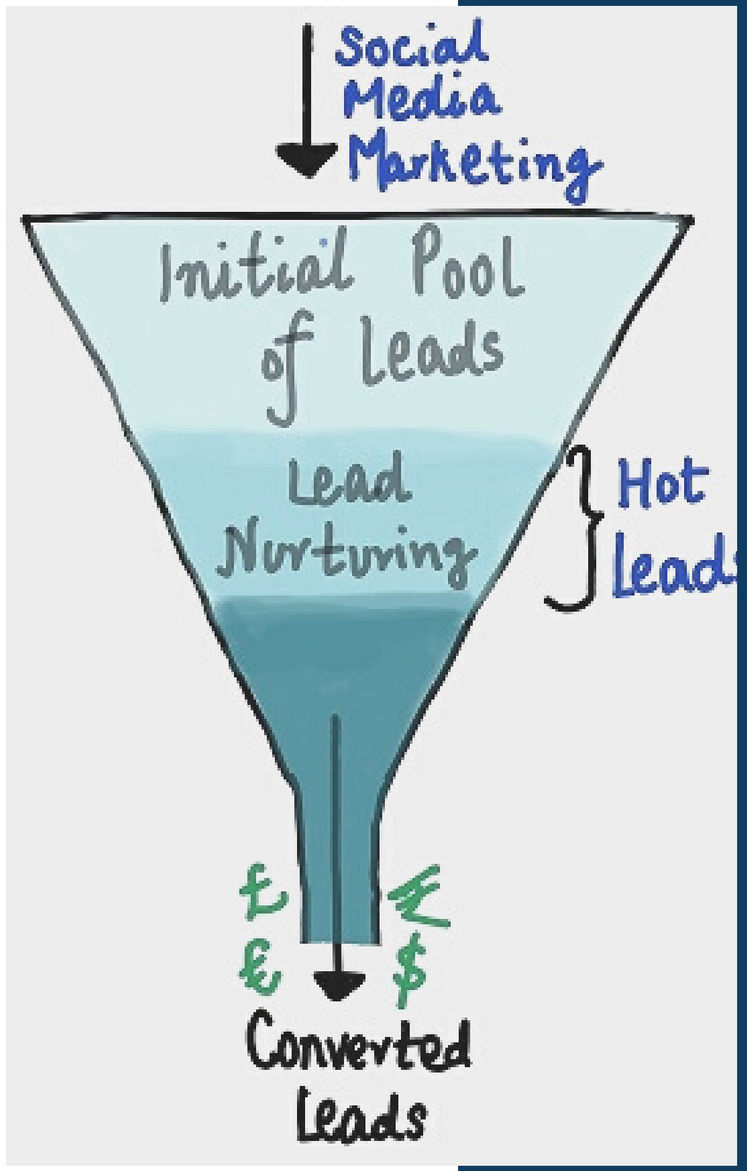
LEAD SCORING ASSIGNMENT

Pragya Deolal & Gaurav Sawant

Contents:



Problem Statement:

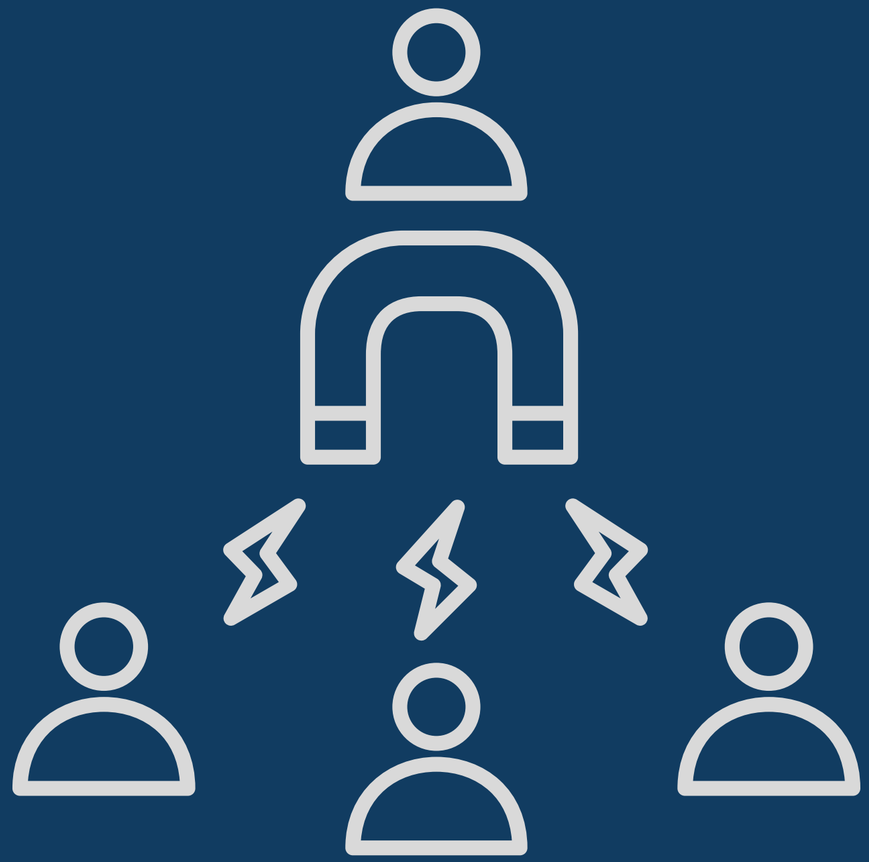
An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Objective:

Build a logistic regression model

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.



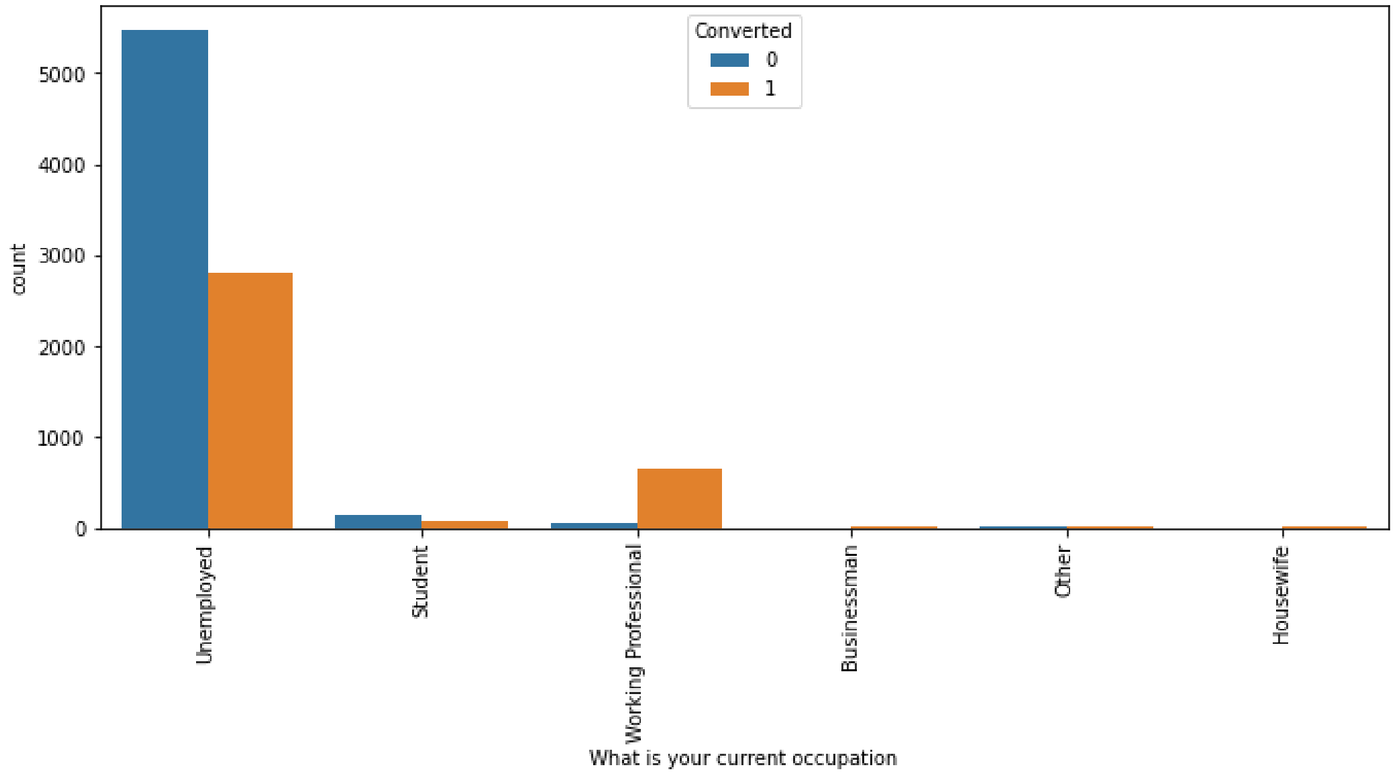
Importing the required libraries and dataset:

```
# importing all the required libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

import warnings
warnings.filterwarnings('ignore')

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
from sklearn.linear_model import LogisticRegression
from sklearn.feature_selection import RFE
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import variance_inflation_factor
from sklearn import metrics
from sklearn.metrics import confusion_matrix
from sklearn.metrics import precision_score, recall_score
from sklearn.metrics import precision_recall_curve
```

```
#import dataset
leads = pd.read_csv('Leads.csv')
leads.head()
```

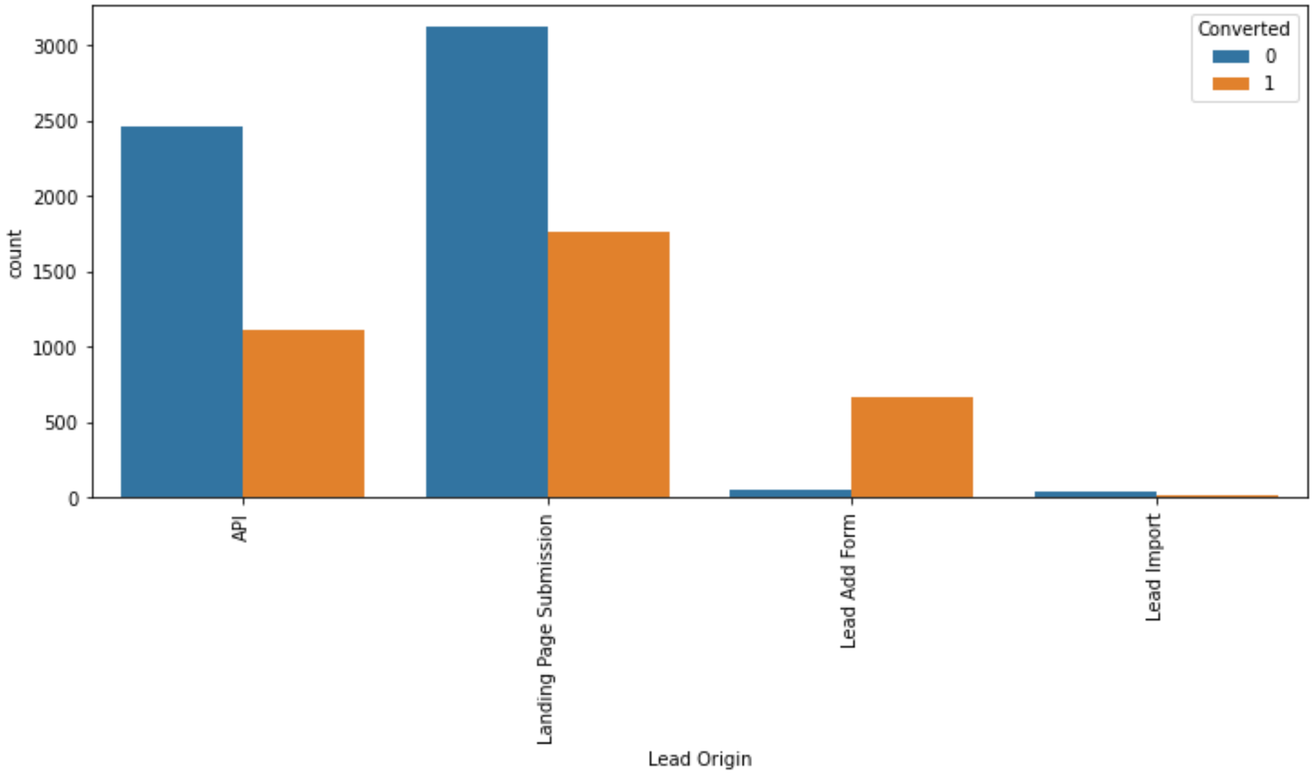
EDA - Data Cleaning & Treatment

- Working Professionals going for the course have high chances of joining it.
- Unemployed leads are the most in terms of Absolute numbers.

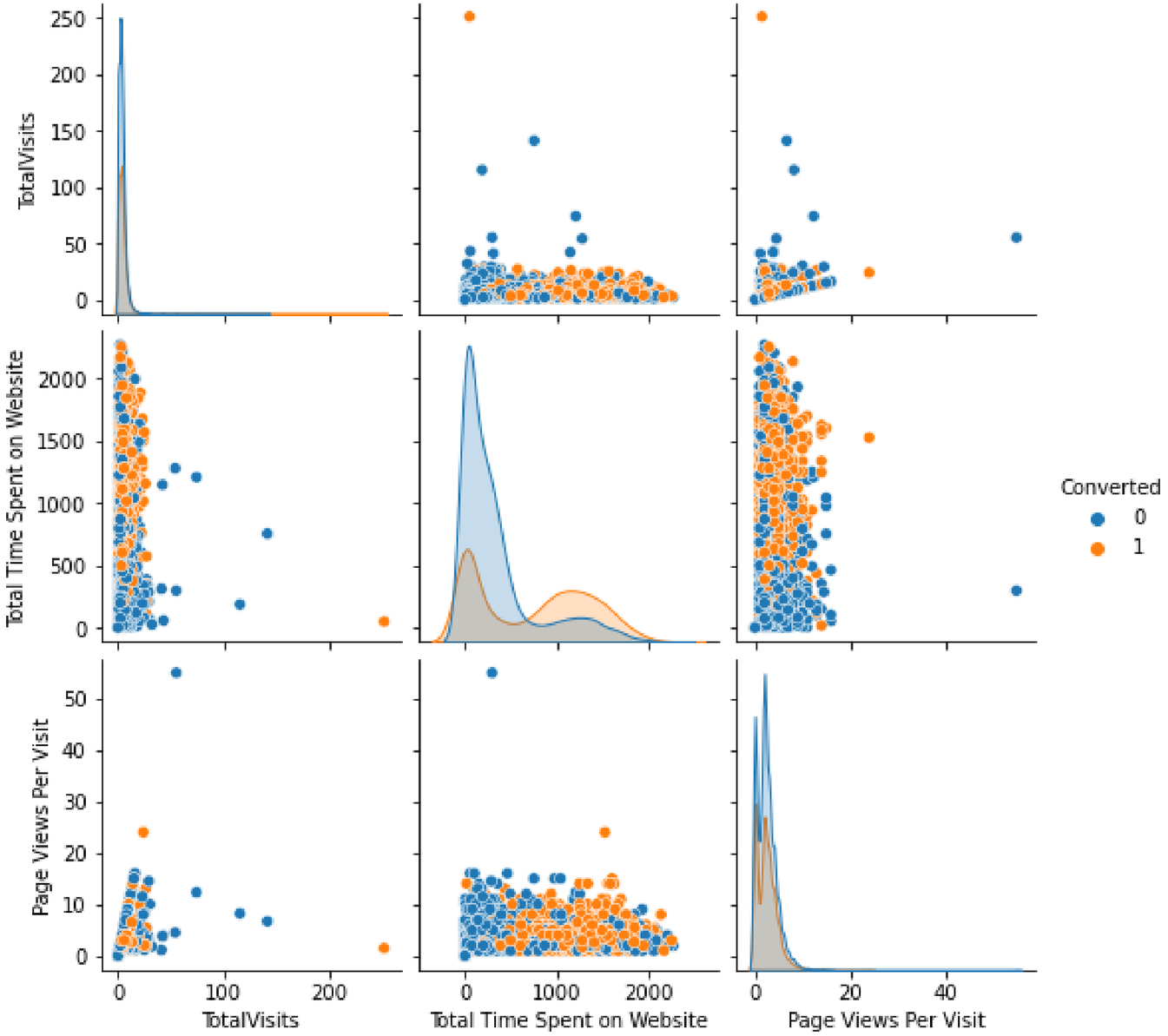


EDA - Data Cleaning & Treatment

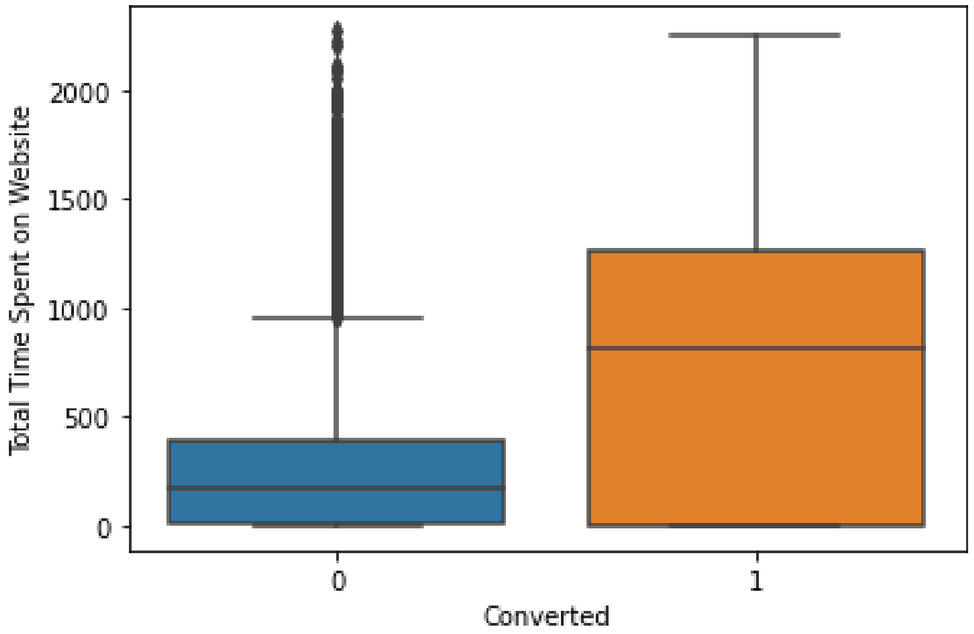
- API and Landing Page Submission bringing higher number of leads as well as conversion.
- Lead Add Form has a very high conversion rate but count of leads are not very high.
- Lead Import and Quick Add Form get less leads.
- In order to improve overall lead conversion rate, improve lead converion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

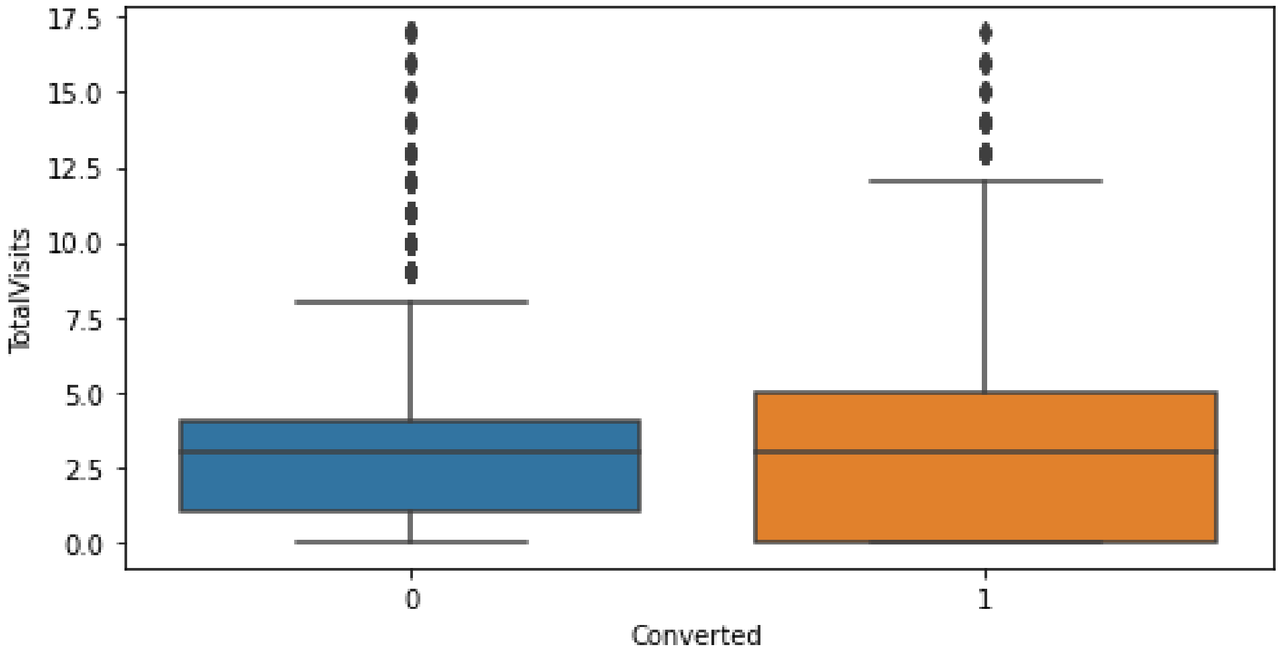


Correlation of numeric values with respect to 'Converted' values



Box plot for Total time Spent on Website and Total Visits on website



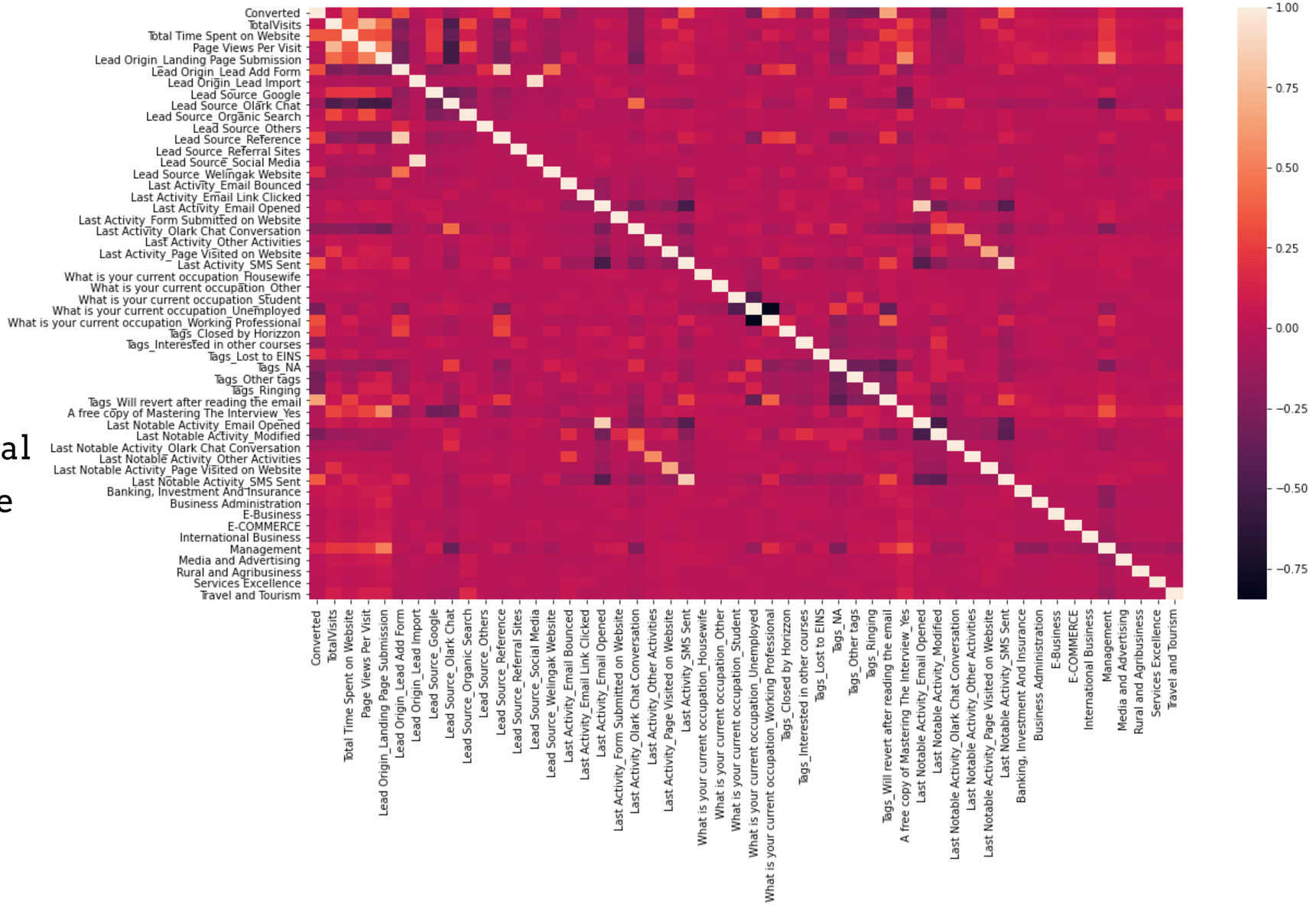


- Leads spending more time on the website are more likely to be converted.
- Website should be made more engaging to make leads spend more time

Correlation Matrix

It can be inferred from the correlation matrix that :
Last Notable Activity_Had a Phone Conversation is an important feature.

Variable What is your current occupation_Working Professional shows a high correlation with the probability of leads conversion whereas For Users with current occupation as Students isn't a better option to go with.



Model 1:

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.7033	0.164	-16.488	0.000	-3.025	-2.382
TotalVisits	1.2605	0.430	2.935	0.003	0.419	2.102
Total Time Spent on Website	-1.4407	0.643	-2.241	0.025	-2.701	-0.181
Page Views Per Visit	4.4214	0.251	17.613	0.000	3.929	4.913
Lead Origin_Lead Add Form	2.0497	0.444	4.620	0.000	1.180	2.919
Lead Source_Olark Chat	1.1999	0.177	6.790	0.000	0.854	1.546
Lead Source_Welingak Website	3.5121	0.849	4.137	0.000	1.848	5.176
Last Activity_SMS Sent	2.1653	0.117	18.448	0.000	1.935	2.395
What is your current occupation_Working Professional	1.0798	0.389	2.774	0.006	0.317	1.843
Tags_Closed by Horizzon	7.2491	1.023	7.083	0.000	5.243	9.255
Tags_Interested in other courses	-1.9443	0.389	-4.998	0.000	-2.707	-1.182
Tags_Lost to EINS	5.9070	0.611	9.669	0.000	4.710	7.104
Tags_Other tags	-2.4733	0.212	-11.664	0.000	-2.889	-2.058
Tags_Ringing	-3.4749	0.244	-14.238	0.000	-3.953	-2.997
Tags_Will revert after reading the email	4.4000	0.195	22.510	0.000	4.017	4.783
Last Notable Activity_Modified	-1.7038	0.127	-13.370	0.000	-1.954	-1.454

For model 1, Total Time Spent on Website has comparatively high p value and it also showed high VIF i.e greater than 5, thus, the variable will be dropped for more feasible model.

Model 2:

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.8309	0.154	-18.390	0.000	-3.133	-2.529
TotalVisits	0.7197	0.357	2.016	0.044	0.020	1.419
Page Views Per Visit	4.4208	0.251	17.643	0.000	3.930	4.912
Lead Origin_Lead Add Form	2.2009	0.439	5.016	0.000	1.341	3.061
Lead Source_Olark Chat	1.3364	0.166	8.058	0.000	1.011	1.661
Lead Source_Welingak Website	3.4945	0.849	4.116	0.000	1.830	5.159
Last Activity_SMS Sent	2.1385	0.116	18.359	0.000	1.910	2.367
What is your current occupation_Working Professional	1.0610	0.390	2.723	0.006	0.297	1.825
Tags_Closed by Horizzon	7.2446	1.023	7.083	0.000	5.240	9.249
Tags_Interested in other courses	-1.9434	0.390	-4.980	0.000	-2.708	-1.179
Tags_Lost to EINS	5.8981	0.608	9.695	0.000	4.706	7.091
Tags_Other tags	-2.4705	0.212	-11.657	0.000	-2.886	-2.055
Tags_Ringing	-3.4994	0.244	-14.331	0.000	-3.978	-3.021
Tags_Will revert after reading the email	4.3863	0.195	22.485	0.000	4.004	4.769
Last Notable Activity_Modified	-1.7140	0.127	-13.458	0.000	-1.964	-1.464

For model 2, the variable Total visits has comparatively high p value, thus, the variable will be dropped for more feasible model.

Model 3:

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.6480	0.123	-21.577	0.000	-2.888	-2.407
Page Views Per Visit	4.4565	0.250	17.840	0.000	3.967	4.946
Lead Origin_Lead Add Form	2.0481	0.431	4.756	0.000	1.204	2.892
Lead Source_Olark Chat	1.1783	0.145	8.140	0.000	0.895	1.462
Lead Source_Welingak Website	3.4791	0.849	4.099	0.000	1.816	5.143
Last Activity_SMS Sent	2.1258	0.116	18.296	0.000	1.898	2.354
at is your current occupation_Working Professional	1.0378	0.389	2.666	0.008	0.275	1.801
Tags_Closed by Horizzon	7.2521	1.022	7.095	0.000	5.249	9.255
Tags_Interested in other courses	-1.9529	0.391	-4.994	0.000	-2.719	-1.186
Tags_Lost to EINS	5.9188	0.610	9.708	0.000	4.724	7.114
Tags_Other tags	-2.4768	0.212	-11.685	0.000	-2.892	-2.061
Tags_Ringing	-3.4929	0.244	-14.303	0.000	-3.972	-3.014
Tags_Will revert after reading the email	4.3924	0.195	22.539	0.000	4.010	4.774
Last Notable Activity_Modified	-1.7346	0.127	-13.669	0.000	-1.983	-1.486

For model 3, the variables show decent p-value and all the variables have VIF < 5.

Model Accuracy

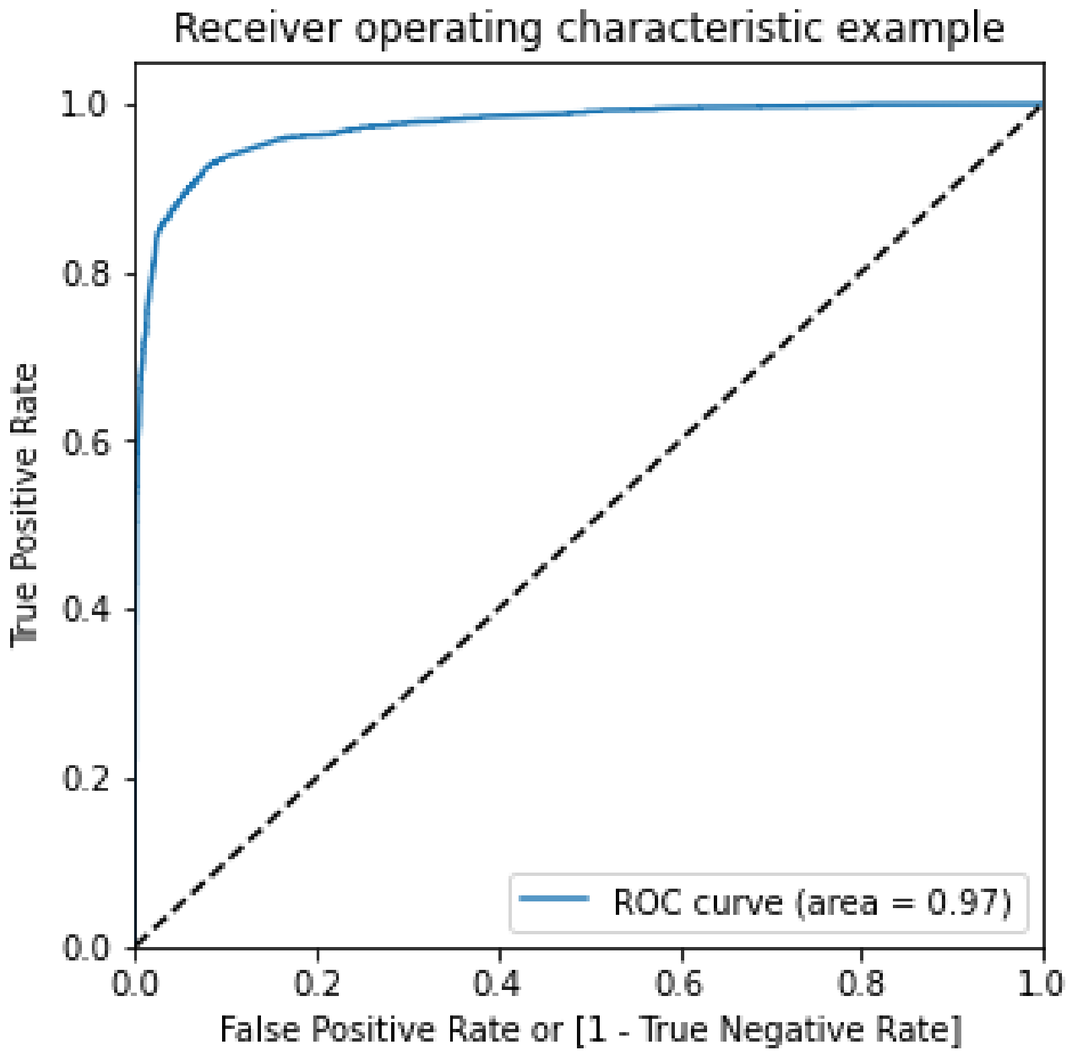
The area under the curve of the ROC is 0.97 which is quite good.

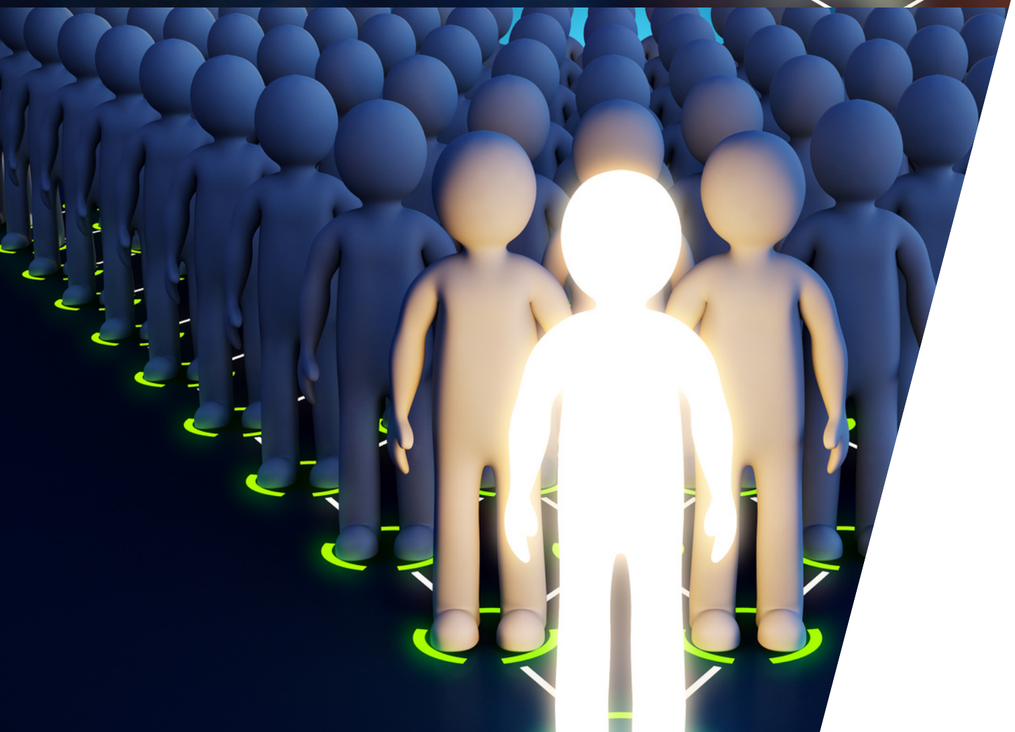
Train Data

- Accuracy: 92.27%
- Sensitivity: 91.41%
- Specificity: 92.78%

Test Data

- Accuracy: 92.31%
- Sensitivity: 88.16%
- Specificity: 94.93%





CONCLUSION:

The accuracy estimates for the model give promising scores in the test set as well as the training set.

Important features that showed a high correlation with the probability of leads conversion:

Last Notable Activity_Had a Phone Conversation

Lead Origin_Lead Add Form and

What is your current occupation_Working Professional