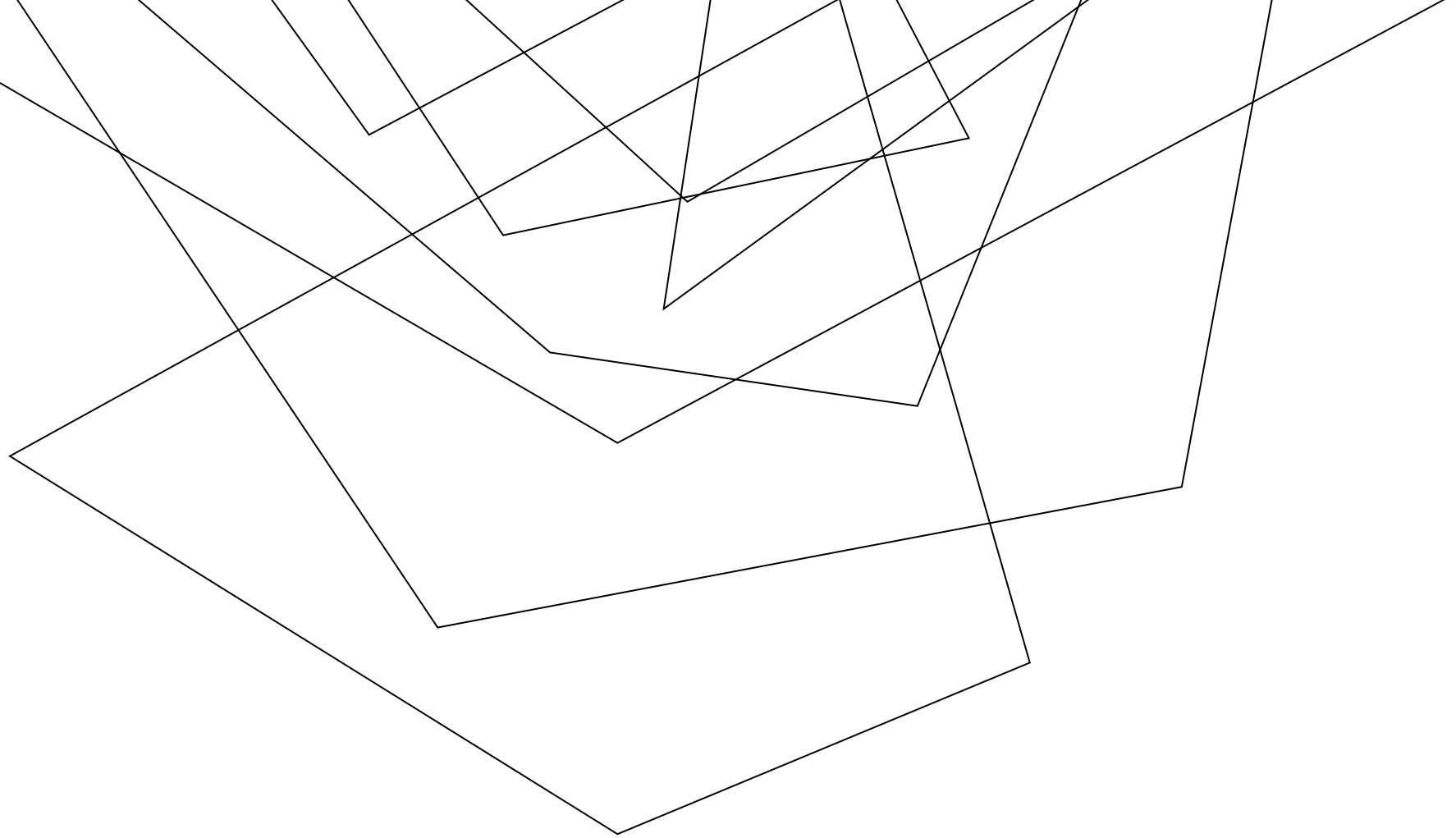


LEADS SCORING CASE STUDY

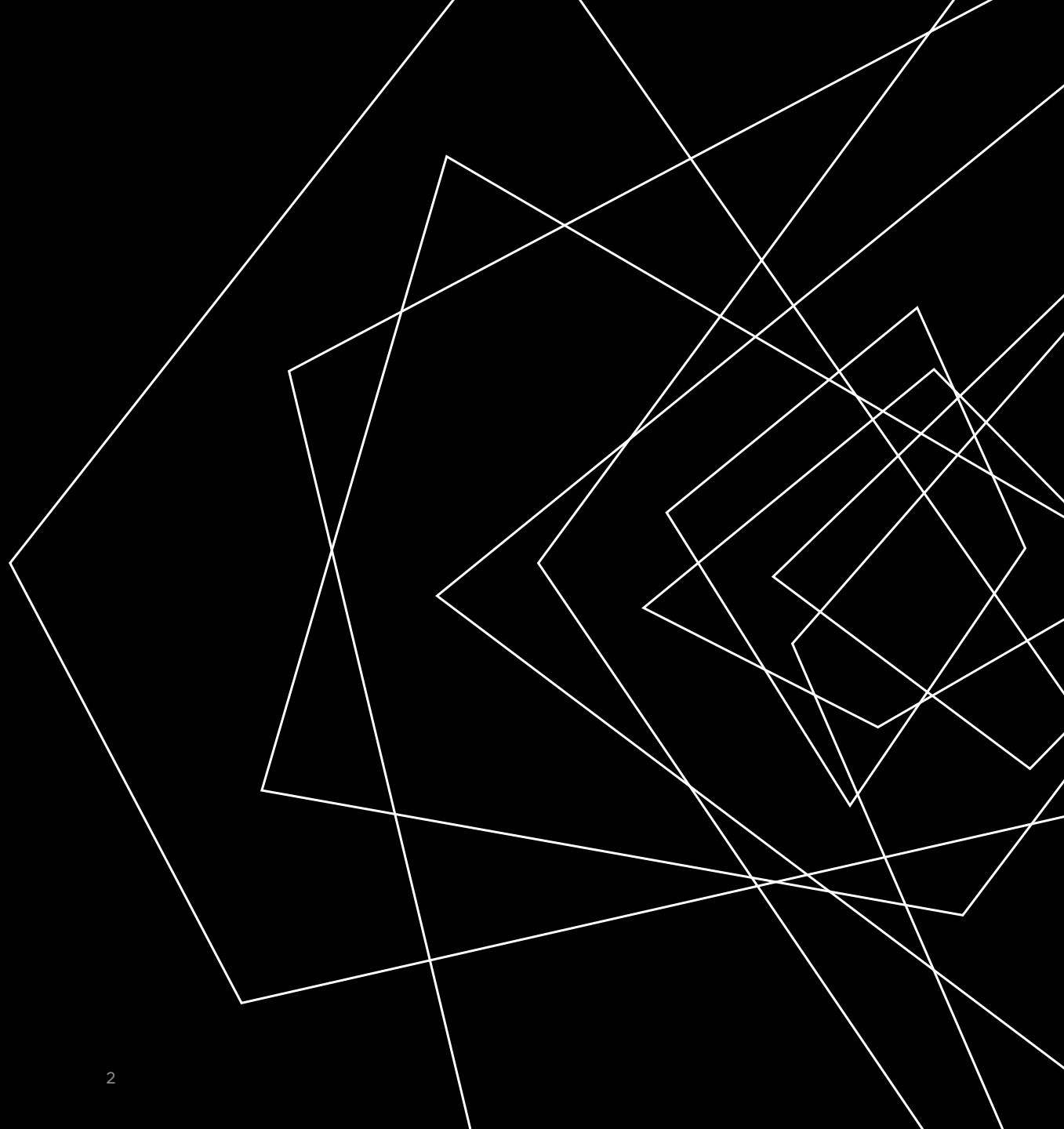
Rahul Gupta , Rahul Davaskar, Rinky Juneja

EPGDS DS May ‘23 C56

BUSINESS OBJECTIVES

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.

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 requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a 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%.



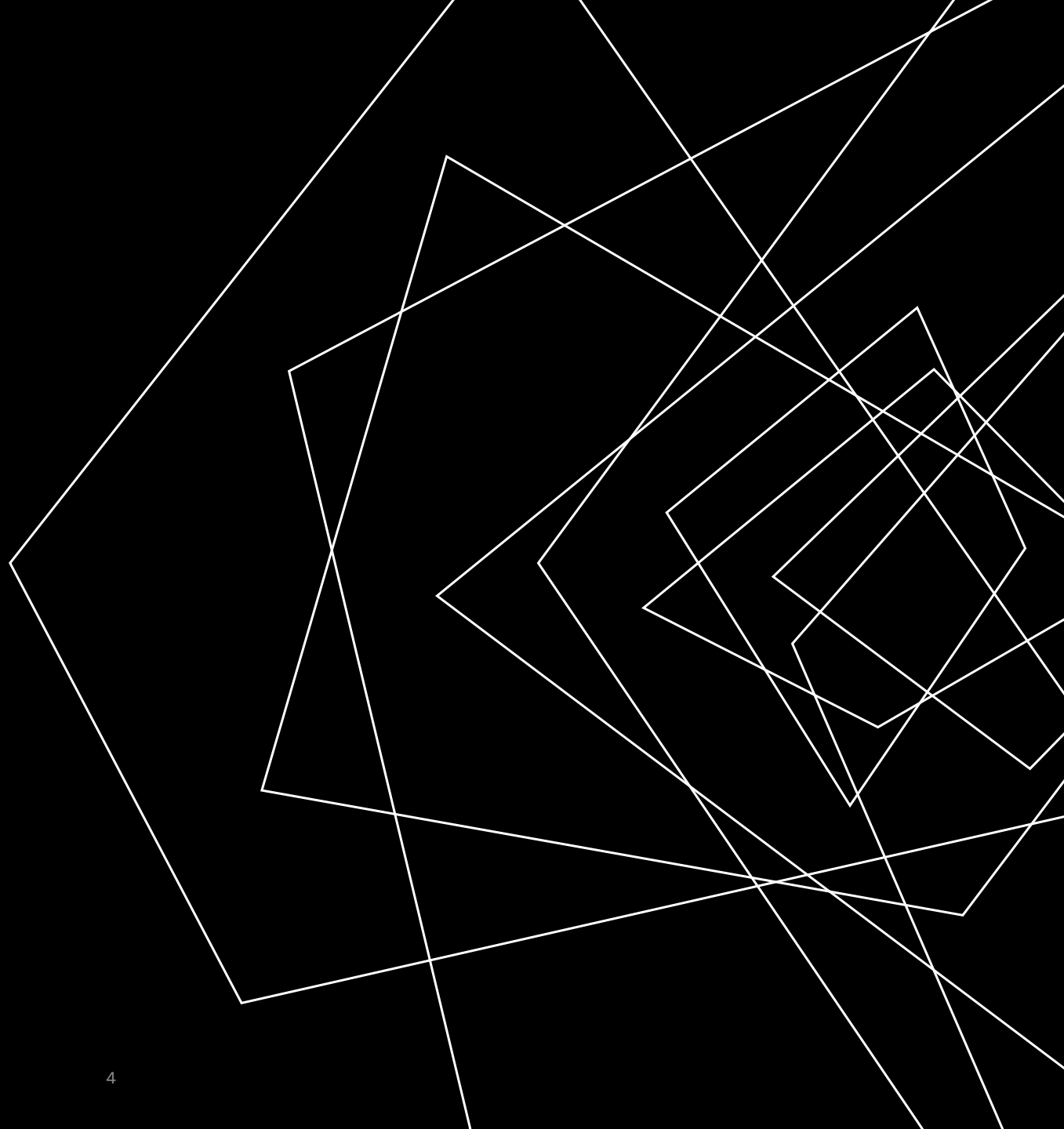
BUSINESS UNDERSTANDING

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. 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%.

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.

ANALYSIS APPROACH

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 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?).

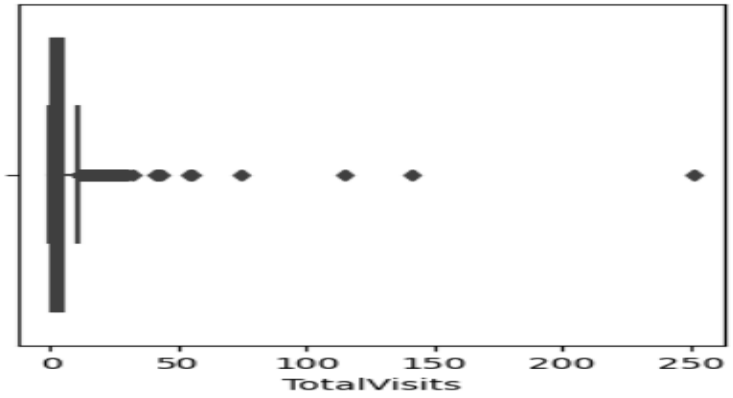


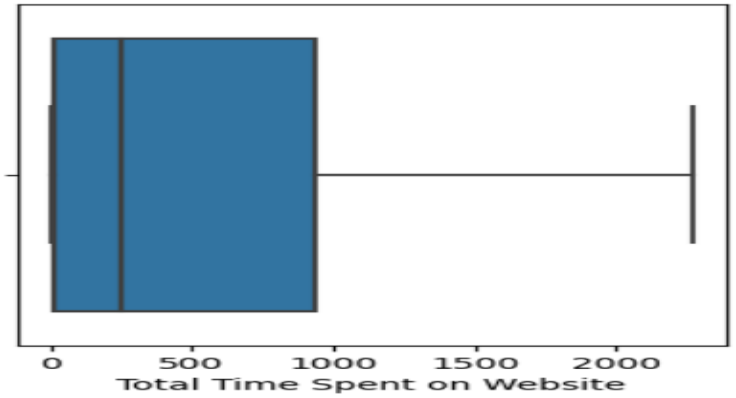
ANALYSIS OF APPLICATION DATA

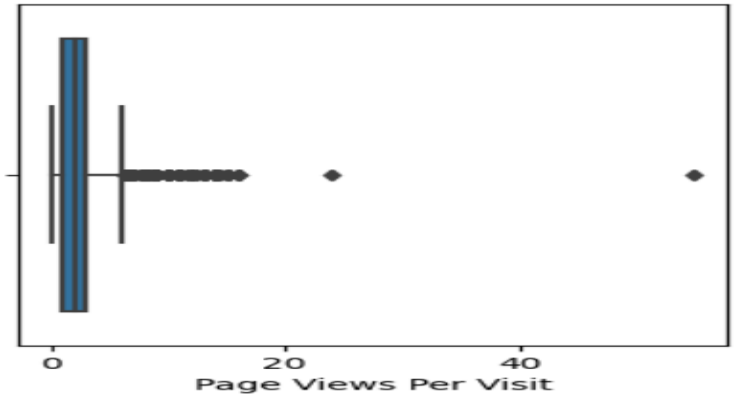
- All the columns in both the data set where null values are more than 35% Tags, Lead Quality, Asymetrique Index are dropped.
- Irrelevant columns like Prospect ID, etc are dropped.
- Lead Number is set as index.
- Select is replaced with Nan and Columns with 40% Nan are dropped.
- Numerical columns with low NA counts are filled with mean value.
- Categorical columns with low NA counts are filled with mode value.
- Skewed columns like Do Not Call, Country, Search, Magazine, Newspaper Article, X Education, etc are dropped.
- Outliers in Total Visits are handled using IQR method.
- Less occurring categorical values are replaced with Others.
- Null Count is checked in entire data set.

UNIVARIATE ANALYSIS (CONTINUOUS)

- TotalVisits and Page Views Per Visit columns showing outliers.
- Rows are dropped using IQR

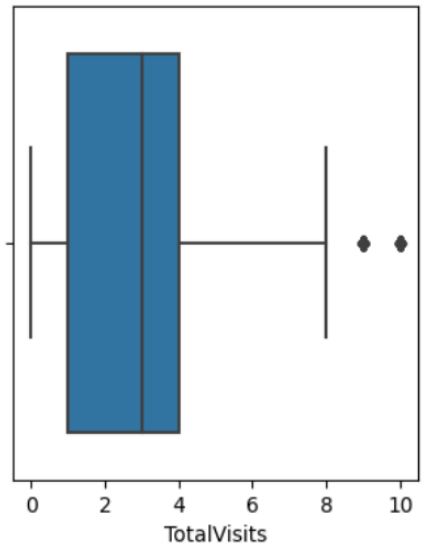


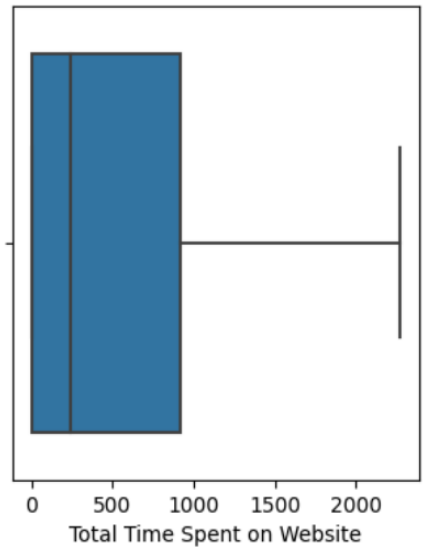


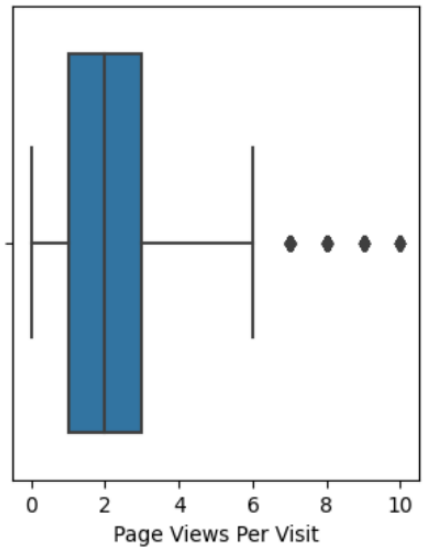


Outliers Treatment

In [52]:
```
def outlier_treatment(data,col):
    Q1 = data[col].quantile(0.25)
    Q3 = data[col].quantile(0.75)
    IQR = Q3 - Q1
    print(Q1,Q3,IQR)
    upper_limit = Q3 + (1.5)*IQR
    lower_limit = Q1 - (1.5)*IQR
    return data[(data[col]<upper_limit)&(data[col]>lower_limit)]
```

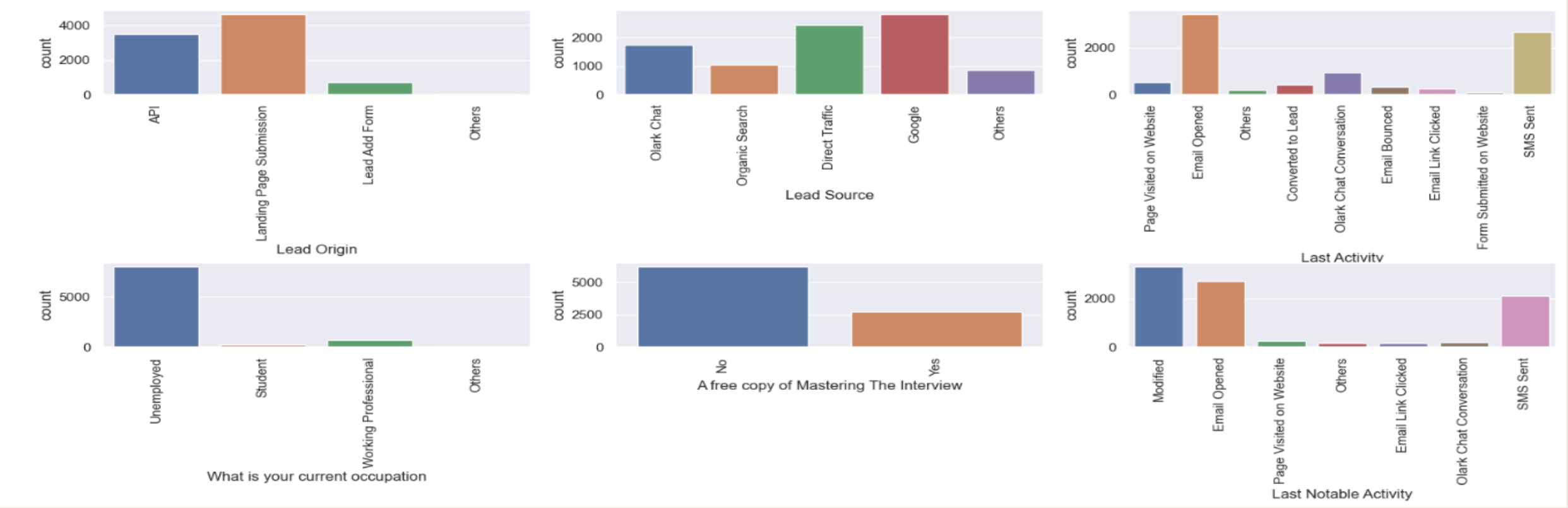






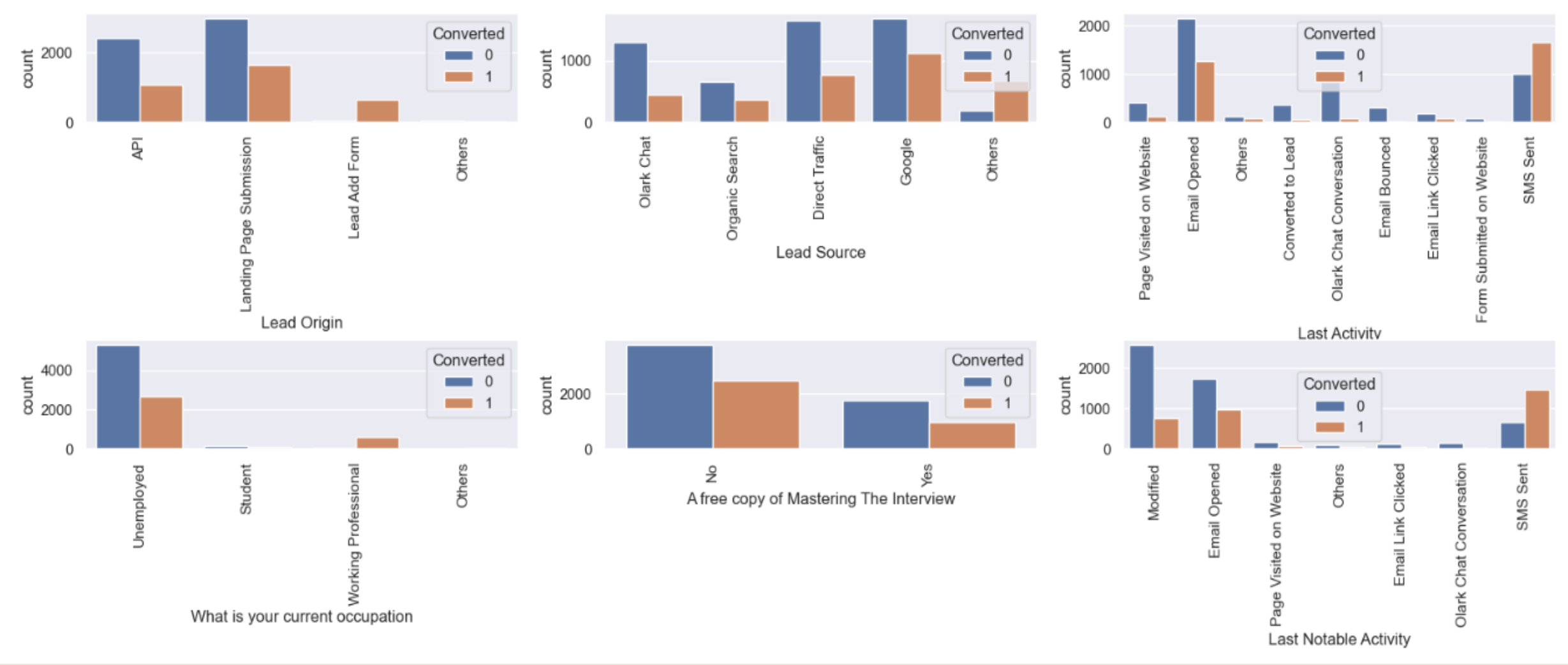
UNIVARIATE ANALYSIS (CATEGORICAL)

- Skewed Categorical Values are dropped.
- Count plot for rest of the columns are plotted.
- Mostly Unemployed People lands on the website
- Google has highest Lead Source
- Most Leads have arrived from Landing Page Submission



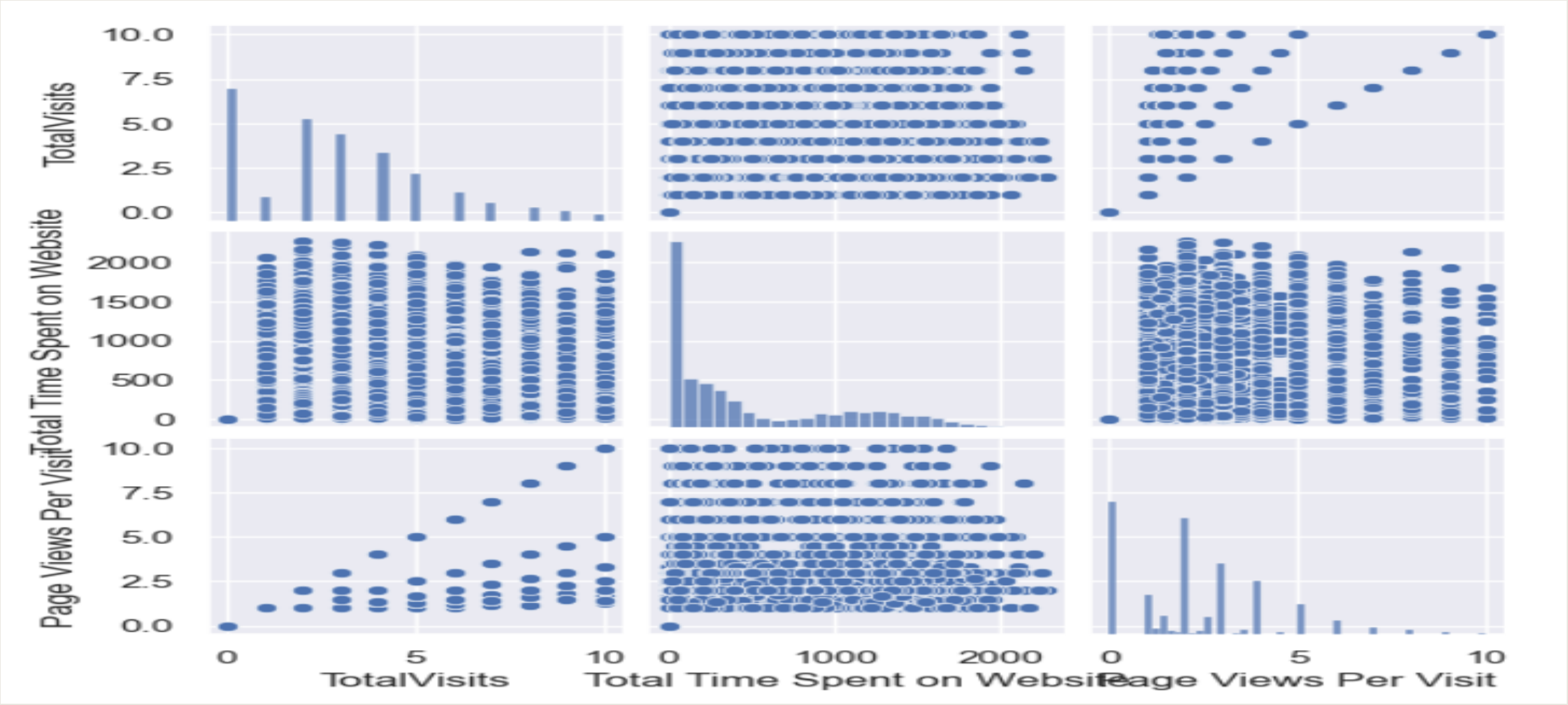
BIVARIATE ANALYSIS (CATEGORICAL)

- Most converted are from Lead Origin Add Form
- Most converted are Working Professional



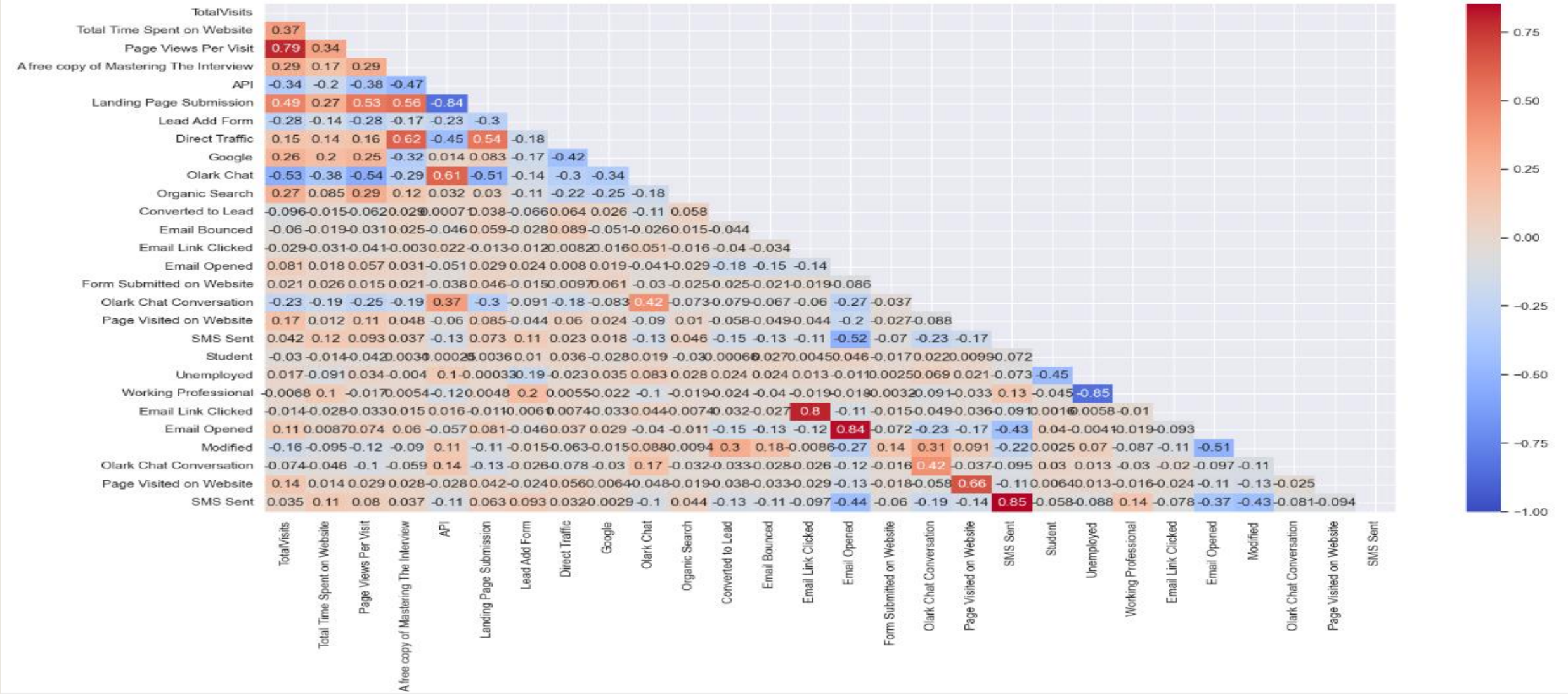
BIVARIATE ANALYSIS (CONTINUOUS)

- TotalVisits and Page Views Per Visit are linearly related



MULTIVARIATE ANALYSIS

- Post Dummies creation most of the columns are correlated.



FINAL MODEL

- Final Model post deleting feature with higher VIP and P-Values.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6220
Model:	GLM	**Df Residuals:**	6211
Model Family:	Binomial	**Df Model:**	8
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2910.4
Date:	Mon, 27 Nov 2023	**Deviance:**	5820.9
Time:	20:57:46	**Pearson chi2:**	8.49e+03
No. Iterations:	6	**Pseudo R-squ. (CS):**	0.3267
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.1103	0.101	-20.879	0.000	-2.308	-1.912
TotalVisits	0.8018	0.219	3.667	0.000	0.373	1.230
Total Time Spent on Website	4.4755	0.158	28.389	0.000	4.166	4.784
Page Views Per Visit	-0.5940	0.286	-2.080	0.037	-1.154	-0.034
Lead Add Form	4.0385	0.195	20.717	0.000	3.656	4.421
Olark Chat	0.8414	0.118	7.107	0.000	0.609	1.073
Converted to Lead	-1.4320	0.213	-6.723	0.000	-1.849	-1.014
Email Bounced	-2.3764	0.308	-7.708	0.000	-2.981	-1.772
Working Professional	3.0413	0.194	15.636	0.000	2.660	3.422

ROC CURVE

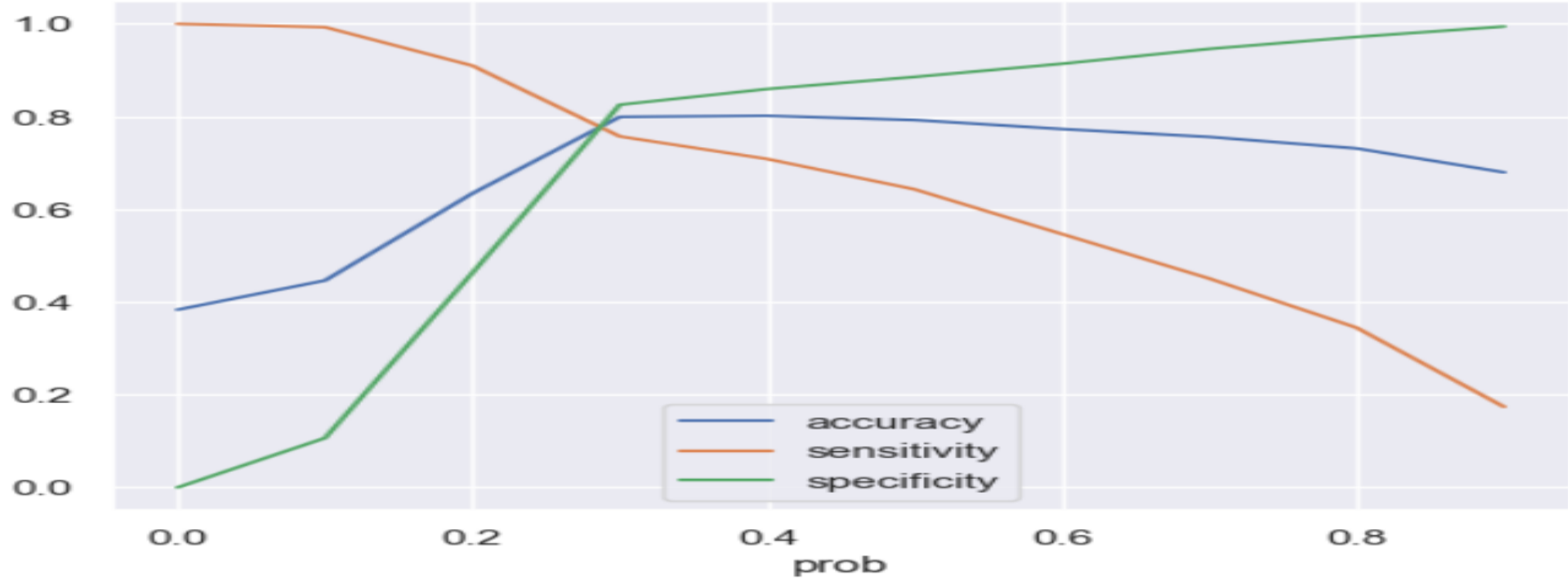

Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.84)

INITIAL OPTIMAL POINT

- Initial Cutoff is at 0.29

```
# Accuracy Sensitivity and Specificity for all probabilities plot
plt.figure(figsize=(20,15))
cutoff_df.plot.line(x='prob', y=['accuracy','sensitivity','specificity'])
plt.show()
```

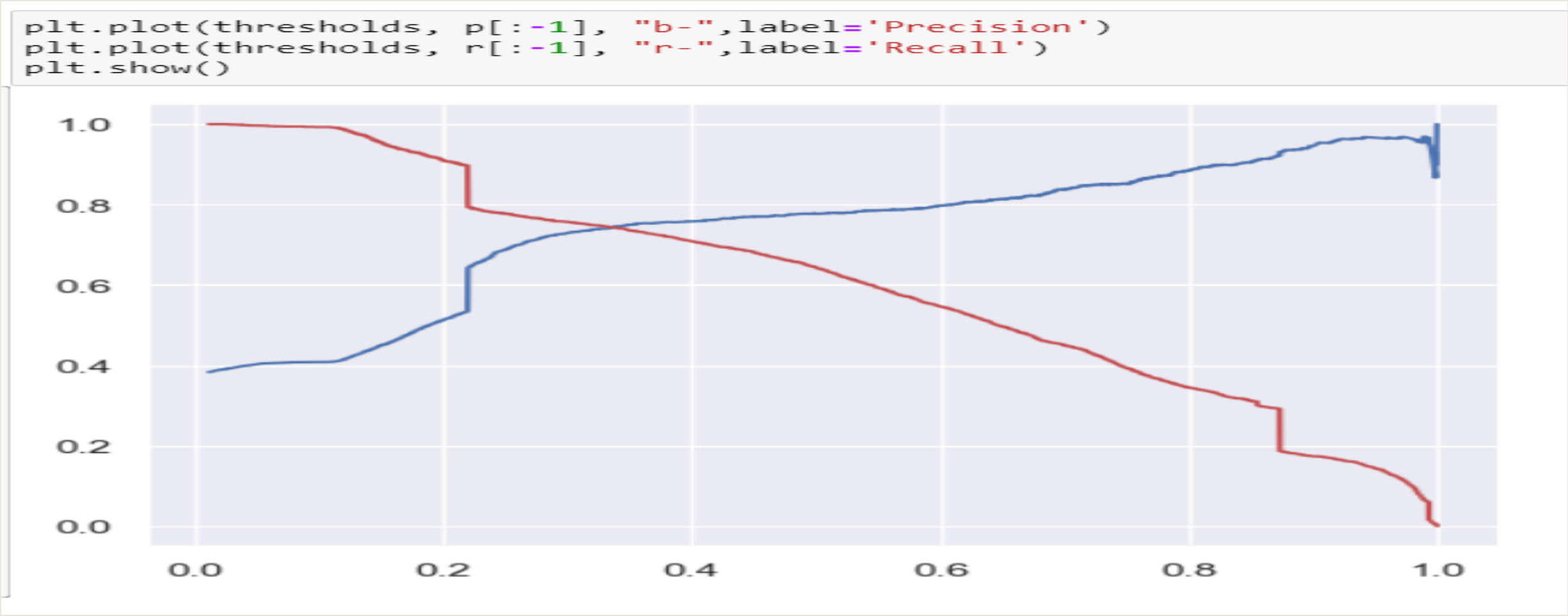
```
<Figure size 2000x1500 with 0 Axes>
```



FINAL OPTIMAL POINT

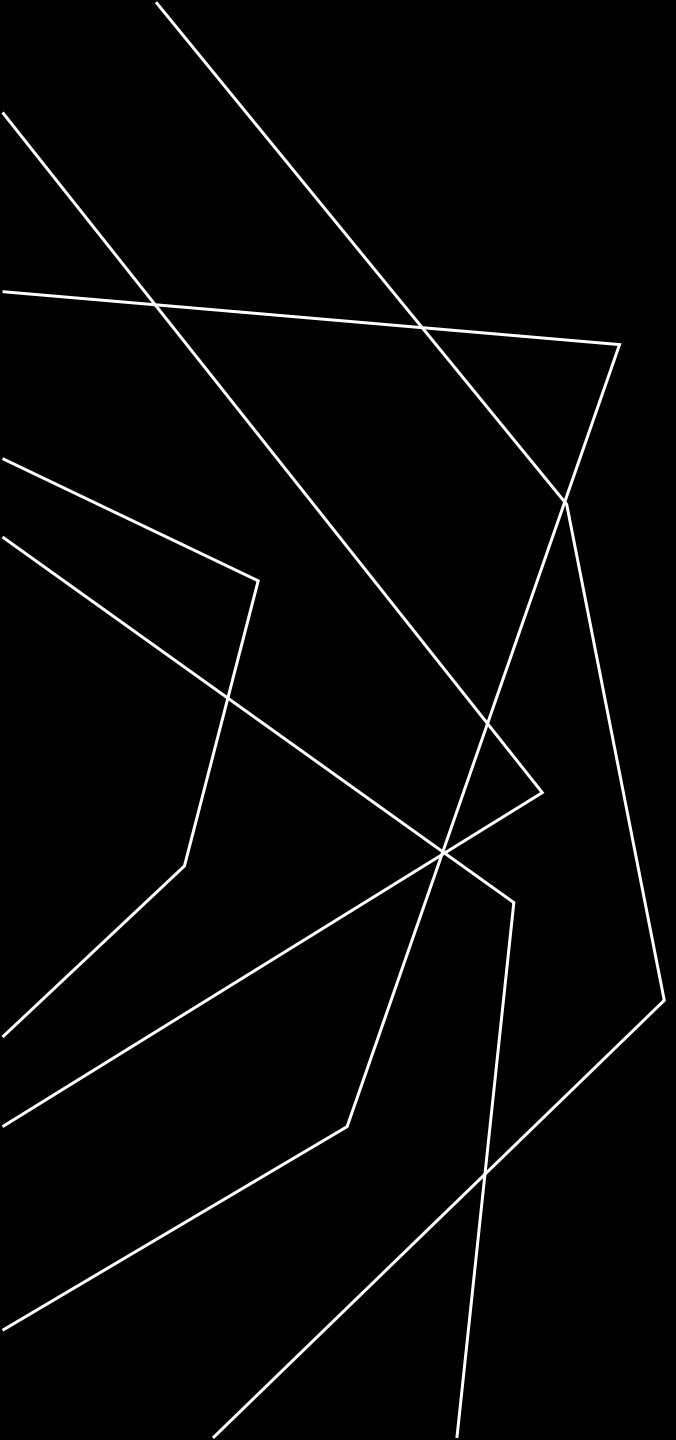
- Final Cutoff is at 0.31

```
plt.plot(thresholds, p[:-1], "b-",label='Precision')
plt.plot(thresholds, r[:-1], "r-",label='Recall')
plt.show()
```



SUMMARY

- Students are visits more but their conversation rate is low.
- Working Professional visits less compared to Students but more likely to take courses.
- Lead source are higher in numbers from Google and Olark Client but people coming via Olark Client converts.
- Mostly people from India visits this website.
- Mostly unemployed person create traffic.
- Optimal cut off point is 0.31
- Converted lead to be avoided by sales team.
- Referred people have high probability to convert.
- Ignore people with bounced email
- Lead Add form are good candidate.



THANK YOU