LEAD SCORING CASE STUDY

PRESENTED BY –

RAHUL CHOPRA

SONAL HEDAOO

PURPOSE OF THE LEAD SCORING CASE STUDY

- The Company X Education require us to 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.

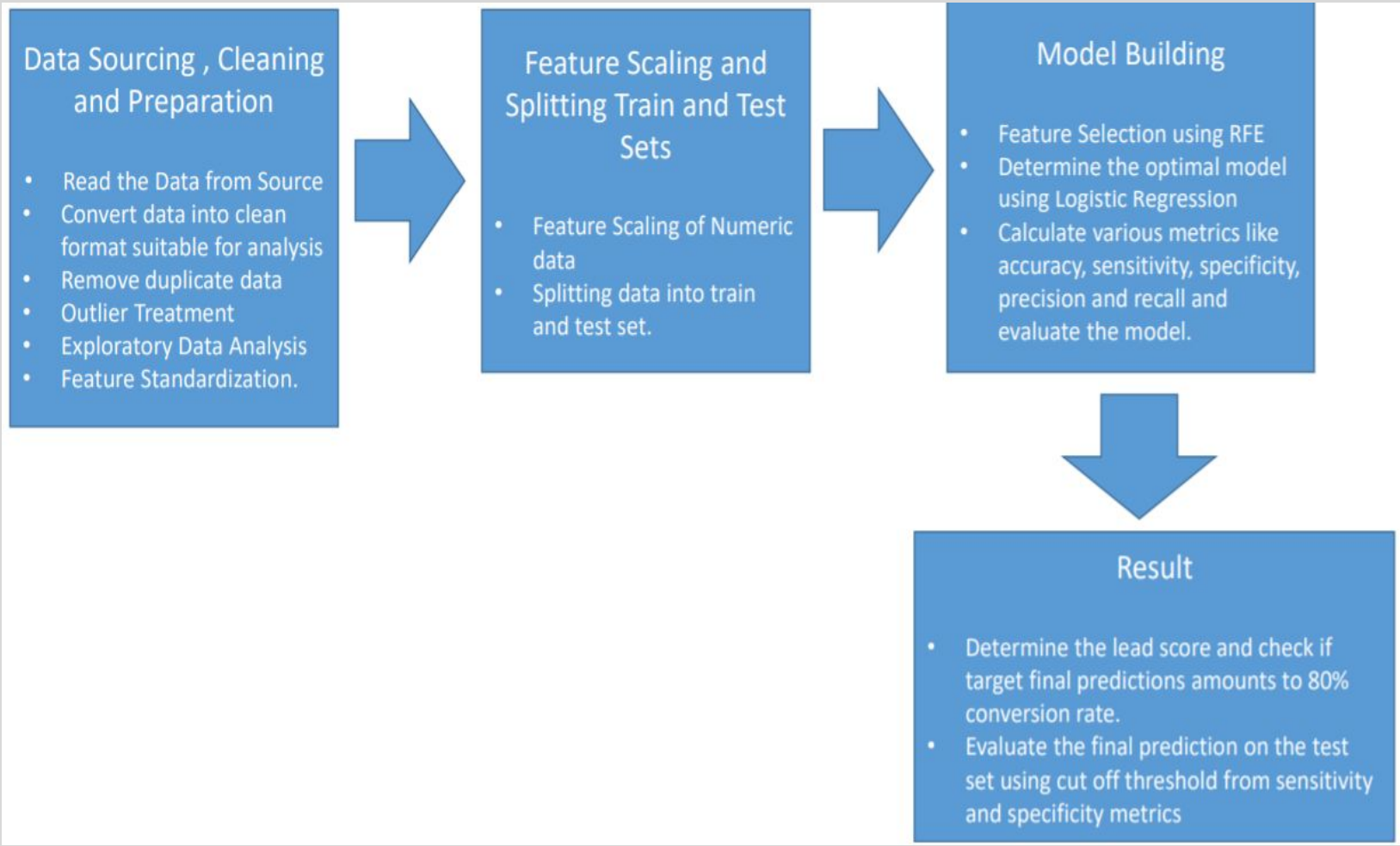
- So we need to need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

In [1]:

OPERATIONS PERFORMED

- Reading the Data
- Data Preparation
- EDA
- Dummy Variable Creation
- Test-Train Split
- Feature Scaling
- Checking Correlation
- Model Building
- Model Evaluation
- Conclusion



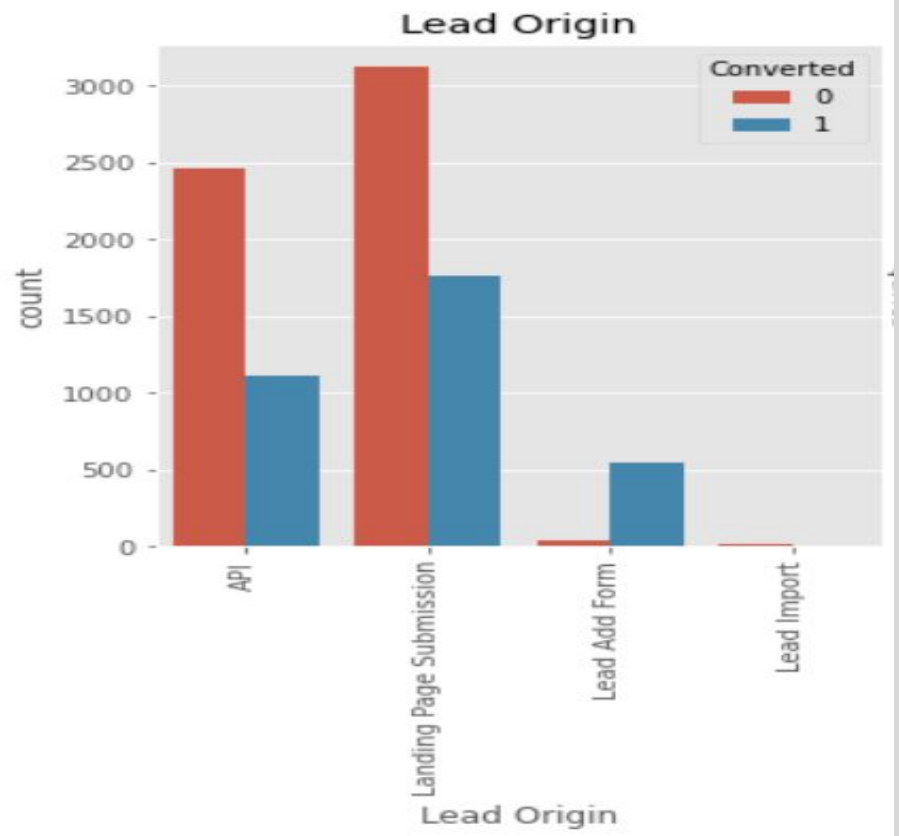
DATA PREPARATION AND DATA CLEANING

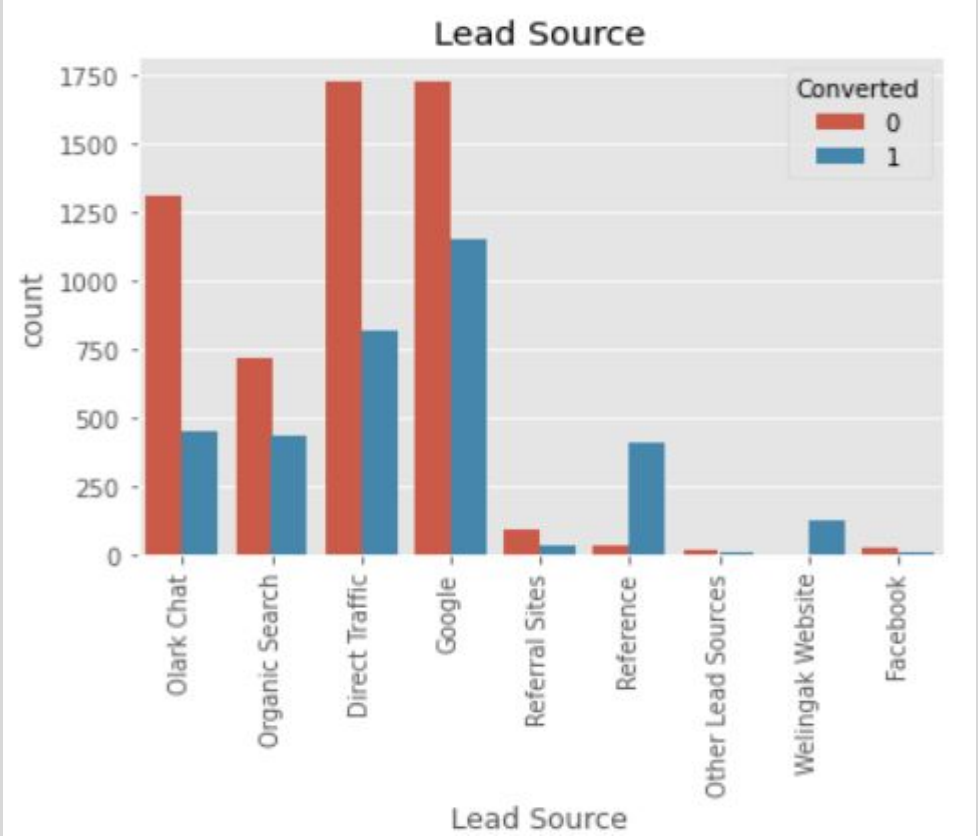
- Encoding the lables with Yes/No to 1s/0s.
- Performed a sanity check on certain columns.
- Replaced the Select data points by null values.
- Checked missing data percentage in the dataset.
- Dropped the columns having more than 40% null or missing values.
- Searched for the missing values and dropped the columns having more than 40% null or missing values.
- Performed a sanity check on certain columns.
- Imputed the null values by new categories for respective variables.
- Dropped the remaining rows directly containing the missing values.
- Handled remaining columns with data imbalance that are binary in nature.
- After dropping certain unwanted columns 98.2% of the data has been retained after the data cleaning.

EXPLORATORY DATA ANALYSIS

We can observed that:

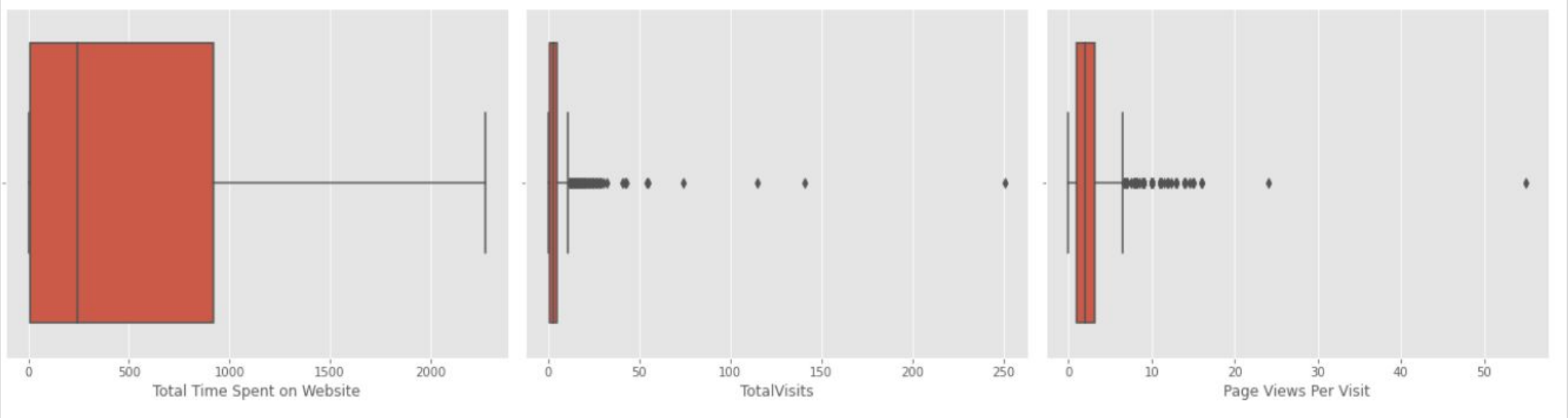
- As per Lead Origin, we can ignore the Lead import as it has very less conversion rate as well as very less count.
- The conversion rate of Lead Add Form is high as compared to it's count which is quite low.
- In Lead Origin, Landing Page Submission has maximum conversion rate.
- As per Lead Source, Major Conversion in the lead source is from Google.
- The count rate of Google and Direct Traffic is more.





CHECKING FOR CONVERSION FOR NUMERICAL VALUES

- We can see that Time Spent on Website, TotalVisits, Page Views Per Visit having outliers.

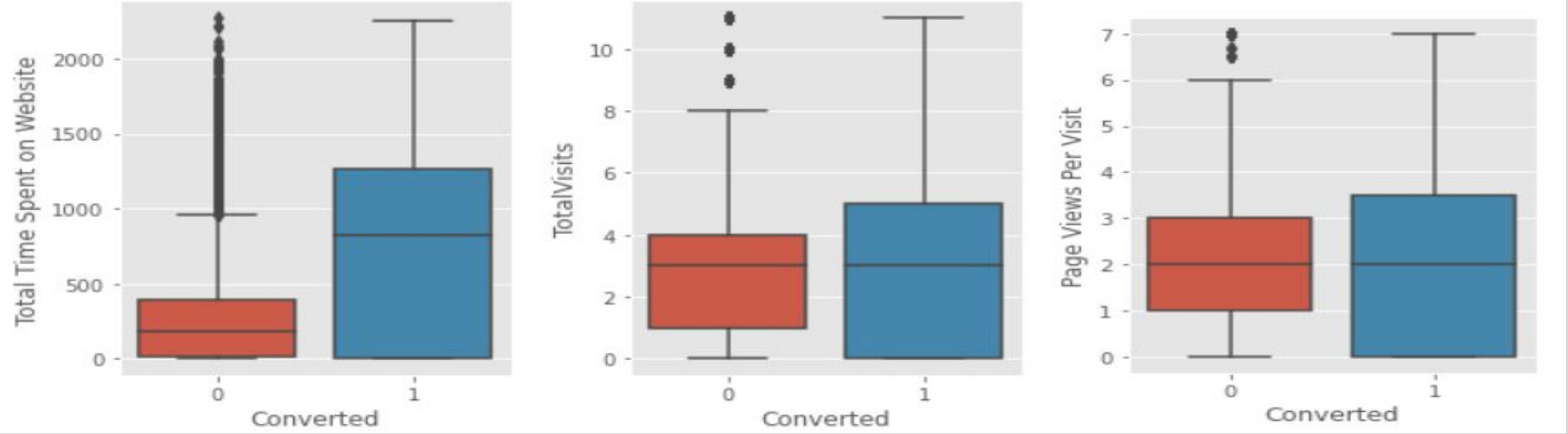


Total Time Spent on Website, TotalVisits, Page Views Per Visit

CHECKING FOR CONVERSION FOR NUMERICAL VALUES

After handling outliers:

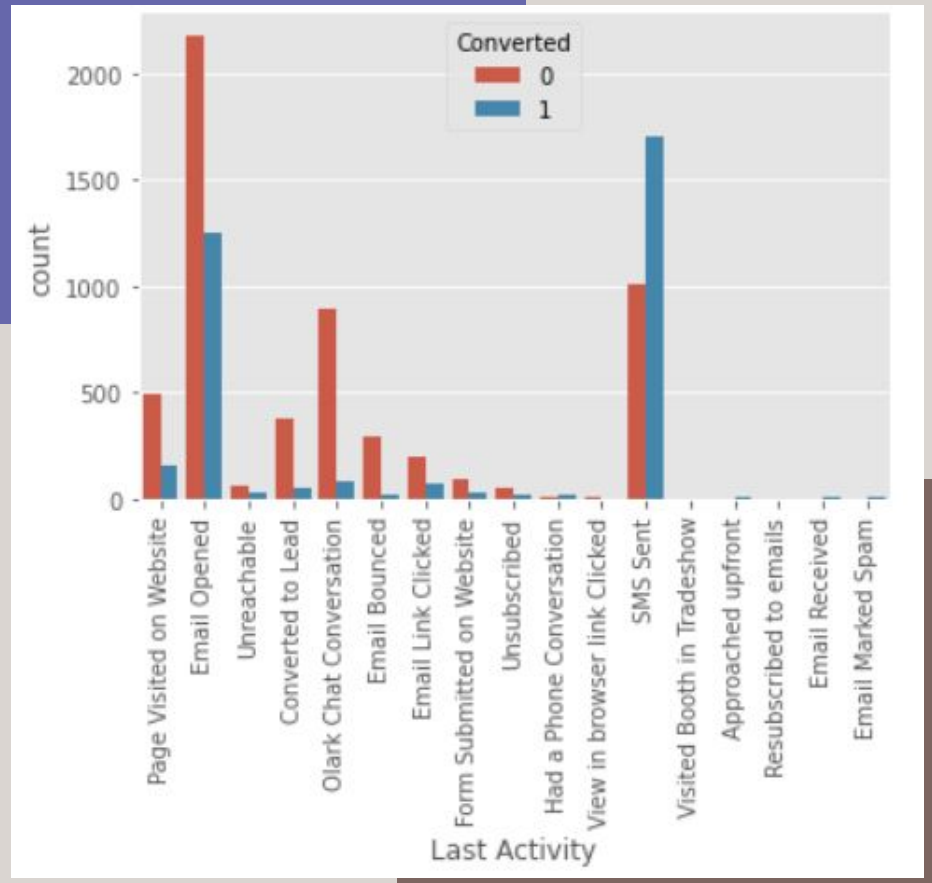
- As per TotalVisits, Median for converted and not converted leads are the close.
- Nothing conclusive can be said on the basis of TotalVisits.
- As per Total Time Spent on Website, spending more time on the website are more likely to be converted.
- Website should be made more engaging to make leads spend more time.
- We can see in Page Views Per Visit, there are no outliers present after handling process.



Total Time Spent on Website, TotalVisits, Page Views Per Visit

LAST ACTIVITY

- We can observed from above plot that the Conversion rate for Email Opened is Highest .
- The Count for SMS Sent is Highest as compared to others activities.



CHECKING TOP CORRELATIONS

- Highly correlated attributes create dependency on various independent factors which will give us inappropriate results.

- We can observe that there are many correlated attributes that needs to be removed.

Last Activity_Resubscribed to emails	Last Notable Activity_Resubscribed to emails	1.000000
Last Activity_Email Marked Spam	Last Notable Activity_Email Marked Spam	1.000000
Lead Origin_Lead Import	Lead Source_Facebook	0.983684
Last Activity_Unsubscribed	Last Notable Activity_Unsubscribed	0.872656
Lead Origin_Lead Add Form	Lead Source_Reference	0.866191
Last Activity_Email Opened	Last Notable Activity_Email Opened	0.861636
Last Activity_SMS Sent	Last Notable Activity_SMS Sent	0.853102
Last Activity_Email Link Clicked	Last Notable Activity_Email Link Clicked	0.800686
TotalVisits	Page Views Per Visit	0.755385
Last Activity_Had a Phone Conversation	Last Notable Activity_Had a Phone Conversation	0.747877
Last Activity_Email Received	Last Notable Activity_Email Received	0.707068
Last Activity_Page Visited on Website	Last Notable Activity_Page Visited on Website	0.691811
Last Activity_Unreachable	Last Notable Activity_Unreachable	0.594369
A free copy of Mastering The Interview	Lead Origin_Landing Page Submission	0.564863
Page Views Per Visit	Lead Origin_Landing Page Submission	0.538577

dtype: float64

AFTER DROPPING THE HIGHLY CORRELATED FEATURES

- Looking at the attachment, it is confirmed that those highly correlated variables were dropped successfully.

TotalVisits	Page Views Per Visit	0.755504
Lead Origin_Lead Add Form	Lead Source_Welingak Website	0.468225
Last Activity_Email Bounced	Last Notable Activity_Email Bounced	0.450911
Lead Source_Olark Chat	Last Activity_Olark Chat Conversation	0.419173
Last Activity_View in browser link Clicked	Last Notable Activity_View in browser link Clicked	0.408088
Last Activity_Olark Chat Conversation	Last Notable Activity_Olark Chat Conversation	0.406150
Total Time Spent on Website	Page Views Per Visit	0.348810
TotalVisits	Total Time Spent on Website	0.342757
Last Activity_Olark Chat Conversation	Last Notable Activity_Modified	0.328700
Page Views Per Visit	Lead Source_Organic Search	0.316105
TotalVisits	Lead Source_Organic Search	0.300473
	A free copy of Mastering The Interview	0.290174
Last Activity_Converted to Lead	Last Notable Activity_Modified	0.288808
Page Views Per Visit	A free copy of Mastering The Interview	0.285645
	Lead Source_Google	0.249018

dtype: float64

MODEL EVALUATION ON TRAIN AND TEST DATA

MODEL EVALUATION : CONFUSION MATRIX

After creating Confusion Matrix

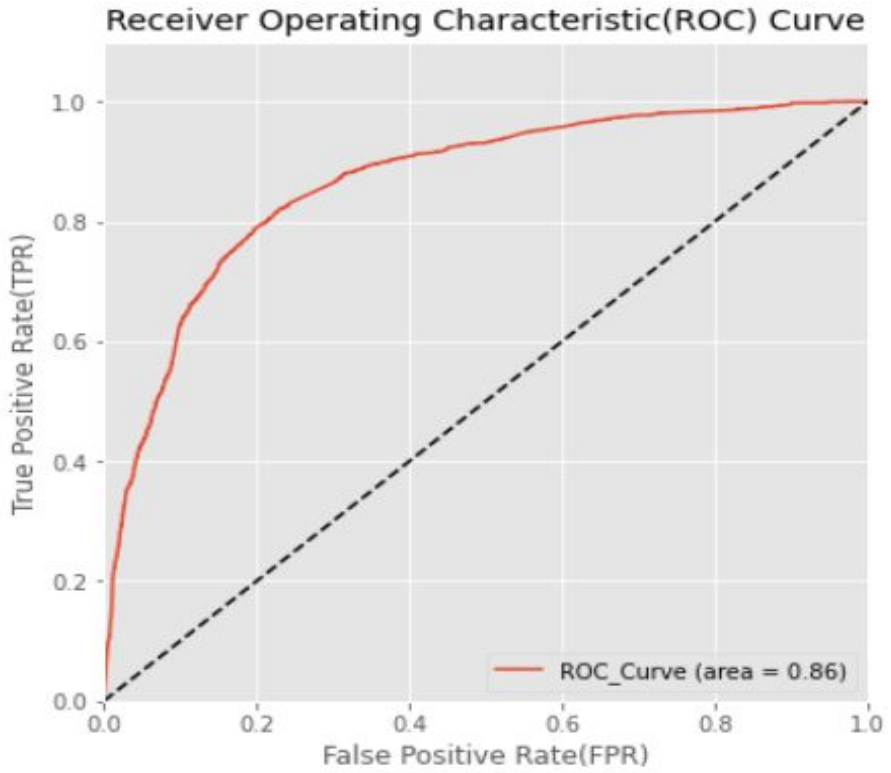
```
[[3435  470]
 [ 810 1636]]
```

- Accuracy – 79.8%
- Sensitivity -66.8%
- Specificity-87.9%

- We can observe that the specificity of the model is higher than the accuracy and sensitivity.
- All the 3 metrix are on similar scale.

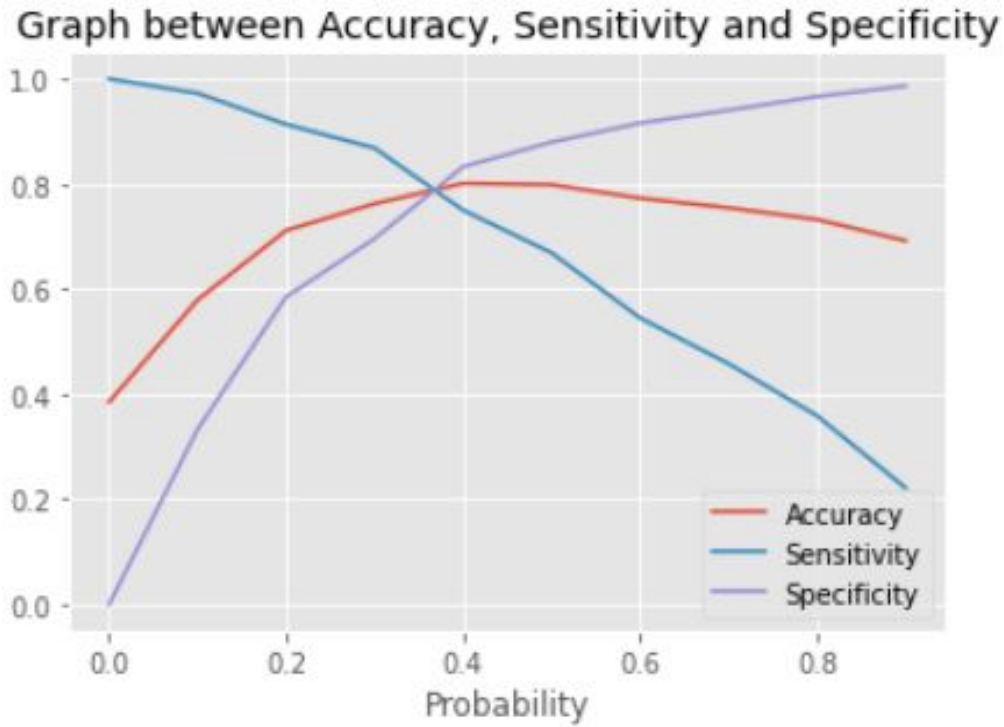
RECEIVER OPERATING CHARACTERISTIC CURVE (ROC CURVE)

- From attached graph, the area under curve is significant and the curve is leaning towards the upper right corner stating that the model has good predictive power.
- The ROC curve area is 0.86 indicating stable accuracy.



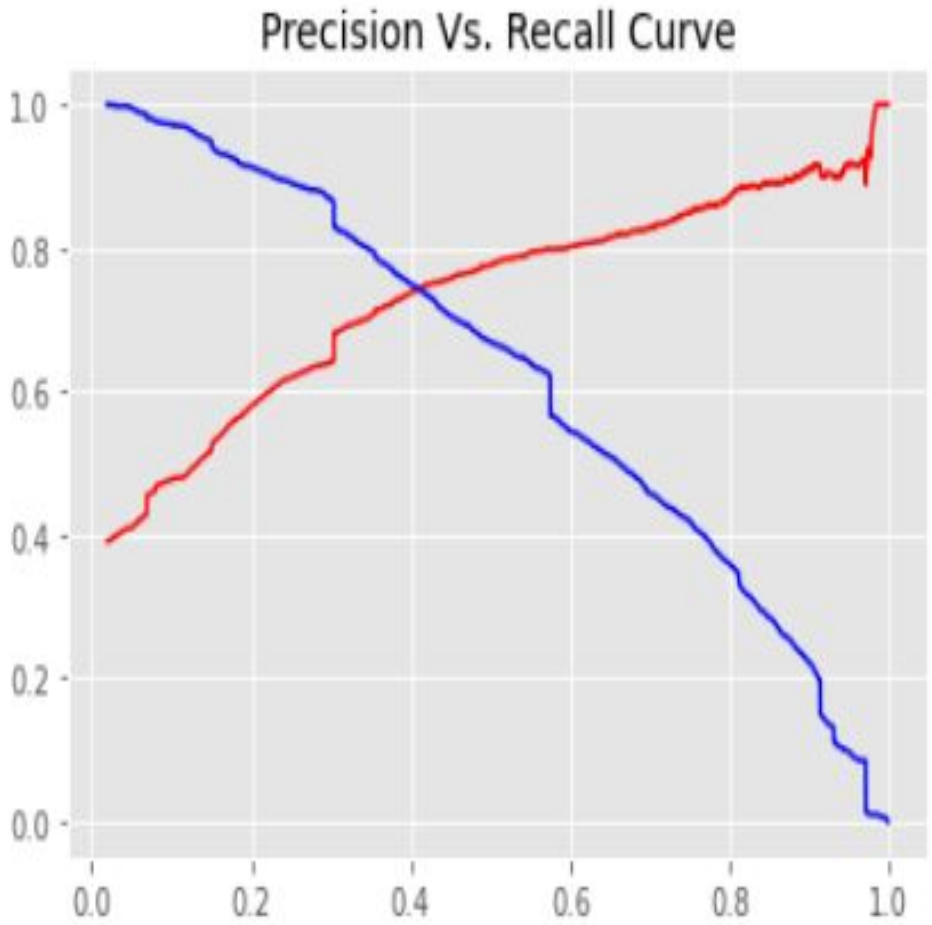
FINDING THE OPTIMAL CUTOFF POINT

- The optimal probability cutoff is where all three curves are intersecting at 0.37 approximately on Accuracy, Sensitivity, Specificity.



Precision Recall Curve

- We created a graph which will show us the trade off between Precision and recall.
- As per graph curve, meeting point is nearly at 0.4



RESULTS FOR TRAIN DATASET AND TEST DATASET

TRAIN DATASET-

- Accuracy - 79.71%
- Sensitivity- 77.55%
- Specificity – 81.05%

TEST DATASET-

- Accuracy - 78.92%
- Sensitivity- 75.16%
- Specificity – 81.08%

CONCLUSION

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction-
- Accuracy, Sensitivity and Specificity values of test set are around 78.92%, 75.16% and 81.08% which are approximately closer to the respective values calculated using trained set.
- It has been concluded that final model has a sensitivity of 75.16% which is an indicator of converted leads being identified about 75.16% correctly.
- Moreover, the values of accuracy, sensitivity and specificity are quite close in the train and test dataset and there is no case of overfitting.
- So we can see that the model is a good fit.

RECOMMENDATIONS

- The lead origins from lead add form, lead import and olark chat seem to have significance on the lead being converted.
- The amount of time lead spends on website is a good indicator and is positive in nature.
- So their last activity like EMail Bounced, Had a phone conversation, SMS Sent are also indicators having significance in explaining whether the lead will convert or not.
- Finally, the last notable activity as modified and olark chat conversation also have a negative significance on conversion.

THANK YOU