LEAD SCORE

CASE STUDY

By-
Subhasis
Pattanayak
Arvind Kumar Patel

Presented to :-

Chief Data Scientist

Business Statement

- X Education sells online courses to industry professionals, the company markets its courses on several websites and search engines like google. 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%.

- ## Problem Statement

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

Role and Objective

X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company wants 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%.

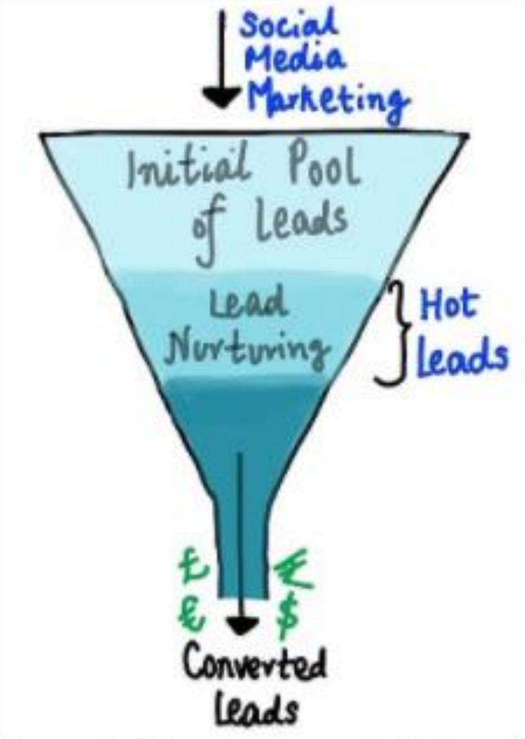
Business Aspect

- The company markets its courses on several websites and search engines like Google.

- Those who fill form are consider as Lead.

- Those who join the course are consider as Converted Lead

- Company is facing issue of low conversion rate i.e if 100 people are leads then only 30 person is successfully converted



ead Conversion Process - Demonstrated as a funnel

- Company want to increase conversion rate from 30 % to 80 %.
- With the help of data scientist, Company want to achieve this.
- Company also want Lead score between 0-100,So it can classify lead as a hot lead and cold lead.
- Hot lead are those who have high lead score and chances of their conversion is very high.

TECHNICAL ASPECT

Sanity Test and Step Taken

- We have dataset with 37 columns and 9240 rows.
- Columns have ‘Select’ value which need to be treated as null.
- Columns which have null value more than 40 % is better to drop and with less than 40%, we can impute them with mean, mode and median.
- Some categorical columns are highly skewed, typo error fixing them.
- We have seen that some categorical columns have many categories, we changed categories which are present in less value to “Other”

Exploratory Data Analysis

- In univariate analysis of categorical columns we have found that, Lead source direct traffic and google are the two main source for leads.
- In bivariate analysis of categorical columns we have found that those lead last activity or last notable activity is ‘SMS sent ’ have higher tendency to convert.

Data Preparation and Features Selection

- Mapping is done for features which have yes or no as 1/0.
- Dummfication of features are done on categorical columns.
- Train and test data is formed in the ratio of 0.7 and 0.3 respectively.
- Scaling of features are done with standard scaler.
- Correlation is checked and then RFE is used to get top 20 features.

Model Building and Evaluation

- With the help stats model we have build total of 6 model.
- In final model we have 15 features and with p value <0.05 and VIF<5.
- Different cutoff values are checked and then 0.37 is choose for optimal cutoff point.
- We have achieved:

	Train	Test
Accuracy	79	78
Sensitivity	81	80
specificity	77	76

Exploratory Data Analysis

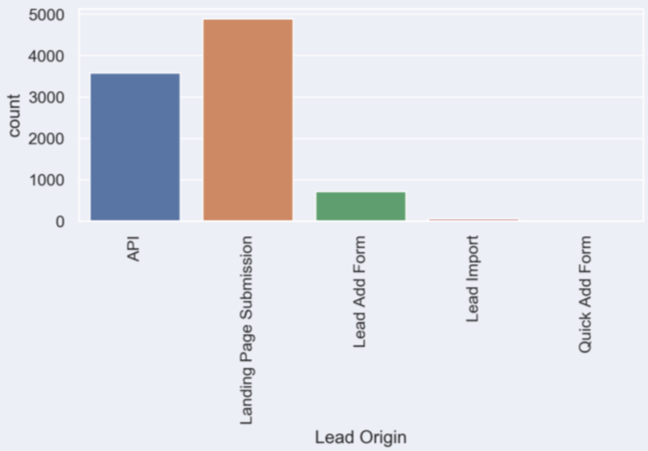
Univariate Categorical analysis

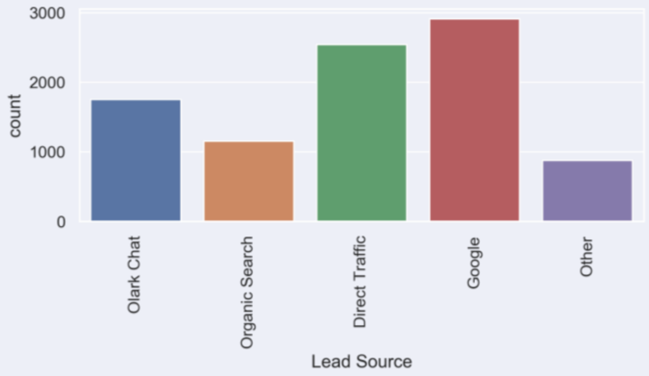
- In Lead Source Direct Traffic and Google are the two main source for Leads

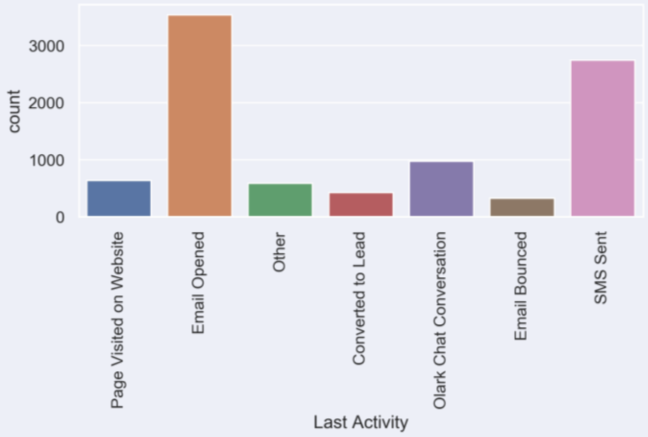
- The Number of values is High in Email Opened and SMS Sent in Last Activity

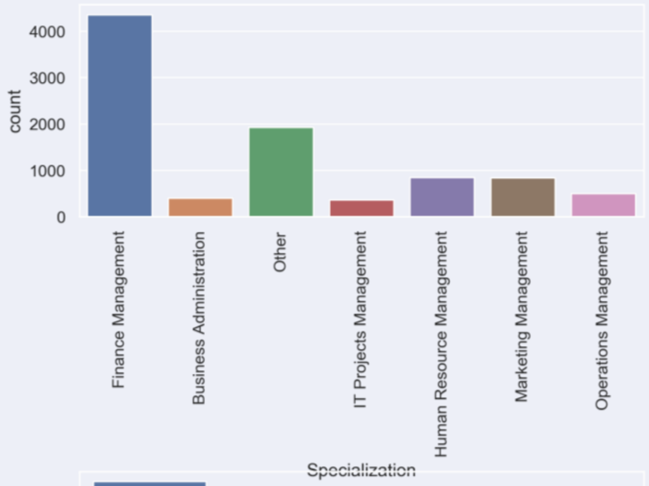
- Most of the people chooses Finance Management Specialization rather than other Specialization

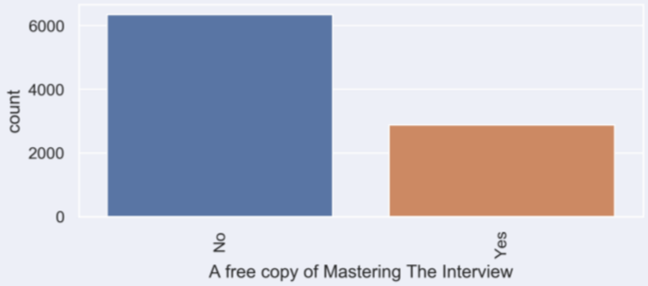
- The IT Project management have very lees so that most of the People not preferred this Specialization

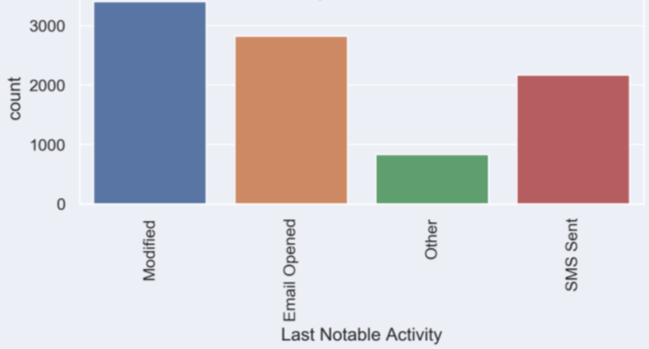










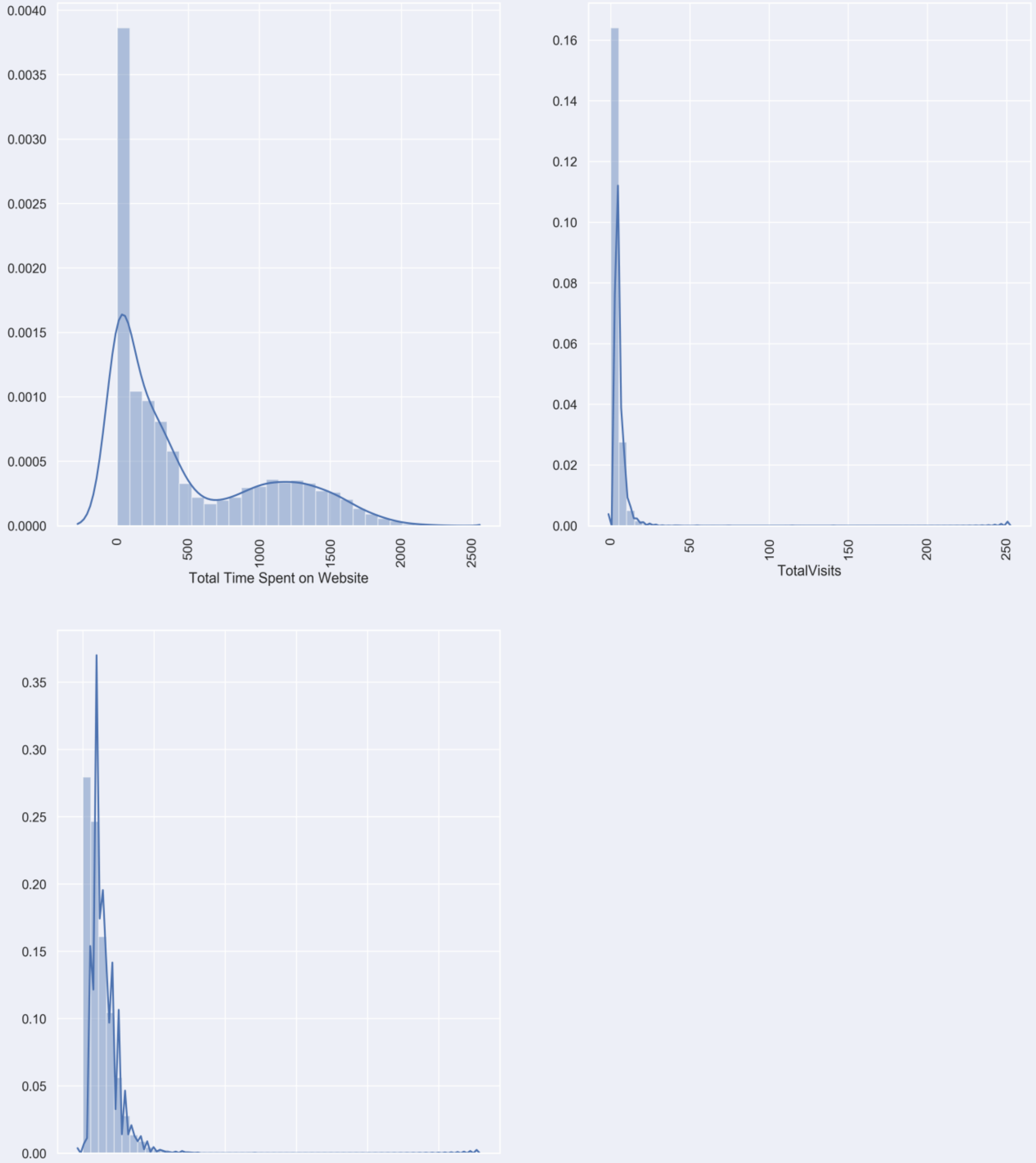


Univariate Continuous analysis

- None of the Continuous Variables are in Normal distribution

- Presence of Outliers in Total Visits and Page Views Per Visit

- In total visits more values is between 0-50 and page views per visits 0-20



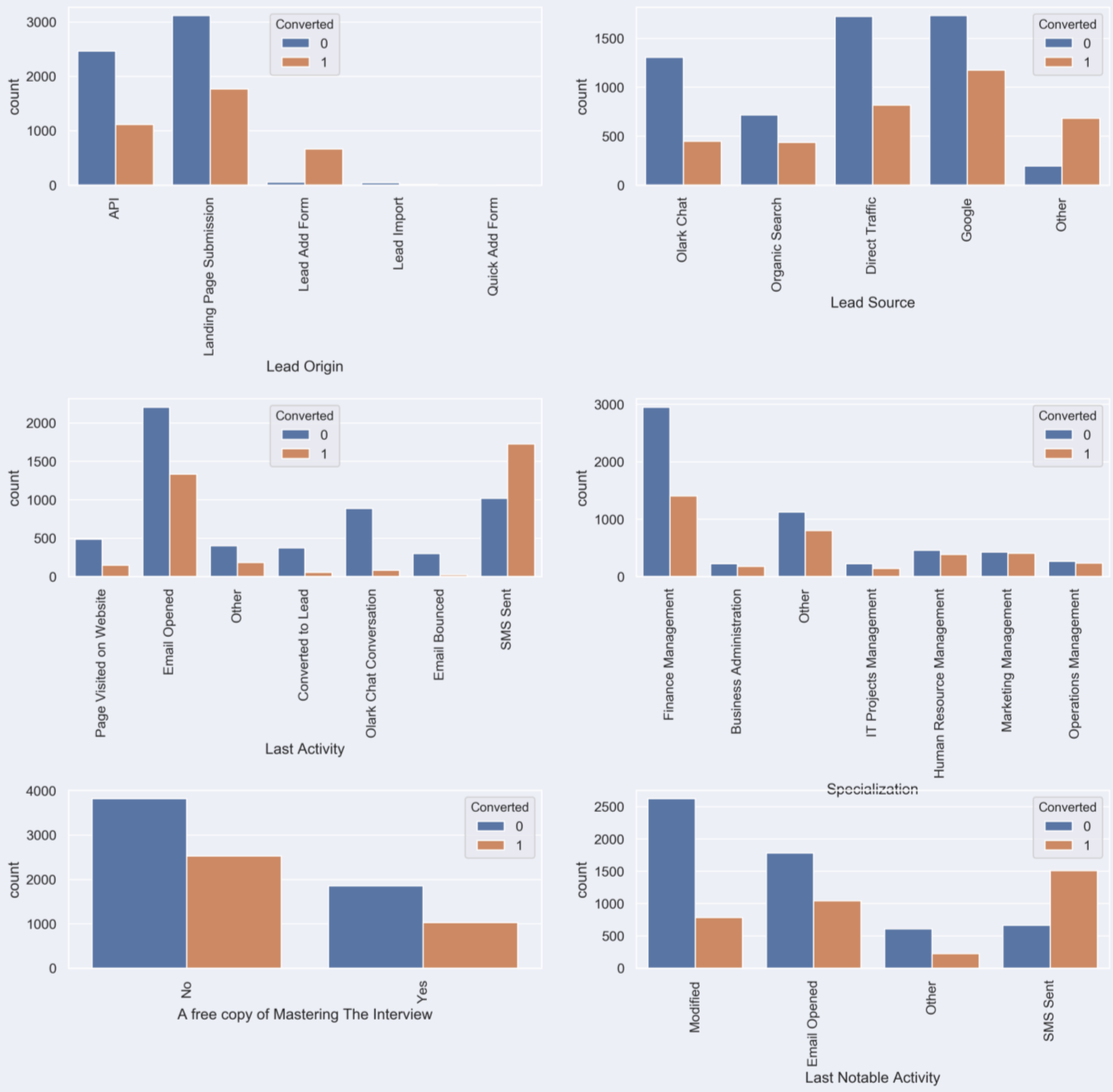
Bivariate Categorical analysis

- In Lead Source The number of Hot leads is higher in Direct Traffic and Google less in Other Category.

- In Last Activity the number of Hot leads is higher in SMS and in EMAIL cold leads is higher than hot leads.

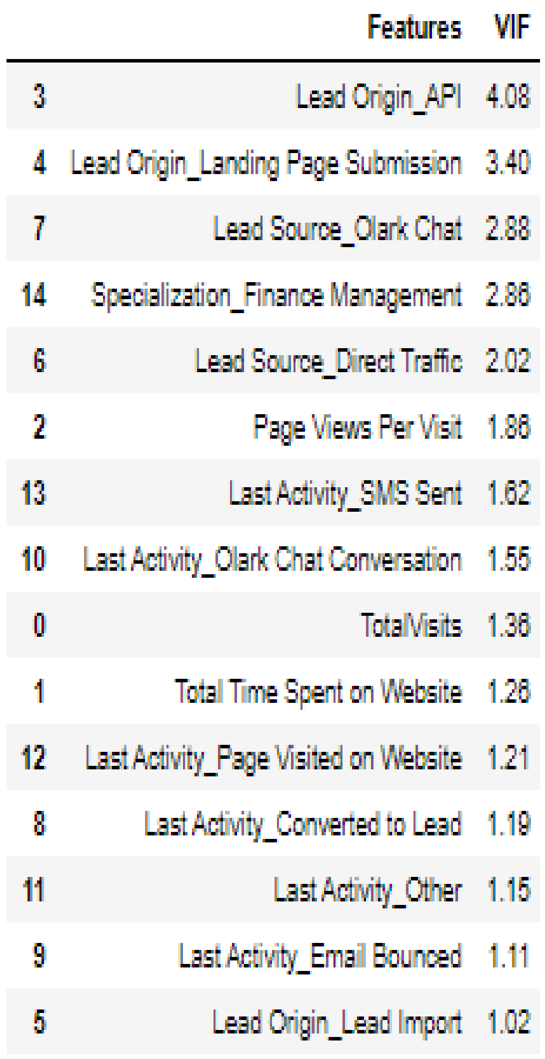
- In Last Notable Activity it's mostly same as Last Activity.

- In Specialization the most of the leads are comes from Finance management but here Hot leads are lesser than Cold leads.

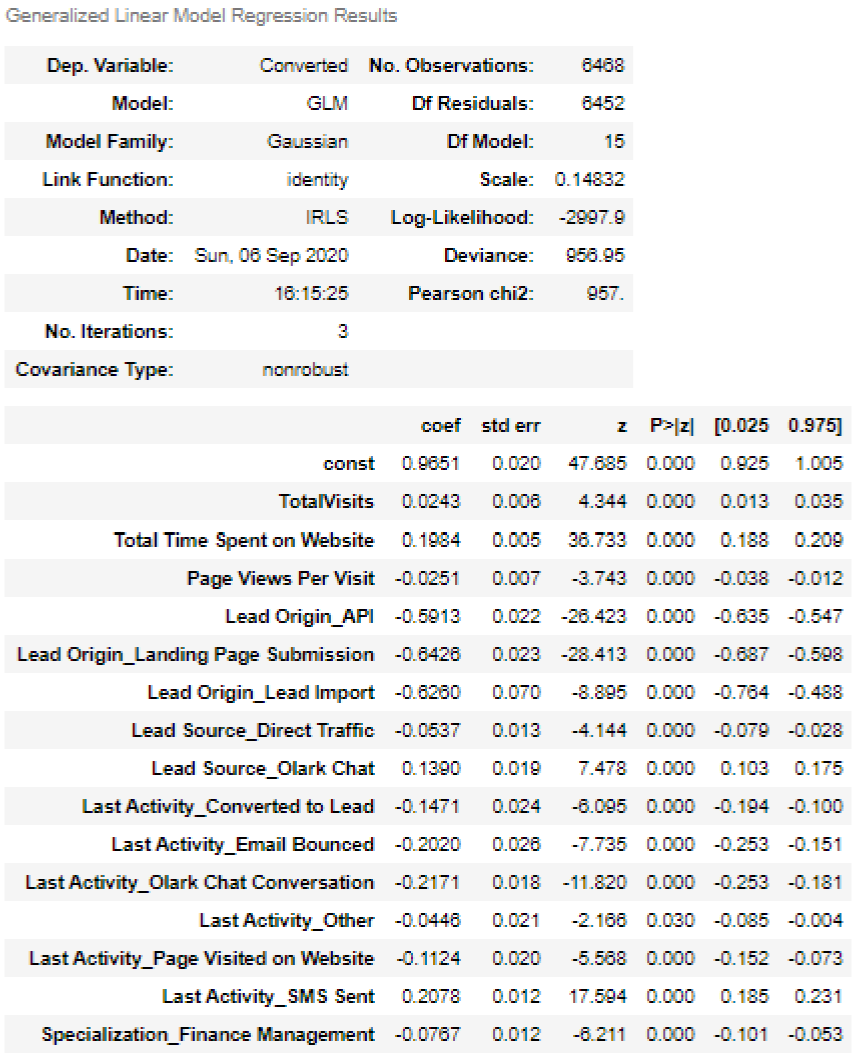


Model And Detail Analysis

- We have build Logistic Regression model with the help statsmodels.api library
- The figures belong to final model which tell about detail statistic.
- We have 15 features in final model.
- VIF of features are less than 5.
- P statistic of all features are under 0.05



	Features	VIF
3	Lead Origin_API	4.08
4	Lead Origin_Landing Page Submission	3.40
7	Lead Source_Olark Chat	2.88
14	Specialization_Finance Management	2.86
6	Lead Source_Direct Traffic	2.02
2	Page Views Per Visit	1.86
13	Last Activity_SMS Sent	1.62
10	Last Activity_Olark Chat Conversation	1.55
0	TotalVisits	1.36
1	Total Time Spent on Website	1.26
12	Last Activity_Page Visited on Website	1.21
8	Last Activity_Converted to Lead	1.19
11	Last Activity_Other	1.15
9	Last Activity_Email Bounced	1.11
5	Lead Origin_Lead Import	1.02



Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6452
Model Family:	Gaussian	Df Model:	15
Link Function:	identity	Scale:	0.14832
Method:	IRLS	Log-Likelihood:	-2997.9
Date:	Sun, 06 Sep 2020	Deviance:	956.95
Time:	16:15:25	Pearson chi2:	957.
No. Iterations:	3		
Covariance Type:	nonrobust		

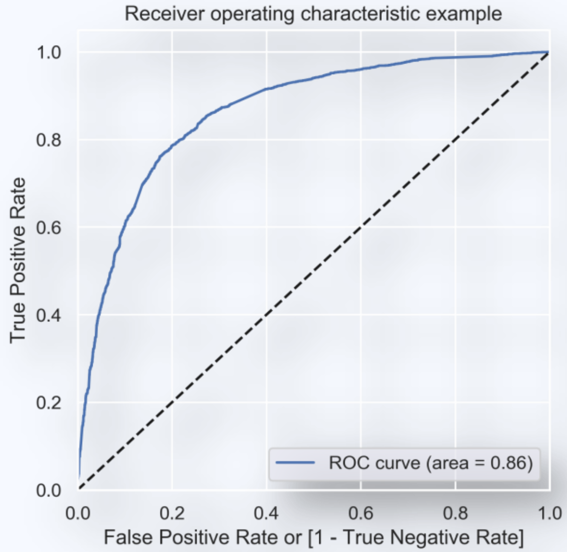
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.9651	0.020	47.685	0.000	0.925	1.005
TotalVisits	0.0243	0.006	4.344	0.000	0.013	0.035
Total Time Spent on Website	0.1984	0.005	36.733	0.000	0.188	0.209
Page Views Per Visit	-0.0251	0.007	-3.743	0.000	-0.038	-0.012
Lead Origin_API	-0.5913	0.022	-26.423	0.000	-0.635	-0.547
Lead Origin_Landing Page Submission	-0.6426	0.023	-28.413	0.000	-0.687	-0.598
Lead Origin_Lead Import	-0.6260	0.070	-8.895	0.000	-0.764	-0.488
Lead Source_Direct Traffic	-0.0537	0.013	-4.144	0.000	-0.079	-0.028
Lead Source_Olark Chat	0.1390	0.019	7.478	0.000	0.103	0.175
Last Activity_Converted to Lead	-0.1471	0.024	-6.095	0.000	-0.194	-0.100
Last Activity_Email Bounced	-0.2020	0.026	-7.735	0.000	-0.253	-0.151
Last Activity_Olark Chat Conversation	-0.2171	0.018	-11.820	0.000	-0.253	-0.181
Last Activity_Other	-0.0446	0.021	-2.166	0.030	-0.085	-0.004
Last Activity_Page Visited on Website	-0.1124	0.020	-5.568	0.000	-0.152	-0.073
Last Activity_SMS Sent	0.2078	0.012	17.594	0.000	0.185	0.231
Specialization_Finance Management	-0.0767	0.012	-6.211	0.000	-0.101	-0.053

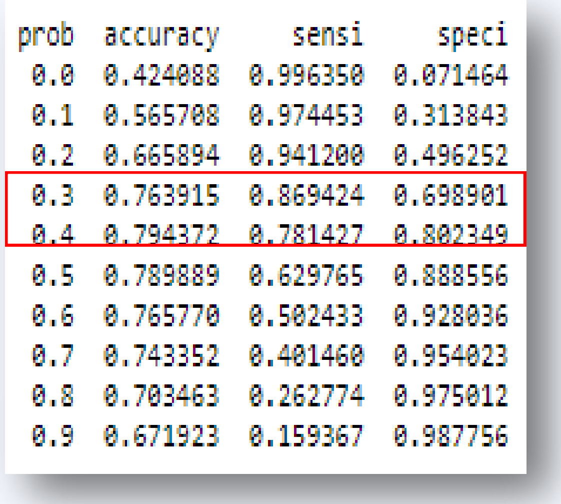
Matrices Evaluation

- We are getting 0.86 area under curve which is quite high and it means that our model is good.

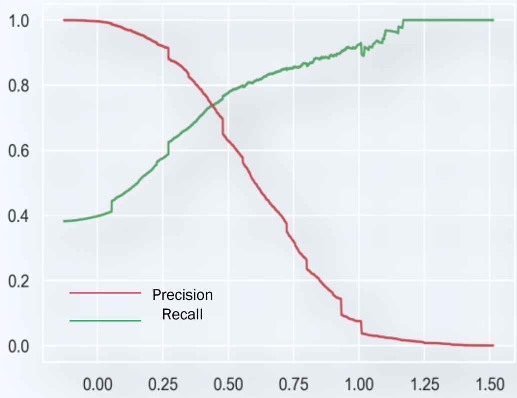
- We can see that value between 0.3 and 0.4 is good for cutoff point.

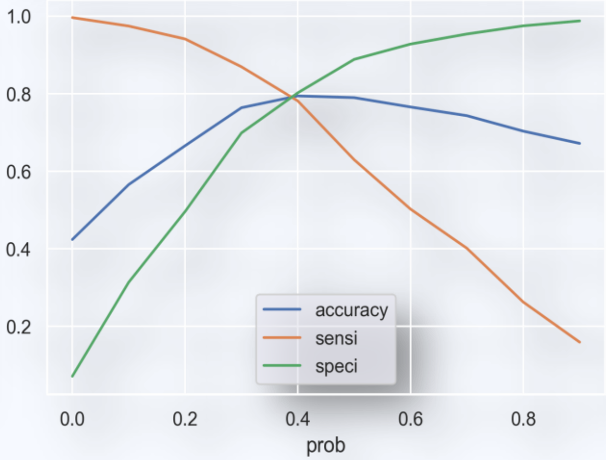
- With the help of accuracy-sensitivity- specificity and precision-recall plot, any value between .35 and .4 will be good, we have choose 0.37 as optimal cutoff point.





```
prob  accuracy     sensi     speci
 0.0  0.424088  0.996350  0.071464
 0.1  0.565708  0.974453  0.313843
 0.2  0.665894  0.941200  0.496252
 0.3  0.763915  0.869424  0.698901
 0.4  0.794372  0.781427  0.802349
 0.5  0.789889  0.629765  0.888556
 0.6  0.765770  0.502433  0.928036
 0.7  0.743352  0.401460  0.954023
 0.8  0.703463  0.262774  0.975012
 0.9  0.671923  0.159367  0.987756
```





Recommendation and Conclusion

Our Model has 80 % sensitivity, Which means it has predictive power 80 %, to tell us who will converted.

The good strategy to employ at this stage to make almost all the potential leads to be converted is to focus on below Continuous and Categories or dummy variables as these features are impacting more on potential lead to be converted.

- Total Time on Website
- Total Visits
- Lead Source with elements Olark Chat
- Last Activity with elements SMS Sent

And not to give more importance on the below Categorical Variables. Because as it’s Coefficient value shows negative values and also these variables have very lower chance to get converted for which you don’t to utilize your effort as our goal is to make most of the customers converted.

- Lead Origin API
- Lead Origin Landing Page Submission
- Lead Origin Lead Import
- Last Activity Email Bounced
- Last Activity Olark Chat Conversation

The below Features that are highly impacted towards the result.

- Total Time on Website
- Total Visits
- Lead Source with elements Olark Chat

The top most 3 Categorical/dummy variables to increase the Probability are:

- Lead Source with elements Olark Chat
- Last Activity with elements SMS Sent
- Last Activity Others.