Summary Report

Problem Statement -

Our main goal is to create a logistic regression model to better identify the most potential leads so that the efforts of the sales team can be focused towards contacting the potential leads.

The company needs the target lead conversion rate to be around 80%.

Summary to the modelling process

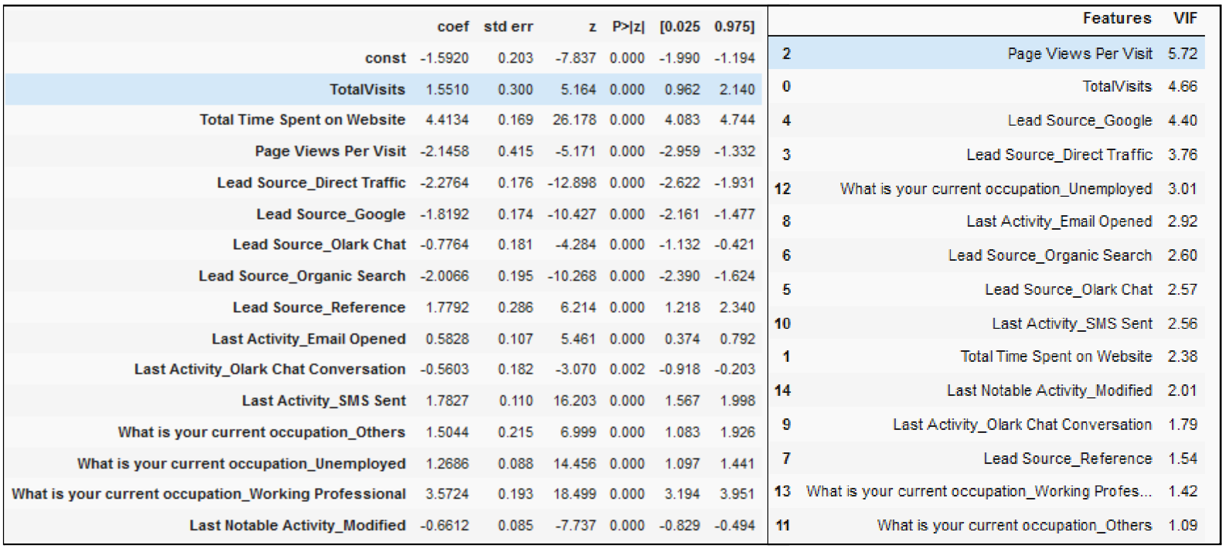
- **Basic analysis.**
    - The data provided to us by the company is a leads dataset with around 9240 data points and 37 attributes.
    - Important attributes of the data like target variable **Converted** were identified,
    - Datatypes of the columns were also checked to get a better understanding of the underlying data.

- **Data Handling and Cleanup.**
    - Null values are detected and processed, columns with more than 36% null values were removed.
    - The 'select' label was detected in many columns indicating that data was not provided by the customer. 'Select' labels were replaced with 'Not Specified' - to preserve data.
    - Remaining attributes in the dataframe are carefully analysed for type & distribution of data and then appropriately imputed.
    - Columns with highly skewed data or with only 1 unique value were removed to make the data insightful.
    - Few sub-categories of attributes were combined to draw meaningful insights from data.
    - Outliers are detected and are treated accordingly to prevent data skewness.

- **Data Visualisation and EDA**
    - Analysis is done on the numerical & categorical variables and inferences & observations are noted and specified.
    - Inferences allow us to get a better understanding of trends and overall behaviour of the customers.

- **Pre- Modelling Data Preparation and Feature Scaling.**
  - For categorical variables having sub-categories,one-hot encoded dummy variables are created.
  - Data is split into Training and Test datasets in a ratio of 70:30.
  - Feature scaling is done using the Min-Max-scaler.
  - Using heatmaps & correlation-matrix, the correlation between predictor variables are analysed and few were dropped too.

- **Model Building**
  - Recursive Feature Elimination is used to detect the 15 best variables for the logistic regression model.

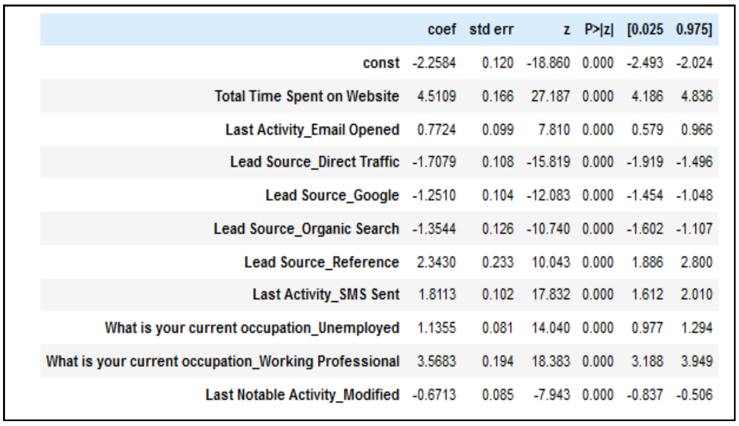


	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.5920	0.203	-7.837	0.000	-1.990	-1.194
TotalVisits	1.5510	0.300	5.164	0.000	0.962	2.140
Total Time Spent on Website	4.4134	0.169	26.178	0.000	4.083	4.744
Page Views Per Visit	-2.1458	0.415	-5.171	0.000	-2.959	-1.332
Lead Source_Direct Traffic	-2.2764	0.176	-12.898	0.000	-2.622	-1.931
Lead Source_Google	-1.8192	0.174	-10.427	0.000	-2.161	-1.477
Lead Source_Olark Chat	-0.7764	0.181	-4.284	0.000	-1.132	-0.421
Lead Source_Organic Search	-2.0066	0.195	-10.268	0.000	-2.390	-1.624
Lead Source_Reference	1.7792	0.286	6.214	0.000	1.218	2.340
Last Activity_Email Opened	0.5828	0.107	5.461	0.000	0.374	0.792
Last Activity_Olark Chat Conversation	-0.5603	0.182	-3.070	0.002	-0.918	-0.203
Last Activity_SMS Sent	1.7827	0.110	16.203	0.000	1.567	1.998
What is your current occupation_Others	1.5044	0.215	6.999	0.000	1.083	1.926
What is your current occupation_Unemployed	1.2686	0.088	14.456	0.000	1.097	1.441
What is your current occupation_Working Professional	3.5724	0.193	18.499	0.000	3.194	3.951
Last Notable Activity_Modified	-0.6612	0.085	-7.737	0.000	-0.829	-0.494

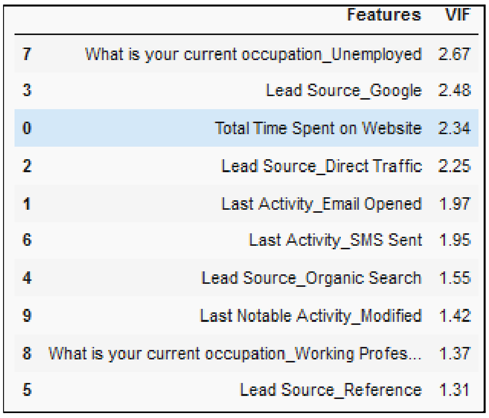
	Features	VIF
2	Page Views Per Visit	5.72
0	TotalVisits	4.66
4	Lead Source_Google	4.40
3	Lead Source_Direct Traffic	3.76
12	What is your current occupation_Unemployed	3.01
8	Last Activity_Email Opened	2.92
6	Lead Source_Organic Search	2.60
5	Lead Source_Olark Chat	2.57
10	Last Activity_SMS Sent	2.56
1	Total Time Spent on Website	2.38
14	Last Notable Activity_Modified	2.01
9	Last Activity_Olark Chat Conversation	1.79
7	Lead Source_Reference	1.54
13	What is your current occupation_Working Profes...	1.42
11	What is your current occupation_Others	1.09

  - Manual Elimination:- Summary and VIF is generated to detect insignificant and highly correlated data. Such variables are carefully removed one at a time while monitoring the p-values and VIF values.

  - The model is finalised when p-value < 0.05 & VIF's are within acceptable range i.e. below 5.



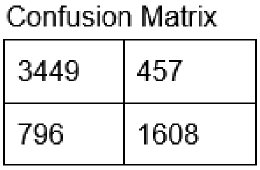
	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.2584	0.120	-18.860	0.000	-2.493	-2.024
Total Time Spent on Website	4.5109	0.166	27.187	0.000	4.186	4.836
Last Activity_Email Opened	0.7724	0.099	7.810	0.000	0.579	0.966
Lead Source_Direct Traffic	-1.7079	0.108	-15.819	0.000	-1.919	-1.496
Lead Source_Google	-1.2510	0.104	-12.083	0.000	-1.454	-1.048
Lead Source_Organic Search	-1.3544	0.126	-10.740	0.000	-1.602	-1.107
Lead Source_Reference	2.3430	0.233	10.043	0.000	1.886	2.800
Last Activity_SMS Sent	1.8113	0.102	17.832	0.000	1.612	2.010
What is your current occupation_Unemployed	1.1355	0.081	14.040	0.000	0.977	1.294
What is your current occupation_Working Professional	3.5683	0.194	18.383	0.000	3.188	3.949
Last Notable Activity_Modified	-0.6713	0.085	-7.943	0.000	-0.837	-0.506



	Features	VIF
7	What is your current occupation_Unemployed	2.67
3	Lead Source_Google	2.48
0	Total Time Spent on Website	2.34
2	Lead Source_Direct Traffic	2.25
1	Last Activity_Email Opened	1.97
6	Last Activity_SMS Sent	1.95
4	Lead Source_Organic Search	1.55
9	Last Notable Activity_Modified	1.42
8	What is your current occupation_Working Profes...	1.37
5	Lead Source_Reference	1.31

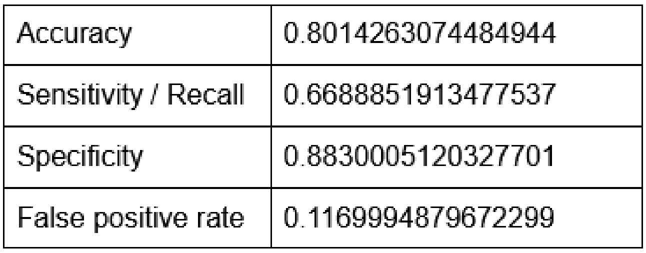
- **Prediction and Evaluation Metrics of Train and Test Data.**

  - **Predictions are done on the Training data**
  - Confusion matrix is generated for the data and evaluation metrics are calculated.
    - Cutoff = 0.5



Confusion Matrix

3449	457
796	1608

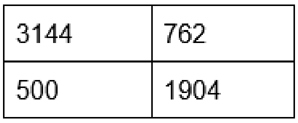


Accuracy	0.8014263074484944
Sensitivity / Recall	0.6688851913477537
Specificity	0.8830005120327701
False positive rate	0.1169994879672299

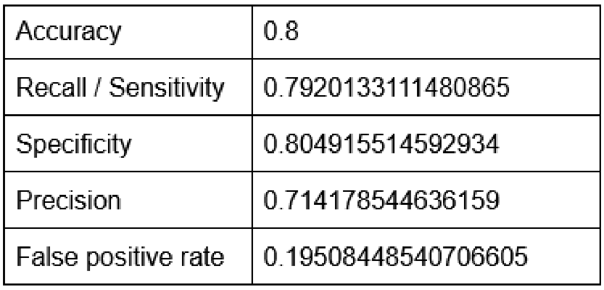
  - ROC Curve area = 0.88

    - Cutoff = 0.40

Confusion matrix

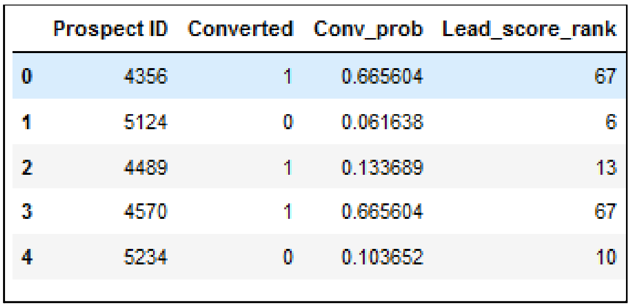


3144	762
500	1904



Accuracy	0.8
Recall / Sensitivity	0.7920133111480865
Specificity	0.804915514592934
Precision	0.714178544636159
False positive rate	0.19508448540706605

- Lead_score_rank is assigned to the data



	Prospect ID	Converted	Conv_prob	Lead_score_rank
0	4356	1	0.665604	67
1	5124	0	0.061638	6
2	4489	1	0.133689	13
3	4570	1	0.665604	67
4	5234	0	0.103652	10

- **Predictions are done on the Test data** and model performance is checked. All metrics are recalculated for the test data and recorded.
  - Cutoff = 0.40

Confusion matrix

1371	327
197	810

Accuracy	0.8062846580406654
Recall / Sensitivity	0.8043694141012909
Specificity	0.8074204946996466
Precision	0.712401055408971
F-Score	0.7555970149253731

- **Recommendation :-** To identify the most potential leads, the sales team should focus towards the below variables :-
    1. Total Time Spent on Website
    2. What is your current occupation
        a. Working Professional
        b. Unemployed
    3. Lead Source
        a. Reference
        b. Google
    4. Last Activity
        a. SMS Sent
        b. Email Opened

The X education company may achieve its target of lead conversion rate of approx 80% if they focus more on above variables.