Lead Score Case Study

By:

Divya Khandelwal
Zeba Afroz
Apurv Pandey

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.

- Now, although 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.

- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you 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%.

Data

- You have been provided with a lead's dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.

- The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

- You can learn more about the dataset from the data dictionary provided in the zip folder at the end of the page. Another thing that you also need to check out for are the levels present in the categorical variables.

- Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value (think why?).

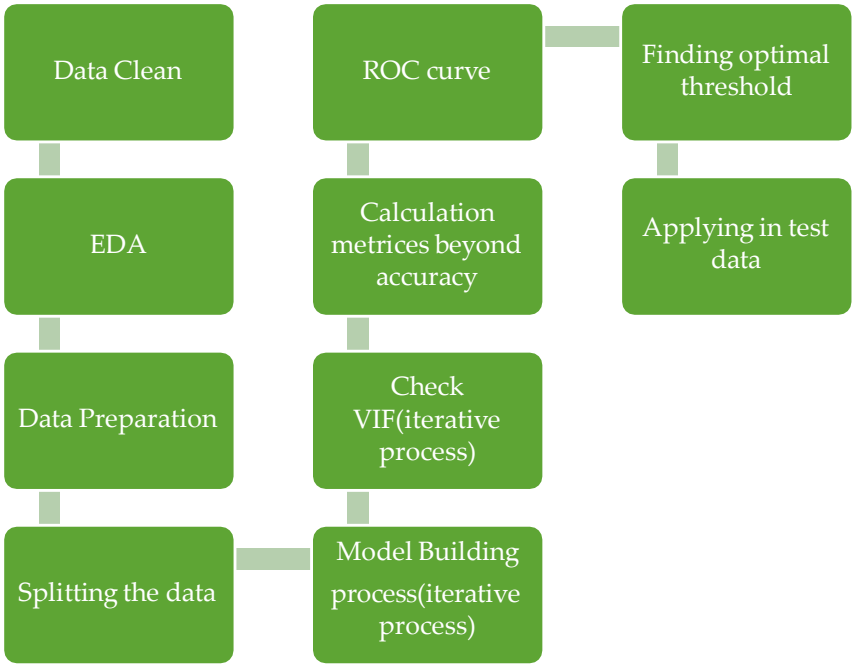
Goals of the case study:

There are quite a few goals for this case study.

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Process


Data Clean
EDA
Data Preparation
Splitting the data
Model Building process(iterative process)
Check VIF(iterative process)
Calculation metrices beyond accuracy
ROC curve
Finding optimal threshold
Applying in test data

Cleaning:

Cleaning the data includes :

- Dealing with missing values.
- Update datatype as per the column if required.
- Change ‘Select” with “NULL”.

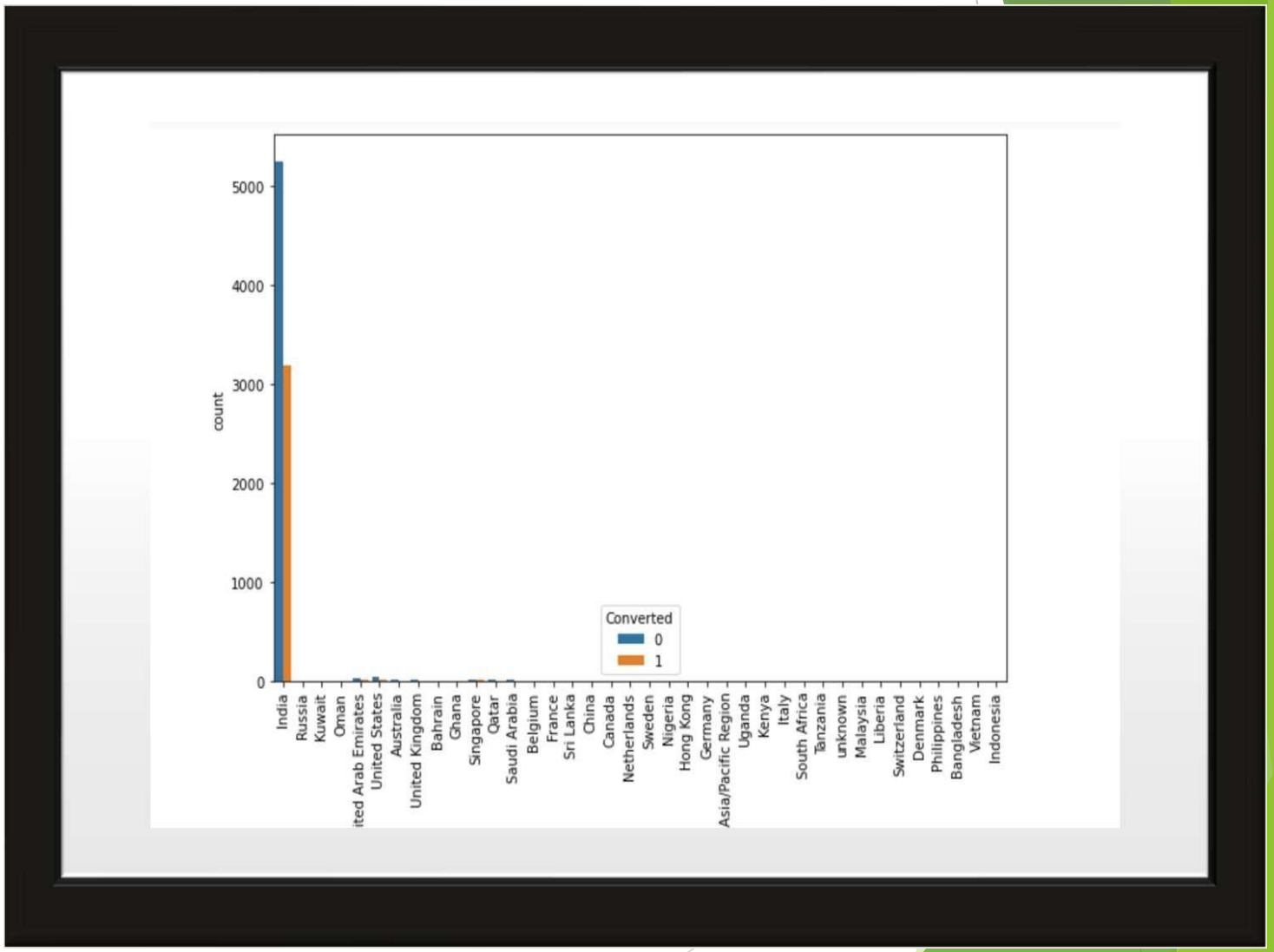
EDA (Univariate and Bi-variate analysis)

- We can perform univariate and bivariate on the data check how the others columns are related with targeted columns.

- As per our observation there are more than 10 columns which are related with targeted column but we have performed univariate on only 10 columns.

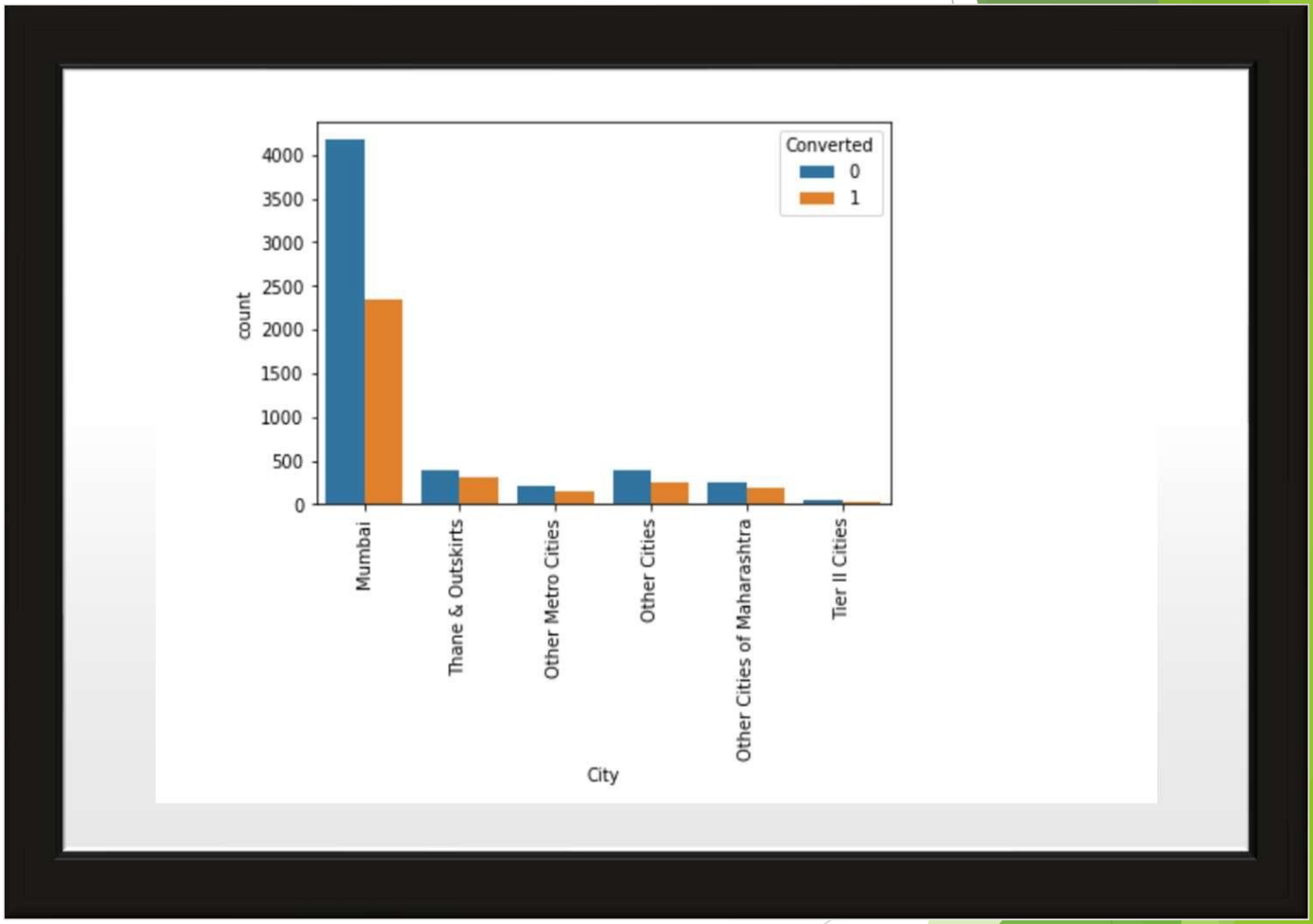
Country:

India is showing the highest result in this



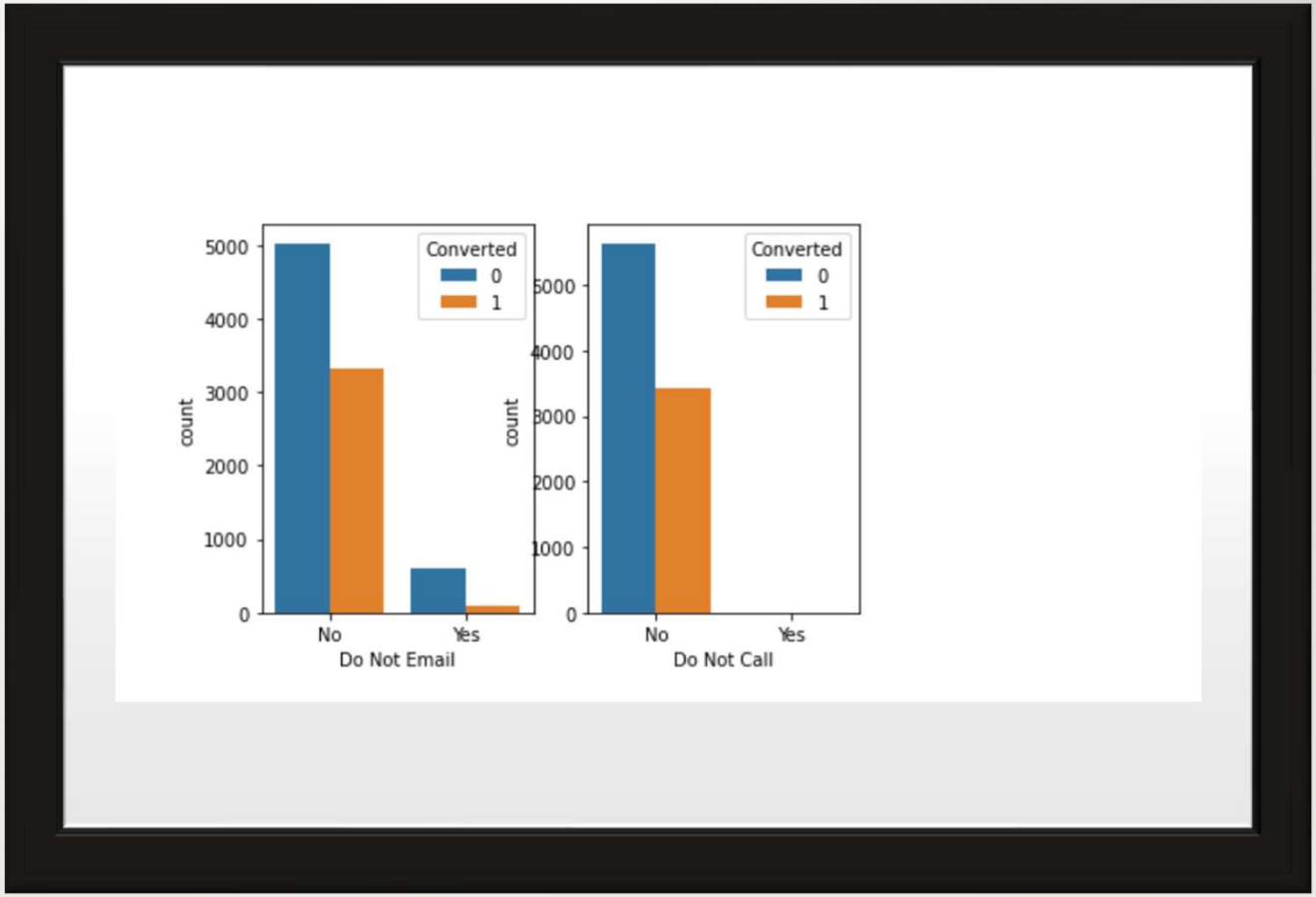
City:

Mumbai is likely to get converted as compared to others.



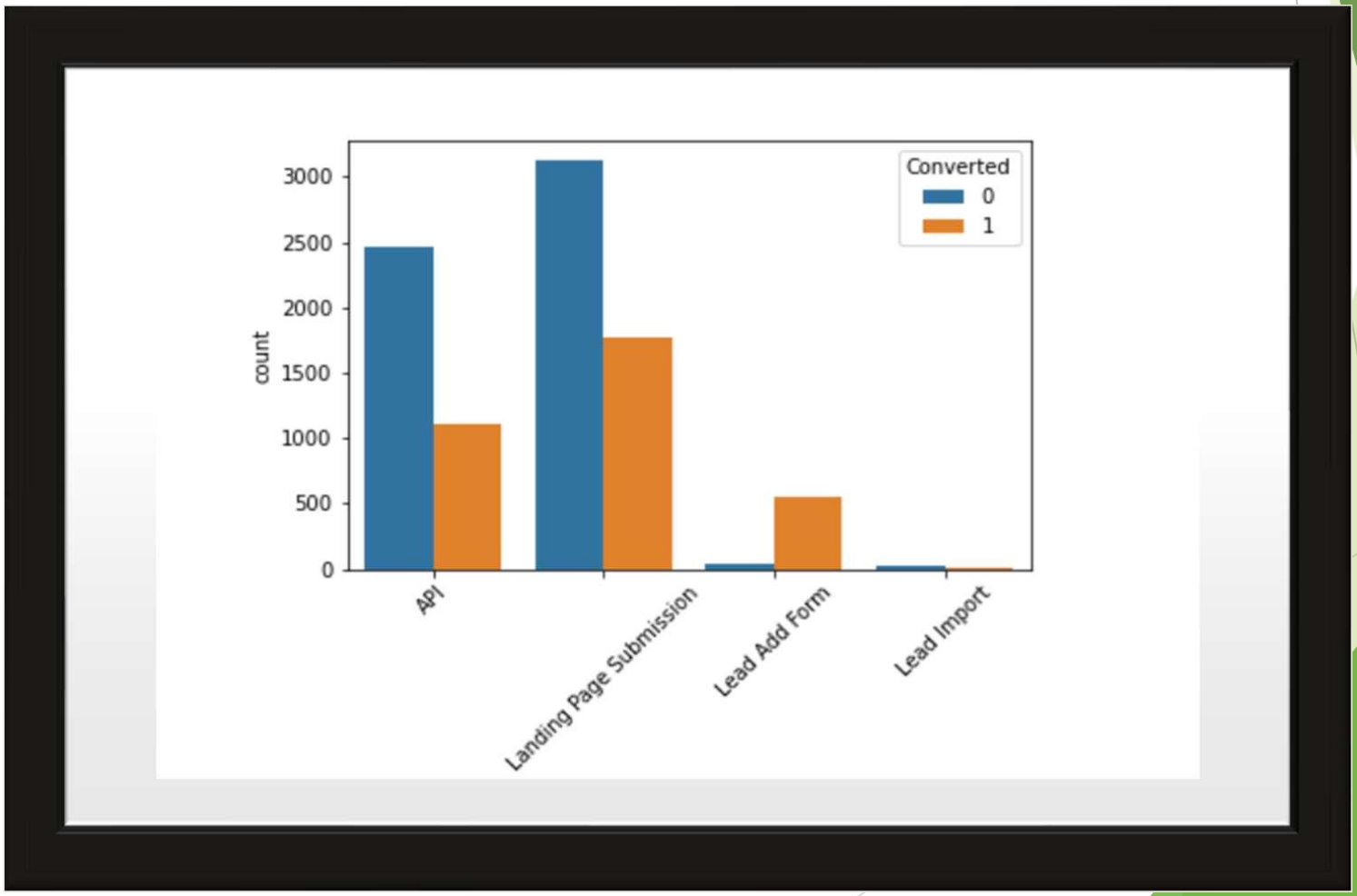
Do not call and Do not mail:

The people who are opting for no for these 2 services are most likely to get converted.



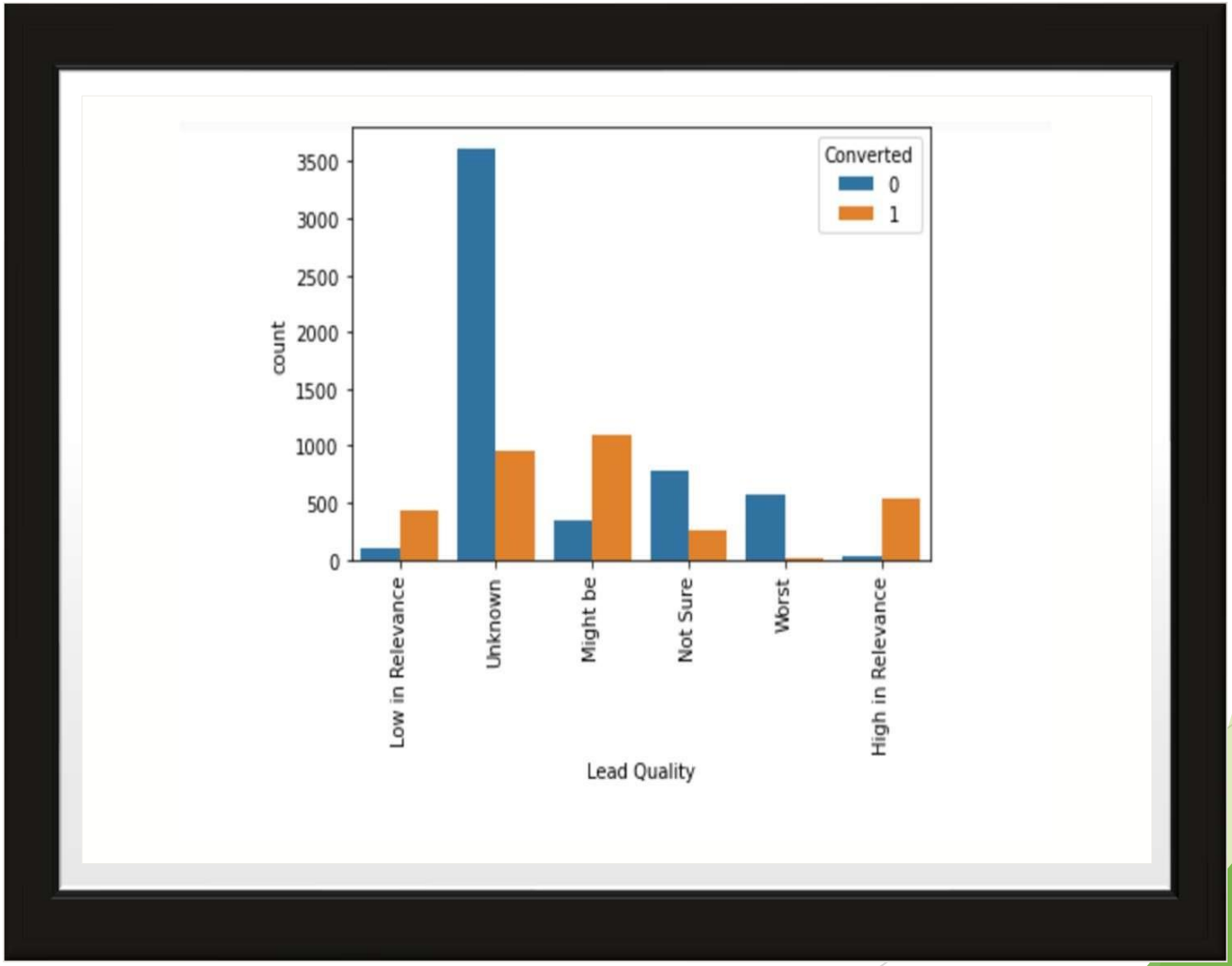
Lead Origin:

"Landing page Submission" origin is most likely to get converted.



Lead Quality:

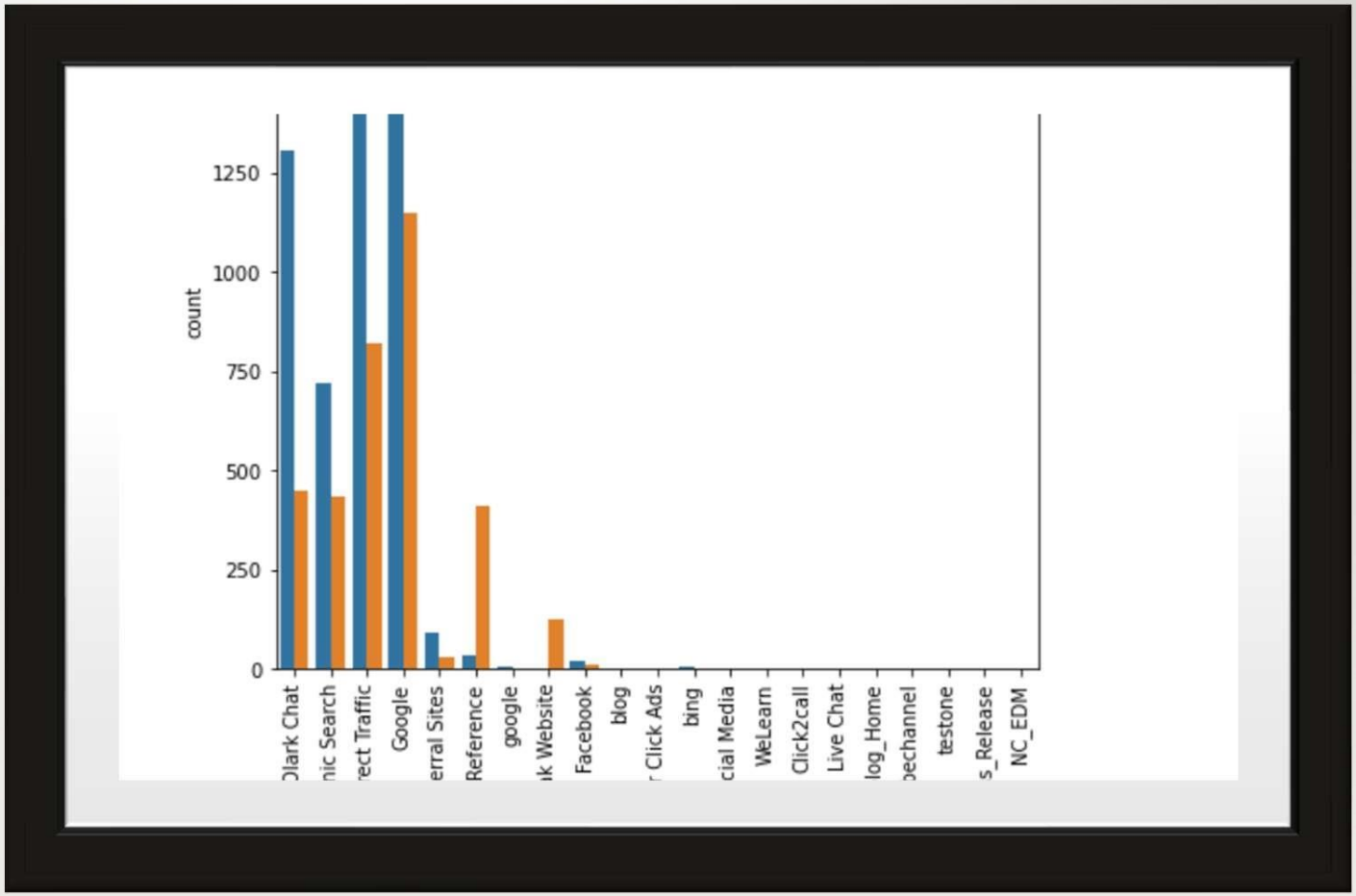
Apart from unknown “Might be” quality is most likely to get converted.



Lead source:

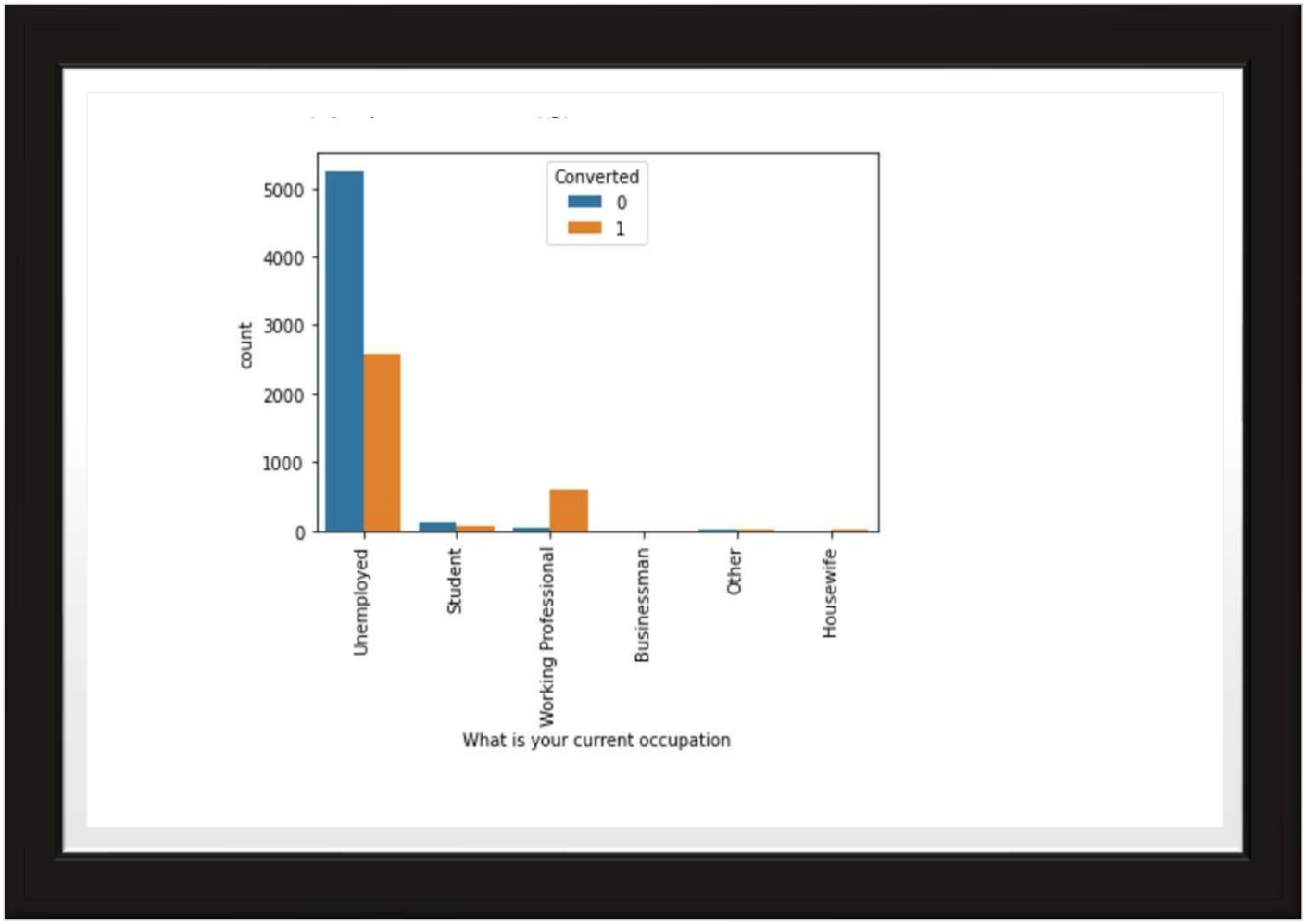
"Direct Traffic" and "Google" source are most likely to get converted compared to others.

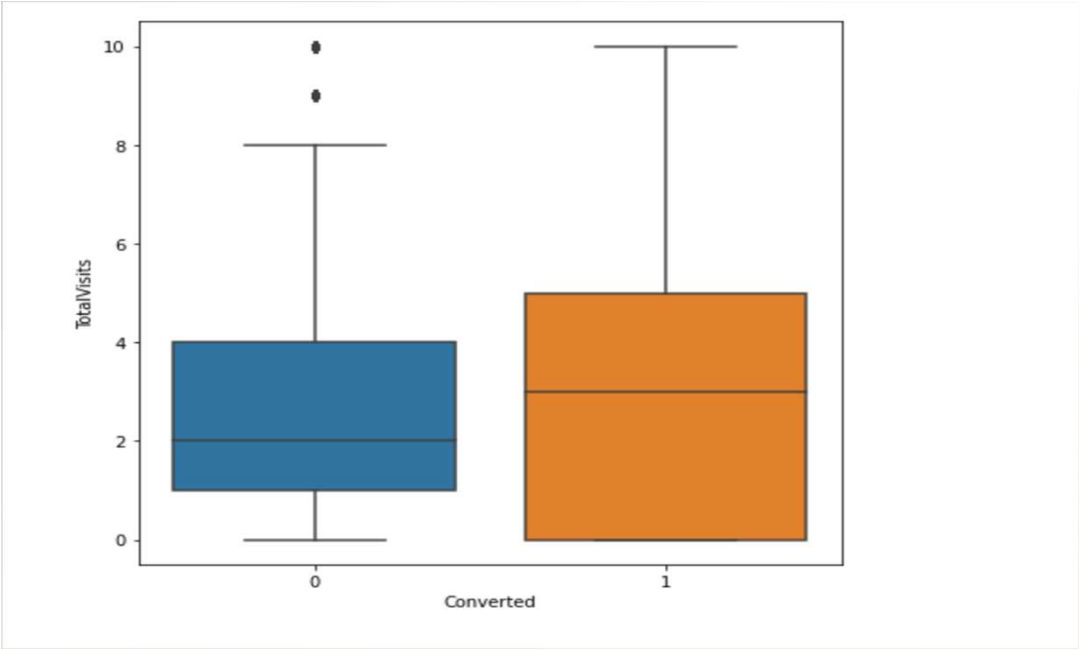
But "Reference" is shoing more positive conversion as compared to negative conversion.

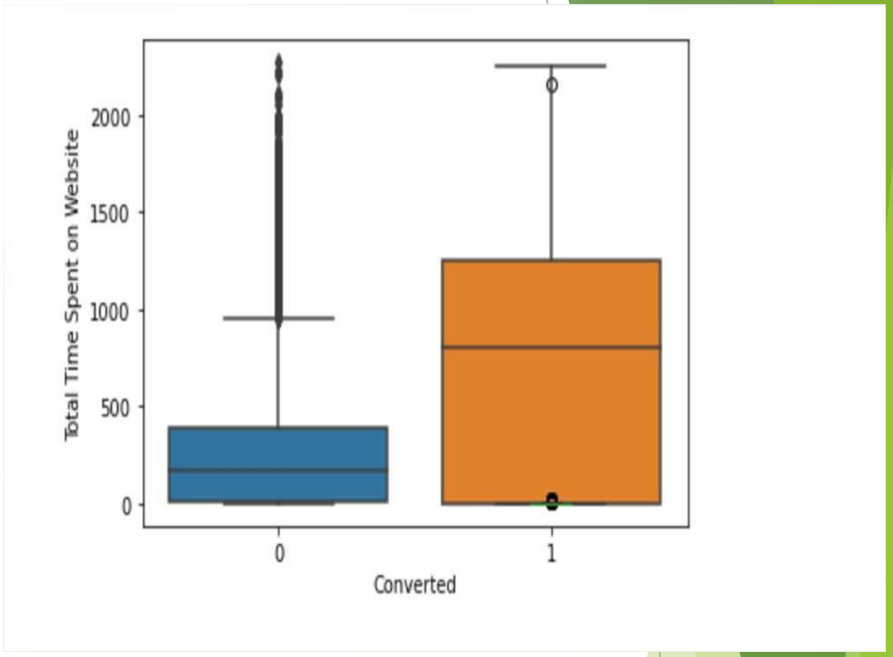


What is your current occupation?

- Unemployed people are most likely to get converted .
- Working professionals are shows more positive conversion than negative







Total time spend and Total visits

Data Preparation:

- After EDA Data preparation is done by converting categorical variable into binary variable or dummies variable.

Splitting the data:

- Once data is prepared, we split the data using test_train_split with 70:30 mor 80:20 ratio.

- Out of which 70/80 would be considered as train data and 30/20 is considered as test data.

Splitting the data:

- Once data is prepared, we split the data using test_train_split with 70:30 mor 80:20 ratio.

- Out of which 70/80 would be considered as train data and 30/20 is considered as test data.

Model Creation:

Model number 3 ➔

	coef	std err	z	P>|z|	[0.025	0.975]
const	-5.9243	1.026	-5.776	0.000	-7.935	-3.914
Lead Number	7.164e-06	1.62e-06	4.434	0.000	4e-06	1.03e-05
Total Time Spent on Website	0.8078	0.040	20.126	0.000	0.729	0.886
Country_India	0.6089	0.220	2.771	0.006	0.178	1.040
Lead Source_Olark Chat	0.4355	0.108	4.050	0.000	0.225	0.646
Last Activity_Olark Chat Conversation	-1.3251	0.176	-7.546	0.000	-1.669	-0.981
Last Activity_SMS Sent	0.6612	0.150	4.394	0.000	0.366	0.956
What is your current occupation_Unemployed	-0.6154	0.214	-2.880	0.004	-1.034	-0.197
What is your current occupation_Working Professional	2.0926	0.288	7.271	0.000	1.529	2.657
Tags_Ringing	-3.3502	0.235	-14.256	0.000	-3.811	-2.890
Tags_Will revert after reading the email	1.0302	0.084	12.226	0.000	0.865	1.195
Lead Quality_Might be	1.4195	0.118	12.022	0.000	1.188	1.651
Last Notable Activity_Modified	-0.7328	0.094	-7.802	0.000	-0.917	-0.549
Last Notable Activity_SMS Sent	1.0127	0.179	5.647	0.000	0.661	1.364
Asymmetrique Profile Index_02.Medium	-0.1711	0.083	-2.064	0.039	-0.334	-0.009

VIF calculation for model 3:

	Features	VIF
0	Lead Number	67.11
6	What is your current occupation_Unemployed	33.60
2	Country_India	31.55
12	Last Notable Activity_SMS Sent	6.51
5	Last Activity_SMS Sent	6.28
7	What is your current occupation_Working Profes...	3.68
9	Tags_Will revert after reading the email	3.45
11	Last Notable Activity_Modified	2.50
3	Lead Source_Olark Chat	1.90
4	Last Activity_Olark Chat Conversation	1.64
13	Asymmetrique Profile Index_02.Medium	1.59
8	Tags_Ringing	1.56
10	Lead Quality_Might be	1.52
1	Total Time Spent on Website	1.23

Final Model:

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.8878	0.213	-4.172	0.000	-1.305	-0.471
Total Time Spent on Website	0.7983	0.040	20.023	0.000	0.720	0.876
Lead Source_Olark Chat	0.3692	0.104	3.553	0.000	0.166	0.573
Last Activity_Olark Chat Conversation	-1.2179	0.174	-6.995	0.000	-1.559	-0.877
Last Activity_SMS Sent	1.3215	0.081	16.231	0.000	1.162	1.481
What is your current occupation_Unemployed	-0.5814	0.213	-2.732	0.006	-0.998	-0.164
What is your current occupation_Working Professional	2.0938	0.289	7.247	0.000	1.528	2.660
Tags_Ringing	-3.2589	0.233	-13.996	0.000	-3.715	-2.803
Tags_Will revert after reading the email	1.0413	0.083	12.518	0.000	0.878	1.204
Lead Quality_Might be	1.4565	0.117	12.461	0.000	1.227	1.686
Last Notable Activity_Modified	-1.0231	0.080	-12.723	0.000	-1.181	-0.866

Final model VIF:

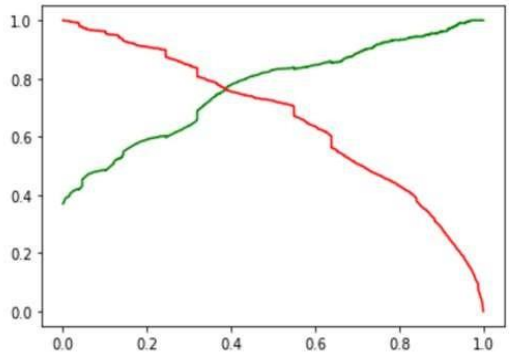
	Features	VIF
4	What is your current occupation_Unemployed	4.81
7	Tags_Will revert after reading the email	3.34
9	Last Notable Activity_Modified	1.88
1	Lead Source_Olark Chat	1.83
3	Last Activity_SMS Sent	1.71
2	Last Activity_Olark Chat Conversation	1.60
6	Tags_Ringing	1.54
5	What is your current occupation_Working Profes...	1.50
8	Lead Quality_Might be	1.50
0	Total Time Spent on Website	1.23

Calculating Precision and Recall tradeoff:

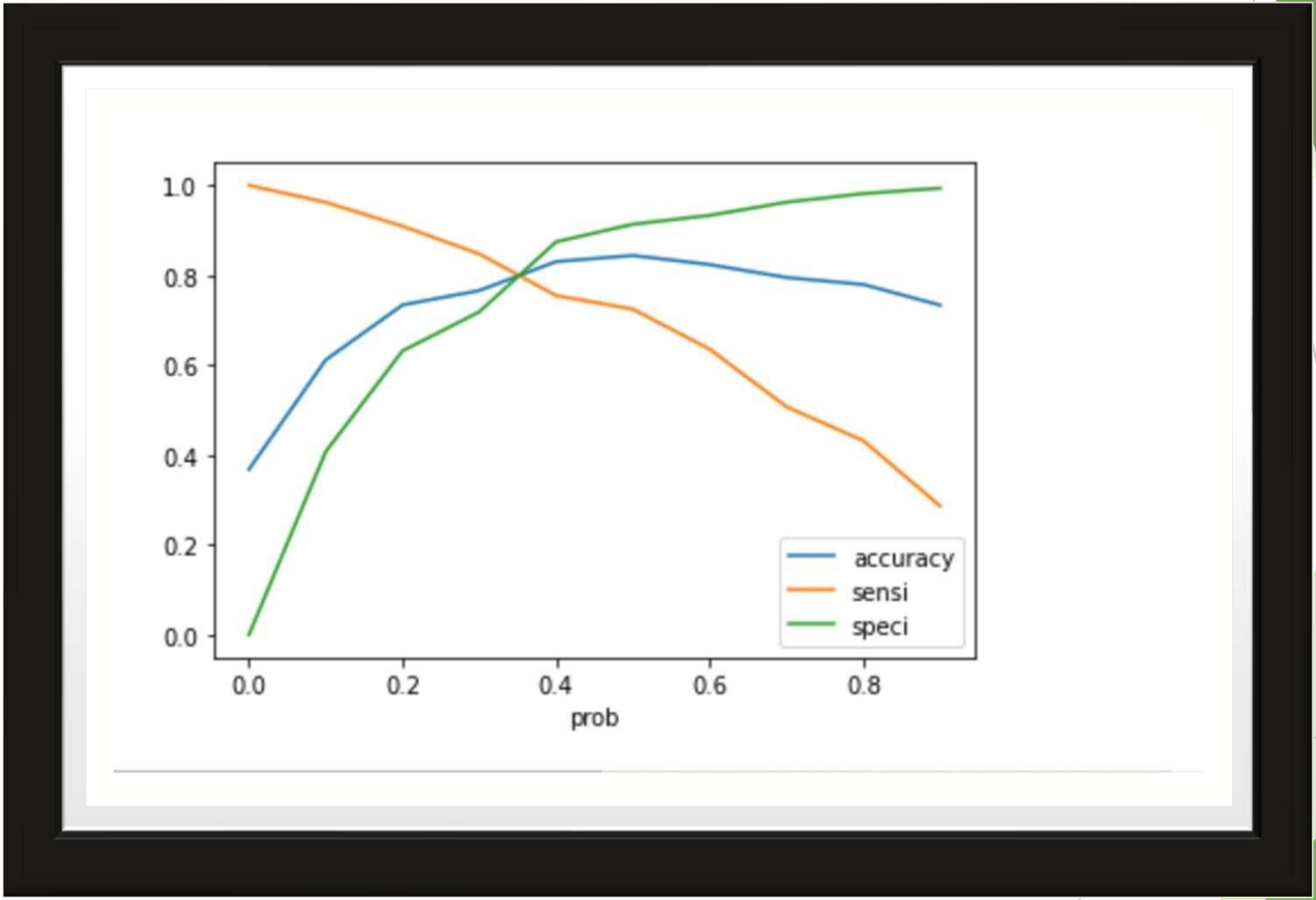
Precision and Recall tradeoff

```
]:  1 from sklearn.metrics import precision_recall_curve
    2
    3 p,r,thresholds=precision_recall_curve(y_train_pv_final.Converted,y_train_pv_final['Conversion Predicted'])
    4
```

```
]:  1 plt.plot(thresholds,p[:-1],"g-")
    2 plt.plot(thresholds,r[:-1],"r-")
    3 plt.show()
```



Threshold calculation:



Using Train model into test

Test Data: Final

	Lead ID	Converted	Converted_prob	Final_predicted
0	926	0	0.467135	1
1	6289	0	0.017222	0
2	330	0	0.012022	0
3	3904	0	0.431329	1
4	4391	0	0.012439	0

Final models

values:

```
1  Sensi=TP/float(TP+FN)
2  Speci=TN/float(TN+FP)
3  print(Sensi)
4  print(Speci)
```

```
0.9013671875
0.6202134337727558
```

Summary

- We have use sklearn and statsmodels library to build this model which has high sensitivity and less specificity which will result in covering almost all the true positives.

- We can change threshold value as per the client requirement .For example: If client don't want to make unnecessary calls because the target is already achieved so we can simply increase the value of threshold which will result in increase of specificity.

- In this model the threshold use was 0.34.

- Overall accuracy of the model is also good hence we can consider that this model fits perfect with pvalue less than 0.05 and vif less than 5.

Thank You!