LEAD SCORING CASE STUDY

PROBLEM STATEMENT

- An X Education sells online courses to industry professional.
- An X education gets a lots of leads, it's lead conversion rate is very poor about 30%.
- The company wants to increases the lead conversion rate to be round 80%. To make this process more efficient the company wished to identify the most potential leads.

GOAL OF THE CASE STUDY

- Building a logistic regression model to assign a lead between 0 and 100 to each of the leads which can be used by the company to target potential leads.
- The higher the lead score the more promising the lead to get converted, the lower it is, the lesser the chances of conversion.
- The Model to be built in lead conversion rate around 80% or more.

SOLUTION METHODOLOGY

- Data cleaning and data manipulation.
    1. Check and handle duplicate data.
    2. Check and handle NA values and missing values.
    3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
    4. Imputation of the values, if necessary.
    5. Check and handle outliers in data.
- EDA
    1. Univariate data analysis: value count, distribution of variable etc.
    2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

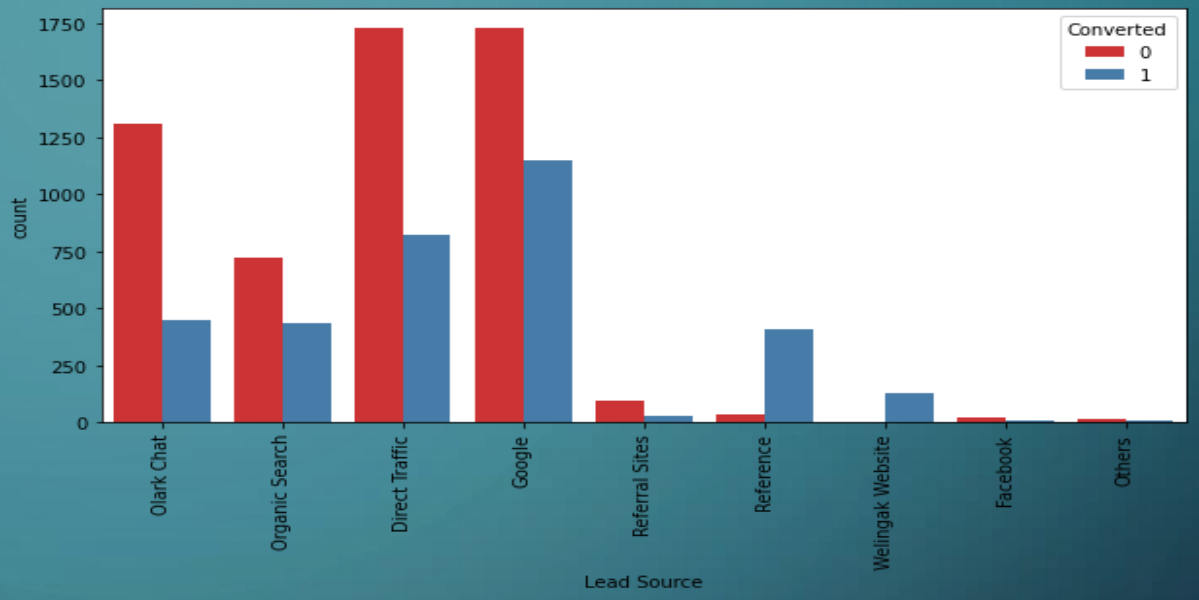
DATA MANIPULATION

- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”
- Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- Dropping the columns having more than 35% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

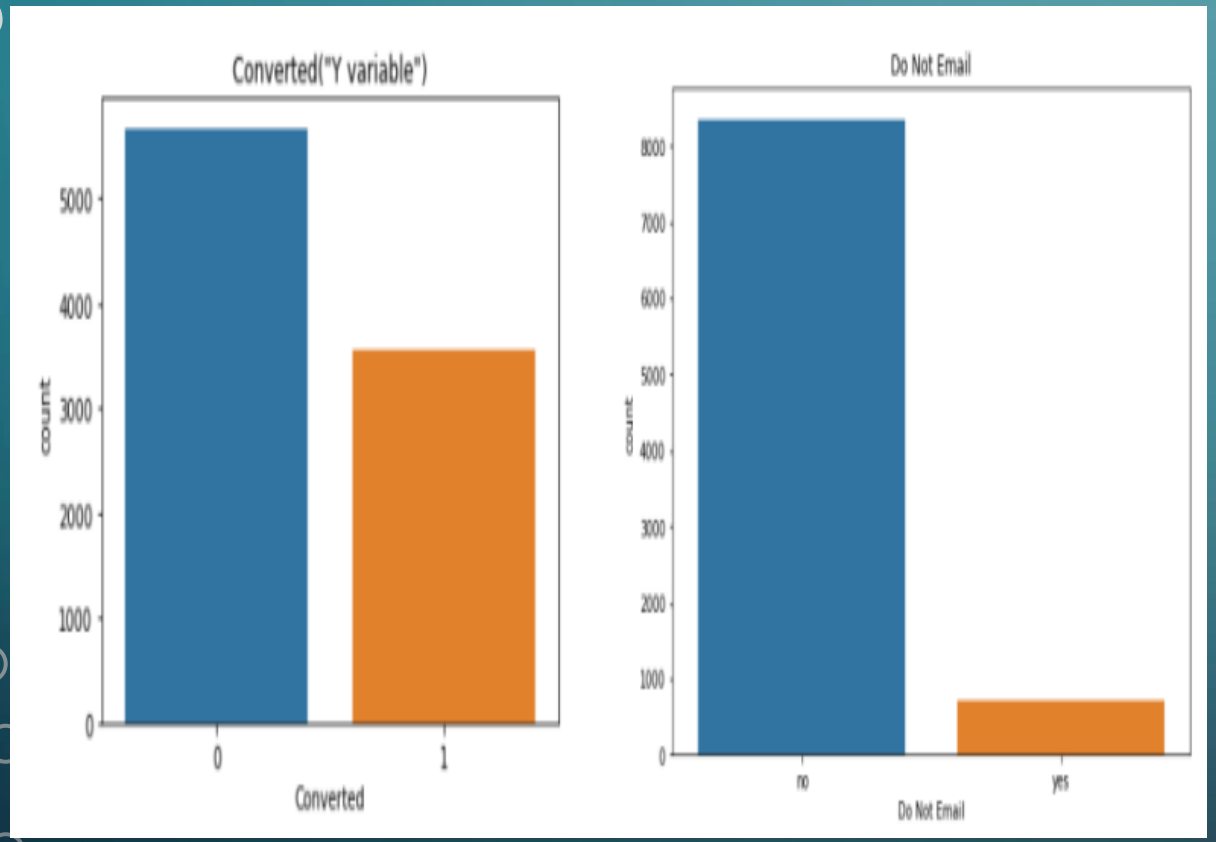
EDA

LEAD SOURCE VS CONVERTED

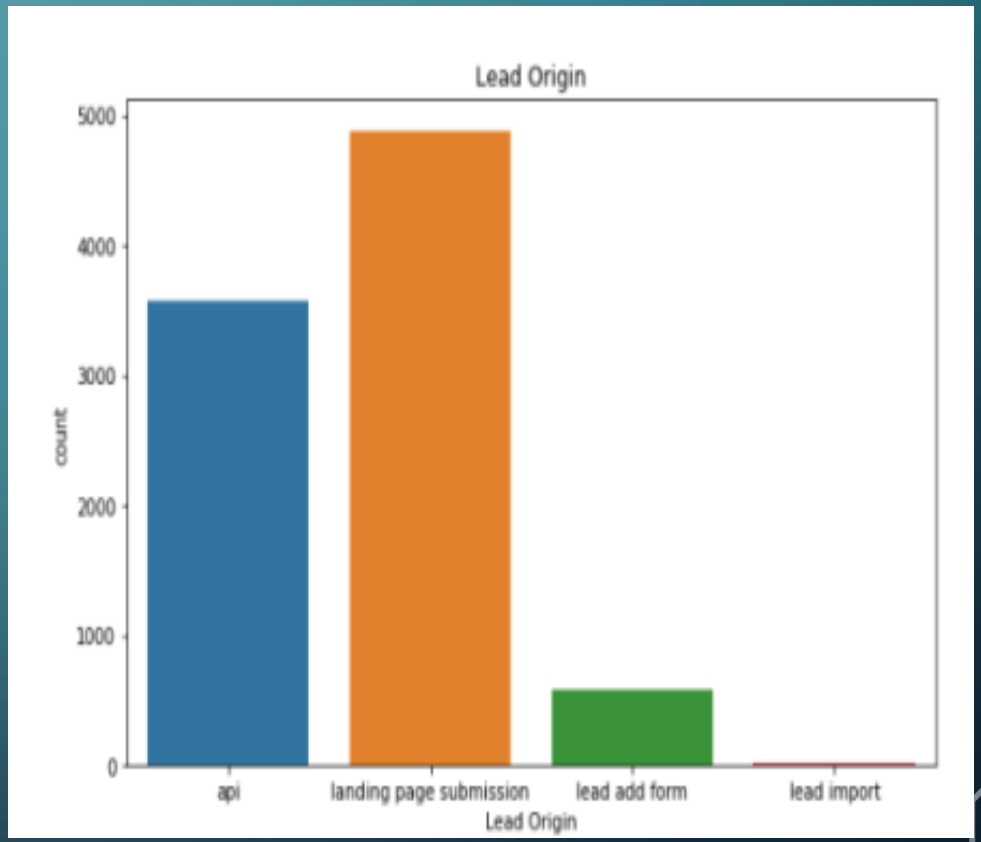
- Google searches has had high conversions compared to other modes, whilst references has had high conversion rate.



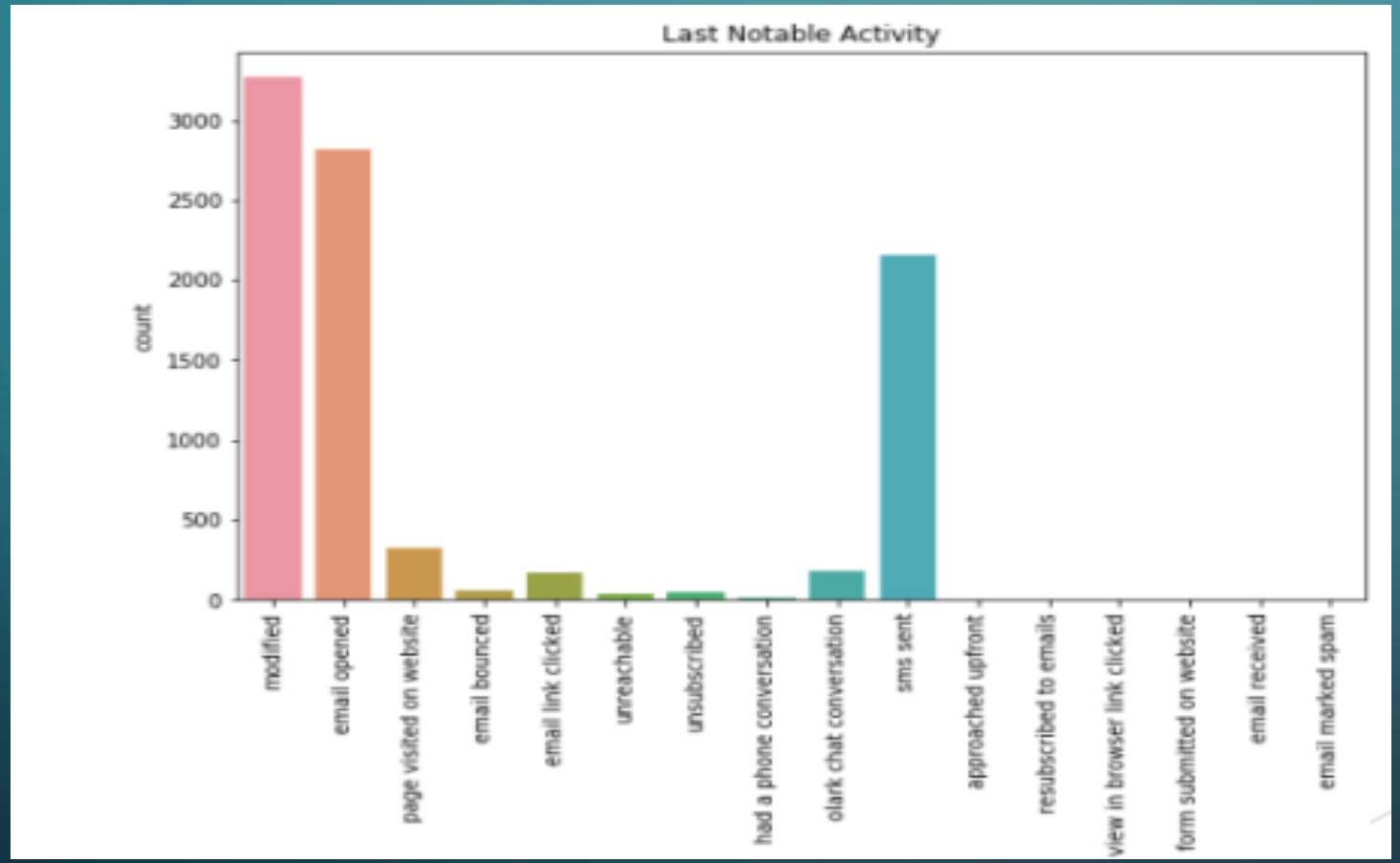
DO NOT EMAIL VS CONVERTED

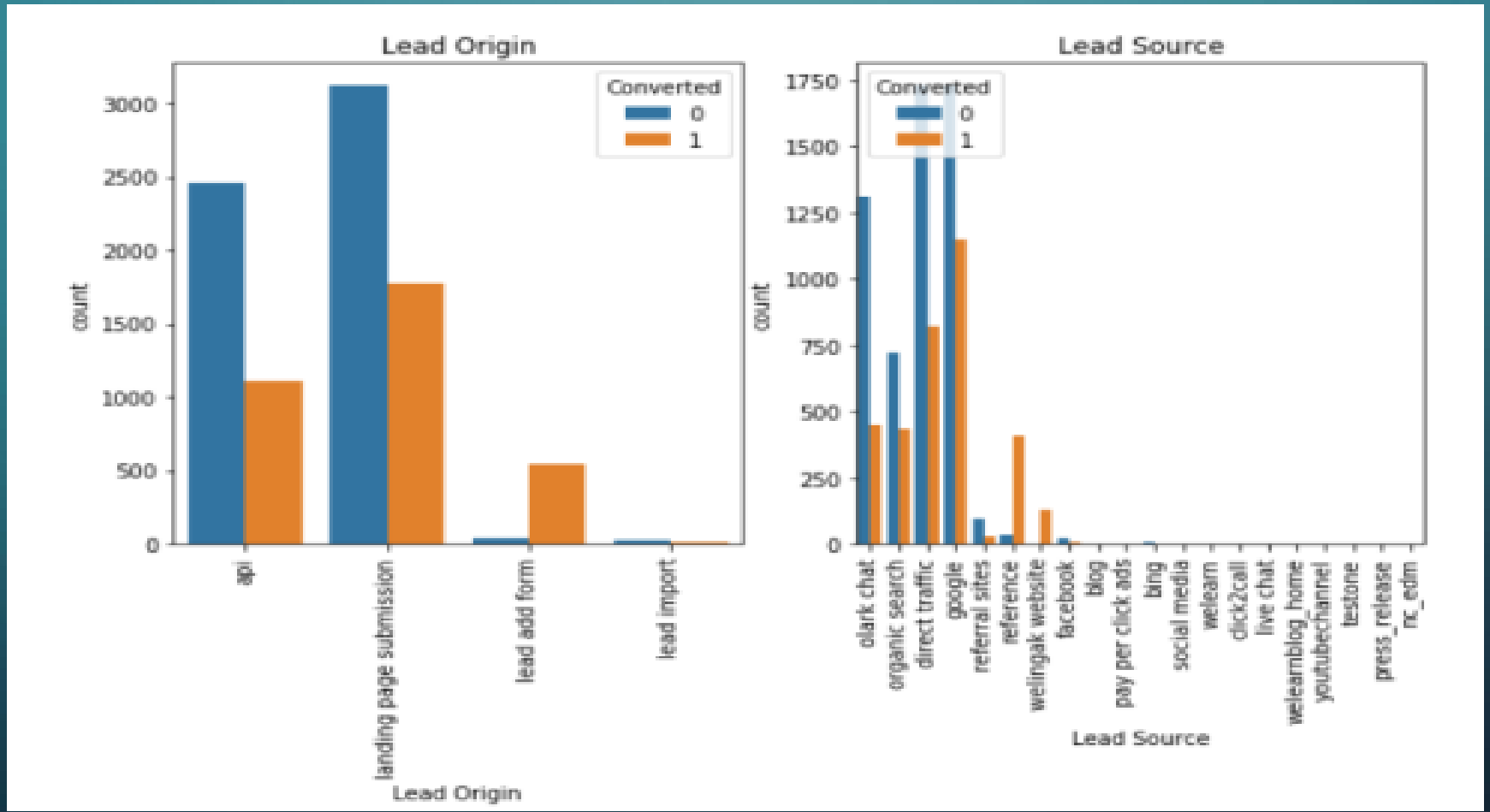

Converted("Y variable")
count
5000
4000
3000
2000
1000
0
0
1
Converted

Do Not Email
count
8000
7000
6000
5000
4000
3000
2000
1000
0
no
yes
Do Not Email

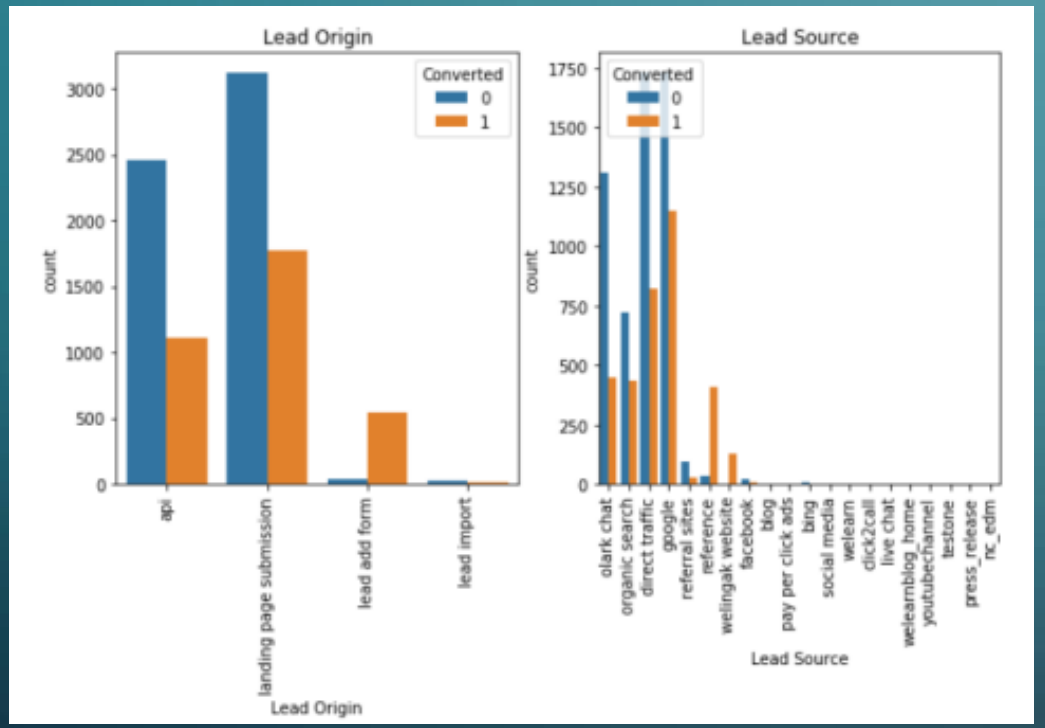

Lead Origin
count
5000
4000
3000
2000
1000
0
api
landing page submission
lead add form
lead import
Lead Origin

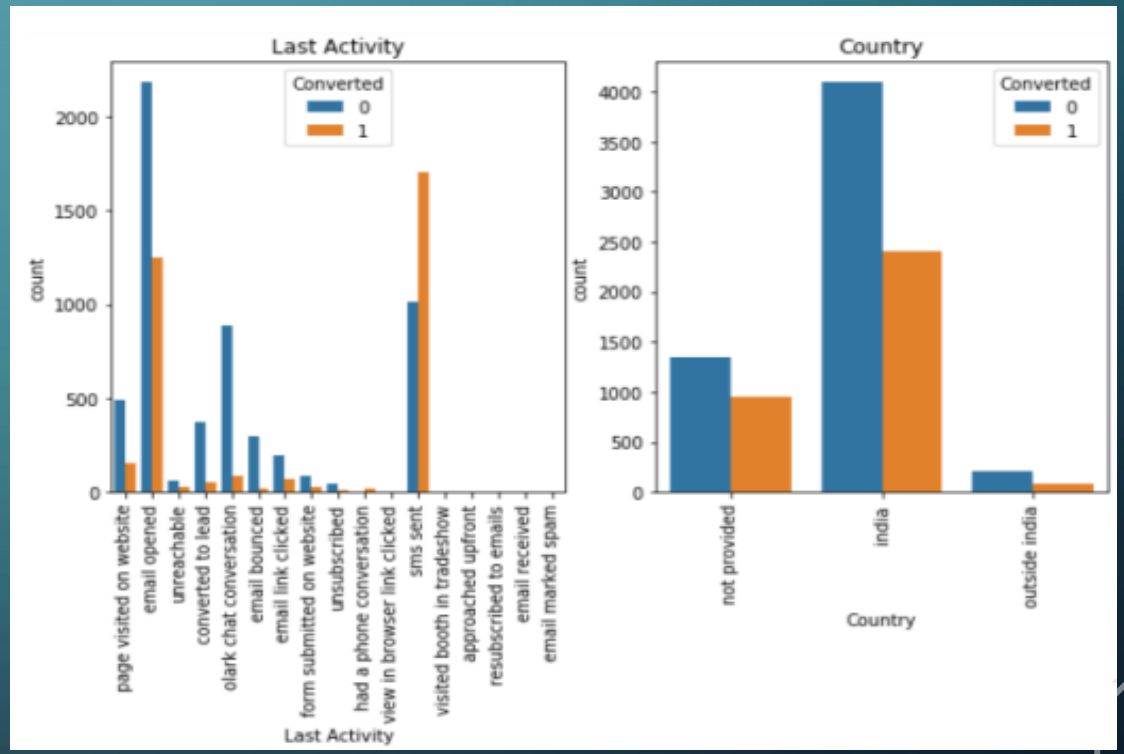
LAST NOTABLE ACTIVITY EDA




Lead Origin
Converted
0
1
3000
2500
2000
1500
1000
500
0
ount
api
landing page submission
lead add form
lead import
Lead Origin
Lead Source
Converted
0
1
1750
1500
1250
1000
750
500
250
0
ount
olark chat
organic search
direct traffic
google
referral sites
reference
welingak website
facebook
blog
pay per click ads
bing
social media
welearn
click2call
live chat
welearnblog_home
youtubechannel
testone
press_release
nc_edm
Lead Source

CATEGORICAL VARIABLE RELATION



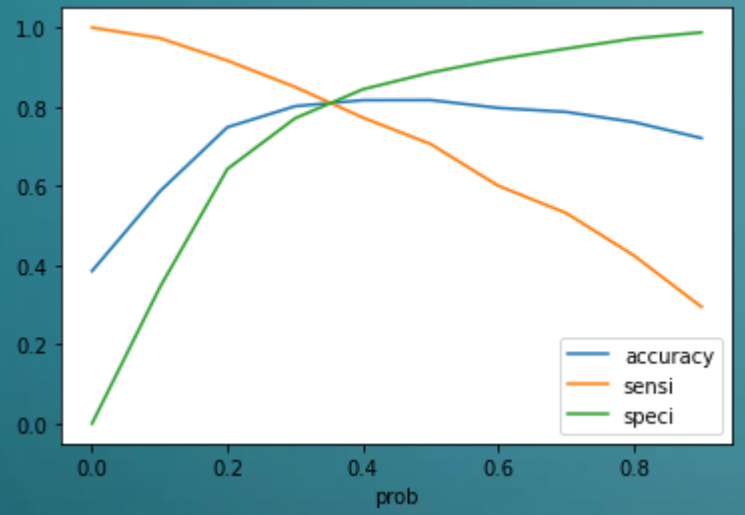


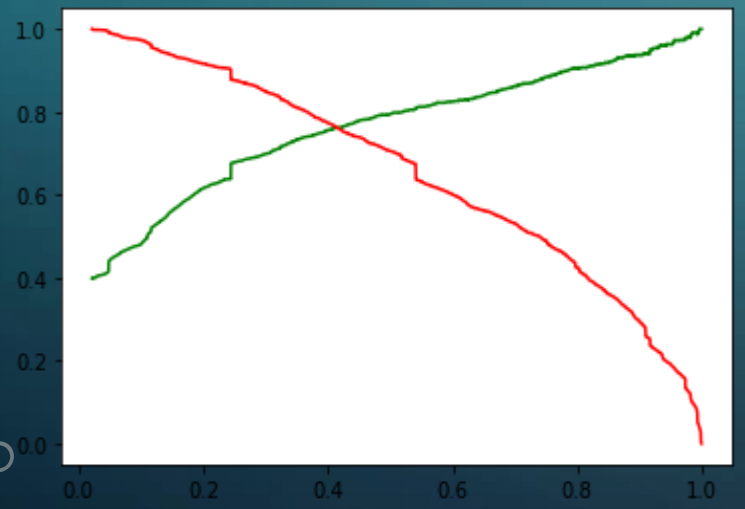
MODEL BUILDING

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and value is greater than 5
- Predictions on test data set
- Overall accuracy 81%

ROC CURVE

MODEL EVALUATION TRAIN

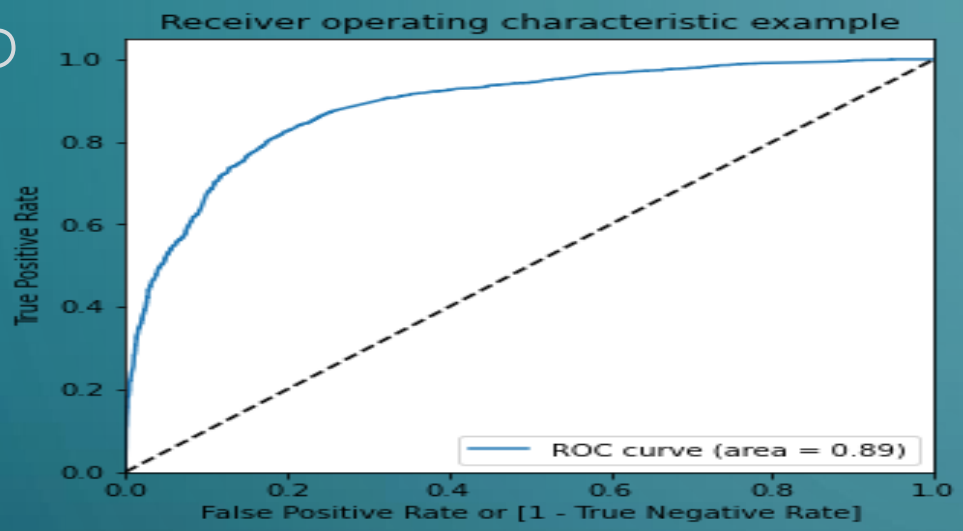


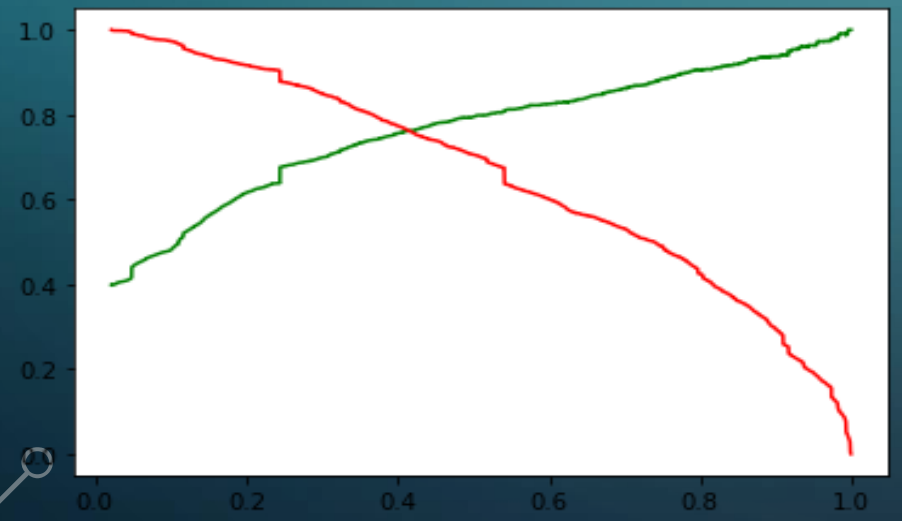


- **ACCURACY SENSITIVITY AND SPECIFICITY**
  - 81.0% Accuracy
  - 81.7% Sensitivity
  - 80.6% Specificity

- **PRECISION AND RECALL**
  - 79.54% Precision
  - 70.61% Recall

MODEL EVALUATION TRAIN (TEST)





- We have higher **(0.89)** area under the ROC curve, therefore our model is a good one.

- **ACCURACY SENSITIVITY AND SPECIFICITY**
  - 80.4% Accuracy
  - 80.4% Sensitivity
  - 80.5% Specificity
  - From the graph to the left, **0.34** is the optimum point to take it as a cut off probability.

- **PRECISION AND RECALL**
  - 79.54% Precision
  - 70.61% Recall

CONCLUSION

- **Exploratory Data Analysis:**
  - People spending higher than average time are promising leads, so targeting them and approaching them can be helpful in conversions.
  - SMS messages can have a high impact on lead conversion.
  - Landing page submissions can help find out more leads.
  - Marketing management, human resources management has high conversion rates. People from these specializations can be promising leads.
  - References and offers for referring a lead can be a good source for higher conversions.
  - An alert messages or information has seen to have high lead conversion rate.
- **Logistic Regression Model:**
  - The model shows high close to 80.4% accuracy.
  - The threshold has been selected from Accuracy, Sensitivity, Specificty measures and precision, recall curves.
  - The model shows 80.4% Sensitivity and 80.5% Specificity.
  - The model finds correct promising leads and leads that have less chances of getting converted.
  - Overall this model proves to be accurate.

THANK YOU