Case Study on Lead Score for X- Education

Prepared By:

Sunder Singh

Dipjit Basak

Dipen Jaysukh Prajapati

Batch : DS C60 –IIIT B(UPGRAD)-2023-24

Problem Statement

1. X Education struggles with a low lead conversion rate despite significant website traffic.

2.A typical lead conversion process shows a significant drop-off from initial leads to paying customers.

3. The company aims to identify 'Hot Leads' with higher conversion potential.

4.X Education needs to prioritize leads for targeted communication to improve conversion rates.

5. The CEO expects a lead conversion rate of around 80%.

6.Many professionals browse X Education's courses but do not convert into paying customers.

7. The sales team faces challenges in effectively communicating with potential leads.

8.X Education receives leads through various channels, including website forms and referrals.

9.The company wishes to build a lead scoring model to prioritize leads based on conversion likelihood.

10. Improving lead conversion efficiency is crucial for X Education's business success.

The Business Objectives for X Education are:

1.Improve lead conversion rate: The company aims to increase the percentage of leads that convert into paying customers from the current rate of around 30% to approximately 80%.

2.Identify 'Hot Leads': X Education seeks to identify the most potential leads, also known as 'Hot Leads', who are most likely to convert into paying customers. This involves building a lead scoring model to prioritize leads based on their conversion likelihood.

3.Optimize communication strategy: To make the lead conversion process more efficient, the company intends to focus more on communicating with potential leads identified as 'Hot Leads' rather than making calls to everyone. This involves nurturing potential leads through effective communication and education about the product.

Solution Methodology

1. Data Cleaning and Data Manipulation:

   - Check for and handle duplicate data.

   - Handle NA values and missing values appropriately.

   - Drop columns with a large number of missing values that are not useful for analysis.

   - Impute missing values if necessary.

   - Check and handle outliers in the data.

2. Exploratory Data Analysis (EDA):

   - Conduct univariate analysis to understand the distribution of variables.

   - Perform bivariate analysis to explore correlations and patterns between variables.

Solution Methodology

3. Feature Scaling & Dummy Variables:

   - Scale features if necessary to ensure uniformity in their range.

   - Create dummy variables and encode categorical data for modeling purposes.

4. Classification Technique:

   - Utilize logistic regression for modeling and prediction, considering its suitability for binary classification problems.

5. Validation of the Model:

- Validate the performance of the model using appropriate validation techniques such as cross-validation or holdout validation.

6. Model Presentation:

   - Present the model along with its key metrics, such as accuracy, precision, recall, and F1-score.

7. Conclusions and Recommendations:

   - Draw conclusions based on the findings from the analysis and model performance.

   - Provide recommendations for potential actions or improvements based on the insights gained from the model.

Data Manipulation

1. Total Number of Rows and Columns: The dataset contains 37 rows and 9240 columns.

2.Drop Single Value Features: Features with only one unique value, such as "Magazine", "Receive More Updates About Our Courses", etc., have been removed as they provide no variability.

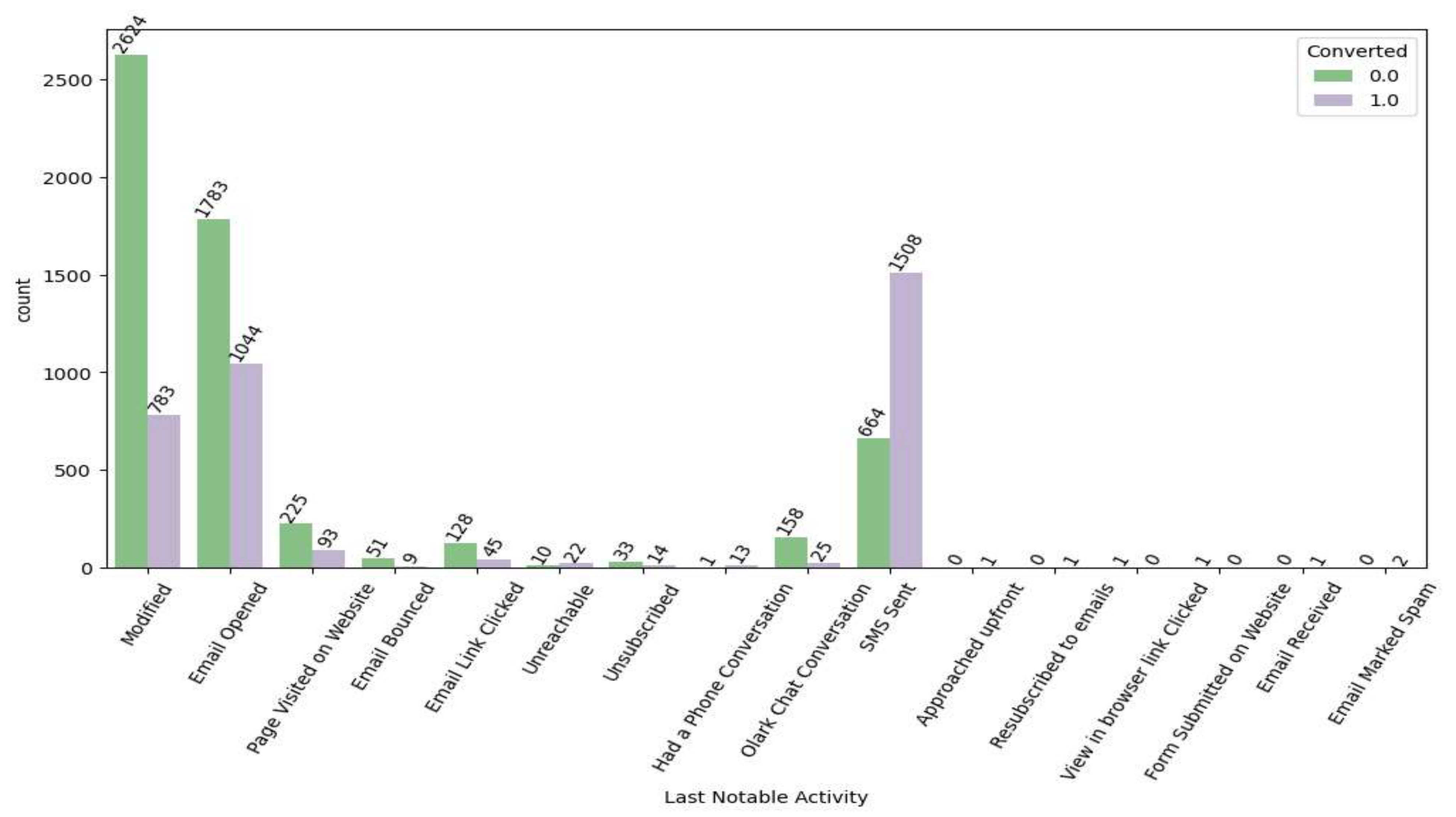
3."Chain Content", "Get updates on DM Content", and "I agree to pay the amount through cheque" have been dropped. These features likely contained information that was not relevant to the analysis or did not contribute significantly to the modeling process. Removing such features helps streamline the dataset and focuses the analysis on the most relevant variables for the problem at hand.

4.Remove Unnecessary Columns: Columns like "Prospect ID" and "Lead Number" are dropped as they are not relevant for analysis.

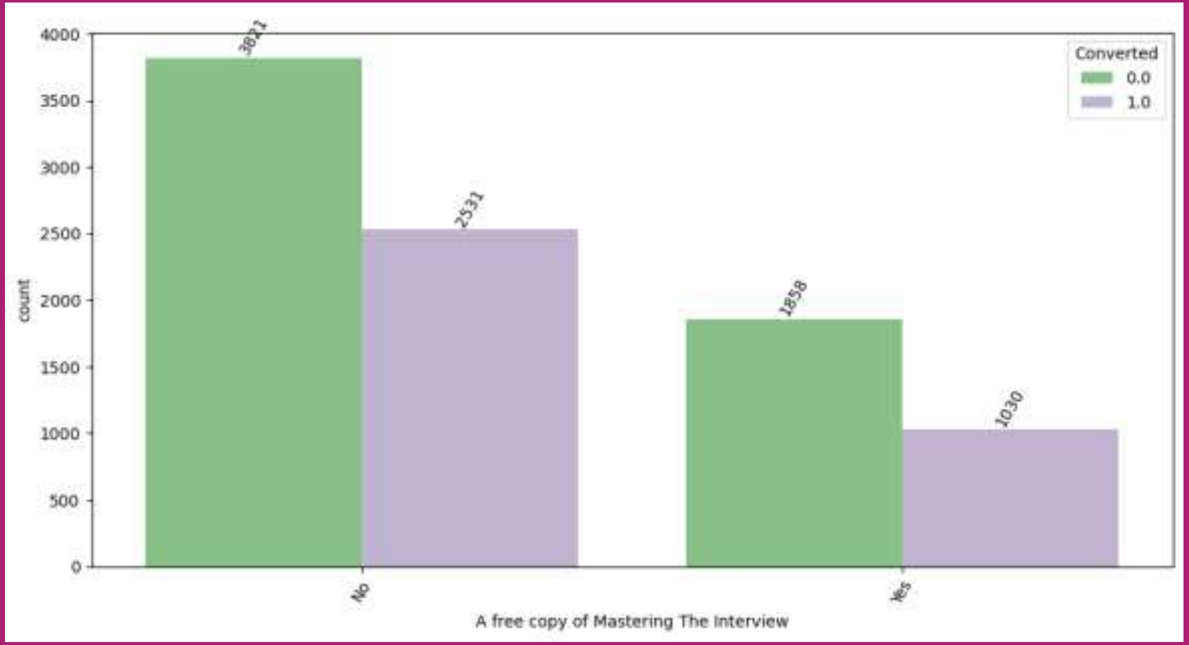
5.Drop Low Variance Features: Object type variables with low variance, such as "Do Not Call", "What matters most to you in choosing course", etc., have been dropped as they do not provide useful information.

6.Drop Columns with High Missing Values: Columns with more than 35% missing values, such as 'How did you hear about X Education' and 'Lead Profile', are removed to avoid bias in the analysis.

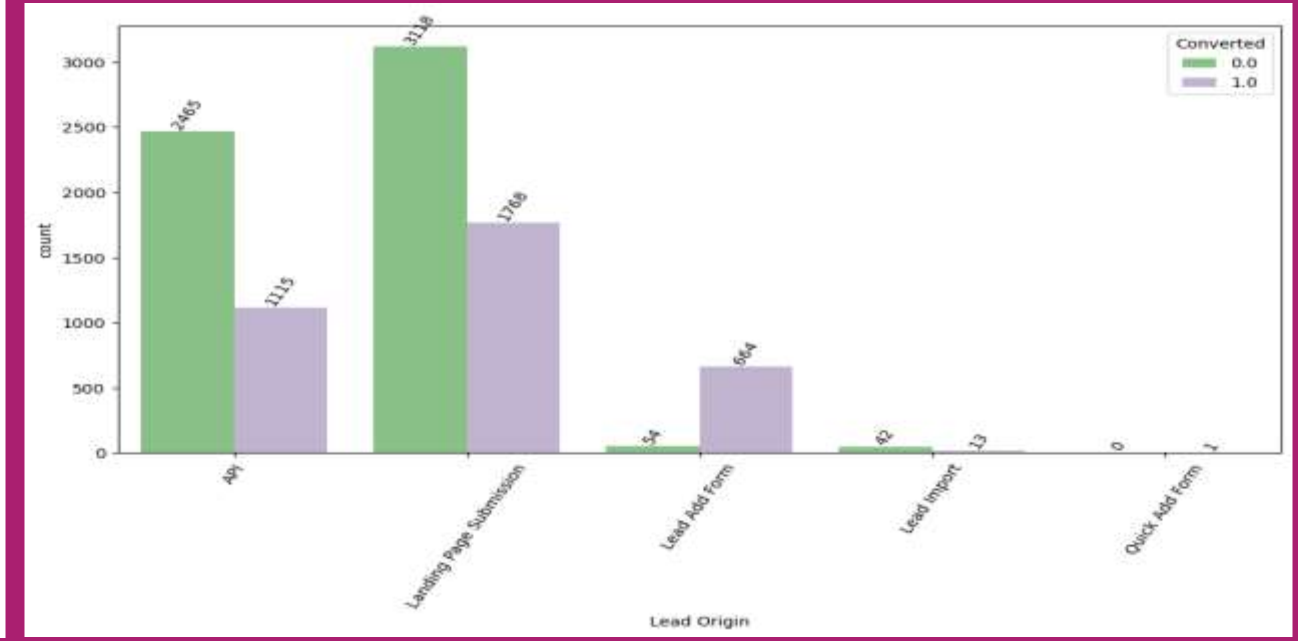
EDA


count
2500
2000
1500
1000
500
0
Converted
🟩 0.0
▨ 1.0
2624
783
1783
1044
225
93
51
9
128
45
10
22
33
14
1
13
158
25
664
1508
0
1
0
1
1
0
1
0
0
1
0
2
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

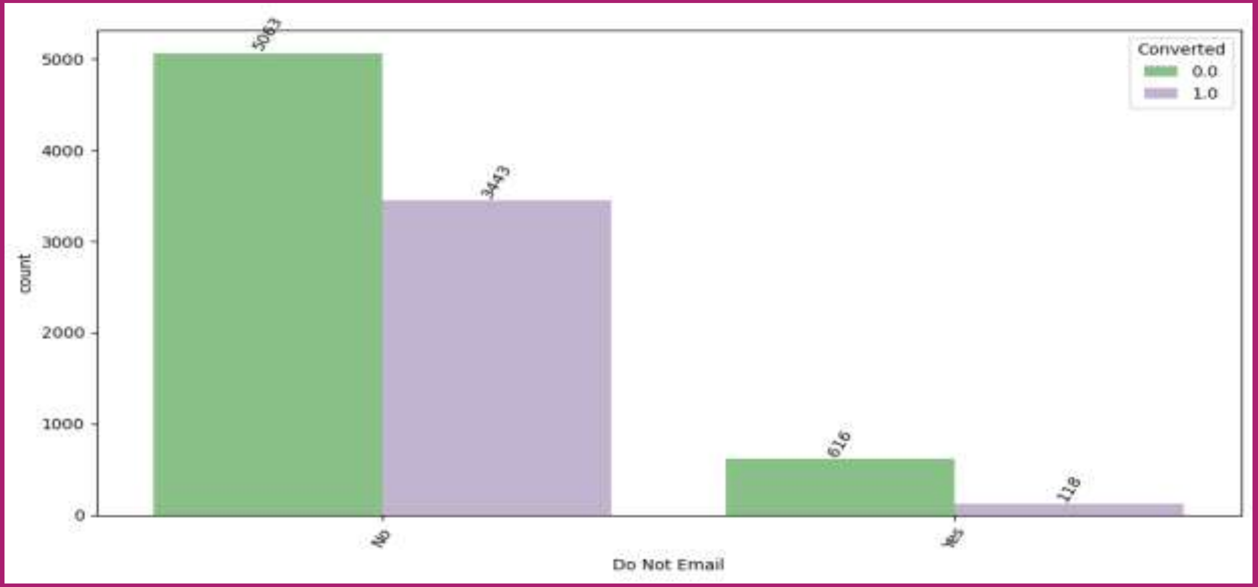
EDA



count
4000
3500
3000
2500
2000
1500
1000
500
0
3821
2531
1858
1030
No
Yes
A free copy of Mastering The Interview
Converted
🟩 0.0
🟪 1.0

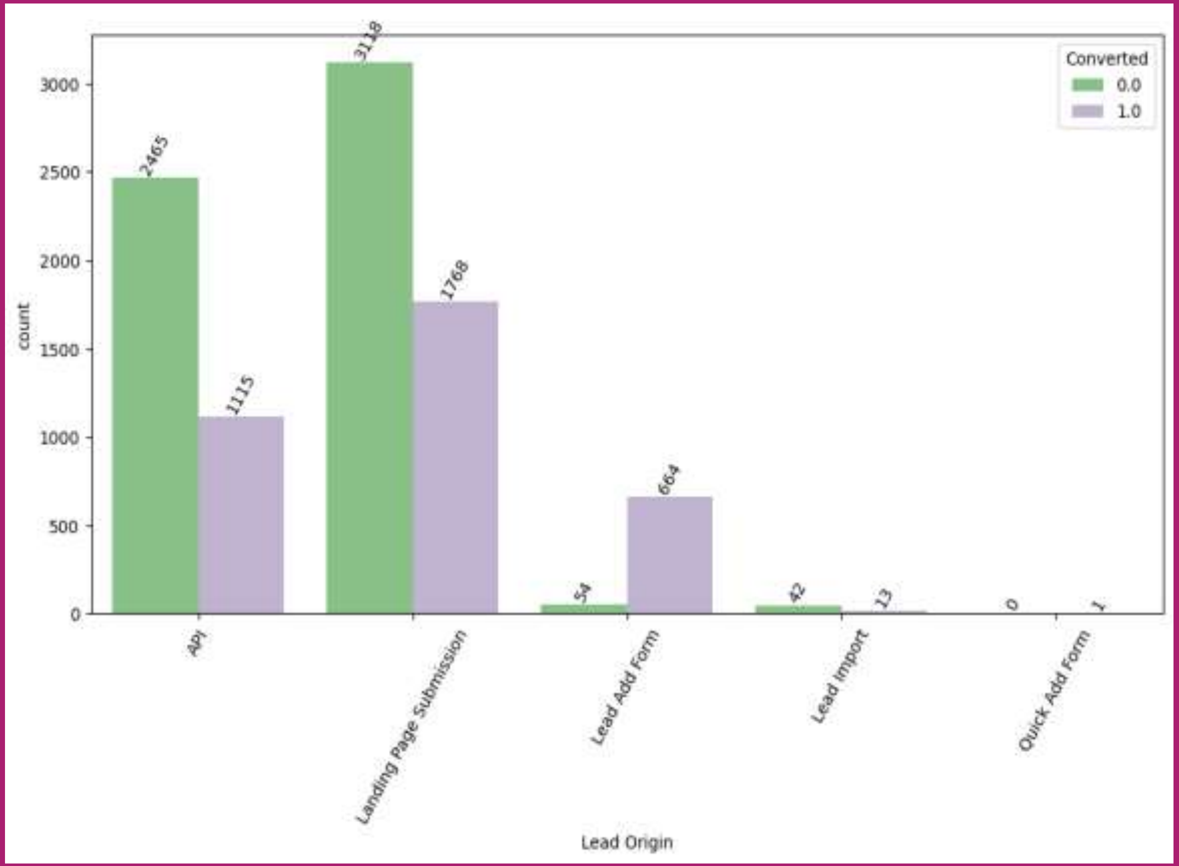


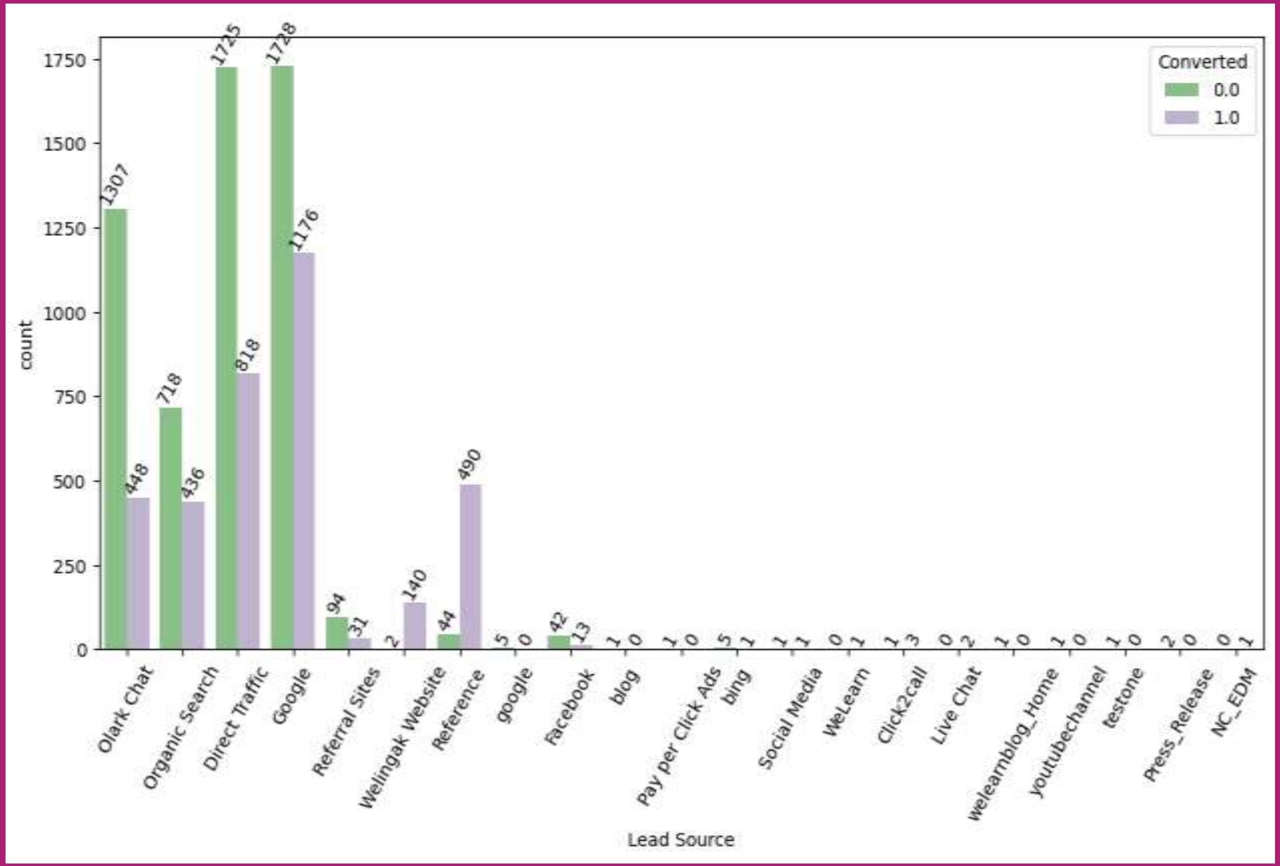
count
3000
2500
2000
1500
1000
500
0
2465
1115
3118
1768
54
664
42
13
0
1
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Origin
Converted
🟩 0.0
🟪 1.0



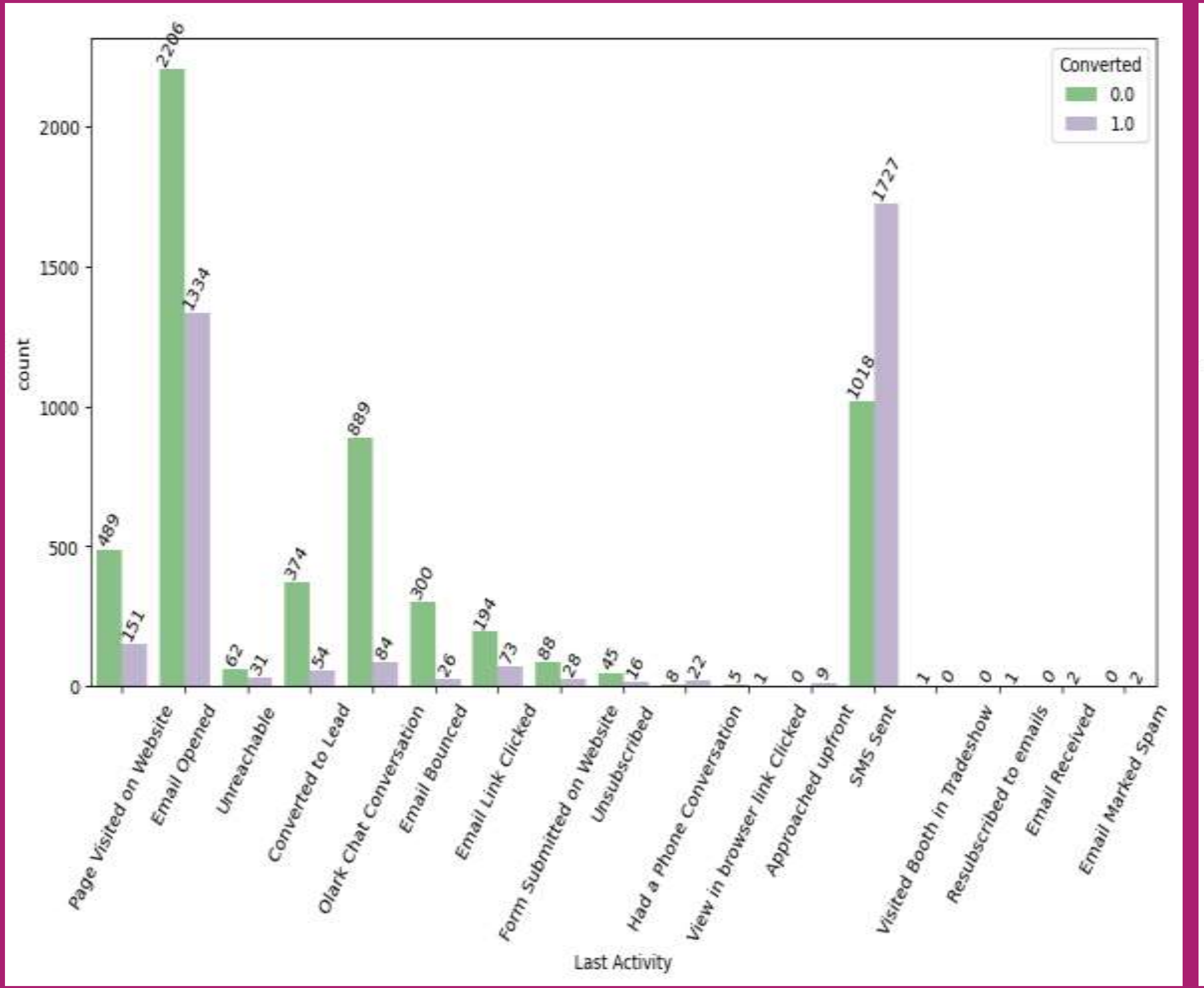
count
5000
4000
3000
2000
1000
0
5063
3443
616
118
No
Yes
Do Not Email
Converted
🟩 0.0
🟪 1.0

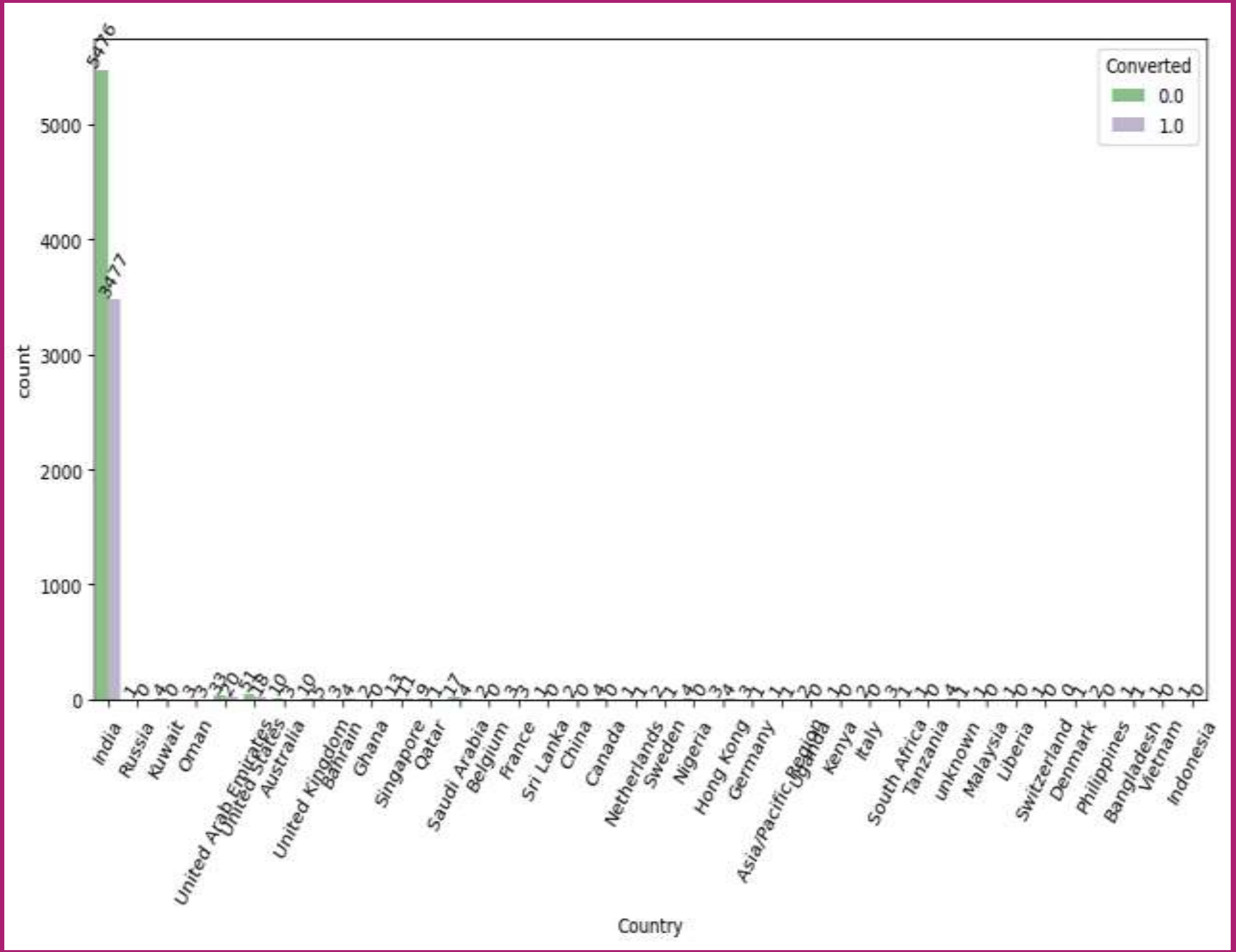
Categorical Variable Relation





Categorical Variable Relation





DATA Conversion

- Numerical variables are normalized to bring them to a standard scale, ensuring consistent comparisons.

- Object type variables are converted into dummy variables, representing categorical data in a format suitable for analysis.

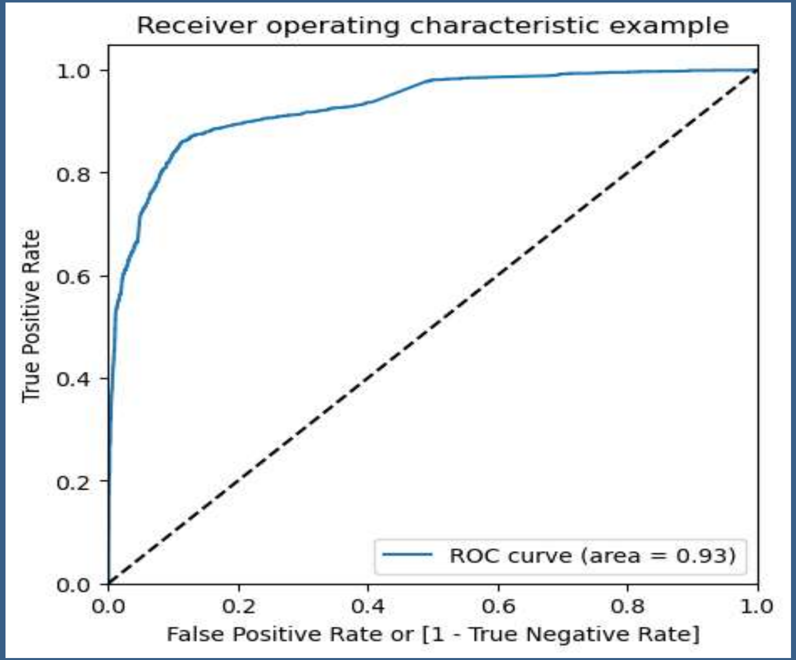
- After conversion, the dataset contains 9240 rows and 150 columns, ready for further analysis.

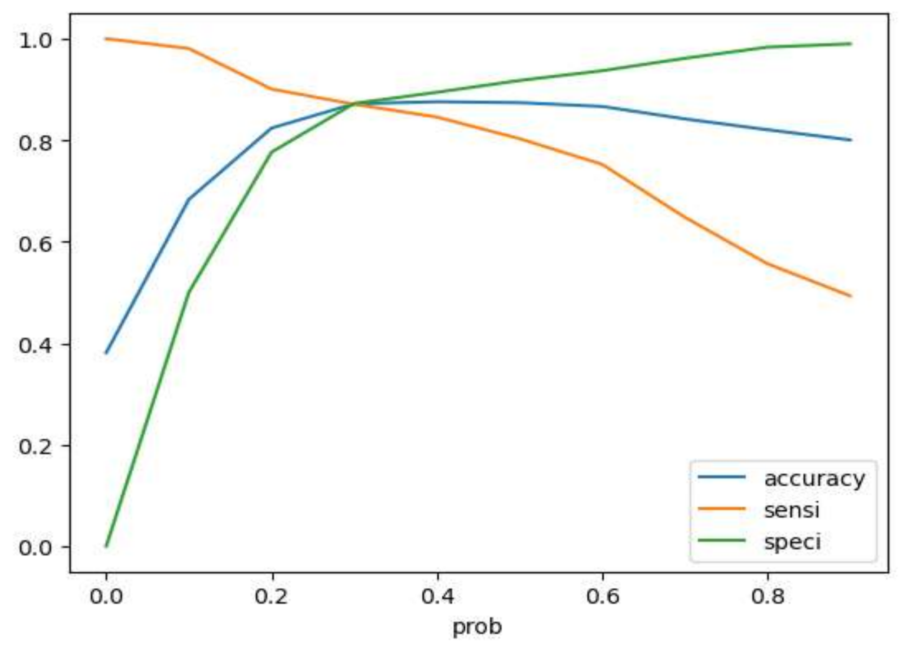
MODEL BUILDING

- The data is split into training and testing sets using a 70:30 ratio.
- Recursive Feature Elimination (RFE) is employed for feature selection, with 15 variables selected as output.
- A model is built by eliminating variables with p-values greater than 0.05 and VIF values greater than 5.
- Predictions are made on the test dataset.
- The overall accuracy achieved is 87%.

Generating the ROC Curve

- The optimal cut-off point is determined based on achieving a balance between sensitivity and specificity.
- This cut-off probability is where we achieve a balanced trade-off between sensitivity and specificity.
- From the ROC curve, it is observed that the optimal cut-off point is at a probability of 0.3.





Conclusion

The list of important features from our final model indicates the variables that have the most significant impact on lead conversion. Based on their coefficients, we can derive the following recommendations:

1. **Prioritize leads from "Welingak Websites" and "Reference"** as they have the highest positive coefficients, suggesting a strong correlation with conversion.
2. **Focus on contacting "working professionals"** as they show a significantly higher likelihood of conversion compared to other occupations.
3. **Allocate resources to engage leads who spend more time on the website**, as indicated by the positive coefficient for "Total Time Spent on Website."
4. **Pay attention to leads from "Olark Chat"**, which also have a positive coefficient, indicating a higher conversion rate associated with this lead source.
5. **Target leads whose last activity was "SMS Sent,"** as they demonstrate a higher propensity to convert compared to other activities.

Conclusion

6. **Avoid contacting leads with last activity as "Olark Chat Conversation,"** as indicated by the negative coefficient, suggesting a lower likelihood of conversion.

7. **Refrain from reaching out to leads with a lead origin of "Landing Page Submission"** as they have a negative coefficient, indicating a lower conversion probability.

8. **Exclude leads with specialization listed as "Others,"** as they are less likely to convert based on the negative coefficient associated with this category.

9. **Do not engage leads who opted for "Do not Email" as "yes,"** as indicated by the negative coefficient, suggesting a lower likelihood of conversion for leads with this preference.

- By implementing these recommendations, X-Education can optimize its lead conversion strategy, focusing resources on leads with the highest probability of conversion, thereby maximizing sales efficiency and improving overall business performance.

THANK YOU

Team

Sunder Singh

Dipjit Basak

Dipen Jaysukh Prajapati

Batch : DS C60 –IIIT B(UPGRAD)-2023-24