Lead Scoring Case study

Submitted By:

Venkatesh R

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.

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. Moreover, the company also gets leads through past referrals. 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, although 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.

Goals of the Case Study

1. 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. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

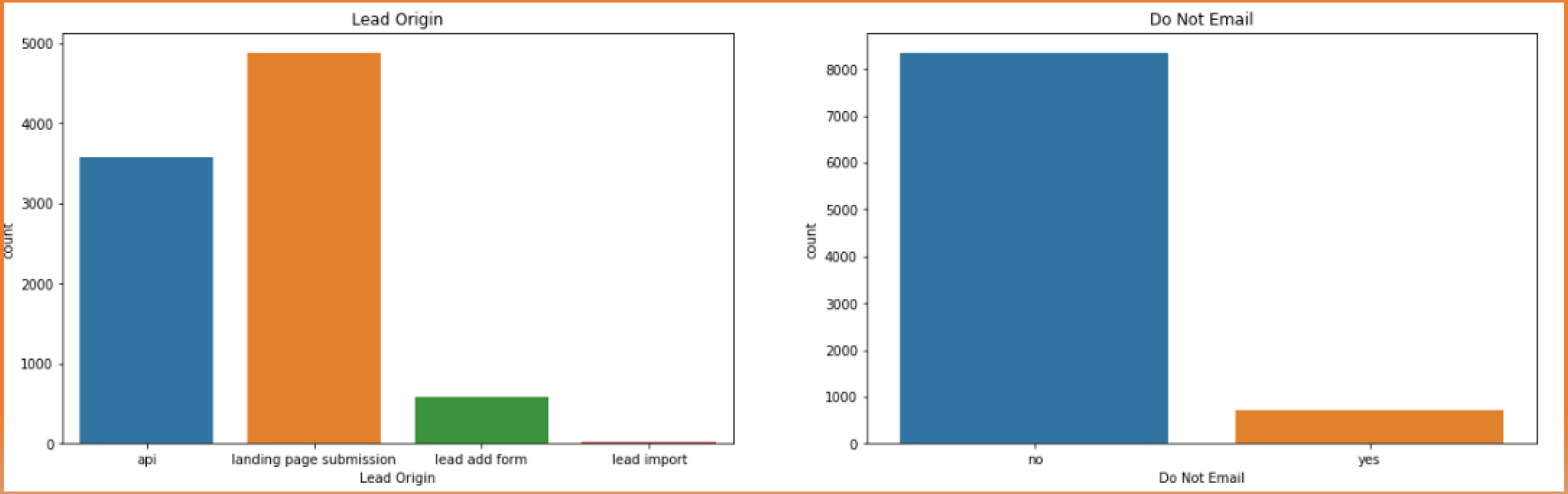
2.There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

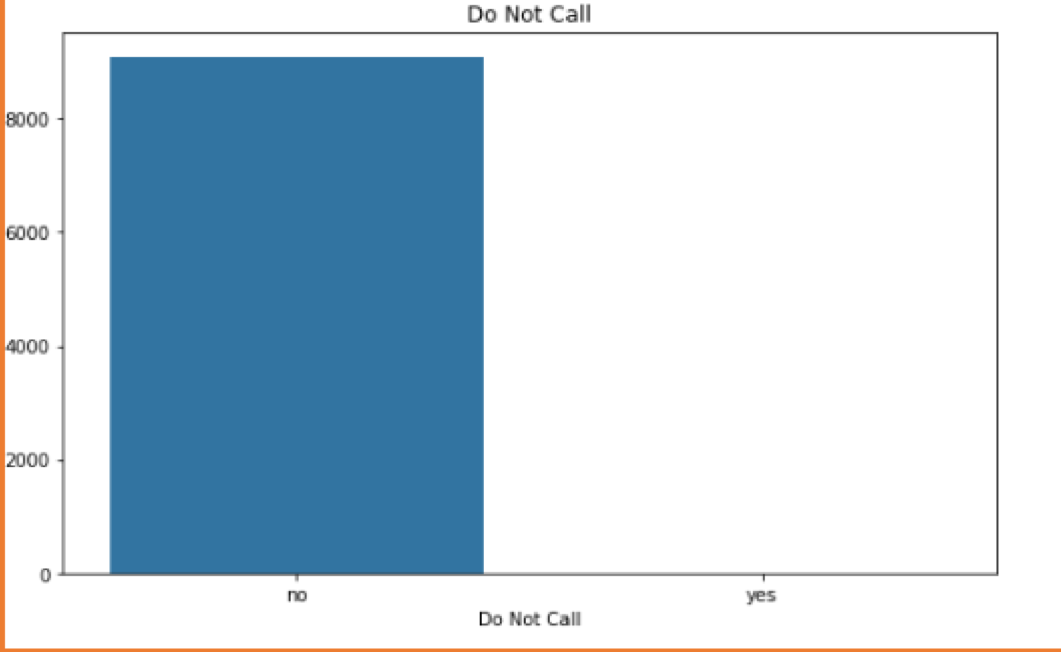
Approach

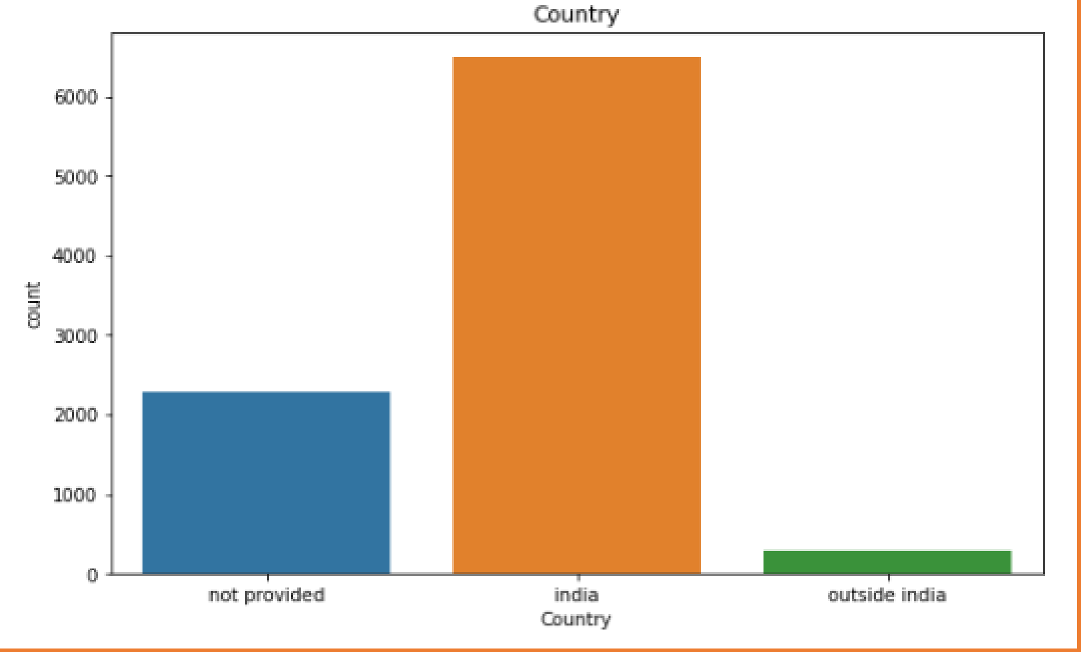
- Source the data for analysis
- Reading and understanding the data
- Data cleaning
- EDA
- Feature scaling
- Splitting the data into train and test dataset
- Prepare the Data for modelling
- Module building
- Module evaluation specificity and sensitivity or precision recall
- Making prediction on the test set

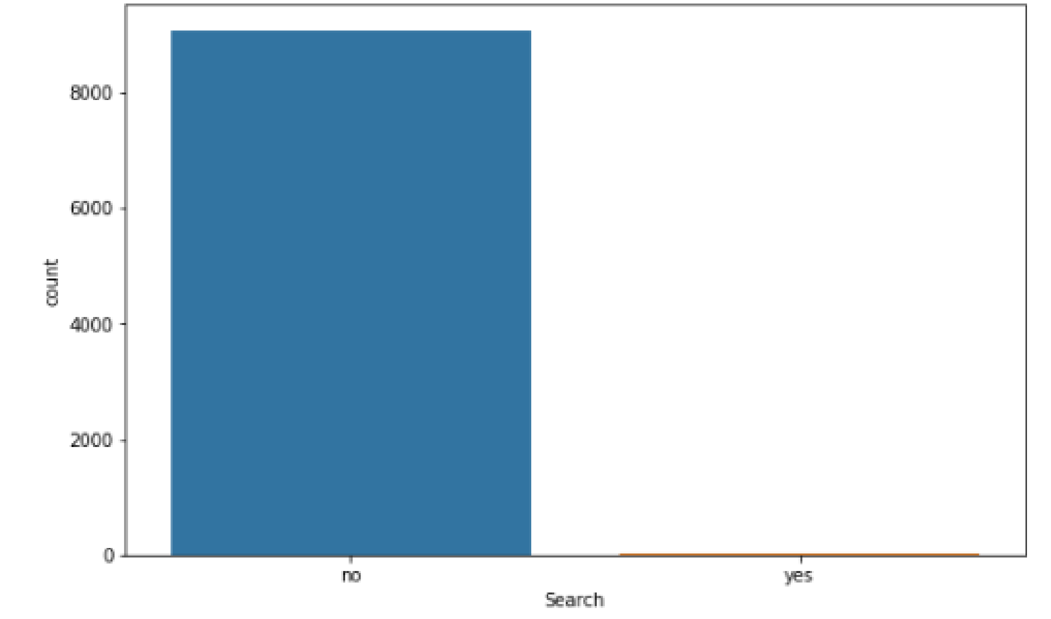
EDA

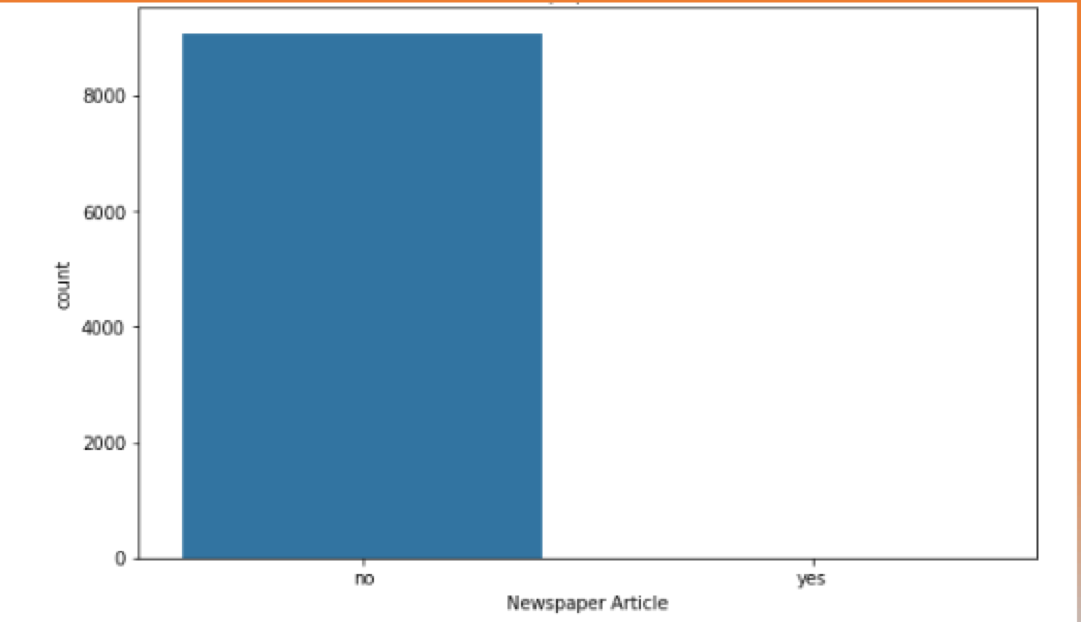
Univariate Analysis -Categorical Variables

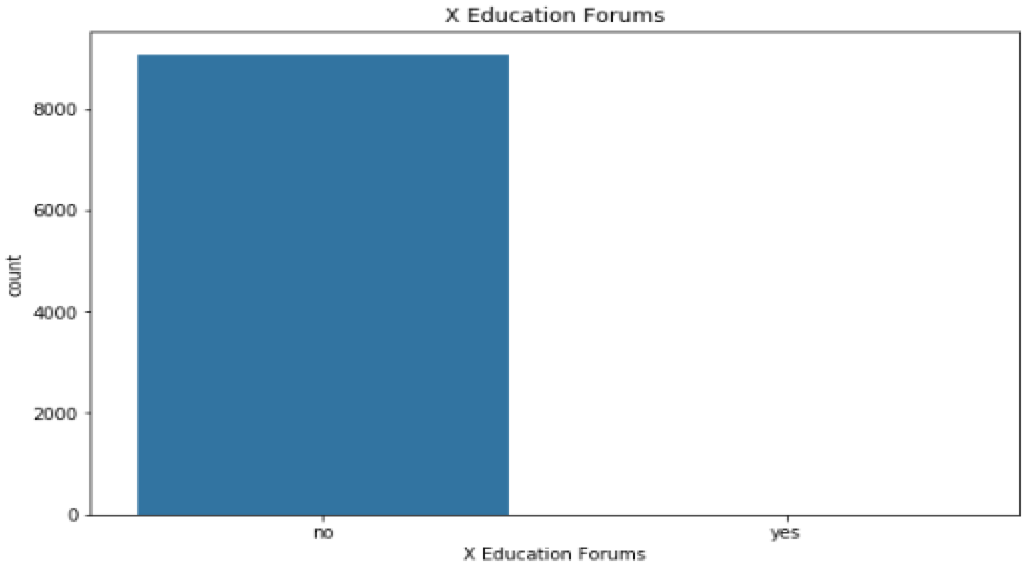



Do Not Call
8000
6000
4000
2000
0
no
yes
Do Not Call

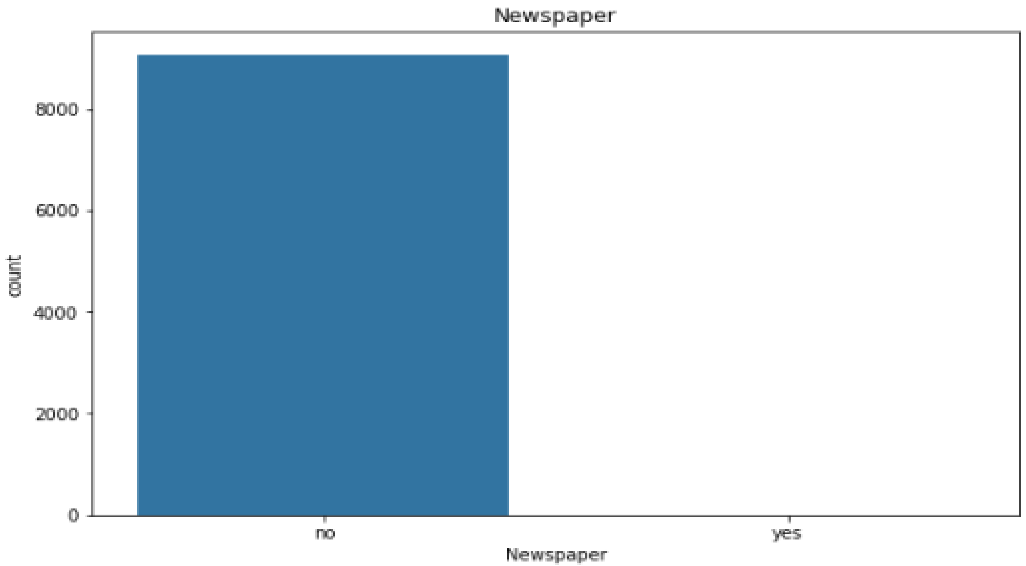

Country
count
6000
5000
4000
3000
2000
1000
0
not provided
india
outside india
Country


count
8000
6000
4000
2000
0
no
yes
Search

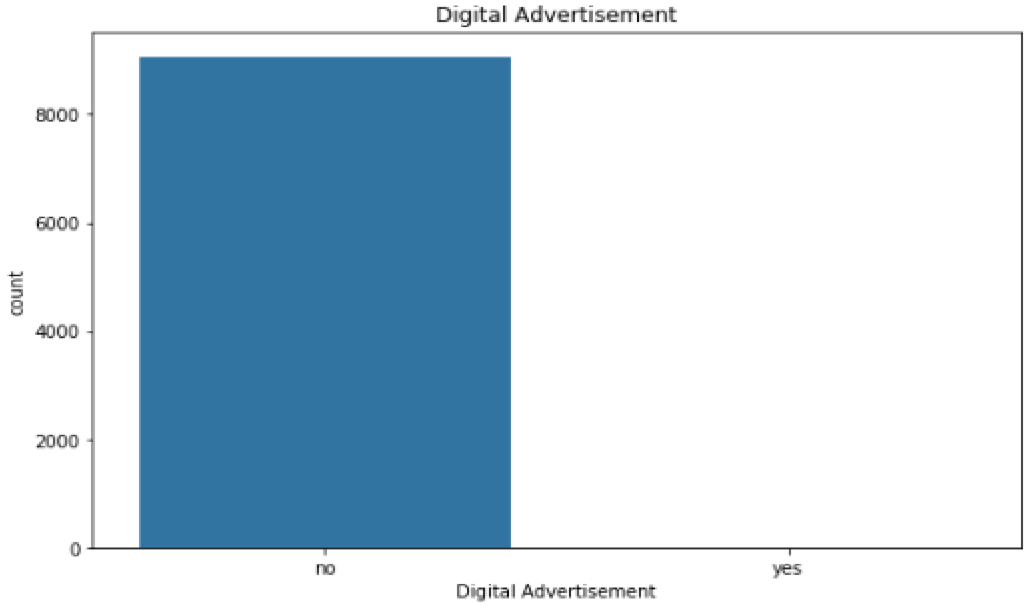

count
8000
6000
4000
2000
0
no
yes
Newspaper Article



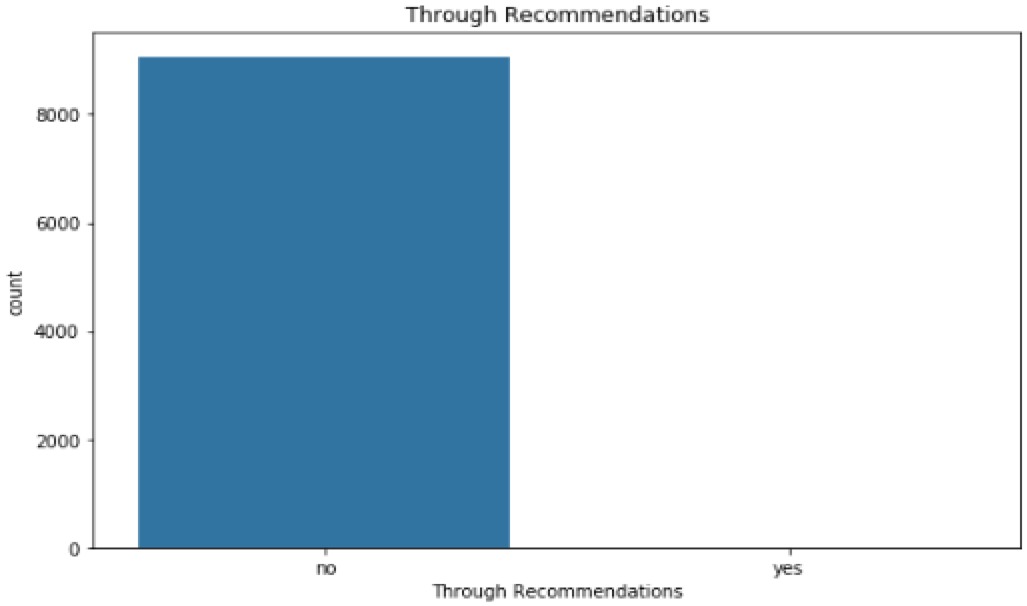
X Education Forums



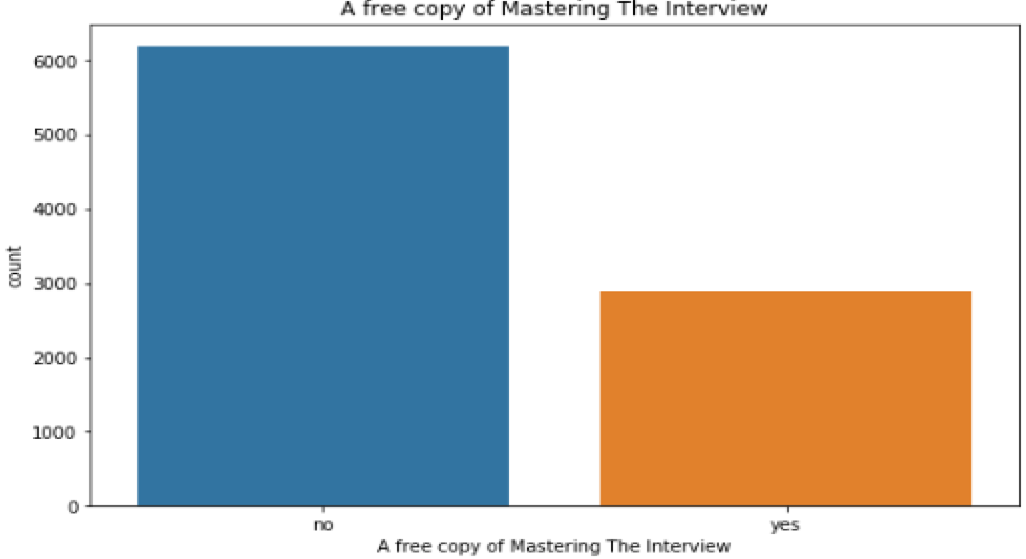
Newspaper



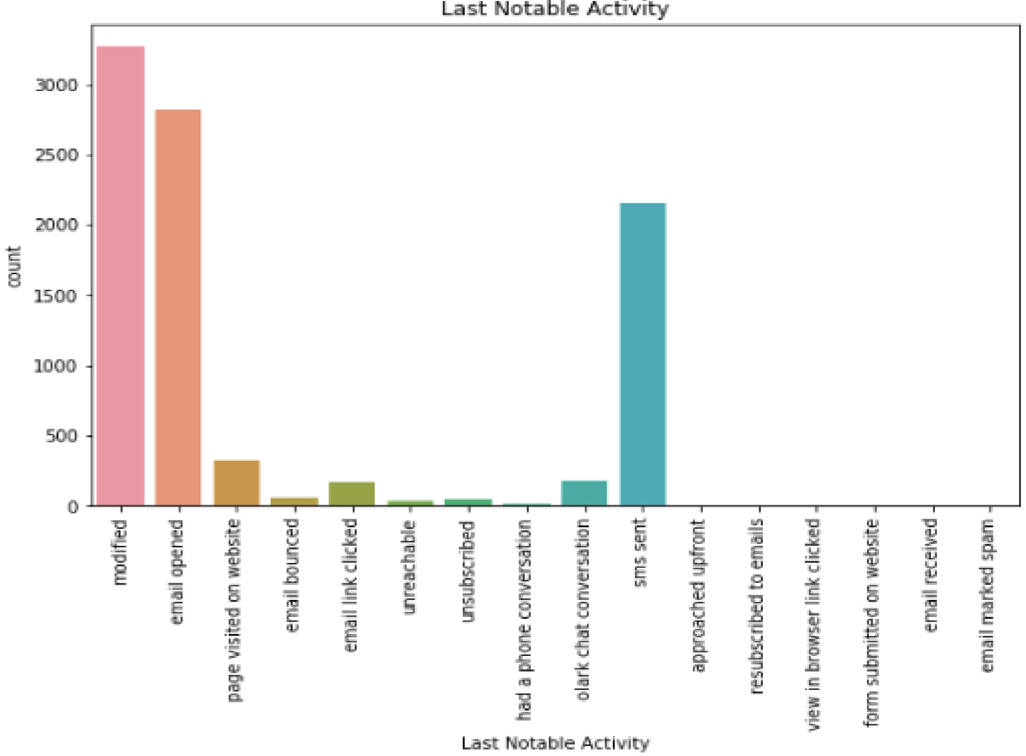
Digital Advertisement



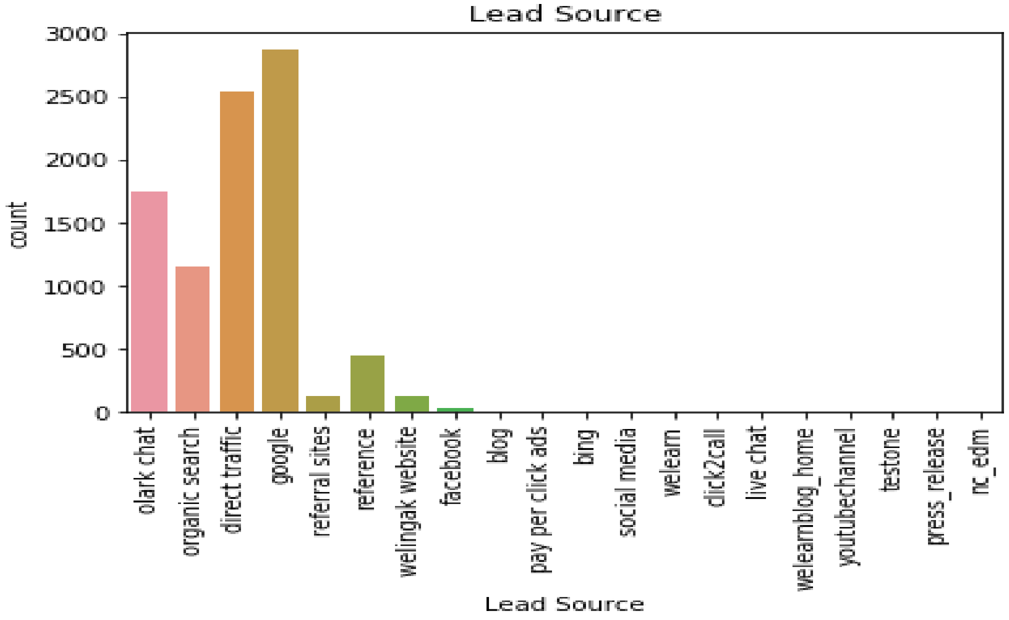
Through Recommendations



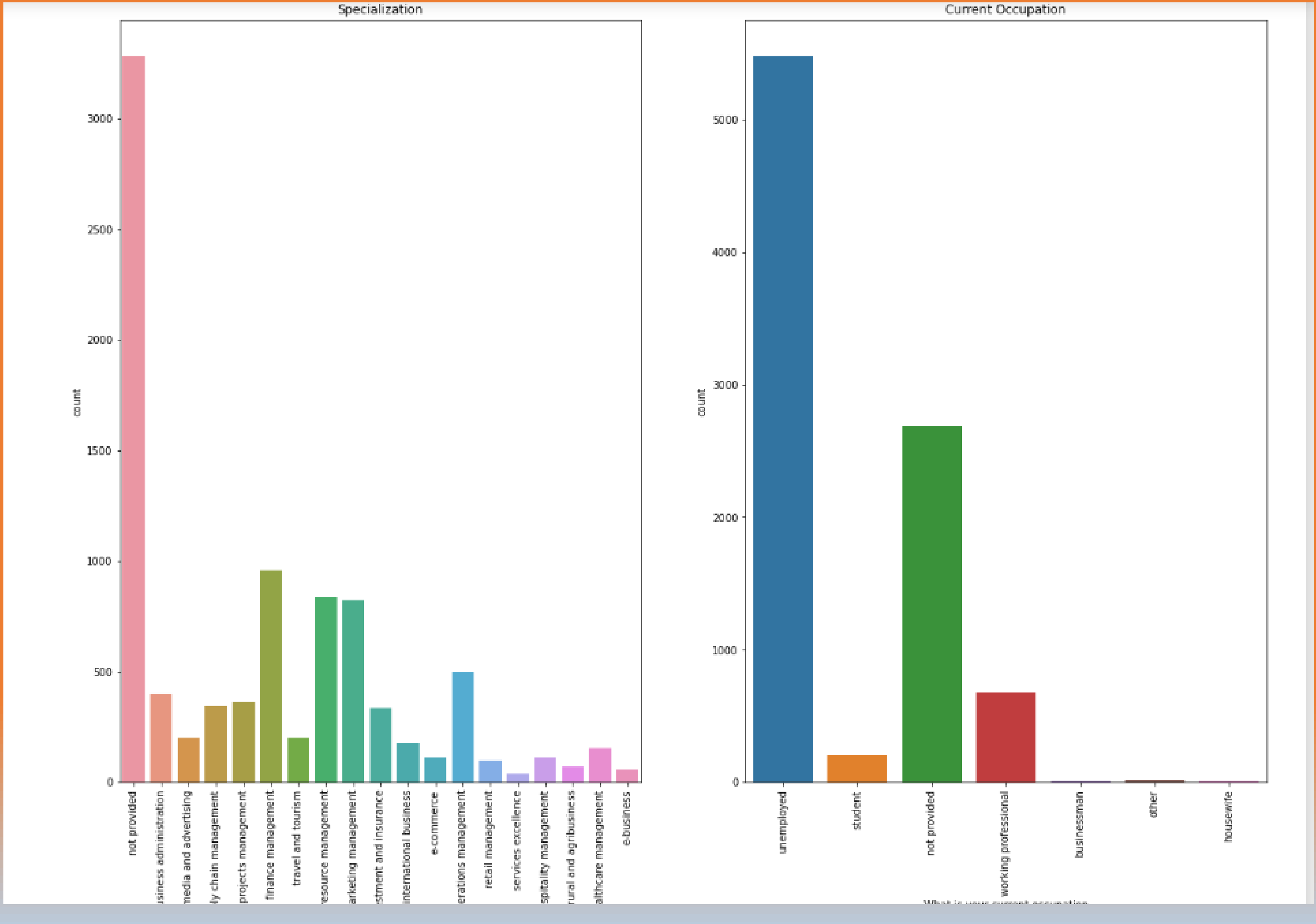
A free copy of Mastering The Interview

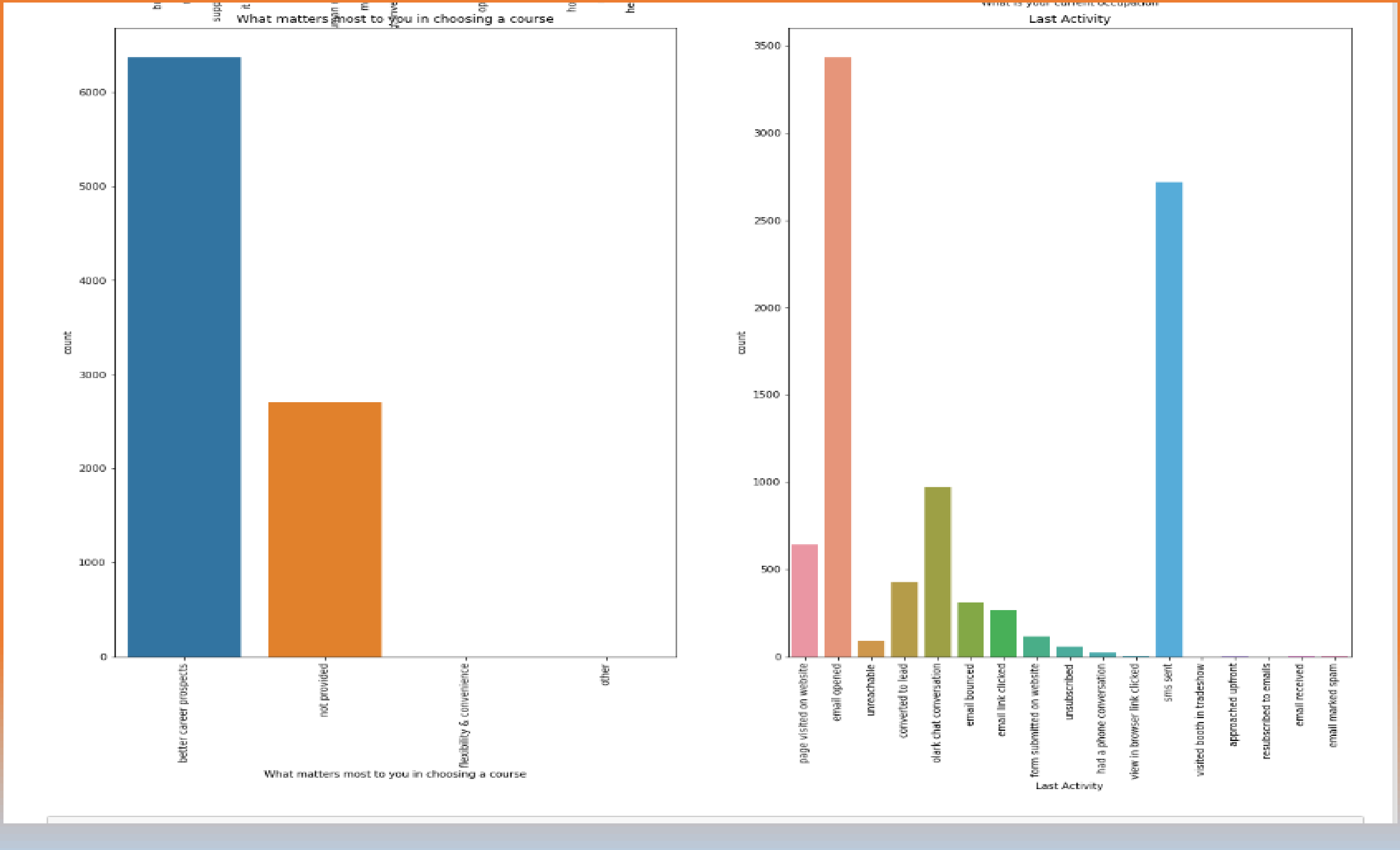


Last Notable Activity

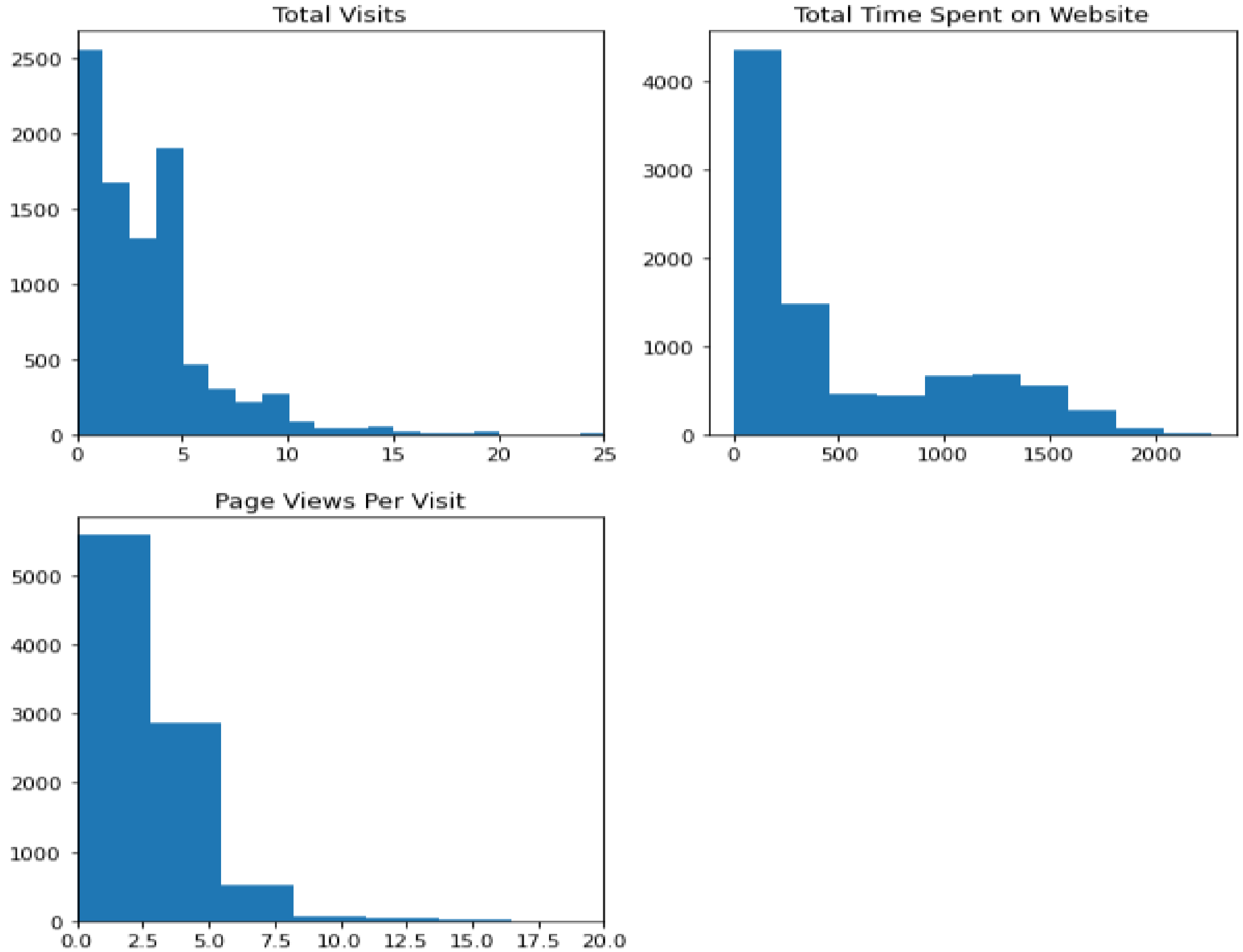


Lead Source

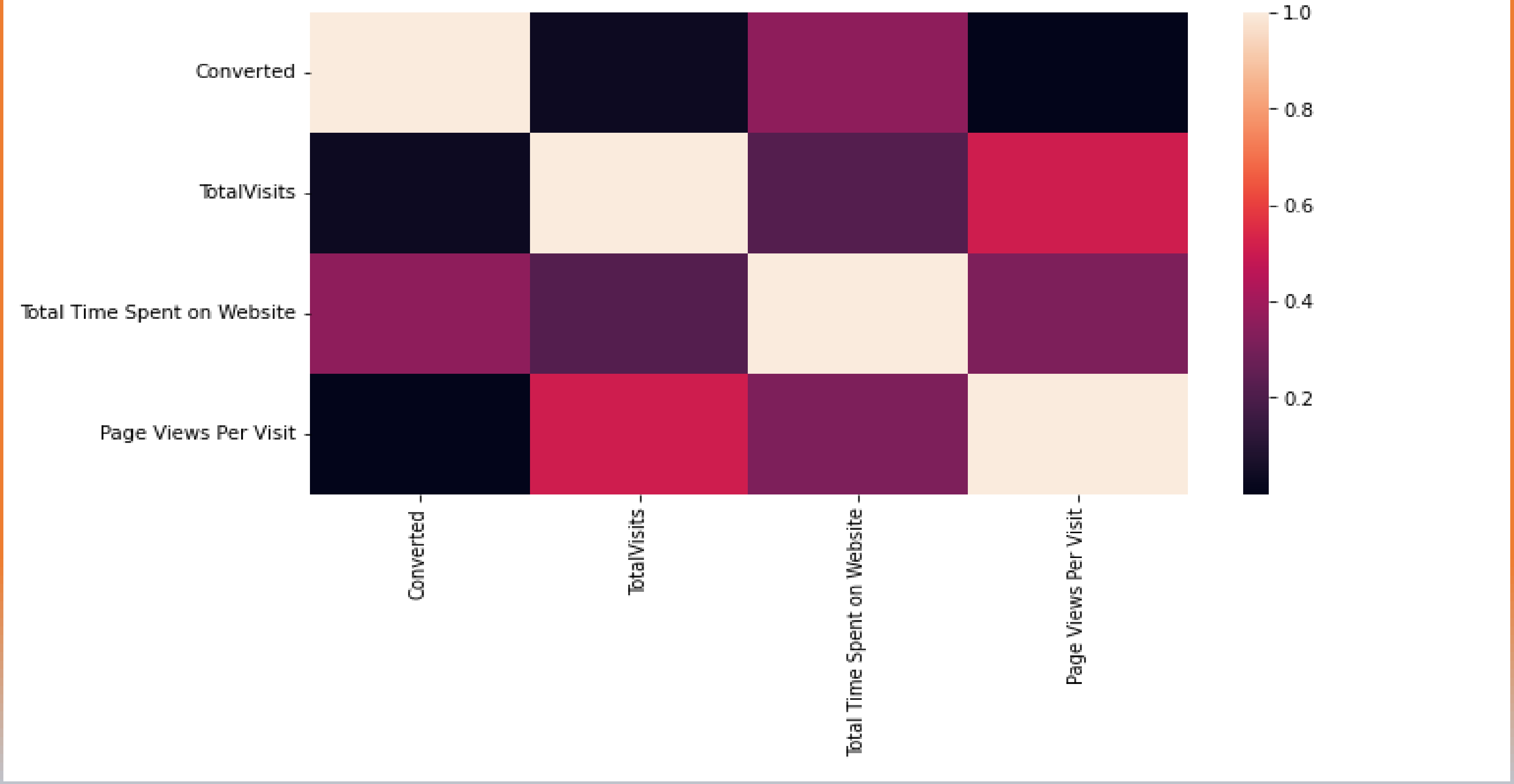

Specialization
count
0
500
1000
1500
2000
2500
3000
not provided
usiness administration
media and advertising
ly chain management
projects management
finance management
travel and tourism
resource management
arketing management
stment and insurance
international business
e-commerce
erations management
retail management
services excellence
spitality management
rural and agribusiness
althcare management
e-business
Current Occupation
count
0
1000
2000
3000
4000
5000
unemployed
student
not provided
working professional
businessman
other
housewife
What is your current occupation


What is your current occupation
What matters most to you in choosing a course
count
0
1000
2000
3000
4000
5000
6000
better career prospects
not provided
flexibility & convenience
other
What matters most to you in choosing a course
Last Activity
count
0
500
1000
1500
2000
2500
3000
3500
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
sms sent
visited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam
Last Activity

Numerical Variables



Correlation among variables

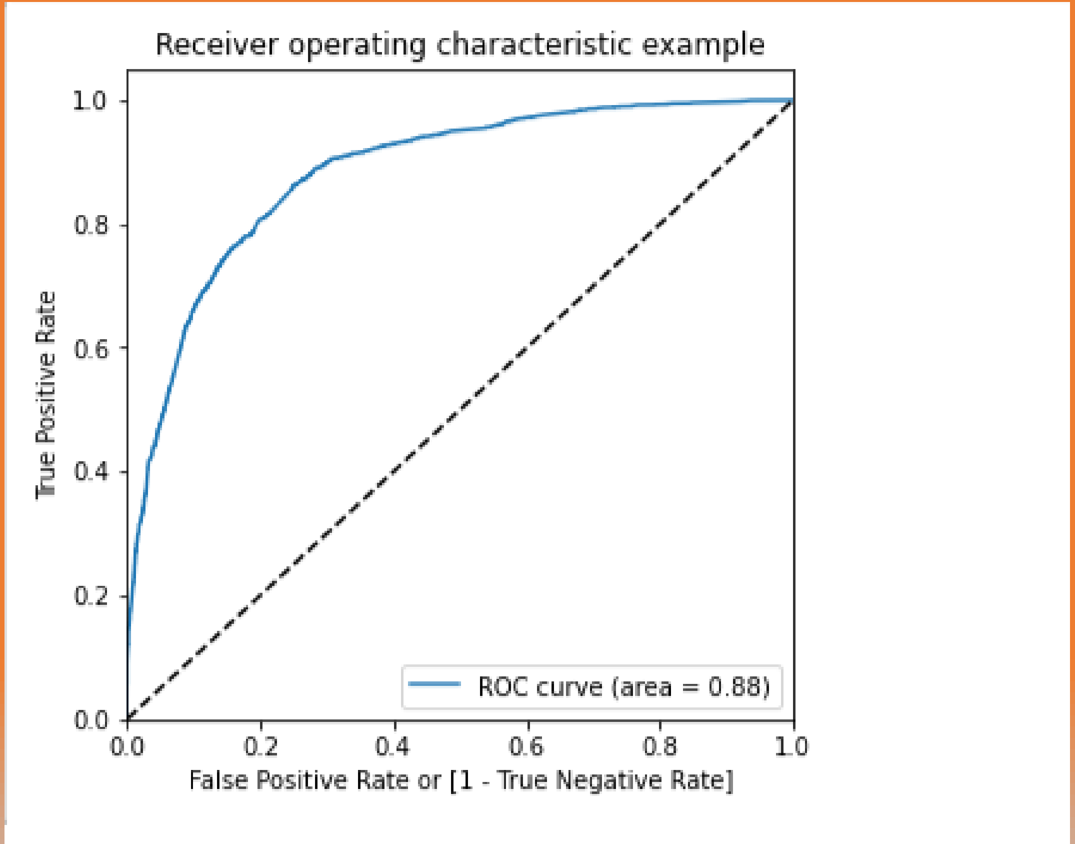


Optimise Cut off (ROC Curve)

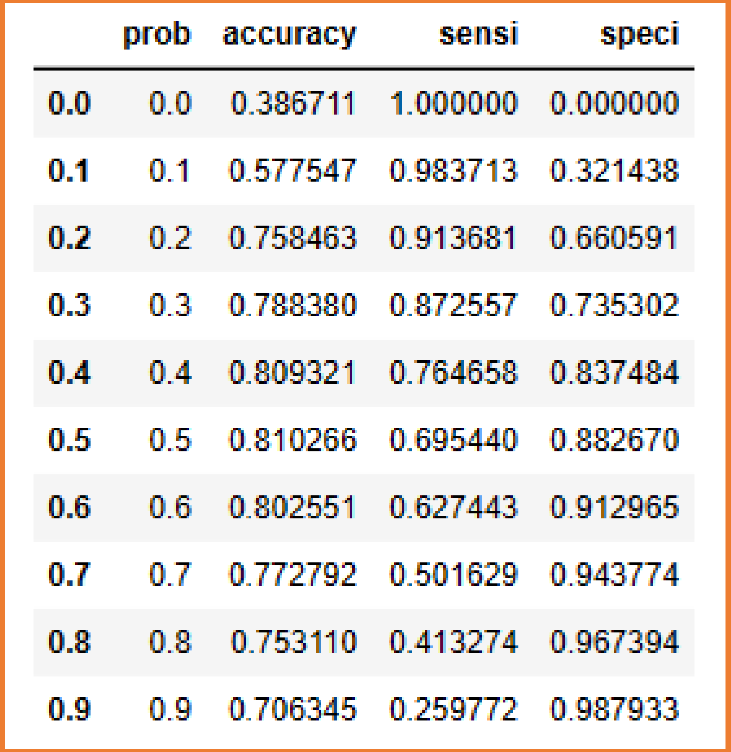
The term "Optimize Cut-off (ROC Curve)" refers to finding the optimal threshold or cut-off value for a binary classification model using the Receiver Operating Characteristic (ROC) curve.

In a binary classification problem, the model's predictions are typically based on a continuous probability score. The ROC curve is a graphical representation of the model's performance at various classification thresholds. It plots the true positive rate (TPR) against the false positive rate (FPR) for different threshold values.

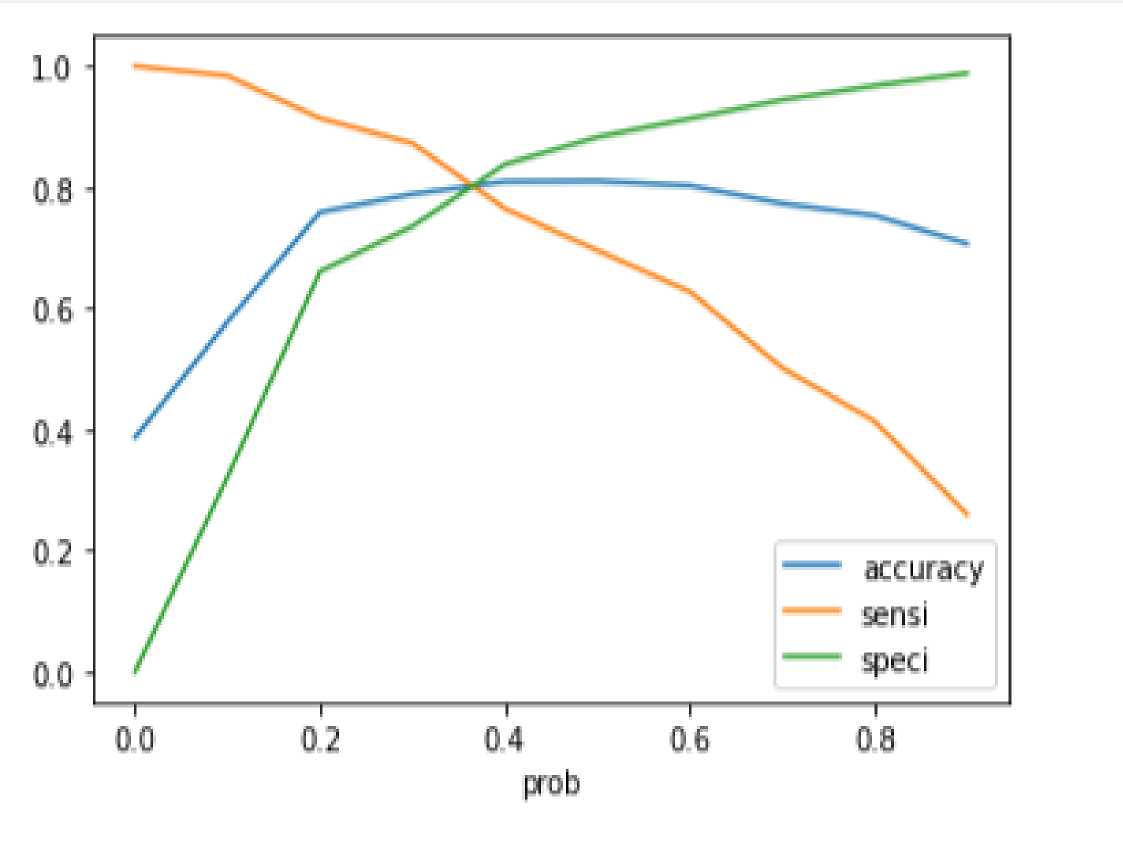
The area under ROC curve is 0.88 which is a very good value

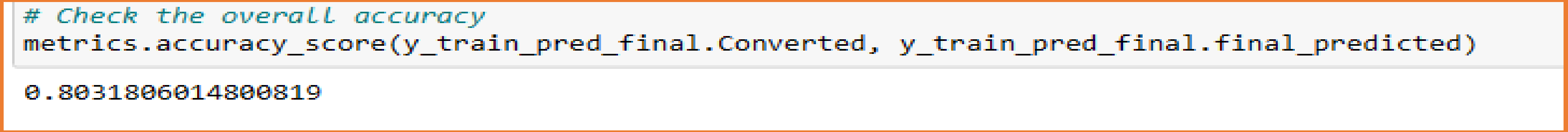


Confusing matrix to find values of sensitivity, accuracy and specificity



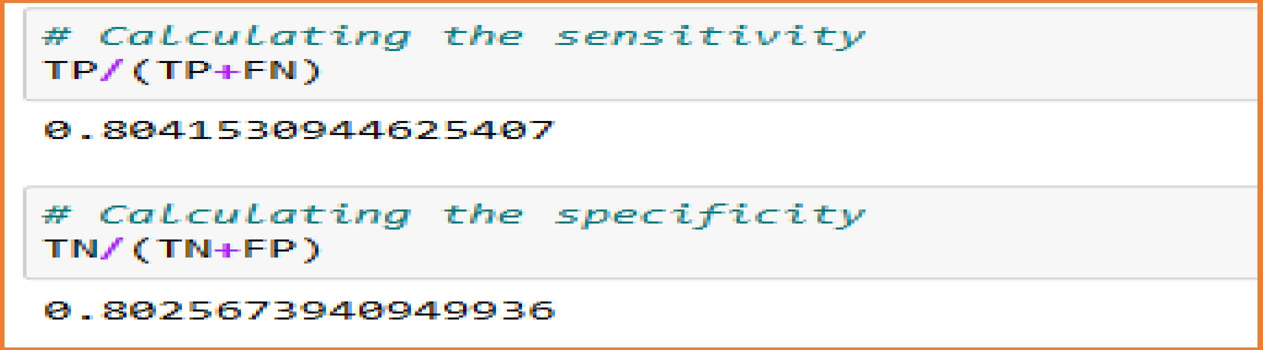
	prob	accuracy	sensi	speci
0.0	0.0	0.386711	1.000000	0.000000
0.1	0.1	0.577547	0.983713	0.321438
0.2	0.2	0.758463	0.913681	0.660591
0.3	0.3	0.788380	0.872557	0.735302
0.4	0.4	0.809321	0.764658	0.837484
0.5	0.5	0.810266	0.695440	0.882670
0.6	0.6	0.802551	0.627443	0.912965
0.7	0.7	0.772792	0.501629	0.943774
0.8	0.8	0.753110	0.413274	0.967394
0.9	0.9	0.706345	0.259772	0.987933





```
# Check the overall accuracy
metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.final_predicted)
```

```
0.8031806014800819
```



```
# Calculating the sensitivity
TP/(TP+FN)
```

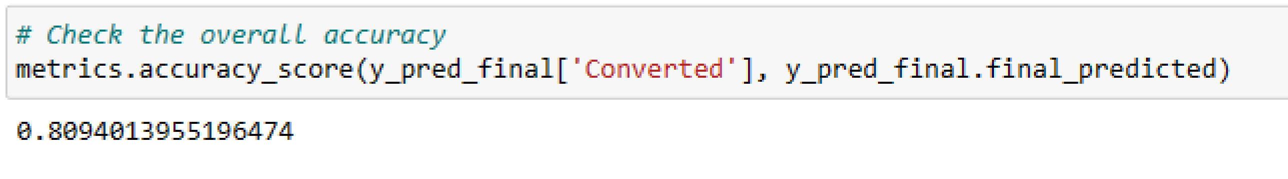
```
0.8041530944625407
```

```
# Calculating the specificity
TN/(TN+FP)
```

```
0.8025673940949936
```

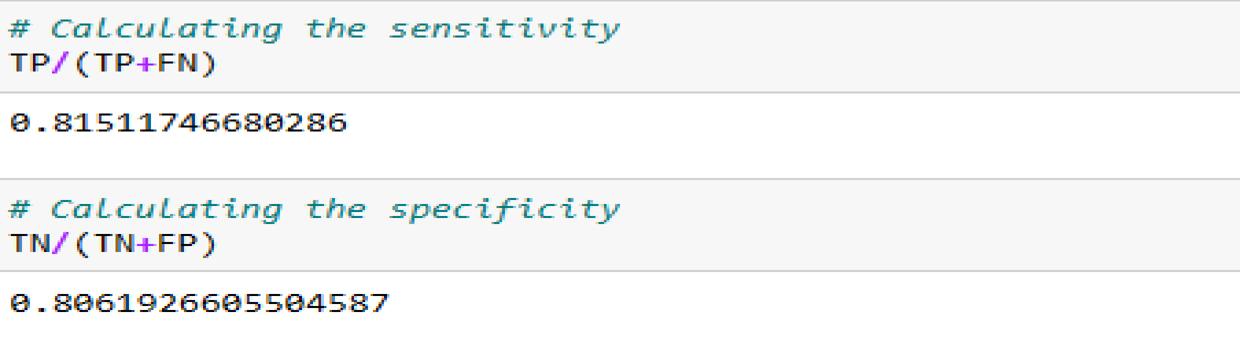
With the current cut off as 0.35 we have accuracy, sensitivity and specificity of around 80%

Prediction on Test set



```
# Check the overall accuracy
metrics.accuracy_score(y_pred_final['Converted'], y_pred_final.final_predicted)
```

```
0.8094013955196474
```



```
# Calculating the sensitivity
TP/(TP+FN)
```

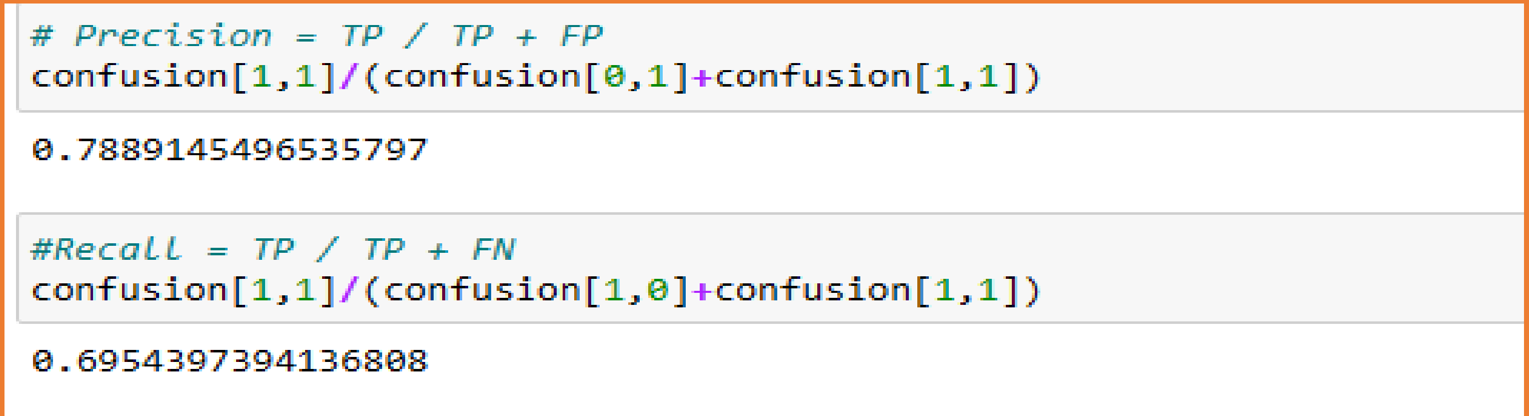
```
0.81511746680286
```

```
# Calculating the specificity
TN/(TN+FP)
```

```
0.8061926605504587
```

With the current cut off as 0.35 we have accuracy, sensitivity and specificity of around 81%

Precision-Recall



```
# Precision = TP / TP + FP
confusion[1,1]/(confusion[0,1]+confusion[1,1])
```

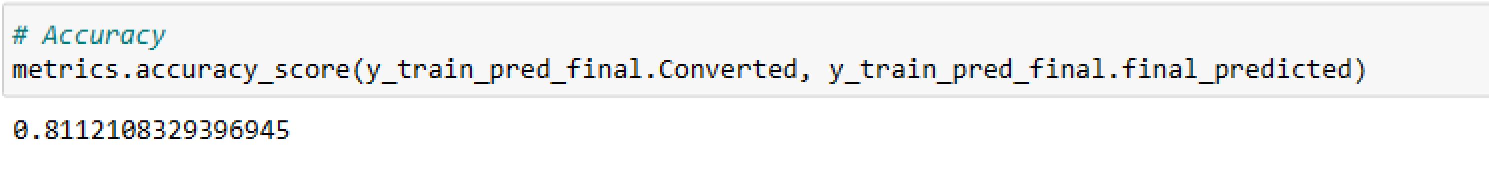
```
0.7889145496535797
```

```
#Recall = TP / TP + FN
confusion[1,1]/(confusion[1,0]+confusion[1,1])
```

```
0.6954397394136808
```

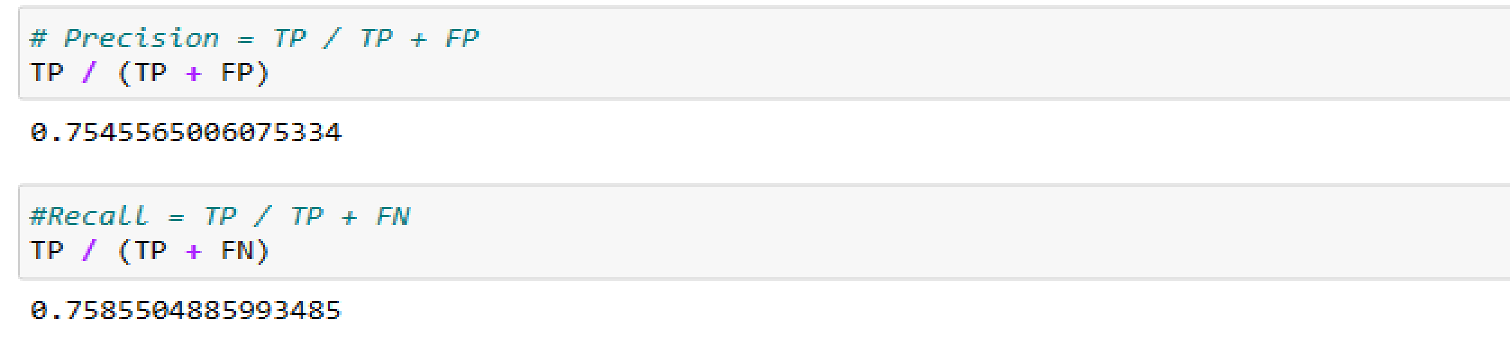
With the current cut off as 0.35 we have Precision around 78% and Recall around 70%

Precision and recall trade-off



```
# Accuracy
metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.final_predicted)
```

```
0.8112108329396945
```

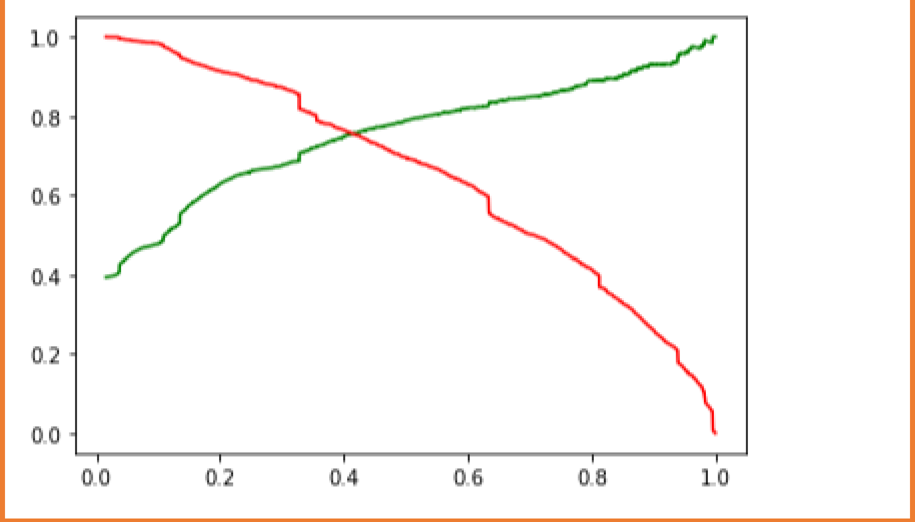


```
# Precision = TP / TP + FP
TP / (TP + FP)
```

```
0.7545565006075334
```

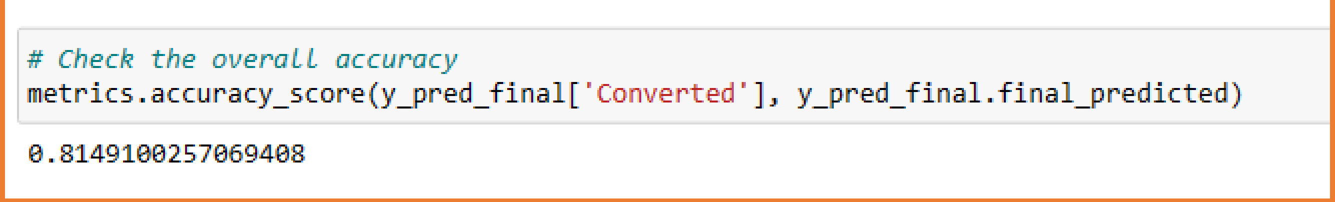
```
#Recall = TP / TP + FN
TP / (TP + FN)
```

```
0.7585504885993485
```



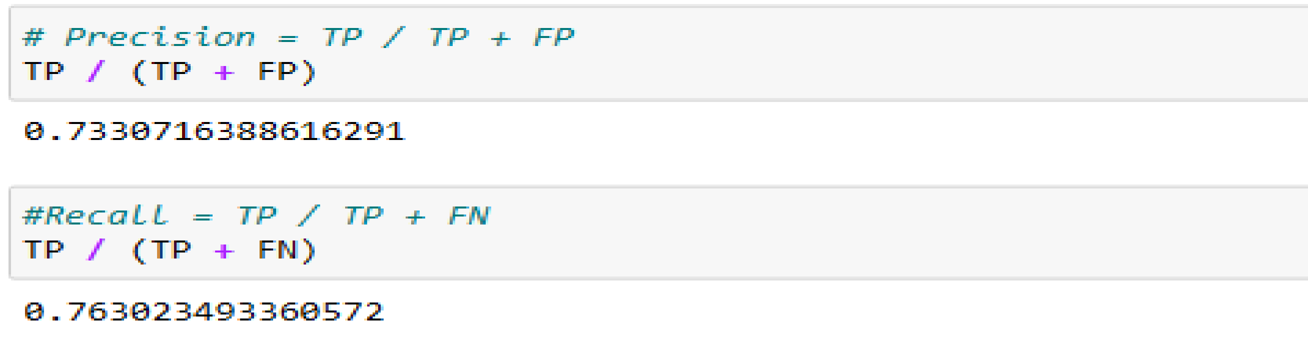
With the current cut off as 0.41 we have Precision around 75% and Recall around 76%

Prediction on Test set



```
# Check the overall accuracy
metrics.accuracy_score(y_pred_final['Converted'], y_pred_final.final_predicted)
```

```
0.8149100257069408
```



```
# Precision = TP / TP + FP
TP / (TP + FP)
```

```
0.7330716388616291
```

```
#Recall = TP / TP + FN
TP / (TP + FN)
```

```
0.763023493360572
```

With the current cut off as 0.41 we have Precision around 73% and Recall around 76%

Conclusion

Based on the analysis, the most influential variables in identifying potential buyers, ranked in descending order of importance, are as follows:

- Total time spent on the website.
- Total number of visits.
- Lead source: a. Google b. Direct traffic c. Organic search d. Welingak website
- Last activity: a. SMS b. Olark chat conversation
- Lead origin as Lead add format.
- Current occupation as a working professional.
- Taking these factors into consideration, X Education has a significant opportunity to convert a majority of potential buyers by focusing on these key aspects. By emphasizing the importance of the website experience, the number of visits, various lead sources, specific last activities, lead origin, and targeting working professionals, X Education can enhance its chances of successfully converting potential buyers into customers.