LEAD SCORING

CASE STUDY

NAME: DANG THI TRUC

CONTENT

01 PROBLEM STATEMENT

02 ANALYSIS APPROACH

03 DATA CLEANING & HANDLING

04 EXPLORATORY DATA ANALYSIS

05 DATA PREPARATION

06 MODEL BUILDING

07 MODEL EVALUATION

PROBLEM STATEMENT

X Education sells online courses to industry professionals who are interested in the courses. The typical lead conversion rate at X Education is around 30%. Although X Education gets a lot of leads, its lead conversion rate is very poor.

The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark estimate of the target lead conversion rate at around 80%.

ANALYSIS APPOACH

Develop Hypothesis

Although the company acquired a lot of leads, customers have to consider many elements before making a decision. The performance of the company is very poor for some reasons:

- Spending more time to take care of poor-quality customers instead of potential customers, as they don't identify who their customers are.
- Poor customer consulting service, etc.

Data Collection

When people fill out a form providing their email address or phone number, they are classified as leads or past referals, we can build up a dataset that contains key information about customers, including demographics, behaviors, and how they engaged with our service.

Problem Mapping

In this case, we will build a Logistic regression Model to predict whether customers have a high chance of being converted or not. It helps the sales team focus on the right leads, achieve the target.

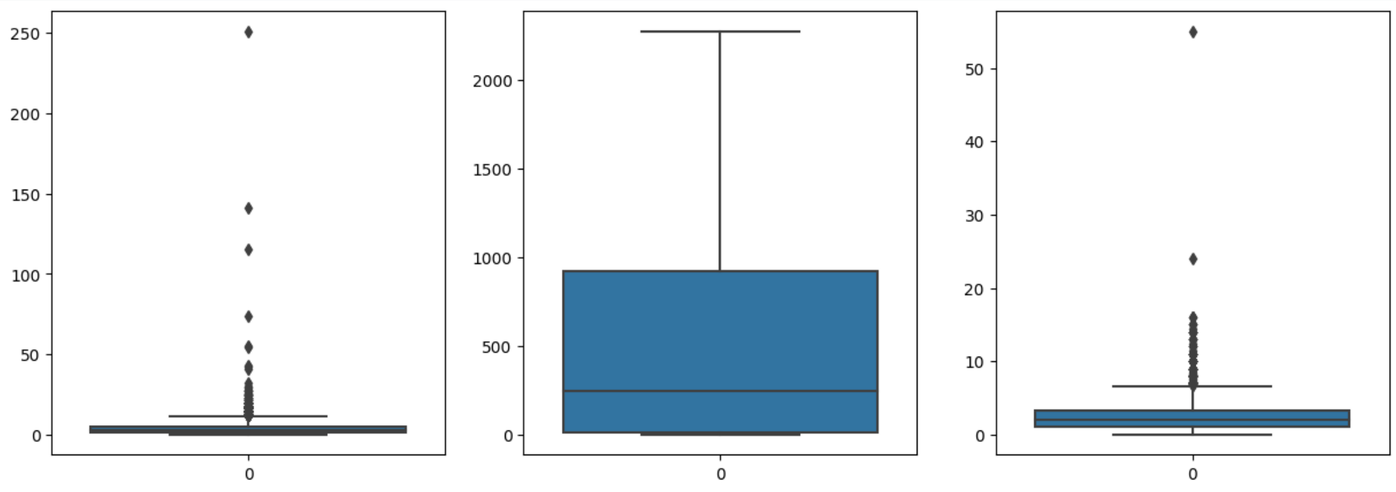
DATA CLEANING & HANDLING

Missing values

Asymmetrique Activity Index	45.65
Asymmetrique Profile Index	45.65
Asymmetrique Activity Score	45.65
Asymmetrique Profile Score	45.65

- Firstly, we drop columns contains more than and equal to 45% as they don't provide enough information.
- In case some columns contains '**Select**' and missing values, we will treat them as the same. If the total number of these is so high, we can drop that columns.

Outliers treatment

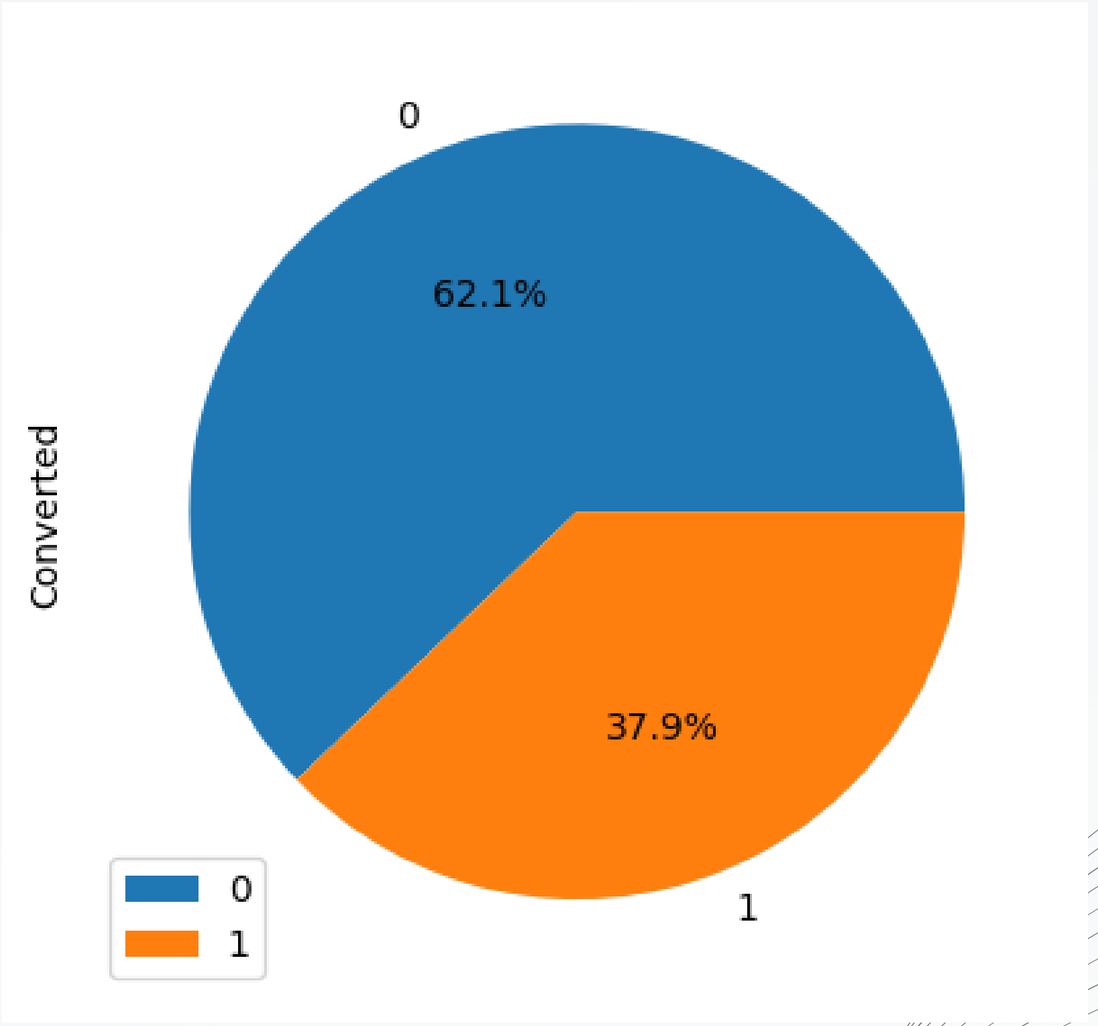


- TotalVists and Page Views Per Visit have outliers in the distribution displayed above. Because the difference between the 95% and 99% percentiles is insignificant, the outliers can be replaced with the 99% percentile value.

EXPLORATORY DATA ANALYSIS

TARGET VARIABLE - CONVERTED

For the purpose of analysis, we determined 'Converted' as the target variable.
Following the distribution above, the percentage of converted and not converted is not significantly different. So we can consider the dataset balanced.



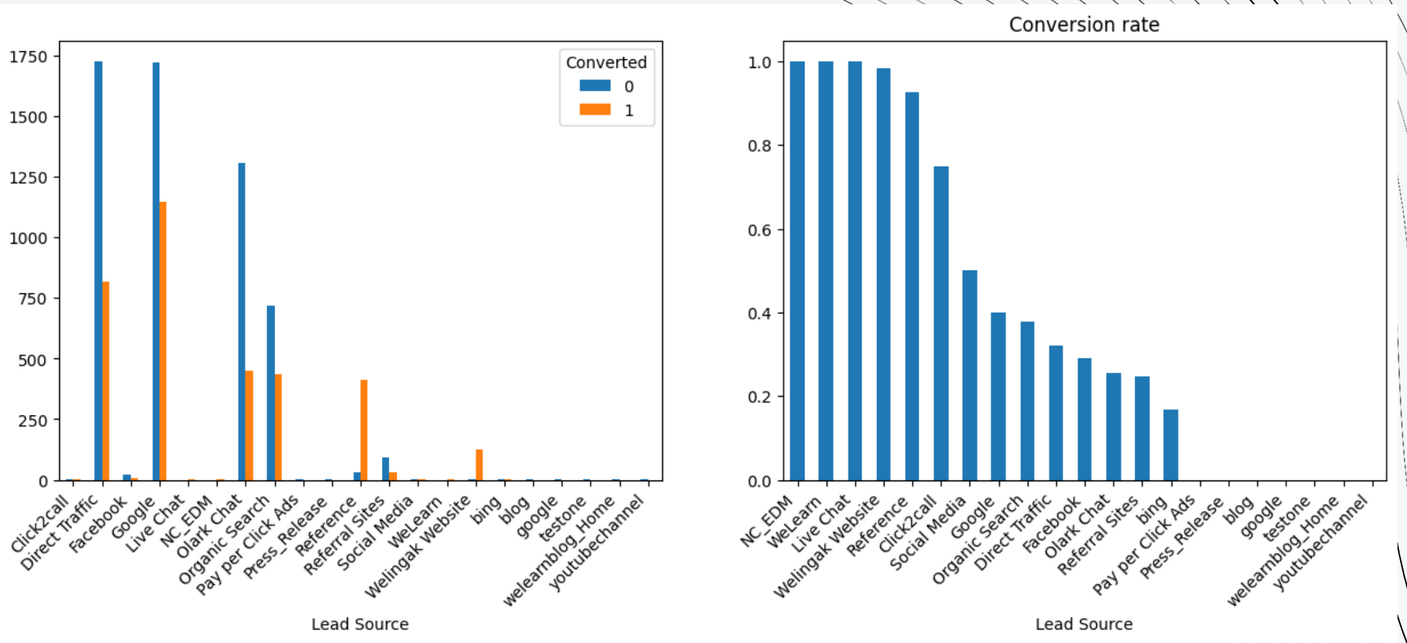
CATEGORY VARIABLE ANALYSIS

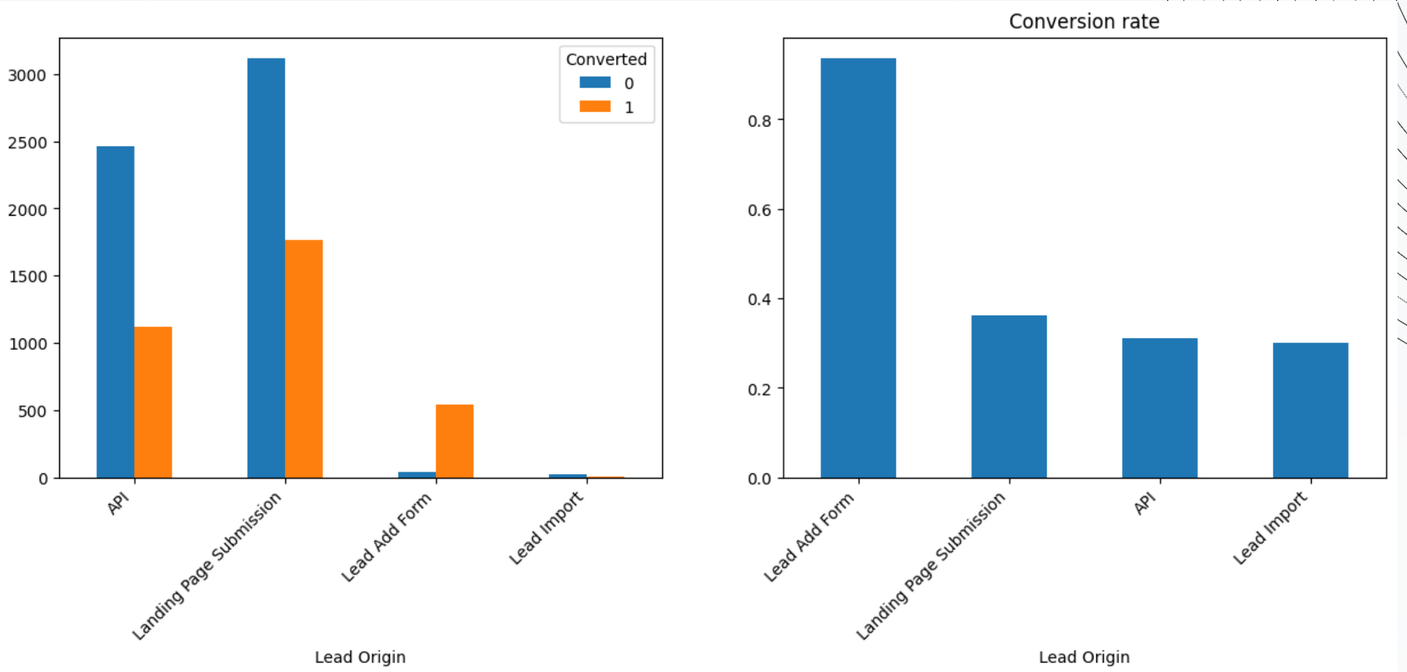
- **Lead Orgin:**

Landing Page Submission and API are two identifiers that have the most leads but a low conversion rate, whereas Lead Add Form has the highest conversion rate. Based on that, we can recommend that the company should improve the conversion rate of the landing page and API, while drive more traffic to Lead Add Form.

- **Lead Source:**

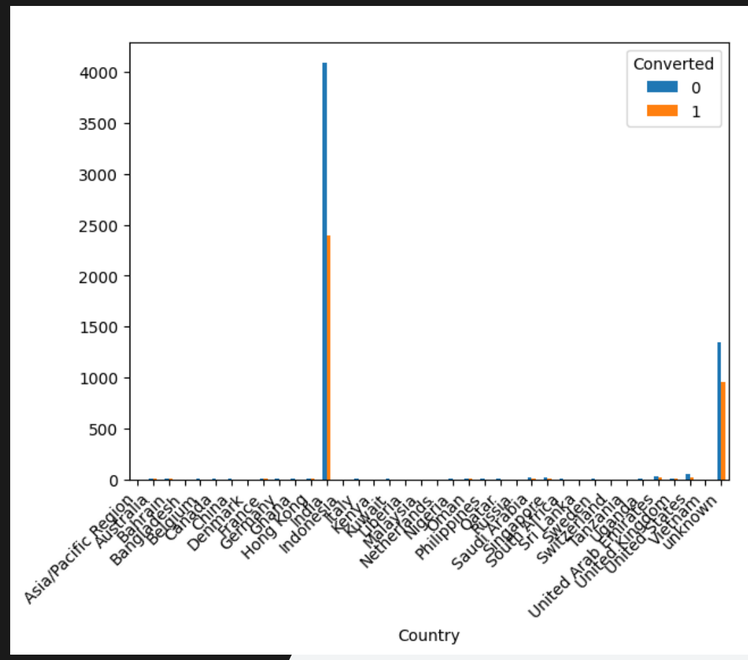
Most leads come from Direct Traffic, Google, Orlack Chat, and pay-per-click ads, but these channels don't have high conversion rates. The performance of these channels should be improved in the future. Reference is a reliable source with high conversion rate.

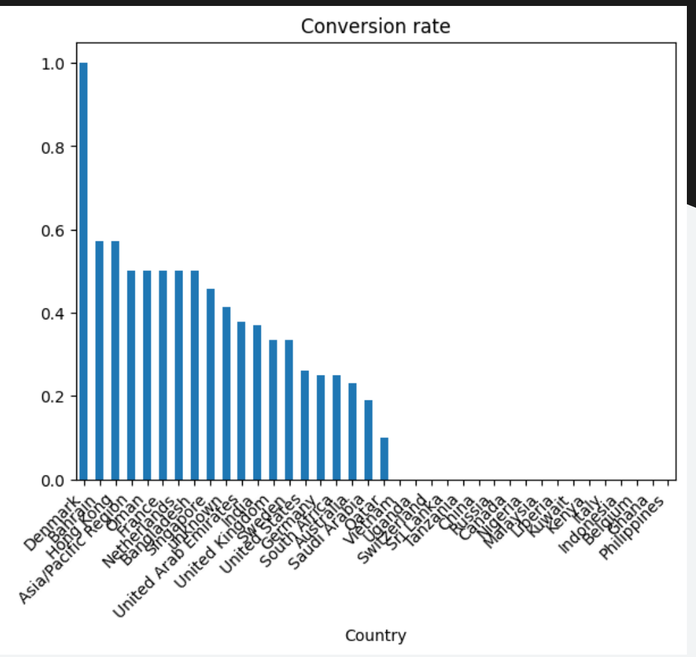


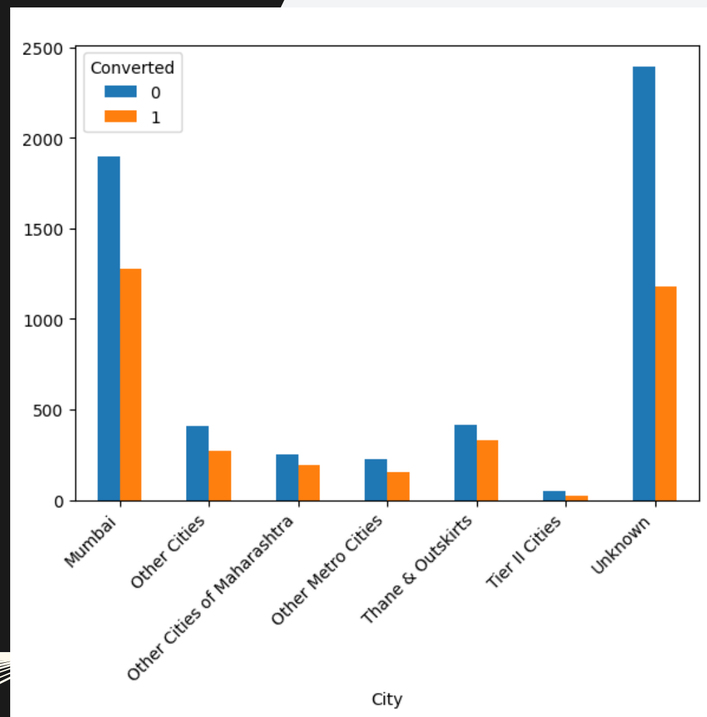


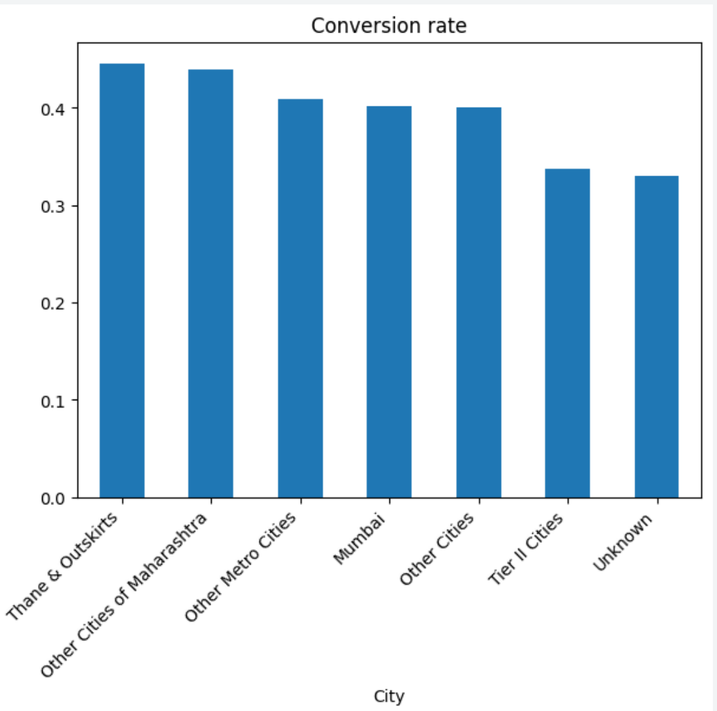
CATEGORY VARIABLE ANALYSIS

Combining the two variables, Country and City, we can determine that the key market for the company is India. While most customers lived in Mumbai and other top-tier cities, the conversion rate decreased from top-tier cities to low-tier cities and the outskirts.









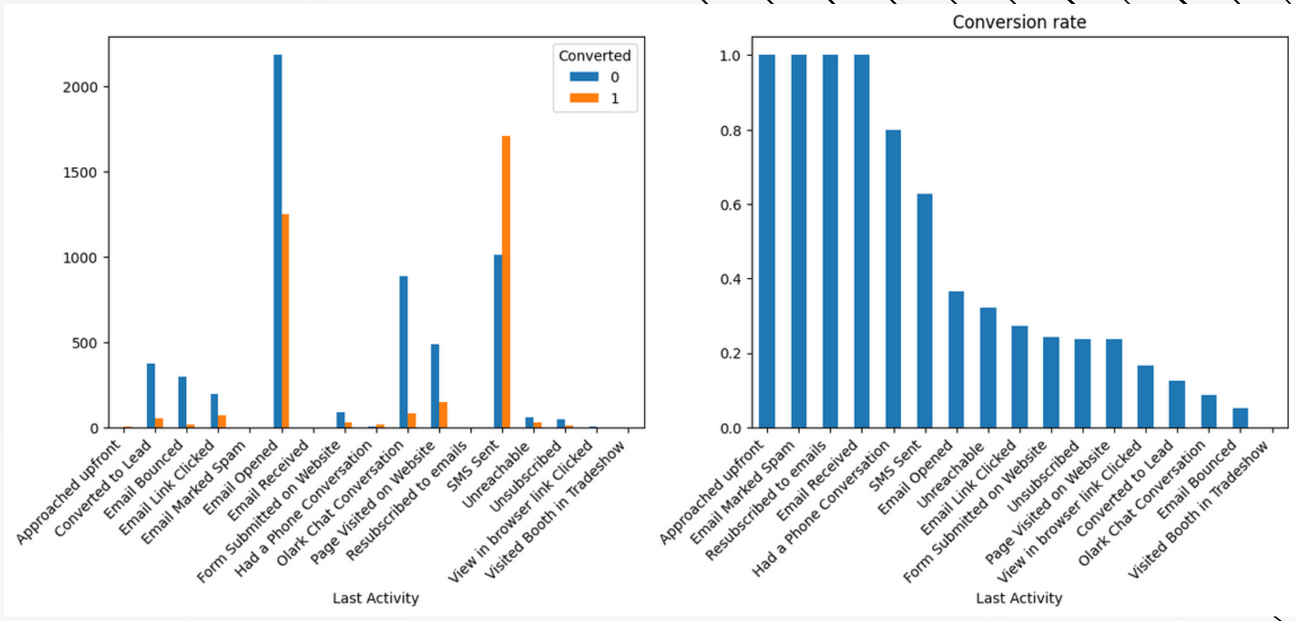
CATEGORY VARIABLE ANALYSIS

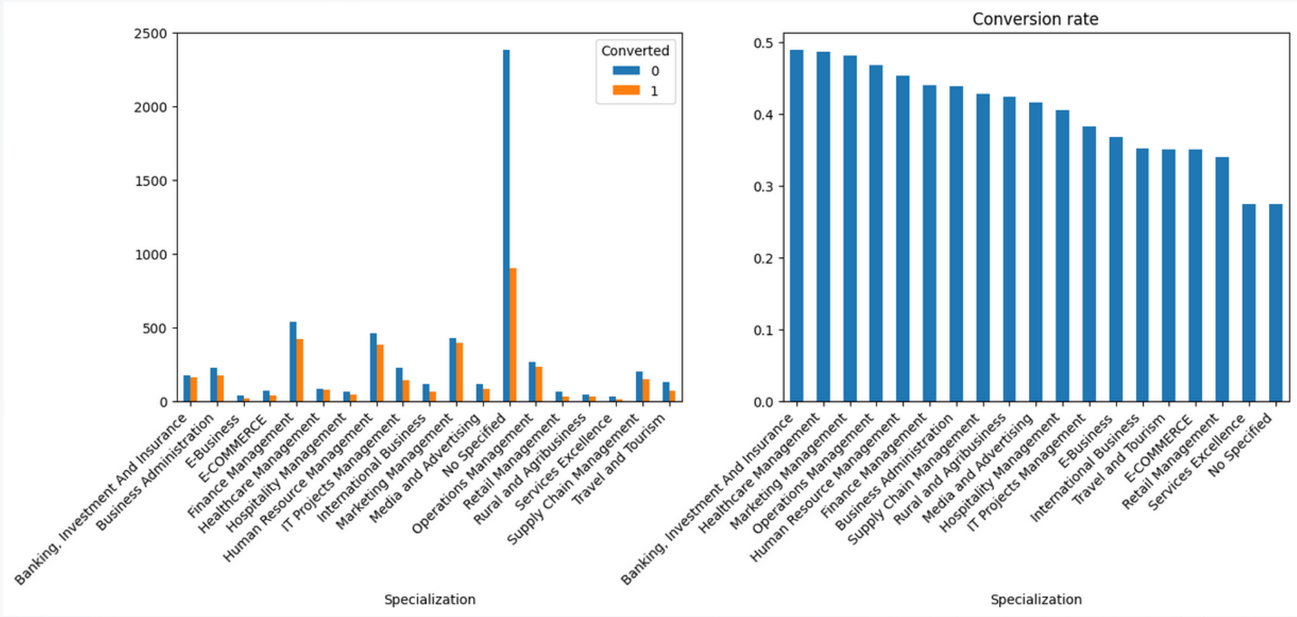
- **Last Acitivity:**

Following the distribution above, most customers engaged with Email, SMS, or Orlak Chat, but the group of customers who sent SMS had a higher conversion rate than the other two groups. It also demonstrated the behavior of customers interested in our online courses through the last activity they performed. For instance, customers who opened an email have a higher conversation rate than those who just received it. Customers who resubscribed to emails or had a phone conversation are more interested in our courses as well.

- **Specialization:**

Most customers don't fill out their specialization in the form, but based on the current distribution, we can see some domains have a good conversation rate, such as Banking, investment, and Insurance, Healthcare management, and Marketing Management. It can be affected by the content of the company's online course.





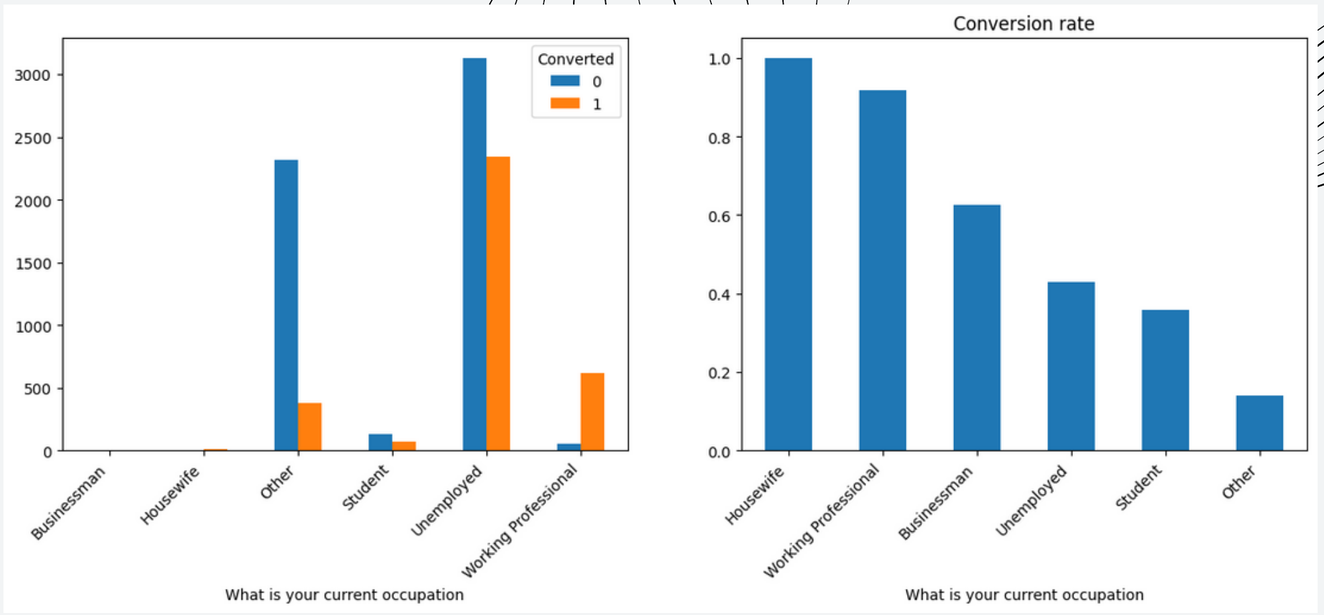
CATEGORY VARIABLE ANALYSIS

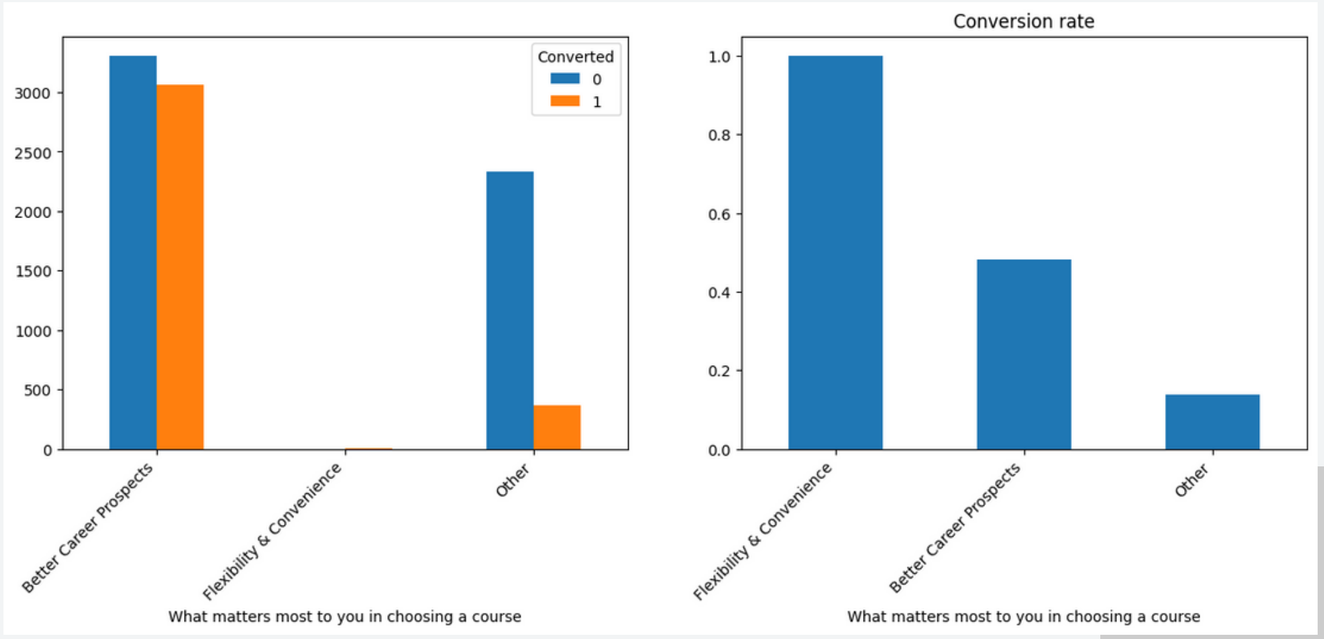
- **Occupation:**

The company reached out to a high number of unemployed customers, but the conversation rate is still low. Instead of that, housewife, working professionals, and businessmen should be the main target audiences. The company should review their approach strategy.

- **What matters most to you in choosing a course:**

Most customers choose our online courses because they want better career prospects. However, the stronger reasons for converted leads are flexibility and convenience. It's also one of the advantages of studying online compared to other ways.





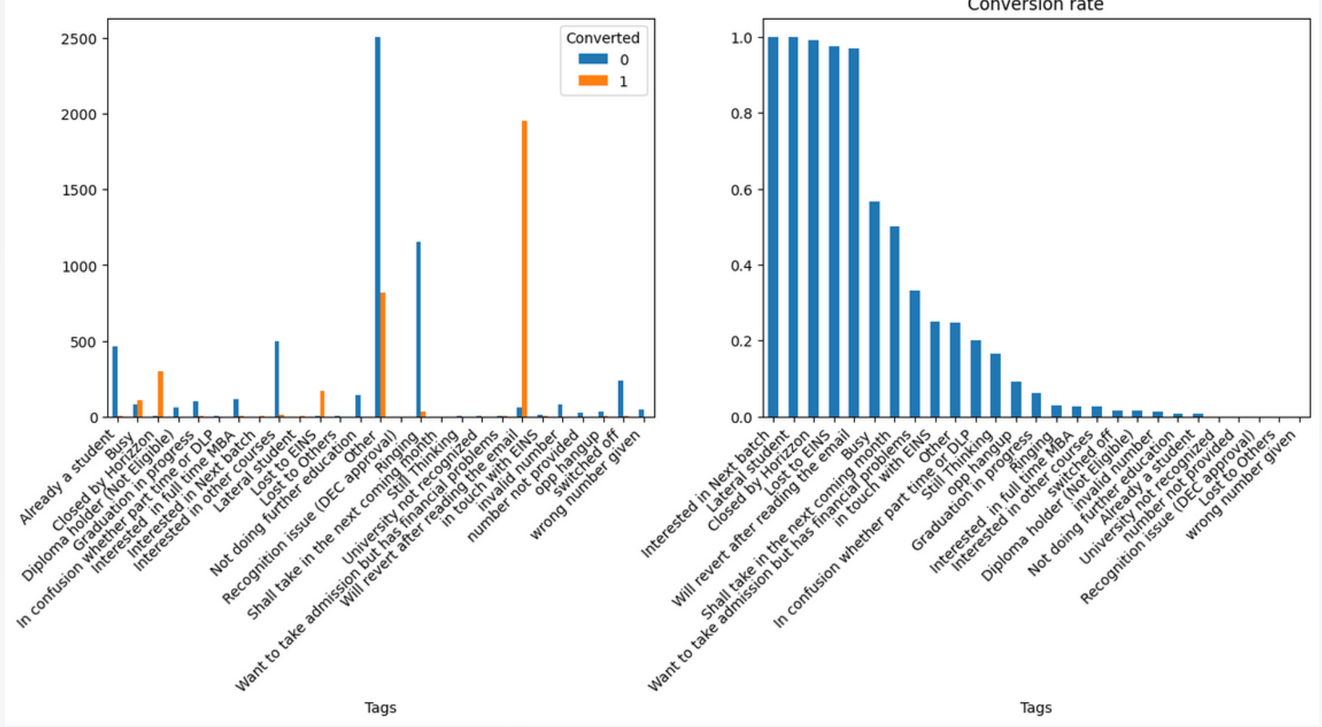
CATEGORY VARIABLE ANALYSIS

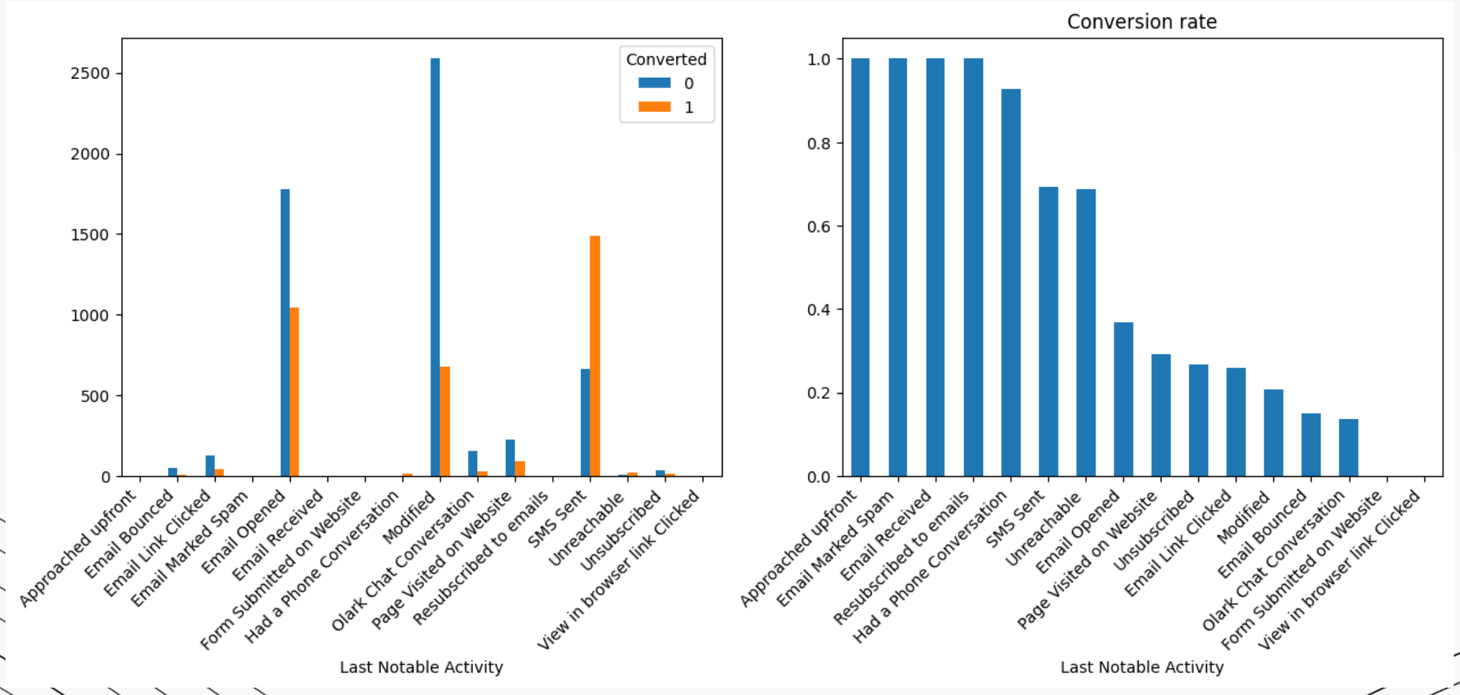
- **Tags:**

The variable indicates the current status of leads. Following each status, the sales team had already contacted them and defined potential leads. Also, the most common current status for leads is 'Will revert after reading the email'. Customers should take care of this group to have next step with them.

- **Last Notable Activity:**

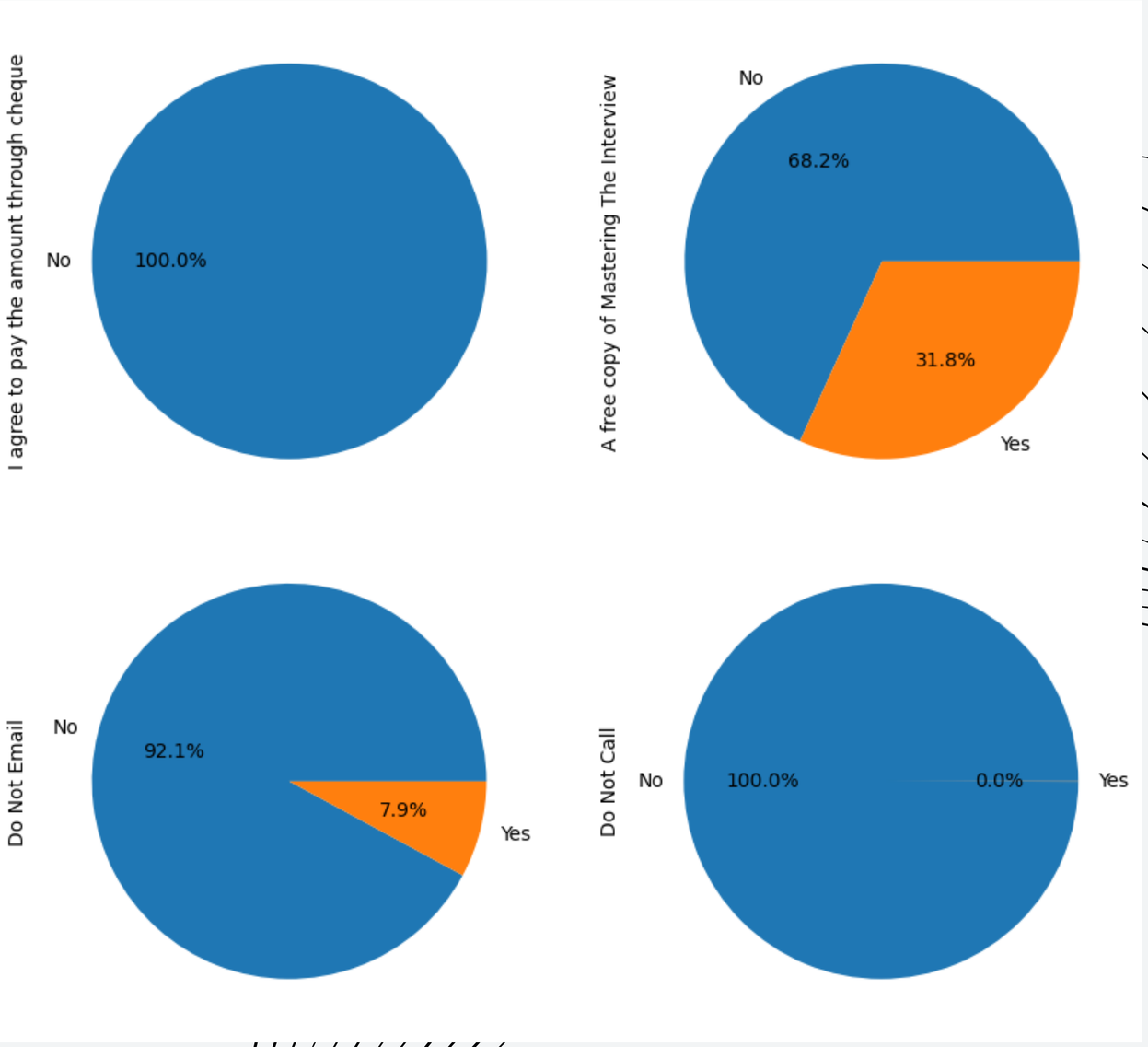
As the purpose of the analysis is to define who are potential leads when we approach them, we will not deep dive into this variable.





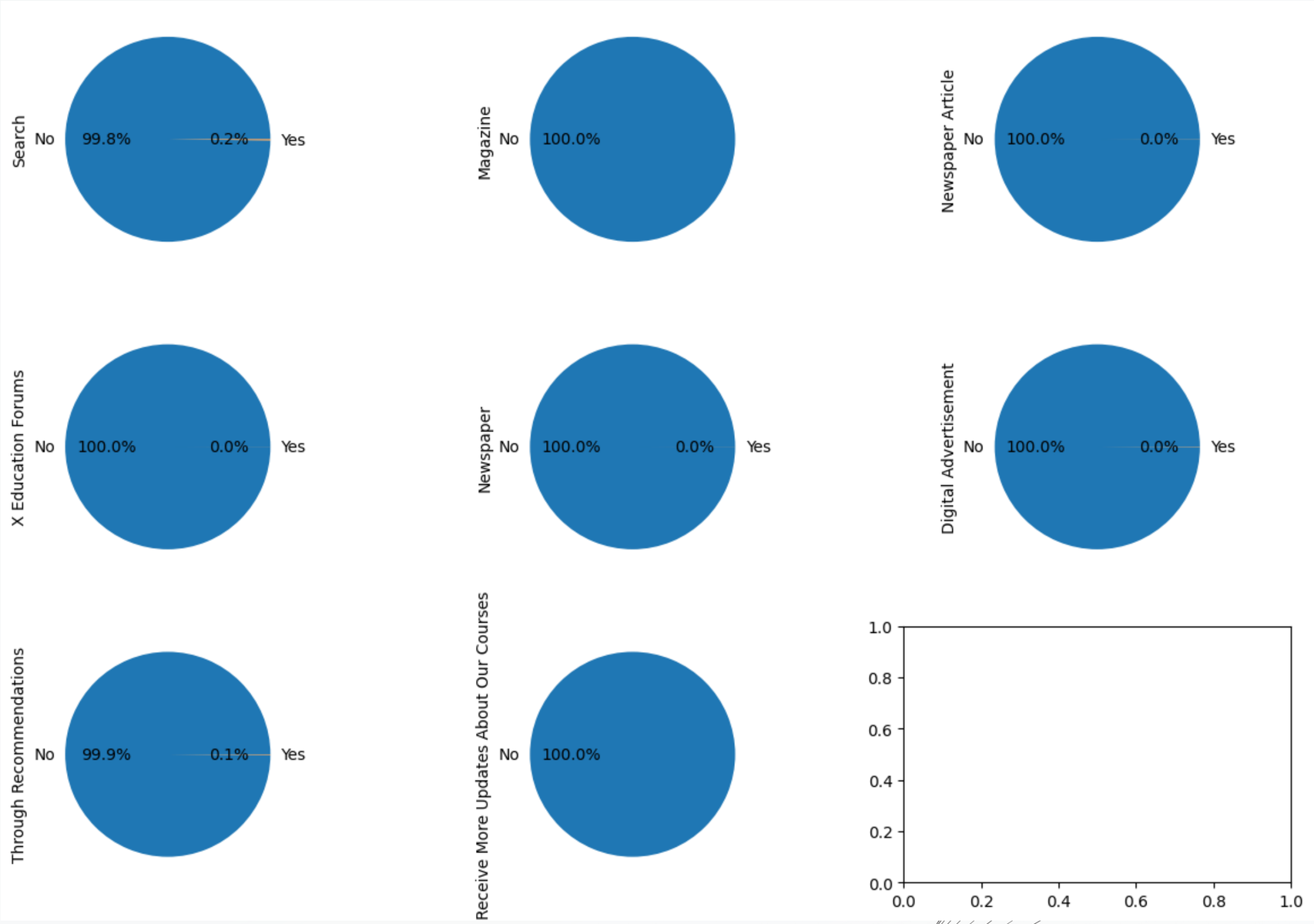
BINARY VARIABLE ANALYSIS

- Although we have too many binary variables, there are only two that are clearly classified. Most customers don't want to receive email about the course or free copy of 'Mastering the Interview'.



BINARY VARIABLE ANALYSIS

All these variables are not classified, so we will drop them out of the dataset in the next step.



NUMERICAL VARIABLE ANALYSIS

- **Total Visits:**

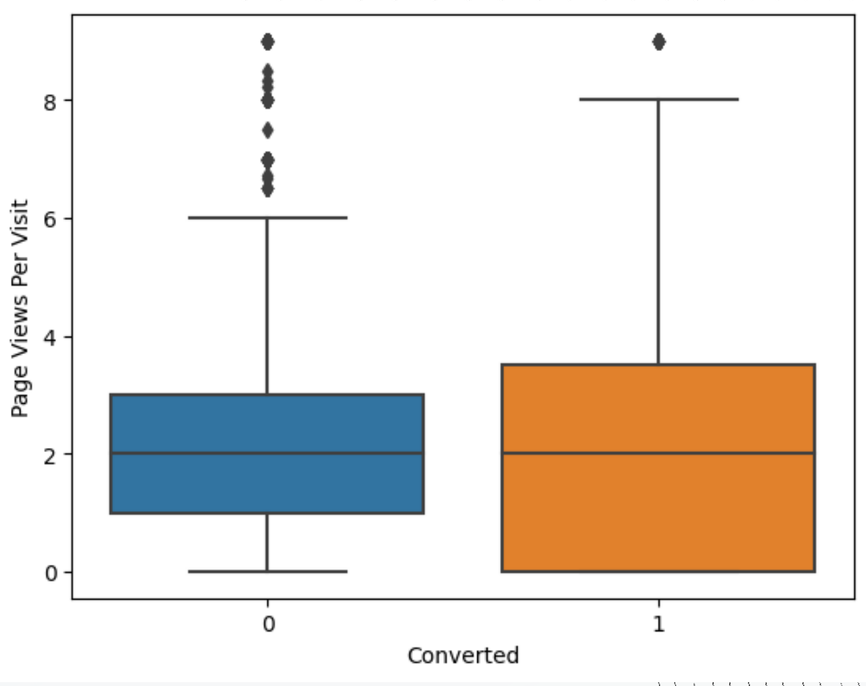
Although the median of converted and non-converted leads is the same, we can determine that customers who are interested in our course will visit the website more often than those who are not.

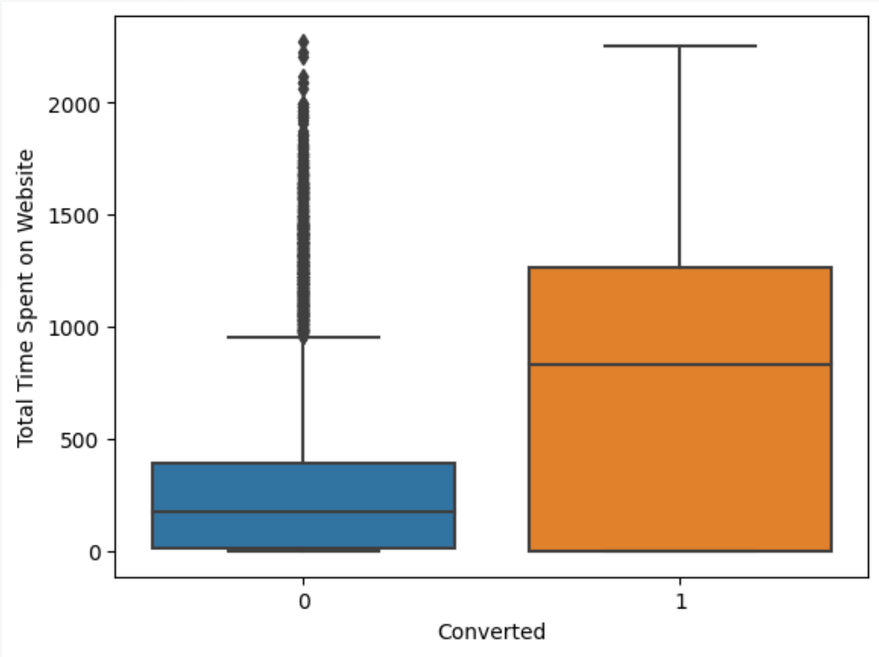
- **Total Time Spent on Website**

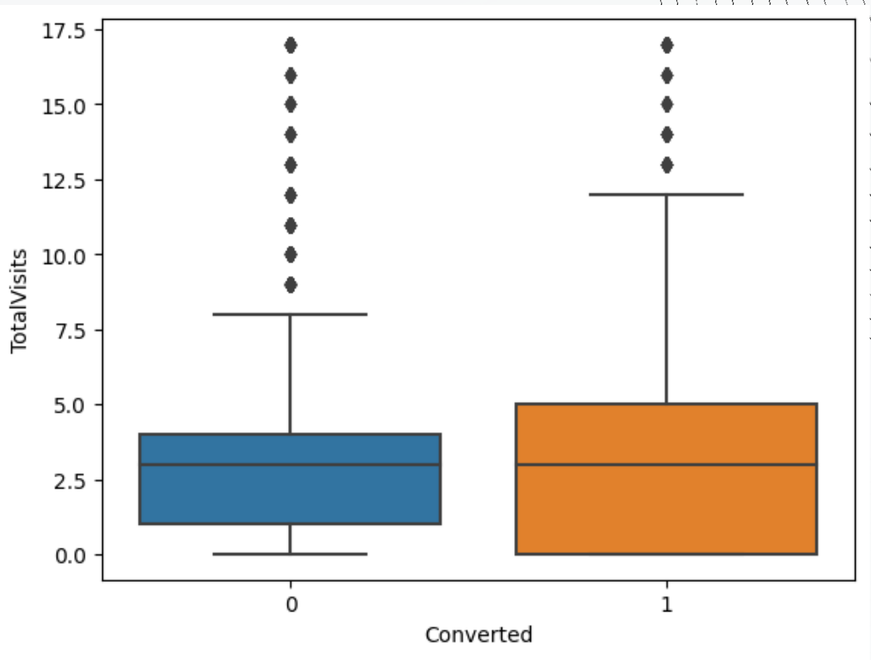
The customers who are interested in our course spent more time on the landing page to carefully read the information.

- **Page Views Per Visit:**

There is no significant difference between the average number of pages viewed during a visit by the two types of customers.

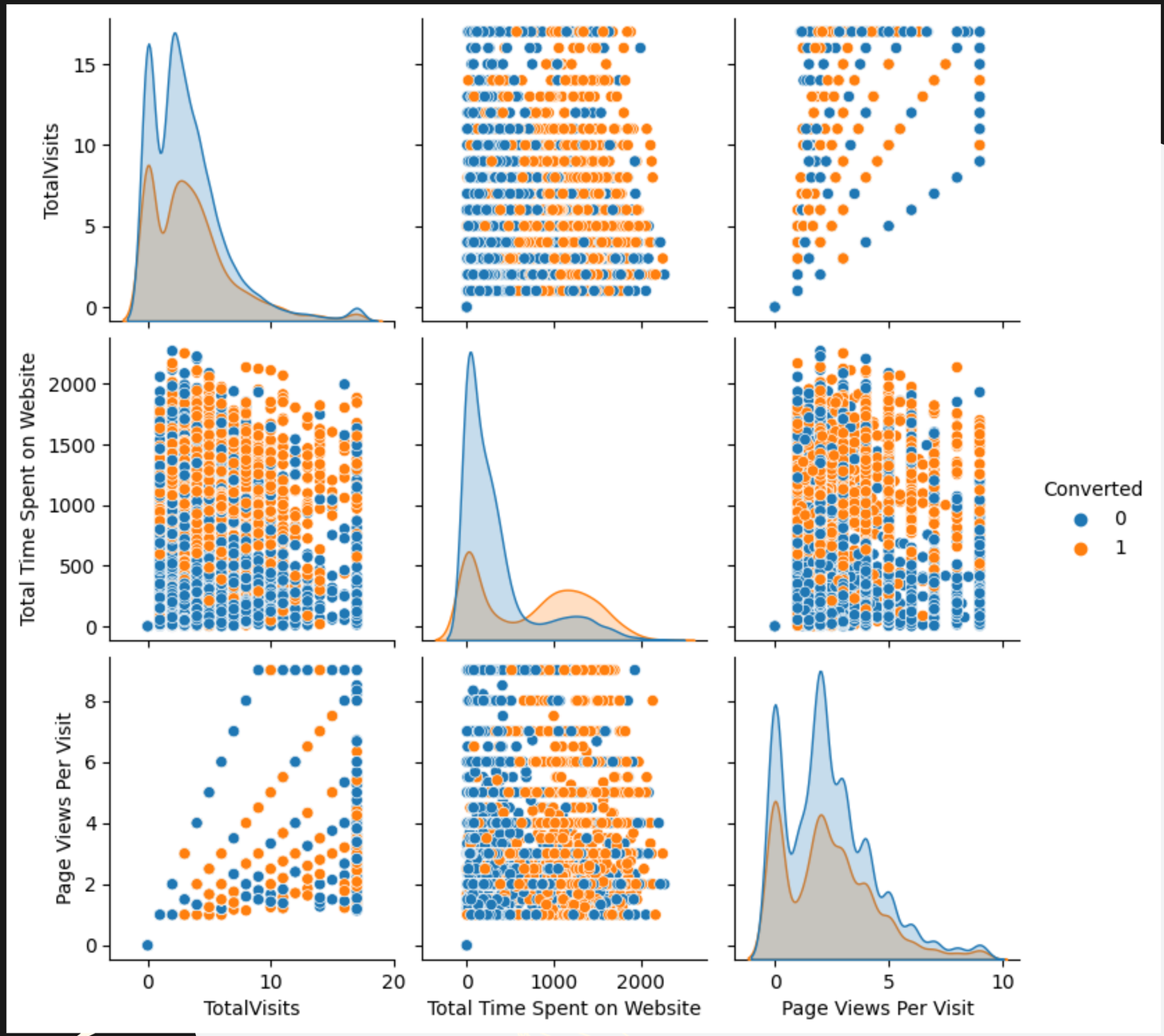






BIVARIATE ANALYSIS

There is a certain correlation between Total Visits and Page Views Per Visit. Actually, we can see that if a customer is interested in the course, they will usually visit the landing page and go to the other pages to find more information.



DATA PREPARATION

No	Step	Describe
1	Drop unimportant columns	After the EDA part, we define that there are some variables that we should remove from the dataset because they will not contribute to our model.
2	Convert Binary Variables	After removing all unimportant variables, we just have 2 binary variables: Do Not Email and A free copy of Mastering The Interview. To add them to the model, we have to convert the value Yes/No to 1/0.
3	Check correlation	We have defined some variables that are correlated with each other, but we can leave out important imformation if we drop them, so we still keep them in the dataset and remove them by applying Feature selection.
4	Train-test split	Apply train_test_split from sklearn, we will have X_train, X_test, y_train, y_test with train_size = 0.7, test_size = 0.3 and random_state = 100
5	Feature Scalling	With all numerical variables, we will apply the standardization method to scale the data.

MODEL BUILDING

FEATURE SELECTION

After the data preparation step, we have a dataset with 76 columns. We have to use RFE to reduce the number of variables and find the top 15 high-contributing variables.

Next, we apply manual feature selection to drop high p-value and high VIF columns. Finally, we have 14 variables.

```
col
✓ 0.0s
Index(['Do Not Email', 'Total Time Spent on Website',
       'Lead Origin_Landing Page Submission', 'Lead Origin_Lead Add Form',
       'Lead Source_Olark Chat', 'Lead Source_Welingak Website',
       'Last Activity_Converted to Lead',
       'Last Activity_Had a Phone Conversation',
       'Last Activity_Olark Chat Conversation', 'Last Activity_SMS Sent',
       'Last Activity_Unsubscribed', 'Specialization_No Specified',
       'What is your current occupation_Housewife',
       'What is your current occupation_Working Professional',
       'What matters most to you in choosing a course_Other'],
      dtype='object')
```

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.1476	0.125	-1.183	0.237	-0.392	0.097
Do Not Email	-1.7852	0.186	-9.612	0.000	-2.149	-1.421
Total Time Spent on Website	1.1107	0.041	27.131	0.000	1.030	1.191
Lead Origin_Landing Page Submission	-0.9941	0.129	-7.735	0.000	-1.246	-0.742
Lead Origin_Lead Add Form	3.0332	0.234	12.969	0.000	2.575	3.492
Lead Source_Olark Chat	1.2318	0.125	9.886	0.000	0.988	1.476
Lead Source_Welingak Website	2.4850	0.756	3.286	0.001	1.003	3.967
Last Activity_Converted to Lead	-1.2502	0.225	-5.568	0.000	-1.690	-0.810
Last Activity_Had a Phone Conversation	2.6143	0.759	3.445	0.001	1.127	4.102
Last Activity_Olark Chat Conversation	-1.3743	0.168	-8.169	0.000	-1.704	-1.045
Last Activity_SMS Sent	1.2650	0.076	16.615	0.000	1.116	1.414
Last Activity_Unsubscribed	1.3847	0.473	2.928	0.003	0.458	2.312
Specialization_No Specified	-0.9161	0.126	-7.245	0.000	-1.164	-0.668
What is your current occupation_Working Professional	2.4130	0.194	12.451	0.000	2.033	2.793
What matters most to you in choosing a course_Other	-1.2259	0.089	-13.753	0.000	-1.401	-1.051

PREDICTION ON TRAIN SET

Based on the final model, we predict y_train_pred and assign it a Converted Probability. This value will help us define potential leads with a lead score in the future.

Firstly, we can evaluate the model with the value of 0.5 and calculate accuracy, sensitivity, specificity, and so on.

- Accuracy: 81,7 %
- Sensitivity: 70,9 %
- Specificity: 88,4 %
- False positive rate: 11,5 %
- Positive predictive value: 79,4%
- Negative predictive value: 82,9 %

With the cut-off point, the sensitivity is quite low (just 0.70), while the CEO expected that the conversion rate would be around 80%. So it can't meet the demand.

```
y_train_pred_final = pd.DataFrame({'Converted':y_train.values, 'Converted_Prob':y_train_pred})
y_train_pred_final['Prospect ID'] = y_train.index
y_train_pred_final.head()
```

[94] ✓ 0.0s

	Converted	Converted_Prob	Prospect ID
0	0	0.072723	3009
1	0	0.037120	1012
2	0	0.609706	9226
3	1	0.673174	4750
4	1	0.914057	7987

```
# Confusion matrix
confusion = metrics.confusion_matrix(y_train_pred_final.Converted, y_train_pred_final.predicted )
print(confusion)
```

✓ 0.0s

```
[[3455  450]
 [ 710 1736]]
```

```
print(metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.predicted))
```

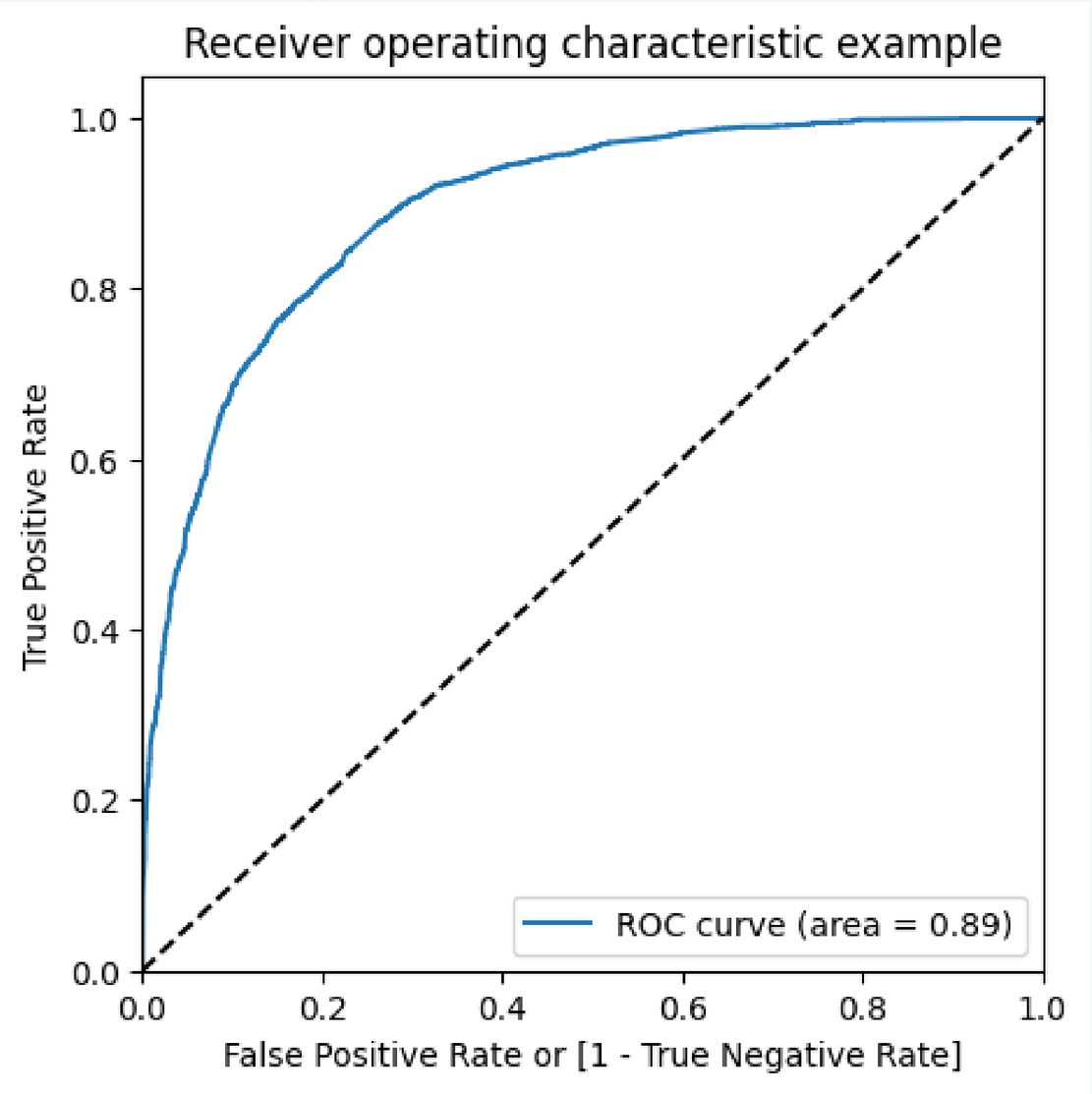
✓ 0.0s

```
0.817351598173516
```

```
TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives
```

ROC CURVE

As the area under the curve indicates how good model is, we can detemine our model is good with area = 0.89

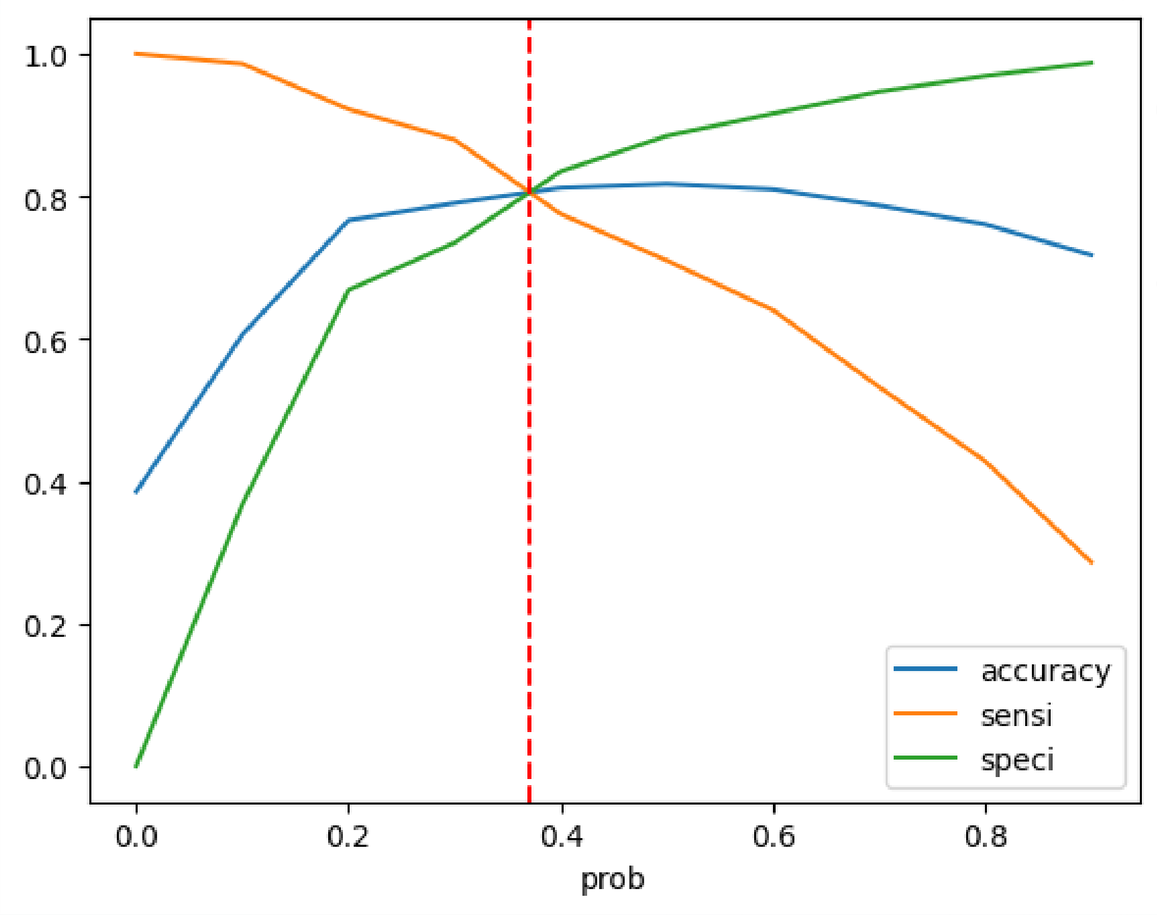


OPTIMIAL CUT-OFF POINT

We define the cut-off point is 0.37 at intersection of accuracy, sensitivity and specificity.
The evaluation matrix we have if we apply to our model:

- Accuracy: 80,7 %
- Sensitivity: 79,2 %
- Specificity: 81,7 %
- False positive rate: 18,2 %
- Positive predictive value: 73,1%
- Negative predictive value: 86,2 %

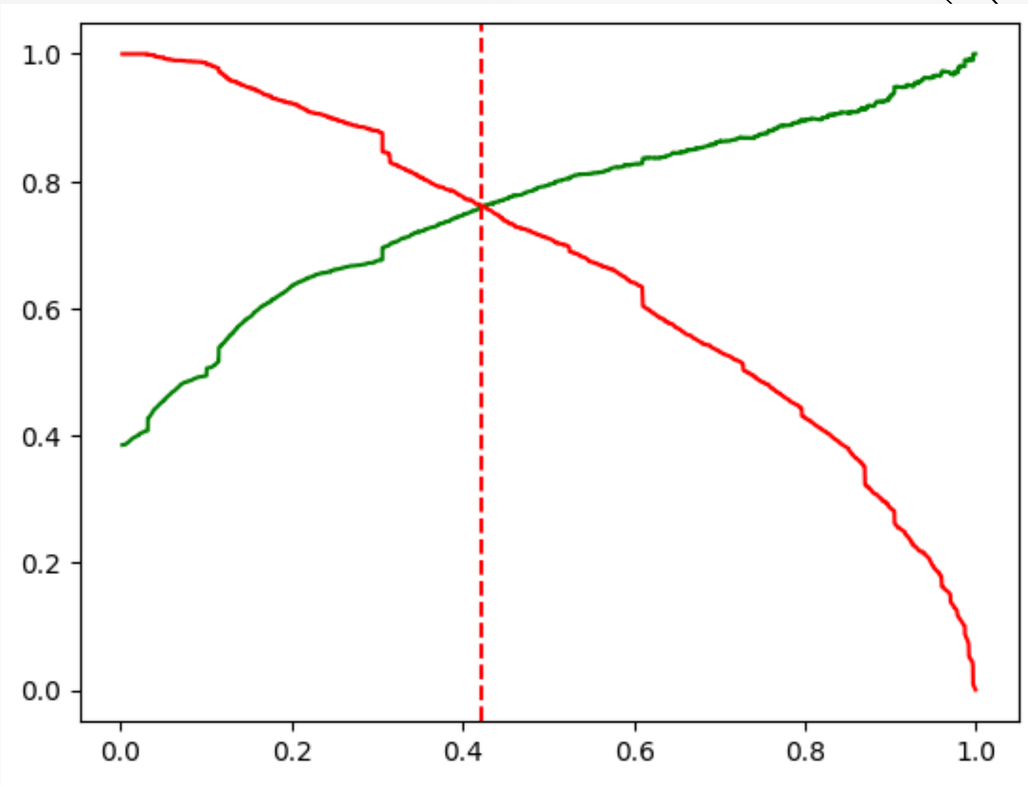
At about a threshold of 0.37, the curves of accuracy, sensitivity and specificity intersect, and they all take a value of around 79-81%.



PRECISION AND RECALL TRADEOFF

We try another method, precision and recall, to find out the possibilities of threshold. Now we have the cut-off point of 0.47.
Precision score: 75.9 %
Recall score: 0.76

The Recall score is only 0.76, lower than the CEO's expectation of 80%. So we just keep the cut-off point at 0.37.



MODEL EVALUATION

ON TEST SET

Based on the model we have on the train set, we evaluate the model with test set. Applying evaluation matrix, we have:

- Accuarcy: 81,1 %
- Sensivitity: 79,06 %
- Specificity: 82,3 %

These values are close to the value of train set.

```
[139]  y_pred_final['final_predicted'] = y_pred_final.Converted_Prob.map(lambda x: 1 if x > 0.37 else 0)
       ✓ 0.0s

[140]  y_pred_final.head()
       ✓ 0.0s
```

	Converted	Prospect ID	Converted_Prob	final_predicted
0	0	3271	0.049419	0
1	1	1490	0.966673	1
2	0	7936	0.042053	0
3	1	4216	0.869731	1
4	0	3830	0.046513	0

CONCLUSION

Finally, we have built Logisitc Regression model to predict customers who are potential to convert.

- The key variables to determine a potential customers or not are:
  + Do Not Email
  + Total Time Spent on Website
  + Lead Origin_Landing Page Submission
  + Lead Origin_Lead Add Form
  + Lead Source_Olark Chat
  + Lead Source_Welingak Website
  + Last Activity_Converted to Lead
  + Last Activity_Had a Phone Conversation
  + Last Activity_Olark Chat Conversation
  + Last Activity_SMS Sent
  + Last Activity_Unsubscribed
  + Specialization_No Specified
  + What is your current occupation_Working Professional
  + What matters most to you in choosing a course_Other

- The evaluation matrix of those is really close to each other, and combined with the area under the curve, it indicates how good the model we built is.
- Based on the converted probability, we can calculate the lead score to help the sales team choose the right customers and enhance their performance.
- The target conversion rate of the CEO is possible to reach.

RECOMMENDATION

Combining EDA and Model building, we can define some issues that the company should address to improve conversion rates:

- Grab more leads from high-quality sources:Lead Add Form is a good choice rather than API and Landing page submission.
- Instead of investing money in a variety of channels, the company should focus on some key channels like Olark Chat and the Welingak website.
- Last Activity and Total time spent on the website are good variables to define the behavior of customers and indicate whether they are interested in our course or not.
- Working professionals are still appropriate target audiences for our online course.
- Olark Chat is a good source to grab leads, but sales teams should communicate with customers through other methods to increase conversion rates.

THANK YOU!