X EDUCATION COMPANY

LEAD SCORE CASE STUDY

BY

APARNA KISHAN

JAYASREE RAMAKRISHNAN

MANU GUNDALA

BUSINESS OBJECTIVE :

To help X Education select most promising leads who can be converted to successful customers.

- Company CEO expectation is converting 80% of leads into customers, taking up their product services.
- History of Lead conversion data is given from the sales team with insights.
- Data set can be used to build a model to predict potential leads and provide useful recommendations to focus and attain maximum benefits and working principles to identify the right category of customers, so as to reduce time and cost.

APPROACH :

To build a Logistic Regression Model that assigns a Lead Score between 0 and 100 to all the leads such that the leads with higher score have a higher probability of becoming a successful customer of X Education.

Target Lead Conversion rate is 80%

STEP 1: DATA IMPORT AND CLEANING

- Import and Understand the data set provided by the Sales Team

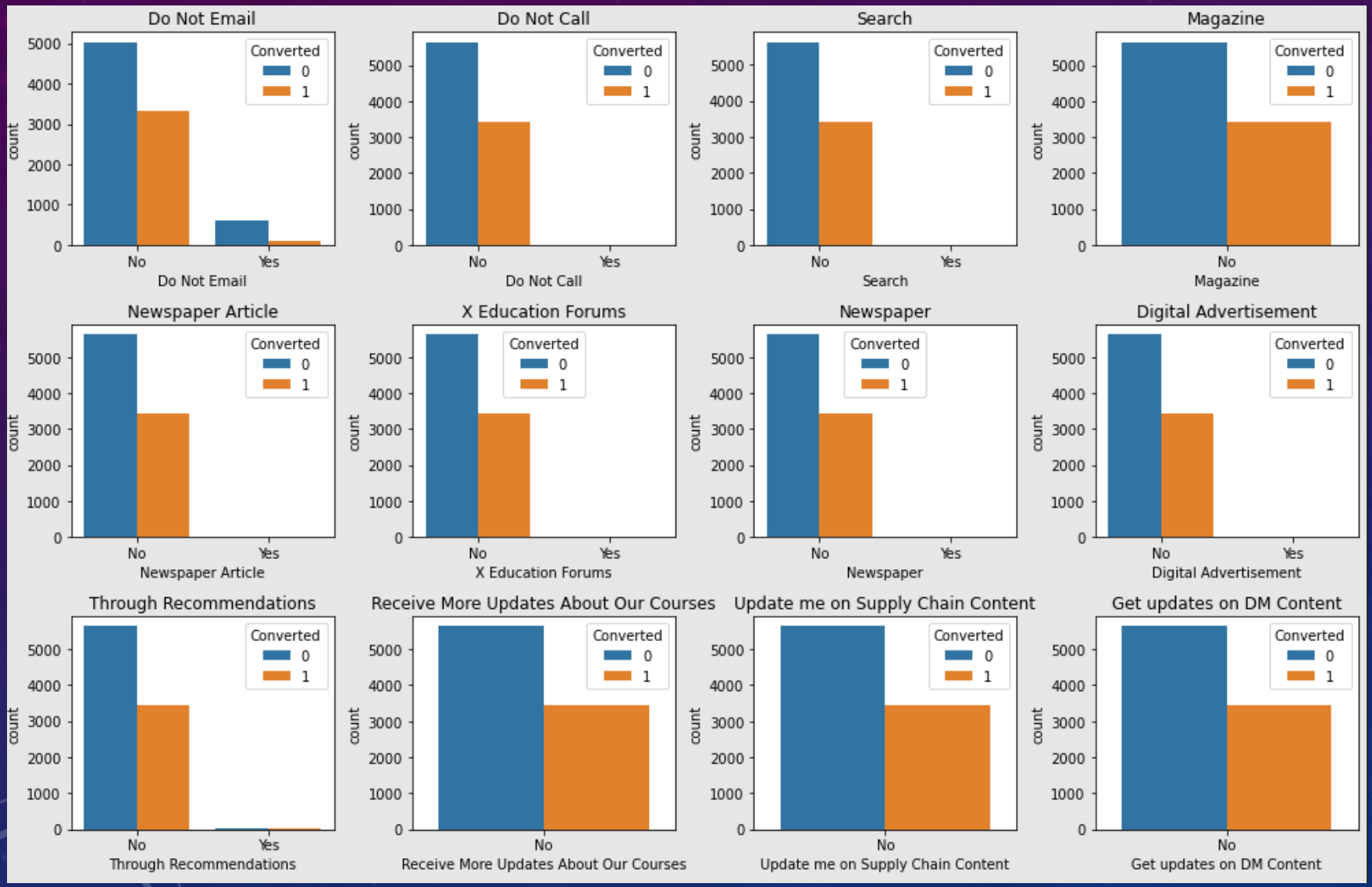
- Data Cleaning
  - ❖ Dropping duplicate rows and unwanted columns
  - ❖ Replacing ‘Select’ with null values
  - ❖ Handling null values

STEP 2: DATA VISUALIZATION AND EXPLORATORY DATA ANALYSIS

- Visualize Categorical variables
  - ❖ Drop variables with low variance or imbalanced data
  - ❖ Group together less frequent categories

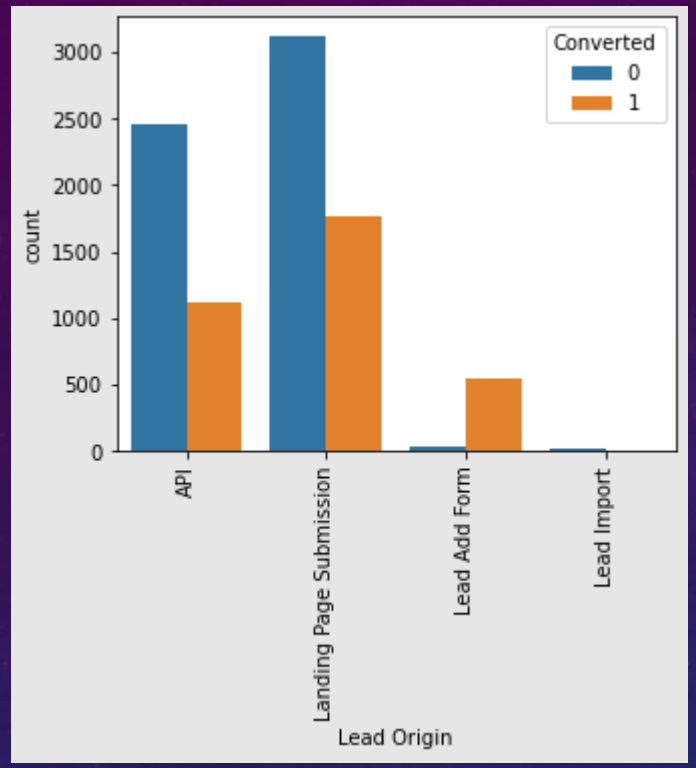
- Visualize Numeric variables
  - ❖ Check for correlations
  - ❖ Handle Outliers

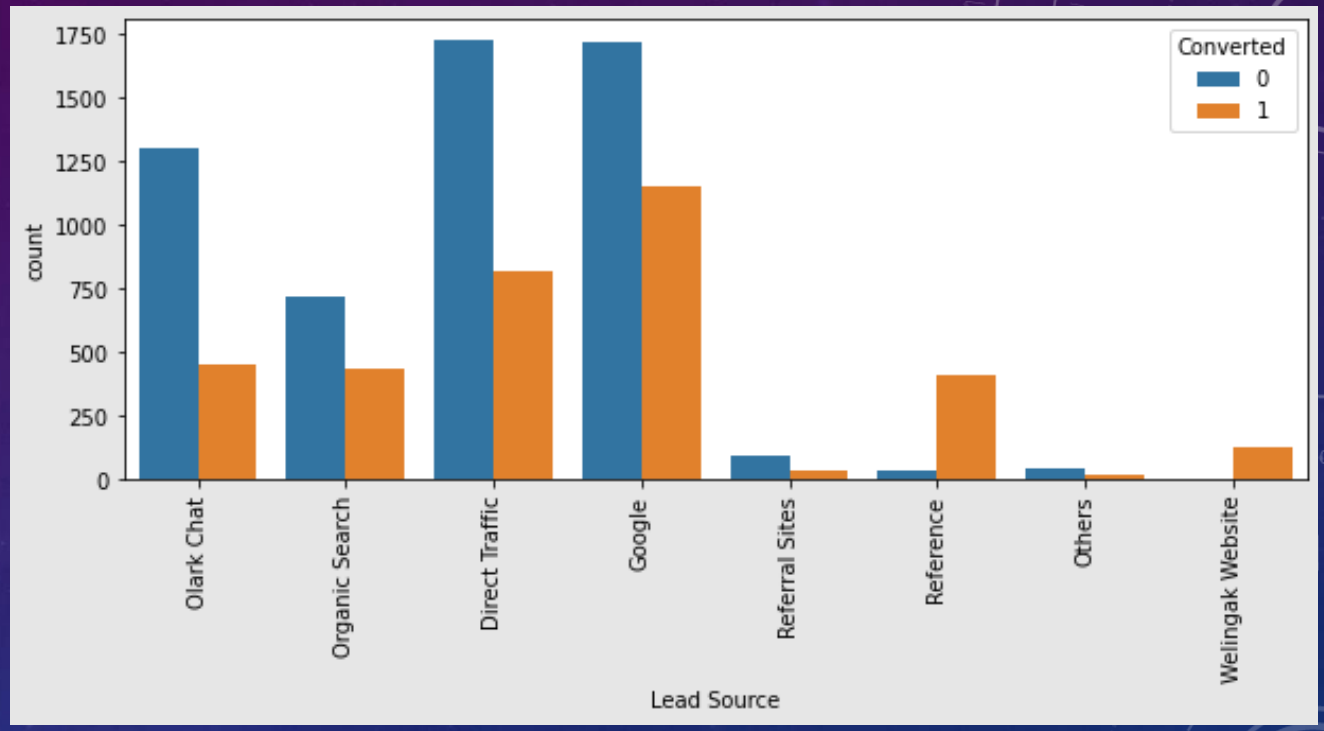
Visualizing Binary Features



The graph shows that many variables have low variance. i.e., majority or all of the values are 'No'.
Since this doesn't provide any insight for our analysis, we dropped such columns.

Visualizing Categorical Features





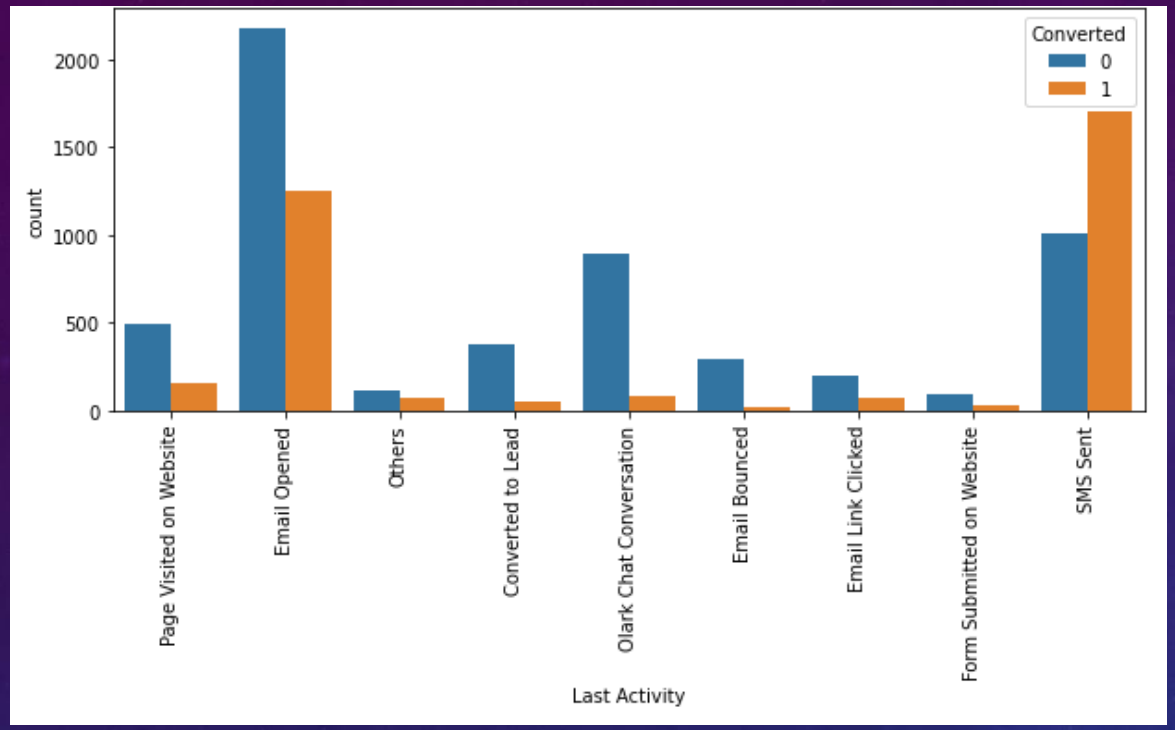
Lead Origin :

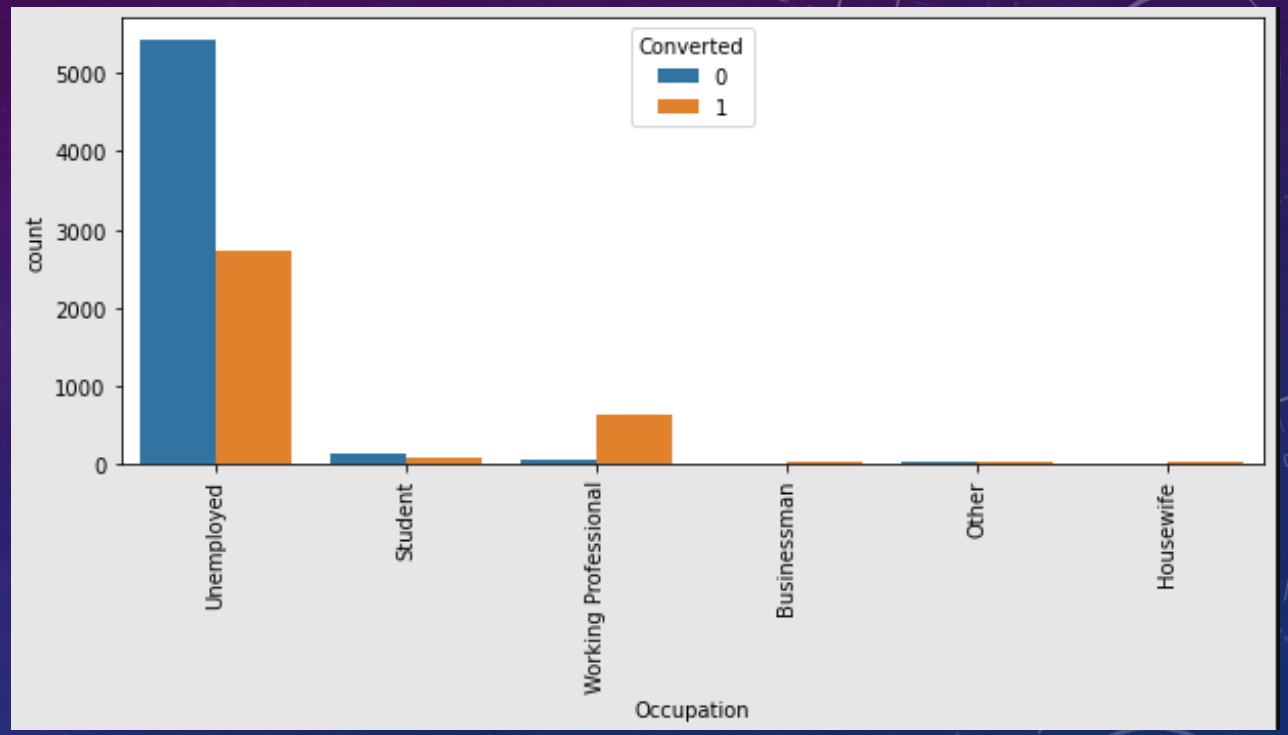
- Landing Page Submissions and API have higher number of conversions.
- Lead Add From has a high conversion rate however the number of leads is less.

Lead Source :

- Direct Traffic and Google have comparatively higher number of conversions.
- Leads obtained by Reference have a very high chance of conversion

Visualizing Categorical Features





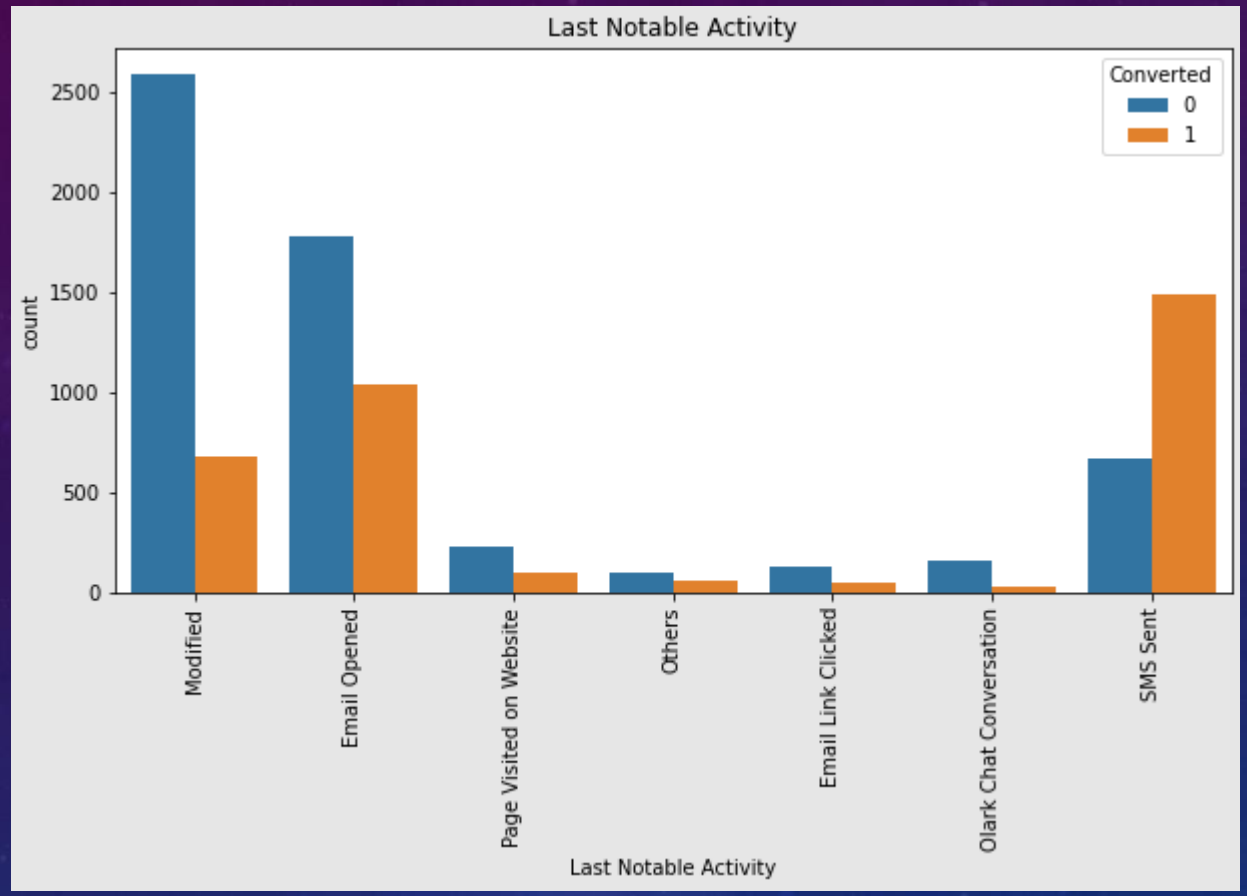
Lead Activity :

- Email opened has the highest number of leads.
- SMS sent indicates a high chance of conversion.
- Olark Chat Conversation has a large number of leads but the conversion rate is too low.

Occupation :

- Leads are high among unemployed people
- Rate of conversion is high among Working Professional

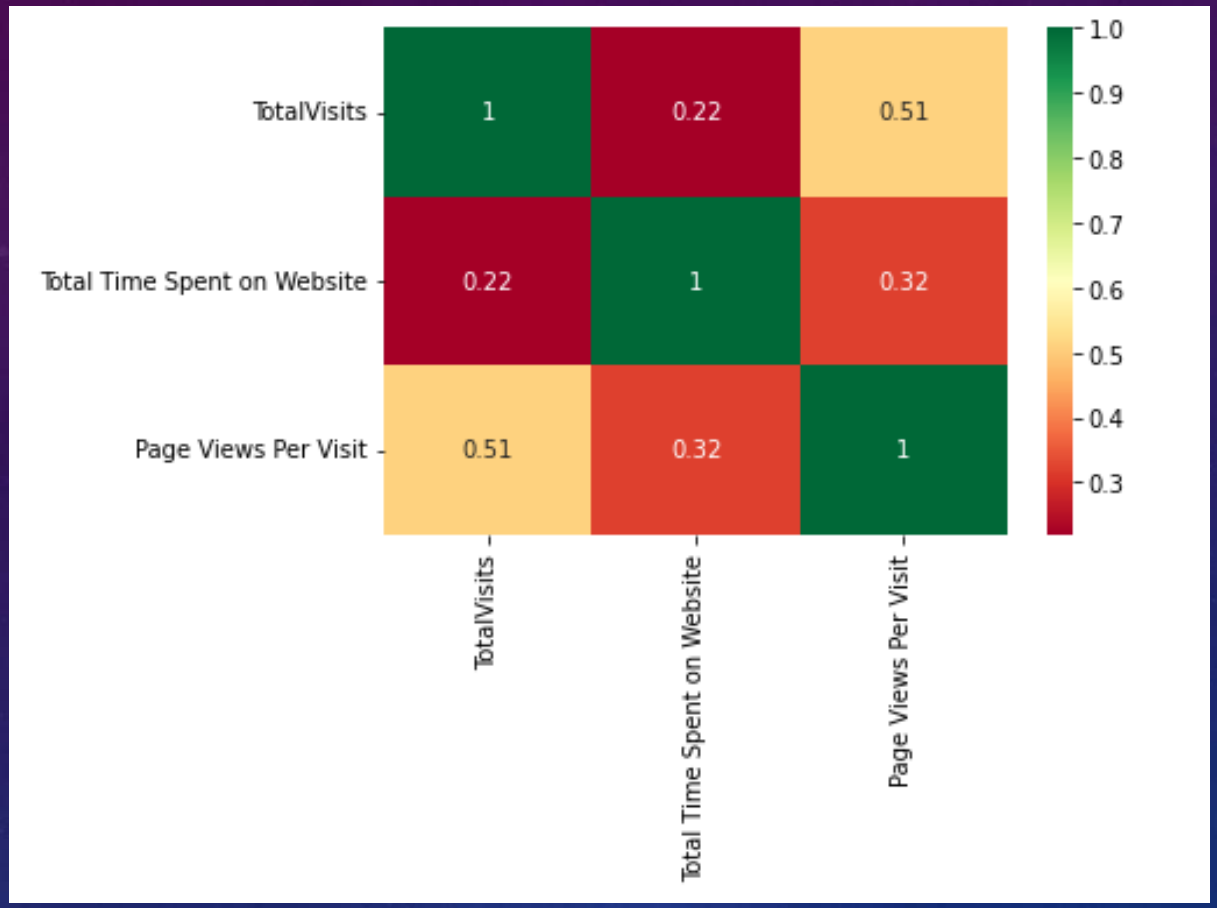
Visualizing Categorical Features



Lead Notable Activity :

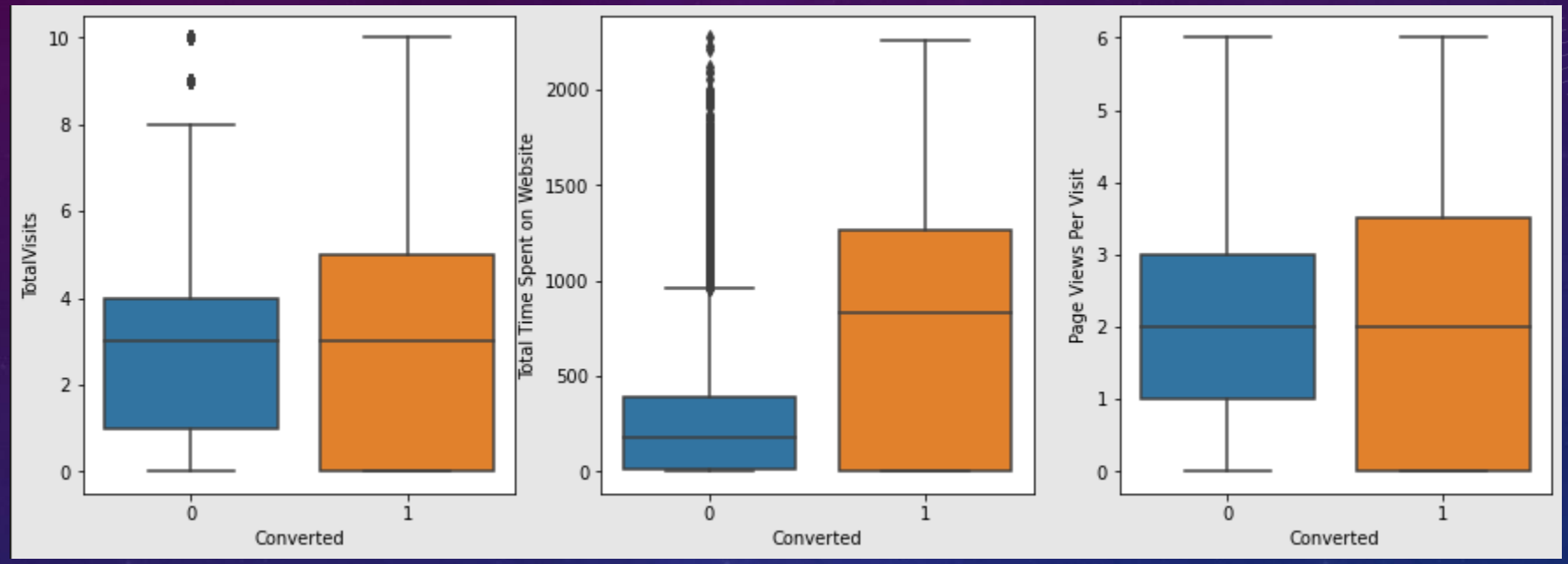
- Modified and Email opened have higher number of Leads
- SMS sent indicates a high chance of conversion

Visualizing Numerical Features



There doesn’t seem to be high correlation between numeric features

Visualizing Numerical Features



People who spent more time on website are more likely to become converted leads

Model Summary

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6337
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2650.8
Date:	Mon, 27 Feb 2023	Deviance:	5301.6
Time:	23:26:15	Pearson chi2:	6.41e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3924
Covariance Type:	nonrobust		

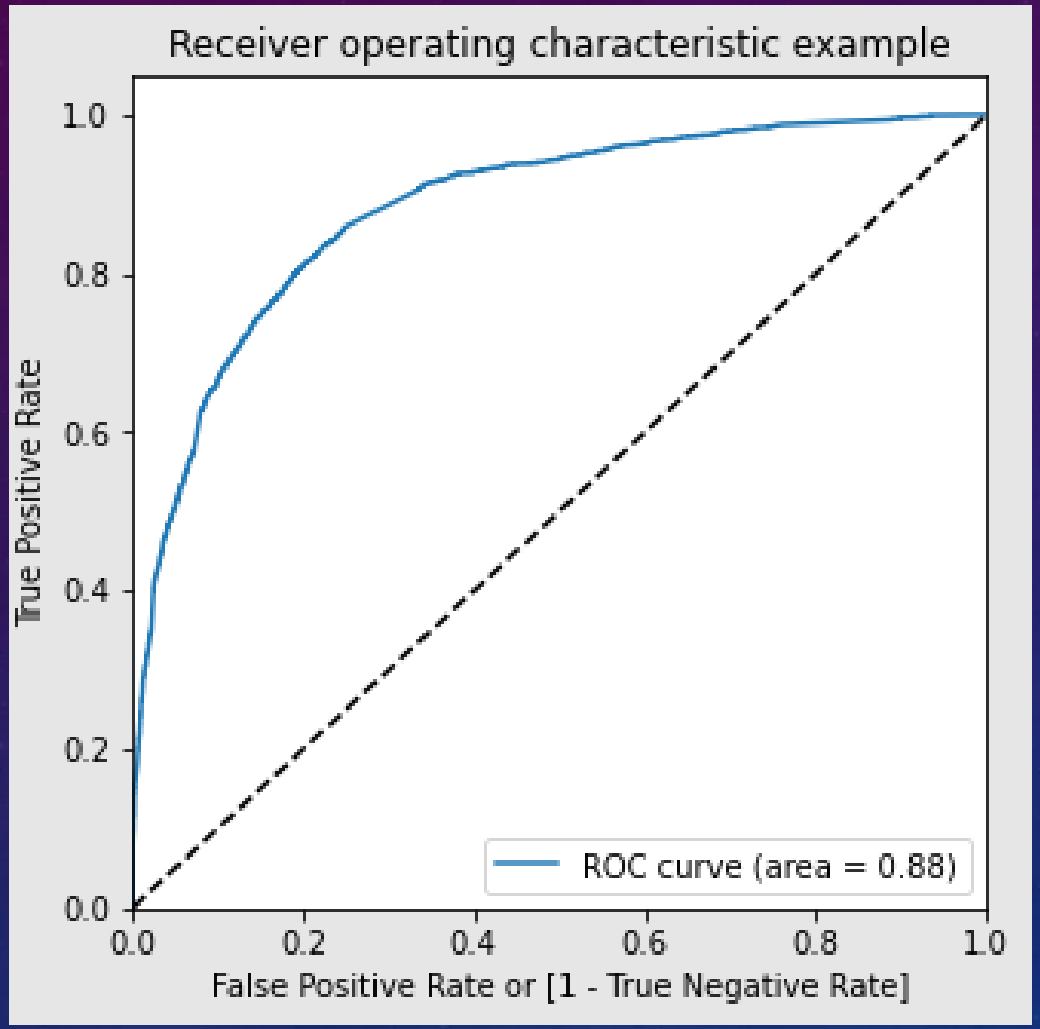
	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.7374	0.099	-27.634	0.000	-2.932	-2.543
Do Not Email	-1.6482	0.206	-7.998	0.000	-2.052	-1.244
TotalVisits	0.8194	0.156	5.239	0.000	0.513	1.126
Total Time Spent on Website	4.5804	0.164	27.885	0.000	4.258	4.902
Lead Source_Olark Chat	1.5861	0.121	13.113	0.000	1.349	1.823
Lead Source_Others	1.3048	0.352	3.704	0.000	0.614	1.995
Lead Source_Reference	4.3403	0.237	18.288	0.000	3.875	4.805
Lead Source_Welingak Website	6.3367	0.730	8.681	0.000	4.906	7.767
Last Activity_Email Bounced	-1.4024	0.447	-3.140	0.002	-2.278	-0.527
Last Activity_Olark Chat Conversation	-1.4669	0.161	-9.092	0.000	-1.783	-1.151
Last Activity_Page Visited on Website	-0.4627	0.148	-3.118	0.002	-0.754	-0.172
Occupation_Working Professional	2.7160	0.187	14.517	0.000	2.349	3.083
Last Notable Activity_Others	1.8453	0.286	6.448	0.000	1.284	2.406
Last Notable Activity_SMS Sent	1.5755	0.081	19.405	0.000	1.416	1.735

VIF Values of features in Final Model

Features	VIF
TotalVisits	2.14
Total Time Spent on Website	1.98
Do Not Email	1.89
Last Activity_Email Bounced	1.78
Last Notable Activity_SMS Sent	1.47
Lead Source_Olark Chat	1.41
Last Activity_Olark Chat Conversation	1.39
Last Activity_Page Visited on Website	1.22
Occupation_Working Professional	1.19
Lead Source_Reference	1.14
Last Notable Activity_Others	1.14
Lead Source_Welingak Website	1.02
Lead Source_Others	1.00

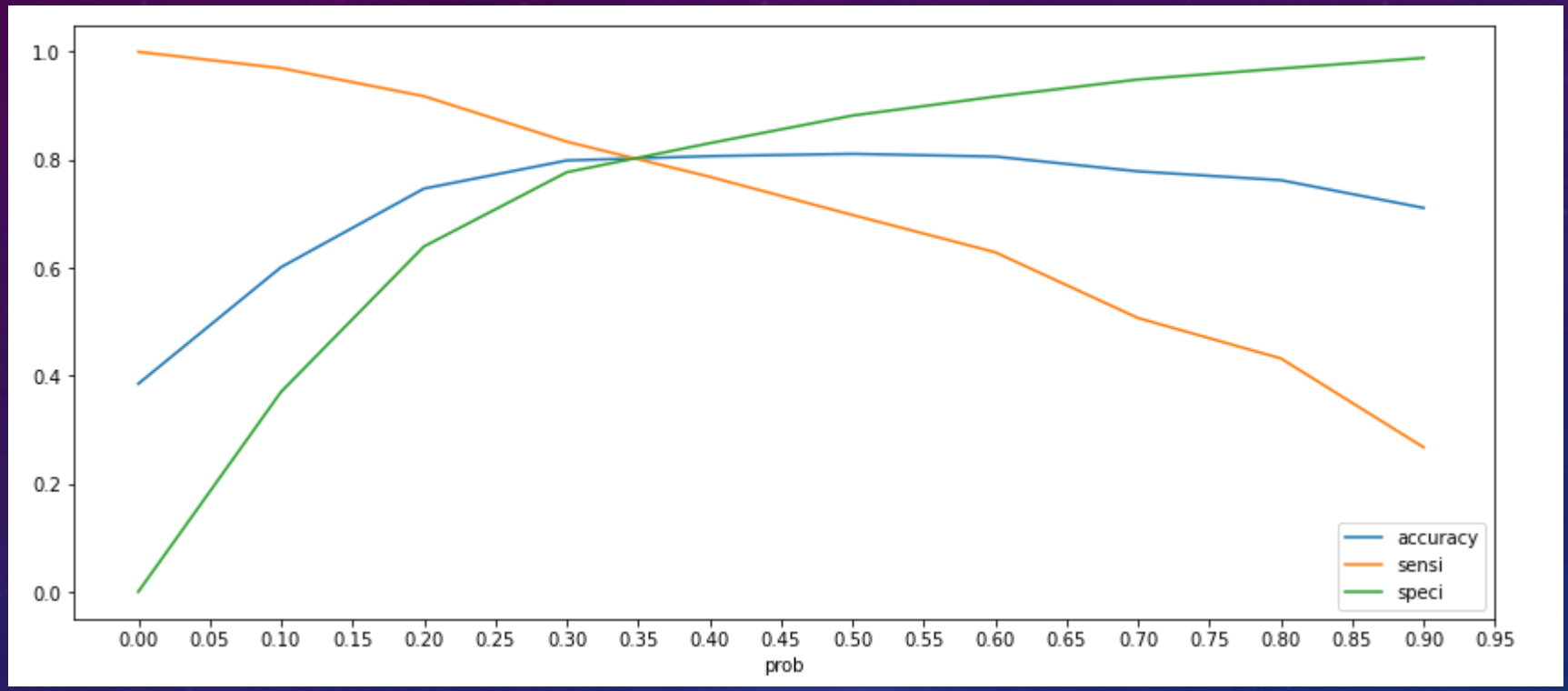
All the features have p value less than 0.05 and VIF less than 5

Model Evaluation : ROC Curve



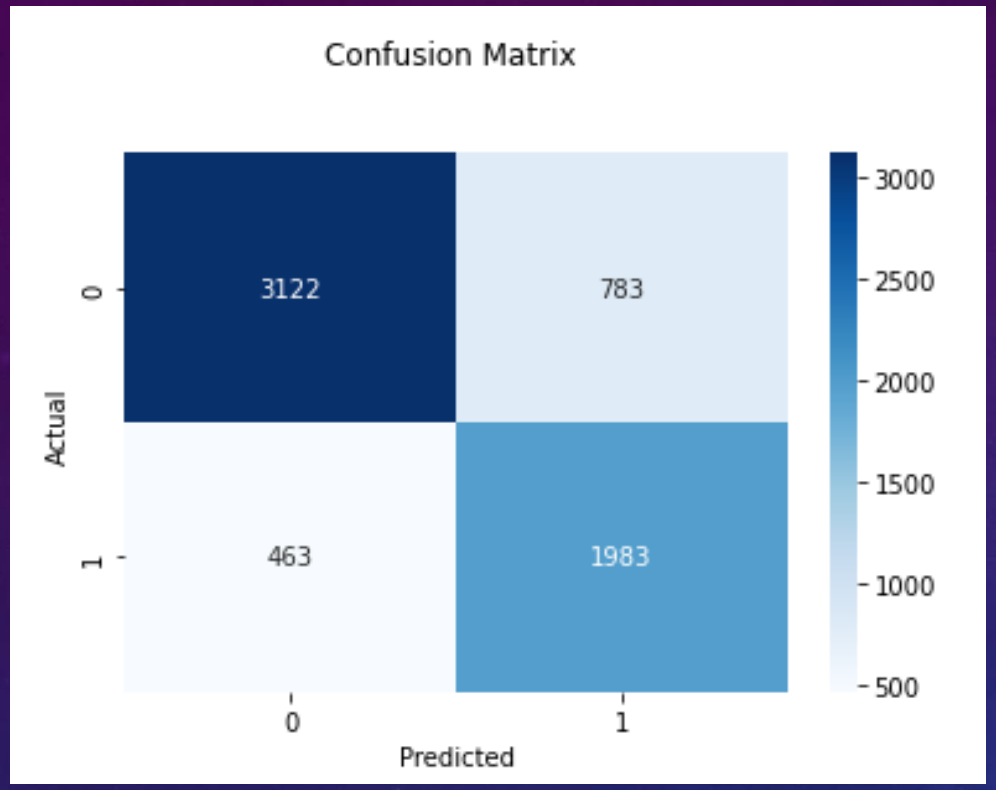
Area under the curve is 0.88, which means the predictive power of the model is good

Model Evaluation : Optimal Cut-Off Threshold

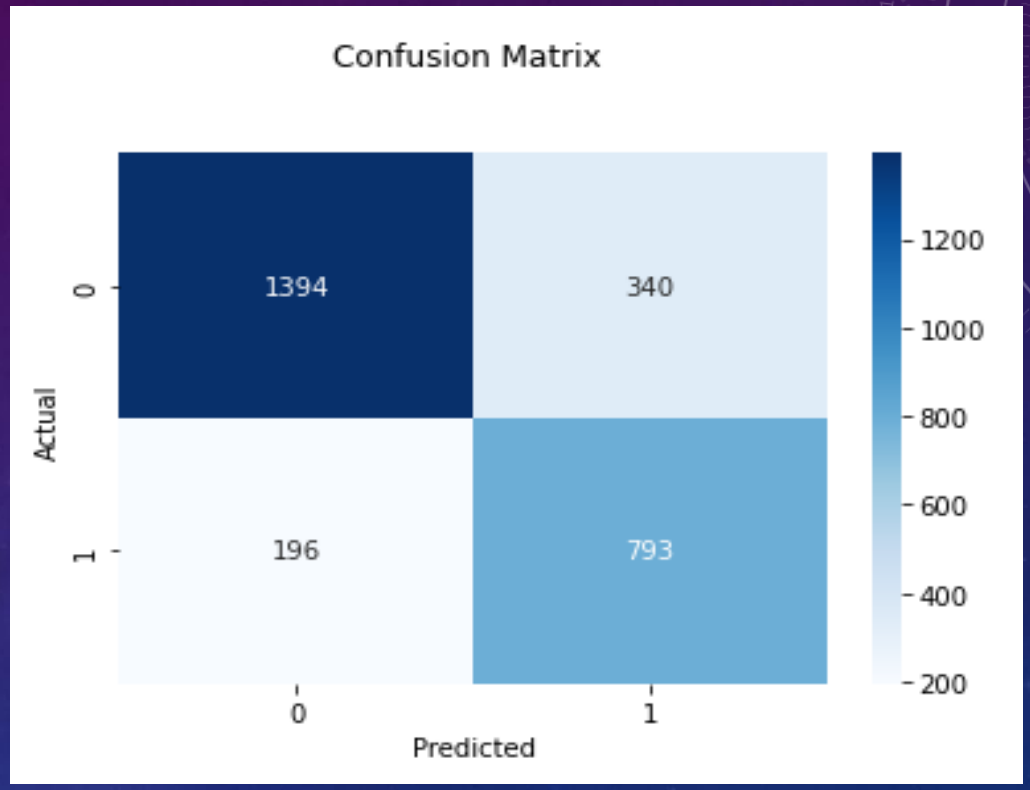


To get a balanced Accuracy, Sensitivity and Specificity, we determined the Optimal Cut-Off Threshold as 0.34

Model Evaluation : Confusion Matrix



Train Data



Test data

Model Evaluation :

Evaluation Data	Train Data	Test Data
Accuracy	0.80	0.80
Sensitivity	0.81	0.80
Specificity	0.80	0.80
False Positive Rate	0.20	0.20
Positive Predictive Value	0.72	0.70
Negative Predictive Value	0.87	0.89
F1-Score	0.76	0.75

Conclusion:

The final model has 13 features and the accuracy of the model is 80.32% at an optimal threshold of 0.34

Features having positive impact on conversion probability

1.Total Visits

2.Total Time Spent on Website

3.Lead Source :

- ❖ Olark Chat
- ❖ Reference
- ❖ Welingak Website
- ❖ Others

4.Occupation

- ❖ Working Professional

5.Last Notable Activity

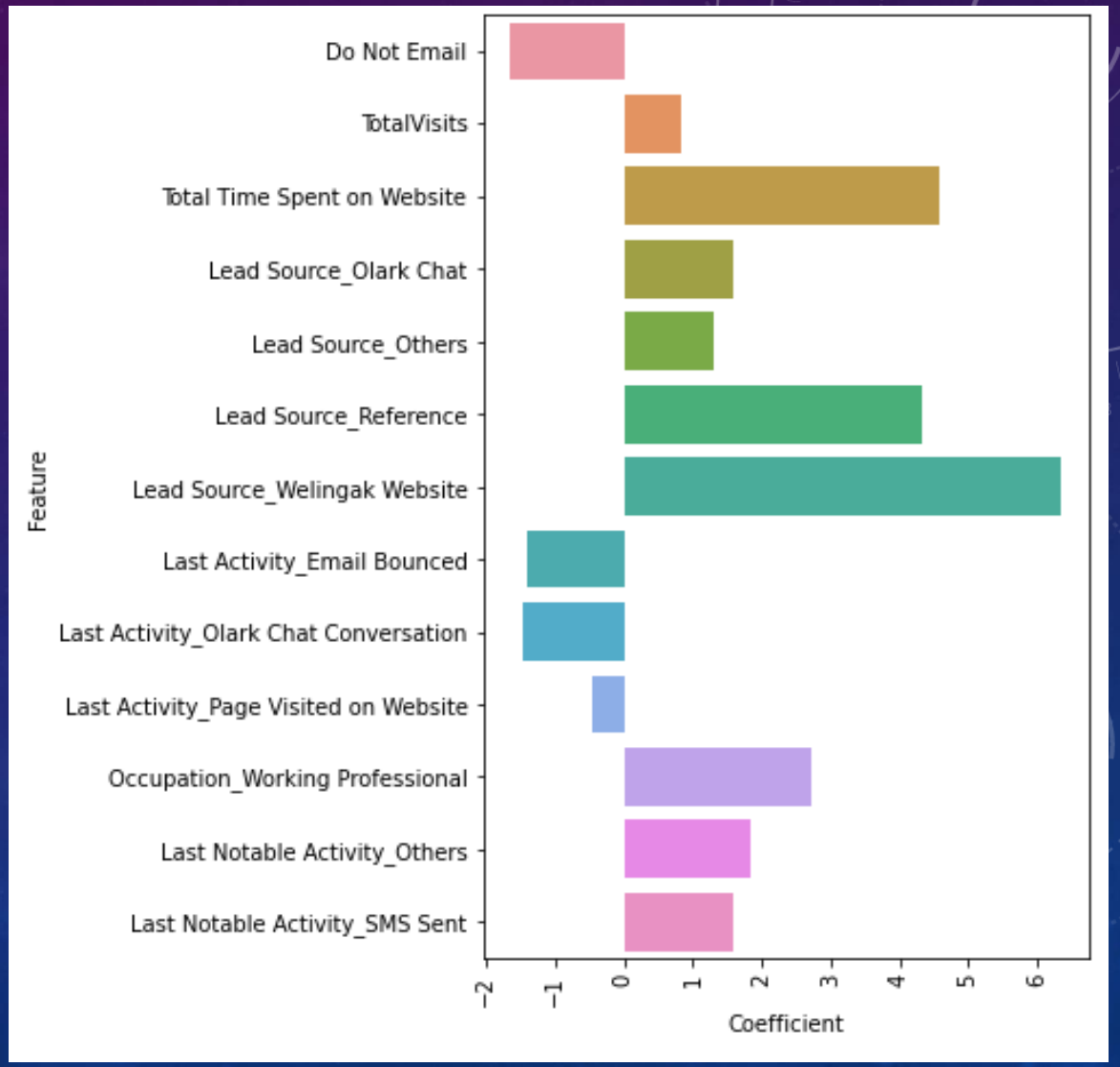
- ❖ SMS Sent
- ❖ Others

Features having negative impact on conversion probability

1.Do Not Email

2.Last Activity

- ❖ Email Bounced
- ❖ Olark Chat Conversation
- ❖ Page Visited on Website



THANK YOU