LEAD SCORING CASE STUDY

Group Members -

Prakhar Kochhar
Ankita Sharma
Ayaaz Hussain

PROBLEM STATEMENT

1. X Education is an education company that offers online courses to professionals in various industries.
2. The company markets its courses on multiple websites and search engines like Google to attract potential customers.
3. When people visit the X Education website, they may browse courses, fill out a form, or watch videos.
4. Those who fill out a form providing their email or phone number are classified as leads, which the sales team follows up on to convert them into customers.
5. The company's typical lead conversion rate is approximately 30%, meaning only about 30 out of 100 leads are converted into paying customers.
6. X Education aims to increase its lead conversion rate by identifying the most promising leads, also known as "Hot Leads."
7. By focusing more on communicating with the most promising leads, the company believes it can improve its lead conversion rate and make the process more efficient.

OBJECTIVE

1. X Education aims to identify the most promising leads for their business.
2. To achieve this goal, they plan to develop a machine learning model.
3. This model will help identify "hot" leads, or those with a high probability of conversion.
4. Once developed, the model will be deployed for future use.
5. This will enable X Education to focus their resources and efforts on leads that are most likely to generate revenue.

APPROACH USED

1. Duplicate data handling: Check for duplicate data in the dataset and remove them to avoid errors in the analysis.
2. Handling of missing values: Check for missing values and handle them using appropriate techniques like dropping or imputing values.
3. Column dropping: If a column contains a large number of missing values and is not useful for analysis, drop the column.
4. Imputation: If necessary, impute missing values using appropriate techniques.
5. Outlier handling: Identify and handle outliers in the dataset using techniques like 1% Winsorization or removal.
6. Univariate data analysis: Analyze variables individually using techniques like value count and distribution.
7. Feature scaling: Scale the data to a common scale for better analysis.
8. Dummy variables and encoding: Convert categorical data to numerical data using dummy variables and encoding.
9. Logistic regression: Use logistic regression as a classification technique to make predictions based on the data.
10. Validation of the model.
11. Model presentation: Present the results and conclusions of the analysis based on the model.
12. Conclusion and recommendations: Draw conclusions and make recommendations based on the analysis and the results of the model.

DATA PREPARATION

1. Initially, we identified redundant variables and removed them from the dataset.
2. We converted the "Select" variable to null values, as it represented the absence of data entry.
3. We then checked the percentage of null values present in each column and dropped any columns that had more than 45% null values.
4. To address the null values in the dataset, we also removed some rows.
5. Overall, these steps helped to clean and prepare the dataset for further analysis.

EXPLORATORY DATA ANALYSIS

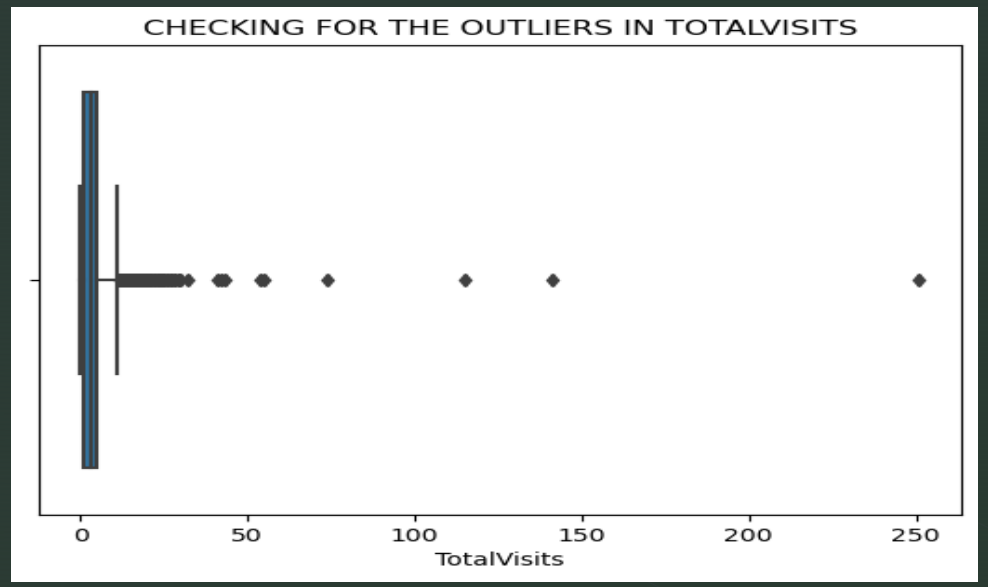
1. To assess the quality of our data, we performed an exploratory data analysis (EDA).
2. During the EDA, we found that several elements in the categorical variables were insignificant.
3. We also observed outliers in the "TOTALVISITS" variable.
4. To address the outliers, we removed the top 1% and bottom 1% of values from the column.
5. By handling the outliers and other issues identified during the EDA, we improved the overall quality of the dataset.
6. This helped ensure that any subsequent analysis or modeling efforts would be based on reliable data.

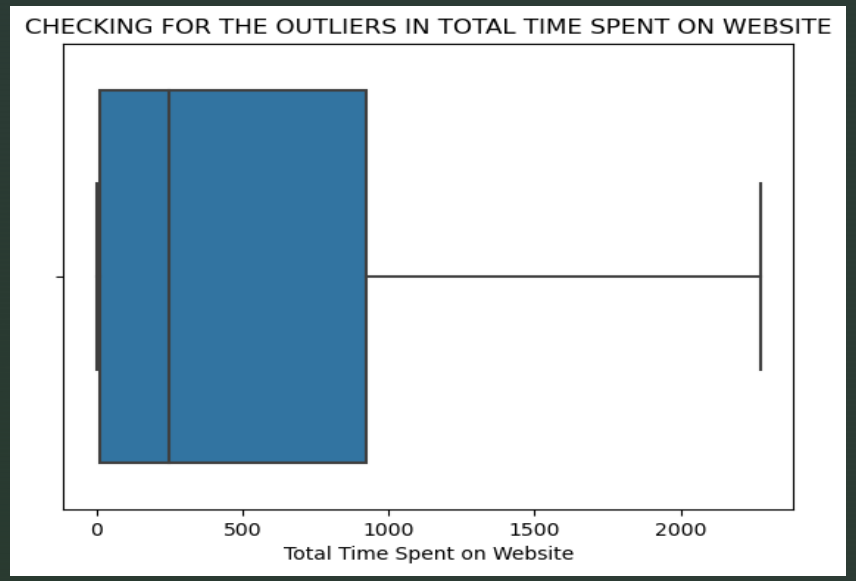
LIST OF COLUMNS DROPPED

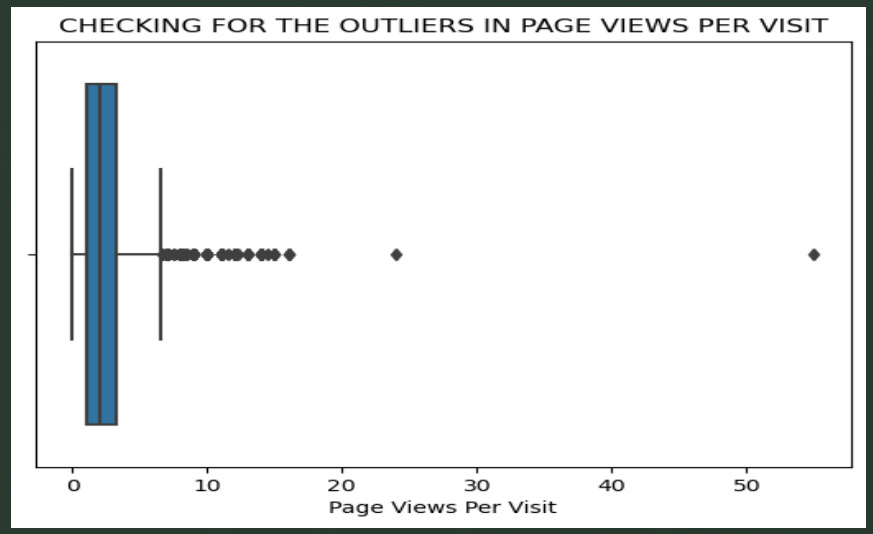
List of columns which we dropped taking into consideration that they were redundant variables.

1. Prospect ID", "Lead Number“ we dropped these columns as we found them to be merely unique customer id.
2. We also deleted few columns which had no unique category we checked that using “nunique” function
3. "Tags", "Country", "Newspaper Article", "X Education Forums", "Newspaper", "Digital Advertisement“.
4. We also deleted columns with high percentage of null values for instance more than 45% of null values in a given column.

OUTLIER ANALYSIS



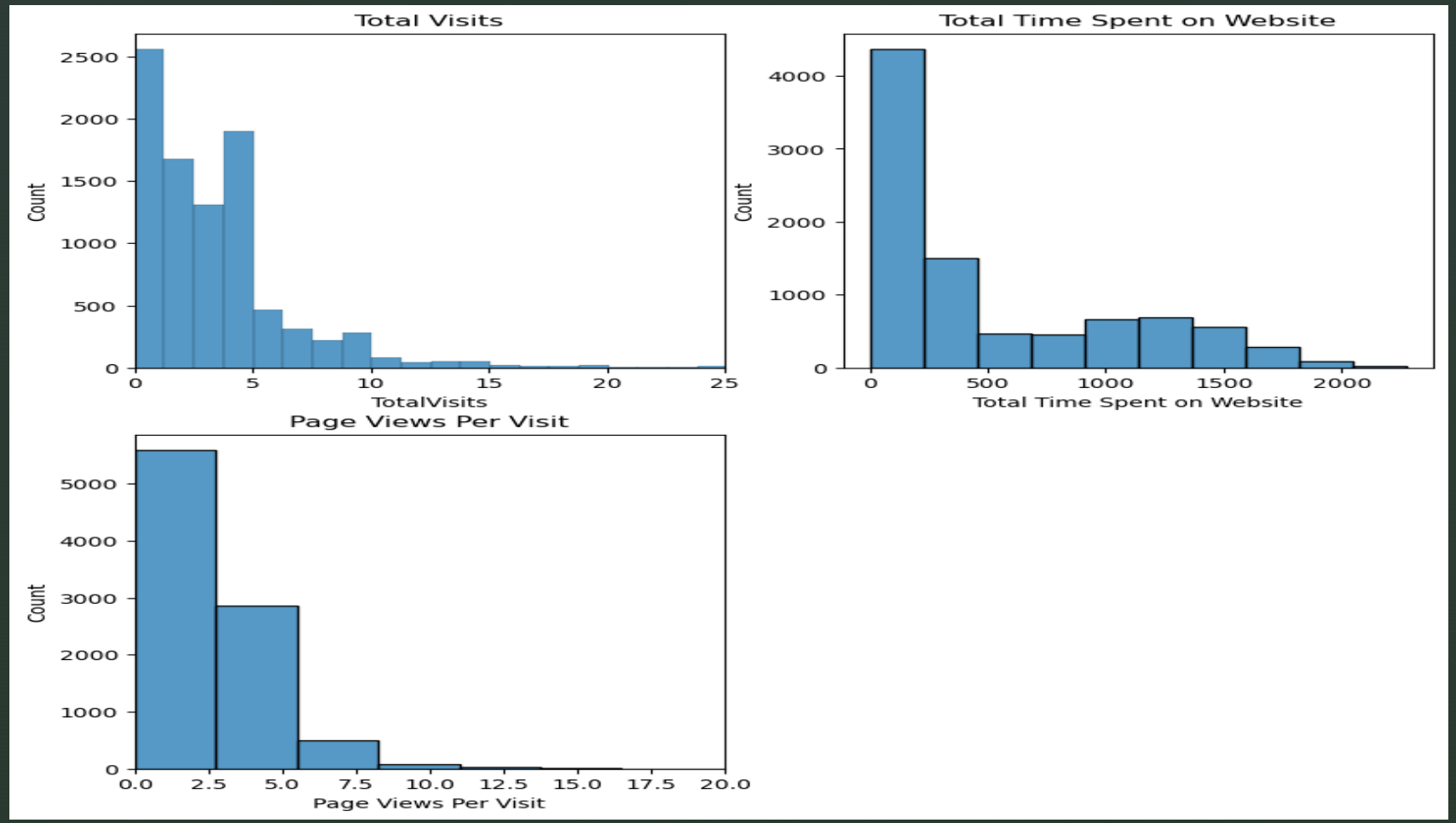




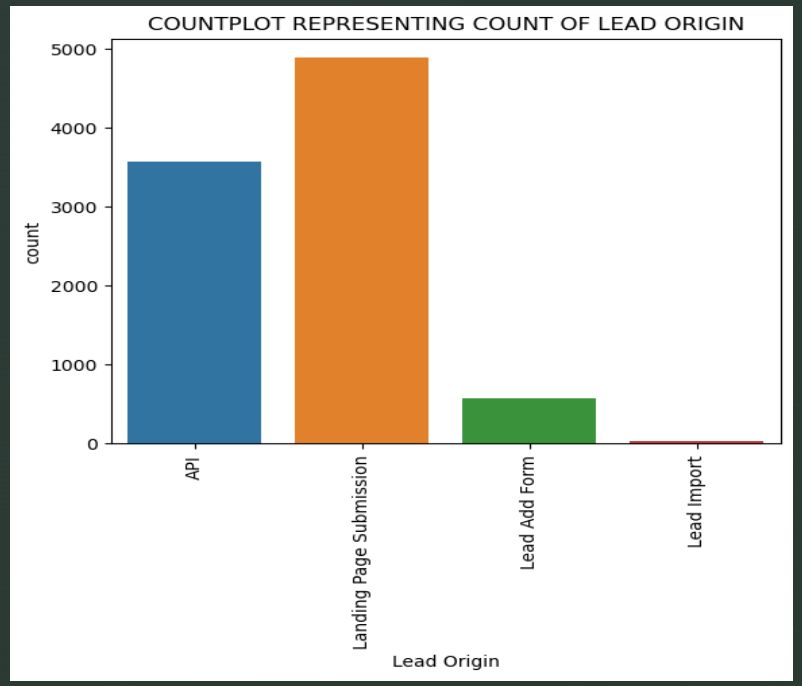
- We dealt with outliers in the “Totalvisits” column using 1% Winsorization technique.
- It means we dealt with it by removing top and bottom 1% of the values from the “Totalvisits” column.

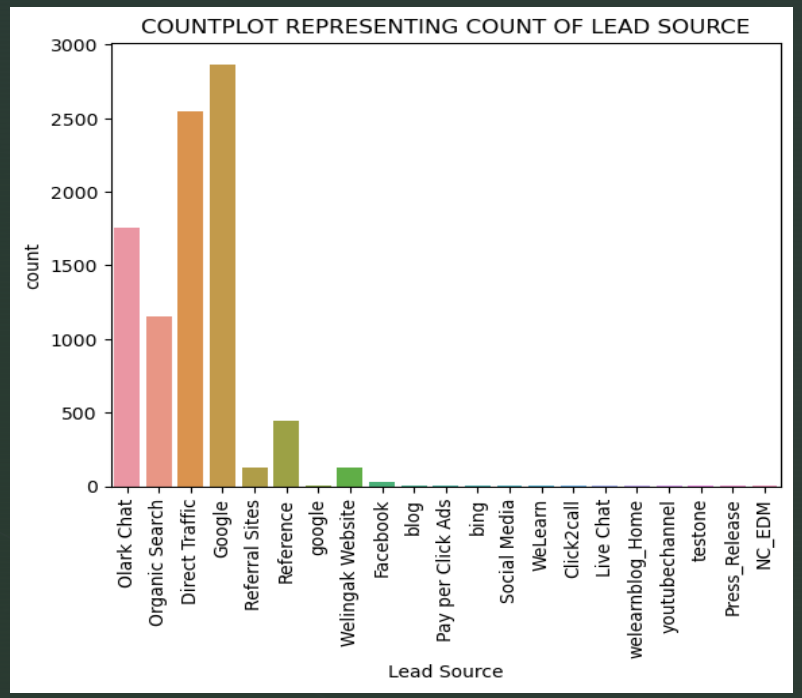
CORRELATION HEATMAP

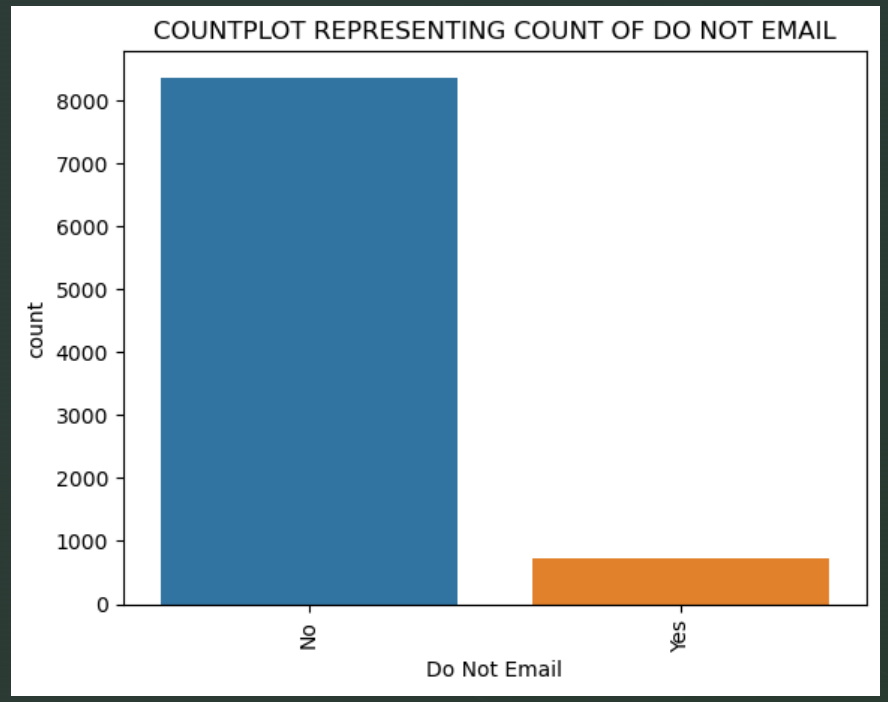
Numerical Variable Analysis

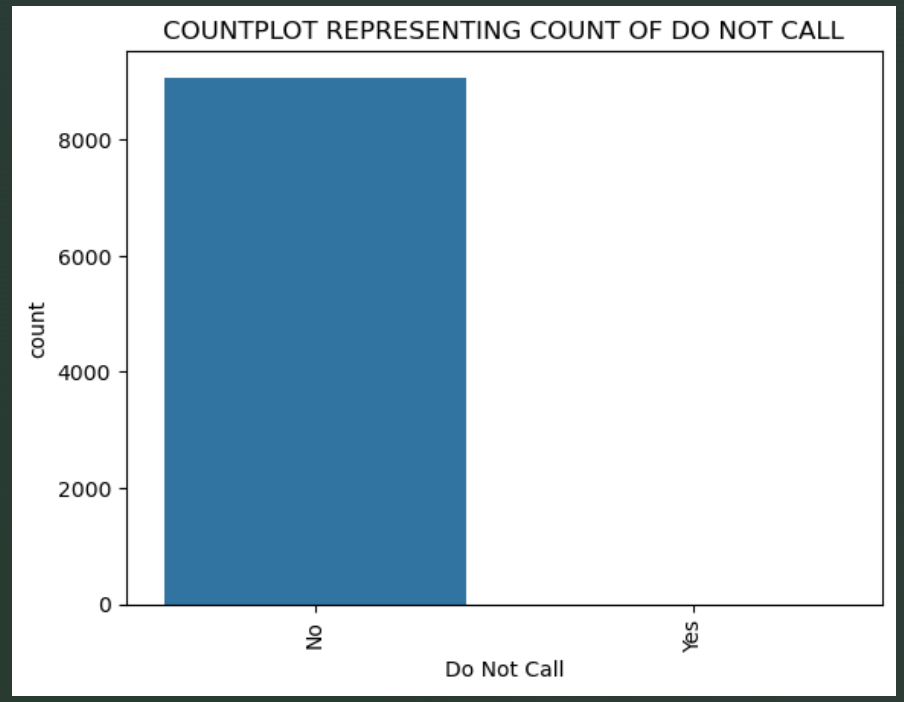


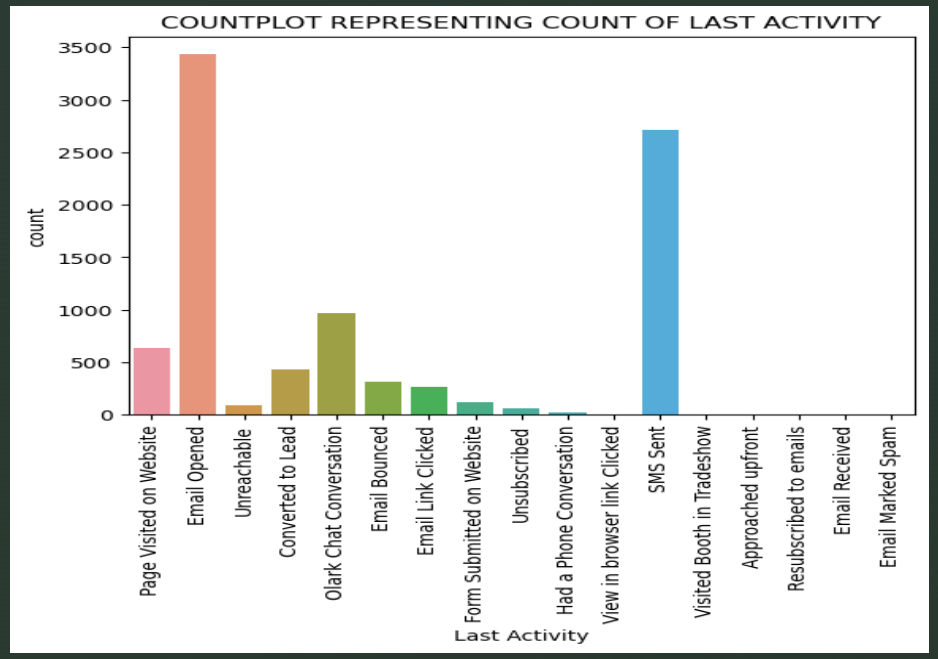
CATEGORICAL VARIABLE ANALYSIS

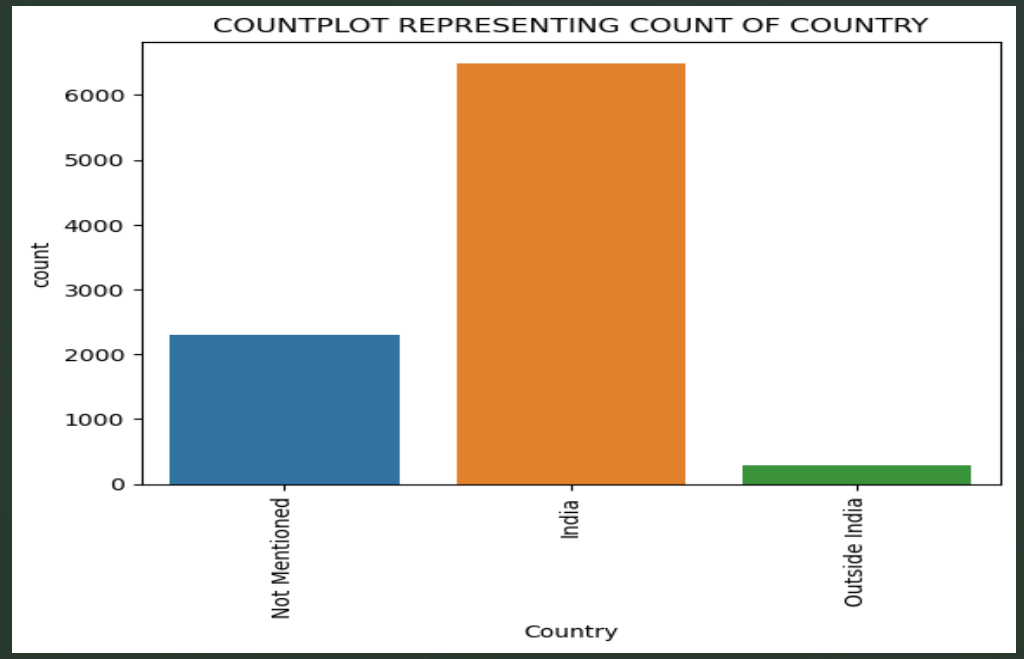


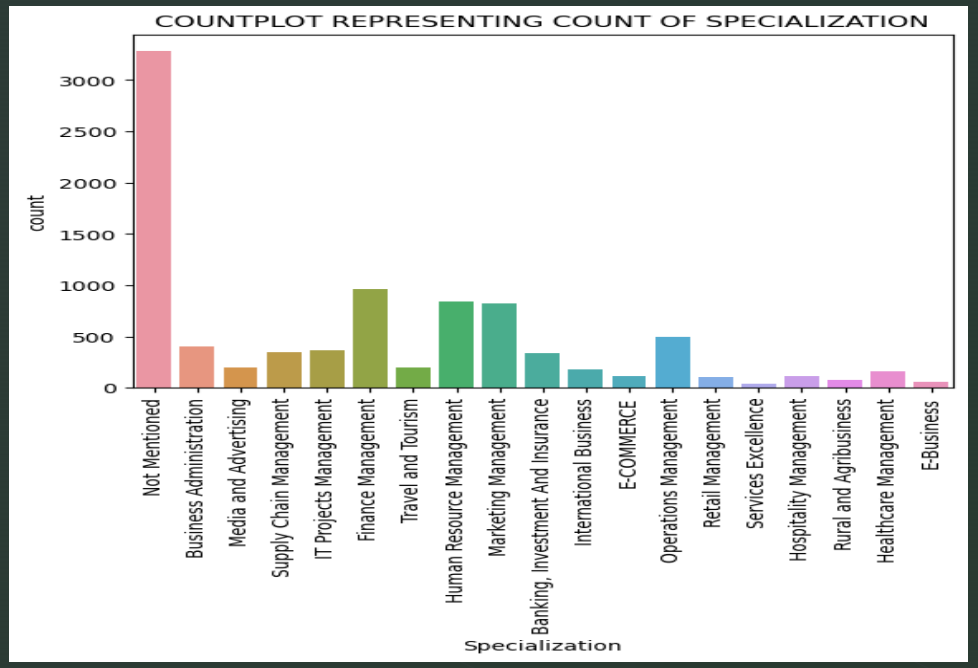


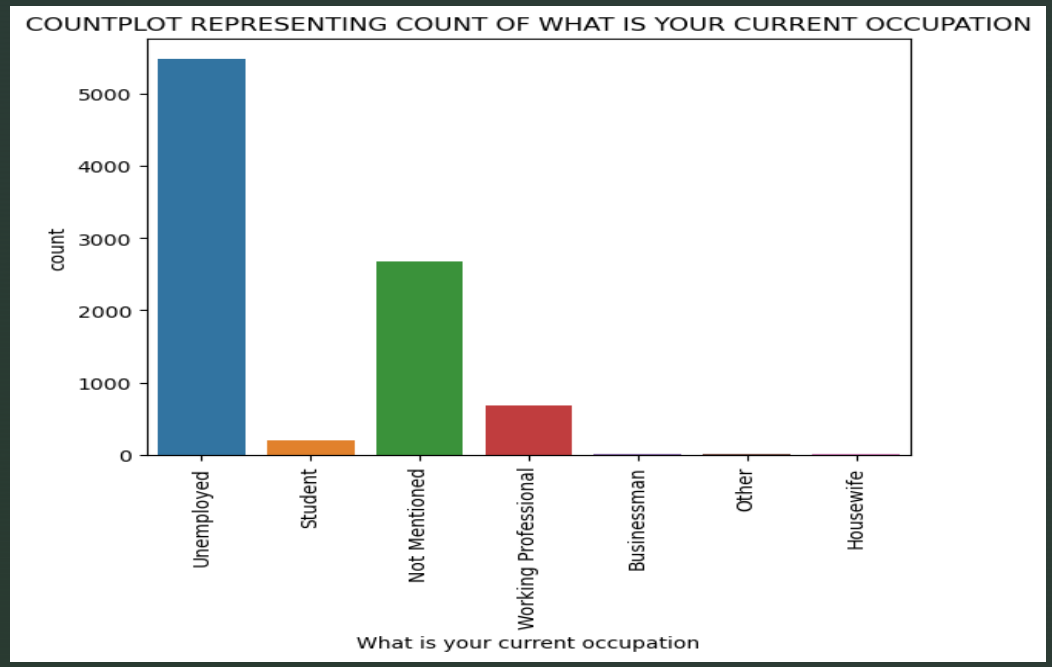

COUNTPLOT REPRESENTING COUNT OF DO NOT EMAIL
count
8000
7000
6000
5000
4000
3000
2000
1000
0
No
Yes
Do Not Email

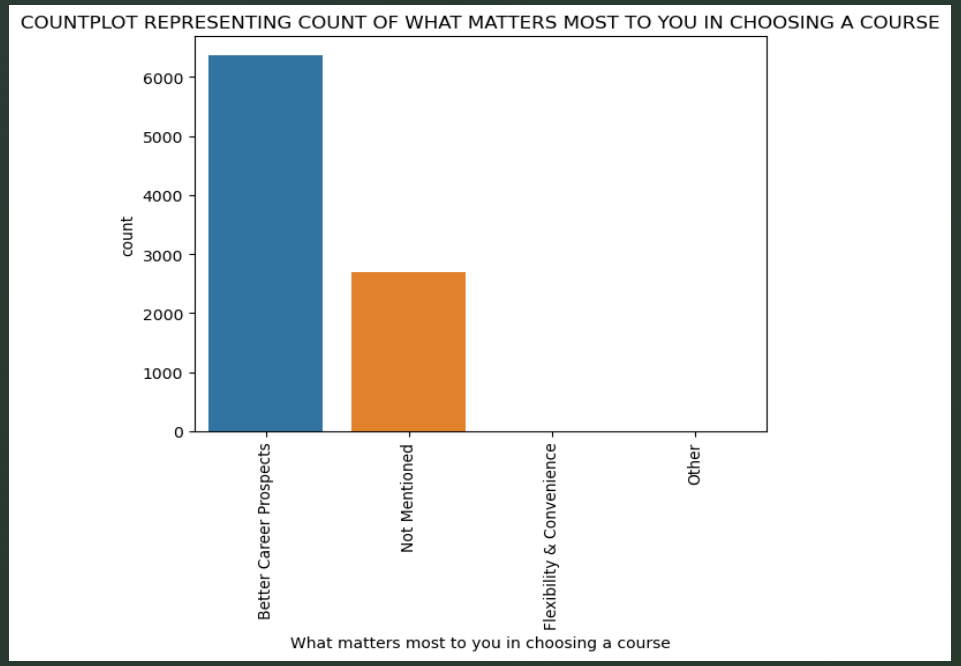

COUNTPLOT REPRESENTING COUNT OF DO NOT CALL
count
8000
6000
4000
2000
0
No
Yes
Do Not Call

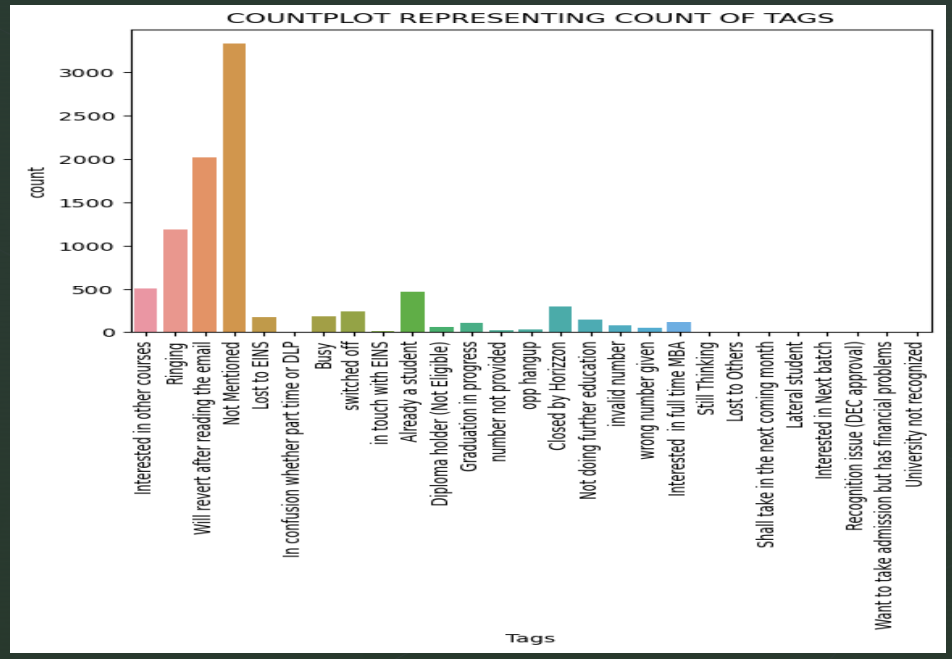

COUNTPLOT REPRESENTING COUNT OF LAST ACTIVITY
count
3500
3000
2500
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
SMS Sent
Visited Booth in Tradeshow
Approached upfront
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity


COUNTPLOT REPRESENTING COUNT OF COUNTRY
count
6000
5000
4000
3000
2000
1000
0
Not Mentioned
India
Outside India
Country


COUNTPLOT REPRESENTING COUNT OF SPECIALIZATION
count
3000
2500
2000
1500
1000
500
0
Not Mentioned
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization


COUNTPLOT REPRESENTING COUNT OF WHAT IS YOUR CURRENT OCCUPATION
count
5000
4000
3000
2000
1000
0
Unemployed
Student
Not Mentioned
Working Professional
Businessman
Other
Housewife
What is your current occupation


COUNTPLOT REPRESENTING COUNT OF WHAT MATTERS MOST TO YOU IN CHOOSING A COURSE
count
6000
5000
4000
3000
2000
1000
0
Better Career Prospects
Not Mentioned
Flexibility & Convenience
Other
What matters most to you in choosing a course


COUNTPLOT REPRESENTING COUNT OF TAGS
count
3000
2500
2000
1500
1000
500
0
Interested in other courses
Ringing
Will revert after reading the email
Not Mentioned
Lost to EINS
In confusion whether part time or DLP
Busy
switched off
in touch with EINS
Already a student
Diploma holder (Not Eligible)
Graduation in progress
number not provided
opp hangup
Closed by Horizzon
Not doing further education
invalid number
wrong number given
Interested in full time MBA
Still Thinking
Lost to Others
Shall take in the next coming month
Lateral student
Interested in Next batch
Recognition issue (DEC approval)
Want to take admission but has financial problems
University not recognized
Tags

DATA HANDLING

- Creating Dummy Variables: - In this step, we created dummy variables with the help of pandas get_dummies function.
- Performing Train-Test Split: - In this step, we split the data into X_train, X_test and y_train, y_split the ratio of split we chose for train and test was 70% and 30% respectively.
- Feature Scaling: - Here, we scaled the data using StandardScaler from sklearn.preprocessing library.

MODEL BUILDING

- Two machine learning models were built.
- The first step was to use Recursive Feature Elimination (RFE) to select the 15 most relevant variables.
- Variables were dropped based on their p-value and VIF score.
- A threshold of 5 was chosen for VIF score and 0.05 for p-value.

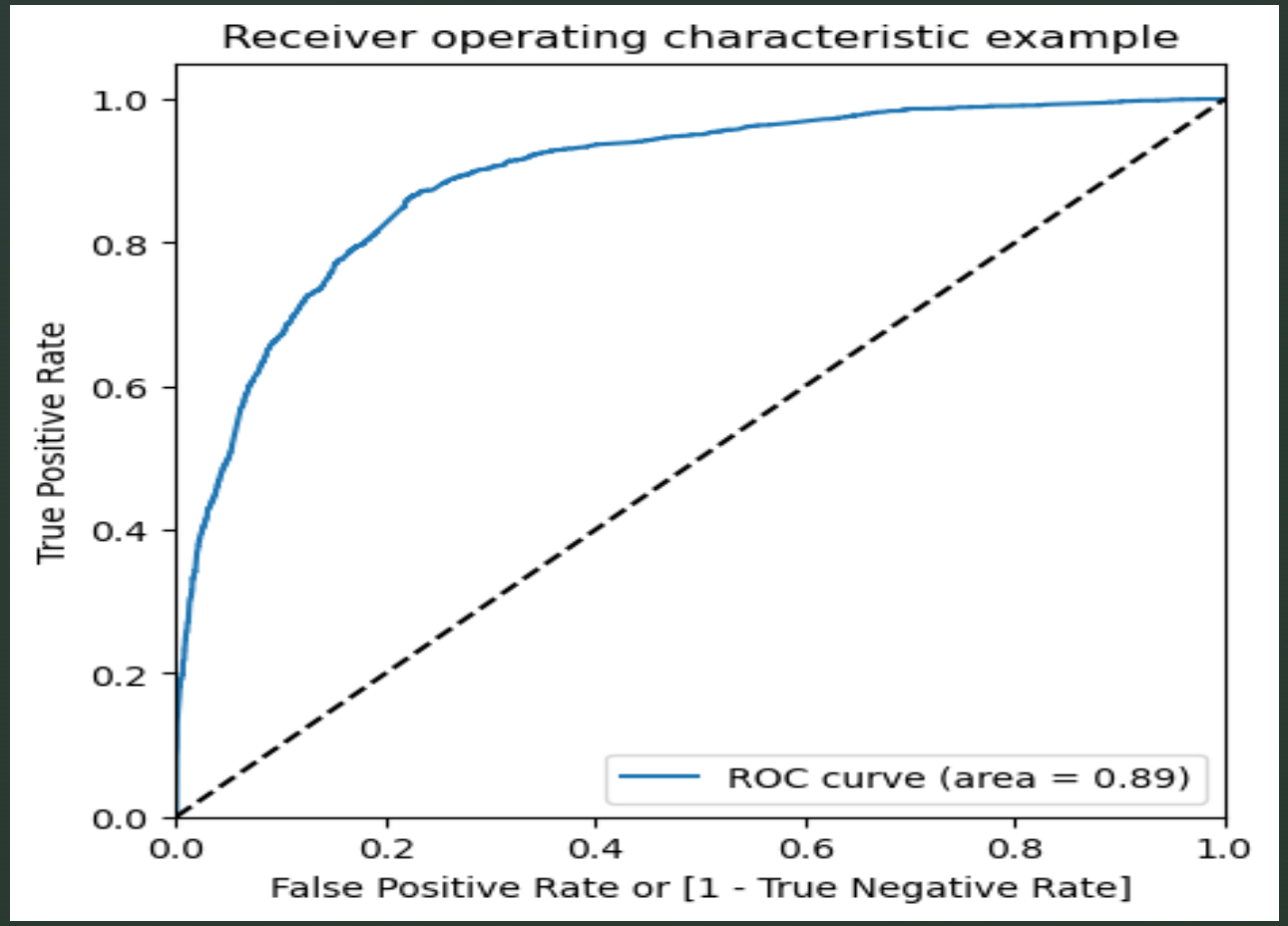
MODEL EVALUATION

- A confusion matrix was created to evaluate the classification model.
- The goal was to find the best cut-off point to balance the sensitivity and specificity of the model.
- The optimal cut-off point was found to be 0.35.
- After determining the optimal cut-off point, the accuracy, specificity, and sensitivity of the model were checked on the test set.
- The accuracy, specificity, and sensitivity were approximately 80%, 81%, and 81%, respectively.

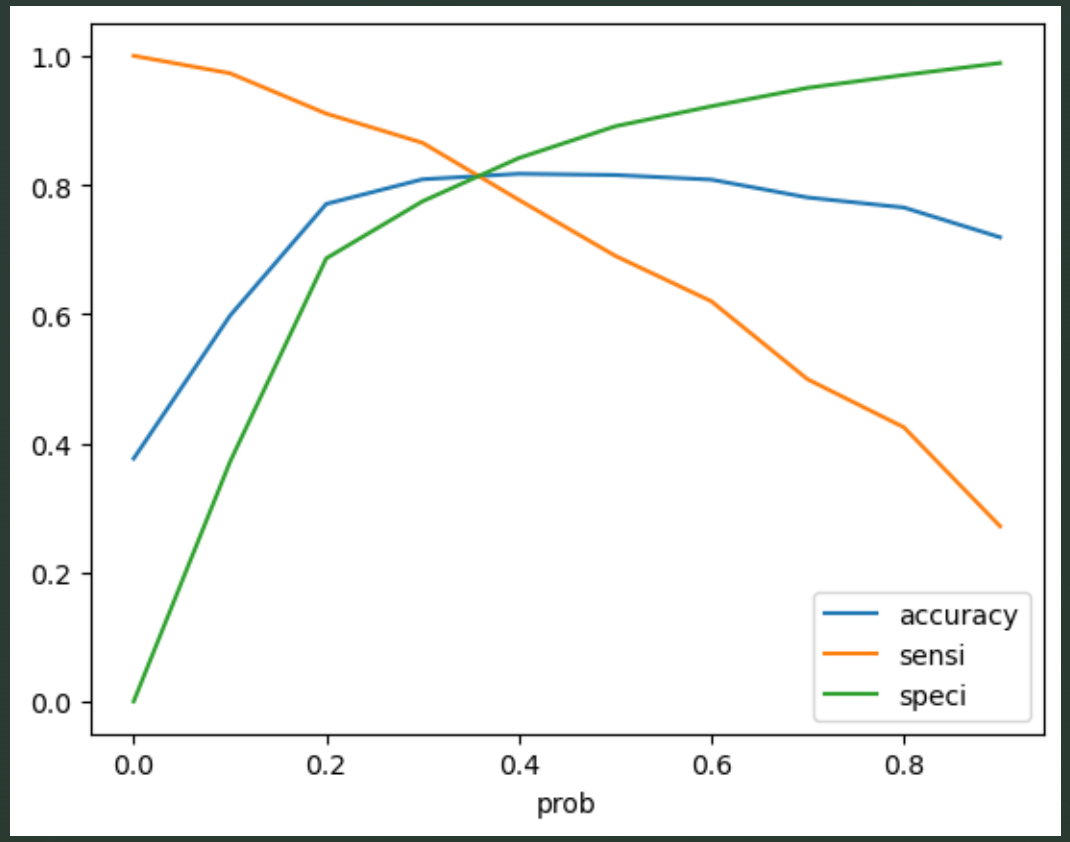
PRECISION AND RECALL

- A precision-recall curve was plotted to evaluate the performance of the classification model.
- An optimal cut-off point of 0.42 was determined from the curve.
- After setting the cut-off point to 0.42, predictions were made on the test dataset.
- The precision and recall values of the model were found to be 74% and 76%, respectively.

ROC CURVE

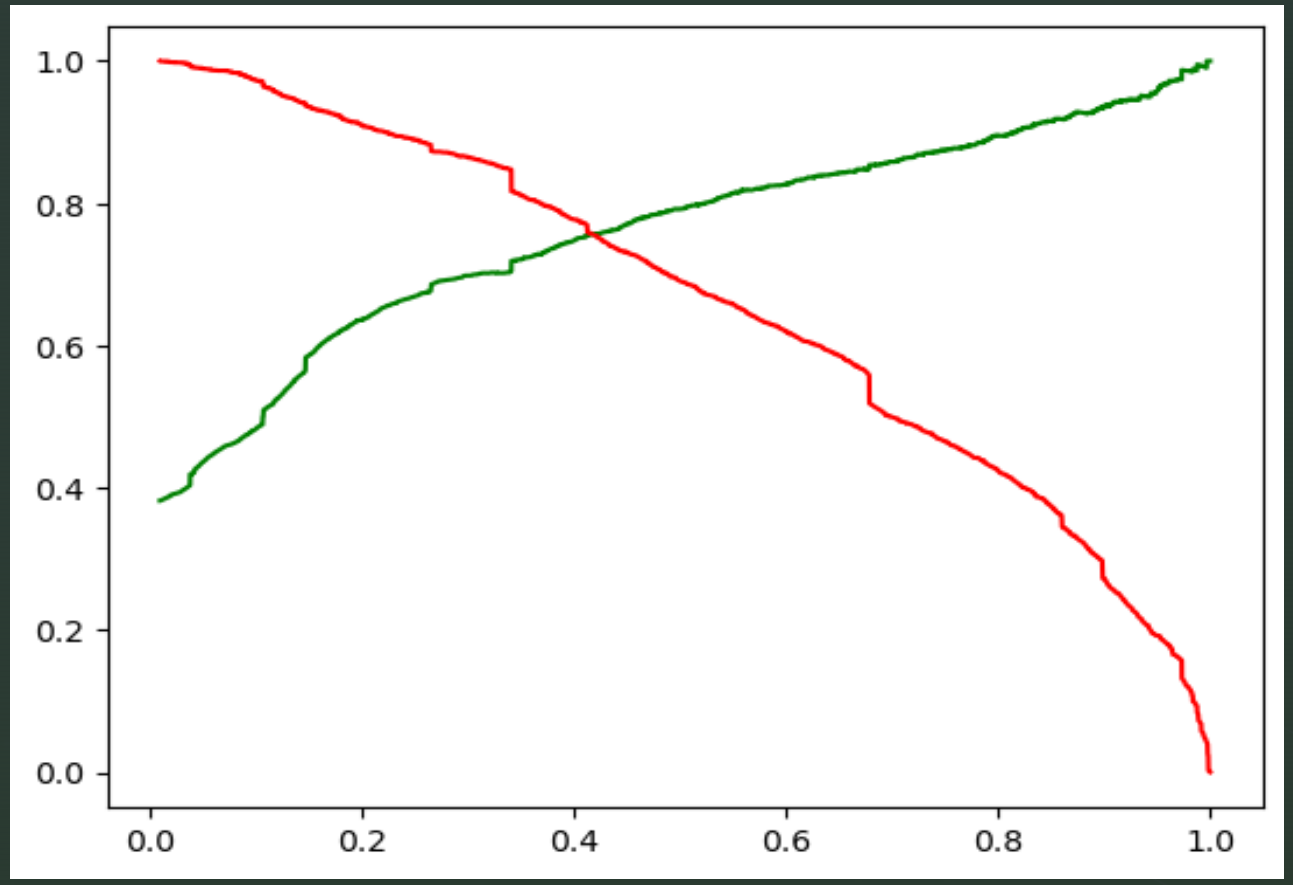

Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.89)

OPTIMAL CUTOFF POINT



- From the curve above we can infer 0.35 is the optimum point to take it as a cutoff probability
- This is an optimal cut off point i.e., a point where we get balanced sensitivity and specificity.

PRECISION RECALL CURVE



- Precision and recall curve suggests us to set a cut of 0.42 which we used in later parts of our analysis.

FINDINGS AND RECOMMENDATIONS

The variables that are most significant in determining whether a lead is a potential buyer or not are listed below.

- Factors that have positive impact in converting the lead :

1. Total Time Spent on Website.
2. If the lead is a Working Professional.
3. If the lead had a Phone Conversation with the sales team.
4. The lead source was:
   a) Welingak Website
   b) Reference

- At the same time customers falling under below mentioned category are less likely to convert :

1. Who have said don't email them.
2. With whom last conversation was through Olark chat.
3. When the client has not mentioned "What matters you the most" when choosing a particular course.

THANK YOU