LEAD SCORING ASSIGNMENT

Aditya Prasad
Naveen Vishwakarma
Priyanka Kumari

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.

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%.

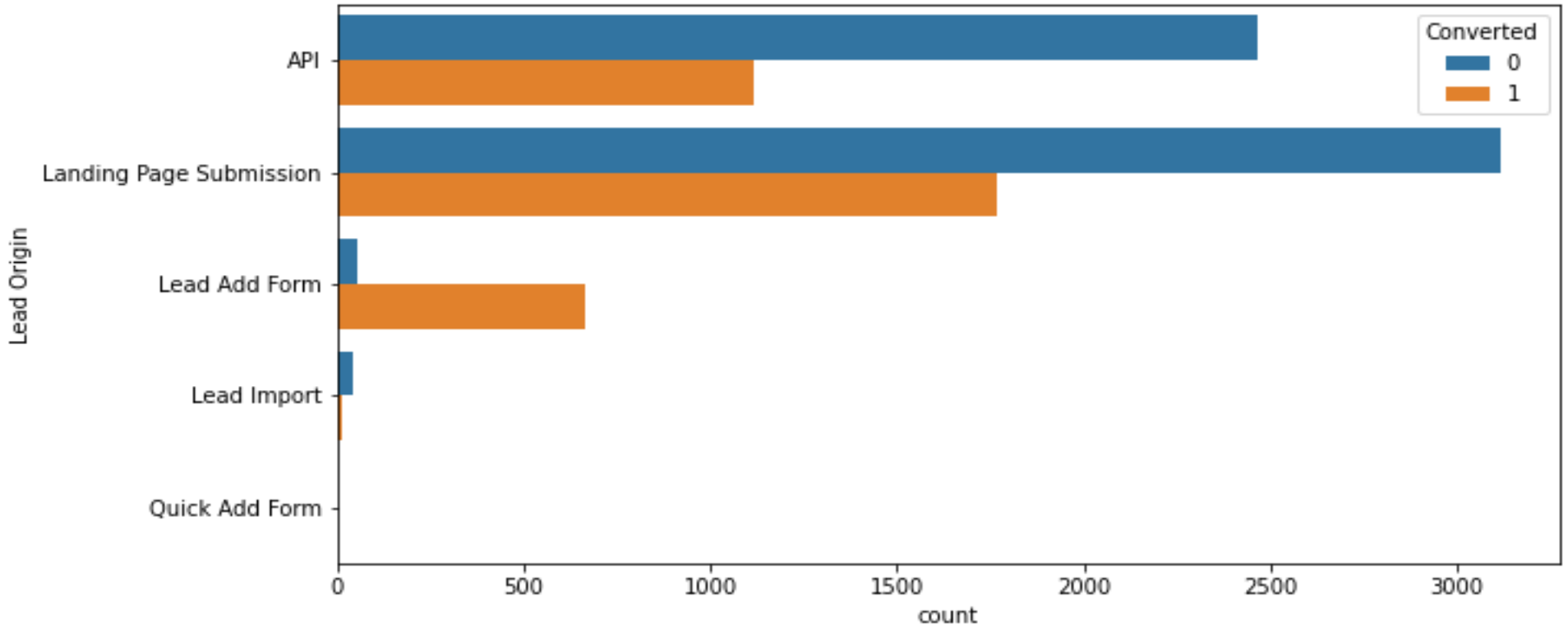
Although X Education gets a lot of leads, its lead conversion rate is very poor.

GOALS TO BE ACHIEVED

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.

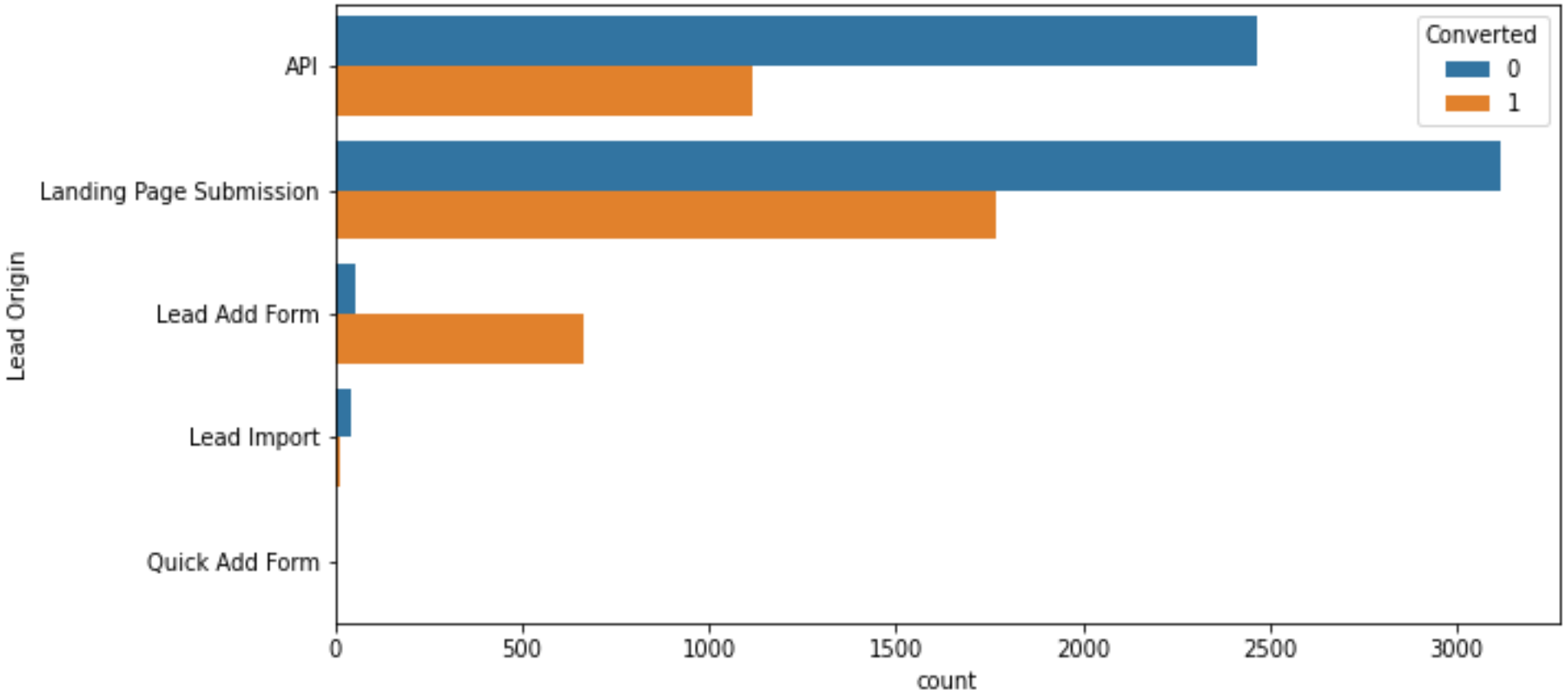
A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

EXPLORATORY DATA ANALYSIS



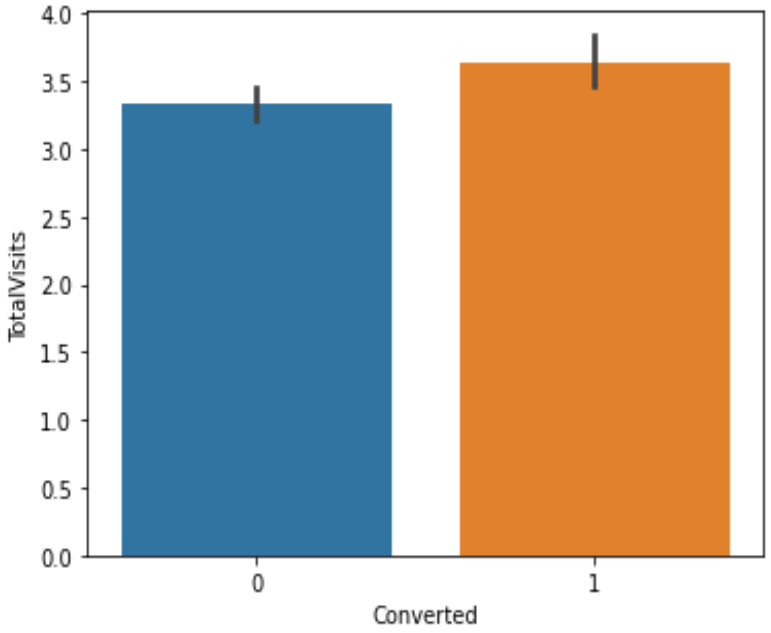
- From Lead Origin finding, maximum lead conversion happened from Landing Page Submission.

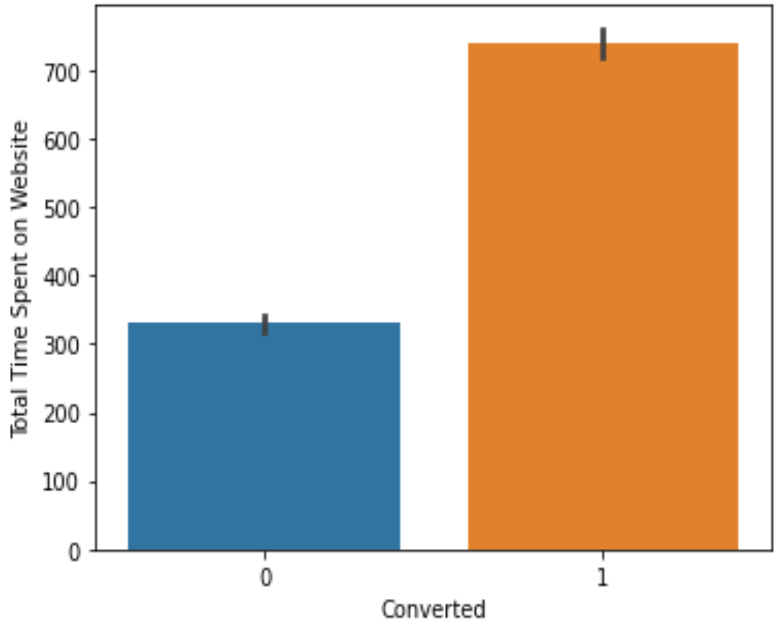
EDA

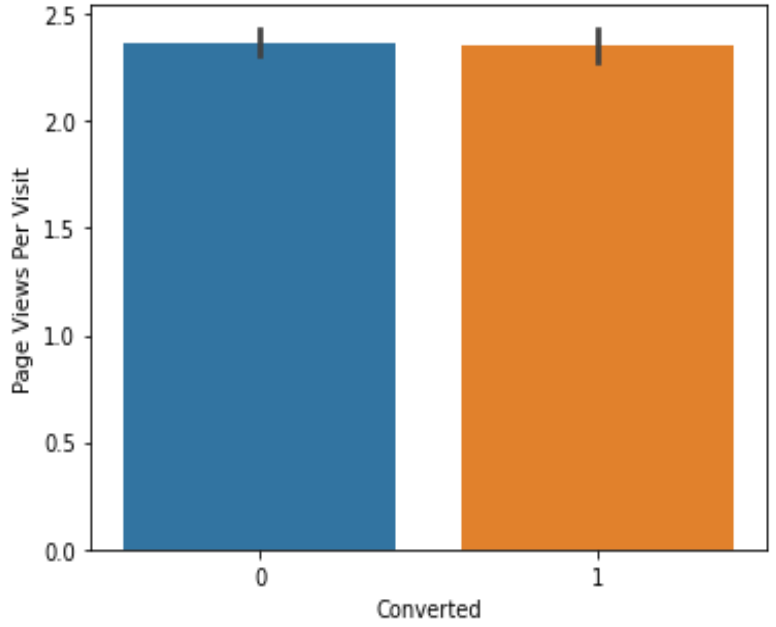


- From Lead Origin finding, maximum lead conversion happened from Landing Page Submission.

EDA

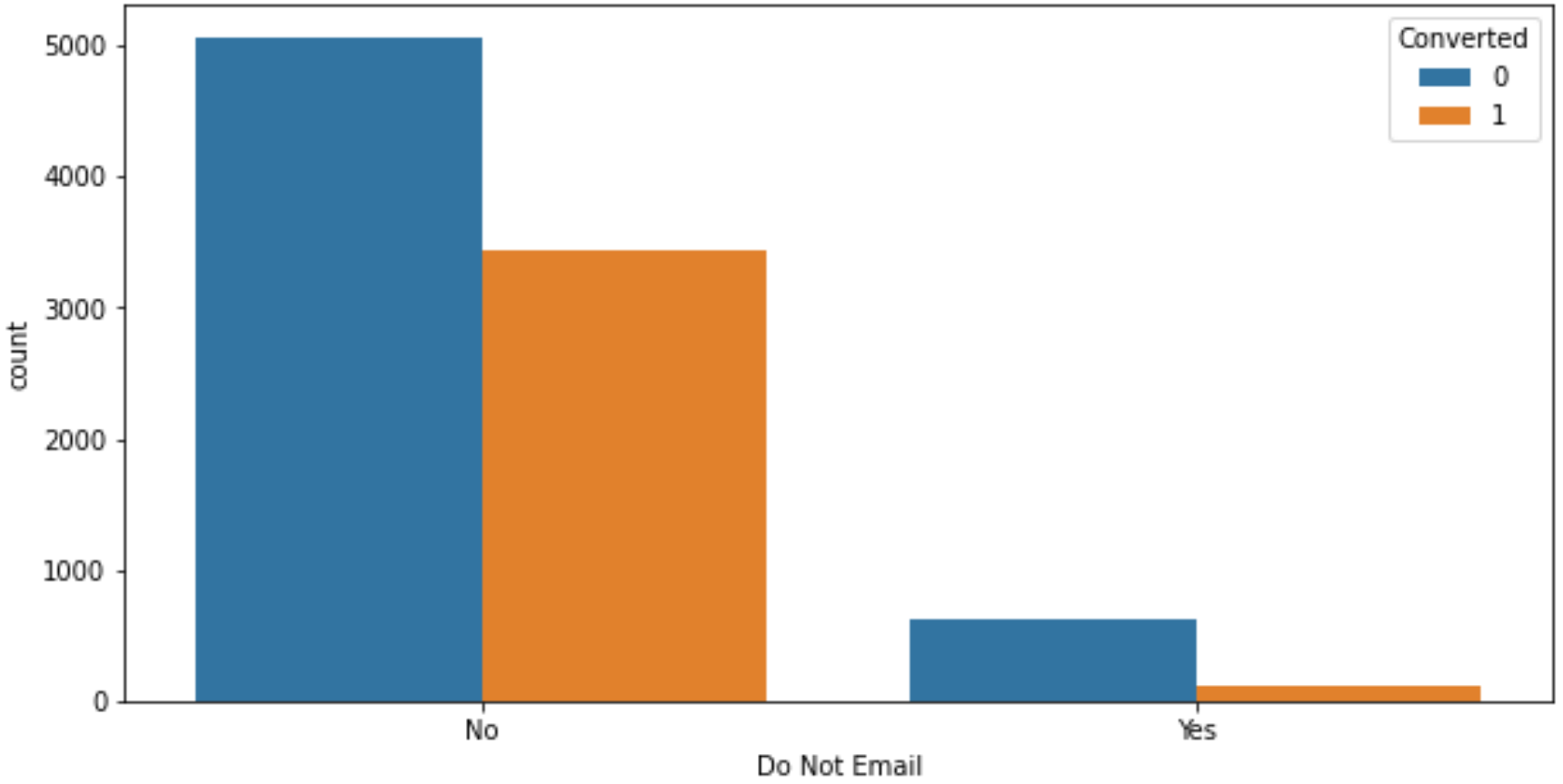






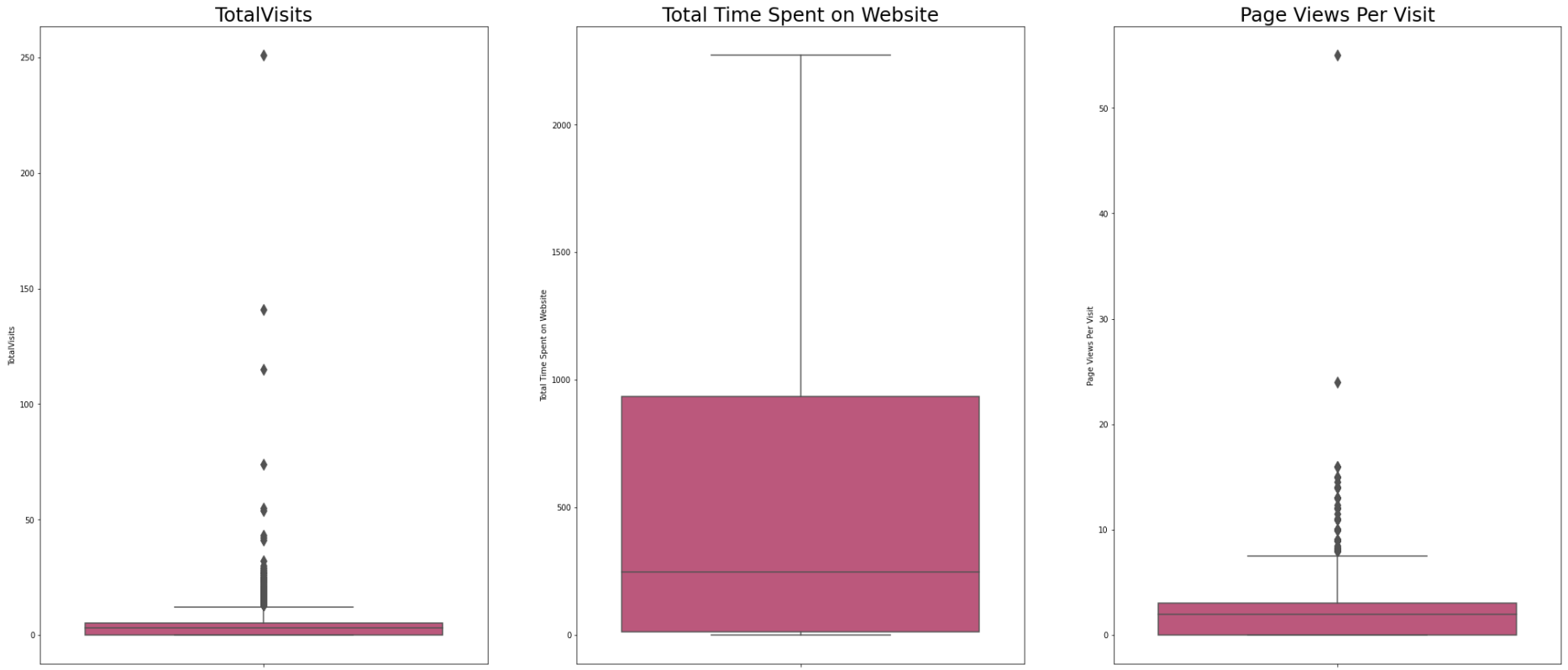
- From the above graph, we have major lead conversion from Total Visits, Total Time Spent on Website, Page Views Per Visit.

EDA



- Based on the above graph , major lead conversion has happened from email that have been sent.

BOXPLOT

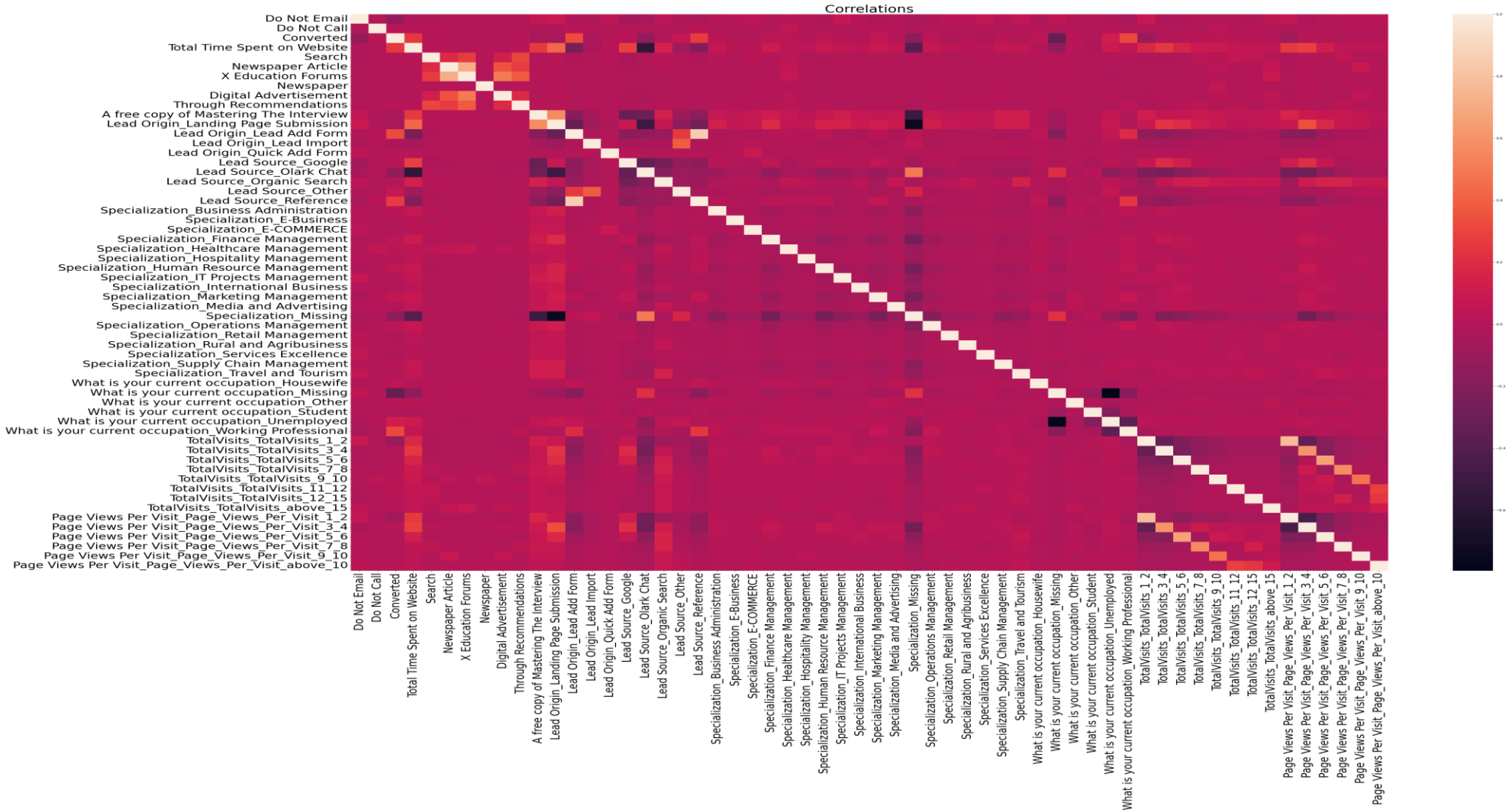


- From the above boxplots we can now confirm that we have two outlier variables in our dataset ('Total Visits' and 'Page Views Per Visit').

CREATING BINS

- After creating bins we removed the outliers and are now good to go. Before creating the dummy variables let's remove redundant columns/ variables.
- Also from above we know columns : 'Last Activity', 'Tags', 'Last Notable Activity' activity columns came from sales team, thus we will drop these redundant columns.

HEATMAP


Correlations
Do Not Email
Do Not Call
Converted
Total Time Spent on Website
Search
Newspaper Article
X Education Forums
Newspaper
Digital Advertisement
Through Recommendations
A free copy of Mastering The Interview
Lead Origin_Landing Page Submission
Lead Origin_Lead Add Form
Lead Origin_Lead Import
Lead Origin_Quick Add Form
Lead Source_Google
Lead Source_Olark Chat
Lead Source_Organic Search
Lead Source_Other
Lead Source_Reference
Specialization_Business Administration
Specialization_E-Business
Specialization_E-COMMERCE
Specialization_Finance Management
Specialization_Healthcare Management
Specialization_Hospitality Management
Specialization_Human Resource Management
Specialization_IT Projects Management
Specialization_International Business
Specialization_Marketing Management
Specialization_Media and Advertising
Specialization_Missing
Specialization_Operations Management
Specialization_Retail Management
Specialization_Rural and Agribusiness
Specialization_Services Excellence
Specialization_Supply Chain Management
Specialization_Travel and Tourism
What is your current occupation_Housewife
What is your current occupation_Missing
What is your current occupation_Other
What is your current occupation_Student
What is your current occupation_Unemployed
What is your current occupation_Working Professional
TotalVisits_TotalVisits_1_2
TotalVisits_TotalVisits_3_4
TotalVisits_TotalVisits_5_6
TotalVisits_TotalVisits_7_8
TotalVisits_TotalVisits_9_10
TotalVisits_TotalVisits_11_12
TotalVisits_TotalVisits_12_15
TotalVisits_TotalVisits_above_15
Page Views Per Visit_Page_Views_Per_Visit_1_2
Page Views Per Visit_Page_Views_Per_Visit_3_4
Page Views Per Visit_Page_Views_Per_Visit_5_6
Page Views Per Visit_Page_Views_Per_Visit_7_8
Page Views Per Visit_Page_Views_Per_Visit_9_10
Page Views Per Visit_Page_Views_Per_Visit_above_10

BUILDING MODEL

- Importing RFE and logistic regression libraries from scikit learn.
- Creating RFE model with 15 variables.
- Fitting the model
- Listing which all columns are selected (True) by RFE & which all are rejected.
- Storing selected columns by RFE in a list.
- Listing features removed by RFE feature selection.
- Creating new train dataframe with RFE selected features.

MODEL 1

- Summary of features for model 1

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.1330	0.136	8.345	0.000	0.867	1.399
Do Not Email	-1.2865	0.160	-8.055	0.000	-1.600	-0.973
Total Time Spent on Website	1.1141	0.039	28.510	0.000	1.037	1.191
Lead Origin_Landing Page Submission	-0.9087	0.119	-7.607	0.000	-1.143	-0.675
Lead Origin_Lead Add Form	2.5615	0.188	13.607	0.000	2.193	2.930
Lead Origin_Lead Import	-1.5579	0.538	-2.897	0.004	-2.612	-0.504
Specialization_Hospitality Management	-1.0284	0.327	-3.141	0.002	-1.670	-0.387
Specialization_Missing	-0.9752	0.119	-8.164	0.000	-1.209	-0.741
What is your current occupation_Housewife	22.3639	1.32e+04	0.002	0.999	-2.58e+04	2.58e+04
What is your current occupation_Missing	-1.1869	0.083	-14.245	0.000	-1.350	-1.024
What is your current occupation_Working Professional	2.4044	0.185	12.982	0.000	2.041	2.767
Page Views Per Visit_Page_Views_Per_Visit_1_2	-1.1833	0.118	-10.005	0.000	-1.415	-0.952
Page Views Per Visit_Page_Views_Per_Visit_3_4	-0.9541	0.127	-7.531	0.000	-1.202	-0.706
Page Views Per Visit_Page_Views_Per_Visit_5_6	-0.8874	0.154	-5.766	0.000	-1.189	-0.586
Page Views Per Visit_Page_Views_Per_Visit_7_8	-0.8564	0.234	-3.655	0.000	-1.316	-0.397
Page Views Per Visit_Page_Views_Per_Visit_9_10	-1.0598	0.373	-2.842	0.004	-1.791	-0.329

- We are dropping 'const',’ What is your current occupation_Housewife' due to high p-value

MODEL 2

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.1352	0.136	8.362	0.000	0.869	1.401
Do Not Email	-1.2910	0.160	-8.083	0.000	-1.604	-0.978
Total Time Spent on Website	1.1129	0.039	28.509	0.000	1.036	1.189
Lead Origin_Landing Page Submission	-0.9055	0.119	-7.585	0.000	-1.139	-0.672
Lead Origin_Lead Add Form	2.5701	0.188	13.665	0.000	2.201	2.939
Lead Origin_Lead Import	-1.5575	0.538	-2.897	0.004	-2.611	-0.504
Specialization_Hospitality Management	-1.0328	0.327	-3.155	0.002	-1.674	-0.391
Specialization_Missing	-0.9786	0.119	-8.196	0.000	-1.213	-0.745
What is your current occupation_Missing	-1.1897	0.083	-14.283	0.000	-1.353	-1.026
What is your current occupation_Working Professional	2.3987	0.185	12.953	0.000	2.036	2.762
Page Views Per Visit_Page_Views_Per_Visit_1_2	-1.1797	0.118	-9.984	0.000	-1.411	-0.948
Page Views Per Visit_Page_Views_Per_Visit_3_4	-0.9525	0.127	-7.523	0.000	-1.201	-0.704
Page Views Per Visit_Page_Views_Per_Visit_5_6	-0.8905	0.154	-5.787	0.000	-1.192	-0.589
Page Views Per Visit_Page_Views_Per_Visit_7_8	-0.8599	0.234	-3.671	0.000	-1.319	-0.401
Page Views Per Visit_Page_Views_Per_Visit_9_10	-1.0632	0.373	-2.852	0.004	-1.794	-0.332

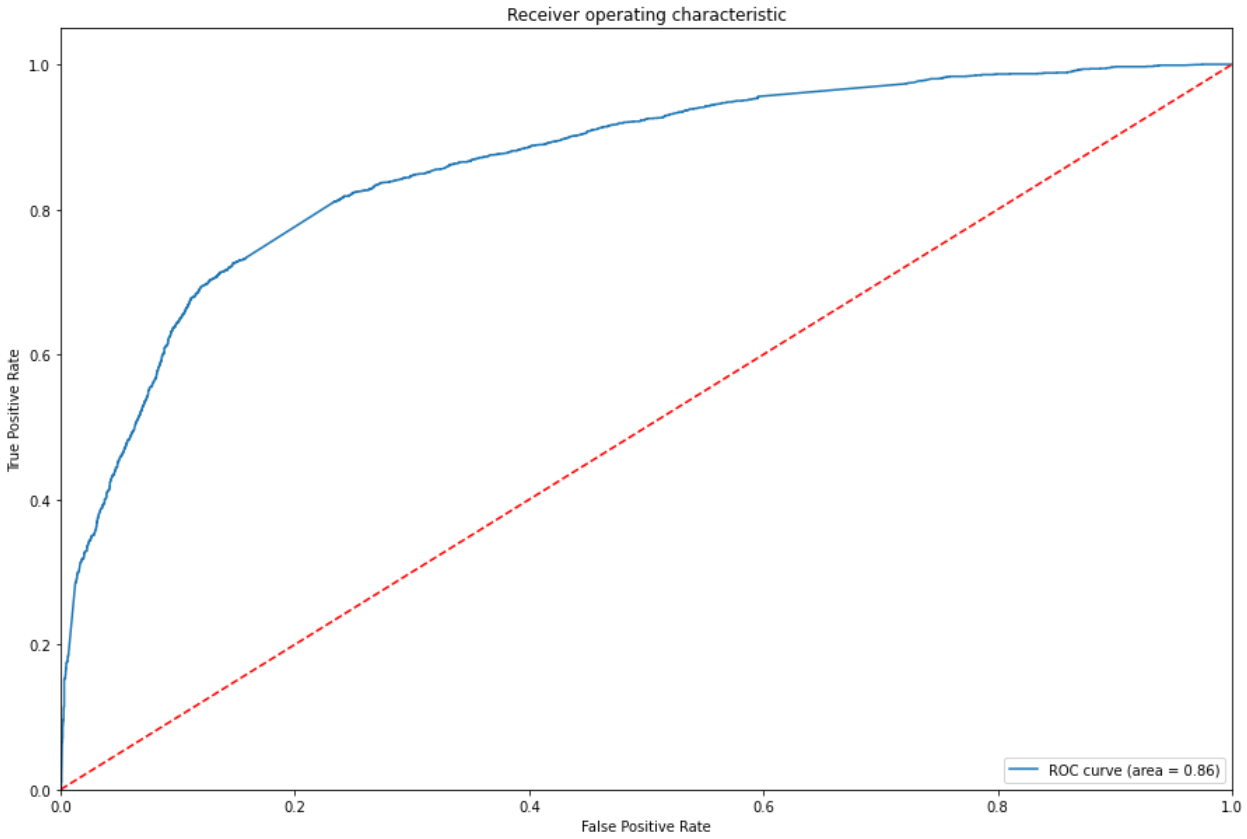
MODEL 3

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.3026	0.087	3.468	0.001	0.132	0.474
Do Not Email	-1.3137	0.160	-8.229	0.000	-1.627	-1.001
Total Time Spent on Website	1.1258	0.039	28.870	0.000	1.049	1.202
Lead Origin_Landing Page Submission	-0.2207	0.084	-2.642	0.008	-0.384	-0.057
Lead Origin_Lead Add Form	2.7839	0.184	15.102	0.000	2.423	3.145
Lead Origin_Lead Import	-1.3740	0.527	-2.605	0.009	-2.408	-0.340
Specialization_Hospitality Management	-0.8809	0.319	-2.761	0.006	-1.506	-0.256
What is your current occupation_Missing	-1.2688	0.083	-15.371	0.000	-1.431	-1.107
What is your current occupation_Working Professional	2.5462	0.181	14.044	0.000	2.191	2.901
Page Views Per Visit_Page_Views_Per_Visit_1_2	-1.0712	0.115	-9.300	0.000	-1.297	-0.845
Page Views Per Visit_Page_Views_Per_Visit_3_4	-0.8097	0.123	-6.577	0.000	-1.051	-0.568
Page Views Per Visit_Page_Views_Per_Visit_5_6	-0.7444	0.151	-4.923	0.000	-1.041	-0.448
Page Views Per Visit_Page_Views_Per_Visit_7_8	-0.6927	0.232	-2.990	0.003	-1.147	-0.239
Page Views Per Visit_Page_Views_Per_Visit_9_10	-0.9893	0.371	-2.665	0.008	-1.717	-0.262

FINAL MODEL

	Converted	Converted_probability	ID
1871	0	0.333107	1871
6795	0	0.272324	6795
3516	0	0.216616	3516
8105	0	0.705671	8105
3934	0	0.333107	3934

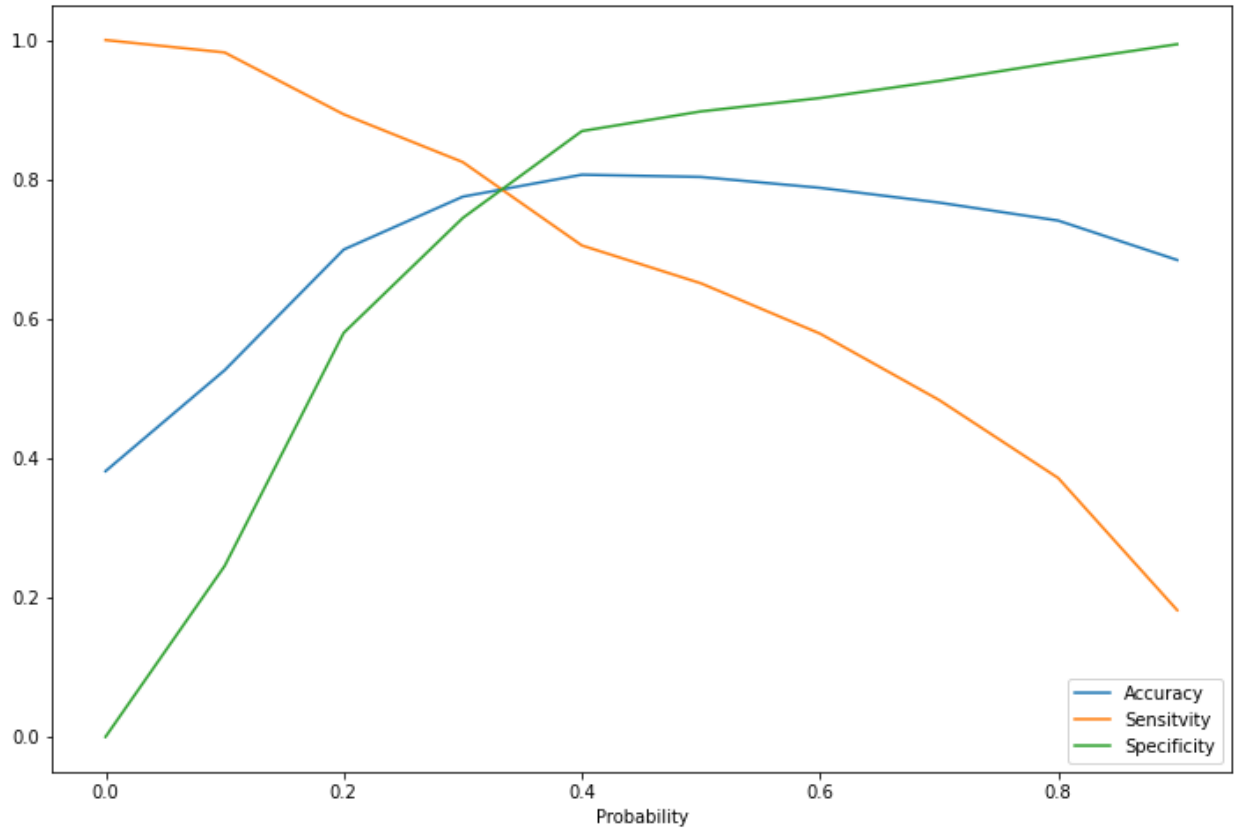
ROC


Receiver operating characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.86)

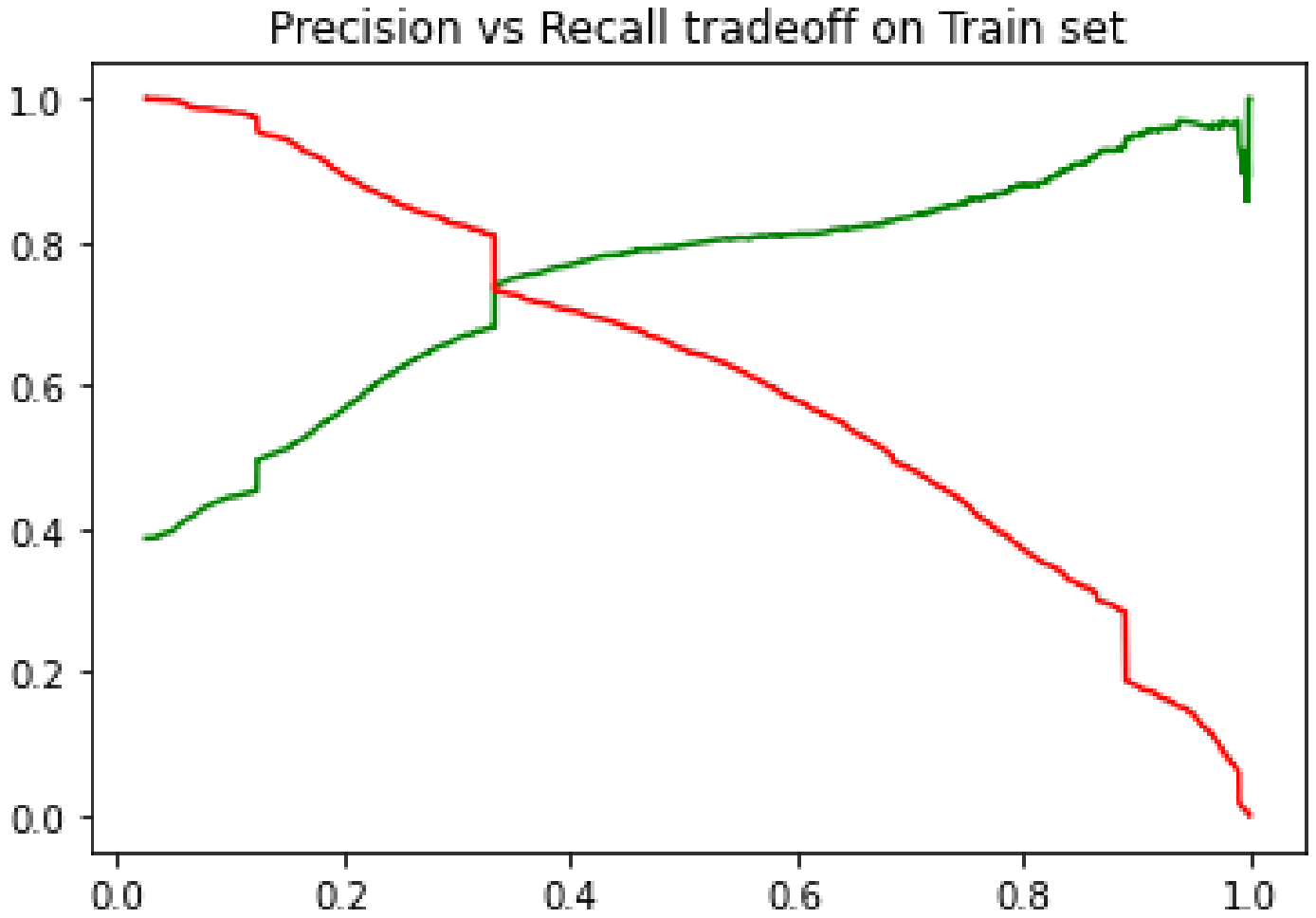
ACCURACY, SENSITIVITY, SPECIFICITY

	Probability	Accuracy	Sensitvity	Specificity
0.0	0.0	0.381262	1.000000	0.000000
0.1	0.1	0.526129	0.982157	0.245127
0.2	0.2	0.699289	0.893350	0.579710
0.3	0.3	0.775201	0.824818	0.744628
0.4	0.4	0.806741	0.705191	0.869315
0.5	0.5	0.803494	0.650852	0.897551
0.6	0.6	0.787879	0.578670	0.916792
0.7	0.7	0.766698	0.483374	0.941279
0.8	0.8	0.740878	0.371452	0.968516
0.9	0.9	0.684292	0.181671	0.994003

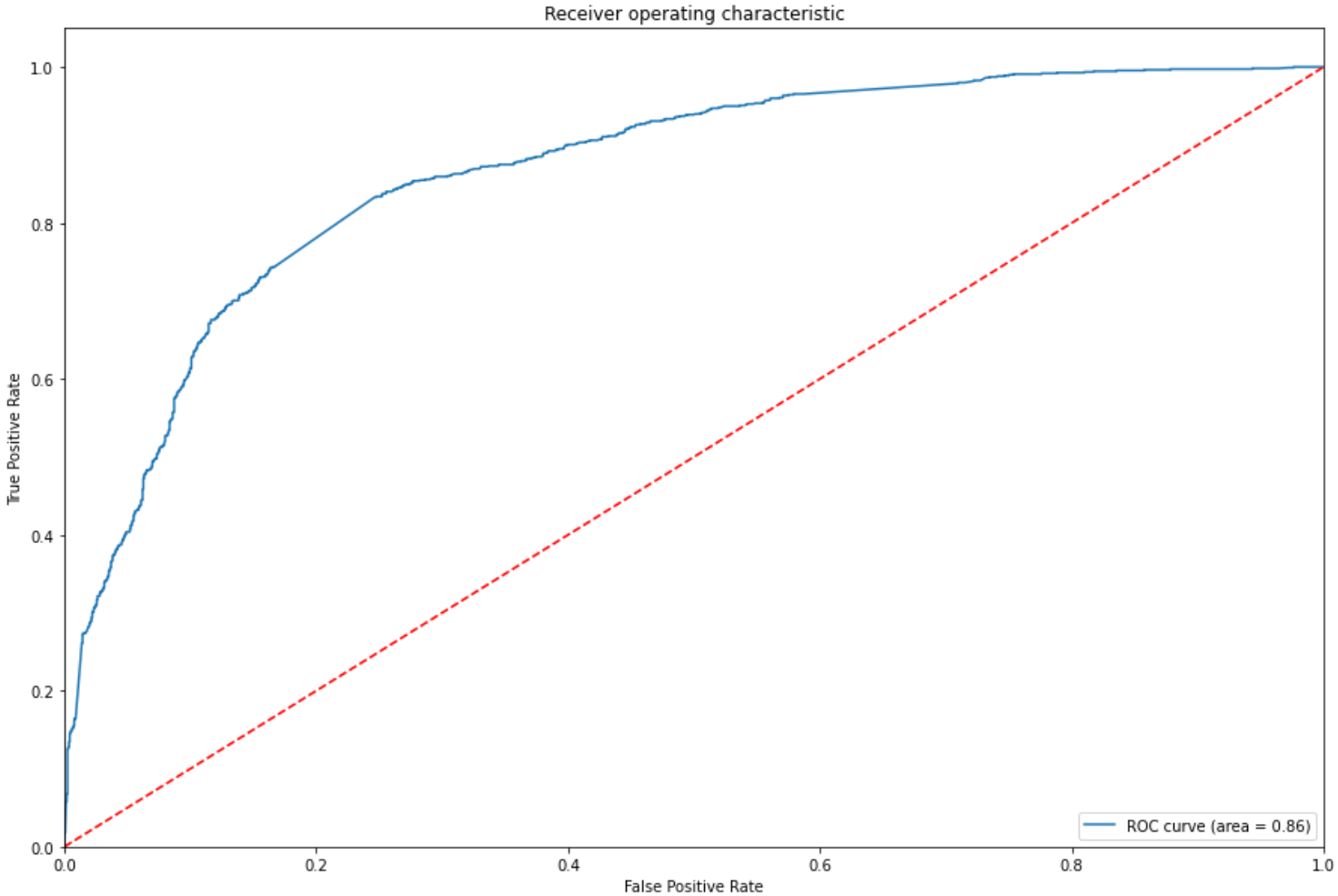
GRAPH


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
Probability
Accuracy
Sensitvity
Specificity

PRECISION VS RECALL TRADEFALL



ROC CURVE


Receiver operating characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.86)

CONCLUSION

- The sensitivity and specificity, Accuracy , Precision and Recall score we got from test set are almost accurate.
- We have high recall score than precision score which is a sign of good model.
- In business terms this model has an ability to adjust with the company's requirement in coming future.
- Important features responsible for good conversion rate or the ones which contributes more towards the probability of a lead getting converted are:
- Lead Origin_Lead Add Form
- Total Time Spent on Website
- What is your current occupation_working professional.