International
Institute of Information
Technology Bangalore

upGrad

Lead Scoring Case Study

|| BUILDING A MACHINE LEARNING MODEL USING LOGISTIC REGRESSION ||

GAURI BHALE | MADAN HM

Problem Statement

An education company named X Education sells online courses to industry professionals.Through marketing and the leads procured by various resources, the conversion rate is around 30%, which is considerably poor.

With this case study, X Education wants to identify the most significant leads and focus on them in near future in order to increase the customers opting for the online offered courses by them.

This study potentially should increase the conversion rate as well as help sales team to dial up those whose probability is higher as per the hot leads instead of identifying in random.

Goals of the Study:

- ***Identifying the most potential leads or hot leads which contribute more in higher conversion rate***
- ***Build a model with the identified potential leads using logistic regression and evaluate the metrics.***
- ***Evaluate the model and deploy the algorithm in order to assist sales team improvise conversion rate***

STEPS FOLLOWED IN ANALYSIS TO ACHIEVE THE GOALS

(1) Importing the necessary mathematical, visual and statistical support libraries to perform the analysis

- *Numpy*
- *Pandas*
- *Matplotlib*
- *Seaborn*
- *Sklearn*
- *Statsmodel*

(2) Reading and understanding the Dataset

(3) Data Cleaning

- *Handle the "Select" level that is present in many of the categorical variables.*
- *Drop columns that are having high percentage of missing values.*
- *Check the percentage of missing values row wise.*
- *Check the number of unique categories in each categorical column. Here you may need to do something.*
- *Finally check the percentage of rows retained in data cleaning process.*

continued….

STEPS FOLLOWED IN ANALYSIS TO ACHIEVE THE GOALS

(4) Exploratory Data Analysis (EDA)

(5) Creation of Dummy Variables

(6) Test Train Split

(7) Build the model using logistic regression technique

(8) Prediction on train dataset

(9) Evaluation of Model

(10) Optimization of Cut off using ROC Curve

(11) Prediction on test dataset

(12) Precision-Recall

(13) Conclusion

DATA CLEANING STEP BY STEP

- *It is observed in previous stage of understanding the dataset step that, during the data collection some of the information are left by default with* ***'Select'*** *as a value which are actually the null values. Let's start the data cleaning process by* ***replacing all the default selections with null values****.*

- *Next, we identified quite a good amount of null values in dataset with highest being* ***'How did you hear about X Education'*** *with* ***78.46%****, which could have troubled and highly affected the analysis. Hence we have dropped the columns with 35% null values straight away.*

- ***Specialization, What matters most to you in choosing a course*** *and* ***What is your current occupation****, as they are the most important factors to understand the behavior of a student enrolling into the course and hence the null values are filled as* ***unknown***

- *Upon checking the null value percentage for* ***Total Visits****,* ***Page Views Per Visit****,* ***Last Activity*** *and* ***Lead Source****, it is noticed to be less than 2% and hence dropping entire rows in dataset.*

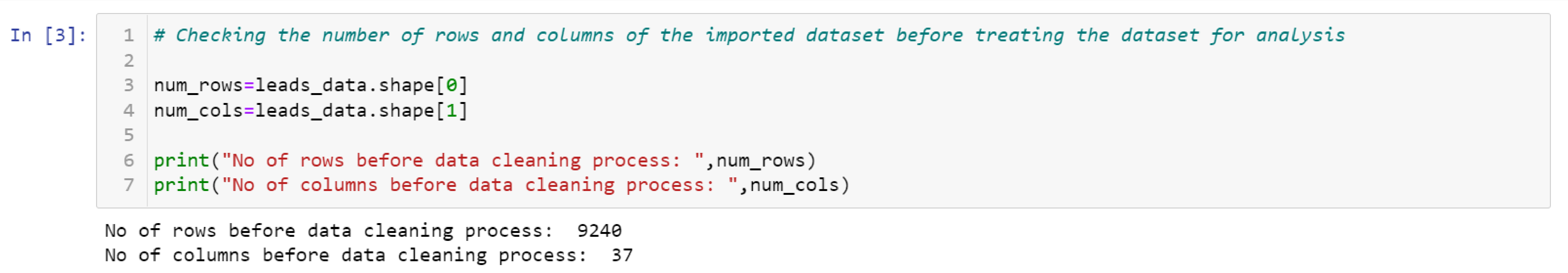
- *we had then reached the stage of data cleaning with no null values in our dataset loaded for analysis. Moving forward, we checked the unique value counts for each of the categorical variables.*

- *Upon results with variables having unique value count one such as* ***Magazine, I agree to pay the amount through cheque, Get updates on DM Content, Update me on Supply Chain Content, Receive More Updates About Our Courses*** *are dropped from dataset including Country as 70% leads are from India only.*

- Last but not least, we have dropped **Prospect ID and Lead Number** as all the values in them are unique and doesn't contribute anything for analysis

SHAPE OF DATASET

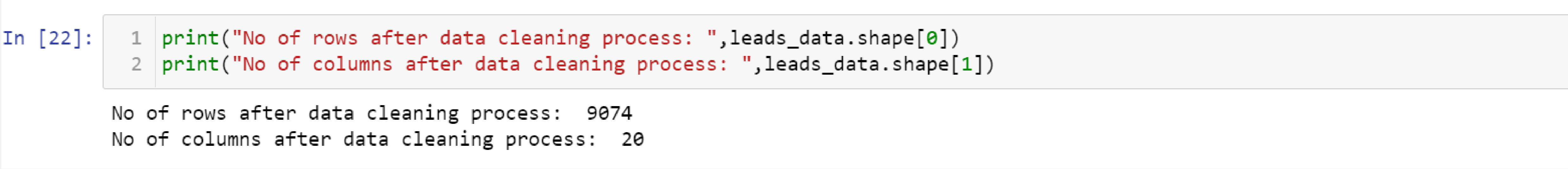
Before data cleaning:



```
In [3]:

# Checking the number of rows and columns of the imported dataset before treating the dataset for analysis

num_rows=leads_data.shape[0]
num_cols=leads_data.shape[1]

print("No of rows before data cleaning process: ",num_rows)
print("No of columns before data cleaning process: ",num_cols)
```

```
No of rows before data cleaning process:  9240
No of columns before data cleaning process:  37
```

After data cleaning:

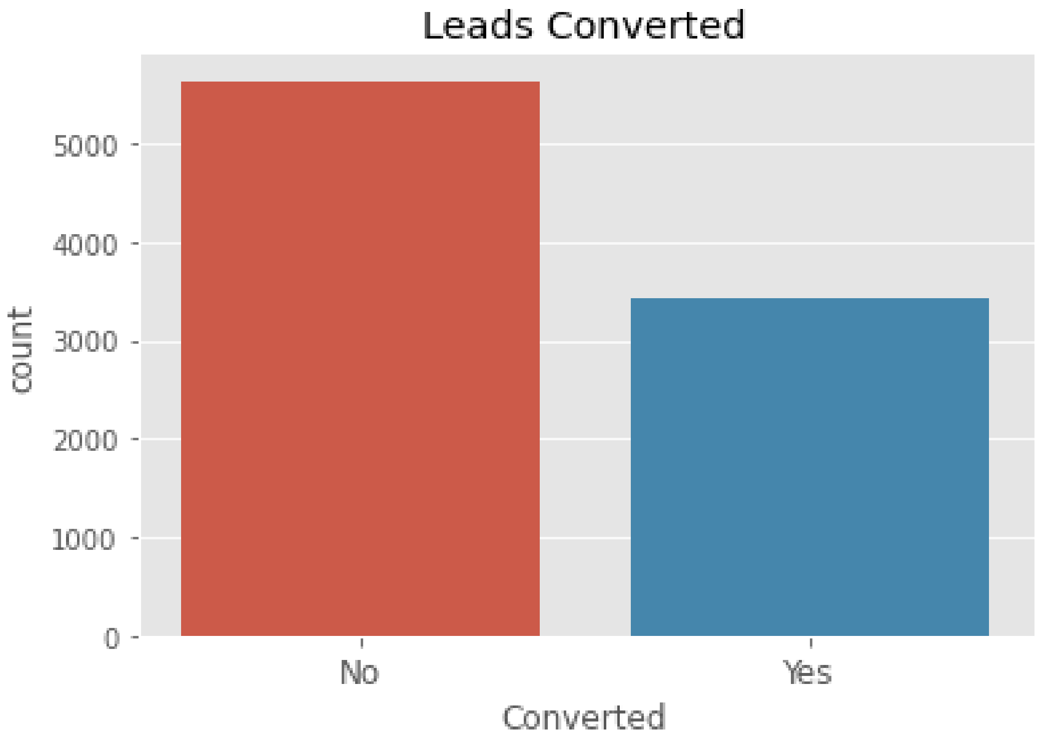


```
In [22]:

print("No of rows after data cleaning process: ",leads_data.shape[0])
print("No of columns after data cleaning process: ",leads_data.shape[1])
```

```
No of rows after data cleaning process:  9074
No of columns after data cleaning process:  20
```

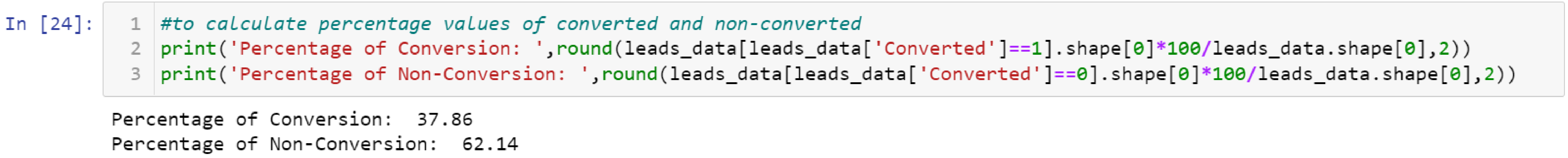
Exploratory Data Analysis

Data balance in Target variable



Inference

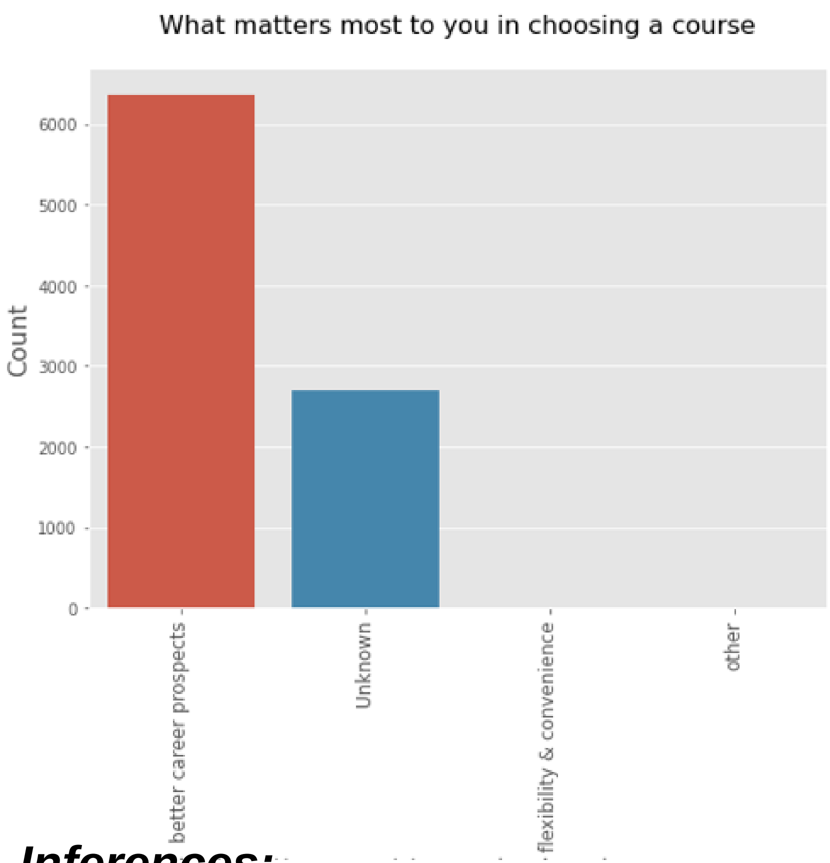
The dataset is not highly imbalanced and hence meaningful insights can be drawn from it with respect to target variable.

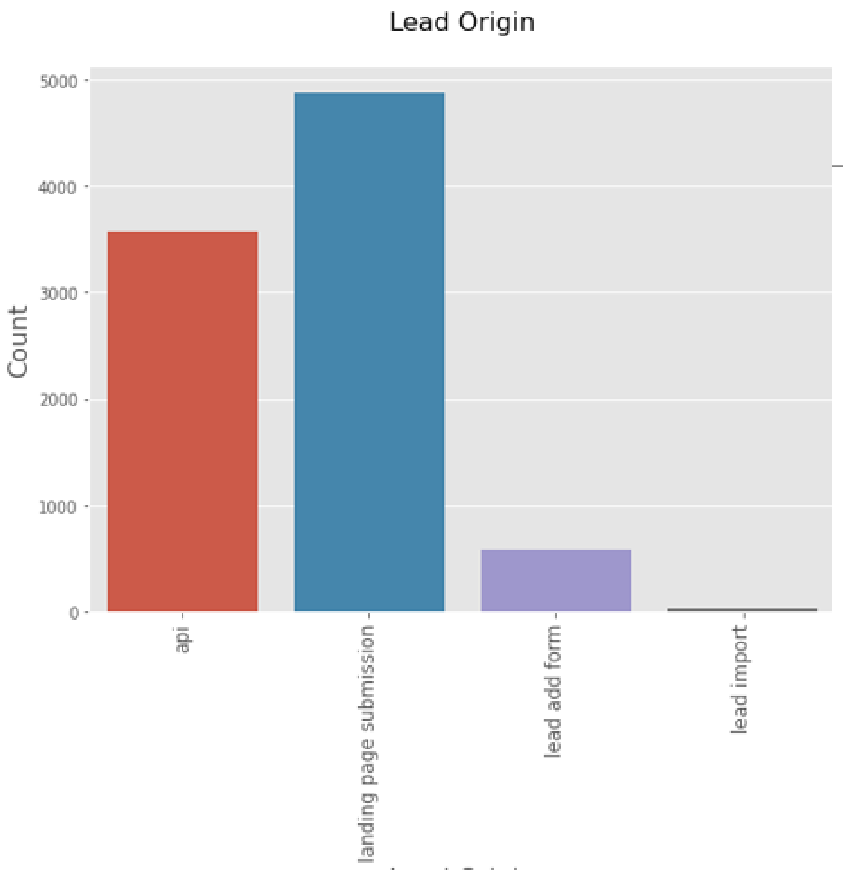


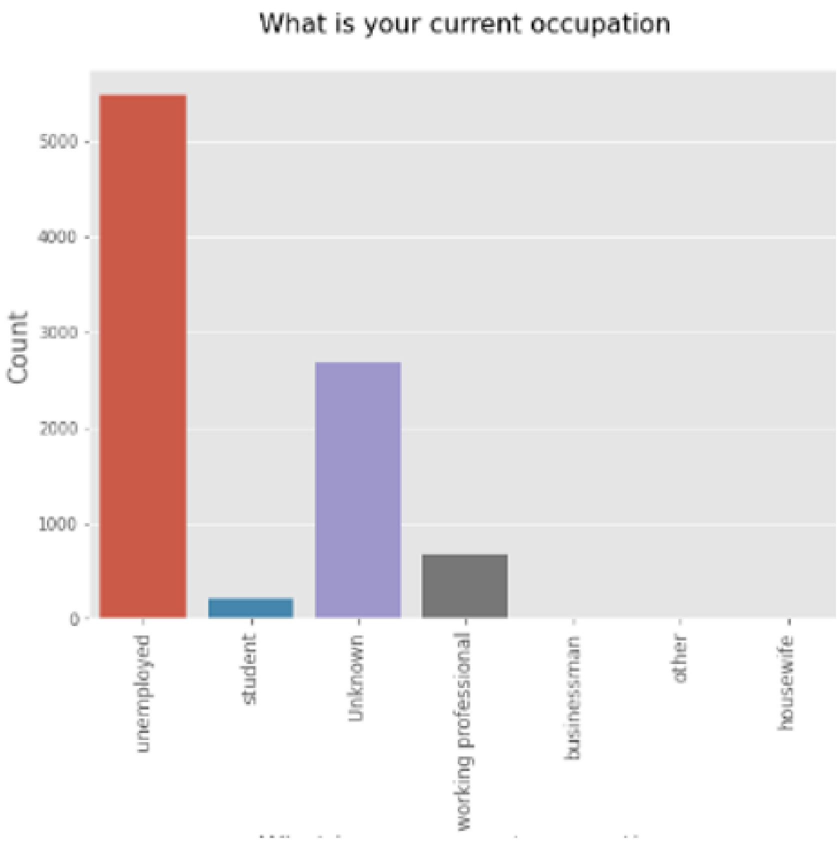
```
In [24]: #to calculate percentage values of converted and non-converted
         print('Percentage of Conversion: ',round(leads_data[leads_data['Converted']==1].shape[0]*100/leads_data.shape[0],2))
         print('Percentage of Non-Conversion: ',round(leads_data[leads_data['Converted']==0].shape[0]*100/leads_data.shape[0],2))
```

```
Percentage of Conversion:  37.86
Percentage of Non-Conversion:  62.14
```

Exploratory Data Analysis


What matters most to you in choosing a course

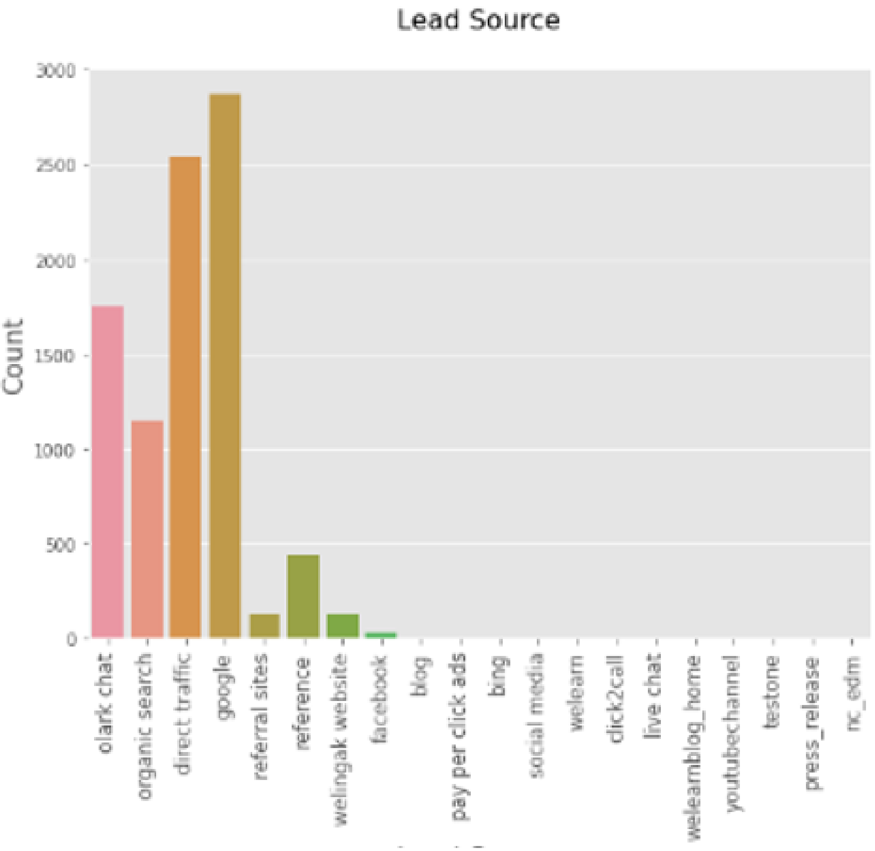

Lead Origin

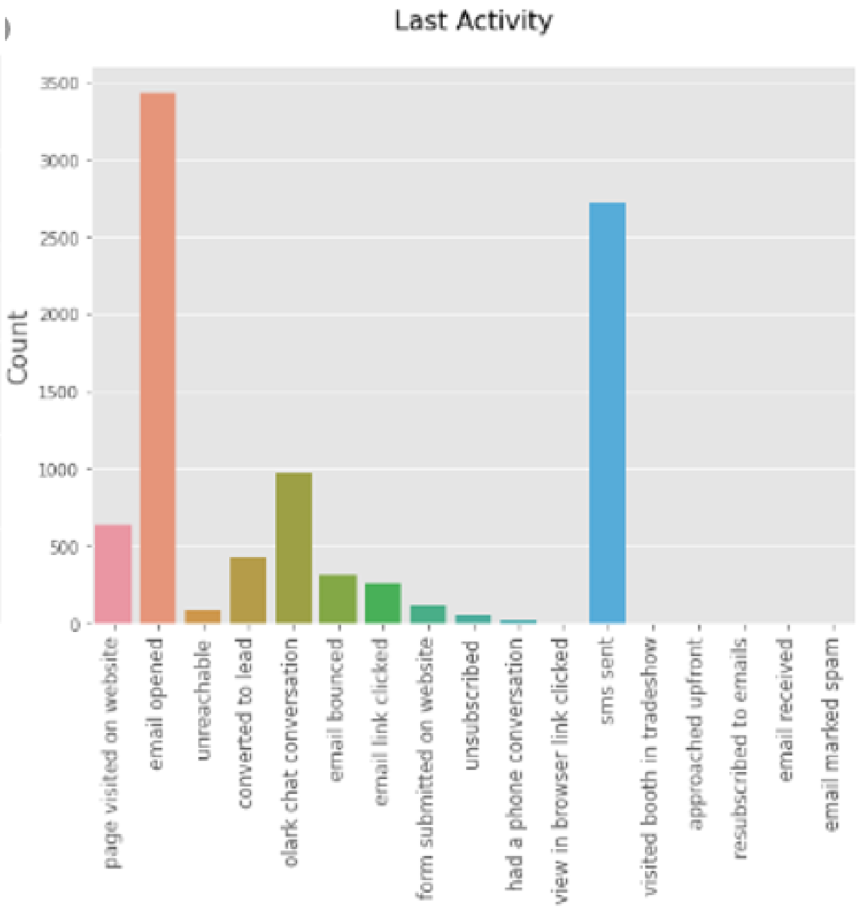

What is your current occupation

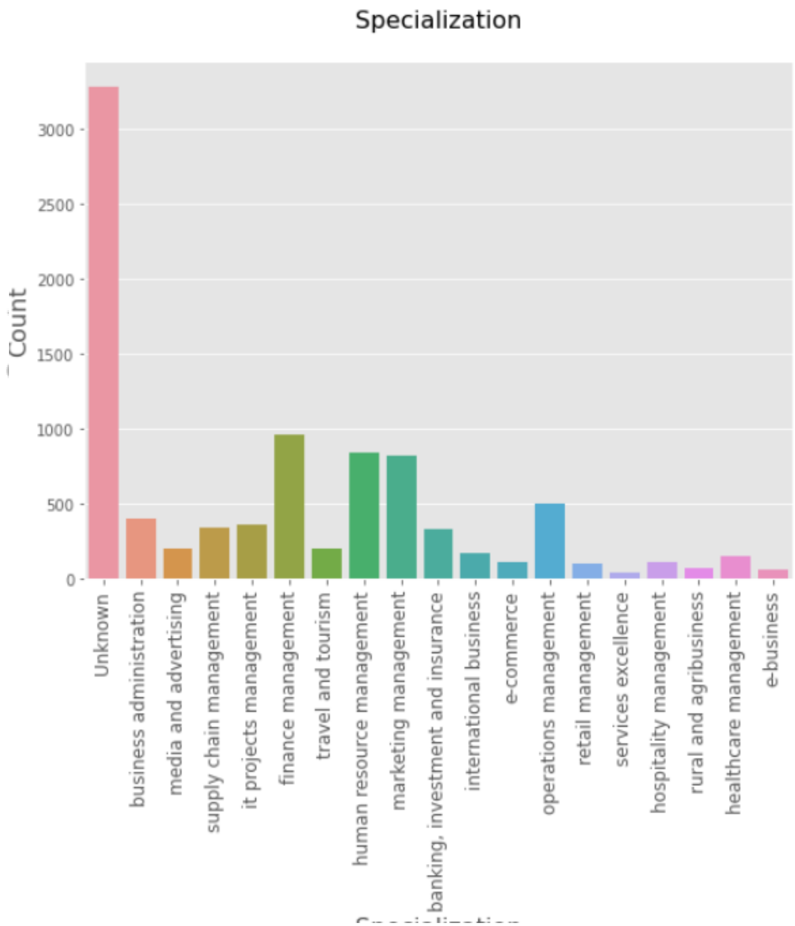
<u>Inferences:</u>

*The major motto which has been mattered to most of the customers in choosing the course is **better career prospects**.*
***Landing page submission** is a major lead origin to identify the customer to be a lead.*
***Working professionals** and **unemployed** are highest amongst the population in occupation variable.*

Exploratory Data Analysis

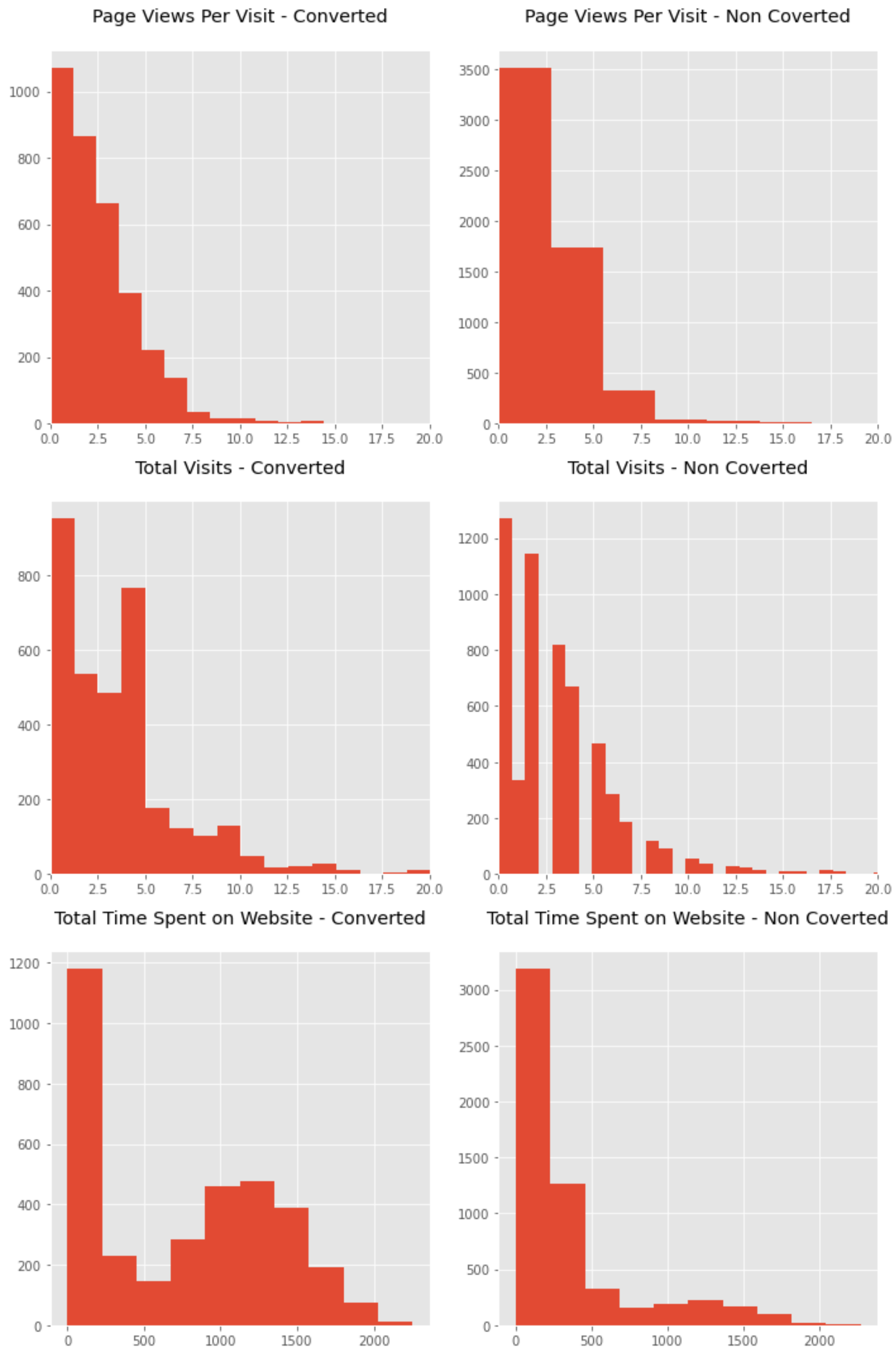






<u>Inferences:</u>

Google, organic search *along with* ***direct traffic*** *has been a main source of lead for customers.*
Most of the customers have enrolled to course after ***SMS sent*** *or* ***Email being read*** *by them as per the latest activity.*
Finance Management *is the highest selected specialization category whereas* ***HR and Marketing*** *are amongst top 3.*



Inferences:

1. *It is recommended to make website more interactive in order to ensure customers spend more time to convert. It is observed that leads spending more time on website are at higher chances of conversion.*

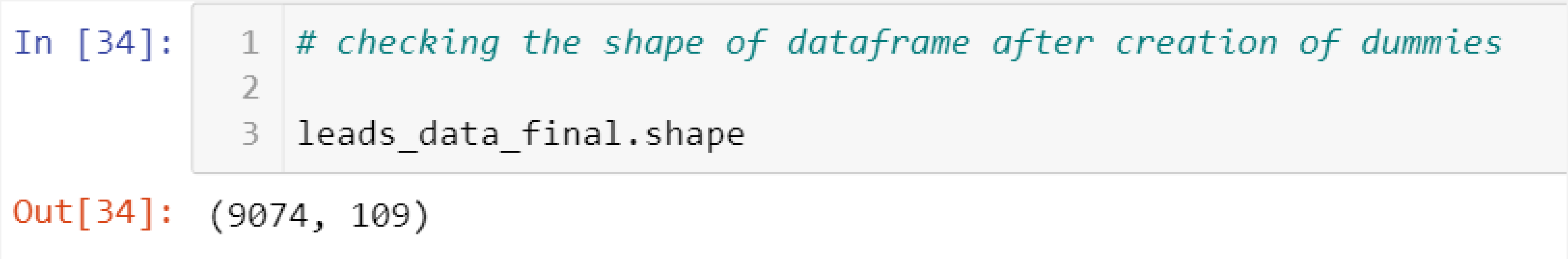
2. *Advertisements to be increased for visitors in order to ensure they engage in website more time, more likely to convert as average visit time for both converted and non-converted are same.*

3. *Average number of pages on the website viewed during the visits is very less for the converted as well as non-converted, it is advised to minimize the pages and increase effective content on home page.*

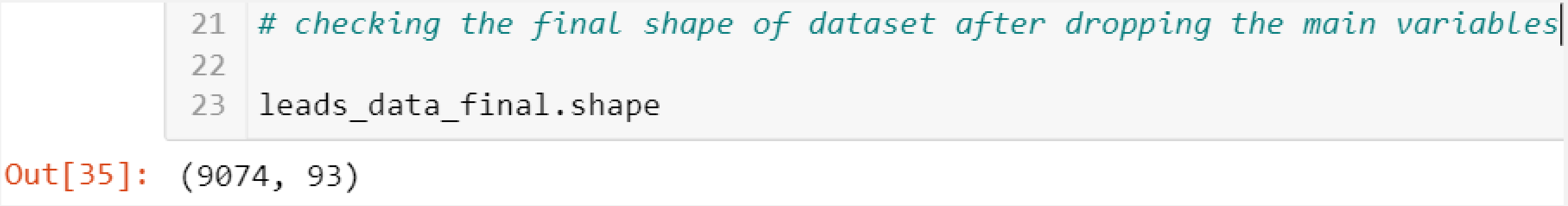
CREATION OF DUMMY VARIABLES

List of Categorical variables considered for creation of dummy variables are in the list below.

['Newspaper’, 'Do Not Call’, 'What matters most to you in choosing a course’, 'Do Not Email’, 'A free copy of Mastering The Interview’, 'X Education Forums’, 'Newspaper Article’, 'Through Recommendations’, 'Last Activity’, 'Lead Origin’, 'Lead Source’, 'What is your current occupation’, 'Search’, 'Digital Advertisement’, 'Last Notable Activity’, 'Specialization']



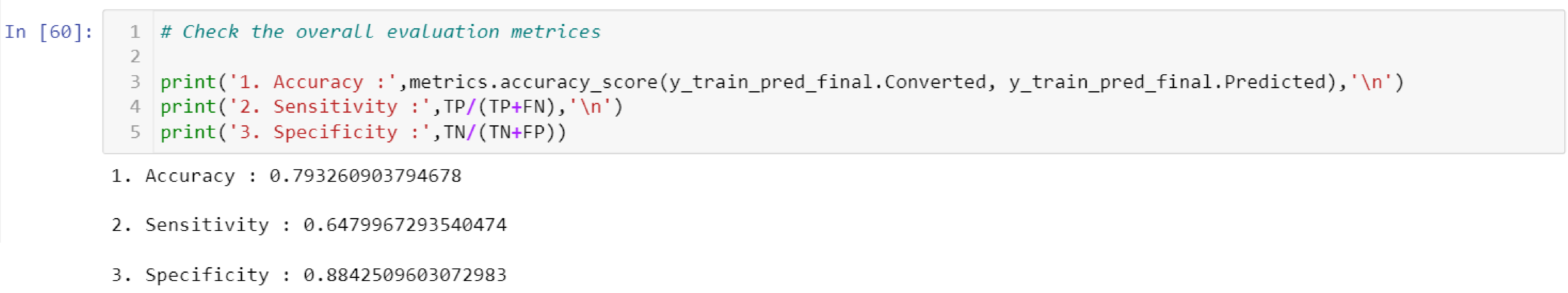
```
In [34]:  # checking the shape of dataframe after creation of dummies

           leads_data_final.shape
Out[34]: (9074, 109)
```



```
           # checking the final shape of dataset after dropping the main variables

           leads_data_final.shape
Out[35]: (9074, 93)
```

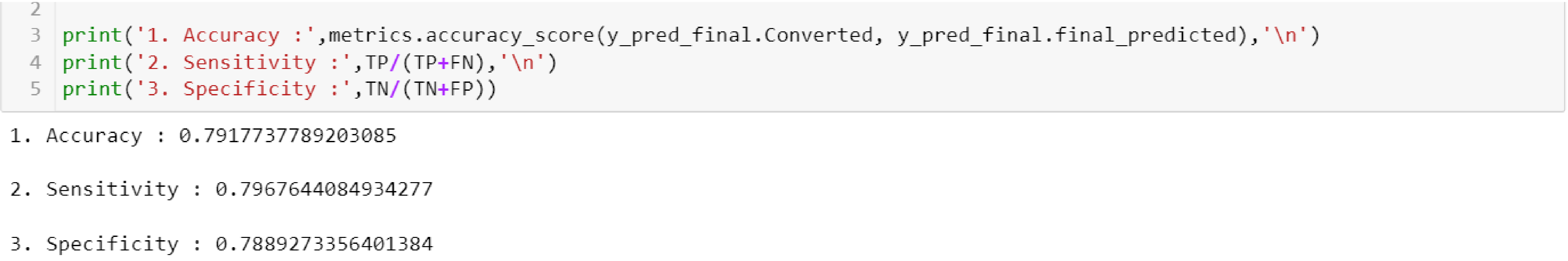
MODEL BUILDING STEP BY STEP

- *Splitting the dataset into **70% train data and 30% test data** for building the model using **train_test_split** function*
- *Scaling the numerical variables using the **MinMaxScaler** function*
- *Using **RFE method of feature elimination** and selection of variables. We have considered to reduce it to 15.*
- *Building model by adjusting or removing the **p-value and VIF** by manual elimination method.*
- ***Predictions** on Test dataset and finding out the **values of evaluation metrices**.*



```
In [60]:
# Check the overall evaluation metrices

print('1. Accuracy :',metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.Predicted),'\n')
print('2. Sensitivity :',TP/(TP+FN),'\n')
print('3. Specificity :',TN/(TN+FP))
```

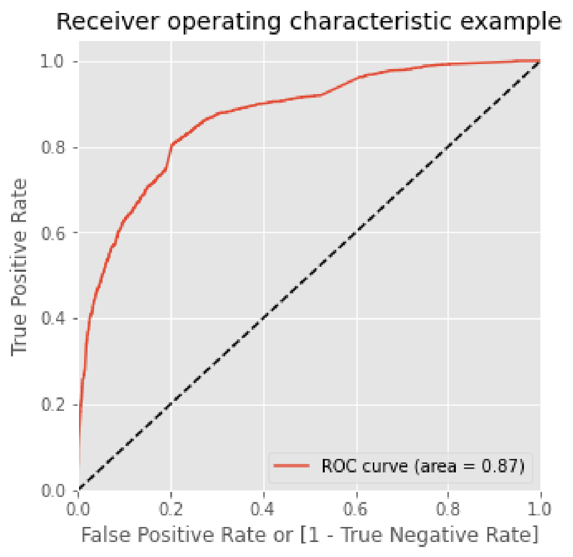
```
1. Accuracy : 0.793260903794678

2. Sensitivity : 0.6479967293540474

3. Specificity : 0.8842509603072983
```

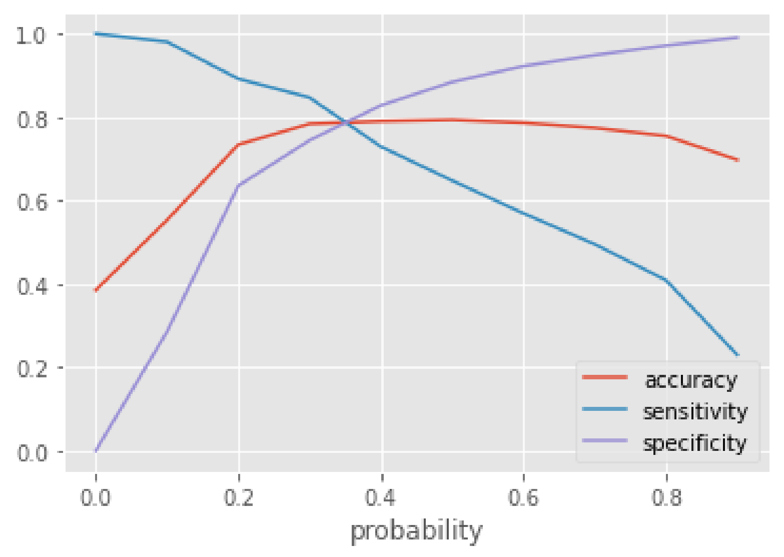


```

print('1. Accuracy :',metrics.accuracy_score(y_pred_final.Converted, y_pred_final.final_predicted),'\n')
print('2. Sensitivity :',TP/(TP+FN),'\n')
print('3. Specificity :',TN/(TN+FP))
```

```
1. Accuracy : 0.7917737789203085

2. Sensitivity : 0.7967644084934277

3. Specificity : 0.7889273356401384
```

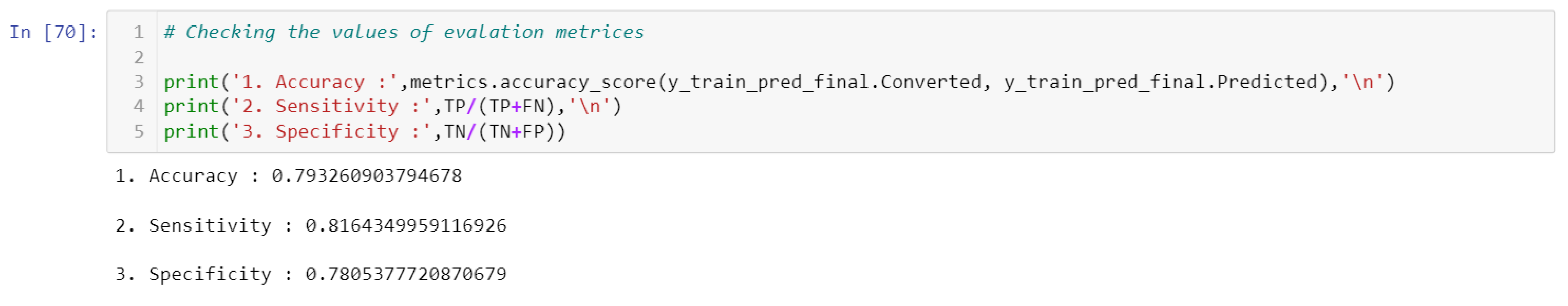
OPTIMIZATION THE CUT OFF WITH ROC CURVE





Optimal Cut off point: 0.35

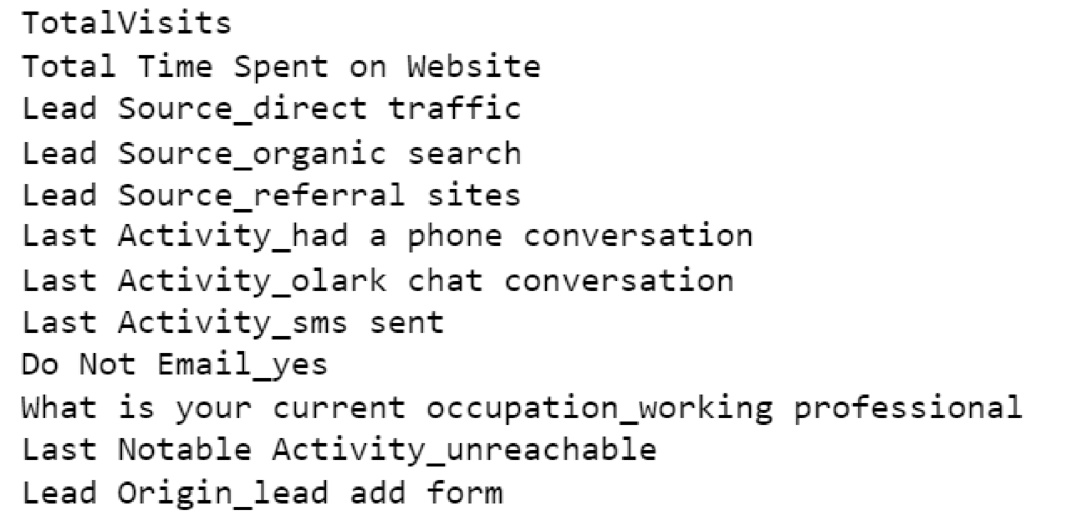
It can be observed that there is an increase in Sensitivity from 64.7% to 81.6% with change in cut off from 0.5 to 0.35 using ROC curve optimization technique



```
In [70]:
# Checking the values of evalation metrices

print('1. Accuracy :',metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.Predicted),'\n')
print('2. Sensitivity :',TP/(TP+FN),'\n')
print('3. Specificity :',TN/(TN+FP))
```

```
1. Accuracy : 0.793260903794678

2. Sensitivity : 0.8164349959116926

3. Specificity : 0.7805377720870679
```

CONCLUSION



```
TotalVisits
Total Time Spent on Website
Lead Source_direct traffic
Lead Source_organic search
Lead Source_referral sites
Last Activity_had a phone conversation
Last Activity_olark chat conversation
Last Activity_sms sent
Do Not Email_yes
What is your current occupation_working professional
Last Notable Activity_unreachable
Lead Origin_lead add form
```

- *It is recommended to focus on the listed features above in order to improvise the conversion rate. Some of the insights are listed below*
- *Total **time spent on website** and total **visits to website** plays a vital role, as more the customer checks the website and spend time, more the chances of enrolling to course.*
- *Source for the leads to look after are **Direct traffic, organic search** and **referral sites**, which has higher probability as per model to enroll the customer.*
- *Next, **Phone conversation, Olark chat conversation, SMS sent** are notable amongst the last activities which can possibly lead to conversion of customer to enroll into course*
- *Last but not the least, Customers who are **working professionals** are the hot leads for enrolling into the course*