Lead Scoring Case Study

Problem Statement

- X Education sells online courses to industry professionals.
- 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

Business Objective:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

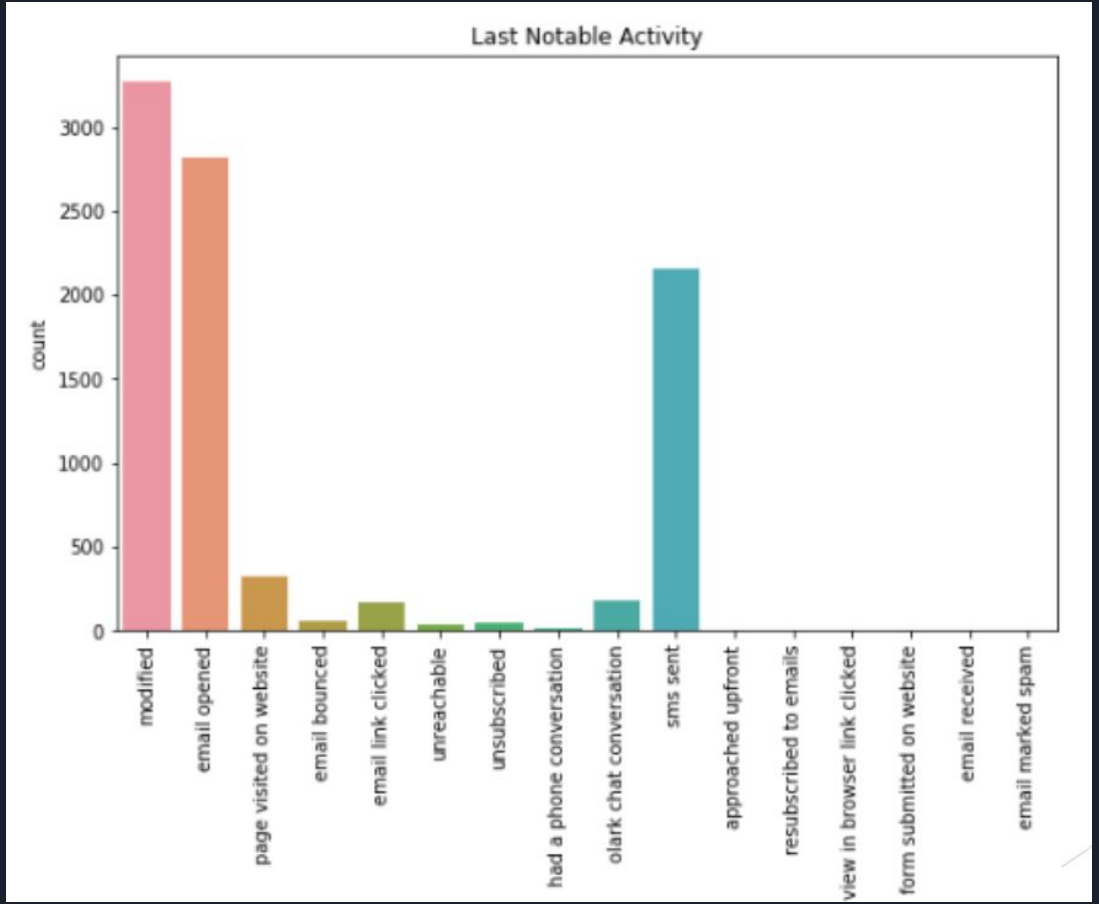
Solution Methodology

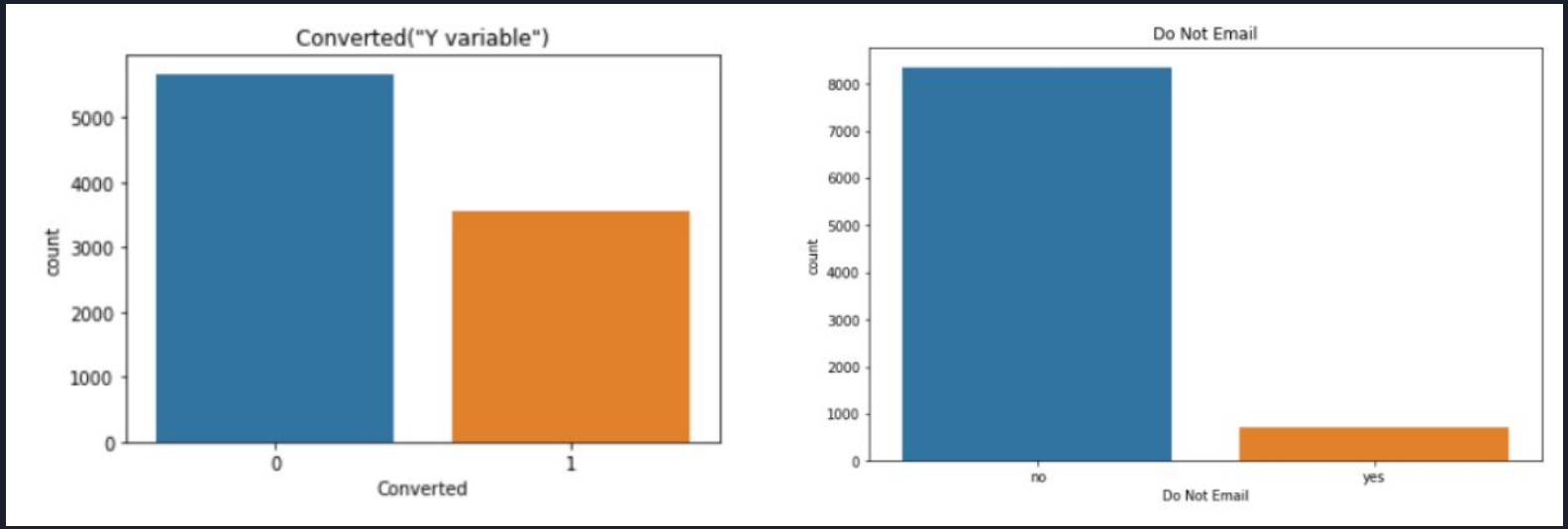
- Data cleaning and data manipulation.
  - Identify and address duplicate data.
  - Identify and manage NA and missing values.
  - Remove columns with a significant amount of missing data if they are not essential for the analysis.
  - Perform imputation for missing values, if needed.
  - Detect and address outliers in the data.
- EDA
  - Univariate data analysis: value count, distribution of variable etc.
  - Bivariate data analysis: correlation coefficients and pattern between the variables etc
- Feature scaling, creation of dummy variables, and encoding of the data.
- Classification technique: Logistic regression for model development and prediction.
- Model validation.
- Presentation of the model.
- Conclusions and recommendations.

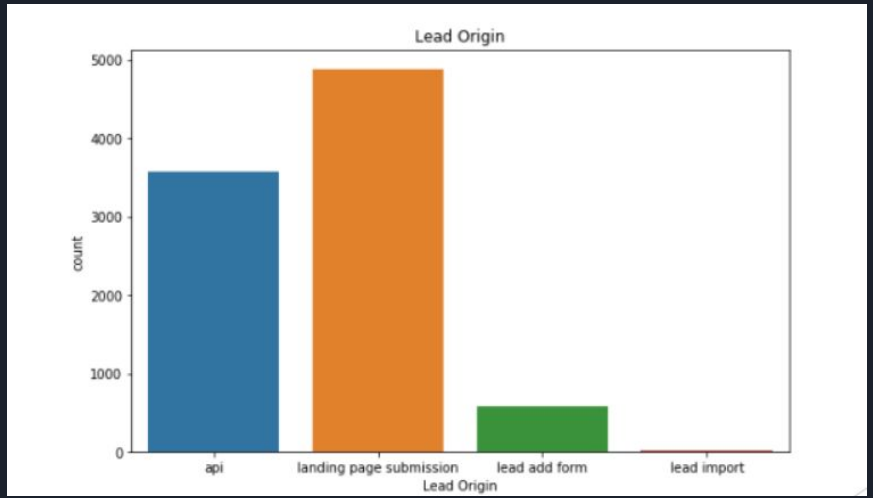
Data Manipulation

- Total Number of Rows: 37, Total Number of Columns: 9240.
- Features with single values, 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,” have been removed.
- The columns “Prospect ID” and “Lead Number,” which are not necessary for the analysis, have been discarded.
- Upon reviewing the value counts for certain object-type variables, we identified features with insufficient variance and removed them. These features include: “Do Not Call,” “What matters most to you in choosing a course,” “Search,” “Newspaper Article,” “X Education Forums,” “Newspaper,” and “Digital Advertisement.”
- Columns with more than 35% missing values, such as ‘How did you hear about X Education’ and ‘Lead Profile,’ have been dropped.

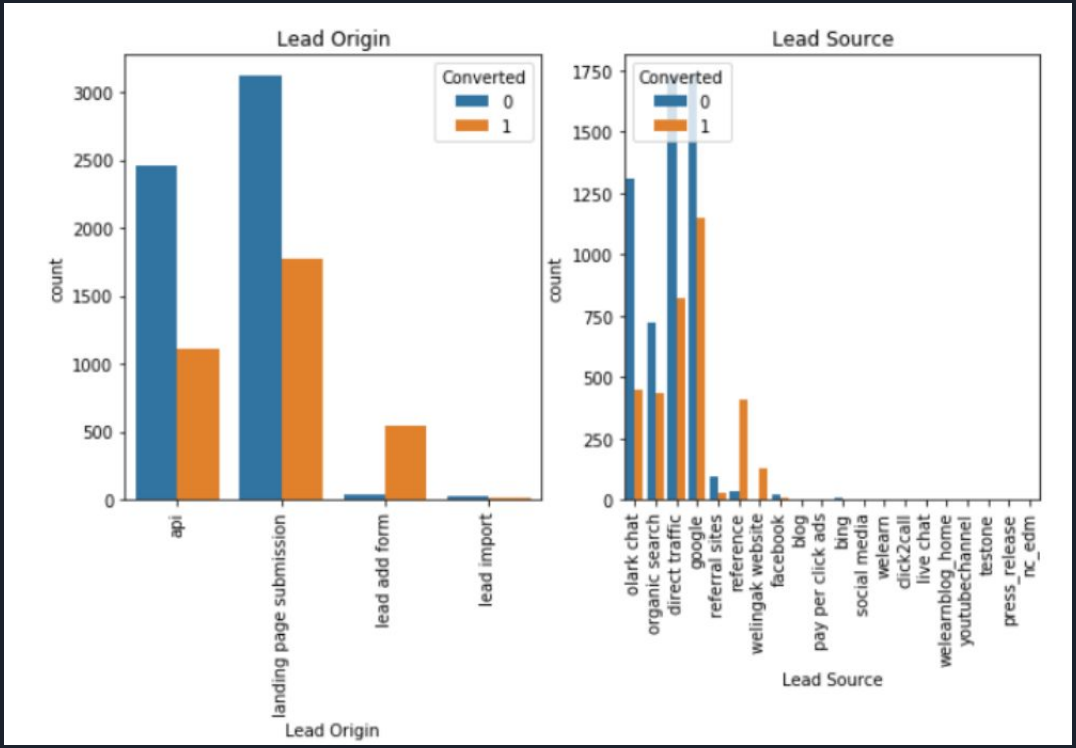
EDA

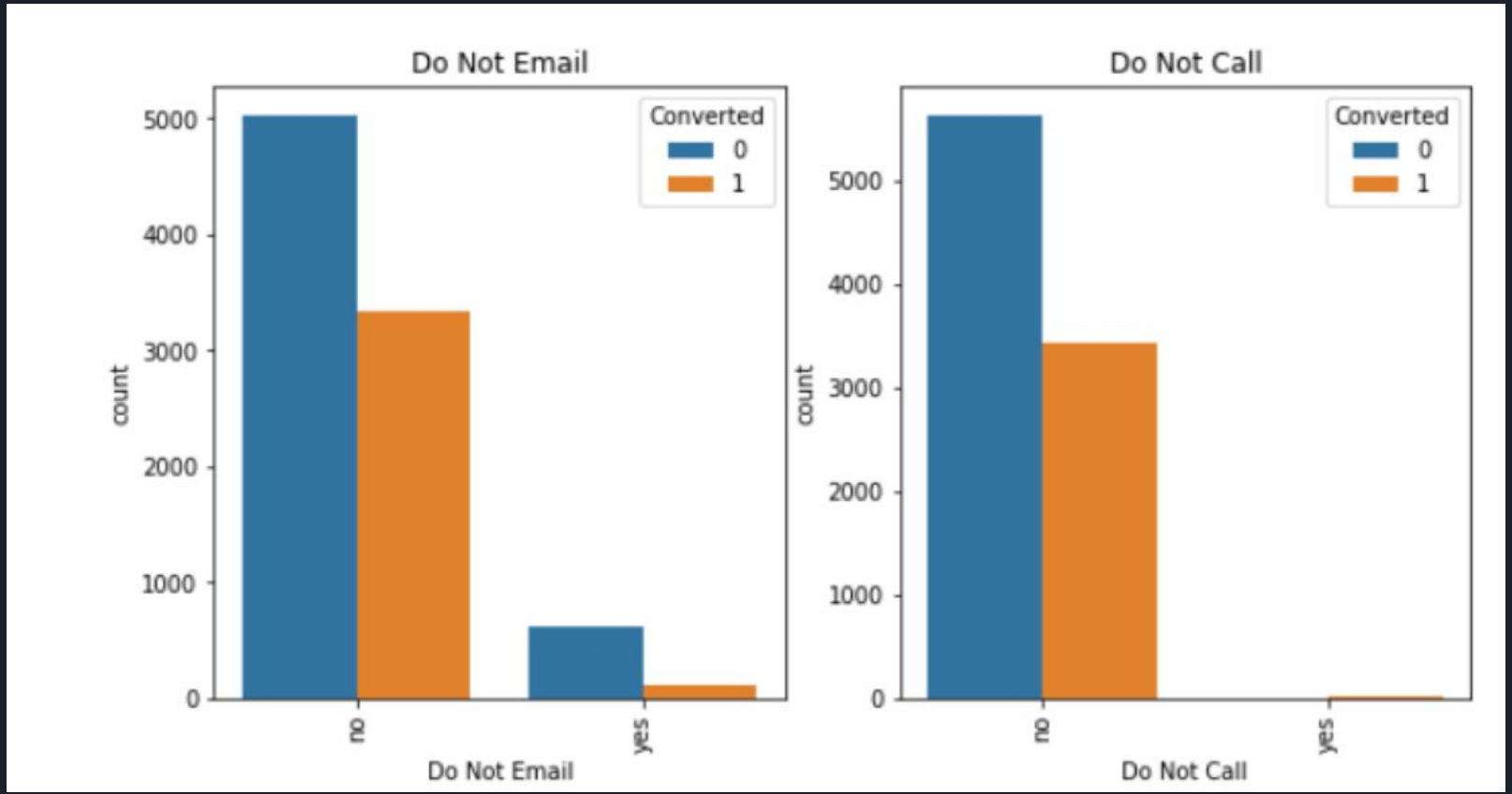

Last Notable Activity
count
3000
2500
2000
1500
1000
500
0
modified
email opened
page visited on website
email bounced
email link clicked
unreachable
unsubscribed
had a phone conversation
olark chat conversation
sms sent
approached upfront
resubscribed to emails
view in browser link clicked
form submitted on website
email received
email marked spam


Converted("Y variable")
count
5000
4000
3000
2000
1000
0
0
1
Converted
Do Not Email
count
8000
7000
6000
5000
4000
3000
2000
1000
0
no
yes
Do Not Email


Lead Origin
count
5000
4000
3000
2000
1000
0
api
landing page submission
lead add form
lead import
Lead Origin

Categorical Variable Relation




Do Not Email
Converted
0
1
5000
4000
3000
2000
1000
0
count
no
yes
Do Not Email
Do Not Call
Converted
0
1
5000
4000
3000
2000
1000
0
count
no
yes
Do Not Call

Data Conversion

- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 8792
- Total Columns for Analysis: 43

Model Building

- Splitting the Data into Training and Testing Sets
- The initial step in regression is to perform a train-test split, with a 70:30 ratio.
- Use Recursive Feature Elimination (RFE) for feature selection.
- Run RFE to select 15 variables as output.
- Build the model by removing variables with a p-value greater than 0.05 and a VIF value above 5.
- Make predictions on the test dataset.
- Achieve an overall accuracy of 81%.

Conclusion

The most significant variables influencing potential buyers, in descending order, are:

- Total time spent on the website
- Total number of visits
- Lead source, specifically:
  - a. Google
    b. Direct traffic
    c. Organic search
    d. Welingak website
- Last activity, particularly:
  - a. SMS
    b. Olark chat conversation
- Lead origin as a lead ad format
- Current occupation as a working professional

Considering these factors, X Education has a strong opportunity to convert nearly all potential buyers into customers for their courses.