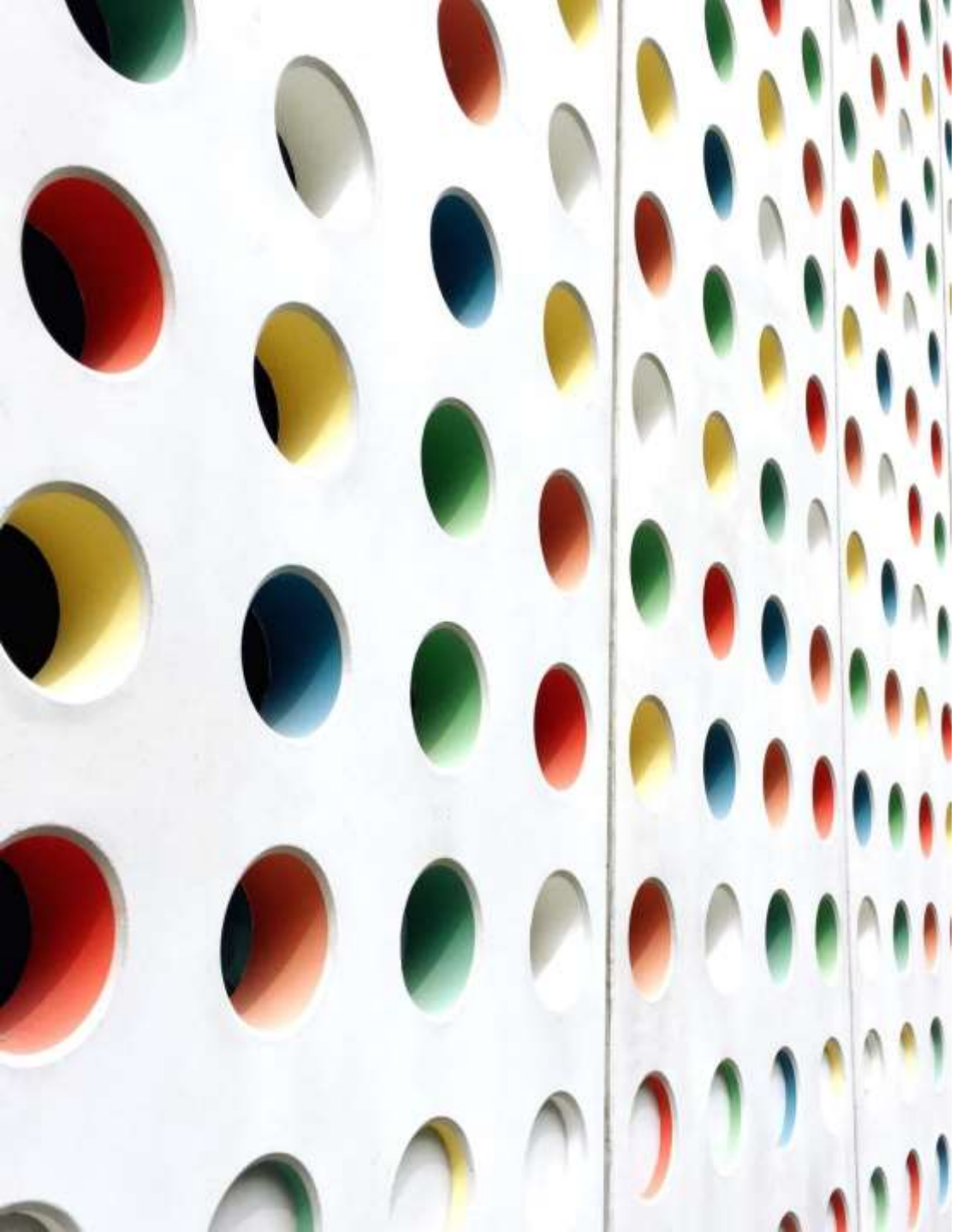


Lead Scoring case Study

Business Problem Statement:

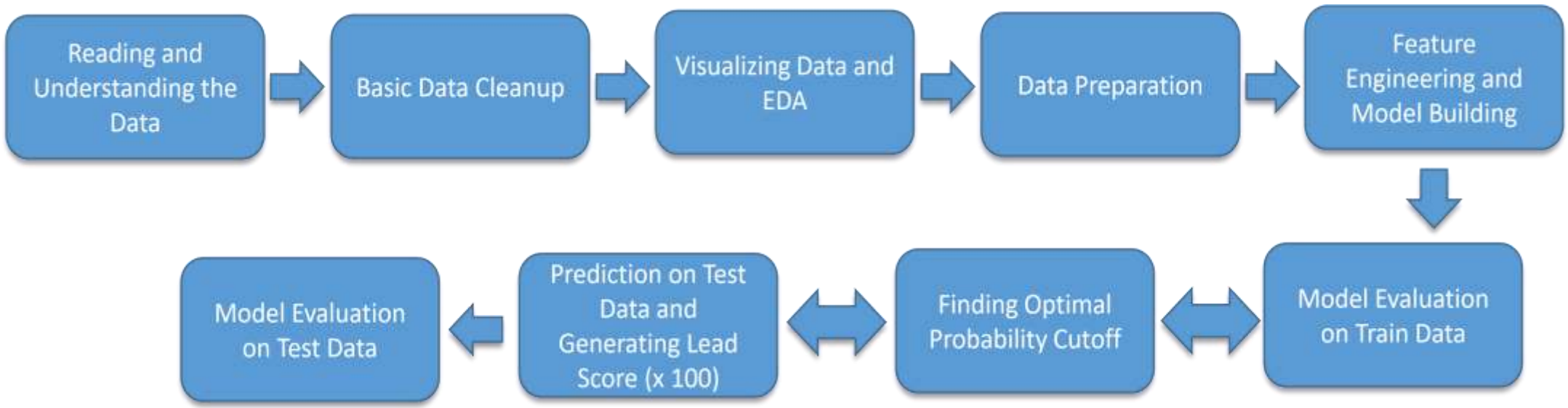
An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

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. Moreover, the company also gets leads through past referrals. 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%.

Goal:

1. To identify the features that contributes to predict Lead Conversion.
2. Identifying Hot Leads by generating Lead Score for all leads, so that leads having higher Lead Scores can be contacted with priority for achieving Higher Lead Conversion Rate.

Overall Approach



Understanding the Data & Basic Data Cleanup:

•There are 37 columns (30 categorical and 7 Numeric) and 9240 observations in the dataset.
•Select is present as a class in different columns like:
- Specialization
- How did you hear about X Education
- Lead Profile
- City

- As Select is not a valid class, we can conclude that the Select might be the default value set in the form dropdown and if the user has not selected any option from the dropdown, then the value remained as Select. We replaced Select with NaN.

- **Magazine, Receive More Updates About Our Courses, Update me on Supply Chain Content, Get updates on DM Content, I agree to pay the amount through cheque** - These columns have no missing data and have only one unique value. So, these columns have no variance and not helpful for our EDA or model building, hence we dropped these columns.

- **How did you hear about X Education, Lead Profile, Lead Quality, Asymmetries Activity Index, Asymmetries Profile Index , Asymmetries Activity Score, Asymmetries Profile Score** – These columns have more than 40% missing value. So, we have dropped these columns from our EDA and model building.

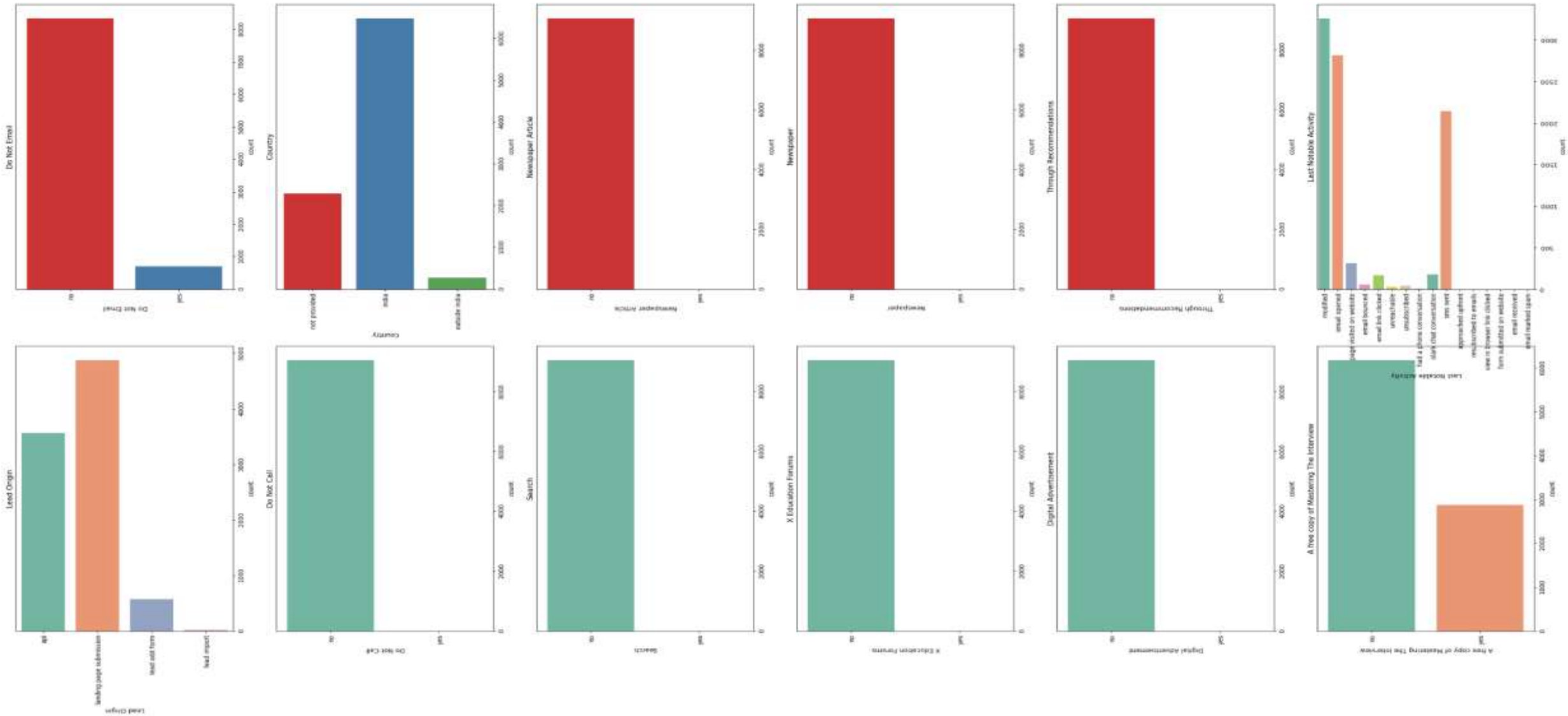
- There is no datapoint/ observation (rows) in our dataset having more than 70% missing values.

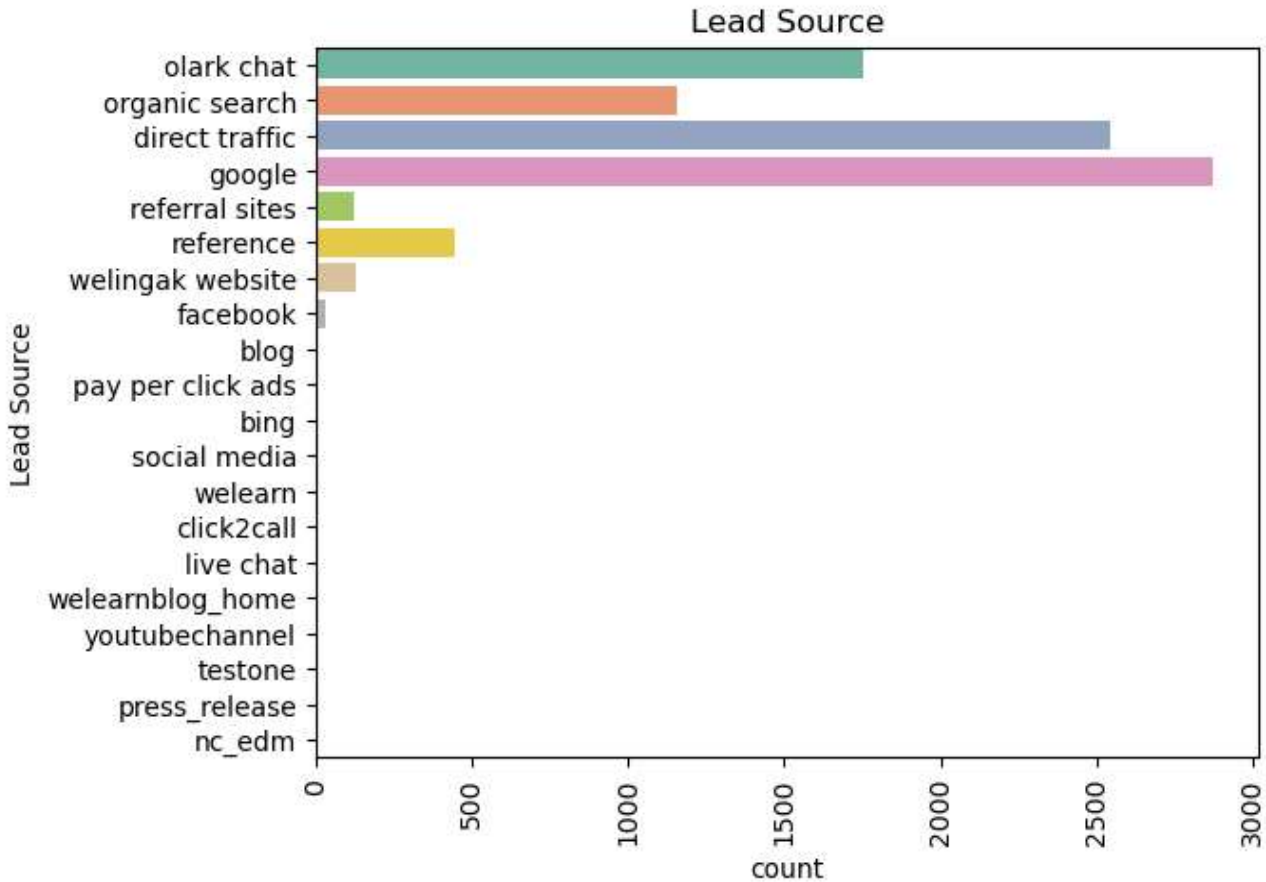
- We have created new buckets/bins for the categorical variables having very high numbers of classes with few datapoints: **Lead Origin, Lead Source, Last Activity**, **Last Notable Activity**, **Country**, **Specialization**, **What is your current occupation.**

- Performed missing value treatment using **Business Understanding.** For **Specialization** and **Occupation** NaN values are replaced with a new category **Not Disclosed.**

- We renamed **What is your current occupation** column to **Occupation** and **What matters most to you in choosing a course** to **Reason_choosing** for our convenience during EDA and Model building .

Data analysis

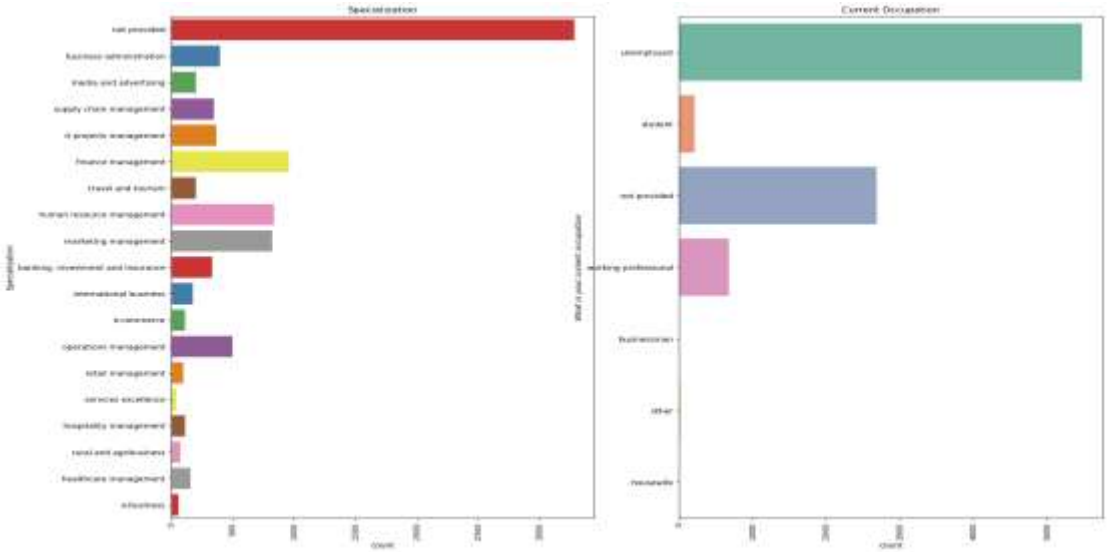


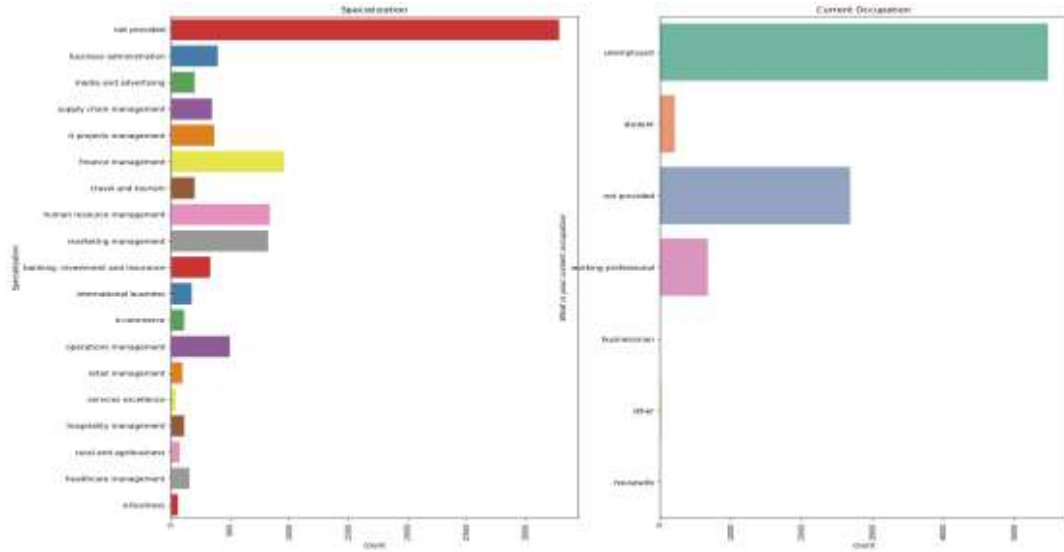

Lead Source
olark chat
organic search
direct traffic
google
referral sites
reference
welingak website
facebook
blog
pay per click ads
bing
social media
welearn
click2call
live chat
welearnblog_home
youtubechannel
testone
press_release
nc_edm
Lead Source
0
500
1000
1500
2000
2500
3000
count

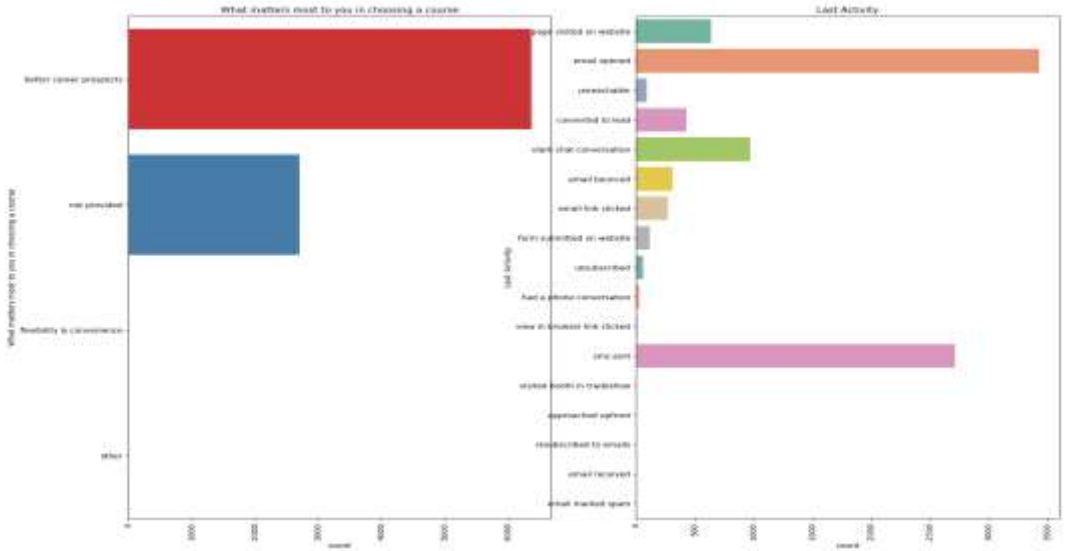
Inferences:

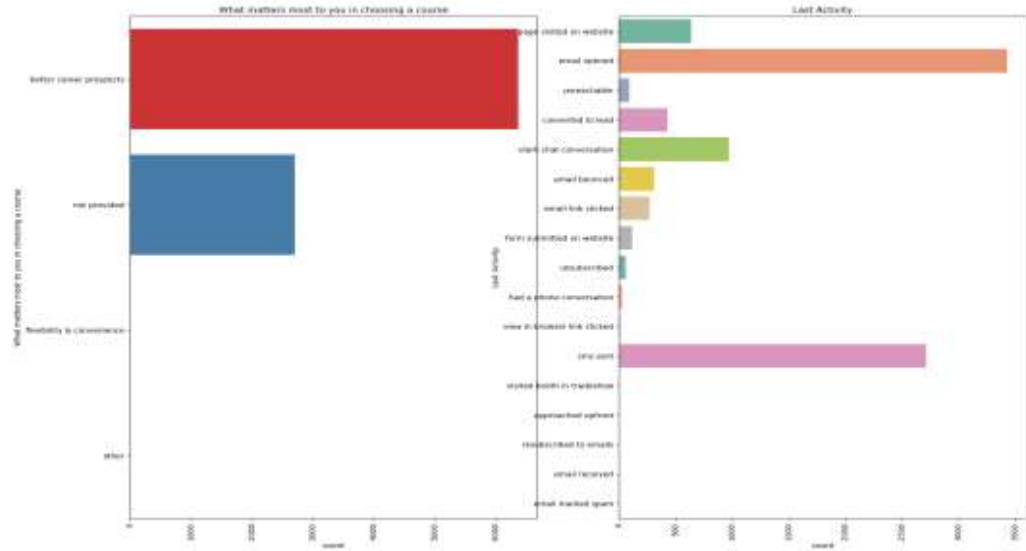
Dominant Lead Sources:

- Certain lead sources, such as “Google” and “Direct Traffic,” seem to have significantly higher counts compared to others. This suggests they are the primary drivers of leads.
- Other sources like “Referral Sites” and “Organic Search” contribute moderately, while some sources like “Press Release” and “YouTube” have minimal impact.
- They may focus more on optimizing high-performing sources (e.g., Google, Direct Traffic).
- They might also consider improving underperforming lead sources by allocating more budget to paid ads or social media marketing.


Specialization
not provided
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
ecommerce
operations management
retail management
service excellence
hospitality management
rural and agribusiness
healthcare management
e-business
Specialization
count
Current Occupation
unemployed
student
not provided
working professional
businessman
other
housewife
What is your current occupation
count


Specialization
not provided
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
ecommerce
operations management
retail management
service excellence
hospitality management
rural and agribusiness
healthcare management
e-business
Specialization
count
Current Occupation
unemployed
student
not provided
working professional
businessman
other
housewife
What is your current occupation
count


What matters most to you in choosing a course
better career prospects
not provided
flexibility & convenience
other
What matters most to you in choosing a course
count
Last Activity
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
count

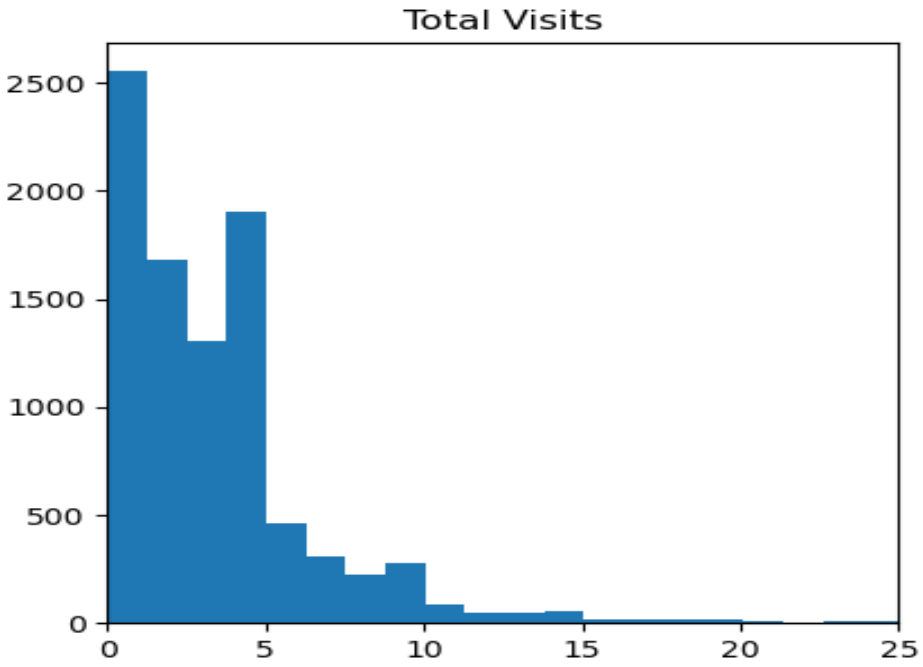

What matters most to you in choosing a course
better career prospects
not provided
flexibility & convenience
other
What matters most to you in choosing a course
count
Last Activity
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
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
count

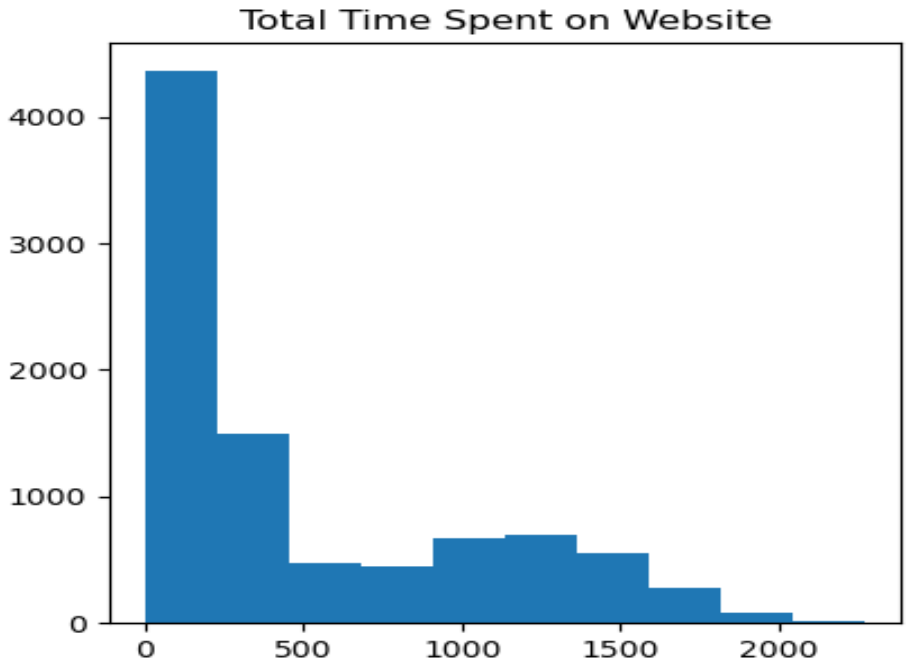
- **Distribution Insights:**

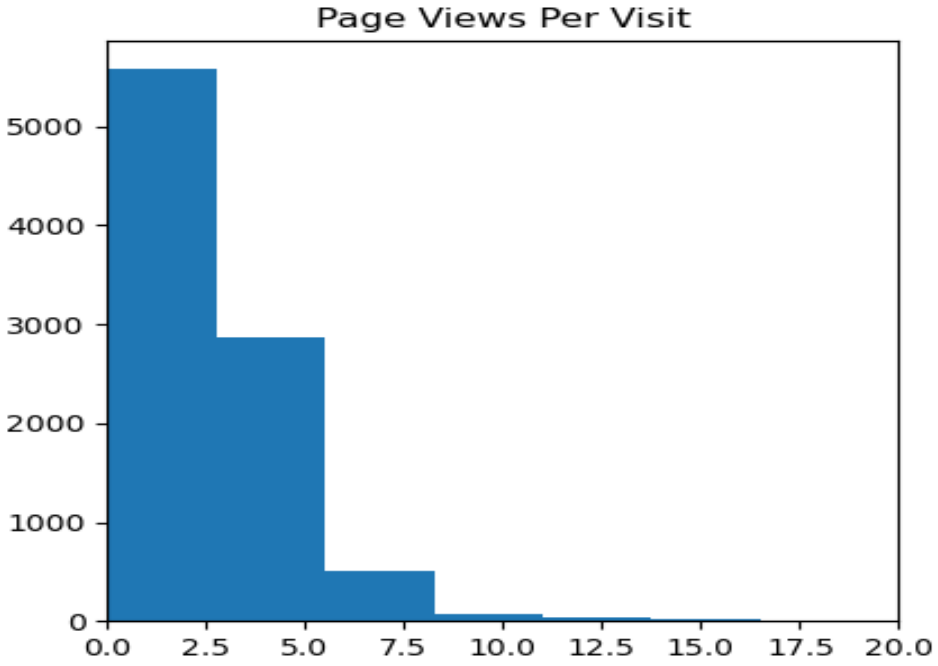
- The charts on the left have many small bars, which might indicate a long-tail distribution where a few categories contribute significantly while others have minimal impact.
- The charts on the right seem to have more concentrated data, possibly showing the most relevant categories.

- **Potential Business Insights:**

- The dominant bars could indicate the most successful channels for customer acquisition.
- The smaller bars might highlight underperforming areas that need strategic changes.





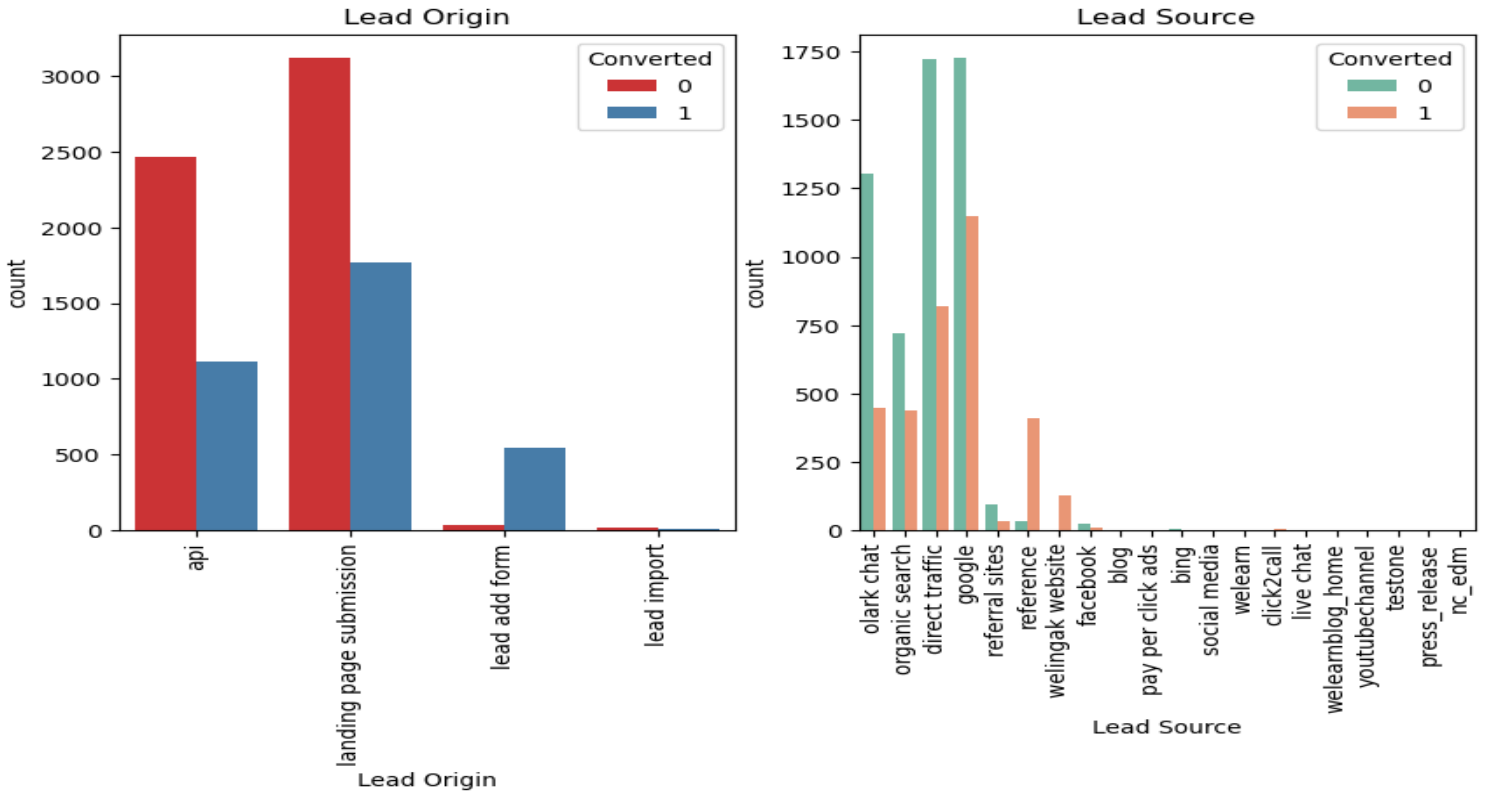


Total Visits: This histogram shows the distribution of the number of visits to the website. The data appears to be skewed right, indicating that most users have a small number of visits, with a few users having a much larger number.

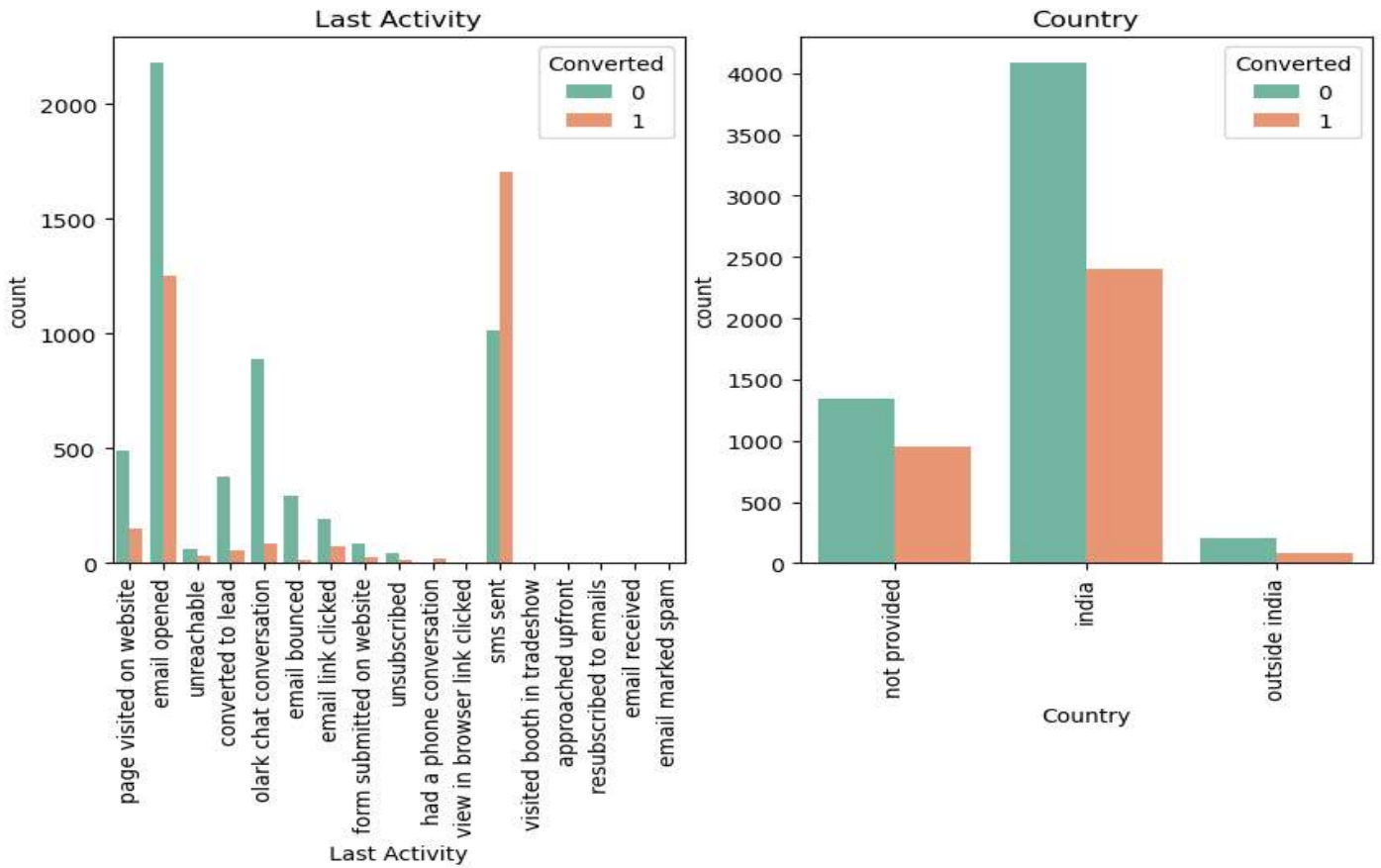
Total Time Spent on Website: This histogram shows the distribution of the total time spent on the website. Similar to the first histogram, it is skewed right, suggesting that most users spend a short amount of time on the website, while a few users spend significantly longer.

Page Views Per Visit: This histogram shows the distribution of the number of page views per visit. It also appears to be skewed right, indicating that most users view a small number of pages per visit, with a few users viewing many more pages.

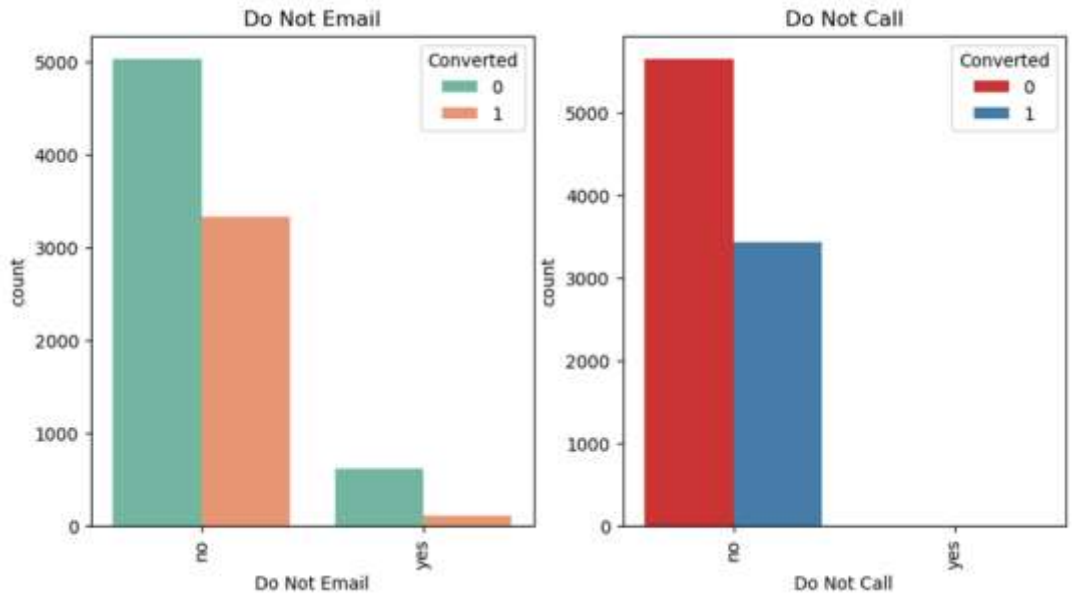
Key Observations:



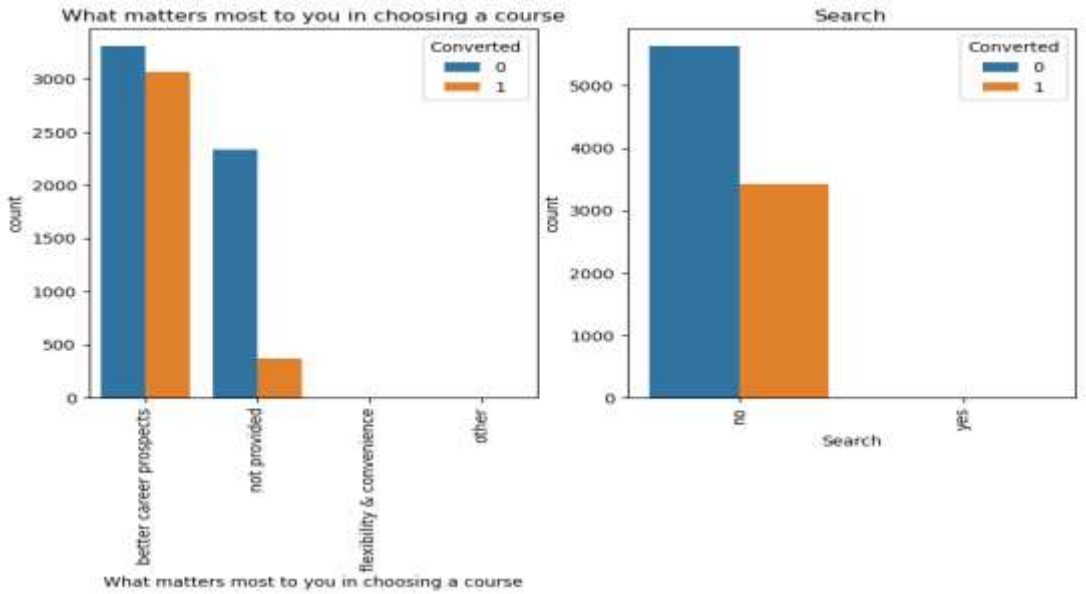
•**Lead Origin:** "Landing Page Submission" appears to have a higher conversion rate compared to other origins.
•**Lead Source:** "Google" and "Direct Traffic" seem to be significant sources of leads, but their conversion rates might vary.

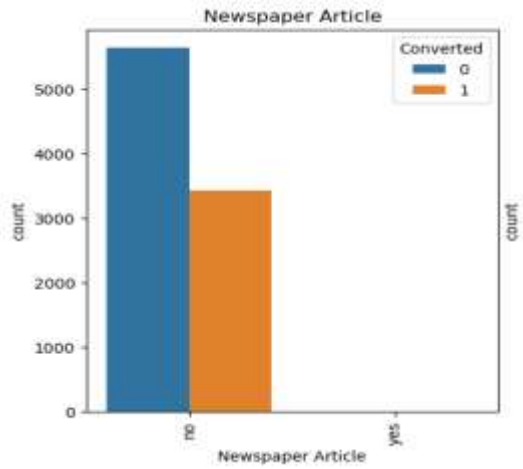


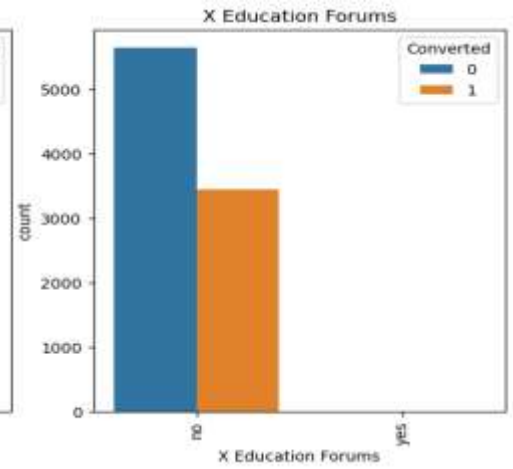
•**Last Activity:** "Converted to Lead" and "Email Opened" appear to be strong indicators of conversion.
•**Country:** India seems to be the dominant country for leads, but the conversion rate might be lower than "Outside India".

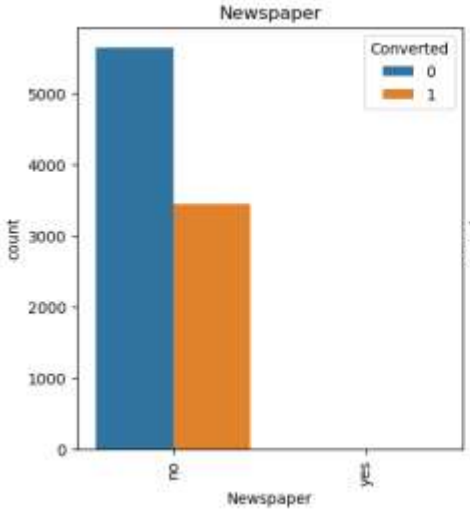


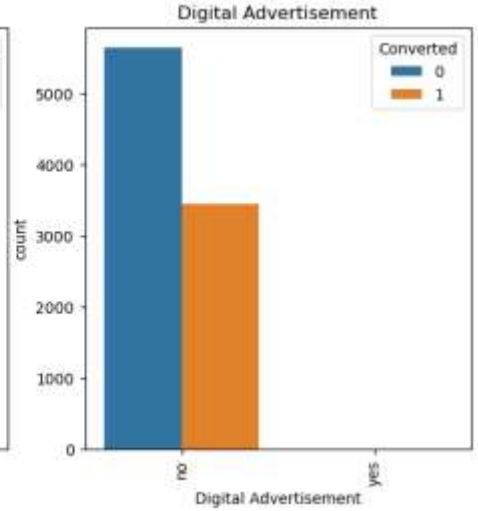
•**Do Not Email:** Leads who opted out of emails ("yes") have a significantly lower conversion rate.
•**Do Not Call:** Leads who opted out of calls ("yes") also have a lower conversion rate, but not as drastically as email opt-outs.



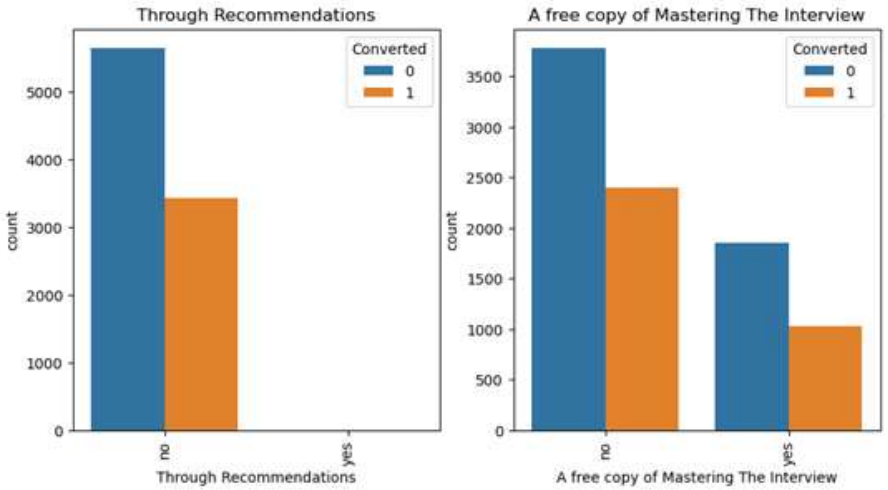


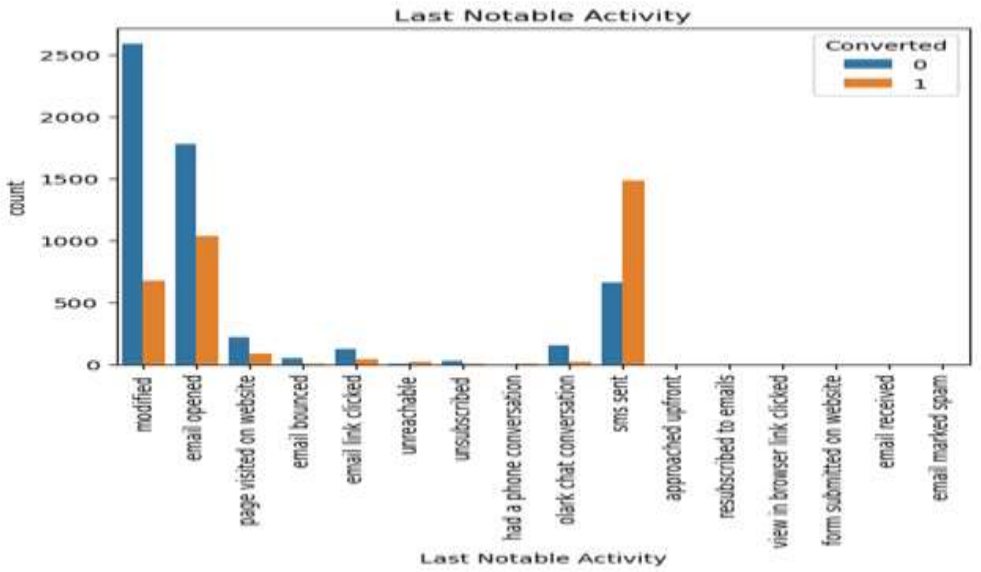






- **Lead Acquisition Channels: Search:** High volume but lower conversion rate.
- **Newspaper/Digital Ads:** Lower volume but higher conversion rates.
- **X Education Forums:** High volume with a moderate conversion rate.
- **Newspaper Article:** Lower volume but potentially higher conversion rate.





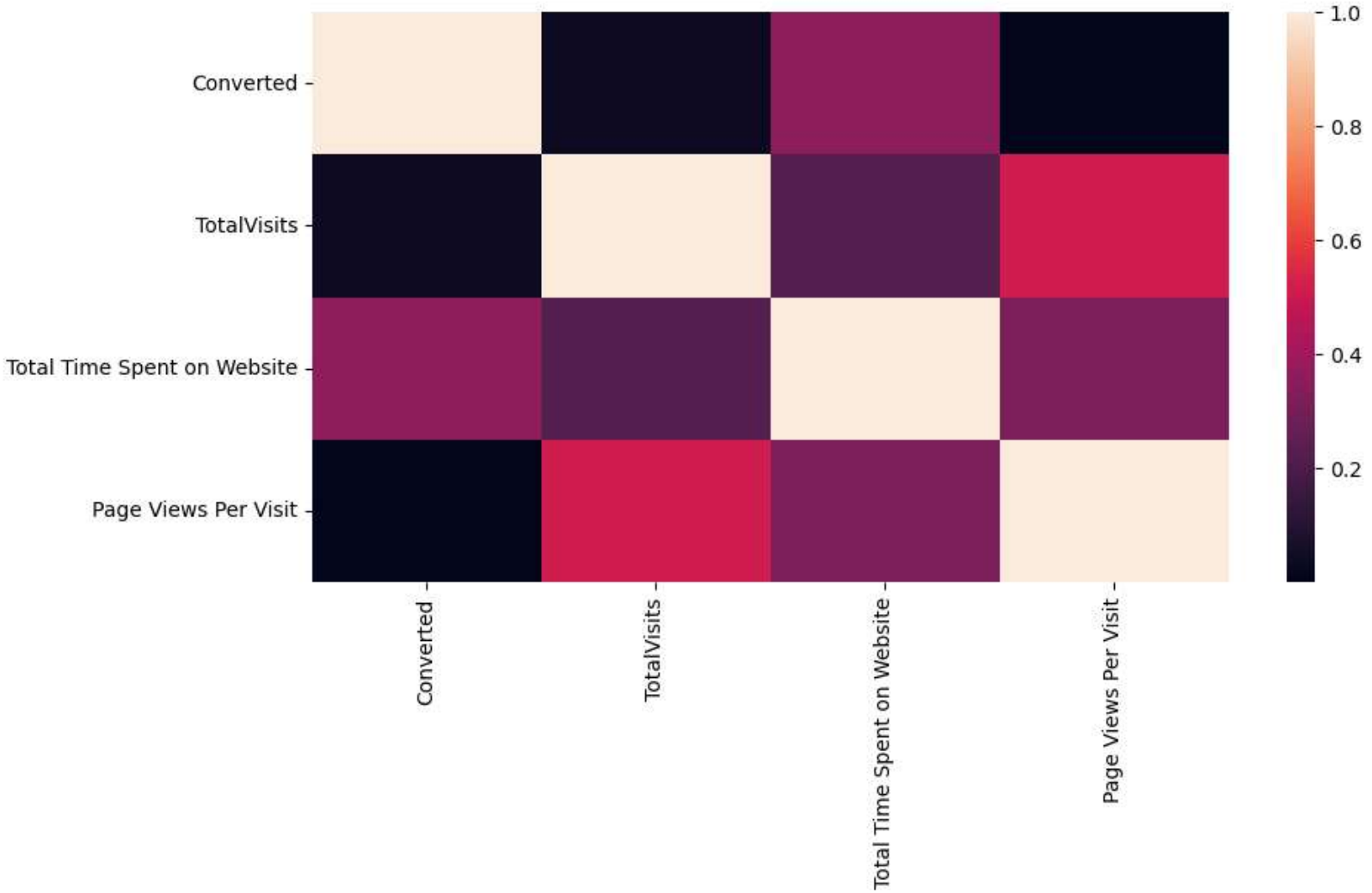
Observations and Potential Actions:

•**Diversify Lead Sources:** While "X Education Forums" brings in a high volume, exploring and optimizing other channels like newspaper and digital ads could improve overall conversion rates.

•**Refine Messaging:** Tailor messaging to better align with the motivations of different lead segments, particularly those driven by "better career prospects".

•**Leverage Effective Incentives:** Continue to leverage recommendations and free resources like "Mastering The Interview".

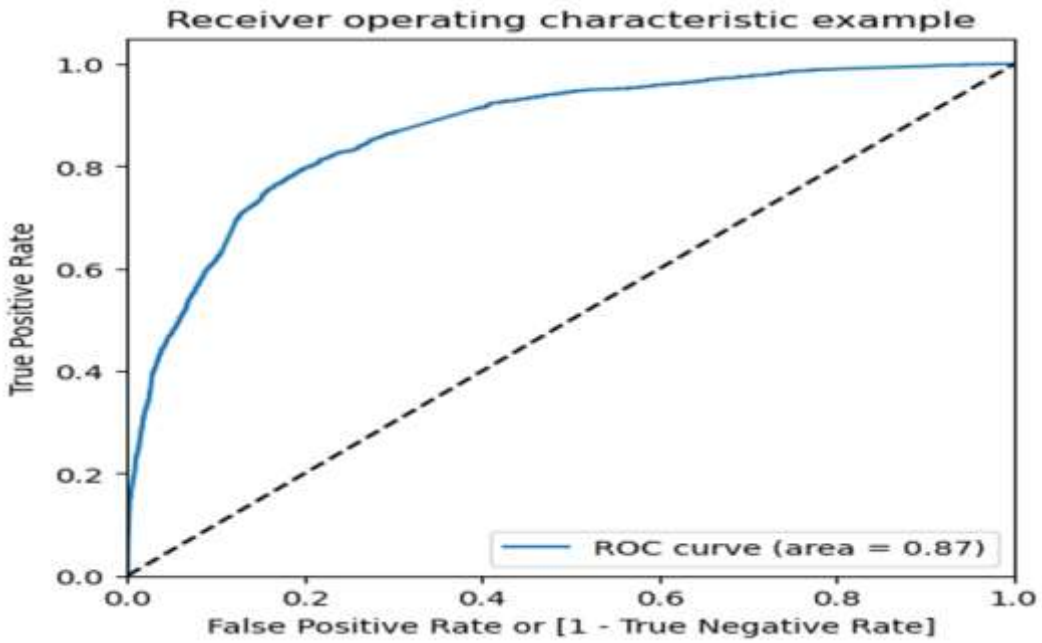
•**Optimize Last Notable Activities:** Focus on strategies that encourage "SMS Sent" and "Page Visited on Website" activities, as these seem to correlate with higher conversions.



•**Analyzing Correlations: Converted vs. Total Time Spent on Website:** The cell at the intersection of "Converted" and "Total Time Spent on Website" is light, indicating a positive correlation. This suggests that users who spend more time on the website are more likely to be converted.
•**Converted vs. Page Views Per Visit:** The cell at the intersection of "Converted" and "Page Views Per Visit" is also light, indicating a positive correlation. This suggests that users who view more pages per visit are more likely to be converted.
•**Total Visits vs. Converted:** The cell at the intersection of "Total Visits" and "Converted" is darker, indicating a weaker correlation. This suggests that the number of visits alone is not a strong predictor of conversion.
•**Total Time Spent on Website vs. Page Views Per Visit:** The cell at the intersection of "Total Time Spent on Website" and "Page Views Per Visit" is light, indicating a strong positive correlation. This suggests that users who spend more time on the website also tend to view more pages per visit.

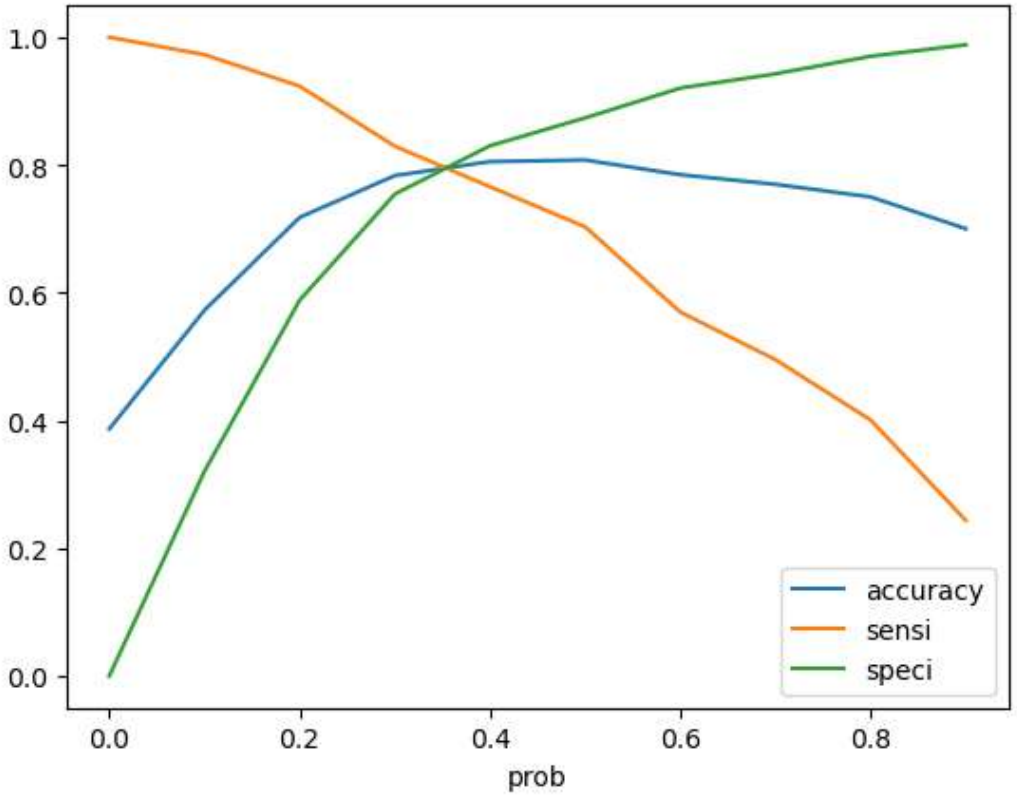
Takeaways:
•**Engagement Matters:** The heatmap highlights the importance of user engagement (time spent and page views) in predicting conversion.
•**Time is More Important Than Volume:** Total time spent on the website seems to be a stronger predictor of conversion than the total number of visits.
•**Further Investigation:** While the heatmap shows correlations, it doesn't prove causation. Further analysis is needed to understand the underlying reasons for these relationships and to develop effective strategies for improving conversion rates.

ROC Curve:



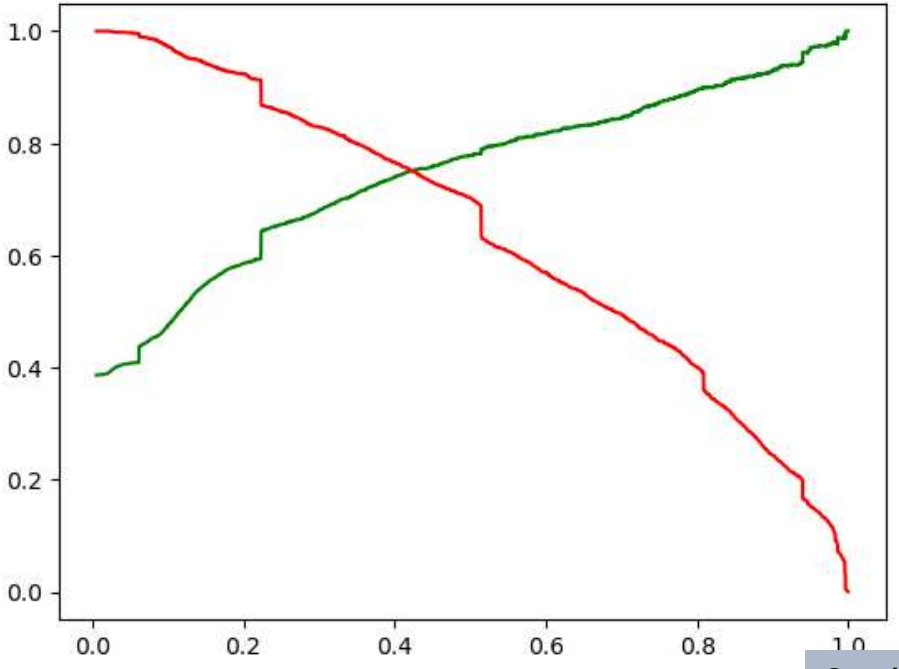
Overall Interpretation:
These plots provide a comprehensive view of the classification model's performance. The ROC curve and AUC (0.87) indicate good overall performance. The other plots help in understanding the trade-offs between different evaluation metrics and selecting an appropriate classification threshold based on the specific needs of the problem.

Finding the Optimal Decision Threshold



- To determine the optimal decision threshold, we analyzed Accuracy, Sensitivity, and Specificity across different thresholds.
- Based on the graph, the ideal threshold appears to be **0.35**, where we achieve a balanced trade-off between sensitivity and specificity while maintaining high accuracy.

Precision-Recall Tradeoff at Different Thresholds



Conclusion
It was found that the variables that mattered the most in the potential buyers are (In descending order) : 1. The total timespend on the Website. 2. Total number of visits. 3. When the lead source was:
a. Google
b. Direct traffic
c. Organic search
d. Welingak website
4. When the last activity was:
a. SMS
b. Olark chat conversation
5. When the lead origin is Lead add format. 6. When their current occupation is as a working professional.
Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers tochange their mind and buy their courses.

Takeaways:

- **Model Performance:** The model shows good performance with an AUC of 0.87.
- **Threshold Selection:** The plots help in understanding the impact of threshold selection on various metrics.
- **Trade-offs:** There is a trade-off between sensitivity and specificity. The optimal threshold depends on the relative importance of minimizing false positives vs. false negatives.

Conclusion:

- It was found that the variables that mattered the most in the potential buyers are (In descending order) :

1. The total time spend on the Website.
2. Total number of visits.
3. When the lead source was:
   a. Google
   b. Direct traffic
   c. Organic search
   d. Welingak website
4. When the last activity was:
   a. SMS
   b. Olark chat conversation
5. When the lead origin is Lead add format. 6. When their current occupation is as a working professional.
   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.

THANK YOU