LEAD SCORING CASE STUDY

By : Yamini K, Prashanth D and Ankith S A

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. 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 most potential leads, also known as ‘Hot Leads’.
- Deployment of the model for the future use.

Solution Approach

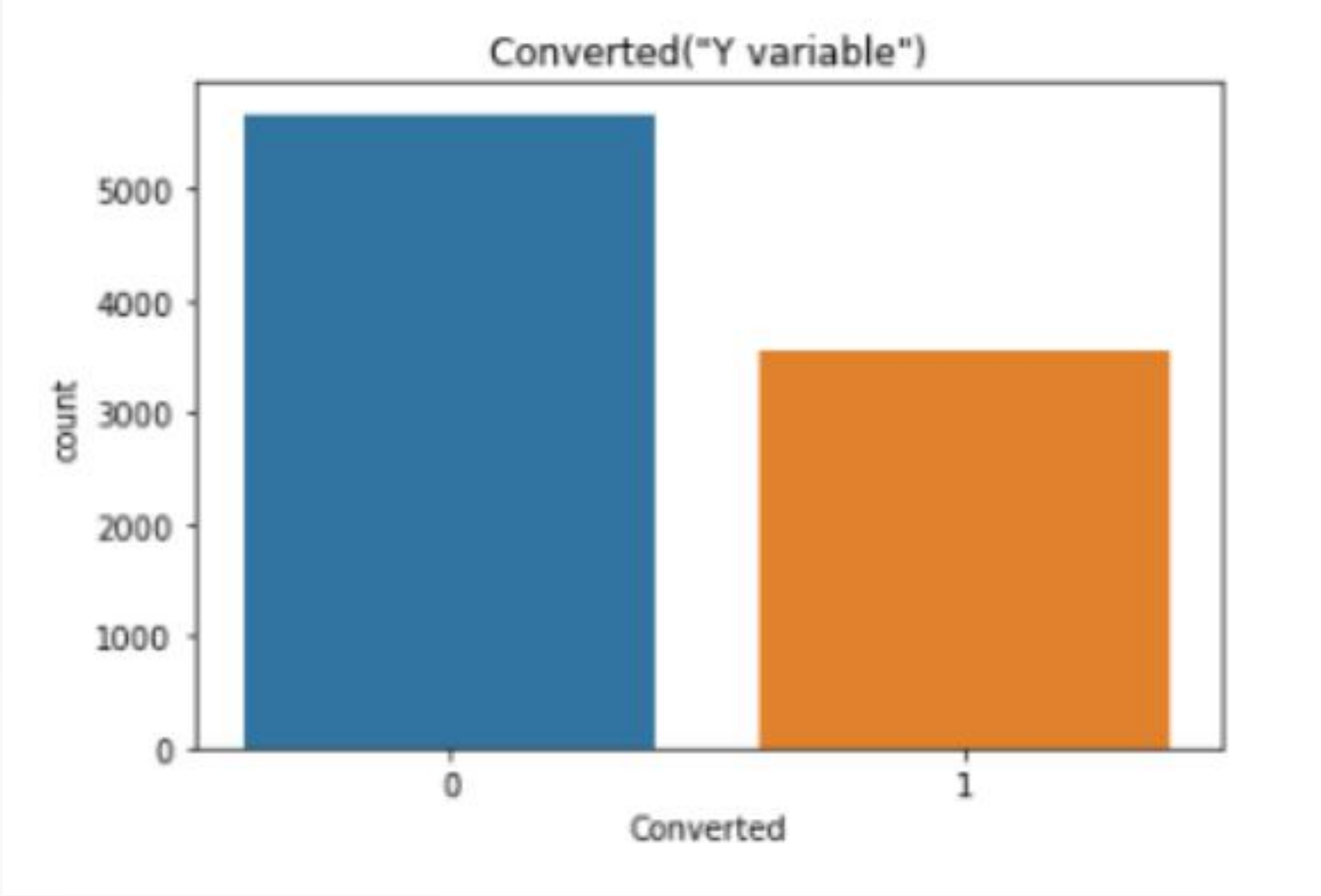
- Data cleaning and data manipulation.
    1. Check and handle duplicate data.
    2. Check and handle NA values and missing values.
    3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
    4. Imputation of the values, if necessary.
    5. Check and handle outliers in data.
- Exploratory Data Analysis :
    1. Univariate data analysis: value count, distribution of variable etc.
    2. 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 the model and Model presentation.
- Conclusions and recommendations.

Data Cleaning and Data Manipulation

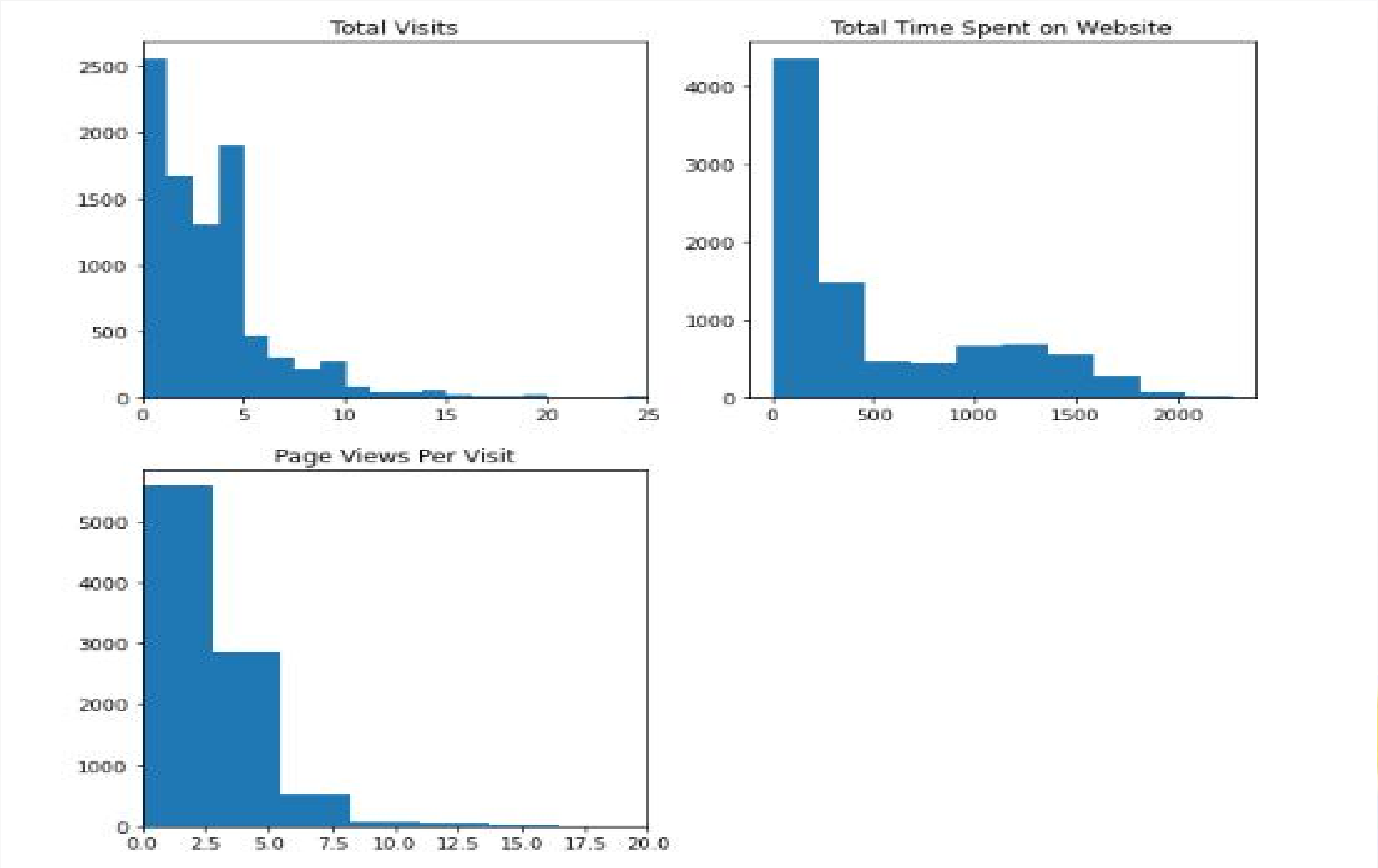
- Total Number of Rows =9240 , Total Number of Columns =37.
- Dropping unique valued columns,Magazine,Receive More Updates About Our Courses,I agree to pay the amount through cheque,Get updates on DM Content, Update me on Supply Chain Content.
- Removing all the columns that are not required like Prospect ID and Lead Number.
- Dropping the columns having more than 35% as missing value such as ‘How did you hear
- about X Education’ and ‘Lead Profile’.
- Dropping columns like Do Not Call, What matters most to you in choosing course, Search, Newspaper, Article, X Education Forums, Newspaper, Digital Advertisement, which do not have enough variance.
- After cleaning and data manipulation, we have 21 rows and 9074 columns .

Exploratory Data Analysis

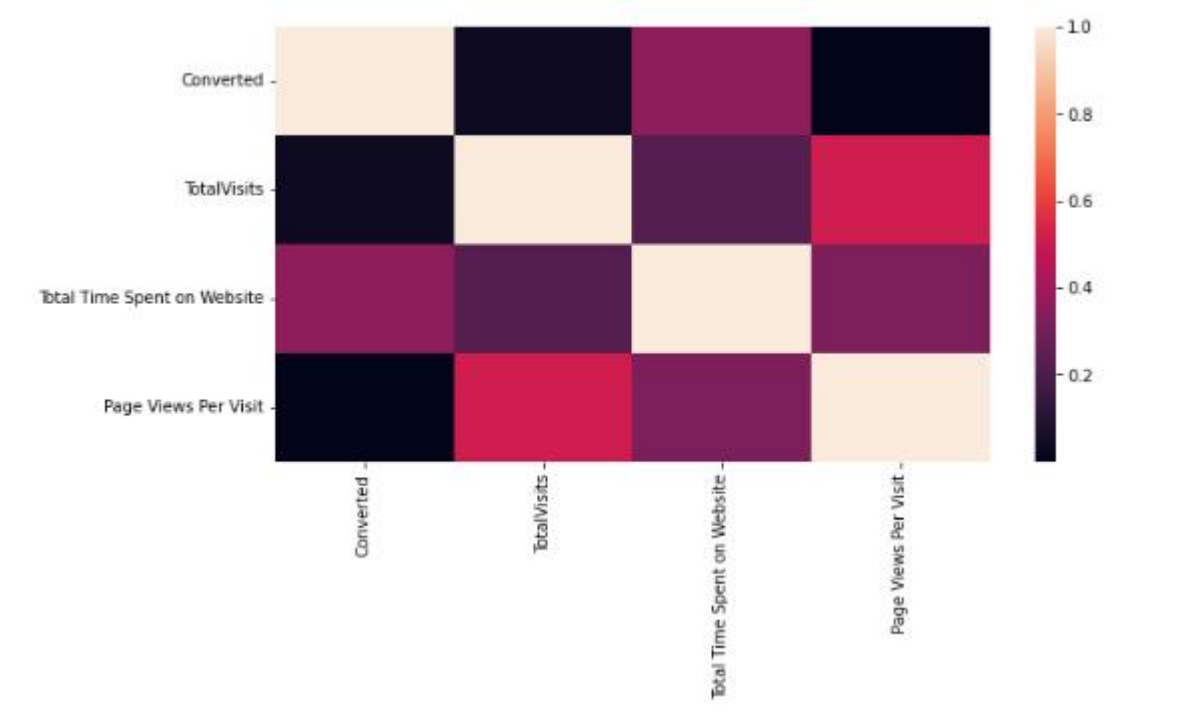
From the graph, we can observe that, the conversion rate is only 37%



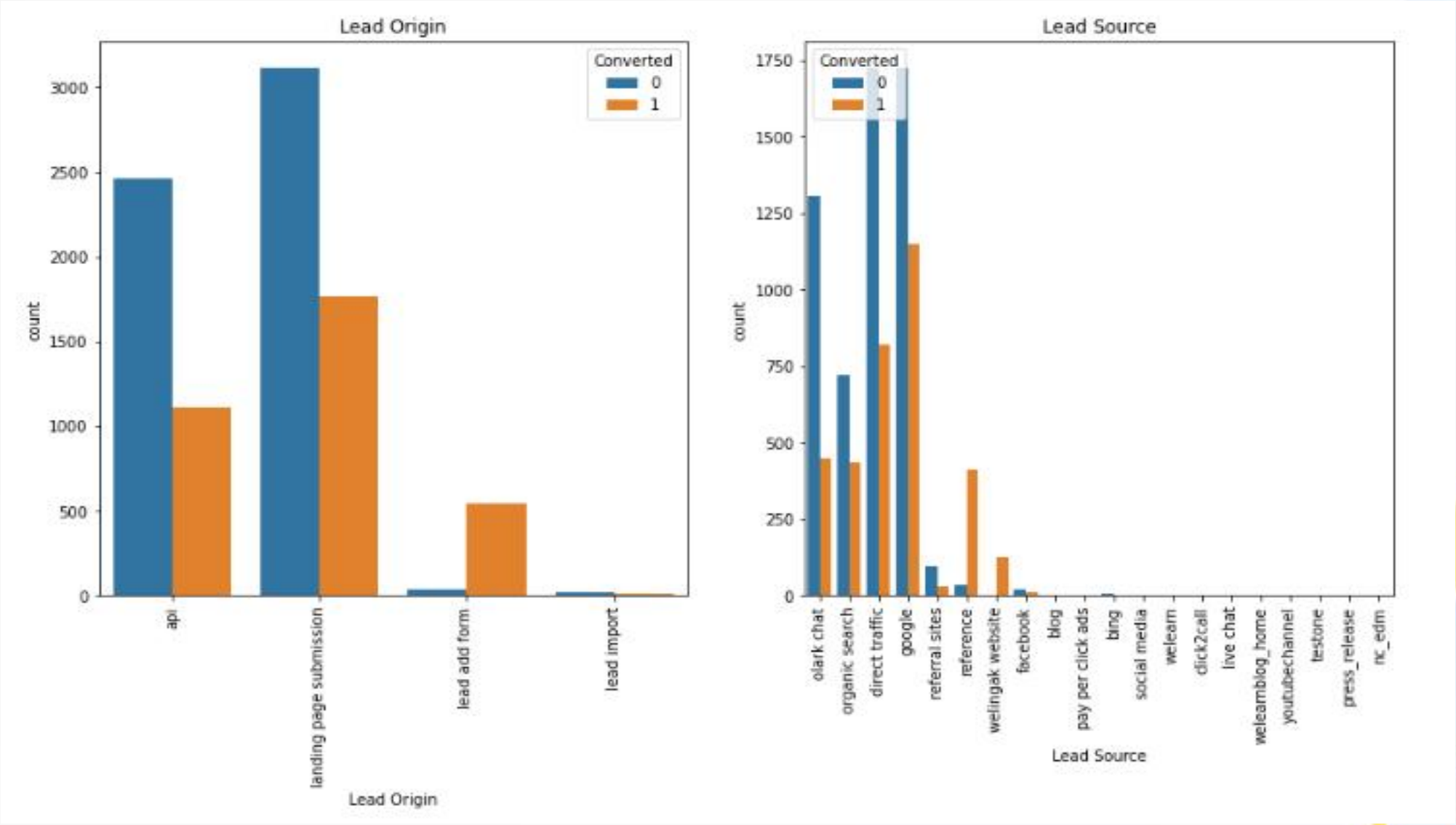
Plots of numerical variables to view the trends :

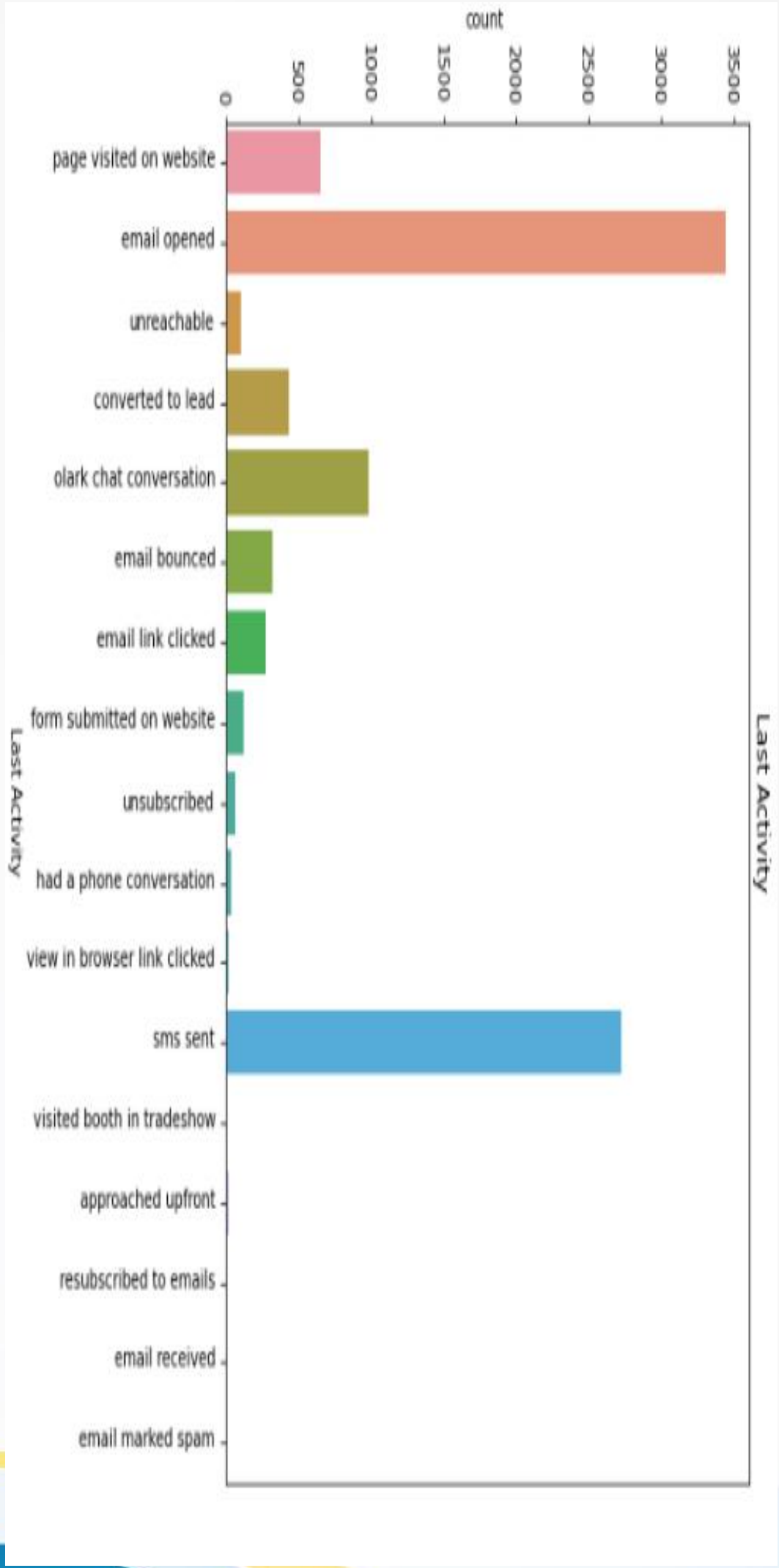


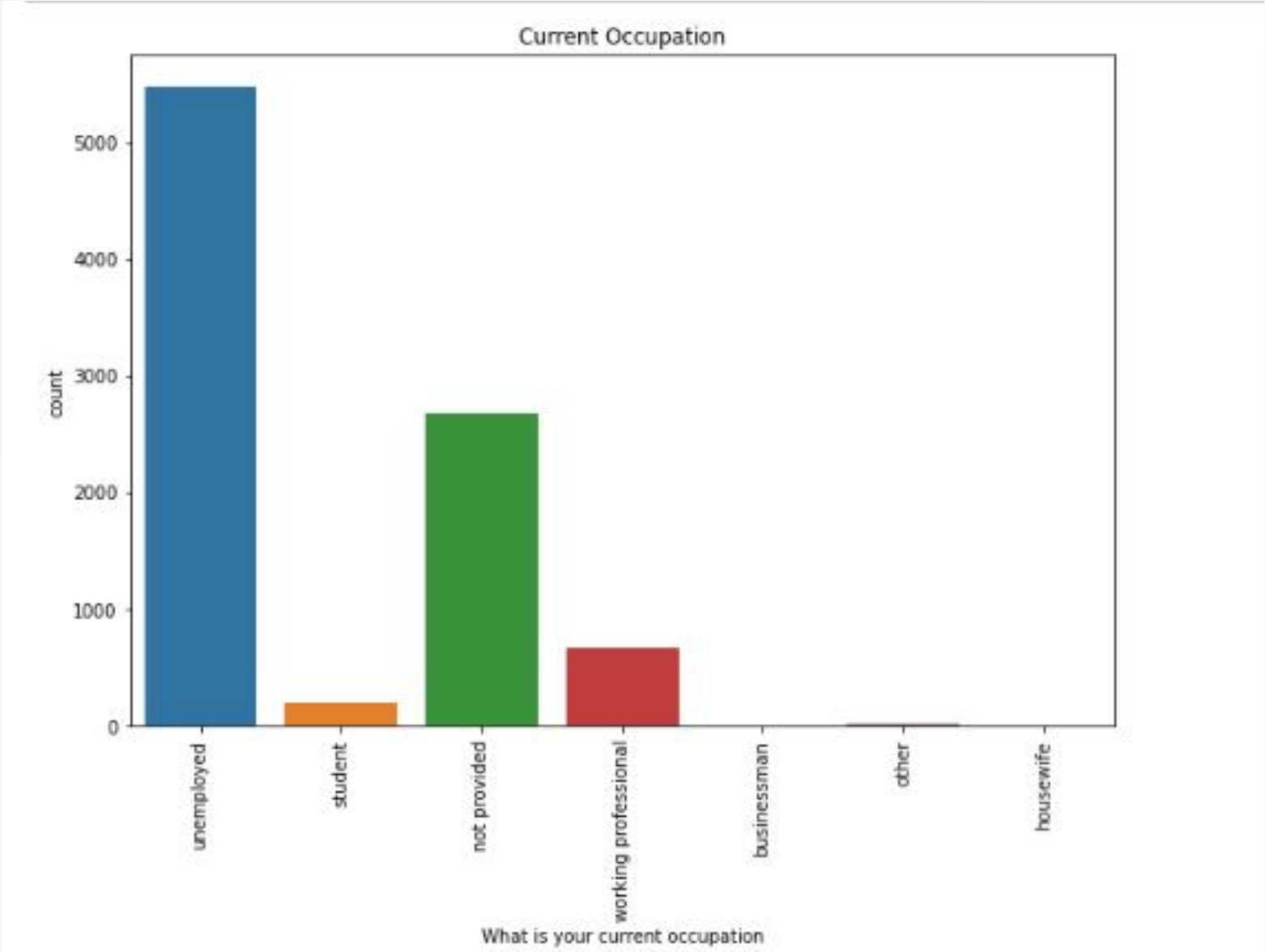
Co-relation of numerical variables :

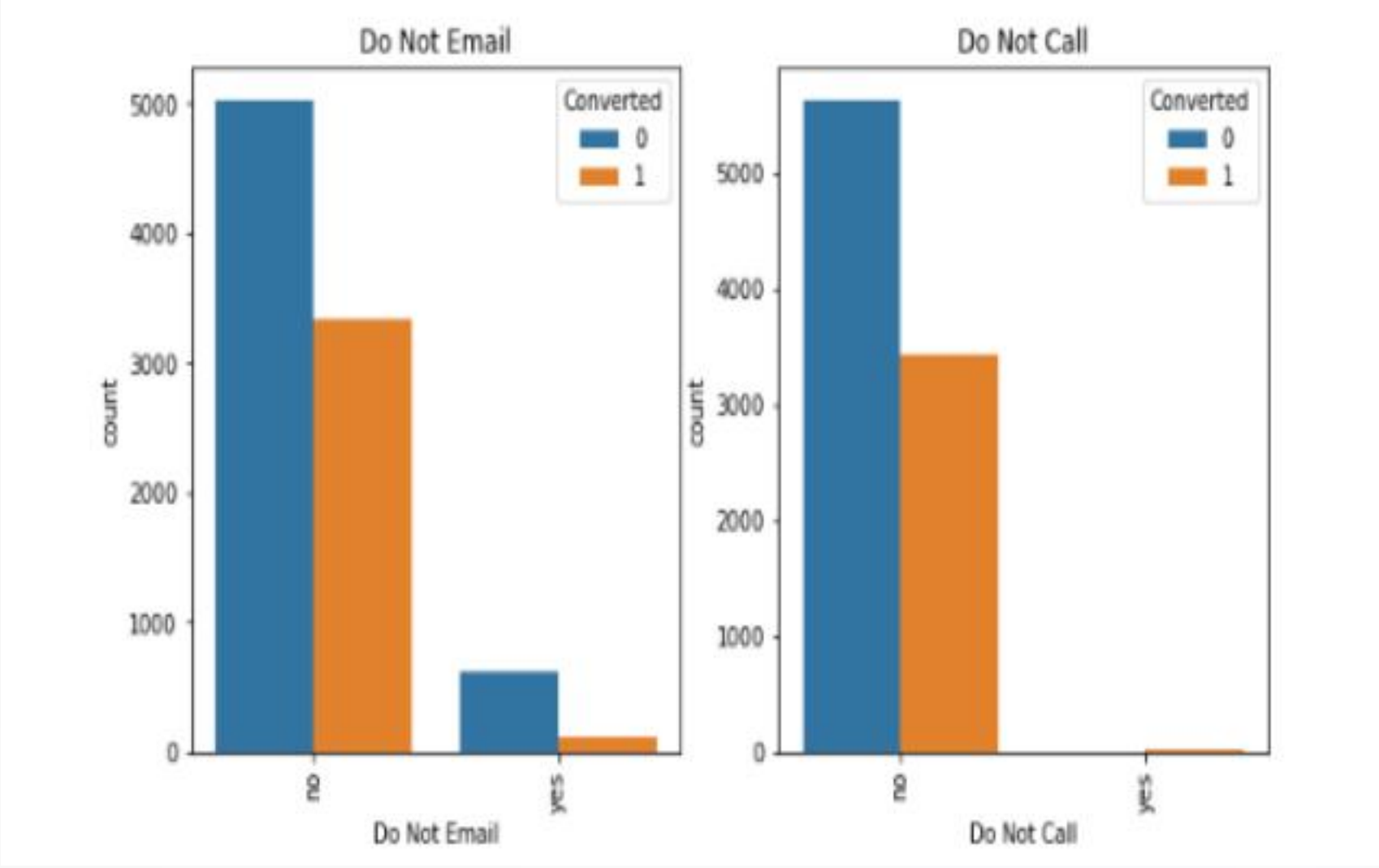


Plots of Categorical variables to view the trends :




count
0
500
1000
1500
2000
2500
3000
3500
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
sms sent
visited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam
Last Activity
Last Activity

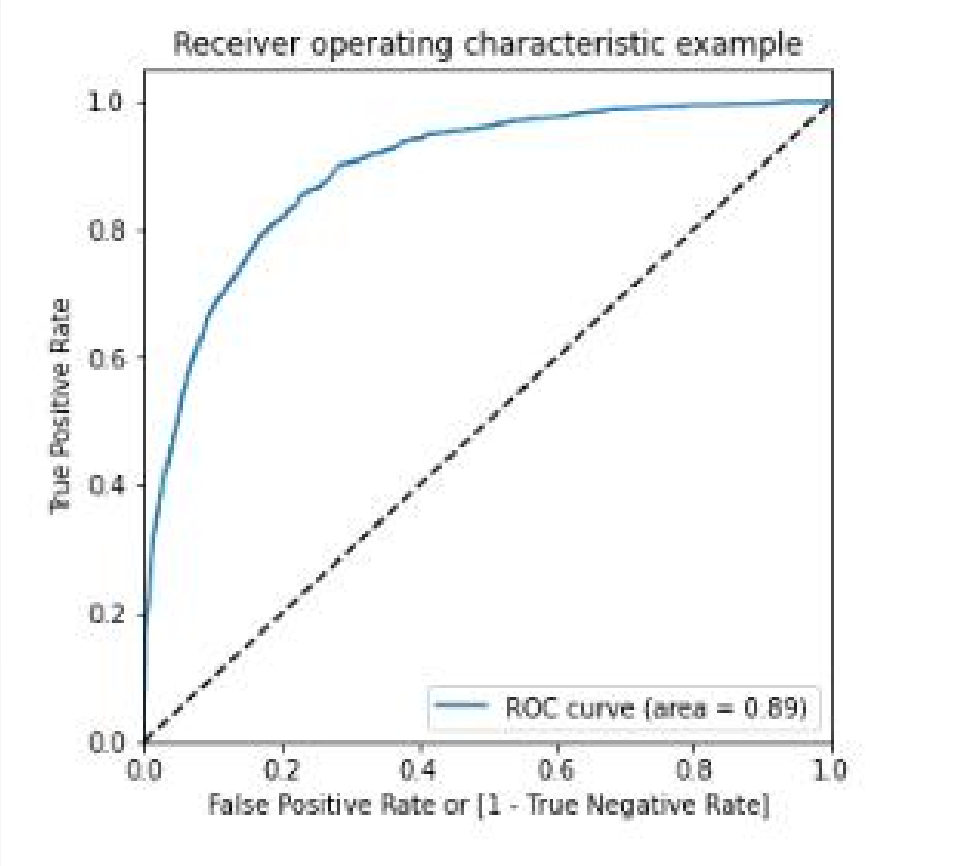

Current Occupation
count
5000
4000
3000
2000
1000
0
unemployed
student
not provided
working professional
businessman
other
housewife
What is your current occupation

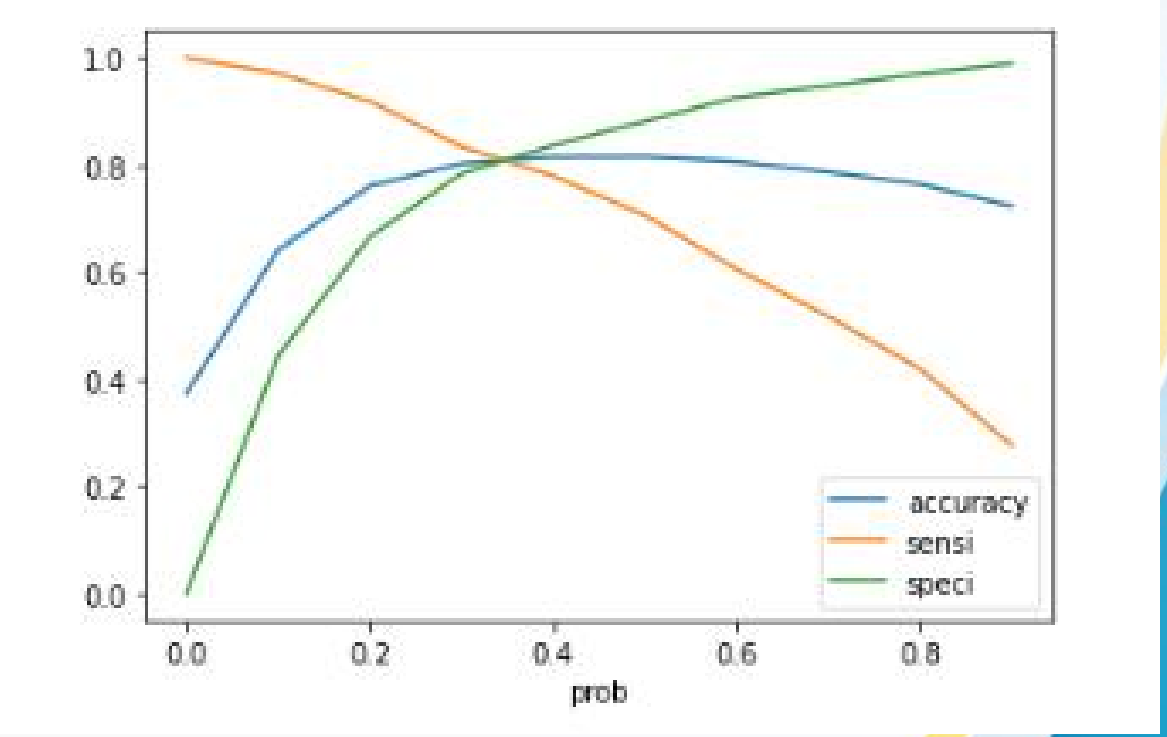

Do Not Email
Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
Do Not Email
Do Not Call
Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
Do Not Call

Model Building

- Splitting the Data into Training and Testing Sets
- The first step for regression is performing a train-test split, we have chosen 70:30 ratio.
- We use Recursive Feature Elimination to reduce the number of features to 15.
- Building Model by removing the variable whose p - value is greater than 0.05 and vif value is greater than 5
- After building the model, we make predictions based on test data.

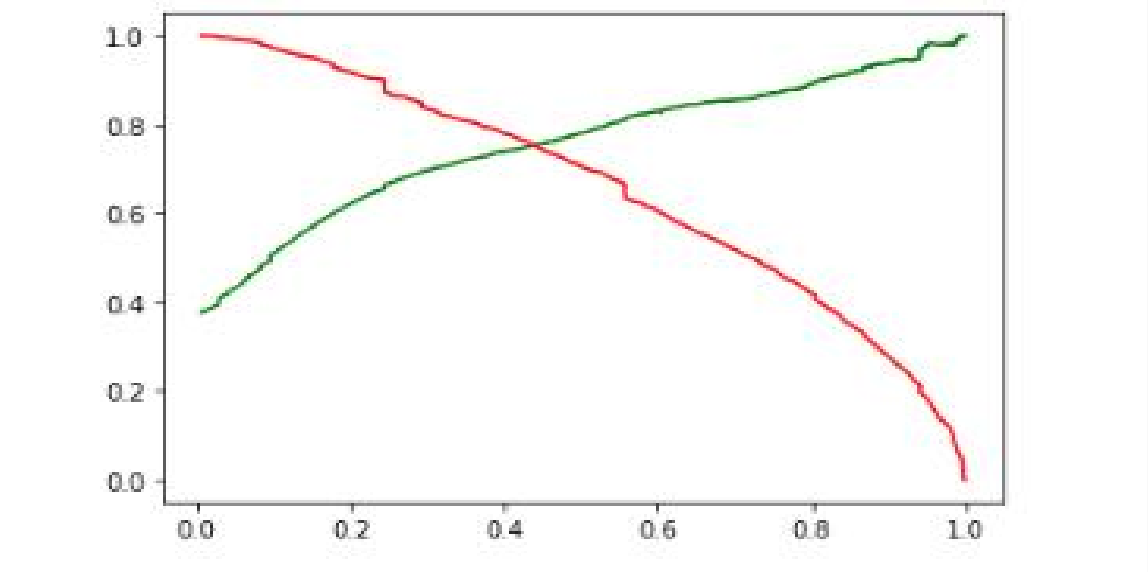
ROC Curve


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.89)
0.0
0.2
0.4
0.6
0.8
1.0


accuracy
sensi
speci
prob
0.0
0.2
0.4
0.6
0.8
1.0

- Optimal cut off probability is that probability where we get balanced sensitivity and specificity.
- From the Precision Recall Curve, we can see that the cut - off point is 0.35.
- With the current cut off as 0.35 ,we get sensitivity of around 79.3% and specificity of around 80%
- We get the overall accuracy to be 80.1%

Precision Recall Curve



- From the Precision Recall Curve, we get neew cut-off to be 0.41.
- With that cut-off, we get the precision and recall , both to be around 75%

Conclusion

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

- The total time spend on the Website.
- Total number of visits.
- When the lead source was:
  - (a) Google
  - (b) Direct traffic
  - (c) Organic search
  - (d) Welingak website
- When the last activity was:
  - (a) SMS
  - (b) Olark chat conversation
- When the lead origin is Lead add format.
- When their current occupation is as a working professional.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.