LOGISTIC REGRESSION- LEAD SCORE CASE STUDY

Pieush Vyas

(DS-27)

CONTENTS

- Objective
- AIM
- Lead Conversion Procedure
- Approach
- Analysis for Application Data
- ROC Curve
- Cut-Off Point
- Comparison of Train and Test data
- Outcomes
- Recommendations

OBJECTIVE

- Analyze the patterns available in the given datasets which helps to understand about the conversion leads.

- With the help of Univariate/Bivariate analysis and by model building we can get insight of our best fitted model.

- Also need to calculate model Accuracy, Specificity, Sensitivity and ROC curve to know the model.

- By assigning Lead Score as per the conversion probability of the lead Company can take decision or make the strategy to convert that lead.

- All such analysis allows to take fare business decisions further to increase lead conversion.
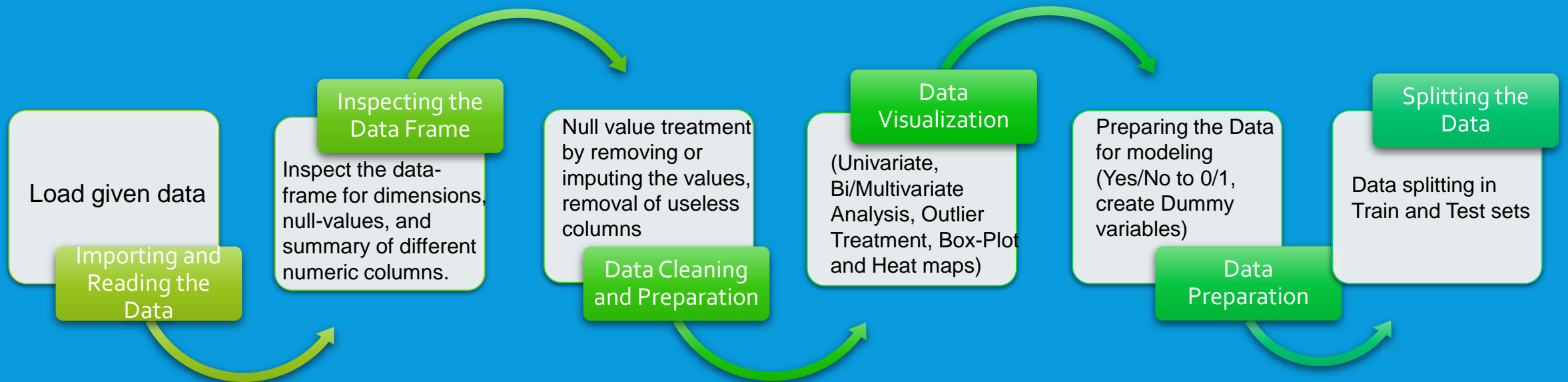
AIM

- Find out better model to increase lead conversion ratio from 38%.

- Find out the Lead Score to differentiate the leads as per their conversion probability and increase lead conversion ratio up-to 80%.
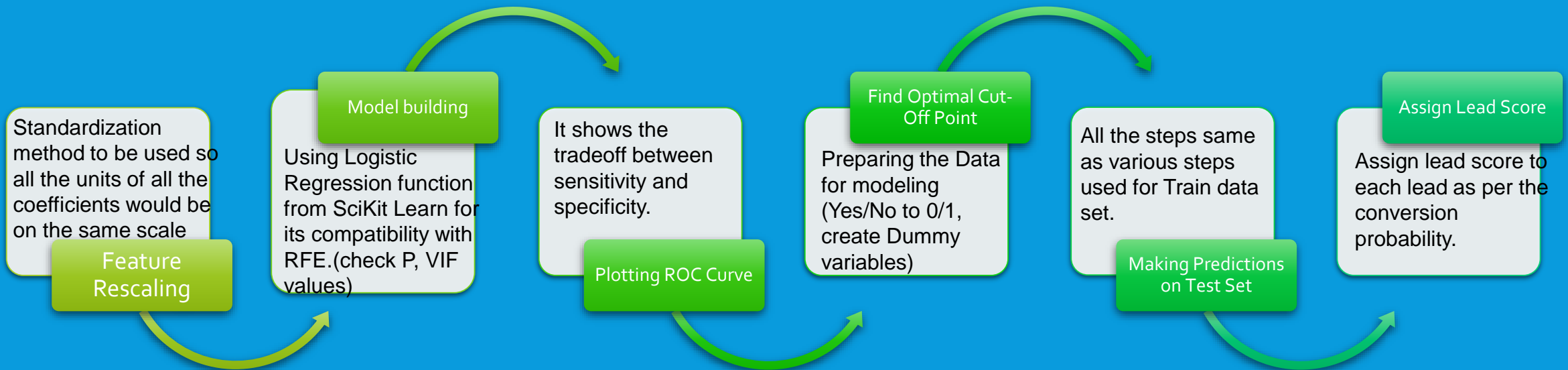
LEAD CONVERSION PROCEDURE

Social Media Marketing → Initial Pool of leads → Lead Nurturing (Hot Leads) → Converted Leads

- Educate the leads about the product, constant communication with the lead.

ANALYSIS APPROACH

Importing and Reading the Data
Load given data

Inspecting the Data Frame
Inspect the data-frame for dimensions, null-values, and summary of different numeric columns.

Data Cleaning and Preparation
Null value treatment by removing or imputing the values, removal of useless columns

Data Visualization
(Univariate, Bi/Multivariate Analysis, Outlier Treatment, Box-Plot and Heat maps)

Data Preparation
Preparing the Data for modeling (Yes/No to 0/1, create Dummy variables)

Splitting the Data
Data splitting in Train and Test sets

ANALYSIS APPROACH

Feature Rescaling
Standardization method to be used so all the units of all the coefficients would be on the same scale

Model building
Using Logistic Regression function from SciKit Learn for its compatibility with RFE.(check P, VIF values)

Plotting ROC Curve
It shows the tradeoff between sensitivity and specificity.

Find Optimal Cut-Off Point
Preparing the Data for modeling (Yes/No to 0/1, create Dummy variables)

Making Predictions on Test Set
All the steps same as various steps used for Train data set.

Assign Lead Score
Assign lead score to each lead as per the conversion probability.

ANALYSIS FOR APPLICATION DATA

Inferences:

- As per the graph it is clear that "Converted" leads are less than the "Non-Converted" leads.

ANALYSIS FOR APPLICATION DATA

Leads Soruce Vs Count
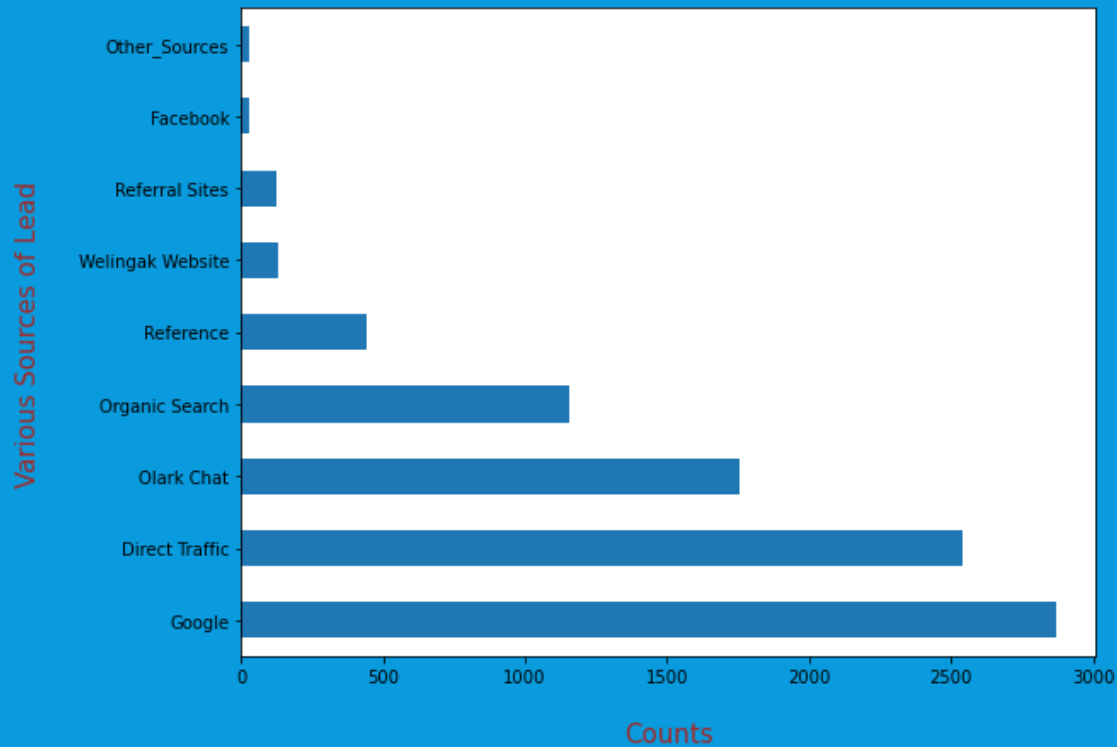
Inferences:

- As per the graph it is clear the maximum no. of leads are getting from 'Google' and then from 'Direct Traffc'.

- Whereas very less from Facebook and Referral Sites and almost Nill from other sources like 'Pay Per Click Ads, Press_Release, testone, WeLearn, Live Chat, Social Media bign, Click2call

ANALYSIS FOR APPLICATION DATA

Last Activity Status
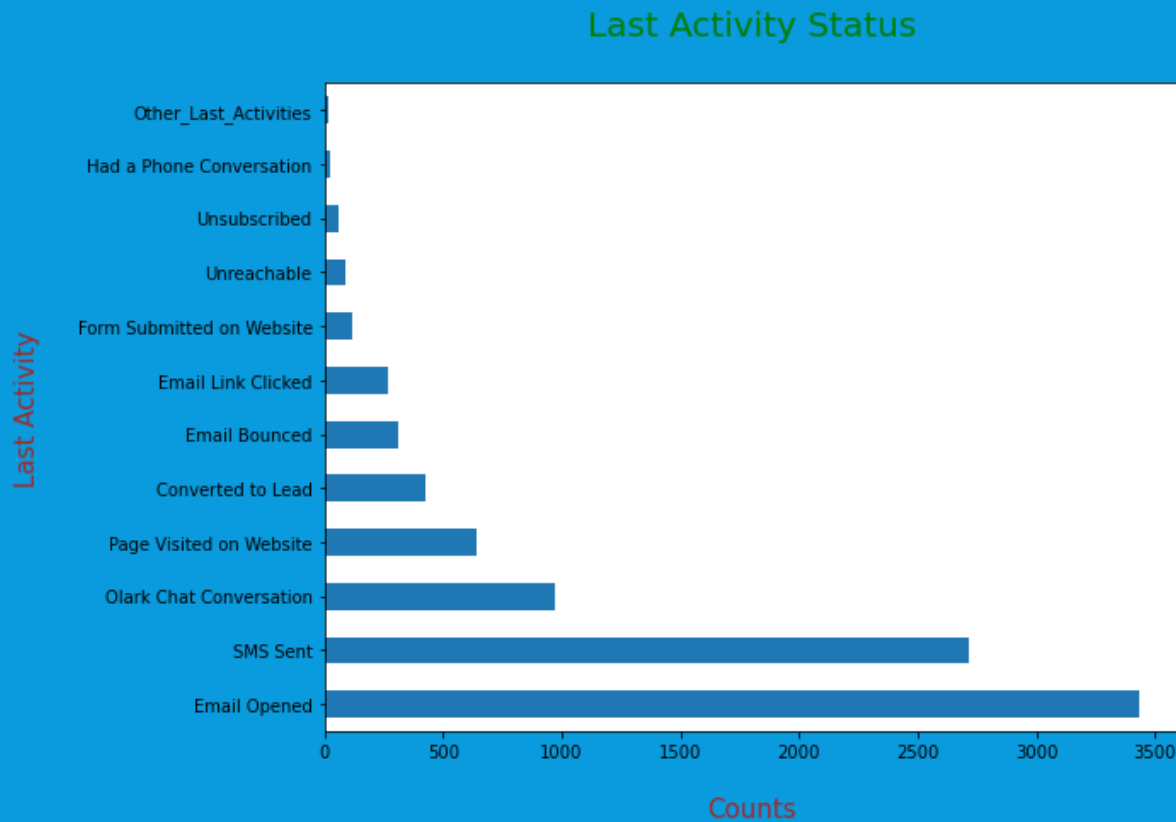
Inferences:

- As per the graph it is clear that the top 2 last activities done by the users are respectively the 'Email Opened', and 'SMS Sent'.

- 'Olark Chat Conversation' and 'Page Visited on Website' are 3rd and 4th mostly done last activities by the user.
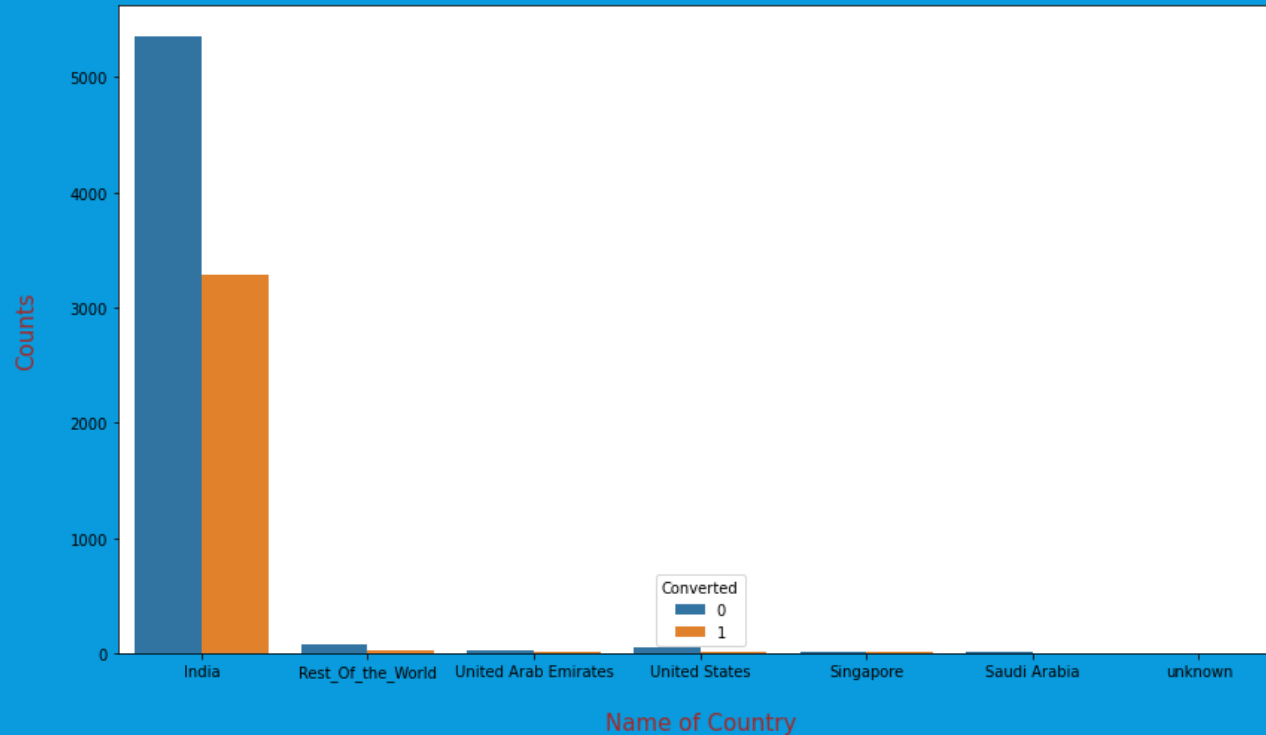
ANALYSIS FOR APPLICATION DATA
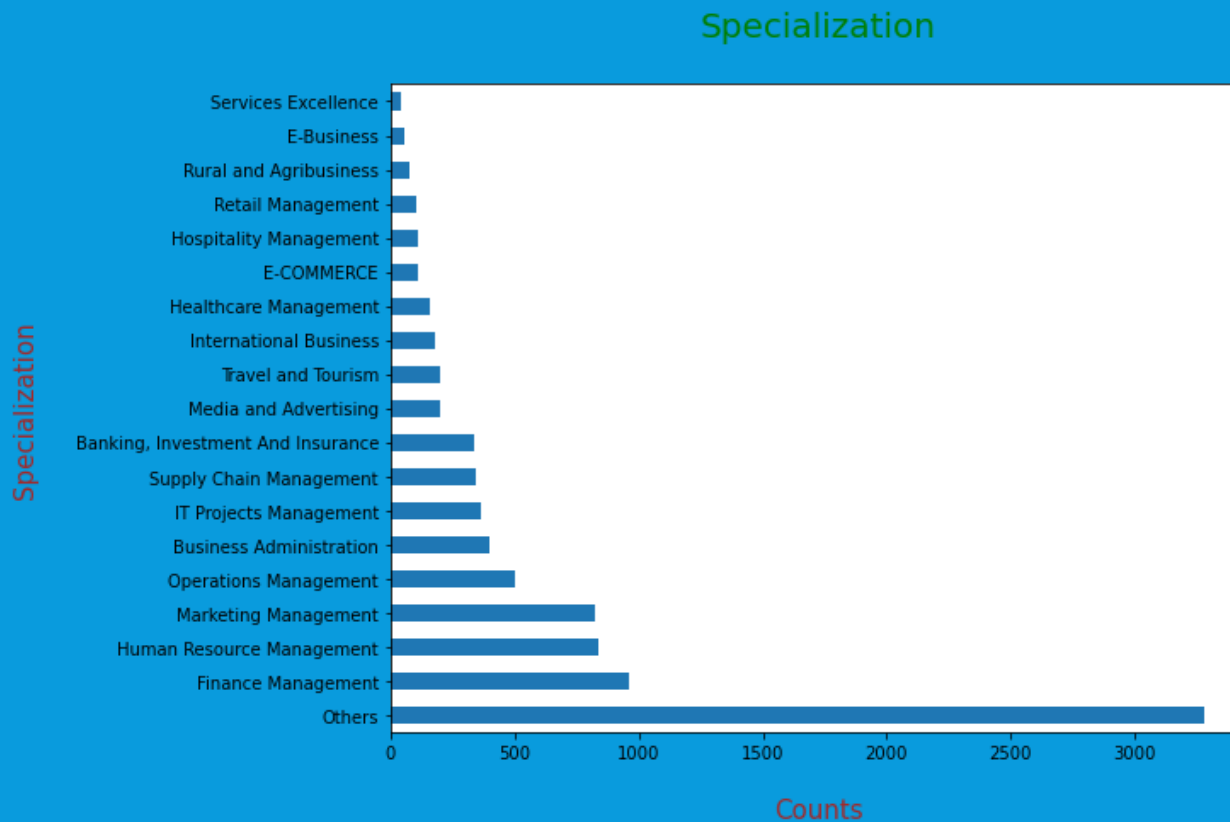
Specialization

Inferences:

- As per the graph it is clear that the maximum no. of leads didn't mentioned anything about Specialization therefore it is considered as 'Others'.

- Top 3 Specializations are 'Finance Management', 'Human Resource Management', and 'Marketing Management', whereas least liked Specialization is 'Services Excellence'.

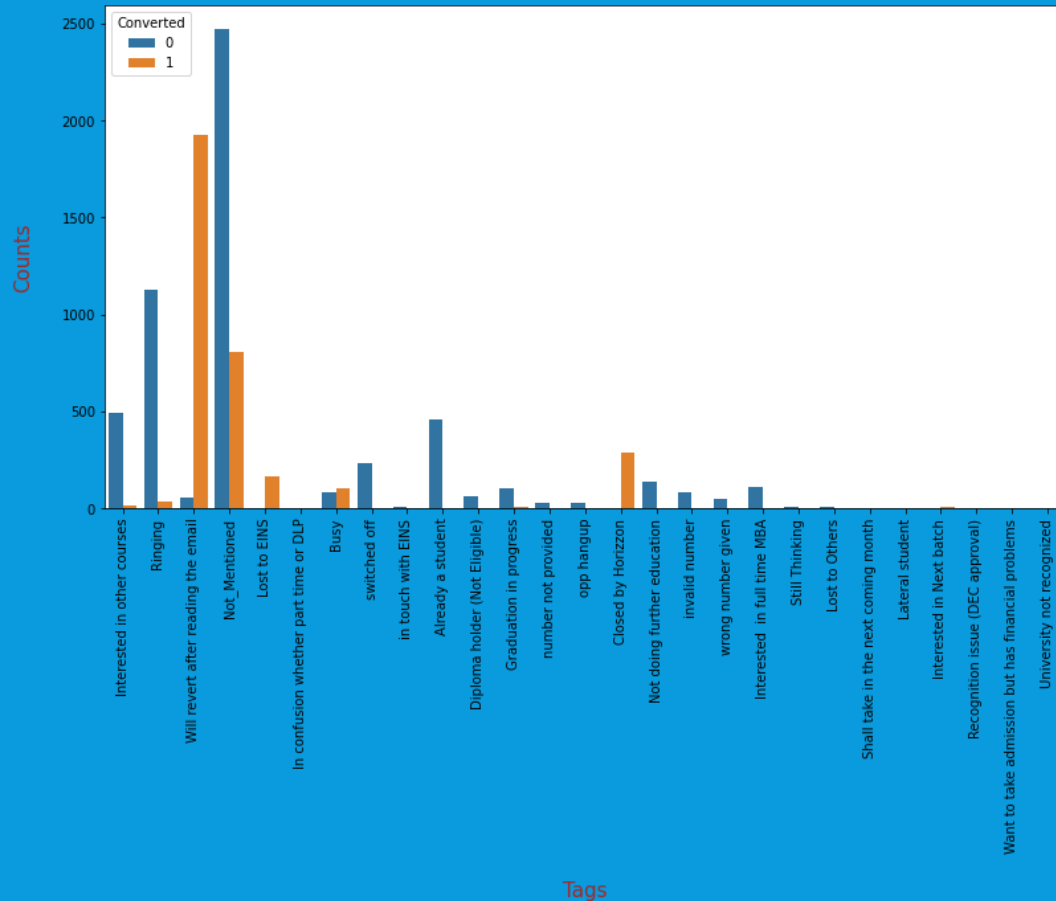# ANALYSIS FOR APPLICATION DATA

Country Wise Leads' Conversion Status

Inferences:

- From the graph it is clear that 'Converted' leads are less than 'Non-Converted' leads for any country.

- It is also clear that whether the leads are 'Converted' or 'Non-Converted' more than 90% are from India itself.

ANALYSIS FOR APPLICATION DATA


Tag wise Leads' Conversion Status

Inferences:

- As per the graph it is clear that the leads which has 'Will revert after reading the email', 'Closed by Horizzon', and 'Lost to EINS'Tags has very high conversion ratio.

- Whereas, 'Ringing', 'Interested in other Courses' Tags has very-2 low conversion ratio.

- Therefore Company should work accordingly to increase the leads conversion.

ANALYSIS FOR APPLICATION DATA

City wise Leads' Conversion Status
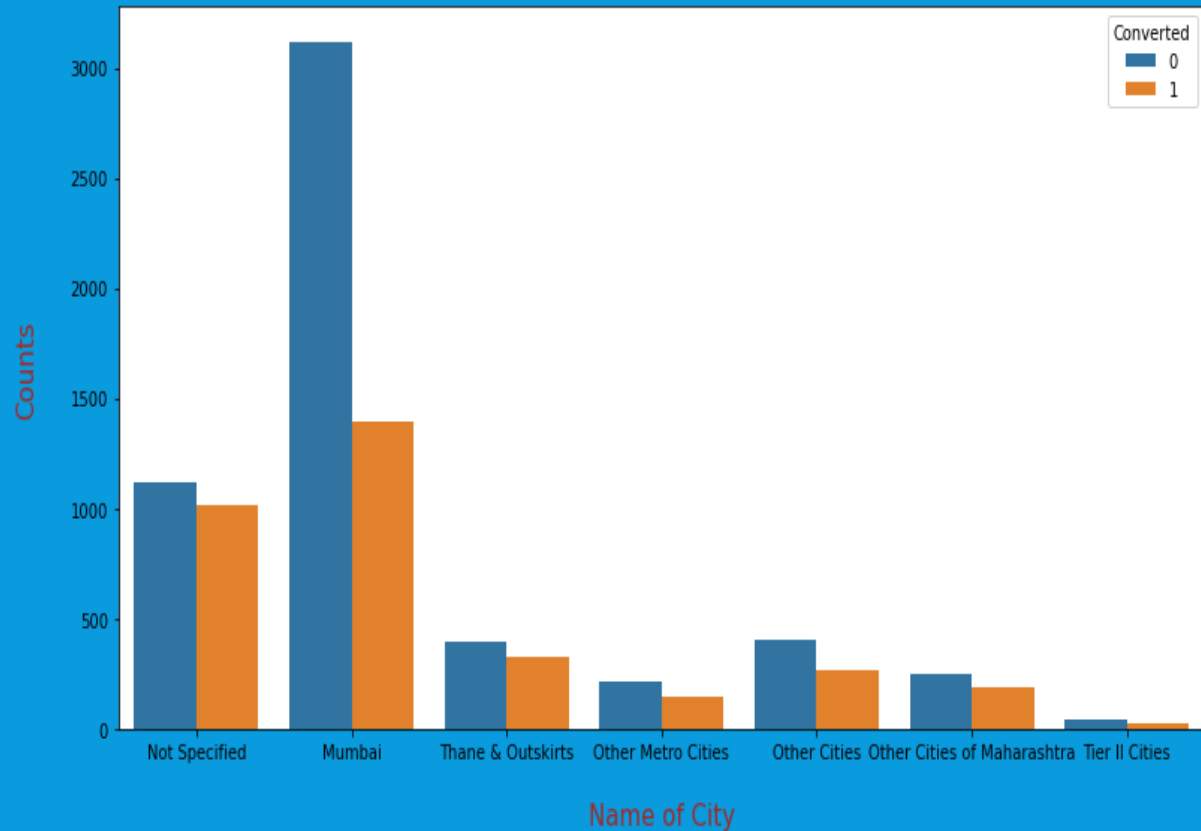
Inferences:

- As per the graph it is clear that 'Mumbai' has highest no. of leads' but conversion ratio among other cities is less.

- Like wise 'Other-Metro Cities', 'Other Cities' and 'Tier-II Cities' also has less converted leads in comparison to Non-Converted leads, but conversion ratio is still better in Tier-II cities.

- Whereas, 'Thane & Outskirts' and 'Cities of Maharashtra' has more converted leads than 'Non-Converted' leads.

ANALYSIS FOR APPLICATION DATA

Occupation wise Leads' Conversion Status
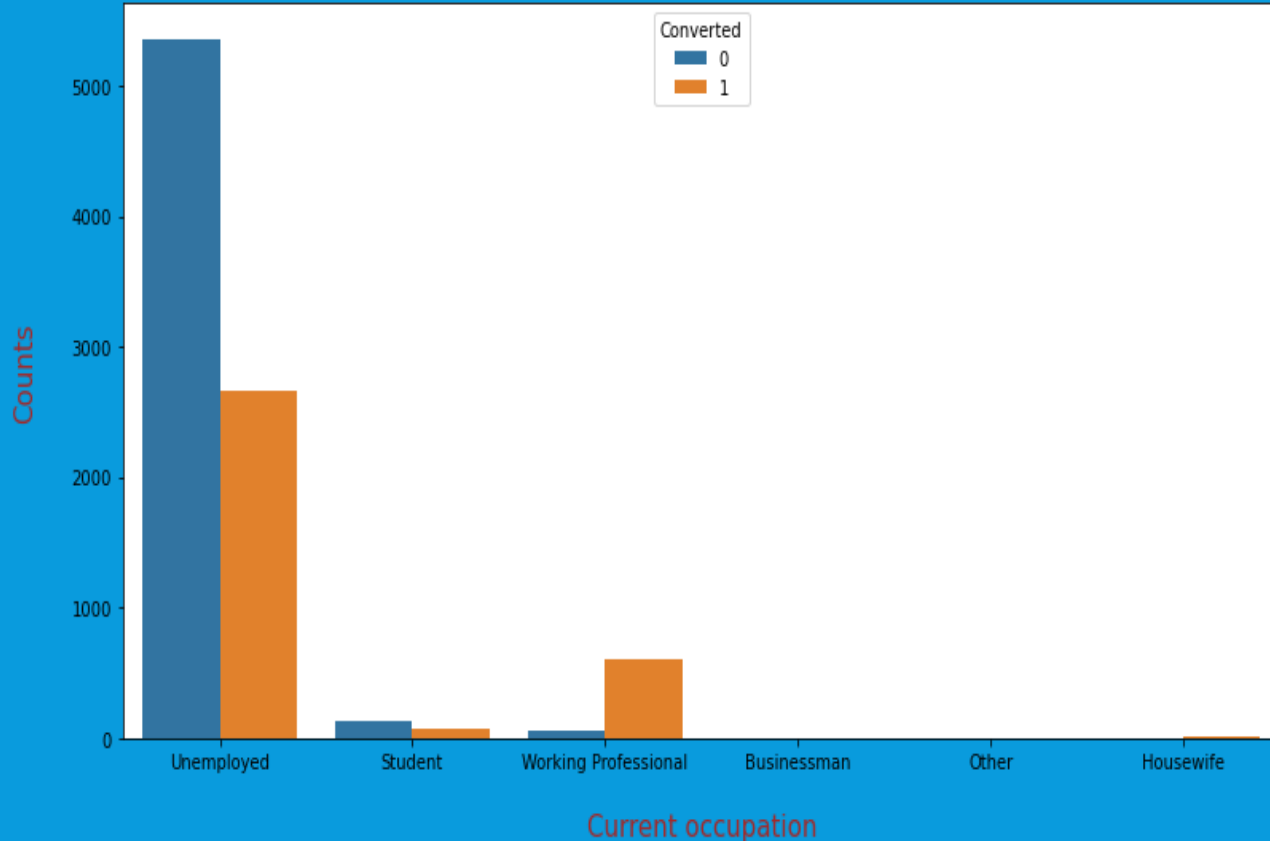
Inferences:

- Among all other professions most of the Converted leads are from 'Working Professional', whereas the no. of leads are higher among 'Unemployed' users.

- But it is surprisingly that 'Non-Converted' leads are also high in 'Unemployed' professional.

- Other professions like 'Businessman', 'House-wife' and 'Other' categories are very less in no.

ANALYSIS FOR APPLICATION DATA
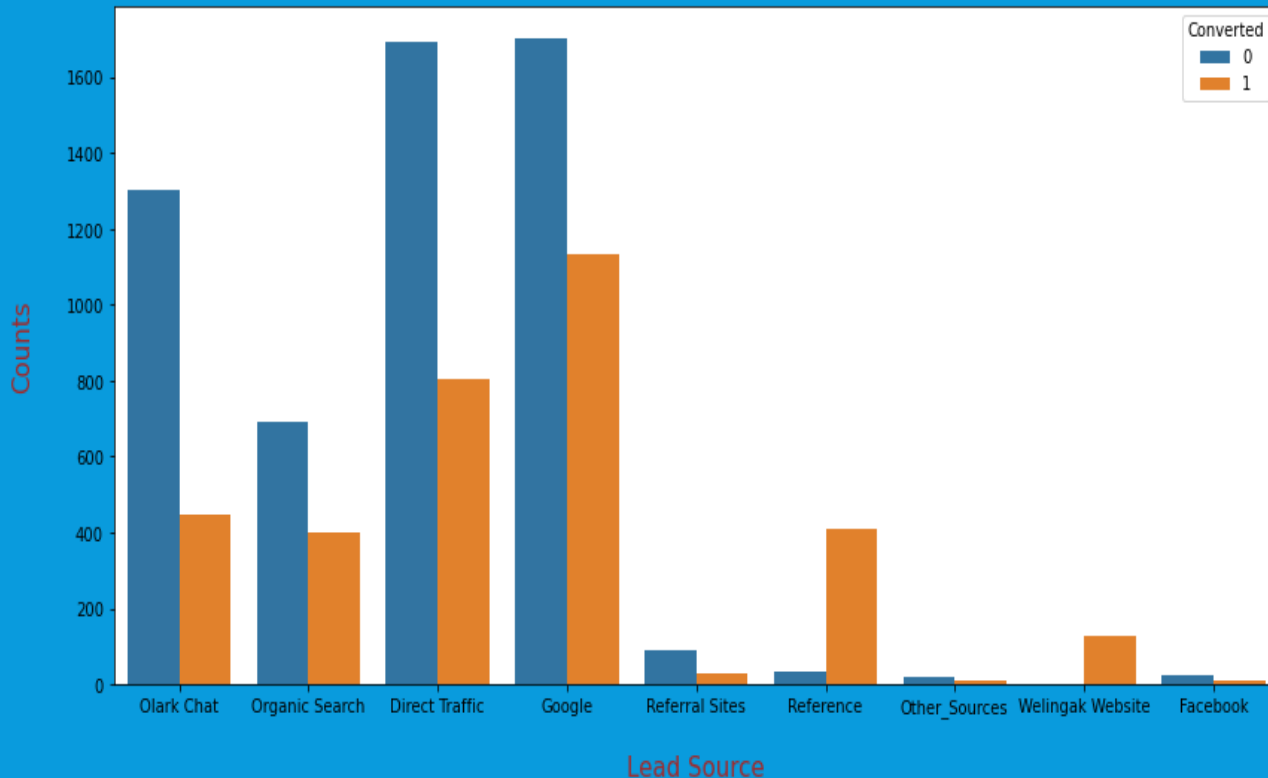

Lead Origin wise Leads' Conversion Status
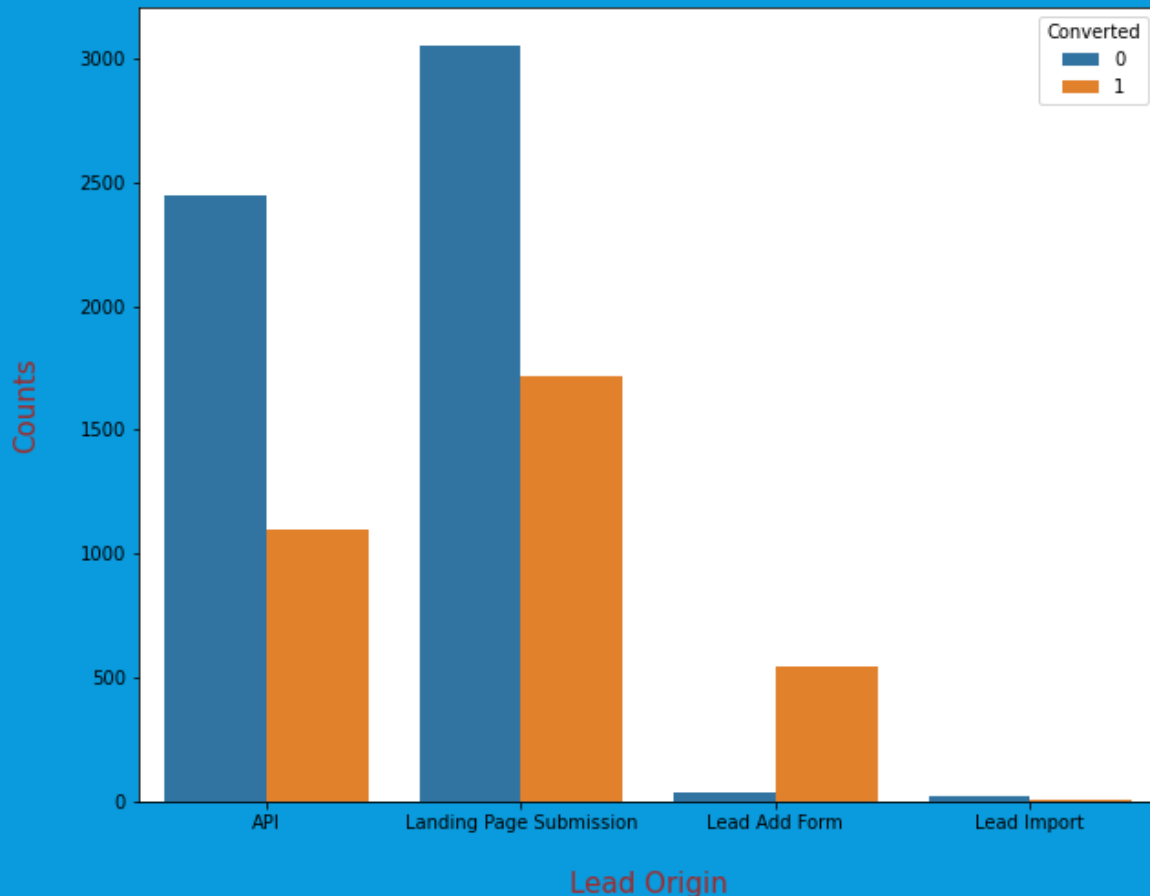
Inferences:

- As per the graph it is clear that most of the leads has been originated from 'Landing Page Submission', whereas 2nd highest Lead's Origin is from 'API'.

- Whereas, among all the Lead Origin categories maximum leads conversion are from 'Lead Add Form'.

- Therefore company should engage the leads by 'Lead Add Form'.

ANALYSIS FOR APPLICATION DATA


Lead Source wise Leads' Conversion Status

Inferences:

- As per the graph it is clear that lead conversion is less in most of the 'Lead Source', except 'Reference', and 'Welingak Website'.

- But on Google and Direct Traffic company got maximum leads but still the conversion ratio is less.

- Whereas 2nd and 3rd highest lead's conversion source are 'Olark Chat' and 'Organic Search' respectively.
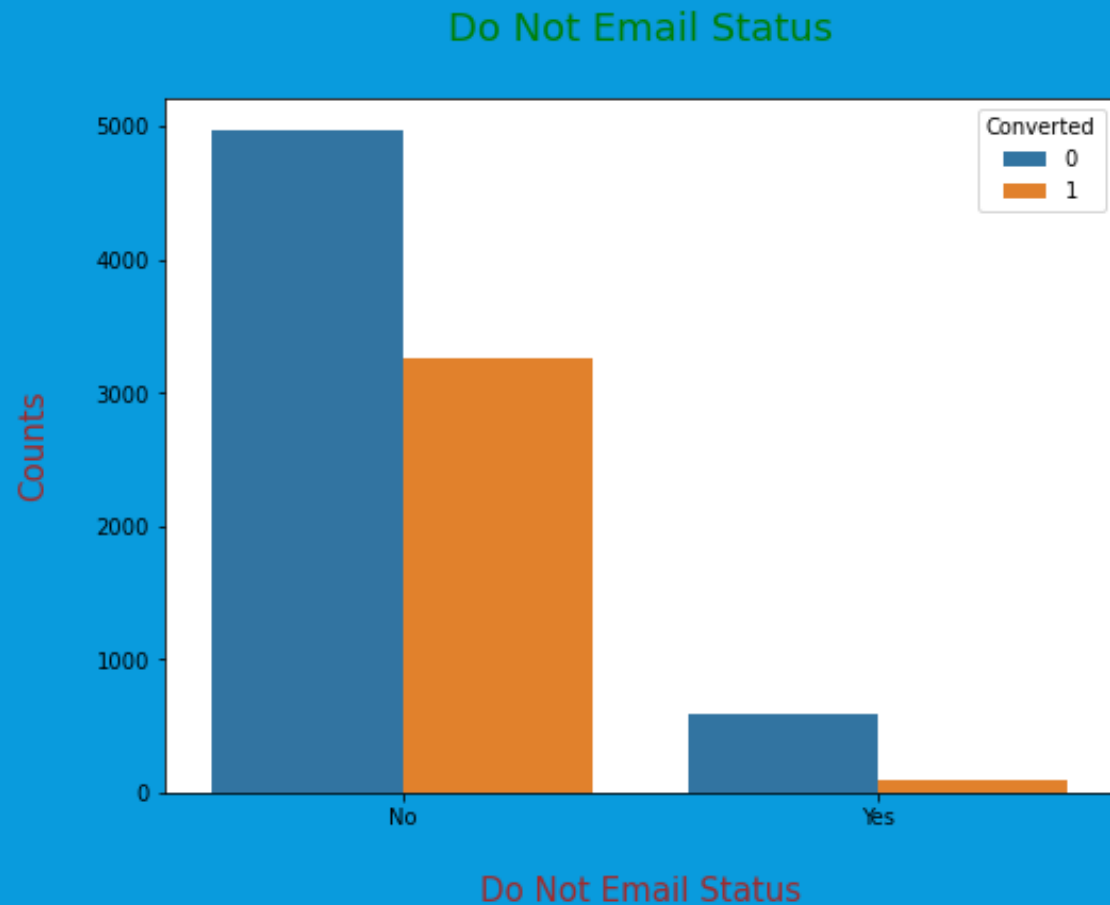
ANALYSIS FOR PREVIOUS APPLICATION DATA

Last Notable Activity wise Leads' Conversion Status
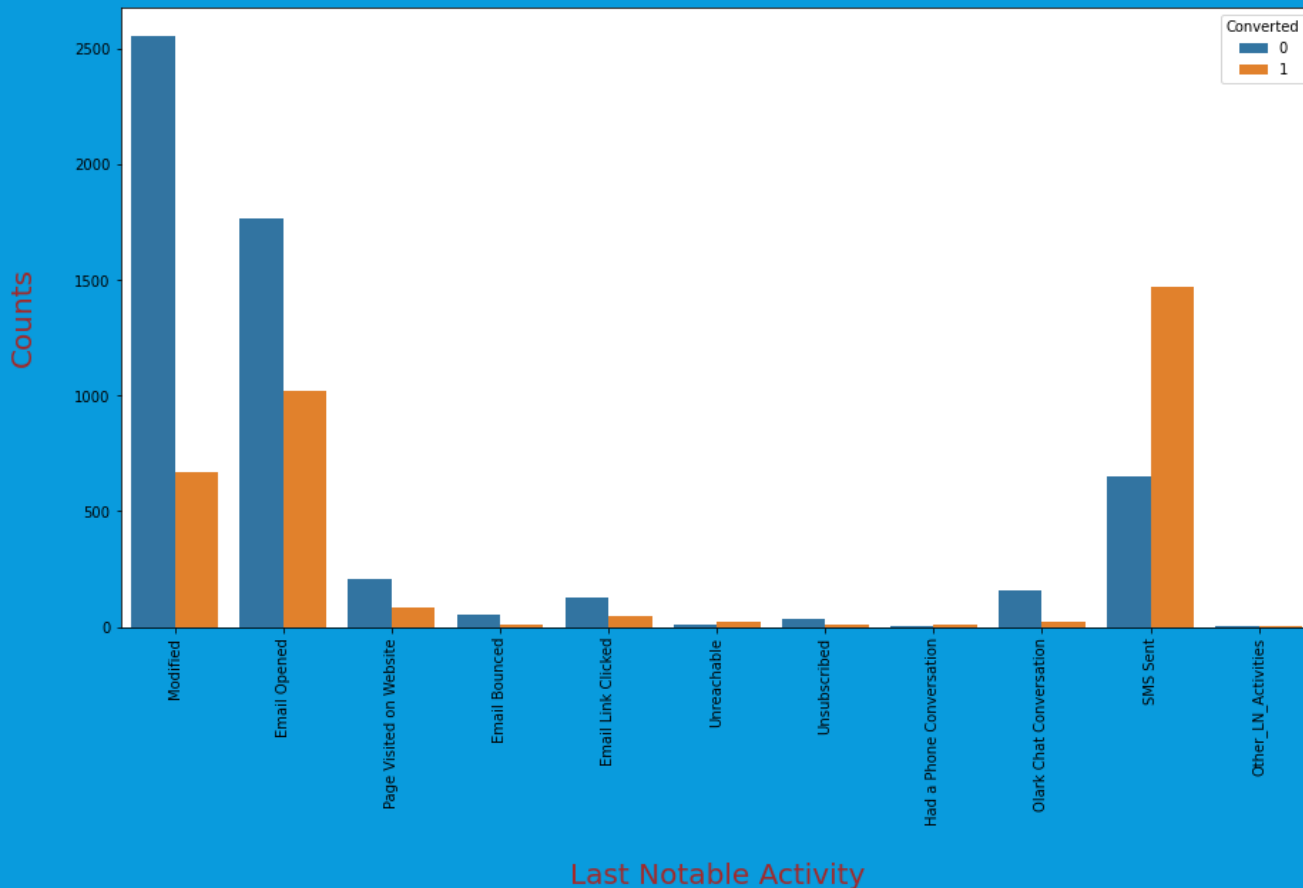
Inferences:

- As per the graph most of the converted leads' last activity was 'Modified' and 'Email Opened', but the lead conversion is less.

- Whereas most of the leads has been converted via the 'SMS Sent'.

- For all other categories of the Last notable Activity are very less for both type of leads.

ANALYSIS FOR PREVIOUS APPLICATION DATA

Inferences:

- No. of emails are more under 'No' category of "Do Not Email"

- Both types of Leads are less converted on 'Do Not Email' status.

- Therefore this graph shows that the both types of leads don't want emails about the course.

ANALYSIS FOR PREVIOUS APPLICATION DATA

Leads Status wise Total Visits

Leads' Total Time Spent on Website

Leads Page Views Per Visit
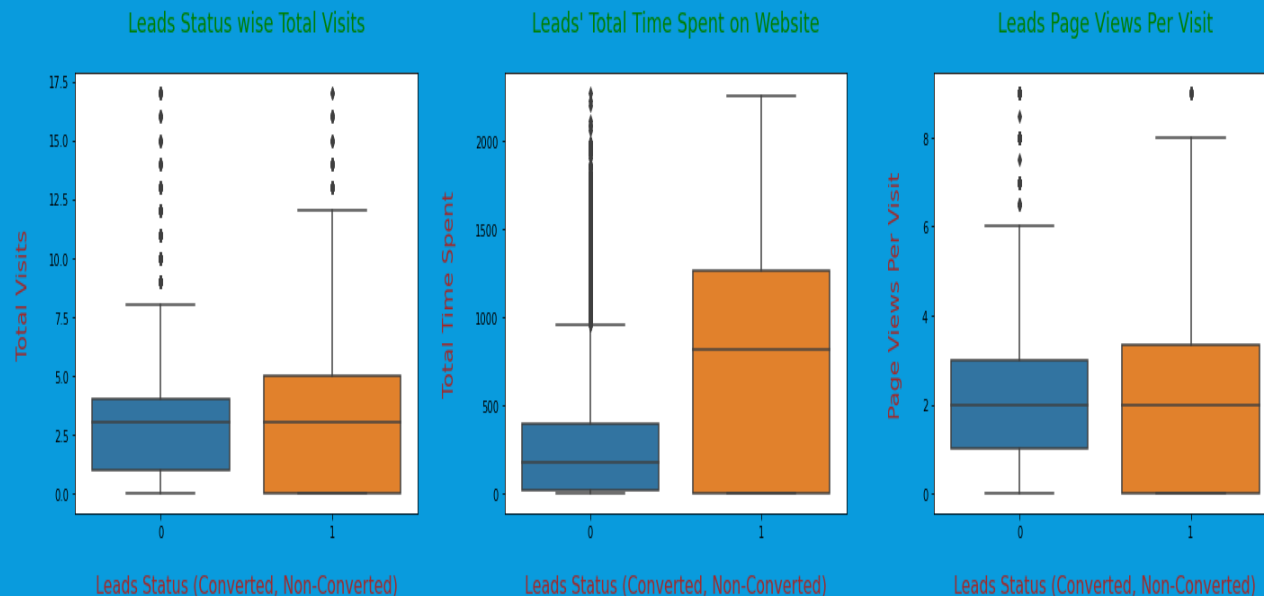
- As per the Box plots it is clear that Total Visits by converted leads is more than not converted leads.

- Whereas for the Converted leads distribution of box is higher in lower part than the upper part..

- Total Time spent on the Website by the converted leads is higher than Non-Converted leads, therefore it's range is also higher, whereas it is almost equally divided for Non-Converted leads.

- Upper limit of Leads Page Views Per visit is same for both type of leads whereas for Converted leads lower part is more.

- In short all the above Box Plots shows that Converted leads visits sites more at initial stage and spent much time at later stage, as well as initially their Page views per visit is higher.

ANALYSIS FOR PREVIOUS APPLICATION DATA

Correlation between various Numerical Variables

Inferences

- Converted leads Vs Page Views Per Visit are inversely proportional to each other.

- Whereas Total Visits are mostly related to the Page Views Per Visit and very less related to the Converted leads.

- Total time Spent is fairly related to 'Page Views Per Visit', and 'TotalVisits'.
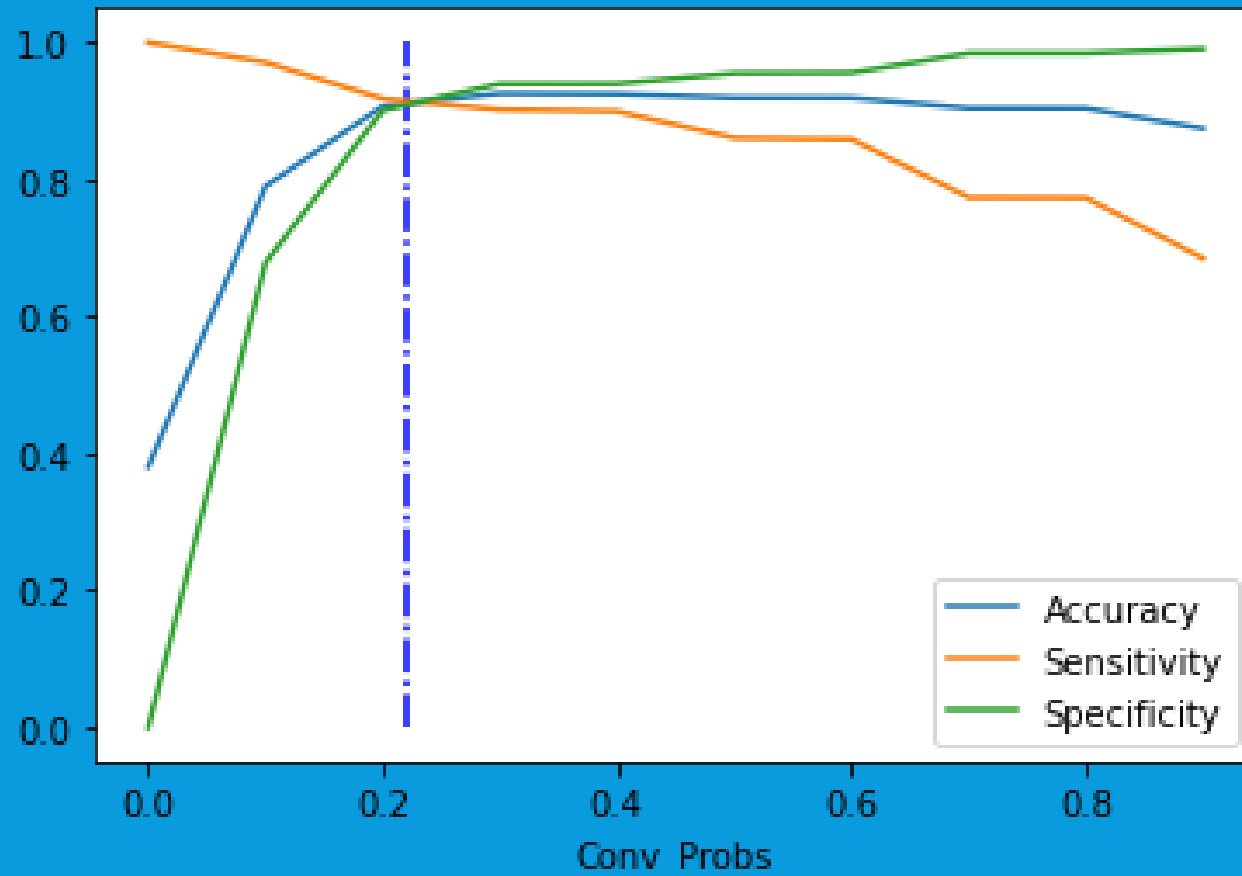
ROC CURVE

Receiver operating characteristic example
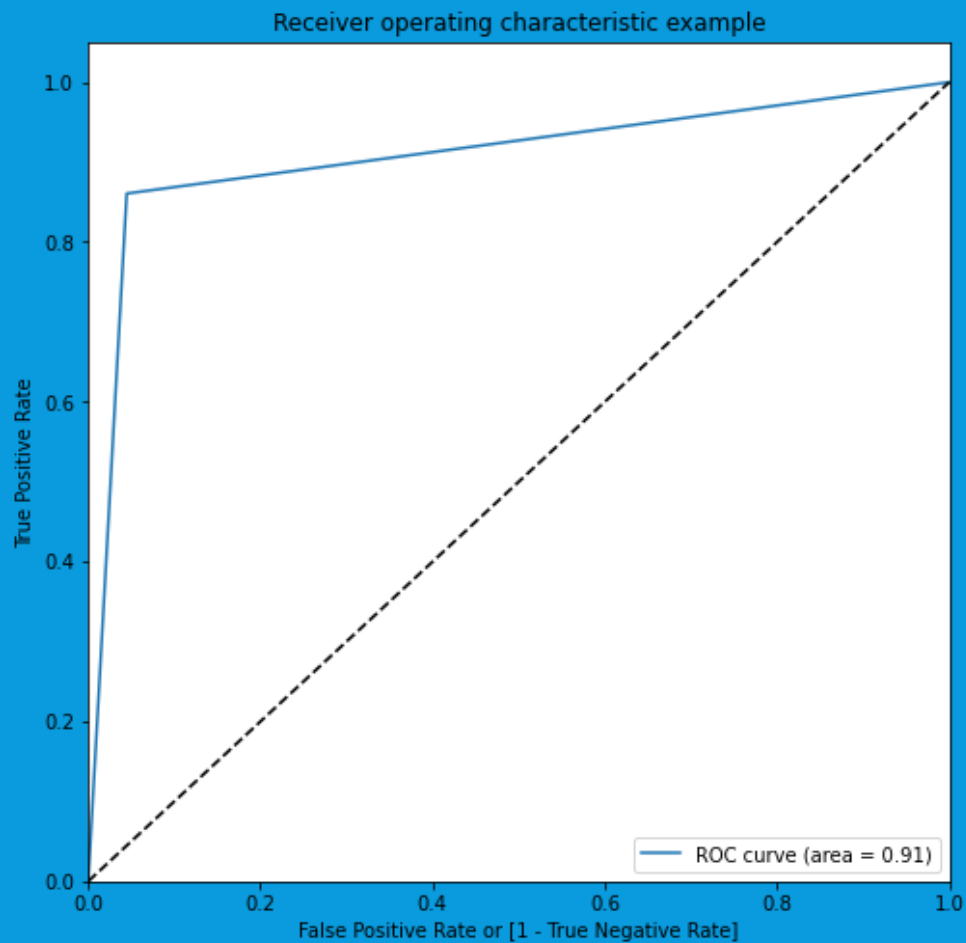
Inferences:

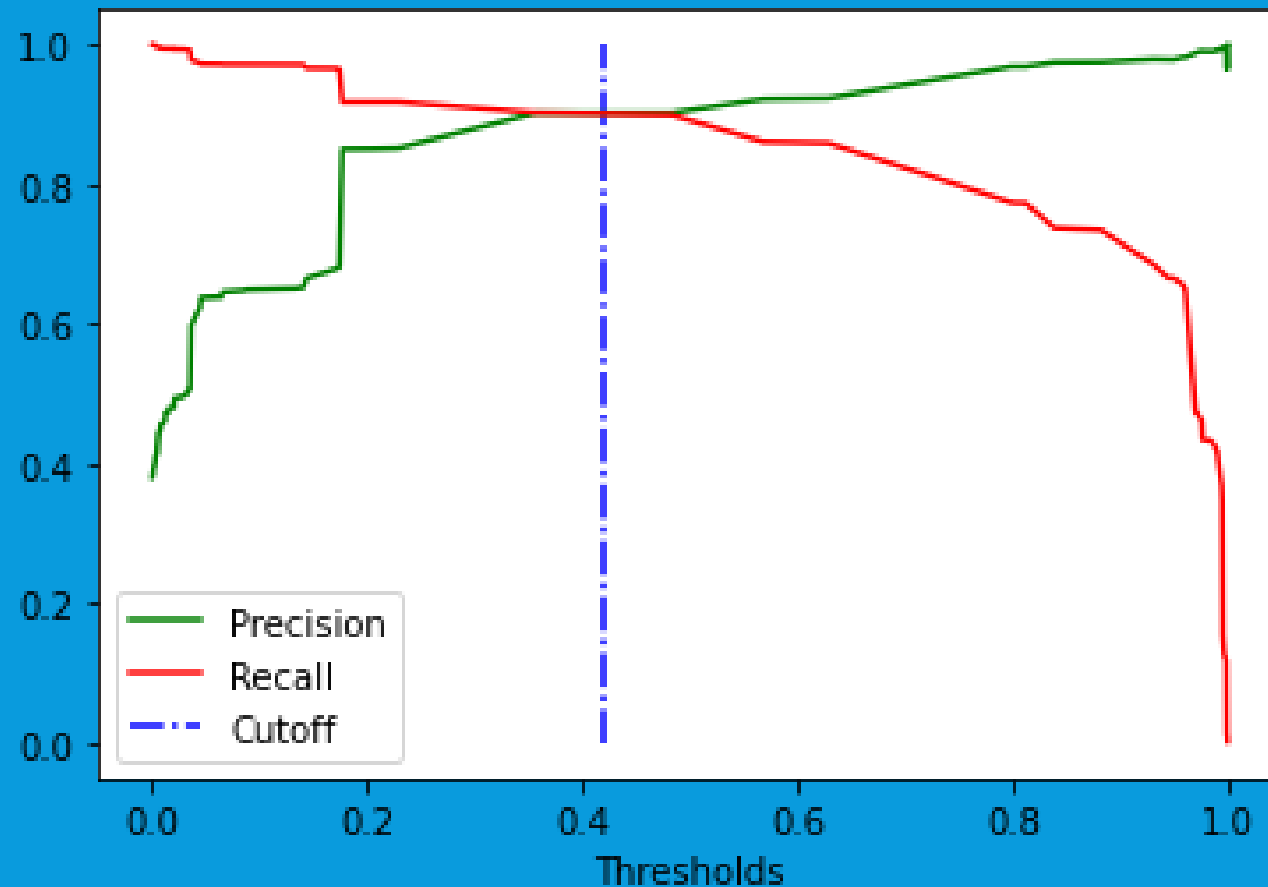- **Area Under the Curve (AUC) is 91%.**

CUT-OFF POINT-1

Inferences

- From the probability curve 0.22 is the optimum point to take it as a cutoff probability.
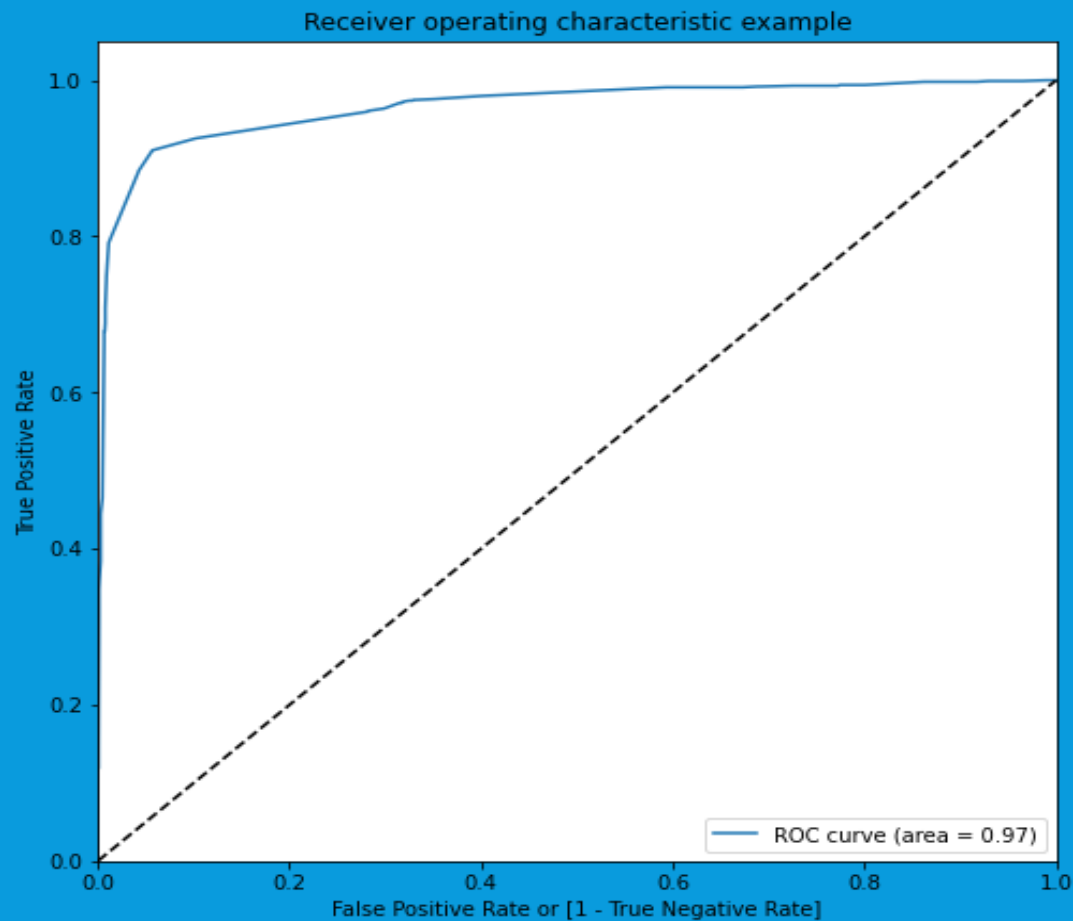
CUT-OFF POINT-2

Inferences

- After varying cut-off point finally new optimum point has been achieved at 0.42.

ROC-CURVE-2

Inferences

- Are under the Curve (AUC) of the ROC curve for the Test predicated data is 97%.

COMPARISON OF TRAIN AND TEST DATA

Train Data:

Accuracy - 91.9 %

Sensitivity – 86 %

Specificity – 96 %

False Positive Rate – 4 %

Positive Predictive Value – 92 %

Negative Predictive Value – 92 %

Test Data:

Accuracy : 93 %

Sensitivity : 91 %

Specificity : 94 %

False Positive Rate – 6 %

Positive Predictive Value – 90 %

Negative Predictive Value – 95 %

OUTCOMES

- Overall Accuracy of the model is almost stable for Test and Train Data. This is 91.9 % and 93 % respectively.

- Whereas False positive rate is only 4-6%.

- ROC for Train and Test data is 91 % and 97 % respectively, which is very good.

- After predications Cut-off values has been updated from 0.22 to 0.42.

- Therefore X Education Company should implement on work on the various suggestions given in next slides for getting higher lead conversion.

- This can be decided according to the Lead Score given to Leads according to their Conversion Probability.

RECOMMENDATIONS

X Education Company should work on the following aspects to improve the overall lead conversion:--

I)   Following Variables affects positively on Lead Conversion ratio or has higher conversion ratio:--

    a) Should increase sending SMS as it has higher conversion ratio.

    b) Increase user time on the Welingak Website as it also has higher conversion rate.

    c) Phone Conversation with the leads as it helps to know the customer in a more

        better way and chances of lead conversion has been increased. This activity also

        has higher conversion ratio.

RECOMMENDATIONS

d) Leads which received via reference has higher conversion ratio, therefore try to increase more no. of leads via reference, to increase lead conversion ratio.

e) Although most of the leads come from Google, and Direct Traffic but still the leads conversion ratio is highest for the leads come from 'Reference', and Welingak Website. Therefore company need to work on this.

f) 'Lead Add Form', has higher conversion ratio therefore, company should increase leads time on the website by engaging them via multiple Activity forms.

g) Leads which has tag of 'Lost to EINS', 'Will revert after reading the email', and 'Closed by Horizzon' has higher conversion ratio, therefore company should consider these tags and attend those leads first.

h) 'Working Professional' has very high conversion ratio therefore company should emphasize more on such leads.

RECOMMENDATIONS

I) Although maximum leads are from 'Mumbai' itself but the conversion ratio is higher in 'Tier-II' city, therefore company should focus on those cities more to increase more no. of leads and therefore lead conversion.

J) Improve 'Olark Chat' as it is affecting negatively to the leads conversion.

K) As 'TotalVisits' are highly related to 'Page Views Per Visit', company should increase Total Visits on the website, which increases

Lastly, Lead Score has been assigned to each lead as per their conversion ratio and this is the most decisive step to identify the Hot lead and increase the lead conversion ratio to achieve the lead conversion target of 80%.

THANKS VERY MUCH FOR YOUR PATIENCE!