LEAD SCORE CASE STUDY

GROUP MEMBERS

1. THERA BHUVANA CHANDRA
2. RAJAT DIXIT

PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

BUSINESS OBJECTIVE

- X Education wants to know most promising leads.
- For that they want to build a model which identifies the hot leads
- Deployment of the model for the future use.

SOLUTION

DATA INFORMATION

- Data set used : Leads.csv
- Total Customers present : 9240
- Total no of features : 37
- Model used: Linear Regression
- Target column in our dataset : “Converted”
- We need to reduce the features to maximize the conversion rate

BASIC STEPS

- Import data and all the libraries, inspect the data frame

DATA CELANING

- ✓ Check and handle duplicate data
- ✓ Check and handle NA values and missing values
- ✓ Drop columns which are having large missing values
- ✓ Imputation of values if necessary

EDA

- ▪ Univariate data analysis: value count, distribution of variable etc.
- ▪ 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 model
- ➢ Model presentation
- ➢ Conclusions
- ➢ Recommendations

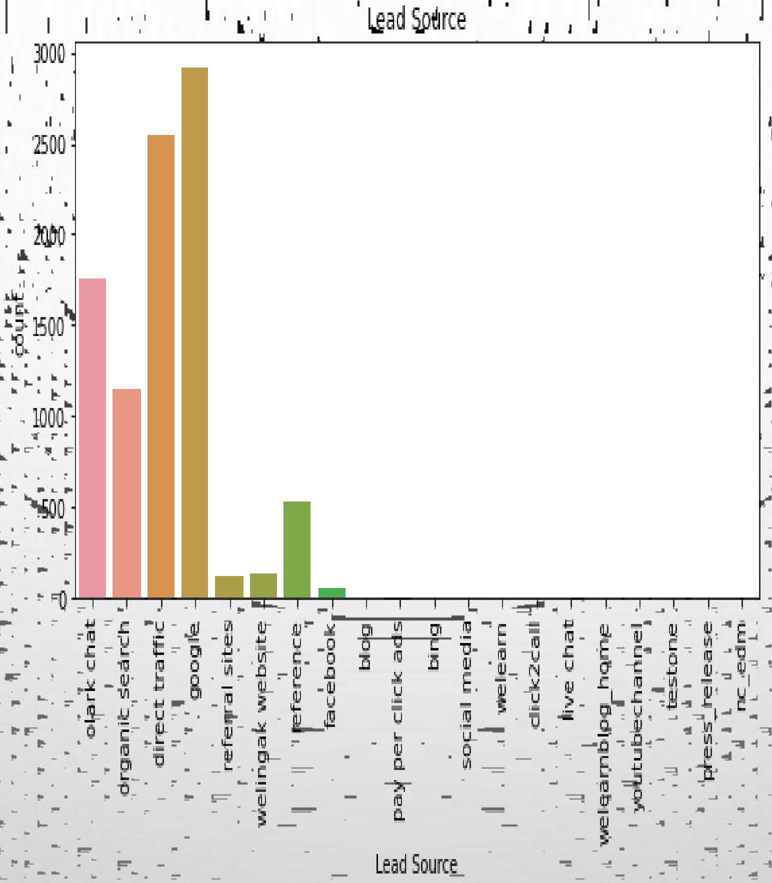
DATA CLEANING

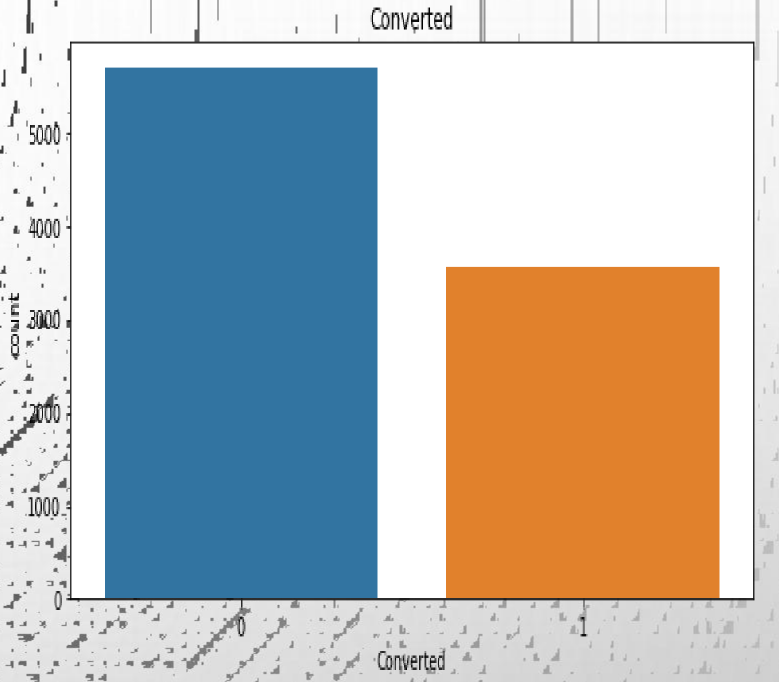
- Total no of rows = 37, total no of columns = 9240
- Single value features like 'Magazine', 'Receive more updates about our courses', 'update me on supply'.
- Dropping the columns having more than 40% missing values impute the missing value with mean & mode.
- Few of the Null values were changed to 'NA' so as to not lose much data. Although they were later removed while making dummies
- 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.

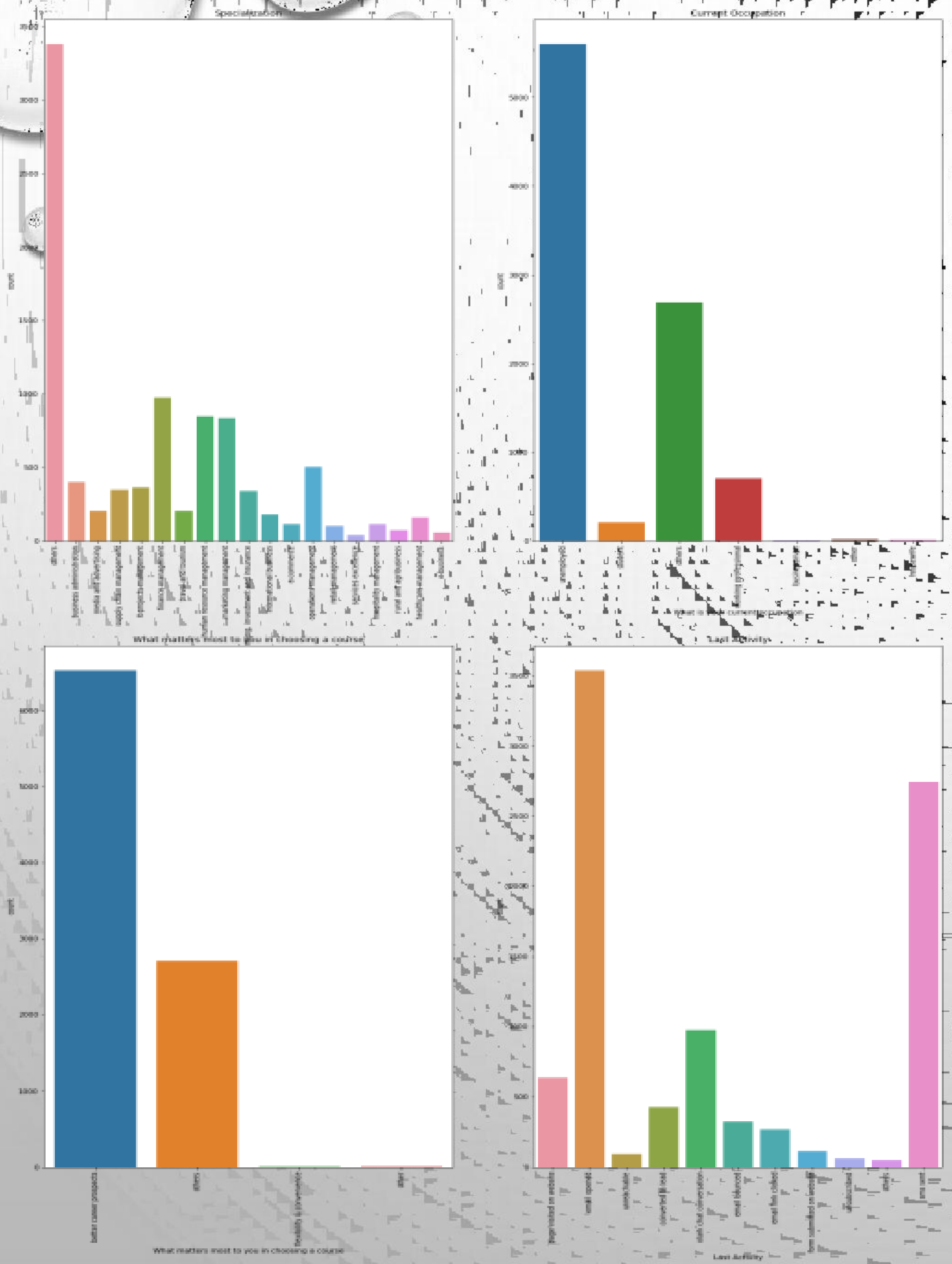
EDA

- EDA CONDUCTED ON OUR DATA FRAME
- LOT OF ELEMENTS IN CATEGORICAL VARIABLES WERE IRRELEVANT
- NUMERICAL VALUES ARE GOOD
- OUTLIERS IN NUMERICAL VARIABLES
- CAPPING ALL NUMERICAL VARIABLES

- CATEGORICAL ANALYSIS





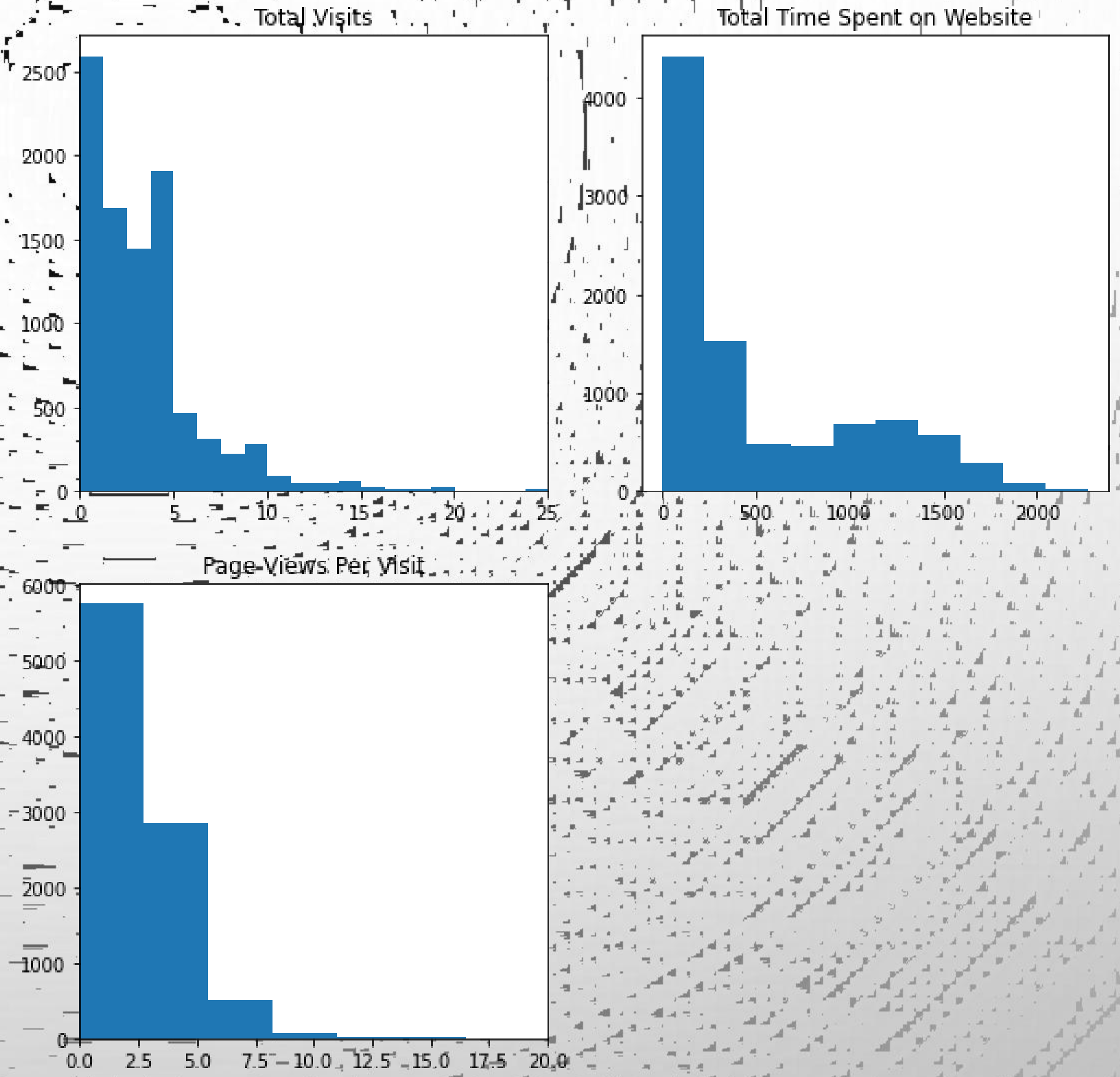


From the above slide the plot represents the Lead sources and from other variables converting into leads are:

- Google
- Direct Traffic
- Organic Search

Many are converting successful leads

NUMERICAL ANALYSIS



From the above plot we can clearly observe that “Total Visits ” and “Total Time Spent On Websites ” are top successful converters

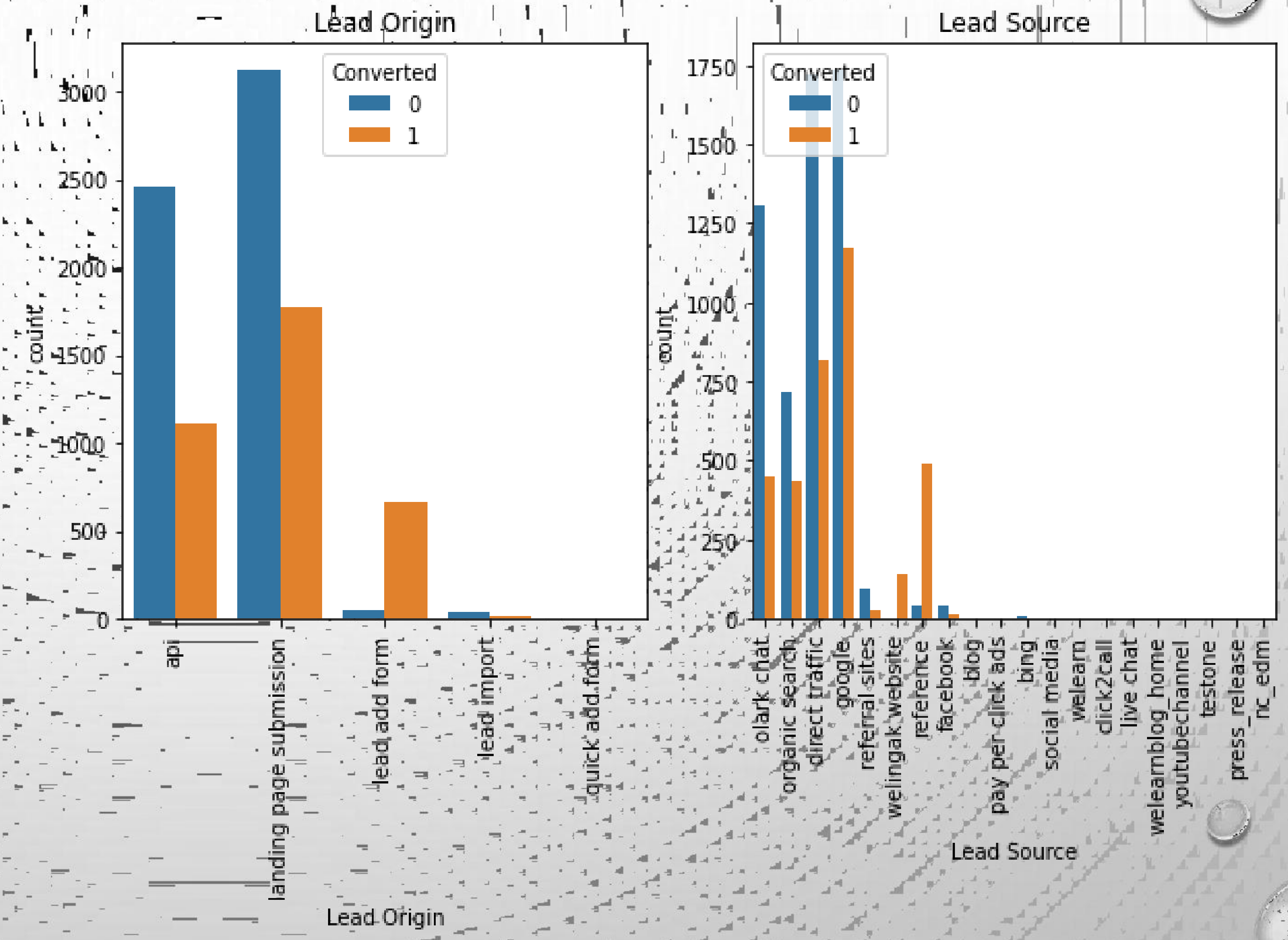
CATEGORICAL VARIABLES TO CONVERTED

LEAD ORIGIN:

- the percentage of converted people is found to be grater for landing page submission.
- we can see that lead origin add form and lead conversion ratio is very high.

LEAD SCORE

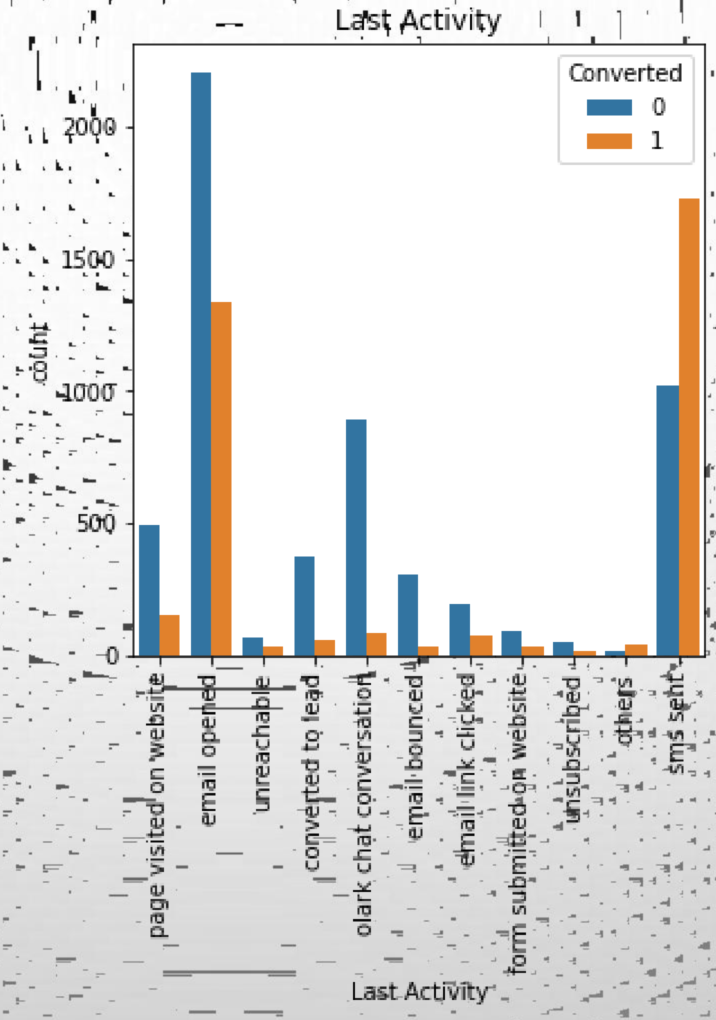
- Google is found to be one of the most important sources for lead conversion.
- Direct traffic also proves to be important to secure leads

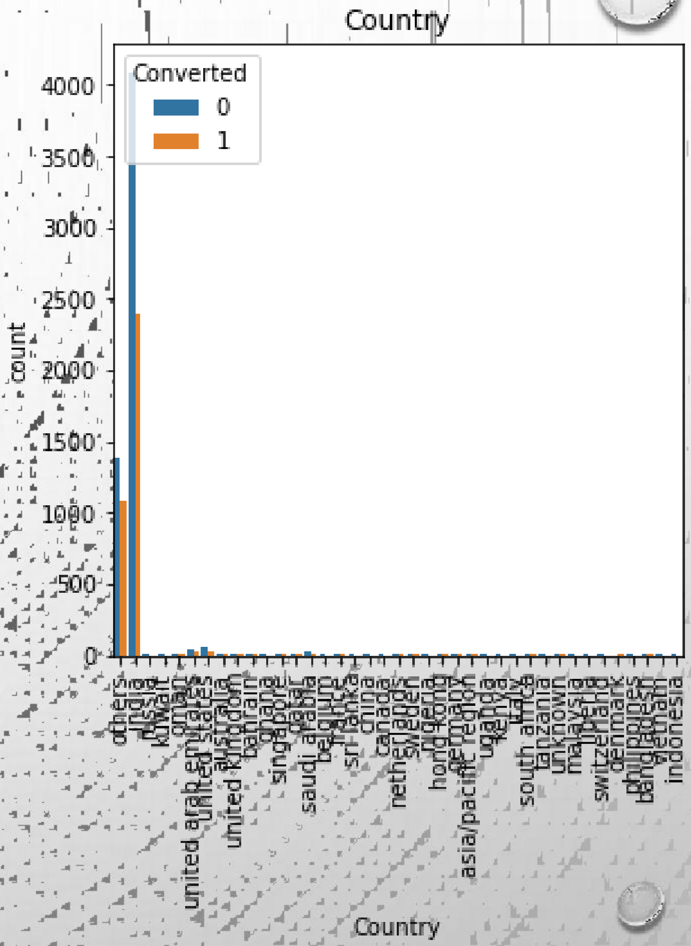


CATEGORICAL VARIABLES TO CONVERTED

LAST ACTIVITY

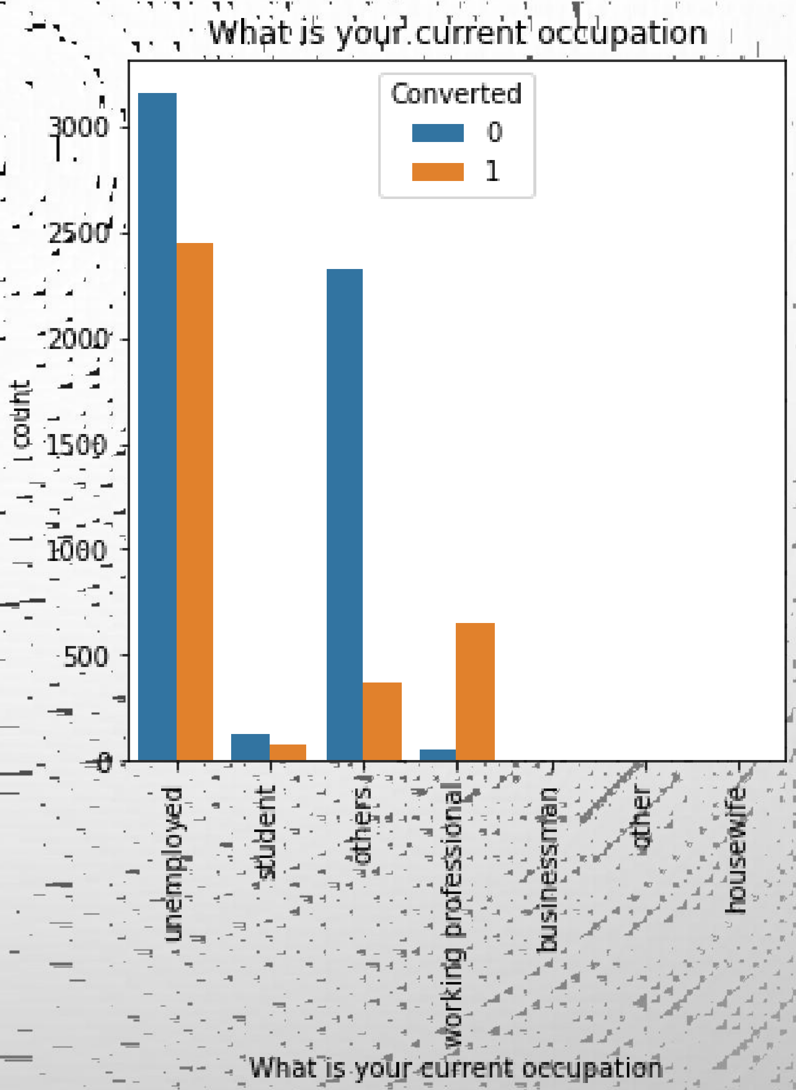
- RESPONSE ACTIVITY FOR SMS IS FOUND TO BE HIGH
- PEOPLE ALSO REPLY TO EMAILS, SO TARGETING PEOPLE VIA EMAILS IS EQUALLY IMPORTANT.





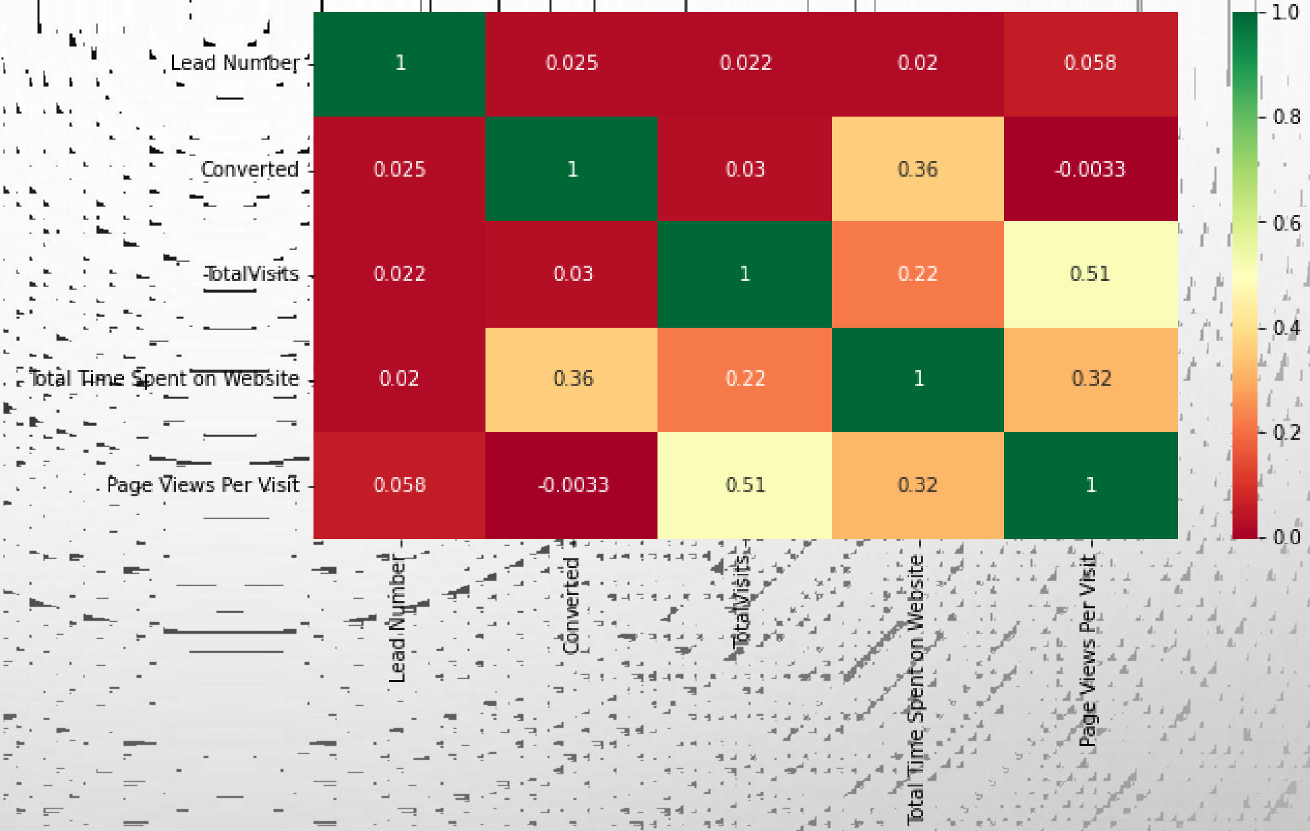
CATEGORICAL VARIABLES TO CONVERTED

As Observed From The Plot We Need To Target Unemployed And Working Professional As The Conversion Rate Is Found To Be Higher In These Cases.



HEATMAP

- From Above Heat Map It Is Observed That “Converted ” Column Has Strong Correlation.
- “Total Time Spent On Website Of Value” Is 0.36



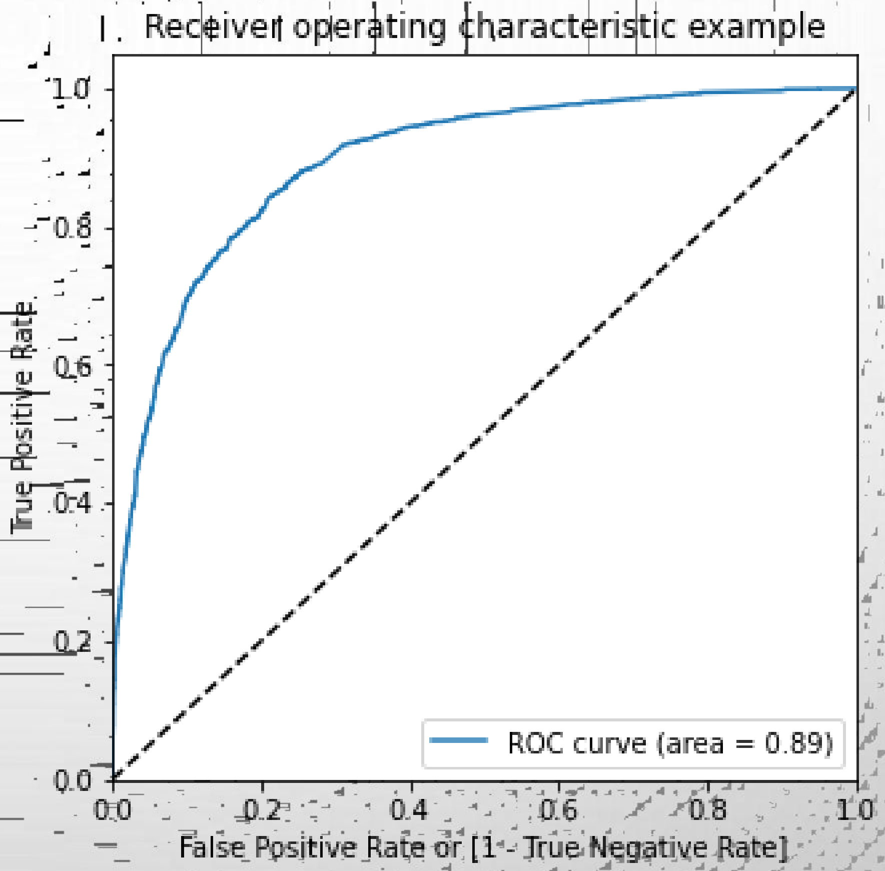
	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
Lead Number	1	0.025	0.022	0.02	0.058
Converted	0.025	1	0.03	0.36	-0.0033
TotalVisits	0.022	0.03	1	0.22	0.51
Total Time Spent on Website	0.02	0.36	0.22	1	0.32
Page Views Per Visit	0.058	-0.0033	0.51	0.32	1

MODEL BUILDING

- Splitting The Data Into Training And Testing Sets
- Basic Step For Regression Is Performing A Train-test Split, We Have Chosen 70:30 Ratio
- Use RFE For Feature Selection
- Building A Model By Removing The Variable Whose p-value Is Greater Than 0.05 And VIF Value Greater Than 5
- Predictions On Test Data Set
- Overall Accuracy 82%

MODEL EVALUATION

- A confusion matrix is made.
- Check overall accuracy found 82%
- Optimum cutoff value 0.5(ROC Curve) was used to find the accuracy
- The area under ROC curve is 0.89 which is very good value.



OPTIMAL PROBABILITY THRESHOLD

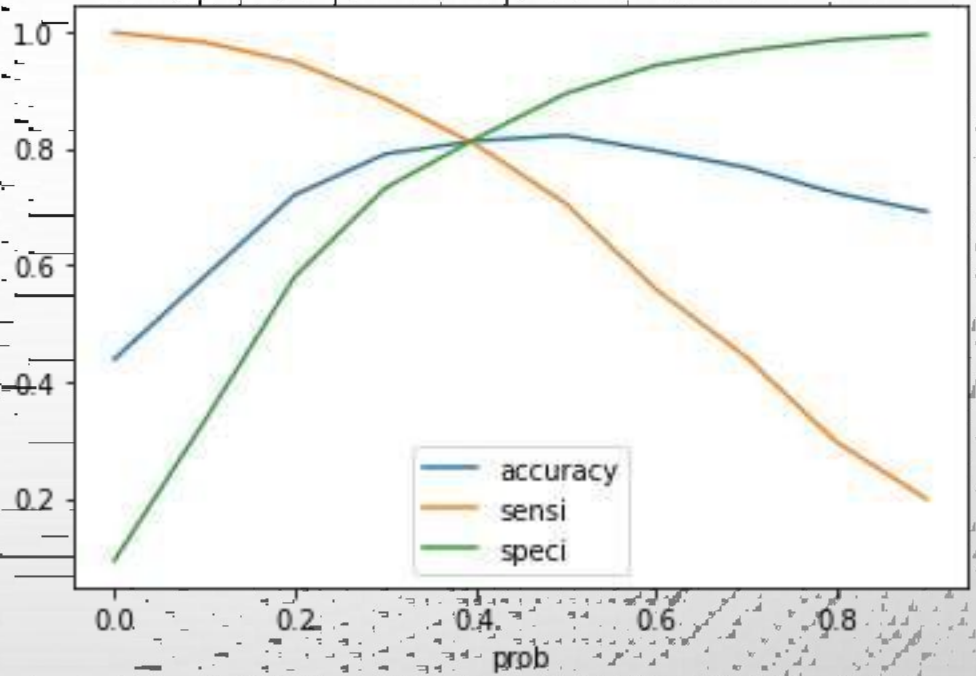
- FROM THE GRAPH IT IS VISIBLE THAT THE OPTIMAL CUT OFF IS 0.37
- CALCULATING SENSITIVITY

TP/(TP + FN) = 80.69%

- CALCULATING SPECIFICITY

TN/(TN+FP) = 81.60%

- OVERALL ACCURACY = 81.26%



PRECISION AND RECALL TRADEOFF

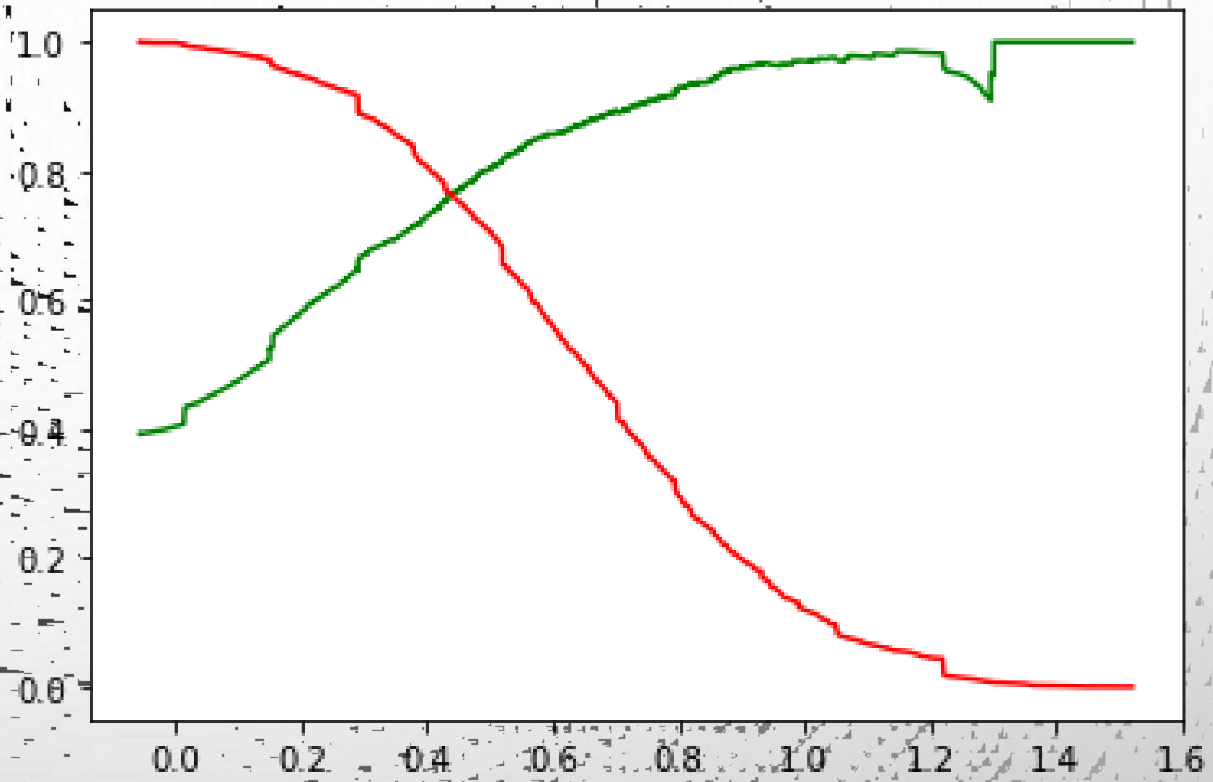
- FROM THE GRAPH IT IS VISIBLE THAT THE CURRENT CUT OFF IS 0.43
- PRECISION

TP/ TP + FP = 75.40%

- RECALL

TP/ TP+FN = 76.99%

- OVERALL ACCURACY = 81.46%



CONCLUSIONS AND RECOMMENDATIONS

- From model, we can conclude following points:

Focus on

- The total time spend on the website.
- Total number of visits.
- When the lead source was:
  A. Google
  B. Direct traffic
  C. Organic search
- We must majorly focus on leads whose last activity is SMS sent or email opened
- The customer/leads who fills the form are the potential leads.
- We must majorly focus on working professionals
- It’s better to focus least on customers to whom the sent mail is bounced back.

CONCLUSIONS AND RECOMMENDATIONS

Focus less

- If the lead source is referral, he/she may not be the potential lead.
- If the lead didn’t fill specialization, he/she may not know what to study and are not right people to target
- Page views per visit
- So, it’s better to focus less on such cases.

THANK YOU