Lead Score Case Study

Atanu Bhowmick
Ashish Ahuja
Abheer Brahme

Contents

- Problem statement
- Problem approach
- EDA
- Model building
- Model evaluation
- Observation
- Conclusion

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. The company looking to build a model where we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a 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%.

Problem approach

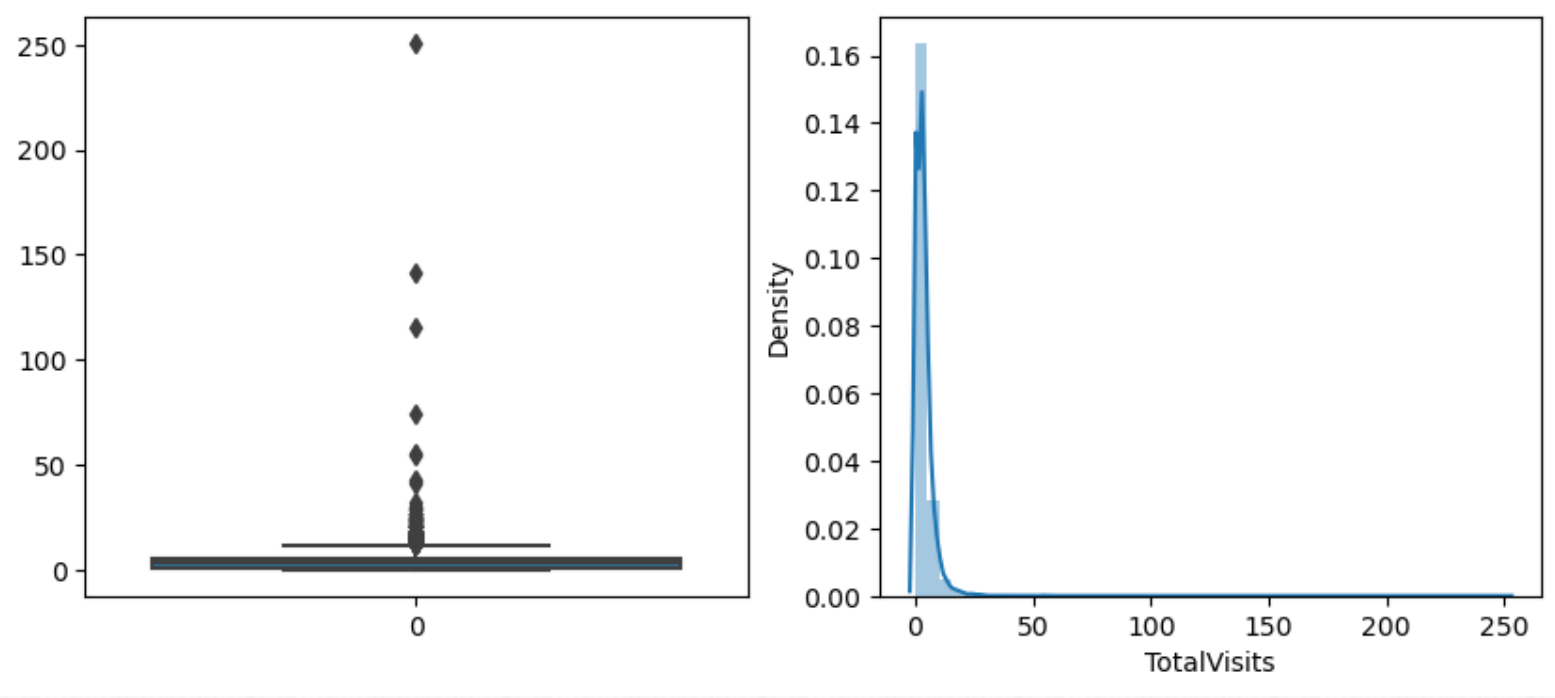
Below steps are involved to approach the solution.

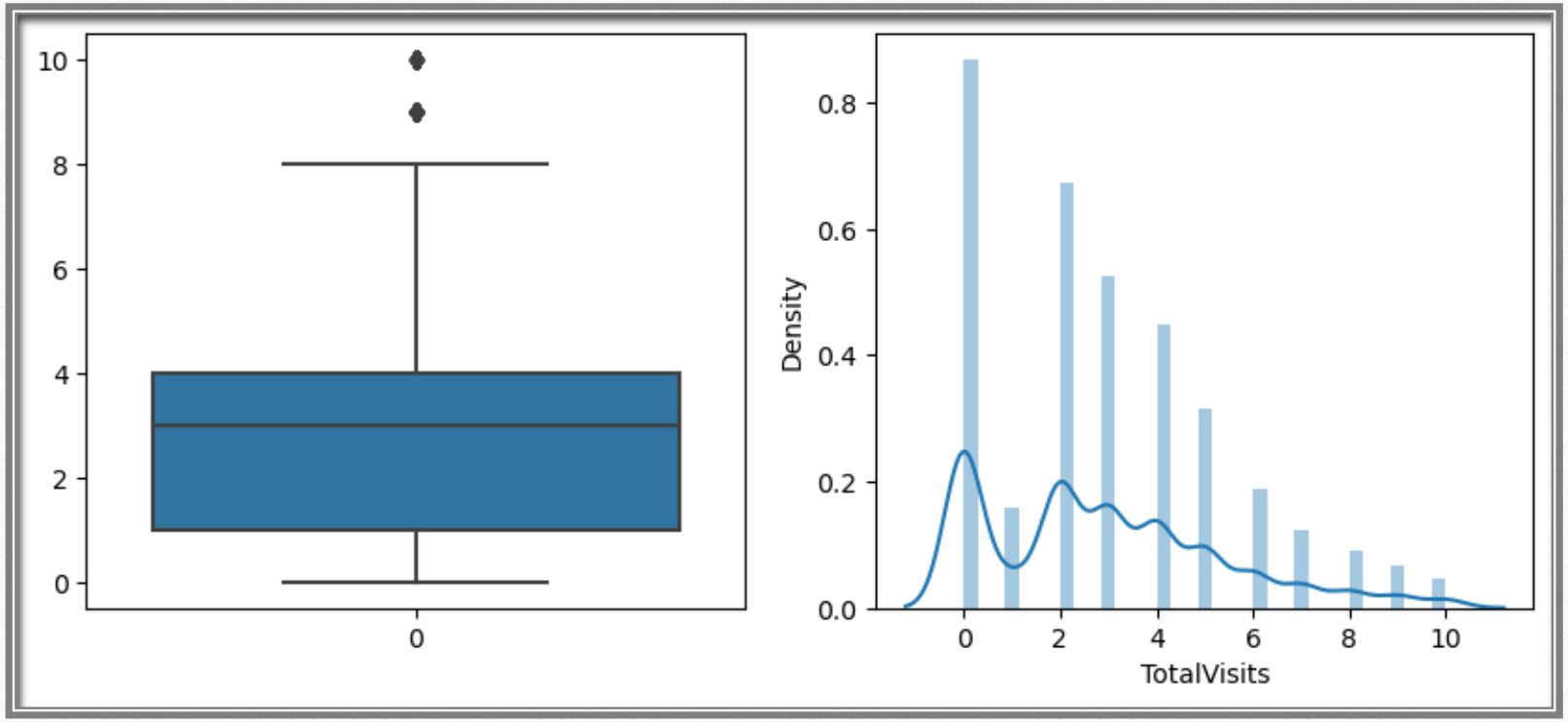
- EDA
- Model building
- Model evaluation
- Observation
- Conclusion

Reading data & Cleanup

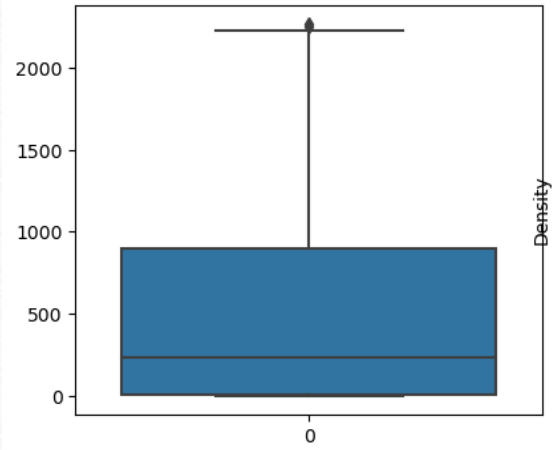
- There ate total 37 rows and 9240 columns
- Columns which contain unique value (Prospect ID, Lead Number) are dropped
- Replaced ‘Select’ value with Null
- Columns for which missing data percentage is more than 33% are dropped
- The columns which represents the ad channels are dropped. Columns are (Search, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations)
- Country column is dropped as most of the user’s country is India or missing

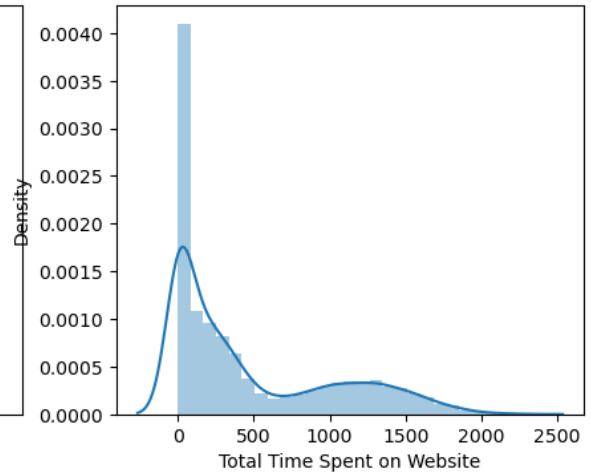
EDA

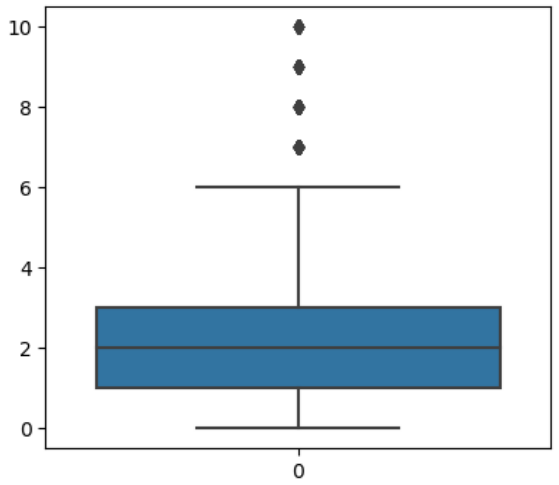

250
200
150
100
50
0
0
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Density
0
50
100
150
200
250
TotalVisits

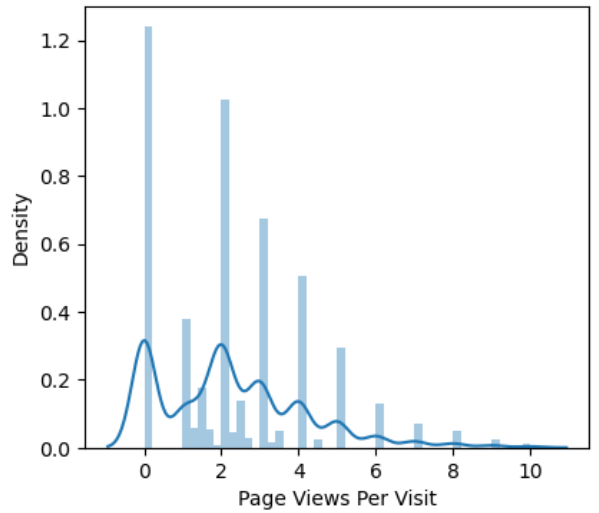


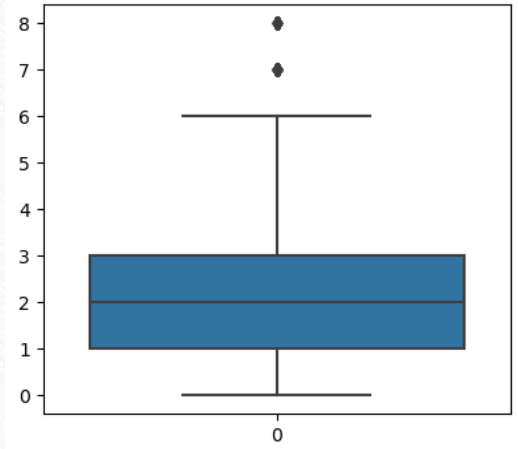
After removing outliers

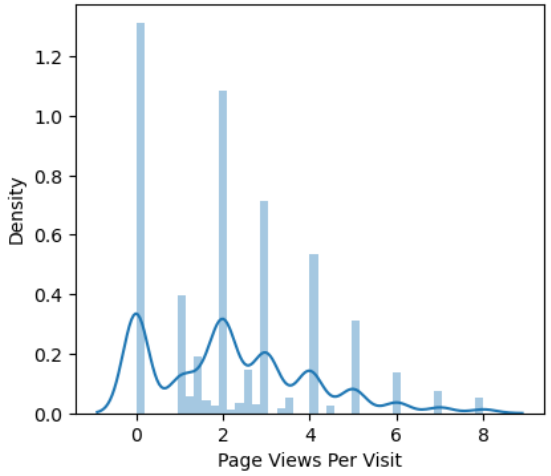

2000
1500
1000
500
0
0

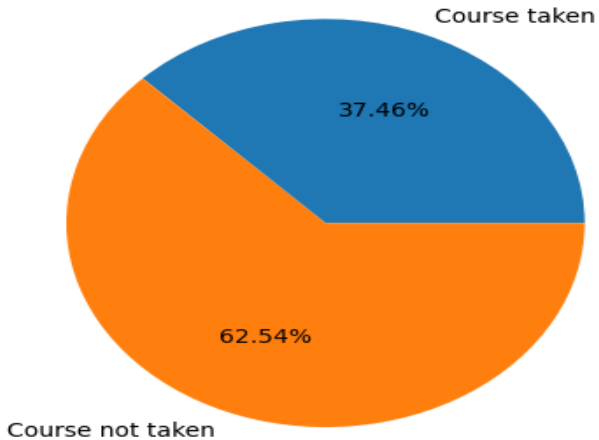

0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
Density
0
500
1000
1500
2000
2500
Total Time Spent on Website

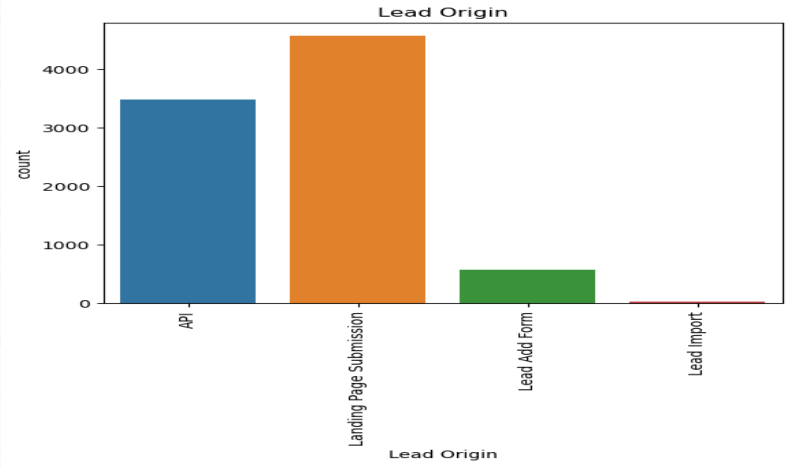

10
8
6
4
2
0
0

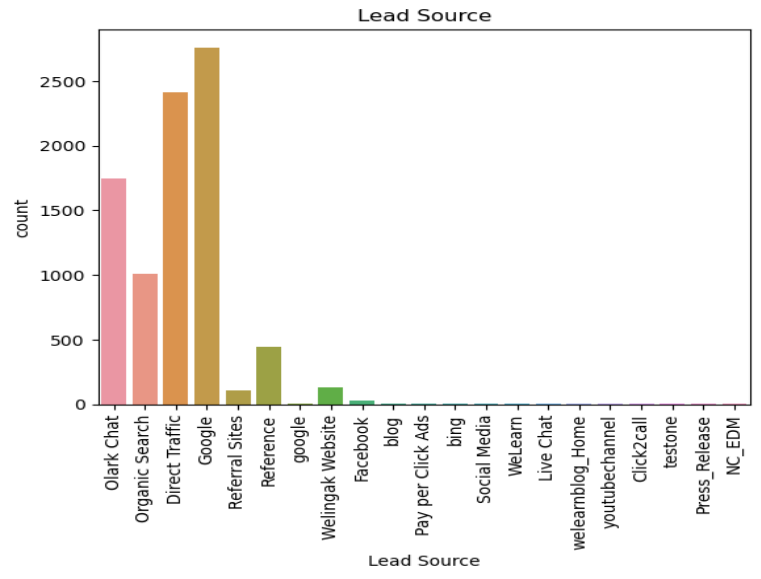

1.2
1.0
0.8
0.6
0.4
0.2
0.0
Density
0
2
4
6
8
10
Page Views Per Visit

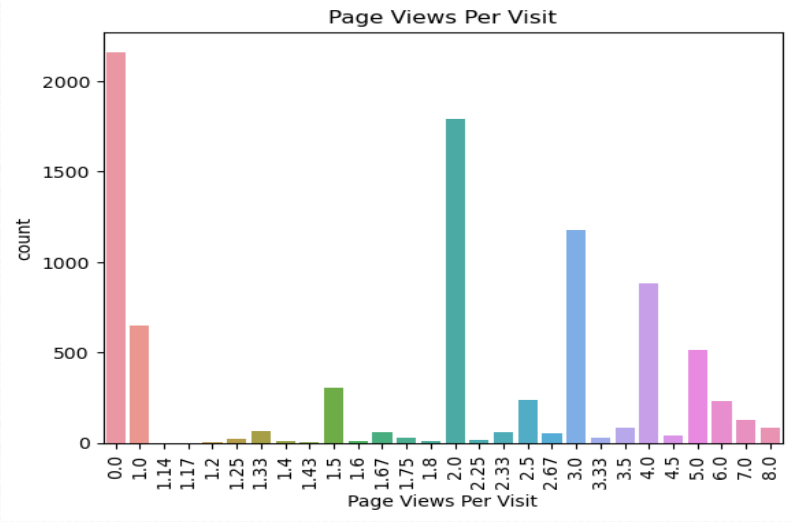

8
7
6
5
4
3
2
1
0
0

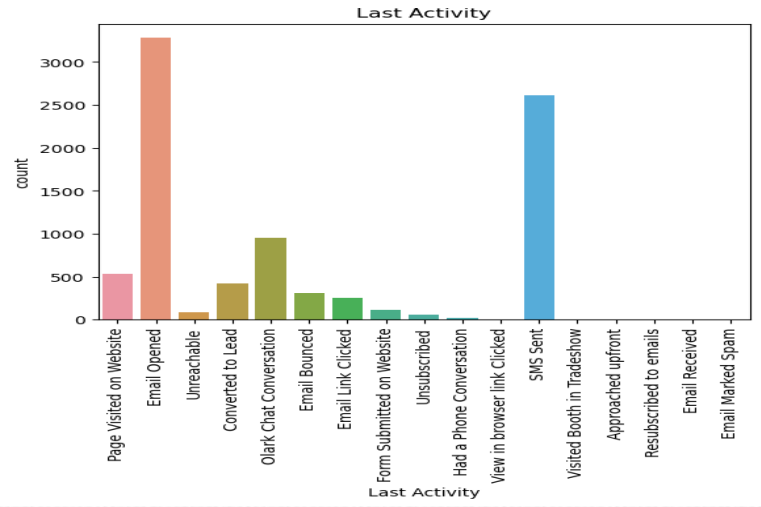

1.2
1.0
0.8
0.6
0.4
0.2
0.0
Density
0
2
4
6
8
Page Views Per Visit


Course taken
37.46%
62.54%
Course not taken

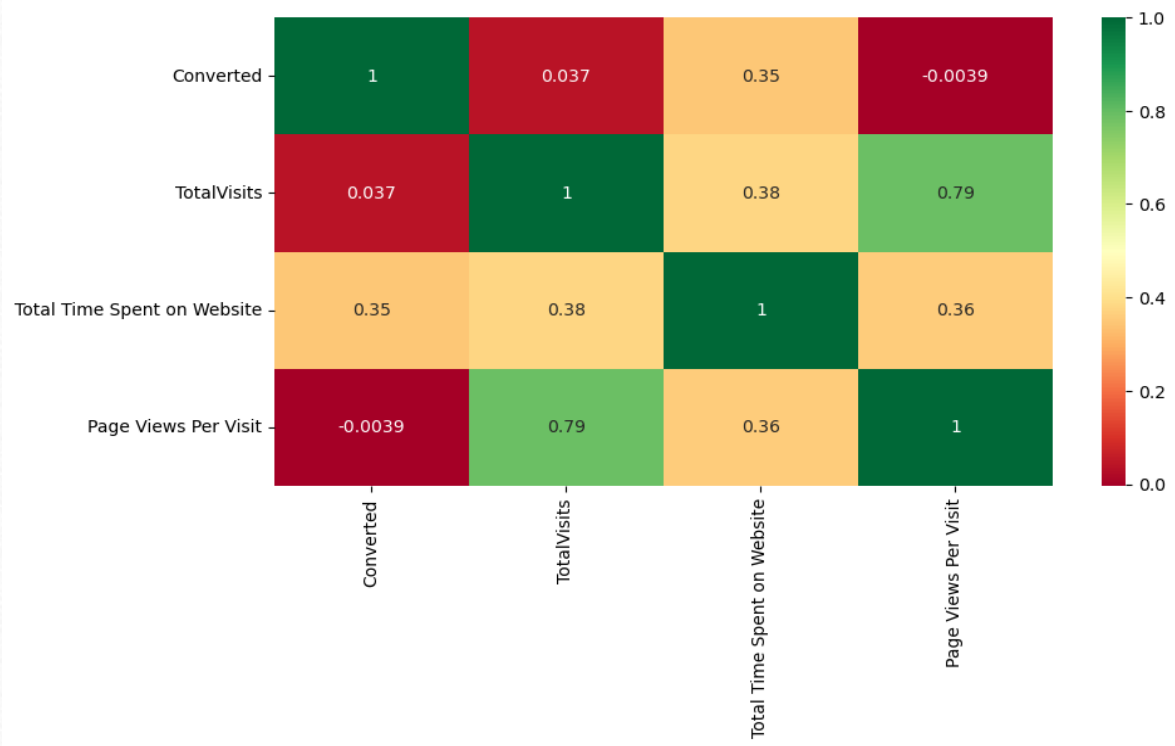

Lead Origin
count
4000
3000
2000
1000
0
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin


Lead Source
count
2500
2000
1500
1000
500
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
google
Welingak Website
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Live Chat
welearnblog_Home
youtubechannel
Click2call
testone
Press_Release
NC_EDM
Lead Source


Page Views Per Visit
count
2000
1500
1000
500
0
0.0
1.0
1.14
1.17
1.2
1.25
1.33
1.4
1.43
1.5
1.6
1.67
1.75
1.8
2.0
2.25
2.33
2.5
2.67
3.0
3.33
3.5
4.0
4.5
5.0
6.0
7.0
8.0
Page Views Per Visit


Last Activity
count
3000
2500
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
SMS Sent
Visited Booth in Tradeshow
Approached upfront
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity

Correlation



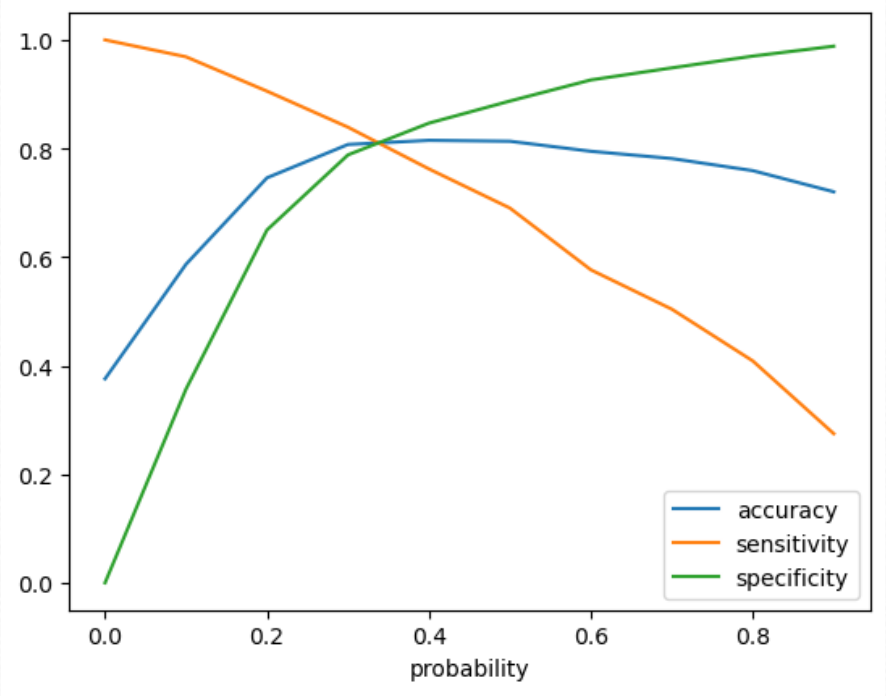
Data Conversion

- Numerical variables are normalized
- Dummy variables are created for object type variables
- Total rows for analysis: 8653
- Total columns for analysis: 11

Model Building

- Split the data into training and test set
- Train-Test split with 70:30 ratio
- The RFE for feature selection is used in Logistic Regression model
- Choosing 10 features out of all the features
- Model selection by dropping features which has p-value greater than 0.00 and VIF value greater than 2.5
- Predicting the test data
- Achieved the model accuracy is 80.20
- Model predicted with a optimum cutoff of 0.36

ROC Curve


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
probability
accuracy
sensitivity
specificity

- From the model we can see that the variables that mattered the most in the conversion are
  - Last notable activity
  - Lead source
  - Total time spent
  - Current Occupation

Thank You