LEAD SCORING CASE STUDY

group members

Rajdeep singh

Rinsha p

Ritu pareek

PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead.
- some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.
- 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 ‘Hot 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.

GOALS OF THE CASE STUDY

- 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.
- The Aim for the whole exercise is to reach a lead conversion rate of 80%.
- Model should be able to adjust to if the company's requirement changes in the future.

STEPS

- Data set understanding and cleaning
  - Handling missing values
- Data analysis(EDA)
  - Univariate Analysis
  - Bivariate Analysis
- Preparing data for modelling
  - Dummy variables
  - Test-train split
  - Feature scaling
- Model Building
- Model Evaluation
- Predictions
- Model validation
- Conclusion/Recommendations

DATA SET UNDERSTANDING AND CLEANING

- 9240 rows and 37 columns.
- Data cleaning and handling missing values.
  - Few columns showing 'Select'. These values are as good as missing values and hence it will convert 'Select' values to Nan, replacing 'Select' values with Nan.
  - Dropping columns having missing value percentages greater than or equal to 35%.
  - Dropping Prospect ID and Lead Number columns because they are irrelevant information or variables for data analysis.
  - Dropping unique valued columns.
  - Imputing columns having missing value percentages more than 20% .
  - Dropping the rows with remaining missing data.

EXPLORATORY DATA ANALYSIS

- **Univariate Analysis:**

➢Visualized categorical variable using count plot.

➢Visualized Numerical Variables using Histograms.

- **Bivariate Analysis:**

➢Visualized the relationship between categorical variables and target variable " converted" using count plot.

EXPLORATORY DATA ANALYSIS (KEY INSIGHTS)

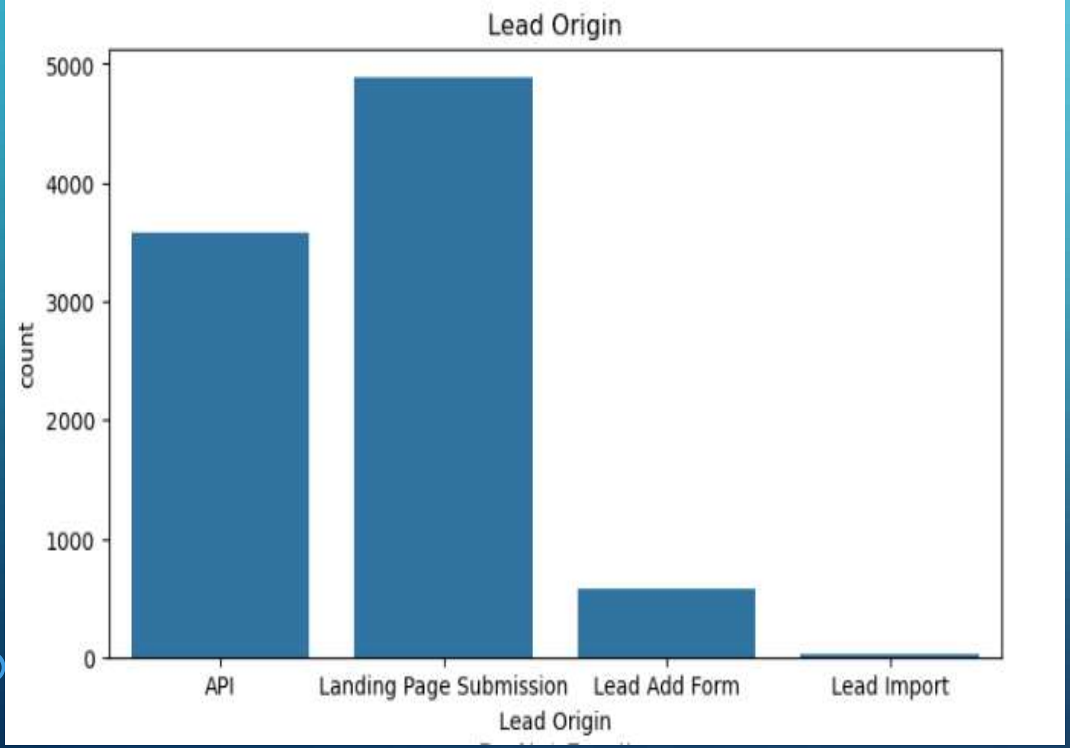
- **Univariate Analysis:**

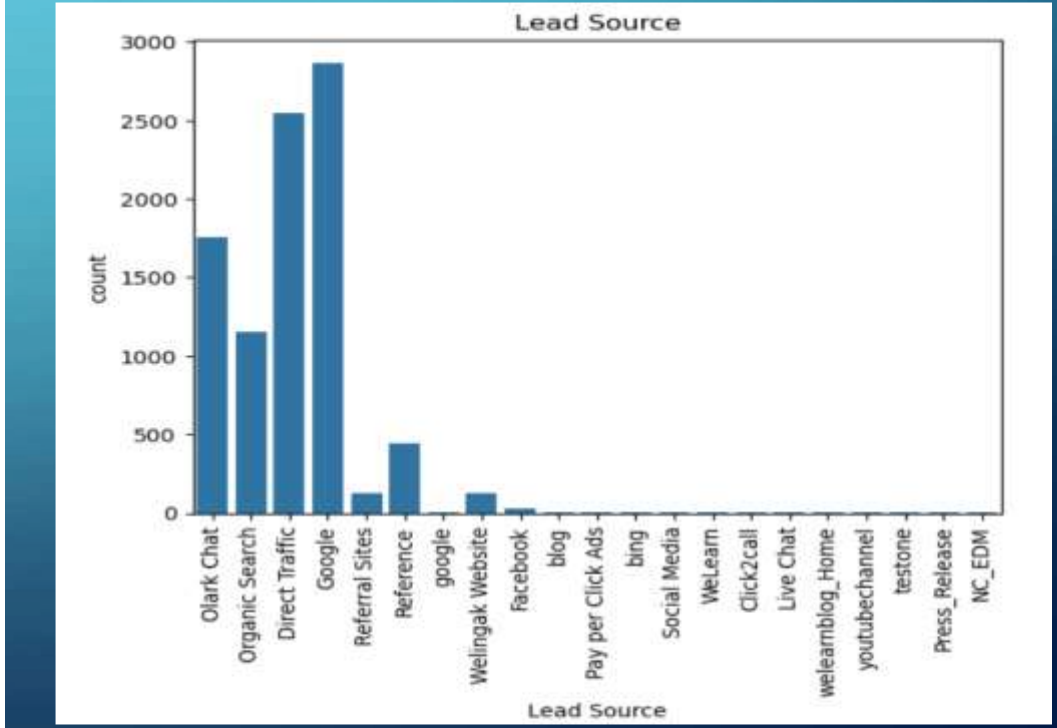
➢ Most of the Leads are generated from Landing Page Submission, followed by APIs.

➢ Most Source of Leads are coming from Google followed by Direct Traffic.

➢ Most Leads generated are from India.

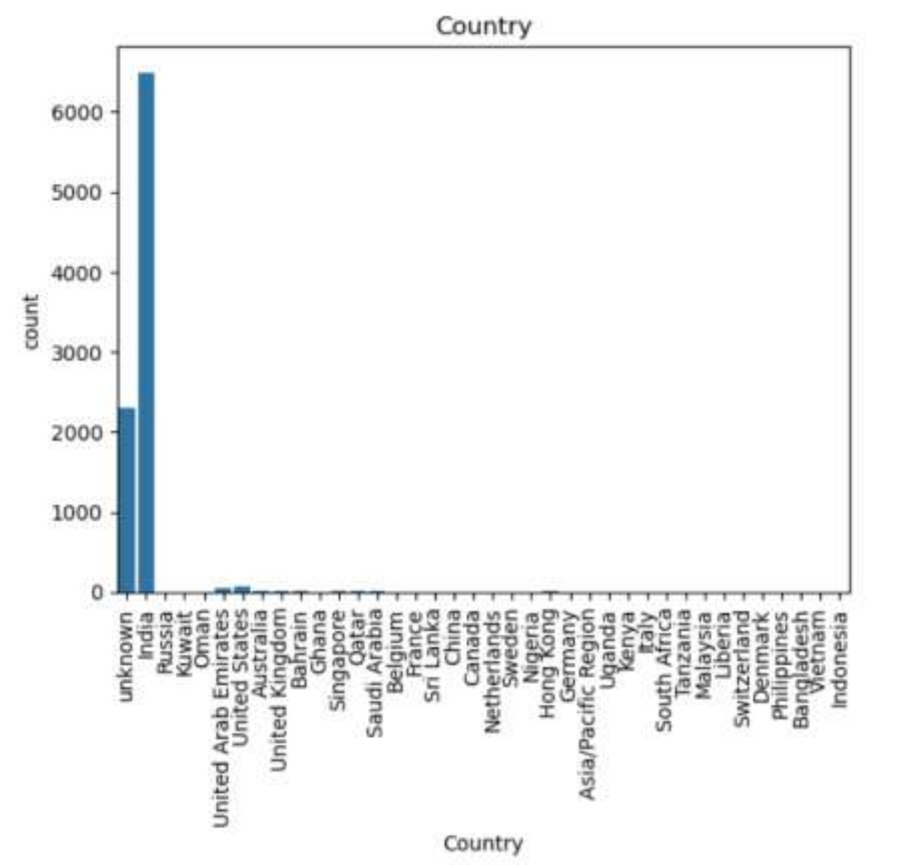
➢ Most Leads are looking for a course to get better career prospects.

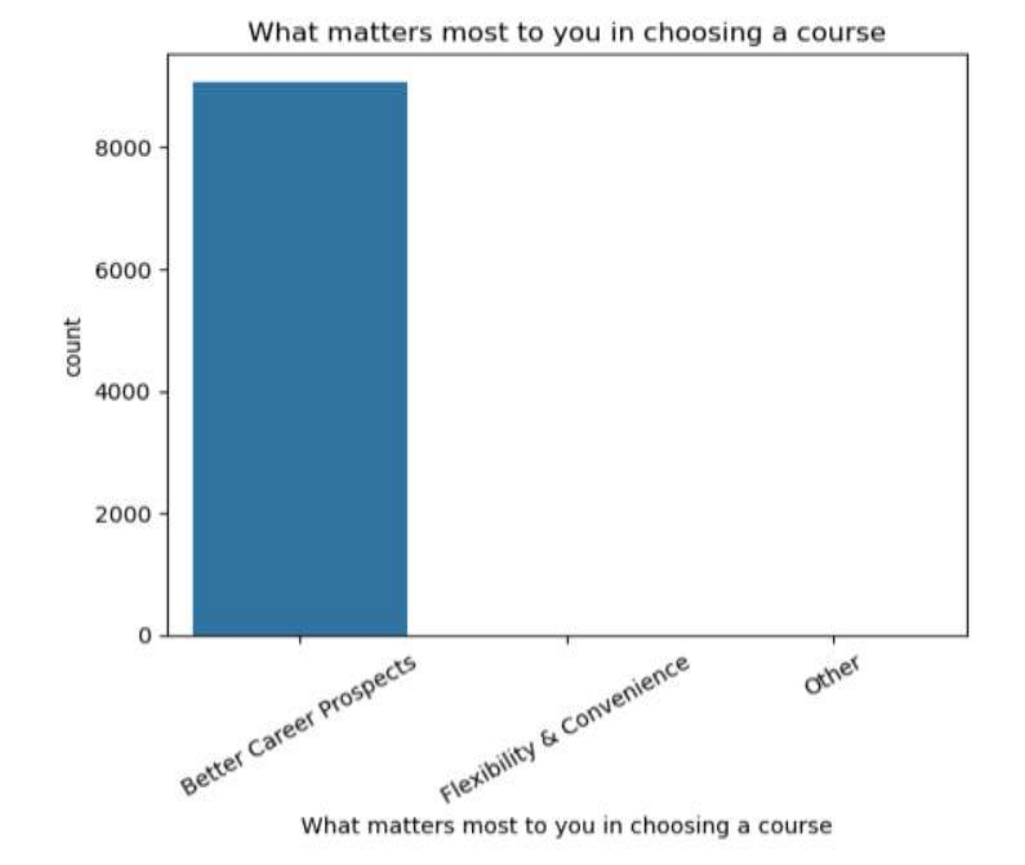




EXPLORATORY DATA ANALYSIS (KEY INSIGHTS)

Univariate Analysis:





EXPLORATORY DATA ANALYSIS (KEY INSIGHTS)

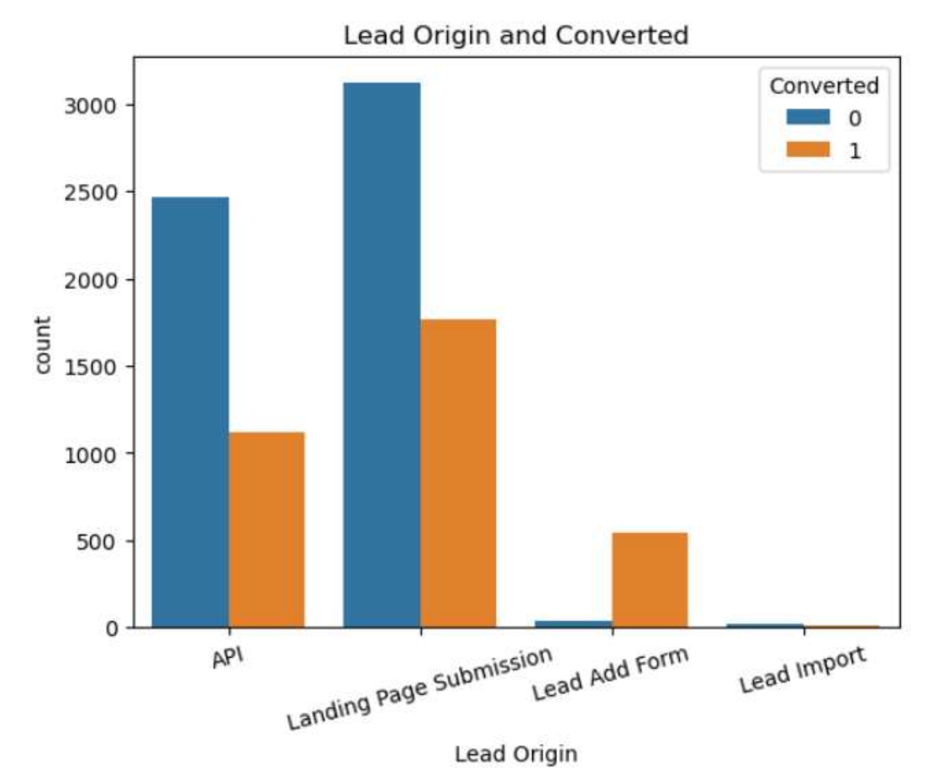
- **Bivariate Analysis:**

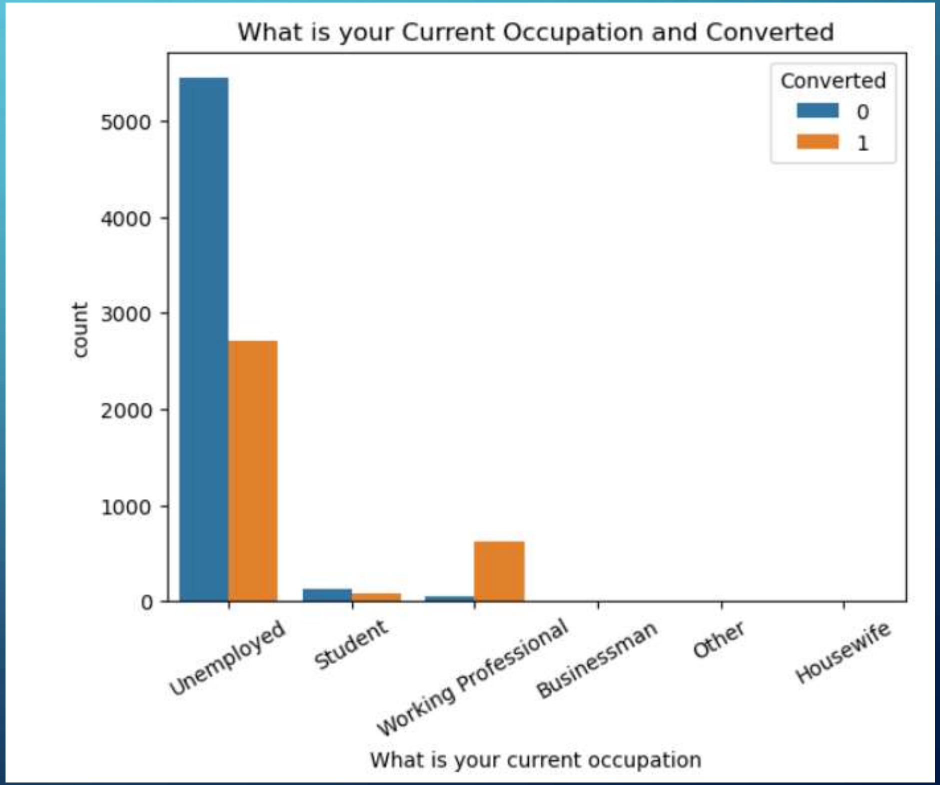
➢ ‘Lead Add form’ is a great origin to identify leads as it has the highest percentage of conversions.

➢ Highest number of leads generated are ‘Unemployed’ whereas highest percentage of conversion Is among the leads who are ‘Working Professionals’.

➢ Leads whose last notable activity was SMS sent have highest percentage of conversion.

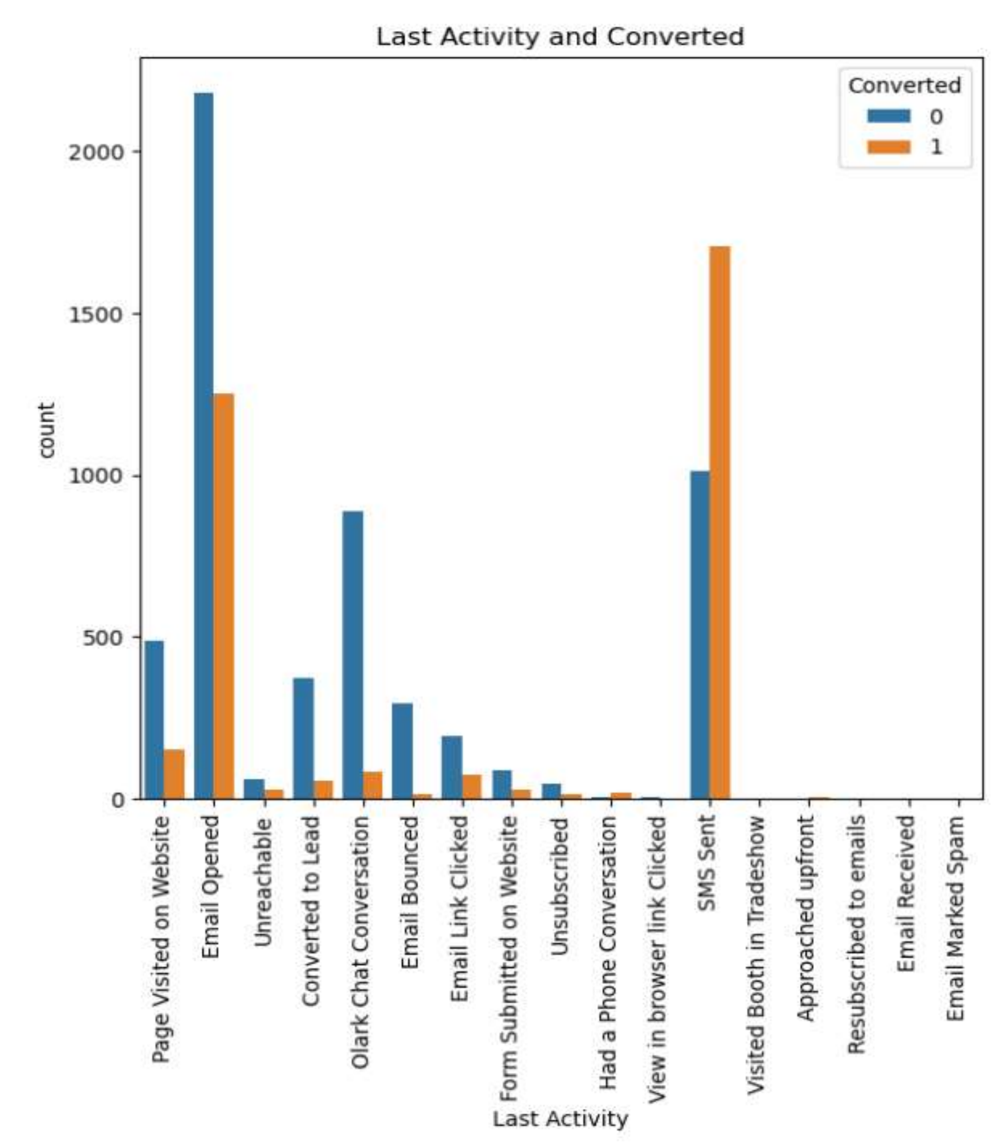
➢ Leads identified through sources ‘Reference’ and ‘Welingak Website’ have pretty high conversion rates.

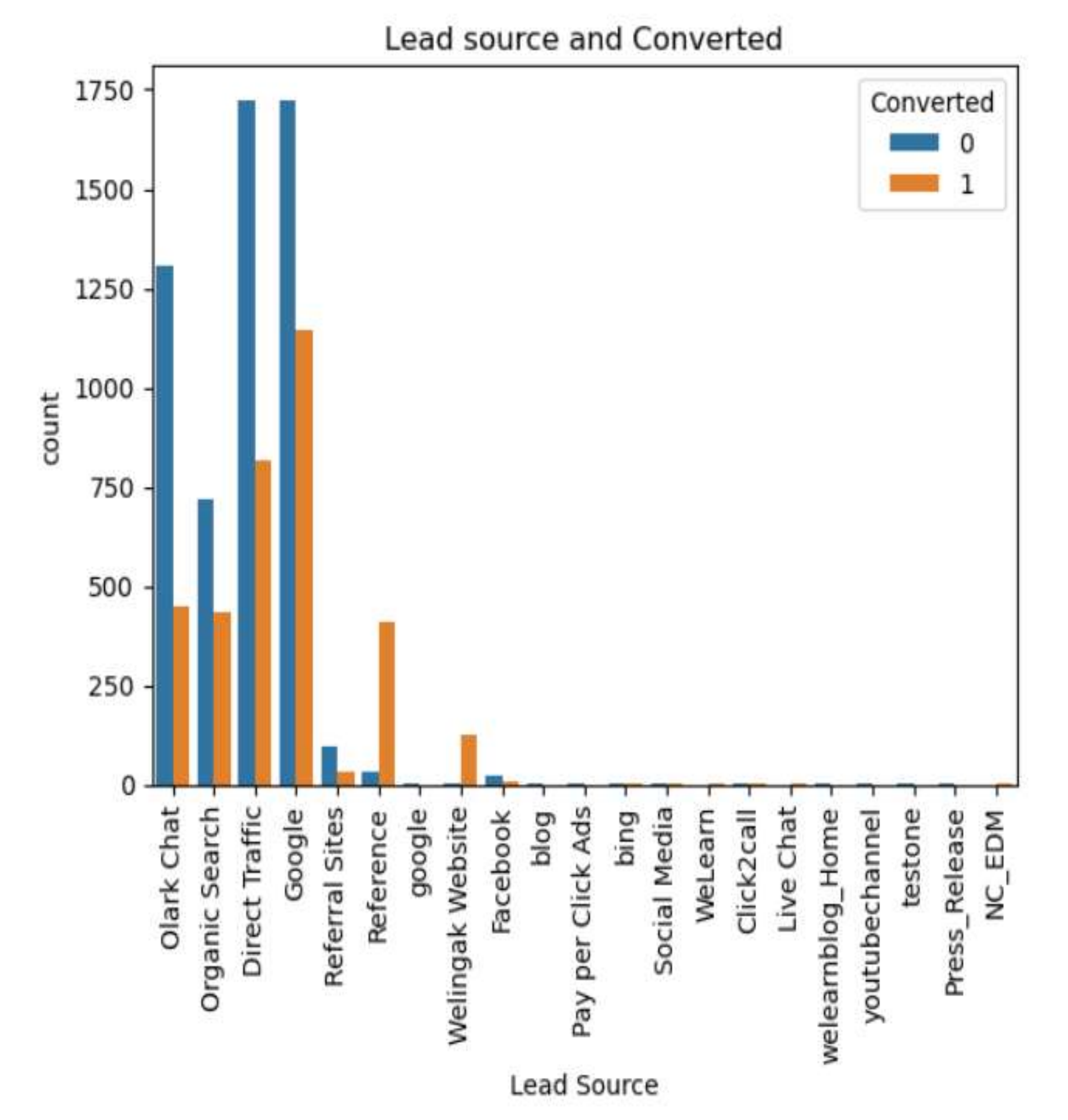




EXPLORATORY DATA ANALYSIS (KEY INSIGHTS)

- **Bivariate Analysis:**



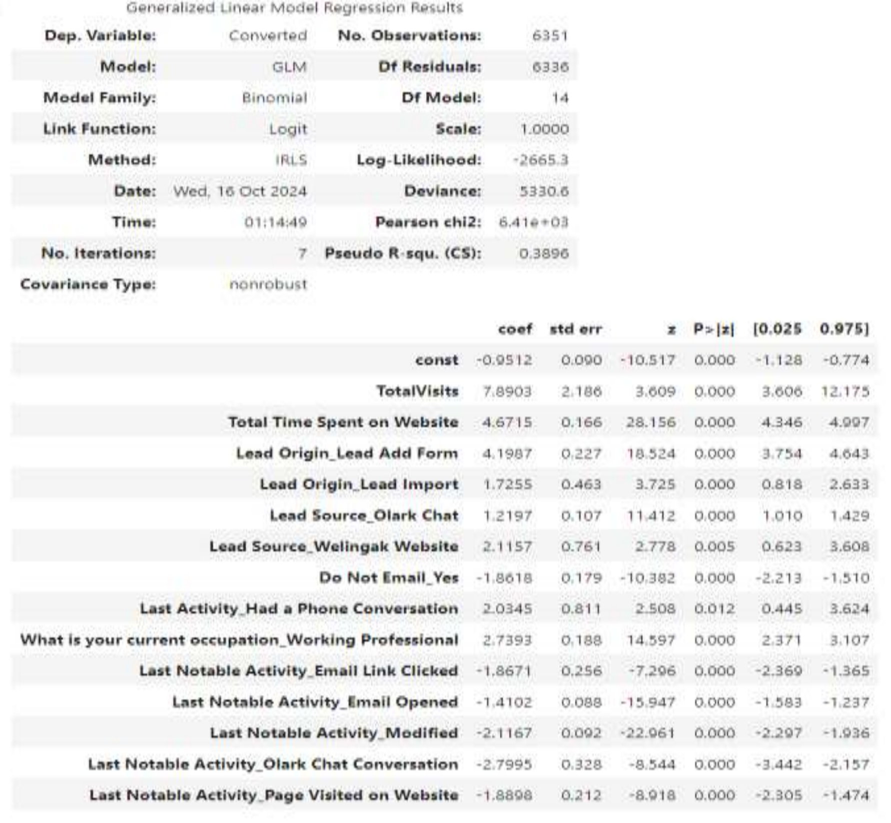


PREPARING DATA FOR MODELLING

- Created dummy variables and dropped the redundant dummy variables.
- Test-Train Split
  - ➢Data set divided into test and train with proportion of 70-30%.
- Feature Scaling
  - ➢Minimax scale is used for scaling numerical variables.
  - ➢Checked the correlation among variables using heatmap(it is difficult to analyze due to lot of variables.).

MODEL BUILDING

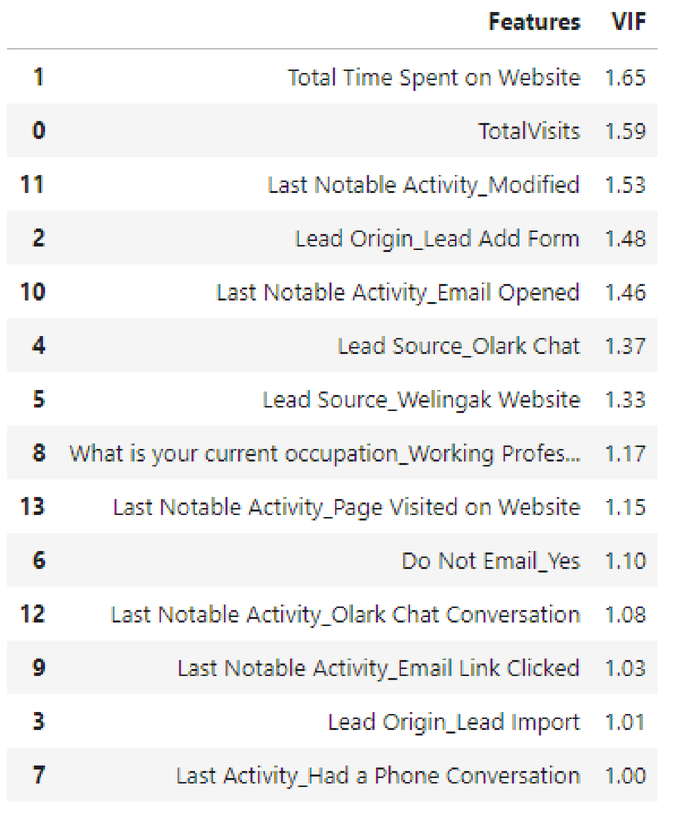
- Imported the 'Logistic Regression' and creating a Logistic Regression object.
- Used Recursive Feature Elimination for feature selection.
- Checked Variance Inflation Factor (VIF) for all variables present in the model.
- Eliminated variables with high p-value (p-value >0.05) or VIF (VIF > 5).
- Repeated the above steps until all the variables were within the acceptable range.



Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6336
Model Family:	Binomial	**Df Model:**	14
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2665.3
Date:	Wed, 16 Oct 2024	**Deviance:**	5330.6
Time:	01:14:49	**Pearson chi2:**	6.41e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3896
Covariance Type:	nonrobust		

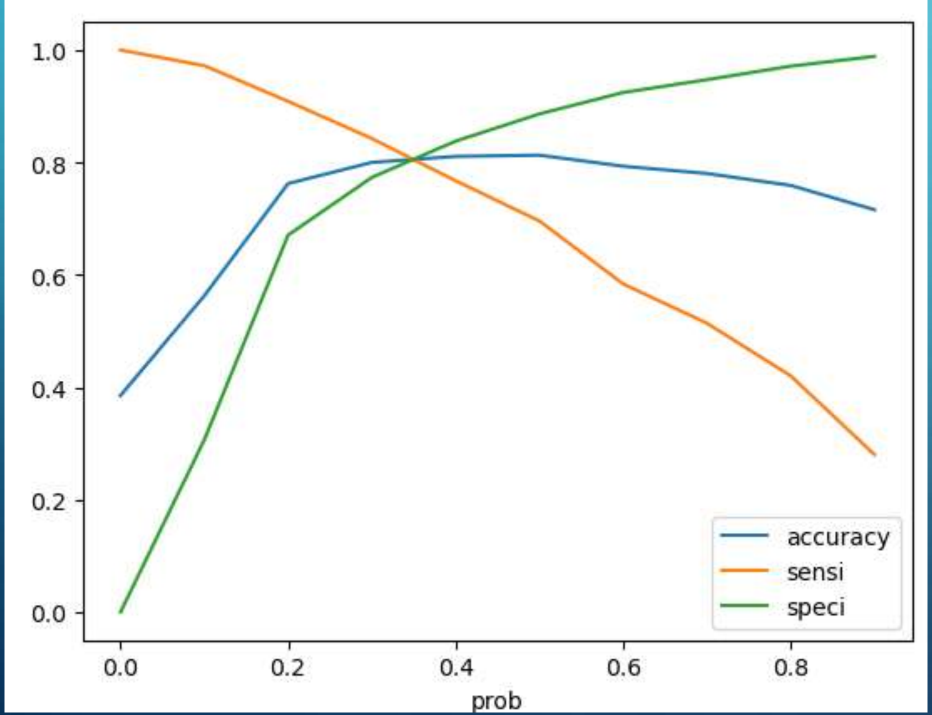
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.9512	0.090	-10.517	0.000	-1.128	-0.774
TotalVisits	7.8903	2.186	3.609	0.000	3.606	12.175
Total Time Spent on Website	4.6715	0.166	28.156	0.000	4.346	4.997
Lead Origin_Lead Add Form	4.1987	0.227	18.524	0.000	3.754	4.643
Lead Origin_Lead Import	1.7255	0.463	3.725	0.000	0.818	2.633
Lead Source_Olark Chat	1.2197	0.107	11.412	0.000	1.010	1.429
Lead Source_Welingak Website	2.1157	0.761	2.778	0.005	0.623	3.608
Do Not Email_Yes	-1.8618	0.179	-10.382	0.000	-2.213	-1.510
Last Activity_Had a Phone Conversation	2.0345	0.811	2.508	0.012	0.445	3.624
What is your current occupation_Working Professional	2.7393	0.188	14.597	0.000	2.371	3.107
Last Notable Activity_Email Link Clicked	-1.8671	0.256	-7.296	0.000	-2.369	-1.365
Last Notable Activity_Email Opened	-1.4102	0.088	-15.947	0.000	-1.583	-1.237
Last Notable Activity_Modified	-2.1167	0.092	-22.961	0.000	-2.297	-1.936
Last Notable Activity_Olark Chat Conversation	-2.7995	0.328	-8.544	0.000	-3.442	-2.157
Last Notable Activity_Page Visited on Website	-1.8898	0.212	-8.918	0.000	-2.305	-1.474

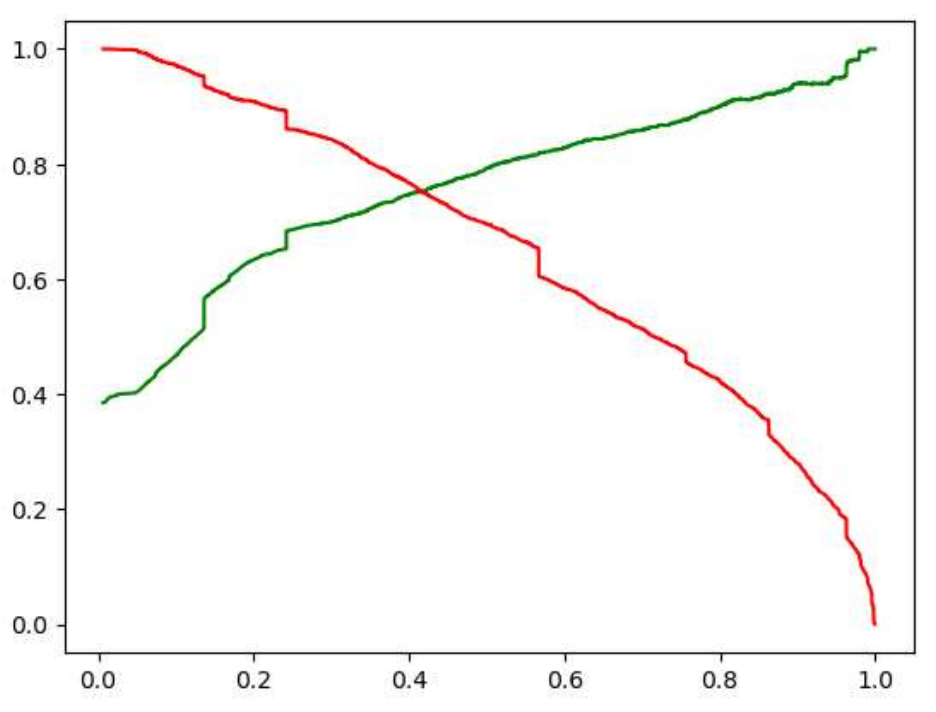


	Features	VIF
1	Total Time Spent on Website	1.65
0	TotalVisits	1.59
11	Last Notable Activity_Modified	1.53
2	Lead Origin_Lead Add Form	1.48
10	Last Notable Activity_Email Opened	1.46
4	Lead Source_Olark Chat	1.37
5	Lead Source_Welingak Website	1.33
8	What is your current occupation_Working Profes...	1.17
13	Last Notable Activity_Page Visited on Website	1.15
6	Do Not Email_Yes	1.10
12	Last Notable Activity_Olark Chat Conversation	1.08
9	Last Notable Activity_Email Link Clicked	1.03
3	Lead Origin_Lead Import	1.01
7	Last Activity_Had a Phone Conversation	1.00

MODEL EVALUATION

- Used 'predict' to predict the probability on the train set.
- Created a dataframe with the actual conversion flag and the predicted probability.
- Created a new column 'Predicted' with 1 if Paid_Prob > 0.5 else 0.
- Created the confusion matrix and checked the Accuracy, Sensitivity and Specificity.
- Determined the optimal cut-off using ROC and Precision-Recall tradeoff.





MODEL OBSERVATIONS

- We can observe a difference in Optimal cut-offs calculated using ROC and Precision-Recall tradeoff.
- Since we are more focused on maximizing the identification True positives we'll consider optimal cut-off obtained through Precision-Recall tradeoff which is 0.44.

Model Evaluation			
Training dataset		**Testing dataset**	
Accuracy	81.01%	**Accuracy**	80.90%
Precision	76.38%	**Precision**	74.20%
Recall	73.38%	**Recall**	72.69%

Features
Total Time Spent on Website
TotalVisits
Last Notable Activity_Modified
Lead Origin_Lead Add Form
Last Notable Activity_Email Opened
Lead Source_Olark Chat
Lead Source_Welingak Website
What is your current occupation_Working Profes...
Last Notable Activity_Page Visited on Website
Do Not Email_Yes
Last Notable Activity_Olark Chat Conversation
Last Notable Activity_Email Link Clicked
Lead Origin_Lead Import
Last Activity_Had a Phone Conversation

CONCLUSION

- Most of the Leads are generated from Landing Page Submission, followed by APIs where as ‘Lead Add form’ is a great origin to identify leads as it has the highest percentage of conversions.
- Most Source of Leads are coming from Google followed by Direct Traffic.
- Most Leads generated are from India.
- Most Leads are looking for a course to get better career prospects.
- Highest number of leads generated are ‘Unemployed’ whereas highest percentage of conversion Is among the leads who are ‘Working Professionals’.
- Leads whose last notable activity was ‘SMS sent’ have highest percentage of conversion. Leads with last notable activity as ‘E-mail opened’ also have a high conversion rate.
- Leads identified through sources ‘Reference’ and ‘Welingak Website’ have pretty high conversion rates. Highest number of leads are being generated through ‘Google’ followed by ‘Direct Traffic’.
- Leads who have adopted to receive calls or e-mails have high rate of conversion.
- Not many leads are being generated through recommendations.