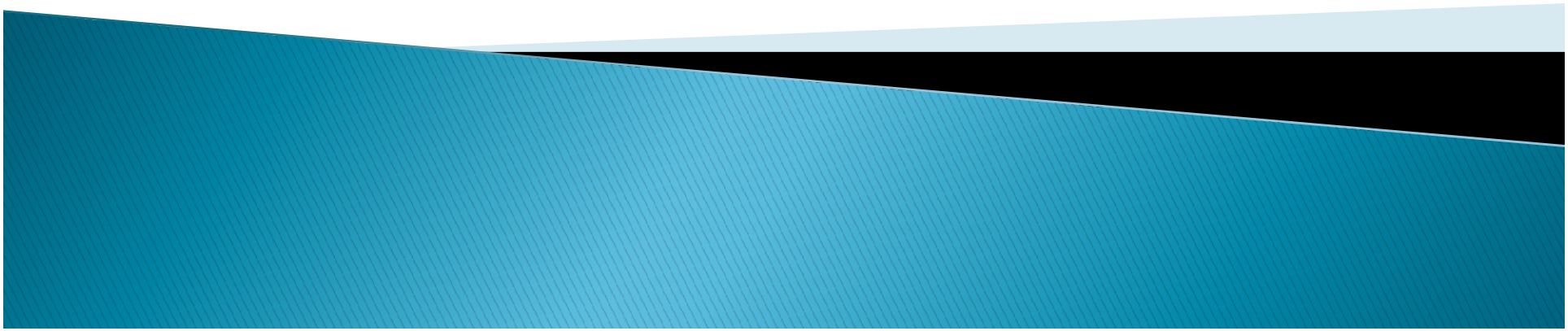
LEAD SCORING CASE STUDY

BY:-PRIYA NARAYAN



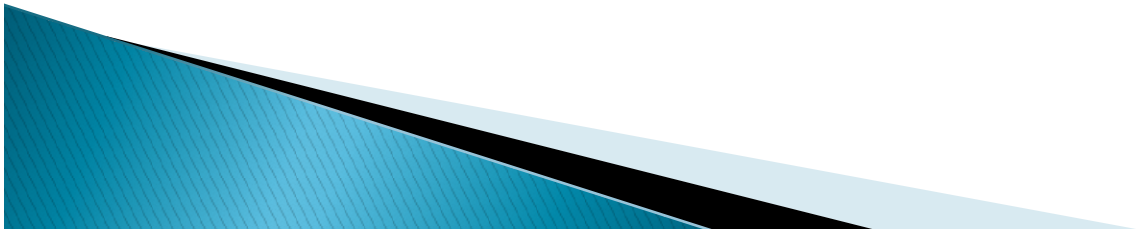
Problem Statement

X education is an organization which provides online courses for industry professionals. The company marks its courses on several popular websites.

X education wants to select most promising leads that can be converted to paying customers.

Although the company generates a lot of leads only a few are converted into paying customers, wherein the company wants a higher lead conversion. Leads come through numerous modes like email, advertisements on websites, Google searches etc.

The company has had 30% conversion rate through the whole process of turning leads into customers by approaching those leads which are to be found having interest in taking the course. The implementation process of lead generating attributes are not efficient in helping conversions.

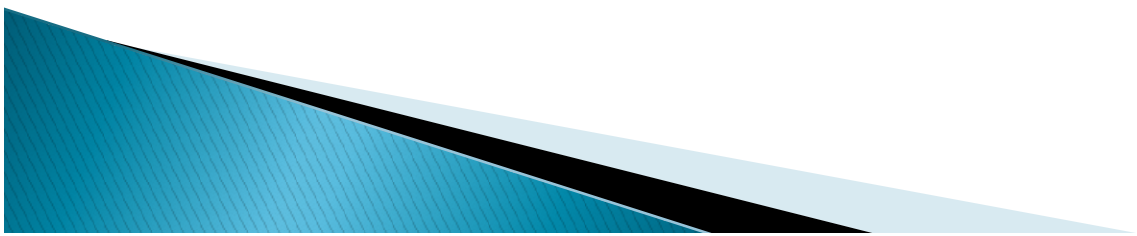


Business Goal

The company require a model to be built for selecting most promising leads.

Lead Score to be given to each leads such that it indicates how promising the lead could be. The higher the lead score the more promising the lead to get converted, the lower is the chances of conversion.

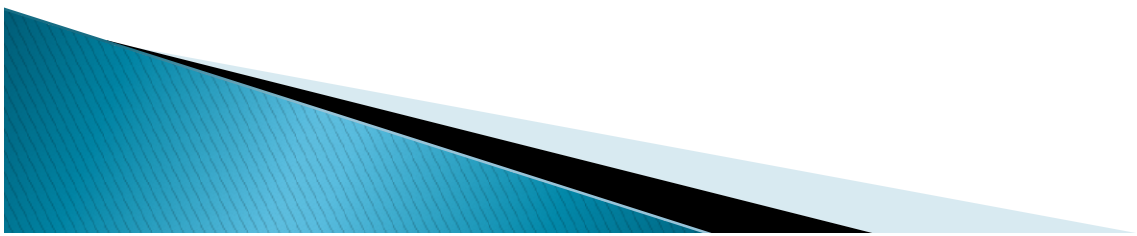
The model to be built in lead conversion rate around 80% or more.

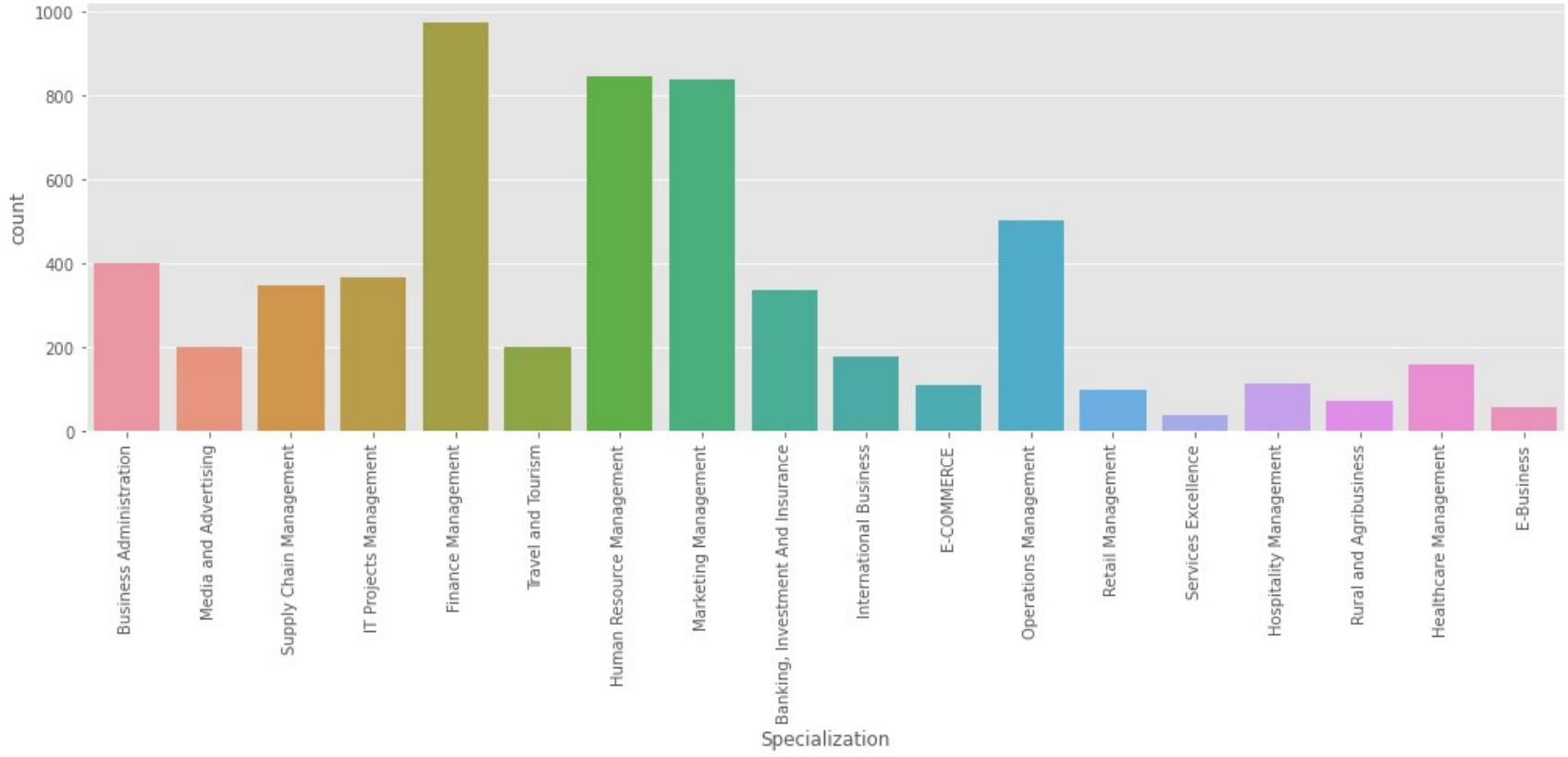


STEPS FOLLOWED

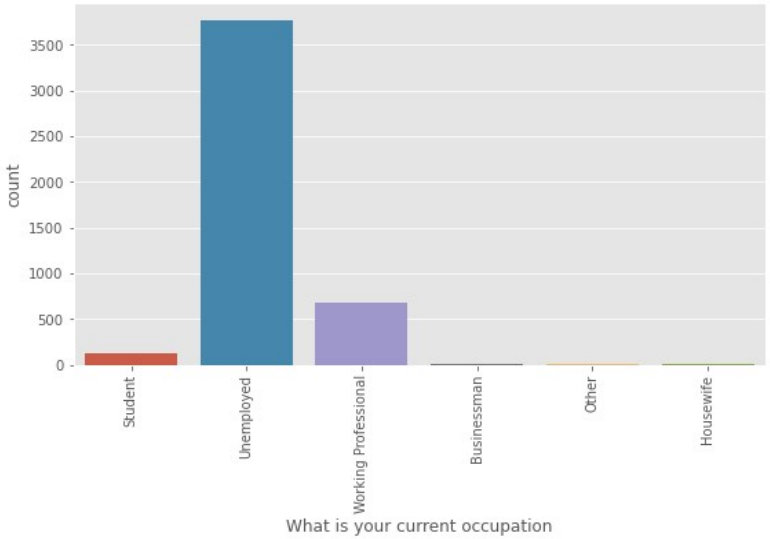
Step -1.Reading and importing the data.
Step-2.Inspecting the Dataset.
Step-3.Cleaning the null values.
Step-4.EDA
Step-5.Transformation/Get Dummies/Label encoding
Step-6:Train-Test Split
    6.1:-Scaling Features
    6.2:- Correlation
Step-7:-Model Building
    7.1:-Feature selection using RFE
    7.2:-Checking for VIF values
    7.3:-Checking for P-values
Step-8:- Making Predictions on the Train and Test Set
Step-9:-Plotting the ROC Curve

Handling Missing Values

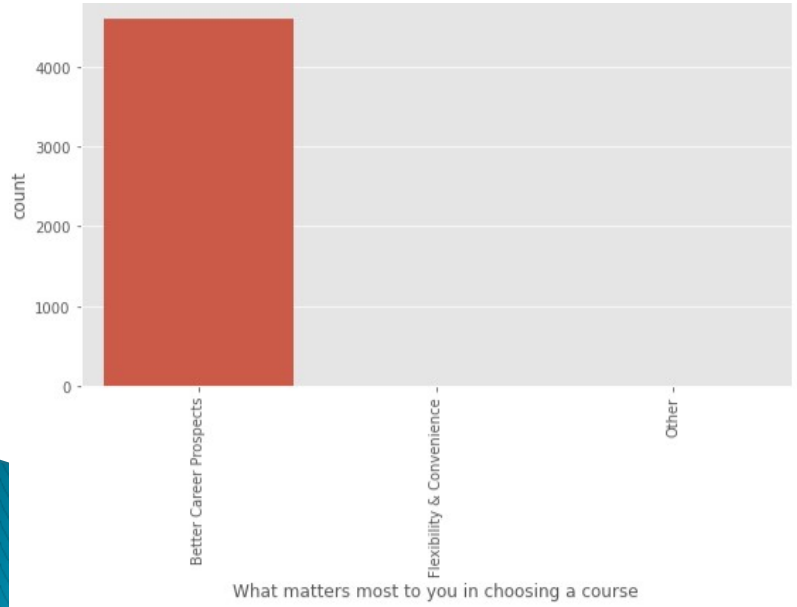




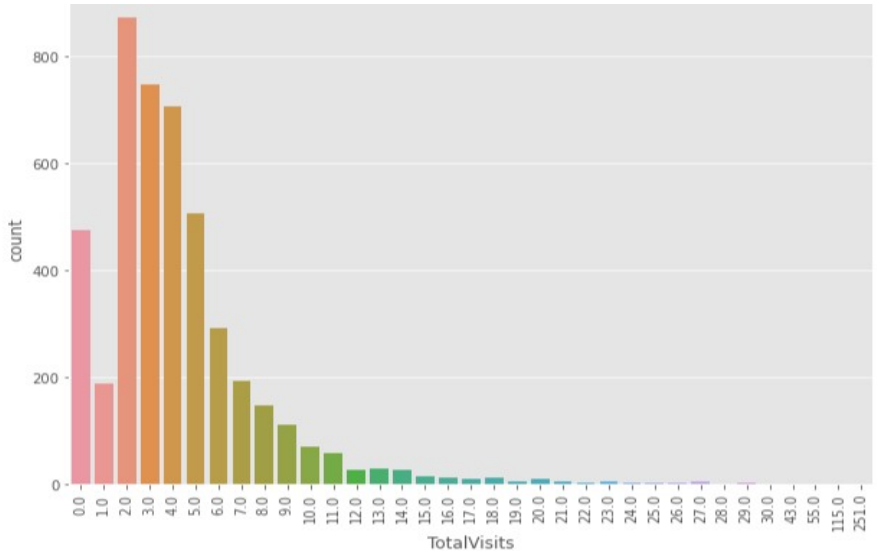
Specialization have 37% missing values present in it. So , it may be possible that the lead may leave this column blank if he may be a student or not having any specialization or his specialization is not there in the options given.



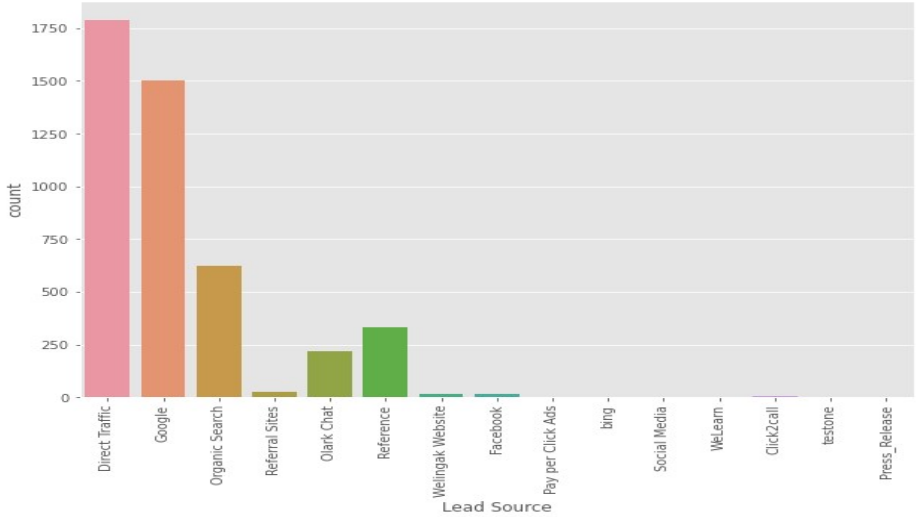
What is your current occupation - This column has 29.11% missing values present.



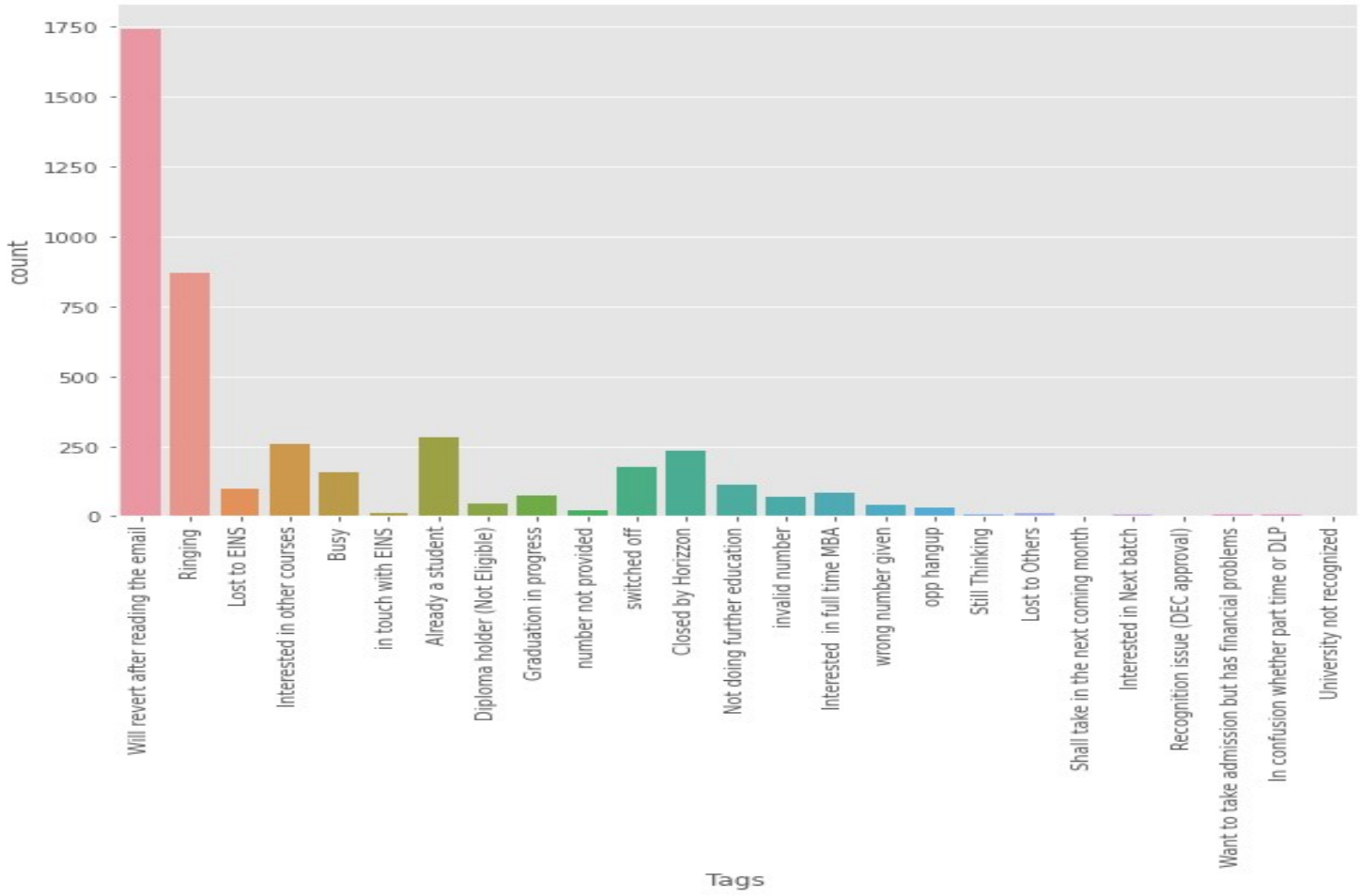
What matters most to you in choosing a course has 26% missing values present.



Total visits showing remarkable growth rate



Lead Source–To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

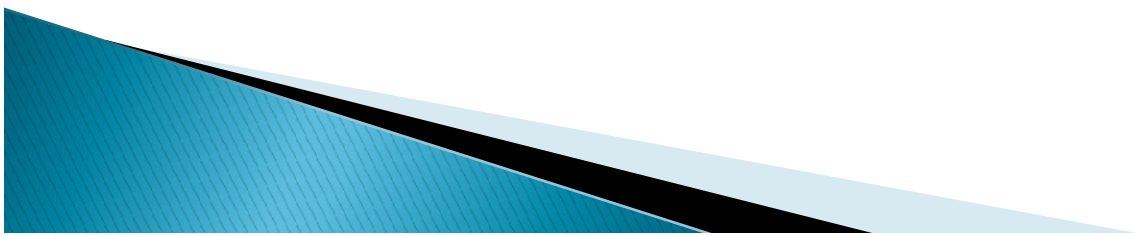


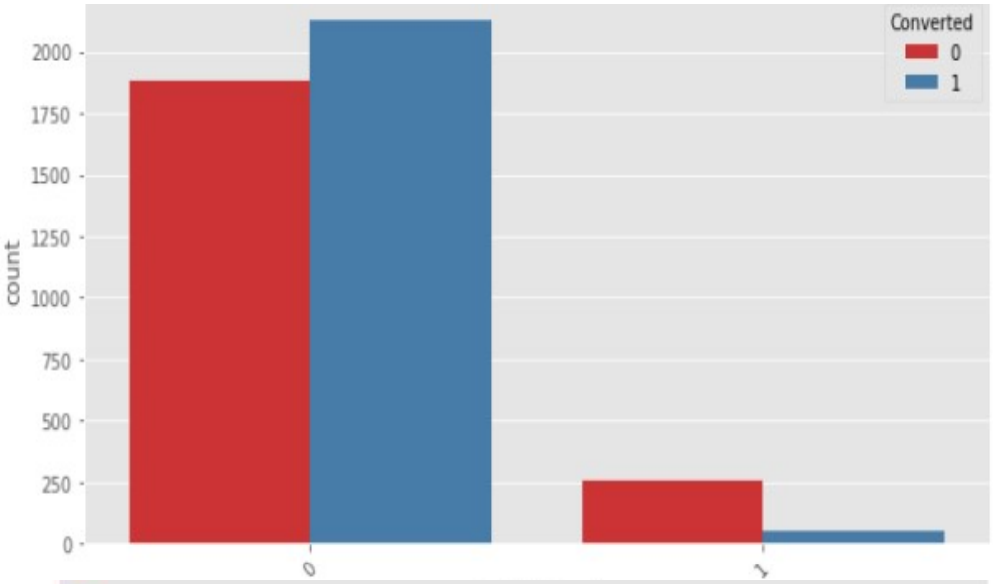
Since most values are 'Will revert after reading the email' , we can impute missing values in this column with this value.

EXPLORATARY DATA ANALYSIS

Pairplot of Lead Dataset

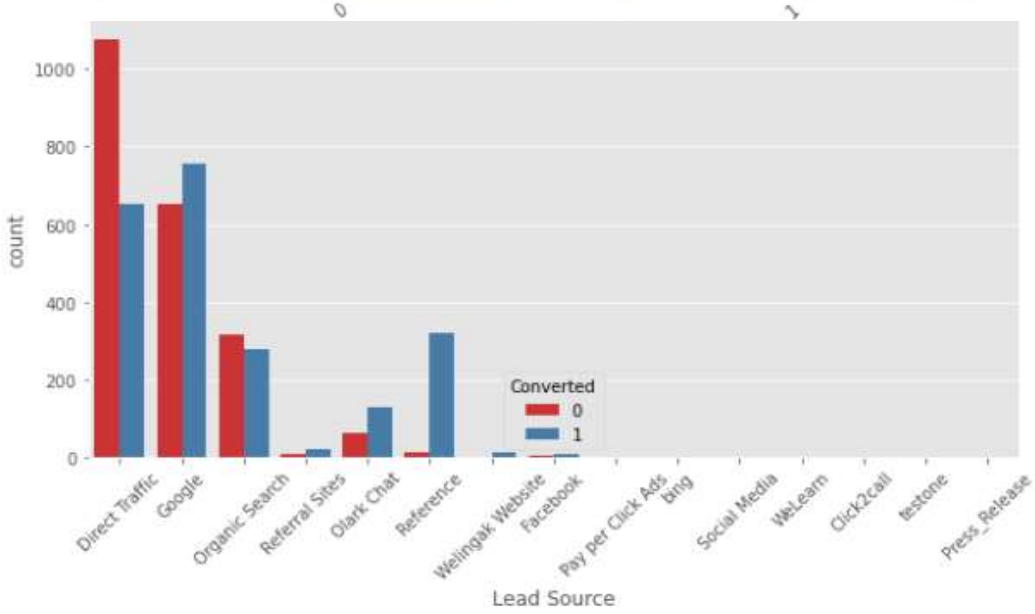
Univariate Analysis





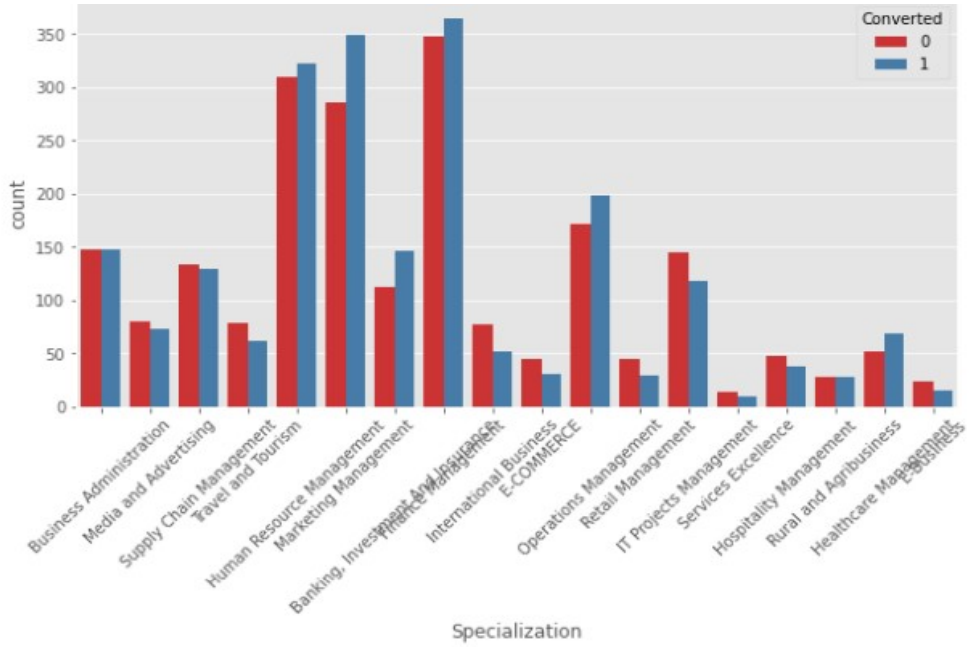
Do Not Email vs Converted

Google searches has high conversion rate in compared to other modes.

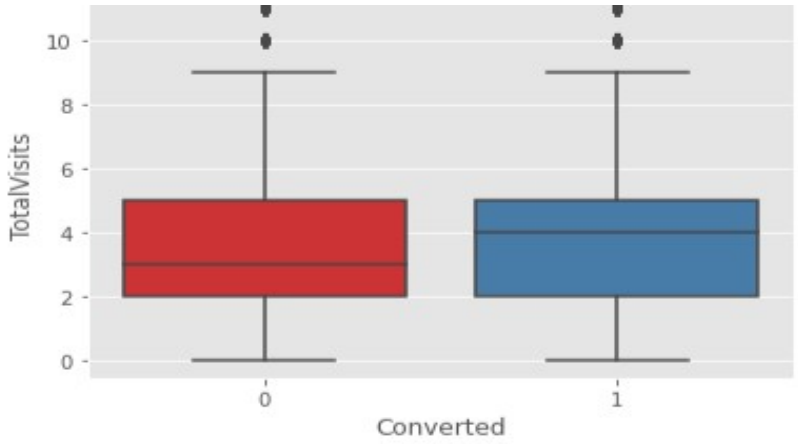


Lead Source Vs Converted

•Google and Direct traffic generates maximum number of leads.
•Conversion Rate of reference leads and leads through welingak website is high.

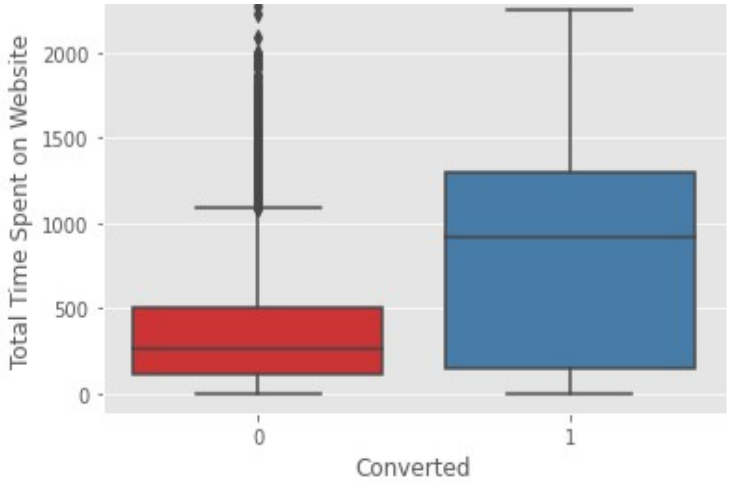


Inference
•**Focus should be more on the Specialization with high conversion rate.**

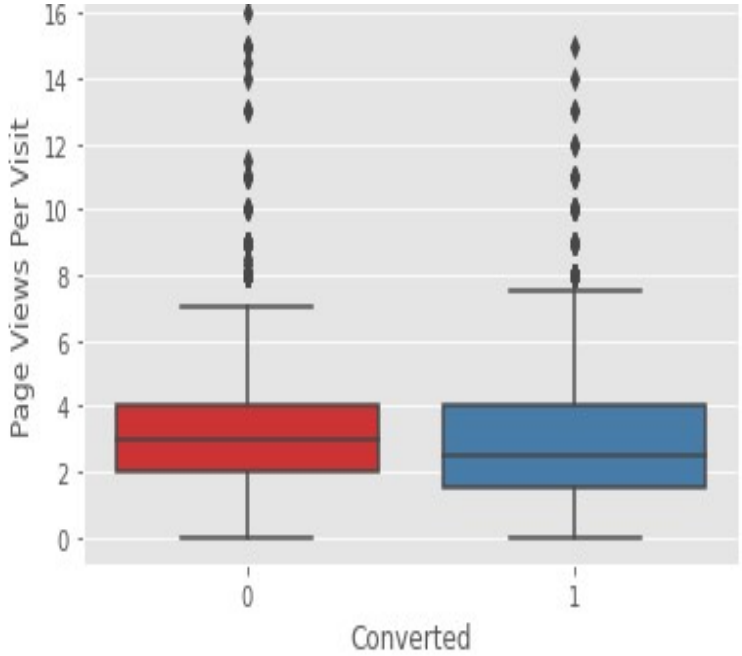


Inference
•Median for converted and not converted leads are the same.
•Nothing can be concluded on the basis of Total Visits.

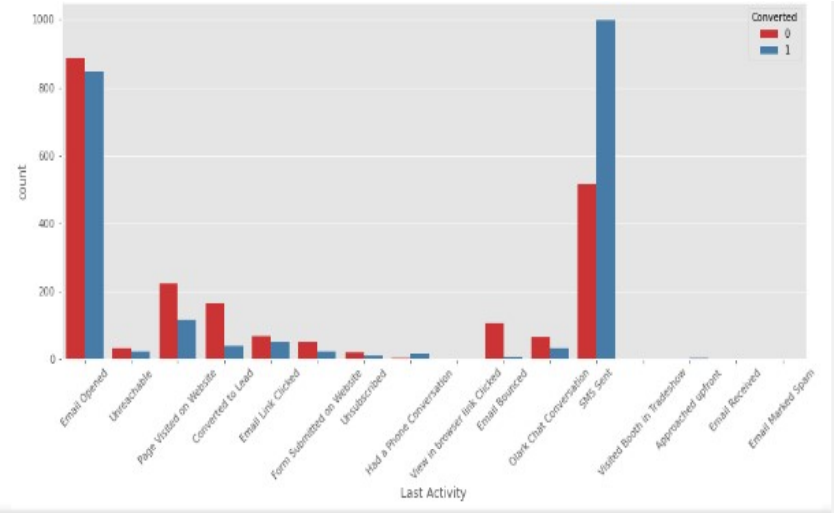
TotalVisits



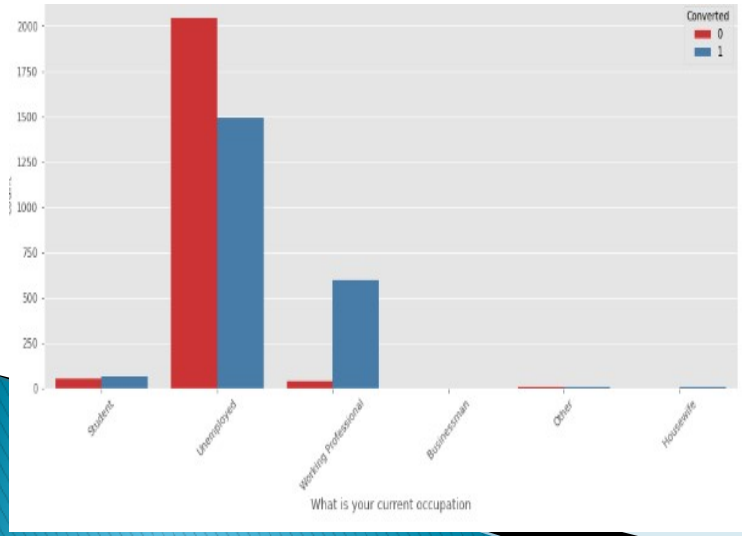
Inference¶
Leads spending more time on the website are more likely to be converted**.**



Inference
•Median for converted and unconverted leads is the same.
•Nothing can be said specifically for lead conversion from Page Views Per Visit

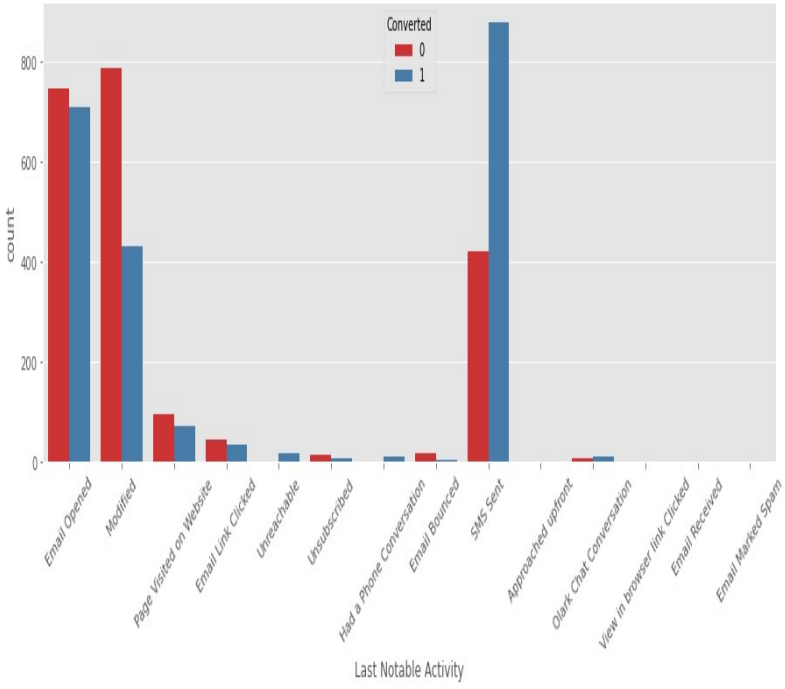


Inference¶
Most of the lead have their SMS SENT as their last activity. Conversion rate for leads with last activity as SMS Sent is almost 100%



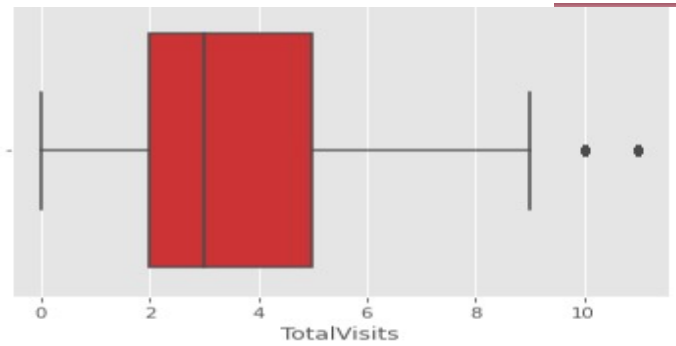
Inference – Most of the leads are unemployed

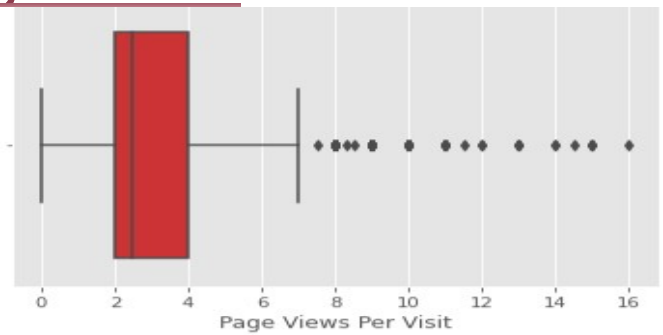
Inferences
Leads refers less free copy of mastering the interview

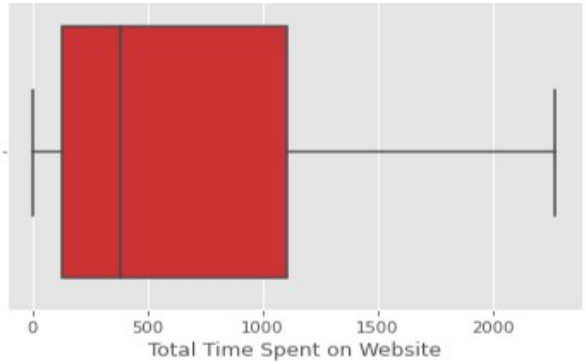


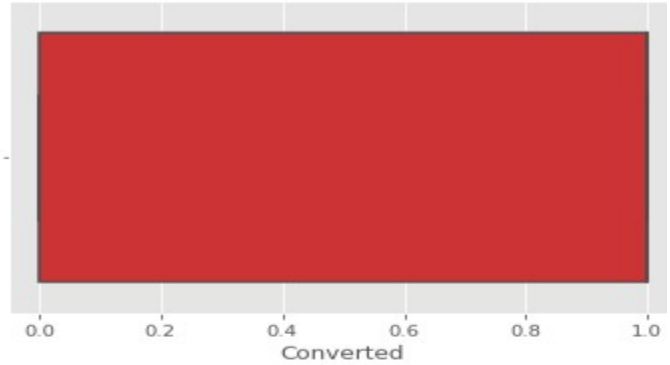
Inference – Last Notable Activity done by the leads are sms sent

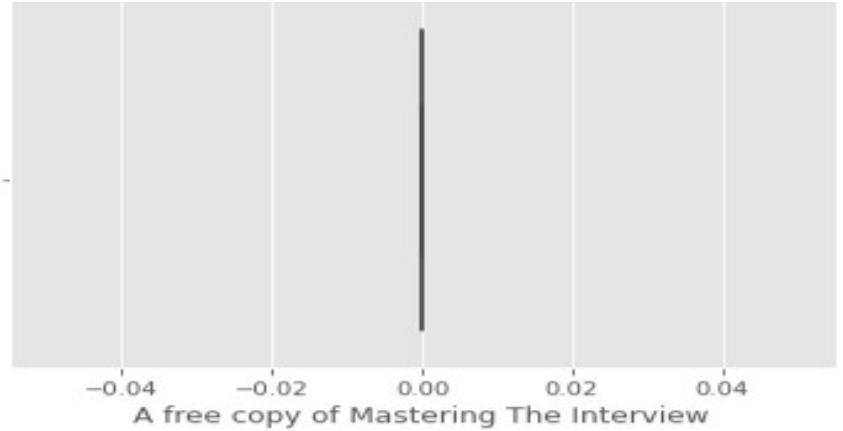
Handling Outliers



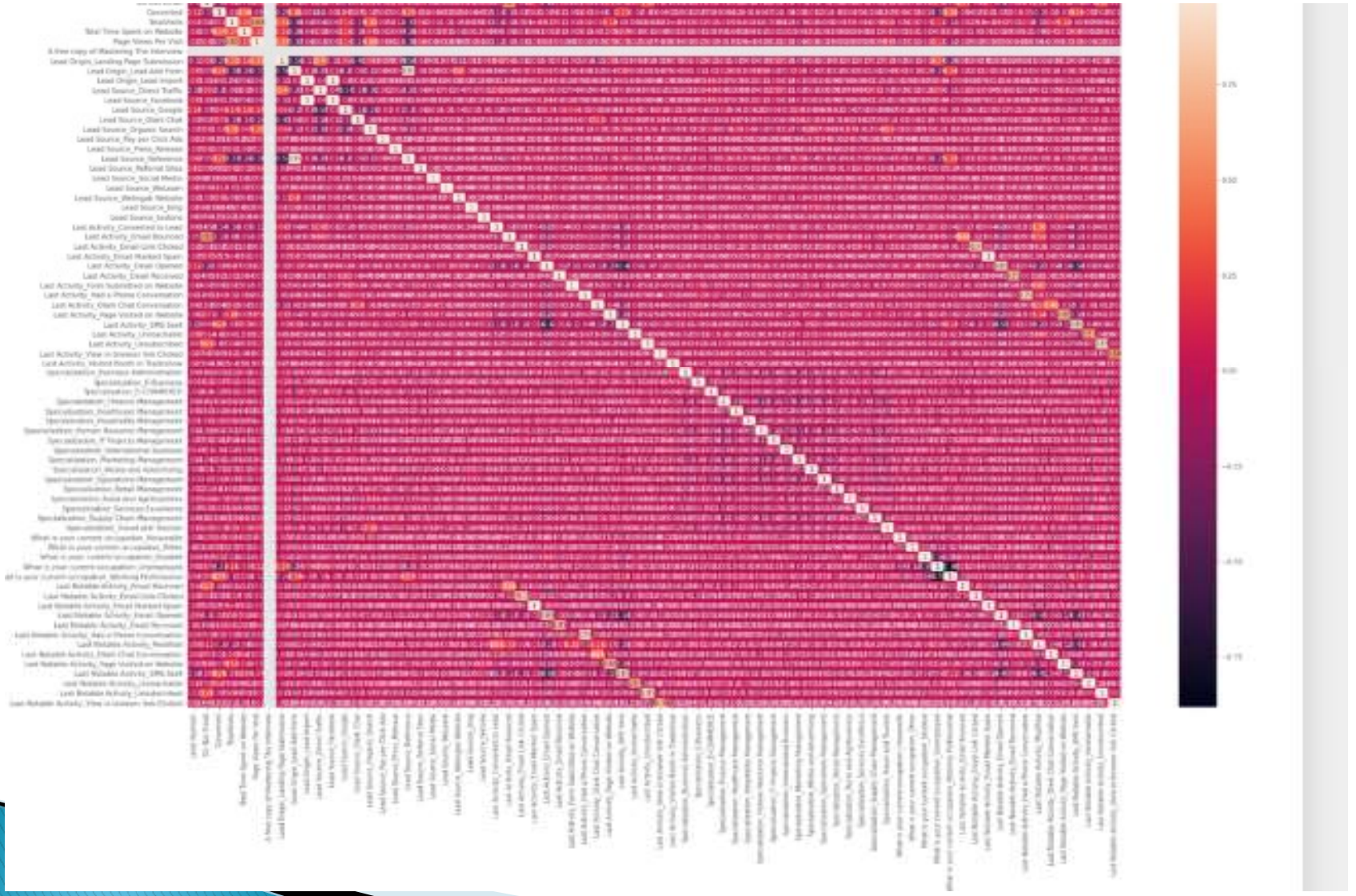




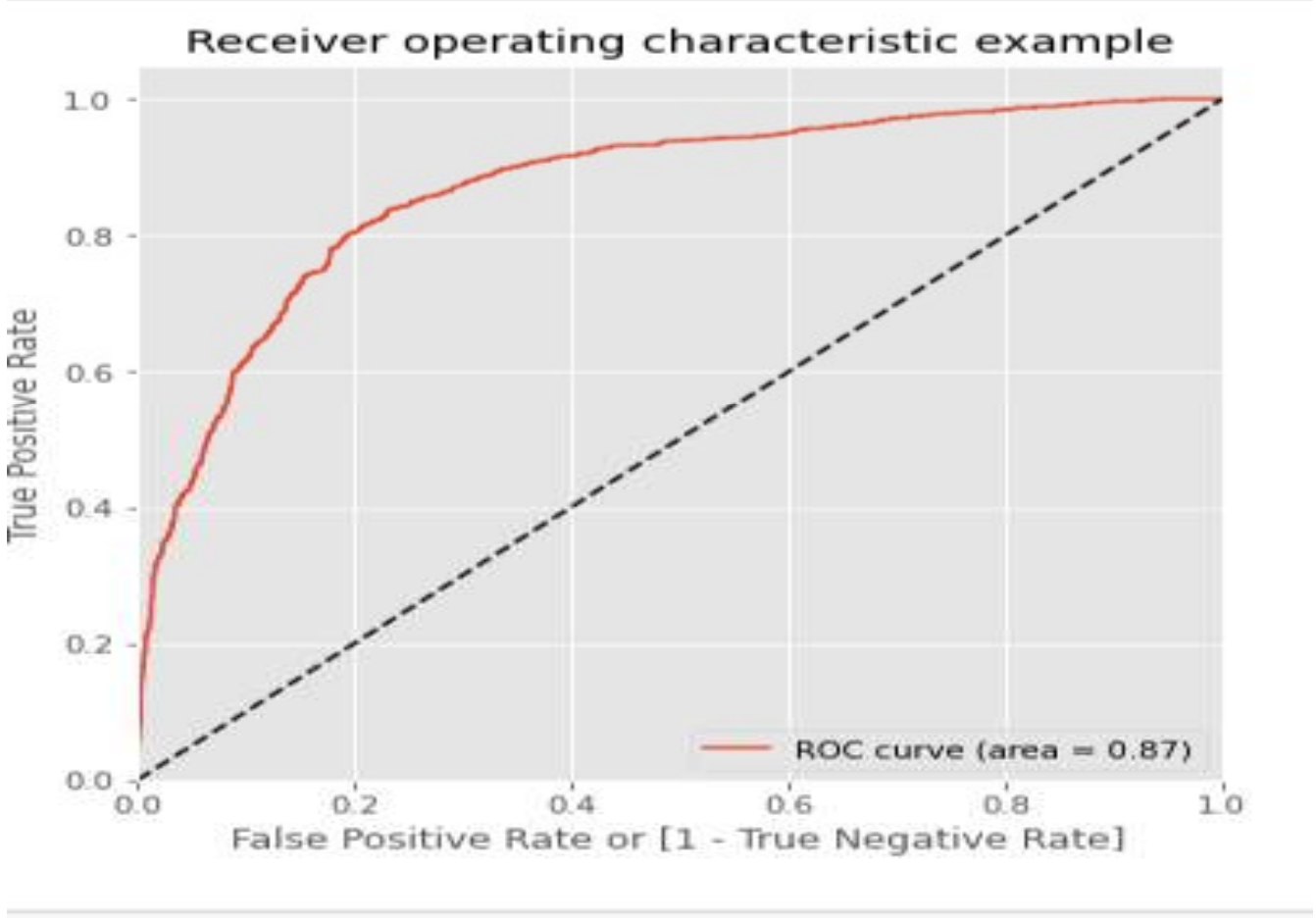




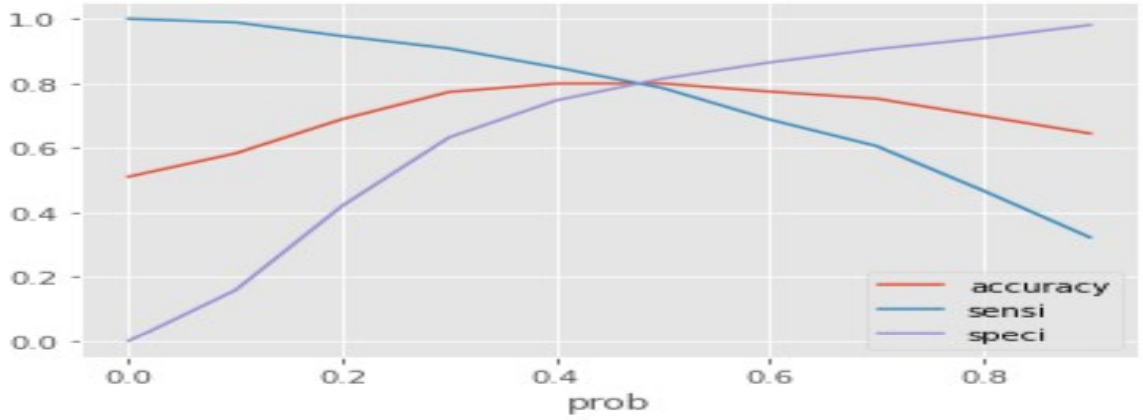
Correlation



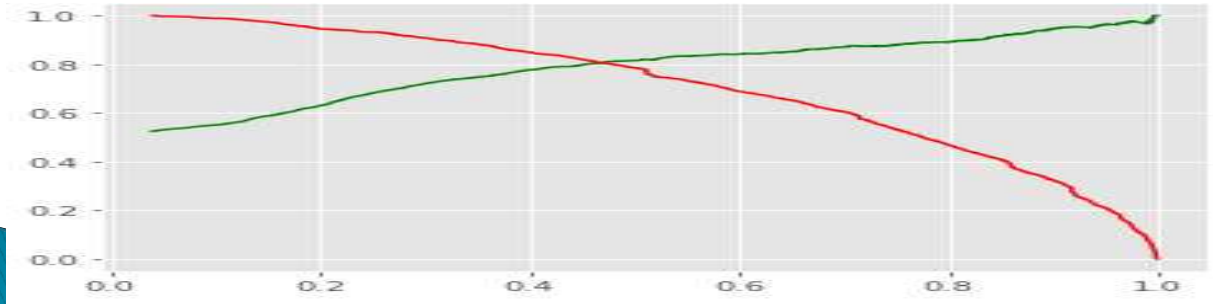
ROC Curve


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.87)
0.0
0.2
0.4
0.6
0.8
1.0

Plotting accuracy sensitivity and specificity for various probabilities

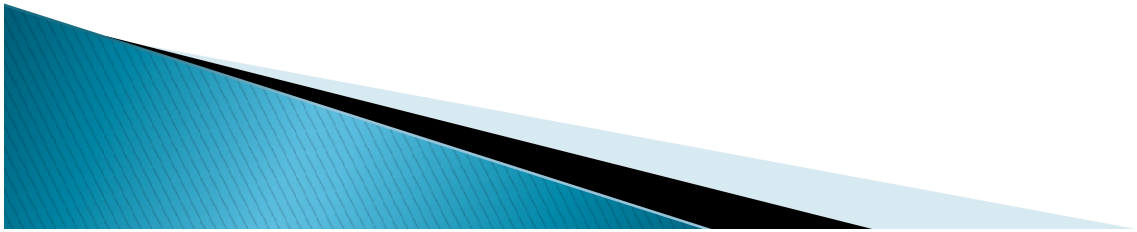


Plotting a trade-off curve between precision and recall



Summary

•The company **should make calls** to the leads coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.
•The company **should make calls** to the leads who are the "working professionals" as they are more likely to get converted
•The company **should make calls** to the leads who spent "more time on the websites" as these are more likely to get converted.
•The company **should make calls** to the leads coming from the lead sources "Olark Chat" as these are more likely to get converted.
•The company **should make calls** to the leads whose last activity was SMS Sent as they are more likely to get converted.
•The company **should not make calls** to the leads whose last activity was "Olark Chat Conversation" as they are not likely to get converted.
•The company **should not make calls** to the leads whose lead origin is "Landing Page Submission" as they are not likely to get converted.
•The company **should not make calls** to the leads whose Specialization was "Others" as they are not likely to get converted.
•The company **should not make calls** to the leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.



THANK YOU