Lead Scoring Case Study

Sainath S & Archana Yadav

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%.

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 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%.

Approach

We have gone through the following steps

1. Data Cleaning

* Handled Unknown values 'Select' which were present in many categorical columns

* Dropped columns that are having more than 40% missing values

* Dropped categorical columns which are having more than 80% class imbalance

* Dropped Rows which are having less than 2% missing values

* Performed Outlier Analysis on the Numerical Columns and limited the data to 99th percentile as the data more than 99th percentile were less than 2%

* Performed Univariate analysis to check the distribution of data in Numerical and categorical columns

Approach (cont.)

2. Data Preparation

* created dummy variables for categorical columns

* performed 80:20 train:test split

* Performed Minmax feature scaling on numerical columns

3. Model Building and Evaluation

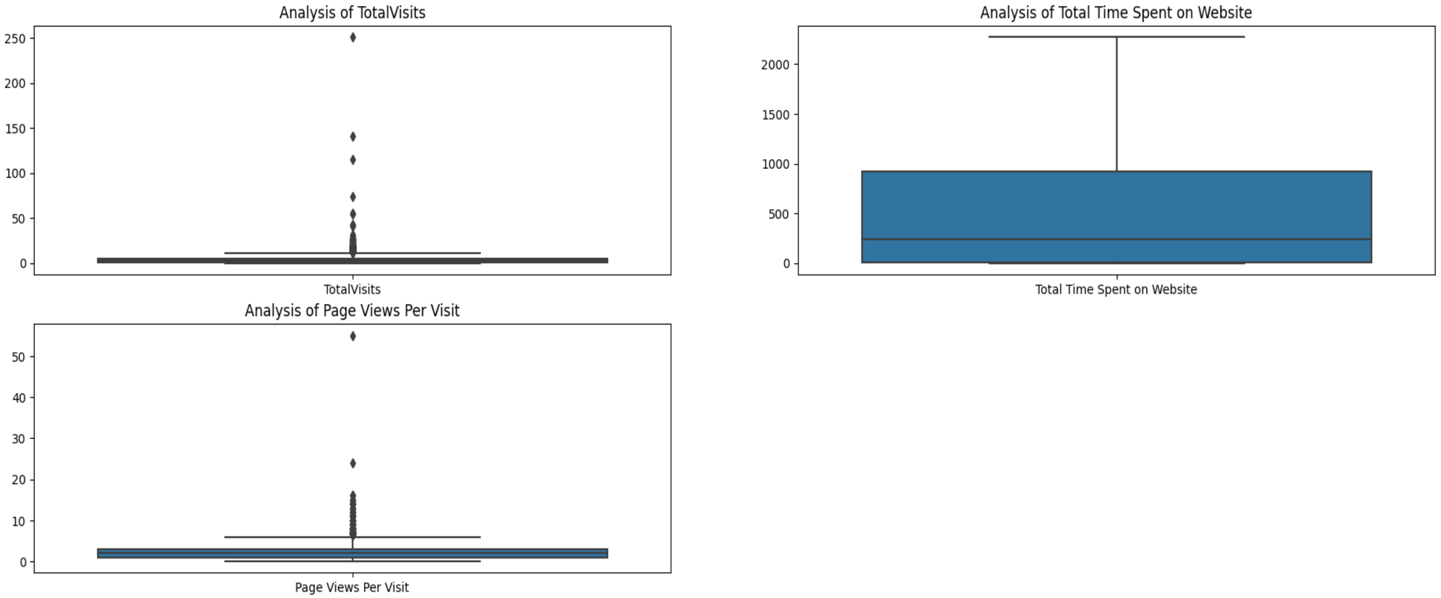
* Started off with GLM from statsmodel and used RFE to perform feature selection and ended up with 15 estimators which are having p-value of < 0.05 and VIF is around 5

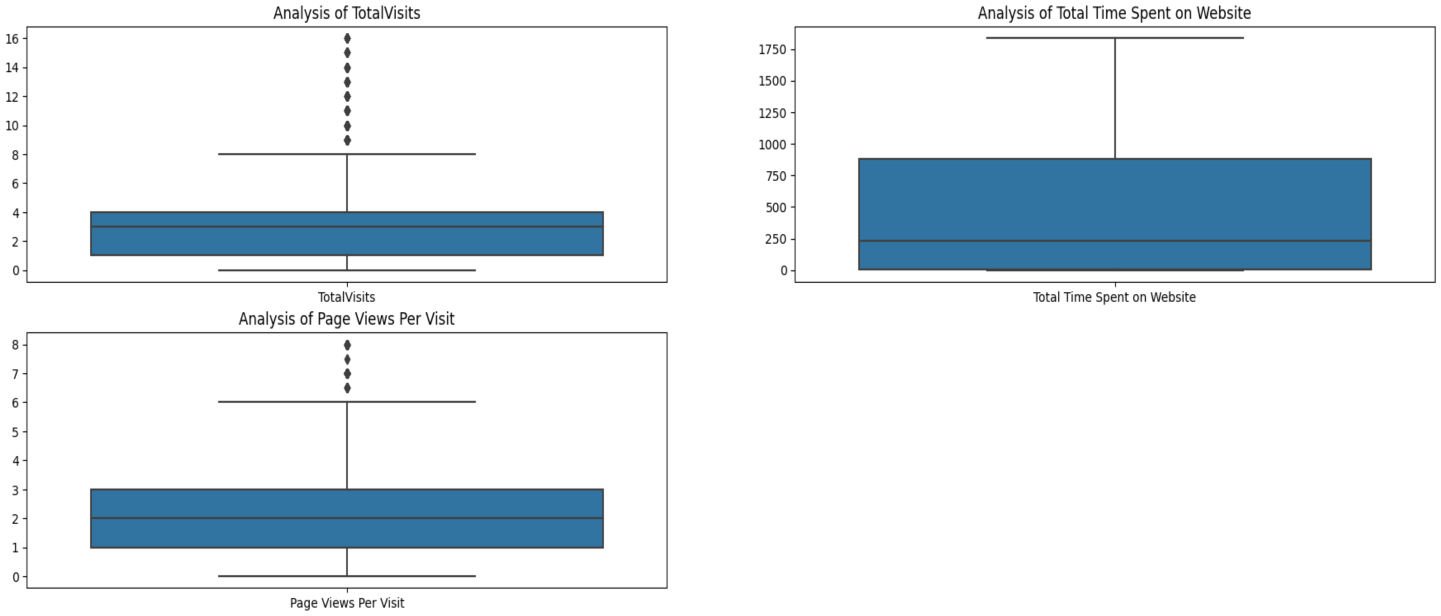
We have started building a logistic Regression Model starting with 60+ feature [after categorical variable conversion] and ended up 15 features with a Recall rate around 80%

Outlier Analysis:

From the box plot of the numerical features we can there is huge difference in value between 99th percent and max value and rows which were having these values contributed to less then 2% rows in dataset.

So instead of imputing , we have dropped those from dataset.





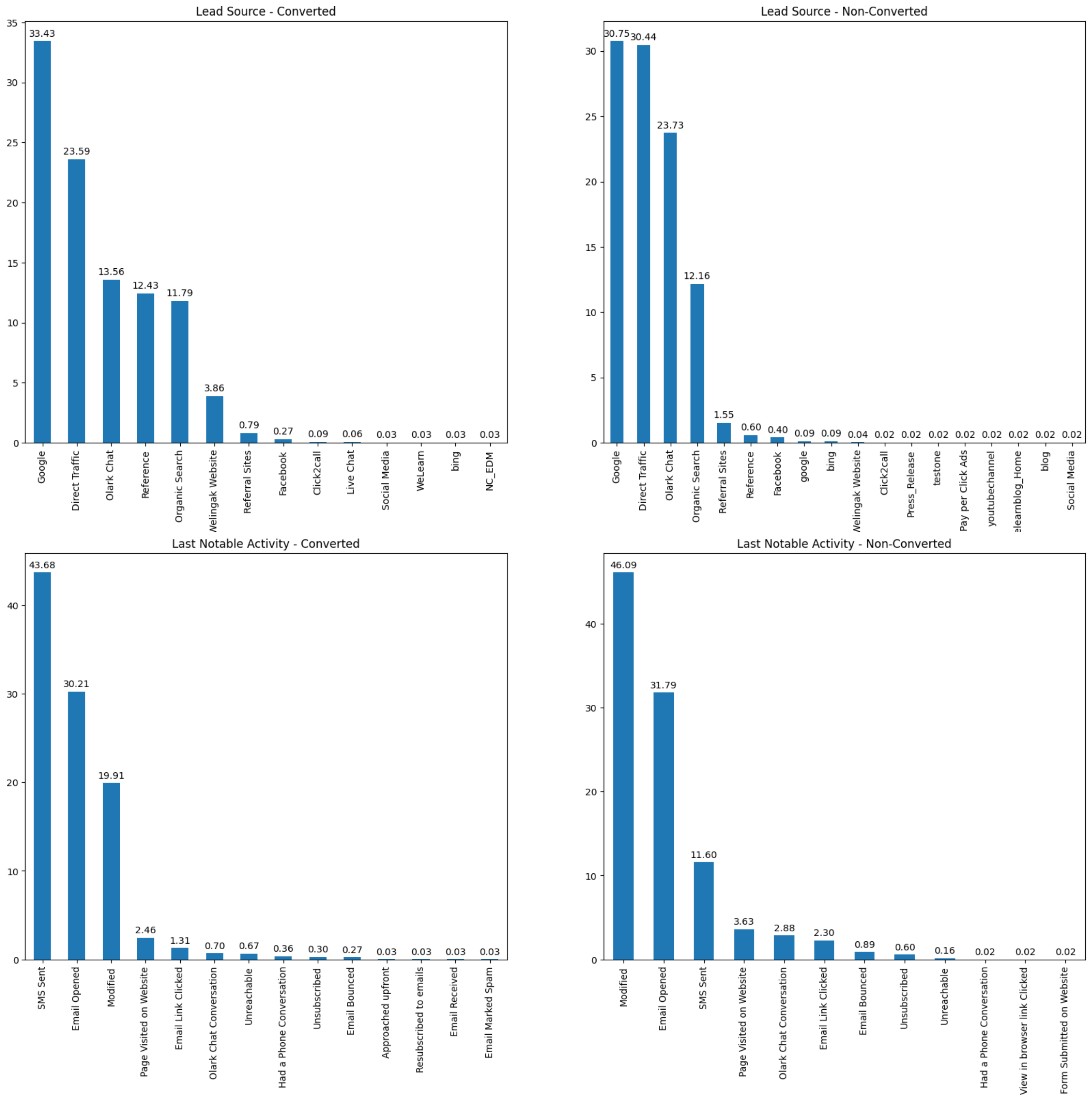
Univariate Analysis:

For lead Source:

Direct Traffic and Olark Chat seems to be less effective in conversion Rate

For Last Notable Activity:

Sending SMS Text[Follow-ups] , user opening the emails seems to be contribution to positive to the conversion rate



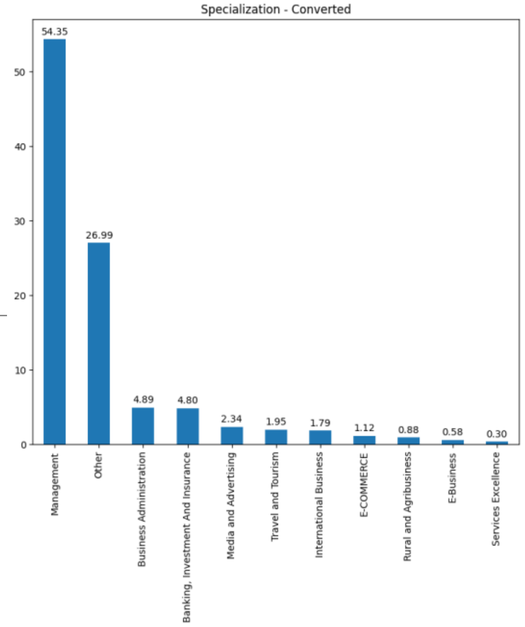
Univariate Analysis:

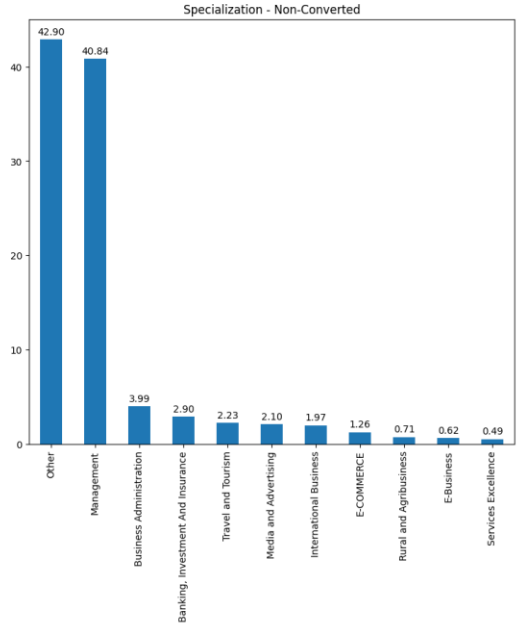
For Specialization:

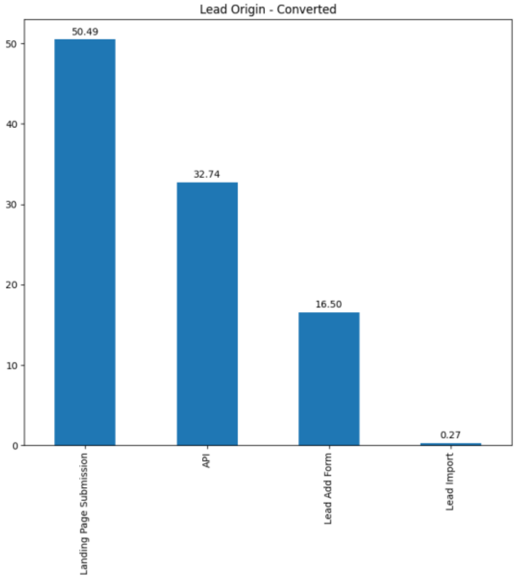
If the customer has not selected the specialization , then the chances of non-conversions is higher

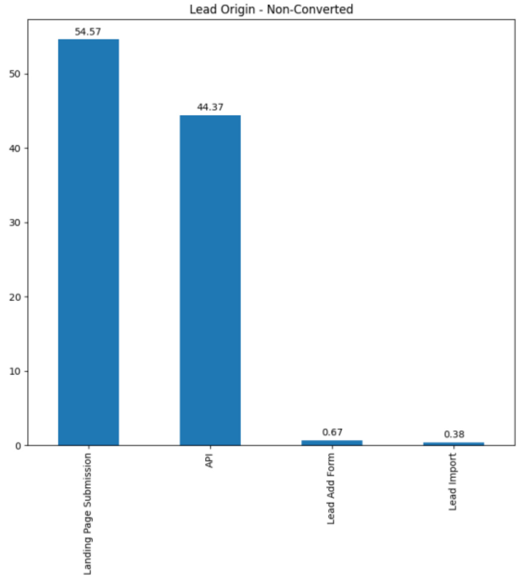
For Lead Origin:

There is not much difference in the conversion rate


Specialization - Converted


Specialization - Non-Converted


Lead Origin - Converted


Lead Origin - Non-Converted

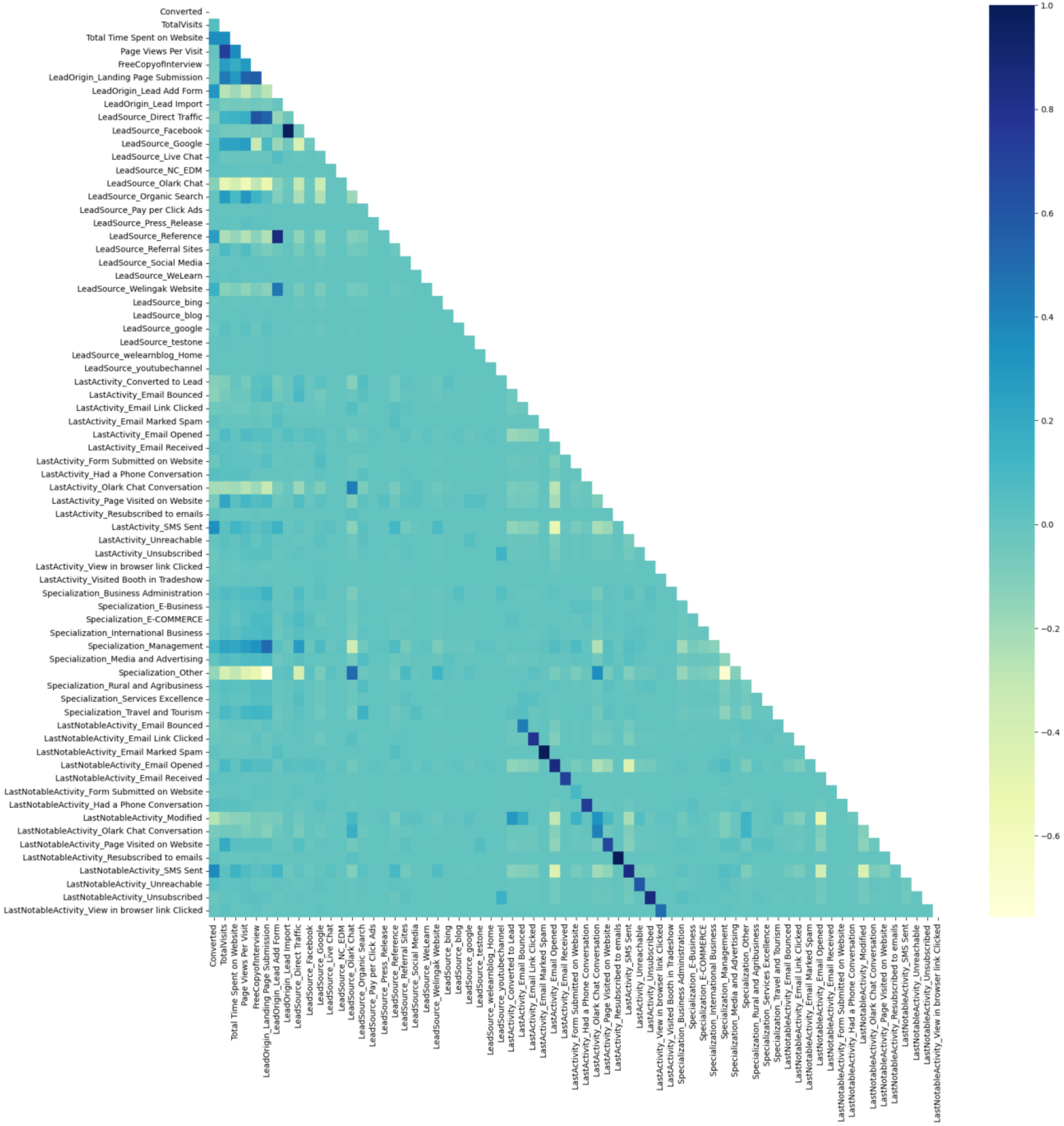
Multivariate Analysis:

Heatmap of correlation between features in the dataset.

There is high correlation [> 0.95] between the dummy variables

'LeadSource_Facebook', 'LastNotableActivity_Email Marked Spam', 'LastNotableActivity_Resubscribed to emails’

We have dropped them before Model Building



Model Building:

Using RFE we have identified the select the feature variables and using GLM from stats model we were able to identify the effect of each feature on the conversion score.

* Top 3 features which has a ***postive effect*** on the outcome is

* Total Time Spent on Website - coefficient 3.7857

* LeadSource_Welingak Website - coefficient 2.7793

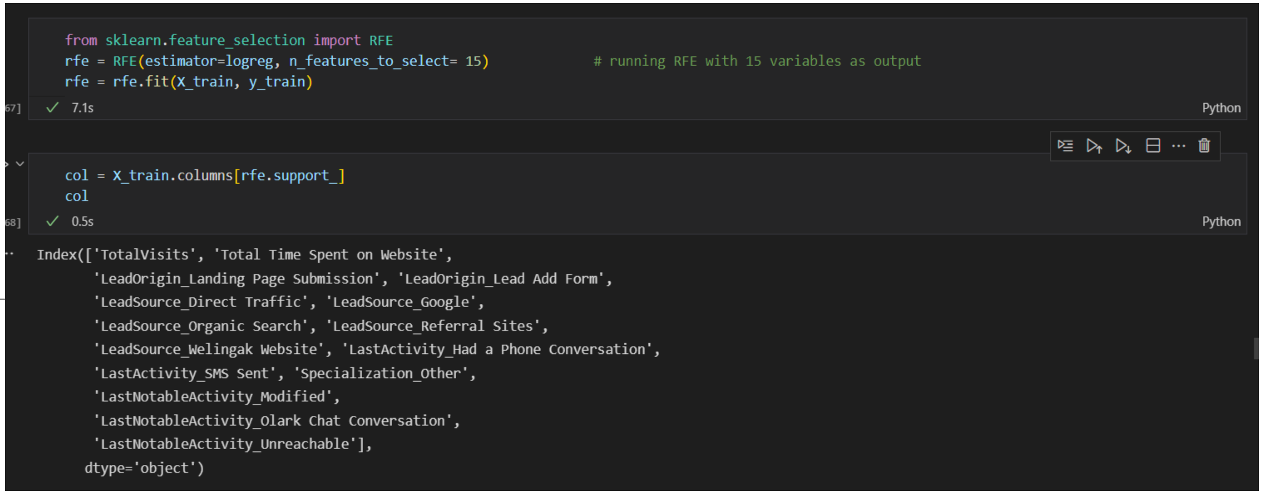
* LastActivity_Had a Phone Conversation - coefficient 2.6726

* Top 3 features which has a ***negative effect*** on the outcome is

* LeadSource_Referral Sites - coefficient -1.4605

* LeadSource_Organic Search - coefficient -1.4441

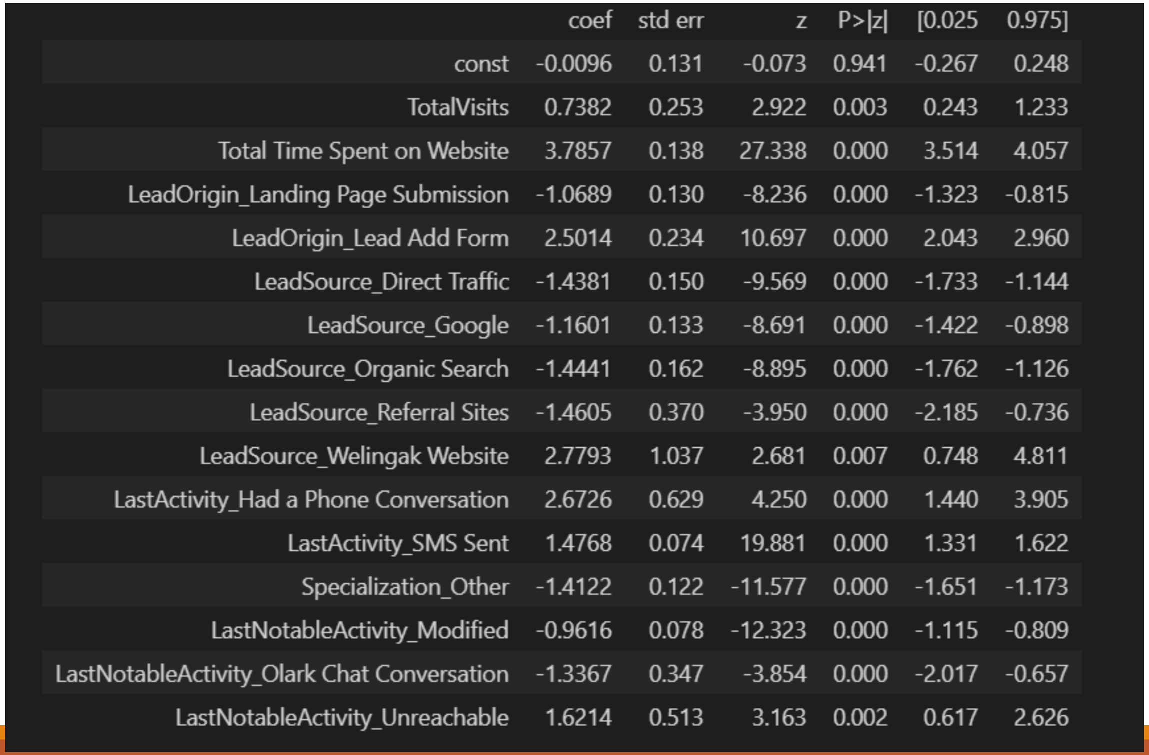
* LeadSource_Direct Traffic - coefficient -1.4381



```
from sklearn.feature_selection import RFE
rfe = RFE(estimator=logreg, n_features_to_select= 15)            # running RFE with 15 variables as output
rfe = rfe.fit(X_train, y_train)
```

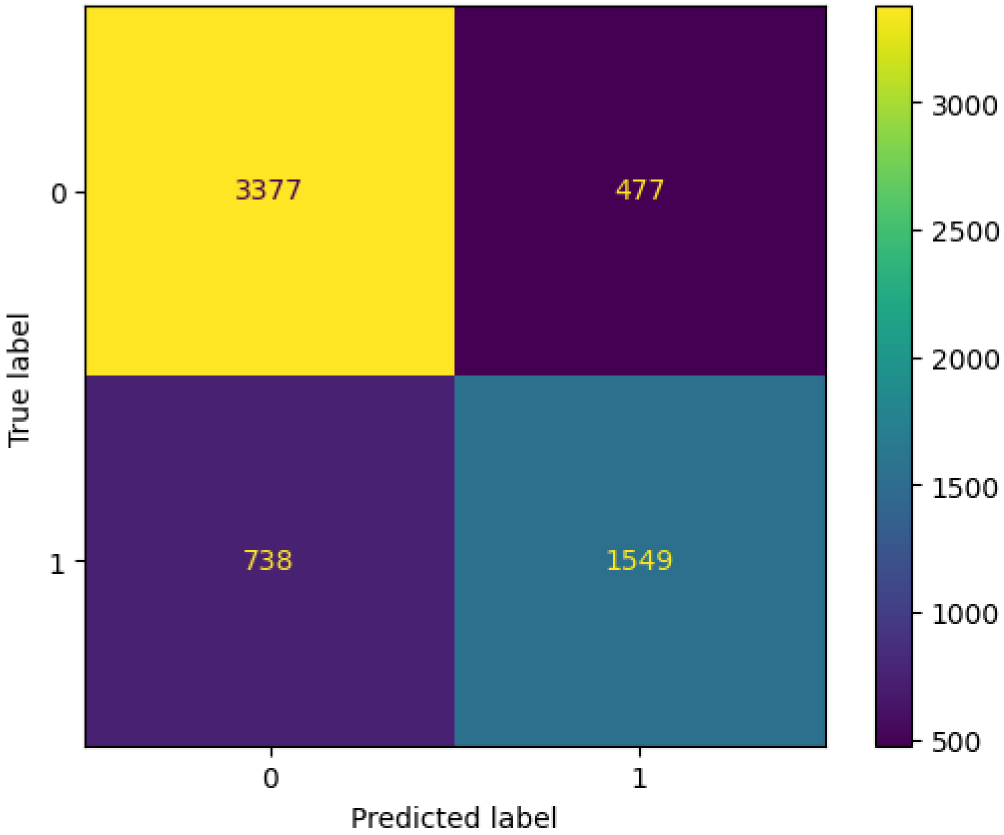
```
col = X_train.columns[rfe.support_]
col
```

```
Index(['TotalVisits', 'Total Time Spent on Website',
       'LeadOrigin_Landing Page Submission', 'LeadOrigin_Lead Add Form',
       'LeadSource_Direct Traffic', 'LeadSource_Google',
       'LeadSource_Organic Search', 'LeadSource_Referral Sites',
       'LeadSource_Welingak Website', 'LastActivity_Had a Phone Conversation',
       'LastActivity_SMS Sent', 'Specialization_Other',
       'LastNotableActivity_Modified',
       'LastNotableActivity_Olark Chat Conversation',
       'LastNotableActivity_Unreachable'],
      dtype='object')
```

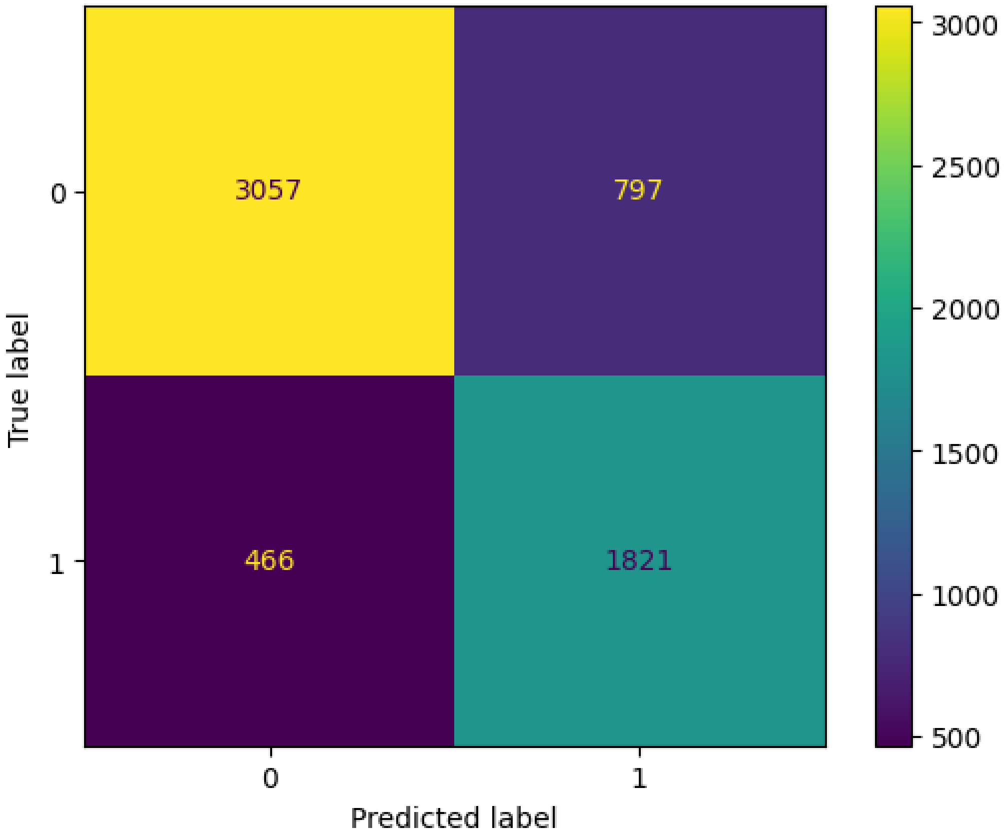


	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0096	0.131	-0.073	0.941	-0.267	0.248
TotalVisits	0.7382	0.253	2.922	0.003	0.243	1.233
Total Time Spent on Website	3.7857	0.138	27.338	0.000	3.514	4.057
LeadOrigin_Landing Page Submission	-1.0689	0.130	-8.236	0.000	-1.323	-0.815
LeadOrigin_Lead Add Form	2.5014	0.234	10.697	0.000	2.043	2.960
LeadSource_Direct Traffic	-1.4381	0.150	-9.569	0.000	-1.733	-1.144
LeadSource_Google	-1.1601	0.133	-8.691	0.000	-1.422	-0.898
LeadSource_Organic Search	-1.4441	0.162	-8.895	0.000	-1.762	-1.126
LeadSource_Referral Sites	-1.4605	0.370	-3.950	0.000	-2.185	-0.736
LeadSource_Welingak Website	2.7793	1.037	2.681	0.007	0.748	4.811
LastActivity_Had a Phone Conversation	2.6726	0.629	4.250	0.000	1.440	3.905
LastActivity_SMS Sent	1.4768	0.074	19.881	0.000	1.331	1.622
Specialization_Other	-1.4122	0.122	-11.577	0.000	-1.651	-1.173
LastNotableActivity_Modified	-0.9616	0.078	-12.323	0.000	-1.115	-0.809
LastNotableActivity_Olark Chat Conversation	-1.3367	0.347	-3.854	0.000	-2.017	-0.657
LastNotableActivity_Unreachable	1.6214	0.513	3.163	0.002	0.617	2.626

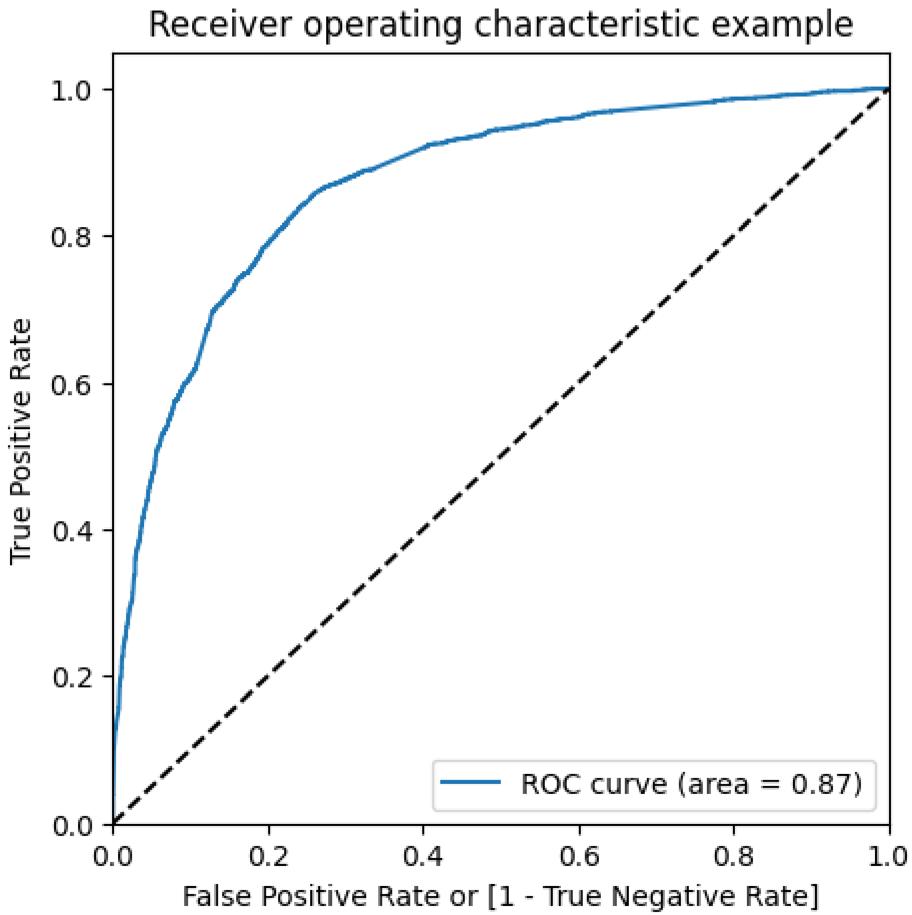
Confusion Matrix on Train Set Prediction when threshold is 0.5



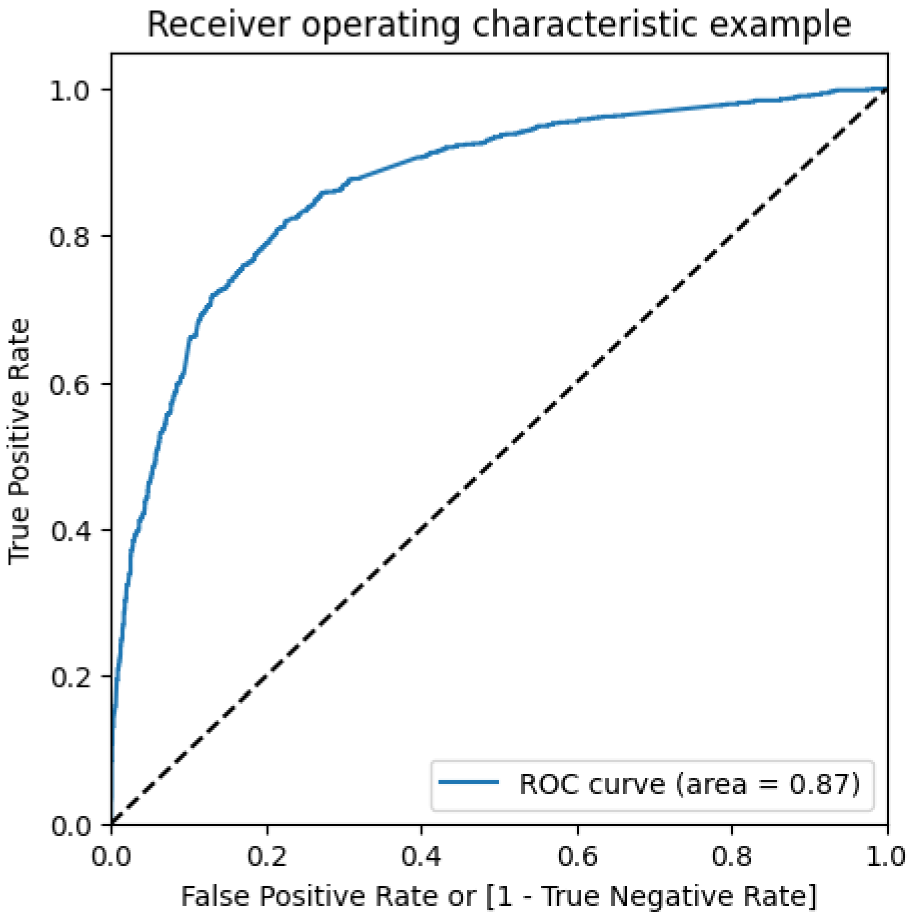
Confusion Matrix on Train Set Prediction when threshold is 0.35.



ROC Curve on the Training to show the Model Performance



ROC Curve on the Test to show the Model Performance



Conclusion

1. We got 15 features towards end of the training, Out of which

- * Top 3 features which has a _*postive effect*_ on the outcome is
    - Total Time Spent on Website - coefficient 3.7857
    - LeadSource_Welingak Website - coefficient 2.7793
    - LastActivity_Had a Phone Conversation - coefficient 2.6726
- Top 3 features which has a _*negative effect*_ on the outcome is
    - LeadSource_Referral Sites - coefficient -1.4605
    - LeadSource_Organic Search - coefficient -1.4441
    - LeadSource_Direct Traffic - coefficient -1.4381

2. Using the ROC curve and brute force calcuation of accuracy,sensitivit and specificity we have identified the threshold to be __**0.35**__ for categorizing the leads as converted and not-converted

3. From the model we can see that having phone conversation AND direct lead add form has a positive effect on the conversion .

Conclusion

3. Metric from the Test set are as below

```
 * Accuracy                        : 0.7978723404255319
 * Sensitivity                     : 0.7786640079760718
 * Specificty                      : 0.8096992019643954
 * FalsePositiveRate               : 0.19030079803560468
 * PositivePredictiveValue         : 0.7158570119156737
 * NegativePredictiveValue         : 0.8559377027903958
 * F1 Score                        : 0.7459407831900668
```

3. Metrics from the Train set are as below

```
 *  Accuracy                       : 0.794333170493405
 *  Sensitivity                    : 0.7962396152164407
 *  Specificty                     : 0.7932018681888947
 *  FalsePositiveRate              : 0.20679813181110535
 *  PositivePredictiveValue        : 0.6955691367456074
 *  NegativePredictiveValue        : 0.867726369571388
 *  F1 Score                       : 0.742507645259939
```