LEAD SCORING CASE STUDY

BY:

DUVVURI SURYATHEJA REDDY

GAURAV RASAL

PROBLEM STATEMENT

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. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. 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. although X Education gets a lot of leads, its lead conversion rate is very poor. We must help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. We are required to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

SOLUTION METHODOLOGY

DATA CLEANING, TRANSFORMATION

- Imported the data
- Checked for null values, duplicates, dropping unnecessary columns
- Transformed data into more usable format for model building

EDA,DUMMY-VARIABLES,OUTLIER HANDLING

- Univariate, segmented univariate and bivariate analysis is done to check for patterns and correlations
- Created dummy variables for categorical variables
- Removed outliers for numerical variables

TRAIN-TEST SPLIT,FEATURE SCALING

- Spitted data into train and test data
- Scaling is done on all numeric features

RFE, MODEL BUILDING,MODEL EVALUATION

- Selecting top 15 features using RFE ,and building a model using these 15 features
- Reduction of columns based on p-values and VIFs of columns and recreating the model
- After selecting final model, evaluating it using several evaluation metrics

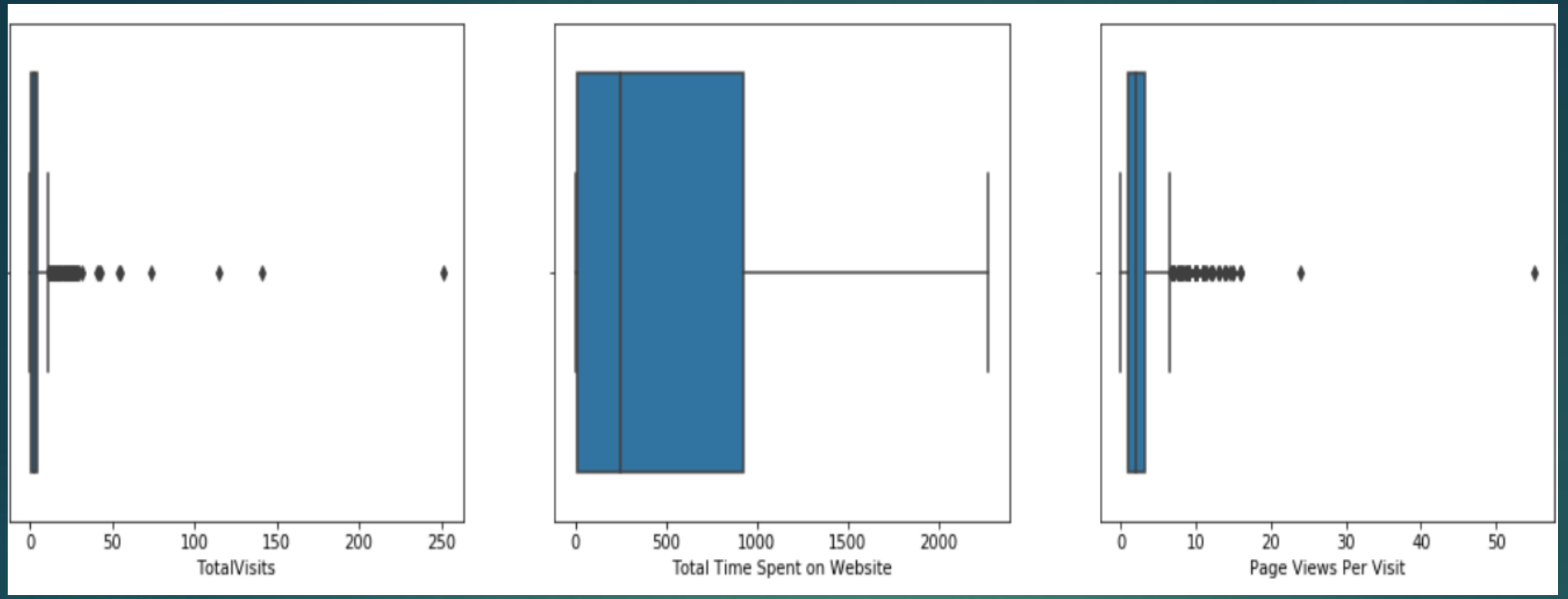
ROC,OPTIMAL CUT-OFF,PRECISION,RECALL

- Plotting ROC curve
- Finding optimal cut-off using sensitivity-specificity curve
- Plotting precision-recall curve

PREDICTIONS ON TEST SET,PERFORMANCE ANALYSIS

- Predictions of test set using final model is done
- Performance of the model based on test set predictions is analyzed
- Summary of the model is written at the end

Visualizations

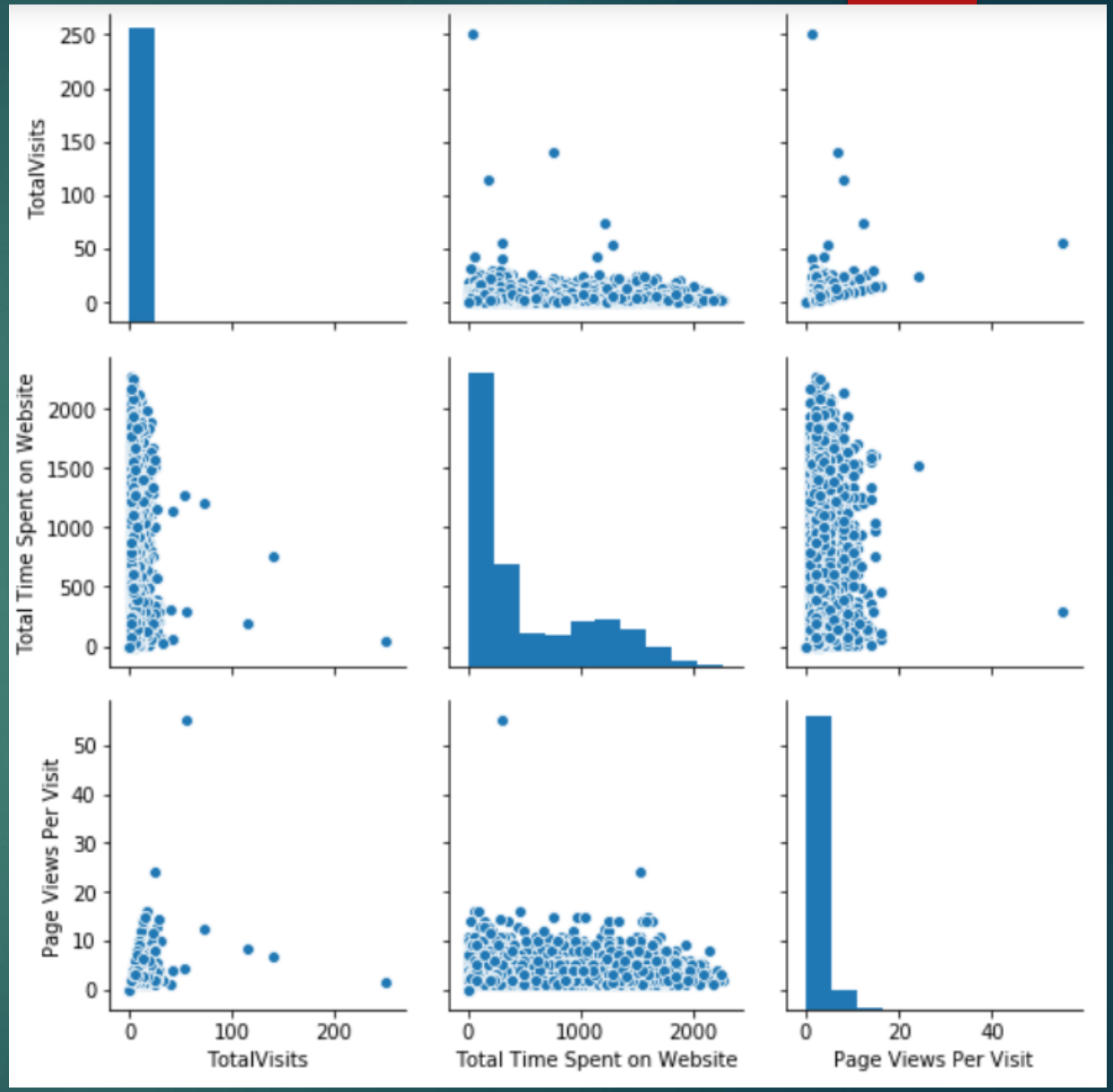


UNIVARIATE ANALYSIS OF NUMERIC COLUMNS:

- Most of the total times spent on website are in between range 250 and 2000

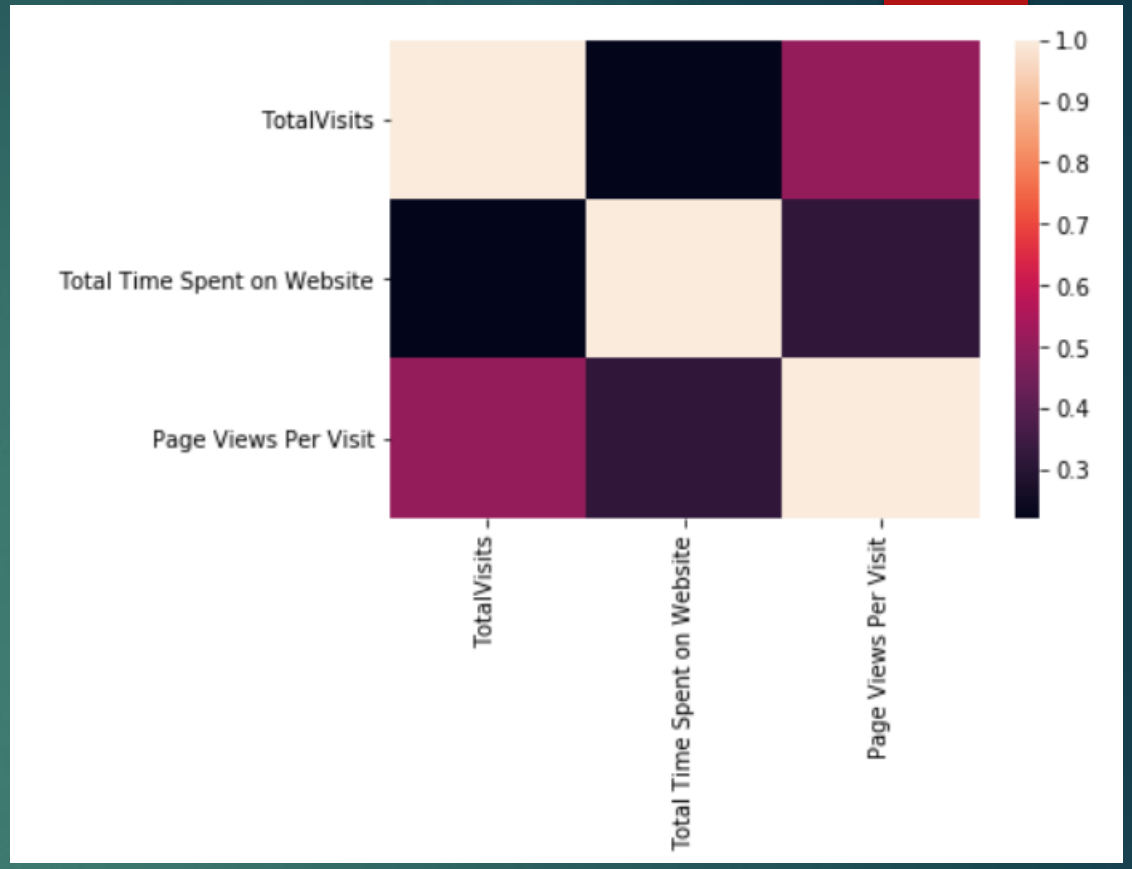
BIVARIATE ANALYSIS OF NUMERIC COLUMNS

- Most of the total visits are in the range between 0 and 100
- Most of the page views per visit are in the range between 0 and 20

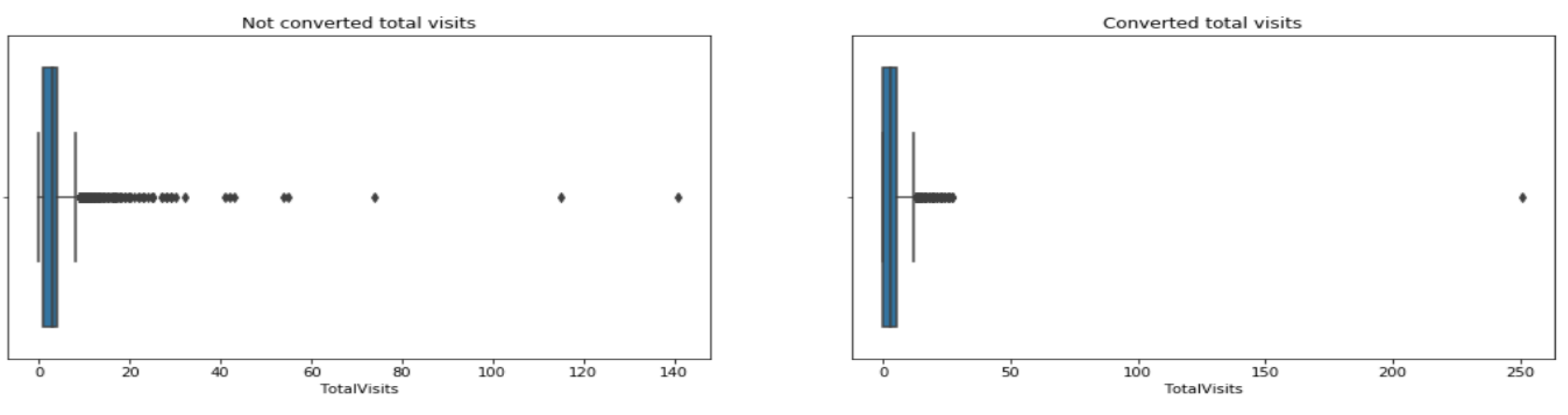


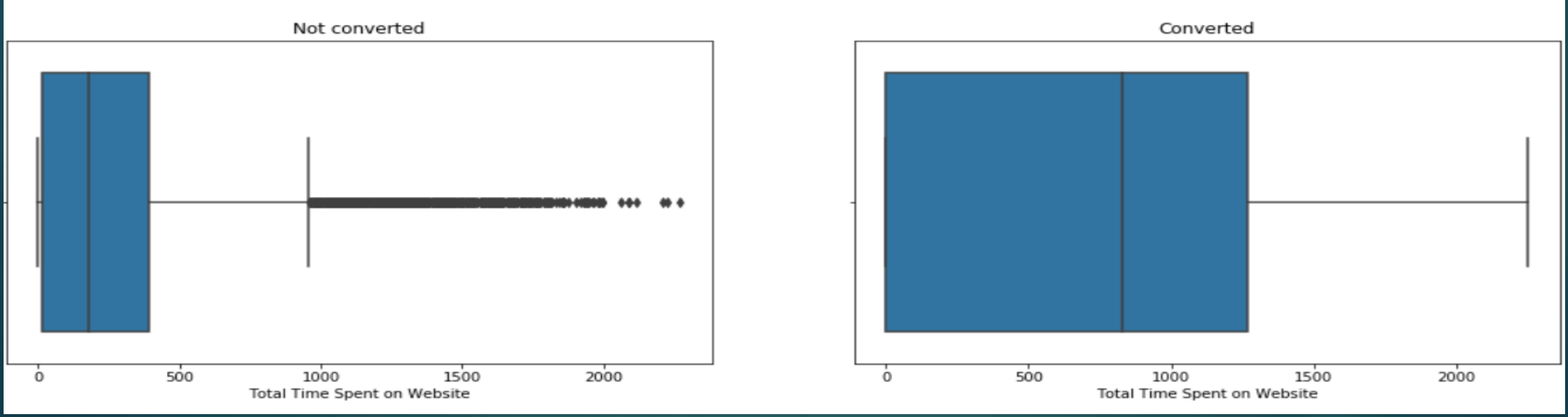
CORRELATION MATRIX OF NUMERIC COLUMNS

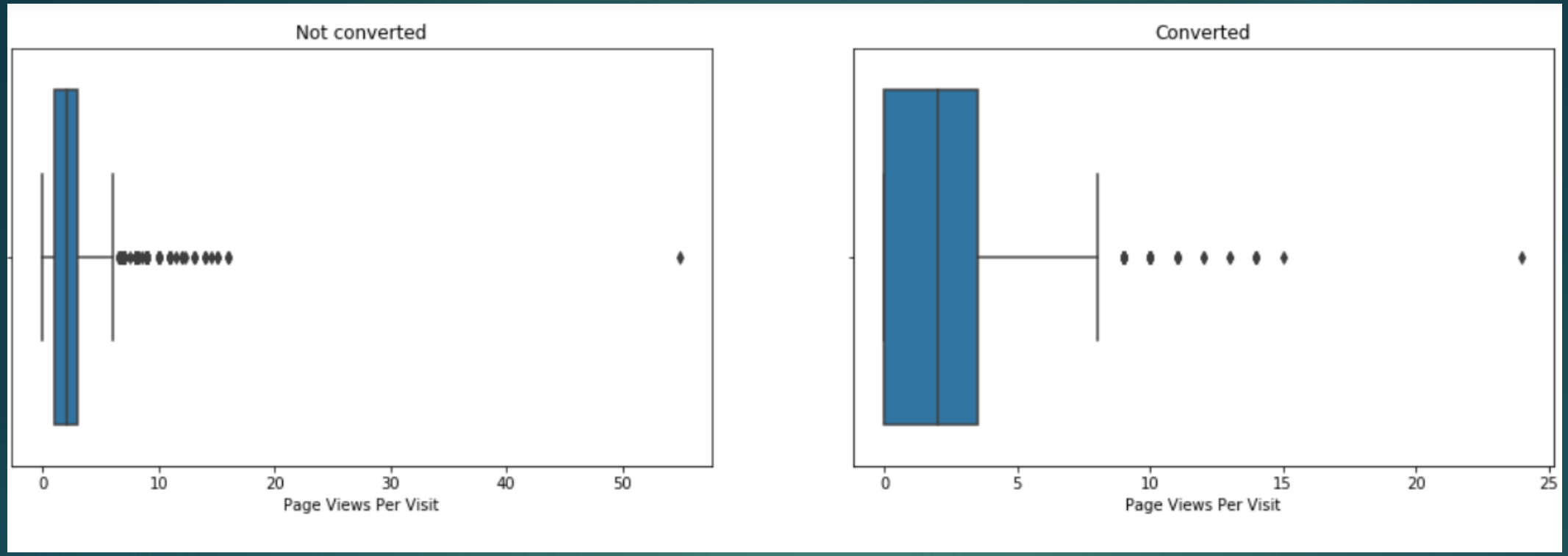
- There is a slight correlation between Totalvisits and page views per visit
- There is a weak correlation between Total time spent on website and total visits



SEGMENTED UNIVARIATE ANALYSIS OF NUMERIC COLUMNS

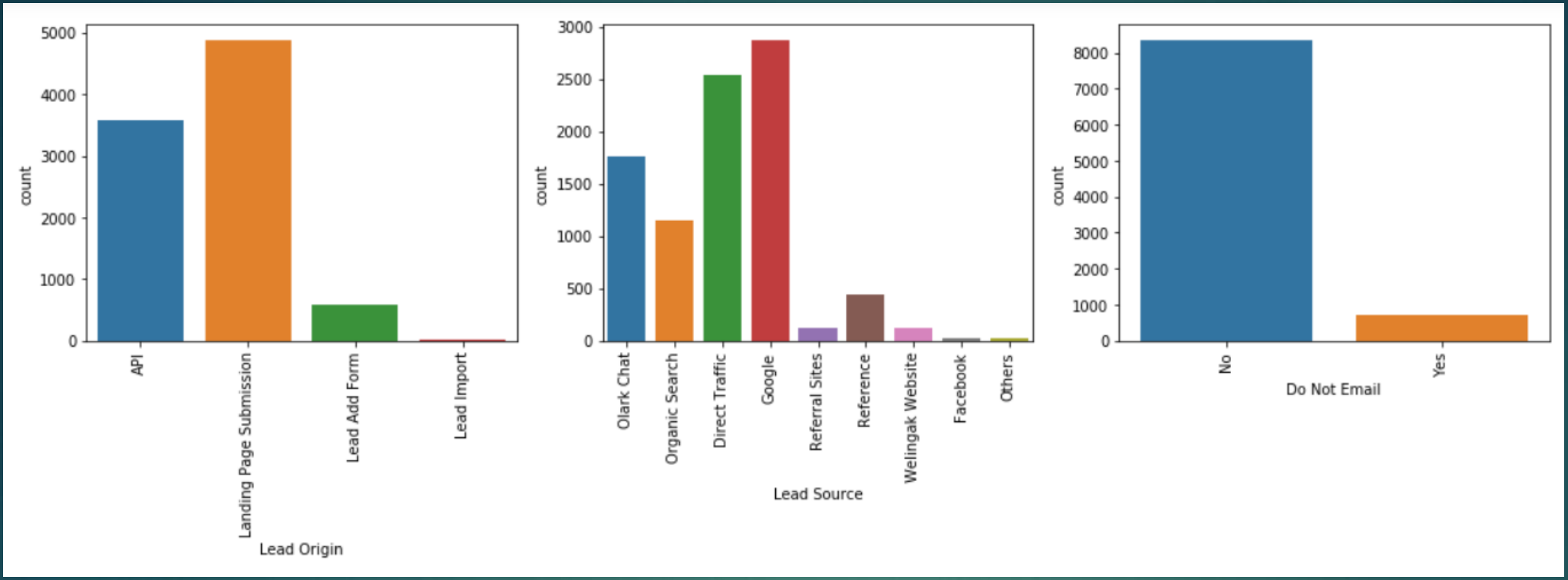






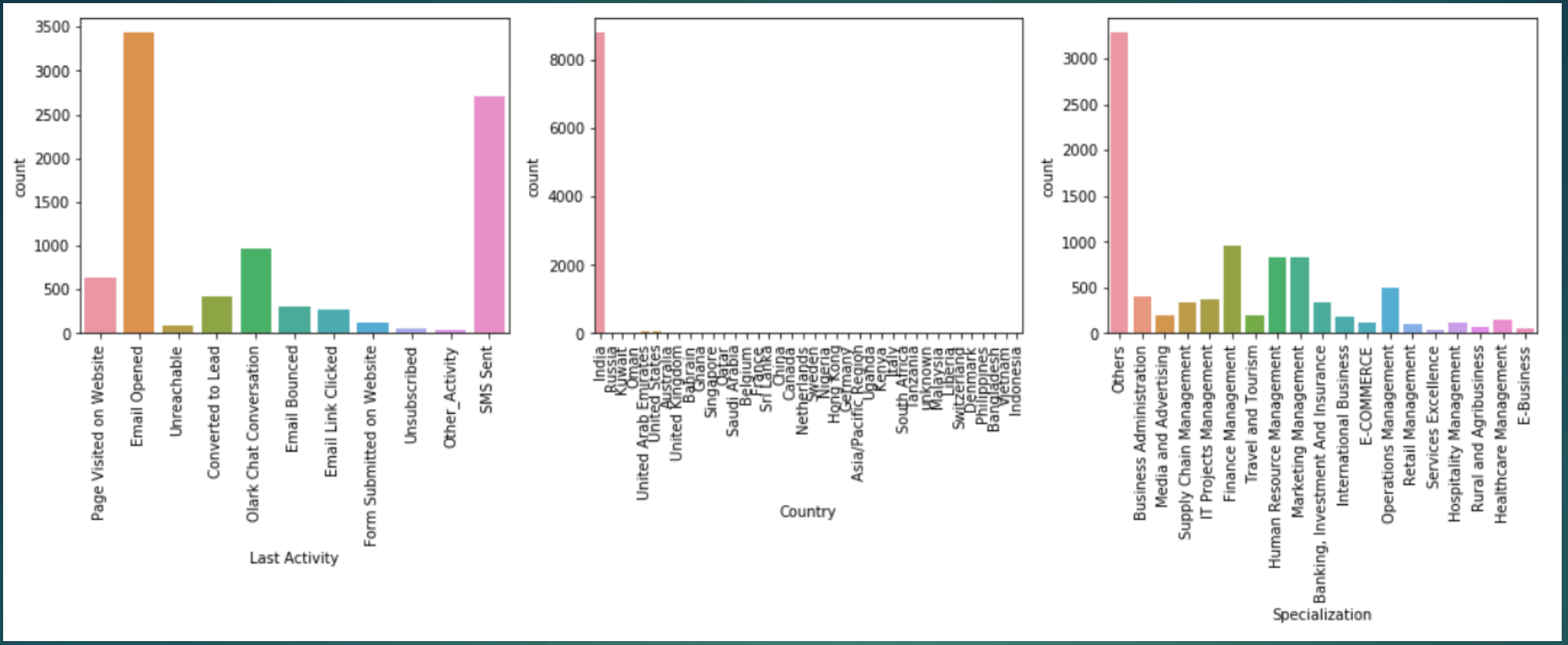
INSIGHTS OF ALL THREE SEGMENTED UNIVARIATE PLOTS OF NUMERIC COLUMNS

- Most of the total times spent on website in converted people are in between 0 and 750
- total time spent Median of not converted is less than total time spent median of converted.
- Most of the Page views per visit in converted are in range between 0 and 2.



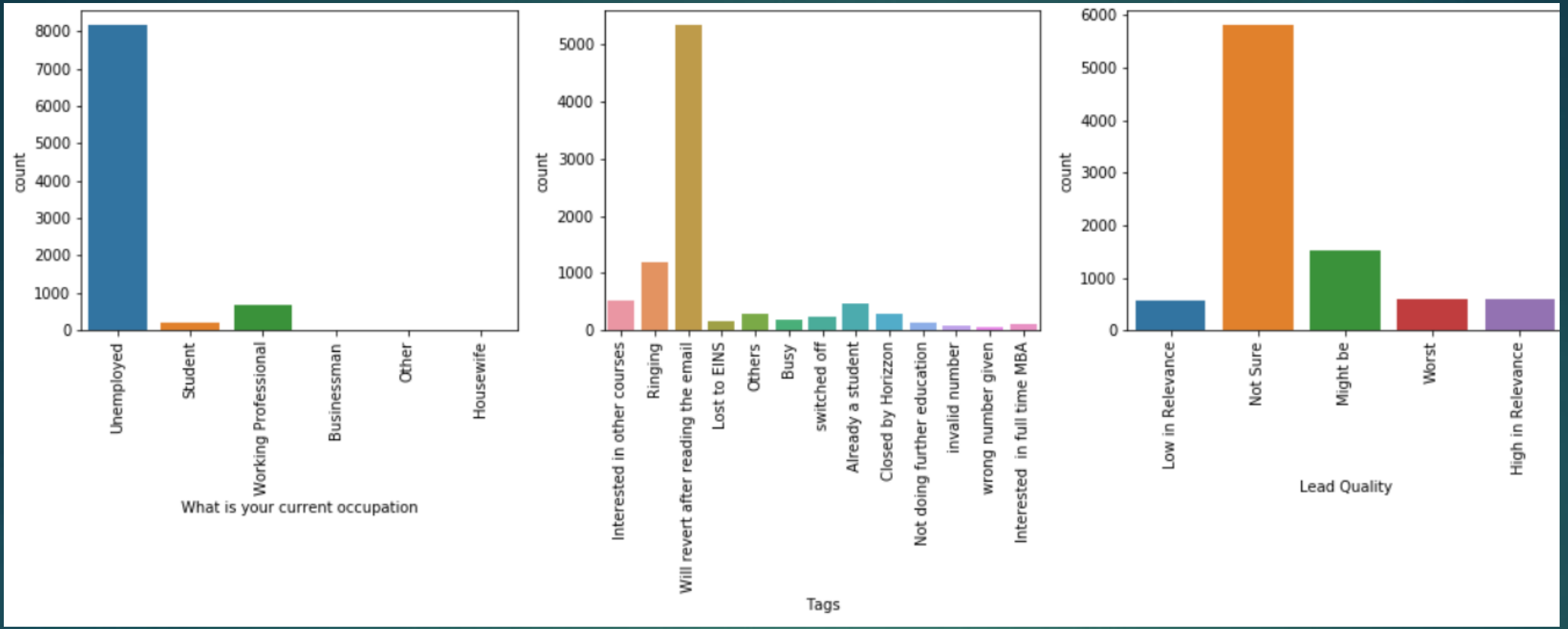
UNIVARIATE ANALYSIS OF CATEGORICAL COLUMNS:

- Most of the lead origins are from landing page submission and least lead origins are from lead import
- Most of the lead sources are from Google,direct traffic
- In Do Not Email, No has major count



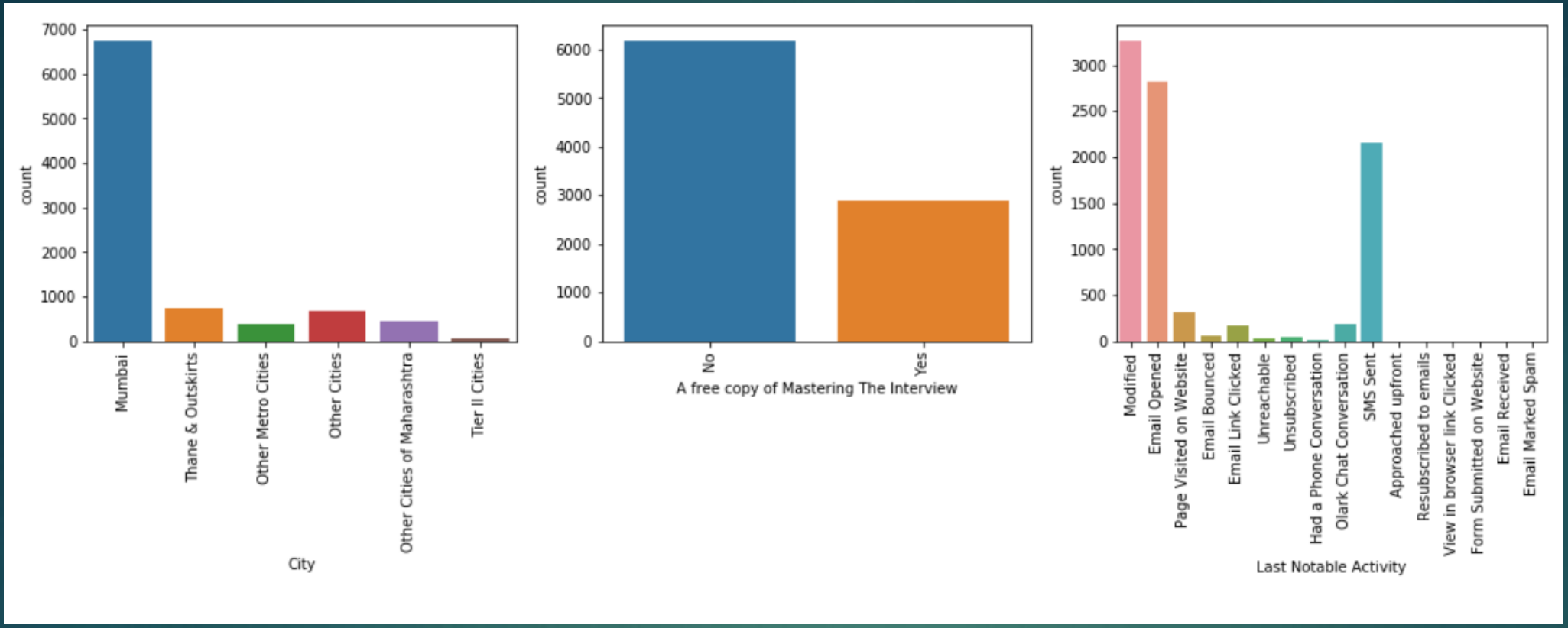
UNIVARIATE ANALYSIS OF CATEGORICAL COLUMNS:

- Among last activities did by customers, Email opening tops the list, next comes SMS sent category
- Most of the customers are from India
- Some of the specializations which have most customers are finance management,human resource management, others



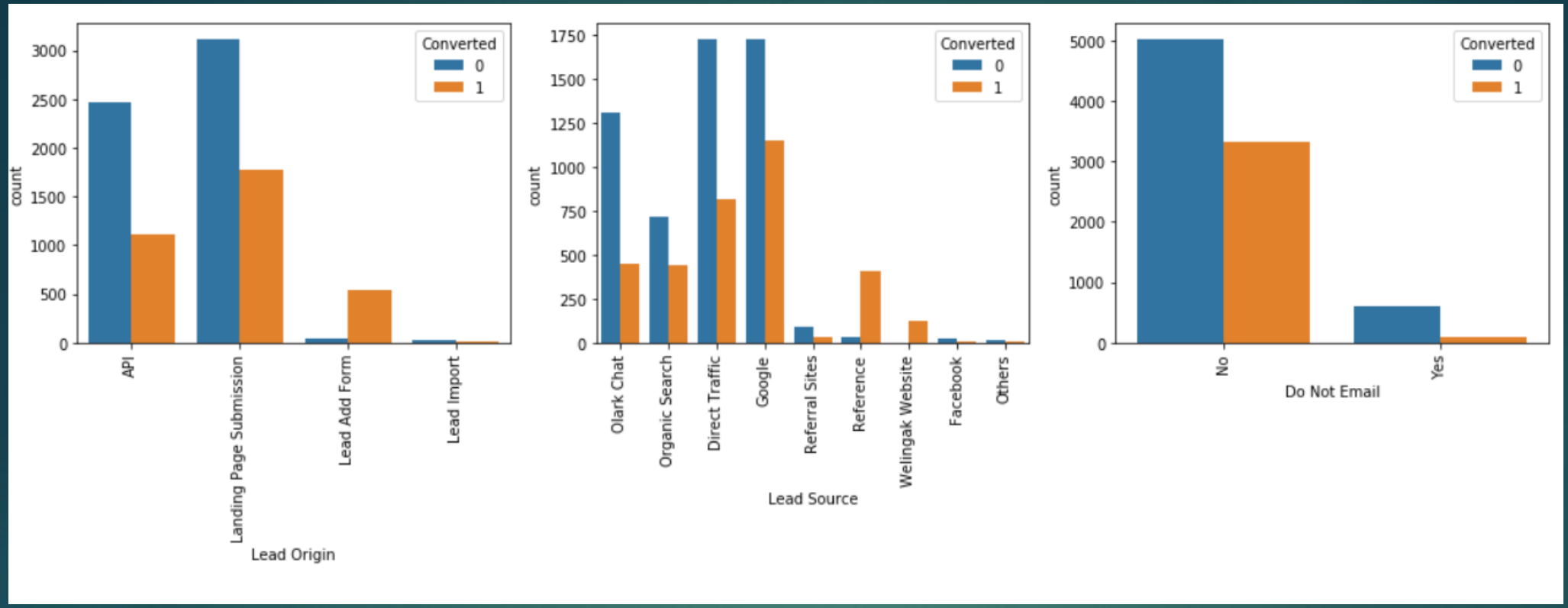
UNIVARIATE ANALYSIS OF CATEGORICAL COLUMNS:

- Most of the customers current occupation is unemployed
- Most of the customers are tagged 'Will revert after reading the email'
- Most of the lead quality from the data we got belongs to not sure category



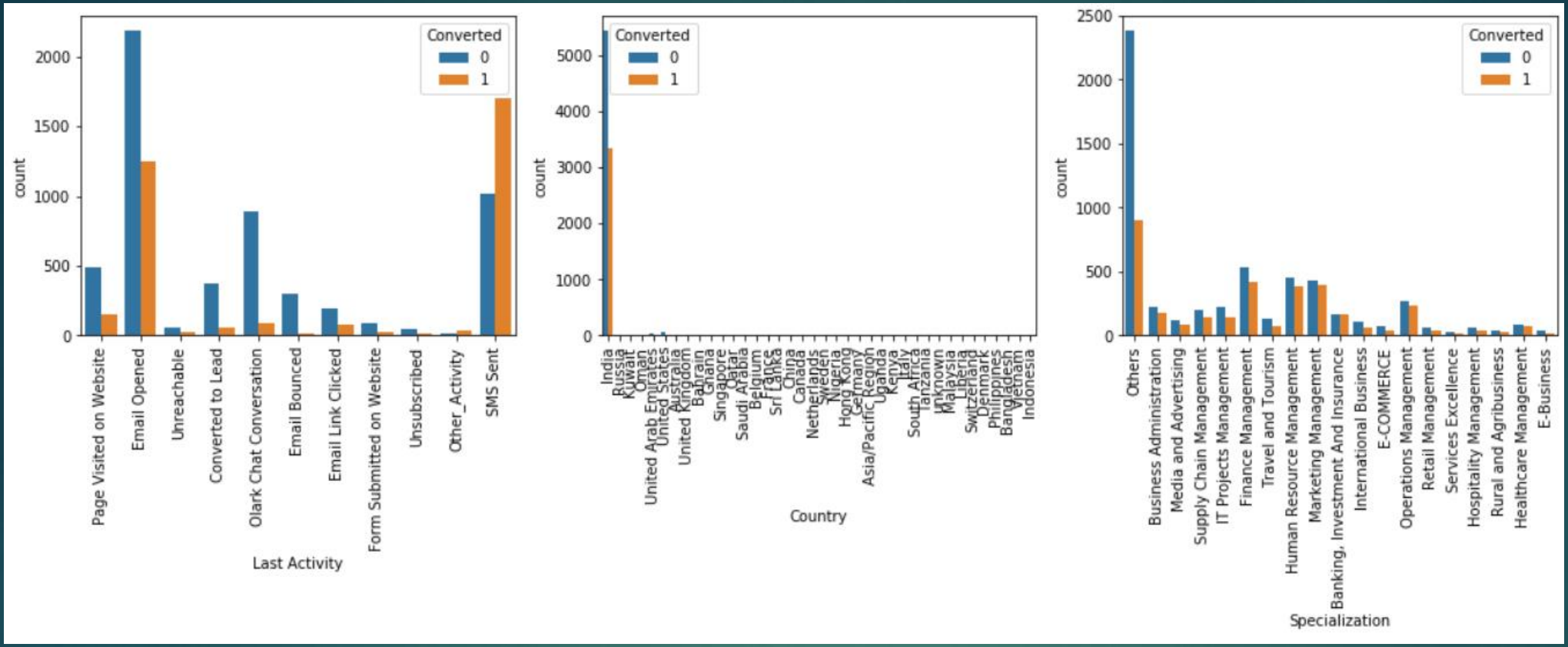
UNIVARIATE ANALYSIS OF CATEGORICAL COLUMNS:

- Most of the customers don't want a free copy of mastering the interview
- Modified is the category which tops in last notable activity, next comes up Email opened and SMS sent category
- most of the customers live in Mumbai



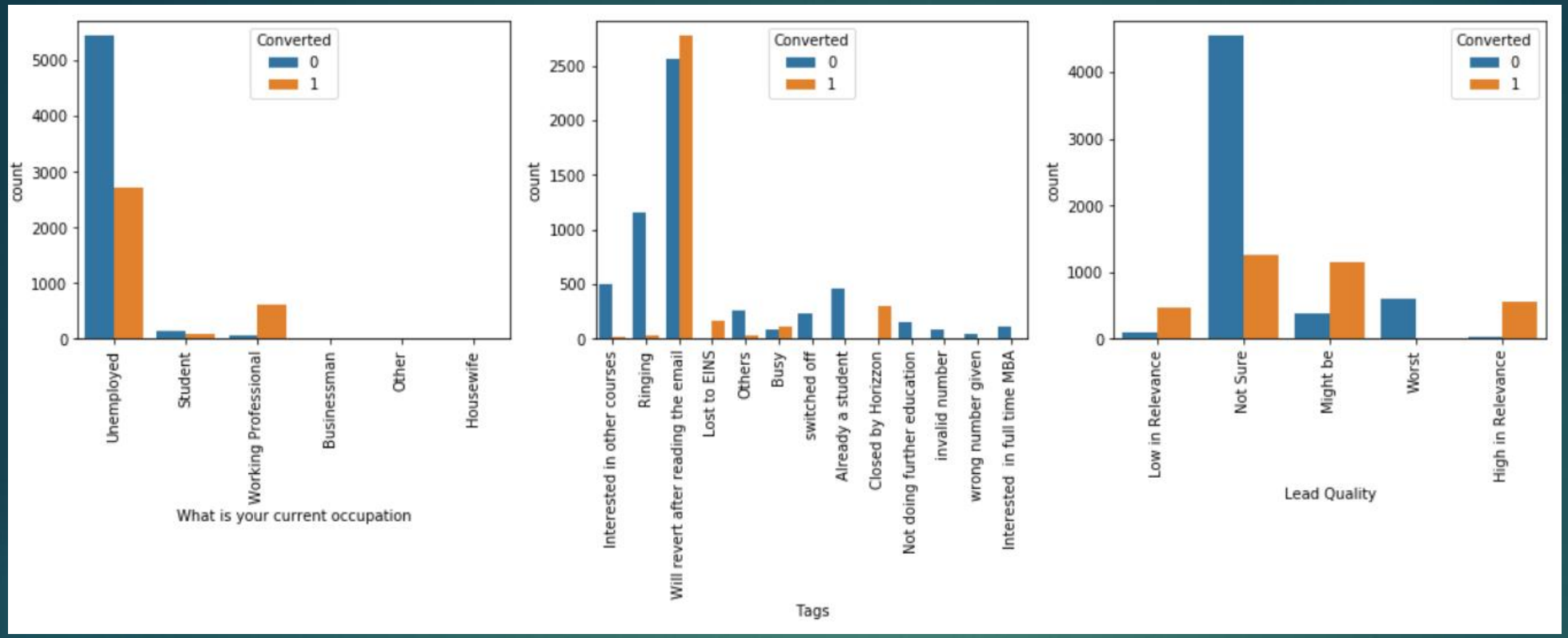
SEGMENTED UNIVARIATE ANALYSIS OF CATEGORICAL COLUMNS:

- Most origins of the lead which are converted come from landing page submission
- We can see that some of the other origins of the lead which are converted come from API,Lead add form
- Most Sources of the lead which are converted are from Google
- We can see that some of the sources of the lead which are converted come from Direct traffic,Reference,Olark Chat
- Most of the converted opted for no in do not email column



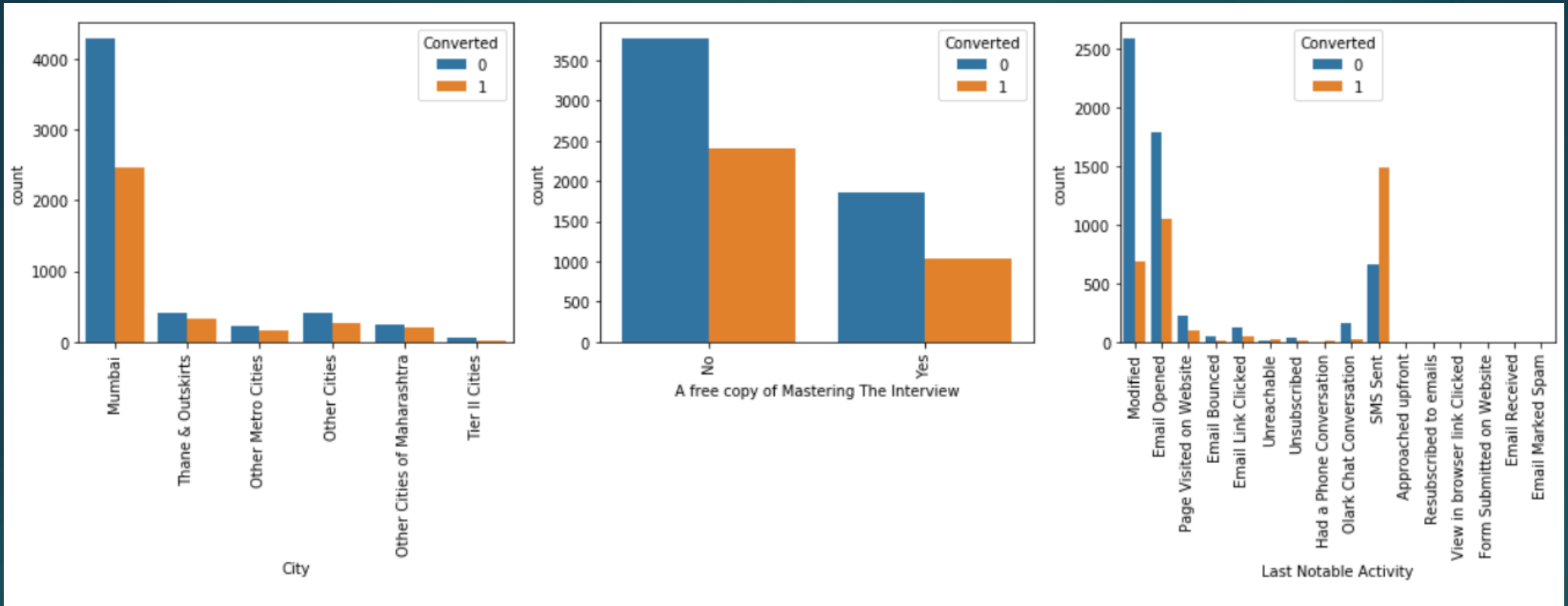
SEGMENTED UNIVARIATE ANALYSIS OF CATEGORICAL COLUMNS:

- Most of the customers who are converted has last activity as SMS sent and most of the customers who are not converted opened their emails as their last activity
- Most of the customers who are converted are from India
- Most of the customers who are converted belong to others, finance management, human resource management specialization categories



SEGMENTED UNIVARIATE ANALYSIS OF CATEGORICAL COLUMNS:

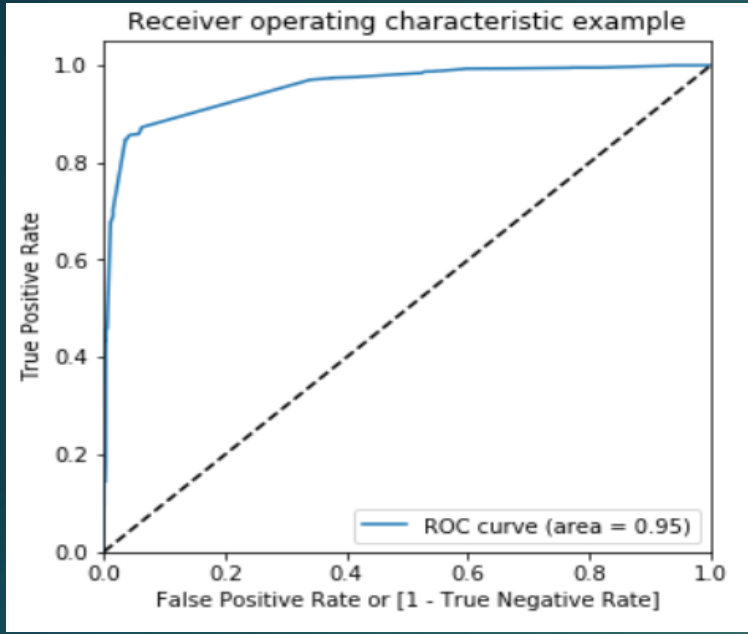
- Most of the customers who are converted are unemployed
- Most of the customers who are converted got 'Will revert after reading the email' tagged.
- Most of the customers who are converted has lead quality has might be or not sure



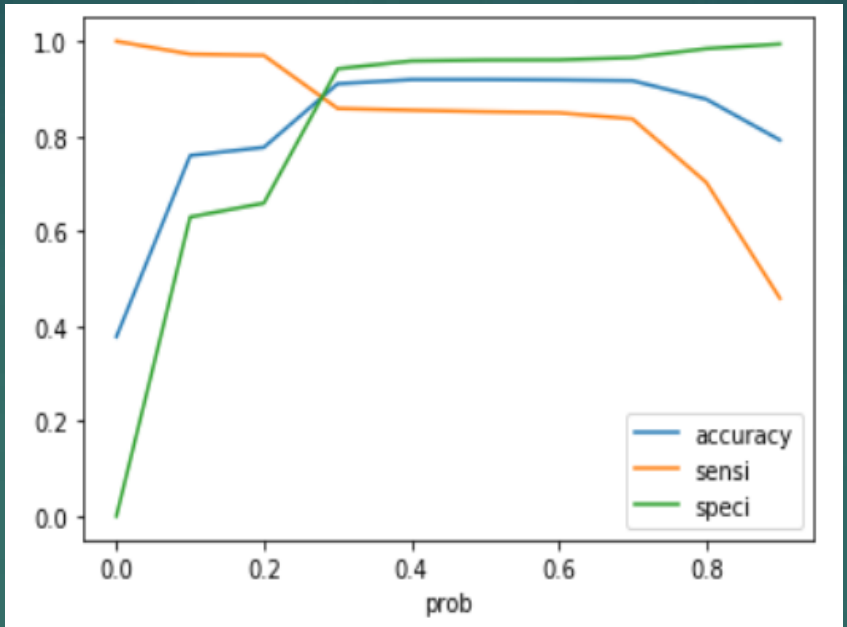
SEGMENTED UNIVARIATE ANALYSIS OF CATEGORICAL COLUMNS:

- Most of the converted customers last activity is SMS sent
- Some of the other converted customers last activities belong to Modified, email opened categories
- Most of the customers who are converted don't want a free copy of mastering the interview
- Most of the customers who got converted live in Mumbai

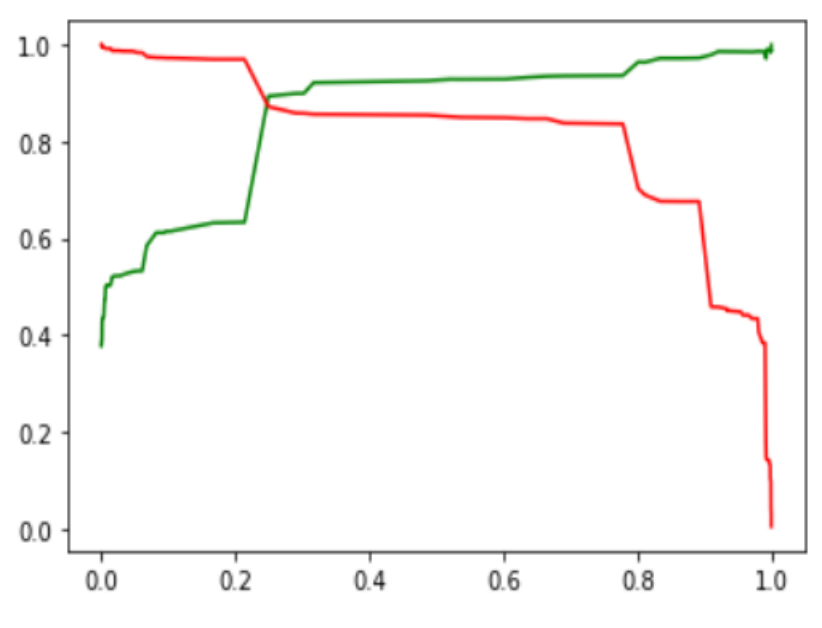
ROC CURVE



SENSITIVITY-SPECIFICTY-CURVE



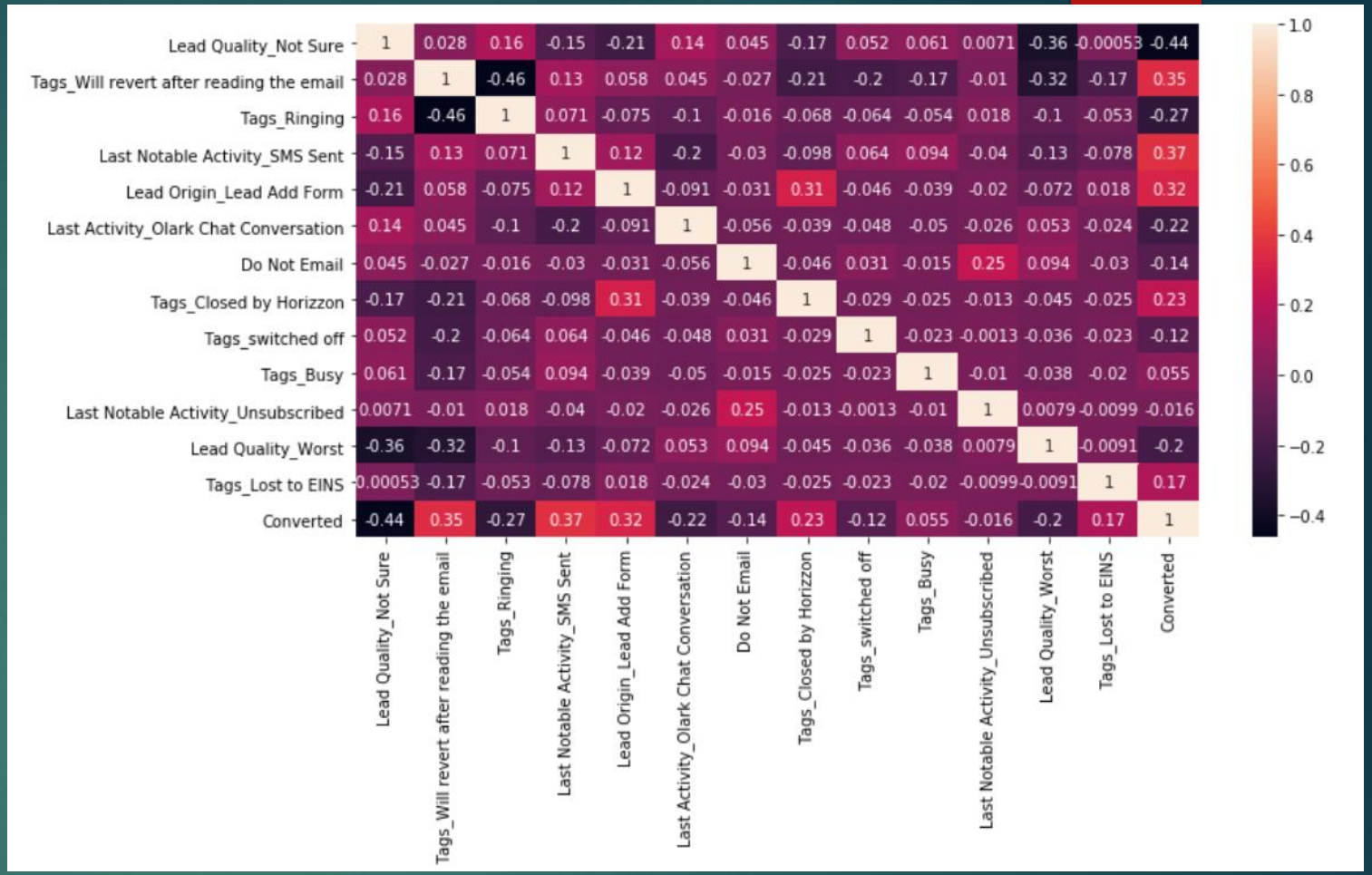
PRECISION-RECALL CURVE



- Area under ROC curve is 0.95 which infers that our model is pretty good model
- Optimal threshold is taken as 0.3 as from 0.3 as we go up accuracy does not change much, also sensitivity and specificity are pretty high at 0.3 so for balanced metrics it is better to go with 0.3
- Optimal threshold is chosen from sensitivity-specificity curve as sensitivity is important measure based on our business objective

CORRELATION MATRIX OF FINAL MODEL COLUMNS

- Top three variables in model, that contribute to lead conversion are :
  - Last Notable Activity SMS Sent
  - Tags_Will revert after reading the email
  - Lead Origin Lead Add Form
- Top three variables in model that should be focused in order to increase the probability of lead conversion:
  - Lead Quality_Not Sure (negatively impacting to conversion)
  - Last Notable Activity_SMS Sent (positively impacting to conversion)
  - Tags_Will revert after reading the email (positively impacting to conversion)



Model Adjustability for Different Scenarios

MODEL ADJUSTABILITY FOR SCENARIO-1

Scenario-1: The company has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.

RECOMMENDATIONS FOR DEALING WITH THIS SCENARIO:

- They should concentrate on customers with a little bit of lower lead conversion probability than threshold for more broader audience.
- The above step is done technically by reducing the threshold by a little bit to get more customers, so that they can call a greater number of customers.
- They should concentrate on customers whose last notable activity is SMS sent, customers who are tagged as 'Will revert after reading the email' first, since there is higher correlation between conversion and those two variables.

By doing all these above things, we can increase chances of the conversion of customers whose lead conversion probability is low as well

MODEL ADJUSTABILITY FOR SCENARIO-2

Scenario-2: Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage.

RECOMMENDATIONS FOR DEALING WITH THIS SCENARIO:

The strategy they should employ at this stage is:

- Since the company does not want to make phone calls unless it is extremely necessary, it is better if the sales team focus on more important leads.
- These more important leads can be obtained by increasing the threshold cut off, so that only leads with higher conversion probability will be considered and leads with lesser conversion probability will be discarded.
- By doing this , there will be minimal phone calls and we will be getting good conversions

MODEL SUMMARY AND OTHER RECOMMENDATIONS

- Test accuracy is 90.8 %, training accuracy is 91.0 %
- Sensitivity for test set is 85.4 %, Sensitivity for training set is 85.8 %
- Specificity for test set is 93.9 %, Specificity for training set is 94.2 %
- Company should focus on sending more SMS, since this helps in higher conversion.
- Company should focus on lead add form since customer identification by that produces more conversion
- Also company should focus on the customers whose current status is 'Will revert after reading the email' ,so customers who are tagged by this type must be monitored since there is a high potential for these type of customers for lead conversion
- Company should improve its techniques for analyzing quality of lead as it is negatively impacting the conversion
- Company should improve Olark chat service since it is negatively impacting the conversion