




# LEAD SCORE CASE STUDY

*BATCH:*  
**DS C67**

*GROUP MEMBERS:*

**YAMINI ANOOSHA VALLURU**  
**YASHWANT BABULAL SADH**  
**YASHWANT MOKKARALA**



# PROBLEM STATEMENT

An X Education is seeking assistance in identifying the most promising leads, meaning those that are most likely to become paying customers. The company has asked us to create a model that assigns a lead score to each lead, ensuring that those with higher scores have a greater chance of conversion, while those with lower scores have a reduced chance. Specifically, the CEO has provided an estimate for the target lead conversion rate, aiming for around 80%.

## GOALS AND OBJECTIVES

- This case study has several objectives.
- Create a logistic regression model to assign a lead score ranging from 0 to 100 for each lead, which the company can utilize to focus on potential candidates. A higher score indicates that the lead is hot, meaning it is more likely to convert, while a lower score suggests that the lead is cold and unlikely to convert.
- The company has additional challenges that your model should adapt to if their needs evolve in the future, so be prepared to address these as well. These challenges are included in a separate document. Please complete this based on the logistic regression model you developed initially. Additionally, ensure that you incorporate this into your final presentation where you'll provide recommendations.

# METHODOLOGY:

- ❑ STEP 1: DATA READING AND CLEANING
  - *EXAMINE THE DATAFRAME.*
  - *CLEANING THE DATAFRAME*
- ❑ STEP:2 EXPLORATORY DATA ANALYSIS
  - *UNIVARIATE ANALYSIS*
    - CATEGORICAL VARIABLES
    - NUMERICAL VARIABLES
    - LINKING ALL THE CATEGORICAL VARIABLES TO CONVERTED
- ❑ STEP:3 DUMMY VARIABLES
- ❑ STEP:4 TEST - TRAIN SPLIT
- ❑ STEP:5 MODEL BUILDING
- ❑ STEP:6 CREATING PREDICTION
- ❑ STEP:7 MODEL EVALUATION
  - *OPTIMISE CUTT OFF (ROC CURVE)*
- ❑ STEP:8 PREDICTION ON TEST SET
- ❑ STEP:9 PRECISION-RECALL
  - *PRECISION AND RECALL TRADEOFF*
- ❑ STEP:10 PREDICTION ON TEST SET

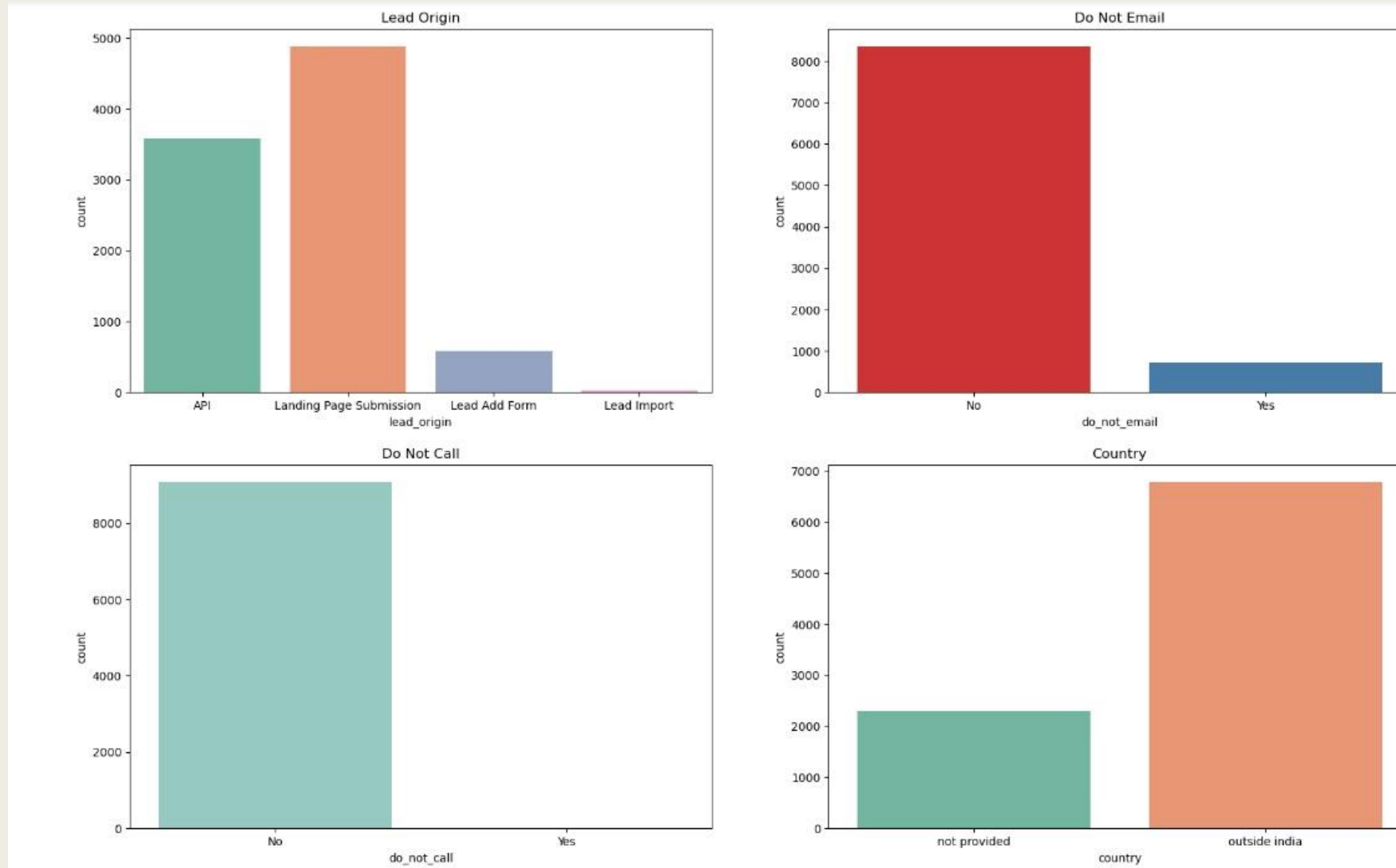
# DATA READING AND CLEANING

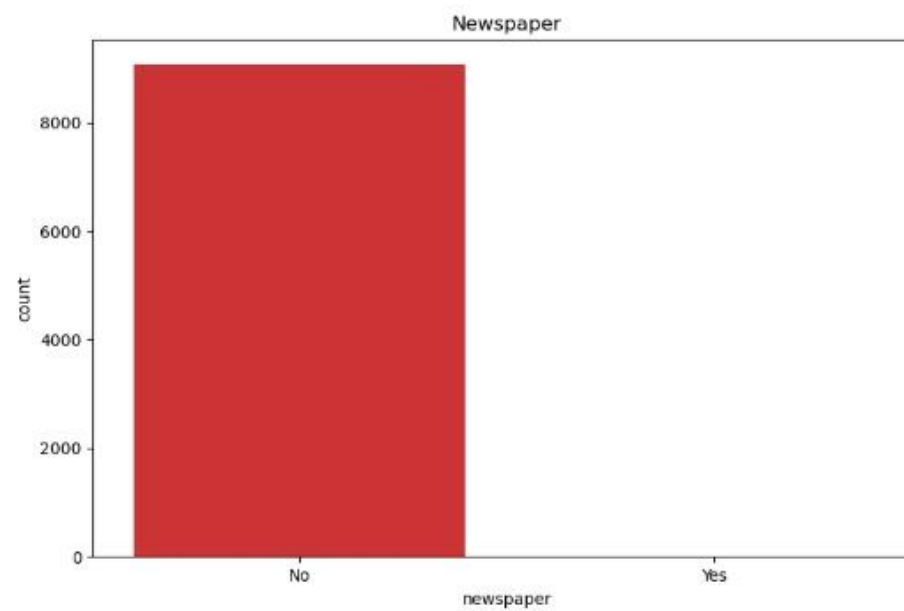
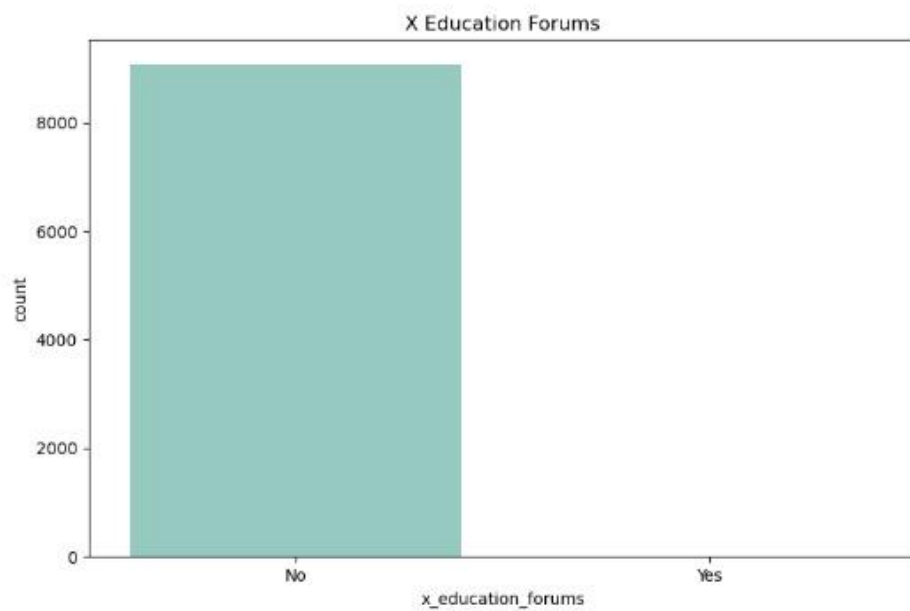
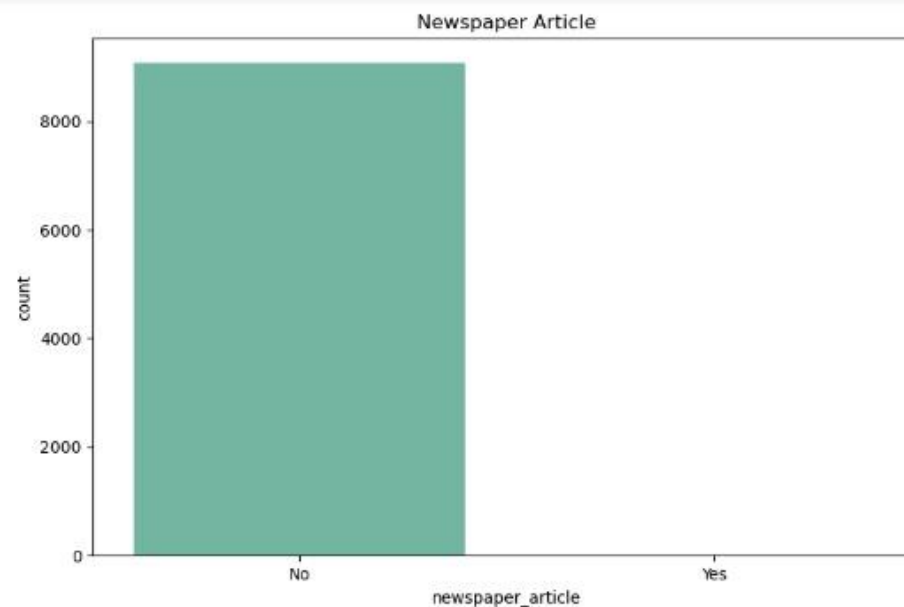
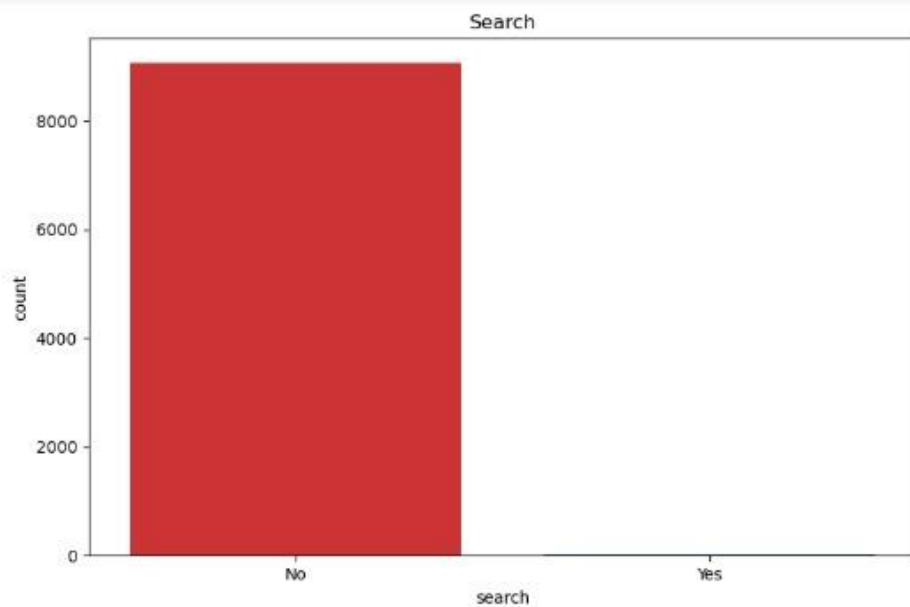
- Total Row Count = 37, Total Column Count = 9240.
- Individual value attributes such as "magazine," "receive\_more\_updates\_about\_our\_courses," and "update\_me\_on\_supply\_chain\_content", "get\_updates\_on\_dm\_content", "i\_agree\_to\_pay\_the\_amount\_through\_cheque" etc. have been dropped.
- Eliminating the "prospect\_id" and "lead\_number" as they are not essential for the analysis.
- Upon reviewing the value counts for several object-type variables, we identified some features with insufficient variance, which we have decided to drop. The features removed include: "do\_not\_call," "what\_matters\_most\_to\_you\_in\_choosing\_course," "search," "newspaper\_article," "x\_education\_forums," "newspaper," and "digital\_advertisement," among others.
- Eliminating the columns with over 35% missing values, including 'how\_did\_you\_hear\_about\_x\_education' and 'lead\_profile'.

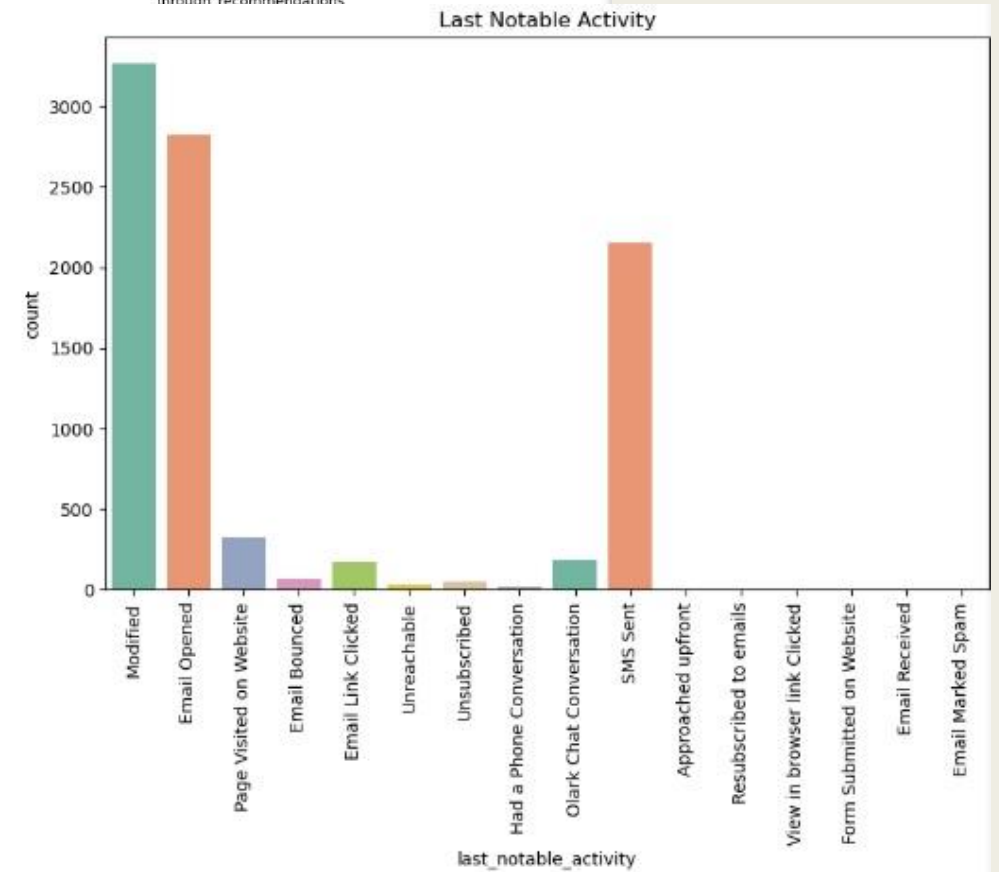
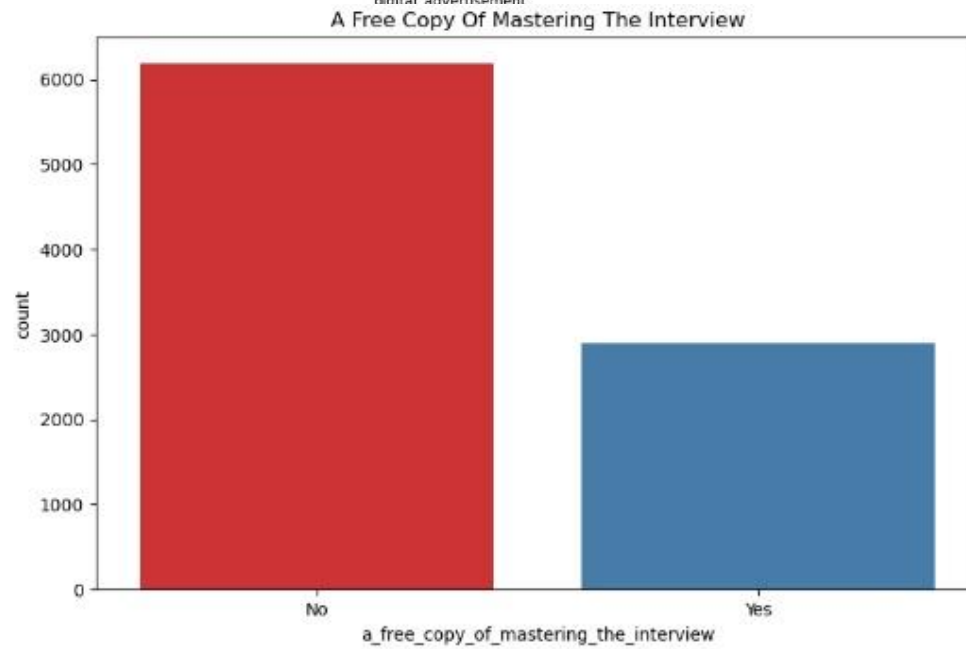
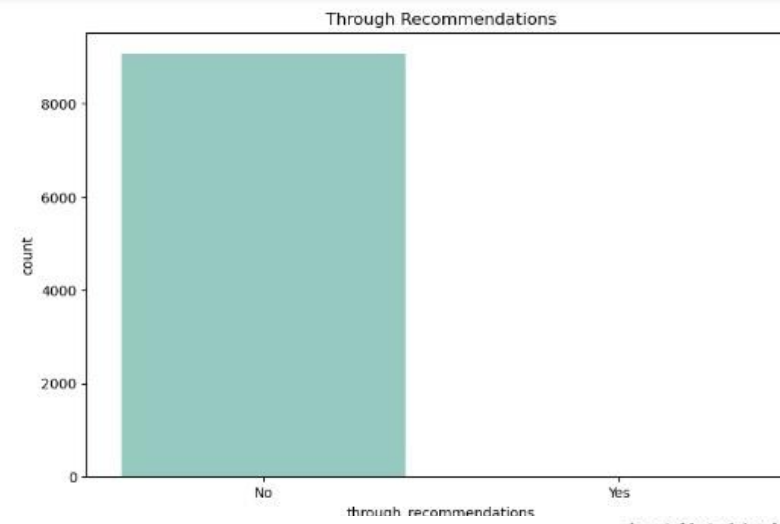
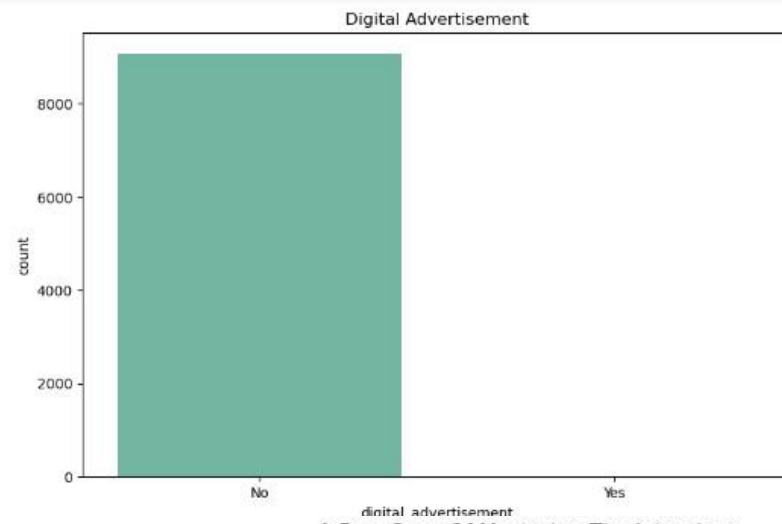
# EXPLORATORY DATA ANALYSIS

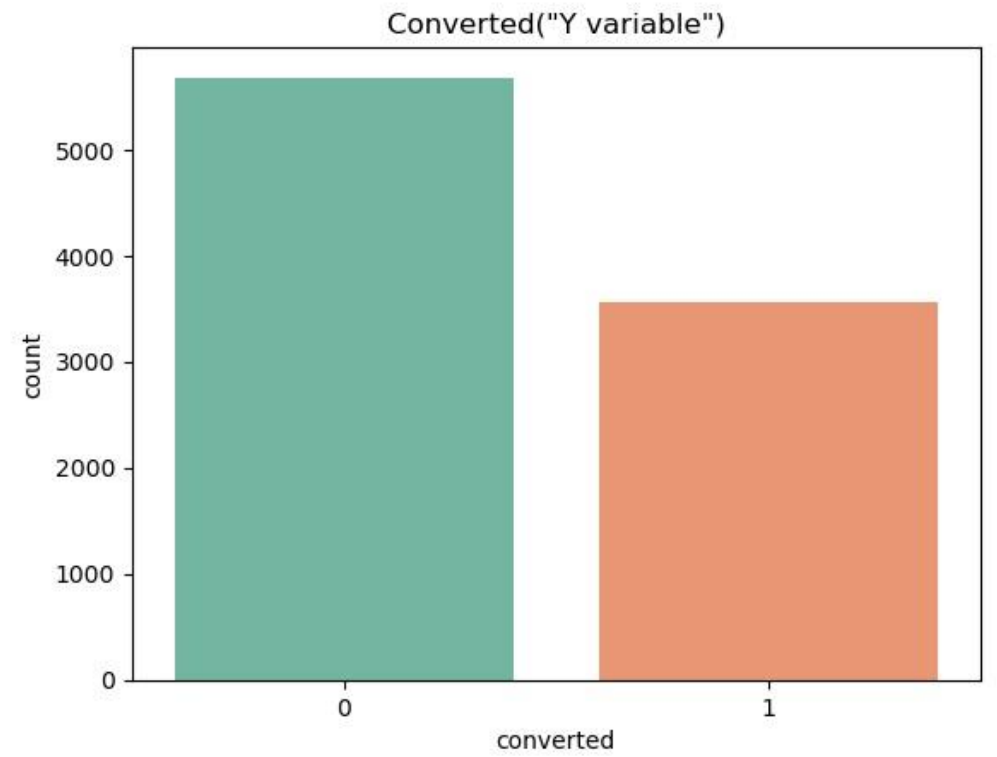
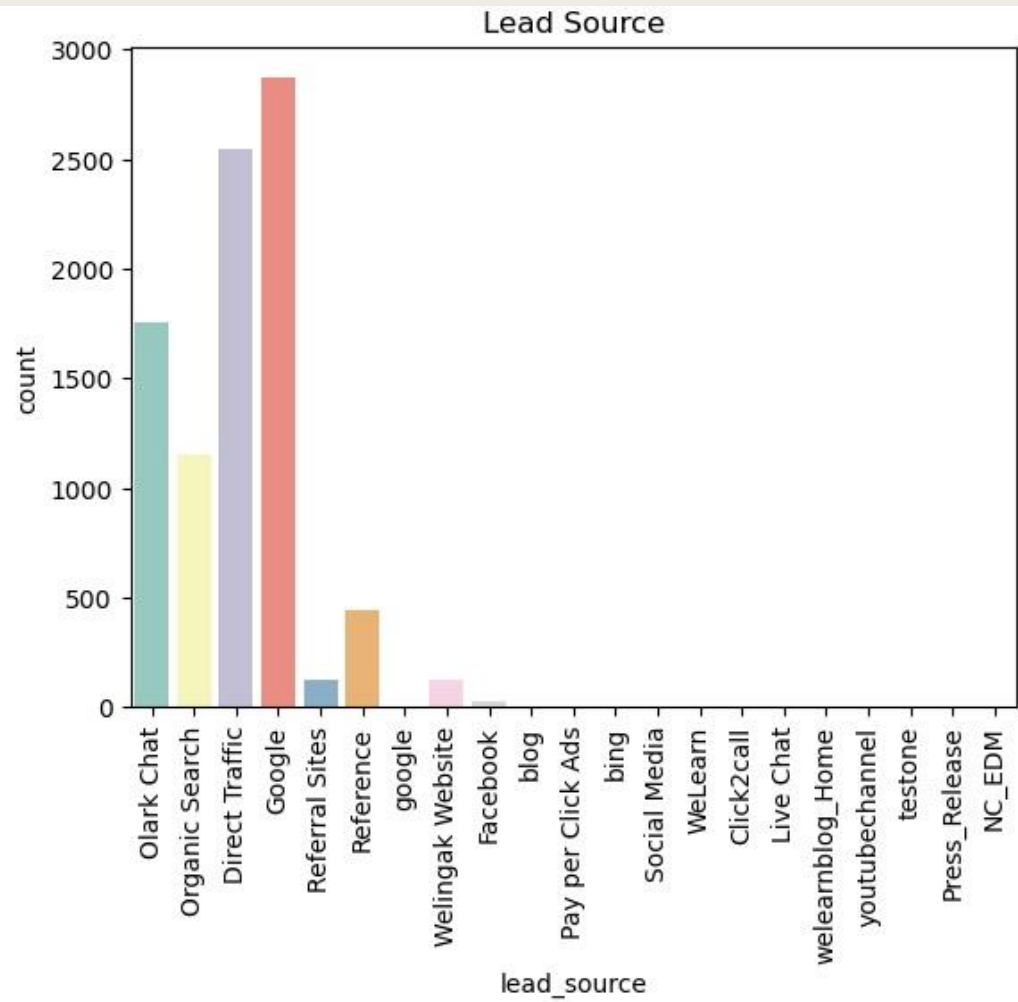
## UNIVARIATE ANALYSIS

### CATEGORICAL VARIABLE RELATION

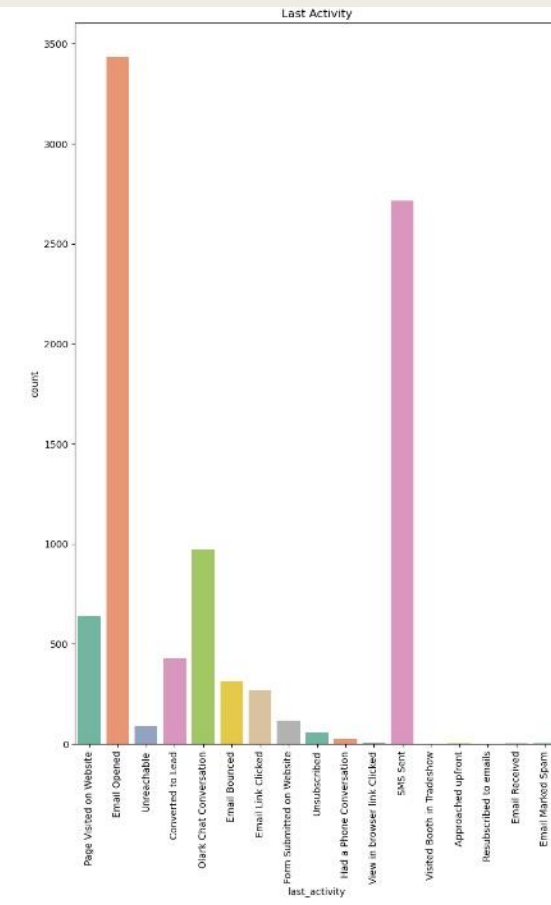
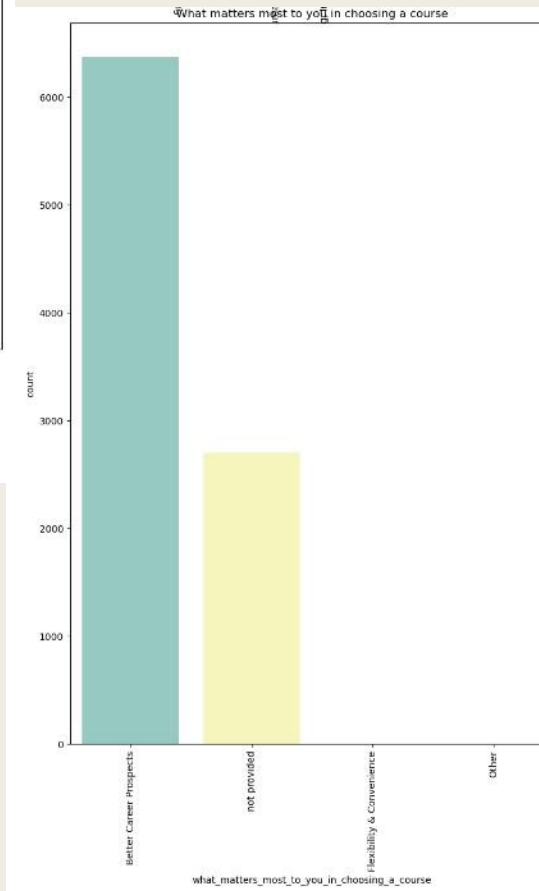
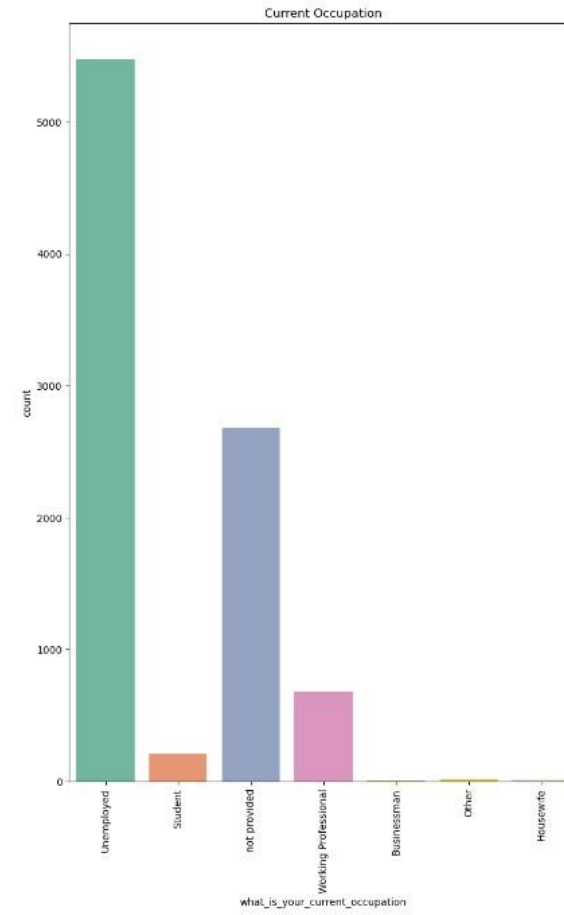
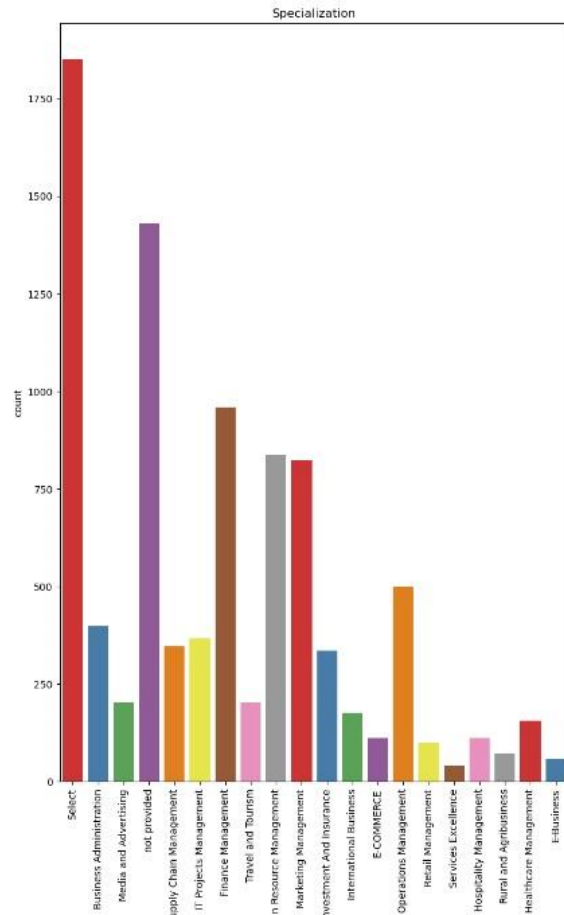




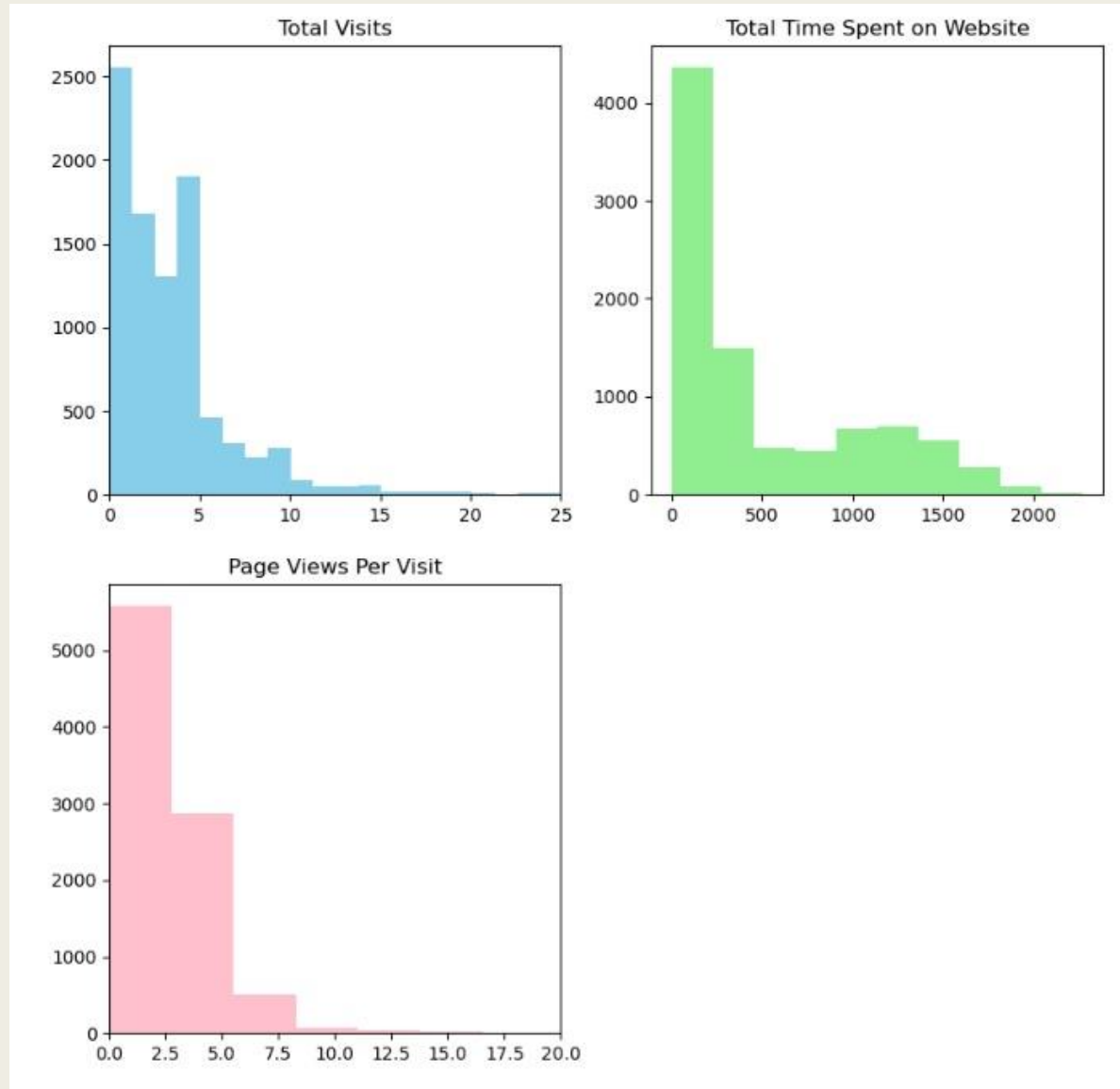




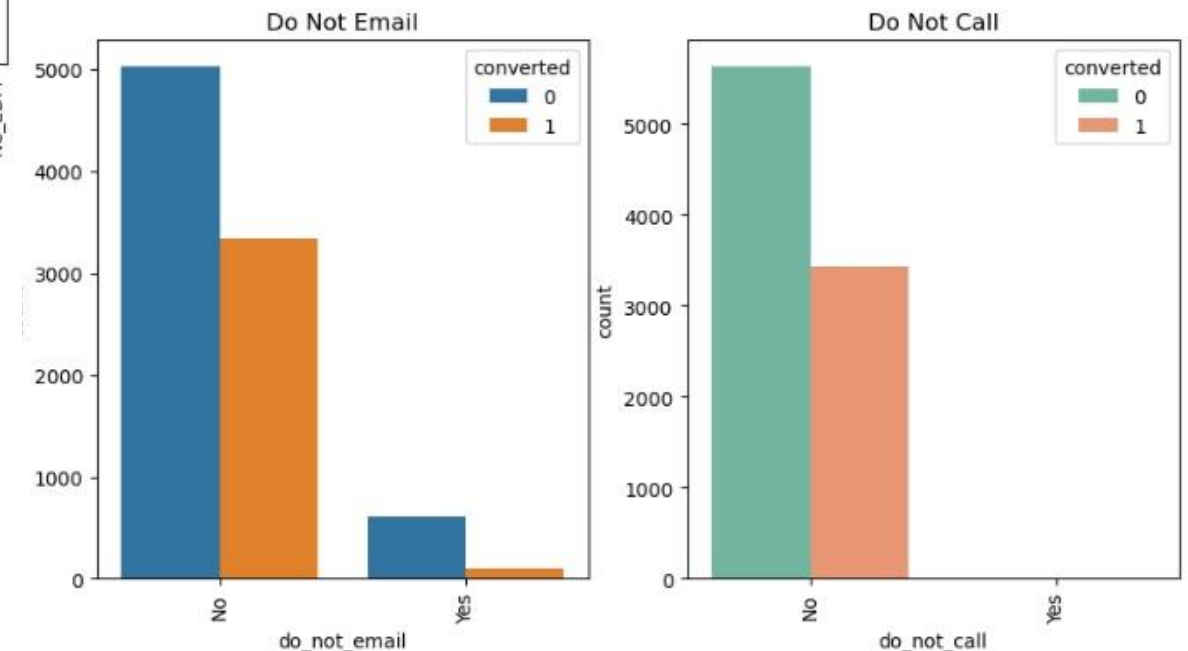
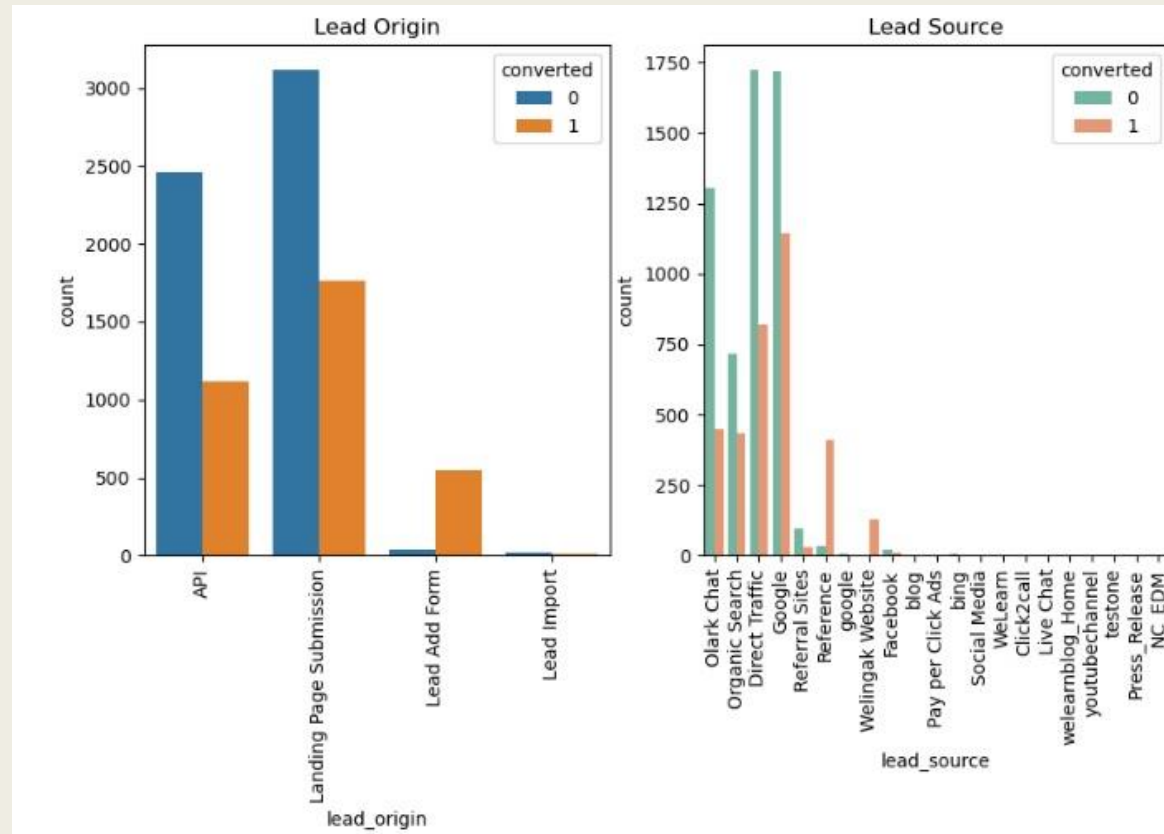


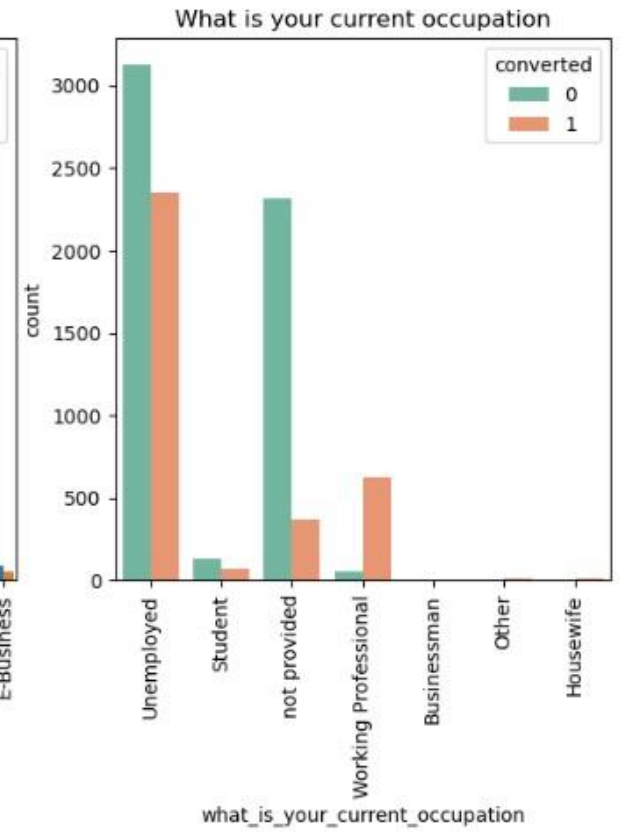
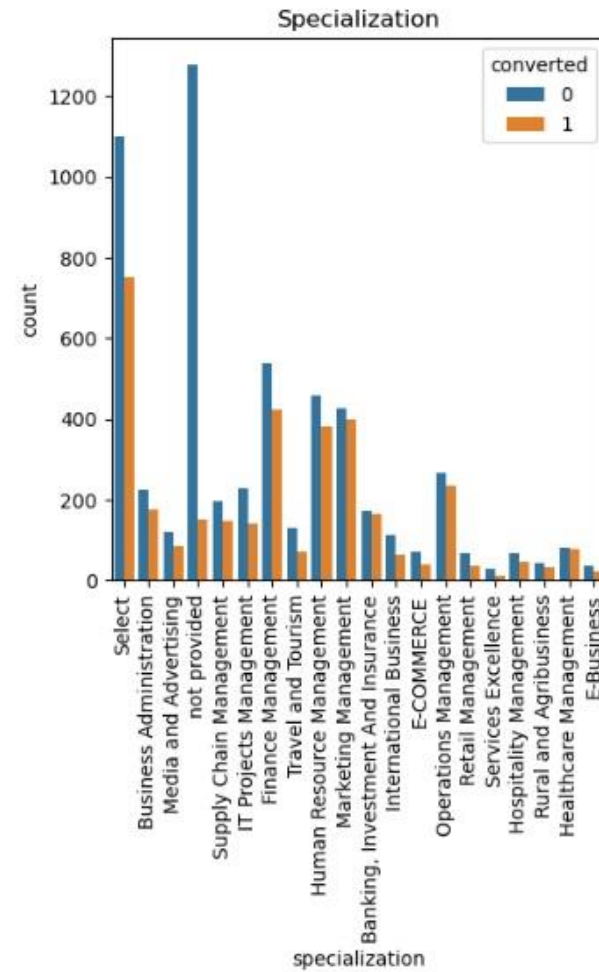
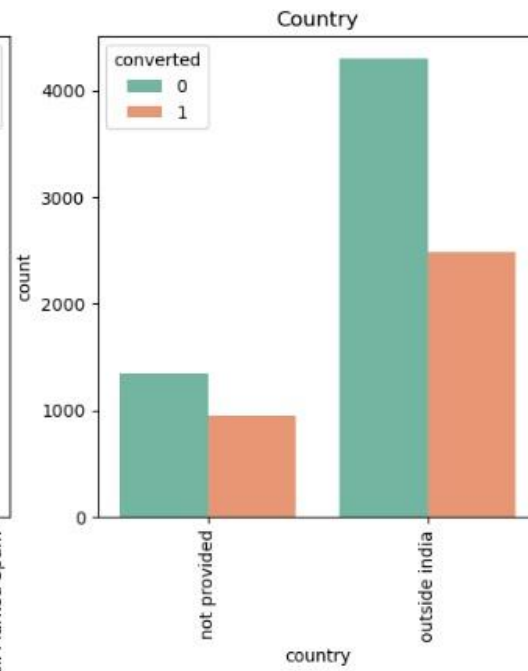
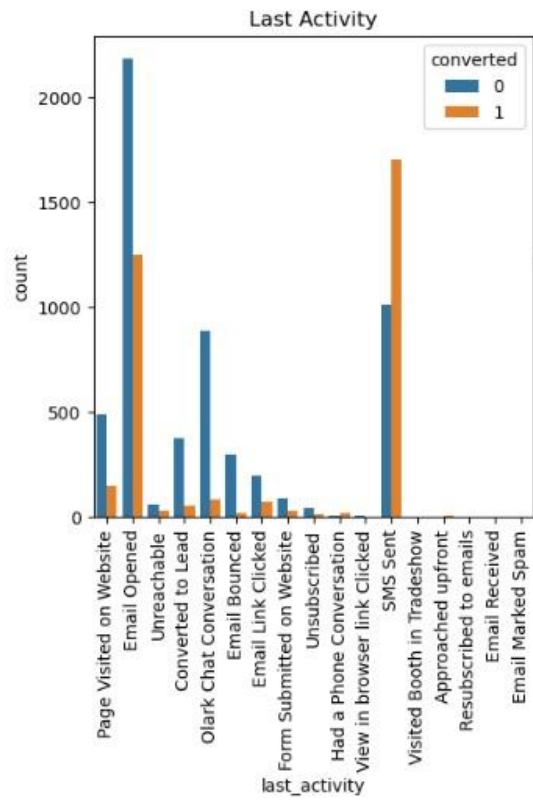


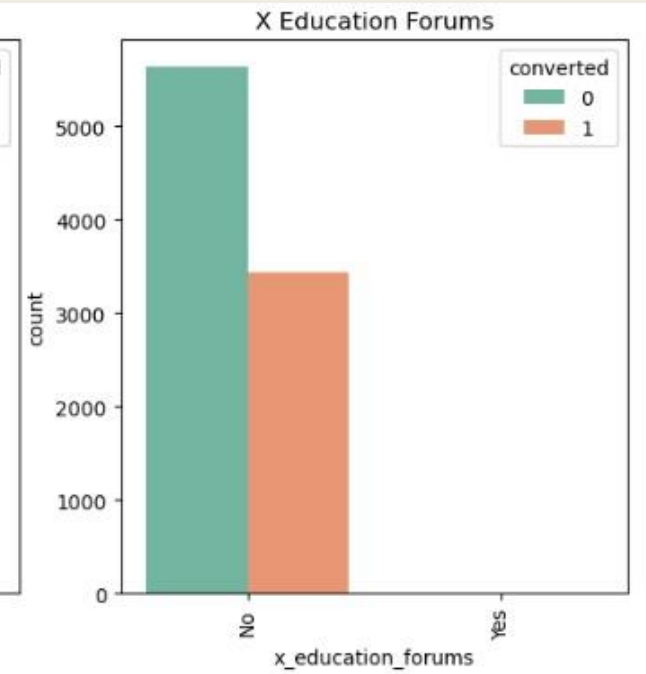
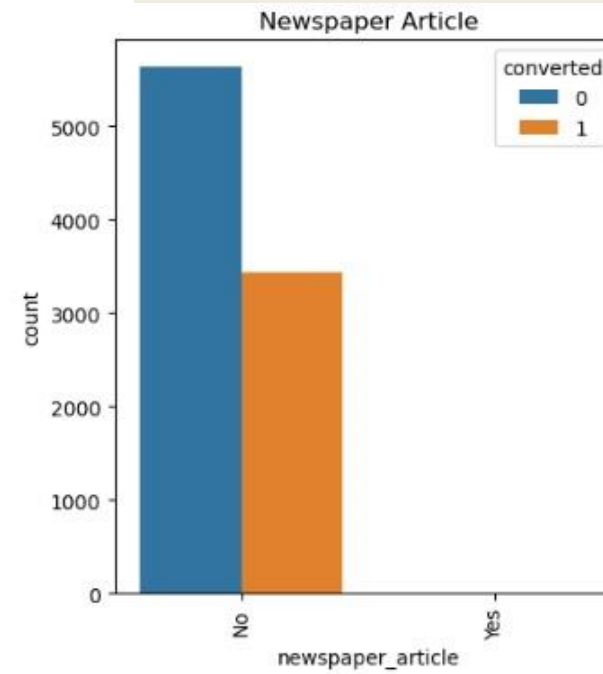
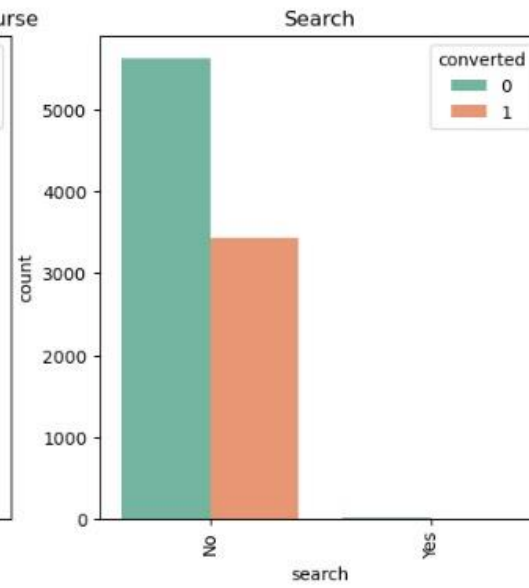
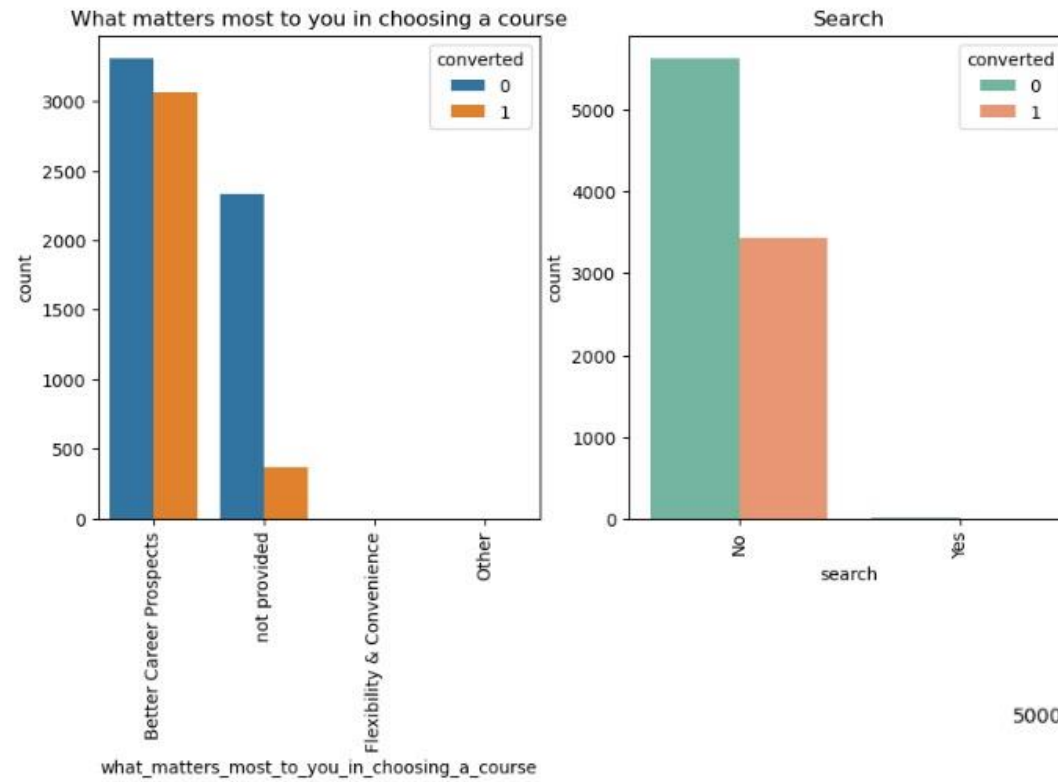
# NUMERICAL VARIABLES

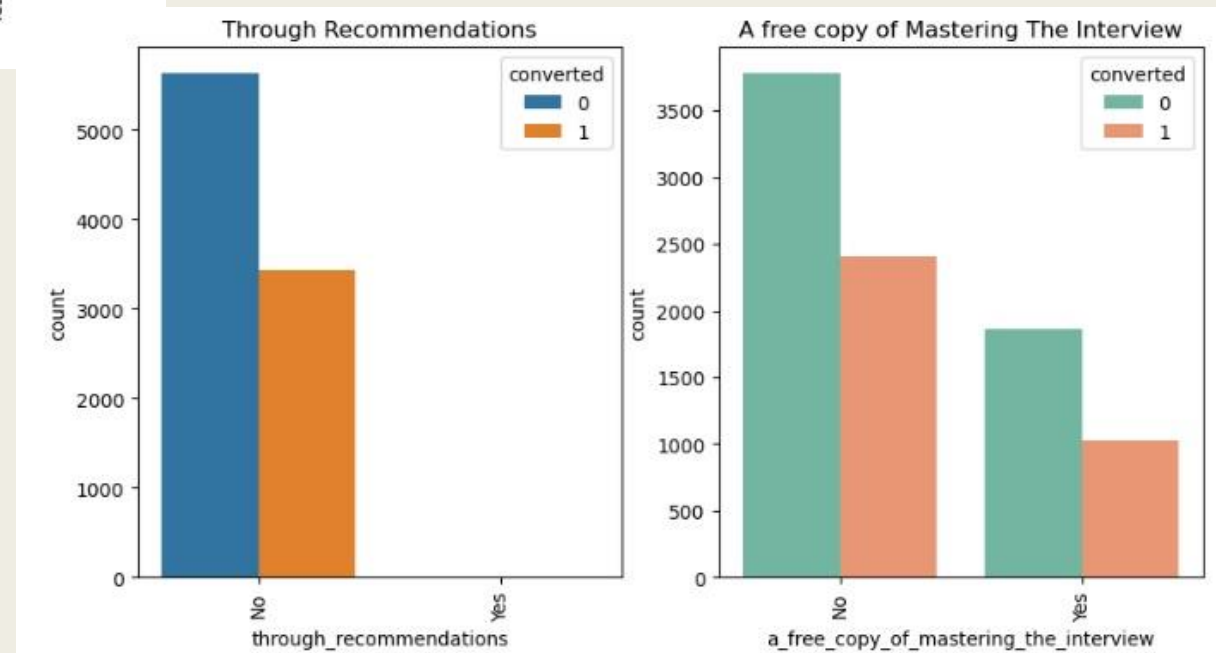
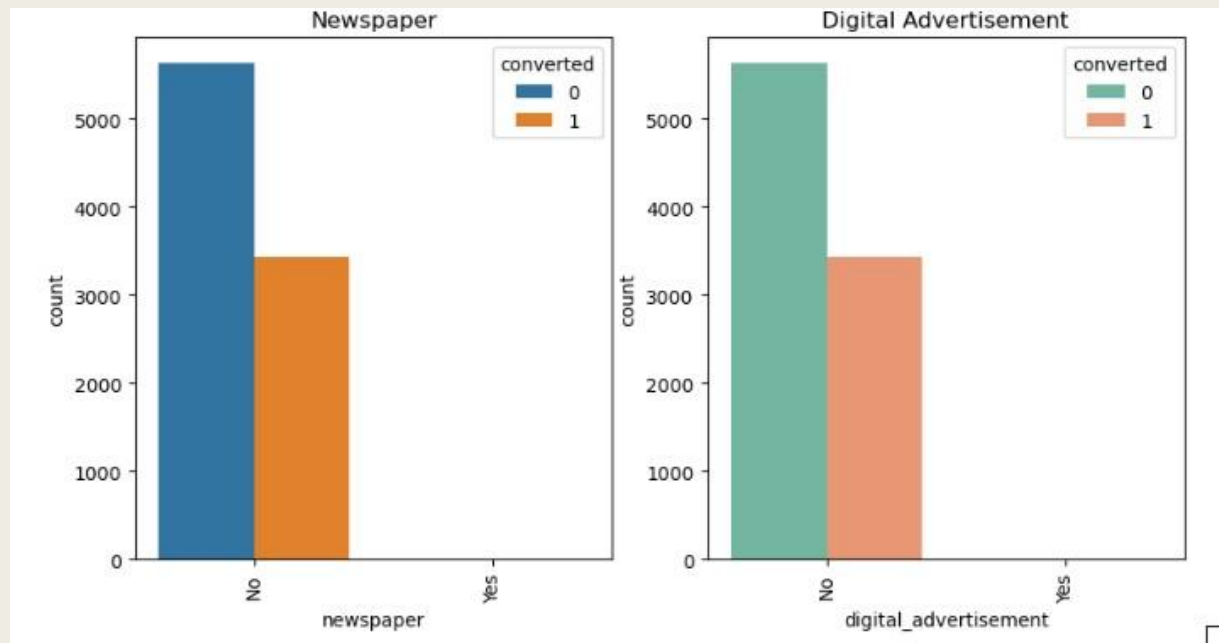


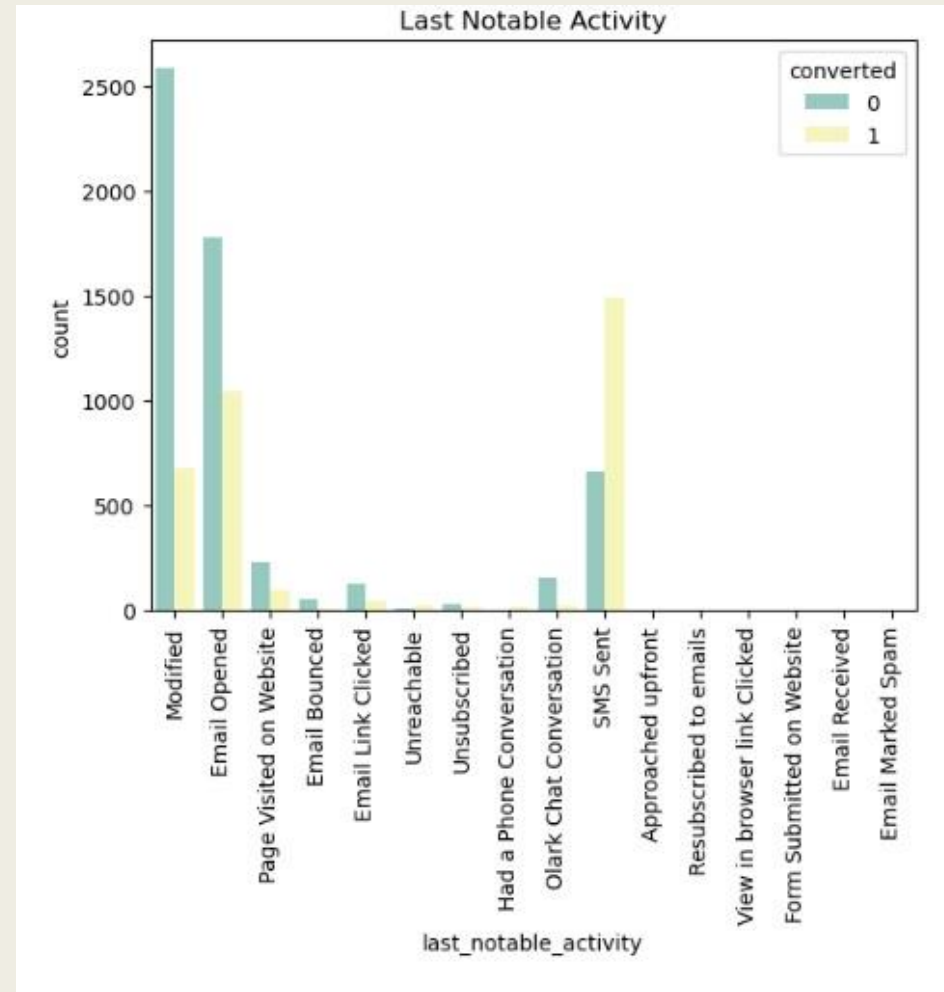
# LINKING ALL THE CATEGORICAL VARIABLES TO CONVERTED











***Based on the preceding exploratory data analysis, it's clear that there are several factors with minimal data, making them less significant to our evaluation.***

# DUMMIES VARIABLES

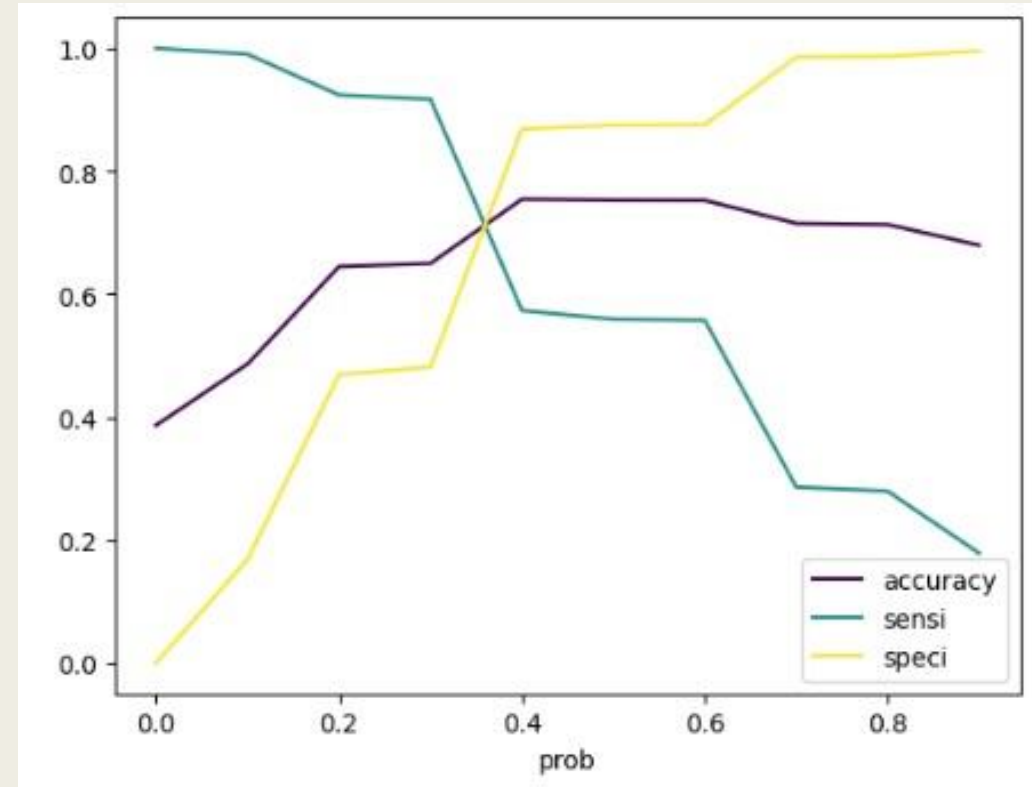
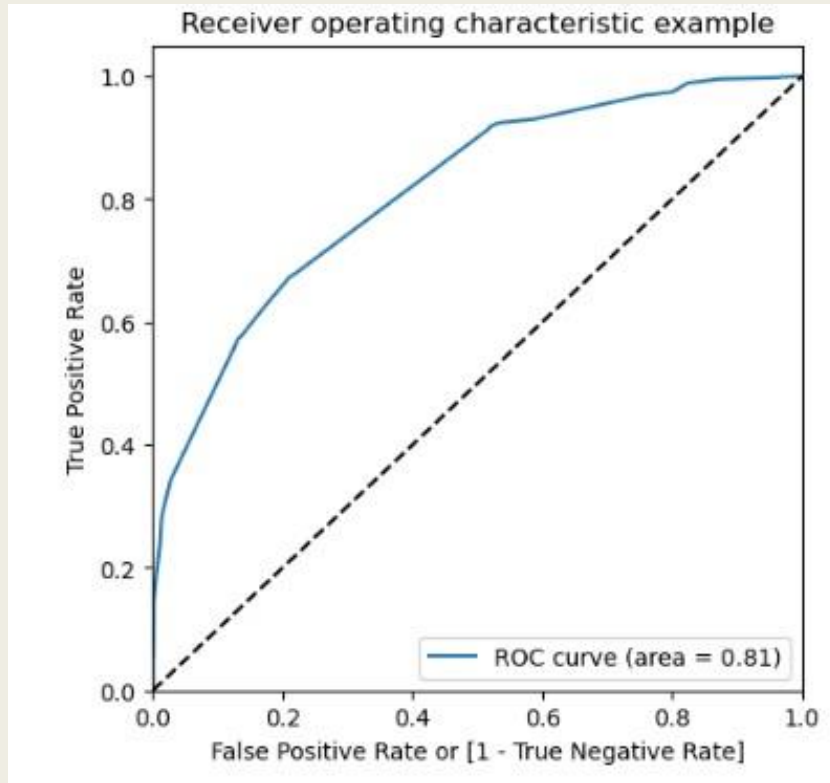
- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 8792
- Total Columns for Analysis: 43
- The following dummies are Dropped
- 'what\_is\_your\_current\_occupation\_not provided', 'lead\_origin', 'lead\_source', 'do\_not\_email', 'do\_not\_call', 'last\_activity', 'country', 'specialization', 'specialization\_not provided', 'what\_is\_your\_current\_occupation', 'what\_matters\_most\_to\_you\_in\_choosing\_a\_course', 'search', 'newspaper\_article', 'x\_education\_forums', 'newspaper', 'digital\_advertisement', 'through\_recommendations', 'a\_free\_copy\_of\_mastering\_the\_interview', 'last\_notable\_activity'



# TEST - TRAIN SPLIT AND MODEL BUILDING

- Dividing the Data into Training and Testing Sets
- The initial fundamental step for regression involves conducting a train-test split, which we have decided to implement using a 70:30 ratio.
- Utilizing RFE for Feature Selection
- Executing RFE with 15 variables as outcomes.
- Creating the Model by eliminating variables that have a p-value exceeding 0.05 and a VIF value greater than 5.
- Making Predictions on the test dataset
- Achieving an overall accuracy of 81%.

# ROC CURVE



- Determining the Ideal Cut-off Point
- The optimal cut-off probability is the point at which sensitivity and specificity are well-balanced.
- It can be observed from the second graph that the ideal cut-off is at 0.35.

# CONCLUSION

The key variables influencing potential buyers are ranked as follows (in descending order):

1. The total duration spent on the website.
2. The overall number of visits.
3. The lead source was identified as:
  - *Google*
  - *Direct traffic*
  - *Organic search*
  - *Welingak website*
4. The timing of the last activity was:
  - *SMS*
  - *Olark chat conversation*
5. The lead origin was in the Lead add format.

Their current occupation falls under working professionals.

With these factors considered, X Education has a significant opportunity to persuade nearly all potential buyers to reconsider and purchase their courses.