# LEAD SCORING CASE STUDY

Group Members:

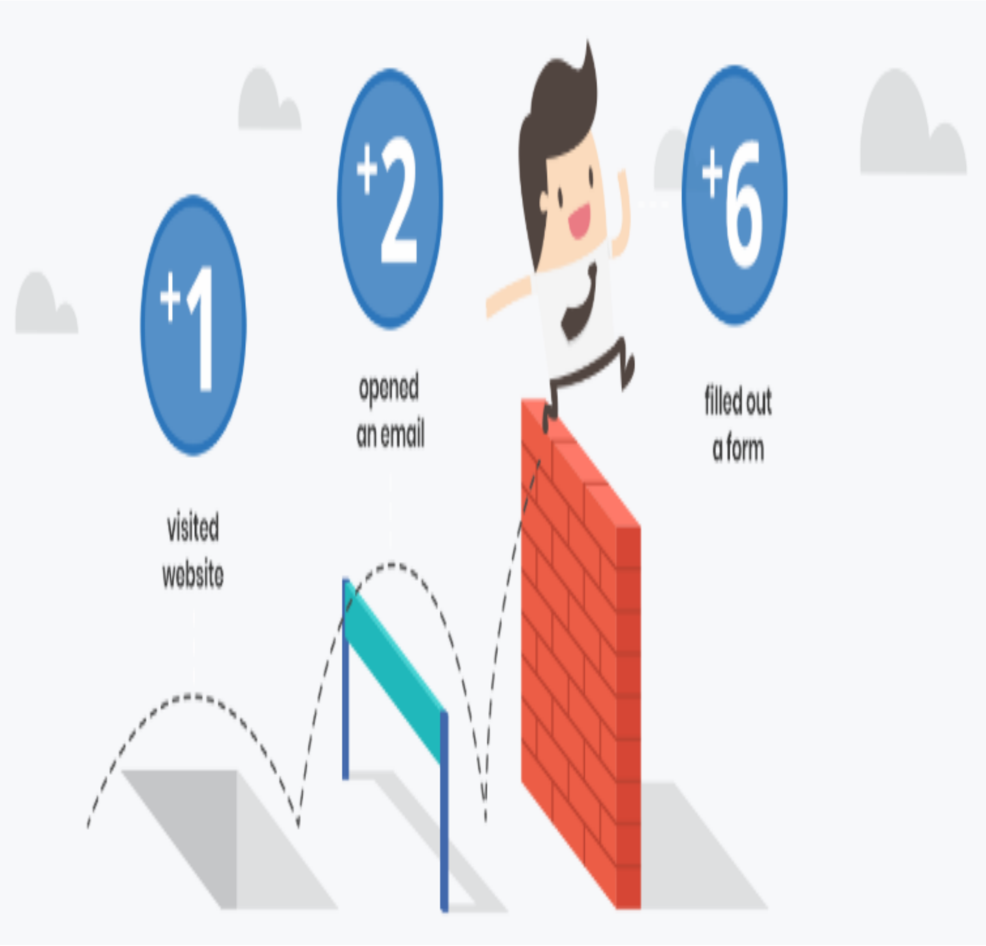
1.Karandeep Malik

2. Vaibhavi Shendge

# Problem Statement:

## Introduction:

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 typical lead conversion rate at X education is around 30% and we have to help them in increasing the rate around 80%.



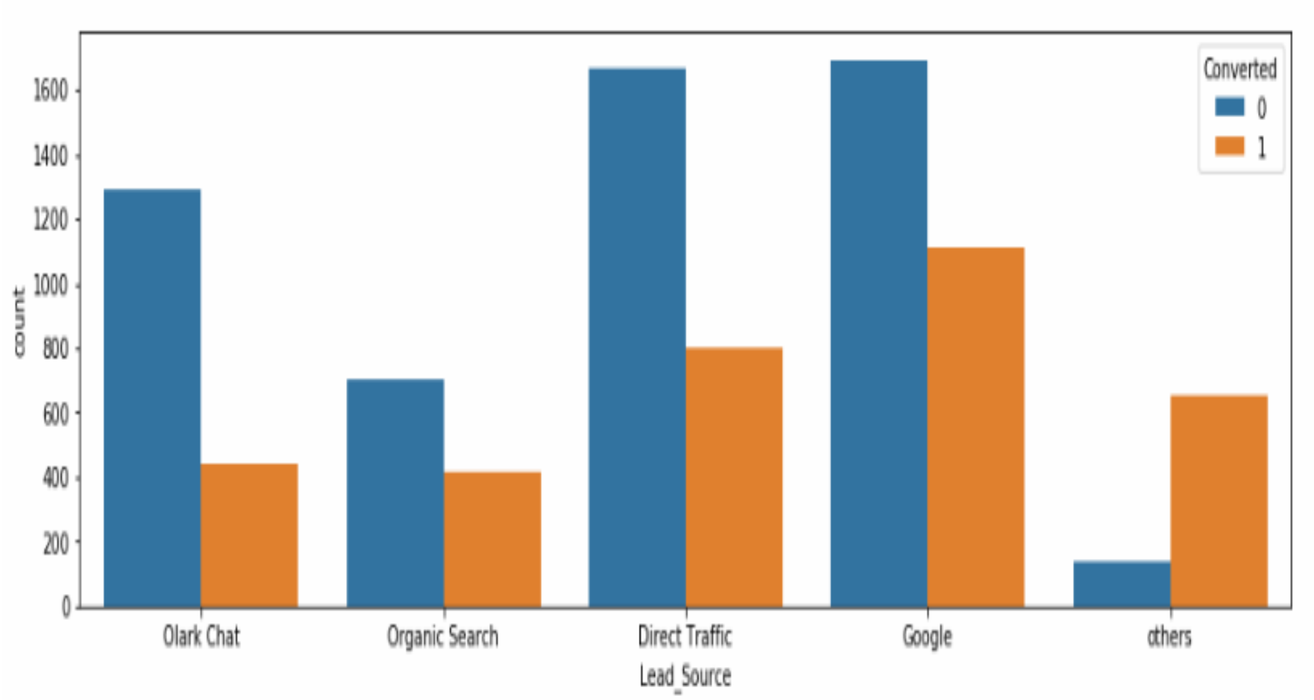
## Objective:

To build a Logistic Regression Model to predict

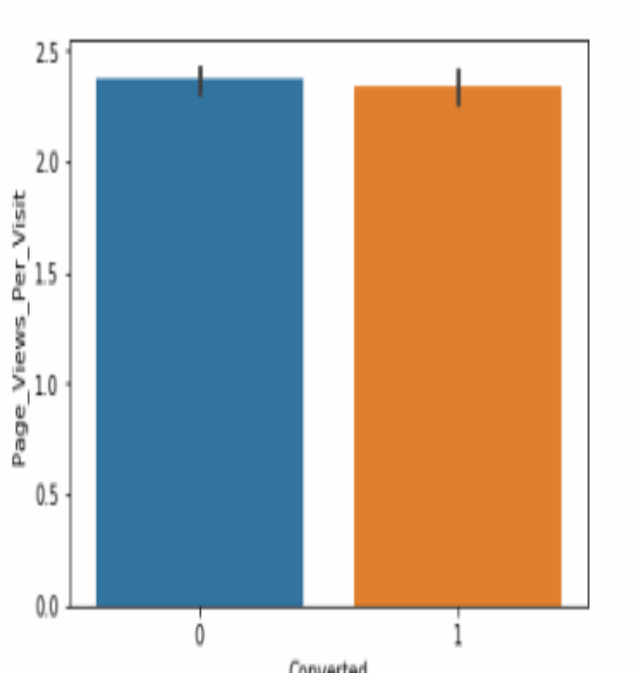
- To find the most promising leads( hot leads)
- Build a model to assign the score based on their probability of conversion.

# Problem Solving Methodology

- Understanding the Data set & Data Preparation
- Applying Recursive feature elimination to identify the best performing subset of features for building the model.
- Building the model with features selected by RFE.
- Eliminate all features with high p-value and VIF values and finalize the model
- Use the model for prediction on the test dataset and perform model evaluation for the test set.
- Decide on the probability threshold value based on optimal cutoff point and predict the dependent variable for the training data.
- Perform model evaluation with various metrics like sensitivity, specificity, precision, recall, etc.

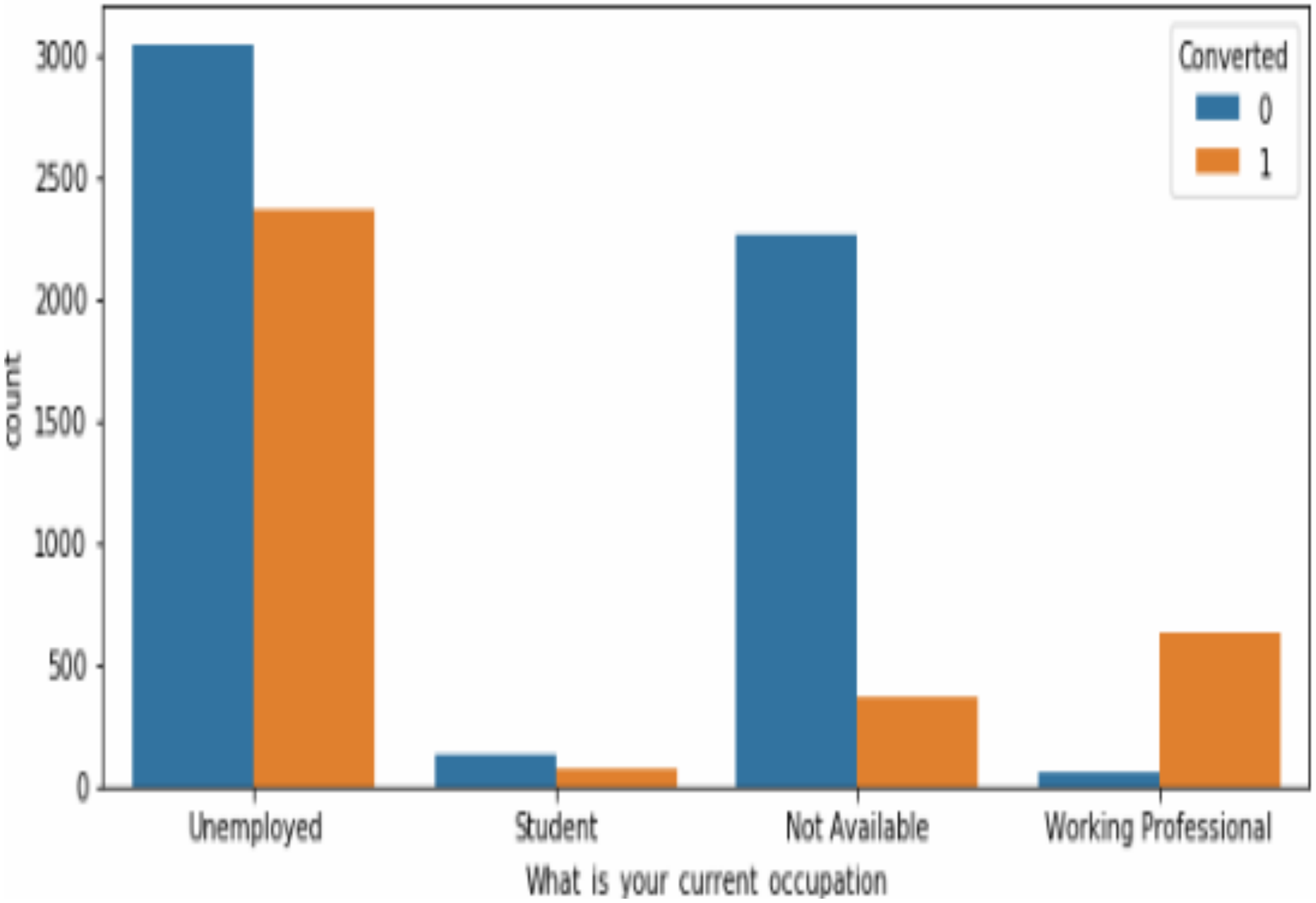


The above graph is plotted between Lead source and conversion count. From this we can say that conversion count at Google lead source is highest.

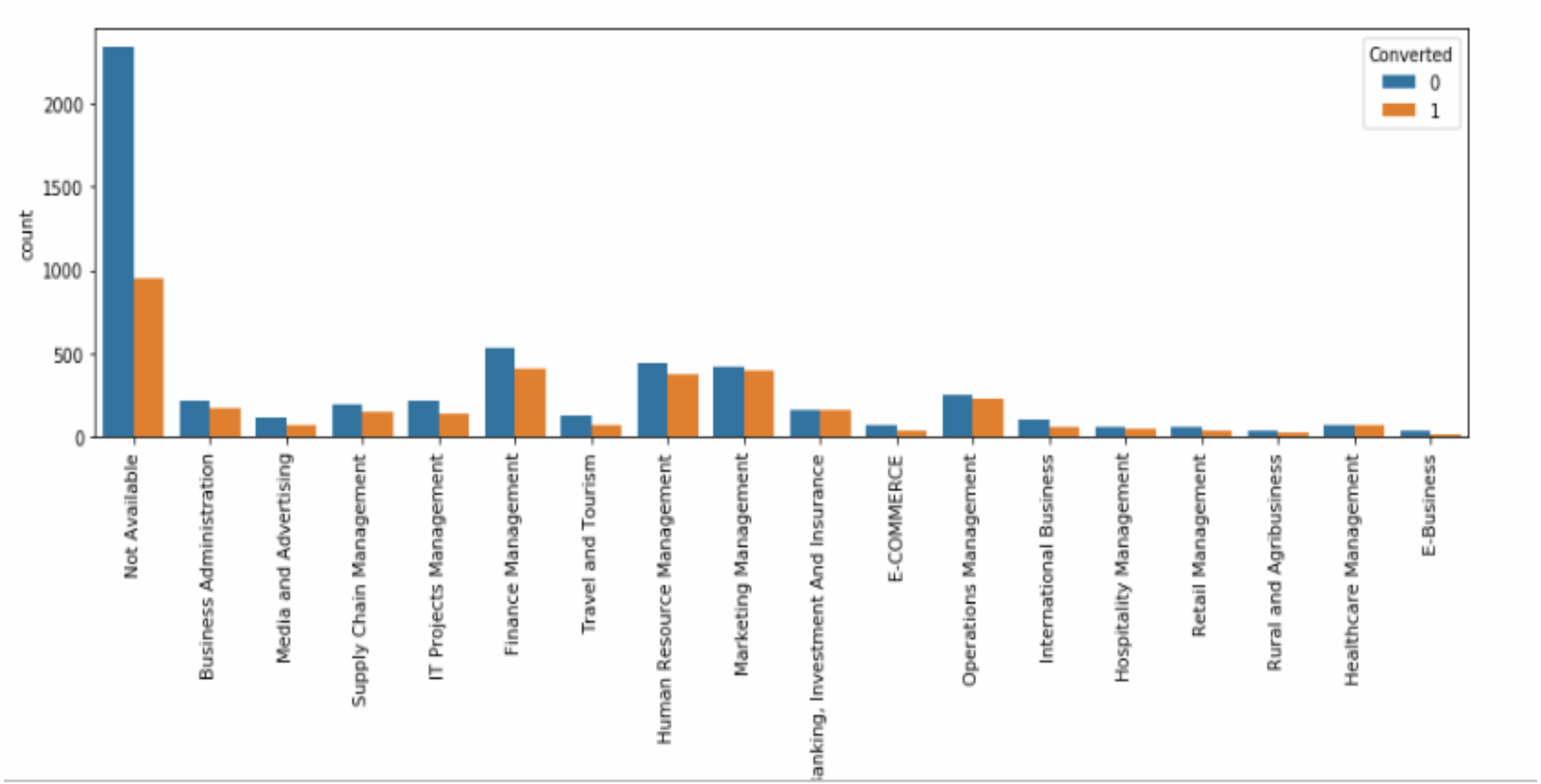


The above graph is plotted between conversion count and page per views. From this it is clear that conversion rate is low and not every visitor is getting converted.

The graph is plotted between current occupation of viewer and count of viewer. The unemployed viewers are much interested in the conversion. And the working profession are on second position on conversion. Where as students are very less interested in conversion.

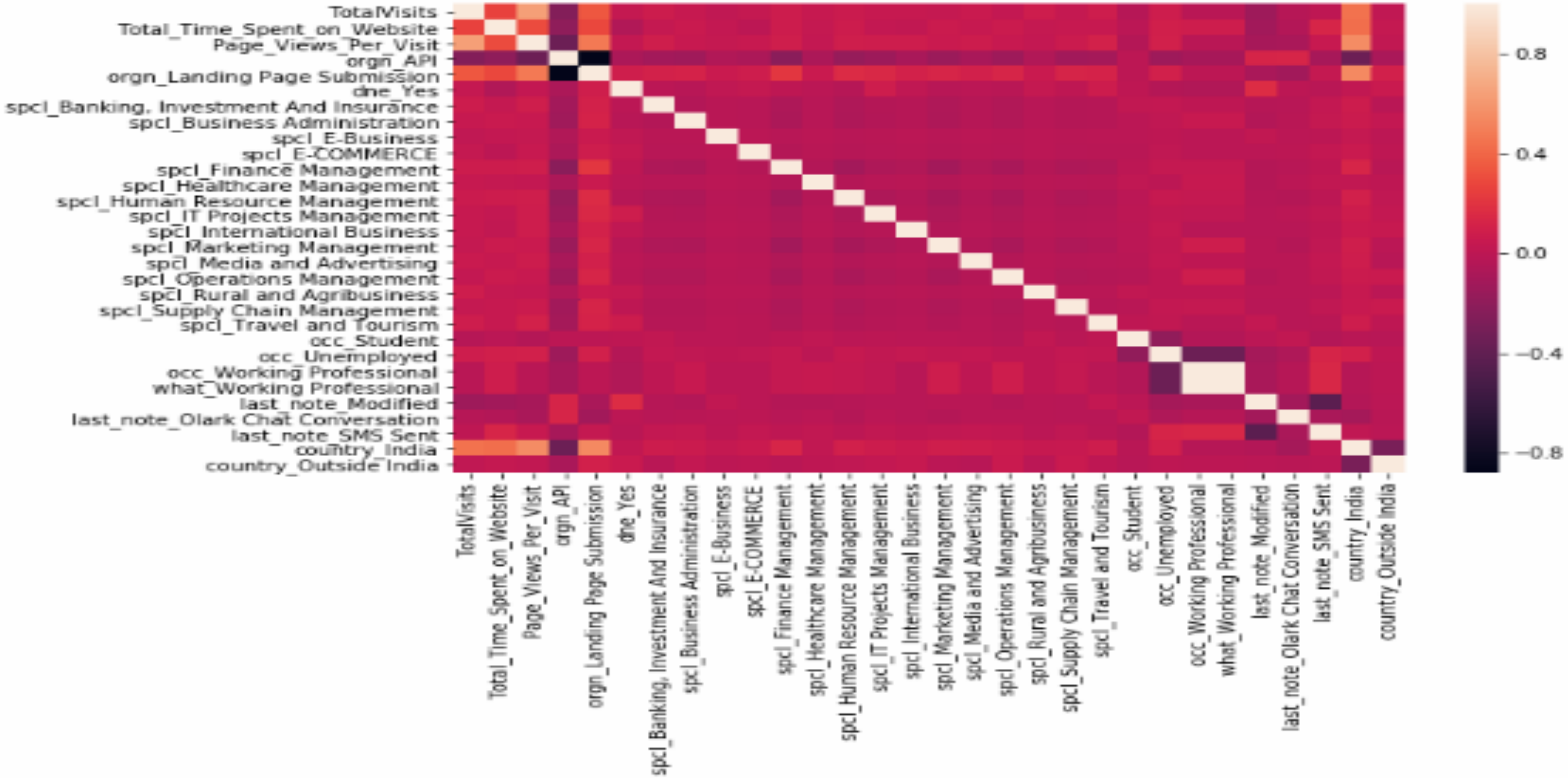


# Plot between Specialization to the converted rate

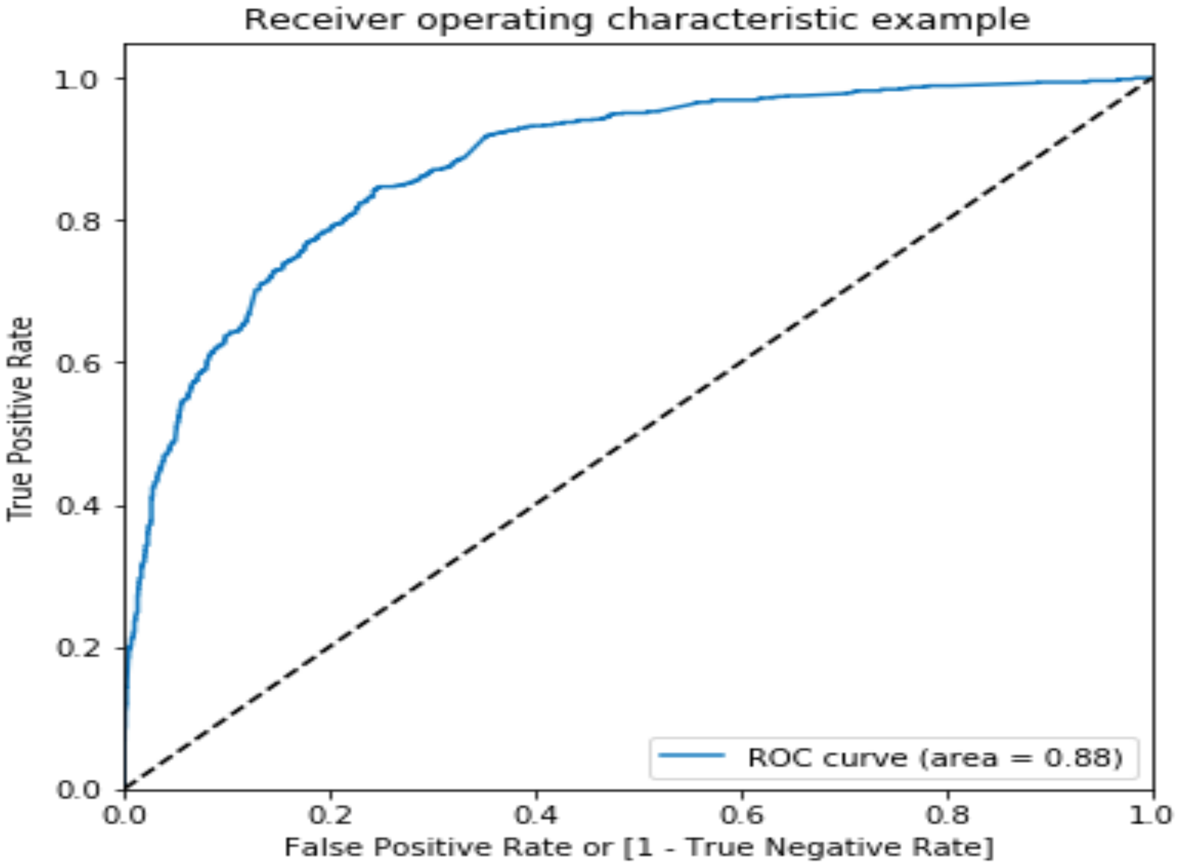


This suggests that people with Business Administration S, Marketing Management, Human Resource Management and Operations Management Specialization are more interested in joining these courses

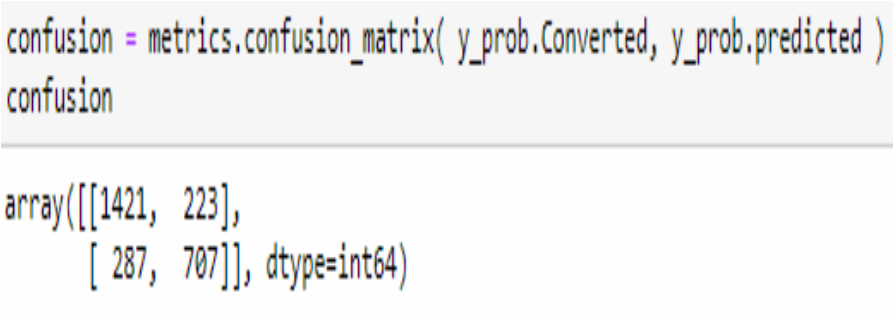
# Heatmap for the correlation:



# ROC CURVE:

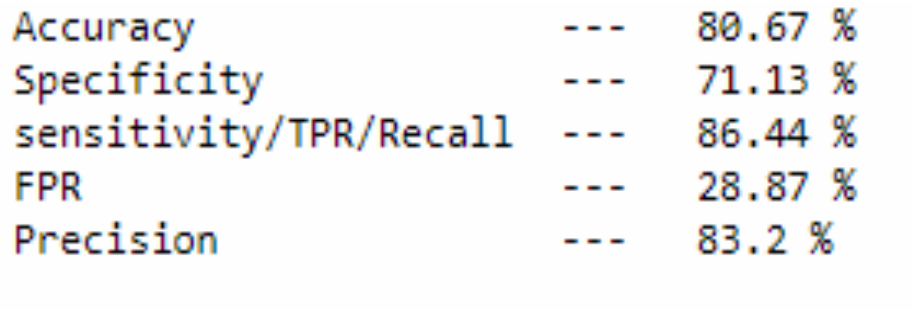

Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.88)

# Model Result:



```
confusion = metrics.confusion_matrix( y_prob.Converted, y_prob.predicted )
confusion
```

```
array([[1421,  223],
       [ 287,  707]], dtype=int64)
```



```
Accuracy                     ---    80.67 %
Specificity                  ---    71.13 %
sensitivity/TPR/Recall       ---    86.44 %
FPR                          ---    28.87 %
Precision                    ---    83.2 %
```

From the above results it can be seen that

- The accuracy is 80.67%
- Specificity is 71.13%
- Sensitivity/TPR/Recall is 86.44%
- FPR is 28.87%
- Precision is 83.2%

# Conclusion:

- Main variables that contributed to Lead conversion were: Total Visits, Page Views Per Visit, Total Time Spent on Website, Lead Origin and Occupation
- The conversion probability of lead increases with increase in Total visits, Total Time Spent on websiteOcuupation being a working Professional
- The conversion probability of lead decreases with increase in Page Views Per Visit, when the Lead Origin is API, Lead Origin is Landing Page Submission

# Recommendation:

- The Sensitivity of the model can be by capturing more data with less gaps and hence better prediction of Conversion
- A lot of gaps existed in data for lot of essential parameters like Lead Profile which could have been significant contributors and would have resulted in better accuracy. Having this information for greater no of leads would have helped.
- The model can be tuned in the future based on cutoff to allow more no of leads or less based on available resources.

.