Lead Score Case Study

Group Members :

1. Sree Badri
2. Sasank Yenduri
3. Vyanktesh Chaturvedi

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 Steps :

Data cleaning and data manipulation.

1. Check and handle duplicate data.
2. Check and handle NA values and missing values.
3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
4. Imputation of the values, if necessary.
5. Check and handle outliers in data.

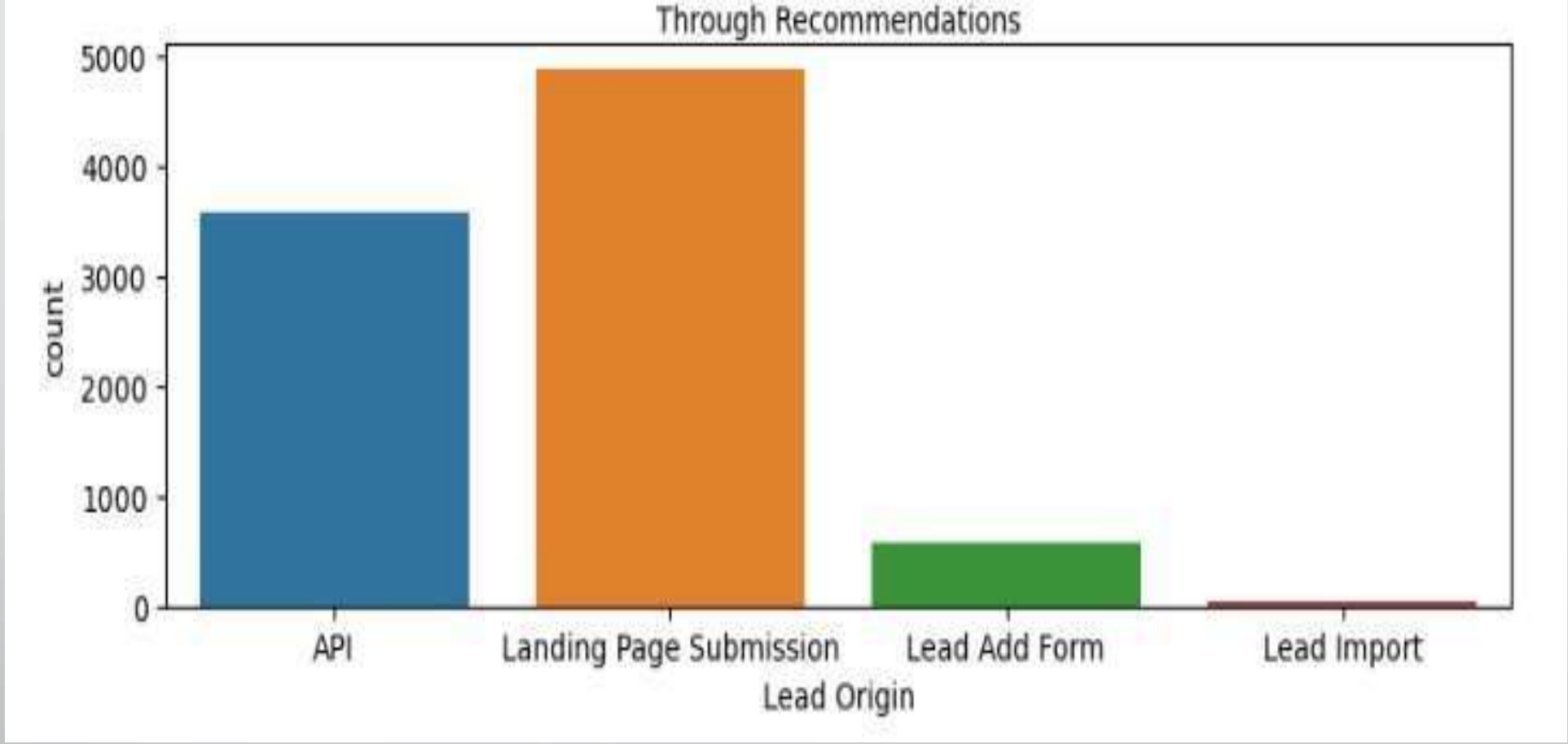
- EDA
  1. Univariate data analysis: value count, distribution of variable etc.
  2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

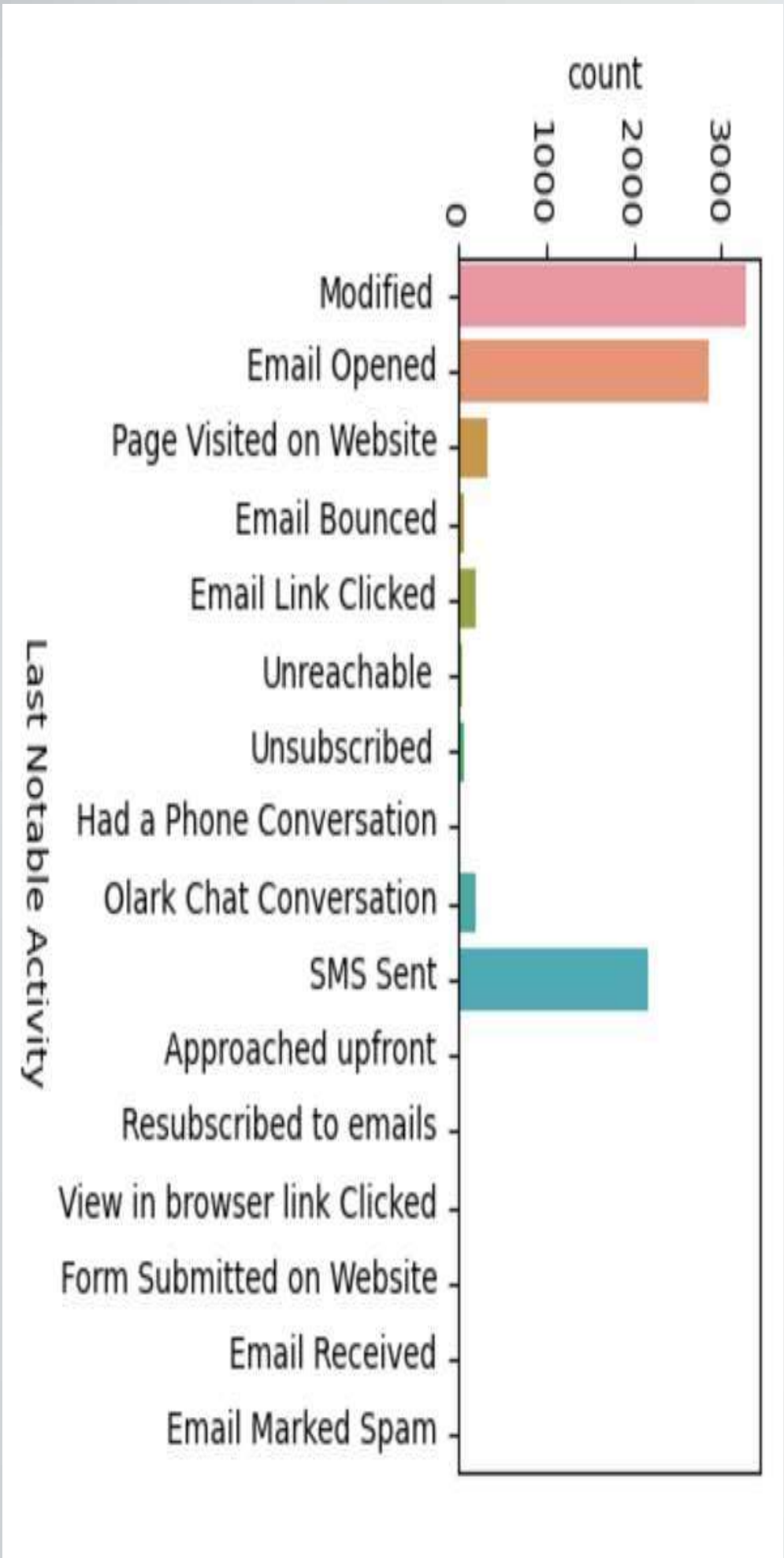
Data Cleaning :

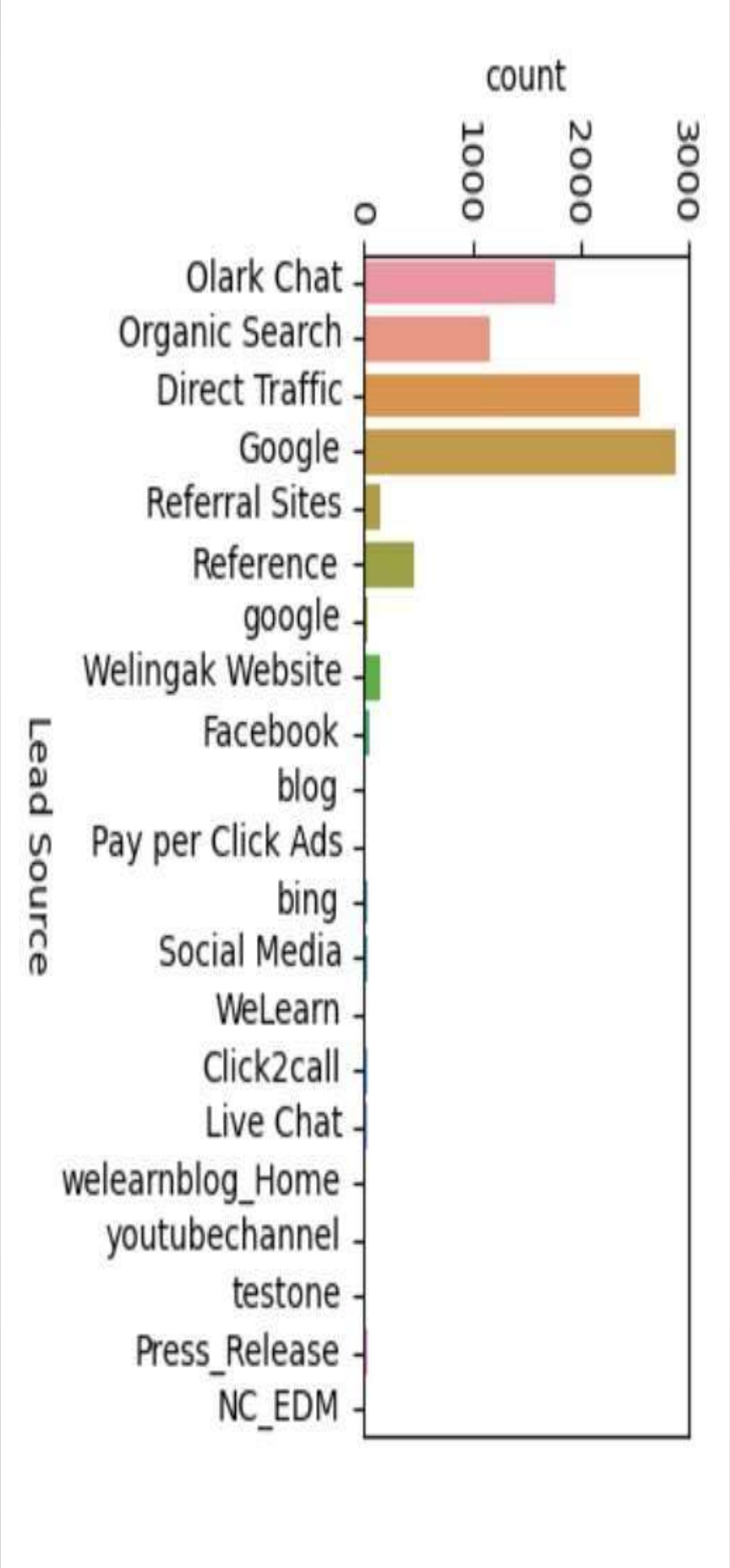
Total Number of Rows =9240, Total Number of Columns =37.

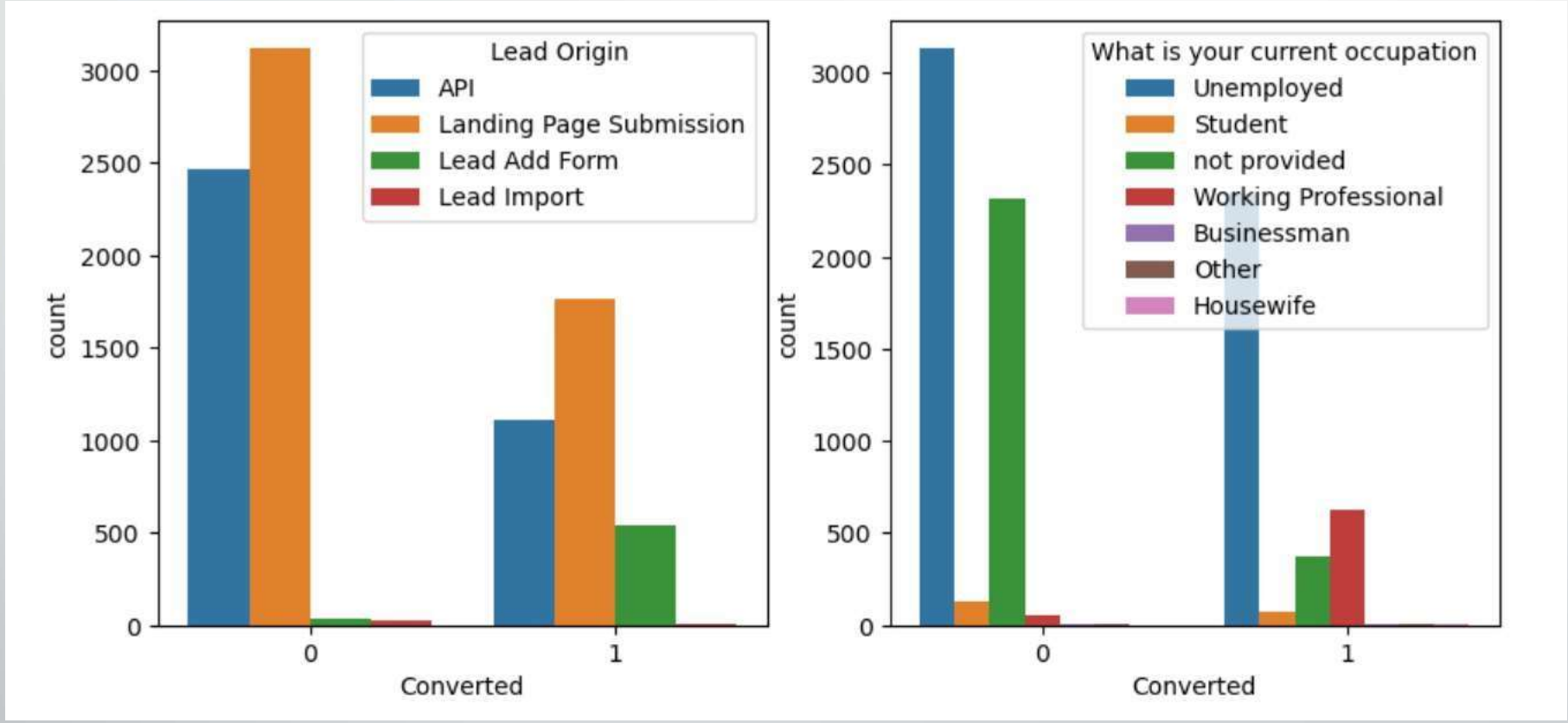
- Removing columns which are not required in analysis like Prospect ID','Magazine','Receive More Updates About Our Courses','I agree to pay the amount through cheque','Get updates on DM Content','Update me on Supply Chain Content'.
- 'Asymmetrique Profile Index','Asymmetrique Activity Index','Asymmetrique Activity Score','Asymmetrique Profile Score','Lead Profile','Tags','Lead Quality','How did you hear about X Education','City','Lead Number' also removed.
- Null values replaced with "Not provided" in these columns ('Specialization', 'What matters most to you in choosing a course', 'Country', 'What is your current occupation').

EDA


Through Recommendations
5000
4000
3000
2000
1000
0
count
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin


count
0
1000
2000
3000
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
Resubscribed to emails
View in browser link Clicked
Form Submitted on Website
Email Received
Email Marked Spam


count
0
1000
2000
3000
Lead Source
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
google
Welingak Website
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM


Lead Origin
API
Landing Page Submission
Lead Add Form
Lead Import
count
3000
2500
2000
1500
1000
500
0
0
1
Converted
What is your current occupation
Unemployed
Student
not provided
Working Professional
Businessman
Other
Housewife
count
3000
2500
2000
1500
1000
500
0
0
1
Converted

Data Conversion :

- NumericalVariables are Normalised
- DummyVariables are created for object type variables
- Total Rows for Analysis: 9074
- Total Columns for Analysis including dummies: 87

Model Building :

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 20 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5
- After building final model we have done model evaluation on test data where we have calculated metrics such as accuracy, sensitivity and specificity.

#Accuracy : 81.5 %
#Sensitivity : 68.5 %
#Specificity : 88.8 %

Conclusion

- The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.
- The company should make calls to the leads who are the "House Wifes" as they are more likely to get converted.
- The company should make calls to the leads whose last activity was SMS Sent as they are more likely to get converted.
- The company should make calls to the leads who spent "more time on the websites" as these are more likely to get converted.
- The company should not make calls to the leads whose last activity was "Olark Chat Conversation" as they are not likely to get converted.
- The company should not make calls to the leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.