LEAD SCORING CASE STUDY

Submitted by :
Chenamma Priya Rangaraj

PROBLEM STATEMENT

- X Education, an online education provider, targets industry professionals through various marketing channels such as websites and search engines like Google. Upon visiting the website, potential customers may explore courses, fill out course forms, or engage with course materials. Leads are identified when individuals provide their contact information, such as email addresses or phone numbers.

- Despite generating a considerable number of leads, X Education faces challenges with lead conversion. For instance, out of 100 leads acquired in a day, only about 30 are converted into paying customers.

- X Education aims to prioritize leads with the highest likelihood of conversion, thereby maximizing their sales opportunities. To achieve this objective, a model needs to be developed to assign a lead score to each prospect. This lead score will serve as an indicator of the probability of conversion, with higher scores indicating a greater likelihood of conversion and lower scores suggesting a lower chance of conversion.

ANALYSIS APPROACH

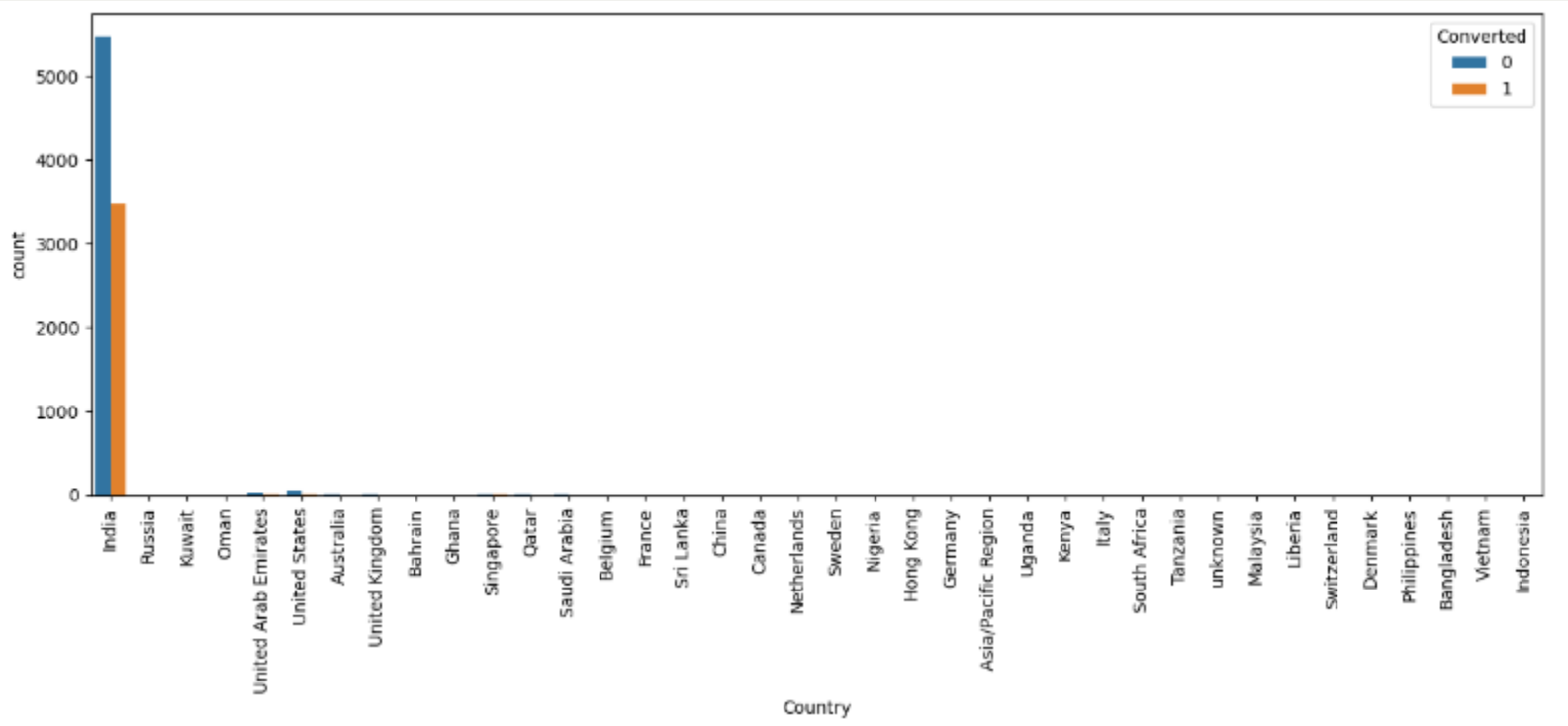
DATA CLEANING AND TREATMENT:

- The "select" level found in numerous categorical variables was addressed.
- Columns with over 45% missing values were removed.
- "Prospect ID" and "Lead Number," which merely serve as identification numbers for contacted individuals, were eliminated.

2. CATEGORICAL ATTRIBUTE ANALYSIS

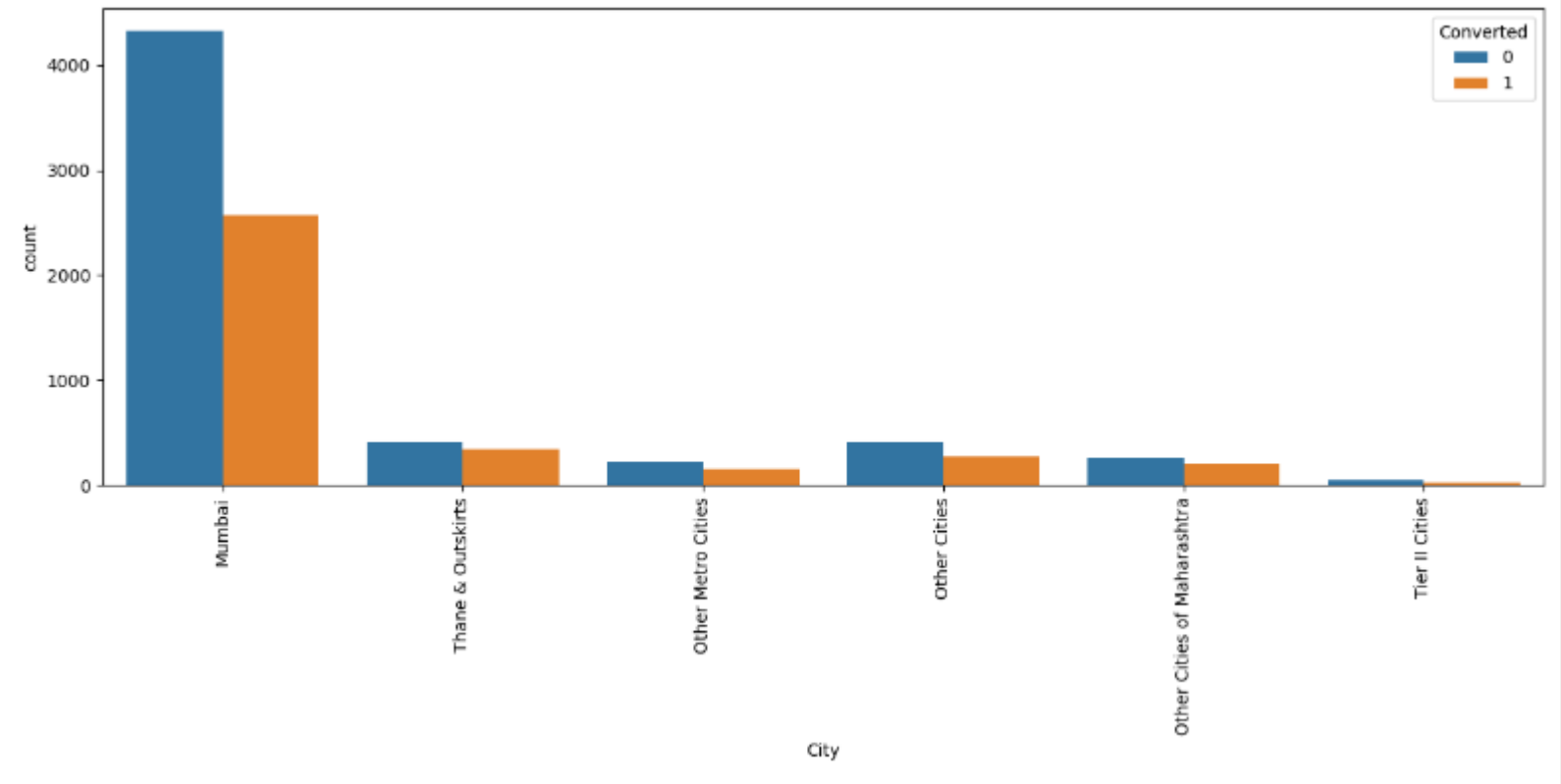
2.1 **Graph illustrating the distribution of converted and non-converted leads by country**

Among the non-missing values, 'India' was the most frequently occurring country. Therefore, we replaced the missing values with 'India'. After imputation, the majority of values in the column are 'India' (approximately 97% of the data), indicating highly skewed data. Consequently, this column can be dropped.



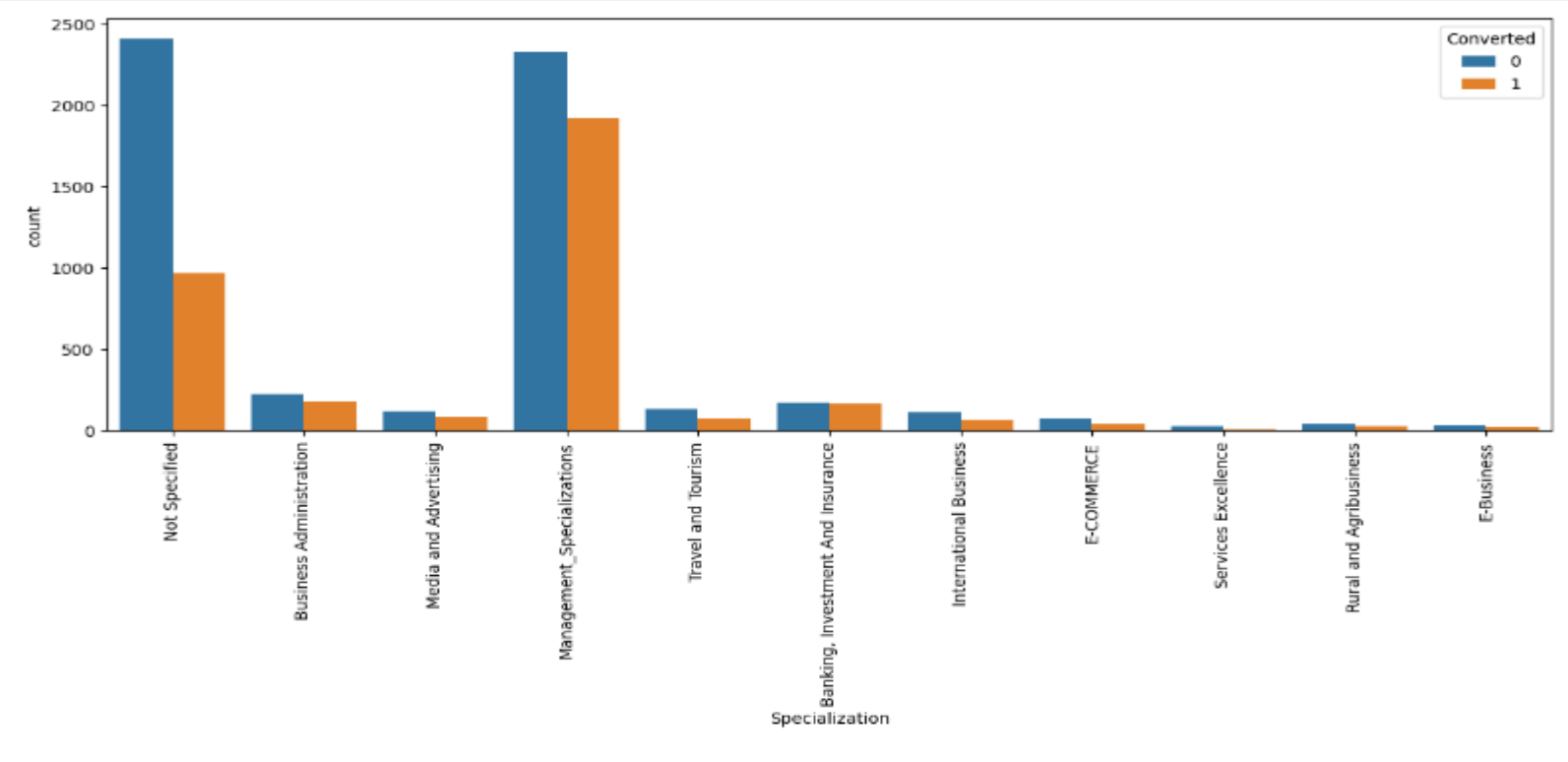
2.2 Creating a visual representation of the distribution of converted and non-converted leads based on city.

- Observing that Mumbai City has the highest number of both converted and non-converted leads.



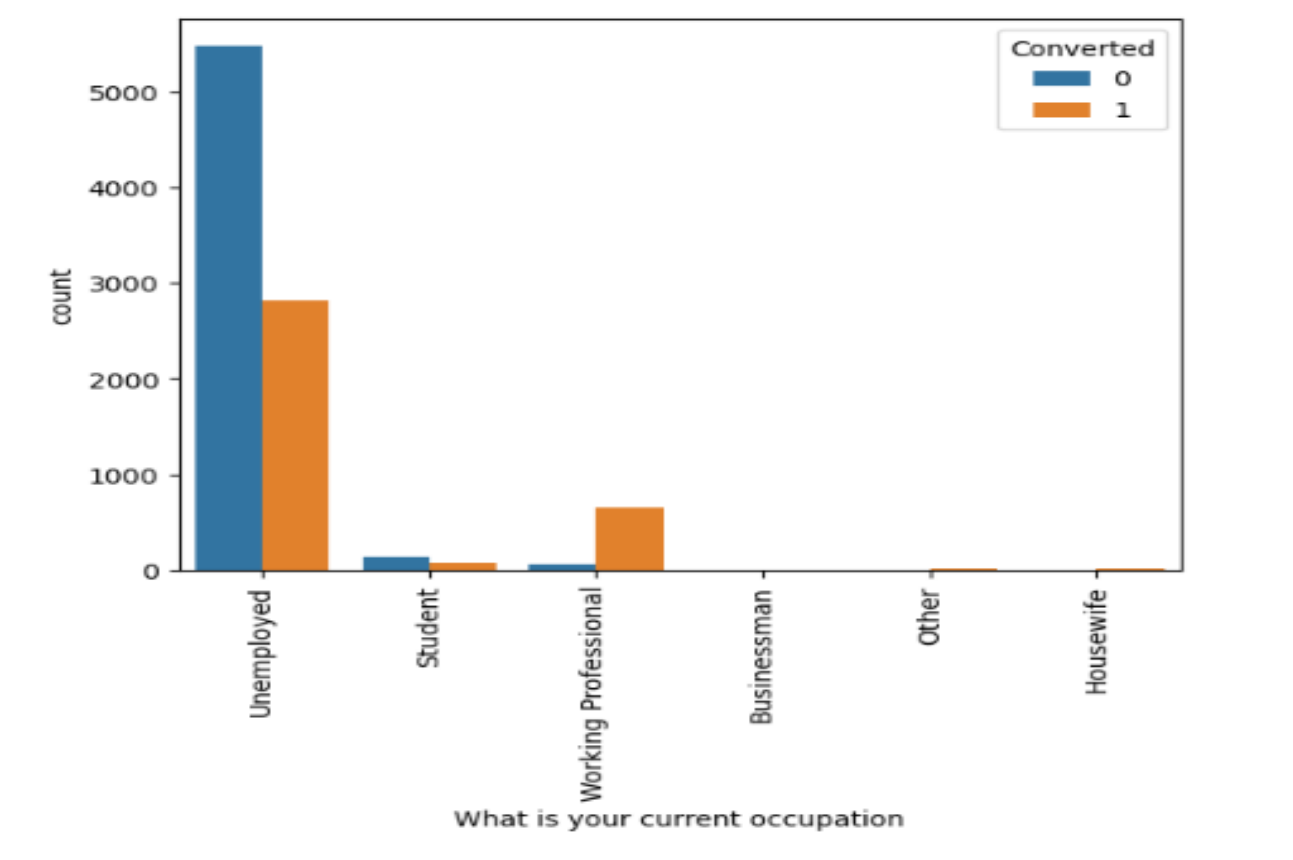
2.3 **Visualizing the distribution of converted and non-converted leads based on specialization**

- We observe that the majority of leads have not specified their specialization, possibly because they may not have provided this information, indicating that they are students.
- Specializations in management exhibit a higher number of both converted and non-converted leads. Therefore, courses related to management should be prioritized.

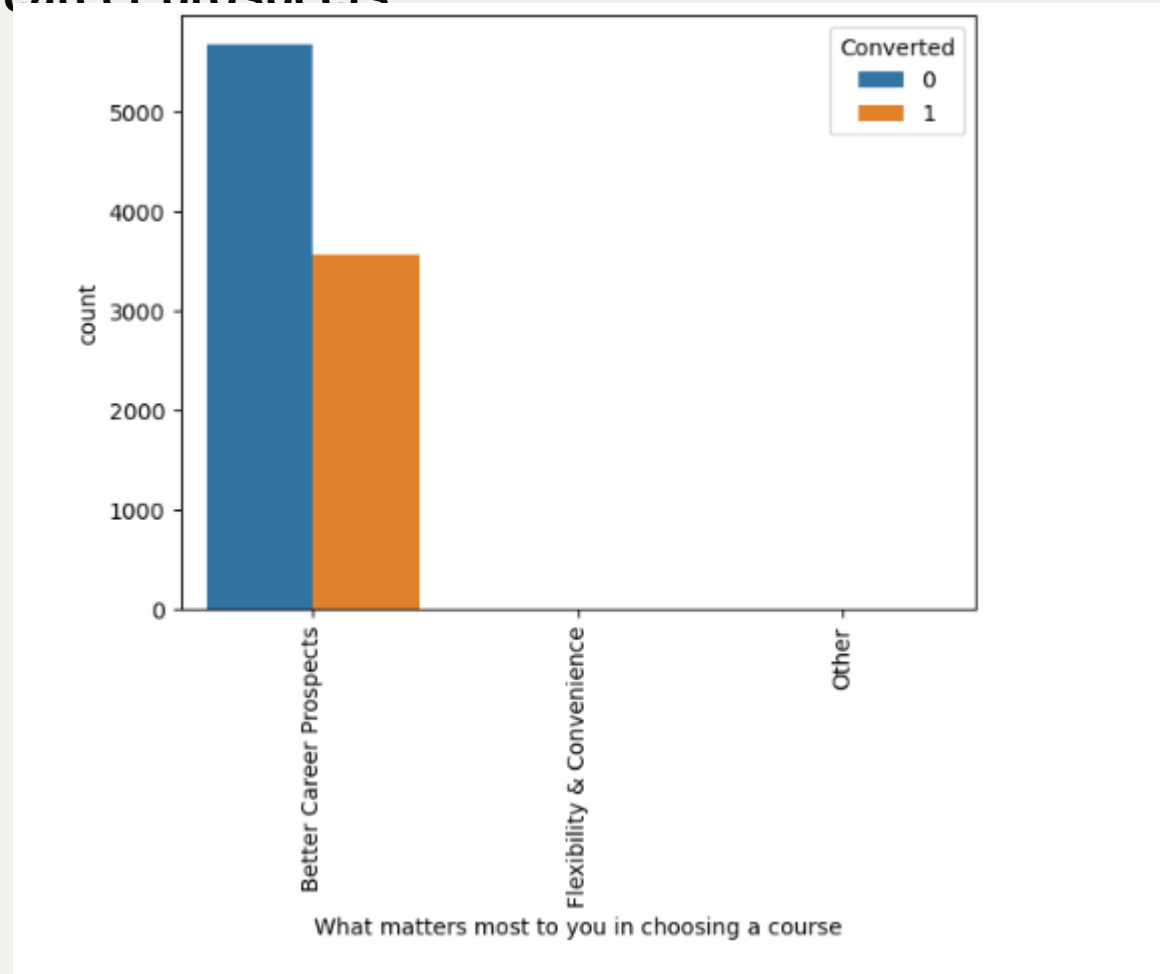


2.4 **Distribution of converted and non-converted leads based on occupation**

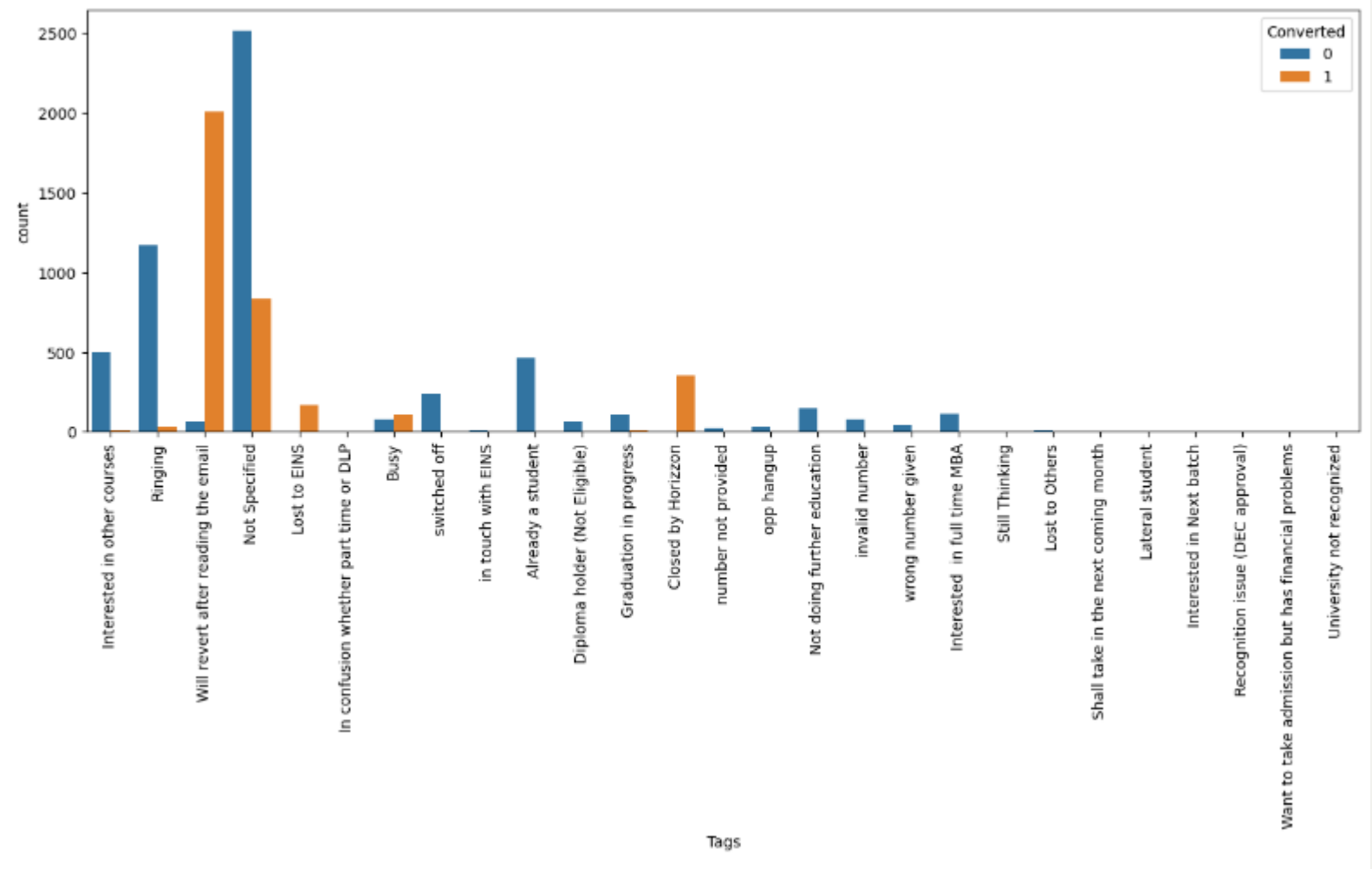
- Working professionals opting for the course have a higher likelihood of enrollment.
- Unemployed leads constitute the highest number in absolute terms.



2.5 **The majority of both converted and non-converted leads select the course due to better career prospects**

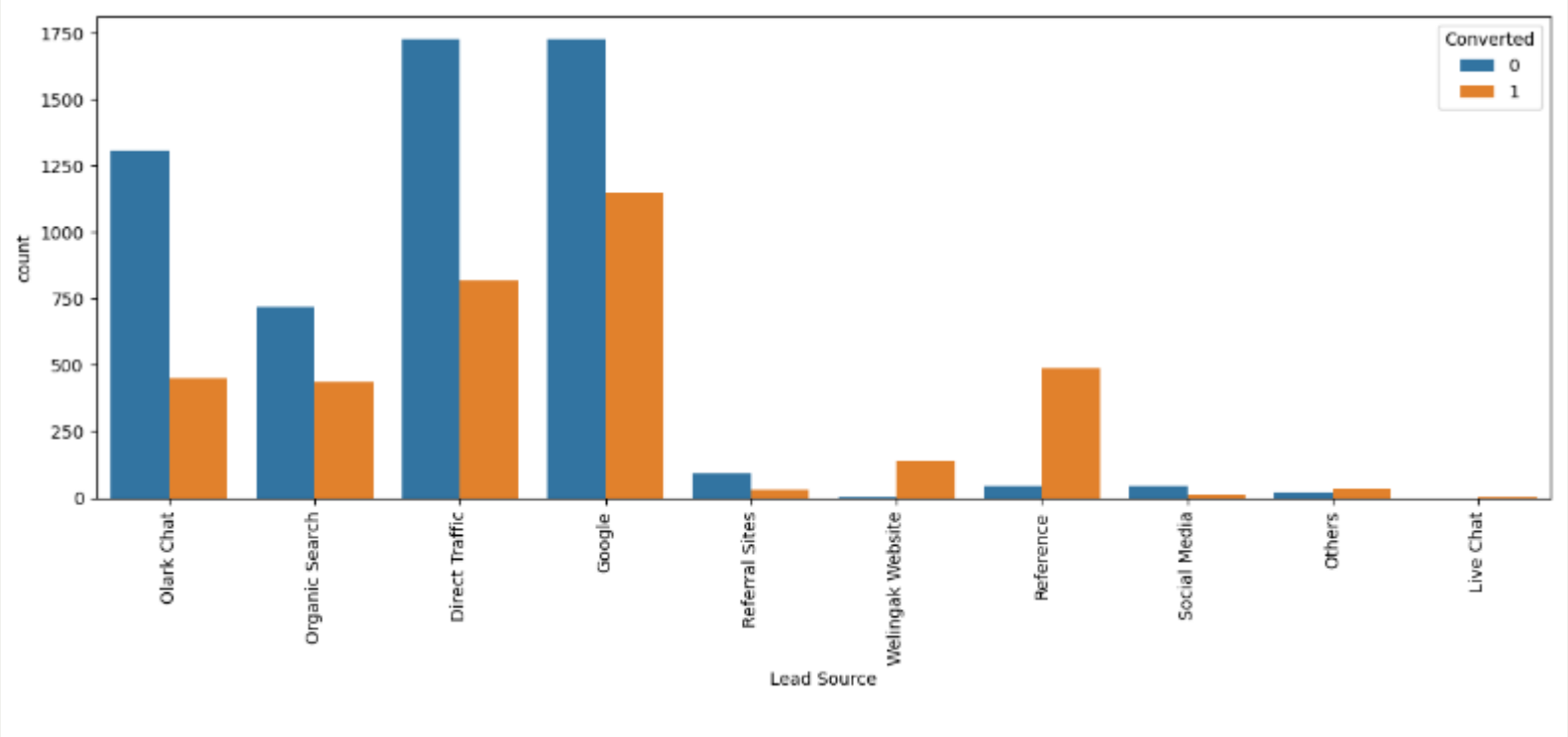


2.6 Many of the leads that were converted are expected to respond after reading the email advertisement



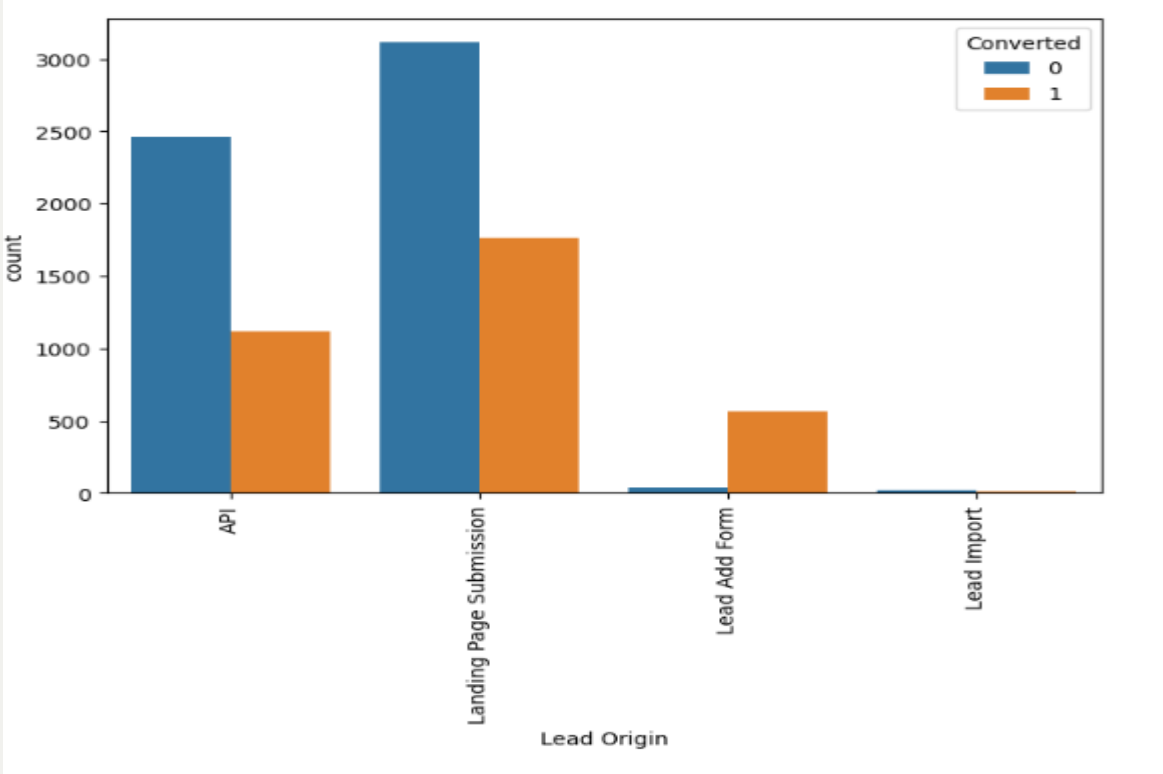
2.7 **Distribution of converted and non-converted leads by source.**

- The highest number of leads originates from Google and direct traffic sources.
- Leads from reference sources and the Welingak website exhibit a high conversion rate.
- To enhance the overall lead conversion rate, efforts should concentrate on improving the conversion rates of leads from Olark chat, organic search, direct traffic, and Google sources, while also increasing the number of leads from reference sources and the Welingak website

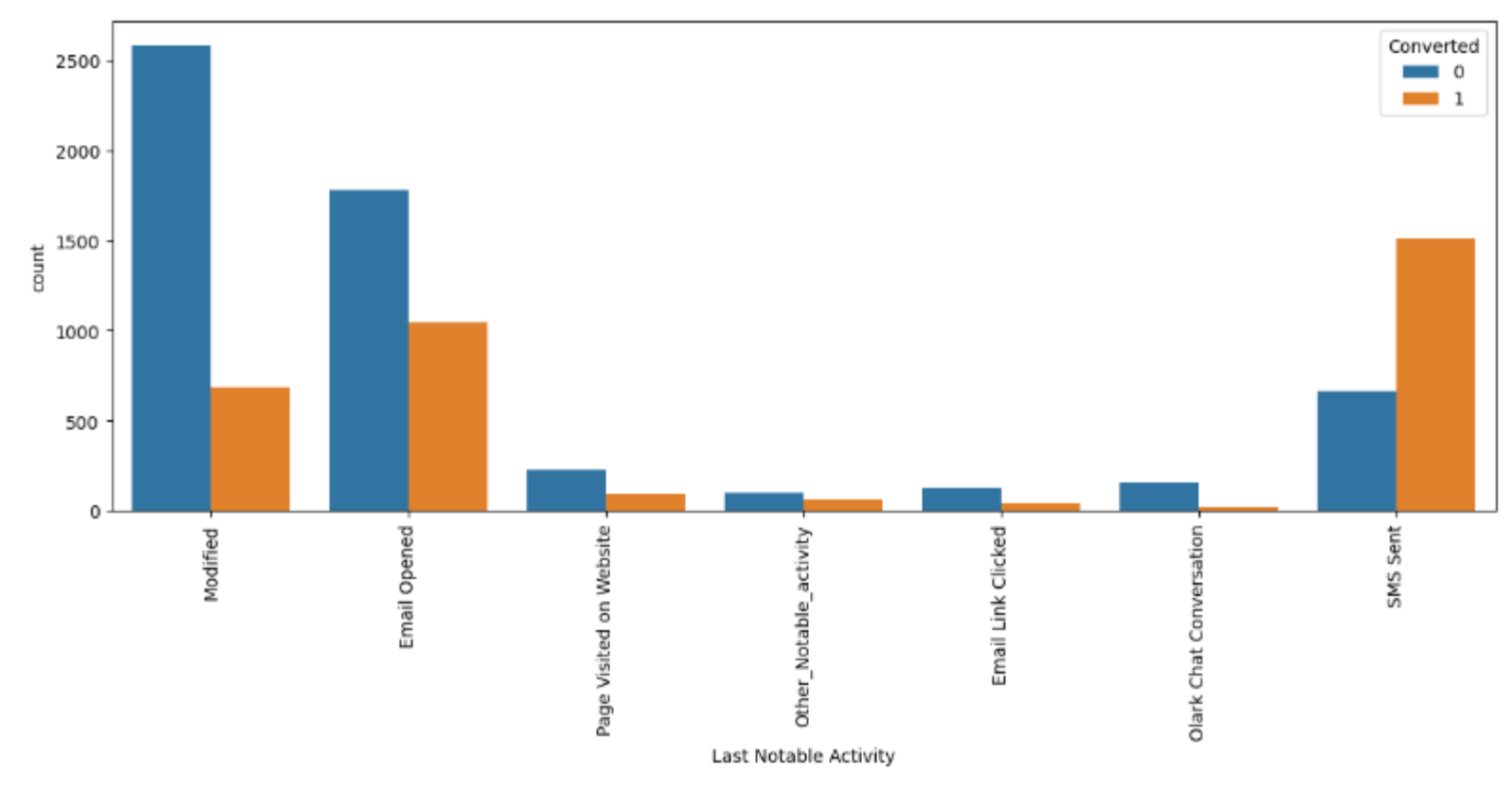


2.8 Lead Origin distribution of Converted and Non-Converted Leads

- API and Landing Page Submission generate a significant number of leads, along with high conversion rates.
- Although Lead Add Form boasts a remarkable conversion rate, the lead count is relatively low.
- Lead Import and Quick Add Form yield minimal leads.
- To enhance the overall lead conversion rate, the focus should be on improving the conversion rates of leads from API and Landing Page Submission sources, while also increasing lead generation from the Lead Add Form.



2.9 Among the last notable activities performed by students, the highest number of both converted and non-converted leads originated from SMS and email.

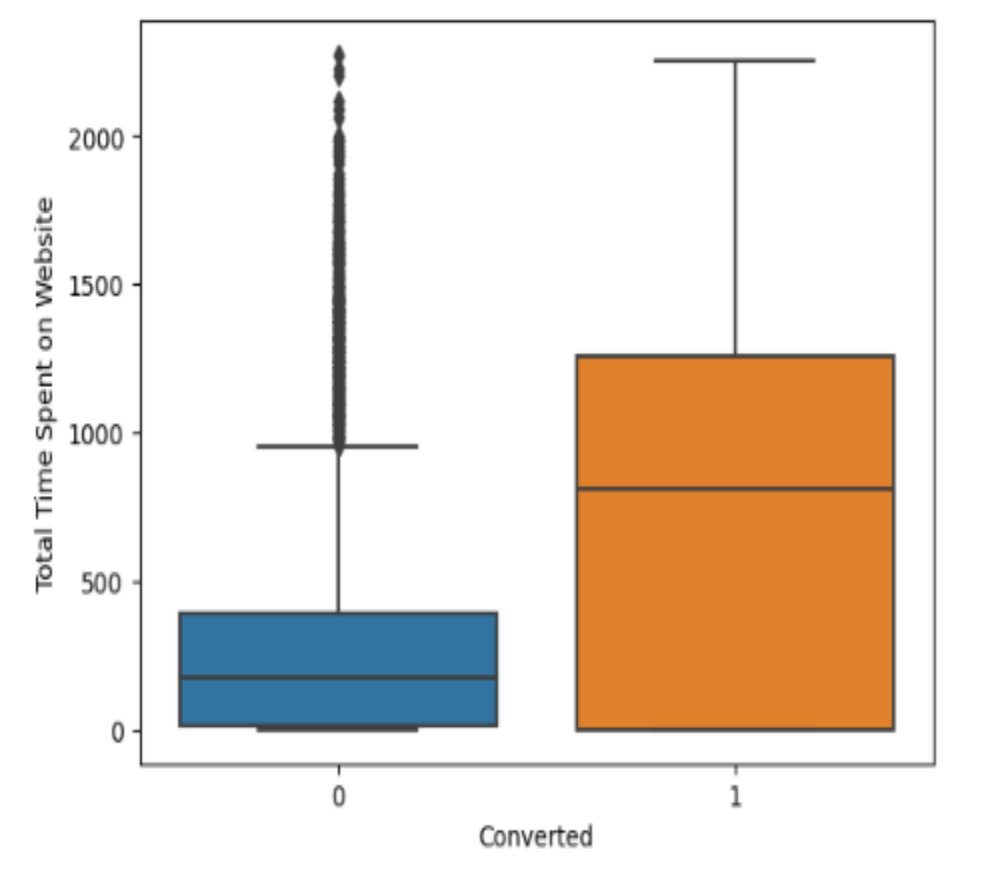


3. NUMERICAL ATTRIBUTES ANALYSIS

3.1 The correlation between numerical columns is visualized using a heatmap

3.2. Plot of Total Time spent on Website Vs the Converted Leads

Leads who spend more time on the website are likelier to convert.
To encourage leads to spend more time, the website should be made more engaging.



4. Logistic Regression Model Building

- To address this issue, we need to develop a logistic regression model that assigns a lead score ranging from 0 to 100 to each lead. This score can then be utilized by the company to target potential leads effectively. A higher score indicates a 'hot' lead, with a higher likelihood of conversion, whereas a lower score suggests a 'cold' lead, less likely to convert.

- The logistic regression model was trained using 70% of the complete dataset and subsequently tested on the remaining 30% of the data. Recursive Feature Elimination (RFE) was employed to select the top 15 variables as output.

Final Output for RFE model building

Generalized Linear Model Regression Results

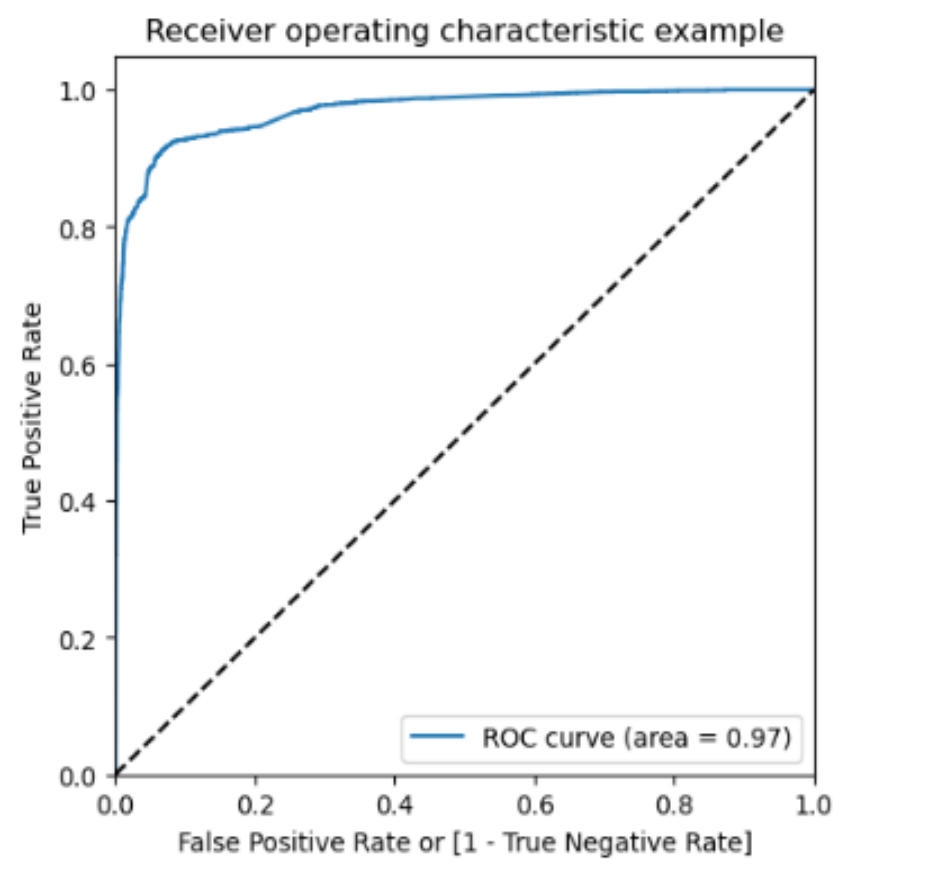
Dep. Variable:	Converted	**No. Observations:**	6267
Model:	GLM	**Df Residuals:**	6253
Model Family:	Binomial	**Df Model:**	13
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1263.3
Date:	Sun, 14 Apr 2024	**Deviance:**	2526.6
Time:	21:46:54	**Pearson chi2:**	8.51e+03
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.6037
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1179	0.084	-13.382	0.000	-1.282	-0.954
Total Time Spent on Website	0.8896	0.053	16.907	0.000	0.786	0.993
Lead Origin_Lead Add Form	1.6630	0.455	3.657	0.000	0.772	2.554
Lead Source_Direct Traffic	-0.8212	0.127	-6.471	0.000	-1.070	-0.572
Lead Source_Welingak Website	3.8845	1.114	3.488	0.000	1.701	6.068
Last Activity_SMS Sent	1.9981	0.113	17.718	0.000	1.777	2.219
Last Notable Activity_Modified	-1.6525	0.124	-13.279	0.000	-1.896	-1.409
Last Notable Activity_Olark Chat Conversation	-1.8023	0.491	-3.669	0.000	-2.765	-0.839
Tags_Closed by Horizzon	7.1955	1.020	7.053	0.000	5.196	9.195
Tags_Interested in other courses	-2.1318	0.406	-5.253	0.000	-2.927	-1.336
Tags_Lost to EINS	5.9177	0.611	9.689	0.000	4.721	7.115
Tags_Other_Tags	-2.3737	0.206	-11.507	0.000	-2.778	-1.969
Tags_Ringing	-3.4531	0.238	-14.532	0.000	-3.919	-2.987
Tags_Will revert after reading the email	4.5070	0.188	24.002	0.000	4.139	4.875

	Features	VIF
1	Lead Origin_Lead Add Form	1.82
12	Tags_Will revert after reading the email	1.56
4	Last Activity_SMS Sent	1.46
5	Last Notable Activity_Modified	1.40
2	Lead Source_Direct Traffic	1.38
3	Lead Source_Welingak Website	1.34
10	Tags_Other_Tags	1.25
0	Total Time Spent on Website	1.22
7	Tags_Closed by Horizzon	1.21
11	Tags_Ringing	1.16
8	Tags_Interested in other courses	1.12
9	Tags_Lost to EINS	1.06
6	Last Notable Activity_Olark Chat Conversation	1.01

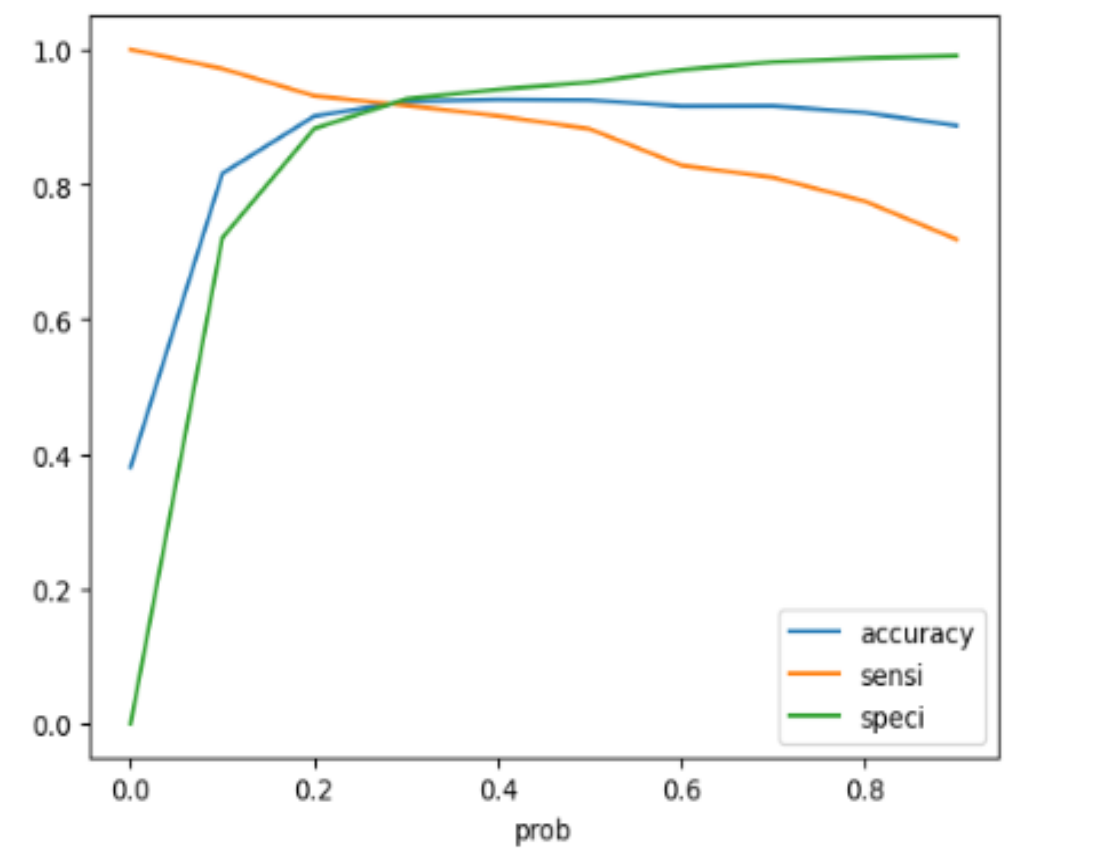
5. ROC Curve

- The ROC curve value, ideally close to 1, is satisfactory at 0.97, indicating a strong predictive model.



Optimal Cut-Off point

- Based on the curve below, the optimal cutoff probability is determined to be 0.3.



Observation

- As observed, the model demonstrates strong performance. With an ROC curve value of 0.97, the model's predictive ability is notable. Additionally, the following metrics were obtained for the training data:
- • Accuracy: 92.29%
- • Sensitivity: 91.70%
- • Specificity: 92.66%
- Further statistics including False Positive Rate, Positive Predictive Value, Negative Predictive Value, Precision, and Recall are provided below.

Final Observation

- Let's contrast the values obtained for the Train and Test datasets:
- • **Train Data**:
- Accuracy: 92.29%
- Sensitivity: 91.70%
- Specificity: 92.66%
- • **Test Data**:
- Accuracy: 92.78%
- Sensitivity: 91.98%
- Specificity: 93.26%
- **The model demonstrates effective prediction of the conversion rate, instilling confidence in the CEO's decision-making based on this model.**

Conclusion

- The top three variables are following:
  - Lead Origin_Lead Add Form
  - Last Activity_SMS Sent
  - Tags_Will revert after reading the email

- The key categorical/dummy variables to focus on in the model are:
  - Lead Origin
  - Lead Source
  - Last Activity
- They should implement the following strategies:
  - Focus on leads where the customer was identified as a lead, such as those from API and Landing Page Submissions.
  - Prioritize leads based on the last activity performed by the customer, including Email Opened and Olark Chat Conversation.
  - Target leads who have engaged with emails by reading or replying, indicating their interest.
  - While approaching students is an option, it's important to note that their conversion probability might be lower due to the industry-based nature of the course. However, this can also serve as motivation for ensuring industry readiness upon completion of their education.