



LEAD\_SCORE\_

CASE\_STUDY

LOGISTIC\_REGRESSION



# 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

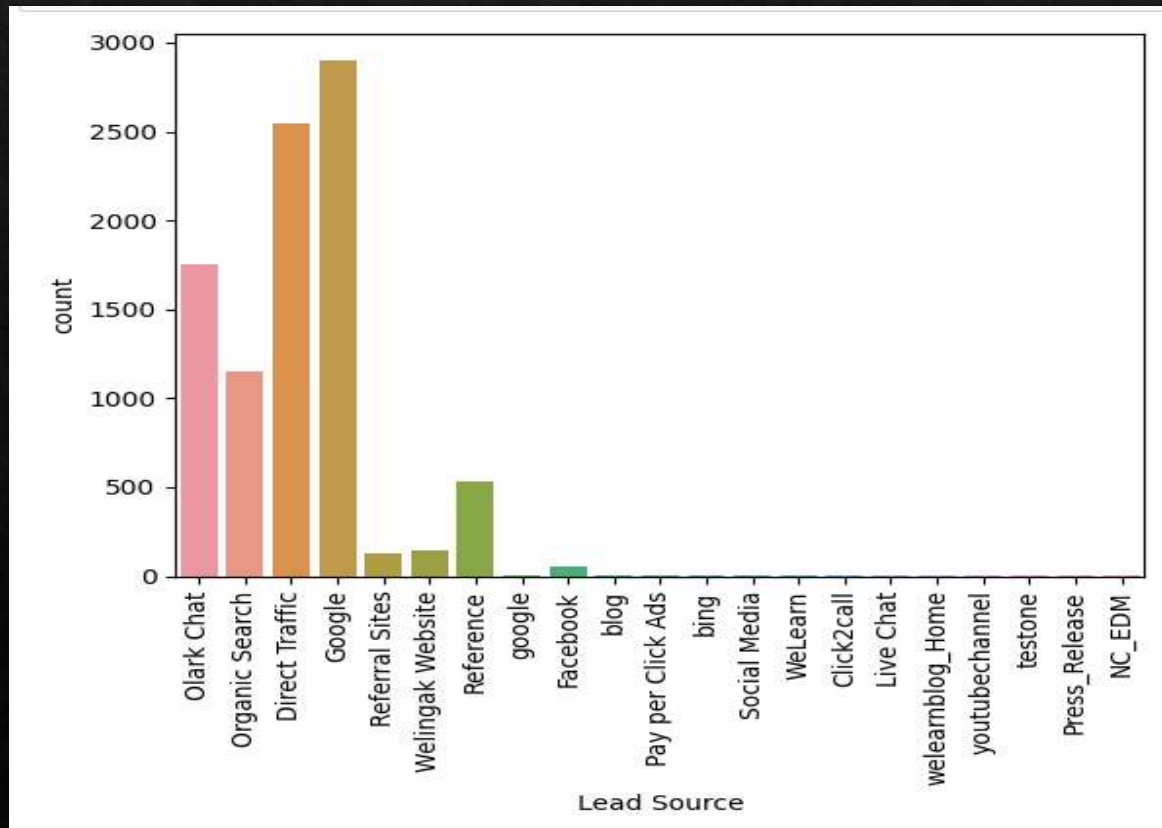
- ❖ Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

# Steps performed for the analysis:

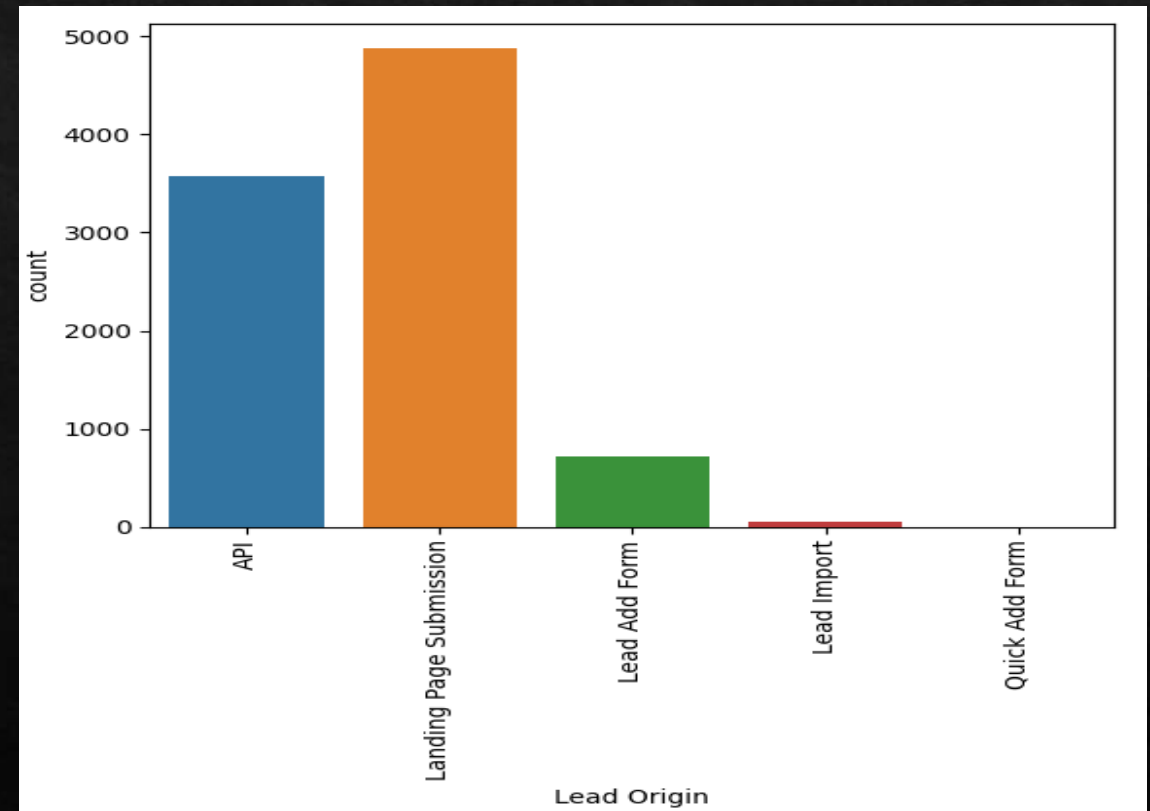
- ▣ Importing the required libraries and the dataset
- ▣ Inspecting the Data Frame
- ▣ Checking for missing values
- ▣ Analyzing columns individually and handling missing values
- ▣ Univariate Analysis and outlier treatment
- ▣ Dropping redundant columns
- ▣ Converting some binary variables (Yes/No) to 0/1
- ▣ Grouping column features
- ▣ Dummy creation
- ▣ Train - Test Split
- ▣ Feature Scaling
- ▣ Model Building
- ▣ Feature Selection Using RFE
- ▣ Plotting the ROC Curve
- ▣ Finding Optimal Cut-Off Points
- ▣ Precision and Recall
- ▣ Making Predictions on the Test Set

# EDA

## LEAD SCORE

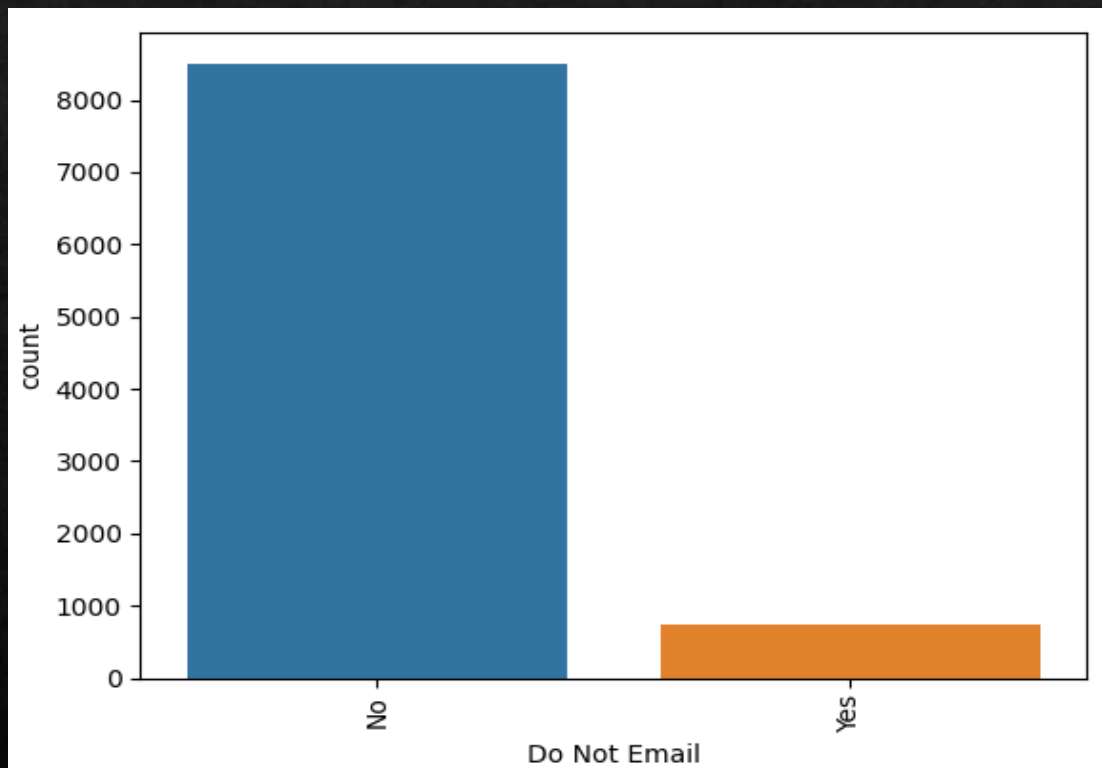


## LEAD ORIGIN

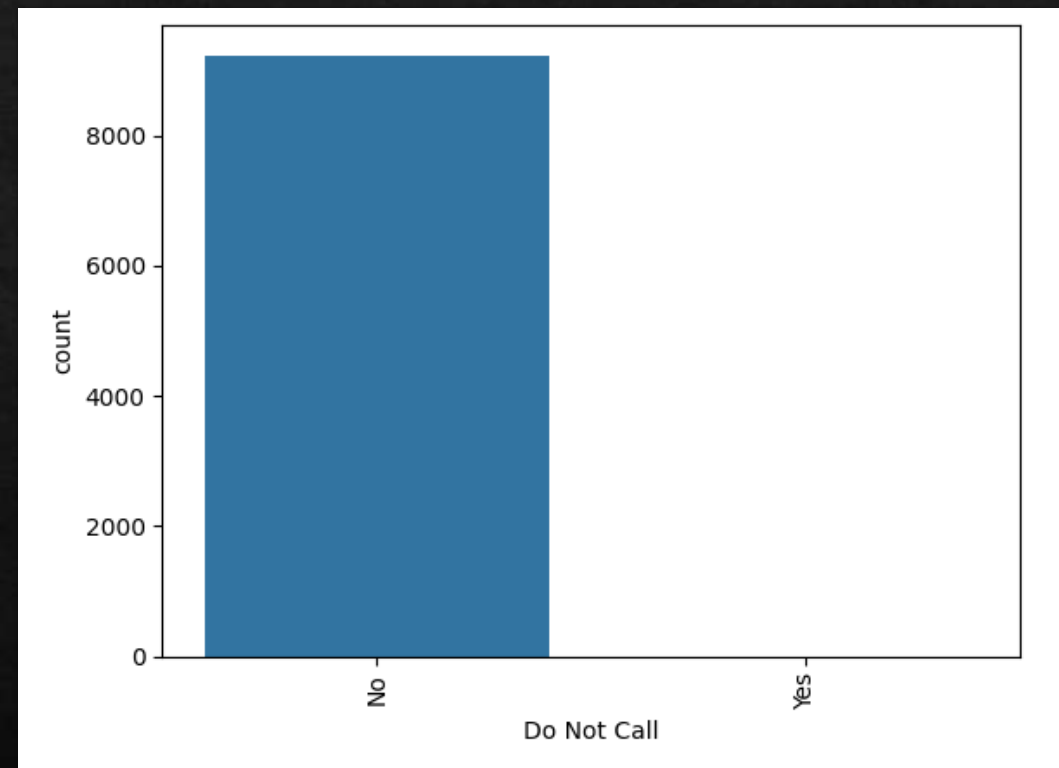




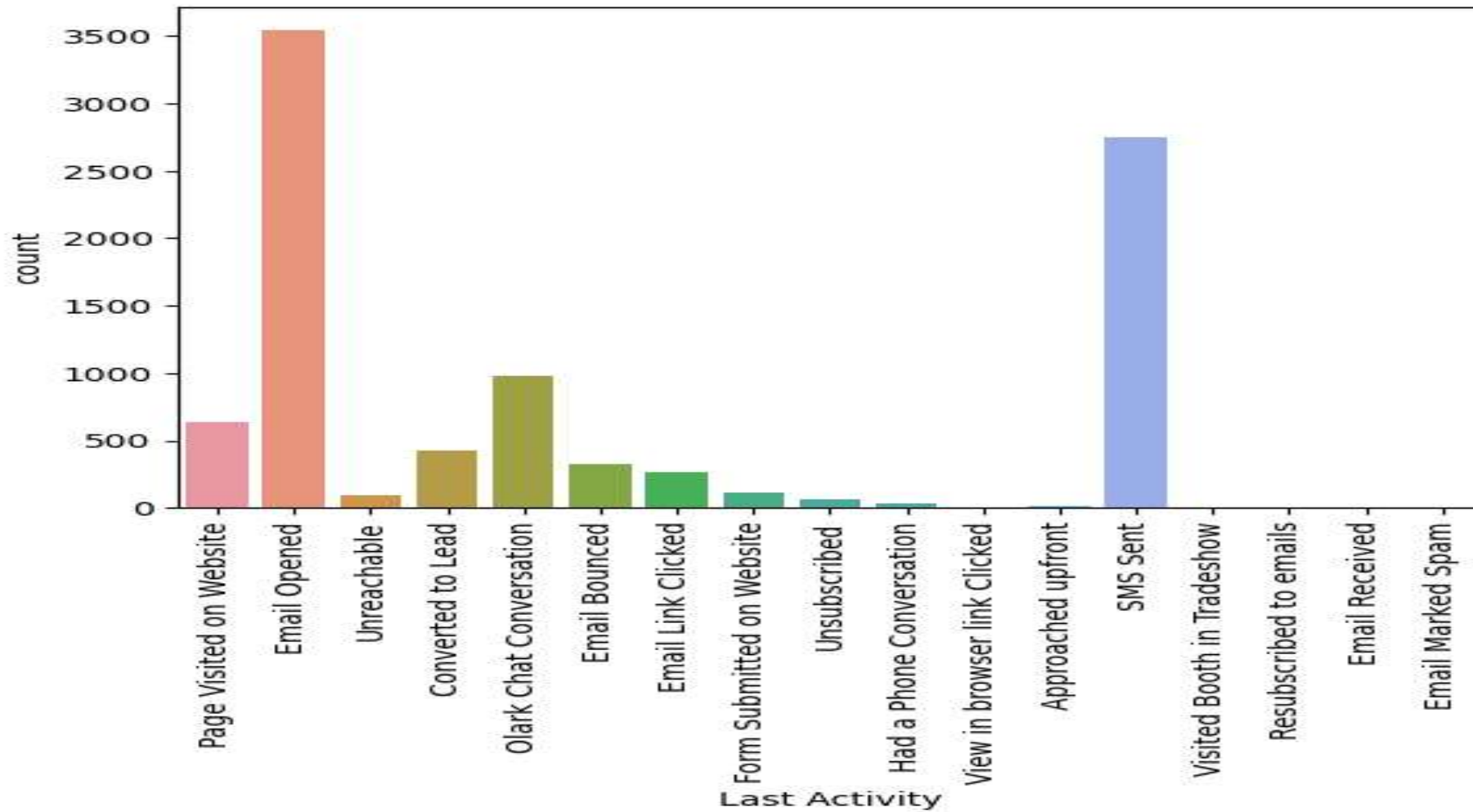
DO NOT EMAIL



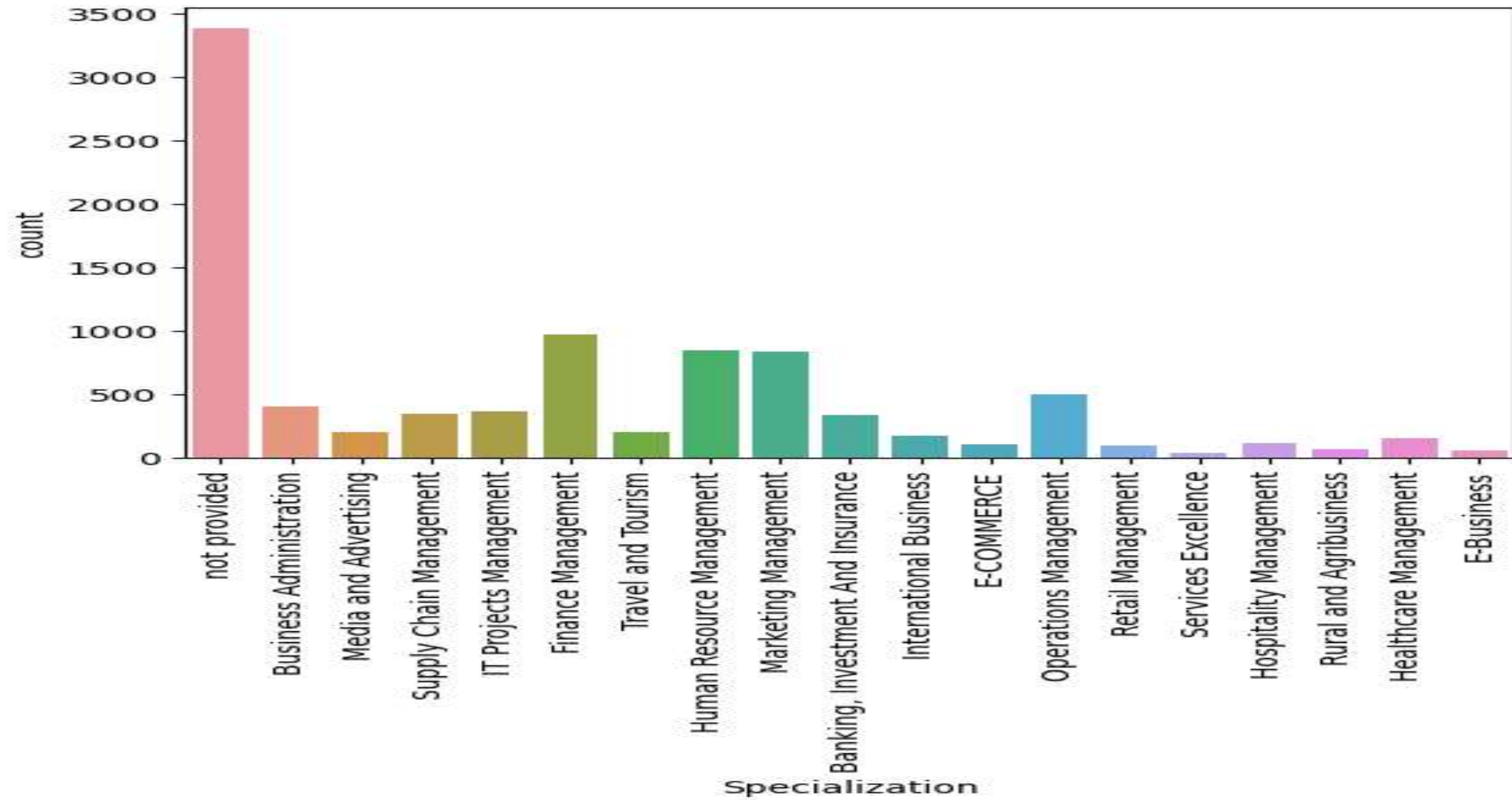
DO NOT CALL



## LAST ACTIVITY

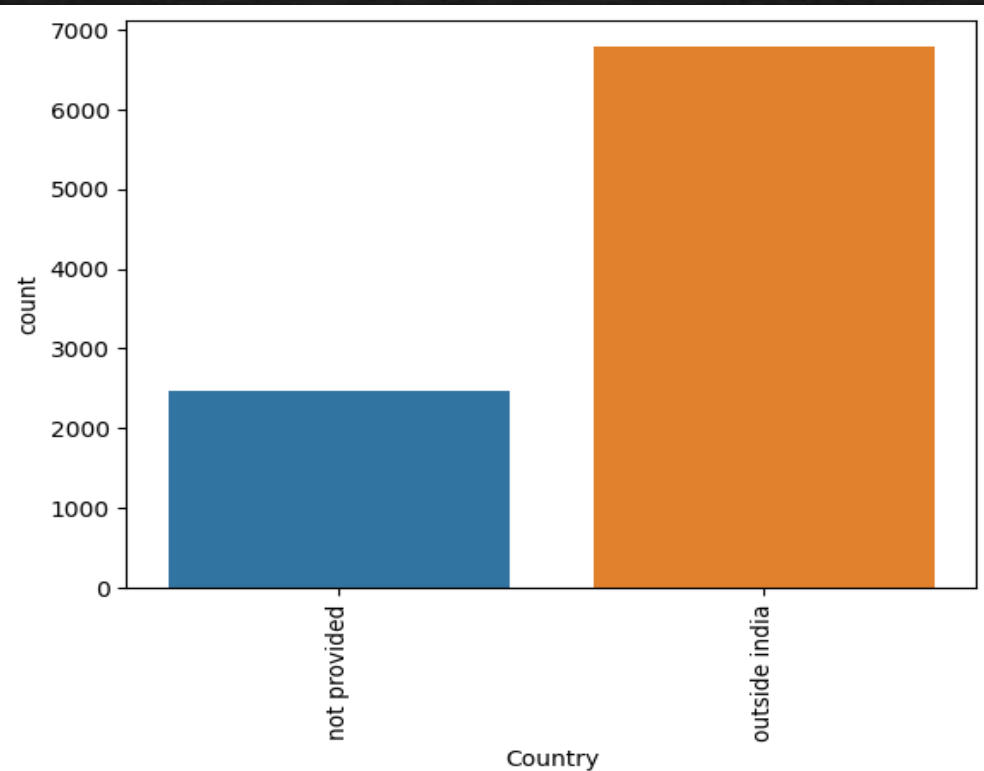


## SPECIALIZATION

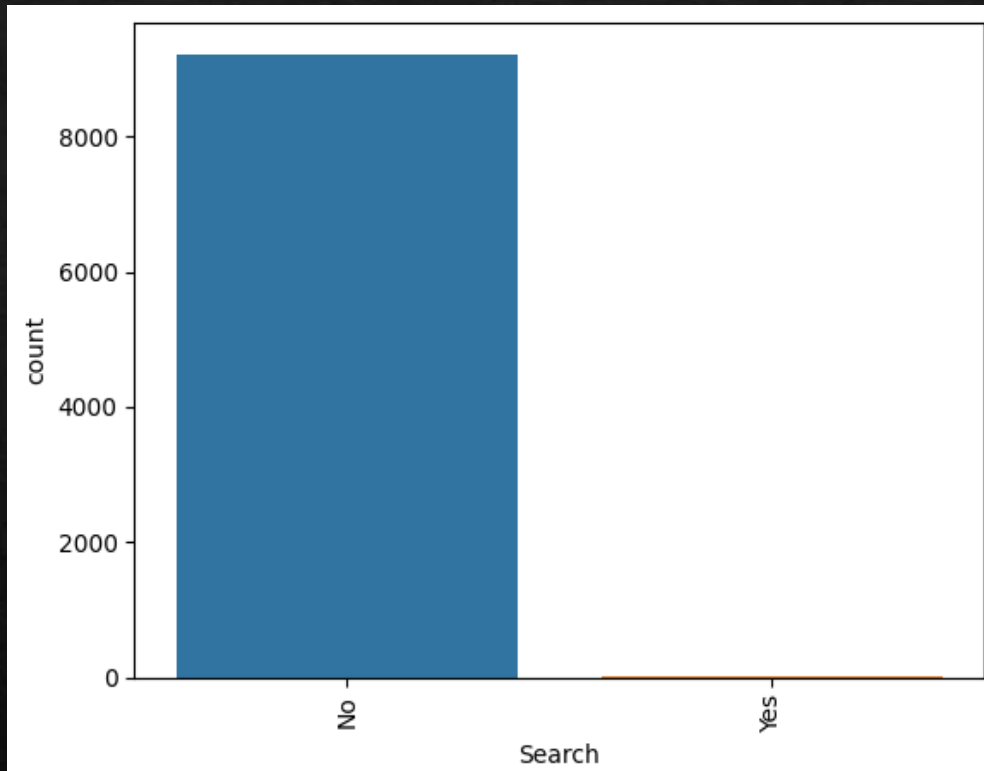




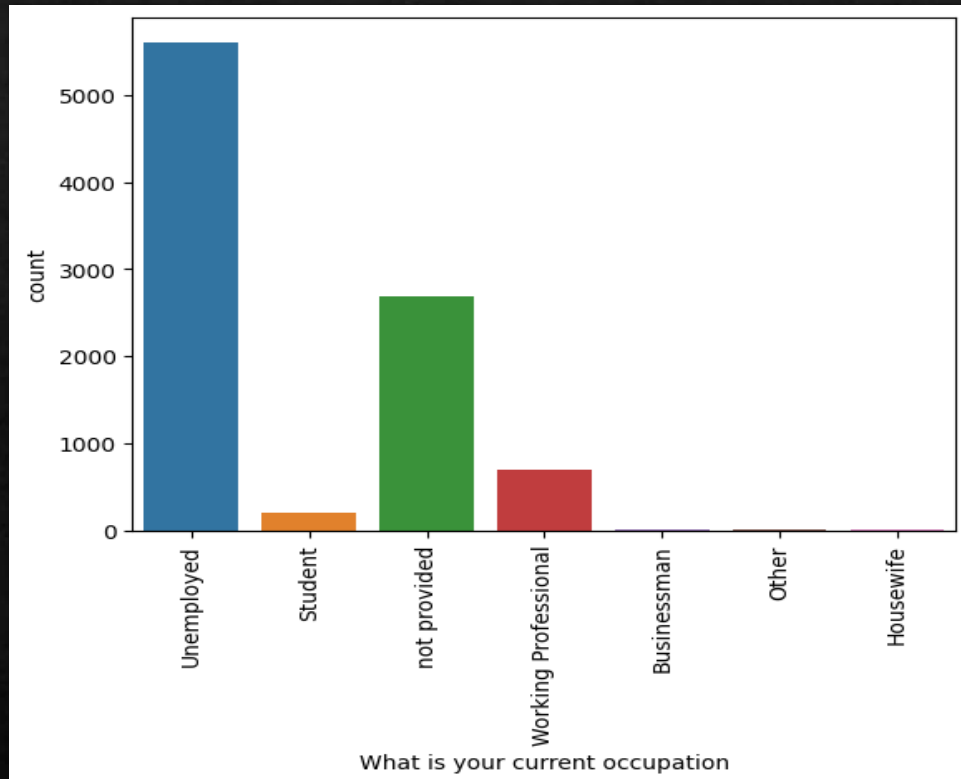
COUNTRY



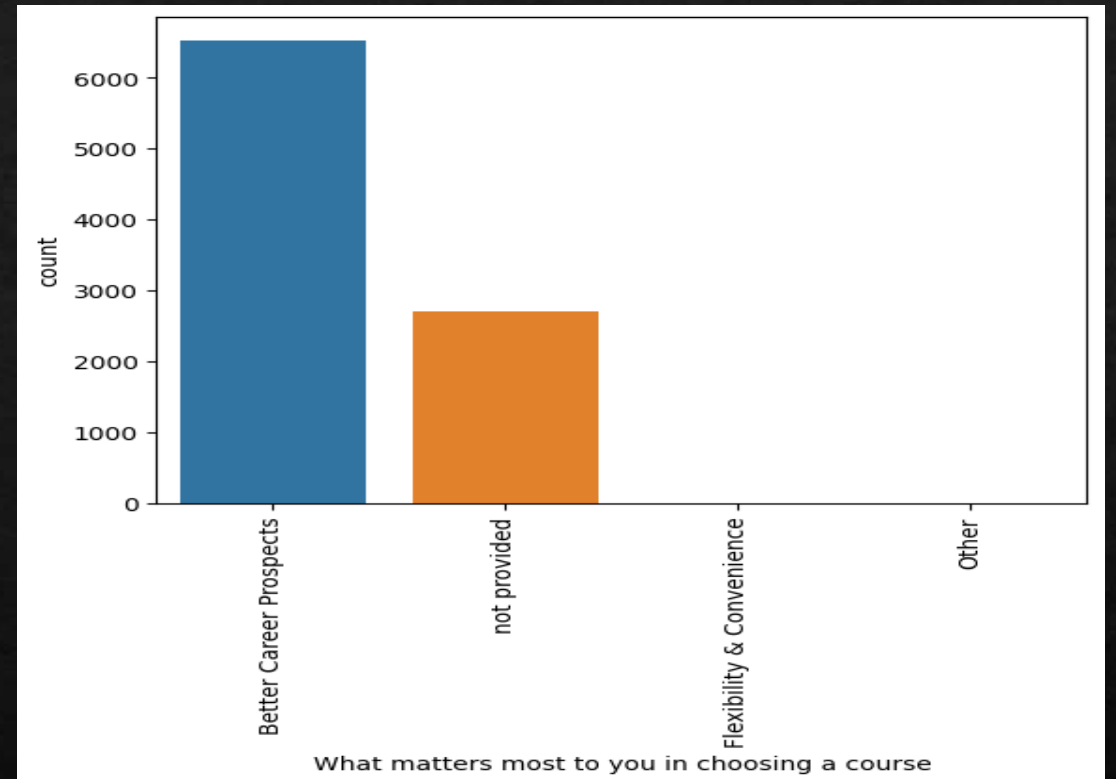
SEARCH



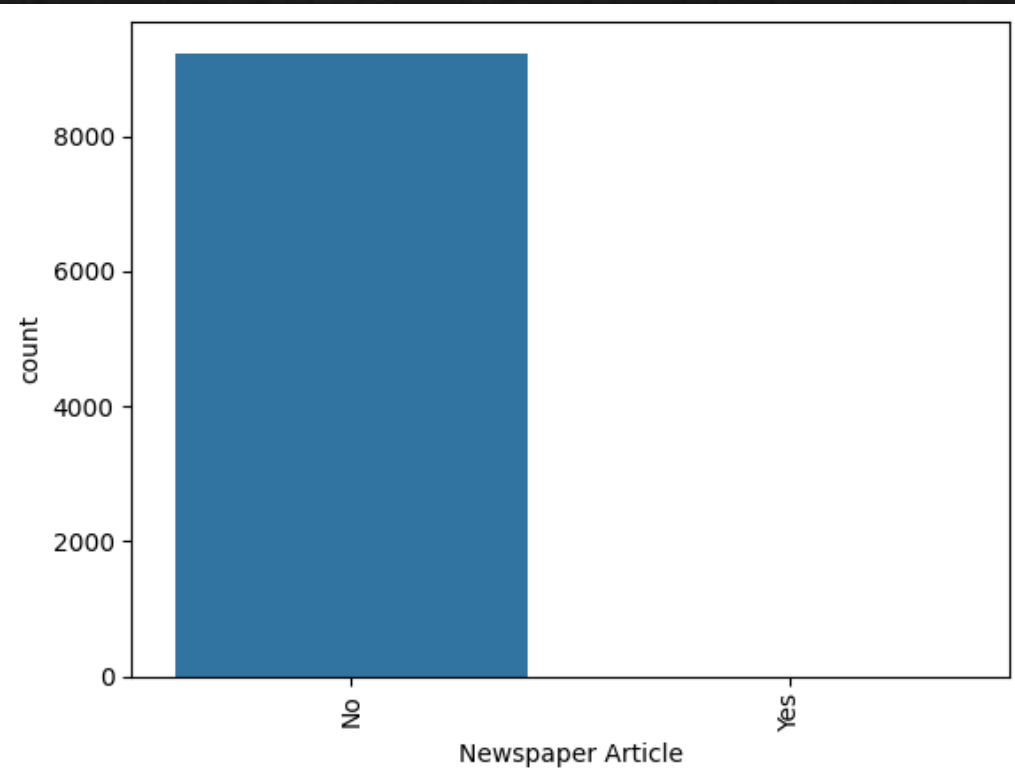
## CURRENT OCCUPATION



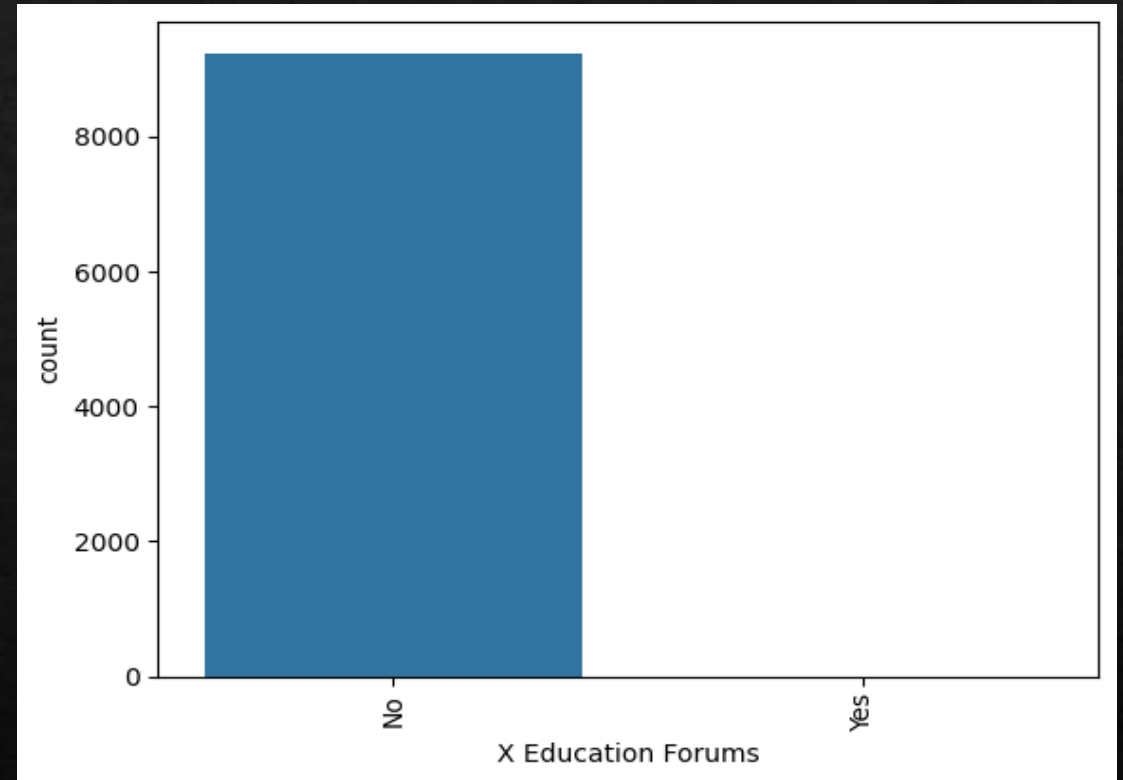
## WHAT MATTERS THE MOST IN CHOOSING A COURSE ?



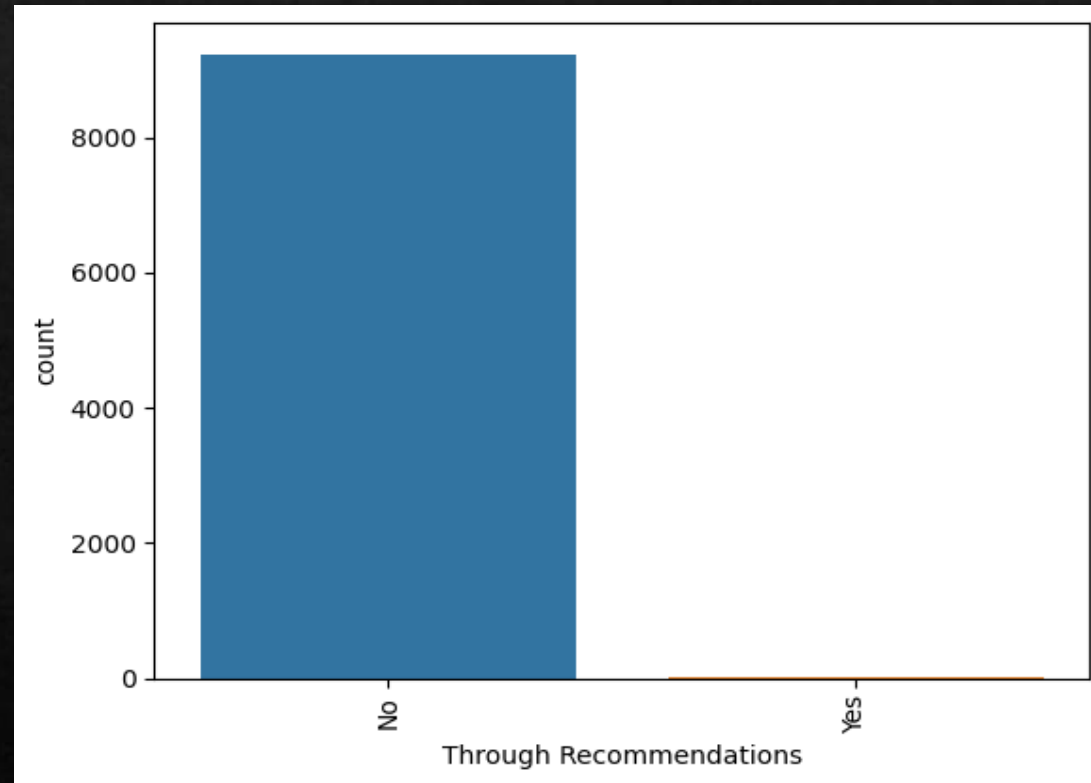
## NEWSPAPER ARTICLE



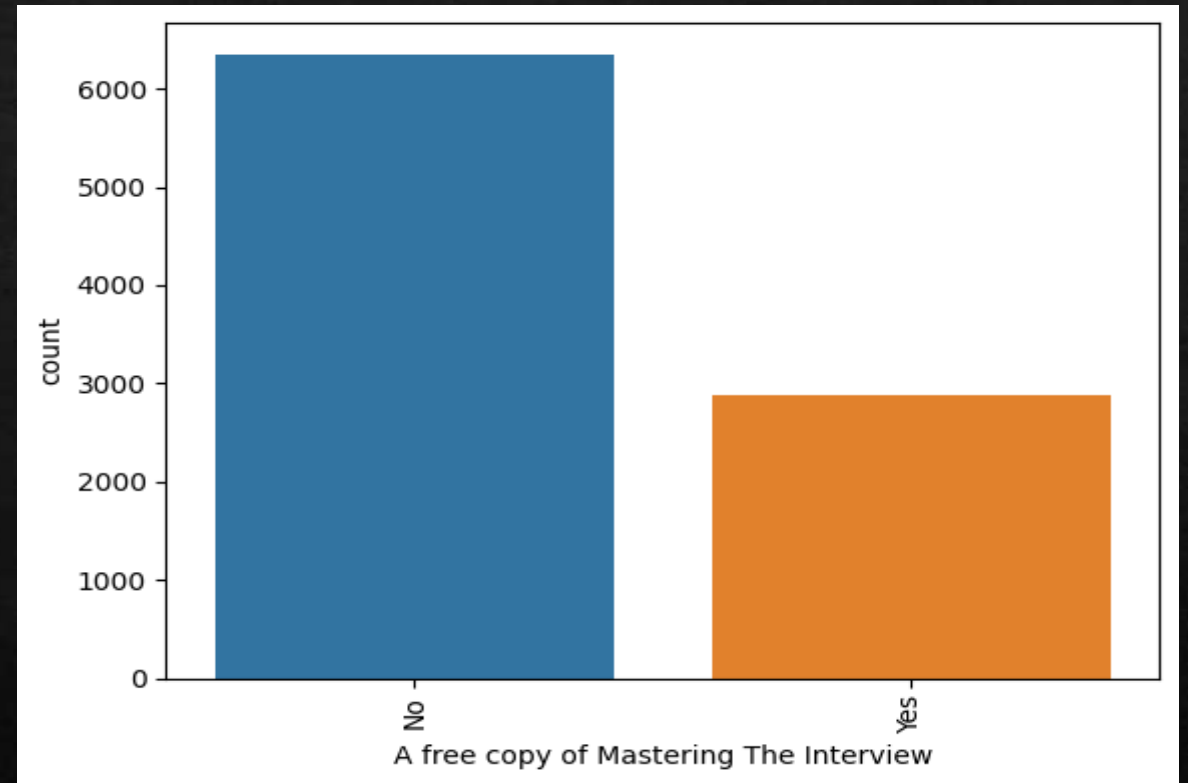
## X EDUCATION FORUMS



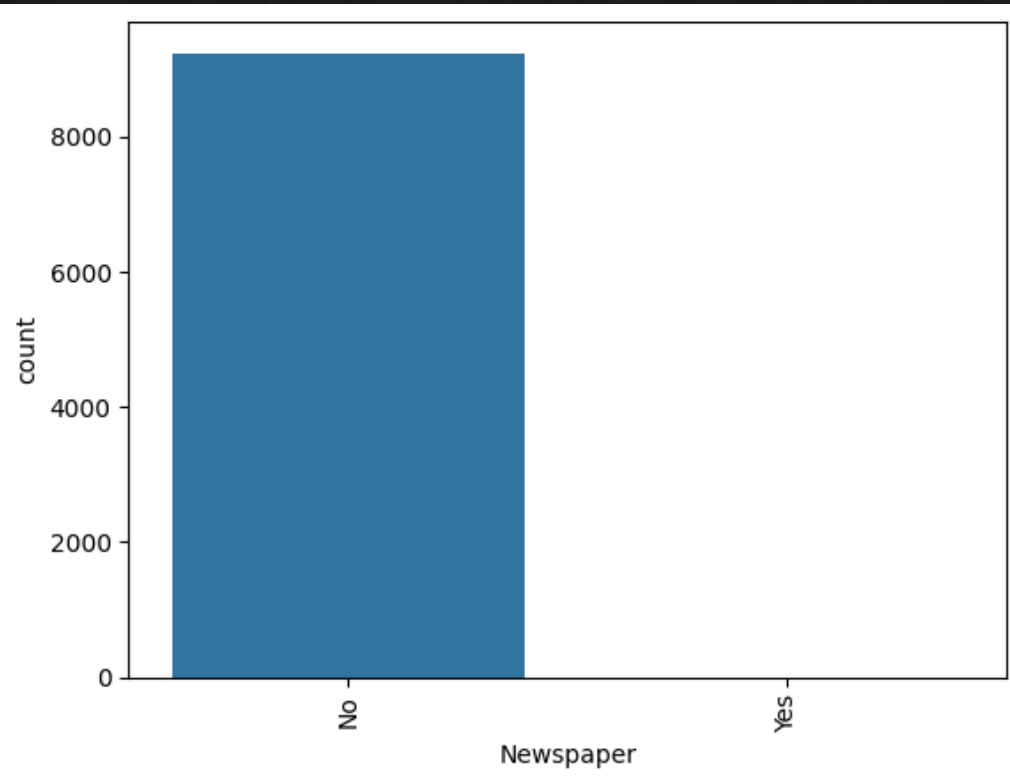
## THROUGH RECOMMENDATIONS



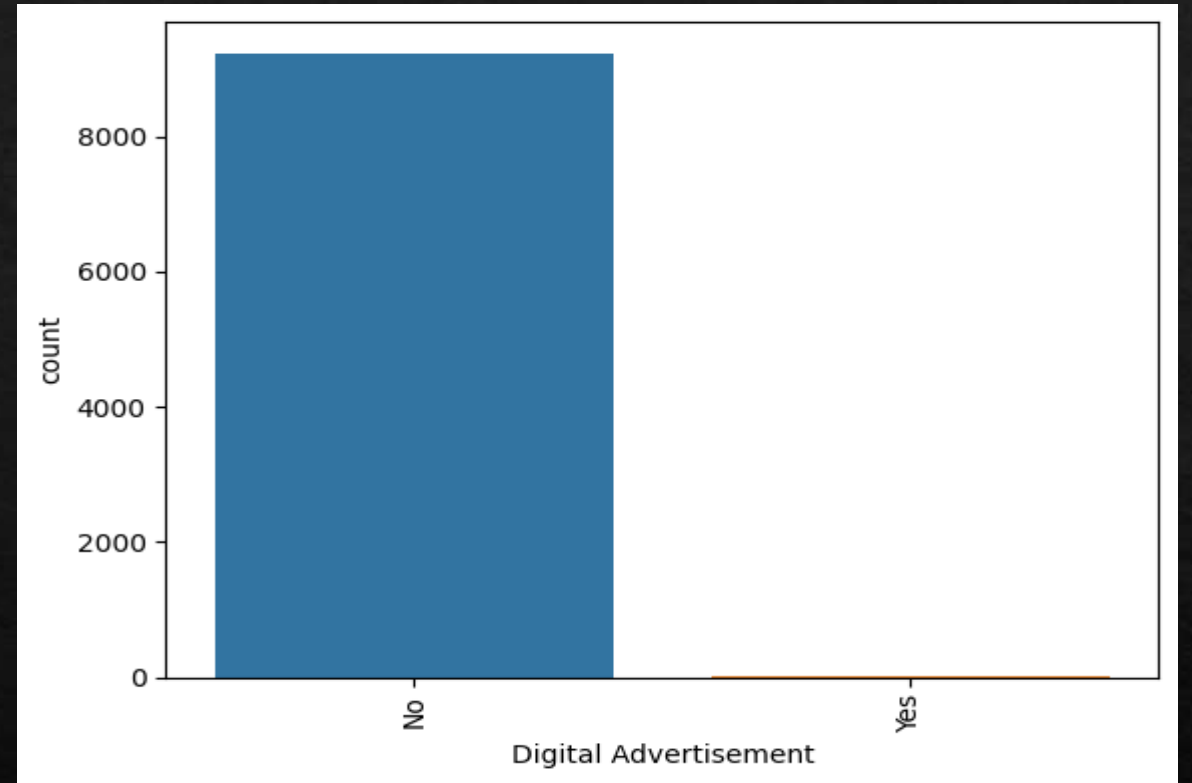
## A FREE COPY OF MASTERING THE INTERVIEW



## NEWSPAPER

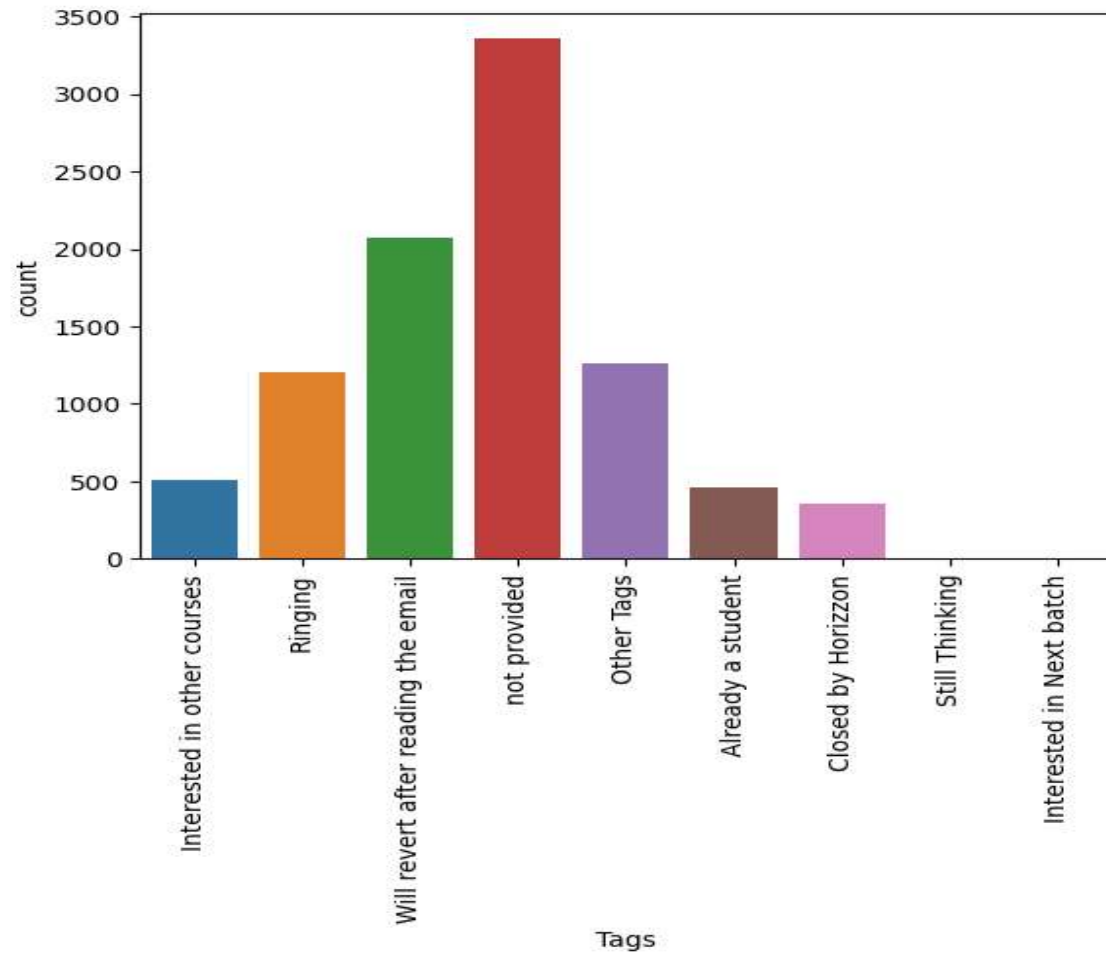


## DIGITAL ADVERTISEMENT

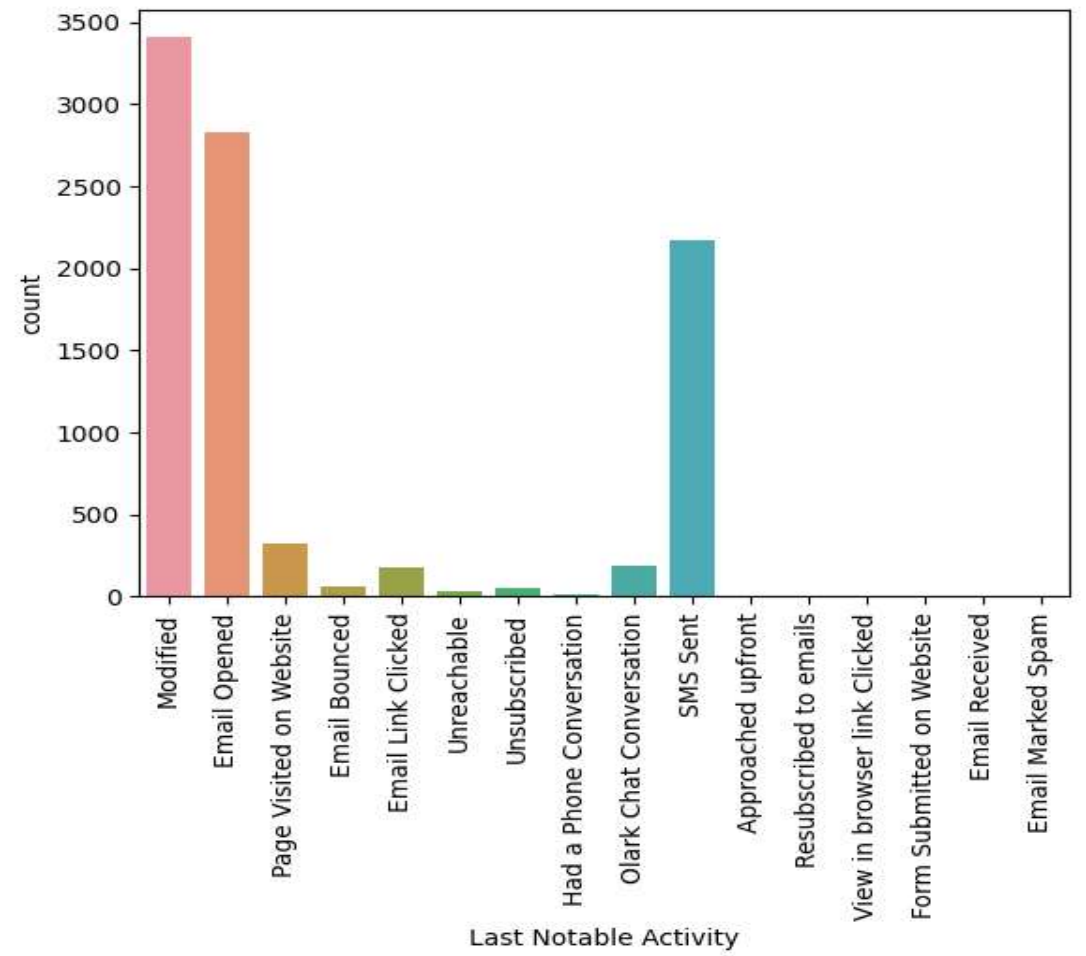




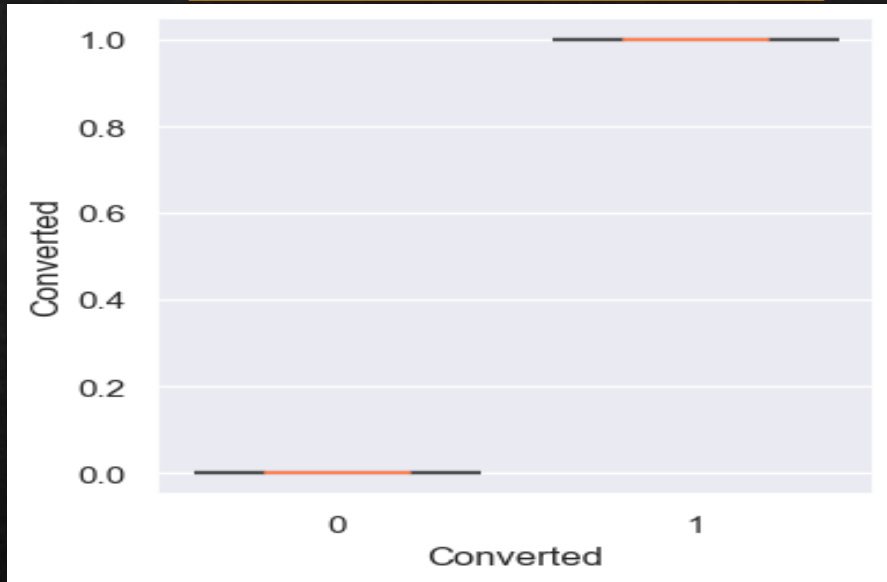
## TAGS



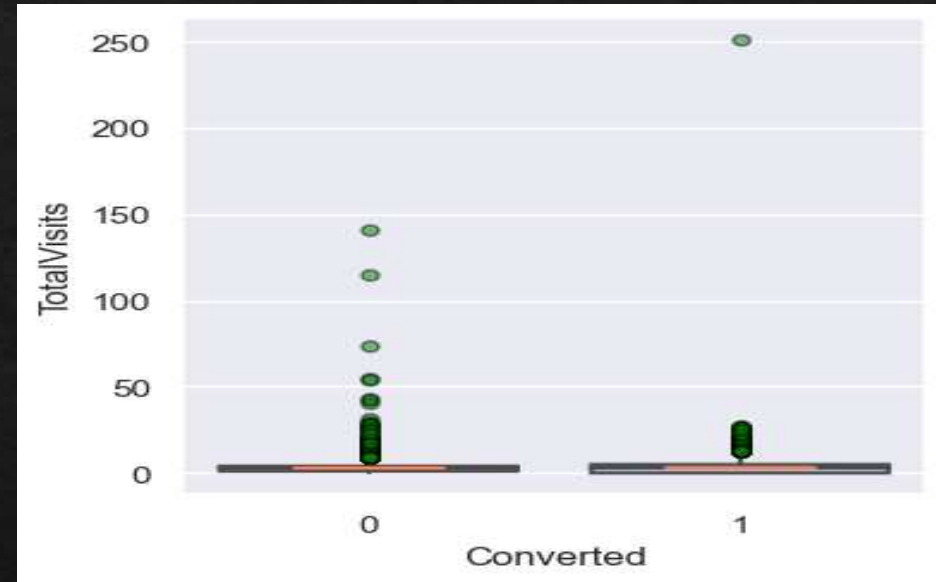
## LAST NOTABLE ACTIVITY



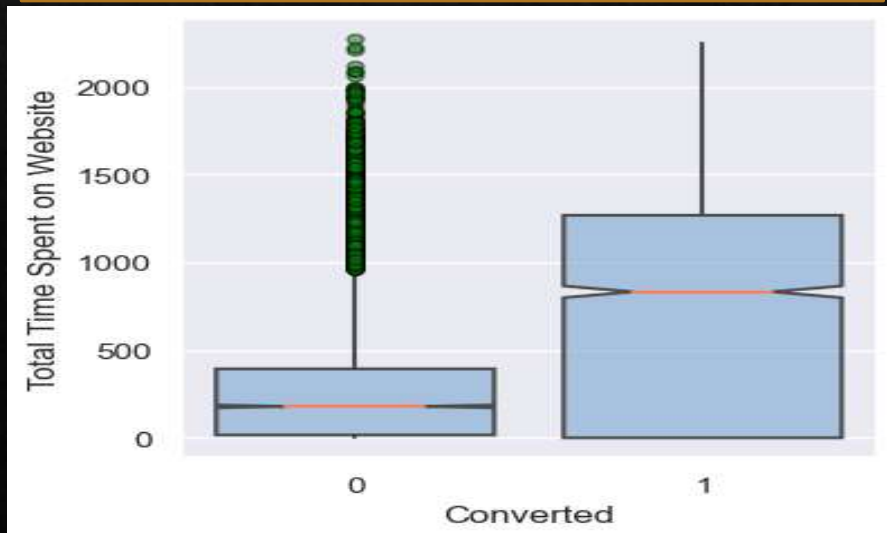
Search vs Converted



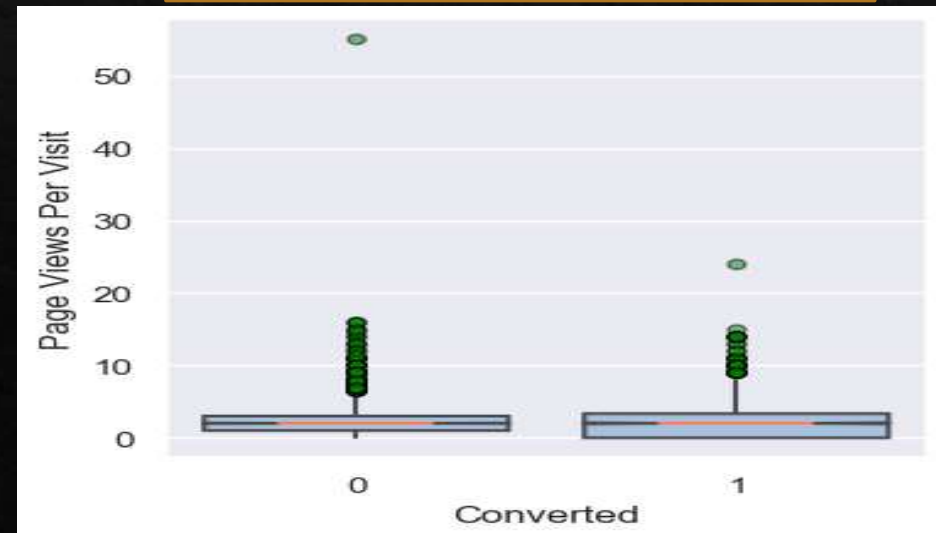
Total\_views vs converted



Total\_Time\_spent on website vs Converted



Page Views per Visit vs Converted



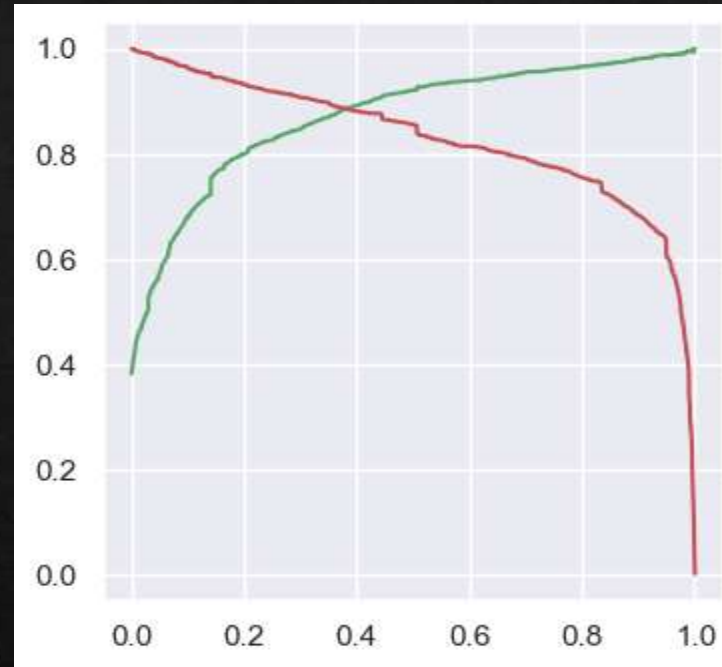
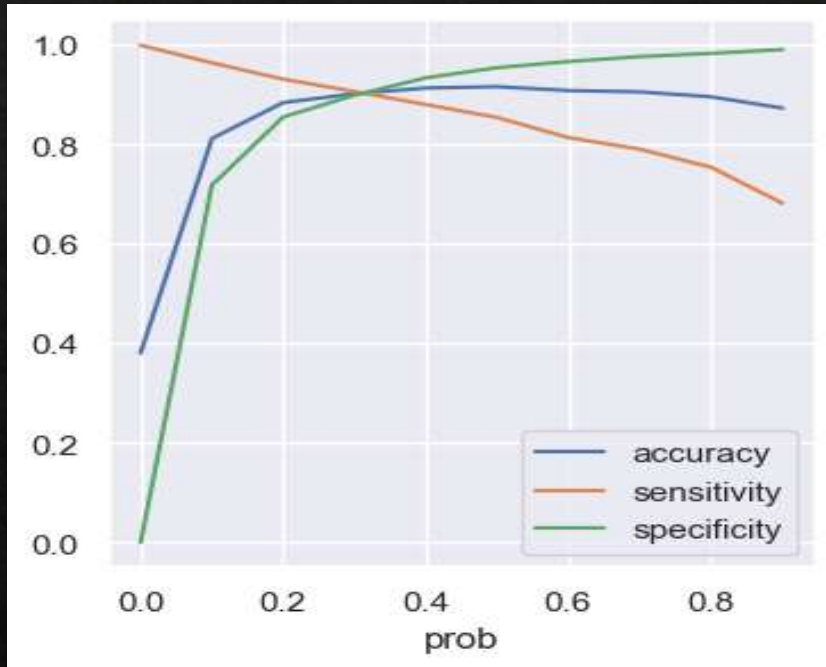
# DATA CONVERSION:

- ◇ Numerical variables are normalised.
- ◇ Dummy variables are created for objective type variables.
- ◇ Total Rows for Analysis = 9240
- ◇ Total columns for Analysis = 37

# 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 70:30 ratio.
- ◆ Use RFE for feature selection.
- ◆ Running RFE with 20 variables as output.
- ◆ Building Model by removing the variables whose p-value is greater than 0.05 and vif value is greater than 5.
- ◆ Predictions on test data set.
- ◆ Overall accuracy is 91%.

# ROC CURVE



- ◇ Finding Optimal Cut off point.
- ◇ Optimal cut off probability is that where we get balanced sensitivity and specificity.
- ◇ From the first graph it is visible that the optimal cut off is at 0.35.



# Conclusion:

- ◇ It was found that the variables that mattered the most in the potential buyers are (In descending order) :
  1. What\_matters\_most\_to\_you\_in\_choosing\_a\_course
  2. Tags\_Will revert after reading the email
  3. Last\_Notable\_Activity\_Modified
  4. a.Tags\_Other Tags  
b.Tags\_Ringing
  5. a. Last\_Activity\_SMS Sent  
b. Last\_Activity\_Olark Chat Conversation
  6. Occupation\_Working Professional
  7. Lead\_Origin\_Lead Import
- ◇ Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.