Lead Score Assignment

Abhishek Vaggar

Hemanth Kumar

Sonia Dhanda

Problem statement

- An X Education need help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

- Goals and Objectives There are quite a few goals for this case study.

- 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. 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.

Step by step approach towards problem

- Data clean up

    - Remove duplicate

    - Check for missing and NA values

    - Remove such columns depending the frequency of occurrence

    - Check for outliers

    - Imputation of selected values if required

- Exploratory data analysis

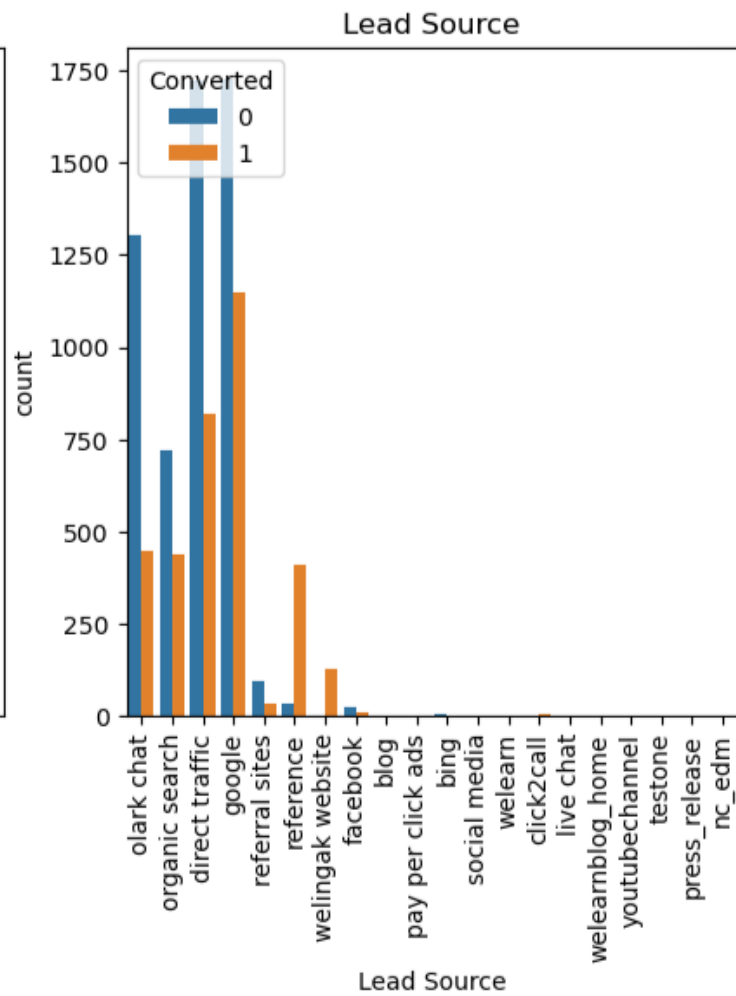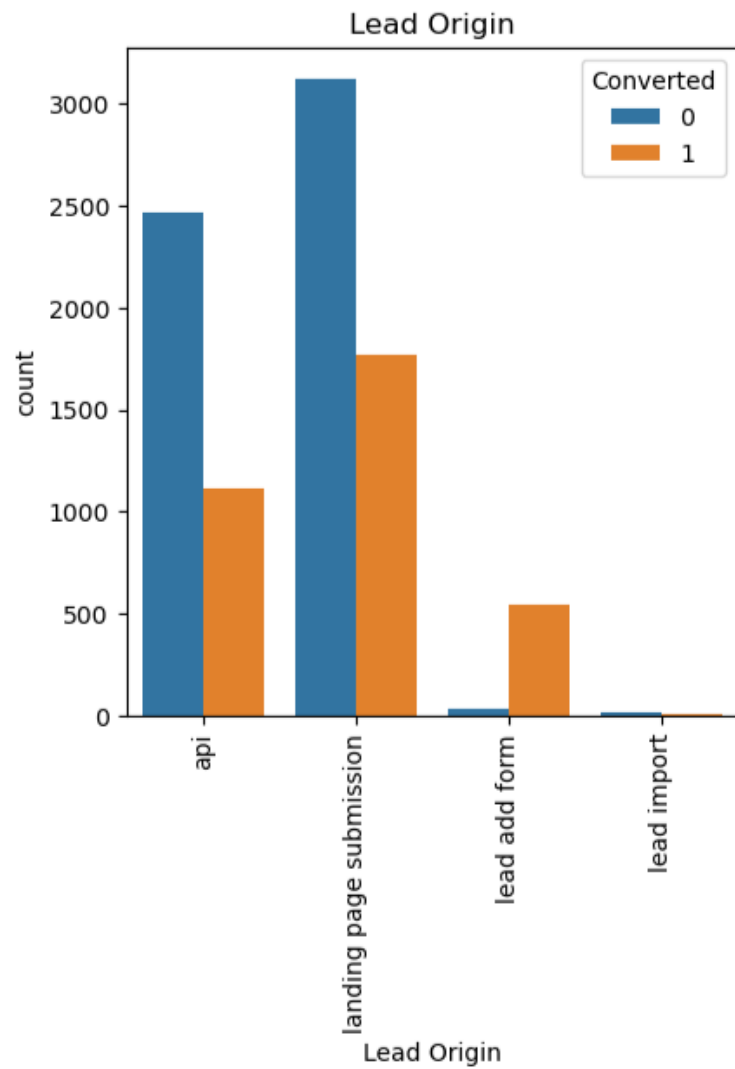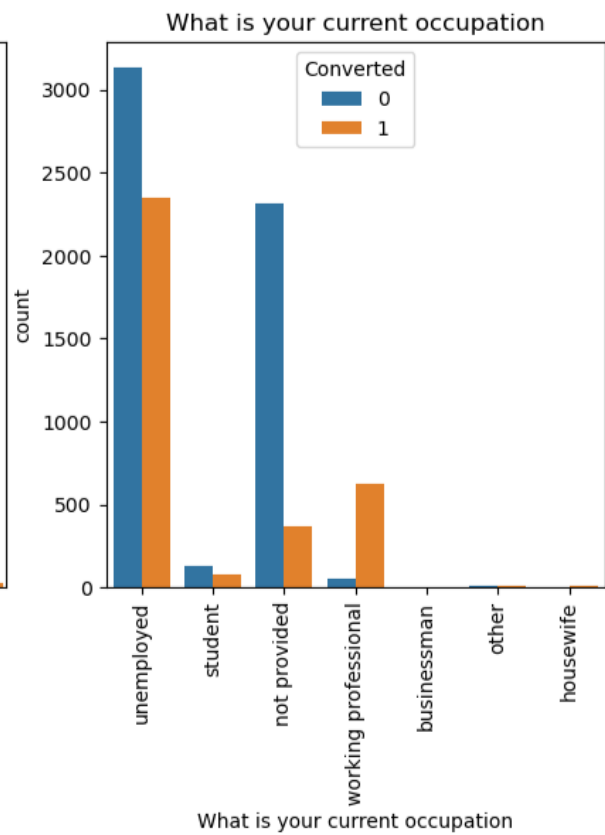- Validate the model
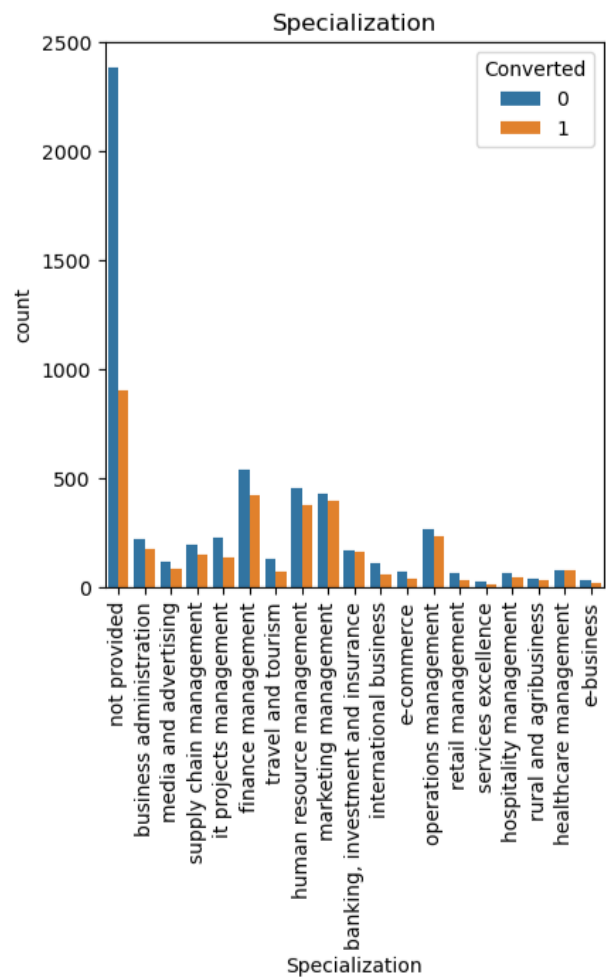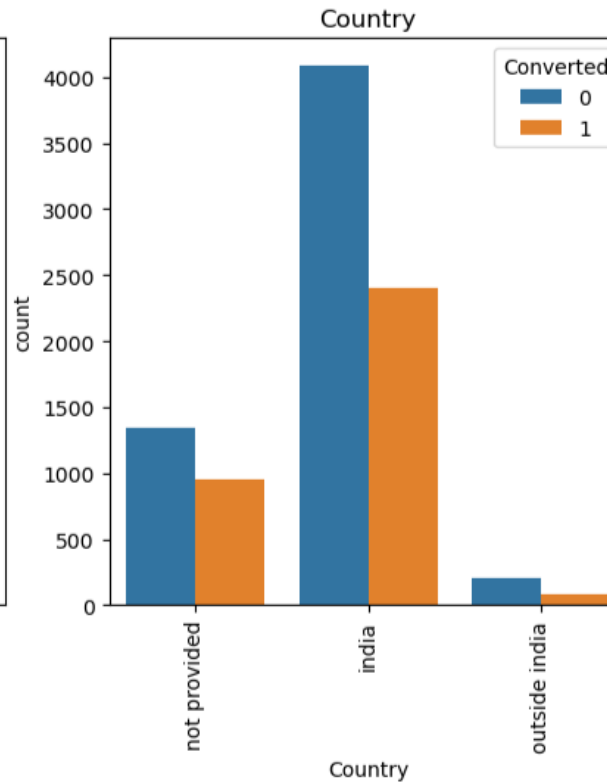
- Conclusion

Data manipulation

- Please review the comments provided in python file script

EDA

Do Not Email

Converted("Y variable")

Country

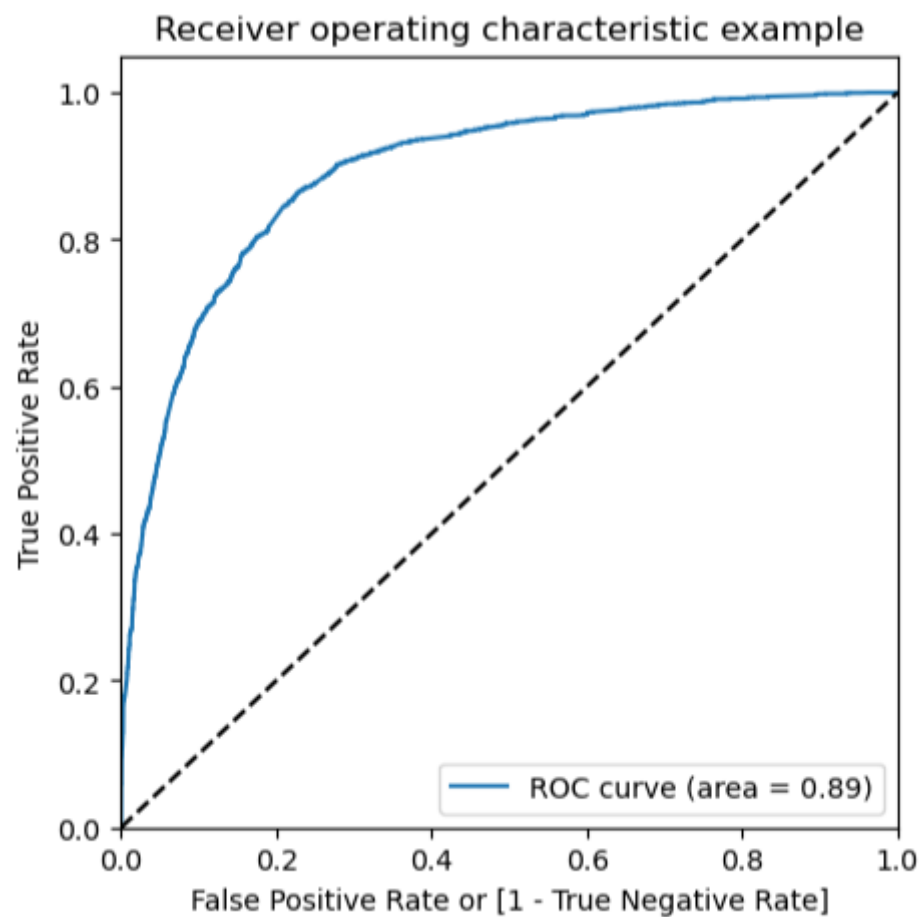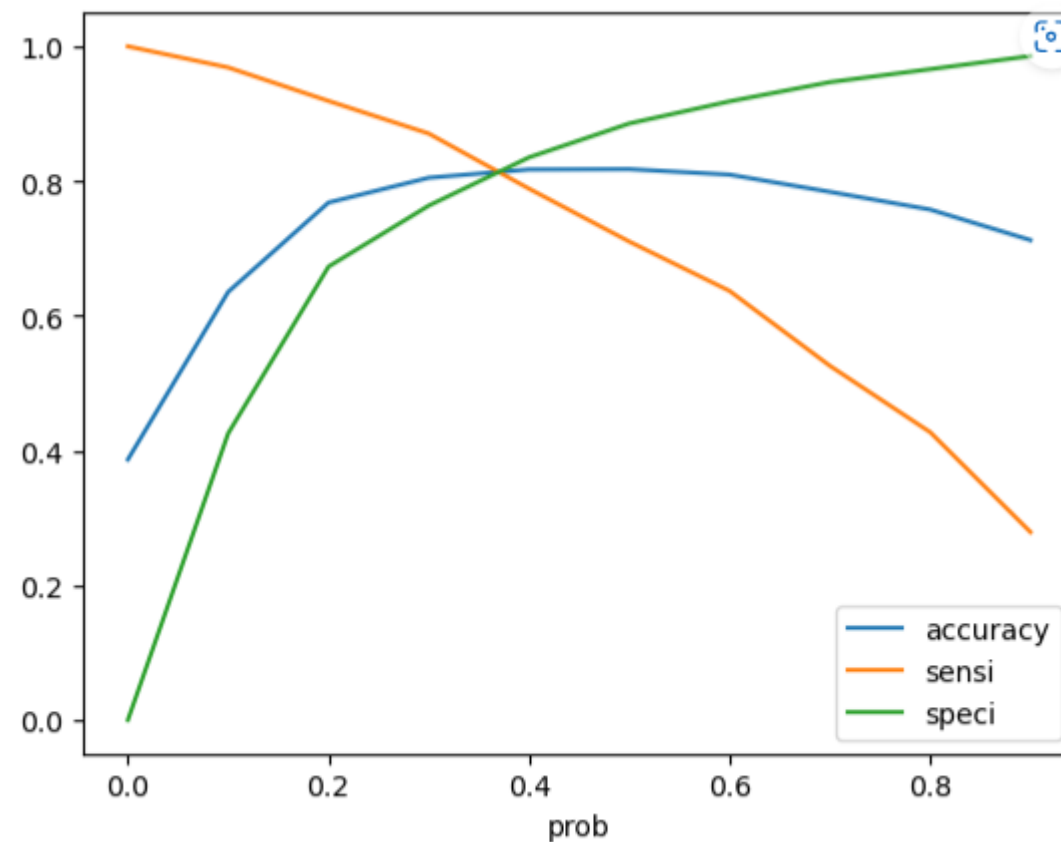## Lead Origin
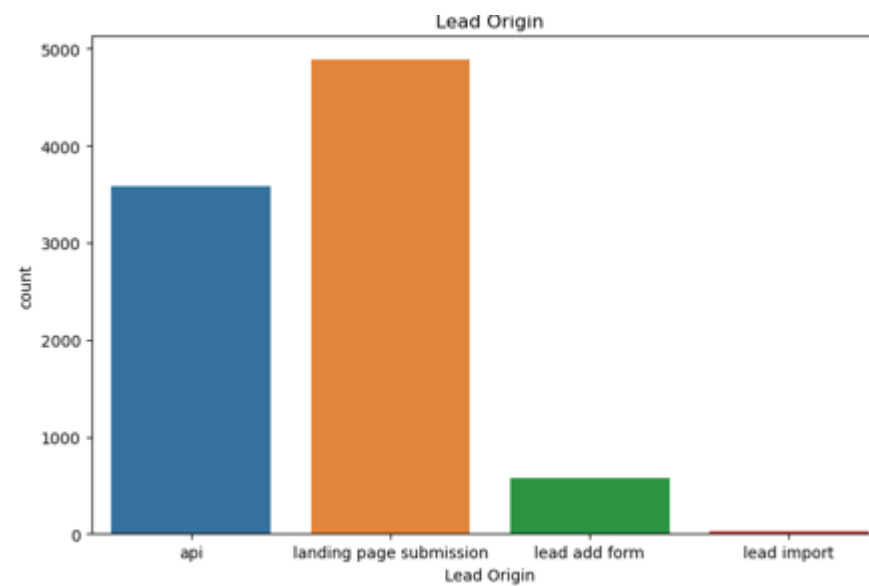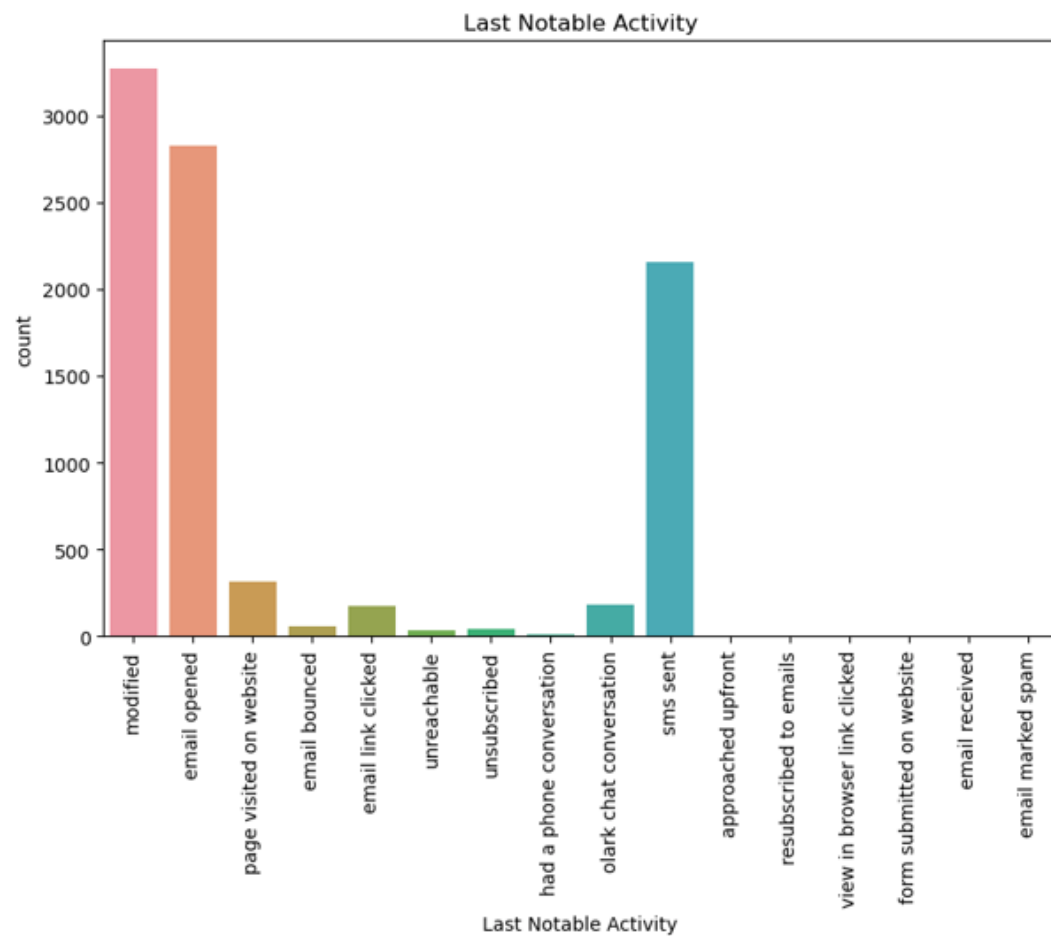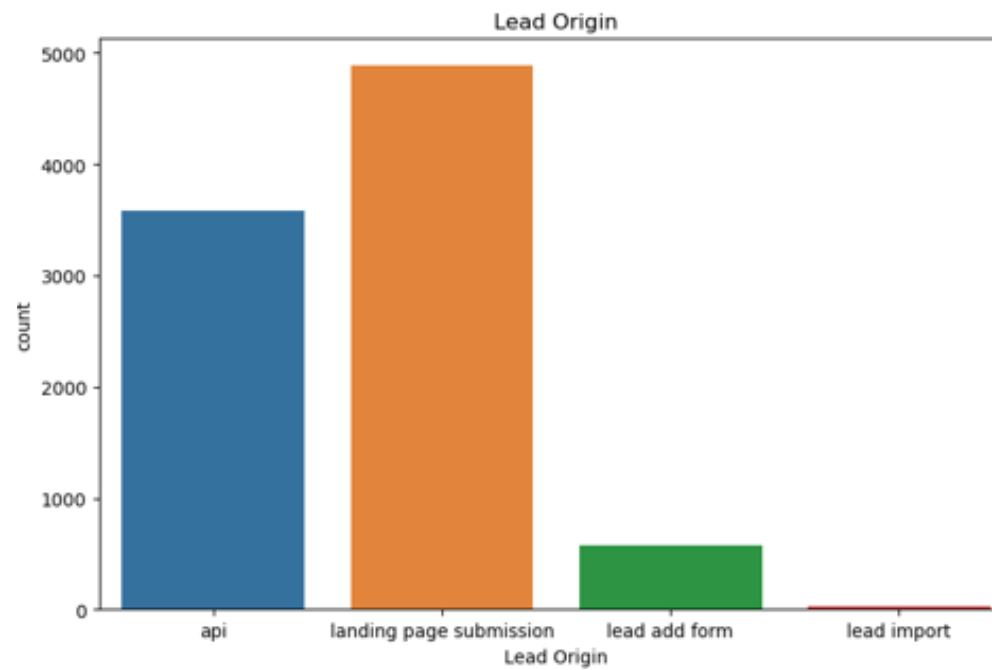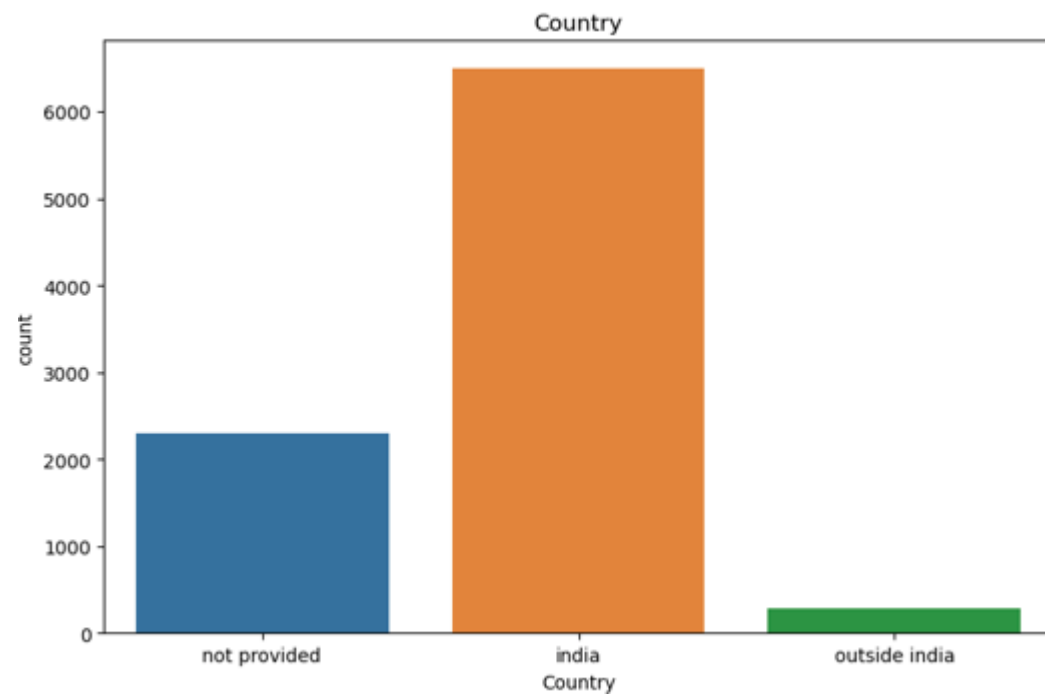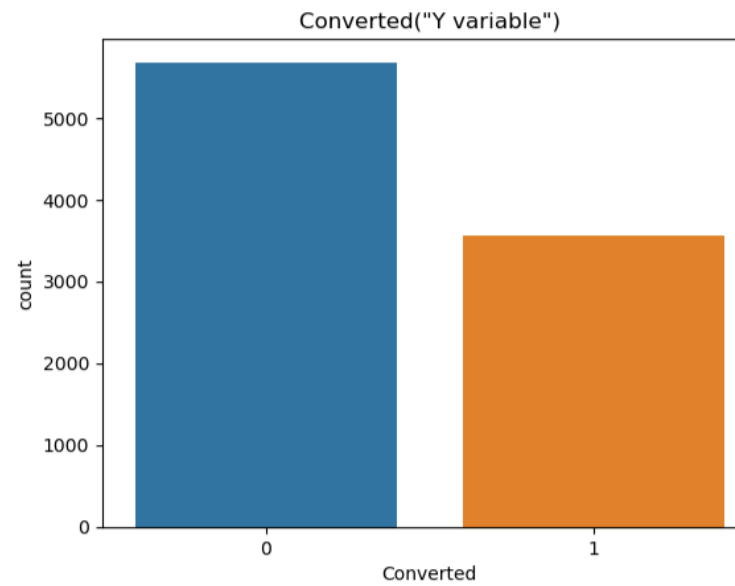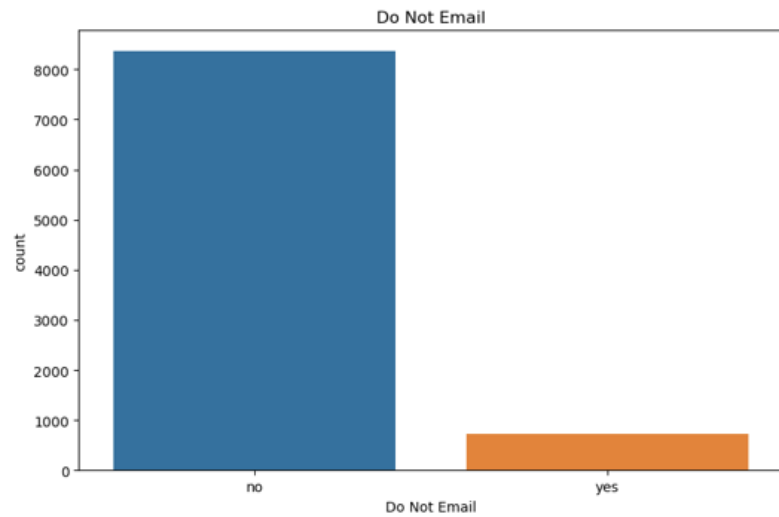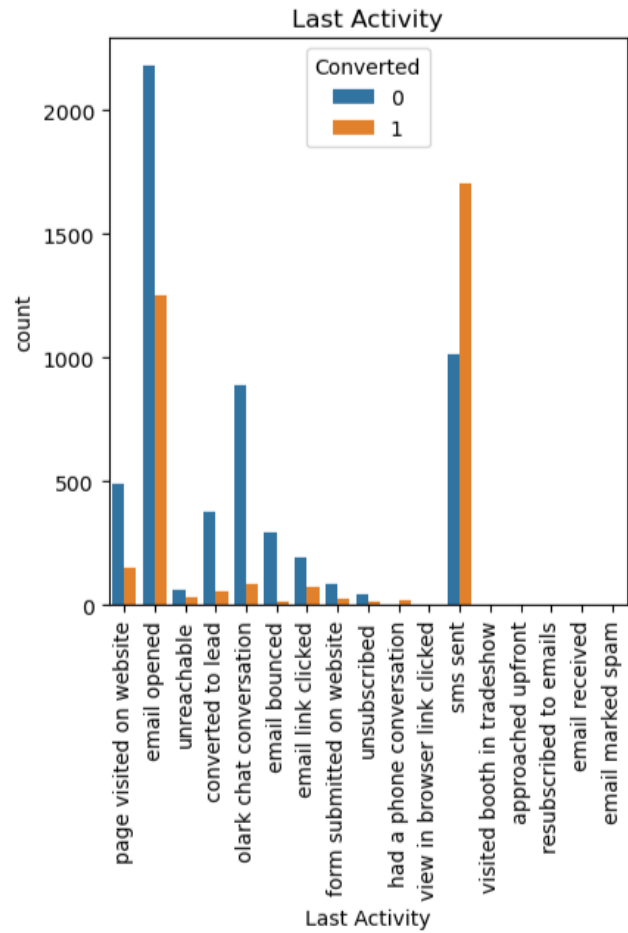
Model building

- Split date into training set and testing set

- Ratio 70:30

- Run RFE with step variable = 15

- Model build with variable –p value greater than 0.05 and vif greater than 5 (standard procedure)

- Overall accuracy : 81.76

ROC

Receiver operating characteristic example

Cut off is at approx. 0.38

Conclusion

- Variables that would be potential buyers are :

1. time spent on the website

2. Total visits to the website

3. lead source was:

        - Google

        - Direct traffic

        - Organic search

        - Welingak website

4. last activities: SMS, Olark chat conversation

5. When the lead origin is Lead add format.

6. When their current occupation is as a working professional.

X education can have success as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.