LEAD SCORING CASE STUDY

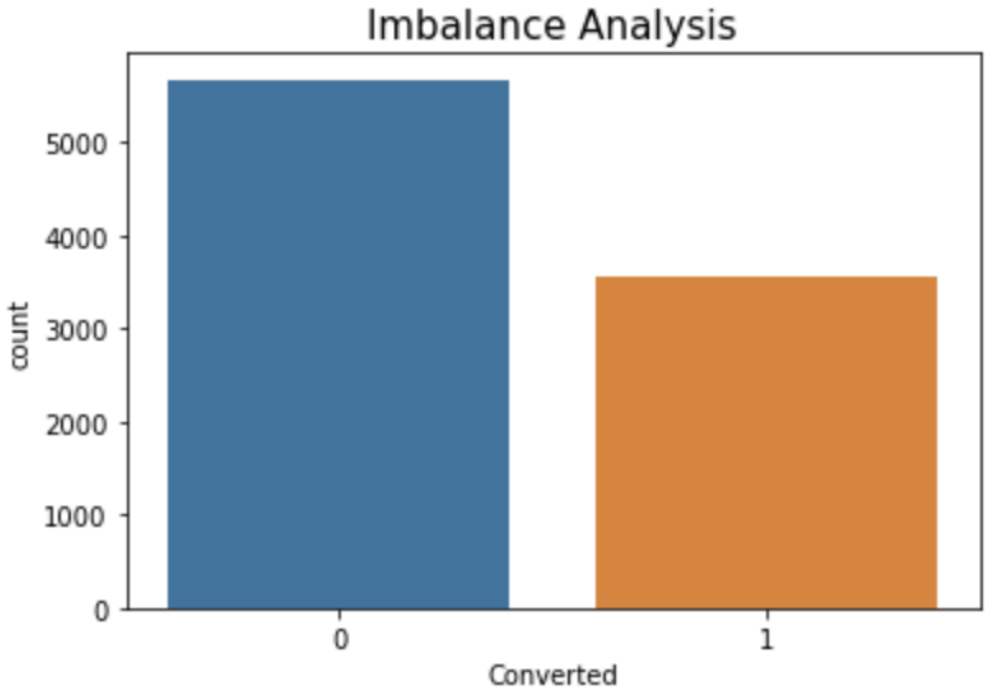
SWETA KUMARI SHAW

SUKEERTHI G

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.
- X Education needs to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company requires to build a model wherein it needs 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 INSPECTION- DATA IMBALANCE ANALYSIS



- Percentage of not converted leads 61.46%
- Percentage of converted successfully 38.54%
- Imbalance Ratio 1.59
- We can see ~ 62:38 ratio . This means data is slightly imbalance . Hence we can proceed.

DATA CLEANING

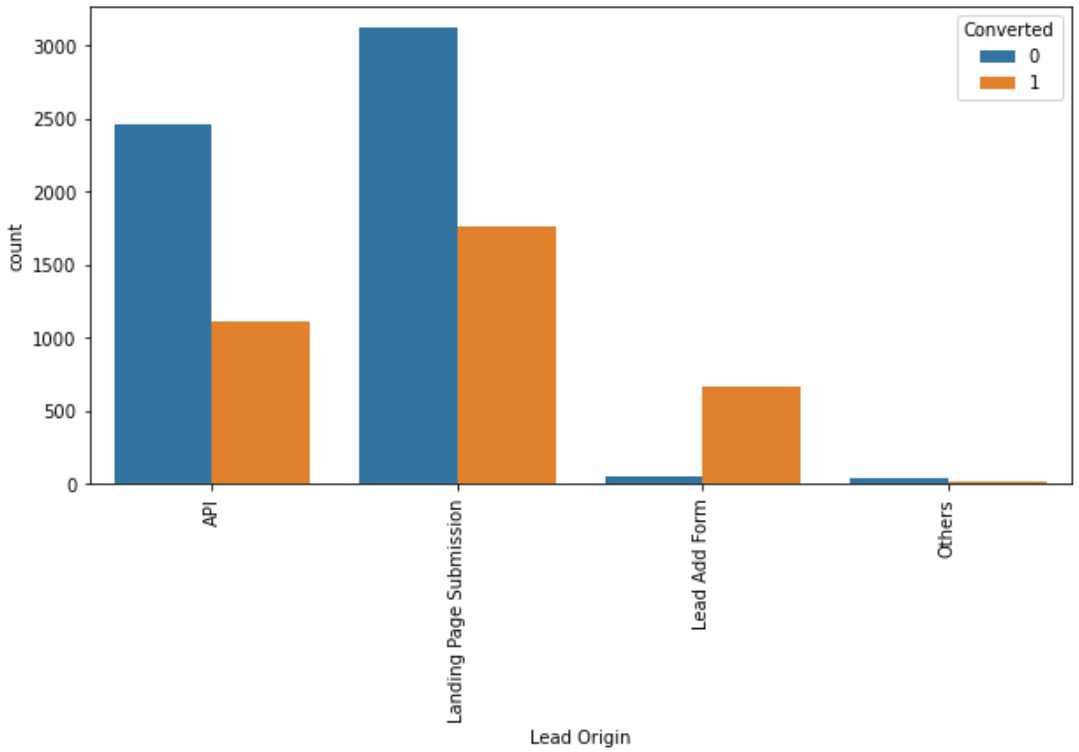
- Dropping these columns Prospect ID and Lead Number as they are unique identifiers.
- Replacing 'Select' value with Null
- Checking for duplicates
- Imputing new levels (merging different levels or using mean/median/mode) in some columns to handle data imbalance:
    1. Specialization
    2. What is your current occupation
    3. TotalVisits
    4. Page Views Per Visit
    5. Lead Source
    6. Lead Origin

DATA CLEANING

- Handling null values: Columns dropped due to high percentage of nulls or skewness of the data-

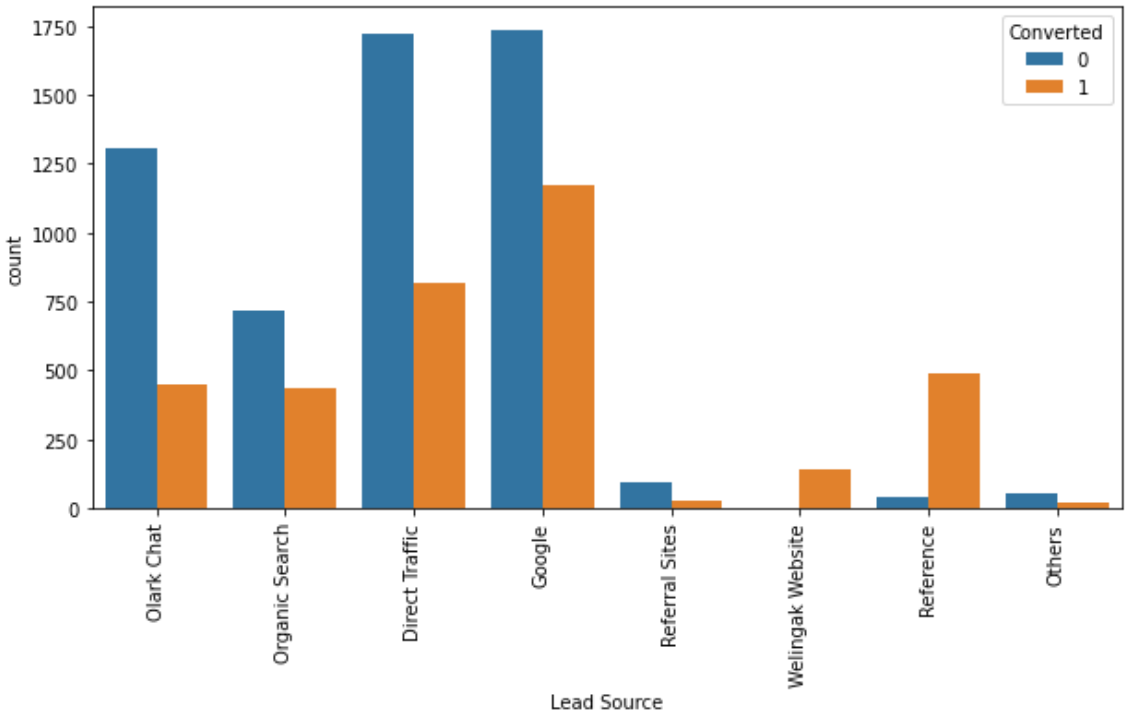
1. City
2. What matters most to you in choosing a course
3. Country
4. Newspaper Article
5. I agree to pay the amount through cheque
6. Get updates on DM Content
7. Update me on Supply Chain Content
8. Receive More Updates About Our Courses
9. Through Recommendations
10. Digital Advertisement
11. Newspaper
12. X Education Forums
13. Magazine
14. Search
15. Do Not Call

UNIVARIATE ANALYSIS- LEAD ORIGIN



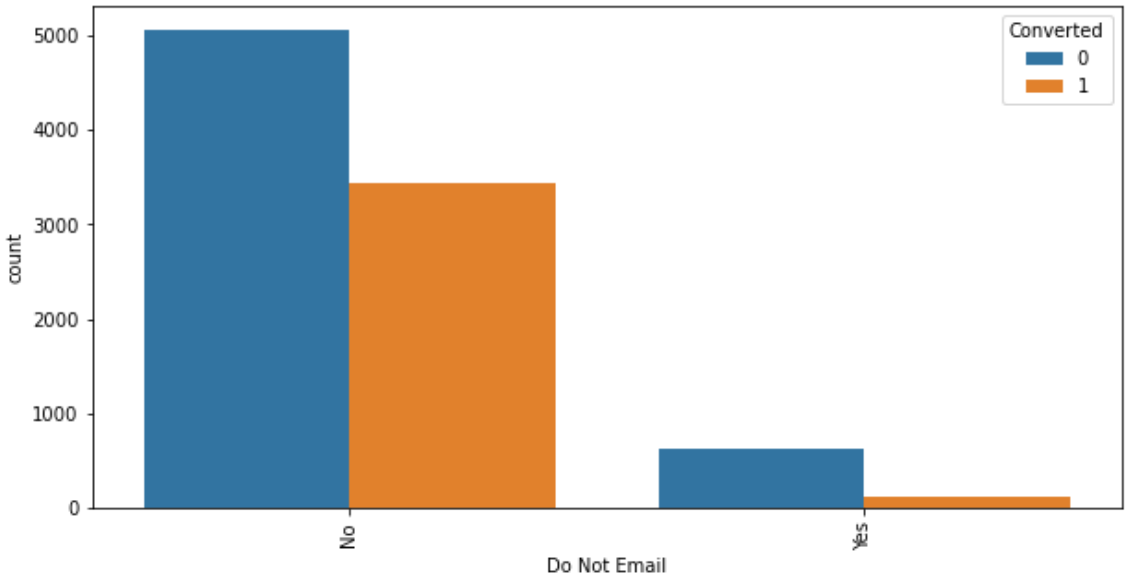
- For API and Landing page submission , lead conversion count is almost half as compared to non-conversion count. Hence we can say , there is only 30 -35 % conversion rate . But count of lead originated from these points are high.
- For 'Lead Add Form' , we can say conversion rate is significantly high. But count of lead originated is not very high.
- To improve conversion rate , we need to focus on below two things:
  - Improving conversion rate of lead origin API and Landing Page Submission.
  - Improving count of lead generation from Lead Add From.

UNIVARIATE ANALYSIS- LEAD SOURCE



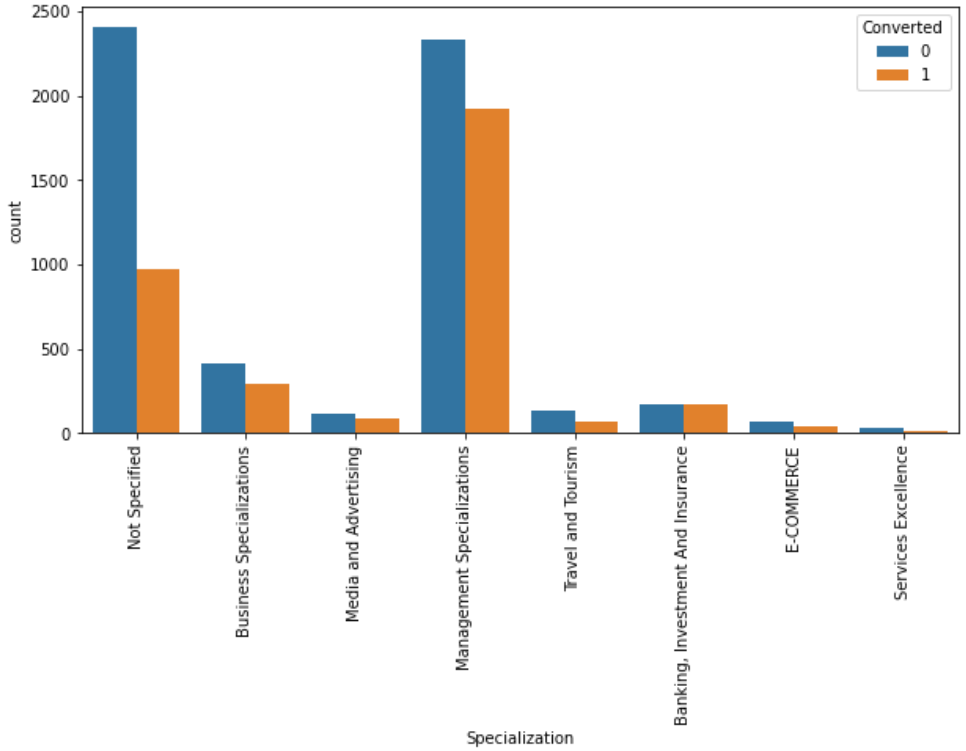
- Maximum leads are getting generated from Google and Direct Traffic. While Google has better leads conversion rate as compare to Direct Traffic.
- Leads conversion rate from Reference and welingak website are significantly high.
- To improve conversion rate , we need to focus on below things:
  - Improving conversion rate of lead source Google and Direct Traffic.
  - Improving conversion rate of lead source olark chat and organic search as we have significant count of lead generation from these two sources.
  - Improving count of lead generation from Reference and welingak website as these are having good conversion rate.

UNIVARIATE ANALYSIS- DO NOT EMAIL



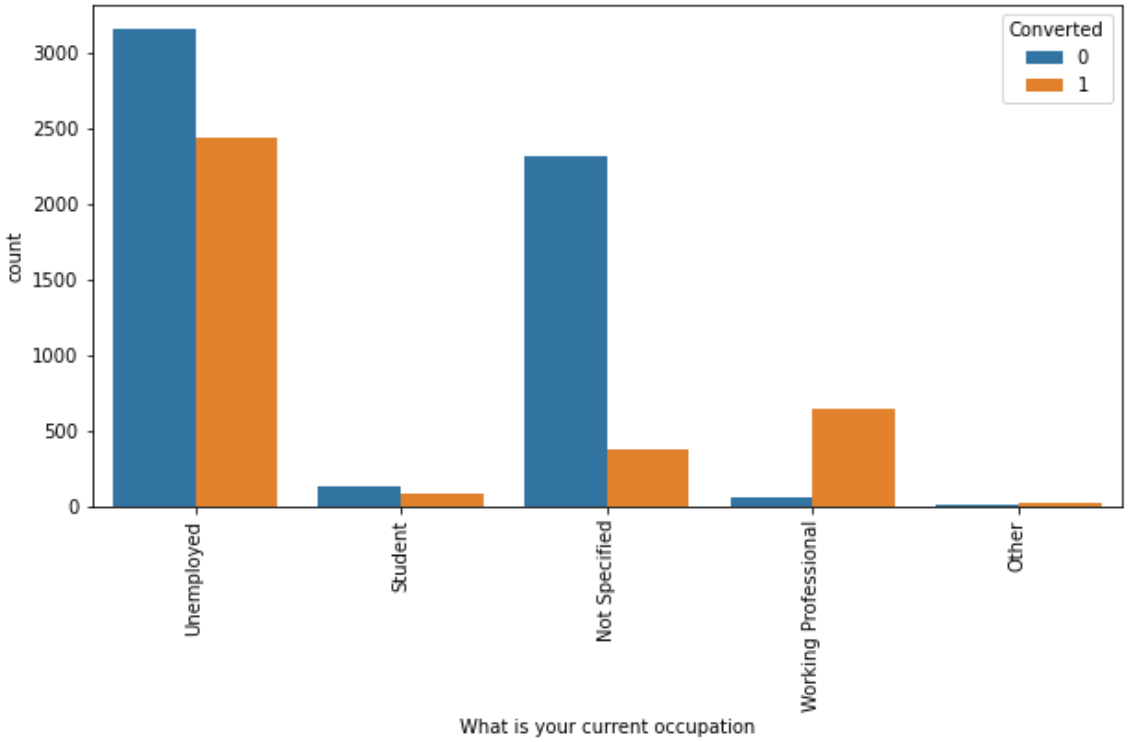
- Maximum leads are getting generated where Do Not Email is opted as No though conversion rate is not that good.
- We need to focus on improving conversion rate of those leads where Do Not Email is opted as No.

UNIVARIATE ANALYSIS- SPECIALIZATION



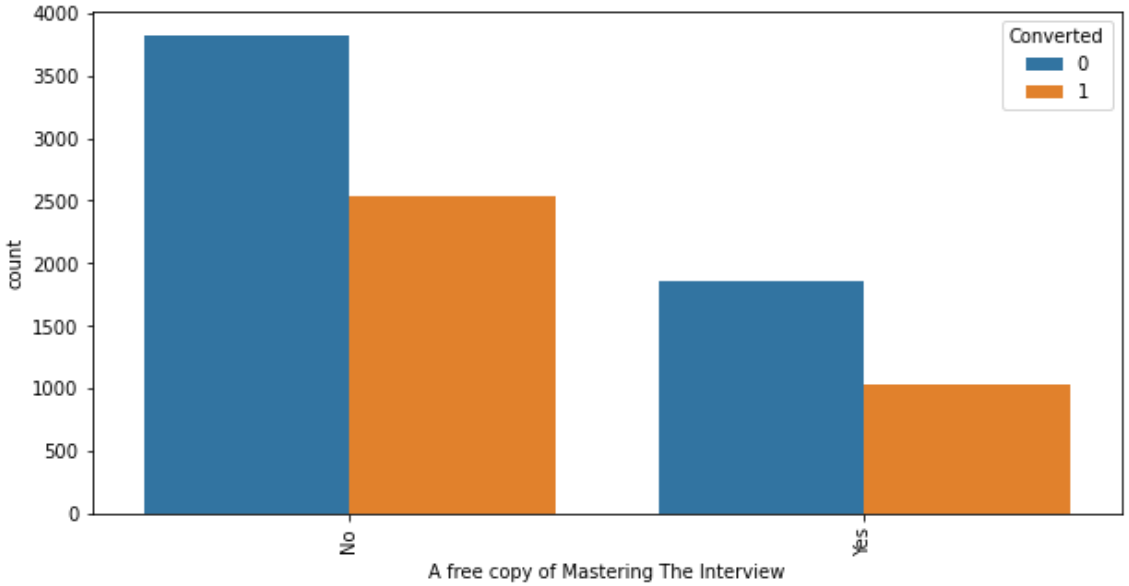
- Customers belonging to Management Specialization are responding good . We have high leads conversion rate there.
- Customers belonging to Banking , Investment and Insurance are having mix response. We have almost 50% conversion rate.
- Customers with Services Excellence specialization are showing very low conversion rate.

UNIVARIATE ANALYSIS- OCCUPATION



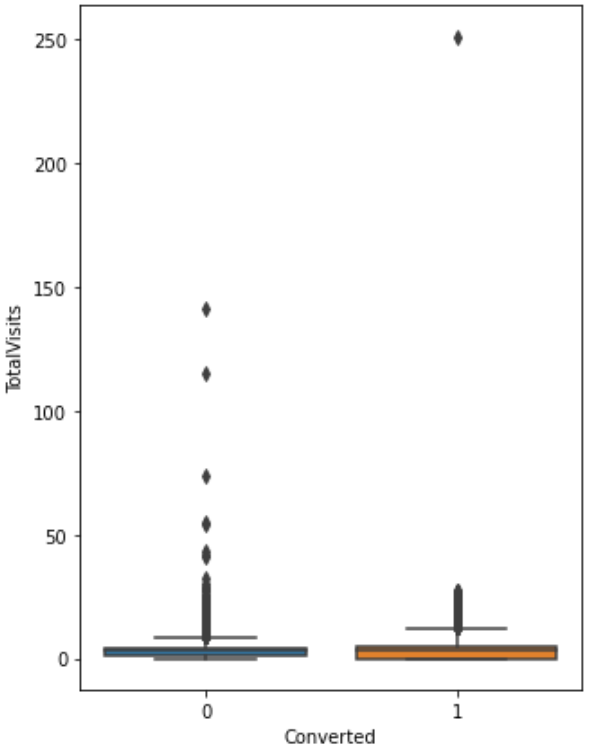
- Working professional are having good conversion rates while from Unemployed are holding most leads
- To improve conversion rate , we need to focus on below things:
  - Improving conversion rate for Unemployed.
  - Improving leads generation from Working professional.

UNIVARIATE ANALYSIS- A FREE COPY OF MASTERING THE INTERVIEW



- For both value Yes and No , leads conversion rate is not that good but 'Yes' is having better conversion rate as compare to No though we have more leads from No.

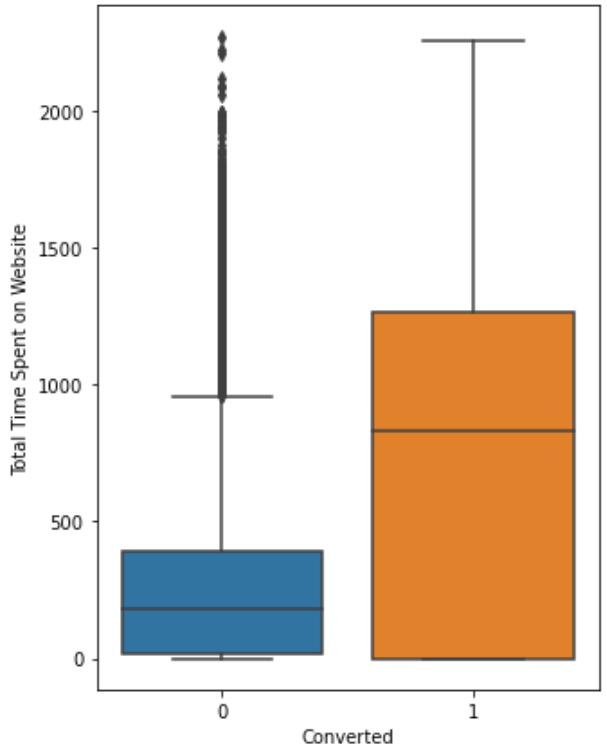
HANDLING OUTLIERS- TOTAL VISITS




```
count    9240.000000
mean        3.438636
std         4.819024
min         0.000000
5%          0.000000
25%         1.000000
50%         3.000000
75%         5.000000
90%         7.000000
99%        17.000000
max       251.000000
Name: TotalVisits, dtype: float64
```

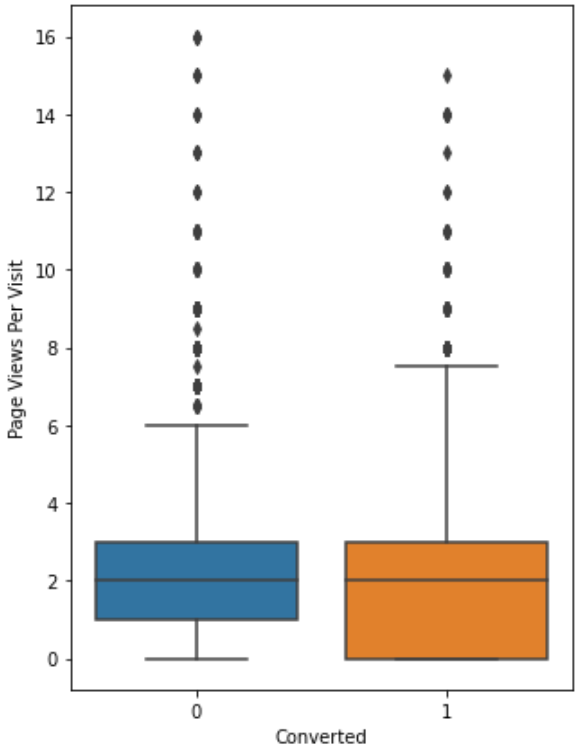
- We can see outliers present .This need to be handled.
- We can see top 1% is holding outliers. We can remove this.

HANDLING OUTLIERS- TOTAL TIME SPENT ON WEBSITE



- Converted Leads are having more Total Time Spent on Website.

HANDLING OUTLIERS- PAGE VIEWS PER VISIT




```
count    9157.000000
mean        2.332225
std         2.047285
min         0.000000
5%          0.000000
25%         1.000000
50%         2.000000
75%         3.000000
90%         5.000000
99%         9.000000
max        16.000000
Name: Page Views Per Visit, dtype: float64
```

We can see some outliers present . This need to be handled.

We can see only top 1% is holding outliers. We can remove this.

DATA PREPARATION

- Dummy variables creation
- Train-Test data split for modelling
- Feature scaling

MODEL BUILDING

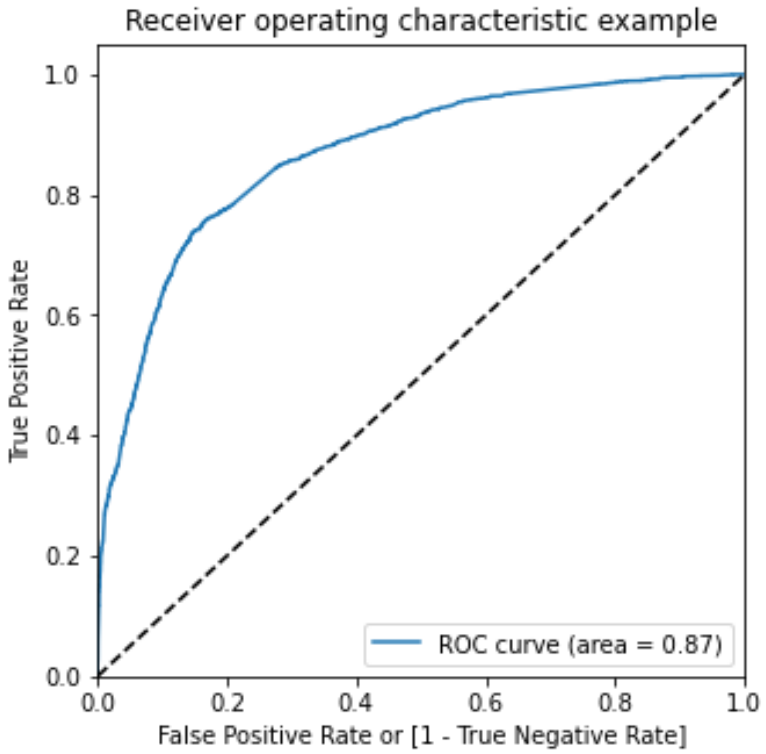
- Initial model using Stats model
- Feature selection using RFE
- Using RFE features and start building the model using sklearn by removing feature with high p-value and/or high VIF
- Finalized the model based on the threshold: p-values<0.05 and VIF<5

MODEL USED FOR PREDICTION

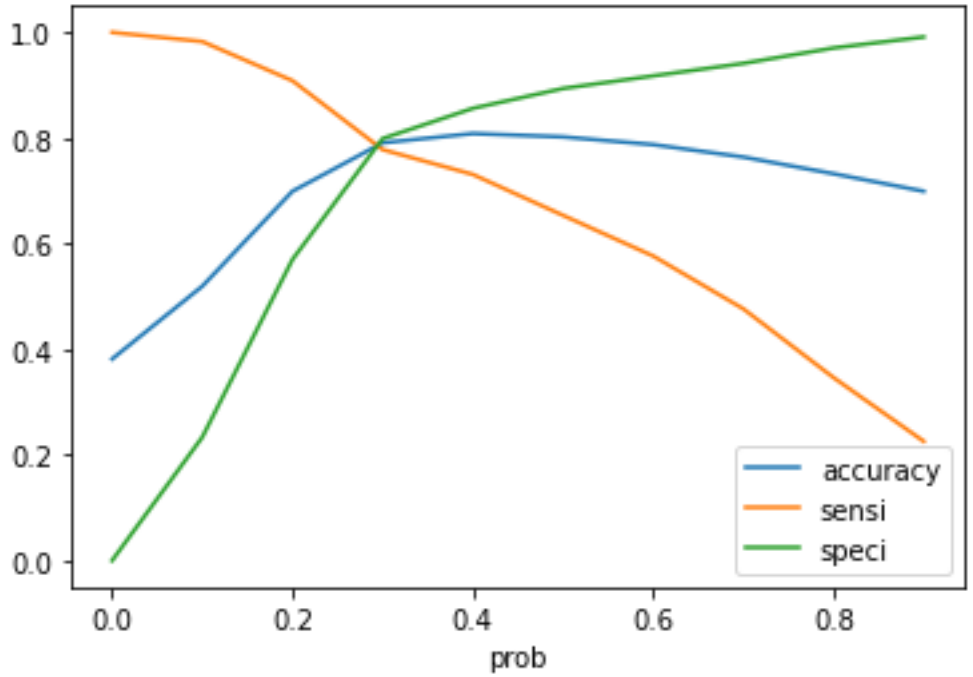
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.5982	0.151	-10.574	0.000	-1.894	-1.302
Do Not Email	-1.2319	0.161	-7.661	0.000	-1.547	-0.917
Total Time Spent on Website	1.0653	0.039	27.526	0.000	0.989	1.141
Lead Origin_Landing Page Submission	-0.5490	0.125	-4.399	0.000	-0.794	-0.304
Lead Origin_Lead Add Form	3.3805	0.210	16.095	0.000	2.969	3.792
Lead Source_Google	0.3262	0.077	4.221	0.000	0.175	0.478
Lead Source_Olark Chat	1.1403	0.125	9.094	0.000	0.895	1.386
Lead Source_Welingak Website	3.2657	1.026	3.185	0.001	1.256	5.276
Specialization_Not Specified	-0.7903	0.118	-6.717	0.000	-1.021	-0.560
What is your current occupation_Other	2.2445	0.543	4.131	0.000	1.180	3.309
What is your current occupation_Student	1.0134	0.223	4.553	0.000	0.577	1.450
What is your current occupation_Unemployed	1.2928	0.085	15.207	0.000	1.126	1.459
What is your current occupation_Working Professional	3.6614	0.195	18.790	0.000	3.279	4.043

	Features	VIF
10	What is your current occupation_Unemployed	2.66
7	Specialization_Not Specified	2.41
2	Lead Origin_Landing Page Submission	2.38
5	Lead Source_Olark Chat	2.02
3	Lead Origin_Lead Add Form	1.70
4	Lead Source_Google	1.63
11	What is your current occupation_Working Professional	1.33
6	Lead Source_Welingak Website	1.32
1	Total Time Spent on Website	1.26
0	Do Not Email	1.11
9	What is your current occupation_Student	1.06
8	What is your current occupation_Other	1.01

MODEL VALIDATION- ROC CURVE



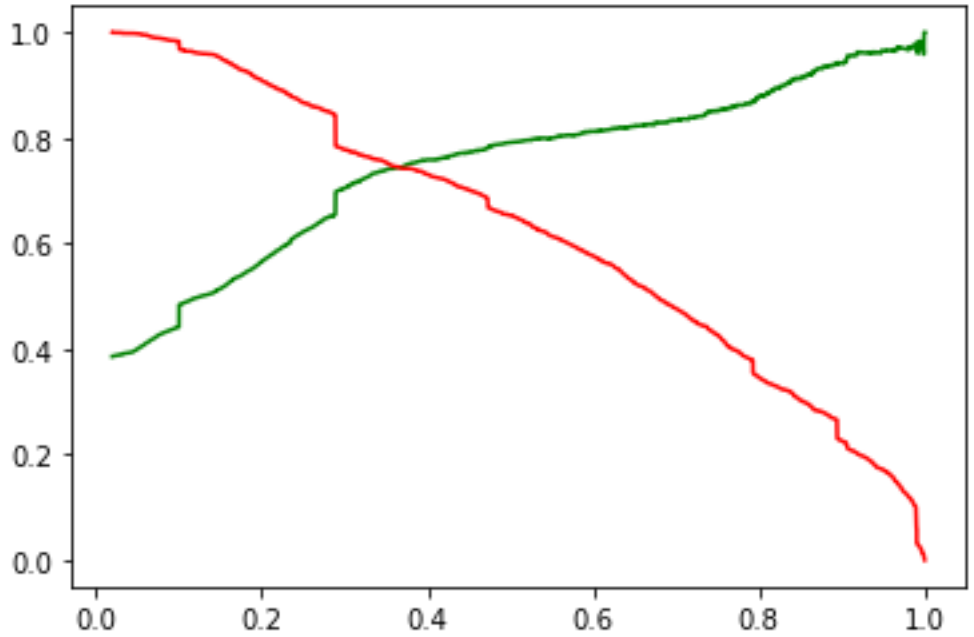
MODEL VALIDATION-FINDING OPTIMAL CUTOFF- 0.3



MODEL EVALUATION- CONFUSION MATRIX METRICS

SL No.	Metrices	Values (%)
1	Accuracy	79.09
2	Sensitivity	77.80
3	Specificity	79.89
4	Positive predictive values	70.48
5.	Negative predictive values	85.36

MODEL EVALUATION- PRECISION AND RECALL



- Precision score: 0.7048
- Recall score: 0.7780

PREDICTION OF THE LEAD PROBABILITY

	Converted	Prospect ID	Converted_Probability	Final_Pred
0	1	6906	0.766787	1
1	0	1873	0.130077	0
2	0	771	0.165343	0
3	0	4495	0.232898	0
4	1	9061	0.599086	1

PREDICTION ON TEST DATA SET

SL No.	Metrices	Values (%)
1	Accuracy	78.58
2	Sensitivity	76.47
3	Specificity	79.93
4	Positive predictive values	71.01
5.	Negative predictive values	84.09

LEAD SCORE ASSIGNMENT

- We have also assigned 'Lead Score' which can help to identify potential of each lead.

	Prospect ID	Converted_Probability	Converted	Final_Pred	Lead Score
2722	1939	0.317795	0	1	32
2723	1540	0.504925	1	1	50
2724	5198	0.094328	1	0	9
2725	8660	0.100340	0	0	10
2726	6219	0.210999	0	0	21

CONCLUSION

- Top 3 variables which are contributing the most to the model
  - What is your current occupation_Working Professional
    - ✓ This indicates that working professionals are more prone to be converting leads.
  - Lead Origin_Lead Add Form
    - ✓ This indicates that Leads originated form add form , are more likely to get converted.
  - Lead Source_Welingak Website
    - ✓ This indicated that from Welingak website are getting good number of convertible leads.

RECOMMENDATIONS FOR THE SALES TEAM

- To improve conversion rate , we need to focus on below things:
  - Improving leads generation from **Working professional**.
  - Improving count of **lead generation from Lead Add From**.
  - Improving count of lead generation from **Reference and welingak website** as these are having good conversion rate.
  - Improving conversion rate for **Unemployed.**
  - Improving conversion rate of **lead source olark chat** and **organic search** as we have significant count of lead generation from these two sources.
  - Improving conversion rate of those leads where we have more **Total Time Spent on Website**.
  - Improving conversion rate for **Student.**
  - Improving conversion rate of **lead origin API** and **Landing Page Submission**.
  - Improving conversion rate of **lead source Google** and **Direct Traffic**.
  - We need to focus on improving conversion rate of those leads where **Do Not Email is opted as No**.

THANK YOU