Table of contents

Background of X Education company

- An education company named xEducation sells online courses to industry professionals.
- On any given day, many professionals who are interested in the courses land on their browse for courses.
- The company markets its courses on several websites and search engines like google.
- Once's 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

Problem statement:

- X education gets a lot of lead, 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 built a model where in we need to assign a lead score to each of the leads.
- The CEO has given a ballpark of the target lead conversion rate to be around 80%.

Suggested ideas for lead conversion

1. Prepare Quality Content

Adding quality content to your website significantly increases the chances of getting more leads for your business.

2. Use Social Media Channels to Generate Leads

Today people are massively engaged on social media platforms like Facebook, Twitter, Instagram, LinkedIn, and many other ones.

3.Create an Appealing Landing Page Design

Your landing page design is the first and foremost element that defines the usability and success of your website

Analysis Approach

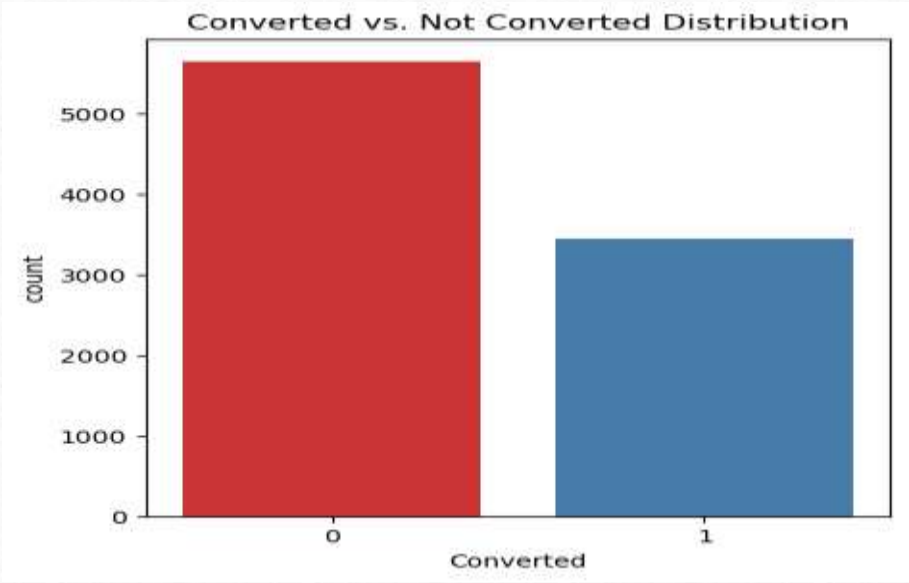
- “Reading Data”—”Cleaning Data”—”Exploratory Data Ananlysis”(EDA)—”Creating Dummy”—”Splitting data into test and train set”—”Building Model”—”Making Prediction”—”Model Evaluation”—”ROC Curve”—”Prediction on test set”—”Precision-Recall”

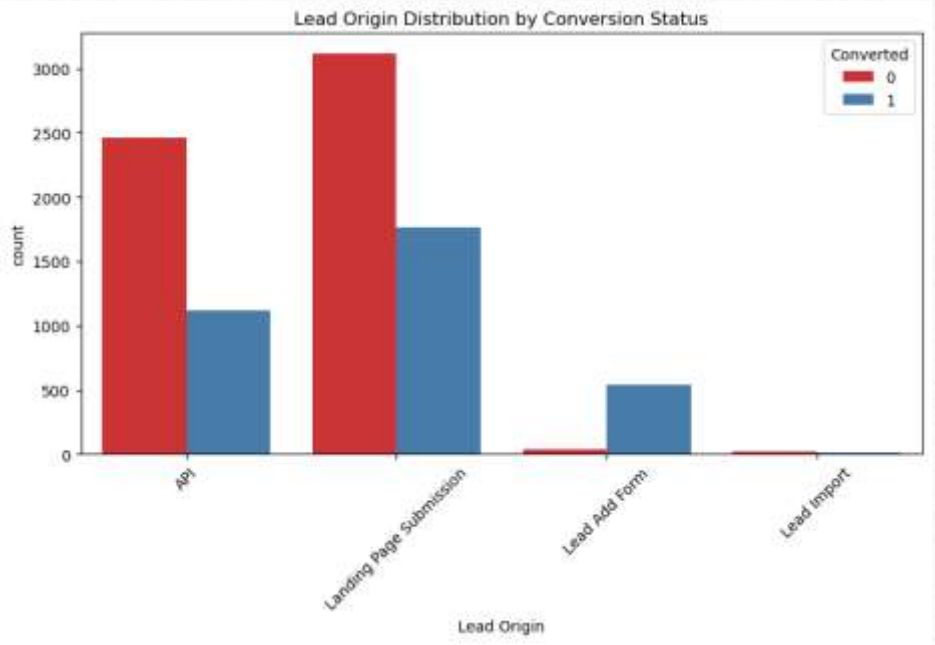
Data Cleaning

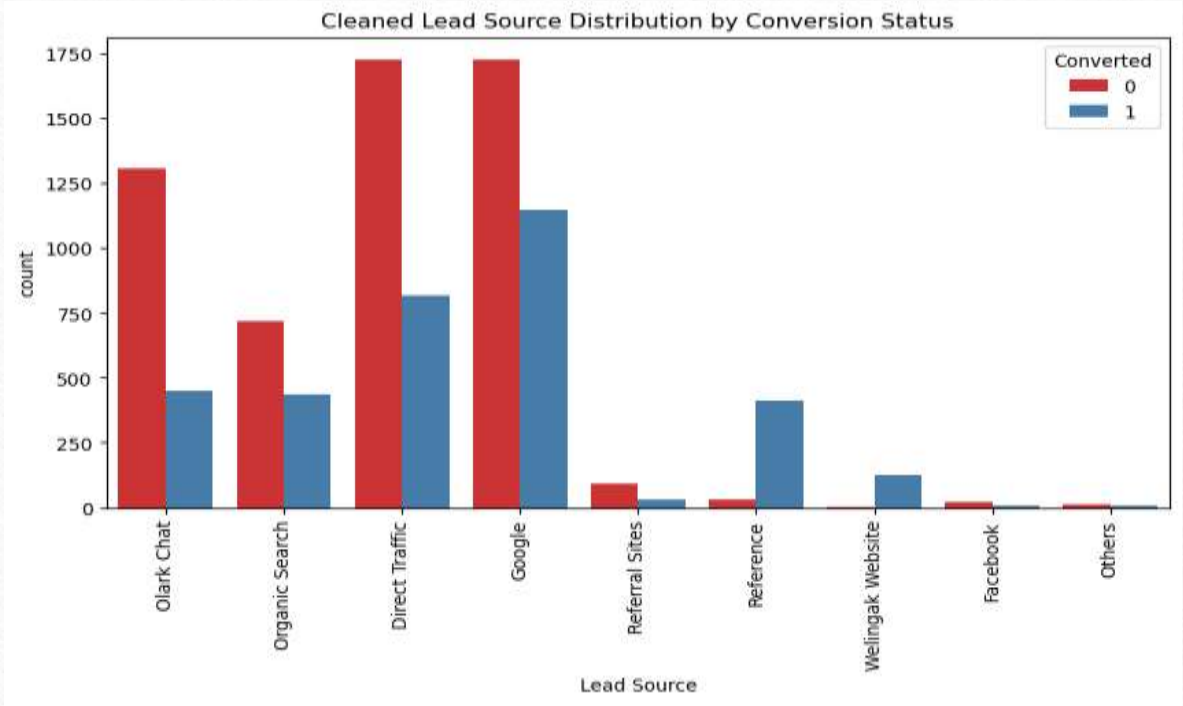
- Select level represents null values for some categorical variables, as customers did not choose any option from the list
- Columns with over 40% null values were dropped.
- Missing values in categorical columns were handled based on values counts and certain considerations.
- Skewed category columns were checked and dropped to avoid bias in logistic regression models.
- Binary categorical variables were mapped.

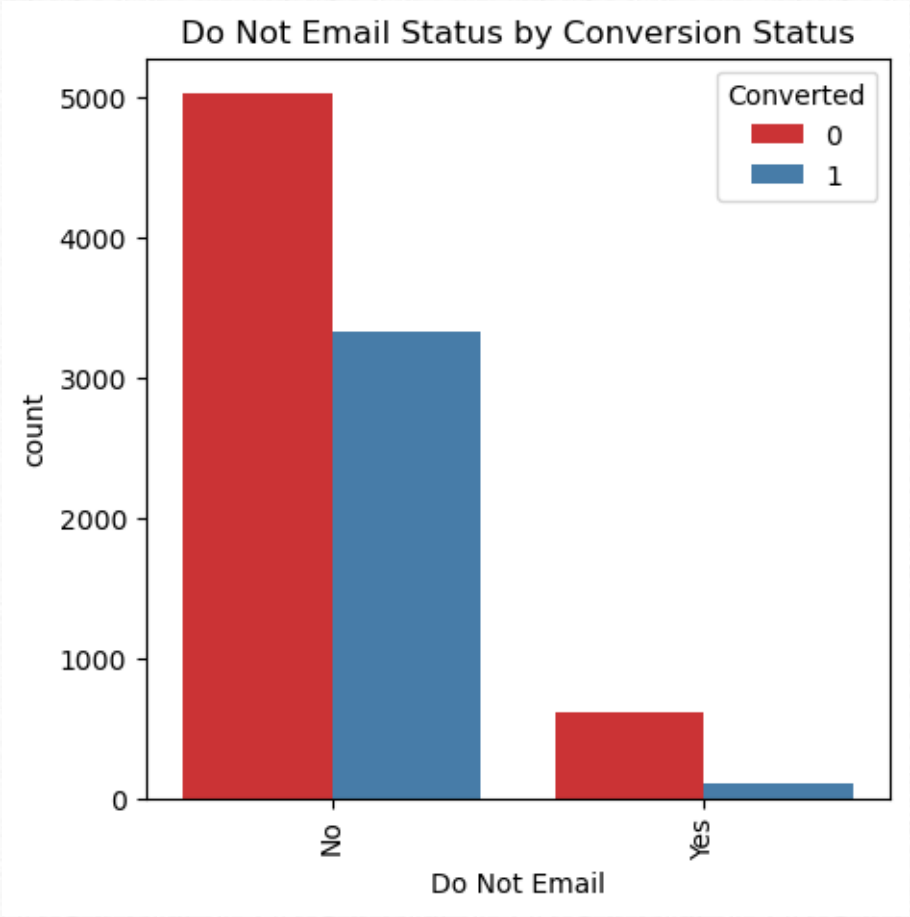
EDA

- Data is imbalanced while analyzing target variable

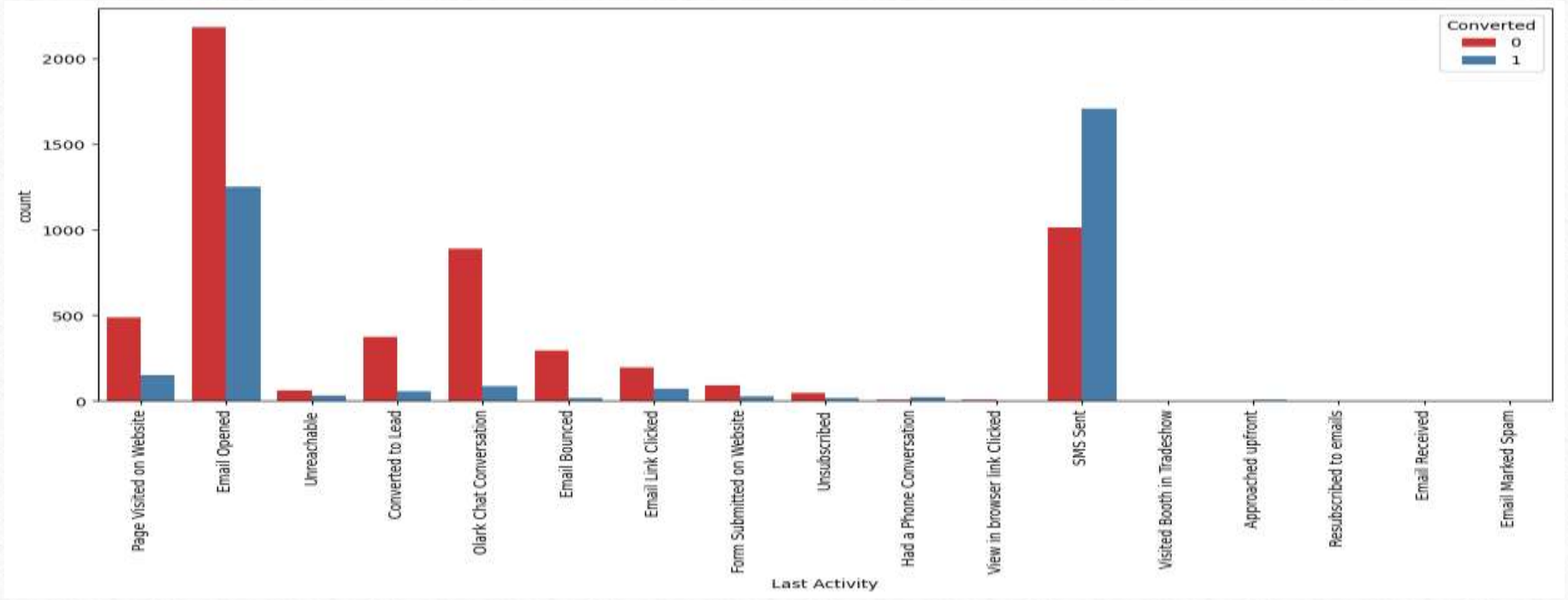


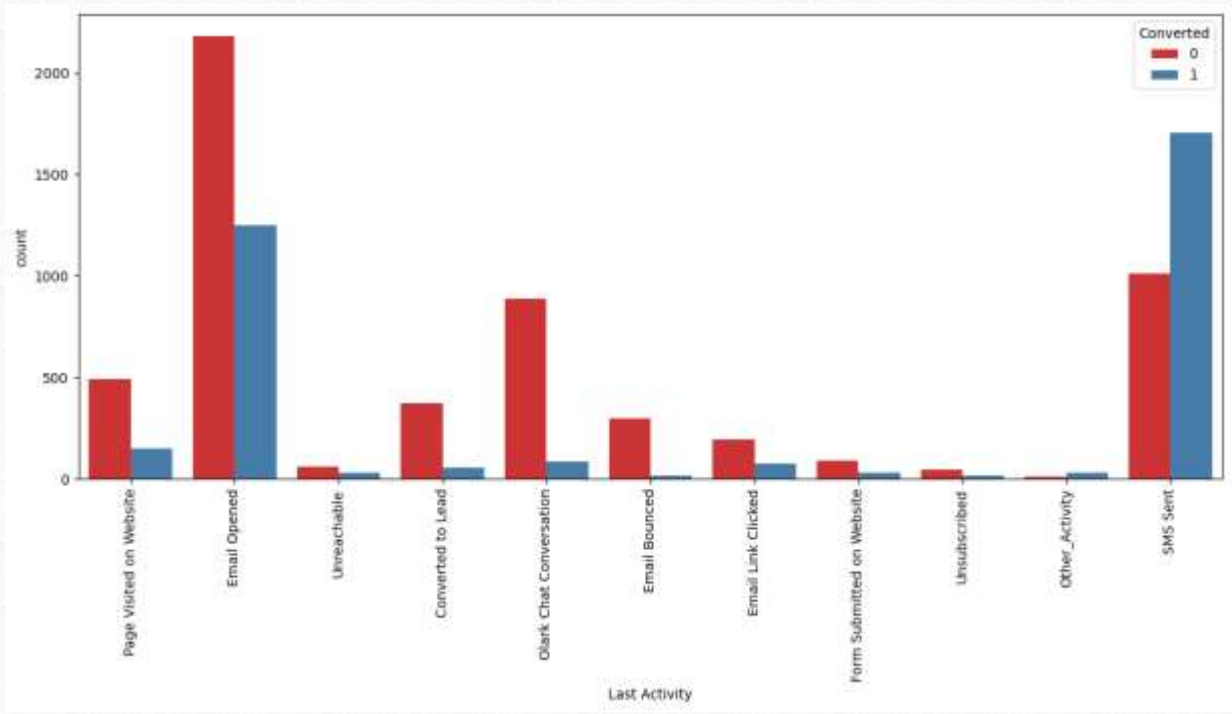


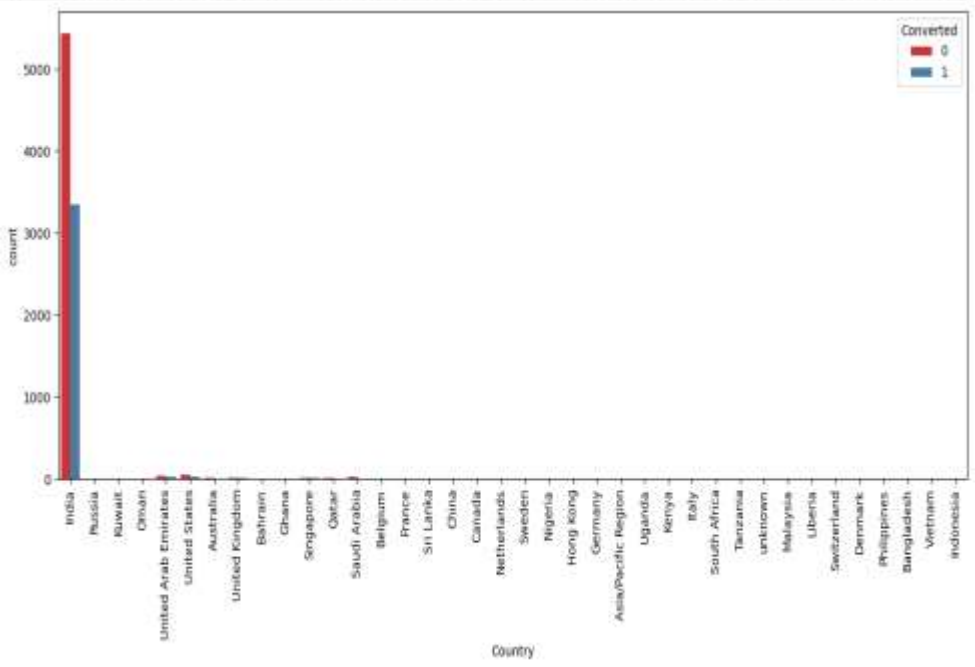

Cleaned Lead Source Distribution by Conversion Status
Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
Welingak Website
Facebook
Others
Lead Source

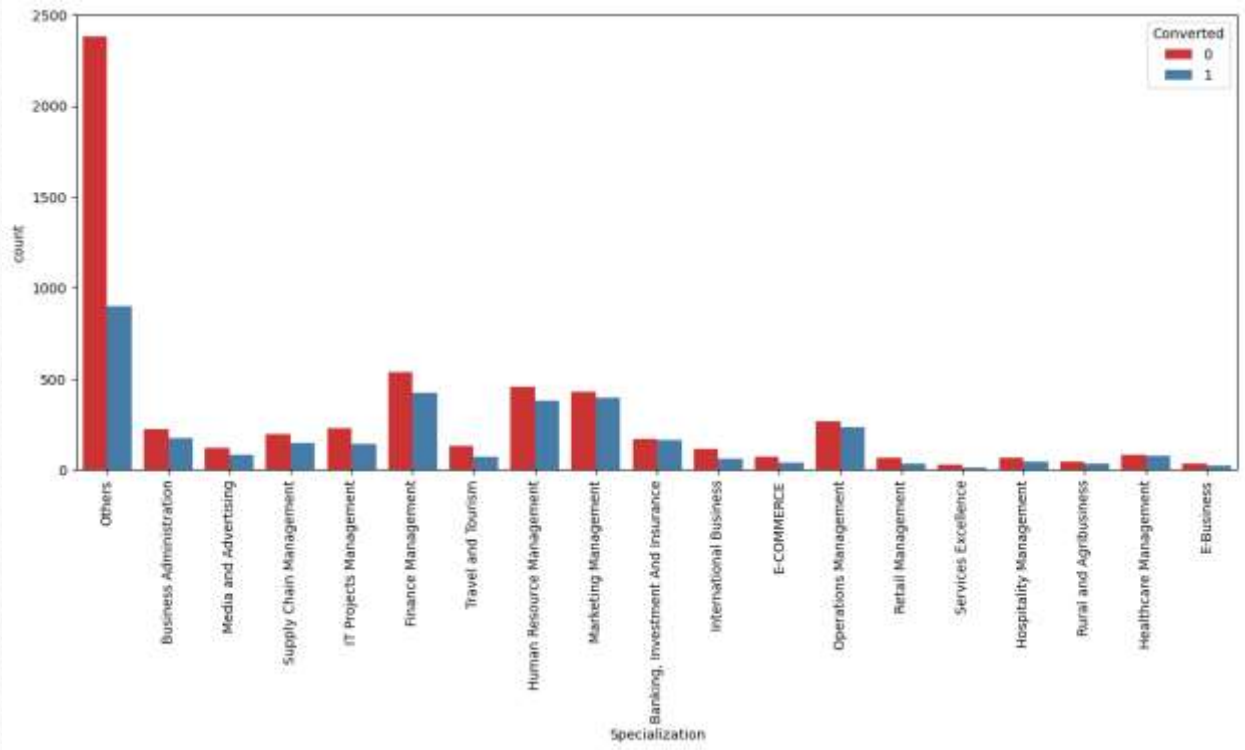

Do Not Email Status by Conversion Status
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Email

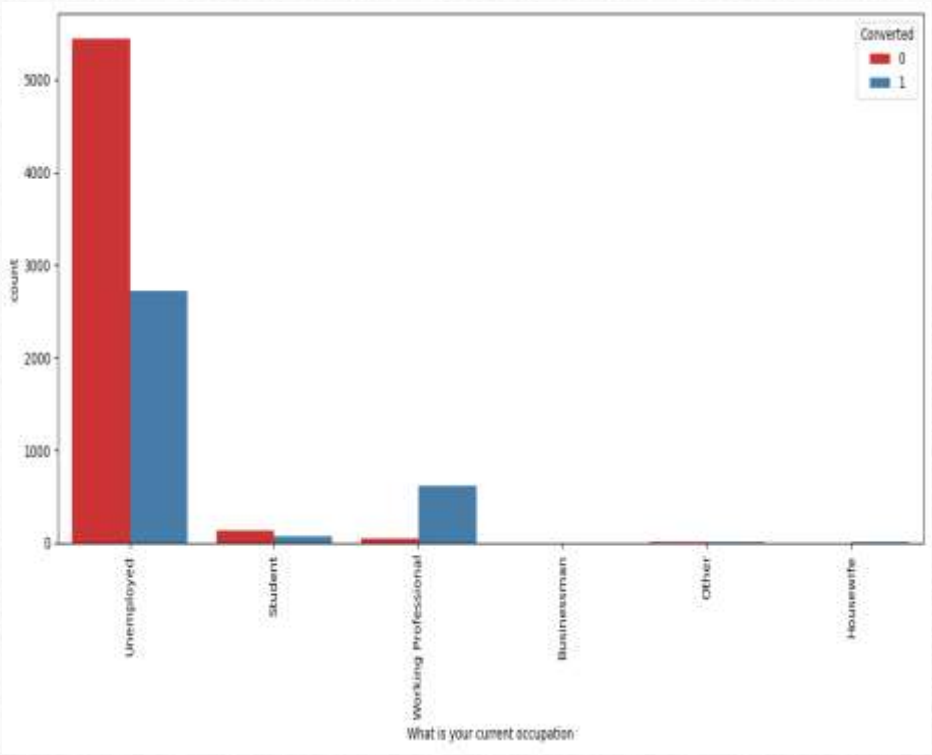
EDA- Bivariate Analysis for categorical variables



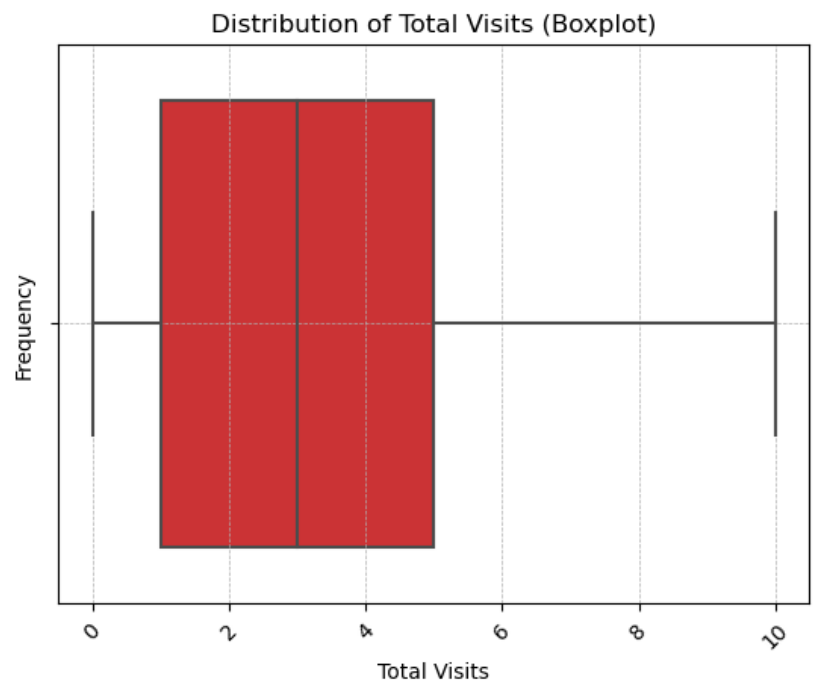

Converted
0
1
count
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Other_Activity
SMS Sent
Last Activity

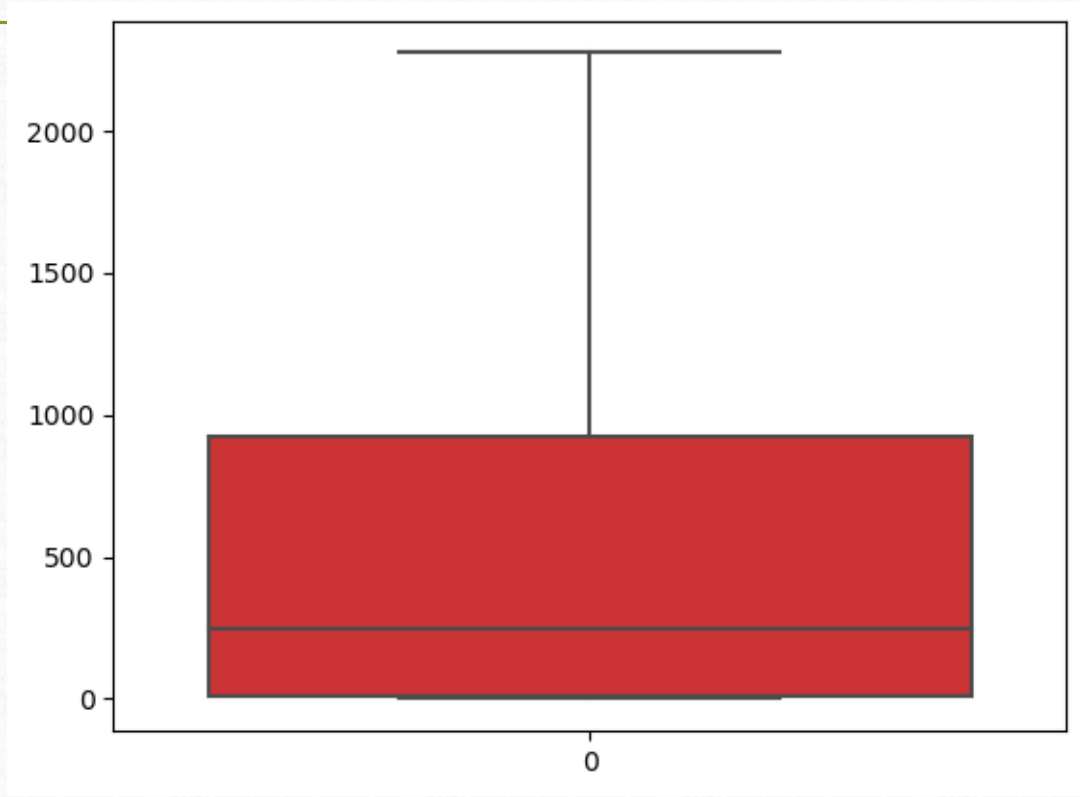

Converted
0
1
count
5000
4000
3000
2000
1000
0
India
Russia
Kuwait
Oman
United Arab Emirates
United States
Australia
United Kingdom
Bahrain
Ghana
Singapore
Qatar
Saudi Arabia
Belgium
France
Sri Lanka
China
Canada
Netherlands
Sweden
Nigeria
Hong Kong
Germany
Asia/Pacific Region
Uganda
Kenya
Italy
South Africa
Tanzania
unknown
Malaysia
Liberia
Switzerland
Denmark
Philippines
Bangladesh
Vietnam
Indonesia
Country

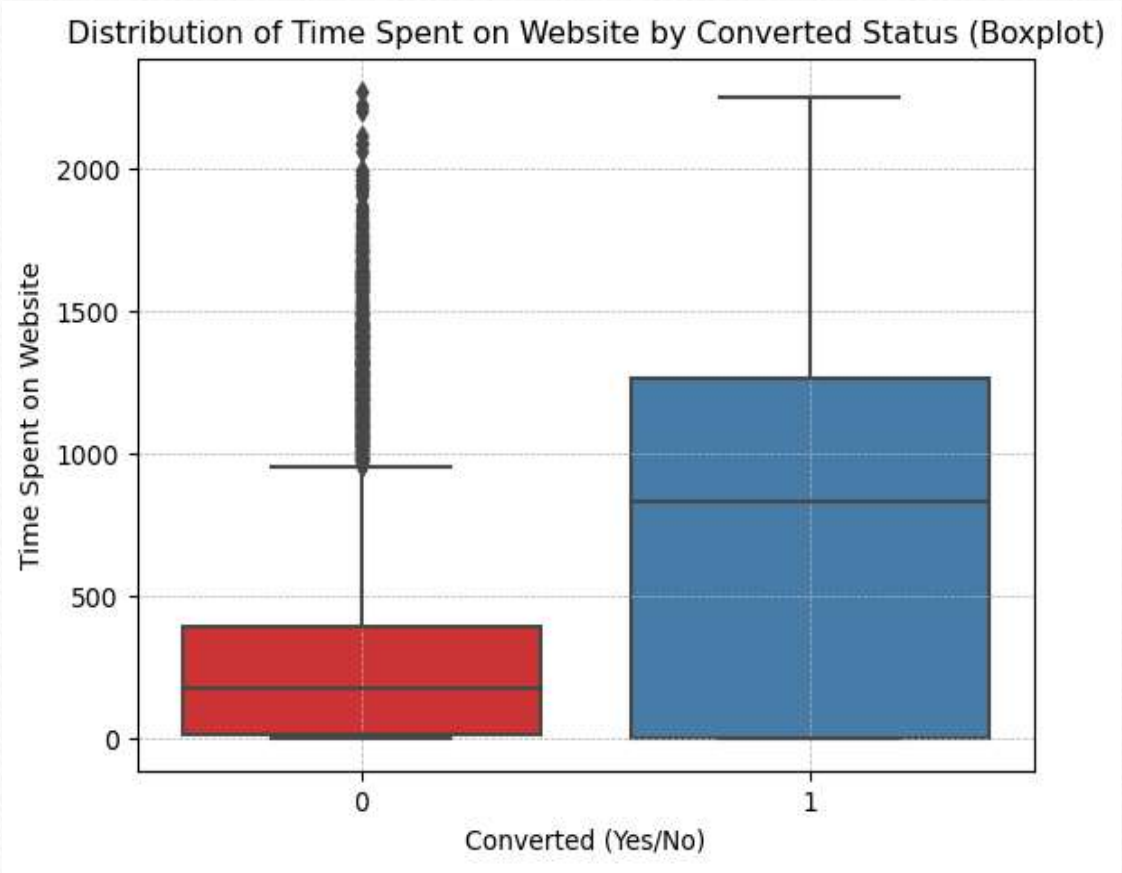

Converted
0
1
count
2500
2000
1500
1000
500
0
Others
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization

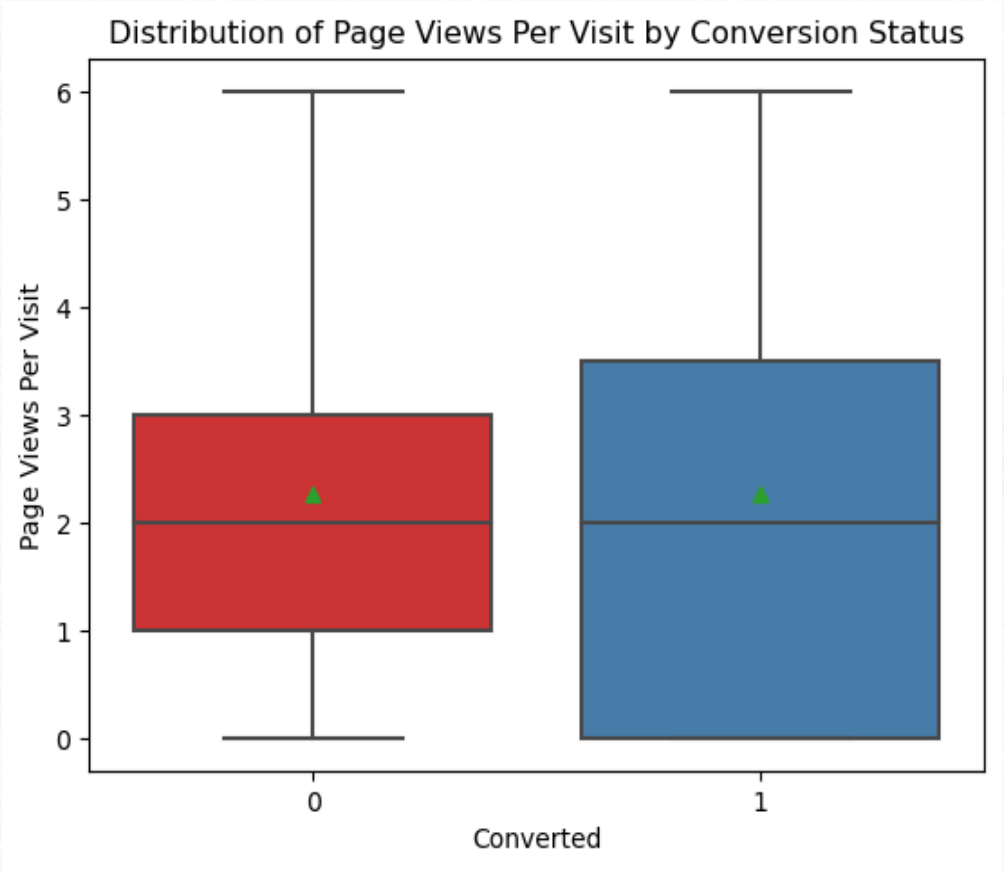

Converted
0
1
count
5000
4000
3000
2000
1000
0
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation

EDA- Bivariate Analysis for numerical variable






Distribution of Time Spent on Website by Converted Status (Boxplot)
Time Spent on Website
2000
1500
1000
500
0
0
1
Converted (Yes/No)


Distribution of Page Views Per Visit by Conversion Status
Page Views Per Visit
6
5
4
3
2
1
0
0
1
Converted

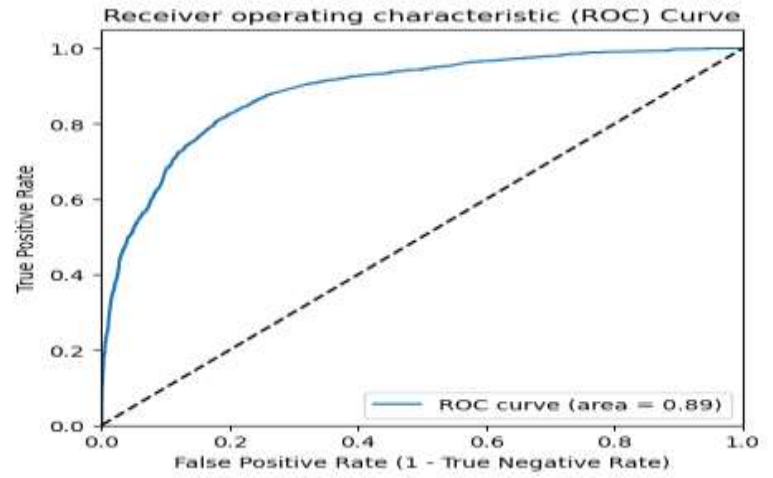
Data preparation before model building

- Binary level categorical columns were already mapped 1/0 in previous steps
- Created dummy features (one-hot encoded) for categorical variables – lead origin, lead source, last activity, specialization, current_occupation.
- Feature scaling – standardization method was used to scale the features.
- Checking the correlations – predicator variables which were highly correlated with each other were dropped.

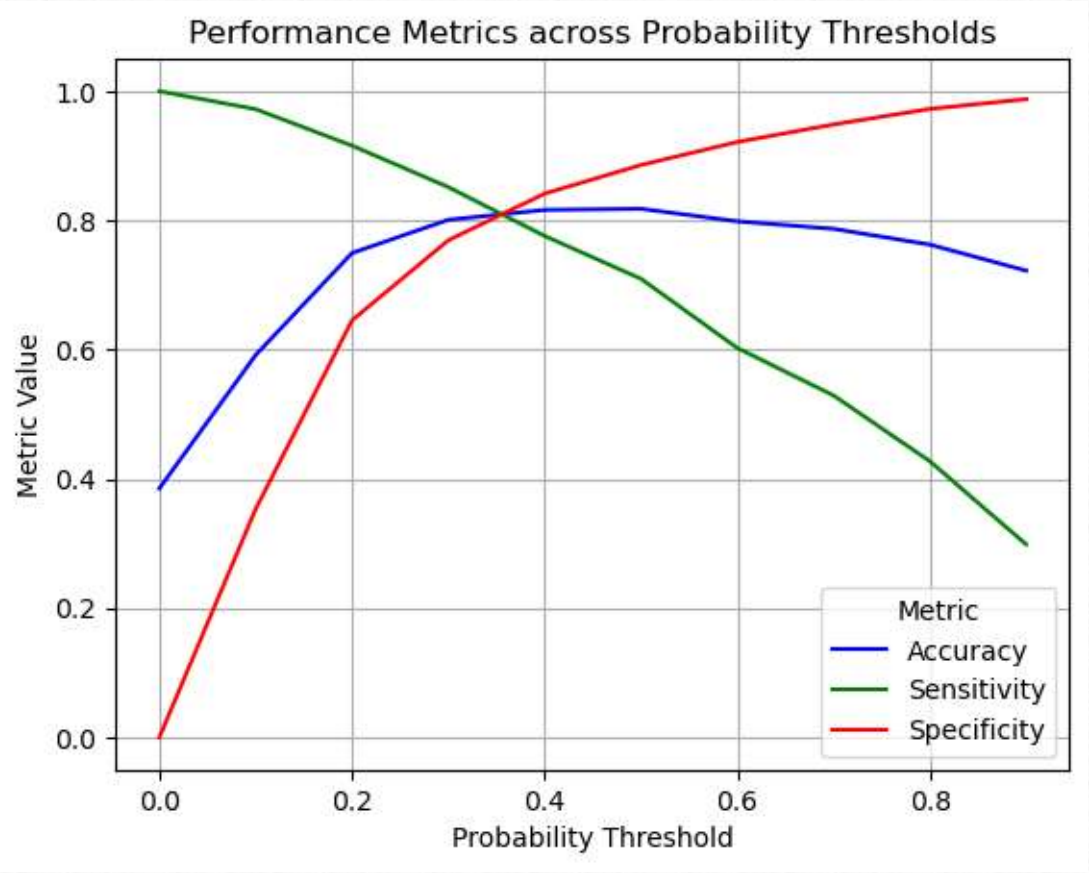
Model Building

- The data set has lots of dimension and large number of features.
- This will reduce model performance and might take high computation time.
- Then we manually fine tune the model
- Hence it is important to perform **Recursive feature elimination** and to select only the important columns.

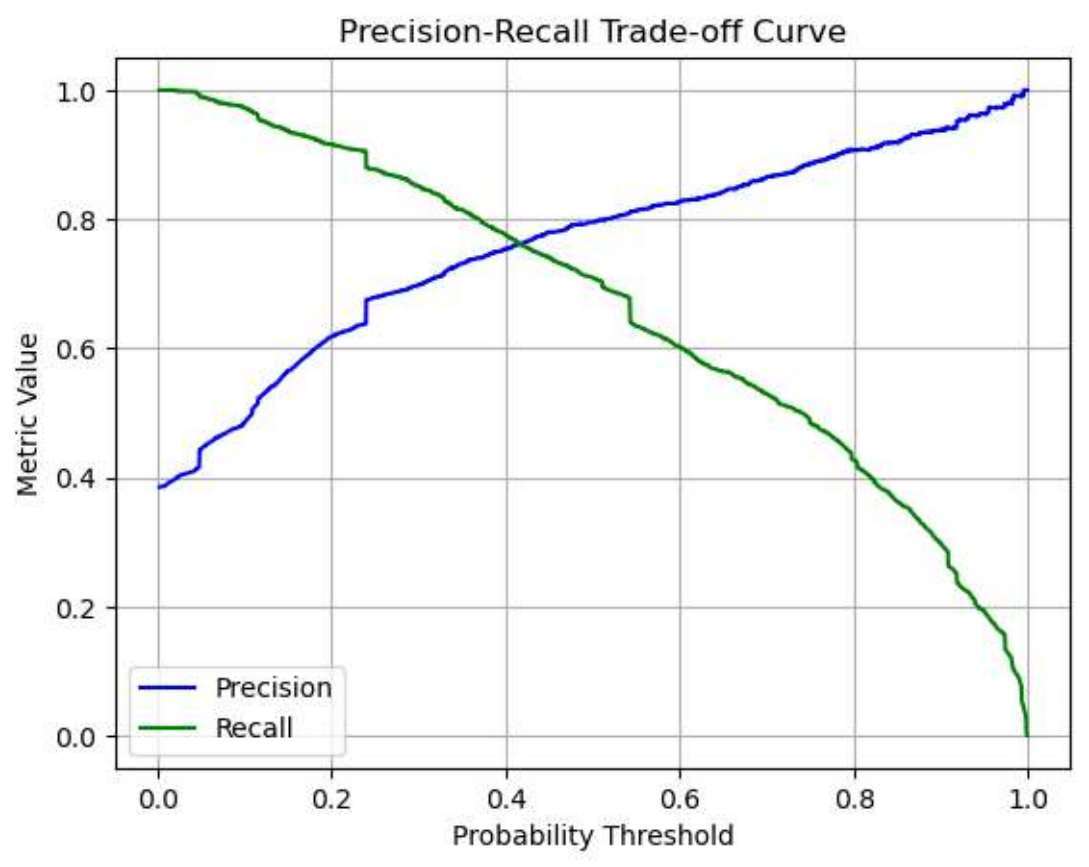
Model evaluation



- The model achieved a promising AUC score of 0.89, indicating good overall performance in distinguishing converted and non-converted prospects. However, it's essential to evaluate other performance metrics (e.g., precision, recall, F1-score) and consider their suitability for your specific task and decision-making criteria. Additionally, analyzing the trade-off between sensitivity and specificity at different thresholds along the ROC curve can provide insights into potential improvements.



- ***The analysis suggests that 0.34 might be the best cutoff point for identifying converted prospects.***



- ***"This plot depicts the inverse relationship between precision and recall across various probability thresholds. Higher recall indicates better success in capturing true positives, but it might come at the cost of lower precision due to potential false positives."***

Recommendation for X education

Prioritize lead outreach based on conversion likelihood:

Do call(high conversion potential):

- Leads from "**welingak websites**" and "**References**"sources "**working professionals**".
- Leads who spent "**more time on websites**".
- Leads from "**olark chat**"
- Leads with last activity of "**SMS sent**"

Avoid calls(Low conversion potential)

- Leads with last activity of "**Olark chat conversation**"
- Leads from "**Landing page submission**" origin.
- Leads with "**others**" specialization.
- Leads who chose " **Do not Email**" as "**Yes**" specialization

_____THANK YOU_____