LEAD SCORE CASE STUDY

P K VIVEK VIDHYA

CONTENTS

PROBLEM STATEMENT

- A company named X Education sells online courses to all the customers and they get lot of leads but their rate of lead conversion is very less.
- For e.g. if say they acquire 100 leads a day only 30 are converted.

PROBLEM STATEMENT

- In order to make this process more efficient or faster the company wishes to identify potential leads known as **hot leads**. They also want to find the **cold leads**.
- If they successfully identify those people who can be converted into successful leads. The rate of lead conversion can go high and the sales team can focus/communicate with them as they are important leads rather than making calls to everyone.

PROBLEM STATEMENT

THIS IS THE TYPICAL LEAD CONVERSION FUNNEL WHICH THE X EDUCATION COMPANY USES.

BUSINESS OBJECTIVE

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. They want to build a model to identify the Hot leads and to deploy this model for the future use.

SOLUTION METHODOLOGY

- Understanding the business problem and the objective
- Importing and knowing the insights of the data
- Data cleaning and manipulation
  1. Treating the missing values by dropping the columns which have high rate of missing values and dropping those columns/features which aren't important for the analysis
  2. Check the outliers

SOLUTION METHODOLOGY

- EDA analysis (exploratory data analysis)
  1. Finding the different classes of certain features by using value_counts()
  2. Visualization among the numerical variables (by using pair plot)
  3. Visualization among the categorical variables with respect to dependent variable (by using box plot)
  4. Find the frequency of distribution among the variables (by using histogram)
  5. Univariate and bivariate analysis

SOLUTION METHODOLOGY

- Create a dummy variable for those categorical feature which is having more than two classes.
- Use min max scaler to scale down the values of the feature which is not in zero and one.

SOLUTION METHODOLOGY

- Use classification technique as our dependent/label as having categorical value in the form of zero and one. So, use logistic regression model.
- Use RFE for auto feature selection of top fifteen
- Validation of the model and model presentation
- Give a conclusion and recommendation

DATA MANIPULATION

- Initially our data shape was 9240 rows and 37 columns.
- We dropped city and country as it was not required for the analysis
- Also found few features having a value select(that means the customer has to chose). So, dropped the columns 'How did you hear about X Education' and 'Lead Profile'.

DATA MANIPULATION

- Do Not Call, Search, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, 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. These columns have values maximum NO .Its not useful for analysis.

DATA MANIPULATION

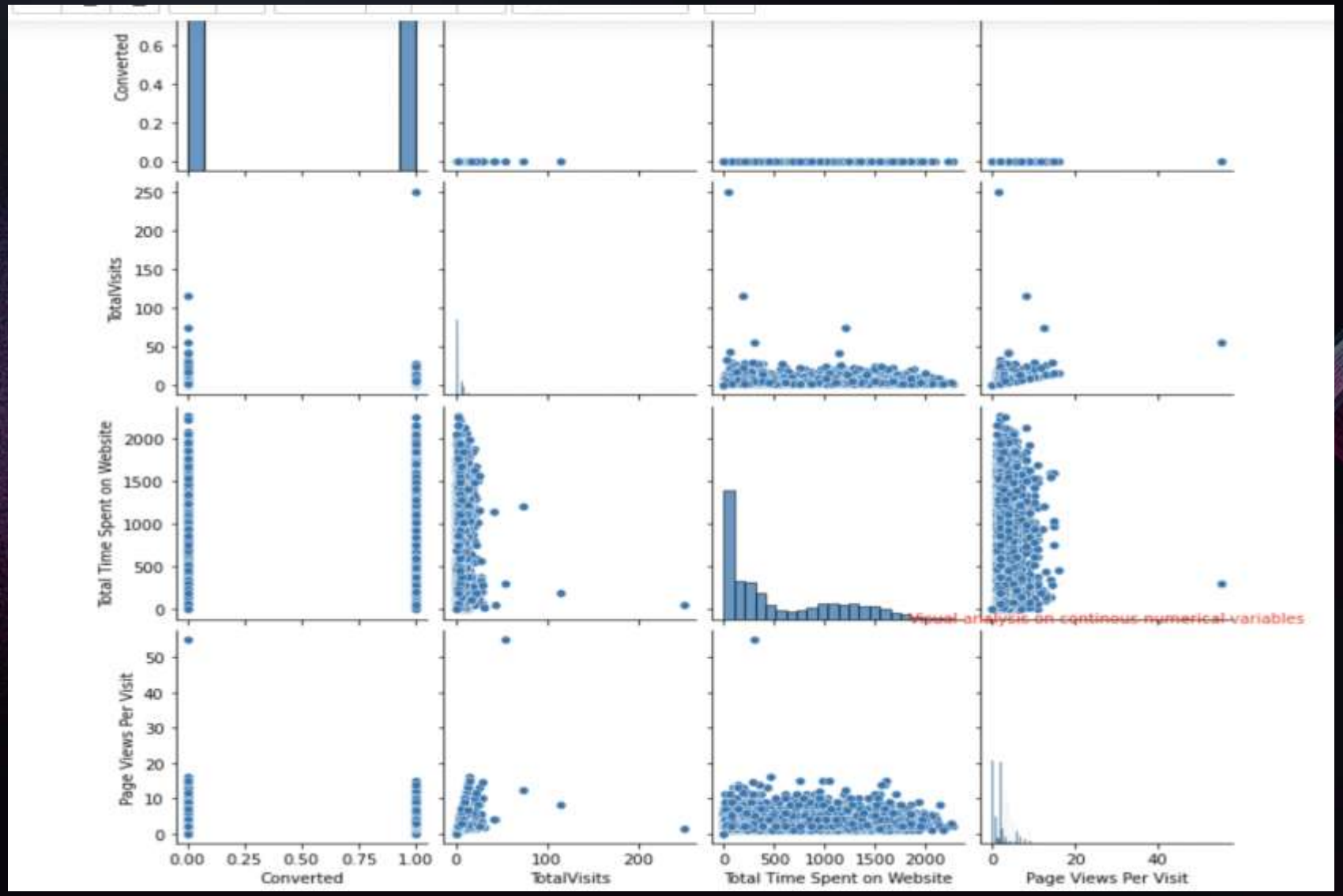
- Also dropped few rows of some columns which have null values.
- Finally left with 12 columns and 6373 rows.

DATA MANIPULATION

- Also created dummy variable for few columns which is having more than two class and imputed into one dataframe and finally concatenated with our original dataframe.
- Also dropped the original columns whose dummy variable is created. So our final dataframe is left with 6373 rows and 75 columns.

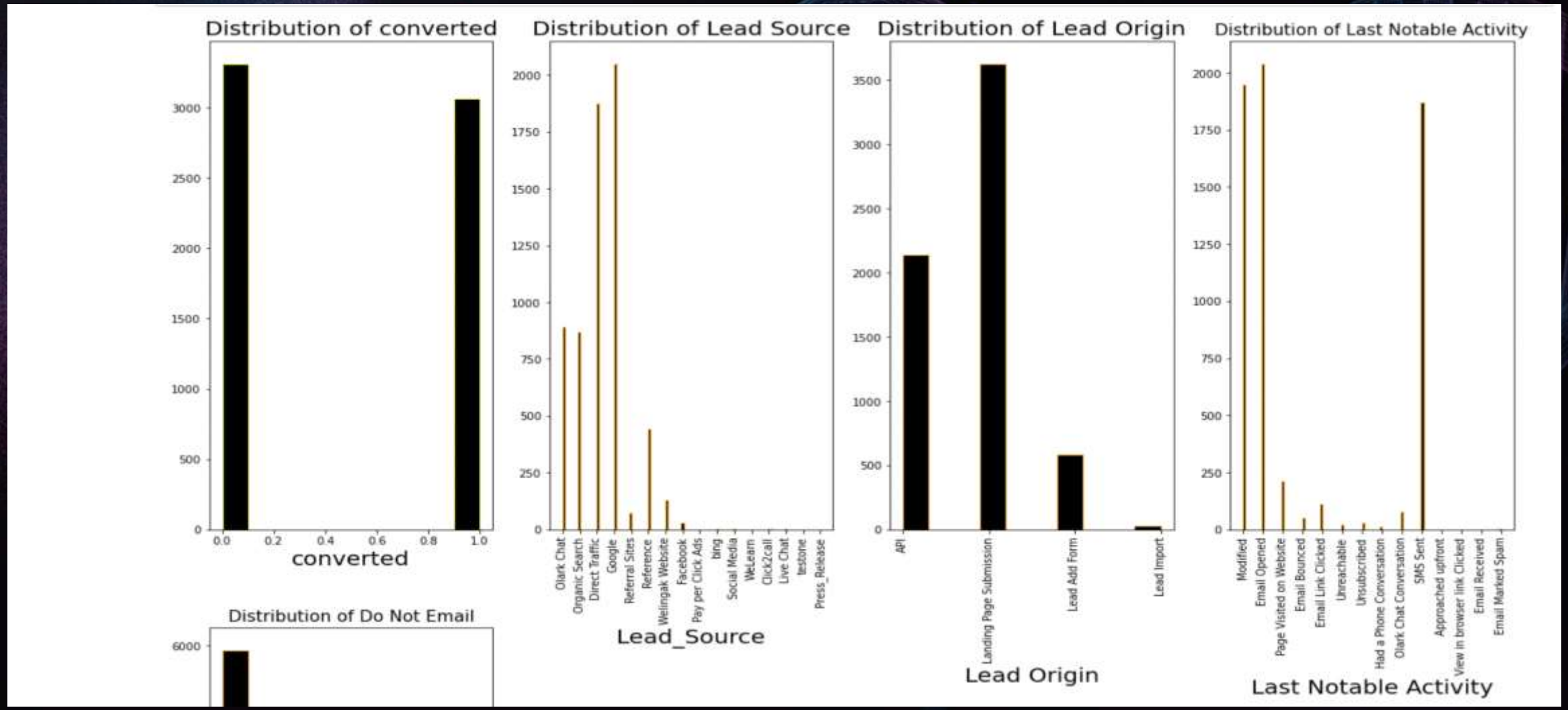
EDA AND CATEGORICAL/NUMERICAL VARIABLE RELATIONSHIP

Analysis and visualizing numerical variable by using pair plot.

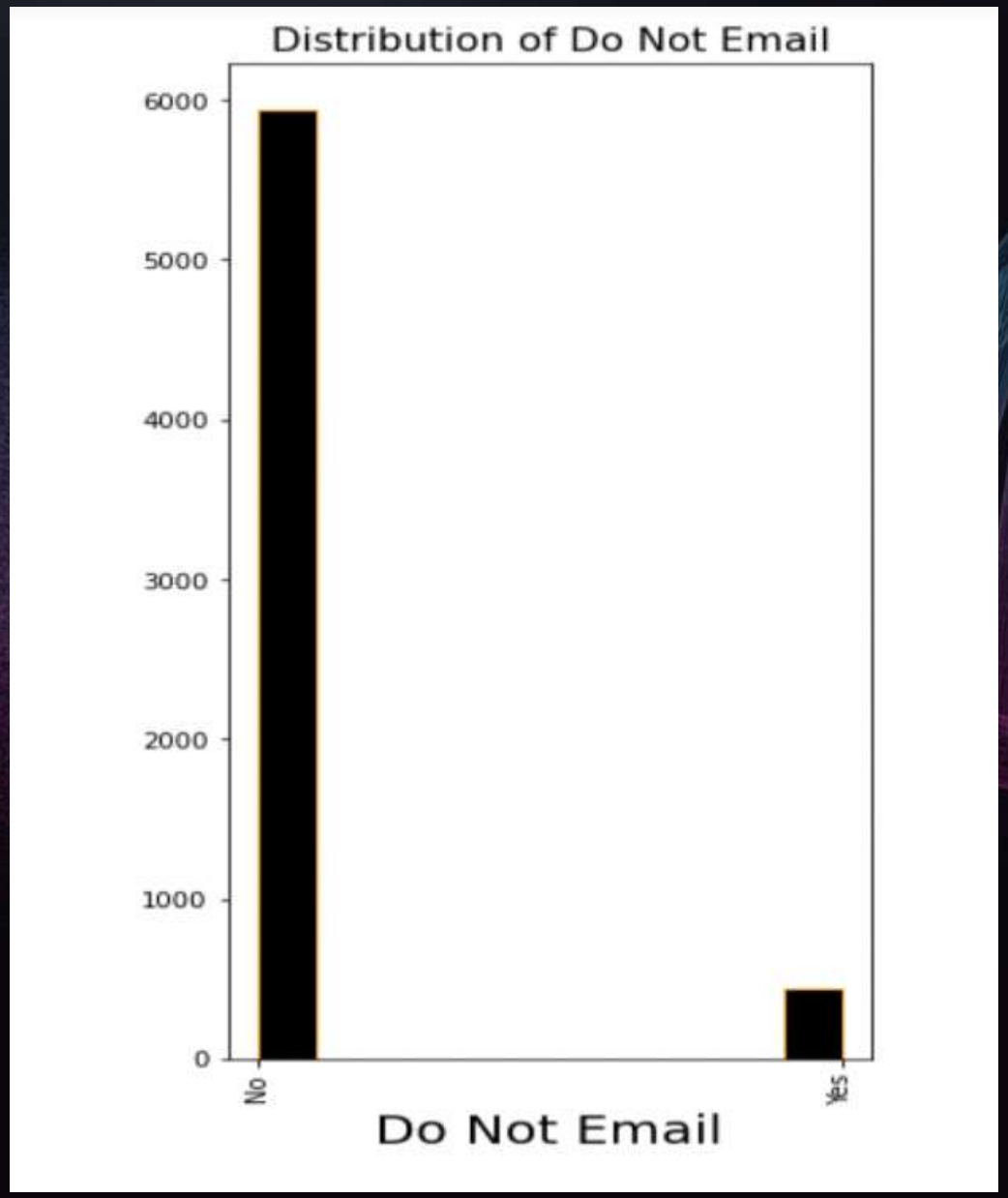


OBSERVATION: FOUND A KIND OF LINEAR RELATIONSHIP BETWEEN TOTAL VISITS AND PAGE VIEWS PER VISIT.

Frequency distribution of all categorical values by using histogram in sub plot.

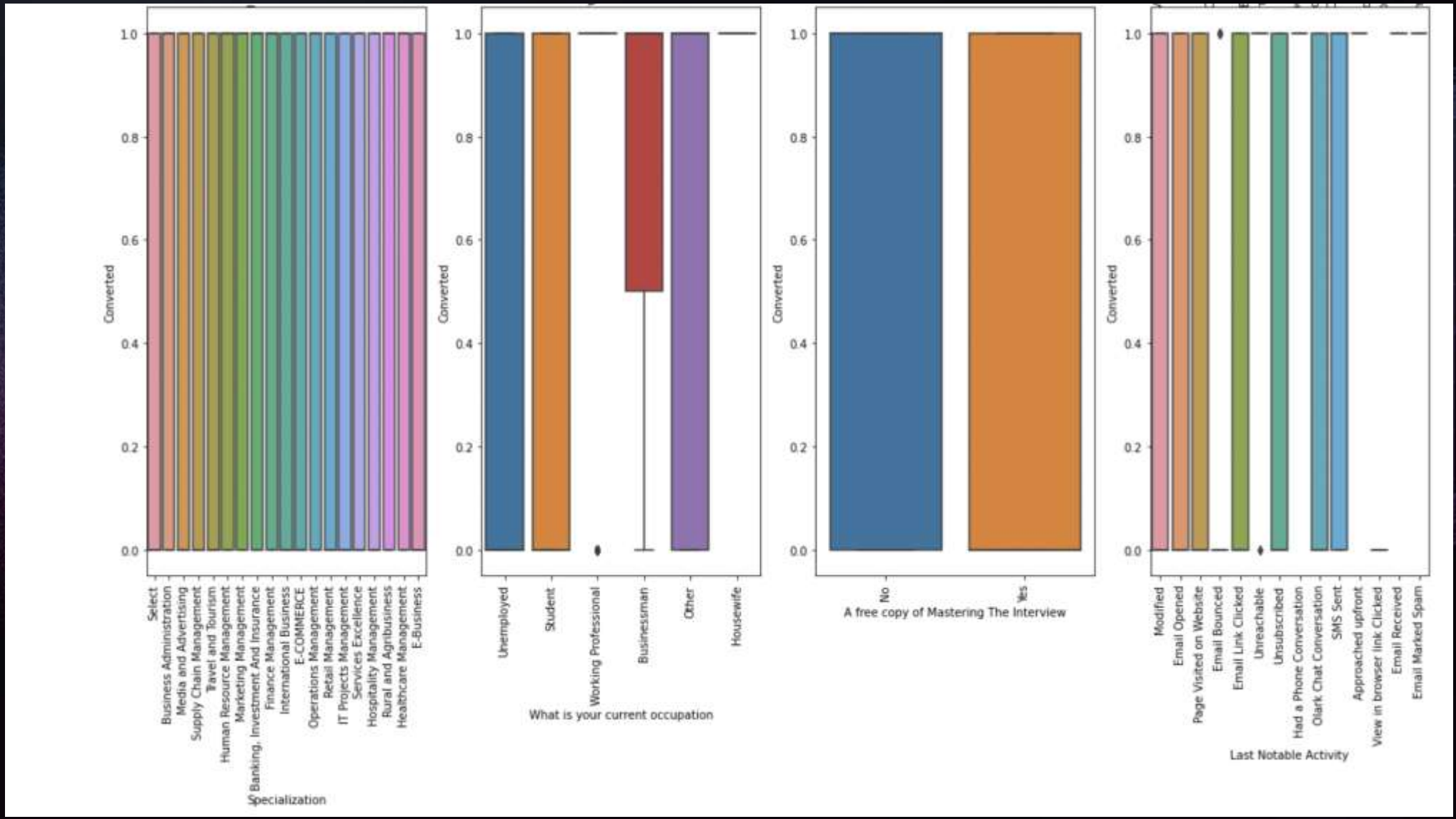


Frequency distribution of all categorical values by using histogram in sub plot.

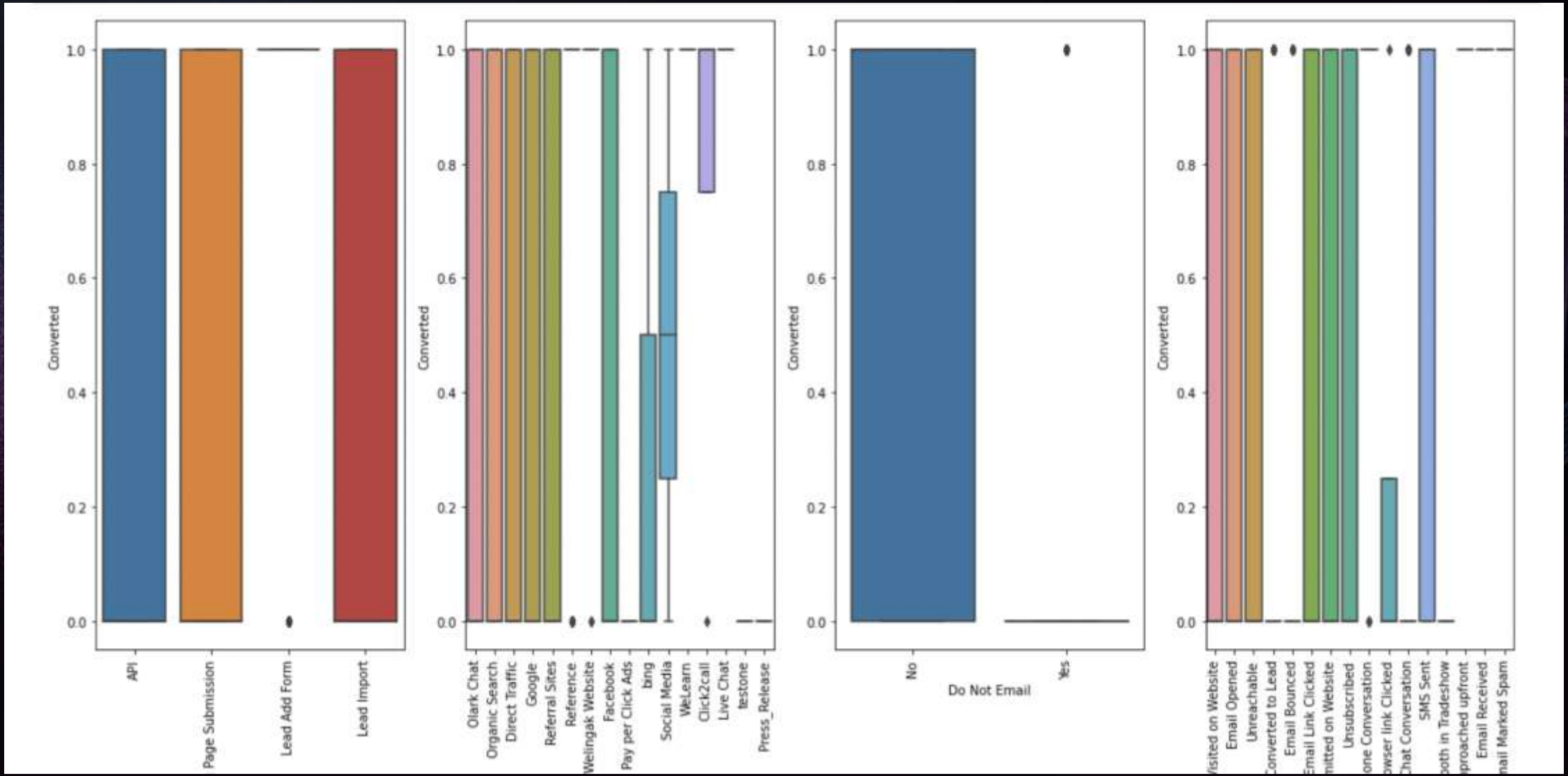


- **OBSERVATION:-** FOUND DISTRIBUTION OF COVERTED RATE IS LESS THAN NO LEAD ,AND IN LEAD SOURCE GOOGLE HAVE MORE DISTRIBUTION ,LEAD ORGIN = LANDING PAGE HAVE MORE DISTRIBUTION ,LN LAST ACTIVITY EMAIL OPENED AND SMS SENT HAVE MORE DISTRIBUTION AND DO NOT EMAIL SENT NO HAVE MORE DISTRIBUTION IN OUR DATASET

Visualization of categorical variable vs dependent variable using box plot in sub plot



Visualization of categorical variable vs dependent variable using box plot in sub plot



Observation on Visualization of categorical variable vs dependent variable using box plot in sub plot

- Free copy of mastering the interview both values yes and no are converted
- Customer who choose to respond by emailing do not email (No) are converted
- Also found customer of different specialization are converted

Observation on Visualization of categorical variable vs dependent variable using box plot in sub plot

- maximum from google search ,organic ,direct traffic etc. are converted
- In the last labeled activity customers who opened email, visited the website, chat with olark and link clicked were converted
- Customers who opened email, SMS, visited the website and clicked the link in their last activity were converted.

MODEL BUILDING

- Split the data into 'x' and 'Y' (independent and dependent variable). Again split 'x' and 'Y' into train and test dataset.
- I have chose 70-30 ratio for splitting, 70 percentage is train part and 30 percentage is test part.
- Used min max scaler to scale down the values for those variable whose values are not in the form of zero and one.

MODEL BUILDING

- USED RFE FEATURE WITH AN OUTPUT OF TOP 15 SELECTED FEATURE
- MADE PREDICTION ON TRAIN DATA FIRST AND ULTIMATELY PREDICTED ON TEST DATA.
- I REPEATEDLY MADE MODELS TO IMPROVE THE ACCURACY AND THE FIFTH MODEL WAS THE FINAL MODEL WITH DESIRED ACCURACY AS THERE IS A MULTICOLLINEARITY BETWEEN THE DEPENDENT VARIABLE AND ALSO WITH A HIGH 'P' VALUE.

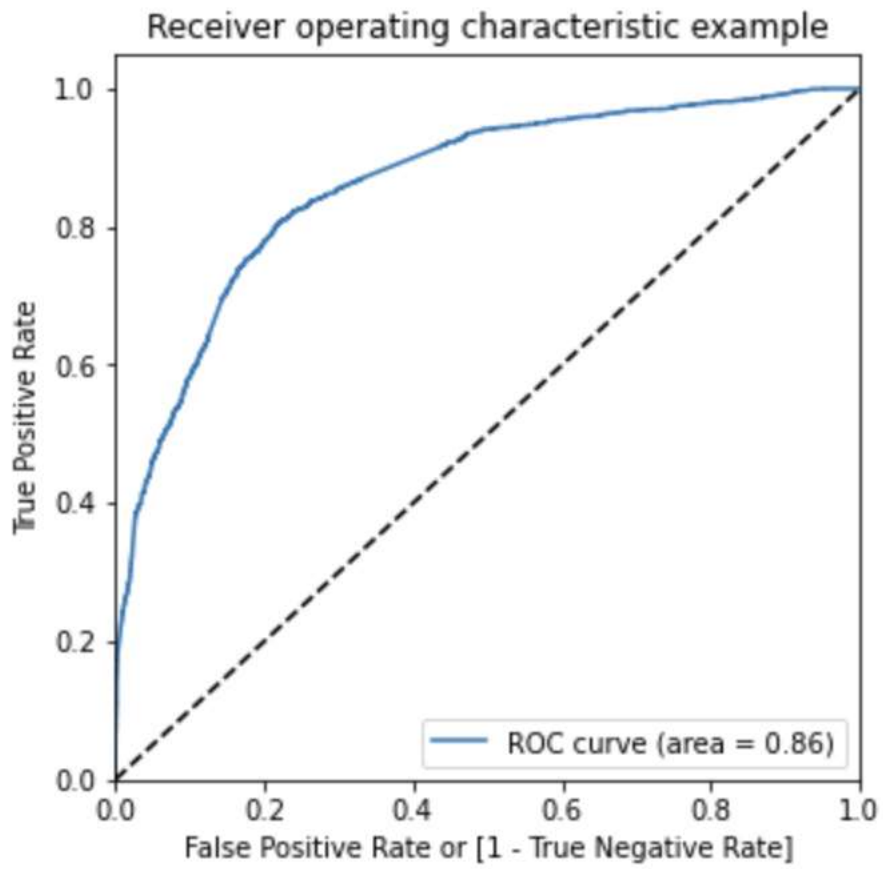
MODEL BUILDING

- I HAVE USED A CLASSIC LIST COMPREHENSION METHOD TO CREATE A DATAFRAME WITH VARIABLES NAMED ACTUAL, PREDICTED AND CONVERTED PREDICTED.
- I HAVE FOUND CUT-OFF BEING 0.4 AND DECIDED TO CUT THE PREDICTION VALUE.
- MY FINAL MODEL ACCURACY SCORE ON TEST DATA IS 0.78

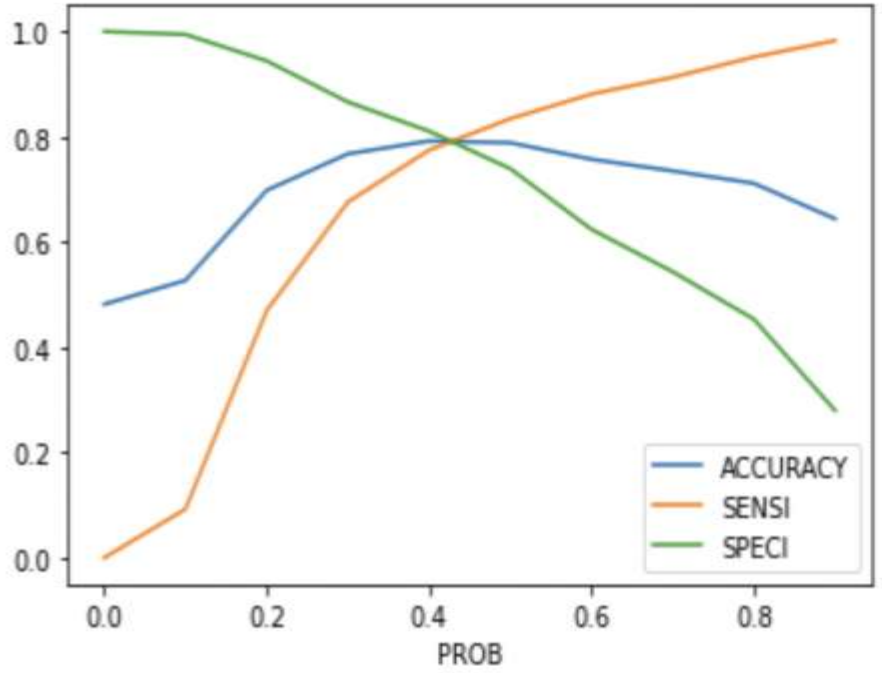
MODEL BUILDING

- SENSITIVITY(TRUE POSITIVE RATE) = 0.78
- SPECIFICITY(TRUE NEGATIVE RATE)=0.79
- RECALL=0.78
- PRECISION=0.77
- F1 SCORE IS 0.78

ROC CURVE AND EXACT CUT-OFF




```
# Ploting for excat cutoff
new_cut.plot.line(x='PROB', y=['ACCURACY','SENSI','SPECI'])
plt.show()
```



ROC CURVE AND EXACT CUT-OFF

Observation

- Found region under curve is 0.86 .It looks like good model .But need to check the in cut off range
- As we found around 0.42 is optimal value where all the three metrics intersect between each other (accuracy, sensitivity and specificity)

CONCLUSION

- Found that the variables which are having negative and positive impact on lead conversion
- **POSITIVE VARIABLES** Used for lead conversion

1. The total time spend on the Website.
2. Total number of visits.
3. When the lead source was:
   a. Welingak website
   b. Olark Chat
4. When the last activity was:
   a. SMS
   b. PHONE CONVERSATION
5. When the lead origin is Lead Add Form.

CONCLUSION

- **NEGATIVE VARIABLES** WHICH HAVE A NEGATIVE IMPACT ON LEAD CONVERSATION
  A. IF THE CURRENT OCCUPATION IS STUDENT AND UNEMPLOYED

Above aspect we should Keep in mind, The X Education can improve by seeing there negative and positive variables .By using Above features in mind X education can achieve their goal.

A wise man once said…

The end is merely the start.

The End