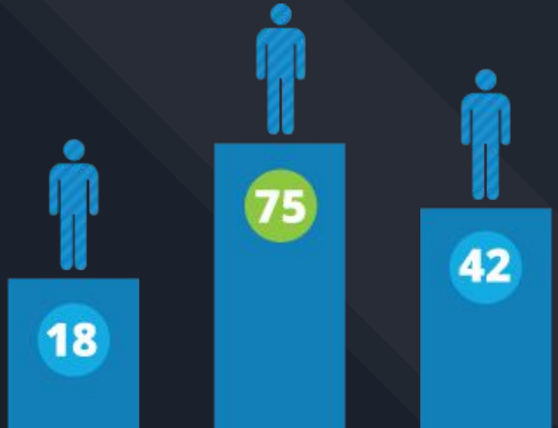
Lead Scoring Assignment

By: Shweta Patil
Amit Pawar

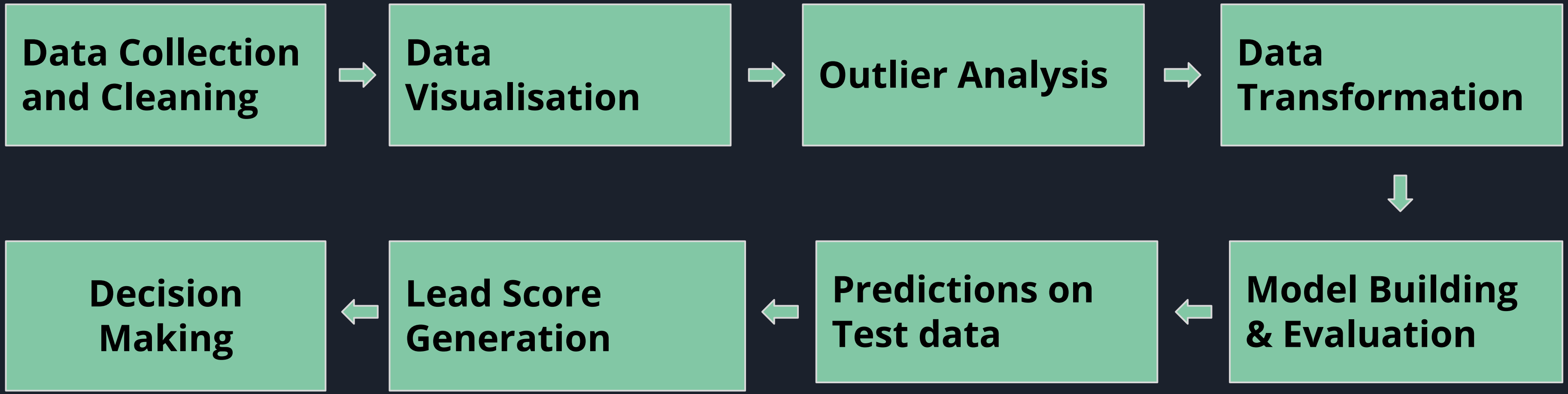


Problem statement:

An education company named X Education sells online courses to industry professionals.

The company requires a model that will assign a lead score 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.

Analysis Approach



Data Collection and Cleaning

Data Visualisation

Outlier Analysis

Data Transformation

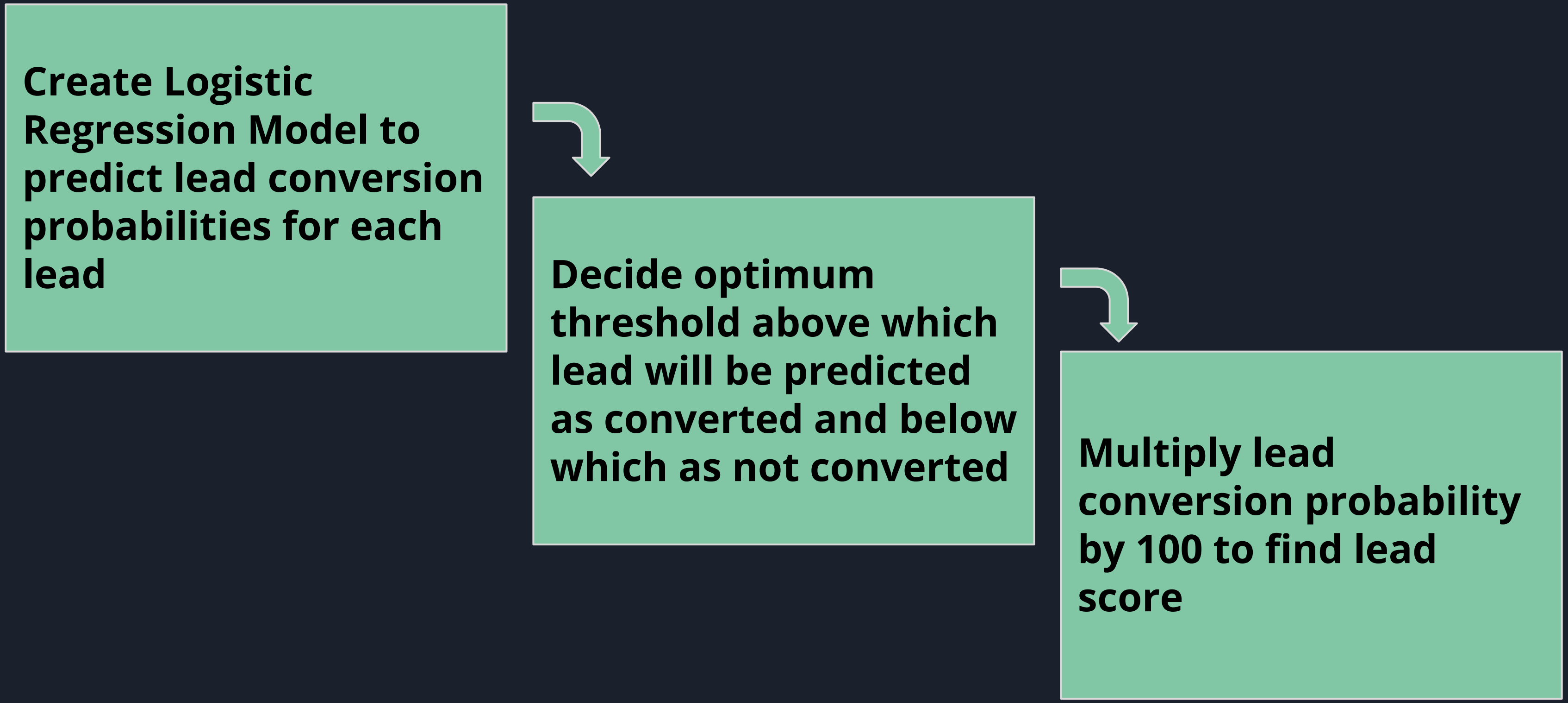
Model Building & Evaluation

Predictions on Test data

Lead Score Generation

Decision Making

Goals



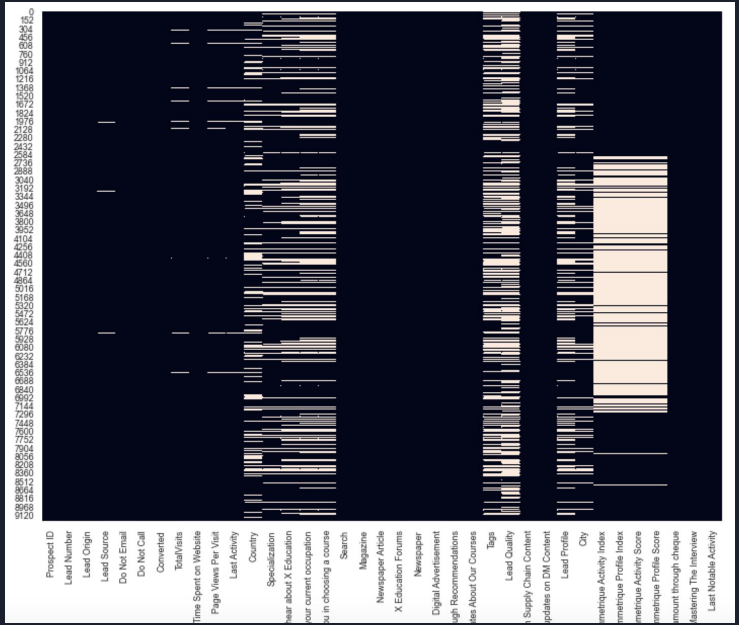
Create Logistic Regression Model to predict lead conversion probabilities for each lead

Decide optimum threshold above which lead will be predicted as converted and below which as not converted

Multiply lead conversion probability by 100 to find lead score

Data Understanding

Plot of Missing values in dataset



1. Drop all the columns having missing values greater than 3000

2. Drop the columns with very low variance

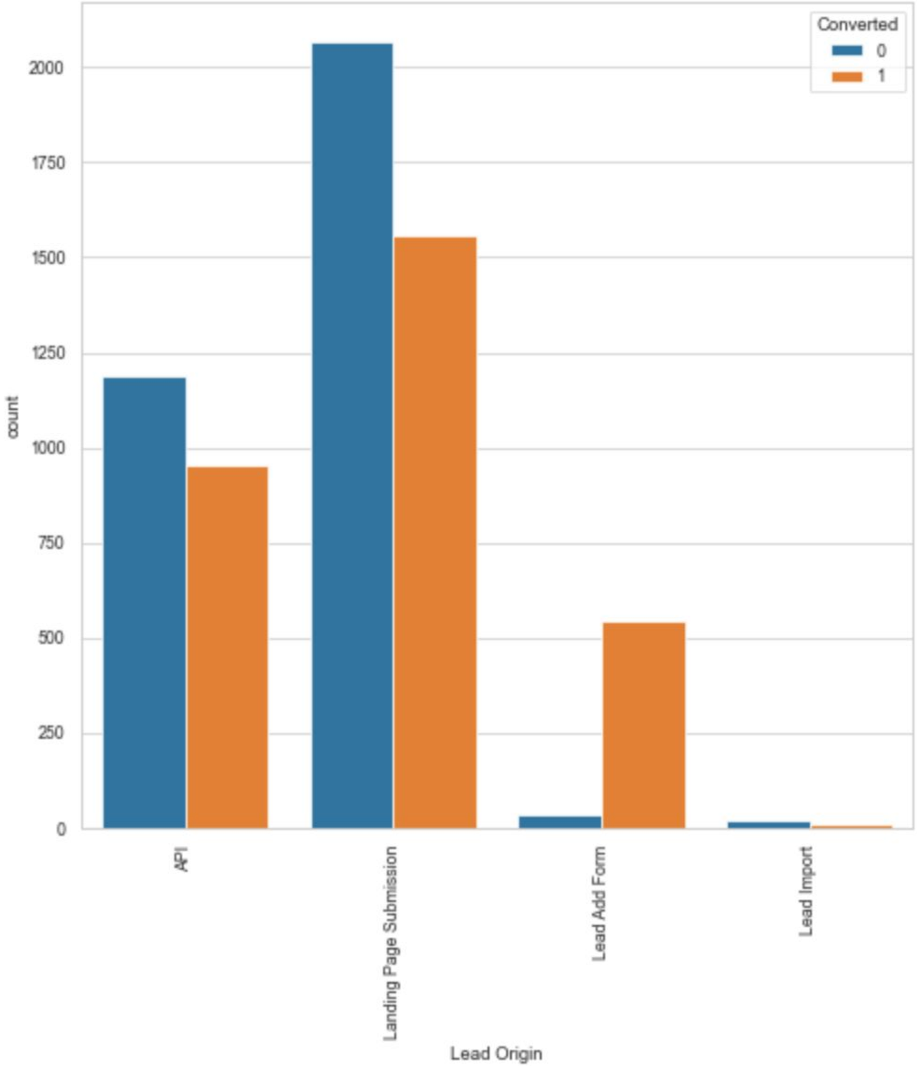
3. Drop the rows where missing values cannot be predicted (eg. TotalVisits)

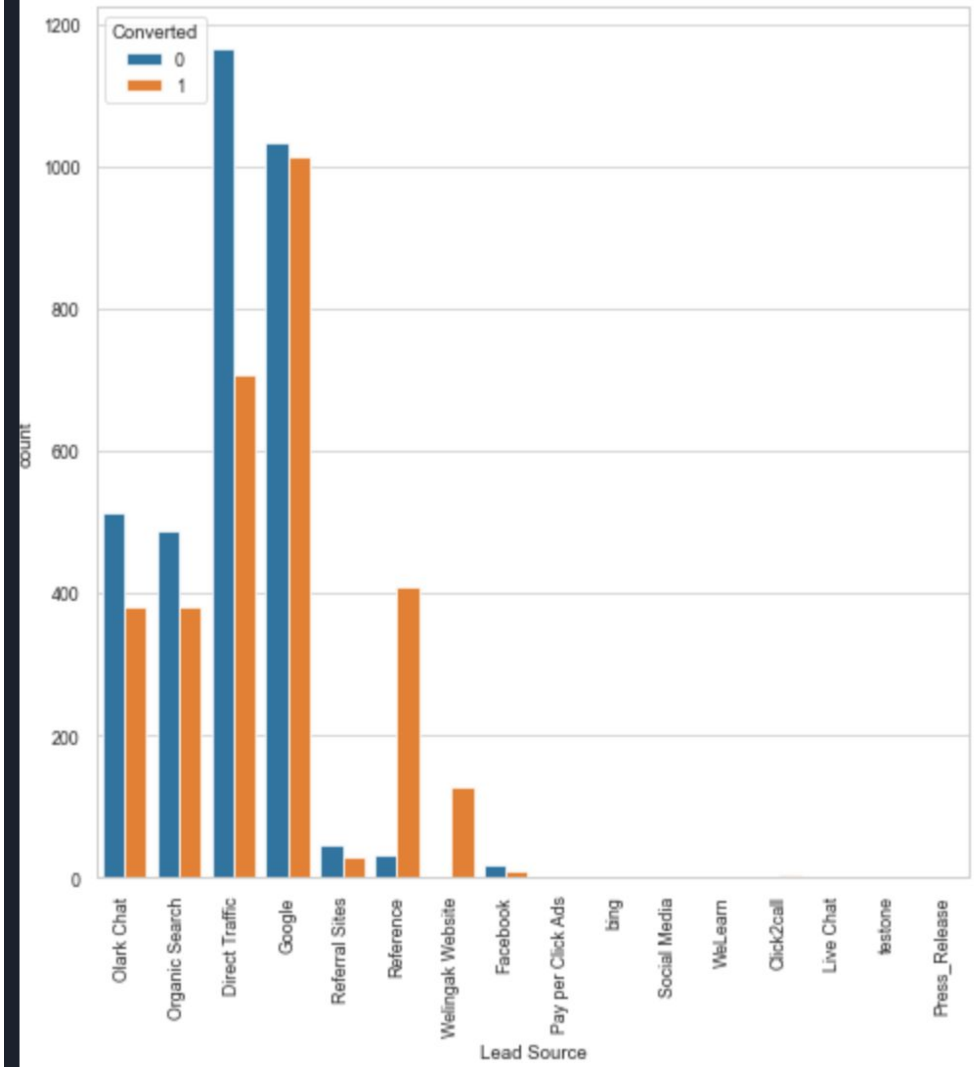
4. Columns with ‘select’ (where user has not provided any input) value are as good as missing values and of no use. So we can drop these values.

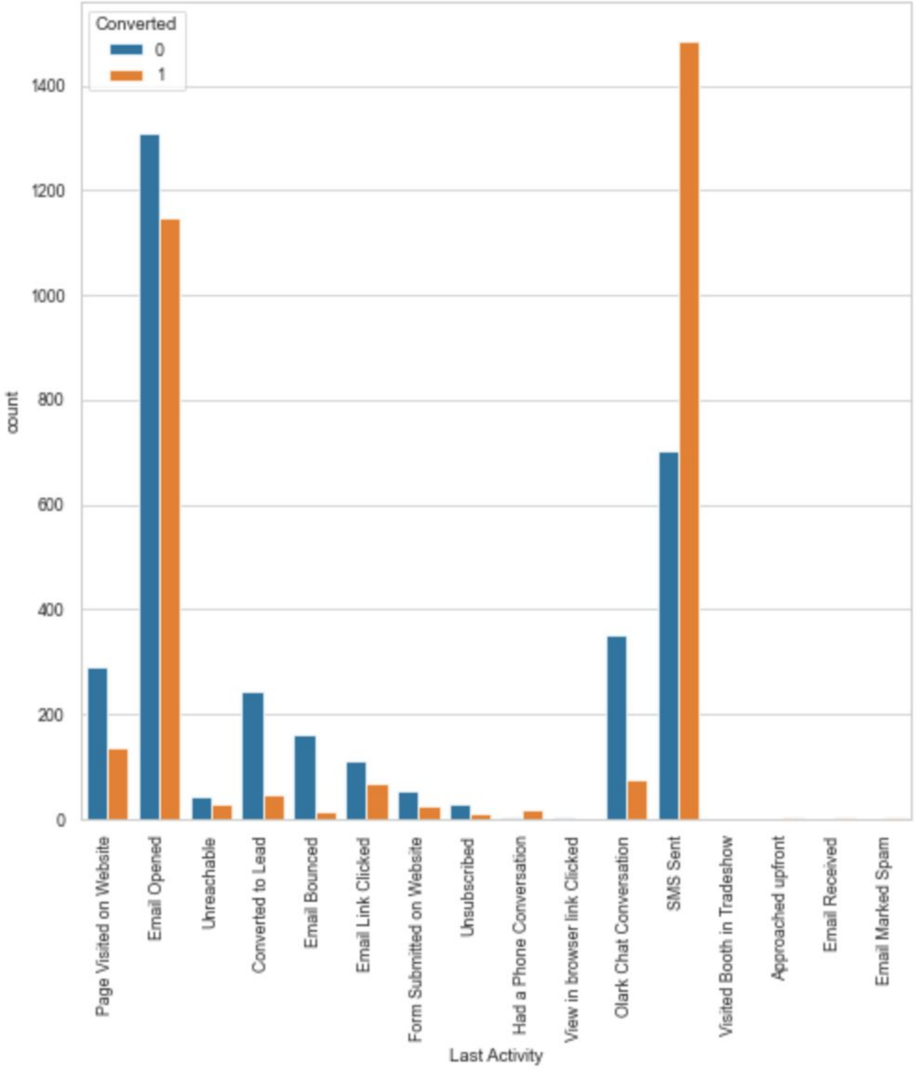
Data Visualization

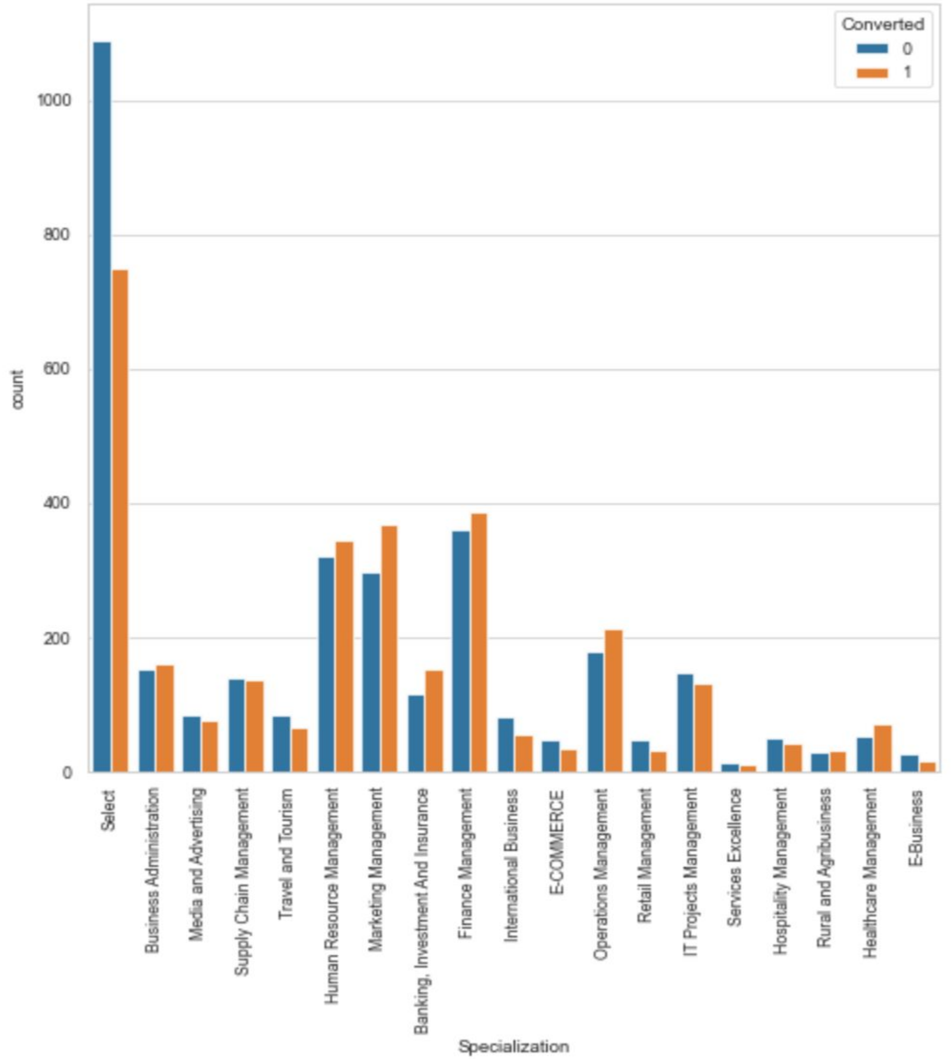
Observations for Categorical columns count plot analysis

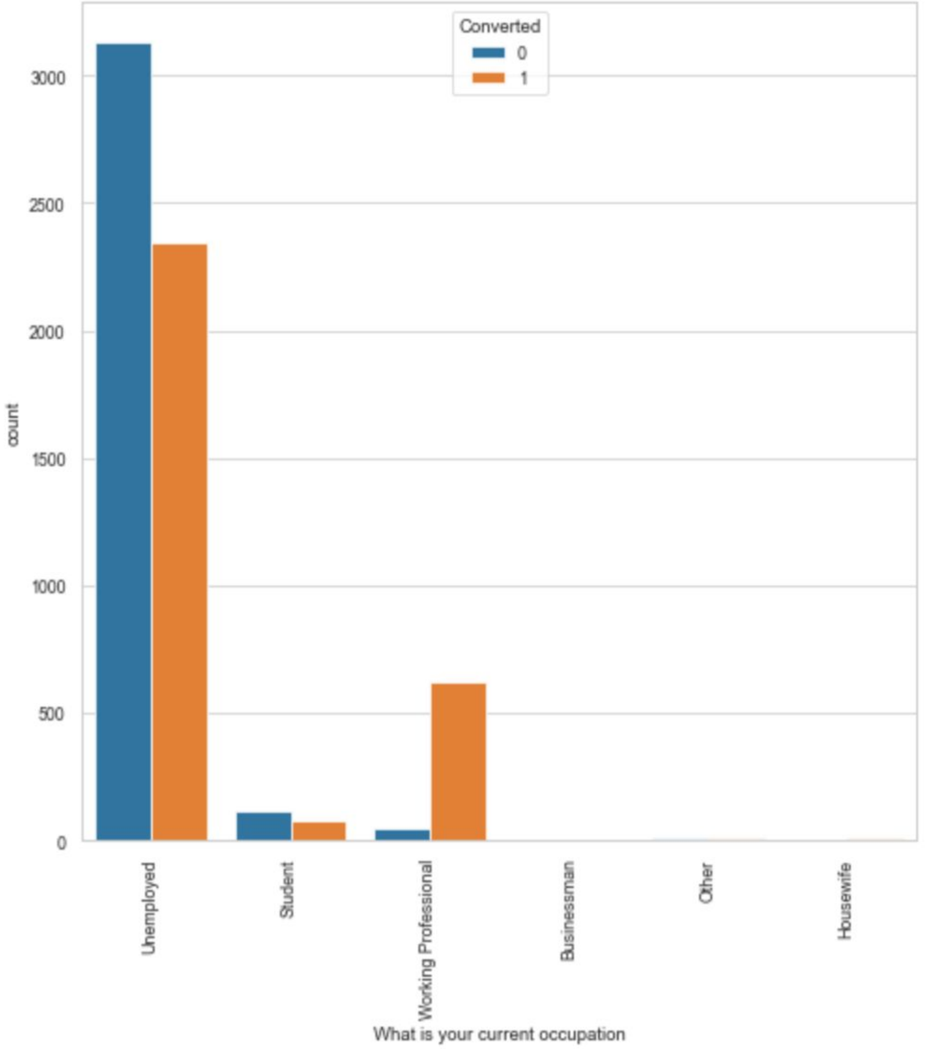
1. Maximum conversion is for students who performed Landing page submission
2. Most of the traffic that is converted is from direct google search engine
3. Last activity in case of converted people is 'SMS sent' where as in case of non converted people is 'email opened'
4. Most of the people have not selected the specialization from dropdown menu
5. Mostly unemployed people are converting and buying courses. This might be to make their profile stronger for further job hunting
6. Dataset is not biased as we have approximately same samples from converted and not converted group of people

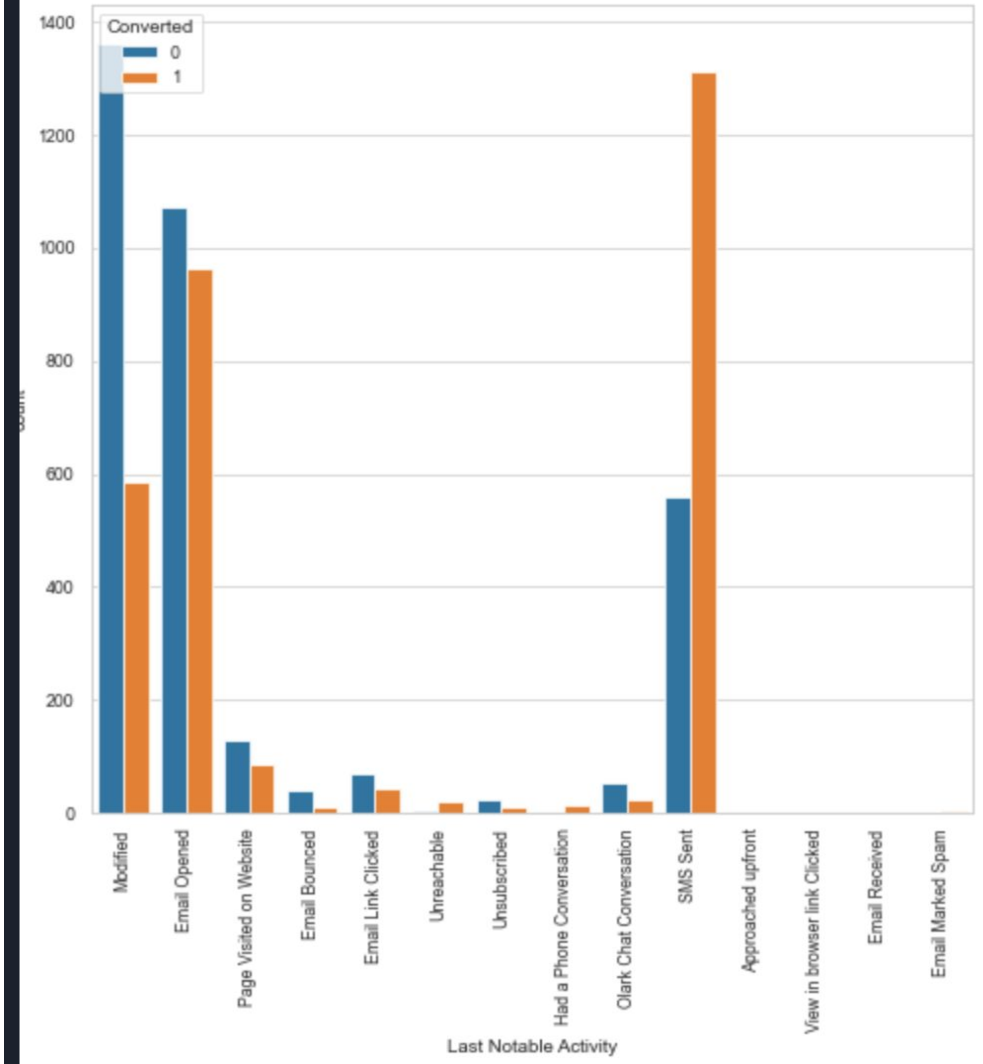

Converted
0
1
count
0
250
500
750
1000
1250
1500
1750
2000
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin


Converted
0
1
count
0
200
400
600
800
1000
1200
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
Welingak Website
Facebook
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
testone
Press_Release
Lead Source

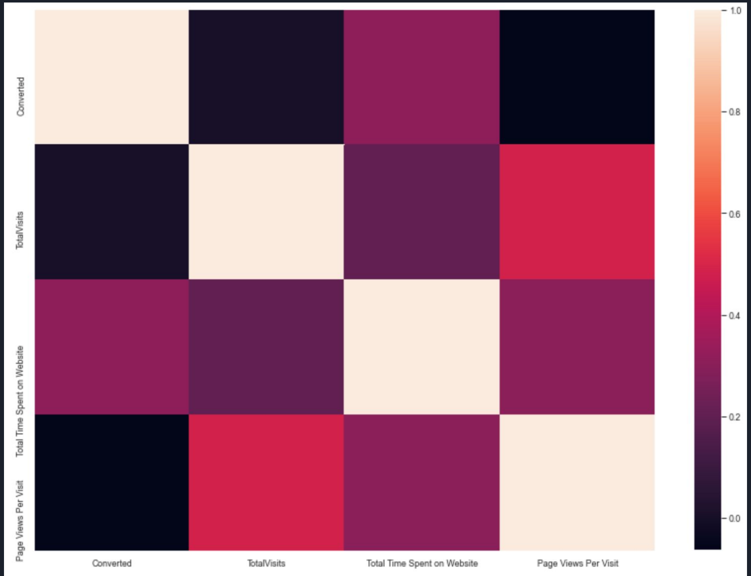

Converted
0
1
count
0
200
400
600
800
1000
1200
1400
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
Olark Chat Conversation
SMS Sent
Visited Booth in Tradeshow
Approached upfront
Email Received
Email Marked Spam
Last Activity

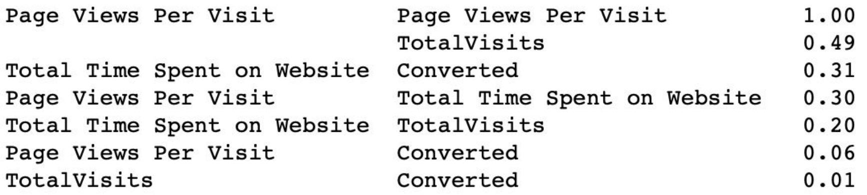

Converted
0
1
count
0
200
400
600
800
1000
Select
Business Administration
Media and Advertising
Supply Chain Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
Finance Management
International Business
E-COMMERCE
Operations Management
Retail Management
IT Projects Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization


Converted
0
1
count
3000
2500
2000
1500
1000
500
0
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation


Converted
0
1
count
1400
1200
1000
800
600
400
200
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
View in browser link Clicked
Email Received
Email Marked Spam
Last Notable Activity

Data Visualization - Correlation matrix

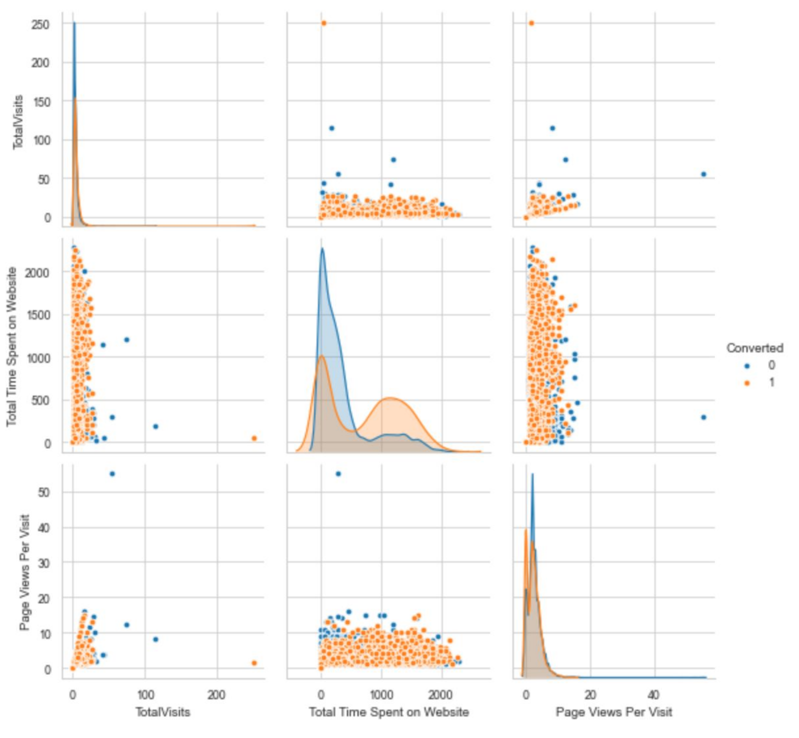




```
Page Views Per Visit         Page Views Per Visit          1.00
                             TotalVisits                   0.49
Total Time Spent on Website  Converted                     0.31
Page Views Per Visit         Total Time Spent on Website   0.30
Total Time Spent on Website  TotalVisits                   0.20
Page Views Per Visit         Converted                     0.06
TotalVisits                  Converted                     0.01
```

We can see that no pairs have very high correlation. Hence we will not drop any columns

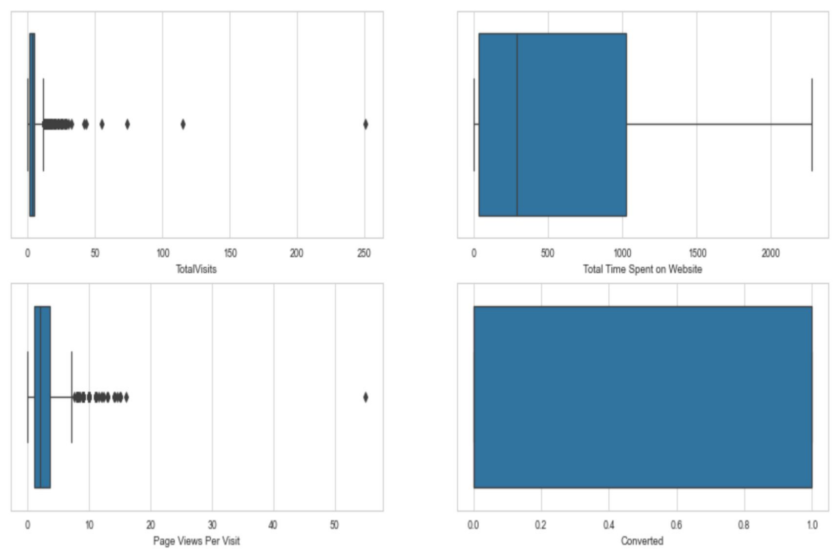
Data Visualization - Numeric columns



There are some strange patterns in these plots.

If the total visits, page views per visit and time spend on website is very high people are less likely to convert

Data Visualization - Outlier Analysis



1. We can see that some people are visiting website many times and hence the graph is right skewed

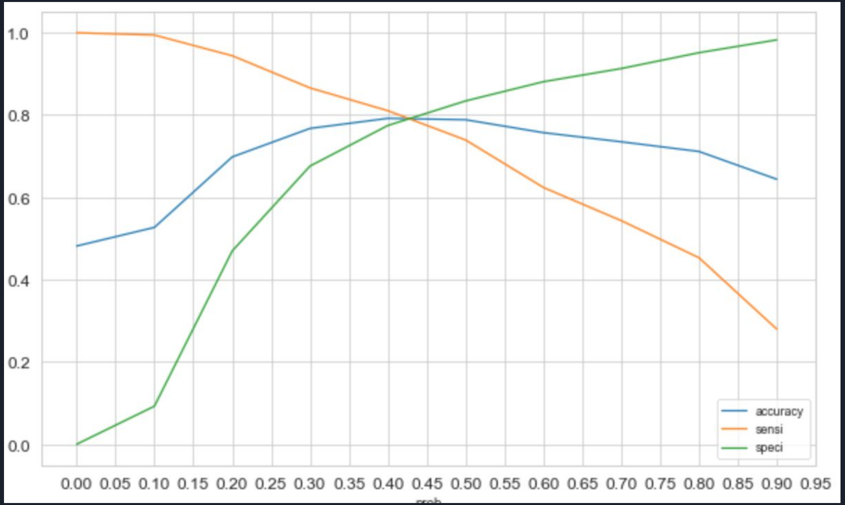
2. There are no outliers in 'Total Time Spent on Website' column

3. Some people seem to spend significantly high time on website

4. In 'Converted' column there is no graph because the value will be either 1(converted) or 0(not converted)

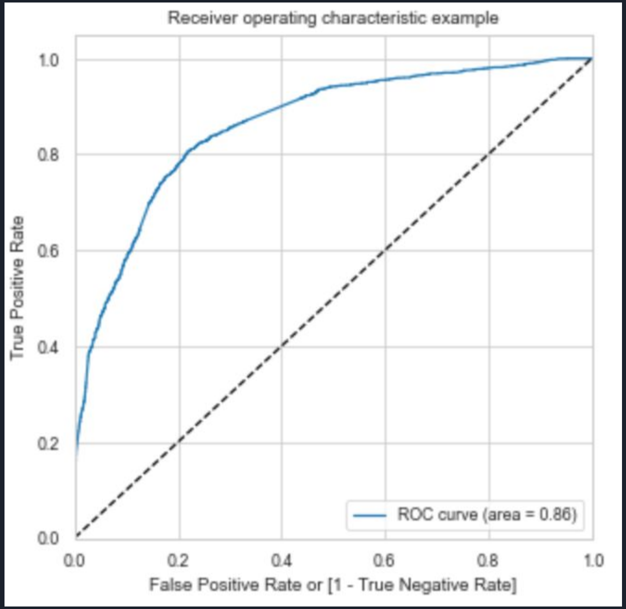
We decide to keep all the outliers as they are important for model building from business point of view

Plot accuracy sensitivity and specificity for various probabilities and find optimal cutoff value on Train set



From the above graph, 0.42 seems to be the optimum cutoff value

ROC for train set

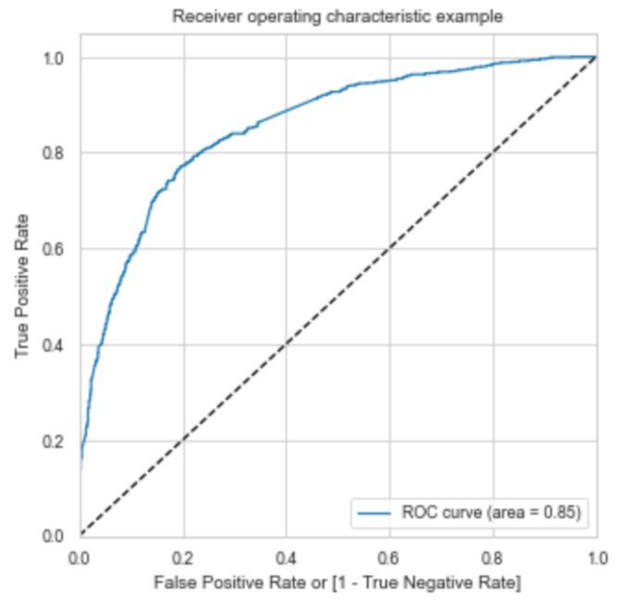


Area under ROC curve comes out to be 0.86 on test set which is good value

Model accuracy comes out to be approximately 79% on Train set with 0.42 cutoff value

Model sensitivity comes out to be approximately 80% on Train set with 0.42 cutoff value

ROC for test set

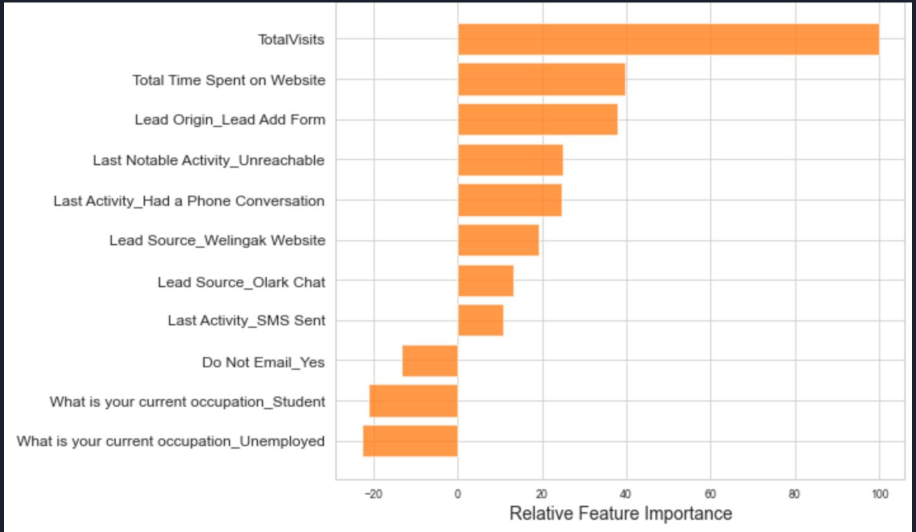


Area under ROC curve comes out to be 0.85 on test set which is good value

Model accuracy comes out to be approximately 79% on Test set with 0.42 cutoff value

Model sensitivity comes out to be approximately 78% on Test set with 0.42 cutoff value

Feature importance



The top 3 features are: ‘TotalVisits’, ‘ Total Time Spent On Website’, ‘Lead Origin_Lead Add Form’

Conclusion

1. Our accuracy and sensitivity is almost same on Train set and Test set which means the model build is stable with adaptive environment skills which will adjust with generic data sets

2. For lead scoring definition, sensitivity is very important from business point of view which comes out to be 78% for Test set in our case

3. The top 3 features responsible for building model with good prediction capability are: ‘TotalVisits’, ‘ Total Time Spent On Website’, ‘Lead Origin_Lead Add Form’

Recommandation

1. If we want to predict maximum leads correctly then we can lower the optimum threshold value for Conversion probability

2. If we keep the optimum threshold value for Conversion probability very high, we can restrict ourselves to getting extremely important leads whose conversion probability will be very high

3. As we observed, 'TotalVisits', ' Total Time Spent On Website', 'Lead Origin_Lead Add Form' play important role in model creation. These factors are related to user interaction with website/landing pages

Hence we must design the website and landing pages which are informative, precise and appealing to users.

“Information is the oil of the 21st century, and analytics is the combustion engine.”

- Peter Sondergaard

Thanks!

PLAN
Success!
Business
Idea and Strategy
Puzzle
Process
Marketing
Today:
SUCCESS
BUSINESS PLAN
Team
Problems
Solutions
messages
GROW
Analysis
computing
50,000.00
Choice
50%
30%
35%