Lead Scoring Case Study using logistic regression

SUBMITTED BY – RAJAN RAM

RAMAN GUPTA

18 JULY 2023

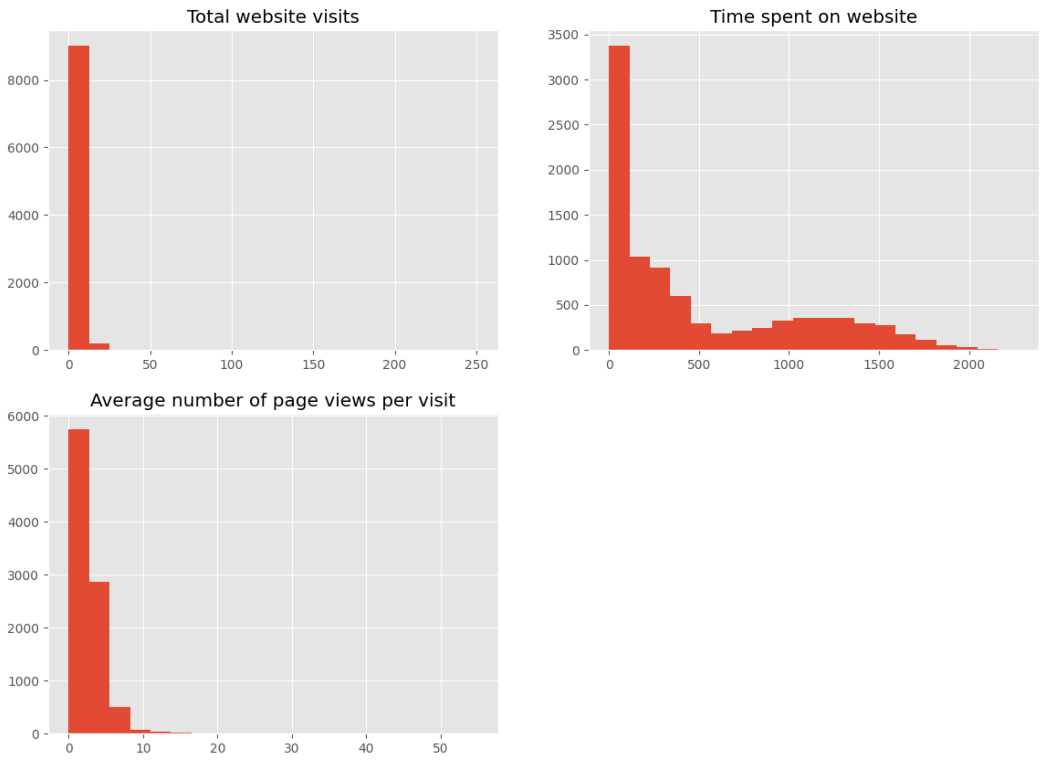
Problem Statement

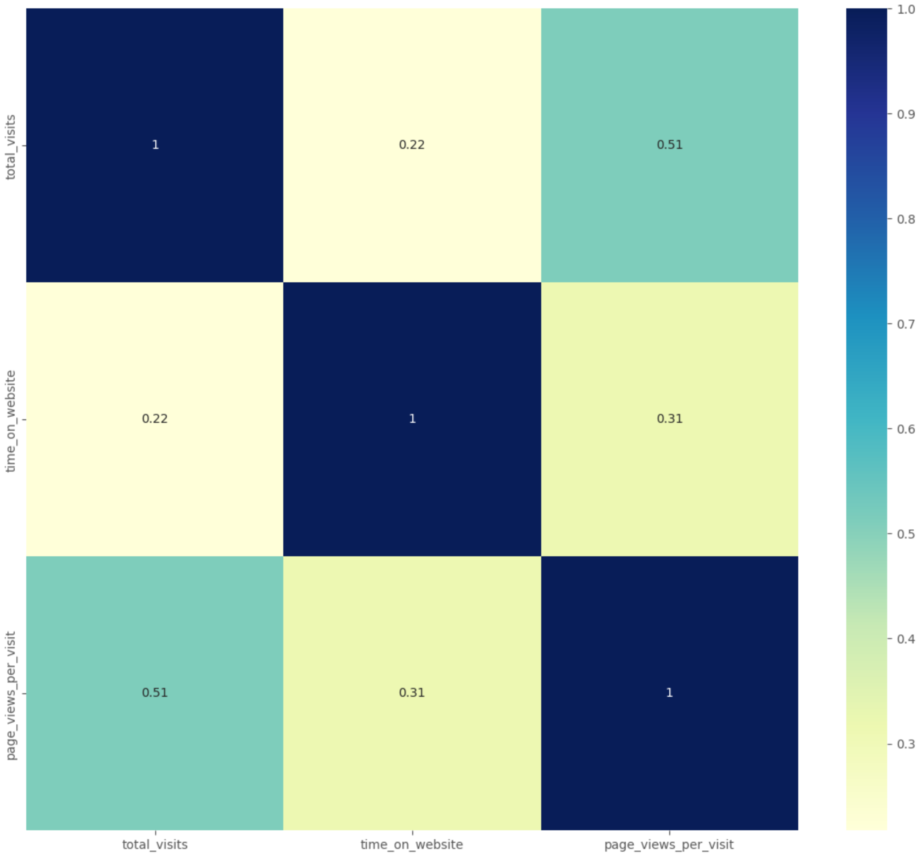
- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- To Build a logistic regression model for X education to assign a lead score between 0 and 100 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.
- X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.
- at times X education reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage

Problem Approach

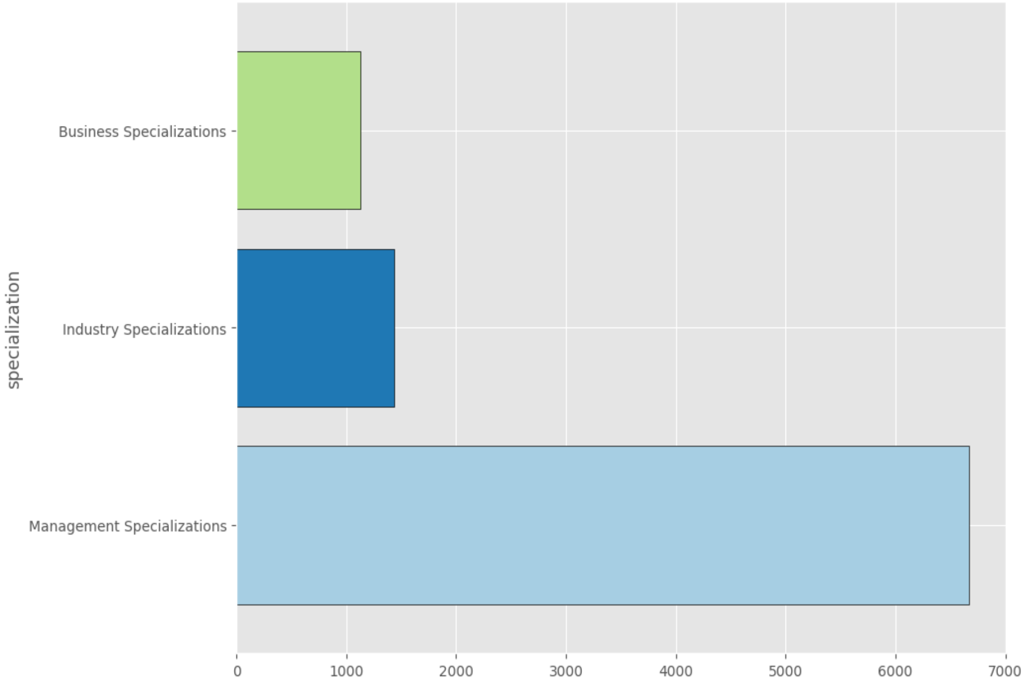
1. Importing necessary libraries.
2. Importing and reading the DATA.
3. DATA cleaning.
4. EDA.
5. DATA Preparation(Binary column, dummy variable etc.)
6. TEST-TRAIN Split.
7. Feature scaling.
8. Correlation.
9. MODEL Building.
10. Feature selection using RFE.
11. Assessing the model with stats model.
12. Manually dropping feature columns on basis of p-value and VIF.
13. calculating overall accuracy.
14. Making predictions on the test set.

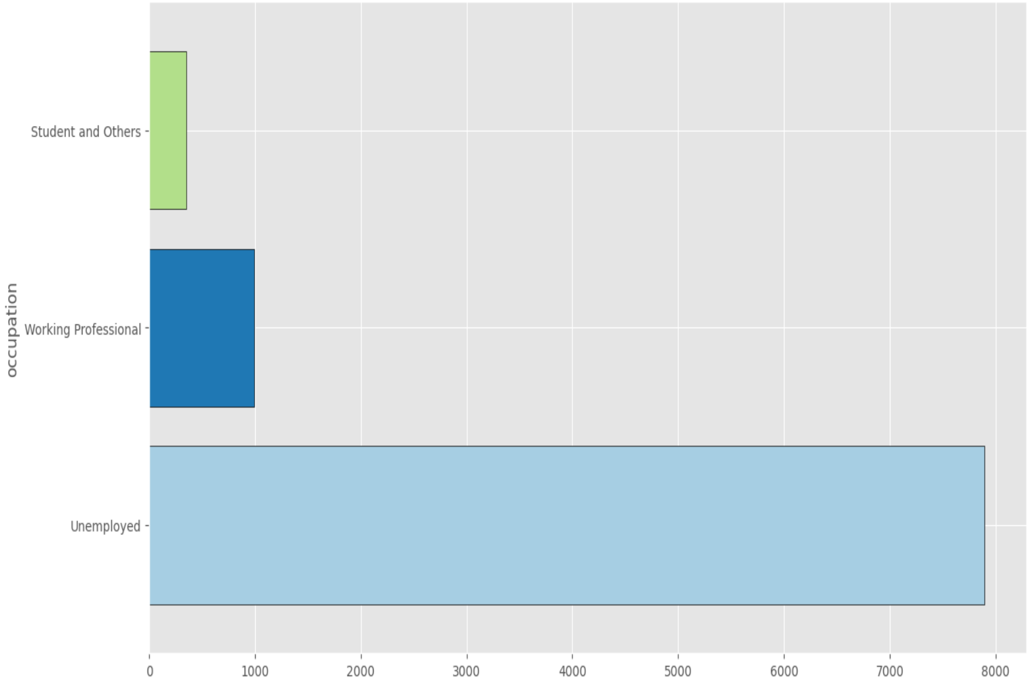
EDA (numerical columns)



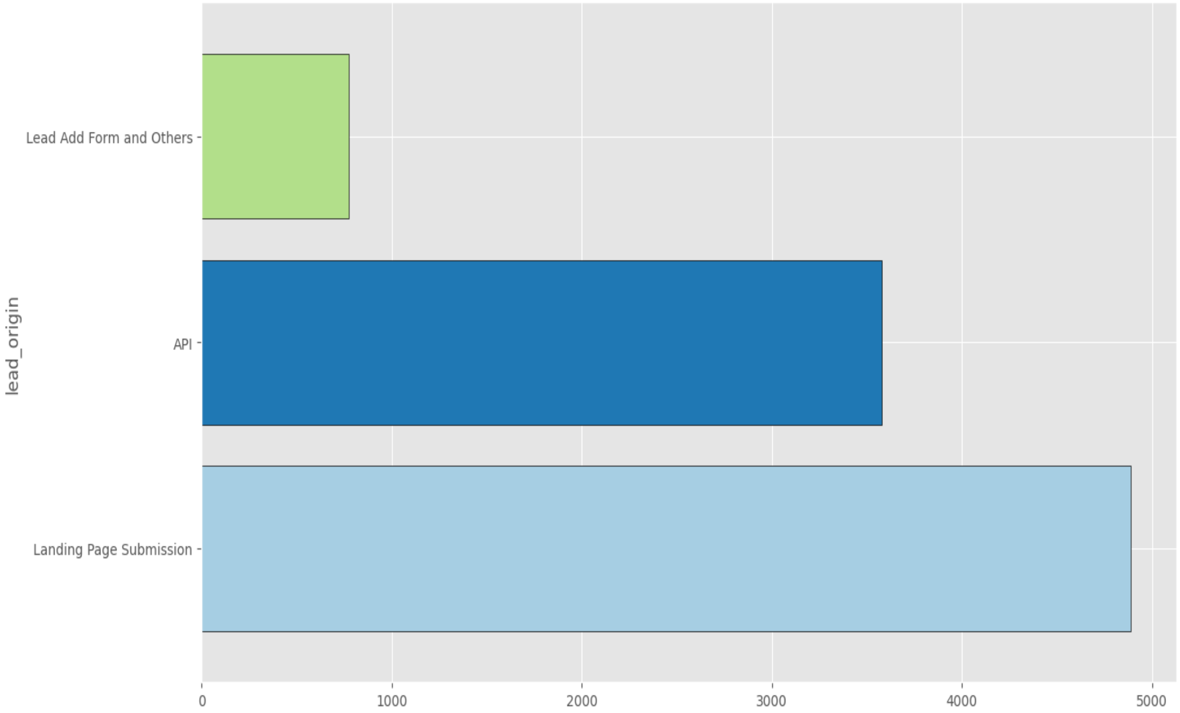


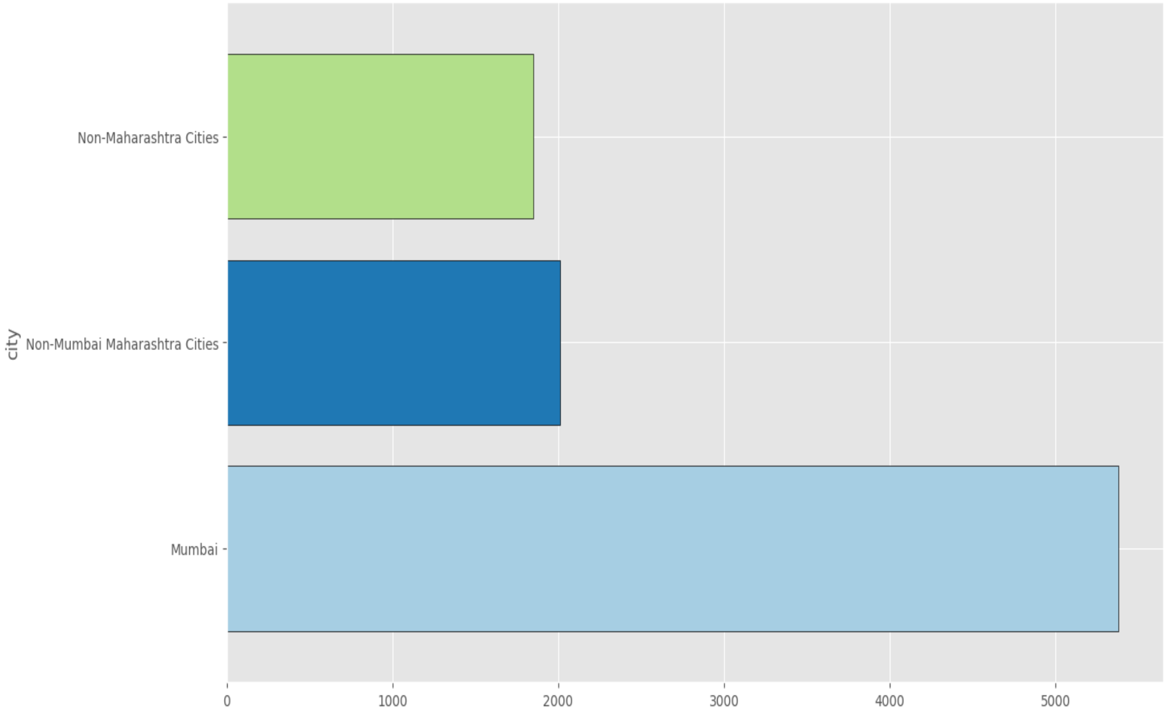
EDA (categorical columns)



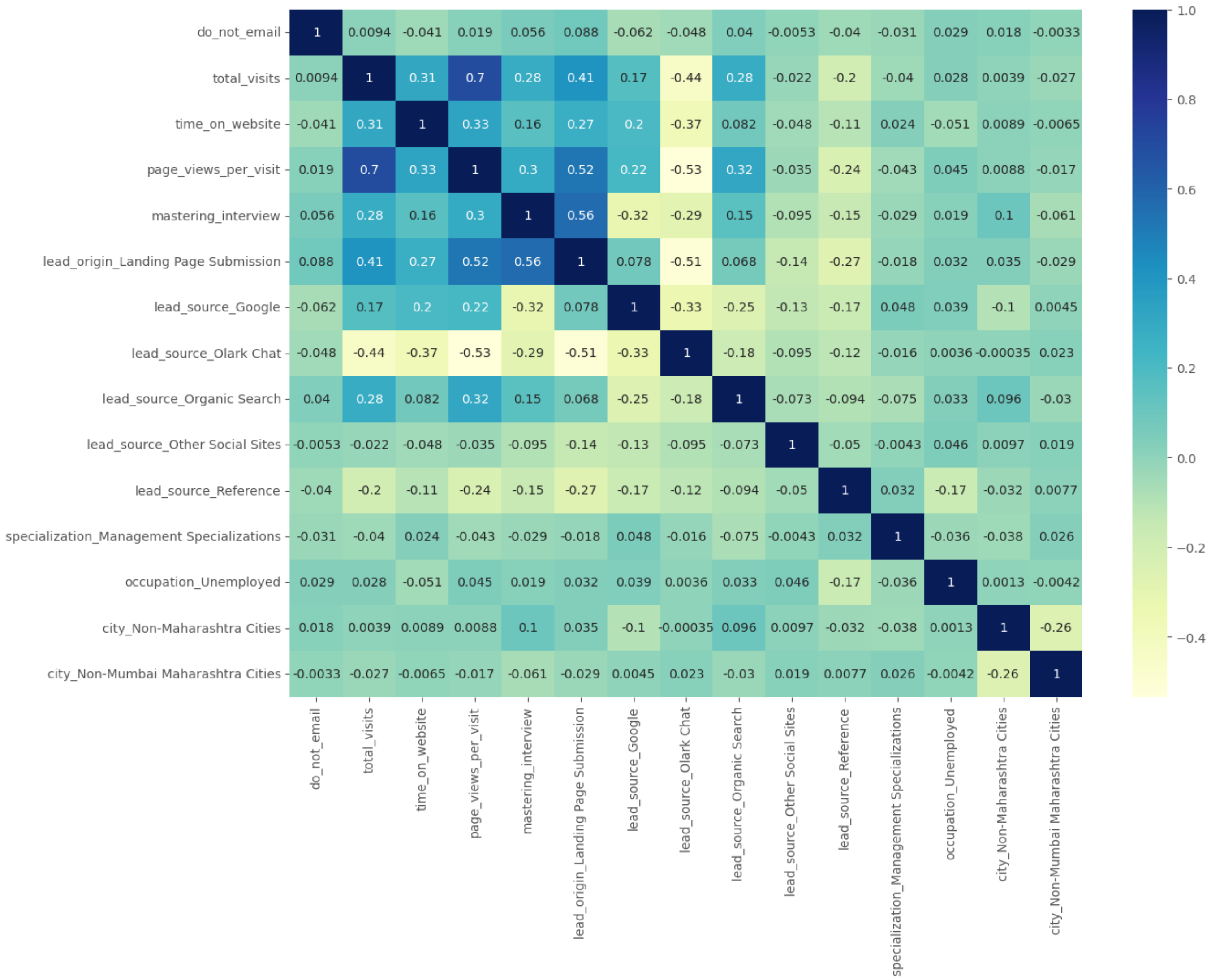


EDA (categorical columns)



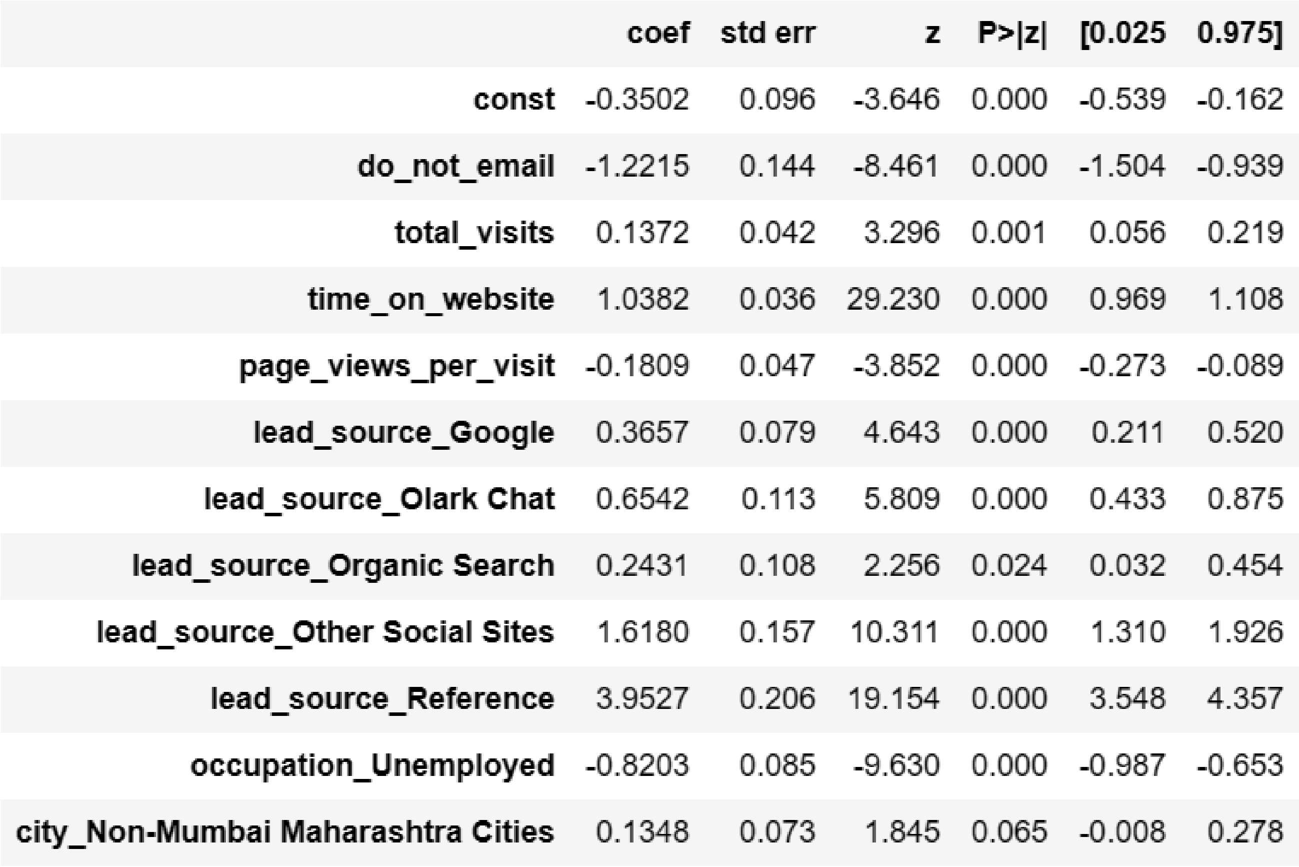


Correlation



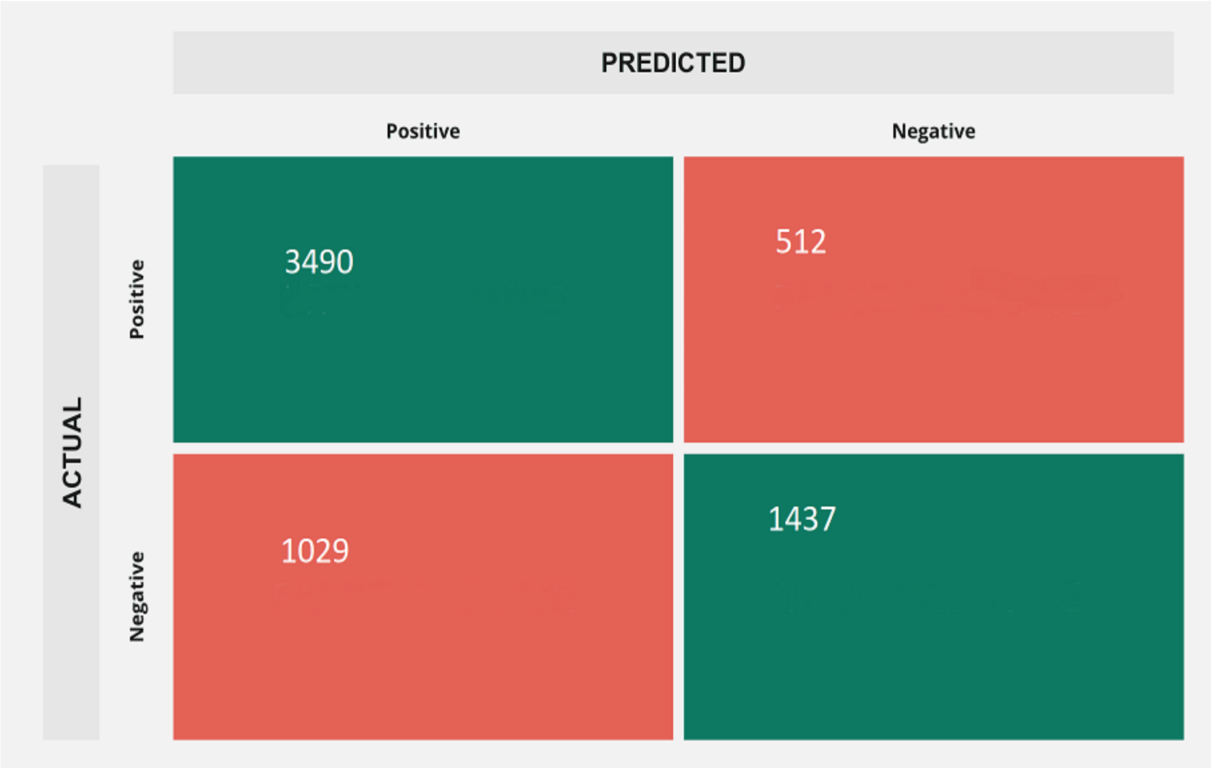
	do_not_email	total_visits	time_on_website	page_views_per_visit	mastering_interview	lead_origin_Landing Page Submission	lead_source_Google	lead_source_Olark Chat	lead_source_Organic Search	lead_source_Other Social Sites	lead_source_Reference	specialization_Management Specializations	occupation_Unemployed	city_Non-Maharashtra Cities	city_Non-Mumbai Maharashtra Cities
do_not_email	1	0.0094	-0.041	0.019	0.056	0.088	-0.062	-0.048	0.04	-0.0053	-0.04	-0.031	0.029	0.018	-0.0033
total_visits	0.0094	1	0.31	0.7	0.28	0.41	0.17	-0.44	0.28	-0.022	-0.2	-0.04	0.028	0.0039	-0.027
time_on_website	-0.041	0.31	1	0.33	0.16	0.27	0.2	-0.37	0.082	-0.048	-0.11	0.024	-0.051	0.0089	-0.0065
page_views_per_visit	0.019	0.7	0.33	1	0.3	0.52	0.22	-0.53	0.32	-0.035	-0.24	-0.043	0.045	0.0088	-0.017
mastering_interview	0.056	0.28	0.16	0.3	1	0.56	-0.32	-0.29	0.15	-0.095	-0.15	-0.029	0.019	0.1	-0.061
lead_origin_Landing Page Submission	0.088	0.41	0.27	0.52	0.56	1	0.078	-0.51	0.068	-0.14	-0.27	-0.018	0.032	0.035	-0.029
lead_source_Google	-0.062	0.17	0.2	0.22	-0.32	0.078	1	-0.33	-0.25	-0.13	-0.17	0.048	0.039	-0.1	0.0045
lead_source_Olark Chat	-0.048	-0.44	-0.37	-0.53	-0.29	-0.51	-0.33	1	-0.18	-0.095	-0.12	-0.016	0.0036	-0.00035	0.023
lead_source_Organic Search	0.04	0.28	0.082	0.32	0.15	0.068	-0.25	-0.18	1	-0.073	-0.094	-0.075	0.033	0.096	-0.03
lead_source_Other Social Sites	-0.0053	-0.022	-0.048	-0.035	-0.095	-0.14	-0.13	-0.095	-0.073	1	-0.05	-0.0043	0.046	0.0097	0.019
lead_source_Reference	-0.04	-0.2	-0.11	-0.24	-0.15	-0.27	-0.17	-0.12	-0.094	-0.05	1	0.032	-0.17	-0.032	0.0077
specialization_Management Specializations	-0.031	-0.04	0.024	-0.043	-0.029	-0.018	0.048	-0.016	-0.075	-0.0043	0.032	1	-0.036	-0.038	0.026
occupation_Unemployed	0.029	0.028	-0.051	0.045	0.019	0.032	0.039	0.0036	0.033	0.046	-0.17	-0.036	1	0.0013	-0.0042
city_Non-Maharashtra Cities	0.018	0.0039	0.0089	0.0088	0.1	0.035	-0.1	-0.00035	0.096	0.0097	-0.032	-0.038	0.0013	1	-0.26
city_Non-Mumbai Maharashtra Cities	-0.0033	-0.027	-0.0065	-0.017	-0.061	-0.029	0.0045	0.023	-0.03	0.019	0.0077	0.026	-0.0042	-0.26	1

FINAL MODEL



	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3502	0.096	-3.646	0.000	-0.539	-0.162
do_not_email	-1.2215	0.144	-8.461	0.000	-1.504	-0.939
total_visits	0.1372	0.042	3.296	0.001	0.056	0.219
time_on_website	1.0382	0.036	29.230	0.000	0.969	1.108
page_views_per_visit	-0.1809	0.047	-3.852	0.000	-0.273	-0.089
lead_source_Google	0.3657	0.079	4.643	0.000	0.211	0.520
lead_source_Olark Chat	0.6542	0.113	5.809	0.000	0.433	0.875
lead_source_Organic Search	0.2431	0.108	2.256	0.024	0.032	0.454
lead_source_Other Social Sites	1.6180	0.157	10.311	0.000	1.310	1.926
lead_source_Reference	3.9527	0.206	19.154	0.000	3.548	4.357
occupation_Unemployed	-0.8203	0.085	-9.630	0.000	-0.987	-0.653
city_Non-Mumbai Maharashtra Cities	0.1348	0.073	1.845	0.065	-0.008	0.278

Confusion matrix AND ACCURACY



		PREDICTED	
		Positive	Negative
ACTUAL	Positive	3490	512
	Negative	1029	1437

- TRAIN DATA ACCURACY - 0.761
- TEST DATA ACCURACY - 0.751

OBSERVATIONS

- A high number of website visits by a lead increases the chance of conversion.
- More time spent on the website is a good sign for a lead conversion.
- A lead from ‘reference’ is more likely to be converted.
- More numbers of leads are from Maharashtra state(other than Mumbai city).
- There is very less chance for an unemployed lead to be converted so Some separate schemes are required for such leads.
- A Lead opted out for the mail-updates or viewing very less pages in a visit is not a good sign.

THANK YOU