LEAD SCORING CASE STUDY

Presented By:

❖ Shwettha RR

❖ Uttam Chatterjee

❖ Saumya Sonal

CONTENT

1. Problem Statement
2. Business Objective/ Goal
3. Solution Methodology
4. EDA
5. Categorical Variable Relation
6. Model Building
7. Top Factors that Impact the Conversion of Leads.
8. Summary.

PROBLEM STATEMENT

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

- ❑ 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.
- ❑ Model to achieve target lead conversion rate to be around 80%.

SOLUTION METHODOLOGY

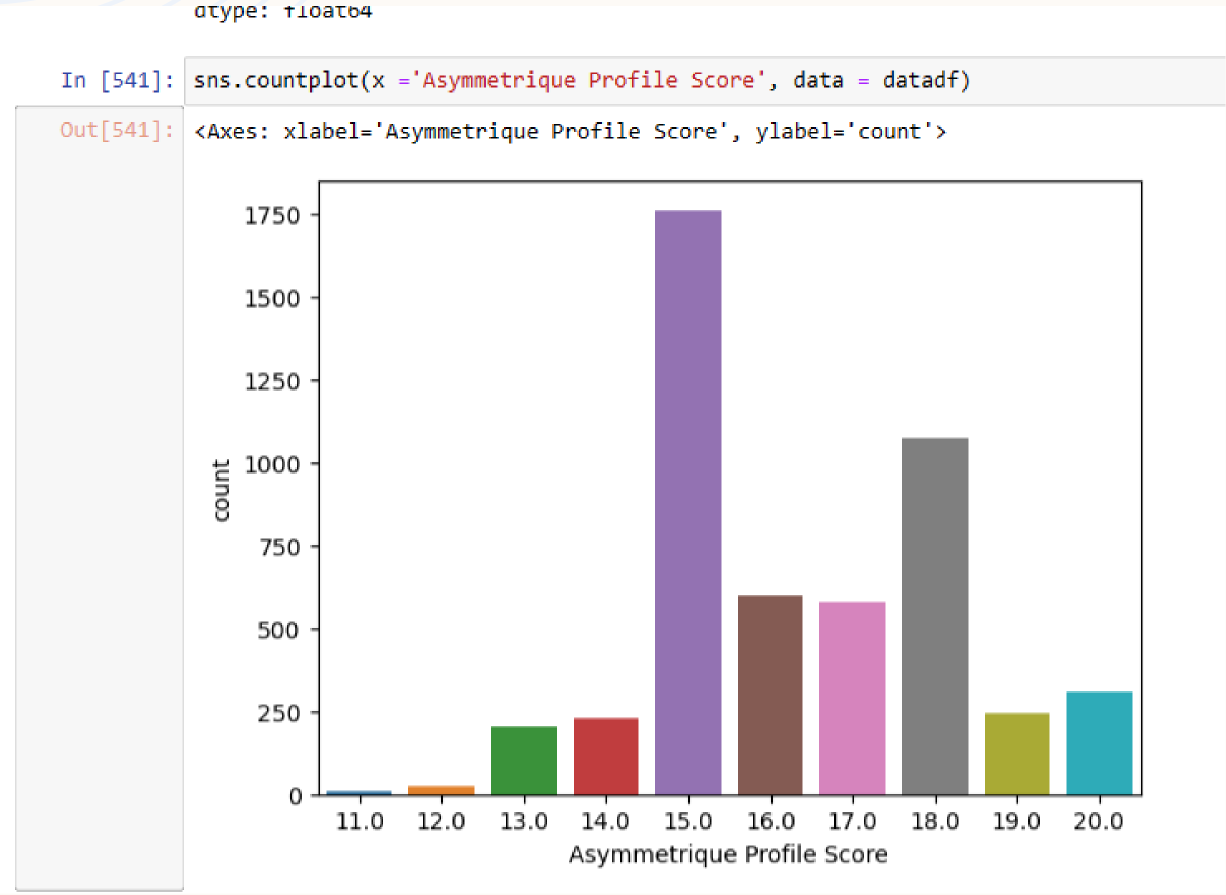
Data cleaning and data manipulation.

1. Check and handle duplicate data.
2. Check and handle NA values and missing values.
3. Drop columns, if it contains a large number of missing values and are not useful for the analysis.
4. Imputation of the values, if necessary.
5. Check and handle outliers in data.

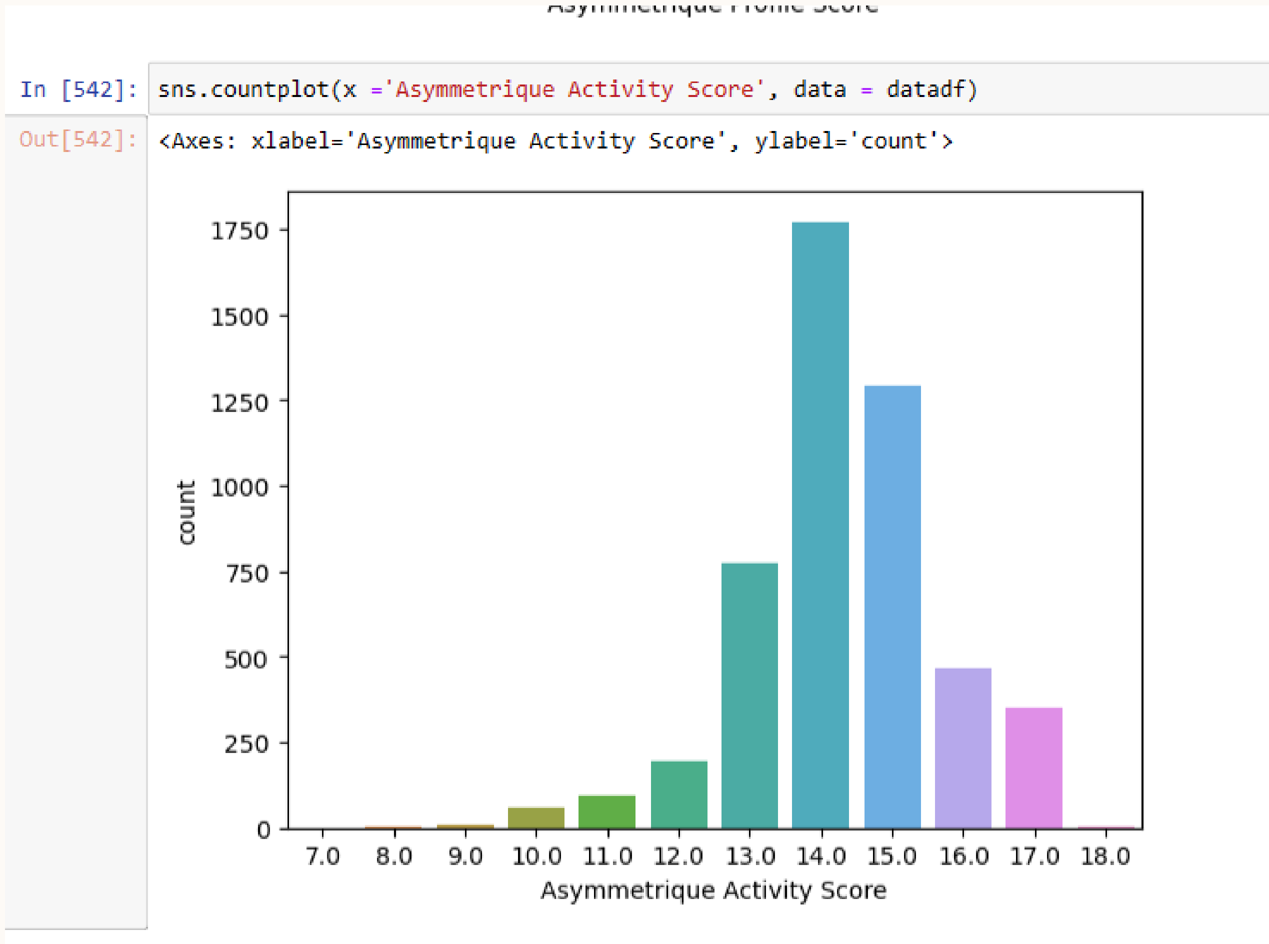
EDA

1. Univariate data analysis: value count, distribution of variables etc.
2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.
3. Feature Scaling & Dummy Variables and encoding of the data.
4. Classification technique: logistic regression is used for the model making and prediction.
5. Validation of the model.
6. Model presentation.
7. Conclusions and recommendations.

EDA

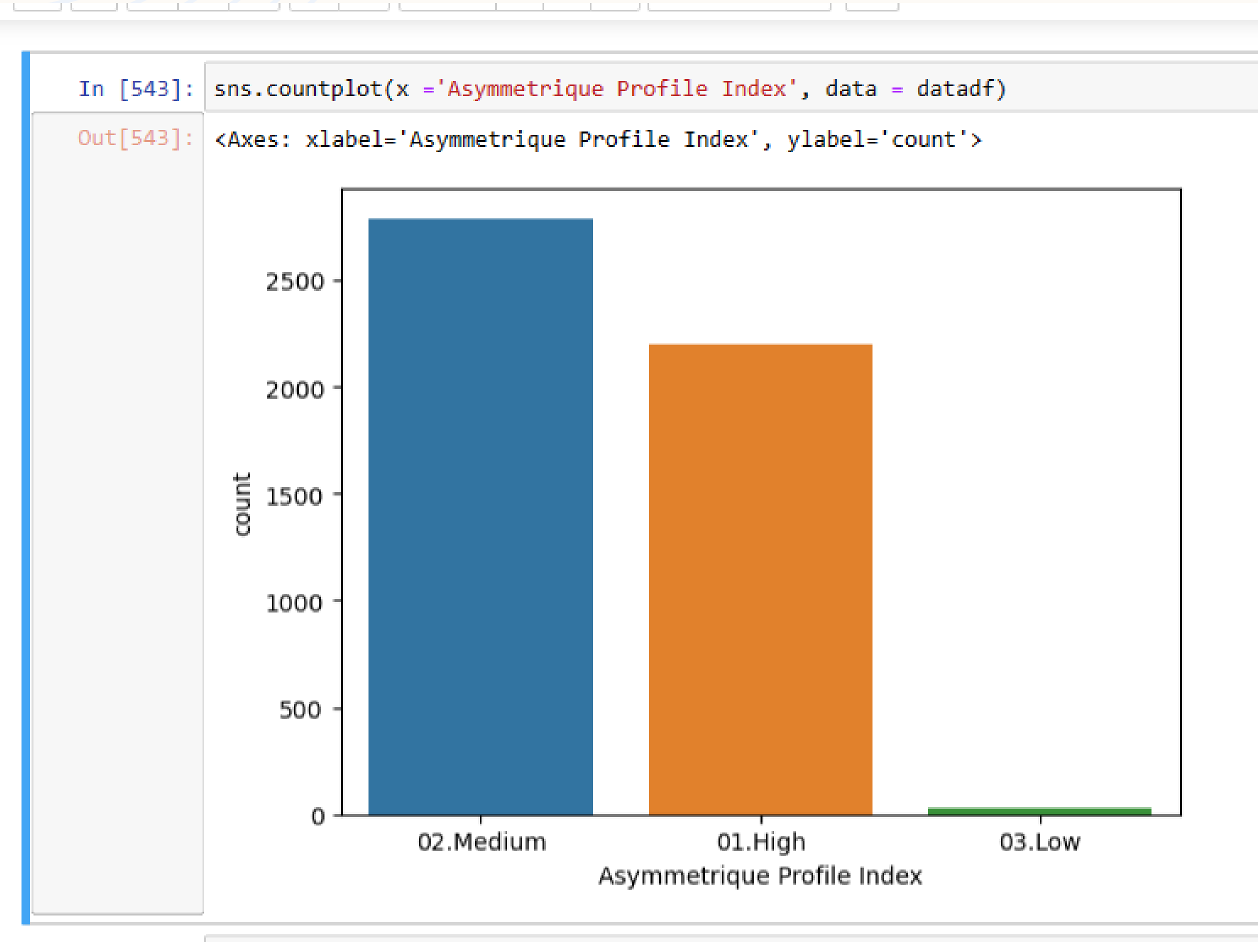


```
dtype: float64

In [541]: sns.countplot(x ='Asymmetrique Profile Score', data = datadf)

Out[541]: <Axes: xlabel='Asymmetrique Profile Score', ylabel='count'>
```

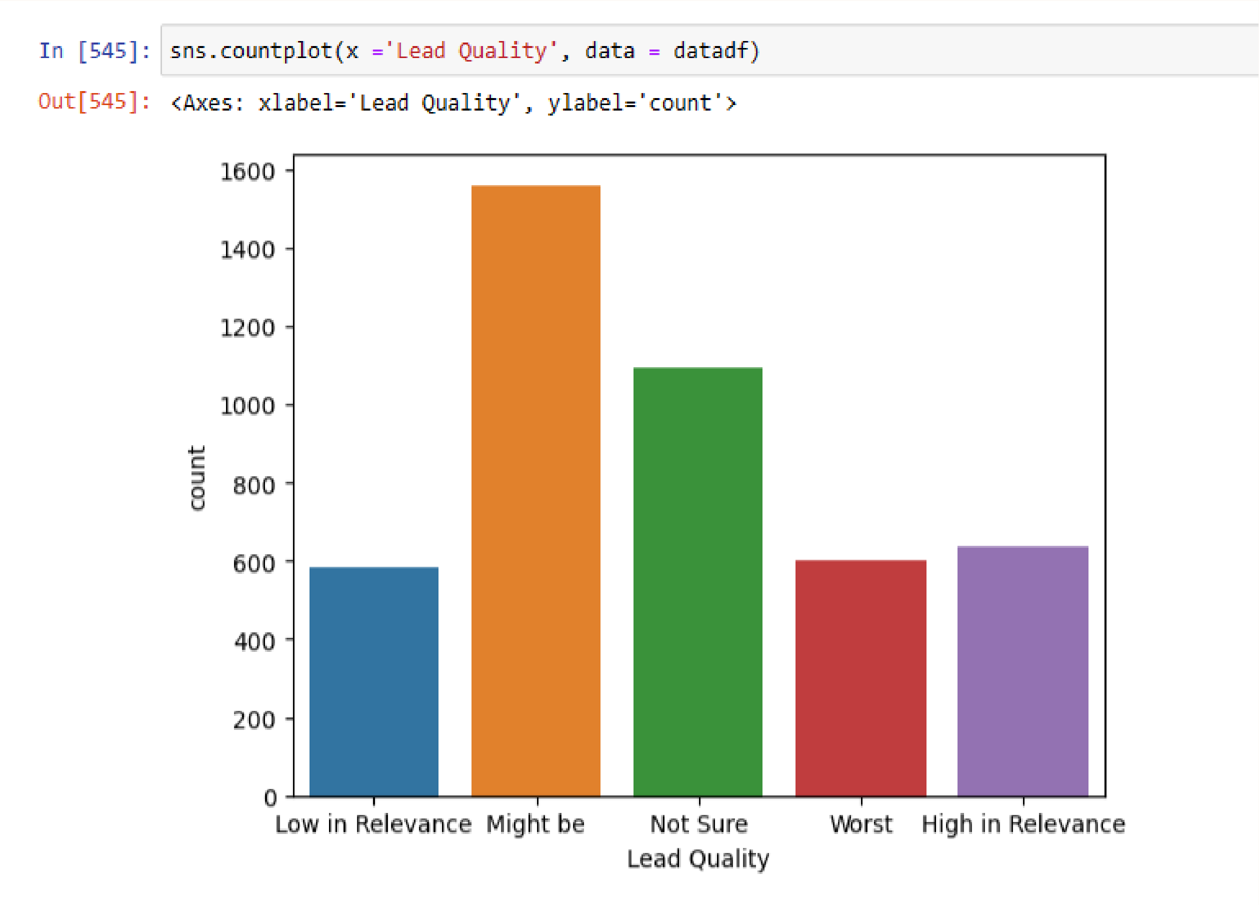


```
Asymmetrique Profile Score

In [542]: sns.countplot(x ='Asymmetrique Activity Score', data = datadf)

Out[542]: <Axes: xlabel='Asymmetrique Activity Score', ylabel='count'>
```

EDA

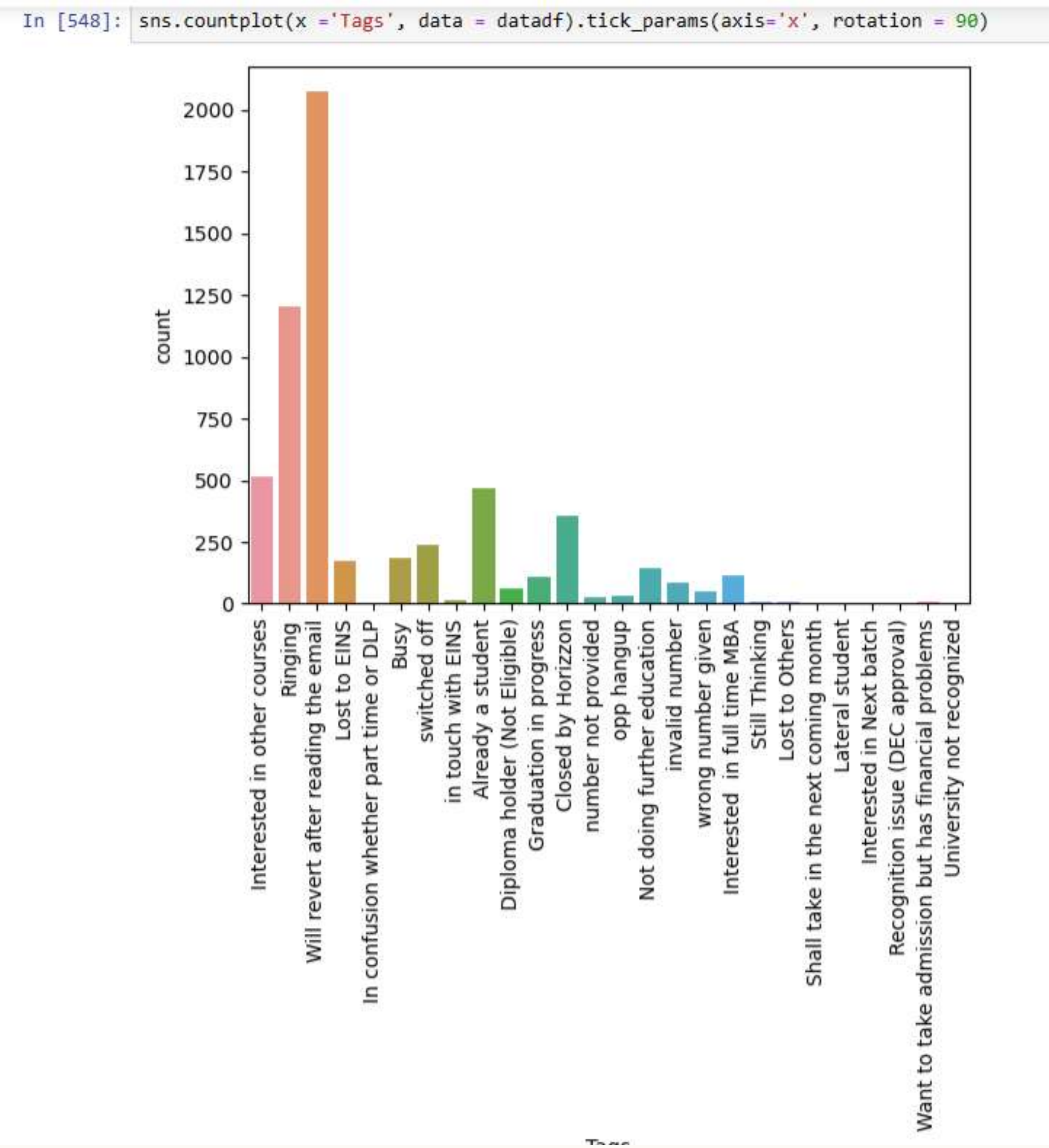


```
In [543]: sns.countplot(x ='Asymmetrique Profile Index', data = datadf)
Out[543]: <Axes: xlabel='Asymmetrique Profile Index', ylabel='count'>
```



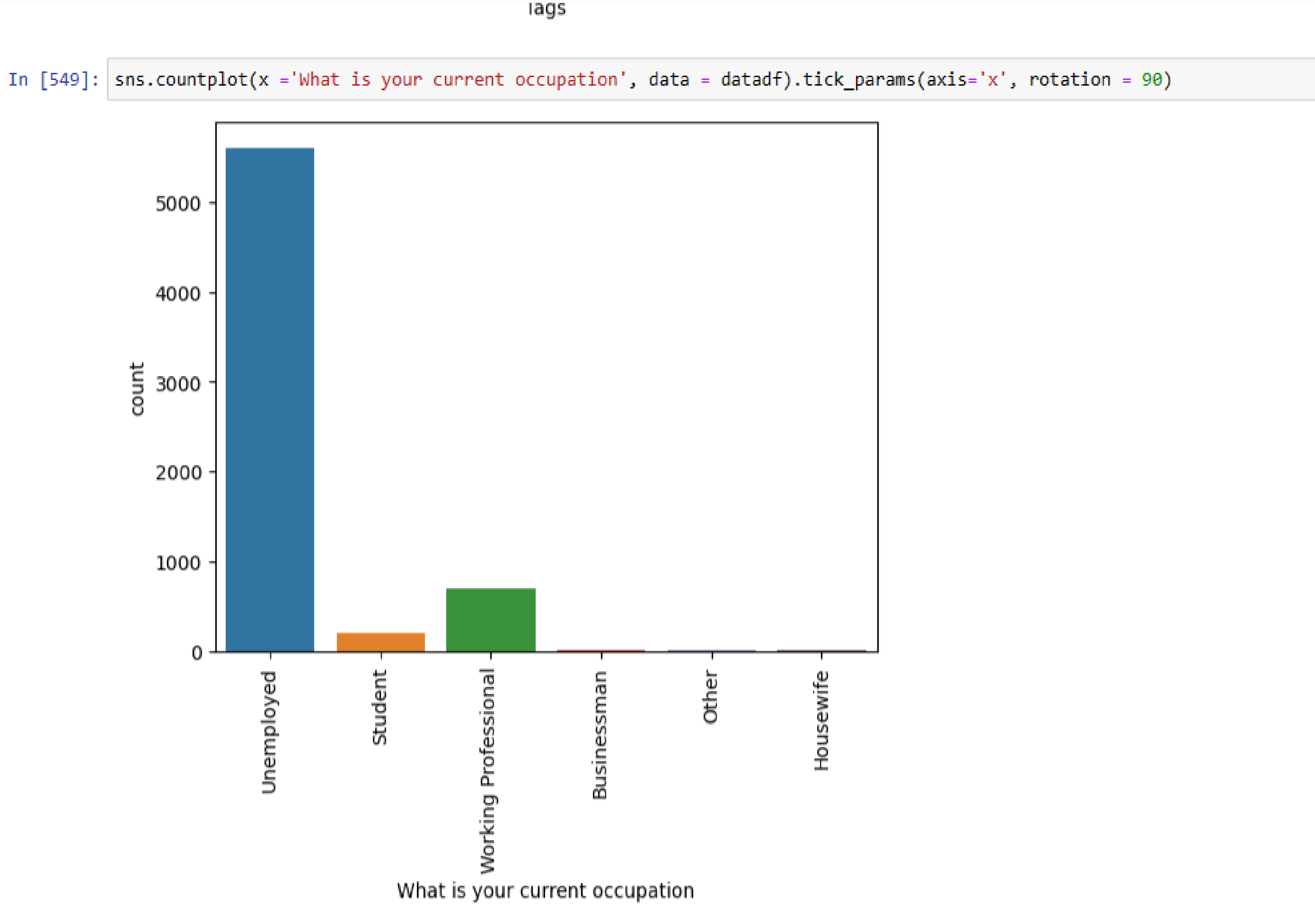
```
In [545]: sns.countplot(x ='Lead Quality', data = datadf)
Out[545]: <Axes: xlabel='Lead Quality', ylabel='count'>
```

EDA



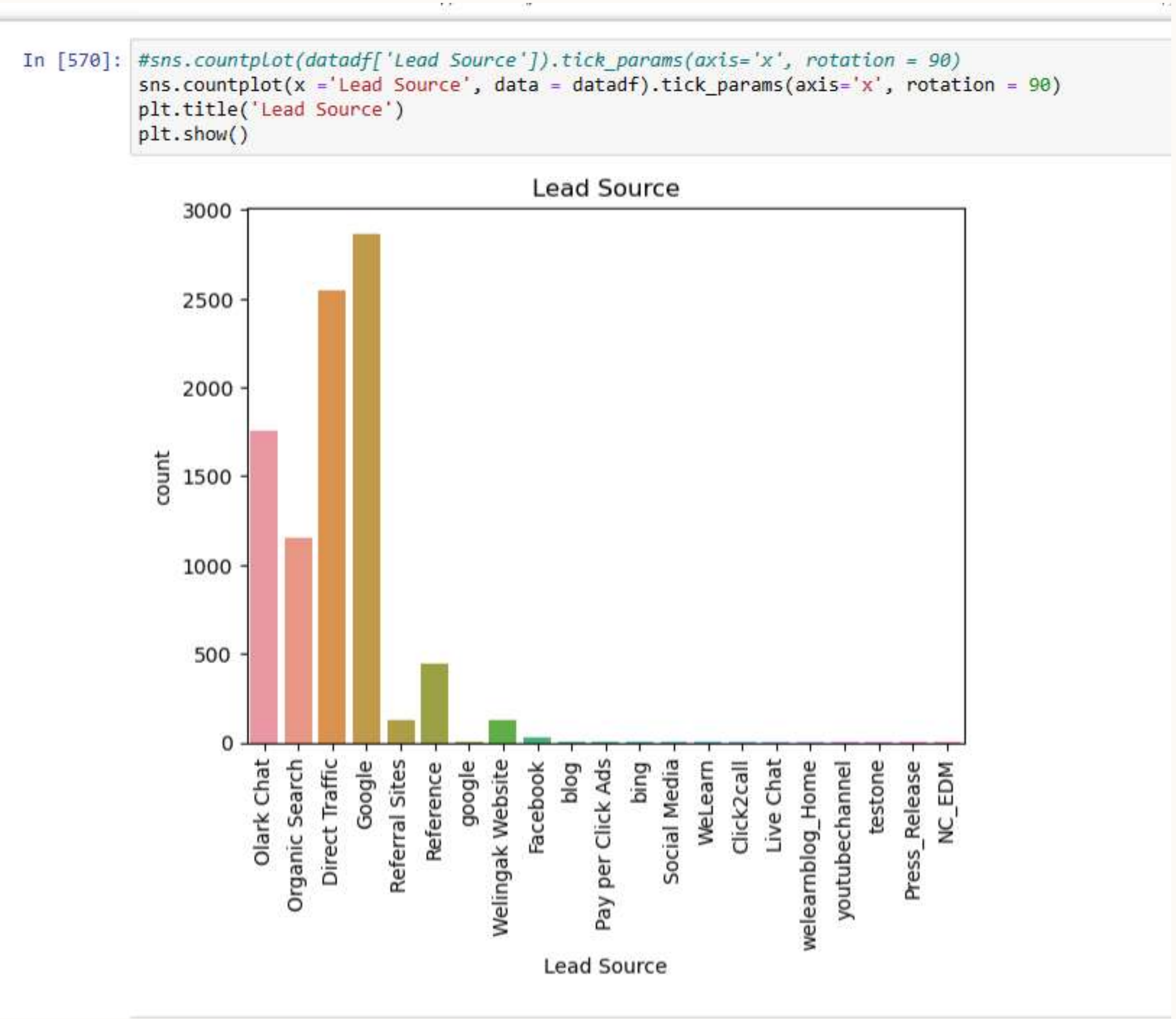
```
In [548]: sns.countplot(x ='Tags', data = datadf).tick_params(axis='x', rotation = 90)
```

count
2000
1750
1500
1250
1000
750
500
250
0
Interested in other courses
Ringing
Will revert after reading the email
Lost to EINS
In confusion whether part time or DLP
Busy
switched off
in touch with EINS
Already a student
Diploma holder (Not Eligible)
Graduation in progress
Closed by Horizzon
number not provided
opp hangup
Not doing further education
invalid number
wrong number given
Interested in full time MBA
Still Thinking
Lost to Others
Shall take in the next coming month
Lateral student
Interested in Next batch
Recognition issue (DEC approval)
Want to take admission but has financial problems
University not recognized
Tags

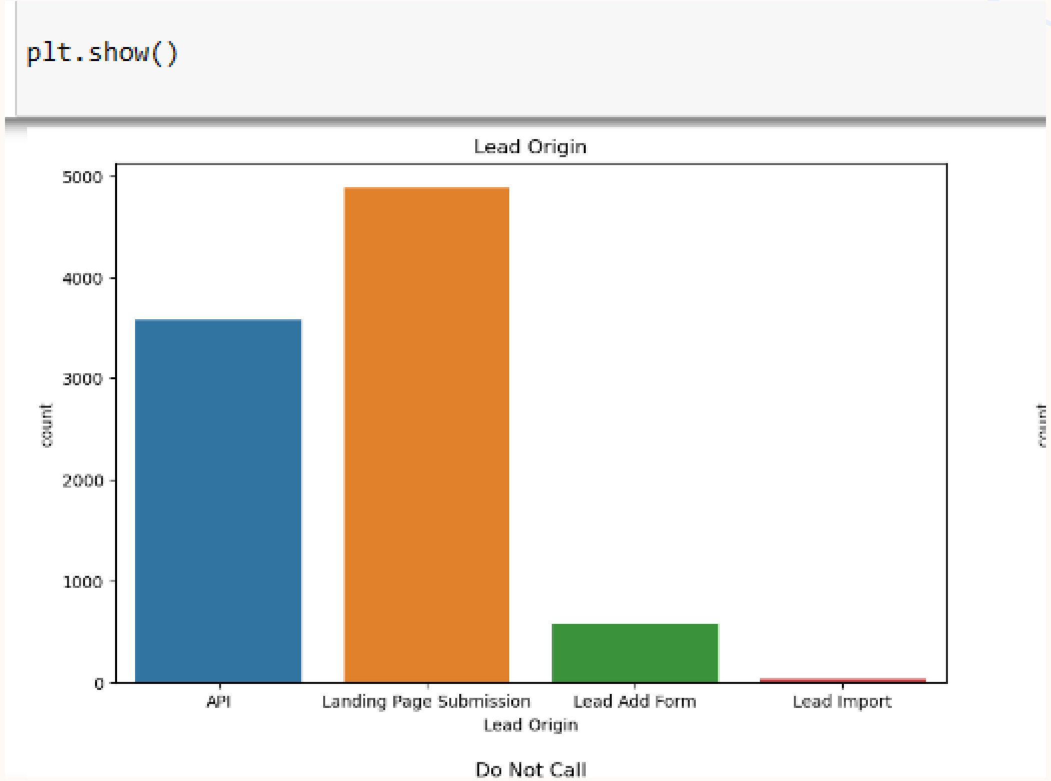


```
In [549]: sns.countplot(x ='What is your current occupation', data = datadf).tick_params(axis='x', rotation = 90)
```

CATEGORICAL VARIABLE RELATION

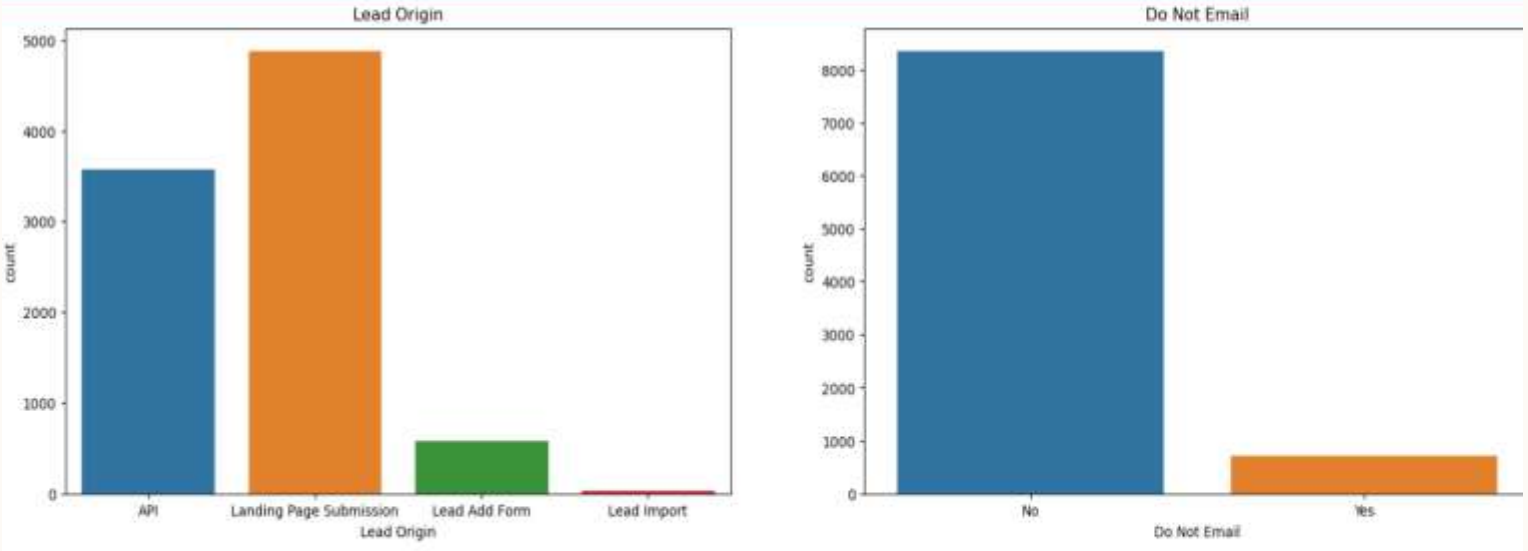


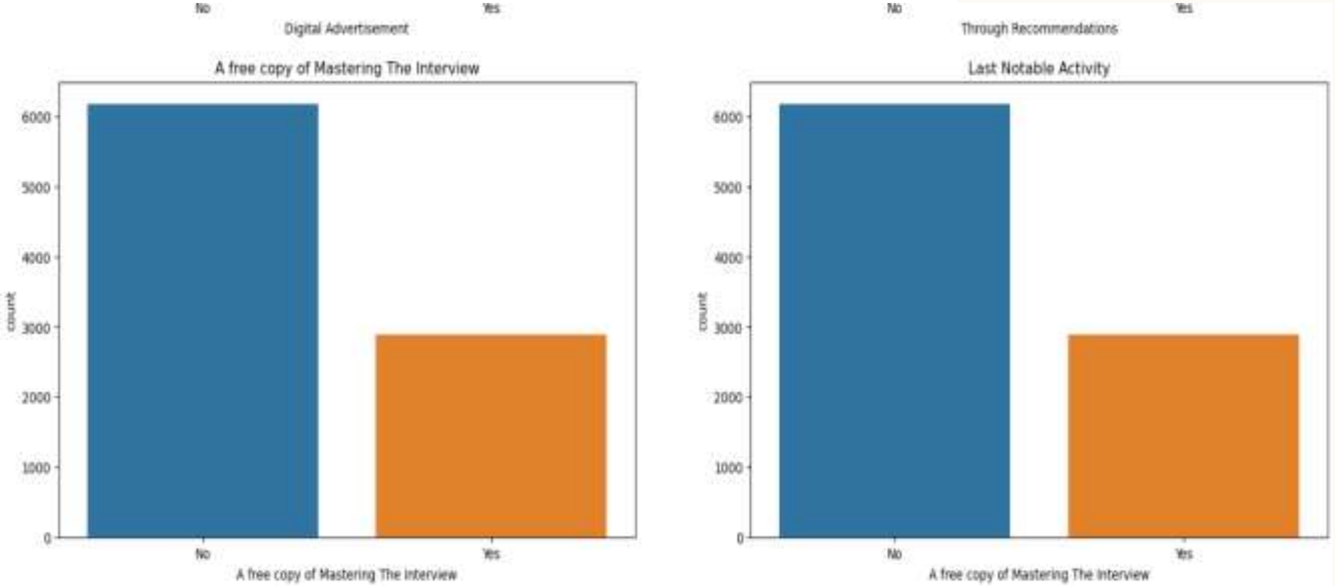
```
In [570]: #sns.countplot(datadf['Lead Source']).tick_params(axis='x', rotation = 90)
          sns.countplot(x ='Lead Source', data = datadf).tick_params(axis='x', rotation = 90)
          plt.title('Lead Source')
          plt.show()
```



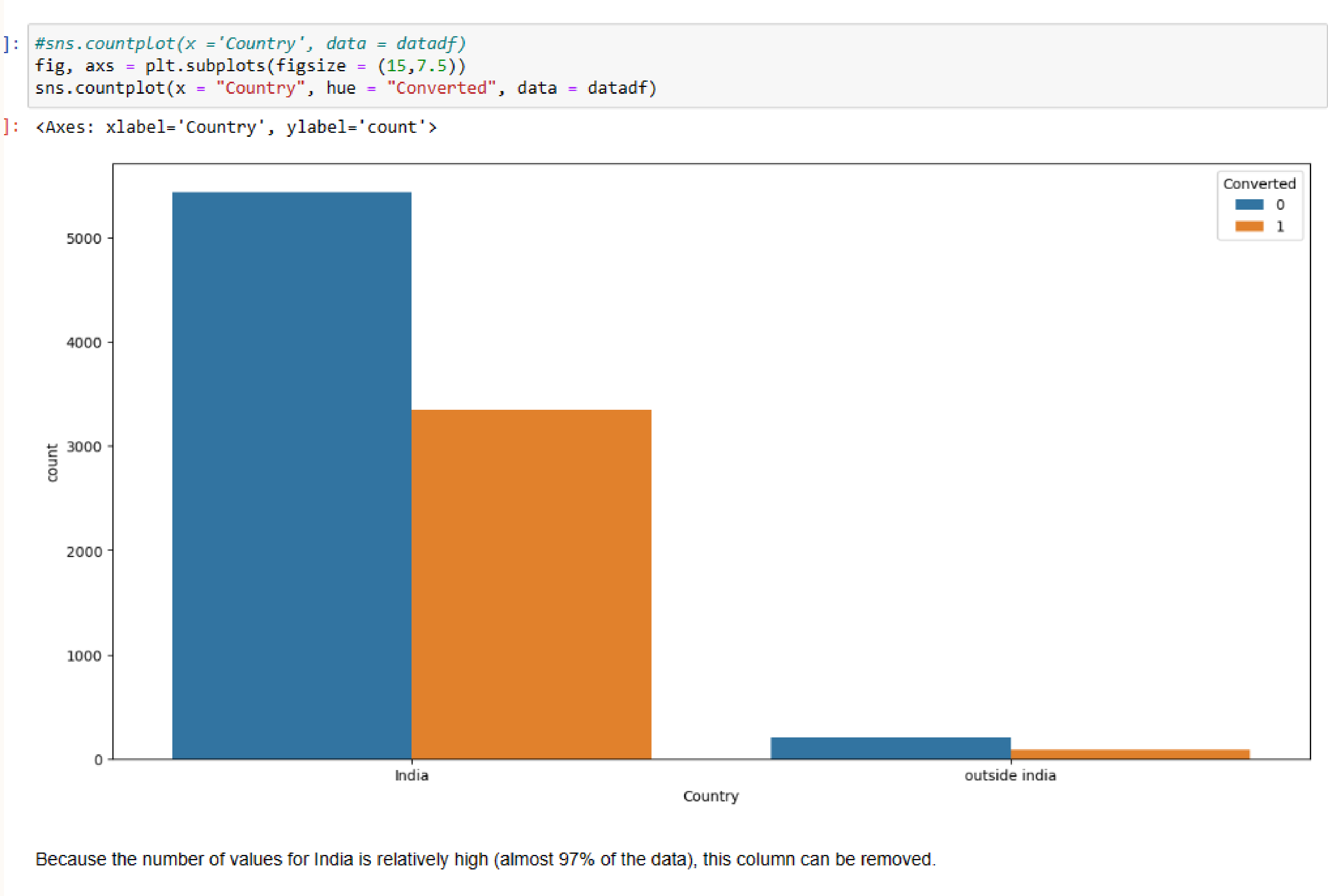
```
plt.show()
```

CATEGORICAL VARIABLE RELATION





CATEGORICAL VARIABLE RELATION



```
]: #sns.countplot(x ='Country', data = datadf)
   fig, axs = plt.subplots(figsize = (15,7.5))
   sns.countplot(x = "Country", hue = "Converted", data = datadf)
```

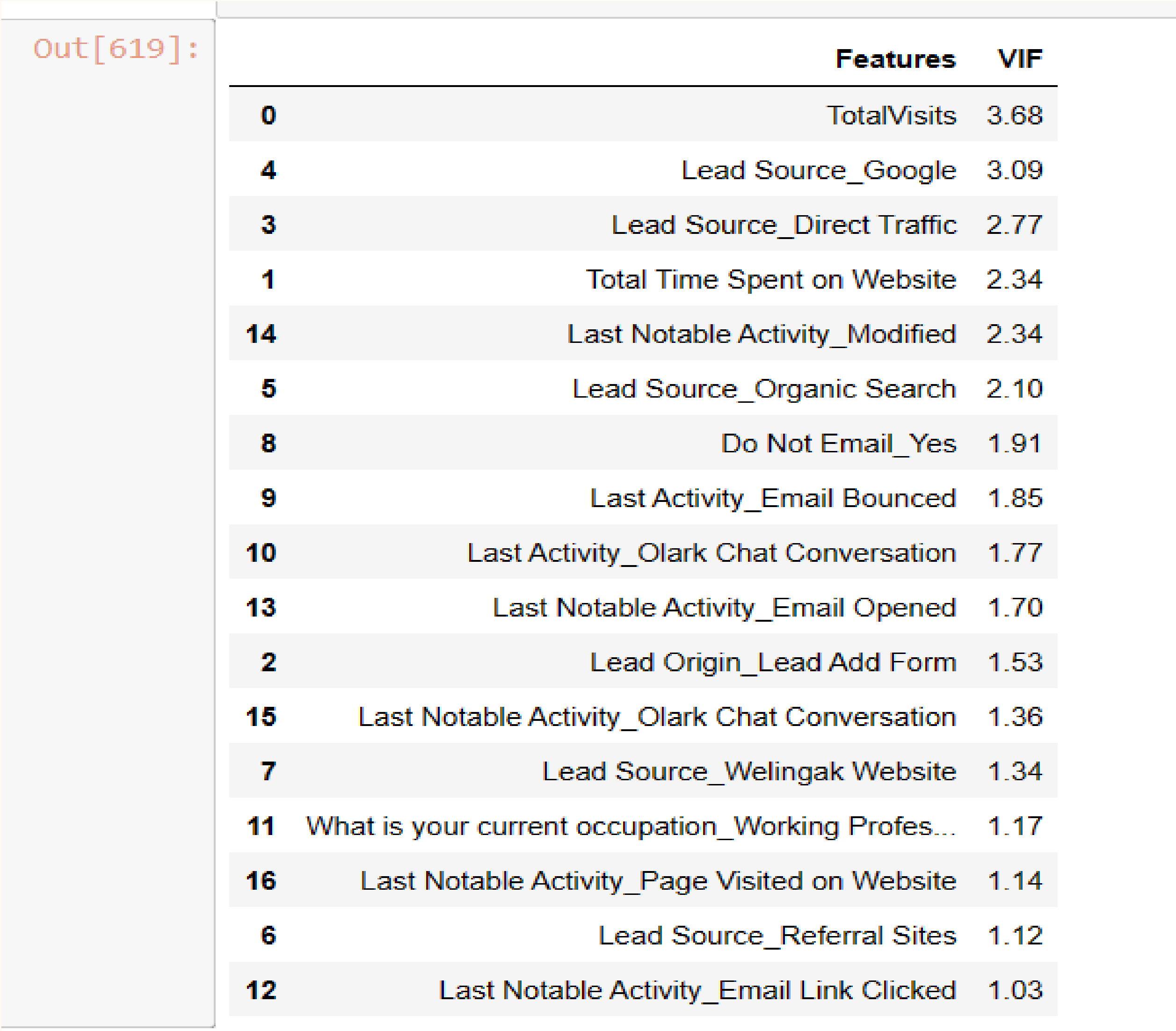
```
]: <Axes: xlabel='Country', ylabel='count'>
```

Because the number of values for India is relatively high (almost 97% of the data), this column can be removed.

MODEL BUILDING

- *Splitting the Data into Training and Testing Sets.*
- *The first basic step for regression is performing a train-test split, we have chosen a 70:30 ratio.*
- *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 whose VIF value is greater than 5.*
- *Predictions on the test data set.*
- *Overall accuracy 82%*

TOP FACTORS THAT IMPACTS THE CONVERSION OF LEADS



Out[619]:

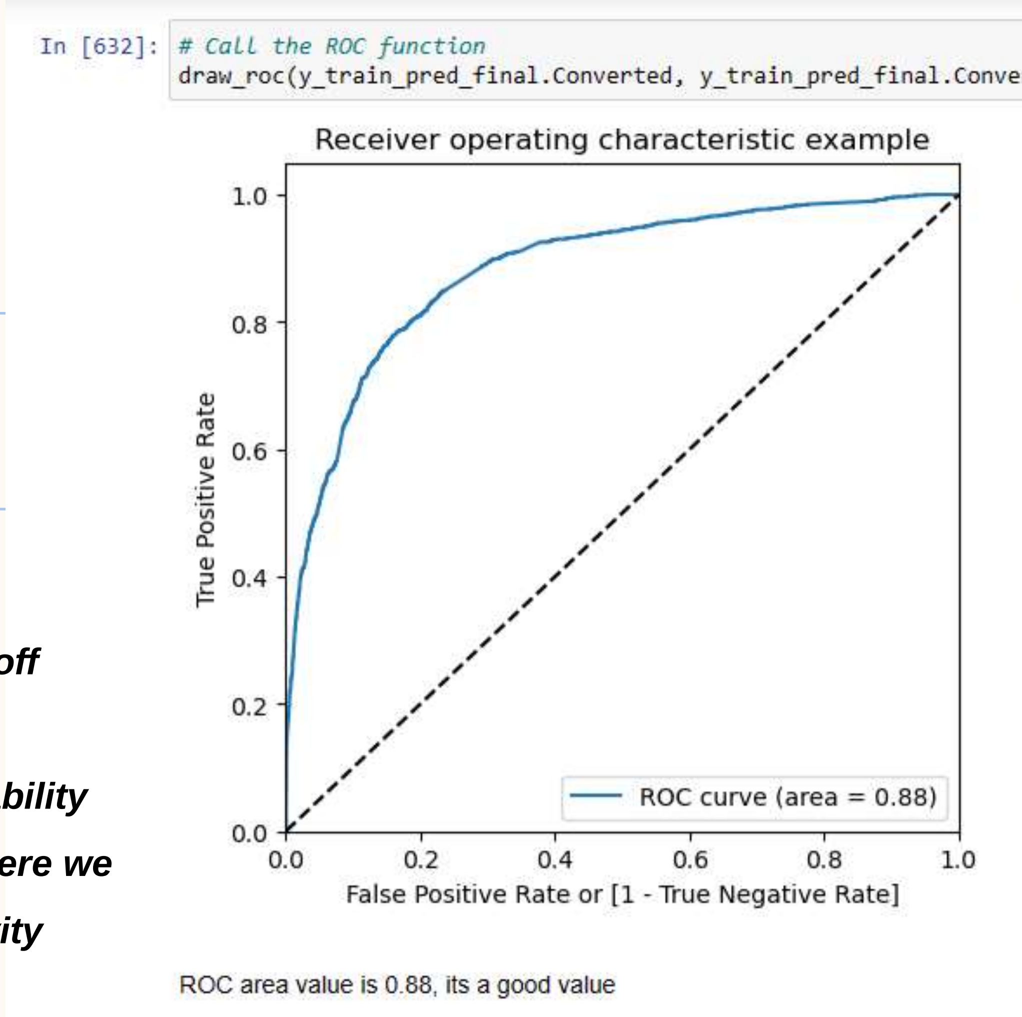
	Features	VIF
0	TotalVisits	3.68
4	Lead Source_Google	3.09
3	Lead Source_Direct Traffic	2.77
1	Total Time Spent on Website	2.34
14	Last Notable Activity_Modified	2.34
5	Lead Source_Organic Search	2.10
8	Do Not Email_Yes	1.91
9	Last Activity_Email Bounced	1.85
10	Last Activity_Olark Chat Conversation	1.77
13	Last Notable Activity_Email Opened	1.70
2	Lead Origin_Lead Add Form	1.53
15	Last Notable Activity_Olark Chat Conversation	1.36
7	Lead Source_Welingak Website	1.34
11	What is your current occupation_Working Profes...	1.17
16	Last Notable Activity_Page Visited on Website	1.14
6	Lead Source_Referral Sites	1.12
12	Last Notable Activity_Email Link Clicked	1.03

P-Values and VIF values are good with this model

HOW WE GET THERE

ROC CURVE

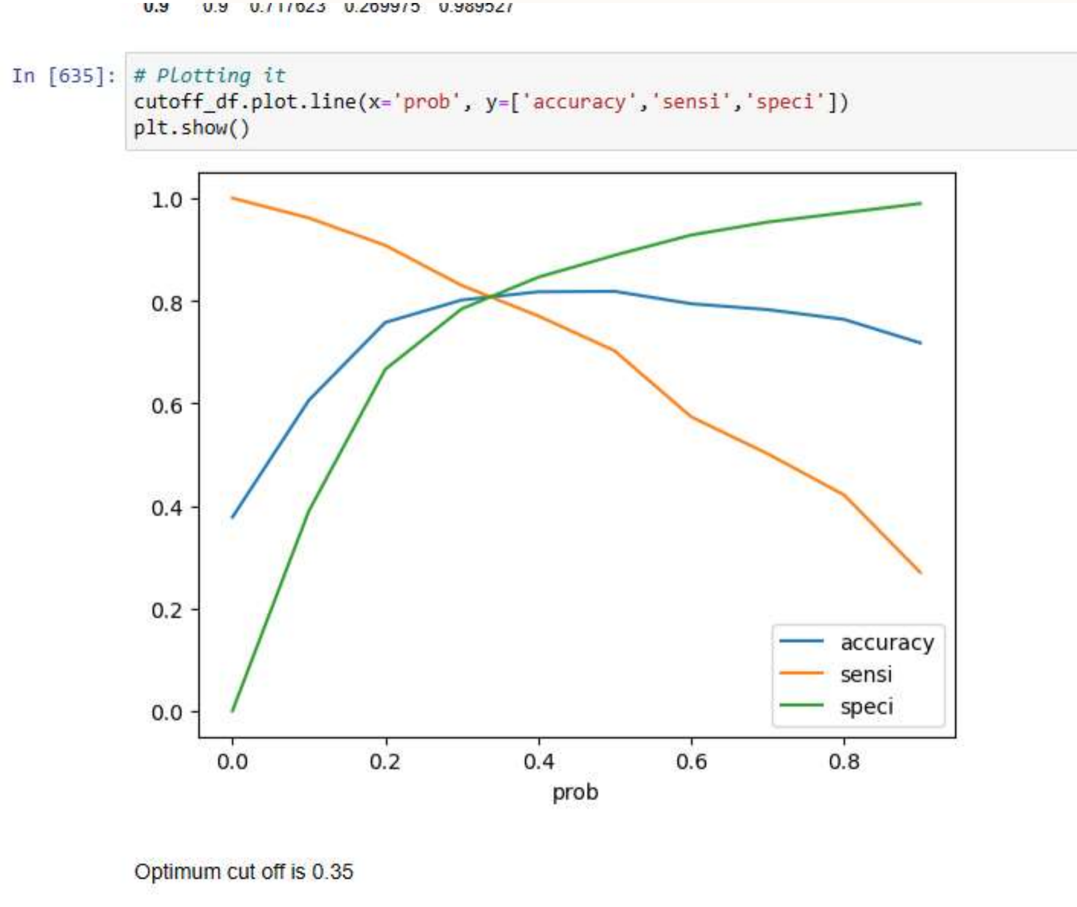
- ❑ *Finding Optimal Cut-off Point*
- ❑ *Optimal cut-off probability is that probability where we get balanced sensitivity and specificity.*
- ❑ *From the second graph it is visible that the optimal cut-off is approximately at*



```
In [632]: # Call the ROC function
          draw_roc(y_train_pred_final.Converted, y_train_pred_final.Conversion_Prob)
```

Receiver operating characteristic example
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.88)

ROC area value is 0.88, its a good value



```
0.9   0.9  0.717623  0.269975  0.989527
```

```
In [635]: # Plotting it
          cutoff_df.plot.line(x='prob', y=['accuracy','sensi','speci'])
          plt.show()
```

Optimum cut off is 0.35

SUMMARY

Conclusion:

The criteria that mattered the most to potential purchasers were discovered to be (in descending order):

1. TotalVisits
2. The total time spent on the Website.
3. Lead Source_Direct Traffic
4. Lead Source_Google
5. Lead Source_Welingak Website #Lead Source_Organic Search
6. Lead Source_Referral Sites Website of
7. Lead Source_Welingak #Do Not Email_Yes #Last Activity_Email Bounced

With these analyses, X Education may thrive since they have a very good possibility of convincing practically all potential buyers to alter their minds and purchase their courses.

THANK YOU