X Education – Lead Scoring Case Study

Team Members :
Vaishnavi Rachita
Raghunath
Priyanka G

Table of Contents

- Background of X Education Company
- Problem Statement & Objective of the Study
- Suggested Ideas for Lead Conversion
- Analysis Approach
- Data Cleaning
- EDA
- Data Preparation
- Model Building (RFE & Manual fine tuning)
- Model Evaluation
- Recommendations

Background of X Education Company

- An education company named X Education sells online courses to industry professionals.
- On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google.
- Once these 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.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc.
- Through this process, some of the leads get converted while most do not.
- The typical lead conversion rate at X education is around 30%.

Problem Statement & Objective of the Study

- **Problem Statement:**
- X Education gets a lot of leads; its lead conversion rate is very poor at around 30%
- X Education wants to make lead conversion process more efficient by identifying the most potential leads, also known as Hot Leads
- Their sales team want to know these potential set of leads, which they will be focusing more on communicating rather than making calls to everyone.

- **Objective of the Study:**
- To help X Education 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 we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- The CEO has given a ballpark of the target lead conversion rate to be around 80%.

Data Cleaning

"Select" level represents null values for some categorical variables, as customers did not choose any option from the list.

Columns with over 40% null values were dropped.

Missing values in categorical columns were handled based on value counts and certain considerations.

Drop columns that don't add any insight or value to the study objective (tags, country)

Imputation was used for some categorical variables.

Additional categories were created for some variables.

Columns with no use for modeling (Prospect ID, Lead Number) or only one category of response were dropped. Numerical data was imputed with mode after checking distribution.

Data Cleaning

Skewed category columns were checked and dropped to avoid bias in logistic regression models.

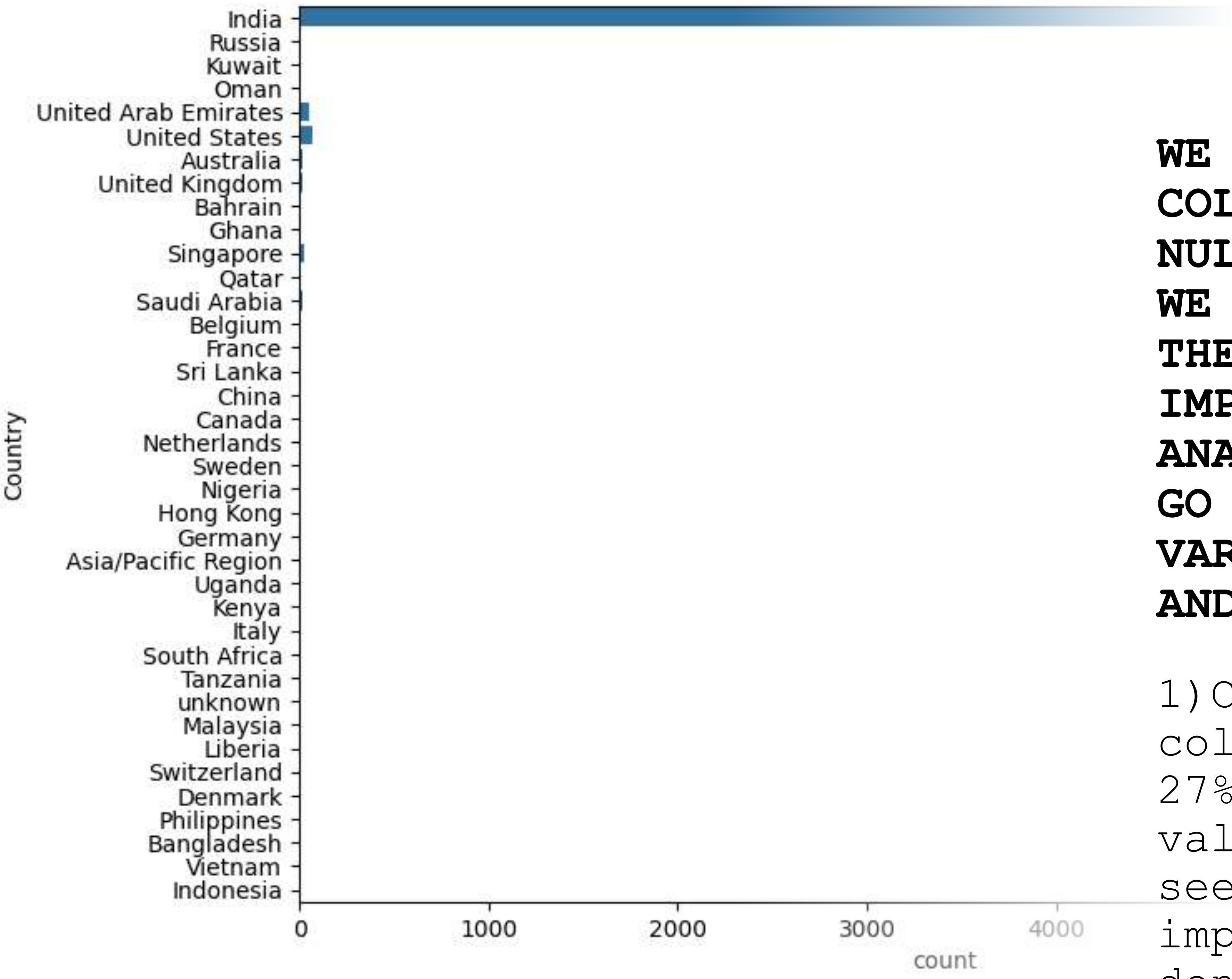
Outliers in TotalVisits and Page Views Per Visit were treated and capped.

Invalid values were fixed and data was standardized in some columns, such as lead source.

Low frequency values were grouped together to "Others".

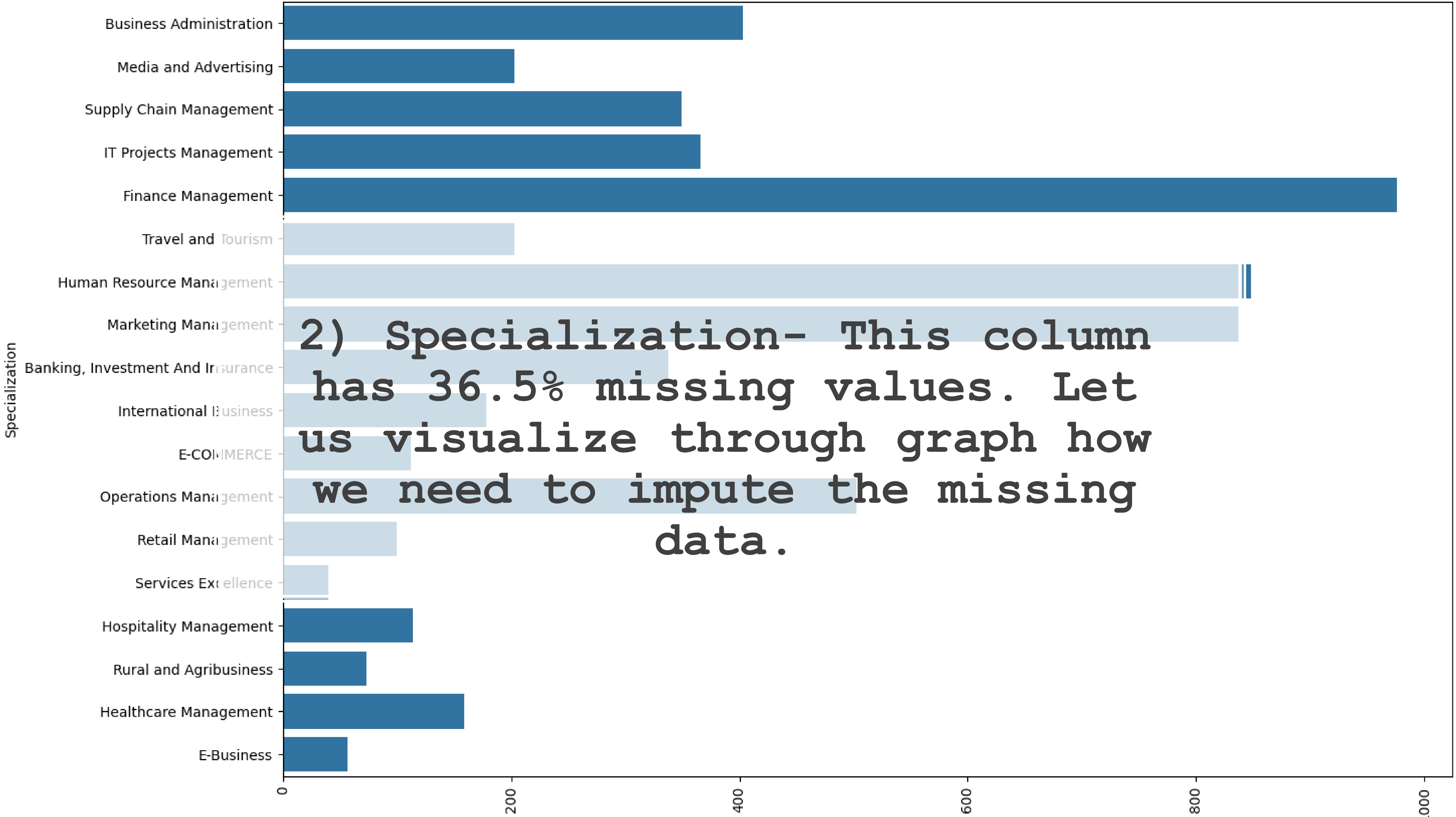
Binary categorical variables were mapped.

Other cleaning activities were performed to ensure data quality and accuracy. o Fixed Invalid values & Standardizing Data in columns by checking casing styles, etc. (lead source has Google, google)



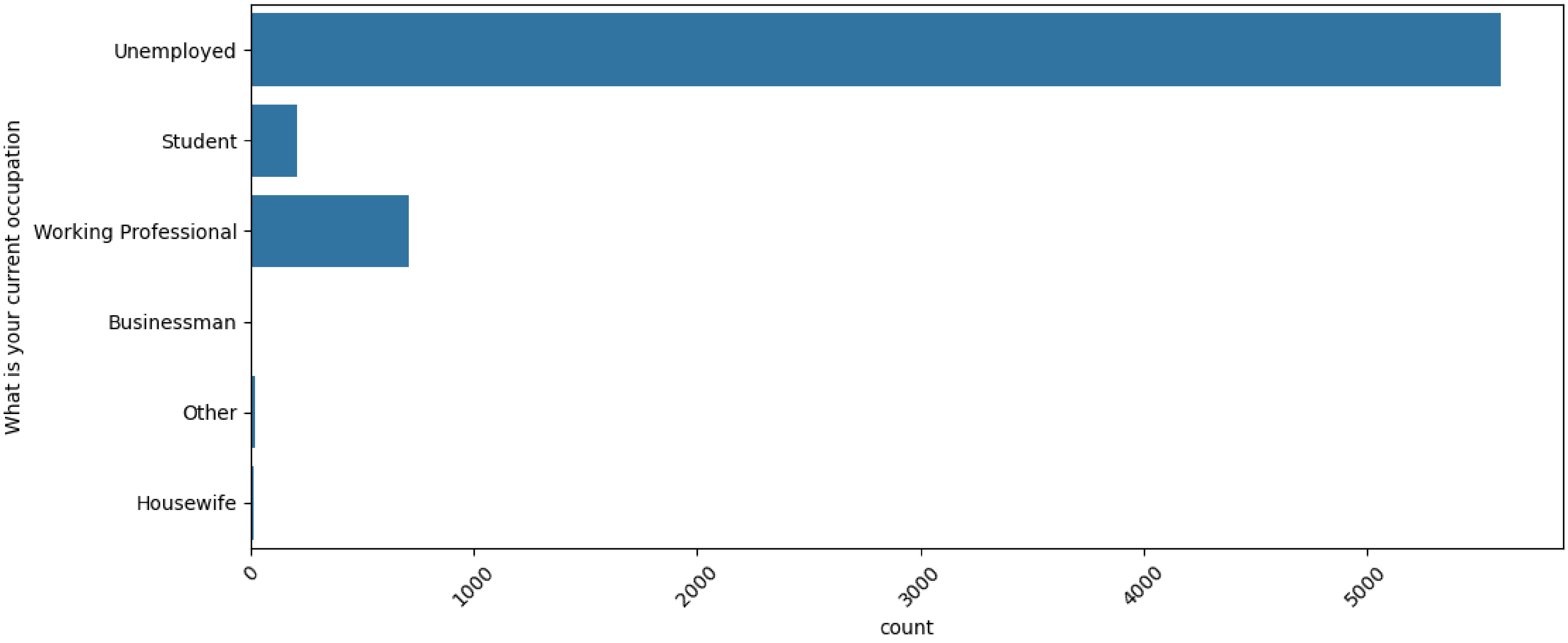
WE SEE CERTAIN COLUMNS WITH HIGH NULL VALUES, BUT WE CANNOT DROP THEM AS THEY ARE IMPORTANT FOR ANALYSIS. SO, WE GO FOR THE VARIABLE ANALYSIS AND IMPUTATION.

1)Country- This column has almost 27% missing values. We will see how its imputation is done

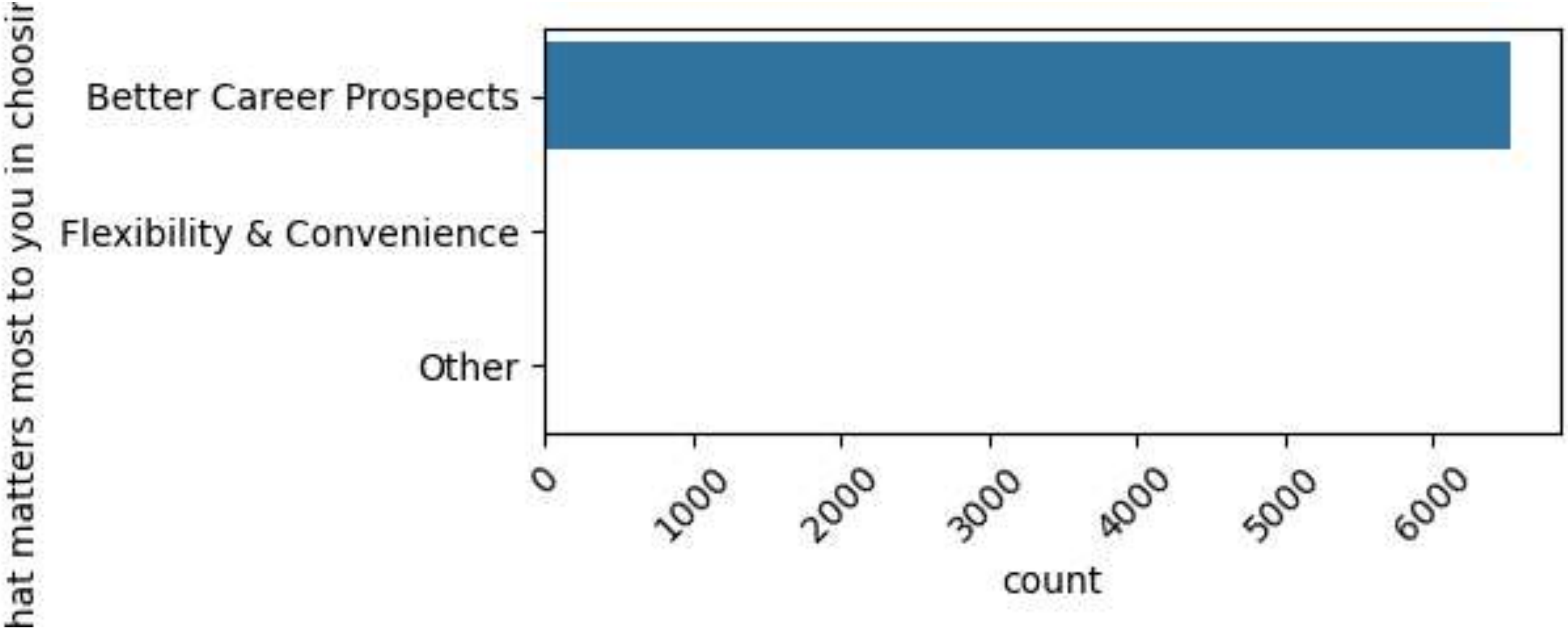


2) Specialization- This column has 36.5% missing values. Let us visualize through graph how we need to impute the missing data.

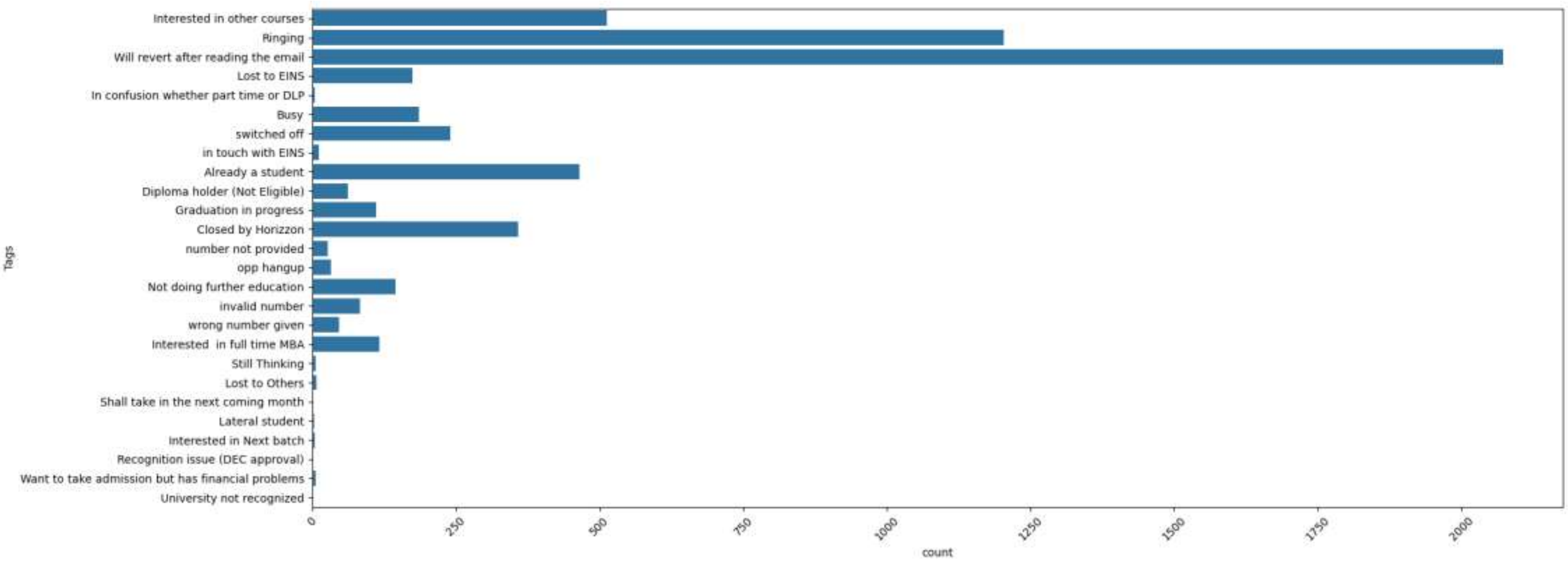
3) What is your current occupation- This variable has almost 30% missing values.



4) 'What matters most to you in choosing a course'- This column has almost 30% missing values. Let us visualize the categories of columns using a plot.

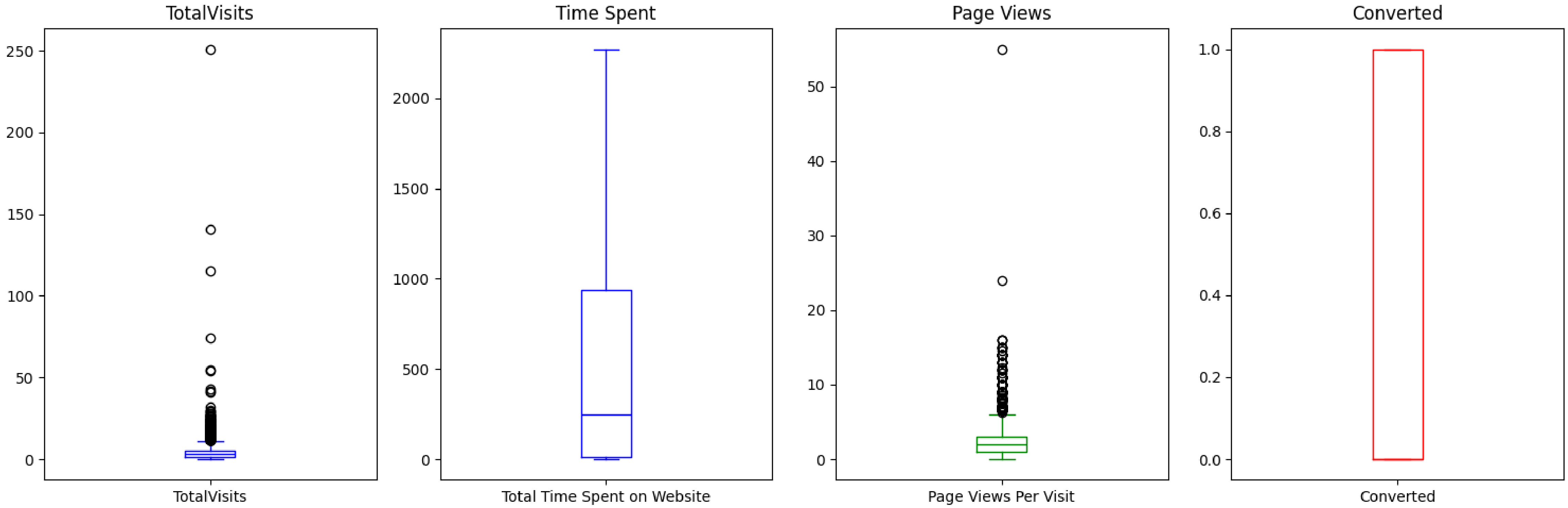


5) 'Tags'- The column Tags has 36% missing values.Let us visualize the column categories

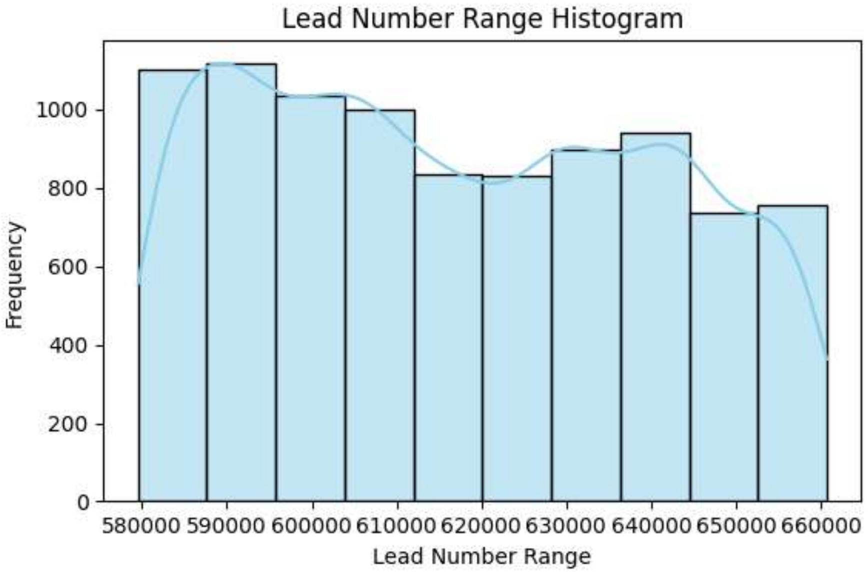


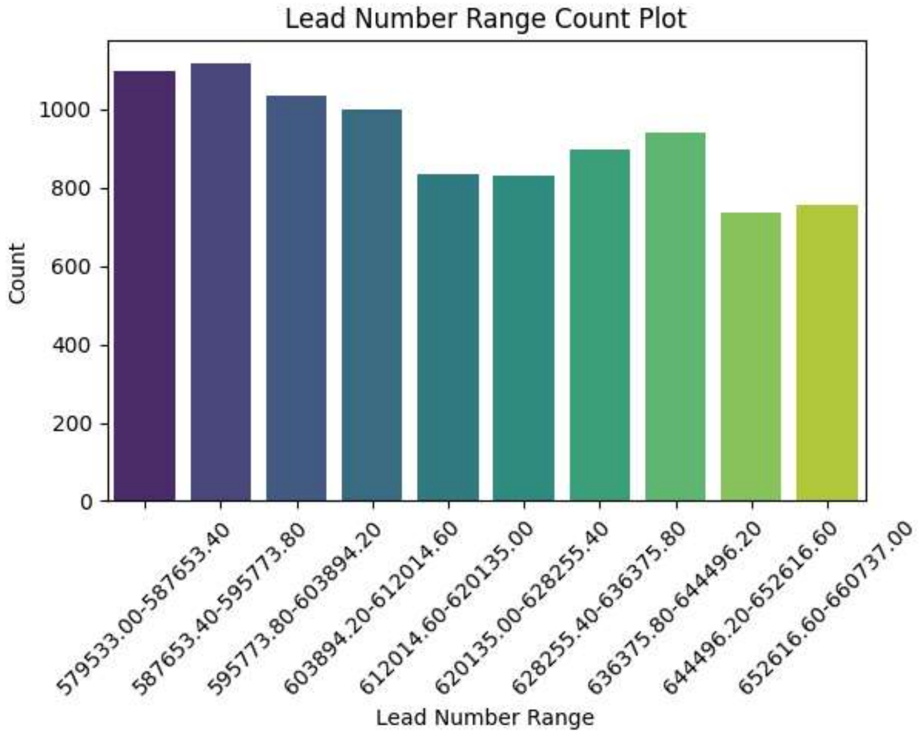
EDA

- Conversion rate is 38.5%, meaning only 38.5% of the people have converted to leads(Minority)

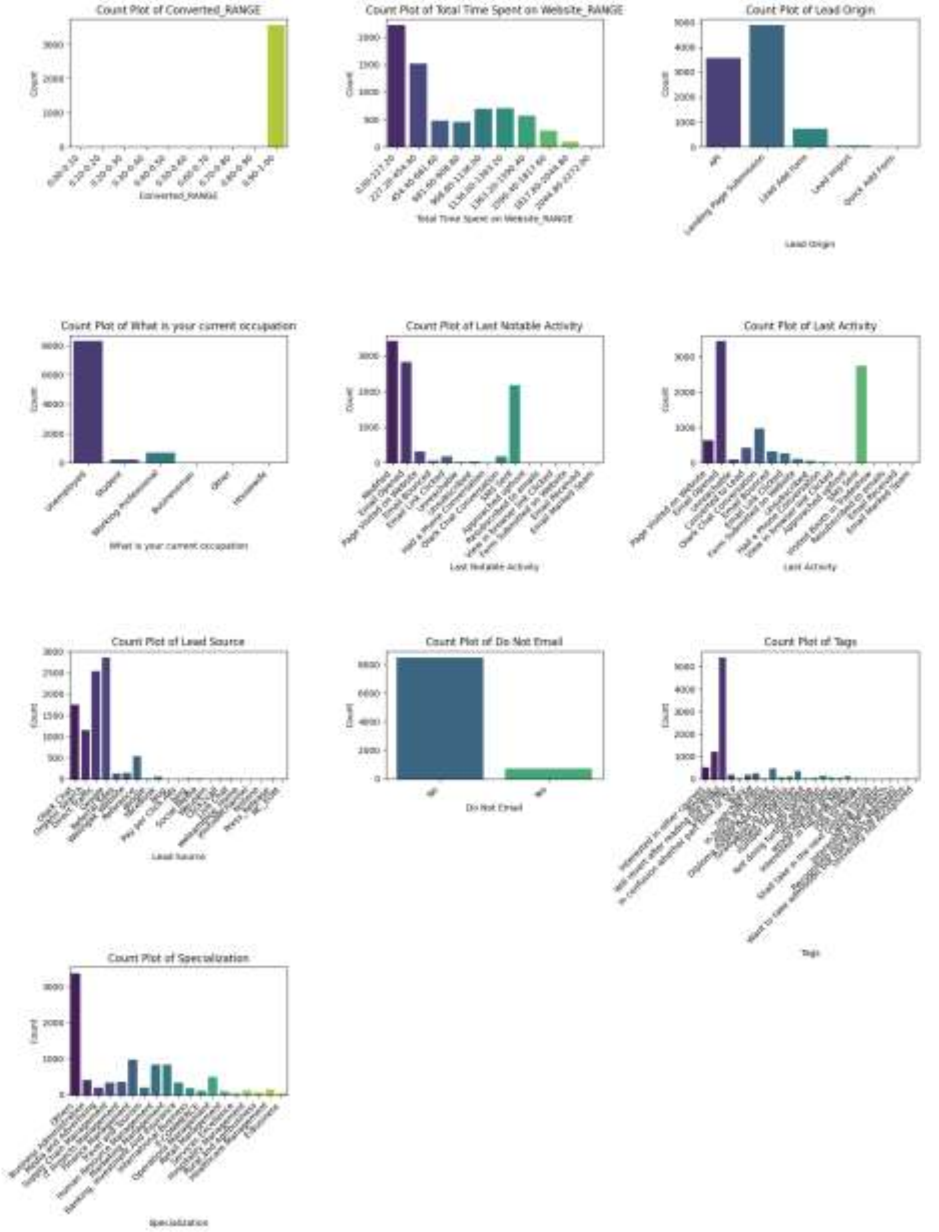


Shows distribution of univariate analysis on numerical Columns





Top N correlated count plots at a shot using subplot Univariate Analysis - Categorical columns



Lead Source: 58% Lead sources is from Google and Direct Traffic Combined

Last Activity : 68% of customers contribution in SMS Sent & Email Opened activites.

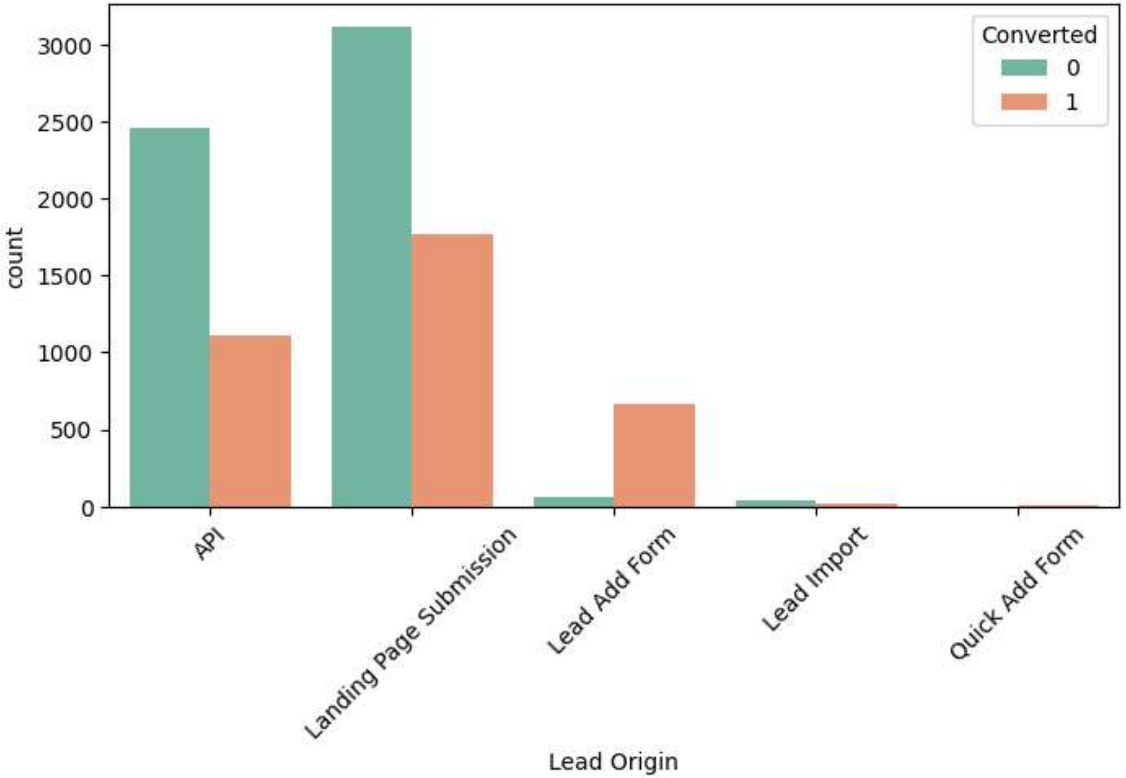
Others Specialization, Will revert, Google search, Email and SMS, Unemployed Count are high Lets Study individual and impute

LEAD ORIGININFERENCE DRAWN:

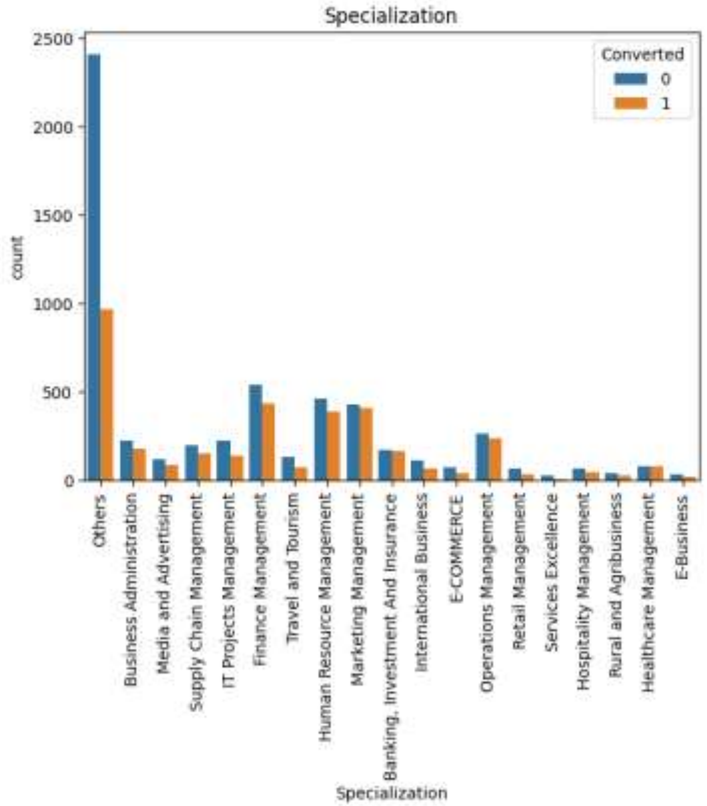
The API and Landing Page Submission have considerable lead counts, whereas the converted leads account to almost 31-35%.

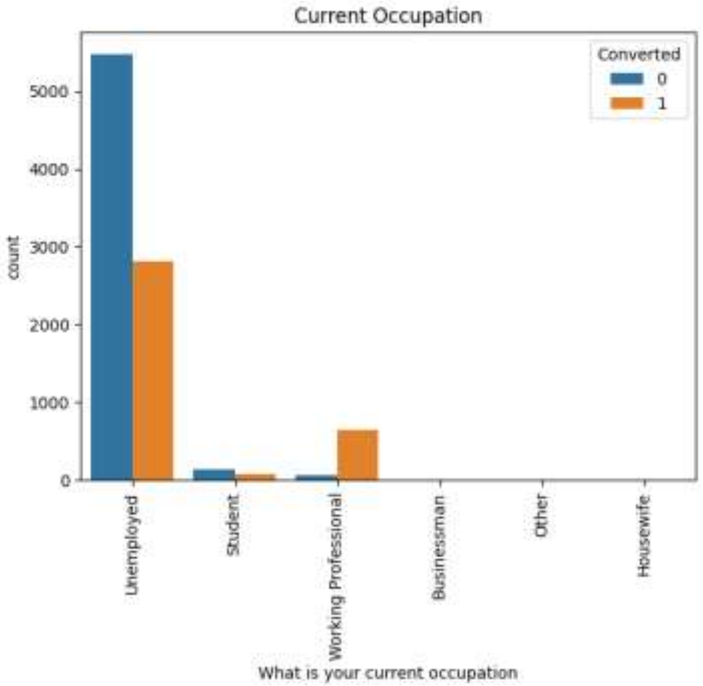
Lead Add form has very high conversion rate, but the count of leads is quite low.

Improving lead conversion rate of API and Landing Page Submission origins via enhancing websites and softwares can help in the business.Since Lead Add Form have higher rate of conversion, focus should be on that too.

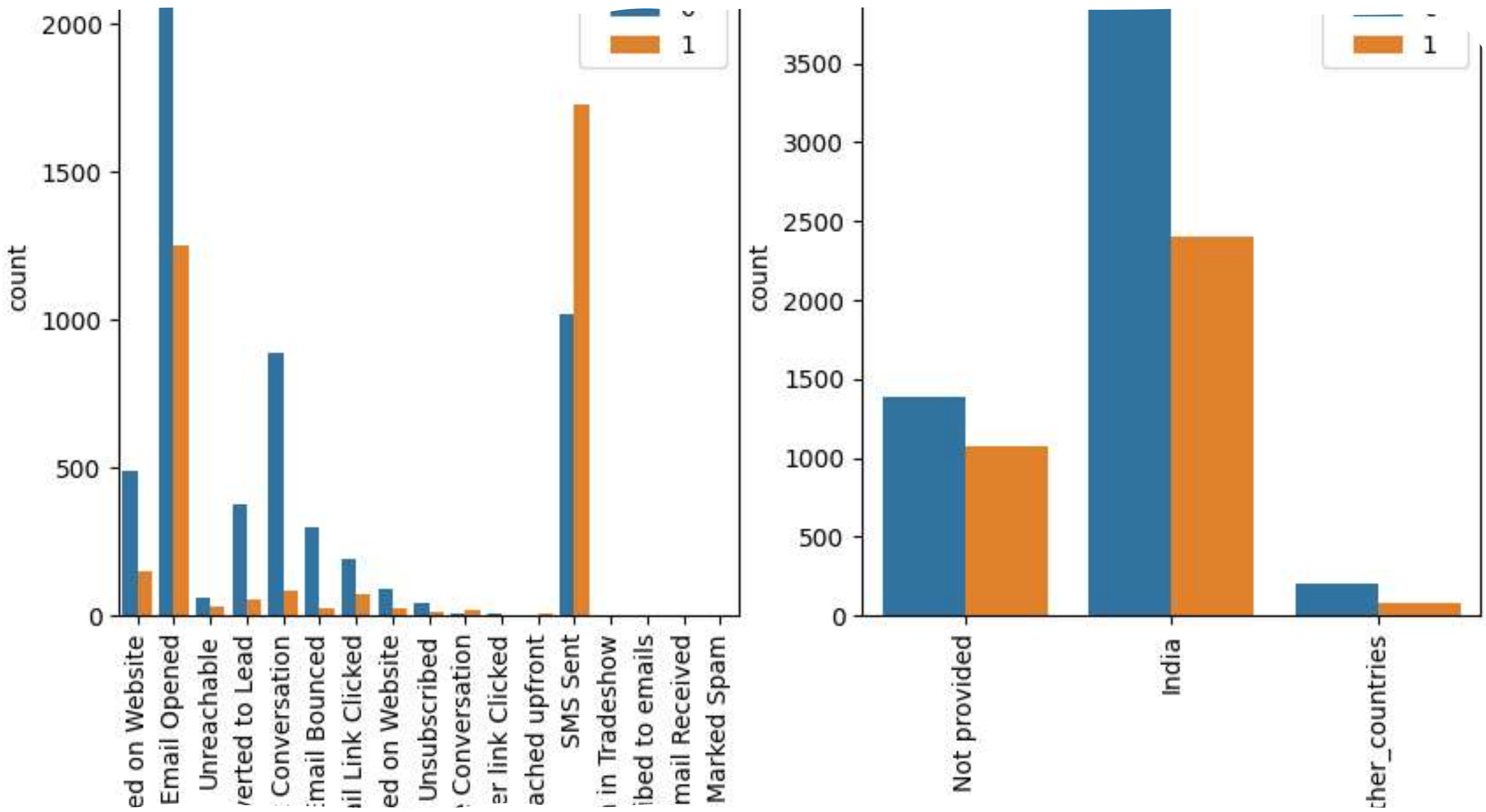


INFERENCE DRAWN-





- Most of the working professional leads converted and almost 30% of unemployed leads converted.
- Focus should be on other specialization apart from the categories in the data which help in generating leads.


count
2000
1500
1000
500
0
1
ed on Website
Email Opened
Unreachable
verted to Lead
Conversation
mail Bounced
il Link Clicked
ed on Website
Unsubscribed
Conversation
er link Clicked
ached upfront
SMS Sent
in Tradeshow
ibed to emails
mail Received
Marked Spam
count
3500
3000
2500
2000
1500
1000
500
0
1
Not provided
India
ther_countries

Data Preparation Before Model building

- Binary Level categorical colums were already mapped to 1/10 in previous steps
- To represent categorical data in numeric format, we create **dummy variables : Lead Origin, Lead Sources, Last Activity, Specialization, Current_occupation.**
- **Test-Train Split**
- **Feature scaling** : Standardization method was used to scale the features : 38.02%
- **Looking at Correlations** : Predictor variables which were highly correlated with each other were dropped (Lead Origin_Lead Import and Lead Origin_Lead Add form)
-

Model Building

- Feature Selection

- The data set has lots of dimension and large number of features
- This will reduce model performance and might take high computation time
- Hence it is important to perform Recursive Feature Elimination(RFE) and to select only the important columns.
- Then we can manually fine tune the model
- RFE outcome : 48 columns & post RFE: 15 columns

Plotting the ROC Curve

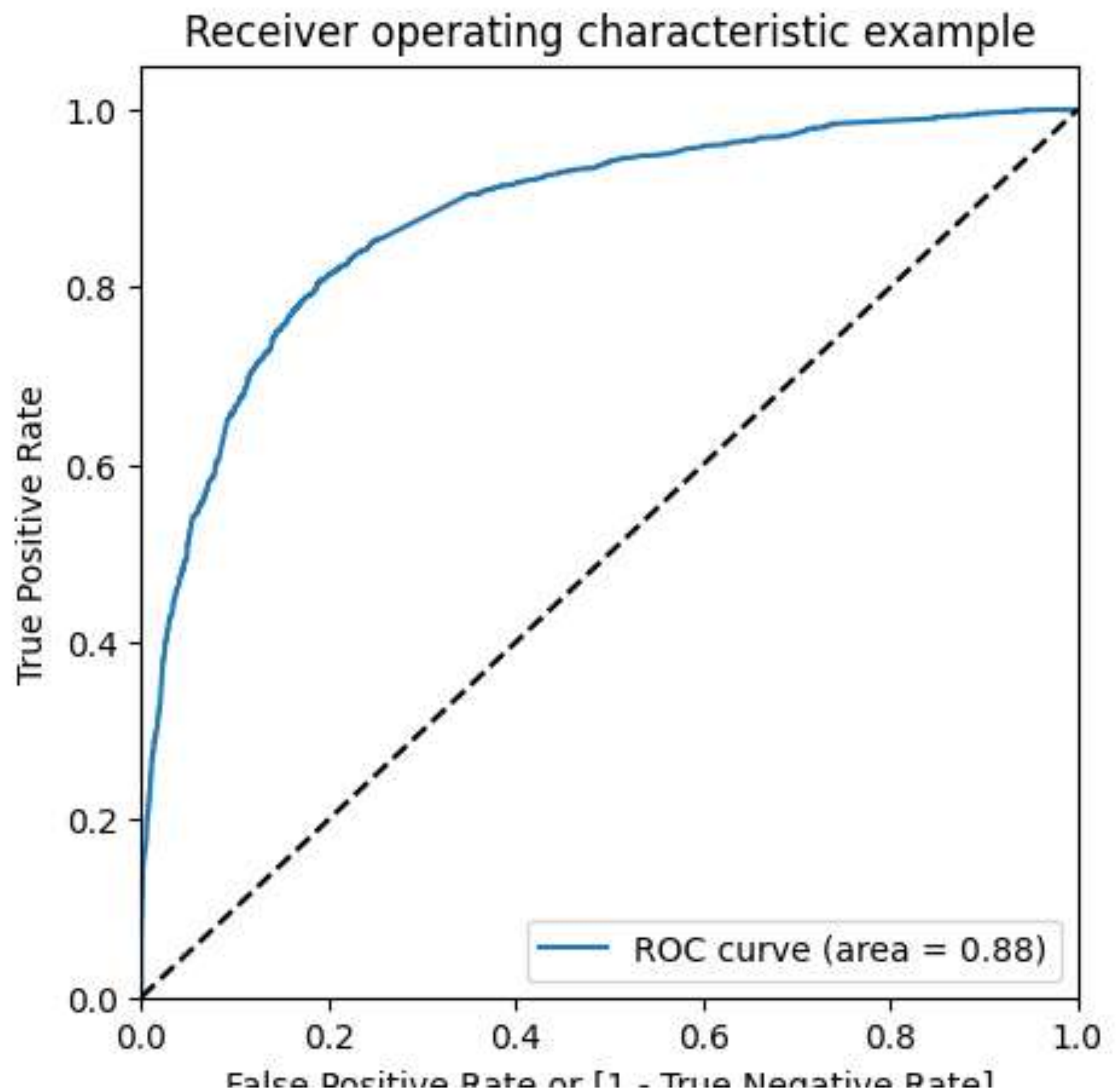
An ROC curve demonstrates several things:

It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

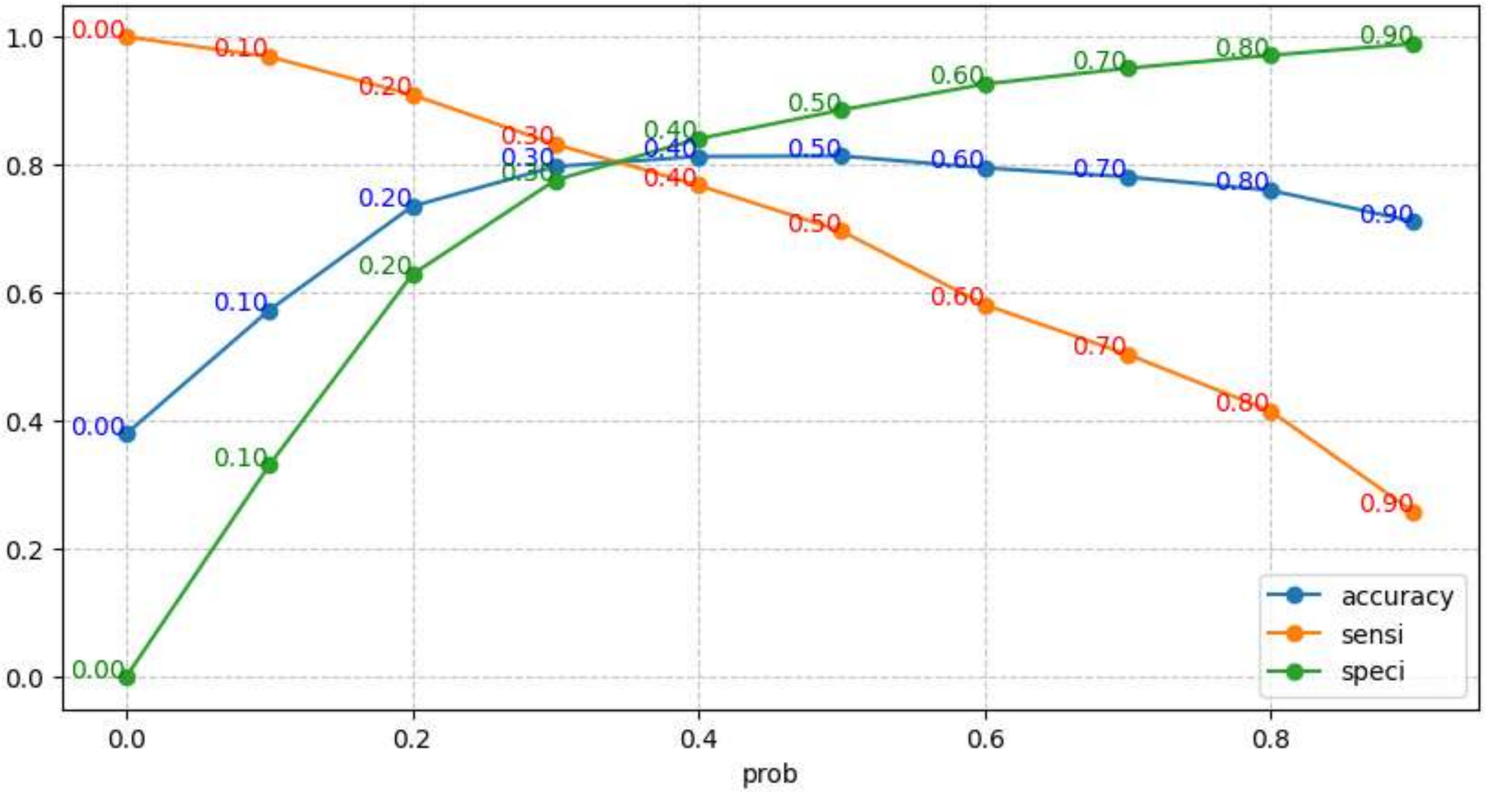
The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

- **Finding Optimal Cutoff Point**
- Optimal cutoff probability is that prob where we get balanced sensitivity and specificity



Module Evaluation

From the curve above, 0.36 is the optimum point to take it as a cutoff probability.



Recommendation based on final model

As per the problem statement, increasing lead conversion is crucial for the growth and succss of X education.

We have determined the following features that have the highest positive coefficients, and these features should be given priority in our marketing and sales efforts to increase lead conversion

Lead Source_welingak websites : 1.32

Lead source_Reference:0.83

Current_occupation_working professional:1.14

Last activity_SMS sent:1.26

Last Activity_others:1.27

Total time spent on website:1.27

Last Activity_Email opened:0.94

Lead source_Olark chat:0.9

We have also identified features with negative coefficients that may indicate potential areas for improvements, These include

Specialization in Hospitality Management:0.01

Lead origin of landing page submission: -0.75