LEAD SCORING CASE STUDY FOR X EDUCATION

BY:

Yashika Singh

Yash Jindal

Yogesh

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 wants its leads generation rate to be increased. For which they need to identify what are the potential leads aka Hot Leads.
- For this company wants to create a model where we can assign lead score to each of the identified leads for higher chances of conversion. Target for the company is 80 percent.

GOALS TO ACHIEVE

- 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.

Problem solving methodology

- Reading and Understanding the data
- Data Cleaning
- Data Visualization using EDA
- Data Preparation for Modelling
- Model Building: Building the model with features selected by RFE. Eliminate all features with high p-values and VIF values and finalize the model
- Model Evaluation: with various metrics like sensitivity, specificity, precision, recall, etc.

UNDERSTANDING THE DATA

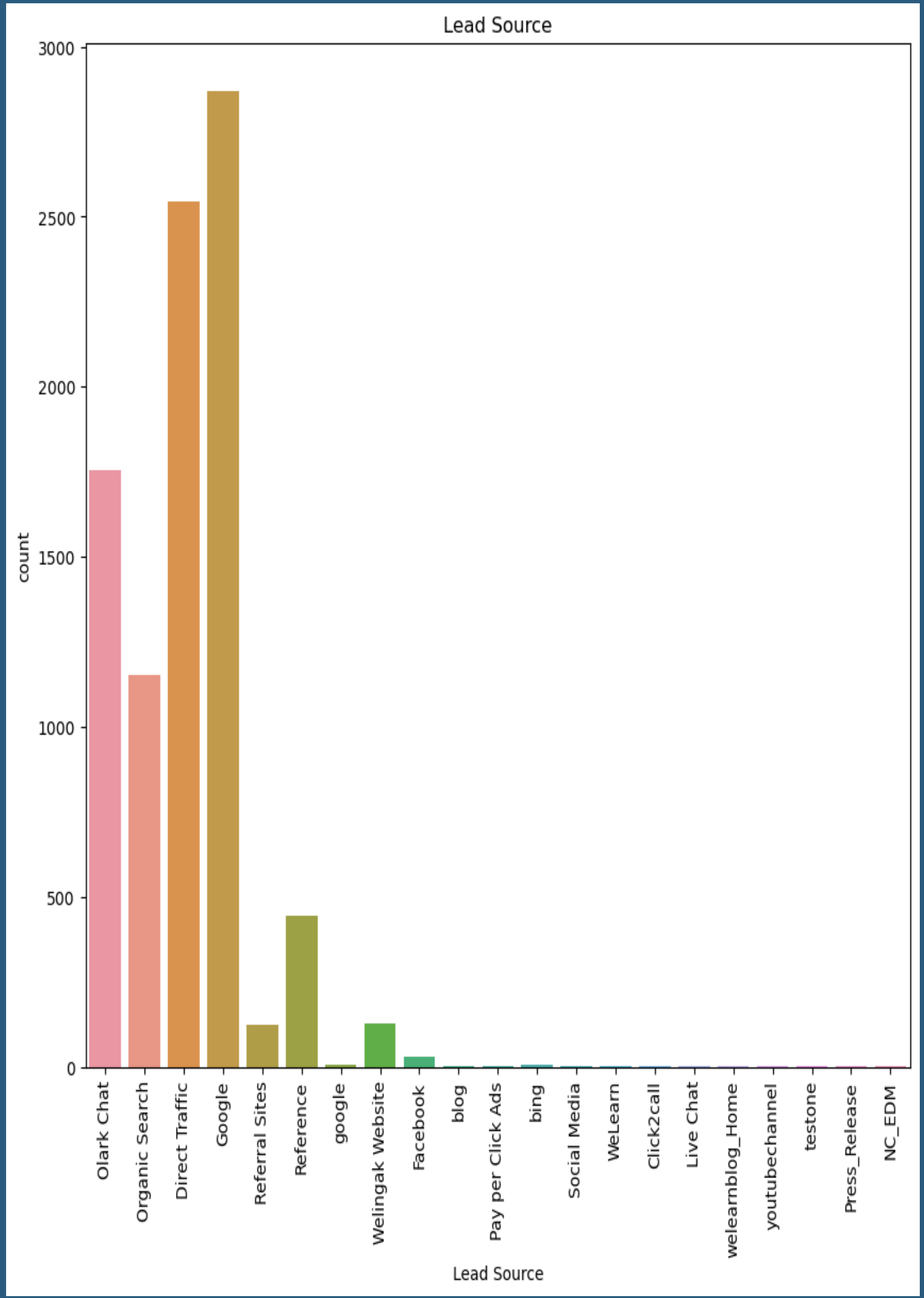
- We import all the necessary libraries for e.g. NumPy , pandas, matplotlib seaborn etc., import all the warnings.
- We read the data and check the no. of rows and columns. We also, check if there are any missing/ null values or not. Afterwards, we see the statistical summary of the data.

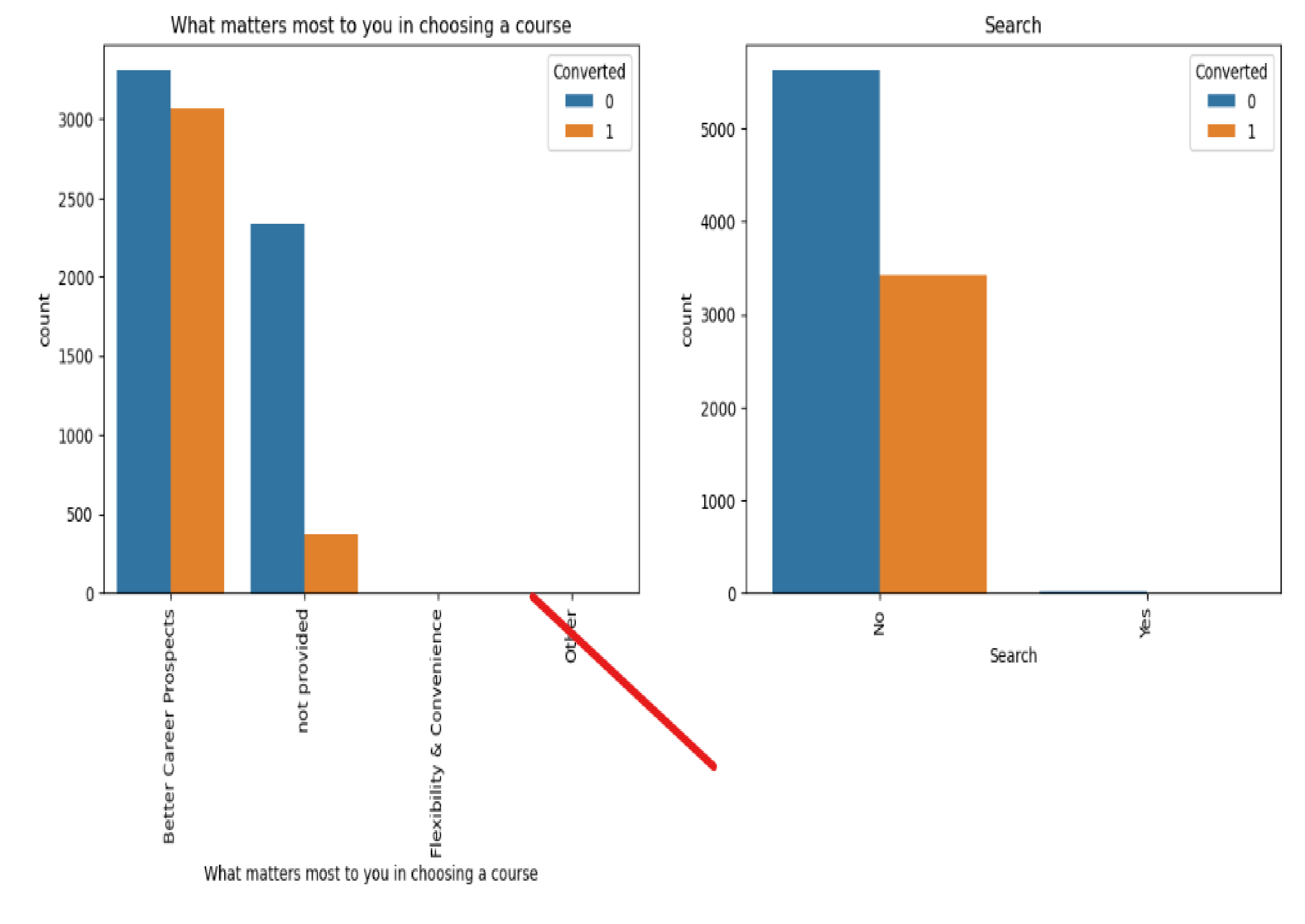
DATA CLEANING

- We saw there were few columns with high percentage of null values, so we decided to drop those columns.
- few columns had null values, but column's were important for analysis so all null values repaced with ‘not provided’.
- Few of the columns had values as ‘Select’ so we replaced it by ‘NaN’.
- Few columns were having outliers and treatment of outlier was performed.

DATA VISUALIZATION USING EDA

- For data visualization we performed exploratory data analysis (EDA). We first did the univariate analysis of categorical and continuous data.
- Quick check was done on % of null value and we dropped columns with more than 45% missing values.
- Moved ahead with the bi-variate analysis of categorical and continuous data.
- We also found in Lead Source data most of the records were from Google and few were from social media platforms and we classified as same.
- We found correlations between variables by using different plots.

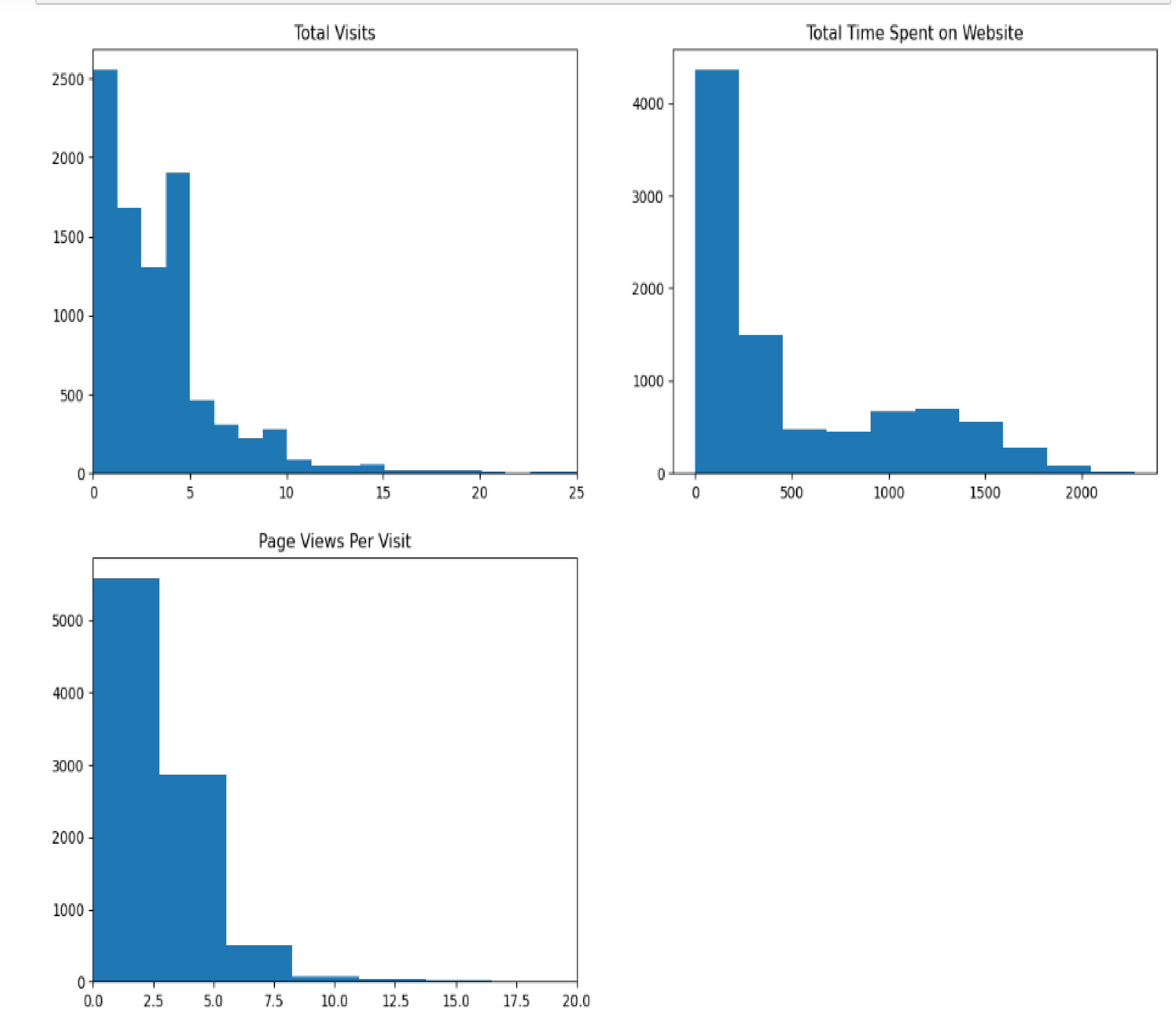




- ❑From this we inferred that candidates focus on better career prospectus while choosing a course.

- ❑Very less number of leads saw the advertisement regarding the course which shows that less visibility of the website.

From this data visualisation graph we can infer that more than 2500 candidates visit the website but spend around two minutes and visit only two pages. Therefore, website should be improved and made more enthusiastic so that candidates spend more time on the website and see maximum pages possible.



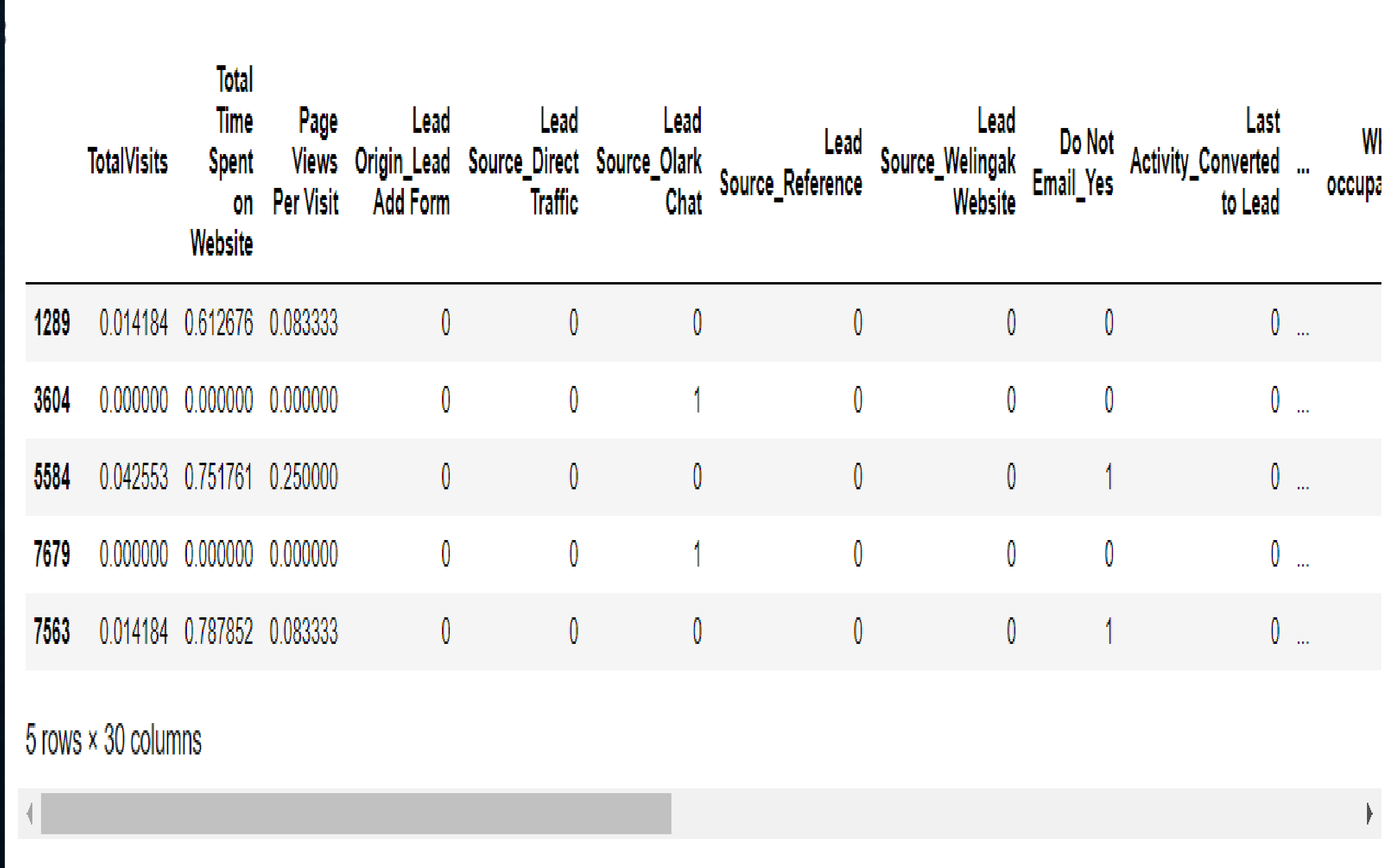
DATA PREPARATION FOR MODELLING

- Data preparation for multiple linear regression involves handling the categorical variables first and then performing dummy encoding.
- We then performed the train and test split using 70%-30% rule and then performed the scaling of variables. Since scaling of variables is an important step, we may have different variables of different scales. So, it's important to have everything on the same scale for the model to be easily interpretable.
- Therefore, we used **MinMaxScaler** for the same.

MODEL BUILDING

- We follow the bottom up approach for this, i.e. we start by building the model with just one variable. Hence, the choice of variables becomes very crucial.
- RFE was done to attain the top 15 relevant variables. First, we will look at the significance of variables and based on the significance.
- We checked VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept). In our model for all the feature VIF < 5 and p-value < 0.05.

FEATURE SELECTION USING RFE

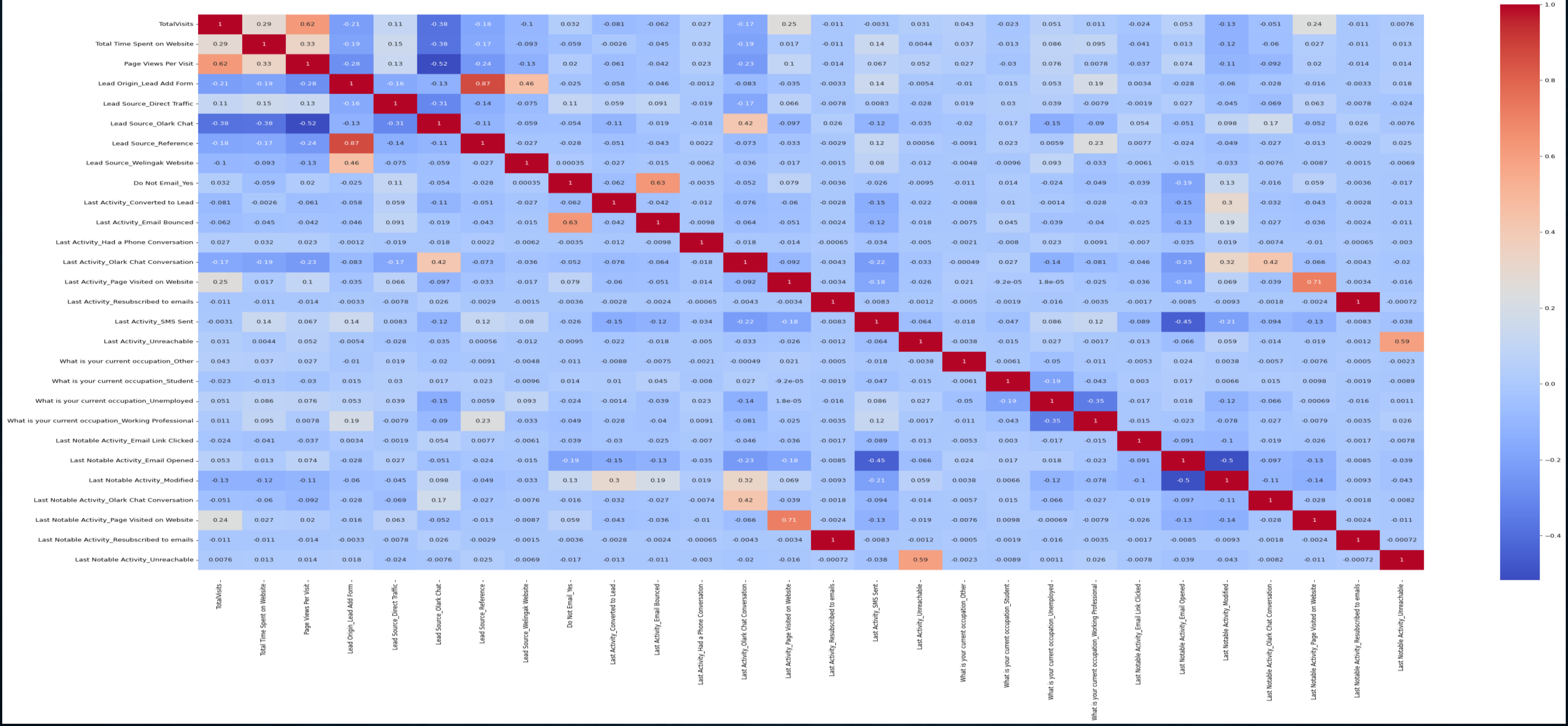


	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_Lead Add Form	Lead Source_Direct Traffic	Lead Source_Olark Chat	Lead Source_Reference	Lead Source_Welingak Website	Do Not Email_Yes	Last Activity_Converted to Lead	...	Wl occupa
1289	0.014184	0.612676	0.083333	0	0	0	0	0	0	0	...	
3604	0.000000	0.000000	0.000000	0	0	1	0	0	0	0	...	
5584	0.042553	0.751761	0.250000	0	0	0	0	0	1	0	...	
7679	0.000000	0.000000	0.000000	0	0	1	0	0	0	0	...	
7563	0.014184	0.787852	0.083333	0	0	0	0	0	1	0	...	

5 rows × 30 columns

- Recursive feature elimination is an optimization technique for finding the best performing subset of features.
- It is based on the idea of repeatedly constructing a model and choosing either the best (based on coefficients), setting the feature aside and then repeating the process with the rest of the features.
- This process is applied until all the features in the dataset are exhausted. Features are then ranked according to their time of elimination.

TO CHECK CORRELATION AMONG VARIABLES



	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_Lead Add Form	Lead Source_Direct Traffic	Lead Source_Olark Chat	Lead Source_Reference	Lead Source_Welingak Website	Do Not Email_Yes	Last Activity_Converted to Lead	Last Activity_Email Bounced	Last Activity_Had a Phone Conversation	Last Activity_Olark Chat Conversation	Last Activity_Page Visited on Website	Last Activity_Resubscribed to emails	Last Activity_SMS Sent	Last Activity_Unreachable	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	Last Notable Activity_Email Link Clicked	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_Resubscribed to emails	Last Notable Activity_Unreachable
TotalVisits	1	0.29	0.62	-0.21	0.11	-0.38	-0.18	-0.1	0.032	-0.081	-0.062	0.027	-0.17	0.25	-0.011	-0.0031	0.031	0.043	-0.023	0.051	0.011	-0.024	0.053	-0.13	-0.051	0.24	-0.011	0.0076
Total Time Spent on Website	0.29	1	0.33	-0.19	0.15	-0.38	-0.17	-0.093	-0.059	-0.0026	-0.045	0.032	-0.19	0.017	-0.011	0.14	0.0044	0.037	-0.013	0.086	0.095	-0.041	0.013	-0.12	-0.06	0.027	-0.011	0.013
Page Views Per Visit	0.62	0.33	1	-0.28	0.13	-0.52	-0.24	-0.13	0.02	-0.061	-0.042	0.023	-0.23	0.1	-0.014	0.067	0.052	0.027	-0.03	0.076	0.0078	-0.037	0.074	-0.11	-0.092	0.02	-0.014	0.014
Lead Origin_Lead Add Form	-0.21	-0.19	-0.28	1	-0.16	-0.13	0.87	0.46	-0.025	-0.058	-0.046	-0.0012	-0.083	-0.035	-0.0033	0.14	-0.0054	-0.01	0.015	0.053	0.19	0.0034	-0.028	-0.06	-0.028	-0.016	-0.0033	0.018
Lead Source_Direct Traffic	0.11	0.15	0.13	-0.16	1	-0.31	-0.14	-0.075	0.11	0.059	0.091	-0.019	-0.17	0.066	-0.0078	0.0083	-0.028	0.019	0.03	0.039	-0.0079	-0.0019	0.027	-0.045	-0.069	0.063	-0.0078	-0.024
Lead Source_Olark Chat	-0.38	-0.38	-0.52	-0.13	-0.31	1	-0.11	-0.059	-0.054	-0.11	-0.019	-0.018	0.42	-0.097	0.026	-0.12	-0.035	-0.02	0.017	-0.15	-0.09	0.054	-0.051	0.098	0.17	-0.052	0.026	-0.0076
Lead Source_Reference	-0.18	-0.17	-0.24	0.87	-0.14	-0.11	1	-0.027	-0.028	-0.051	-0.043	0.0022	-0.073	-0.033	-0.0029	0.12	0.00056	-0.0091	0.023	0.0059	0.23	0.0077	-0.024	-0.049	-0.027	-0.013	-0.0029	0.025
Lead Source_Welingak Website	-0.1	-0.093	-0.13	0.46	-0.075	-0.059	-0.027	1	0.00035	-0.027	-0.015	-0.0062	-0.036	-0.017	-0.0015	0.08	-0.012	-0.0048	-0.0096	0.093	-0.033	-0.0061	-0.015	-0.033	-0.0076	-0.0087	-0.0015	-0.0069
Do Not Email_Yes	0.032	-0.059	0.02	-0.025	0.11	-0.054	-0.028	0.00035	1	-0.062	0.63	-0.0035	-0.052	0.079	-0.0036	-0.026	-0.0095	-0.011	0.014	-0.024	-0.049	-0.039	-0.19	0.13	-0.016	0.059	-0.0036	-0.017
Last Activity_Converted to Lead	-0.081	-0.0026	-0.061	-0.058	0.059	-0.11	-0.051	-0.027	-0.062	1	-0.042	-0.012	-0.076	-0.06	-0.0028	-0.15	-0.022	-0.0088	0.01	-0.0014	-0.028	-0.03	-0.15	0.3	-0.032	-0.043	-0.0028	-0.013
Last Activity_Email Bounced	-0.062	-0.045	-0.042	-0.046	0.091	-0.019	-0.043	-0.015	0.63	-0.042	1	-0.0098	-0.064	-0.051	-0.0024	-0.12	-0.018	-0.0075	0.045	-0.039	-0.04	-0.025	-0.13	0.19	-0.027	-0.036	-0.0024	-0.011
Last Activity_Had a Phone Conversation	0.027	0.032	0.023	-0.0012	-0.019	-0.018	0.0022	-0.0062	-0.0035	-0.012	-0.0098	1	-0.018	-0.014	-0.00065	-0.034	-0.005	-0.0021	-0.008	0.023	0.0091	-0.007	-0.035	0.019	-0.0074	-0.01	-0.00065	-0.003
Last Activity_Olark Chat Conversation	-0.17	-0.19	-0.23	-0.083	-0.17	0.42	-0.073	-0.036	-0.052	-0.076	-0.064	-0.018	1	-0.092	-0.0043	-0.22	-0.033	-0.00049	0.027	-0.14	-0.081	-0.046	-0.23	0.32	0.42	-0.066	-0.0043	-0.02
Last Activity_Page Visited on Website	0.25	0.017	0.1	-0.035	0.066	-0.097	-0.033	-0.017	0.079	-0.06	-0.051	-0.014	-0.092	1	-0.0034	-0.18	-0.026	0.021	-9.2e-05	1.8e-05	-0.025	-0.036	-0.18	0.069	-0.039	0.71	-0.0034	-0.016
Last Activity_Resubscribed to emails	-0.011	-0.011	-0.014	-0.0033	-0.0078	0.026	-0.0029	-0.0015	-0.0036	-0.0028	-0.0024	-0.00065	-0.0043	-0.0034	1	-0.0083	-0.0012	-0.0005	-0.0019	-0.016	-0.0035	-0.0017	-0.0085	-0.0093	-0.0018	-0.0024	1	-0.00072
Last Activity_SMS Sent	-0.0031	0.14	0.067	0.14	0.0083	-0.12	0.12	0.08	-0.026	-0.15	-0.12	-0.034	-0.22	-0.18	-0.0083	1	-0.064	-0.018	-0.047	0.086	0.12	-0.089	-0.45	-0.21	-0.094	-0.13	-0.0083	-0.038
Last Activity_Unreachable	0.031	0.0044	0.052	-0.0054	-0.028	-0.035	0.00056	-0.012	-0.0095	-0.022	-0.018	-0.005	-0.033	-0.026	-0.0012	-0.064	1	-0.0038	-0.015	0.027	-0.0017	-0.013	-0.066	0.059	-0.014	-0.019	-0.0012	0.59
What is your current occupation_Other	0.043	0.037	0.027	-0.01	0.019	-0.02	-0.0091	-0.0048	-0.011	-0.0088	-0.0075	-0.0021	-0.00049	0.021	-0.0005	-0.018	-0.0038	1	-0.0061	-0.05	-0.011	-0.0053	0.024	0.0038	-0.0057	-0.0076	-0.0005	-0.0023
What is your current occupation_Student	-0.023	-0.013	-0.03	0.015	0.03	0.017	0.023	-0.0096	0.014	0.01	0.045	-0.008	0.027	-9.2e-05	-0.0019	-0.047	-0.015	-0.0061	1	-0.19	-0.043	0.003	0.017	0.0066	0.015	0.0098	-0.0019	-0.0089
What is your current occupation_Unemployed	0.051	0.086	0.076	0.053	0.039	-0.15	0.0059	0.093	-0.024	-0.0014	-0.039	0.023	-0.14	1.8e-05	-0.016	0.086	0.027	-0.05	-0.19	1	-0.35	-0.017	0.018	-0.12	-0.066	-0.00069	-0.016	0.0011
What is your current occupation_Working Professional	0.011	0.095	0.0078	0.19	-0.0079	-0.09	0.23	-0.033	-0.049	-0.028	-0.04	0.0091	-0.081	-0.025	-0.0035	0.12	-0.0017	-0.011	-0.043	-0.35	1	-0.015	-0.023	-0.078	-0.027	-0.0079	-0.0035	0.026
Last Notable Activity_Email Link Clicked	-0.024	-0.041	-0.037	0.0034	-0.0019	0.054	0.0077	-0.0061	-0.039	-0.03	-0.025	-0.007	-0.046	-0.036	-0.0017	-0.089	-0.013	-0.0053	0.003	-0.017	-0.015	1	-0.091	-0.1	-0.019	-0.026	-0.0017	-0.0078
Last Notable Activity_Email Opened	0.053	0.013	0.074	-0.028	0.027	-0.051	-0.024	-0.015	-0.19	-0.15	-0.13	-0.035	-0.23	-0.18	-0.0085	-0.45	-0.066	0.024	0.017	0.018	-0.023	-0.091	1	-0.5	-0.097	-0.13	-0.0085	-0.039
Last Notable Activity_Modified	-0.13	-0.12	-0.11	-0.06	-0.045	0.098	-0.049	-0.033	0.13	0.3	0.19	0.019	0.32	0.069	-0.0093	-0.21	0.059	0.0038	0.0066	-0.12	-0.078	-0.1	-0.5	1	-0.11	-0.14	-0.0093	-0.043
Last Notable Activity_Olark Chat Conversation	-0.051	-0.06	-0.092	-0.028	-0.069	0.17	-0.027	-0.0076	-0.016	-0.032	-0.027	-0.0074	0.42	-0.039	-0.0018	-0.094	-0.014	-0.0057	0.015	-0.066	-0.027	-0.019	-0.097	-0.11	1	-0.028	-0.0018	-0.0082
Last Notable Activity_Page Visited on Website	0.24	0.027	0.02	-0.016	0.063	-0.052	-0.013	-0.0087	0.059	-0.043	-0.036	-0.01	-0.066	0.71	-0.0024	-0.13	-0.019	-0.0076	0.0098	-0.00069	-0.0079	-0.026	-0.13	-0.14	-0.028	1	-0.0024	-0.011
Last Notable Activity_Resubscribed to emails	-0.011	-0.011	-0.014	-0.0033	-0.0078	0.026	-0.0029	-0.0015	-0.0036	-0.0028	-0.0024	-0.00065	-0.0043	-0.0034	1	-0.0083	-0.0012	-0.0005	-0.0019	-0.016	-0.0035	-0.0017	-0.0085	-0.0093	-0.0018	-0.0024	1	-0.00072
Last Notable Activity_Unreachable	0.0076	0.013	0.014	0.018	-0.024	-0.0076	0.025	-0.0069	-0.017	-0.013	-0.011	-0.003	-0.02	-0.016	-0.00072	-0.038	0.59	-0.0023	-0.0089	0.0011	0.026	-0.0078	-0.039	-0.043	-0.0082	-0.011	-0.00072	1

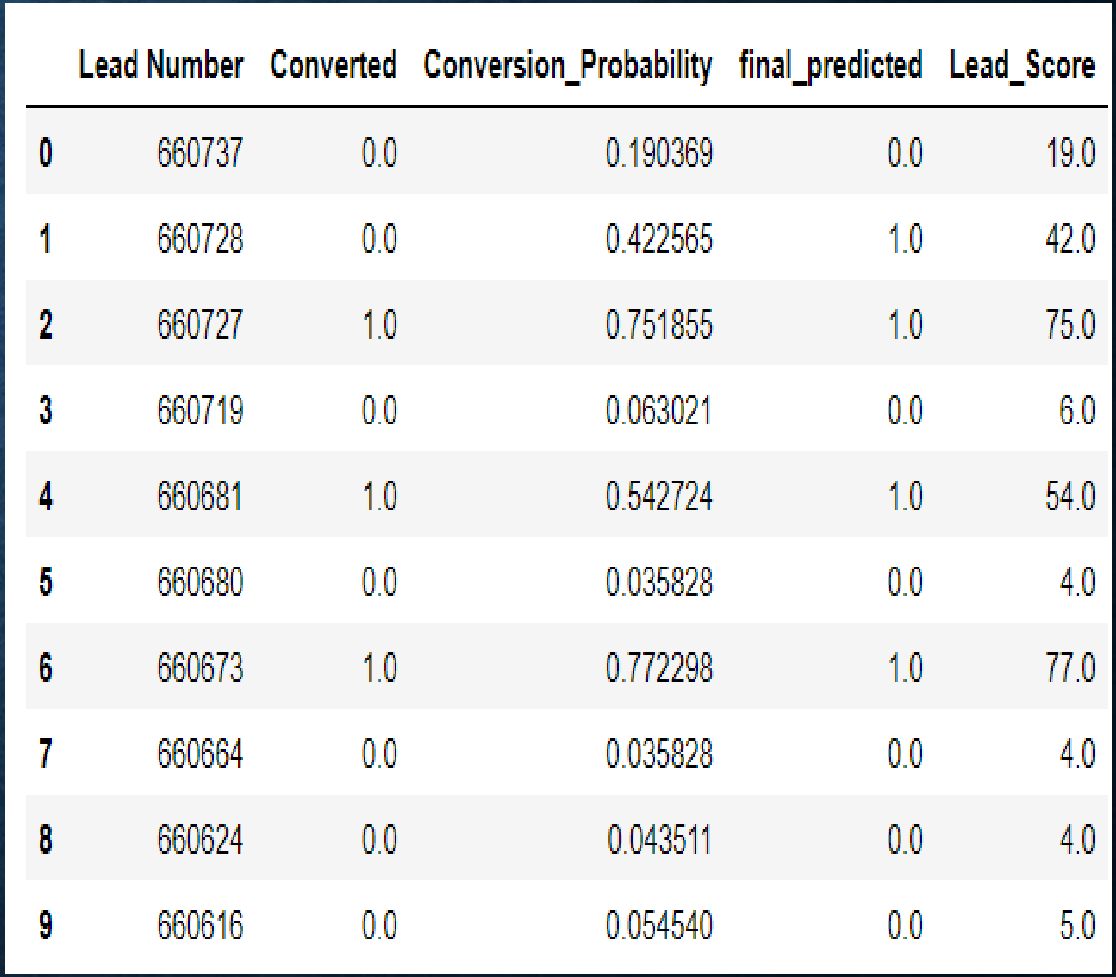
LEAD SCORE CALCULATION

- Lead score is calculated for all the leads in the original data frame.

- Formula for Lead Score calculation is
  Lead Score = 100* Conversion Probability.

- Higher the lead score, higher is the probability of a lead getting converted and vice versa.

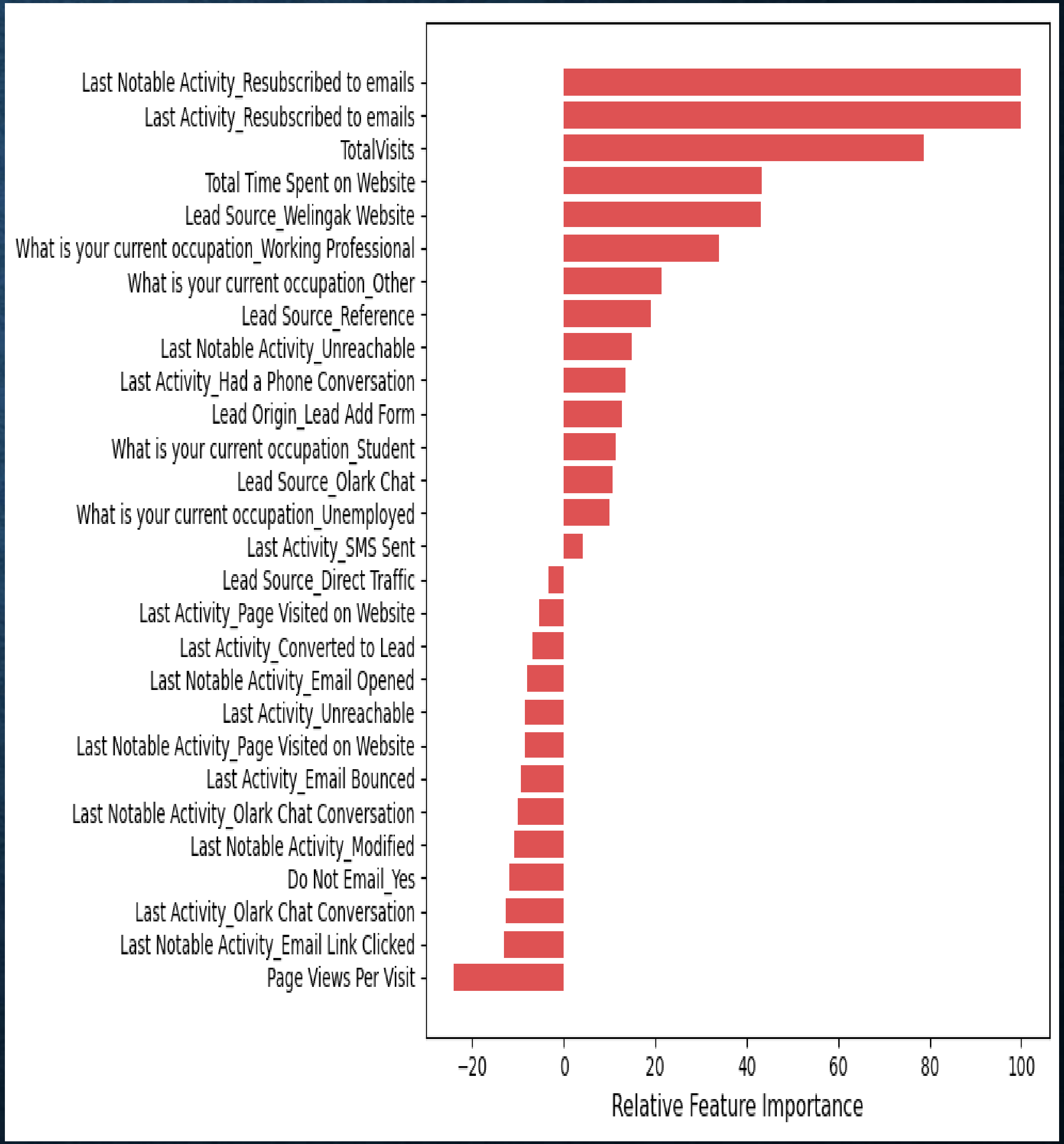
- Since, we had used 0.42 as our final probability threshold for deciding if a lead will convert or not, any lead with a lead score of 34 or above will have a value of ‘1’ in the final predicted column.



	Lead Number	Converted	Conversion_Probability	final_predicted	Lead_Score
0	660737	0.0	0.190369	0.0	19.0
1	660728	0.0	0.422565	1.0	42.0
2	660727	1.0	0.751855	1.0	75.0
3	660719	0.0	0.063021	0.0	6.0
4	660681	1.0	0.542724	1.0	54.0
5	660680	0.0	0.035828	0.0	4.0
6	660673	1.0	0.772298	1.0	77.0
7	660664	0.0	0.035828	0.0	4.0
8	660624	0.0	0.043511	0.0	4.0
9	660616	0.0	0.054540	0.0	5.0

FEATURE DETERMINATION

- The Coefficient (beta) values for each of these features from the model parameters are used to determine the order of importance of these features.
- Features with high positive beta values are the ones that contribute most towards the probability of a lead getting converted.
- Similarly, features with high negative beta values contribute the least



MODEL EVALUATION

- We first created confusion matrix to find the TP, TN, FP and FN. We calculated the sensitivity and specificity.
- We plotted the graph of sensitivity, accuracy and specificity for each level of probability. We found that 0.42 was the cut-off point.
- We then used the cut-off point 0.42 to select the person and see if he would be converted or not then ,
- We again created the confusion matrix to calculate TP, TN,FP & FN calculated the sensitivity and specificity and got 0.84984 and 0.78669 respectively.

Thank You