Lead Scoring Case Study

Prepared and Presented by:

Shakti Singh

Sarat Shankar

Problem Statement

- X Education offers online courses designed for professionals across various industries.
- Despite generating a substantial number of leads, X Education is grappling with a low conversion rate. For instance, out of 100 leads acquired in a day, only about 30 are successfully converted.
- To enhance efficiency, the company aims to identify the leads with the highest potential, often referred to as ‘Hot Leads’.
- If the company can accurately identify this group of leads, it would result in an increased conversion rate as the sales team could then focus their efforts on these potential leads instead of reaching out to everyone.

Business Objectives

- X Education intends on identifying the most promising leads.
- To achieve this, they plan to construct a Model that can effectively identify these ‘Hot Leads’.
- The model, once developed, will be deployed for future use to streamline the lead conversion process.

Problem Resolution Process

- Data Cleaning and Transformation :

  - Check for duplicate data and Missing Values.

  - Handling the missing values such as relacing or dropping them.

  - Imputing values if required

  - Checking and handling outliers

Problem Resolution Process

- EDA:

  - Univariate Analysis: Value Count, Distribution, etc.,

  - Bi-variate Analysis: Evaluating correlation coefficients and identifying patterns among the variables, among other things.

Problem Resolution Process

- Feature Scaling, Dummy Variables and Encoding
- Classification of Data: Using Logistic Regression, model building and prediction.
- Valuation of Model
- Model Presentation
- Outcomes and Recommendation

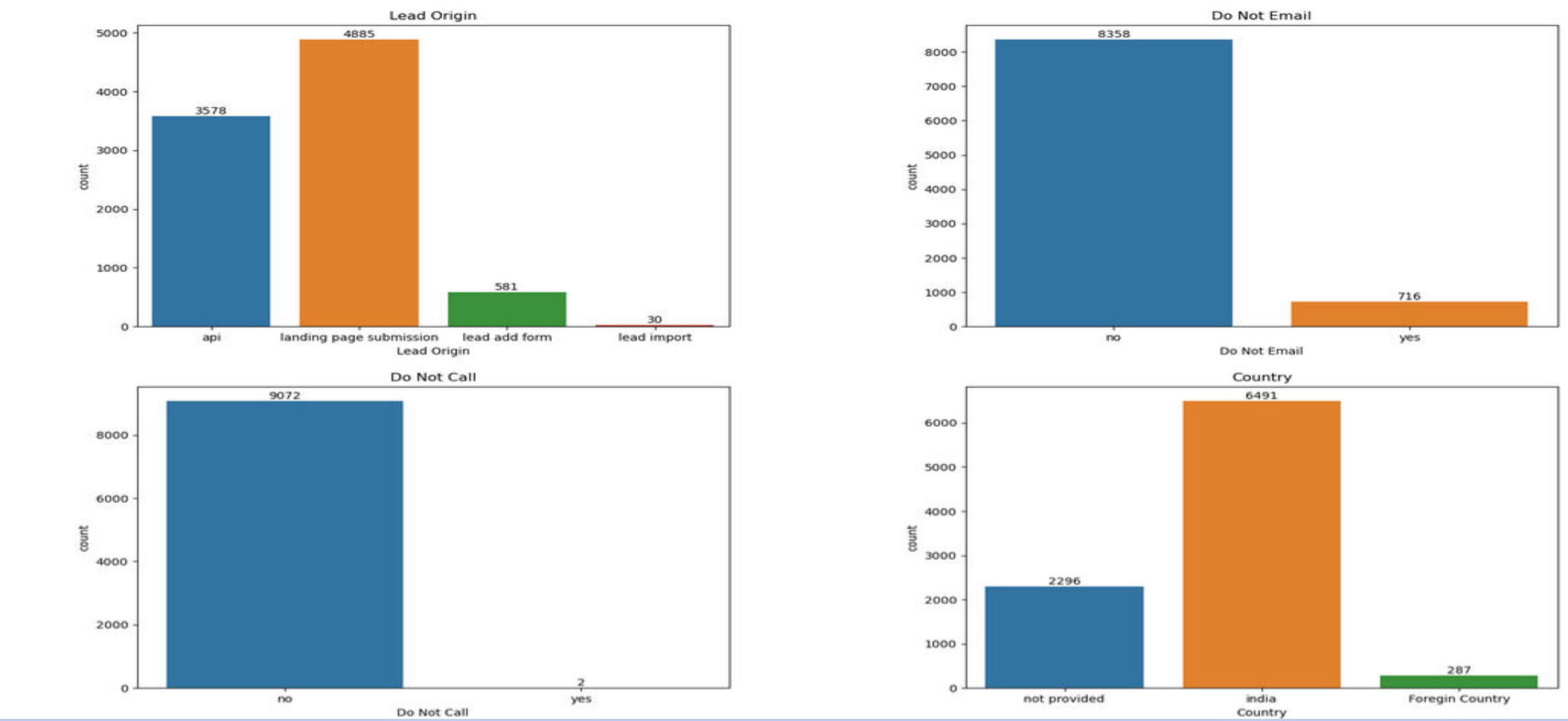
Data Transformation

- The dataset under consideration comprises 37 rows and 9240 columns.
- Certain features, such as “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply Chain Content”, “Get updates on DM Content”, and “I agree to pay the amount through cheque”, which only contain a single value, have been excluded from the dataset.
- The identifiers “Prospect ID” and “Lead Number” have been removed as they do not contribute to the analytical process but have been later added.

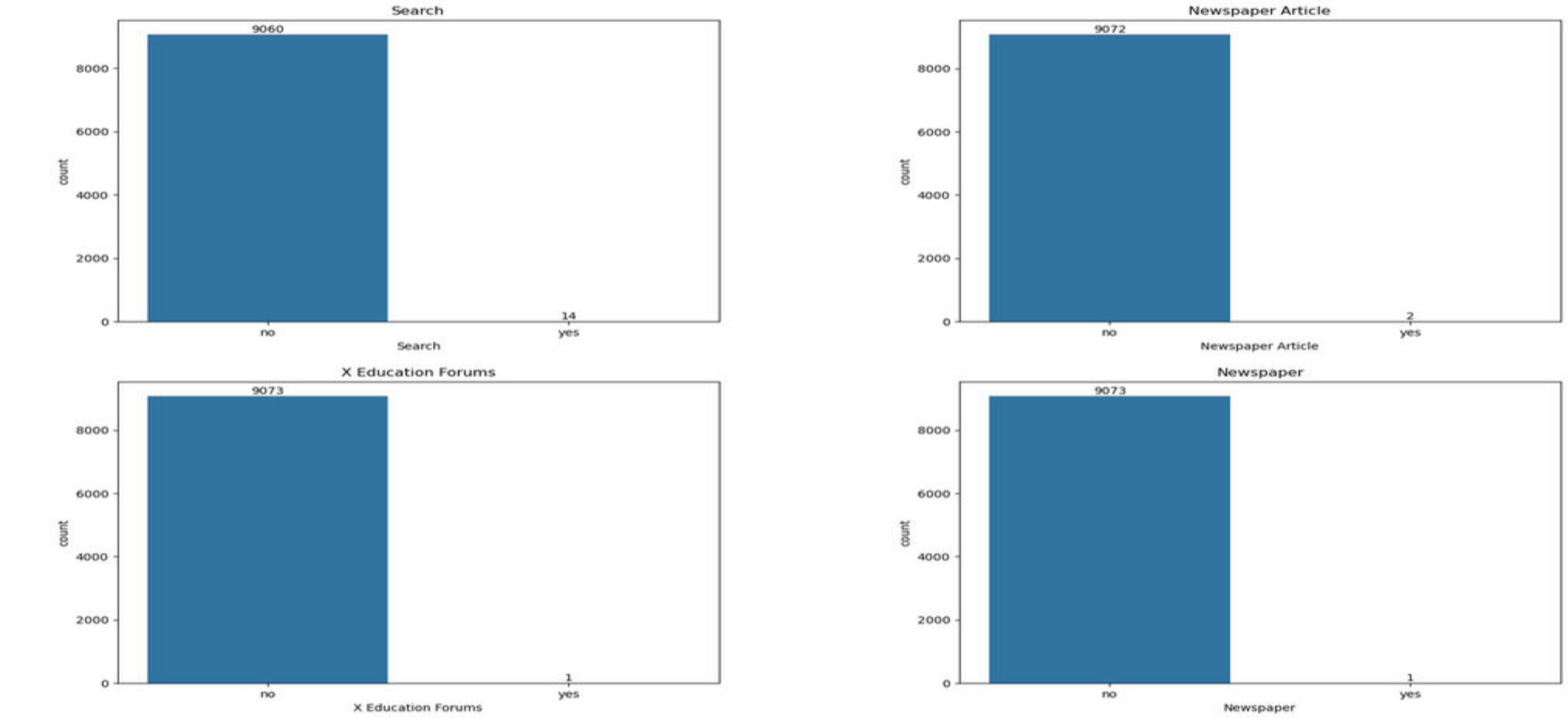
Data Transformation

- Upon evaluating the value counts for some of the object type variables, we identified certain features that exhibit insufficient variance. These features, including “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement”, and others, have been dropped from the dataset.
- Columns with more than 35% missing values, such as ‘How did you hear about X Education’ and ‘Lead Profile’, have been eliminated to maintain the integrity and reliability of the analysis.
- This meticulous data preprocessing ensures a robust and reliable foundation for subsequent analytical procedures.

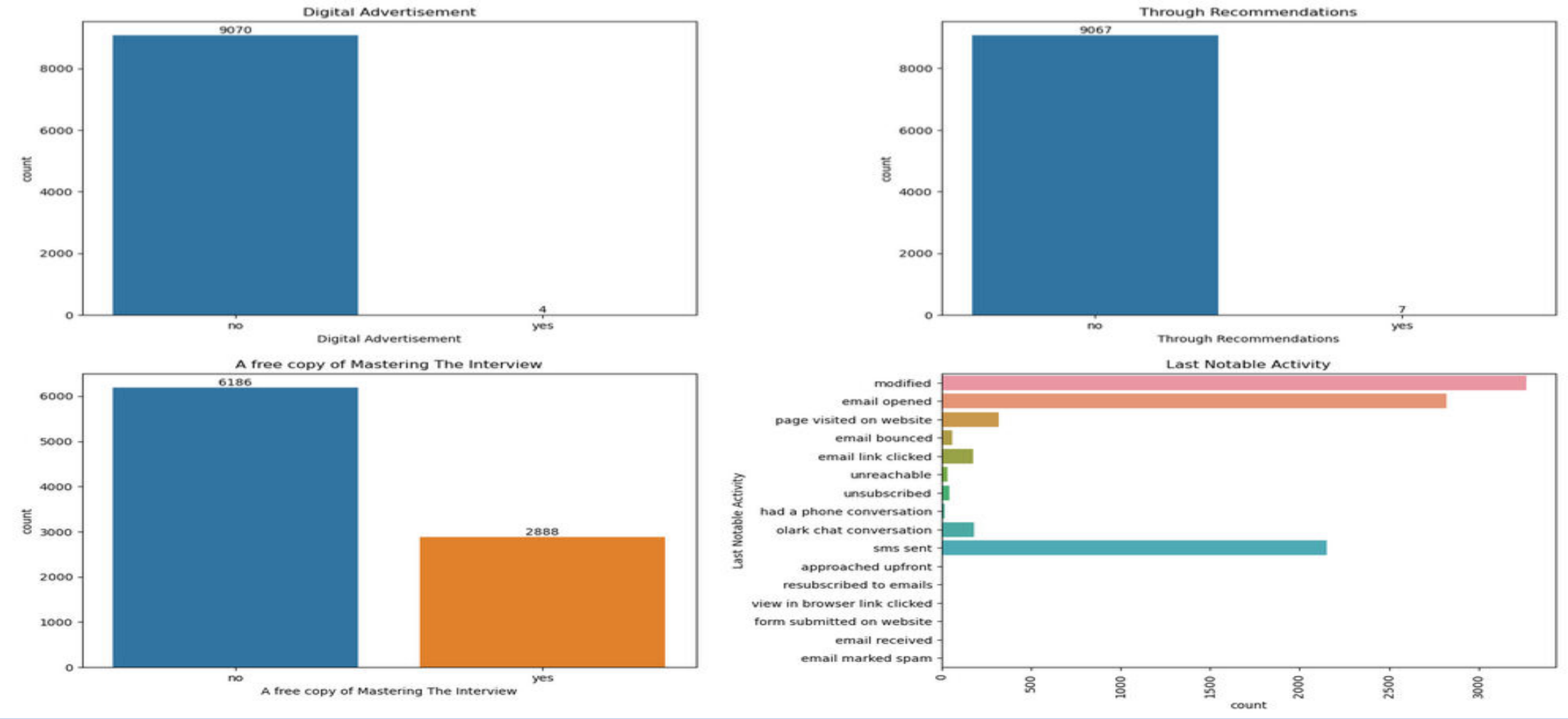
EDA


Lead Origin
count
Lead Origin
api
landing page submission
lead add form
lead import
3578
4885
581
30
Do Not Email
Do Not Email
no
yes
8358
716
Do Not Call
Do Not Call
no
yes
9072
2
Country
Country
not provided
india
Foregin Country
2296
6491
287

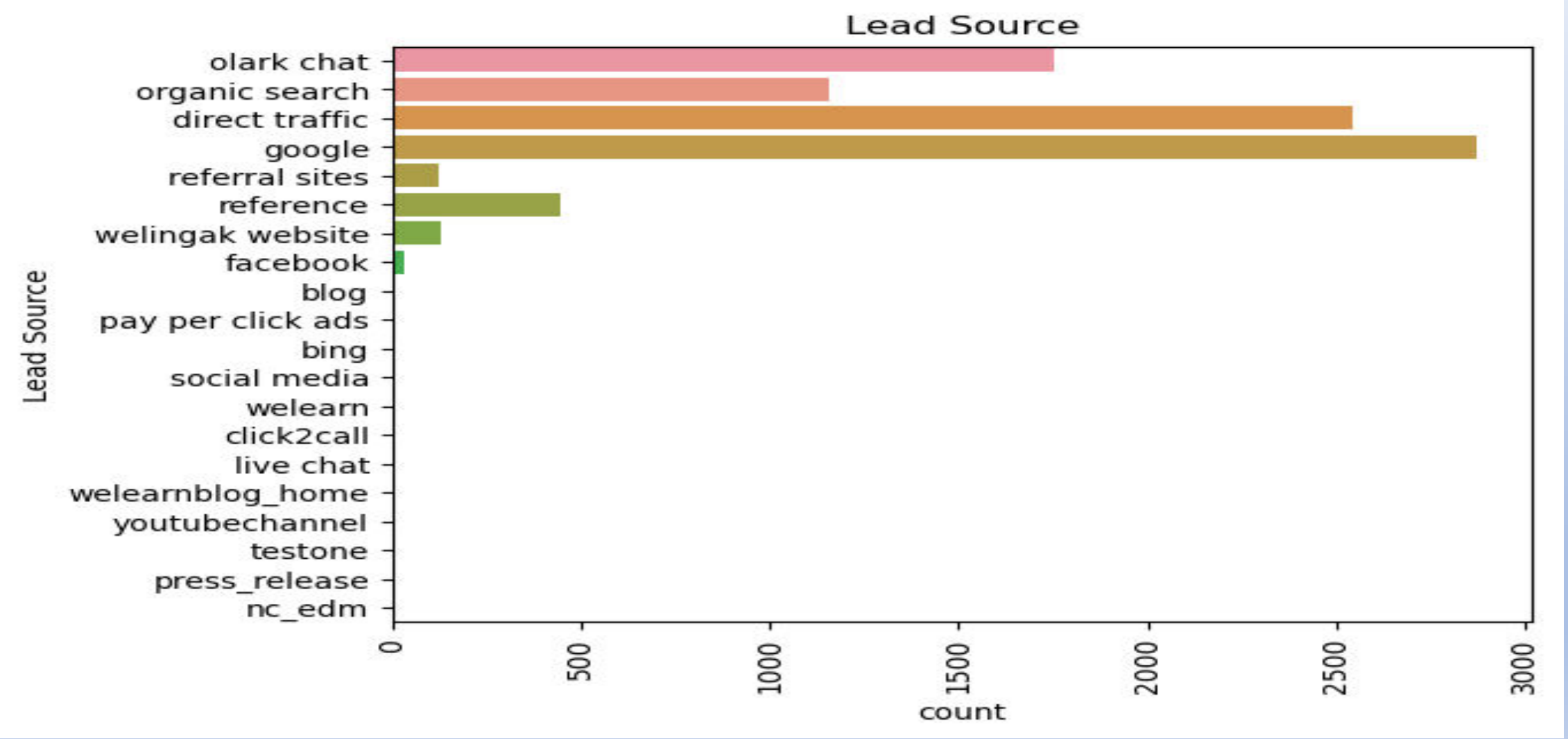
EDA


Search
9060
14
no
yes
Search
count
Newspaper Article
9072
2
no
yes
Newspaper Article
X Education Forums
9073
1
no
yes
X Education Forums
Newspaper
9073
1
no
yes
Newspaper
0
2000
4000
6000
8000

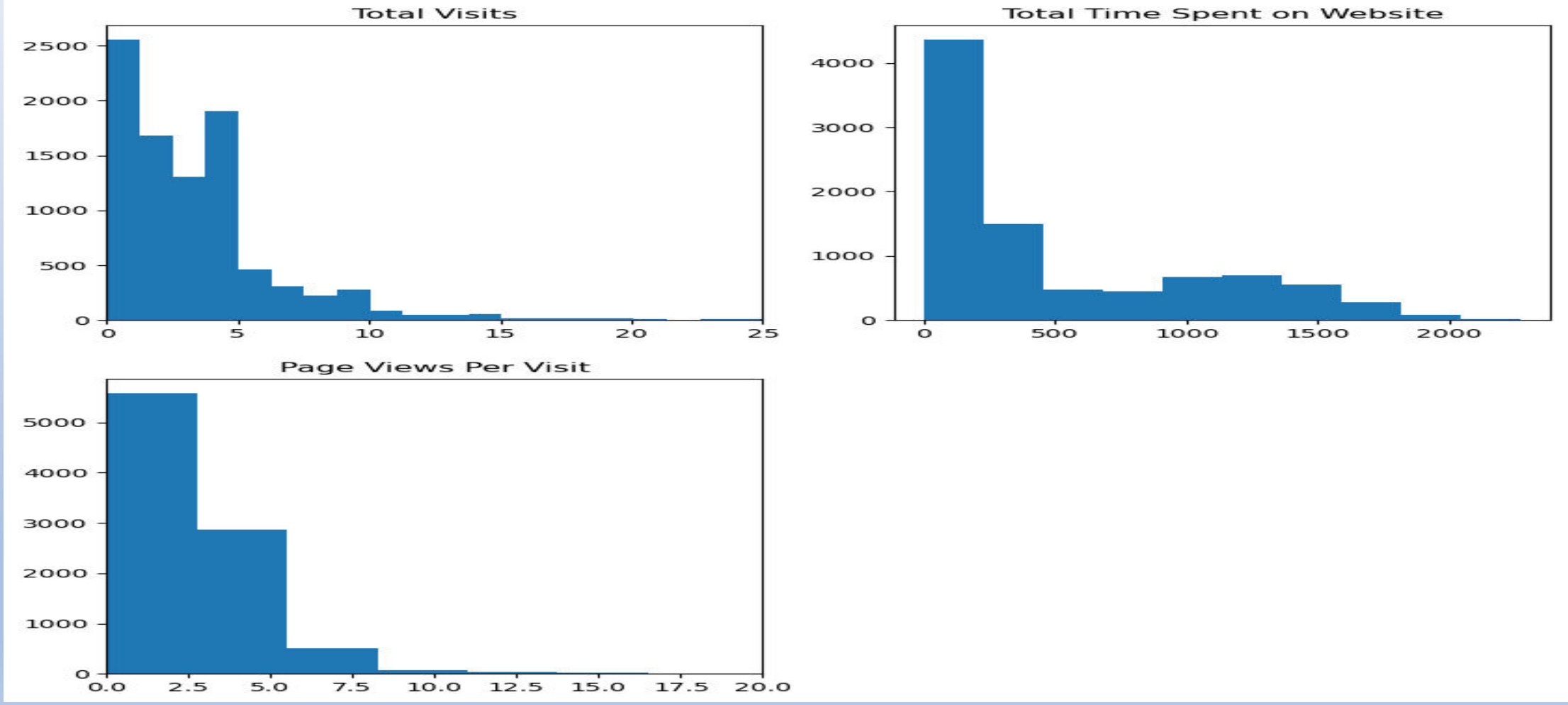
EDA


Digital Advertisement
9070
4
no
yes
count
Digital Advertisement
Through Recommendations
9067
7
no
yes
count
Through Recommendations
A free copy of Mastering The Interview
6186
2888
no
yes
count
A free copy of Mastering The Interview
Last Notable Activity
modified
email opened
page visited on website
email bounced
email link clicked
unreachable
unsubscribed
had a phone conversation
olark chat conversation
sms sent
approached upfront
resubscribed to emails
view in browser link clicked
form submitted on website
email received
email marked spam
Last Notable Activity
count

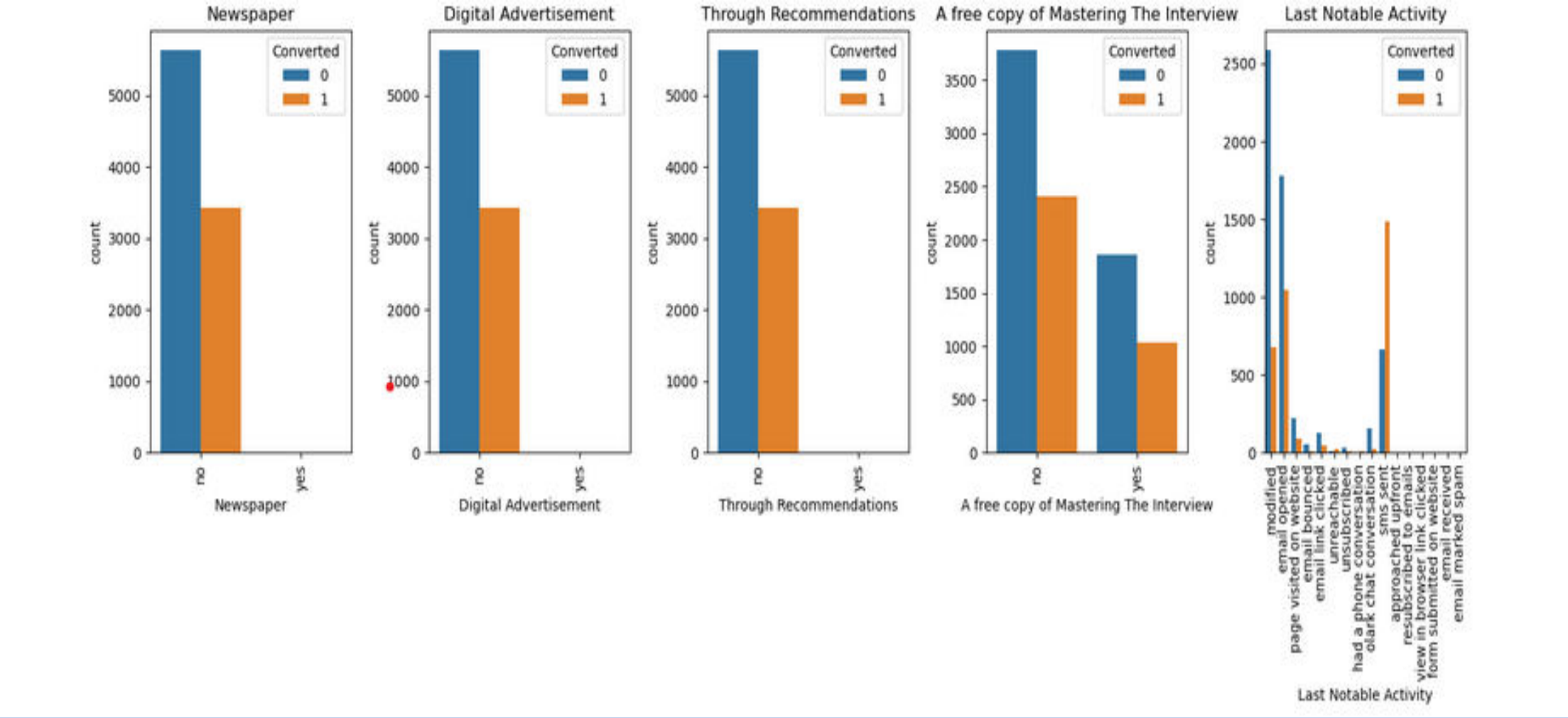
EDA


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

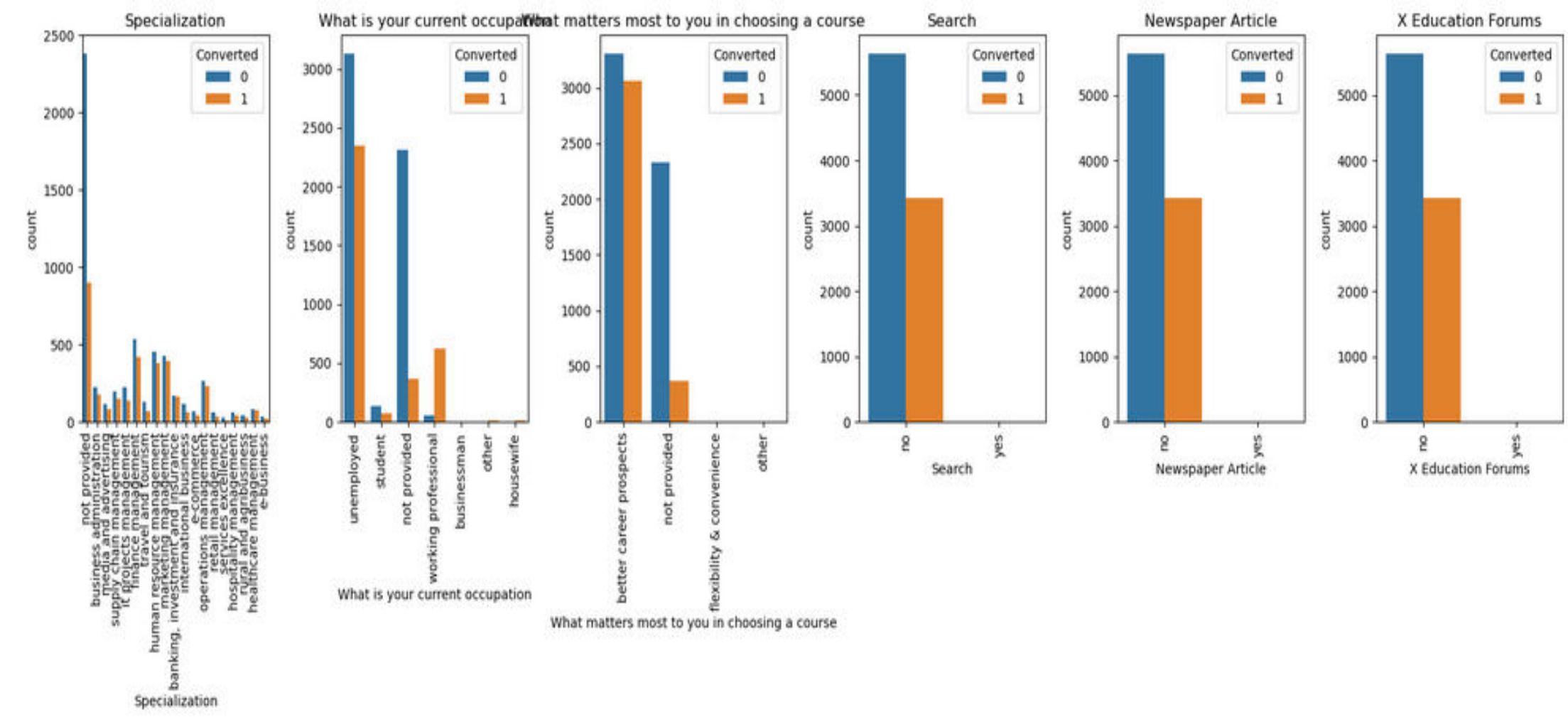
Numerical Data Representation



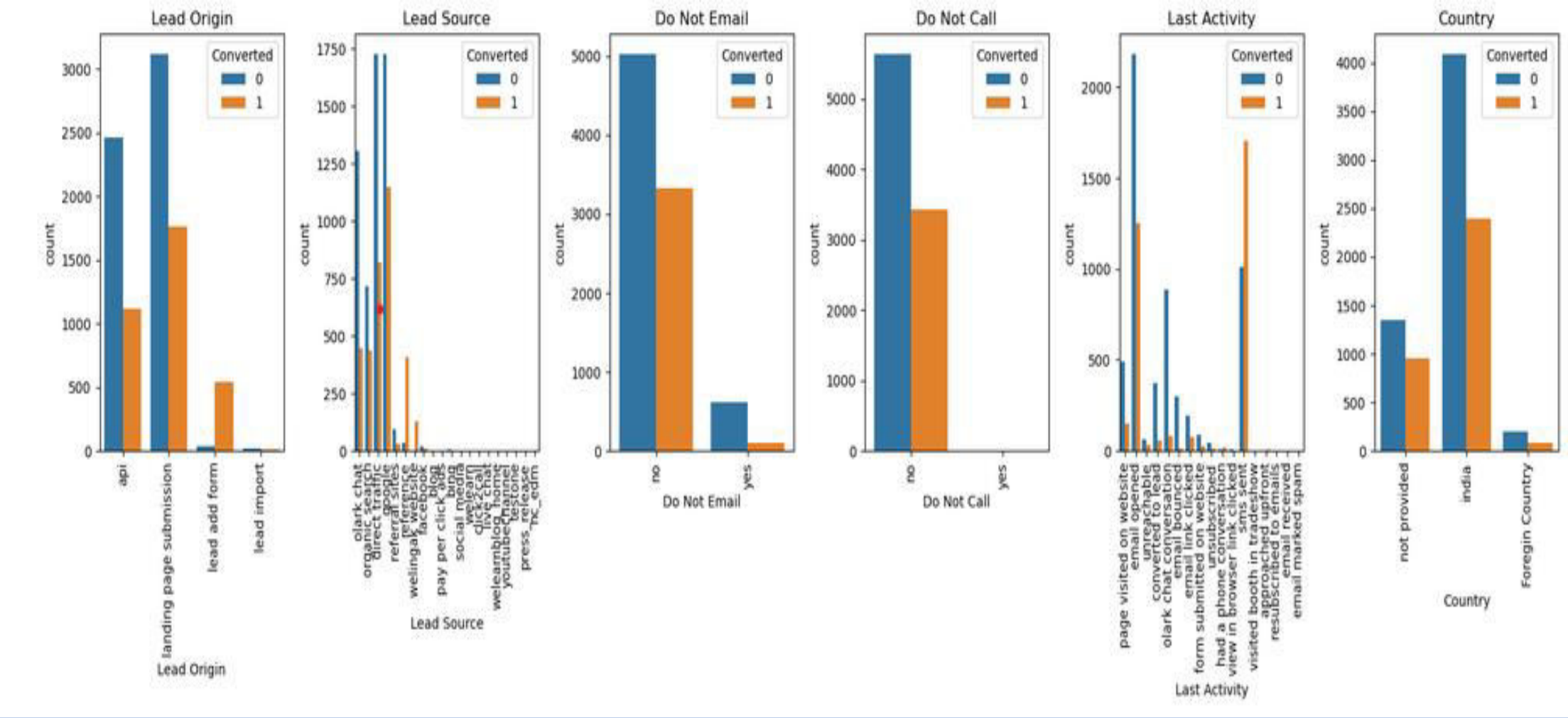
Categorical Variables in-relation-to the Variable "Converted"



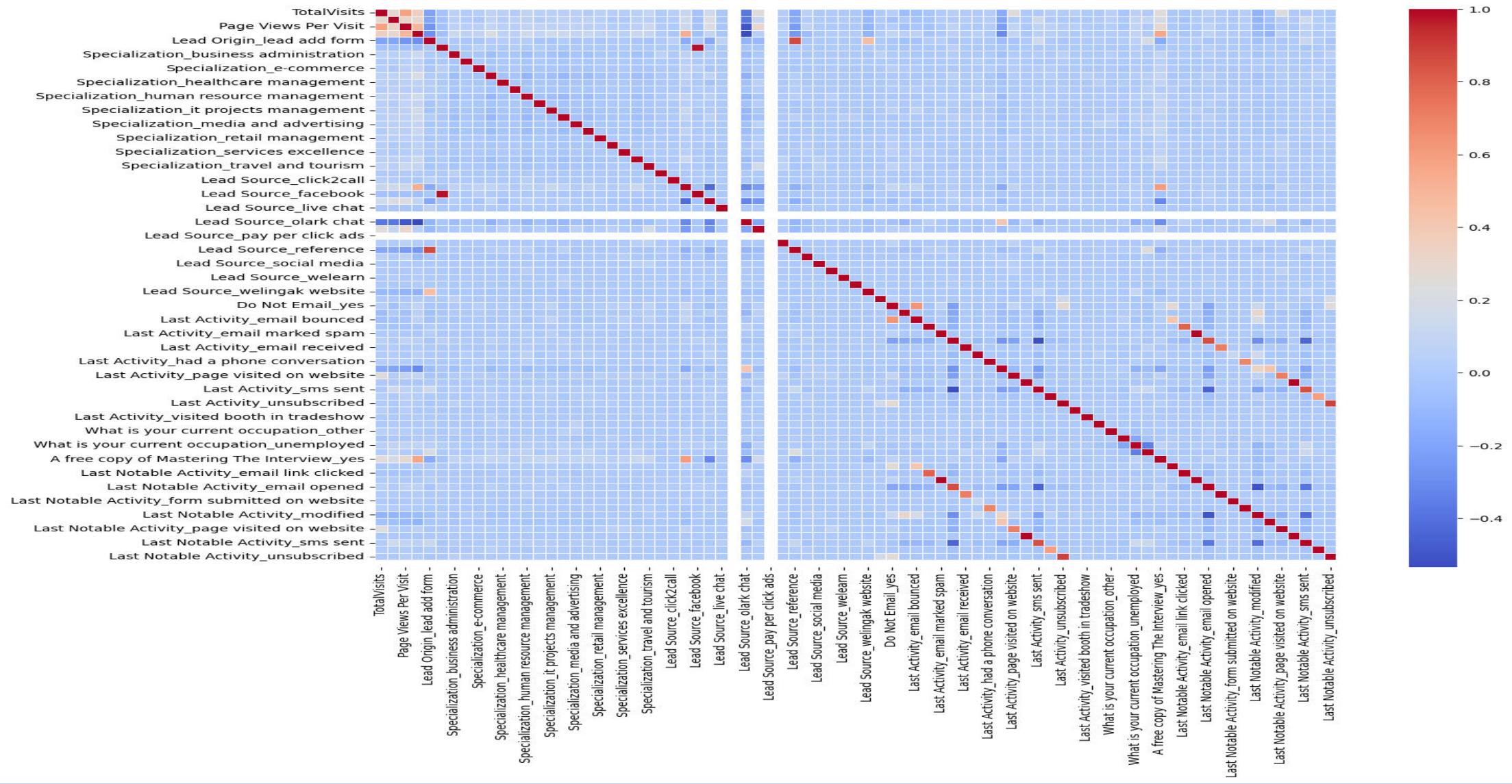
Categorical Variables in-relation-to the Variable "Converted"



Categorical Variables in-relation-to the Variable "Converted"



Heat Map showing Correlation



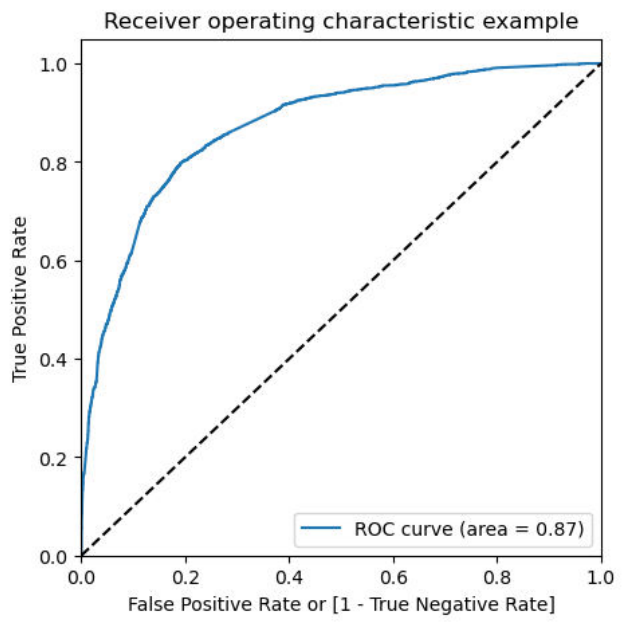
Data Transformation

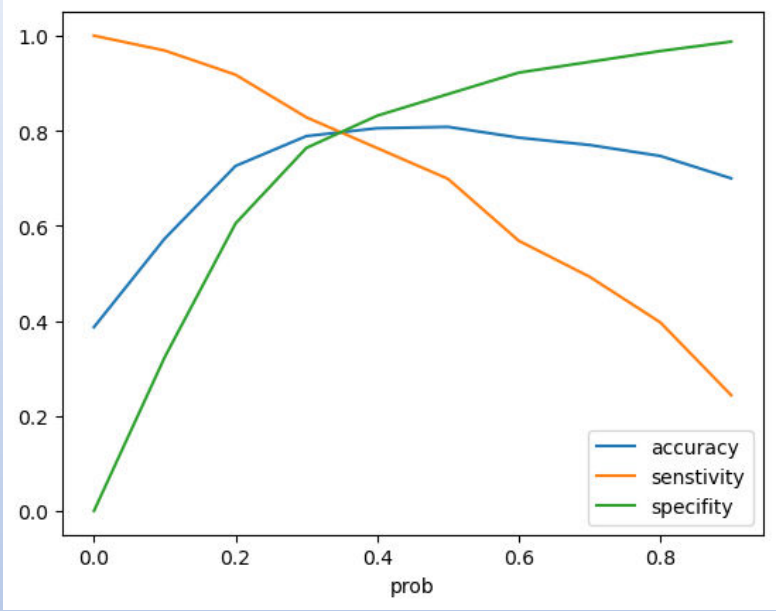
- Numerical variables have been standardized.
- Dummy variables have been generated for variables of object type.
- The total number of rows available for analysis is 9074.
- The total number of columns available for analysis is 81.

Model Creation

- The data is partitioned into training and testing sets.
- The initial fundamental step for regression involves executing a train-test split, with a chosen ratio of 70:30.
- Recursive Feature Elimination (RFE) is employed for feature selection.
- RFE is executed with an output of 15 variables.
- The model is constructed by eliminating the variable with a p-value greater than 0.05 and a Variance Inflation Factor (VIF) value exceeding 5.
- Predictions are made on the test dataset.
- The model achieves an overall accuracy of about 81%.

ROC Curve and Cut-Off plot





- The area under the ROC Curve is high at 0.87 which is very good. Therefore, it is safe to say that our model is a good model.
- From the above graph we can determine our probability cut-off to be at around 0.35 and we will proceed forward with this threshold or optimum point.

Observation and Results

- Observations:

  After running the model on the Test Data, it is determined that:

  Accuracy : 80.30 % Sensitivity : 80.08 % Specificity : 80.40 %

- **Results :**

  A) Comparing the values obtained for Train & Test:

  Train Data:

  Accuracy : 80.38 % Sensitivity : 80.04 % Specificity : 80.59 %

  Test Data:

  Accuracy : 80.30 % Sensitivity : 80.08 % Specificity : 80.40 %

Observation and Results

- We have successfully reached our objective of estimating the target lead conversion rate to be approximately 80%. The model appears to forecast the Conversion Rate accurately, which should instill confidence in the CEO to make informed decisions using this model to achieve a higher lead conversion rate of 80%.

B) Leads that can be contacted:

- Customers with a “Lead Score” of 85 or above should be reached out to. These customers can be classified as ‘Hot Leads’.

Conclusion

- The company should make calls to the Leads from 'TotalVisits' forthwith as they are more likely to be converted that also shows or rather suggests that they have interest as the Visits are High as well as their conversion probability is the highest.
- The company should make calls to leads generated from 'Total Time Spent on Website' as their conversion probability is the second highest.
- 'Lead Origin_lead add form' leads coming from this source have expressed their interest, hence they should be given a call as well and have the third highest conversion rate.
- 'What is your current occupation_working professional' leads from this source also have a good chance of bring converted hence they should be given a call as well.
- 'Lead Source_welingak website' leads from this source also have a decent chance of being converted. Therefore, giving a call is advisable.

Conclusion

- 'Last Notable Activity_unreachable' Leads from this source have a positive chance of conversion as well mainly because they were essentially not spoken to in the first place. A follow up call can result in a positive outcome.
- 'Lead Source_olark chat' leads from this source stand a positive chance of conversion as they might have some interest.
- 'Last Activity_sms sent' leads from this source also can be converted as the customer did respond a follow up call might just make a difference.
- 'Do Not Email_yes', 'Last Activity_olark chat conversation' Leads from this sources should not be contacted as it would be a waste of both company's time as well as the customer as the customer has already been informed/contacted and have expressed their disinterest.

The End