Lead Score Case Study

CHAITANYA UGALE
KAPILA GAUR

Problem Statement

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, 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

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

Lead Conversion Process

OLD PROCESS

Lead Generation → Visit To Website by potential Customers → Visitors Either Provide E-mail, Contact Details → Calling/ Emailing to All Leads → Conversion Score Of Leads : **39%**

NEW PROCESS

Lead Generation → Visit To Website by potential Customers → Visitors Either Provide E-mail, Contact Details → **Creating Model To Filter Leads** → Conversion Score Of Leads : **79%**

CLEANING THE DATA

TOTAL CURRENT APPLICATION DATA COLUMN: 37

NULL VALUES

17

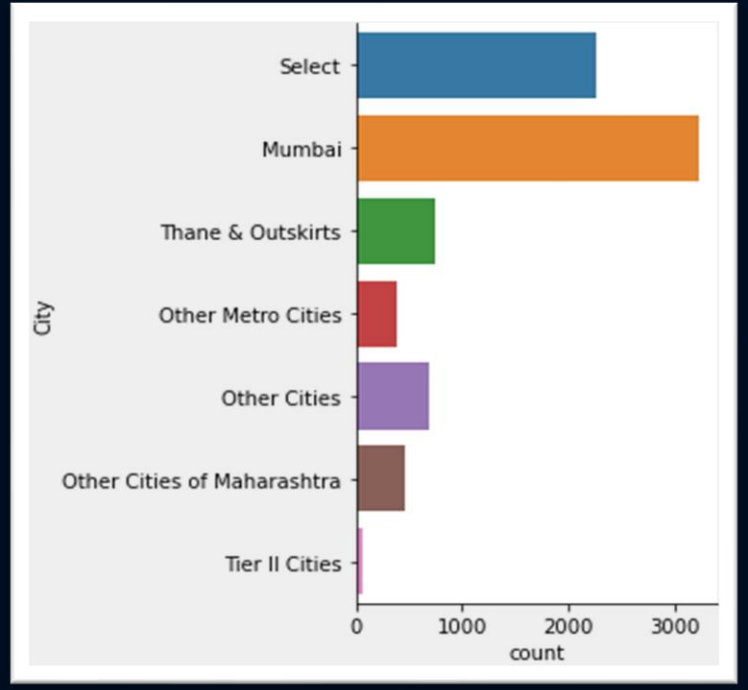
MORE THAN 30% NULL VALUE

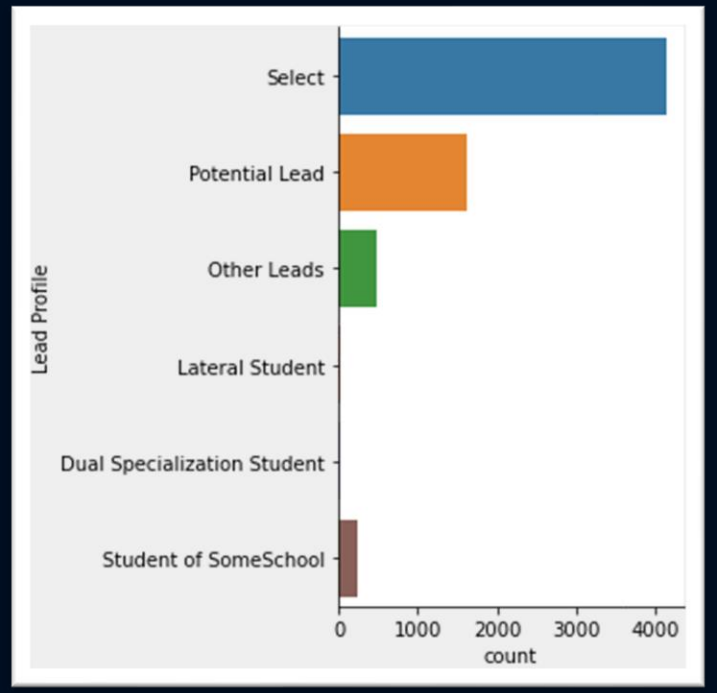
6

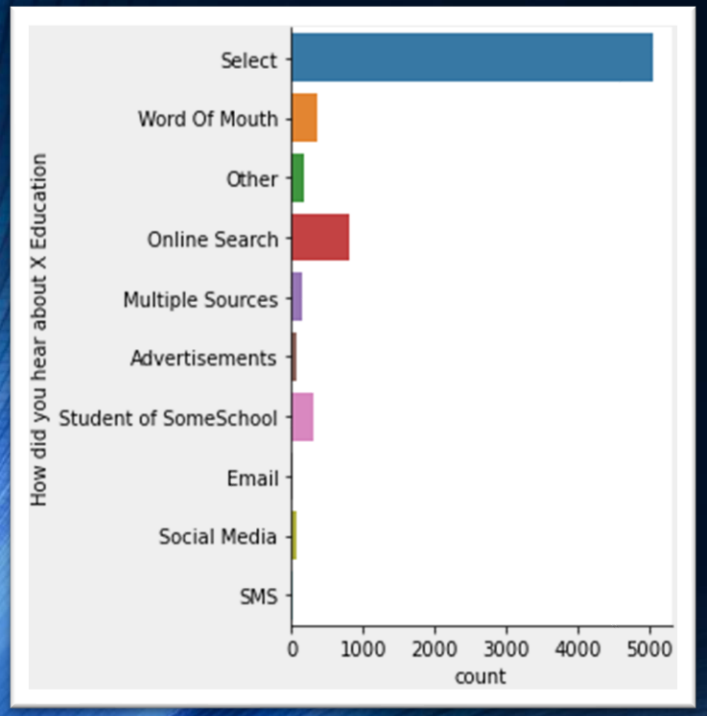
UNWANTED COLUMNS

24

EDA





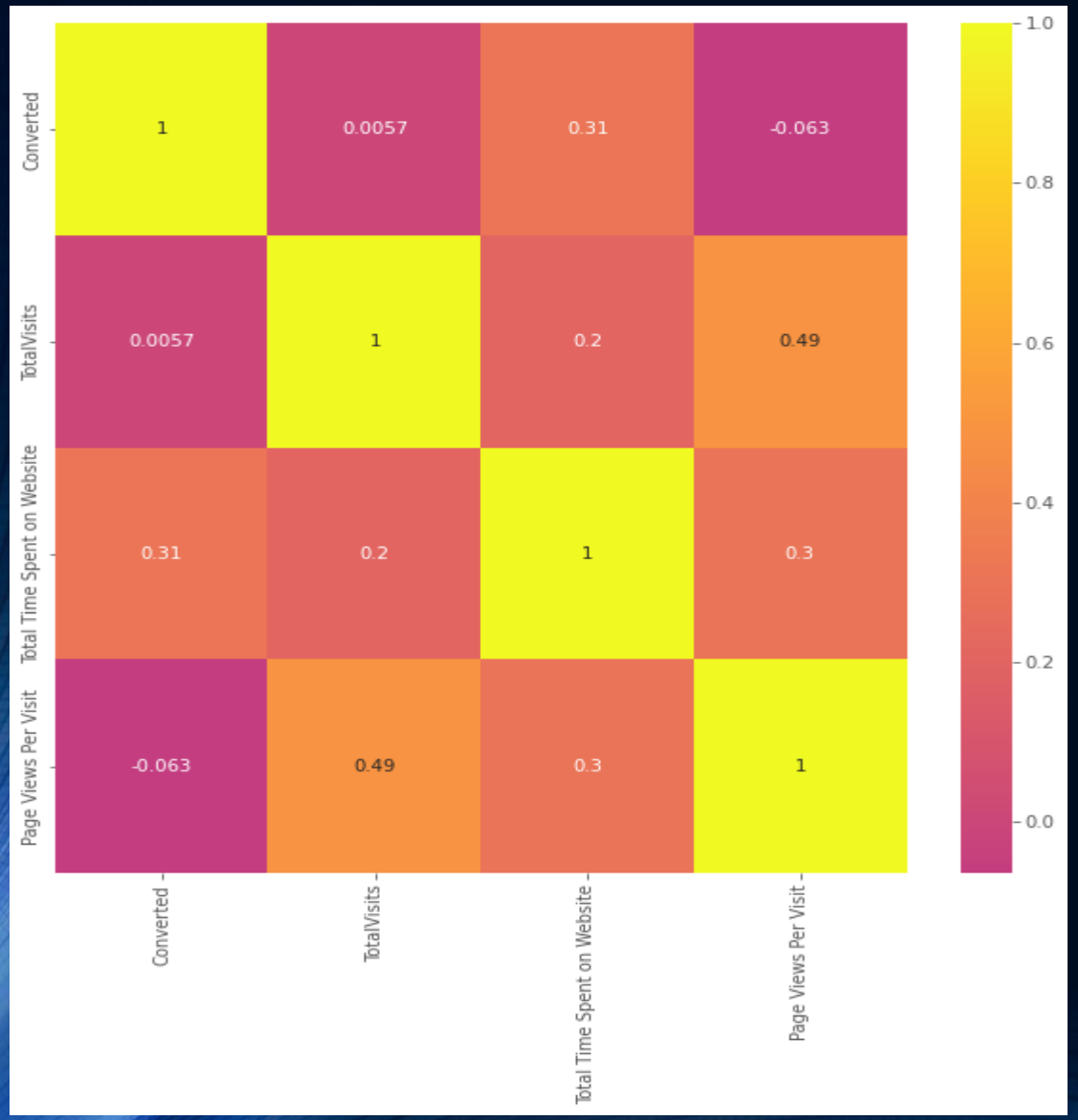


Categorical Variable

There are 4 columns where SELECT values are more available which is due to the mentioned variables are not selected.

For this analysis we have removed the mentioned columns

Correlation

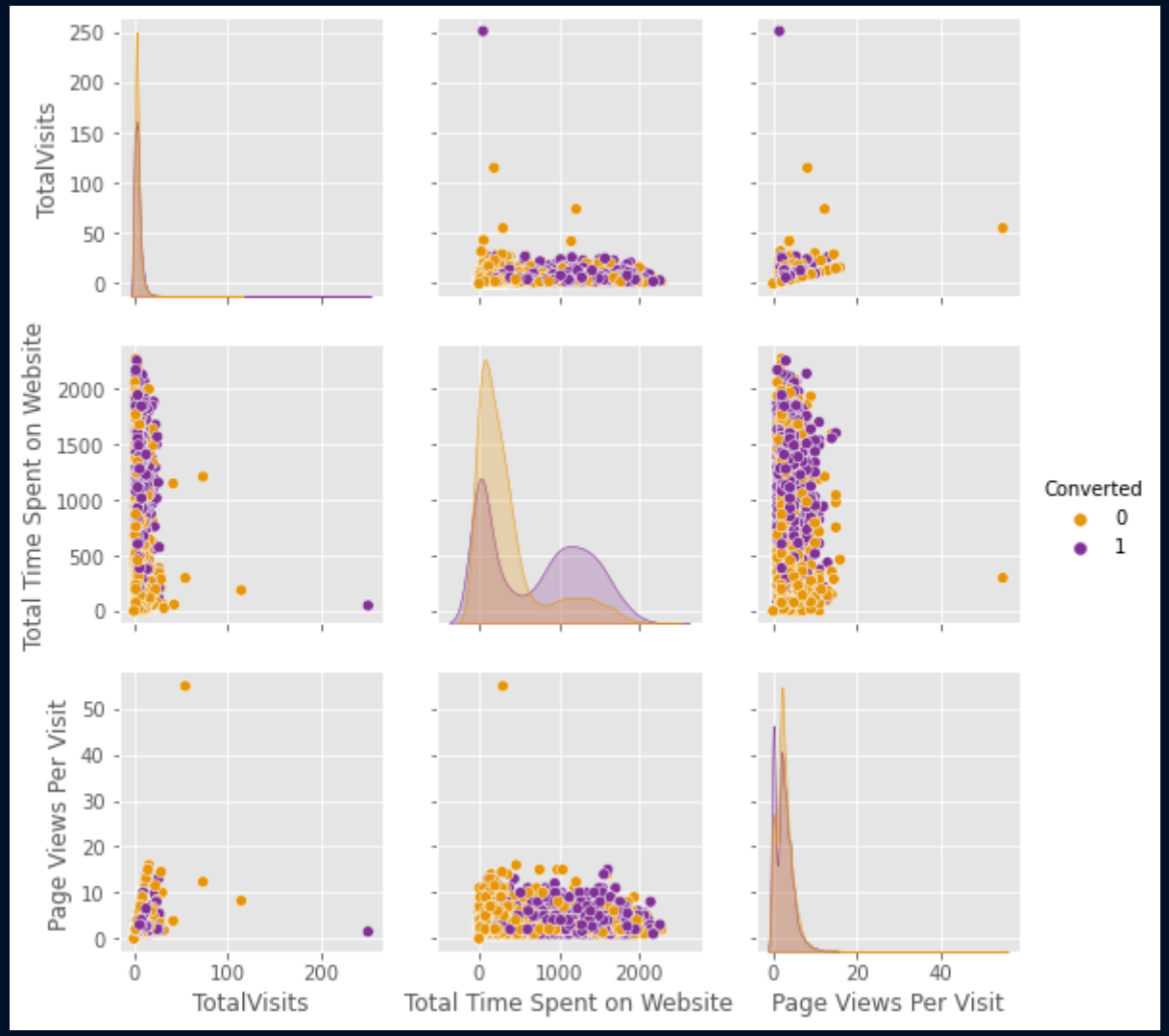


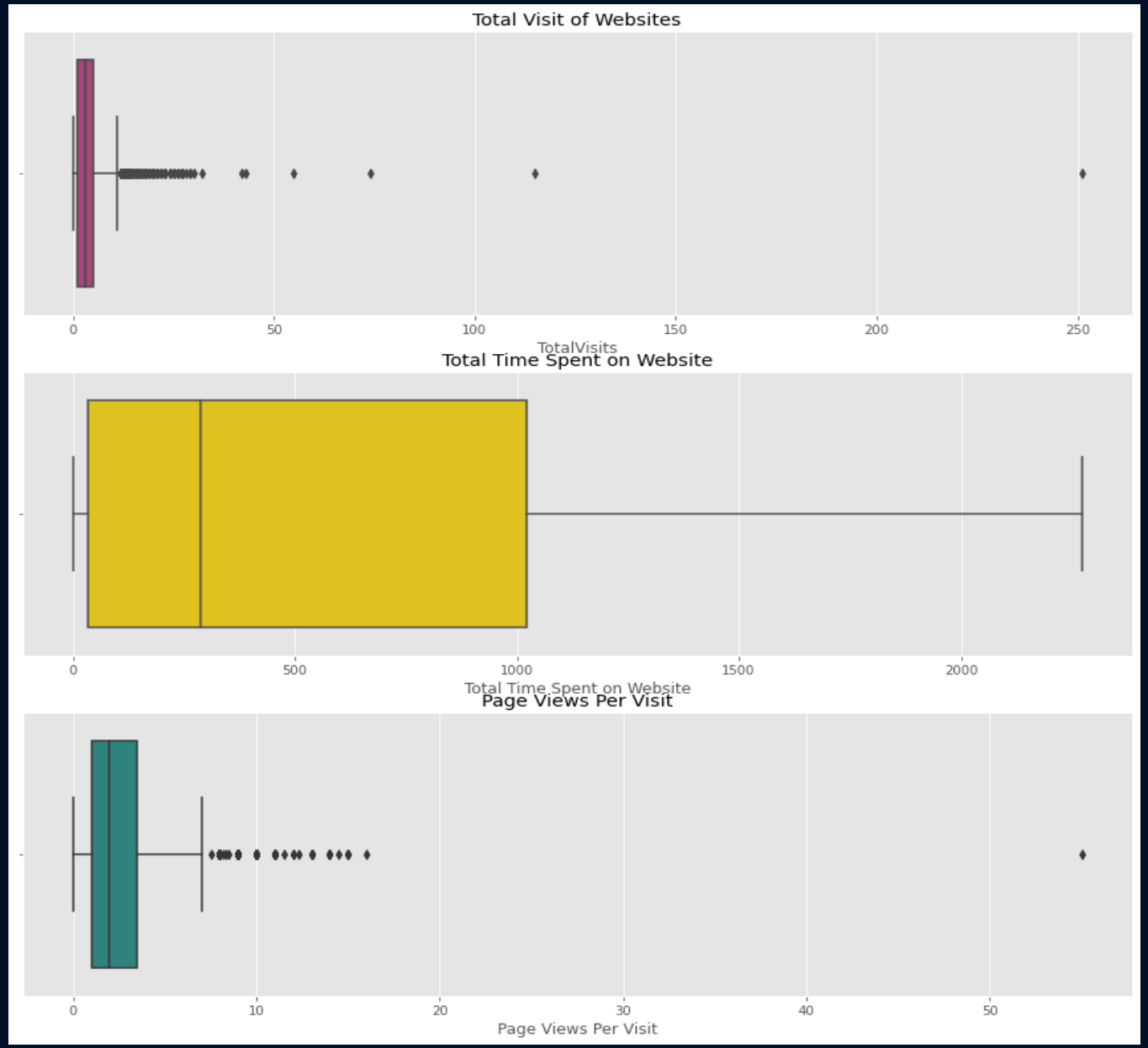
TotalVisits are **highly correlated** with **Page Views Per Visit** with value **0.49**

Page View Per Visit has **Negative correlation** with **Converted** with value **-0.063**

Total Time Spent On Website has **Positive correlation** with **Converted** with value **0.3**

Pair Plot based on Conversion





MAX VISIT OF WEBSITE

60K

OUTLIERS

251

MEDIAN

3

MAX TIME SPENT ON WEBSITE

2272

OUTLIERS

NA

MEDIAN

287

MAX PAGE VIEW PER VISIT

55

OUTLIERS

55

MEDIAN

2

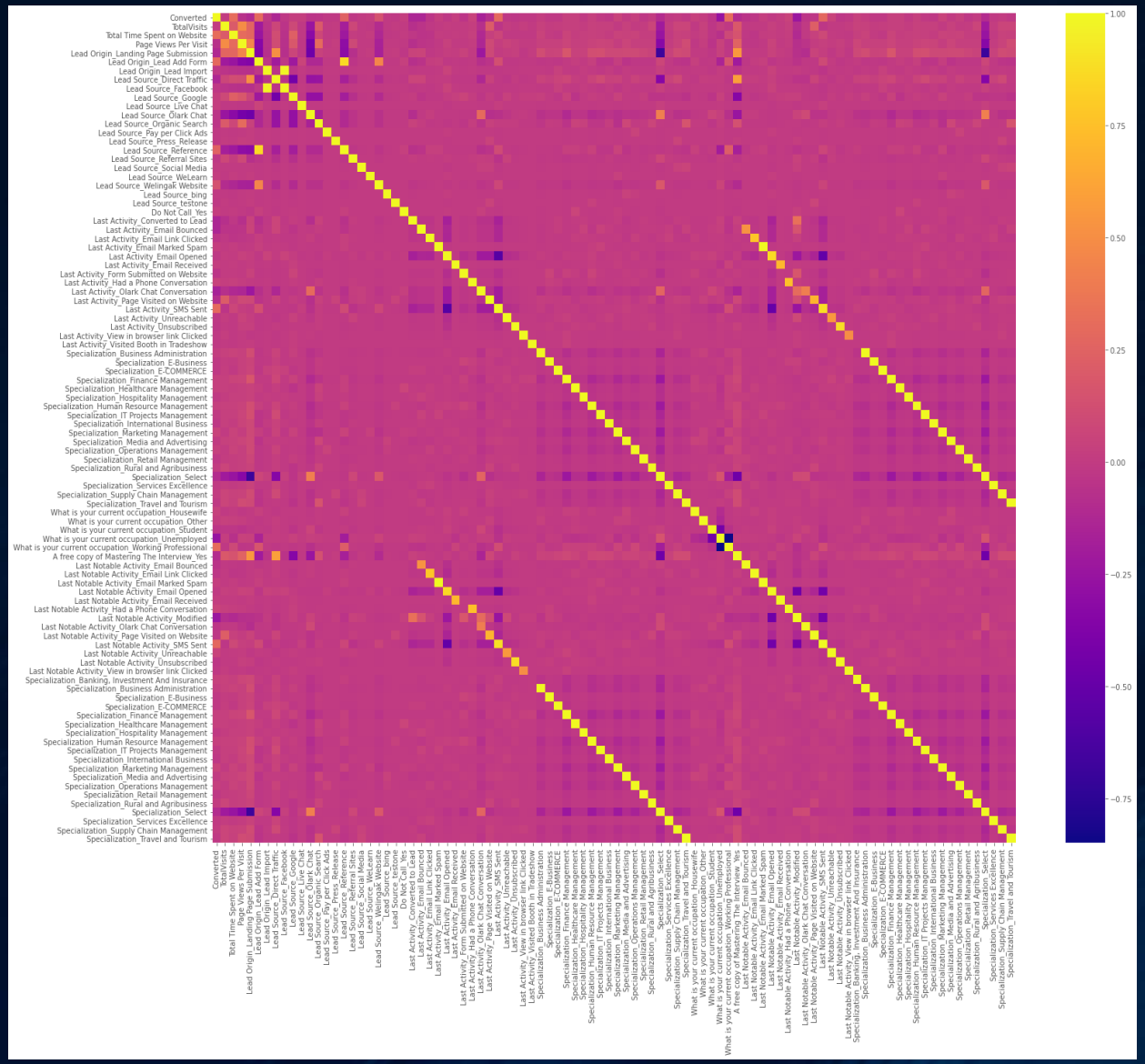
MODEL BUILDING

AND EVALUATION

STEPS

- Split the data into Training and Test datasets
- Build Logistic Regression Model
- Feature selection using RFE
- Assessing the model with **StatsModels**
- Create a Data frame with the actual Conversion Flag and the predicted probabilities
- Create new column ‘Predicted' with **1** if **Conversion Probability > 0.5** else **0**
- Check VIFs and **Drop columns** based on **higher VIF/higher P** values and check metrics after each column drop

Correlation of Models

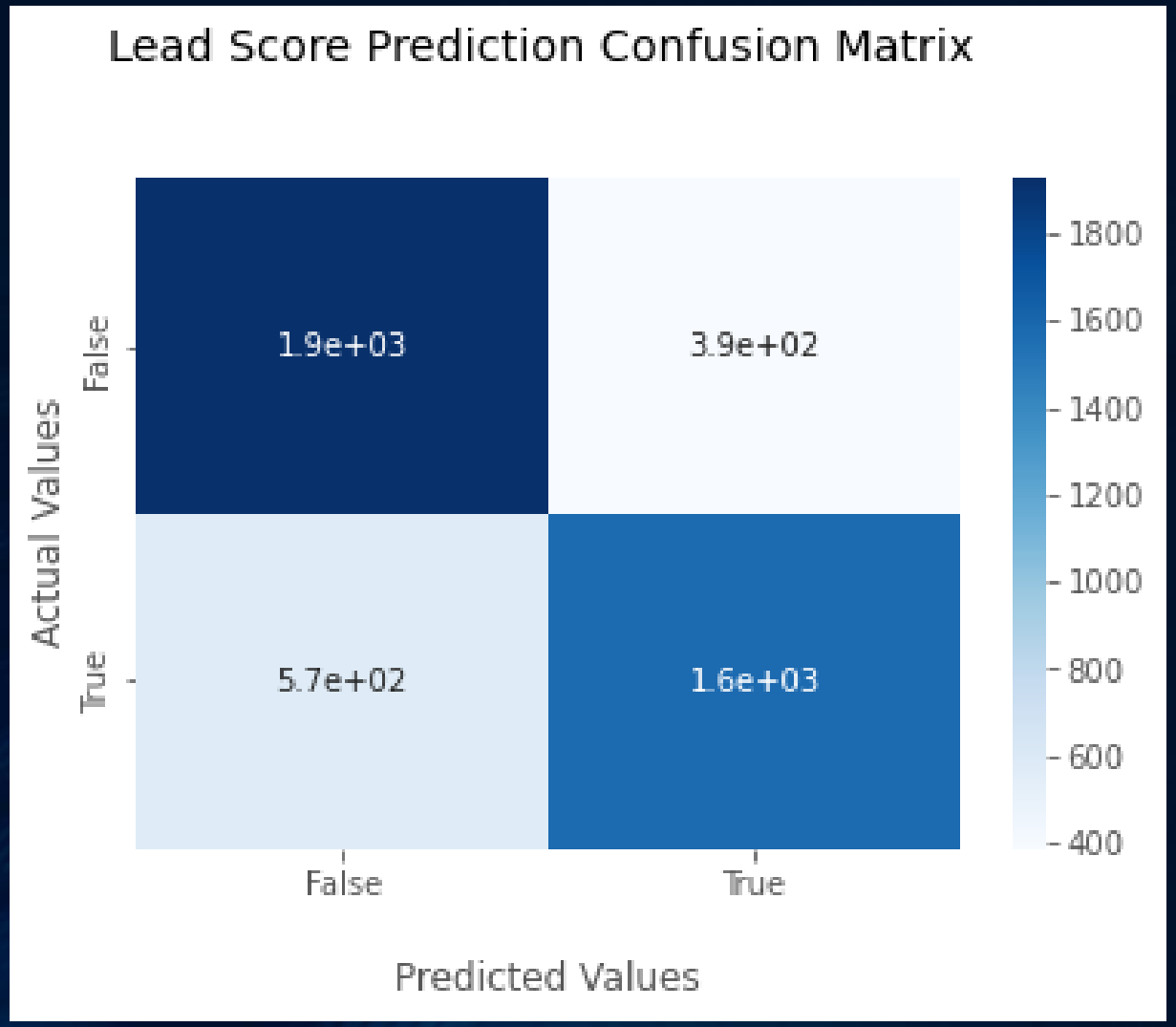


Lead Origin_Landing Page Submission are highly correlated with **Lead Source_Direct Traffic** with value **0.50**

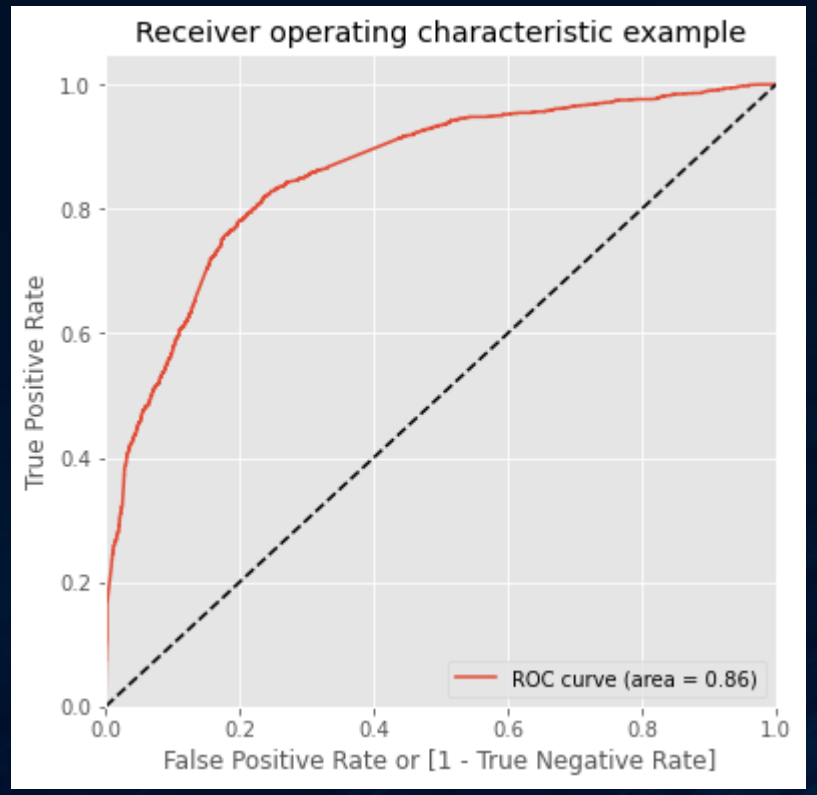
Specialization_Select has Negative correlation with **Lead Origin_Landing Page Submission** with value **-0.688**

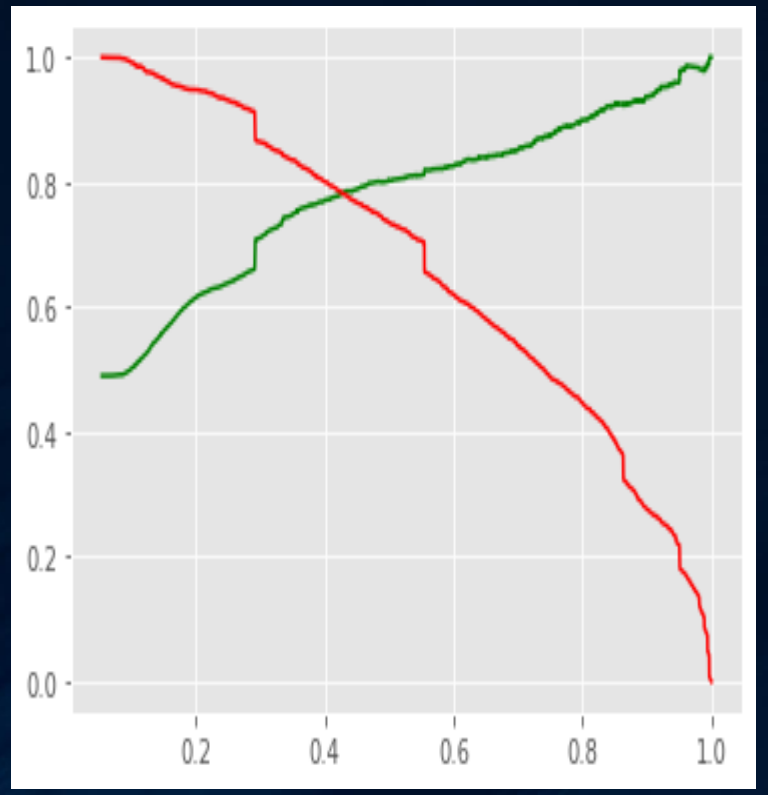
Lead Origin_Landing Page Submission has positive correlation with **Lead Source_Direct Traffic** with value **0.50**

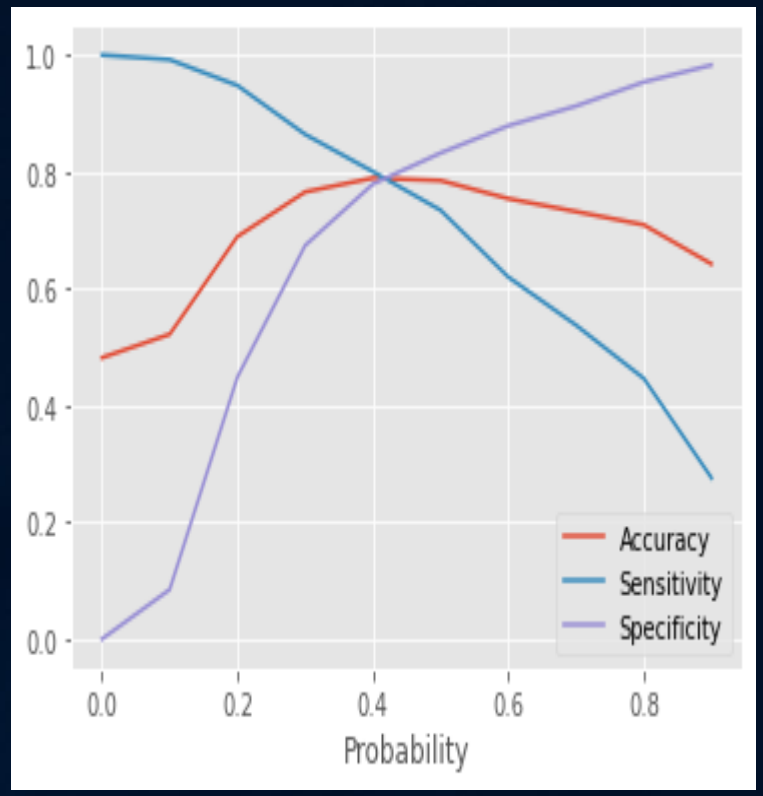
CONFUSION METRICS



MODEL PREDICTION







AREA UNDER ROC CURVE

0.86

PRECISION RECALL CUTOFF

0.44

PRECISION RECALL CUTOFF

0.42

MODEL PREDICTION

ACCURACY : 79%

PRECISION : 77%

SENSITIVITY : 78.4%

RECALL : 80%

SPECIFICITY : 79.6%

SUMMARY

Variables which contributed most for leads conversion

- Lead Origin_Lead Add Form
- Total Time Spent on Website
- Total Visits

Business need to focus

- Last Activity_Had a Phone Conversation
- What is your current occupation_Working Professional
- Total Visits

SUMMARY

- Focus on wider set of lead audience (inclusion of slightly lower conversion probable leads)
- Technically, we can generate this new set of leads by altering (moving down) the value of cut off so as to include more leads as the hot leads from our Logistic Regression Model
- Doing so, we will be better utilizing resources and improving chance of converting a lead whose lead conversion probability might be low as well.

THANK YOU