Lead Score - Case Study

Submitted By:
Jaidip Ghosh
Shubham Kaser

Data Explorarion

- Dimension of the Dataset is **(9240, 37)**
- Out of **37** columns, **17** columns have null values
- Dataset has **7 numeric** and **30 categorical** variables

Problem Statement: 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.

Data Cleaning and Manipulation

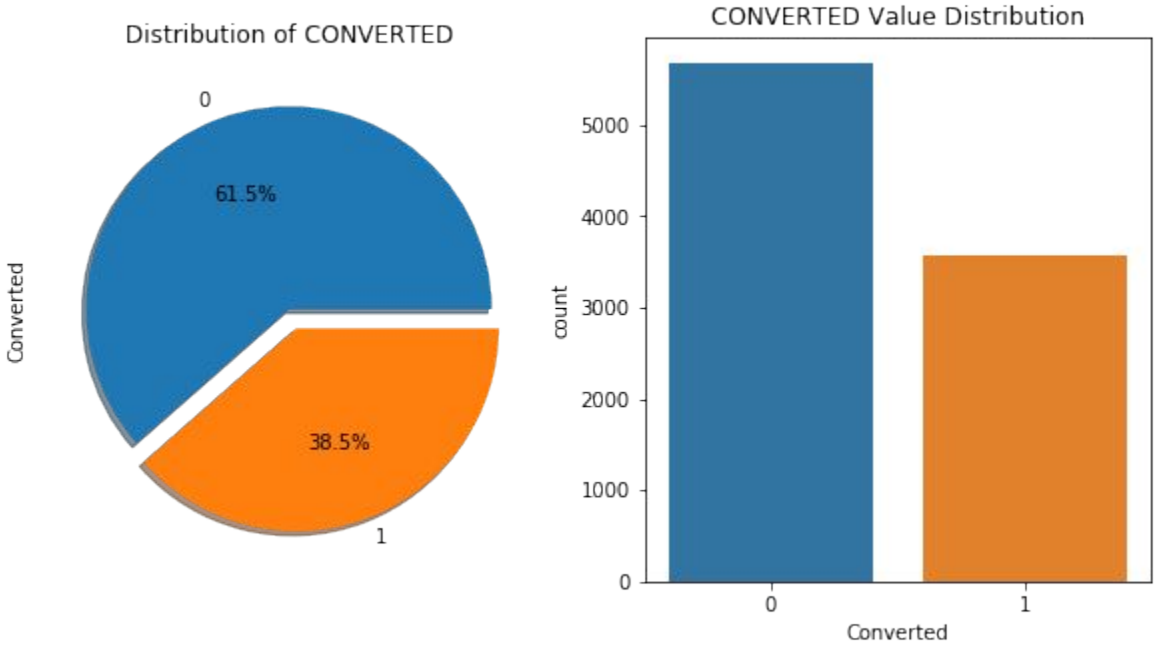
Possible Data inconsistencies:

- NaN values in the dataset
- Unacceptable number of outliers

Other Issues:

- Many numerical variables needs to be converted to categorical variables
- Many variables have a lot of categories that needs to be segregated as single category
- Same category has two different names e.g., ‘Google’ & ‘google’
- ‘Select’ category in various variables has to be treated as NaN.

Data Imbalance of the Target Variable



Variables with missing values (in %)

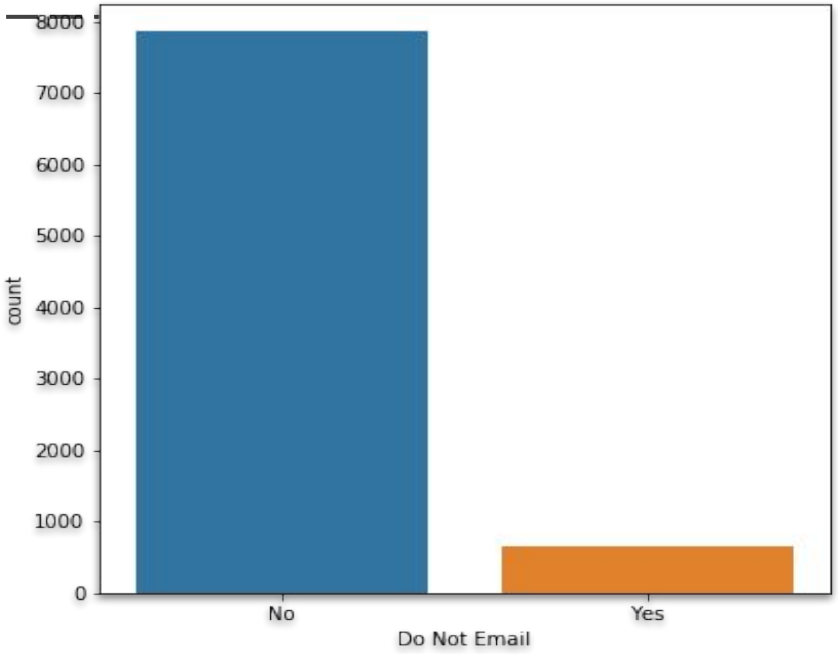
	Null Values
Lead Quality	51.59
Asymmetrique Activity Index	45.65
Asymmetrique Profile Score	45.65
Asymmetrique Activity Score	45.65
Asymmetrique Profile Index	45.65
Tags	36.29
Lead Profile	29.32
What matters most to you in choosing a course	29.32
What is your current occupation	29.11
Country	26.63
How did you hear about X Education	23.89
Specialization	15.56
City	15.37
Page Views Per Visit	1.48
TotalVisits	1.48
Last Activity	1.11
Lead Source	0.39

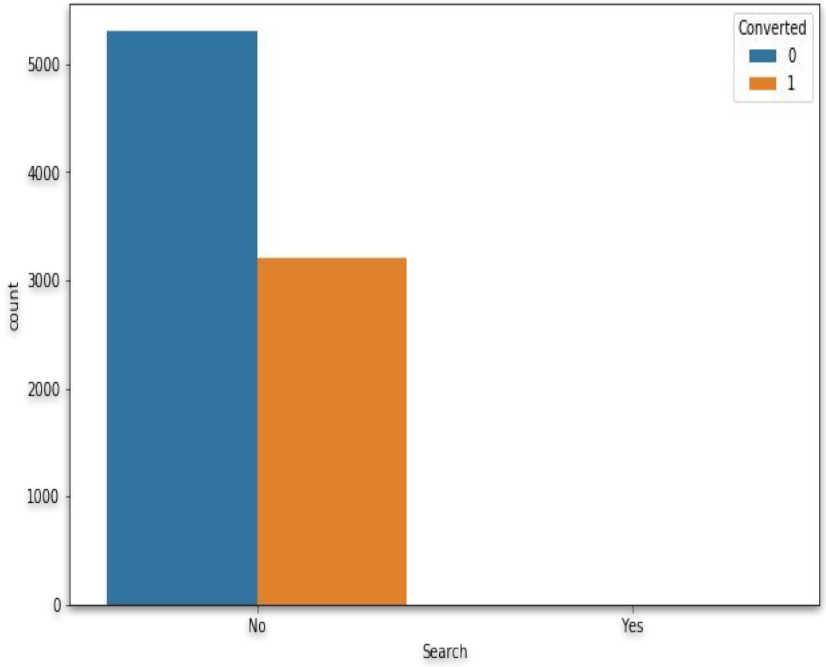
Which columns were dropped?

- As size of the dataset was limited, Columns with more than 30% of missing values with no scope of imputation were dropped:

	Null Values
Lead Quality	51.59
Asymmetrique Activity Index	45.65
Asymmetrique Profile Score	45.65
Asymmetrique Activity Score	45.65
Asymmetrique Profile Index	45.65
Tags	36.29

- Columns that were highly imbalanced were also dropped





```
No    8513
Name: Newspaper Article, dtype: int64
```

```
No    8513
Name: X Education Forums, dtype: int64
```

```
No     8512
Yes       1
Name: Newspaper, dtype: int64
```

```
No     8511
Yes       2
Name: Digital Advertisement, dtype: int64
```

```
No     8509
Yes       4
Name: Through Recommendations, dtype: int64
```

```
No    8513
Name: Receive More Updates About Our Courses, dtype: int64
```

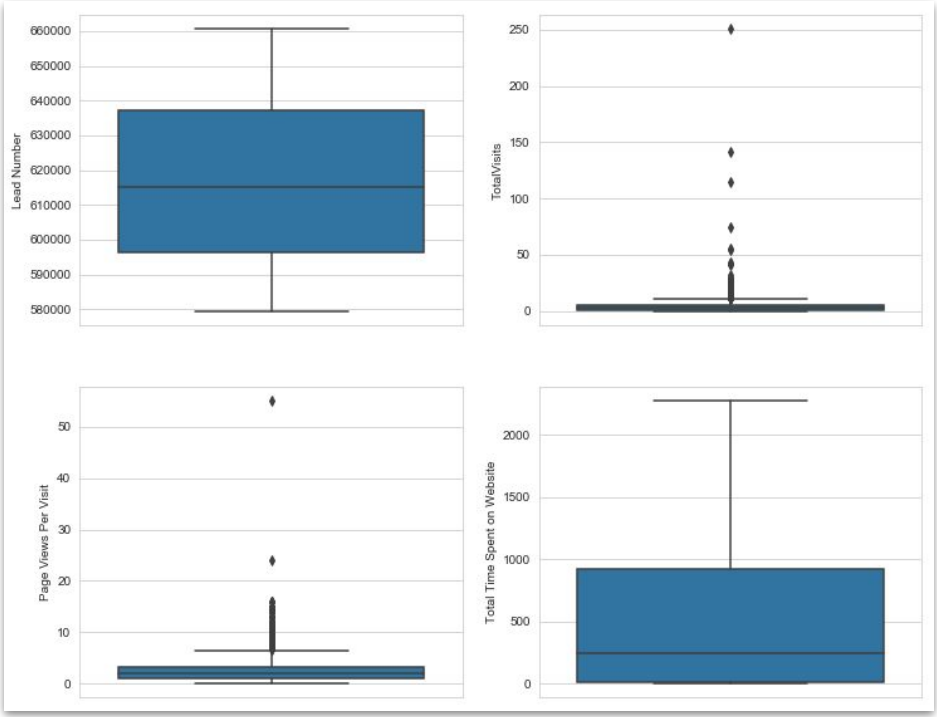
```
No    8513
Name: Update me on Supply Chain Content, dtype: int64
```

```
No    8513
Name: Get updates on DM Content, dtype: int64
```

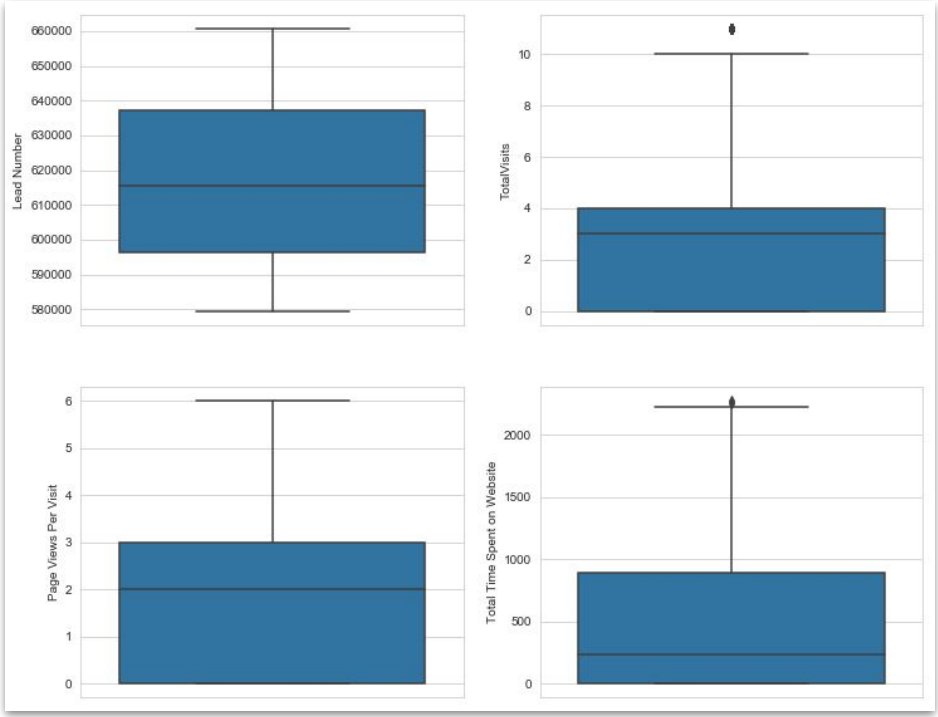
```
No    8513
Name: I agree to pay the amount through cheque, dtype: int64
```

These columns turned out to be contributing towards single category making the variable highly imbalanced.

Outlier Analysis



Before Removing Outliers

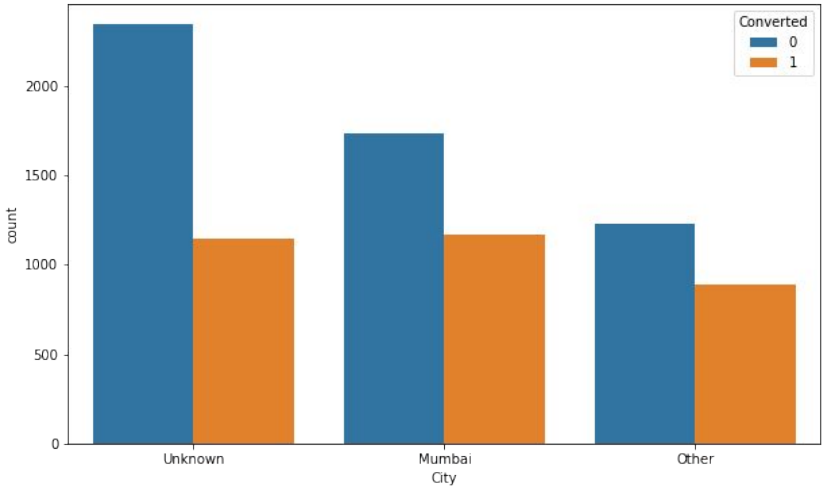


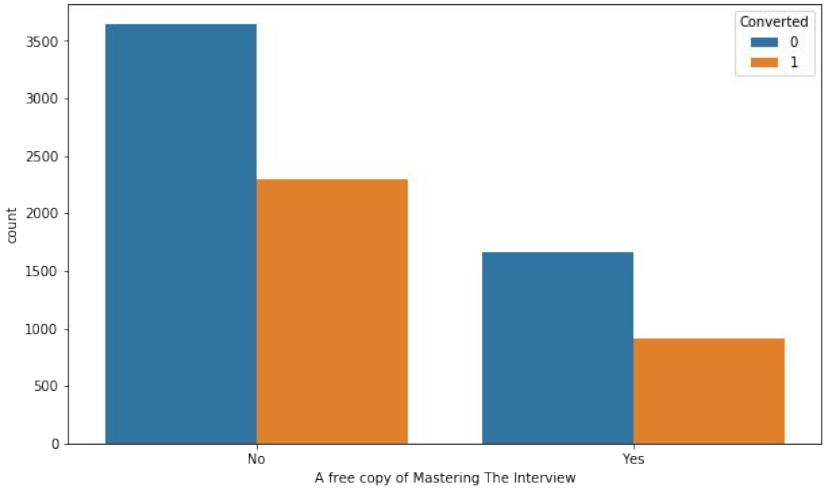
After Removing Outliers

Univariate Analysis

Inference:
1. We can see 'Mumbai' and 'Other' has high and similar conversion rate.

2. Leads with 'Unknown' cities are comparatively less likely to converted but have significant rate of conversion





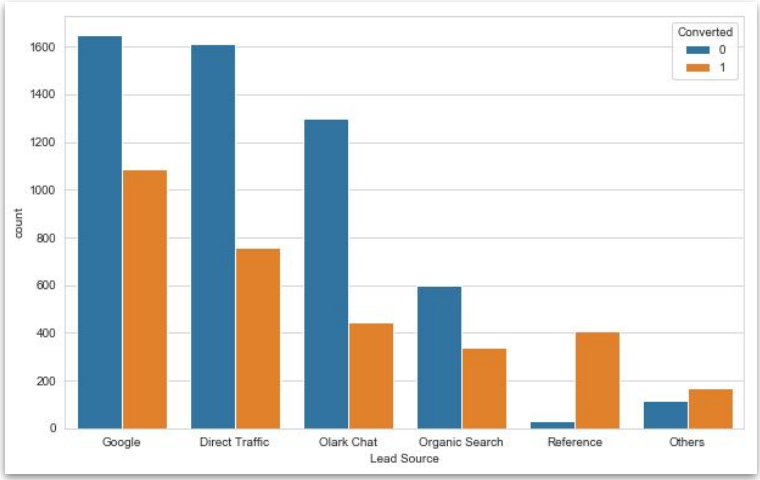
Inference:
1. There are almost similar rate of conversion for both the categories

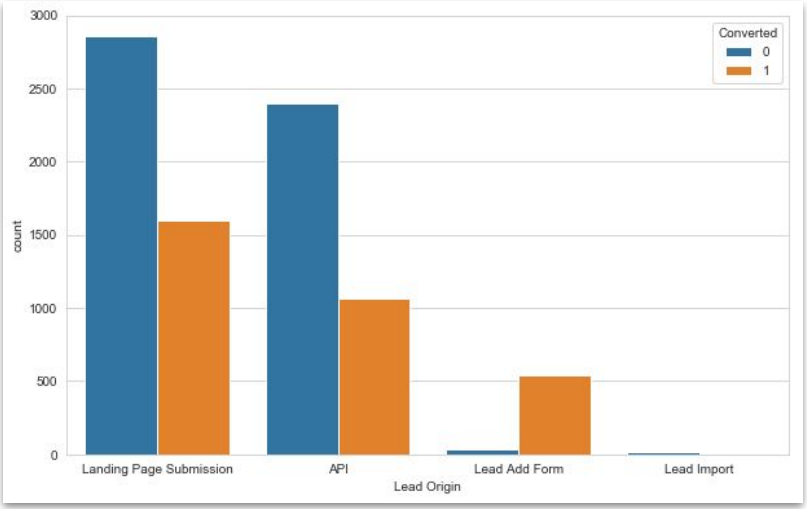
2. Most of the leads don't opt for free copy of mastering the Interview

Univariate Analysis

Inference:

1. Sorces like 'Olark Chat', 'Organic Search', 'Direct Traffic', 'Google' brings most of the leads with significant conversion rate of around 30% -60% with 'Google' bringing the most conversion.
2. Leads from the source 'Reference' and 'Others' seems to bring the leads that are only to be converted.

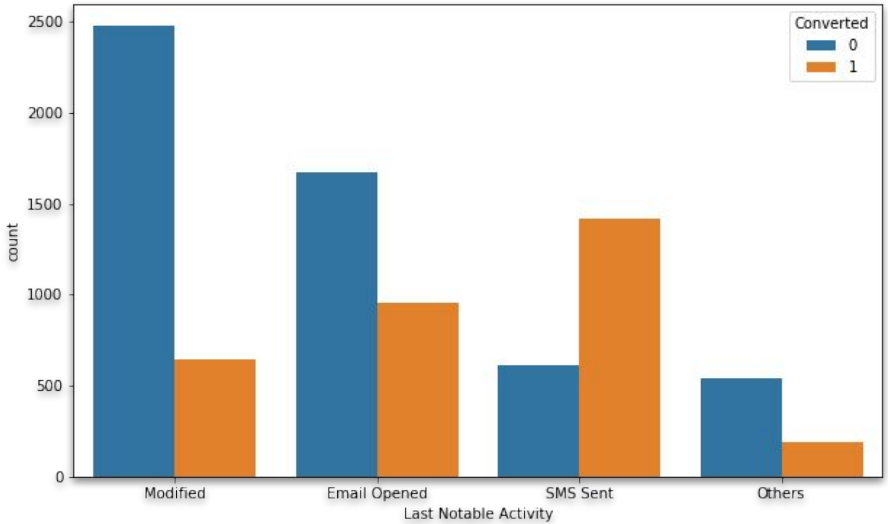


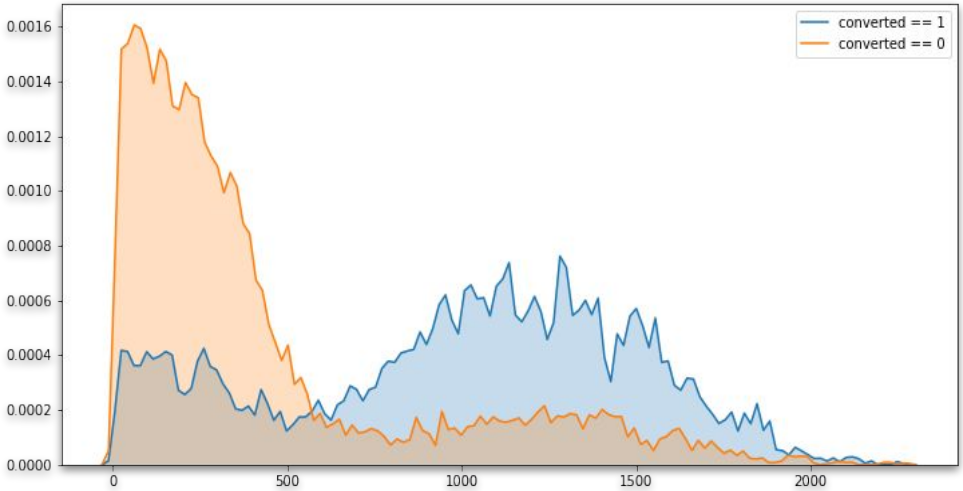


Inference:

1. Origins that bring most of the leads are 'API' and 'Landing Page Submission' with the conversion rate around 40% - 50%
2. From the origin 'Lead Add Form' it is most likely the lead to be converted.

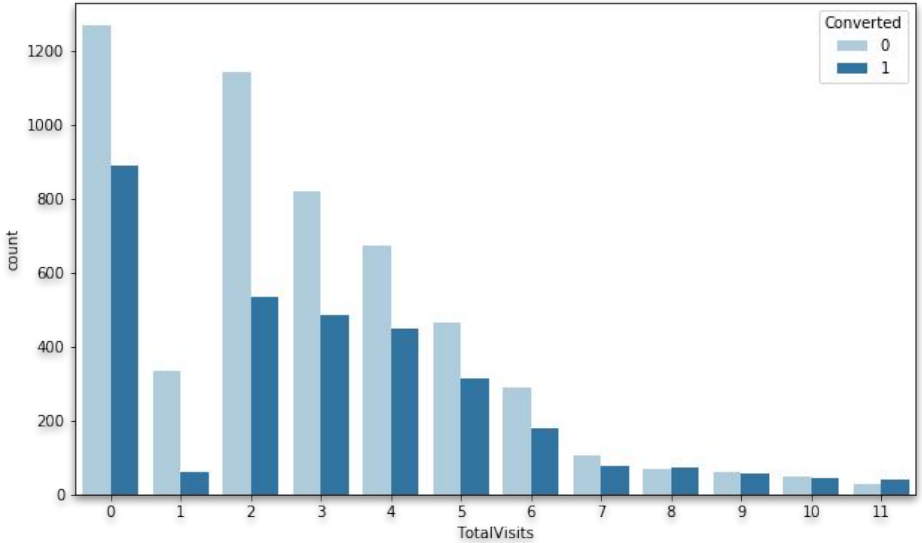
Inference:
1. For the category 'SMS Sent', gets the highest conversion.
2. Coversion rate for 'Modified' is comparitively low but has significant number of lead counts.

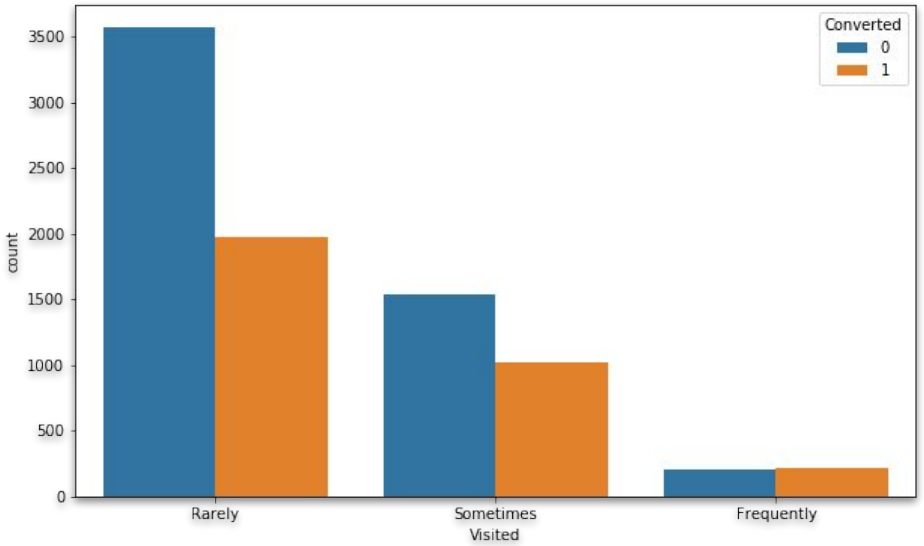


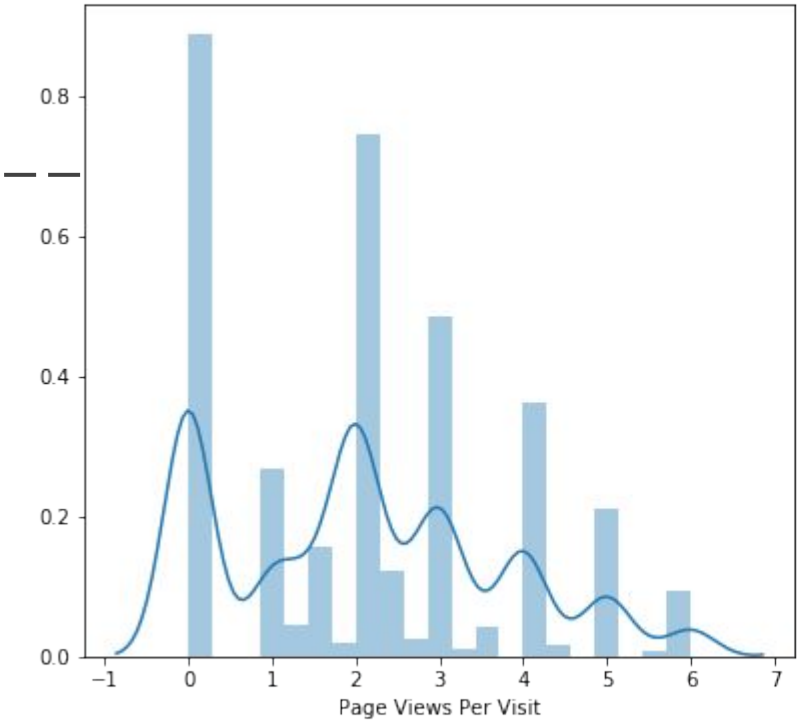


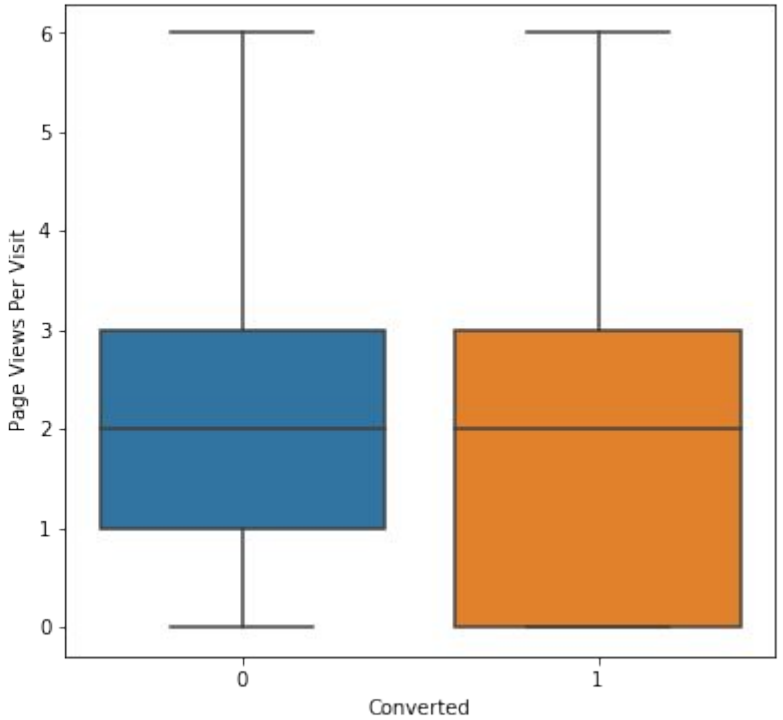
Inference:
1. Leads who spend more than 500, are more likely to get converted.
2. Leads spending less than 500 seems to be converted very less

Variables that were changed from Numerical to Categorical









Inference: 1. It is more of a continuous variable for the average.
2. Most of the leads that are Not Converted, lies between 1-3 and the leads that are converted lies between 0-3.

Dummy Variables and Train Test Split

- Shape of Dataframe before dummies: (8513, 13)
  - ‘Prospect ID’ and ‘Lead Number’ were moved to separate Dataframe.
- Shape of Dataframe after dummies: (8513, 67)
- Shape of X_train: (5959, 66)
- Shape of y_trian: (5959,)
- Shape of X_test: (2554, 66)
- Shape of y_test: (2554,)

Correlation Heatmap

Inference:
As we can see in the corners, there are negative as well as positive correlations among the variables.

Logistic Regression (Model - 1)

The model was made without removing any feature

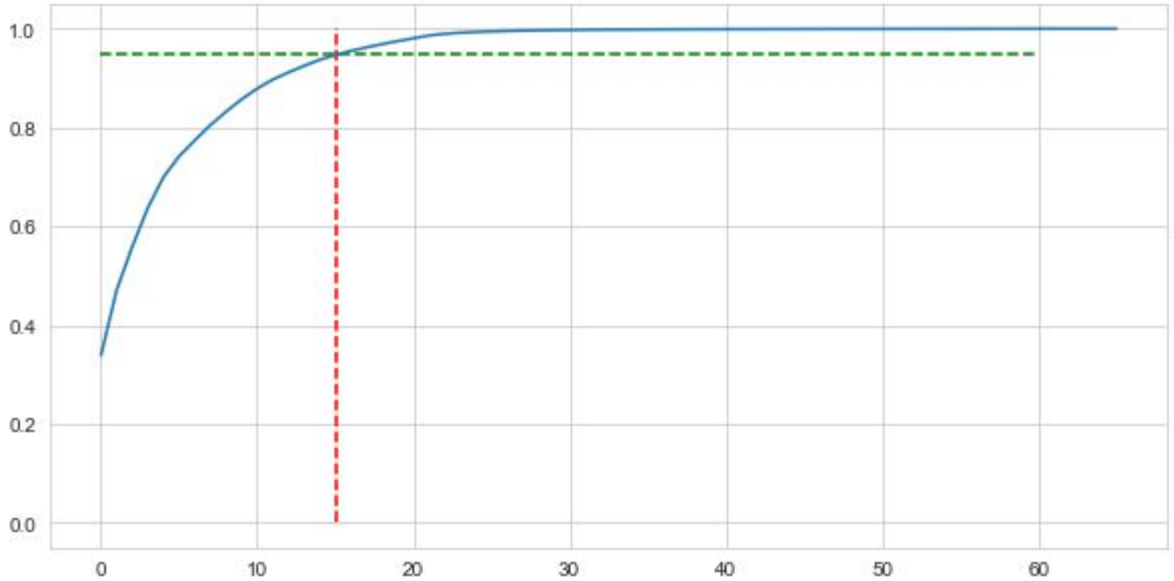
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	5959
Model:	GLM	Df Residuals:	5895
Model Family:	Binomial	Df Model:	63
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2406.3
Date:	Thu, 14 Nov 2019	Deviance:	4812.6
Time:	13:57:04	Pearson chi2:	6.71e+03
No. Iterations:	22	Covariance Type:	nonrobust

1. We saw, for most of the variables, p value is very high or 1.
2. Correlation among the some of the variables are high and negative.
3. Data seems to be linear

Accuracy Score	0.817

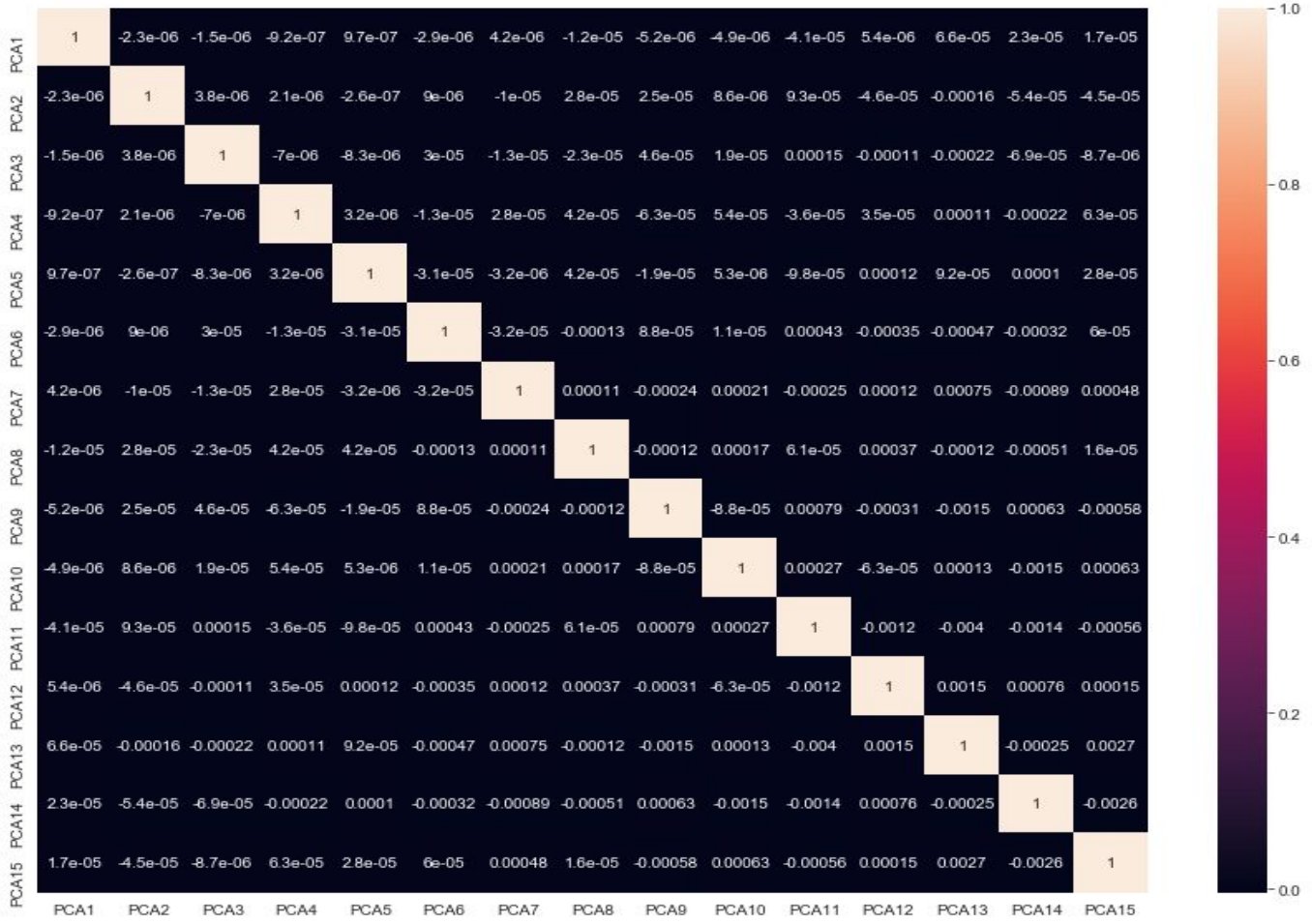
Dimensionality Reduction Using PCA



From the above plot, it can be concluded that 15 components will be able to explain 95% of the variance of the data.

Visualising Correlation among the obtained Variable after PCA

The heatmap shows there are zero correlation among the variables obtained after PCA.



Logistic Regression (Model - 3)

On model 2, variables with p-value > 0.05 were removed.

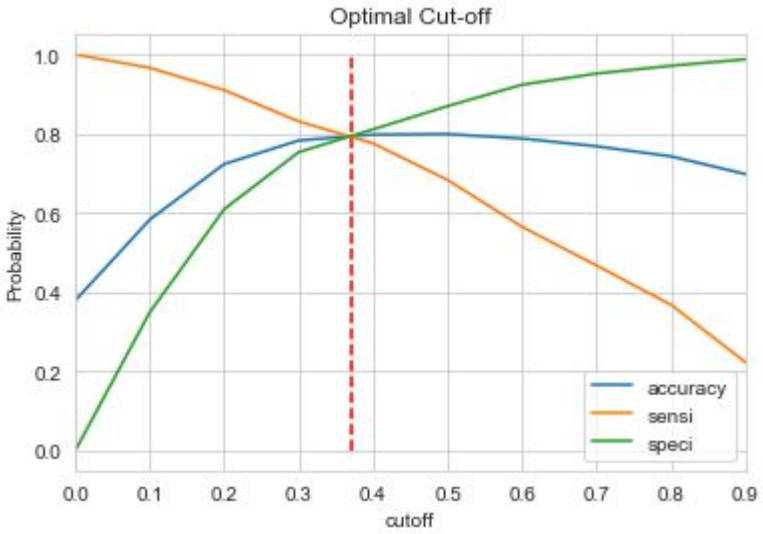
The model -3 when fit on the PCA reduced dimensions, it gave the summary as above where we can see for every variables obtained has value of p-value < 0.05.

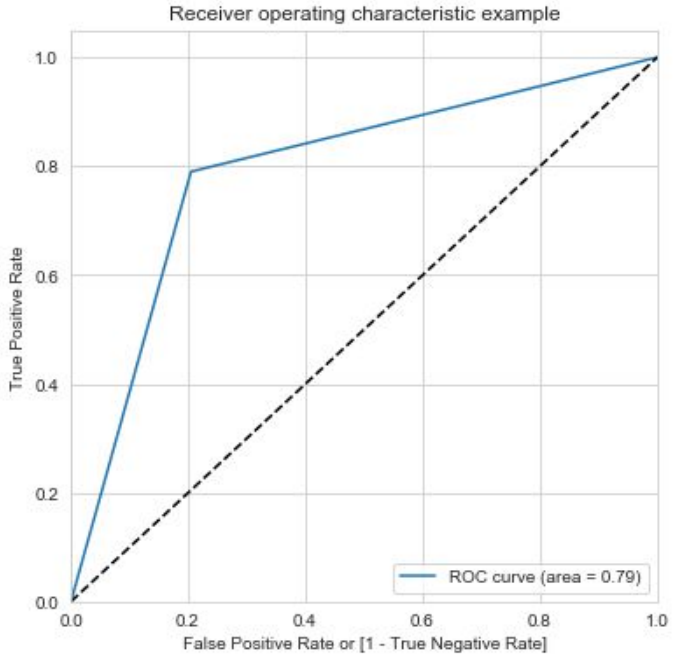
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	5959
Model:	GLM	Df Residuals:	5944
Model Family:	Binomial	Df Model:	14
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2610.5
Date:	Thu, 14 Nov 2019	Deviance:	5221.0
Time:	14:21:41	Pearson chi2:	6.18e+03
No. Iterations:	6	Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.6249	0.036	-17.166	0.000	-0.696	-0.554
x1	0.3489	0.027	12.843	0.000	0.296	0.402
x2	1.3168	0.046	28.426	0.000	1.226	1.408
x3	1.2221	0.051	24.100	0.000	1.123	1.322
x4	0.3941	0.054	7.332	0.000	0.289	0.499
x5	-1.3152	0.068	-19.295	0.000	-1.449	-1.182
x6	0.2177	0.073	2.976	0.003	0.074	0.361
x7	-1.7994	0.103	-17.460	0.000	-2.001	-1.597
x8	0.8478	0.094	9.040	0.000	0.664	1.032
x9	0.7951	0.093	8.554	0.000	0.613	0.977
x10	-0.3603	0.102	-3.515	0.000	-0.561	-0.159
x11	0.3105	0.102	3.050	0.002	0.111	0.510
x12	-0.3108	0.126	-2.469	0.014	-0.557	-0.064
x13	-1.0108	0.133	-7.579	0.000	-1.272	-0.749
x14	-0.3767	0.149	-2.527	0.011	-0.669	-0.085

Optimal Cut-Off, ROC Curve and Accuracy





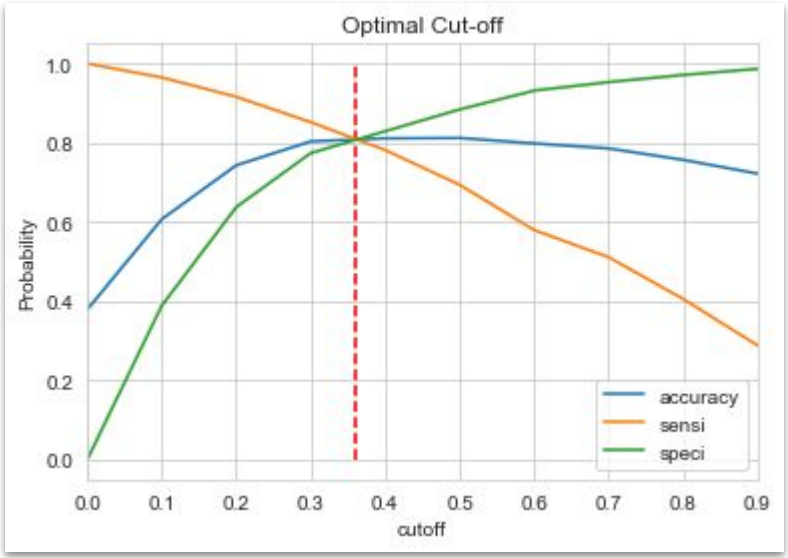
Accuracy Score	0.793

Model Building after Applying RFE

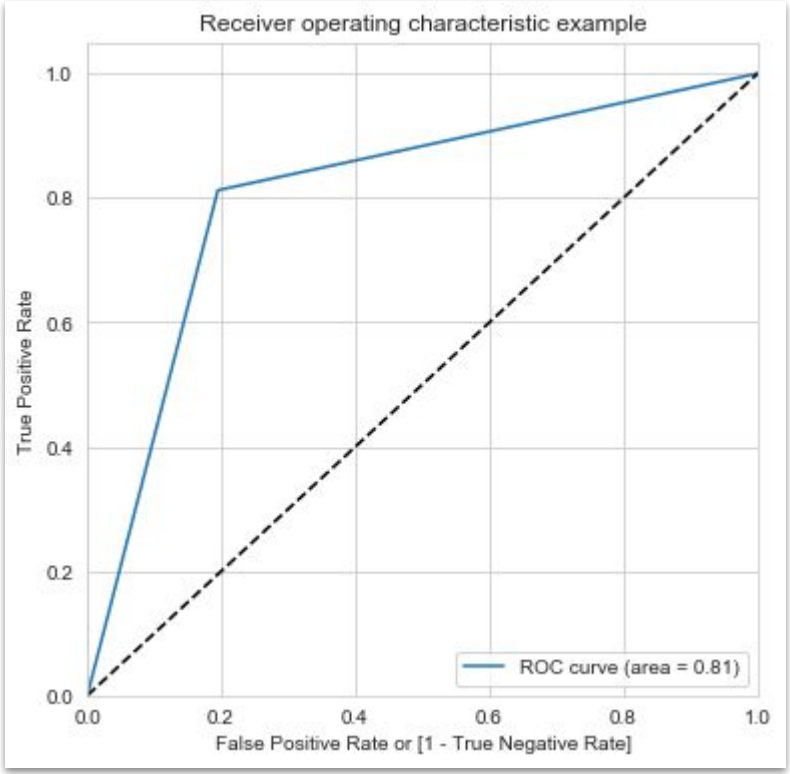
- From the 15 variables after RFE, several models were built and features with p-value > 0.05 and vif > 5 were removed.
- The final model that we obtained was from 10 variables.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5385	0.134	-4.021	0.000	-0.801	-0.276
Total Time Spent on Website	1.0634	0.041	25.814	0.000	0.983	1.144
Lead Origin_Landing Page Submission	-1.0489	0.139	-7.563	0.000	-1.321	-0.777
Lead Origin_Lead Add Form	3.9615	0.219	18.050	0.000	3.531	4.392
Lead Source_Olark Chat	1.2072	0.124	9.729	0.000	0.964	1.450
Do Not Email_Yes	-1.3888	0.172	-8.073	0.000	-1.726	-1.052
Last Activity_Olark Chat Conversation	-1.4467	0.173	-8.357	0.000	-1.786	-1.107
What is your current occupation_Working Professional	2.9005	0.201	14.411	0.000	2.506	3.295
City_Unknown	-1.1216	0.133	-8.456	0.000	-1.382	-0.862
Last Notable Activity_SMS Sent	1.7545	0.082	21.294	0.000	1.593	1.916
Visited_Frequently	0.7557	0.155	4.881	0.000	0.452	1.059

Optimal Cut-Off, ROC Curve and Accuracy



Accuracy Score	0.807

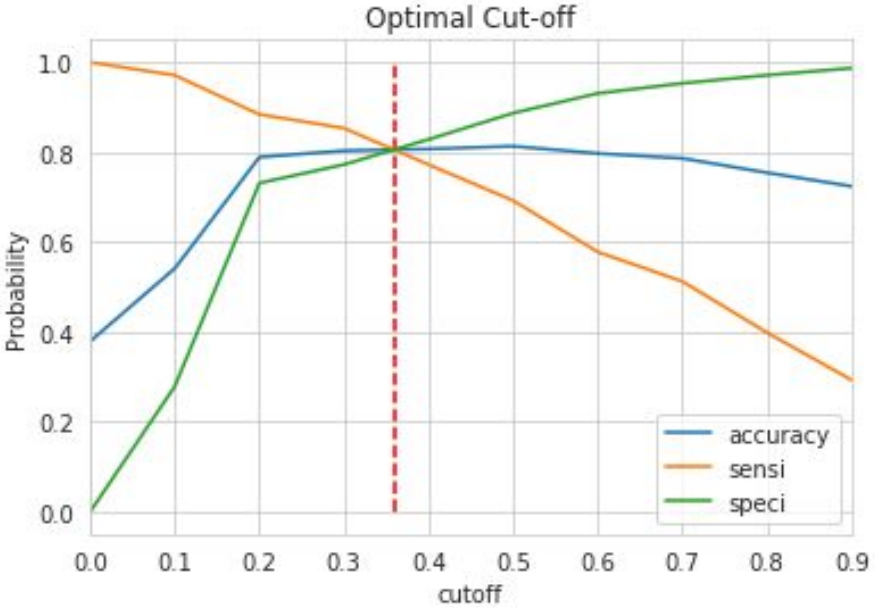


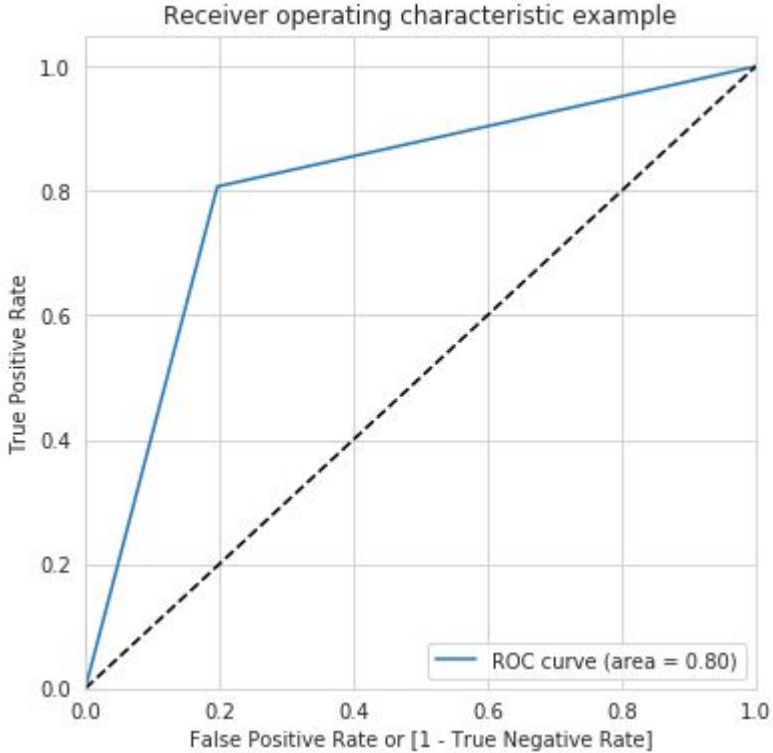
Applying SVM after RFE

To find the best model, GridSearchCV was applied with different hyper-parameters such as:

C: 1, 10, 100, 1000

The best test score was obtained as 0.807 corresponding to hyperparameters {'C': 1}





Accuracy Score	0.807

Different Scores obtained as per Different Models

	LR_TRAIN_RFE	LR_TEST_RFE	LR_TRAIN_PCA	LR_TEST_PCA	SVM_TRAIN_RFE	SVM_TEST_RFE
Sensitivity	0.81	0.78	0.79	0.76	0.81	0.77
Specificity	0.81	0.81	0.80	0.81	0.80	0.81
False_Positive_Rate	0.19	0.19	0.20	0.19	0.20	0.19
Positive_Predictive_Value	0.72	0.71	0.70	0.70	0.71	0.71
Negative_Predictive_Value	0.88	0.86	0.86	0.85	0.87	0.86
Precision	0.72	0.71	0.70	0.70	0.71	0.71
Recall	0.81	0.78	0.79	0.76	0.81	0.77

Conversion Rate of the Predicted Values from the model : **38.54%**

Final DataFrame with Lead Scores

	Prospect ID	Lead Number	Actual	Probability	Opt Cutoff	Score
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	0	0.150046	0	15.00
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	0	0.145008	0	14.50
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	1	0.789203	1	78.92
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	660719	0	0.130323	0	13.03
4	3256f628-e534-4826-9d63-4a8b88782852	660681	1	0.764058	1	76.41
...	...	...	...	...	...	...
8508	19d6451e-fcd6-407c-b83b-48e1af805ea9	579564	1	0.152552	0	15.26
8509	82a7005b-7196-4d56-95ce-a79f937a158d	579546	0	0.186014	0	18.60
8510	aac550fe-a586-452d-8d3c-f1b62c94e02c	579545	0	0.094641	0	9.46
8511	5330a7d1-2f2b-4df4-85d6-64ca2f6b95b9	579538	1	0.338120	1	33.81
8512	571b5c8e-a5b2-4d57-8574-f2ffb06fdeff	579533	1	0.701767	1	70.18