Lead Scoring Case Study

Authors: Ashik Saibabu T, Saurabh Agrawal, Nobin Bera

Contents

Problem Statement

X Education is an online course selling company. They have collected potential leads through various channels. But at present their lead conversion rate is roughly 30% which is very low and the business is spending lots of money for lead follow ups. We need to create a machine learning solution which will increase the lead conversion rate by identifying the potential leads and non-potential leads. This will save the company money on lead follow up activities and potentially give the business an understanding on the profile of a lead who might purchase a course from X Education.

Data Analysis

Our dataset contains 9240 data points each with 36 features

```
Data columns (total 37 columns):
 #   Column                                         Non-Null Count  Dtype
---  ------                                         --------------  -----
 0   Prospect ID                                    9240 non-null   object
 1   Lead Number                                    9240 non-null   int64
 2   Lead Origin                                    9240 non-null   object
 3   Lead Source                                    9204 non-null   object
 4   Do Not Email                                   9240 non-null   object
 5   Do Not Call                                    9240 non-null   object
 6   Converted                                      9240 non-null   int64
 7   TotalVisits                                    9103 non-null   float64
 8   Total Time Spent on Website                    9240 non-null   int64
 9   Page Views Per Visit                           9103 non-null   float64
 10  Last Activity                                  9137 non-null   object
 11  Country                                        6779 non-null   object
 12  Specialization                                 7802 non-null   object
 13  How did you hear about X Education             7033 non-null   object
 14  What is your current occupation                6550 non-null   object
```

Data Analysis

Most of the features are categorical values

	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Asymmetrique Activity Score	Asymmetrique Profile Score
count	9240.000000	9240.000000	9103.000000	9240.000000	9103.000000	5022.000000	5022.000000
mean	617188.435606	0.385390	3.445238	487.698268	2.362820	14.306252	16.344883
std	23405.995698	0.486714	4.854853	548.021466	2.161418	1.386694	1.811395
min	579533.000000	0.000000	0.000000	0.000000	0.000000	7.000000	11.000000
25%	596484.500000	0.000000	1.000000	12.000000	1.000000	14.000000	15.000000
50%	615479.000000	0.000000	3.000000	248.000000	2.000000	14.000000	16.000000
75%	637387.250000	1.000000	5.000000	936.000000	3.000000	15.000000	18.000000
max	660737.000000	1.000000	251.000000	2272.000000	55.000000	18.000000	20.000000

Data Analysis

Removed unwanted features and features with ~0 variance

> Project ID, Lead Number
> Do not call, What matters most to you when choosing a course, Magazine, Newspaper Articles, 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, Search

Data Analysis

> Many categorical columns has labels with very less percentage of occurrence

> Grouped those items to single category to avoid large number of dummy variables

Data Analysis

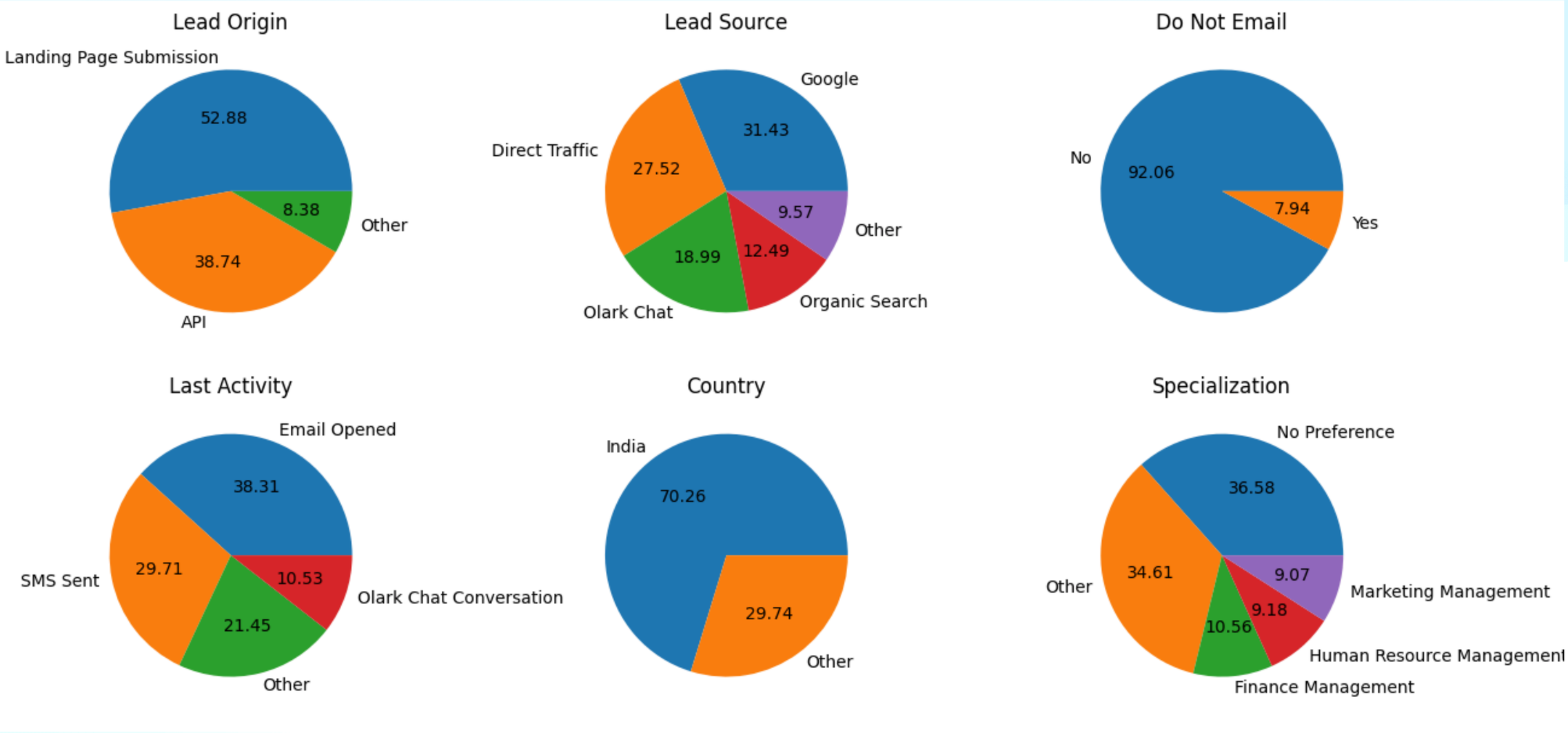
> Country has majority items in India. Grouped all other countries together

Country	
India	70.259740
United States	0.746753
United Arab Emirates	0.573593
Singapore	0.259740
Saudi Arabia	0.227273
United Kingdom	0.162338
Australia	0.140693
Qatar	0.108225
Hong Kong	0.075758
Bahrain	0.075758
Oman	0.064935
France	0.064935

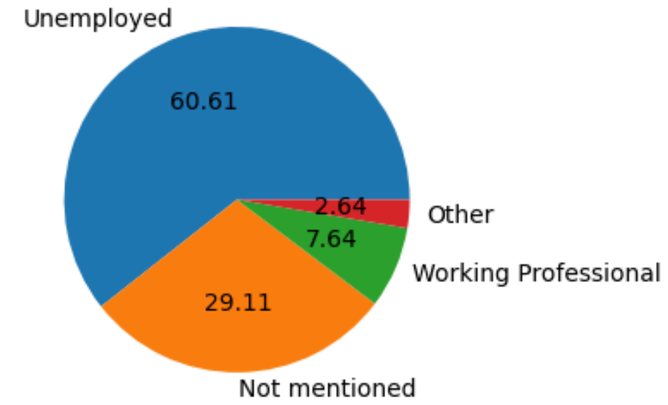
Data Analysis

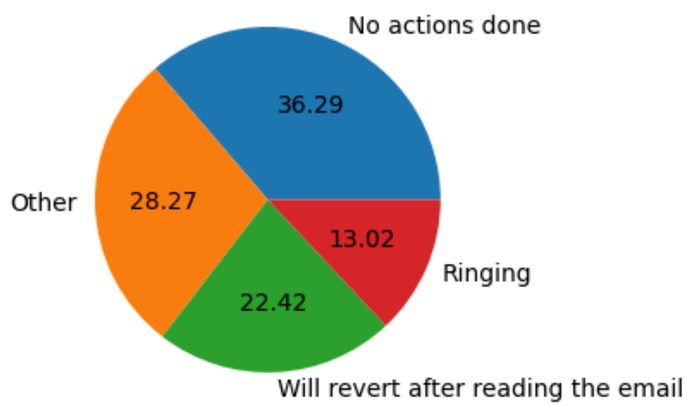
> Columns with more than 40% null values are dropped
> Other null values are imputed with central tendency measures
> In case of ‘City’, since it can be grouped with country we used central tendency measures of each group.
> We could identify that even though country is not India, many cities are assigned as Mumbai. This should be further analysed on the user input side.

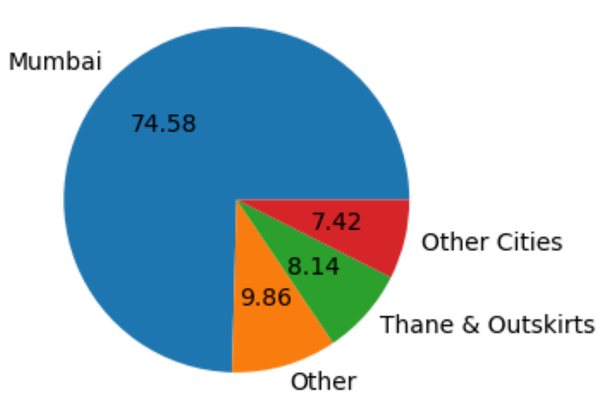
Data Analysis

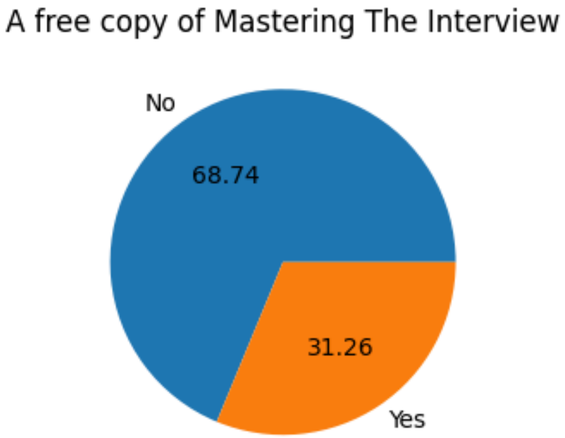


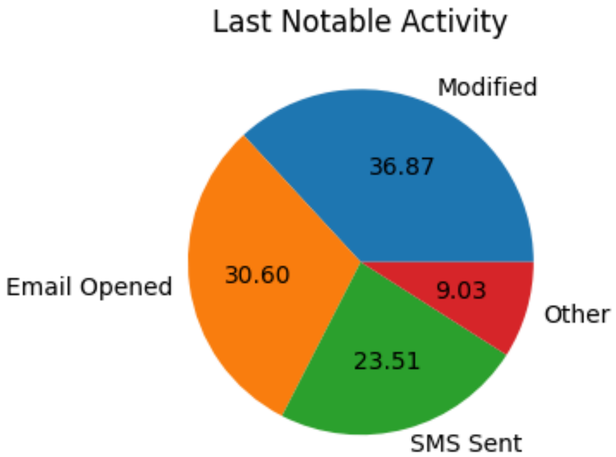
Data Analysis

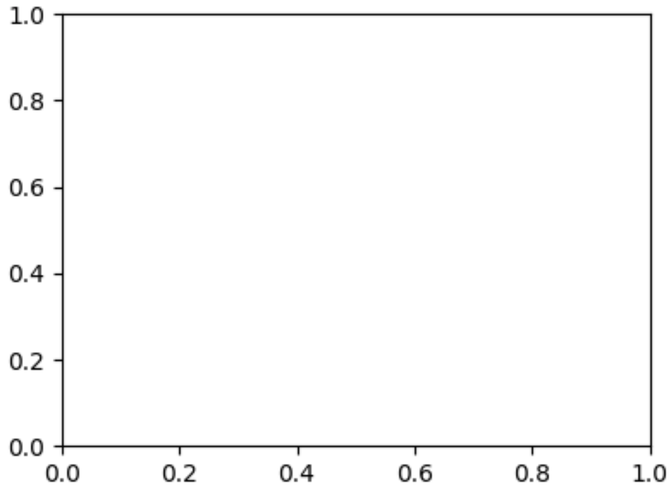






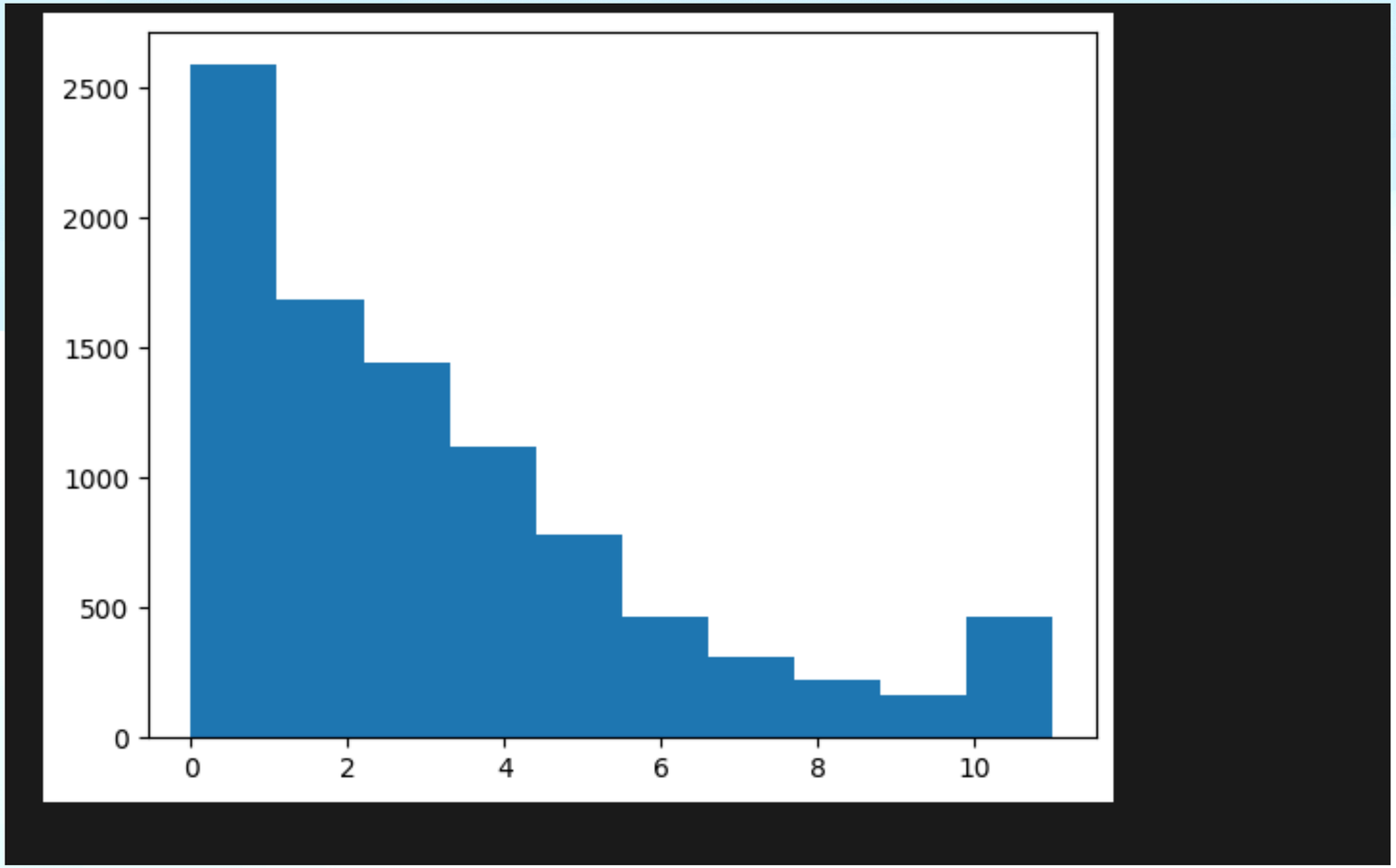






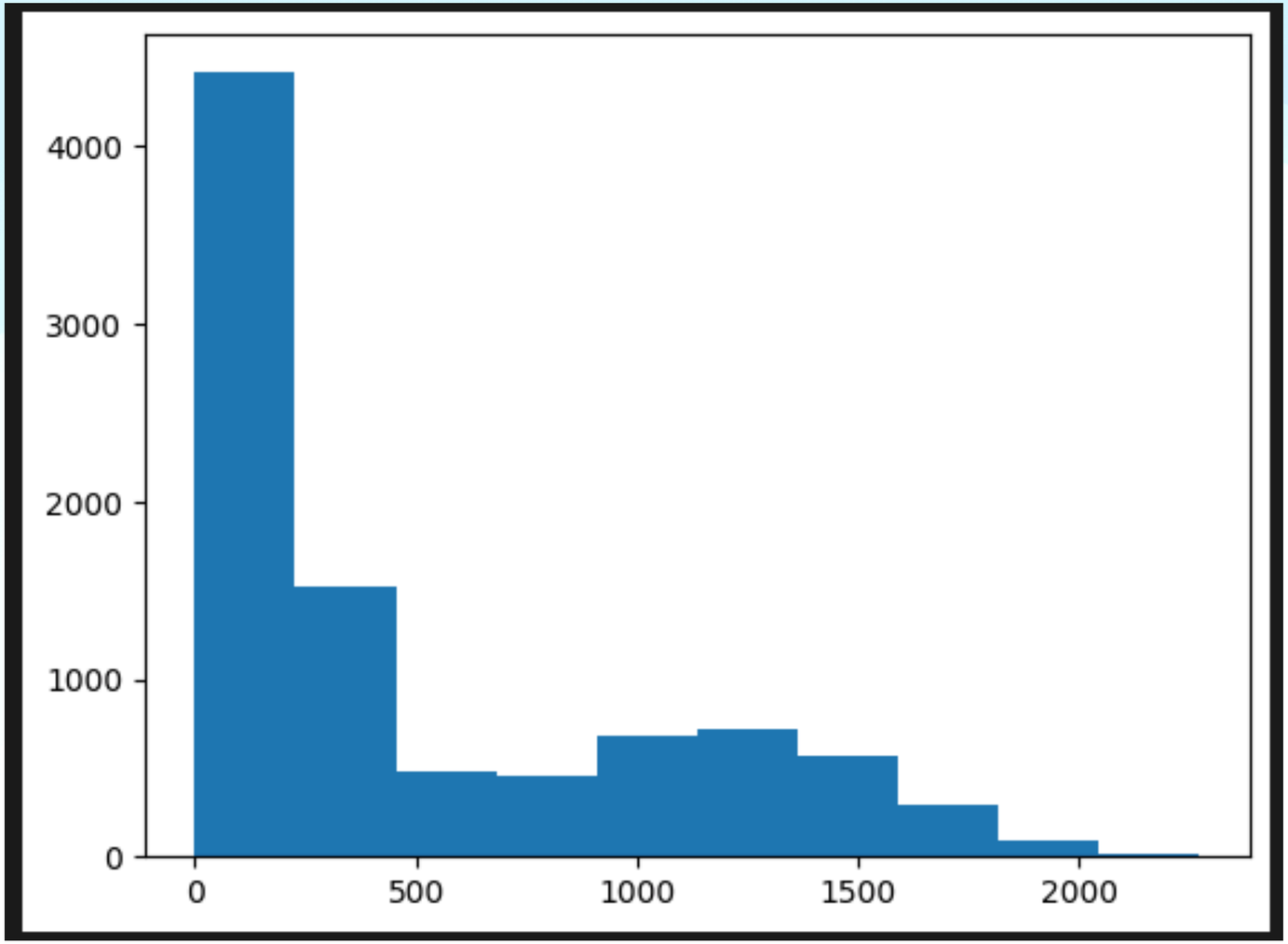
Data Analysis

Most Total Visits are under 5 and is continuously decreasing



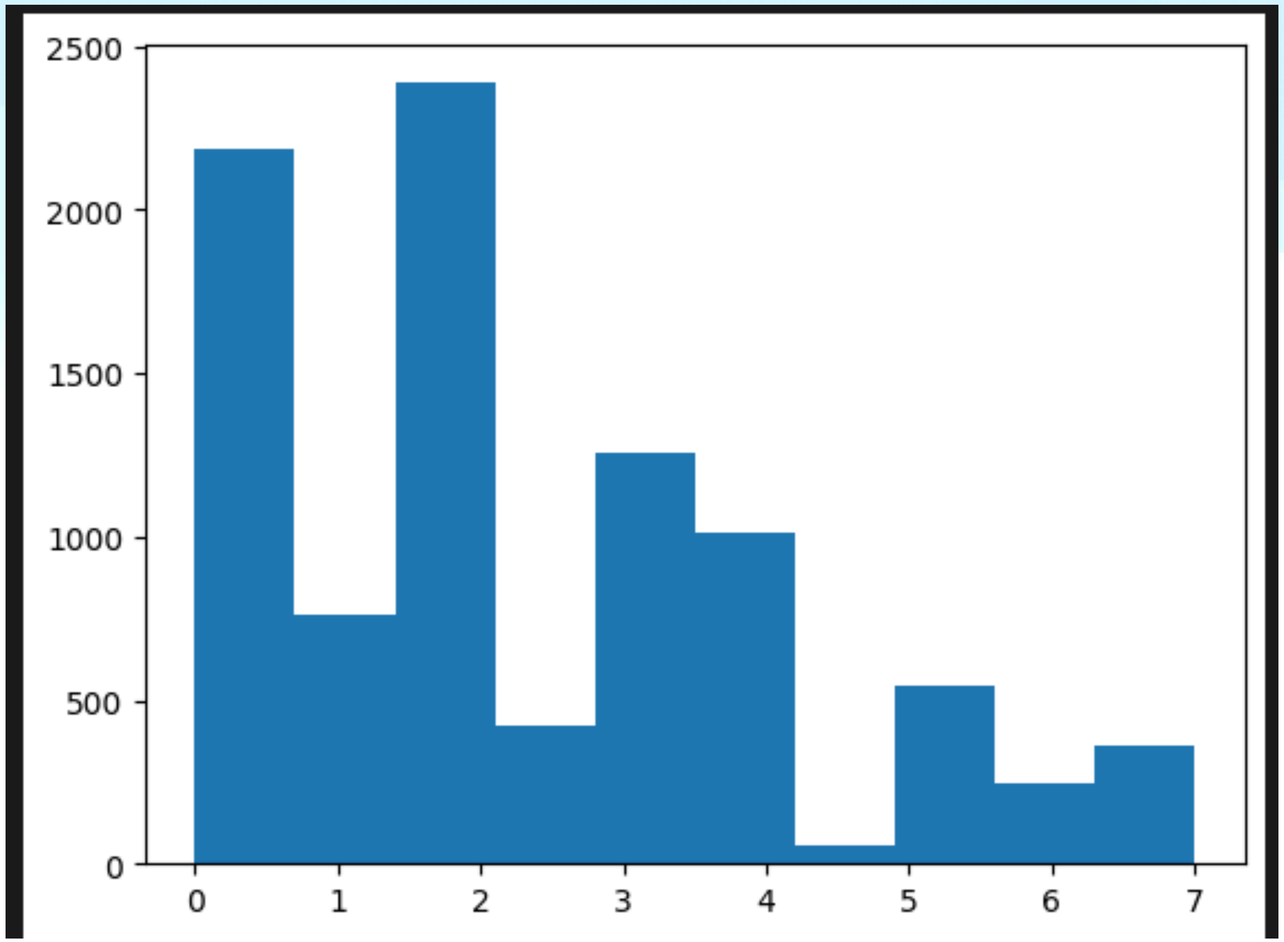
Data Analysis

Most students spend <500min on site and is continuously decreasing



Data Analysis

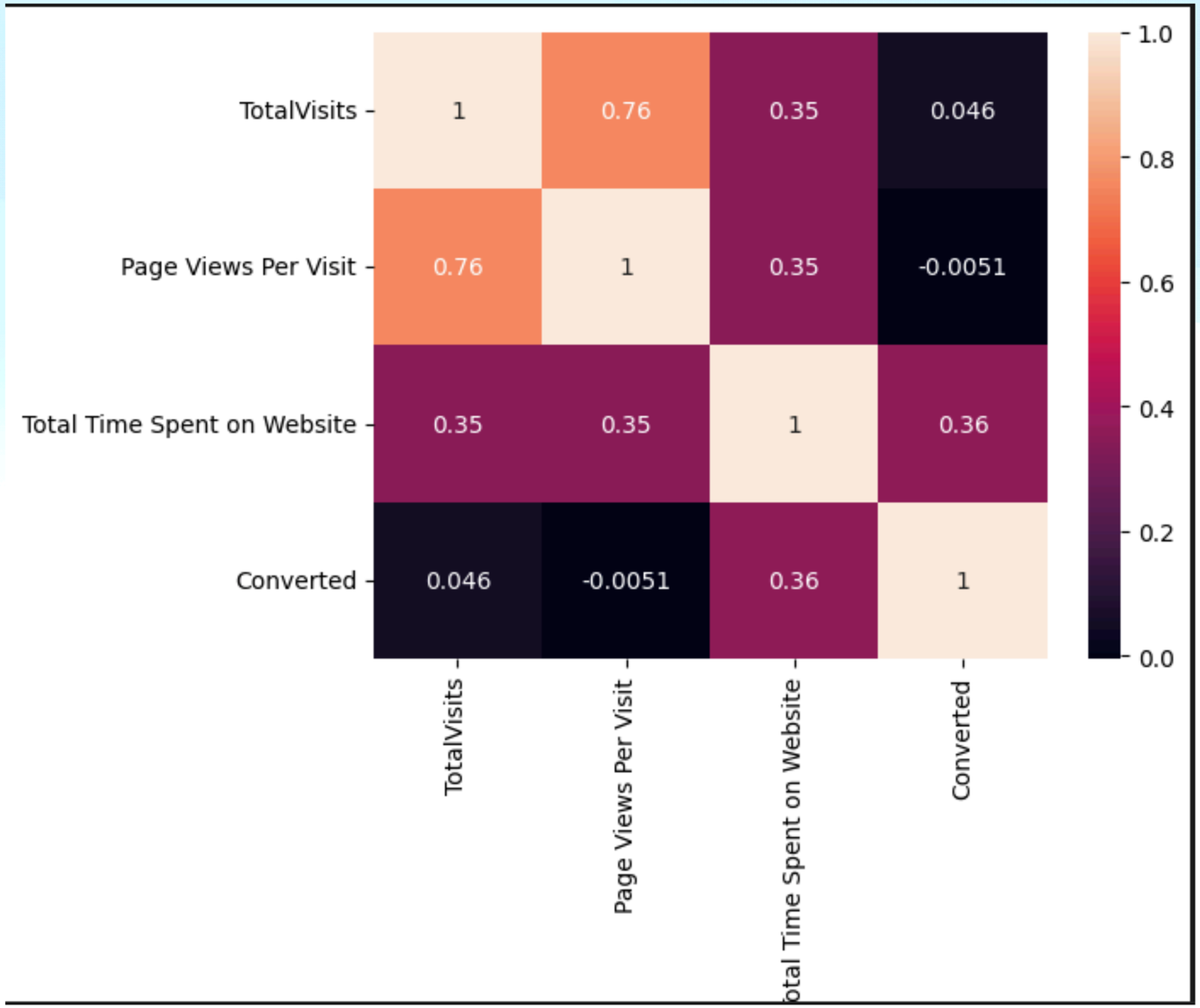
Most students view only less than 2 pages and this variable does not have a linear pattern



Data Analysis

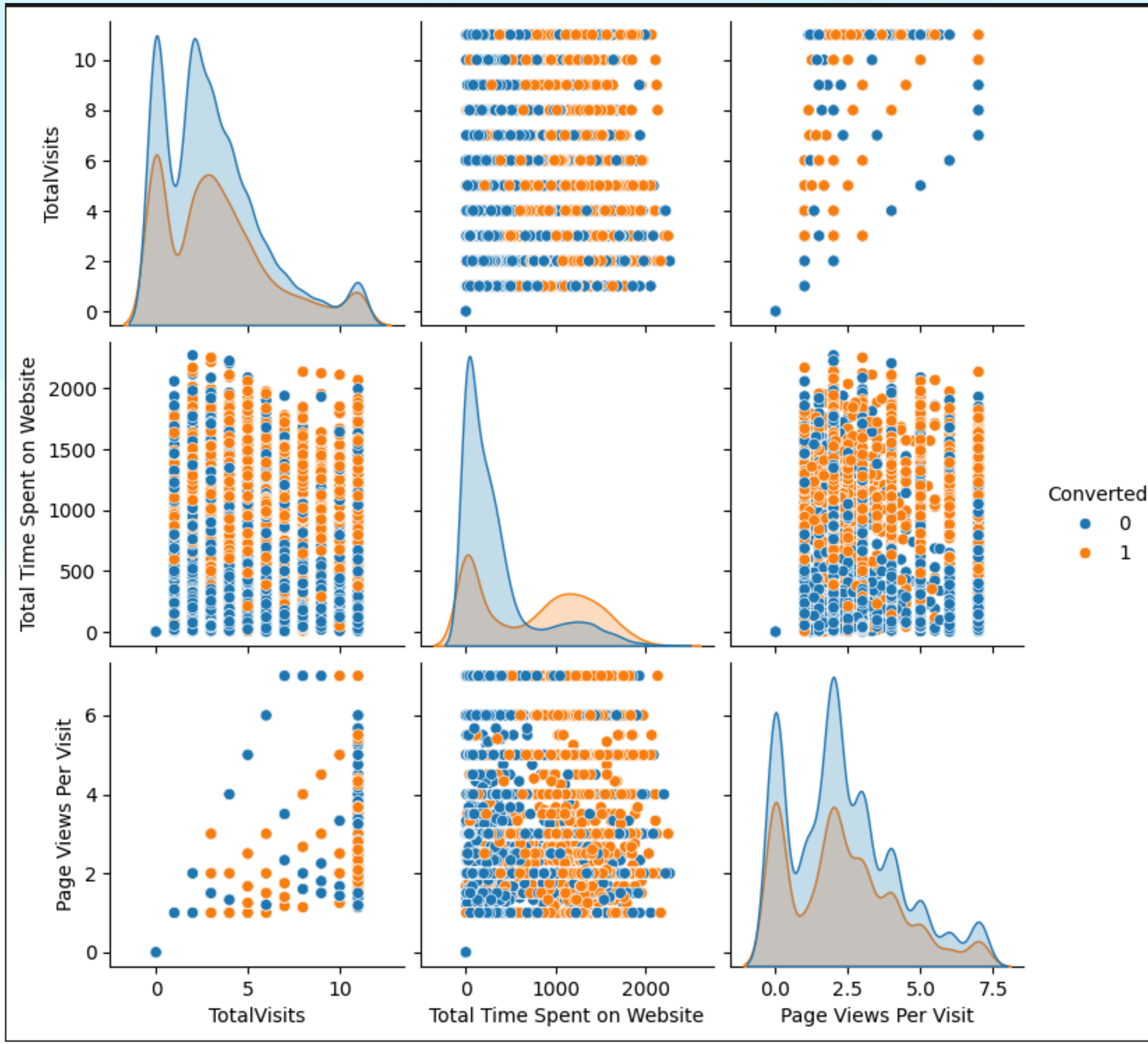
Correlation Matrix:

Time spent on website have a good +ve correlation with convertion. Total visits and Page views per visit have high correlation. We might need to drop them after checking multicollinearity.



Data Analysis

Unable to observe any useful patterns in pair plot of continuous variables.



Modeling

From coef column, we can identify the influence of each feature used in the model

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.2204	0.102	-21.756	0.000	-2.420	-2.020
totalvisits	0.4268	0.061	6.951	0.000	0.306	0.547
total_time_spent_on_website	1.0357	0.048	21.749	0.000	0.942	1.129
page_views_per_visit	-0.4184	0.071	-5.891	0.000	-0.558	-0.279
lead_origin__other	3.0763	0.200	15.395	0.000	2.685	3.468
lead_source__olark_chat	0.9549	0.153	6.245	0.000	0.655	1.255
do_not_email__yes	-1.4629	0.207	-7.063	0.000	-1.869	-1.057
last_activity__olark_chat_conversation	-1.4953	0.186	-8.060	0.000	-1.859	-1.132
last_activity__other	-0.6813	0.137	-4.965	0.000	-0.950	-0.412
last_activity__sms_sent	0.4261	0.160	2.658	0.008	0.112	0.740
what_is_your_current_occupation__other	1.4716	0.312	4.713	0.000	0.860	2.084
what_is_your_current_occupation__unemployed	1.8900	0.131	14.387	0.000	1.633	2.147
what_is_your_current_occupation__working_professional	3.1072	0.254	12.217	0.000	2.609	3.606
tags__other	-1.2839	0.128	-10.057	0.000	-1.534	-1.034
tags__ringing	-4.6593	0.242	-19.245	0.000	-5.134	-4.185
tags__will_revert_after_reading_the_email	2.7827	0.192	14.525	0.000	2.407	3.158
last_notable_activity__other	0.5547	0.173	3.204	0.001	0.215	0.894
last_notable_activity__sms_sent	1.4219	0.173	8.197	0.000	1.082	1.762

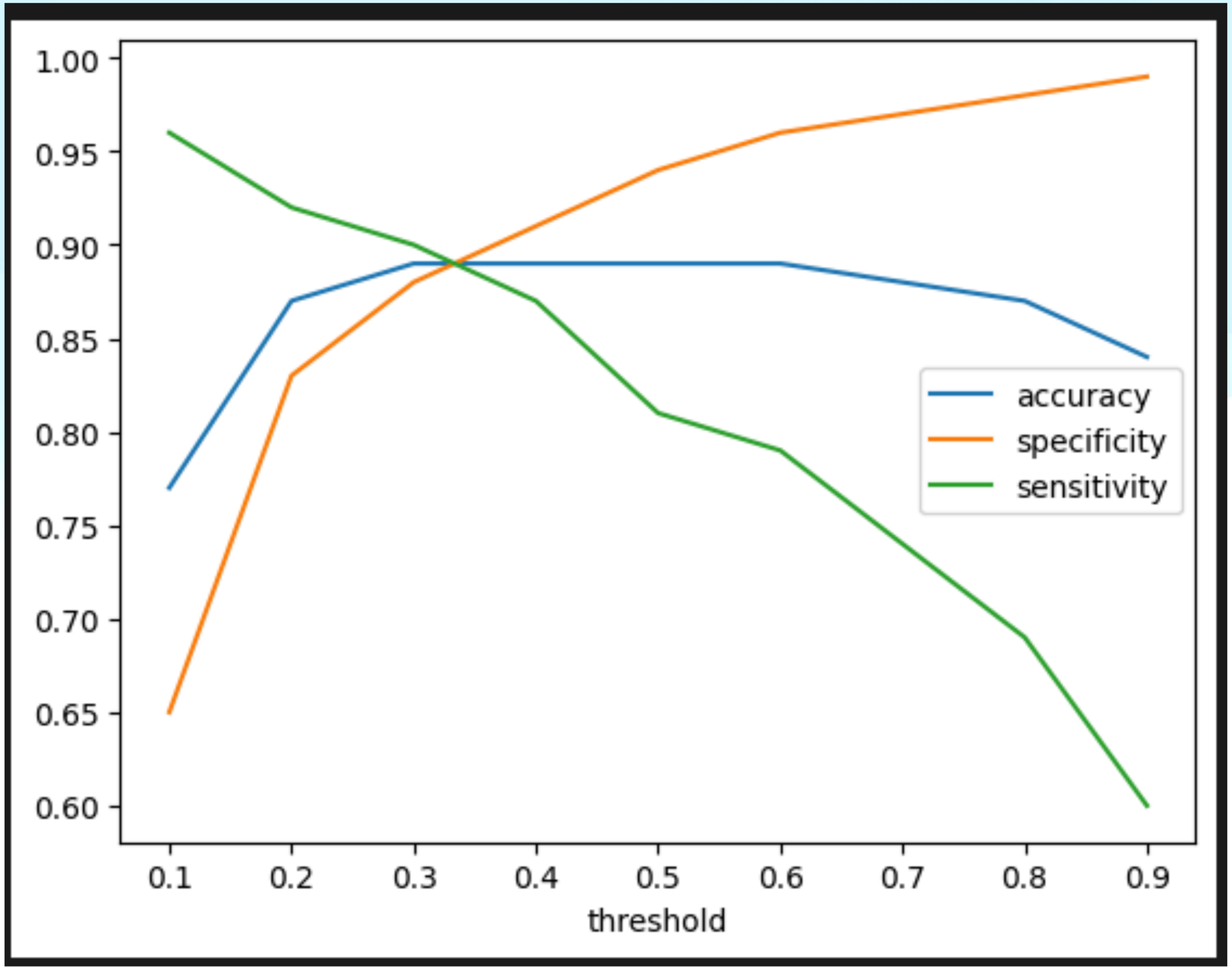
Modeling

We have all VIF values less than 5 indicating no multicollinearity among the variables used in the model.

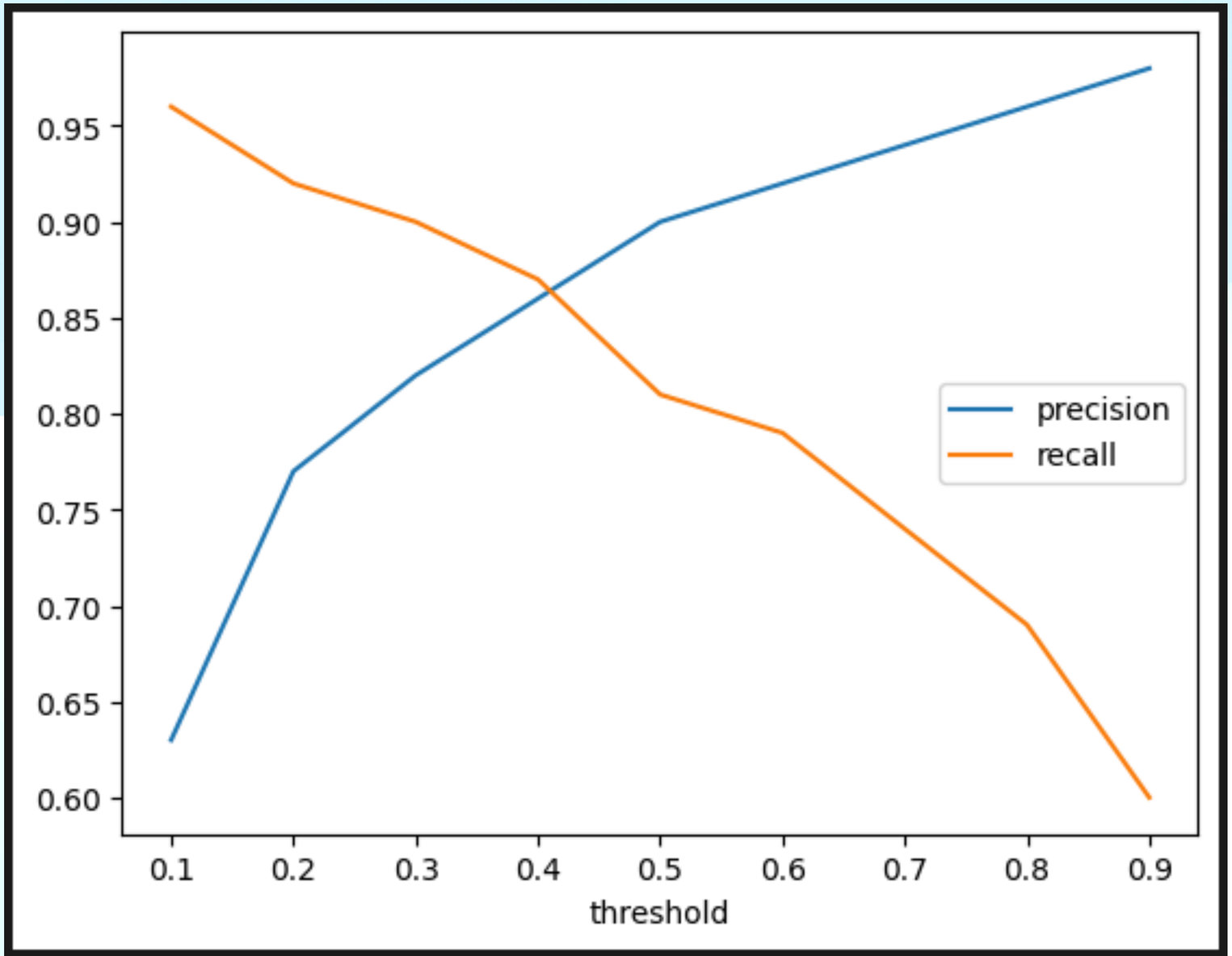
	Features	VIF
0	const	6.32
9	last_activity__sms_sent	3.99
17	last_notable_activity__sms_sent	3.88
11	what_is_your_current_occupation__unemployed	3.62
13	tags__other	3.08
15	tags__will_revert_after_reading_the_email	3.05
3	page_views_per_visit	2.92
1	totalvisits	2.54
12	what_is_your_current_occupation__working_profe...	2.27
14	tags__ringing	2.27
5	lead_source__olark_chat	2.16
8	last_activity__other	1.76
4	lead_origin__other	1.46
7	last_activity__olark_chat_conversation	1.43
2	total_time_spent_on_website	1.38
16	last_notable_activity__other	1.38
10	what_is_your_current_occupation__other	1.34
6	do_not_email__yes	1.14

Modeling

Around 0.33 we get an intersection. We'll be using this point as the threshold.



Precision Recall



Evaluation

> Training model evaluation matrix
> In test set we are able to achieve
    * Accuracy 0.9
    * Specificity 0.9
    * Sensitivity 0.9

	threshold	accuracy	sensitivity	specificity	precision	recall
1	0.1	0.77	0.96	0.65	0.63	0.96
2	0.2	0.87	0.92	0.83	0.77	0.92
3	0.3	0.89	0.90	0.88	0.82	0.90
4	0.4	0.89	0.87	0.91	0.86	0.87
5	0.5	0.89	0.81	0.94	0.90	0.81
6	0.6	0.89	0.79	0.96	0.92	0.79
7	0.7	0.88	0.74	0.97	0.94	0.74
8	0.8	0.87	0.69	0.98	0.96	0.69
9	0.9	0.84	0.60	0.99	0.98	0.60

Conclusion

> We created a logistic regression model with 17 features
> We where able to get a specificity of 0.9 and sensitivity of 0.9 in the test set created.
> We can change the threshold value to get more sensitivity of specificity according to the requirement.