LEAD SCORING CASE STUDY

BY:

Sayantan Singha

Sagar Rishi

Shreeragh C V

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%.

X EDUCATION HAS APPOINTED US TO HELP THEM SELECT THE MOST PROMISING LEADS, I.E. THE LEADS THAT ARE MOST LIKELY TO CONVERT INTO PAYING CUSTOMERS. THE COMPANY REQUIRES US 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%.

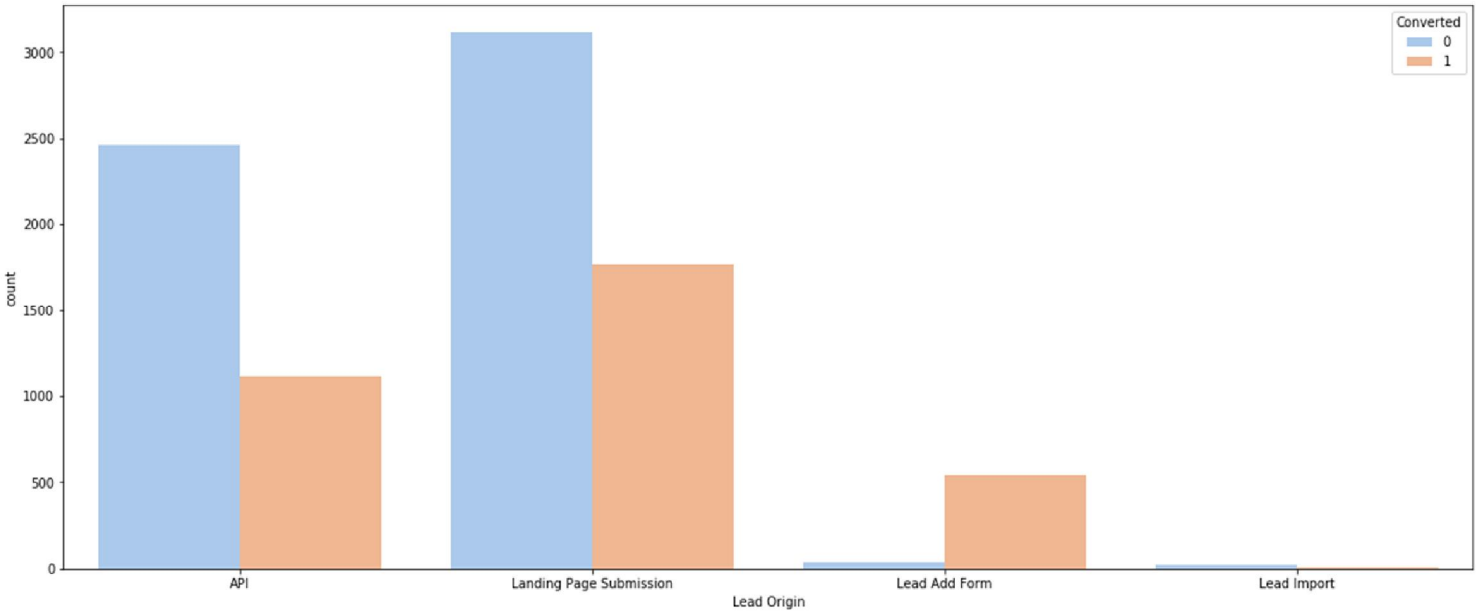
ANALYSIS APPROACH

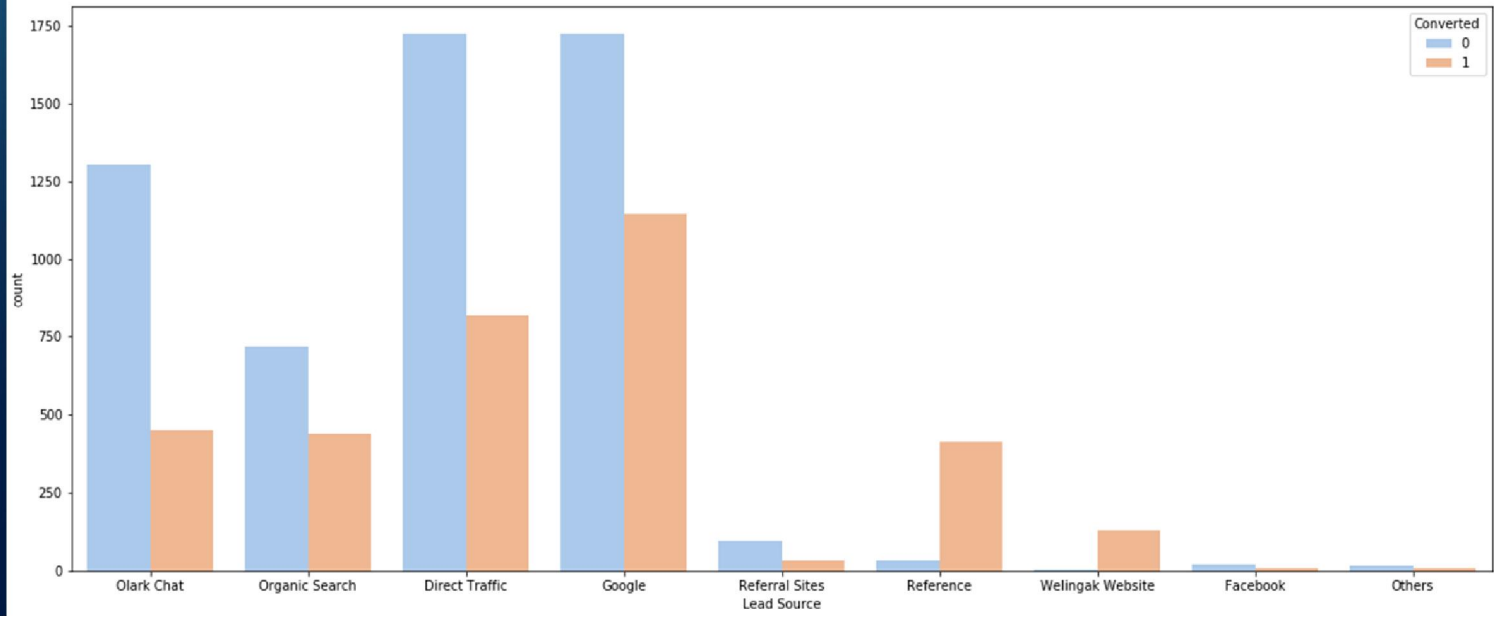
STEP 1: MISSING VALUE TREATMENT

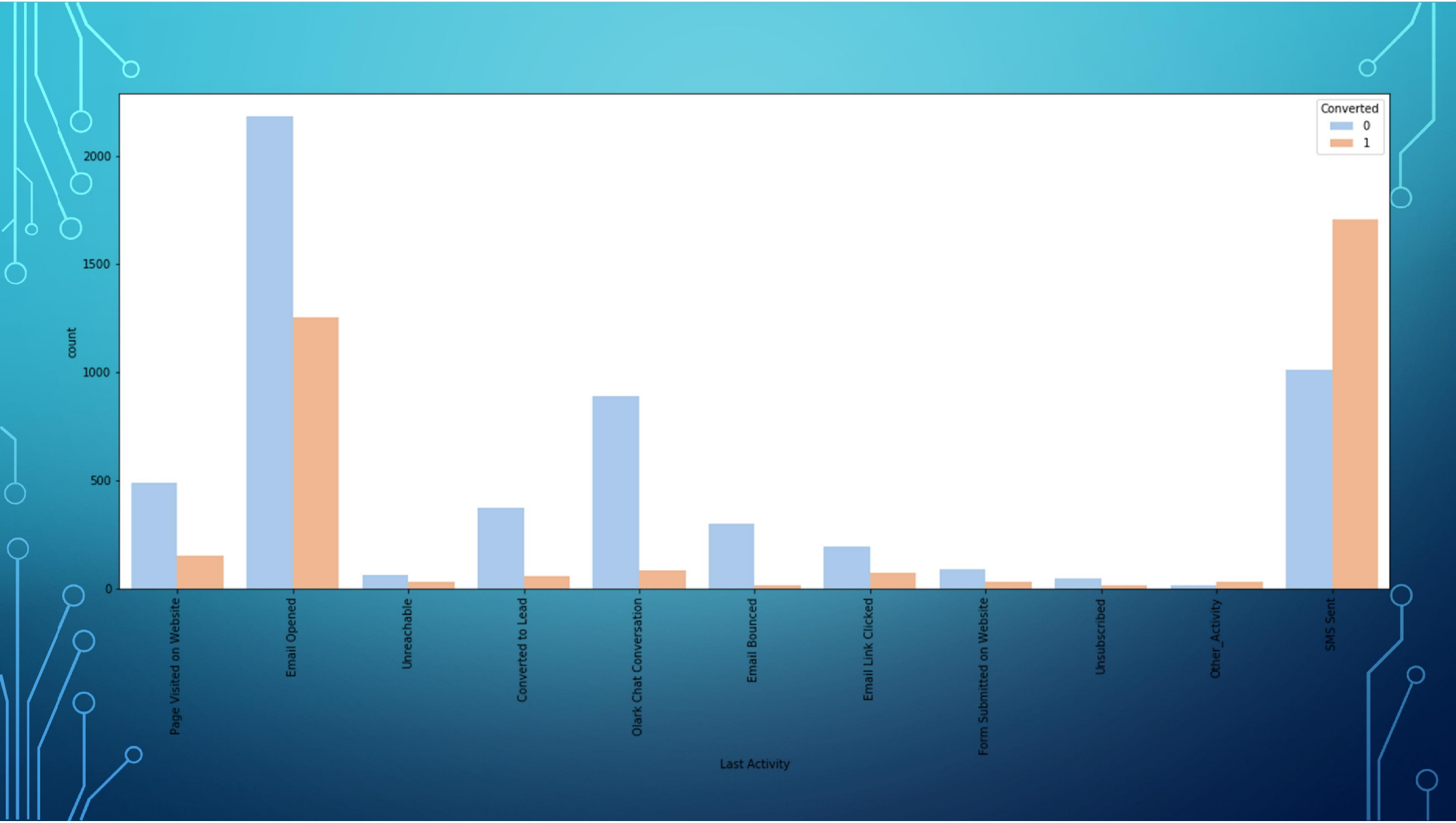
- Missing Values in each columns were treated.
- Columns with missing value > 70% were dropped straight away
- Column 'Lead Quality' had nearly 50% of missing values. As this column had category 'Not Sure', all missing values were imputed with this category
- Other columns with missing values > 40% were also dropped
- For rest of the columns, missing values were imputed with most frequently occurring value.

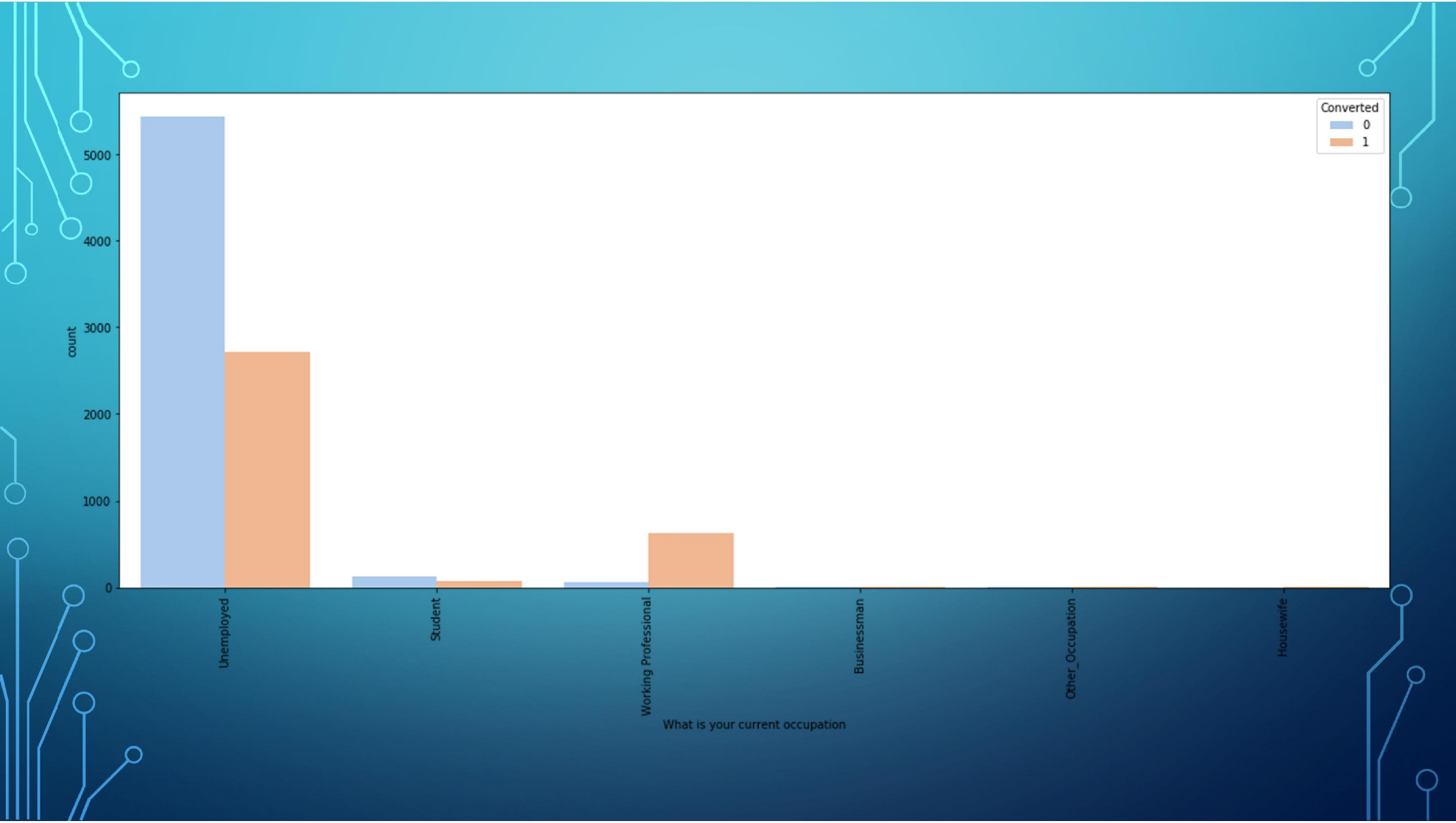
STEP 2: EDA

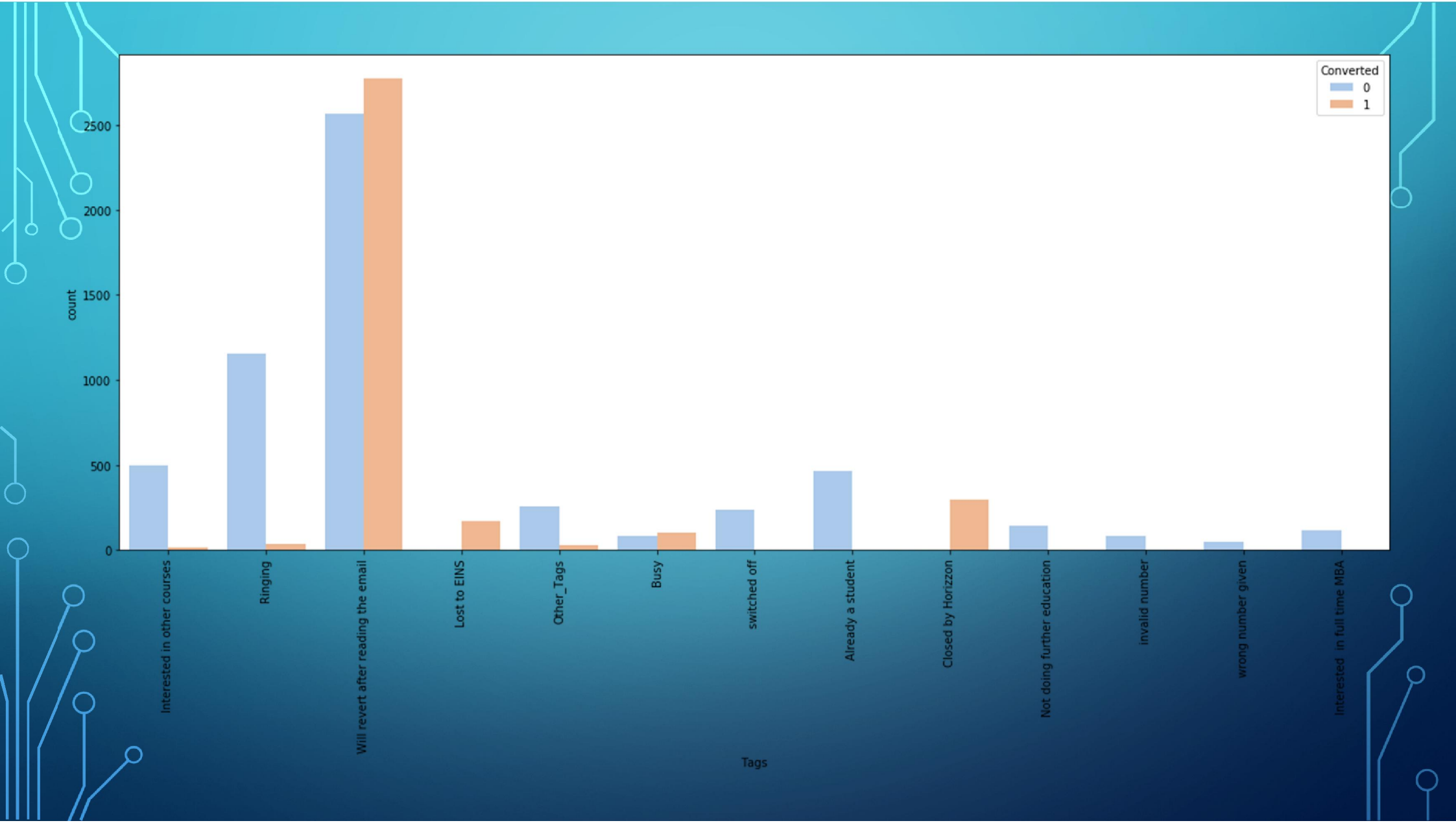
- Each columns was analyzed with respect to the target column, i.e., 'Converted' column
- Importance of the column for building the logistic model was determined by analyzing each graph
- Those column which did not have any significant affect on the target column were dropped at the end of EDA before moving on to model building

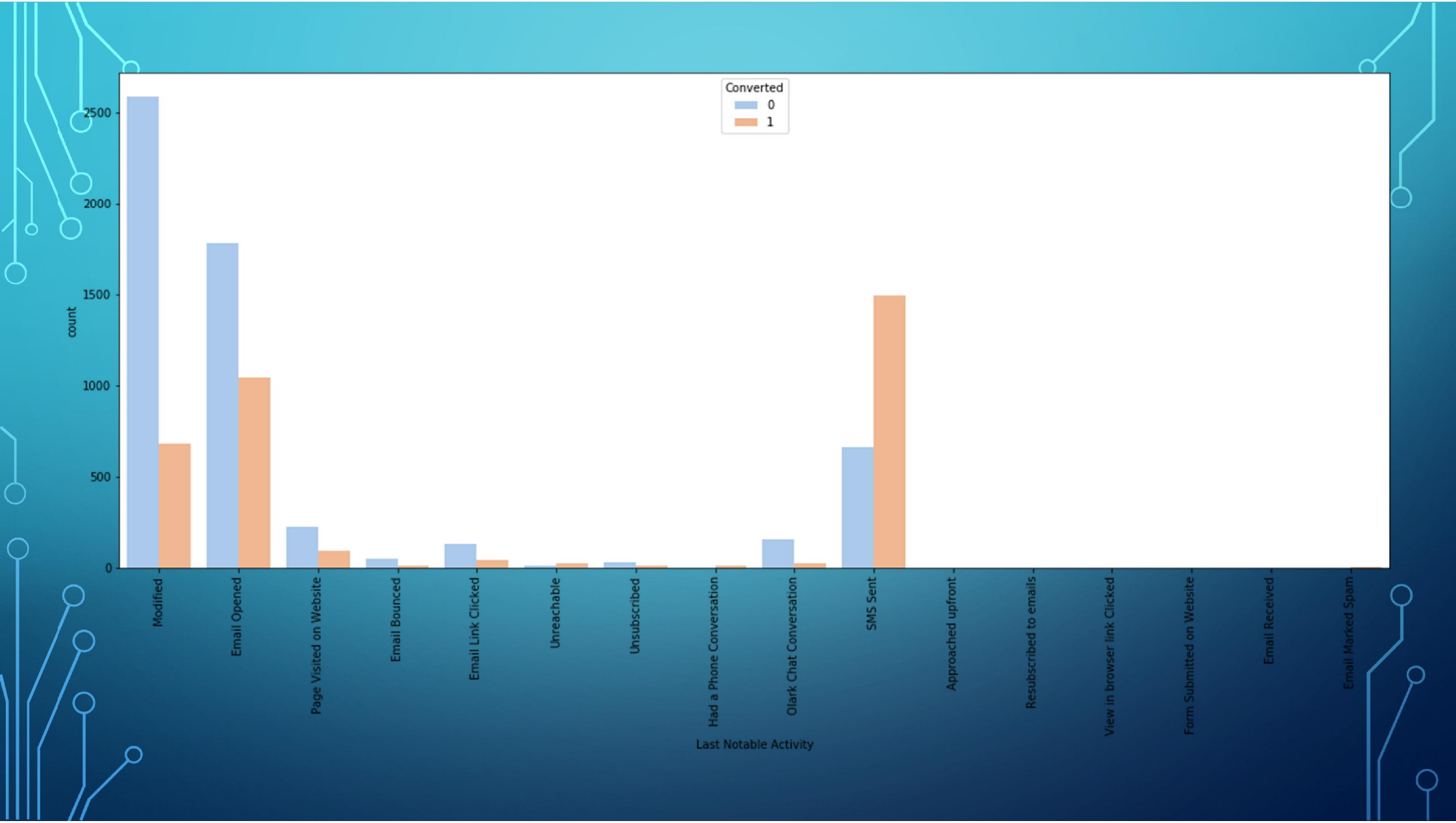


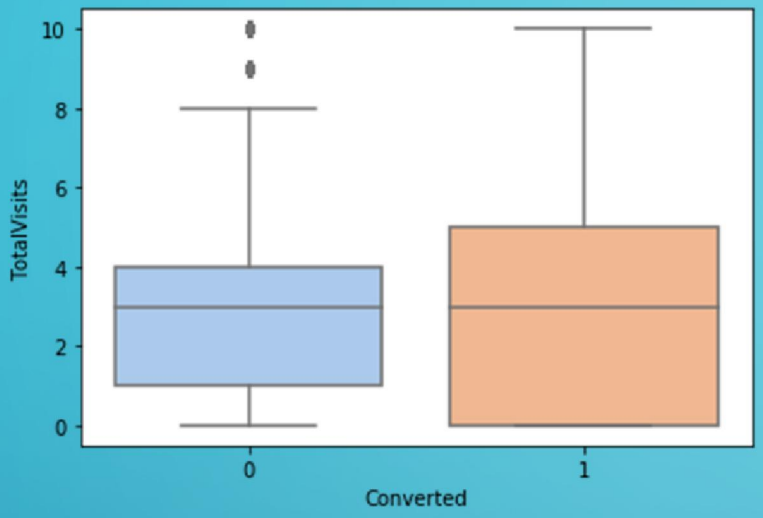


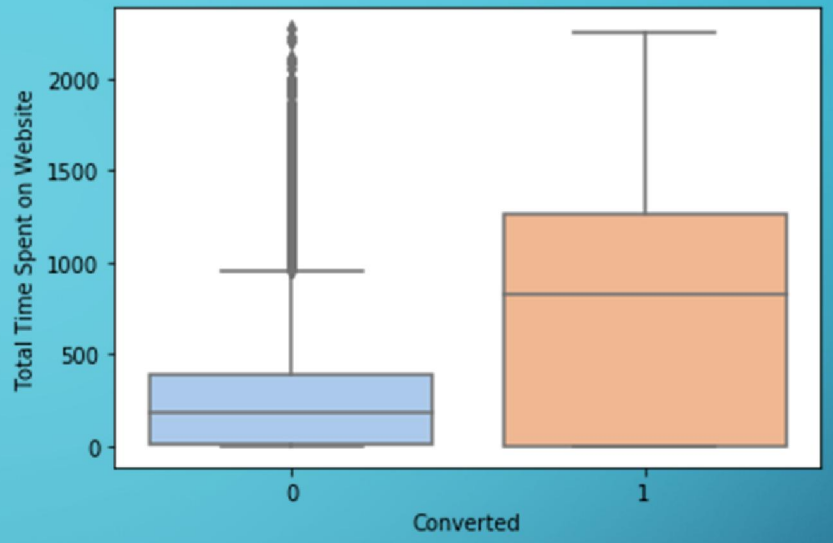

Converted
0
1
2000
1500
1000
500
0
count
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Other_Activity
SMS Sent
Last Activity

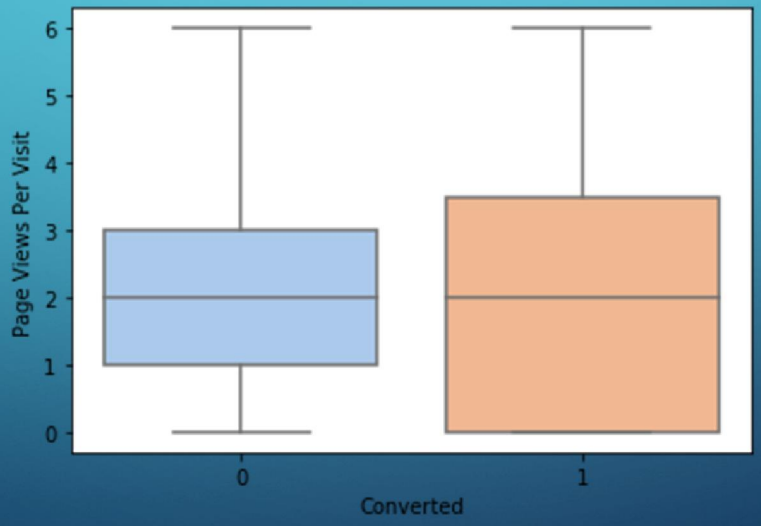

Converted
0
1
count
5000
4000
3000
2000
1000
0
Unemployed
Student
Working Professional
Businessman
Other_Occupation
Housewife
What is your current occupation


Converted
0
1
count
2500
2000
1500
1000
500
0
Interested in other courses
Ringing
Will revert after reading the email
Lost to EINS
Other_Tags
Busy
switched off
Already a student
Closed by Horizzon
Not doing further education
invalid number
wrong number given
Interested in full time MBA
Tags


Converted
0
1
count
2500
2000
1500
1000
500
0
Modified
Email Opened
Page Visited on Website
Email Bounced
Email Link Clicked
Unreachable
Unsubscribed
Had a Phone Conversation
Olark Chat Conversation
SMS Sent
Approached upfront
Resubscribed to emails
View in browser link Clicked
Form Submitted on Website
Email Received
Email Marked Spam
Last Notable Activity


TotalVisits
10
8
6
4
2
0
0
1
Converted


Total Time Spent on Website
2000
1500
1000
500
0
0
1
Converted

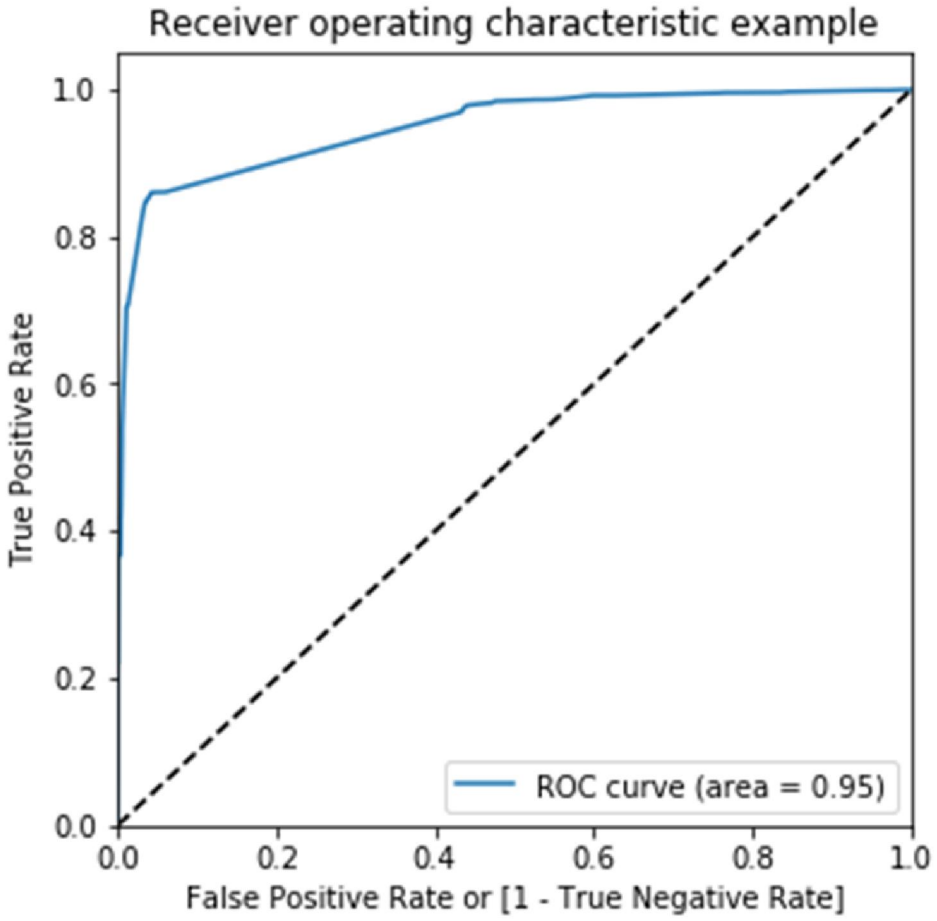

Page Views Per Visit
6
5
4
3
2
1
0
0
1
Converted

STEP 3: DATA PREPARATION & MODEL BUILDING

- For all categorical column, dummy columns were created, and original column was dropped.
- Dataset was split into Train and Test.
- For train dataset, numerical columns were standardized.
- Model was build using RFE, considering Top-15 columns.
- After that, each column's p-value was checked and those column with higher p-values were dropped manually.
- Final model was arrived with 13 feature variable.
- VIF of the feature variables were also checked and found ok.

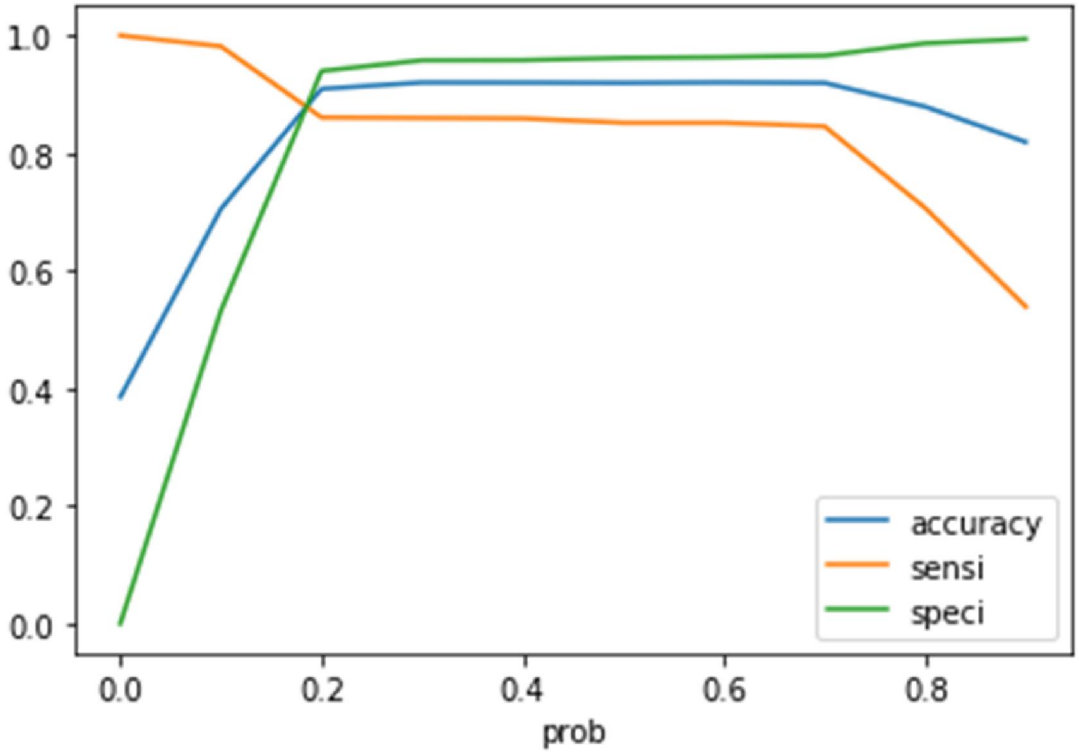
STEP 4: ROC CURVE PLOT

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test



STEP 5: FINDING OPTIMAL CUTOFF POINT

- **Optimal cutoff probability is that prob where we get balanced sensitivity and specificity**



STEP 6: MAKING PREDICTION ON TEST SET

- Numerical columns of Test dataset were standardized.
- Final model developed was used to predict target variable of the test dataset

Comparison of Accuracy, Sensitivity & Specificity of Train and Test dataset

	Sensitivity	Specificity
Train	86.05	93.95
Test	84.42	93.88

CONCLUSION

Following are the variables / dummy variables which should be focused on

- 'Do Not Email'
- 'Lead Origin - Lead Add Form'
- 'Lead Source - Welingak Website'
- 'What is your current occupation - Unemployed'
- 'Tags - Busy'
- 'Tags - Closed by Horizzon'
- 'Tags - Lost to EINS'
- 'Tags - Ringing'
- 'Tags - Will revert after reading the email'
- 'Tags - switched off'
- 'Lead Quality - Not Sure'
- 'Lead Quality - Worst'
- 'Last Notable Activity - SMS Sent'

THANK YOU