Lead Scoring Case Study

Shuchi Singh
Manish Singh
Shriya Chintawar

Problem Statement

- An Education company named X Education sells online courses to industrial professionals.
- Once they receive the leads, sales team will start making calls, writing emails. Throughout the process, some leads will convert and some may not.
- They get a lot of leads but their lead conversion rate is very poor. For ex. If they receive 100 leads in a day, only 30 of them are converted.
- If they want more leads to be converted, they should start focusing more on communicating with potential leads rather than making calls to everyone.

Business Goal

- The company wants to know most promising leads.
- For this they want to build model which will identify the hot leads.

Problem Approach

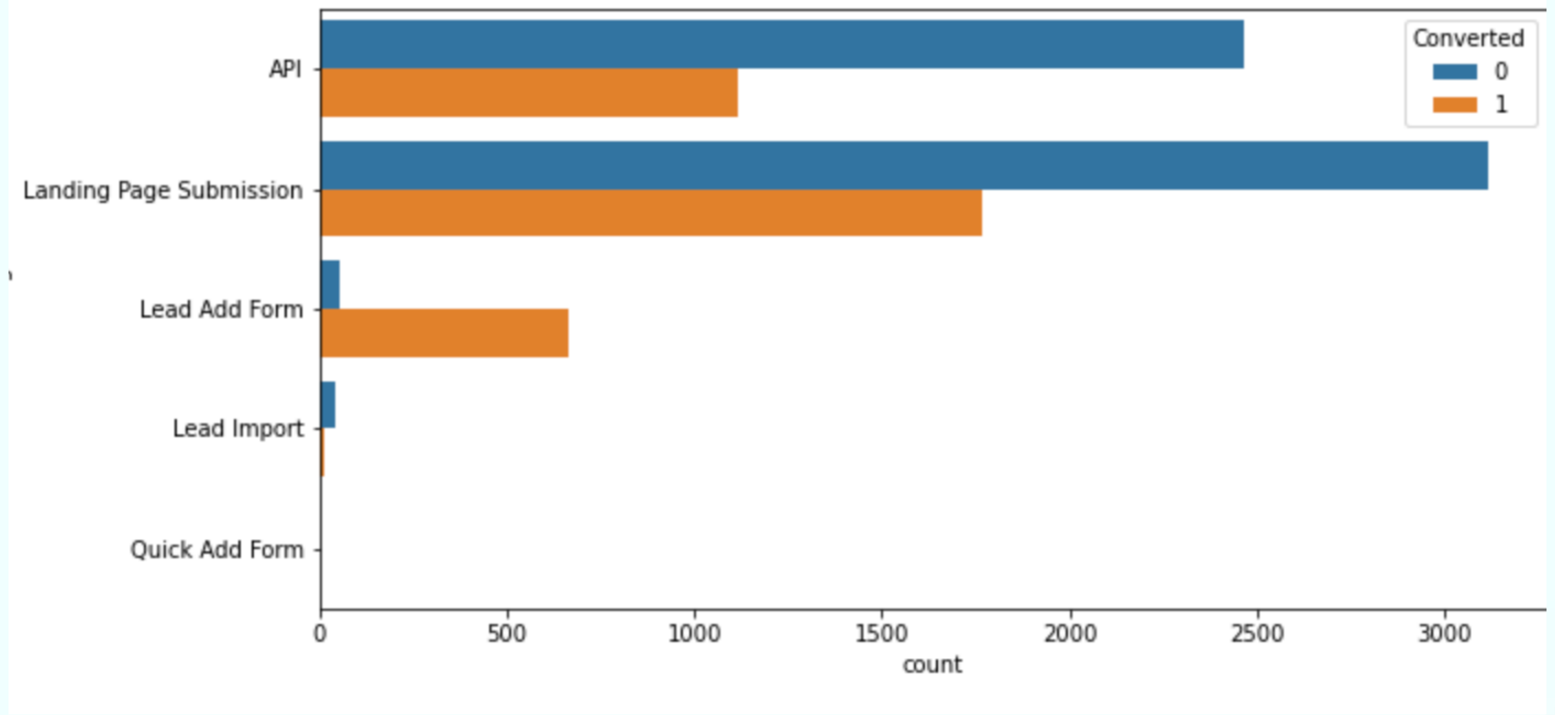
- First we will start with data cleaning by checking missing values and handle duplicate data, dropping columns, handling outliers.
- Dummy variable creation,Feature scaling
- Test-train split, Correlations
- Model Building
- Model Evaluations
- Conclusion

Understanding Data

- There are total no of 37 rows and 9240 columns.
- 16 columns have been dropped .

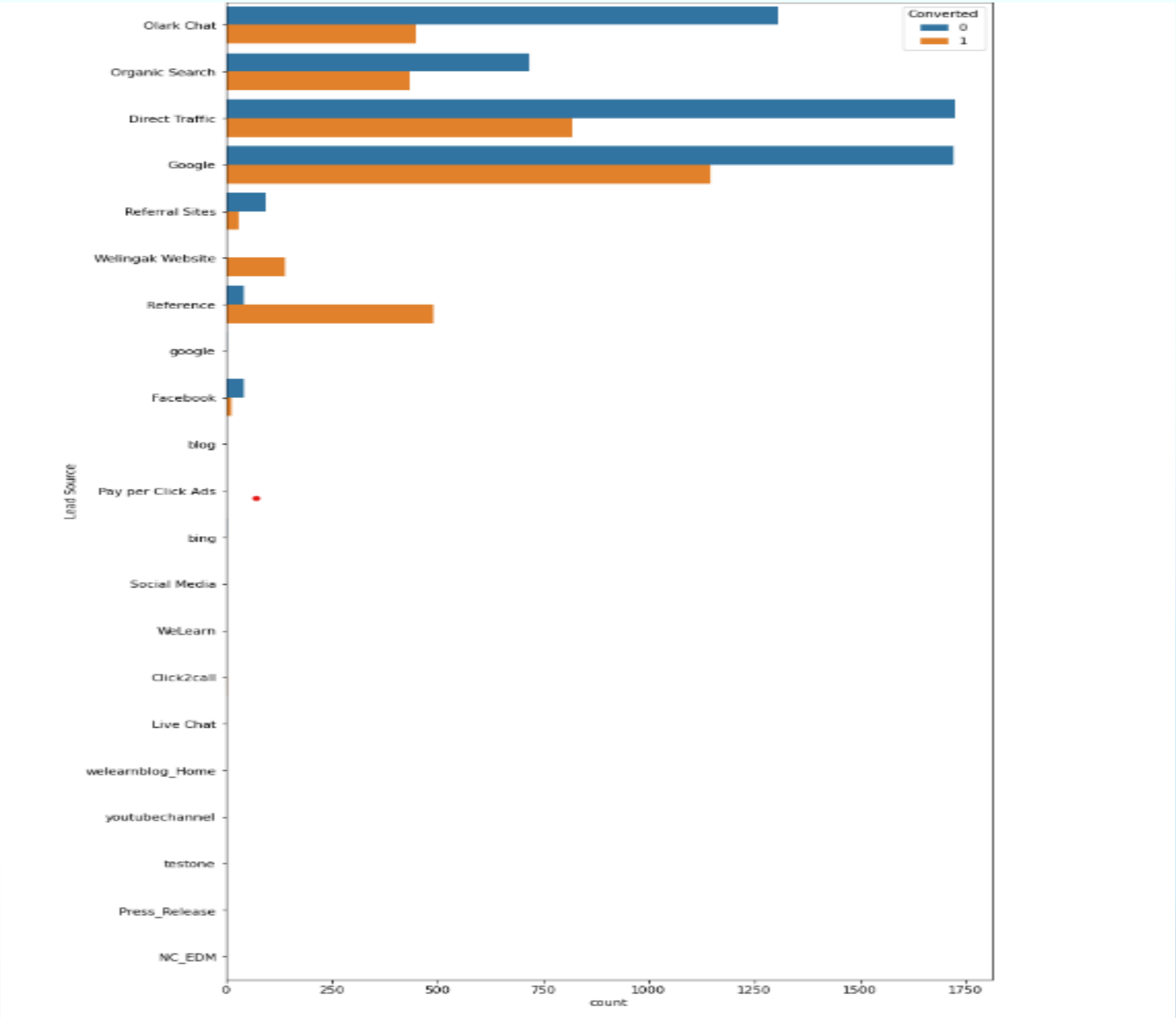
Lead Conversion and Lead Source

- This graph indicates the Lead conversion from lead Origin who are converted and non-converted.



Lead Conversion and Lead Source

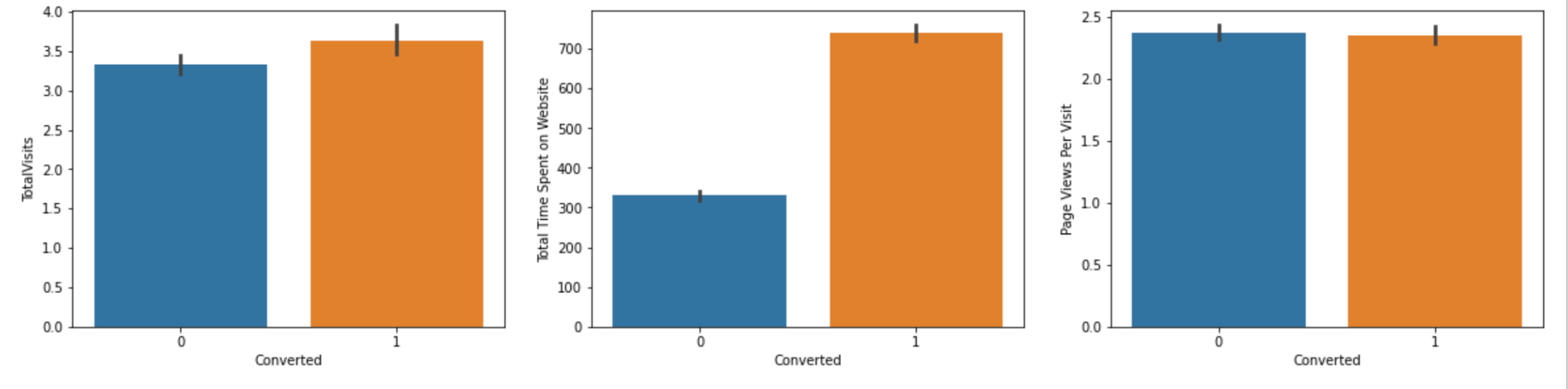
- This graph indicates the Lead conversion from lead Source who are converted and non-converted.

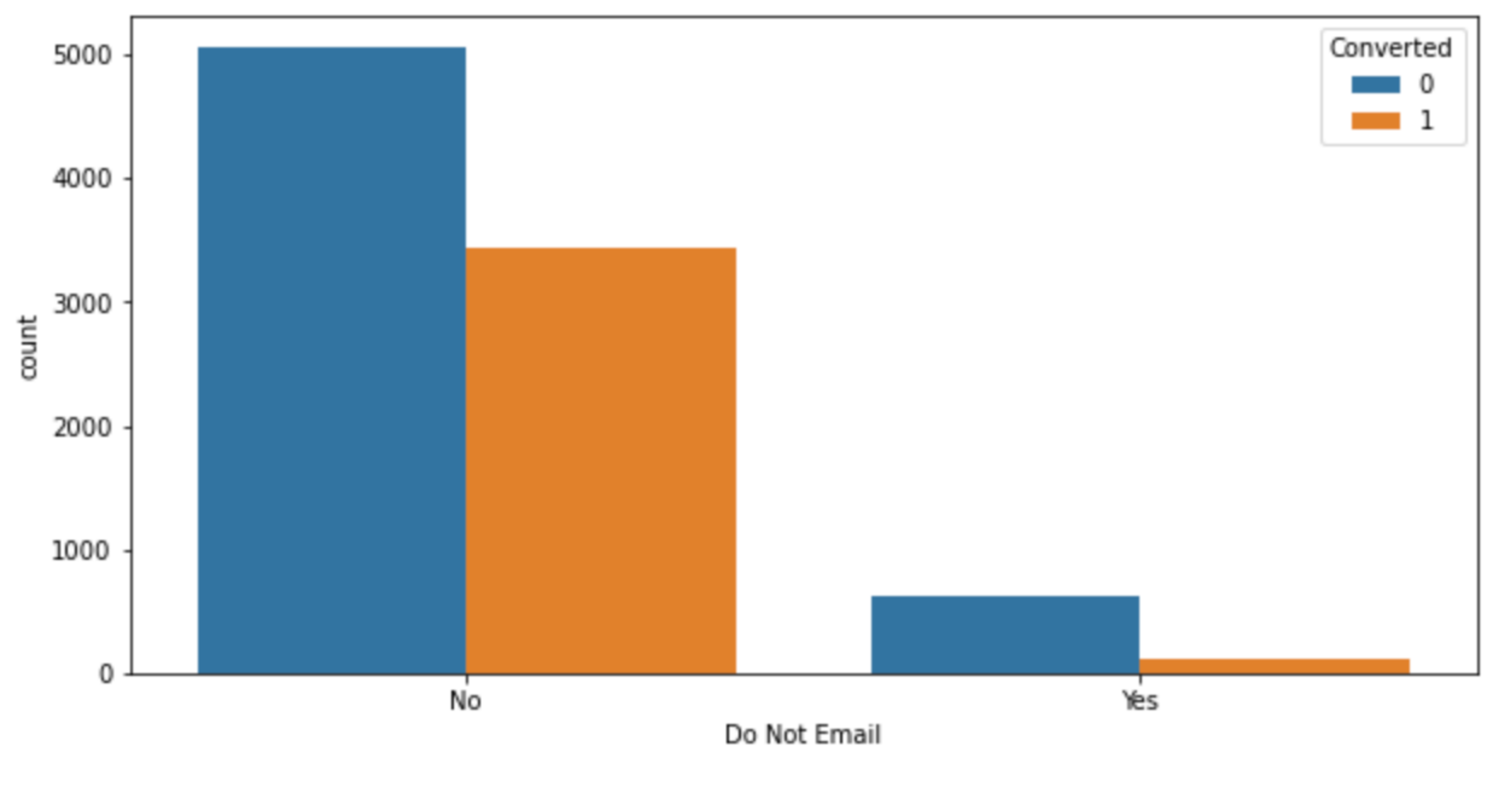


Total Visits, Total Time Spent on Website, Page Views

- Here we can see in first graph the Total Visits, Total Time Spent on Website, Page Views per visit which are converted and not converted.

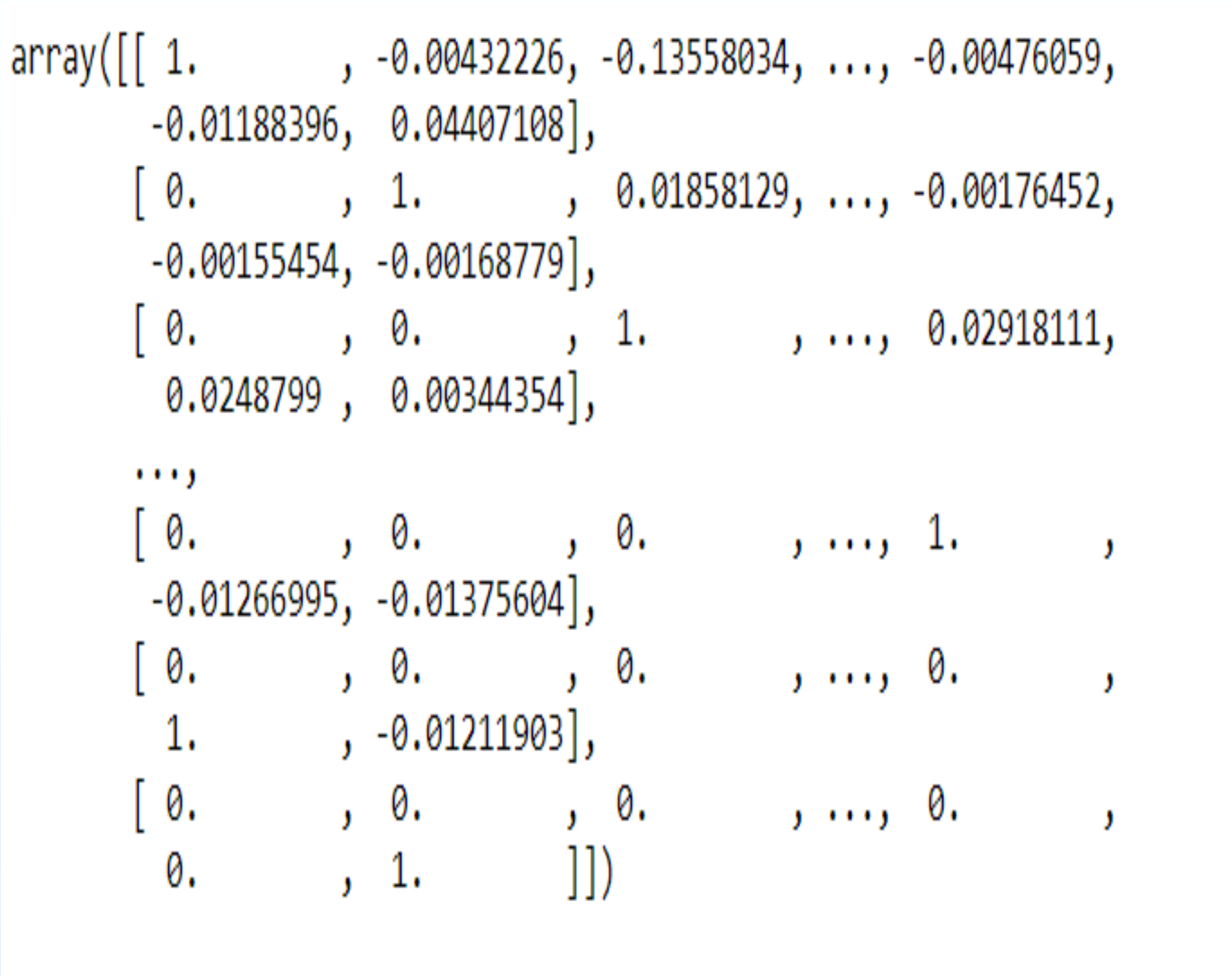
- In second graph Lead conversion customer who has mailed , converted and vice versa.



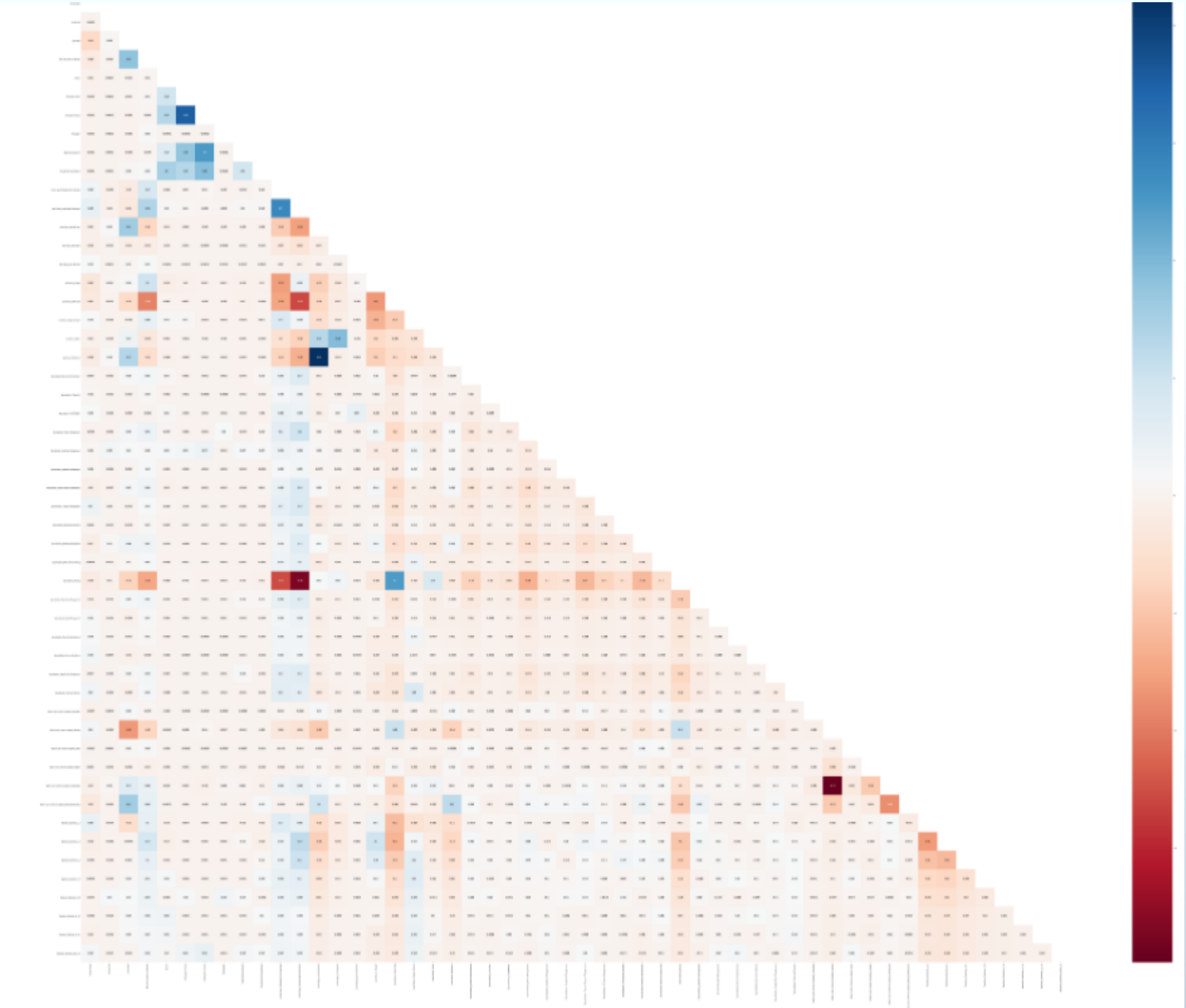


Correlation

- Here we can see the correlation between the variables. Light shaded indicates that there is less correlation.



```
array([[ 1.        , -0.00432226, -0.13558034, ..., -0.00476059,
        -0.01188396,  0.04407108],
       [ 0.        ,  1.        ,  0.01858129, ..., -0.00176452,
        -0.00155454, -0.00168779],
       [ 0.        ,  0.        ,  1.        , ...,  0.02918111,
         0.0248799 ,  0.00344354],
       ...,
       [ 0.        ,  0.        ,  0.        , ...,  1.        ,
        -0.01266995, -0.01375604],
       [ 0.        ,  0.        ,  0.        , ...,  0.        ,
         1.        , -0.01211903],
       [ 0.        ,  0.        ,  0.        , ...,  0.        ,
         0.        ,  1.        ]])
```

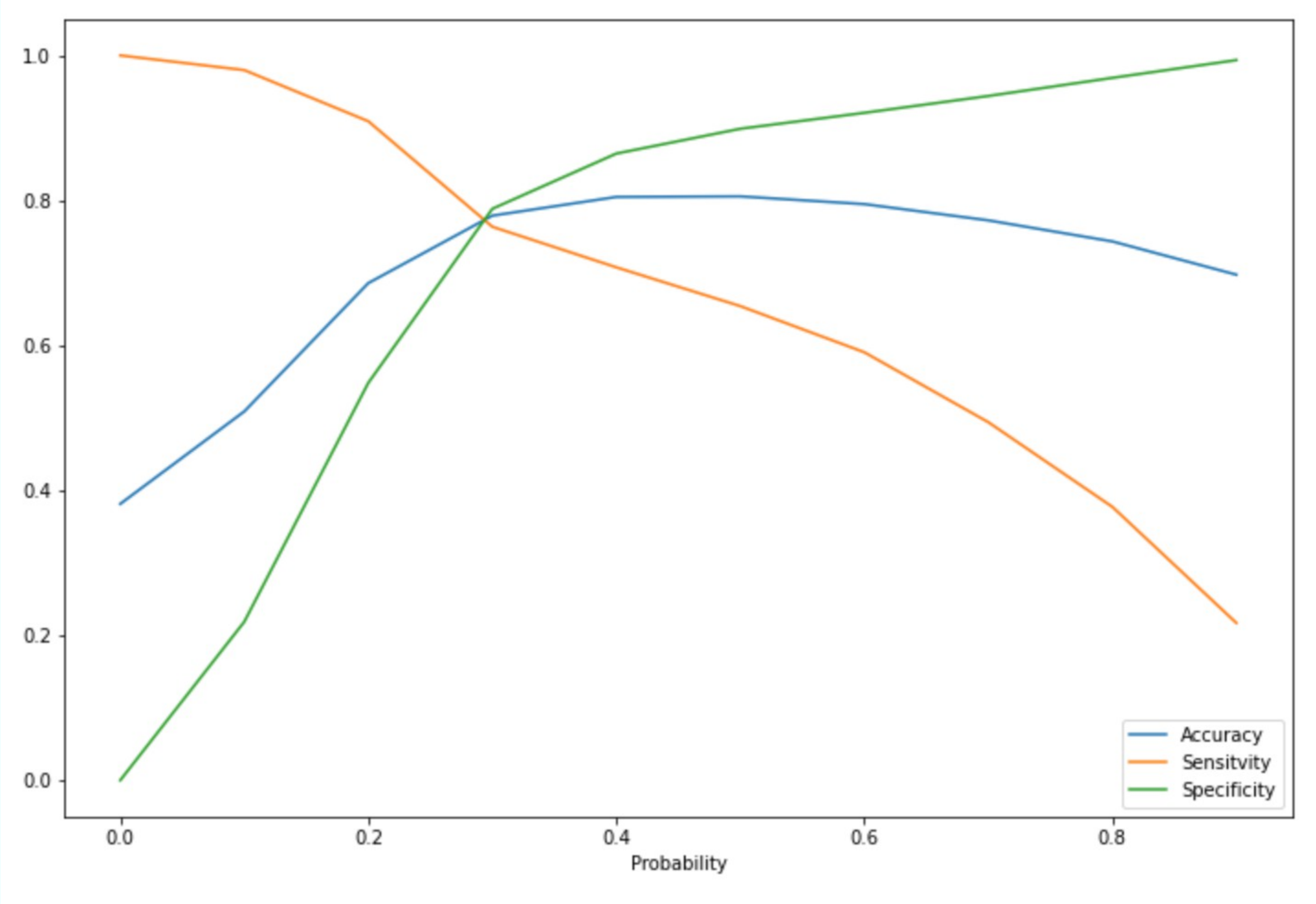


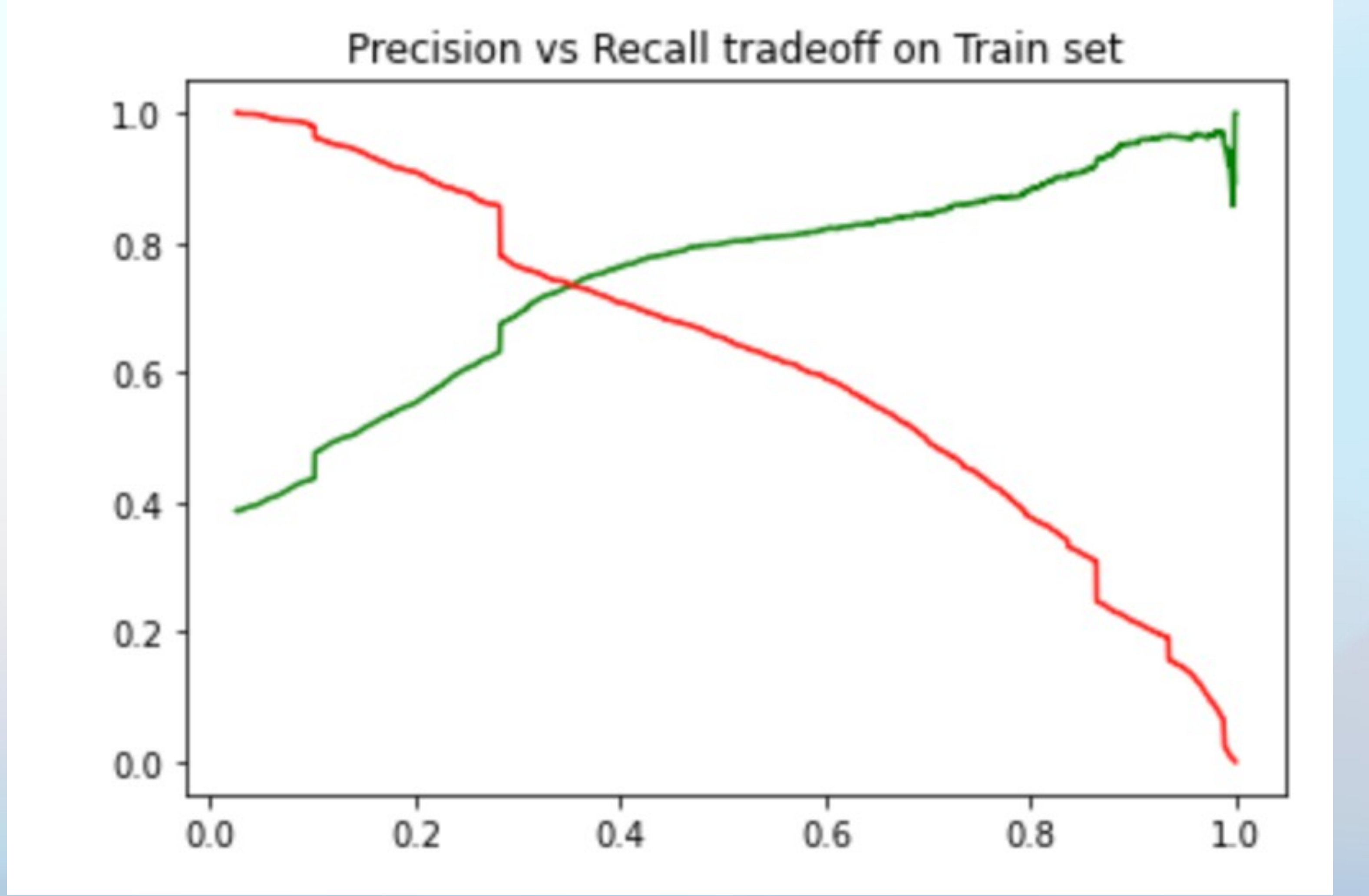
Model Evaluation

- Splitting the data into train and test data sets in ratio of 70:30.
- Using RFE for feature selection.
- Running RFE with 15 variables.
- Building model by removing the variable whose p-value is greater than 0.05 and VIF is greater than 5.
- Prediction on test data set.
- Overall accuracy is around 77%

Model Evaluation

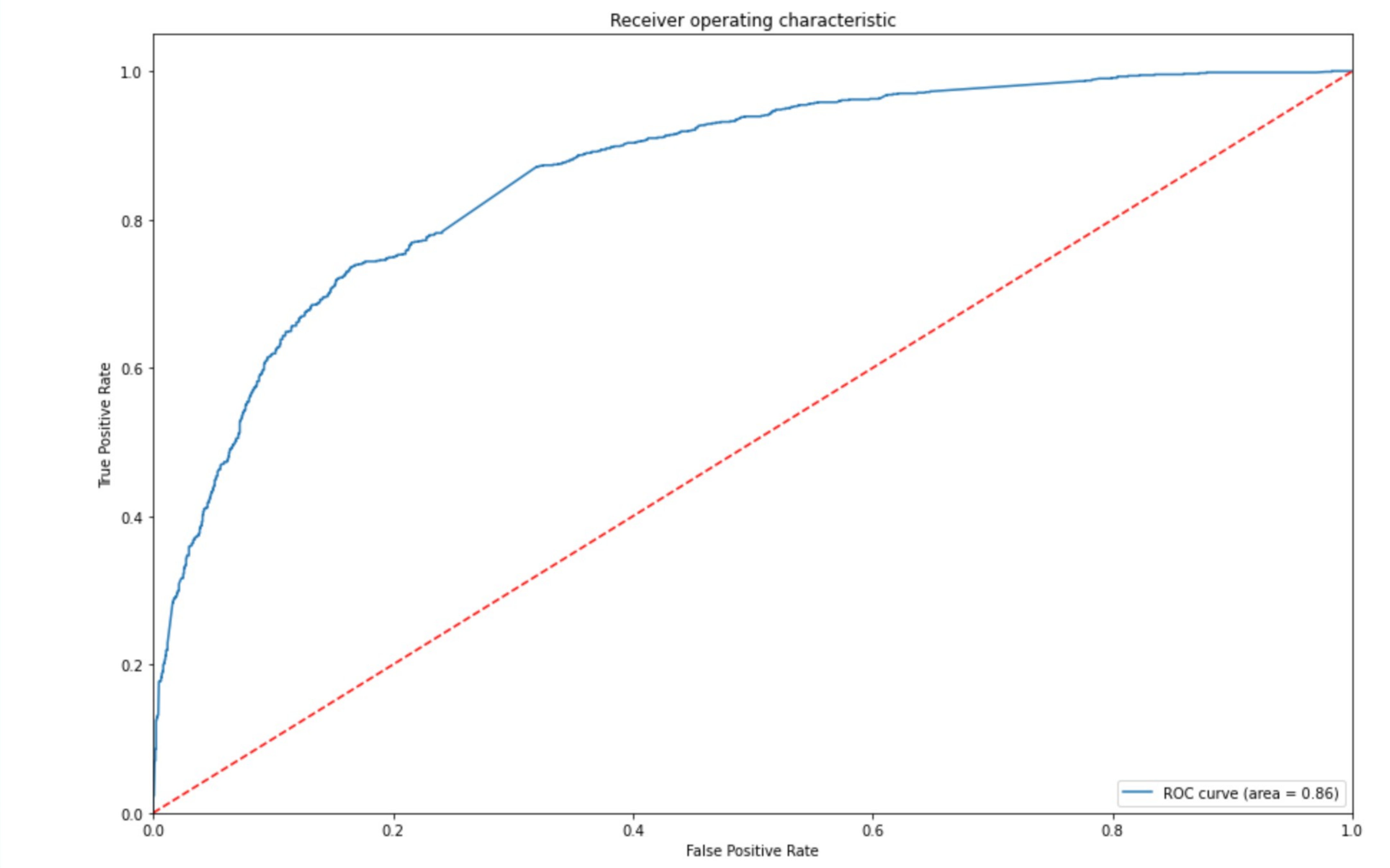
- 0.38 is the tradeoff between precision and recall.
- We can say that 38% is the probability for the lead to be hot lead.




Precision vs Recall tradeoff on Train set

ROC Curve

- Calling ROC curve function for plotting TP and FP .



Observations

- **Train Data**
- Sensitivity : 76.36
- Specificity : 78.84
- Precision : 68.97
- Recall : 76.36
- Accuracy : 77.89
- **Test Data**
- Sensitivity : 77.08
- Specificity : 77.58
- Precision : 69.18
- Recall : 77.08
- Accuracy : 77.38

Conclusion

- The accuracy we got from test data is 77% approximately and therefore we can consider it as accurate.
- High recall score than precision score is a sign of good model.
- Leads who spent more time on website is more likely to convert.
- People spending higher than average time can be hot leads, so targeting them can be helpful in conversions.
- When the current occupation is working professional the company has high chance to get a potential buyer which will buy the course.
- Maximum lead conversion happened from Landing Page Submission.
- We can conclude that model is in stable state.