Lead Scoring Case Study

SUBMITTED BY:

ADITYA SATIJA

KHUSHBOO SINGH

Problem statement

- An education company named x education sells online courses to industry professionals. Many professionals lands on their website and browse for courses.
- When these people fill up a form providing their email address or phone number, they are classified as leads. Moreover, the company also gets leads through past referrals.
- Once these leads are acquired , employees from sales team start making calls, writing emails , etc. The typical lead conversion rate is very poor(around 30%)

BUSINESS GOAL:

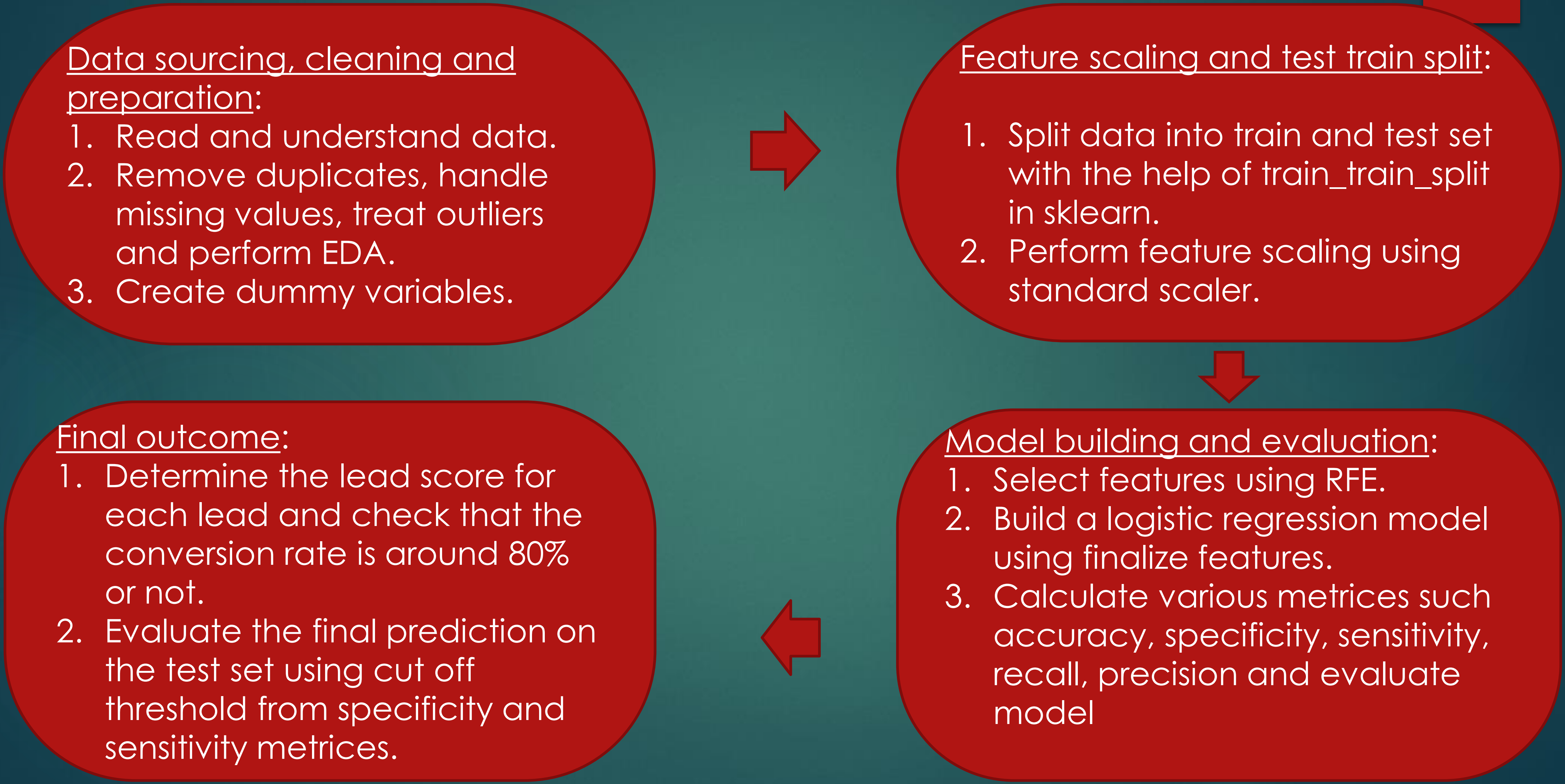
- To make this process efficient, you are appointed to help them select the most promising leads known as hot leads which are more likely to convert into paying customers with the help of model building wherein you need to assign a lead score to each of the leads such that the customers with high lead score has a high conversion chance and the customer with the lower lead score have a lower conversion chance.
- This will make the sales team focus on the potential customers rather than spending time and money on cold leads. Target is to make this lead conversion rate around 80%.

Analysis approach

With the given data we have followed these steps:

1. Reading and understanding data
2. Data cleaning and data preparation
3. Eda- univariate and bivariate analysis
4. Splitting the data into test train split
5. Feature scaling
6. Model building and calculating lead score
7. Model evaluation using metrices
8. Prediction on the test set

Methodology



Data sourcing, cleaning and preparation:

1. Read and understand data.
2. Remove duplicates, handle missing values, treat outliers and perform EDA.
3. Create dummy variables.

Feature scaling and test train split:

1. Split data into train and test set with the help of train_train_split in sklearn.
2. Perform feature scaling using standard scaler.

Model building and evaluation:

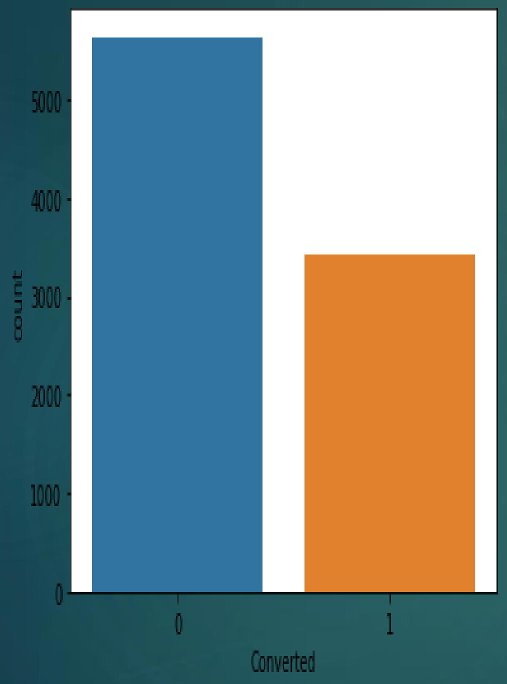
1. Select features using RFE.
2. Build a logistic regression model using finalize features.
3. Calculate various metrices such accuracy, specificity, sensitivity, recall, precision and evaluate model

Final outcome:

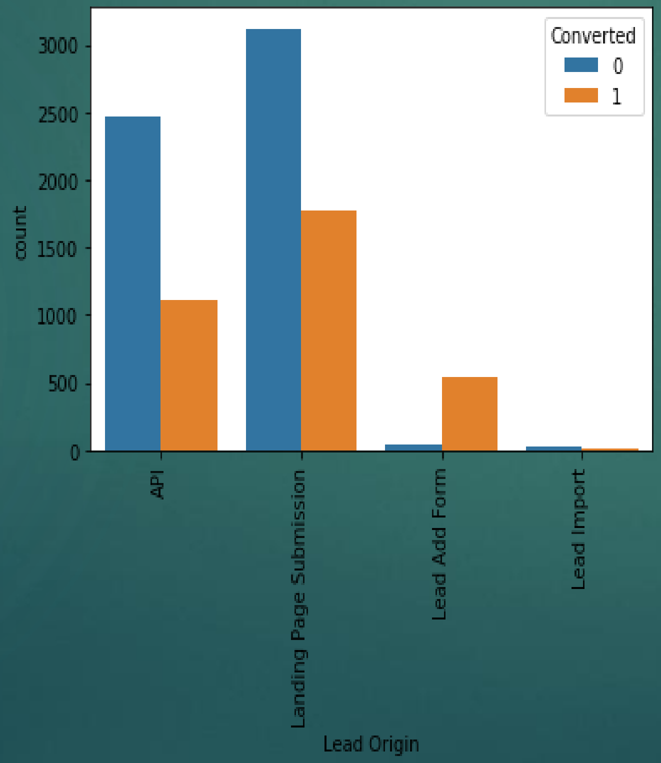
1. Determine the lead score for each lead and check that the conversion rate is around 80% or not.
2. Evaluate the final prediction on the test set using cut off threshold from specificity and sensitivity metrices.

Impact of variables on conversion rate

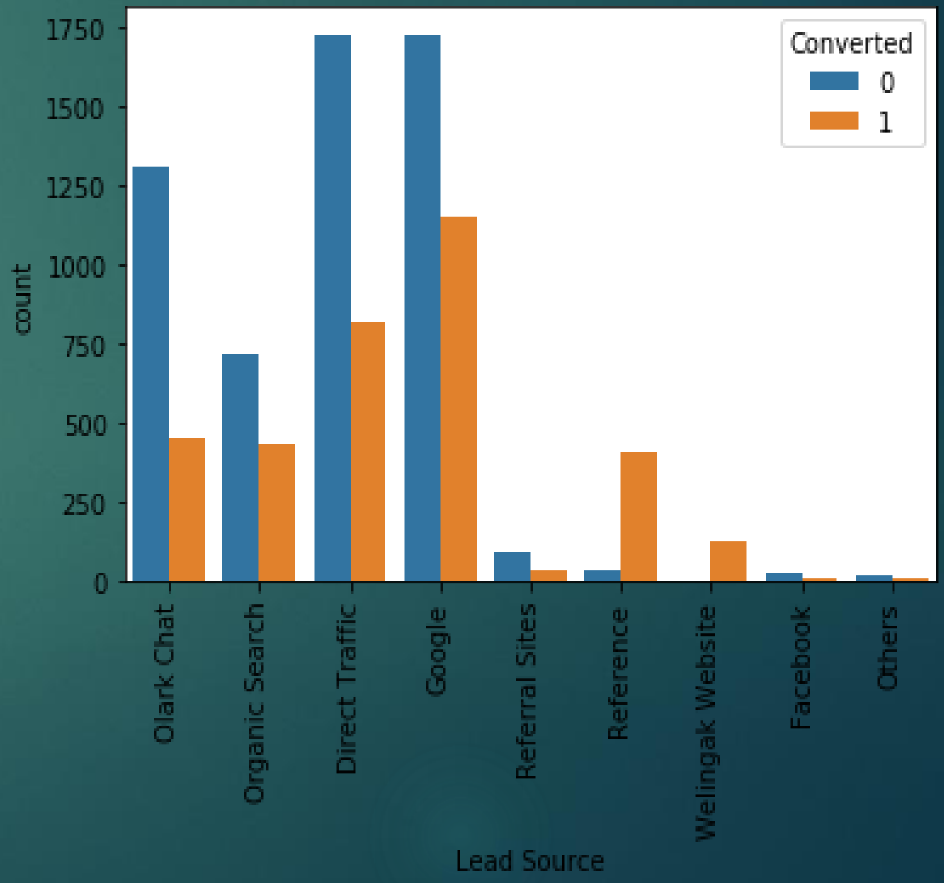
In current data set conversion rate is 38%



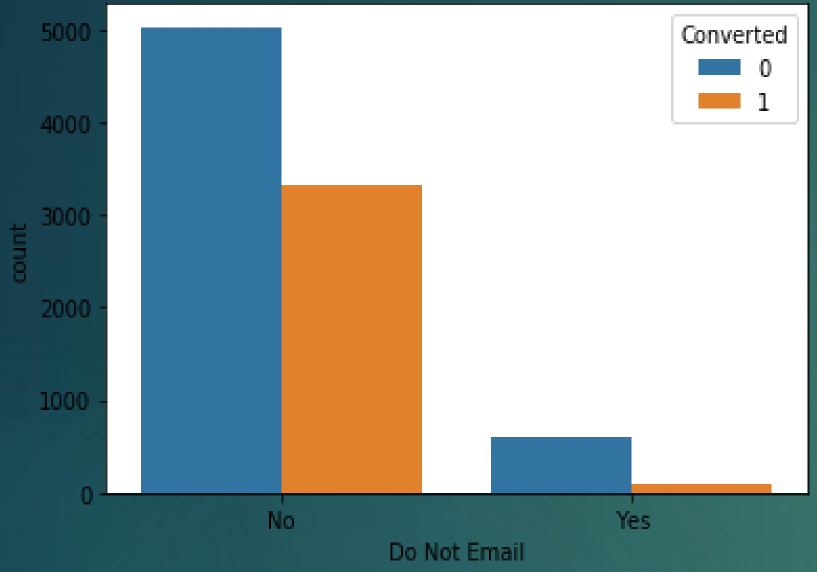
In lead Origin, maximum conversion happen from landing page submission

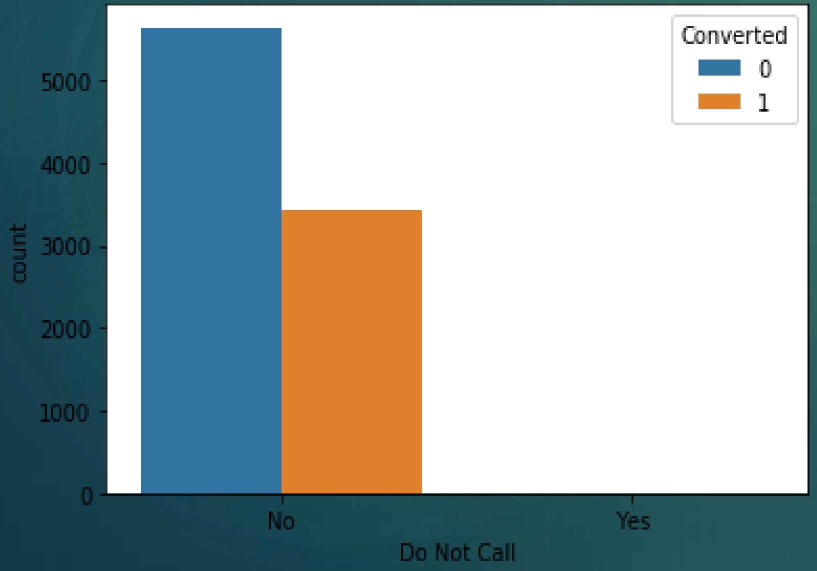


In Lead Source, maximum conversion happen from Google

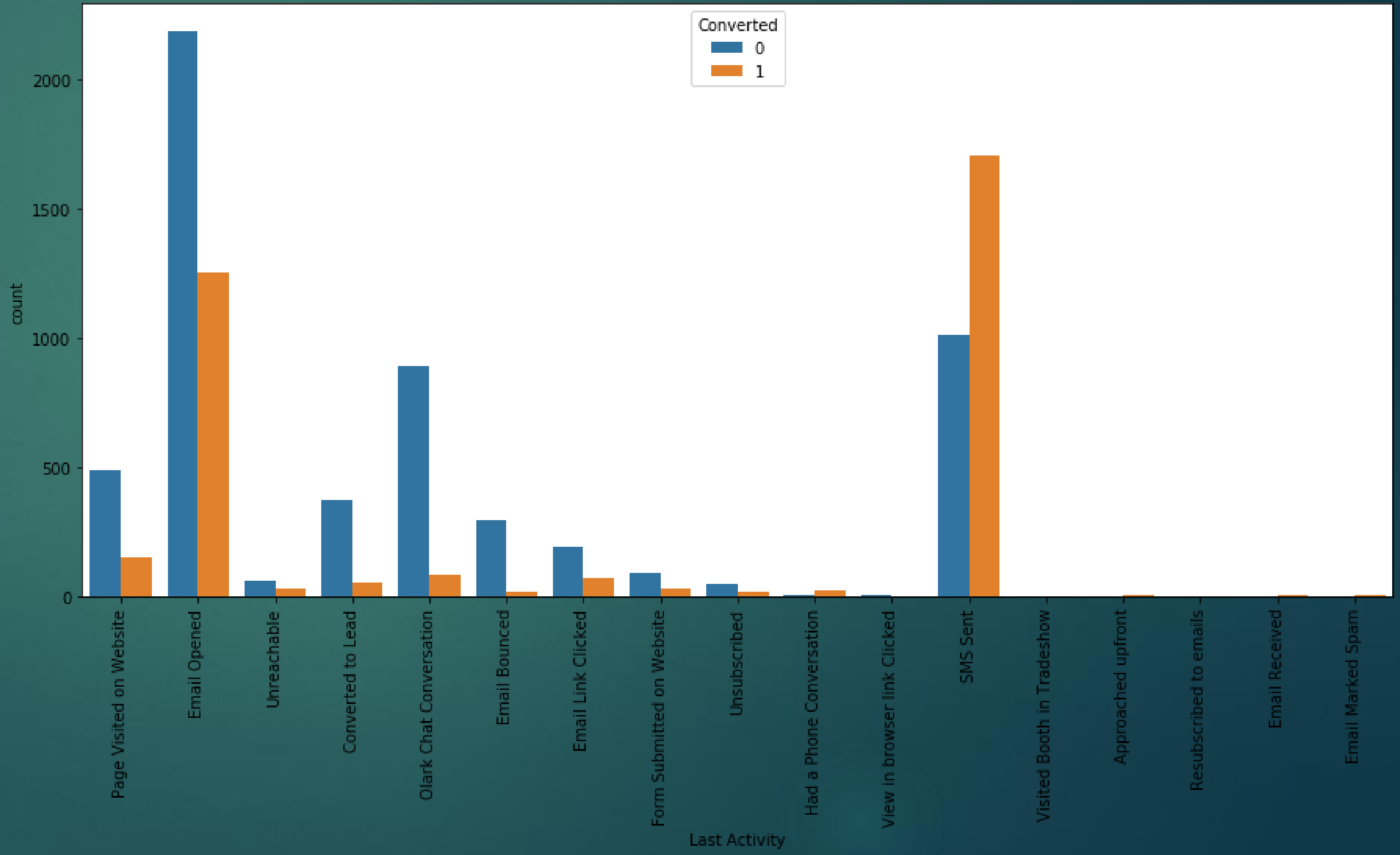


This signifies conversions due to emails and calls.

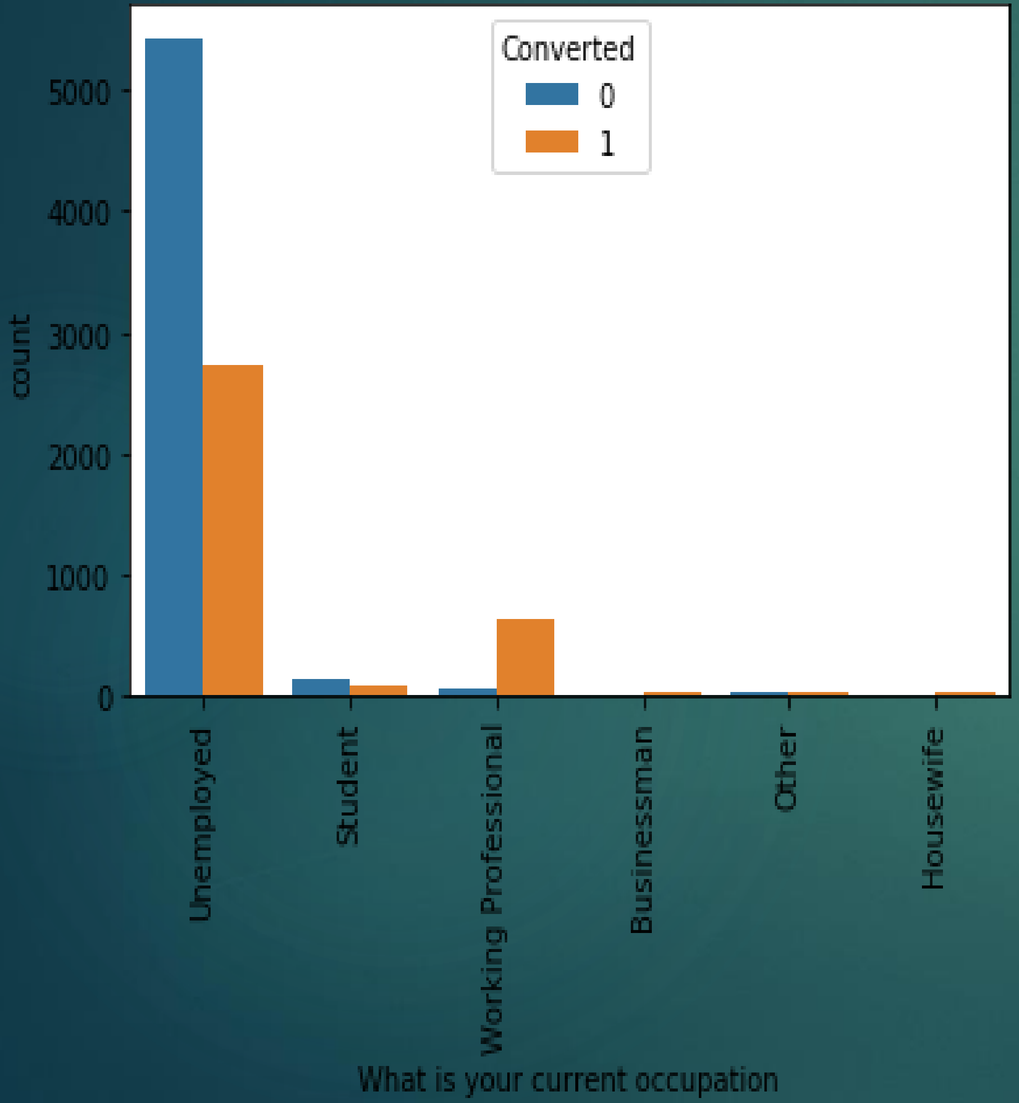




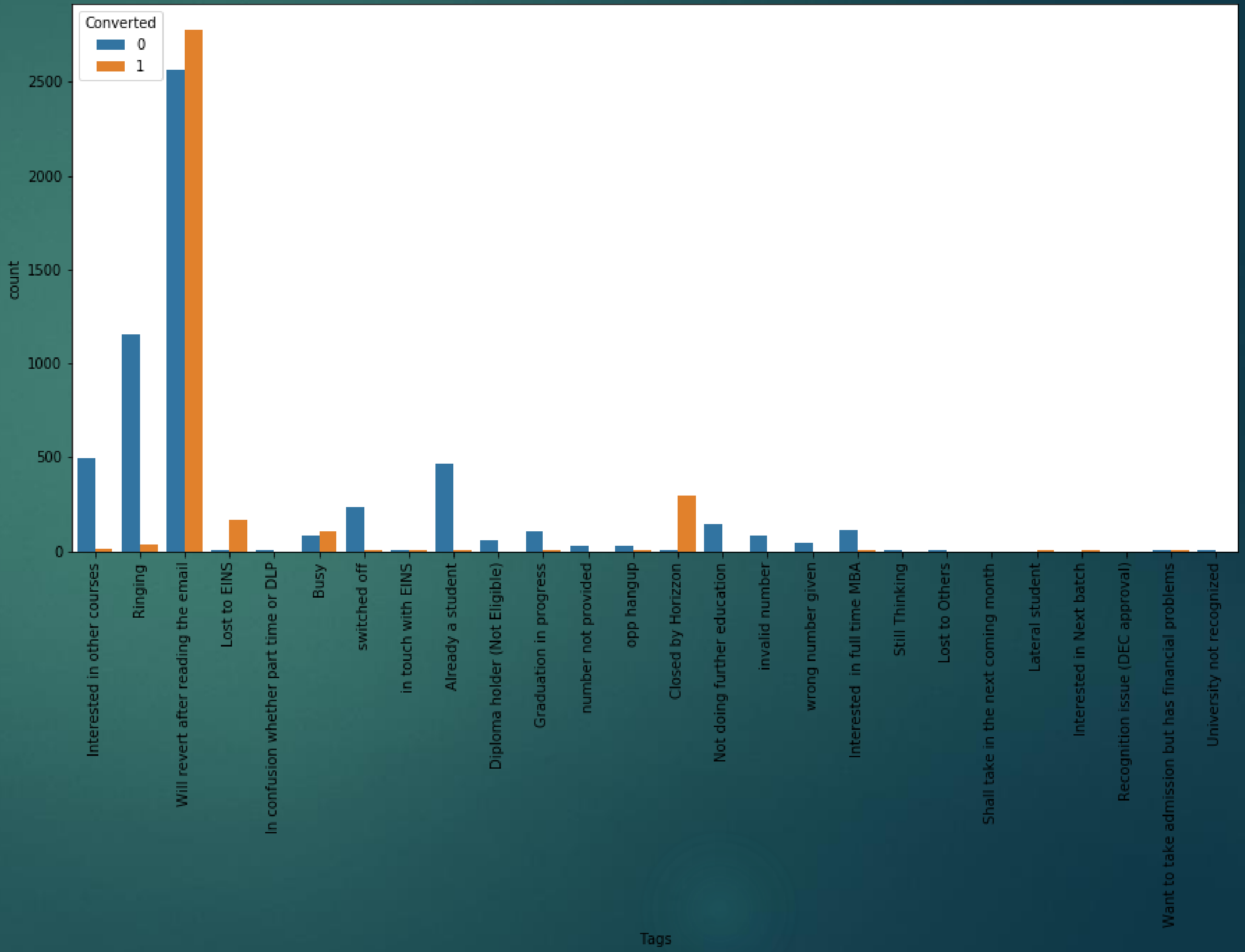
As we can see, SMS Sent has a pretty high conversion of the Last Activity variable.



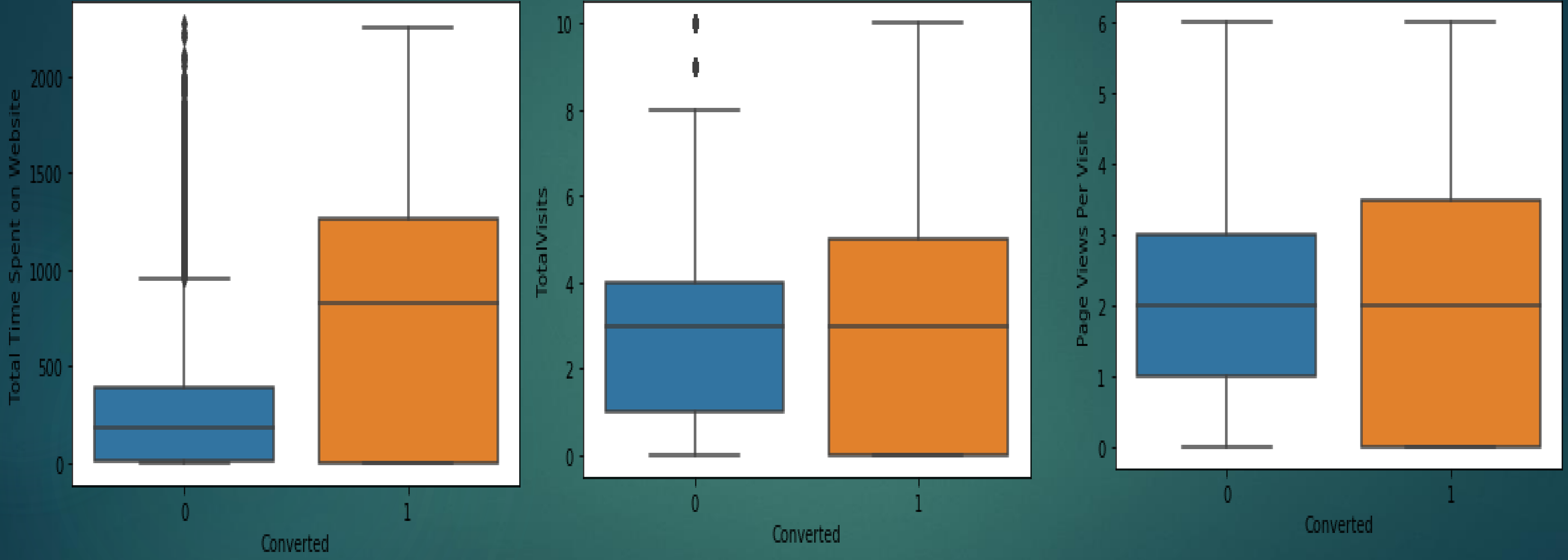
More conversion are happened to people who are unemployed.



Most conversion are for ‘Will revert after reading the email’ tags.

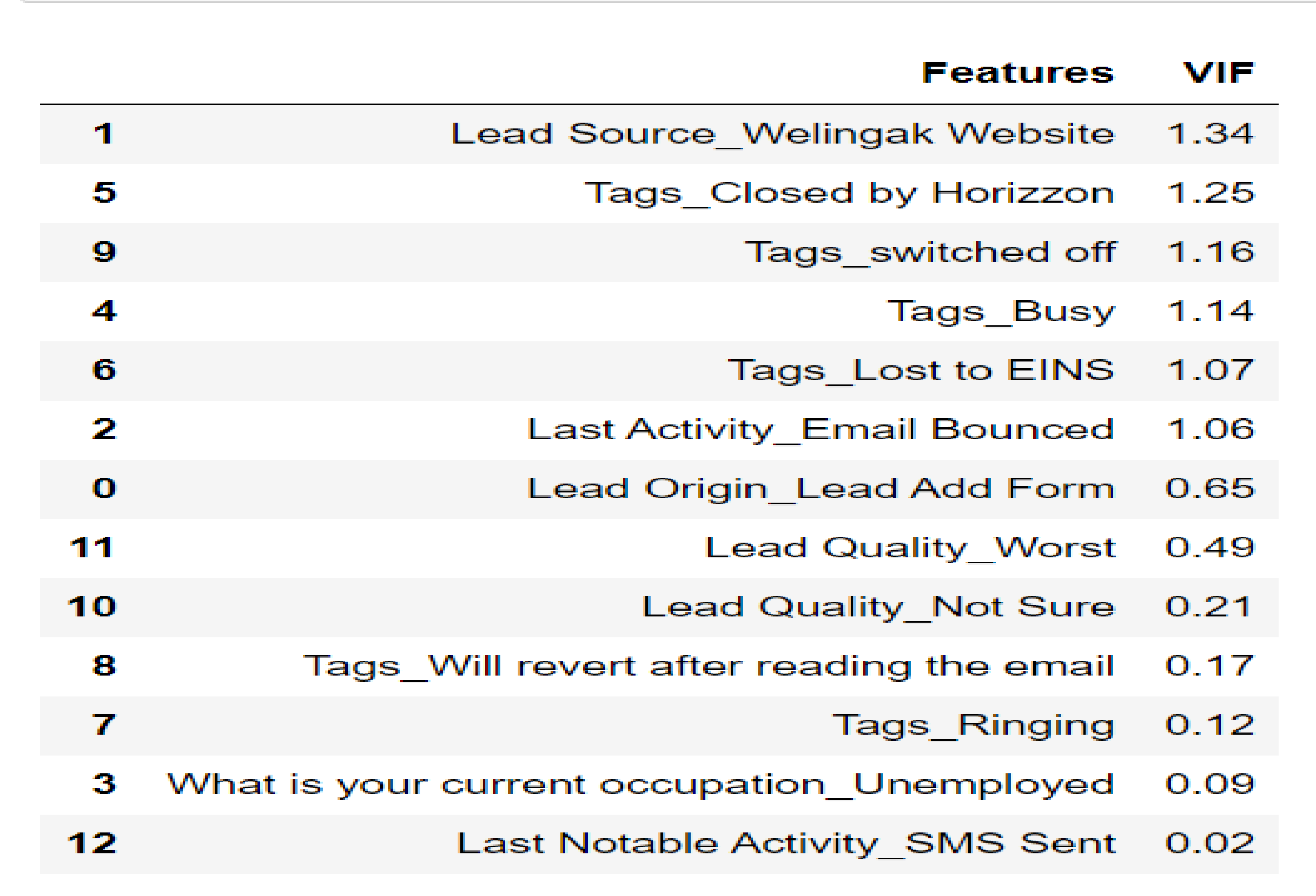


These three variables: Total Time Spent on Website, TotalVisits, Page Views Per Visit have very high conversion rates.



These are the features which significantly affect conversion rate.

Final Features of our Model

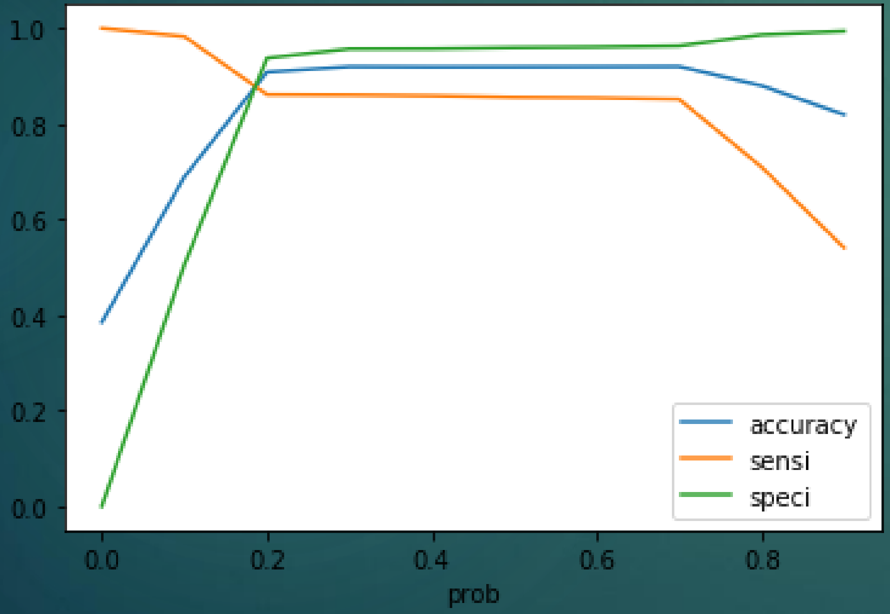


	Features	VIF
1	Lead Source_Welingak Website	1.34
5	Tags_Closed by Horizzon	1.25
9	Tags_switched off	1.16
4	Tags_Busy	1.14
6	Tags_Lost to EINS	1.07
2	Last Activity_Email Bounced	1.06
0	Lead Origin_Lead Add Form	0.65
11	Lead Quality_Worst	0.49
10	Lead Quality_Not Sure	0.21
8	Tags_Will revert after reading the email	0.17
7	Tags_Ringing	0.12
3	What is your current occupation_Unemployed	0.09
12	Last Notable Activity_SMS Sent	0.02

Model Evaluation

Accuracy, Specificity and Sensitivity on Train Data Set

The optimal cut-off turns out to be 0.20.



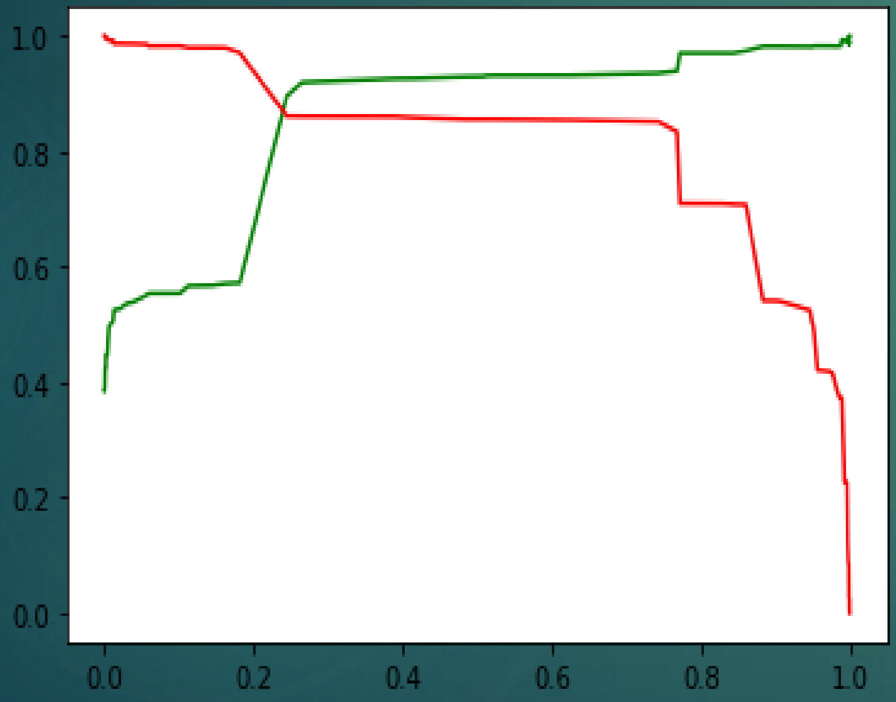
Confusion Metrics:

[[3747, 158],
  [353, 2093]],

- Accuracy – 90%
- Sensitivity – 86%
- Specificity – 93%
- False Positive Rate – 6%
- Positive Predictive Value – 89%
- Negative Predictive Value – 91%

Precision and Recall on Train Data Set

The optimal cut-off turns out to be 0.24.



Confusion Metrics:

[[3747, 158],
  [353, 2093]],

- Precision – 92%
- Recall – 85%

Accuracy, Specificity and Sensitivity on Test Data Set

Confusion Metrics:

```
[ [ 1626, 108],
[      153, 836] ],
```

- Accuracy – 90%
- Sensitivity – 84%
- Specificity – 93%
- False Positive Rate – 6%
- Positive Predictive Value – 88%
- Negative Predictive Value – 91%

Conclusion

- We have considered the optimal cut-off based on specificity and sensitivity for calculating final prediction on the test set.
- Accuracy, Specificity, Sensitivity for the test set turned out to be 90%, 84% and 93% which is approximated closer to the values calculated for train data set.
- Conversion Rate has now been increased up to 86%.
- The top 3 variables that contribute for lead getting converted are:
  - Lead Origin_Lead Add Form
  - Lead Source_Welingak Website
  - Last Activity_SMS Sent
- Hence overall model seems to be good.