

# LEAD SCORE CASE STUDY

Teja Shashank

# PROBLEM STATEMENT

## INTRODUCTION:

An education company, X Education sells online courses to industry professionals. The company markets its courses on various websites and search engines such as Google.

Once people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals.

Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. The typical lead conversion rate at X education is around 30%

# Business Goals

- Company wishes to identify the most potential leads, also known as “Hot Leads”
- The company needs a model wherein a lead score is assigned to each of the leads such that the customer with higher lead score have a higher conversion chance and customer with lower lead score have a lower conversion chance
- The CEO, in particular, has given a ballpark number for the lead conversion rate i.e. 80%

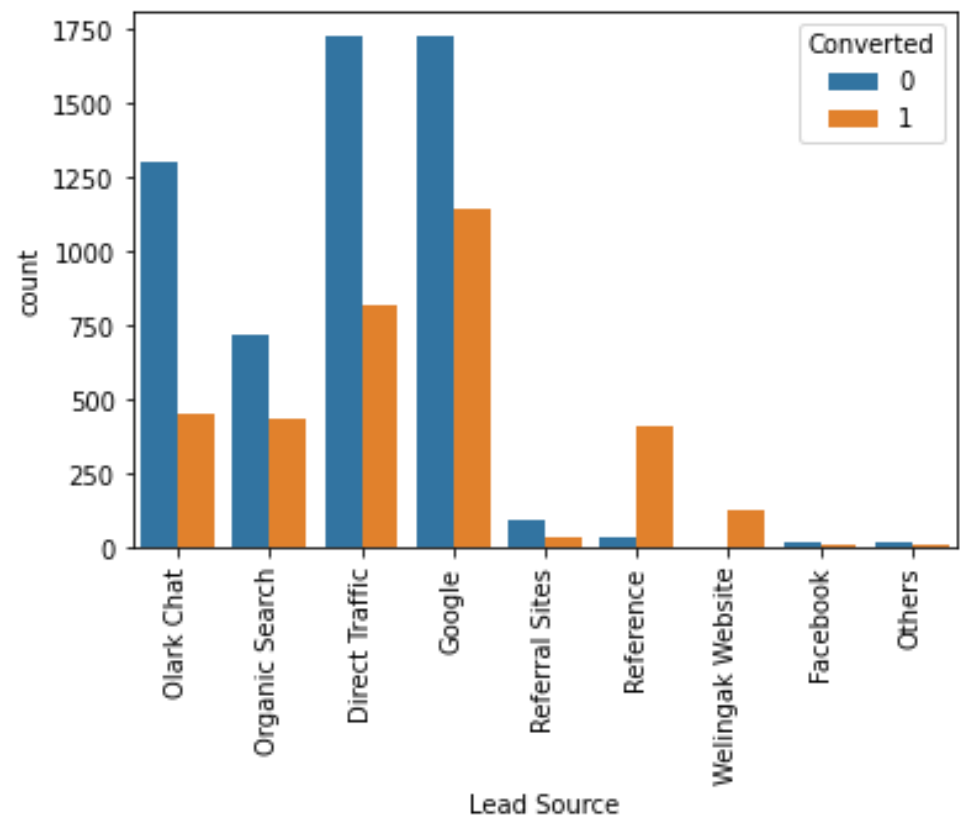
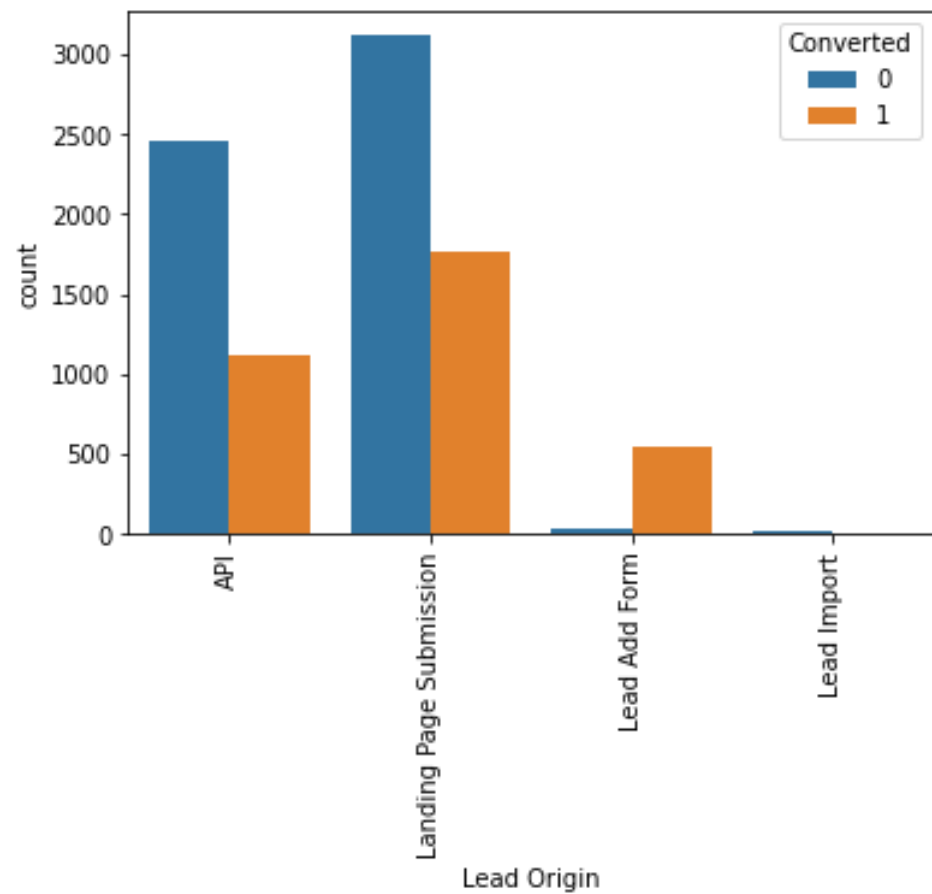
# Approach

- Data Cleaning
- EDA
- Dummy variable creation
- Logistic regression model building
- Model Evaluation
- Conclusion & Recommendation

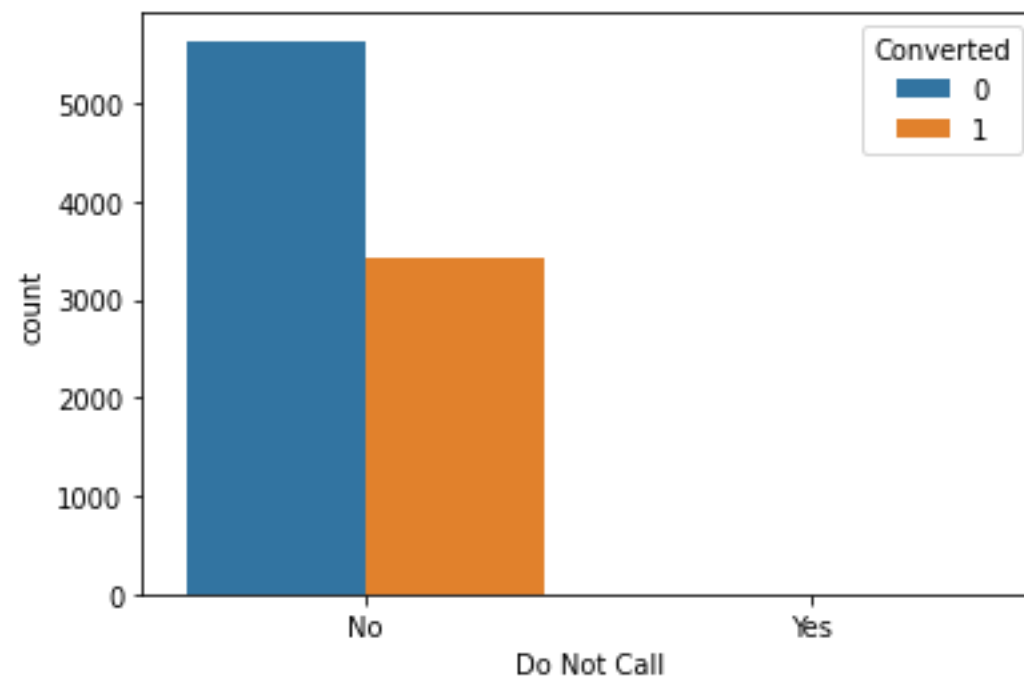
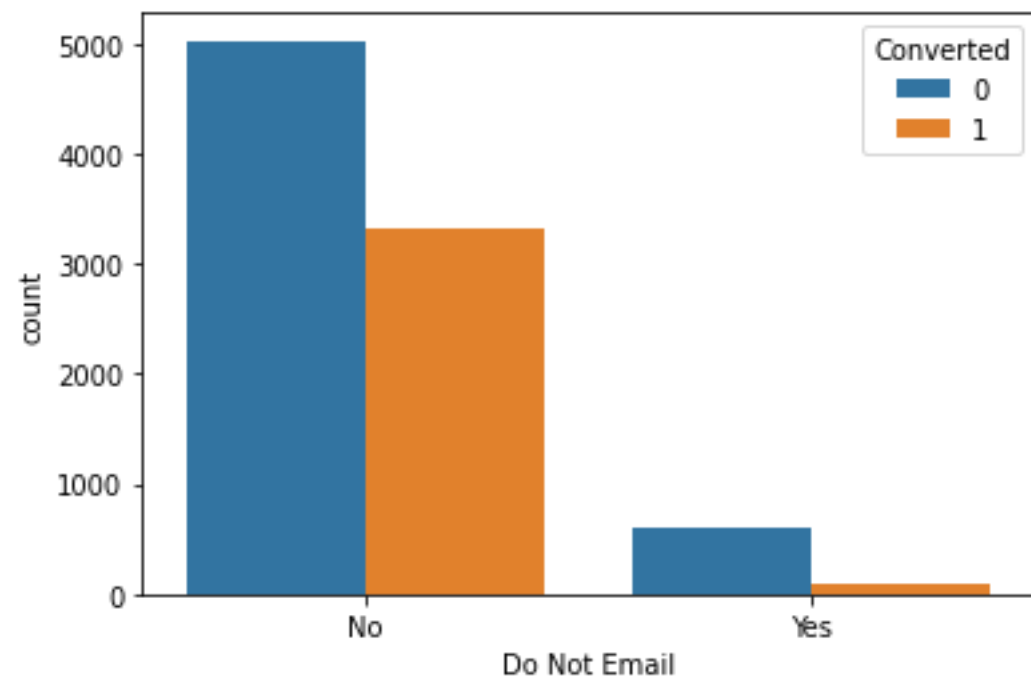
# DATA CONVERSION

- CONVERTING THE VARIABLE WITH VALUES YES/NO to 1/0s
- CONVERTING THE 'SELECT' VALUES WITH NaNs
- DROPIING THE COLUMNS HAVING >70% OF NULL VALUES
- DROPPING UNNECESSARY COLUMNS
- DROPPING THE ROWS AS THE NULL VALUES WERE <2%

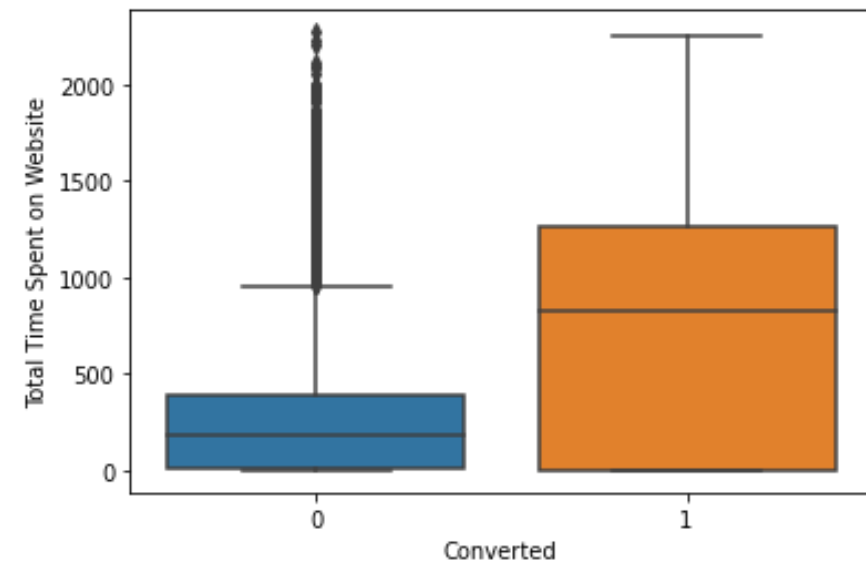
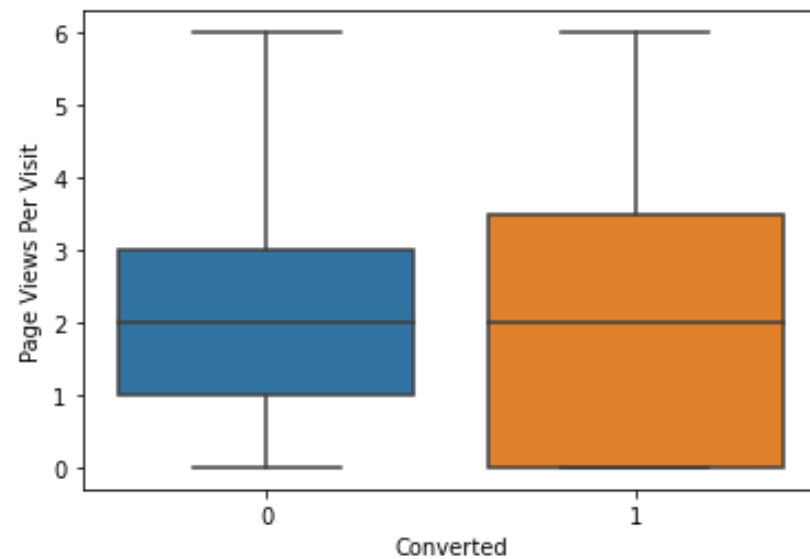
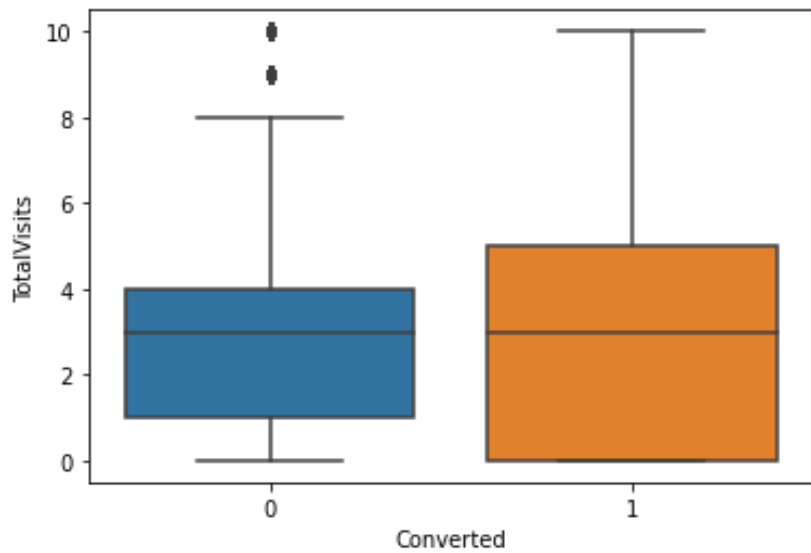
# EDA



# EDA

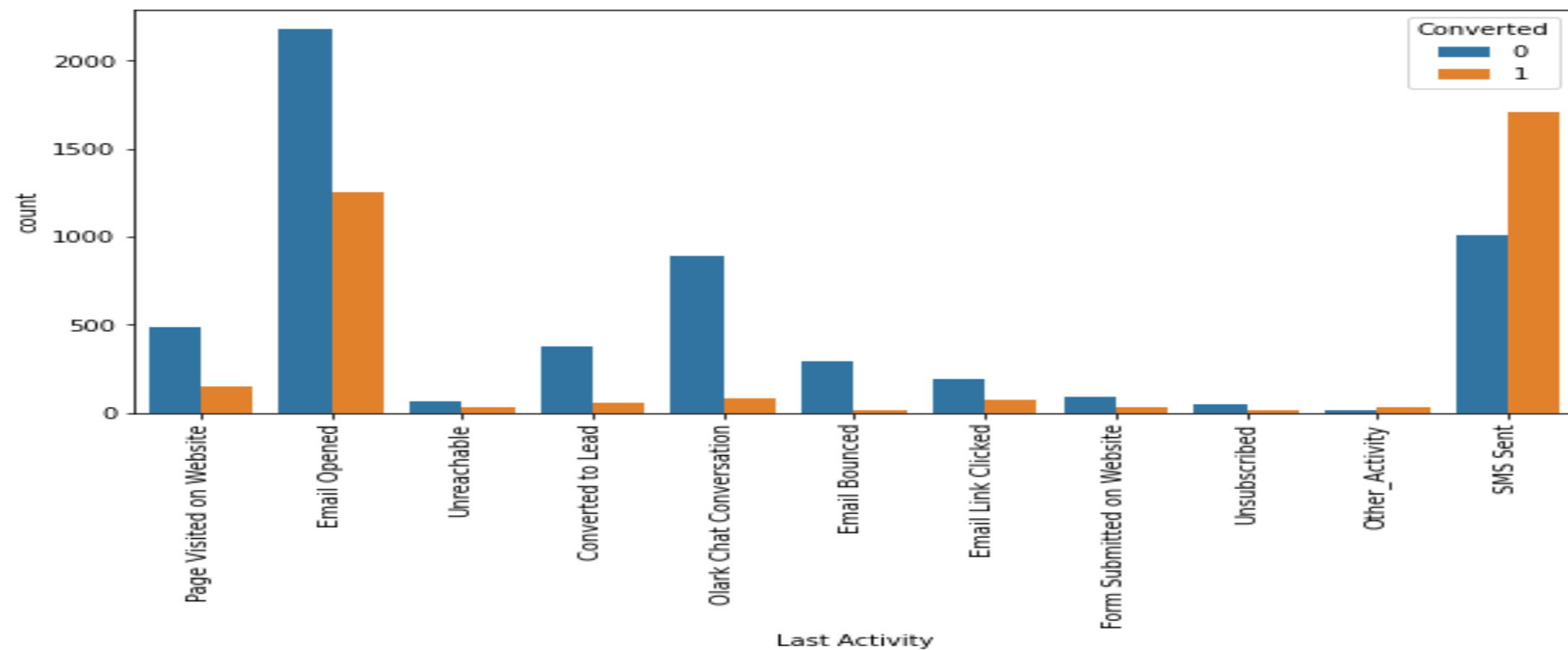


# EDA

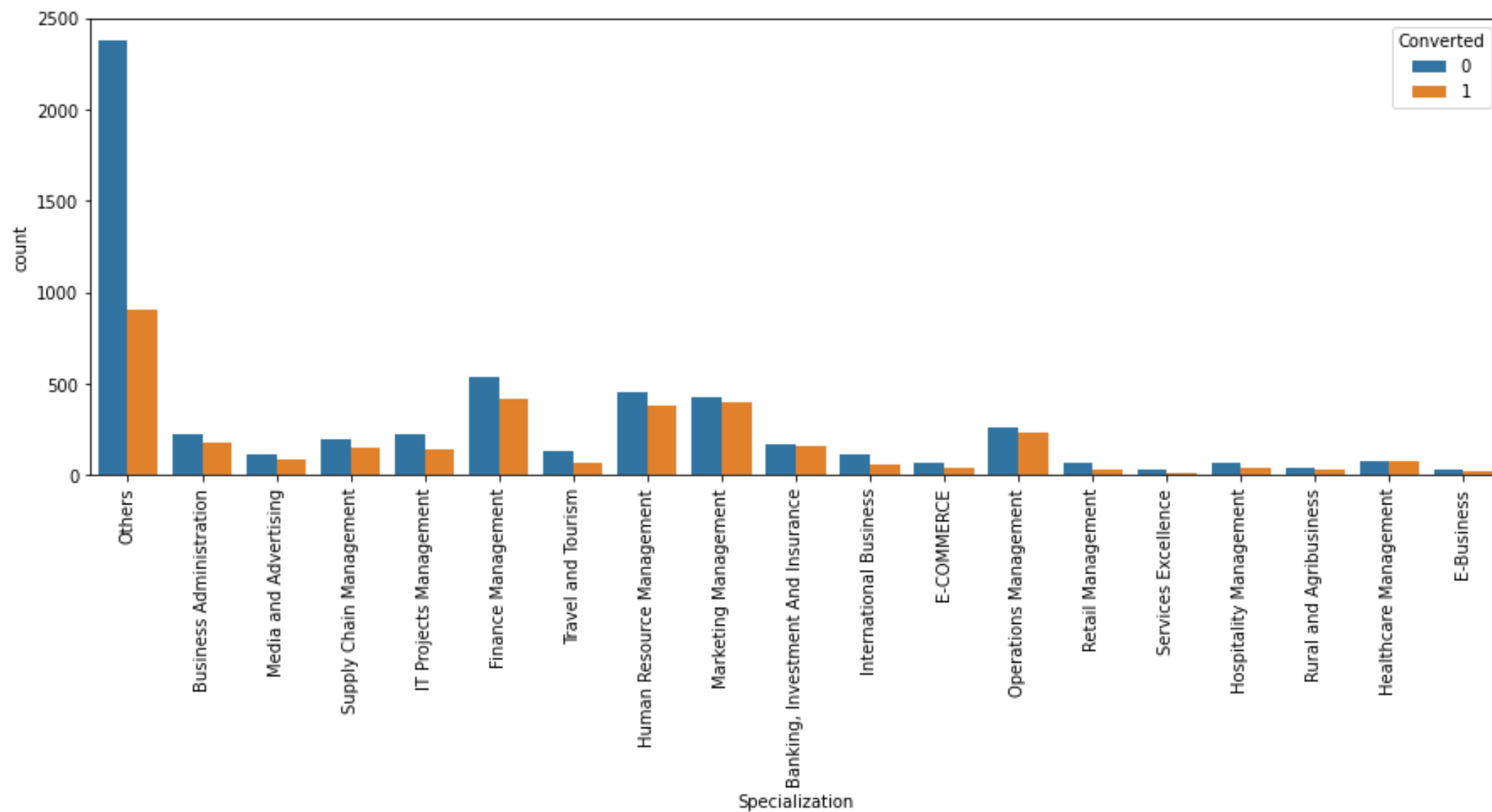




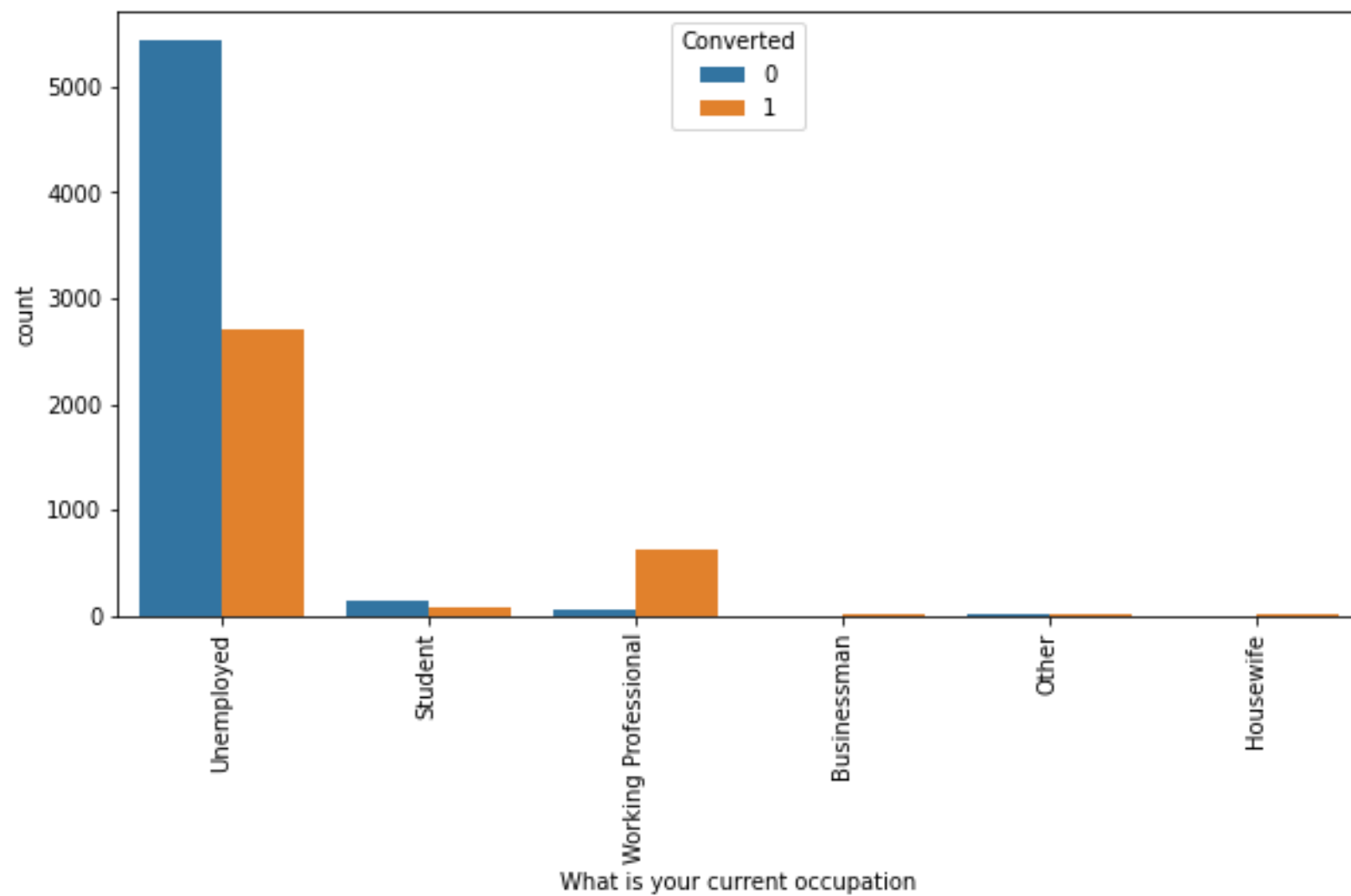
# EDA



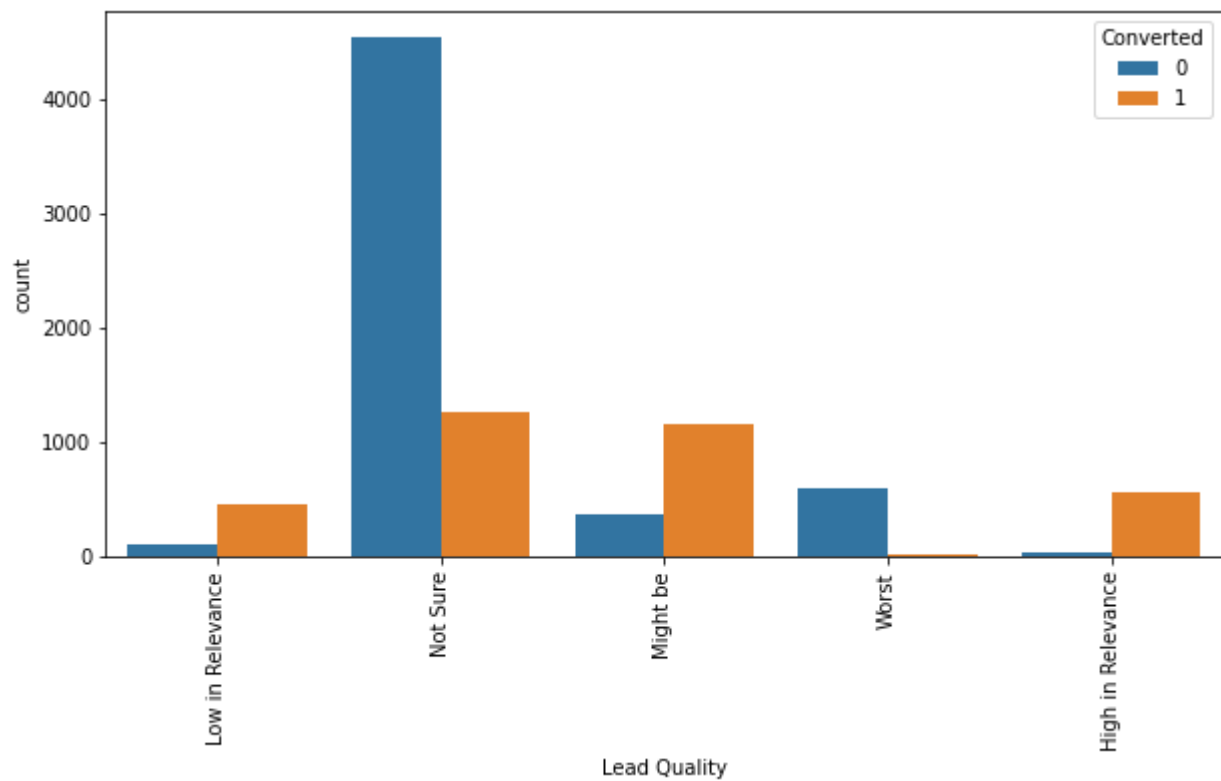
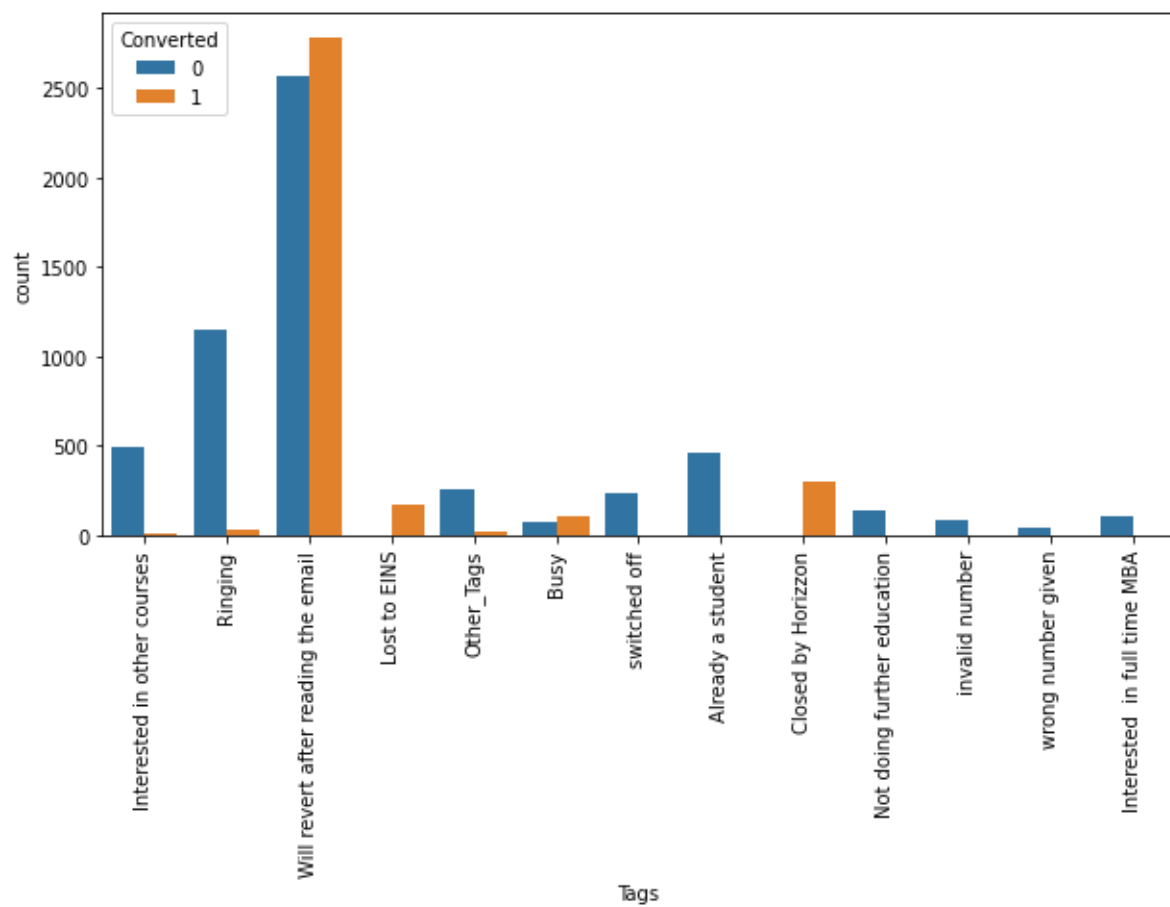
# EDA



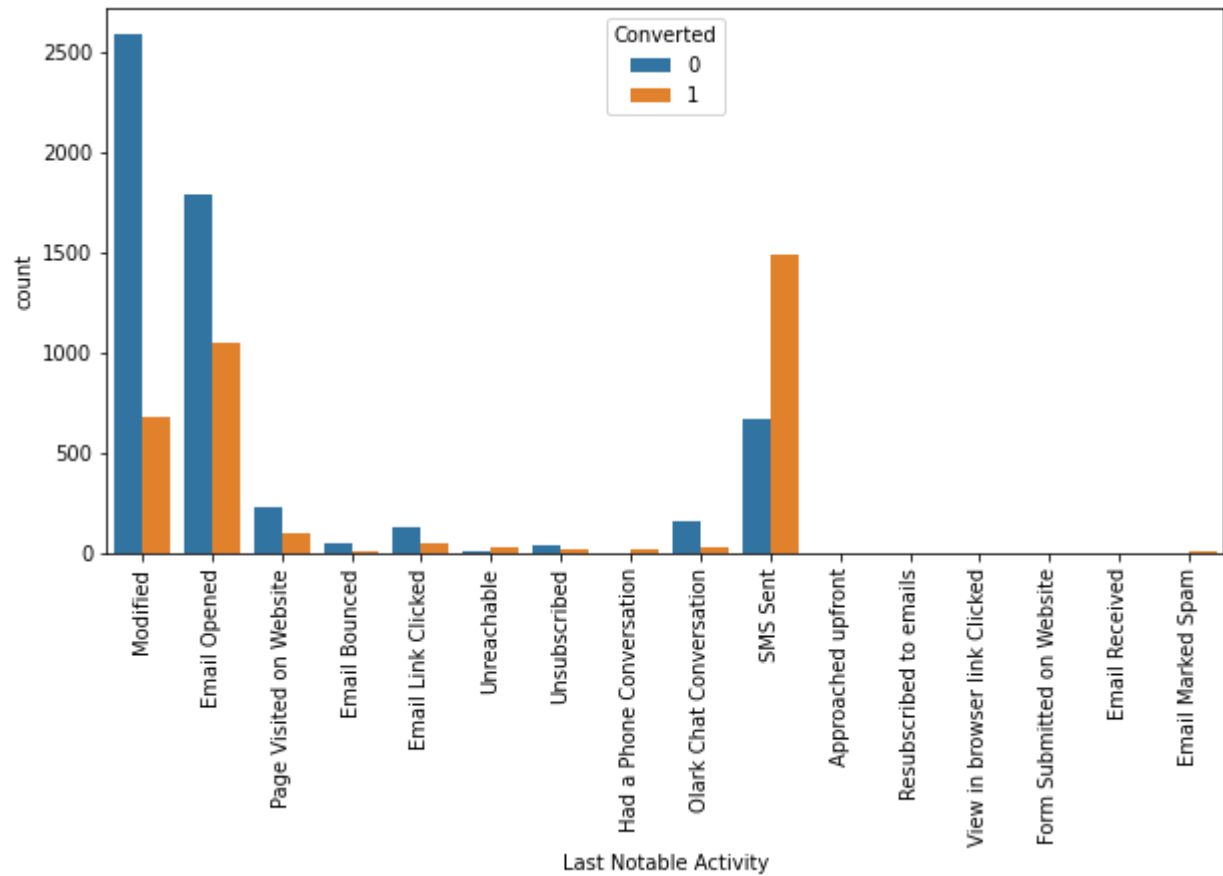
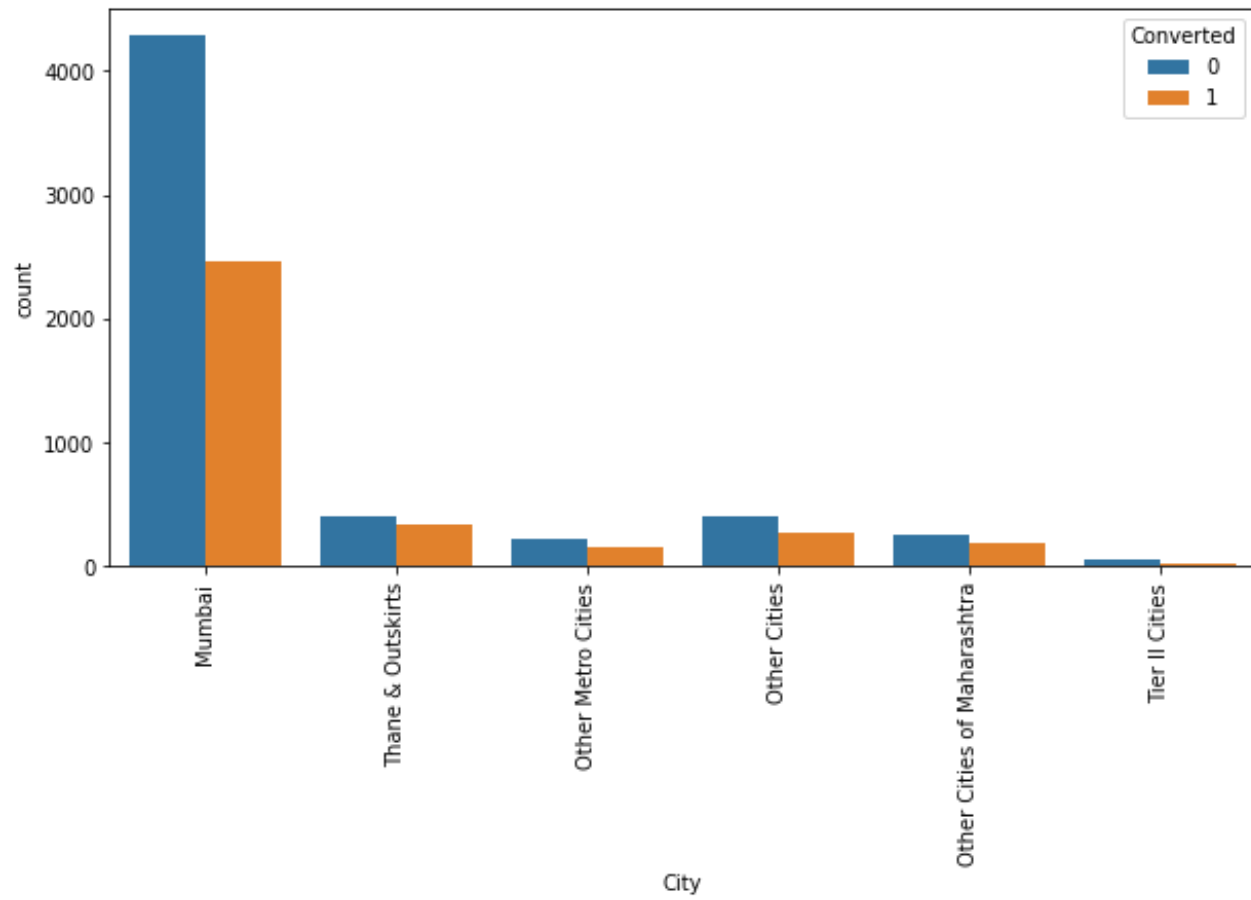
# EDA



# EDA

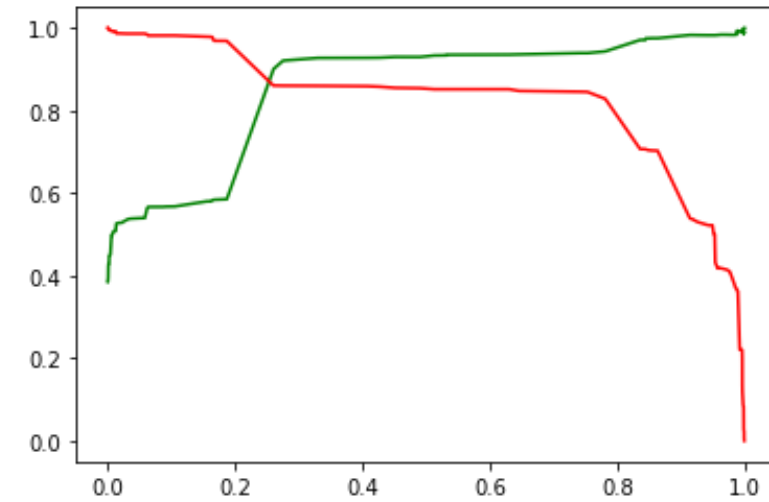
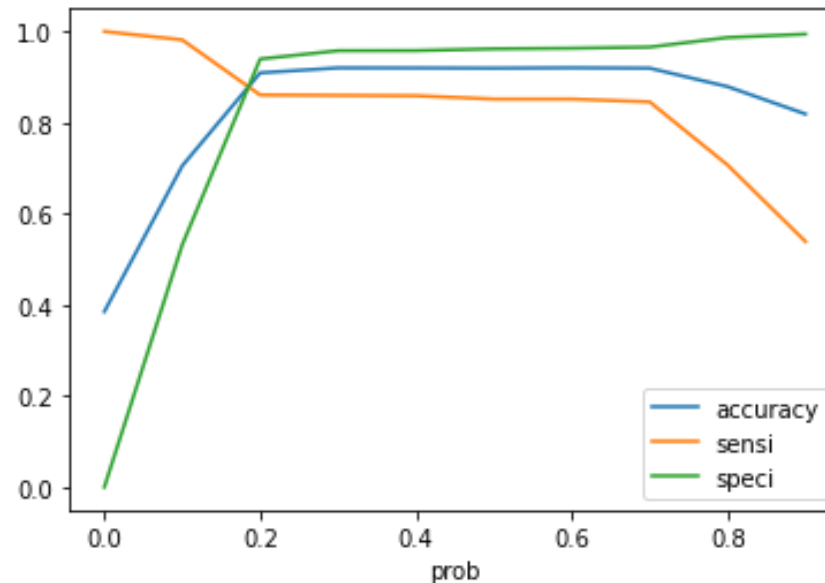
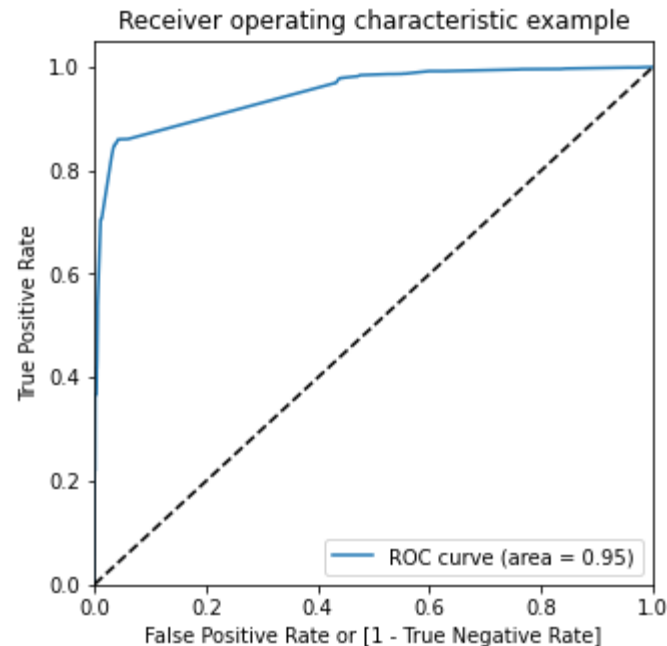


# EDA



# Model Building

- Splitting the data into test and training sets
- Using RFE to choose top 15 variables
- Overall Accuracy is 90.4%



# Model Prediction

	<b>coef</b>
<b>const</b>	-0.9661
<b>Do Not Email</b>	-1.3127
<b>Lead Origin_Lead Add Form</b>	1.0963
<b>Lead Source_Welingak Website</b>	3.4147
<b>What is your current occupation_Unemployed</b>	-1.1746
<b>Tags_Busy</b>	3.8305
<b>Tags_Closed by Horizzon</b>	7.9914
<b>Tags_Lost to EINS</b>	9.2178
<b>Tags_Ringing</b>	-1.6472
<b>Tags_Will revert after reading the email</b>	3.9881
<b>Tags_switched off</b>	-2.2412
<b>Lead Quality_Not Sure</b>	-3.3158
<b>Lead Quality_Worst</b>	-3.9600
<b>Last Notable Activity_SMS Sent</b>	2.7443

# CONCLUSION

- The logistic regression model is used to predict the probability of conversion of a customer.
- This model has capability to adjust with the company's requirements in coming future
- TOP variables that contribute for lead getting converted in the model are:
  1. Tags\_Closed by Horizon
  2. Tags \_Lost to EINS
  3. Leads Quality\_Worst
- Hence Overall this model seems to be good