LEAD SCORING CASE STUDY

Submitted By: Parth Patpatiya

Problem Statement

- A model is to be formed that will allocate a lead score to each lead, wherein consumers who have higher lead scores exhibit a greater likelihood of conversion, while customers with lower lead scores demonstrate a diminished likelihood of conversion.

Importing Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from sklearn.model_selection import train_test_split, GridSearchCV, cross_validate
from statsmodels.stats.outliers_influence import variance_inflation_factor
from sklearn.preprocessing import MinMaxScaler
from sklearn.feature_selection import RFE
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, roc_curve, roc_auc_score, precision_score, recall_score, precision_recall_curve, f1_score, accuracy_score
```

Importing Data

In [2]:
```
# Importing data
df = pd.read_csv("D:\\DataSets\\UPGRAD\\Assignments\\Lead+Scoring+Case+Study\\Leads.csv")
df.head()
```

Out[2]:

	Prospect ID	Lead Number	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Last Activity	Country	Specialization	How dic you hea about X Educatior
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	API	Olark Chat	No	No	0	0.0	0	0.0	Page Visited on Website	NaN	Select	Select
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	No	No	0	5.0	674	2.5	Email Opened	India	Select	Select
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	Email Opened	India	Business Administration	Select
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	660719	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0	Unreachable	India	Media and Advertising	Word Of Mouth
4	3256f628-e534-4826-9d63-4a8b88782852	660681	Landing Page Submission	Google	No	No	1	2.0	1428	1.0	Converted to Lead	India	Select	Other

Checking Null Values

```
In [3]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
Prospect ID                                    9240 non-null object
Lead Number                                    9240 non-null int64
Lead Origin                                    9240 non-null object
Lead Source                                    9204 non-null object
Do Not Email                                   9240 non-null object
Do Not Call                                    9240 non-null object
Converted                                      9240 non-null int64
TotalVisits                                    9103 non-null float64
Total Time Spent on Website                    9240 non-null int64
Page Views Per Visit                           9103 non-null float64
Last Activity                                  9137 non-null object
Country                                        6779 non-null object
Specialization                                 7802 non-null object
How did you hear about X Education             7033 non-null object
What is your current occupation                6550 non-null object
What matters most to you in choosing a course  6531 non-null object
Search                                         9240 non-null object
Magazine                                       9240 non-null object
Newspaper Article                              9240 non-null object
X Education Forums                             9240 non-null object
Newspaper                                      9240 non-null object
Digital Advertisement                          9240 non-null object
Through Recommendations                        9240 non-null object
Receive More Updates About Our Courses         9240 non-null object
Tags                                           5887 non-null object
Lead Quality                                   4473 non-null object
Update me on Supply Chain Content              9240 non-null object
Get updates on DM Content                      9240 non-null object
Lead Profile                                   6531 non-null object
City                                           7820 non-null object
Asymmetrique Activity Index                    5022 non-null object
Asymmetrique Profile Index                     5022 non-null object
Asymmetrique Activity Score                    5022 non-null float64
Asymmetrique Profile Score                     5022 non-null float64
I agree to pay the amount through cheque       9240 non-null object
A free copy of Mastering The Interview         9240 non-null object
Last Notable Activity                          9240 non-null object
dtypes: float64(4), int64(3), object(30)
```

Checking mean, median and mode

```
In [4]: df.describe()
```

Out[4]:

	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Asymmetrique Activity Score	Asymmetrique Profile Score
count	9240.000000	9240.000000	9103.000000	9240.000000	9103.000000	5022.000000	5022.000000
mean	617188.435606	0.385390	3.445238	487.698268	2.362820	14.306252	16.344883
std	23405.995698	0.486714	4.854853	548.021466	2.161418	1.386694	1.811395
min	579533.000000	0.000000	0.000000	0.000000	0.000000	7.000000	11.000000
25%	596484.500000	0.000000	1.000000	12.000000	1.000000	14.000000	15.000000
50%	615479.000000	0.000000	3.000000	248.000000	2.000000	14.000000	16.000000
75%	637387.250000	1.000000	5.000000	936.000000	3.000000	15.000000	18.000000
max	660737.000000	1.000000	251.000000	2272.000000	55.000000	18.000000	20.000000

Data Cleaning

- There are several columns that exist inside a dataset that contain only one category, resulting in redundancy.
- In the column ***lead_score:*** *WeLearn* and *WeLearnblog_home* are similar.
- 'Select' in ***How Did You Do, Specialization, City*** is NaN.
- There is a potential for the presence of overlaps in the City column.
- There are two distinct categories for "*Asia/Pacific Region*" and "*Asian Countries*" in the table's column labeled "Australia" for the nation.
- The entry '*select' are equivalent to NaN*.

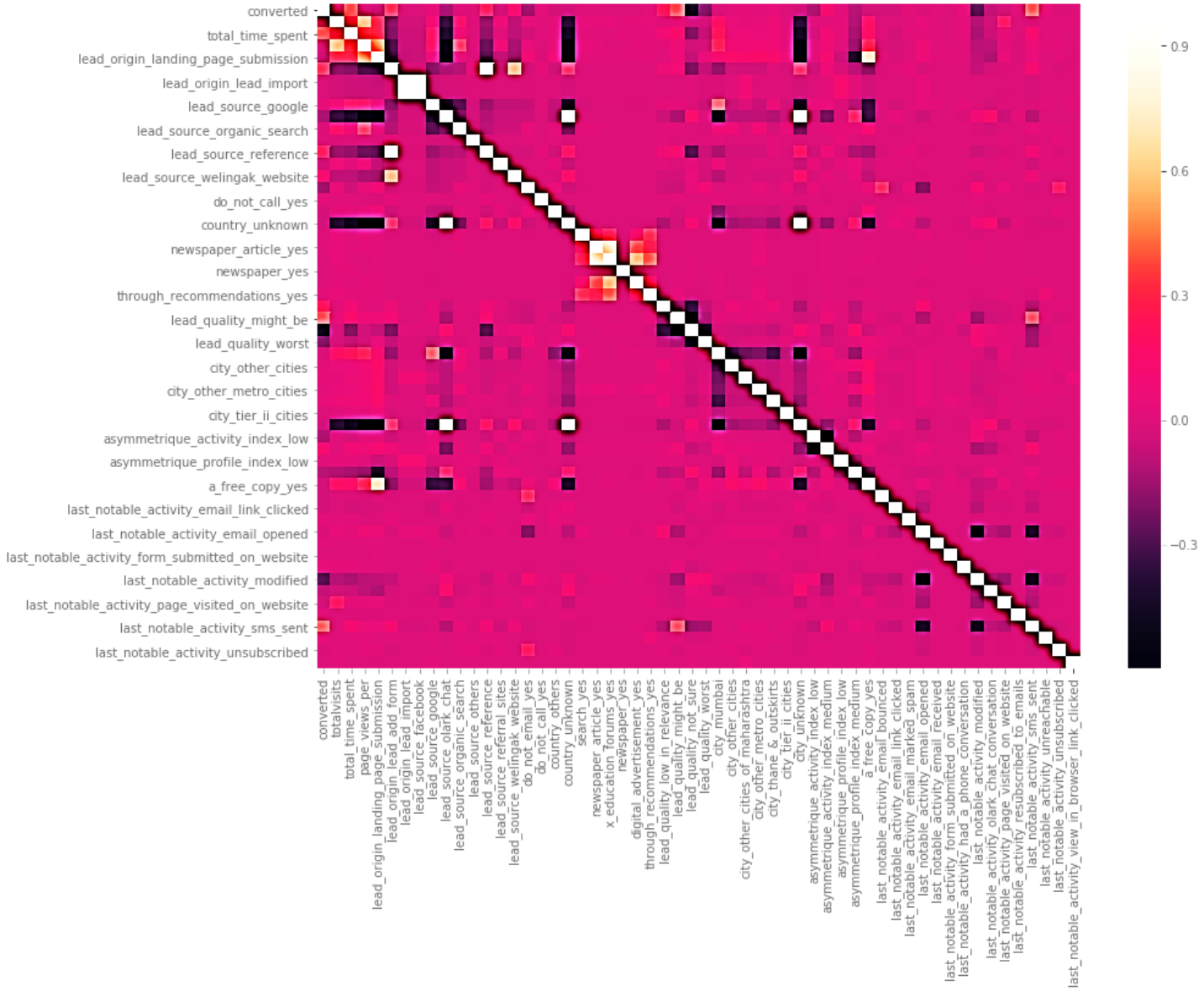
Working with Nan (not-a-number) values

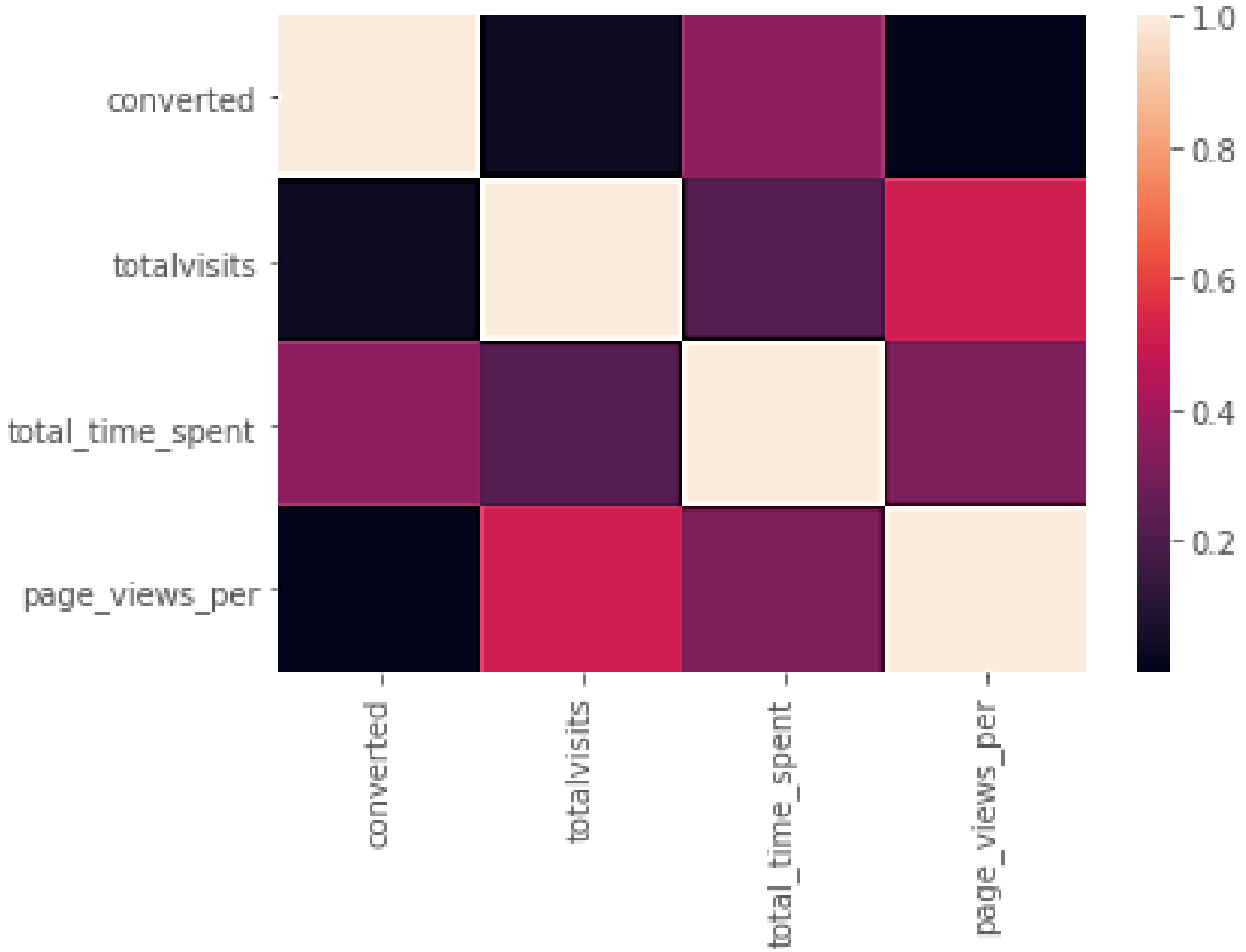
- Imputed Values: 'how_did_you', *lead_profile*
- Asymmetrique scores: 'profile_score', 'activity_score'
- Dropping Values: 'asymmetrique_profile_score', 'asymmetrique_activity_score
- Eliminating prefixes from 'asymmetrique_profile_index'
- Replacing NaN with mode value in 'asymmetrique_activity_index'
- Droping na values in ['totalvisits', 'page_views_per', 'lead_source'

Creating Dummy variables

- Creating dummy entries for 'object‘ column

Exploratory Data Analysis (EDA)

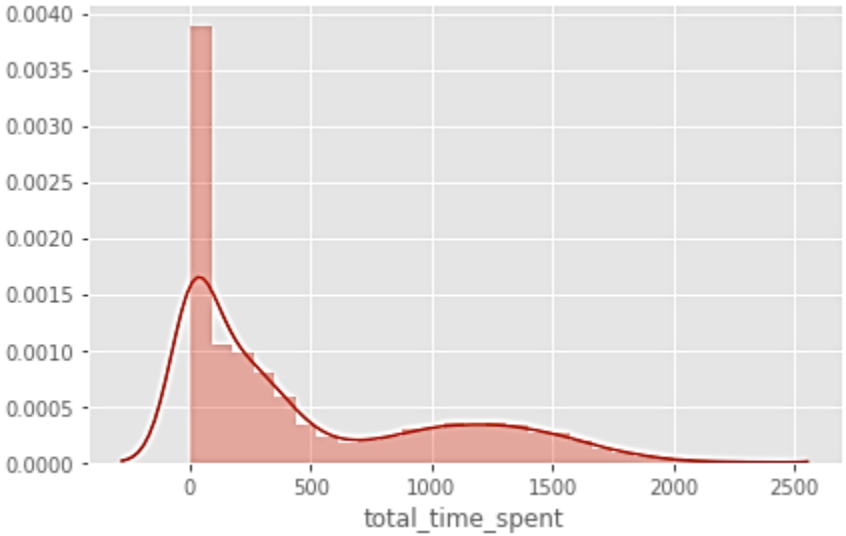


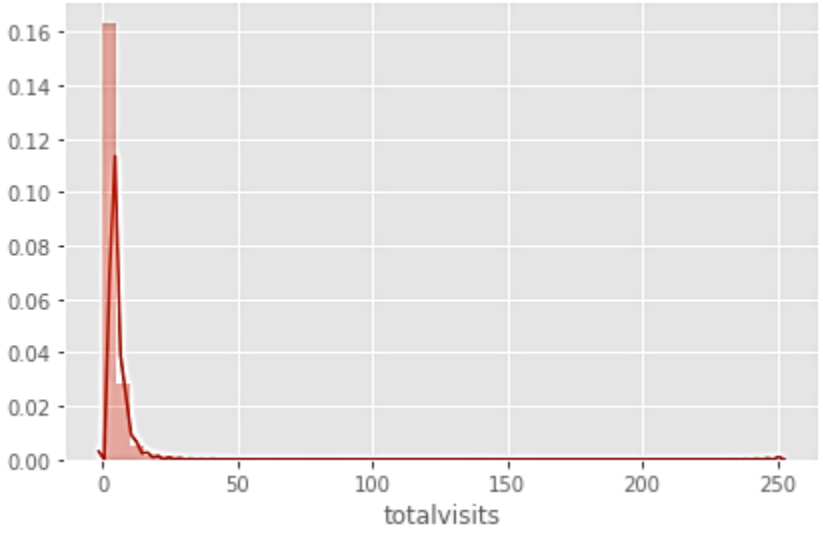

converted
totalvisits
total_time_spent
page_views_per
converted
totalvisits
total_time_spent
page_views_per
1.0
0.8
0.6
0.4
0.2

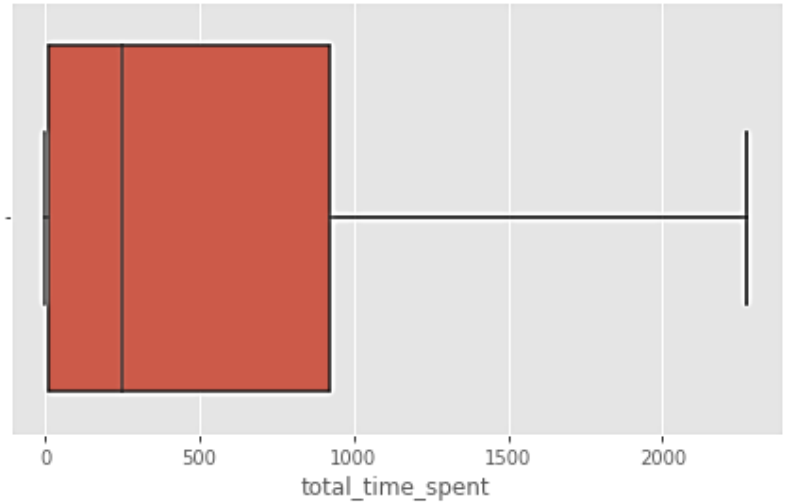
Removing highly correlated variables

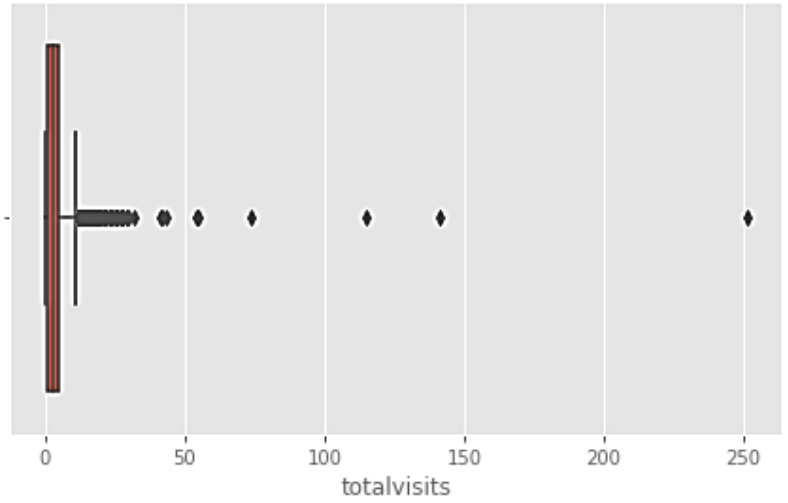
- ['lead_source_facebook', 'lead_origin_lead_add_form', 'lead_source_olark_chat']

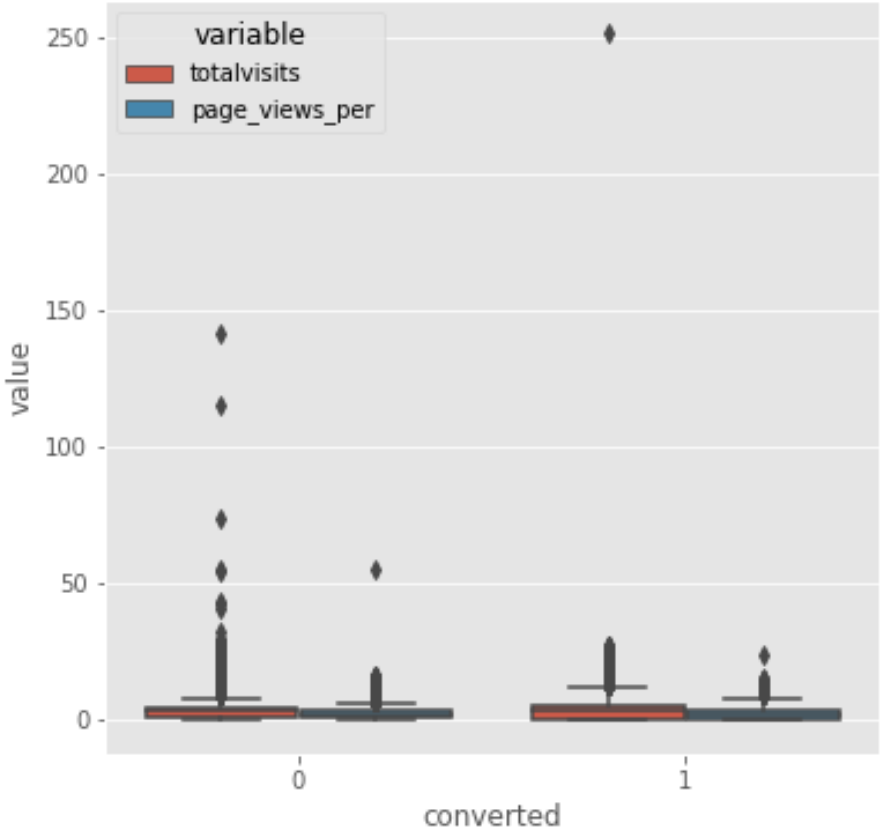
Outlier Detection and Univariate Analysis (Bar Plot and Subplot)






0
500
1000
1500
2000
total_time_spent


0
50
100
150
200
250
totalvisits


variable
totalvisits
page_views_per
250
200
150
100
50
0
value
0
1
converted

Bivariate Analysis

- Pair Plot framed for the variables:
- ['totalvisits', 'page_views_per', 'total_time_spent']

Linear Predictive Model

- Steps for building the predictive model:

1. Splitting the model(train_test_split)
2. scaler=MinMaxScaler() [X__train, X_test]
3. Recursive Feature Elimination (RFE) & Cross Validation
4. Variable Influence Factor Analysis
5. Receiving Operating Characteristic Curve
6. Sensitivity

Variable Influence Factor Analysis

	Features	VIF
1	total_time_spent	19.12
0	totalvisits	14.11
17	PC2	10.39
8	lead_quality_not_sure	3.72
5	country_unknown	2.51
7	lead_quality_might_be	1.94
14	last_notable_activity_sms_sent	1.52
9	lead_quality_worst	1.23
4	do_not_email_yes	1.19
2	lead_source_reference	1.18
16	last_notable_activity_unsubscribed	1.09
3	lead_source_welingak_website	1.07
10	asymmetrique_activity_index_low	1.06
12	last_notable_activity_olark_chat_conversation	1.06
11	last_notable_activity_had_a_phone_conversation	1.01
15	last_notable_activity_unreachable	1.01
6	search_yes	1.00
13	last_notable_activity_resubscribed_to_emails	1.00

Statistical Model Assesment

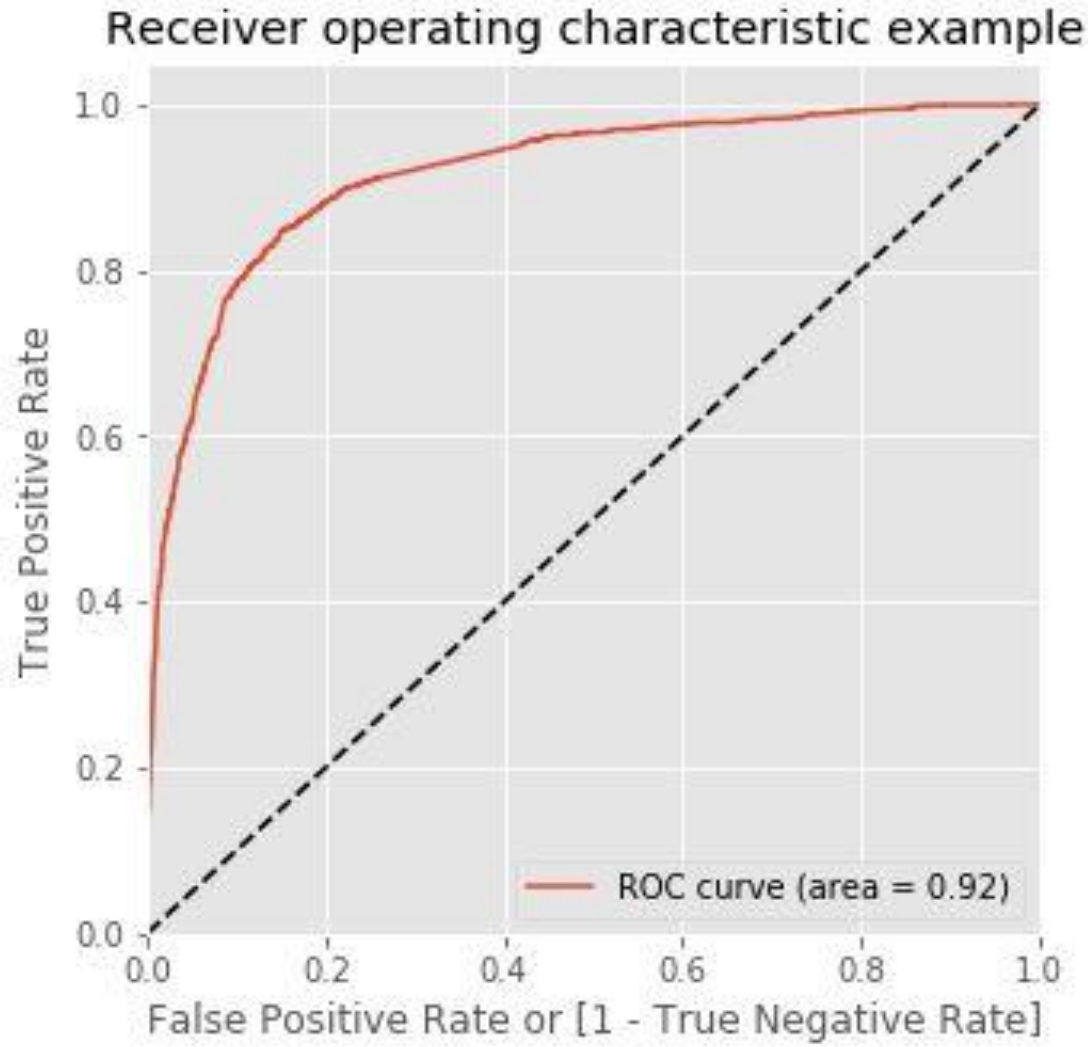
- Statsmodels.api

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.5615	0.154	3.647	0.000	0.260	0.863
totalvisits	3.0614	0.243	12.579	0.000	2.584	3.538
lead_source_reference	1.7742	0.273	6.509	0.000	1.240	2.308
do_not_email_yes	-1.2528	0.209	-5.993	0.000	-1.662	-0.843
country_unknown	1.3248	0.130	10.171	0.000	1.069	1.580
search_yes	-2.1603	1.321	-1.635	0.102	-4.749	0.429
lead_quality_might_be	-1.5194	0.166	-9.180	0.000	-1.844	-1.195
lead_quality_not_sure	-3.5546	0.152	-23.415	0.000	-3.852	-3.257
lead_quality_worst	-5.6548	0.414	-13.655	0.000	-6.466	-4.843
asymmetrique_activity_index_low	-1.8210	0.297	-6.122	0.000	-2.404	-1.238
last_notable_activity_had_a_phone_conversation	2.4401	1.251	1.950	0.051	-0.012	4.892
last_notable_activity_olark_chat_conversation	-1.0916	0.363	-3.009	0.003	-1.803	-0.380
last_notable_activity_sms_sent	1.8976	0.092	20.558	0.000	1.717	2.078
last_notable_activity_unreachable	2.1003	0.642	3.272	0.001	0.842	3.359
last_notable_activity_unsubscribed	1.1198	0.611	1.834	0.067	-0.077	2.316
PC2	4.2052	0.192	21.892	0.000	3.829	4.582

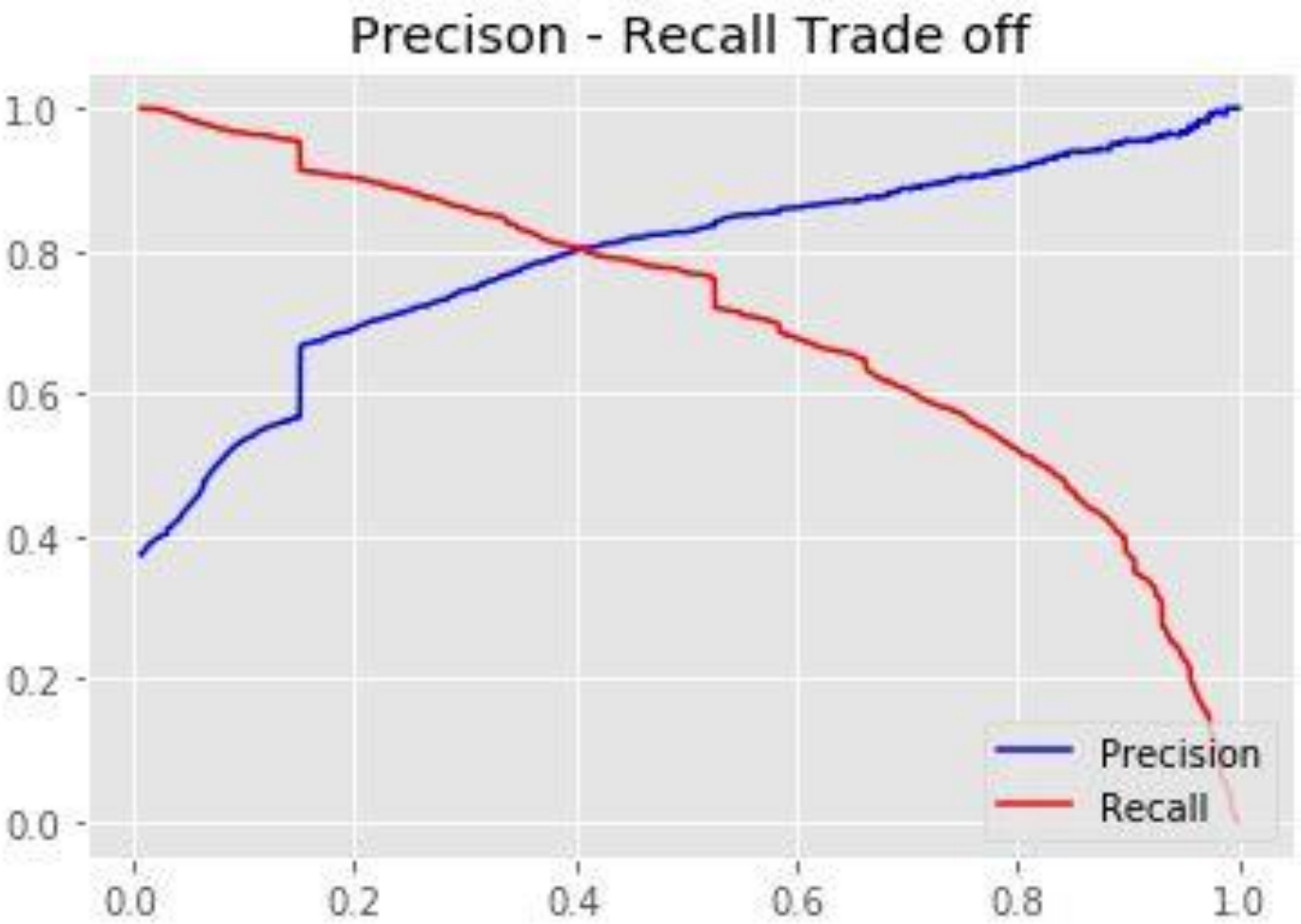
Optimal Cut off value

- Optimal Cut off value=0.32
- Sensitivity (Recall): 0.78
- Specificity: 0.91
- Precision: 0.83
- F-Score: 0.79

Receiving Operating Characteristic Curve



Precision & Recall



Conclusion

- Ultimately, our Logistic Regression model has an overall accuracy of about 0.85. It may be inferred that there exists an 85% probability of successful conversion for the leads that have been anticipated. This achieves the objective of achieving a minimum of 80% lead conversion.