Lead Scoring Case Study

- **Y PRADEEP KUMAR**

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%.
- Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Approach

- The model is being built for Company X Education to identify strategies for converting potential users. The goal is to understand and validate the data to draw conclusions that help target the correct audience and improve the conversion rate. Let's discuss the steps followed in this process:

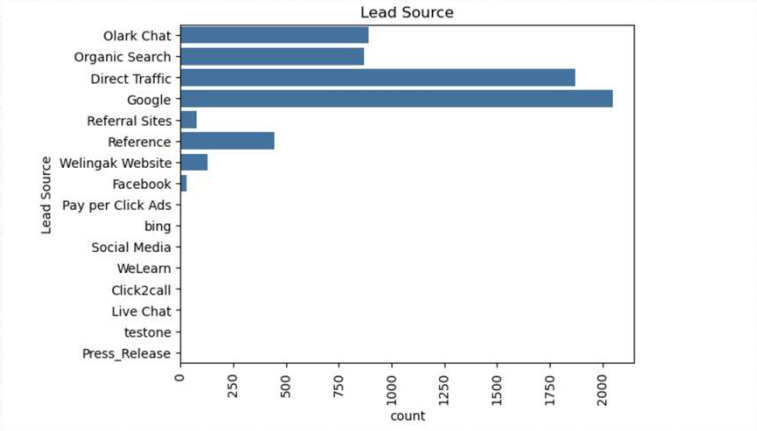
EDA:

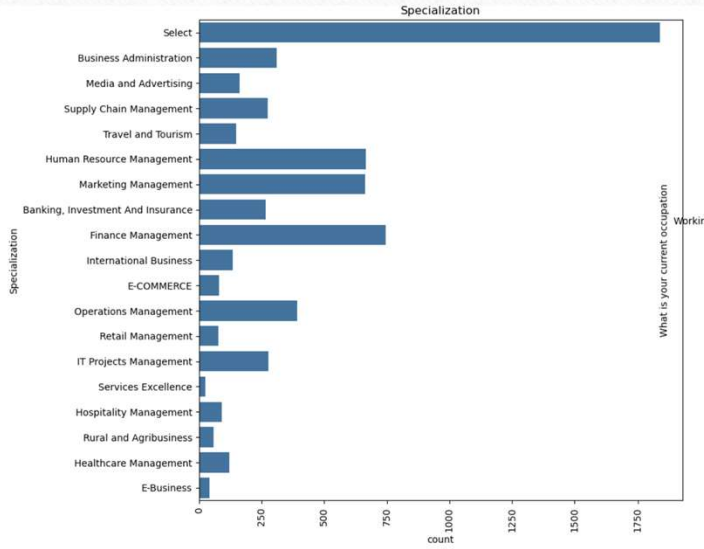
- Checked for missing values in the data set and then Dropped all the columns which is having missing values greater than 3000.
- Here we can see Prospect ID and Lead Number are having 0 null values and having unique values, but these will not be helpful to us. So, we dropped these 2 columns.

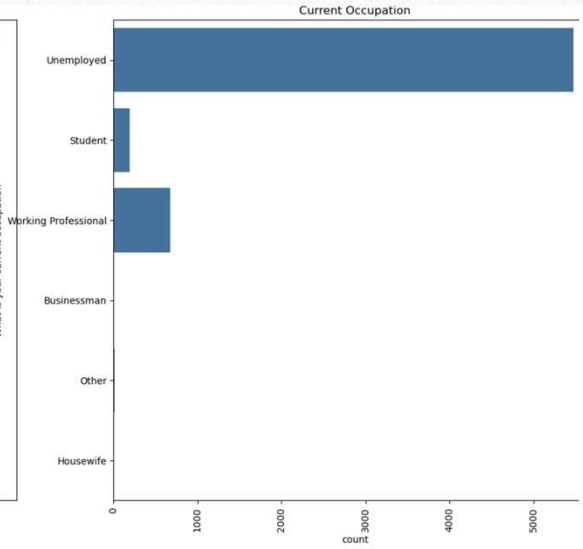
- Also noticed that when we got the value counts of all the columns, there were a few columns in which only one value was majorly present for all the data points. These include Do Not Call, Search, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, 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. Since practically all of the values for these variables are No, it's best that we drop these columns as they won't help with our analysis.
- Dropped the null value rows present in the variable 'What matters most to you in choosing a course', 'Total Visits', 'Lead Source',' What is your current occupation', ' Specialization'.
- We also worked on numerical variable and dummy variables.

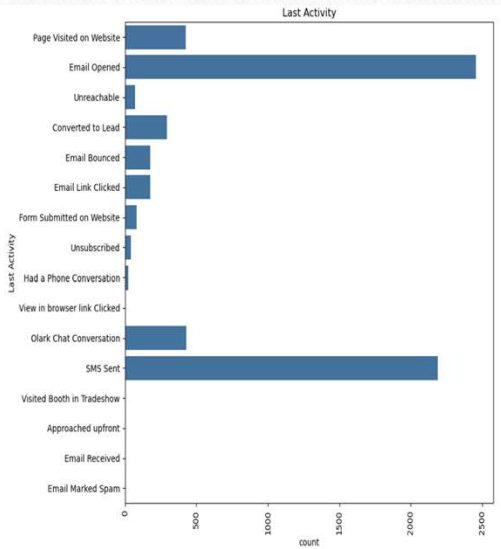
After cleaning the data we have represented the data which is present in visualized way as follows:

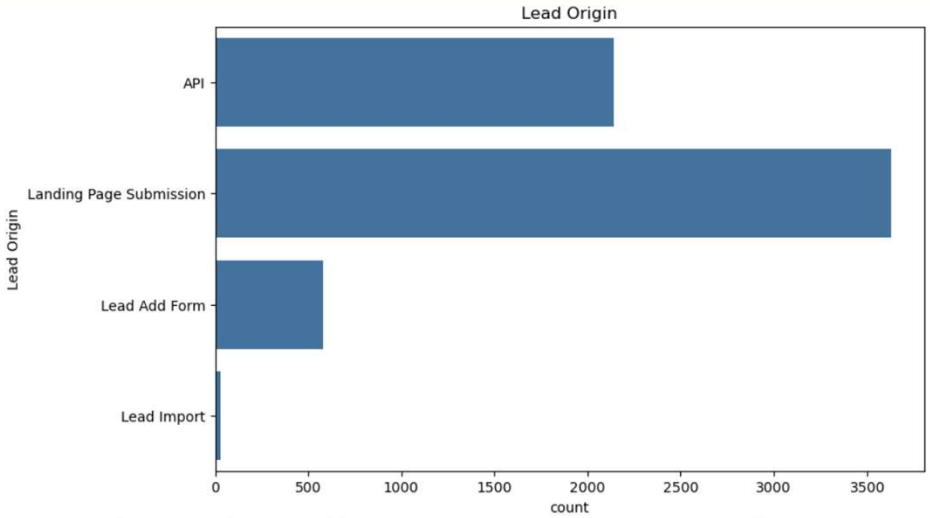
Let us see the categorical columns representation:

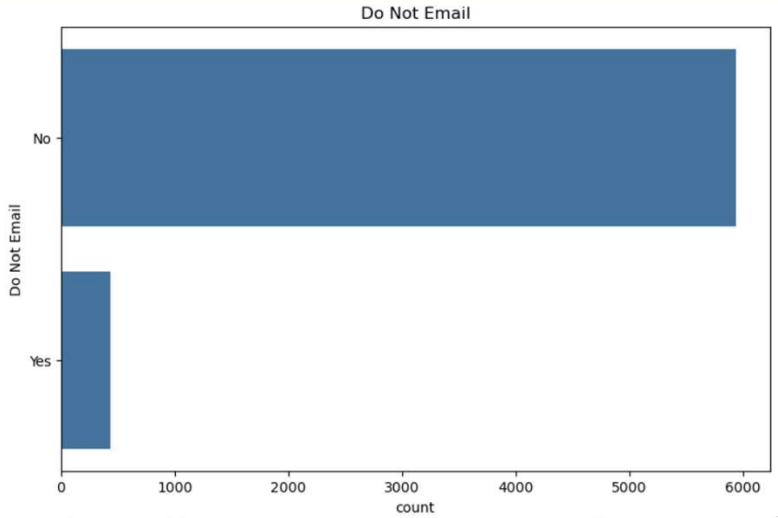


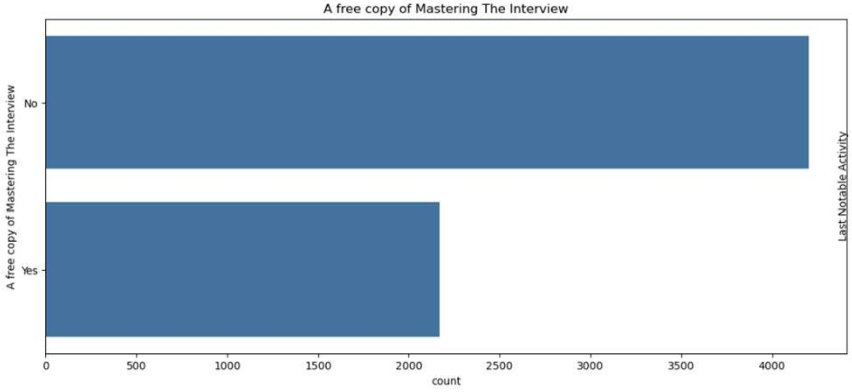

Specialization
Select
Business Administration
Media and Advertising
Supply Chain Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
Finance Management
International Business
E-COMMERCE
Operations Management
Retail Management
IT Projects Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization
0
250
500
750
1000
1250
1500
1750
count

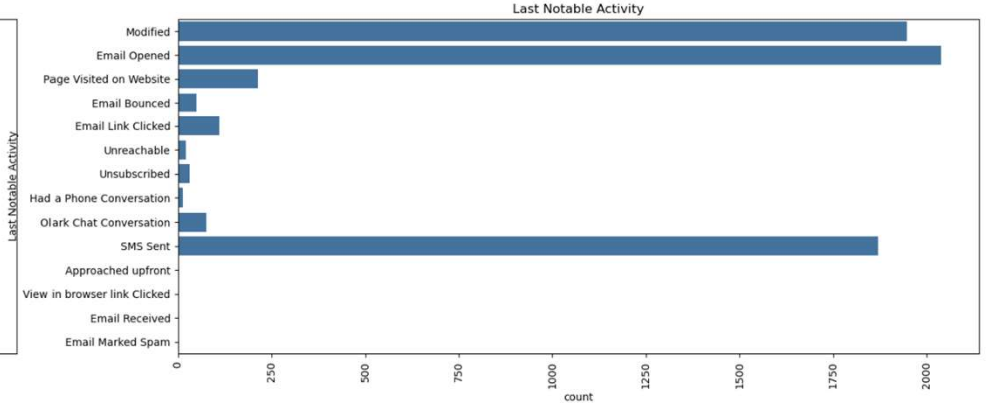

Current Occupation
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation
0
1000
2000
3000
4000
5000
count


Last Activity
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
Olark Chat Conversation
SMS Sent
Visited Booth in Tradeshow
Approached upfront
Email Received
Email Marked Spam
Last Activity
0
500
1000
1500
2000
2500
count

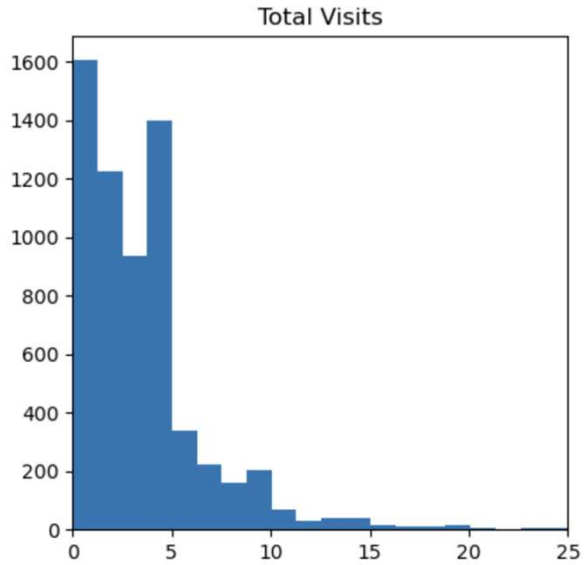

Lead Origin
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin
0
500
1000
1500
2000
2500
3000
3500
count

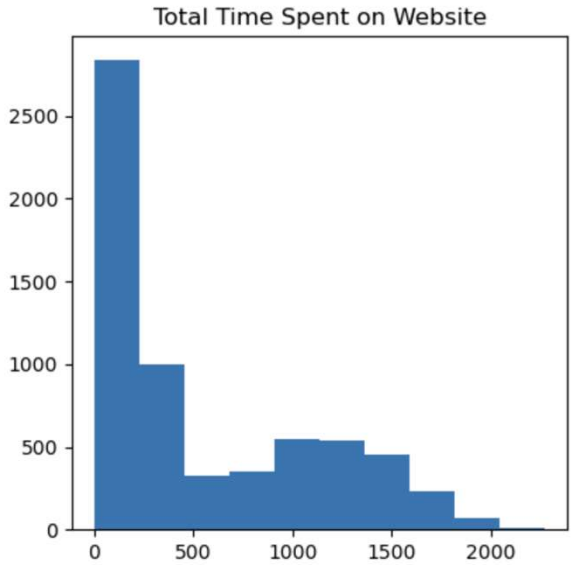

Do Not Email
No
Yes
Do Not Email
0
1000
2000
3000
4000
5000
6000
count

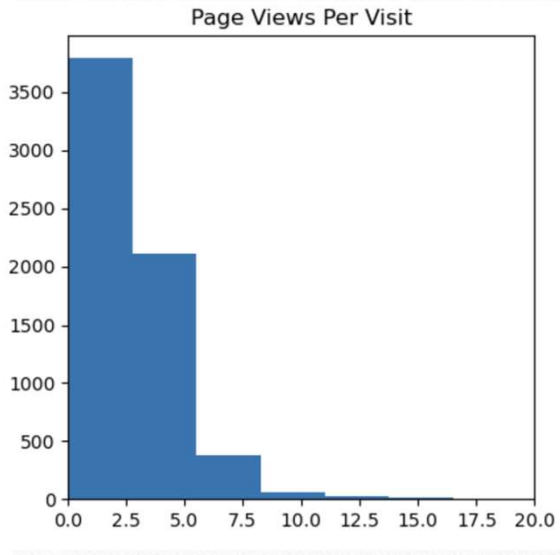

A free copy of Mastering The Interview
No
Yes
A free copy of Mastering The Interview
0
500
1000
1500
2000
2500
3000
3500
4000
count


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
View in browser link Clicked
Email Received
Email Marked Spam
Last Notable Activity
0
250
500
750
1000
1250
1500
1750
2000
count

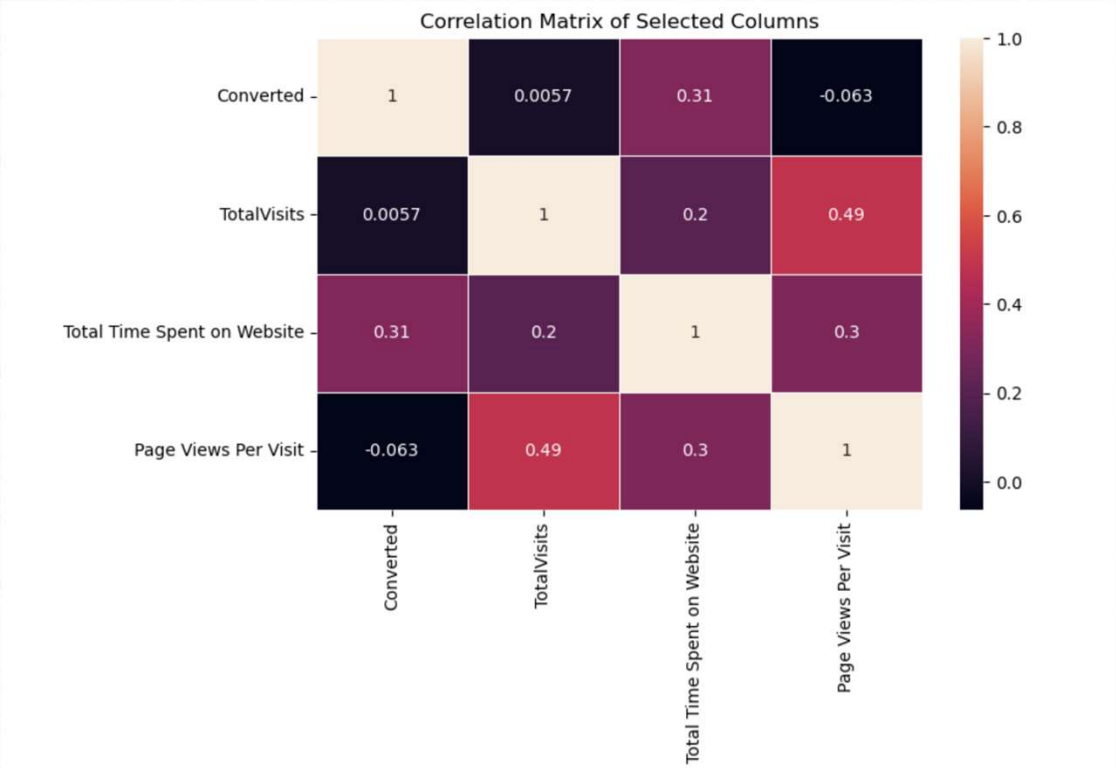
Numerical columns representation using Histogram







Correlation Matrix



Train-Test split & Scaling:

- The data was split into 70% for training and 30% for testing.
- Min-max scaling was applied to the variables: ['TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit'].

Model Building

- RFE was used for feature selection.
- Then RFE was done to attain the top 15 relevant variables.
- Later the rest of the variables were removed manually depending on the VIF values and p-value.

Model Evaluation

Sensitivity – Specificity Evaluation:

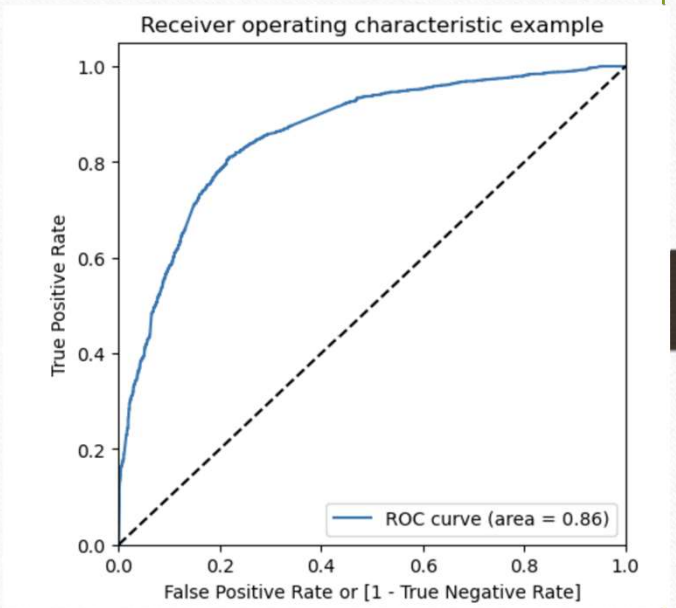
For the Sensitivity-Specificity evaluation, we utilized the ROC curve to determine the optimal cutoff point.

Training Data:

- The area under the ROC curve (AUC) was 0.86, indicating strong model performance.
- After plotting the ROC curve, the optimal cutoff value was found to be 0.42, resulting in the following metrics:
  - **Accuracy: 79.5%**
  - **Sensitivity (True Positive Rate): 80.9%**
  - **Specificity (True Negative Rate): 78.2%**

Test Data:

- When evaluating the model on the test data, we observed:
  - **Accuracy: 78.4%**
  - **Sensitivity: 79.6%**
  - **Specificity: 77.3%**



Precision – Recall Evaluation:

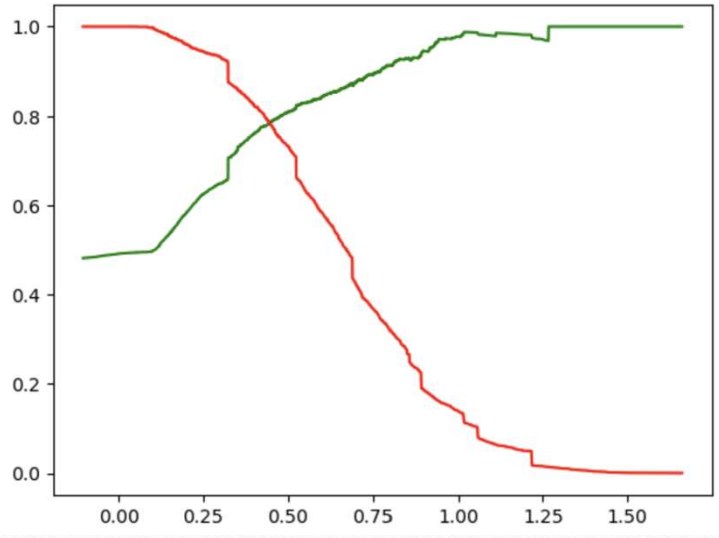
For Precision-Recall evaluation, we focused on optimizing the balance between precision and recall.

Training Data:

- **Initially, the precision and recall values were:**
    - **Precision: 80.9%**
    - **Recall: 73.1%**
- To improve these values, we adjusted the cutoff value. After plotting, the optimal cutoff was determined to be 0.44, yielding:
    - **Accuracy: 79.1%**
    - **Precision: 78.09%**
    - **Recall: 78.9%**

Test Data:

- On the test data, we obtained:
    - **Accuracy: 78.6%**
    - **Precision: 77.5%**
    - **Recall: 77.8%**



- When evaluating with Sensitivity-Specificity, the optimal cutoff value is 0.42.
- When evaluating with Precision-Recall, the optimal cutoff value is 0.44.

The choice of evaluation metric depends on the business objective—whether maximizing sensitivity (recall) or precision is more important for conversion rates at Company X Education.

CONCLUSION

TOP 3 VARIABLE CONTRIBUTING TO CONVERSION:

1. Total Time Spent on Website: This feature shows the most significant impact on lead conversion. A higher time spent on the website indicates greater interest and engagement, which increases the likelihood of conversion.

2. Lead Origin_Lead Add Form: Leads generated through this origin tend to have a much higher conversion rate. This reflects a clear intent to enroll or inquire further.

3. TotalVisits: Although not as significant as the other two, this feature still plays a crucial role, as more visits often correlate with higher engagement, leading to conversion.

The model appears to predict the conversion rate effectively, providing the company with confidence to make informed decisions based on its outcomes