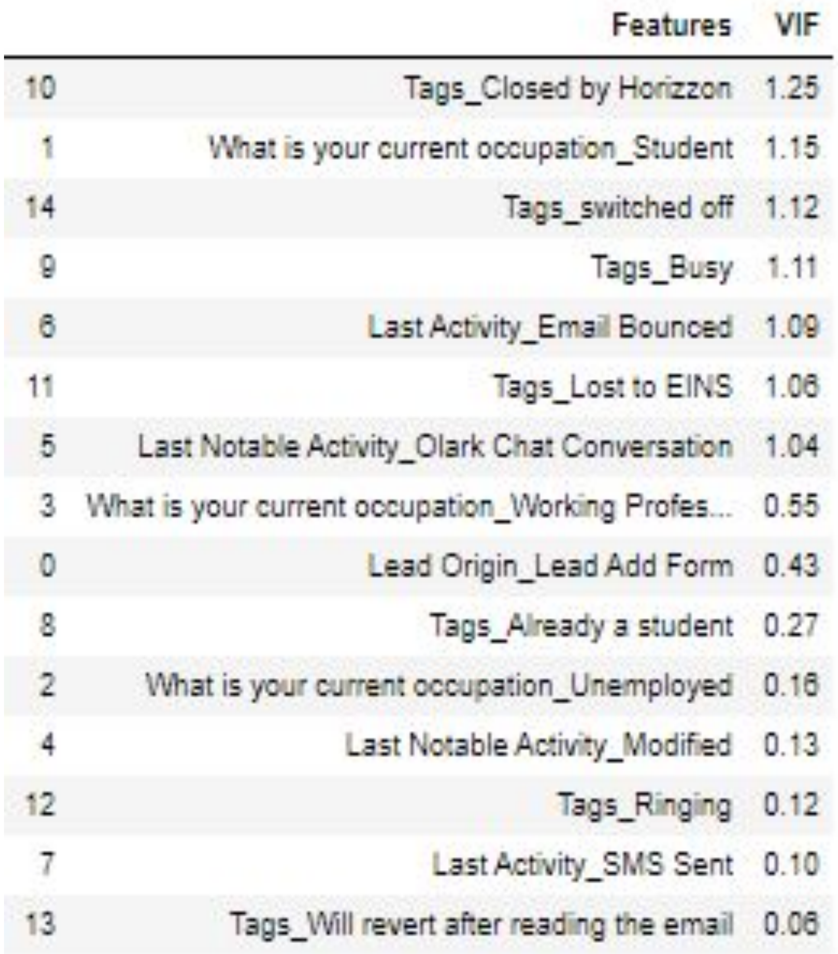
Accuracy & Confusion matrix:

Overall accuracy = 90.70%

Confusion matrix =



```
[[3781  145]
 [ 440 1927]]
```

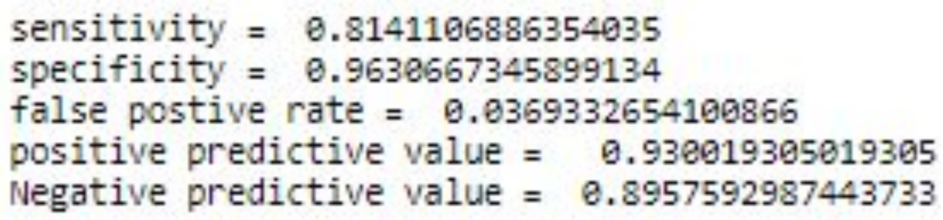


	Features	VIF
10	Tags_Closed by Horizzon	1.25
1	What is your current occupation_Student	1.15
14	Tags_switched off	1.12
9	Tags_Busy	1.11
6	Last Activity_Email Bounced	1.09
11	Tags_Lost to EINS	1.06
5	Last Notable Activity_Olark Chat Conversation	1.04
3	What is your current occupation_Working Profes...	0.55
0	Lead Origin_Lead Add Form	0.43
8	Tags_Already a student	0.27
2	What is your current occupation_Unemployed	0.16
4	Last Notable Activity_Modified	0.13
12	Tags_Ringing	0.12
7	Last Activity_SMS Sent	0.10
13	Tags_Will revert after reading the email	0.06

MODEL EVALUATION & OPTIMIZATION

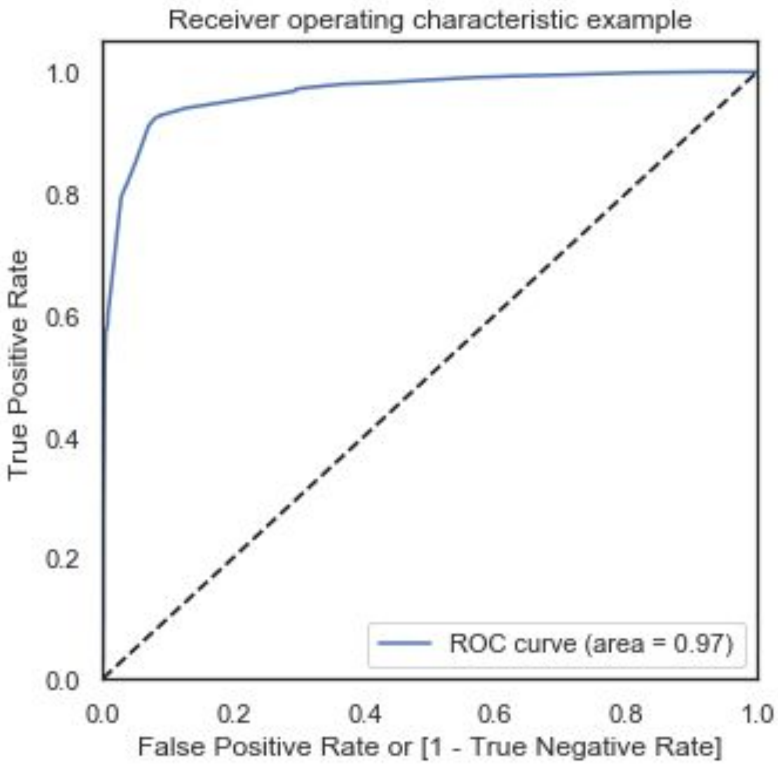
By using confusion matrix following metrics

was calculated to evaluate the model.



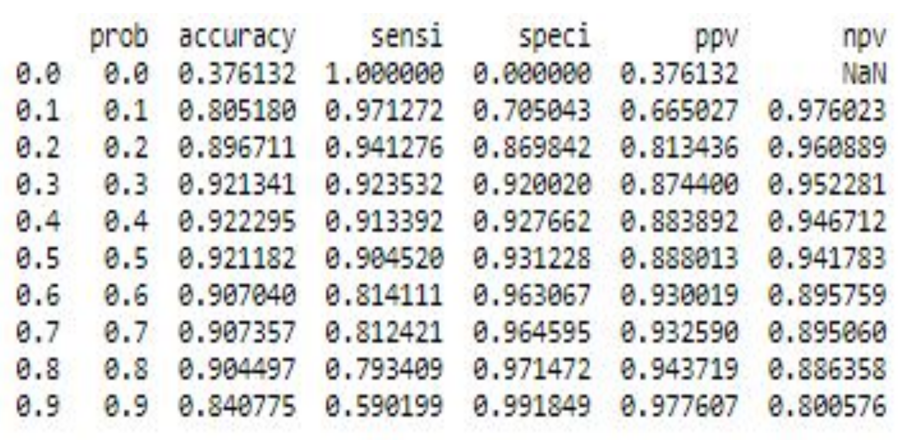
```
sensitivity =  0.8141106886354035
specificity =  0.963066734589134
false postive rate =  0.036933265410866
positive predictive value =   0.930019305019305
Negative predictive value =  0.8957592987443733
```

ROC CURVE

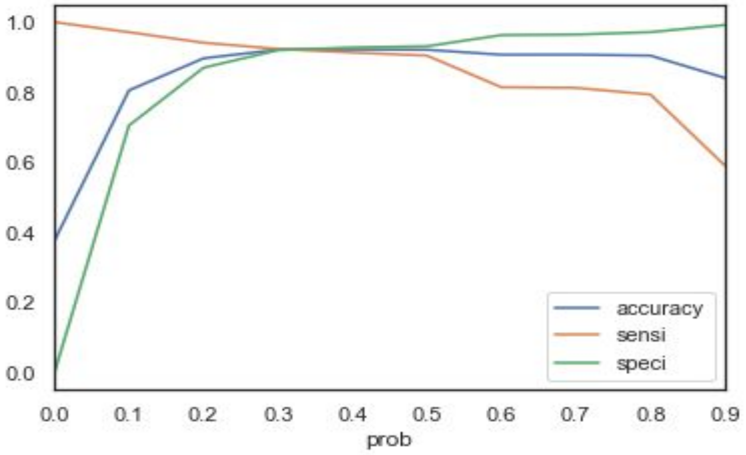


ROC curve area = 0.97 , pretty close to 1 which is good.

Finding Optimal Cutoff Point



```
     prob  accuracy     sensi     speci       ppv       npv
0.0   0.0  0.376132  1.000000  0.000000  0.376132       NaN
0.1   0.1  0.805180  0.971272  0.705043  0.665027  0.976023
0.2   0.2  0.896711  0.941276  0.869842  0.813436  0.960889
0.3   0.3  0.921341  0.923532  0.920020  0.874400  0.952281
0.4   0.4  0.922295  0.913392  0.927662  0.883892  0.946712
0.5   0.5  0.921182  0.904520  0.931228  0.888013  0.941783
0.6   0.6  0.907040  0.814111  0.963067  0.930019  0.895759
0.7   0.7  0.907357  0.812421  0.964595  0.932590  0.895060
0.8   0.8  0.904497  0.793409  0.971472  0.943719  0.886358
0.9   0.9  0.840775  0.590199  0.991849  0.977607  0.800576
```



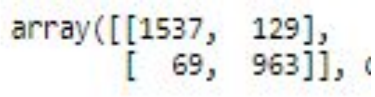
From the accuracy , sensitivity & specificity curve intesection , optimal cutoff probability came out to be 0.3

Final Model building on Test data set

Using 0.3 cutoff probability final model was build on test data set.

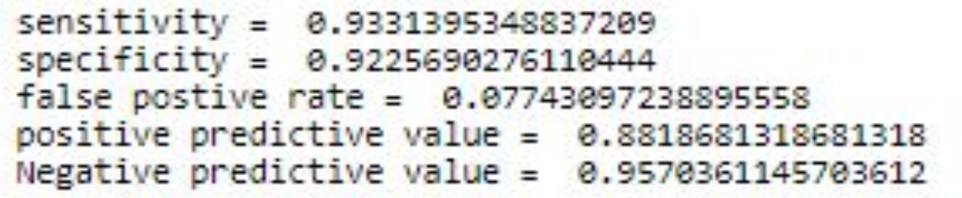
Model accuracy - 92.66%

Confusion matrix -



```
array([[1537,  129],
       [  69,  963]],
```

Evaluation metric -



```
sensitivity =  0.9331395348837209
specificity =  0.9225690276110444
false postive rate =  0.07743097238895558
positive predictive value =  0.8818681318681318
Negative predictive value =  0.9570361145703612
```

THANK YOU!!!!