Lead Scoring Case Study

- Rohit Patnaik
  - rohitpatnaik27@gmail.com
  - +91-8895-291-619

- Yoshitha Palavali
  - yoshi.palavali@gmail.com
  - +91-8096-579-756

Problem Statement

- X Education sells online courses to industry professionals and Leads are genberated from various sources are captured.
- To make this process more efficient, the X Education company wants to identify the potential leads so that the rate of conversion can be higher.
- If the company identifies the potential leads, the employees will be focusing more on communicating with potential leads rather than writing emails or giving call to every user.

Solution Approach

1. Understanding the data

2. Clean the data

2.1. Handle the duplicate data, missing values and drop columns which are not relevant for our analysis.

3. Model Building Preparation

3.1. Univariate and Bivariate Data analysis.

4. Model Building

4.1. Scaling and Dummy variables

4.2. Split the data into Train and Test

5. Model Evaluation

5.1. Creation of Confusion matrix and find the overall accuracy, specificity and sensitivity of the data.

6. Prediction on Test set

6.1. Prediction performed on the Test data and then find the overall accuracy, specificity and sensitivity of the data.

7. Conclusion

Understand the Data

Leads.csv contains all the leads generated through various sources. This file contains the following:

- The file contains 9240 rows and 37 columns
- Out of 37 columns, 7 are numerical columns and rest 30 are categorical columns.

Leads Data Dictionary.csv file describes the meaning of all the variables involved the Leads dataset.

Data Cleaning and Preparation

The columns were analysed and the columns where only one unique value is present have been dropped. Below are the columns:

1. Magazine
2. Receive More Updates About Our Courses
3. Get updates on DM Content
4. Update me on Supply Chain Content
5. I agree to pay the amount through cheque

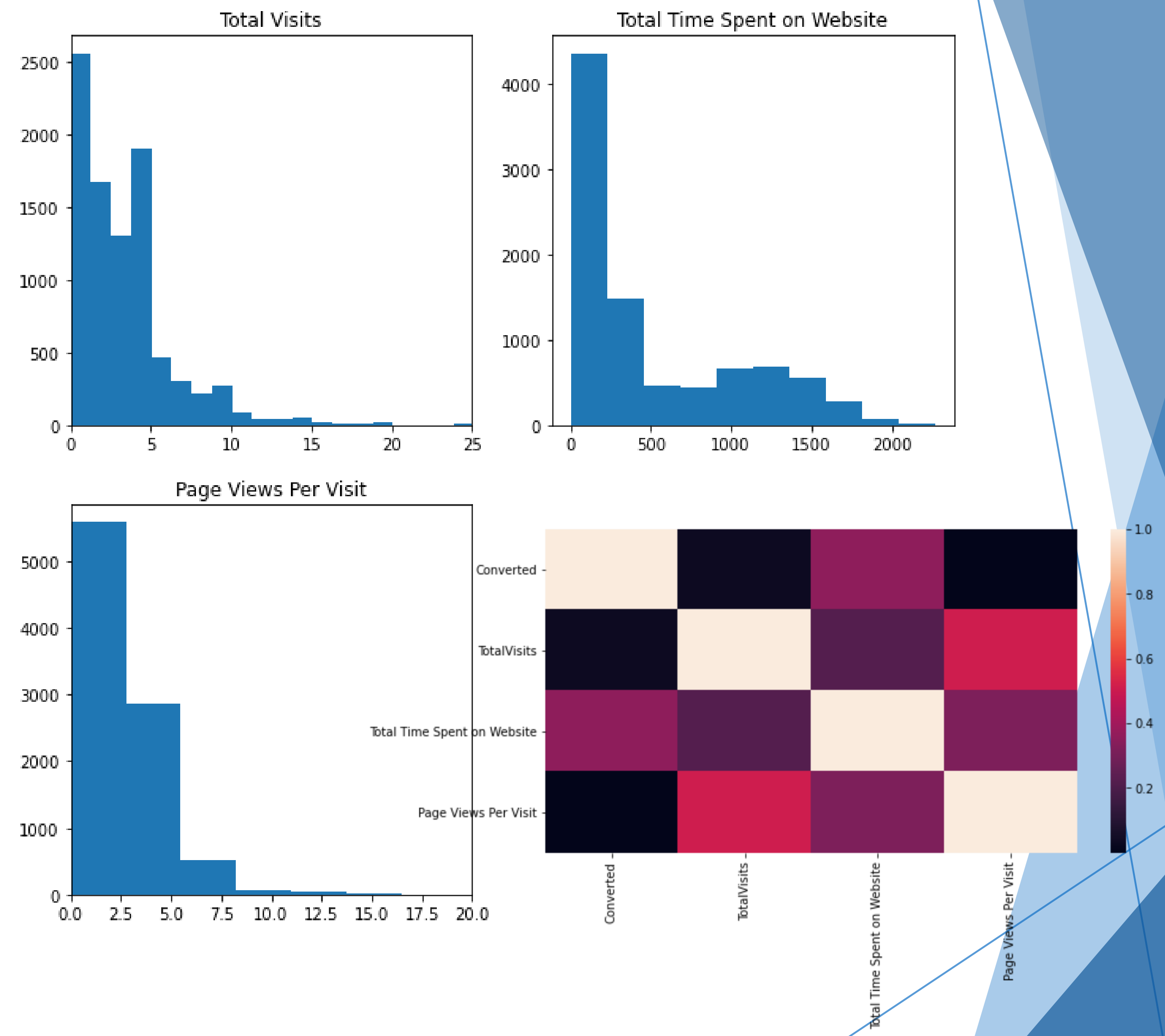
Below are the columns which are having more than 35% of the null values and the columns were dropped:

1. How did you hear about X Education
2. What is your current occupation
3. What matters most to you in choosing a course
4. Tags
5. Lead Quality
6. Lead Profile
7. City
8. Asymmetrique Activity Index
9. Asymmetrique Profile Index
10. Asymmetrique Activity Score
11. Asymmetrique Profile Score

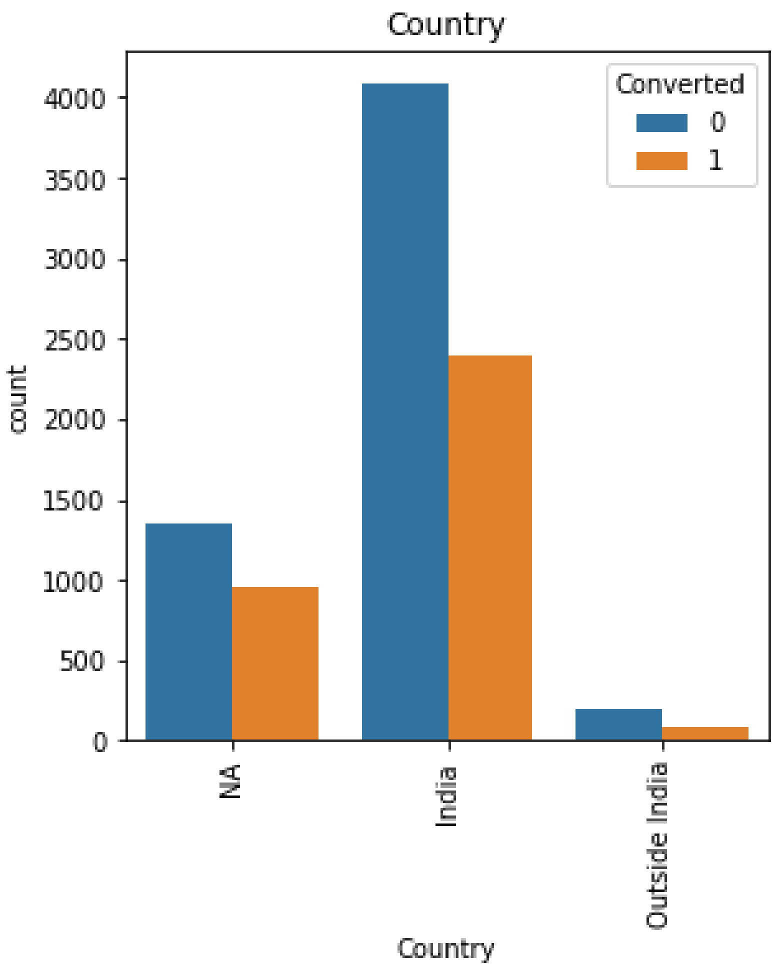
Following are the columns where Select value has been replaced with NA and later removed these data from the analysis:

1. Specialization
2. How did you hear about X education
3. Lead Profile
4. Country

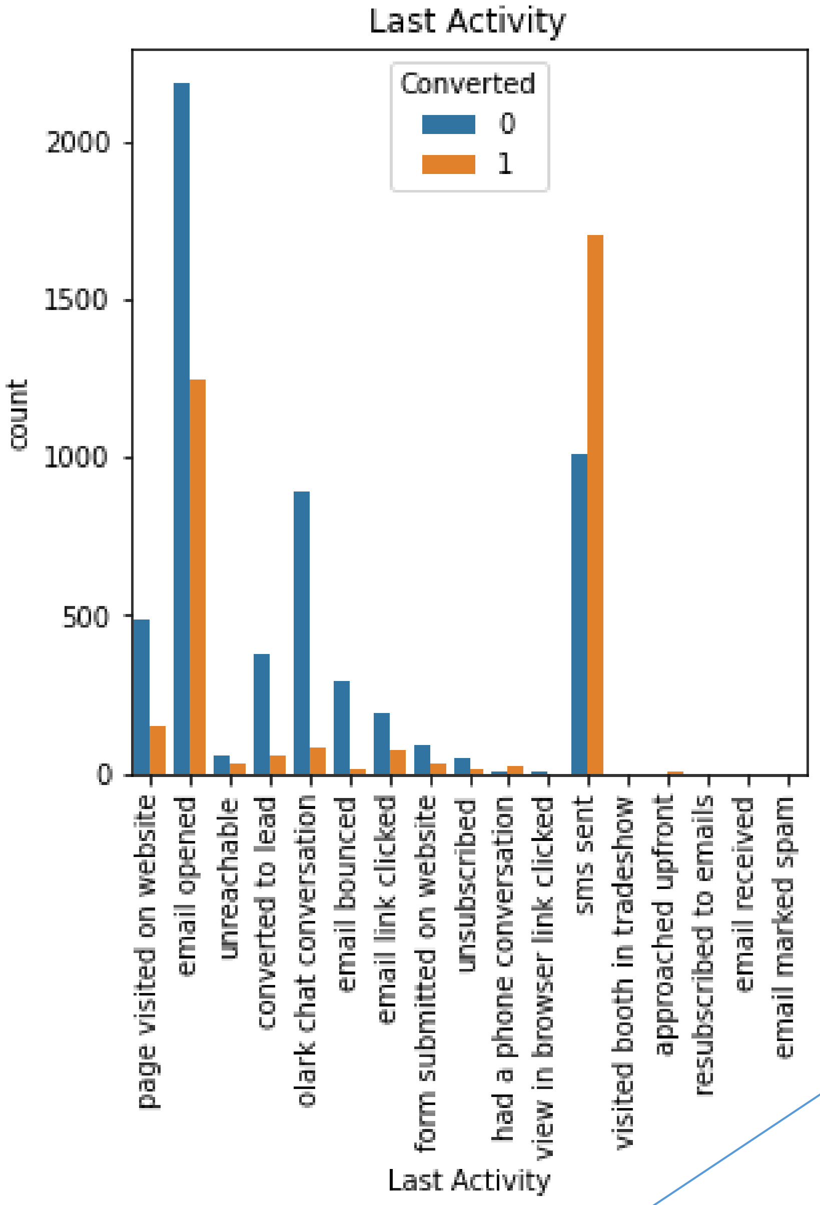
Exploratory Data Analysis



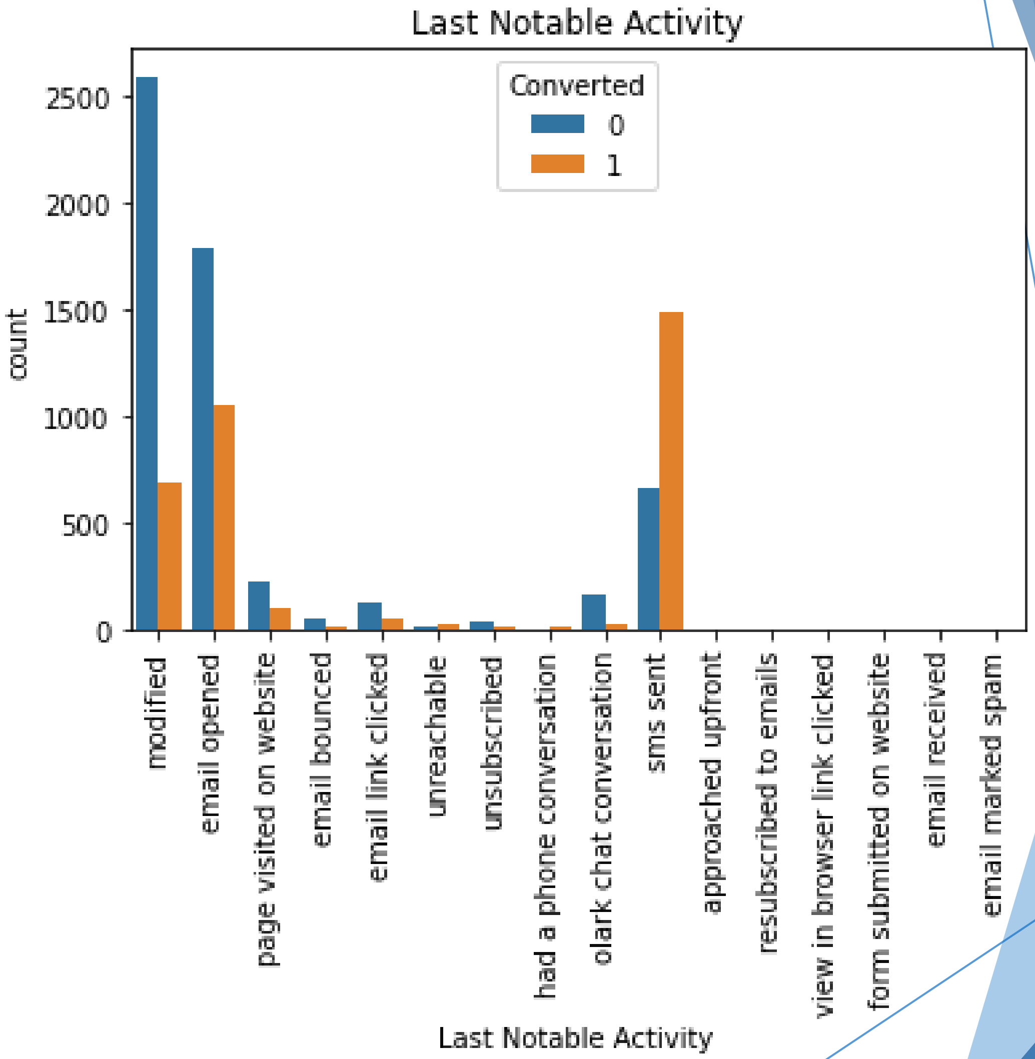
Exploratory Data Analysis



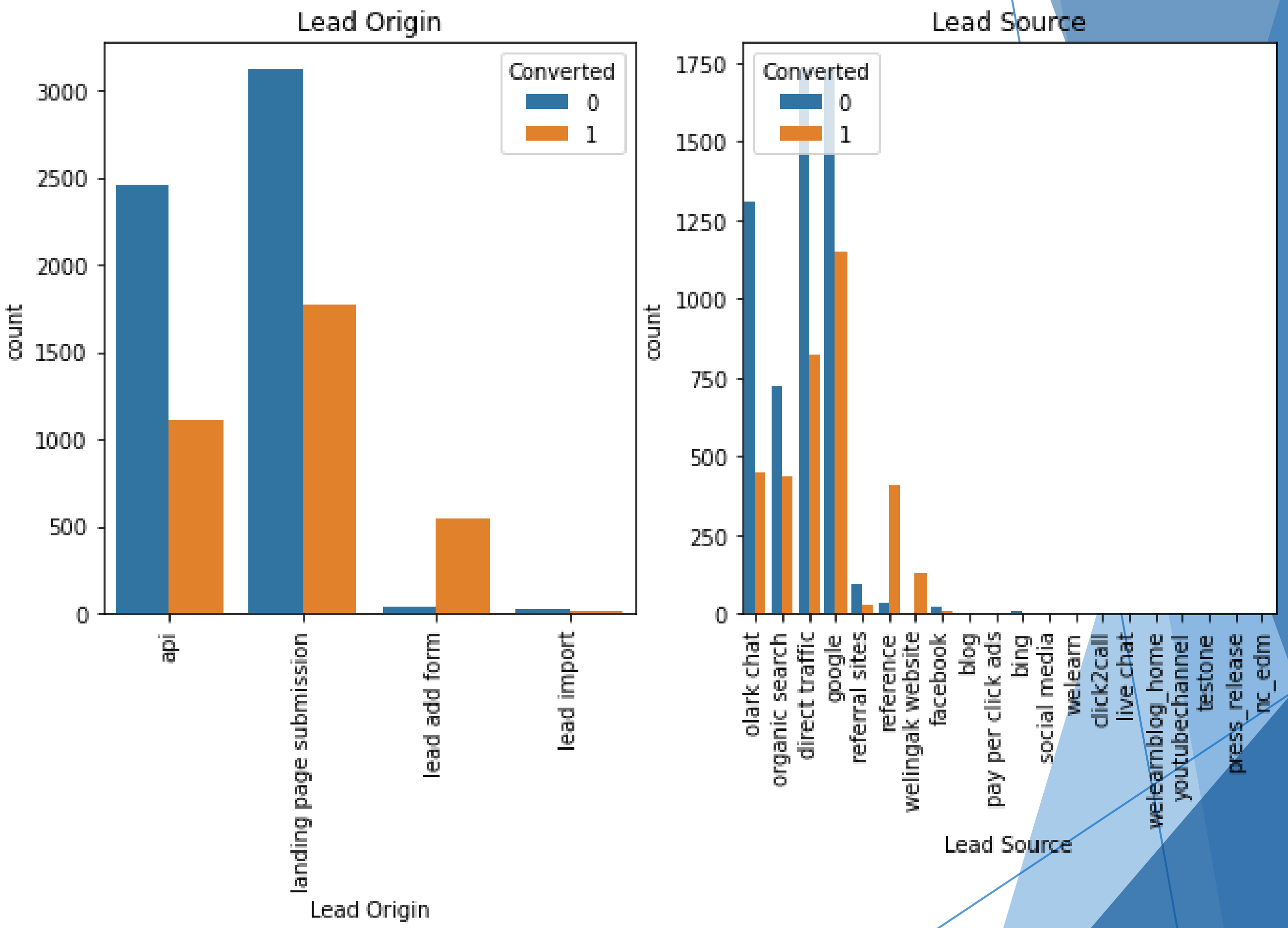
Exploratory Data Analysis



Exploratory Data Analysis



Exploratory Data Analysis



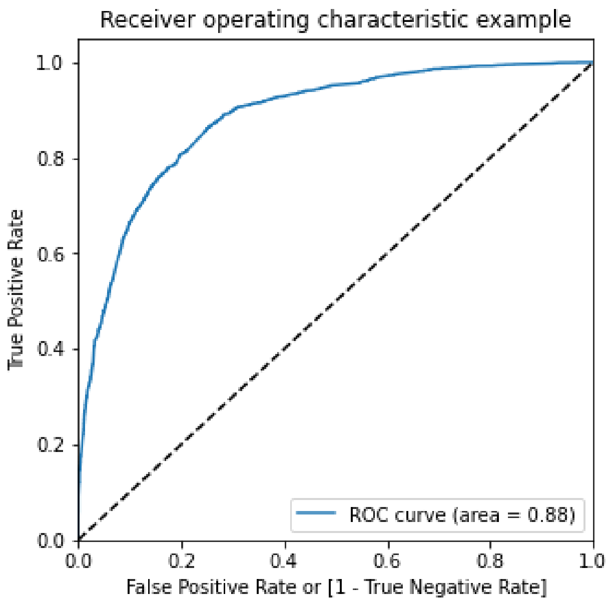
Model Building

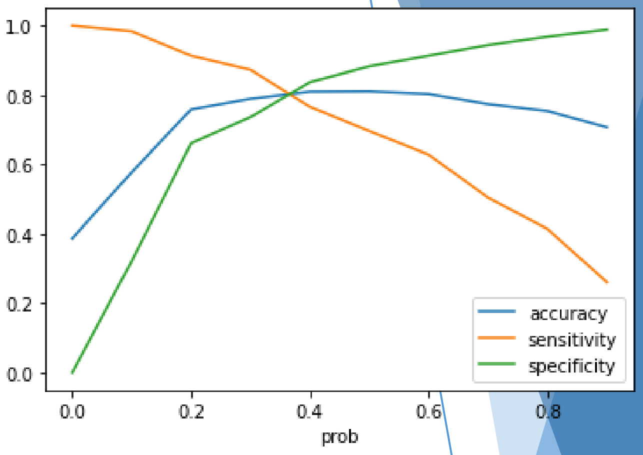
- To build the model we first created dummy variables for all the categorical variables using get_dummies method and all the NA variables were removed. For numerical variables we have used MinMaxScaler.
- Then we have split the data into Train and Test where 70% of the data was for Train and 30% of the data was for Test.
- Then we have performed RFE to attain top 15 relevant variables. Depending on the VIF and the P values i.e., VIF < 5 and P value < 0.05, we have also dropped some columns manually using drop.

Model Evaluation

- A confusion matrix was created and with an optimum cut off value using ROC curve we found the overall accuracy, sensitivity and specificity which turns around 80% on an average.
- A prediction was also performed on the Test data with cut value as 0.35 and found the overall accuracy, sensitivity and specificity which turns around 80% on an average. This method was also used to recheck with an optimum cut off value of 0.4 and the same turns around 75% on the test data frame.

ROC Curve





From the second chart it is viewed that optimal cut off is at 0.35 and the overall accuracy, specificity and sensitivity comes around to be around 80%

CONCLUSION

- Below are the list of variables which impact the most:
  1. Total number of visits
  2. When their current occupation is working professional
  3. The total time spend on the Website
  4. When the Lead Source is :
     1. Direct Traffic
     2. Google
     3. Organic Search
     4. Wellingak Website
     5. Olark Chat
     6. Reference

X Education can keep the above points to get more conversion rate for their courses.

Thank you