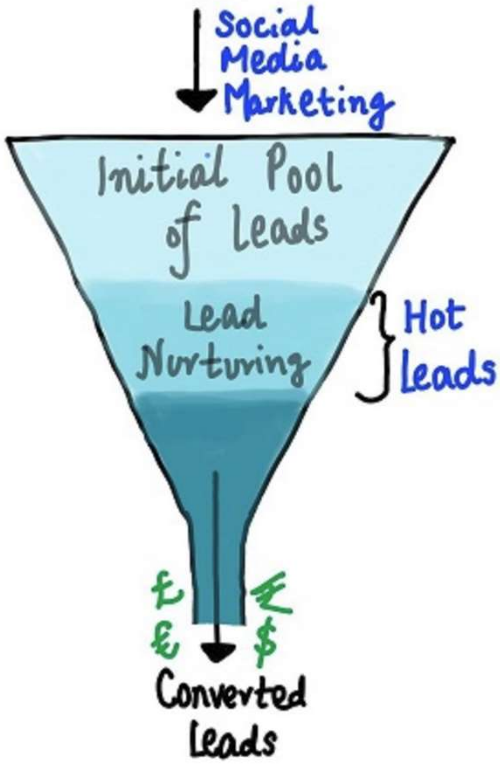
HOT LEAD

Lead Scoring Case Study

By:
Nitanshu Joshi
Anshika Dua



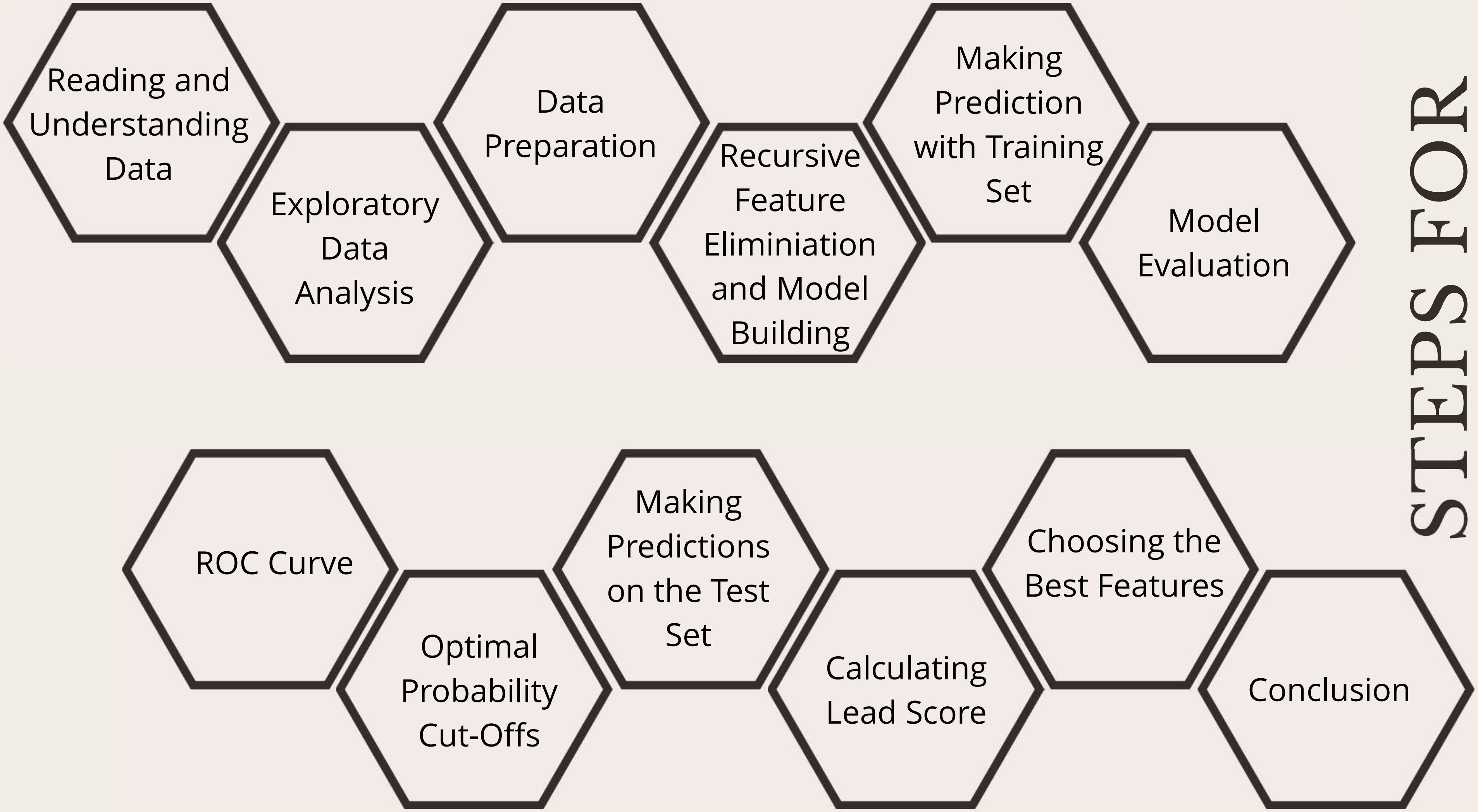
Problem Statement

To build a model 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 ballpark of the target lead conversion rate to be around 80%.

The model should be able to adjust to if the company's requirement changes in the future.

STEPS FOR ANALYSIS

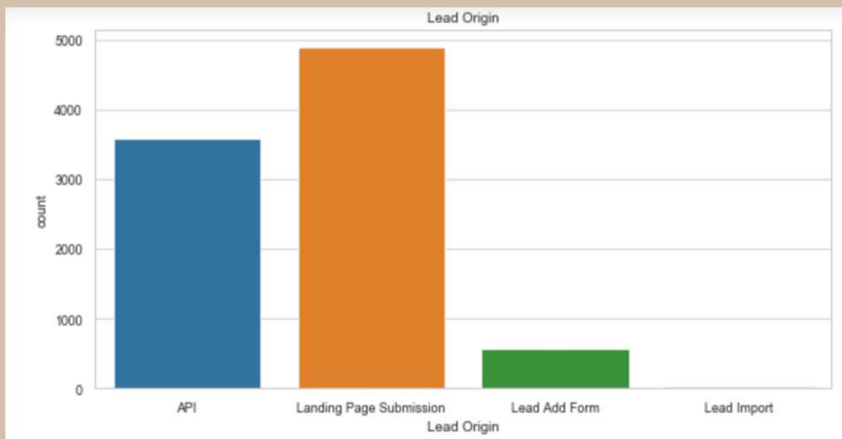


EXPLORATORY DATA ANALYSIS-

DATA CLEANING-

- Dropping Sales Team Generated columns
- Dropping unrequired columns
- Handling NULL Values

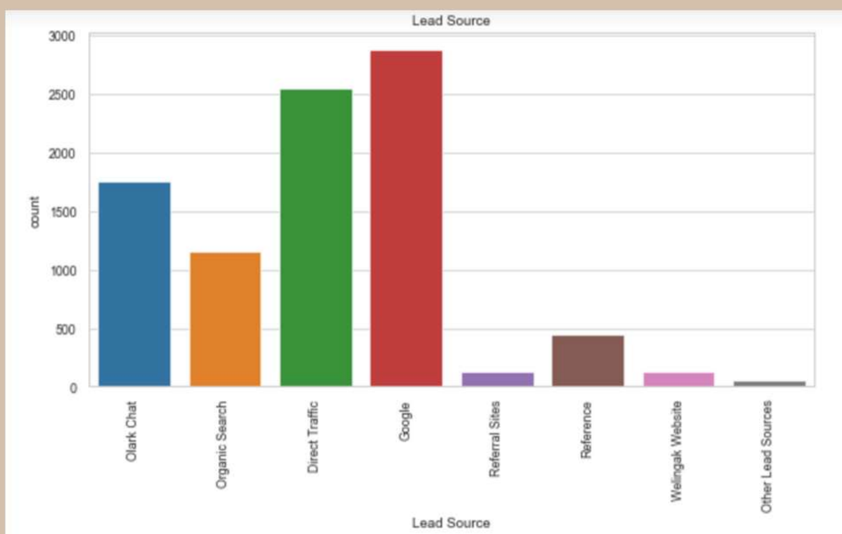
DATA VISUALISATION

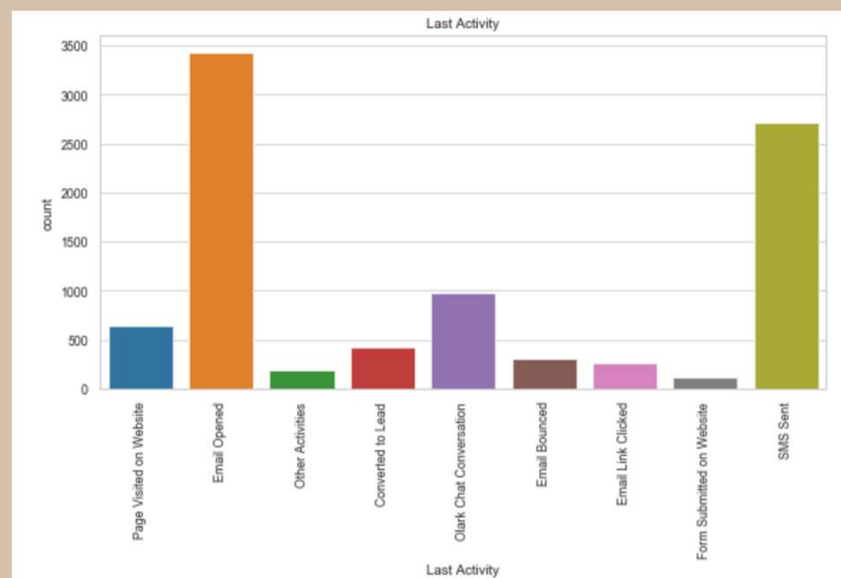


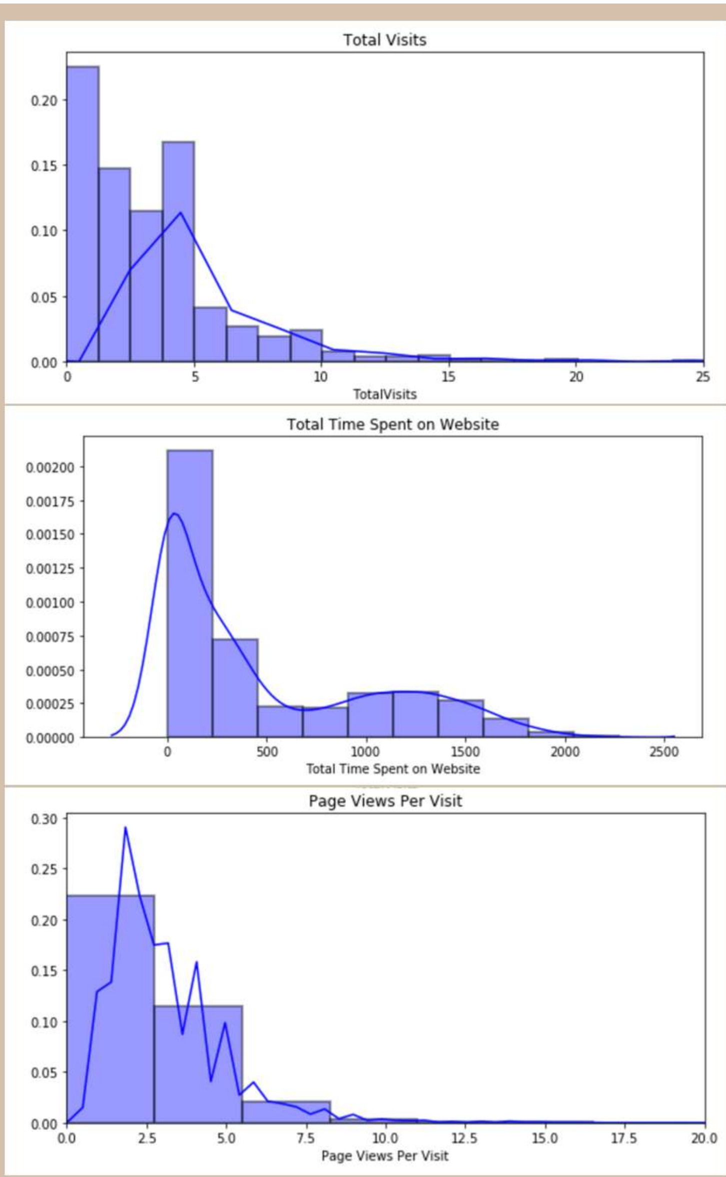
- Most of origin of the leads are from 'Landing Page Submission',while the least is from 'Lead Import'.

- Most of the Lead Sources are from 'Google' followed by 'Direct Traffic' whereas 'Referral Sites' have the least.

- People who opened their email are highly seen by the company as a possible lead.





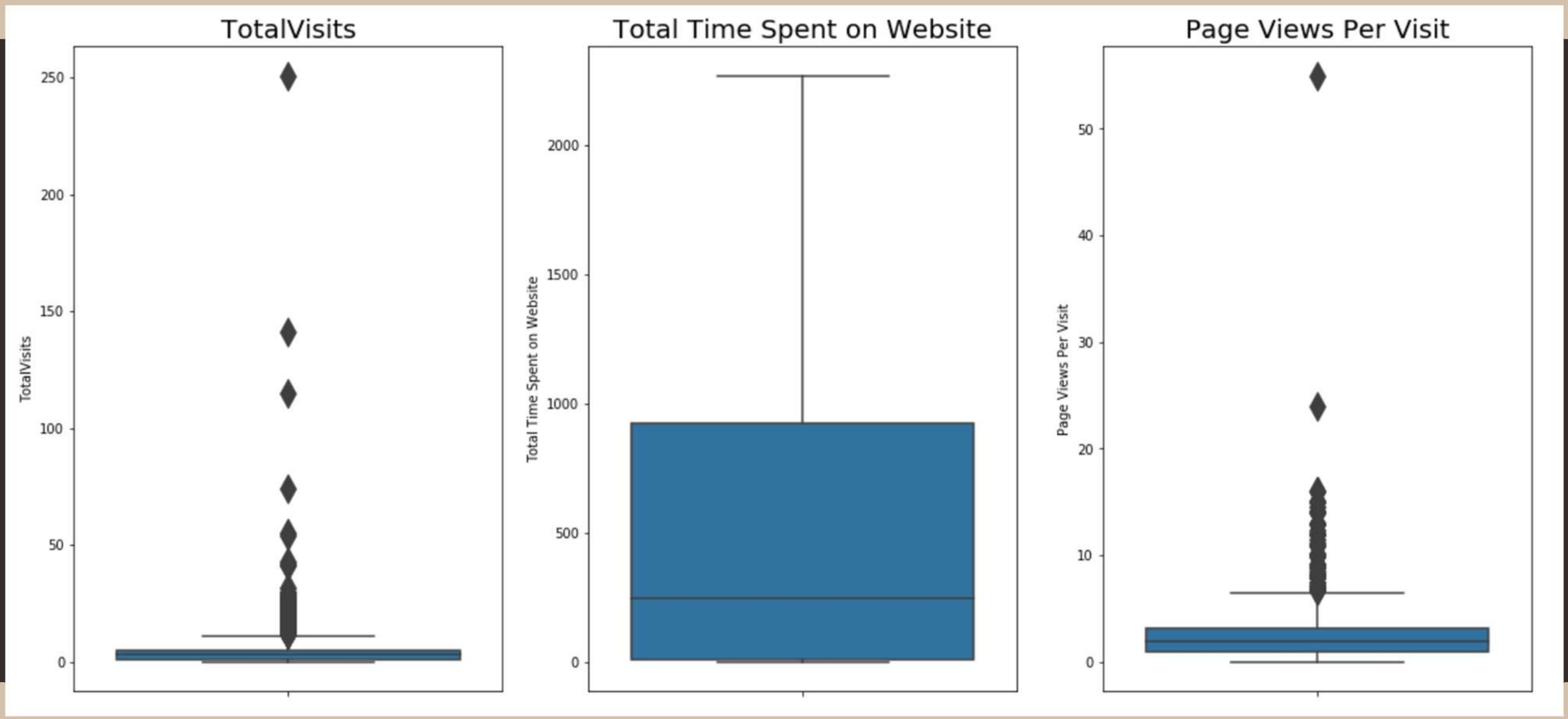


- We observe that the average number of total visits by a customer on the website is on the lower end.

- We also observe that average total time spent on the website by customes is also less.

- We also observe that the average number of page view for the website on one visit is very less.

OUTLIER ANALYSIS & TREATMENT



Clearly, we can see there are outliers in the columns `Page Views Per Visit` and `TotalVisits`, so we created bins in these columns.

DATA PREPARATION

There are 3 non-binary categorical columns:
- Lead Origin
-Lead Source
-Last Activity

For the '**Lead Origin**' column, no conversion is required.

For the '**Lead Source**' column, we can make the following changes:
- We see that there is a repetition of Google and google. We will merege these cell values.
- Also there various labels with very low counts, thus we can convert them to 'Other Lead Sources'. All those column labels with 100 or less values could be converted to 'Other Lead Sources'.

For the '**Last Activity**' column, we can make the following changes:
- We see that there are various labels with very low counts, thus we can convert them to 'Other Activities'. All those column labels with 100 or less values could be converted to 'Other Activities'.

For binary categorical variables, we converting the values of variables into 0s and 1s.

There are only 3 categorical variables that require dummies to be formed:
- Lead Origin
- Lead Source
- Last Activity

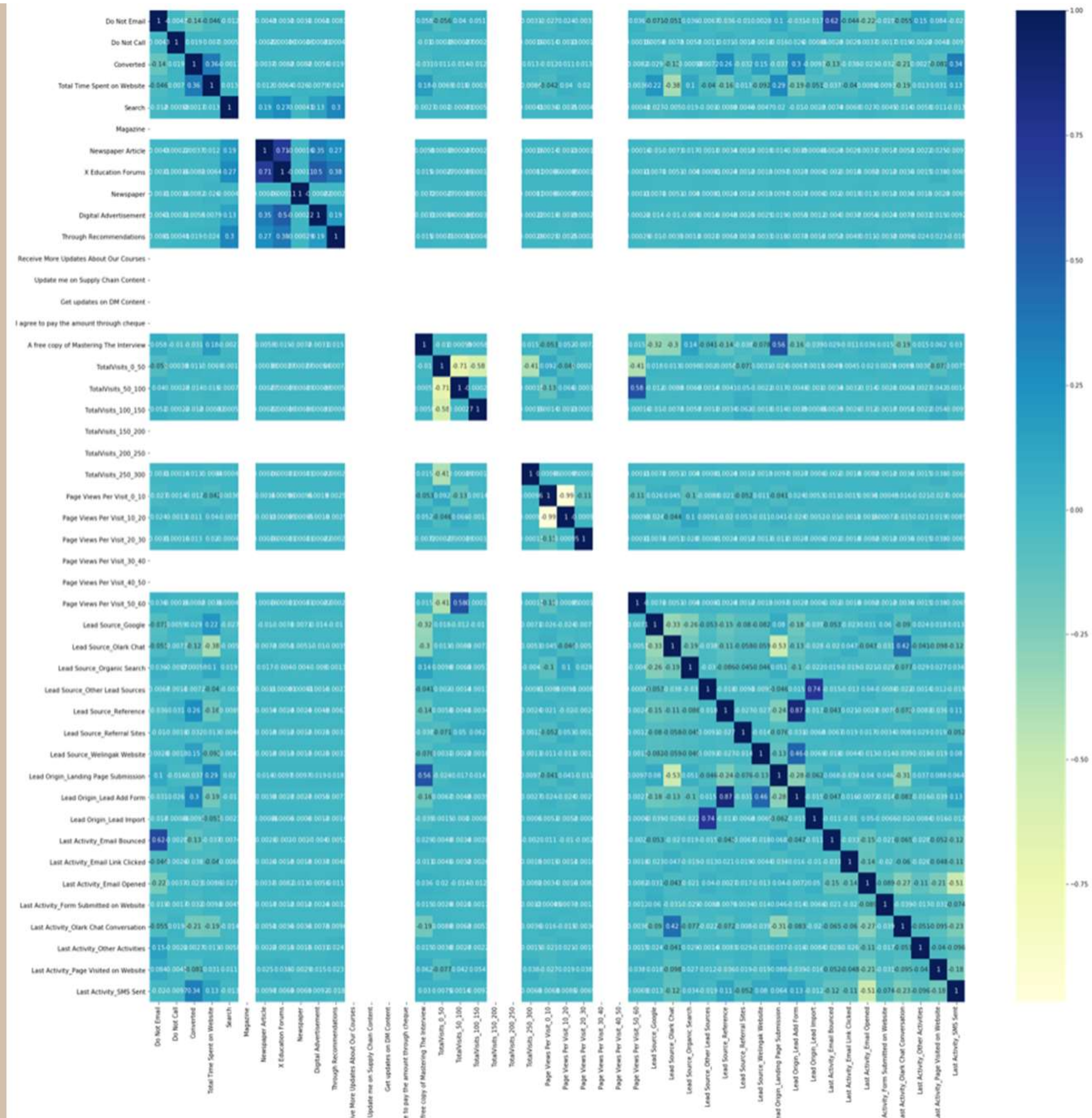
We then split the data into test and train sets and perform Rescaling using Standard Scalar.

On checking the lead conversion rate, we have about 38% churn rate. This is neither exactly 'balanced' (which a 50-50 ratio would be called) nor heavily imbalanced.
So we don't do any special treatment for this dataset.

Correlation:

We observe that quite a few variables have a high correlation between themselves.

We observe that there are a lot of variables present in the dataframe, and it will be very difficult to drop these variables. Thus we will minimize the variables using the RFE (Recursive feature elimination) process.



Recursive Feature Elimination

We run the RFE with output number of variables equal to 15

```
Index(['Do Not Email', 'Total Time Spent on Website', 'Newspaper',
       'TotalVisits_0_50', 'TotalVisits_250_300', 'Page Views Per Visit_0_10',
       'Lead Source_Olark Chat', 'Lead Source_Reference',
       'Lead Source_Welingak Website', 'Lead Origin_Lead Add Form',
       'Lead Origin_Lead Import', 'Last Activity_Email Opened',
       'Last Activity_Olark Chat Conversation',
       'Last Activity_Other Activities', 'Last Activity_SMS Sent'],
      dtype='object')
```

Model Building

We initially build the model with
15 variables selected by RFE.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6335
Model Family:	Binomial	Df Model:	15
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2860.4
Date:	Mon, 11 Jan 2021	Deviance:	5720.7
Time:	16:37:49	Pearson chi2:	6.52e+03
No. Iterations:	20		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-19.3375	1.98e+04	-0.001	0.999	-3.88e+04	3.87e+04
Do Not Email	-1.5817	0.172	-9.182	0.000	-1.919	-1.244
Total Time Spent on Website	1.1559	0.039	29.662	0.000	1.080	1.232
Newspaper	-23.2590	2.92e+04	-0.001	0.999	-5.73e+04	5.73e+04
TotalVisits_0_50	18.2051	1.98e+04	0.001	0.999	-3.87e+04	3.88e+04
TotalVisits_250_300	42.5275	3.53e+04	0.001	0.999	-6.91e+04	6.92e+04
Page Views Per Visit_0_10	-0.7006	0.354	-1.977	0.048	-1.395	-0.006
Lead Source_Olark Chat	1.2235	0.100	12.269	0.000	1.028	1.419
Lead Source_Reference	2.0971	0.920	2.279	0.023	0.294	3.900
Lead Source_Welingak Website	3.6659	1.153	3.180	0.001	1.406	5.926
Lead Origin_Lead Add Form	2.1089	0.896	2.354	0.019	0.353	3.865
Lead Origin_Lead Import	1.4280	0.437	3.268	0.001	0.571	2.285
Last Activity_Email Opened	0.7102	0.104	6.822	0.000	0.506	0.914
Last Activity_Olark Chat Conversation	-0.9095	0.177	-5.128	0.000	-1.257	-0.562
Last Activity_Other Activities	1.5928	0.221	7.199	0.000	1.159	2.026
Last Activity_SMS Sent	1.8630	0.106	17.561	0.000	1.655	2.071

	Features	VIF
3	TotalVisits_0_50	125.63
5	Page Views Per Visit_0_10	120.33
9	Lead Origin_Lead Add Form	62.40
7	Lead Source_Reference	47.97
8	Lead Source_Welingak Website	15.47
11	Last Activity_Email Opened	3.33
14	Last Activity_SMS Sent	2.84
12	Last Activity_Olark Chat Conversation	1.91
6	Lead Source_Olark Chat	1.77
1	Total Time Spent on Website	1.30
0	Do Not Email	1.24
13	Last Activity_Other Activities	1.12
4	TotalVisits_250_300	1.02
10	Lead Origin_Lead Import	1.02
2	Newspaper	1.00

Upon removing variables with high VIF and p-values, we get the following final model:

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6351
Model:                            GLM   Df Residuals:                     6340
Model Family:                Binomial   Df Model:                           10
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2870.1
Date:                Mon, 11 Jan 2021   Deviance:                       5740.2
Time:                        16:48:12   Pearson chi2:                 6.53e+03
No. Iterations:                     7
Covariance Type:            nonrobust
===========================================================================================================
                                          coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------
const                                  -1.8113      0.093    -19.394      0.000      -1.994      -1.628
Do Not Email                           -1.5755      0.171     -9.187      0.000      -1.912      -1.239
Total Time Spent on Website             1.1496      0.039     29.627      0.000       1.074       1.226
Lead Source_Olark Chat                  1.2099      0.099     12.172      0.000       1.015       1.405
Lead Source_Reference                   4.1877      0.221     18.940      0.000       3.754       4.621
Lead Source_Welingak Website            5.7550      0.729      7.896      0.000       4.327       7.184
Lead Origin_Lead Import                 1.4115      0.437      3.230      0.001       0.555       2.268
Last Activity_Email Opened              0.6994      0.104      6.745      0.000       0.496       0.903
Last Activity_Olark Chat Conversation  -0.9203      0.177     -5.197      0.000      -1.267      -0.573
Last Activity_Other Activities          1.5937      0.221      7.219      0.000       1.161       2.026
Last Activity_SMS Sent                  1.8531      0.106     17.541      0.000       1.646       2.060
===========================================================================================================
```

	Features	VIF
2	Lead Source_Olark Chat	1.74
7	Last Activity_Olark Chat Conversation	1.38
1	Total Time Spent on Website	1.29
9	Last Activity_SMS Sent	1.22
3	Lead Source_Reference	1.16
6	Last Activity_Email Opened	1.16
0	Do Not Email	1.06
4	Lead Source_Welingak Website	1.05
8	Last Activity_Other Activities	1.03
5	Lead Origin_Lead Import	1.02

MAKING PREDICTIONS ON THE TRAIN SET:

Data frame with given convertion rate and probablity of predicted ones

	Converted	Conversion_Prob	LeadID
0	0	0.21	3009
1	0	0.02	1012
2	0	0.56	9226
3	1	0.87	4750
4	1	0.91	7987
5	1	0.75	1281
6	0	0.11	2880
7	1	0.90	4971
8	1	0.88	7536
9	0	0.90	1248

Creating a new column 'Predicted' with 1 if conversion rate>0.5, else 0.

	Converted	Conversion_Prob	LeadID	Predicted
0	0	0.21	3009	0
1	0	0.02	1012	0
2	0	0.56	9226	1
3	1	0.87	4750	1
4	1	0.91	7987	1
5	1	0.75	1281	1
6	0	0.11	2880	0
7	1	0.90	4971	1
8	1	0.88	7536	1
9	0	0.90	1248	1

Model Evaluation

Confusion Matrix

Predicted Values >>>	Lead Not Converted	Lead Converted
Actual Values	--	--
Lead Not Converted	TN = 3425	FP = 480
Lead Converted	FN = 805	TP = 1641

Accuracy: 0.7976696583215241

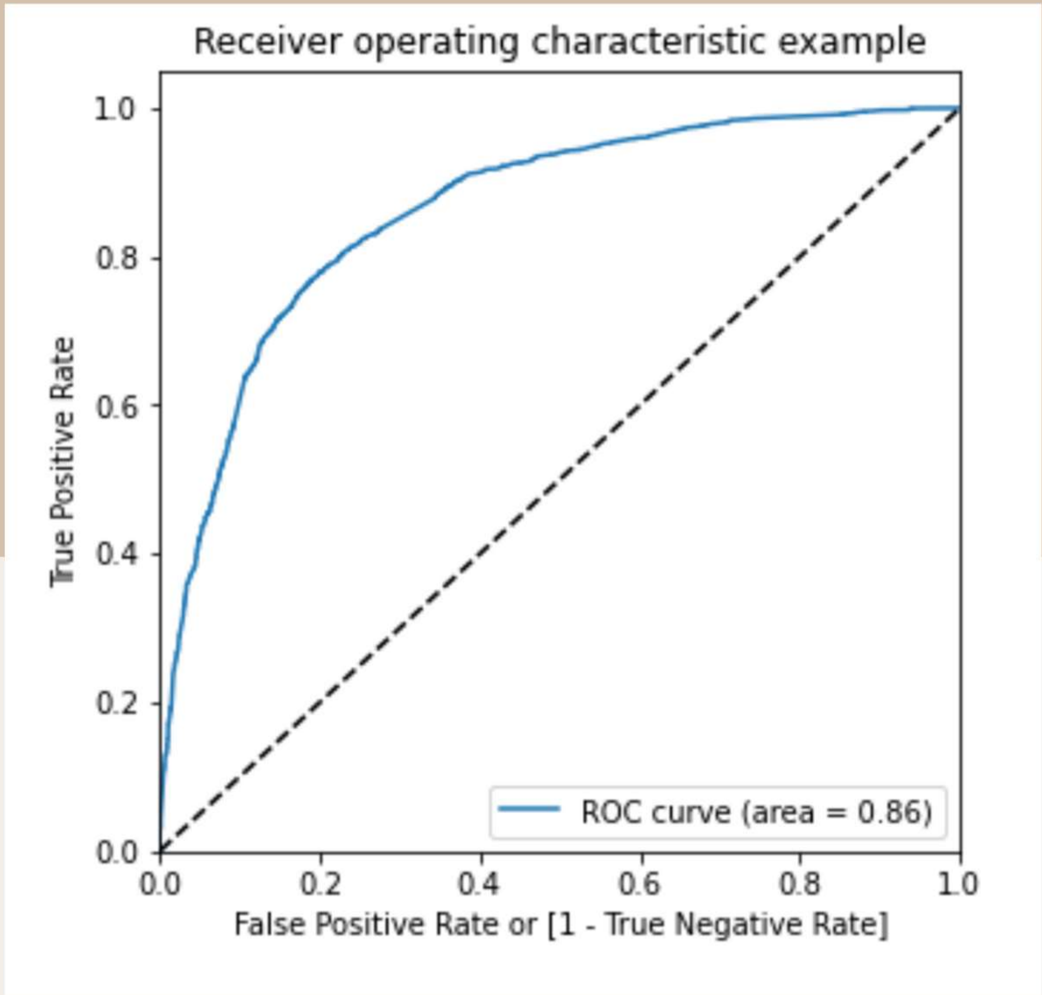
Sensitivity: 0.6708912510220768

Specificity: 0.8770806658130602

Precision: 0.7736916548797736

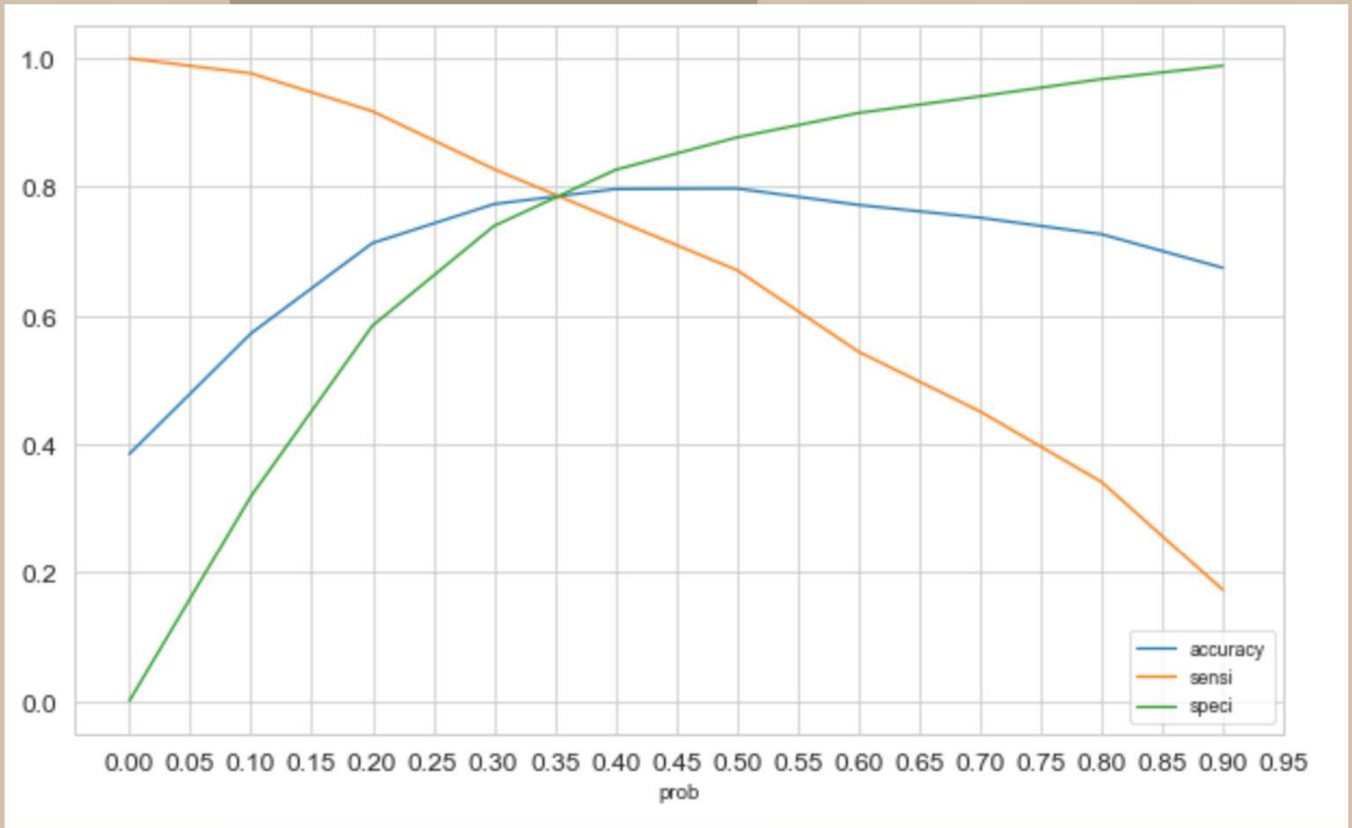
ROC Curve:

An ROC curve demonstrates several things:
- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



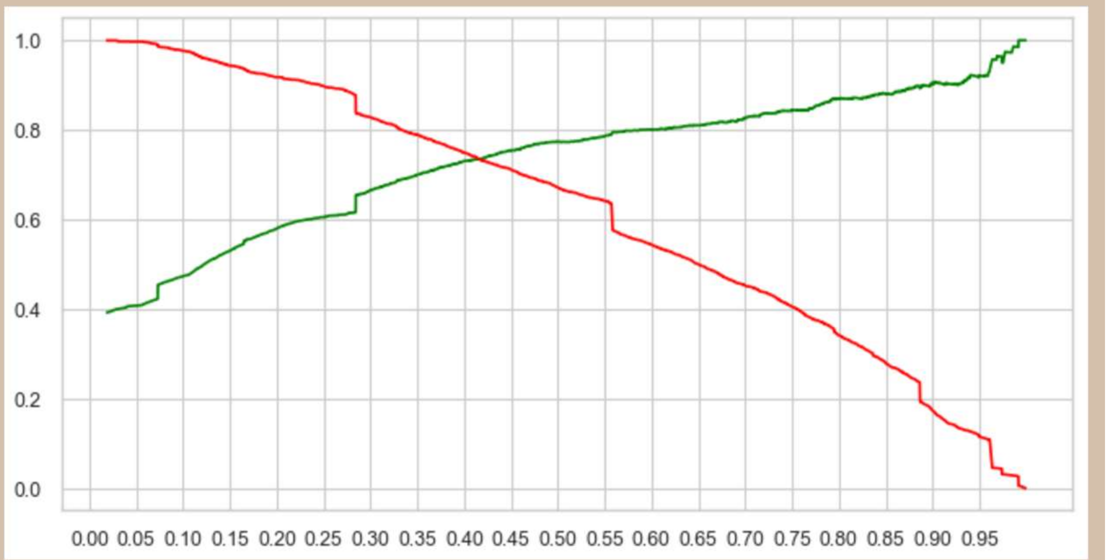
The area under the curve comes out to be **0.86** which is considered a very good value

OPTIMAL PROBABILITY CUT-OFFS



From the curve we find that the optimum point to be taken as a cut-off for the probability values is **0.35**.

Precision Recall Trade-off



From the above curve we observe that the precision-recall curve gives us a cut-off of 0.42, but we already fulfilled our business requirement of lead conversion-rate of about 80%

Confusion Matrix

Predicted Values >>>	Lead Not Converted	Lead Converted
Actual Values	--	--
Lead Not Converted	TN = 3011	FP = 894
Lead Converted	FN = 479	TP = 1967

Accuracy: 0.7882223271925681
Sensitivity: 0.7882256745707277
Specificity: 0.7882202304737516
Precision: 0.6998185117967333

F1 score: 0.7413934771840115

Making Predictions on Test Set

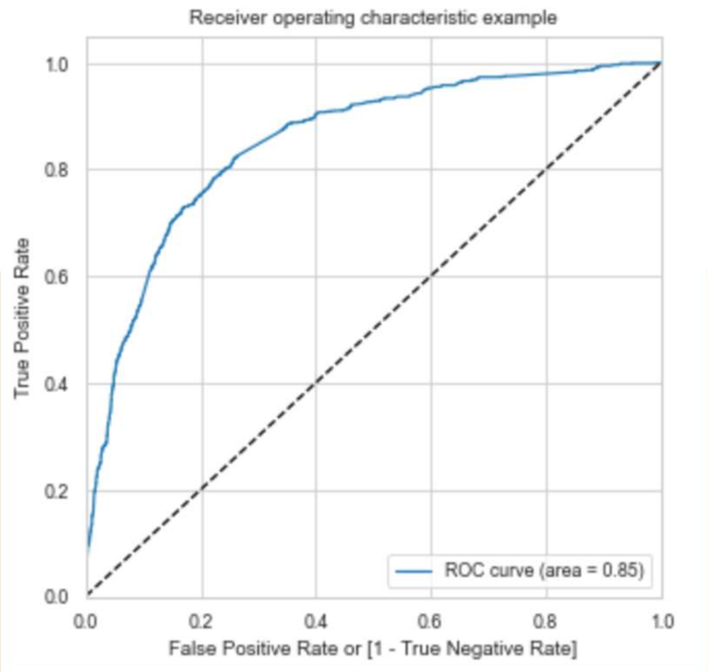
Prediction using cut-off 0.35

	Converted	LeadID	Conversion_Prob	final_predicted
0	0	3271	0.14	0
1	1	1490	0.74	1
2	0	7936	0.12	0
3	1	4216	0.89	1
4	0	3830	0.14	0

Confusion matrix

Predicted Values >>>	Lead Not Converted	Lead Converted
Actual Values	--	--
Lead Not Converted	TN = 1361	FP = 373
Lead Converted	FN = 230	TP = 759

ROC Curve:



Since we got a value of 0.85, our model seems to be doing well on the test dataset.

Metrics:

Accuracy: 0.7785530664708042

Sensitivity: 0.7674418604651163

Precision: 0.6704946996466431

Recall: 0.7674418604651163

Specificity: 0.7848904267589388

LEAD SCORE FOR ENTIRE DATASET

Lead Score = Conversion Probability * 100

LeadID	Converted	Conversion_Prob	final_predicted	Lead_Score
0	0	0.16	0	16
1	0	0.33	0	33
2	1	0.75	1	75
3	0	0.36	1	36
4	1	0.54	1	54

Choosing the Best Features

The top features which contribute most towards the probability of lead getting converted are:

The conversion probability of a lead increases with increase in values of the following features in descending order:

- Lead Source_Welingak Website
- Lead Source_Reference
- Last Activity_SMS Sent
- Last Activity_Other Activities
- Lead Origin_Lead Import
- Lead Source_Olark Chat
- Total Time Spent on Website
- Last Activity_Email Opened

The conversion probability of a lead increases with decrease in values of the following features in descending order:

- Do Not Email
- Last Activity_Olark Chat Conversion

Conclusion

The model we made using the logistic regression can be considered a good model. It has the following characteristics -
- All the features/variables have a P-value of less than 0.05.

- The VIF scores for all the variables are very low and are less than 5, thus there is hardly any multi-collinearity between the variables.

- The overall accuracy of the model is around 78%, with a threshold probability of 0.5. Thus the accuracy is very acceptable.

- Also the specificity of the model is around 78.5% which is also acceptable.

- The sensitivity/recall of the model is around 76% which is also acceptable.

- The precision of the model is about 68% which could be considered decent.

- Also, when we plotted the ROC curve, the area under the curve we got was aroung 86%, which could be considered good.

Overall this model meets our business requirement, where we can say we got a lead conversion rate of nearly 80%

Apart from the model there were some variable which greatly influenced our model.

The top 3 variables which influenced our model in a positive way are -

1 - Lead Source - Welingak Website

2 - Lead Source - Reference

3 - Last Activity - SMS Sent

The top 2 variables which influenced our model in a negative way are -

1 - Do Not Email

2 - Last Activity - Olark Chat Conversion

THANK YOU