LEAD SCORING

CASE STUDY AND ASSIGNMENT

By

Ankur Singh &
Pratik Chouhan

"IT'S ONE SMALL STEP FOR MAN, ONE GIANT LEAP FOR MANKIND."

- Neil Armstrong

PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, but its lead conversion rate is very poor, around acquired 100 leads in a day, and only about 30 of them are converts.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’ whose conversion rate should be higher than 80%.
- 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.

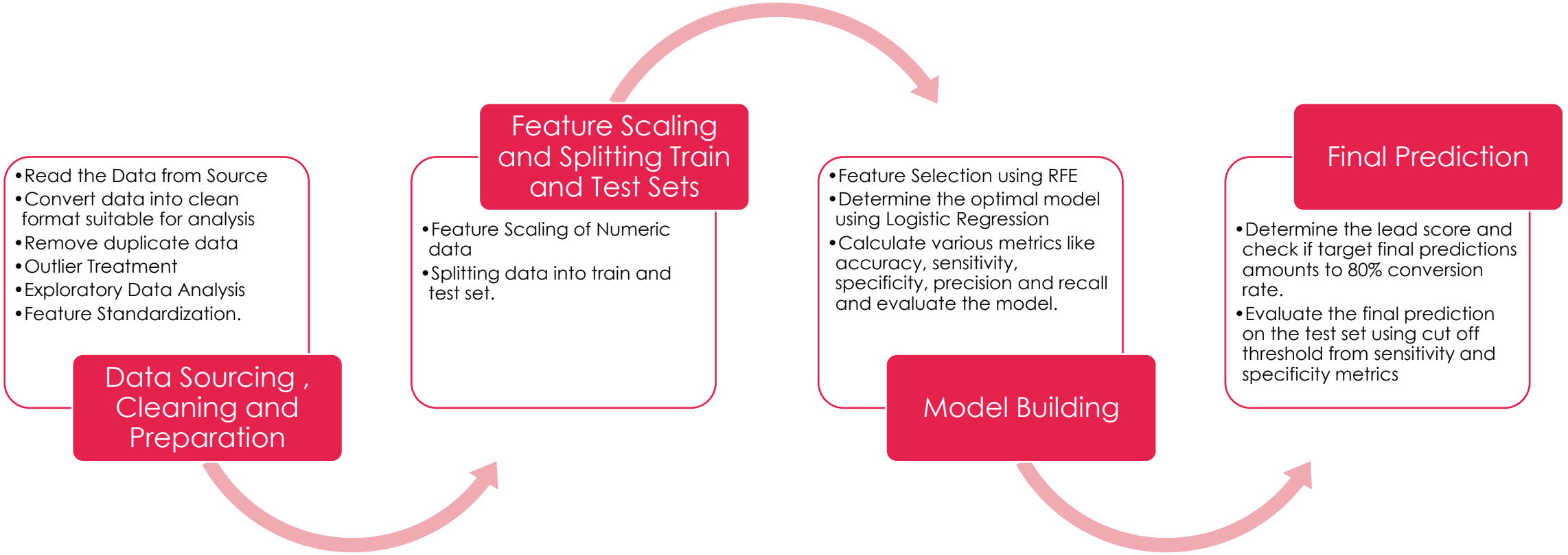
BUSINESS OBJECTIVE

- X education wants to know the most promising leads.
- For that they want to build a model which identifies the hot leads.
- Deployment of the model for future use.
- The model should have an accuracy of more than 80%.

STRATEGY AND APPROACH

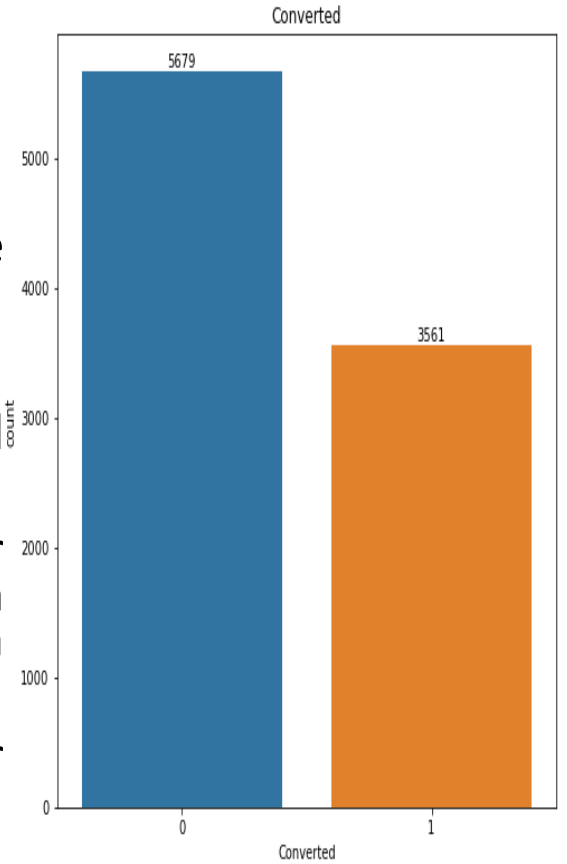
- Source the data for analysis
- Clean and prepare the data
- Exploratory Data Analysis.
- Feature Scaling
- Splitting the data into Test and Train datasets.
- Building a Logistic Regression model and calculating Lead Score.
- Evaluating the model by using different metrics -Specificity and Sensitivity or Precision and Recall.
- Applying the best model in Test data based on the Sensitivity and Specificity Metrics.

SOLUTION METHODOLOGY



DATA SUMMARY

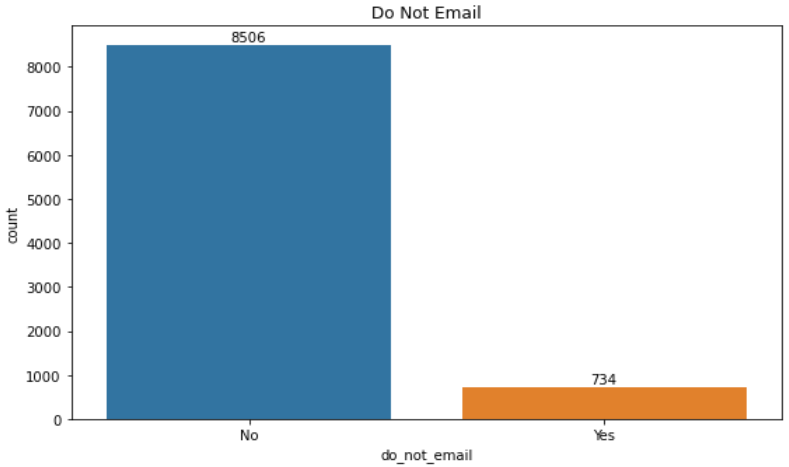
- Shape of Data: 9240 rows and 37 columns.
- There are no duplicate values.
- Prospect Id and Lead Number serve the same purpose. Prospect ID Should be dropped.
- Lead Number is data type int should be converted to object.
- Several Columns have a high missing value percentage. Should be dropped accordingly.
- As per the problem statement: Few categorical columns have "Select" in their entries. Those select are essentially null values because Select appears when someone does not select anything from the dropdown. So need to replace them with a null value and do the missing value analysis again.
- Columns have very long names. These columns can be renamed for better understanding.
- Current Conversion Rate is 38.53.

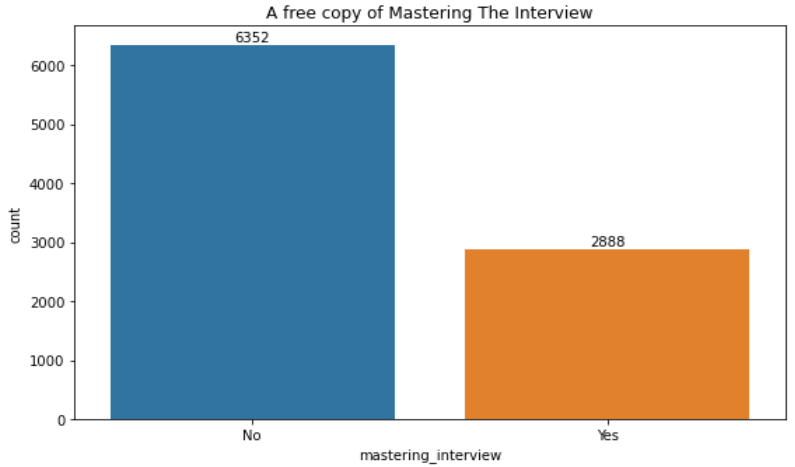


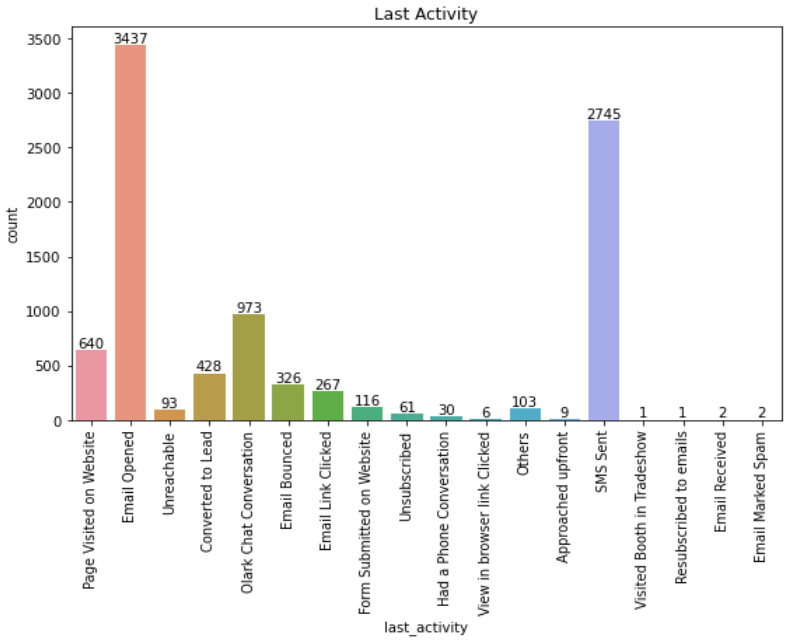
EXPLORATORY DATA ANALYSIS(EDA)

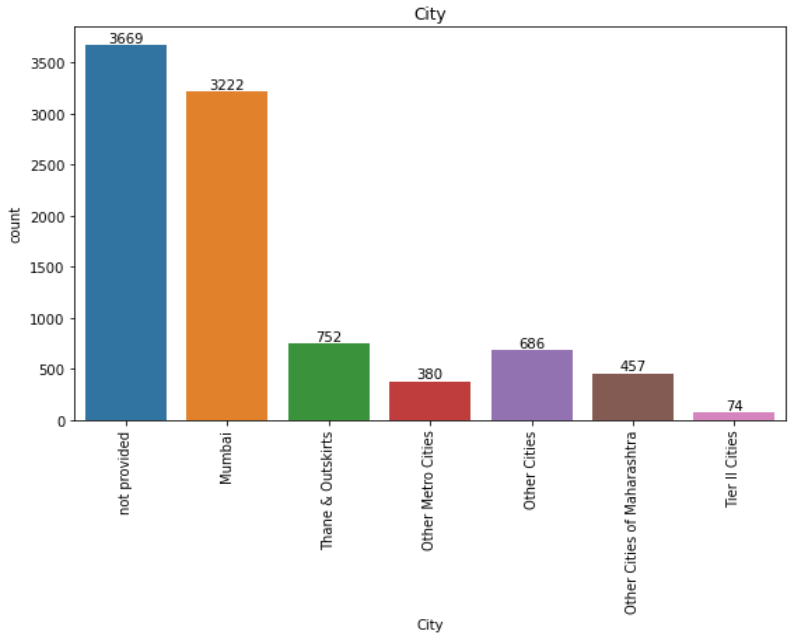
- **Univariate**

- **Do not Email**: 92% of leads said ‘No’
- **Mastering Interview Book**: Only 31.25 % said they want a copy of the book.
- **Last Activity**: Emailed opened and SMS sent are the two highest with 37.19% and 29.70%.
- **City**: 34.87% are from Mumbai





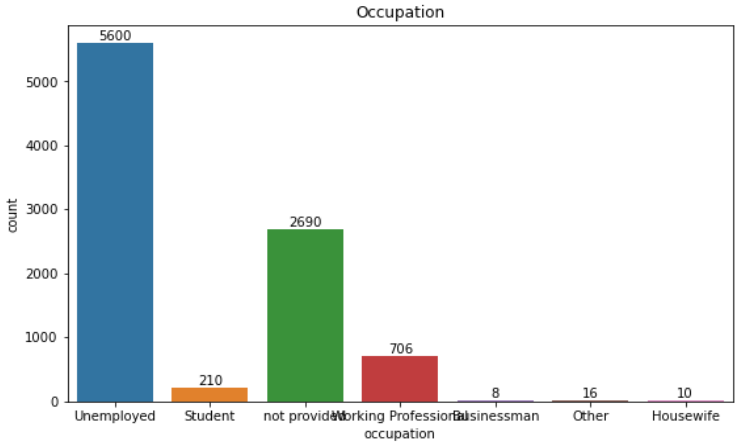


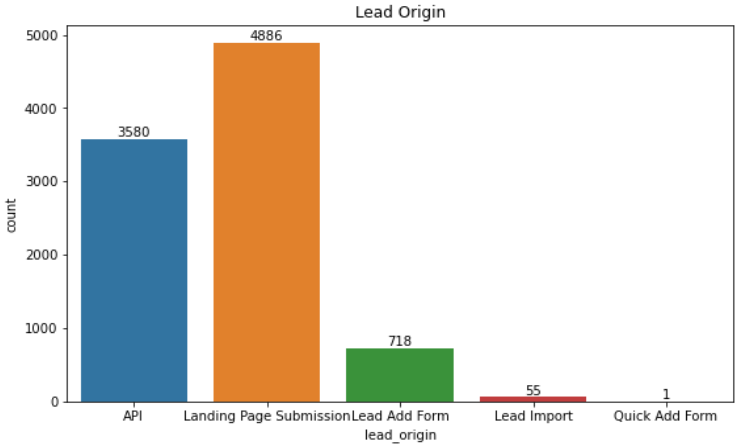


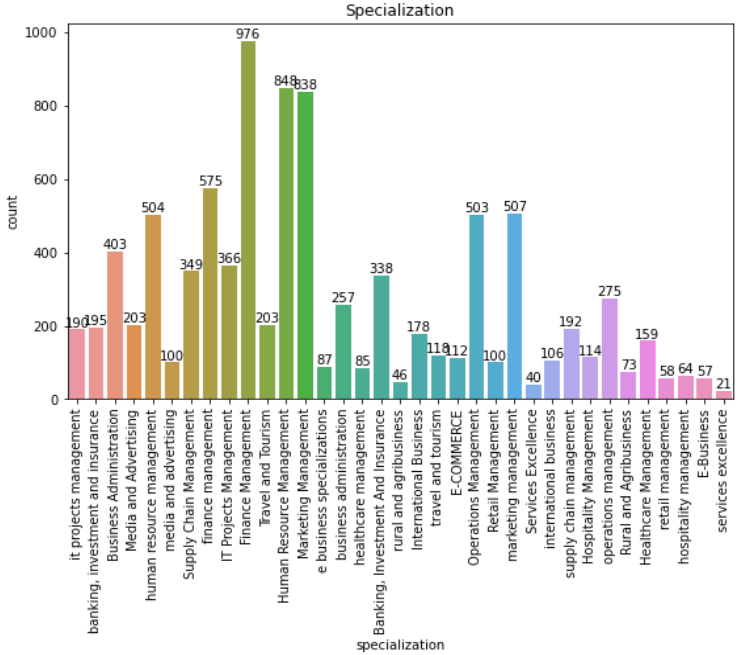
EXPLORATORY DATA ANALYSIS(EDA)

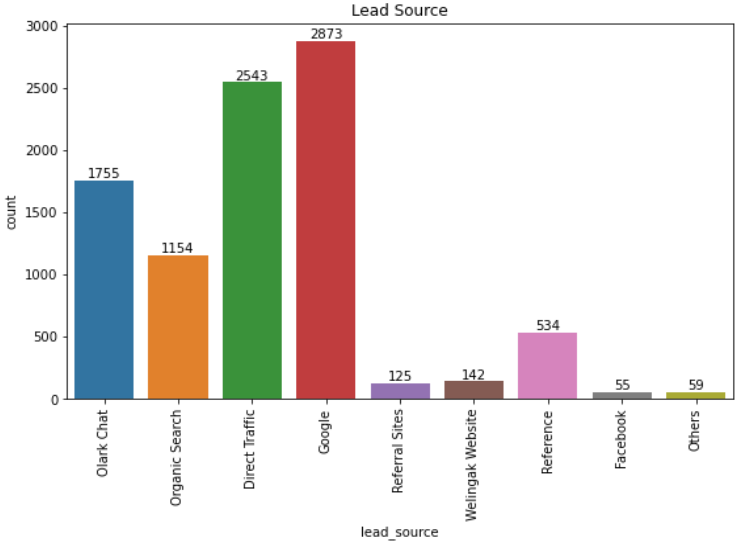
Univariate

- **Occupation**: 60.60 % of leads are unemployed.
- **Lead Origin**: Maximum leads i.e. 52.88 % originated from Landing Page.
- **Specialization**: Maximum leads have a background in Finance Management followed by HR and Marketing Management i.e. 10.56%, 9.17%, and 9.06% respectively.
- **Lead Source**: 58.61% of lead come from two source i.e. Google and Direct Traffic with 31.09% and 27.52% respectively.





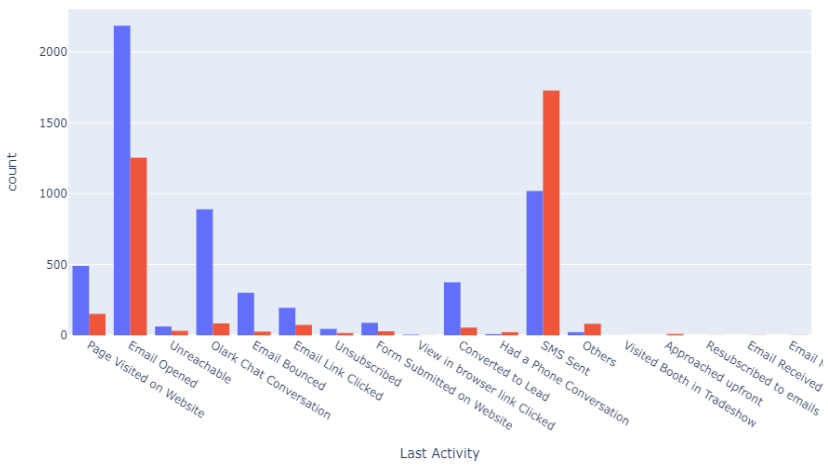


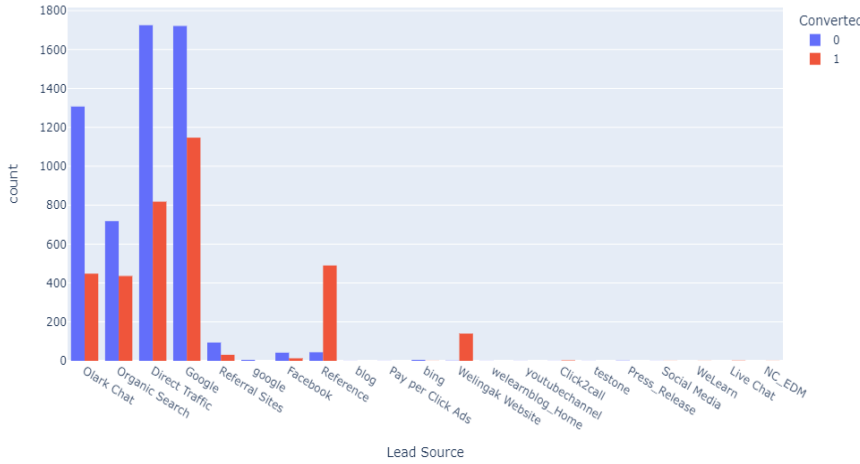


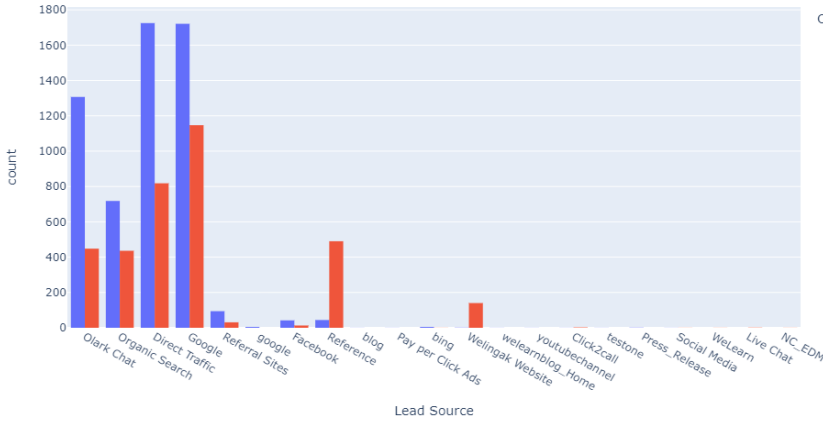
EXPLORATORY DATA ANALYSIS(EDA)

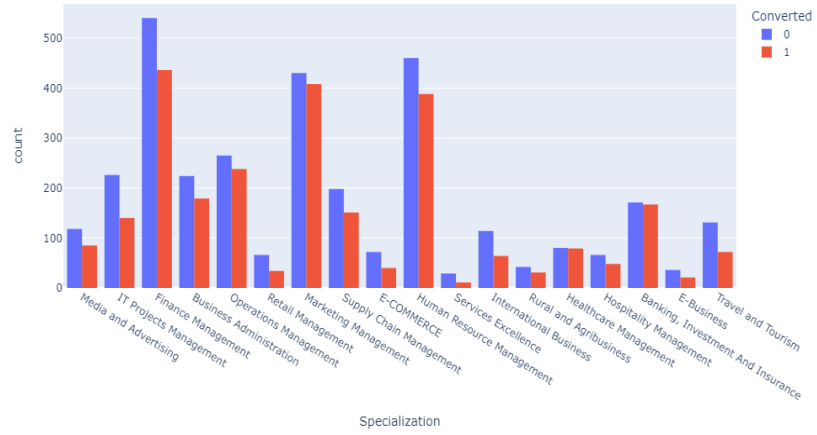
Bivariate and Multivariate

- **Lead Activity**: Maximum conversion is from Email opened and SMS sent
- **Lead Source**: Maximum conversions are from google, but Reference has the highest lead-to-conversion ratio.
- **Specialization**: Maximum lead converted is from Finance followed by HR and Marketing management.
- **Lead Source**: Maximum leads are converted from Google followed by Direct traffic, but reference has an exception with lower lead generation and high lead conversion.





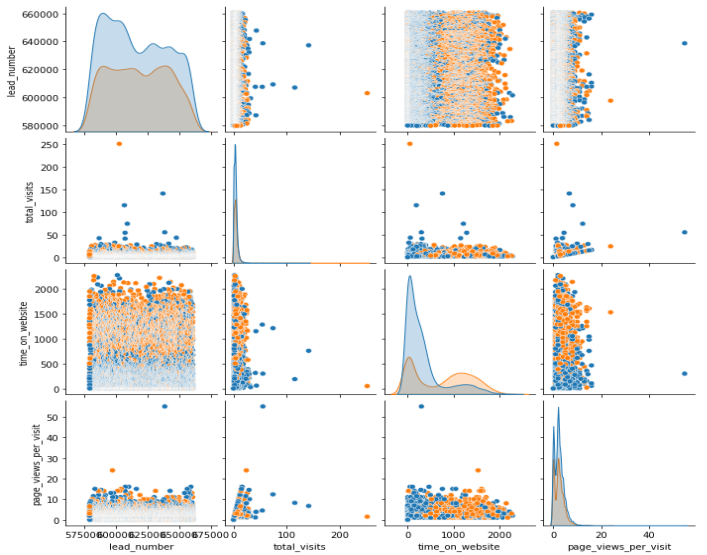


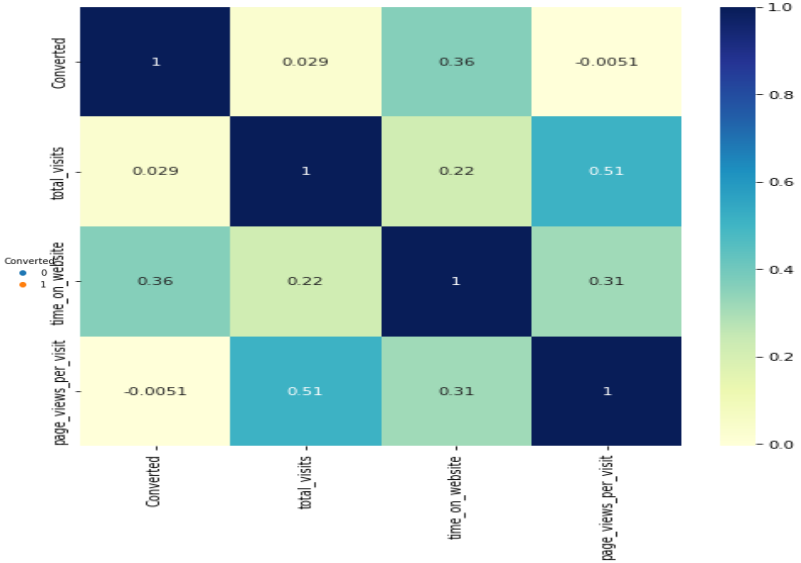


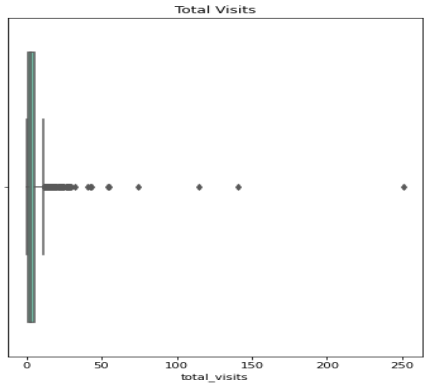
EXPLORATORY DATA ANALYSIS(EDA)

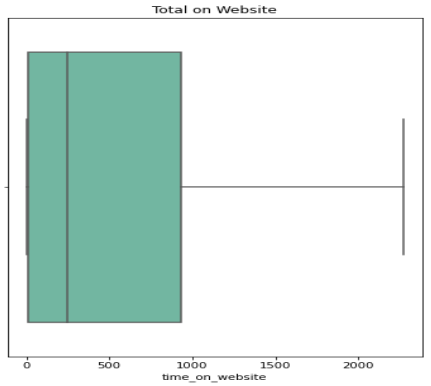
Bivariate and Multivariate

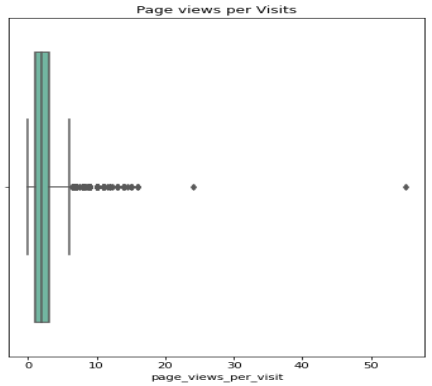
- **Total Visits, Time on Website, and Page View per Visits**: These numerical variables are not correlated, Time on the Website has no outlier and there are few outliers in Total visits and Page views per visit.











DATA PREPARATION

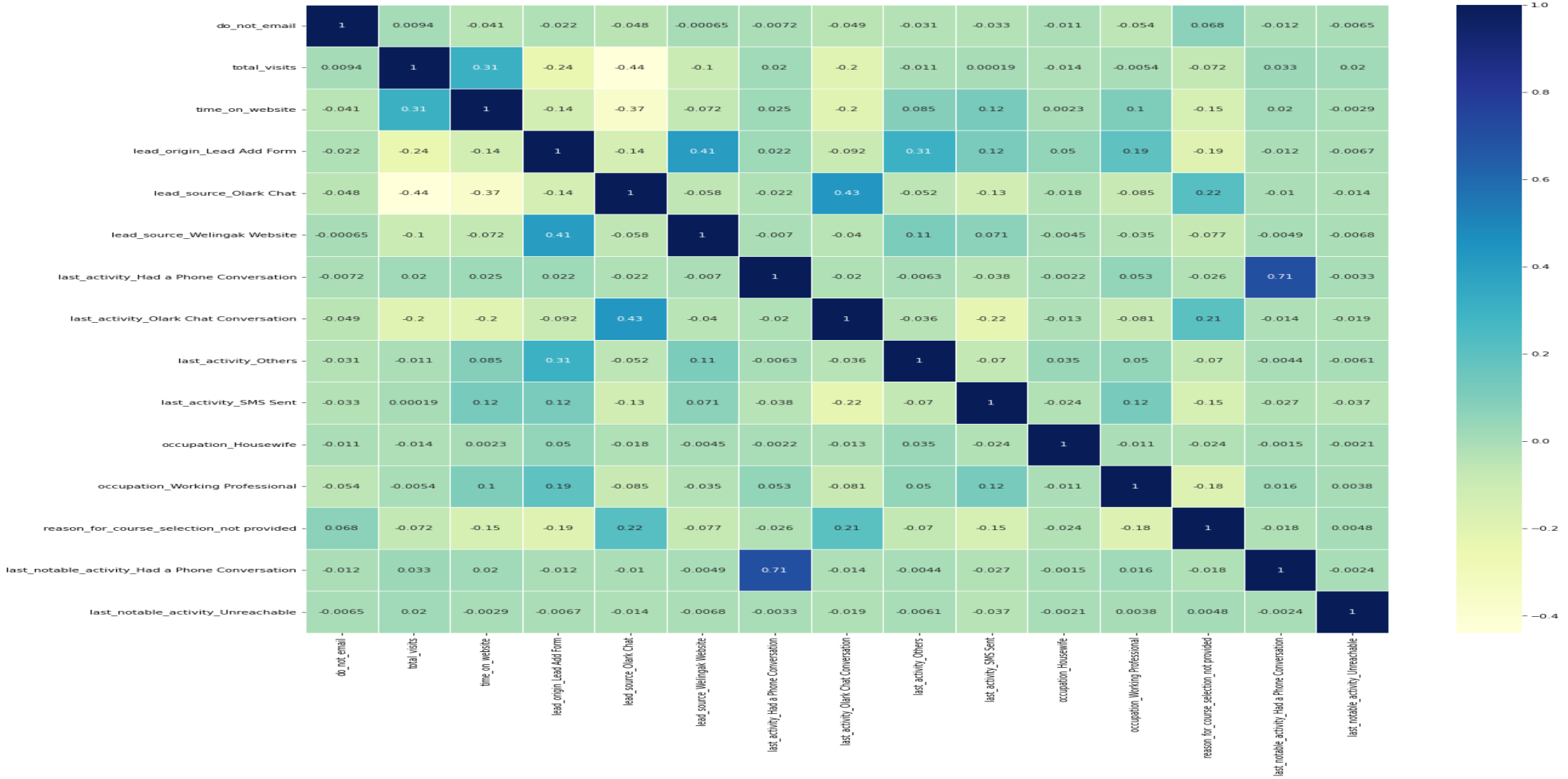
- Numerical Variables are Normalised using binary mapping to 0 and 1.
- MinMaxScaler is used for Numerical Feature scaling.
- Dummy Variables are created for object-type variables.
- ‘Tags’ column is dropped as it is generated by the Sales team after the lead is generated.
- Total Rows for Analysis: 9240
- Total Columns for Analysis: 100

MODEL BUILDING

- Splitting the Data into Training and Testing Sets
- Regression is performing a train-test split, we have chosen a 70:30 ratio with a random state of 100.
- RFE is used for Feature Selection with 15 variables. The Variable selected by RFE are: occupation_Housewife, time_on_website , lead_origin_Lead Add Form, last_notable_activity_Had a Phone Conversation, occupation_Working Professional, last_notable_activity_Unreachable, lead_source_Welingak Website, lead_source_Olark Chat, last_activity_SMS Sent, total_visits, last_activity_Had a Phone Conversation, last_activity_Olark Chat Conversation, reason_for_course_selection_not provided, do_not_email, last_activity_Others.
- Building Model by removing the variable whose p-value is greater than 0.05 and VIF value is greater than 5.

MODEL BUILDING

- **Correlation Matrix for Variable selected by RFE:** No such Major Correlation was observed



MODEL BUILDING

- **Final Model Summary with p-Values and VIF**

Generalized Linear Model Regression Results

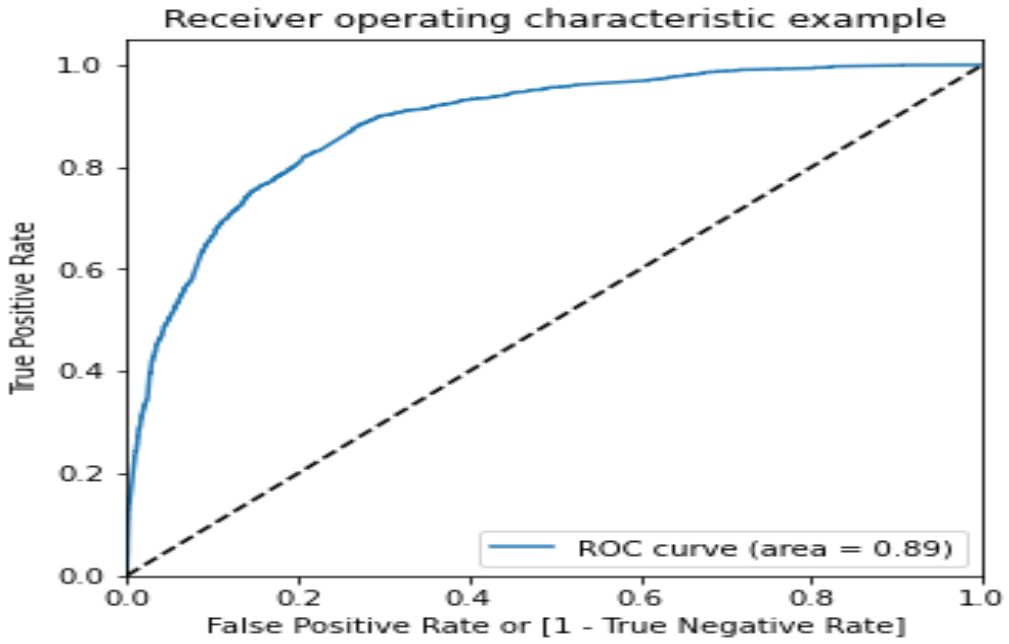
Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6456
Model Family:	Binomial	Df Model:	11
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2664.2
Date:	Sun, 12 Jun 2022	Deviance:	5328.4
Time:	13:08:05	Pearson chi2:	6.85e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.4183	0.093	-26.090	0.000	-2.600	-2.237
do_not_email	-1.3753	0.165	-8.324	0.000	-1.699	-1.051
total_visits	1.0050	0.205	4.905	0.000	0.603	1.407
time_on_website	4.4952	0.164	27.350	0.000	4.173	4.817
lead_origin_Lead Add Form	4.0393	0.199	20.258	0.000	3.649	4.430
lead_source_Olark Chat	1.5757	0.116	13.632	0.000	1.349	1.802
last_activity_Olark Chat Conversation	-1.1825	0.167	-7.074	0.000	-1.510	-0.855
last_activity_Others	-1.5809	0.439	-3.603	0.000	-2.441	-0.721
last_activity_SMS Sent	1.2837	0.074	17.383	0.000	1.139	1.428
occupation_Working Professional	2.5234	0.187	13.511	0.000	2.157	2.889
reason_for_course_selection_not provided	-1.1784	0.086	-13.667	0.000	-1.347	-1.009
last_notable_activity_Unreachable	2.0033	0.539	3.716	0.000	0.947	3.060

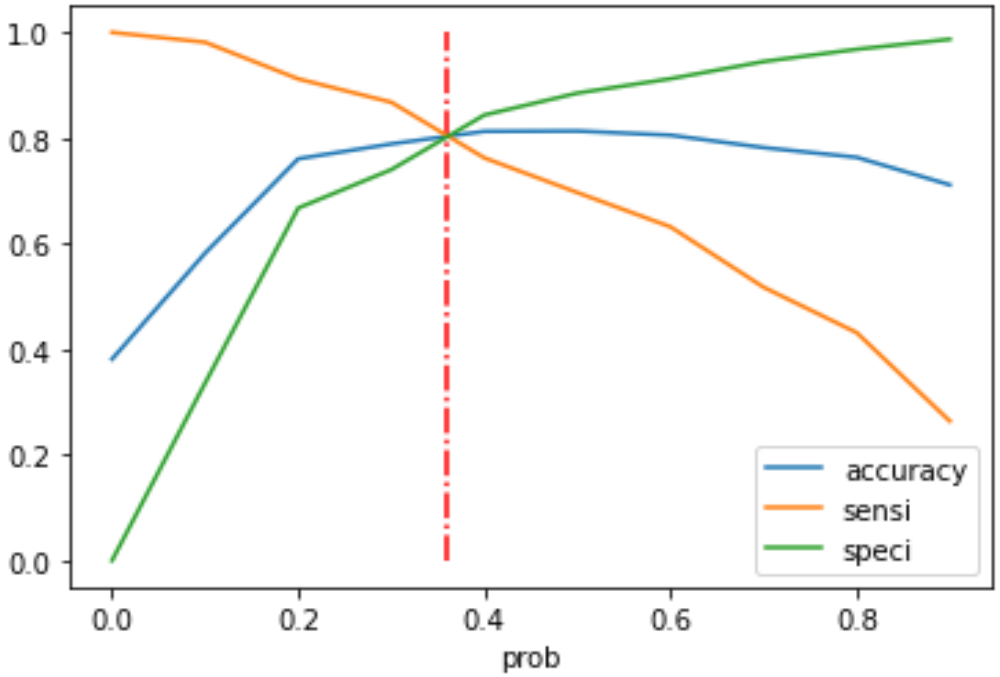
	Features	VIF
2	time_on_website	1.89
1	total_visits	1.88
4	lead_source_Olark Chat	1.55
7	last_activity_SMS Sent	1.46
9	reason_for_course_selection_not provided	1.45
5	last_activity_Olark Chat Conversation	1.42
3	lead_origin_Lead Add Form	1.28
8	occupation_Working Professional	1.17
6	last_activity_Others	1.16
0	do_not_email	1.07
10	last_notable_activity_Unreachable	1.00

MODEL PREDICTION

- **ROC Curve**: Area under the curve is 0.89.



- **Plotting Accuracy, Sensitivity, and Specificity:** 0.36 is the optimum point to take as a cut-off probability.



MODEL PREDICTION

Data with an optimal Cut-off of 0.36

- Confusion Matrix :

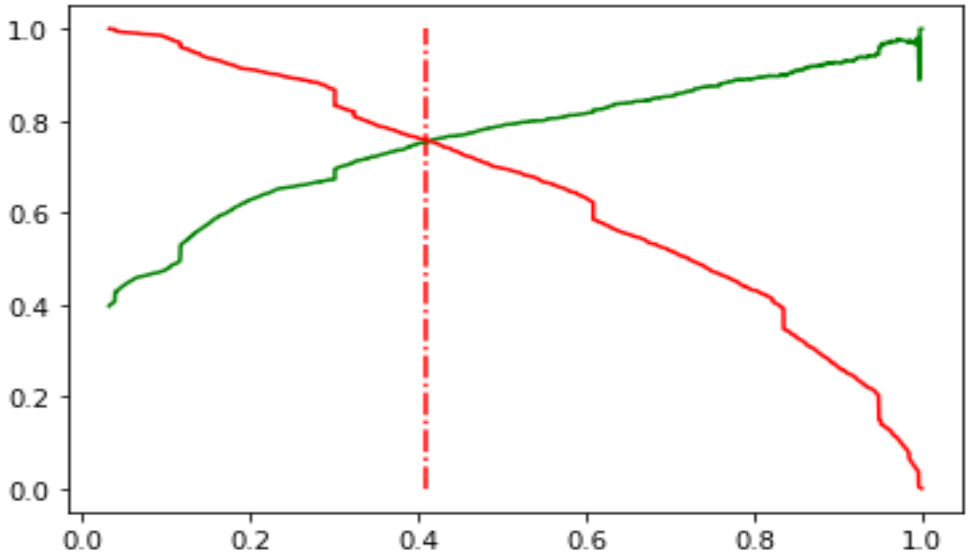
3279	723
527	1939

Confusion Matrix

- Accuracy : 80.67%
- Sensitivity : 78.63%
- Specificity : 81.93%
- False Positive Rate : 18.06%
- Positive Predictive Value : 72.84%
- Negative predictive value : 86.15%

Precision and Recall:

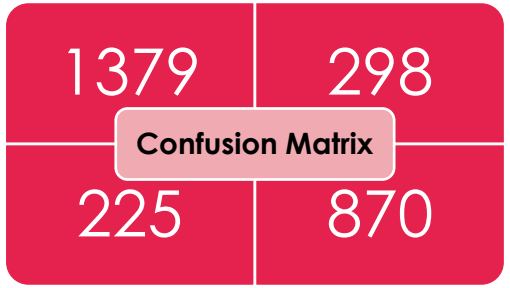
- Precision: 78.88% and Recall: 69.67%
- The graph depicts an optimal cut-off of 0.41 based on Precision and Recall



MODEL EVALUATION

Prediction on Train Data with an optimal Cut-off of 0.36

- Confusion Matrix :



- Accuracy : 81.13%
- Sensitivity : 79.45%
- Specificity : 82.23%
- False Positive Rate : 17.76%
- Positive Predictive Value : 74.49%
- Negative predictive value : 85.97%

Sample Final Table

	lead_number	Converted	prob_lead_conversion	final_Prediction	lead_score
0	4269	1	0.544179	1	54
1	2376	1	0.948081	1	95
2	7766	1	0.925650	1	93
3	9199	0	0.116594	0	12
4	4359	1	0.834938	1	83

MODEL FEATURES

Final model Features

Features	Co-efficient	Impact
time_on_website	4.495224	Positive
lead_origin_Lead Add Form	4.039318	Positive
occupation_Working Professional	2.523387	Positive
last_notable_activity_Unreachable	2.003294	Positive
lead_source_Olark Chat	1.575666	Positive
last_activity_SMS Sent	1.283716	Positive
total_visits	1.005008	Positive
reason_for_course_selection_not provided	-1.17837	Negative
last_activity_Olark Chat Conversation	-1.18247	Negative
do_not_email	-1.37531	Negative
last_activity_Others	-1.58092	Negative
const	-2.41828	Negative

CONCLUSION

- **Final Formula for Logistic Regression Model is**

ln (p/(1-p)) = -2.418279 + time_on_website*4.495224 + lead_origin_Lead Add Form*4.039318 + occupation_Working Professional*2.523387 + last_notable_activity_Unreachable*2.003294 + lead_source_Olark Chat*1.575666 + last_activity_SMS Sent*1.283716 + total_visits*1.005008 + reason_for_course_selection_not provided*-1.178366 + last_activity_Olark Chat Conversation*-1.182474 + do_not_email*-1.375312 + last_activity_Others*-1.580917

- The top 3 variables that contribute to lead getting converted in the model are:
  - Total time spent on the website
  - Lead Add Form from Lead Origin
  - Having occupation as Working Professional
- Train Data:
  - Accuracy: 80.67%
  - Sensitivity: 78.62%
  - Specificity: 81.93%
- Test Data:
  - Accuracy: 81.13%
  - Sensitivity: 79.45%
  - Specificity: 82.23%
- The Model has an accuracy of >80%, which seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model to get a higher lead conversion rate of 80%.