Lead Scoring: Case study analysis

- A Anurag
- Vijaya R

Problem Statement and Approach

- 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.
- 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%.
- 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 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%.

- The given dataset is: *'leads.csv'.* After loading the required libraries we load the dataset in our ipynb file for performing actions on it.
- Then we clean the dataset(Like treating the missing values, deleting the columns which will not be part of our analysis, conversion of the columns to right format, taking care of negative value, checking for outliers etc.)
- Univariate, Bivariate and Multivariate analysis on some of the columns will done to trends or insights from them.
- Then after that we start building model and make predictions and evaluate using metrics like accuracy etc.

Summary of Dataset

Dataset: Leads data

- Initially we had 37 columns, but it was reduced to 13 columns which will be used for analysis.
- Null values percentage greater than 40% values columns were removed.
- We did following things on the dataset
  - Removed unwanted columns for the analysis
  - Treated the null values wherever required.
  - Conversion of the values of some columns to Others and Correcting some spelling mistakes
- With all these changes this dataset now is ready for the analysis.

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
15	What matters most to you in choosing a course	6531 non-null	object
16	Search	9240 non-null	object
17	Magazine	9240 non-null	object
18	Newspaper Article	9240 non-null	object
19	X Education Forums	9240 non-null	object
20	Newspaper	9240 non-null	object
21	Digital Advertisement	9240 non-null	object
22	Through Recommendations	9240 non-null	object
23	Receive More Updates About Our Courses	9240 non-null	object
24	Tags	5887 non-null	object
25	Lead Quality	4473 non-null	object
26	Update me on Supply Chain Content	9240 non-null	object
27	Get updates on DM Content	9240 non-null	object
28	Lead Profile	6531 non-null	object
29	City	7820 non-null	object
30	Asymmetrique Activity Index	5022 non-null	object
31	Asymmetrique Profile Index	5022 non-null	object
32	Asymmetrique Activity Score	5022 non-null	float64
33	Asymmetrique Profile Score	5022 non-null	float64
34	I agree to pay the amount through cheque	9240 non-null	object
35	A free copy of Mastering The Interview	9240 non-null	object
36	Last Notable Activity	9240 non-null	object

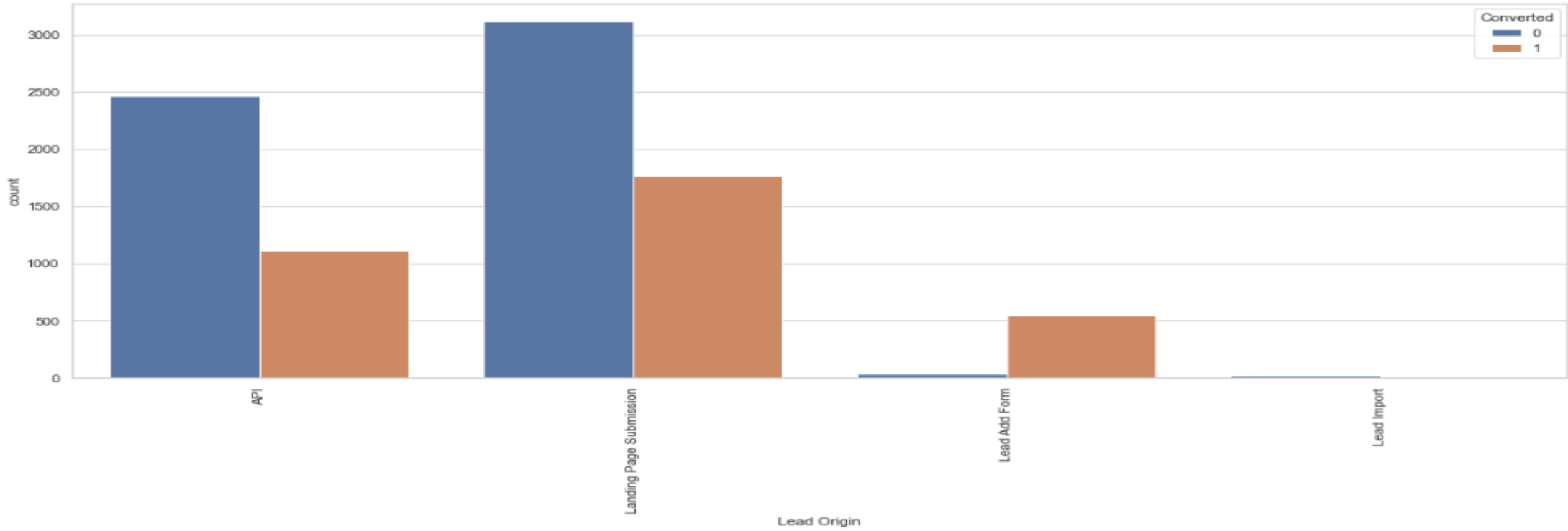
dtypes: float64(4), int64(3), object(30)

0	Lead Origin	9074 non-null	object
1	Lead Source	9074 non-null	object
2	Do Not Email	9074 non-null	object
3	Converted	9074 non-null	int64
4	TotalVisits	9074 non-null	float64
5	Total Time Spent on Website	9074 non-null	int64
6	Page Views Per Visit	9074 non-null	float64
7	Last Activity	9074 non-null	object
8	Specialization	9074 non-null	object
9	What is your current occupation	9074 non-null	object
10	Tags	9074 non-null	object
11	Lead Quality	9074 non-null	object
12	Last Notable Activity	9074 non-null	object

dtypes: float64(2), int64(2), object(9)

Univariate and Bivariate Analysis

- Column Lead Origin



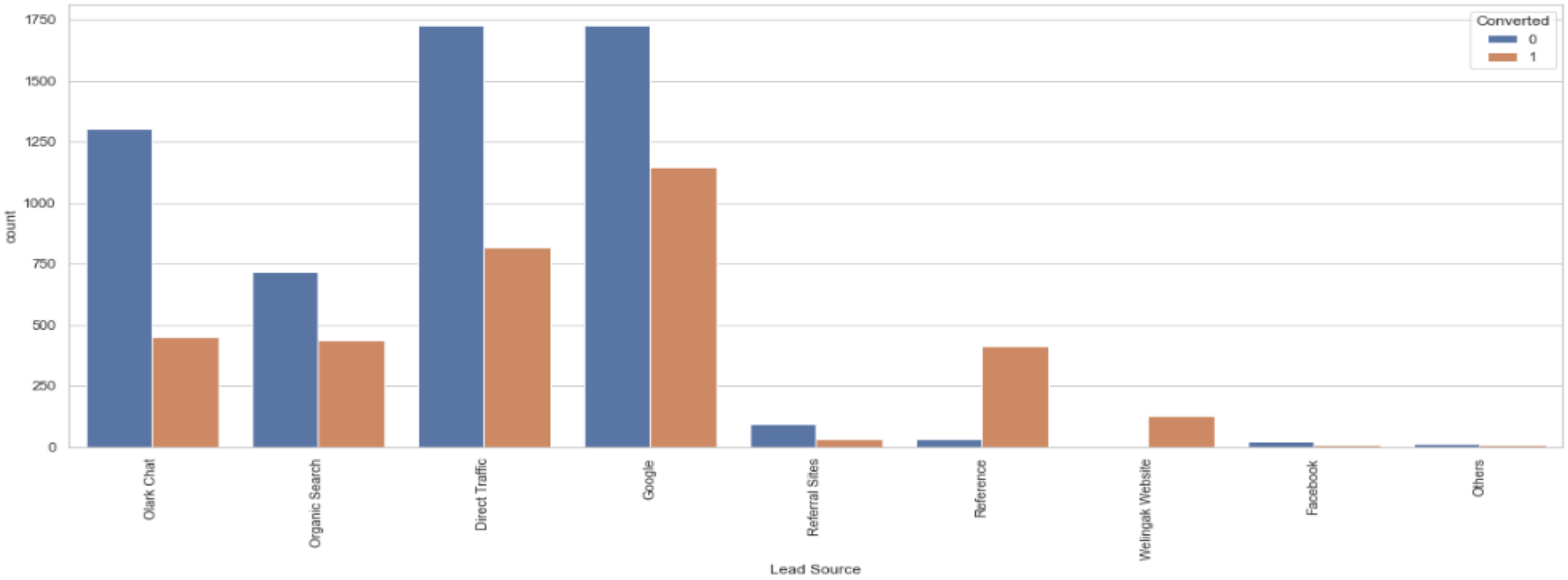
Upon inspection

- API and Landing Page Submission generate most of the leads but conversion rate for them are close to 40 to 50 percent.
- Lead Add Form generate less number of leads as compared to API and Landing Page Submission but the conversion rate is good.
- Lead Import does not so signicant

So we should try to increase the conversion rate for API and Landing Page Submission and increase leads using Lead Add Form.

Contd.

- Column Lead Source

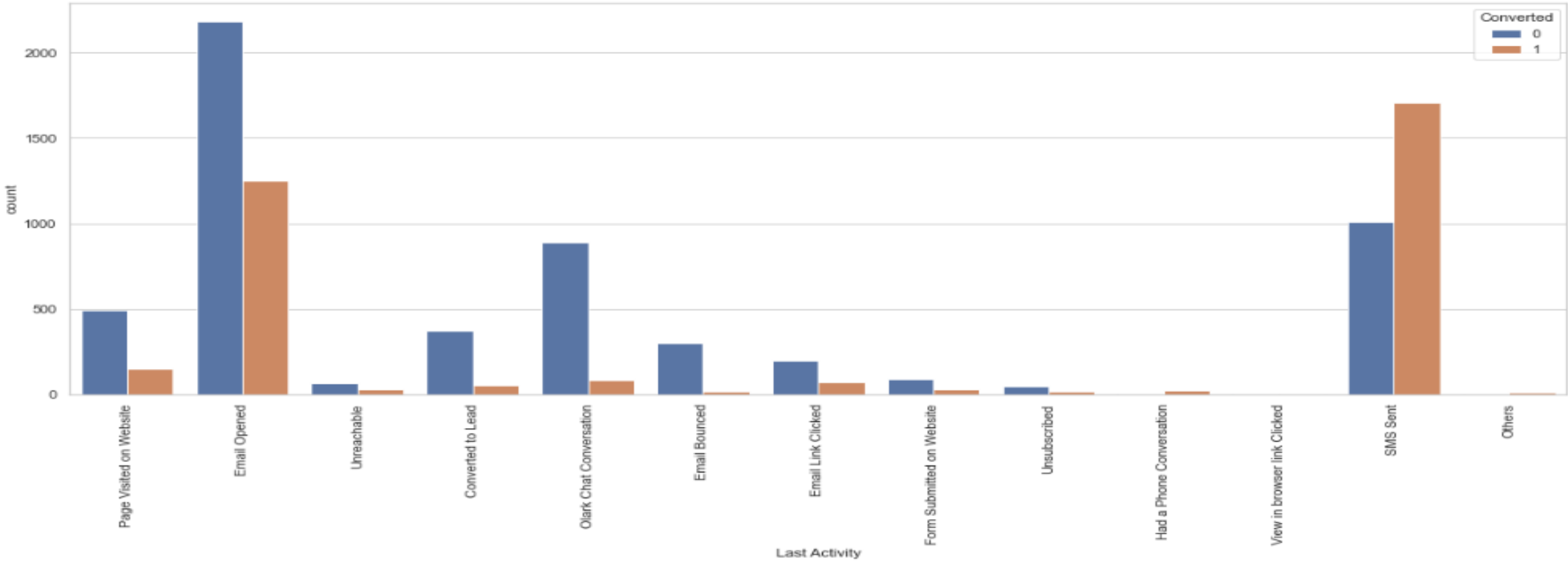


Upon Inspection

- There is a spelling error for google, has it should be Google. So we need to convert these to right form.
- Lead generation after the facebook source are very neglible so lets see the piechart of it and put all of them into others category.
- Direct Traffic and Google generate most number leads but their conversion rate is low.
- Reference and Welingak Website generate less number fo leads compared to Direct trafiic and Google but their conversion rate is high.

Contd.

- Column Last Activity

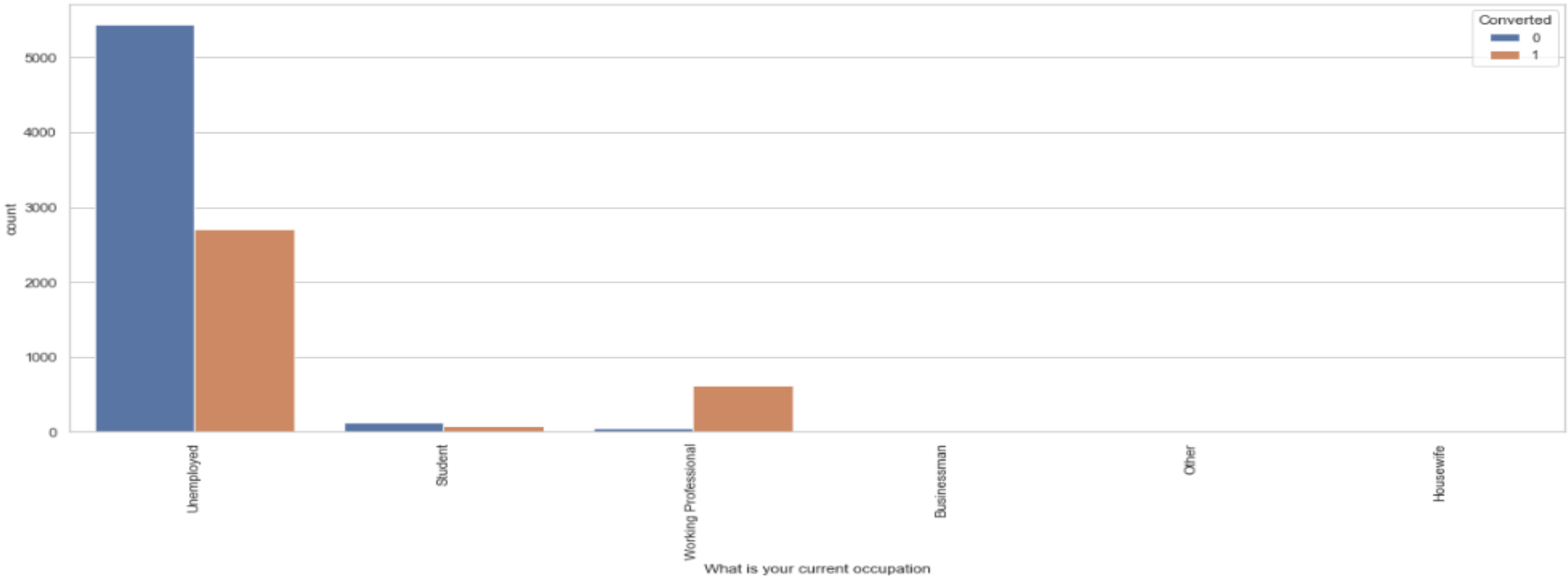


Upon Inspection

- More number of leads are generated from Email opened and SMS Sent. For Email Opened conversion rate is less but for SMS Sent conversion rate is high.
- And categpories after SMS Sent are having negligible values so lets put them in Others categories.

Contd.

- Column What is your current occupation

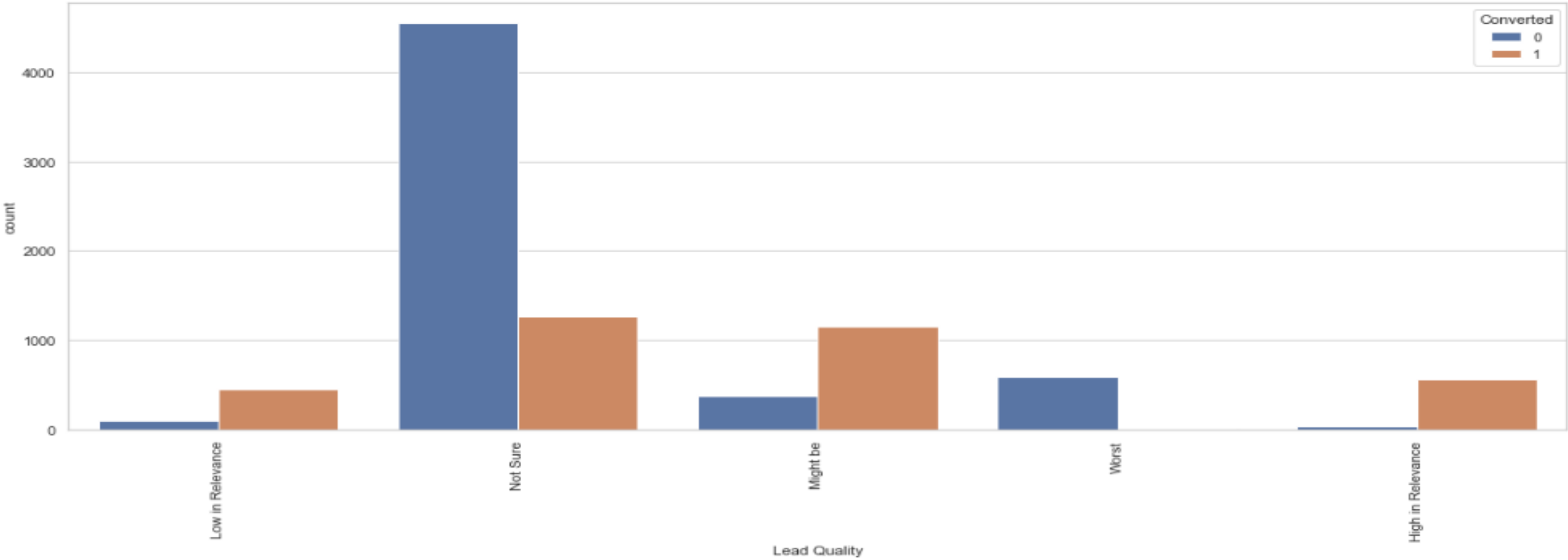


Upon Inspecting

- More number of leads are generated by unemployed profile but their conversion rate is low, but Working Professional has very high conversion rate.

Contd.

- Column Lead Quality

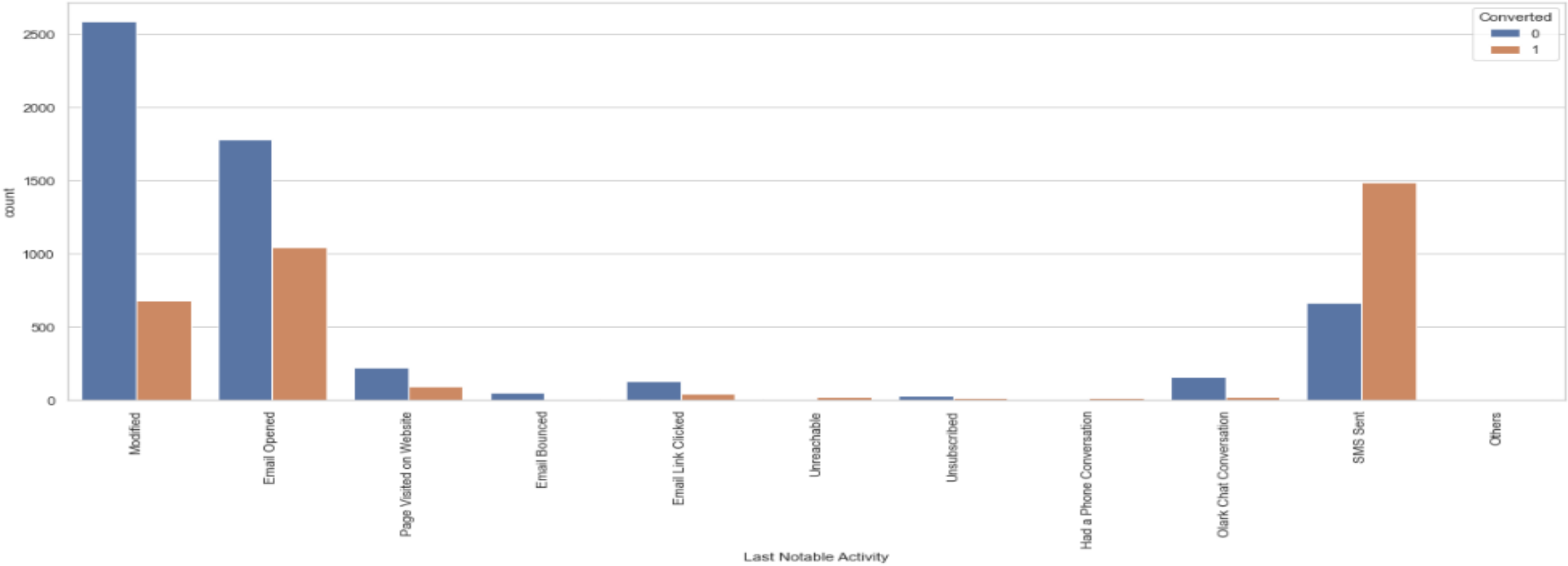


Upon Inspecting

- Not Sure has more number of leads generated but their conversion rate is very low.
- Might be even though it has less number of leads generated as compared with Not Sure but it has high conversion rate.
- And Worst has very low conversion rate close to negligible.

Contd.

- Column Last Notable Activity

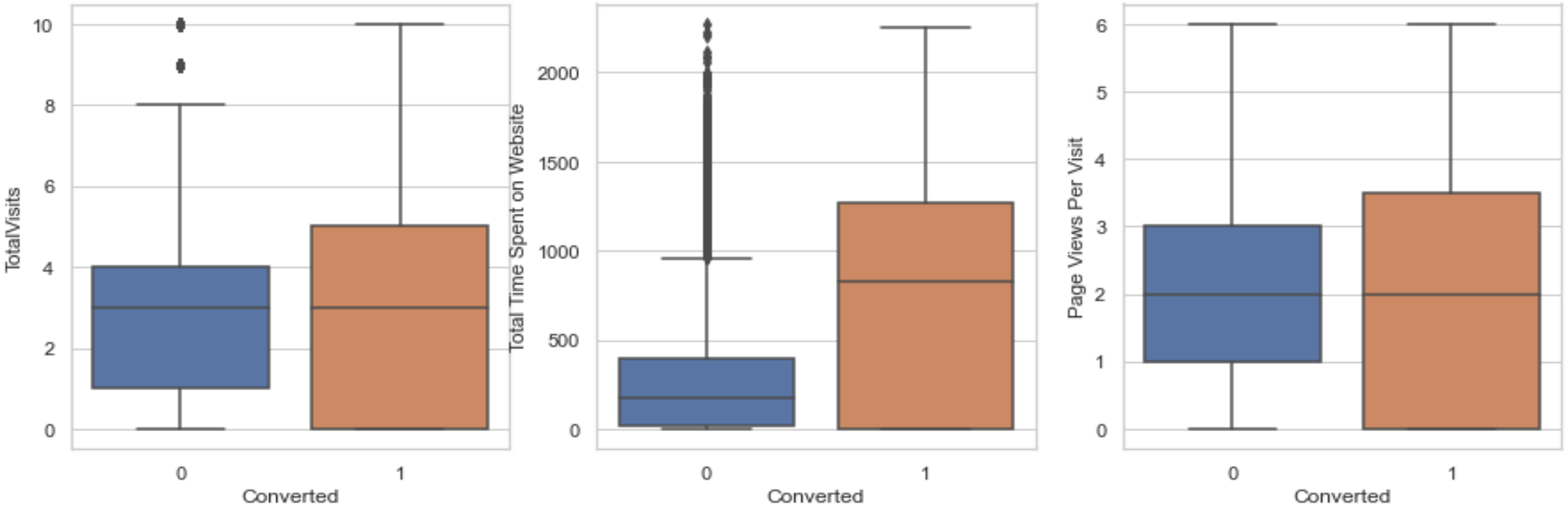


Upon Inspecting

- Modified has high number of leads generated but conversion rate is very low.
- SMS Sent has very high conversion rate.
- After SMS Sent we can combine all others to Others category

Contd.

- Numerical columns after outlier treatment (capping) against Converted column



Our observations:

- TotalVisits has same median value for both converted and not converted, so not much conclusion can be drawn from this.
- In the Total Time Spent on Website are more likely to get converted. As peopple who are interested will spend more time on the website. This can also be general knowledge.
- Page View Per Visit also as same median so this also not much conclusion can be drawn.

Model Building

- First the creation of the dummy variables was done, then we used standardscaler to scale the numerical variables. Then we split the data into train and test.
- Then used RFE to build the model with 15 features in it.
- After building we removed two more features one by one based on the p-value and VIF values.
- After building the metrics are:
  - Accuracy = 92%
  - Sensitivity = 85%
  - Specificity = 96%

```
Confusion Matrix:
[[3756  149]
 [ 363 2083]]
```

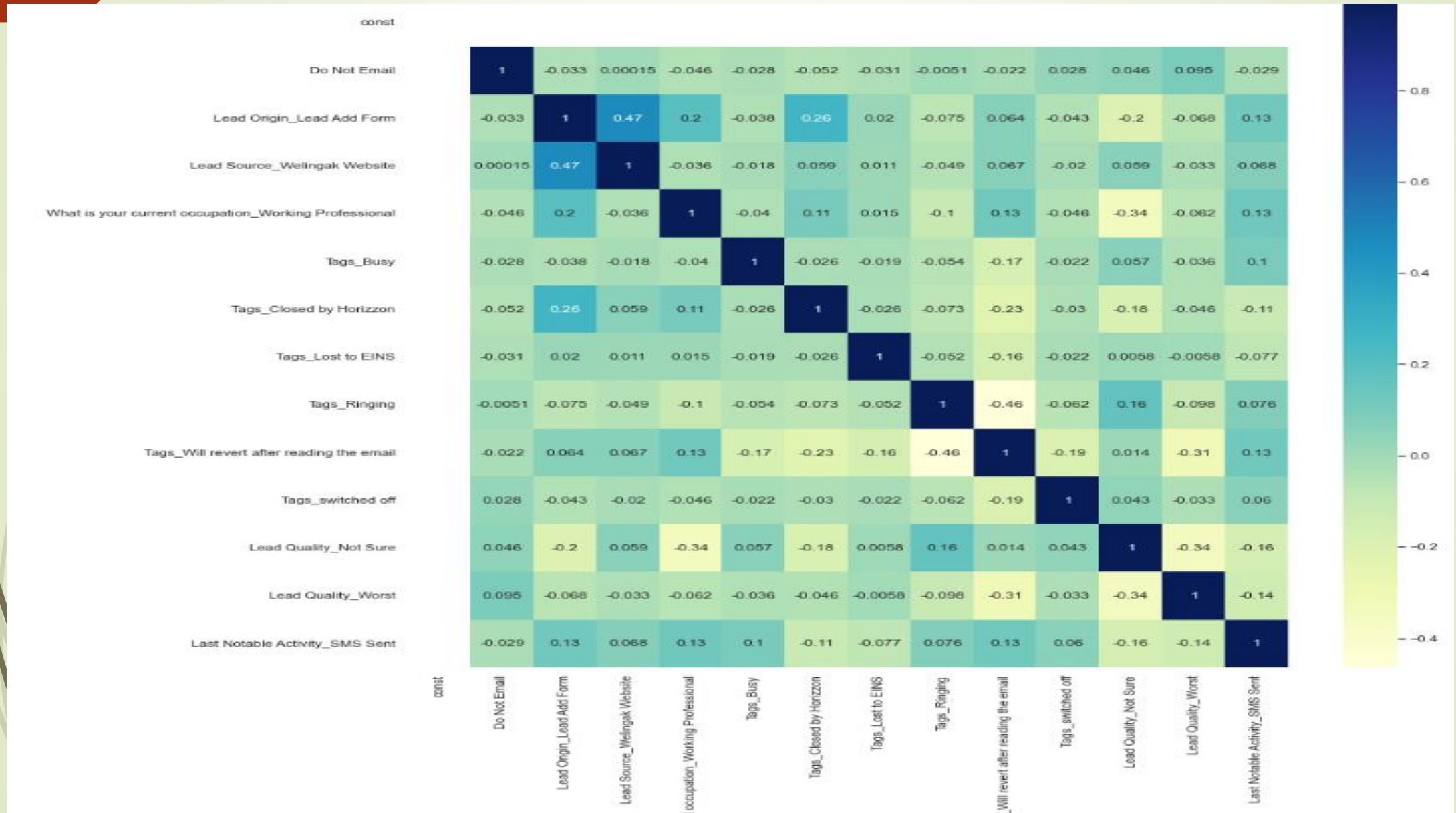
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6337
Model Family:	Binomial	Df Model:	13
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1588.8
Date:	Tue, 19 Dec 2023	Deviance:	3177.6
Time:	19:16:55	Pearson chi2:	3.08e+04
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.0888	0.216	-9.654	0.000	-2.513	-1.665
Do Not Email	-1.3012	0.212	-6.134	0.000	-1.717	-0.885
Lead Origin_Lead Add Form	1.0894	0.363	3.001	0.003	0.378	1.801
Lead Source_Welingak Website	3.4138	0.818	4.173	0.000	1.810	5.017
What is your current occupation_Working Professional	1.3403	0.291	4.602	0.000	0.769	1.911
Tags_Busy	3.8040	0.330	11.532	0.000	3.157	4.450
Tags_Closed by Horizzon	7.9562	0.763	10.433	0.000	6.461	9.451
Tags_Lost to EINS	9.1785	0.754	12.177	0.000	7.701	10.656
Tags_Ringing	-1.6947	0.337	-5.036	0.000	-2.354	-1.035
Tags_Will revert after reading the email	3.9665	0.229	17.311	0.000	3.517	4.416
Tags_switched off	-2.2882	0.587	-3.900	0.000	-3.438	-1.138
Lead Quality_Not Sure	-3.3406	0.128	-26.026	0.000	-3.592	-3.089
Lead Quality_Worst	-3.7624	0.850	-4.426	0.000	-5.428	-2.096
Last Notable Activity_SMS Sent	2.7406	0.120	22.847	0.000	2.506	2.976

Contd.

- Correlation

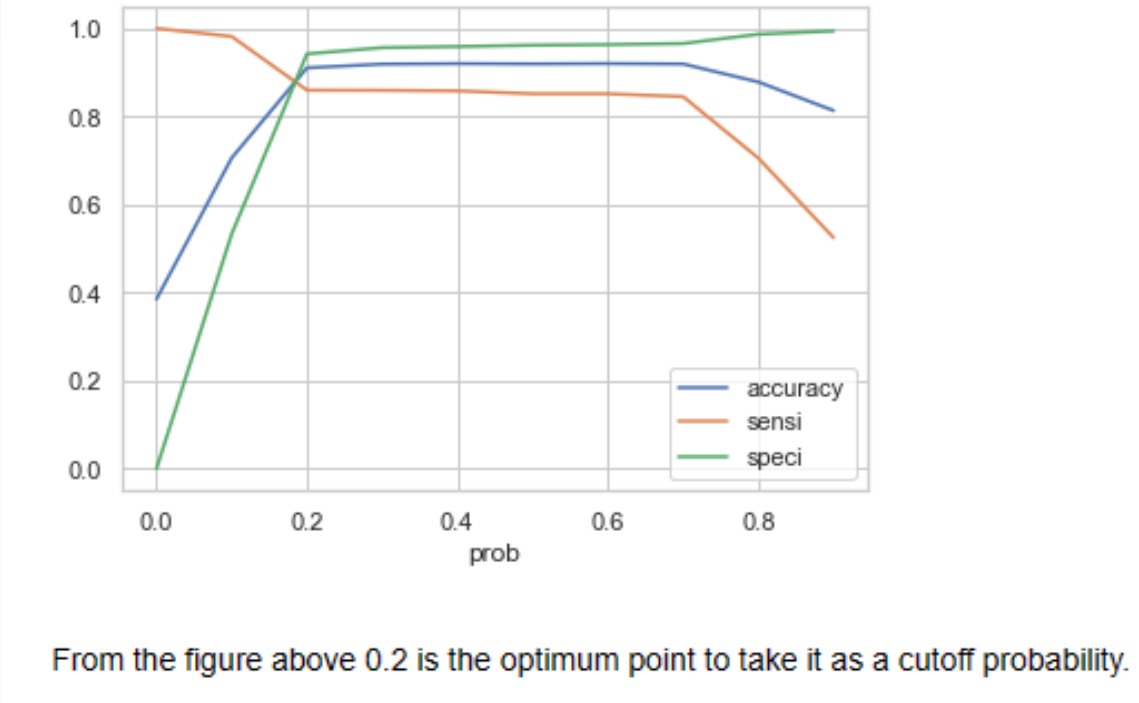


Model Evaluation

- The graph depicts the optimal cutoff probability to be around 0.2
- Now with 0.2 we got:
  - Confusion matrix

```
Confusion Matrix:
[[3679  226]
 [ 343 2103]]
```

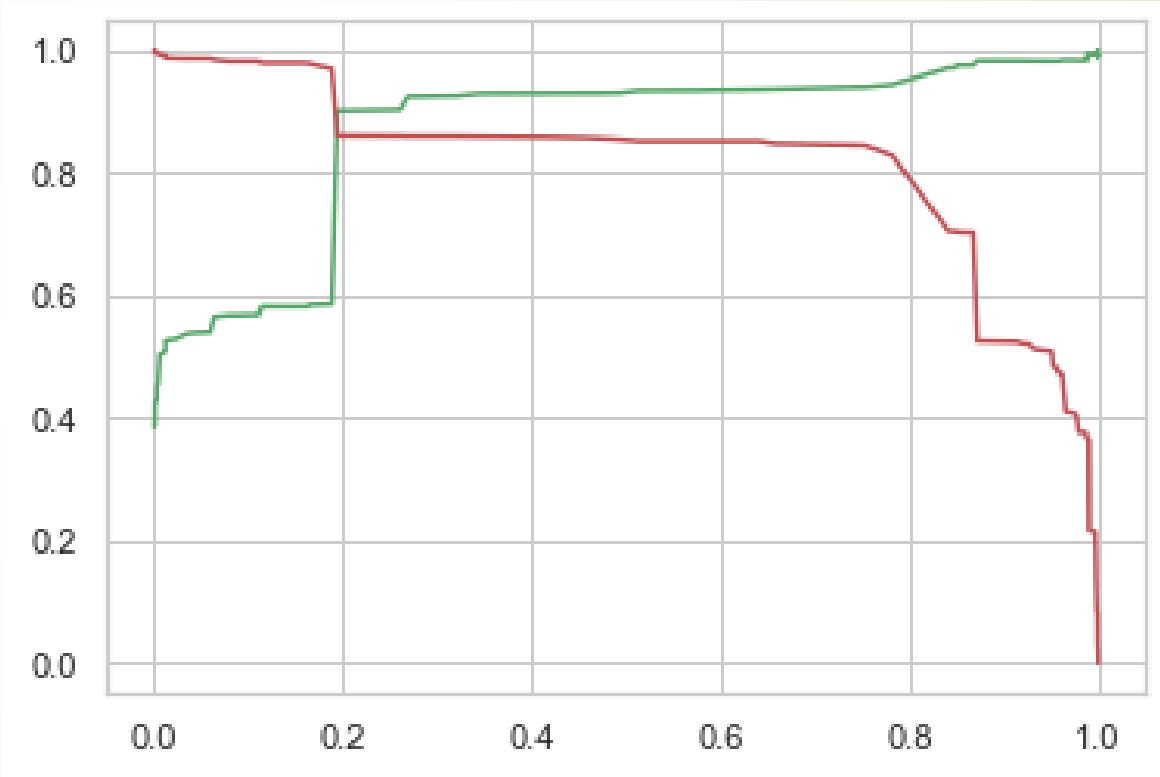
  - Accuracy = 91%
  - Sensitivity = 86%
  - Specificity = 94%



From the figure above 0.2 is the optimum point to take it as a cutoff probability.

Model Evaluation

- Precision and Recall
  - Precision score = 93%
  - Recall score = 85%



Model Evaluation

- On Test data set we got:
  - Accuracy = 90%
  - Sensitivity = 84%
  - Specificity = 94%

```
Confusion Matrix:
[[1635   99]
 [ 155  834]]
```

Conclusions

- Accuracy, sensitivity and the specificity of the test dataset are close to the ones of the train dataset respectively.
- Some of the important features to be focused upon are:
  - Lead Origin – Lead Add Form
  - Working professionals in What is your current occupation
  - Lead source from Reference and welingak website seems to have high conversion rate
  - In last notable activity is SMS Sent seems to have high conversion rate.
- These are some of the important features which have high conversion rate and should be focused upon for calling.
- The model seems good to go given the results.