LEAD SCORING CASE STUDY

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.
- 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%.
- 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.
- 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%.

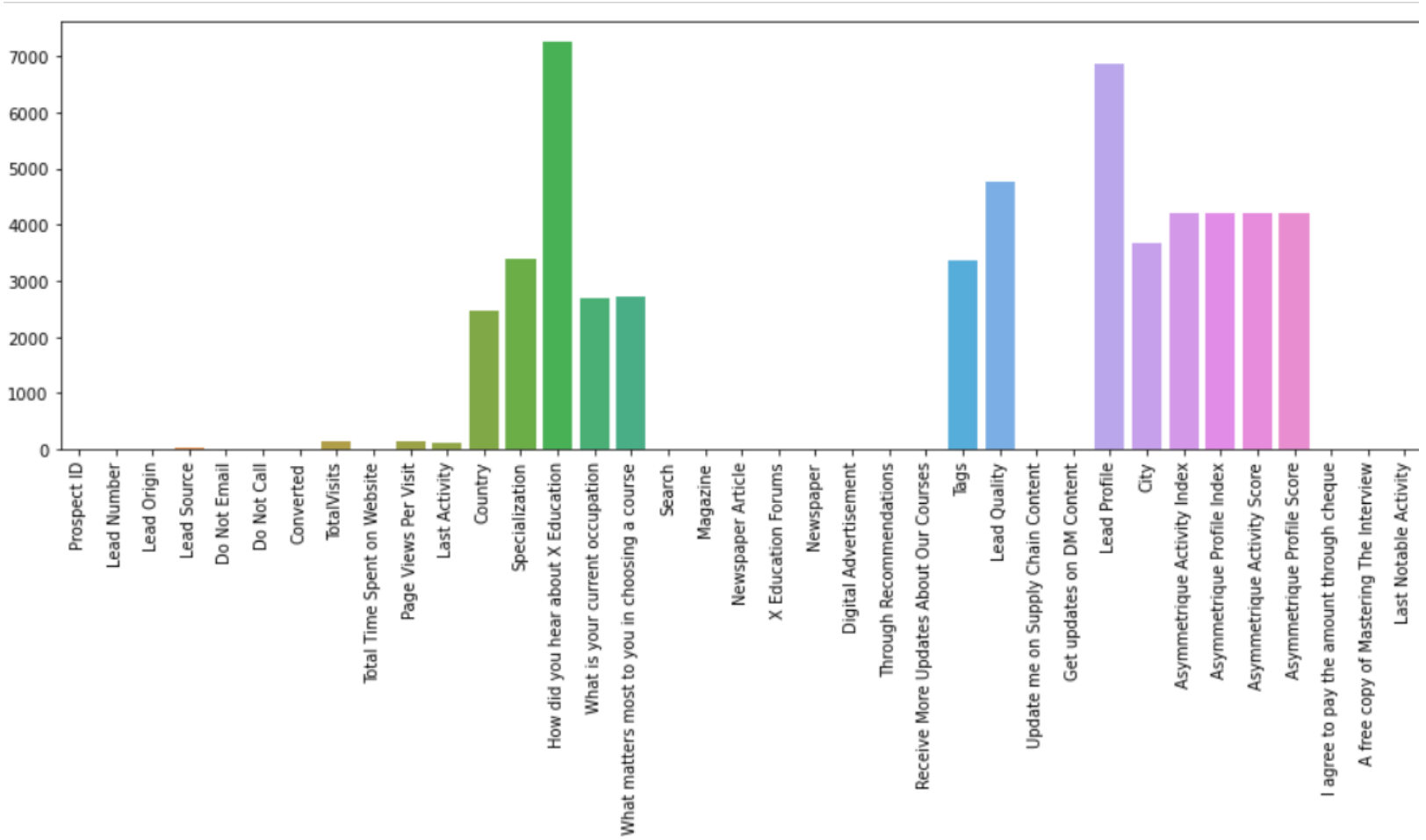
GOALS

- There are quite a few goals for this case study.
- To 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.

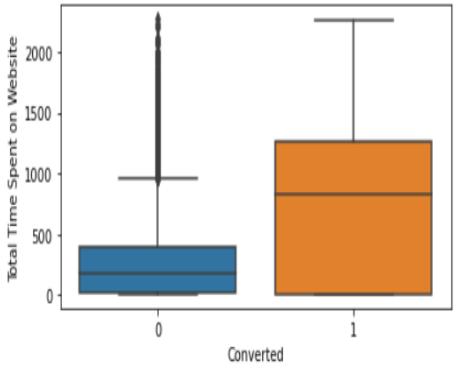
ANALYSIS

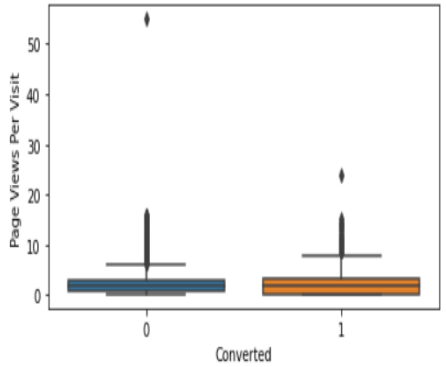
- Check missing values or Nan values. Reporting how to handle them.
- Checking outliers in the variables.
- Exploratory Data Analysis

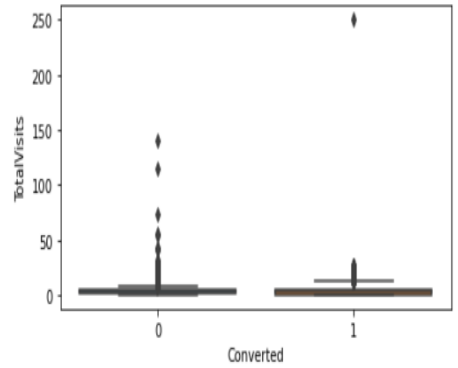
MISSING VALUE COLUMNS



IMPUTATION





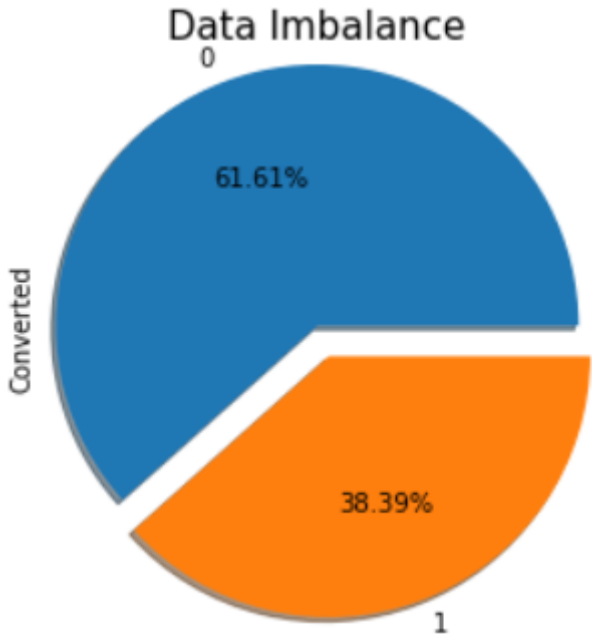


People who are sending more time on websites are prone to conversion. Outliers are present in for the ones not converting which means there are some people who are spending more time but still not taking the course.

For total number of visits, the median value is almost same for both although the upper quantile is higher for the ones converting. Also there are outliers present here.

For Page Views per visits, the median value is almost same for both. There is one outlier who has visited a lot pages every time he/she visits but has not converted yet.

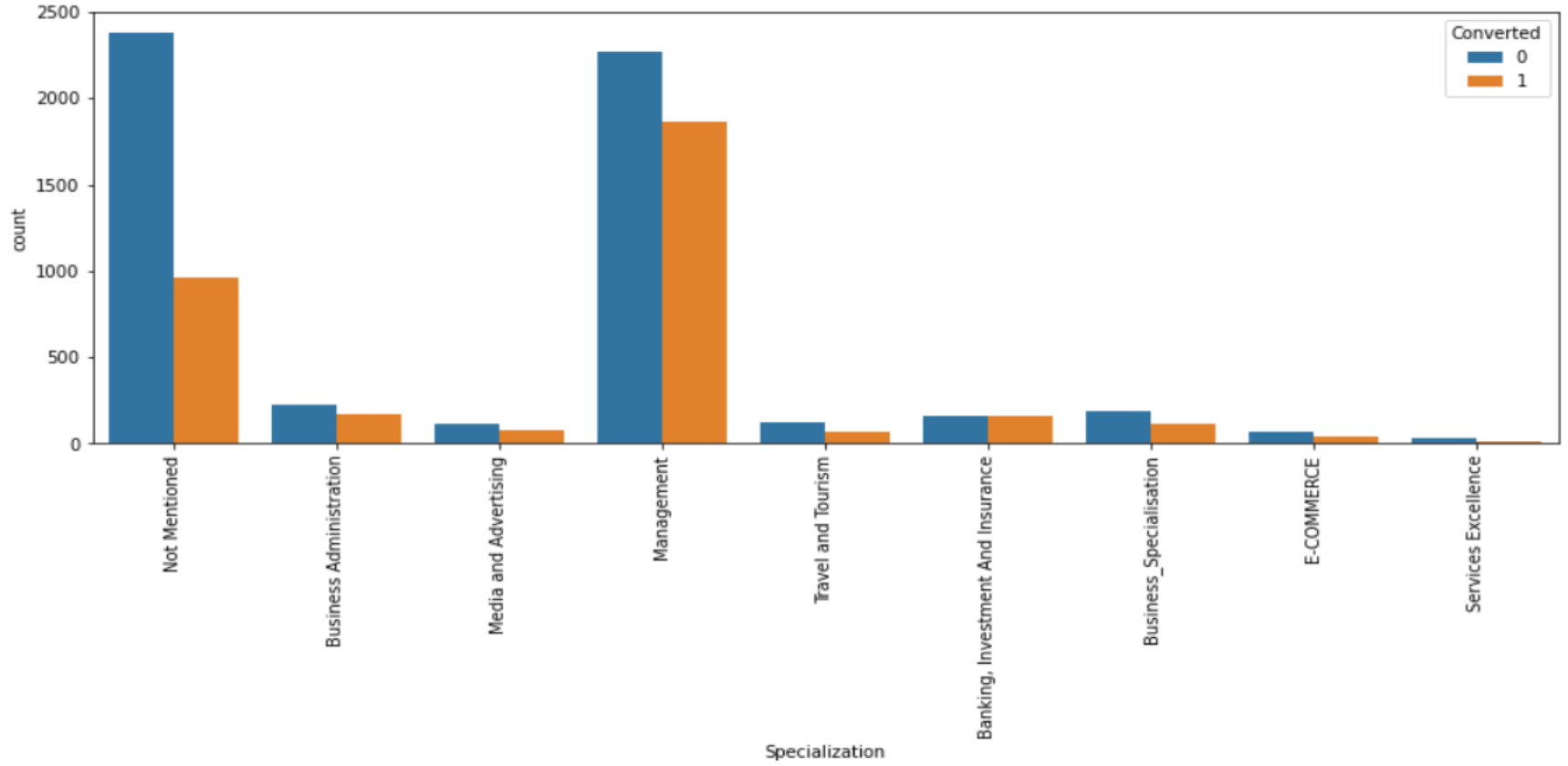
IMBALANCE CHECKING



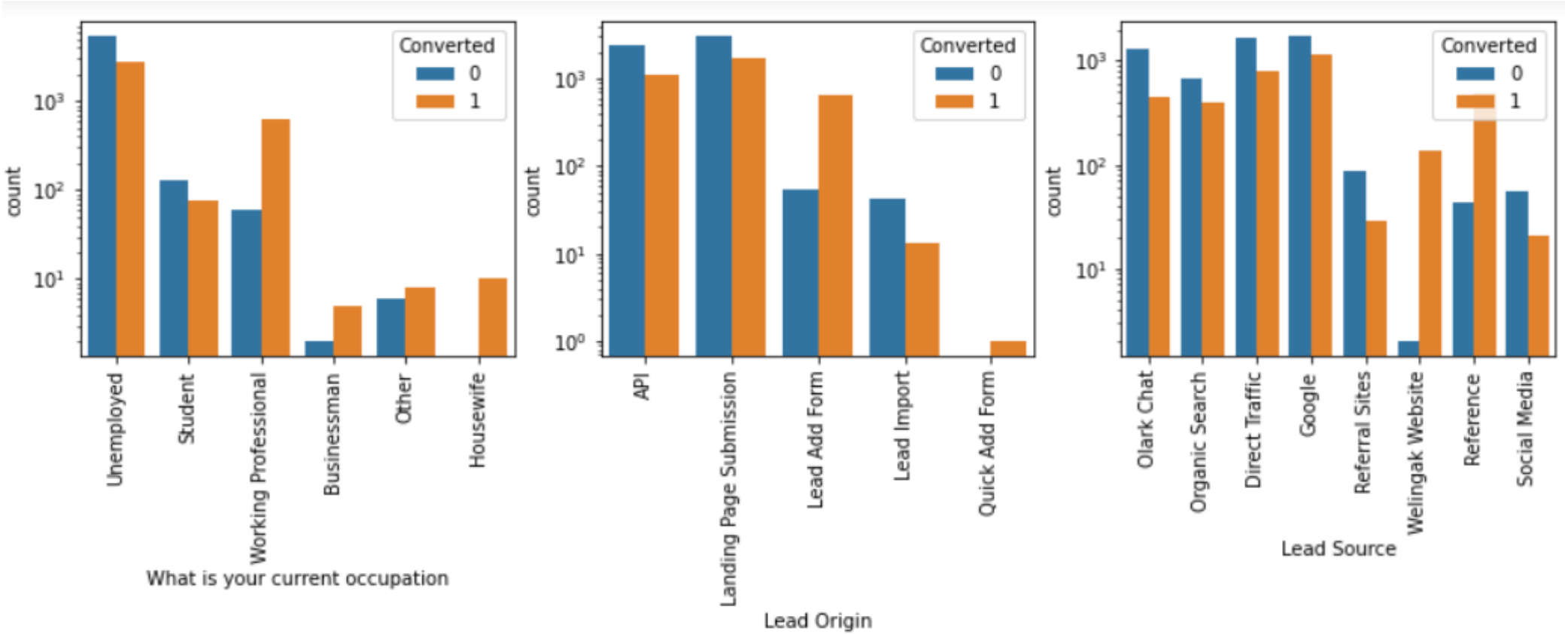
lead conversion rate is 39%

EXPLORATORY DATA ANALYSIS

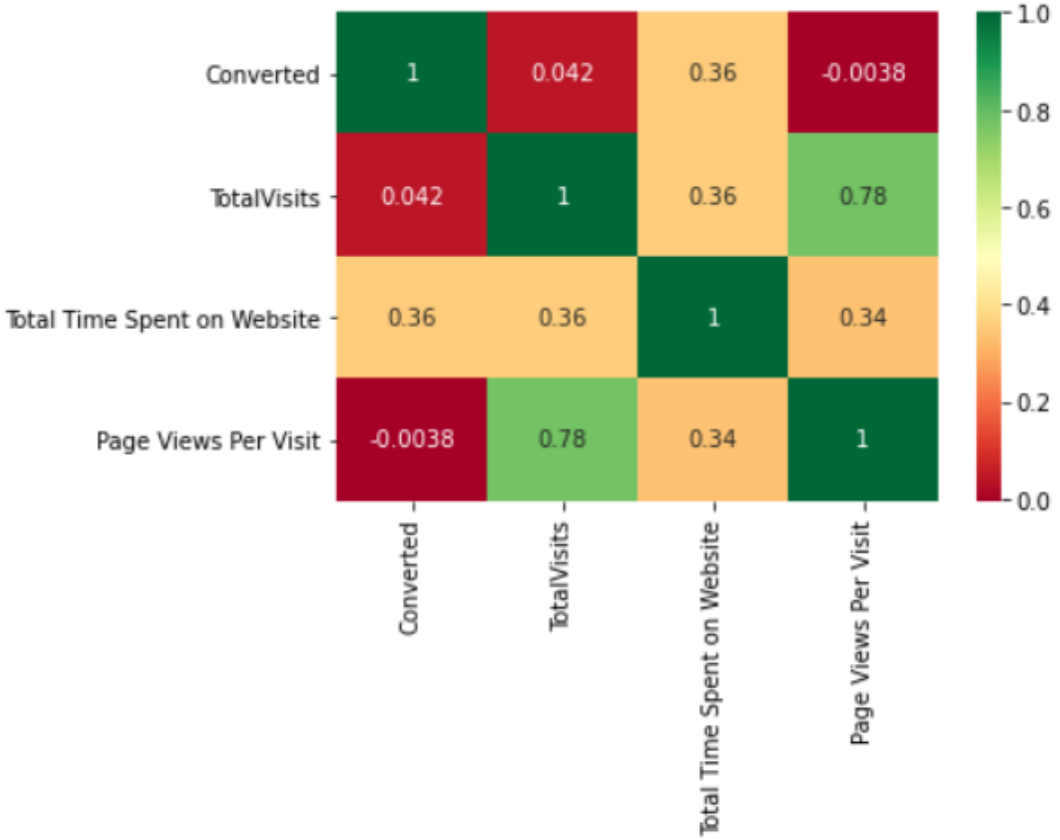
CATEGORICAL ANALYSIS



CATEGORICAL ANALYSIS



MULTIVARIATE ANALYSIS



DUMMY CREATION

```
Data columns (total 36 columns):
 #   Column                                        Non-Null Count  Dtype
---  ------                                        --------------  -----
 0   Do Not Email                                  9018 non-null   int64
 1   Do Not Call                                   9018 non-null   int64
 2   Converted                                     9018 non-null   int64
 3   TotalVisits                                   9018 non-null   float64
 4   Total Time Spent on Website                   9018 non-null   int64
 5   Page Views Per Visit                          9018 non-null   float64
 6   Search                                        9018 non-null   int64
 7   Newspaper Article                             9018 non-null   int64
 8   Newspaper                                     9018 non-null   int64
 9   Digital Advertisement                         9018 non-null   int64
 10  Through Recommendations                       9018 non-null   int64
 11  A free copy of Mastering The Interview        9018 non-null   int64
 12  Lead Source_Google                            9018 non-null   uint8
 13  Lead Source_Olark Chat                        9018 non-null   uint8
 14  Lead Source_Organic Search                    9018 non-null   uint8
 15  Lead Source_Reference                         9018 non-null   uint8
 16  Lead Source_Referral Sites                    9018 non-null   uint8
 17  Lead Source_Social Media                      9018 non-null   uint8
 18  Lead Source_Welingak Website                  9018 non-null   uint8
 19  Lead Origin_Landing Page Submission           9018 non-null   uint8
 20  Lead Origin_Lead Add Form                     9018 non-null   uint8
 21  Lead Origin_Lead Import                       9018 non-null   uint8
 22  Lead Origin_Quick Add Form                    9018 non-null   uint8
 23  Specialization_Business Administration        9018 non-null   uint8
 24  Specialization_Business_Specialisation        9018 non-null   uint8
 25  Specialization_E-COMMERCE                     9018 non-null   uint8
 26  Specialization_Management                     9018 non-null   uint8
 27  Specialization_Media and Advertising          9018 non-null   uint8
 28  Specialization_Not Mentioned                  9018 non-null   uint8
 29  Specialization_Services Excellence            9018 non-null   uint8
 30  Specialization_Travel and Tourism             9018 non-null   uint8
 31  What is your current occupation_Housewife     9018 non-null   uint8
 32  What is your current occupation_Other         9018 non-null   uint8
 33  What is your current occupation_Student       9018 non-null   uint8
 34  What is your current occupation_Unemployed    9018 non-null   uint8
 35  What is your current occupation_Working Professional  9018 non-null   uint8
dtypes: float64(2), int64(10), uint8(24)
memory usage: 1.4 MB
```

DATA CONVERSION

- Rows for analysis – 9018
- Columns for analysis – 36
- Numerical variables are normalized
- Dummy variables are created

MODEL BUILDIING

- Splitting data into train and test set
- Train size is 70% and test size is 30%
- We are using RFE for feature selection. We are using 15 variables for output
- Building model by removing variables with p-value greater than 0.05 and VIF greater than 5.
- Predictions on test set with final model.

MODEL 1

Dep. Variable:	Converted	**No. Observations:**	6312
Model:	GLM	**Df Residuals:**	6296
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2929.8
Date:	Sat, 10 Apr 2021	**Deviance:**	5859.7
Time:	23:54:21	**Pearson chi2:**	9.10e+03
No. Iterations:	21		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.6146	1.010	-0.609	0.543	-2.594	1.364
Do Not Email	-1.4022	0.161	-8.687	0.000	-1.719	-1.086
TotalVisits	0.7122	0.188	3.796	0.000	0.344	1.080
Total Time Spent on Website	4.3851	0.158	27.738	0.000	4.075	4.695
Newspaper	-24.4444	4.82e+04	-0.001	1.000	-9.45e+04	9.44e+04
Lead Source_Olark Chat	0.9047	0.121	7.493	0.000	0.668	1.141
Lead Source_Referral Sites	-0.6772	0.334	-2.026	0.043	-1.332	-0.022
Lead Source_Welingak Website	3.1932	1.028	3.107	0.002	1.179	5.208
Lead Origin_Landing Page Submission	-0.8999	0.122	-7.399	0.000	-1.138	-0.662
Lead Origin_Lead Add Form	3.7773	0.219	17.271	0.000	3.349	4.206
Specialization_Not Mentioned	-0.9921	0.118	-8.430	0.000	-1.223	-0.761
What is your current occupation_Housewife	21.8627	1.43e+04	0.002	0.999	-2.8e+04	2.81e+04
What is your current occupation_Other	-1.9138	1.308	-1.463	0.143	-4.477	0.649
What is your current occupation_Student	-0.7162	1.027	-0.697	0.486	-2.729	1.296
What is your current occupation_Unemployed	-0.6980	1.003	-0.696	0.487	-2.665	1.269
What is your current occupation_Working Professional	1.9825	1.019	1.946	0.052	-0.014	3.979

	Features	VIF
13	What is your current occupation_Unemployed	16.63
7	Lead Origin_Landing Page Submission	7.24
9	Specialization_Not Mentioned	4.78
1	TotalVisits	3.95
4	Lead Source_Olark Chat	2.58
2	Total Time Spent on Website	2.27
14	What is your current occupation_Working Profes...	2.14
8	Lead Origin_Lead Add Form	2.12
12	What is your current occupation_Student	1.39
6	Lead Source_Welingak Website	1.33
0	Do Not Email	1.11
5	Lead Source_Referral Sites	1.05
11	What is your current occupation_Other	1.03
10	What is your current occupation_Housewife	1.02
3	Newspaper	1.00

MODEL 2

Dep. Variable:	Converted	**No. Observations:**	6312
Model:	GLM	**Df Residuals:**	6297
Model Family:	Binomial	**Df Model:**	14
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2930.1
Date:	Sat, 10 Apr 2021	**Deviance:**	5860.2
Time:	23:54:21	**Pearson chi2:**	9.11e+03
No. Iterations:	21		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3098	0.142	-9.201	0.000	-1.589	-1.031
Do Not Email	-1.4029	0.161	-8.691	0.000	-1.719	-1.087
TotalVisits	0.7124	0.188	3.797	0.000	0.345	1.080
Total Time Spent on Website	4.3860	0.158	27.748	0.000	4.076	4.696
Newspaper	-24.4455	4.82e+04	-0.001	1.000	-9.45e+04	9.44e+04
Lead Source_Olark Chat	0.9046	0.121	7.492	0.000	0.668	1.141
Lead Source_Referral Sites	-0.6784	0.334	-2.029	0.042	-1.334	-0.023
Lead Source_Welingak Website	3.1926	1.028	3.106	0.002	1.178	5.207
Lead Origin_Landing Page Submission	-0.9025	0.122	-7.419	0.000	-1.141	-0.664
Lead Origin_Lead Add Form	3.7773	0.219	17.272	0.000	3.349	4.206
Specialization_Not Mentioned	-0.9942	0.118	-8.445	0.000	-1.225	-0.763
What is your current occupation_Housewife	22.5600	1.43e+04	0.002	0.999	-2.8e+04	2.81e+04
What is your current occupation_Other	-1.2176	0.842	-1.447	0.148	-2.867	0.432
What is your current occupation_Student	-0.0190	0.224	-0.085	0.932	-0.457	0.419
What is your current occupation_Working Professional	2.6796	0.183	14.640	0.000	2.321	3.038

	Features	VIF
1	TotalVisits	3.18
7	Lead Origin_Landing Page Submission	2.95
9	Specialization_Not Mentioned	2.32
2	Total Time Spent on Website	2.10
4	Lead Source_Olark Chat	1.85
8	Lead Origin_Lead Add Form	1.43
6	Lead Source_Welingak Website	1.31
13	What is your current occupation_Working Profes...	1.18
0	Do Not Email	1.11
5	Lead Source_Referral Sites	1.05
12	What is your current occupation_Student	1.03
10	What is your current occupation_Housewife	1.01
3	Newspaper	1.00
11	What is your current occupation_Other	1.00

MODEL 3

Dep. Variable:	Converted	**No. Observations:**	6312
Model:	GLM	**Df Residuals:**	6298
Model Family:	Binomial	**Df Model:**	13
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2932.1
Date:	Sat, 10 Apr 2021	**Deviance:**	5864.2
Time:	23:54:22	**Pearson chi2:**	9.09e+03
No. Iterations:	21		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3019	0.142	-9.153	0.000	-1.581	-1.023
Do Not Email	-1.4016	0.161	-8.687	0.000	-1.718	-1.085
TotalVisits	0.6985	0.187	3.726	0.000	0.331	1.066
Total Time Spent on Website	4.3760	0.158	27.716	0.000	4.067	4.686
Lead Source_Olark Chat	0.8981	0.121	7.446	0.000	0.662	1.134
Lead Source_Referral Sites	-0.6758	0.334	-2.023	0.043	-1.331	-0.021
Lead Source_Welingak Website	3.1926	1.028	3.106	0.002	1.178	5.207
Lead Origin_Landing Page Submission	-0.9039	0.122	-7.433	0.000	-1.142	-0.666
Lead Origin_Lead Add Form	3.7703	0.219	17.246	0.000	3.342	4.199
Specialization_Not Mentioned	-0.9953	0.118	-8.457	0.000	-1.226	-0.765
What is your current occupation_Housewife	22.5584	1.43e+04	0.002	0.999	-2.8e+04	2.81e+04
What is your current occupation_Other	-1.2132	0.842	-1.441	0.149	-2.863	0.437
What is your current occupation_Student	-0.0182	0.223	-0.081	0.935	-0.456	0.420
What is your current occupation_Working Professional	2.6798	0.183	14.644	0.000	2.321	3.038

	Features	VIF
1	TotalVisits	3.18
6	Lead Origin_Landing Page Submission	2.95
8	Specialization_Not Mentioned	2.32
2	Total Time Spent on Website	2.10
3	Lead Source_Olark Chat	1.85
7	Lead Origin_Lead Add Form	1.43
5	Lead Source_Welingak Website	1.31
12	What is your current occupation_Working Profes...	1.18
0	Do Not Email	1.11
4	Lead Source_Referral Sites	1.05
11	What is your current occupation_Student	1.03
9	What is your current occupation_Housewife	1.01
10	What is your current occupation_Other	1.00

MODEL 4

Dep. Variable:	Converted	**No. Observations:**	6312
Model:	GLM	**Df Residuals:**	6299
Model Family:	Binomial	**Df Model:**	12
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2936.5
Date:	Sat, 10 Apr 2021	**Deviance:**	5873.1
Time:	23:54:22	**Pearson chi2:**	9.11e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2944	0.142	-9.107	0.000	-1.573	-1.016
Do Not Email	-1.4053	0.161	-8.711	0.000	-1.721	-1.089
TotalVisits	0.6897	0.187	3.681	0.000	0.323	1.057
Total Time Spent on Website	4.3697	0.158	27.697	0.000	4.061	4.679
Lead Source_Olark Chat	0.8942	0.121	7.418	0.000	0.658	1.130
Lead Source_Referral Sites	-0.6749	0.334	-2.021	0.043	-1.329	-0.020
Lead Source_Welingak Website	3.1850	1.028	3.099	0.002	1.171	5.200
Lead Origin_Landing Page Submission	-0.9016	0.122	-7.418	0.000	-1.140	-0.663
Lead Origin_Lead Add Form	3.7749	0.218	17.282	0.000	3.347	4.203
Specialization_Not Mentioned	-0.9988	0.118	-8.491	0.000	-1.229	-0.768
What is your current occupation_Other	-1.2148	0.842	-1.443	0.149	-2.865	0.435
What is your current occupation_Student	-0.0213	0.223	-0.095	0.924	-0.459	0.417
What is your current occupation_Working Professional	2.6753	0.183	14.622	0.000	2.317	3.034

	Features	VIF
1	TotalVisits	3.18
6	Lead Origin_Landing Page Submission	2.94
8	Specialization_Not Mentioned	2.32
2	Total Time Spent on Website	2.10
3	Lead Source_Olark Chat	1.85
7	Lead Origin_Lead Add Form	1.42
5	Lead Source_Welingak Website	1.31
11	What is your current occupation_Working Profes...	1.18
0	Do Not Email	1.11
4	Lead Source_Referral Sites	1.05
10	What is your current occupation_Student	1.03
9	What is your current occupation_Other	1.00

MODEL 5

Dep. Variable:	Converted	**No. Observations:**	6312
Model:	GLM	**Df Residuals:**	6300
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2936.5
Date:	Sat, 10 Apr 2021	**Deviance:**	5873.1
Time:	23:54:22	**Pearson chi2:**	9.11e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2951	0.142	-9.121	0.000	-1.573	-1.017
Do Not Email	-1.4052	0.161	-8.711	0.000	-1.721	-1.089
TotalVisits	0.6900	0.187	3.683	0.000	0.323	1.057
Total Time Spent on Website	4.3697	0.158	27.697	0.000	4.061	4.679
Lead Source_Olark Chat	0.8940	0.121	7.417	0.000	0.658	1.130
Lead Source_Referral Sites	-0.6746	0.334	-2.020	0.043	-1.329	-0.020
Lead Source_Welingak Website	3.1853	1.028	3.099	0.002	1.171	5.200
Lead Origin_Landing Page Submission	-0.9015	0.122	-7.418	0.000	-1.140	-0.663
Lead Origin_Lead Add Form	3.7748	0.218	17.281	0.000	3.347	4.203
Specialization_Not Mentioned	-0.9986	0.118	-8.491	0.000	-1.229	-0.768
What is your current occupation_Other	-1.2143	0.842	-1.442	0.149	-2.864	0.436
What is your current occupation_Working Professional	2.6758	0.183	14.631	0.000	2.317	3.034

	Features	VIF
1	TotalVisits	3.18
6	Lead Origin_Landing Page Submission	2.93
8	Specialization_Not Mentioned	2.32
2	Total Time Spent on Website	2.10
3	Lead Source_Olark Chat	1.85
7	Lead Origin_Lead Add Form	1.41
5	Lead Source_Welingak Website	1.31
10	What is your current occupation_Working Profes...	1.18
0	Do Not Email	1.11
4	Lead Source_Referral Sites	1.05
9	What is your current occupation_Other	1.00

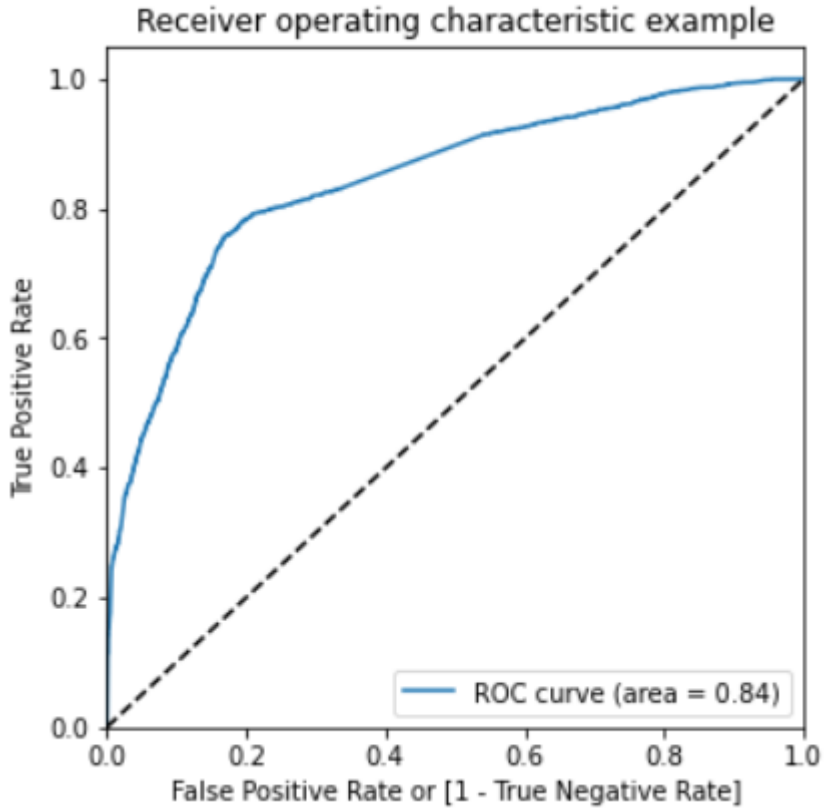
MODEL 6

Dep. Variable:	Converted	**No. Observations:**	6312
Model:	GLM	**Df Residuals:**	6301
Model Family:	Binomial	**Df Model:**	10
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2937.7
Date:	Sat, 10 Apr 2021	**Deviance:**	5875.3
Time:	23:54:22	**Pearson chi2:**	9.08e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2979	0.142	-9.142	0.000	-1.576	-1.020
Do Not Email	-1.4099	0.161	-8.738	0.000	-1.726	-1.094
TotalVisits	0.6813	0.187	3.638	0.000	0.314	1.048
Total Time Spent on Website	4.3637	0.158	27.685	0.000	4.055	4.673
Lead Source_Olark Chat	0.8928	0.120	7.410	0.000	0.657	1.129
Lead Source_Referral Sites	-0.6708	0.334	-2.009	0.045	-1.325	-0.016
Lead Source_Welingak Website	3.1876	1.028	3.101	0.002	1.173	5.202
Lead Origin_Landing Page Submission	-0.8958	0.121	-7.373	0.000	-1.134	-0.658
Lead Origin_Lead Add Form	3.7717	0.218	17.264	0.000	3.343	4.200
Specialization_Not Mentioned	-0.9932	0.118	-8.450	0.000	-1.224	-0.763
What is your current occupation_Working Professional	2.6781	0.183	14.648	0.000	2.320	3.036

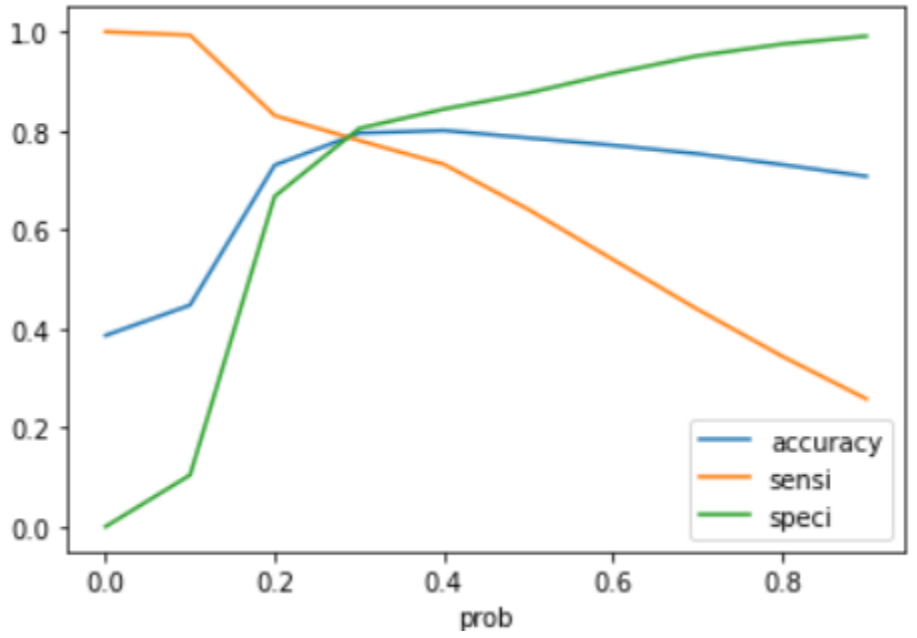
	Features	VIF
1	TotalVisits	3.18
6	Lead Origin_Landing Page Submission	2.93
8	Specialization_Not Mentioned	2.31
2	Total Time Spent on Website	2.10
3	Lead Source_Olark Chat	1.85
7	Lead Origin_Lead Add Form	1.41
5	Lead Source_Welingak Website	1.31
9	What is your current occupation_Working Profes...	1.18
0	Do Not Email	1.11
4	Lead Source_Referral Sites	1.05

ROC CURVE



Above 45 degree line. So it is a good curve.

ACCURACY SENSITIVITY SPECIFICITY CURVE



MODEL EVALUATION

- From the sensitivitiy, specificity and accuracy curve we took 0.3 as the optimum point

Confusion Matrix(Converted=1)	Not converted	Converted
Not converted	3117	757
Converted	536	1902

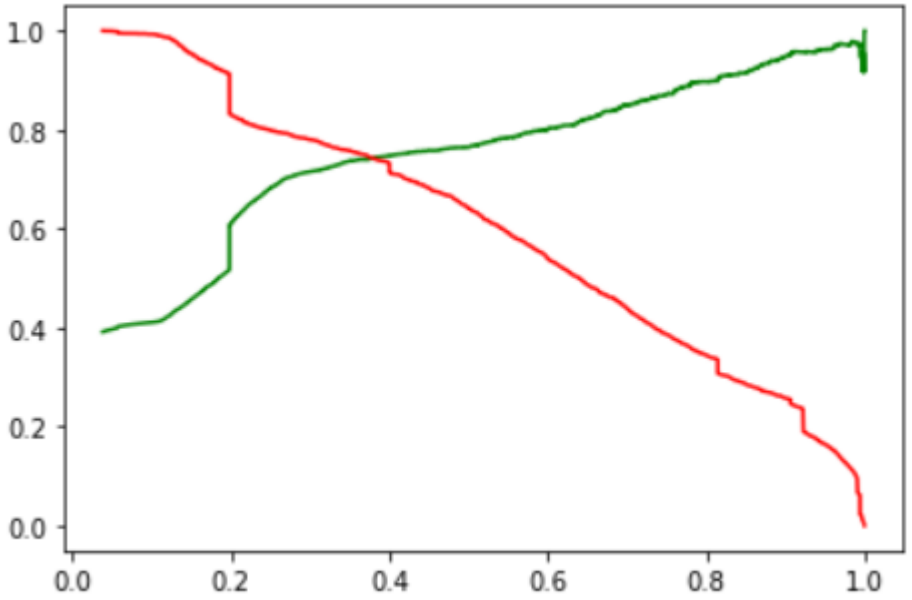
- Accuracy = 79.5%
- Sensitivity = 78%
- Specificity = 80%
- False Positive rate = 19%
- Positive predictive value = 71%
- Negative predictive value = 85%

PRECISION-RECALL

CM	Not converted	converted
Not converted	3394	480
converted	875	1563

PRECISION –RECALL CURVE

- Precision Score = 76%
- Recall score = 64%



TEST SET EVALUATION

- Accuracy = 79.6%
- Sensitivity = 78.8%
- Specificity = 80%
- Precision = 70%
- Recall = 78%

CONCLUSION

The company should focus more on:

1. People who are visiting their websites often and spending time
2. Lead Source is Olark Chat, Wellingak website and referral sites.
3. Lead origin is from Landing Page submission, Lead add form
4. Working professionals should be focused more.

Following these steps the company can get maximum conversion output.