LEAD SCORING CASE STUDY

Kumar Mayank
Mayank Jain
Mangesh Deepak Bagul

AGENDA

Problem Statement

Methodology

Data Manipulation

Exploratory Data Analysis

Model Building

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.
- ❑ Building a suitable model which identifies the hot leads.

METHODOLOGY

❑ **Data cleaning and data manipulation.**

- Check and handle duplicate data.
- Check and handle NA values and missing values.
- Drop columns, if it contains large amount of missing values and not useful for the analysis.
- Imputation of the values, if necessary.
- Check and handle outliers in data.

❑ **Exploratory data Analysis**

- Univariate data analysis: value count, distribution of variable etc.
- Bivariate data analysis: correlation coefficients and pattern between the variables etc.

❑ **Feature Scaling & Dummy Variables and encoding of the data**.

❑ **Classification technique:** Logistic regression used for the model making and prediction.

❑ **Validation of the model.**

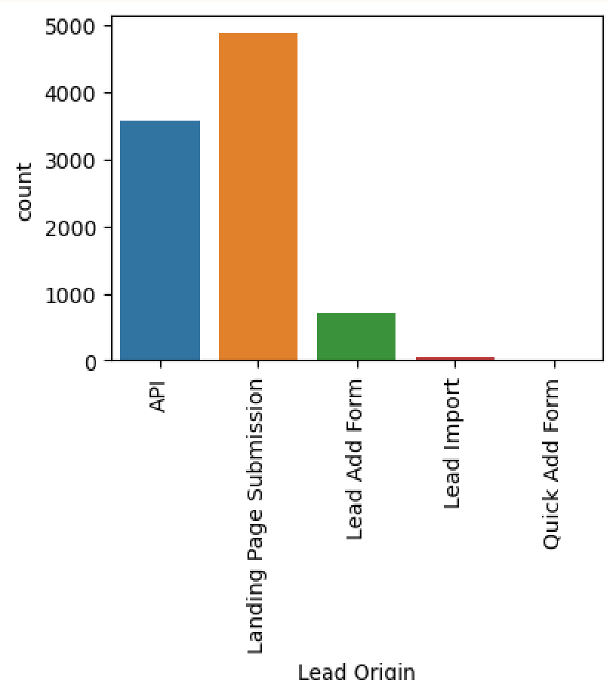
DATA CLEANING AND MANIPULATION

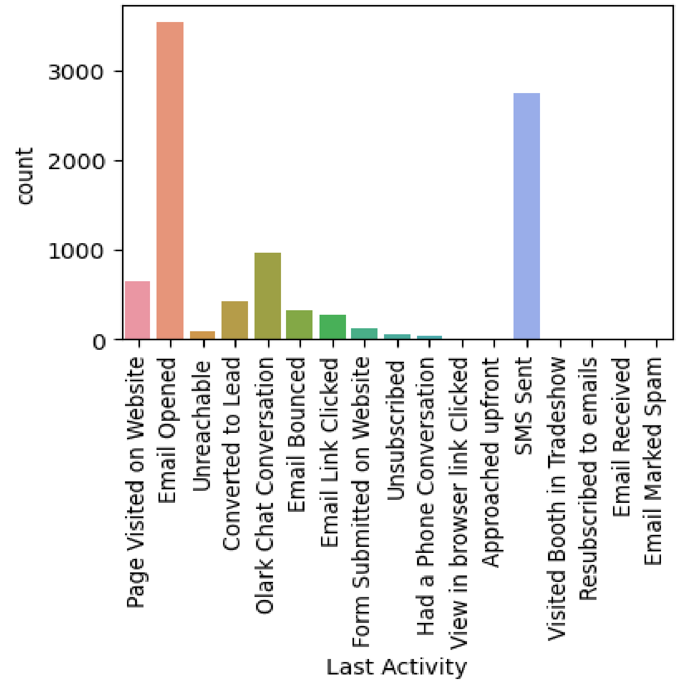
- ‘Select’ values were present which we have replaced with Null Values.
- Columns having minimum 40% Null values were dropped off
- Columns which had only one unique value were dropped off
- Imputing values with mean and mode in which less than 5% null values are present.
- Checking for Duplicate data points and rectifying the same.

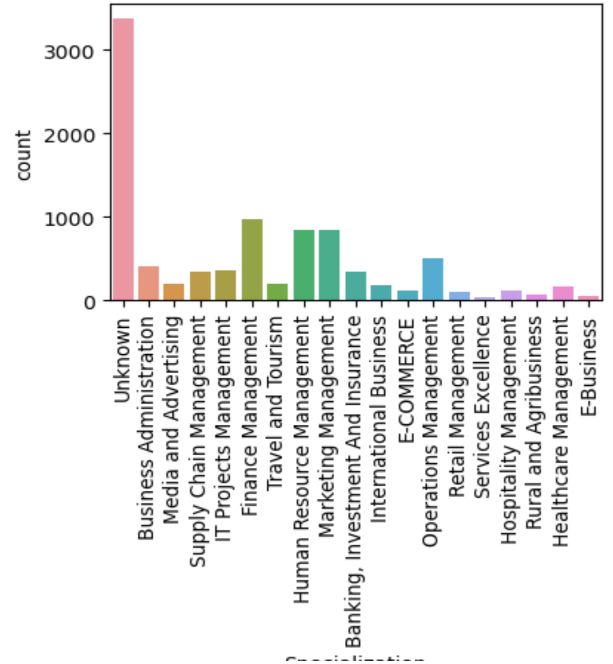
PERFORMING EDA

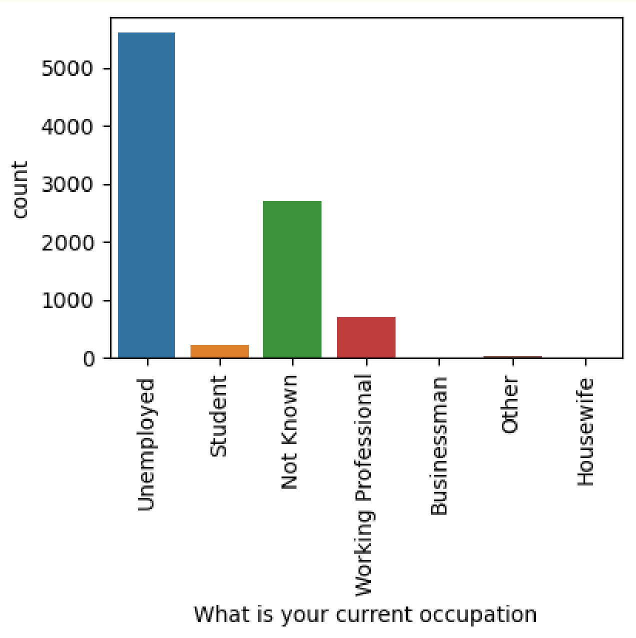
Univariate Analysis

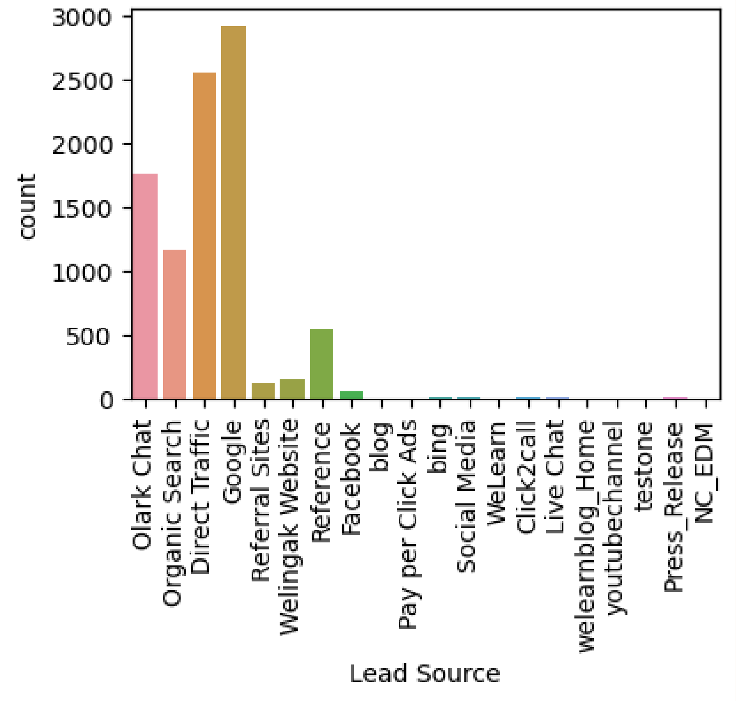
In Univariate analysis, mainly count plots and histograms were plotted to ge the understanding of the counts and frequencies of the different variables present in the data

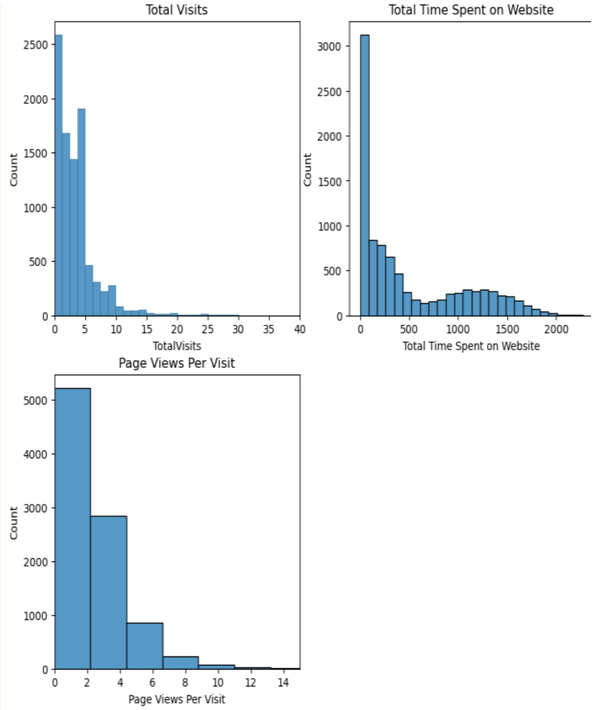

5000
4000
3000
2000
1000
0
count
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Origin

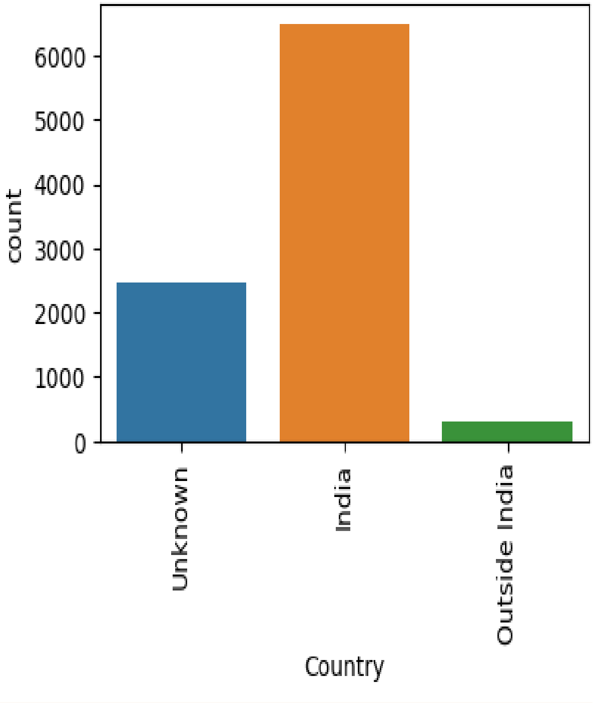

3000
2000
1000
0
count
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
Approached upfront
SMS Sent
Visited Booth in Tradeshow
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity

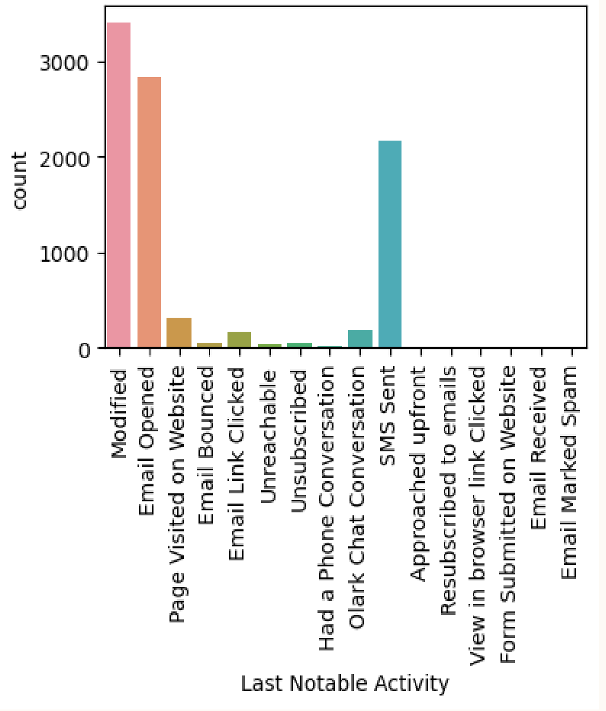

3000
2000
1000
0
count
Unknown
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization


5000
4000
3000
2000
1000
0
count
Unemployed
Student
Not Known
Working Professional
Businessman
Other
Housewife
What is your current occupation


3000
2500
2000
1500
1000
500
0
count
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source


Total Visits
2500
2000
1500
1000
500
0
Count
0
5
10
15
20
25
30
35
40
TotalVisits
Total Time Spent on Website
3000
2500
2000
1500
1000
500
0
Count
0
500
1000
1500
2000
Total Time Spent on Website
Page Views Per Visit
5000
4000
3000
2000
1000
0
Count
0
2
4
6
8
10
12
14
Page Views Per Visit

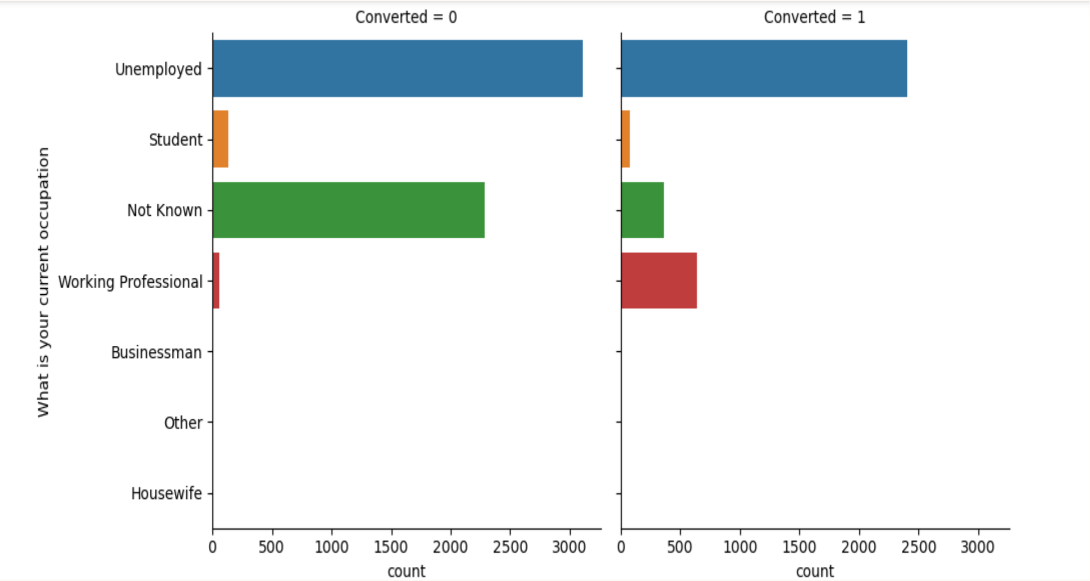

6000
5000
4000
3000
2000
1000
0
count
Unknown
India
Outside India
Country

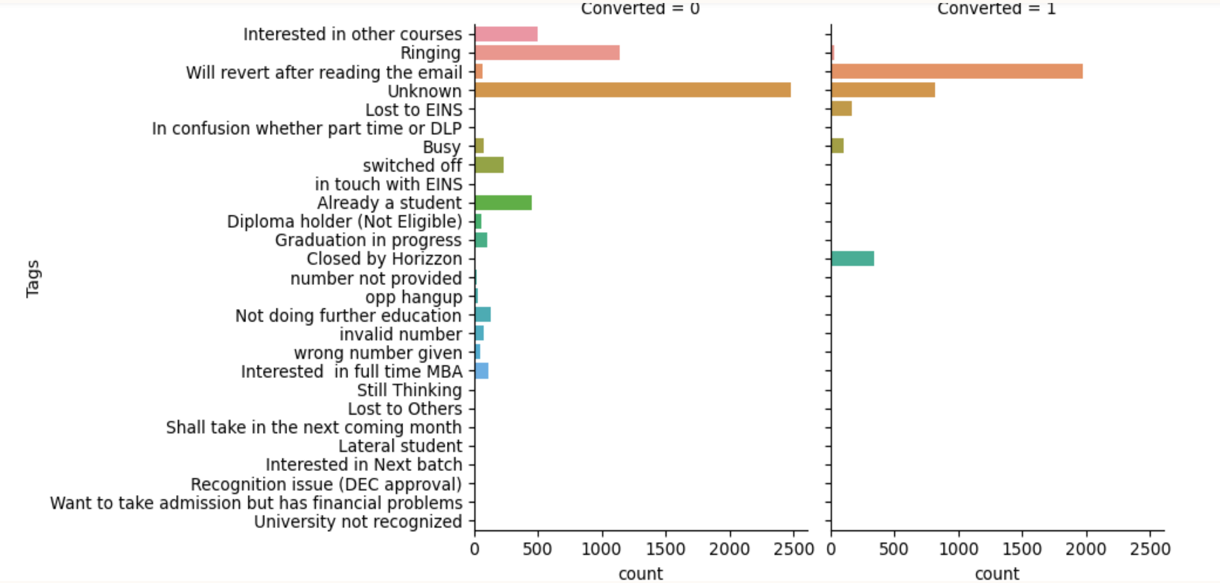

3000
2000
1000
0
count
Modified
Email Opened
Page Visited on Website
Email Bounced
Email Link Clicked
Unreachable
Unsubscribed
Had a Phone Conversation
Olark Chat Conversation
SMS Sent
Approached upfront
Resubscribed to emails
View in browser link Clicked
Form Submitted on Website
Email Received
Email Marked Spam
Last Notable Activity

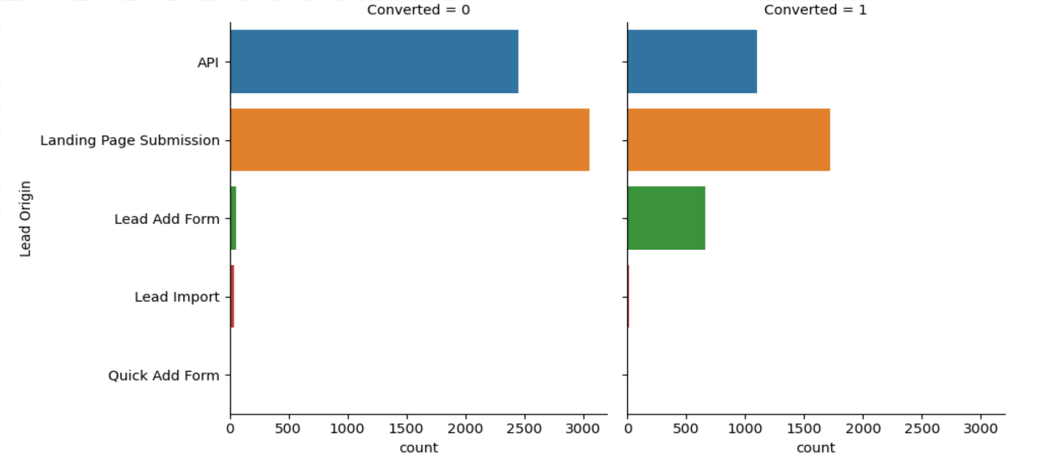
PERFORMING EDA

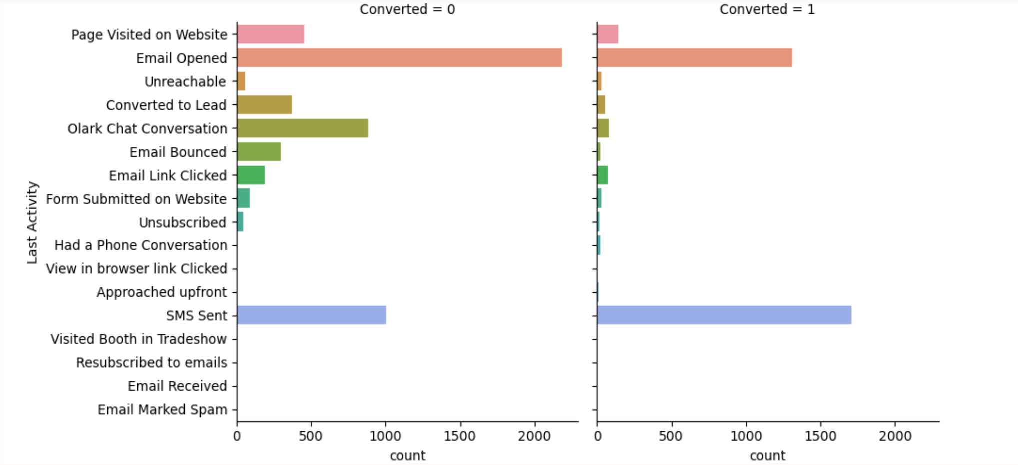
Bivariate Analysis

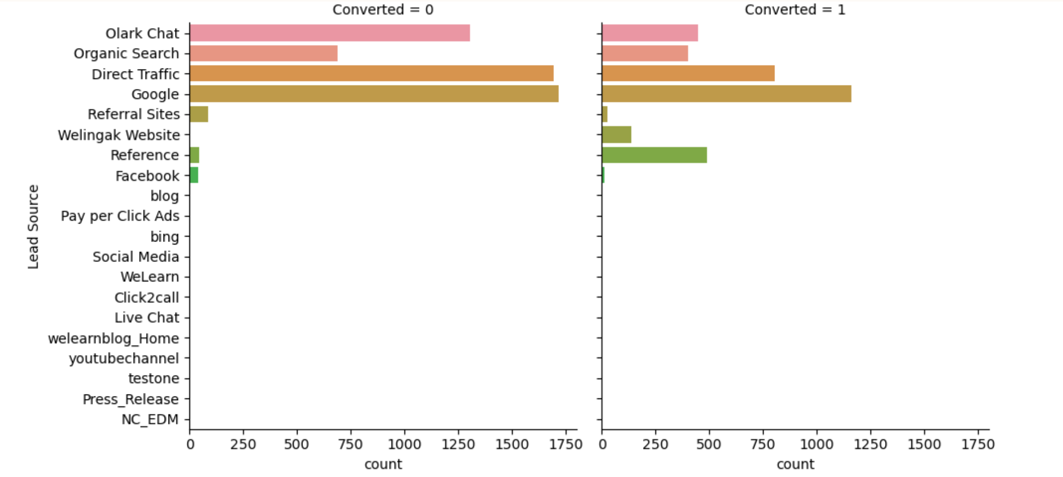
In Bivariate analysis, count plots were plotted keeping ‘Converted’ as the target variable and other independent variables

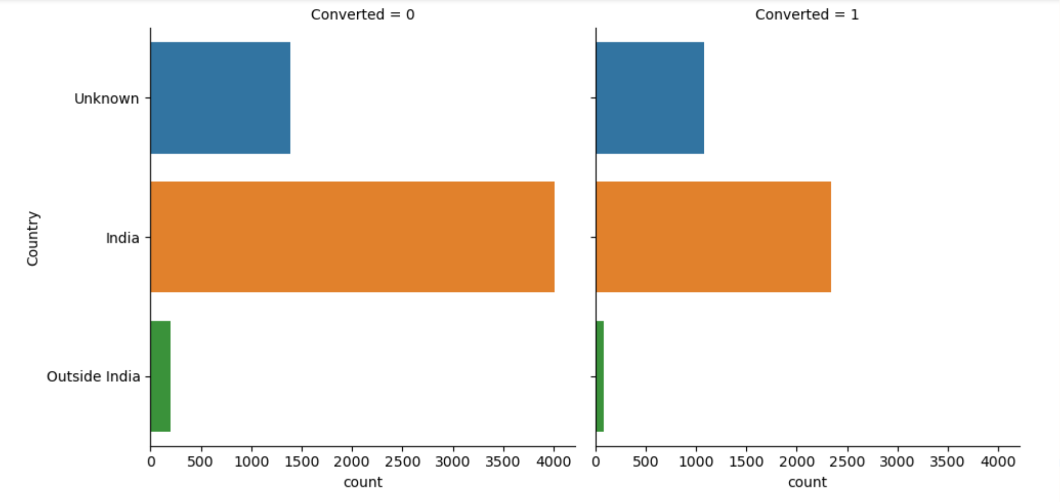





Converted = 0
Converted = 1
Lead Origin
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
0
500
1000
1500
2000
2500
3000
count


Converted = 0
Converted = 1
Last Activity
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
Approached upfront
SMS Sent
Visited Booth in Tradeshow
Resubscribed to emails
Email Received
Email Marked Spam
0
500
1000
1500
2000
count


Converted = 0
Converted = 1
Lead Source
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
0
250
500
750
1000
1250
1500
1750
count


Converted = 0
Converted = 1
Country
Unknown
India
Outside India
0
500
1000
1500
2000
2500
3000
3500
4000
count

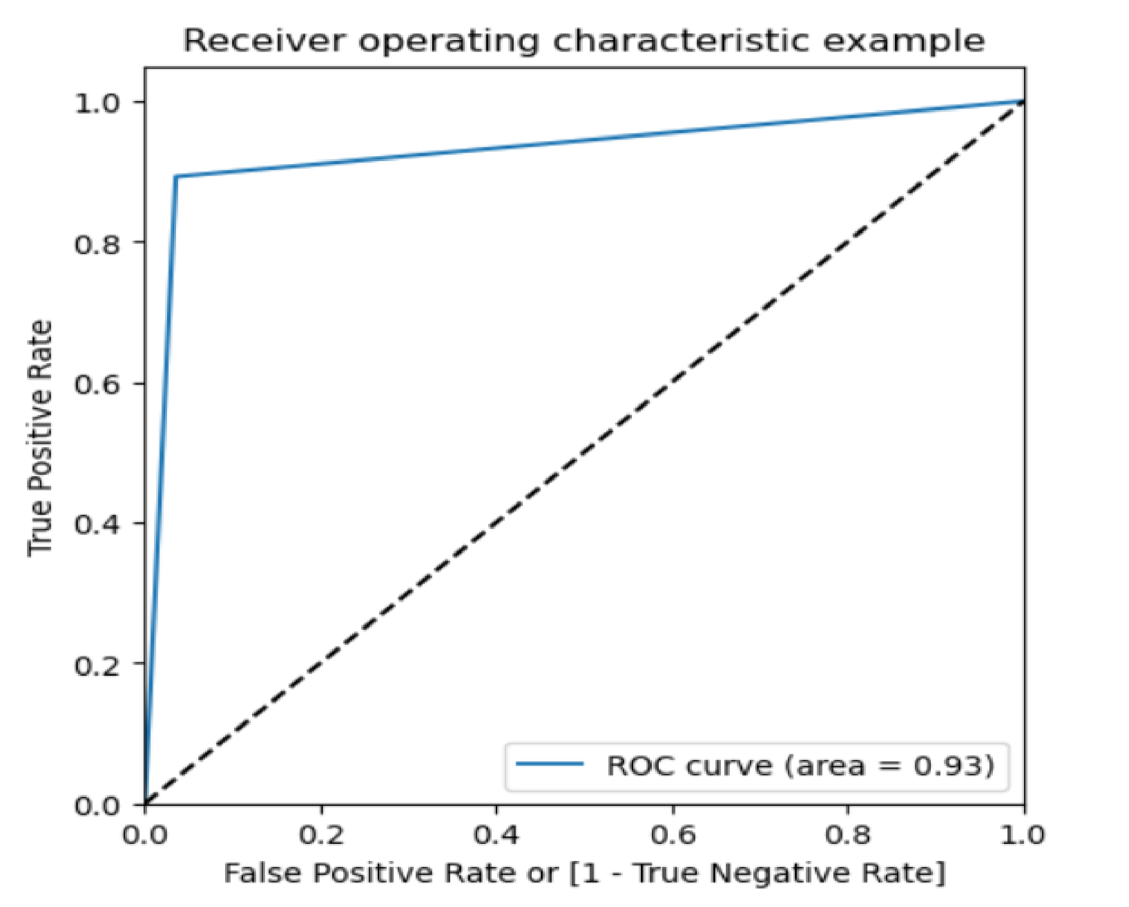
DATA CONVERSION

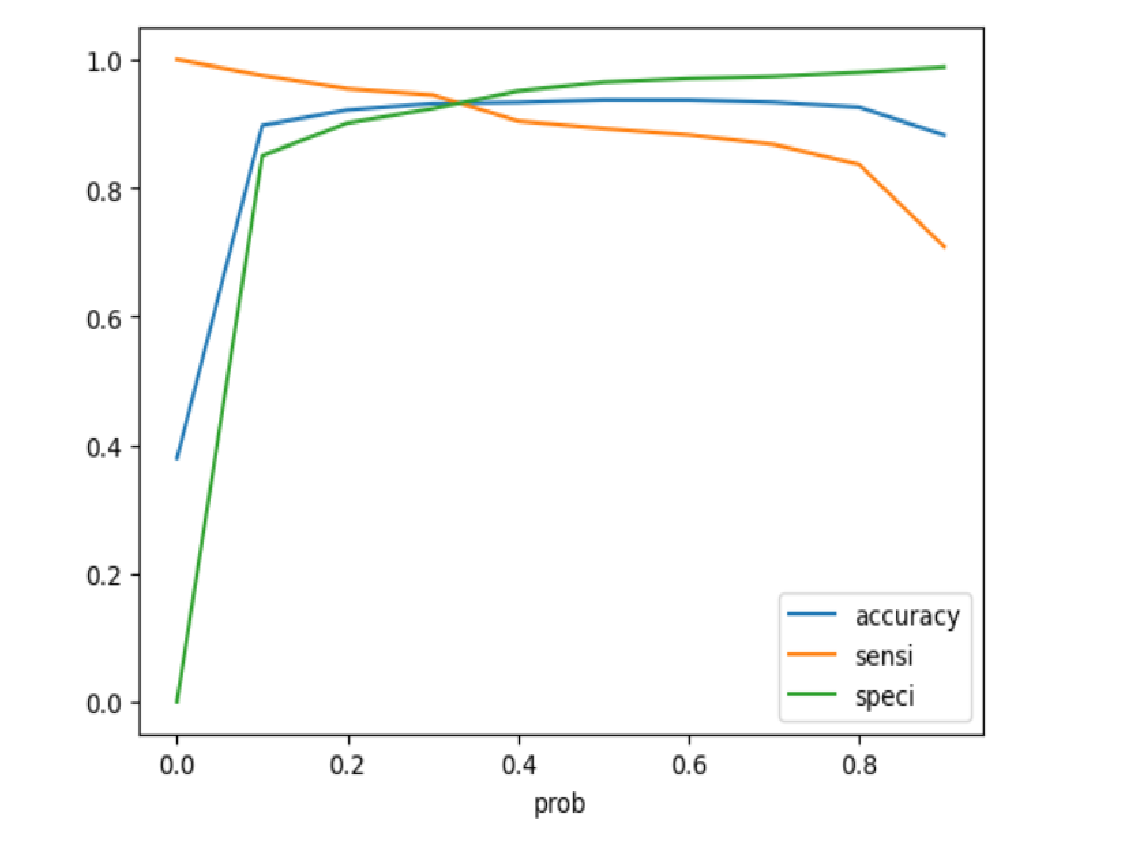
- Before initiating the model building, the dummy variables were created using ‘get dummies’
- The data type of Boolean columns was also converted into integer type

MODEL BUILDING (1/3)

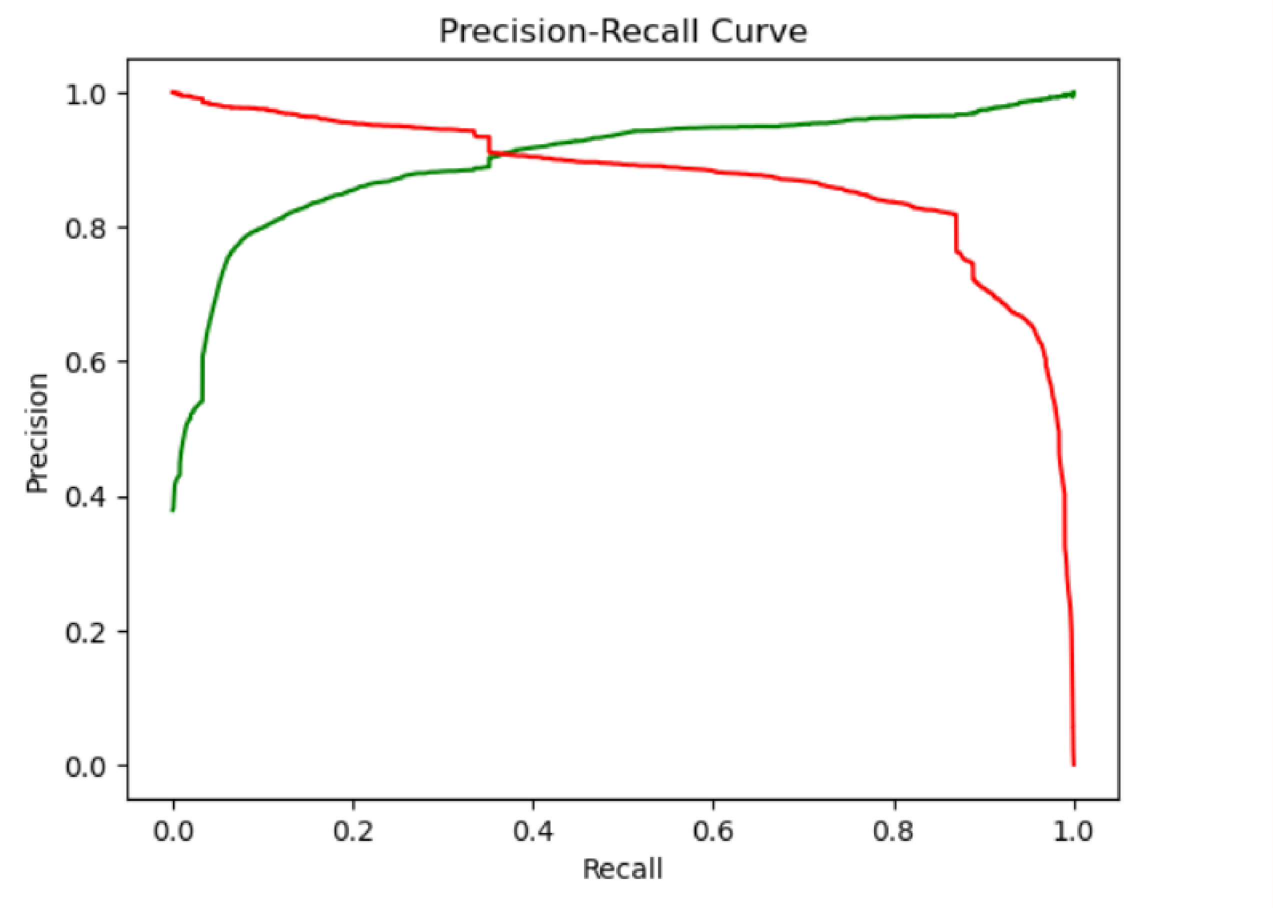
- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Feature scaling
- Checking correlation of variables using heat maps
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and VIF value is greater than 5
- Predicting the probabilities on train set and evaluating the model using accuracy, sensitivity specificity.
- Predictions on test data

MODEL BUILDING (2/3)

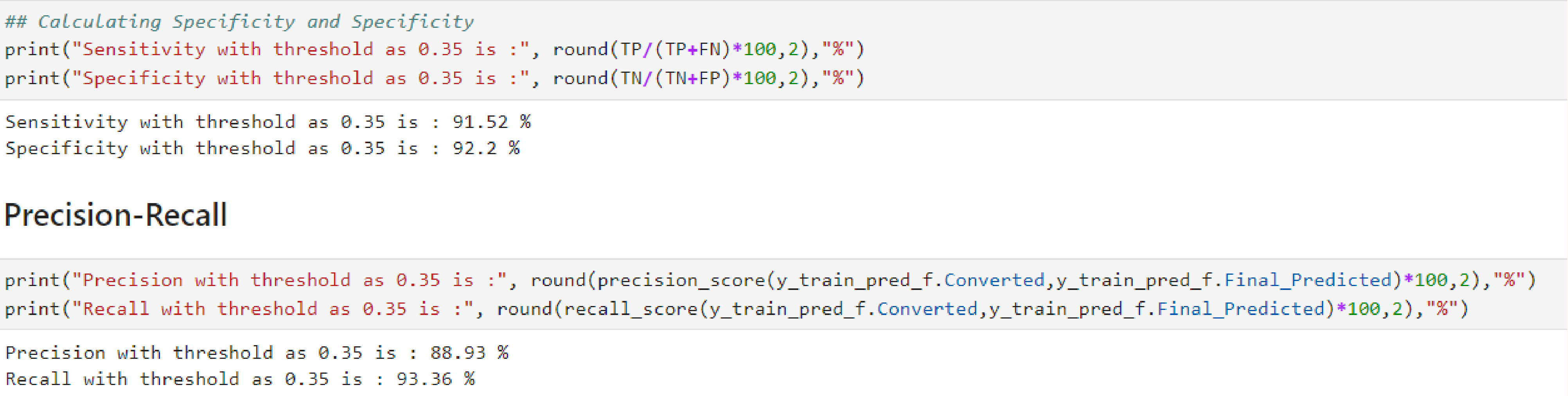




MODEL BUILDING (3/3)



MODEL BUILDING (4/4)



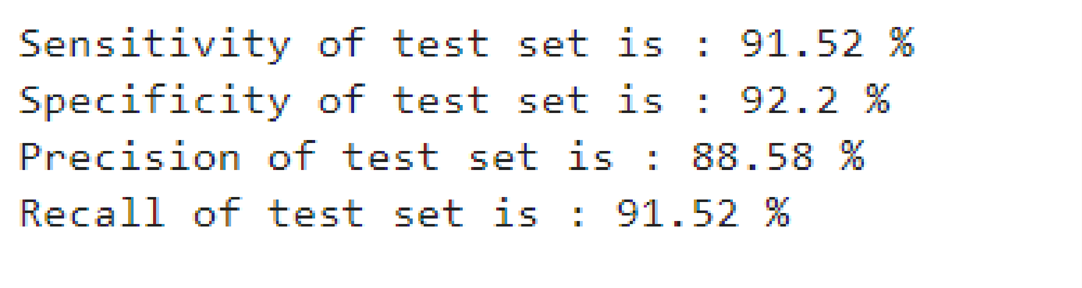
```
## Calculating Specificity and Specificity
print("Sensitivity with threshold as 0.35 is :", round(TP/(TP+FN)*100,2),"%")
print("Specificity with threshold as 0.35 is :", round(TN/(TN+FP)*100,2),"%")
```

```
Sensitivity with threshold as 0.35 is : 91.52 %
Specificity with threshold as 0.35 is : 92.2 %
```

Precision-Recall

```
print("Precision with threshold as 0.35 is :", round(precision_score(y_train_pred_f.Converted,y_train_pred_f.Final_Predicted)*100,2),"%")
print("Recall with threshold as 0.35 is :", round(recall_score(y_train_pred_f.Converted,y_train_pred_f.Final_Predicted)*100,2),"%")
```

```
Precision with threshold as 0.35 is : 88.93 %
Recall with threshold as 0.35 is : 93.36 %
```



```
Sensitivity of test set is : 91.52 %
Specificity of test set is : 92.2 %
Precision of test set is : 88.58 %
Recall of test set is : 91.52 %
```

KEY INSIGHTS

The variables identified as most influential in predicting potential buyers, ranked in descending order of importance, are:

- Tags
- Lead Source
- Total Time Spent on Website
- Last Notable Activity
- Last Activity.
- Current occupation as a working professional.

In conclusion, focusing efforts on these key variables presents an opportunity for X Education to significantly enhance its conversion rates by effectively engaging and persuading potential buyers.

THANK YOU