LEAD SCORE CASE STUDY

Presented by
Chandan Singh
Diwakar Mishra
Shreeharsh Pande

PROBLEM STATEMENT

Problem:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

So, the company wants to know:

- Leads that are most likely to convert into paying customers.
- Expecting the target lead conversion rate to be around 80%

Objective:

- This company wants to know the most promising leads.
- A new model need to be created which identifies the hot leads.
- Logistic regression model has to be created for business problems and solution.

APPROCH USED

Data cleaning and data manipulation :

- Check and handle duplicate data
- Check and handle NA values & missing values
- Drop columns if it has large amount of missing values
- Imputation of the values if needed
- Outlier handling if exist.

EDA:

- Univariate Analysis
  - Check value count, distribution of the variable etc.

- Bivariate Analysis
  - Correlation coefficients and look for the pattern between the variables etc.

Feature selection & Dummy variables and encoding

Classification technique :

- Use of Logistic regression to predict model

Model validation

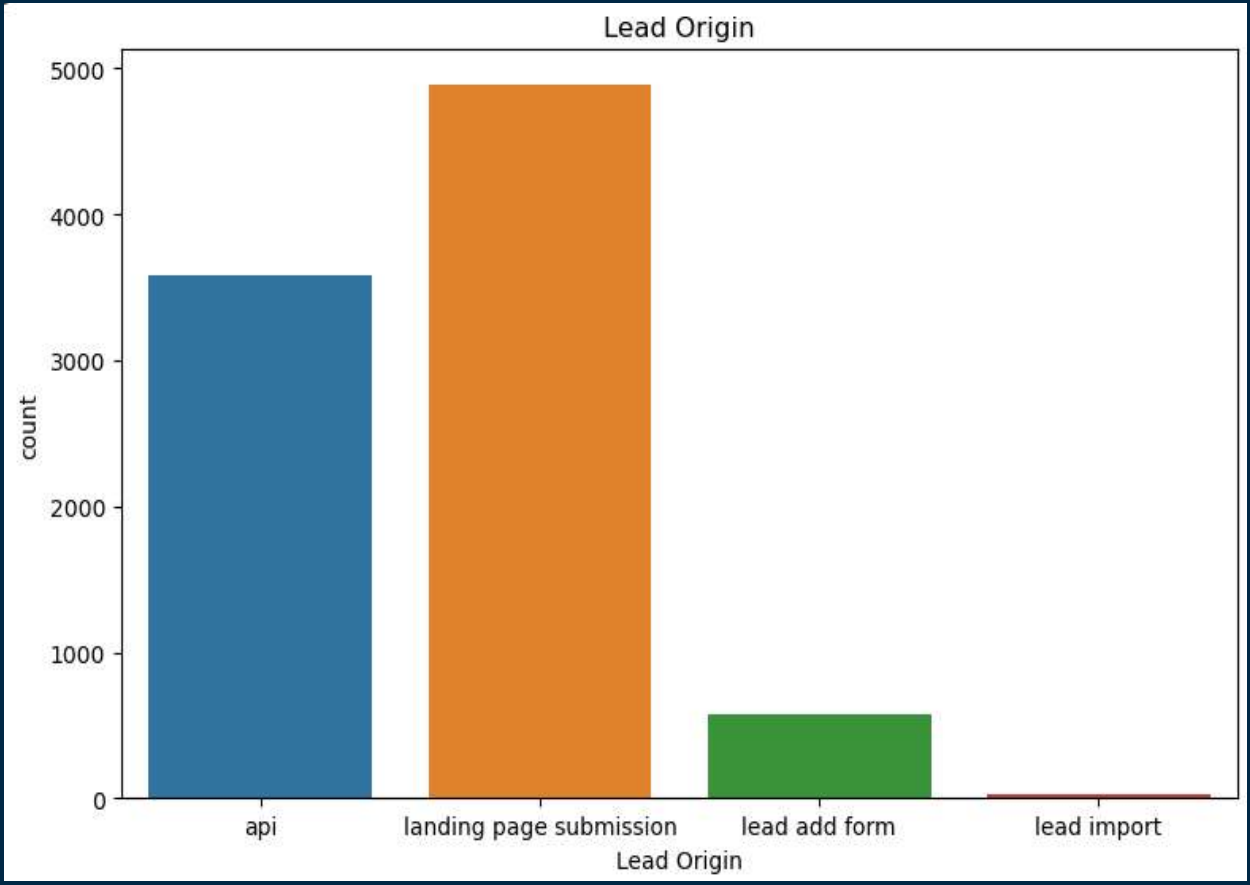
Model Presentation

Conclusions

DATA MANIPULATION

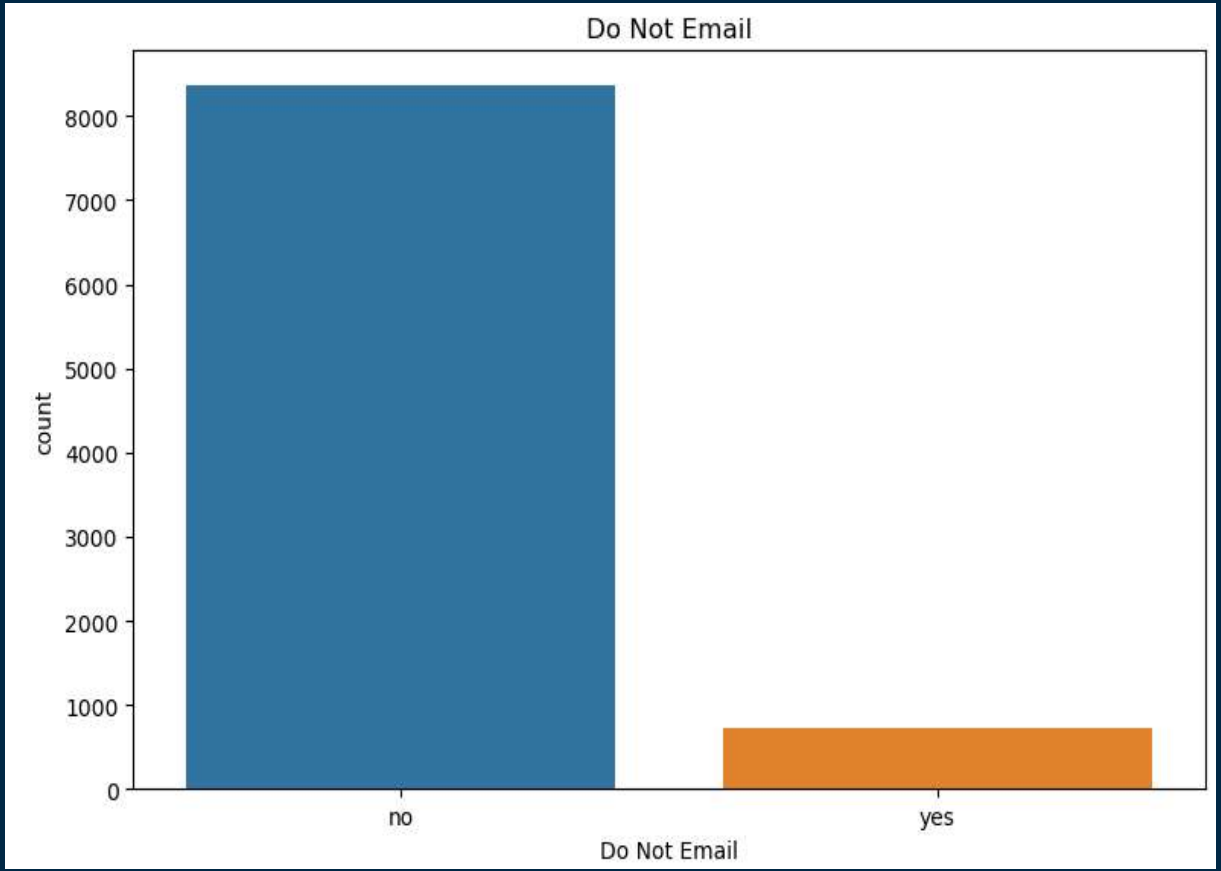
- Total Number of Rows =37, Total Number of Columns =9240.
- Removing the 'Prospect ID' and 'Lead Number' which is not necessary for the analysis.
- Single value features like 'Magazine', 'Receive More Updates About Our Courses', 'Update me on Supply'.
- Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque' etc. have been dropped.
- 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-1



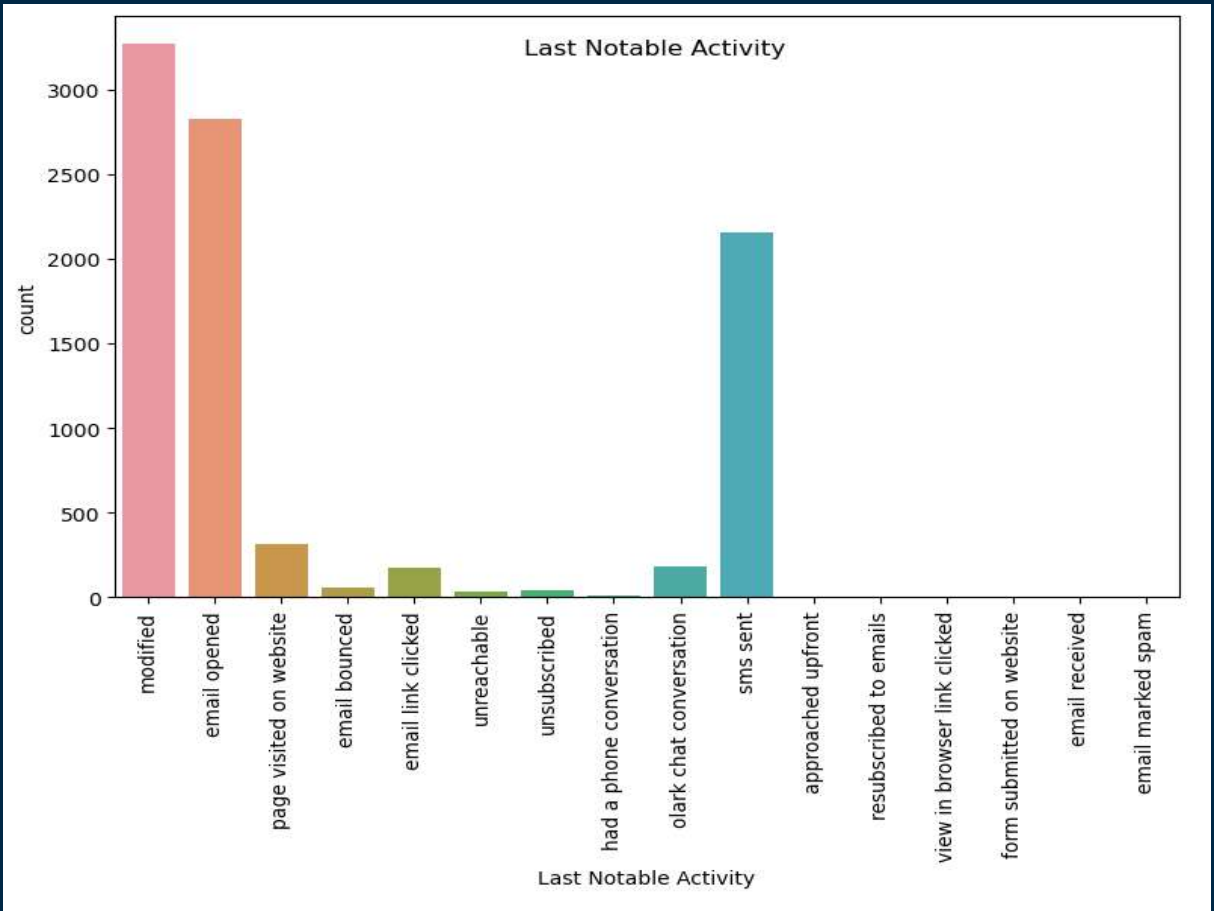
If you will see this Lead origin plot, it gives an idea that ‘Landing page submission’ has the highest value and ‘Lead Import’ has the lowest value.

EDA-2



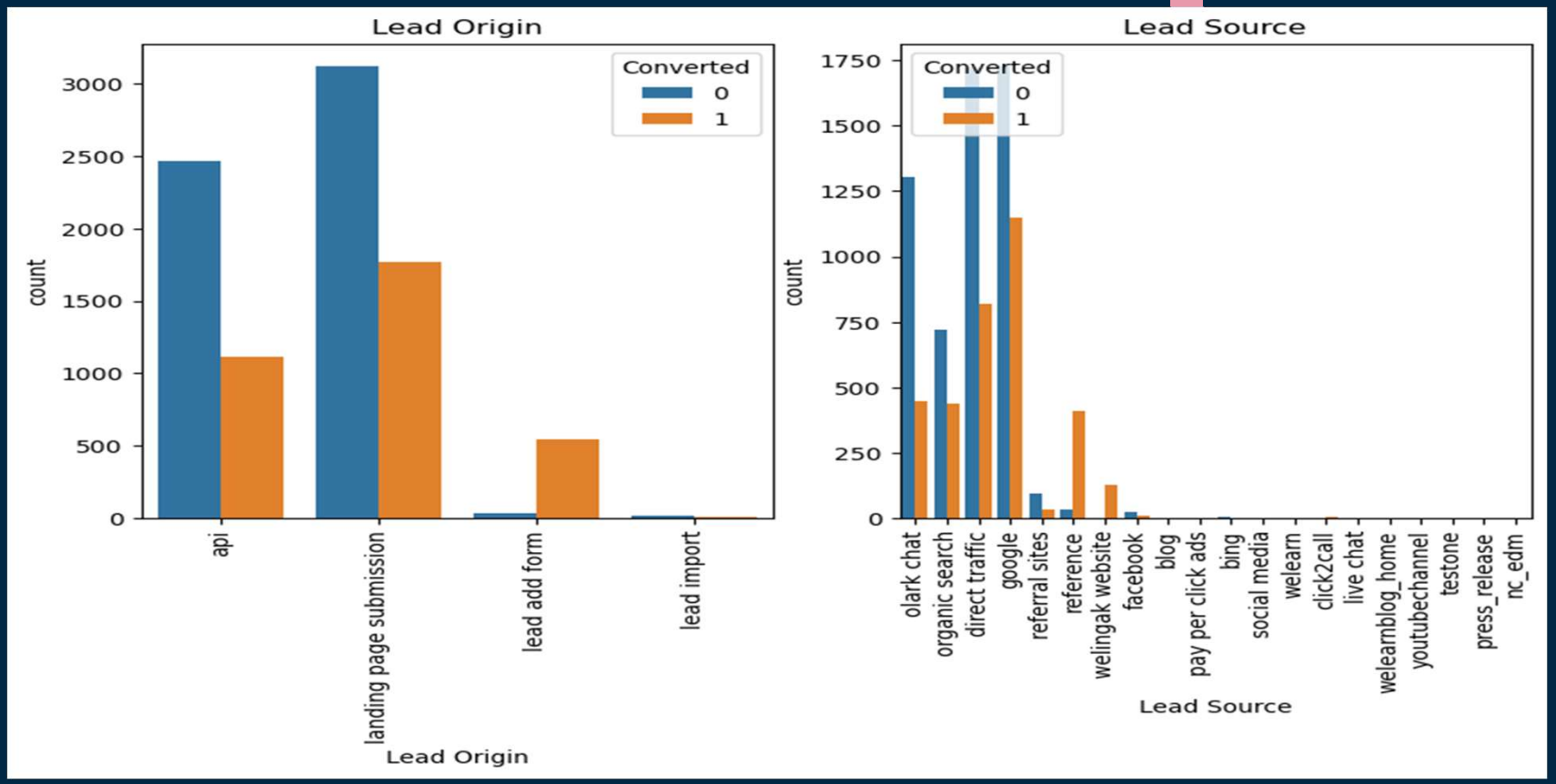
If you will see the ‘Do not Email’ plot, it gives an idea that ‘No’ has the highest value and ‘yes’ has the lowest value and we have act accordingly as per the plot.

EDA-3

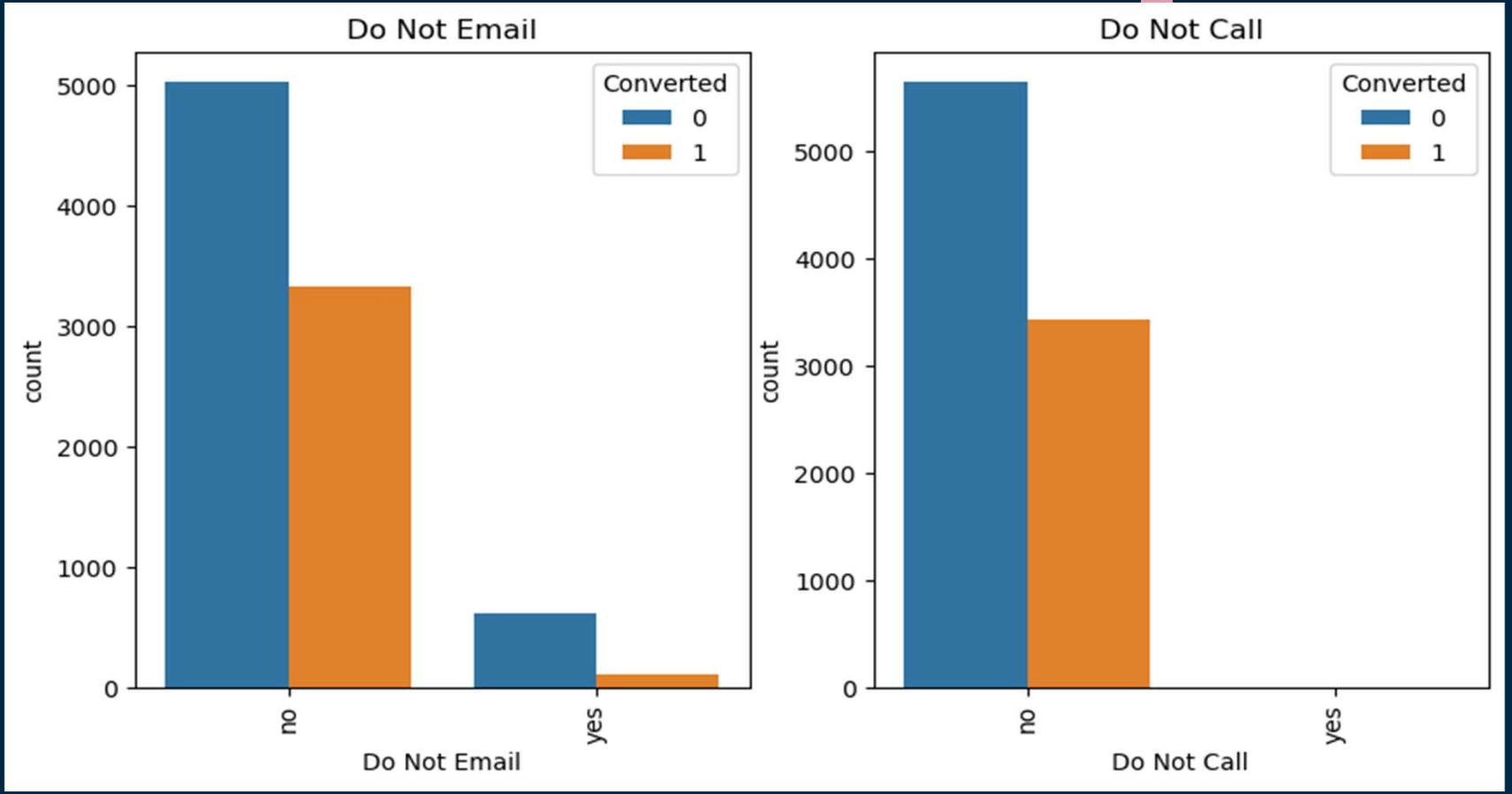


If you will see the ‘lost Notable Activity’ plot, it gives an idea that ‘Modified’, ‘Email opened’ and ‘SMS sent’ have the highest value and ‘others’ has the lowest value and we have act accordingly as per the plot.

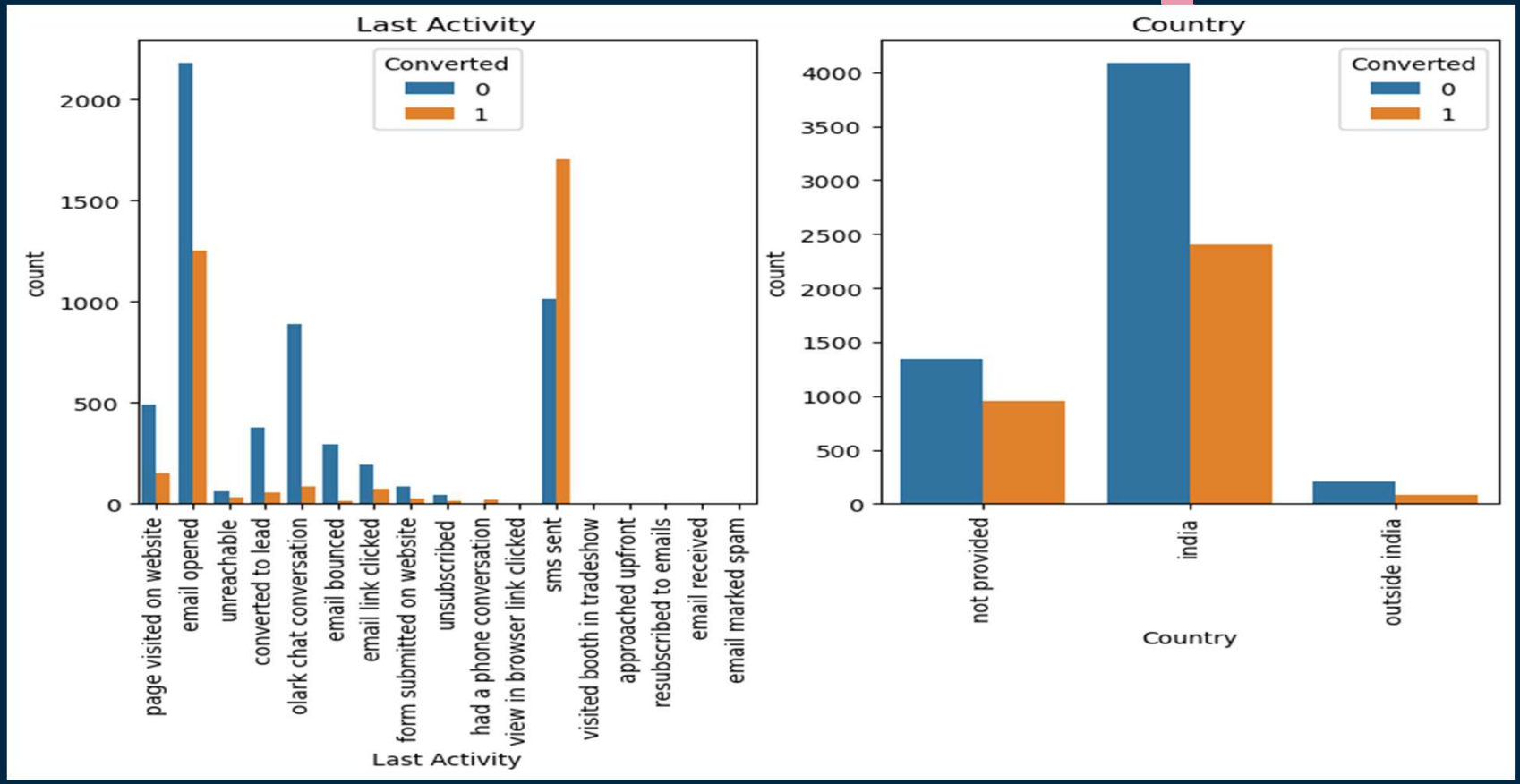
CATEGORICAL VARIABLE RELATION-1



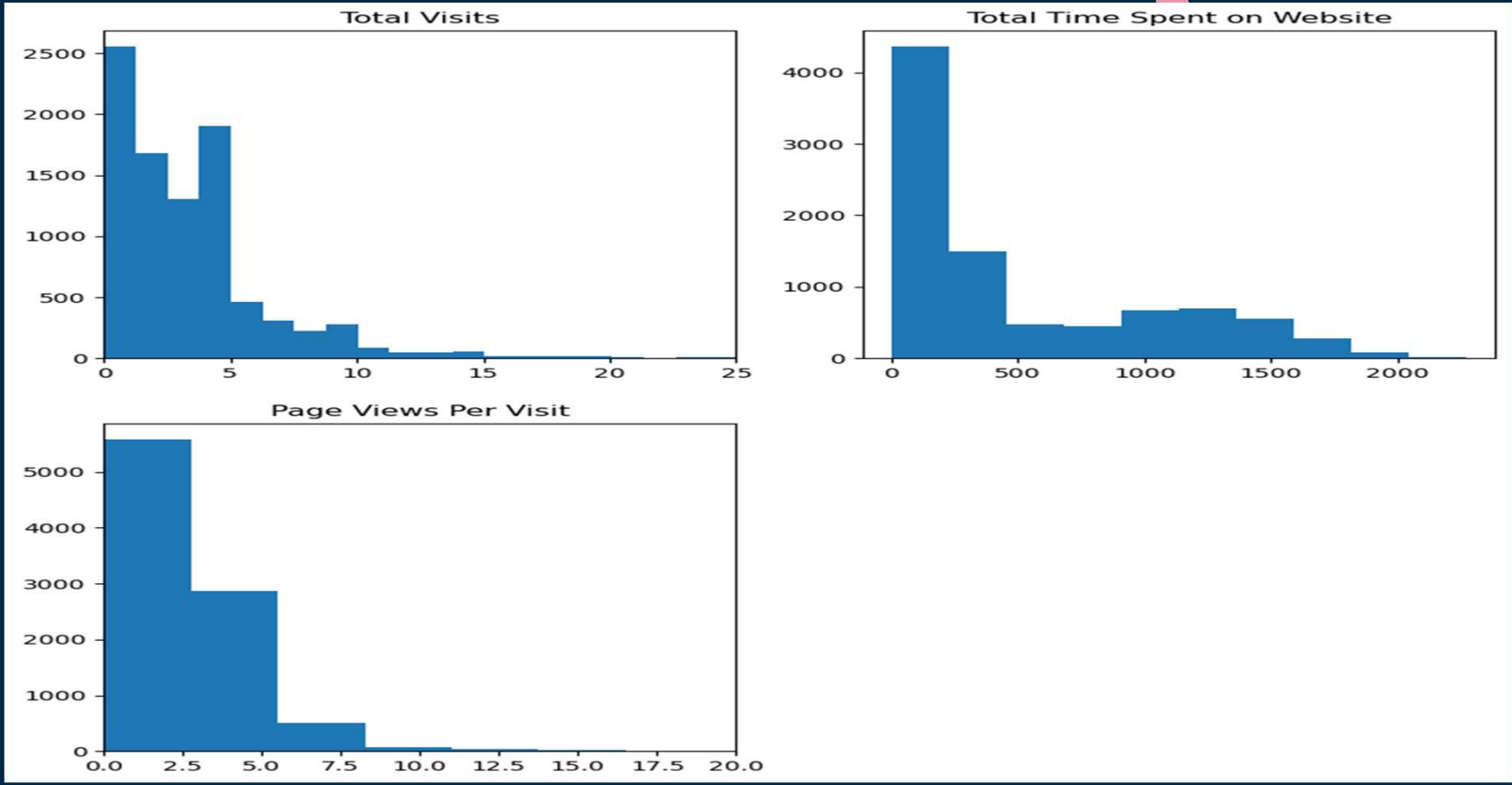
CATEGORICAL VARIABLE RELATION-2



CATEGORICAL VARIABLE RELATION-3



NUMERICAL VARIABLE RELATION



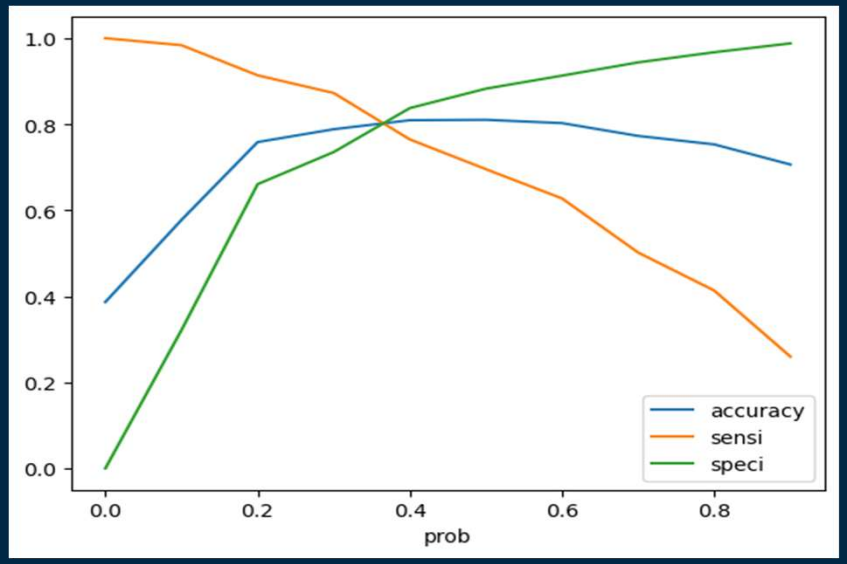
DATA CONVERSION

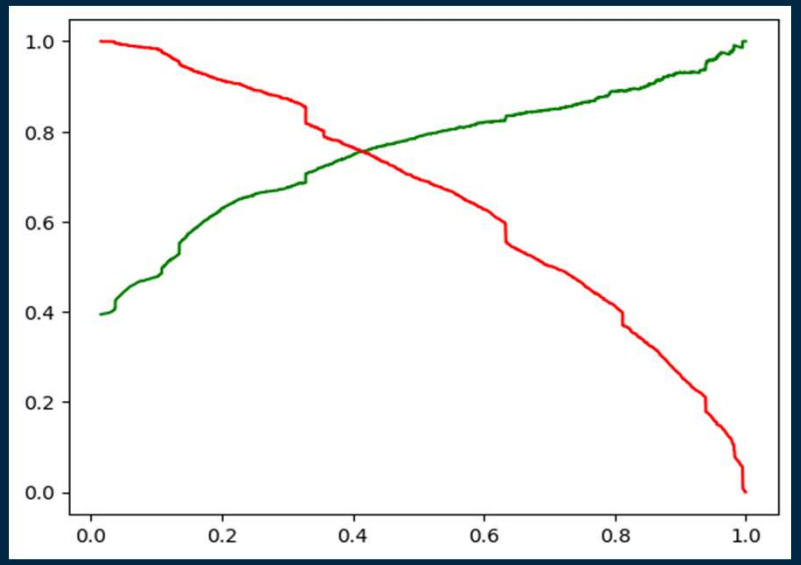
- Numerical Variables are Normalized
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 8792
- Total Columns for Analysis: 43

MODEL BUILDING

- Splitting into train and test data (70:30 ratio)
- Scale variables in train set
- Build the first model
- Use RFE to eliminate less relevant variables
- Running RFE on 15 variables
- Build the next model
- Eliminate the variable based on high p-values(greater than 0.05)
- Check VIF value for all the existing columns(remove if VIF > 5)
- Predict using train set
- Evaluate accuracy and other metric
- Predict using test set
- Precision and Recall Analysis on test predictions

MODEL EVALUATION(train)





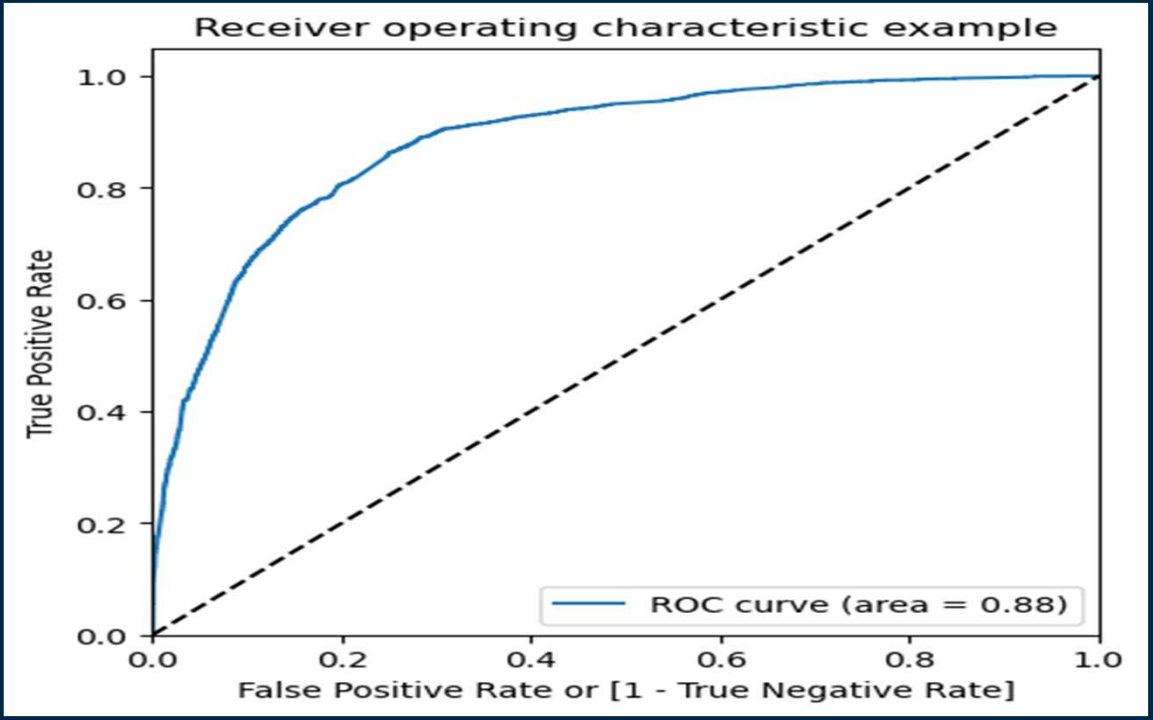
ACCURACY SENSITIVITY & SPECIFICITY

- 81% Accuracy
- 70% Sensitivity
- 88% Specificity

PRECISION AND RECALL

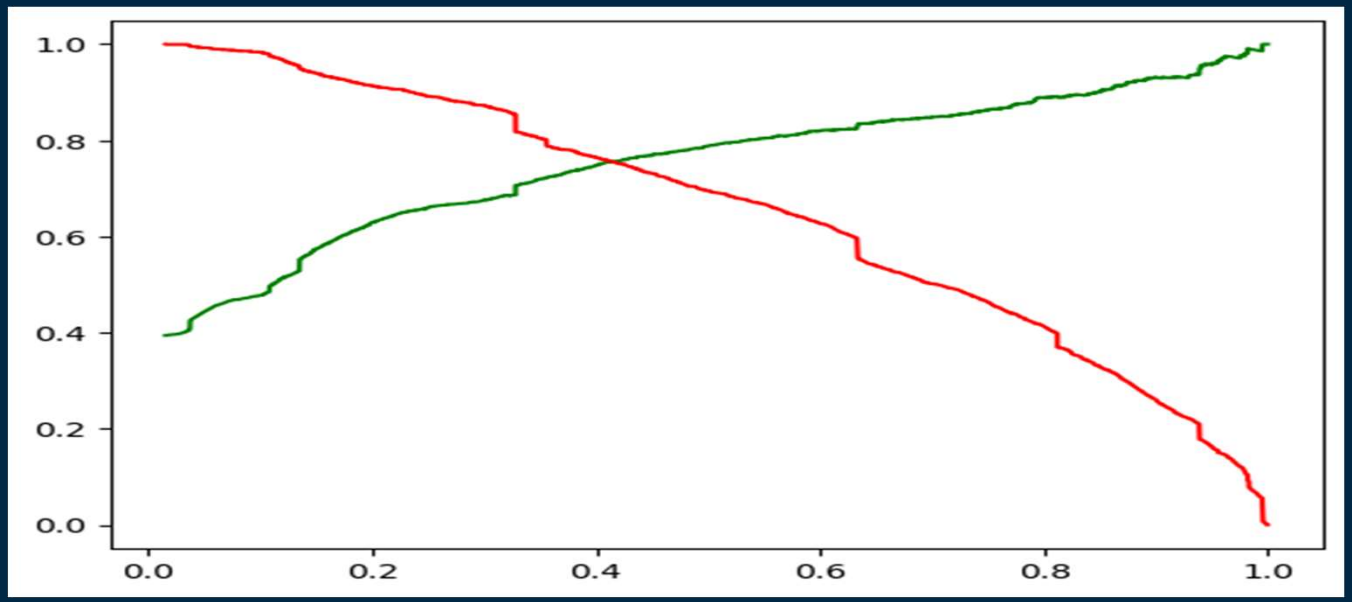
- 73% Precision
- 77% Recall

ROC CURVE



The area under ROC curve is 0.88 which is a very good value.

MODEL EVALUATION(test)



ACCURACY SENSITIVITY & SPECIFICITY

- 80% Accuracy
- 81% Sensitivity
- 80% Specificity

PRECISION AND RECALL

- 73% Precision
- 76% Recall

CONCLUSION-1

EDA:

- **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 good source for higher conversions.**
- **An alert messages or information has seen to have high lead conversion rate.**

CONCLUSION-2

LOGISTIC REGRESSION MODEL:

- **The model shows high close to 80% accuracy.**
- **The threshold has been selected from Accuracy, Sensitivity, specificity measures and precision, recall curves.**
- **The model shows 81% sensitivity and 80% specificity.**
- **The model finds correct promising leads and leads that have less chances of getting converted.**
- **Overall this model proves to be accurate.**


THANKS