Lead Score Case study

Submitted By- Deepak Bhardwaj and Ashis Kumar Dubey

Problem Statement and Goals

Problem Statement :

X Education sells online courses to industry professionals. The company has multiple Lead Generation methodology and markets its courses on several websites and search engines.

The professionals might browse the courses getting information or may fill up a form or indulge in other website activities. A query is considered to be a lead once the people fill up a form providing their genuine credentials like email address/ phone number or through referrals.

The typical lead conversion rate at X education is around 30% which ABC wants to improve to to have more qualified leads in the system for conversion that would account to be 80%

Business Goal:

CEO of X Education wants help to build a model a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

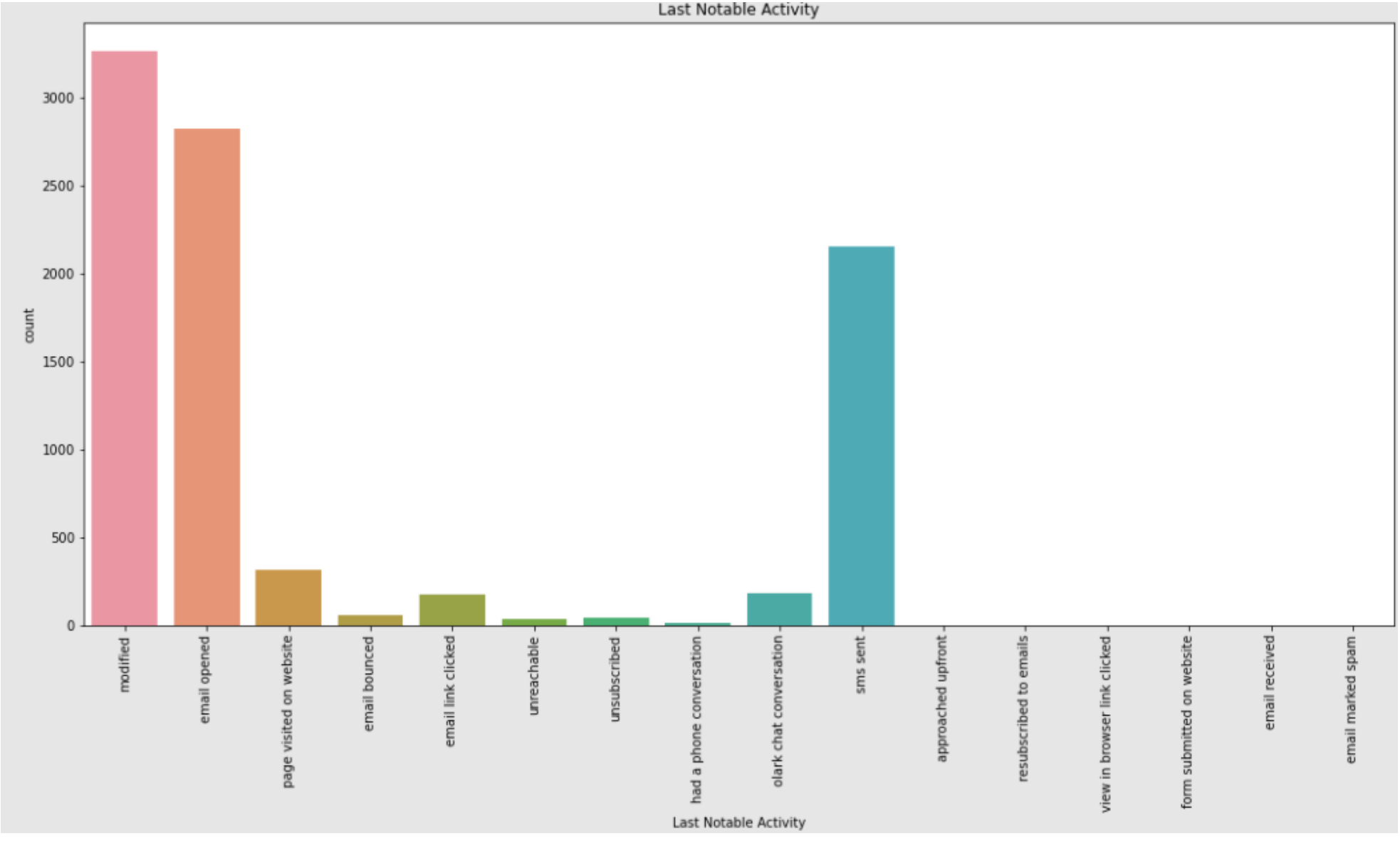
Solution Approach

- Source the data (Leads.csv)
- Data Cleaning and Data Manipulation
  - Check and handle duplicate data
  - Check and handle NA and missing values
  - Impute columns, if it contains large amount of missing values and not useful for the analysis.
  - Check and handle outliers in data.
- EDA
  - Univariate Analysis: value count, distribution of variable etc.
  - Bivariate Analysis: correlation coefficient and pattern between the variables.
- Feature Scaling & Dummy variable and encoding of the data.
- Classification method: logistic regression used for the model making and prediction.
- Validation of the model
- Conclusion and recommendation

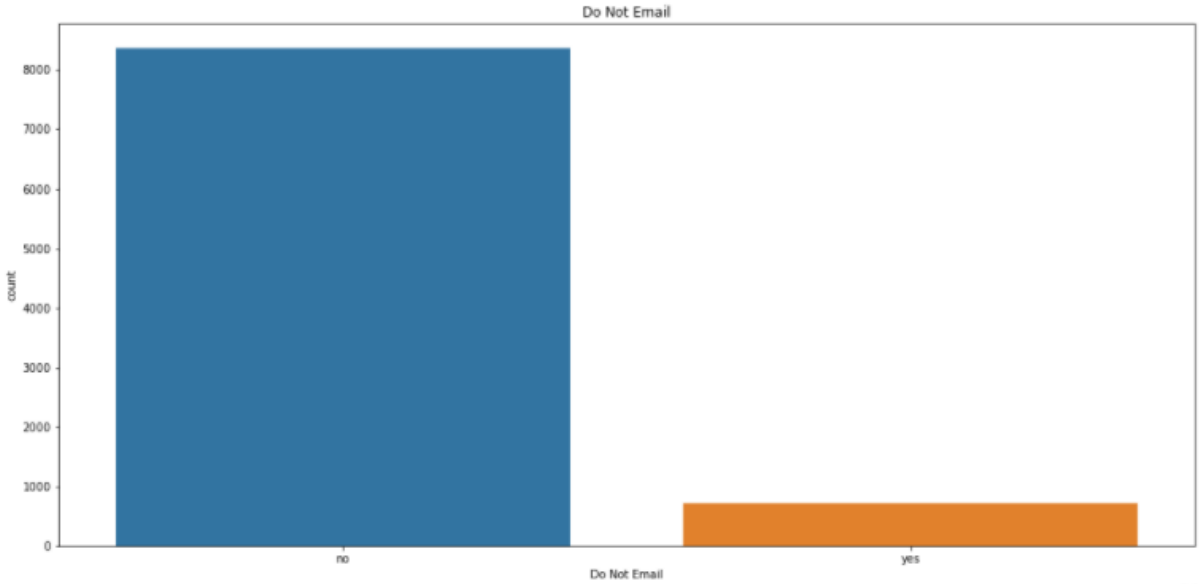
Data Manipulation

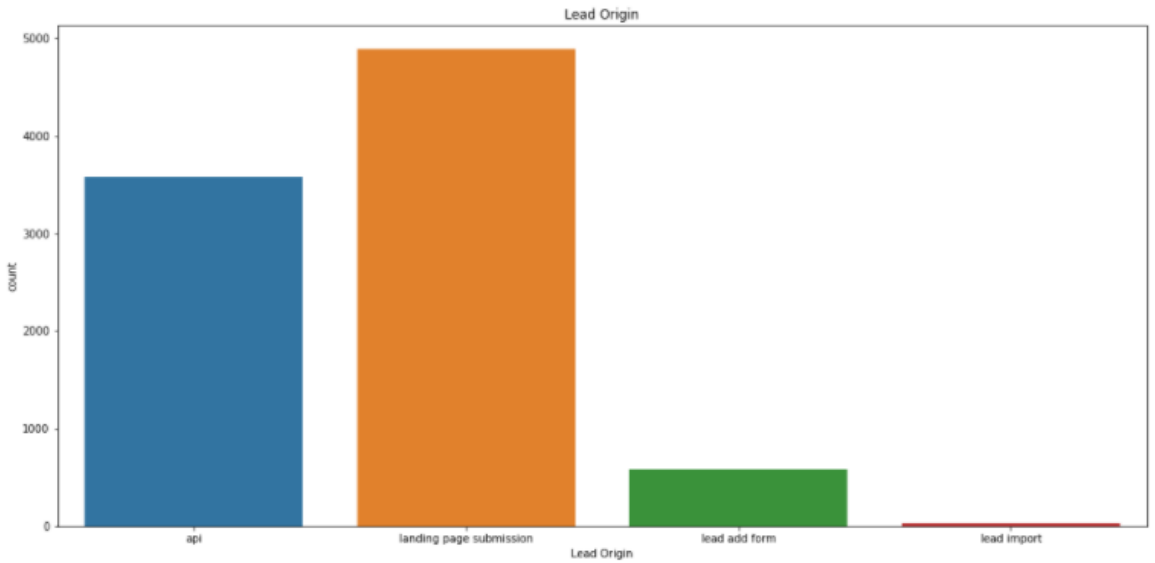
- Total number of Rows = 37 , Columns=9240
- Single value features like 'Magazine', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content',
- 'Get updates on DM Content', 'I agree to pay the amount through cheque'
- Removing 'Prospect ID' which is not necessary for analysis
- Dropping columns having more than 35% of missing values like 'Asymmetrique Profile Index',
- 'Asymmetrique Activity Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score', 'Lead Profile', 'Tags',
- 'Lead Quality', 'How did you hear about X Education', 'City', 'Lead Number'
- We found that there were few columns with high number of NaN values but seems to be important hence we kept them for further analysis and also replaced their value with 'not provided' and those columns are: 'Specialization' , 'Country' ,
- 'What matters most to you in choosing a course' , 'What is your current occupation'

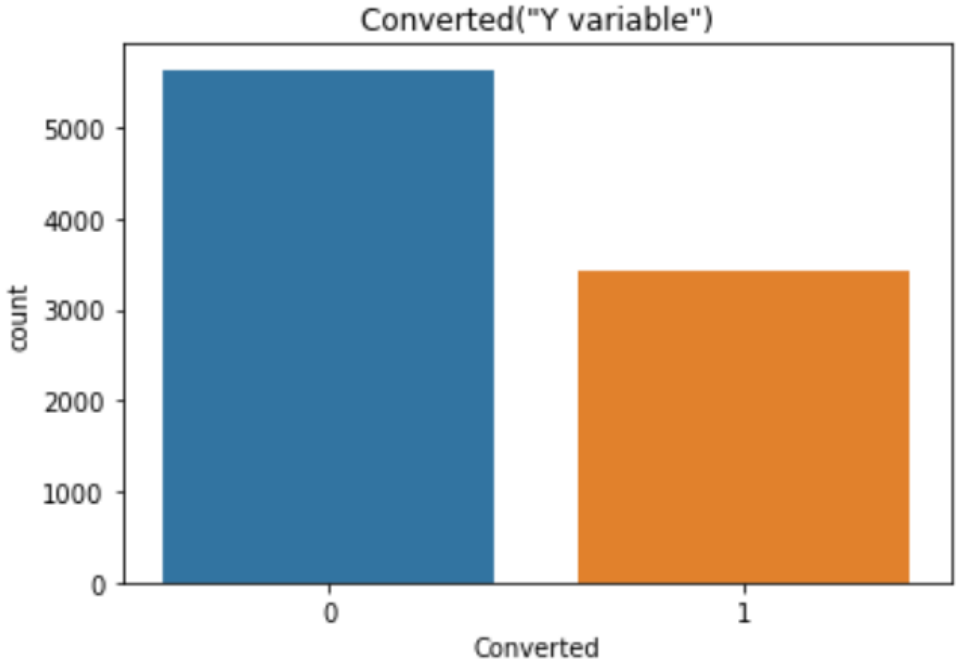
EDA


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
Last Notable Activity

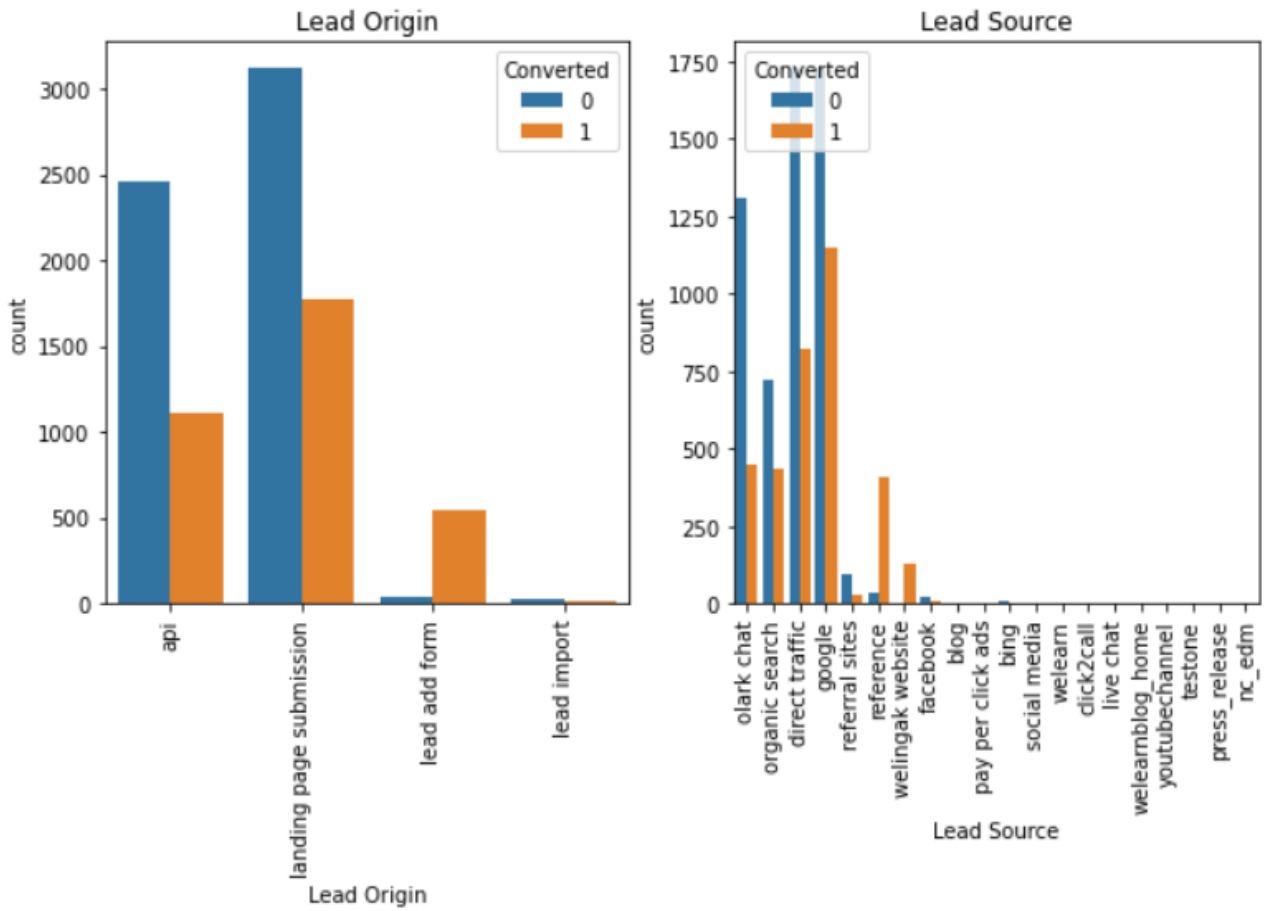
EDA


Do Not Email
count
8000
7000
6000
5000
4000
3000
2000
1000
0
no
yes
Do Not Email

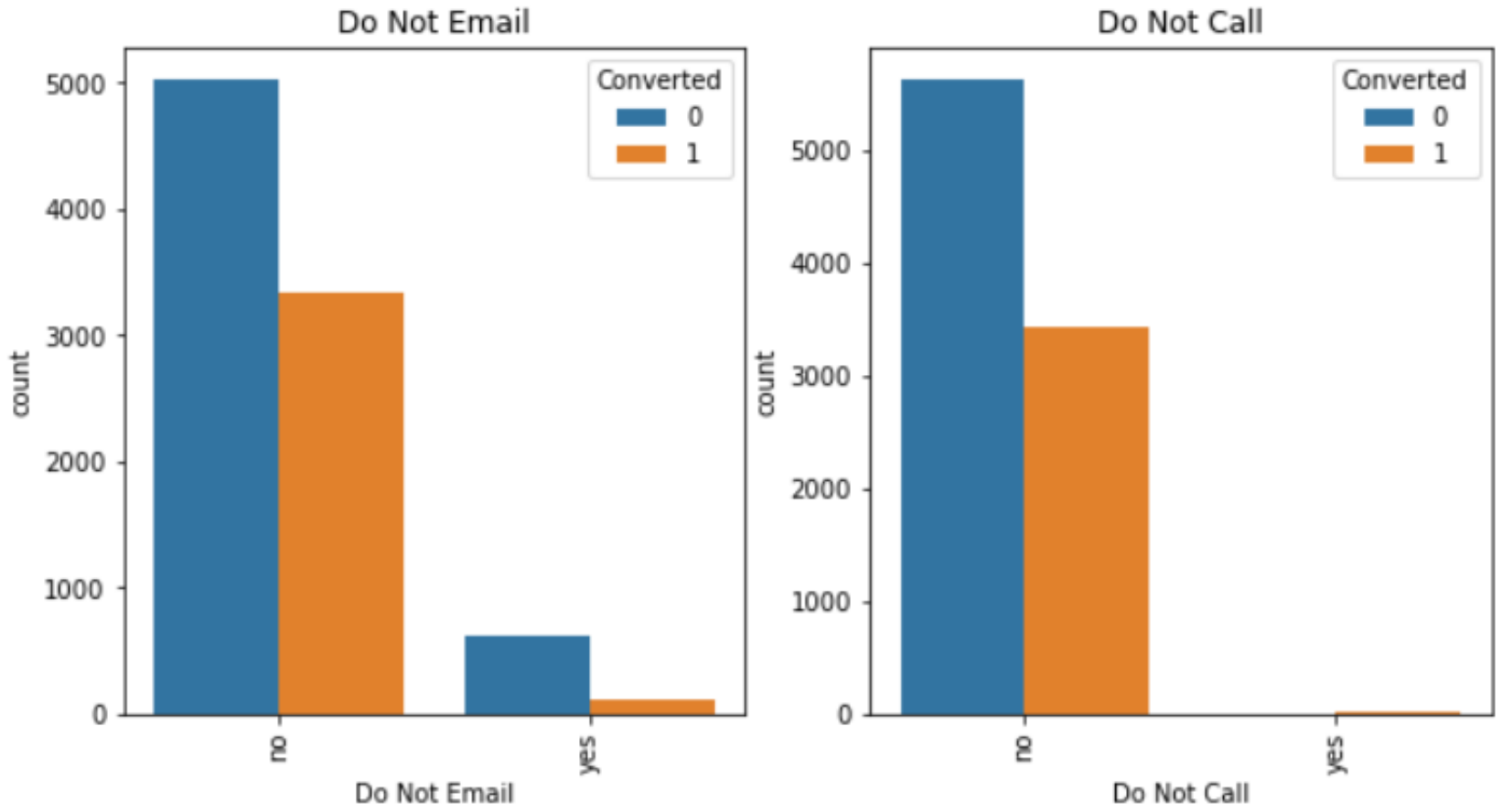

Lead Origin
count
5000
4000
3000
2000
1000
0
api
landing page submission
lead add form
lead import
Lead Origin


Converted("Y variable")
count
5000
4000
3000
2000
1000
0
0
1
Converted

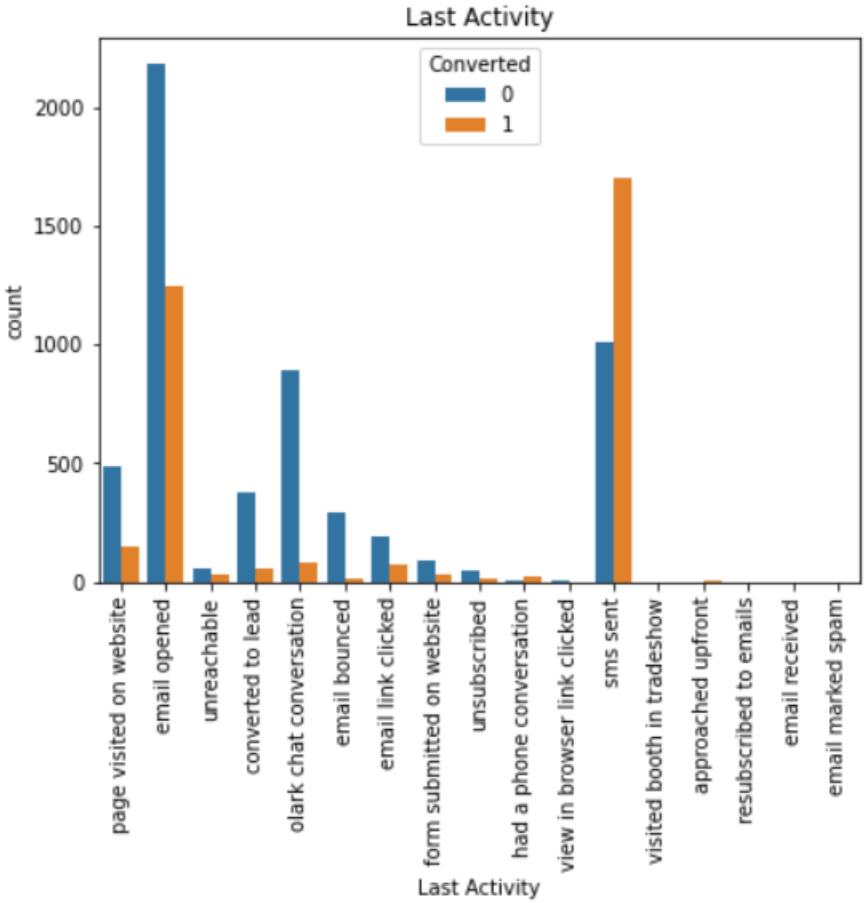
Categorical variable relation

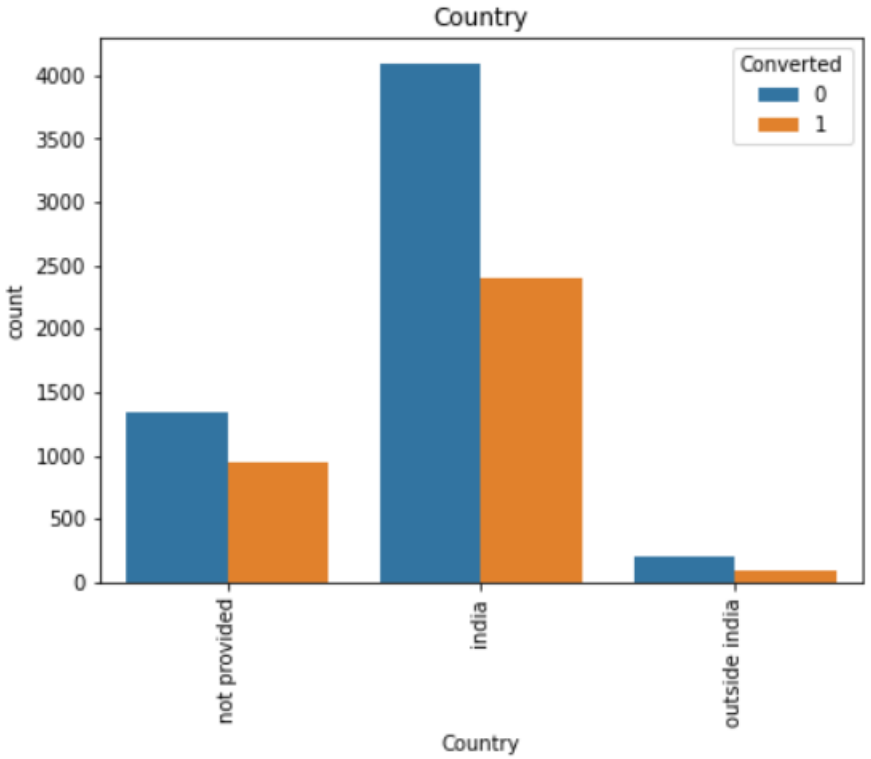


Categorical variable relation

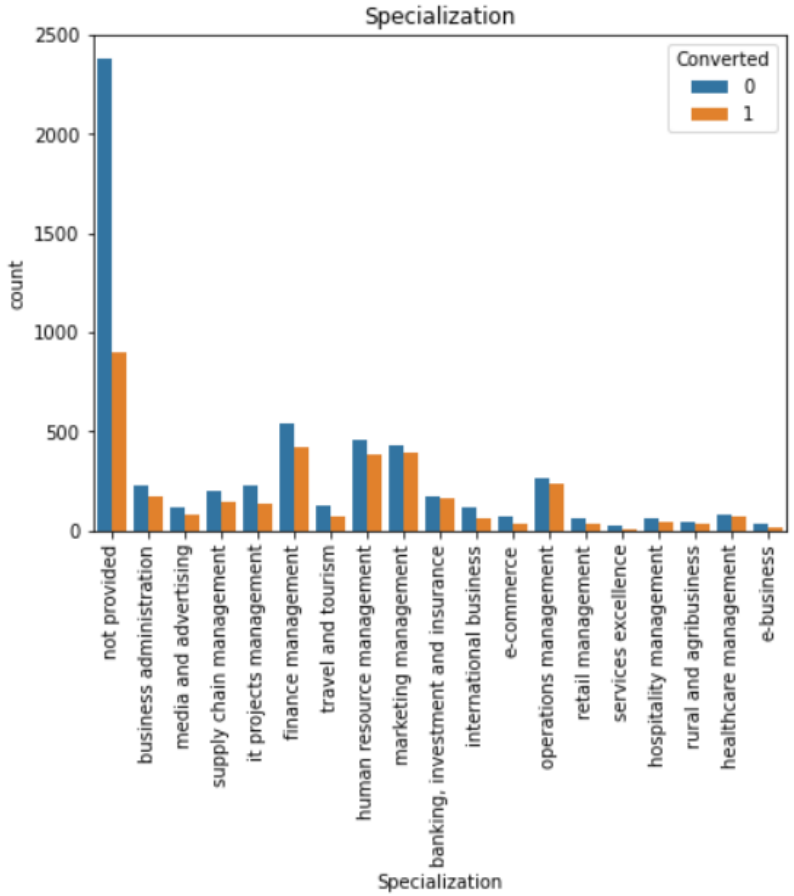


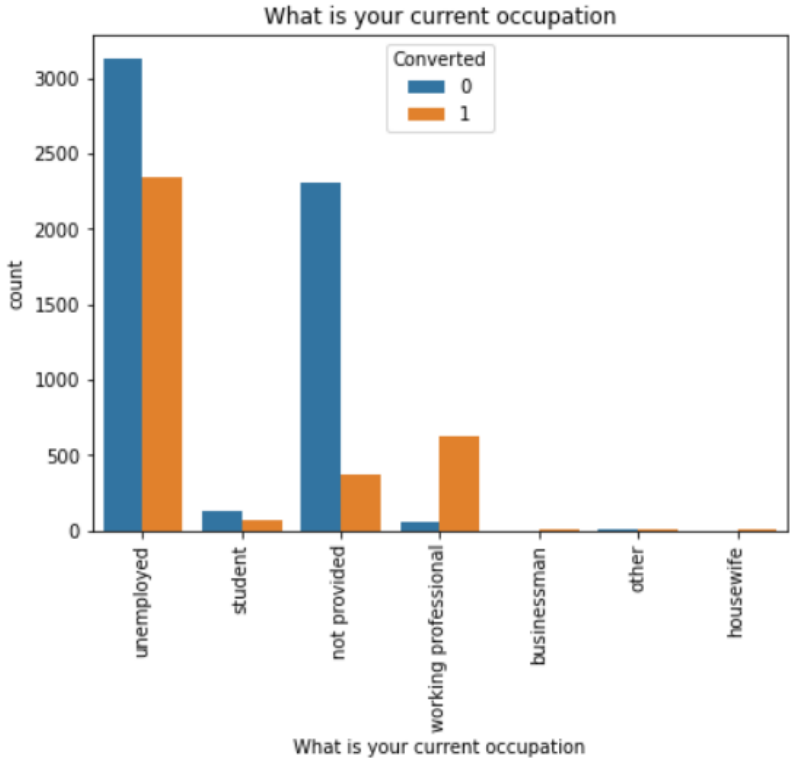
Categorical variable relation



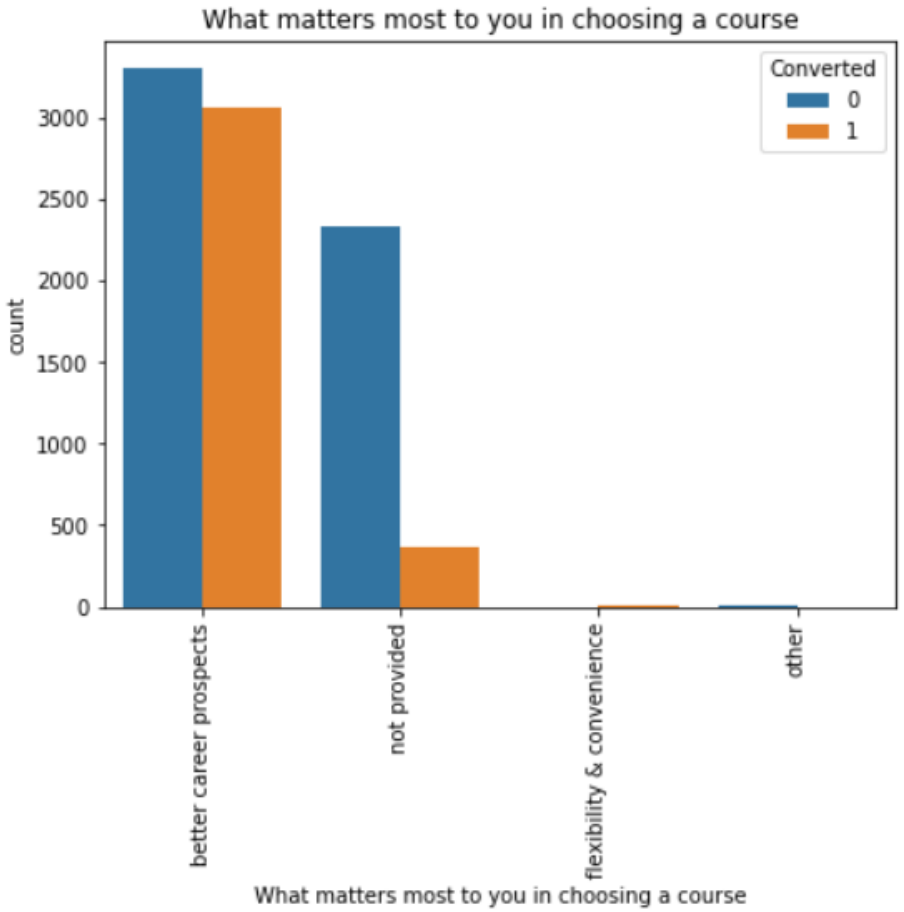


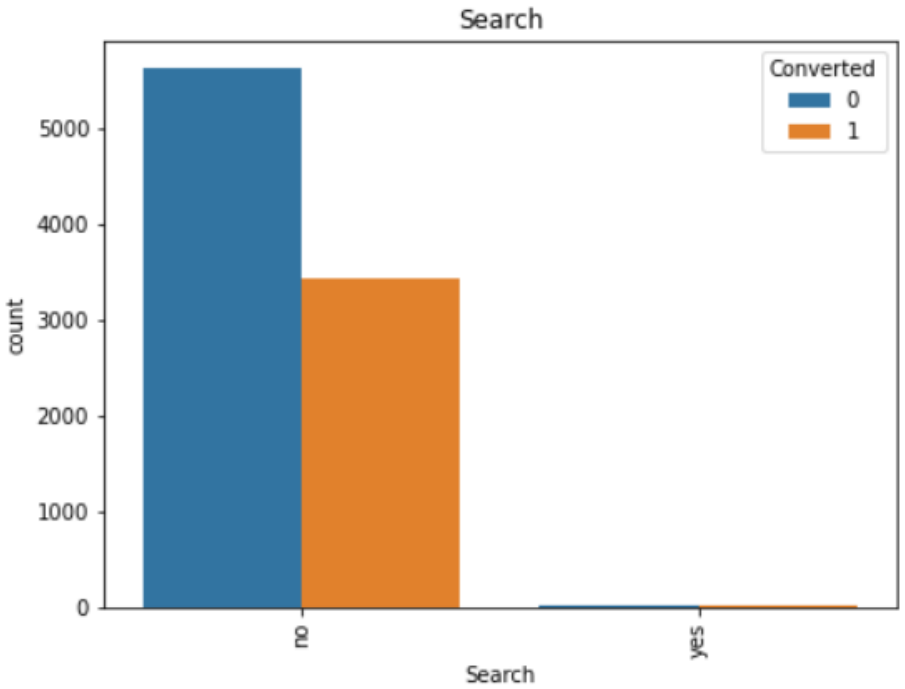
Categorical variable relation





Categorical variable relation





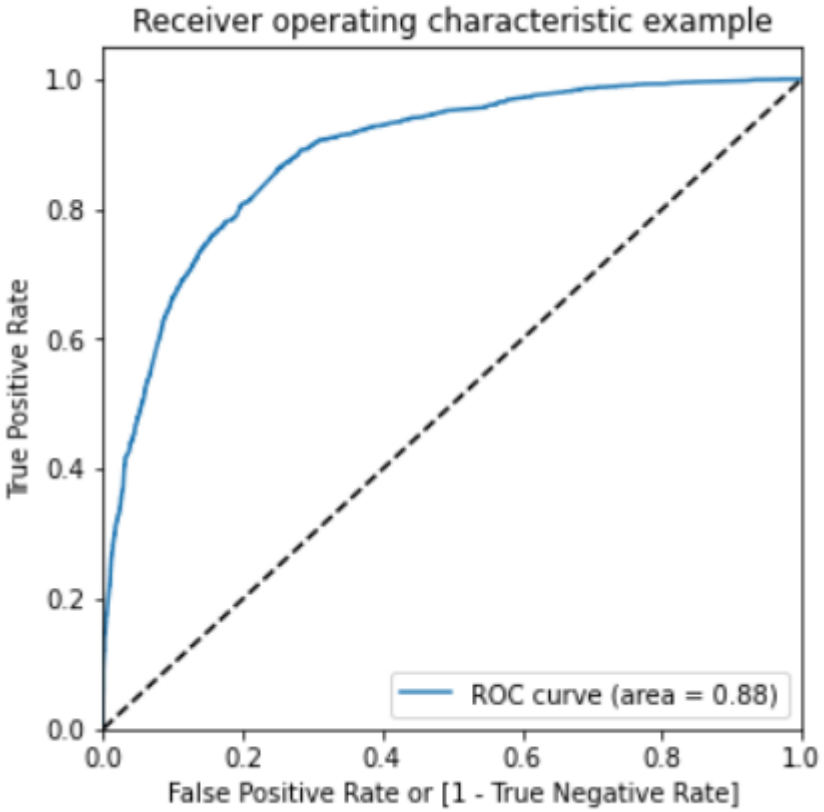
Data Standardizations

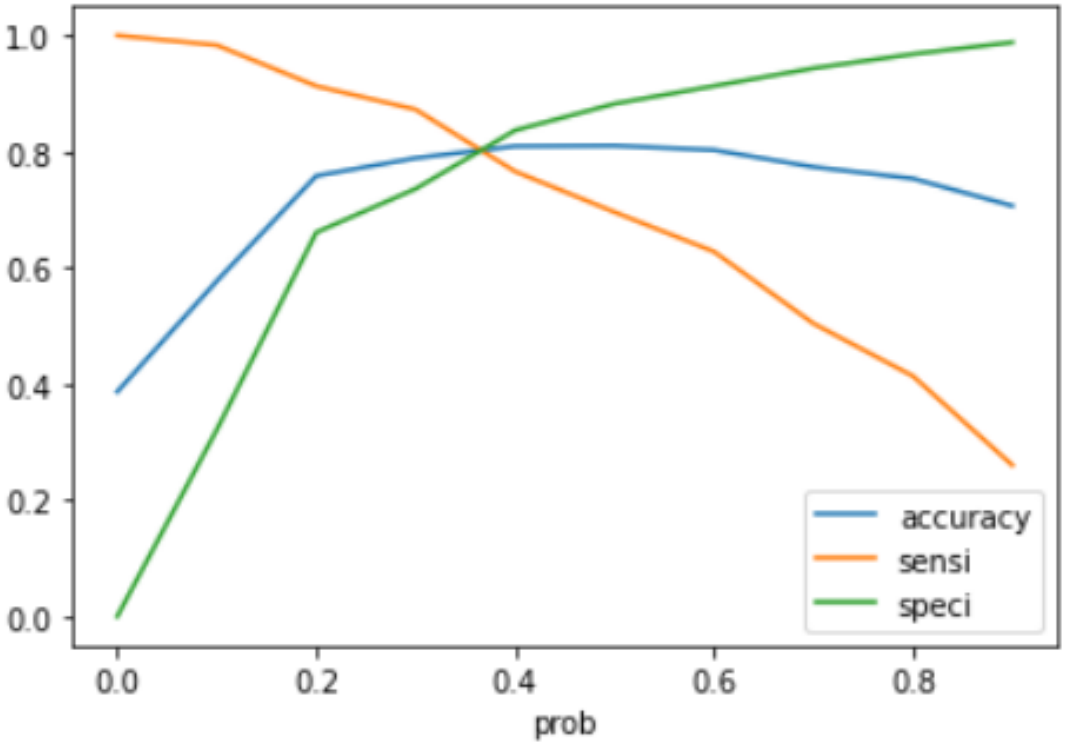
- ✓ Numerical variables like 'TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website' are normalised using "MinMaxScaler".
- ✓ Allocating Dummy variables for object data type variables.

Model Building

- Splitting data into Training and Test Set.
- Performing Train-Test split with a ratio of 70:30.
- Using RFE for automatic feature selection.
- Running RFE with 15 variables as output.
- Building model by removing the variable whose p-value is greater than 0.05 and VIF less than 5.
- Making predictions on test data.
- Overall accuracy achieved is 81%.

ROC Curve


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.88)
0.0
0.2
0.4
0.6
0.8
1.0


accuracy
sensi
speci
prob
0.0
0.2
0.4
0.6
0.8
1.0

Conclusion

Below mentioned variables are the ones that mattered most in potential buyer (Desc. Order).

- TotalVisits
- Total Time Spent on Website
- Lead Origin_lead add form
- Lead Source_olark chat
- Lead Source_welingak website
- Do Not Email_yes
- Last Activity_olark chat conversation
- Last Activity_sms sent
- What is your current occupation_other
- What is your current occupation_student
- What is your current occupation_working professional
- Last Notable Activity_unreachable