Lead Scoring Case Study using logistic regression

Submitted by

1.Karishma Sahay

2.Keshav Raj

3. Swati Mehta

Contents

- Problem Statement
- Problem Approach
- EDA
- Correlations
- Model Evaluation
- Observations
- Conclusion

Problem Statement

- An education company named X Education sells online courses to industry professionals. Many professionals who are interested in the courses land on their website and browse for courses. They have process of form filling on their website after which the company that individual as a lead.
- 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, this means 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, that are 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.

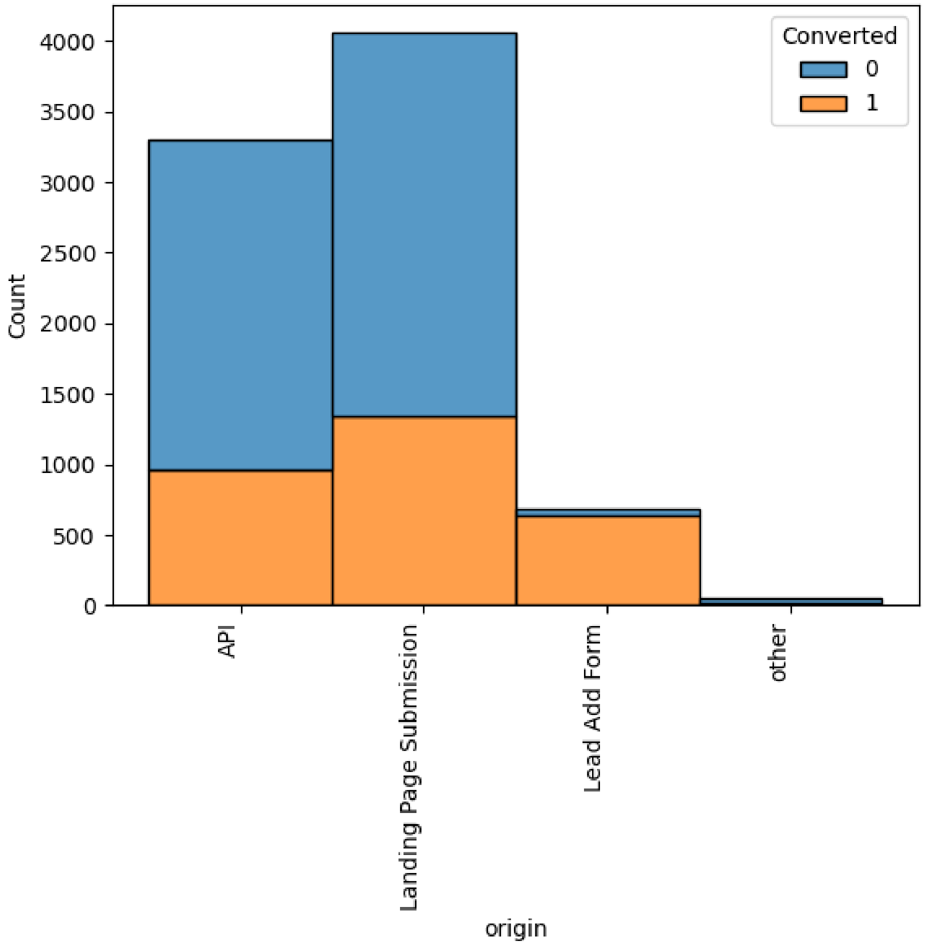
Objective

- Lead X wants us to build a model to give every lead a lead score between 0 -100.
- So that they can identify the Hot leads and increase their conversion rate as well.
- The CEO want to achieve a lead conversion rate of 80%.
- They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full man power and after achieving target what should be the approaches.

Problem Approach

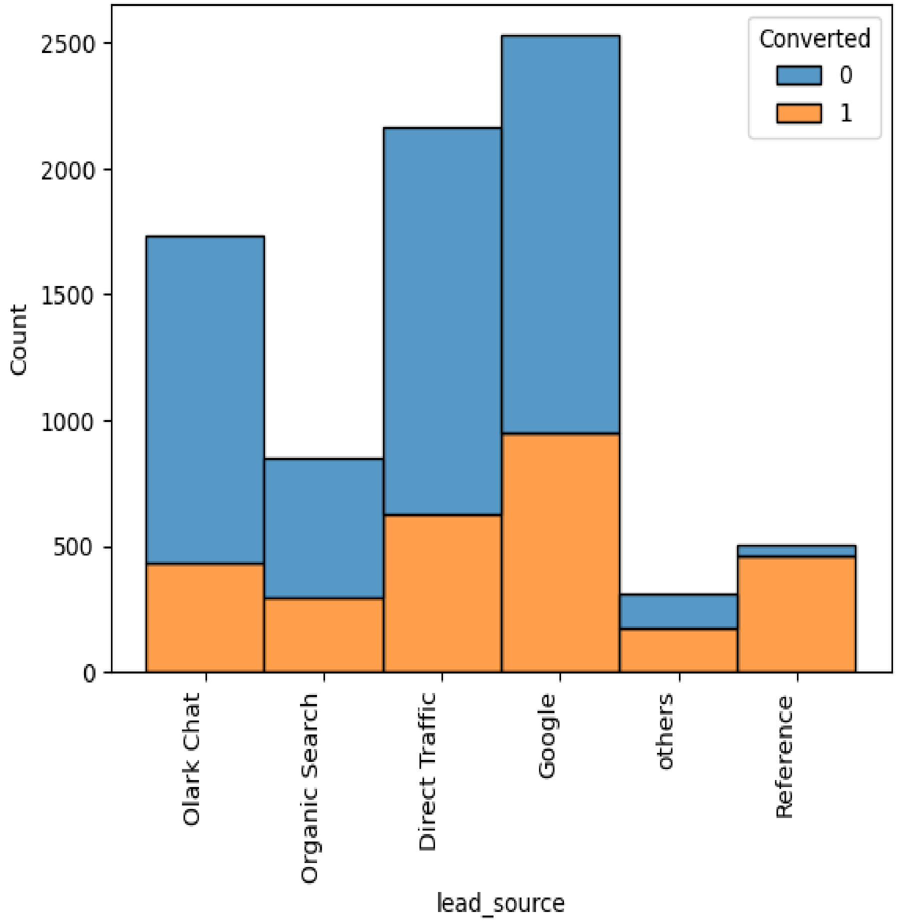
- Importing the data and inspecting the
- data frame
- Data preparation
- EDA
- Dummy variable creation
- Test-Train split
- Feature scaling
- Correlations
- Model Building (RFE,VIF and pvalues)
- Model Evaluation (Specificity, Sensitivity,Accuracy)
- Making predictions on test set

Exploratory Data Analysis



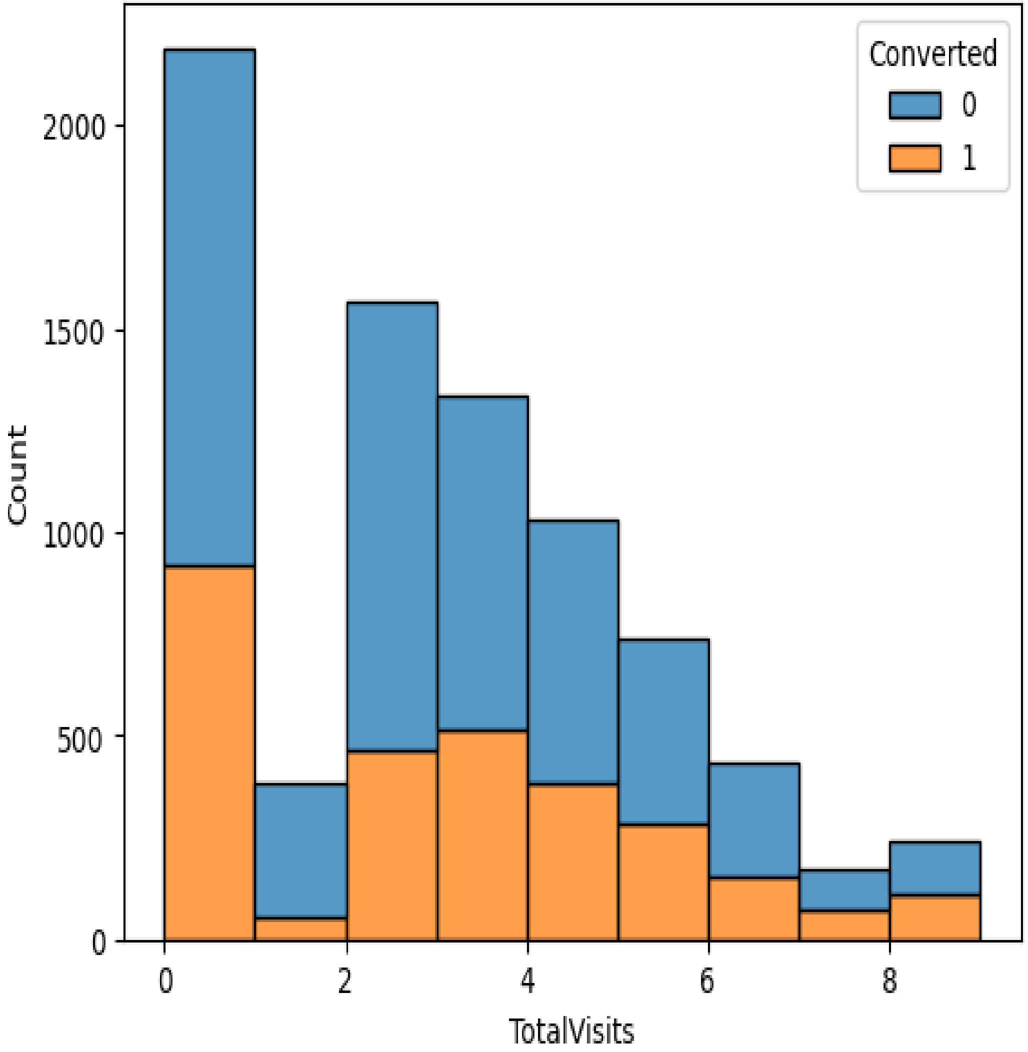
- In these graph we can see the higher conversation rate is in ‘Lead Add Form’.
- And there is no conversation in ‘other’.

Lead Source



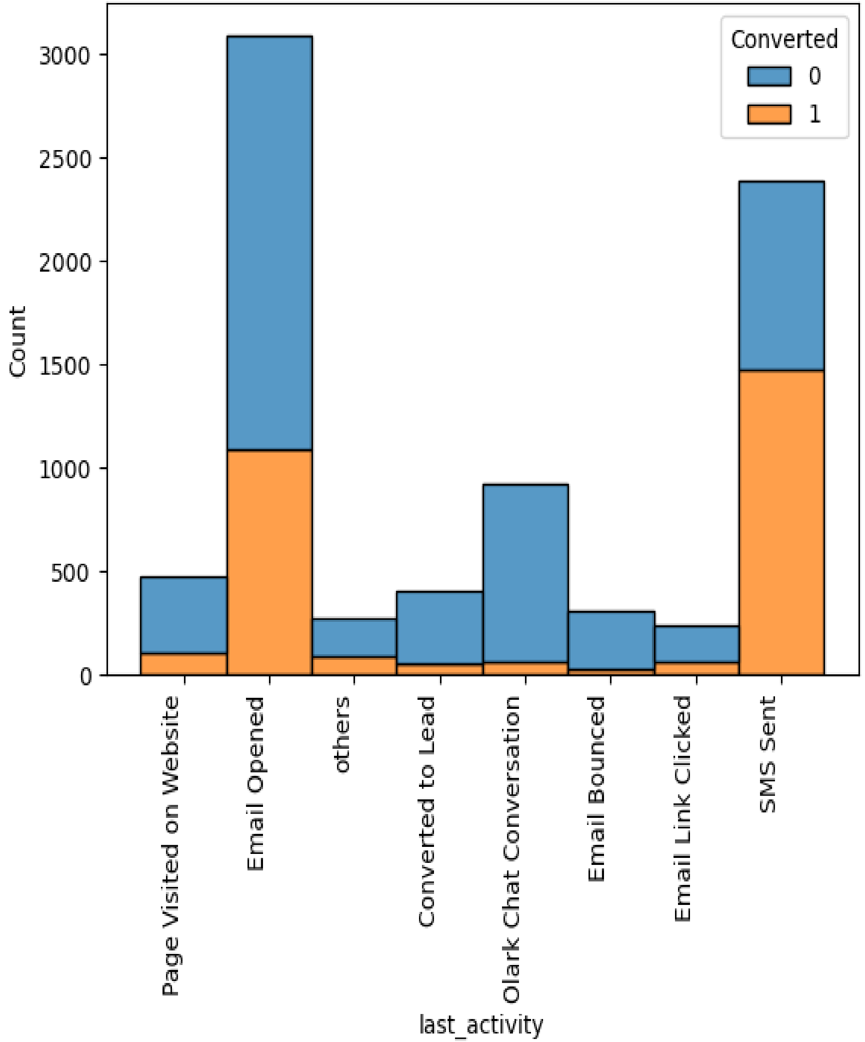
- In these graph lead_source, we can observe that the higher rate of conversation is in ‘Reference’.

Total Visit



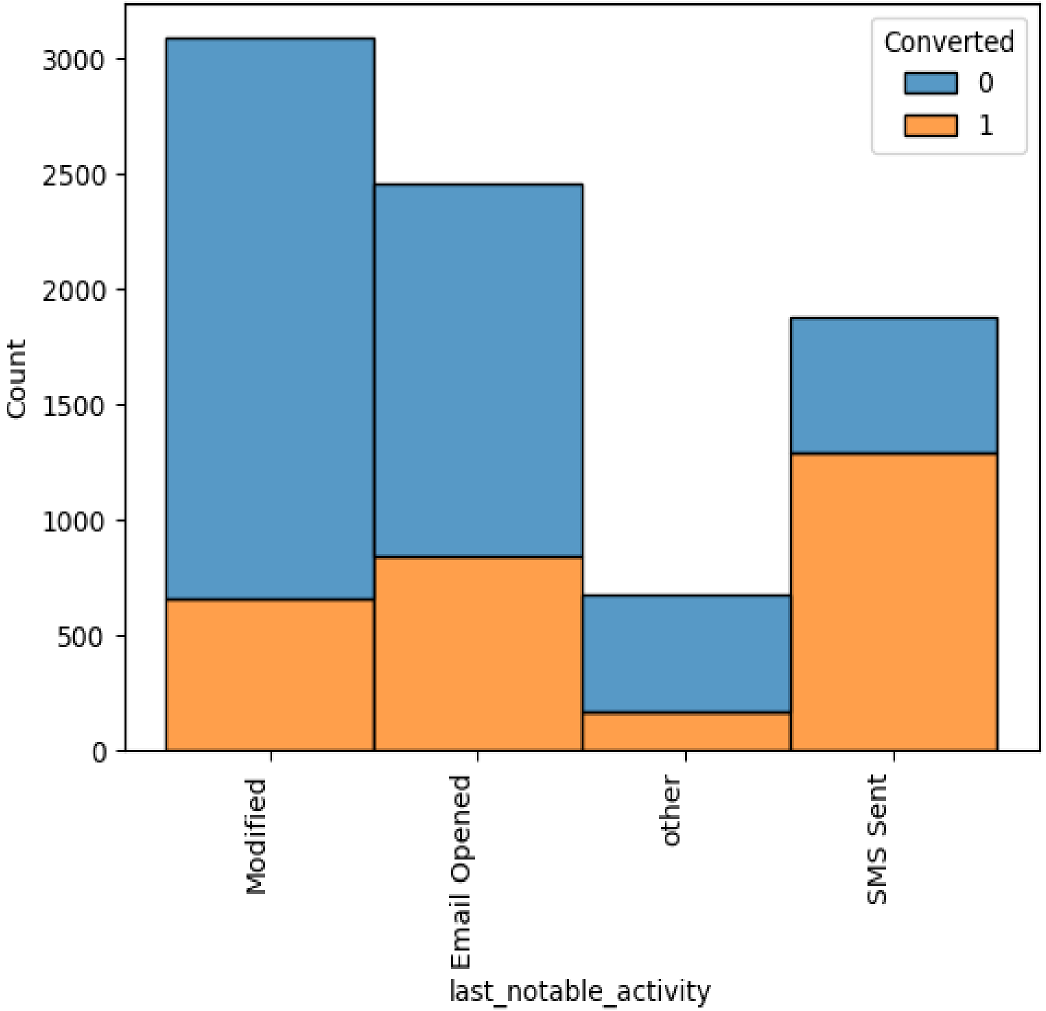
- In these graph, Total visit does not show a great impact on whether a lead was converted or not.

Last Activity



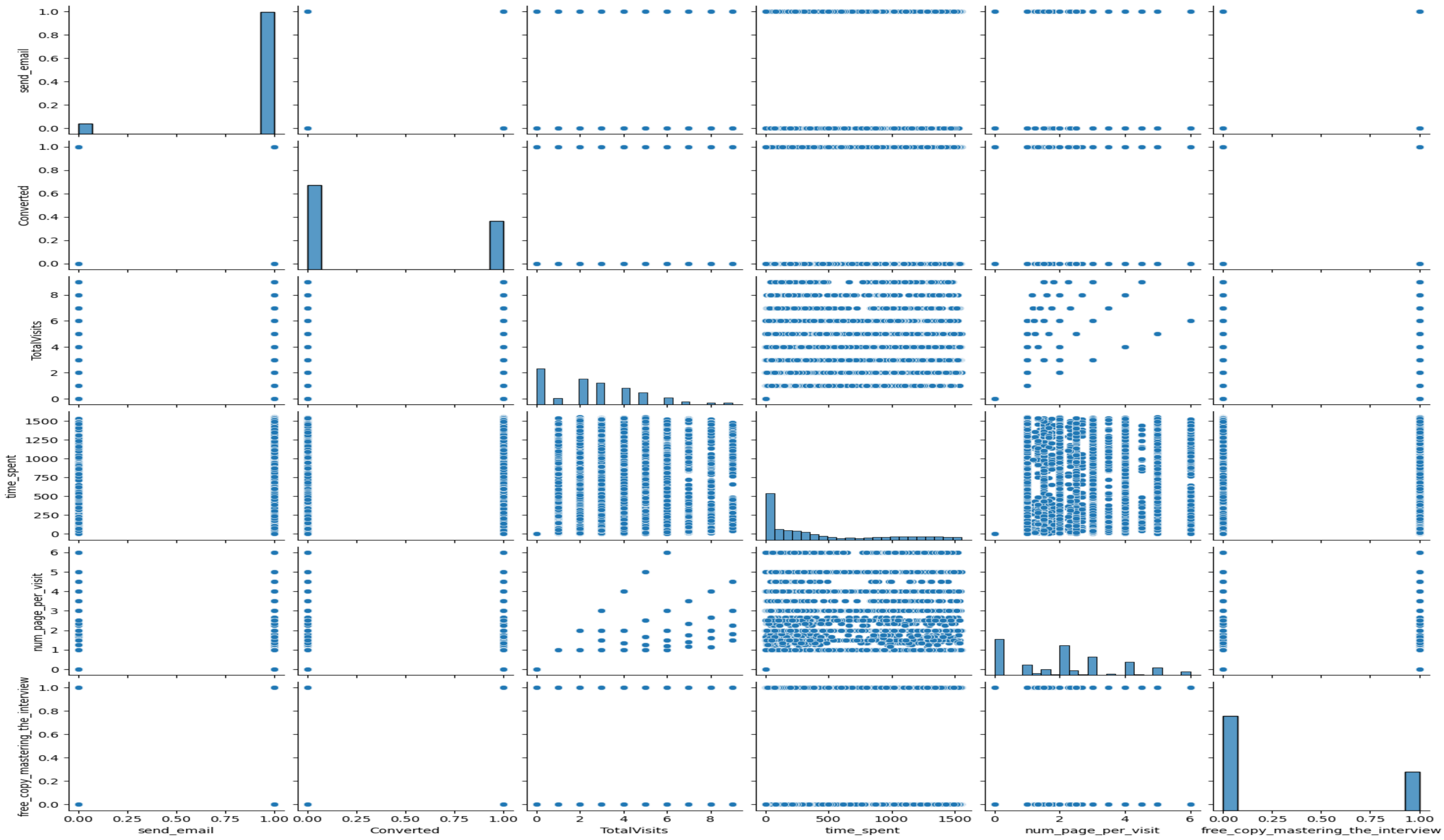
- By these graph we understand that ,one whose last activity was SMS sent are converted more.
- Also Email opened has also almost 33% conversion rate.

Last Notable Activity



- In these graph of last_notable_activity ‘SMS Sent’ seems to have higher impact on conversion rate.

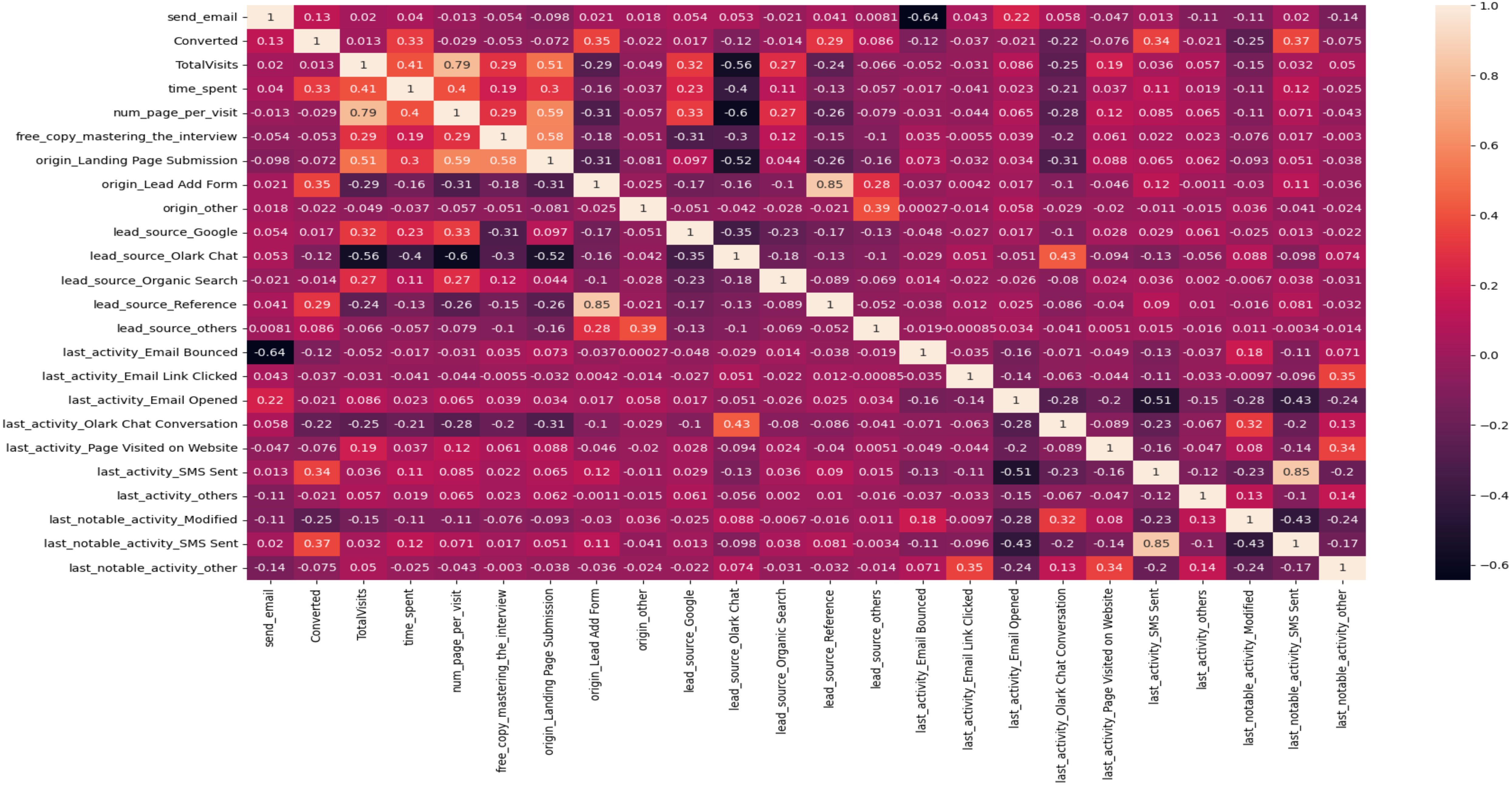
Observation After EDA



Observation After EDA

- Conversion rate is very less. i.e. almost 38%
- We had to drop a lot of features because either null_percentage > 10% or more than 95% values were same.
- Most of the lead origin is API or Landing Page Submission. But their conversion rate is very low. However Lead Add Form is having very high conversion rate
- Google, OLARK Chat and Direct traffic are major lead sources but having low conversion rates. Referral is having very high conversion rate.
- If user has asked not to send the email, then he is not going to convert for sure.
- Total visits, Time spent and num of pages per visit was having outliers. We have taken values up to the 98th percentile.
- Users who have spent more time on the website are more likely to be converted.
- Users who have opted for free copy of mastering the interview does not seem.

Correlations



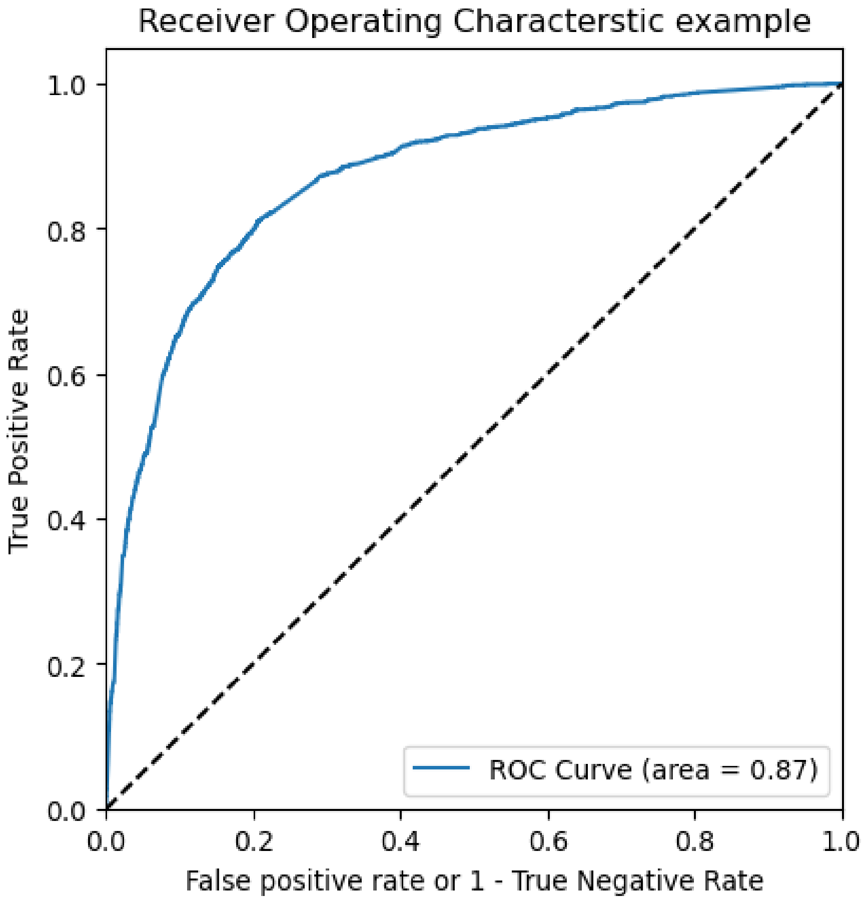
	send_email	Converted	TotalVisits	time_spent	num_page_per_visit	free_copy_mastering_the_interview	origin_Landing Page Submission	origin_Lead Add Form	origin_other	lead_source_Google	lead_source_Olark Chat	lead_source_Organic Search	lead_source_Reference	lead_source_others	last_activity_Email Bounced	last_activity_Email Link Clicked	last_activity_Email Opened	last_activity_Olark Chat Conversation	last_activity_Page Visited on Website	last_activity_SMS Sent	last_activity_others	last_notable_activity_Modified	last_notable_activity_SMS Sent	last_notable_activity_other
send_email	1	0.13	0.02	0.04	-0.013	-0.054	-0.098	0.021	0.018	0.054	0.053	-0.021	0.041	0.0081	-0.64	0.043	0.22	0.058	-0.047	0.013	-0.11	-0.11	0.02	-0.14
Converted	0.13	1	0.013	0.33	-0.029	-0.053	-0.072	0.35	-0.022	0.017	-0.12	-0.014	0.29	0.086	-0.12	-0.037	-0.021	-0.22	-0.076	0.34	-0.021	-0.25	0.37	-0.075
TotalVisits	0.02	0.013	1	0.41	0.79	0.29	0.51	-0.29	-0.049	0.32	-0.56	0.27	-0.24	-0.066	-0.052	-0.031	0.086	-0.25	0.19	0.036	0.057	-0.15	0.032	0.05
time_spent	0.04	0.33	0.41	1	0.4	0.19	0.3	-0.16	-0.037	0.23	-0.4	0.11	-0.13	-0.057	-0.017	-0.041	0.023	-0.21	0.037	0.11	0.019	-0.11	0.12	-0.025
num_page_per_visit	-0.013	-0.029	0.79	0.4	1	0.29	0.59	-0.31	-0.057	0.33	-0.6	0.27	-0.26	-0.079	-0.031	-0.044	0.065	-0.28	0.12	0.085	0.065	-0.11	0.071	-0.043
free_copy_mastering_the_interview	-0.054	-0.053	0.29	0.19	0.29	1	0.58	-0.18	-0.051	-0.31	-0.3	0.12	-0.15	-0.1	0.035	-0.0055	0.039	-0.2	0.061	0.022	0.023	-0.076	0.017	-0.003
origin_Landing Page Submission	-0.098	-0.072	0.51	0.3	0.59	0.58	1	-0.31	-0.081	0.097	-0.52	0.044	-0.26	-0.16	0.073	-0.032	0.034	-0.31	0.088	0.065	0.062	-0.093	0.051	-0.038
origin_Lead Add Form	0.021	0.35	-0.29	-0.16	-0.31	-0.18	-0.31	1	-0.025	-0.17	-0.16	-0.1	0.85	0.28	-0.037	0.0042	0.017	-0.1	-0.046	0.12	-0.0011	-0.03	0.11	-0.036
origin_other	0.018	-0.022	-0.049	-0.037	-0.057	-0.051	-0.081	-0.025	1	-0.051	-0.042	-0.028	-0.021	0.39	0.00027	-0.014	0.058	-0.029	-0.02	-0.011	-0.015	0.036	-0.041	-0.024
lead_source_Google	0.054	0.017	0.32	0.23	0.33	-0.31	0.097	-0.17	-0.051	1	-0.35	-0.23	-0.17	-0.13	-0.048	-0.027	0.017	-0.1	0.028	0.029	0.061	-0.025	0.013	-0.022
lead_source_Olark Chat	0.053	-0.12	-0.56	-0.4	-0.6	-0.3	-0.52	-0.16	-0.042	-0.35	1	-0.18	-0.13	-0.1	-0.029	0.051	-0.051	0.43	-0.094	-0.13	-0.056	0.088	-0.098	0.074
lead_source_Organic Search	-0.021	-0.014	0.27	0.11	0.27	0.12	0.044	-0.1	-0.028	-0.23	-0.18	1	-0.089	-0.069	0.014	-0.022	-0.026	-0.08	0.024	0.036	0.002	-0.0067	0.038	-0.031
lead_source_Reference	0.041	0.29	-0.24	-0.13	-0.26	-0.15	-0.26	0.85	-0.021	-0.17	-0.13	-0.089	1	-0.052	-0.038	0.012	0.025	-0.086	-0.04	0.09	0.01	-0.016	0.081	-0.032
lead_source_others	0.0081	0.086	-0.066	-0.057	-0.079	-0.1	-0.16	0.28	0.39	-0.13	-0.1	-0.069	-0.052	1	-0.019	-0.00085	0.034	-0.041	0.0051	0.015	-0.016	0.011	-0.0034	-0.014
last_activity_Email Bounced	-0.64	-0.12	-0.052	-0.017	-0.031	0.035	0.073	-0.037	0.00027	-0.048	-0.029	0.014	-0.038	-0.019	1	-0.035	-0.16	-0.071	-0.049	-0.13	-0.037	0.18	-0.11	0.071
last_activity_Email Link Clicked	0.043	-0.037	-0.031	-0.041	-0.044	-0.0055	-0.032	0.0042	-0.014	-0.027	0.051	-0.022	0.012	-0.00085	-0.035	1	-0.14	-0.063	-0.044	-0.11	-0.033	-0.0097	-0.096	0.35
last_activity_Email Opened	0.22	-0.021	0.086	0.023	0.065	0.039	0.034	0.017	0.058	0.017	-0.051	-0.026	0.025	0.034	-0.16	-0.14	1	-0.28	-0.2	-0.51	-0.15	-0.28	-0.43	-0.24
last_activity_Olark Chat Conversation	0.058	-0.22	-0.25	-0.21	-0.28	-0.2	-0.31	-0.1	-0.029	-0.1	0.43	-0.08	-0.086	-0.041	-0.071	-0.063	-0.28	1	-0.089	-0.23	-0.067	0.32	-0.2	0.13
last_activity_Page Visited on Website	-0.047	-0.076	0.19	0.037	0.12	0.061	0.088	-0.046	-0.02	0.028	-0.094	0.024	-0.04	0.0051	-0.049	-0.044	-0.2	-0.089	1	-0.16	-0.047	0.08	-0.14	0.34
last_activity_SMS Sent	0.013	0.34	0.036	0.11	0.085	0.022	0.065	0.12	-0.011	0.029	-0.13	0.036	0.09	0.015	-0.13	-0.11	-0.51	-0.23	-0.16	1	-0.12	-0.23	0.85	-0.2
last_activity_others	-0.11	-0.021	0.057	0.019	0.065	0.023	0.062	-0.0011	-0.015	0.061	-0.056	0.002	0.01	-0.016	-0.037	-0.033	-0.15	-0.067	-0.047	-0.12	1	0.13	-0.1	0.14
last_notable_activity_Modified	-0.11	-0.25	-0.15	-0.11	-0.11	-0.076	-0.093	-0.03	0.036	-0.025	0.088	-0.0067	-0.016	0.011	0.18	-0.0097	-0.28	0.32	0.08	-0.23	0.13	1	-0.43	-0.24
last_notable_activity_SMS Sent	0.02	0.37	0.032	0.12	0.071	0.017	0.051	0.11	-0.041	0.013	-0.098	0.038	0.081	-0.0034	-0.11	-0.096	-0.43	-0.2	-0.14	0.85	-0.1	-0.43	1	-0.17
last_notable_activity_other	-0.14	-0.075	0.05	-0.025	-0.043	-0.003	-0.038	-0.036	-0.024	-0.022	0.074	-0.031	-0.032	-0.014	0.071	0.35	-0.24	0.13	0.34	-0.2	0.14	-0.24	-0.17	1

Observation from correlation

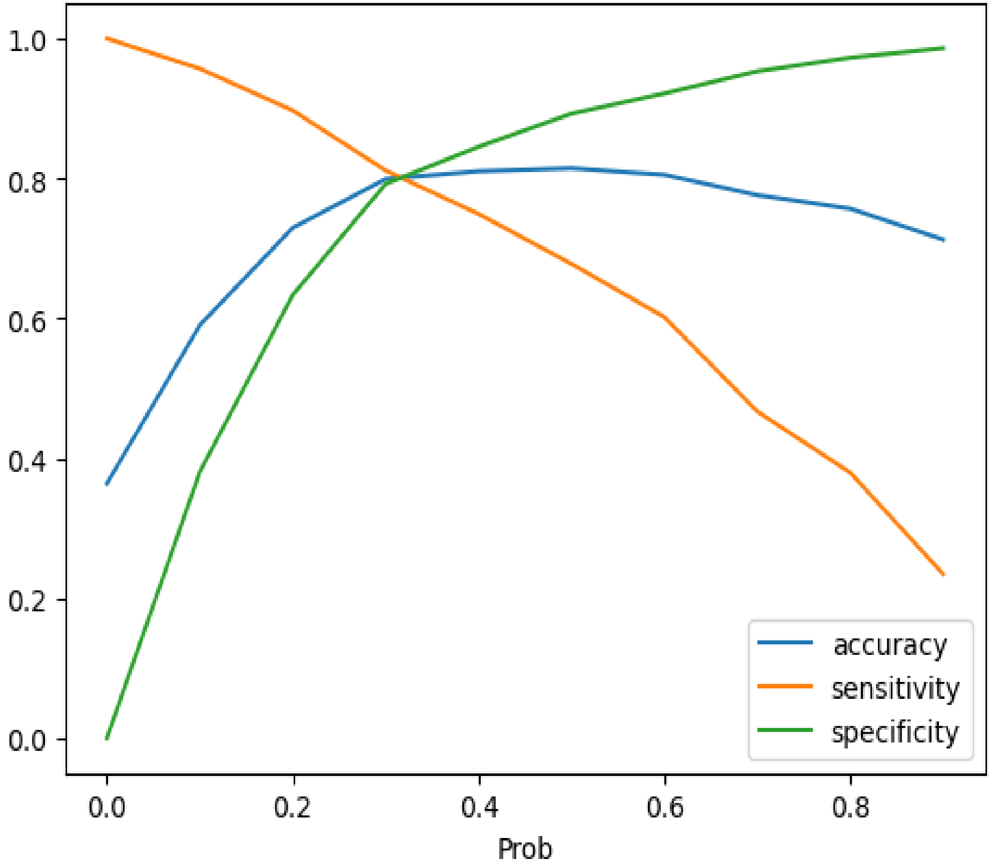
- Send_email and last_activity_email_bounced high correlation
- lead_source_olark_chat has high correlation with origin_landing_page_submission, num_pages_per_visit and total_visits
- num_page per visit has high correlation with num_pages_per_visit
- origin_landing_page_submission has high correlation with free_copy_mastering_the_interview and num_pages_per_visit
- Hope that these collinearity will be dealt while RFE and VIF Step. Just to keep eye.

Model Evaluation

- Here model looks good as AUC is 0.87 for ROC



- As looking at the below plot we choose 0.3 as cutoff point



Conclusion

- We created a model with finally 9 Features.
- We choose 0.3 as cutoff value.
- Model Evaluation on Train data :

1. accuracy = 0.79
2. sensitivity = 0.81
3. specificity = 0.79

- Model Evaluation on Test Data :

1. accuracy = 0.79
2. sensitivity = 0.80
3. specificity = 0.79

- We see max number of leads are generated by google/direct traffic.
- Most common last activity is email opened, highest rate = SMS Sent.
- Finally we merged predicted data frame with test data on index and assigned lead score by multiplying with 100

Business Recommendation

- In order to increase probability of more lead conversion. we should focus on what actions on organization end result in more lead conversion. As per our model –

1) Lead source Olark chat has a positive coefficient of 1.4738. So, we should try to put more people on Olark chat to generate more lead source from there.
2) Last activity Olark chat conversation has negative coefficient of -1.3043. It means a lot of our leads drop off after the Olark chat conversation. So, we need to find out if our employees are not trained enough who are there on Olark chat and decrease the drop off from there itself.
3) Email bounced has a negative coefficient of -1.5312. So, it means we are not validating the email and entering the wrong email on the lead form. Let it be website or past referrals. On website we should check if email validation is there before form submission.

Business Recommendation

- When business wants to go more aggressive on lead conversion , they should increase the sensitivity and recall of the model. In simpler words, No of actual leads predicted to total no of actual leads and our model does not predict a hot lead in to cold lead should be more.
- Also when company wants to minimize the rate of useless phone calls, we should focus more on specificity and precision(probability that a predicted lead is actual lead). We can compromise on sensitivity of the model. In our model at cutoff = 0.6 sensitivity = 0.60 specificity = 0.92 precision = 0.81.