



Lead Scoring Case Study


Submitted by :
Simran
Jyoti

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. They have process of form filling on their website after which the company that individual as a lead.
- ▶ Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.
- ▶ The typical lead conversion rate at X education is around **30%**. Now, this means if, say, they acquire 100 leads in a day, only about 30 of them are converted. 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 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

- ▶ Lead X wants us to build a model to give every lead a lead score between 0 -100 . So that they can identify the Hot leads and increase their conversion rate as well.
- ▶ The CEO want to achieve a lead conversion rate of 80%.
- ▶ They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full man power and after achieving target what should be the approaches.



Problem Approach

- ▶ **Importing the data and inspecting the data frame**
- ▶ **Data Understanding and Cleaning**
- ▶ **Outliner Analysis**
- ▶ **EDA**
- ▶ **Dummy variable creation**
- ▶ **Test-Train split**
- ▶ **Feature scaling**
- ▶ **Correlations**
- ▶ **Model Building (RFE)**
- ▶ **Model Evaluation**
- ▶ **Conclusion and recommendation**

Data Understanding and Cleaning

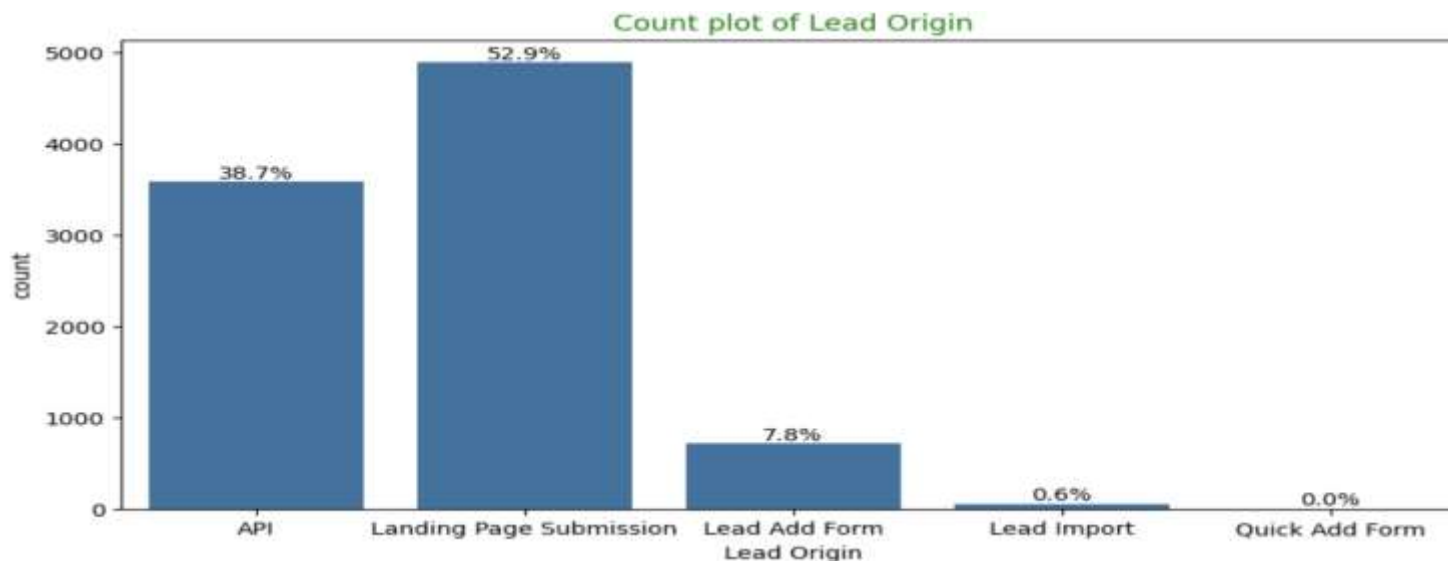
- ▶ - Many of the categorical variables have a level called select which needs to be handled. Customer did not select any option from the list for such columns the data remained as default 'Select' for Select.
- ▶ Replaced the missing values
- ▶ Dropping the columns which has more number of missing values and which would not add onto the model prediction

```
TotalVisits      1.48
Page Views Per Visit 1.48
Prospect ID      0.00
Magazine         0.00
A free copy of Mastering The Interview 0.00
I agree to pay the amount through cheque 0.00
Get updates on DM Content 0.00
Update me on Supply Chain Content 0.00
Receive More Updates About Our Courses 0.00
Through Recommendations 0.00
Digital Advertisement 0.00
Newspaper        0.00
X Education Forums 0.00
Newspaper Article 0.00
Search           0.00
Lead Number      0.00
What is your current occupation 0.00
Specialization   0.00
Last Activity    0.00
Total Time Spent on Website 0.00
Converted        0.00
Do Not Call      0.00
Do Not Email     0.00
Lead Source      0.00
Lead Origin      0.00
Last Notable Activity 0.00
dtype: float64
```

Univariate Analysis – Outliers

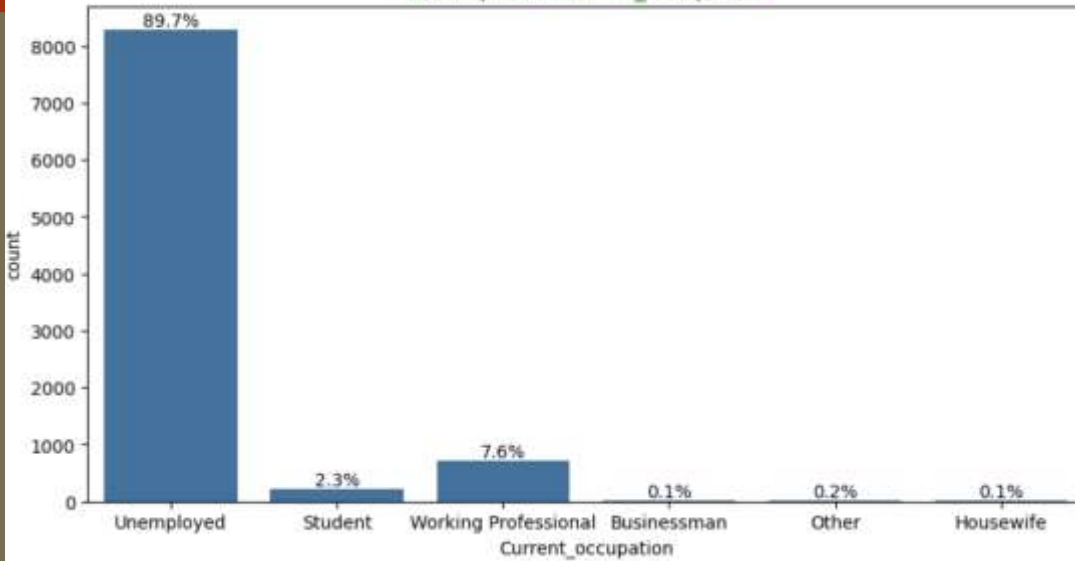
List of features from variables which are present in majority (Converted and Not Converted included)

- ▶ Lead Origin: "Landing Page Submission" identified 53% customers, "API" identified 39%.
- ▶ Current_occupation: It has 90% of the customers as Unemployed
- ▶ Do Not Email: 92% of the people has opted that they dont want to be emailed about the course.
- ▶ Lead Source: 58% Lead source is from Google & Direct Traffic combined
- ▶ Last Activity: 68% of customers contribution in SMS Sent & Email Opened activities

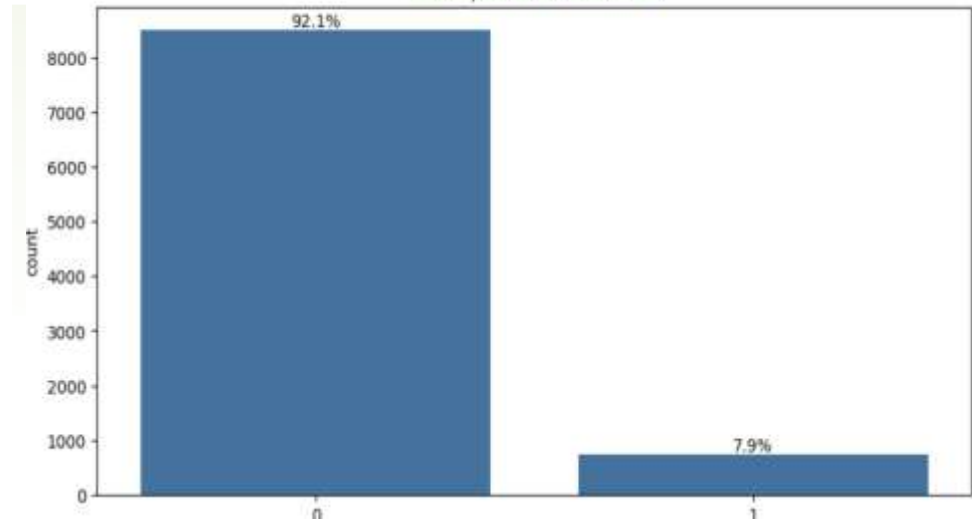


Univariate Analysis – Outliers

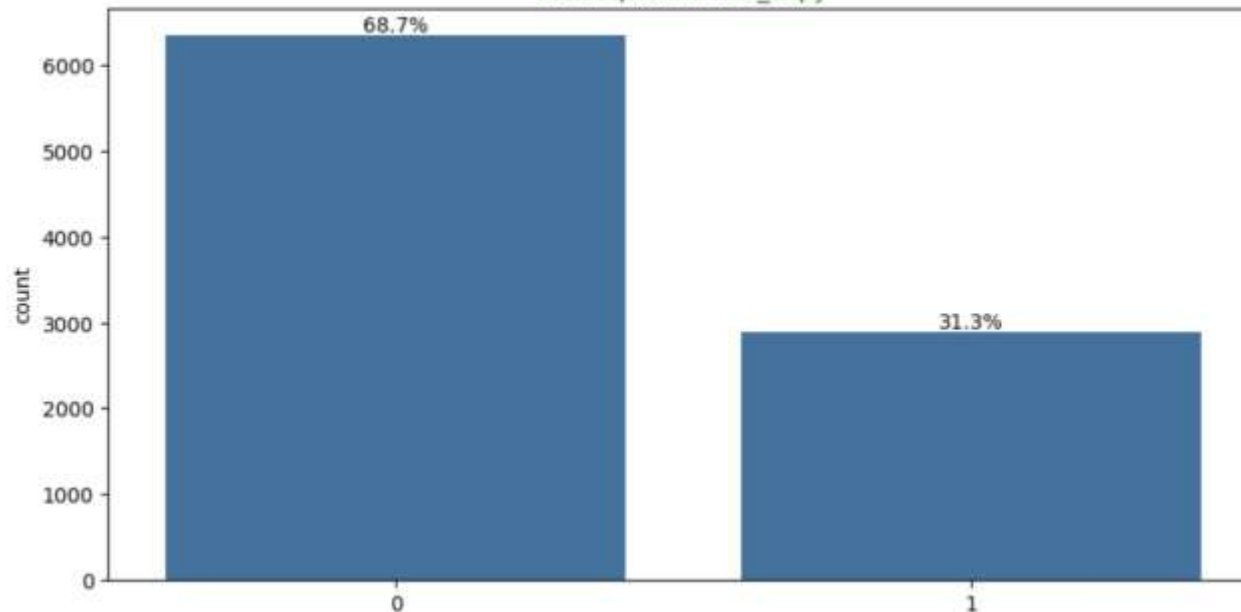
Count plot of Current_occupation



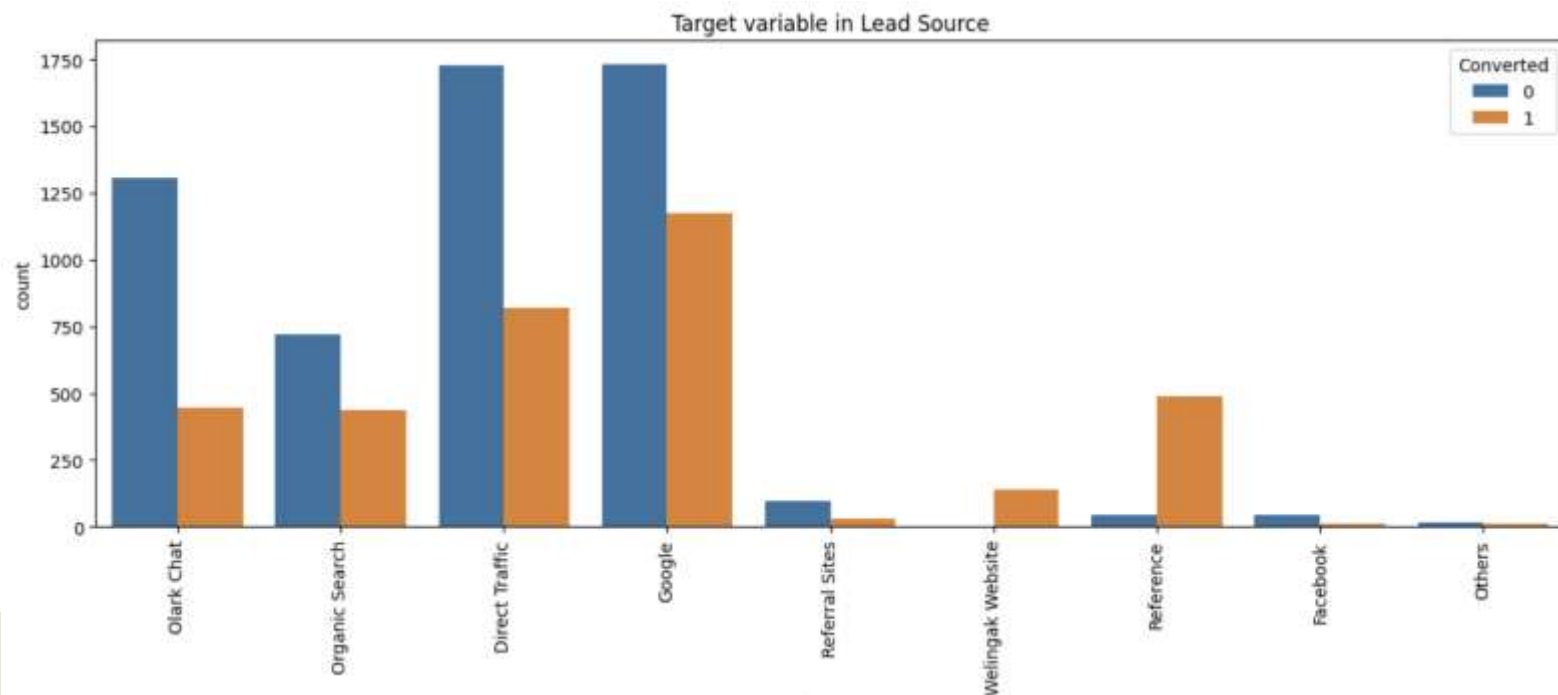
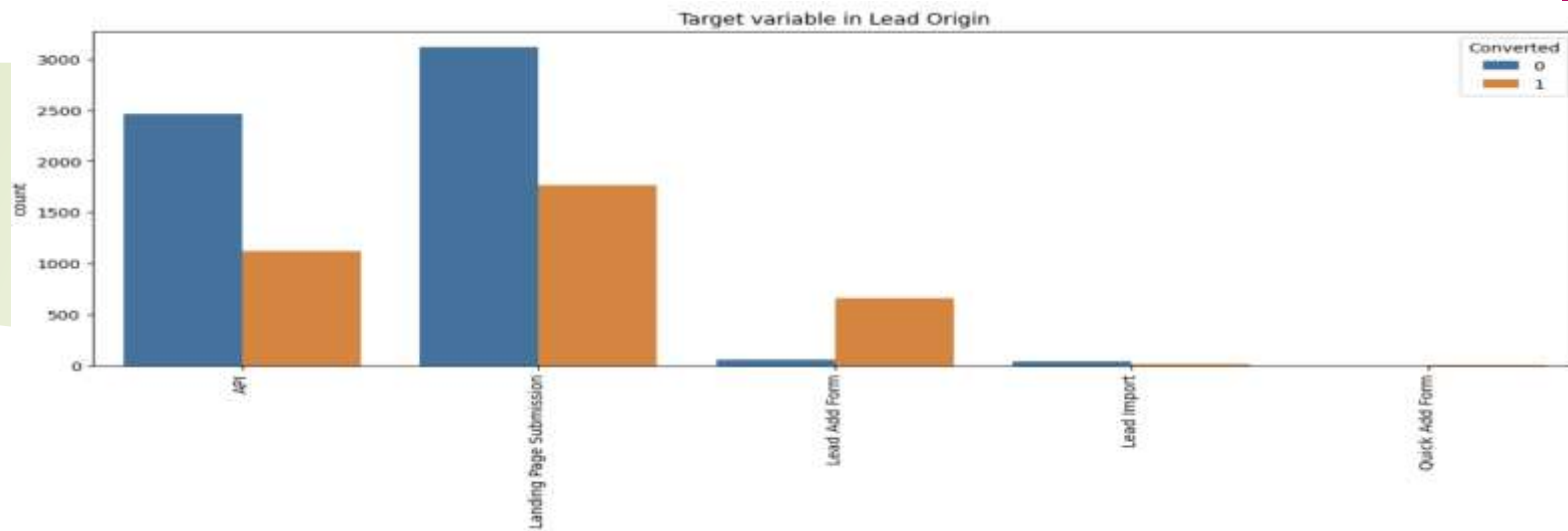
Count plot of Do Not Email



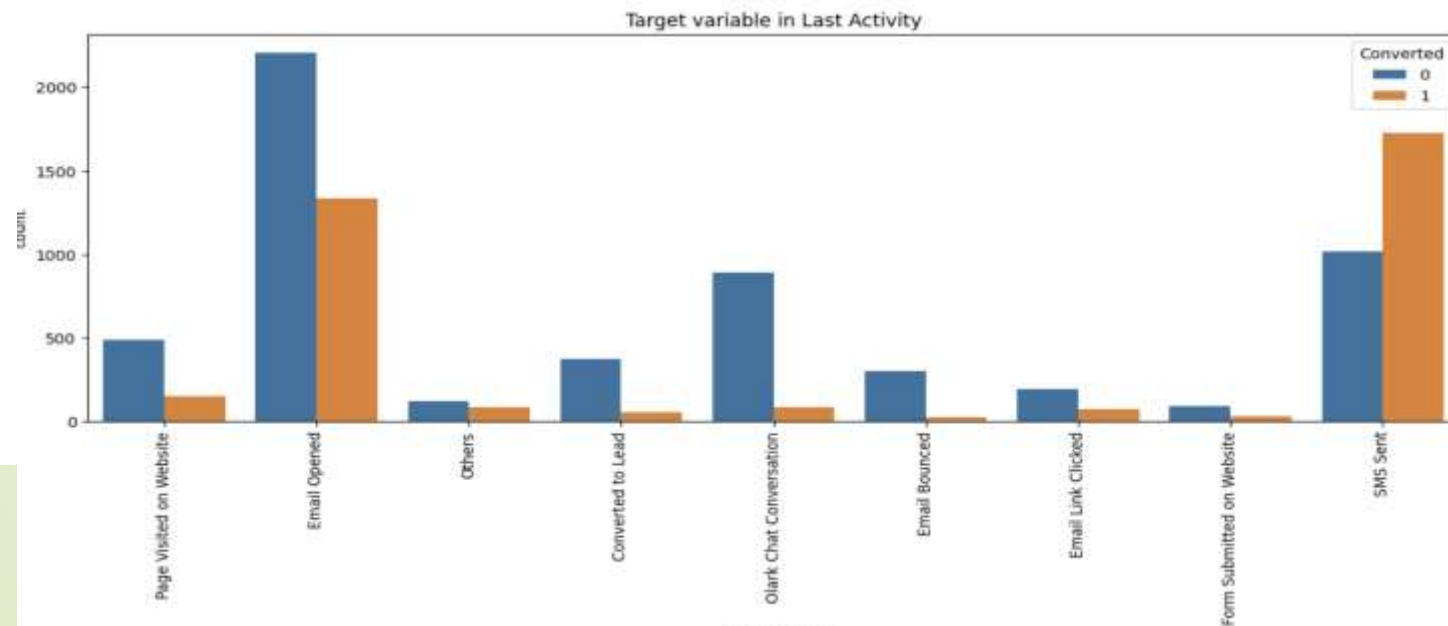
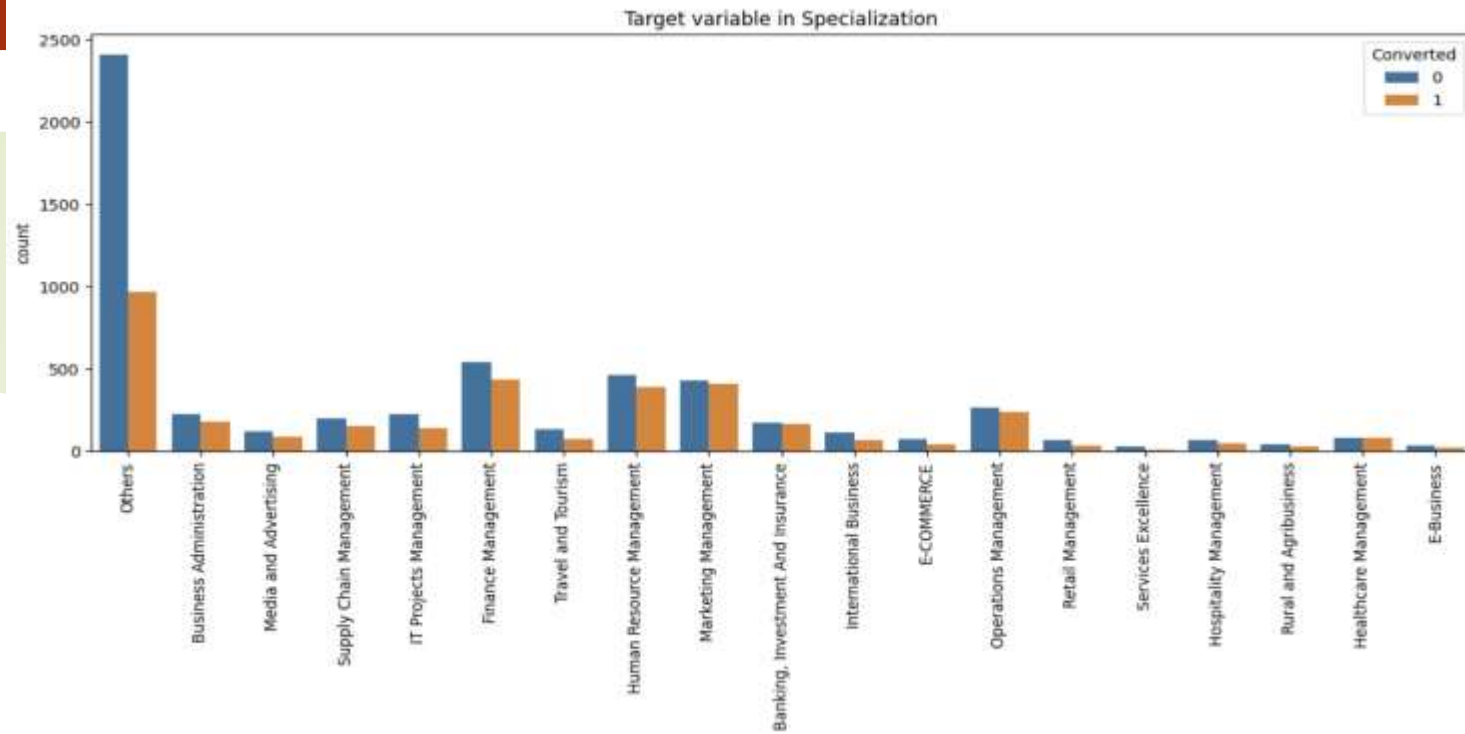
Count plot of Free_copy



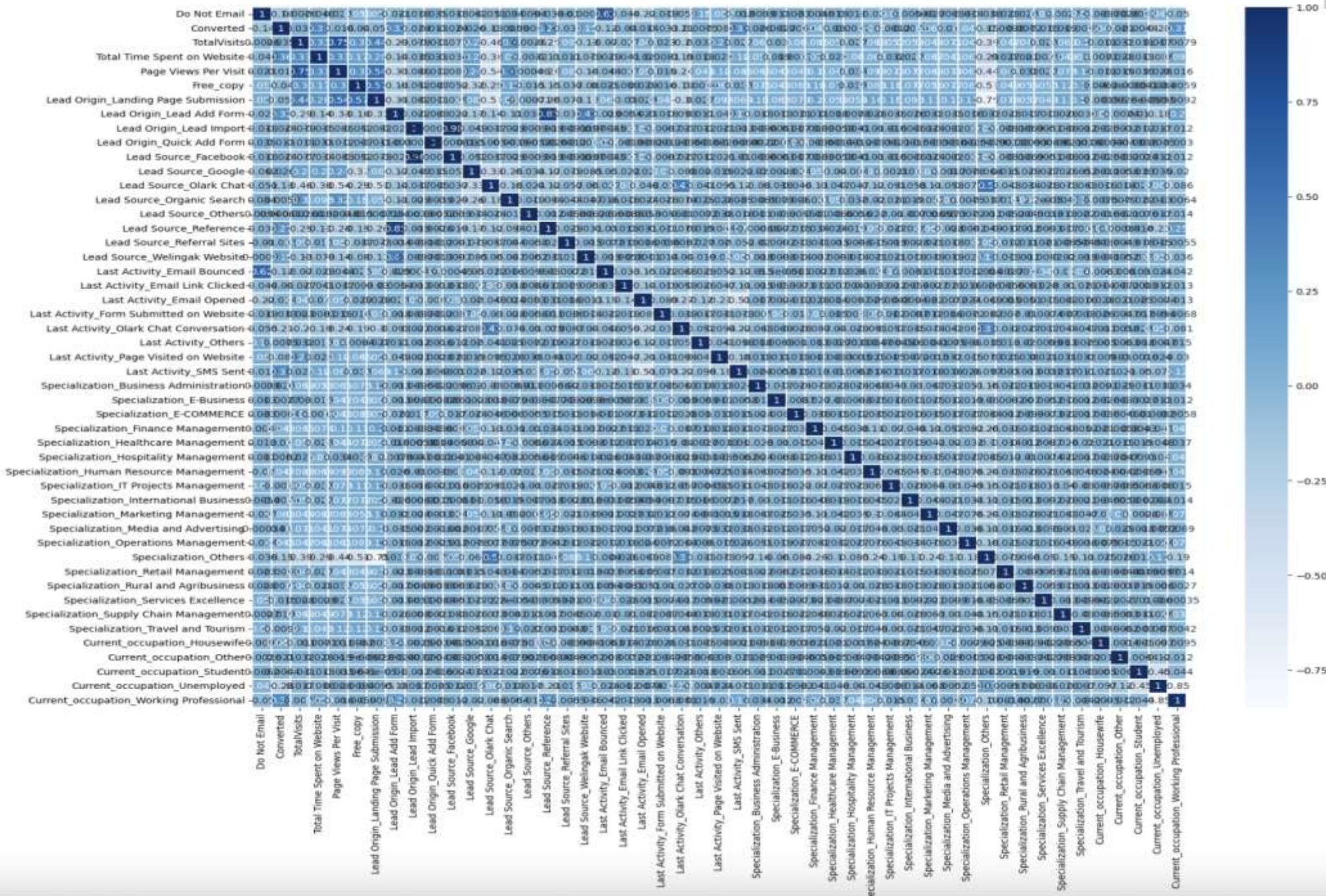
Bivariate Analysis – Outliers



Bivariate Analysis – Outliers



Bivariate Analysis – Checking correlation



Dummy variable creation

- Independent variables as dummy variables allows easy interpretation and calculation of the odds ratios, which increases the stability and significance of the coefficients.
- Dummy variables have been created for following columns:
 1. Lead Origin
 2. Lead Source
 3. Last Activity
 4. Specialization
 5. What is your current occupation

Test-Train split

➤ Train – Test Split:

- The modified 'Leads' dataset has been split into Train and test dataset in the ratio 70:30.
- Train dataset has been used to train the model whereas Test dataset has been used to evaluate the model

➤ Feature Scaling:

- It is important to have all variables on the same scale in order to avoid the dominance of variables with high magnitude in the model.
- "StandardScaler" function has been used to scale the data for modeling which brings all the data points into a standard normal distribution with mean at '0' and standard deviation at '1'.

Model Building

We choose the model for following reasons :

- p-values for all variables is less than 0.05
- This model looks acceptable as everything is under control (p-values & VIFs).

Generalized Linear Model Regression Results

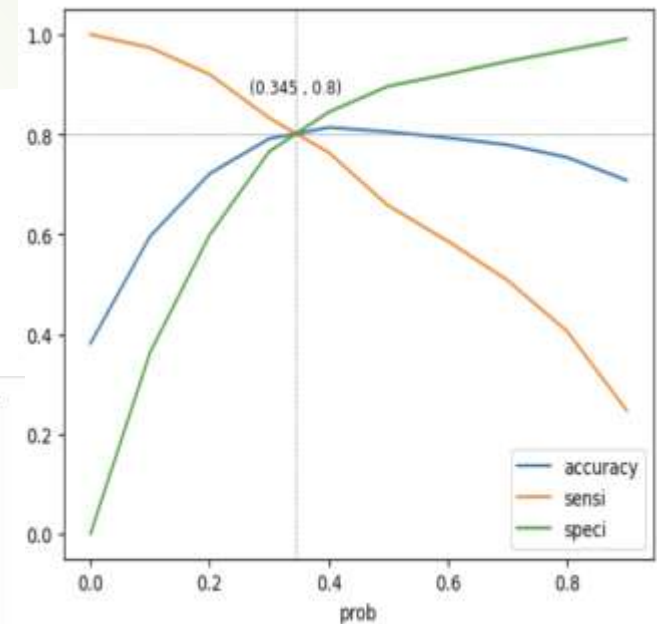
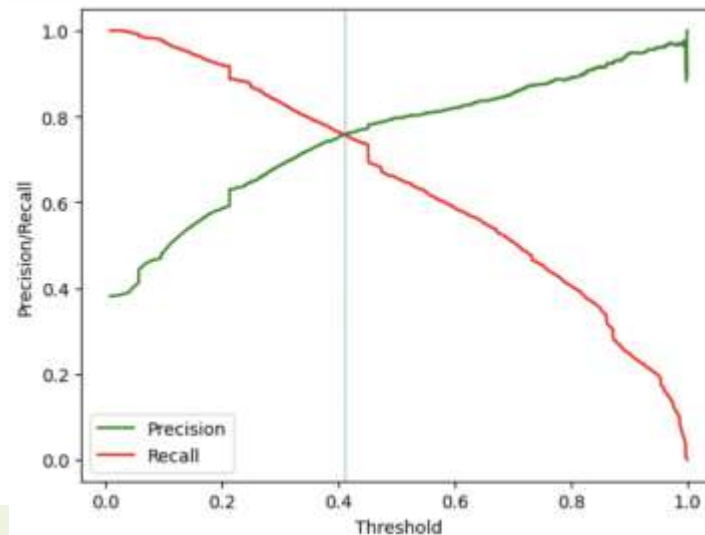
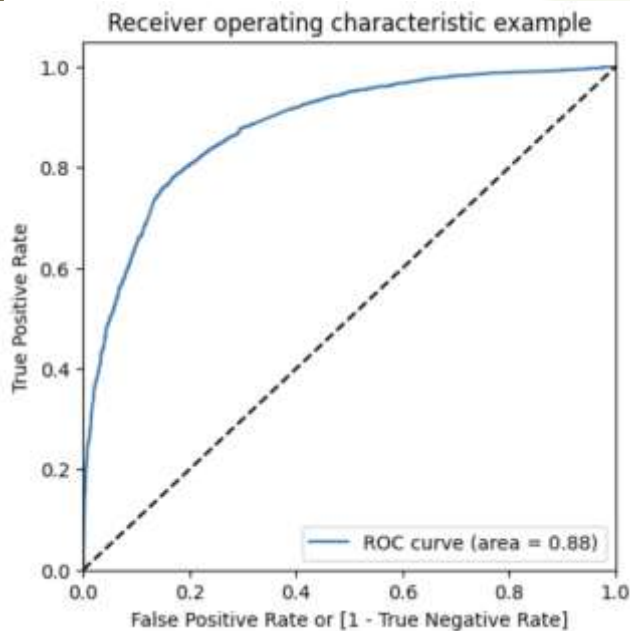
Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6455
Model Family:	Binomial	Df Model:	12
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2743.1
Date:	Tue, 19 Nov 2024	Deviance:	5486.1
Time:	16:37:14	Pearson chi2:	8.11e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3819
Covariance Type:	nonrobust		

	coef	std err	z	P> z	[0.025	0.975]
const	-1.0236	0.143	-7.145	0.000	-1.304	-0.743
Total Time Spent on Website	1.0498	0.039	27.234	0.000	0.974	1.125
Lead Origin_Landing Page Submission	-1.2590	0.125	-10.037	0.000	-1.505	-1.013
Lead Source_Olark Chat	0.9072	0.118	7.701	0.000	0.676	1.138
Lead Source_Reference	2.9253	0.215	13.615	0.000	2.504	3.346
Lead Source_Welingak Website	5.3887	0.728	7.399	0.000	3.961	6.816
Last Activity_Email Opened	0.9421	0.104	9.022	0.000	0.737	1.147
Last Activity_Olark Chat Conversation	-0.5556	0.187	-2.974	0.003	-0.922	-0.189
Last Activity_Others	1.2531	0.238	5.259	0.000	0.786	1.720
Last Activity_SMS Sent	2.0519	0.107	19.106	0.000	1.841	2.262
Specialization_Hospitality Management	-1.0944	0.323	-3.391	0.001	-1.727	-0.462
Specialization_Others	-1.2033	0.121	-9.950	0.000	-1.440	-0.966
Current_occupation_Working Professional	2.6697	0.190	14.034	0.000	2.297	3.042

Model Evaluation

We choose the model for following reasons :

- p-values for all variables is less than 0.05
- This model looks acceptable as everything is under control (p-values & VIFs).



Observations

Train Data Set:

Accuracy: 80.46%

Sensitivity: 80.05%

Specificity: 80.71%

Test Data:

Accuracy: 80.34%

Sensitivity: 79.82% \approx 80%

Specificity: 80.68%

Recommendation

- ▶ Leads with high 'lead score' can be more focused on.
- ▶ Marketing from Google has the highest conversion rates as the traffic volume is high
- ▶ Encouraging referrals with exciting incentives
- ▶ Mumbai has the major leads, can focus on marketing to other cities to achieve higher results.
- ▶ Unemployed category has the focus, and finance with specialization
- ▶ Students having the least rate.