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.
- The company markets its courses on several websites and search engines like Google. 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. The typical lead conversion rate at X education is around 30%.
- Now, although X Education gets a lot of leads, its lead conversion rate is very poor. 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

Goal

- Build 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.
- A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

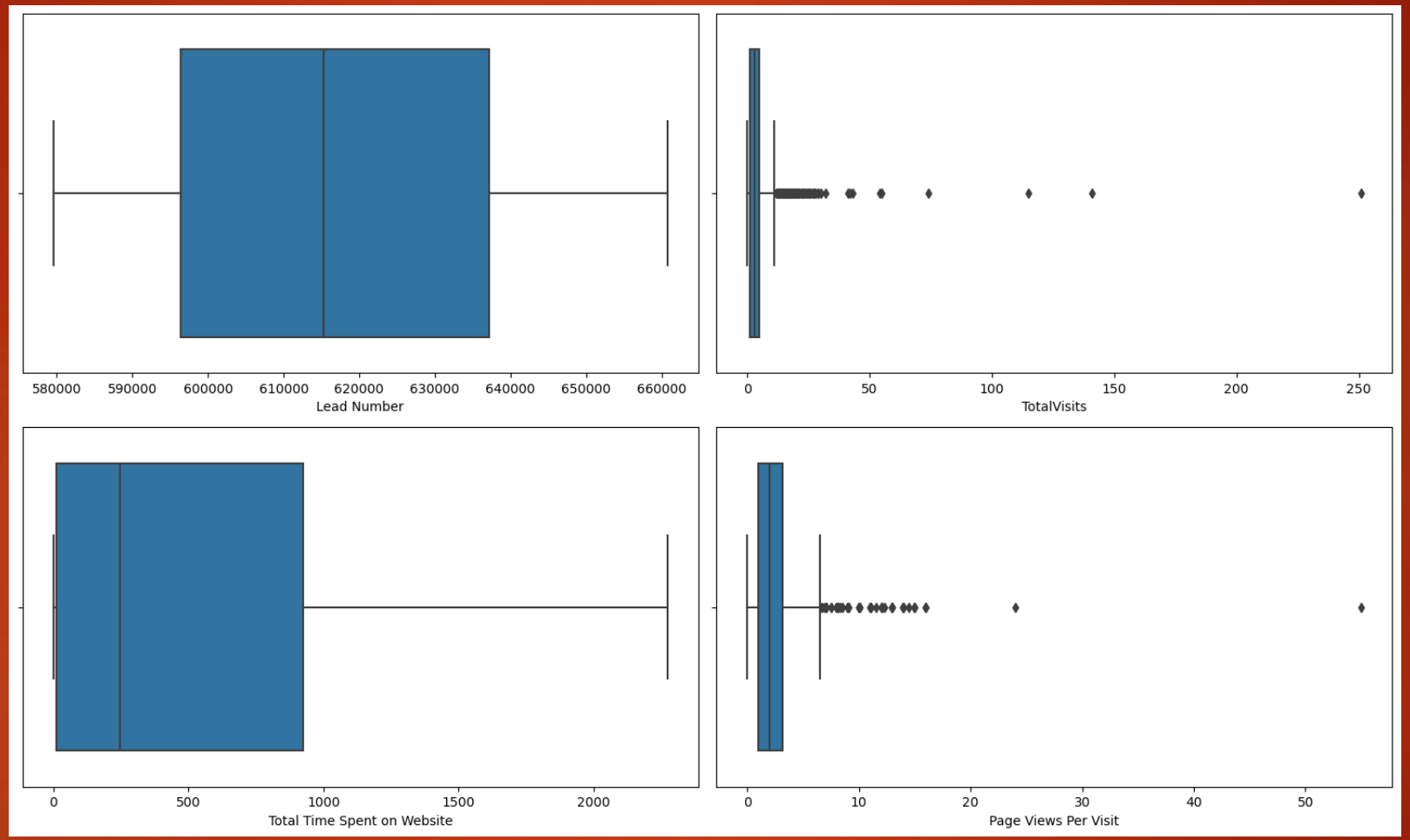
Approach

- - 1- Data Understanding
- - Importing Data and Check Statistics
- - 2- Data Cleaning
- - Check missing values/checking outliers and fix those by checking their statistics
- - 3- Exploratory Analysis
- - Uni-Variate, Bi-Variate and Correlation or pair plots
- - 4- Data Preparation
- - Convert in binary column and dummy variables creation
- - Feature Scaling
- - 5- Build Model
- - Split the data in train and test, features scaling, check correlation matrix,
- - - Features Selection using RFE and manual
- - 7- Model Evaluation
- - Confustion Matrix
- - Accuracy , specificity, Precision and recall, ROC Curve
- 8- Prediction of test Data

Data Understanding and Cleaning

- Read CSV file
- Check the data shape, information and data types
- Check missing data and Inaccurate data type columns
- Impute with some values in some columns
- Remove the columns which are having high missing data
- Check for the outliers in data and cap them
- Columns remove- "Do Not Email", "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", "A free copy of Mastering The Interview“ Most the values are in one variable only so cant get any inference from it.

Outlier Treatment



- As per boxplot, TotalVisits and PageViewPerVisit having outliers
- After checking the percentile of both, find the interquartile range first and then cap on minimum and maximum values.

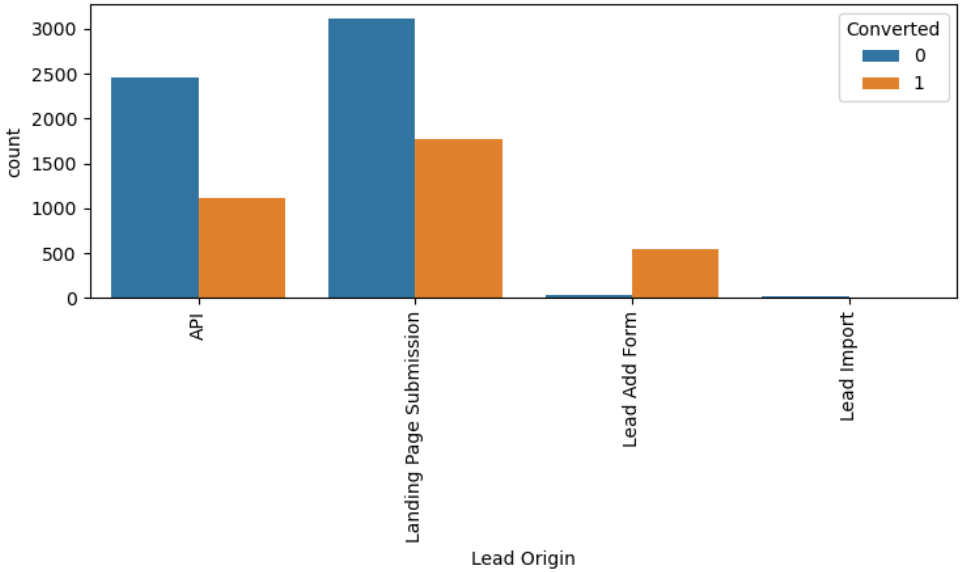
Exploratory Analysis

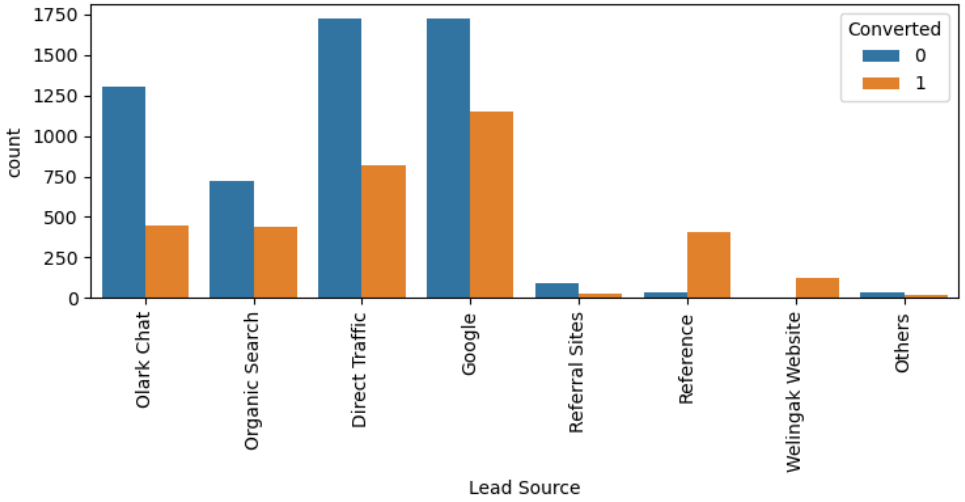
- Group low frequency data in other columns
- First Uni-variate analysis for categorical and continuous columns perform
- Remove those column which are not giving any inferences
- Run bi-variate analysis with target variable “converted”
- Multi variate Analysis

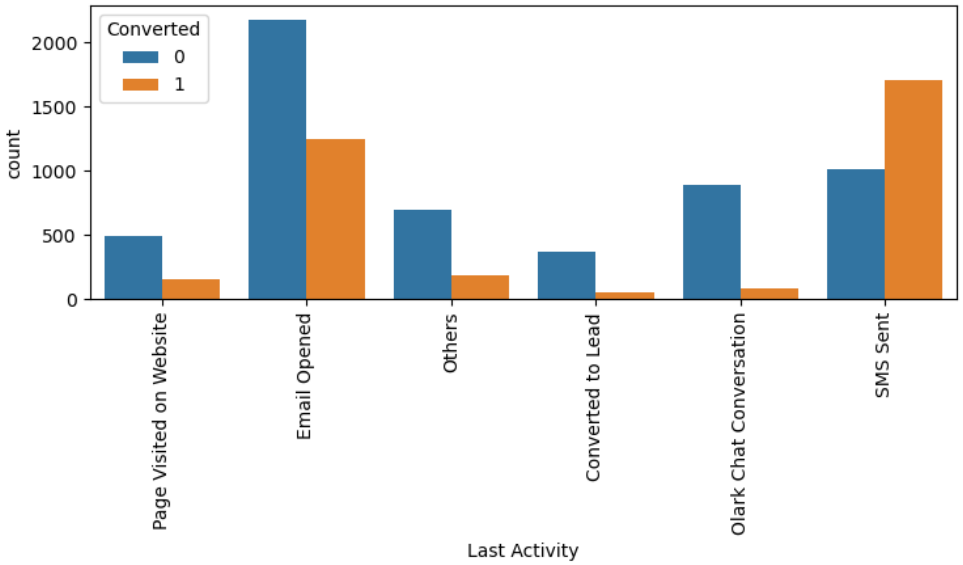
Inferences from uni and bi-variate analysis

- More converted Leads are coming from Landing Page submission, Lead Origin
- Most converted Lead Source are from Google
- Most Converted Leads are coming from Last Activity from SMS sent
- Converted Leads are not choosing's specialization, either they don't have in list or they don't want to opt
- Most Converted leads are searching online for this X Education
- Most Converted Unemployed are opting for education
- Tags will revert after reading the email having most converted lead
- Potential Leads are from Lead Profile
- Mumbai City having most converted leads
- Converted Leads mostly are done last activity as SMS Sent
- 50 percentile for Lead Number for Converted and non- Converted are almost same, so cant get any inferences, we can remove this column
- 50 percentile for TotatVisits for Converted and non- Converted are almost same, so cant get any inferences, although converted leads having more counts
- 50 percentile for Page View per Visit for Converted and non- Converted are almost same, so cant get any inferences, although converted leads having more counts
- Total Time spent on Website having more converted leads

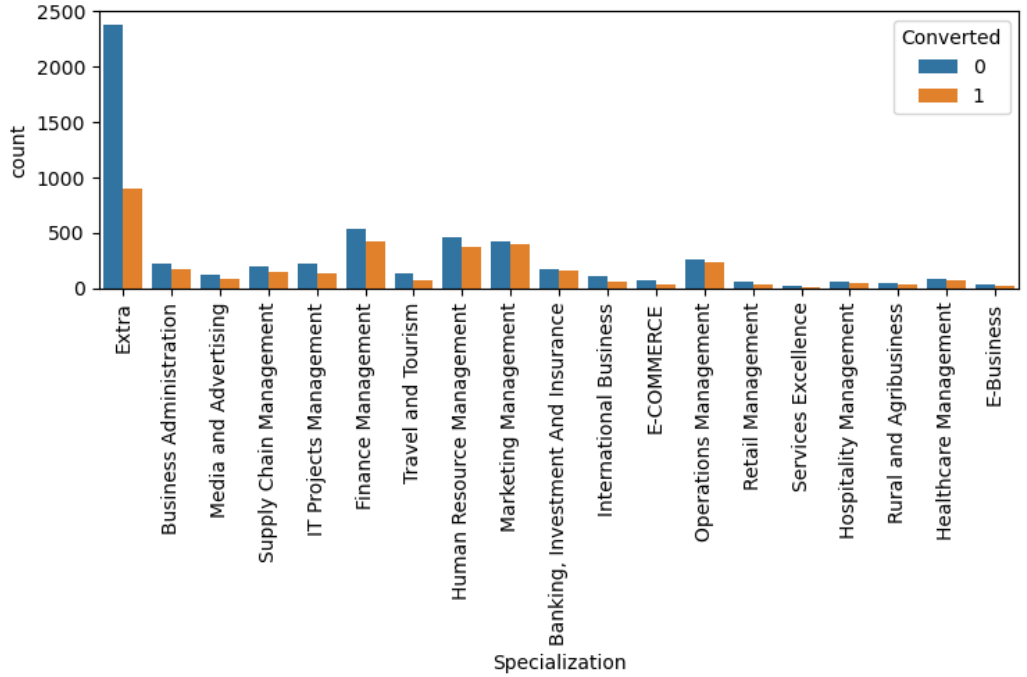
Visualization

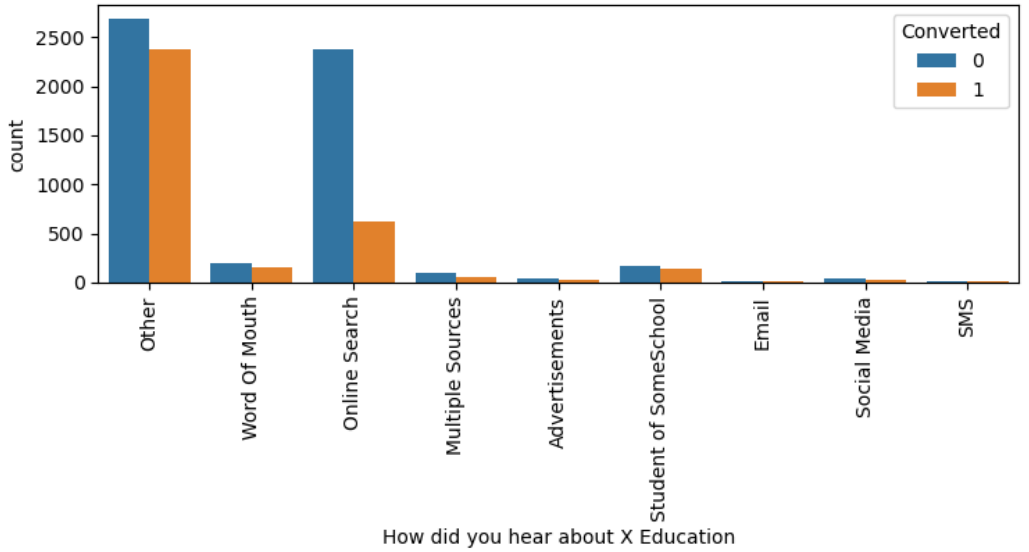


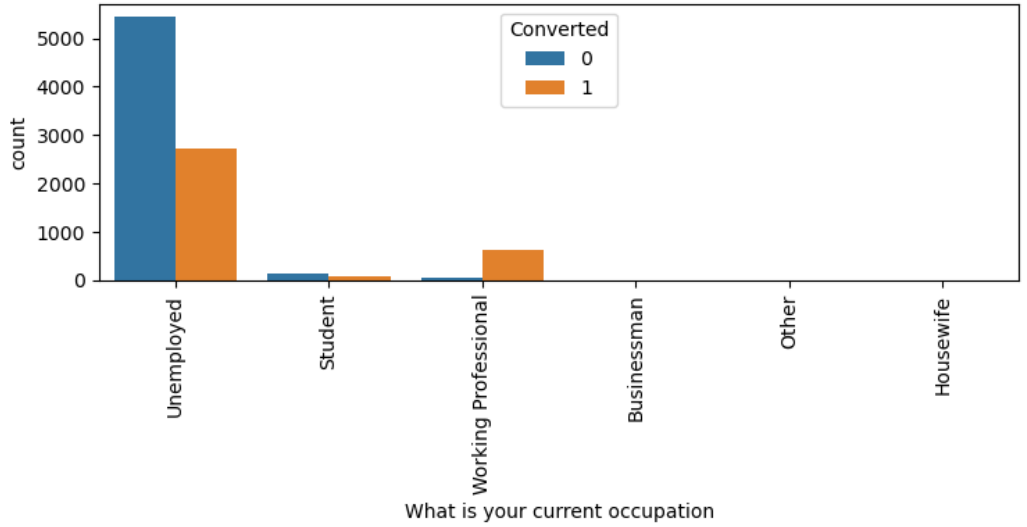




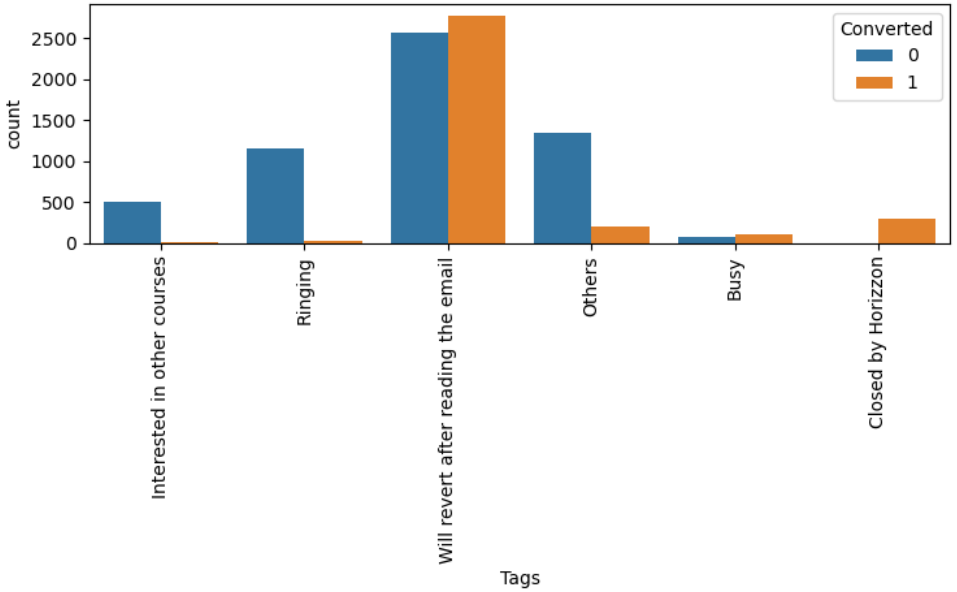
Visualization

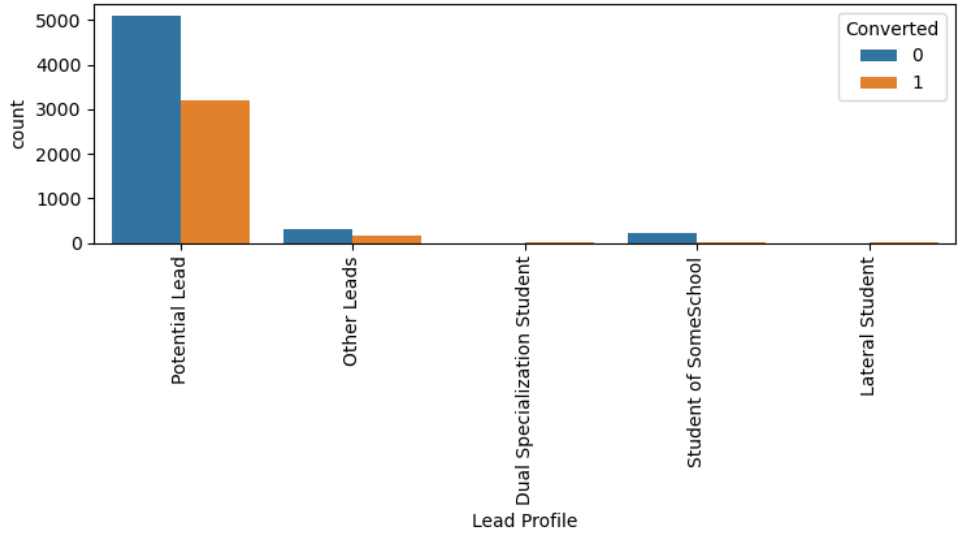


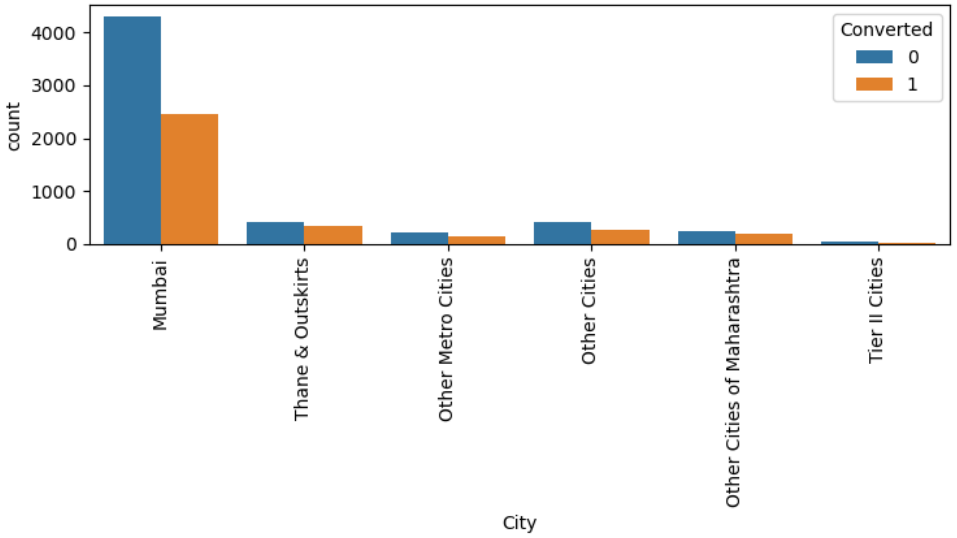


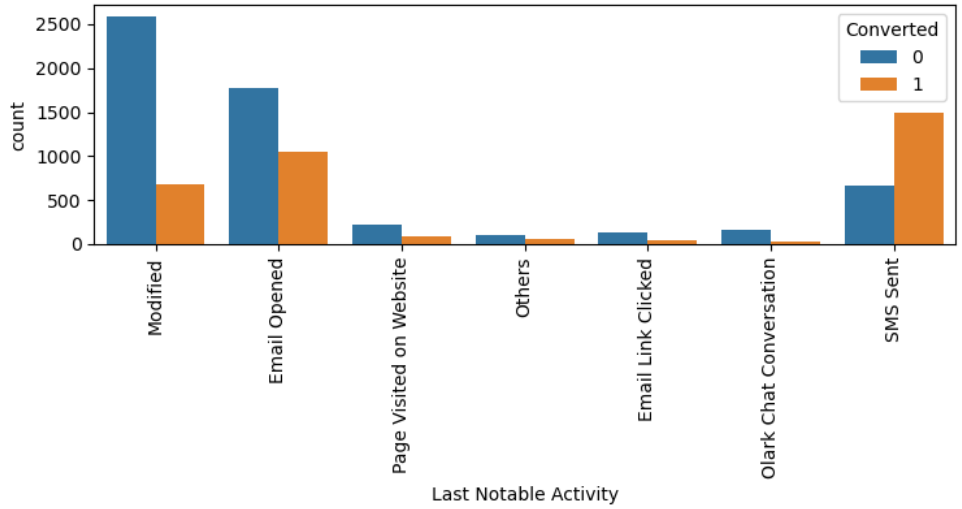


Visualization

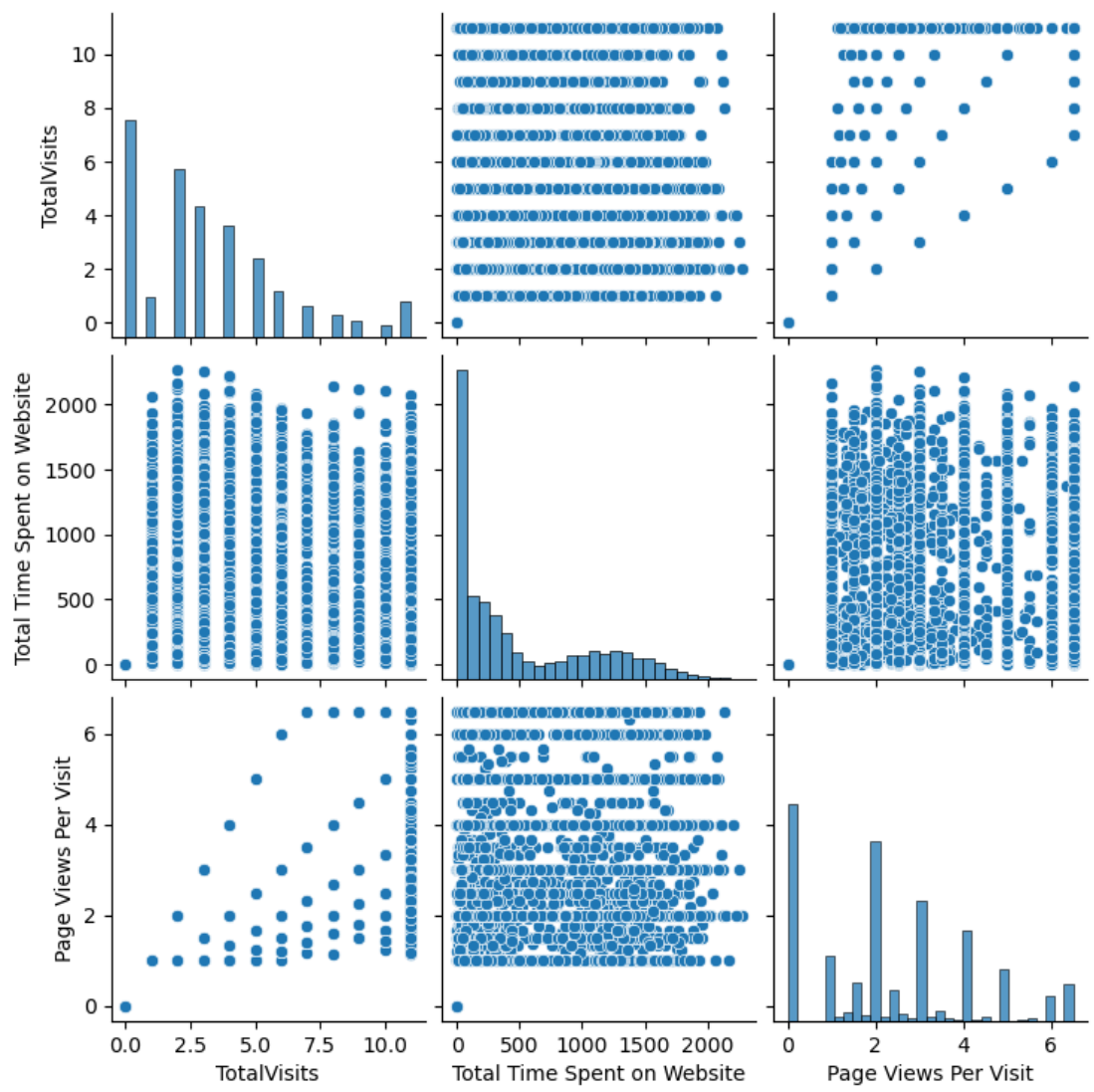








Visualization



Data Preparation

- - Converted column into binary values (0/1)
- Columns are - "Do Not Email","Do Not Call","Get updates on DM Content","I agree to pay the amount through cheque","A free copy of Mastering The Interview","Search","Magazine","Newspaper Article","X Education Forums", "Newspaper","Update me on Supply Chain Content","Digital Advertisement","Through Recommendations", "Receive More Updates About Our Courses","Get updates on DM Content“
- Creation of dummy variables in categorical columns
- Feature Scaling

Model Building

- Feature Selection using RFE
- Run the logistics regression model iterative and check VIF
- Stop where all variable are below 0.05 and high VIF
- Variable those are participating for converting leads are-
- 'Origin__Lead Add Form', 'Source__Welingak Website', 'Activity__Email Bounced', 'Activity__Had a Phone Conversation', 'Activity__Olark Chat Conversation', 'Education__Online Search', 'Education__SMS', 'occupation__Working Professional', 'Tags__Closed by Horizzon', 'Tags__Interested in other courses', 'Tags__Others', 'Tags__Ringing', 'Tags__Will revert after reading the email', 'Profile__Student of SomeSchool', 'notableActivity__SMS Sent'

Run Model

- - Get the predicted value on train set
- Created a data frame with actual converted and converted probability
- Probability with 50 % above are consider as converted predicted lead
- Calculate confusion matrix

Model Evaluation

- On train Data

  Accuracy- 80%

  Sensitivity- 84%

  Specificity- 91%

  Precision- 85%

  Recall-84%

Optimal cut-off- 0.3

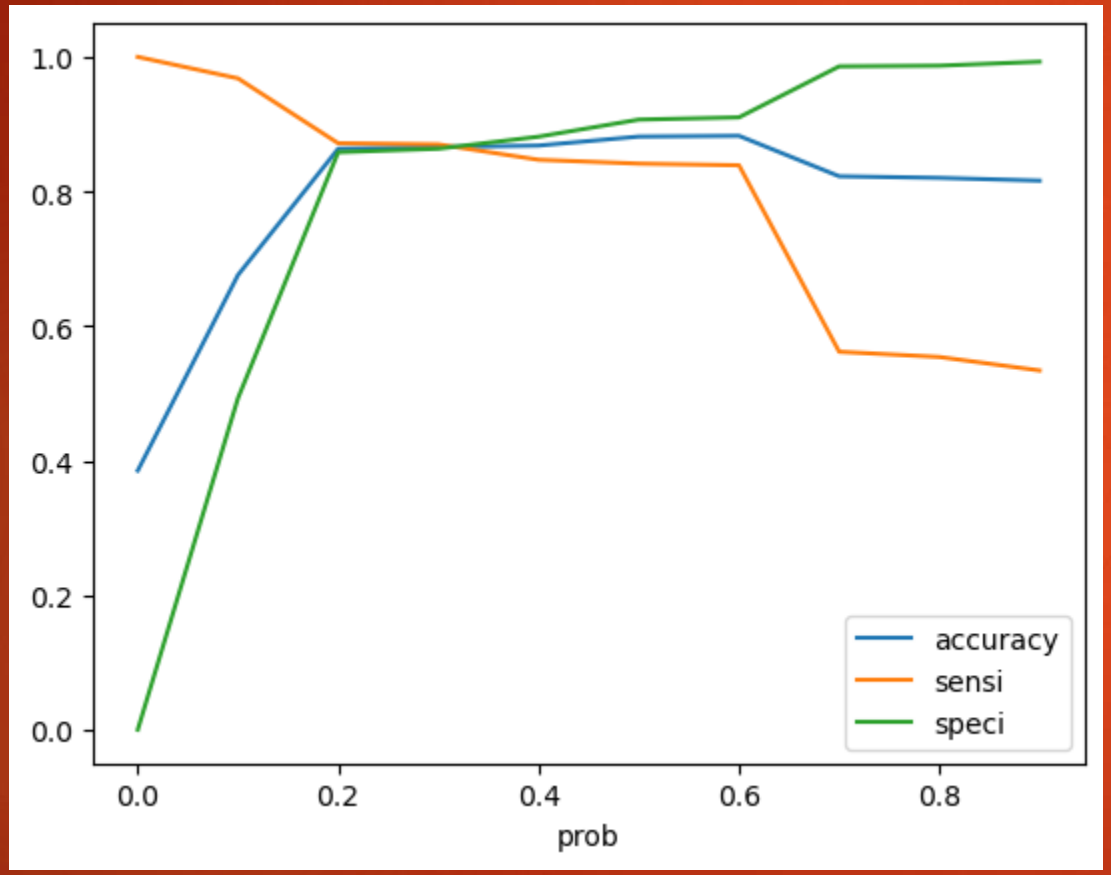
- On test Data

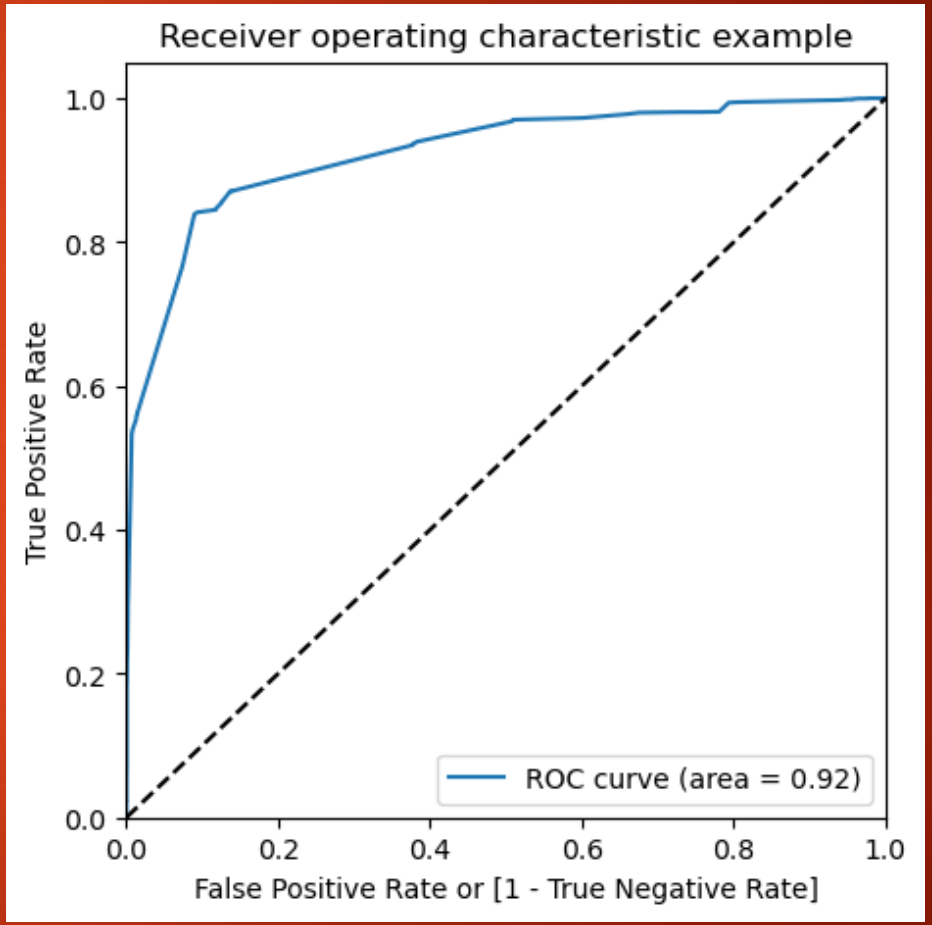
  Accuracy-85%

  Sensitivity- 84%

  Specificity- 85%

ROC Curve


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
prob
accuracy
sensi
speci


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.92)
0.0
0.2
0.4
0.6
0.8
1.0

Thank You