LEAD SCORE CASE STUDY

Group members :-

1. Pramod Poojary
2. Shivani Gupta

PROBLEM STATEMENT

X Education sells online courses to industry professionals

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.

BUSINESS OBJECTIVE

- 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 we 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.
- Target lead conversion rate to be around 80%.
- Deployment of the model for the future use.

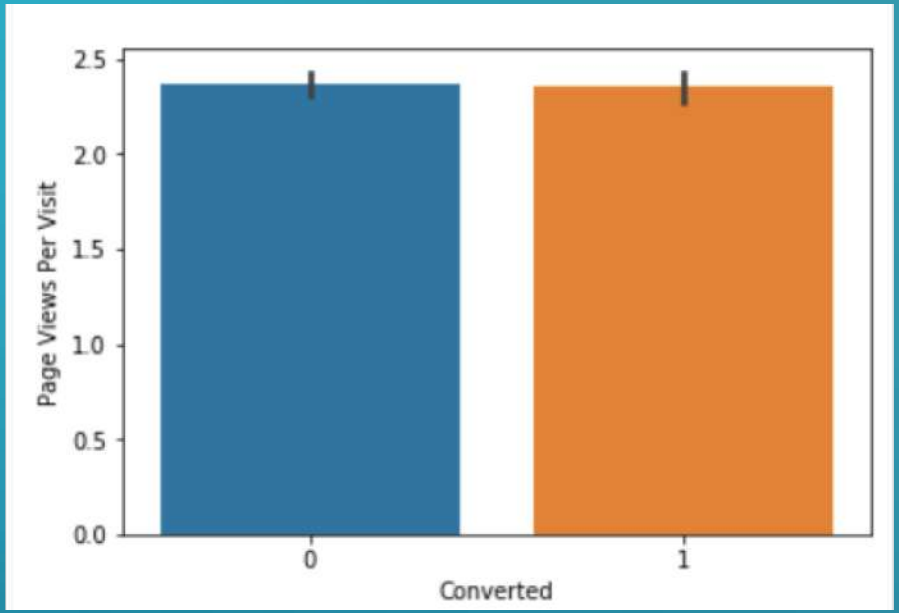
SOLUTION METHODOLOGY

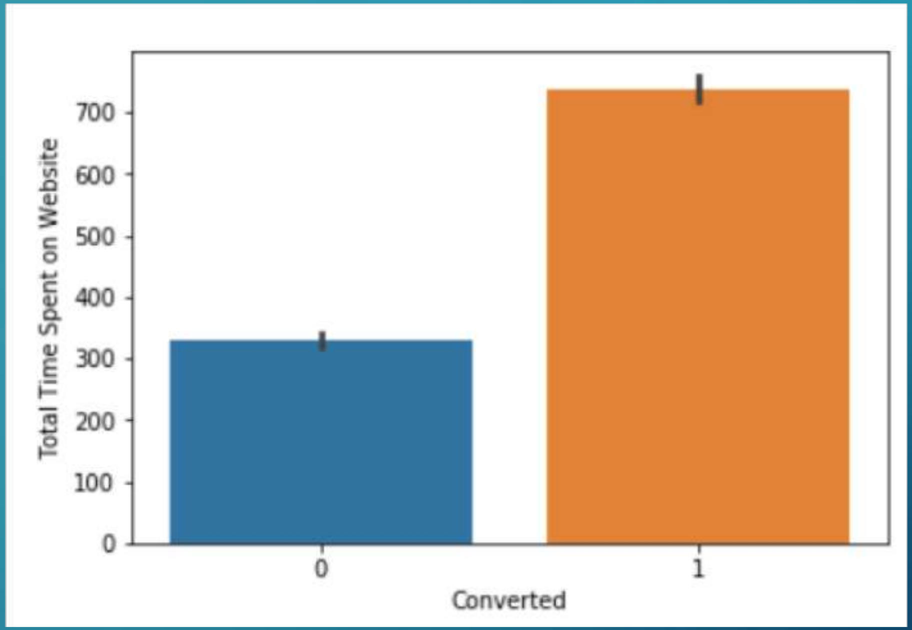
1. Data cleaning and Data manipulation (Check and handle duplicate data, null values, missing values and outliers in data).
2. Mapping Yes/No values to 0/1, dummy variable creation and splitting data into Test-Train.
3. Feature scaling and finding correlations between variables.
4. Building Logistic Regression model
5. Validation of the model
6. Conclusion

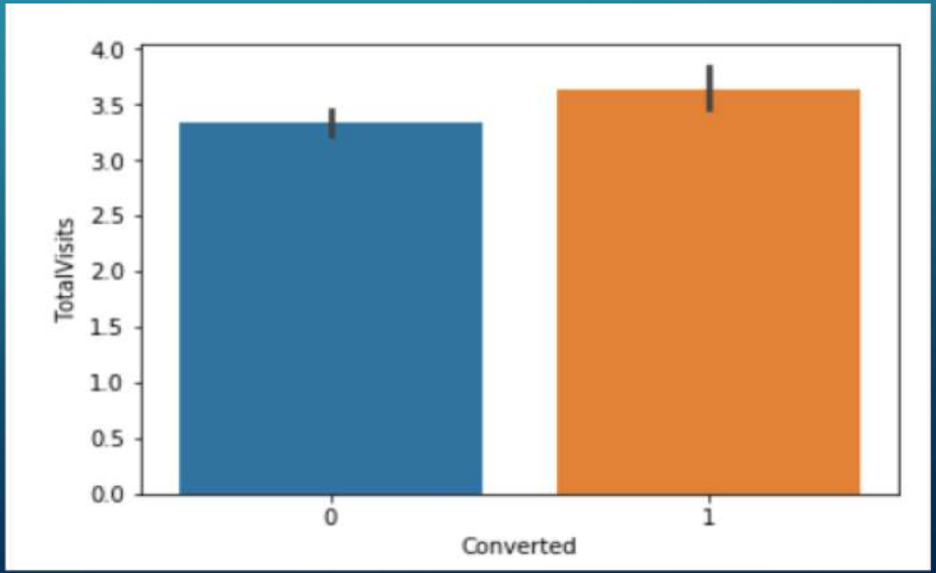
DATA CLEANING AND MANIPULATION

- ‘Select’ present in many columns replaced with NaN value as it is useless.
- Insignificant columns like “City”, “Country” and “Prospect ID” are dropped.
- Columns containing more than 3000 columns has been dropped.
- Retained the rows having <=5 NaN’s.
- After removing above all and outliers , we left with 8997 Rows and 24 Columns.

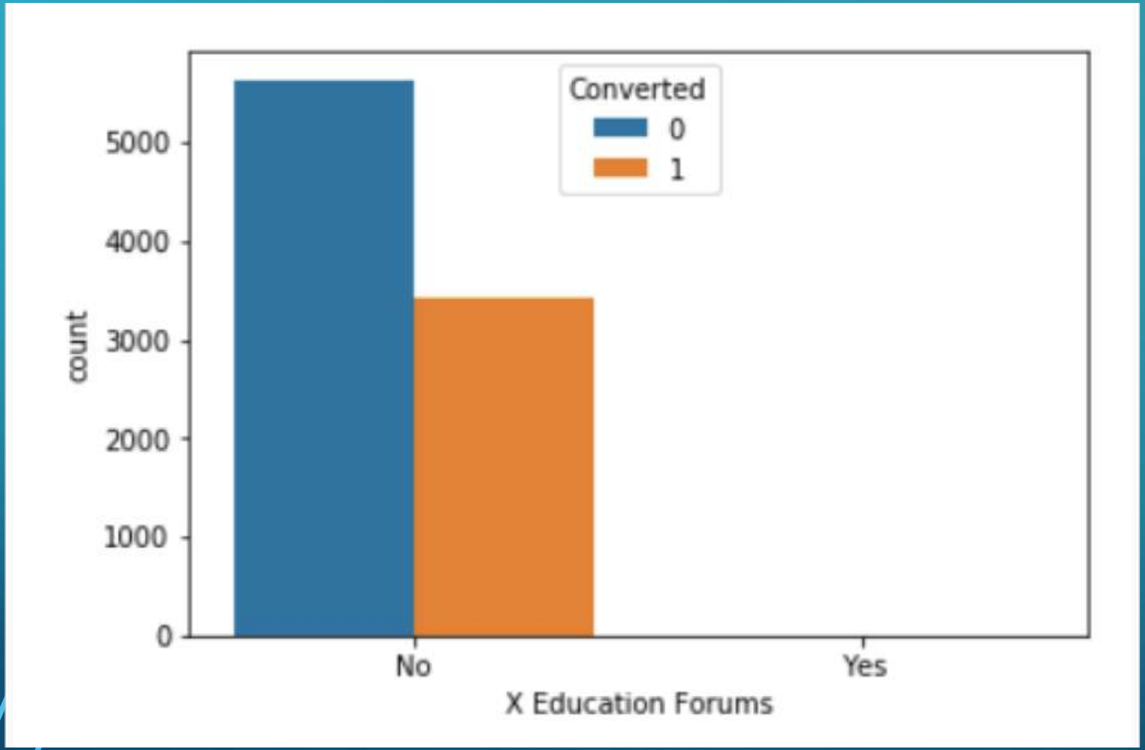
DATA VISUALIZATION OF CONTINUOUS VARIABLES

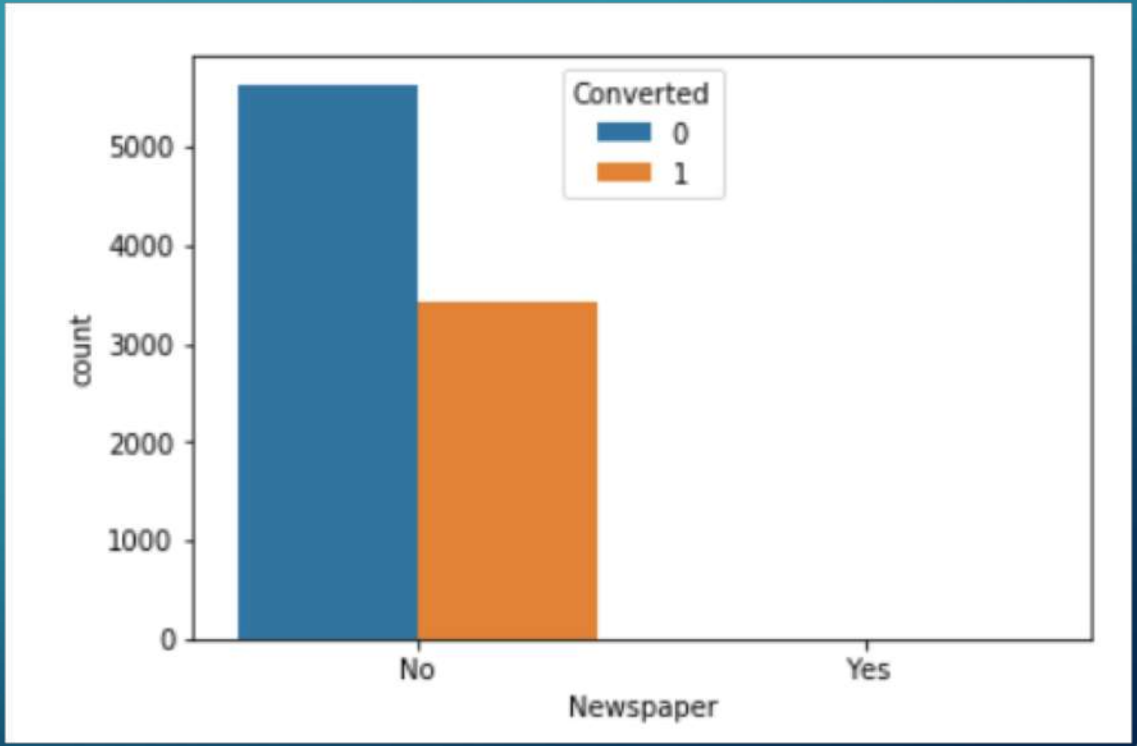




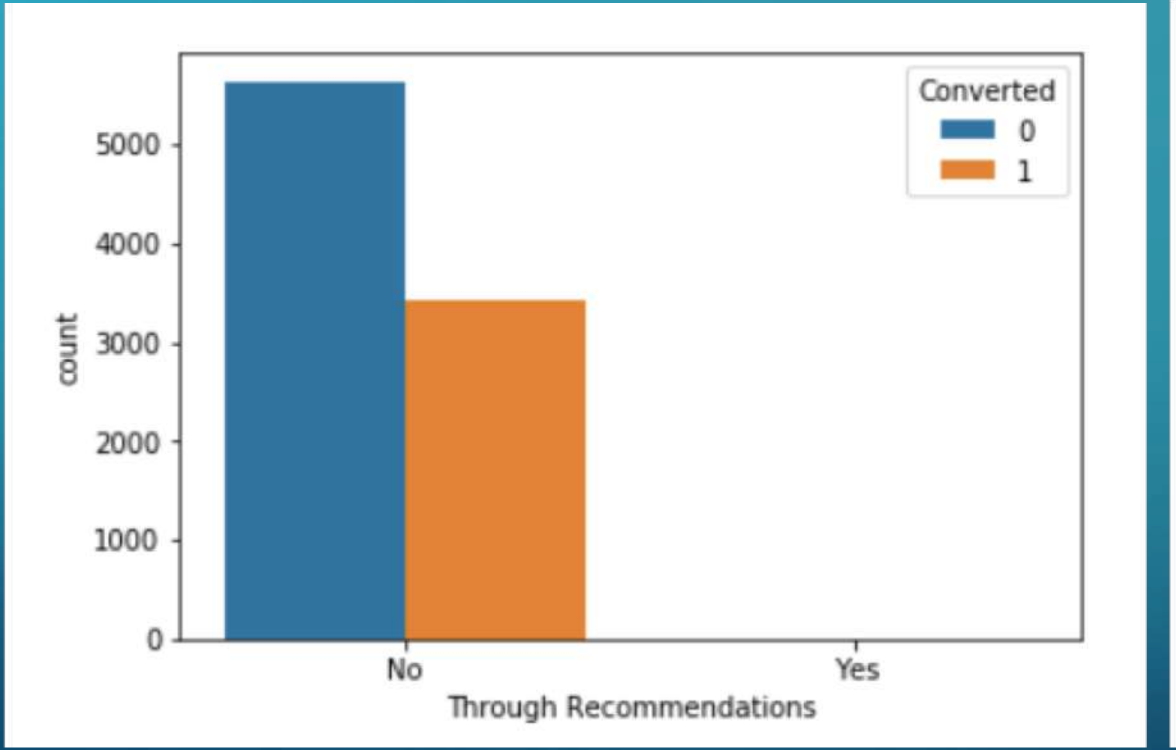


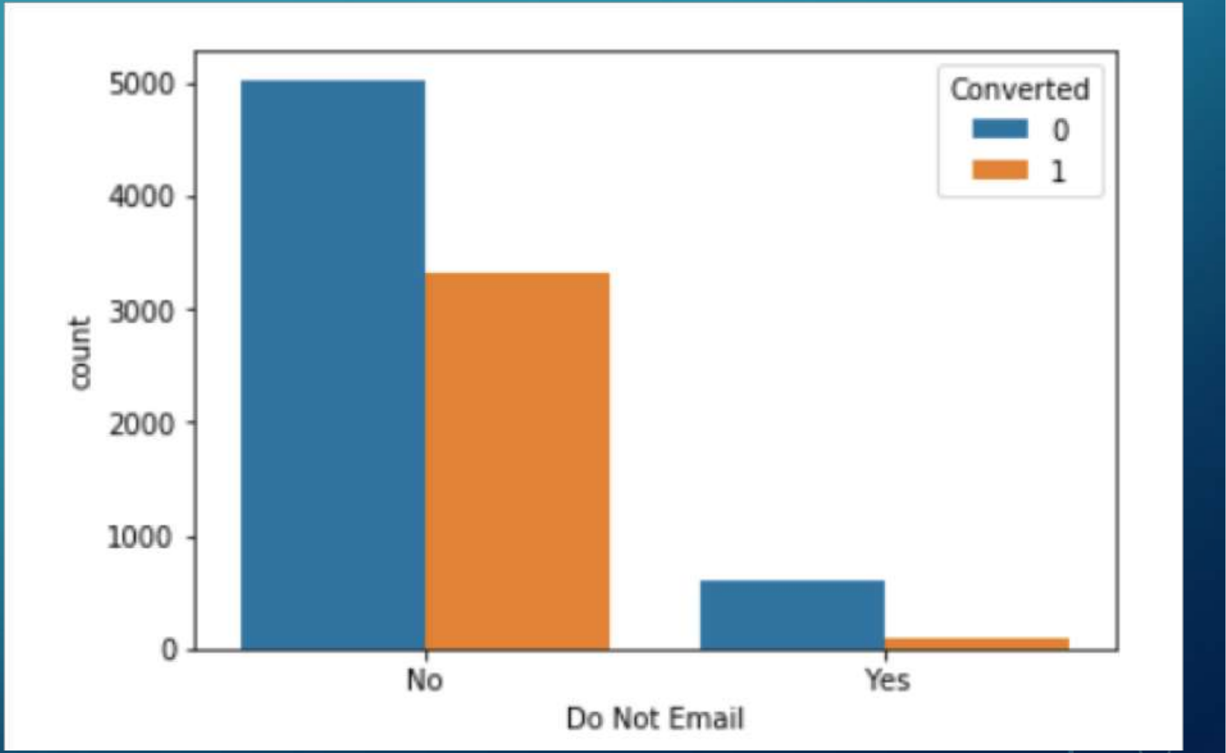
VISUALIZATION OF VARIABLES





VISUALIZATION CONTINUED..

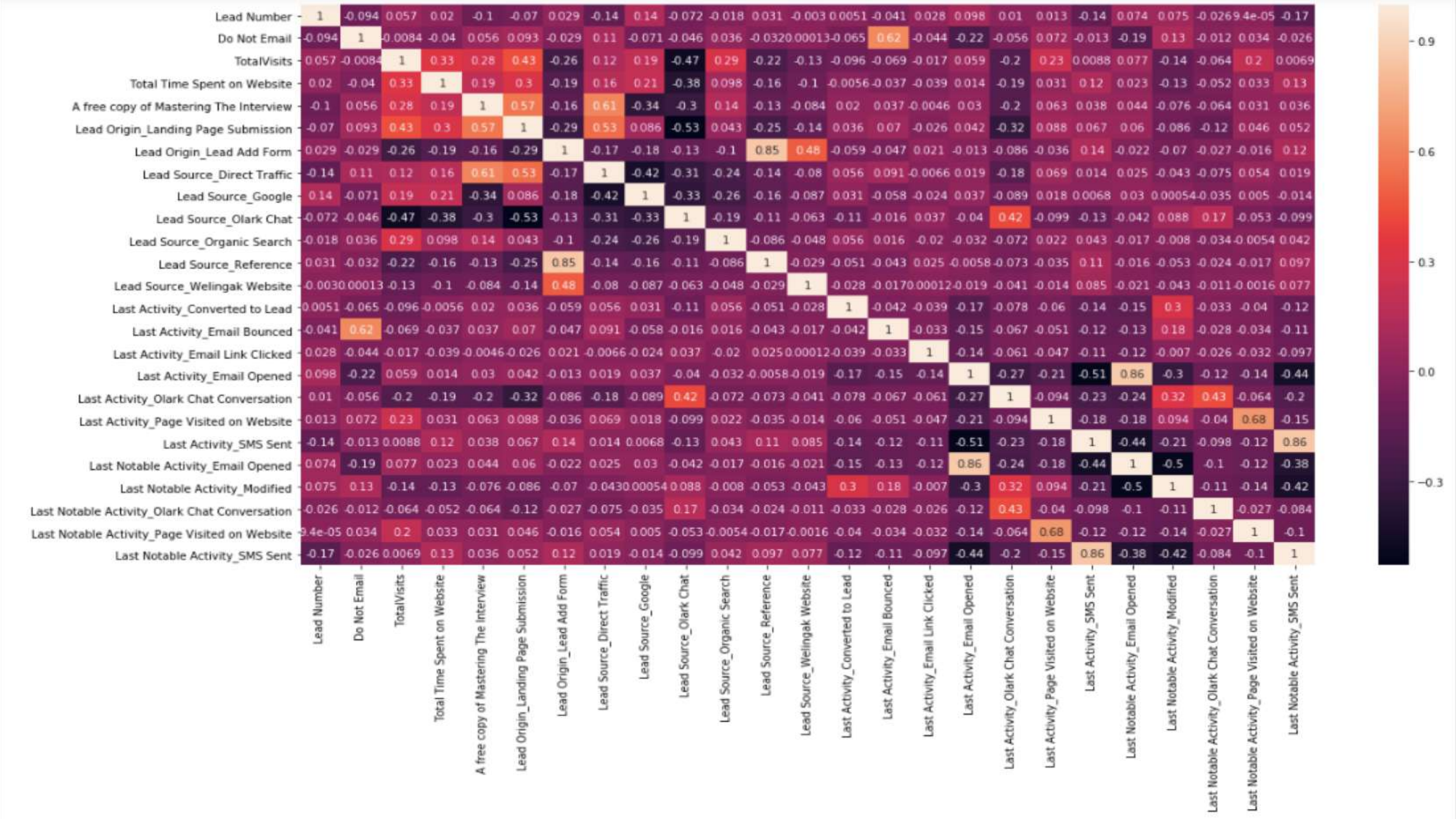




MODEL BUILDING

- Basic step to splitting the data into training and test sets and we choosen 70:30 ratio.
- Used RFE for feature selection.
- Running RFE with 24 variables as output.
- Removing the variable whose P-value is greater than 0.05 and VIF is greater than 0.02.
- Prediction on test data set.

CORRELATION BETWEEN VARIABLES



	Lead Number	Do Not Email	TotalVisits	Total Time Spent on Website	A free copy of Mastering The Interview	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Source_Direct Traffic	Lead Source_Google	Lead Source_Olark Chat	Lead Source_Organic Search	Lead Source_Reference	Lead Source_Welingak Website	Last Activity_Converted to Lead	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 Notable Activity_Email Opened	Last Notable Activity_Modified	Last Notable Activity_Olark Chat Conversation	Last Notable Activity_Page Visited on Website	Last Notable Activity_SMS Sent
Lead Number	1	-0.094	0.057	0.02	-0.1	-0.07	0.029	-0.14	0.14	-0.072	-0.018	0.031	-0.003	0.0051	-0.041	0.028	0.098	0.01	0.013	-0.14	0.074	0.075	-0.026	9.4e-05	-0.17
Do Not Email	-0.094	1	-0.0084	-0.04	0.056	0.093	-0.029	0.11	-0.071	-0.046	0.036	-0.032	0.00013	-0.065	0.62	-0.044	-0.22	-0.056	0.072	-0.013	-0.19	0.13	-0.012	0.034	-0.026
TotalVisits	0.057	-0.0084	1	0.33	0.28	0.43	-0.26	0.12	0.19	-0.47	0.29	-0.22	-0.13	-0.096	-0.069	-0.017	0.059	-0.2	0.23	0.0088	0.077	-0.14	-0.064	0.2	0.0069
Total Time Spent on Website	0.02	-0.04	0.33	1	0.19	0.3	-0.19	0.16	0.21	-0.38	0.098	-0.16	-0.1	-0.0056	-0.037	-0.039	0.014	-0.19	0.031	0.12	0.023	-0.13	-0.052	0.033	0.13
A free copy of Mastering The Interview	-0.1	0.056	0.28	0.19	1	0.57	-0.16	0.61	-0.34	-0.3	0.14	-0.13	-0.084	0.02	0.037	-0.0046	0.03	-0.2	0.063	0.038	0.044	-0.076	-0.064	0.031	0.036
Lead Origin_Landing Page Submission	-0.07	0.093	0.43	0.3	0.57	1	-0.29	0.53	0.086	-0.53	0.043	-0.25	-0.14	0.036	0.07	-0.026	0.042	-0.32	0.088	0.067	0.06	-0.086	-0.12	0.046	0.052
Lead Origin_Lead Add Form	0.029	-0.029	-0.26	-0.19	-0.16	-0.29	1	-0.17	-0.18	-0.13	-0.1	0.85	0.48	-0.059	-0.047	0.021	-0.013	-0.086	-0.036	0.14	-0.022	-0.07	-0.027	-0.016	0.12
Lead Source_Direct Traffic	-0.14	0.11	0.12	0.16	0.61	0.53	-0.17	1	-0.42	-0.31	-0.24	-0.14	-0.08	0.056	0.091	-0.0066	0.019	-0.18	0.069	0.014	0.025	-0.043	-0.075	0.054	0.019
Lead Source_Google	0.14	-0.071	0.19	0.21	-0.34	0.086	-0.18	-0.42	1	-0.33	-0.26	-0.16	-0.087	0.031	-0.058	-0.024	0.037	-0.089	0.018	0.0068	0.03	0.00054	-0.035	0.005	-0.014
Lead Source_Olark Chat	-0.072	-0.046	-0.47	-0.38	-0.3	-0.53	-0.13	-0.31	-0.33	1	-0.19	-0.11	-0.063	-0.11	-0.016	0.037	-0.04	0.42	-0.099	-0.13	-0.042	0.088	0.17	-0.053	-0.099
Lead Source_Organic Search	-0.018	0.036	0.29	0.098	0.14	0.043	-0.1	-0.24	-0.26	-0.19	1	-0.086	-0.048	0.056	0.016	-0.02	-0.032	-0.072	0.022	0.043	-0.017	-0.008	-0.034	-0.0054	0.042
Lead Source_Reference	0.031	-0.032	-0.22	-0.16	-0.13	-0.25	0.85	-0.14	-0.16	-0.11	-0.086	1	-0.029	-0.051	-0.043	0.025	-0.0058	-0.073	-0.035	0.11	-0.016	-0.053	-0.024	-0.017	0.097
Lead Source_Welingak Website	-0.003	0.00013	-0.13	-0.1	-0.084	-0.14	0.48	-0.08	-0.087	-0.063	-0.048	-0.029	1	-0.028	-0.017	0.00012	-0.019	-0.041	-0.014	0.085	-0.021	-0.043	-0.011	-0.0016	0.077
Last Activity_Converted to Lead	0.0051	-0.065	-0.096	-0.0056	0.02	0.036	-0.059	0.056	0.031	-0.11	0.056	-0.051	-0.028	1	-0.042	-0.039	-0.17	-0.078	-0.06	-0.14	-0.15	0.3	-0.033	-0.04	-0.12
Last Activity_Email Bounced	-0.041	0.62	-0.069	-0.037	0.037	0.07	-0.047	0.091	-0.058	-0.016	0.016	-0.043	-0.017	-0.042	1	-0.033	-0.15	-0.067	-0.051	-0.12	-0.13	0.18	-0.028	-0.034	-0.11
Last Activity_Email Link Clicked	0.028	-0.044	-0.017	-0.039	-0.0046	-0.026	0.021	-0.0066	-0.024	0.037	-0.02	0.025	0.00012	-0.039	-0.033	1	-0.14	-0.061	-0.047	-0.11	-0.12	-0.007	-0.026	-0.032	-0.097
Last Activity_Email Opened	0.098	-0.22	0.059	0.014	0.03	0.042	-0.013	0.019	0.037	-0.04	-0.032	-0.0058	-0.019	-0.17	-0.15	-0.14	1	-0.27	-0.21	-0.51	0.86	-0.3	-0.12	-0.14	-0.44
Last Activity_Olark Chat Conversation	0.01	-0.056	-0.2	-0.19	-0.2	-0.32	-0.086	-0.18	-0.089	0.42	-0.072	-0.073	-0.041	-0.078	-0.067	-0.061	-0.27	1	-0.094	-0.23	-0.24	0.32	0.43	-0.064	-0.2
Last Activity_Page Visited on Website	0.013	0.072	0.23	0.031	0.063	0.088	-0.036	0.069	0.018	-0.099	0.022	-0.035	-0.014	-0.06	-0.051	-0.047	-0.21	-0.094	1	-0.18	-0.18	0.094	-0.04	0.68	-0.15
Last Activity_SMS Sent	-0.14	-0.013	0.0088	0.12	0.038	0.067	0.14	0.014	0.0068	-0.13	0.043	0.11	0.085	-0.14	-0.12	-0.11	-0.51	-0.23	-0.18	1	-0.44	-0.21	-0.098	-0.12	0.86
Last Notable Activity_Email Opened	0.074	-0.19	0.077	0.023	0.044	0.06	-0.022	0.025	0.03	-0.042	-0.017	-0.016	-0.021	-0.15	-0.13	-0.12	0.86	-0.24	-0.18	-0.44	1	-0.5	-0.1	-0.12	-0.38
Last Notable Activity_Modified	0.075	0.13	-0.14	-0.13	-0.076	-0.086	-0.07	-0.043	0.00054	0.088	-0.008	-0.053	-0.043	0.3	0.18	-0.007	-0.3	0.32	0.094	-0.21	-0.5	1	-0.11	-0.14	-0.42
Last Notable Activity_Olark Chat Conversation	-0.026	-0.012	-0.064	-0.052	-0.064	-0.12	-0.027	-0.075	-0.035	0.17	-0.034	-0.024	-0.011	-0.033	-0.028	-0.026	-0.12	0.43	-0.04	-0.098	-0.1	-0.11	1	-0.027	-0.084
Last Notable Activity_Page Visited on Website	9.4e-05	0.034	0.2	0.033	0.031	0.046	-0.016	0.054	0.005	-0.053	-0.0054	-0.017	-0.0016	-0.04	-0.034	-0.032	-0.14	-0.064	0.68	-0.12	-0.12	-0.14	-0.027	1	-0.1
Last Notable Activity_SMS Sent	-0.17	-0.026	0.0069	0.13	0.036	0.052	0.12	0.019	-0.014	-0.099	0.042	0.097	0.077	-0.12	-0.11	-0.097	-0.44	-0.2	-0.15	0.86	-0.38	-0.42	-0.084	-0.1	1

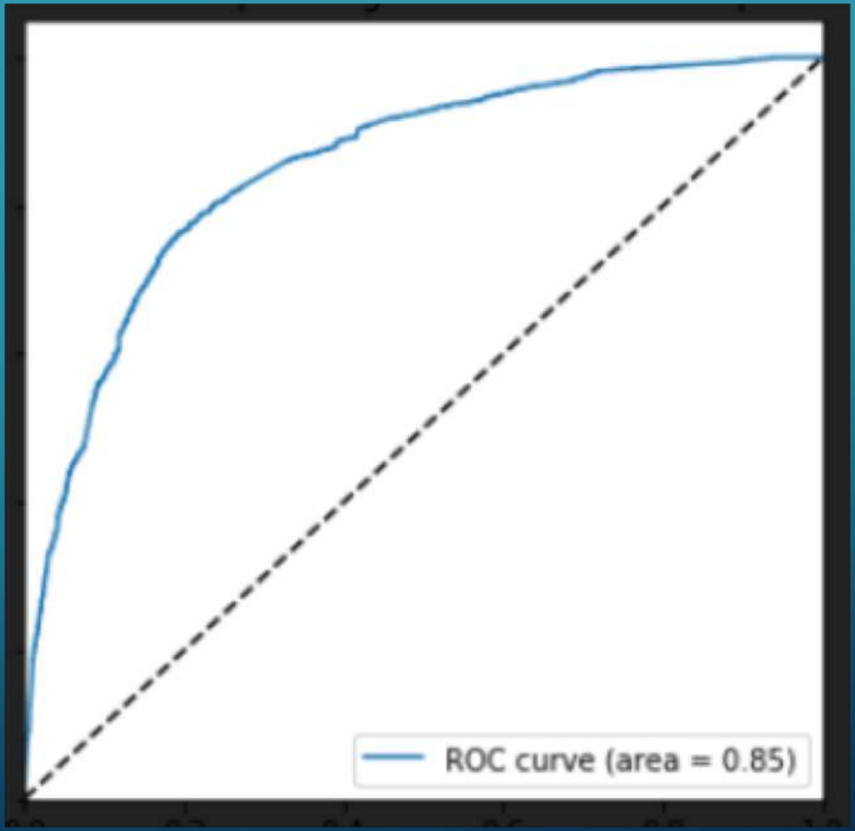
RESULTS

After building the stable model ,we made the prediction and checked the statistics of the model -

- Sensitivity – 66%
- Specificity – 86%
- False positive rate – 13%
- Positive predicted value – 75%
- Negative predicted value – 80.5%

ROC CURVE

ROC curve is Area under the curve is around 0.85 which shows a good model.



CONCLUSION

- Main variables that contribute to analysis are “**Lead Source_Reference**”, “**Total Time Spent on Website**” and “**Last Notable Activity_SMS Sent**”.
- For more conversion we need to consider lower lead score, If we want to cover 100% of the leads, then we should consider score from 0. And If we want to cover around 97% of the leads then we should consider lead scores above 10.
- And we observed from the model if we want to precise conversion we need to consider higher lead score, If we want to be 98% accurate , then we should consider score from 90 or above. And to be around 100% accurate leads then we should consider lead scores above 99.