LEAD SCORING CASE STUDY USING LOGISTIC REGRESSION

HARIHARASUDHAN D

CONTENT

- Problem statement
- Methodology
- EDA
- Model Building
- Model Evaluation
- Observation
- Conclusion

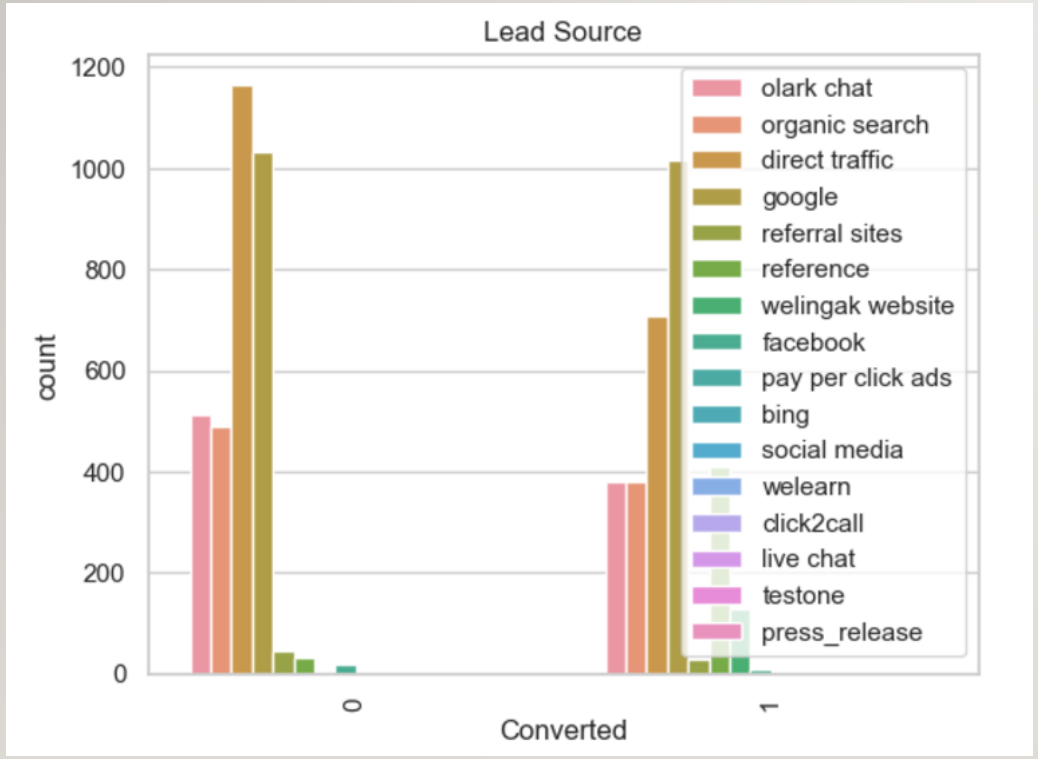
PROBLEM STATEMENT

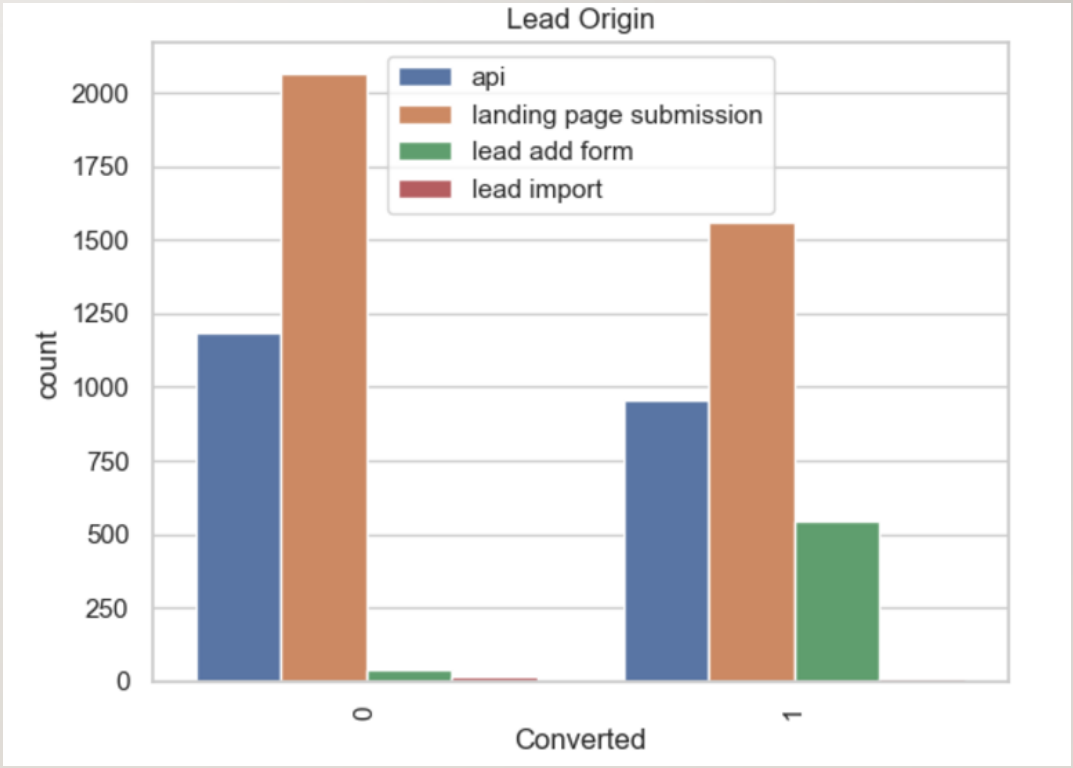
- X Education company sells online courses to industry professionals. Professionals who interested in the course land on their websites. They fill in their details in the website.
- Currently company has lead conversion rate around 30%.
- Company wanted higher lead conversion rate around 80% and understand the which are the factor impacting conversion rate.
- Goal of this case study : build machine learning model , identify the features which helps to improve the lead conversion rate.

METHODOLOGY

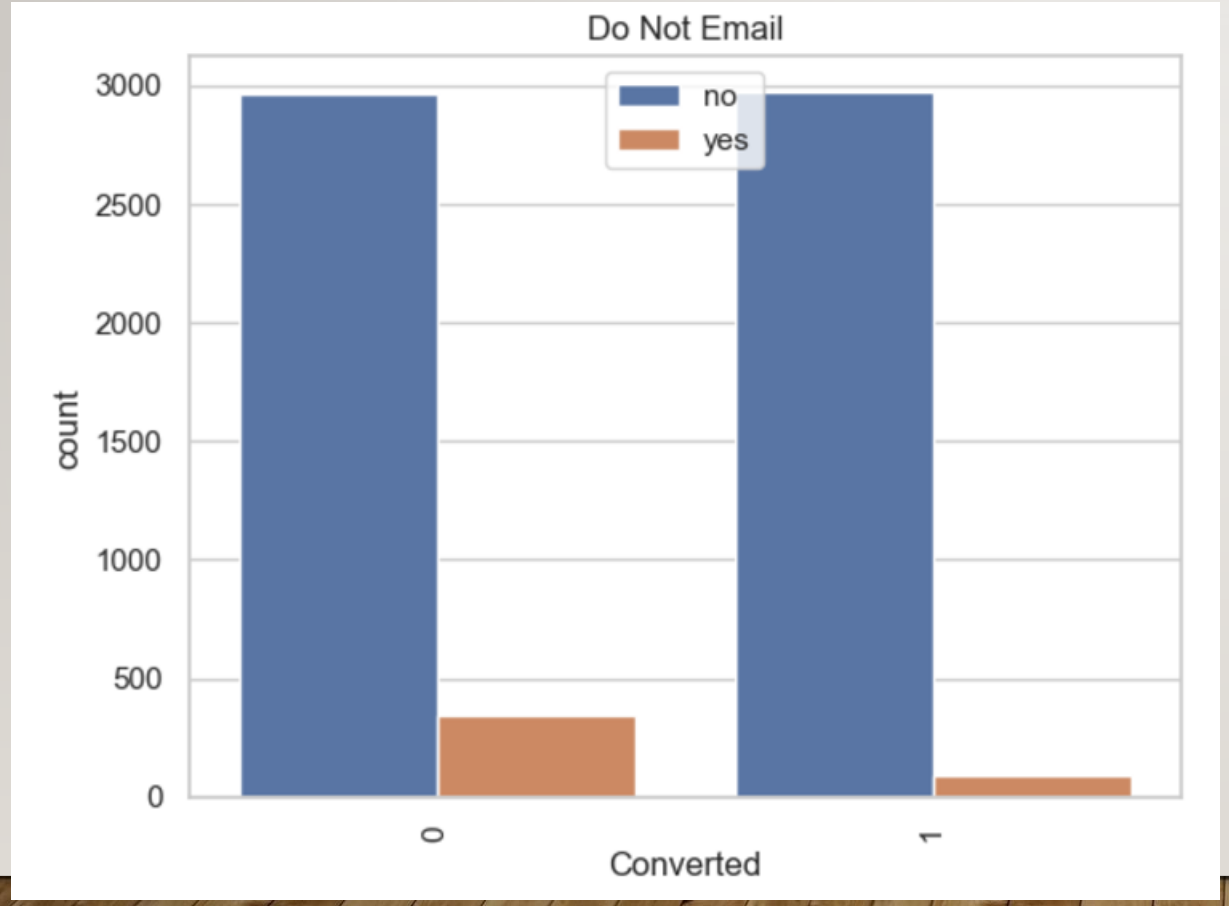
- Data understanding
- Data cleaning
- EDA
- Feature rescaling
- Test – Train split
- Feature selection using RFE
- Model building and iteration (using VIF, P value)
- Model evaluation
- Prediction using test dataset

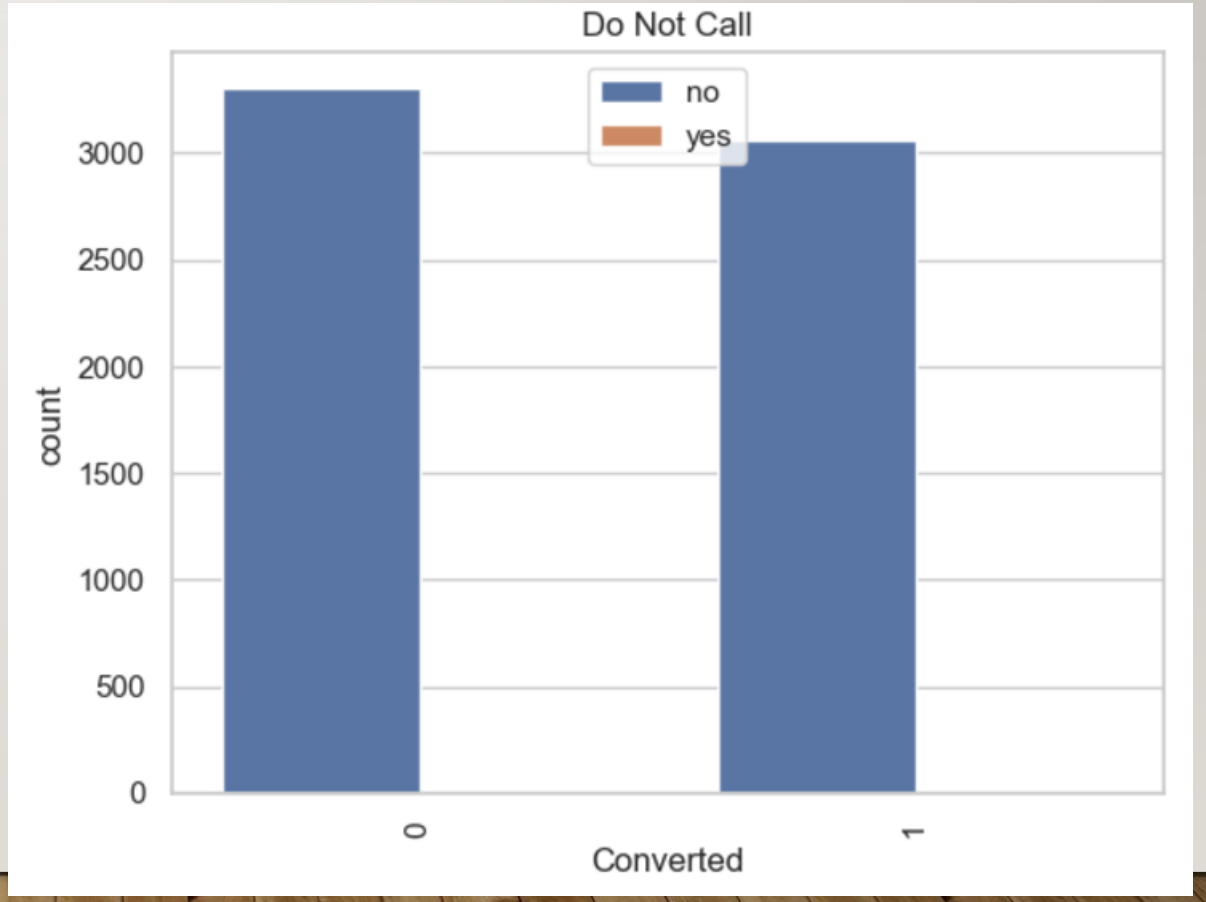
EDA


Lead Source
count
Converted
0
1
0
200
400
600
800
1000
1200
olark chat
organic search
direct traffic
google
referral sites
reference
welingak website
facebook
pay per click ads
bing
social media
welearn
click2call
live chat
testone
press_release

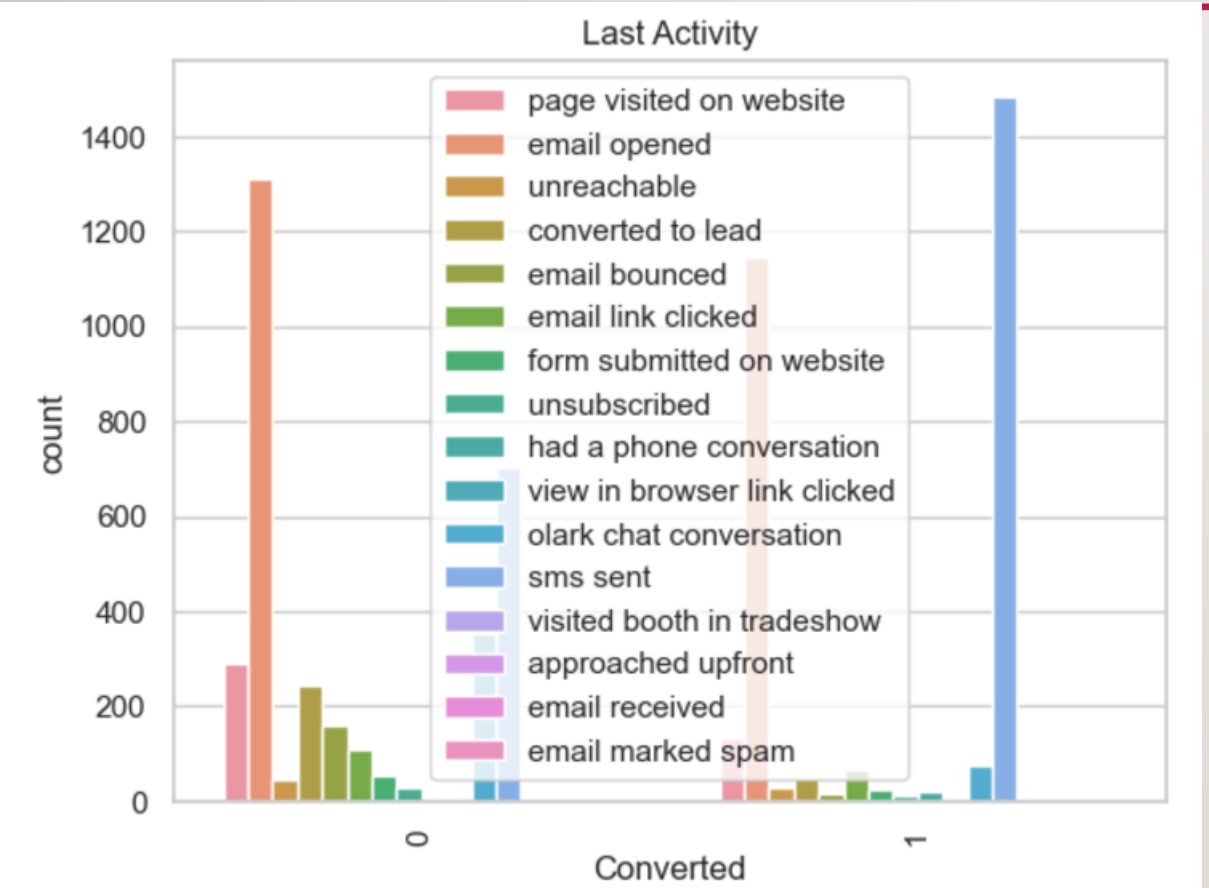

Lead Origin
count
Converted
0
1
0
250
500
750
1000
1250
1500
1750
2000
api
landing page submission
lead add form
lead import

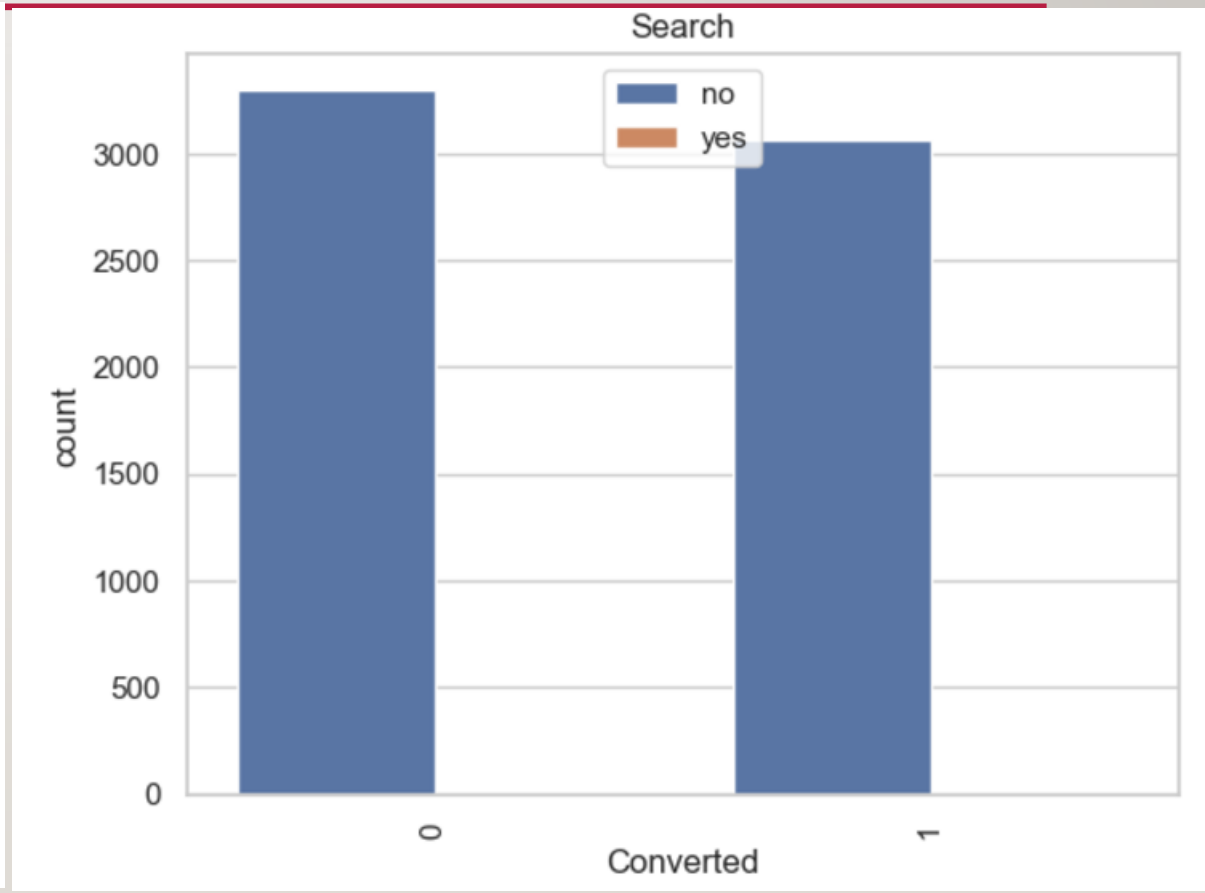
EDA


Do Not Email
no
yes
count
3000
2500
2000
1500
1000
500
0
0
1
Converted


Do Not Call
no
yes
count
3000
2500
2000
1500
1000
500
0
0
1
Converted

EDA


Last Activity
page visited on website
email opened
unreachable
converted to lead
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
olark chat conversation
sms sent
visited booth in tradeshow
approached upfront
email received
email marked spam
count
0
200
400
600
800
1000
1200
1400
0
1
Converted


Search
no
yes
count
0
500
1000
1500
2000
2500
3000
0
1
Converted

LOGISTIC MODEL BUILDING

[304]:

Generalized Linear Model Regression Results

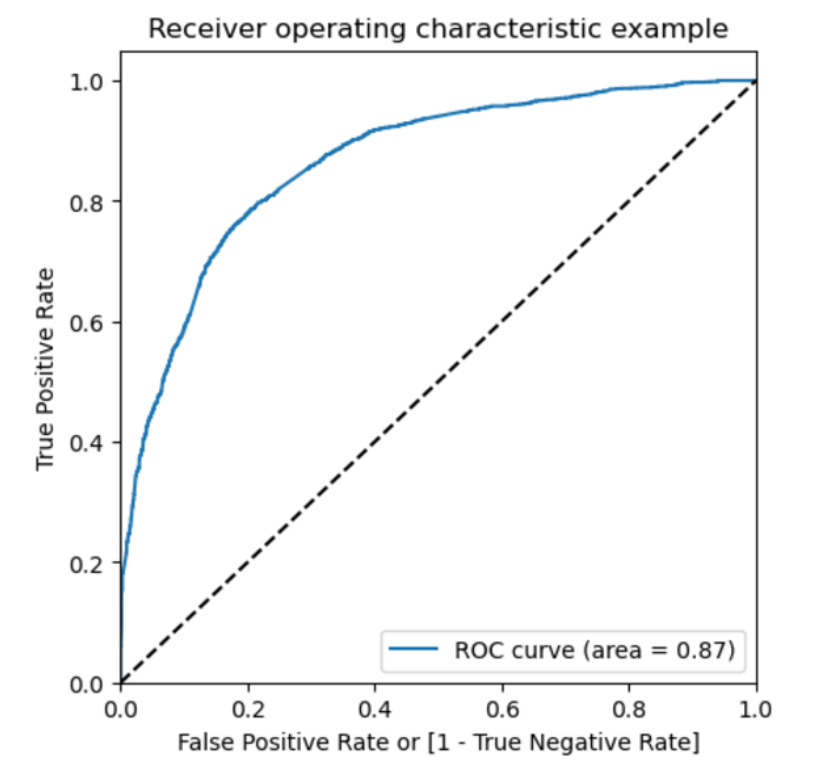
Dep. Variable:	Converted	**No. Observations:**	4460
Model:	GLM	**Df Residuals:**	4447
Model Family:	Binomial	**Df Model:**	12
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2051.2
Date:	Thu, 28 Nov 2024	**Deviance:**	4102.4
Time:	22:19:14	**Pearson chi2:**	4.59e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3718
Covariance Type:	nonrobust		

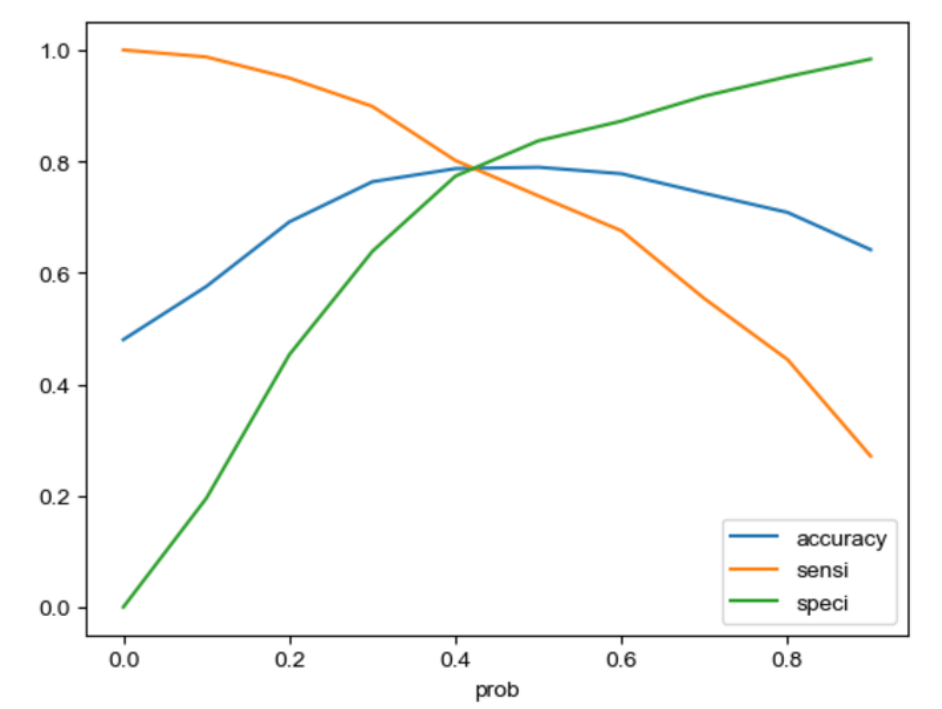
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.4121	0.202	2.043	0.041	0.017	0.807
TotalVisits	8.8860	2.945	3.017	0.003	3.114	14.658
Total Time Spent on Website	4.4755	0.188	23.793	0.000	4.107	4.844
Lead Source_olark chat	1.5094	0.125	12.111	0.000	1.265	1.754
Lead Source_reference	3.9999	0.248	16.132	0.000	3.514	4.486
Lead Source_welingak website	6.2263	1.012	6.152	0.000	4.243	8.210
Do Not Email_yes	-1.3425	0.190	-7.067	0.000	-1.715	-0.970
Last Activity_had a phone conversation	2.8211	0.802	3.517	0.000	1.249	4.393
Last Activity_sms sent	1.0070	0.084	12.000	0.000	0.842	1.171
What is your current occupation_student	-2.4888	0.286	-8.716	0.000	-3.048	-1.929
What is your current occupation_unemployed	-2.4332	0.187	-13.039	0.000	-2.799	-2.067
Last Notable Activity_modified	-0.8502	0.090	-9.445	0.000	-1.027	-0.674

	Features	VIF
9	What is your current occupation_unemployed	3.42
1	Total Time Spent on Website	2.03
0	TotalVisits	1.58
7	Last Activity_sms sent	1.57
10	Last Notable Activity_modified	1.49
2	Lead Source_olark chat	1.35
3	Lead Source_reference	1.13
5	Do Not Email_yes	1.09
4	Lead Source_welingak website	1.08
8	What is your current occupation_student	1.08
6	Last Activity_had a phone conversation	1.01
11	Last Notable Activity_unreachable	1.01

- Splitting dataset into train and test with 70:30 ratio.
- RFE used for feature selection. Started with 15 features, then using high VIF , high P values. It is reduced to 10 features.
- Model build with VIF less than 5 and P value less than 0.005.
- Accuracy = 79%

MODEL EVALUATION





ROC curve – AUC = 0.87 which shows it is good model.

Cut off optimization done using Accuracy, Sensitivity, Specificity.

New cut off – 0.42.

OBSERVATION

Train data set
Accuracy – 79%
Sensitivity – 79%
Specificity – 79%

Test dataset
Accuracy – 76%
Sensitivity – 92%
Specificity – 67%

	Features
9	What is your current occupation_unemployed
1	Total Time Spent on Website
0	TotalVisits
7	Last Activity_sms sent
10	Last Notable Activity_modified
2	Lead Source_olark chat
3	Lead Source_reference
5	Do Not Email_yes
4	Lead Source_welingak website
8	What is your current occupation_student
6	Last Activity_had a phone conversation
11	Last Notable Activity_unreachable

CONCLUSION

- Features plays important to help lead conversion. (descending order)
- Total time spend on website
- Lead source – olark chat, reference, welingak website.
- Google, direct traffic has more leads. But, conversion is less when comparing with above feature.
- Last Activities on phone conversation, Last activity sms sent

- X Education should concentrate on above features to improve the lead conversion rate.