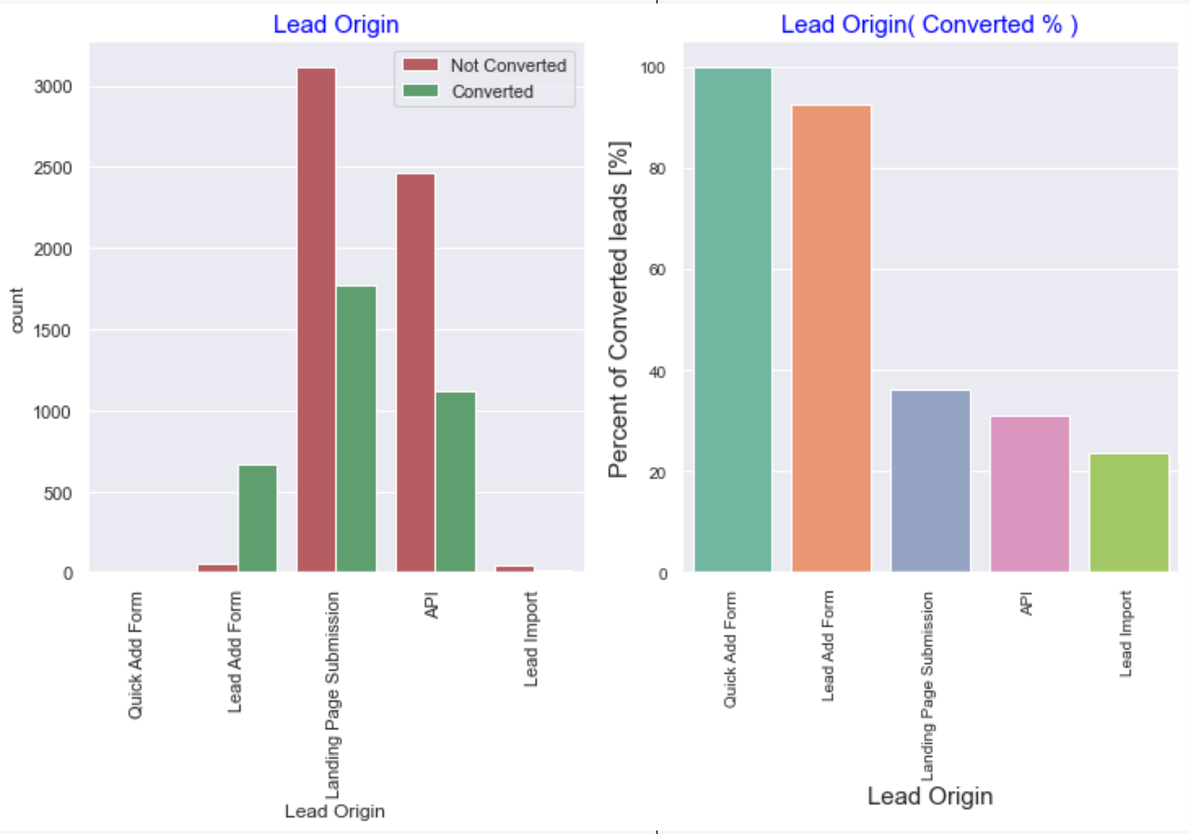
LEAD SCORING CASE STUDY

Reema David

Approach

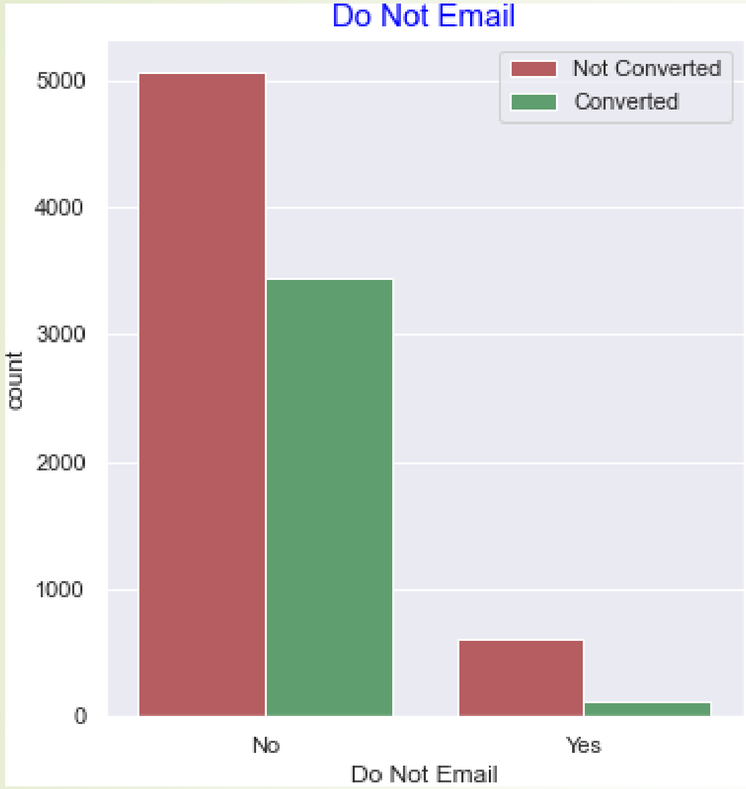
- **Data cleaning and data manipulation.**
- **EDA**
- **Feature Scaling & Dummy Variables**
- **Logistic Regression**
- **Validation of Model**
- **Model Presentation**
- **Conclusion**

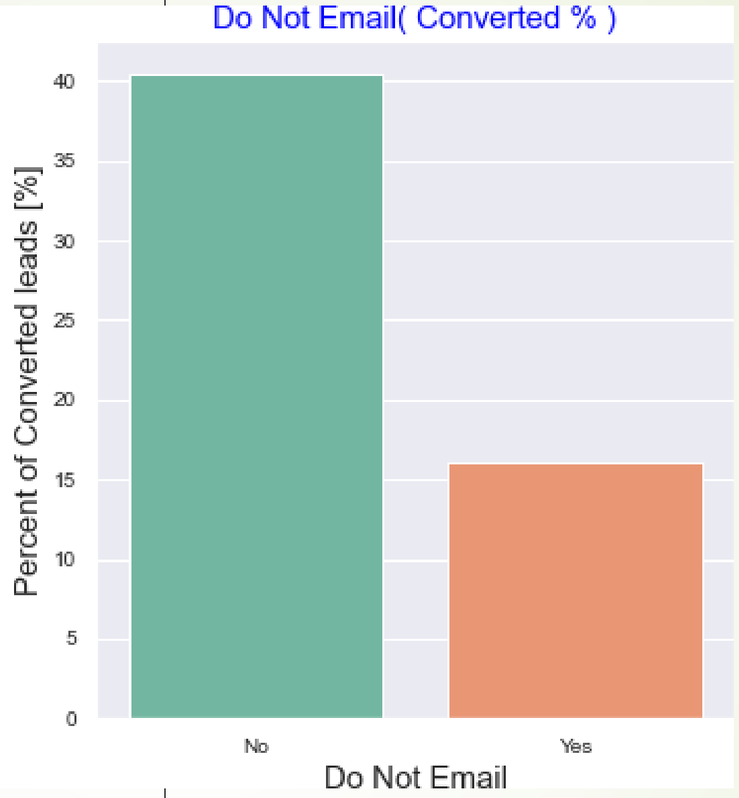
Univariate Analysis -Lead Origin



- Most Leads are from "Landing Page submissions" out of which around 38% got converted, followed by "API", where around 32% are converted.
- Leads from the "Lead Add Form" have third highest conversions with conversion rate around 90%.
- "Lead Import" has only 55 records with the lowest conversion rate if around 22%
- "Quick Add Form" are 100% Converted with just 1 lead from this category.

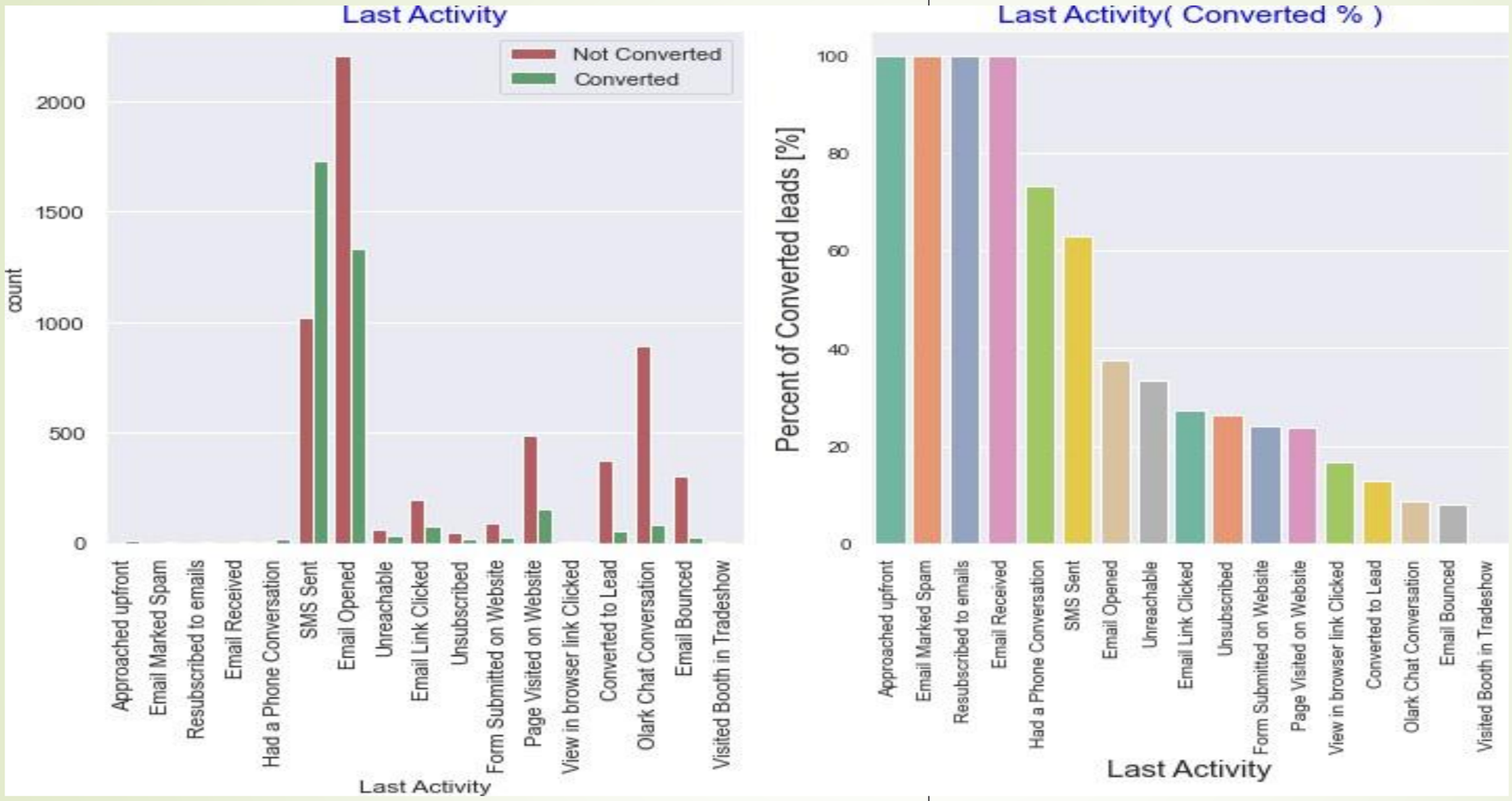
Univariate- Do Not Email





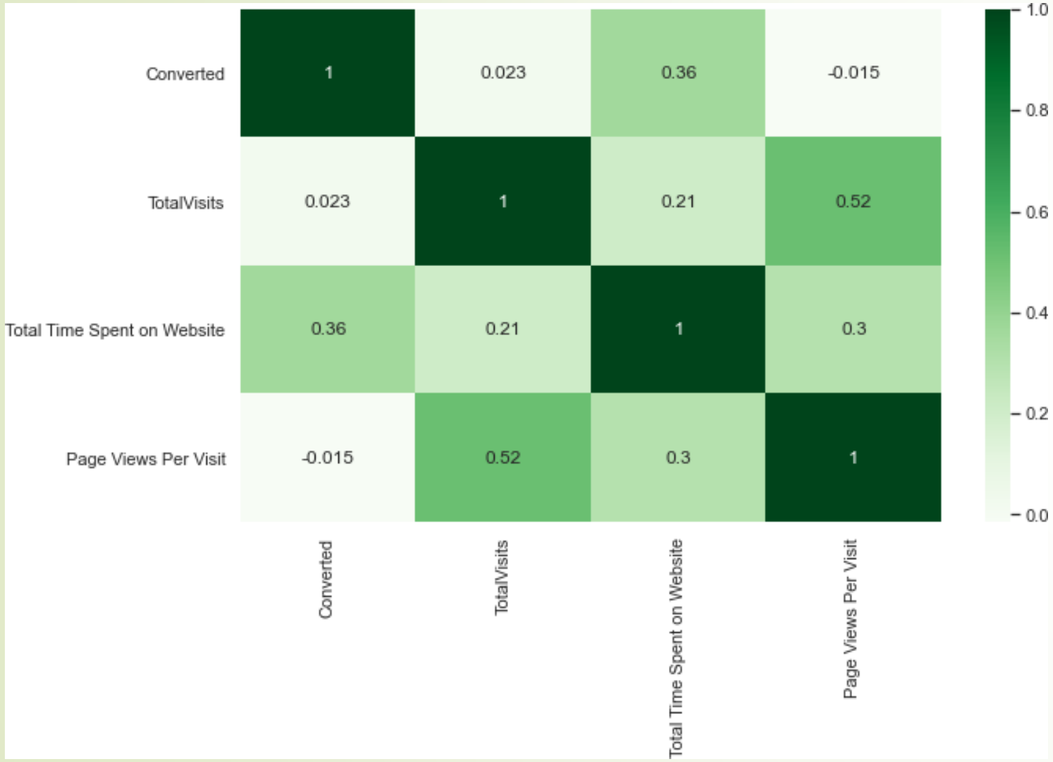
- **Majority of the people(approx. 92%) are fine with receiving email.**
- **People who are ok with email has conversion rate of 40%**
- **People who have opted out of receive email has lower number of records and also have rate of conversion (only 15%)**

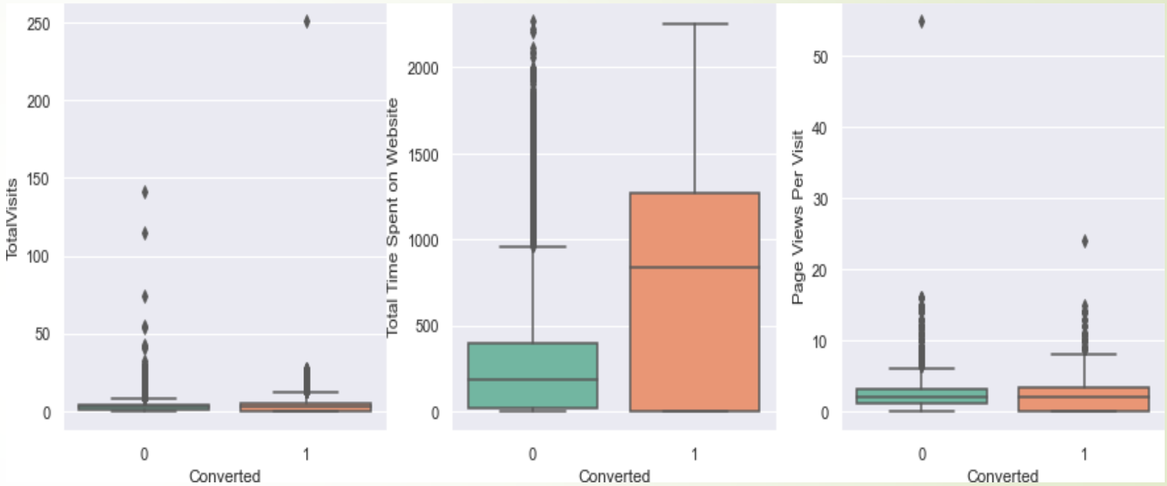
Last Activity



- **"Email opened" is the last activity for most of the leads with conversion rate 38%.**
- **"SMS Sent" is the second highest last activity with Conversion rate of around 62%**
- **We can considering all other smaller Last Activity types as Other Activity.**

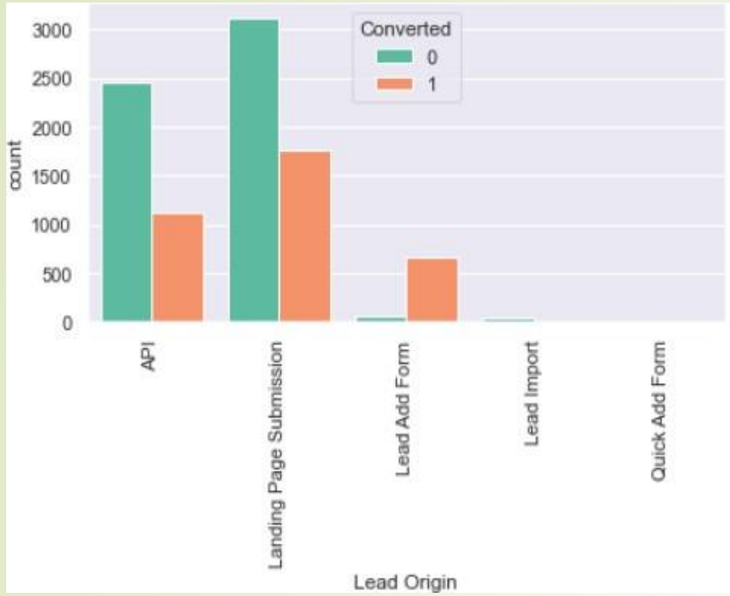
Univariate Analysis - Numerical

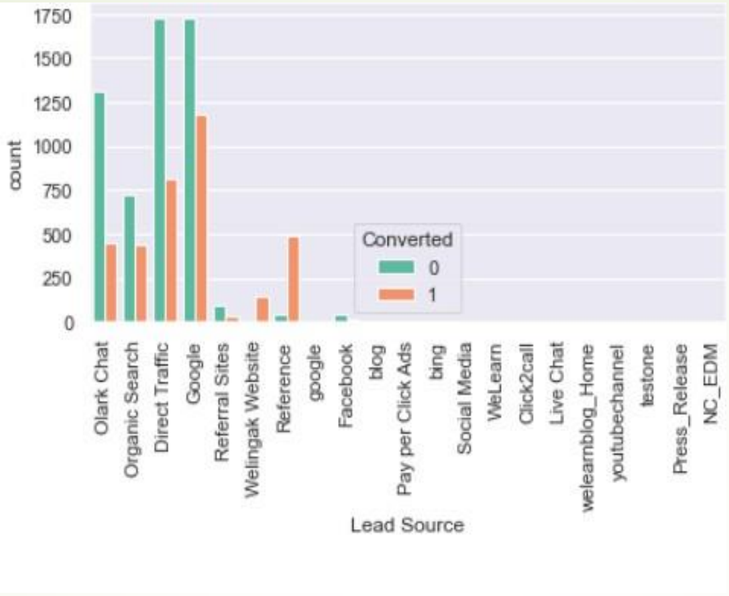


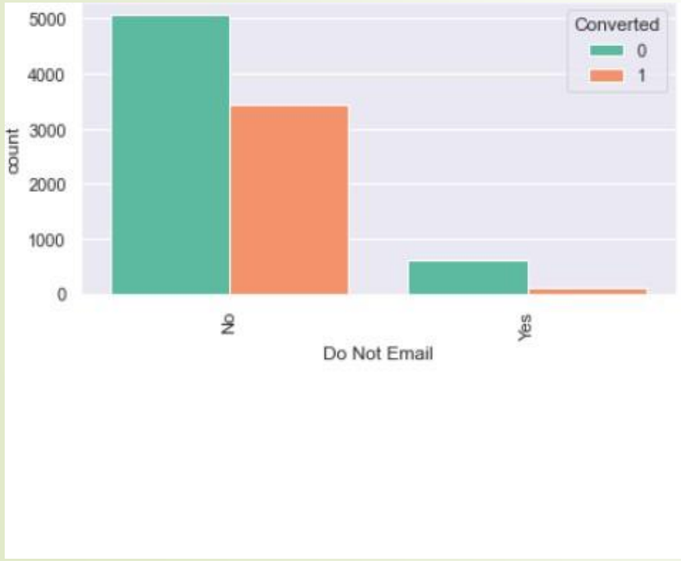


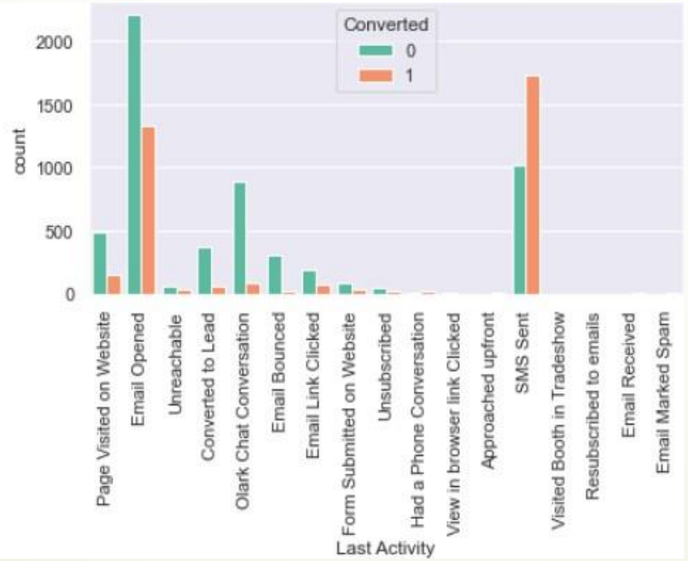
Insights:-

- TotalVisits: It has some outliers which needs to be treated.
- Total Time Spent on Website: People whose spend more time has higher chance of getting converted.
- Page Views Per Visit: It has some outliers which needs to be treated.





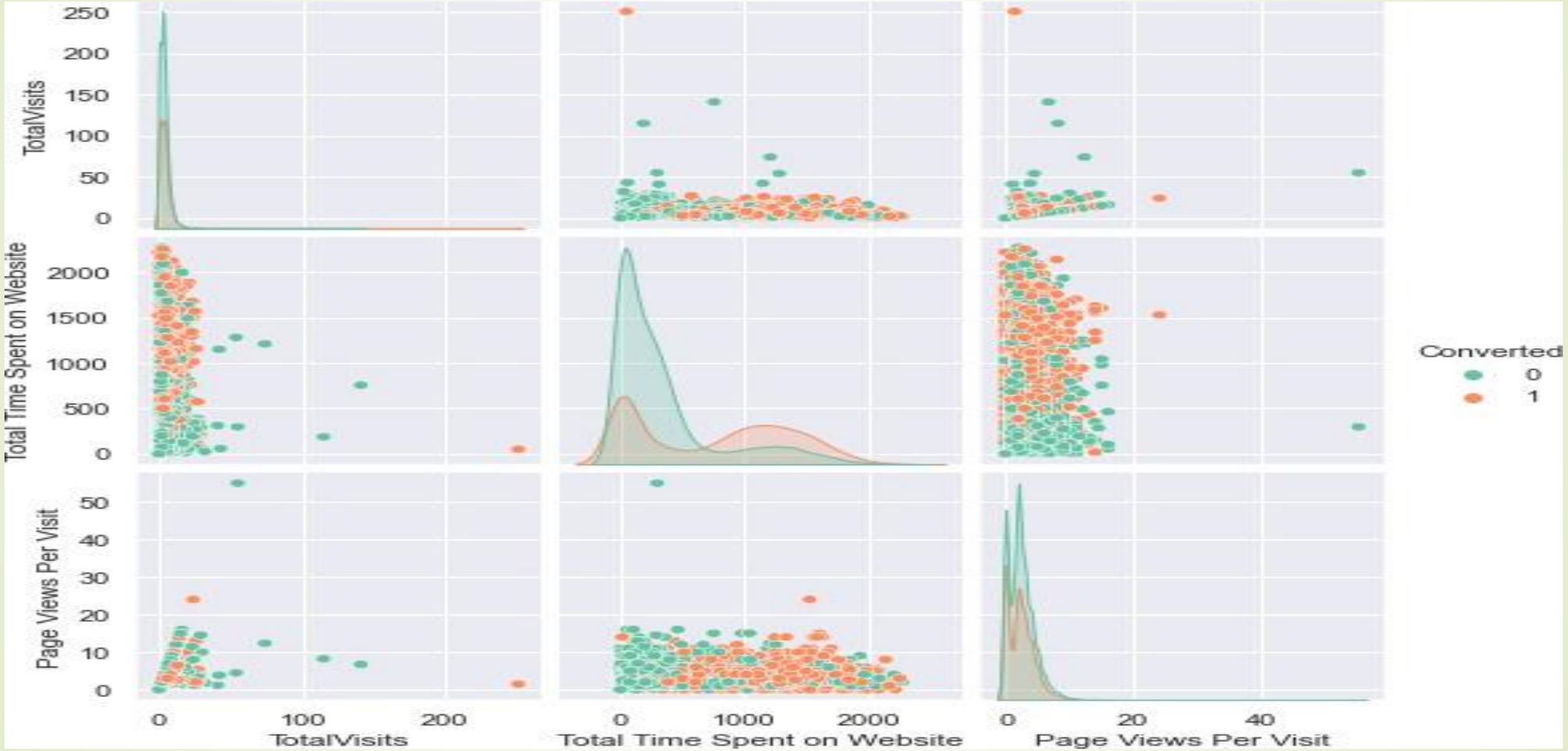




Bivariate Analysis (Categorical)

- Lead Origin: Higher leads in "Landing Page Submission" and "API" category
- Lead Source: leads are higher in "Direct Traffic" and "Google" Category
- Do not email: No has higher converted as well as non- converted population
- Last Activity: The number of Hot leads is higher in SMS and in EMAIL category.

Bivariate Analysis - Numerical

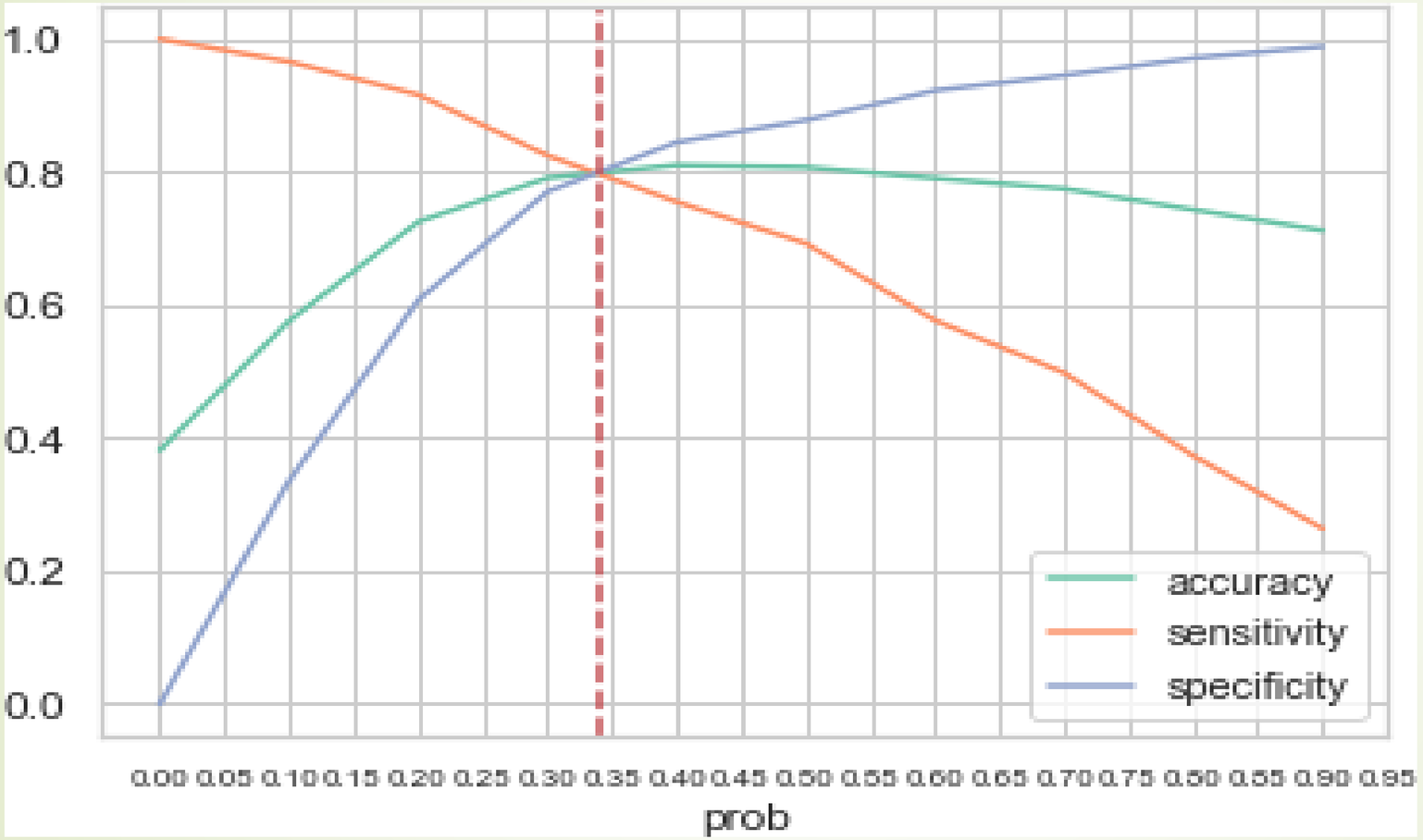


Insights:-

- Data is not normally distributed.
- Total Visits and Page Views Per Visit has positive correlation among each other.

Model Evaluation

Optimal Cutoff Point



Insights:-

Optimal cut-off probability is that probability where sensitivity and specificity and accuracy meet. We are getting cut-off of 0.34

Specificity = TN/(TN+FP) Sensitivity = TP/(TP+FN) Accuracy = (TP+TN)/(TP+TN+FP+FN)

Confusion Matrix and Logistic Regression Metrics

Train

```
################################
  Confusion Matrix
[[3225  777]
 [ 492 1974]]
################################
True Negative                    :  3225
False Positve                    :  777
False Negative                   :  492
True Positive                    :  1974
Model Accuracy value is          :  80.38 %
Model Sensitivity value is       :  80.05 %
Model Specificity value is       :  80.58 %
Model Precision value is         :  71.76 %
Model Recall value is            :  80.05 %
Model True Positive Rate (TPR)   :  80.05 %
Model False Positive Rate (FPR)  :  19.42 %
################################################
```

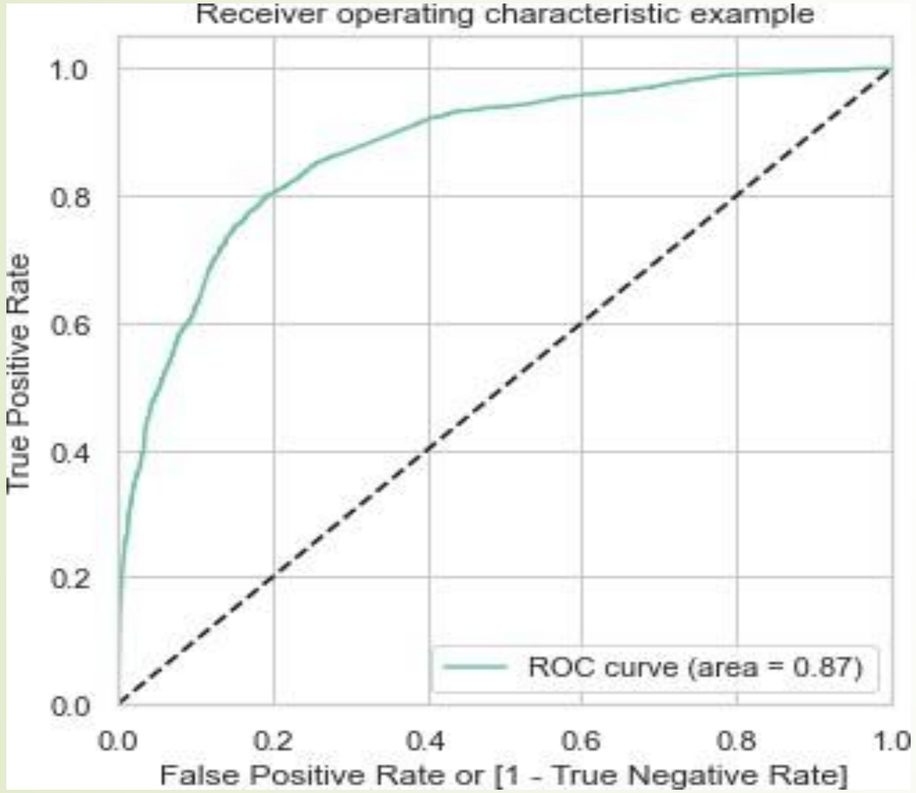
Test

```
################################
  Confusion Matrix
[[1358  319]
 [ 217  878]]
################################
True Negative                    :  1358
False Positve                    :  319
False Negative                   :  217
True Positive                    :  878
Model Accuracy value is          :  80.66 %
Model Sensitivity value is       :  80.18 %
Model Specificity value is       :  80.98 %
Model Precision value is         :  73.35 %
Model Recall value is            :  80.18 %
Model True Positive Rate (TPR)   :  80.18 %
Model False Positive Rate (FPR)  :  19.02 %
################################################
```

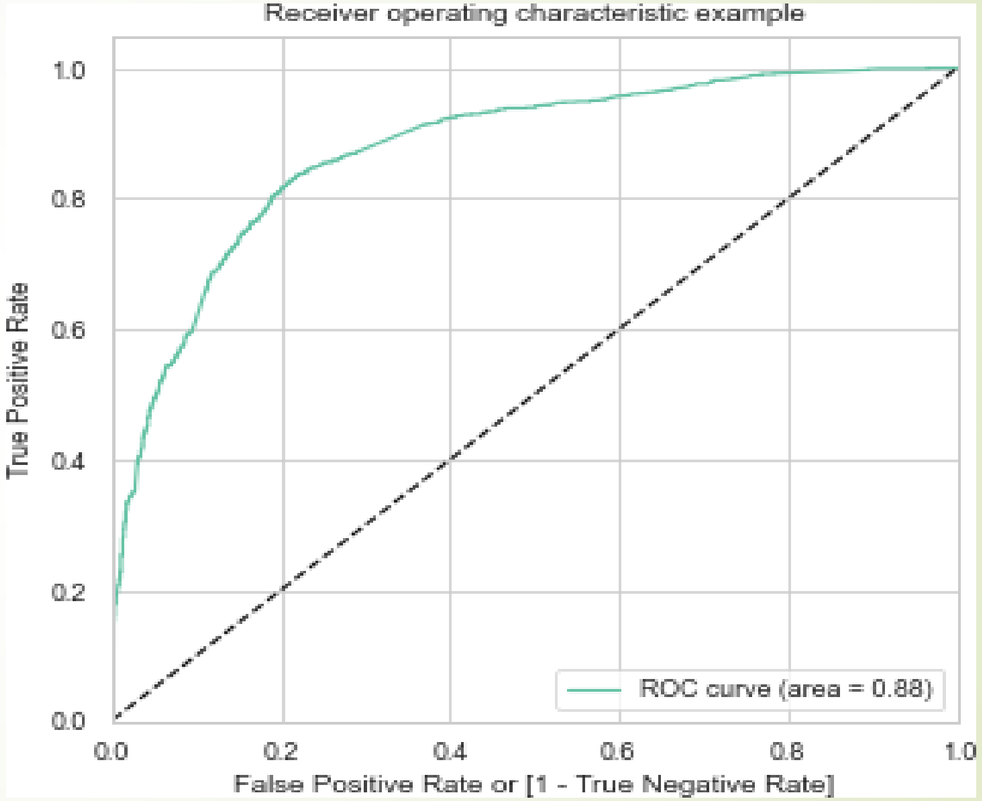
Using cut off value at 0.34 Sensitivity of 80.05% in Train and 80.18% in Test. Sensitivity in this case indicates how many leads the model identify correctly out of all potential leads which are converting. More than 80% is what the CEO has requested in this case study. Accuracy is also 80.38% which is also good.

ROC Curve

Train

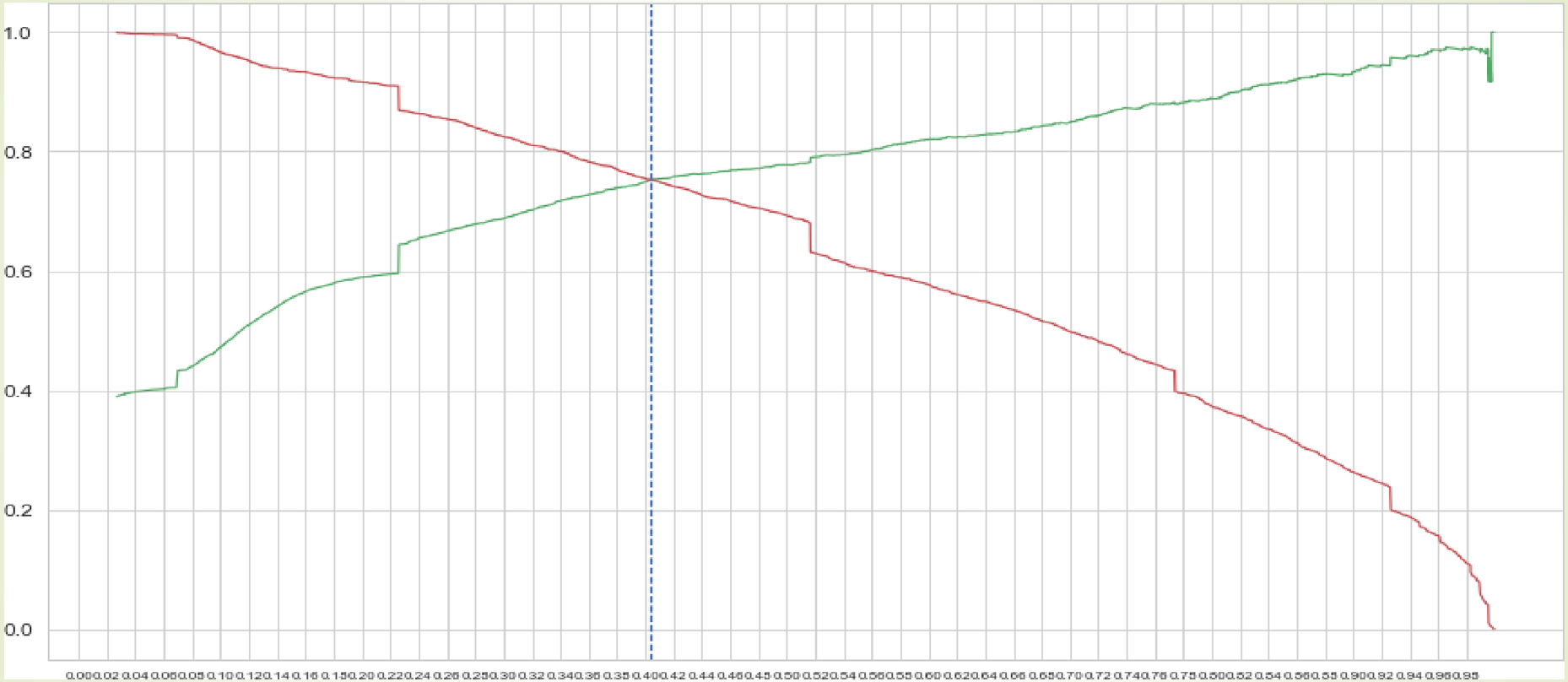


Test



ROC Curve area is 0.87 for Train and 0.88 for Test model, which indicates that the model is good because the closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate is it.

Precision – Recall Trade-off



Based on Precision-Recall Trade off curve, the cut-off point is0.404. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%. He wants people to be correctly identified as leads with 80% success i.e. True Positive , True Positive Rate ,Sensitivity, Recall should be close to 80%, which we are getting using the previous cut off of 0.34. These number decreased by using 0.404 cut-off . Hence we will go for 0.34 cut-off.

Conclusion

- Leads from the "Lead Add Form" have third highest conversions with conversion rate. Hence we should try to put Lead Add Forms on the social media websites specially on the Welingak Website and we should give more importance to customers you came through this channel.
- More focus should be given in engaging with the Working professionals because of high conversion rate
- More adds should be given on Welingak Website to cater the leads from their, as it has higher chance to conversion.
- Leads that came through a "reference" has over 90% conversion, we should encourage and incentivize existing members to bring more of their referrals.